

Data Mining and Predictive Analysis

Intelligence Gathering and
Crime Analysis

Second Edition

Colleen McCue



AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Butterworth-Heinemann is an imprint of Elsevier



Acquiring Editor: Sara Scott
Editorial Project Manager: Marisa LaFleur
Project Manager: Punithavathy Govindaradjane
Designer: Mark Rogers

Butterworth-Heinemann is an imprint of Elsevier
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2015, 2007 Elsevier Inc. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: www.elsevier.com/permissions.

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data

McCue, Colleen.

Data mining and predictive analysis : intelligence gathering and crime analysis / Colleen McCue. -- 2 Edition.
pages cm

ISBN 978-0-12-800229-2

1. Crime analysis. 2. Data mining in law enforcement. 3. Law enforcement--Data processing.
4. Criminal behavior, Prediction of. I. Title.

HV7936.C88M37 2015

363.250285'6312--dc23

2014031816

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN: 978-0-12-800229-2

For information on all Butterworth-Heinemann publications
visit our website at <http://store.elsevier.com/>



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Dedication

**This book is dedicated to Naval Criminal Investigative Service,
Supervisory Special Agent (Ret.) Richard J. McCue, my
partner in crime and everything else that matters.**

Foreword

If ever there was any doubt about the presence of evil in our world, one need only conduct a quick Internet search for Joseph Kony and his Lord's Resistance Army. Over a horrific generation, Kony has marauded his way across central Africa—Uganda, South Sudan, Democratic Republic of Congo, and the Central African Republic—killing and maiming uncounted thousands, many of them children, almost all of them innocents. To its great credit, the government of Uganda, along with its neighbors and a dedicated cohort of international and nongovernmental organizations, has been relentlessly pursuing Kony and his henchmen in an effort to protect local populations from the LRA's attacks and to aid survivors and escapees.

The United States had long, but inconsistently, supported anti-Kony efforts, but that changed with President Obama's signing into law the Lord's Resistance Army Disarmament and Northern Uganda Recovery Act of 2009. And in 2011 Mr. Obama ordered the deployment of approximately 100 U.S. Special Operations personnel to aid in the multinational effort to bring Kony and his top leaders to justice. Among the many challenges, including language, culture, logistical support in remote regions, and many more, was the fundamental difficulty of finding Kony and his top lieutenants in a vast area of Africa—an area roughly the size of the state of Colorado that is densely forested with little infrastructure and even less governmental reach. At United States Africa Command, intelligence analysts and seasoned Foreign Service Officers aggressively sought methodologies and processes to more quickly and accurately predict where and when LRA activities might occur. Traditional pattern analysis and tracking procedures just weren't working. Enter Dr. Colleen McCue.

Dr. McCue's work with the Richmond, Virginia Police Department had demonstrated the value of more detailed, refined predictive analysis. It appeared that her approach might prove useful in a vastly different region and in a military, vice law enforcement endeavor. That approach was quickly proven accurate. Dr. McCue, using the same methodologies as she has so successfully applied in Virginia, was able to help military analysts sift through mounds of data and incident reports in the effort to find the real nuggets of information that

would allow the forces in pursuit of the LRA to predict future attacks and even their heretofore clandestine routes of travel. Within just a few months, using Dr. McCue's methods, Ugandan and American forces were able to interdict LRA routes, deter village attacks, and capture or cause the surrender of several key Kony associates. While Joseph Kony himself remains at large, the results of Dr. McCue's work mean that this notoriously vicious warlord is operating largely in survival mode rather than roaming the region with impunity.

In this new book, an update from her initial 2007 publication, Dr. McCue makes a compelling case for the effectiveness of predictive analysis in a widening array of functional communities. She clearly and concisely lays out the processes she has developed, affording analysts and academics the opportunity to thoroughly assess and examine her work. But, she does so in a way easily understood by operators (like me) who possess neither the academic nor research credentials of those who normally work in this space. It is this aspect of Dr. McCue's writing that appeals to me and, I have found, to others across a wide variety of operational interests—police work, to be sure, but also disaster preparedness and relief specialists, the counter-illicit trafficking community, even those who focus agricultural and medical trends. One can see ready applicability for commercial enterprises as well. Essentially, what Dr. McCue offers is a now well-tested and proven method for decision-makers, private or governmental, to choose how to most effectively apply scarce resources to address a given problem.

Dr. McCue's well-crafted second edition not only provides additional and more current examples of how her processes have been applied operationally in an ever-expanding array of activities, but also addresses how developing concepts and capabilities aid in data mining and predictive analysis. The art of her work lies in the manner in which she takes complex analytical capabilities from the scientific and academic worlds and translates them into real-world issues of understanding and predicting human behavior in support of operational decision makers. It is this blending of analytics with operational experience and expertise that will be of greatest interest to those in law enforcement, military, or other security fields. In short, when operators gain an appreciation of the power of data mining and predictive analysis and when analysts better understand the needs of operators, a synergy is obtained that benefits all (well, maybe not criminals or terrorists). The essence of Dr. McCue's work is to translate science into meaningful action and she makes a powerful case for doing so.

General Carter Ham
U.S. Army (Retired) Former Commander, U.S. Africa Command

Preface

So many things have changed since the first edition of this textbook, particularly as relates to data, technology, and tradecraft. Some things have not changed, however, including my love of science and desire to develop innovative solutions to some of the really challenging public safety and security challenges. Operational security analytics, at its core, is designed to effectively characterize bad behavior in support of information-based approaches to anticipation and influence. Whether “influence” entails prevention, thwarting, mitigation, response, or consequence management, we are trying to change outcomes for the better.

In the beginning of my operational security analytics journey, I became profoundly intrigued by how many of the seasoned detectives I worked with were often able to generate quick yet accurate hypotheses about their cases, sometimes only moments after they had arrived at the scene. Like the “profilers” on television and in the movies, many of them seemed to have an uncanny ability to accurately describe a likely motive and related suspect based merely on a review of the crime scene and some preliminary knowledge regarding the victim’s lifestyle and related risk factors. Over time, I started to acquire this ability as well, although to a lesser degree. It became much easier to read a report and link a specific incident to others, predict future related crimes, or even calculate the likelihood that a particular case would be solved based on the nature of the incident. Drawing on my training as a scientist, I frequently found myself looking for some order in the chaos of crime, trying to generate testable hypotheses regarding emerging trends and patterns, as well as investigative outcomes. Sometimes I was correct. However, even when I was not, I was able to include the information in my ever-expanding internal rule sets regarding crime and criminal behavior.

Prior to working for the Richmond Police Department, I spent several years working with that organization. Perhaps one of the most interesting aspects of this early relationship with the Department was my weekly meeting with the officer in charge of violent crimes. Each week we would discuss the homicides from the previous week, particularly any unique or unusual behavioral

characteristics. Over time, we began to generate casual predictions of violent crime trends and patterns that proved to be surprisingly accurate. During the same time period, I also began to examine intentional injuries among incarcerated offenders. As I probed the data and drilled down in an effort to identify potentially actionable patterns of risk, it became apparent that many of the individuals I looked at were not just in the wrong place at the wrong time, as they frequently indicated. Rather, they were in the wrong place at the wrong time *doing the wrong things with the wrong people* and were assaulted as a result of their involvement in these high-risk activities. As I explored the data further, I found that different patterns of offending were associated with different patterns of risk. This work had immediate implications for violence reduction efforts. It also had implications for the analysis of crime and intelligence data. Fortunately, the field of data mining and predictive analytics had evolved to the point that many of the most sophisticated algorithms were available in a PC environment, so that everyone from a software-challenged psychologist like myself to a beat cop could begin to not only understand but also use these incredibly powerful tools.

Although I did not realize it at the time, a relatively new approach to marketing and business intelligence was emerging at the same time we were engaging in this lively speculation about crime and criminals at the police department. Professionals in the business community were exploiting artificial intelligence and machine learning to characterize and retain customers, increase sales, focus marketing campaigns, and perform a variety of other business-related tasks. For example, each time I went through the checkout counter at my local supermarket, my purchasing habits were coded, collected, and analyzed. This information was aggregated with data from other shoppers and employed in the creation of models about purchasing behavior and how to turn a shopper into a buyer. These models were then used to gently mold my future behavior through everything from direct marketing based on my existing preferences to the strategic stocking of shelves in an effort to encourage me to make additional purchases during my next trip down the aisle. Similarly, data and information were collected and analyzed each time I perused the Internet. As I skipped through web pages, I left cookies, letting the analysts behind the scenes know where I went and when and in what sequence I moved through their sites. All of this information was analyzed and used to make their sites more friendly and easier to navigate or to subtly guide my behavior in a manner that would benefit the online businesses that I visited. The examples of data mining and predictive analytics in our lives are almost endless, but the contrast between my professional and personal lives was profound. Directly comparing the state of public safety analytical capacity to that of the business community only served to underscore this shortcoming. Throughout almost every aspect of my life, data and information were being collected on me and analyzed using

sophisticated data mining algorithms; however, the use of these very powerful tools was severely limited or nonexistent in the public safety arena in which I worked. With very few exceptions, data mining and predictive analytics were not readily available for the analysis of crime or intelligence data, particularly at the state and local levels.

Like most Americans, I was profoundly affected by the events of September 11th. In the week of September 10th, 2001, I was attending a specialized course in intelligence analysis in northern Virginia. Like many, I can remember exactly what I was doing that Tuesday morning when I saw the first plane hit the World Trade Center and how I felt as the horror continued to unfold throughout the day. As I drove back to Richmond, Virginia that afternoon (the training had been postponed indefinitely), I saw the smoke rise up over the Beltway from the fire at the Pentagon, which was still burning. Those of us working in the public safety community were inundated with information over the next several days, some of it reliable, much of it not. Like many agencies, we were swamped with the intelligence reports and BOLOs (be on the lookout reports) that came in over the teletype, many of which were duplicative or contradictory. Added to that were the numerous suspicious situation reports from concerned citizens and requests for assistance from the other agencies pursuing the most promising leads. Described as the “volume challenge” by former CIA director George Tenet, the amount of information threatened to overwhelm us. Because of this, it lost its value. There was no way to effectively manage the information, let alone analyze it. In many cases, the only viable option was to catalog the reports in three-ring binders, with the hope that it could be reviewed thoroughly at some later date. Like others in law enforcement, our lives as analysts changed dramatically that day. Our professional work would never again be the same. In addition to violent crimes and vice, we now have the added responsibility of analyzing data related to the war on terrorism and the protection of homeland security, regardless of whether we work at the state, local, or federal level. Moreover, if there was one take-home message from that day as an analyst, particularly in Virginia, it was that the terrorists had been hiding in plain sight among us, sometimes for years, and they had been engaging in a variety of other crimes in an effort to further their terrorist agenda, including identity theft, forgery, and smuggling; not to mention the various immigration laws they violated. Many of these crimes fall within the purview of local law enforcement.

As we moved through the days and weeks following the attacks, I realized that we could do much better as analysts. The subsequent discussions regarding “connecting the dots” highlighted the sad fact that quite a bit of information had been available before the attacks; however, flaws in the sharing and analysis of information resulted in tragic consequences. Although meaningful information sharing remains an important goal, advanced analytical techniques

are available now. The same tools that were being used to prevent people from switching their mobile telephone service provider and to stock shelves at our local supermarkets on September 10th can be used to create safer, healthier communities and enhance homeland security. The good news is that these techniques and tools are being used widely in the business community. The key is to apply them to questions or challenges in public safety, law enforcement, and intelligence analysis.

Again, I thoroughly enjoy science and particularly like the new concept of “data science,” which really captures the creative aspects of analysis and associated promise of transdisciplinary approaches. As someone who likes to color outside the lines and explore novel approaches to analysis, I am intrigued by the use of advanced analytics to improve other aspects of my life and see data science as a means to an end; as a means by which to better understand behavior—good, bad and otherwise—so that we can use it to anticipate and influence outcomes, particularly in support of enhanced public safety and security. Almost everything in my professional life for the previous 20 years has been in direct support of that mission. The second edition of this textbook is no exception.

Although I say “I” quite a bit in this book, it certainly was not created in a vacuum. Countless individuals have helped me throughout my career, and a few have truly inspired me. What follows is a very brief list of those that contributed directly to this effort in some way.

I am tremendously honored by General Carter Ham’s willingness to write the Foreword to the second edition. General Ham has been a great mentor and guide, particularly as relates to improving my understanding of the challenges facing the people of Africa. Our work modeling violent extremism in Africa has been some of the most rewarding for me professionally. The ability to successfully apply western models of crime analysis to the Lord’s Resistance Army (LRA), underscores the importance of foundation-level concepts in understanding violent crime and other predatory behavior; concepts that will enable us to effectively respond to other challenging situations, including those that have not yet emerged. This particular problem space is complex and there will be no easy solutions; however, the saying “African solutions to African problems” reinforces the importance of a local approach in support of meaningful and sustainable answers to some of our hardest problems. Moreover, the more that I learn about Africa, the more that I see parallels, not only in our understanding of challenging behavior, but also in the importance of local solutions to problems in other communities struggling with violence, including those in the United States.

I would like to thank Pam Chester from Elsevier for originally approaching me about a second edition. Marisa LaFleur, my new editor, has brought a fresh

perspective and approach, which has been a great benefit. Nancy Coleman and Turner Brinton from DigitalGlobe, and Brian Wagner from McBee Strategic Consulting, have provided great insight and guidance regarding the importance of narrative and context in conveying the critical points in the new case material.

Most of the early work referenced came out of some very lively discussions that began several years ago with my colleagues at the Federal Bureau of Investigation. In particular, Supervisory Special Agents Charlie Dorsey and Dr. Wayne Lord provided considerable guidance to my early research. Over time, they have become both colleagues and friends, and my work definitely reflects a level of quality that is attributable directly to their input. Also with the FBI, Mr. Art Westveer taught me almost everything that I know about death investigation. I learned a tremendous amount from his lectures, which were punctuated with his dry sense of humor and wonderful anecdotes from a very successful career with the Baltimore Police Department. His untimely passing was a significant loss to our community. Rich Weaver and Tim King graciously allowed me to attend their lectures and training at International Training, Inc. on surveillance detection in support of my research. They also provided some very unique opportunities for field testing many of my ideas in this area to see how well they would play in the real world.

Although many of my former employers merely tolerated my analytical proclivities, the Project Safe Neighborhoods folks provided funding, as well as ongoing support and encouragement for much of the early work outlined in this book. In particular, Paul McNulty, the United States Attorney for the Eastern District of Virginia, carried the message of our success far beyond the audience that I could reach alone.

I also would like to thank Dr. Harvey Sugarman. I still remember the day when he called me out of the blue and told me that he thought that I should be paid for the work I had been doing. A single mother, I had been responding to homicide calls on my own time in the evenings in an effort to gain additional knowledge and insight into violent crime and the investigative process. That particular act made a tremendous positive impact in my life. I gained invaluable experience through my affiliation with the university, but his gentle mentoring and decision to offer me compensation for my work only begins to underscore the kindness in his heart.

I owe a tremendous debt of gratitude to the software and consulting companies that provided me with excellent case study material, without which the second edition would be very thin and not terribly interesting. In particular, David Korn, Allen Sackadorf, and John Tomaselli from SAP NS2; Kevin Merguen, and Ted Desaussure from Information Builders; Bill Wall from Praescent Analytics; Dr. Rick Adderly from A E Solutions (BI) Ltd.; Dave Roberts from the

International Association of Chiefs of Police (IACP); Frank Stein from IBM and Sarah Dunworth from SPSS/IBM; and Mark Moorman and Trent Smith from SAS. I also am very grateful to former Richmond PD, and current Charlotte-Mecklenburg PD Chief Rodney Monroe, and his crime analysis lead, Crystal Cody for making their crime analysis data and results available to me to illustrate the value of advanced analytics in the operational law enforcement environment. Chief Monroe has the unique distinction of implementing this model in two different organizations. It is an amazing task given that many agencies are still trying to do it once. I would also like to thank Special Agent BJ Kang and USMC Staff Sergeant Tom Ferguson for giving me permission to use their photographs throughout this book. Their photographs graphically illustrate our recent history as a nation and serve to further underscore the importance of fighting the good fight, and doing so with honor.

My former colleagues at the Richmond, Virginia Police Department taught me almost everything that I know about police work and law enforcement. To name every individual that has contributed to my training and life would resemble a roll call of the current and previous command, as well as the line staff, who frequently know as much if not more than their supervisors. In particular, I would like to thank Jerry Oliver, the former Chief of the Richmond Police Department, who, with Teresa Gooch, recruited me for the most rewarding yet challenging position I have ever enjoyed. I also owe a tremendous debt to the Virginia Homicide Investigators Association, where I received some outstanding training in death investigation and was fortunate enough to meet my husband. My colleagues in law enforcement have taught me as much, if not more, about life in the many years that I worked with them. In particular, the late Major Hicks was fond of referring to every challenge as a tremendous opportunity to succeed; wisdom that I continue to carry with me. I have seen some truly innovative solutions to some exceptionally hard problems with no easy or obvious way out. Finding those answers and being able to change outcomes is the really rewarding part of this work.

Since completing the first edition I have worked on some very significant projects with exceptionally talented individuals and great teams including the ORSA cell at the COIC; Dr. John Elder and his team at ERI; and Dave Porter, a colleague from Detica and Now SAS. Currently, I have the great pleasure of working with Ken Campbell, the original leader of the "Four Horsemen of the Analysis," Wes Hildebrandt, and Jim Stokes, who originally recruited me to join the team and to whom I will always be grateful for doing so. I am also extremely grateful to my colleagues Curran Runz, Matt Molumby, and Jimbob Skelton, the real domain experts who did most of the recent work outlined in this text. Finally, I continue to serve under inspiring leadership including Mark Dumas (SPADAC); Matt O'Connell (GeoEye); and Jeff Tarr, Walter Scott, and Tony Frazier (DigitalGlobe).

In many ways, my husband, Supervisory Special Agent (Ret.) Rick McCue, has contributed more than enough to earn the right to be a co-author. Through him, I have gained first-hand insight into the needs of operational personnel and the importance of making analytical products accessible to the folks that need them the most: those on the front lines. Whether with outright encouragement or a vacant stare when I became long-winded or obtuse, he has provided invaluable guidance to my skills as an analyst. I am forever grateful for the experiences that I have had vicariously through my husband. As one of the team assigned to the Pentagon recovery immediately after September 11th, my husband saw first-hand the devastation that the terrorist agenda can rain down on innocent lives. I know that neither of us will ever be the same. In his subsequent missions overseas, I began to truly understand the value that good intelligence and analysis can bring to the operational mission.

Our children continue to keep me humble and remind me daily what is most important in life. Like many folks in public safety, there have been more than a few times that I have come home and hugged my children a little bit harder because of what I have seen or done at work. I am so grateful to be blessed with such a wonderful life and family, which makes me work that much harder for those who are not. I believe that other women love their children just as I do. Unfortunately, too many of their children will not be coming home again. Whether it is the result of drugs, gang violence, violent extremism, or the war on terrorism, there is too much pain and suffering in our world, too much killing. For that reason, as a homicide researcher, it always has been important for me to remember that every one of the “subjects” in my studies is a lost life, a devastated family, and a loss to our community, whether local or global. In all humility, it is my sincere wish that the techniques and approaches outlined in this book will help us improve the health and well-being of our communities and create safer neighborhoods for all of our children.

“If there must be trouble, let it be in my day, that my child may have peace.”

Thomas Paine

Colleen McLaughlin McCue, PhD

Digital Assets

Thank you for selecting Butterworth-Heinemann's *Data Mining and Predictive Analysis, Second Edition*. To complement the learning experience, we have provided a number of online tools for instructors to accompany this edition.

Please consult your local sales representative with any additional questions.

FOR THE INSTRUCTOR

Qualified adopters and instructors need to register at the following link for access: <http://textbooks.elsevier.com/web/manuals.aspx?isbn=9780128002292>

- *Test bank* composes, customizes, and delivers exams using an online assessment package in a free Windows-based authoring tool that makes it easy to build tests using the unique multiple choice and true or false questions created for *Data Mining and Predictive Analysis*. What's more? This authoring tool allows you to export customized exams directly to Blackboard, WebCT, eCollege, Angel, and other leading systems. All test bank files are also conveniently offered in Word format.
- *PowerPoint lecture slides* reinforce key topics with focused PowerPoint presentations, which provide a perfect visual outline with which to augment your lecture. Each individual book chapter has its own dedicated slideshow.
- *Instructor's manual* designs your course around customized learning objectives, critical thinking questions, and key terms.

Introduction

Good analysts are like sculptors. They can look at a data set and see the underlying form and structure. Data mining tools can function as the chisels and hammer, allowing the analysts to expose the hidden patterns and reveal the meaning in a data set so that others can enjoy its composition and beauty.

Whether it is called data mining, predictive analytics, sense making, knowledge discovery, or data science, the rapid development and increased availability of advanced computational techniques have changed our world in many ways. There are very few, if any, electronic transactions that are not monitored, collected, aggregated, analyzed, and modeled. Data are collected about everything, from our financial activities to our shopping habits. Even casino gambling is being analyzed and modeled in an effort to characterize, predict, or modify behavior.

One area that has been somewhat limited in its acceptance and use of these powerful new techniques is the public safety community, particularly in crime analysis and operations. This is somewhat surprising because, in many ways, analysts, detectives, agents, professionals in the intelligence community, and other operational personnel embody many of the principles of data mining or knowledge discovery. For example, the process of training detectives in investigative techniques and practices bears a strong resemblance to case-based reasoning.¹ In addition, the characterization, modeling, and prediction associated with the behavioral analysis of violent crime are very similar to the categorization, linking, and anticipatory intelligence associated with data mining and predictive analytics.

Although the relationship between the two areas seems to be natural, the law enforcement community in particular has not enjoyed many of the analytical benefits from these powerful new tools. It is unclear whether this is due to cost, training, or just a lack of knowledge of the existence and availability of these tools, but when they are adopted the increased quality of life for law enforcement personnel, as well as the communities that they serve, is remarkable. In these times of dwindling economic and personnel resources, no

agency can afford to deploy carelessly. As organizations compete for qualified personnel, a candidate's final decision often comes down to quality-of-life and job satisfaction issues. Just a few of the questions potential employees ask themselves before making a final decision are: Will I have a reasonable work schedule? Will I be able to manage my workload effectively? Will my time be used productively? Can I make a difference in my community? Similar decision processes are associated with maintaining a satisfied work force and long-term retention—something that is increasingly difficult, given the rapidly emerging employment opportunities for law enforcement personnel.

At the same time, requirements for accountability and outcome studies are coming from funding agencies and constituents alike. It is no longer acceptable to run programs without the outcome indicators and metrics necessary to demonstrate their efficacy. The emphasis on these measures of accountability highlights the need for new methodologies to document progress and change in response to new initiatives and strategies.

Given the infinitely increasing amounts of information, “connecting the dots” will be possible only with automated systems. Perhaps addressing the gaps in information and information sharing will be more important than trying to create these associations. Only after these challenges have been addressed will we be able to identify and characterize trends and patterns so that future events can be predicted, anticipated, and perhaps even thwarted or prevented. The emphasis needs to shift from describing the past—counting, reporting, and “chasing” crime—to anticipation and influence in support of prevention, thwarting, mitigation, response, and informed consequence management. Only then will we have the possibility to enhance public safety and create safe neighborhoods for all.

SKILL SET

Analysts are deluged with information on a daily basis. The ability to bring some order into this informational chaos can have a huge impact on public safety and the quality of life in the communities that they serve. On the other hand, the opportunity to bring analytical and predictive models directly into the operational environment holds the promise of giving public safety and intelligence professionals the ability to maneuver within the decision and execution cycles of their opponent. Whether it is the war on terrorism, the war on drugs, or the war on crime, enhanced knowledge and the ability to anticipate future actions can afford operational personnel essential situational awareness, while providing unique opportunities to influence outcomes.

Knowledge of advanced statistics is not a prerequisite for using predictive analytics. In fact, the discovery process associated with data mining also could

be viewed as after-the-fact explanations for unpredicted outcomes, something somewhat distasteful in inferential statistics. When examined under the intense scrutiny of the analyst's domain knowledge, however, these unanticipated or surprising findings can have significant value and greatly enhance our understanding of crime and intelligence data. For those who are analytically inclined, it can be a wonderful and exciting process of data exploration and discovery. Those with a strong background in statistics, though, might be somewhat handicapped by the comparatively rigid nature of inferential statistics, with all of its associated rules and assumptions. With a little confidence and practice, even statisticians will be able to overcome their previous training and perform what they once considered to be unnatural acts with data and information.

On the other hand, data mining brings powerful analytics to those who really need them, including operational personnel. In my experience, it is far easier to teach someone with interest who knows something about crime and criminals how to effectively use these tools. With some guidance regarding a few "rules of the road" for data mining, and the application of off-the-shelf software tools, data mining is well within the reach of any organization with an interest and willingness to put more science and less fiction into crime and intelligence analysis. Moreover, many of the new tools and service delivery models, including managed service, software as a service, and cloud computing, can be deployed in a web-based environment and are no more difficult than making a purchase or completing a survey over the Internet. These advancements have created the opportunity for "24/7" analytical capacity,² even within smaller agencies with comparatively limited personnel resources.

The more operational personnel, managers, and command staff understand the information requirements and possible outcomes from analytical products, the more likely they will be to contribute data that is meaningful, detailed, and valuable. They also will be in a better position to work with the analyst and participate in the analytical process, requesting output that has increased value for them as they acquire a better understanding of what is available. By understanding the importance of the data inputs and the potential range of outputs, operational personnel, managers, and command staff alike can become informed information consumers and increase the likelihood of identifying actionable output from the analytical process. This subtle change in relationships and understanding can greatly enhance analysts' ability to gather the necessary data and information, ultimately increasing their ability to support operational personnel, policy decisions, managers, and command staff.

At a security expo several years ago, the author Tom Clancy advised the security and intelligence professionals in the audience to seek out the "smart people," observing that, "[t]he best guys are the ones who can cross disciplines ... [t]he

smartest ones look at other fields and apply them to their own.”³ In my opinion, this advice still holds and many of the “smart people” Clancy refers to will rise out of the operational ranks, given the intuitive nature and relative ease of use associated with the new generation of data mining and predictive analytics software tools. Although most analysts probably do not need to fear for their jobs just yet, increasingly friendly and intuitive capabilities will allow data and information to serve as a fluid interface between analytical and operational personnel. At some point in the future, that distinction will become almost meaningless with the emergence of increasingly powerful software tools and systems and the “agent/analysts” that employ them.

HOW TO USE THIS BOOK

The examples included in this book come largely from real experience. In some cases, though, the specifics have been changed to protect ongoing investigations, sensitive data, or methods. Whenever possible, I have tried to distinguish between real cases, particularly those taken from published work, and those generated specifically as examples. Given the nature of some topics covered in this book, however, it would be inappropriate to provide too much specific detail and compromise sources and methods. To be sure, though, while the names might have been changed to protect the “not so innocent,” the examples are based on real scenarios and experience.

This book is divided into five main sections: “Introduction,” “Methods,” “Applications,” “Case Examples,” and “Advanced Concepts/Future Trends.” The third and fourth sections include annotated examples focusing on the why and how, as well as the limitless possibilities for data mining and predictive analytics in crime and intelligence analysis. Although this organization is relatively logical for training purposes, many readers will choose to read the book out of sequence. In particular, managers, command staff, supervisors, policy makers, and operational personnel interested in learning more about data mining and predictive analytics but not expecting to use these tools first hand will have neither an interest in nor a need for detailed information on specific methods and algorithms. These readers could benefit from reading and understanding the annotated examples if they make acquisition and purchasing decisions for analytical products and determine the focus of their analytical personnel. Moreover, operational personnel can more fully exploit the new technology and work more effectively with analytical personnel if they understand the vast array of possibilities available with these new tools. With the opportunity to deploy data mining and predictive analytics directly into the field, an increasing number of operational personnel will be using the results of data mining and predictive analytics to support operations. Although they might not be generating the specific algorithms or models, a general understanding of data

mining and predictive analytics will certainly enhance their ability to exploit these new opportunities.

Similarly, many analysts will use this book to explore the possibilities for data mining in their environment, identifying ideas and strategies from the annotated examples in the third section, and then returning to the methods section for specific information regarding the use and implementation of these approaches. This book is not intended to provide detailed information about specific software packages or analytical tools, but to merely provide an overview of them. It should serve as a starting point, using terminology, concepts, practical application of these concepts, and examples to highlight specific techniques and approaches in crime and intelligence analysis that employ data mining and predictive analytics, which each law enforcement or intelligence professional can tailor to their own unique situation and responsibilities. Although the basic approaches will be similar, the available data, specific questions, and access to technology will differ for each analyst and agency, requiring unique solutions and strategies in almost every setting.

Perhaps one of the most challenging aspects of writing this book was keeping abreast of the new developments and data mining applications that now appear on an almost daily basis. It is both frustrating and exciting to consider how much this field is likely to change even in the short time between completion of the manuscript and actual publication of the text. Therefore, the final section, "Advanced Concepts/Future Trends," should not be viewed as inclusive. Rather, this particular section is intended to serve as a beginning for ascending to the next level of training for those interested in this field. This rapid pace of innovation, however, is what keeps the field of analysis fresh and exciting, particularly for those with the interest and creativity to define the cutting edge of this new and evolving field.

Bibliography

- 1 Casey E. Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and internet crime investigation. Knowledge Solutions. www.corpus-delicti.com/case_based.html; 2002.
- 2 McCue C, Parker A. Web-based data mining and predictive analytics: 24/7 crime analysis. *Law Enforcement Technology* 2004;31:92–99.
- 3 Fisher D. Clancy urges CIOs: seek out the "smart people." *eWeek*, www.eweek.com; 2003.

Basics

“There are three kinds of lies: lies, damned lies, and statistics.”

Benjamin Disraeli (1804–1881)

1.1 BASIC STATISTICS

Some of my earliest work using data mining and predictive analytics on crime and criminals employed the use of relatively advanced statistical techniques that yielded very complex models. While the results were analytically sound, and even of interest to a very small group of similarly inclined criminal justice and forensic scientists, the outcomes were so complicated and arcane that they had very little utility to those who needed them most, particularly those on the job in the public safety arena. Ultimately, these results really contributed nothing in a larger sense because they could not be translated into the operational environment. My sworn colleagues in the law enforcement world would smile patiently, nodding their heads as if my information held some meaning for them, and then politely ask me what it really meant in terms of catching bad guys and getting the job done. I rarely had an answer. Clearly, advanced statistics was not the way to go.

Data mining, on the other hand, is a highly intuitive, visual process that builds on an accumulated knowledge of the subject matter, something also known as domain expertise. While training in statistics generally is not a prerequisite for data mining, understanding a few basic principles is important. To be sure, it is well beyond the scope of this book to cover statistics with anything more than a cursory overview; however, a few simple “rules of the road” are important to ensure methodologically sound analyses and the avoidance of costly errors in logic that could significantly confound or compromise analysis and interpretation of the results. Outlined here are some simple statistical terms and concepts that are relevant to data mining and predictive analytics, as well as a few common pitfalls and errors in logic that a crime analyst might encounter. These are by no means all inclusive, but they should get analysts thinking and adjusting the way that they analyze and interpret data in their specific professional domain.

1.2 INFERENCE VERSUS DESCRIPTIVE STATISTICS AND DATA MINING

Descriptive statistics, as the name implies, is the process of categorizing and describing information. Inferential statistics, on the other hand, includes the process of analyzing a sample of data and using it to draw inferences about the population from which it was drawn. With inferential statistics, we can test hypotheses and begin to explore causal relationships within data and information. In data mining, we are looking for useful trends, patterns, or relationships in the information. Therefore, data mining more closely resembles descriptive statistics. Predictive analytics, on the other hand, can be seen as a logical extension or complement to the data mining process, and can be used to create models, particularly those that can be used to anticipate or predict future events, surface and characterize hidden patterns or trends, and provide meaningful insight in support of anticipation and influence.

It was not that long ago that the process of exploring and describing data, descriptive statistics was seen as the necessary though unglamorous prerequisite to the more important and exciting process of inferential statistics and hypothesis testing. In many ways, though, the creative exploration of data and information associated with descriptive statistical analysis is the essence of data mining, a process that, in skilled hands, can open new horizons in data and our understanding of the world. For example, summative content analysis is a descriptive technique used for the analysis of unstructured narrative that includes the counting and comparison of key words or content.¹ Figure 1.1 illustrates the results of this technique when applied to media content associated with the Tahrir Square protests during the Arab Spring.² Through the additional segmentation of English language sources into the period immediately before and after the resignation of President Hosni Mubarak, the analyst can compare and contrast perspective and viewpoint, as reflected by language frequency and use.

1.3 POPULATION VERSUS SAMPLES

It would be wonderful if we could know everything about everything and everybody, and have complete access to all of the data that we might need to answer a particular question about crime and criminals. If we had access to every criminal, both apprehended and actively offending, we would have access to the entire *population* of criminals and be able to use population-based statistics. Similarly, if we had access to all of the information of interest, such as every crime in a particular series, this also would resemble a population because it would be all inclusive. Obviously, this is not possible, particularly given the nature of the subject and the questions. It is a common joke that everything

(a)

	Jan. 25, 2011 to Feb. 11, 2011	Feb. 12, 2011 to May 23, 2011	Total
Al-Ahram	25	92	117
Al Jazeera articles	7	27	34
Al Jazeera blog posts	19	61	80
Al-Masry Al-Youm	168	218	386
Antiwar.com article collections	241	166	407
Antiwar.com viewpoint collections	25	10	35
The New York Times	122	225	347
Voice of America	69	38	107
Total	676	837	1513

(b)

Gunfire	Ahram	Al Jazeera	Al Jazeera blogs	Al -Masry Al-Youm	Antiwar articles	Antiwar viewpoints	New York Times	VOA news
Ahram	0.00	0.00	2.58	-1.07	1.84	0.10	1.69	1.68
Al Jazeera		0.00	0.00	0.00	0.00	0.00	0.00	0.00
Al Jazeera blogs			-1.34	-3.65	-0.74	-2.48	-0.89	-0.89
Al-Masry Al-Youm				0.00	2.91	1.17	2.76	2.75
Antiwar articles					-0.83	-1.74	-0.15	-0.15
Antiwar viewpoints						0.00	1.59	1.58
New York Times							0.36	-0.01
VOA news								0.00
All other sources	-1.53	0.00	1.18	-2.78	0.59	-1.45	0.24	0.20
Early count	1	0	13	2	51	2	9	10
Late count	0	0	5	0	13	0	22	0
Total count	1	0	18	2	64	2	31	10

FIGURE 1.1

The results of summative content analysis illustrate the insight that can be generated through the use of descriptive statistics and meaningful segmentation of media content related to the Tahrir Square protests during the Arab Spring.²² The results in panel (a) depict the various English-language media sources used for the analysis, and related frequency counts for documents including reports on the Tahrir Square protests. These reports were further segmented into two date ranges: before and after the resignation of President Hosni Mubarak (January 25, 2011 to February 11, 2011 and February 12, 2011 to May 23, 2011, respectively). The results depicted in panel (b) illustrate differential usage of the word “Gunfire” as represented by frequencies and the related odds ratios between sources for the two reporting periods (“Early count” is before the resignation; “Late count” is after the resignation). *DigitalGlobe, used with permission.*

that we know about crime and criminals is based on the unsuccessful ones, those who got caught. Most criminal justice research is based on correctional populations, or offenders that have some sort of relationship with the criminal justice system. Related to this, most of what we know about terrorists and terrorist groups comes from those who have been caught, infiltrated, or otherwise compromised; in other words, those who have made mistakes. Research on

the so-called “hidden” populations can be extremely difficult, even dangerous in some cases, as these hidden populations frequently include criminals who are still criminally active. Moreover, any time that we extend beyond official documents and records, we step into a gray zone of potentially unreliable information.

Similarly, we have the disadvantage of relying almost exclusively on official records or self-report information from individuals who are not very reliable in the first place. Consequently, we frequently have access to a very limited amount of the total offense history of a particular offender, because generally only a relatively small fraction of criminal behavior is ever identified, documented, and adjudicated. Criminal justice researchers often are limited in this area because offender interviews regarding nonadjudicated criminal activity approach the “third rail” in criminal justice research. For example, criminal justice researchers must obey existing laws requiring the reporting of known or suspected child abuse. Similarly, researchers should consider the ethical issues associated with uncovering or gaining knowledge of unreported, ongoing, or planned criminal activity. Because this information can cause potential harm to the offender due to legal reporting requirements and ethical considerations, research involving the deliberate collection of unreported crime frequently is prohibited when reviewed by institutional review boards and others concerned about the rights of human research subjects. Similar to drug side effects, there are those crimes and behaviors that we know about and those that we do not. Also like drug side effects, it is generally true that the ones that we do not know about will come up and strike us eventually. In addition, security classification or other compartmentalization of data also can limit access to data or otherwise constrain the analyst.

What we are left with, then, is a *sample* of information. In other words, almost everything that we know about crime and criminals is based on a relatively small amount of information gathered from only a fraction of all criminals – generally the unsuccessful ones. Similarly, almost everything that we work with in the operational environment also is a sample, because it is exceedingly rare that we can identify every single crime in a series or every piece of evidence. In many ways, it is like working with a less than perfect puzzle. We frequently are missing pieces, and it is not unusual to encounter a few additional pieces that do not even belong and try to incorporate them. Whether this is by chance, accident, or intentional misdirection on the part of the criminal, it can significantly skew our vision of the big picture.

We can think of samples as *random* or *nonrandom* in their composition. In a random sample, individuals or information are compiled in the sample based exclusively on chance. In other words, the likelihood that a particular individual or event will be included in the sample is similar to throwing the dice. In

a nonrandom sample, some other factor plays a significant role in group composition. For example, in studies on correctional samples, even if every relevant inmate were included, it still would comprise only a sample of that particular type of criminal behavior because there would be a group of offenders still active in the community. It also would be a nonrandom sample because only those criminals who had been caught, generally the unsuccessful ones, would be included in the sample. Despite what incarcerated criminals might like to believe, it generally is not up to chance that they are in a confined setting. Frequently, it was some error on their part that allowed them to be caught and incarcerated. This can have significant implications for the analytical outcomes and generalizability of the findings.

In some cases, identification and analysis of a sample of behavior can help to illuminate a larger array of activity. For example, much of what we know about surveillance activity is based on suspicious situation reports. In many cases, however, those incidents that arouse suspicion and are reported comprise only a very small fraction of the entire pattern of surveillance activity, particularly with operators highly skilled in the tradecraft of covert surveillance. In some cases, nothing is noted until after some horrific incident, and only in retrospect are the behaviors identified and linked. Clearly, this retrospective identification, characterization, and analysis is a less than efficient way of doing business and underscores the importance of using information to determine and guide surveillance detection efforts. By characterizing and modeling suspicious behavior, common trends and patterns can be identified and used to guide future surveillance detection activities. Ultimately, this nonrandom sample of suspicious situation reports can open the door to inclusion of a greater array of behavior that more closely approximates the entire sample or population of surveillance activity.

These issues will be discussed in Chapters 5 and 14; however, it always is critical to be aware of the potential bias and shortcomings of a particular data set at every step of the analytical process to ensure that the findings and outcomes are evaluated with the appropriate level of caution and skepticism.

1.4 MODELING

Throughout the data mining and modeling process, there is a fair amount of user discretion. There are some guidelines and suggestions; however, there are very few absolutes. As with data and information, some concepts in modeling are important to understand, particularly when making choices regarding accuracy, generalizability, and the nature of acceptable errors. The analyst's domain expertise, or knowledge of crime and criminals, however, is absolutely essential to making smart choices in this process.

1.5 ERRORS

No model is perfect. In fact, any model even advertised as approaching perfection should be viewed with significant skepticism. It really is true with predictive analytics and modeling that if it looks too good to be true it probably is; there is almost certainly something very wrong with the sample, the analysis, or both. Errors can come from many areas; however, the following are a few common pitfalls.

1.5.1 Infrequent Events

When dealing with violent crime and other patterns of bad behavior like terrorism, the fact that it is a relatively infrequent event is a very good thing for almost everyone, except the analysts. The smaller the sample size, generally, the easier it is to make errors. These errors can occur for a variety of reasons, some of which will be discussed in greater detail in Chapter 5. In modeling, infrequent events can create problems, particularly when they are associated with grossly unequal sample distributions.

While analyzing robbery-related aggravated assaults, we found that very few armed robberies escalate into an aggravated assault.³ In fact, we found that less than 5% of all armed robberies escalated into an aggravated assault. Again, this is a very good thing from a public safety standpoint, although it presents a significant challenge for the development of predictive models if the analyst is not careful.

Exploring this in greater detail, it becomes apparent that a very simple model can be created that has an accuracy rate of greater than 95%. In other words, this simple model could correctly predict the escalation of an armed robbery into an aggravated assault 95% of the time. At first blush, this sounds phenomenal. With such a highly accurate model, it would seem a simple thing to proactively deploy and wipe out violent crime within a week. Examining the model further, however, we find a critical flaw: There is only one decision rule, and it is “no.” By predicting that an armed robbery will never escalate into an aggravated assault, the model would be correct 95% of the time, but it would not be very useful. What we are really looking for are some decision rules regarding robbery-related aggravated assaults that will allow us to characterize and model them. Then we can develop proactive strategies that will allow us to prevent them from occurring in the future. As this somewhat extreme example demonstrates, evaluating the efficacy and value of a model is far more than just determining its overall accuracy. It is extremely important to identify the nature of the errors and then determine which types of errors are acceptable and which are not.

Another example of rare events relates to pirate attacks, which have been associated with several high-profile incidents including the attack on the Maersk

Alabama.⁴ To put the numbers into perspective, though, at the time of this particular incident, the US 5th Fleet in Bahrain reported that there were a total of 122 raids on vessels making passage through the Gulf of Aden.⁵ Of these attacks, 42 were “successful” from the perspective of the pirates, resulting in a “success” rate for the pirates of 34%. Providing additional context, though, approximately 33,000 vessels made passage during 2008 without incident. Less than 1/2 of 1% of all the vessels were attacked, either successfully or not. Again, we could develop a model that would say that a vessel safely makes passage through the Gulf of Aden and would be correct more than 99% of the time; however, this would have no value to enhancing maritime security in the region.

One way to evaluate the specific nature of the errors is to create something called a confusion or confidence matrix. What this does is break down and depict the specific nature of the errors and their contribution to the overall accuracy of the model. Once it has been determined where the errors are occurring, and whether they impact significantly the value of the overall error rate and model, an informed decision can be made regarding acceptance of the model. Confusion matrices will be addressed in greater detail in Chapter 8, which covers training and test samples.

The confusion matrix is an important example of a good practice in analysis. It can be extremely valuable to challenge the results, push them around a bit analytically and see what happens, or look at them in a different analytical light. Again, the confusion matrix allows analysts to drill down and examine what is contributing to the overall accuracy of the model. Then they can make an informed decision about whether to accept the model or to continue working on it until the errors are distributed in a fashion that makes sense in light of the overall public safety or intelligence objective. While this process might seem somewhat obscure at this point, it underscores the importance of choosing analysts with domain expertise. Individuals that know where the data came from and what they will be used for ultimately can distinguish between those errors that are acceptable and those that are not. Someone who knows a lot about statistical analysis might be able to create extremely elegant and highly predictive models, but if the model consistently predicts that an armed robbery will never escalate into an aggravated assault because the analyst did not know that these events are relatively infrequent, there can be serious consequences. Although this might seem like an extreme example that would be perfectly obvious to almost anyone, far more subtle issues occur regularly and can have similar harmful consequences. The ultimate consequence of this issue is that the folks within the public safety community are in the best position to analyze their own data. This is not to say that it is wrong to seek outside analytical assistance, but totally deferring this responsibility, as seems to be occurring with increasing frequency, can have serious consequences due to the subtle

nature of many of these issues that permeate the analytical process. This point also highlights the importance of working with the operational personnel, the ultimate end users of most analytical products, throughout the analytical process. While they might be somewhat limited in terms of their knowledge and understanding of the particular software or algorithm, their insight and perception regarding the ultimate operational goals can significantly enhance the decision-making process when cost/benefit and error management issues need to be addressed.

Given the nature of crime and intelligence analysis, it is not unusual to encounter infrequent events and uneven distributions. Unfortunately, many default settings on data mining and statistical software automatically create decision trees or rules sets that are preprogrammed to distribute the cases evenly. This can be a huge problem when dealing with infrequent events or otherwise unequal distributions. Another way of stating this is that the program assumes that the prior probabilities or “priors” are 50:50, or some other evenly distributed ratio. Generally, there is a way to reset this, either automatically or manually. In automatic settings, the option generally is to set the predicted or expected probabilities to match the prior or observed frequencies in the sample. In this case, the software calculates the observed frequency of a particular event or occurrence in the sample data, and then uses this rate to generate a model that results in a similar predicted frequency. In some situations, however, it can be advantageous to set the priors manually. For example, when trying to manage risk or reduce the cost of a particularly serious error, it might be necessary to create a model that is either overly generous or very stringent, depending on the desired outcome and the nature of misclassification errors. Some software programs offer similar types of error management by allowing the user to specify the “cost” of particular errors in classification, in an effort to create models that maximize accuracy while ensuring an acceptable distribution of errors.

1.5.2 “Black Swans”

Some events are so unique or rare that they are referred to as “Black Swans.” “True” Black Swans cannot be predicted or anticipated; and by extension, they cannot be prevented or thwarted. More recently, though, the concept of “Anticipatory Black Swans” has been introduced. In contrast to the True Black Swans, an Anticipatory Black Swan “can be known beforehand” as compared to “what truly is a surprise.”⁶ Like data mining generally, the proposed approach to both types of Black Swans is to, “ask the right question at the right time (and be wise enough to understand the response or appreciate the signs)...[which] could lead to success in the matter of anticipating what can be anticipated, and at least understanding sooner the impact of what cannot be anticipated.”⁷

1.5.3 Identifying Appropriate Comparison Groups and Establishing the Denominator

Continuing the analysis of infrequent or rare events is the challenge associated with establishing a meaningful denominator for the calculation of rate. Being able to understand these events and identify appropriate comparison groups and context, though, requires some insight and understanding of the “nonevents.” While this might seem obvious, there is a general bias against reporting of nonevents.⁸ For example, how often do people call the local police department to report that they successfully purchased a rock of crack cocaine without incident, or that they drove through a blighted neighborhood and made it out safely? While this would be highly unlikely,⁹ they represent common examples of the challenge to the analyst regarding the identification of other individuals “exposed” to the same risk or environment, including appropriate comparison groups, and the establishment of a reasonable denominator to calculate rates. This lack of insight regarding these “nonevents” also limits our ability to effectively identify what makes the “incidents” different from these other nonevents. Comparing “successful” to “unsuccessful” attacks is important, but being able to compare these groups to other vessels that made passage through the Gulf of Aden without incident provides additional, operationally relevant insight regarding these attacks that can be used for prevention, thwarting, and response. While it is not common at this point to collect those data, analysis of significant “nonevent” data becomes increasingly viable with the increased deployment of sensors and the ability to effectively analyze geospatial and related route data.

One type of reporting metric that the analyst may encounter includes casualty data, which generally are comprised of injuries and fatalities and are reported as “Wounded in Action” (WIA) and “Killed in Action” (KIA). Unfortunately, it is not unusual for the analyst to receive reports that include lists of these data associated with a specific area of interest or operation (Table 1.1) with little additional information regarding total numbers involved or context. The *a priori* assumption in this type of analysis is that any casualty is bad,¹⁰ but a deeper understanding of the nature, circumstances, and context associated with casualties, including how many other individuals were involved, can provide the

Table 1.1 Total Number of Wounded in Action (WIA) and Killed in Action (KIA) for Four Different Groups

	WIA	KIA
Group 1 ²³	1	3
Group 2 ²⁴	467	383
Group 3 ²⁵	670,846	405,399
Group 4	4,214,200 ²⁶	5,734 ²⁷

insight necessary to develop information-based approaches to force protection going forward. The challenge when presented with these data, therefore, is to understand what they really mean.

As can be seen in [Table 1.1](#), there are marked differences in the numbers of individuals injured or killed in each of the different groups. The total number of casualties in the first group is four, while the number of injuries in Group 4 exceeds four million, and the fatalities in Group 3 exceeds 400,000. Without knowing anything else, one might assume that Group 1 represents the safest option. We need additional information, though, to make a meaningful determination regarding differences between groups 3 and 4.

[Table 1.2](#) includes the population data for the four groups, which provides additional context to the casualty numbers reported. Clarifying further, Group 1 represents Operation Red Wings, which “was the single largest loss of life for Naval Special Warfare since World War II.”¹¹ In addition to the four-man Navy SEAL team included in the table, 16 members of a Quick Reaction Force (QRF) also lost their lives trying to extract the original team. This was a tragic loss to the community that underscores the dangerous nature of Special Operations. Group 2 includes casualty statistics for the Persian Gulf War (1990–1991), and Group 3 includes statistics from World War II (1941–1946).¹² Finally, the data from Group 4 include all work-related injuries, illnesses, and fatal occupational injuries in 2005, the same year as Operation Red Wings. The denominator for this group is the entire civilian labor force, which included people working or actively looking for work.¹³ With this additional information, we realize that the actual casualty rate for Operation Red Wings was 100%; and while every injury and fatality is important, the casualty rate for all other examples is considerably lower despite higher total numbers. To be sure, actually identifying the correct denominator can represent a significant challenge to the analyst, particularly given that we frequently work in fluid environments where operational resources are flexed rapidly in response to changing operational requirements and conditions. Moreover, there frequently are many “hidden populations” and there is a significant lack of candor regarding participation in illegal activities that limit our ability to truly know the actual number exposed.

Table 1.2 Total Number of Wounded in Action (WIA) and Killed in Action (KIA). The Denominator Includes the Total Population Exposed

	WIA	KIA	Denominator
Group 1 ²⁸	1	3	4
Group 2 ²⁹	467	383	2,225,000
Group 3 ³⁰	670,846	405,399	16,112,566 ³¹
Group 4	4,214,200 ³²	5,734 ³³	149.3 million ³⁴

However, an understanding of the importance of these data is critical to good analysis and subsequent interpretation of the results.

1.5.4 Remember the Baseline

It is important to consider baseline data when analyzing and interpreting crime and intelligence information and what might skew or otherwise impact that information. Failure to consider baseline data is an error that occurs frequently, and relates back to the incorrect assumptions that samples are representative of the larger population and that variables tend to be distributed evenly. During the sniper investigation in October 2002 when 10 people were killed around the Washington, D.C. metropolitan area, one of the first assumptions made was that the suspect would be a white male because almost all serial killers are white males. When it turned out that the snipers were black, there was great surprise, particularly among the media. As one stops to consider the likely racial distribution among serial killers, it is important to note the relative distribution of race in the population of interest, in this case the United States, which is approximately 12% black according to the 2000 census data.¹⁴ Taking this information into consideration, we would not expect a 50:50 split along race lines when examining serial killers. Population statistics would indicate fewer black serial killers, if the distribution mirrored the overall population. Moreover, serial killers are relatively rare, which further confounds our calculations for reasons similar to those addressed earlier regarding small sample sizes and infrequent events. Further confounding the “conventional wisdom” regarding this subject is the highly skewed racial distribution of homicide offenders, which are 51.5% black and 46.4% white.¹⁵ When adjusted to per-capita rates, the FBI Uniform Crime Reports indicate that blacks are eight times more likely to commit homicides than whites.¹⁶ These numbers are based on cleared cases and arrests, though, which have their own unique limitations. Therefore, when viewed in light of these apparently contradictory statistics, possible reasons for the apparent bias in the initial demographic predictions of the D.C. sniper case start to make sense. Clearly, baseline information should be used to filter data and outcomes; however, this simple exercise demonstrates that even determining the appropriate baseline can be a challenge in many cases.

This example also highlights the importance of keeping an open mind. Seasoned investigators understand that establishing a mindset early in an investigation can significantly affect interpretation of subsequent leads and clues, allowing important evidence to be overlooked, such as the “white van” emphasized by the media in the sniper investigation, which artificially filtered many leads from concerned citizens and cooperating public safety agencies alike. Similarly, analysts can fall prey to these same challenges if they are not careful and consider appropriate comparative information with a clear mind that is open to alternative explanations for the data. Again, knowledge of the

potential pitfalls is almost as important as the analysis, because ignorance can have a significant impact on the analysis and interpretation of the data.

Arrest data is another area in which considering variances in population distribution can be essential to thoroughly understanding trends and patterns. When we think logically about where and when many arrests occur, particularly vice offenses, we find that officer deployment often directly affects those rates. Like the proverbial tree falling in the woods, it follows that if an officer is there to see a crime, it is more likely that an arrest will be made. This goes back to the earlier discussion regarding the crime that we know about and the crime that we do not know about. Locations associated with higher levels of crime also tend to be associated with heavier police deployment, which concomitantly increases the likelihood that an officer will either be present or nearby when a crime occurs, ultimately increasing the arrest rate in these locations. Unfortunately, the demographics represented among those arrested might be representative of the residents of that specific area but differ greatly from the locality as a whole. This can greatly skew our interpretation of the analysis and findings. What does this mean to data mining and predictive analytics? Simply, that it can be an error to use population statistics to describe, compare, or evaluate a relatively small, nonuniform sample, and vice versa. Remember the baseline, and give some thought to how it was constructed, because it might differ significantly from reality.

1.5.5 Where Did Your Data Come From?

Further segmenting the military casualty data listed earlier, out of a total number of 1929 reported service-related deaths in 2005, 646 were listed as “accident,” 739 were associated with “hostile action,” 280 were illness related, and 54 were homicides.¹⁷ Which number do we use? The total number? Should we exclude the illness-related deaths and accidents? Are “hostile action” and “homicide” similar enough to group together? Related to the importance of knowing the denominator and baseline is knowing the decision rules that were used originally to collect the data and/or establish the sample. For example, there has been considerable controversy regarding the actual number of civilian casualties during the Iraq War.¹⁸ While “death” would appear to be a relatively unambiguous metric, there are a number of different approaches to determining whether a fatality was related to the War. This includes the concept of “excess” casualties during the conflict that are attributed either directly or indirectly to the military presence and associated brutalizing influence and putative normalization of violence in an active conflict currently. Similar discussions are occurring presently regarding the conflict in Syria,¹⁹ particularly as relates to the provision of trauma services, as well as routine medical care and even immunizations to children. While this approach is not inherently right or wrong, it is always important to understand the underlying assumptions

associated with the creation of your source, particularly when interpreting the results.

1.5.6 Magnified or Obscured Effects

Uneven distributions also can create errors in the interpretation of link analysis results, which is discussed in Chapter 3. Briefly, link analysis can be a great way to show relationships between individuals, entities, events, or almost any variable that could be considered in crime and intelligence analysis. Some of the new software tools are particularly valuable, in that actual photos of individuals or elements of interest can be inserted directly into the chart, which results in visually powerful depictions of organizational charts, associations, or events. Beyond just demonstrating an association, however, link analysis frequently is employed in an effort to highlight the relative strength of relationships. For example, if Bob calls Joe 15 times, but Joe calls Paul 52 times, we might assume that the relationship between Joe and Paul is stronger than the relationship between Joe and Bob based on the relative difference in the amount of contact between and among these individuals (Figure 1.2).

These programs often allow the user to establish thresholds for link strength; however, this can provide a false sense of security. For example, in Figure 1.3, it appears that Paul has a stronger relationship with Pete, as compared to his relationship with Joe, based on the relative levels of contact. Bob, on the other hand, appears to have relatively similar relationships with both Joe and Pete, based on relatively equal levels of contact, as depicted in the link chart. Reviewing the related association matrix, however, indicates that this might not be true (Figure 1.4). The actual numbers of contacts indicates that both Paul and Bob had contact with Pete almost twice as much as they did with Joe. The relationship is skewed somewhat in the link analysis chart (Figure 1.2) because the relative levels of activity associated with Bob were much higher than those associated with Paul. As a result, the settings used in the link analysis skewed the visual representation of the relative strength of the relationships noted. For example, in this particular situation, it might be that weak links include

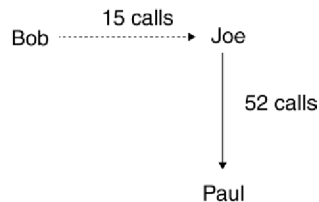
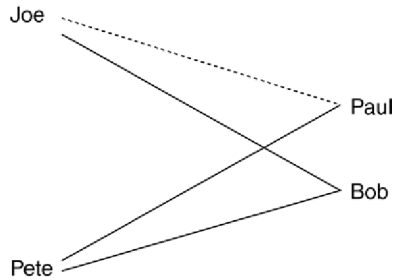


FIGURE 1.2

In addition to surfacing relationships between individuals or events, the relative strength of a relationship also can be inferred based on relative differences in the amount of contact.

**FIGURE 1.3**

Examination of this link chart suggests that Paul has a stronger relationship with Pete compared to his relationship with Joe, while Bob appears to have relatively similar relationships with both Joe and Pete, based on relatively equal levels of contact. These apparent differences in the relationships are based on differences in the strength of the association illustrated by relative differences in the lines in the link chart.

10 associations or less, while strong links require 20 associations or more. Unfortunately, unequal distributions can skew the relative importance of certain associations. In this example, both Paul and Bob had similar ratios of contact with Pete and Joe; a 2:1 relationship. However, this difference was magnified in Paul because he was associated with a lower overall frequency of contact. This allowed the difference in his contact with Pete and Joe to be revealed. On the other hand, the same relative difference in the number of contacts Pete and Joe had with Bob was obscured due to the larger number of contacts overall.

Signal-to-noise issues like this can be particularly tricky for at least two reasons. First, they can magnify differences in less-frequent events. Because it takes less to show a difference, it is relatively easy to cross the arbitrary thresholds established either by the user or preset in the software. Second, they can obscure differences in the events associated with greater frequencies. This is particularly true when simultaneously comparing relationships that are associated with

	Paul	Bob	Joe	Pete
Paul				
Bob	0			
Joe	10	35		
Pete	20	70	0	

FIGURE 1.4

This simple association matrix depicts the number of contacts between a group of individuals, and highlights the errors in the associated link chart depicted in [Figure 1.3](#).

very different levels of activity. Again, if the thresholds are not set thoughtfully with an understanding of relative frequencies, some associations can be magnified while other relationships can be obscured. There are a variety of mechanisms available to address this potential confound, including the use of percentages or ratios, which are discussed in Chapter 5; however, the key to addressing this issue generally is awareness and caution when interpreting these types of results.

1.5.7 Outliers

“Outliers,” unusual subjects or events, can skew dramatically an analysis, model, or outcome with a small sample, as is found with relatively infrequent events. For example, if we analyze a sample of three armed robbers, one of whom likes fruitcake, we might assume erroneously that a preference for fruitcake is a good indicator of criminal behavior; after all, in our current sample, 1/3 of the subjects likes fruitcake. Perhaps we further expand our sample, though, to include a total of 100 armed robbers. Again, this one subject has a preference for fruitcake, but he remains the only one. In this case, a preference for fruitcake is associated with only 1% of the sample, which is not nearly as exciting. While this is a simple example, similar errors in judgment, analysis, and interpretation of results based on small, nonrandom samples have been made throughout history, sometimes with tragic consequences. All too frequently, public safety programs and policies are based on relatively small samples with unusual characteristics.

There is a saying in medicine that there are the side effects that you know about, and those that you do not. It is the side effects that you do not know about that will get you every time. Similarly, when doing data mining and constructing models, it is absolutely imperative to remember that you are only working with a sample of the total information. Even if you believe that you have gathered the total universe of information related to a particular organization, investigation, or case, it is unlikely that you have. There is always that one little tidbit of missing information that will get you in the end. Be prepared for it. Maintaining a healthy degree of realism or skepticism regarding the information analyzed can be extremely important, particularly when new information emerges that must be integrated. So keep in mind as you deal with potentially nonrandom samples that “outliers” need to be considered seriously when analyzing these types of data.

1.6 OVERFITTING THE MODEL

Remember the caution: If it looks too good to be true, it probably is too good to be true. This can occur when creating models. One common pitfall is to keep tinkering with a model to the point that it is almost too accurate. Then

when it is tested on an independent sample, something that is critical to creating meaningful predictive models, it falls apart. While this might seem impossible, a model that has been fitted too closely to a particular sample can lose its value of representing the population. Consider repeatedly adjusting and altering a suit of clothes for a particular individual. The tailor might hem the pants, take in the waist, and let out the shoulders to ensure that it fits that particular individual perfectly. After the alterations have been completed, the suit fits its owner like a second skin. It is unlikely that this suit will fit another individual anywhere near as well as it fits its current owner, however, because it was tailored specifically for a particular individual. Even though it is still the same size, it is now very different as a result of all of the alterations.

Statistical modeling can be similar. We might start out with a sample and a relatively good predictive model. The more that we try fit the model to that specific sample, though, the more we risk creating a model that has started to conform to and accommodate the subtle idiosyncrasies and unique features of that particular sample. The model might be highly accurate with that particular sample, but it has lost its value of predicting for similar samples or representing the characteristics of the population. It has been tailored to fit perfectly one particular sample with all of its flaws, outliers, and other unique characteristics. This can be referred to as “overfitting” a model. It is not only a common but also a tempting pitfall in model construction. After all, who would not love to create THE model of crime prediction? Because this issue is so important to good model construction, it will be discussed in greater detail in Chapter 8.

1.7 GENERALIZABILITY VERSUS ACCURACY

It might seem crazy to suggest that anything but the most predictive model would be the most desirable, but sometimes this is the case. Unlike other areas in which data mining and predictive analytics are employed, many situations in law enforcement and intelligence analysis require that the models be relatively easy to interpret or actionable. For example, if a deployment model is going to have any operational value, there must be a way to interpret it and use the results to deploy personnel. We could create the most elegant model predicting crime or criminal behavior, but if nobody can understand or use it, then it has little value for deployment. For example, we might be able to create a greater degree of specificity with a deployment model based on 30-min time blocks, but it would be extremely difficult and very unpopular with the line staff to try and create a manageable deployment schedule based on 30-min blocks of time. Similarly, it would be wonderful to develop a model that makes very detailed predictions regarding crime over time of day, day of week, and relatively small geographic areas; however, the challenge of conveying that information in any sort of meaningful way would be tremendous. Therefore,

while we might compromise somewhat on accuracy or specificity by using larger units of measure, the resulting model will be much easier to understand and ultimately more actionable.

The previous example highlighted occasions where it is acceptable to compromise accuracy somewhat in an effort to develop a model that is relatively easy to understand and generalize. There are times, however, when the cost of an inaccurate model is more significant than the need to understand exactly what is happening. These situations frequently involve the potential for some harm, whether it is to a person's reputation or to life itself. For example, predictive analytics can be extremely useful in fraud detection; however, an inaccurate model that erroneously identifies someone as engaging in illegal or suspicious behavior can seriously affect someone's life. On the other hand, an inaccurate critical incident response model can cost lives and/or property, depending on the nature of the incident. Again, it is just common sense, but any time that a less-than-accurate model would compromise safety, the analyst must consider some sort of alternative. This could include the use of very accurate, although relatively difficult to interpret, models. Attesting to their complexity, these models can be referred to as "black box" or opaque models because we cannot "see" what happens inside them. As will be discussed in subsequent chapters, though, there are creative ways to deploy the results of relatively opaque algorithms in an effort to create actionable models while maintaining an acceptable level of accuracy.

Deciding between accuracy and generalizability in a model generally involves some compromise. In many ways, it often comes down to a question of public safety. Using this metric, the best solution is often easy to choose. In situations where public safety is at stake and a model needs to be interpretable to have value, accuracy might be compromised somewhat to ensure that the outcomes are actionable. In these situations, any increase in public safety that can be obtained with a model that increases predictability even slightly over what would occur by flipping a coin could save lives. Deployment decisions provide a good example for these situations. If current deployment practices are based almost exclusively on historical precedent and citizen demands for increased visibility, then any increase over chance that can be gained through the use of an information-based deployment model generally represents an improvement.

When an inaccurate model could jeopardize public safety, though, it is generally better to go without than risk making a situation worse. For example, automated motive determination algorithms require a relatively high degree of accuracy to have any value because the potential cost associated with misdirecting or derailing an investigation is significant, both in terms of personnel resources and in terms of the likelihood that the crime will go unsolved. Investigative delays or lack of progress tend to be associated with an ultimate

failure to solve the crime. Therefore, any model that will be used in time-sensitive investigations must be very accurate to minimize the likelihood of hampering an investigation. As always, domain expertise and operational input is essential to fully understanding the options and possible consequences. Without a good understanding of the end user requirements, it can be very difficult to balance the often mutually exclusive choice between accuracy and generalizability.

Analytically, the generalizability versus accuracy issue can be balanced in a couple of different ways. First, as mentioned previously, some modeling tools are inherently more transparent and easier to interpret than others. For example, link analysis and some relatively simple decision rule models can be reviewed and understood with relative ease. Conversely, other modeling tools, like neural nets, truly are opaque by nature and require skill to interpret outcomes. In many ways, this somewhat limits their utility in most public safety applications, although they can be extremely powerful. Therefore, selection of a particular modeling tool or algorithm frequently will shift the balance between a highly accurate model and one that can be interpreted with relative ease. Another option for adjusting the generalizability of a model can be in the creation of the model itself. For example, some software tools actually include expert settings that allow the user to shift this balance in favor of a more accurate or transparent model. By using these tools, the analyst can adjust the settings to achieve the best balance between accuracy and interpretability of a model for a specific need and situation.

1.8 INPUT/OUTPUT

Similarly, it is important to consider what data are available, when they are available, and what outputs have value. While this concept might seem simple, it can be extremely elusive in practice. In one of our first forays into computer modeling of violent crime, we elected to use all of the information available to us because the primary question at that point was: Is it possible to model violent crime? Therefore, all available information pertaining to the victims, suspects, scene characteristics, and injury patterns were used in the modeling process. Ultimately, the information determined to have the most value for determining whether a particular homicide was drug-related was victim and suspect substance use patterns.²⁰ In fact, evidence of recent victim drug use was extremely predictive in and of itself.

The results of this study were rewarding in that they supported the idea that expert systems could be used to model violent crime. They also increased our knowledge about the relative degree of heterogeneity among drug-involved offenders, as well as the division of labor within illegal drug markets.

Unfortunately, the findings were somewhat limited from an investigative standpoint. Generally, the motive helps determine a likely suspect; the “why” of a homicide often provides some insight into the “who” of a homicide. Although a particular model might be very accurate, requiring suspect information in a motive-determination algorithm is somewhat circular. In other words, if we knew who did it, we could just ask them; what we really want to know is why it happened so we can identify who did it. While this is somewhat simplistic, it highlights the importance of thinking about what information is likely to be available, in what form, and when and how all of this relates to the desired outcome.

In a subsequent analysis of drug-related homicides, the model was confined exclusively to information that would be available early in an investigation, primarily victim and scene characteristics.²¹ Supporting lifestyle factors in violent crime, we found that victim characteristics played a role, as did the general location. The resulting model had much more value from an investigative standpoint because it utilized information that would be readily available relatively early in the investigation. An added benefit to the model was that victim characteristics appear to interact with geography. For example, employed victims were more likely to have been killed in drug markets primarily serving users from the suburbs, while unemployed victims tended to be killed in locations associated with a greater degree of poverty and open-air drug markets. Not only was this an interesting finding, but it also had implications for proactive enforcement strategies that could be targeted specifically to each type of location (see Chapter 13 for additional discussion).

In the drug-related homicides example, the model had both investigative and prevention value. The importance of reviewing the value of a model in light of whether it results in actionable end products cannot be understated in the public safety arena. A model can be elegant and highly predictive, but if it does not predict something that operational personnel or policy makers have a need for, then it really has no value. In certain environments, knowledge for knowledge’s sake is a worthy endeavor. In the public safety community, however, there is rarely enough time to address even the most pressing issues. The amount of extra time available to pursue analytical products that have no immediate utility for the end users is limited at best. Similarly, analysts who frequently present the operational personnel or command staff with some esoteric analysis that has no actionable value will quickly jeopardize their relationship with the operational personnel. Ultimately, this will significantly limit their ability to function effectively as an analyst. On the other hand, this is not to say that everything should have an immediate operational or policy outcome. Certainly, some of my early work caused many eyes to roll. It is important, though, to always keep our eyes on the prize: increased public safety and safer neighborhoods.

Bibliography

- 1 Hsieh HF, Shannon SE. Three approaches to qualitative content analysis. *Qual Health Res* 2005; 15(9): 1277–88.
- 2 Hildebrandt W, McCue C. Unbiased analytics for the COCOMs. AHFE 2012.
- 3 McCue C, McNulty PJ. Gazing into the crystal ball: data mining and risk-based deployment. *Violent Crime Newsletter*, U.S. Department of Justice, September, 1–2; 2003.
- 4 McKnight T, Hirsh M. *Pirate Alley: commanding Task Force 151 off Somalia*. Annapolis, MD: Naval Institute Press; 2012.
- 5 Michaels J, Dilanian K, Leinwand D. Pirates another problem for Obama. *USA Today*. http://usatoday30.usatoday.com/printedition/news/20090409/1apirate09_st.art.htm; 2009 [accessed 09.04.2009].
- 6 Hunt C. Black Swans. In: Cabayan H, editor. *Anticipating rare events: Can acts of terror, use of weapons of mass destruction or other high profile acts be anticipated? A scientific perspective on problems, pitfalls and prospective solutions. Topical strategic multi-layer assessment (SMA) multi-agency/multi-disciplinary white papers in support of counter-terrorism and counter-WMD*; 2008.
- 7 Hunt C. 2008; p. 60.
- 8 Due to the bias in favor of reporting poor outcomes that may skew data, the medical community has moved to establish registries in an effort to collect data from all “exposed” individuals including good outcomes or “nonevents” in support of more accurate outcome data (Antiepileptic Drug Pregnancy Registry,). While creative exploitation of existing data resources may provide some insight regarding “nonevents” in the operational public safety and national security domain, analysts frequently must infer or speculate regarding critical variables without the benefit of true comparison data. This is particularly true in fluid or rapidly changing environments. Similar data collection efforts in the operational public safety and national security domain would be invaluable to providing greater depth of understanding and insight regarding incidents, especially as relates to key differentiators between incidents and non-events, in support of truly information-based approaches to prevention, thwarting, mitigation, and response. <http://www.aedpregnancyregistry.org/>.
- 9 I no longer say “never” after reviewing a report where the complainant advised that his drugs had been stolen.
- 10 As the examples in this section underscore, the “data” that we use in crime and intelligence analysis frequently involves some sort of loss, including the ultimate sacrifice on the part of the many young men and women who serve and protect through law enforcement and our military. A colleague reinforced this point a few years ago when we were trying to set a delivery date for an analytic report. This particular work involved the analysis of improvised explosive device (IED) incidents, which at the time were associated with tremendous loss of life and many horrific injuries. As we struggled to establish a “reasonable” date for delivery, a member of our team quietly suggested that we should try to complete the task before the next young person was killed, underscoring the poignant nature of the “data” that the crime and intelligence analyst routinely encounters.
- 11 United States Navy. Michael LT Murphy P. USN.
- 12 Leland A. Table 2; 2012.
- 13 United States Department of Labor, Bureau of Labor Statistics. *Employment status of women and men in 2005: Civilian Labor Force*. <http://www.dol.gov/wb/factsheets/Qf-ESWM05.htm>.
- 14 U.S. Census Bureau. 2000 Census. www.census.gov; 2000.
- 15 Source: FBI, Uniform Crime Reports, 1950–2000; Bureau of Justice Statistics, U.S. Department of Justice, Office of Justice Programs. Homicide rates recently declined to levels last seen in the late 1960s. www.ojp.usdoj.gov/bjs/homicide/hmrt.htm
- 16 Ibid.
- 17 Leland A. Congressional Research Service (CRS), CRS Report for Congress. *American War and Military Operations Casualties: Lists and Statistics*. Table 6. U.S. Active Duty Military

- Deaths, 1980 Through 2010, Part II, Cause of Death (as of November 2011). <http://www.hsdl.org/?view&did=727510>; 2012.
- 18 Reynolds P. Huge gaps between Iraq death estimates. BBC News. http://news.bbc.co.uk/2/hi/uk_news/6045112.stm and Tirman J. Iraq: the human cost. <http://web.mit.edu/humancostiraq/2006> [accessed 20.10.2006].
 - 19 Gladstone R. Report cites 'devastating toll' on health of Syria's children. The New York Times. http://www.nytimes.com/2014/03/10/world/middleeast/report-cites-devastating-toll-on-health-of-syrias-children.html?_r=0; 2014 [accessed 09.03.2014].
 - 20 McLaughlin CR, Daniel J, Joost TF. The relationship between substance use, drug selling and lethal violence in 25 juvenile murderers. *J Forensic Sci* 2000; 45: 349–53.
 - 21 McCue C, McNulty PJ. Guns, drugs and violence: breaking the nexus with data mining. *Law and Order* 2004; 51: 34–36.
 - 22 Hildebrandt W, McCue C 2012.
 - 23 United States Navy. Michael LT, Murphy P. USN: Operation Red Wings, Summary of Action. <http://www.navy.mil/moh/mpmurphy/soa.html>; [accessed 28.06.2005].
 - 24 Leland A. Congressional Research Service (CRS), CRS Report for Congress. American War and Military Operations Casualties: Lists and Statistics. Table 2. Principal Wars in Which the United States Participated: U.S. Military Personnel Serving and Casualties (1775–1991). <http://www.hsdl.org/?view&did=727510>; 2012.
 - 25 Leland A. Table 2; 2012.
 - 26 United States Department of Labor, Bureau of Labor Statistics. Workplace Injuries and Illnesses in 2005. Table 2: Numbers of nonfatal occupational injuries and illnesses by selected industries and case types. <http://www.bls.gov/iif/oshwc/osh/os/osnr0025.pdf>; 2005.
 - 27 United States Department of Labor, Bureau of Labor Statistics. Table A-1. Fatal occupational injuries by industry and event or exposure, All United States. <http://www.bls.gov/iif/oshwc/foi/cftb0205.pdf>; 2005.
 - 28 United States Navy. Michael LT, Murphy P. USN: Operation Red Wings, Summary of Action. <http://www.navy.mil/moh/mpmurphy/soa.html>; [accessed 28.06.2005].
 - 29 Leland A. Table 2; 2012.
 - 30 Leland A. Table 2; 2012.
 - 31 Leland A. Table 2; 2012.
 - 32 United States Department of Labor, Bureau of Labor Statistics. Workplace Injuries and Illnesses in 2005. Table 2: Numbers of nonfatal occupational injuries and illnesses by selected industries and case types. <http://www.bls.gov/iif/oshwc/osh/os/osnr0025.pdf>; 2005.
 - 33 United States Department of Labor, Bureau of Labor Statistics. Table A-1. Fatal occupational injuries by industry and event or exposure, All United States. <http://www.bls.gov/iif/oshwc/foi/cftb0205.pdf>; 2005.
 - 34 United States Department of Labor, Bureau of Labor Statistics. Employment status of women and men in 2005: Civilian Labor Force. <http://www.dol.gov/wb/factsheets/Qf-ESWM05.htm>.

Domain Expertise

“Data + Context = Insight”

Christian Gheorghe

I do not need a degree in mechanical engineering to drive a car. I do need some training in its operation, as well as knowledge of the rules of the road. This is similar in many ways to data mining. The software has progressed to the point where it is no longer necessary to be a statistician or artificial intelligence (AI) engineer. There is some training required to understand how to use the software, however, and some additional knowledge regarding the data mining “rules of the road.” This will help the user avoid some of the common analytical pitfalls covered in other chapters of this book. The most important knowledge for successful data mining is domain expertise. It has been my experience that it is relatively easy to teach crime and intelligence analysts, even those with no formal statistical training, how to use data mining software. The converse is not true. I have found it extremely challenging to teach statisticians and other analytical folks about crime and criminals and what has value to police operations. Almost all of this comes back to domain expertise. When people know crime and criminals, the questions come easily. When they do not, the questions and answers frequently are misguided and reveal errors in logic that seriously compromise the value of the output.

2.1 DOMAIN EXPERTISE

One of the critical prerequisites for data mining is something called “domain expertise.” Generally defined, domain expertise implies knowledge and understanding of the essential aspects of a specific field of inquiry. In other words, you need to know your stuff. This is absolutely essential in data mining because so much of the discovery and evaluation process is guided by an intuitive knowledge of what has value, both in terms of input and output, as well as of what makes sense. With a poor understanding of where the information came from and what the results will be used for, the analytical products are unlikely

to have much, if any, value. Briefly stated, domain expertise is used to evaluate the inputs, guide the process, and evaluate the end products within the context of value and validity.

Operational personnel think quickly and make rapid decisions because they have to. They also possess extreme confidence in their abilities and knowledge – again, because they have to. To behave any other way would make them inherently unsafe in their profession. If they stopped to ponder all of the possible alternative hypotheses and outcomes like analysts would, they would not last long on the street. They would either be killed by the bad guys or lose the support of their own troops after waiting too long to make a decision.

In most situations, operational personnel know more than anyone else about crime, criminals, crime trends, and patterns, what is “normal,” and what is cause for concern. Given this definition, operational personnel should be natural data miners. Unfortunately, one area where operational personnel seem to lack confidence is in the area of data and analysis. Many have an aversion to statistics and seem to be somewhat intimidated by the whole process. This is really unfortunate because most of them have excellent analytical skills. In many ways, a good investigator is an excellent analyst and a natural data miner. In fact, investigative training and process resembles case-based reasoning in many ways. Investigators typically “understand new [cases] in terms of past ones” that they have investigated.¹

For example, who better understands the limitations of crime and intelligence data than the people responsible for collecting it? Who knows better what the analytical products will be used for and what they should look like? Similarly, who better to distinguish between suspicious data and data that are both valid and reliable? The answer to all three questions is operational personnel. Our sworn partners in the good fight are perfectly suited in many ways to do our jobs as analysts, or at least to partner more closely with us in the analytical process.

2.2 DOMAIN EXPERTISE FOR ANALYSTS

Crime and intelligence analysts who spend all of their time in front of a computer can become so separated from the data and end users that they have little value to the organization. Getting out into the field whenever possible serves at least four separate and important functions. First, fieldwork helps analysts understand the data and where it comes from. This work helps them enhance or, in some cases, begin to acquire their domain expertise. It is very difficult to analyze crime or intelligence data without some understanding of the larger context. Again, it is important to know your subjects/suspects. Certainly, there are situations where it would be dangerous or inappropriate for an analyst to

tag along, but periodic ride-alongs, regular attendance at roll call or command staff meetings, and frequent interaction with the organization's operational personnel provide invaluable education regarding local trends and patterns, as well as insight into historical information and institutional memory. Some of the most teachable moments I have experienced were standing over the victim at a crime scene at 2:00 A.M. or sitting in the back of a sweltering surveillance vehicle in July.

Fieldwork also can be particularly useful in identifying limitations to reliability and validity in the data. Similarly, it is very helpful to understand the operational limitations placed on data collection. In many situations, the operators are the individuals responsible for collecting the data and information. Whether incident reports, surveillance information, informant interviews, or forensic evidence, the data collection task almost always resides with the operational personnel. Unfortunately, this frequently creates a tension between collection of complete, accurate, and reliable information and getting the job done on the street. As nice as it would be to have each and every offense report completed accurately with detailed narrative summaries, good behavioral descriptors, and neat penmanship, this is unlikely in most situations. Given current staffing shortages and workload issues, many sworn personnel are so busy responding to calls and other pressing issues that they end up completing many of their reports during their meal break or at the end of their tour. It would be impossible to completely understand the unique challenges that confront overworked operational personnel; however, until analysts "walk the walk," the gulf between them and their sworn counterparts will be huge.

Second, by getting out in the field, analysts get a better understanding of what the operational personnel need. For example, the last thing that most sworn folks need or want is more paperwork. Filling out a pile of lengthy field interview reports, particularly if they are cumbersome, duplicative of other reports, and unnecessarily detailed, really pales in importance when faced with multiple pending calls. By getting out into the field, analysts might be able to identify opportunities to streamline reporting and otherwise become part of the solution. This benefits everyone, including the analysts, who are more likely to receive help, guidance, and valuable input from their colleagues in the field when their relationships are enhanced in this manner.

Third, analysts can better target their analytical products by working more closely with the operational units that they support. Getting out from behind the computer increases the give and take. Some of the best research and analysis that I have had the pleasure to be involved with have come from informal conversations with folks who were on the job and said, "Have you ever thought about looking into this?" There is no disgrace in going directly to the end users and working with them to create an analytical product that will meet their

needs. It certainly saves time and effort when compared to the all-too-common approach of successive approximations.

Finally, fieldwork helps build the relationship between analysts and operational personnel. There is nothing like standing outside at 2:00 A.M. in freezing rain to create camaraderie and bonding. Fieldwork helps analysts understand the unique responsibilities, limitations, and time constraints that the operational personnel face in the line of duty. It also sends a strong positive message to the folks working in the field, who generally receive little praise for doing a difficult job under often miserable circumstances.

2.3 THE INTEGRATED MODEL

Clearly, most operators are not about to give up their lives of excitement and adventure to devote the remainder of their professional careers to analyzing data and information, but there are several avenues for collaboration and compromise.

First, viewing the analysis of crime and intelligence data as a partnership offers the unique opportunity to achieve the best of all worlds. As indicated in [Figure 2.1](#), in many agencies data and information arrive at the desk of the analyst, who reviews and analyzes the information and then prepares some sort of analytical product, which is sent up through the command staff and/or out into the field.

A revised model that integrates analysis and operations into a seamless, self-perpetuating cycle is outlined in [Figure 2.2](#). By working together, the information comes to the analyst within an operational context. The analyst has some indication regarding where the information came from, its reliability, and its validity, as well as what type of analytical product would be most desirable. The information is then processed and analyzed in a much more meaningful way than if the analyst had been working in an informational void. Similarly, the output, rather than representing some arcane statistical analysis or simple crime count, has operational value that can be appreciated and employed directly by the operators. Certainly, there are situations when it is not possible to share everything with the analyst; however, these situations can be mitigated



FIGURE 2.1

In the traditional model, analysts prepare reports and other analytical output with little input or feedback from the operational end users.

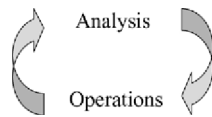


FIGURE 2.2

By establishing information as a fluid interface between operational and analytical domains, it is analyzed within an operational process. Sworn personnel are able to guide the process, including the nature, structure, and format of analytical output, which increases the likelihood that analytical products will be actionable. In the meantime, the analyst frequently receives better information from the field, while gaining better insight regarding data reliability and validity.

somewhat with even minimal interaction and guidance on the part of the operational personnel or other end users.

The potential value of operator–analyst interaction and collaboration has received increased attention from a variety of national security organizations. In their influential paper entitled, “Fixing Intel,”² General Mike Flynn and his colleagues make the point that, “[a]nalysts’ Cold War habit of sitting back and waiting for information to fall into their laps does not work in today’s warfare and must end,”³ and outline a paradigm shift in the operator–analyst model that includes the innovative use of forward deployed, distributed analysts, and creative solutions to knowledge gathering, management, analysis, and dissemination. This model enables forward-deployed analysts the ability to meet with the primary, grassroots collectors in the field, while also providing field-level gathering of information, collation, integration, analysis, and dissemination. The ability to leverage a fluid, rapidly adaptable, end-user-focused model has been associated with clear intelligence successes, underscoring the importance of domain expertise and the value of the embedded analyst model. Similarly, the Office of the Director of National Intelligence (ODNI) developed programs like the Rapid Analytic Support and Expeditionary Response (RASER) as part of their Analytic Transformation, with the goal being to enable analysts to “know their customers and bring them into the process,” in order to “get the right analysis to the right people at the right time, in a form they can use.”⁴

Further underscoring the benefits of an integrated model, key attributes of successful crime and intelligence analysis units as outlined in *Fixing Intel*⁵ include a proactive analyst cadre. Analysts “advertise” their collection and analysis capabilities, and work with the end users to ensure that they know what can be provided while concomitantly seeking to identify additional services that would be of value. Moreover, analysts go out into the field to support the operational end users. This action in particular benefits not only the end users, but the analysts who return more informed regarding the data and operational requirements, as well.

Data mining and predictive analytics, therefore, offer a unique opportunity for analytical and operational personnel to work together in new and exciting ways. By exploiting the intuitive nature of this analytical process, these two groups, with their complementary domain expertise, can more fully utilize existing information resources while creating and guiding novel approaches to enhancing public safety. Although this certainly represents a paradigm shift in how these two relatively diverse professional domains currently function, data mining and predictive analytics afford a unique opportunity to achieve analytical critical mass, taking crime and intelligence analysis into the future.

Bibliography

- 1 Casey E. Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and Internet crime investigation. Knowledge Solutions. www.corpus-delicti.com/case_based.html; 2002.
- 2 Flynn MT, Pottinger M, Batchelor PD. Fixing intel: a blueprint for making intelligence relevant in Afghanistan. January 5, 2010; 2009.
- 3 Flynn MT, Pottinger M, Batchelor PD. Fixing intel: a blueprint for making intelligence relevant in Afghanistan. January 5, 2010; 2009.
- 4 Office of the Director of National Intelligence. Analytic transformation: unleashing the potential of a community of analysts. <http://www.hsdl.org/?view&did=29867>; 2008. p. 2.
- 5 Flynn MT, Pottinger M, Batchelor PD. Fixing intel: a blueprint for making intelligence relevant in Afghanistan. January 5, 2010; 2009.

Data Mining and Predictive Analytics

“Originally, data mining was a statistician’s term for overusing data to draw invalid inferences.”

J. Peterson

To continue the quote, “Our definition will be...‘The extraction of implicit, previously unknown, and potentially useful information from data.’”¹ Revealing its origins and widespread use in business, data mining and predictive analytics go by many names, including knowledge discovery, sense making, and more recently, data science.² Data mining is “the discovery of new and unexpected patterns from large data sets that can be used to solve problems.”³ In other words, data mining involves the systematic analysis of large data sets using automated methods. By probing data in this manner, it is possible to prove or disprove existing hypotheses or ideas regarding data or information, while discovering new or previously unknown information. In particular, unique or valuable relationships between and within the data can be identified and used proactively to categorize or anticipate additional data or subsequent event. Through the use of exploratory graphics in combination with advanced statistics, machine learning tools, and artificial intelligence, critical “nuggets” of information can be mined from large repositories of data.

WHAT CAN WE LEARN FROM WAL-MART AND AMAZON ABOUT FIGHTING CRIME?⁴

I firmly believe that the answers to at least some of our public safety and national security challenges can be found in the e-commerce and marketing community and have been very open about that fact that very little of my own work in this space is particularly innovative or unique. Rather, I have identified and adopted successful approaches used in other professional domains to public safety and national security analysis. With that in mind, Los Angeles Police Department Chief Charlie Beck and I posed the question, what can we learn from Wal-Mart and Amazon about fighting crime in a recession?⁵ While the title was intended to be provocative, there were several very serious points embedded in the paper regarding the value predictive analytics developed

for use in the commercial sector in the development of information-based tactics, strategy, and policy decisions in the operational public safety and security environment.

One of the frequently used examples to illustrate the “confirmation” and “discovery” aspects of predictive analytics is the Wal-Mart “emergency response plan.”⁶ Described as an “analytic competitor,”⁷ Wal-Mart as an organization has effectively leveraged large amounts of historical point-of-sale data in an effort to anticipate and effectively respond to their customers. In one particular example, the analysts at Wal-Mart noted unique patterns of purchasing behavior in advance of a major weather event. Specifically, sales of bottled water, duct tape, and Pop-Tarts increased in the period of time immediately preceding a storm. While the bottled water and duct tape are obvious choices, and included in recommendations prepared by the government,⁸ the increased sales of Pop-Tarts is surprising. Therefore, in this particular example, the bottled water and duct tape represent “confirmation,” and the increased sales of Pop-Tarts would be “discovery” of new and ideally actionable relationships. It is important to note that while we can consider the possible motivations behind this behavior (e.g., Pop-Tarts are easy to store and prepare, do not require refrigeration, can be eaten directly out of the box), understanding of the underlying reason for this is not necessary for it to be actionable. Knowing that consumers will be purchasing Pop-Tarts in anticipation of a major weather event is sufficient for Wal-Mart to adjust their supply chain and meet the need. While this particular example might appear to be trivial or light hearted, Wal-Mart’s ability to quickly anticipate need and stage essential resources in support of a timely post-storm response surpassed the local, state, and federal agencies during Hurricane Katrina.⁹ Similarly, the ability to identify, characterize, and anticipate trends, patterns, associations, and sequences can enable information-based approaches to prevention, thwarting, mitigation, and response.



FIGURE 3.1

Data mining and predictive analytics developed for use in the commercial sector can be adopted and used by public safety and security decision makers to support the development of information-based tactics, strategy, and policy decisions in the operational environment. The perceptive analyst frequently can identify novel approaches and innovative new techniques by looking beyond their professional domain to other settings, particularly those occupied by “analytic competitors.”

Other lessons learned from the commercial sector include the concept of “analytic maturity”¹⁰ or organizational readiness, which is necessary to move agencies from counting and reporting to being able to effectively anticipate events and influence outcomes. Moreover, acquiring analytic capabilities and developing capacity requires more than data and technology; attracting and retaining talent has emerged as a significant constraint for organizations seeking to effectively implement a program of predictive analytics.¹¹ Tom Davenport, a thought leader in the effective use of analytics, has identified key attributes of “analytic competitors” and the associated developmental pathways for early adoption and improvement.¹² Again, reporting, collecting, and compiling data are necessary but not sufficient. Ultimately, being able to better understand and influence behavior enable us to get in front of incidents and change outcomes. By identifying and observing the “analytic competitors” in the commercial sector, the public safety and security community can create the requisite organizational readiness and associated “culture of analytics” that can be used to realize the promise of predictive analytics in support of information-based approaches to prevention, thwarting, mitigation, and response (Figure 3.1).

3.1 DISCOVERY AND PREDICTION

When examining drug-related homicide data several years ago, we decided to experiment with different approaches to the analysis and depiction of the information. By drilling down into the data and deploying the information in a mapping environment, we found that the victims of drug-related homicides generally did not cross town to get killed. While it makes sense in retrospect, this was a very surprising finding at the time. This type of analysis of homicide data had not been considered previously, although after it had been completed it seemed like a logical way to view the information.

After further analysis of the data, we were able to generate a prediction regarding the likely location and victim characteristics of one of the next incidents. Within the next 12 h, a murder was committed with characteristics that were strikingly similar to those included in the prediction, even down to the fact that the victim had not crossed town to get killed.

This embodies the use of data mining and predictive analytics in law enforcement and intelligence analysis. First, the behavior was characterized and, through this process, new information was “discovered.” The idea of looking at the information in this fashion to determine the relationship between the victim’s residence and subsequent murder location made sense, but had not been done before. Adding value to crime information in this manner deviates significantly from the traditional emphasis on counting crime and creating summary reports. By looking at the data in a different way, we were able to discover new facets of information that had significant operational value.

Second, by characterizing the behavior, it could be modeled and used to anticipate or predict the nature of future events. The ability to anticipate or predict events brings a whole new range of operational opportunities to law

enforcement personnel. Much as in the movie *Minority Report*, once we can anticipate or predict crime, we will have the ability to prevent it. Unlike the movie, however, crime prevention can be effected through the use of proactive deployment strategies or other operational initiatives, rather than proactive incarceration of potential offenders. On the other hand, the ability to characterize risk in potential victims provides an opportunity for targeted, risk-based interventions that ultimately can save lives and provide safer neighborhoods for all, a topic that will be covered in Chapter 11.

This example, although a somewhat odd and inelegant use of “brute force” analytics, embodies the essence of data mining and predictive analytics within the public safety arena. Through the use of these powerful tools, we can understand crime and criminal behavior in a way that facilitates the generation of actionable models that can be deployed directly into the operational environment.

3.2 CONFIRMATION AND DISCOVERY

At a very simple level, data mining can be divided into confirmation and discovery. Criminal investigation training is similar to case-based reasoning.¹³ In case-based reasoning, each new case or incident is compared to previous knowledge in an effort to increase understanding or add informational value to the new incident. In addition, each new incident is added to this internal knowledge base. Before long, an investigator has developed an internal set of rules and norms based on accumulated experience. These rules and norms are then used, modified, and refined during the investigation of subsequent cases. Analysts and investigators will look for similarities and known patterns to identify possible motives and likely suspect characteristics when confronted with a new case. This information is then used to understand the new case and investigate it.

These internal rule sets also allow an investigator to select suspects, guide interviews and interrogations, and ultimately solve a case. These existing rule sets can be evaluated, quantified, or “confirmed” using data mining. In addition, internal rule sets can be modified and enhanced as additional information is added and integrated into the models. Finally, as predictive algorithms are developed, we can extend beyond the use of data mining for simple characterization of crime and begin to anticipate, predict, and even prevent crime in some cases.

Many seasoned homicide investigators can identify a motive as the call comes in, based on the nature of the call, geographic and social characteristics of the incident location, and preliminary information pertaining to the victim and injury patterns. For example, a young male killed in a drive-by shooting in an area known for open-air drug markets is probably the victim of a drug-related homicide. Additional information indicating that the victim was known to be involved in drug selling will further define the motive and suggest that likely

suspects will include others involved in illegal drug markets. Postmortem information indicating that the victim had used drugs recently before his death will add additional value to our understanding of the incident.

Law enforcement personnel, particularly those who have acquired both experience and success working on the streets, have internal “rule sets” regarding crime and criminal behavior that can be invaluable to the data mining process. In some initial research on juveniles involved in illegal drug markets, we found results that differed significantly from the prevailing opinions in the literature, which indicated that most drug sellers are involved in illegal drug markets in order to support their personal use.¹⁴ The common scenario involved a poor individual who experimented with drugs, rapidly became hooked, escalated to “hard” drugs, and then needed to rely on drug sales to support a rapidly growing, expensive habit. Our results indicated a very different scenario. The data that we reviewed indicated that drug sellers actually functioned very well, tended to have excellent social skills, and rarely used illegal drugs beyond some recreational use of marijuana. It was only when we looked at the relatively small group of drug traffickers who had been shot previously that we found relatively high levels of substance use and generally poor functioning. Our findings were somewhat confusing until we had the opportunity to discuss them with law enforcement professionals who were still actively working illegal narcotics. These individuals were not surprised by our findings. They pointed out to us that most successful drug dealers do not use what they sell because it cuts into their profits and, perhaps more importantly, impairs their ability to function in an extremely predatory criminal environment. Moreover, those drug dealers that do not function well generally do not live very long. From this point on, we made a point of using this type of reality testing not only to evaluate or confirm our findings but also to guide our research in this area. In many ways, this approach embodies data mining as a confirmation tool. By learning more about the internal rule sets that detectives used to investigate cases, we were able to structure and guide our data mining. In most cases we were able to confirm their instincts; however, in other cases the results were truly surprising.

3.3 SURPRISE

“The most interesting phrase in Science – that often heralds new discoveries – is not ‘Eureka!’ but ‘That’s Funny...’

Asimov

By using automated search and characterization techniques, it also is possible to discover new or surprising relationships within data. The ability to characterize large databases far exceeds the capacity of a single analyst or even a team of analysts.

The Commonwealth of Virginia has been a pioneer in the use of DNA databases to identify suspects and link cases based on the use of DNA evidence. Having achieved considerable success in this area, the Commonwealth boasts a record of approximately one “cold hit” per day.¹⁵ One noteworthy feature of the Virginia database is that it includes DNA from all convicted felons, as opposed to only those known to be violent or sexually violent. An informal conversation with the director of forensic sciences revealed that a large number of their DNA cold hits had come from offenders with no prior history of either violent or sex-related offenses. Many of these offenders had been incarcerated previously for property crimes, particularly burglary. This was a particularly surprising finding because it had been assumed that most of the cold hits would come from offenders previously convicted of sexual or violent crimes. In fact, some states had restricted their inclusion criteria to only those felons convicted of violent or sexually related crimes. The assumption was that these would be the only individuals of interest because they would be the most likely to recidivate in a violent manner.

Having spent considerable time reviewing the case materials associated with murderers, we could recall anecdotally several cases where a specific offender escalated from nonviolent to violent offending. Perhaps most noteworthy was the Southside Strangler case in Virginia, which subsequently became the first case to use DNA evidence to convict a suspect in court. Prior to committing several horrific murders in Richmond, Virginia, Timothy Spencer was known to have committed burglaries in northern Virginia.

Challenged by this seemingly spurious finding, we embarked on an analysis of several large correctional databases to determine whether there was something unusual about the sample of DNA cold hits that could explain this apparent anomaly, or whether it was real. Using discriminant analysis, a classification technique, it was determined that a prior burglary was a better predictor than a prior sex offense of a subsequent stranger rape, a very surprising finding. Subsequent review of the sex offender literature confirmed our findings.

It is important to note, however, that in many cases the type of nonviolent offending was different than crimes perpetrated by offenders who did not escalate. The use of data mining to identify and characterize “normal” criminal behavior has turned out to be an extremely valuable concept and is discussed in detail in Chapter 10.

3.4 CHARACTERIZATION

Using data mining, we can begin to further characterize crime trends and patterns, which can be essential in the development of specific, targeted

approaches to crime reduction. For example, we know that violence can take many forms, which are addressed through different approaches. This is the first step in the modeling process. A program to address domestic violence might employ social service workers as second responders to incidents of domestic violence. Victim education, offender counseling, and protective orders also might be implemented. Drug-related violence, on the other hand, requires a different approach. In fact, different types of drug-related violence will require different solutions, depending on their specific nature. By delving into the data and identifying associated clusters or groups of crime, we can gain additional insight into the likely causes. Ultimately, this facilitates the identification and development of meaningful, targeted intervention strategies.

Analysis of the data in this manner does not involve the use of a crystal ball. Rather, it requires an understanding of the data and the domain expertise necessary to know when, where, and how to dig into the data, what data to use, and what questions to ask about it. The importance of solid domain expertise cannot be overstated. Without knowing what has value and meaning to an understanding of the data within the context of crime and intelligence analysis, processing the information, and investigating, it will add little meaning and might result in bogus findings.

3.5 “VOLUME CHALLENGE”¹⁶

Although this phrase first emerged during the period immediately after the events of September 11, the law enforcement and intelligence community have been trying to address staggering increases in data and information for many years. The number of tips, reports, complaints, and other public-safety-related information confronting law enforcement and intelligence professionals on a daily basis is phenomenal. This particular information challenge can be illustrated well by following major case investigations that have been in the news.

On January 9, 2003, KXTV reported that investigators had received more than 2600 tips in response to the Laci Peterson disappearance.¹⁷ Considering that Ms. Peterson was reported missing on Christmas Eve, the local authorities received these 2600 tips in less than 17 days, or approximately 162 tips per day, assuming that the rate of tips was distributed uniformly, which is unlikely. Similarly, during the D.C. sniper investigations, tips were being received at rates as high as 1000 per hour.¹⁸ Given the nature of these incidents and the volumes of associated information, perhaps the most important task associated with crime tips is logging them into a database in some sort of systematic fashion. This challenge is followed closely by the need to analyze or make some sense out of the information, identifying and clustering those that are similar, and at the same time highlighting any patterns and trends in the information.

How do we even begin to analyze this volume of information, though? In many cases, the tips are initially logged and then shared as leads with investigators. In some cases, tip information might be maintained in electronic databases but, even under the best of circumstances, automated search and analysis is limited by available analytical tools and capacity. Until recently, it was almost impossible for a single analyst or even an analytical task force to thoroughly review and assimilate this amount of information in any sort of meaningful or systematic fashion. Unfortunately, this approach significantly limits the value of tips, which ultimately can compromise public safety and cost lives.

For example, subsequent review of the D.C. sniper investigation revealed that the actual vehicle used by the snipers had been seen and reported. In other words, many of the answers to solving the case resided in the tip databases, but the volume of information precluded their detection. The key nuggets of information essential to identifying a suspect or impending event often are identified in retrospect, which frequently is too late. As the review of high-profile cases often reveals, the information necessary to closing a case or preventing a tragedy might be hidden in plain sight within the large, unmined tip databases residing in law enforcement and intelligence organizations throughout this country and throughout the world. Unfortunately, as time passes and the number of tips continues to grow, the ability to efficiently and effectively review and analyze the information using traditional methodologies decreases concomitantly. That is why it is so important that public safety agencies adopt and employ the automated search strategies and data mining techniques that are now available.

Clearly, no case will be decided exclusively based on the use of computer programs and analytics, but these tools can be brutally objective, beyond the most seasoned detective. Data mining can also transcend the media reports, focus, hype, information overload, and even the brutality and violence associated with the crime scene, focusing exclusively on the compiled information and facts. As such, it offers a tremendous advantage to the public safety community over traditional methodologies.

3.6 EXPLORATORY GRAPHICS AND DATA EXPLORATION

AVAILABLE SOFTWARE

As has been said more than a few times throughout this text, the need to increase the analytical capacity in the crime and intelligence community within the United States, coupled with increasing interest in the area of data mining, has supported a flurry of new products and even some renamed old products. Therefore, one goal of this text is to create an informed consumer, for two reasons. First, and perhaps most obvious, is that data mining software can be very expensive. It is important, therefore, to ensure that you are getting what you have paid for.

Second, and perhaps more important, is the fact that most readers of this text will be considering a purchase in support of some sort of public safety application. Whether for crime or intelligence analysis, it is extremely important to ensure that the outcomes are reliable and valid. An inferior product or one that does not have the analytical muscle to back its advertisements represents a failure not only in purchasing and budgetary decisions but also in the support of public safety. In other words, an error in this realm can cost not only very scarce dollars, it can also cost lives.

It is unlikely that every agency will need to purchase the most expensive and high-powered tools available. Rather, some consideration should be given to the nature of the need within the organization and the best analytical approach and associated tools for the job. One good place to start might be the development of information-based deployment strategies because, if used appropriately, these tools can pay for themselves relatively quickly in personnel savings alone. Another area to consider is investing some time and effort into information management, which can facilitate the full exploitation of predictive analytics. Ideally, systems designed specifically to deploy this information visually or through mapping, or directly into the analytical environment will continue to be developed and enhanced. It is not necessary to create an analytical unit that is outfitted to look like NASA mission control. Identifying some initial, manageable areas for improvement that can be enhanced and expanded over time represents best practice in the acquisition of new technologies.

As we consider exploratory graphics and data visualization, [Figure 3.2](#) illustrates conceptually why this is an important first step and how it is possible to narrow the focus on the data in an effort to identify relatively homogenous subsets of information. For example, the category of “crime” is large and relatively heterogeneous. Included within this category is everything from misdemeanor theft to murder. Trying to do anything with such a diverse array of behavior is almost certain to fail; people often realize this when they try to evaluate a “crime” prevention strategy and find out later that the problem is too large for any single program to make a meaningful impact.

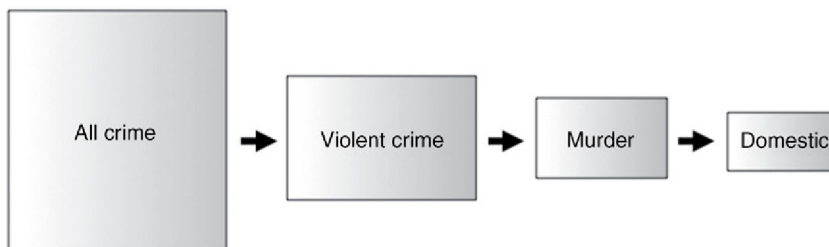


FIGURE 3.2

Narrowing the focus on the data frequently can reveal smaller groups that are relatively similar in their attributes, which facilitates subsequent analysis.

If we divide the data somewhat, the “violent crime” data can be selected. This is still a relatively large, heterogeneous category that is likely to include aggravated and sexual assaults as well as robberies and murders. It would be difficult to generate many useful models or predictions on such a wide range of information. A still more detailed focus on the data provides an opportunity to add further value to a thoughtful characterization and analysis of the data.

Through further investigation of the violent crime data, another subcategory of “murder” can be identified. Again, however, this is still relatively generic. For example, a robbery-related homicide is likely to differ in many significant ways from a domestic homicide. A similar discontinuity could be revealed when domestic homicides are compared directly to drug-related homicides. Dividing “murder” based on motive or victim–perpetrator relationship will further increase the relative homogeneity of the grouping.

One particularly prescient market forecast indicated that an area of growth in the data mining field would include specialty niche markets.¹⁹ Products developed for these niche markets would be tailored toward domain experts. These analytical tools would require less technical expertise and training, relying instead on the end user’s knowledge of their field and the use of innovative graphical interfaces and other visualization techniques. The availability of tools developed specifically for law enforcement, security, and intelligence analysts has increased markedly and continues to advance the accessibility and concomitant use of data mining and predictive analytics in the applied setting.

With this in mind, the first question might be: Do I need power tools for this? The answer is a most definite “maybe.” Exploring the data to identify actionable patterns and trends almost certainly requires the use of computerized approaches. The uncertainty generally involves the specific nature of the tools required. Because this is an important consideration that can impact your success with these tools, it has been addressed throughout this section in an effort to support the “informed consumer” in the acquisition process.

By way of example, [Figures 3.3 and 3.4](#) illustrate the value that exploratory graphics and visualization can bring to the analysis of crime data. [Figure 3.3](#), visualizes crime frequency and clearance metrics for a number of different patterns of offending, which conveys information regarding relative differences in crime frequency and investigative performance in an intuitive and easy to understand interface. On the left panel of the figure, we can see the number of incidents as compared to those that are “resolved” or cleared, which provides a great visual tool to quickly understand crime trends and patterns, and associated investigative performance. In this particular example, the crime categories of vehicle theft, robbery, and burglary are associated with a much lower clearance rate, as compared to other patterns of time. On the right panel, crime incidents and associated clearance data are illustrated by day of week and hour of the

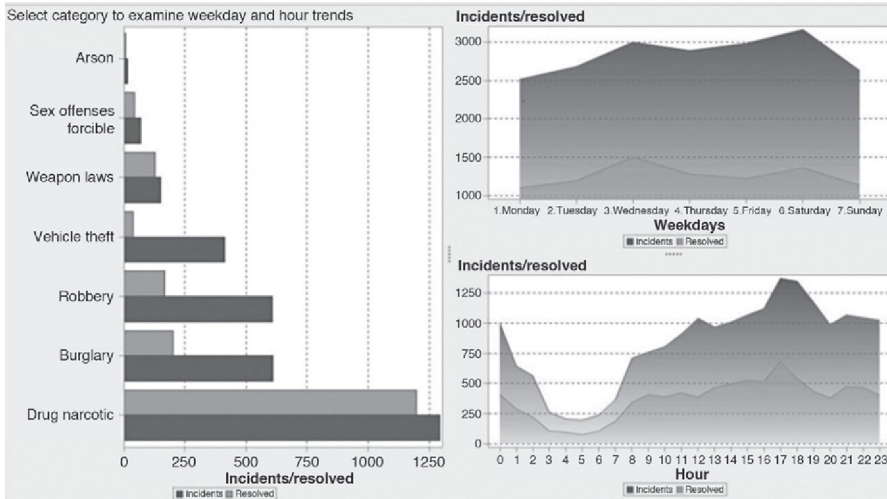


FIGURE 3.3

The use of effective visualization tools to quickly convey information regarding crime trends and patterns in an easy to use, intuitive format. The left panel illustrates different patterns of major crimes, which are further segmented as “resolved” or cleared (darker bars). The figures in the right panel illustrate differences in clearance rates as compared to incidents by day of week (upper panel), and time of day (lower panel). Copyright © 2014 SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.

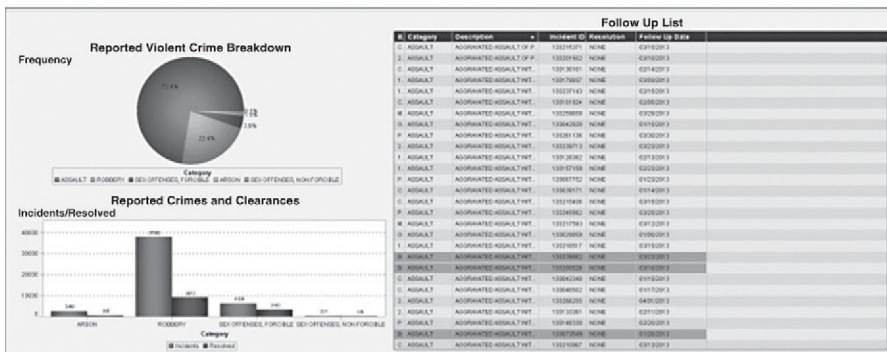


FIGURE 3.4

A different model of visualization to focus on violent crime trends and patterns. The left upper panel illustrates reported violent crime frequencies, broken out into specific patterns of violent crime, which visually surfaces aggravated assaults as accounting for the majority of reported violent crimes. The lower panel segments violent crime types by clearance; confirming robberies as a pattern of crimes with a low resolution rate as compared to other violent crimes. Specific incident information is provided in a listing report in the right panel to highlight specific incidents for follow up. Copyright © 2014 SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.

day. Figure 3.4 provides an operational example of the violent crime segmentation discussed earlier. Figure 3.4 also illustrates information similar to that presented in Figure 3.3, albeit in a slightly different way, which underscores the flexibility associated with different approaches to visualization. In the left upper panel, violent crimes are segmented by type, quickly surfacing differences in the frequency of these different patterns, and underscoring the fact that assaults represent the majority of reported violent crimes (72.4% of all violent crimes). The left lower panel confirms that the clearance rate for robberies is lower than that for other violent crimes. Finally, the information depicted in the table in the right panel includes specific incidents identified for follow up.

Why are exploratory graphics and visualization important? Basic knowledge regarding crime trends and patterns represents an important first step in information-based approaches to resource allocation, and other operational decisions. Moreover, in order to effectively characterize data so as to reveal important associations and create accurate and reliable models, it is important to understand the data and be able to generate samples that are relatively homogenous, except with regard to those unique features that have value from a modeling perspective. For example, when reviewing victims of prior firearms injuries, a first pass through the data revealed no association between the risk of being shot and the likelihood that the victim carried a weapon.²⁰ Dividing the sample based on the pattern of criminal offending, however, revealed an entirely different story. Those victims previously involved in aggressive or violent patterns of offending were much more likely to have been shot if they also were known to carry a weapon, which might be related to or indicative of particularly aggressive interactional patterns of behavior.

Injured drug dealers, on the other hand, were much less likely to carry a weapon, possibly indicating poor defensive skills in a very predatory criminal activity. Therefore, the association between getting shot and carrying a weapon was obscured when the data were in aggregate form. Although it initially appeared that the data were relatively homogeneous in that they were confined to juvenile offenders, important relationships within the data were not revealed until it had been analyzed further. In this case, the associated pattern of offending was an important factor in determining the relationship between sustaining a firearms-related injury and weapon possession.

OFFICER SAFETY

Characterization of different victim risk patterns also has officer safety implications. Anecdotal reports link weapon selection to the reason for using a weapon. For example, those criminals electing to carry a weapon as an extension of an aggressive or violent approach to the world are more likely to select a weapon that is similarly menacing. These offenders frequently prefer something large and scary-looking with an increased capacity. They are willing to compromise accuracy in an effort to acquire something that fits their perceived image and lifestyle.

On the other hand, criminals electing to carry a weapon for defensive purposes, such as those involved in illegal drug markets, generally prefer something that can be readily concealed and is reliable. After being shot, a 15-year-old drug dealer revealed his decision-making process for weapon selection. Interestingly, many of the factors that he considered were similar to the ones cited by a large federal agency that had recently switched manufacturers and chosen the same brand that this juvenile drug seller had selected.

Aside from being ironic, this finding has significant implications for officer safety. By characterizing the likely weapon selection process, operational personnel are able to gain added insight into what they might encounter on the street when confronting a particular type of offender. The knowledge that many young, violent offenders are willing to select form over function while juvenile drug sellers have a preference for easily concealed, very reliable weapons has the potential to determine which party is likely to walk away from a violent encounter. To quote Miguel de Cervantes, "forewarned is forearmed, to be prepared is half of the victory."

Why do we care about this? First, these findings have direct implications for treatment programs. Given that the differences noted were behavioral patterns and styles associated with the risk for injury, it would be inappropriate to develop a generic "firearms injury survivor" group. Just as it would be crazy to create a "drug-involved offenders" group that included drug dealers as well as drug users, it would be similarly risky to combine victims who likely had been shot because of an extremely aggressive interactional style with those who had been shot as a result of poor defensive skills. These findings also have officer safety implications, in that foreknowledge of an offender's likely weapon preference can be extremely valuable on the street.

3.7 LINK ANALYSIS²¹

Sometimes we want to ask the question, "What things go together?" Typically, these might be features of a crime, such as where and when the crime occurred, what types of property, people, or vehicles were involved, the methods used, and so on. One way of answering such "link analysis" questions is to use web graphs, which show associations between items (such as individuals, places, or any other element of interest) by points in a diagram, with lines depicting the links between them. These tools can have added value in that the strength of the association can be depicted by the strength of the line. For example, a solid line conveys a much stronger relationship than a dashed line.

One common pitfall in link analysis is to overinterpret the identified relationships or results, particularly those with unequal distributions. This issue is illustrated further in Chapter 1, but the best way around this issue, like most others in data mining, is to know your domain and know your data. It always is extremely important to explore the data initially and interpret any results cautiously. Potential options for addressing this can include the use of percentages rather than the actual frequencies. Again, this is illustrated in greater detail in Chapter 1.

Many software tools also provide a toggle option or sliding scale that gives the analyst the opportunity to adjust the thresholds used to determine the relative strengths of multiple relationships. This can be a tremendous tool, particularly during the exploration process, as it allows the analyst to reduce the “noise” so that the important relationships can be visualized easily. Other products allow the user to adjust whether the strength of the relationship is based on frequencies or percentages. Again, this can be a tremendous asset when evaluating and comparing relationships in the data.

Related to the challenge associated with accurately interpreting the results of link analysis is the increasing recognition that almost any two entities can be linked with relative ease, given sufficient connections or “hops”.²² Developed decades ago, the “Kevin Bacon”²³ game is a Hollywood parlor game that illustrates the general concept of “six degrees of separation” between people, which effectively underscores this particular challenge in the interpretation of link analysis. In this particular game, almost any other actor can be linked to Kevin Bacon using six connections; and in many cases, far less. Similarly, most entities (people, places, telephone numbers, etc.) can be linked using six or fewer connections. Figure 3.5 notionally illustrates a link chart that I was

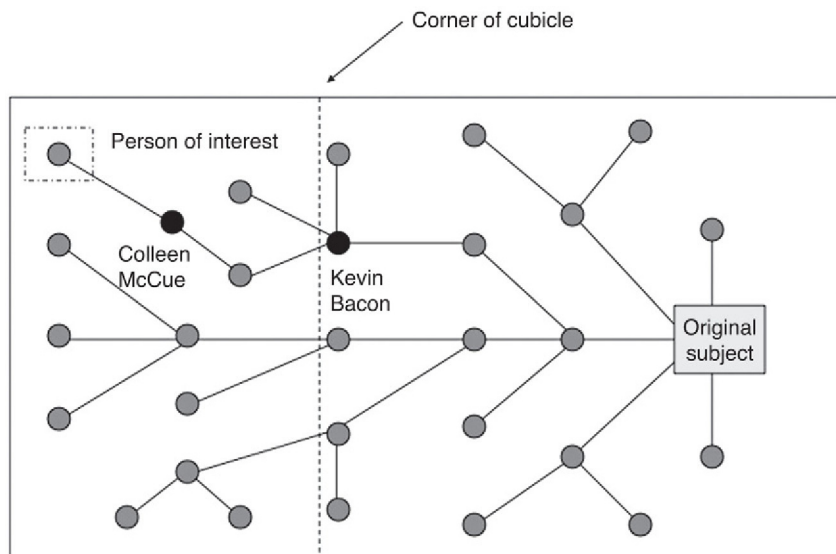


FIGURE 3.5

Notional illustration of a link chart that the author reviewed highlights the challenges associated with interpreting connections surfaced using link analysis. In this particular example, the link chart not only spanned an entire wall of the analyst’s cubicle, but also turned the corner and extended to a second wall; ultimately linking the Original Subject of a background investigation to a known Person of Interest. The actor Kevin Bacon²⁵ and the author of this text have been embedded between the Original Subject and the person of interest to underscore the ease with which individual entities can be linked with increasing connections or “hops.”

briefed on several years ago that involved a general background investigation into an employee. In this particular case, the results of the “analysis” spanned an entire wall of the cubicle, turned the corner, and extended onto a second wall, resulting in a discovered “link” between the original subject to a known person of interest that resulted in additional scrutiny and deeper investigation of the employee. Given the number of connections required to establish this putative link, however, this particular result should be viewed with considerable caution.

3.8 NON-OBVIOUS RELATIONSHIP ANALYSIS (NORA)²⁴

Unfortunately, life generally is not as simple as a web graph or link analysis would indicate. For example, it is not unusual for a suspect to intentionally alter the spelling of his or her name or attempt to vary his or her identity slightly in an effort to avoid detection. Similarly, Richard, Dick, Rick, Rich, Ricky, and so on, all are legitimate variations of the same name. This challenge becomes even more of an issue in investigative and intelligence databases where the information can be even less uniform and reliable.

In addition, it is not uncommon in crime, particularly in organized crime or terrorism, for individuals to try to avoid having a direct relationship with other members of the group or organization. In fact, the Al Qaeda handbook, which is available on the Internet, specifically advises that operatives significantly limit or avoid contact with others in the cell in an effort to reduce detection and maintain operational security. Clearly, this creates a significant limitation for the use of standard automated association detection techniques.

To address this challenge, however, automated techniques referred to as Non-Obvious Relationship Analysis, or NORA, have emerged from the gambling industry in Las Vegas. Used to identify cheaters, these tools have obvious implications for law enforcement and intelligence analysis given their ability to identify links and relationships not readily identifiable using traditional link analysis software. These tools also can identify subtle changes in numeric information, such as social security numbers. In many cases, these transpositions are unintentional keystroke errors. In others, however, numeric information is changed slightly to reduce the likelihood that information will be linked directly, which can be indicative of identity theft or similar types of fraud.

3.9 TEXT MINING

Most information that has value in law enforcement and intelligence analysis resides in unstructured or narrative format. Frequently, it is the narrative portion of a police report that contains the most valuable information pertaining to motive and modus operandi, or MO. It is in this section of the report that

the incident or crime is described in the behavioral terms that will be used to link it to others in a series, to similar crimes in the past, or to known or suspected offenders. In addition, crime tip information and intelligence reports almost always arrive in unstructured, narrative format. Because it is unlikely that an informant will comply with a structured interview form or questionnaire, the onus is upon the analyst either to transcribe and recode the information, or to identify some automated way to analyze it.

Until recently, this information was largely unavailable unless recoded. This is time-consuming, and can alter the data significantly. Recoding generally involves the use of arbitrary distinctions to sort the data and information into discrete categories that can be analyzed. Unfortunately, many aspects of the data, information, and context can be compromised through this process.

Recent advances in text mining tools that employ natural language processing now provide access to this unformatted text information. Rather than crude keyword searches, the information pulled out through the use of text mining incorporates syntax and context. As a result, more complex concepts can be mined, such as “jumped the counter” or “passed a note,” valuable MO characteristics that could be associated with takeover robberies or bank robberies, respectively.

Like the suspect debriefing scenario outlined in the Introduction, these tools promise to advance the analytical process in ways not considered until very recently. For example, the information obtained through the interview process can be inputted directly into the analysis and integrated with other narrative and categorical data. Moreover, tip databases can be reviewed, characterized, and culled for common elements, themes, and patterns. These tools promise to significantly enhance statement analysis, as they can identify common themes and patterns, such as those associated with deception or false allegation. It will be truly exciting to see where these tools take the field of crime and intelligence analysis in the future.

3.10 CLOSING THOUGHTS

Many of the new and emerging data mining and predictive modeling programs are highly intuitive, powerful, and incredibly fast. Capabilities including Social Network Analysis and Sentiment Analysis represent powerful extensions of the approaches reviewed in this chapter and will be revisited in the Case Examples, as well as the discussion of Advanced Concepts in Chapter 15. Ultimately, the use of these analytical capabilities promises to bring more science and less fiction to law enforcement and security operations, policy, and planning. The ability to surface previously unknown or hidden patterns, and model possible scenarios and outcomes including displacement will not only enhance

operational strategy and deployment but also can save lives as potential risks are identified, characterized, and anticipated in support of information-based prevention, thwarting, mitigation, and response. In closing, though, it is important to remember that data mining, predictive analytics, and data science writ large is more than math, particularly in the operational public safety and national security environment; it almost always goes back to behavior – generally, bad behavior or intentions. Even the most sophisticated data sources are reflections of behavior, including attack planning, surveillance, theft of tangible assets, data or intellectual property to name a few. While the data, including sources and methods, can be interesting and compelling, we ultimately are trying to surface trends, patterns, relationships, affinities, and even intentions or indicators of future or planned activity, but it almost always goes back to behavior. Losing sight of that, including the operational context, requirements, and constraints, can result in spurious findings and faulty interpretation of the results. While it is wonderful that our peers in the commercial sector can use these same methods to more effectively segment their market space, sell more products, and increase efficiencies in other areas, our mission is to protect and serve. Therefore, the algorithms, technology and tradecraft employed to surface these trends and patterns are important, but ultimately it should always be more than math.

Bibliography

- 1 Peterson J. Data mining. <http://people.sc.fsu.edu/~pbeerli/classes/isc5315-notes/jpeterson/clustering.pdf>.
- 2 Siegel E. Predictive analytics: the power to predict who will click, buy, lie or die. Hoboken, NJ: John Wiley & Sons; 2013; Helberg C. Data mining with confidence. 2nd ed. Chicago, IL: SPSS, Inc.; 2002; Two Crows Corporation. Introduction to data mining and knowledge discovery. 3rd ed. <http://www.twocrows.com/intro-dm.pdf>; n.d.
- 3 Howard C. Data mining – what is it and why would anyone want to do it? Kilmer Middle School Career Day, http://datamininglab.com/media/pdfs/Data_Mining_%28Career_Day%29.pdf [accessed 18.11.08].
- 4 Beck C, McCue C. Predictive policing: what can we learn from Wal-Mart and Amazon about fighting crime in a recession? Police Chief November; 2009.
- 5 Beck C, McCue C; 2009.
- 6 Beck C, McCue C; 2009.
- 7 Davenport TH, Harris JG. Competing on analytics: the new science of winning. Boston: Harvard Business School Press; 2007.
- 8 Basic Disaster Supplies Kit. <http://www.ready.gov/basic-disaster-supplies-kit>
- 9 Leonard D. The only lifeline was the Wal-Mart: The world's biggest company flexed its massive distribution muscle to deliver vital supplies to victims of Katrina. Inside an operation that could teach FEMA a thing or two. Fortune 2005; October 3. http://money.cnn.com/magazines/fortune/fortune_archive/2005/10/03/8356743/
- 10 Davenport TH, Harris JG. Competing on analytics: the new science of winning. Boston: Harvard Business School Press; 2007.
- 11 Talent makes a difference. c.f., Davenport T. Analyze this: more companies are using analytics to drive their decision-making processes. But there's a right and wrong way to do it; 2005. <http://www.sas.com/offices/europe/uk/technologies/analytics/davenport.pdf>

- 12 Davenport TH, Jarvenpaa SL. Strategic use of analytics in government. IBM Center for the Business of Government: Managing Performance Series; 2008. <http://www.businessofgovernment.org/sites/default/files/Strategic%20Analytics.pdf>
- 13 Casey E. Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and Internet crime investigation. Knowledge Solutions; 2002. www.corpus-delicti.com/case_based.html.
- 14 McLaughlin CR, Reiner SM, Smith BW, Waite DE, Reams PN, Joost TF, et al. Firearm injuries among Virginia juvenile drug traffickers, 1992 through 1994 (Letter). *Am J Public Health* 1996;86:751–752; McLaughlin CR, Smith BW, Reiner SM, Waite DE, Glover AW. Juvenile drug traffickers: characterization and substance use patterns. *Free Inquiry Creative Sociol* 1996;24:3–10; McLaughlin CR, Reiner SM, Smith BW, Waite DE, Reams PN, Joost TF, et al. Factors associated with a history of firearm injuries in juvenile drug traffickers and violent juvenile offenders. *Free Inquiry Creative Socio Special Issue: Gangs Drugs Violence* 1996;24:157–165.
- 15 McCue C, Smith GL, Diehl RL, Dabbs DE, McDonough JJ, Ferrara PB. Why DNA databases should include all felons. *Police Chief* 2001;68:94–100.
- 16 Tabussum Z. CIA turns to data mining, www.parallaxresearch.com/news/2001/0309/cia_turns_to.html; 2003.
- 17 Despite avalanche of tips, police stymied in Laci Peterson case. www.KXTV10.com; 2003 [accessed 09.01.03].
- 18 Eastham T. Washington sniper kills 8, truck sketch released. October 12. www.sunherald.com; 2002.
- 19 METAspectrumSM Market Summary. Data mining tools: METAspectrumSM evaluation. META Group, Inc.; 2004.
- 20 McLaughlin CR, Daniel J, Reiner SM, Waite DE, Reams PN, Joost TF, et al. Factors associated with assault-related firearms injuries in male adolescents. *J Adolesc Health* 2000;27:195–201.
- 21 Helberg C. 2002.
- 22 Harris D Six degrees of separation, NSA-style. GIGAOM, 2013. <http://gigaom.com/2013/07/17/six-degrees-of-separation-nsa-style/> [accessed 17.07.13].
- 23 Terdiman D. At SXSW, thousands get a Kevin Bacon number of one. CNET, 2014. <http://www.cnet.com/news/at-sxsw-thousands-get-a-kevin-bacon-number-of-one/> [accessed 08.03.14].
- 24 Franklin D. Data miners: new software instantly connects key bits of data that once eluded teams of researchers. *Time*, December 23, 2002.
- 25 Terdiman D. 2014.

Process Models for Data Mining and Predictive Analysis

“Traditional scientific method has always been at the very best, 20 - 20 hindsight. It’s good for seeing where you’ve been. It’s good for testing the truth of what you think you know, but it can’t tell you where you ought to go.”

Robert M. Pirsig

This chapter includes an overview of three complementary analytical process models: the Central Intelligence Agency (CIA) Intelligence Process,¹ the Cross Industry Standard Process for Data Mining (CRISP-DM),² and SEMMA,³ as well as an integrated process model for Actionable Mining and Predictive Analysis that is specific to the application of data mining and predictive analytics in the operational public safety and security setting.

All of these models emphasize the analytical *process* over specific tools or techniques. In addition, they have been conceptualized as *iterative* processes, meaning simply that the analytical process can and should be repeated as conditions change or new information becomes available. Rather than representing a failure of the analysis or created model, the need for an update or refresh can serve to validate successful analysis, particularly if it results in a change in the trend or pattern, including displacement. When used to support information-based operations, data mining tools are similar to a public safety time machine in that they offer the ability to characterize, anticipate, influence, and even prevent certain crimes. For example, “risk-based deployment” strategies are based on the concept that identifying and characterizing what is likely to happen in new areas supports proactive deployment. Once crime has been suppressed in a particular area, the next steps could include analysis and evaluation of displacement, which occurs when a particular crime pattern or trend has been moved to another location. This would include a similar analytical process on a new set of data that accurately reflect the current conditions, including the positive changes associated with an earlier iteration of the model and the resulting operational plan. This development of effective, information-based tactics and strategies can allow managers and command staff to target their

resources specifically and more effectively, which increases the likelihood of successful operations.

An iterative process also is important because crime and criminals change. As the players change, so do the underlying patterns and trends in crime. For example, preferences for illegal drugs often cycle between stimulants and depressants. Markets associated with distribution of cocaine frequently differ from those associated with heroin. Therefore, as the drug of choice changes, so will the associated markets and related crime. Similarly, seasonal patterns and even weather can change crime patterns and trends, particularly if they affect the availability of potential victims or other targets. For example, people are less likely to stroll city streets during a torrential downpour, which limits the availability of potential “targets” for street robberies; temperature extremes might be associated with an increased prevalence of vehicles left running to keep either the air conditioning or heat on, which increases the number of available targets for auto theft. Successful police work also will require periodic “refreshing” of the model. The models will change subtly as offenders are apprehended and removed from the streets. These revised models will reflect the absence of known offenders, as well as the emergence of any new players. For example, illegal drug markets frequently experience changes in operation and function associated with changes in players. Similarly, serial killers may have a unique “signature” associated with their crimes that can be used to link several crimes and segment them into a separate and distinct series.⁴

In many ways, identifying changing patterns or players is the most exciting outcome of the data mining process because it underscores the surprise and discovery that can be associated with the analysis. This process is not linear, with a clear beginning and end. Rather, it is better represented by an iterative cycle in which the answers to the initial questions almost inevitably beget additional questions, representing the beginning of the next analysis cycle. Another way to visualize this process of forward iteration is to consider a funnel-shaped spiral rather than a flat circle. The spirals get increasingly tight as the solution moves closer to the idealized target. The concepts of spiral processing, integration, and development are increasingly being used to describe an iterative process with forward progression. Although this language may change over time, the important feature of a spiral model of iterative crime or intelligence analysis is that the subsequent iterations reflect progress and movement toward the best fit.

Sequential iterations of the analysis process can be used to further refine models, which ultimately may result in more specific, targeted tactics, strategies, and responses. For example, it is common knowledge that there are different types of homicides that can be defined by their motives (e.g., domestic, drug related, sexual).⁵ Categorizing homicides has value from an investigative

perspective in that knowledge of the type of homicide or motive generally serves to shorten the list of potential suspects, which ultimately enhances investigative efficacy. Similarly, analysis of the victims of violence has resulted in the identification of groups. This underscores the finding that different people are at risk for different reasons at different times and that broad violence injury prevention programs might be less than adequate if they do not address the unique constellation of risk associated with specific victim groups.⁶ Therefore, additional analysis and characterization of the unique factors associated with specific groups of victims can be used to guide the creation of meaningful prevention and response strategies. This subject is addressed in greater detail in Chapter 11.

Data mining and predictive analytics are as much analytical process as machine learning and math. In fact, the general rule is that the analysis process is 80:20 – 80% preparation and 20% analysis. The specific elements or tasks associated with the process will be addressed separately and in greater detail in subsequent chapters; however, a general overview of these models is provided here in an effort to highlight their similarities and functional relationships. Specific analytical protocols based on these process models also will be provided in relevant chapters.

4.1 CIA INTELLIGENCE PROCESS⁷

To highlight the multiple steps or detailed process associated with the transformation of data and information into intelligence, the CIA has developed an Intelligence Process model. The CIA intelligence model has been divided into six stages: Requirements, Collection, Processing and Exploitation, Analysis and Production, Dissemination and Consumption, and Feedback.⁸

4.1.1 Requirements

During the requirements or “planning and direction” phase,⁹ intelligence information priorities are determined. It is during this phase that conflicting or competing priorities are identified and resolved or rank ordered. The CIA model underscores the dynamic and changing nature of the Intelligence Process, emphasizing that the “answers” to questions frequently represent the starting point for subsequent iterations of the process. Therefore, these identified needs or requirements can and should be reevaluated as conditions or priorities change.

4.1.2 Collection

The intelligence community places particular emphasis on the collection of raw data and information that form the basis for finished intelligence products, creating agencies assigned exclusively to the collection, processing and

exploitation, and analysis of specific intelligence sources. The CIA model specifies five basic collection modalities¹⁰:

1. Open-source information (OSINT) – OSINT includes information available publicly and can include but is not limited to newspapers, radio, television, and the Internet.
2. Human-source intelligence (HUMINT) – HUMINT, as the name implies, includes intelligence gathered from human sources. This collection discipline has been divided further and can include clandestine activities as well as overt collection efforts, debriefing, and official contacts.
3. Signals intelligence (SIGINT) – SIGINT is a general category that includes information obtained from intercepted signals. Subdisciplines within this category include Communications Intelligence (COMINT) and Electronic Intelligence (ELINT).
4. Geospatial intelligence (GEOINT) – GEOINT includes information obtained through satellite, aerial, and ground-based collection methods that is used to describe, visualize, and accurately locate physical features and human activity on the Earth. GEOINT includes imagery intelligence (IMINT).
5. Measurement and signature intelligence (MASINT) – MASINT includes technical data that are not SIGINT or IMINT. Sources for MASINT intelligence can include, but are not limited to, radar, nuclear, seismic, and chemical and biological intelligence.

4.1.3 Processing and Exploitation

The processing and exploitation phase includes the preparation and transformation of data into a format that can be analyzed. The inclusion of this step underscores the complexity of some forms of collected intelligence information. Single agencies may be almost entirely responsible for the processing and exploitation of specific categories of technically derived intelligence, which supports the critical importance of subject matter or domain expertise in the process.

4.1.4 Analysis and Production

It is during the analysis and production phase of the process that raw data and information are converted into finished intelligence products. These products may be relatively brief and limited in depth or coverage, or they may be longer and represent a more comprehensive study of a particular topic or issue. These finished intelligence studies also may include the integration of multiple sources of information into derived products, which affords a greater depth of analysis and insight.

4.1.5 Dissemination

Dissemination includes the distribution of intelligence products to the intelligence community, policy makers, the military, or other consumers of intelligence. Intelligence products may be developed rapidly, based on emerging or rapidly changing events; they take the form of regular reports like the President's Daily Briefing; or they may reflect the results of a long-term study or analysis, such as the National Intelligence Estimates.

4.1.6 Feedback

The inclusion of feedback in the model supports the continuous and iterative nature of the Intelligence Process. Information provided by the consumers of finished intelligence can be used to guide new areas of inquiry or identify gaps in information that need to be filled or otherwise addressed. Feedback also may be used to adjust priorities or emphasis.

4.1.7 Summary

The CIA Intelligence Process is well suited to the functions and needs of the intelligence community. The scope, breadth, and applicability of this approach to such a diverse range of functions and responsibilities within the intelligence community are admirable, and have been highlighted by the model's relative longevity, as well as the frequency with which it has been adopted, cited, and imitated. The level of detail associated with this process model, however, is not sufficient to support specific analytical strategies or approaches, including data mining and predictive analytics. In all fairness, we should point out that it would be extremely difficult if not impossible to develop a general analytical process model or strategy that addressed accurately, and in specific detail, the unique challenges and idiosyncrasies associated with each collection discipline or modality.

4.2 CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING

What the CIA model brings in terms of specificity to intelligence, and by extension applied public safety and security analysis, the Cross-Industry Standard Process for Data Mining (CRISP-DM) process model contributes to data mining as a process, which is reflected in its origins. Several years ago, representatives from a diverse array of industries gathered to define the best practices, or standard process, for data mining.¹¹ The result of this task was the CRISP-DM. The CRISP-DM process model was based on direct experience from data mining practitioners, rather than scientists or academics, and represents a "best practices" model for data mining that was intended to transcend professional

domains and operationalize the fact that data mining and predictive analytics are as much analytical process as they are specific algorithms and models. Like the CIA Intelligence Process, the CRISP-DM process model has been broken down into six steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment.¹²

4.2.1 Business Understanding

Perhaps the most important phase of the data mining process includes gaining an understanding of the current practices and overall objectives of the project. During the business understanding phase of the CRISP-DM process, the analyst determines the objectives of the data mining project. Included in this phase are an identification of the resources available and any associated constraints, overall goals, and specific metrics that can be used to evaluate the success or failure of the project.

4.2.2 Data Understanding

The second phase of the CRISP-DM analytical process is the data understanding step. During this phase, the data are collected and the analyst begins to explore and gain familiarity with the data, including form, content, and structure. Knowledge and understanding of the numeric features and properties of the data (e.g., categorical versus continuous data) will be important during the data preparation process and essential to the selection of appropriate statistical tools and algorithms used during the modeling phase. Finally, it is through this preliminary exploration that the analyst acquires an understanding of and familiarity with the data that will be used in subsequent steps to guide the analytical process, including any modeling, evaluate the results, and prepare the output and reports.

4.2.3 Data Preparation

After the data have been examined and characterized in a preliminary fashion during the data understanding stage, the data are then prepared for subsequent mining and analysis. This data preparation includes any cleaning and recoding as well as the selection of any necessary training and test samples. It is also during this stage that any necessary merging or aggregating of data sets or elements is done. The goal of this step is the creation of the data set that will be used in the subsequent modeling phase of the process.

4.2.4 Modeling

Specific methods will be reviewed in Chapter 7, but the actions at this step can range from simple summary or descriptive statistics and visualization techniques, to extremely complex automated methods including neural nets and powerful classification methods, as well as ensemble methods that enable the analyst to strategically combine or “bundle” models in order to improve

overall accuracy and performance.¹³ Selection of the specific algorithms employed should be based on the nature of the question and outputs desired, but modeling algorithms can be categorized generally into two groups: supervised learning or rule induction models and decision trees; and unsupervised learning or clustering techniques. Supervised learning methods to include scoring algorithms or decision tree models are used to create decision rules based on known categories, associations or relationships that can be applied to unknown data. On the other hand, unsupervised learning or clustering techniques are used to uncover natural trends, patterns or relationships in the data when group membership or category has not been identified previously. Again, the selection of specific methods and approaches should be based on the nature of the data and the analytic question to be answered. Ideally, the analyst will let the question drive the analytic strategy and approach.

4.2.5 Evaluation

During the evaluation phase of the project, the models created are reviewed to determine their accuracy as well as their ability to meet the goals and objectives of the project identified in the business understanding phase. Put simply: Is the model accurate, and does it answer the question posed?

4.2.6 Deployment

Finally, the deployment phase includes the dissemination of the information. The form of the information can include tables and reports as well as the creation of rule sets or scoring algorithms that can be applied directly to other data.

SEMMA

More recently, the analytic software company SAS has developed SEMMA, an acronym for: Sample, Explore, Modify, Model, and Assess. In contrast to the CRISP-DM process model, which is application independent, SEMMA represents the “logical organization of the functional toolset of SAS Enterprise Miner [the SAS data mining work bench] for carrying out the core tasks of data mining.”¹⁴ As illustrated in Table 4.1, CRISP-DM and SEMMA parallel each other in many ways, underscoring essential aspects of the knowledge discovery process including an emphasis on data preparation and exploration.

Table 4.1 Comparison of the CRISP-DM and SEMMA Process Models

CRISP-DM	SEMMA
Business understanding	
Data understanding	Explore
Data preparation	Sample, Modify
Modeling	Model
Evaluation	Assess
Deployment	

4.3 SAMPLE

The Sample step is optional, but involves the extraction of a portion of the data that is small enough to enable quick and easy manipulation and analysis, but large enough to include enough records or cases of interest to support meaningful exploration and results. This step also includes partitioning of the data to create the training, validation, and test samples that are used to build the model, assess the model and guard against overfitting, and test the model to achieve an “honest assessment” of how well it generalizes, respectively.

4.4 EXPLORE

The Explore step includes the use of visualization, summary statistics, and statistical analysis to discover trends, patterns, and relationships in the data, and to enable confirmation and discovery.

4.5 MODIFY

The Modify steps builds on the data exploration, and involves additional selection and transformation of the data to better focus the modeling step. This may include additional segmentation of the sample, introduction of new variables, and the creation of derived variables.

4.6 MODEL

The Model step includes the use of machine learning algorithms to create models of the sample data that can be used to reliably classify unknowns and/or predict outcomes.

4.7 ASSESS

The Assess step in the SEMMA process applies directly to the evaluation of the model’s performance against samples reserved for validation and testing.

4.7.1 Summary

The CRISP-DM model has worked very well for many business applications¹⁵; however, law enforcement, security, and intelligence analysis can differ in several meaningful ways. Analysts in these settings frequently encounter unique challenges associated with the data, including timely availability, reliability, and validity. Moreover, the output needs to be comprehensible and easily understood by nontechnical end users while being directly actionable in the applied setting in almost all cases. Ideally, the end user will be able to quickly and intuitively understand the results, and be able to incorporate their tacit knowledge, domain expertise and experience, and extend from the results in support of novel insight and action. Finally, unlike in the business community,

Table 4.2 Comparison of the CRISP-DM and CIA Intelligence Process Models

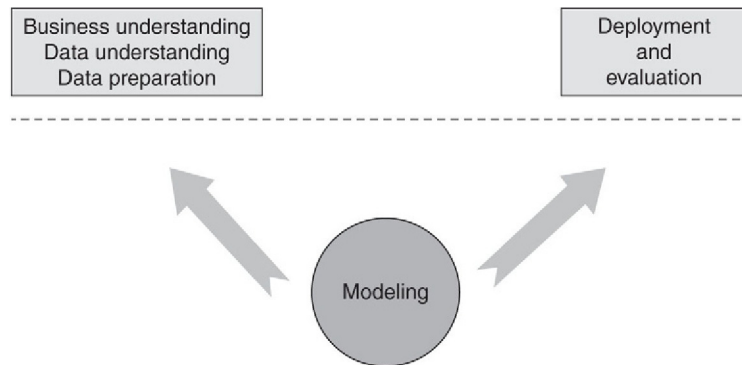
CRISP-DM	CIA Intelligence Process
Business understanding	Needs
Data understanding	Collection
Data preparation	Processing and exploitation
Modeling	Analysis and production
Evaluation	Dissemination
Deployment	Feedback

the cost of errors in the applied public safety setting frequently is life itself. Errors in judgment based on faulty analysis or interpretation of the results can put citizens as well as operational personnel at risk for serious injury or death.

The CIA Intelligence Process has unique features associated with its use in support of the intelligence community, including its ability to guide sound policy and information-based operational support. The importance of domain expertise is underscored in the intelligence community by the existence of specific agencies responsible for the collection, processing, and analysis of specific types of intelligence data. The CRISP-DM process model highlights the need for subject matter experts and domain expertise, but emphasizes a common analytical strategy that has been designed to transcend professional boundaries and that is relatively independent of content area or domain. The CIA Intelligence Process and CRISP-DM models are well-suited to their respective professional domains; however, they are somewhat limited in directly addressing the unique challenges and needs related to the direct application of data mining and predictive analytics in the operational public safety and security arena. Therefore, an integrated process model specific to public safety and security data mining and predictive analytics is outlined next. Like the CIA model, this model recognizes not only a role but also a critical need for analytical tradecraft in the process; and like the CRISP-DM process model, it emphasizes the fact that effective use of data mining and predictive analytics truly is an analytical process that encompasses far more than the mathematical algorithms and statistical techniques used in the modeling phase (Table 4.2).

4.8 ACTIONABLE MINING AND PREDICTIVE ANALYSIS FOR PUBLIC SAFETY AND SECURITY

The CRISP-DM model highlights the importance of domain expertise and analytical tradecraft. As depicted in Figure 4.1, the steps both preceding and following the modeling phase require significant domain expertise and understanding of operational requirements. If we assume, as mentioned earlier, that 80% of the data mining process is in the data preprocessing and preparation steps, an effective data mining process model for public safety and security

**FIGURE 4.1**

In the CRISP-DM process model, the steps preceding and following the modeling phase require significant domain expertise and understanding of the operational requirements.

will specifically address these steps. This preparation should include focusing on the unique limitations and challenges associated with applied public safety and security analysis. In most situations, once the data preprocessing and output have been addressed, commercially available software packages can be used for the actual modeling. To address this requirement for operationally relevant data preprocessing and output, the Actionable Mining and Predictive Analysis for Public Safety and Security model has been created. The Actionable Mining model includes the following steps:

1. Question or challenge
2. Data collection and fusion
3. Operationally relevant preprocessing
 - a. Recoding
 - b. Variable selection
4. Identification, characterization, modeling
5. Public-safety-specific evaluation
6. Operationally actionable output

4.8.1 Question or Challenge

Sometimes the analyst is faced with specific questions: Are these crimes linked? When are burglaries most frequent? Do people buy drugs in the rain? Other times, however, the task initially manifests itself as a vague question or challenge that requires some preliminary work to identify or structure a specific question. For example, it is not unusual to be presented with a series of telephone calls or financial transactions and then to be asked whether there is any sort of pattern worthy of additional investigation. Therefore, during the initial phase of the process, the general question or challenge is identified and converted into a specific question that will be answered by the data mining

process. This question will be used to structure the analytical design plan, guide the process, and ultimately evaluate the fit and value of the answer.

It is also during this stage that current procedures and reports should be reviewed. In the business community, it is desirable to work directly with the client or recipient of the data mining results at this phase to ensure that the end product addresses the specific business questions or challenges and otherwise meets their needs. In law enforcement, intelligence, and security analysis, it is imperative to collaborate directly with the anticipated recipient or end user of the analytical products, particularly operational personnel. Working with the intended recipients of the data mining results can make the difference between generating analytical end products that might be interesting but have little to no value to their recipients, and those analytical results that can be translated directly into the operational environment to support and enhance information-based decisions. One possible consequence of overlooking or omitting this step includes “producing the right answers to the wrong questions.”¹⁶ In the applied setting, if the results cannot be used in the field, then data mining becomes little more than an academic exercise. Therefore, it is never too early to begin to consider what the output should look like and how it will be used, as this can have implications for the remaining steps. The question or challenge phase also is a good point in the process to identify evaluation criteria or other metrics of success that can be reviewed later to evaluate the success of the process. Again, these criteria should include the operational value of the analytical products, as well as traditional measures of accuracy and reliability.

“WICKED” PROBLEMS¹⁷

Asking the “right” questions sets the trajectory for the analysis, and is a key step in initiating the process.¹⁸ In many ways, this can be like finding the “word problem” embedded in an investigation, dataset, or other source. Again, these questions can be specific and very focused, or more vague and open-ended. One common thread, though, is that public safety and security questions tend to be complex and have no simple or obvious answers.

In their paper, which was written originally for the social policy planning community, Rittel and Weber list 10 characteristics of so-called “wicked” problems, many of which have significant implications for and are directly applicable to crime and intelligence analysis. One of the characteristics of wicked problems is that they frequently involve a diverse array of stakeholders who bring different perspectives. Rittel and Weber extend this to suggest that these different professionals also may extend their unique perspectives of the problem space to their preferred approaches to the solution, and that in some cases these differing approaches may contrast markedly. In other words, the characterization of the problem frequently is defined by the chosen solution. A review of the frequently circular relationship between food security and conflict in Africa supports this point.¹⁹ In addition, wicked problems are fluid and difficult to define, have no easy answers, and most of the likely solutions are almost invariably associated with consequences, unintended or otherwise. Moreover, you only get “one shot” with wicked problems because “every implemented solution is consequential [and] leaves traces that cannot be undone.”²⁰ Actions are irreversible

and have consequences, including the fact that the situation is irrevocably altered and the problem going forward is fundamentally changed as a result of previous interventions; a concept familiar to professionals associated with community-based interventions from Community-Oriented Policing to the Anbar Awakening. Wicked problems also are “essentially unique,” which means that while we can leverage some knowledge regarding similar or comparable situations, wicked problems are defined by their complexity and each situation will be new and unique. Therefore, as the financial services disclaimers state, “past performance is not an indicator of future results”; each situation is fundamentally new and unique. Related to the complexity and essentially unique nature of wicked problems is the general interrelatedness of wicked problems. Each wicked problem is comprised of other, interconnected problems, and may be characterized as a symptom of another problem. Finally, the favored solution tends to drive the definition and characterization of a wicked problem, which may result in circular logic; “[t]he analyst’s ‘world view’ is the strongest determining factor... in resolving a wicked problem.”²¹ This last point is particularly relevant to the analyst, especially as it relates to the selection of analytic tools, technology, and tradecraft and will be revisited in Chapter 7.

4.8.2 Data Collection and Fusion

In the CIA Intelligence Process model, data collection is a separate and distinct step; however, data collection is merged with preliminary analysis and exploration in the CRISP-DM process model. This difference in emphasis most likely speaks to the different professional disciplines associated with the two analytical process models and the associated cost and difficulty associated with their respective collection efforts. In the intelligence community, the collection of data and information for analysis can represent a significant function of an entire agency and consume a major portion of the budget, particularly as the technical complexity and required resources associated with the collection process increase. As outlined earlier, collection is so important to the entire intelligence process that it has been divided further into separate collection disciplines. The data collected for analysis in the business setting generally are less difficult to obtain and may even reflect some foresight and analytical input regarding structure, form, and content.

Public safety and security data generally lie somewhere in between these two perspectives. Most public safety and security organizations do not have dedicated collection efforts or the ability to effectively utilize some technically challenging sources currently available to the intelligence community (e.g., SIGINT). Public safety and security data and information generally assume the form of standard incident reports, citizen complaint data, and some narrative information. That is not to say that unusual or unorthodox data resources cannot play a significant role in public safety analysis. It is not unreasonable to consider that the economy, special events, seasonal changes, or even weather might affect crime trends and patterns, particularly if these trends significantly impact the movement and associated access to victim populations. For example, street robberies in a nightclub area might decrease during heavy rain if the robberies normally are associated with patrons leisurely strolling around. Similarly, auto theft might increase when it is extremely cold, as citizens leave

keys in their cars while preheating them. Therefore, thinking outside the box regarding useful data can result in more comprehensive and accurate models of criminal activity. In this case, the size of the box is limited only by the creativity of the analysts, their willingness to explore additional approaches, and their legitimate and ethical access to data and information.

Most, if not all, data analyzed in the public safety and security arena were collected for some other purpose, which can affect data form, content, and structure. Crime incident reporting forms generally are not created with data mining in mind. Moreover, some of the most valuable information in an incident report frequently is included in the unstructured narrative portion of the report. It is in this narrative section that information relating to modus operandi and other important behavioral indicators can be found. Unfortunately, it is this section of the report that also contains misspellings, typographical errors, and incomplete and missing information, as well as other inconsistencies, all of which significantly limit the analysts' ability to effectively exploit the information.

Integration or fusion of multiple data resources also is started during this data collection and fusion phase and can be continued through the data preprocessing stage, when the data set is created for modeling and analysis. Fusion of data and information across collection modalities, data subsets, or separate locations can be desirable or even necessary. Common types of data integration include any necessary linking of required tables with relational data resources, including incident-based reporting systems, as well as any required linking of data that have been stored in separate files. This can include files that are maintained in time-limited samples due to the amount of information. For example, citizen complaint data or calls for service might be stored in monthly files, which will need to be combined to support analysis of longer patterns and trends. Similarly, separate victim and suspect tables might need to be linked to support an analysis of victim selection or victim–perpetrator relationships.

Fusion and integrated analysis of multiple data resources may add value to the process, or may be required to explore a single series or pattern of crime. For example, bank and telephone records can be linked to reveal and model important patterns associated with illegal sales, distribution, or smuggling, while weather data might provide clues to patterns of crime that are affected by seasonal changes or localized weather patterns. Again, the only limitations to the data used are the creativity and insight of the analysts and the legal authority to access and use the information.

Public safety officials in many areas now are recognizing the value of regional analysis of crime trends and patterns.²² By linking data that span jurisdictional boundaries, individual localities can gain an understanding of regional trends and patterns that is not possible with locality-specific data.²³ Regional fusion centers also may represent a unique path for the acquisition of more sophisticated analytical software if the expense is distributed over a region. While the

cost of some powerful data mining tools might exceed a local budget, the cost could be distributed across localities through the establishment of a regional fusion center or coordinated analytical effort.

Finally, linking regional data resources also can be used to increase the frequency of rare events and support effective analysis. For example, some terrorist groups have shown a preference for multiple, simultaneous, yet geographically distinct attacks. While incidents of hostile surveillance are extremely rare, the ability to combine data across similar or otherwise linked locations provides a unique opportunity to more fully characterize and model a larger pattern of behavior.

4.8.3 Operationally Relevant Preprocessing

Life rarely presents the analyst with operationally relevant and actionable data. Therefore, some preprocessing of the data is required. As mentioned earlier, data preprocessing and preparation generally account for approximately 80% of the data mining process. This phase of the data mining process assumes even greater importance in public safety and security analysis, given the limitations frequently associated with public-safety-related data as well as the need for operationally relevant analytical products. Moreover, an additional limitation encountered in applied public safety and security analysis is the fact that not all data resources and variables are available when they are needed. Therefore, to address these issues, we divide the preprocessing step into operationally relevant recoding and variable selection.

4.8.3.1 Recoding

The recoding phase of data preparation includes both transformation and cleaning. Perhaps the most important function in this step is the creation of a data inventory. This data inventory helps the analysts identify what they know as well as what they do not know or what might be missing. The data organization and management function can be extremely powerful, particularly in analytical tasks supporting the investigative process. In the behavioral analysis of violent crime or cold case investigation, one of the first tasks conducted during the preliminary case review and evaluation is to organize the evidence and identify any gaps, inconsistencies, or missing information. In some cases, this review and organization of the case materials is sufficient to solve the crime by revealing information or clues that had been masked by disorganization. Sometimes, just identifying the fact that there is a missing piece in the investigative puzzle can provide new insight, which further underscores the importance of this relatively unglamorous task.

Similar to the CRISP-DM model and other analytical strategies, the data inventory should include a listing of the various data elements and any attributes that might be important to subsequent steps in the process (e.g., categorical

versus continuous data). Also important is the identification of missing data, as well as what the missing data actually mean. For example, do blank fields in a report indicate a negative response or the absence of a particular feature or element, or do they indicate a failure to ask a question or gather the relevant information? The true meaning of missing data can have significant implications not only for the analysis but also for the interpretation of the results. Therefore, decisions about missing data and the interpretation of any subsequent analyses and derived results should be made with considerable caution. That being said, data and information available for applied public safety and security analysis almost always arrive with at least some missing data. This is an occupational hazard that requires subject matter expertise and knowledge of what effect it will have on operational decisions.

Additional data quality issues also are evaluated at this stage. Some quality issues can be addressed, while others cannot. One particularly challenging issue includes the duplication of records, which is addressed in greater detail in Chapter 5. Moreover, the incident-based reporting rules now create a situation in which multiple crimes, victims, and/or suspects can be included in a single crime incident. While this increases the richness of crime data, it also significantly increases the complexity of the data and associated analytical requirements. Crimes with more than one victim and/or perpetrator can be counted more than once. This system makes it difficult to count crimes, and it affects the analyst's ability to analyze crime. Again, it frequently is up to the analyst to make informed decisions regarding data quality, cleaning, and the decision to include or disregard duplicate records.

Other data quality issues include the reliability and validity of the data. Although this is covered in detail in subsequent chapters, it is important to note that victims and witnesses frequently are unreliable in their reporting. Poor lighting, the passage of time, extreme fear, and other distractions can alter recall and reporting accuracy. Furthermore, some suspects, witnesses, and even victims have been known to intentionally distort the facts or to outright lie. While some of these data quality issues can be addressed through cleaning and recoding of the data, many will remain unresolved, contributing to a level of uncertainty and necessary caution regarding the interpretations. Therefore, this step includes any necessary cleaning and recoding as well as the selection of training and test data. Ideally, the training and test samples will be constructed using random assignment methodologies. Given the extremely low frequency of some patterns of criminal behavior, however, alternate sampling methods may be required to ensure adequate representation of data in each sample.

Most of the data and information available for public safety and security analysis require some level of recoding. Whether it involves categorizing crimes by motive or creating new variables based on MO or some other behavioral feature, recoding the data in an operationally relevant manner is essential to effective

analysis, as well as to the creation of meaningful analytical output that will have direct value in the applied setting. In response to the unique importance of time, space, and the nature of the incident or threat to most public safety and security analytical tasks and associated operational decisions, our research group at RTI International has developed the Trinity Sight™ analytical framework, which is described in greater detail in Chapter 6.

Finally, it is during this phase that the analysts begin to explore and probe the data. It is during this very important process that the analysts gain familiarity with the data, particularly the idiosyncrasies or limitations that will affect their interpretation of the findings. Therefore, the data understanding phase can be truly creative as analysts begin to identify and reveal interesting patterns or trends, which might have an impact on the analytical strategy, or even refine or change the original question.

4.8.3.2 Variable Selection

This step includes an assessment of the data resources available, based on the inventory created as well as any existing constraints on the process and assumptions made. Again, the applied setting puts a considerable number of constraints on the ability to identify and access data in a timely fashion and translate analytical products back into the operational environment. Consequently, the selection of the variables that will be used in subsequent modeling steps is extremely important in applied data mining and predictive analytics, and requires significant domain knowledge. Factors that should be considered in the selection of variables include not only the operational value of the variables selected but also their availability.

4.8.3.3 Operational Value

Many relationships identified or models created are interesting, but have no value in the applied setting because they are not operationally actionable. For example, we found that the use of a sawed-off shotgun was related to an increased likelihood that a victim would be assaulted during an armed robbery. A very interesting finding, but one with limited value to the overall objective of the analytical process, was to develop information-based patrol deployment strategies. Social scientists might examine the relationship between weapon selection and the propensity for violence, but it is very difficult to proactively deploy resources for sawed-off shotguns, significantly limiting the operational value of this finding. Therefore, significant domain expertise is required in the variable selection process to ensure that the variables selected will support the creation of operationally actionable models and output. Like the tree that falls in the woods with nobody there to hear it, no matter how interesting some analytical output might be to the analyst, it has little to no value if colleagues in the field cannot use it.

On the other hand, a related finding that the amount of money taken during an armed robbery was associated with an increased likelihood of assault initially appeared to be similarly limited in its value in deployment decisions, yet additional review of the findings suggested otherwise. Discussion with the operational personnel revealed that two specific victim populations were noteworthy for the amount of cash that they carried and their risk for robbery-related assault: drug dealers and illegal immigrants. It is not unusual for street-level drug dealers and other players in illegal drug markets to carry large amounts of cash. Moreover, violence frequently is used to enforce and regulate behavior in this setting,²⁴ so it is not surprising to find an increased likelihood of assault associated with drug-related robberies. Illegal immigrants frequently carry large amounts of cash because their immigration status limits their access to traditional financial institutions. In many cases, they are targeted by robbers specifically for this reason and are assaulted when they resist efforts to steal their money. This issue underscores the importance of domain expertise and a close working relationship with the ultimate recipients of the analysis.

4.8.3.4 Availability and Timeliness

One of the biggest challenges in translating the data mining process to the applied setting of public safety and security has been creating models with operational value and relevance. Elegant, very precise models can be created in the academic setting when accurate and reliable data are readily available and the outcomes are known. In the applied setting, however, suspects lie; incident reports frequently are incomplete; victims can be confused; witnesses are less than forthcoming; and information is limited, unreliable, or otherwise unavailable when it is needed. All of these limit the availability of and timely access to information, not to mention its reliability and validity. Ultimately, these factors can restrict the analytical pace, process, and interpretation, as well as the overall value of the results. Therefore, to increase the likelihood for success, a good understanding of what data are available and when they are available, including how the results will fit into the investigative pace or affect the tempo, how the analytical products will be used, and any other key assumptions or constraints are important to structuring the analysis.

4.8.4 Identification, Characterization, and Modeling

During the identification, characterization, and modeling phase of the project, specific statistical algorithms are selected and applied to the data in an effort to identify, characterize, and model the data of interest. Although the unique aspects of the data collected will guide selection of the specific modeling algorithms, the statistical algorithms used in data mining can be categorized into two general groups: unsupervised learning or clustering techniques, and rule induction models or decision trees. Unsupervised learning or clustering techniques group the data into sets based on similar attributes or features. These

techniques also can be used to detect data that are anomalies or significantly different from the rest of the sample.

Rule induction models capitalize on the fact that criminal behavior can be relatively predictable or homogeneous, particularly regarding common or successful MO features. Specific attributes or behavioral patterns can be characterized and modeled using rule induction models, which resemble decision trees. These models can be based on empirically determined clusters identified using the unsupervised learning techniques or those predetermined by the analyst. Rule induction models can be used to characterize and model known patterns of behavior. These models then can be applied to new data in an effort to quickly identify previously observed, known patterns and categorize unknown behavior.

Although it can be helpful to categorize the specific modeling tools into two groups, they are not mutually exclusive. Unsupervised learning and decision tree models can be used in sequence or in successive analytical iterations on the same data resources to identify, characterize, and model unique patterns, clusters, attributes, or events. For example, in some situations it is enough to know that “something” is in the data, whether it is unusual events, or trends and patterns. In other situations, identification of a case, pattern, or trend of interest represents only the first step in the analytical process. Subsequent analytical steps and processes then will be used to further characterize and/or model the data or event of interest, so it is entirely possible to use unsupervised learning approaches to initially explore and characterize the data, followed by rule induction models or decision trees to further characterize and model these preliminary findings. Again, ensemble methods can enable the analyst to strategically combine or “bundle” models in order to improve accuracy and performance.²⁵ Overall, though, the available data and resources, as well as the operational requirements, analytical tradecraft and preferences, and domain expertise are involved in the modeling approach selected.

Additional considerations in model selection and creation include the ability to balance accuracy and comprehensibility. Some extremely powerful models, although very accurate, can be relatively opaque and difficult to interpret and thus validate and use operationally. On the other hand, models that generate output that can be understood and validated may compromise overall accuracy in order to achieve this. Therefore, an understanding of the nature, costs, and consequences of model errors is important to the selection of the specific approach, particularly as relates to the operational requirements and constraints.

4.8.5 Public Safety and Security-Specific Evaluation

During the evaluation phase of the process, the models created are reviewed to determine whether they answered the question or challenge identified at

the beginning of the process. It is also during this step that the models are evaluated to determine whether the analytical output meets the needs of the end users and is actionable in the applied setting. Some modeling methods are extremely complex and can only be deployed as automated scoring algorithms, while results generated by other models can be interpreted readily and are directly actionable. Of particular importance to data mining in the applied public safety and security setting is the ability to translate the analytical output to the field in support of operational tactics and strategy. Overly complex models, while accurate and reliable, can be somewhat limited if they are too difficult to interpret. Therefore, analysts should work closely with the end users during the data evaluation phase of the process to ensure that this particular goal is achieved.

Included in the evaluation phase of the process is a review of the overall accuracy of the model, as well as the type and nature of errors. Predicting low-frequency events like crime can be particularly challenging, and overall accuracy of the models created can be somewhat misleading with these low-frequency events. For example, a model would be correct 97% of the time if it always predicted “no” for an event with an expected frequency of 3%. Clearly, overall accuracy would be an unacceptable measure for the predictive value of this type of model. In these cases, the nature and direction of errors can provide a better estimate of the overall value of the model. By adjusting the “costs” associated with false positives or misses, the model can be refined to better predict low-frequency events. These costs can be balanced to create a model that accurately identifies cases of interest while limiting the number of false alarms. Unfortunately, the analysts are often in the position of attempting to model infrequent or rare events, events that can change rapidly. Therefore, specific attention to errors and the nature of errors is required. In some situations, anything that brings decision accuracy above chance is adequate. In other situations, however, errors in analysis or interpretation can result in misallocated resources, wasted time, and even can cost lives. As always, significant domain expertise and extensive knowledge of the operational objectives, data resources, procedures, and goals are essential in creating predictive models that are operationally reasonable. It is essential that the analysts work closely with the operational end users during this phase of the process to ensure that the models are valuable and actionable in the applied setting and that any necessary compromises in accuracy are acceptable.

Finally, it is important to evaluate the models created and relationships identified to ensure that they make sense. The importance of domain expertise and tacit knowledge in the interpretation and evaluation of analytical results cannot be overstated. On the other hand, it does not necessarily indicate a failure of the process if the analysis raises as many or more questions than it answers. The data mining process includes confirmation of known or suspected relationships as

well as surprise and discovery. The knowledge discovery associated with unanticipated outcomes can greatly increase our understanding of crime and criminal behavior, and result in novel approaches to enhancing public safety.

4.8.6 Operationally Actionable Output

The ability to translate complex analytical output into a format that can be directly used in the operational setting to support prevention and enforcement strategies is critically important to effective data mining in the applied public safety and security setting. Sophisticated analytical tools, including data mining software applications, have been commercially available for several years, and complex analytical strategies are commonplace in academic criminal justice research. It has been relatively recently, however, that these tools and approaches have started to be used in the applied public safety and security arena, in large part because the analytical output generated by sophisticated algorithms and tools have had little direct relevance to the applied setting. As discussed earlier, overly complex models, while accurate and reliable, can be somewhat limited if they are too difficult to interpret to be useful to the end users. On the other hand, innovative approaches to conveying complex analytical output in a format that is not only readily interpreted and understood by the end user but also leverages their tacit knowledge and domain expertise can add significant value to the analytical process and outcomes. Therefore, the critical importance of “operationally actionable” analysis and output will be referred to repeatedly throughout this text and addressed in more detail in Chapter 8.

4.8.7 Additional Considerations

Use of predictive analytics in the operational public safety and national security domain is neither straightforward nor easy. Challenges relate to the requirement for operationally relevant and actionable analytic products, as well as the legal and ethical access to data. Moreover, as recent examples from the commercial sector underscore,²⁶ just because something can be done does not mean that it should and that caution should be exercised, particularly as relates to context, messaging, and the use of derived analytic products.

4.8.7.1 Privacy

Privacy will be discussed in greater detail in Chapter 16, but recent revelations regarding both the United States government and commercial access to and use of data, including data mining and predictive analytics, have increased public debate regarding the use of advanced analytics in a number of domains. High profile data breaches have further underscored the potential consequences of improper data handling, access and misuse.²⁷ The issue is complex and frequently seems to be situation-dependent, even transactional in nature as individuals who may express strong concern regarding access to and use of their personally identifiable information (PII), including location-specific data, may

be willing if not eager to provide this same information in exchange for benefits like personalized discounts, coupons, and insight regarding amenities in their immediate environment.²⁸

4.8.7.2 Security

Related to privacy concerns, the increased use of derived products, including location data, has raised concerns throughout the community regarding potential security issues. In particular, the use of analytics for conflict mapping and crisis response has increased awareness regarding the potential consequences, especially for vulnerable populations who are the victims of natural or man-made disasters, including conflict.²⁹ Following the adage, “first do no harm,” data sensitivity will be reinforced as a consideration during relevant steps in the analytic process, including preprocessing, evaluation, and output steps.

4.8.7.3 Other Hazards

Developments in technology have made increasingly sophisticated capabilities less expensive and more user friendly; however, this ease of use and increased access should not be confused with “simple” and “easy.” Even “machine learning” can be prone to bias. As Heuer notes, cognitive bias can be particularly elusive to identify and challenging to address, but it can “affect the evaluation of evidence, perception of cause and effect, estimation of probabilities, and retrospective evaluation of intelligence.”³⁰ Heuer specifically reinforces the idea that correlation or covariance does not imply causation, which is particularly relevant to predictive analytics. The leap from correlated data to causal relationships can be tempting, particularly when machine learning or “artificial intelligence” capabilities are employed. Unfortunately, these errors in logic may lead to faulty inference and wrong conclusions, and are not uncommon to find. Therefore, common hazards and pitfalls will be covered throughout the text.

4.8.8 Summary

The Actionable Mining and Predictive Analysis model just presented differs from the first two models in its specificity to the public safety and security domains, as well as in the inclusion of operationally relevant preprocessing and output. Specifically, this model includes operationally relevant recoding and variable selection, public safety and security-specific model evaluation, and an emphasis on operationally actionable output. [Table 4.3](#) compares three of the analytical process models covered in this chapter: CRISP-DM, CIA Intelligence Process, and Actionable Mining and Predictive Analysis.

Data mining is as much analytical process and tradecraft as it is specific mathematical algorithms and statistics. This process of data exploration and the associated surprise and discovery, which are the hallmarks of data mining, can be as exciting as they are challenging; analysts are rewarded with a progressively

Table 4.3 Comparison of the CRISP-DM, CIA Intelligence Process and Actionable Mining, and Predictive Analysis Analytical Process Models

	CRISP-DM	CIA Intelligence Process	Actionable Mining and Predictive Analysis
Business understanding	Y	Y	Y
Data understanding	Y	Y	Y
Data preparation	Y	Y	Y
Modeling	Y		Y
Evaluation	Y	Y	Y
Deployment	Y	Y	Y
Needs	Y	Y	Y
Collection		Y	Y
Processing and exploitation		Y	Y
Analysis and production		Y	Y
Dissemination		Y	Y
Feedback		Y	Y
Question or challenge	Y	Y	Y
Data collection and fusion		Y	Y
Operationally relevant recoding			Y
Variable selection			Y
Identification, characterization, and modeling	Y		Y
Public-safety-specific evaluation			Y
Operationally actionable output			Y

evolving list of questions to be answered as the data reveal additional insights and relationships.

The challenge facing public safety and security analysts lies in being able to craft an analytical process model that can accommodate differences in collection methodologies and functional domains, yet also transcend these differences in support of global applicability. The limitation in that approach, particularly in such a functionally diverse field as applied public safety and security analysis, is that different questions, sources, tactics, and strategies require different analytical approaches and sometimes significantly different analytical processes. Therefore, the Actionable Mining and Analysis Process Model can be thought of as being similar to a building code, which outlines the specific

elements that should be addressed and offers a suggested sequence of steps that should be covered within the larger process. Following this analogy, the specific protocols for each unique analytical task are the blueprints that operationalize these broader elements and concepts for specific public safety, intelligence, and security analyses.

In keeping with this concept, the next five chapters address specific elements or phases in the Actionable Mining and Analysis Process Model. Following that, specific public safety and security questions, topics, and challenges are addressed in greater detail. Specific analytical “blueprints” are provided in each chapter, outlining a specific application of data mining in the applied public safety setting. While it is unlikely that these recommended analytical strategies will fit perfectly with every situation in any agency or department, they should represent a reasonable approximation or template that analysts can apply to their particular situation. Hopefully, as the use of data mining and predictive analytics becomes more widespread in the applied setting and a critical mass of end users is attained, the availability of these blueprints will increase concomitantly.

Bibliography

- 1 United States, Office of the Director of National Intelligence (US ODNI). National intelligence: a consumer’s guide. National Technical Information Service, Springfield, VA. https://ia700506.us.archive.org/7/items/nationalintelligenceconsumersguide/IC_Consumers_Guide_2009.pdf; 2009.
- 2 Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0: Step-by-step data mining guide. <http://www.the-modeling-agency.com/crisp-dm.pdf>; 2000.
- 3 SAS Enterprise Miner. SEMMA. <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>.
- 4 Douglas J, Burgess AW, Burgess AG, Ressler RK. Crime classification manual: a standard system for investigating and classifying violent crime. Hoboken, NJ: Wiley; 2013.
- 5 Ibid.
- 6 Lord WD, Boudreaux MC, Lanning KV. Investigation of potential child abduction cases: a developmental perspective. FBI Law Enforc Bull; April, 2001.
- 7 The intelligence community analytical processes and strategies are historically rich and very interesting. The following overview is not meant to be inclusive. Other process models include the FBI Intelligence Cycle (<http://www.fbi.gov/about-us/intelligence/intelligence-cycle>), which is very similar to the CIA model. For additional reading in this area, see: Lowenthal MM. Intelligence: from secrets to policy. Washington, D.C: CQ Press; 2012.
- 8 Lowenthal MM. 2012.
- 9 US ODNI. 2009; Lowenthal MM. 2012.
- 10 US ODNI. 2009; Lowenthal MM 2012. This list of different “INTs” has been updated from the first edition of this text and continues to evolve as different collection capabilities and sources are developed. For example, emerging “INTs” include Open Source Intelligence (OSINT, Lowenthal MM.; 2012), and Social Media Intelligence (SOCMINT; Omand D, Bartlett J, Miller C. A balance between security and privacy online must be struck...#Intelligence. Demos, ISBN 978-1-909037-08-3, <http://www.demos.co.uk/publications/intelligence>; 2012). Stretching this concept to the ridiculous, the concept of “LOVEINT,” which includes the use of the other INTs to conduct searches on current and/or former targets of romantic interest, was surfaced

- by the National Security Agency (NSA) Office of the Inspector General during a review of NSA collection and use of data after the Edward Snowden leaks (<http://www.grassley.senate.gov/judiciary/upload/NSA-Surveillance-09-11-13-response-from-IG-to-intentional-misuse-of-NSA-authority.pdf>; Moyer E. NSA offers details on 'LOVEINT' (that's spying on lovers, exes). CNET. http://news.cnet.com/8301-13578_3-57605051-38/nsa-offers-details-on-loveint-thats-spying-on-lovers-exes/; 2013 [accessed 27.09.2013].)
- 11 Chapman et al. 2000.
 - 12 Ibid.
 - 13 Seni G, Elder J. Ensemble methods in data mining: Improving accuracy through combining predictions. (Synthesis Lectures on Data Mining and Knowledge Discovery) Grossman R, editor. Morgan & Claypool; 2010.
 - 14 SAS Enterprise Miner, SEMMA. http://faculty.smu.edu/tfomby/eco5385/data/SPSS/SAS%20_%20SEMMA.pdf.
 - 15 For review, see, Piatetsky-Shapiro G. CRISP-DM: a proposed global standard for data mining. *DS Star* 1999; 3: 15; <http://www.tgc.com/dsstar/99/0413/100687.html>; Dnuggets K. What main methodology are you using for data mining?; July, 2002. <http://www.kdnuggets.com/polls/2002/methodology.htm>; Shearer C. The CRISP-DM model: the new blueprint for data mining. *J Data Warehousing* 2002; 5: 13–22.
 - 16 Chapman et al. 1999.
 - 17 Rittel H, Webber M. Dilemmas in a general theory of planning. *Policy Sciences* 1973;4: 155–169.
 - 18 Nisbet R, Elder J, Miner G. Handbook of statistical analysis & data mining applications. Boston: Academic Press; 2009.
 - 19 Buffett HG. Forty chances: finding hope in a hungry world. Simon & Schuster, New York; 2013.
 - 20 Rittel H, Webber M. 1973, p. 163.
 - 21 Rittel H, Webber M. 1973, p. 166.
 - 22 McCue C, Miller L, Lambert S. The Northern Virginia military shooting series: operational validation of geospatial predictive analytics. *Police Chief* 2013; February, http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=2871&issue_id=22013.
 - 23 Faggiani D, McLaughlin CR. A discussion on the use of NIBRS data for tactical crime analysis. *JOQC* 1999; 15: 181–191.
 - 24 Goldstein PJ. The drugs/violence nexus: a tripartite conceptual framework. *J Drug Issues* 1985;15:493–506.
 - 25 Seni G, Elder J. 2010.
 - 26 Duhigg C. How companies learn your secrets. *The New York Times*. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0; 2012 [accessed 16.02.2012].
 - 27 Rosenblum P. The Target data breach is becoming a nightmare. *Forbes*. <http://www.forbes.com/sites/paularosenblum/2014/01/17/the-target-data-breach-is-becoming-a-nightmare/>; 2014 [accessed 17.01.2014].
 - 28 Brustein J. If your phone knows which aisle you're in, will it have deals on groceries? *Bloomberg Business Week*. <http://www.businessweek.com/articles/2014-01-06/apples-ibeacon-helps-marketer-beam-ads-to-grocery-shoppers-phones>; 2014 [accessed 06.01.2014].
 - 29 McCue C. Crisis mappers webinar series: analytics for conflict mapping and crisis response. <http://www.youtube.com/watch?v=U9OgZ30AXUg&feature=youtu.be>; 2013.
 - 30 Heuer RJ. Psychology of intelligence analysis. Center for the Study of Intelligence, Central Intelligence Agency. <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/psychology-of-intelligence-analysis/PsychofIntelNew.pdf>; 1999, p. 112.

Data

“It is a capital mistake to theorize before one has data.”

Arthur Conan Doyle

One of the most important tasks associated with analyzing crime and intelligence information is to know your data. Data and information are the currency within the analytical community; however, very little if any of the information that falls under the purview of crime and intelligence analysis is ever collected with that purpose in mind. Unlike the business and academic communities, in which data sets are designed and constructed with thought given to the ultimate analysis at almost every step of the way, some of the data that falls in the analysts’ laptops are so ugly that only a mother could love them. Without data and information, however, we analysts would perish. Data truly are the lifeblood of the analytical community, and some of the most unruly data sets can prove to be the most rewarding once they have been tamed and analyzed.

WHERE DO ANALYSTS COME FROM?

“Imagination is more important than knowledge ...”

Albert Einstein

“A senior intelligence official used to ask his subordinates two questions about new analysts they wished to hire: ‘Do they think interesting thoughts? Do they write well?’ This official believed that, with these two talents in hand, all else would follow with training and experience.”

Lowenthal, p. 79¹

In my experience, statistical degrees might make a good scientist, but a good analyst needs much more. Over time, I have almost always found it is easier to teach someone with an understanding of crime and criminals, and a need to know “why” how to analyze crime than it is to give domain knowledge to a statistician. Other data scientists have shared similar observations, citing a “spark of keen intellect and curiosity” as key attributes in great analyst and noting that PhD-level

scientists tend to “study things to death,” while someone with a Master’s degree “shows interest without obsession.”² This trait in particular is absolutely necessary in the rapidly changing, fluid environment that characterizes operational public safety and security.

As indicated in the preceding quotes, the ability to “think interesting thoughts,” or think creatively, also seems to be a prerequisite. It often appears that the bad guys are always working on new ways to commit crimes and escape detection, and examples from natural disaster response and recovery efforts like Hurricane Katrina underscore the fact that it frequently is difficult to imagine just how many things might go wrong, how quickly a situation can deteriorate, and just how bad things can get.³ Getting in front of these challenge, natural or manmade, involves going above and beyond, often to the outer limits of our understanding of what can be expected, or even possible, in an effort to stay ahead in the game. Good analysts often have a dark side that they not only are in touch with but also are relatively comfortable tapping. To anticipate the “what if,” “how,” and “why,” we can find ourselves considering the proverbial “perfect crime,” placing ourselves in a criminal’s mind in an effort to unlock the secrets of the crimes that he has committed and to anticipate what might lie ahead, what it is that he would really like to do if given the chance.

One outstanding example of this ability to anticipate the unthinkable is Rick Rescorla, the director of security at Morgan Stanley Dean Witter at the World Trade Center.⁴ While there is no evidence that he used formal analytic tradecraft or enabling technology, after the first attack on the Trade Center in 1993 Rescorla believed that the terrorists would not be satisfied with this “failure”; that they would continue to plan and make another attempt on the towers. With this in mind, Rescorla conducted regular evacuation drills. From the senior executives to administrative staff and even visitors, all were expected to participate in the drills. On September 11, 2001, Morgan Stanley Dean Witter personnel knew exactly what was required to evacuate the Tower and had done so on numerous occasions. As a result, all but 13 employees, including Rescorla made it safely out of the building; 2687 survived. While data mining and predictive analytics may not enable us to prevent, thwart, or otherwise mitigate every incident, it can support information-based response and consequence management. In this particular case, Rescorla’s ability to anticipate the unthinkable was directly attributed to the high survival rate for his organization.

Data mining has been referred to as “sense making,” which often is how the request for analysis is phrased. It is not unusual to find some random lists of telephone numbers, a couple of delivery dates, and a bank statement delivered along with a plea for some clue as to the “big picture” or what it all means: “Please help me make some sense of this!” There are a few tips for understanding and managing data and information that will be addressed in this chapter, but it would be impossible to anticipate every data source that is likely to surface. It is better to develop the ability to work with and through data in an effort to reveal possible underlying trends and patterns. This is a craft in some ways. In my experience, there is a degree of creativity in the type of pattern recognition that is associated with a good analyst: the ability to see the meaning hidden among the general disorder in the information. Data mining truly can be described as a discovery process. It is a way to illuminate the order that often is hidden to all, but the skilled eye.

5.1 GETTING STARTED

When I worked at a lab as a scientist, one of the most important lessons that I learned was to know your subjects. The same is true for crime and intelligence analysts. As it is not advisable to experience crime firsthand, either as a victim or a perpetrator, there are other ways to know and understand the data and information. This is particularly true for information that arrives with some regularity, like offense reports. Subject knowledge has been addressed in greater detail in Chapter 2; in brief, it can be extremely useful for the analyst to develop some familiarity with the operational environment. Whether through routine ride-alongs, attendance at roll call, field training, observing suspect interviews, or as an actual embed within an operational unit, the more that the analyst can understand about where the data and information came from, as well as the operational requirements or constraints, the better the subsequent analysis.

5.2 TYPES OF DATA

Several years ago, I responded to a homicide scene during the very early hours of the morning. Everyone was pretty tired, but as I walked into the house to view the scene, a member of the command staff who was taking a graduate course in criminal justice statistics stopped me and asked for a quick tutorial on the difference between continuous and nominal data. I have been asked some really strange questions at crime scenes, but this one was memorable by its sheer absurdity. What did distinctions between different types of data have to do with death investigation? In some ways nothing, but in other ways it has everything to do with how data should be analyzed and used to enhance the investigative process.

Leave it to researchers to figure out a way to describe and characterize even data, creating subcategories and types for different kinds of information. There is a reason for this, other than the pain and agony that most students experience when trying to figure out what it all means during a class in statistics and probability (also known as “sadistics and impossibilities”). Data and information are categorized and grouped based on certain mathematical attributes of the information. This can be important, because different types of analytical approaches require certain properties of the data being used. As a result, it is important to have at least a basic understanding of the different types of data and information that might be encountered, and how this might guide selection of a particular analytical approach or tool.

5.3 DATA⁵

Continuous variables can take on an unlimited number of values between the lowest and highest points of measurement. Continuous variables include such

things as speed and distance. Continuous data are very desirable in inferential statistics; however, they tend to be less useful in data mining and are frequently recoded into discrete data or sets, which are described next.

Discrete data are associated with a limited number of possible values. Gender and rank are examples of discrete variables because there are a limited number of mutually exclusive options. Binary data are a type of discrete data that encompass information that is confined to two possible options (e.g., male or female; yes or no). Discrete and binary data also are called sets and flag data, respectively.

Understanding the different types of data and their definitions is important because some types of analyses have been designed for particular types of data and may be inappropriate for another type of information. The good news is that the types of information most prevalent in law enforcement and intelligence, sets and flag data, tend to be the most desirable for data mining. With traditional, inferential statistics methodologies, on the other hand, discrete variables are disadvantageous because statistical power is compromised with this type of categorical data. In other words, a bigger difference between the groups of interest is needed to achieve a statistically significant result.

We also can speak of data in terms of how they are measured. Ratio scales are numeric and are associated with a true zero – meaning that nothing can be measured. For example, weight is a ratio scale. A weight of zero corresponds to the absence of any weight. With an interval scale, measurements between points have meaning, although there is no true zero. For example, although there is no true zero associated with the Fahrenheit temperature scale, the difference between 110 and 120°F is the same as the difference between 180 and 190°, which is 10°. Ordinal scales imply some ranking in the information. Although the data might not correspond to actual numeric figures, there is some implied ranking. Sergeant, lieutenant, major, and colonel represents an ordinal scale. Lieutenant is ranked higher than sergeant, and major is ranked higher than lieutenant. Although they do not correspond directly to any type of numeric values, it is understood that there is a rank ordering of these categories. Finally, nominal scales really are not true scales because they are not associated with any sort of measurable dimension or ranking; the particular designations, even if numeric, do not correspond to quantifiable features. An example of this type of data is any type of categorical data, such as vehicle make or numeric patrol unit designations.

Finally, unformatted or text data truly are unique. Until recently, it was very difficult to analyze this type of information because the analytical tools necessary were extremely sophisticated and not generally available. Frequently, text data were recoded and categorized into some type of discrete variables. Recent advancements in computational techniques, however, have opened the door to

analyzing these data in their native form. By using techniques such as natural language processing, syntax and language can be analyzed intact, a process that extends well beyond crude keyword searches.

5.3.1 Big Data

While the term “Big Data” has become increasingly popular, crime and intelligence analysts have been working with massive datasets for year. Citizen Calls for Service (CFS), SIGINT, and geospatial data sources frequently fall into the “Big Data” realm, particularly when they are linked to other sources in relational databases or combined to form derived products. Perhaps the one noteworthy benefit associated with the recent emphasis on big data, though is that the data management and associated analytic capabilities have advanced to the point where these big data resources are increasingly accessible to the analyst. With the development of cloud computing, in-memory appliances, and other capabilities, the analyst can now effectively explore and exploit these resources in a meaningful way in support of real time or near real time analysis and actionable insight.

Big data has been described by the three “Vs” – volume, velocity, and variety – each of which brings unique benefits and challenges to the analyst.⁶

5.3.1.1 Volume

Volume is the most obvious of the “Vs” and refers to the fact that we increasingly are able to collect a lot of data. The availability of large amounts of data generally gives us more observations of interest, which concomitantly enables us to create better models. In addition, as is addressed frequently throughout this text, the crime and intelligence analyst frequently is looking for and/or attempting to model extremely infrequent or rare events. Similar to the high-energy physicists with rooms of sensors collecting data over extended periods of time in an effort to capture that one, single rare event, massive volumes of data increase the likelihood that we will have a sufficient sample to capture that very rare event. As crime and intelligence analysts, if we are looking for the needle in the haystack, then massively increasing the size of the haystack concomitantly increase the likelihood that we will have a sample of sufficient size to find the needle. In some cases, the haystack might be big enough to hold more than one or multiple needles, which provides additional opportunities for characterization and modeling. Again, the ability to collect and aggregate large amounts of data enables us to increase the likelihood of collecting infrequent or rare events.

5.3.1.2 Velocity

The second attribute of big data with relevance to the analyst is velocity. Again, our ability to collect data rapidly has existed for a considerable period of time.

The important change recently has been the ability to analyze data in real time or near real time. One example is fraud detection. Many models rely currently on a “pay and chase” approach, which involves paying the claim, collecting data, and then analyzing them for indicators of improper payments or fraud.⁷ Related to the “volume” benefits associated with big data, the “pay and chase” model enables the aggregation of sufficient events to effectively identify and analyze infrequent or rare transactions; however, it is horribly inefficient. Because the resources required to recover the fraudulent payments tend to be significant, decision rules or thresholds generally are established to make sure that the potential recovery meets or exceeds the investigative resources, time, and effort, as well as any related prosecutorial resources supporting identification of the fraud and related recovery. The alternative, however, is unacceptable for most consumers. Imagine for example dropping off a prescription at the pharmacy or submitting an insurance claim and being advised that you could come back in three weeks after they have analyzed your request to make sure that you are not engaging in prescription drug fraud, identity theft, or insurance fraud. This would be completely unacceptable. The ability to deploy scoring algorithms in real time or near real time against transactional data, therefore, enables the identification of fraudulent or otherwise suspicious transactions in support of timely detection and prevention of fraud.

Another benefit to effectively leveraging big data velocity is real time management of resource deployment. The law enforcement community is familiar with the scenario where a traffic wreck, homicide, or unpopular jury verdict turns into an angry mob, which may escalate into a riot or even widespread civil unrest. While we may not be able to prevent these situations from happening, being able to identify them early and react quickly to fluid and rapidly emerging events enables us to better anticipate and respond in support of timely thwarting, mitigation, and consequence management. Like the “just-in-time” supply chain analytics used to rapidly adjust delivery in response to rapid changes in consumer demand, the ability to not only collect real time data but analyze it in real time or near real time provides new options for getting within the decision cycle of our opponents and influencing the trajectory of events and related outcomes.

5.3.1.3 Variety

The third “V” of big data is variety. Again, crime and intelligence analysts have relied on a wide array of data resources; however, advances in knowledge management and discovery tools have enabled us to effectively fuse and integrate multiple sources in support of more complex analysis and the creation of novel derived products in support of meaningful insight. One such example includes the ability to collect location data associated with entities, transactions, and other sources of interest. Related to this, geospatial predictive analysis, which

will be covered in Chapter 7 and examples throughout the text, relies on hundreds or even thousands of geospatial variables or factor layers to effectively model the complexity of the environment, including place preferences that criminals and other adversaries demonstrate as they look for victims or targets, and enabling environments in which to perpetrate their acts. These factor layers may include both physical and human geography, as well as the interaction between the two.

“Successes like these take a huge effort from some really smart people with access to the right tools and processes to create actionable insights from all of that big data. Otherwise, it’s all just a bunch of ‘stuff’.”

Gene De Libero⁸

Big data and related advances in knowledge management and discovery can enable some truly powerful analysis and related insight. However, without domain expertise and a solid understanding of the context, operational requirements, and constraints, big data becomes increasingly limited in value and even a possible hazard if it engenders overconfidence in the results or spurious findings.⁹

5.4 TYPES OF DATA RESOURCES

WORKING WITH THE OPERATORS TO GET GOOD DATA

The business community provides great insight regarding successful approaches to accurate, complete, and reliable data collection. In some cases, information gathering is anything but obvious, often hidden behind the guise of some sort of discount or incentive. For example, many supermarkets have adopted discount cards in an effort to provide incentives for gathering information. Rather than using coupons, the customer need only provide the store discount card to obtain reduced prices on various items. It is quick and easy, and the customer saves money. In exchange, the store gains very detailed purchasing information. Moreover, if a registration form is required to obtain the discount card, the store has access to additional shopper information that can be linked to purchasing habits. For example, stores might request specific address information so that they can mail flyers regarding sale items and specials to customers. Additional financial and demographic information might be required for check cashing privileges, which can be appended to other sources including demographic information associated with the address. This additional information provides significant value when added to the actual shopping information. Direct mailing campaigns can be targeted to select groups of shoppers or geographic areas, while common shopping patterns can be used to strategically stock shelves and encourage additional purchasing. Overall, this underscores the increasingly transactional relationship between people and their data. Whether privacy or “work,” people frequently will provide data or otherwise enable collection for some sort of benefit.

This same principle can be applied in crime and intelligence analysis data collection. One frequently voiced frustration from operational personnel is that they do not receive any benefits from

all of the paperwork that they complete. Field interview reports can take a considerable amount of time and may confer benefits only to the analyst or other investigative personnel. Analysts can greatly improve data quality and volume by engaging in some proactive work to highlight the value of accurate, reliable, and timely data collection to the specific personnel units most likely to be tasked with the majority of this work. Providing maps or other analytical products on a regular basis can strengthen the partnership between these often overworked, yet underappreciated, line staff and the analytical personnel. These folks also can be a tremendous source of knowledge and abundant domain expertise, given their direct proximity to the information source. Again, embedding analytic personnel in operational units can greatly facilitate direct interaction in support of this goal.¹⁰

It also can be important, whenever possible, to highlight any quality-of-life increases that might be associated with a particular analytical product or initiative. For example, the New Year's Eve initiative in the Richmond Police Department was associated with significant reductions in random gunfire and increased weapons seizures,¹¹ which were achieved with fewer personnel resources than originally anticipated. These results in and of themselves were impressive; however, one additional benefit was that approximately 50 members of the department originally expecting to work that evening were able to take the night off. While this might not be the most notable finding from a public safety perspective, it was pretty important to the department members who were able to take the night off and spend it with their families and friends.

Partnering with the operational personnel reaps many benefits. Having a colleague in the field can provide direct feedback regarding the reliability and validity of specific sources of information, as well as guidance regarding the direction and need for future analytical products. Some of the most interesting projects that I have been involved with started out with the following question from someone working in the field: "Hey doc, have you ever thought about this ...?" Moving closer to an integrated operator-analyst approach benefits all participants. The analyst obtains access to better information and guidance regarding actionable analytical end products, and the operational personnel can not only better understand but even play a role in guiding the analytical process. Once information has been established as a thin, fluid interface between the analytical and operational domains, the process becomes even more dynamic, operators and analysts working handcuff-in-glove to achieve information dominance and operational superiority. Data mining and the associated technologies provide the tools necessary to realize this goal.

5.4.1 Data Sources

The crime analyst is likely to encounter two types of data that have their own sets of issues and challenges: Records Management Systems (RMS), and police Calls for Service (CFS) or citizen complaint data. These data are used frequently in public safety analysis; therefore, their unique features and challenges are worth addressing.

5.4.1.1 Records Management Systems

Most departments maintain large records management systems (RMS) that contain crime incident data. In most cases, however, these "databases" were not necessarily designed to be analyzed. Rather, these databases were created and are used for case management and general crime counting. As a result,

these databases frequently have standard or “canned” queries that facilitate gathering frequently used information or reports; however, these often have limited utility for crime analysis.

The benefit of these databases is that they have a known, stable structure. Queries can be developed and reused repeatedly because the structure associated with this type of database generally does not change frequently. The disadvantages associated with using a records management system, however, can include the reliability and validity of the data, as well as the detail, type, and nature of information, and even access. Common limitations associated with generic records management systems include incomplete or inaccurate data. Unfortunately, reliability, validity, and completeness seem to be most compromised in the information required for crime analysis, particularly information relating to MO. MO characteristics generally are not included in routine crime reports or canned queries, so it is not until the analyst attempts to use it that the holes in the data are revealed. Moreover, MO information can be compromised if the particular categories of information are not anticipated and collected. For example, information relating to the type of weapon, nature of the offense, and time of day frequently are collected, but specific details regarding the behavioral nature of the crime, including the type of approach (e.g., con/ruse, blitz, etc.), verbal themes, and other behavioral characteristics used to analyze and link crimes frequently are incomplete or absent. Much of this information can be found in the narrative portion of the report. However, in most agencies, the narrative section has limited availability and utility, given the degree of complexity associated with the analysis of this type of information. Recent advantages in natural language processing and text mining offer great promise for the retrieval of this information.

Another limitation associated with large records management systems can be timely data entry. Ideally, the data and information would be entered into the database during the collection process, or immediately thereafter. Unfortunately, reports frequently are completed and then wait for review before the information is transcribed and entered. This can be particularly time-consuming if the information is collected in a location that is geographically distinct from the data entry and analysis location, or if field reports must be collected and shipped to another location for entry and subsequent analysis in another part of the world. Even processing records in another part of town can introduce a level of delay that is undesirable.

On the other hand, initial data entry can be almost simultaneous in departments with mobile data computers, which facilitate direct data entry in the field. Although data review and validation might be delayed somewhat, the basic crime incident information generally is entered in a timely fashion. In agencies where reports are completed with paper and pencil and then forwarded for

data entry, significant delays can occur. While this might not be a problem with historical reviews or long-term trend analysis and modeling, for certain types of analysis, including the behavioral analysis of violent crime and/or analysis of a rapidly escalating series, any delay is unacceptable because it can cost lives. Often in these cases, the analyst is required to create specialized databases for certain types of crime or series under active investigation.

Technology has improved to the point where direct data entry in the field can be associated with a concomitant rapid analytical response, even in the absence of a live analyst. Crime frequently occurs at inconvenient times, particularly during the evenings and weekends when most civilian analysts are off duty. The ability to rapidly integrate crime incident information into the context of historical data as well as rapidly emerging trends and patterns can provide the investigator valuable analytical support at the very beginning of an investigation when it is needed the most, rather than later when the analytical staff returns. While still at the scene, an investigator can enter the relevant information and receive a rapid analytical response, which can provide timely access to associated cases, existing leads, and investigative guidance. By adding an analytical overlay or inserting an analytical filter into remote data entry, an organization can add value to the mobile data and analytical capacity while at the same time increasing investigative efficacy.

5.4.1.2 Calls for Service

One set of data maintained by most law enforcement agencies is police dispatch, or Calls for Service (CFS) data. This information can have tremendous value when examining how police resources are deployed. These data also provide some insight into general crime trends and patterns in the community in that they reflect complaint data, or citizen-initiated police work.

Like most public safety data, however, CFS data have significant limitations that are not distributed uniformly in many cases. For example, a complaint of “man down” can mean almost anything from a sleeping vagrant to a murder. It is not unusual for the nature of the complaint to have very little in common with what actually happened.

[Figure 5.1](#) depicts a vehicle driving through a neighborhood, engaging in random gunfire – something not unusual in many high-risk communities, particularly those with active gang rivalries. In this situation, the vehicle enters the neighborhood in the 600 block of Elm Street and one of the occupants starts firing a weapon in the vicinity of 545 Elm Street. The vehicle travels north on Elm, east on Fourth, and north again on Maple Avenue, with five associated bursts of gunfire.

As can be seen on the spreadsheet in [Table 5.1](#), the calls start coming into the dispatch center almost immediately. Three separate calls come in within



FIGURE 5.1 This street map illustrates a vehicle driving through a neighborhood engaging in random gunfire.

Table 5.1 This Table was Created from the Citizen Complaints Associated with the Random Gunfire Illustrated in [Figure 5.2](#)

Date	Time	Location	Nature	Call Number
Feb. 27, 2004	2313	545 Elm St.	Shots fired	258
Feb. 27, 2004	2314	545 Elm St.	Shots fired	258
Feb. 27, 2004	2315	545 Elm St.	Shots fired	258
Feb. 27, 2004	2315	200 W 6th St.	Shots fired	259
Feb. 27, 2004	2318	150 W 4th St.	Shots fired	260
Feb. 27, 2004	2320	148 W 4th St.	Shots fired	260
Feb. 27, 2004	2321	50 W 4th St.	Shots fired	260
Feb. 27, 2004	2326	242 Maple Ave.	Shots fired	261
Feb. 27, 2004	2328	242 Juniper St.	Shots fired	262
Feb. 27, 2004	2329	105 Maple Ave.	Shots fired	263
Feb. 27, 2004	2345	545 Elm St.	Shots fired	264

minutes of the first burst of gunfire at 545 Elm Street. These are immediately recognized as being the same incident and are all assigned the same call number. A fourth call comes in from someone at the corner of Sixth and Elm, who reports having heard random gunfire in the neighborhood. Without a specific address, the caller's location is entered and the report is given a new call number, reflecting the new address. A few minutes later, three additional calls come in from Fourth Street. These calls are seen as related to each other but not to the earlier incident, so a new call number is assigned to these three complaints. As the car turns north on Maple Avenue, three additional calls come in to the dispatch center. Again, these calls are not identified as being part of the same incident and are given unique call numbers. Finally, the first caller at 545 Elm Street, frustrated at not seeing a patrol car in the neighborhood yet, calls again to report random gunfire in the neighborhood. Because this call came in later than the first call from that address, it is not linked to the earlier calls and is given another unique call number. By the time this incident ends, the dispatch center has received a total of 11 calls from citizens reporting random gunfire (Table 5.2), which are aggregated into seven distinct calls, yet all are associated with a single series of random gunfire spanning both geography and time.

Which is the more accurate measure: complaints, CFS, or unique incidents? Further comparison of the map and the call sheet highlights some of the challenges associated with analyzing CFS. For example, should the total number of complaints (11) or unique calls (7) be used for analysis? Depending on the nature of the analysis, either option might be correct. For an analysis of workload and deployment, it might be necessary only to know how many citizen-initiated complaints resulted in the dispatch of police personnel. In this case, the number of unique calls generally would suffice. But, what if the question related to community violence and how many incidents of random gunfire occurred in a community? Clearly, neither the number of complaints nor the number of calls would address that question directly. Further examination of

Table 5.2 In this Table, the 11 Citizen Complaints from a Single Series of Random Gunfire have been Aggregated into 7 Distinct Calls, Highlighting Some of the Challenges Associated with Using Citizen Complaint Data as well as Other Public-Safety-Related Information

Date	Time	Location	Nature	Call Number
Feb. 27, 2004	2313	545 Elm St.	Shots fired	258
Feb. 27, 2004	2315	200 W 6th St.	Shots fired	259
Feb. 27, 2004	2318	150 W 4th St.	Shots fired	260
Feb. 27, 2004	2326	242 Maple Ave.	Shots fired	261
Feb. 27, 2004	2328	242 Juniper St.	Shots fired	262
Feb. 27, 2004	2329	105 Maple Ave.	Shots fired	263
Feb. 27, 2004	2345	545 Elm St.	Shots fired	264

the data would be required, perhaps with the use of maps and a timeline. What happens when gunfire erupts and nobody calls? In some locations it is not unusual to receive a call of a man down, only to find a dead body surrounded by evidence of gunfire and to have received no citizen complaints of gunfire, whether due to fear or a lack of social efficacy.¹²

This issue of incident counting becomes further complicated if multiple units are dispatched to the same incident and are recorded as unique records in the database, or if a unit needs to disengage from a specific call to answer another call with greater urgency and then returns to the initial call. Depending on how the data are recorded and maintained, the latter situation might be recorded as a single unit responding twice to the same call, two unique dispatch records for a single call, or two unique calls for service. Which measure is correct really depends on the nature of the question. The most important point, though, is that the analyst clearly understands the limitations of the measure selected and is able to convey those limitations to the command staff or other end users of the analysis so that they can be taken into account when any decisions are made based on those data.

On the other hand, meaningful analysis of CFS data also may offer the insight necessary to optimize resources, particularly personnel. Figure 5.2 represents a notional illustration of a relatively common pattern of complaint data that

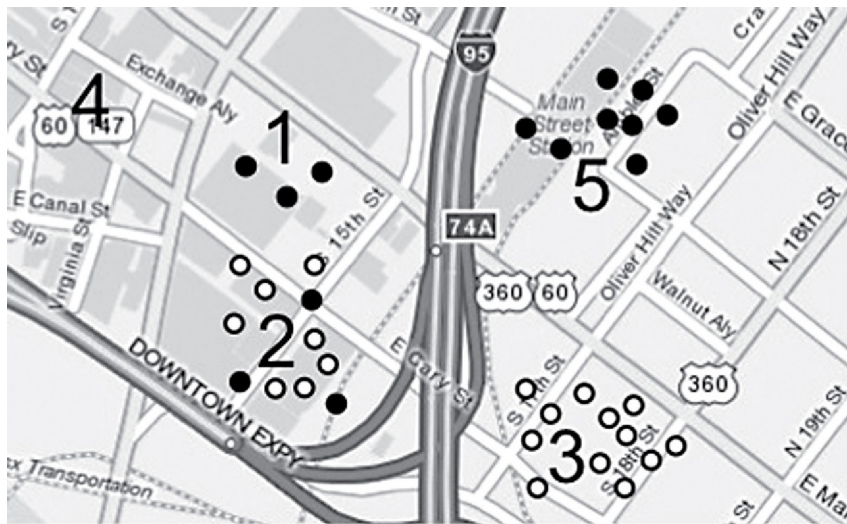


FIGURE 5.2 Map illustrating citizen complaint data or Calls for Service (CFS). The white dots depict complaints for disorderly conduct and fights; the black dots provide the locations for armed robbery complaints. The numbers 1–5 indicate the patrol area.

include clusters of citizen complaints for disorderly conduct and fights (the white dots in the figure), and armed robberies (the black dots in the figure) for a typical Saturday. While all of these incidents were in the same general area, review of the CFS data revealed that the fights and disorderly conduct complaints in the locations marked as 2 and 3 on the map generally increased during the evening hours, escalating as midnight approached, and the street robberies occurred shortly thereafter in areas 1 and 5. Area 4 remained relatively quiet. Standard procedure was to assume a reactive posture; responding to complaints as they were received, which was resource intensive.

Figure 5.3 illustrates a hypothetical patrol deployment schedule or “heat map”¹³ that visualizes relative differences in the frequency of citizen complaints over time¹⁴ and location. While the CFS appear to be different in time, location, and the nature of the complaint, further examination of the context revealed that all of the CFS were generally located in an area associated with nightclubs and restaurants. On further review of the complaints, the parties involved in many of these incidents – regardless of location, time, or nature of the complaint – were revealed to be patrons of these nightclubs. A larger, unifying scenario quickly emerged, which involved individuals who may have overindulged at the nightclubs and bars, and were boisterous, disorderly, and easily provoked into fights when they exited these establishments. These same individuals then represented relatively easy targets for street crimes given their intoxicated state. The complaint times were generally consistent with closing time for the bars (disorderly conduct and fights), and the related travel time to street parking and outlying lots (armed robberies). The solution included the use of mounted patrol deployed to patrol areas 2 and 3 shortly before closing time,

Sector	Shift					
	0000–0359	0400–0759	0800–1159	1200–1559	1600–1959	2000–2359
1						
2						
3						
4						
5						

FIGURE 5.3

This figure illustrates a hypothetical police deployment schedule for the citizen complaint data depicted in Figure 5.2 that includes the time of data and police patrol area for a particular day of the week (Saturday).

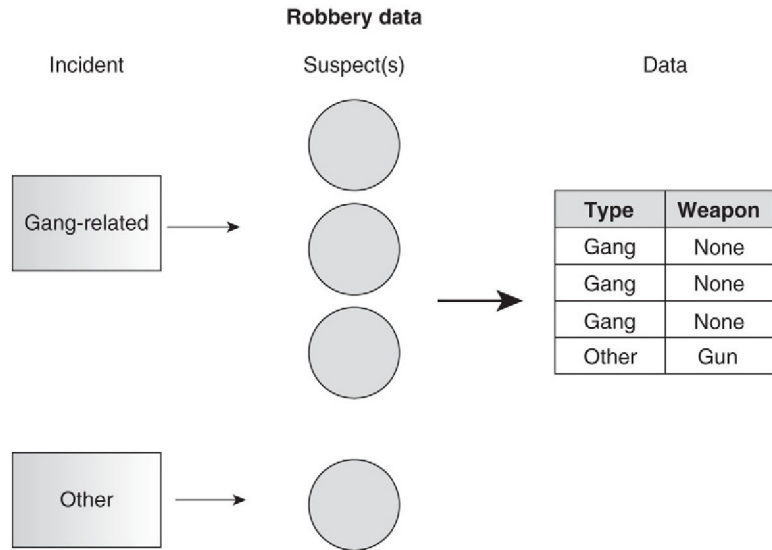
which resulted in a good, very visible police presence and exerted a deterrent effect on the fights and disorderly conduct. These resources were then flexed to the outlying areas where they similarly represented a visible police presence and created an unattractive environment for armed robberies and other street crimes. Overall, analysis of police CFS data enabled a novel insight regarding interrelated patterns of crime, which were effectively addressed through the fluid allocation, redeployment, and effective optimization of patrol resources. Therefore, by using the same resources to address what originally were considered multiple, independent patterns requiring deployment of multiple, independent resources, CFS analysis resulted in a virtual force multiplier for patrol.

5.4.2 Relational Data

Relational databases are comprised of a series of associated or linked tables.¹⁵ This facilitates the inclusion of a greatly expanded amount of information; however, it can create some challenges in analysis and interpretation of data, as outlined next. An example of a relational database that is frequently encountered in law enforcement and public safety is the National Incident-Based Reporting System, or NIBRS.¹⁶ In marked contrast to Uniform Crime Reports (UCR), with NIBRS data a single incident can be associated with multiple offenses, victims, and suspects. For example, if two suspects broke into a home, vandalized the kitchen, stole a television set, and assaulted the homeowner when he came home early, it would be a single incident associated with as many as four separate offenses, two suspects, and one victim under NIBRS rules. With UCR, only the most serious offense would be reported. The other, “lesser included” crimes would not be reported. While NIBRS results in more complete reporting of crime statistics, there are challenges associated with databases of this nature.

First, it is possible to greatly magnify the prevalence of crime if crimes are counted by offense rather than incident. In the example above, simply adding up the offenses would greatly exaggerate total crime statistics. As many as four “crimes” could be reported for what really is one single crime incident that was associated with multiple offenses. Similarly, it is possible to underreport crime when there are multiple victims. For example, using an incident-based reporting method, a double homicide counts as one homicidal incident with two victims. This can be confusing for those accustomed to reporting homicide totals in terms of a body count.

In terms of crime analysis, however, one particular challenge associated with relational data as opposed to a simple “flat file” is that it can be difficult to maintain relationships between variables and to ensure that certain variables are not overemphasized during an analysis, particularly using standard methods. For example, gang-related crime frequently involves multiple suspects associated with a single incident. As illustrated in [Figure 5.4](#), a gang robbery

**FIGURE 5.4**

Relational data are associated with some unique challenges. In this example, a gang-related robbery was perpetrated by three suspects while another robbery was associated with only one suspect. In a relational database, there could be as many as three suspect records associated with the gang-related robbery, compared to a single suspect record associated with the other robbery. The gang-related information will be overrepresented in the analysis unless some precautions are taken to ensure that each incident is counted only once.

was perpetrated by three suspects, while another robbery was associated with a single suspect. When this information is entered into a spreadsheet, there are three records associated with gang-related robbery suspects as compared to a single suspect record associated with the other robbery. At a minimum, the gang-related information will be overrepresented in an analysis of robbery-related suspect characteristics unless some precautions are taken to ensure that each incident is counted only once.

For example, if the gang members used intimidation while the suspect in the other robbery used a gun, a simple suspect-based count of weapon involvement would indicate that weapons were used in only 25% of the robberies, when in fact a gun was used in one of the two, or 50%, of the incidents. On the other hand, the prevalence of gang-related crime might be skewed if suspect-based statistics are used to generate that information. It is always important to think logically about what the question really is and what is the best way to count it. Some crimes, including those involving juveniles and gangs, frequently involve multiple suspects. Obviously, this is not much of an issue with only two incidents; however, when faced with hundreds of incidents, it is important to ensure that these issues are considered and that crime is counted accurately.

5.4.3 Revisiting the INTs

The different “INTs” – OSINT, HUMINT, SIGINT, GEOINT, and MASINT – were introduced in Chapter 4, but it is important to consider these as starting points. For example, Social Media Intelligence (SOCMINT)¹⁷ has increased markedly in use and value recently, particularly given the role that social media has played in the organization, management, and reporting on such events as the Arab Spring. As the activists noted, “[w]e use Facebook to schedule the protests and [we use] Twitter to coordinate, and YouTube to tell the world.”¹⁸ Similarly, as the Internet increasingly is used for surveillance, target selection, and operational planning, the ability to engage in meaningful analysis of Internet activity including web mining is becoming increasingly important.

5.4.4 GEOINT¹⁹

Geospatial Intelligence (GEOINT) is becoming increasingly available to the analyst given the ease of use and accessibility of tools that were until recently confined to experts. Again, GEOINT includes information obtained through satellite, aerial, and ground-based collection methods that is used to describe, visualize, and accurately locate physical features and human activity on the Earth. There are two basic data models for geospatial data: raster and vector. The raster data model represents space as a regular grid of equally sized cells, and the vector data model represents space as points, lines and polygons. Beyond raster and vector, geospatial data also can be further segmented into physical and human geography.

5.4.4.1 Physical Geography

Physical geography includes physical features and associated phenomena of the earth including landforms, hydrology, oceanography, climate, vegetation, soils, and fauna.

5.4.4.2 Human Geography

The National Geospatial-Intelligence Agency (NGA) has defined human geography or “human terrain” as, “[t]he spatial differentiation and organization of human activity and its interrelationships with the physical environment.”²⁰

The NGA has further subdivided human geography into 13 elements or themes that “best characterize the people and their culture within the context of their environment”²¹:

- Demographics and Population Measures
- Language
- Religion
- Ethnicity
- Education
- Medical/ Health

- Political Affiliation and Ideology
- Economy
- Land Use, Cover and Ownership
- Transportation
- Water Supply and Control
- Communications and Media
- Significant Events

Although defined and described here as unique and separate elements, physical and human geography frequently interact and combine to define space and structure, enable, and/ or constrain movement and use. This is particularly true as relates to availability of and access to victims or targets, and the selection of an environment where the individual believes that they will be able to successfully perpetrate their desired acts. This interaction between physical and human terrain will be explored in greater detail in Chapter 7 in the section on Geospatial Predictive Analysis.

5.4.5 Ad Hoc, Self-Generated, and Other Specialized Databases²²

In some situations, the department's records management system does not meet the needs of timely, complete crime analysis. Specialized databases can be created to address this need. These may be ongoing or case-specific. For example, in a department without direct field entry of data or limited availability of MO characteristics in existing data resources, the analytical team might construct databases specifically designed for crime or intelligence analysis. These databases might either be offense-specific, such as a homicide database or a robbery database, or associated with a unique series or pattern of crimes, such as a serial rapist or an unusual series of burglaries. These databases will include standard information likely to be found in the department's records management system (e.g., case number and date, time, and location of offense), as well as information related to specific or unique MO characteristics or behavioral themes.

It is not unusual for a task force or department to establish a specialized database for a particular series of crimes or a high-profile crime. For example, child abduction cases frequently receive large amounts of data and information that need to be entered, managed, and analyzed very rapidly in support of these often fast-breaking investigations. Tip databases are created in an effort to manage the rapid accumulation of data and information in response to a high-profile incident. For example, in the first few weeks of the Laci Peterson investigation, more than 2600 tips were received by law enforcement authorities.²³ Similarly, at the height of the Washington D.C. sniper investigation, authorities received as many as 1000 tips an hour.²⁴ The sheer volume of information associated with these cases requires some sort of automated system of information management.

Extending the casualty data analysis example in Chapter 1, another noteworthy example of the value that an ad hoc database can provide regarding meaningful insight and informed action comes from the U.S. Department of Defense (DOD).²⁵ The DOD provides daily, online updates regarding American military casualties that include fatalities, as well as nonfatal injuries. These nonfatal injury data are segmented further into injuries where the service member survived the injury but could not return to duty, and those injuries where the service member was returned to duty within 72 h;²⁶ segmentation that provides important insight regarding troop strength and readiness, as well as health care resource allocation decisions. Again, the challenge is not only collecting accurate and reliable data, but also segmenting the data in a meaningful way to reveal detail and context, and inform fact-based decision-making.

Deeper analysis of these data, however, reveals a promising reduction in the lethality of combat injuries as indicated by a decreased rate of fatalities. Similar to the trends in U.S. homicide rates, though, trends in combat fatalities frequently reflect access to timely and/or high-quality trauma services, rather than changes in overall casualty rates – underscoring the importance of understanding trends and patterns embedded within the data. Further analysis of the data suggests that improved outcomes for the complex injuries associated with Improvised Explosive Devices (IEDs) are at least partially responsible for the reduced fatalities rates. Also, like U.S. violent crime trends, these improvements in combat casualty outcomes frequently were made by doing more with less – without new technology or treatments. Additional background reveals that three senior military physicians at the 31st Combat Support Hospital in Baghdad took the initiative to start collecting more than 75 different data points on every casualty in an effort to document important trends and patterns associated with outcomes and make a “science out of performance.” They also used these data to demonstrate that civilian trauma criteria were not reliable indicators in combat injuries given marked differences in the nature, complexity, and severity of the injuries. These results demonstrated the value that segmentation by injury type and the collection of additional data may provide in creating the insight necessary to develop information-based decisions and improved outcomes. Moreover, this example further underscores the value of innovative approaches to the collection of the data, as well as the importance of knowing the denominator and larger context in support of meaningful analysis, interpretation, and operational use of the results.

Limitations to this approach include the initial time commitment necessary to create the database, an ongoing commitment to data entry, and the possibility of errors associated with data entry. While creation and maintenance of specific analytical databases require a commitment from the analyst or analytical team, it can reap huge benefits in terms of detail, additional information, analytically relevant information, and timely data entry. Moreover, as the casualty example

highlights, the additional insight gained from these ad hoc databases can surface better understanding and support innovative approaches and change outcomes.

What do you include in a self-generated database? As much as you think that you will need, and then some. When it is time to conduct the analysis, it always seems that no matter how many variables and how much detail was included in the original data set, it would have been better to have included just a little more. Behavioral characteristics and themes and MO features are always a good starting point. Some variables will become standard (e.g., approach, demeanor), while others might be unique to a specific pattern of offending or crime series. It generally is a good idea to include the information in as close to its original form as possible. Automated recoding later can be more efficient and accurate than trying to recode during the data entry process. In addition, early exploration of the data might indicate a need for one form of recoding over another. For example, during an analysis of police pursuits, preliminary analysis of the data revealed potential differences among high-speed pursuits as compared to those at lower speeds. This was not readily apparent, however, until the data had been entered in their original form and then explored.

One challenge associated with tip databases is the fact that the analyst and investigative team typically have very little information early in the investigation, while the database is being created. With such limited information, it might be unclear which variables should be included because the overall direction of the investigation might not have emerged or been developed. Moreover, although the team always tries to be objective, a favored outcome, suspect, or interpretation can dramatically impact the structure of the database as well as the interpretation of individual data elements or reports.

The analyst is challenged further by an overwhelming amount of information with an associated need for rapid analysis. During the D.C. sniper investigation, the investigative process was complicated by the involvement of multiple localities, jurisdictions, states, and task forces. The net result of this type of situation is a data repository that is beyond the analytical capacity of a single analyst or even an analytical team or task force. There is just too much information to absorb, categorize, remember and draw meaning from, and this significantly compromises the overall investigation. In these cases, the best solution is to employ automated search strategies and analysis. Software does not favor any suspects or outcomes and does not become overwhelmed by the amount of information or the nature of the case.

Another challenge associated with tip databases is that the information frequently arrives in narrative format. Tipsters rarely call the hotline with detailed, well-categorized information that has been recoded to match the existing database structure. Rather, they tend to provide information that is in a narrative

format, frequently is incomplete, and sometimes even is inaccurate. Armed with natural language processing and text mining tools, an analyst can explore and analyze large amounts of narrative information extremely efficiently.

Consistency, consistency, consistency! This can be an issue when multiple analysts are entering the same data. Even subtle differences in data entry can greatly increase the variability, with concomitant reductions in the reliability and analytical value of a data set. The decision to develop and maintain a database is huge; do not sabotage your efforts with inconsistency or variability in the data entry process. There is enough variability in crime and intelligence data; the analyst does not need to introduce more in the data categorization and entry process.

The analyst frequently is called to analyze data or information that is collected specifically for a particular case. These data often embody the most difficult yet interesting work that an analyst can become involved in. The opportunity to pull meaning and investigative value from a seemingly unintelligible mass of information can be one of the most exciting challenges that an analyst encounters. While it is impossible to anticipate the exact nature of these data sources, learning some basic data management techniques can help to not only evaluate the reliability and validity but also do any data cleaning or recoding that might be required. Some common data resources are addressed later in this chapter and throughout the text.

5.4.6 Nontraditional Sources (e.g., Weather)

Data collection should be compliant with legal requirements and ethical constraints, but many novel or nontraditional sources are available and can add value to the analysis. For example, normal seasonal weather trends, as well as dramatic changes, can have an impact on crime trends and patterns, particularly as they may influence the behavior of and access to potential victims or targets. For example, during the colder months, many individuals preheat their vehicles. Since it is easier to steal a vehicle when the ignition key is available, these cold temperatures can be associated with an increased number of vehicle thefts during weekday mornings when people are preparing to leave for work. Similarly, people are frequently tempted to leave their vehicles running during the summer months in an effort to keep their vehicles cool when they run into the convenience store to make a quick purchase, and the “fighting season” in Afghanistan has become so well known as to be the subject of public commentary and discussion. These anecdotes support the fact that weather data can add significant value to the analysis of crime and intelligence data, as well as associated operational planning to include anticipation, prevention, thwarting, and influence. Data mining and predictive analytics tools are particularly well suited to fully exploit the possibilities associated with this type of novel analysis.

As outlined in Chapter 3, the retail giant and analytic competitor Wal-Mart has used their point of sale data to document marked differences in purchasing behavior associated with major weather events.²⁷ Similarly, less frequent but significant weather events also can have an effect on criminal behavior. For example, looting after a devastating weather disaster is not uncommon. Several years ago, we noticed that violent crime, particularly street violence, decreased to almost nothing during a major snow event and that subsequent clearing of the weather and roads was associated with a concomitant spike in crime, particularly random gunfire complaints. While it makes sense that anything that alters routine activities, including the distribution of or access to potential victims would have an impact on something like street crimes, the increase in random gunfire after the quiet period associated with a major weather event was puzzling. In other words, the decrease in street crimes represented “confirmation” of what we knew; however, the increase in random gunfire complaints was surprising, representing a “discovered” relationship. Again, it is not necessary to understand why this relationship exists. Rather, knowledge of the relationship between a break in the weather and random gunfire complaints is sufficient to deploy resources and act.

This relationship had become so reliable in one locality that it was considered conventional wisdom in the police department that any sort of major weather event that restricted travel and confined people to their homes would be followed by a sharp increase in random gunfire. This point was mentioned to a new chief during preparation for a hurricane. Incredulous, he told the crime analysts to document this effect and report back to him when they had the numbers, which they did. After compiling complaint data, the analysts were able to confirm that period during the storm was relatively quiet, save for a few folks trying to buy drugs, while the period immediately after the storm was associated with marked increase in the number of citizen complaints for random gunfire in the community.

Again, while I could speculate regarding the true meaning behind the association between inclement weather and random gunfire, the truth is that I have absolutely no idea why this occurred in that particular community. The important thing is, however, that the relationship was noted and documented, and could be anticipated and effectively responded to in the future. New Orleans, Louisiana, also reported changes in criminal activity in the aftermath of Hurricane Katrina. Some of this was likely related to dramatic changes in geography, including the fact that large sections of the city were under water. Other emerging patterns of crime likely were related to efforts to acquire food and water, and fear associated widespread misinformation and rumors regarding civil unrest and violence, as well as outright looting. Similarly, Houston, Texas, noted significant storm-related changes in its crime patterns and trends, including marked increases in some patterns of offending associated with the

large numbers of displaced persons, which also included known criminals and gang members embedded within the evacuees.

Although largely anecdotal, these findings highlight the value that nontraditional crime measures like weather can have with regard to forecasting and strategic crime analysis. Historical weather data are relatively easy to get. For example, local television stations often maintain archival information, which is available over the Internet in many locations. Again, be creative. Crime patterns frequently are as unique as the individuals involved and can be affected by a variety of obvious, as well as not so obvious, factors in the community. By transcending the analytical boundaries associated with a single type of data or information, the analyst can begin to identify and analyze a larger array of factors and potential consequences associated with the original information of interest.

5.5 DATA CHALLENGES

5.5.1 Reliability and Validity

Regardless of how perfect a data set might seem to be, it almost always has some shortcomings. In law enforcement and intelligence analysis, the data and information generally are anything but perfect. In fact, I often experience “data envy” when reading scientific or business-related papers on data mining because of the tremendous amount of quality control the authors often have over their data. In contrast, it is not unusual for an analyst to receive some data or information in a format completely unsuitable for analysis. For example, a billing invoice, complete with headers and other meaningless formatting, might contain information critical to the development of a timeline or identification of routine expenditures. Not only do these data come in a less than desirable format, there frequently is the expectation that they will be turned around very quickly. Any delay in the investigative process can make the difference between a case that is cleared and one that languishes with no proper closure for a long time, if ever. The last thing that anyone wants is a delay in the investigative process because the analyst can neither tolerate nor accommodate less than perfect data.

Data layout and design often can be addressed. Beyond format, however, are far more sinister issues that need to be addressed: reliability and validity. Reliability implies a degree of stability in the measure. In other words, reliability means that if you conduct repeated measurements of the same thing over and over again, the measurements will not change significantly. For example, a witness statement is reliable, in statistical terms, if the witness says approximately the same thing at each interview. This same statement has absolutely no value, however, if the witness is not telling the truth. Therefore, the second measure of interest is validity. Validity simply means that the measure, in this case the witness statement, is an accurate measure of truth, or what actually happened. Another term for validity is accuracy.

These definitions differ somewhat from the traditional law enforcement or intelligence definitions of reliability. For example, a reliable informant is one that is both dependable and accurate. In many ways, this working definition of reliability represents a composite of statistical reliability and validity, which only serves to highlight the importance of each measure. A dependable witness with poor information would be of little value, just as one who has good information but cannot be counted on also would have limited utility.

Inaccurate or unreliable data can arise from a variety of sources, including everything from keystroke errors to intentional corruption of data and information. Because these issues are important to accurate analysis and the development of meaningful and reliable models, some of the more common challenges will be addressed in detail.

5.5.2 Data Entry Errors

What happens when you run across a “juvenile offender” with a listed age of 99 or a male with a previous pregnancy in his medical data? Several years ago, when analyzing juvenile offender data, I was assured that the medical section was both accurate and reliable. It turns out that the data were neither.

To convince myself of this, I ran a quick frequency distribution that listed the number of occurrences for each possible value associated with a particular measure. What I found was that a significant number of male offenders were listed as having experienced pelvic inflammatory disease, something uniquely female. It was unlikely that the physician performing the intake medical examination made the error. Rather, it appeared to be an error that occurred later during the data compilation or entry phase. Either way, it significantly compromised the value of the information.

Data entry errors happen. It is a monotonous process, and people get fatigued. There might be incentives for speedy data entry without the necessary quality control, although this seems to be changing, with some automated reliability and validity checks now being included in some records management systems. Sometimes people just do not care. It is not the most glamorous job in law enforcement, and generally is not well compensated. Even under the most stringent conditions, however, data entry and keystroke errors happen. The solution, then, is identifying and correcting them.

Frequently, running a quick frequency distribution can highlight information that appears to be grossly out of range. For example, running a frequency distribution on age will tell us how many 25-year-olds there are, how many 26-year-olds there are, and so on. This method will highlight any data points that are well beyond what one would normally expect and that should be investigated further. As in the earlier example, a 99-year-old juvenile offender clearly is incorrect. In many cases, however, the value “99” is used for missing data or

when the information is not known. Therefore, an entry of “99” might mean that the information was unavailable or unknown. This can be clarified and addressed.

It is important to note that there are cases where the information is unknown. Developing ways to indicate this within a data set can be extremely important. If those methods were incorporated, the ages listed as “99” would be excluded automatically, rather than contaminating subsequent analysis. Using indicator variables for missing data also is important because blank fields in a data set can indicate many things. For example, if work history or current employment is left blank in a file or data set, does it mean that the subject in question has never worked, or that this information has not been collected? Similarly, does it mean anything that this information could not be found? For example, it can be significant to know that a listed address does not exist. Understanding the importance and implications of missing, inaccurate, or unreliable data can greatly enhance the value and understanding of the information collected, as well as the subsequent analyses.

Returning to the juvenile offender data, there also were greatly differing reports of prior drug history in the section on substance use history. Further examination revealed that the nurses were streetwise and savvy. Few inmates lied to them about their drug use, and those who did were challenged by these health care providers. As a result, the medical history information collected by the nurses was determined to be relatively accurate. The physicians, on the other hand, frequently came from different environments and tended to have less experience with juvenile offenders. As a consequence, they tended to believe what they were told. One physician in particular was extremely nice and soft-spoken. Interestingly enough, her physical examinations revealed very little substance use and even less sexual activity among the juvenile offenders who she interviewed, which stood in stark contrast to information collected in other environments by other personnel. Unfortunately, this finding cast a shadow of doubt over all of the other information that she collected during the intake process. This situation highlights how easy it is to encounter unreliable information. One detective might be particularly adept in an interview situation, while others might have less skill in eliciting information. All of these factors can significantly affect the reliability of the information.

This highlights another important concept – the reliability check. Whenever possible, it is extremely valuable to cross-check information for consistency and accuracy. In the example with the physician, we were able to check her information with that of other interviewers and found hers to be different. Further examination of the data that she collected proved it to be inaccurate, which cast doubt over almost everything else that she had collected. Similarly, if we can compare some data against known information, it gives a greater degree of comfort with information that cannot be validated or checked. This

is not an uncommon practice in law enforcement or intelligence information collection, and represents a form of best practice in the assurance of information integrity.

It also is possible to gain some information from response rates alone. For example, an unmanageable crowd at a crime scene can be unnerving, but the spookiest scenes for me were the ones where nobody was out. The total absence of spectators said something. Similarly, increases in citizen complaints can be a troubling issue to address in an evaluation of a crime reduction initiative. In some communities truly ravaged by crime, the citizens can become so discouraged that they give up: “What is the point in calling if nothing ever changes?” Crime becomes so expected and normal that the outrage is gone. In these situations, one of the first indicators of improvement can be a spike in complaints, as members of the community begin to reengage and participate in community public safety. As in music, sometimes the spaces between the notes can be important to identifying the tune.

5.5.3 Misrepresentation, Fabrication, and Poor Recall

Criminals often misrepresent the facts, both when it matters and when it does not. For that matter, so do witnesses, informants, and many other assets that we use on a regular basis to help us to gather information and data. Similarly, other people withhold information because they have something to hide, are concerned that they might implicate themselves or someone else, or just because. In addition, victims can be emotional or confused, and often make poor historians. Filling out an accurate and reliable offense report generally is not what is going through their mind as they are getting robbed at gunpoint. It is not unusual to receive a very detailed, if not somewhat exaggerated, description of the weapon and little to no good information regarding the suspect. This is not surprising and reflects victims’ focus during the incident. In addition, the incident might be very brief and occur under less than optimal lighting. While convenience stores generally have done a good job of providing good illumination and height markers by the door, the average victim of a street robbery is at somewhat of a disadvantage regarding all of the information that we might like to include in the analysis.

This is not unique to law enforcement or intelligence data. Respondents to marketing surveys lie as well. Think again about the last survey that you may or may not have completed. Were you honest? How do you think your less-than-honest responses might have affected the results? Data science professionals in other domains encounter “professional” survey takers who maintain numerous identities in an effort to collect money for their surveys and profiles, and another group that likes to make mischief with data. All of this diminishes the reliability and particularly the validity of the data that we encounter and confirms that there are limited options for the “perfect” data set.

5.5.4 Unsuccessful Criminals

It bears repeating that almost everything that we know about crime and criminals is based on those who have been caught – unsuccessful criminals. We all laugh when we hear about the bank robber who used his own deposit slip to write the note or the burglar who dropped his wallet on his way out of the house, but the sad fact is that many criminals are caught because they make mistakes. Research on nonadjudicated samples is extremely difficult. For example, asking individuals about involvement in nonadjudicated criminal activity, particularly felonies, is challenging both legally and ethically, due to mandated reporting requirements for some crimes. As a result, studies that involve gathering information on nonadjudicated crimes frequently are difficult to get approved by human subject review committees. There also are challenges in even identifying these populations, particularly those with limited or no contact with the criminal justice system. Certain types of ethnographic research, which frequently involve going out into the communities and locations where criminals operate, can be extremely risky. They also are prone to artifact in that they often rely on subject referrals from other criminals. Since criminals tend to be less than honest and frequently are unreliable, depending on them to help establish a solid research sample can be tenuous at best. Therefore, because most criminal justice research is conducted on identified offender populations, such as those already in prison, there are some significant gaps in our knowledge regarding crime and criminals. Knowing what is normal and what is not is absolutely essential to developing the domain expertise necessary to evaluate the results. This is covered in greater detail in Chapter 10.

5.5.5 Outliers with Value

All outliers are not created equal. Should outliers universally be removed from the analysis or otherwise discounted? Or is an outlier or some other anomaly in the data worth considering? While most outliers represent some sort of error or other clutter in the data, some are extremely important. In my experience, deviation from normal when considering criminal justice or intelligence data often indicates something bad or a situation or person with significant potential for escalation.

5.5.6 Duplication

What happens when duplicate records are encountered in a data set? This is not at all uncommon, as multiple citizens can call about the same fight or shooting, multiple officers might respond to the same complaint, and so on. In the random gunfire example outlined previously, some complaints were obviously duplicative. As such, they received identical call numbers. In [Table 5.2](#), this same data set is depicted without the duplicates. Obviously, there was duplication beyond what could be culled through simple identification of duplicate call numbers, but this would represent a good start.

When does duplication have value? Perhaps one of the more common areas is in workload studies and deployment. Knowing when multiple units are dispatched to a single incident is important in determining how personnel resources are being used. For example, knowing that multiple officers are dispatched to domestic complaints, while only one officer generally responds to an alarm call, is essential to a complete understanding of police personnel workload. It would be difficult to anticipate every possible situation when duplication occurs and when it is necessary to answer a particular question.

A somewhat more complex example of both necessary and unnecessary duplication in a data set occurred in the following example. Briefly, an organization of interest was linked to a billing invoice that included hundreds of individual telephone conference calls comprised of thousands of individual telephone call records. Further examination of the invoice revealed that some individuals regularly participated in calls with a similar group and that certain individuals appeared to have been involved in multiple conference calls. This was important information, in that identifying the key players and linked individuals helped begin to reveal an organizational structure and associated relationships. Review of the records also revealed that some individuals might have dialed into a particular conference call multiple times. This could have been related to bad or unreliable connections or a variety of other reasons. At this point, though, the underlying cause for the duplication in calls is not nearly as important as the fact that it exists. Calculating frequencies for individual callers had to be delayed until this unnecessary duplication was addressed. Unfortunately, culling the unnecessary duplication within calls while maintaining the necessary duplication between calls can be a very complex task, particularly with a large data set. Automated methods for accurately culling these data were available and saved time while maintaining valuable relationships and features of the data set. In this example, the data set was reduced by almost 1/2 by removing the unnecessary duplication in the data.

5.5.7 Other Considerations

Another important issue in this discussion is that crime is not distributed evenly. In fact, it is rare to find evenly distributed crime. For example, criminals, at least successful criminals, will tend to select affluent areas to look for expensive jewelry and electronics, while open-air drug markets tend to be located in higher-risk areas. Subtle attributes of the environment, which will be discussed in greater detail later, will guide victim or target selection. Even something as simple as traffic stops can be skewed by police patrol deployment, since areas associated with high crime also tend to be associated with increased police deployment. Logic follows that the greater the police presence, the more likely someone is to get caught with an expired city decal or burned-out headlight. Similarly, while it can be hard, although not impossible, to skew homicide rates,²⁸ statistics for vice offenses tend to be driven largely by arrest rates.

Therefore, a reduction in or absence of arrests for drugs or prostitution does not necessarily mean that those crimes do not exist. Drugs, guns, and violence typically go together, and certain types of vehicles tend to be stolen more than others. Whether this is due to a preference for certain vehicles or the fact that some vehicles might just be easier to steal than others is important for prevention efforts. Overall, though, it is important to consider possible heterogeneity in the distribution of crime when drawing conclusions regarding differences between locations or over time.

5.5.8 We're Not the Only Ones Using the Data...

It is also important to note that the bad guys frequently have access to and use the same data sources as we do including school schedules, known supply routes, pay dates, and fraud “pay and chase” decision rules, just to name a few. We are not the only ones collecting data in support of informed anticipation and influence. In some cases, their knowledge is almost certainly better than ours, and they do not work under many of the same requirements and constraints that face the crime and intelligence analyst, including restrictions on the use of personally identifiable information, the creation and use of some derived products, and other restrictions that may be associated with access to and use of classified material. Examples of our adversary’s effective use of open source data include the Mumbai attackers who used Internet resources to plan and actively manage the attack,²⁹ while more recently members of al Shabaab have used social media to recruit followers and manage their attack on the Westgate Mall in Nairobi, Kenya. Moreover, incidents like the Beslan School siege, and the increasingly sophisticated complex attacks perpetrated throughout the Middle East and Africa underscore a deep and nuanced understanding of human behavior, particularly in response to critical incidents and assaults that is used to structure extremely complex and effective attacks. Finally, recent retailer data breaches have included more than point-of-sale data. The derived variables and additional information related to “pattern of life” included in many of these files may be even more sensitive than the financial data if they end up in the wrong hands, particularly given their potential value to structuring sophisticated spear phishing attacks and related patterns of fraud.

5.6 HOW DO WE OVERCOME THESE POTENTIAL BARRIERS?

Establishing as many converging lines of evidence as possible for validity checks can be helpful. This can include verification with known, reliable sources of information like arrest records. We also might look for similarities in false stories. For example, “the bullet came out of nowhere” explanation is likely to arouse suspicion among investigators. An analyst can benefit from the development of similar internal norms, or domain expertise regarding normal

trends and patterns, as well as common themes of deception. People just are not that creative when it comes to prevarication, and sometimes information about inaccurate or false statements can be as valuable as credible information. For example, in some types of statement analysis, the linguistic structure of a statement can be as important as the actual content in evaluating the validity of the statement. Most valid statements have a beginning, middle, and an end. False allegations frequently deviate from valid statements in the emphasis of certain portions of the story. Once this type of deviation has been identified, it is possible to begin to evaluate content in a different light and look for potential secondary gain related to a possible false allegation. Data mining and predictive analytics can be extremely valuable, particularly regarding the analysis of unstructured narrative and identification of commonly occurring patterns associated with false allegations.

It is important to note, though, that the lack of good data in this space is not meant to imply that there is an actual lack of data.³⁰ In fact, the exact opposite frequently is true. The public safety and security domain generally does not lack for data. Rather, these analysts lack reliable, accurate, and valid data. In other words, what they need are GOOD data. Overall, though, the trends are heading in the right direction. There are increasing opportunities to play a role in the development of key analytic resources and the analysts can use their insight, experience, and tacit knowledge to help their IT colleagues get it right. Knowing certain data are questionable or false, however, does not mean that they can be ignored. Just the opposite, it is extremely important to know as many details as possible about any limitations in the data so that they can be dealt with during the analytical process and interpretation of the results. It is like medical side effects. There are those that you know about and those that you do not. Generally, those that you are not aware of create the most trouble in the end.

Finally, the best rule for data collection is to plan ahead, if at all possible. Like evidence collection, you rarely get a second chance to collect the data. Moreover, the more that you can anticipate the recoding requirements or strategy, the more you can streamline and increase the efficiency of the collection, data entry, recoding, and even fusion. It is very sad to encounter a dataset created and populated by a well-intentioned department or agency only to encounter a serious flaw that severely limits the value of the data, or even precludes their ability to answer the questions that prompted the original creation of the database. Think about what you are trying to do with the data. Mentally, perform the analysis or even use preliminary data or notional data to test the database, its functions, and structure. The ability to perform analyses early in the process and related flexibility to adjust or refine collection strategies, categories or coding schemes, and even data access is invaluable, if possible. Again, we frequently have little to no involvement in the creation of the data resources

or databases that we use. Much of our data is collected on an ad hoc basis or is uncovered during the course of an investigation. However, with the increasing emphasis on refining collection strategies, actionable knowledge management, and analysis-ready architecture, there are increasing opportunities to play a role in the development of these key analytic resources. The analyst can use their insight, experience and tacit knowledge to get it right from the beginning.

Bibliography

- 1 Lowenthal MM. *Intelligence: From secrets to policy*. CQ Press, Washington, DC; 2000.
- 2 Elder J. Modeling analytics dream teams. Team Analytics, webinar, 11 September 2013. http://www.talentanalytics.com/wp-content/uploads/2013/08/FINAL_Modeling_Webinar_TalentAnalytics-91113.pdf; 2013.
- 3 Fink S. *Five Days at Memorial*. Crown, New York; 2013.
- 4 Ripley A. *The unthinkable: who survives when disaster strikes – and why*. Three Rivers Press, New York; 2009.
- 5 Howell D. *Statistical methods for psychology*. 3rd ed. Belmont, CA: Duxbury Press; 1992. Those without recent, or even any, statistical training might benefit from acquiring a basic introductory statistics text that can be used as a reference. This is a very good introductory text that would serve well in this capacity.
- 6 Dumbill E. What is big data? O'Reilly Radar. <http://strata.oreilly.com/2012/01/what-is-big-data.html>; 2012 [accessed 11.01.12].
- 7 Many programs, intentionally or otherwise, telegraph their “pay and chase” business rules, including investigative and prosecutorial thresholds, which concomitantly inform the decision rules for the fraudsters. For example, if the investigative threshold is established at two hundred thousand dollars, it is not unusual to see patterns of distributed fraud that fall under, sometimes just below, this threshold. The fraudsters generate their desired revenue by distributing the fraud over multiple transactions and/ or multiple individuals or entities; hence the name, distributed fraud.
- 8 De Libero G. Using big data to invent the future. Big Data Forum. <http://www.big-dataforum.com/98/using-big-data-invent-future>; 2014 [accessed 10.02.14].
- 9 Brown ED. Context and big data. Big Data Forum. <http://www.big-dataforum.com/103/context-and-big-data>; 2014 [accessed 12.02.14].
- 10 Flynn MT, Pottinger M, Batchelor PD. *Fixing Intel: A Blueprint for Making Intelligence Relevant in Afghanistan*. 5 January 2010; 2009.
- 11 McCue C, Parker A, McNulty PJ, McCoy D. Doing more with less: Data mining in police deployment decisions. *Violent Crime Newsletter*, U.S. Department of Justice, Spring 2004;1: 4–5.
- 12 Sampson RJ, Raudenbush SW, Earls F. Neighborhoods and violent crime: a multi-level study of collective efficacy. *Science* 1997;277:918–924.
- 13 “Heat Maps” will be covered in greater detail in Chapter 13.
- 14 Law enforcement complaint and incident data generally use a 24-h clock. With this in mind, the complaints for a single date go from 0001 to 2400. Viewing these data as a standard “day of the week,” however, would include the very early morning hours immediately after midnight through the late evening hours immediately preceding midnight, which may not align well with a traditional “day of the week” model (i.e., “Saturday night” would include very late “Saturday” and very early “Sunday”), and impact visualization and analysis. This is not uncommon but is important to note, particularly as relates to crimes that run from the late evening to very early morning hours, spanning midnight (i.e., most crimes!), including weekends and

- special events like New Year's Eve. For example, in the analysis for the New Year's Eve initiative (McCue C, Parker A, McNulty PJ, McCoy D. Doing more with less: data mining in police deployment decisions. U.S. Department of Justice: Violent Crime Newsletter; Spring 2004. p. 4–5), the initial calculations suggested a 48 h “holiday” as indicated by increased complaints for random gunfire on December 31st and January 1st. When the actual distribution of random gunfire complaints were reviewed, though, it was clear that this activity was confined to a very brief period of time immediately preceding and following midnight; albeit, a period of only a few hours but spanning two “days” when dates were used.
- 15 Agosta L. The essential guide to data warehousing. Upper Saddle River, NJ: Prentice Hall; 2000.
 - 16 The IBR Resource Center can be found at www.jrsa.org. This is an excellent resource that includes many references, as well as analytical syntax specific to IBR data and contact information for others working with this type of information.
 - 17 Omand D, Bartlett J, Miller C. A balance between security and privacy online must be struck... #Intelligence. Demos, ISBN 978-1-909037-08-3, <http://www.demos.co.uk/publications/intelligence>; 2012.
 - 18 Kassim S. Twitter revolution: how the Arab Spring was helped by social media. <http://www.policymic.com/articles/10642/twitter-revolution-how-the-arab-spring-was-helped-by-social-media>; 2012 [accessed 03.07.12].
 - 19 A comprehensive review of GEOINT is well beyond the scope of this text. The interested reader is directed to an overview of basic concepts, http://www.youtube.com/watch?v=O5rdWilAq_s, and the National Institute of Justice Crime Mapping resources, <http://nij.gov/topics/technology/maps/pages/welcome.aspx>, as well as options for free online training. <http://training.esri.com/gateway/index.cfm?fa=seminars.gateway>.
 - 20 For a comprehensive review of Human Geography, c.f., National Geospatial Intelligence Agency. Incorporating human geography into GEOINT, student guide (SG). The School of Geospatial-Intelligence: NGA College, <http://info.publicintelligence.net/NGIA-HumanGeography.pdf>; 2011 [accessed 12.09.11].
 - 21 Ibid.
 - 22 Additional, specialized data resources worth exploring include the Armed Conflict Location & Event Data Project (ACLED, www.acleddata.com), which includes publicly available data on political violence for developing states; Invisible Children's LRA Crisis Tracker (www.lracrisistracker.com), which includes reliable, vetted, and validated data, maps, and related analysis on attacks perpetrated by the Lord's Resistance Army (LRA); and Ushahidi (ushahidi.com), which supports a variety of different capabilities that enable crowdsourced collection of information in support of crisis and conflict mapping.
 - 23 www.KXIV10.com; 2003. Despite avalanche of tips, police stymied in Laci Peterson case, January 9.
 - 24 Eastham T. Washington sniper kills 8, truck sketch released. 2002. www.sunherald.com [accessed 12.10.02].
 - 25 Gawande A. Better: a surgeon's notes on performance. New York: Picador; 2007.
 - 26 For a great review of this case study, as well as several others where data are used to guide information-based approaches to medicine, including the judicious allocation and optimization of scarce health care resource, see: Gawande A. Better: a surgeon's notes on performance. New York: Picador; 2007.
 - 27 Beck C, McCue C. Predictive policing: what can we learn from Wal-Mart and Amazon about fighting crime in a recession? Police Chief 2009;November.
 - 28 Homicides in some vulnerable populations (e.g., the homeless, runaways, prostitutes) can go unreported if nobody is found and no one notices or reports their absence. Outside the US, homicide data can be especially challenging to compile, particularly in active conflict zones and/ or locations with official corruption.
 - 29 Kahn J. Mumbai terrorists relied on new technology for attacks. New York Times, http://www.nytimes.com/2008/12/09/world/asia/09mumbai.html?_r=0; 2008.
 - 30 Flynn MT, Pottinger M, Batchelor PD. Fixing intel: a blueprint for making intelligence relevant in Afghanistan. January 5, 2010.

Operationally Relevant Preprocessing

“Does anybody really know what time it is?”

Chicago

As the famous song suggests, the concept of “time” is an abstract, often user-defined convention and there are numerous ways to calculate it. We can consider nanoseconds, minutes, days, weeks, seasons, or even millennia, just to name a few units of measuring time. Extending this, we also may consider time from a relational perspective. For example, in Chapter 5, we found that the period immediately after a major weather event was associated with increases in random gunfire. Similarly, juvenile justice researchers have found that the time between school dismissal and when parents come home from work can be associated with all manner of mischief.

Location also can be defined and measured in a number of different ways including geographic coordinates (latitude and longitude), Military Grid Reference System (MGRS), Universal Transverse Mercator coordinate system, Central Address Systems, or light years. Also like time, there are relational measures for time including the Euclidian distance between points and cost surface, as well as qualitative measures (e.g., “in the adult beverage aisle”). Moreover, the perspective of time and location may differ based on perspective and role (i.e., good guy, bad guy, victim, consumer, retailer). With this in mind, some discussion of the “when, where, and what” of crime, particularly as it relates to operationally relevant preprocessing of the data is in order.

As a researcher, I gathered as much data as possible in an effort to study and characterize the causes, consequences, and correlates of violent crime; particularly, drug-related violent crime. This included reviewing endless offense reports, autopsy protocols, and correctional databases. Like many other scientists, I focused almost exclusively on closed cases because there was generally accurate, reliable, and relatively detailed information available for the victim, offender, motive, crime, and scene characteristics. I was able to use statistical methods to address duplicative data, and could even in some situations surface

less accurate data and discard it in favor of information determined to be more accurate and reliable. Using this approach, I was able to develop some relatively good models of violent crime; confirming some known “truths” in the field, while also surfacing some “surprising” results.

The challenge with this approach, however, was that it had extremely limited value in the applied setting. For example, while it might be interesting to be able to make reliable predictions about the substance use patterns among individuals that perpetrate violence in illegal drug markets, and may have some implications in the correctional setting, it has limited value in solving or preventing those crimes. If I wanted to develop models that would have operational relevance, then what I really needed to do was confine the analysis to only those variables that could serve as predictive variables to support information-based prevention, or variables available immediately after a crime was committed to support enhanced investigative efficacy. Unfortunately, limiting the variables reduced the accuracy of the models. This is not unusual in the applied public safety and national security setting, and underscores why it is so important for the operational requirements and constraints to be included in the model evaluation process.

Similarly, while our modeling tools might be particularly adept at handling time as a continuous variable, temporal sets that align with existing deployment structures or other qualitative aspects of time that are associated with victim activity or target access might be more operationally relevant and actionable. All else being equal, it will be much easier for the end user to use the analytic results if existing deployment schedules and geographic boundaries are used. On the other hand, there are times when it makes sense to look at the data from the perspective of the perpetrator. Factors including shift changes, troop rotations, convoy movements, and routes may have appeal if they result in decreased attention, routine or otherwise predictable activities, or a target-rich environment. A good analyst will explore the data, examining it from more than one perspective in an attempt to find the best model.

Just because we have access to the data and can analyze something does not mean that it will be actionable or even have value to the end user. Again, this is not meant to imply that data should not be considered or discarded quickly. Analysis in the applied setting frequently represents a dynamic balance between considering as much information as possible, yet generating output that is both relevant and actionable.

What does this mean and how do we consider everything yet confine the analysis to only that information determined to be relevant and actionable? The key to this process is a good understanding of the problem, knowledge of the operational options being considered or available, and most

importantly, a close working relationship with the end user. Therefore, this chapter covers operationally relevant preprocessing to include recoding and variable selection. This chapter also covers some frequently used data resources that have unique value as well as significant challenges (e.g., telephone and Internet data).

6.1 OPERATIONALLY RELEVANT RECODING

In general, analysts should expect to spend approximately 80% of their time preparing the data and 20% of their time analyzing them.¹ While this sounds like a terribly unattractive prospect, if the data preparation is done well, huge benefits in the overall quality of the analysis can be reaped. Moreover, the analysts will gain additional insight into the data, which can further refine the analysis.

It frequently is advisable to categorize and recode as much continuous information as possible. Even categorical data can be further aggregated into smaller sets or binary data, depending on the requirements and overall objectives of the analysis. For example, motive might be reduced to “drug-related” and “nondrug-related,” or “domestic” and “other.” In data mining, sets or binary data (e.g., yes/no) seem to work well and often are preferable to continuous data. Sets or binary data also tend to increase the likelihood that the generated model will be actionable. Of course, if the data logically are continuous (e.g., pursuit speed), it does not make any sense to recode them into sets this early. But if it is possible to recode continuous data into categories, this can facilitate subsequent exploration and analysis. This issue will be revisited due to its importance to operationally relevant data mining.

This is not to say that the original data should be discarded during the recoding process. Recoding generally should represent the creation of new variables that are derived from the original data rather than replacement of the data. The original data may become particularly useful when the results are being evaluated and validated. For example, it is not unusual for a large number of calls for service to be generated for the local emergency department. The ability to link those calls back to a specific location may be important to accurate interpretation of the findings. Similarly, when data have been derived from unstructured narrative, it may seem reasonable to discard the original information. It is impossible to know, however, how subsequent information or analysis may change the interpretation or meaning of the original information. Like the data mining process, data recoding also may represent an iterative process. As the data are probed and explored, different strategies for recoding and analysis might emerge. Therefore, new variables should be created and the original information retained.

6.2 WHEN, WHERE, WHAT?

Most problems or challenges in public safety and security can be reduced to an analysis of time, space, and the nature of the incident or threat. As always, the emphasis is on operationally meaningful, actionable output, which frequently begins with operationally relevant preprocessing.

6.2.1 Time

There are many ways to begin exploring data for recoding; however, there are some standard techniques that can be used to start the process. Preliminary steps include parsing the data by various temporal measures (e.g., time of day, date, day of week, month, season). For ease of use, time of day can be divided into time blocks or shifts. This is particularly useful when considering deployment issues, as it is much easier to staff from midnight to 0800 h than 0258 h \pm 53 min. Four- to eight-hour time blocks work well for deployment analysis. Personnel generally do not work 4-h shifts, but using a 4-h level of analysis does afford some flexibility, as 8- or 12-h shifts can be overlapped to provide additional coverage for time periods associated with greater anticipated workload. Shorter than 4-h time blocks becomes cumbersome, as the number of time blocks within the day increases and it is unlikely that very brief time blocks will have any value from a deployment standpoint.

One exception to this is periods of time that are associated with specific incidents or anticipated events. For example, juvenile delinquency may spike for the relatively short period between school dismissal and when parents return home from work. In this case, establishing a time-limited deployment strategy around this relatively short yet high-risk time period makes sense. Similarly, as outlined in the example in Chapter 5, it is not unusual to observe transient increases in aggravated assaults associated with the closing of bars and nightclubs, followed by an increase in armed robberies. In this situation, these relatively transient spikes in crime can be related to the movement of a common victim population – bar patrons. Aggravated assaults, brawls, and tussling frequently are associated with the generalized mayhem of bar closings in areas densely populated with nightclubs. As these individuals make their way back to their vehicles, they make good targets for street robberies given the low lighting associated with the time of night and the increased likelihood that the victims' judgment has been impaired from a night of drinking. Therefore, an effective response to these related patterns could include relatively brief, targeted deployment to the specific areas in question. The nightclub area would be addressed first, with an established police presence in anticipation of bar closings and the associated crowd control issues. These same resources could then be flexed to parking lots, side streets, and other areas associated with street robberies of these same patrons. This type of fluid deployment strategy can serve as a functional force multiplier because two seemingly different crime patterns

are linked and addressed with the same resources. Because a common victim population is identified, the same personnel could be used to address two, relatively brief, time-limited challenges that appear to be very different at first glance (aggravated assaults and street robberies). Strategies like these require significant domain expertise and an excellent working relationship with operational personnel to validate the interpretation of the results and associated approach. The use of creative analytical strategies and fluid deployment can optimize public safety and security resource allocation.

Time blocks longer than 8 h often yield diminishing returns, as important fluctuations in activity are diminished with the increased amount of data, which can be referred to as regression toward the mean.² Similarly, it does not make much sense to establish time blocks, even if they are the appropriate length, that do not match the existing or desired times for shift change. For example, in the development of a strategy to reduce random gunfire on New Year's Eve, we found that the majority of the random gunfire occurred during a 4-h period that spanned from 10:00 P.M. on New Year's Eve to 2:00 A.M. on New Year's Day. While it might be attractive from a cost standpoint to craft a 4-h initiative to address this issue, it is not good personnel management to ask the staff assigned to the initiative to come in and work for only 4 h. In that situation, it made sense to expand the time block somewhat to make it more attractive to the folks working that night. If there is not some pressing need to change existing times, it works best if the data are analyzed and the models constructed to reflect existing shift change times.

This scheduling issue can be managed at several points along the analytical process. During data entry and recoding, the analyst should consider what particular time blocks would make sense from a scheduling standpoint. Does the department use 8-h shifts, 12-h shifts, or is there some opportunity for overlap during particularly busy periods throughout the day? The answer to this question will dictate to a certain degree what level of data aggregation more closely reflects the staffing preferences, and therefore be the most easy to interpret and use. This is not to suggest that everything should remain the same because "that is the way it always has been done." This type of thinking can really squander resources, particularly personnel resources. Rather, working within or relatively close to the realistic, real-world parameters significantly increases the value of a model and the likelihood that it will be used.

Recoding specific dates into days of the week is a relatively standard practice. It is important to remember, however, that time of day can be important in the analysis of daily trends. For example, it was puzzling to discover a large number of street robberies on Sundays until the specific times were examined. This analysis almost always revealed that these robberies occurred during the early morning hours and actually reflected a continuation of activity from Saturday night. Seasonal variations also can be important, particularly if they are related

to the migratory patterns of victim populations (e.g., tourists). Other temporal recoding to consider may include school hours and holidays, as the unsupervised time between school dismissal and when parents return home from work can be associated with considerable mischief. Curfew violations and truancy also are associated with unique time periods that might have value if recoded appropriately.

Recoding should be considered an iterative process. As patterns and trends are revealed, different approaches to recoding or emphasis will emerge. For example, it is not unusual to find increases in criminal activity associated with payday. Therefore, recoding the data to reflect paydays could add value to modeling efforts and related operations. Other events including concerts and sporting events also may be related to public safety challenges or issues. Revealing those relationships and creating derived variables that document these events could result in the creation of more accurate or reliable models that will support better operational decisions. The ultimate question to be answered, however, will determine what time parameters are most appropriate.

In one particularly clever example, a local crime analyst was considering a series of bank robberies. This particular series has been notionally illustrated in [Figure 6.1](#). Analysis of the timeline revealed no readily identifiable pattern, which baffled the team. Days of the week, holidays, and even special events were considered but nothing emerged until they included the amount of money taken during the robberies in the analysis ([Table 6.1](#)). When “cash flow” was considered the pattern finally emerged. Larger amounts were associated with a longer period between incidents, while smaller sums were associated with a shorter time period between incidents. Subsequent interview of the suspect after his apprehension confirmed that these bank robberies represented his sole source of income and that he needed to maintain a certain household cash flow to meet his financial obligations. When he was able to obtain a larger amount of money he could delay the next robbery. Conversely, smaller amounts stolen were associated with a need to rob another bank more quickly.

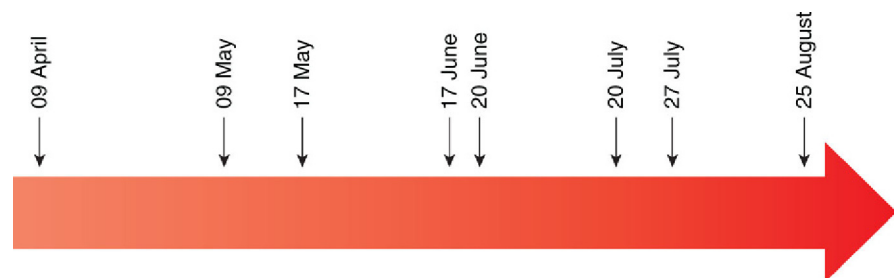


FIGURE 6.1

Timeline illustrating a series of armed bank robberies.

Table 6.1 Notional Bank Robbery Data (Incident Date, Amount Taken, and Days to Next Incident)

Date	Bank	Amount (\$)	Days to Next Incident
09 Apr.	Citizen's National Bank	10k	30
09 May	People's Bank (south)	5k	8
17 May	People's Bank (north)	10k	31
17 Jun.	First National Bank	2k	3
20 Jun.	State Credit Union (north)	10k	30
20 Jul.	Banker's Bank & Trust (north)	5k	7
27 Jul.	First Federal (north)	10k	29
25 Aug.	Bailey Savings & Loan (north)	2k	Arrested

Similarly, segmentation of the medical fraud “timeline” into the time before a bill is submitted for payment, the period of time between submission of the claim and payment, and after payment has been issued (the “pay and chase” model), was used to better understand the fraud lifecycle and also guide relevant approaches to prevention and thwarting.³

6.2.2 Space

The advancement of geographic information system (GIS) capacity has enabled analysts to have access to very precise spatial data. While this information can be extremely helpful for mapping tasks, it may not confer the same benefit to analysis. In fact, it might even hinder analysis if it causes the analyst to focus on specific locations rather than trying to identify general spatial patterns and trends. The criminals involved in many patterns of offending, particularly those involved in serial crimes (e.g., burglary, robbery, rape), generally do not target the same location repeatedly unless they are focusing on a common location for the identification of or access to victims. Rather, they tend to select locations that are similar in nature and/or geography.⁴ These similarities could be as simple as a series of armed robberies in the same general location or as complicated as a group of similar locations that span a broad geographic range. Therefore, a second task during preprocessing is spatial recoding.

Differences in reporting can further confound this issue. For example, a location could be reported as a specific address (401 Main Street), a hundred block (the 400 block of Main Street), an intersection (4th and Main), a specific landmark (in front of the convenience store), or even in terms of its longitude and latitude. Hierarchical organizational strategies that can be expanded or collapsed are quite useful in these situations. For example, 401 Main Street also could be recoded and analyzed as the “400 block of Main Street,” “Main Street,” and a “convenience store.” Additional variables could be created that correspond to various patrol areas, precincts, traffic zones, dispatch regions, or census tracts. Any or all of these can have value if they help to further define

and characterize a trend or pattern, and if they are actionable for the end user. While apparently simple, these decisions may not be easy or initially obvious. For example, census data are rich with information that can be used to further characterize crime trends and patterns. Unfortunately, census boundaries change over time and may not be linked directly to any recognized public safety patrol boundaries. Therefore, census data and associated tract boundaries are limited in their ability to guide very specific deployment strategies or approaches.

Additional spatial attributes to consider include the functional aspects of the location or space, particularly those attributes or features that would confer some tactical or strategic value or advantage to the suspect. Factors to consider include the nature of the space or facility. Is it a park, a single family dwelling, or a multiunit apartment? Is it a business? Is there known, ongoing criminal activity such as an open-air drug market near the location of interest? Is the behavior associated with a common person or event? Are the events or incidents in question associated with occupied or unoccupied dwellings? Considerations also include the identification of spatial features or attributes that could provide cover, concealment, or easy access and/or exit for the criminal. For example, proximity to highway on-ramps may be an important qualitative feature to include as a variable. Banks that are in grocery stores also represent a unique target and tend to show up in series. The use of orthophotography images or other detailed mapping layers can provide additional value to the analysts as they try to reveal attributes.

It is important to remember that locations may not be fixed in space. The exact physical location may have little to no value in the analysis of unusual or suspicious behavior on a moving location like a commuter train or airplane. For example, several of the 9/11 hijackers were seated in first class, which presumably facilitated their access to the cockpit. In this situation, the relative position within the plane is the relevant spatial variable. Similarly, analysis of cargo theft incident data suggests that the highest risk location for in-transit cargo theft falls within 200 miles of the departure point.⁵

Figure 6.2 provides an example that reinforces the concepts in introduced in Chapter 1 regarding the importance of knowing the source, sample size, denominator, and related decision rules regarding collection and use of spatial data in order to support informed analysis. In this particular example, geolocated social media data associated with the Tahrir Square protests were collected, analyzed, and mapped.⁶ The results indicated that the 5272 individual “tweets” included in the sample were associated with 402 unique sources, providing some insight regarding the relative “breadth” of the voice. Additional metrics included information regarding the nature of the source (e.g., mobile as compared to stationary), as well as a calculated distance metric to support

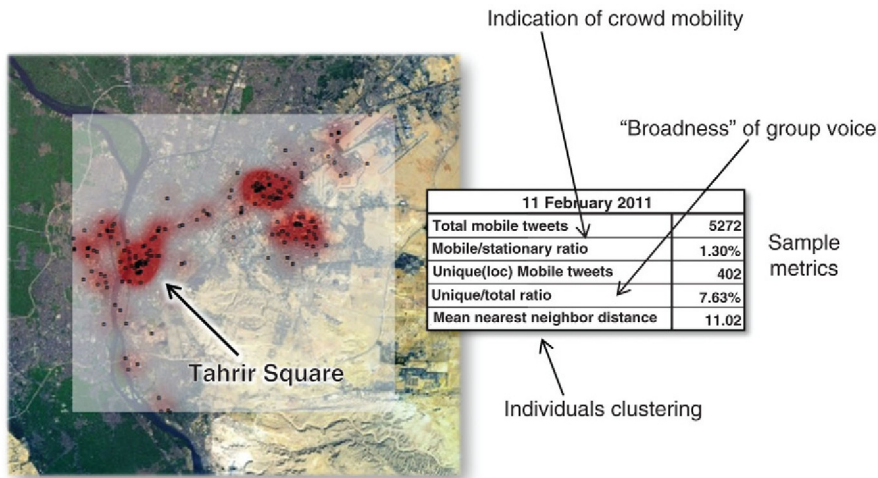


FIGURE 6.2

Illustrates the use of geospatial analysis and summary statistics to visualize and describe social media data associated with the Tahrir Square protests resulting from the resignation of Egyptian President Hosni Mubarak.²⁶ Individual metrics describe the breadth of the voice, and inferred crowd size and density. *DigitalGlobe, used with permission.*

some insight regarding the density of the crowd. Overall, these data were used to characterize the nature and relative activity of the crowd, including formation, distribution, and interaction, as well as the nature and breadth of the voice.

Notable exceptions to the “crime generally does not occur in exactly the same location” rule include open-air drug markets and hostile surveillance. Repeated identification of similar behavior or activity associated with a common location could suggest possible surveillance activity. Terrorists, burglars, and predatory violent offenders may spend considerable time observing possible target locations or individuals. It is unlikely that the suspect in these situations will use the exact same location repeatedly. Rather, suspicious or unusual behavior will be correlated in the same general location or focused on a common area or target, so it is important to identify qualitative aspects of these locations. For example, one person may be observed watching the same critical facility from several different vantage points. In this case, the observation point is important, but the target of the observation can be even more critical to the analysis.

Two people to consider when recoding spatial data are the suspect and the end user. Particularly in threat assessment and surveillance detection, it is important to identify the object of the suspect’s interest. For example, if a suspicious

person was observed filming a fire drill in front of a bank, spatial variables of interest would include not only the location of the suspect but also the target of interest – the bank (the red zone and potential target, respectively). Thinking in terms of the end user, the information regarding the position of the suspect can be used to guide surveillance detection operations. The object of the suspect's interest, however, has significant implications for threat assessment and related activities. Therefore, clarification of the potential target can help refine the analysis further, as well as provide specific guidance for the threat assessment. For example, noting that the suspect was observing the north face of a school is good, but specifying interest in a specific aspect of the building, like the school bus loading zone, provides even more information regarding likely intentions and related operational value.

Always remember, though, to recode into a new or derived variable rather than replace. In some situations, the focus will change over time as the suspect refines his or her approach or plan. Being able to document these changes can be extraordinarily valuable from an operational perspective, particularly if it can be used to highlight increasing specificity of a threat or efficacy of target hardening and deterrence. So be sure to retain the original spatial measures to support subsequent analyses as the behavior or interpretation of the behavior changes.

6.2.3 Nature of the Incident or Threat

The line between the spatial attributes of the incident and the nature of the incident or threat becomes increasingly blurred as the location is examined with progressively more thought to the potential threat. For example, in a review of stranger rapists, we found prior burglaries to be a reliable predictor.⁷ Further analysis revealed that preferential targeting of occupied dwellings was an even better indicator. While this feature could be considered as a spatial qualifier, it also had relevance in the evaluation of the nature of the incident or threat. A link between the behavioral aspects of an incident and the location can be a particularly relevant variable when common behaviors are observed in the same or similar locations. For example, observing someone taking a photograph of a building may not be relevant until it is noted that the same or similar buildings have been photographed multiple times previously.

Some qualitative aspects can be used to document changes or infer escalation. For example, in an analysis of suspicious behavior in and around a critical facility, I categorized the behaviors generally into "photography," which was confined to still photography; "video," which included videotaped observation; "approach," which included probing the perimeter, attempting to gain access, or asking security-related questions of facility personnel; and "suspicious situations," which encompassed everything that did not fit into another

category.⁸ The results of the analysis are covered in more detail in Chapter 14; however, recoding the nature of the activity into four discrete categories facilitated characterization, analysis, and interpretation of the behavior, which would not have been possible if the data remained in their native, unstructured form. Perhaps more importantly, this recoding strategy and analysis revealed a shift in the suspicious behavior over time from still photography to more operationally relevant surveillance, including video and security probes – highlighting this particular facility as worthy of additional evaluation and focus.

By better characterizing the nature of the threat or incident, the operational response can be tailored to address the specific threat. Therefore, another example of the value associated with segmenting the data is illustrated in Figure 6.3. In this particular example, drug-related violence was analyzed in an effort to guide risk-based deployment decisions for a special initiative. One variable that emerged was related to victimology. Victims in the lighter shaded areas were found to me

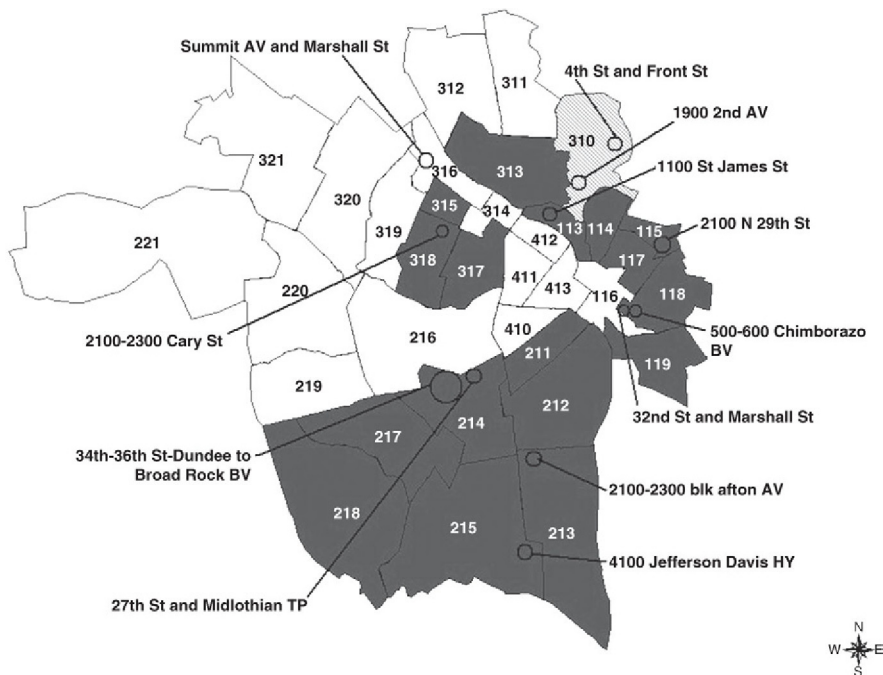


FIGURE 6.3

Map illustrating the results of an analysis of drug-related violent crimes. The victims in the lighter shaded areas were found to have been employed at the time of the incident, while those in the darker shaded areas were listed as unemployed. This information was then used to structure different operational responses specifically targeting different patterns of drug-related violence based on the nature of the incident.

more likely to be employed, while those in the darker areas generally were listed as unemployed. Additional analysis revealed that many of the “unemployed” victims were known drug sellers so they were “employed” but not in an occupation that would be captured by official records. In this particular case, the victim characteristic provided additional insight regarding the nature of the assaults, which was used to guide operations using an approach that parallels “treatment matching” in mental health services. For example, additional review of the incident narratives revealed that many of the “employed” victims were assaulted during a drug transaction. Keeping them out of these relatively dangerous illegal narcotics markets represented the best way to prevent the assaults. Therefore, in this particular situation, demand reduction approaches including “reversals”⁹ were implemented in an effort to make the location unattractive to buyers. The “unemployed” victims, on the other hand, frequently were assaulted as the result of the violence used to enforce rules and norms, and generally regulate behavior in illegal narcotics markets.¹⁰ Approaches aimed at dismantling the drug selling organization and distribution network were developed for those other locations. Again, insight regarding the specific nature of the incident or threat can be used to guide operational strategy and approach.

Recoding offense information to highlight the nature of the incident or threat can be particularly time consuming, as this information is almost always included in the unstructured, narrative portion of the offense report, as well as in supplementary investigative notes and even case briefings. Moreover, it is not always readily apparent what is important, or even what the most relevant recoding strategy might be at the beginning of an analysis. Over time, however, as the data are analyzed and/or additional cases come to the attention of the analyst, a recoding strategy or organization generally will emerge. Most analysts realize the importance of this information, particularly as it relates to MO, and already engage in a certain amount of recoding. Common strategies and themes are addressed in Chapter 10.

6.2.4 [Functional] Interoperability

As noted in the previous chapter, it is frustrating to encounter a database with some essential flaw that significantly limits the ability to use it for analysis.

For example, historic review underscores the consequences of incompatible electrical systems in maintaining service during catastrophic floods in New Orleans.¹¹ Moreover, after action review of the response to the first World Trade Center bombing in 1993 revealed that the police and fire service radios were not interoperable,¹² and many legacy major case response and data management resources in the community cannot be effectively searched or integrated with other common, potentially useful resources.¹³

While efforts have been made to address obvious incompatibility, particularly as relates to technology, many more subtle limitations still exist. For example,

while interoperable radios were distributed to police and fire after the bombing in 1993, many of the new handsets were sitting unused in fire service lockers and police department vehicle trunks because agreements regarding chain of command and who would control the interagency frequencies had not been resolved.¹⁴ Similarly, during Hurricane Katrina, agency software constraints limited the ability to sort data in a way that could enable interagency sharing of information and collaboration; while hospital evacuation lists were prioritized differently by different agencies and responders, which resulted in significant confusion and noteworthy system failures related to evacuation resource allocation responsibilities.¹⁵ Therefore, the new goal relies on the ability to achieve “functional interoperability.” Functional interoperability involves more than just common technology and information access, and can be achieved only when the shared resources are used effectively to support joint decisions and complementary responses. Meaningful data management and analysis, including an emphasis on “operationally actionable” output increases the likelihood that the relevant public safety and security resources can be deployed efficiently and effectively in support of public safety and security objectives.

6.2.5 Other Considerations

The increased use of derived products has raised concerns regarding potential security issues.¹⁶ In addition, the use of location data for analytics,¹⁷ including conflict mapping and crisis response has elevated awareness regarding the potential consequences for privacy, civil liberties, safety, and security, especially for vulnerable populations. Following the adage, “first do no harm,” data sensitivity will be reinforced as a consideration during relevant steps in the analytic process, including the preprocessing and evaluation steps.

6.3 DUPLICATION

As mentioned in Chapter 5, law enforcement and security data frequently contain significant amounts of duplication, which can inflate the numbers for certain types of calls and skew analysis. For example, shots fired in an area can prompt several individuals to call 911. While it can be interesting to speculate about what it means when several people call in one neighborhood while another neighborhood has less citizen involvement, police executives and command staff generally are more interested in what happened. That is, their focus is on how many specific incidents of random gunfire occurred in a particular area at a specific time, not on how many people called in to report the incident. Working within the parameters of the particular data set (How are calls numbered? Are duplicate calls identified and flagged?) will depend on the specifics for each complaint database. The important thing, though, is that the analyst always is aware of the potential for duplication within data and can identify ways to address it if it will skew the analysis or results.

6.4 DATA IMPUTATION

Data imputation is the technical term for “filling in the blanks” when missing data are encountered in a sample. There are multiple methods for data imputation, but they all generally can be described as methods of determining a likely value for missing data based on an analysis of available data. Many data mining software tools and statistical packages include methods and tools for “filling in the blanks” associated with missing data. Like almost anything, though, just because you can do something does not necessarily mean that you should.

Missing data are a frequent occurrence in public safety and security analysis. In some situations, missing data can be important in and of itself. For example, an individual perpetrating fraud might complete every section on a fraudulent credit application in the mistaken belief that this is routine practice for most applicants. The presence of missing data or common skip patterns in these applications could be an important indicator of a valid application. Similarly, some interviewing techniques or statement analysis tools (e.g., Scientific Content ANalysis, or SCAN), specifically look for missing information in a statement or interview as a possible indicator of deception. In those particular situations, the missing data are noteworthy in their absence, and to use data imputation techniques would obscure that finding.

On the other hand, many incidents or events of interest to public safety and security analysts are extremely rare. Missing data can seriously limit or even preclude meaningful analysis of these data, and it would seem that data imputation would represent the only recourse in these situations. Further compounding the challenge associated with infrequent events, however, is the finding that these rare events frequently tend to be heterogeneous. In other words, they may differ between each other in important ways. Data imputation methods are based on the assumption that the available data can be used to identify a reasonable proxy for the missing data. By using data imputation techniques, which fill in missing data based on an analysis of complete records, unusual findings might be magnified or overrepresented. Again, just because a technique is available to the analyst does not mean that it should be used. In some cases it is better to accept the existing limitations of the data, even if this means termination of the analysis, rather than to misdirect resources or otherwise compromise public safety.

6.5 TELEPHONE DATA

Analysis of telephone data can be extremely tedious. However, it is one area in which data mining and predictive analytics can make a huge difference in analytical capacity. By using reverse lookup programs or websites, some value can be added to telephone numbers, even if specific information or identifiers associated with a particular number is unobtainable. For example, in the

absence of specific subscriber information or content, the telephone numbers themselves can be decomposed, aggregated, and recoded to reveal additional information regarding location. While staring at a page full of numeric data is not much fun, manipulation of these data can reveal a considerable amount about relationships, timelines, transactions, and a variety of other information that holds value to public safety and security analysis. Therefore, telephone records comprise an extremely valuable data resource that can be exploited very well through the use of automated methods.

There are several initial recoding steps that can add value to telephone data almost immediately. Many of these can be time intensive, but the major time commitment in data mining frequently involves the initial cleaning and preparation of data. This time investment almost always pays high dividends by increasing both the data quality and our understanding of the data. Through the data preparation process, knowledge is increased, and subsequent analyses frequently become apparent as the data are explored and prepared.

Most analysts know the value of reverse lookup tables and websites. In addition, many analysts also know that telephone numbers can be decomposed into separate components that have value, even if the specific subscriber cannot be identified. For example, a fair amount of information can be obtained from the following telephone numbers:

011-202-633-XXXX
011-201-228-XXXXX

The first number is associated with a telephone number in Cairo, Egypt, while the second number is linked to a cell phone in Egypt. As illustrated in [Table 6.2](#), some generic recoding can be done by breaking the numbers up into their component parts in a database. This will add value to subsequent analyses.

By using reverse lookup tables, information can be added to provide geographic and regional specificity to the data. In some cases, individual subscribers can be linked to a specific number. It is also frequently possible to identify or link a particular service provider to a telephone number, even if the number is unpublished or associated with a cell phone. This service provider information can then be used to generate a subpoena for additional subscriber information, if necessary. Building on [Table 6.2](#), information can be added through recoding, as shown in [Table 6.3](#).

Table 6.2 Example of Preliminary Recoding of Telephone Numbers

International Prefix	Area or Country Code	Region Code	Individual Subscriber
011	202	633	XXXX
011	201	228	XXXXX

Table 6.3 Example of Additional Telephone Number Recoding

Prefix		Country		Region		Subscriber
011	International	202	Egypt	633	Cairo	XXXX
011	International	201	Egypt	228	Mobile	XXXXX

While this can seem tedious, recoding telephone numbers in this fashion adds value to the data that can be used later to identify and cluster seemingly unique numbers based on geography and regions. This can be of tremendous value in and of itself and can be essential if additional, geographically specific data (e.g., shipping records) are added to the analysis.

Additional points to consider include the fact that telephone numbers can vary in length and how the numbers are divided into sections. For example, some international numbers, like the Egyptian mobile phone number listed earlier, do not have seven digits. Additional numbers associated with various extensions or routing within a system can increase the variability in telephone numbers. This information comes at the end of a telephone number and may (but does not always) include the “#” or “*” symbols to indicate the selection of additional extensions. These differences can be accommodated, however, if the analyst is aware of or can anticipate them and includes some flexibility in the data set.

The analysis of telephone data can become complicated further when the target of an investigation utilizes multiple telephone numbers. For example, many people now have cell phones in addition to home telephones. While adding a work or business phone greatly increases the amount of information that needs to be integrated and correlated, many new analytical packages can readily accommodate this. In these cases, the use of date as a common linking variable or key generally represents a good option. Additional information including meetings, delivery schedules, and financial transactions might all add to the complexity of analyzing criminals and related organizations; however, it can greatly enhance the analytical process. Again, the use of some common linking variable or key, such as date, can greatly facilitate analysis and interpretation of the results.

6.6 CONFERENCE CALL EXAMPLE

The next example is based on an actual analysis, but many of the details have been changed to protect the confidentiality of the information and any associated investigations. This example highlights how complicated the analysis of telephone records can become, while highlighting the insight that data mining and predictive analytics can provide. In short, it would not have been possible to analyze these data without the application of data mining and predictive analytics.

Recent advancements in telecommunications have influenced the way that many of us do business, and criminals are no exception. It now is possible to hold a meeting with individuals from the United States, South America, and the Middle East without any of the participants needing to leave their home or office.

In the past, traditional surveillance techniques were used to determine relationships and organizational structure by documenting liaisons and activities. While these techniques will always have value, the same telephone and Internet conferencing techniques that save time and money for businesses also afford a greater degree of anonymity to those wishing to keep their relationships and activities hidden.

The following example is based on real case materials. While many of the details have been changed, the analytical approach and techniques are identical.

The local police department received a 37-page invoice that was associated with a large, unpaid bill (Figure 6.4). The telephone conference call service quickly determined that the information used to establish the account was fraudulent, and they had no additional leads to pursue. Additional information from the company suggested that this series of calls might have been associated with a particular criminal enterprise. Therefore, it was determined that

P.O. Box 923875 Anytown, NE	Invoice #: 837018			
Payment is due upon receipt. Payments not received before September 25 will be subject to a late charge of 3.8% on the outstanding balance.	For billing inquiries please call Customer Service at: 800-555-1212.			
S T A T E M E N T				
Mr. Robert White 2752 North President St Anywhere, NE				
NAME	CONFERNC	NUMBER	DURATION	DATE
BOB WHITE	04428769	201-615-XXXX	6	6/1/2002
BOB WHITE	04428769	201-615-XXXX	27	6/1/2002
MR R WHITE	04364606	201-615-XXXX	63	6/1/2002

FIGURE 6.4

Billing invoice that was associated with a large, unpaid bill.

analysis of these data might provide some clues regarding the identity of the participants, so that the teleconference company might attempt to recover its losses. This case also gave us an opportunity to gain additional insight into this organization and how similar criminal organizations might use and exploit teleconferencing to plan operations and disseminate information.

The first step in the process was to obtain an electronic copy of the bill, which arrived in text format. Unfortunately, this is not always possible. Rekeying data is both tedious and fraught with error, but it is necessary in some cases.

Like most invoices, this bill had a large amount of extraneous information that needed to be removed, including header information and additional text related to the organization's invoice process (Figure 6.5). After this information had been removed, the resulting document included the conference IDs

Easy Dial Conference Calls Page 1 of 37				
STATEMENT				
NAME	CONFERNC	NUMBER	DURATION	DATE
BOB WHITE	04428769	201-615-XXXX	6	6/1/2002
BOB WHITE	04428769	201-615-XXXX	27	6/1/2002
MR B WHITE	04364606	201-615-XXXX	63	6/1/2002
MR COOK	04428769	972-821-XXXX	54	6/1/2002
MR GRAY	04428769	972-821-XXXX	4	6/1/2002
MR GRAY	04428769	972-821-XXXX	25	6/1/2002
MR GREY	04364606	972-821-XXXX	4	6/1/2002
MR GREY	04364606	972-821-XXXX	7	6/1/2002
MR GREY	04364606	972-821-XXXX	213	6/1/2002
MR YELLOW	04364606	312-261-XXXX	5	6/1/2002
MR YELLOW	04364606	312-261-XXXX	6	6/1/2002
MR YELLOW	04364606	312-261-XXXX	12	6/1/2002
MR YELLOW	04364606	312-267-XXXX	204	6/1/2002
MR YELLOW	04428769	312-821-XXXX	5	6/1/2002
MR YELLOW	04428769	312-821-XXXX	82	6/1/2002
BOB WHITE LDR	04429392	201-615-XXXX	8	6/2/2002
MR B WHITE	04364607	201-615-XXXX	61	6/2/2002
MR GRAY	04429392	972-821-XXXX	15	6/2/2002
MR GRAY	04429392	972-821-XXXX	53	6/2/2002
MR GRAY	04429392	972-821-XXXX	152	6/2/2002
MR GREY	04364607	972-821-XXXX	163	6/2/2002
MR YELLOW	04364607	312-261-XXXX	2	6/2/2002
MR YELLOW	04364607	312-261-XXXX	2	6/2/2002
MR YELLOW	04364607	312-261-XXXX	2	6/2/2002
MR YELLOW	04364607	312-261-XXXX	159	6/2/2002
MR YELLOW	04429392	312-267-XXXX	4	6/2/2002
MR YELLOW	04429392	312-267-XXXX	4	6/2/2002

FIGURE 6.5

An electronic version of the invoice was used as the starting point in the creation of a database. This invoice included a large amount of unnecessary information and formatting that had to be removed, including headers, which have been highlighted.

(a unique number assigned by the conference call company), the participants' telephone numbers, the duration of the calls, and the dates. A name was present in less than 5% of the cases. While it was assumed that these names were fraudulent, they were retained in the data because they could be used for additional linking.

This information was then pulled into a statistical package. At this point, the area code was separated from the rest of the information, because area codes can be recoded to unique locations and used for additional linking or aggregating of the information (Figure 6.6).

Initial recoding included linking a location to the area code and telephone prefix (the first three digits of the telephone number). The date was converted into day of the week. Date frequently is important for determining

NAME	CONF	AREA	NUMBER	DURATION	DATE
BOB WHITE	04428769	201	615XXXX	6	06/01/2002
BOB WHITE	04428769	201	615XXXX	27	06/01/2002
MR B WHITE04364606	201	615XXXX	63	06/01/2002	
MR COOK	04428769	972	821XXXX	54	06/01/2002
MR GRAY	04428769	972	821XXXX	4	06/01/2002
MR GRAY	04428769	972	821XXXX	25	06/01/2002
MR GREY	04364606	972	821XXXX	4	06/01/2002
MR GREY	04364606	972	821XXXX	7	06/01/2002
MR GREY	04364606	972	821XXXX	213	06/01/2002
MR YELLOW04364606	312	261XXXX	5	06/01/2002	
MR YELLOW04364606	312	261XXXX	6	06/01/2002	
MR YELLOW04364606	312	261XXXX	12	06/01/2002	
MR YELLOW04364606	312	267XXXX	204	06/01/2002	
MR YELLOW04428769	312	821XXXX	5	06/01/2002	
MR YELLOW04428769	312	821XXXX	82	06/01/2002	
BOB WHITE LDR	04429392	201	615XXXX	8	06/02/2002
MR B WHITE04364607	201	615XXXX	61	06/02/2002	
MR GRAY	04429392	972	821XXXX	15	06/02/2002
MR GRAY	04429392	972	821XXXX	53	06/02/2002
MR GRAY	04429392	972	821XXXX	152	06/02/2002
MR GREY	04364607	972	821XXXX	163	06/02/2002
MR YELLOW04364607	312	261XXXX	2	06/02/2002	
MR YELLOW04364607	312	261XXXX	2	06/02/2002	
MR YELLOW04364607	312	261XXXX	2	06/02/2002	
MR YELLOW04364607	312	261XXXX	159	06/02/2002	
MR YELLOW04429392	312	267XXXX	4	06/02/2002	
MR YELLOW04429392	312	267XXXX	4	06/02/2002	
MR YELLOW04429392	312	267XXXX	5	06/02/2002	
MR YELLOW04429392	312	267XXXX	55	06/02/2002	

FIGURE 6.6

After the headers and other unnecessary information were removed, the data were pulled into a spreadsheet program for additional cleaning and recoding. During this step, the country code or regional area code was separated from the rest of the telephone number to facilitate additional geographic recoding.

NAME	CONF	AREA	LOCATION	NUMBER	DURATION	DATE	DAY
BOB WHITE	04428769	201	Hackensack NJ	615XXXX	6	6/1/2002	SAT
BOB WHITE	04428769	201	Hackensack NJ	615XXXX	27	6/1/2002	SAT
MR B WHITE	04364606	201	Hackensack NJ	615XXXX	63	6/1/2002	SAT
MR COOK	04428769	972	Irving TX	821XXXX	54	6/1/2002	SAT
MR GRAY	04428769	972	Irving TX	821XXXX	4	6/1/2002	SAT
MR GRAY	04428769	972	Irving TX	821XXXX	25	6/1/2002	SAT
MR GREY	04364606	972	Irving TX	821XXXX	4	6/1/2002	SAT
MR GREY	04364606	972	Irving TX	821XXXX	7	6/1/2002	SAT
MR GREY	04364606	972	Irving TX	821XXXX	213	6/1/2002	SAT
MR YELLOW	04364606	312	Chicago	261XXXX	5	6/1/2002	SAT
MR YELLOW	04364606	312	Chicago	261XXXX	6	6/1/2002	SAT
MR YELLOW	04364606	312	Chicago	261XXXX	12	6/1/2002	SAT
MR YELLOW	04364606	312	Chicago	267XXXX	204	6/1/2002	SAT
MR YELLOW	04428769	312	Chicago	821XXXX	5	6/1/2002	SAT
MR YELLOW	04428769	312	Chicago	821XXXX	82	6/1/2002	SAT
BOB WHITE LDR	04429392	201	Hackensack NJ	615XXXX	8	6/2/2002	SUN
MR B WHITE	04364607	201	Hackensack NJ	615XXXX	61	6/2/2002	SUN
MR GRAY	04429392	972	Irving TX	821XXXX	15	6/2/2002	SUN
MR GRAY	04429392	972	Irving TX	821XXXX	53	6/2/2002	SUN
MR GRAY	04429392	972	Irving TX	821XXXX	152	6/2/2002	SUN
MR GREY	04364607	972	Irving TX	821XXXX	163	6/2/2002	SUN
MR YELLOW	04364607	312	Chicago	261XXXX	2	6/2/2002	SUN
MR YELLOW	04364607	312	Chicago	261XXXX	2	6/2/2002	SUN
MR YELLOW	04364607	312	Chicago	261XXXX	2	6/2/2002	SUN
MR YELLOW	04364607	312	Chicago	261XXXX	159	6/2/2002	SUN
MR YELLOW	04429392	312	Chicago	267XXXX	4	6/2/2002	SUN
MR YELLOW	04429392	312	Chicago	267XXXX	4	6/2/2002	SUN
MR YELLOW	04429392	312	Chicago	267XXXX	5	6/2/2002	SUN
MR YELLOW	04429392	312	Chicago	267XXXX	55	6/2/2002	SUN

FIGURE 6.7

An additional variable was recreated, which indicated the country or general geographic area associated with the recorded country and area codes. The date also was converted into day of the week.

timelines and sequencing, while the day of the week can reveal other patterns (Figure 6.7).

An initial review of the data indicated 2017 unique calls or records. A quick visual check of the data, however, suggested that within particular conferences, the same individual might have dialed in more than once. Frequently, one of the calls was much longer while the others were of a minute's duration or less. While this might be meaningful in some way, the most likely explanation was that these individuals had difficulty connecting to or maintaining a connection with the teleconference. Duplication between unique conferences, on the other hand, had value, as it was important in the characterization of particular individuals as well as the various conference calls. Therefore, a decision was made to remove the duplicate calls within a conference while retaining the duplication across conferences.

As can be seen in Figure 6.8, duplicate numbers within a unique conference call were deleted, while duplicated numbers across different conference calls were retained. Culling the duplicates revealed 1042 unique calls. Again, these calls were confined exclusively to those without duplication of the same telephone number within a single conference, while maintaining duplication across conferences.

MR YELLOW	04364606	312	261XXXX	5	06/01/2002
MR YELLOW	04364606	312	261XXXX	6	06/01/2002
MR YELLOW	04364606	312	261XXXX	12	06/01/2002
MR YELLOW	04364606	312	267XXXX	204	06/01/2002
MR YELLOW	04428769	312	821XXXX	5	06/01/2002
MR YELLOW	04428769	312	821XXXX	82	06/01/2002
MR YELLOW	04364607	312	261XXXX	2	06/02/2002
MR YELLOW	04364607	312	261XXXX	2	06/02/2002
MR YELLOW	04364607	312	261XXXX	2	06/02/2002
MR YELLOW	04364607	312	261XXXX	159	06/02/2002



MR YELLOW	04364606	312	267XXXX	204	06/01/2002
MR YELLOW	04428769	312	821XXXX	82	06/01/2002
MR YELLOW	04364607	312	261XXXX	159	06/02/2002

FIGURE 6.8

The data were culled to remove some duplicative calls that were determined to be unnecessary for subsequent analyses.

Finally, the cleaned and recoded data set was analyzed. Using an unsupervised learning process, which is covered in Chapter 7, three groups or clusters of similar calls were identified based on the day of the month that the conference occurred and the number of participants involved in a particular call (Figure 6.9). Further analysis of the participants involved in these calls suggested the possibility that the short calls early in the month involved the key participants or leaders in the process. The gap in activity noted in the middle of the month possibly allowed the organizers of this criminal enterprise to ensure that their activity had not been detected. After it was determined that it was safe to continue, activity resumed later in the month, which escalated to a brisk pace.

The small and medium groups again involved many of the key participants from early in the month, which might have been associated with the relative importance of these calls, their purpose, and some of the other participants. The extremely large conference calls were consistent with the dissemination of information to large groups, such as lectures or fund raising. This activity

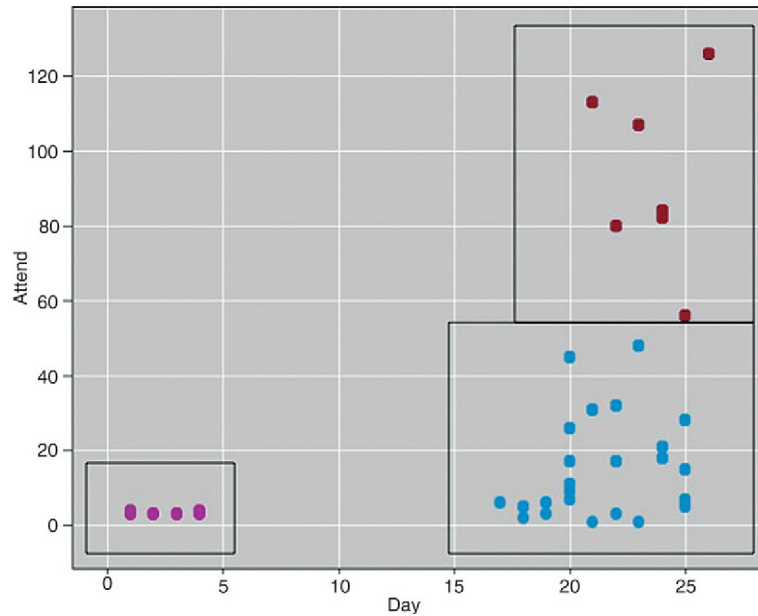


FIGURE 6.9

An unsupervised learning algorithm was used to cluster the calls into similar groups. This resulted in the identification of three distinct clusters of calls, based on the number of participants and the day of the month that the conference occurred. Additional information suggested possible operational differences between the clusters.

continued until the end of the month, at which time it ceased abruptly. Termination of activity at this time probably was preplanned, because the end of the month also represented the end of the billing cycle, which was when this fraud was identified. By terminating activity, access to those involved was also terminated, which significantly limited investigative options.

This example highlights the increased complexity associated with telephone records, as well as the value of recoding information and the fact that quite a bit of valuable information can be elicited from telephone data without access to the actual content of the calls or even specific subscriber information. This information can be modeled and applied to new data sets, which can reveal new information regarding the possible nature of the activity and related participants.

Key points include the importance of recoding telephone numbers in as much detail as possible. In addition, reverse lookup tables can provide information on country, area, and regional coding within telephone numbers, as well as specific subscriber information in some cases. Again, it is not always necessary to identify the specific subscriber. Mining the data to identify regional

geographic specificity can be adequate. This is particularly true in the development of scoring algorithms or classification systems. In those cases, specific subscriber information might limit the identification of a meaningful model that can be applied to new data. If it is essential to identify a specific subscriber, provider information associated with the number can facilitate the information request process.

6.7 INTERNET DATA

Surveillance detection is addressed in Chapter 14, but it is worth mentioning here. Methods of physical surveillance detection are very good; however, large categories of information might be overlooked if surveillance detection is confined exclusively to physical surveillance. Increasingly, terrorists and extremist groups are utilizing Internet resources for preoperational surveillance and information collection. "Correlation" in surveillance detection frequently refers to seeing the same person or vehicle in space or time. Given the interest in technology, it might be time to extend this definition to include correlation between physical surveillance and surveillance activities on the Internet. For example, what happens if physical surveillance has been detected, and vigorous correlated activity is noted on a related website? Data mining tools have the analytical muscle necessary to combine these relatively disparate and unrelated data resources, integrate them, and analyze them in the same environment. By combining web mining tools with analysis of the products of traditional physical surveillance detection, a more complete model of surveillance activity can be developed.

One aid in the task of characterizing and modeling web browsing patterns are "cookies." Briefly, Internet "cookies" are similar to electronic breadcrumbs that we leave behind as we move through the Internet. There are two types of cookies: session cookies and persistent cookies. Session cookies are temporary and only track your movement through a web site during a single visit or session. Persistent cookies, on the other hand, track your movement throughout the Internet. In some ways, persistent cookies have more value to law enforcement because they can help link information from a variety of websites or call out repeat visits to the same site over a longer duration of time. These tend to be intrusive, however, particularly from a privacy standpoint, and many people turn them off as a matter of general principle. Unfortunately, many law enforcement and intelligence agencies do not set cookies on their websites, either for privacy reasons or because they have not thought of it. Not to worry, though; if your agency does not set cookies, it still is possible to analyze and even characterize activity, since this type of behavior tends to be relatively unique yet specific, and very infrequent. These features allow the analyst to putatively link activity based on common Internet Protocol (IP) address or browsing patterns.

Although not perfect, it can represent a viable option for further analysis of suspicious activity.

Web logs contain standard metrics including date, time of day, specific pages visited, and amount of time spent viewing them. In addition, Internet-specific data including IP address, IP address country of origin, search terms and key phrases, and referring page may provide insight regarding the nature and source of the Internet activity associated with a particular page or site (Figure 6.10).

It is important to note, though, that what you see may or may not be real on the Internet. As with most crime and intelligence data, misrepresentation and prevarication frequently are the norm in the Internet. Information including IP addresses can be forged or spoofed,¹⁸ and only a small fraction of the Internet is visible to standard search engines like Google, Bing, and Yahoo and indexed. The so-called “dark” or deep web accounts for the vast majority of the Internet.¹⁹

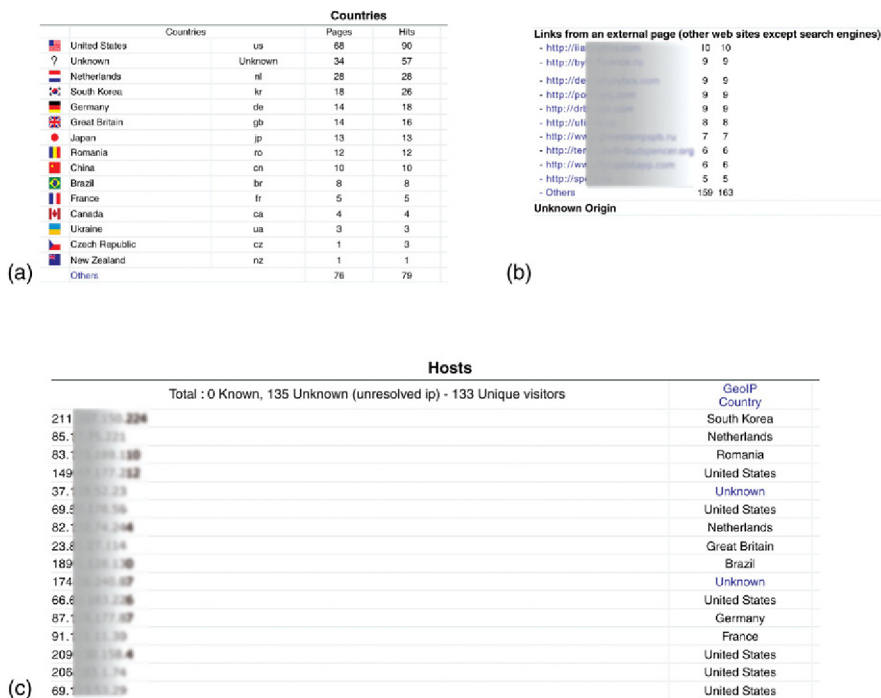


FIGURE 6.10

Illustrates web-mining data to include summary statistics for Internet Protocol (IP) address country of origin (a), referring URL (b), and unique IP addresses associated with activity on a web site (c; screenshot taken by the author).

Related to the topic of clandestine activity on the Internet is The Onion Router (Tor). Originally developed by the Naval Research Laboratory as a means by which to protect government communication. Tor is a distributed, anonymous network that limits the ability to conduct traffic analysis by masking the source and destination of Internet traffic. Tor is used by a variety of groups, including dissidents, journalists, law enforcement and security professionals, and criminals. Much has been written about the use of Tor for illegal activity, including illegal narcotics trafficking, the sex trade, pornography, and insider financial trading just to name a few.

6.8 OPERATIONALLY RELEVANT VARIABLE SELECTION

After recoding, another area for consideration is determining which variables will have value to the investigation, proposed operation, or analysis and should be included in subsequent steps. This is not to suggest that some information should be discarded or excluded, but not every piece of information will have the same amount of value in the analytical process. For example, middle names frequently are collected and entered into law enforcement records management systems. I can honestly state that I have never seen this particular piece of information have any predictive value whatsoever. I certainly would not start a movement to discard this information, because it does have value in terms of identifying unique individuals, particularly those with common first and last names, but it is not something that I would ever consider using in an analysis other than in some sort of link analysis or organizational chart.

Like data mining and analysis in other professions, data quality should be considered first. Issues regarding reliability and validity as well as the frequency of missing data will directly influence confidence in the results and interpretation of the findings. While analysts in the applied public safety and security setting generally need to go with what they have, there will be situations where the data quality issues so significantly limit their ability to effectively analyze the data and trust the results that they must question whether it is even prudent to proceed. These decisions will almost always be situation dependent, but analysts should always exercise caution regarding less than optimal data.

Two additional factors to consider in the selection of operationally relevant variables are determining whether the variables of interest are available and are actionable. Even the most relevant variable has limited value if it is not available when it is needed. For example, studies have shown that significant progress needs to be made quickly on a death investigation if the case is going to be solved.²⁰ Therefore, any information required for a motive determination model should be available quickly if the model is intended to provide investigative support. This is situation dependent, though. The identification of predictive variables that are not available in time for the immediate investigation

could be considered for information-based prevention strategies or cold case investigation, where time is not an issue.

There are other data elements with little or no value in analysis that are not as readily apparent. Many times these elements emerge in a model as a result of errors in logic or as “leaks from the future.”²¹ For example, we found that the suspect’s involvement in substance use and drug selling was a strong predictor of motive in drug-related murders.²² Knowledge of the suspect’s substance use patterns and criminal history would require knowing the suspect’s identity, however. Generally, identification of a specific motive is used to direct the investigation toward a possible suspect, rather than the other way around. Motive becomes important in a homicide investigation for its ability to create a short list of possible suspects. Requiring detailed knowledge of the suspect significantly limits the value of the model by creating a somewhat circular argument (i.e., you try to determine the motive to identify the suspect; if you need suspect information to determine the motive...). In other words, if we knew who did it, then we could just ask them why they did it and would not need advanced analytics to surface likely motive. Therefore, while inclusion of specific details regarding the suspect might result in a highly predictive motive determination model, it would have little value to the investigative process because the model requires specific suspect information to predict the motive, which is being designed to predict likely suspects. In retrospect, the circular logic is obvious, but in the analytical environment, errors in logic such as these are not always obvious. It is always important to keep focused on the ultimate goals of analysis and what is likely to be available for inclusion in the applied setting or operational environment.

The next consideration is whether inclusion of the variable will result in a model that is actionable. Referring back to the use of census data in models predicting crime, we noted that census tract boundaries change and may not match existing patrol boundaries. Therefore, while the information contained in census records might result in more accurate models, if the results cannot be used in the applied setting, they have limited value. It is important to remember that even if the relationships identified are not immediately actionable, they may be considered for other venues. For example, in the analysis of drug-related violence described earlier, we found that victims’ employment status was related to the risk for drug-related violence in one specific location. The victims assaulted in one particular location were more likely to be employed, while in almost every other location studied the victims tended to be unemployed. Although it was possible to identify the victims’ employment status in a timely fashion, it appeared that this variable had limited value for deployment, because it was not clear how resources could be deployed to specifically address the employment status of potential victims. After additional consideration, however, we speculated that the assaults in these areas were related to

robberies of individuals buying drugs. Using this working hypothesis, it was determined that the risk to these victims could be reduced if they did not come into this particular area to buy drugs. The resulting operational strategy then focused on demand reduction in this location in an effort to reduce the risk for this specific group of victims.

Related to this point is the importance of understanding how the model will be used. This is particularly true given the very different requirements associated with analysis that will be used to guide information-based prevention and thwarting, and those approaches designed to enable effective response and investigative support. For example, it has been noted that al Qaeda has demonstrated a preference for multiple, simultaneous yet geographically distinct attacks. While this might have value for retrospectively assigning responsibility or linking attacks, it has extremely limited utility from a prevention perspective because by the time the fact that one is dealing with “multiple, simultaneous yet geographically distinct attacks” has manifest itself, the window of opportunity for prevention or thwarting has closed. On the other hand, knowing that “Bob” likes the convenience of victim-operated improvised explosive devices (VOIED), while “Ed” has a preference for remotely detonated devices so that he can initiate the detonation himself and film the event may have considerable value if a camera crew is spotted on a rooftop or indicators of a pressure plate device are found before the device is triggered. Similarly, “signature” and modus operandi (MO) can be used to link incidents and does have value investigatively. Unfortunately, they generally have limited value in being able to accurately anticipate and prevent events, unless knowledge of likely MO characteristics can be used to quickly surface and thwart an incident or support information-based response plans. So the conundrum is that some variables are valuable some of the time, and very few can be used all of the time. This issue will be revisited in Chapter 8 as approaches to operationally relevant evaluation are considered.

Additional options for variable selection are available. These include stepwise selection approaches, where the user provides an array of possible variables and the most relevant or predictive variables are selected, and user-defined variable inclusion strategies, where the user specifically selects the variables that will be included in the analysis. Each approach has its own strengths and weaknesses, which are covered in Chapter 7.

6.8.1 Challenges

Again, variable selection is not trivial and relies as much on analytical tradecraft as on science. This is particularly true in the applied public safety and security setting, where additional consideration regarding the availability, data quality, and operational relevance regularly come into play. It certainly is possible to create extremely elegant, highly predictive models, but if they require

information that is not readily available, have no relevance to the operational setting, or are so obtuse as to be inactionable, then they have no value. This can be a difficult concept to stay on top of, but the operational personnel will yank the analyst right back into reality with a roll of their eyes and a request for how they should know the suspect's middle name when they have not yet determined the possible motive. Again, "time," "space," and the "nature" of the incident or threat generally work. Beyond those, it depends largely on the data available and the nature of the question.

Ultimately, variable selection is an area in which analysts must rely on their domain expertise to identify and select variables that are both appropriate and relevant for inclusion in the analysis. Because only those variables selected will be considered, the analysts' preconceived notions and biases can play a role in this process. All of the variables selected for inclusion in the analysis, whether appropriate or not, will at least be considered, if not included in any models developed. Going back to the concept of Wicked Problems introduced in Chapter 4, the analyst's "world view" and desired solution tends to frame the problem description and related solution, which includes variable selection.²³ As the researcher Mark Sageman continues, "the empirical world is rarely as tidy as we want it to be...does not easily fit into our analytical categories and requires us to make difficult decisions on the selection of our data."²⁴ This observation is particularly important as we continually watch for and protect against bias in favor of a popular hypothesis, model, or outcome. The variables that we select can have a very definite impact on the analytic outcome so we always need to protect against the introduction of bias. Ultimately, while correlation does not mean causality, inclusion of a particular variable in a predictive model may have implications that extend well beyond the specific analysis. Again, any preconceived notions or biases will be reflected in the variables selected. These are important issues to consider, as criminal justice research is an area of science that has experienced its share of controversy. Throughout history, individuals have used questionable research supported by loosely correlated relationships to confirm theories of criminality and deviance based on prejudice and bias.²⁵ As always, common sense and judgment is an excellent partner to probable cause and ethics in the analytical process.

Bibliography

- 1 Helberg C. Data mining with confidence. 2nd ed. Chicago, IL: SPSS, Inc.; 2002.
- 2 Howell D. Statistical methods for psychology. 3rd ed. Belmont, CA: Duxbury Press; 1992.
- 3 Recommended Requirements for Enhancing Data Quality in Electronic Health Record Systems Final Report prepared for The Office of the National Coordinator for Health Information Technology, US Department of Health and Human Services. June 2007.
- 4 Spatial similarities or criminal place preferences will be addressed in greater detail in Chapter 7, Geospatial Predictive Analysis.

- 5 Burges D. Cargo theft, loss prevention, and supply chain security. Burlington, MA: Butterworth-Heinemann; 2012.
- 6 Hildebrandt W, McCue C. Unbiased analytics for the COCOMs. AHFE; 2012.
- 7 McCue C, Smith GL, Diehl RL, Dabbs DE, McDonough JJ, Ferrara PB. Why DNA databases should include all felons. *Police Chief* 2001;68:94–100.
- 8 McCue C. Lecture presented to Diplomatic Security Service personnel at U.S. Department of State (ArmorGroup, International Training), Rosslyn, VA, May 14, June 25; 2004.
- 9 “Reversals” or a “reverse sting” operation involves undercover operations where investigators pose as dealers in an effort to identify drug buyers and/or reduce illegal narcotics activity in a particular area through demand reduction. Lynn MD. *Practical drug enforcement*. 3rd ed. Boca Raton: CRC Press; 2006.
- 10 Goldstein PJ. The drugs/violence nexus: a tripartite conceptual framework. *J Drug Issues* 1985;15:493–506.
- 11 Fink S. Five days at memorial. New York: Crown; 2013.
- 12 Dwyer J, Flynn K. 102 minutes: the untold story of the fight to survive inside the twin towers. New York: Times Books; 2005.
- 13 Matthews W. FBI struggles with data management. FCW. <http://fcw.com/articles/2002/12/01/fbi-struggles-with-data-management.aspx>; 2002 [accessed 01.12.02].
- 14 Dwyer J, Flynn K. 102 minutes: the untold story of the fight to survive inside the twin towers. New York: Times Books; 2005.
- 15 Fink S. Five days at memorial. New York: Crown; 2013.
- 16 Wells CE. Discussion of the “Mosaic Theory.” In: CIA v. Sims: Mosaic Theory and Government Attitude. *Administrative Law Review*, p. 845–879, as accessed in: University of Missouri School of Law Scholarship Repository, Faculty Publications. <http://scholarship.law.missouri.edu/cgi/viewcontent.cgi?article=1392&context=facpubs>; 2006.
- 17 The Geolocation Privacy and Surveillance Act (H.R.1312: <http://beta.congress.gov/bill/113th-congress/house-bill/1312>) and (S.639: <http://beta.congress.gov/bill/113th-congress/senate-bill/639>); overview at: <http://www.gps.gov/policy/legislation/gps-act/>
- 18 After being the victim of one of the recent high-profile data breaches, I received a series of undeliverable or “bounced” e-mail messages that I had not sent and did not actually originate from my account but that included my e-mail address as sender. Subsequent correspondence from my service provider confirmed that these were spoofed messages that had impacted approximately 2% of their customer base.
- 19 The ultimate guide to the invisible web. Open Education Database (OEDB). <http://oedb.org/librarian/invisible-web/> [accessed 11.11.13]; Bergman MK. White paper: the deep web: surfacing hidden value. *Taking License* 2001;7(1), August. <http://quod.lib.umich.edu/j/jep/3336451.0007.104/--white-paper-the-deep-web-surfacing-hidden-value?rgn=main;view=fulltext>.
- 20 Wellford C, Cronin J. Clearing up homicide clearance rates. *Natl Ins Justice J*; 2000;243:1–7.
- 21 Nisbet R, Elder J, Miner G. *Handbook of statistical analysis & data mining applications*. Boston: Academic Press; 2009.
- 22 McLaughlin CR, Daniel J, Joost TF. The relationship between substance use, drug selling and lethal violence in 25 juvenile murderers. *J Forensic Sci* 2000;45:349–353.
- 23 Rittel H, Webber M. Dilemmas in a general theory of planning. *Policy Sci* 1973; 4:155–169.
- 24 Sageman M. *Understanding terror networks*. Philadelphia: University of Pennsylvania Press; 2004. p. 62.
- 25 See Gould SJ. *The mismeasure of man*. New York: WW Norton & Company; Lewontin RC, Rose S, Kamin LJ. *Not in our genes*. New York: Pantheon Books; 1984.
- 26 Hildebrandt W, McCue C. Unbiased analytics for the COCOMs. AHFE; 2012.

Identification, Characterization, and Modeling

“The most incomprehensible thing about how the world is that it is comprehensible.”

Albert Einstein

Examples throughout the book will underscore the value of effective visualization and summary statistics. Again, approximately 80% of the analyst’s time is spent preparing the data for analysis. The descriptive statistics and exploration associated with the data preparation process provide unique opportunities to gain insight regarding embedded trends, patterns, and relationships in the data that can be used to inform subsequent analysis and modeling, as well as the interpretation and use of the results. The premise behind the use of advanced analytics for crime and intelligence analysis, though, is that behavior, even very bad or extremely aberrant behavior, can be relatively homogeneous and even predictable if viewed in the proper light. Therefore, the goal of predictive analytics in the public safety and security domain is to identify and characterize bad behavior in support of meaningful insight, and effective anticipation and influence. Perhaps even more importantly, predictive analytics move the public safety and security community from counting and reporting, or “chasing” bad behavior, to the use of modeling to better understand it in support of informed anticipation and influence. In other words, advanced analytics offer the opportunity to effectively leverage the fact that behavior tends to be relatively homogeneous and predictable in support of proactive rather than reactive responses.

7.1 PREDICTIVE ANALYTICS

Effectively illustrating the difference between “counting” crimes and analyzing for risk, the example in [Figure 7.1](#) illustrates the results of an analysis designed to model robbery-related aggravated assaults.¹ Going on the premise that if an armed robbery is bad, then getting shot or shot at during the robbery is worse, the goal of the analysis is to identify the factors associated with these riskier robberies in support of information-based operations and response. Similar to

**FIGURE 7.1**

This example illustrates the difference between “counting and reporting” and “risk-based” deployment. Robbery incidents are depicted as dots, while the areas associated with the robbery-related aggravated assaults, the “riskier” robberies, are indicated by the shaded areas.²³

the piracy example described earlier, only 3% of the armed robberies escalated into an aggravated assault. While this is a good thing for the victims, it creates a challenge for the analyst. We could create a model that accurately predicts the outcome 97% of the time by just saying, “No, a robbery will never escalate.” Clearly, that “model” would have extremely limited value.

In [Figure 7.1](#), previous robberies are depicted as “dots” on the map, while the locations identified as being at greater risk for a robbery-related aggravated assault have been shaded. As can be seen in the figure, the “dots” associated with reported robbery incidents do not necessarily align well with the locations identified as being at greater risk for a robbery-related aggravated assault. Therefore, standard approaches to deployment that rely on reported robbery incidents to guide deployment – “putting cops on dots” – would miss these riskier robberies. Moreover, reinforcing the treatment matching concept

introduced in Chapter 6, end users were able to incorporate their tacit knowledge and domain expertise in the interpretation and operational use of these results; identifying three separate and distinct locations associated with three separate and distinct patterns of robbery-related risk. Again, paralleling the example in Chapter 6, one particular location was associated with open-air drug markets where buyers were the most likely victims. A second location was associated with bar and restaurant patrons. Finally, the third area was associated with undocumented workers who did not use organized financial institutions; opting instead to carry their entire net worth with them. The local thugs referred to them as “walking ATMs” and these individuals fought, sometimes to the death in order to protect their cash and other valuables.² Again, the ability to specifically characterize the risk enables the development of strategy and related interventions designed to directly address the unique constellation of risk associated with each different pattern of offending; something that would be missed with a traditional counting and reporting approach to deployment.

The use of advanced analytics to support risk-based deployment will be described in additional detail in Chapter 13. Contrary to the portrayal of advanced analytics in the media, though, this work does not require psychic ability or crystal balls. Rather, it requires the ability to create a statistical model that can be used to meaningfully characterize behavior in support of informed anticipation and influence. Again, there is a very important role for descriptive statistics and visualization, but the ability to move from counting and reporting or “chasing” to modeling bad behavior in support of the insight necessary to effectively anticipate it and influence outcomes is game changing for the public safety and security community.

7.2 HOW TO SELECT A MODELING ALGORITHM, PART I

Although a complete understanding of exactly how these algorithms work is well beyond the scope of this text,³ a general understanding of the broad categories of modeling tools can help the analyst select the proper tool for the job. Just as you do not want to bring a knife to a gunfight, you probably do not want to use a neural net for a deployment model.

Selection should represent a balance between availability and appropriateness of the particular modeling tool. It would be naïve to insist that the algorithm selected should be based exclusively on the best fit for the particular data set and desired outcome, because most agencies do not have unlimited access to modeling tools. However, confining analysis to only what is available *because* it is available is probably just as inappropriate. Ideally, the analyst will let the problem guide the solution and associated technology, but this may not be realistic. Relegating analysis to last place, though, and relying only on what is

inexpensive or readily available is frequently the most expensive “cheap fix” in the public safety community. The best compromise is to anticipate routine tasks and purchase the necessary analytical software to address this work appropriately. The personnel savings associated with the use of data mining and predictive analytics in deployment strategies is documented in Chapters 8 and 13. Similarly, the emergence of regional fusion centers has highlighted not only the enhanced analytical capacity but also the critical need for powerful analytical tools given the increasing complexity of the data, as well as opportunities for cost sharing across jurisdictions and/or agencies.⁴ Therefore, the savings associated with information-based decisions and shared resources can be used to expand analytical capacity.

7.2.1 Supervised versus Unsupervised Learning Techniques⁵

While this is somewhat simplistic, modeling algorithms can be divided generally into supervised and unsupervised learning techniques, based on the availability of known incident or training data and the overall objectives of the analysis.

7.2.1.1 Supervised Learning Algorithms

Briefly, with supervised learning techniques, the goal is to develop a group of decision rules that can be used to determine a known outcome. These also can be called rule induction models, and they include classification and regression models. Supervised learning algorithms can be used to construct decision trees or rule sets, which work by repeatedly subdividing the data into groups based on identified predictor variables, which are related to the selected group membership. In other words, these techniques create a series of decision rules that can be used to separate data into specific, predetermined groups. The use of classification models in automated motive determination is described in Chapter 11. Some modeling algorithms are designed specifically for categorical data, while others can accommodate continuous data. Rule induction models that use continuous data, though, still end up parsing them into categories by identifying breaks, or establishing “cut points,” in the range.

7.2.1.2 Unsupervised Learning Algorithms

Unsupervised learning algorithms are used to group cases based on similar attributes, or naturally occurring trends, patterns, or relationships in the data. These models also are referred to as self-organizing maps. Unsupervised models include clustering techniques and self-organizing maps. Different algorithms use different strategies for dividing data into groups. Some methods are relatively straightforward, quickly dividing the cases into groups based on common attributes or some other similarity. The two-step clustering method differs somewhat in that an optimal number of clusters is

determined in an initial pass through the data, based on certain statistical criteria. Group assignment is then made on a second pass through the data; hence the name “two-step.” Neural networks are more complicated than some of the other unsupervised learning algorithms and can yield results that are relatively opaque and difficult to interpret. An example of an unsupervised learning algorithm includes the analysis of the conference call data outlined in Chapter 6.

7.2.2 Generalizability versus Accuracy

Another important consideration in the selection of a specific modeling tool includes the anticipated or desired use of the results. The topic of accuracy versus generalizability was addressed in Chapter 1; however, it is worth revisiting within this context. Neural networks are truly amazing. That software engineers can even approximate human cognitive processing is a phenomenal achievement in the field of artificial intelligence. The fact that these networks can be used in a PC environment with limited analytical training, albeit abundant domain expertise, was unthinkable even a few years ago. Unfortunately, neural networks have somewhat limited utility in many of the necessary public safety functions because they are relatively opaque. In other words, it is not possible to just look at a neural net and understand the nature of the associations, which significantly limits their applicability in certain tasks. Therefore, in many situations it is important to compromise somewhat on accuracy in an effort to identify an actionable model that can be used in the operational setting. While it is possible to run a scoring algorithm behind the scenes using web-based or remote analytical applications, the balance between accuracy and generalizability frequently guides model selection.

On the other hand, rule sets or decision trees can be relatively intuitive, such as “If X happens, then Y is likely to follow,” “Indicators suggesting overkill generally imply anger or a personal relationship between the victim and perpetrator,” and so on. Even rule sets, however, can become extremely difficult to interpret as the number of variables and options increase and the associated model becomes progressively more complex. These rule sets often need to be relatively transparent to ensure that they will be actionable and have value for the end users.

There are numerous algorithms that can be used, some specific to their associated analytical tool sets or software packages. Those described in the following sections include only a sampling but should represent a good starting point for consideration of specific application, desired outcome, and what is likely to benefit each particular organization. Specific examples of these modeling algorithms are highlighted in other chapters throughout the text.

7.3 EXAMPLES

7.3.1 Link Analysis

Link analysis tools can be used to identify relationships in the data. With a limited number of observations, association matrices and link charts can even be done by hand. As the number of observations increases, though, automated methods usually are required. These tools can be relatively inexpensive and may represent an economical point of entry into data mining. Given this particular benefit, many public safety agencies already use some sort of link analysis tool to analyze their data. There are some limitations to link analysis; however, domain expertise and a good understanding of the concept behind link analysis can help the analyst interpret the results. Some common pitfalls associated with link analysis and their remedies are outlined in Chapter 3.

7.3.2 Neural Networks

With few exceptions, the wonderful complexity that is the human brain is created with two simple elements: neurons and synapses, the connections between neurons. All of our memories, our ability to engage in routine tasks like driving or playing a musical instrument, our capacity to think, and even our sense of humor comes down to unique combinations of neurons or synapses.

As the field of cognitive neuroscience has developed and progressed, scientists have been able to replicate some elements of human cognition. The brain is composed of a relatively small number of common elements. It is the complex arrangement of these fundamental building blocks, the neurons, that achieves the tremendous complexity that we associate with the human brain and cognition. In some ways, the brain can be compared to the “Kevin Bacon” game discussed in Chapter 3. Just as you can connect Kevin Bacon to any other actor with six or fewer links, neuroscientists often brag that they can connect any two locations in the brain with only a few synapses. It is these connections that add the complexity necessary to model complex processes and data. [Figure 7.2](#) depicts a very simple neural network. This particular network includes an input layer with four neurons, a hidden layer with three neurons, and an output layer with two neurons. More complex models could incorporate additional hidden layers, which greatly increase the possible numbers of connections and associated complexity of the model. This ability to layer connections adds tremendous opportunities for additional complexity; again, using only a few common elements. Complexity is achieved through the nature of the relationships and the relative strength of the associations, which can result from repeated use or learning.

We see this repeated throughout nature and even behavior. Computer scientists have been able to replicate certain aspects of neural processing through the development of neural network algorithms. While perhaps simple in their

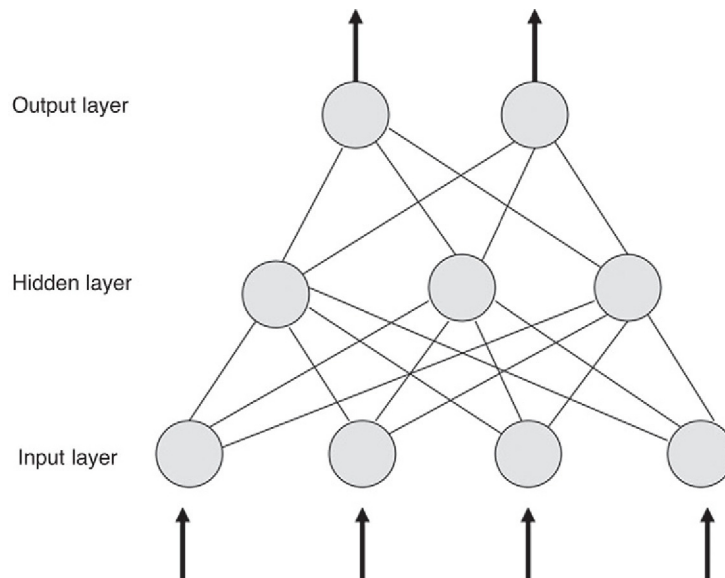


FIGURE 7.2
Simplified neural net model.

basic elements, these sophisticated algorithms can be used to model extremely complex associations and relationships.

7.3.3 Kohonen Network Models

Kohonen network models are a type of neural network. The two unique features associated with Kohonen networks are unsupervised learning and competition. Unsupervised learning models do not create models based on pre-existing groups, clusters, categories, or classification schemes. Rather, these pattern recognition tools seek to identify and characterize the underlying form and natural structure in a given data set, based on the attributes selected for inclusion in the model. In other words, the “correct” output is not known *a priori*, but is determined through the analysis. “Competition” refers to how the structure of the model is determined, which is based on how the human brain learns and is modified by the learning process. Also like the brain, which organizes similar or related functions in distinct and interconnected anatomical locations, Kohonen networks group similar clusters in close proximity and dissimilar clusters at greater distances. Therefore, unlike other pattern recognition algorithms, the relative position of the clusters identified in a Kohonen network have additional value in that clusters that are relatively close share more similarities than those positioned at greater distance on the map. Perhaps for these reasons, the Kohonen network, or self-organizing map, is one of the more popular neural network modeling techniques.

7.3.4 Geospatial Predictive Analytics

The importance of location intelligence and increased use in the commercial sector was introduced in Chapter 3. By identifying and characterizing place preferences, sales and marketing professionals are able to anticipate and influence behavior. Criminals also demonstrate place preferences, particularly as they relate to the availability of and access to victims or targets, as well as the selection of an environment where the individual believes that they will be able to successfully perpetrate their desired acts. As Willie Sutton once said when asked why he robbed banks, because “that’s there the money is.”

Building on the general premise behind predictive analytics as relates to the ability to characterize behavior in support of informed anticipation and influence, geospatial predictive analysis⁶ similarly exploits the fact that behavior is not uniformly or homogeneously distributed across the environment.⁷ Moreover, as introduced in Chapter 5, physical and human geography frequently interact with each other and combine to uniquely define space, which concomitantly may enable, and/or constrain movement and use. Again, offender place preferences generally relates to the distribution of potential victims or target locations, and selection of an environment where they believe that they can be successful. It is important to note, however, that while certain attributes of the environment may directly and obviously meet the offender’s needs, other attributes of the environment may be more nuanced and subtle to the point where the perpetrators themselves may not even be consciously aware of it. These attributes also tend to be offender and offense specific. For example, while proximity to a police station would likely be a deterrent for a bank robber, frequent attacks on police recruiting stations and security barracks in postconflict Iraq demonstrate the concept that the same location may serve as either a repeller or an attractor, depending on the nature of the offense.

Using a supervised learning approach, known locations are statistically characterized in an effort to identify the geospatial attributes that make the training events similar to each other, but different from locations not associated with known events. The resulting geospatial statistical model can then be applied to other locations – even completely new “noncontiguous” locations – in an effort to identify the locations of future incidents. Crime often displaces, in some cases markedly, in response to the increased patrol that is almost invariably used in response to crime. Because it incorporates predictive analytics rather than counting and reporting, geospatial predictive analysis is particularly adept at modeling displacement and identifying hidden patterns of unreported or underreported crime. Specific examples utilizing geospatial predictive analysis as relates to the behavioral analysis of violent crimes and deployment will be presented in Chapters 11 and 13, respectively.

7.3.5 “Artificial Artificial Intelligence”

While there has been exciting development in new and increasingly powerful machine learning algorithms (as well as enabling technologies to include big data management and processing, intuitive user interface, and visualization capabilities), tasks still require a human; so-called “artificial artificial intelligence.”⁸ Examples of these approaches include TomNod⁹ and the Amazon Mechanical Turk,¹⁰ which have developed novel approaches to crowdsourced image analysis and Human Intelligence Tasks (HITs), respectively. Exciting examples that underscore the rapid processing speed and incredible scale associated with these approaches include rapid processing of imagery after natural disasters like tornadoes,¹¹ wildfires, and hurricanes, as well as coordinated searches of large areas.¹² While advances in machine learning continue to be developed, “artificial artificial intelligence” underscores a foundation level assumption that the human will (or should) always remain in the loop. Automated methods may enable a human to effectively manage and process more information at a faster pace, but there is no automated shortcut to meaningful analysis; no free lunches.

7.4 HOW TO SELECT A MODELING ALGORITHM, PART II

With the increased availability of comprehensive data mining suites, there is a dizzying array of modeling algorithms available to the analyst. Even after decisions have been made regarding analytical strategy and the numeric features of the data have been evaluated, there still may be more than one modeling algorithm that would be a good match. It is entirely appropriate to run the data more than one way in an effort to find the analytical approach or algorithm that works best. Again, it is important to remember that we are not hypothesis testing. The most likely error that you will encounter in predictive modeling is that you will not find something.¹³ It is unlikely that one particular modeling technique will emerge as a runaway leader, but subtle differences are not only possible, they are expected. As can be seen in [Figure 7.3](#), the Enterprise Miner platform includes a feature that allows the analyst to run the data using more than one tool and then to compare the results. While this automated approach makes direct comparison relatively easy, it is still possible for the analyst to run the data using several different approaches and expert settings and then compare the outcomes using the strategies for evaluating the results described earlier.

7.4.1 Tools of the Trade and Workflow

Individual modeling algorithms and approaches are illustrated throughout the text, and specific strategies also are suggested in each chapter. The following

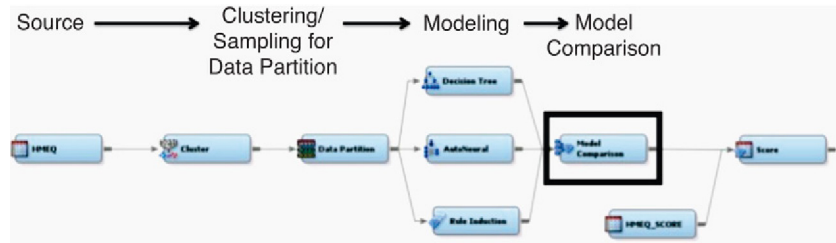


FIGURE 7.3

A model comparison feature that can be used to compare directly different modeling approaches or strategies. Copyright © 2014 SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.

section will go through the general layout and associated workflow of two popular data mining tools: Enterprise Miner (SAS) and IBM SPSS Modeler.¹⁴

Figure 7.4 depicts a sample analytical IBM SPSS Modeler “stream” that was used to analyze the conference call data reviewed in Chapter 6. This example effectively illustrates the role that the data mining platform plays in establishing and following a good analytic process. Figure 7.3 depicts a similar analytical pathway that was generated using the SAS Enterprise Miner™ data mining solution. Moving from left to right in Figure 7.4, the first icon indicates the source of the data. This source node specifies the location of the data to be analyzed in this particular analytical stream. Although not depicted in this example, there are times where it is useful or even necessary to merge data

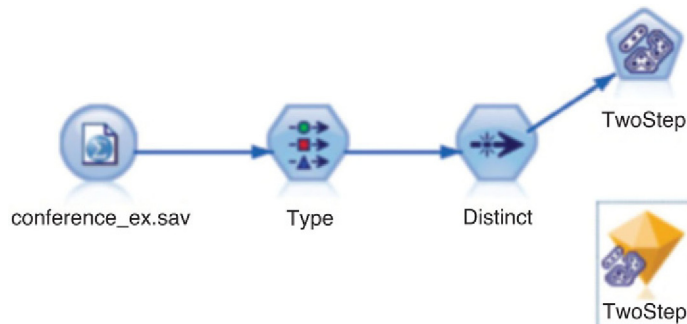


FIGURE 7.4

Sample analytic workflow. Not all analytical packages use the same tools or format for cleaning and recoding data; this illustration has been provided to depict the basic analytical process. IBM® SPSS® Statistics software (“SPSS”). Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

sources, which would require the inclusion of more than one source node. The next node in the IBM SPSS Modeler stream is the “type” node (Figure 7.4), which specifies the nature of the data to be analyzed. Different analytical packages address data specification in different ways, some with greater degrees of flexibility than others. Data definition is important in modeling because certain analyses require specific mathematical properties of the data. The node that follows is a distinct node. As discussed in Chapter 6, there was unnecessary duplication in the data set that would potentially compromise the analysis. Therefore, the data were culled to remove the unnecessary duplication within the data set. The next node in the IBM SPSS Modeler stream is the modeling node (Figure 7.4). In this particular example, a clustering algorithm was used to create groups of similar cases. This also is referred to as an unsupervised learning approach because it does not start with a predetermined outcome or classification system.

The Enterprise Miner workflow illustrated earlier in Figure 7.3 includes a segmentation icon (Data Partition), which randomly assigns the data into samples. In this particular project, this function was used to segment and randomly assign the data into three different modeling algorithms, which were evaluated and compared using the “Model Comparison” feature. The Data Partition function also can be used to create the training and test samples discussed in the next chapter. As illustrated in Figure 7.5, however, new High-Performance (HP) modules have been designed specifically to allow the analyst to realize the full potential of big data and distributed processing without the clustering and sampling required previously. The ability to analyze big data using an in-memory model enables the end user to solve complex problems quickly and efficiently, fully exploiting the benefits of big data.

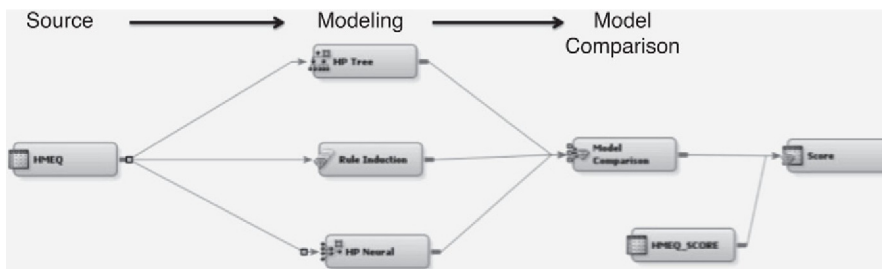


FIGURE 7.5

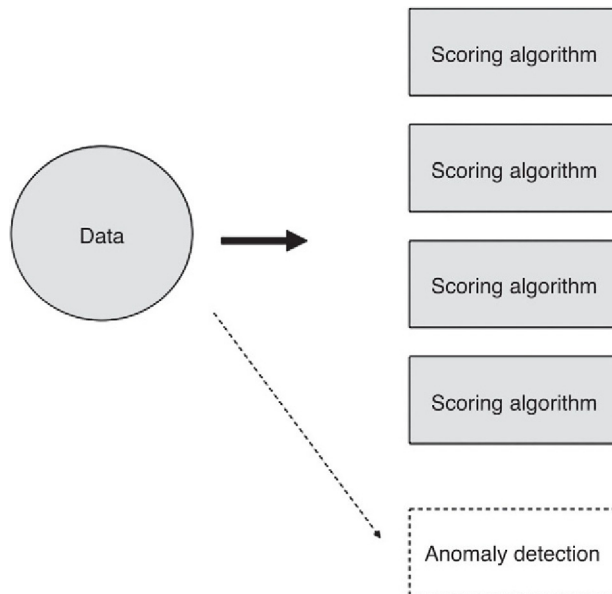
Using “High-Performance” (HP) modules designed specifically to realize the potential of big data and distributed processing, the analyst can now use all of their data quickly and efficiently, running models without the clustering and sampling required previously. *Copyright © 2014 SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.*

7.4.2 Putting it all Together...

In a perfect analytical world, there would be a variety of scoring systems running that would automatically look at information presented (regarding a case, a medical claim, etc.) and analyze that information in terms of its deviation from the normal. In this perfect world, anomaly detectors also would be running quietly in the background, constantly “sniffing” for something unusual that might indicate that trouble is brewing. More and more we realize that criminals study our methods and techniques. It is not unusual to find books and papers outlining police methods and procedures among criminals’ personal effects. In one particular case, a teenage murderer was caught because the investigator was able to elicit a confession by manipulating the suspect’s knowledge of police procedures and forensic techniques. Subsequent search of the suspect’s bedroom revealed several books on serial killers and murder investigation. More recently, evidence has emerged indicating that Al Qaeda operatives have been studying the principles of Fourth-Generation Warfare,¹⁵ while the Iraqis were researching psyops and related topics in the days prior to the most recent Gulf War.¹⁶ How do we accommodate this constantly evolving game of cat and mouse? By building a better mousetrap.

Many scoring algorithms are designed to detect known patterns of criminal offending or unusual behavior. When a new pattern of criminal or suspicious behavior is identified, it can be characterized and modeled. The resulting model can then be used as a scoring algorithm to evaluate each event for signs that indicate similarities with known patterns of criminal or suspicious behavior. If similarities are noted, the incident can be flagged and evaluated further. How likely is it, though, that we can anticipate every possibility for suspicious or criminal behavior? Generally, it is the ones that we do not know about or have not anticipated that catch us each time. Rather than trying to “connect the dots” after something has happened, would it not be better to develop some system that would alert us to patterns of behavior and activities that are unusual or out of the norm, particularly within the public safety setting?

If there is one constant in crime analysis, it is that the creativity of the criminal mind seems unbounded. It is often amazing to see what lengths some individuals will go to in an effort to break the law. Therefore, if detection systems are required to be based exclusively on known patterns of criminal or suspicious behavior, we are always going to be playing the game of catch-up. While it is not likely that all of the possibilities can be anticipated, there is another solution. Running anomaly detection in parallel with traditional scoring algorithms further increases the likelihood that we will identify criminal or suspicious behavior that we do not know about (Figure 7.6). For example, the fraud detection platform RADR uses a combination of supervised learning and anomaly detection in an effort to identify known patterns of fraud, waste, and abuse, while also identifying suspicious or otherwise unusual cases that merit

**FIGURE 7.6**

This figure illustrates the concept of running anomaly detection in parallel with traditional scoring algorithms in an effort to further increase the likelihood that criminal or suspicious behavior will be detected.

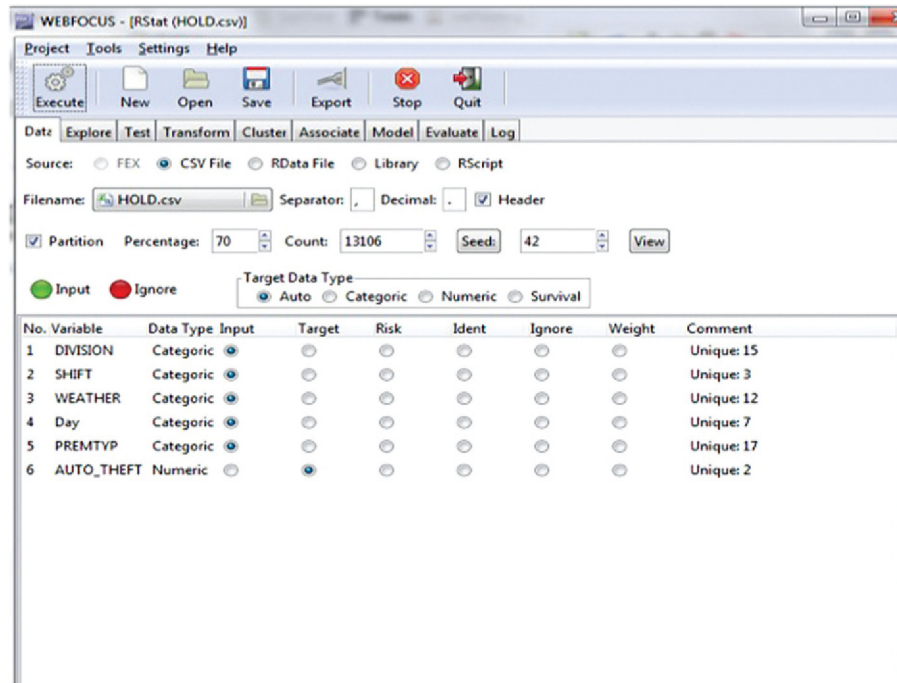
additional review.¹⁷ Clearly, this system will not catch everything, but it does represent an analytical safety net for those patterns of activities or behaviors that we do not know about currently or have not anticipated.

7.5 GENERAL CONSIDERATIONS AND SOME EXPERT OPTIONS

Despite the rapid proliferation of powerful and increasingly intuitive software capabilities, there are no “easy button” solutions or shortcuts. The role of the human in the analysis process continues to be critical and absolutely necessary. To be sure, these new tools can permit new access and insight, while also providing the processing speed and scale for meaningful analysis of very big data – data deemed inaccessible, previously – in real time or near real time. Their primary benefit, though, is to optimize the analyst’s time and talent by surfacing interesting trends, patterns, and relationships in the data that the analyst can further analyze and pursue.

7.5.1 Variable Entry

Some packages will “read” the data and identify preliminary numeric properties. Additional steps include selection of the target variables and the

**FIGURE 7.7**

Sample dialog box illustrating the selection of variables. *Used with permission, Information Builders WebFOCUS RStat.*

identification of which variables should be included for further consideration and modeling. While it might seem foolish to leave anything out,¹⁸ most data sets generally include information that would be inappropriate or irrelevant for modeling (e.g., case ID numbers). Figure 7.7 illustrates the Information Builders WebFOCUS RStat dialog box that allows the analyst to select variables for inclusion in subsequent analyses.

After the initial, “manual” selection process, automated methods can be used to select variables based on their predictive value and relevance to the model. For example, stepwise entry of data allows the inclusion of only those variables contributing to an increase in accuracy, in order of importance. Once model accuracy has been optimized, no additional variables are included. There are other options that allow the analyst to select additional strategies for variable inclusion, but the stepwise entry method is the most common. Following creation of the first model, the analyst then may elect to manually adjust or cull the variables in an effort to minimize unintentional circularity in the model (e.g., “leaks from the future”),¹⁹ or colinearity in the predictive variables, as well as review for operational relevance.

7.5.2 Prior Probabilities

The issue of prior probabilities and its particular relevance in modeling rare or infrequent events was mentioned in Chapter 1. In some modeling packages, the prior probabilities are preset to 50:50 and the analyst must determine the frequency of the target and adjust the prior probabilities manually to accurately model and predict rare events. This generally is not a problem unless the analyst forgets to do it. In other packages, however, the program automatically determines the prior probabilities of the target and accordingly sets the expected probabilities in the model.

7.5.3 Costs

Costs were also mentioned in Chapter 1. As much art as technology, the analyst may manually adjust the costs of certain types of errors and then review the accuracy achieved to determine the best trade-off for the particular analysis. It frequently takes a series of successive approximations before the “sweet spot” is found and the most favorable distribution of errors achieved.

As can be seen from the examples earlier, the new technology greatly facilitates analysis. Moreover, the inclusion of some of the expert options, including the abilities to determine prior probabilities and adjust costs, can help the analyst construct models that directly address some of the unique challenges and needs associated with applied public safety and security analysis. These tools are very likely to improve even further between the time of this writing and the actual publication of this text, as each new release of a data mining tool or predictive analytics suite includes greater functionality and capacity packaged in a more intuitive interface. All that being said, though, domain expertise always will be the essential ingredient in applied public safety and security analysis. A quick review of the Applied Mining and Predictive Analysis model outlined in Chapter 4 underscores the fact that analytical tradecraft and domain expertise are key – all the math in the world cannot create the insight necessary to solve the really hard problems without solid domain expertise.

7.5.4 Combining Algorithms

Not only are data mining and predictive analytics tools becoming increasingly accessible and easy to use, the specific models also can be combined in a number of different ways. For example, standalone technology can be effectively “combined” through brute force use of output from one tool as input to a second tool. The mapped modeling output at the opening to this chapter involved the use of a data mining tool to identify the locations associated with increased likelihood for a robbery-related aggravated assault. These results were then transferred as input to a separate mapping tool as a means by which to visualize them in an operationally relevant and actionable format. Different modeling algorithms also can be used together in the same analytic

Table 7.1 Notional Data Illustrating the Number of Different Ways That a Single Individual May be Represented in a Law-Enforcement Records Management System

URN	Surname	Forename	DOB	Address
TY123456/09	Smith	Billy	12/04/1976	123 High Street
GH734569/11	Smythe	Billy	12/04/1976	123 High Street
TY123456/09	Smith	Billy	04/12/1976	123 High Street
TY123456/09	Smith	Billy	12/04/1976	132 High Street
ZZ765833/10	Smith	Billy	12/04/1976	123 High Street

Differences may include typographic and keystroke errors, common misspelling, errors in file merges, and intentional misrepresentation on the part of the subject. In this particular example, a single individual is associated with five different Unique Record Numbers (URN; Provided by Dr. Rick Adlerly).

workflow and recent advances in data science have resulted in the development of ensemble methods that enable the analyst to gather or strategically “bundle” models as a means by which to leverage the unique benefits associated with each individual algorithm and use them in combination to improve overall accuracy and performance.²⁰

Domain-specific tools also may incorporate multiple, complementary, or interrelated methods in the same platform. For example, Authority Miner[®] effectively combines a suite of related investigative tools that includes an entity resolution tool and an associated social network analysis (SNA) tool.²¹ As discussed throughout this text, crime and intelligence data frequently are less than pristine. One particular challenge is related to duplication of records. As can be seen in [Table 7.1](#), spelling and typographical errors may result in the same individual being associated with multiple records in the same system. The ability to effectively reconcile these records and associate multiple entries with a common entity is essential to accurate and reliable analysis, particularly as related to social network analysis. By using an array of complementary entity resolution capabilities, the tool selects and matches data entries that have a high probability of relating to the same person, place, or thing. This then provides better output for their SNA capability, which identifies and prioritizes relationships and roles for individual offenders in a criminal network ([Figure 7.8](#)).

7.5.5 Refresh Your Models

In the business world, it seems that marketers are always working to further refine their models and improve their predictive value. We should be no exception. Again, there is no such thing as the perfect algorithm or “set it and forget it” in public safety and security modeling. Particularly, in the homeland security and counter terrorism world we know that our adversaries always are

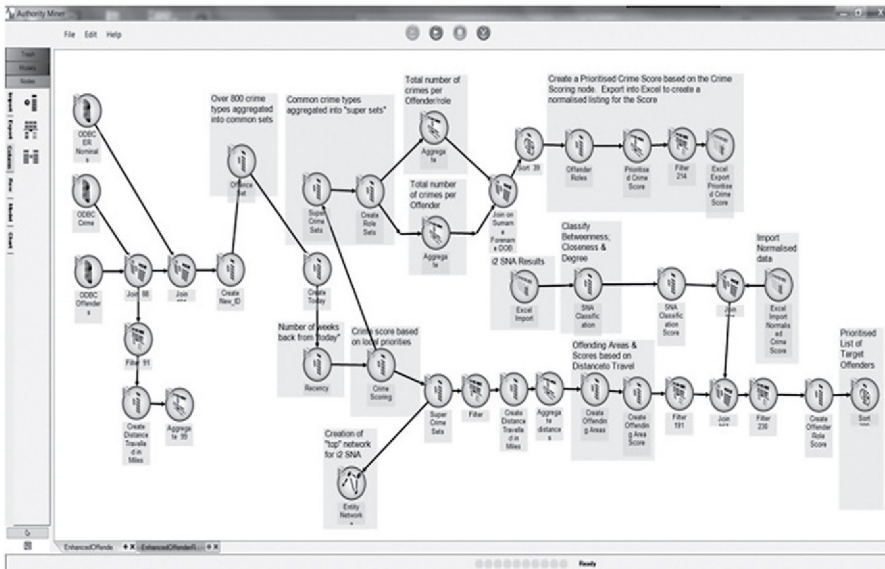


FIGURE 7.8

Sample Authority Miner[®] social network analysis (SNA) output, which identifies and prioritizes relationships and roles for individual offenders in a criminal network.²⁴

looking for a weakness or vulnerability; a better, more efficient way to attack us and increase the damage.

Things change. As discussed in Chapter 6, seasonal changes, weather, changes in access to and distribution of victim populations, and even local events may be associated with changes in local crime patterns and trends. For example, crime patterns and trends were markedly altered after Hurricane Katrina devastated New Orleans.²² Not only were large sections of the city under water, which significantly changed their ability to deploy resources and respond, but the crime trends and patterns also changed in response to shortages of food and water, and general social disorder that emerged in response to the extreme conditions.

One conundrum that the analyst also may encounter is that the more effective their model is in identifying actionable trends and patterns, the more frequently that they may need to refresh it. Whether due to displacement and/or arrest or apprehension of key players, effective enforcement frequently changes trends and patterns by creating an inhospitable environment for crime, or removing key players. In the example used to open the next chapter, displacement of crime in response to aggressive deployment altered the local distribution of crime and the associated “hot spots.” Failure to identify and account for this resulted in a mismatch between crime and law enforcement resources, and created a situation where the local law enforcement agency was “chasing”

crime rather than proactively addressing it. In the marketing world, stale models may result in diminished sales or loss of market share. In applied public safety and security, though, stale models can result in ineffective deployment of scarce resources, as well as more significant consequences including the loss of life.

7.5.6 Warning! Screening versus Diagnostic

In medicine, like many other professions, there is a difference between a screening tool and a diagnostic tool. A screening test highlights possible cases, while a diagnostic test provides confirmation. A screening test is not presumptive evidence of anything other than that further evaluation is warranted. Similarly, in crime and intelligence analysis, anomaly detection should be considered a screening process. While unusual or unlikely events often indicate something more serious, they are not infallible. They pick up a number of other things, including equipment malfunctions, data entry error, and garden-variety outliers. As such, they need to be interpreted with extreme caution until additional information has been collected and evaluated.

Ultimately, it is important to remember that these tools are just math being used to model trends, patterns, relationships, affinities, and perhaps even surface intentions as expressed in the data. With that in mind, domain expertise, knowledge of the operational end user requirements and goals, and a solid understanding of and grounding in context is necessary for accurate interpretation, meaningful insight, and effective use of the results.

Bibliography

- 1 McCue C, Parker A, McNulty PJ, McCoy D. Doing more with less: data mining in police deployment decisions. *Violent Crime Newsletter*, U.S. Department of Justice, Spring; 2004, 4–5.
- 2 The crime analysis capability based on this work received the 2007 Gartner BI Excellence Award; Henschen D. Police department wins Gartner’s 2007 BI Excellence Award. *Information Week Software*. <http://www.informationweek.com/software/business-intelligence/police-department-wins-gartners-2007-bi/198100280>; 2007 [accessed 20.03.07].
- 3 For a more in-depth treatment of this subject, the interested reader is directed to, Nisbet R, Elder J, Miner G. *Handbook of statistical analysis & data mining applications*. Boston: Academic Press; 2009.
- 4 McCue C, Miller L, Lambert S. The Northern Virginia military shooting series: operational validation of geospatial predictive analytics. *Police Chief* February. http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=2871&issue_id=22013; 2013.
- 5 Helberg C. *Data mining with confidence*. 2nd ed. Chicago, IL: SPSS, Inc.; 2002; Two Crows (www.twocrows.com), which is an excellent source of accurate, yet easy to understand information on data mining and predictive analytics. (probably want to replace/add Elder).
- 6 Dalton JR, Porter MD. *Geospatial preference models in signature analyst* (white paper, McLean, VA: SPADAC, Inc.).
- 7 Guidetti R, Morentz JW. Geospatial statistical modeling for intelligence-led policing. *Police Chief* 2010;77(8):72–76, http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=2152&issue_id=82010.

- 8 Mausam PD, Weld DS. Artificial intelligence for artificial intelligence. Association for the Advancement of Artificial Intelligence. http://homes.cs.washington.edu/~mausam/apers/hcomp11b.pdf?_sm_au_=iHV7tFn0RFrsHjwQ; 2011.
- 9 TomNod. <http://tomnod.com/>.
- 10 Amazon Mechanical Turk. <https://www.mturk.com/mturk/>.
- 11 DigitalGlobe. Lending Eyes for Moore Oklahoma. <http://www.digitalglobeblog.com/2013/06/14/moore/>; 2013.
- 12 Martinez M, Newsome J. Crowdsourcing volunteers comb satellite photos for Malaysia Airlines Jet. CNN. <http://www.cnn.com/2014/03/11/us/malaysia-airlines-plane-crowdsourcing-search/index.html>; 2014 [accessed 12.03.14].
- 13 Klecka WR. Discriminant analysis. Quantitative Applications in the Social Sciences; 1980.
- 14 IBM® SPSS® Statistics software (“SPSS”). SPSS Inc. was acquired by IBM in October, 2009.
- 15 Papyrus News. Fourth-generation wars: Bin Laden lieutenant admits to September 11 and explains Al-Qa’ida’s combat doctrine, February 10; <https://maillists.uci.edu/mailman/listinfo/papyrus-news>; 2002.
- 16 McWilliams B. Iraq’s crash course in cyberwar. Wired News, May 22, 2003.
- 17 Elder Research Inc. Elder Research Builds Custom Tool to Help Reduce Fraud, Waste, and Abuse at the Postal Service IG. http://datamininglab.com/images/case-studies/ERI_USPS_OIG_Case_Study.pdf; 2014.
- 18 Nisbet R, Elder J, Miner G. Handbook of statistical analysis & data mining applications. Boston: Academic Press; 2009 [chapter 20].
- 19 Nisbet R, et al. 2009.
- 20 Seni G, Elder J. Ensemble methods in data mining: improving accuracy through combining predictions. In: Grossman R, series editor. Synthesis lectures on data mining and knowledge discovery. Morgan & Claypool; 2010.
- 21 Authority Miner® background material and screenshots provided to the author by Dr. Rick Adderly. For additional information on this capability, c.f., <http://www.a-esolutions.com/index.php?id=authority-miner#.U2ERP9wcWMM>.
- 22 Fink S. Five days at memorial. New York: Crown; 2013.
- 23 McCue C, Parker A, McNulty PJ, McCoy D. Doing more with less: data mining in police deployment decisions. Violent Crime Newsletter, U.S. Department of Justice, Spring; 2004, 4–5.
- 24 Authority Miner® screenshots provided to the author by Dr. Rick Adderly.

Public-Safety-Specific Evaluation

“It ain’t what you don’t know that gets you into trouble. It’s what you know for sure that just ain’t so.”

Mark Twain

Predictive analytics includes confirmation and discovery – confirmation, operationalization, and extension of what we know or *think that we know*, and discovery of new trends, patterns, and relationships. While evaluation may sound simple – either it was there or it was not there, the program worked or it did not – uncertainty, confusion, and bias can dip into almost any part of the process and significantly limit our ability to effectively interpret and evaluate the results.

By way of example, I had a long discussion with a small-town police chief at an airport after a predictive policing conference. This chief came from a larger city and was somewhat disappointed by the relative lack of technology in his new department. At the time of his arrival, the department was struggling with a significant violent crimes problem, with associated unhappiness in the community. The department was working hard, but could not seem to get ahead of the problem and seemed to spend most of their time chasing complaints. They just could not get in front of it and the citizens were angry and tired.

Coming from a larger department that had been using proactive, risk-based deployment, this chief knew that information-based approaches to policing would represent a key first step to turning his department into an analytically competitive organization. With this in mind, one of the first things that he did was to query his command staff regarding the locations associated with the greatest criminal activity – the “hot spots” – in an effort to update his deployment strategy and ensure that they were effectively anticipating crime rather than just reacting. When the responses came back he was impressed by the strong alignment among his command staff. [Figure 8.1](#) represents a notional example of his findings. As can be seen in the illustration, their assessments of the problem locations were remarkably similar with each member of his

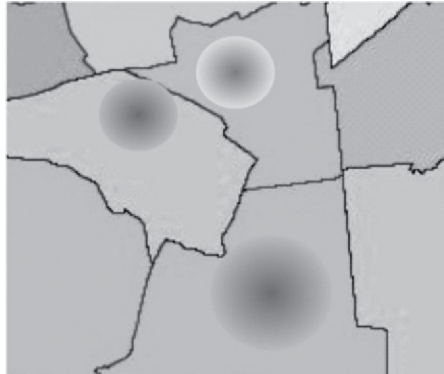


FIGURE 8.1

Notional example of command staff institutional knowledge and “gut instincts” regarding the location of problem locations or “hot spots” within a community.

leadership team picking almost exactly the same locations, which matched their current approach to deployment.

Almost as an afterthought, the new chief decided to review the Calls for Service (CFS) for the previous months. Because the department did not have newer technology, the chief had to request a spreadsheet that listed these complaints sequentially and then create a traditional pin map for the data; manually placing a dot on the location for each citizen complaint. The pattern that emerged stunned the chief. As can be seen in [Figure 8.2](#), the CFS were almost perfectly offset from the locations that the command staff had identified as the problem locations. Perhaps more important, however, resources had been deployed to locations where they thought crime had been occurring, which did not align well at all with the actual complaint data and explained why they were frequently “chasing” crime throughout the community in a largely reactive mode.

Candid discussion with his team revealed that the locations that they had identified for the chief had been associated hotspots historically. Through aggressive deployment, however, they had reduced crime in these locations, effectively displacing it to these new locations. Unfortunately, the institutional knowledge and associated deployment decisions had not been as quick to change. By redeploying his resources to these new areas, however, and continuing to monitor CFS data going forward, the chief was able to make marked reductions in crime, which resulted in a great response from the community and tremendous satisfaction over a job well done among the troops.

One question that comes up with increasing frequency is: How effective is a particular deployment strategy or operational plan at reducing crime? What

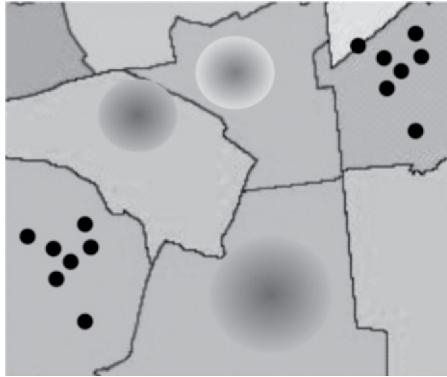


FIGURE 8.2

The identified “hot spots” in relation to citizen Calls for Service (CFS). This mismatch between institutional knowledge and data underscores the “ready, fire, aim” challenge that can be encountered when what we think we “know” proves to be incorrect or misleading.

works? More and more, funding agencies, government organizations, even citizen groups are expecting outcome information demonstrating the efficacy of a particular crime fighting or prevention program. This is particularly true with novel, innovative, or costly approaches.

Evaluating the predictive efficacy of a particular model or scoring algorithm has been addressed throughout this text; if the model cannot be used in the applied setting, it does not matter if it is elegant and highly predictive. On the other hand, if the model is transparent enough to be understood in the operational environment, its only value may be with regard to *post hoc* descriptions of what happened. In other words, it can be used to “connect the dots” only after something happened; it cannot predict and thereby prevent. Similarly, some enhancements are extremely difficult to measure. For example, enhanced investigative efficacy can be as difficult to document as it is to measure.

Outcome evaluation can be an extremely challenging proposition in the public safety arena. In a perfect world, our particular operation or intervention would be the only public-safety-related program being implemented and there would be no other variables outside of our control. We would have a perfectly matched comparison group or location with all factors, except the variables of interest, held constant. In the real world, however, it is very likely that a new deployment strategy will be implemented at the same time as a variety of other crime prevention strategies are being considered and/or implemented. At this very moment, there are multiple crime prevention strategies and programs running simultaneously in almost every community throughout this country, very few of which have been coordinated to ensure that they are complementary or at least not conflicting.

As underscored by the earlier example, very little can be held constant, including crime trends and patterns, criminals, and local citizens. One major event, such as a hurricane, local fluctuations in illegal narcotics markets, or a factory closing, can ruin the most well-planned crime prevention strategy. One example of a completely unexpected event that skewed local crime trends and patterns was the attacks of September 11. This event and the changes related to the subsequent war on terrorism have been associated with wide-ranging effects, including everything from the immediate psychological impact of the event to radical changes in deployment in response to heightened alert status, travel restrictions, and military activations. Entire agencies have been created and added to the public safety community. Even now, ongoing homeland security issues, as well as the periodic changes in the alert status, tax law enforcement organizations.

Similarly, the impact of Hurricane Katrina on public safety in New Orleans has been almost unimaginable. Rampant looting and intermittent violence plagued the days following the hurricane, while the evacuation centers themselves, most notably the Superdome, became mini communities that experienced their own crime trends and patterns. Although extreme, these examples highlight the reality that the only constant in public safety is that everything that can change probably will, and those things that we see as being constants probably also will change, most likely when we least expect it.

8.1 OUTCOME MEASURES

There is no specific response to the question, “how accurate is accurate enough.” It depends. The likely distribution of errors (false positives versus misses) must be considered in light of the potential consequences of an error, and tolerance for opacity in a model versus the need for transparency or interpretability. While model accuracy certainly is important, though, public-safety-specific evaluation goes well beyond traditional model performance metrics. Tools that can be used to directly measure differences in performance between two models were introduced in Chapter 7.

In the drug-related violence example outlined in Chapter 6, requirements for an accurate model varied greatly depending on the nature of the operational requirements. From a prevention perspective, we were able to associate different patterns of drug-related violence with different victim lifestyles and risk factors, in different locations. Specifically, we found that certain locations were associated with violence related to a drug transaction, rather than the systematic, intra/intermarket violence noted in other locations. Knowing this, the best approach to preventing violence under those circumstances was to keep the potential victims from coming into the area and attempting to purchase drugs; so-called demand reduction methods. We used this knowledge to

create risk-based deployment models based on the victim lifestyle information that proved invaluable for selecting specific tactics and strategy. Ultimately, the operational requirements did not demand a high degree of accuracy in the models; however, we did need to be able to deploy the information through maps in order to convey the information visually in support of risk-based deployment. Therefore, the temporal and spatial data were parsed in a relatively gross fashion (police beats, days of the week, and relatively large time blocks), in support of deployment. The strategy (demand reduction) was based on the earlier analysis, which identified the victim characteristics and likely scenario. Extending this work to support automated motive determination as discussed in Chapter 11, on the other hand, did require a much greater degree of accuracy in order to ensure that resources were allocated appropriately and that any error did not misdirect or stall the investigation.

Perhaps most important, though, the time to consider outcome measures is before the plan is implemented; preferably while the operational plan is being developed. Briefly, an outcome measure is something that can be quantified and is expected to change as a direct result of the operational plan or intervention. Reaching back to the work on best practices,¹ the outcome measures should be relatively specific in time, space, and nature of the outcome, and relatively easy to operationally define and assess.

8.1.1 Time

It is important to ensure that the measured change coincides with the implementation of the operation or intervention. For example, in the New Year's Eve initiative² first discussed in Chapter 6, one additional piece of the evaluation was to measure the number of random gunfire complaints in the time period immediately preceding the initiative in an effort to document the specificity of the operational plan. While the comparison between the two time periods indicated marked reductions in random gunfire associated with the initiative, it was important to ensure that this was not merely the result of a generalized decrease in illegal weapons use that began before the operational plan was implemented.

There are some situations in which advance publicity might impact crime rates prior to deployment of the intervention of interest. The Project Exile program in Richmond, Virginia, exploited this to the benefit of the program. Using the reasoning that law enforcement could enhance the initiative by telling the bad guys that they were going to crack down on weapons violations, Project Exile effectively used advertising outlets to enhance the aggressive prosecutorial and law enforcement strategies that formed the core of this program. While this innovative approach resulted in a very successful gun violence reduction program, unintended consequences can be associated with advanced notice highlighting an impending program. Some programs

might be associated with a lag. In other cases, the implementation of a new program might result in the exact opposite of what was predicted, hoped for, and expected.

Project Exile initially was associated with a huge increase in the recovery of illegal firearms, which later slowed to a trickle. In this example, there was confusion about what was the expected and more desirable outcome: a significant increase in the weapon recovery rate, which illustrated directly that illegal firearms were being taken off of the street; or lower recovery rates, which reflected a reduced number of guns on the streets. Consideration of the goals of the program suggested that everything was going according to plan, although many researchers and policy makers were mildly confused. This was because the real goal of the program was to reduce the carry rate of illegal firearms. In many ways, accomplishing this goal involved changing the decision-making process for the criminals. The program was designed to increase the penalty associated with carrying an illegal firearm, so that the criminals would elect to leave them at home rather than face the consequences. It took a little while for this message to trickle down to the streets. Prosecutors knew that the new program was going to be rough for the criminals, but the criminals did not realize it until their colleagues started doing long, hard time in the federal system. Thoughtful analysis and abundant domain expertise frequently can be used to identify and evaluate these possibilities.

8.1.2 Space

It is important to consider where the particular operational strategy has been deployed and to measure the outcome accordingly. Looking at the New Year's Eve initiative, certain areas expected to be associated with an increased prevalence of random gunfire complaints were identified and targeted for heavy deployment. Although the citizen complaints of random gunfire were reduced citywide, it is important to note that the intervention specifically targeted only a portion of the city. If the data only had been analyzed at the aggregate level, it would have been entirely possible that any differences associated with the specific intervention areas would have been lost in the noise of the other areas.

Focusing the analysis on the targeted location can be particularly important if crime displacement is a possibility. Some patterns of criminal offending resemble a bubble under the carpet. When particular areas become inhospitable to crime due to a specific intervention, crime might just move over to the next block. For example, illegal narcotics markets can be extremely fluid. If a particular corner becomes the subject of aggressive enforcement strategies, the dealers are very likely to move over to another area and resume business. Therefore, if the outcome is measured based on aggregate, citywide measures, this type of

displacement could potentially offset any benefits that might otherwise have been realized.

Another way of looking at this issue would be if a wonderful, very effective violence reduction strategy was developed and deployed in Chicago, but statistics for the whole nation were used to evaluate the efficacy of the program. It would be unlikely that an initiative in Chicago would confer a benefit to the entire nation, but this is similar to how many localities approach evaluation at the community level. An intervention is deployed in a specific location, but crime statistics are collected regionally. By using the wrong measures, promising interventions could be abandoned due to a failure in the evaluation, not the program. Therefore, it is important to drill down and evaluate specifically what worked and where.

8.1.3 Nature

Most communities are heterogeneous. There is some variance to how the population is distributed. Generally speaking, there are areas associated with lower crime rates, better schools, and greater affluence. Conversely, other areas might be challenged with open-air drug markets, poverty, unemployment, and elevated school dropout rates. Crime control strategies for these two different localities are going to be very different. Just as it would be inappropriate to schedule a series of aggressive, highly visible approaches including jump-outs or reversals in an area with relatively low crime rates, it would be irresponsible to rely on community meetings and personal safety lectures for crime prevention in higher risk areas. While this makes sense from an operational standpoint, it is important not to sabotage the evaluation by utilizing aggregate crime rates to measure the efficacy of a specific, targeted crime prevention strategy.

8.1.4 Specific Measure

“Overall, serious crime rate down, but arsons, burglaries, homicides up”³

It is very important to select the specific outcome measure with thoughtful consideration. As the quote suggests, the way that we count crime matters. One of the most popular violence prevention outcome measures is homicide rate. While focus on the homicide rate frequently reflects a significant concern over needless loss of life associated with a violent crime problem in a community, it can be a terrible outcome measure. These numbers tend to be relatively low, which is a good thing for the community, but a challenge from an evaluation standpoint. A homicide also can reflect other factors, including access to timely, competent medical care. Aggravated assaults, on the other hand, are more frequent and often represent incomplete or poorly planned homicides. As such, they represent a good proxy for homicides and are a more effective measure of violent crime.

Similar situations can occur with a variety of measures. For example, arrest-based crime reporting can incorrectly make it look as if crime is increasing in response to a particular initiative. For example, aggressive drug enforcement strategies generally are associated with an increased arrest rate for narcotics offenses. Because arrests are used as the measure of crime, an increased arrest rate can suggest that the problem is getting worse. The truth actually might be that the aggressive enforcement strategy is getting drug dealers off of the street and making the community inhospitable to illegal drug markets, which by almost any standard would be a measure of success. The arrest rate also can be a good process measure, as it definitely shows that folks are out there doing something. Unfortunately, arrest rates can create particular challenges when used as outcome measures. This is not necessarily bad, but it is important to understand what might impact this to ensure that the information is interpreted appropriately and within the proper context.

Deeper understanding of the true goals of a particular intervention also can be used to guide the related performance metrics. For example, there has been an increased deployment and use of closed-circuit television (CCTV) to support security and surveillance efforts in areas deemed to be at high risk for crime and terrorist attacks. Perhaps, the most noteworthy application of this model is the “ring of steel” in London, and more recently the venues for the 2014 Olympics in Sochi, Russia, which include massive deployment of CCTV cameras. The potential value of this model was demonstrated during the London bombings in July 2005. Extensive video footage of the terrorists purchasing supplies, engaging in dry runs, and even planting and detonating the devices associated with the 7/7 attacks in London was associated with brisk investigative pace, including rapid identification and apprehension of the suspects.⁴ Similarly, surveillance and bystander video footage was used in 2013 to quickly identify the Boston Marathon bombers.⁵ Unfortunately, they were not able to use the video footage proactively to identify the preoperational planning in support of prevention, thwarting, or consequence management in either of these incidents. So is CCTV in crime prevention a tool, or not? While preoperational planning, to include purchase of the bomb components, rehearsal and dry runs, and even emplacement of the devices was readily apparent in retrospect, the video footage played no role in early identification and prevention or thwarting of the attacks. Therefore, in considering an appropriate metric for CCTV as an effective counterterrorism tool, would it be rapid identification of possible suspects and enhanced investigative efficacy; or prevention, thwarting, consequence management, mitigation, changed outcomes? Metrics matter.

Related to this, scientists at IBM⁶ posed a similar question regarding the best metric for fraud and financial crimes programs: early detection and a timely response, or anticipation, prevention and influence. In the real world, the best

answer probably is both. Ideally, fraud and other patterns of bad behavior will be detected early and prevented or thwarted. In the absence of early intervention, though, the ability to effectively anticipate in support of information-based response and consequence management might be equally important. Again, rigorous evaluation of the outcome will be contingent on selection of the appropriate metric.

Consider how particular measures might change over time. For example, the Project Exile weapon recovery rate rose initially and then fell off as criminals got the message. Similarly, complaint data can change during an intervention. A community experiencing regular drive-by shootings might be somewhat less motivated to report gang-related tagging or graffiti in the area; graffiti might be bothersome, but it pales in comparison to the amount of lead flying through the air each night. As the violent crime rate is addressed, however, and the community becomes reengaged, residents might be more motivated to begin reporting some lesser crimes, which could appear to be an increase in these crimes. Citizens also might be more likely to report random gunfire after an initiative has been established and shown some promise. Prior to deployment of the initiative, there might have been a sense that nothing would change or that there was danger associated with becoming involved. However, as improvements are noticed following the initiative, crime reporting might increase as neighborhoods become revitalized and the residents begin to reengage and participate in the enforcement efforts.

This point is related to a similar issue: All crime is not created equally. How many aggravated assaults equal a homicide? Is an armed robbery equal to a sexual assault? How about a drug-related murder? While these might seem like absurd questions, law enforcement agencies frequently compile and aggregate these numbers and create a composite “violent crime index.” Formerly referred to as “Part I” crimes, these various measures generally are lumped together and used as a generic measure of violent crime in a community. This is unfortunate because combining all this information together increases the likelihood that something important will be obscured.

Many of the crimes frequently included in composite violent crime indices occur with differential frequency. Generally, there are far more aggravated assaults in a community than murders, and far more robberies than sexual assaults. A decrease in a relatively low-frequency crime might be lost when it is considered within the context of all of that additional information. Moreover, these crimes are not equivalent in terms of their impact on a community. Few would argue that homicide is far more serious than an aggravated assault. Why throw them together in a composite violent crime index that weights them equally?

This probably makes sense to most, but an important extension of this issue arises with the use of generic offense categories to evaluate an initiative

targeting a specific pattern of offenses. For example, why create a specific model of drug-related homicide if you are going to base the outcome evaluation on the entire murder rate? It is very rare to develop and deploy a crime prevention strategy that addresses everything, even all the crime within a general category. A particular robbery initiative might target street robberies, but the entire robbery rate traditionally is used to evaluate the efficacy of the initiative. Similarly, an initiative targeting commercial robberies is not likely to affect carjackings, but it is very likely that carjackings will be included in the “robbery” outcome measure.

Another point to consider is whether it is even possible to measure what the program is designed to impact. An interesting question emerged out of the Project Exile work: How do we measure the firearms carry rate? This is a very important question, because the stated goal of Project Exile was to reduce the carry rate of illegal firearms. The illegal carry rate, however, would be very difficult, if not impossible, to measure. As a result, additional proxy measures were selected in an effort to measure the efficacy of the program. The proxy measures included the number of illegal firearms recovered, as well as other measures of gun-related violent crime. While it was not possible to accurately measure the true carry rate of illegal firearms, these other measures turned out to be just as important in terms of quantifying community public safety and were linked intrinsically to the original measure of interest.

So, while it might not be possible to directly measure the outcome of interest, there generally are other indicators linked to the original measure that can be documented in its place. For example, investigative efficacy as a measure is likely to be elusive. Case clearance rates, however, can be documented and used as a reflection of an improved investigative process.⁷ Serious thought to the specific goals of the operational plan and some creativity in the selection of outcome measures can address these challenges, particularly if these decisions are made as part of the operational planning process.

8.1.5 Challenges and Bias

Another factor to consider is the introduction of bias into the selection of outcome metrics. Again, the best approach is to establish metrics at the start of the analysis process when the original question is being framed. This provides some hedge against shifting the outcome and adjusting it to accommodate less than stellar results. In one particular example,⁸ an organization implemented a new approach to deployment. In response to questions regarding the efficacy of the approach, the team cited results from community surveys that revealed that the public perceived an increased police presence in locations implementing the new approach; suggesting favorable public opinion as an ad hoc outcome metric that supported the value of the approach.

In this particular example, increased public perception of police presence was not one of the intended outcomes of the intervention. Rather, the goals of the approach were to use information-based approaches to proactively identify the “when, where, and what” of crime in support of informed deployment decisions to prevent, thwart, mitigate, or respond more effectively. Ideally, proactive deployment will deter crimes from occurring through increased presence. If that does not happen, the second best are decreased response times and a concomitant increase in arrests by predeploying or staging resources in locations where we expect bad things to happen. In a perfect world, the public would be aware of these efforts only through increased efficacy of their public safety resources through crime reductions; not by seeing a cop on every corner. In fact, used properly, risk-based deployment (aka “just-in-time policing”), has been demonstrated to reduce costs by enabling agencies to do more with less by supporting smarter resource allocation and optimization.⁹

Related to this is the importance of watching for bias in implementation of the effort and/or collection of the results. For example, it is not unusual to have a significant amount of attention and fanfare associated with the “new and improved” operational plan. This can include heightened focus on the effort within the organization, as well as attention from the community and even the media. There frequently is a strong desire to ensure that the effort is effective. The organization wants their chief and other members of the command staff to look good, and significant amounts of money, status, and community goodwill might be on the line. With this in mind, it may not be unusual for the troops and other line staff to “pay a little extra attention” to the focus areas, which may influence or otherwise skew the results, whether intentional or not. There may be heightened awareness of and attention to identified hot spots or other focus areas by the law enforcement agency using the new capability that may be associated with increased patrol in those area, whether intentional or self-directed, as well as a desire on the part of the police agency to demonstrate value and look good.

8.1.6 The “Crystal Ball” Method

In some situations, working an active series provides the opportunity to validate model performance by accurately anticipating subsequent incidents. For example, in the Northern Virginia military shooting series,¹⁰ 3 days after our team briefed the initial assessment, a subsequent shooting occurred in a location identified by the model, providing validation for the assessment. In a related example, our team created a model of the Lord’s Resistance Army (LRA) attacks in the spring of 2011.¹¹ Shortly after briefing the work that summer, US and African forces were inserted into the triborder region of Africa¹² – our original area of interest (AOI). This influx of security forces into a region that had been associated with frequent and ongoing activity resulted in a natural test of

the model. Analysis of incident data collected after the initiative in the triborder region revealed an increase in LRA activity in the Central African Republic, which was consistent with displacement into an area previously identified by the model. Additional detail regarding the LRA work and the Virginia military shooting series will be provided in Chapters 11 and 13, respectively.

While being able to accurately “predict” subsequent incidents does provide some statistical validation for the model, there generally is limited satisfaction in being “right” because it generally means that a crime has occurred. While a model that performs well and reliably “predicts” subsequent events might be satisfying from a statistical perspective, the knowledge that bad things are happening does not provide much comfort. Related to this point, even the most accurate models may not translate directly to improved public safety outcomes; being able to reliably predict the location of future incidents does not necessarily mean that we can prevent them. By way of example, our team worked a challenging crime series with a prolific offender and one of the investigators asked me how it was going. The “good news” was that the model was performing exceptionally well; reliably “predicting” the incidents. The “bad news” was that the model was performing exceptionally well; reliably “predicting” the incidents, but the offender was still active. This offender subsequently was apprehended in a location identified by the model, but it underscores the challenge associated with translating predictive models into effective operational plans.

8.1.7 “Proving” the Negative

Reviewing the basic premise behind risk-based deployment, the primary goal is to prevent crime. In the event that crime does occur, then a second goal is to respond more rapidly by prepositioning resources. Unfortunately, in this model the preferred outcome is crime prevention, which requires the ability to effectively identify and document the absence of crime, and then link it back to a specific operation or intervention. For example, in the analysis of the Northern Virginia military shooting series,¹³ the incidents stopped after the original models were disseminated and used to guide deployment. While the break in the series was a very positive development, it was not clear whether the offender stopped because of the increased, targeted deployment or for some other reason. For example, did the offender move out of the area? Was he arrested? Did he die? Or, did the risk-based deployment influence his decision process? In this particular example, the shooter was apprehended 6 months later in a location identified by the model as being at high likelihood for a future incident, but it underscores the challenge associated with evaluating the efficacy of many models in operational public safety and security: how do we show that nothing happened, and how do we link these nonevents to our intervention? While there are no perfect answers, the same creativity and domain expertise

necessary for the use of data mining and predictive analytics in operational planning can be a huge asset during the evaluation process.

8.1.8 Unintended Consequences

Again, the increased use of derived products, including location data, has raised concerns regarding potential privacy and security issues, particularly in conflict mapping and crisis response. Evaluation of potential unintended consequences, particularly as relates to vulnerable populations, also may be considered during this stage.

8.2 THINK BIG

It is unfortunate, but violent crime often is evaluated based on one measure: the homicide rate. Generally, these tend to be low-frequency events compared to other types of violent crime. While this is not a bad thing, it can seriously hamper evaluation efforts. Murders are committed for a variety of reasons, and it is unlikely that a single violence prevention effort will address the entire range of motives. For example, an initiative targeting domestic violence is unlikely to address drug-related violence. Moreover, the provision of skilled medical services in a timely manner can make the difference between an aggravated assault and a murder, and therefore greatly affect the homicide rate.

It is important not to start making broad, general claims based on differences among low-frequency events. For example, while it is tempting to report a 50% reduction in the homicide rate when the numbers drop from four murders to two, you must also be prepared to assume a 100% increase when the numbers change from two to four – again, a difference of two. Clearly, each murder is important, but it can be a very tough measure to use given the mathematical limitations associated with the evaluation of low-frequency events. On the other hand, all assaults can be thought of as incomplete or poorly planned homicides. As such, they are very similar to murders. They also tend to occur with greater frequency and tend to be a better outcome measure.

What does all of this have to do with data mining and predictive analytics? Some things work, others do not, and some things make the problem worse. Those programs that work should be identified and replicated, while those that do not should be modified or discarded. There is no place for programs that make things worse within the public safety arena. Outcome evaluation is critical for identifying programs that exacerbate problems so they can be addressed quickly. But what happens when a particular initiative does all three? [Figure 8.3](#) illustrates just such an intervention. The first panel depicts hypothetical random gunfire complaints in several distinct geographic locations within a particular community. Note that the distribution of complaints varies

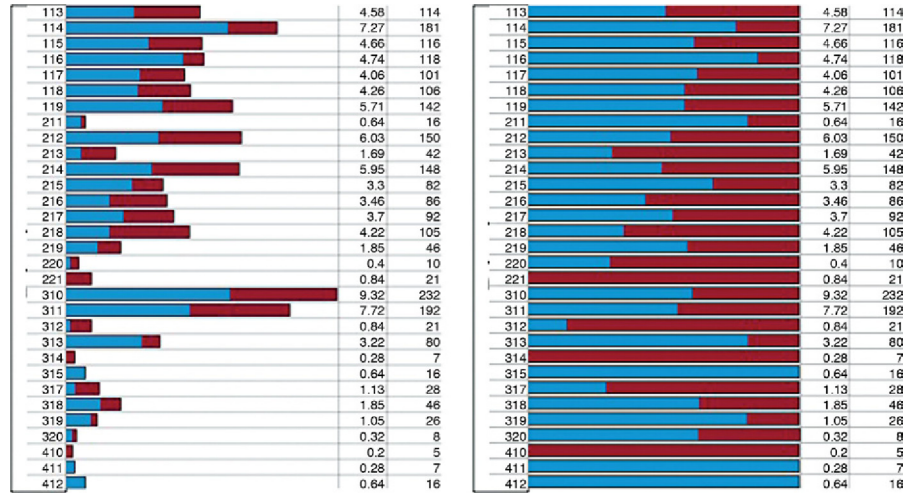


FIGURE 8.3

This figure illustrates hypothetical random gunfire complaints in several distinct geographic locations within a particular community.

greatly among the individual districts. The lighter portion of the bars, which is on the left, indicates the number of complaints during the period immediately preceding the intervention, while the darker bars on the right depict the number of complaints after the intervention has been implemented. The panel on the right depicts exactly the same information as the one on the left, with one exception: The data have been normalized to facilitate comparison of the differences. The raw numbers have been transformed into percentages, which assist review of the intervention despite the differences in overall numbers.

As can be seen, some areas showed improvement after the initiative while others appeared to show an increase in random gunfire. The encouraging finding was that there was an overall reduction – things generally got better after the initiative. But what about areas that got worse? A potential first approach would be to map those areas in an effort to evaluate whether the reduction represented crime displacement. If the areas that showed increases were geographically contiguous with the specific target areas that showed improvement, we might continue to explore this as a possible explanation for our findings. Drilling down in an attempt to further evaluate this hypothesis might be warranted.

On the other hand, if crime displacement does not seem to provide a worthy explanation for the findings, then additional data exploration and evaluation definitely is necessary. Drilling down from the aggregate level statistics is an essential component of this evaluation process. The ratio between random

gunfire complaints pre- and postintervention must be conducted at the specific district level in an effort to gain a complete understanding of how the intervention worked. Confining the analysis to aggregate or overall numbers would give a somewhat skewed perception of the effectiveness of the strategy. By drilling down, it is possible to determine specifically where this particular program worked. Fortunately, in this case, it happened to be in the districts specifically targeted by the intervention. Increases were also noted in other areas, which need to be explored further to address them in future strategies.

ROI

One concept in the business world is “return on investment,” or ROI. In other words, if I spend money upgrading my analytical capacity, what type of return can I expect? In these times of diminishing economic resources, it is difficult to justify big-ticket purchases, particularly when there is no direct and tangible public-safety-related increase. For example, how can a public safety organization justify investing in expensive software resources when the fleet needs maintenance, when there are ongoing training requirements, and when law enforcement professionals are so poorly compensated for their time? Put another way, how many ballistic vests could this same amount of money purchase? These are difficult questions, but they certainly are fair given the ongoing decreases in public safety resources.

ROI is not an easy concept to measure in public safety. For example, how do we measure fear, lost revenues, and lost opportunities in a community inundated with violence? Although many have tried, can we really put a price tag on human lives? How do you measure enhanced investigative efficacy? More and more, public safety agencies are encountering calls for accountability. Communities and funding agencies alike now expect outcome measures and demonstrable effects. Yet increases in public safety can be very difficult to quantify and measure.

Some frequently asked questions about the use of data mining and predictive analytics in law enforcement and intelligence analysis are: How do you know it works? What have you improved? Can you clear cases faster? How many lives have been saved? In response to these questions, specific outcome measures were documented during the Richmond, Virginia, New Year’s Eve initiative¹⁴ discussed previously.

In some ways, identifying the specific outcome measures was relatively easy for this initiative. The deployment strategy was based largely on citizen complaints of random gunfire for the previous year. The primary goal of the initiative was to reduce the number of random gunfire complaints in these locations through the use of targeted deployment. It was anticipated that using heavy deployment in the specific areas previously associated with random gunfire would serve to suppress that activity. Therefore, the number of random gunfire complaints represented one outcome measure. A second expectation of this initiative was that by proactively deploying police units in the locations expected to have an increased prevalence of random gunfire, officers would be able to respond quickly to complaints and make rapid apprehensions. Therefore, a second outcome measure was the number of weapons recovered during the initiative.

Both of these measures documented the success associated with this type of risk-based deployment strategy. Two other benefits also were achieved that night. First, while the original deployment plan called for complete staffing, the risk-based deployment strategy required fewer personnel on the streets. Because personnel resources were used more efficiently, fewer were

needed. This resulted in the release of approximately 50 sworn employees that night and a savings of approximately \$15,000 in personnel costs during the 8 h associated with the initiative. This figure does not include the fact that data mining also was used to confine the initiative to an 8-h time period. Above and beyond the quantified cost savings, the intangible benefit of being able to allow that many sworn personnel the opportunity to take leave on a major holiday was enormous.

Ultimately, data mining and risk-based deployment provided a significant return by several measures. Random gunfire in the community was decreased with fewer personnel resources. This yielded an increase in public safety, at a lower cost, with a concomitant increase in employee satisfaction. By any measure this was an effective outcome.

More recently, the team at Elder Research has demonstrated an impressive ROI in their work with the U.S. Postal Service Office of Inspector General,¹⁵ and Information Builders has documented marked savings in total costs to the community associated with reductions in robberies and larcenies from vehicles achieved by the Charlotte-Mecklenburg Police Department.¹⁶ These examples represent only one approach to documenting the value of advanced analytics in a law enforcement environment. Each locality is different; therefore, specific deployment and evaluation plans will likely differ as well. There are, however, a few elements in these examples that were directly linked to the successful evaluation of the strategies and are worth highlighting. First, the outcome measures could be counted with relative ease. While this sounds very simple, it is not. Think about some of the public-safety-related “measures” that are often tossed around in casual conversations and speeches. For example, fear in a community and investigative efficacy can be very difficult to measure. Even decreases in “crime” can be difficult to define and measure. Identifying an objective, quantifiable outcome measure can have a significant impact on the success of the evaluation, as well as on the initiative being measured. Second, the outcome measures were relatively high-frequency events, which provided greater opportunities for change. As mentioned, homicide rates, although a popular outcome measure, can be extremely unforgiving. They tend to be relatively infrequent, which is a good thing, but which also means that it will take longer to achieve a meaningful difference in the rates. Moreover, many aspects of this measure are completely outside of the control of law enforcement, such as timely medical intervention.

These examples highlight an important issue for analysts, managers, and command staff alike. The ability to incorporate new technology, such as data mining and predictive analytics, into the public safety community will not only be based on a willingness to interact with data and information in novel and creative ways. At some level, the organizations choosing to incorporate these exciting new technologies also will be required to justify their acquisition and use. Being able to proactively identify measurable outcomes and use staged deployment of these powerful tools might be as related to their successful incorporation and implementation as the associated analytical training.

Public safety professionals are told repeatedly that crime prevention is economical. Crime, particularly violent crime, can be extraordinarily expensive when the associated medical costs, pain and suffering, and lost productivity are tallied. Aggressive law enforcement strategies and incarceration can be similarly expensive. One strategy to consider when determining the value of data mining in your organization is the ROI of data-mining-based operational or deployment strategies. Although expensive, data mining software often can pay for itself by preventing even a single firearms-related aggravated assault.

In addition, effective use of investigative or patrol resources can represent a unique approach to the evaluation of a particular strategy or operation. Ultimately, the savings of human lives and suffering is priceless.

As with program results, it is important to evaluate results related to ROI, particularly in the public safety arena. No matter how great the software or deployment plan, if it is not making a difference, then changes need to be made. Radical changes in deployment, large operational plans, and data mining software can be expensive. Justification for the cost associated with these endeavors frequently is, and should be, required. Some things work, others do not, and some things make crime worse. Data mining and predictive analytics can be used to enhance and guide the evaluation process by helping us identify what works – for whom, when, and under what circumstances.

8.3 TRAINING AND TEST SAMPLES

If we work long enough and hard enough, we frequently can generate a model so specific that we get almost complete accuracy when testing it on the original sample; but that is not why we generate models. Ultimately, models are created with the expectation that they will facilitate the accurate classification or prediction of future incidents, subjects, or events. Ideally, the sample data used to create the model truly will be representative of the larger population. What happens, though, if you have something unique or weird in the sample data? Perhaps the sample includes a couple of odd or unusual events that skew the results. When a model is overtrained, it might describe the sample well, but it will not make accurate or reliable future predictions because it is based, at least in part, on the specific features of the sample data, including outliers and other idiosyncrasies. In other words, if a model is modified and refined repeatedly to the point where it has been fit perfectly to the original training sample, then it probably has limited utility beyond that initial sample.

The impact of outliers on the overall results is a particular challenge with smaller samples. When there are a large number of data points, a couple of unusual events or outliers probably will not have a tremendous impact on the outcome. This is similar to baseball; a pitcher can have an off day without significantly impacting his career performance statistics. When the data are confined to a relatively small number of observations, however, anything unusual or out of range can greatly affect the analysis and outcomes. Similarly, a kicker who misses several field goals, even during a single game, can compromise his statistics for the entire season.

One way to address this issue is through the use of training and test samples. If the sample is sufficiently large, it can be divided into two smaller samples: one for the development of the model, the “training” sample, and a second

one that is reserved and used to evaluate the model, the “test” sample. Generally, revising and adjusting a model will continue to increase its accuracy up to a certain point, at which time further modifications to the model result either in no change to the accuracy of the model or actually decreases the accuracy on the test sample. At this point it is wise to stop or back off somewhat in the effort to select the model with the greatest predictive value for the population as a whole, rather than just the training sample. Through the use of training and test samples, a model can be developed, tested, and altered while managing the risk of overtraining the model. When a model works equally well on independent test data, we call this property “generalizability.”

The best way to divide a sample into training and test samples is by using some sort of random selection process. By using random selection, the likelihood that a particular record will be included in either the training or test sample is 50%, or approximately chance. As illustrated in Figure 8.4, some software programs will do this automatically; however, it is possible to split a sample manually as long as there is some assurance that the data have been entered and are selected in a way that maximizes the possibility that the training and test samples will be selected randomly and are as similar as possible. For example,

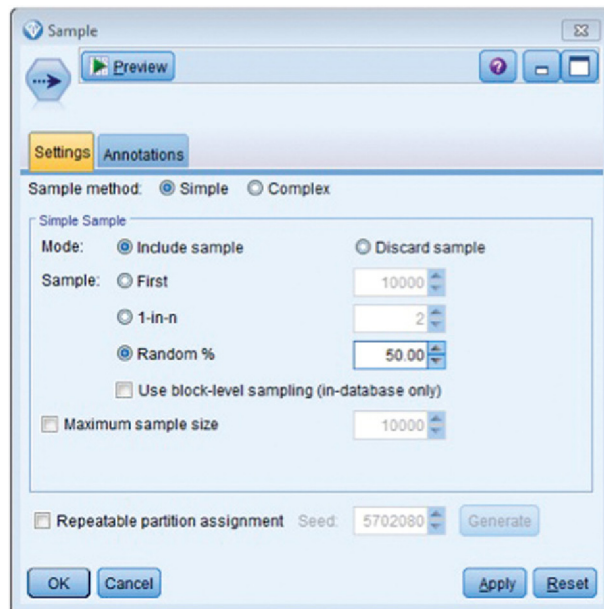


FIGURE 8.4

This dialog box illustrates the random assignment of data into training and test samples. (IBM® SPSS® Statistics software (“SPSS”). Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation. SPSS Inc. was acquired by IBM in October, 2009).

dividing subjects based on the final digit of their social security number, even or odd, generally results in a random assignment to two different groups. Using the first digit, however, would not achieve the goal of random assignment, as the first three digits in a social security number are associated with the location in which the number was issued initially. Therefore, these numbers are not random and could skew the selection of the samples.

Another consideration is when and how frequently to conduct the random assignment to the groups. In a perfect world, it would not matter. The sample could be split each and every time the model is adjusted, modified, and tested because the samples would be similar each time and the model would be robust enough to accommodate any slight differences. In fact, this is how some random assignment software procedures function. In reality, however, it can happen that the samples are comparable, the model development is progressing nicely, and on the next iteration something happens, the training sample differs significantly from the test sample, and everything falls apart. When this occurs, the analyst might be perplexed and goes back and refines the model in an effort to accommodate this new development, but on the next iteration nothing works again. This can go on almost indefinitely. Because of this possibility, it is generally best to split the sample, ensure that they are comparable, particularly on the dimensions of interest, and then move forward into model development and testing with a confidence that the training and test samples are similar, yet randomly divided.

In modeling a particularly low-frequency event, it is important to ensure that the factors of potential interest are distributed evenly. Because they are assigned randomly to the training and test samples, it is possible that the analyst could end up with an uneven distribution that would skew the model based on certain features or attributes that are associated with a few, unique cases. For example, [Table 8.1](#) depicts a small sample of ages that were entered randomly into a database. The data shown are confined to the portion of the sample of interest, or the cases that we would like to model and predict. A quick check reveals that the average age of the entire list of cases is 31 years; however, the average age of the training sample is 24 years, while the average age of the test sample is 39 years. Is this a problem? It depends. If age was included as a discriminating or predictive variable in the model because of these artificial differences in the averages then, yes, this nonrandom selection could have a significant impact on the created model. It could be that age makes no difference, that the comparison sample, which has not been shown here, also has an average age of 31 years. The uneven selection of the training sample, however, generated an average age of 24 years, which might be different enough to be included as a predictive factor in the model. When we come back with the test sample, however, the average age is 39 years. With differences like these, it is relatively easy to see how the performance of the model might suffer when

Table 8.1 A Small Sample of Age Data That Has Been Divided into Training and Test Samples

Age	Sample
25	Train
45	Test
24	Train
38	Test
23	Train
32	Test
21	Train
45	Test
24	Train
35	Test
27	Train
39	Test
24	Train
37	Test

It is important to ensure that training and test samples are similar, particularly when they are relatively small.

tested with a sample that differs so significantly on a relevant variable. Similarly, if an average age of 24 was included in the model as a predictive variable, it would perform very poorly when deployed in an operational setting. This highlights further the importance of using training and test samples during the development process. Therefore, running a few quick comparisons between the training and test samples on variables of interest can be helpful in ensuring that they are as similar as possible and ultimately in protecting the outcomes.

Remember that it is also important to consider the distribution of the outcome of interest between the training and test samples. For example, if the model being created is to predict which robberies are likely to escalate into aggravated assaults, there might be hundreds of robberies available for the generation of the model but only a small fraction that actually escalated into an aggravated assault. In an effort to ensure that the model has access to a sufficient number of cases and that the created model accurately predicts the same or a very similar ratio of incidents, it is important to ensure that these low-frequency events are adequately represented in the modeling process. When working with very small numbers, training and test sample differences of even a few cases can really skew the outcomes, both in terms of predictive factors and predicted probabilities.

Sometimes, the events of interest are so rare that splitting them into training and test samples reduces the number to the point where it seriously compromises the ability to generate a valid model. Techniques, such as boosting, that increase the representation of low-frequency events in the sample, can be used,

but they should be employed and interpreted cautiously, as they also increase the likelihood that unusual or unique attributes will be magnified. It is important to monitor the prior probabilities when boosting is used to ensure that the predicted probabilities reflect the prior, not boosted, probabilities. In addition, other methods of testing the model should be considered, such as bootstrapping, in which each case is tested individually against a model generated using the remainder of the sample. These techniques are beyond the scope of this text, but interested readers can refer to their particular software support for additional information and guidance.

8.4 EVALUATING THE MODEL

While overall accuracy would appear to be the best method for evaluating a model, it tends to be somewhat limited in the applied public safety and security setting given the relative infrequency of the events studied. As mentioned previously, it is possible to attain a high, overall level of accuracy in a model created to predict infrequent events by predicting that they would never happen. Revisiting the piracy example from Chapter 1,¹⁷ we recall that we could create a model that would predict that piracy would never happen and be correct almost 100% of the time; however, we would miss the most important events. Infrequent and rare events represent standard fare for the crime and intelligence analyst.

Moreover, it is often the case that the behavior of interest is not only infrequent but also heterogeneous, because criminals tend to commit crime in slightly different, individual ways. Increasing the fidelity of the models in an effort to overcome the rarity of these events and accurately discriminating the cases or incidents of interest may include accurate measurement and use of the expected or prior probabilities and subtle adjustment of the costs to shift the nature and distribution of errors into an acceptable range. Prior probabilities are used in the modeling process to ensure that the models constructed reflect the relative probability of incidents or cases observed during training, while adjusting the costs can shift the distribution and nature of the errors to better match the overall requirements of the analytical task. The following section details the use of confusion matrices in the evaluation of model accuracy, as well as the nature and distribution of errors.

Confusion matrices were introduced in Chapter 1. A confusion matrix can help determine the specific nature of errors when testing a model. While the overall accuracy of the model has value, it generally should be used only as an initial screening tool. A final decision regarding whether the model is actionable should be postponed until it can be evaluated in light of the specific distribution and nature of the errors within the context of its ultimate use. Under the harsh light of this type of thoughtful review, many “perfect” models have been

Table 8.2 This figure illustrates a typical confidence matrix

		Predicted	
		False	True
Actual	False	33	5
	True	1	14

discarded in favor of additional analysis because the nature of the errors was unacceptable.

[Table 8.2](#) provides an example of a typical confusion matrix. In this example, a total of 53 cases were classified with 47 of them, or almost 89%, being classified correctly. The overall accuracy of this model is good, but it is important to determine the exact nature of the errors. By drilling down and reviewing the specific errors, we see that 33 of 38 “false” predictions, or 87%, were classified correctly. Further review indicates that 14 of 15 “true” predictions, or 93%, also were classified correctly. In public safety and intelligence modeling, these results would be extremely impressive, so much so that additional review probably would be required to ensure that there were no errors in logic that could have contributed to such an accurate model.

The confusion matrix for a second model with the same sample is depicted in [Table 8.3](#). The overall accuracy of this model is much lower, 70%, but there

Table 8.3 The Overall Accuracy of This Model Is Much Lower Than That Seen in [Table 8.2](#), But There Has Been No Degradation in the Accuracy of Predicting “True” Events

		Predicted	
		False	True
Actual	False	23	15
	True	1	14

Most of the decrease in overall accuracy of this model is associated with false positives.

has been no degradation in the accuracy of predicting “true” events, which is still 93%. Most of the errors with this model occur when the model incorrectly predicts that something will happen. In other words, most of the decrease in overall accuracy of the model is associated with false positives.

Clearly, if we had to choose between the two models, the first model with its greater overall accuracy would be the obvious choice. What happens, however, if we are developing a deployment model and the first model is too complicated to be actionable? Can we use the second model with any confidence? It is important to think through the consequences to determine whether this model will suffice.

The second model accurately predicts that something will occur 93% of the time. As a deployment model, this means that by using this model we are likely to have an officer present when needed 93% of the time, which is excellent. As compared to many traditional methods of deployment, which include historical precedence, gut feelings, and citizen demands for increased police presence, this model almost certainly represents a significant increase over chance that our officers will be deployed when and where they are likely to be needed.

Conversely, most of the errors in this model are associated with false positives, which means that the model predicts deployment for a particular time or place where 39% of the time nothing will occur. From a resource management standpoint, this means that those resources were deployed when they were not needed and were wasted. But were they? When dealing with public safety, it is almost always better to err on the side of caution. Moreover, in all reality, it is highly unlikely that the officers deployed to locations or times that were not associated with an increased risk wasted their time. Opportunities for increased citizen contact and self-initiated patrol activities abound. Therefore, although the model was overly generous in terms of resource deployment, it performed better than traditional methods. Thoughtful consideration of the results still supports the use of a model with this pattern of accuracy and errors for deployment.

In the next confusion matrix (Table 8.4), we see an overall classification accuracy of 78%. While this is not bad for a model that will be used in policing, a potential problem occurs when we examine the “true” predictions. In this example, the “true” classifications are accurate only about 50% of the time. In other words, we could attain the same level of accuracy by flipping a coin; the model performs no better than chance in predicting true occurrences. If this performance is so low, then why is the overall level of accuracy relatively high, particularly in comparison? Like many examples in law enforcement and intelligence analysis, we are trying to predict relatively low-frequency events. Because they are so infrequent, the diminished accuracy associated with predicting these events contributes less to the overall accuracy of the model.

Table 8.4 The Overall Classification Accuracy of This Model Is Not Terrible; However, the “True” Classifications Are Accurate Only About 50% of the Time

		Predicted	
		False	True
Actual	False	148	38
	True	5	5

This poor performance does not significantly degrade the overall accuracy because it is associated with a low-frequency event.

As can be seen in the confusion matrix in [Table 8.5](#), which is associated with a revised model for the previous example, the overall accuracy has not been increased substantially, but the “true” predictions or classifications have been improved to an accuracy of 67%. Additional thought will be required to determine whether this level of accuracy is acceptable. For deployment purposes, almost anything above chance has the potential to increase public safety; however, it must be determined whether failing to place resources 1/3 of the time when they are likely to be needed is acceptable. On the other hand, if this model was associated with automated motive determination or some sort of relatively inconsequential anomaly detection, then the comparatively low false positive rate might be of some benefit. A good understanding of the limitations of this model in determining actual events is imperative to ensuring that it is used properly.

Table 8.5 This Confidence Matrix Depicts a Revised Model for [Table 8.2](#)

		Predicted	
		False	True
Actual	False	151	33
	True	3	6

The overall accuracy has not been increased substantially but the “true” predictions or classifications have been improved.

8.5 UPDATING OR REFRESHING THE MODEL

In our experience, models need to be refreshed on a relatively regular basis. In some ways, this is a measure of success. Many features, particularly behavioral characteristics, are unique to particular offenders. As these suspects are identified and apprehended, new criminals take their place, and the patterns and trends change. In most situations, the changes are minimal, although they still frequently warrant a revised model.

Seasonal changes can have an effect on some patterns of offending. As mentioned, during the colder months, people often heat their vehicles in the mornings before they leave the house. Therefore, lower temperatures would be associated with an increased prevalence of vehicles stolen from residential areas, with an increased frequency in the mornings when both the weather and vehicles are colder. When temperatures climb, people often leave their vehicles running while they make quick stops, in an effort to keep their cars cool. A model tracking these incidents might predict an increase in motor vehicle thefts near convenience stores and day care centers, possibly later in the day, when temperatures are higher. Therefore, a motor vehicle theft model that takes into account these trends might need to be refreshed or rotated seasonally. It also might require a greater sampling frame, or time span of data, to ensure that all of the trends and patterns have been captured.

8.6 THERE ARE NO FREE LUNCHES

It is relatively easy to develop a model that describes existing data very well but that is overtrained and therefore relatively inaccurate with new data. The model has been refined to the point where it describes all of the little flaws and idiosyncrasies associated with that particular sample. It is highly predictive of the training data, but when checked against an independent sample of test data, the flaws are revealed through compromised accuracy. This is called “overfitting” the model. The “No free lunch” theorem, which comes out of the machine learning literature, advises generally that there is no one learning algorithm that is better than others. Extending this, any solution that perfectly fits the problem is likely to be so specific as to have extremely limited ability to generalize to other problems, while any solution that addresses all problems is likely to be so general as to have limited predictive value. In other words, “[t]here is no algorithm which is better than all the rest for all kinds of data.”¹⁸ The best advice when evaluating your work is to remember that, “essentially, all models are wrong, but some are useful,”¹⁹ which serves to remind us that these are human-derived mathematical constructs, not reality.

Again, we also need to guard against “leaks from the future”²⁰ and other circular logic. A few years ago I was shown an elegant model used to predict

motive in gang-related homicides. The predictive ability of the model was so impressive as to be suspect. Like many things in the public safety and security world, if it looks too good to be true it probably is. After looking at the details of the model it became apparent that it had been created using closed cases; “suspect gang affiliation” was a key predictive variable. On seeing that, my first question was, “if you know who did it, why don’t you just ask them!” As discussed in Chapter 6, motive becomes important in a homicide investigation for its ability to create a short list of possible suspects. Requiring detailed knowledge of the suspect significantly limits the value of the model by creating a somewhat circular argument (i.e., you try to determine the motive to identify the suspect; if you need suspect information to determine the motive...).

It is also important to remember that the ultimate “cost” of errors in the applied public safety setting can be lives. Therefore, specific attention to errors and the nature of errors is absolutely required. In some situations, anything that brings decision accuracy above chance is adequate. In other situations, however, errors in analysis or interpretation can result in misallocated resources, wasted time, and even loss of life. Therefore, I have at times declined to provide the results of an analytic assessment if I was less than confident in the findings, deferring instead to the tacit knowledge and domain expertise of the operational end user over questionable analytic results. Overall, though, while we cannot prevent everything, advanced analytics brings great promise to crime and intelligence analysis in its ability to create the insight necessary to anticipate, influence, and respond more effectively in support of consequence management and changed outcomes.

Finally, it is important not to get discouraged and remember that, “disappointment, in science, is sometimes a gateway to insight.”²¹ In other words, data mining and predictive analytics are about confirmation AND discovery. Sometimes, it is the “surprising” findings that yield the greatest insight and most profound leaps forward in our understanding of the really hard problems in operational public safety and security.

Bibliography

- 1 Davenport TH, Jarvenpaa SL. Strategic use of analytics in government. IBM Center for the Business of Government: Managing Performance Series. <http://www.businessofgovernment.org/sites/default/files/Strategic%20Analytics.pdf>; 2008.
- 2 McCue C, Parker A, McNulty PJ, McCoy D. Doing more with less: Data mining in police deployment decisions. *Violent Crime Newsletter*, U.S. Department of Justice, Spring 1; 2004, 4–5 and Beck C, McCue C. Predictive policing: what can we learn from Wal-Mart and Amazon about fighting crime in a recession? *Police Chief*; November, 2009.
- 3 Gregory DC. Overall, serious crime rate down. *Chesterfield Observer* 2013; 18: 10.
- 4 Report of the Official Account of the Bombings in London on 7th July 2005 (HC 1087). https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/228837/1087.pdf; [accessed 11.05.2006].

- 5 O'Brien M. Manhunt – Boston Bombers. NOVA, original air date 29 May 2013. <http://www.pbs.org/wgbh/nova/tech/manhunt-boston-bombers.html>; 2013.
- 6 Hardy W. IBM's big hope for fraud. The New York Times. http://bits.blogs.nytimes.com/2014/03/20/ibms-big-hope-for-fraud/?_php=true&_type=blogs&_php=true&_type=blogs&_php=true&_type=blogs&mid=tw-share&r=2&; 2014 [accessed 20.03.2014].
- 7 Even relatively "objective" clearance rates are subject to manipulation. Several years ago I encountered a violent crimes unit commander who suggested a particularly clever approach to boosting his department's clearance rates, which were the subject of some controversy. According to Uniform Crime Reporting (UCR) rules, a case can be cleared by "exceptional means" if the death of the offender can be documented. Using actuarial tables, this commander had gone back in time and identified all cases where it was certain that the offender had died given the age of the case. His notional plan was to periodically add these "clearances" to his statistics in an effort to boost the department's problematic clearance rates. Clearly, this is an absurd example and the commander did not manipulate his rates, but it underscores how easy it is to manipulate even "objective" outcomes with a little bit of insight and creativity.
- 8 Talbot D. Crime software: still awaiting a verdict. MIT Technology Review. <http://www.technologyreview.com/view/512121/crime-software-still-awaiting-a-verdict/>; 2013 [accessed 06.03.2013].
- 9 Beck C, McCue C. Predictive policing: what can we learn from Wal-Mart and Amazon about fighting crime in a recession? Police Chief; 2009.
- 10 McCue C, Miller L, Lambert S. The Northern Virginia military shooting series: Operational validation of geospatial predictive analytics. Police Chief. http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=2871&issue_id=22013; 2013.
- 11 McCue C, Hildebrandt W, Campbell K. Pattern analysis of the Lord's Resistance Army and internally displaced persons. Human Social Culture Behavior (HSCB) Modeling Program Winter 2012 Newsletter, Spotlights 2012;12: 9.
- 12 The triborder region in Africa includes the Central African Republic, Democratic Republic of the Congo, and Sudan (South Sudan separated and became an independent country during the course of our study).
- 13 McCue C., Miller L. Lambert S. The Northern Virginia military shooting series: Operational validation of geospatial predictive analytics. Police Chief. http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=2871&issue_id=22013; 2013.
- 14 Ibid.
- 15 Elder Research Inc. Elder Research Builds Custom Tool to Help Reduce Fraud, Waste, and Abuse at the Postal Service IG. http://datamininglab.com/images/case-studies/ERI_USPS_OIG_Case_Study.pdf; 2014.
- 16 Charlotte-Mecklenberg Police Department Reduces Crime, Improves Efficiency, and Gains \$7.8 Million in Net Benefits with WebFOCUS Law Enforcement Analytics. Case Study Forum. http://www.informationbuilders.com/pdf/products/webfocus/IBI_Charlotte-Mecklenburg%20PD%20Case%20Study_2-1-12.pdf
- 17 Michaels J, Dilanian K, Leinwand D. Pirates another problem for Obama. USA Today. http://usatoday30.usatoday.com/printedition/news/20090409/1apirate09_st.art.htm; 2009 [accessed 09.04.2009].
- 18 This essay provides an excellent overview of the common pitfalls and challenges in data science. Braun ML. Data analysis: the hard parts. Marginally Interesting: Machine Learning, Computer Science, Jazz, and All That 2014. <http://blog.mikiobraun.de/2014/02/data-analysis-hard-parts.html>; [accessed 17.02.2014].
- 19 Box GEP, Draper NR. Empirical model building and response surfaces. New York: John Wiley & Sons, p. 424.
- 20 Dr. John Elder has compiled many of the common errors and mistakes into a wonderful primer on the topic entitled, "Top 10 Data Mining Mistakes," in: Nisbet R, Elder J, Miner G. (Eds.). Handbook of statistical analysis & data mining applications (Chapter 20). Boston: Academic Press; 2009. pp. 733–54. I was fortunate enough to receive a standalone pamphlet that I keep on my desk as a key "go to" and reality check.
- 21 Quammen D. Spillover. New York, N.Y.: W.W. Norton, 2012.

Operationally Actionable Output

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

John Tukey

In many ways, the creation of operationally relevant and actionable output is the most important step in the analysis process because it is where we translate the data and math into knowledge, insight and action. Whether you are visualizing descriptive statistics, or the results of a sophisticated machine learning algorithm, the effective visualization of the output is where we operationalize the equation, “data + context = insight.”¹ At a minimum, the goal is for the end user to be able to accurately interpret and understand the results of the analysis. Ideally, they will be able to incorporate their tacit knowledge and domain expertise, and effectively extend from the results in support of meaningful insight and novel responses to some of our most difficult problems.

Albert Einstein is quoted as saying, “most of the fundamental ideas of science are essentially simple, and may, as a rule, be expressed in a language comprehensible to everyone.” Again, everything that preceded this step in the process means nothing in applied data mining and predictive analysis unless it can be translated directly into the operational setting and used for decision support. The ability to translate the analytical products from the data mining process directly into the operational environment holds the promise for significantly enhancing situational awareness, guiding information-based decisions, and ultimately changing outcomes. Unfortunately, analytical output, particularly output generated by advanced modeling techniques, tends to be relatively opaque and generally does not translate well into the applied setting, which significantly limits its operational value. Visualization techniques, though, can be invaluable in conveying complex information quickly. The ability to create visual representations of the analytical results in an operationally relevant format holds tremendous potential for bridging the gap between analytical science and operational practice. Therefore, one of the hallmarks of the Actionable Mining and Predictive Analysis model is the use of operationally

actionable visual representation of the data and analytic results that can significantly enhance the knowledge discovery while guiding operational strategy and tactics.

Ideally, the objective is to address the “I’ll know it when I see it” goal of effective visualization. Anything less is the analytic equivalent of a tree falling in the woods with no one there to hear it. Ultimately, effective visual output should say to the end user, “Go here now and expect this,” with “this” being immediately obvious to the end user. Again, it does not matter how prescient or insightful the analysis is if it cannot be translated into action in the operational environment. Operationally relevant and actionable are the nonnegotiable requirements of effective output in the applied public safety and security environment, enabling the end users to incorporate their tacit knowledge and domain expertise in the understanding and use of the content.

9.1 ACTIONABLE OUTPUT

While the importance of operationally relevant and actionable analysis may seem obvious, this requirement was not always apparent. Moreover, it is important to remember that achieving this goal can represent a significant challenge to the analyst as they may see things in the data not readily apparent to others given their immersion in the data during the preprocessing and analysis steps. In addition, I continue to believe that, “operators are from Mars and analysts are from Venus.” While we serve mutually beneficial, complementary functions, our view of the world and interaction with data frequently tend to be very different. To paraphrase a line from the movie *The Sixth Sense*, “I see data.” With that in mind, graphical representations and other visualization techniques can help the analyst share their vision of the patterns and trends that are embedded in the data with others. Unfortunately, translation of modeling results into operationally relevant and actionable output can be particularly challenging in the applied public safety and security setting.

Again, while the importance of operationally relevant and actionable analytic output seems obvious now, it was not always so. The following series of examples illustrates our journey – including some missteps, false starts, and course corrections – while we worked to align the analytic output with the operational requirements and constraints. The specific examples are covered in greater detail in other chapters, but the overall development process concomitantly illustrates our maturation in thinking, while also underscoring the importance of operationally relevant and actionable output, as well as the tradeoff between model accuracy versus generalizability.

The output shown in [Figure 9.1](#) illustrates the results from a supervised learning algorithm. As an analyst, I would spend hours reviewing results like these;

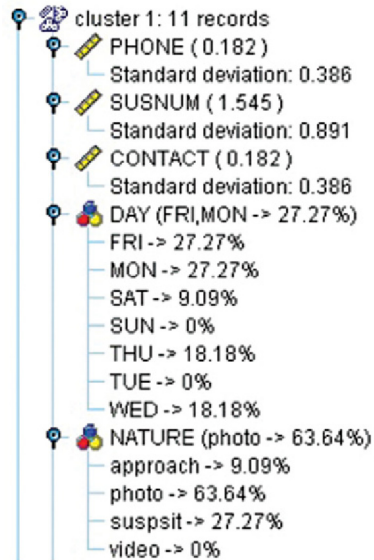


FIGURE 9.1

Some analytical output products, while accurate, can be very difficult to interpret and deploy within an operational environment. *IBM® SPSS® Statistics software (“SPSS”). Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation. SPSS Inc. was acquired by IBM in October, 2009.*

examining and reexamining subtle details in the model generated, making adjustments in the parameters and approach in an effort to improve performance, and achieving that sense of analytic nirvana when all of the factors aligned and the created model performed well on a set of test data. While these models were accurate and reliable, the results had limited value in the operational setting in their current form. Results like those depicted in [Figure 9.1](#) represented something that only a data scientist could easily interpret, appreciate, and use. In order for these models to bring value to the operational setting they needed to be actionable; the best course of action needed to be readily apparent to a nontechnical end user and domain expert.

Taking a cue from the e-commerce community, our team developed a capability that enabled us to deploy a scoring algorithm through a user-friendly web interface. Modeled after an online shopping experience, the web interface was designed to be intuitive, very simple, and easy to use ([Figure 9.2](#)). The end user simply provided the necessary inputs to the model, hit the “predict” button, and the scoring algorithm ran against the data behind the scenes. Unlike many of the cumbersome incident reporting tools that the operational personnel struggled to use, only those variables that were relevant and required inputs

Edit View Favorites Tools Help
 Back → → Search Favorites History
 http://rpd-as-06/sv/viewer?category=objectid%3A%2F1104&object-id=oid%3A1105&catalog=true&backid=cat&catalogaction=view&view=detail&idprivate=true
 Citizen Business Visitor eCitizen
 The Official City of Richmond, Virginia Website
 Home Publisher Log Out
 FAQ | Directory | Help | Search | Site Map | Email Us
 << Back to Category | Add to Quick Links
 rmed robbery - related assault risk assessment
 created by detective on 2003.03.12 02:41:33 PM
Robbery Turning Violent Predictor
 Incident Start Date: 314 SUNDAY
 TYPE: Highway
 WEAPON: Shotgun
 Shift: 0001 - 0400
 Predict likelihood: 0401 - 0800 Start over
 Home Publisher
 Local Intranet

FIGURE 9.2

Early methods of scoring algorithm deployment included the use of a web-based model that was based on e-commerce websites. *Cleo, IBM® SPSS® Statistics software (“SPSS”). Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

for the model needed to be entered. In addition, the use of pull-down menus and auto population features facilitated data entry, while also increasing accuracy. This greatly improved the simplicity and ease of use, and concomitantly reduced the time required to use the tool.

In this particular example, the likelihood for a robbery escalating into an aggravated assault is calculated and provided to the end user in an easy to interpret format (Figure 9.3). At the time, we thought that this was great and would revolutionize crime analysis and policing. Unfortunately, it did not translate well into practice. While everyone had very positive feedback regarding the concept, we quickly found out that it was not being used in the field. This was somewhat puzzling until we learned that one of the few members of the command staff actually using the tool was a watch commander who would sit in his police cruiser at the beginning of his shift and run through every possible permutation of incident time and location (Figure 9.4). In doing so, he assessed the likelihood for every location during each 4-h time block; effectively creating a “schedule” of likely risk (Figure 9.5). Using this schedule, he then

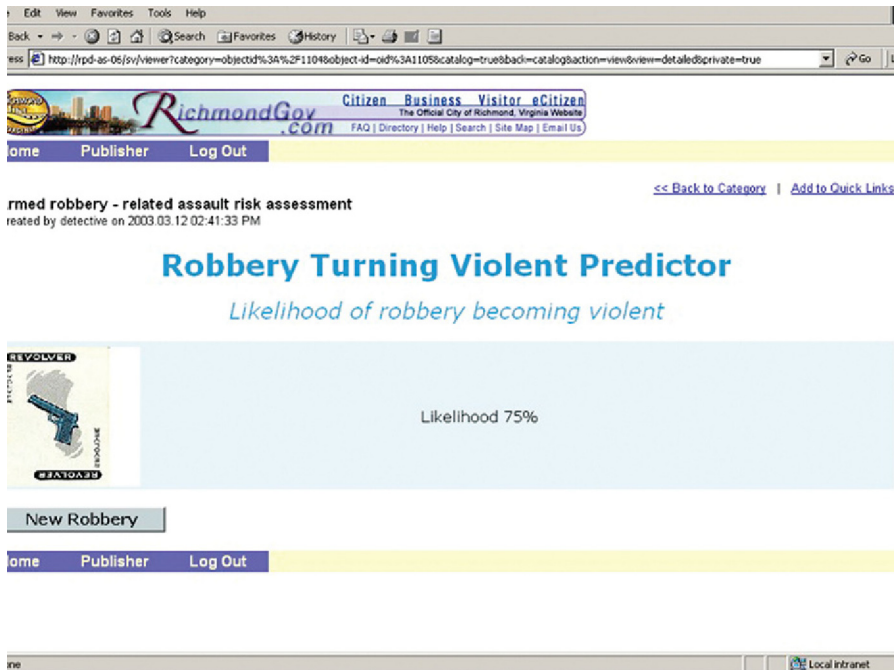


FIGURE 9.3

Sample output generated by the tool depicted in Figure 9.2. Cleo, IBM® SPSS® Statistics software ("SPSS"). Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.



FIGURE 9.4

Sometimes plans change. While early feedback regarding the web-enabled tool was positive, further investigation revealed that the actions necessary to generate operationally actionable results were time consuming and cumbersome; requiring field commanders to run every possible permutation of time and space at the beginning of their tour in an effort to derive output that they could use to inform deployment.

Area	Shift					
	0000–0359	0400–0759	0800–1159	1200–1559	1600–1959	2000–2359
1						
2						
3						
4						
5						

FIGURE 9.5

Sample deployment “schedule” created by an early adopter using the results from the automated tool depicted in [Figures 9.2 and 9.3](#).

deployed his patrol units when and where he expected the greatest risk in an effort to prevent crime, or respond more rapidly.

Recognizing that deployment represents the allocation of resources across time and space based on likely or anticipated risk, we captured and then extended the watch commander’s approach by translating the model output into a map. At the time, there were no mapping tools resident in the data mining platform that we were using, so using brute force techniques we leveraged existing patrol boundaries in the creation of categorical spatial “sets” that were used as inputs in a supervised learning algorithm. Going back to the original model, we then captured the modeled likelihood for each patrol area and rendered it on a map using a standard geospatial tool. The generated output basically said to the end user, “Go here now and expect this” ([Figure 9.6](#)). This particular example is covered in greater detail in [Chapters 7 and 13](#); however, the novel strategy of illustrating the results of a supervised learning algorithm in a mapping environment, albeit crude, highlights the analyst’s ability to convey the necessary information to the end user; ultimately converting relatively complex analytical output into something both intuitive and actionable.

It is important to note in this particular example that there was some loss of accuracy in the translation of the model from a web-enabled “black box” scoring algorithm into a mapping environment. Again, without the ability to directly import or link the results of the model into the mapping tool, we were

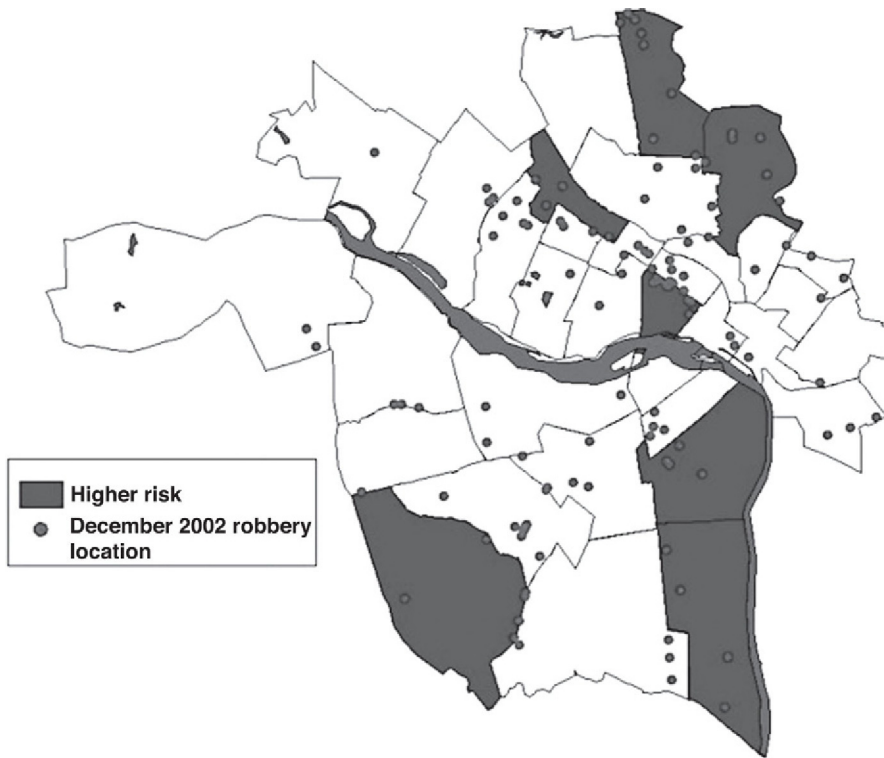


FIGURE 9.6

Example of operationally relevant and actionable output. Map illustrating the results of a supervised learning algorithm identifying locations associated with an increased likelihood of robbery-related aggravated assaults. Additional detail for this particular example can be found in Chapters 7 and 13.

required to create a categorical spatial set that allowed us to manually translate the results from the modeling algorithm into the mapping tool. In patrol deployment, however, almost anything that improves our ability to accurately anticipate incidents above chance can be useful so the tradeoff between model accuracy in favor of interpretability and use was acceptable. Other types of incidents, however, require far greater accuracy and will be described later.

These original findings formed the foundation for additional innovation and related technology development, including the development of capabilities that enable the analyst to directly convey the results of sophisticated modeling algorithms in a geospatial environment. The following example illustrates the Law Enforcement Analytics capability developed by Information Builders (Figures 9.7 to 9.10), which effectively captures and extends the development process described earlier. Using real-time scoring algorithms, the analyst is able to input current conditions to include day of the week, time of day, and

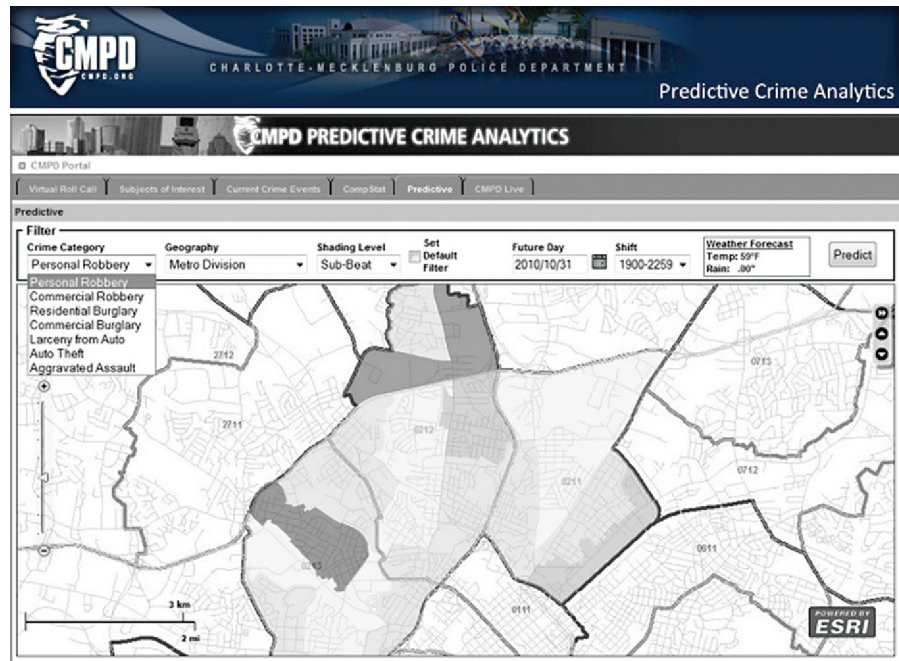


FIGURE 9.7

Map depicting probability of crime by dispatch zone. The shading illustrates the probability of Personal Robbery between the hours of 1900 and 2259 in Metro Division. *Used with permission, Information Builders LEA WebFOCUS, and Chief Rodney Monroe, Charlotte-Mecklenburg Police Department.*

weather conditions to generate an estimate of risk, which is rendered in a map. [Figure 9.7](#) depicts the probability of crime (personal robberies) by dispatch zone (“Sub-Beat”) under specific conditions. This information can be used directly by operational personnel for deployment planning purposes. As seen in [Figure 9.8](#), a specific area has been captured, and individual offenses have been rendered by location for the previous 7 days. This image provides additional detail about the area, allowing the end user to access specific details regarding a particular incident portrayed on the map. This feature enables them to incorporate their tacit knowledge to identify particular environmental features, landmarks, or other attributes in the interpretation of the results, as well as any operational planning, thereby enriching and extending the analysis.

[Figures 9.9 and 9.10](#) provide additional insight regarding the role that time of day may play in crime trends and patterns. [Figure 9.9](#) illustrates the distribution of auto thefts by beat for the time period 1100–1459, on 16 February. [Figure 9.10](#) also illustrates the distribution of auto thefts by beat, but for a different time period (1900–2259), on the same date (16 February). As can be seen when comparing the two different figures, auto theft frequency and geospatial

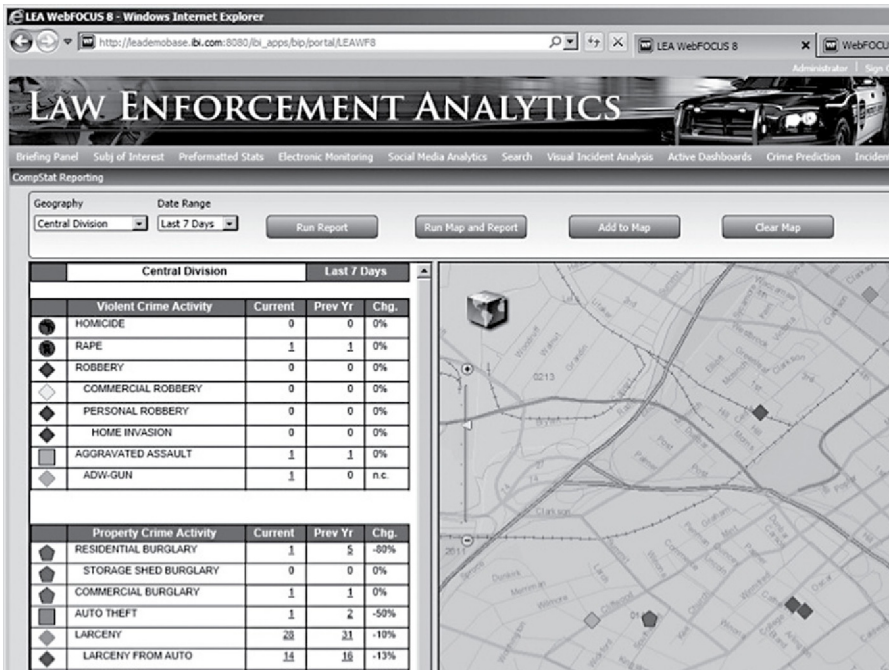


FIGURE 9.8

Specific offenses are rendered by location for the previous 7 days. By providing additional detail regarding the area, the end user can incorporate their tacit knowledge in the interpretation and operational use of the results. *Used with permission, Information Builders LEA WebFOCUS, and Chief Rodney Monroe, Charlotte-Mecklenburg Police Department.*

distribution differs markedly between these two time periods on the same day, which has significant implications for the allocation of police resources. Again, the end users can incorporate their tacit knowledge of the area and domain expertise as they add value to the interpretation and operational use of these results. Ultimately, capabilities like these enable the police manager and command staff to effectively characterize and anticipate crime trends and patterns in support of information-based approaches to resource allocation, prevention, thwarting, and response.

9.2 GEOSPATIAL CAPABILITIES AND TOOLS

The savvy reader will notice that there are a lot of maps in this book. This is particularly true with regard to the operational examples. This fact underscores the value that geospatial capabilities bring to the creation of operationally relevant and actionable analysis, and is especially true in the visualization of complex analytic output in the operational public safety and security setting.



FIGURE 9.9

Distribution of auto thefts by beat for the time period 1100–1459. *Used with permission, Information Builders LEA WebFOCUS, and Chief Rodney Monroe, Charlotte-Mecklenburg Police Department.*

Again, the end user can incorporate their tacit knowledge and domain expertise to not only understand and interpret the analytic results, but also to extend from them and develop a novel insight. Also, maps are operationally relevant and actionable. The analyst can easily convey the “when, where, and what” (i.e., “go here now and expect this”) of an analysis using geospatial tools. Finally, maps represent a unique environment for transdisciplinary collaboration.² Domain expertise with differing backgrounds, training, expertise, and experience can effectively collaborate in the geospatial environment in support of novel, transdisciplinary solutions to the really difficult problems. Therefore, geospatial capabilities and tools deserve special mention in a discussion of operationally relevant and actionable analysis.

Perhaps the earliest example of the effective use of maps to surface novel insight goes back to a map created by Dr. John Snow in 1854 to illustrate the geographic distribution of cholera cases during the Broad Street outbreak.³ While Snow was not the first to use a map to illustrate the location and distribution of the cholera incidents, his map was different in that he incorporated additional detail to convey the location of associated infrastructure, including

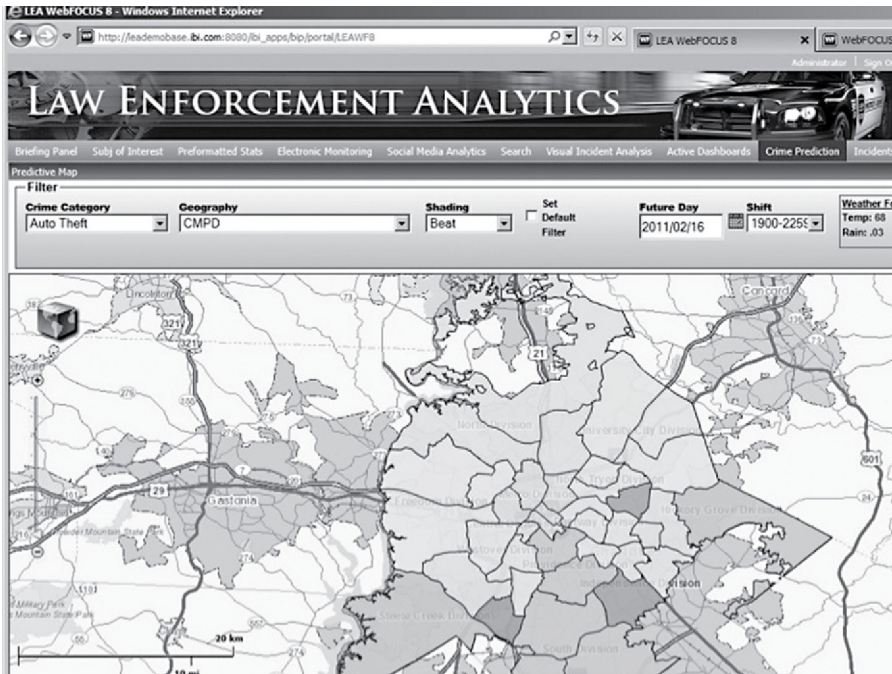


FIGURE 9.10

Illustration of auto thefts by beat for the same date as depicted in [Figure 9.9](#), but a different time period (1900–2259). The ability to identify and characterize temporal trends and patterns in crime frequency and geospatial distribution enables proactive, fluid deployment models and information-based approaches to resource allocation. *Used with permission, Information Builders LEA WebFOCUS, and Chief Rodney Monroe, Charlotte-Mecklenburg Police Department.*

the positions of the 13 water pumps that served the neighborhood. It was this information that highlighted the now infamous Broad Street pump as the focal point of the outbreak, which was subsequently confirmed as the source of the disease. While Snow’s map did not specifically identify the pump handle as the source of the outbreak, his work provided the context necessary to localize the source and essentially solve the mystery. A second innovation embedded in Snow’s cholera map was the inclusion of “street-level knowledge” and local experts to provide additional information and enhanced context to the results. Again, groups like Ushahidi are revisiting this concept and effectively leveraging the power of the crowd to capture and process local data in a timely manner.

[Figures 9.11 to 9.14](#) illustrate a modern extension of the Snow example. In this notional example, which is based on real experience, the analytic question related to a cluster of Improvised Explosive Device (IED) incidents. The incidents

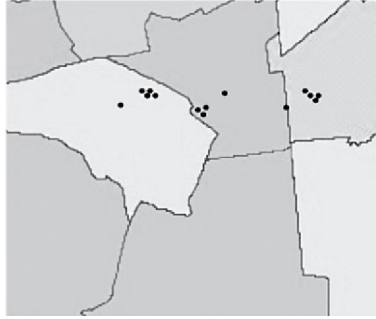


FIGURE 9.11

Notional example of Improvised Explosive Device (IED) incident data. While there appears to be some clustering of the incidents, the events transcend administrative boundaries and no obvious pattern is evident.

were occurring in a relatively broad area that spanned several administrative districts (Figure 9.11). While it appears that the incidents are occurring in clusters, there is no readily apparent pattern. Refining the analysis to include road networks, however, suggested that the IED incidents were associated with a main thoroughfare (Figure 9.12). Again, while the clusters were readily apparent and appeared to be associated with intersections, all intersections were not equally affected. Incorporating additional context regarding the physical infrastructure to include traffic control devices revealed that the IED incidents appeared to be associated with traffic circles (Figure 9.13). In fact, consideration of the traffic flow revealed that the specific location of the incidents was associated

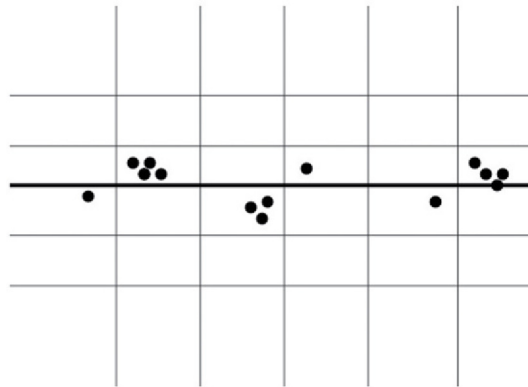
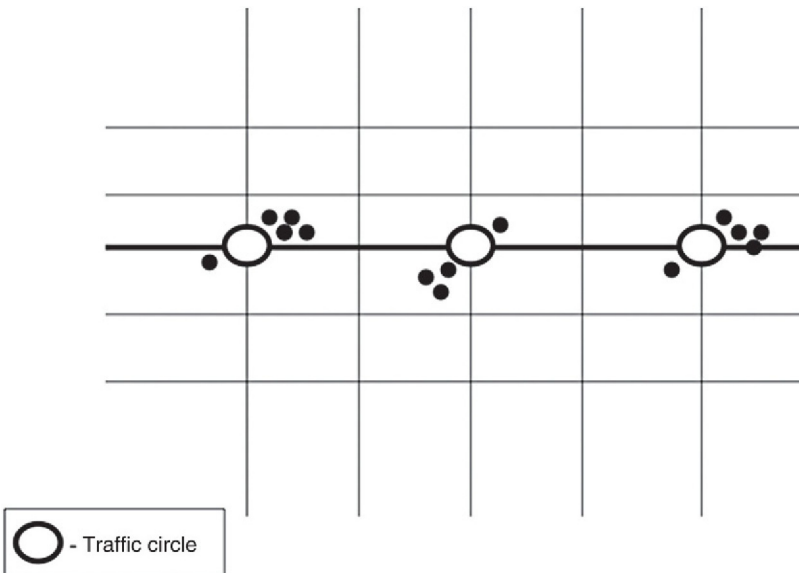
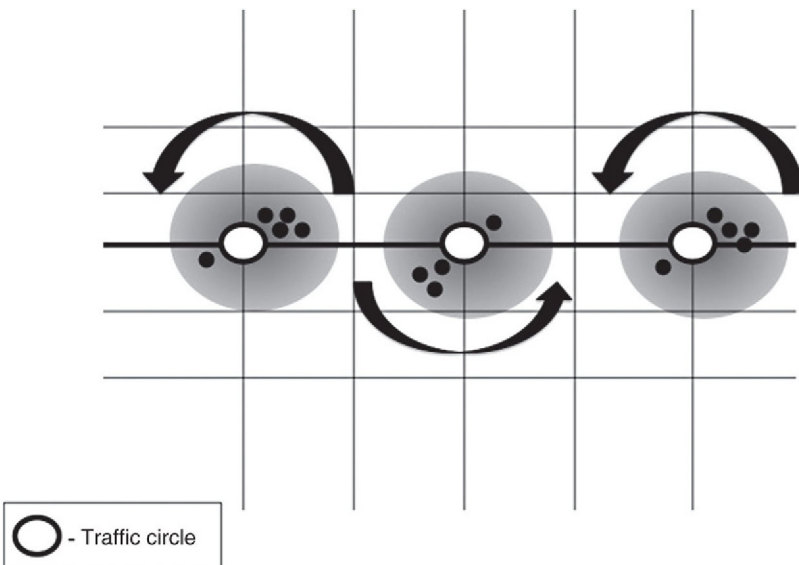


FIGURE 9.12

Adding a road network to the visualization reveals that the incidents appear to be associated with the main thoroughfare. While the incidents might be associated with intersections, not all intersections are similarly affected suggesting some other factor might be influencing IED emplacement.

**FIGURE 9.13**

Inclusion of additional infrastructure information reveals that the IED incidents are associated with intersections that have traffic circles to control speed, flow, and movement through the intersection.

**FIGURE 9.14**

Consideration of traffic flow and direction suggests that the vehicles are most vulnerable on entering the traffic circle and they slow down and work to negotiate the new traffic pattern.

with traffic entering the traffic circle (Figure 9.14). Additional analysis of the results suggested that the vehicles slowed down to enter and negotiate the traffic circle, creating congestions. As the traffic congestion increased, the slow moving vehicles represented easier and more numerous targets.

Again, the importance of context cannot be understated when considering the interpretation and use of our results.⁴ Once Snow placed the water pumps on the map, it created the key insight necessary to identify the Broad Street pump as the source of the epidemic. Similarly, adding the traffic circles to the map provided a possible explanation for the clustering of the incidents; vehicles slow down, particularly when entering the traffic circles, which makes a better target. These examples effectively illustrate the basic equation, “data + context = insight.”⁵

“I’m just trying to create the space for wisdom”

Merritt McKinney⁶

As stated earlier, another benefit of the use of geospatial capabilities to convey results is that it represents a unique transdisciplinary collaboration environment.⁷ Again, referring to the cholera maps created by John Snow in the mid-nineteenth century London, the ability to enable the “lateral, cross-disciplinary flow of ideas”⁸ can facilitate the development of truly novel insight regarding some of our most challenging problems, particularly those that span domains. In other words, maps can be used to not only say, “go here now and expect this,” but different players can bring their specific domain expertise, knowledge and experience, and work together to surface novel insight and truly unique solutions to some of our hardest problems. Geospatial capabilities and tools also are particularly well suited to creating context,⁹ “[b]y providing deeper, contextual analysis of places informed not only by the earth’s physical features and imagery intelligence, but also by ‘human geography’.”¹⁰ Representing a solid response to the fragmentation, heterogeneity, and related complexity of the MultiINT environment,¹¹ geospatial capabilities and tools can create a space where the whole is truly greater than the sum of the parts, which enables analysts to effectively integrate disparate sources, and fluidly transcend domains in support of novel insight and innovative solutions.

9.3 OTHER APPROACHES

After reading the preceding sections, should we assume that all of our analytic results be conveyed as map? Of course not! The key is to select the visualization technique that fits the question posed and operational requirements, including the frequently competing requirements for accuracy and interpretability of the results. For example, association matrices and link analysis are particularly well suited for the analysis of networks. Similarly, while the relatively opaque

deployment of a scoring algorithm did not work well for the patrol deployment example at the opening of this Chapter, this approach is used frequently for the analysis of fraud and other financial crimes where the requirements for model transparency are not as great and deployed scoring algorithms can be extremely complex.

Going back to the series of examples at the beginning of the Chapter, the watch commander had essentially created a heat map. A heat map can be a relatively intuitive visual representation of the data. Basically, the more intense the color, the greater the number represented on the figure. This can be used to depict simple frequencies or risk. In Figure 9.15, different types of crime (INCDGRP) are depicted over different 4-h time blocks (SHIFT). It is very easy to identify the most frequent crimes and when they occur. In this analysis, larcenies were the most frequent pattern of offending, which is not surprising, given their relative frequency in most communities. Examining the chart further, however, reveals that larcenies were most frequent during the 1200–1600 and

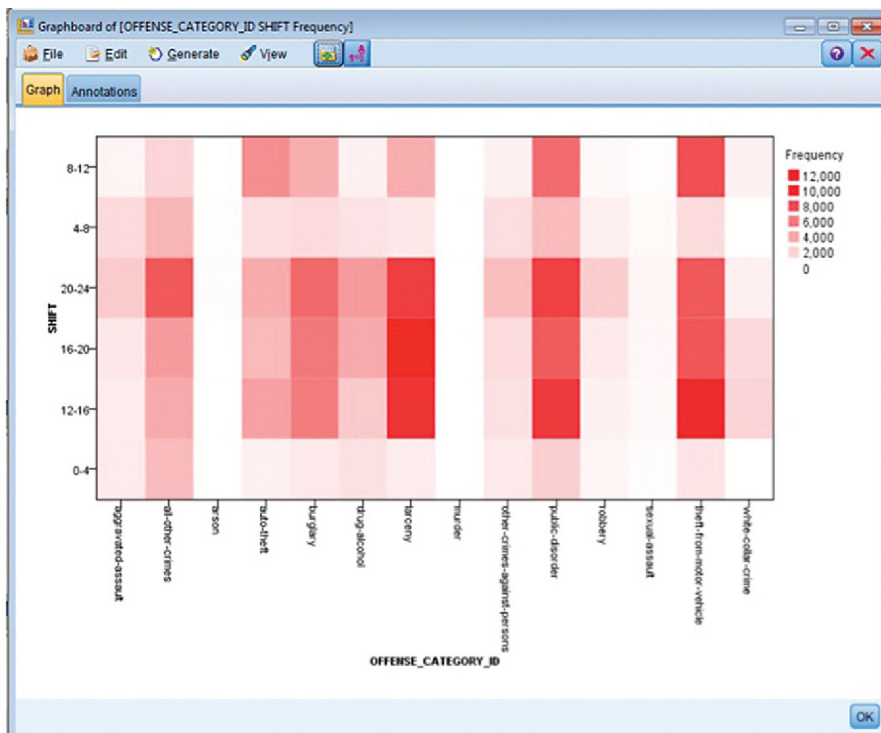


FIGURE 9.15

“Heat map” example. IBM® SPSS® Statistics software (“SPSS”). Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.

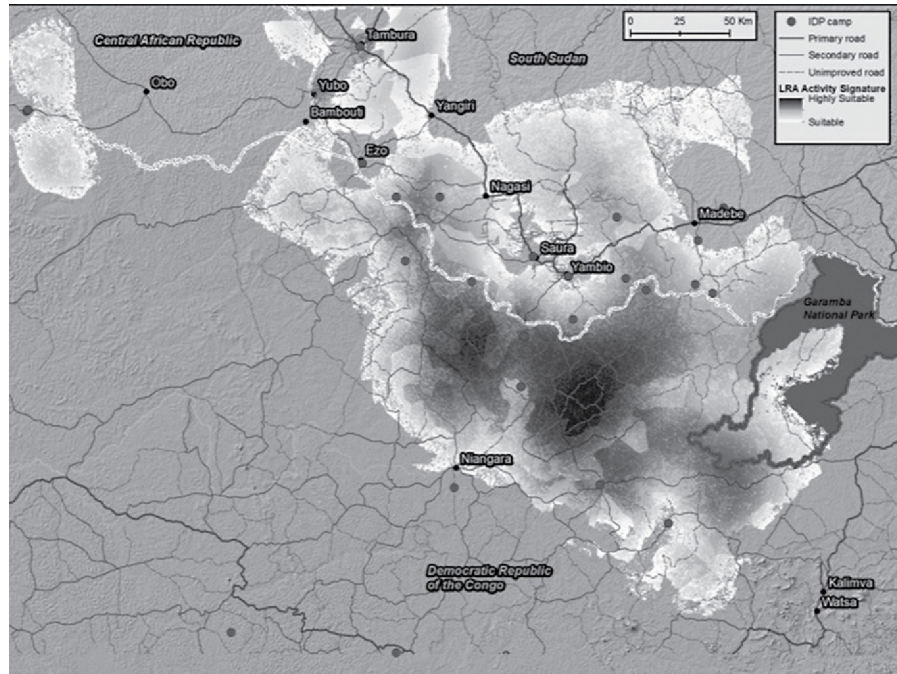


FIGURE 9.16

Illustration of the results of a geospatial predictive model of the Lord's Resistance Army activity in Africa. The results are rendered as a "heat map" where differences in the gray scale are used to convey differences in suitability or likelihood for a future LRA attack (additional detail is provided in Chapter 11).¹⁷ *DigitalGlobe, used with permission.*

1600–2000 hours time blocks. Similarly, assault and battery (A&B) offenses were more frequent during the evening time blocks. Heat maps also can be deployed in a geospatial environment to convey differences in likelihood or risk associated with specific locations (Figure 9.16).

As discussed in Chapter 7, the importance of balancing accuracy with transparency, or the ability to directly interpret the results, must be taken into consideration when creating analytic output. For example, the results of a drug-related violence model have been illustrated in a mapping environment to support deployment decisions (Figure 9.17). In this particular map, different patterns of drug-related violence have been differentiated visually in support of operational approaches that can be designed to specifically address the unique constellation of risk associated with a particular area. Again, the requirements for accuracy in deployment frequently are more flexible than other types of analysis, and the addition of context and opportunity for the end user to incorporate their tacit knowledge and domain expertise into the interpretation and

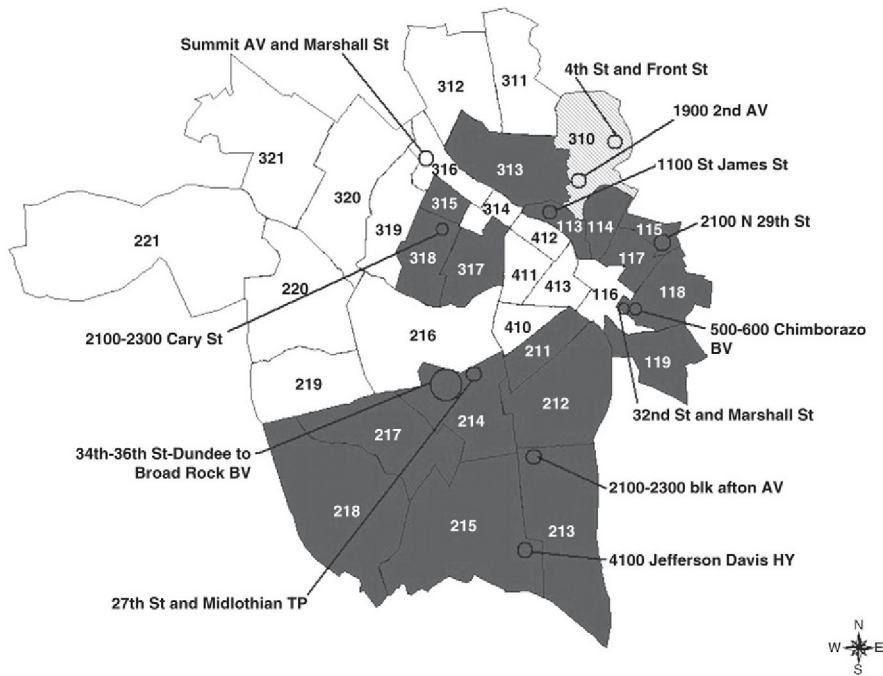


FIGURE 9.17

Map illustrating the results of a predictive model of drug-violence in a community. The dark areas indicate locations at increased likelihood for drug-related violence; the cross-hatched areas are associated with a different constellation of victim risk factors (additional detail is provided in Chapter 6).

use of these results more than offsets any reduction in accuracy that might have occurred by translating these results into a map.

Motive determination, on the other hand, generally requires greater accuracy given the potential consequences of misdirecting an investigation and/or identifying the wrong suspect. Moreover, many crimes occur at inconvenient times and inconvenient locations where immediate or timely access to analytic support may not be realistic. While we used the web-based deployment model as a negative example earlier in the chapter, there is greater tolerance for an opaque model in an investigative support role, which enables the analyst to deploy a more complex scoring algorithm. As illustrated in Figure 9.18, a scoring algorithm created by an analyst can be made directly available to field-based personnel, providing access to an agency's analytical capacity when and where it is needed most. The end user can enter a few relevant details and receive a risk assessment, automated motive determination, or some other analytical output without direct access to an analyst. Figure 9.19 illustrates the relevant, easy-to-understand output generated in response to the input illustrated in

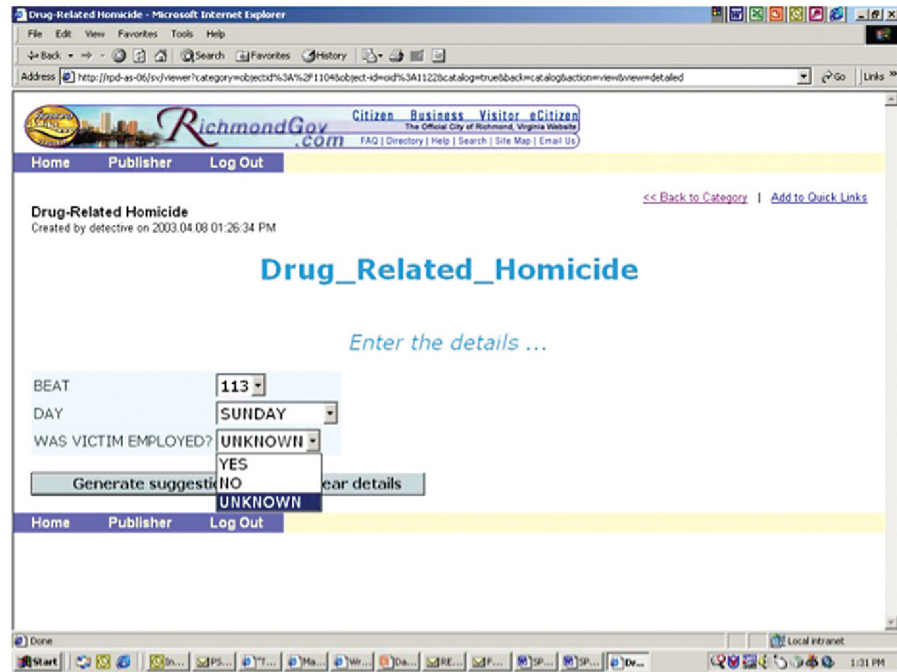


FIGURE 9.18

By using web-based deployment of data mining models, the investigator can benefit from 24/7 crime analysis, which means that analytical capacity will be available when and where it is needed, even at the crime scene. *Cleo, IBM® SPSS® Statistics software (“SPSS”). Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Figure 9.18. Therefore, identifying ways to translate data mining and predictive analytics output into a format with direct relevance to operational personnel provides the opportunity to greatly increase access to and related optimization of analytical resources.

Ultimately, using these web-based tools and a secure Internet connection, analytical capacity can be made available to operational personnel when and where they need it, providing increased situational awareness and investigative support on a 24/7 basis. Whether at a crime scene (Figure 9.20), or in a remote, forward-deployed location several time zones away, analytical support is just a keystroke away.

9.3.1 Future Trends in the Visualization of Output

As we reconsider the tremendous benefits that geospatial capabilities and tools bring to our ability to generate operationally relevant and actionable output – Go here now and expect this! – we realize that the community is at the cusp of

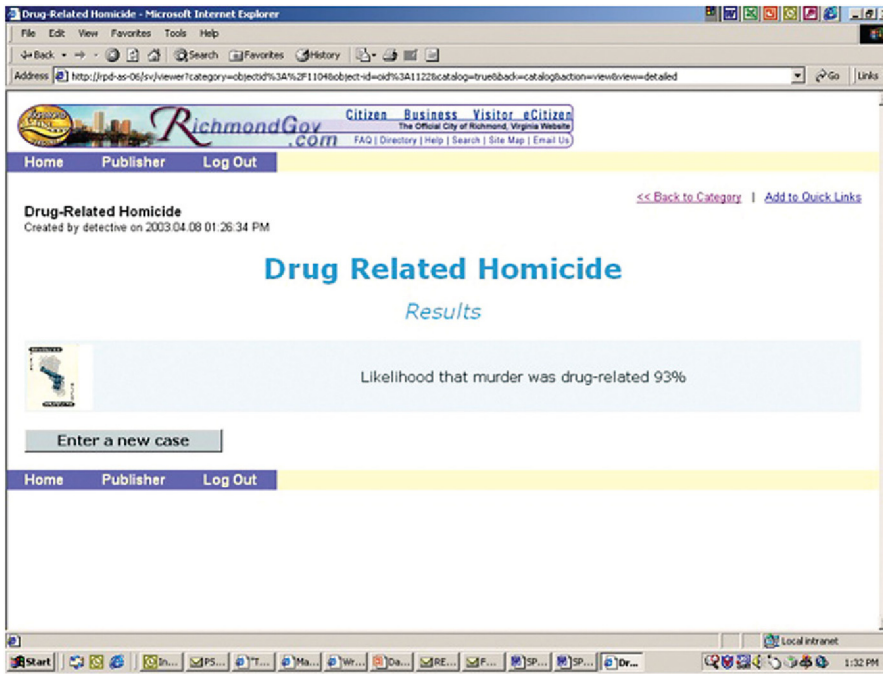


FIGURE 9.19

Sample output generated by the tool depicted in Figure 9.18. Cleo, IBM® SPSS® Statistics software (“SPSS”). Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.



FIGURE 9.20

Analytic results can be deployed through handheld devices, providing analytical support when and where it is needed.

a truly disruptive development in this space. In her keynote address at GEOINT 2013*, NGA Director Letitia Long emphasized the “immersive” environment that enables analysts and operators alike to experience the data and analytic results in new and exciting ways. By moving our analytic reports from static products to dynamic, interactive experiences we will be able to enable novel approaches to insight and discovery.

At least three approaches come to mind as I consider this revised model. The first includes exploitation of new capabilities that enable real time access to and manipulation of big data. The ability to conduct “analysis at the speed of thought”¹² represents a game changing paradigm shift in the way that an analyst can interact with material seamlessly and at scale; dynamically “pulling threads” and following them in support of novel insight and discovery. The ability to leverage these capabilities in support of a truly interactive, immersive experience holds great promise for significantly enhancing data exploration, and analysis for analysts, as well as their operational consumers.

Second, humans, like many other animals, are first derivative organisms, which means that we work very well as change detectors. A major benefit of the use of the geospatial environment in that it enables us to convey complex statistical results and relationships in an operationally relevant and actionable format; telling the end user, “go here now and expect this.” One way to exploit this further is by incorporating motion into our work. For example, the use of motion to depict change over time and space enables the end user to not only intuitively understand complex trends, patterns, and relationships, but also to extend from the analytic product in support of novel insight and discovery, effectively enabling them to anticipate what is likely to come next and influence outcomes. [Figure 9.21](#) includes a notional illustration of the results of an unsupervised learning algorithm that was used to provide insight regarding a series of suspicious situation reports associated with a critical facility. In this particular example, different colored dots were used to convey membership in different clusters of associated, operationally relevant patterns of behavior. Additional shading of the dots was used to convey progression of the incidents over time; however, the resulting image became so complicated that many of the subtle attributes suggesting escalation in behavior were lost; requiring serious hand waving on the part of the briefer in order to fully convey the nature of the behavioral change over time and space. The use of motion, however, would have enabled the end user to better appreciate the change in behavior and related escalation over time and space, allowing them to better understand the development of the behavior within the context of the physical infrastructure, and anticipate the “next steps” similar to the way that meteorologists use animated weather maps to enable intuitive “forecasting” in their viewing audience.

Again, the human brain is uniquely suited to function as a change detector. Extending this, there is an increasing role for the incorporation of cognitive

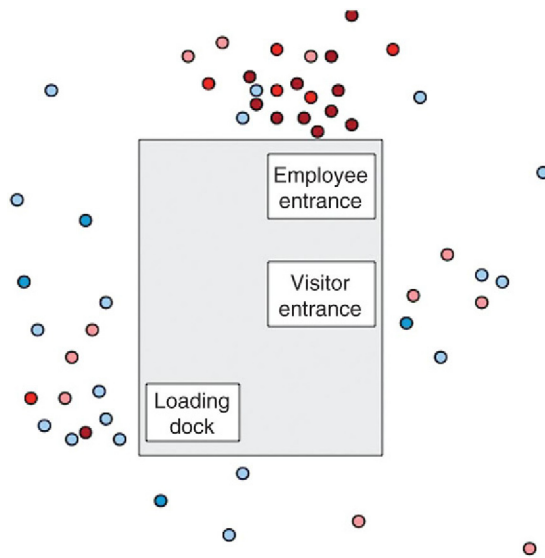


FIGURE 9.21

Facility map with overlay of unsupervised learning output. The original illustration used color to convey change over time. The ability to incorporate motion, however, would represent an excellent fit for the brain's ability to detect change.

neuroscience in the development of analytic process, tradecraft, and particularly output. The ability to effectively leverage these concepts enables us to create analytic products that address the "I'll know it when I see it" goal of good analysis. Software increasingly is able to support the development and easy viewing of complex images. Included in this improved capacity is the ability to use motion or movement to convey changes over time, including escalation in activity, and enable active exploration on the part of both analysts and operators, which directly addresses the call for dynamic as compared to the static experience of analytic output and results.

Finally, returning to the "Agent/Analyst" referenced in the Introduction, the ability to deploy real-time scoring algorithms and associated collection tools directly to the field increasingly blurs the line between operators and analysts (Figure 9.22). The opportunity to concomitantly inform both operations and collections in the forward environment holds the promise of increased situational awareness, and greatly improved collection and related knowledge management; particularly as we use these capabilities to fluidly enable incorporation of end user domain expertise and tacit knowledge, as well as operational requirements and constraints; ultimately moving analysis from the creation of static reports and documents, to a truly dynamic, interactive, and immersive experience.



FIGURE 9.22

The ability to deploy analytic content, including scoring algorithms, and concomitantly support collection provides increased situational awareness to the operational end user, while enhancing collection capabilities.

9.3.2 Cautions

Again, “when the combination of interesting data and clever display are properly aligned, remarkable outcomes can result.”¹³ It is important to remember though, that the data and related analysis need to be solid, and the results conveyed responsibly. The following list includes a few quick cautions regarding the potential dark side of visualization and other methods to visually display data and analytic results.

The increased access to powerful graphics tools and beautiful color printing devices has created the expectation for the use of color in everything that we produce. There are a few things to remember, though, when using color to visualize your results. First, while it can be tempting to be creative or “go wild” with color, specific colors mean specific things to the end user. For example, I encountered one software package several years ago that flipped the standard color ramp; green represented greater frequency or “high,” while red was used to convey infrequent or “low.” When I asked the vendor about their decision to do this, their response was that they wanted to set themselves apart from the other software companies in this space and saw their unique use of color as a differentiator in a crowded field. Unfortunately, what actually happened was that it was very difficult to interpret the results that their software tool generated.¹⁴ Whether it comes from expectations and standards, or actual sensory physiology and cognitive processing,¹⁵ end users tend to view red as “high” and green as “low.” Reversing this relationship creates confusion that requires the end user to actively think about the output, and violates the “I’ll know it when I see it” goal of great visualizations.

In addition to using “standard” or expected color, maintaining consistency in your visualization methods within and across analytic output also enables the end user to immediately “know” what you are trying to convey and can use this foundation knowledge to immediately move into the interpretation of the results. While it can be tempting to change things in an effort to keep the report interesting or add some variety, keeping with a standard theme enables the end user to easily move between the results in pursuit of larger trends and patterns, particularly those that may transcend multiple reports. It is also important to remember that not everyone perceives color completely or even at all. Creating a compelling figure in color for an end user who cannot perceive color well really limits their ability to completely appreciate your work. While color is nice and can provide unique options for visualization, it is not essential to the creation of operationally relevant and actionable output. In fact, in case you have not noticed, this book is illustrated entirely in black and white. While several of the figures were created originally in color, shading, texture, placement, and other techniques can be used to visualize important elements.

Again, increased access to and use of “big data,” including personally identifiable information (PII), the use of derived products, and the ability to accurately and precisely identify location has raised concerns throughout the community regarding potential privacy and security issues. This is particularly true as relates to vulnerable populations, which may include the victims of natural and manmade disasters, crisis and conflict zones, crime victims, and even suspects. Similarly, inclusion of an entity in a map, model, association matrix, or link chart implies a relationship. While these issues have been addressed in detail throughout the text, the importance of reviewing the output to ensure that the visualization accurately represents the findings, and that privacy and security have been properly considered, especially within the context of the intended end user or audience, cannot be overstated. This can be particularly true with automatically generated results where it is easy to switch to “autopilot” and run with whatever has been produced assuming that it is correct and appropriate for the intended recipient. As always, question, evaluate, check, and double-check your results to ensure that you “say” what you mean and mean what you “say” in your analytic output.

Finally, while visualization represents a powerful tool that can be used to effectively convey complex analytic output in an operationally relevant and actionable manner, be cautious of using it to oversell your results. Like the importance of being mindful of the sample size and denominator, adjustment of graph axes and scale can be used to magnify desired results, or obscure embarrassing trends. While it can be tempting to use creative visualization to convey what you “know” to be there, be cautious of engaging in “Gladwellism – the hard sell of a big theme supported by dubious, incoherent but dramatically presented evidence.”¹⁶

Bibliography

- 1 Gheorghe C. Big idea 2014: the warm embrace of context. <https://www.linkedin.com/today/post/article/20131210125224-1336307-big-idea-2014-the-warm-embrace-of-context>; 2013 [accessed 10.12.2013].
- 2 Usher A, Bowers J. Human geography: an evolving discipline. *Geospatial Intelligence Forum* 2012; March: 22–24. http://issuu.com/kmi_media_group/docs/gif_10-2_final.
- 3 Johnson S. *The Ghost Map*. London: Penguin Books. 2006.
- 4 Brown ED. Context and big data. *Big Data Forum*. <http://www.big-dataforum.com/103/context-and-big-data>; 2014 [12.02.2014].
- 5 Gheorghe C. Big idea 2014: the warm embrace of context. <https://www.linkedin.com/today/post/article/20131210125224-1336307-big-idea-2014-the-warm-embrace-of-context>; 2014 [accessed 10.12.2013].
- 6 Fictional character played by Woody Harrelson in the movie, *Now You See Me*; 2013.
- 7 Usher A, Bowers J. Human geography: an evolving discipline. *Geospatial Intelligence Forum* 2012; March: 22–24. http://issuu.com/kmi_media_group/docs/gif_10-2_final.
- 8 Johnson D. 2006, p. 225.
- 9 Usher A, Bowers J 2012.
- 10 Long LA Putting the Power of GEOINT in Your Hands. *GEOINT Symposium*, 2 November 2010. In: National Geospatial Intelligence Agency (2011). *Incorporating Human Geography into GEOINT, Student Guide (SG)*. The School of Geospatial-Intelligence: NGA College 12Sep11 (1-1-3); 2013. <http://www.scribd.com/doc/113126238/NGA-Incorporating-Human-Geography-Into-GEOINT-NGA-College-12Sep11>.
- 11 Konkel F. The intelligence community's big-data problem. *FCW*. <http://fcw.com/articles/2014/03/13/ic-big-data.aspx>; 2014 [accessed 13.03.2014].
- 12 Coined by SAP NS2, the concept of “analysis at the speed of thought” has been embraced by others in the community given the immediate and direct appeal of the concept (SAP NS2, *In-Memory Computing for Intelligence Missions, Executive White Paper*. <http://www.sapns2.com/files/white-paper/Exec%20White%20Paper%20-%20In-Memory%20Computing%20for%20Intel%20Missions%20Paper.pdf>).
- 13 Wainer H, Lysen S. That's Funny... A window on data can be a window on discovery. *Am Sci* 2009; July–August. <https://www.americanscientist.org/issues/pub/2009/4/thats-funny>.
- 14 It was almost unsettling and caused me as the end user to really struggle to interpret the results – definitely not the goal of good visualization! In this case, the reversed color ramp similarly did not look “right” and was very distracting, reminding me of the “disgust” research where they found that people did not think that blue chicken tastes “right” and in some cases would not eat it (Poon L. *Tasting with our eyes: why bright blue chicken looks so strange*. NPR. <http://www.npr.org/blogs/thesalt/2014/04/16/303215873/tasting-with-our-eyes-why-bright-blue-chicken-looks-so-strange>; 2014 [accessed 16.04.2014]).
- 15 There are a number of suggested reasons why red is viewed as “high” or more intense, many of which go back to nature and include the fact that red, particularly when viewed against a green background, is highly visible at a distance, which makes it an excellent “danger” cue (Khounsary A. *Ask a scientist: color red and danger*. Again, working with rather than against or around biology really increases the potential value of the visualization. <http://www.newton.dep.anl.gov/askasci/gen99/gen99420.htm>; 1999.).
- 16 Appleyard B. Why futurologists are always wrong – and why we should be skeptical of technopians. *NewStatesman*. <http://www.newstatesman.com/culture/2014/04/why-futurologists-are-always-wrong-and-why-we-should-be-sceptical-techno-utopians>; 2014 [accessed 10.04.2014].
- 17 McCue C, Hildebrandt W, Campbell K. Pattern analysis of the Lord's Resistance Army and internally displaced persons. *Human Social Culture Behavior (HSCB) Modeling Program Winter 2012 Newsletter*. *Spotlights* 2012; 12: 9.

Normal Crime

“When people are free to do as they please, they usually imitate each other.”

Eric Hoffer

Ultimately, crime is behavior. Like many other patterns of behavior, even the most serious violent crime frequently can be characterized, categorized, and described, and hopefully even anticipated and influenced. In many ways, the behavioral analysis of violent crime is based on this assumption. If it were not predictable, we would not have the field of criminal profiling. We also would not have as many interesting movies or books.

While it is relatively interesting to sit back and assume the role of an armchair behavioral psychologist, the field truly is much more complicated than the popular press would lead one to believe. This can be a source of frustration for analysts and other professionals in the field when confronted by lay “experts.” While most of us would never dream of suggesting a certain approach or technique to a cardiac surgeon, it seems that everyone has an opinion regarding crime and criminals, regardless of their training or education in the area. This certainly is not meant to imply that crime or intelligence analysis approaches heart surgery or other similarly complicated medical procedures. However, crime and intelligence analysis is not nearly as simple and straightforward as the movie of the week might suggest.

Why has this subject been included in a text on data mining and predictive analysis? The relative degree of complexity associated with crime and criminal behavior, even garden-variety offenses, is something that public safety and security professionals can exploit in an effort to detect crime and other unusual or suspicious behavior. Moreover, this area in particular can benefit from the sophisticated analysis and characterization associated with data mining and predictive analytics.

In keeping with this idea, we can begin to develop a concept of “normal” crime. Again, like most patterns of behavior, crime can be seen as being somewhat homogenous and relatively predictable if viewed in the proper

light. People often get tripped up and caught when they try to behave normally or “fly under the radar.” This is especially true with criminal and other forms of bad behavior. In many cases, however, they often do not have a good sense of what normal truly looks like and get caught out of ignorance or because they stand out by trying to be inconspicuous. Recalling the Strawberry Pop-Tarts example from Chapter 3,¹ it is difficult to completely understand what normal looks like until we capture and characterize it. It is for this reason that understanding normal trends and patterns, as well as “normal” abnormal trends and patterns, can be a valuable component of public safety domain expertise.

Further underscoring the importance of “normal” in crime and intelligence analysis, deviation from normal also may indicate the potential for escalation or the presence of something really “bad.” For example, the crime of burglary generally is economically motivated. In “normal” burglaries, the goal of the criminal is to get in and out without being detected, and take, or at least attempt to take something of value. In our experience, deviations from this “normal” pattern generally are associated with the potential for serious escalation and/or violence. This is particularly true of offenders that appear to be targeting occupied dwellings and is addressed in greater detail in Chapter 11. Suffice it to say, though, that criminal behavior, although deviant by many definitions, can be described in terms of “normal” patterns of offending and certain routine baseline behaviors.

10.1 INTERNAL NORMS

Seasoned investigators generally have excellent internal norms or gut instincts. It is not unusual when things are a little odd to have a group of detectives standing around a crime scene commenting on how something “just doesn’t feel right” or reviewing something in an offense report and getting second opinions. Frequently, what they are saying is that their internal anomaly detector has gone off, although they probably will use less delicate terminology.

Supporting this model, the investigative training process resembles case-based reasoning in many ways. Investigators come to understand a new experience or a new case based on their prior experiences.² By accumulating an internal database of previous cases and associated outcomes, they can attempt to match each new experience to their internal library. If an experience matches a previous case, they have an internal scenario that can be used to structure the current investigation. For example, a husband calls and reports his wife missing; wife found murdered with signs of overkill; previous cases indicate domestic homicide; interview husband. If something new does not fit into their past experiences in any sort of logical fashion, then they have encountered an anomaly, which

requires further inquiry to either fit it into an existing norm or create a new category. In many cases, listening closely to these internal anomaly detectors frequently can highlight situations or individuals that merit further scrutiny.

10.2 KNOWING NORMAL

“Science is nothing but trained and organized common sense.”

Thomas Henry Huxley

Again, criminals often get caught because they try to fly under the radar but end up sticking out like a sore thumb because they do not have a good understanding of “normal.” For example, in one case of suspected embezzlement, the accounting ledgers had gone missing – only the bank statements were available for analysis, and the information on these were limited to the date and amount of each transaction. Any detailed information pertaining to the transaction recipient had disappeared with the ledger books. The account in question was established and maintained to pay bills and reimburse expenses. It would appear that this would be a difficult situation to identify even the most brazen fiscal impropriety because the information was so limited. In fact, as can be seen in [Table 10.1](#), the initial review of the bank statements suggested that nothing seemed to be amiss in the accounting. After the data

Table 10.1 Transaction Records From a Suspicious Bank Account

	Date	Amount	Chk. No.	Bal.
38	Oct. 26, 20XX	10.50	543	1640.46
39	Sep. 29, 20XX	120.90	544	3041.58
40	Mar. 16, 20XX	165.00	545	2406.63
41	Oct. 20, 20XX	39.12	546	1900.96
42	Nov. 17, 20XX	421.20	548	2536.81
43	Oct. 19, 20XX	42.00	549	1940.08
44	Oct. 30, 20XX	250.00	550	1390.46
45	Nov. 20, 20XX	75.58	551	1705.23
46	Nov. 21, 20XX	54.96	552	1584.82
47	Nov. 20, 20XX	756.00	553	1780.81
48	Nov. 17, 20XX	500.00	554	2958.01
49	Nov. 21, 20XX	65.45	555	1639.78
50	Dec. 26, 20XX	157.86	556	2068.47
51	Nov. 28, 20XX	39.49	557	1499.33
52	Nov. 24, 20XX	46.00	558	1538.82
53	Jan. 10, 20XX	11.40	559	1677.26
54	Dec. 26, 20XX	118.25	560	1950.22
55	Jan. 3, 20XX	71.56	561	1878.66
56	Jan. 16, 20XX	378.00	562	1299.26
57	Jan. 31, 20XX	8.00	563	1291.26
58	Feb. 28, 20XX	197.87	564	1620.39

Table 10.2 Frequency Distribution of the Withdrawals Associated with the Bank Account Illustrated in Table 10.1

Amount	Frequency	Percentage calculated	
80.05	1	0.4	0.4
81.00	1	0.4	0.4
87.86	1	0.4	0.4
88.78	1	0.4	0.4
89.00	1	0.4	0.4
89.50	1	0.4	0.4
95.70	1	0.4	0.4
96.00	2	0.8	0.8
97.50	1	0.4	0.4
100.00	23	9.3	9.3
105.00	1	0.4	0.4
110.00	1	0.4	0.4
117.74	1	0.4	0.4
118.25	1	0.4	0.4
120.00	1	0.4	0.4
120.90	1	0.4	0.4
125.00	2	0.8	0.8
125.25	1	0.4	0.4
126.43	1	0.4	0.4
126.50	1	0.4	0.4
128.00	1	0.4	0.4
130.00	3	1.2	1.2

An unusual distribution of expenditures in round amounts was suspicious and warranted additional investigation.

were put into a spreadsheet and graphed, however, it became apparent that the transaction amounts were very unusual. As depicted in Table 10.2, 23 of the 248 transactions reviewed were for \$100. Similarly, 20 of the checks written were for exactly \$200 and 12 were for \$50. More than one in five, 22%, of the checks written from this account to pay bills and reimburse expenses were for the amounts of \$200, \$100, and \$50. A quick review of my own bank statement reveals no such pattern of even transactions. While it was not clear exactly who the checks were written to, the amounts certainly raised questions regarding the activity in that account.

In fact, this pattern of checks is similar to what one would expect from someone writing a series of checks for “cash” in relatively even amounts. Armed with this suspicion, the data were reviewed again, and it was noted that initially there were a series of “counter” transactions in which the suspect was simply withdrawing funds from the bank teller; again, in even amounts. Generation of a quick timeline revealed that these counter transactions stopped abruptly after a duplicate statement was requested, suggesting that someone else might have been suspicious about the account activity.

This quick analysis provided the following information, which provided investigative leads for follow-up:

- The pattern of activity was unusual for an account maintained to pay bills and reimburse expenditures.
- The monetary total of suspicious transactions was approximately \$6900.
- The behavior changed after the duplicate statement was requested. The individual who requested the duplicate statement might have been suspicious and would be someone worth identifying and interviewing.

This relatively simple exercise demonstrates the value of characterizing data and drilling down to identify hidden details. It did not involve any sophisticated analyses and would be relatively easy to replicate in any setting. The important features were exploring the data and comparing it against what we know to be normal. Moreover, this example, like the work with telephone call data discussed in Chapter 7, was able to highlight actionable investigative details without specific information. Because the ledgers were missing, it was not possible to document specific details regarding the disbursements. The use of data mining, however, revealed patterns of financial activity that were at least curious. The deviation from “normal” financial transaction patterns certainly suggested that something unusual was occurring with this particular account, and supported the need for additional investigation.

10.3 “NORMAL” CRIMINAL BEHAVIOR

In the next example, a more rigorous statistical approach was employed to determine the relationship between different patterns of criminal offense, particularly as they relate to the potential for significant escalation.

The relationship between property crimes and stranger rapes will be discussed in Chapter 11; however, the salient feature for anomaly detection is that even criminal behavior is associated with normal trends and patterns. For example, if the primary goals of a burglary are economic gain and to escape without detection, breaking into an occupied dwelling and taking something of little or no value would be unusual and counter to the assumed primary motivations of the crime. Through a casual conversation with the Director of the Virginia Division of Forensic Science, it was noted that most of the offenders identified through the use of DNA cold hits had not been included in the Commonwealth of Virginia’s DNA database for previous violent or sexual crimes.³ Rather, many of these individuals were in the database for prior property-related crimes. Additional review of these cases revealed, however, that their property crimes differed qualitatively from “normal” trends and patterns associated with property crimes. On subsequent examination of these “abnormal” burglaries, there

frequently were unique attributes that suggested the potential for significant escalation into violent crime, particularly sexually violent crime.⁴ For example, in many situations, these incidents appeared to represent “near-miss” rapes, or even seemed to indicate that the offenders were developing their “victim access” methodology.

In other words, these crimes were not normal; they were not what one would expect if the motive in the associated property crime were economic gain with minimal risk. Through more detailed review of the offense reports, it was determined that these incidents differed from other property crimes in at least two significant elements. First, these crimes frequently were associated with occupied dwellings, which significantly increases the risk for detection and subsequent apprehension. Second, the offenders generally took something of limited value, if anything was even taken during the burglary. In some situations, the item taken could best be described as something having souvenir or trophy value, rather than any sort of worthwhile monetary value that would justify the risk associated with the offense. This “anomaly” suggests some other or additional secondary gain associated with the crime.

In our experience, anything that deviates from “normal” patterns of offending generally is cause for concern. Identifying the usual “primary” and “secondary” gain associated with a particular type of crime can be very important in interpreting the analytical results, which further highlights the importance of domain expertise in the analysis of crime and intelligence data. Behavior that deviates from or generally appears to be inconsistent with the expected secondary gain in a particular crime or series of crimes warrants further analysis.

10.4 GET TO KNOW “NORMAL” CRIME TRENDS AND PATTERNS

There is huge value in taking a few minutes periodically to mine routine crime data. Through this process, conventional wisdom, gut instincts, and common knowledge can be explored, confirmed, or reevaluated. This also is an excellent time to “discover” information that can add value to existing knowledge or investigative practice. Sometimes pulling back and looking at standard crime data with a fresh eye, drilling down, and categorizing it a slightly different way, can reap huge benefits for the analyst and the investigator.

While it would be nice if crime would just go away, this is unlikely to happen any time soon. Rather, a certain baseline of criminal offending in a community generally is to be expected. This probably is not something that policy makers and the command staff would want to discuss openly, but historical review supports the existence of at least some unacceptable or criminal behavior that emerges whenever and wherever groups of individuals congregate and coexist.

One would imagine that back in the days of Australopithecus, certain individuals could be found stealing brontoburgers from their neighbor or writing bad checks on clay tablets to purchase a newer grass hut. We certainly have knowledge of crime that was present during the earliest days of recorded history and biblical times.

Identifying and understanding normal rates of crime can help identify unusual spikes that warrant additional consideration. An unexpected spike in normal or expected criminal behavior can indicate a brewing problem. Certain data mining algorithms have been developed that monitor normal patterns of crime and offending. These are programmed to alert the analyst when significant deviations in crime frequency occur. This approach, which is called a control-charting function, has been incorporated into the Regional Crime Analysis Program (RECAP), developed by researchers at the University of Virginia.⁵ Effective, timely response to these localized fluctuations in crime will require a greater degree of fluidity in patrol deployment; however, it is much easier and more efficient to address emerging rather than established changes in the local crime climate.

On the other hand, some patterns of crime might fluctuate with weekly or seasonal periodicity, which also can be anticipated. Therefore, an understanding of normal crime patterns and trends can provide a necessary baseline for the comparison and evaluation of any perceived changes. In addition, characterization of normal patterns also can facilitate forecasting, which enables a proactive rather than a reactive response to any changes in offense rates. For example, seasonal changes frequently are associated with fluctuations in the number and location of auto thefts. Identifying this pattern can facilitate the development of proactive enforcement strategies, which are discussed in greater detail in Chapter 12.

10.4.1 Anomaly Detection

The subject of outliers was addressed briefly in Chapter 5. Within that context, outliers were seen as a hindrance – something that needed to be addressed or overcome. What happens, however, when these outliers have significance or meaning? In law enforcement and intelligence analysis, sometimes the most interesting aspects of the job are these outliers. These anomalies in the data can also be cause for significant concern.

10.4.1.1 How It Works

Briefly, there is a variety of clustering algorithms that group cases based on similarities between them.⁶ This clustering is something that is done in law enforcement and public safety on a regular basis. For example, homicides are grouped based on motive, while robberies can be grouped based on the location of the incident (e.g., street, commercial, bank), whether the suspect

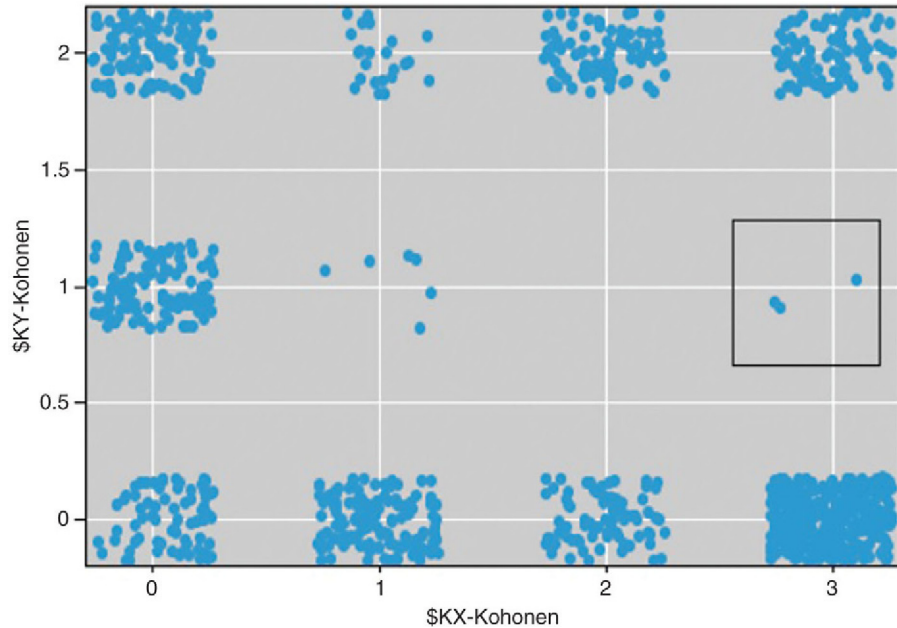


FIGURE 10.1

Results of a clustering algorithm. The box highlights three anomalous cases in the data set, which do not fit into any of the other clusters. *Screenshot taken by the author.*

was armed, or the value of what was taken (e.g., petit versus grand). What happens, though, when something is outside the norm, when it does not fit into any of these predetermined categories? These cases fall into the category of outliers, which can have a significant impact on model construction and evaluation if they are not identified and addressed. As shown in the framed area in [Figure 10.1](#), three cases do not fit into any of the larger clusters. These represent anomalies or possible outliers in the data, something that is difficult to evaluate until these cases have been examined in closer detail.

Sometimes, however, an anomaly represents more than just statistical clutter. Particularly in law enforcement and intelligence detection, anomalies often are cause for concern because they frequently indicate that something is where it does not belong, is doing something unusual, or has a potential for escalation. In general, deviations from normal in law enforcement and intelligence analysis indicate cause for concern and further evaluation.

Anomaly detection can have significant value in law enforcement and intelligence analysis and should be included in the core analytical functions. While automated detection systems can be wonderful, those without access to sophisticated software resources are encouraged to at least develop some

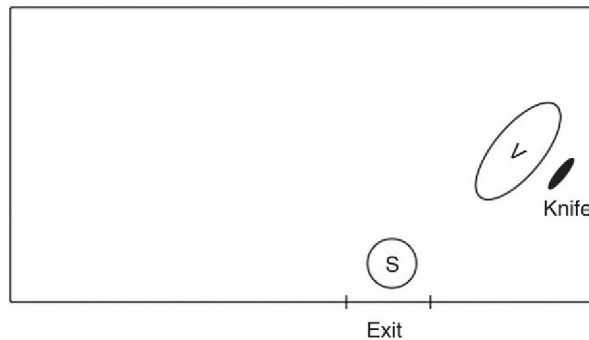
understanding of the “normal” crime within their purview. Many of the examples highlighted in other chapters were developed without access to sophisticated data mining software. Periodically running frequency distributions and characterizing crime trends and patterns using descriptive statistics can greatly increase an analyst’s ability to detect unusual or emerging patterns. These “brute force” techniques, while not terribly elegant, still get the job done when nothing else is available and should be included as an essential analytical function.

Outlined in the following sections are a few examples of the many potential uses for anomaly detection within the public safety setting. Again, knowledge of “normal” represents a very important component of the analyst’s acquired domain expertise because in the public safety environment almost any deviation from normal or expected is cause for concern and further investigation. Moreover, it is extremely important to have a valid baseline upon which these deviations can be characterized and evaluated.

10.5 STAGED CRIME

One group of individuals that frequently gets it wrong with “normal” is people who attempt to stage crimes. Again, having a good sense of what generally occurs in crime can be of enormous benefit to the analyst encountering something unusual. Strange things happen, but most law enforcement and intelligence professionals become suspicious when encountering too many coincidences or strange occurrences. Everyone “knows” how crime works, or at least think they know, but very few actually get it right.

The following example illustrates how a good understanding of normal criminal behavior can reveal a possible staged crime or false allegation. Several years ago, a call came in advising that an older gentleman had shot and killed his caretaker. The suspect reported that he had to shoot his caretaker because she had grabbed a knife and attempted to stab him during an argument. Almost immediately the story was suspicious. As can be seen in [Figure 10.2](#), the suspect was standing next to the door when he shot the victim. Not only was he close to an exit, but he had cornered the caretaker, effectively blocking her escape. People occasionally make unusual decisions when involved in a violent confrontation, but choosing not to flee the aggressor, as the caretaker was portrayed, would be unusual. This is particularly true given the proximity to the exit and the relative ease with which the suspect could have escaped this particular situation and gone for help. People lie, however, especially if they believe that it will cast them in a favorable light. Perhaps the suspect had not been completely honest regarding his role as an unwilling participant in the argument, fearing only for his life when faced with the attacking caretaker.



S – Suspect

V – Victim

FIGURE 10.2

Scene diagram of the staged crime described in the text.

In most situations, it takes at least two for an argument. It is entirely likely that the argument had been more reciprocal than the gentleman had originally conveyed in this statement; however, something still felt strange.

Review of the crime scene was consistent with the story. There was a knife on the floor next to the victim. Further examination of the scene, though, revealed that the knife was on the right side of the victim's body, next to her left hand. Left-handedness is relatively infrequent within the population, though, and an easy mistake in staging a crime would be to take a knife in the right hand and drop it on the right side of the victim, not realizing that it would be next to her left hand – an error in logic. As it turned out, the suspect had staged the scene, planting the knife in a hurry without consideration for the handedness of the victim.

Sometimes things just look too good, too consistent, or too homogeneous. Again, one tipoff in the embezzlement case described earlier was that the frequency of checks written for whole dollar amounts was extremely high. This was especially unusual given that the checks ostensibly were written to pay bills and reimburse other routine expenditures. When we ran a frequency distribution on the amounts, the pattern became even more unusual: The number of checks written for exactly \$100 was unusually high, particularly compared to what we would expect for usual expenses.

10.5.1 What Are the Odds?

Things that occur with unusually high frequency also can be suspect. In 2005, a college coed staged her disappearance shortly after she reported being the victim of a rather unusual abduction. Abductions by strangers are extremely rare. Therefore, one might ask what the likelihood is that the same individual

would be abducted by a stranger twice within such a short period of time – particularly abductions associated with unusual circumstances. While not definitive indicators of wrongdoing, most investigators and analysts tend to become extremely suspicious when encountering coincidences like these.

Again, most analysts and investigators have developed a keen sense of what is “normal” and expected in crime and criminal behavior through experience. These internal norms can be confirmed and characterized through the use of data mining and predictive analytics, and the resulting models can be used to identify more subtle patterns of deception in future cases.

Similarly, through the discovery process, these definitions of “normal” can be expanded to include additional, unexpected features. In its native state, crime frequently deviates from what we come to expect from its portrayal in the popular media. Those trying to simulate a particular crime frequently base their manipulations on what they think crime should look like, which may deviate significantly from what actually occurs. Knowledge based on ongoing review and analysis of crime trends and patterns can reinforce internal norms while enhancing existing domain expertise for the analyst, as well as for the investigator. In combination, the characterization and modeling of normal crime and behavior represent an extremely valuable tool in the investigative arsenal, particularly when criminals attempt to manipulate the investigative process in an effort to elude detection or misdirect an investigation.

CAN THIS BE USED AGAINST US

“Observe your enemies, for they first find out your faults.”

Antisthenes

One concern in writing a book like this is that the bad guys will buy it in an effort to learn something new that will give them the upper hand. It is not unusual for criminals to study police methods and procedures – perhaps as a means to stay one step ahead of law enforcement, or perhaps to enhance the thrill of the chase. More recently, we find that our adversaries in the war on terror have been studying our methods and tactics.⁷ We study their tactics, methods, culture, and goals. Why should it be surprising that they do the same? The issue of “normal” behavior and “normal” crime is one of the things, however, that makes me relatively confident that the information contained within this book will not provide an unfair advantage to our adversaries. We can characterize crime and criminal behavior forever, and certain common features, trends, and patterns almost certainly will be identified. On the other hand, the more I study behavior, the more I continue to be impressed by the subtle fluidity and adaptation that emerge over time. In many ways, change and transition truly are the status quo. Often I have been totally surprised to find out what “normal” is in a certain situation. Whether examining routine traffic on the Internet or how criminals typically function, I almost always have been perplexed by at least a few things that could not be explained readily. While this makes life interesting, it also gives me a relative degree of confidence that it would be difficult to slip through undetected because, ultimately, “normal” is so strange.

Bibliography

1. Beck C, McCue C. Predictive policing: what can we learn from Wal-Mart and Amazon about fighting crime in a recession? *Police Chief* November, 2009.
2. Casey E. Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and Internet crime investigation. *Knowledge Solutions*, www.corpus-delicti.com/case_based.html; 2002.
3. McCue C, Smith GL, Diehl RL, Dabbs DE, McDonough JJ, Ferrara PB. Why DNA databases should include all felons. *Police Chief* 2001;68:94–100.
4. McCue C, Smith GL, Diehl RL, Dabbs DE, McDonough JJ, Ferrara PB. Why DNA databases should include all felons. *Police Chief* 2001; 68: 94–100.
5. Brown DE. The Regional Crime Analysis Program (RECAP): a framework for mining data to catch criminals. University of Virginia; 1998.
6. Helberg, C. *Data mining with confidence*, 2nd ed. SPSS, Inc., Chicago, IL; 2002.
7. Papyrus News (2002). Fourth-generation wars: Bin Laden lieutenant admits to September 11 and explains Al-Qa'ida's combat doctrine. February 10; <https://maillists.uci.edu/mailman/list-info/papyrus-news>; 2002.

Behavioral Analysis of Violent Crime

“Outside of the killings, Washington has one of the lowest crime rates in the country.”

Marion Shepilov Barry, Jr.

Although the above quote might seem humorous or absurd, it makes a good point. In many ways, the key to reducing the crime rate is *preventing* crime. To effectively prevent crime, it is important to characterize and understand it so that it can be anticipated and influenced. Therefore, the use of data mining and predictive analytics to characterize and predict crime represents the first steps in the preparation of a comprehensive, information-based approach to enhanced investigative efficacy and meaningful, targeted crime prevention.

For better or worse, people generally develop their impressions of crime in a community based on the number of murders or on the overall violent crime rate (Figure 11.1). The ability to reduce community violence can reap tremendous benefits in terms of quality of life; rampant violent crime can impact almost every aspect of life in a community. Aside from the immediate, direct impact on the victims and their families, the overall quality of life, including economic opportunities, decreases significantly in a crime-ridden community. As the cycle of violence spins out of control, those residents that can, leave; existing businesses relocate; and new opportunities for development are lost. Even the young people are affected.¹ Those growing up in urban combat zones (Figure 11.2) acquire the short-term approach to life so frequently observed in people with diminished perceptions of their value to the community, reduced access to opportunity, and shortened overall life expectancy.

Several years ago, I began a conversation with some colleagues who possessed exceptional skills in the behavioral analysis of violent crime. The basic thesis of this discussion was that the behavioral analysis of violent crime or criminal “profiling” works because violent crime is relatively homogeneous and predictable.



FIGURE 11.1

Homicide rates can alter perceptions of the relative safety and associated desirability of a community.

Violence is a behavior, albeit an extreme behavior, when reduced to its fundamental form. Social and behavioral scientists have been studying and categorizing behavior for many years and have found that in many ways, nature is economical. Biologists frequently find the same common elements repeated within an organism and across species. In other words, one of nature's rules seems to be, if it works, keep using it. Complexity often is achieved by unique combinations of simple common elements.

This conservation mechanism also can be applied to behavior. In many ways, the behavioral analysis of violent crime involves describing, aggregating, and categorizing behavior, similar to the behavioral categories or taxonomies developed and employed in the laboratory setting. Many of the same behavioral concepts developed in pigeons or rats can be applied to humans. While this comment might prompt a few uncharitable comments from those who work with criminal behavior on a daily basis, the concept of behavioral reduction can have profound implications for the use of data mining and predictive analytics in the analysis and prediction of criminal behavior, particularly violent behavior, as well as victim selection, victim response, and victim–perpetrator interactions.

People tend to get into ruts with regard to their behavior, which is the basic foundation for the routine activity theory of crime.² By way of example,



FIGURE 11.2

Many communities inundated with drugs and violence resemble urban combat zones.

a systematic review of drug-related homicide data in Richmond, Virginia, revealed that the victims of drug-related homicides generally did not get killed on the opposite side of the river from where they lived. We have conducted similar analyses since that time and have found a relatively consistent pattern of results. People still tend to get killed over drugs on the same side of the river where they reside.

In some ways, this should not have been a particularly surprising finding. The Richmond metropolitan area is divided by the James River. While people frequently will cross the river for work and work-related functions, they generally tend to stay on the same side of the river as their residence for most of their other routine activities. Additional analysis of the illegal drug markets in the city revealed a similar pattern. At the time of the analysis, there were drug markets that had evolved to serve the population from each particular side of the river. They tended to be in locations that were convenient for the particular market or clientele being served, but a quick review of vehicle identification stickers revealed very little crossover, particularly among users. Perhaps the one noteworthy exception to this finding was that dealers and other individuals involved with the illegal networks were killed on the other side of town, possibly because they did not belong in that area.

What does this mean? First, it has some implications for drug enforcement strategies. If people were not crossing the river and getting killed buying drugs,

a logical assumption would be that they did not cross the river to buy drugs. These homicide data, in many ways, provided a snapshot into illegal drug buying habits in the city. Triangulating between the residence of the victims and the locations of the murders also provided some insight regarding normal traffic patterns and routes associated with the various drug markets. This information has considerable value for drug enforcement strategies.

What about folks working outside of the Richmond metropolitan area? What does this mean for them? Reducing this example to its common elements reveals three features that can be applied to other settings. First, it highlights the use of the discovery and modeling process embodied within data mining and predictive analytics. Although using very simple techniques, this example demonstrates that we can learn new things about violent crime through the process of characterization and categorization. By drilling down into the information, new trends and patterns can be revealed or discovered. Moreover, this information can be used to anticipate and predict future similar events, which opens the door to meaningful enforcement and prevention strategies.

Second, humans are creatures of habit. If I can buy milk, watch a movie, and get my car washed near my home, why should I cross the river to buy drugs? Drug markets can be extremely adaptable and fluid, particularly when responding to the preferences of the users. In fact, it is really amazing to see the differences between some of the inner-city open-air drug markets and those serving folks from the outlying areas. During various tours of the illegal markets, the differences in structure, setup, and function were noteworthy. At that time, the dealers associated with the inner-city markets tended to be savvier, able to quickly distinguish the unmarked vehicles – frequently after young children on bicycles called out, “Five-oh!,” heralding our impending arrival.³ The other markets, on the other hand, resembled a fast-food restaurant in terms of convenience and product availability, as well as setup and function.

The third benefit of this example is that it highlights the use of data mining and predictive analytics in the analysis of violent crime. If nothing else, this example should encourage other agencies to step outside the box in the analysis of violent crimes. Simply “counting crime” is not enough. To prevent crime, we need to be able to anticipate and predict it. This extremely low-tech yet powerful example highlights the value of creative analysis. It also further supports the use of spatial analysis or mapping. As illustrated repeatedly throughout this text, mapping should not be confined to pin maps showing what happened. Maps can and should be used to elucidate novel insight and convey additional details or information regarding crime trends and patterns of interest.

11.1 BEHAVIOR 101⁴

Behavior or the analysis of behavior merits its own chapter because ultimately all of the analysis, math, and algorithms described in this text go back to behavior. In public safety and national security analysis, the revealed relationships, affinities, transactions, sequences, patterns, and trends all reflect behavior – usually, bad behavior. Therefore, some understanding of normal behavior, abnormal behavior, and the associated challenges with analyzing it should be addressed.

Signature and Modus Operandi (MO) represent foundational organizing concepts in the behavioral analysis of crime.⁵

11.1.1 Modus Operandi

Simply, these are the actions required to commit the crime. This may change over time, particularly as the suspect becomes more adept, and may incorporate learning through direct experience and/or observed or shared with other offenders.

11.1.2 Signature

Signature includes behavior that extends beyond the direct requirements to commit the crime, and may include elements that are psychologically important to the criminal. Unlike MO, the signature will remain fundamentally stable and will not change in any sort of meaningful way over time.

11.1.3 Case-Based Reasoning

In many ways, the process of criminal investigative analysis, or behavioral profiling of violent crime, truly is an amazing example of data mining and predictive analytics. At first, the process seems like magic. When confronted with a jumbled mass of clues, the investigator is able to identify a likely suspect, like a magician pulling a rabbit out of the hat. On further investigation, though, it becomes clear that there are two important elements functioning in this investigative process. The first is good case management. By reviewing the case thoroughly, items that have been overlooked or lost can be identified and addressed. Similar to identifying where missing puzzle pieces belong, this process in and of itself can result in tremendous clarification in the direction of an investigation and significant progress toward closing the case. It allows the investigator to reveal the big picture and identify any readily apparent trends and patterns. In many ways, this is similar to the process of data cleaning and management. By thoroughly reviewing and organizing the case, missing details or data are identified and, if possible, addressed.

Similarly, using data mining and other automated methodologies can prompt consideration of evidence in a different light. By distilling the collected narrative information through a categorical filter, the information can be further characterized and categorized into smaller, readily identifiable groups that have value from an investigative perspective. How does this work in the field? Signature and MO features, as well as many other factors considered during investigation can be reduced to binary “yes/no” decisions, or other categorical variables. For example, it does not matter whether a victim was stabbed 5 times or 50 times, if the number of wounds exceeded what was necessary to kill the victim then it represents overkill, which is a “yes/no” decision. Using this type of approach, decision trees can be developed that guide the investigative process, which ultimately can be used in motive determination and models of likely suspect attributes. These and other data management techniques also can add more structure and meaning to the collection and organization of investigative information.

The second key piece in the behavioral analysis process is pattern recognition. While this might sound easy, any good investigator knows that it can be much more difficult than it sounds. Investigators who are particularly good at this task are those able to move beyond some of the obvious details and identify underlying themes and patterns. In addition, they will begin trying to match these elements in the current case to their internal database of previously solved cases in an effort to identify a common theme or pattern that might give them some insight. This process of comparing current evidence against historical memory to see what fits can be described as case-based reasoning. Case-based reasoning is a model of learning in which people comprehend new experiences within the context of previous ones,⁶ a process that can be modeled using expert systems or artificial intelligence.

In some ways, a good investigator can be thought of as a closet analyst. The process and approach associated with criminal investigations is one of the most analytical aspects of law enforcement. Attributes that are associated with a good patrol person or an excellent member of the SWAT team are very different from those factors that are associated with a good investigator. The comparison between the *Hawaii Five-O* and *Columbo* detective shows provides a nice parallel. McGarrett and Danno are out there in the thick of things, catching bad guys, while Detective Columbo seems to be stumbling around in a clumsy fashion. Yet the viewer can almost see Columbo’s wheels turning and watch him mentally chewing on the evidence as he chews on his cigars, trying to make the pieces fit in an effort to understand the puzzle. In addition, Columbo always was “bothered by” the pieces that did not fit, highlighting gaps in an alibi or oddities in the crime scene. In many ways, Columbo is a walking example of data mining, case-based reasoning, and anomaly detection rolled into one baggy trench coat. Again, not very elegant, but it works.

Cold case investigation can be seen as an extreme manifestation of the analytical side of the investigative process. These very special investigators and analysts are truly unique in their ability to move beyond the superficial nature of the evidence and look for the underlying form or commonalities with other solved cases in the investigator's repertoire. While this may sound almost like a Zen-like approach to investigation, much of it comes back to excellent case management, anomaly detection, and some superb case-based reasoning.

Replicating this process through the use of computer programs or artificial intelligence algorithms certainly confers some benefits. First, a computerized database has the ability to store and quickly recover a large amount of information, with an associated capacity that extends well beyond the memory storage and retention abilities of a human. Expert systems have the ability to store, retain, and simultaneously consider information from more cases than any single investigator is likely to encounter in his or her entire career. In addition, the information is not subject to the same memorial decay and errors in interpretation over time that can occur with the human memory.

Second, an expert system theoretically has no bias – bias that can favor a particular scenario and color the course of future evidence collection and investigative process. It is not swayed by the suspicious boyfriend, or by the shifty looks and dubious alibi given by the jilted lover. Similarly, it is not likely to even subtly discard or overlook evidence that is contrary to a favored hypothesis or outcome. This certainly is not to suggest any lack of professionalism or ethics on the part of sworn investigators. They are some of the most honor-bound, committed professionals in law enforcement. Rather, human investigators bring their internal norms, life experiences, and feelings to every case. While this can make them incredible champions for the victims and their families, it also shaves away some of their objectivity. It is almost impossible to look at the face of a dead child and remain objective; however, even a little compromise in objectivity can be associated with a concomitant decrease in efficacy. Computers, on the other hand, do not care.

Finally, expert systems are not bothered by crime scene details; they can get beyond the "yuck" factor that occasionally can catch even the most seasoned investigator. Certain victims and certain scenes affect investigators in different ways. Those that are affected by everything that they see are limited in their ability to effectively investigate violent crimes. On the other hand, those who do not respond to anything probably have been on the job too long and should consider a change of pace in an effort to remain effective and retain their humanity. In short, even the most staid investigators are likely to be bothered by something that they encounter. This limits their objectivity and can compromise their ability to reduce the crime to common elements that can be characterized and compared to previous cases. Artificial intelligence systems

truly embody the philosophy of “just the facts, ma’am” because they have no capacity for anything else (at least at the time of this writing).

Does this mean that we should discard detectives in favor of computer modeling programs and expert systems? Absolutely not! Expert systems, no matter how “smart,” lack one element critical to effective and meaningful data mining – domain expertise. Without an “expert” in the field to evaluate the nature of the evidence and information collected and to evaluate critically the value and validity of any created models, the risk for significant errors in logic and interpretation would seriously limit our ability to use these tools. Moreover, it is absolutely essential that professionals within law enforcement and public safety have significant involvement in the analytical process. The best scenario would be one in which analytical and investigative personnel work together in the data mining and modeling process, perhaps with some outside help and support from statistical experts. It bears repeating that it seems to be relatively easy to teach data mining and predictive analytics to law enforcement personnel. These folks know where the data come from and how the models will be used. They intuitively know what is available for models and when. As such, they are much less likely to make critical errors that result in the creation of a model based on circular logic, which essentially requires information that is dependent upon the output as an input or indicator variable. Many investigators are natural data miners, given the intuitive nature of their work, and seem to embrace the approach when given the opportunity. On the other hand, trying to convey the internal norms, historical knowledge, and accumulated domain expertise from an active investigative career to data mining experts has proven to be extremely difficult.

In sum, data mining and predictive analytics can enhance the investigative process, particularly through many of the automated pattern recognition programs and scoring algorithms. However, the use of expert systems alone has limited value and could significantly compromise the investigative process.

WHERE DO MURDERERS COME FROM?

“I actually think I may be possessed with demons, I was dropped on my head as a kid.”

Dennis Lynn Rader, the “BTK” Killer

The “nature versus nurture” question has swirled around in behavioral biology through several generations of scientists at this point, and the most reasonable answer seems to be that human development probably incorporates a little bit of both. One area where this question has some particular urgency, however, involves juvenile murderers. Many people look at a particularly heinous crime that has been committed by an individual who meets both the legal and chronological definitions of a child and wonder just where this individual came from to be able to commit such a heinous crime. In the course of my research, I have reviewed more than a few cases that fit these criteria and have been impressed again by the relative degree of behavioral homogeneity between the crimes committed by young people with relatively limited access to information regarding

their chosen field of criminal expertise, and those committed by others who seem to have an abundant source of examples and mentoring. Beyond the impulsive juvenile murderers who kill secondary to the commission of another felony, there lies another group that has taken murder to a level that seems to truly transcend their age and relative amount of life experience. As a result of this anecdotal experience, I have tended to informally subdivide juvenile murderers into two groups: those who learn how to kill, and those who seem to have an intuitive sense or need to kill.⁷

The first group generally uses violence or murder to achieve some sort of secondary objective. For this group, violence frequently is a means to an end; a way to achieve a particular goal. This type of juvenile murderer is especially prevalent in illegal drug markets. For example, in illegal drug markets, violence frequently is used to enforce rules and norms, particularly as there is limited access to legal enforcement mechanisms.⁸ In other words, if Bob sold Joe some bad dope, Joe generally could not expect to receive much help from the economic crimes unit at the local police department. It is not unusual for these offenders to commit multiple murders, and even use postmortem mutilation or positioning to send a message to the community; behaviors and practices frequently associated exclusively with serial killers. However, killing for these youthful offenders is a means to an end. Any additional behavior or manipulation of the victim's remains often represents a punctuation mark to the underlying message sent.

This group seems to have acquired their skills through a process of social learning. These kids learn by watching others commit violent crimes and use violence to achieve secondary goals and objectives. This also relates to the idea that drug selling, like law enforcement, is a 24/7 profession. If one is to succeed in the extremely predatory world of illegal drug sales, then it is important to convey a sense of power and strength in every life domain. It would be extremely dangerous to be perceived as weak in a social setting, for example. This has been described previously as the "outlaw" lifestyle, and it can be linked to some murders involving those linked to illegal drug markets but not directly to a drug-selling transaction. Drug-involved violence, in particular, has been studied and characterized in some detail, which forms a great foundation and framework for the use of data mining and predictive analytics in the analysis of violent crime.

Consistent with the social learning model, Bennet, Dilulio, and Walters wrote a book entitled *Bodycount*.⁹ At the time of its publication, their concept of a juvenile "superpredator" received a tremendous amount of press. At first, the concept of a superpredator seemed alien, almost offensive. He was almost like a "robocriminal," relatively automatic, with slim prospects for rehabilitation. Over time, however, as I encountered more young killers like those involved in illegal drug markets, it became clear that there were some kids who had been changed significantly and perhaps irrevocably by their environments.¹⁰ Ongoing and repeated insults during critical periods of emotional and moral development had changed their view of the world and related approach to life to the point where they employed different rules of the road as they negotiated the twists and turns on their life path. That is why I am so sad to see young children in the arms of adults at crime scenes, because these events often represent the first steps in the social learning process that ultimately turns out juvenile murderers and victims in our urban combat zones. Unfortunately, this behavior is not confined to the United States, and subsequent observations in postconflict Iraq¹¹ and other locations reveal.¹²

The second group of juvenile murderers just seems to like to kill, whether to fulfill a need to attain the ultimate power over another human being or to gain the opportunity to engage in unlimited exploration of the human body. This group is particularly intriguing, given the intuitive sense that they seem to have for what they want or need to do and the relative homogeneity of their behavior. Their intuition frequently is associated with very little outside input, although this has been changing in recent years.

It is interesting to study historical cases of serial sexual homicide, many of which were used as the basis for the creation of criminal investigative analysis, and note the eerie similarities between many of the crimes. The relative degree of homogeneity between the crimes committed by these different individuals is uncanny, particularly given the lack of contact between these individuals and the relatively limited public knowledge and understanding of this behavior even a few years ago. Similarly, the ability to take these concepts and apply them successfully to groups outside the U.S. underscores the relative homogeneity of seriously violent, predatory behavior; particularly as it seamlessly transcends culture, language, rule of law, and national identity. The fact that this behavior is so similar and predictable that it can be used to enhance the investigative process really begs the question, "Where does violence come from?"

Again, by using data mining and predictive analytics, it is possible to transcend human bias and opinion in an effort to reveal the underlying elements of a crime, prepare a strategic characterization of the likely offender, and close cases that formerly challenged even the most seasoned investigators.

11.2 MOTIVE DETERMINATION

Frequently in an actual investigation, the nature of the victim–perpetrator relationship is unknown. It is the crime scene characteristics and victim lifestyle factors or “victimology” that suggest a possible motive, which then is used to identify possible suspects. In fact, it is this type of characterization of likely suspects that is embodied in the behavioral analysis process.

Some of our earliest work using advanced statistics for characterizing violent crime involved the development of automated motive determination models.¹³ Again, this work arose out of some lively discussions regarding whether it is possible to accurately model violent crime using automated methods. The cases used for this study included 25 juvenile murderers incarcerated in the Commonwealth of Virginia Juvenile Correctional Centers from February 1992 to July 1996. Information pertaining to the victims, suspects, injury patterns, and the behavioral characteristics of the assault were identified and analyzed in an effort to determine whether it was possible to determine the motive using automated methods.

For the first analysis, all of the information that could possibly be obtained was put into the model. The analytical approach selected was discriminant analysis, which is a classification or supervised learning algorithm. This analytic approach is covered in greater detail in Chapter 7; in brief, discriminant analysis is a multivariate statistical approach that can be used to identify factors that are useful in determining group membership. Generally, one of the assumptions with the use of discriminant analysis is that the variables used to create the model are continuous. While some of the data available for this study were continuous (e.g., age), most of the relevant information was either binary or categorical. This assumption could be violated with a certain degree

of confidence given the relative strength of the algorithm and the nature of the errors likely to occur. Specifically, the type of error more likely to occur if the assumptions are violated with discriminant analysis is a failure to find a relationship in the data even though one may exist.¹⁴

In crime and intelligence analysis, it is almost always better for the analysis to come up empty than to identify a spurious or false relationship in the data. I mention this point not to suggest that the analyst should habitually violate rules and assumptions associated with modeling algorithms. Rather, I wish to highlight two key points about predictive analytics and the associated modeling algorithms. First, some of the rules and assumptions associated with these techniques are more important than others, and it is possible to exercise some discretion with the statistical algorithms. And second, these tools are designed to identify and model relationships in the data. The type of error most likely to occur is a failure to identify a relationship when one actually exists. While this may be frustrating and even limiting if the analyst is being asked to provide information-based support for a particular operation or investigation, unreliable, inaccurate, or spurious findings generally carry a far greater risk to public safety in most situations. The issue of errors and how they should be evaluated is addressed in Chapter 4; however, in a motive determination model, any error that would misdirect an investigation could potentially be very significant as it would misdirect or waste investigative resources. Moreover, delay in the investigative process could seriously compromise the ability to solve the case, ultimately allowing the suspect to continue offending. Therefore, failure to identify a model due to the violation of the data type assumption was determined to be acceptable after thoughtful consideration of the operational requirements.

The initial results were extremely promising. Using information related to recent victim drug use and suspect substance use history, it was possible to accurately categorize 85% of the cases as drug-related or not drug-related. One interesting finding regarded the direct relationship between the suspects' use of illegal drugs, particularly marijuana, and their involvement in a drug-related homicide. Our earlier work had confirmed conventional wisdom among narcotics detectives: Most successful drug dealers do not use the drugs that they sell.¹⁵ In fact, substance use has been associated with an increased risk for firearms assaults among drug sellers. Whether our finding was a cause or a consequence was not entirely clear, but substance use was not a healthy choice, particularly among those involved in the sale and distribution of illegal narcotics. Therefore, this result was somewhat surprising. Further examination revealed that the suspects' substance use generally involved marijuana, while the recent victim use included cocaine and opiates (e.g., heroin). This finding was consistent with a model of drug-related violence that proposed different subtypes of drug-related violence associated with different types of suspects,

similar to a division of labor within drug distribution networks.¹⁶ In particular, the suspects who emerged in this preliminary study were very similar to the “enforcers” described in this model.

The end result of this preliminary study was support for the notion that violent crime could be modeled using advanced statistics, as well as additional knowledge regarding our understanding of drug distribution networks and the proposed division of labor. From an investigative standpoint this was somewhat helpful in that it offered additional information pertaining to the types of individuals likely to commit drug-related violence; however, it provided very little in the way of enhanced investigative efficacy. In other words, the results of this study indicated that recent victim use and suspect use were good indicators of drug-related violence. From an investigative standpoint, though, this creates circular logic: To identify the suspect, it is helpful to know the motive; to determine the motive, we need information regarding the suspect’s substance use; to know the suspect’s substance use habits, we need to know who the suspect is, which is the original question. Clearly, further work was needed.

More recent work in this area has been confined exclusively to the information that is available early in an investigation. Again, drug-related violence represents a good area of study for several reasons. First, drug-related violence frequently drives the overall violent crime trends in many communities plagued by serious increases in violent crime¹⁷; the homicide rate often rises and falls as a direct consequence of the drug-related homicide rate. Therefore, addressing drug-related violence can significantly reduce the violent crime rates in these communities.

Drug-related crimes, particularly violent crimes, can be difficult to solve in a timely fashion. Witnesses may be reluctant to come forward or may be unreliable. In some cases, drug-related violence is seen as “the price of doing business,” reducing sympathy for the victim and making others less likely to get involved. The more time that elapses, however, the less likely it is that a murder will be solved. If significant progress is not made in the first few days of an investigation, it becomes increasingly unlikely that the case will be solved. Finally, and perhaps most importantly, drug-related violence appears to be relatively homogeneous and amenable to modeling, which makes it a good candidate to evaluate automated motive determination scoring algorithms.

Using classification techniques similar to those described earlier, drug-related homicides were analyzed in an effort to develop a model that could be used to automatically determine a motive using information available early in an investigation. Like the axiom in real estate, the relevant variables that emerged were location, location, location. While this is an overly simplistic interpretation of the results, the most salient fact to emerge from this analysis was that certain areas were associated with an increased risk for drug-related violence.

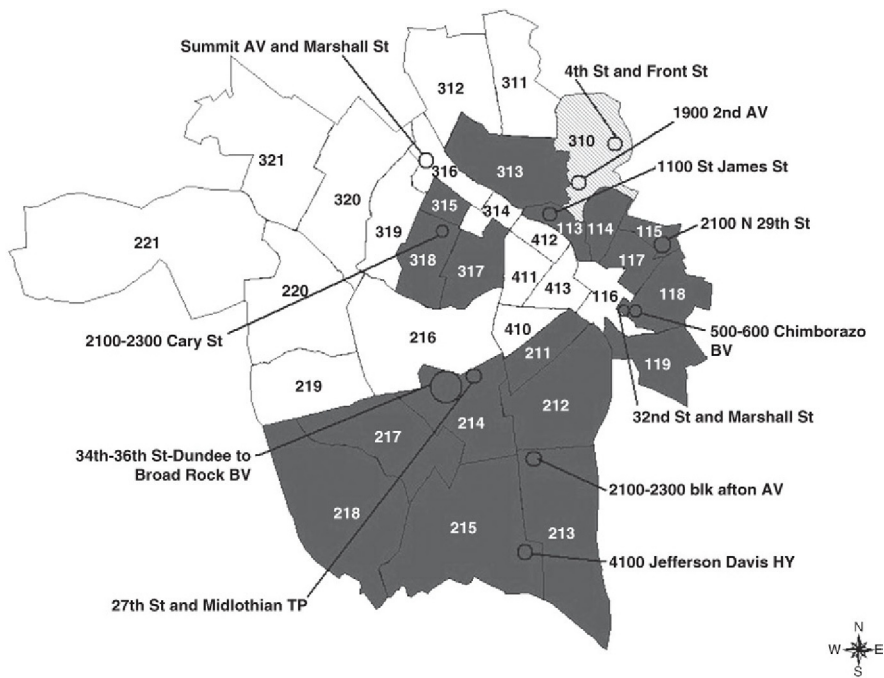


FIGURE 11.3

Drug-related violence models can be deployed through the use of mapping tools. These facilitate the development of proactive deployment strategies. This particular map highlights differences in victim characteristics, which might have additional value from an operational perspective. Characterizing victim risk or “victimology” can be used to create operational plans that directly address specific patterns of risk.

The results were distilled and deployed through maps (Figure 11.3), which allowed for proactive, risk-based deployment specifically targeting drug-related violence. One interesting finding that emerged with the maps was that one particular area was associated with victims who were more likely to be employed. Further analysis of the location revealed a drug market frequented by users from the surrounding localities, which was consistent with employed victims. While this initially appeared to be nothing more than an interesting factoid, the operational implications quickly became apparent. Drug violence associated with employed victims was consistent with victims who were buyers. One way to reduce the violence in these areas was to reduce the number of potential victims. Demand reduction approaches offered the perfect solution to this finding, while different approaches could be used in other markets associated with different types of drug-related violence.

When the time of day, day of the week, and victims’ employment status were included in the analysis, the accuracy of the model increased concomitantly. Again, this model included only that information available early in an investigation;

however, when the new information was added to the model, it became too opaque to have any value in the deployment process. Therefore, as described in Chapter 9, these results were deployed through a web-based package, which provided for 24/7 analytical capacity.¹⁸ As most homicide detectives can attest, people generally do not kill each other when it is convenient for us, particularly when drugs are involved. Most analytical teams, however, work normal business hours. By deploying algorithms through these web applications, operational personnel gain access to analytical support when they most need it – early in an investigation, when the case is progressing most rapidly.

The opportunity to more fully characterize crime and tailor specific crime reduction approaches based on a thoughtful analysis of the problem is one powerful benefit of using data mining and predictive analytics in the development of operational strategies. Just as we cannot please all of the people all of the time, there is no universal “crime reduction” strategy that will work for every situation all of the time. While this seems entirely logical, it is very difficult to even begin to match an appropriate operational plan to a particular issue if the nature of the problem has not been characterized and defined. For those who would say, “Utilize, don’t analyze,” this approach offers a solution for evaluating operational approaches in a meaningful way – something that has become a nonnegotiable requirement as resources are increasingly limited.

11.3 BEHAVIORAL SEGMENTATION

Taking a cue from the business community, we know that customer populations tend to be relatively heterogeneous groups and that marketing strategies frequently are targeted specifically to a unique customer segment. For example, in marketing mobile phone plans, different advertising campaigns are designed to specifically target young people, families, and even senior citizens by highlighting different features and benefits determined to be of interest to these respective groups. Similarly, segmentation of crime based on type, nature, and motive can be invaluable to more specific characterization in support of enhanced investigative efficacy,¹⁹ as well as anticipation and influence.

This concept was applied to the analysis of Lord’s Resistance Army (LRA) incidents.²⁰ The LRA has been engaging in bad behavior in Africa for decades with literally tens of thousands of victims. LRA attacks, however, include a broad array of incidents that include looting, abduction, and killing. Not only do these different attack patterns present markedly different consequences to the victims, but the operational responses and even security considerations for individuals working in these areas differ broadly, based on the nature of the incidents and related threat. Alternatively, the operational requirements facing the LRA also differ markedly based on the nature of the incident. For example, looting provisions from an Internally Displaced People (IDP) camp is relatively easy as

compared to abduction, which brings unique operational demands associated with abducting, managing, and moving hostages. Therefore, segmenting LRA incidents based on the nature of the attack represents an opportunity to gain additional insight that can be used to support information-based approaches to prevention, thwarting, and response.

Following the general schema outlined in the Crime Classification Manual,²¹ the incidents were divided into looting, abduction, “incidental homicide,” and murder. Incidental homicide was defined as a homicide that occurred during another incident type (e.g., looting, abduction), and generally involved a homicide that was related to resistance to looting and/or abduction. “Murder” was the only exclusive category. Geospatial predictive analysis,²² which has been reviewed in Chapter 7, was used to create a model of “suitable” or likely locations for other attacks, which includes future attacks, unreported or “hidden” incidents, and displacement should the current area of operation become unfavorable.

Figure 11.4 illustrates the pattern analysis of all LRA incidents.²³ As can be seen, areas identified as being suitable for other attacks align well with locations of

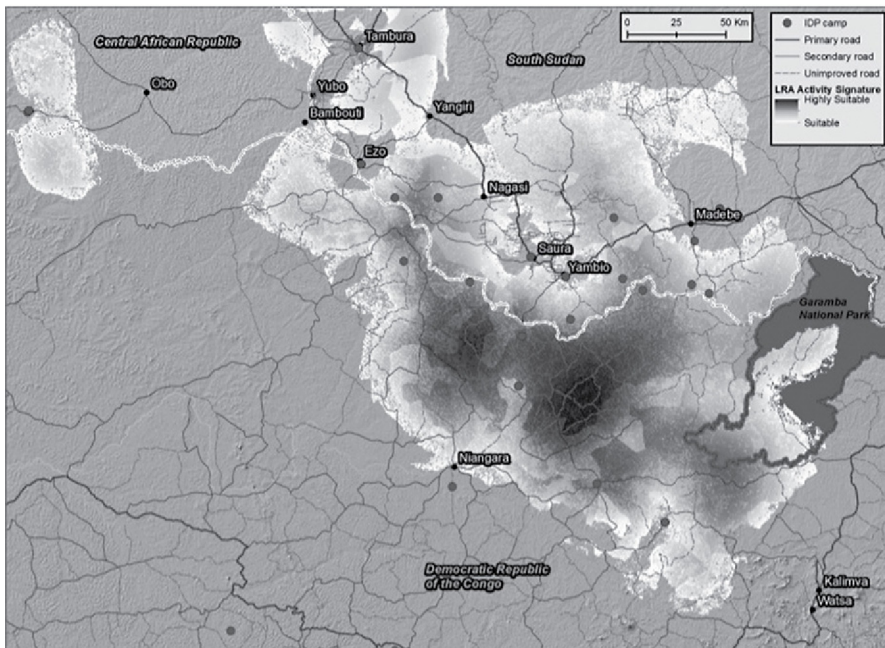


FIGURE 11.4

Pattern analysis of the Lord's Resistance Army (LRA) incidents in Spring 2011, prior to the insertion of security forces into the region.⁴³ Dots on the figure indicate the location of Internally Displaced People (IDP) camps. *DigitalGlobe, used with permission.*

previously known attacks, although areas in the Central African Republic also were identified as being suitable for other attacks. Again, the areas identified by the model include locations suitable for future attacks, as well as locations associated with unreported or “hidden” incidents, and locations for possible displacement. Dots on the map indicate the location of IDP camps, which represented a factor reliably associated with an increased likelihood for attack. Again, place preferences generally reflect access to potential targets or victims, and a favorable environment. The favorable environment may include factors that facilitate or otherwise enable, and/or avoidance of factors that may inhibit or otherwise thwart the behavior of interest.

Figure 11.5 depicts the results of the analysis of abduction incidents. As can be seen in the figure, the abduction incidents represent a distinct subset of the total incidents, confirming segmentation of these attacks as a unique pattern that differs from the aggregate sample. The locations of IDP camps have been indicated on the figure, which were also associated with an increased likelihood for abduction, which is not surprising. Given the requirement for

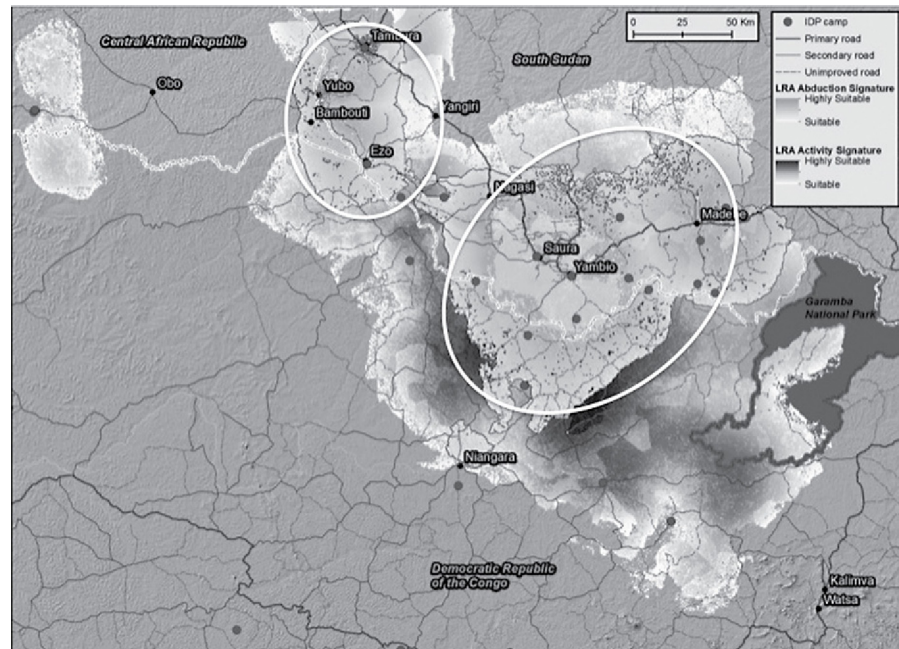


FIGURE 11.5

Segmentation of LRA abduction incidents as compared to all LRA incidents, depicted in Figure 11.6. The lighter areas of the map, which have been circled, identify locations suitable for abduction incidents and include future incidents, hidden or unreported incidents, and displacement. Dots on the figure indicate the location of Internally Displaced People (IDP) camps. *DigitalGlobe, used with permission.*

victim access, IDP camps represent a “target-rich” environment for abduction, replete with large numbers of particularly vulnerable victims. In addition, abduction locations also were found to be associated with local population centers, which further underscores access to possible victims as an attractor. Additional analysis of the results indicates that the geospatial distribution of abduction incidents tends to be more concentrated, which may reflect the fact that abducting, managing, and moving hostages is operationally challenging. These abduction incidents could be further subdivided based on the motive for abduction, which includes the use of abductees as porters to move looted goods,²⁴ forcible conscription, human shields, and sex slaves. Similar to the research on child abduction,²⁵ these different motives also are associated with differential outcomes for the victims, as well as different approaches to prevention and consequence management.

Figure 11.6 illustrates the result of incidental homicides. Again, by definition, these incidents overlapped with looting and abduction incidents to some

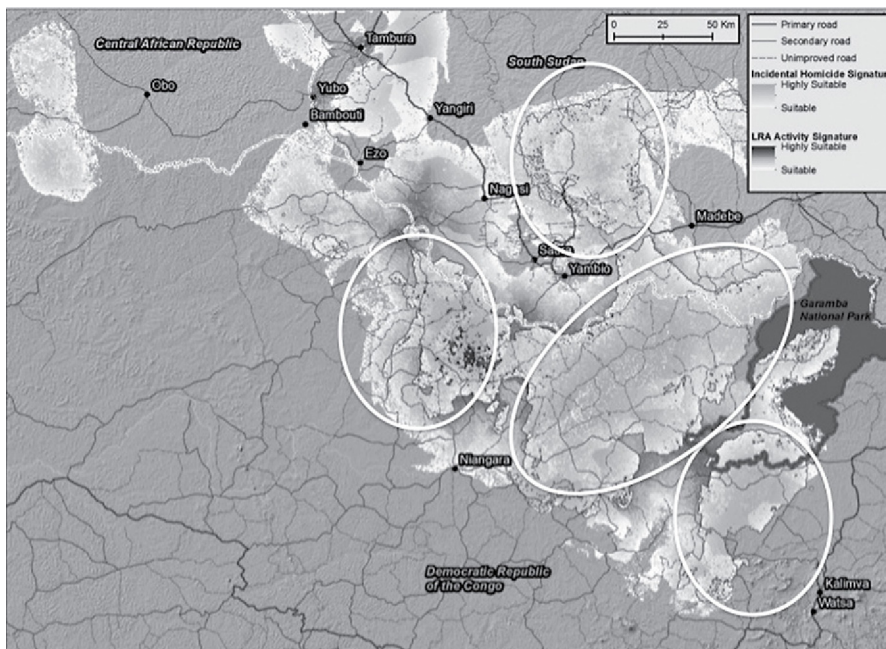


FIGURE 11.6

Segmentation of incidental homicide incidents as compared to all LRA incidents, depicted in the figure. Incidental homicides have been defined as homicides perpetrated in association with another pattern of attack (looting and/or kidnapping). The lighter areas of the map, which have been circled, depict locations suitable for incidental homicide incidents to include future, hidden or unreported, and displacement.

DigitalGlobe, used with permission.

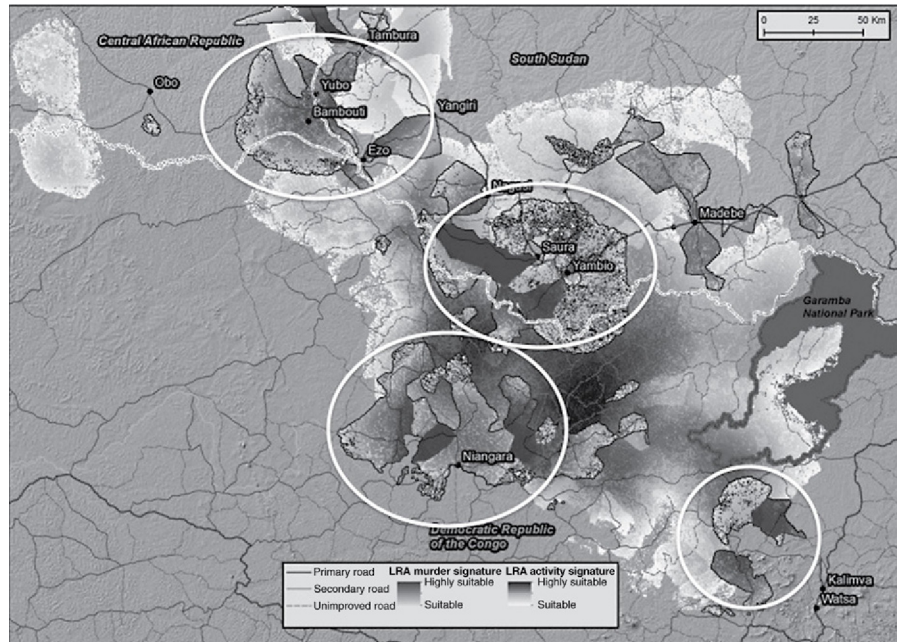


FIGURE 11.7

Segmentation of LRA-perpetrated murders as compared to all LRA incidents, depicted in [Figure 11.6](#). Murder incidents represented the only mutually exclusive category. The areas identified as being suitable for other murder incidents have been circled, and include future incidents, hidden or unreported incidents, and displacement. *DigitalGlobe, used with permission.*

degree, and tended to reflect resistance to the primary incident type (looting, abduction). Analysis of the murders ([Figure 11.7](#)), however, reveals a marked offset with the incidental homicides, confirming these patterns as separate and distinct. Reaching back to the work on drug-related violence,²⁶ we recall that violence can be used to enforce rules and norms. In this particular example, the murders associated with the periphery of the LRA area of operation or influence may reflect a similar use of violence as a means by which to establish and enforce boundaries.

Overall, the results of the analysis reveal different geospatial patterns and associated factors associated with the different incident types studied: looting, abduction, incidental homicide, and murder. Segmenting LRA incidents based on the nature of the attack provided a unique insight regarding the geospatial distribution and associated factors, which has significant implications for meaningful approaches to prevention, thwarting, response, and security. Of particular interest was the finding that incidental homicide and murder were almost completely offset suggesting that the LRA kills for very different reasons,

and that this was a meaningful and relevant segmentation of these incidents. Distribution of the murders in particular was consistent with the use of violence to enforce social rules, norms, and boundaries or territory.²⁷ While not perfect, this example underscores the potential value of operationally relevant and actionable segmentation of incidents based on behavioral attributes and offender primary gain. Similar to the “treatment matching” concept proposed for the drug-related violence example, insight regarding the nature of the incident or threat can be used to inform security decisions, prioritization of operations, prevention, thwarting, and response.

Perhaps as important, however, these results effectively illustrate core elements of predatory and violent behavior that may transcend national boundaries, rule of law, race, culture, religion, and language. The fact that analytic schema developed on western patterns of crime and criminal behavior represented a valid approach to behaviorally segmenting extremist behavior in Africa illuminates foundation level concepts and underscores the importance of behavioral analysis in effectively characterizing and segmenting violent behavior in support of informed characterization, anticipation, and influence.

11.4 VICTIMOLOGY

In some ways, victimology, or the study of the relationship between certain individual attributes or behaviors and the risk for violent crime, can be seen as a logical extension of risk assessment. Examining victim characteristics can be used in two ways. First, it frequently has value from an investigative standpoint because identification of victim lifestyle issues or risk factors often can suggest a possible motive and, by extension, a likely suspect.

On the other hand, some insight regarding potential risk factors associated with violent crime opens the door to meaningful, specifically targeted prevention strategies. Many victims of violence suggest that their injury was the result of “bad luck” or that they were merely in the “wrong place at the wrong time.” While this may be true for some, we have come to realize that certain victim attributes and behaviors can increase or even be related directly to an individual’s risk for violent victimization. For example, prostitutes are at increased risk for sexual assault and other violent crimes due to their involvement in that pattern of criminal activity. Similarly, substance users are at increased risk for violence related to their involvement with illegal drug markets. These frequently are referred to as “lifestyle” factors, meaning that some aspect of the victim’s lifestyle increased his or her risk for violence.

This is a very important point, because if violent victimization truly is related only to “bad luck,” then we are very limited in our ability to effectively prevent violent crime beyond just identifying the perpetrators and getting them off the

streets as quickly and for as long as possible in an effort to reduce crime. Crime prevention would rely almost exclusively on investigative efficacy. But if we can identify particular behaviors or activities that are linked to an increased risk for violent crime, then we have an opportunity to intervene and change those behaviors, which might reduce their risk for violent crime.

11.4.1 Victim Risk Factors

While it certainly is true that different people get killed for different reasons, what factors increase someone's risk for being the victim of a violent crime, and what does this have to do with data mining? Analysis of aggregate victim data generally reveals little, if anything, regarding specific victim risk factors. By drilling down into the data, researchers have found that one of the best predictors for violent victimization is involvement in criminal offense. However, this remains too broad a category to identify any specific risk factors. Therefore, it is necessary to parse the data even further in an effort to identify relatively homogenous groups that would be acceptable for modeling purposes.

By examining firearms injuries among juvenile offenders, certain patterns of associated risk could be identified and characterized. One attribute of particular interest was involvement with firearms. In particular, it was found that violent offenders who had been shot previously were more likely to admit to possessing a weapon, while juvenile drug sellers were much less likely to indicate that they carried a weapon. Additional information collected during this study, including information related to weapon selection and carry rates ultimately had implications for officer safety in the field.

Analysis of aggregate statewide injury rates provided additional insight, particularly as related to the important role that geospatial variables may play in risk. For example, when the data were analyzed by community, important differences began to emerge. In particular, juvenile drug sellers were much more likely to be shot in poorer communities than in those that were more affluent. Perhaps this was because there were more open-air drug markets in the impoverished communities, or perhaps robbery was the real motive. Regardless of the cause, this differential risk associated with specific community characteristics further illustrated the finding that victim risk represented a complex array of individual and community factors that interacted to determine the composite threat associated with a particular individual.

11.4.2 Child Abduction

Another area in which the characterization of victim attributes can indicate the likely suspect and even probable outcome is child abduction. In a series of excellent papers, members of the Federal Bureau of Investigation's National Center for the Analysis of Violent Crime have characterized this pattern of

offending in great detail.²⁸ Briefly, these researchers have identified a reliable association between victim age and gender and the likely offender, reason for the abduction, and probable outcome. Again, this type of victim characterization and modeling has tremendous implications for enhanced investigative efficacy in cases where investigative speed can be related directly to the likelihood of a good outcome.

This research also underscores the point that consideration of victimology and victim risk, or even lifestyle factors, is meant to suggest that the victim is somehow at fault. Rather, analysis of victim-related variables represents an attempt to better understand offender target selection and related preferences in an effort to focus the investigation, particularly early, as well as any related operational planning to include crime reduction programs.

11.5 VIOLENT CRIMES

Common patterns of violent crimes that the analyst may encounter are outlined next.

11.5.1 Homicide

At its most basic, homicide as a crime can be categorized and divided in many ways, most frequently based on victim–perpetrator relationship and motive.²⁹ Figure 11.8 depicts a possible decision tree for murder. The first branch of the

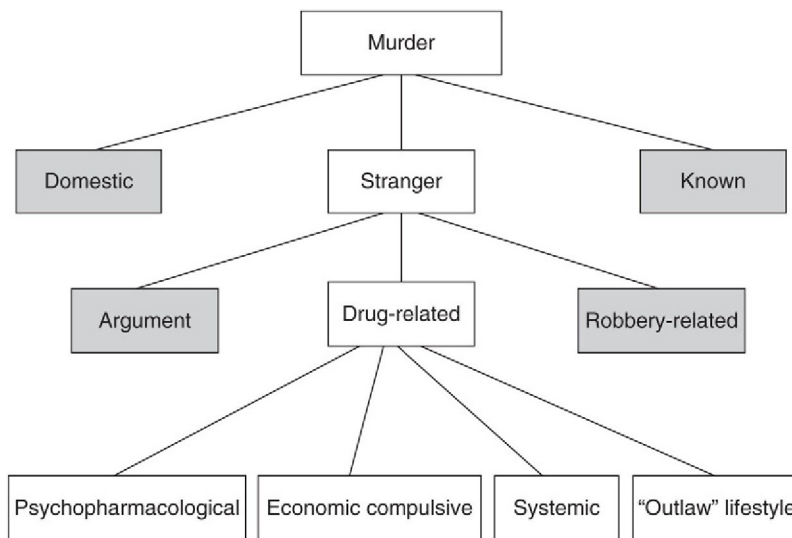


FIGURE 11.8

Possible decision tree for categorizing murder.

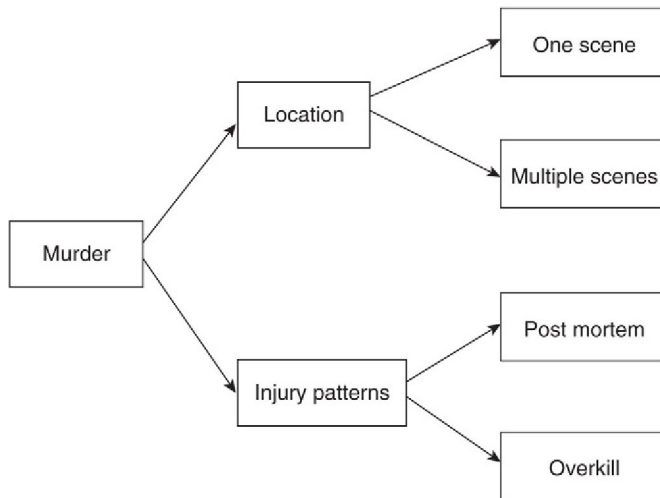
tree is divided by three possible victim–perpetrator relationships: familial or domestic, acquaintance or known, and stranger. The second division in the branches includes possible motives associated with stranger murder. Finally, within the various motives, it is possible to characterize the motive even further. In this case, the drug-related homicides have been divided using the Tripartite Model of Drug-Related violence developed by Goldstein.³⁰

Briefly, this model describes three types of drug-related violence: Psychopharmacological, economic compulsive, and systemic violence, which are related to the different roles that drugs and/or the drug-related lifestyle may play in violent crime. A fourth type of drug-related violence also should be considered, which has been described as the “outlaw lifestyle.”³¹ Briefly, this type of drug-related violence is related to the fact that drug dealers, like police officers, generally are “on the job 24/7.” What this means is that to function effectively in an extremely predatory environment, the individual needs to always be “on.” A drug dealer involved in a violent drug market is going to have difficulty maintaining his safety when he is selling if he is perceived as being weak in other domains of his life.

When asking why this is relevant, it is important to remember that the more an investigator knows about what happened and why, the higher the likelihood that a suspect will be developed, effectively investigated, and prosecuted. A second benefit of analyzing violent crime in this fashion is that the outcomes can be used to guide the development of enforcement strategies. For example, a finding that in a certain area associated with open-air drug markets, most of the victims of drug-related homicides were employed is consistent with drug users being killed. A possible enforcement strategy in this situation could include aggressive demand reduction techniques in an effort to keep potential victims out of harm’s way. On the other hand, victim characteristics consistent with dealer rivalries or gang disputes over markets or territories would require a different approach.

Information collected during the investigative process can be subdivided and used to determine the nature of the crime and possible suspects. As can be seen in [Figure 11.9](#), much of the information collected and considered in the investigative process can be reduced to simple binary choices or sets, reinforcing the fact that data preparation is one of the most important steps in the data mining process.

While startling to many, violent crime can be very homogeneous when viewed in the proper light. One feature of the behavioral analysis of violent crime is the characterization of an incident based on common elements. These include victim as well as scene characteristics, which ultimately are associated with common perpetrator characteristics. The goal is to develop a profile or model of the type of person likely to have committed the crime, based on the known

**FIGURE 11.9**

Possible decision tree for categorizing behavioral and forensic evidence.

characteristics of the crime. In many ways, the behavioral analysis of violent crimes is similar to algebra. Ultimately, the objective is to solve for “X,” the suspect’s identity, which is done by systematically revealing and examining the other elements of the equation.

Whether consciously or not, the death investigator goes through a series of “yes or no” questions, often as early as when the first call comes into the dispatch center. This information is used to begin characterization and categorization of the crime based on characteristics of the victim(s) and crime scene(s). These include, but are not limited to the following:

- Was there more than one crime scene? Was the victim moved or the body dumped?
- Was it a weapon of opportunity, or did the suspect bring the weapon to the scene?
- Was the victim at high risk for violence, or did the suspect assume risk in selecting this particular victim?

The list goes on, but a considerable amount of information can be described as binary in nature, meaning that it can be answered by a simple “yes” or “no.” Other information can be divided into categories or sets (e.g., occupation) – information ideal for data mining and predictive analytics.

These similarities can be used to link similar crimes perpetrated by the same individual or group of individuals. They also can be used to link similar crimes perpetrated by similar individuals or groups of individuals. In the former case,

it is important to link crimes in a series because, like in a puzzle, individual elements present in one incident that were missing or overlooked in other incidents can contribute to a greater understanding of all crimes in the series. In other words, the whole crime series often is greater than the sum of the individual incidents. It also is important to identify all of the crimes associated with a particular individual for prosecutorial purposes. Finally, presenting a linked series can be helpful in jogging the memory of a reluctant witness or suspect.

Associating a crime or series of crimes to known suspects associated with solved cases can be helpful in at least two ways. First, it can provide guidance as to the type of individual likely to have perpetrated a particular crime or series. In contrast to television and movie portrayals of “profilers,” the behavioral analysis of violent crime does not identify a specific individual. Rather, it associates a crime or series with perpetrator characteristics that can be used to guide the investigation. Also, by relating a current investigation to an earlier, solved case, additional elements can be highlighted (e.g., things that you might want to look for, why they did something, and investigative strategies, particularly interviewing).

11.5.2 Aggravated Assault

Many aggravated assaults can be viewed as incomplete or poorly planned homicides. Similarly, the line that separates a lethal from a nonlethal assault often reflects the quality of medical care or timeliness of response, rather than some specific intent on the part of the suspect. With that in mind, homicides and aggravated assaults can be viewed more as a continuum, rather than as two separate and distinct entities. When we explore this a little further, it makes sense based on what we know of violent crime. For example, the number of innocent victims associated with drive-by shootings supports the fact that drive-bys frequently do not go according to plan. Vehicle movement, unreliable weapons, and random bystanders all contribute to the variability associated with this type of violence. Similarly, it is not unusual for there to be several nonlethal assaults that precede a domestic homicide. Therefore, many of the same approaches to the analysis of murder can be applied to nonlethal assaults. One important difference from a data mining perspective, however, is that there usually are more nonlethal assaults than homicides. While this is a great thing for the victims, it also is a very good thing for the analyst, as there generally are more incidents available for analysis and modeling.

11.5.3 Sexual Assault

Stranger rapists can be some of the most disturbing predatory criminals a detective will encounter. Even the hint of a serial stranger rapist can create a climate of fear in a community. Several years ago, during a casual conversation,

Dr. Paul Ferrara, the director of forensic science in Virginia, noted that a surprising number of the DNA “cold hits” for predatory sex crimes had come from criminals without prior histories of sex offending. At the time, Virginia was noteworthy for having created a very successful offender DNA database with broad inclusion criteria. While some states were confining their samples to convicted sex or violent criminals, the Commonwealth of Virginia obtained DNA samples from all convicted felons.

Dr. Ferrara noted that several of the offenders identified by cold hits were associated with prior property crimes, particularly burglaries. In some ways, this finding was not surprising, in light of what we know about sexual predators and some violent criminals. In reviewing violent crimes, it is not unusual to find a pattern of escalation that includes crimes that do not appear violent initially. For example, Timothy Spencer, the “Southside Strangler” and first person convicted with DNA evidence, had a history of burglaries in northern Virginia that preceded the homicides that he committed later. Similar cases have revealed a pattern of burglaries or trespassing that preceded escalation into more serious patterns of offending.

An initial review of the data revealed that approximately 40% of the cold hits were associated with offenders who had no documented history of either sex crimes or violent offending. Perhaps more importantly, had the database been confined to these patterns of offending, approximately 40% of the criminals might not have been caught and thus would have been allowed to continue to prey on their communities.³²

Exploring this finding offered at least two benefits. First, the ability to characterize and identify patterns of offending that indicate an increased likelihood for escalation offers the promise of early detection, enhanced investigative efficacy, and increased community safety. This also creates increased opportunities for early detection and intervention for sex offenders, who have a pattern of offending that is noteworthy for its high recidivism rate and resistance to treatment. Second, an increased understanding of sex offenders, how they escalate, and how they prey on the community gives behavioral scientists an opportunity to better understand this particularly challenging form of criminal behavior. This greater understanding offers the promise for early intervention and the concomitant increase in public safety.

The initial study involved reviewing large correctional databases. Using discriminant analysis, models were created to determine which factors were predictive of subsequent stranger rapes. Not surprisingly, prior offense history reliably emerged as the most predictive variable. What was a shock, however, was that a prior property crime actually was a better predictor for a stranger rapist than a prior sex offense. It is important to note that discriminant analysis ideally is used with continuous variables, while offense history frequently is

confined largely to categorical variables. Because discriminant analysis is such a robust statistical test,³³ however, it was permissible to violate this assumption. The most likely error to occur when this assumption is violated would be a failure to identify a model, which was acceptable given the nature of the question.

Subsequent manual review of the paper files associated with these offenders revealed some differences in the nature of the property crimes they perpetrated. In several cases, these offenders appeared to specifically target occupied dwellings. When someone was home at the time of the break-in, it frequently was a female alone, or a female with small children. In addition, these offenders frequently took items of little or no value, if they removed anything at all from the residence. This behavior is inconsistent with a purely economic motive for the crime.

In many ways, these crimes differed qualitatively from traditional burglaries. Reduced to its simplest form, a burglary is an economic crime in which the offender tries to maximize the yield while managing the risk of being caught. The individuals associated with a subsequent stranger rape distinguished themselves from “normal” burglars in that they incurred a greater degree of risk in preferentially targeting occupied dwellings, and generally had little to show for their efforts other than a few items that could be viewed as souvenirs. In other words, their crimes were abnormal.

The discovery and confirmation processes associated with data mining expanded our understanding of stranger rapists. This new understanding was used to generate a brute force anomaly detection system for identifying crime patterns and trends determined to be at increased risk for escalation. Perhaps more importantly, it also resulted in the development of a general principle regarding “normal” crime, which is discussed in Chapter 10. Since the original study, this concept has been applied successfully to other nonviolent crimes that deviate in some way from “normal.” In other words, anything that suggests some type of secondary gain beyond or instead of the economic motive generally associated with the crime is cause for concern because it frequently indicates potential for escalation or more serious patterns of offending. Similarly, any preferential behavior on the part of the offenders that increases their risk of detection or apprehension also is cause for concern.

In England, the West Midlands police have conducted very successful work using predictive analytics to characterize and apprehend sexual predators. Through the use of self-organizing maps or clustering algorithms, they were able to identify clusters or groups of crimes that were similar. These similarities were based on a variety of relevant dimensions, which included method of approach, verbal themes, precautions taken to prevent detection, and victim characteristics.³⁴

This work is particularly encouraging because many of the category clusters identified in this study match classifications described previously in the United States.³⁵ This similarity suggests commonalities in some patterns of offending that might transcend national and perhaps even cultural boundaries. Therefore, unlike some business, health care, and educational applications of data mining and predictive analytics, work in the public safety and intelligence arena promises to transcend national boundaries. This further increases the number of potential end users for predictive algorithms, while enhancing the opportunities for increased resource exploitation of criminal justice data resources and predictive algorithms.

11.5.4 Abduction

As the LRA and child abduction examples illustrated earlier, victims can be kidnapped for a variety of reasons that pose differential risk to the victim and likelihood for a favorable resolution. Additional patterns of abduction that the analyst may encounter include the following³⁶:

“Tiger” Kidnapping

Another variation on the kidnap for financial gain model includes the “Tiger” kidnapping,³⁷ which actually represents two linked crimes: kidnapping and robbery. In this model, the offenders kidnap someone with access to the target of interest and/or their family members and coerce the person with inside knowledge or access to commit the actual robbery. The name refers to the fact that these complex, well organized crimes frequently involve extensive periods of preoperational surveillance of the victims and potential target that resembles the stalking behavior exhibited by predatory animals. Perhaps the most well-known Tiger kidnapping involved the Securitas cash station in Tonbridge, Kent, UK.³⁸ Several years ago a series of similar “Tiger” kidnapping robberies were perpetrated in Maryland, raising concern that this pattern would soon establish itself in the United States. After an initial series of incidents, however, the pattern was ended by arrest,³⁹ although Tiger kidnapping still represents a challenge to law enforcement outside the United States.⁴⁰

“Express” Kidnapping

The creativity of criminals never ceases to amaze me. In the “express” kidnapping, criminals exploit the challenge associated with midnight discussed in Chapter 5 wherein two dates are captured in one “night.” In this particular model, a victim is kidnapped and taken to an ATM where they are forced to withdraw the daily limit. The kidnappers then hold the victim, taking them to the ATM again after midnight and force them to withdraw the daily limit a second time; effectively doubling the yield by holding the victim past midnight and exploiting the fact that each “night” essentially captures two dates.

11.6 CHALLENGES

One common theme throughout this text can be summed up as “caveat emptor,” or let the buyer beware. As with other sources, behavior may or may not be what it appears. Again, victims frequently are traumatized, witnesses might be confused, and suspects frequently lie. While a complete treatment of this subject is well beyond the scope of this text, it is important for the analyst to understand that the information that they receive is associated with varying degrees of accuracy and reliability, both unintentional and deliberate.

11.6.1 Deception

There have been several popular television programs and even more technology solutions based on the idea that deception can be inferred from physical indicators, behavior, or otherwise “read.” In a discussion of deception, several colleagues and I explored the relative complexity of deception and some of the challenges associated with accurately measuring and countering it. In fact, entire programs of research are devoted to surfacing and modeling the neural substrates and cognitive processes associated with deception. Of interest to the public safety and national security community are questions regarding how fraudulent statements may differ from those that are true. Are there detectable differences between deception that involves creating new information versus just denying the truth; so-called sins of “commission” versus “omission”? And, are there measurable, physiological indicators of deception not susceptible to countermeasures?

Related to this, other researchers are delving into other factors that may compromise the reliability and validity of information, including eyewitness testimony and recall. One interesting message implicit in the popular treatment of deception, as well as “scientific” approaches to detecting deception is that humans attempt to deceive in relatively predictable ways. In fact, the polygraph is based on the underlying assumption that the physiological response associated with attempts to deceive is not only detectable, but relatively homogenous in nature. It can be measured and even predicted; changes in heart rate, perspiration, and so on, serve as indicators of detection. Newer approaches look at facial features and expression, including changes in blood flow. Similarly, other approaches used to identify deceptive statements and reports are based on the finding that the verbal content also tends to be predictable. In other words, people just are not that original when they are confabulating. Whether they are creating a story to explain a set of circumstances or fabricating financial transactions to perpetrate fraud, they tend to differ from “normal” behavior, frequently in an unusual degree of homogeneity or lack of detail.

As mentioned in Chapter 6, certain investigative techniques, including Scientific Content Analysis (SCAN), specifically look for gaps or differentially

missing data as indicators of deception. While I am sure that there are some very interesting psychological or even biological reasons for why people revert to common themes and predictable patterns in deception, the important thing is that they can be predictable and detected.

11.6.2 Bias

We all bring our life experiences, belief systems, history, and context to analysis. The challenge is to not let it color or otherwise alter our analysis to include variable selection, analytic approach, and interpretation and presentation of the results. When I lecture on juvenile murderers, I generally highlight the fact that one of the biggest challenges in investigating violent juveniles includes what the detective “knows,” or thinks that they know about children. For better or worse, most of us have direct experience with children. Whether they are our own, nieces, nephews, neighbors, or even characters on television; we have a wealth of knowledge regarding child behavior. Unfortunately, this “knowledge” frequently limits our ability to consider what seriously violent young people might be capable of doing. Those of us with experience in violent crimes understand the importance of compartmentalizing our personal from professional lives in order to protect against personalizing the victims and harming ourselves. Those who see their children or other loved ones in the faces and lives of crime victims cannot effectively function in this space. Similarly, those who let their experiences and personal knowledge color or otherwise guide their understanding of the acts that people, particularly young people, are capable of committing are limited in their ability to consider all likely victims and effectively investigate these types of crime.

One particular case that illustrates this point involved a 10-year-old who killed a family member and then altered the scene to make it appear as if the victim had been killed in a drive-by shooting. In this particular case, the young person ejected the spent round and tossed it into the yard. He then moved the victim outside onto the front stoop and called 911 to report a shooting. At an age when most children have a difficult time even understanding the concept of death, this youngster killed someone and then altered the forensic evidence in an effort to stage a crime and deflect attention away from him as the perpetrator.

Is this case unusual? Most certainly. In fact, the statutes in most states guiding the age at which a child can be arrested for certain crimes, prosecuted, or transferred to the adult system not only reflect the rarity of this type of event, but also implicitly underscore the presumption that very young children cannot and do not perpetrate seriously violent crimes. Unfortunately, they can and do with increasingly frequency. Incidents involving sexualized violence perpetrated by young people can be even more difficult to understand, particularly when most of our direct experience with early adolescent sexual experience involves broken hearts or an unplanned pregnancy, not sexual homicide.

The point that I make to investigators with these cases is that the question for them is not, “are children capable of committing seriously violent, or other heinous acts,” because the answer is a resounding, “yes.” The real question for the investigator is, “is the child suspect that I am considering capable of committing the crime that I am investigating?” The related point for the analyst is that we need to be cautious about the difference between what we know and what we *think* we know. We need to protect against bias, subtle or otherwise, which can influence our analysis and interpretation. Again, the confirmation aspects of data mining and predictive analytics can be particularly important in reality testing conventional wisdom.

11.6.3 Other Influencers

While generally not intentional, there is an extensive body of research demonstrating that eyewitness testimony can be notoriously, and in some cases, tragically unreliable. Whether due to psychological trauma, confusion, or by power of suggestion – intentional manipulation or just poor interview technique – recall can be very labile, and prone to contamination. Similarly, demand characteristics and a desire to fill in details or find “closure,” whether conscious or unconscious, also can contaminate eyewitness and victim reporting, as well as analytic objectivity. As always, maintaining a healthy skepticism and ability to test, retest, validate, and repeatedly question the findings will serve the analyst well.

11.6.4 Just the Facts, Ma’am

“Sometimes a cigar is just a cigar.”

Sigmund Freud

While the quote probably represents a behavioral sciences urban legend, it underscores an important point that can be best illustrated by example. Several years ago I was reviewing a sexual homicide case. After sexually assaulting and killing the victim, the perpetrator in this particular case moved her to the bathroom, placed her in the bathtub and filled it with water. I was intrigued because this postmortem activity was not related directly to the crime. The victim had been killed earlier by other means so placing her in the bathtub was not necessary to committing the crimes of rape and murder. Because the perpetrator had engaged in this “extra” behavior, I assumed that it was meaningful and started to create some very complex, psychologically rich explanations for it. I reached far back into my brain to an abnormal psychology course that I had taken several years earlier and decided that the perpetrator actually felt very guilty about the crime that he had committed and had placed the victim in the bathtub filled with water because he was trying to “wash away his sins” in some sort of primal, cleansing ritual; similar to the undoing that is sometimes

seen in crimes perpetrated by people familiar with, or having some affection toward the victim. I had an opportunity to test these hypotheses with a seasoned federal investigator who had far more expertise in behavioral analysis than I. His explanation for the behavior was that the perpetrator hoped the water would wash away or otherwise contaminate the forensic evidence left behind after the assault. He was very likely correct and his point very important as we analyze behavior – keep it simple. With the exception of signature, most behavior is related directly to successfully executing the crime. Behavior that is not, frequently was unintentional; a response to something unplanned, or a spurious finding that was accidentally included during the investigation. There is a saying in medicine that applies equally to behavioral analysis: “when you hear hoof beats, think horses, not zebras.” While it is intriguing to brainstorm complex or exotic explanations for behavior, it is important not to let the “zebras” contaminate or overly influence the analysis of behavior or interpretation of your findings.

11.7 MOVING FROM INVESTIGATION TO PREVENTION

Improving investigative efficacy is very important. But what if violent crime could be characterized, anticipated, and even predicted? If this were possible, then we would have an opportunity to engage in proactive strategies that would prevent crime before it happened. In death investigations, the act already has been committed. But what if it was possible to anticipate who was next? Minimally, the high rate of subsequent assaults and murders documented in victims of violent crimes⁴¹ identifies them as a group at extraordinarily high risk for violent assault and murder. It is unknown whether this increase is associated with the idea that each aggravated assault really represents an incomplete or poorly planned homicide, or because the same lifestyle factors that resulted in the first assault are still present. By identifying who is at risk, where, and why, traditional enforcement strategies can be matched and targeted to specific patterns of risk. Through the use of predictive analytics to create a rule set for drug-related homicides, for example, it was determined that victims killed near a particular drug market were more likely to be employed. Additional information may indicate that this particular market was frequented by younger, relatively affluent, recreational drug users from neighboring communities. One possible approach in this situation would be a demand reduction strategy such as “reversals,” where police officers play the role of drug dealers in an effort to identify and arrest users, ultimately keeping the potential victims away from a dangerous activity or market. A different approach might be warranted for drug-related homicides involving individuals associated with other illegal drug markets. Identification, characterization, and modeling of victim risk factors represent another use for data mining and predictive analytics in our efforts to reduce violent crime. Ultimately, being able to enter the decision cycle of

our opponent increases the opportunities for influence, and moved us to a proactive posture rather than simply responding to crimes only after they have occurred.

11.7.1 Anticipation and Influence

Taking a cue from the commercial sector, we can use predictive analytics to characterize behavior in support of informed anticipation and influence. The use of data mining and predictive analytics to characterize crime and other patterns of bad behavior in support of targeted approaches to prevention, thwarting, mitigation, and response represents the first steps in the preparation of a comprehensive, information-based approach to enhanced investigative efficacy and meaningful, targeted crime prevention. While we cannot prevent everything, meaningful analysis of behavior and the related insight can be used to change response planning and mitigate consequences of those incidents that do occur despite our best efforts. Again, there was no evidence that security manager Rick Rescorla used anything other than tacit knowledge and domain expertise, but his ability to effectively anticipate bad behavior and plan accordingly changed outcomes.⁴² This concept will be considered in greater detail in the next chapter.

Bibliography

- 1 McLaughlin CR, Yelon JA, Ivatury R, Sugerma HJ. Youth violence: a tripartite examination of putative causes, consequences and correlates. *Trauma Violence Abuse* 2000;1:115–27.
- 2 Cohen LE, Felson M. Social change and crime rate trends: a routine activity approach. *Am Sociol Rev*;44:588–608.
- 3 McLaughlin et al. 2000.
- 4 The interested reader is directed to, Douglas J, Burgess AW, Burgess AG, Ressler RK. *Crime classification manual: a standard system for investigating and classifying violent crime*. Hoboken, NJ: Wiley; 2013.
- 5 A full review of this particular topic is well beyond the scope of the current text. The interested reader is strongly encouraged to read the seminal text on this topic: Douglas J, Burgess AW, Burgess AG, Ressler RK. *Crime classification manual: a standard system for investigating and classifying violent crime*. Hoboken, NJ: Wiley; 2013.
- 6 Casey E. Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and internet crime investigation. Knowledge Solutions, www.corpus-delicti.com/case_based.html; 2002.
- 7 McCue, C. Juvenile murderers. In: Arthur E. Westveer, Ed., *Managing death investigation*. Washington, D.C., U.S. Department of Justice, Federal Bureau of Investigation; 2002. pp. 481–489.
- 8 Goldstein; 1985.
- 9 Bennett, W.J., Dilulio, J.J., Walters, J.P. *Body count*. Simon & Schuster, New York; 1996.
- 10 McLaughlin et al. 2000.
- 11 McCue, C. Haahr, K. The impact of global youth bulges on Islamist radicalization and violence. *CTC Sentinel*, 2008, 1(11) 12-14. <https://www.ctc.usma.edu/posts/the-impact-of-global-youth-bulges-on-islamist-radicalization-and-violence>
- 12 McCue, C., Hildebrandt, W. Campbell, K. *Pattern Analysis of the Lord's Resistance Army and Internally Displaced Persons*. Human Social Culture Behavior (HSCB) Modeling Program Winter 2012 Newsletter, Spotlights, 2012;12, 9.

- 13 McLaughlin CR, Daniel J, Joost TF. The relationship between substance use, drug selling and lethal violence in 25 juvenile murderers. *J Forensic Sci* 2000;45:349–353.
- 14 Klecka WR. *Discriminant analysis. Quantitative Applications in the Social Sciences*; 1980.
- 15 McLaughlin CR, Reiner SM, Smith BW, Waite DE, Reams PN, Joost TF, et al. Firearm injuries among Virginia juvenile drug traffickers, 1992 through 1994 (Letter). *Am J Public Health* 1996;86:751–752.
McLaughlin CR, Smith BW, Reiner SM, Waite DE, Glover AW. Juvenile drug traffickers: characterization and substance use patterns. *Free Inquiry Creative Sociol* 1996;24:3–10.
McLaughlin CR, Reiner SM, Smith BW, Waite DE, Reams PN, Joost TF et al. Factors associated with a history of firearm injuries in juvenile drug traffickers and violent juvenile offenders. *Free Inquiry Creative Sociol, Special Issue: Gangs, Drugs and Violence* 1996;24:157–165.
- 16 Goldstein, P.J. The drugs/violence nexus: A tripartite conceptual framework. *J Drug Issues*; 1985, 15, 493–506.
- 17 McLaughlin CR, Robinson DW, Faggiani D. Declining homicide rates in the 1990s: not everywhere! *ACJS* 1998.
- 18 McCue C, Parker A. Web-based data mining and predictive analytics: 24/7 crime analysis. *Law Enforcement Technol* 2004;31:92–99.
- 19 Douglas J, Burgess AW, Burgess AG, Ressler RK. *Crime classification manual: a standard system for investigating and classifying violent crime*. Hoboken, NJ: Wiley; 2013.
- 20 McCue C. Pattern analysis of LRA & IDPs. ICCM 2012. <http://www.youtube.com/watch?v=TRkva9NAjTw>; 2012.
- 21 Douglas J, Burgess AW, Burgess AG, Ressler RK. *Crime classification manual: a standard system for investigating and classifying violent crime*. Hoboken, NJ: Wiley; 2013.
- 22 Dalton JR, Porter MD. *Geospatial Preference Models in Signature Analyst* (white paper, McLean, VA: SPADAC, Inc.);2013.
- 23 McCue C, Hildebrandt W, Campbell K. Pattern analysis of the Lord's Resistance Army and internally displaced persons. *Human Social Culture Behavior (HSCB) Modeling Program Winter 2012 Newsletter, Spotlights* 2012;12:9.
- 24 Consistent with the use of abductees as porter, analysis of the looting incidents revealed that abduction areas represented a factor reliably associated with looting; McCue C. 2012.
- 25 Lord WD, Boudreaux MC, Lanning KV. Investigation of potential child abduction cases: a developmental perspective. *FBILEB*; April 2001.
- 26 Goldstein PJ. The drugs/violence nexus: a tripartite conceptual framework. *J Drug Issues* 1985;15:493–506.
- 27 For a discussion of the use of violence to regulate behavior and enforce informal social rules and norms, Goldstein PJ. The drugs/violence nexus: a tripartite conceptual framework. *J Drug Issues* 1985;15:493–506.
- 28 Lord WD, Boudreaux MC, Lanning KV. Investigation of potential child abduction cases: a developmental perspective. *FBILEB* April 2001.
- 29 See Westveer AE, editor. In: *Managing death investigation*. Washington, DC: U.S. Department of Justice, Federal Bureau of Investigation; 2002; or Geberth VJ. *Practical homicide investigation: tactics, procedures, and forensic techniques*. 3rd ed. New York: CRC Press; 1996.
- 30 Goldstein PJ. The drugs/violence nexus: a tripartite conceptual framework. *J Drug Issues* 1985;15:493–506.
- 31 McCue C. 2002.
- 32 McCue C, Smith GL, Diehl RL, Dabbs DF, McDonough JJ, Ferrara PB Why DNA databases should include all felons. *Police Chief* 2001;68:94–100.
- 33 Klecka WR. 1980.
- 34 Adderly R, Musgrove PB. Data mining case study: modeling the behavior of offenders who commit serious sexual assault. *ACM Special Interest Group on Knowledge Discovery and Data Mining*; 2001.
- 35 Ressler RK, Burgess AW, Douglas JE. *Sexual homicide: patterns and motives*. New York: Lexington Books; 1988.

- 36 For additional insight and background on different kidnapping models, and their related outcomes, the interested reader is encouraged to read, Auerbach AH. Ransom: the untold story of international kidnapping. New York: Henry Holt; 1998; and Hallums R. Buried alive: the true story of kidnapping, captivity, and a dramatic rescue. Nashville, TN: Thomas Nelson; 2009.
- 37 For a very good review of the model, see: Control Risks Group Limited. Tiger kidnap – the threat to the UK banking sector; 2007.
- 38 Macdiarmid P. British police arrest 2, continue hunt in \$87M heist. USA Today. http://usatoday30.usatoday.com/news/world/2006-02-23-bank-heist_x.htm; 2006 [accessed 23.02.2006].
- 39 Sentementes GG. Bank holdup trend: kidnapping manger's family. The Baltimore Sun. http://articles.baltimoresun.com/2008-12-31/news/0812300125_1_bank-robberies-kidnapping-trust-bank; 2008 [accessed 31.12.2008].
- 40 Mulgrew J. Tiger kidnapping: cash stolen after family held hostage by gang in west Belfast. Belfast Telegraph. <http://www.belfasttelegraph.co.uk/news/local-national/northern-ireland/tiger-kidnapping-cash-stolen-after-family-held-hostage-by-gang-in-west-belfast-29983287.html>; 2014 [accessed 05.02.2014].
- 41 Sims DW, Bivins BA, OBeid FN, Horst HM, Sorenson VJ, Fath JJ. Urban trauma: a chronic recurring disease. J Trauma 1989;29:940–947.
- 42 Ripley A. The unthinkable: who survives when disaster strikes – and why. New York: Three Rivers Press; 2009.
- 43 McCue C, Hildebrandt W, Campbell K. Pattern analysis of the Lord's Resistance Army and internally displaced persons. Human Social Culture Behavior (HSCB) Modeling Program Winter 2012 Newsletter, Spotlights 2012;12:9.

Risk and Threat Assessment

“Prediction is difficult, especially of the future.”

Neils Bohr

The world is filled with risk. Every action that we consider or take has some degree of risk associated with it. We conduct risk or threat assessments internally almost constantly as we move through our lives. “Should I pull out in front of this car?” “Will I run out of gas?” “Can I skip my annual checkup?” These are only a few of the many calculated risks that we take each day.

Part of risk assessment is weighing the likelihood of a particular outcome associated with a particular action or inaction against the potential seriousness of the outcome. For example, when making a decision to fly, we know that the risk of a crash is extremely small; however, the potential consequences associated with an airplane crash can be very serious. It is much riskier to drive, but the perceived consequences associated with a crash are much less. Also factoring into the equation in many cases are the perception of control. If I drive to my destination, there is the perception that I have greater control over the outcome, which highlights the fact that some, if not most, of the information that we use in personal risk or threat assessment might be inaccurate or unreliable.

In the public safety community, we are asked to evaluate and mitigate risk on a regular basis. For example, is the potentially hostile crowd depicted in [Figure 12.1](#) likely to riot, or will a show of force only escalate a relatively stable situation? Are there other ways of dealing with this potentially volatile situation that will reduce the risk? Perhaps one of the greatest challenges is that we usually are dealing with very unlikely events and with incomplete information. The likelihood of being a victim of crime generally is very low. As discussed in Chapter 11, however, risk can be increased by certain lifestyle factors or other related issues known to impact crime victimization rates. Data mining and predictive analytics can greatly assist the process of identifying factors associated with risk, and are particularly adept at addressing the many issues that make accurate risk and threat assessment such a challenge.



FIGURE 12.1

Potentially hostile crowd outside the Coalition Provisional Authority headquarters in Basra, Iraq. (Courtesy of Staff Sergeant Tom Ferguson, USMC.)

FOURTH-GENERATION WARFARE (4GW)

Modern warfare has been divided into three distinct generations by Lind *et al.*¹ The first generation is based on line and column and was largely driven by the weapons technology of the time – smoothbore muskets. Second-generation warfare was driven by changes in technology, which included the introduction of rifled muskets and the breechloader, barbed wire, the machine gun, and indirect fire. Second-generation warfare remained predominantly linear, incorporating the tactics of fire and movement; however, it relied on massed firepower rather than massed personnel. Second-generation warfare remained the foundation of US doctrine until the 1980s, and is still used today.

While first- and second-generation warfare tended to be technology-driven and linear, third- and now fourth-generation warfare represent the first nonlinear tactics, which reflected changes in ideas rather than technology, and emphasizes maneuver over attrition. Several elements attributed to 4GW have significant implications for local law enforcement. These include decreasing reliance on centralized logistics and the use of agile, compartmentalized units, which are similar to the cells frequently associated with international terrorist organizations or domestic hate groups. 4GW also emphasizes the goal of collapsing the enemy from within as opposed to annihilating him physically, and targets may include social and cultural objectives as well as support for a war effort, economic stability, power plants, and industry. One needs only look at the stock market and the airline industry after 9/11 to see the larger impact of these tactics. Perhaps one

of the most important issues for local law enforcement will be the blurred distinction between civilian and military, and the absence of a traditional front or battlefield. According to Lind and colleagues, the enemy will be targeted at every level, including social and cultural, and traditional “battlefields” might be difficult to identify in a 4GW scenario.

12.1 BASIC CONCEPTS

Meaningful risk and threat assessment requires an integrated approach that includes topics covered throughout the text. These subjects merit repetition given their relevance to risk and threat assessment and will be reviewed next.

12.1.1 Knowing “Normal”

In many ways, solid internal norms or domain expertise is essential to effective risk and threat assessment. Knowing what is “normal,” particularly for crime and criminal behavior, can be used to identify abnormal behavior, which generally indicates an increased risk for escalation. Data mining and predictive analytics can function as invaluable tools in the characterization of normal, as well as in the identification of abnormal incidents or trends that are worthy of additional investigation. For example, a series of suspicious situation reports could mean totally different things, depending on the time of day and local crime patterns. Suspicious behavior occurring during the daytime could indicate possible surveillance associated with burglaries or other property crimes. On the other hand, surveillance during the night, when residents are likely home alone, could suggest something far more sinister. This example underscores the importance of knowing the community, normal patterns, and even normal crime when evaluating analytical output. Again, characterization and analysis of normal behavior, including crime, can be invaluable to identifying abnormal or potentially threatening behavior. This concept is addressed in greater detail in Chapter 10.

12.1.2 Behavioral Analysis

As discussed in Chapters 11 and 14, understanding the tactics, techniques, and procedures (TTP), including preoperational surveillance, required for an attack is important to information-based approaches to early identification, prevention, thwarting, and incident response. Perhaps as important, the ability to understand victim behavior, including the anticipated and possibly exploited victim response, can be essential to meaningful planning and consequence management. Similarly, the ability to effectively anticipate likely methods, TTP, and scenarios supports information-based response planning that may mitigate the impact of an attack that cannot be prevented.

Again, a solid understanding of “normal” is essential to being able to surface unusual, suspicious, or otherwise concerning behavior. Related to this is the

importance of understanding “normal” human response to critical incidents and other threats. For example, review of airline accidents found that many passengers lost their lives trying to exit through the door that they had entered the plane. Now, during the preflight safety briefing, you may notice that the attendants specifically call out that the closest exit might be behind you in an effort to mitigate this “normal” but potentially hazardous response.² Similarly, a very primal response to danger or a perceived threat is the desire to flee. While this may have represented an adaptive behavior in our ancestors, it is becoming increasingly clear that our adversaries also are students of behavior and are leveraging the anticipated responses of their victims and/or first responder in support of complex attacks, which frequently are based on anticipated response to an attack, emergency, or some other precipitating incident.

The complex attack that leverages the human response to a critical incident has been used for decades. Examples of this model include the Beslan school siege where the attackers engaged in an extensive period of preoperational surveillance and attack planning in support of information-based anticipation and manipulation of the responses of the parents, school officials, and children; effectively exploiting it to channel the victims into a location where they could be easily managed – the auditorium.³ Similarly, many of the complex attacks perpetrated by extremist groups in Africa include an initial attack that results in the fleeing victims being funneled into a secondary “kill zone” where they are sprayed with small arms fire. Alternatively, complex attacks in other locations may involve the use of an improvised explosive device (IED), or some other attack method that stops or otherwise constrains movement, effectively creating a stationary target.

“It seems to me that the whole virtue of good forecasting is not merely to predict the obvious but to predict the exceptional.”

Ernest Clowes⁴

It is also important to note that being able to anticipate an event may not be sufficient to preventing it. Identifying the likely, “when, where, and what” of an attack, however, can guide information-based approaches to prevention, thwarting, response, and consequence management, as well as informed public safety and security decisions, prioritization, and resource allocation. Again, the role that information-based response planning may play in reducing casualties was underscored during the September 11 attacks. Rick Rescorla, Vice President for Security at Morgan-Stanley/Dean-Witter, believed that terrorists would continue to attack the World Trade Centers until they were able to bring those buildings down.⁵ Knowing that only practice would increase the ability of the employees to quickly and safely evacuate the building, Rescorla repeatedly staged evacuation drills for his personnel to the point where it became a standing joke within the organization. While his foreknowledge of the likely method

for the attack was startling in its accuracy, it was not sufficient to prevent the attack. His insistence on routine evacuation drills, however, was credited for saving the lives of 2700 of his colleagues in the South Tower.⁶ On September 11, 2001 when the planes hit the building, the employees knew exactly what to do.

12.1.3 Surveillance Detection

This topic is covered in Chapter 14; however, the ability to detect when a person or place is being watched or evaluated can be extremely important in risk and threat assessment. Not only does the identification of possible surveillance activity indicate the possibility of some type of preoperational planning or activity, but it also can highlight previously undetected vulnerabilities in either the physical or operational security of a particular location or activity. For example, consistent reports of suspicious activity indicating possible surveillance associated with an area frequented by employees who smoke might reveal that a door is propped open regularly to facilitate return into the building without having to utilize distant or inconvenient points of access. This surveillance activity, while focusing primarily on a particular area, ultimately reveals potential security vulnerability that could be addressed.

The ability to identify, model, and characterize possible hostile surveillance⁷ provides at least three direct operational benefits. First, it allows us to identify the times and location of interest to our adversary, which then supports targeted surveillance detection efforts. If we know when and where we are being watched, then we also know when and where to watch them (watch us). This can be invaluable in revealing larger patterns of hostile surveillance and attack planning.

The second benefit is that it can reveal potential vulnerabilities or areas of interest to likely predators. The risk associated with a particular facility, location, or individual is unique and can fluctuate in response to prevailing conditions and a wide array of external events. I can speculate as to what might be of interest to someone; however, I am likely to be wrong, as I do not have sufficient information to see the big picture from another's perspective. For example, someone interested in a particular facility because his or her spouse works there presents a very different risk than someone interested in a facility because it supports critical infrastructure or represents the potential for a high number of casualties. The potential threat, strategy, and required tactics associated with the domestic situation would be expected to be very different in time, space, and method than the threat associated with someone interested in the entire facility. To try to assume what might happen can blind the analyst to what is being considered. It is generally better to let their behavior reveal their intentions to us.

Finally, the insight gained through an analysis of preoperational surveillance indicators can provide important information regarding the possible nature or

TTP of an attack. Again, letting our adversaries reveal their intentions through their behavior can help mitigate the “failure of imagination” frequently cited after the 9/11 attacks.

12.2 VULNERABLE LOCATIONS

Again, location matters to our adversaries. Access to victims or targets and a favorable environment can be used to reliably identify incident locations. Some locations are uniquely vulnerable to attacks that could disrupt life or generate a large number of casualties, either by the nature of the business conducted or by the value of the occupants. These locations include critical infrastructure and locations where large groups of people congregate. For example, we have seen increasing evidence of hostile surveillance on critical financial facilities and the transportation industry, while the developing security and military forces in Iraq have been the target of ongoing attacks by the insurgents. Similarly, Israel has experienced attacks in crowded locations, such as shopping malls and dining establishments frequented by civilians, for many years. Recent attacks throughout the world have targeted hotels and resorts, further underscoring the increased risk associated with locations like these.

12.2.1 Schools

One location of increasing concern to public safety and security experts is our schools. Children represent something very important to us as individuals and society. Their innocence and vulnerability has been exploited by individuals and terrorist groups that hope to further an agenda or create fear in a community. As someone that was affected directly by the wave of random violence associated with the Washington, DC sniper in the autumn of 2002, I can speak directly to the abject fear that can be created by the potential for risk to our children. Even the mere suggestion that school children in central Virginia might be at risk resulted in school closings and widespread panic. Unlike types of risk that are associated with involvement in high-risk activities, occupations, or lifestyles, there is something terribly unsettling about predators randomly targeting children, particularly in the school setting.

In addition to being a source of particularly innocent and vulnerable targets, schools also provide reliable access to specific children, as underscored by the number of familial kidnappings that are associated with schools. During the academic year, the most reliable location to find a specific child is the school, which can be easily exploited. Also, until recently, schools in the United States generally represented very soft targets. The physical security was lacking, as were related approaches to operational security. The increased prevalence of school shootings, however, has caused many schools to revisit physical security, as well as response planning to include lockdown drills, but these locations

still represent very attractive targets. Finally, schools may represent a compelling target given their purpose, mission, and what they represent – locations for the sharing of ideas and education. For example, the extremist group Boko Haram frequently targets schools given their opposition to Western education and culture.

SHEEP, WOLVES, AND SHEEPDOGS

Lieutenant Colonel (Ret.) Dave Grossman, a global expert on violence and terrorism, is a hero to many in the operational world, and deservedly so. He is not only a dynamic lecturer who has studied violence and our response to it, but he speaks with considerable authority about honor and the warrior lifestyle. In these lectures, Colonel Grossman frequently categorizes people as sheep, wolves, and sheepdogs.⁸ The vast majority of people in the world can be categorized as sheep. Grossman advises that this label is not derogatory. Rather, it refers to the fact that most people are kind and gentle with few inclinations toward violence; however, like sheep, they require protection from predators. The wolves, as one might imagine, are the predators among us. Grossman describes these people as truly evil, preying on the weak and defenseless sheep at will. Finally, the sheepdogs are relatively few in number; it is their role to protect the flock by confronting the wolves. According to Colonel Grossman, these are the “warriors,” the operational personnel in the public safety, security, and military fields that protect us against the predators in the world. The sheepdog also has the capacity for violence, but only in its role as protector of the flock.

This analogy is particularly relevant to several of the topics in this book. For example, Grossman describes research based on interviews with convicted violent offenders that suggests these predators look for weak victims: those off by themselves, demonstrating a lack of confidence, or poor survival skills. Similarly, predators often target prey that wander off or get separated from the rest of the herd, or those that are weak, show poor survival skills, or show a lack of situational awareness. These interview data are consistent with some of the research on victim risk factors and victimology outlined in Chapter 11. Grossman goes on to describe sheepdog behavior as being vigilant, “sniffing around out on the perimeter;” behaviors that are very similar to the surveillance detection described in Chapter 14.

This is a wonderful model for understanding the relationship between predators, public safety and security personnel, and the people that they protect. As an analyst, however, the role of the shepherd immediately comes to mind. Like analysts, the shepherd is not there to tell the sheepdog how to do its job; rather, the shepherd brings a unique perspective that can enhance the sheepdog’s situational awareness and provide additional guidance. Ultimately, like the sheepdog and shepherd, analysts and operators must work together and support each other. To be truly effective, the sheepdog and shepherd must work together as a team. Although each functions in a different capacity, they both share a common goal of protecting the sheep, something that neither of them can do alone.

Michael and Chris Dorn have compiled a historical accounting of attacks in and on schools and schoolchildren throughout the world.⁹ Going back to the 1960s, terrorists and other predators have specifically targeted schools, underscoring their value as potential targets to extremists and other individuals.

More recently, Chechen terrorists attacked a school in Beslan, Russia, moving the children and their parents into the auditorium and planting 10 to 30 explosive devices throughout the crowd.¹⁰ When the siege ended 52 h later, 300 to 400 were dead.¹¹ Colonel Dave Grossman also has studied the potential vulnerability associated with schools and confirms that schools are a particularly desirable target for extremists; citing several examples underscoring the number of international terrorist attacks specifically targeting schools.¹² Colonel Grossman suggests that while preparedness for weapons of mass destruction is important, all indicators suggest that terrorist groups will continue to use conventional explosives, particularly car bombs, given their ongoing success with these methods.¹³ This is not to suggest that they have no capacity for improvement to their methods or that they do not learn. Rather, as John Giduck asserts in his analysis of the Beslan siege,¹⁴ these groups constantly are improving their tactics and strategy in an effort to address operational flaws or limitations, as well as countermeasures. The terrorists involved in the school hostage taking and subsequent siege and massacre had incorporated lessons learned from the Nord-Ost Theater hostage siege and massacre two years earlier, as well as al-Qaeda tactical training, particularly as relates to how to deal with hostages.¹⁵ In his review of previous incidents, Colonel Grossman highlights the terrorists' use of the complex attack model that includes an initial assault that is used to increase the number of victims by massing them in a common location (e.g., outside a building), or otherwise channel the victims to an area where they can be managed more easily (e.g., the Beslan school auditorium).

Why is it important for an analyst to understand terrorist tactics and strategy and how these incidents play out? There are several very important reasons. First, the complexity associated with some of the larger terrorist operations underscores not only the amount but also the prolonged duration of the pre-attack planning cycle. In many situations, this requires significant gathering of intelligence regarding the facility or person of interest. While this may include searches of open-source materials related to the potential target, it also frequently requires extensive on-site observation and collection. As a result of this extensive preattack planning activity, it is not unusual for the hostile surveillance or intelligence collection to be observed and reported in the form of suspicious situation reports, which can be exploited by the analyst to identify and characterize potential preoperational surveillance and attack planning (see Chapter 14). The longer the planning cycle, theoretically the more opportunities for identification, which ultimately supports proactive, information-based prevention, deterrence, and response efforts. Again, prior to the attack on the school in Beslan, the terrorists had collected intelligence on this and related facilities in support of target selection, as well as preparation of the operational plan, tactics, and strategy. This preattack planning activity in particular represents opportunities for identification, characterization, and proactive responses to potential threats, including prevention.

12.2.2 Critical Infrastructure

Critical Infrastructure Protection (CIP) represents a relatively fluid and increasingly broad topic. As the name implies, “Critical Infrastructure” includes, “the assets, systems, and networks, whether physical or virtual, so vital to the United States that their incapacitation or destruction would have a debilitating effect on security, national economic security, national public health or safety, or any combination thereof.”¹⁶ Informal estimates suggest that 75–80% of our nation’s critical infrastructure is privately owned, which further complicates the coordination necessary to ensure that these critical resources are protected. The total number and nature of the individual Critical Infrastructure sectors have changed over time as the concept and related operations have evolved, but the most recent list promulgated by the US Department of Homeland Security, includes the following:¹⁷

- Chemical sector
- Commercial facilities sector
- Communications sector
- Critical manufacturing sector
- Dams sector
- Defense industrial base sector
- Emergency services sector
- Energy sector
- Financial services sector
- Food and agriculture sector
- Government facilities sector
- Healthcare and public health sector
- Information technology sector
- Nuclear reactors, materials, and waste sector
- Transportation systems sector
- Water and wastewater system sector

While all of these sectors are by definition important, some are more important than others in that they embody differential criticality and/or resilience. Therefore, a shorter list of sectors, which tend to be user-defined, is often referred to as “Lifeline CIP” in reference to direct link to sustaining life. As the literature on 4WG highlights, though, disruption of infrastructure and resources deemed to be critical but not defined by or perceived as “lifeline” still holds the power to create significant social disorder and related civil unrest. This was clearly evident during the aftermath of Hurricane Katrina. As illustrated by the events at Memorial Hospital and other health care facilities and nursing homes in the region; loss of power, disrupted communications, and even perceived but not actual deprivation for a relatively short period of time resulted in increasingly poor decisions, including euthanasia of patients.¹⁸ Therefore, in addition to actual attacks on

and intentional disruption of critical infrastructure, other possible scenarios include the “The Next Big One”¹⁹ as relates to global pandemic illness. Similarly, infrastructure disruptions associated with a major earthquake or other natural disasters are also likely to result in panic, lawlessness, and general social disorder.

12.2.3 Other Vulnerable Places

Other locations like houses of worship and other religious facilities, public markets, stadiums, shopping malls, and hotels represent a unique challenge to the analyst and their operational partners given the frequently open nature of the facility, relatively transient nature of the occupants, and the inability to effectively prepare for and drill as a group. For example, unlike cruise ships or even airplanes where the trip is kicked off with a security briefing and in some cases actual drills, religious facilities, hotels, markets, sporting events, and shopping malls generally do not do this, relying almost exclusively on signage and placards, security personnel, and *in situ* extemporaneous consequence management and response.

In one conversation that I had with a senior security professional whose practice included shopping centers in Europe, he shared with me the results of an informal survey of the security personnel assigned to his facilities that made him extremely concerned. He asked them, “What would you do if you knew that an attack was imminent; that there was a suicide bomber in the mall or that the detonation of a device was imminent?” To an individual, they indicated that they would pull the fire alarms in an effort to quickly evacuate the building. Unfortunately, in many well-planned complex attack scenarios, this action would align almost perfectly with the plans of the attackers by actively moving the patrons into a potential kill zone.²⁰ Similarly, hotels have increasingly become targets²¹; however, evacuation frequently relies on individual response as the opportunity to coordinate with possible victims and engage in actual dress rehearsals or drills is unrealistic. Moreover, with a constant churn of visitors, it becomes increasingly difficult to characterize “normal” in support of informed approaches to risk and threat assessment, and coordinated incident response. Again, behavioral analysis, including review of likely victim behavior can represent an important tool in the identification and characterization of preattack indicators and meaningful response planning in support of informed approaches to prevention, thwarting, and consequence management.

12.3 PROCESS MODEL CONSIDERATIONS

Some specific issues to consider in information-based risk and threat assessment are given next.

12.3.1 Data

The data used for risk and threat assessment are especially poor and are almost exclusively comprised of narrative reports. One of the first challenges associated with collecting the data required for effective and thorough risk and threat assessment is to encourage people to report things. In his book *The Gift of Fear*, Gavin de Becker posits that people do not just “snap”; rather, there generally are signs and indicators preceding the event that often go unanswered.²² “As predictable as water coming to a boil,”²³ these signs can be observed and predicted. He even notes that in cases of workplace violence, coworkers often know exactly who the perpetrator is the moment the first shot rings out, further underscoring the leading indicators present in these cases.

As the book title suggests, De Becker observes that most people have the gift of fear. Getting them to acknowledge and heed their fear can save their life. As an analyst, however, getting people to not only acknowledge but also report their suspicions or concerns can save other lives as well.

Colonel Grossman has recommended providing digital cameras to personnel working in and around schools to accurately and reliably collect information suggestive of hostile surveillance or some other threat.²⁴ Not only would this approach be relatively simple for these folks, who are noteworthy for their level of responsibility and lack of spare time and extra hands for the added responsibility of surveillance detection, but this method also provides an opportunity to retain the data for additional review, analysis, comparison, and follow-up. This approach is a relatively easy, low-cost one that could be applied to other locations, particularly those identified as being at risk.

Maintaining open communication and an open attitude is key to making people feel comfortable about reporting their suspicions. Although most attention is on terrorism and homeland security issues, a facility is far more likely to experience violence related to a domestic situation or a disgruntled employee. Also, people carry their personal risk with them, and two of the most predictable locations are school and work. We are generally expected to arrive at a particular time and leave at a particular time. Frequently, our routes to these locations are as set as our schedules, which makes school and work some of the easiest locations to find individuals.

12.3.2 Accuracy versus Generalizability

The issue of model accuracy versus generalizability and error types have been covered in Chapter 1, but are worth addressing again within the context of risk and threat assessment.

One might think that it is always desirable to create the most accurate model possible, particularly in the public safety arena. Further examination of the

issue, however, reveals several trade-offs when considering accuracy. First, and perhaps of greatest practical importance, is the fact that it is unlikely that we will have all of the information necessary to either generate or deploy models with 100% accuracy. When analyzing historical data in the development of models, it is extremely rare to be privy to all of the relevant information. In most cases, the information is similar to a puzzle with many missing pieces as well as the inclusion of a few extra ones that do not belong. On the other hand, it is possible to generate some very accurate models, particularly when using previously solved closed cases in which most of the pieces that have been identified fit into place. It is extremely important, however, to be aware of what information is likely to be available and when. For example, when developing models regarding drug-related homicides, we were able to achieve a high degree of accuracy when suspect information was included in the model. In an investigative setting, however, a model has considerably more value to investigators if it relies on information available early on in an investigation (see Chapter 13).

The second point to consider is how the model will be used. Very accurate models can be developed using some of the “black box” modeling tools currently available; however, those models are not very user friendly. In other words, they cannot be pulled apart and reviewed in an effort to generate actionable output like some of the decision tree models. Even some of the more complex decision trees are relatively opaque and will be difficult to interpret. If the model is intended to be used to guide an operational plan or risk reduction, like the risk-based deployment models, then some consideration to generalizability of the model will need to be given. The ability to clearly interpret a model generally increases at the cost of accuracy. Each circumstance will require thoughtful review and consideration of possible consequences and nature of errors.

12.3.3 “Cost” Analysis

No matter how accurate, no model is perfect. In an effort to manage these inaccurate predictions, the nature of the errors in a model needs to be evaluated. Again, all errors are not created equal. The cost of each type of mistake needs to be evaluated and a “confusion” matrix can be generated to determine the nature of these errors (see Chapter 4).

The cost analysis for risk and threat assessment should include the cost of responding, as compared to cost associated with a failure to respond. In many cases, the potential cost associated with a failure to respond or evacuate in a timely fashion can be enormous. One only needs to look at fatality rates associated with hurricanes prior to accurate prediction models and evacuation to see the enormous cost that can be associated with a failure to act in the face of an imminent threat. Much of the discussion regarding the possible intelligence failures leading up to 9/11 have focused on the number of lives that could have been saved had the threat been recognized and been acted upon in a timely fashion.

When making a decision to evacuate in response to a predicted hurricane, officials also include the potential cost associated with a false alarm. Unnecessary calls for evacuation can be extremely expensive, as the economic costs associated with evacuating an area can be enormous. Perhaps more importantly, they also can cause public safety personnel to lose credibility, which can impact future calls for evacuation. Concern regarding this type of “alert fatigue” has been raised regarding the number of times the terrorist threat level has been raised within the United States. Again, activation of an emergency response system or threat level that is associated with a null event also can compromise public safety if individuals begin to ignore a system that has been associated with repeated false alarms.

12.3.4 Evaluation

Colonel Grossman has discussed US regional responses to mass killings in schools that include mandated “lockdown” drills and a requirement for emergency response plans in schools.²⁵ These responses have included enhanced efforts to identify possible incidents of preoperational surveillance. He has indicated that some schools have distributed digital cameras to school employees to encourage reporting and increase the accuracy of the collected information and has noted that these strategies can serve a deterrence function by creating an inhospitable environment for the necessary preoperational surveillance and planning. This particular strategy also would deter pedophiles and could discourage school-based domestic child abductions. Should an incident occur, however, the photographic evidence collected *a priori* may concomitantly enhance and accelerate investigative efforts, similar to the Boston Marathon and London bombings.

This combined approach highlights the metrics that can be used to evaluate effective risk and threat assessment: early identification, prevention, and response. Ideally, the identification and characterization of a potential threat will support the development of effective prevention strategies. Again, forewarned is forearmed. Although sometimes difficult to measure, one of the goals of data mining and predictive analysis in public safety and security is the identification and characterization of potential threats in support of effective, specifically targeted prevention and deterrence efforts. Unfortunately, these approaches frequently are imperfect, falling well short of the crystal ball each analyst secretly covets. Therefore, a second measure of the efficacy of risk and threat assessment is effective response planning. The transportation attacks in London during the summer of 2005 underscore this point well. Although the signs and indicators of an impending attack were not discovered until the subsequent investigation, the methodical response planning and high state of preparedness resulted in a response to those incidents that was enviable. The interagency coordination and collaboration in support of an integrated response was flawless, and almost certainly limited the loss of life to that associated with the blasts.

Colonel Grossman has created a very interesting analogy to support preparedness for violence in schools and other public venues by citing the amount of resources and time devoted to fire safety. He highlights the number of fire alarms and sprinklers, the use of fire-retardant materials, and the signage marking exits and posted escape plans, noting that the likelihood of a fire is very small, yet there are considerable resources devoted to it. He extends the comparison to the school setting and notes that the number of students killed in a school during a fire during the last 25 years was zero, while the number of kids killed as the result of violence (either an assault or a school mass murder) was in the hundreds, yet fire drills and response plans are mandated while similar planning for violence (the greater threat) generally does not exist.²⁶

12.3.5 Output

Figure 12.2 illustrates possible hostile surveillance activity in and around a critical facility and demonstrates an important point regarding the generation of operationally actionable output in risk and threat assessment. As can be seen in the figure, the concentration of activity associated with a specific aspect of the building highlights the location of greatest interest to the individual or group involved in the hostile surveillance. This information can be used to further refine the threat assessment of this building by focusing on the areas associated with the greatest activity and highlighting particular spatial features or attributes worthy of additional review. Moreover, the analysis of the nature of surveillance activity (outlined in Chapter 14) can further underscore the escalation in the

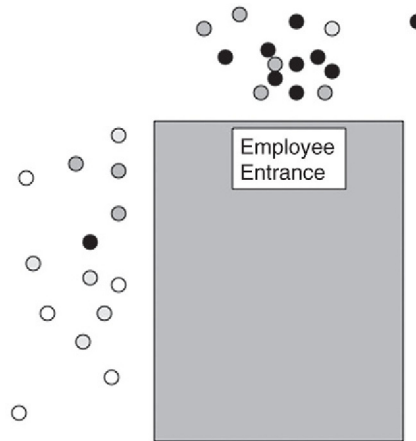


FIGURE 12.2

Figure depicting suspected preoperational surveillance activity associated with a critical facility. The darker dots represent the hypothesized increase in operational value of the surveillance methods employed.

operational relevance of the behavior observed. This information, together with [Figure 12.2](#), which indicates spatial refinement and focusing of the behavior, suggests an increased level of risk associated with this facility.

An important aspect of surveillance detection is to identify and characterize a possible threat so that effective countermeasures can be used. Ideally, the analytical output should build on the end users' tacit knowledge and increase their situational awareness in support of effective prevention, deterrence, and response planning. There is a fine line, though, that separates thoughtful analysis and interpretation of the findings, and reading too much into the data. Errors in interpretation can misdirect resources and potentially cost lives. Therefore, it is almost always a better strategy to let the behavioral trends and patterns speak for themselves and reveal the suspect's intentions than to try to presuppose or second-guess what they might be considering.

12.4 EXAMPLES

Studying previous attacks provides greater insight into tactics and strategy, which can be parlayed into the tacit knowledge frequently required to effectively identify potential threats. For example, an understanding of how predatory pedophiles select and acquire their victims may give the analyst a greater ability to identify likely predators before they harm a child, rather than after the fact. Forewarned truly is forearmed. Identifying an impending attack during the planning phase allows public safety and security professionals an opportunity to move within their adversary's decision cycle and to change outcomes.

By studying incidents and outcomes, the analyst can contribute to the identification and understanding of changes in tactics and strategy. As illustrated by operational revisions incorporated into the Beslan siege that were associated with lessons learned after the attack on the Nord-Ost Theater, terrorist methods are constantly changing and evolving in response to previous failures as well as improved countermeasures and response. Data mining and predictive analytics are well suited to identifying and capturing fluid changes in behavior and modus operandi in a timely fashion. The powerful modeling algorithms incorporated in the tools are able to accommodate and adjust to changes and refresh predictive models accordingly. Again, the ability to stay within our adversary's decision cycle can be game changing in terms of the options available for prevention and deterrence. To support this function, however, the analyst should maintain current knowledge regarding tactics and strategy, particularly as they apply to the assessment of risk and threat.

Finally, knowledge of the operational aspects of risk and threat assessment, as well as response strategies and tactics, is necessary to the production of operationally relevant and actionable output. For example, in terms of surveillance detection,

it is important to consider the nature of the activity and to create a variable that could be used to illustrate changes or escalation in the hostile surveillance. Being able to depict this information in an operationally relevant manner increases the value of the analysis and allows the end users to incorporate their tacit knowledge in the development of surveillance detection and response operations. Similarly, the identification of specific risk factors or victims attributes associated with an increased risk for victimization can be used to develop targeted operational tactics and strategies that directly address the risk or threat. For example, the finding that drug-related violence was associated with the robbery of drug users coming into a particular area to purchase drugs supported the use of a specific operational plan – demand reduction. By identifying why victims were at risk, the command staff was able to structure an operational strategy that kept potential victims out of the area and thereby reduced their risk. To support operational plans like these, though, the analyst needs a solid understanding of crime and criminals as well as operational tactics and strategy to create actionable output.

Although I have highlighted schools as a vulnerable location, it is important to remember that the predator selects the location. Although we can anticipate when and where they might attack, they ultimately select the location based on access, availability, personal preference, secondary gain, and a host of other factors known only to them. Therefore, risk and threat assessment is a fine balance between identifying locations worthy of additional attention and vigilance, and remaining open to subtle indicators and signs that reveal the predator's intentions. This is one area where data mining and predictive analysis can be a tremendous asset. Identification of preoperational surveillance can not only illuminate interest in a particular facility but also be used as a starting point for risk-based threat assessment and response.

The ability to identify, model, and characterize possible hostile surveillance provides at least two direct operational benefits. First, it allows us to identify the times and location of interest to our adversary, which then supports targeted surveillance detection efforts. If we know when and where we are being watched, then we also know when and where to watch them (watch us). This can be invaluable in revealing larger patterns of hostile surveillance and attack planning.

The second benefit is that it can reveal potential vulnerabilities or areas of interest to likely predators. The risk associated with a particular facility, location, or individual is unique and can fluctuate in response to prevailing conditions and a wide array of external events. I can speculate as to what might be of interest to someone; however, I am likely to be wrong, as I do not have sufficient information to see the big picture from another's perspective. For example, someone interested in a particular facility because his or her spouse works there presents a very different risk than someone interested in a facility because it supports critical infrastructure or represents the potential for a high number of casualties.

The potential threat, strategy, and required tactics associated with the domestic situation would be expected to be very different in time, space, and method than the threat associated with someone interested in the entire facility. To try to assume what might happen can blind the analyst to what is being considered. It is generally better to let predators reveal their intentions to us.

12.4.1 Risk-Based Deployment

The concept of risk-based deployment was developed as part of the Project Safe Neighborhoods initiative in the Eastern District of Virginia,²⁷ and it has been used repeatedly throughout this text to illustrate various features of data mining and predictive analytics. Essential to development of the deployment model, however, was creation of the risk model.

Briefly, it was determined that while armed robberies were bad, an armed robbery that escalated into an aggravated assault was worse. By developing a model of robbery-related aggravated assaults, it would be possible to identify potential risk factors associated with this serious pattern of offending and develop targeted law enforcement initiatives that were designed specifically to reduce the risk for these crimes.

Perhaps one of the first challenges associated with this task was that robbery-related aggravated assaults, like many other examples in risk and threat assessment, were a very low-frequency event. Less than 5% of all armed robberies escalated into an aggravated assault. This issue was addressed in the modeling process in two ways. First, it was important to collect a sample that included a sufficient number of the events of interest so that this pattern of crime could be modeled adequately. Therefore, the sampling frame for the analysis was 6 months. This represented an adequate number of events of interest for modeling purposes, yet it did not extend so long as to start incorporating a greater amount of variability, which also would compromise model preparation.

Crime trends and patterns tend to fluctuate over time, some more than others. Robberies frequently vary throughout the year, and can change as various players come and go. Although this issue is addressed in Chapter 13, a sampling frame of approximately 6 months seemed to work well for this pattern of offending. Much beyond that, the amount of variability in the information associated with these long-term patterns and trends really compromised the model construction. Moreover, it also was questionable how much value this type of model would have, as it would be based on relatively old information.

The second method of addressing low-frequency events was to adjust the predicted probabilities of the model. This has been addressed previously, but when modeling low-frequency events, it is important to ensure that the predicted probabilities reflect the actual probabilities. In other words, a good model in this case

should predict that less than 5% of armed robberies will escalate into an aggravated assault. In risk and threat assessment, we are asked to develop a model that will predict an event that generally is relatively infrequent. It is important, therefore, that the model is created to predict future events with a frequency that is relatively close to what would be anticipated based on the rate of incidents historically.

12.4.2 UK Riots

In the summer of 2011, rioters from a variety of backgrounds and locations converged on the streets of England; openly and brazenly confronting law enforcement. More than 3000 arrests were made in response to widespread looting, arson, and violence.²⁸ In the following example, Praescient Analytics utilized the Palantir analytic platform to analyze geospatial and temporal elements of 388 riot incidents and 976 alleged rioters throughout England during early August of 2011. Utilizing Kapow software, analysts explored an exhaustive database constructed entirely of open-source, public information from reputable sources including the Guardian and BBC.²⁹ The goal of the analysis was to provide insight regarding the origin and nature of these riots in support of enhanced situational awareness and informed response.

Figure 12.3 illustrates the riots by incident type. As can be seen in the figure, “looting” and “vandalism” represented the most frequent type of incidents.

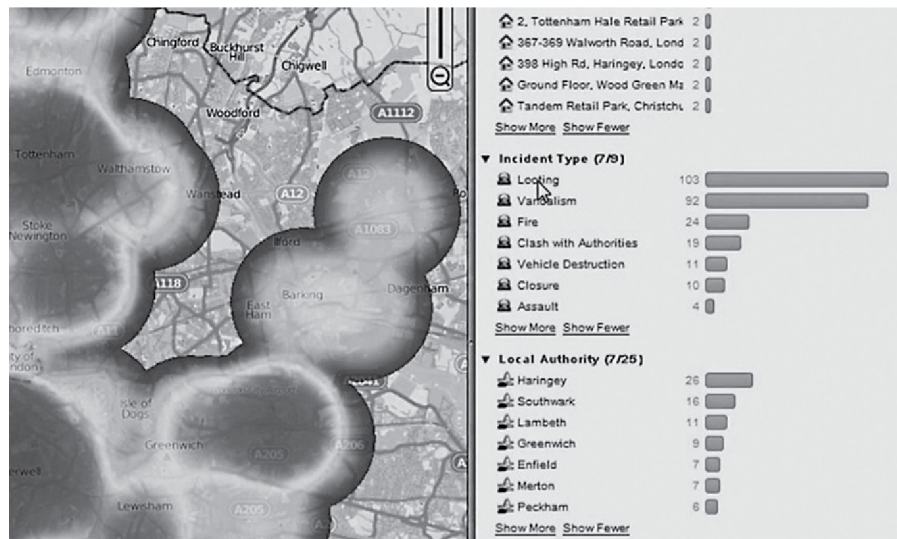


FIGURE 12.3

Illustrates the geospatial distribution of the UK riots by incident type. (Created by Praescient Analytics utilizing the Palantir analytic platform. All rights reserved. Reproduced with permission of Praescient Analytics.)

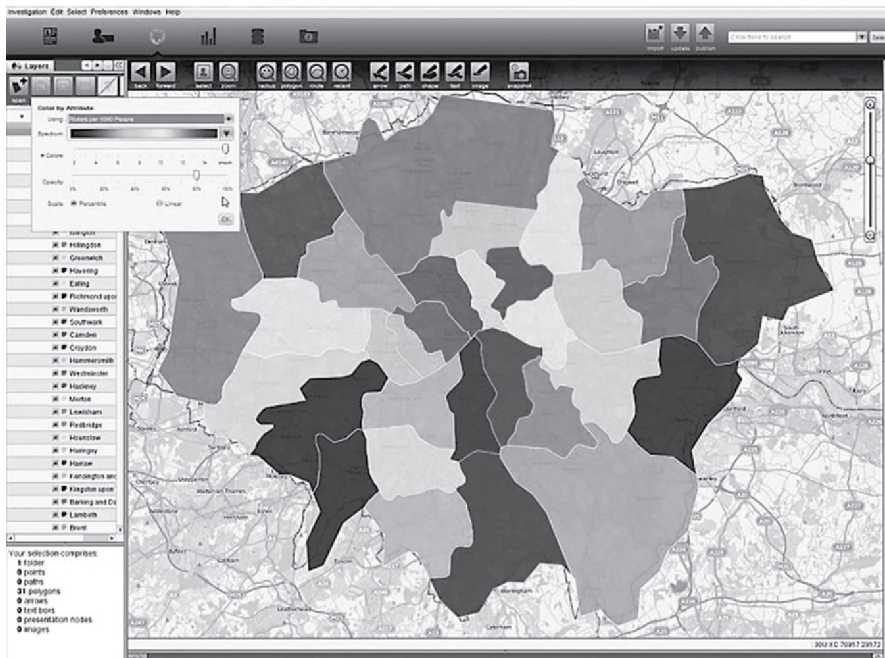


FIGURE 12.4

Per capita distribution of rioters by borough per 1000 people. (Created by Praescient Analytics utilizing the Palantir analytic platform. All rights reserved. Reproduced with permission of Praescient Analytics.)

Specific types of incidents can be selected from the aggregate to further elucidate geospatial distribution by the nature of the incident, which can be used to inform specific response.

While information regarding the total number of rioters is important from a primary response perspective, additional information pertaining to the number of rioters as compared to total population can provide insight regarding the per capita rate of participation in social disorder. As discussed in Chapter 1, information on the total population is essential to establishing a meaningful denominator and truly understanding rates. Therefore, Figure 12.4 illustrates the number of rioters per 1000 people, which enables direct comparison across the entire area of interest. This also can be used for direct comparison to future outbreaks of social disorder in the same or other locations. Again, accurate calculation of a per capita rate based on a known denominator allows for more confident direct comparison between events, time, and location.

We know, however, that it is not unusual to have individuals from outside the community participate in demonstrations, particularly in large demonstrations

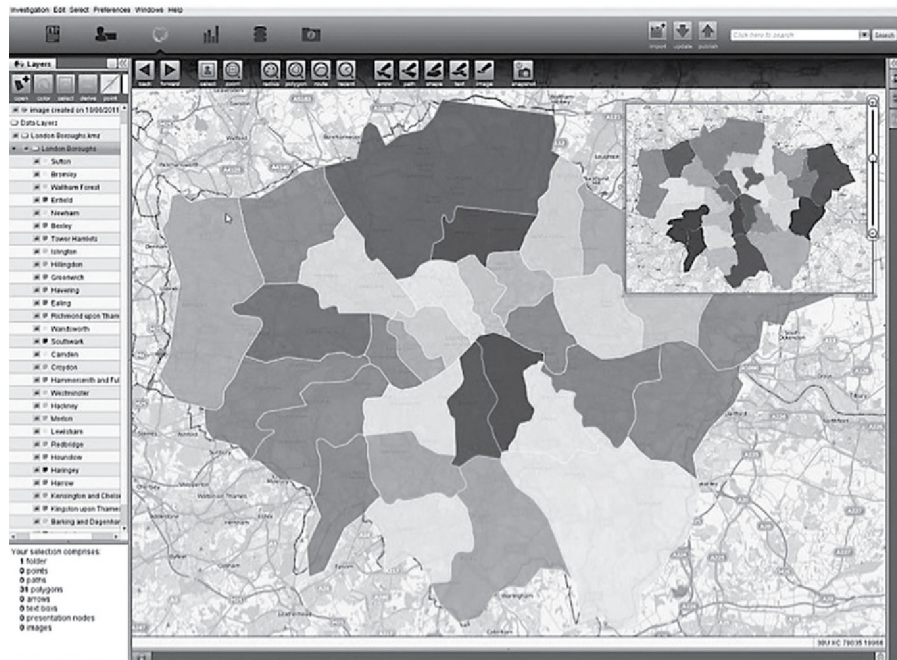


FIGURE 12.5

Location of UK riot origins by borough as compared to riots per 1000 people (Created by Praescient Analytics utilizing the Palantir analytic platform. All rights reserved. Reproduced with permission of Praescient Analytics.)

and other events. These outside agitators can influence and incite crowds, and represent a significant challenge to the public safety community. Figure 12.5 illustrates the origins by borough of the rioters, which reflects migration or travel to the location of the riots. Again, this is important to understanding the origin, nature, and composition of the rioters, which can concomitantly inform prevention, thwarting, mitigation, and response.

Finally, social disorder of the breadth and depth documented during the UK riots is complex in nature and origin. Therefore, by way of evaluating larger social context and putative etiology, the analysts included an assessment of the “Big Society”³⁰ as a measure of social health, cohesion, and efficacy. This was done in an effort to provide additional context and insight regarding the deeper causes for the riots, particularly, it may relate to suppression and/or mitigation of antisocial behavior. While no strong associations between a “Big Society” score and the origin of rioters was found overall (Figure 12.6), this information could be used proactively to guide specifically targeted approaches to preventing or effectively responding to future incidents by leveraging

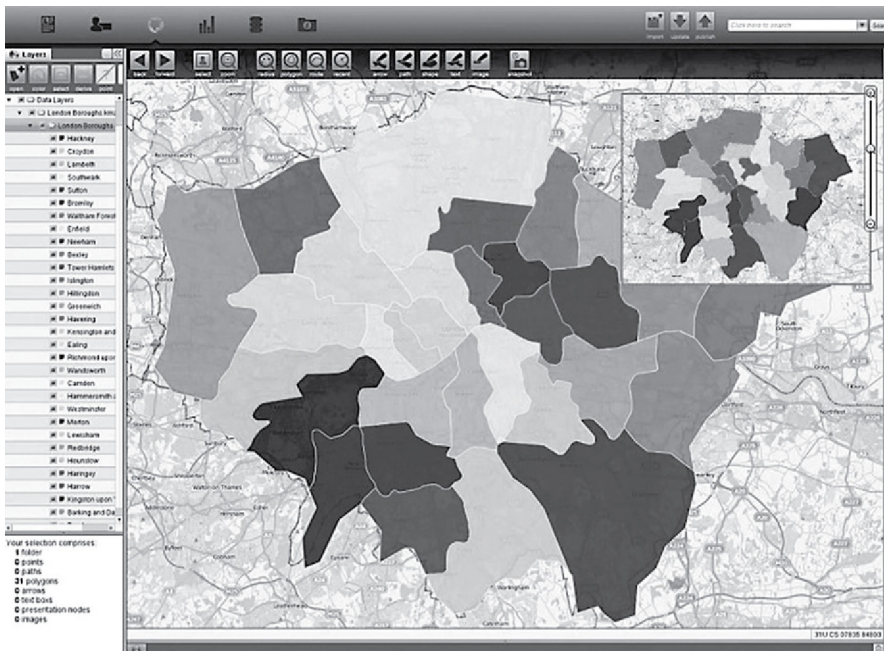


FIGURE 12.6

Illustrates the UK riots according to Big Society³⁸ average score. (Created by Praescient Analytics utilizing the Palantir analytic platform. All rights reserved. Reproduced with permission of Praescient Analytics.)

local social networks and ties to the community in locations that were associated with high Big Society scores. Conversely, in locations with concentrated disadvantage and low Big Society scores, community-based civil society efforts would not be expected to represent an effective approach to reducing social disorder.

As discussed in Chapter 9, this approach aligns well with a fluid, interactive approach to knowledge discovery and insight, enabling the analyst to explore the data and informally develop and test working hypotheses regarding relationships, trends, and patterns. Moreover, the use of multiple capabilities to collect, process, explore, and analyze the data reinforces the importance of letting the problem guide the solution, particularly in a complex challenge, such as that represented by the UK riots and other patterns of social disorder. Again, if the only tool that the analyst carries is a hammer, then every problem looks like a nail. The ability to fluidly move between analytic capabilities, however, better resembles the master carpenter who is able to select and effectively use the appropriate tool, and also can readily adopt new capabilities as they are developed. Finally, reports from the analysts indicate that this entire analysis was completed in 4 h by two analysts; effectively illustrating how the combination

of subject matter expertise with good technology delivers the greatest impact for decision making in real time.

12.4.3 Tahrir Square Protests

By way of a different, yet complementary approach, Figure 12.7 illustrates crowd behavior associated with the Tahrir Square protests.³¹ By analyzing Twitter data, crowd size, density, and relative breadth of voice can be inferred. While breadth of voice and message content can be invaluable sources of data, additional insight regarding the “physics” of the crowd can be used to identify potentially dangerous conditions.³² Historically, large gatherings – spontaneous and well organized – have been associated with tragic outcomes, including panic, stampedes, and related crush deaths. In response to several high-profile incidents, significant research has been conducted on the physical properties of crowds, including indicators and warning associated with pending catastrophe. This research suggests that as long as humans have at least one square yard per person, the crowd remains stable. When the density drops below this threshold, however, the crowd becomes unstable and tragic outcomes become increasingly likely. Therefore, information regarding crowd formation, density, and overall behavior also can be used to anticipate panic, and inform crowd management and response.

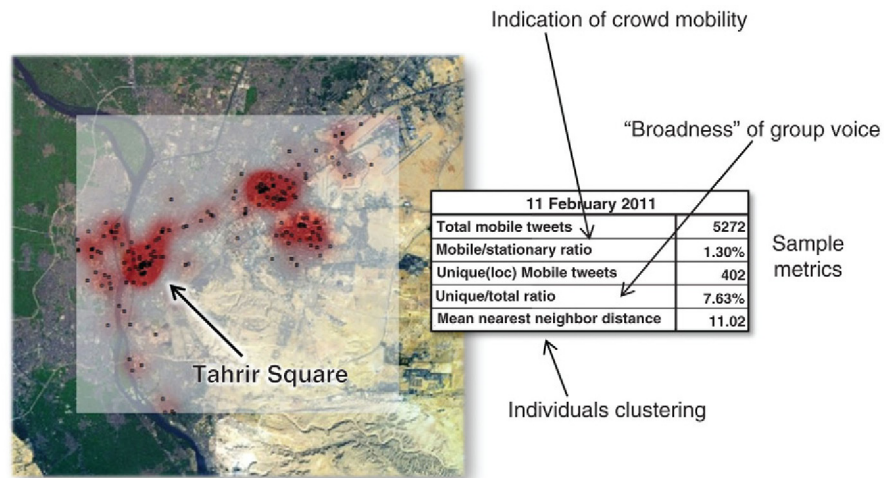


FIGURE 12.7

Illustrates the use of geospatial analysis and summary statistics to visualize and describe social media data associated with the Tahrir Square protests.³⁹ Individual metrics describe the breadth of the voice, and inferred crowd “physics” to include size and density. (*DigitalGlobe, used with permission.*)

12.5 NOVEL APPROACHES TO RISK AND THREAT ASSESSMENT

In *The Gift of Fear*, de Becker describes the small voice most of us have that speaks up and tells us when things are not right or that we are in danger.³³ This is the “gift” of fear. As previously mentioned, it can be very difficult to encourage people to act on their intuition. Some novel approaches to risk and threat assessment, however, use nontraditional means to tap into these gut feelings and intuition, as well as insider information in some cases.

For example, an interesting extension of the “gut feeling” is the finding that crowds tend to be smarter than individuals. Based on his experience with the popular TV game show *Who Wants to Be a Millionaire*, Michael Shermer reviewed the literature that supports the accuracy of group decisions as compared to those made by individuals.³⁴ He found that the audience was correct 91% of the time, as compared to the “experts,” who were correct only 65% of the time on the show. In explaining this finding, Shermer notes that individual errors on either side of the correct response tend to cancel each other out, bringing the group response closer to the truth than an individual response. It is important to note that this finding does not apply to all groups. Critical features of the group include autonomy, diversity, and decentralization to ensure the range of knowledge and opinion for this phenomenon to occur.

This has not gone unnoticed by the US Department of Defense. The Pentagon’s Defense Advanced Research Projects Agency (DARPA) supported research in this area, which included their Electronic Market-Based Decision Support and Futures Markets Applied to Prediction (FutureMAP) programs.³⁵ Briefly, these programs were designed to artificially create groups that incorporated the diversity of thinking found to be required for accurate group-based decision making. By tapping into the knowledge from a varied array of experts, it was hypothesized that the consensus opinions would be superior to those generated by individuals, even if these individuals were experts in their respective fields. The DARPA scientists used market-based techniques to compile and consolidate these diverse opinions and generate a unified response. This concept was not new to the Department of Defense. In 1968, naval scientist John Craven assembled a group of submarine commanders in an effort to find the missing submarine *Scorpion*.³⁶ Using Bayes’ Theorem and consensus expert opinions generated by the commanders, Craven was able to construct an effective search strategy and find the submarine.

The DARPA programs used these market-based approaches to generate estimates of the likelihood of specific events of interest to the Department of Defense. These included estimates for the development or acquisition of certain technologies, as well as estimates of political stability in certain regions. The ultimate goal of these programs was to consolidate opinion from a variety of sources,

including expert opinion and insider information, in an effort to accurately predict future events and avoid surprise attacks. Unfortunately, this program came under attack and was cancelled when it became known publicly that terrorist attacks and assassinations were included in the events of interest. The public outcry that ensued in response to the idea that the United States government was essentially betting on tragedy was more than enough to terminate the program.³⁷

Interestingly, this concept still exists. The website www.Tradesports.com supports an electronic market that includes subjects of interest to those tasked with preventing future terrorist attacks and supporting homeland security. Tradesports.com describes itself as a “person-to-person trading ‘Exchange’” where individuals can trade directly on a variety of events including sports, weather, entertainment, legal outcomes, and politics, to name just a few categories. Tradesports.com also accepts “contracts” on current events, including anticipated events related to the war on terrorism. For example, at the time of this writing, current contracts relate to whether Osama Bin Laden and Abu Musab al-Zarqawi will be “captured/neutralized” by a certain date. Their market-based predictions tend to be highly accurate, most likely due to the same factors that DARPA was trying to exploit. Tradesports.com taps into a very large sample that includes a diverse array of individuals with expertise in a variety of areas. These opinions very likely include insider information in a variety of areas that can further enhance the accuracy of the consensus opinion generated. The exchange consolidates these opinions and generates a consensus probability. These market-based approaches incorporate the speed and agility necessary to effectively track issues that may fluctuate rapidly. Public opinion can change on a dime, far faster than most existing collection methods. As a result, tools like these bring the speed and agility required to instantly document changes and effectively track fluid trends.

It is important to remember, though, that groups also are able to generate some very bad consensus opinions. For example, expectations regarding how significant events may affect gasoline prices or stock prices can actually alter these events, albeit temporarily. Like any risk assessment tool, group opinion is only as reliable as the inputs. Bad information results in inaccurate and unreliable predictions, regardless of the method used to calculate the risk. The value that can be added by compiling and integrating diverse expert opinions cannot be underestimated, however, and supports the importance of a close working relationship between the analytical and operational personnel. As always, solid domain expertise and a healthy dose of skepticism are necessary tools in the evaluation of risk and threat.

Bibliography

- 1 Lind, W.S., Nightengale, K., Schmitt, J.F., Sutton, J.W., and Wilson, G.I. (1989). The changing face of war: Into the fourth generation. *Marine Corps Gazette*, October, 22–26. <http://www.blackwaterusa.com/btw2004/articles/0726sheep.html>

- 2 Ripley A. *The unthinkable: who survives when disaster strikes – and why*. New York: Three Rivers Press; 2009.
- 3 Giduck J. *Terror at Beslan*. Bailey, CO: Archangel Group; 2005.
- 4 Clowes E. as cited in: Scotti RA. *Sudden sea: the Great Hurricane of 1938*. New York: Chapter and Verse; 2003.
- 5 Stewart J, Stewart JB. *Heart of a soldier*. New York: Simon & Schuster; 2003.
- 6 Ripley A. *The unthinkable: who survives when disaster strikes – and why*. New York: Three Rivers Press; 2009.
- 7 Characteristics of possible preincident surveillance or behavior have been compiled, including various types of surveillance activity and security probes (International Training, Inc., ArmorGroup, Department of Homeland Security Surveillance Detection Training for Commercial Infrastructure Operators and Security Staff Course Syllabus. https://www.fbiic.gov/public/2009/jan/SD_CI_Syllabus.pdf; 2006).
- 8 Grossman, D. and Christensen, L.W. (2008). *On Combat, The Psychology and Physiology of Deadly Conflict in War and in Peace* (3rd ed.). Warrior Science Publications: Milstadt, IL.
- 9 Dorn M, Dorn C. *Innocent targets: when terrorism comes to school*. Macon, GA: Safe Havens International; 2005.
- 10 See, Dorn, M. and Dorn, C. (2005). *Innocent targets: When terrorism comes to school*. Safe Havens International, Macon, GA.; and Giduck, J. (2005). *Terror at Belsan*. Archangel Group, Royersford, PA.
- 11 Dorn, M. and Dorn, C. (2005). *Innocent targets: When terrorism comes to school*. Safe Havens International, Macon, GA.
- 12 Grossman D. Lecture for ArmorGroup, International Training, Richmond, VA; 2005, October 31.
- 13 Grossman D. 2005.
- 14 Giduck J. *Terror at Beslan*. Bailey, CO: Archangel Group; 2005.
- 15 Giduck J. 2005.
- 16 U.S. Department of Homeland Security. What is critical infrastructure? <http://www.dhs.gov/what-critical-infrastructure>. For additional insight, c.f., The White House. Presidential Policy Directive – Critical Infrastructure Security and Resilience. Presidential Policy Directive/PPD-21. <http://www.whitehouse.gov/the-press-office/2013/02/12/presidential-policy-directive-critical-infrastructure-security-and-resil>; 2013.
- 17 Additional description and related detail regarding each sector can be found at: U.S. Department of Homeland Security. Critical Infrastructure Sectors. <http://www.dhs.gov/critical-infrastructure-sectors>
- 18 Fink S. *Five days at memorial: life and death in a storm-ravaged hospital*. New York: Crown; 2013.
- 19 Quammen D. *Spillover: Animal Infections and the Next Human Pandemic*. New York: W.W. Norton; 2013.
- 20 In a real-world test of this model, many of the victims of the Nairobi Westgate Mall attack spontaneously sheltered in place.
- 21 Several years ago, one particular hotel chain was associated with a series of attacks. While it was not immediately clear whether this reflected a known vulnerability in their security, targeting on the basis of some attribute of the chain and/or their ownership, or just bad luck, there was some very deliberate travel planning on the part of the community familiar with the pattern.
- 22 De Becker G. *The gift of fear*. New York: Dell; 1997.
- 23 Ibid.
- 24 Grossman D. 2005.
- 25 Ibid.
- 26 Grossman D. 2005.
- 27 McCue C, McNulty PJ. Gazing into the crystal ball: data mining and risk-based deployment. *Violent Crime Newsletter* 2013, September, 1–2.

- 28 Akwagyiram A. England riots: one year on. BBC News. <http://www.bbc.com/news/uk-19077349>; 2012 [accessed 06.08.2012].
- 29 Lasko A. Analyzing the UK riots – Praescient Analytics using Kapow Software and Palantir. <http://www.youtube.com/watch?v=hCszSkhiojo>; 2012.
- 30 Civil Exchange. The Big Society Audit 2012. http://www.civilexchange.org.uk/wp-content/uploads/2012/05/THE-BIG-SOCIETY-AUDIT-2012_Civil-ExchangeFinal8May.pdf; 2012.
- 31 For additional discussion of this example, cf., Chapter 6 (Hildebrandt W, McCue C. Unbiased analytics for the COCOMs. AHFE 2012).
- 32 For a good overview of the research, cf., Ripley A. The unthinkable: who survives when disaster strikes – and why. New York: Three Rivers Press; 2009.
- 33 De Becker G. 1997.
- 34 Shermer M. Common sense: surprising new research shows that crowds are often smarter than individuals. ScientificAmerican.com; <http://www.sciam.com/article.cfm?chanID=sa006&articleID=00049F3E-91E1-119B-8EA483414B7FFE9F&collID=13>; 2004.
- 35 DARPA – FutureMAP Program. Policy analysis market (PAM) cancelled. IWS – The Information Warfare Site; <http://www.iwar.org.uk/news-archive/tia/futuremap-program.htm>; 2003, July 29.
- 36 Sontag S, Drew C. Blind man’s bluff: the untold story of American submarine espionage. New York: HarperCollins; 1999.
- 37 Yeh PF. Using prediction markets to enhance US intelligence capabilities: a “Standard & Poors 500 Index” for intelligence. Studies in Intelligence 2006; 50(4). <https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol50no4/using-prediction-markets-to-enhance-us-intelligence-capabilities.html>; 2006.
- 38 Civil Exchange. The Big Society Audit 2012. http://www.civilexchange.org.uk/wp-content/uploads/2012/05/THE-BIG-SOCIETY-AUDIT-2012_Civil-ExchangeFinal8May.pdf; 2012.
- 39 Hildebrandt W, McCue C. Unbiased analytics for the COCOMs. AHFE 2012.

Deployment

“Chance favors the prepared mind.”

Louis Pasteur

In many ways, the goal of force deployment is similar to most resource allocation problems – try to do the most work possible with existing, or even fewer resources. The deployment challenges facing police executives and command staff currently are multifaceted and complex, and include ensuring public safety through proactive approaches to crime prevention and deterrence, maintaining the ability to respond to those incidents that do occur and all other citizen calls for service in a timely manner, and serving as responsible stewards of increasingly limited public safety resources by efficiently allocating frequently scarce and overburdened personnel resources when and where they are likely to do the most good. Unfortunately, police deployment decisions often are made based on historical precedence, gut instincts, vocal constituent groups, or uniformed policy makers. When crime is increasing the first response frequently includes calls for increased deployment of personnel or a “cop on every corner”¹; however, most public safety agencies do not have the luxury of addressing rising crime rates with the indiscriminate use of resources, particularly given shrinking law enforcement budgets, mission creep, and even outright reductions in force.

Information-based approaches to the deployment of police resources can address all three challenges facing the public safety and security executive, and other command staff. By identifying areas likely to be associated with increased demands for service, police managers and command staff can proactively allocate resources when and where they are likely to be needed. The ability to preposition resources in anticipation of need increases the likelihood that they will be able to prevent or otherwise thwart crime from occurring. Moreover, by being prepositioned in anticipation of need, they also are able to respond more rapidly to incidents that do occur, which increases the likelihood that they will apprehend the perpetrator, thereby limiting his or her activity going forward. This approach concomitantly reduces the likelihood that assets

will be placed in locations where they are not needed, which not only wastes valuable resources but also may limit their ability to respond to incidents that do occur in a timely manner. Ultimately, information-based approaches to deployment offer improved public safety outcomes by very specifically characterizing anticipated need and influencing outcomes through targeted tactics and strategy, thereby increasing the likelihood that chance will favor the prepared mind.

13.1 RISK-BASED DEPLOYMENT

Several years ago the Richmond Police Department began operationalizing this concept through the use of predictive analytics in support of “risk-based” deployment strategies as part of their role in the Project Safe Neighborhoods initiative with the United States Attorney General’s Office in the Eastern District of Virginia.² Taking a cue from the business community, including the concept of “just-in-time” supply chain management, the team developed a novel, information-based approach to more efficient and effective allocation and optimization of resources, enabling police managers and command staff to do more with less in an increasingly complex but resource-constrained threat environment.³ A detailed example of this approach is discussed later in this chapter, but the fundamental idea behind the concept was that locations could be identified and characterized as being at increased likelihood or risk for crime or other calls for service through the use of predictive analytics. Personnel resources could then be proactively deployed using the modeling output; setting tactics and strategy to the specific pattern of offending or risk identified and characterized. Ideally, crime would be prevented or otherwise thwarted through proactive approaches, including heavy deployment, by creating an environment unfavorable to crime. The second best would be to preposition resources in order to respond more rapidly and influence future outcomes by apprehending criminals, thereby taking them off of the streets and reducing their criminal activity. Ultimately, increased use of this method provides an opportunity to move from counting and reporting crime into proactive policing, whereby crime can be thwarted or prevented.

13.2 GENERAL CONCEPTS

In reviewing the timing of citizen-initiated complaints, we noted that most agencies experience a typical cycle. On weekdays, calls might be steady throughout the day, picking up to a brisk pace in the evening, and then slowing down after midnight. This pattern might be altered somewhat on the weekends, particularly if the community has an active entertainment district or special events. For example, calls might continue after midnight, spiking for a period

around the time that the various nightspots close and the patrons begin flowing out into the streets.

Complicating the model, however, is the fact that the specific nature of these calls frequently differs throughout the day. For example, there might be more commercial robberies during the day, an increase in domestic complaints when people return home from work in the evening, and more street robberies and vice complaints in the later evening. Activity after midnight might be confined almost exclusively to alarm calls. Moreover, the time required to clear a call and the personnel requirements for a malfunctioning alarm as compared to a highly charged domestic situation will differ greatly, so we cannot rely exclusively on the number of calls for service. The nature of the complaint must be included in any analysis.

Seasonal fluctuations might bring an increase in the number of street robberies related to a large, transient tourist population, while bad weather can alternately suppress some types of crime while increasing others, and special events will bring unique issues all their own. The number of variables that can affect patrol deployment requirements is almost limitless, and each community will have its own unique array of issues and circumstances that impact police workload. It is easy to see that we have quickly exceeded the computational capacity of a pocket calculator or a simple spreadsheet program for determining demand for police services and related deployment strategies.

By using data mining and predictive analytics tools, analysts are able to consider multiple factors simultaneously and drill down to determine and characterize further the unique constellation of risk or activity associated with a particular area, location, or time period. It often can be interesting to see how the manifest patterns of activity and risk flow throughout time and space as the analyst drills down further, revealing additional detail and added refinement. This can be particularly true for relatively arbitrary distinctions like day of week. For example, what might appear to be unusual activity associated with a particular day of the week could merely reflect the continuation of activity from the previous day. Ultimately, through the use of classification models or scoring algorithms, decision rules can be created for the specific pattern of risk or deployment needs, which can guide the development of effective deployment strategies.

13.2.1 Data

There are some obvious data that should be included in a deployment model. For example, citizen complaint data frequently are the most direct representation of citizen-initiated police work. Beyond absolute number of calls, though, it also is important to include the number of officers required for each call, as well as the estimated time to clear a call. Any additional crime-related data also should be included in the model. For example, illegal narcotics or vice-related

arrests might further enhance a deployment model by incorporating underlying crime that might not be manifest in citizen complaint data. This can be particularly true in areas where the issues associated with open-air drug selling pale in comparison to high rates of violent crime and might not be included in complaints.

This also is an opportunity to think outside of the box regarding what types of additional information might have value with respect to crime prediction. These resources have been mentioned previously but are worth repeating, as they can add considerable value to our analysis of crime data and ultimately result in more accurate, complete predictive models.

13.2.1.1 Weather and Climate Data

As discussed in Chapter 5, weather may play a role, sometimes significantly, in influencing crime trends and patterns. This is particularly true as relates to the distribution of and access to potential victims or targets. Weather data usually are readily available and can add considerable value to a deployment strategy. For example, inclement weather might suppress some patterns of criminal activity that exploit victim routine activities. For example, a severe winter snowstorm that significantly limits travel would be expected to decrease the number of street robberies since fewer potential victims will be out on the streets. On the other hand, the number of traffic wrecks associated with bad weather or hazardous driving conditions might place additional demands on traffic units and patrol. The ability to anticipate these changes in demand can facilitate a fluid transition from one operational strategy and associated deployment plan to another. Therefore, insight regarding the role that weather may play in influencing crime and other patterns requiring law enforcement response can be used to inform patrol deployment and resource allocation.

Seasonal changes also can be associated with changes in demand for police services. Especially cold weather might be associated with an increased number of vehicle thefts from residential areas as people preheat their cars in the morning, and very hot weather might be associated with a similar increase in stolen vehicles from convenience store parking lots or daycare centers as citizens leave their cars running in an effort to keep them cool. These and other trends can be identified, characterized, and modeled to support proactive approaches to crime prevention and response. The associated scoring algorithms can be deployed with triggering mechanisms that would prompt an immediate modification in a deployment plan in response to changing conditions that are predicted to be associated with changing needs for police services. Personnel resources then can be flexed proactively, rather than being placed in a reactive position in response to changing conditions, which can significantly impair the efficient use of resources and result in poor provision of services to the community. In our experience, electronic copies of archival weather data

suitable for inclusion in deployment modeling and analysis are available from a variety of sources, including the local news stations.

Finally, one advantage associated with some of the more powerful data mining and predictive analytics tools is their speed and agility. This gives the analyst the ability to process large, rapidly changing data sets quickly in response to rapidly changing situations like those associated with large manmade and natural disasters and the associated recovery and rebuilding periods. These new tools also allow for flexibility in recoding variables, including spatial data, which can be critical when the existing geography has been changed dramatically by a natural or manmade event. Ultimately, these capabilities can be used to develop fluid deployment models that can accommodate up-to-the-minute information regarding conditions and allow public safety and security organizations to assume proactive rather than reactive approaches to deployment.

13.2.1.2 Human Terrain

As discussed in Chapter 5, information related to population density, relative wealth, and other population-based measures can guide deployment strategies. For example, some patterns of larceny and economic crimes are more likely to occur in wealthier environments. On the other hand, increased population density will require heavier deployment just based on the sheer number of people living in a particular area. Again, location matters in crime, particularly as relates to the distribution of and access to potential victims and/or targets, as well as the identification of a favorable environment in which to successfully perpetrate the act. Criminals frequently will search for a “target-rich” environment in which to commit their crime; whether a known open-air drug market or a residential area with an abundance of high-end electronics and vehicles. Therefore, while “target rich” might be defined as population density in one community and relative affluence in another, census data and other sources of human terrain can provide a valuable addition to crime modeling.

Other information to consider when creating deployment models includes changes associated with transient population fluctuations. For example, resort communities or college towns can experience extremely large population fluctuations. Tourists often make easy targets for pickpockets and robbers, while college student populations might be associated with increases in illegal narcotics trafficking, alcohol and other vice-related offenses, and sex offenses, including date rapes. Anticipation of these population fluctuations could trigger associated modifications in patrol deployment plans, as well as proactive crime prevention strategies targeting these anticipated crime trends. Similarly, special events can create transient increases in patrol demand. Concerts or sporting events that are associated with disorderly crowds can be characterized and modeled, creating deployment strategies to target specific issues and challenges known to be associated with these particular events.

Domain expertise, as always, is critically important in the creation of deployment strategies. For example, communities with a thriving entertainment district or seasonal trends in tourism will require very different deployment strategies when compared to locations with rampant drug-related crime. There is no “one size fits all” deployment strategy that will have value and meaning for every locality. Even within the same city, crime might flex and flow around different areas during different times and be based on very different factors. Therefore, it is essential that all deployment-related data mining and predictive models be viewed and reviewed within the context of prevailing community issues and needs. Resource deployment represents one of the most critical functions within any public safety organization, not only due to the personnel and economic assets involved, but because it can have such a critical impact on public safety.

13.3 HOW TO

The first task when examining issues related to patrol deployment generally includes characterization of the data and information through the use of exploratory graphics,⁴ which allows the analyst to drill down into the data and convey relative differences visually. The use of graphics permits visual review and analysis of the information in a format that is inherently actionable from an operational standpoint.

13.3.1 Time

Figure 13.1 represents a notional example of the use of descriptive statistics to inform deployment. In this particular example, more than half of the 266 citizen calls for service during the week selected for analysis (58%) occurred during the 3-day period between Friday and Sunday. This initial analysis

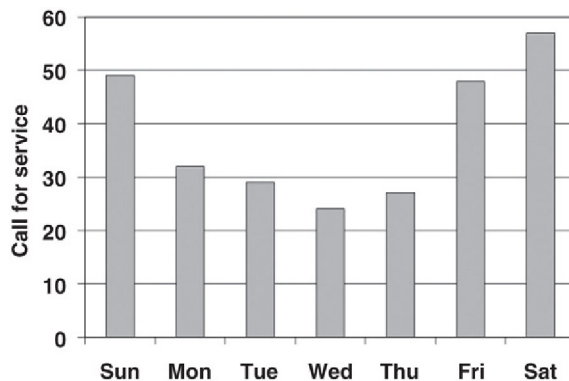


FIGURE 13.1

Frequency distribution depicting citizen calls for service by the day of week during a 1-week period.

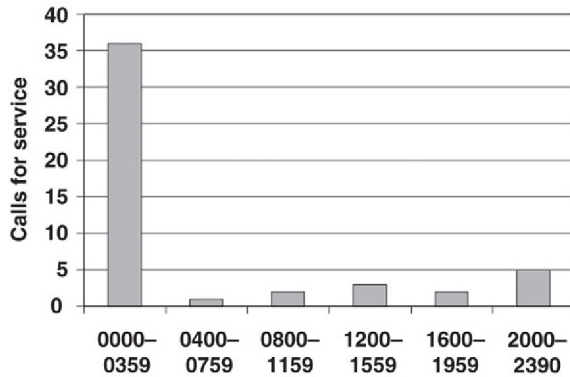


FIGURE 13.2

Citizen calls for service from [Figure 13.1](#) by time of day.

already adds some value and specificity to the understanding of the differential distribution of police work throughout the week, which has implications for deployment. By looking at this distribution, the command staff or supervisors can determine that 3 days out of the week accounted for more than 1/2 of the citizen-initiated police work during this week. Further analysis of the types of complaints, the number of units required, and the time to clear each call would add additional value to this analysis, although each of those factors would add another level of complexity to the analytical task and model, which would require increasingly powerful analytical resources. At a minimum, though, this analysis has value from a deployment perspective.

To further refine the specific deployment strategy, the information can be subdivided by time of day ([Figure 13.2](#)). For example, drilling down into the service calls reported on Sunday reveals that most of them occurred between midnight and 0400 h. In other words, they represented a continuation of activity from Saturday night. This information, which indicates that the complaints are not distributed uniformly throughout the day in many cases, would have significant implications for deployment strategies. It also would be worthwhile to examine what specific types of calls were associated with different periods throughout the day, and what implications this might have for deployment. For example, alarm reset calls that occur in the middle of the night generally do not require extensive personnel resources and can be relatively quick to clear. A disorderly disturbance call associated with closing time at a nightclub, however, might require multiple units and take a considerable amount of time to clear, particularly if arrests are involved. Again, a specific analysis of time is important to ensure that adequate police resources are available when and where they are needed and that deployment is limited during times when the need for police presence is less.

Friday

District	Shift					
	0000–0359	0400–0759	0800–1159	1200–1559	1600–1959	2000–2359
1						
2						
3						
4						
5						

FIGURE 13.3

A hypothetical police deployment schedule that includes time of day and police district for a particular day of the week, in this case, Friday. This schedule was created using the results of a self-organizing network called a Kohonen network.

13.3.2 Space

Determining deployment requirements across day of the week and time of day certainly puts more science and less fiction into resource allocation, but a good deployment model also needs to take into account spatial differences in the need for a police presence. [Figure 13.3](#) depicts a hypothetical police deployment schedule, which was extracted from a self-organizing network called a Kohonen network, that includes time of day and police district for a particular day of the week, in this case Friday. In this notional example, the model was created through the use of an algorithm, which associates a relative degree of risk for crime associated with a particular time and location for the day of interest. Relative levels of risk for crime have been depicted visually as relative densities to fill in the various time blocks.

As can be seen in the [Figure 13.3](#), Third District is associated with the greatest need on Fridays, from 1200 to 1359 h. Examination of the data revealed that Gotham High was located in Third District, and that there had been a large

fight during a football game with their uptown rivals, the Spartans from Groverville East. Further review indicated that this was a regular challenge associated with this time and location, and that heavier proactive deployment and some collaboration with the schools probably would address this issue.

The increased demand in Second District was linked to a regular after-work party each Friday near the business corridor, while the increased demand in First and Fifth Districts was associated with a large block of nightclubs that transcend the boundary between those two districts.

By incorporating time and space into a deployment schedule like the one illustrated in [Figure 13.3](#), the analytical team was able to deploy the results of their analysis into a format that was inherently actionable for the operational command staff. While additional options were available for further enhancements to the models by the inclusion of additional call-related details, this deployment model represents an information-based schedule that has considerable value over what had been used previously.

This example was created to be relatively simple in an effort to highlight specific points. In the applied setting, a sampling frame longer than 1 week almost certainly would be used unless there were very specific reasons for choosing otherwise, such as the creation of a focused, short-term deployment model or initiative that was linked to a specific time period, similar to the New Year's Eve initiative described later. In addition, deployment models should be evaluated and refreshed periodically to ensure that they continue to address requirements for police-related work appropriately. In many ways, it can be a sign of success that the models need to be adjusted periodically. As crime patterns are addressed, suppressed, or displaced, the model needs to be refined and redeployed to accommodate the successes as well as new or rapidly emerging trends.

13.3.3 Nature of the Incident or Threat

In addition to knowing the “when” and “where” of crime trends and patterns, insight regarding the specific nature or the “what” of crime is important to selecting appropriate tactics and strategy to effectively address the specific challenge or threat.

Several years ago, a review of the homicide data from Richmond, Virginia, showed an extremely high homicide rate that placed it repeatedly in the top 10 in the nation for its per capita murder rate. Several initiatives were created in an effort to address this problem, including a federally funded homicide reduction program. Research into the homicide rates confirmed what everyone knew to be true that the homicide rate in Richmond was increasing rather dramatically when rates in other locations were decreasing.⁵ In and of itself, however, the overall homicide rate in a community generally contributes little value to

a thoughtful understanding of the possible causes, nor are aggregate numbers likely to be associated with the development of any meaningful or long lasting solutions. Further inquiry was necessary.

Drilling down into the data, it became rapidly apparent that almost all of the increase noted over the examined time period could be attributed directly to increases in the drug-related homicide rates. This indicated that proactive measures designed specifically to address drug-related homicides would go a long way toward reversing these trends. Further parsing of the data indicated that at the same time that the drug-related homicide rate was increasing rapidly, the average age of the victims and suspects was decreasing. In other words, if violence is a disease, as it has been characterized so frequently, then drug-related violence is a disease that differentially impacts the young.⁶

Why is this important from a deployment standpoint? If the goal of a particular initiative is violence reduction, then it is important to understand in some level of detail what is driving the trends and who is involved. In this example, the rapid increase in the murder rate could be attributed directly to an increased prevalence of drug-related homicide. Therefore, characterization of the problem provided some direction regarding possible solutions. Had the increases in the homicide rates been associated with the activity of a serial killer or with domestic violence, different strategies would have been indicated. So, characterizing the problem is frequently the first step in identifying a meaningful solution.

13.3.4 Operationally Actionable Output

As has been stated several times before, analytic output, no matter how elegant, sophisticated, or accurate, has no value in an operational environment if it is not actionable. This is particularly true with deployment.

Again, the ability to effectively convey the analytic results can determine whether the results are used operationally or not. If the command staff, operators, or other end users are not able to interpret and use the results, then it significantly limits the utility of the analysis. The “generalizability versus accuracy” challenge has been addressed earlier; however, even a relatively simple predictive algorithm has limited value if it cannot be used in the operational environment. One excellent method that can be used to convey complex results in an operationally relevant and actionable manner is mapping. Adding to the value of this approach, the context provided in the mapping environment also enables the end user to incorporate their tacit knowledge and domain expertise in the interpretation and even extension of the results in support of novel insight and discovery. Examples of data mining and predictive analytic products that are deployed through geospatial tools are highlighted throughout this text. From a deployment standpoint, however, there are few methods that even

approach the effectiveness and functional utility associated with mapping. Moreover, many patterns of offending tend to be geographically dependent. Therefore, geospatial information has tremendous value from a deployment perspective because, at its most primitive, deployment generally involves prepositioning personnel and other assets in time and space. Ideally, personnel are deployed in response to anticipated patterns of offending or calls for service so that they can influence outcomes, something that data mining and predictive analytics can facilitate. Through the creative use of geospatial tools, meaningful differences in time and space can be conveyed in an actionable format to operational personnel.

13.4 RISK-BASED DEPLOYMENT CASE STUDIES

“Telling the future by looking at the past assumes that conditions remain constant. This is like driving a car by looking in the rearview mirror.”

Herb Brody

The idea behind the use of highly mobile tactical units is their ability to respond quickly to rapidly changing events. Unlike traditional patrol units, which tend to be anchored to a particular geographic area, tactical units can be deployed proactively to areas in anticipation of an increased need or a rapidly emerging situation. With this in mind, it would seem ideal to identify a way in which trouble could be anticipated. This would support the concept of proactive deployment; at a minimum, permitting a rapid response to incidents. Ideally, heavy deployment in these areas would result in crime deterrence.

In many situations, “proactive” deployment decisions are based on a historical review of crime patterns, including pin maps. While these can be great for counting crime and depicting it within the context of geography, they do little to inform us of the future (Figure 13.4). By using data mining and predictive analytics, however, areas associated with an increased risk for certain types of crime can be modeled and mapped. This might seem like such a subtle distinction as to have no value in law enforcement, but read on.

13.4.1 Robbery-Related Aggravated Assaults

The Richmond, Virginia, Police Department, as part of the Project Safe Neighborhoods strategy, began developing the use of risk-based deployment strategies in an effort to reduce gun violence.⁷ As the thinking goes, if armed robberies are bad, then an armed robbery where the victim gets hurt is worse. Is there a way to model this so that we can deploy our tactical units proactively and keep people safe? The challenge was that the created model had to be simple enough to be actionable. It also had to be confined to variables that had value



FIGURE 13.4

Relying exclusively on historic events to guide future deployment is similar to trying to drive a car by looking only in the rearview mirror. It tells you where you have been, but not necessarily where you are going. *Courtesy of SA BJ Kang.*

from a deployment standpoint. For example, it might be that robbers who are verbally aggressive with their victims are more likely to assault them, but a police department cannot proactively deploy for aggressive robbers, so this particular variable had little value from a deployment standpoint.

Information pertaining to geography, time of day, and day of week were included in the analysis, while almost everything else, including type of weapon and suspect-related information, were excluded. The resulting model was relatively accurate, although not perfect. It should be noted that more accurate models were developed and deployed through the web-based analytics described earlier,⁸ but the associated algorithms were too opaque to be deployed directly into the operational environment.

At this point, some very good questions might be, “Does it matter that the accuracy is reduced?” or “How low can you go?” This issue has been addressed in much greater detail in previous chapters because it is extremely important. Briefly, in this case it was determined that anything that would increase the efficacy of the deployed resources above chance would be considered acceptable, because even a slight increase in public safety associated with the use of the model could potentially save lives. For that reason, the errors associated with the model were shifted toward being somewhat generous in terms of deployment.

The consequences associated with missing a potentially high-risk circumstance were determined to be much more serious than deploying resources to an area where they might not be needed. So the model was adjusted to permit more “false positives” in an effort to increase the likelihood that the personnel would be in place when and where they were needed.

Another challenge associated with this deployment initiative was the fact that armed-robbery-related aggravated assaults are relatively infrequent. While this generally is a very good thing for a community, particularly for potential victims, it can be a significant challenge from a modeling perspective. There were a relatively limited number of incidents of interest for use in the creation of the model. In addition, it was important to ensure that the number of incidents predicted were similar to the frequency of observed events. Many modeling algorithms are preset to a 50:50 distribution, which would be extremely inaccurate in the current situation.

For this same reason, close attention to the nature of the errors was extremely important. Because less than 5% of all armed robberies in our sample escalated into an aggravated assault, it would have been possible to create a very simple algorithm that was correct 95% of the time by simply predicting that an armed robbery would never escalate. While the accuracy associated with this model would be enviable, particularly in an applied setting like ours, it would do little to inform the deployment process, which would mean that it has no value.

Another challenge associated with such a low incident rate was the creation of separate training and test samples, which is addressed in detail in Chapter 8. In the current example, the sample was randomly split into training and test samples. These samples were then evaluated to ensure that the factor of interest, robbery-related aggravated assaults, was distributed evenly and that there were no unusual differences between the two groups. Because the samples were so small, these group assignments were subsequently maintained throughout the modeling process.

Figure 13.5 illustrates the results, highlighting areas that were predicted to be at greater risk for a robbery-related firearms assault based on a review of the armed robbery data for a 6-month period. It is important to note, however, that the areas identified by the model and highlighted on the map are not predicted to be associated with an increased number of armed robberies; rather, the armed robberies in these areas are predicted to be more risky. This is an important consideration for the type of deployment and this particular strategy. Traditional deployment models generally consider only the frequency of crime and deploy accordingly. With specialized units, however, there is tremendous advantage in being able to deploy based on predicted risk of more serious crime, rather than relative frequencies of multiple patterns of offending.



FIGURE 13.5

A risk-based deployment map for robbery-related aggravated assaults. The dots represent armed robberies, while the shaded areas are associated with a greater likelihood for a robbery-related firearms assault.²⁰

Therefore, by using a risk-based deployment strategy, areas predicted to be at greater risk for more serious patterns of offending are identified and highlighted for selected, tactical enforcement strategies, while patrol resources can be deployed in response to other types of crime and citizen-initiated work.

Finally, in our experience it has been necessary to periodically refresh the models. Conditions change in the community and criminals are apprehended, which can serve to diminish the predictive efficacy of the models. This can be a sign of tremendous success.

13.4.2 Auto Thefts

New improvements in the deployment of predictive analytics output and scoring algorithms now allow the use of web-based analytics.⁹ With these systems, an analyst is able to analyze data and create rule sets or scoring algorithms that

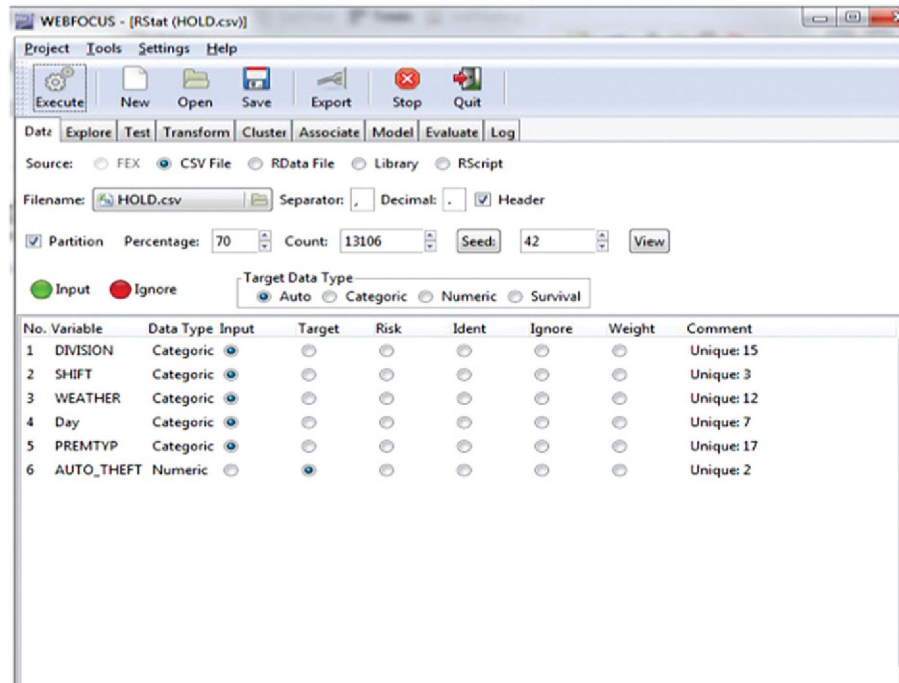
can be deployed remotely. These systems are no more difficult to use than making an online purchase. In many ways, it is much easier than using the remote data entry systems that many law enforcement organizations have adopted in that the amount of information required for this transaction has been limited significantly. The only data required by the model is the information determined to be relevant and necessary by the deployed algorithm. Additional nonessential data entry is minimized, which can be extremely important in an operational setting. After the information has been entered, the output is available quickly, with form and content that is relevant to the operational needs of the end user.

The following example illustrates the use of such a system in the analysis of auto thefts.¹⁰ Using the process model outlined in Chapter 4, the relevant crime incident data are collected. As can be seen in Figure 13.6, basic crime incident data have been compiled. Using indicator variables, auto thefts are identified by a “1,” while all other incidents are coded as “0.” Some preliminary recoding

<u>DIVISION</u>	<u>SHIFT</u>	<u>WEATHER</u>	<u>Day</u>	<u>PREMTYP</u>	<u>AUTO THEFT</u>
Airport Division	4PM-MID	CLEAR	Monday	APARTMENT	1
Airport Division	4PM-MID	CLEAR	Monday	COMMERCIAL PARKING LOT/GA	0
Airport Division	4PM-MID	CLEAR	Monday	HOTEL/MOTEL/ETC.	0
Airport Division	4PM-MID	CLEAR	Monday	RESIDENCE/HOUSE	0
Airport Division	4PM-MID	CLEAR	Monday	ROAD/STREET/SIDEWALK	0
Airport Division	4PM-MID	CLEAR	Tuesday	APARTMENT	0
Airport Division	4PM-MID	CLEAR	Tuesday	COMMERCIAL PARKING LOT/GA	0
Airport Division	4PM-MID	CLEAR	Tuesday	ROAD/STREET/SIDEWALK	1
Airport Division	4PM-MID	CLEAR	Tuesday	SERVICE/GAS STATION	0
Airport Division	4PM-MID	CLEAR	Wednesday	COMMERCIAL PARKING LOT/GA	0
Airport Division	4PM-MID	CLEAR	Wednesday	ROAD/STREET/SIDEWALK	0
Airport Division	4PM-MID	CLEAR	Thursday	COMMERCIAL PARKING LOT/GA	1
Airport Division	4PM-MID	CLEAR	Thursday	HOTEL/MOTEL PARKING LOT	0
Airport Division	4PM-MID	CLEAR	Thursday	HOTEL/MOTEL/ETC.	0
Airport Division	4PM-MID	CLEAR	Thursday	RESIDENCE/HOUSE	0
Airport Division	4PM-MID	CLEAR	Thursday	ROAD/STREET/SIDEWALK	0
Airport Division	4PM-MID	CLEAR	Friday	COMMERCIAL PARKING LOT/GA	0
Airport Division	4PM-MID	CLEAR	Friday	ROAD/STREET/SIDEWALK	0
Airport Division	4PM-MID	CLEAR	Friday	SERVICE/GAS STATION	0
Airport Division	4PM-MID	CLEAR	Saturday	COMMERCIAL PARKING LOT/GA	0
Airport Division	4PM-MID	CLEAR	Sunday	APARTMENT PARKING LOT	0
Airport Division	4PM-MID	CLEAR	Sunday	COMMERCIAL PARKING LOT/GA	0
Airport Division	4PM-MID	CLEAR	Sunday	HOTEL/MOTEL/ETC.	0
Airport Division	4PM-MID	CLEAR	Sunday	RESIDENCE/HOUSE	0
Airport Division	4PM-MID	CLEAR	Sunday	ROAD/STREET/SIDEWALK	1
Airport Division	4PM-MID	CLEAR	Sunday	SERVICE/GAS STATION	0
Airport Division	4PM-MID	CLOUDY	Monday	APARTMENT	0
Airport Division	4PM-MID	CLOUDY	Monday	COMMERCIAL PARKING LOT/GA	0
Airport Division	4PM-MID	CLOUDY	Monday	ROAD/STREET/SIDEWALK	0

FIGURE 13.6

Sample input data for the analysis of auto thefts. Auto thefts have been designated using a binary indicator variable (“1” = auto theft, “0” = all other incidents). Variables indicating incident time (“Shift” and “Day”), location (“Division” and premises type [“PREMTYP”]), and weather also are listed. *Information Builders WebFOCUS RStat, used with permission.*

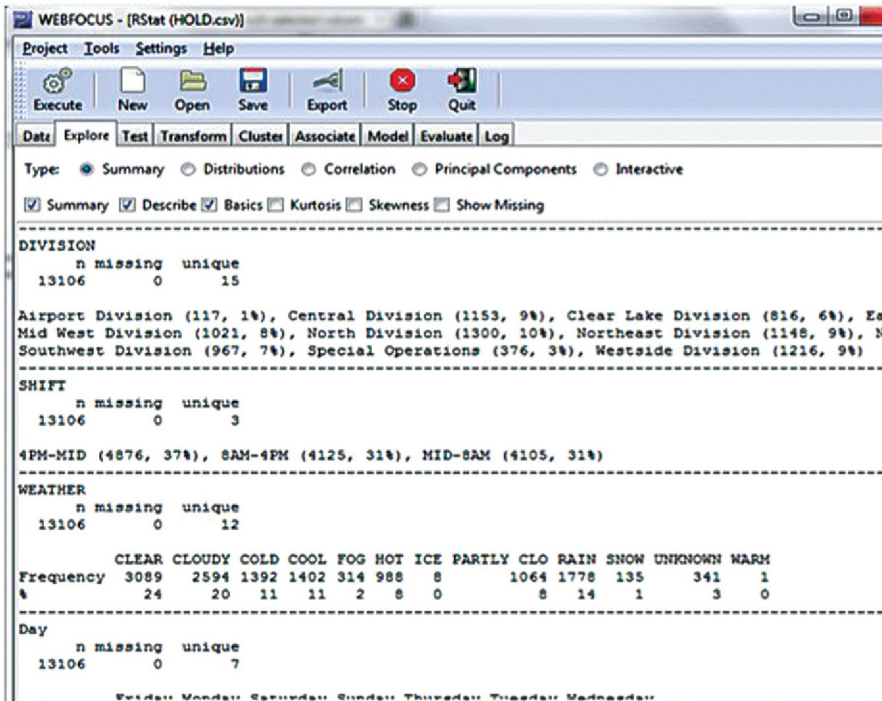
**FIGURE 13.7**

In this example, the “Data” tab enables easy instantiation of data including identification of the data source, and the establishment of input variables and targets, as well as options for partitioning of the data set and weighting of specific variables. *Information Builders WebFOCUS RStat, used with permission.*

of the data is already visible, including location (“Division” and “PREMTYP” [premises type, e.g., residence, commercial, etc.]), time (“Day” and “Shift,” which includes recoding of time into 4-h time blocks or “shifts”), and weather. Additional data sources including citizen calls for service, code enforcement records, local events, the school calendar, and census data also can be collected and used as variables in the model.

Figure 13.7 illustrates instantiation of the data. In this particular example, the analyst is able to use the “Data” tab to identify a data source, specify the data structure, instantiate the variables, and partition the data set. In addition, the analyst is able to set specific variables as inputs to the model and targets, and also exclude¹¹ variables (“Ignore”), and even weight-specific variables. In this example, auto theft has been set as the target, and police division, premises type, shift, day of the week, and weather have all been included as possible input variables.

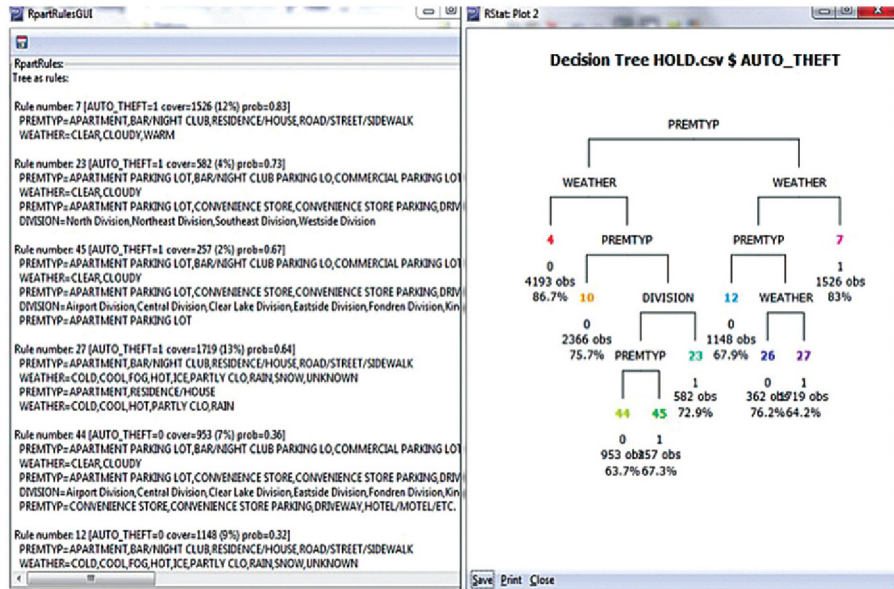
Figure 13.8 illustrates the preliminary descriptive statistics generated for the analysis. In some situations, summary descriptive statistics provide sufficient

**FIGURE 13.8**

Using the “Explore” tab the analyst is able to generate preliminary descriptive and summary statistics to support additional preprocessing and segmentation of the data and analysis. *Information Builders WebFOCUS RStat, used with permission.*

insight to support action. For example, in the New Year’s Eve initiative described later, preliminary analysis of the time of day associated with random gunfire complaints revealed that almost all of the citizen calls for service were received during the time period immediately preceding and then following midnight. In addition to better understanding the data, these results also enable the analyst to explore the data and further refine the approach to modeling, including additional segmentation and/or recoding of the data.

Figure 13.9 illustrates the results from the modeling step. In this example, a rule induction model or “Decision Tree” was used to create a set of decision rules that would reliably characterize auto theft incidents. As can be seen in Figure 13.9, a subset of the decision rules has been listed on the left, while the decision tree is illustrated on the right. While not suitable for direct use in the operational environment, there is enough transparency in this particular modeling approach for the analyst to review, vet, and validate the results. In this particular model, division, weather, and the type of premises were identified as the most important variables for predicting auto thefts. Again, location,

**FIGURE 13.9**

Sample scoring algorithm results that the analyst can review, vet, and validate. The left panel illustrates the decision rules, while the panel on the right depicts the decision tree generated by the rule induction model employed. The output highlights patrol division, nature of the premises (e.g., residence, commercial, parking lot, etc.), and weather as the variables most important in predicting auto thefts. *Information Builders WebFOCUS RStat, used with permission.*

as defined by the police division and type of premises, and weather are operationally relevant and actionable factors that can be used to inform deployment decisions.

Finally, [Figure 13.10](#) illustrates the results of the modeling algorithm as deployed in a mapping environment. Using an easy to navigate interface with pull down menus for the relevant variables, the end user can explore the model in a very easy to use and efficient manner; easily identifying times, locations, and conditions where heavier deployment might be required. This enables targeted troop deployment in direct response to changes in anticipated risk or workload. Command staff, police managers, and other members of the planning team are able to interactively explore, update, and refresh these maps, dynamically investigating different outcomes associated with new or evolving conditions, and crafting deployment strategies while considering a variety of different scenarios. Dissemination of advanced analytics output through a web-enabled geospatial environment can enable a greater degree of deployment fluidity; optimizing personnel resources in a manner not possible using hard copy paper maps or other static resources. Moreover, sharing this



FIGURE 13.10

Results from the analysis are deployed in a web-enabled mapping environment where the end user can adjust the input parameters in order to calculate specific risk based on current and/or anticipated conditions. *Used with permission, Information Builders LEA WebFOCUS, and Chief Rodney Monroe, Charlotte-Mecklenburg Police Department.*

capability with field-based personnel provides actionable analysis directly to the operational end user, thereby concurrently increasing situational awareness and informed decision-making. Finally, the ability to deploy scoring algorithms directly to the end user through a web-enabled mapping environment decreases the amount of work required to create each separate and distinct analytic product, shifting the focus to fluid exploration and discovery in an immersive, transdisciplinary environment.

The advantages of a system like this are numerous. First, operational personnel gain access to analytical support on a 24/7 basis. Crime frequently occurs at times when civilian analytical personnel are not on duty. To wait until they return can create an unacceptable delay, particularly in fast-breaking cases or those requiring analysis in a timely fashion. In addition, these systems can be deployed remotely. For example, the data can be analyzed and the models developed in centralized areas far behind the front lines. Scoring algorithms can then be deployed directly to operational personnel in their environment. This maximizes analytical personnel resources, while diminishing associated

analytical support requirements through the establishment of analytical fusion cells that can be used by multiple operational units. As a result, analysis functions can be centralized and available to remote locations, as well as centrally located command units. This analytical utilization model ensures that even remotely located operational units have access to the full analytical capacity of the organization, without unnecessary duplication of analytical personnel, resources, and support.

Furthermore, utilization of centralized analytical capacity and remote deployment of scoring algorithms and models also facilitates the integration of different types of data and information. As described in the Northern Virginia military shooting series shown next, crime frequently transcends traditional jurisdictional boundaries and functional domains.¹² Moreover, seemingly disparate patterns of crime often are interrelated. Prostitutes use drugs, drug users rob convenience stores, and gang members commit violent crimes. While a certain degree of operational specificity often is required, a common analytical resource can integrate data and information on related crime patterns and trends. The resulting models will be significantly enriched through the use of these various related informational resources, which then can be deployed directly to the various operational end users and other stakeholders through the use of web-based tools.

Finally, as exemplified by the earlier auto theft example, web deployment of advanced analytics output, particularly through a geospatial environment, also permits the use of more complex scoring algorithms because the model output does not need to be interpreted directly to have value. This provides the opportunity to use relatively opaque or “black box” models with a greater degree of accuracy. The end user need only enter a limited amount of information to receive the model-derived insight. This arrangement affords a high degree of accuracy through the use of relatively sophisticated modeling techniques, while requiring limited end user training and data entry requirements.

13.4.3 The Northern Virginia Military Shooting Series¹³

The war on terrorism has generated a variety of new challenges for law enforcement agencies attempting to protect our homeland, while addressing routine crime issues that generally defined their purview prior to 9/11. Perhaps one of the biggest challenges is stretching already diminished personnel and budget resources to accommodate the additional responsibilities associated with the war on terrorism. The concept of fourth-generation warfare and implications for local law enforcement is discussed in Chapter 12; however, the direct impact on resource allocation and deployment can be understood regardless of the cause.

Prior to 9/11, most agencies were in the unenviable position of doing more with less, particularly with diminishing economic resources. After that date,

local agencies increasingly became responsible for collecting and compiling additional data and information, increased deployment related to sensitive or high-profile locations, and periodic escalation in readiness associated with heightened threat levels. Agencies already coping with limited troop strength lost additional personnel to military activation, federal hiring, and reallocated resources to homeland security tasks and task forces. In addition, these new responsibilities frequently transcend jurisdictional boundaries and functional domains requiring cooperation and collaboration between local, state, and federal law enforcement organizations, as well as various Department of Homeland Security (DHS) component agencies.

Doing more with less requires smart, data-based, results-driven deployment strategies. Personnel resources, in particular, need to be allocated judiciously to ensure complete coverage and maintain the ability to respond adequately. This is true not only for routine patterns of offending and enforcement but for rapidly emerging homeland-security-related functions as well. Predictive analytics and information-based approaches to deployment facilitate the provision of more science and less fiction in personnel deployment. Similarly, enhancements to and integration of predictive analytics capabilities and mapping tools offer additional opportunities for the development of operationally relevant and actionable analytic output that can move from the analysis unit directly into the operational environment.

In October of 2010, the law enforcement community in Virginia was faced with a series of shootings into facilities of interest to the United States military including the Pentagon, a United States Marine Corps (USMC) recruiting facility in Chantilly, and two separate incidents involving the National Museum of the Marine Corps. In addition to the significance of the facilities targeted, this series was eerily reminiscent of the DC Sniper series, which also occurred in October 8 years earlier in the same general location. Adding to the concern regarding this particular series, autumn in northern Virginia is associated with several high-profile, well-attended events of interest to the military community including Veteran's Day, the Marine Corps Birthday, and the Marine Corps Marathon. Given this context, the question on everyone's mind was, was this a prelude to something more sinister? Would the shooter be content shooting into unoccupied buildings, or would they move from shooting at things to targeting people and were these high-profile events at risk?

Figure 13.11 illustrates the location of the first four incidents in the series, which covered an area of approximately 750 square miles¹⁴ and spanned multiple police jurisdictions to include local, county, state, and federal. Moreover, given the nature of the targets, various DHS component agencies and United States Department of Defense (DoD), including the Pentagon Police and force protection elements also became involved. Despite the broad agency response and involvement, though, the area of interest was still too broad to cover

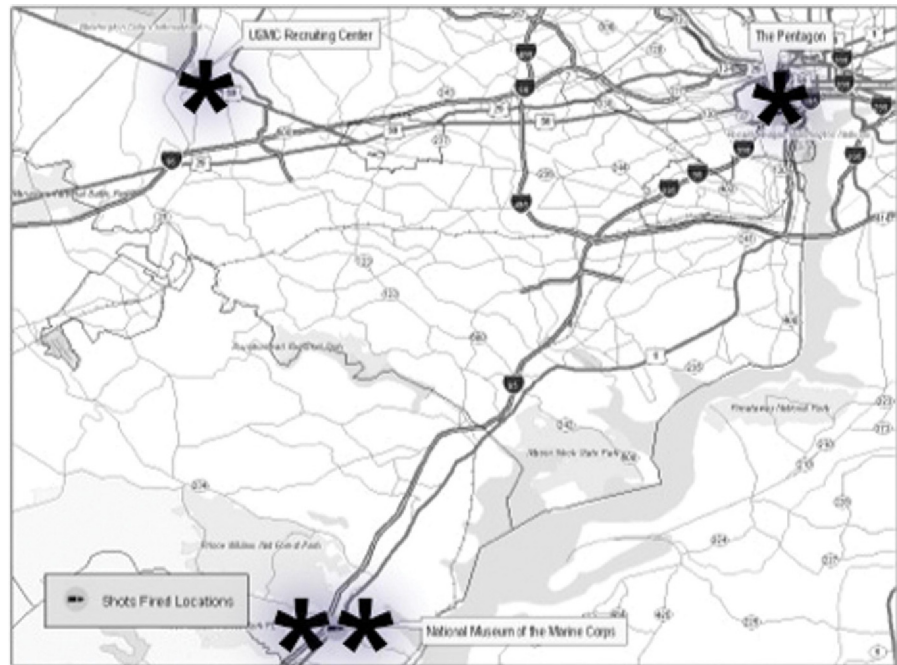


FIGURE 13.11

Location of the first four shootings in the Northern Virginia military shooting series, including the US Marine Corps Recruiting Station in Chantilly, Virginia; the Pentagon; and two separate incidents at the National Museum of the Marine Corps. *Reprinted from The Police Chief 2013; 80(2): 48–52. Copyright held by the International Association of Chiefs of Police, Inc. Further reproduction without express permission from IACP is strictly prohibited.*

effectively and included thousands of possible targets given the heavy DoD presence in the area. Therefore, in an effort to more efficiently allocate and optimize resources geospatial predictive analysis was employed by a multiagency team at the Virginia Fusion Center.

Geospatial predictive analysis is a rule induction model that is based on the fact that behavior generally is not homogenous or uniformly distributed.¹⁵ Rather, people tend to develop place preferences that can be statistically characterized, and the generated models used to reliably anticipate future behavior. Criminal place preferences tend to reflect two related needs. First is a requirement for access to a victim and/or potential target. Second is selection of an environment where they believe that they can successfully perpetrate their desired act by leveraging enabling factors and/or avoiding possible deterrents or other factors that would thwart or otherwise mitigate the consequences of their act. Again, these factors tend to be both offense and offender specific. One offender's deterrent could represent the preferred target for another. For example,

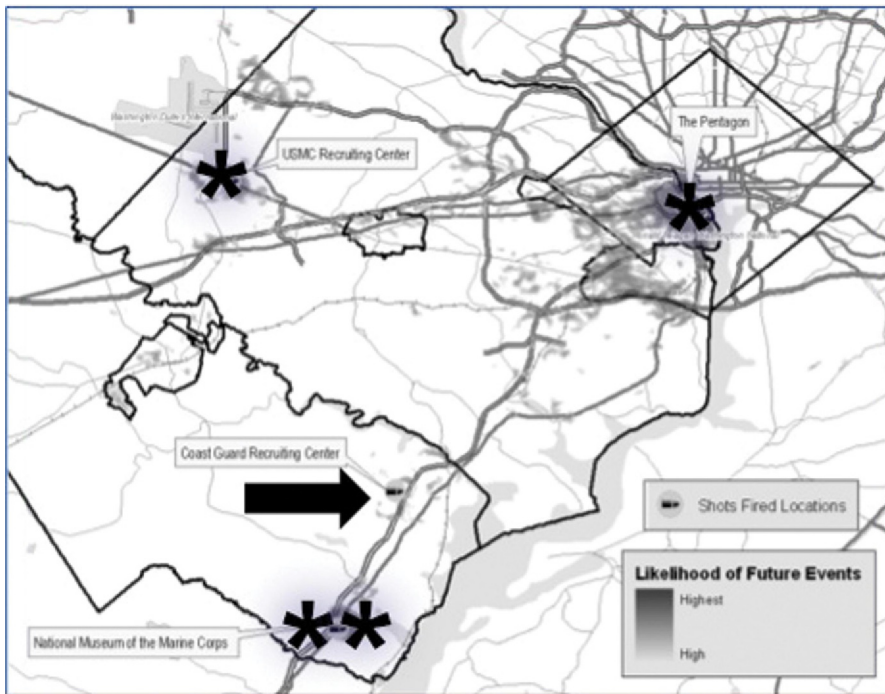


FIGURE 13.12

Map depicting the results of the geospatial predictive analysis that was conducted on the first four incidents in the series. The arrow highlights a location in Woodbridge, Virginia, which was included in the high likelihood area, and was the location of a shooting incident at the US Coast Guard Recruiting Center that occurred 3 days after the analysis was disseminated and briefed. *Reprinted from The Police Chief; 2013; 80(2): 48–52. Copyright held by the International Association of Chiefs of Police, Inc. Further reproduction without express permission from IACP is strictly prohibited.*

while many criminals actively avoid law enforcement and other security forces, extremist groups in the Middle East and Africa may preferentially target these resources.

The results of the analysis are illustrated in [Figure 13.12](#). Shaded areas in the map indicate an increased likelihood for a future incident. Again, the first four incidents occurred over a broad area covering approximately 750 square miles. Given the large area of interest and resource limitations, the model thresholds were adjusted to reveal the top 2% most likely locations, enabling the team to reduce the overall search space by 98%. Of special relevance to this particular series, the statistical model created also could be projected to novel, noncontiguous locations rather than confining the analysis to those locations that are in close proximity to or contiguous with previous incidents; a critical feature given the geographically dispersed nature of the previous incidents.

Three days after the original analysis was briefed, a new shooting was reported at the US Coast Guard Recruiting Center in Woodbridge, Virginia. Although there were no direct ties to the USMC, this location had been identified previously by the model as being at high likelihood for a future incident. Moreover, while correlational, these models also can surface factors associated with an increased likelihood for a future incident. In this particular series, variables identified were associated with easy egress from the targets, in addition to cemeteries and motels. While the easy egress is relatively common, identification of cemeteries and motels were particularly concerning given the planned activities for Veteran's Day and the Marine Corps Birthday celebrations, respectively.

Operationally relevant and actionable, the geospatial predictive analysis was loaded onto laptops and shared with personnel in the participating agencies on a "need to know" basis to support information-based deployment decisions, including allocation of patrol resources, as well as surveillance assets.¹⁶ Again, the primary goal for risk-based deployment is to create an unattractive environment and suppress future incidents. Based on the analysis, deployment was adjusted. The operational end users were able to use the output to support information-based decisions regarding allocation and optimization of their resources in an effort to prevent future incidents.

After these results were deployed, the shootings stopped. As with many cases like this, it is impossible to know exactly why the shooter stopped. Did the shooter get sick, die, move out of the area, or was he or she arrested for another crime? Even if the heavy deployment in locations the shooter preferred that was guided by the analysis did cause him or her to stop, he or she may not have been consciously aware of it; criminal place preferences can be extremely subtle and nuanced, and the shooter may only have been aware of a general level of comfort, or conversely a new uneasiness with a particular environment or location. Ultimately, though, whether this effort caused the series to stop or some other factor was driving the shooter's activity, the important point is that the shootings stopped and no one got hurt. The crime analyst can live with the uncertainty regarding the actual cause.

Approximately 6 months after the last shooting incident, on June 17, 2011 a USMC Reservist, Yonathan Melaku was apprehended in the Arlington National Cemetery. Search of the backpack that he was carrying at the time of his arrest and a subsequent search of his residence revealed supplies and instructions for making improvised explosive devices (IEDs), extremist materials including references to Osama bin Laden and The Path to Jihad, videotaped and other physical evidence linking him to the previous shootings, and spray paint, which he was going to use to deface the grave markers of service personnel from Iraq and Afghanistan.¹⁷ Investigators familiar with the case reinforced

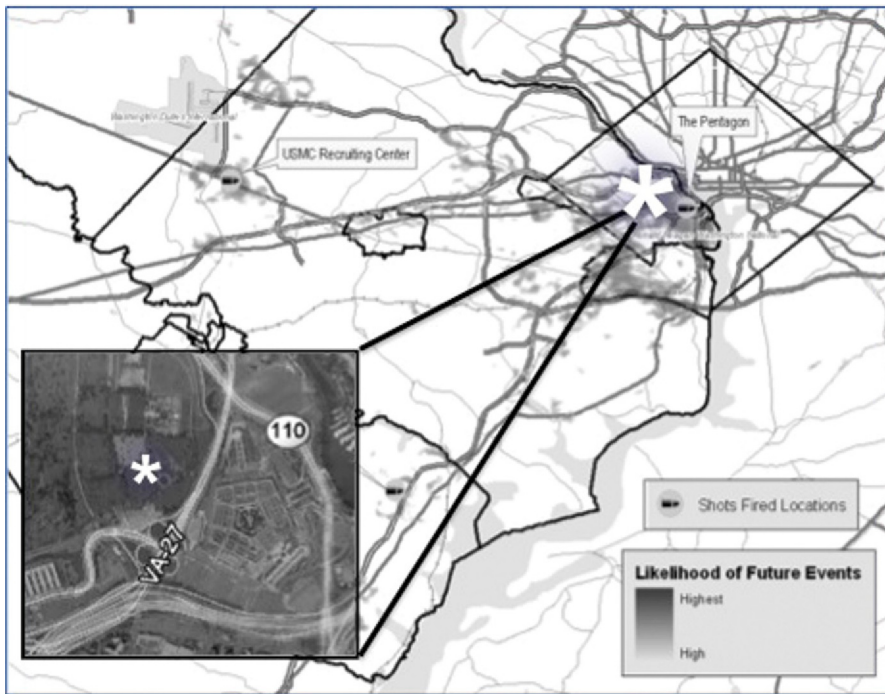


FIGURE 13.13

Map illustrating the location where Yonathan Melaku was apprehended in the Arlington National Cemetery on July 17, 2011. The inset box illustrates the location in the Cemetery in greater detail, including its relationship to the Pentagon. *Reprinted from The Police Chief 2013; 80(2): 48–52. Copyright held by the International Association of Chiefs of Police, Inc. Further reproduction without express permission from IACP is strictly prohibited.*

earlier concerns regarding the potential for escalation in this series, noting that it was, “unclear what might have been coming had he not been caught.”¹⁸ As can be seen in [Figure 13.13](#), he was apprehended in an area determined previously to be at high likelihood for a future incident, which provides additional validation for the model.

In addition to the direct analytic value of geospatial predictive analysis, this particular case also highlights the benefit associated with the fusion center model. Given the cross-jurisdictional nature of the series, local, county, state and federal agencies, as well as DHS and DoD elements were actively supporting the investigation. Leveraging the fusion center model, the Virginia Fusion Center was able to provide the vertical and horizontal integration of data and other resources necessary to effective respond to the series. Moreover, in a unique public–private partnership, the Virginia Fusion Center was

able to secure and functionally optimize advanced analytics technology and other resources, to include trained data scientists in support of coordinated action and mutual analytic support, creating an analytic force multiplier by establishing their fusion center as a managed service cell supporting the various agencies responding to the series and validating the fusion center concept.

13.4.4 Evaluating the Concept: The New Year's Eve Initiative

The risk-based deployment models described earlier evaluated well using training and test samples, as well as future incidents; however, it was difficult to generate a measurable outcome, particularly given the nature of this relatively low-frequency event. The question inevitably was asked, "How has this approach improved public safety?" Therefore, a different type of risk-based deployment strategy was developed for the 2004 New Year's Eve holiday; one that incorporated embedded outcome measures.¹⁹

The New Year's Eve holiday frequently is associated with increased reports of random gunfire. Therefore, in an effort to increase public safety over New Year's Eve, a risk-based deployment strategy was developed as part of the Project Safe Neighborhoods initiative with the United States Attorney's Office in the Eastern District of Virginia.

To create the deployment model, random gunfire complaints from the previous year were examined. By drilling down into the data, a unique array of activity across time and space emerged that resulted in the development of a specific, targeted, risk-based deployment strategy. Examination of the activity patterns revealed that almost all of the increase in random gunfire complaints occurred between 2200 h on New Year's Eve and 0200 h the following day. While this made intuitive sense, this was the first time that the temporal patterns of activity had been examined in this manner. As a result of this finding, the risk-based deployment initiative was confined to an 8-h period bracketing the period anticipated to be associated with the most work.

Within this time period, activity across specific policing beats was analyzed further. The beats were rank-ordered by relative activity during the previous New Year's Eve holiday, and the top locations were selected for increased deployment. Recent trends and patterns also were analyzed in an effort to identify any areas that might be ramping up or experiencing significant increases in activity that might require more attention during the holiday. Through this approach, additional locations were identified and added to the list.

A final list was developed that included areas previously associated with increased random gunfire complaints during the New Year's Eve holiday as well as additional areas showing recent increases in random gunfire. An operational plan was developed and implemented using a "beat-stacking" approach, which

included heavy patrol and the deployment of additional tactical units in the areas determined to be at increased risk for random gunfire.

The results of the initiative supported the use of this type of risk-based deployment strategy for targeted deployment. Random gunfire complaints were decreased by 47% on New Year's Eve and by 26% during the entire 2-day holiday. Moreover, the number of weapons recovered during the initiative was increased from 13 the previous year to 45 during the initiative – an increase of 246%. To ensure that the random gunfire reductions were specific to the initiative, the period immediately prior to New Year's Eve was analyzed. A comparison between the random gunfire complaints revealed no differences between the 2 years.

Perhaps the most encouraging outcome measure involved the personnel resources used for the initiative. As a direct result of confining the initiative to an 8-h period and the use of a risk-based deployment strategy, the number of required personnel was decreased significantly. By specifically targeting personnel resources, approximately 50 sworn employees were released from duty over the holiday, which resulted in an economic savings of approximately \$15,000 in personnel costs and associated holiday pay during an 8-h period. These results support the hypothesis that information-based allocation of resources not only increases public safety, but also may enable police managers and command staff to optimize personnel resources, particularly in a resource-constrained environment. Further information related to the outcome evaluation associated with this very successful initiative has been addressed in greater detail in Chapters 6 and 8.

Bibliography

- 1 It is interesting to note, however, that when things are going well, citizen interest in a strong (i.e., visible) police presence is not as great. In fact, a highly visible police presence might even be perceived as intrusive when crime is low.
- 2 McCue C, McNulty PJ. Gazing into the crystal ball: data mining and risk-based deployment. *Violent Crime Newsletter*, September 1–2; 2003.
- 3 Beck C, McCue C. Predictive Policing: what can we learn from Wal-Mart and Amazon about fighting crime in a recession? *Police Chief*, November, 2009.
- 4 Helberg C. *Data mining with confidence*, 2nd ed. Chicago, IL: SPSS, Inc.; 2002.
- 5 McLaughlin CR, Robinson DW, Faggiani D. Declining homicide rates in the 1990s: Not everywhere! *ACJS* 1998.
- 6 McLaughlin CR, Yelon JA, Ivatury R, Sugeran HJ. Youth violence: a tripartite examination of putative causes, consequences and correlates. *Trauma Violence Abuse* 2000; 1: 115–127.
- 7 McCue C, McNulty PJ. 2003.
- 8 McCue C, Parker A. 2004.
- 9 McCue C, Parker A. Web-based data mining and predictive analytics: 24/7 crime analysis. *Law Enforc Technol* 2004; 31: 92–99.
- 10 Information Builders WebFOCUS RStat, used with permission.

- 11 Variables likely to be excluded during the initial instantiation of the data may include offense or incident numbers, or any other variable identified *a priori* as being unlikely to bring predictive value to the model and/or those that are not operationally actionable (e.g., suspect characteristics).
- 12 Faggiani D, McLaughlin CR. A discussion on the use of NIBRS data for tactical crime analysis. *JOQC* 1999; 15: 181–191.
- 13 McCue C, Miller L, Lambert S. The Northern Virginia military shooting series: Operational validation of geospatial predictive analytics. *Police Chief*, February. http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=2871&issue_id=22013; 2013.
- 14 Klein A. Police enlist war tech in crime fight. *The Washington Post*. http://articles.washingtonpost.com/2013-02-18/local/37158510_1_plate-readers-shootings-data-police; 2013 [accessed 18.02.2013].
- 15 Dalton JR, Porter MD. Geospatial preference models in signature analyst (white paper). McLean, VA: SPADAC, Inc.; 2009 .
- 16 The interested reader is encouraged to read the primary source for this particular analysis, which includes full color reproductions of the shape files and other analytic products created for this investigation: McCue, C., Miller, L. and Lambert, S. The Northern Virginia military shooting series: operational validation of geospatial predictive analytics. *Police Chief*, February. http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=2871&issue_id=22013; 2013.
- 17 White J. “Yonathan Melaku Admits Shooting at Pentagon, Military Buildings,” *Washington Post*. http://www.washingtonpost.com/blogs/crime-scene/post/plea-agreement-hearing-for-alleged-pentagon-shooter/2012/01/25/gIQAYduHRQ_blog.html; 2012 [accessed 26.01.2012].
- 18 White J. 2012.
- 19 McCue C, Parker A, McNulty PJ, McCoy D. Doing more with less: Data mining in police deployment decisions. *Violent Crime Newsletter*, U.S. Department of Justice, Spring 2004; 1: 4–5.
- 20 McCue C, and McNulty, P.J. Gazing into the crystal ball: data mining and risk-based deployment. *Violent Crime Newsletter*, September, 1–2; 2003.

Surveillance Detection

“We can learn even from our enemies.”

Ovid

In the days following the 9/11 attacks, information, speculation, rumors, “be on the lookout” or BOLOs, and “suspicious situation” reports flooded into every public safety agency, which generally compiled these reports in notebooks and clipboards. Many of these reports were investigated; however, the vast flow of information made it difficult to conduct any sort of analysis. Similarly, information went up, down, over, around, and through almost every public safety agency in this country, whether large, small, local, state, or federal; however, there were limited opportunities to ensure that this information-sharing process was organized or even complete. As things have slowed down somewhat from that initial frenzy, two information-based challenges have emerged: information stovepipes and the failure to identify meaningful relationships and patterns. One of the goals of this text is to encourage analytical and operational personnel to work together more closely, even within the same organization. Addressing information stovepipes in law enforcement and intelligence analysis is well beyond the scope of this book. The emerging emphasis on identifying meaningful patterns and relationships however, is well within the purview of data mining and predictive analysis. In my opinion, “connecting the dots” merely tells us what happened. To create safer neighborhoods for our children and ensure our homeland security requires us to look forward in an effort to anticipate and ultimately prevent bad things from happening. Whether it is a street corner drug-related shooting or the next cataclysmic terrorist attack, figuring out what happened in retrospect is a costly approach to public safety.

Preoperational or hostile surveillance generally is intended to be covert or to appear relatively innocuous to uninformed observers. Frequently, it is only when an attack occurs or a larger pattern of suspicious behavior or presumptive

preoperational surveillance activity has been identified, compiled, and characterized that the true nature of the activity is revealed. For example, reports in the media suggested increased interest in facilities in northwest Washington State.¹ These reports outlined several incidents of unusual or suspicious behavior, including photo and video surveillance of sensitive locations and facilities as well as attempts to obtain regional survey materials. This repeated and ongoing occurrence of suspicious and unusual behavior in and around Anacortes, Washington, the Deception Pass Bridge, and Whidbey Island has particular relevance given the critical infrastructure and military assets located in that area. These assets include the Whidbey Island Naval Station and the Washington state ferry system, which provides critical access to many of the islands in the Puget Sound, as well as the neighboring oil refineries. This unusual behavior takes on added significance given the fact that the Millennium bomber was apprehended at the Anacortes ferry terminal. When reviewed in isolation, these reports might not be cause for concern. Analyzed as a larger pattern, however, these incidents suggest a coordinated effort to acquire information about a particular geographic region.

While there are no crystal balls in law enforcement and intelligence analysis, data mining and predictive analytics can help characterize criminal behavior so that we can make accurate and reliable predictions regarding future behavior or actions, which is absolutely essential to effective crime prevention. One area where this has tremendous potential is surveillance detection. In many ways, surveillance is a systematic review of a person, route, facility, or some other item of interest. Data mining and predictive analytics thrive on homogeneous and coordinated behavior, such as that which is embodied in the aforementioned “systematic review.” Therefore, advanced analytics can represent a powerful tool in the identification and characterization of putative surveillance activities in the operational public safety and security environment.

14.1 SURVEILLANCE DETECTION AND OTHER SUSPICIOUS SITUATIONS

Preoperational surveillance is relatively common in many patterns of offending. A criminal planning a bank robbery might drive by several banks looking for those with physical characteristics that appeal to him. Easy access and egress might be imperative. Proximity to major highways or multiple escape routes might be important considerations. Once a specific location has been selected, the suspect might spend time watching the bank to determine routine operations. When is it busy? When is it relatively slow? Are there security personnel? If so, do they take breaks? In short, the potential bank robber is interested in information that will maximize his gain while minimizing the risk of apprehension.

On the other hand, the suspect might have been noticed several times during this process. The bank tellers may have noticed the same vehicle sitting outside the bank on multiple occasions. The suspect might even have come into the bank and then left without transacting any bank business. In some cases the suspect might engage in conversation with bank employees or make inquiries regarding the security procedures. Unfortunately, this information often comes to the attention of law enforcement personnel only after something happens or if awareness has been heightened due to a high-profile event or series of events. The important point, though, is that preoperational surveillance is associated with many patterns of offending, and that in many cases this behavior is noted. In some cases, preoperational surveillance is reported, but it is rare for it to be compiled and analyzed on a routine basis. If a particular agency understands the value of this information and proactive analysis, they might be able to anticipate the type of location to be targeted next, respond proactively, and prevent the crime from occurring. Unfortunately, law enforcement agencies generally do not receive information that is proactive and specific unless they are in the midst of a particular series. Regardless, “suspicious situation” reports of this nature should be analyzed whenever possible, as they frequently provide a window into the criminal planning process.

14.2 GENERAL CONCEPTS

Suspicious actions or behavior suggestive of preoperational planning or surveillance are both infrequent and subtle by their very nature. Trying to identify unusual or suspicious behavior indicative of something far more sinister often resembles looking for the proverbial needle in the haystack. Frequently, indications of these types of activities almost always occur only when the potential suspect makes a mistake, which further highlights their rarity. What would be helpful in revealing these activities, the “needle in the haystack,” would be some sort of magnet. In many ways, the technique of anomaly detection can serve that function.

Building on the concept of risk-based deployment described in Chapter 13, similar analytic strategies can be used to optimize surveillance detection resources.² Like patrol deployment, the use of advanced analytics takes advantage of the nonrandom or systematic nature of preoperational surveillance. Characterizing and predicting when and where this activity is likely to occur can guide proactive deployment of surveillance detection resources in a way that increases the likelihood that these personnel resources will be in place when and where the behavior of interest occurs. Moreover, this strategy also decreases the likelihood that resources will be deployed when and where they are not needed – a feature that supports the thoughtful allocation of resources.

14.2.1 Data

It is not unusual to interview witnesses after a major event and have them recount unreported suspicious behavior that indicated something bad was about to happen. In fact, Gavin de Becker in *The Gift of Fear* recounts cases of workplace violence in which the event was so anticipated that as soon as the shooting started people correctly identified the suspect before the actual nature of the event was even known.³ Similarly, during the 9/11 inquiries, reports surfaced outlining unusual or troubling preincident behavior that was not taken seriously, investigated, or linked. The challenge implicit in this public safety predicament, therefore, is threefold. First, the information needs to be reported and compiled. While this is not a challenge specific to analytical personnel, they can be greatly impacted by incomplete or inaccurate reporting. It is not possible to analyze what does not exist, so analysts have a vested interest in ensuring that suspicious situations and other indicators of preoperational surveillance are reported and compiled. Second, the information needs to be effectively analyzed. Compiling and storing suspicious situation reports in a three-ring binder is a waste of a potentially valuable resource. The information should be entered into a database, analyzed, periodically reviewed, and analyzed again. Third, the results of the analysis need to be used operationally. This extends beyond preliminary investigation of the suspicious situation reports. Preoperational surveillance is designed to look innocuous. Frequently, it is only when the larger pattern of suspicious activity or surveillance has been revealed that it becomes actionable. If suspicious situation reports reflect mistakes on the part of the potential bad guy, then those can be used to reveal the larger pattern of surveillance. Using a model of suspicious situation reports to guide additional surveillance detection efforts can maximize often limited personnel resources. By determining when and where the most activity is occurring, operational personnel can proactively deploy and increase the chances that they might identify additional, less obvious behavior. Moreover, it also increases the likelihood that specific individuals or vehicles will be identified, which further enhances the investigative effort.

Again, while increasing natural surveillance is not really a problem of data mining or predictive analytics, gathering information that is as complete as possible is essential to creating accurate and reliable models. In many ways, enhancing information collection is an essential first step in creating a program of surveillance detection and threat assessment. Most, if not all, suspicious situation reports should be analyzed for any consistent behavior, unusual patterns, or indications of possible intensification or escalation, even if they have been investigated already. The analyst frequently can provide preliminary information illuminating what the suspect might be considering. For example, ongoing repeated observations of the same facility might indicate an interest in that location as a particular target. Preliminary analysis of these reports might

suggest clustering in the weekends. By separating the activity by different time blocks, the analyst might notice increased surveillance activity around closing time. By drilling down into the data, we might identify two types of suspicious activity. Perhaps there are two groups interested in this location. One is considering a robbery, while the other might be interested in an after-hours burglary. Characterizing and modeling this behavior can guide additional coordinated surveillance detection activities, including identification of hostile surveillance positions or the “red zone,”⁴ or it might establish a likely time frame and possible type of incident, which could be addressed by heavy deployment or some other proactive, targeted operation or response planning.

14.2.1.1 Suspicious Activity Reports (SARs)

Consistency in reporting can be a powerful enabler to the analysis process. One response to this challenge has been the Nationwide SAR Initiative (NSI), which is a collaborative effort involving DHS, the FBI, and local, state, tribal, and territorial law enforcement agencies that outlines data collection, data organization and structure, processing and analysis, and dissemination and sharing.⁵ In addition to providing training and facilitation of technology solutions, the NSI also engages in stakeholder outreach and privacy protection.

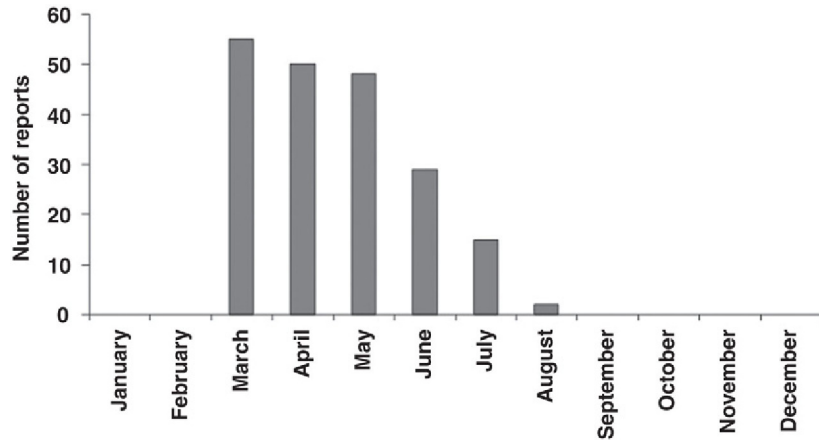
14.3 HOW TO

Similar to the analysis of other patterns of behavior, initial exploration and characterization of the data generally represents the first task in the analysis of potential hostile surveillance.

14.3.1 Time

Figure 14.1 represents a notional example of the use of descriptive statistics to explore suspicious situation reports in support of surfacing possible hostile surveillance. As can be seen in the figure, a marked increase in the number of suspicious situation reports was noted in March. The first question that should be addressed is “What happened in March?” If employee personal safety training had been offered in March, or if there had been a major incident in late February that had heightened awareness, the increased number of reports received during the month of March would be viewed somewhat cautiously. However, if nothing obvious had changed, then it would be important to quickly assess the nature of these reports in an effort to determine whether there is cause for concern.

Similarly, the trend in the number of reports received appears to have decreased over time. Again, it is important to put this information into a context to determine whether this decline is real or something that needs to be addressed. For example, additional information indicating that each report had

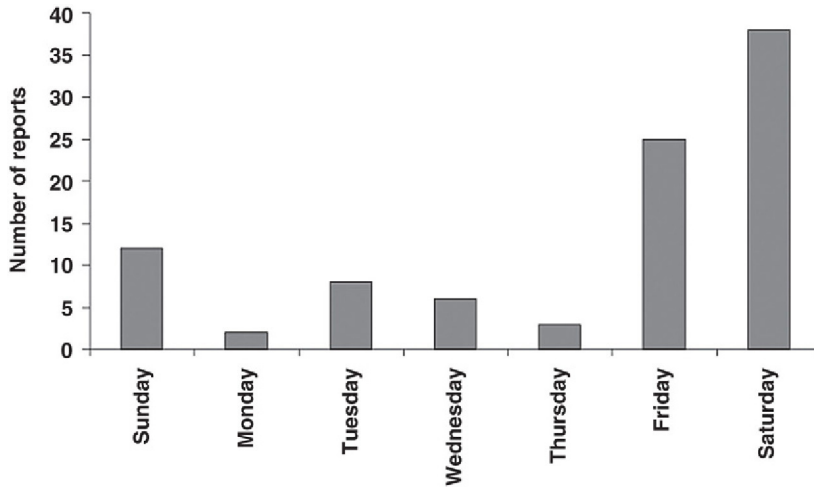
**FIGURE 14.1**

Graph of suspicious situation reports over time.

generated a rapid and aggressive security response would suggest that perhaps this location has become a difficult target. If this is the case, a more complete review of the specific reports would help to further define the nature of the potential threat and might even form the basis for a security-related “after action” report. On the other hand, a decline in reporting with no obvious change in security might indicate apathy or frustration on the part of the staff. Again, it is important to thoroughly review the reports and possibly to conduct a survey to ensure that reporting is encouraged within the organization.

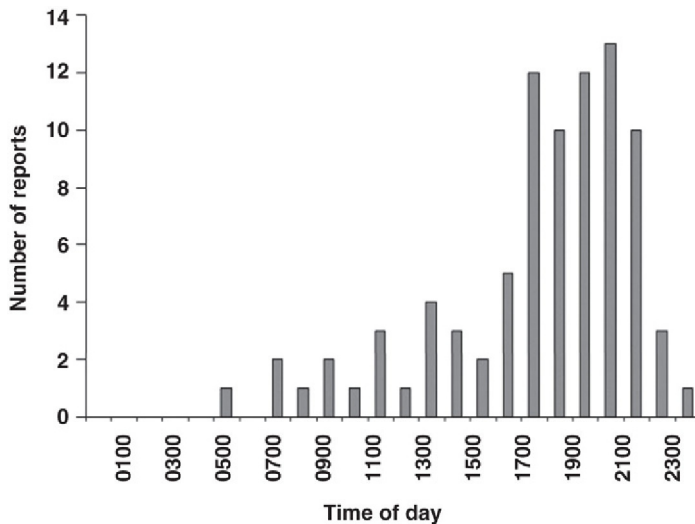
In a more complex example, we evaluate a series of reports of suspicious activity around a shopping mall. Several reports were received, but it was not clear whether this should be cause for concern. By creating a simple spreadsheet and graphing the data, it becomes apparent that most of the activity is occurring on Fridays and Saturdays (Figure 14.2). By further drilling down by time of day, it also becomes obvious that most of the activity is occurring when the mall is open and that the activity increases during the evening (Figure 14.3). Several questions come to mind at this point. For example, how does the mall activity differ on the weekend as compared to weekdays, and what is different about the evening?

More importantly, though, does this really mean anything? Is there anything of significant concern at this point? One could certainly argue that the mall population increases on the weekend. Is the increased reporting related to a transient increase in the mall population observed on Fridays and Saturdays? We also might expect more young people in the mall on the weekends because they are out of school. Is the increased reporting an artifact of an increased number of kids who have been sensitized to “stranger danger” and all of the

**FIGURE 14.2**

Graph of suspicious situation reports at a shopping mall, by day of week.

other victimization-prevention programming that is available today? In other words, can this apparent increase in activity be attributed to reporting bias? While this might explain the increase noted on the weekends, it does not necessarily explain the increased number of reports associated with the evening hours.

**FIGURE 14.3**

Graph of same suspicious situation reports depicted in [Figure 14.2](#), by time of day.

On the other hand, the pattern of results could have nothing to do with anything special or unique about the mall. Rather, it could represent a convenient time or place or something unique about a potential suspect. It is not unusual for the timing and even location of crime to be related to the convenience or routine schedule of the criminals. In fact, this is referred to as their “comfort zone” and is not at all unusual with certain patterns of offending. We frequently focus on the location of the reports or some unique feature of the victim or location targeted. For example, as outlined in Chapter 6, a series of bank robberies was analyzed several years ago using regression analysis in an effort to determine the length of time between incidents. The results of the analysis revealed that the time between robberies was related to the amount taken in the previous robbery. The criminal in this case needed to maintain a certain cash flow to meet his expenses, so if he was able to obtain a large amount from one bank, the time to his next robbery was decreased. If, on the other hand, his take was relatively small, he would need to go back out and rob another bank sooner. The relationship between amount taken and the crime interval is relatively common among drug addicts. Due to the compulsive nature of drug use and/or the need to stave off withdrawal, many addicts commit economic crimes to support their drug habits. Consequently, the frequency with which they commit crimes might be related to the cost of maintaining their habit, or cash flow, and the monetary yield from each crime. While this might be the best explanation for the activity noted at the mall, it is generally a good idea to play the devil’s advocate and consider an alternate hypothesis for a particular set of data or information, because it is not at all unusual for the particular time, location, or the victim selection to be related to some unique but unknown feature of the suspect.

Even if nothing more is done analytically at this point, by compiling the information and conducting this quick analysis, operational deployment can be altered to respond specifically to the reported behavior. This results in three possible benefits. First, if there is something unusual going on, by specifically deploying operational personnel when and where it has been occurring, the likelihood is increased that they also will observe this behavior or be able to respond more quickly should it occur again. Second, increased deployment in the area might deter any additional suspicious or unusual behavior. Finally, targeted deployment in response to these reports visibly projects an increased police presence, while concomitantly enhancing the perception of increased public safety in that area.

14.3.2 Space

Preoperational surveillance requires a certain amount of time for observation of the potential target, time during which the operator is vulnerable to detection. The ability to not only identify but also characterize and model this

behavior has tremendous tactical and strategic value. Most frequently, this information arrives in the form of suspicious situation reports, which provide a general descriptive characterization of suspicious activity (e.g., photographing or videotaping a facility). Although suspicious situation reports rarely include specific location information (e.g., exact location) in a standardized format, this attribute can be considered within the context of spatial sets, which are particularly well suited for data mining and operational planning. For example, relatively general information characterizing what the subject of a suspicious situation report was observing can provide invaluable guidance regarding that individual's likely intentions and the possible vulnerabilities associated with a particular location. Similarly, information pertaining to the individual's general location or observation point can guide the placement of surveillance detection resources. Neither of these analyses requires specific information. Rather, spatial sets match the available data resources and are sufficient for not only analysis but operational action as well. With this in mind, we can consider three possible locations associated with suspicious activity: the location where the hostile surveillance is occurring – the red zone; what the suspicious person is observing – the potential target; and the observer's location, which frequently is the location indicated in citizen complaint or other crime analysis databases. All three locations are important to effectively analyzing and interpreting the behavior, particularly as relates to risk and threat assessment of a potential target, and structuring effective surveillance detection.

It is important to note, however, that certain types of space, including those associated with transportation (e.g., trains, airplanes, trucks), are not easily linked to absolute spatial indicators, such as longitude and latitude, center lines, patrol regions, or assessment boundaries. Rather, relative indicators (e.g., first class section, behind the tree line, 300 miles from the departure point) may have far greater value when characterizing movement and identifying a potential risk or threat. It is also worth considering the fact that the space in and around a facility is not homogenous. Unusual or suspicious behavior around a facility frequently is informally weighted. Different types of activity are given greater or lesser value, and the value associated with a potential behavior is an interaction between the location, existing rules, norms, and boundaries, and the specific attributes of the behavior itself. Some movement is of more consequence than others. For example, forcing through a checkpoint would be an obvious transgression. Any incursion into this area would be cause for concern. Other behaviors are not as obvious. For example, photographing or videotaping a facility might not raise as much suspicion, but repeated activity or focus on a particular facility or specific aspect of a facility might be cause for concern.

Perhaps more subtle, buildings and facilities also are associated with normal flow patterns and boundaries. Delineation of these spatial boundaries might

be explicit through signage or implicit through group norms and behaviors. Similar to “personal space,” some buildings or facilities have an invisible buffer. Also like “personal space” violations, transgressions of these spatial rules and norms can attract attention. Therefore, by creating spatial sets associated with the facility of interest, it is possible to look for possible focusing or spatial specificity. Similarly, by weighting different locations within the created spatial sets, as well as the behaviors, it is possible to identify and document potential escalation.

Again, it can add great value to an analysis to identify not only when, but where. To identify specific locations or areas that are associated with increased interest can greatly assist in the spatial refinement of surveillance detection efforts by further defining the true or active zone of unusual or suspicious activity. The exact physical location of a possible incident of surveillance generally may not be as important as where the person was and what he or she was looking at. Extending this, five different individuals could occupy five slightly different locations, but if they were all observing the same person, building, or specific aspect of a facility, then it is important. In fact, it could add even greater value to know that the same facility, person, or location was observed multiple times in multiple locations or from several different vantage points. Correlations in behavior across time and space can be powerful indicators of coordinated surveillance efforts.

14.3.3 Behavior

As in other patterns of crime, it is important to let behavior reveal intentions, rather than engaging in the popular public safety “what if” parlor game (i.e., “if I was the bad guy, I would...[insert favorite next incident, target, etc., here]”). Again, the behavior in question may not appear obviously aberrant or otherwise unusual. Rather, it stands out in contrast to “normal” or expected patterns and trends when analyzed and compared against context. In many of the examples outlined in this chapter, the behavior observed was not overtly unusual or dangerous so as to prompt an immediate call for service. Only when viewed in its entirety and within the context of “normal” behavior, however, did the hostile surveillance patterns emerged.

14.3.4 Operationally Actionable Output

Similar to other types of analysis, providing a map of suspicious activity to the operational personnel for use in surveillance detection planning can be greatly appreciated and results in a much better operational plan by visually refining and depicting potential target areas. This can be particularly helpful with multibuilding facilities or complexes, like the shopping mall example, by further refining the specific areas of interest. It also can begin to provide additional insight into the true nature of the suspicious activity, or the “why” of

the behavior. Mapping or otherwise providing some sort of visual depiction of any identified spatial patterns or trends can be especially useful in conveying this type of information.

Sophisticated mapping software, although generally very beneficial and frequently used by most public safety agencies, is not entirely necessary for an analysis of this nature. Internet-based mapping tools, orthophotography images, or even line drawings, such as those included in this chapter, all convey the necessary information and can be more than adequate for this type of analysis. In many ways, a map can be viewed simply as a specialized figure or graph, a unique way to visually depict data or information. Although many mapping programs have sophisticated mathematical tools associated with them, in this situation, visually depicting the information so that the operational personnel can guide their efforts and begin to determine what is occurring and why is the most important aspect of this exercise and does not require any additional analytical software.

Mapping the data over time also can be especially valuable in determining whether the location associated with the greatest activity or most marked increases appears to move, change, or otherwise refine itself over time. In some cases it is more useful to think of a relatively fluid “cloud” of potential risk that has moved into or settled over a particular area, rather than struggling to identify and define discrete areas. Thinking of the edges as being somewhat fuzzy will limit restricting the area too much and missing potentially significant activity in the future.

Returning to the shopping mall example, let us assume that the increased number of suspicious situation reports at the mall is not related to any reporting bias and that the pattern of results is related to something associated with the mall itself. We could stop at this point and suggest that the mall increase patrol during the weekend, particularly during the evening hours, but there is some additional work that can be done to further refine the scope and add value to our understanding of what might be happening at the mall. Any additional trends or patterns that we can reveal in the data can provide additional insight into a possible motive for the unusual activity, which then translates into greater definition and refinement of the response options.

By creating a map of the report locations throughout the shopping mall example, an obvious pattern emerges. The majority of the activity seems to be centered in the vicinity of the cinema (Figure 14.4). This finding also is consistent with the day and time of the reports. The cinema tends to be more active in the evening hours, particularly on Friday and Saturday nights.

The mall’s suspicious situation reports could reflect preoperational surveillance for anything from robbers to sexual predators surveying a target-rich

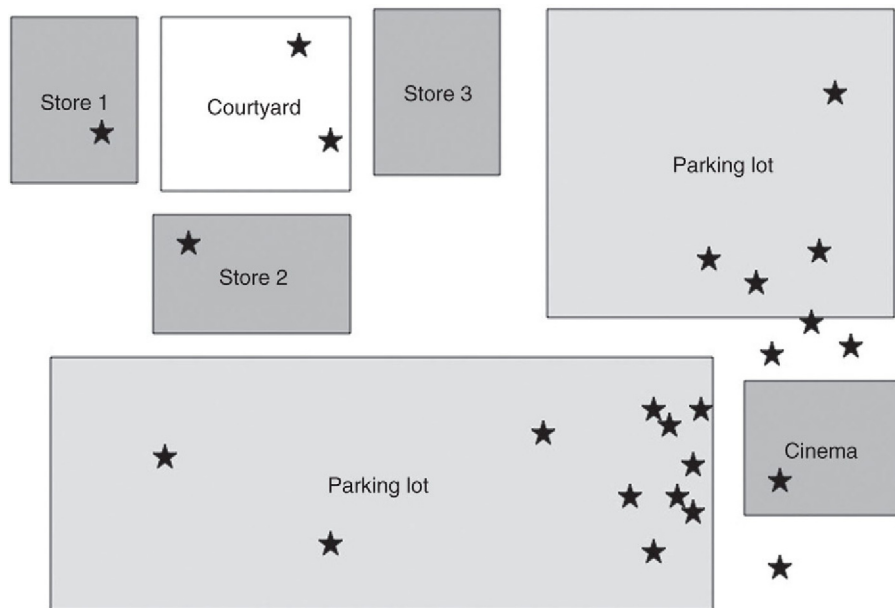


FIGURE 14.4

Map illustrating specific locations associated with the situation reports from [Figures 14.2 and 14.3](#).

environment for potential victims to an extremist group interested in calling attention to its agenda. Our analysis does not necessarily address the “who” or specific “why” of this activity. What it does, though, is characterize the behavior sufficiently that coordinated surveillance detection efforts and operational deployment can be targeted specifically to the time and location associated previously with possible surveillance activity. This limits the personnel resources required for formal surveillance detection and increases the likelihood that surveillance detection activities will be placed when and where they are most beneficial. Moreover, routine patrol can be concentrated when and where activity is greatest. Minimally, this increases the opportunity for rapid response should something bad happen. Ideally, placement of operational personnel when and where they are likely to be needed gives us the opportunity to anticipate and possibly even prevent crime.

In the shopping mall example, a series of suspicious situation reports create a very simple database. We were then able to characterize the data and drill down to extract additional details that could be used to create a focused surveillance detection plan while guiding additional public safety and crime prevention approaches. Sometimes, however, additional steps will need to be taken to further characterize the data, link possible associated events, identify potential transitions or escalation in surveillance activity, and make predictions about

possible future behavior. This is particularly a challenge in high-profile sites or in locations where a variety of information has been compiled and needs to be culled for meaning.

14.3.5 Other Sources and Methods

Other sources and methods are being developed continuously that can be used to collect and analyze putative hostile surveillance. A few are described next.

In the Northern Virginia military shooting series outlined in Chapter 13,⁶ additional analyses included the use of high-resolution geospatial data (Light Detection and Ranging, LiDAR) to create a line of site or “viewshed” analysis. Briefly, viewshed creates a three-dimensional model or viewshed of what can be seen from a particular location by taking into account the terrain, elevation, and other physical features that could constrain or otherwise limit the shooter’s line of site. Similarly, by creating a three-dimensional model of what one could see or could be seen by a specific location, surveillance detection, including identification of the red zone and related surveillance detection locations can be further informed.

At the time that the first edition was written, access to high-resolution imagery generally was limited. Many municipalities collected and made available orthophotography imagery, and there was some effort made to understand its use from a customer service perspective. Now, however, GoogleEarth and related capabilities have made high-resolution imagery readily available, which may be both good and bad from a crime and intelligence analysis perspective. While not deployed explicitly to support surveillance, it can be used to discreetly understand space, including line of sight or viewshed, and context in support of operations. Unfortunately, as discussed later, this resource also has obvious operational value for our adversaries. In addition to the geospatial content and imagery deployed over the Internet, the ability to conduct relatively sophisticated geospatial analysis, including route planning, and the increased availability of street views and related geospatial capabilities can be used to create sophisticated casing reports and operational plans in support of target selection and preattack planning.

Another downside to the deployment of potentially sensitive information over the Internet is the high degree of anonymity associated with it. Just as child predators have been able to exploit the anonymity of the Internet, other individuals with malevolent intentions have been able to take advantage of the vast amounts of information available, easily concealing their activities. By using anonymizers or spoofed IP addresses, it can be extremely difficult to identify a particular source or individual. This allows an individual or group to easily exploit open-source material to conduct preliminary surveillance virtually undetected. On the other hand, these measures might not even be necessary

given the tremendous amount of traffic currently on the Internet. The amount of information contained in weblogs alone can be staggering, and it is increasing continuously.

Finally, there are the cameras that you know about and those that you do not. In addition to security surveillance cameras and ubiquitous webcams, unmanned aerial vehicles (UAVs) or drones are increasingly available. While debate continues regarding access to and lawful use of UAVs, it appears that the bad guys have embraced the technology and are using it for all manner of surveillance including high tech peeping.⁷

14.3.6 Red Teaming

It is important to consider the other side in a discussion of surveillance detection sources and methods. In their discussion of fourth-generation warfare, Lind *et al.*⁸ noted that, "Terrorists use a free society's freedom and openness, its greatest strengths, against it." Many organizations, agencies, and localities deploy a tremendous amount of sensitive information over the Internet in a misguided attempt to achieve the ideal of "transparent government." For example, a cursory review of municipal websites reveals everything from specific details regarding emergency response equipment, including equipment model numbers, to detailed, high-resolution images of sensitive locations. As early as 2001, the Israelis reported that their adversaries were exploiting the increased availability of orthophotography images freely available over the Internet,⁹ and many localities have reported suspicious or unusual activity on their websites, especially as relates to local infrastructure and public safety.¹⁰

As an analyst, therefore, an important step in the analysis of potential hostile surveillance would be to review and assess what other sources of information are available and what value they might have for someone with less than altruistic intentions. Why is this process important? First, as always, the key to effective and meaningful data mining is domain expertise. Knowledge of what information is deployed through other sources including the Internet, how it is organized, and what it might mean from either a tactical or strategic perspective is critical to understanding the value that it may have to our adversaries. It also is important to consider how this information might be combined with other information on the same website or with other resources, including identified patterns of hostile surveillance. For example, patrol boundaries can be very helpful in extrapolating average response times. Information pertaining to workload, including crime rates or calls for service, would add value to the calculation of deployment and potential response times.

Discussion of Internet data has been included in Chapter 6, but approaches and tools similar to those employed by online retailers to characterize online

shopping behavior and create models of potential buyers, so-called web mining tools, can be utilized by analysts in an effort to identify and monitor possible Internet surveillance and protect critical infrastructure. Using these same web-mining tools, Internet activity can be analyzed in a timely, comprehensive fashion. These data can then be merged with additional surveillance detection information and used to support a comprehensive, integrated approach to surveillance detection.

WHAT ABOUT BOB?

“Just because you’re paranoid doesn’t mean they aren’t after you”

Joseph Heller

It has become common practice, ostensibly in an effort to personalize service, to ask for the customer name during even relatively minor transactions. By way of example, industry professionals now encourage fast food establishments to ask for the customer’s first name in an effort to “personalize” their order by calling it out when their food is ready or writing it on their bag.¹¹ While it might be true that using a customer’s first name may enhance the experience and make them feel special, it also creates a unique collection opportunity for potential predators.

The dark side of this is that by sharing even a first name in this relatively anonymous environment, I have seeded the information collection process for anyone with less than positive intentions. Predators are especially adept at using social context and informal social “requirements” to increase intimacy with and gain access to potential targets. Stalkers and other predators are particularly adept at creating and exploiting situations involving social obligation. By sharing a first name in these situations, it facilitates approach. While the potential victim is trying to recall how they might “know” this stranger who has just greeted them with a friendly smile, using their name, the predator already has a foot in the door.

Related to this is over-sharing on the Internet. Social media has markedly changed patterns of social discourse and interaction; not always for the better. Many people who go to great lengths to set timers on lamps, stop their mail and paper delivery, and even make arrangements to park extra cars in their driveway to create the perception that they are home, think nothing of posting vacation plans, including destinations and itineraries, and photos taken throughout their trip, which effectively convey critical information to would-be home burglars. While this has been highlighted repeatedly in the media,¹² over-sharing has become so common as to be second nature to most.

Underscoring the value associated with ostensibly innocuous personal information, several years ago I saw a “get to know me” e-mail message that was circulating on the Internet. While at first glance it appeared to have been written by a teen, subtle clues revealed that it actually was a cleverly worded collection instrument used by a child predator to familiarize them with and groom a potential target. In addition to the standard questions regarding their age, activities, favorite music groups, dreams, aspirations, and goals, this particular note included questions regarding points of tension between the potential child victim and their parents that could be exploited and used as a wedge during the grooming process. While most teens would balk at such aggressive questioning in any other public place, the Internet in particular creates the perception of safety and privacy in an environment that is anything but that. Moreover, while the chatty tone was

intended to mimic the language of a peer and create the perception of familiarity, the recipient of the questionnaire had absolutely no idea who the sender was or what their intentions truly were.

In contrast to unintentional or otherwise passive over-sharing, people increasingly see their personal data in a transactional model and are willing to exchange even very private information for perceived savings or a personalized experience.¹³ While I may balk at intrusive government collection, I am more than happy to provide detailed insight regarding my purchasing behavior, financial transactions, and routine activities if it will optimize my Internet searches and secure personalized discounts. Unfortunately, these same, detailed pattern of life data that support complex analytics conducted on my behalf by commercial entities that I currently or may patronize, also may enable specifically targeted attacks in the hands of bad actors.

Getting back to the fast food establishment that asks for my first name in an effort to personalize the experience and ensure that I come back because I feel wanted, I personally view this as an unacceptable violation of personal security and use the name “Bob” whenever asked to provide my name in these situations. While it frequently causes the cashier to pause for a moment, it is a name that I recognize when my order is called, but also creates an immediate red flag for someone trying to leverage this situation to create a sense of familiarity with me. Perhaps I am being paranoid, but it keeps me sharp and good operational security or OPSEC only works if you are consistent.

Unsupervised learning or clustering techniques also may be used to characterize “normal” patterns of activity in an effort to identify possible surveillance activity. This method of anomaly detection frequently can reveal patterns of unusual behavior. People often get tripped up and caught when they try to behave normally or “fly under the radar.” In many cases, however, they do not have a good sense of what “normal” truly looks like and get caught out of ignorance or because they stand out even more in their attempts to be inconspicuous. It is often difficult to completely understand what “normal” looks like until we characterize it and then analyze it in some detail. Similarly, language or cultural differences can impair an individual’s ability to melt into the background noise. Ignorance of cultural subtleties, nuances, or norms can serve as a spotlight, highlighting unusual or suspicious behavior. It is for this reason that characterizing normal trends and patterns can have value, as it provides a baseline against which unusual or suspicious behavior can be measured.

14.4 SURVEILLANCE DETECTION CASE STUDIES

At a minimum, the ability to characterize suspicious behavior provides invaluable guidance for those interested in establishing surveillance detection. Operational resources almost always are in short supply and must be deployed as efficiently as possible. When, where, and what in support of SD resource allocation and optimization, information-based risk and threat assessment, prevention and thwarting, and informed approaches to response and consequence management.

14.4.1 The Complex Environment

The ability to take a series of suspicious situation reports and identify trends and patterns gives us the opportunity to deploy surveillance detection when and where it is most likely to gather additional information. But what happens when there are multiple potential locations of apparent interest? A multibuilding complex or facility with several layers of physical security is going to require more complex surveillance activity, and concomitantly more sophisticated surveillance detection to accurately detect, dissect, and convey the overall pattern of activity.

In this fictitious example, there is a multibuilding complex, which is depicted in [Figure 14.5](#). The facility is surrounded by a 6-ft perimeter fence ([Figure 14.6](#)). There is only one point of access to the facility, through the front sally port, which is continuously manned. Due to the nature of the complex, suspicious activity is aggressively reported and investigated. The reports are then compiled for historical archiving ([Figure 14.7](#)).

After an incident at a related facility, the security manager decides that the suspicious activity reports should be reviewed, characterized, and analyzed. Using data mining and predictive analytics, the reports were analyzed and classified into four separate groups. The analyst assigned to the task selected an

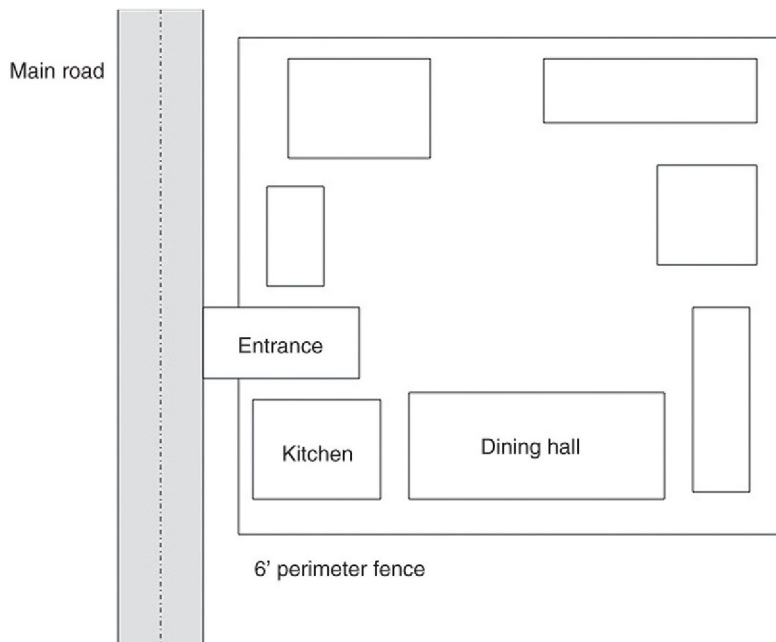


FIGURE 14.5

Map of a fictitious multifacility complex associated with suspicious activity.



FIGURE 14.6

Six-foot fence surrounding the perimeter of the compound. *Staff Sergeant Tom Ferguson, USMC; used with permission.*

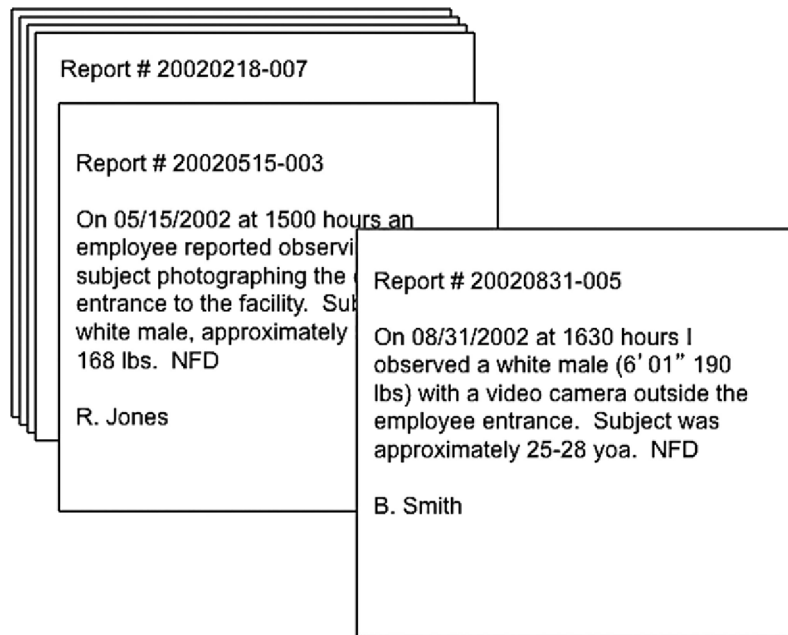


FIGURE 14.7

Samples of suspicious situation reports received and logged by security.

unsupervised learning technique, which clustered the incident reports based on similar characteristics. In an effort to convey the information to the operational personnel in an actionable format, a facility map diagram was prepared in which the locations associated with the different clusters of activity were marked and highlighted.

The location indicators on the map were intentionally depicted as vague areas rather than solid areas in an effort to convey a general area of risk, rather than specific indicators or points, which might indicate specific locations. Again, using these “clouds” of risk conveys increased activity associated with this general location that might be associated with a concomitant elevation in associated risk. Similarly, size, color, and even relative differences in color saturation or intensity can be used to convey additional information, such as frequency of activity, or temporal variance. By using these techniques, the analyst can convey a relatively large amount of information through the use of a two-dimensional map.

The first cluster of activity, which is indicated by the number “1” in [Figure 14.8](#), was characterized by activity outside the perimeter. This was frequent, as

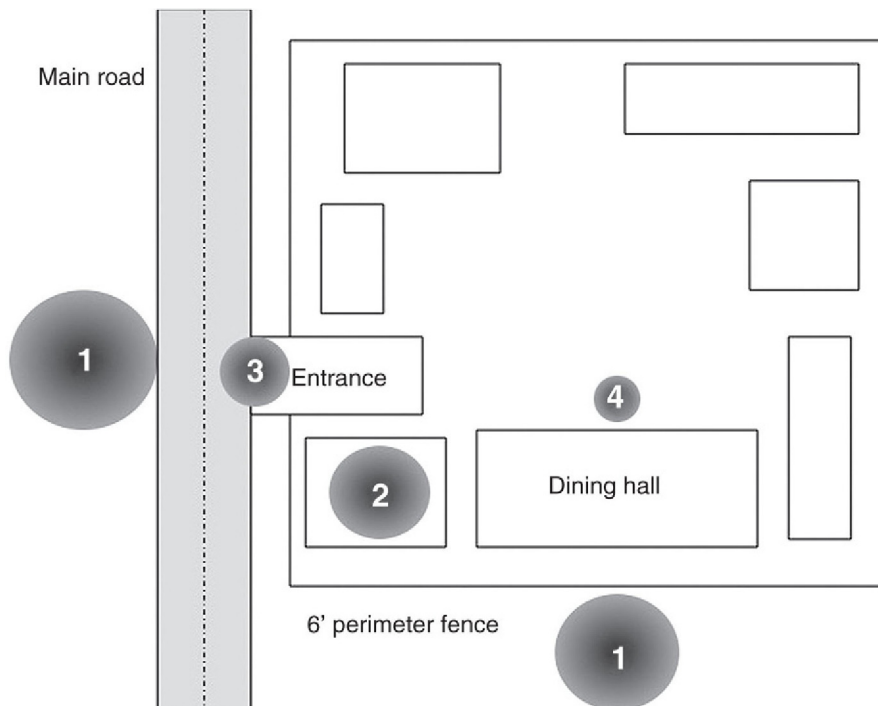


FIGURE 14.8

Map depicting “clouds” of risk associated with various locations within the compound.

**FIGURE 14.9**

Front gate of compound. Staff Sergeant Tom Ferguson, USMC; used with permission.

indicated by larger clouds of risk on the diagram. In particular, significant activity was associated with the front gate (Figure 14.9). Analysis revealed that the activity associated with this cluster not only increased in frequency over time but appeared to intensify as well. Additional surveillance activity was associated with the area outside the fence (Figure 14.10) in relative proximity

**FIGURE 14.10**

Area outside the fence near the dining hall. Staff Sergeant Tom Ferguson, USMC; used with permission.

to the dining hall. Further refinement associated with the time of this activity was noted, which initially appeared random and subsequently appeared to coincide with meals.

The second cluster of activity was associated with the kitchen. This also was associated with relatively frequent reporting of unusual behavior, and even included one situation where an unauthorized person gained access to the facility in a delivery truck. The activity in the second cluster differed from the first in that it represented very little overt surveillance, but did include several suspicious telephone calls and inquiries regarding delivery and dining schedules.

The third cluster of activity was associated with the entrance. Again, there was not much overt visual surveillance of the facility. This cluster was associated with security probes, which included conversations and inquiries involving the personnel manning the entrance. This pattern of activity also distinguished itself in that it started to occur after the perimeter surveillance had already been operating for a period of time.

The fourth cluster of activity was by far the least frequent and the last to occur in the time series. In many ways, the incidents included in this “cluster” comprised such a diverse array of incidents that they were almost discarded as outliers or anomalies. They occurred much later than all other incidents, after a break in activity. They differed significantly in terms of the nature of the behavior and time of day, and included an unauthorized person who tried to gain access to the dining hall during a meal, as well as a triggered alarm at the entrance to the same dining facility one night. The only consistent factor was the location – the entrance to the dining hall. After the other clusters were mapped and evaluated, however, it was determined that this loose array of incidents might represent the final preoperational planning stage to an incident.

In response to this analysis, surveillance detection, physical security enhancements, and proactive deployment operational plans were developed. These were based on the specific decision rules associated with each identified cluster, which ultimately were linked to a particular set of vulnerabilities identified in the fictitious complex. This permitted the specific targeting of resources, as well as the development of additional security enhancements that were based on the associated risks related to each specific location within the complex.

By using operationally actionable mining and predictive analysis, force protection resources and strategies can be deployed in direct response to the analytical output. This includes the specific targeting of resources, as well as the development of additional security enhancements that are based on the unique constellation of associated risks related to each specific location within a multifacility base or complex.

14.4.2 Escalating Hostile Surveillance

As outlined in Chapter 12, risk and threat assessment includes the identification of the likelihood of risk, as well as characterization of the potential threat in support of informed approaches to prevention, thwarting, mitigation, response, and consequence management. In this particular example, which is based on an actual analysis,¹⁴ suspicious situation reports were analyzed in support of information-based risk and threat assessment.

After a meeting to discuss a suspected threat, the facility security manager indicated that they had a number of suspicious situation reports, which had been investigated, although never analyzed. These reports had been compiled into the standard security Records Management System (i.e., a three-ring binder). This particular facility had large stores of valuable assets in the subterranean portion of the building, and the working assumption was that these valuables were the primary targets.

Data mining and predictive analytics were used to characterize possible surveillance activity associated with the facility. As can be seen in [Figure 14.11](#), a

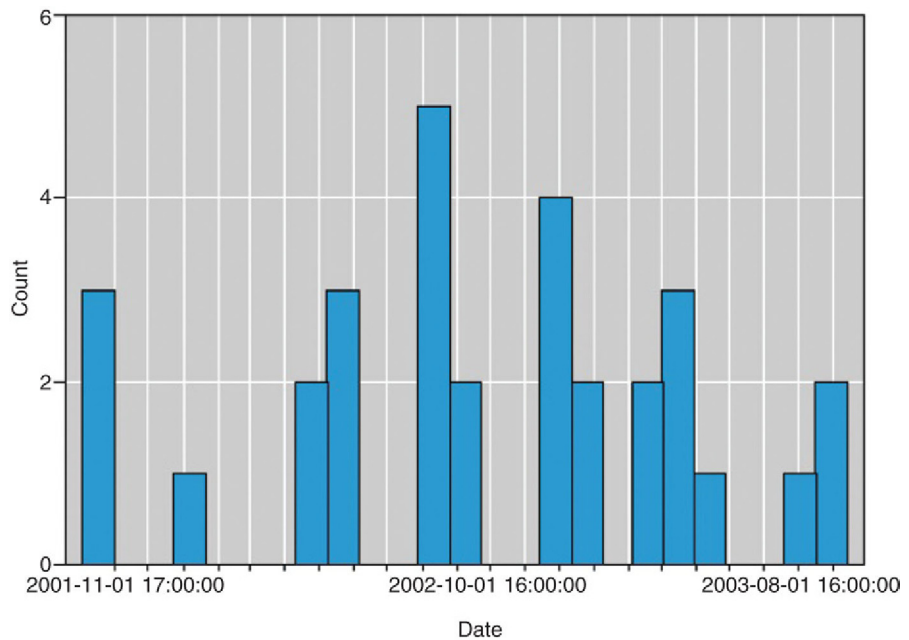


FIGURE 14.11

Frequency distribution depicting the relative occurrence of suspicious activity reports associated with a facility of interest.

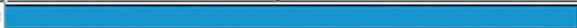






Value	Proportion	%	Count
WED		25.0	11
TUE		15.91	7
THU		15.91	7
FRI		15.91	7
MON		13.64	6
SAT		9.09	4
SUN		4.55	2

FIGURE 14.12

Distribution of suspicious activity by day of week.

quick review of the frequency of reports over time revealed increasing activity consistent with growing interest in the facility. Analysis of the activity by day of week further highlighted the nonrandom nature of this activity; 25% of the reported incidents occurred on Wednesdays (Figure 14.12). Of note, this finding correlated with other casing reports on financial institutions within the United States that had been recovered from other bad actors, which also indicated increased activity on Wednesdays.¹⁵ Reviewed in isolation, these events might indicate nothing more than something idiosyncratic or unique to the facility, or even reporting bias. On the other hand, the finding of increased activity on Wednesdays across multiple facilities increases the value of that observation and supports the possibility of coordinated activities and common planning.

The incidents were recoded into operationally relevant categories that more accurately described the suspect behavior. These included still photography (“photo”); video photography (“video”); any movement toward the facility, attempted interaction with the security personnel, or probing of the perimeter (“approach”); and all other behaviors not appropriate for inclusion in any of the previous categories (“Suspsit”). These recoded incidents were plotted over time, but, as can be seen in Figure 14.13, any interpretation of these results was limited by the complexity of the graph created.

Moving beyond simple descriptive statistics and characterization, a clustering technique was used to determine whether the events could be grouped based on their time, nature, or location. Unsupervised learning or clustering techniques can be used to identify naturally occurring groups, associations, relationships, and behavior. These approaches also may be used to characterize “normal” patterns of activity in an effort to identify possible unusual or otherwise suspicious activity – putative indicators of hostile surveillance. Anomaly detection frequently can reveal patterns of unusual behavior as people may reveal themselves when they try to behave normally or “fly under the radar.”

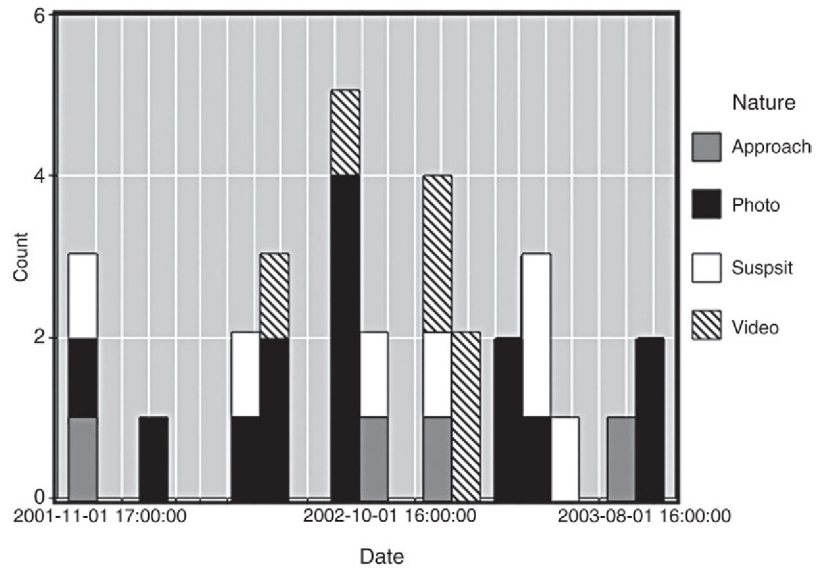


FIGURE 14.13

This figure depicts the distribution of suspicious behavior over time. “Approach” indicates that the suspect physically approached the facility or attempted to probe the security features or personnel, “Photo” indicates the suspect use of still photography, “Video” refers to the suspect use of video photography, and “Suspsit” includes all other behavior not included in the previous groups.

In many cases, however, they do not have a good sense of what “normal” truly looks like and get caught out of ignorance or because they stand out even more in their attempts to be inconspicuous. In fact, it is often difficult to completely understand what “normal” looks like until we actually study and characterize it. It is for this reason that characterizing normal trends and patterns can bring significant value to the analyst as it provides a baseline against which unusual or suspicious behavior can be measured.

In this particular example, we were looking for actionable patterns within the suspicious behavior that had been reported. The analysis revealed two different clusters of suspicious situation incidents, which generally were associated with different types of observed behavior of differential relevance from a surveillance detection perspective (e.g., still photography versus videotaping and other operationally oriented surveillance to include security probes). As can be seen in [Figure 14.14](#), graphing these groups across time reveals a transition in the nature of suspicious activity from relatively simple behavior to more operationally oriented surveillance, suggesting an escalation in the nature of surveillance activity that paralleled the increase in frequency over time.

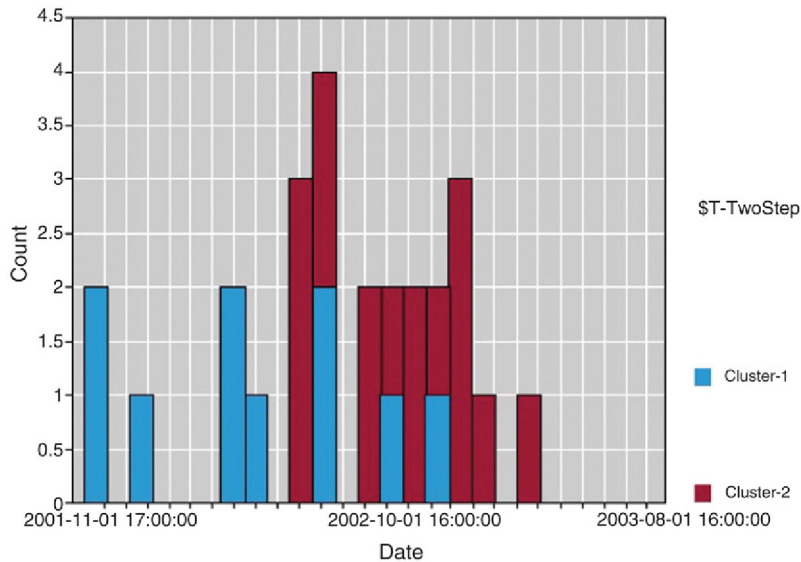


FIGURE 14.14

Pattern of suspicious behavior. This figure depicts an identified pattern of suspicious behavior over time, as revealed through the use of a clustering or unsupervised learning technique. Group membership was determined largely by the nature of the activity. Cluster 1 generally was associated with still photography of the facility, while the incidents in Cluster 2 tended to be associated with more operationally oriented activity, including video surveillance.

Using relatively simple techniques, it was possible to generate operationally actionable output from the analysis. As illustrated in Figure 14.15, preparation of a crude facility map highlighted the relative spatial distribution of the incidents. Additional value was added to the map through the use of different shades of gray to depict the nature of the activity and different intensities to convey relative differences across time. This simple technique also can serve to highlight the emerging geographic specificity of the suspected surveillance activity.¹⁶

This and the previous example underscore that mapping does not require sophisticated technology or need be confined exclusively to traditional geographic boundaries. Relatively simple “maps” can be created for facilities or complexes, or even single buildings associated with differentially distributed risk. These maps can facilitate the identification of specific patterns of offending or risk associated with particular times and/or specific locations, which can be used to provide insight regarding specific targets and/or attack scenarios. Moreover, this type of map can be extremely useful for personnel deployment or specifically targeted crime prevention strategies, particularly if it reveals increased attention to a unique attribute or feature, or some other actionable aspect of the pattern.

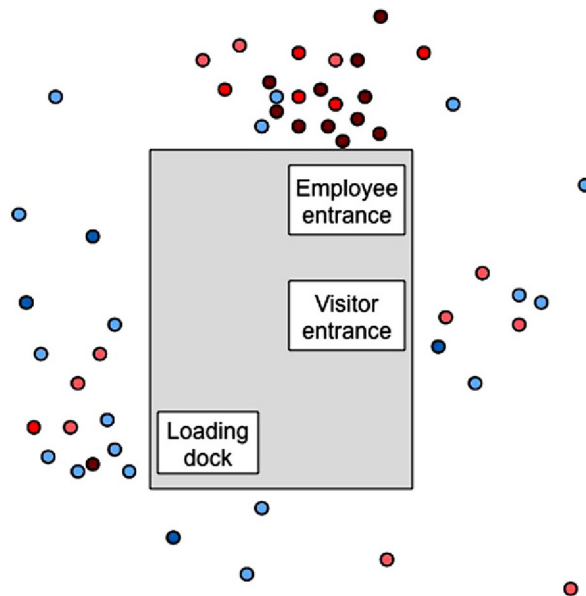


FIGURE 14.15

“Map” depicting the spatial distribution of suspicious activity around the facility of interest. Note that the shade of the icon is associated with the cluster membership, which visually highlights the spatial focusing of interest over time.

Overall, this analysis provided three direct benefits. First, we were able to identify a coordinated pattern of hostile surveillance on the facility, which confirmed it as a potential target. Second, through the use of predictive modeling, we were able to characterize the pattern of surveillance in support of informed approaches to surveillance detection resource allocation and optimization. Ideally, by prepositioning these assets we will increase the likelihood that our surveillance detection resources will be able to collect even more information in support of informed prevention. Finally, the analysis also revealed escalation in the nature of the hostile surveillance, as well as spatial refinement in the activity; documenting a shift from locations associated with the valuable assets stored in the facility to the employee entrance. By letting behavior reveal intentions, the analysis provided additional insight regarding the specific location targeted and possible nature of the attack. While it is unclear whether this revealed change in activity on the part of the adversary reflected a true shift in focus, or merely revealed uninformed early assumptions regarding the nature of the attack, the analytic results supported information-based approaches to deployment, prevention, and response planning.

14.4.3 “Food Truck”

The previous examples highlighted the value that thoughtful, well-considered analysis can bring to the identification and characterization of putative hostile surveillance in support of informed surveillance detection resource allocation, risk and threat assessment, prevention, and response planning. While these results provided increased insight and actionable results regarding trends and patterns embedded within a series of suspicious situation reports, the delay required to collect and process these data in support of finished intelligence products can be unacceptable, particularly in situations where forward deployed operational personnel require immediate feedback to support decisions in the field, or when an attack might be imminent. Similar to the model outlined for web-enabled deployment of deployment models outlined in Chapter 13, a surveillance detection web services model would enable a deployed operator to identify behavior or a person of interest, enter their report directly, run it against existing “knowns” and deployed algorithms, and receive immediate feedback regarding whether and how this information fits within the context of existing intelligence, including implications for officer safety. Based on this feedback, the operator could then make a truly informed decision in real time regarding next steps.

In keeping with this model, the following notional example was developed by the SAP National Security Solutions (NS2) team to illustrate the use of advanced analytics in support of a common surveillance detection training scenario. In this particular example, the goal is to assess the activity of food truck vendors in Washington, DC, based on classification, location, frequency of activity, putative affiliation and coordinated action, and customer sentiment. By gathering information on food truck activity and using predictive analytics in the SAP NS2 HANA in-memory appliance, an analyst can analyze large amounts of data in real time to characterize “normal” trends and patterns, as well as identify unusual or otherwise suspicious activity worthy of additional follow up and review.

As can be seen in [Figure 14.16](#), the analyst can effectively integrate different intelligence types to include geospatial intelligence (GEOINT) and human intelligence (HUMINT). By visualizing the data in a geospatial environment, the analyst can easily explore the data, incorporating their tacit knowledge and domain expertise in the interpretation of the data, and beginning to identify trends and patterns of behavior. New reports to include food truck sightings also can be easily updated and added directly using this interface, which enables field operatives to quickly and easily capture and log locations and related content to update data and related analytics.

[Figure 14.17](#) illustrates the text analytics functionality, which enables the direct incorporation and analysis of unstructured data to include the location and

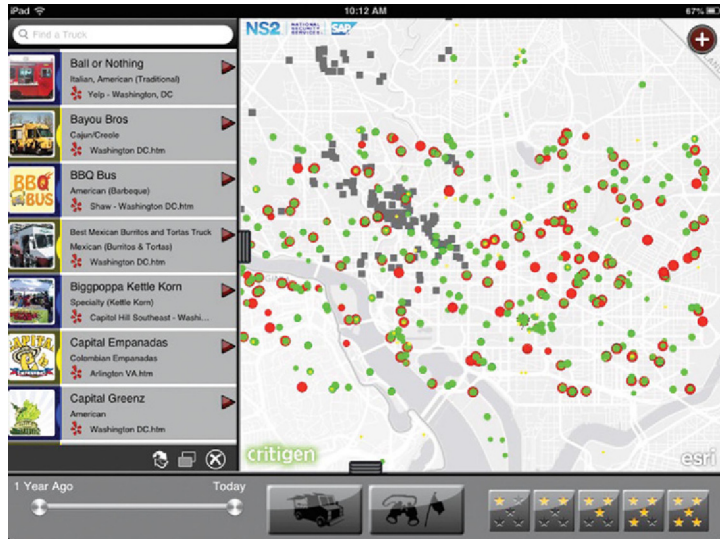


FIGURE 14.16

Depicts the “Food Truck” example developed by SAP National Security Solutions (NS2) to illustrate the use of advanced analytics in the analysis of putative surveillance activity. This image depicts the visualization of geospatial intelligence (GEOINT) and human intelligence (HUMINT) in a geospatial environment, which enables the fluid exploration of the content. *SAP NS2, used with permission.*

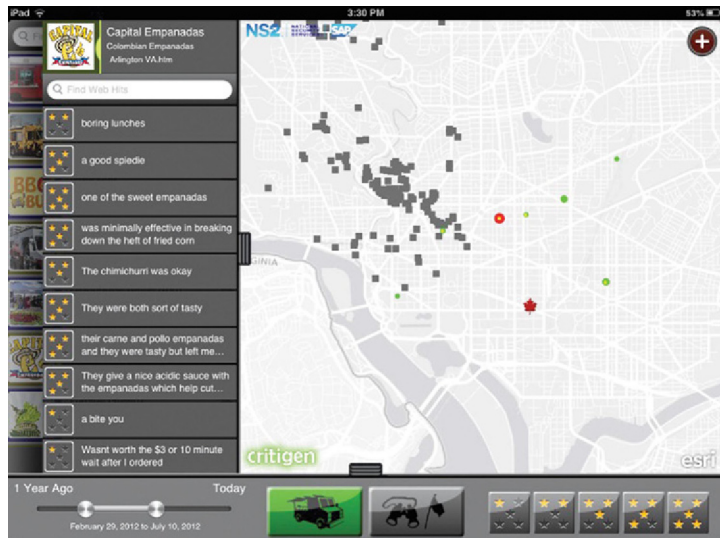


FIGURE 14.17

Illustrates the analysis of unstructured narrative collected from social media for sentiment. *SAP NS2, used with permission.*

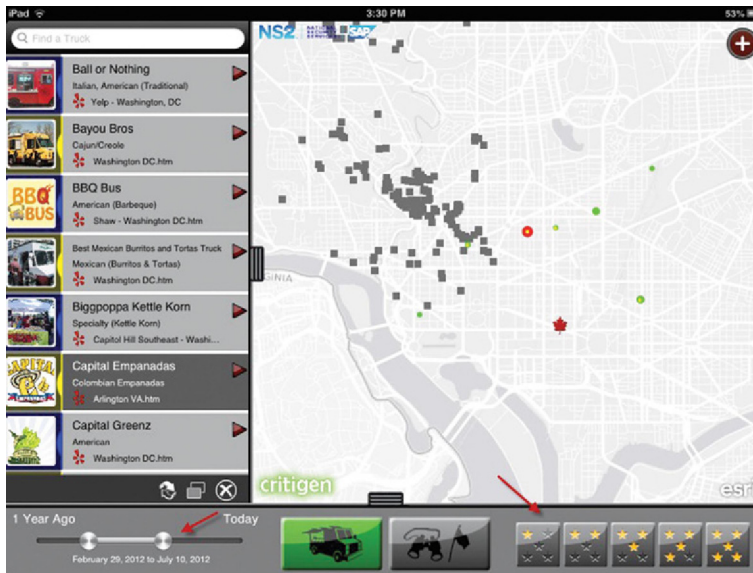


FIGURE 14.18

Depicts refinement of the analysis to include sentiment (lower right), selection of a temporal sampling frame (slider bar on lower left), and selection of a specific food truck from the list on the left side of the panel. *SAP NS2, used with permission.*

additional narrative regarding food trucks in the area of interest (AOI), as well as sentiment analysis posted by food truck patrons on social media sites including Yelp and Twitter. The ability to directly incorporate real time social media or other transactional data greatly enhances the timeliness and value of the related analysis, particularly regarding operational situational awareness to include enhanced collection efforts.

In [Figure 14.18](#), the visualization and related analysis can be refined through the selection of a specific food truck, sampling frame or time period of interest, and sentiment rating or other narrative. These functions also can be used to quickly classify and visualize activity, including common patterns and trends, within a geospatial environment, which provides additional context in support of meaningful interpretation and operational use of the results.

[Figure 14.19](#) depicts the use of additional geospatial tools, which enable the analyst to capture the activity of interest within a specific distance, to include food trucks and other points of interest. Similarly, the analyst can use geospatial analysis to select multiple trucks and review their associated locations, temporal activity, and HUMINT in an effort to explore possible relationships, and either confirm or deny putative collusion or other patterns of associated behavior.

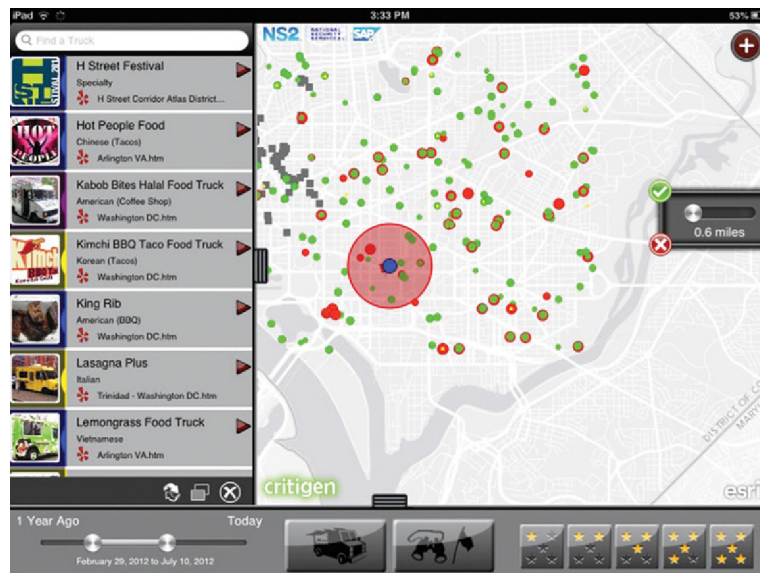


FIGURE 14.19

Illustration of the use of additional geospatial capabilities to calculate and use distance functions in an effort to identify food trucks and additional points of interest within a certain distance from a particular location. *SAP NS2, used with permission.*

Finally, [Figure 14.20](#) illustrates a dashboard that can be used to view, analyze, and explore the collected data. Again, this capability enables real time sharing of incidents and other significant intelligence, providing both analysts and operators timely feedback, enhanced decision support, and situational awareness in an easy-to-use, intuitive mobile map interface that enables direct operator, analyst, and command staff interaction, cooperation, and collaboration.

As outlined in the case studies, the key in operationally relevant and actionable analysis of hostile surveillance and other indicators of preoperational planning is to convey analytical output and information in a format that is relevant to the end user and immediately actionable in the applied setting. Again, the use of effective visualization tools, including maps and other geospatial capabilities, provides operationally actionable analytical products that can be given directly to personnel in the field. Moreover, these capabilities increasingly are able to concomitantly serve as collection and analysis tools, enabling the operator or other end user to input data directly and receive real time feedback in the field; significantly enhancing their situational awareness. Moreover, the ability to effectively integrate and analyze data from multiple, disparate locations can further inform and extend our understanding of complex, multifaceted attacks, particularly regarding those groups and organizations with

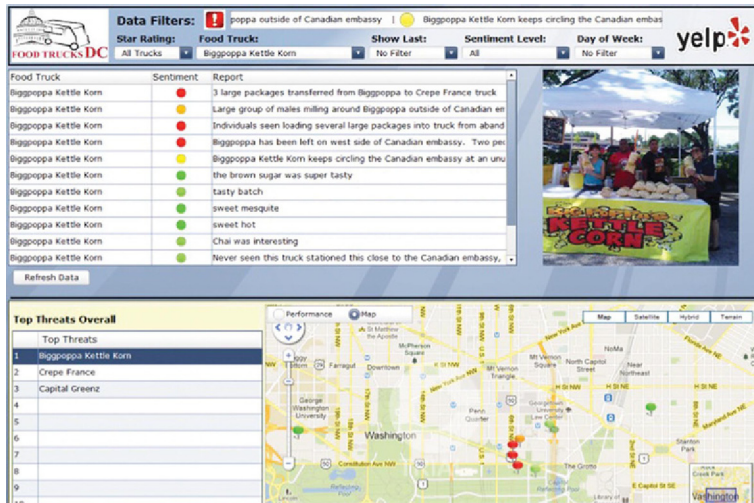


FIGURE 14.20

Screenshot of the dashboard with enabled real-time sharing, exploration, and analysis of data in support of transdisciplinary collaboration, enhanced situational awareness, and information-based operational decisions. *SAP NS2, used with permission.*

a historical preference for multiple, simultaneous, geographically distinct attacks. Access to and analysis of integrated data resources can be used to identify infrequent events and reveal subtle trends or patterns. Moreover, determining “when” and “where” often can provide insight regarding “why.” Therefore, the identification and characterization of surveillance activity can not only refine surveillance detection planning and deployment but also can be used to highlight potential vulnerabilities and threats, which ultimately can be used to support the information-based deployment of countermeasures.

14.5 SUMMARY

A good understanding of “normal” can be invaluable in the detection and identification of possible preoperational surveillance activities at the local level. Just as staging can be detected in violent crime because most criminals do not have a good working knowledge of “normal crime trends” and patterns, those unfamiliar with “normal” in other environments also might reveal themselves when they fail to blend into the surrounding environment. Complex attacks, such as robberies, kidnapping, and terrorist attacks, frequently include a period of preoperational surveillance. Analysis of this activity can be particularly challenging given its relatively low frequency, as well as issues associated with obtaining accurate, complete, and reliable reporting before an attack. Moreover, since the analyst generally is trying to identify and characterize an

attack in the planning stages in support of information-based prevention and thwarting, they may never know whether their analysis was successful. While it is difficult to prove a negative, however, no analyst wants to be “right” about predicting an attack.

Again, identifying suspicious or unusual activity can be compared to finding a needle in a haystack. This is where anomaly detection, which is a very powerful, automated process that can be used to identify and characterize extremely low-frequency events, can have tremendous value. Once a single event has been identified and characterized, it can be modeled and used as a veritable data “magnet” to identify additional needles in the informational haystack.

As described earlier, characterization and analysis of suspicious situation reports can guide future surveillance detection operations by highlighting the times and/or locations that are generating the greatest apparent interest. It is always important to remember, however, that there are the incidents that the analysts know about, and those that they do not. Suspicious situation reports generally reflect only a small percentage of all surveillance behavior. By identifying the times and/or locations associated with the greatest degree of apparent interest, as indicated by an increased number of suspicious situation reports, operational resources can be deployed. This has the potential to increase the amount of behavior that is documented through good, targeted surveillance detection.

In addition to supporting general risk and threat assessment, thoughtful analysis of preoperational surveillance reports also can provide some indications regarding the nature of the planned attack to include timing and location, as well as other indicators of the adversary’s intentions. This information can be used to prevent or otherwise thwart the planned incident, while also supporting informed response planning and consequence management in the event that the attack cannot be prevented.

It is always essential to review surveillance detection or suspicious situation reports within a larger context. Obvious changes, including surveillance detection training, reports that heighten awareness, or major incidents, can greatly impact natural surveillance and reporting and concomitantly influence the data. While high-profile events or recent training can increase awareness, apathy, complacency, or frustration can decrease reporting. Efforts to maintain reporting and surveillance detection efforts can include reminders and refresher courses, particularly if personnel changes are frequent, to ensure that the information is valued and that attitude is conveyed to the frontline personnel.

Any analytical program, regardless of the sophistication of the analytical tools employed, will be severely compromised by incomplete, inaccurate, or unreliable reporting. You cannot analyze what you do not have, so it behooves the analyst to work with the larger team in an effort to ensure data quality to

whatever degree possible. Finally, all results should be interpreted cautiously. Abundant domain expertise and a certain degree of caution is an asset to reviewing these data.

Bibliography

- 1 Carter M. Why feds believe terrorists are probing ferry system. *The Seattle Times*, October 10. http://seattletimes.com/html/localnews/2002058959_ferry10m.html; 2004
- 2 McCue C, Parker A, McNulty PJ, McCoy D. Doing more with less: Data mining in police deployment decisions. *Violent Crime Newsletter*, Spring; 2004, p. 1, 4–5; McCue C, McNulty PJ. Gazing into the crystal ball: data mining and risk-based deployment. *Violent Crime Newsletter*, September, 1–2; 2003; McCue C, McNulty PJ. Guns, drugs and violence: breaking the nexus with data mining. *Law and Order* 2004; 51: 34–36.
- 3 De Becker G. *The gift of fear*. New York: Little, Brown and Company; 1997.
- 4 International Training, Inc., ArmorGroup. Department of Homeland Security Surveillance Detection Training for Commercial Infrastructure Operators and Security Staff Course Syllabus, https://www.fbiic.gov/public/2009/jan/SD_CI_Syllabus.pdf; 2006.
- 5 Nationwide SAR Initiative (NSI). <http://nsi.ncirc.gov/?AspxAutoDetectCookieSupport=1>; Nationwide SAR Initiative Fact Sheet, http://nsi.ncirc.gov/documents/Nationwide_SAR_Initiative_Fact_Sheet_2014.pdf
- 6 McCue C, Miller L, Lambert S. The Northern Virginia military shooting series: Operational validation of geospatial predictive analytics. *Police Chief* 2013, February. http://www.policechief-magazine.org/magazine/index.cfm?fuseaction=display&article_id=2871&issue_id=22013
- 7 Small J. Woman calls police to report drone ‘spying’ outside Seattle apartment, KIROTV.com, June 23. http://www.kirotv.com/news/news/woman-sees-drone-outside-apartment-window/ngRCd/?__federated=1; 2014.
- 8 Lind WS, Nightengale K, Schmitt JF, Sutton JW, Wilson GI. The changing face of war: into the fourth generation. *Marine Corps Gazette* 1989, October 22–26.
- 9 Shapira R. We are on the Palestinians’ map. *Maariv* (Tel Aviv) 2001, May 18.
- 10 Gellman B. Cyber-attacks by Al Qaeda feared. *Washington Post*, June 27. <http://www.washingtonpost.com/wp-dyn/content/article/2006/06/12/AR2006061200711.html>; 2002.
- 11 Bergold, R.T. Let’s get personal. *QSR Magazine*, October; 2012. <http://www.qsrmagazine.com/roy-bergold/let-s-get-personal>
- 12 Siegler, M.G. Please rob me makes Foursquare super useful for burglars. *Techcrunch.com*, February 17, 2010. <http://pleaserobme.com>; <http://techcrunch.com/2010/02/17/please-rob-me-makes-foursquare-super-useful-for-burglars/>
- 13 Morris, T. 3 ways to personalize the customer experience without getting to personal, *Parature*, June 10; 2013. <http://www.parature.com/personalize-cx/>. Cisco. Cisco customer experience research: automotive industry global data; 2013. http://www.cisco.com/web/about/ac79/docs/ccer_report_manufacturing.pdf
- 14 To the best of my knowledge this particular facility has not been attacked, although that is not to say that planned attacks have not been disrupted, prevented or otherwise thwarted.
- 15 Joint DHS and FBI Advisory. Homeland security system increased to orange for financial institutions in specific geographic areas. August 1. www.dhs.gov/interweb/assetlibrary/IAIP_AdvisoryOrangeFinancialInst_080104.pdf; 2004.
- 16 In this particular example, the ability to incorporate motion in the visualization of the surveillance activity over time would better show refinement of the pattern, as well as escalation over time.

Advanced Topics

“Any sufficiently advanced technology is indistinguishable from magic.”

Arthur C. Clarke

I have attempted to outline some of the more common public safety and security challenges in the preceding chapters; however, these topics represent only a limited glimpse of the potential for data mining and predictive analytics in law enforcement and intelligence analysis. Therefore, the goal of this chapter is to provide an overview of additional work in this area and to whet the reader’s appetite for further study.

The following list, while not complete, highlights several areas in which analysts are actively using these techniques with considerable success.

15.1 ADDITIONAL “EXPERT OPTIONS”

Several expert options, including prior probabilities and costs, have been discussed earlier. While it would be impossible to address every option available with each tool, two additional options are worth mentioning at this point given their potential value and relatively common use.

15.1.1 Boosting

Boosting methods can be used to address extremely small sample sizes or infrequent events. These methods confer additional weight or emphasis to infrequent or underreported events. While these frequently can yield greater overall accuracy, like the limitations associated with the data imputation techniques described in Chapter 6, the heterogeneous nature of many patterns of criminal activity can limit the ability to use approaches like this, particularly if they magnify unusual or spurious findings.

15.1.2 Data Partitioning

The importance of using training and test samples was covered in Chapter 8. Different approaches to training and validating models exist, however, which use slightly different partitioning techniques. For example, a three-sample approach to data partitioning also is used, which includes training, validation, and test samples. Like the partitioning method outlined in Chapter 8, the training sample is used to train or build the model. The difference between this approach and the one described earlier resides in the inclusion of a validation sample. The validation sample is used to provide the first estimate of the accuracy of the model created using the training data. These results frequently are also used to fine-tune the model. Finally, as described earlier, the test sample is used to evaluate the performance of the model on a new set of data.

Additional approaches to data partitioning include the use of different percentages of data to the training and test samples. For example, a model can be trained on 80% of the data and tested on 20%, rather than the 50:50 approaches outlined earlier. This approach to data partitioning can be particularly useful when modeling infrequent or rare events, as it results in an increased number of cases of interest from which to create the model, without over representing unusual or spurious findings, which is a limitation with boosting methods.

15.2 UNSTRUCTURED DATA

Text mining is realizing its full potential in applied public safety and security mining and analysis. The ability to tap directly into and use unstructured narrative data has been game changing in many ways. Most analysts understand the value represented in those resources; however, the work required to manually extract that information and recode it is extremely time consuming and generally not as accurate as automated methods. In my own initial experience with a text mining tool several years ago, I was able to quickly search a large number of robbery reports in an effort to identify a series defined by a unique MO. In that first foray into text mining, the tool identified several incidents that I knew about and a few more that were new to me. After this experience, I was a true believer in the power and capacity embodied in text mining tools.

The ability to collect and process unstructured data, including narrative, has increased markedly. From approaches that automate traditional methods like summative content analysis in order to enable rapid processing of large amounts of content,¹ and increasingly extensive foreign language capabilities, text mining has become a powerful tool for the analyst. Extending this

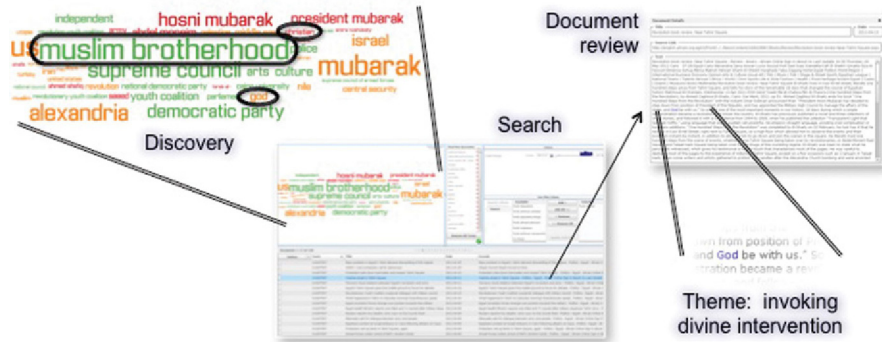


FIGURE 15.1

Illustrates the use of text mining to support exploration, characterization, and description of open source media content associated with the Tahrir Square protests associated with the resignation of Egyptian President Hosni Mubarak.²⁸ *DigitalGlobe, used with permission.*

capability, sentiment and narrative analysis provide greater access to unstructured language content and related insight into the emotional valence and tacit social knowledge in human language. Examples of sentiment analysis include the “Food Truck” case study from the previous chapter, and the social media analysis use case outlined later. Figure 15.1 illustrates the use of text mining and related summary statistics to support the exploration and analysis of open source media content associated with the Tahrir Square protests during the Arab Spring.² The use of summative content analysis to explore and describe this narrative material was reviewed in Chapter 1; however, this content was subsequently explored further through the use of narrative analysis, which “focuses on the ways in which people make and use stories to interpret the world.”³

15.3 GEOSPATIAL CAPABILITIES AND TOOLS

In the first edition of this textbook, crime mapping was relatively new, confined largely to pin mapping and requiring creativity, and in some cases brute force techniques to take the results from predictive analytics modeling algorithms and visualize them in a geospatial environment. Currently, the use of geospatial tools to visualize results has become standard, almost expected practice that uniquely enables interpretation of the results, as well as extension and novel insight through the incorporation of context and transdisciplinary collaboration. Examples throughout the text have demonstrated the efficacy of this approach to visualization, as well as rapid developments in the enabling technology. Moreover, we now are able to use geospatial data in their native form as independent variables in geospatial

predictive analysis; characterizing complex place preferences at an increasingly granular and detailed level in support of information-based anticipation and influence.

In addition to visualization and analysis, derived geospatial products like the line-of-site or viewshed analysis discussed in Chapter 14 further enhance our ability to effectively incorporate location intelligence into analysis. Additional geospatial capabilities are given next.

15.3.1 Automated Feature extraction

Machine learning algorithms increasingly are able to process imagery and extract features including structures, water, vegetation, and debris fields, which enables very rapid processing of large amounts of imagery in support of real-time or near real-time insight. In addition to the clear implications for enhanced situational awareness, the ability to leverage these capabilities in support of change detection can facilitate rapid, informed decision-making regarding resource allocation. Moreover, these results can be used as independent variables in geospatial predictive analysis, or integrated with other derived content to include calls for service in support of informed emergency response and resource allocation. Conversely, sometimes an absence of behavior is key to understanding risk or need. With this in mind, identifying communication “void spaces” within a debris field may highlight greater need than a location associated with a high number of calls for assistance. In these cases where speed is of the essence, the rapid response enabled by automated feature extraction may truly change outcomes.

15.3.2 Full-Motion Video (FMV) and Wide-Area Motion Imagery (WAMI)⁴

The use of closed-circuit television (CCTV) as an investigative tool in the London and Boston Marathon bombings was discussed in Chapter 8. In these cases, CCTV enhanced the investigative pace and efficacy by recording preoperational planning to include dry runs, which facilitated rapid identification and apprehension of the suspects. As discussed in Chapter 8, while the use of this content to support investigation has been good, being able to use streaming feeds to proactively identify attacks during the planning stages would be better. Unfortunately, the large number of sensors deployed and the rapidly growing amount of streaming video content greatly transcends human capacity for meaningful analysis. Therefore, automated methods able to effectively exploit the extremely large amounts of streaming data required to complete the processing, exploitation, and dissemination (PED) cycle for FMV and WAMI content within an operationally relevant and actionable time frame are being developed.

15.4 SOCIAL MEDIA

As discussed in the previous chapter, people increasingly share, and even over-share in social media. The law enforcement community is moving into this area in an effort to collect and share information.

Like many other public-facing organizations, the law enforcement community is beginning to embrace social media in an effort to understand the public's perspective of the organization or the efficacy of specific marketing campaigns for specialized functions including recruitment. [Figure 15.2](#) illustrates metrics providing insight regarding traffic on a specific department's Facebook page. As can be seen in the upper left panel, overall activity on the page can be monitored and analyzed. Content in the upper right panel includes posted messages with the associated number of "likes." Information in the bottom panel includes the analysis of unstructured narrative that is visualized in a number of different ways to provide insight regarding word frequency (lower left panel), as well as sentiment (lower middle panel). Finally, the content visualized in the lower right panel includes information regarding Facebook "friends" of the agency and their activity on the page.



FIGURE 15.2

Example of a dashboard created to support the visualization and analysis of social media data, which includes metrics regarding total activity (upper left panel), specific comments and "likes" (upper right panel), key word use (lower left panel), sentiment (lower center), and summary information regarding posting (lower right). *Information Builders WebFOCUS, used with permission.*

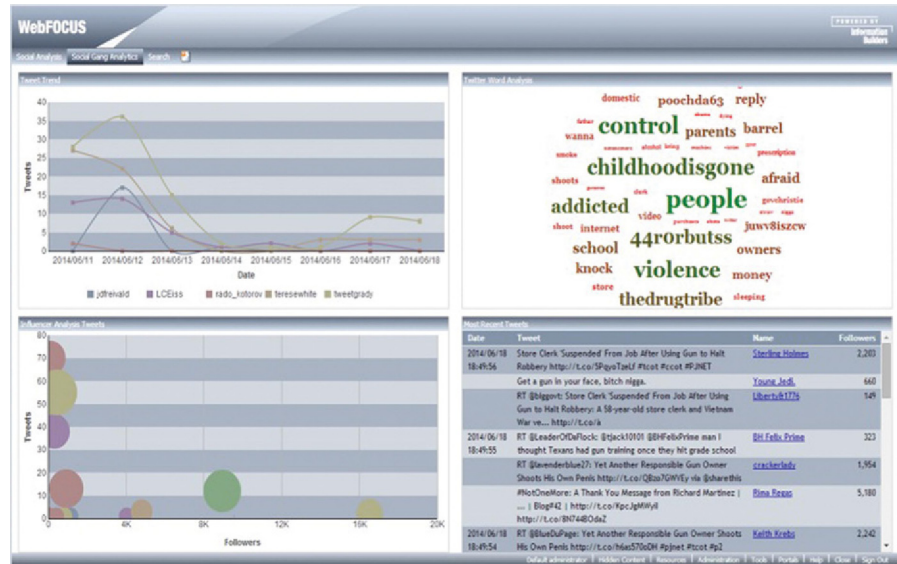


FIGURE 15.3

Illustration of a dashboard designed to visualize Twitter content. *Information Builders WebFOCUS*, used with permission.

Figure 15.3 illustrates a dashboard designed to monitor Tweets for specific accounts, which enables an organization to better follow and understand online conversations regarding their agency including recent, breaking, or pending events, or other incidents or activities of interest to their organization and its mission. Figure 15.4 depicts a related dashboard that enables the fluid exploration and analysis of unstructured content, including social media or investigative notes.

It would be presumptuous to assume that we are the only ones using these capabilities. For example, the Islamic State of Iraq and Syria (ISIS), a Sunni militant group, launched an Arab-language Twitter app called, *The Dawn of Glad Tidings*, which is promoted as a means by which to follow the group. In addition to the provision of news, updates, and other content, messaging and related activities are carefully crafted to support branding and other narrative themes. The overall tone and related activity also reveal a deeper appreciation for Twitter functions and rules.

Similarly, many other groups of interest maintain an active social media presence including Facebook profiles, and the extremist group al Shabaab launched a Twitter account in December of 2011.⁵ Adopting the Twitter handle @HSMPress (Harakat al-Shabaab al-Mujahideen Press Office),⁶ the content was remarkably Western-facing in tone and content to include irony and nuance in its interaction

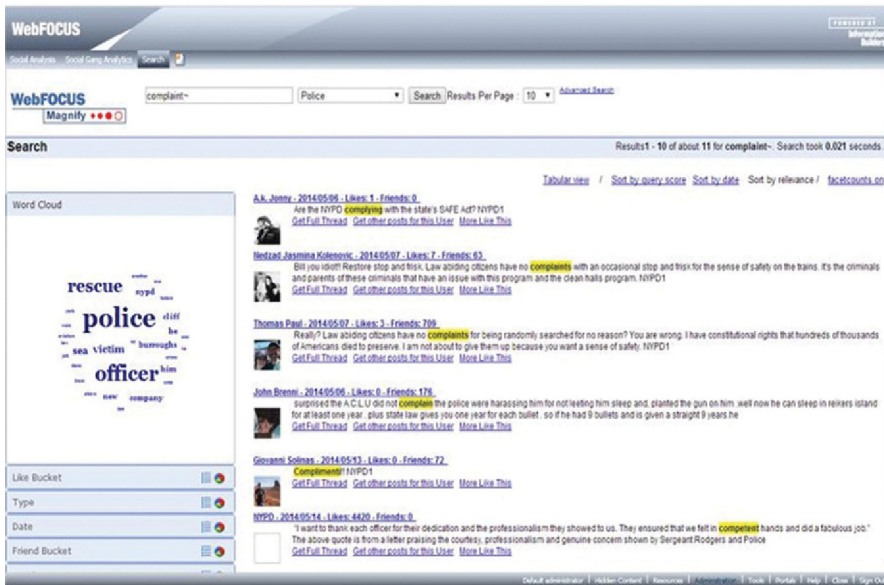


FIGURE 15.4

Capability to enable exploration and analysis of unstructured data to include social media content, as well as investigative notes. *Information Builders WebFOCUS, used with permission.*

with their Kenya Defence Forces counterpart. Ultimately shut down, it was interesting to follow and provided insight regarding their perception of events, as well as larger goals and objectives, which were woven throughout their frequent Tweets.

15.5 SOCIAL NETWORK ANALYSIS

A brief example of social network analysis was presented in Chapter 7; however, this type of analysis has advanced considerably from the days of using pencil and paper to construct a simple association matrix, which could be visually illustrated using a link chart. At its foundation, social network analysis is about mapping and measuring relationships between entities. The methods for executing this task, however, can be computationally intense. This is particularly true for relationships that are not obvious and/or those that are intentionally hidden.

Figure 15.5 illustrates sample Authority Miner[®] output,⁷ which identifies and prioritizes relationships and roles for individual offenders in a criminal network. Therefore, while the underlying computational science might be intense, the ability to visualize and interactively explore the results of social network

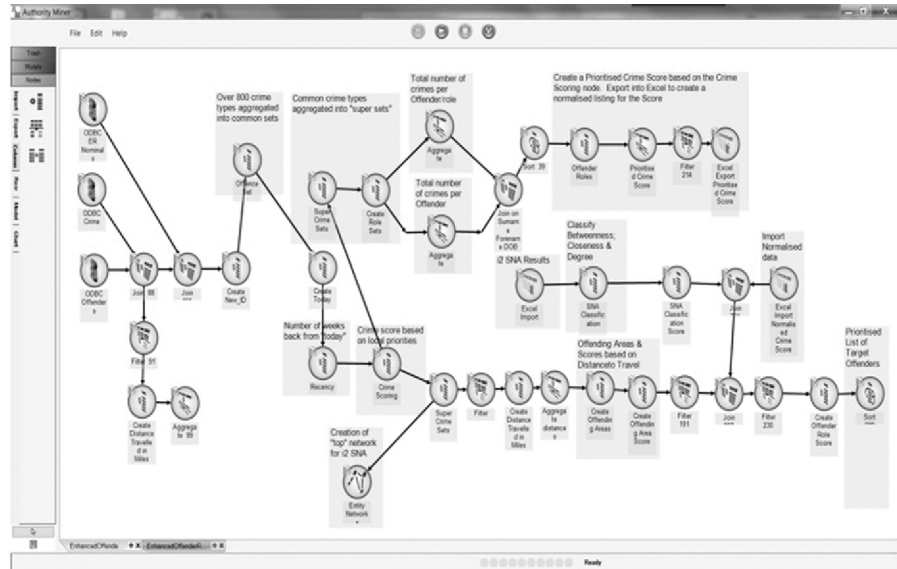


FIGURE 15.5

Sample Authority Miner® social network analysis (SNA) output, which identifies and prioritizes relationships and roles for individual offenders in a criminal network.²⁹

analysis can provide meaningful insight regarding relationships, affinities, and organizational centers of gravity that are operationally relevant and actionable.

15.6 FRAUD DETECTION

The topic of fraud detection is so large that entire textbooks, training programs, and even companies are devoted to it exclusively. In addition to the complexity associated with this pattern of offending, there are many different “flavors” of fraud to include phishing, spear phishing, breakout fraud, and ATM skimming just to name a few. Specific patterns of fraud and associated consequences also may differ based on the specific domain. For example, in addition to the financial consequences associated with medical identity theft, material changes to the medical record also can have significant consequences for the patient.

The basic methods for fraud detection generally are similar to those outlined throughout this textbook, but the domain expertise required is significant. Adding to the challenge, this pattern of offending tends to be very fluid and opportunistic, with the criminals exploiting vulnerabilities and weaknesses as soon as they are identified. Moreover, it is not unusual for the decision rules to be effectively telegraphed to the criminal community, particularly as relates to thresholds related to so-called “pay and chase” models where fraud detection

and related investigation occurs only after the payment has been made. In these particular cases, the cost associated with investigation and recovery must be balanced against the amount likely to be recovered. Therefore, audit or investigative thresholds are set; ideally, above the anticipated recovery amount to ensure that there is a return on the investigative investment. Losses falling below these thresholds frequently are disregarded and absorbed as the cost of doing business. Any specific insight regarding these thresholds, however, gives the criminals guidance regarding how high they can go before an audit is triggered and the crime is investigated. Specific information regarding thresholds that trigger fraud investigation would seem to represent very sensitive business intelligence. It is not unusual, though, for these decision rules to be openly discussed and debated, particularly when they involve fraud perpetrated against government programs. For example, the claim threshold for crop insurance fraud is \$200k per crop per entity per county, which automatically triggers an audit.⁸ With this actionable intelligence in hand, the criminals are able to set their fraud targets below the threshold in an effort to avoid an audit or investigation. In some cases, though, fraudsters also may use this information to perpetrate multiple smaller acts of distributed fraud that yield a higher combined return, while avoiding these thresholds.

Data mining and predictive analytics represent effective approaches to addressing this pattern of illegal behavior. Specifically, modeling algorithms that incorporate clustering techniques and anomaly detection can be used to identify patterns of behavior or activity that deviate from established patterns and trends. Data diverging or deviating from “normal” can be identified for further evaluation. On the other hand, rule induction models capitalize on the fact that people frequently are not creative or unique when they commit fraud. Although there are important individual differences in this type of criminal behavior, the secondary gain or desired goal generally structures the approach somewhat, which may limit the options for committing fraud. Therefore, rule induction models can be used to characterize and model known patterns of fraudulent behavior that can be applied to new data in an effort to quickly identify these patterns. Finally, the use of integrated approaches that utilize both scoring algorithms and unsupervised learning models can allow the analyst to exploit knowledge regarding previously identified or otherwise known or suspected patterns of criminal behavior, while remaining open to discovering unknown or unanticipated patterns of suspicious behavior. This combined approach of confirmation and discovery represents one of the more powerful aspects of data mining.

15.6.1 Identity Theft

Identity theft is not a new problem. In fact, my own grandfather assumed the identity of a deceased older sibling so that he could go to work in the central

Illinois coal mines before he was old enough. More recently, however, many consumers have had their financial lives ruined by thieves who assumed their identities in an effort to commit economic fraud.

This problem escalated in importance when the 9/11 attacks were investigated and it was determined that the hijackers were able to obtain false credentials necessary to move throughout our country with relative ease. Underscoring the breakdown in our ability to identify and prevent identity theft, 7 of the 19 terrorists involved in the attacks were able to obtain valid Virginia State ID cards, although they lived in Maryland hotels. Subsequent investigation uncovered a group of individuals who had provided hundreds of false ID cards to individuals with questionable documents, underscoring the new reality that the line between “general” crimes and terrorism or homeland security-related threats is increasingly blurred, particularly in the area of financial crimes or other economic crimes used to support terrorism.

Detection of identity theft frequently occurs only after something bad has happened, either fraud or some other misuse of a person’s identity. Proactive efforts involving manual searches of credit records and personal data in an effort to proactively identify cases of identity theft or misuse, however, are not only difficult but also very inefficient, given the extremely large amount of information involved. Searching public records for duplicate social security numbers or birth dates would be extremely laborious and inefficient using existing methods. Alternatively, automated searches of these databases using data mining and predictive analytics could not only flag invalid, suspicious records but also could create models associated with common patterns of identity theft or fraud. Additional information exploitation for illegal purposes, including the use of aliases and fraudulent addresses, could be identified with data mining tools. Ultimately, this approach would facilitate the development of proactive strategies in an effort to identify identity theft before serious consequences occur. While it is not likely that even automated methods will identify every case of identity theft, it might detect enough identity theft and fraud to make this type of illegal activity more difficult, deterring some illegal use of valid credentials in the future.

15.7 CYBER

The cyber threat has received considerable attention. Again, data breaches have become almost commonplace. At the time of this writing, I personally had two credit monitoring service subscriptions that were provided to me at no cost by retailers after my personally identifiable information was compromised in a data breach. One challenge associated with the rapid development and deployment of new technologies includes a concomitant increase in the risk. Again, there are the vulnerabilities that you know about, and those yet to be surfaced

and exploited by our adversaries. Moreover, with increasing ease of use and accessibility comes increasing comfort – perhaps, a false sense of security – wherein capabilities might be adopted outside of normal IT security process, without the appropriate protection and constraints. Peter ODell has created a list of potential vulnerabilities that should be of special concern to organizational leadership and their IT professionals:⁹

- Legacy systems – It is not unusual for organizations that were created years, or even decades ago that have not been upgraded in response to security threats that have been identified since their original development.
- Emerging technologies – Again, new technologies may bring undiscovered vulnerabilities. The number of “fixes” that appear on a regular basis underscore the fact that it frequently takes deployment into the operational setting for many vulnerabilities to be surfaced (frequently, by our adversaries).
- Personal devices – While most secure facilities restrict personal devices, other organizations do not have the same requirements. It is not unusual for employees to purchase and use their own devices in the workspace, particularly as technology may develop beyond their company-issued devices. These devices generally operate outside the control of the IT team and may introduce vulnerabilities. Again, a significant percentage of our critical infrastructure is privately held and controlled, leaving these facilities potentially vulnerable.
- Internal networks and facilities – Working within the walls of an organization’s IT system may convey a false sense of security. Portable devices, removable media, and other tools, however, can introduce malware and also be used to carry sensitive information out of the facility. Recent high profile examples of this particular vulnerability include the massive Wikileaks data breach, which involved the theft of classified documents through the use of a thumb drive,¹⁰ and the Stuxnet virus, which was believed to have been introduced into Iranian nuclear plant through an infected memory stick.¹¹
- The Internet – How many of us have picked up malware surfing the web? Virus protection software can be effective, but only for known vulnerabilities. Again, the new and emerging vulnerabilities frequently are the ones that hit hard and fast, and even protected systems may be vulnerable.
- E-mail – [Figure 15.6](#) illustrates an e-mail that I received, ostensibly from a colleague. The subject line “Important Document” and document name “Secure Document.html” apparently were used to create a sense of urgency and confidence in the note and its contents. The message, written in less than perfect English, encouraged me to download the document so that I could read the contents. “Nuff said...”

From: [redacted]
 Subject: Important Document
 Date: July 4, 2014 at 5:15 PM
 To: [redacted]

Please download attachment to see the Doc i uploaded Via Google Docs.
 Thanks. Happy Celebration!



Secure Document.html

FIGURE 15.6

Suspicious e-mail with attachment (screenshot taken by the author).

- Physical infrastructure – Access control systems and almost everything else within modern facilities are controlled by computers; systems that are only as secure as the supporting IT infrastructure. Again, the Stuxnet virus, which specifically targeted the programmable logic controllers on centrifuges in an Iranian nuclear plant,¹² was carried into a facility on a thumb drive.
- People – People represent the single greatest vulnerability. Everything from weak passwords to clever social engineering can compromise even the most secure system. Going back to the earliest known compromises,¹³ breakdowns can range from the overtly intentional like the Wikileaks compromise, to laziness and naïveté.

Again, as soon as we had the Internet, even before most of us were even using it regularly, there were efforts to exploit it. It is important to remember, though, that even the most sophisticated cyber-attack ultimately goes back to behavior, very bad behavior in some cases, but behavior nonetheless.

15.8 APPLICATION TO OTHER/ADJACENT FUNCTIONAL DOMAINS

While the emphasis throughout this book has been on public safety and security, other related fields where data mining and predictive analytics have shown promise are outlined here.

15.8.1 Syndromic Surveillance

As a behavioral scientist and avid follower of the television series, “Walking Dead” I am keenly aware of the shock to social order and civil society that would occur in the face of the next “Big One.” Whether pandemic illness, a large natural disaster, solar storms, or a zombie apocalypse, disruption in

power, loss of communication, and/or shortages of food or medicine would be expected to be associated with serious social disorder and lawlessness. By way of example, the chaos that followed Hurricane Katrina very effectively underscored the significant public safety and security challenges resulting from a critical incident that disrupted essential services.

With this in mind, syndromic surveillance systems have been developed specifically for the detection of disease outbreak and bioterrorism in support of early identification and informed intervention. By using pattern detection, as well as anomaly detection, these automated systems are able to identify clusters of symptoms or unanticipated changes in normal disease rates as measured by official reporting, purchase of over-the-counter medications, and even online searches for health-related information. There are many challenges associated with this type of monitoring, including similar signal-to-noise issues discussed previously that might be associated with an isolated or unevenly distributed outbreak occurring within a large association of monitored facilities.¹⁴ In addition, while tools like Google Flu Trends have shown significant promise in the near real-time detection of regional outbreaks of influenza, they do not differentiate between those who actually are sick and those merely searching for information regarding a particular illness;¹⁵ something that might occur during a well-publicized epidemic. By enhancing standard anomaly detection with decision rules, though, the performance of these screening algorithms can be increased.¹⁶

15.8.2 Military Applications and the Irregular Threat

Behavioral analysis and predictive analytics increasingly are being incorporated into military operations, particularly as relates to addressing the irregular or asymmetric threat. As outlined in Chapter 2, the influential paper, "Fixing Intel"¹⁷ frames a plan for the creative use of knowledge gathering, management, analysis, and dissemination. Moreover, examples like the behavioral analysis of the Lord's Resistance Army (LRA) outlined in Chapter 11, underscore foundation-level concepts regarding violent or other patterns of predatory behavior that transcend language, culture, and national boundaries, and the applicability of these concepts to other problem spaces. Similarly, the concept of human terrain analysis and other social-science-based approaches increasingly are being incorporated into risk and threat assessment, as well as operational planning and strategy.

Related to these parallels, several concepts outlined in the USMC Approach to Counterinsurgency Operations¹⁸ align well with the unique capabilities and approach of data mining and predictive analytics. For example, one of the opening comments includes the admonition that the "objective is not to crush but to influence ideas and wills,"¹⁹ which fits well with the goals of predictive analytics to anticipate and influence. Moreover, the authors call

for a segmentation of the problem space, advising that the adversary is not a homogenous group of actors. Rather, there are separate factions and embedded groups that require differential action and approach. Additional guidance regarding the complexity of the problem space includes a call to, “[u]nderstand the complex dynamics of the threat, including the wider environment [emphasis in original];”²⁰ effectively establishing the importance of context in analysis and interpretation of the results. Finally, in discussing the “Thinking Adversary,” the authors describe the dynamic relationship between the Marine forces and the adversary, noting that the “adversary will change his methods, operations, and strategy in order to stay ahead of friendly forces.”²¹ The response to this is a call for the active use of anticipatory intelligence to get within their decision cycle and establish a “superior tempo for adaptation”²² – concepts that align well with the model of proactive approaches and spiral development that are implicit in the dynamic use of predictive analytics to inform decision making and operations.

15.8.3 Private/Commercial Security

As was noted in Chapter 12, informal estimates suggest that 75–80% of our critical infrastructure is privately owned, which means that nongovernmental personnel frequently are responsible for the safety, security, and continuity of service. Moreover, the increased prevalence of workplace violence, active shooter incidents, and other high-profile events has increased the burden on the corporate security manager.

While they may not have access to classified or law enforcement sensitive information, other data sources, as well as the use of technical surveillance and other sensors to monitor facilities and track assets are relatively common. In addition, former operational personnel, including local, state, and federal law enforcement, and former military, frequently run corporate security programs.²³ These professionals bring their bring experience and related expectations regarding sources and methods, and recognize that reporting and compiling data are necessary but not sufficient to increasing security. As a result, many corporate security programs have expanded to include a layered approach that includes physical security and operations, with an overlay of intelligence and analysis. These intelligence and analysis functions are used to monitor, evaluate and characterize risk in support of fact-based approaches to security tactics, strategy, training, and policy. In this model, the intelligence and analysis capabilities are used to guide, and in many cases optimize the physical and operational security elements, while identifying emerging trends and patters. While we cannot mitigate the risk to zero, information-based approaches to prevention, thwarting, mitigation, and response can be used to change outcomes in the private sector, reducing losses while concomitantly increasing employee and customer safety.

15.8.4 Brand Integrity

While “brand integrity” may not immediately appear to be relevant to crime and intelligence analysis, contraband and counterfeit products represent economic crimes that also are linked directly to organized crime, terrorism finance, tax evasion, and consumer safety. Historically, increases in crime associated with increases in product costs are almost immediate. Within months, legal sales of consumer products plummet as theft, diversion, and sales of gray and black market, and counterfeit products increase.²⁴

Counterfeiting of consumer products has become so sophisticated that it frequently is difficult to distinguish counterfeit products from genuine brand. An example of this includes counterfeit cigarettes, which may require product content analysis to accurately identify counterfeit products.²⁵ In addition to lost revenue and other financial impacts, counterfeit products also represent a significant product safety and consumer protection issue. Not only do these products often fail to perform as expected, but they also may represent a significant threat to consumers given potentially toxic contents. Widely reported incidents associated with counterfeit pharmaceuticals, baby formula, and children’s toys underscore the potential threat associated with this very serious pattern of crime.

Gray market distribution of regulated products like alcohol and tobacco, in particular, may represent significant lost tax revenue, particularly when criminals are able to exploit local difference in tax rates or other fees associated with regulated products. For example, differential participation in the Master Settlement Agreement, as well as differences in tax rates can result in marked differences in retail prices on cigarettes, which makes an attractive environment for diversion and gray market sales.²⁶ Moreover, the proceeds of these crimes may be used to support other patterns of illegal activity. For example, Operation Smokescreen, which involved the legal purchase of cigarettes by a Hezbollah cell in Charlotte, North Carolina, a relatively low tax state, and illegal transportation to and sale in Michigan, where the taxes were significantly higher. Estimates suggest that the diversion ring made \$2 million in profit, at least some of which was used to fund Hezbollah.²⁷ Finally, as has been discussed previously, crime frequently begets more crime. The same illegal supply chain may support multiple commodities including drugs, guns, people, and bulk cash with links to money laundering and other patterns of illicit finance.

15.9 SUMMARY

The topics just listed represent a small sampling of the additional work that has been done using data mining and predictive analytics in the applied public safety and security setting. Work in this area is developing at a rapid pace,

which underscores its value to the law enforcement and intelligence domains, as well as to the analysts who benefit directly from the enhanced capacity that these approaches provide.

Bibliography

- 1 Hildebrandt W, McCue C. Unbiased analytics for the COCOMs. AHFE 2012.
- 2 Hildebrandt W, McCue C. Unbiased analytics for the COCOMs. AHFE 2012.
- 3 Hunt J, Romero P, Good J. Storytelling in interaction: agility in practice. In: Abrahamsson P, Marchesi M, Succi G, editors. *Extreme programming and agile processes in software engineering*. 7th International Conference, XP 2006/2006. p. 196.
- 4 Blasch E, Seetharaman G, Palaniappan K, Ling H, Chen G. Wide-area motion imagery (WAMI) exploitation tools for enhanced situation awareness. *Applied Imagery Pattern Recognition Workshop (AIPR) IEEE*; 2012, October.
- 5 Gettleman J. U.S. considers combating Somali militants' Twitter use. *The New York Times*, December 19. http://www.nytimes.com/2011/12/20/world/africa/us-considers-combating-shabab-militants-twitter-use.html?_r=3&; 2011; Oremus W. Twitter of terror: Somalia's al-Shabaab unveils a new social media strategy for militants. *Slate*, December 23. http://www.slate.com/articles/technology/technocracy/2011/12/al_shabaab_twitter_a_somali_militant_group_unveils_a_new_social_media_strategy_for_terrorists_.html; 2011.
- 6 Pearlman L. Tweeting to win: al-Shabaab's strategic use of microblogging. *The Yale Review of International Studies*, November, <http://yris.yira.org/essays/837>; 2012.
- 7 Authority Miner® screenshots provided to the author by Dr. Rick Adderly.
- 8 Taylor MZ. Brace for crop insurance audits. *The Progressive Farmer*, July 31. <http://www.dtn-progressivefarmer.com/dtnag/common/link.do?symbolicName=/ag/blogs/template1&blogHAndle=business&blogEntryId=8a82c0bc3865298c0138de7cb86f04ea>; 2012.
- 9 ODell P. Cyber 24/7: risks, leadership and sharing. Amazon CreateSpace, USA; 2014.
- 10 Sachtman N. Military bans disks, threatens courts-martial to stop new leaks. *Wired*, December 09. <http://www.wired.com/2010/12/military-bans-disks-threatens-courts-martials-to-stop-new-leaks/>; 2010.
- 11 Terdiman D. Stuxnet delivered to Iranian nuclear plant on thumb drive. *CNET*, April 12. <http://www.cnet.com/news/stuxnet-delivered-to-iranian-nuclear-plant-on-thumb-drive/>; 2012.
- 12 Kushner D. The real story of Stuxnet, *IEEE Spectrum*, February 26. <http://spectrum.ieee.org/telecom/security/the-real-story-of-stuxnet>; 2013.
- 13 Stoll C. *The cuckoo's egg*. New York: Pocket Books; 1989.
- 14 Reis BY, Mandl KD. Integrating syndromic surveillance data across multiple locations: effects on outbreak detection performance. *Proc AMIA Symp*; 2003: 549–553.
- 15 Carneiro HA, Mylonakis E. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases* 2009; 49(10): 1557–1564. <http://cid.oxfordjournals.org/content/49/10/1557.full>
- 16 Wong WK, Moore A, Cooper G, Wagner M. Rule-based anomaly pattern detection for detecting disease outbreaks. *AAAI* 2002.
- 17 Flynn MT, Pottinger M, Batchelor PD. Fixing Intel: a blueprint for making intelligence relevant in Afghanistan. 2009; January 5, 2010.
- 18 Marine Corps Combat Development Command. Tentative manual for countering irregular threats: an updated approach to counterinsurgency operations.
- 19 Marine Corps Combat Development Command 2006. p. 12.
- 20 Marine Corps Combat Development Command 2006. p. 15.
- 21 Marine Corps Combat Development Command 2006. p. 43.
- 22 Marine Corps Combat Development Command 2006. p. 43.
- 23 Javers E. *Broker, trader, lawyer, spy*. New York: Harper Business; 2011 .

- 24 Grabosky PN. Unintended consequences of crime prevention. In: Holmel R, editor. *The politics and Practice of Situational Crime Prevention Studies*, vol. 5. St Louis, MO: Willow Tree Press 1998; 25–56.
- 25 McCue C, Charchman RA, Zedler BK, McCue RJ. Family smoking prevention and tobacco control act: implications for the law enforcement community. *Food Drug Law Inst Update* 2009, November/December.
- 26 McCue C, Charchman RA, Zedler BK, McCue RJ. Family smoking prevention and tobacco control act: implications for the law enforcement community. *Food Drug Law Inst Update* 2009, November/December.
- 27 Levitt M. *Hezbollah: the global footprint of Lebanon's Party of God*. Washington, DC: Georgetown University Press; 2013.
- 28 Hildebrandt W, McCue C. Unbiased analytics for the COCOMs. AHFE 2012.
- 29 Authority Miner® screenshots provided to the author by Dr. Rick Adderly.

Future Trends

“Progress has not followed a straight ascending line, but a spiral with rhythms of progress and retrogression, of evolution and dissolution.”

Johann Wolfgang von Goethe

“[I]t is pretty easy to go far astray when trying to project the future of technology. In the cases of data mining and predictive analytics, however, the future is becoming reality at such a rapid pace that almost anything that I write will be outdated before the first copy of this book is purchased. Therefore, I will confine my comments to a few areas that I am particularly excited about, even if that “future” represents current reality. In many ways, that is one of the features that make this area of research and practice so exciting.”¹

Just as I “predicted” in the first edition of this text, several of the emerging technologies and “future trends” have become mainstream capabilities and the same cautions apply now. Text mining, fusion centers, and data warehouses – all relatively new capabilities a few years ago – have been extended and in some cases even replaced with tools like sentiment analysis, Big Data, and cloud computing. The following sections include an overview of new or emerging capabilities that I believe are particularly interesting and/or promising. This list is not meant to be all inclusive or comprehensive, particularly given the high probability that some of the most promising capabilities that we are likely to incorporate into the applied public safety and security setting over the next few years are still sitting on the work benches of data science “imagineers” as I write this chapter.

16.1 [REALLY] BIG DATA

Really massive data sources have been around for many years. Particularly as we look outside of our own domain, we can see other professionals, including those in marketing and financial services, who have been working to effectively manage and exploit very large datasets. Similarly, scientists working in astrophysics and high-energy physics have been developing more effective methods

of weak signal detection in extremely large data; capabilities with direct applicability and benefit to us.

While Big Data really are not new, the capabilities and associated tradecraft developed to effectively exploit these resources are relatively new and increasingly accessible to the data science community. Moreover, the formalization of concepts associated with big data, particularly as relates to the three “Vs” (volume, velocity, and variety)² can help us better manage and use this resource. Finally, considering “big data” as a unique space also offers new opportunities for transdisciplinary collaboration to identify potentially useful data science capabilities from other disciplines facing similar challenges in the management and meaningful exploitation of “big” data that can be applied to operational public safety and security analysis.

Looking ahead, the increased deployment of sensors will markedly grow both the amount and associated granularity of data in support of even deeper analysis. In particular, the increase in wearable devices like Google Glass and other persistent collection tools, including geoenabled devices like smartphones will contribute to a “people as sensors” model. This will generate massive amounts of location intelligence, including streaming video and other content that can be merged and integrated in support of activity-based intelligence (ABI) and pattern of life analysis. Extending this concept, Google cofounder and CEO Larry Page has envisioned the seamless integration of Google in the human brain, so that “[w]hen you think about something...you will automatically get information,”³ a model of “supreme artificial intelligence” that approaches the cyborgs of science fiction. While these sources are promising as relates to our understanding of the real complexity of behavior, the data collected will be truly massive, and the associated management and exploitation requirements will be staggering.

16.1.1 Biometrics

Biometrics to include fingerprints and DNA have become accepted standard practice in the operational public safety and security domain. DNA databases in particular represent a great source for the identification, as well as exclusion of criminal suspects.⁴ Ongoing development of additional biometric capabilities include face and voice recognition, as well as iris scans.

As technology has developed, biometrics are increasingly easy to collect and use for identification purposes. Many of us are familiar with, and may even have direct experience with the use of fingerprints for access control at secure facilities. Other benefits and uses, however, quickly come to mind. For example, the perpetration of identity theft would be significantly curtailed, because while I might be able to steal your credit card number, or even your personally identifiable information in an effort to assume your identity, I cannot easily

steal and use your fingerprints, iris scan or genetic profile. In response, insurance companies and financial services institutions increasingly are moving to the use of biometrics for identification purposes.⁵

As we consider the expanded use of biometrics, however, several additional opportunities and related questions emerge. For example, will there be a day when we totally abandon credit and ATM cards, and use biometric data exclusively for this purpose? Will saying “Colleen McCue” into a voice recognition device at the point of sale be sufficient to make a purchase? While it would ostensibly make theft more difficult, criminals could still perpetrate fraud in other ways. It also may be increasingly difficult to establish and/or assume an identity or legend in support of undercover or covert operations if biometric data and technology become prevalent.

Again, these uses of biometric data could be used to create some tremendously granular data in support of activity-based intelligence; however, what privacy considerations would need to be considered regarding access to and use of these data? Data privacy already is being actively debated as relates to the use of automated license plate recognition (ALPR) and toll collectors. However, these data sources are relatively crude and unreliable for establishing actual location for a specific individual. On the other hand, while I can loan you my car, I cannot loan you my fingerprint or DNA. Biometrics would represent a significant step in the ability to accurately identify and locate an individual. Similar to earlier debates regarding the routine collection and use of DNA for law enforcement purposes, these are all issues that are likely to be debated before any widespread public safety and security adoption and use.

16.2 ANALYSIS

Big data is not so much about big data as it is about an enhanced ability to extend and more effectively realize the promise of predictive analytics. Therefore, concomitant developments in analytics to include new tradecraft, technology, and even better approaches to processing will enable us to realize the promise of big data.

As we consider future trends and capabilities in analysis, it is important to revisit basic concepts regarding data mining and predictive analysis. While exciting new methods, technology, tradecraft, and even nomenclature (e.g., “data science”) have been developed, predictive analytics at its core still goes back to confirmation and discovery; *confirmation*, operationalization, and extension of what we know or think we know, and *discovery* of new trends, patterns, and relationships. Within this basic framework of characterization, confirmation and discovery, some analytic methods and capabilities merit additional attention given their promise for enhanced approaches to crime and intelligence analysis.

16.2.1 Geospatial

At the time of the first edition of this text, I had some familiarity with the use of GPS data to identify location. This knowledge was based on a specific case, and at the time that it was used to identify the location of a stalker it was still relatively unusual in local law enforcement, and represented a protected, law enforcement sensitive capability. Since this time, the use of GPS and other location-based capabilities to find people, including missing persons, represents a resource that is discussed openly and described in detail in the media. Underscoring the growth and development of geospatial capabilities, as well as the marked increase in the distribution and use of geoenabled mobile phones, it is not unusual for the public to call early and often for the analysis of mobile phone data to triangulate location in an abduction or other missing persons case; the same “sensitive” capability that was used just a few years earlier in the stalking case.

16.2.2 Activity-Based Intelligence

Building on the geospatial capabilities introduced and used throughout this text, activity-based intelligence is an emerging methodology introduced in response to the need to effectively leverage GEOINT into actionable intelligence, ultimately yielding “big value” from “big data”⁶ by effectively capturing and modeling pattern of life. By using advanced analytics to “identify patterns, trends, networks, and relationships hidden within large, data collections from multiple sources: full motion video, multispectral imagery, infrared, radar, foundation data, as well as SIGINT, HUMINT and MASINT information,”⁷ we can better anticipate and influence future behavior, particularly within a geospatial context.

16.2.3 Experts, Expert Systems, and the Power of the Crowd

If data mining and predictive analytics truly are game changing, why have they not been universally adopted? It would seem that increased public safety is something that everyone could get behind; however, there has been a lag in the acceptance of automated tools in some areas. Research from the political science community may provide an answer to this apparent disconnect between science and practice. It seems that people are more inclined to trust an “expert” despite the finding that the accuracy of “expert” predictions does not differ from those of mere mortals, both of which perform well below predictions derived using statistics and mathematical modeling.⁸

I have direct personal experience with this phenomenon. Several years ago, I attended a scientific meeting that included a lively debate over expert opinion versus statistical estimates of risk for future violence. Despite the fact that the data overwhelmingly supported the accuracy and reliability of the statistical estimates, the attendees found a number of exceptions that would have been

missed by computer models and ultimately elected to stay with the human judgments. One possible explanation for this is that people may find comfort in the authority that an “expert” conveys, rather than believing that human nature can be reduced to math and equations.⁹ Given the capacity that data mining and predictive analysis can bring to support public safety and security, however, this disconnect between science and practice really needs to be addressed. Perhaps the best model for the paradigm shift required lies somewhere in between those two extreme positions and could include *domain* experts using the expert systems embodied in data mining and predictive analysis software.

16.2.3.1 Consensus Opinions

Although the Defense Advanced Research Projects Agency (DARPA) FutureMAP program was cancelled due to public outrage over government-sponsored wagering on future terrorist attacks and assassinations, consensus opinions have been used with some success. In a unique application of Bayes’ theorem, naval scientist John Craven used consensus expert opinions to locate the US nuclear submarine *Scorpion*.¹⁰ Bayesian inference is particularly appealing for applied public safety and security analysis because it supports the incorporation of tacit knowledge and domain expertise from experts representing diverse backgrounds, potentially bringing the “best of all worlds” to the analytical process.

16.2.3.2 Crowd Sourcing

Again, Ushahidi, TomNod, and even astronomy efforts like the Andromeda Project¹¹ and Galaxy Zoo,¹² are leveraging the power of the crowd to address big data processing tasks. Similarly, the DARPA Red Balloon Challenge effectively used crowd sourcing to solve a “distributed, time-critical, geolocation problem”¹³ with implications for public safety and security, as well as search and rescue. Not only can these capabilities be used to break up tasks, but as TomNod has demonstrated, there are certain tasks that humans are uniquely suited to perform. Using so-called, “artificial intelligence” TomNod has been able to effectively leverage the crowd in support of search and rescue, feature extraction, and a number of other challenging geospatial tasks.

16.3 OTHER USES

Again, looking outside the specific public safety and security domain, innovation in other professional disciplines frequently can be leveraged to support crime and intelligence analysis techniques. For example, with access to incredibly complex data resources, Google has developed a business model that enables them to infer age, gender, and interest based on online behavior in support of increasingly targeted advertisements.¹⁴ In keeping with the concept of ABI or pattern of life analysis, location intelligence also is being

increasingly incorporated into these models.¹⁵ The ability to effectively characterize the “when, where, and what” of consumer behavior supports optimization, including microtargeting; something that is also being used in political campaigns¹⁶ as staff develop approaches to anticipation and influence. All of these capabilities have direct and obvious implications for crime and intelligence analysis

Unfortunately, like many other tools, these capabilities also can be used against us. For example, if our adversaries can obtain sufficient data either through direct collection or theft (e.g., stolen marketing databases), can they also use analytics to infer identity, preferences, and/or pattern of life in support of increasingly complex attacks? Will ABI be used to support, augment, or even replace hostile surveillance? While the credit card companies are able to generate new credit cards and provide credit monitoring services in response to data theft, it is not quite clear how we will address the theft or compromise of more complex data sources, particularly those relating to pattern of life.

16.4 TECHNOLOGY AND TOOLS

In addition to data and analysis, new technology and tools, as well as more efficient analytics service delivery models will greatly increase both the capacity and capability of crime and intelligence analysis.

16.4.1 Domain-Specific Tools

The emergence of tools designed specifically for public safety and security analysis continues to parallel general advancements in data science. While many of the examples included in this book were generated using technology and tradecraft that were developed originally for some other domain, the development of applications designed specifically for public safety and security analysis makes them even more accessible to the crime and intelligence analyst. Similarly, advances in the visual depiction of complex analytical output continue to be the focus of research and development. Analytical output that directly addresses the “I’ll know it when I see it” metric and build on the end user’s tacit knowledge in support of direct transfer to and use in the applied setting continue to represent a powerful trend in the industry.

16.4.2 Processing

New capabilities like the SAS High Performance analytics and SAP HANA leverage in-memory processing in support of real-time or near-real-time analysis of the really big data that we are increasingly encountering – effectively enabling “analysis at the speed of thought.”¹⁷ As illustrated by the “Food Truck” example in Chapter 14, the ability to provide real-time feedback regarding suspicious activity will significantly inform and enhance surveillance

detection operations, while also providing meaningful insight and situational awareness directly to forward deployed operational personnel. Similarly, the use of SAS HP analytics brings the promise of real-time analysis of transactional data in support of truly effective approaches to fraud detection and prevention, including real-time scoring of transactional data, thereby eliminating many of the serious inefficiencies associated with the traditional “pay and chase” model.

16.4.3 IBM Watson

Since its public debut on Jeopardy, Watson has evolved from a very exciting science project into something that is being used to solve real problems in a variety of different and disparate domains. Of particular interest to the operational public safety and security community is Watson’s agility with transactional data, as well as its powerful capabilities to infer associations and relationships. By using Semantic Analysis Technology (SAT), Watson is able to identify subtle and nuanced associations in data, including patterns of behavior suggestive of financial crimes and fraud.¹⁸ Similarly, Watson is able to leverage life-event detection and psycholinguistic tools in support of more comprehensive and nuanced pattern of life analysis, including inference regarding which social media accounts might be yours.¹⁹ Again, the applicability of these capabilities to operational crime and intelligence analysis is exciting.

16.4.4 Managed Service, Software as a Service, and Cloud Computing

Managed service delivery models, Software as a Service (SaaS), and cloud computing minimize, if not completely eliminate the “burden of ownership” associated with purchasing, installing, and maintaining sophisticated analytic software locally. Moreover, leverage of the analytic fusion center model also offers the promise of optimized analytic resources, including skilled personnel, as well as the benefits associated with data and operations that are both vertically and horizontally integrated.²⁰

16.5 POTENTIAL CHALLENGES AND CONSTRAINTS

In addition to new data sources, technology, and tradecraft, a number of potential challenges and constraints have emerged including transnational crime, and concerns regarding the protection of privacy, civil rights, and civil liberties. In addition, the rapid proliferation of analytic capabilities also creates new challenges for the crime and intelligence analyst as they work to acquire technical proficiency with specific sources and methods, while remaining open to the “art of the possible” as embodied in new data, technology, and tradecraft. Again, any potential challenge also represents a unique opportunity to succeed

and I look forward to the development of meaningful solutions that will further enhance, inform, and strengthen crime and intelligence analysis going forward.

16.5.1 Globalization of Crime

The increasingly global nature of our world has created numerous opportunities as relates to fluid, continuous pathways for commerce, speedy response to manmade and natural disasters, and a larger sense of community that transcends national boundaries. Unfortunately, this increased interconnectedness has concomitantly cultivated and grown the transnational nature of crime, which will represent a significant challenge going forward. We know that criminals in the United States frequently exploit jurisdictional boundaries and differential enforcement as a means by which to enable their criminal activity, particularly organized crime. Similarly, national boundaries, differential enforcement, and even differences in language, culture, and the rule of law create an attractive environment that can be exploited by criminals. Further complicating this challenge is the role that international crime, terrorism, and violent extremism play in the creation of ungoverned and under-governed spaces that support or otherwise enable crime by limiting the ability to effectively and consistently enforce the rule of law.²¹

Again, location matters and the important role that location plays in crime has been highlighted by example throughout this text. Transnational crime not only exacerbates but also perpetuates the challenge of ungoverned or under-governed space by creating veritable “enforcement-free zones” for all manner of crime including illegal supply chains that move drugs, guns, people, bulk cash, and other natural resources, and the violence used to enforce the rules, norms, and boundaries associated with these locations, further threatening vulnerable populations and fragile states. Solutions will not be simple or easy; likely requiring an unprecedented level of global collaboration and cooperation in support of meaningful and long lasting approaches.²²

16.5.2 Let the Process Guide the Solution

“If all you have is a hammer, everything looks like a nail.”

Abraham Maslow

The rapid proliferation of crime and intelligence analysis tools has been both a blessing and a curse. A blessing in their ability to effectively translate high-quality advanced analytics for use in the operational public safety and security environment. A curse because they have created a growing population of crime and intelligence analysis “technicians,” or individuals with deep expertise

regarding a specific source or method but limited knowledge regarding analysis as a process and the importance of context in the interpretation and effective use of the results. While this may provide immediate, short-term benefit to the organization, it also threatens to limit the ability to effectively respond to new or emerging threats, as well as the ability to successfully adopt new sources and/or methods as they become available.

It is not unusual for an analyst to become exceptionally proficient in a specific technology, method, tool, or tradecraft to the exclusion of others. In this situation, however, if all I have is the equivalent of an analytic hammer, then every question begins to assume the form of a nail (or is forced to fit that model). This may occur out of comfort and complacency on the part of the analyst. On the other hand, organizational commitment to a specific tool or platform, particularly if it is perceived as being “cutting edge” and/or was expensive, can similarly constrain access. Unfortunately, this situation does not enable the analyst to effectively adapt and respond to the exceptionally diverse and evolving nature of our problem space. The better approach would be to train the equivalent of analytic master carpenters who would be able to select the tool most appropriate for the question. Ideally, these analysts also would be uniquely suited to quickly embrace and use new tools as they become available, and may even be involved in the development of next-generation tools; either through transdisciplinary adoption and use of existing capabilities from other domains, or even *de novo* development of novel capabilities that would benefit our community.

As discussed in Chapter 4, “wicked” problems also create a unique hazard for the analyst in that the favored solution tends to drive the definition and characterization of the problem, which may result in analysis confounded by “circular logic.”²³ Moreover, there are no perfect solutions, no “free lunches” in analysis.²⁴ Ultimately, it is important to remember that these tools support the data mining *process*, and while they might be necessary for analysis, they certainly are not sufficient for the insight required to support meaningful and effective anticipation and influence in the operational public safety and security environment. Therefore, the primary objective for most technology solutions is that it will optimize human time and attention by surfacing interesting patterns, trends, and relationships. Again, it is the domain expertise and ability to create operationally relevant and actionable output that is the priceless element in the applied public safety and security analytical process. As noted in the opening to this chapter, new data sources, technology, and tradecraft are being developed daily. The well-trained analyst will be able to seamlessly adopt these new capabilities as they are made available, as well as those that have not even been invented yet; assuming a fluid approach to technology and allowing the questions to guide the analytic approach, thereby letting the problem guide the solution.

16.5.3 Privacy and Civil Liberties

The challenge of applying eighteenth century legal statutes and concepts to twenty-first century technology has created tension for policy makers and analysts alike. Unfortunately, common misperceptions regarding data mining and predictive analytics, and related concerns regarding privacy and civil liberties, as well as uncertainty regarding the overall value of this approach in operational public safety and security threaten to severely curtail the use of advanced analytics for crime and intelligence analysis.

Other new capabilities have encountered resistance when they were first introduced, including the use of DNA evidence and criminal investigative analysis, or the behavioral analysis of violent crime. In these cases, the community was well served by individuals and professional organizations that were not only willing to go out and advocate, but also document successes in support of these capabilities. These groups established standard practice and related credentialing requirements for expert testimony, while concomitantly educating the legal community, juries, their peers, and the public. Underscoring the success of these efforts, juries have come to expect the introduction of DNA evidence at trial, regardless of the crime being tried, as a result of the so-called "CSI Effect." Perhaps we will see a similar embrace on the part of operational security analytics.

Therefore, as we increasingly use data mining and predictive analysis to anticipate, predict, and prevent crime, we must be sensitive to the concerns regarding the protection of privacy and civil liberties. Some of the specific issues are described here.

16.5.3.1 Data

In response to several high-profile data breaches, as well as revelations regarding government collection and use of data, the public is becoming increasingly savvy about their data. Areas of new or increased concern include personally identifiable information (PII), sensitive financial information, social media,²⁵ and communications "metadata," as well as location data.²⁶ On the other hand, individuals increasingly view their data from a transactional perspective, expressing a willingness in both word and action to exchange their information for something of value.²⁷ In other words, they are willing to give up some privacy in exchange for a benefit. Whether it is a financial incentive or information regarding restaurants within their vicinity based on geolocation data, they frequently are willing to barter their data in exchange for desired information, goods, or services.

Automatic license plate recognition and automated toll collection systems provide a good example of the complexity of these issues, particularly as relates to automated collection capabilities. Even recently, it was not unusual for police

departments and other agencies to collect the license plate information in an effort to identify specific vehicles associated with a particular event or location. While this was not incontrovertible evidence, it was possible to use this information to make some inferences regarding specific individuals. For example, there is no expectation of privacy associated with a vehicle in a public location and it was not unusual to collect license plate numbers from organized crime or gang funerals in an effort to identify putative membership and better understand these networks. ALPR systems and other sensors, however, can automate this process, significantly increasing the amount of data collected, as well as concomitantly increasing the speed and relative complexity of the associated analysis.

Similarly, is there a difference between seeing an individual in a particular public location or using GPS-enabled capabilities like those embedded in a smartphone or other device to infer location? And does this matter if it is a vendor or other commercial entity that is collecting this information for market analysis purposes, or a law enforcement or security organization supporting public-safety-related analysis or an active investigation? In response to these concerns, there has been increasing public debate, as well as proposed legislation²⁸ that will curtail or otherwise restrict access to information. Again, law enforcement agencies had been collecting these data for years; however, the introduction of automated systems has prompted discussion of what is appropriate and what may require additional legal review and permission, particularly as relates to persistent collection.

16.5.3.2 Analysis

Recently, data scientists in a number of disparate domains are learning painful lessons regarding the power of advanced analytics and that just because you can do something does not necessarily mean that you should. While it appeared that most attention in previous years was directed at public safety and security use of these tools, more recently marketing and retail use of predictive analytics have received considerable negative attention. This has been particularly true of derived products and the use of advanced analytics for decision support. Perhaps one of the most high profile examples came from Target,²⁹ a retailer that developed an algorithm to infer the early stages of pregnancy based on purchasing behavior. They then used this information to create personalized packages of coupons and related product offerings in an effort to gain market share in the lucrative pregnancy and childcare market. This practice came to light after the retailer mailed a brochure with pregnancy and baby-related coupons to a high school student, which prompted an uncomfortable discussion in her home and subsequent public outrage over Target's use of advanced analytics when the story came to light.

This example has surfaced a number of issues regarding the legal, ethical, and even practical boundaries related the use of predictive analytics as the public

increasingly views some of the results as inappropriate transgressions of privacy. Many of the comments from the data science community suggested that there is nothing inherently “bad” about what Target did, noting that it was just math. Rather, it was how the results were presented and used that created the issue. Again, all of the data used in the Target analysis had been collected legally; however, the subsequent analysis created a derived product that was significantly more sensitive than the original sources. Moreover, analysis, even if based entirely on open source data, that reaches into private family matters including marriage and family planning crosses important social boundaries; transgressions that may have the unintended consequence of repelling possible customers rather than expanding market share. Therefore, at least one of the lessons learned in response to the Target case has been that just because you can do something does not necessarily mean that you should, and related but more subtle considerations regarding operational use, and associated messaging of analytic results and context.

We have a similar challenge in operational security analytics that emerged in 2003 when the Data Mining Moratorium Act was introduced, which proposed that the use of advanced analytics in the applied public safety and national security domains should be significantly regulated or even curtailed. As with the marketing example, the relevant data sources and analytic techniques were legally and ethically available. Rather, it was the use of the results that created concern. Therefore, the related take home message for the crime and intelligence analyst is to think about what you are doing – particularly as relates to the creation of derived products. These are sensitive for a number of reasons. Second, these are increasingly important capabilities. Misuse, whether intentional or otherwise, creates challenges for all of us and may limit our ability to use some of the most powerful tools available to us if they are not used properly, and with sensitivity and respect.

16.5.3.3 “Prediction” and the Minority Report Concern

Additional public concern relates to the use of data mining and predictive analytics to guide action, particularly as it relates to taking action based on statistical “predictions” or what *might* happen versus addressing actions already committed (e.g., the “Minority Report” concern).³⁰ Most of the examples outlined in this text rely on the use of analysis to identify general patterns and locations of future risk based on behavior; however, data mining and predictive analytics also are being used to develop models designed to identify specific individuals who may merit additional scrutiny (e.g., behavioral screening at airports), as well as individual risk assessment models (e.g., recidivism). Again, people were being investigated, questioned, and detained based on what they might do, sometimes wrongly, rather than what they have done before data mining and predictive analytics came into use.³¹ This issue is not particularly new or

unique to data mining and predictive analytics. Rather, it goes back to responsible decision making in the operational law enforcement and public safety community; something that should be respected regardless of the source of the opinion or related analysis.

16.5.4 First Do No Harm...

“Rather than thinking about exceptional moral rules for exceptional moral situations we should almost always see exceptional moral situations as opportunities for us to show exceptionally-deep commitment to our deepest moral values.”³²

Again, the unique circumstances of crisis and conflict mapping merit special consideration. As has been noted throughout the text, adding location not only increases the potential value of the data but also may concomitantly increase the sensitivity of the data and related derived products. In crisis and conflict mapping in particular, this “value add” may pose harm to already vulnerable populations. This is an issue that the community is working through currently;³³ however, as discussed with the Target example, just because you can do something does not always mean that you should. The analyst should always maintain awareness of the situation and larger context associated with their efforts, and consider the broader implications and use of their work.

While these very effective tools hold great promise, they may not be available to us for long given concerns about privacy and civil liberties. It is our responsibility, therefore, to use these tools ethically and responsibly, adhering to the spirit as well as the letter of the law. Adoption and use of advanced analytics in the operational public safety and security setting represents a major paradigm shift for the community. It is important that we clearly recognize what these tools can and cannot accomplish, though. There are no crystal balls, no “Minority Report.” At their foundation, this is all just math. These tools are a means to an end, with the major objective being enhanced public safety and security. With that in mind, we need to understand and respect the issues and related concerns. In addition, we need to educate ourselves and the operational end users, as well as the public regarding the value and limits of these tools, and the important protections in place, to ensure truly informed debate and appropriate use within the applied setting. Going forward, please be good stewards of these capabilities and make an effort to create “informed consumers” among your clients, end users, and the general public, and remember the following:

- Reporting, collecting, and compiling data are necessary, but not sufficient to increasing public safety.
- Advanced analytics are used in almost every segment of society to improve service delivery and optimize resources.

- Operational security analytics support the meaningful exploitation of public safety and security data necessary to information-based anticipation and influence, including prevention, response, and consequence management.
- Used responsibly, operational security analytics can enhance public safety, prevent crime and change outcomes.

IS DATA MINING EVIL?

Further confounding the question of whether to acquire data mining technology is the heated debate regarding not only its value in the public safety community but also whether data mining reflects an ethical, or even legal, approach to the analysis of crime and intelligence data. The discipline of data mining came under fire in the Data Mining Moratorium Act of 2003.

Unfortunately, much of the debate that followed has been based on misinformation and a lack of knowledge regarding these very important tools. Like many of the devices used in public safety, data mining and predictive analytics can confer great benefit and enhanced public safety through their judicious deployment and use. Similarly, these same assets also can be misused or employed for unethical or illegal purposes.

One of the harshest criticisms has addressed important privacy issues. It has been suggested that data mining tools threaten to invade the privacy of unknowing citizens and unfairly target them for invasive investigative procedures that are associated with a high risk of false allegations and unethical labeling of certain groups. The concern regarding an individual's right to privacy versus the need to enhance public safety represents a long-standing tension within the law enforcement and intelligence communities that is not unique to data mining. In fact, this concern is misplaced in many ways because data mining in and of itself has a limited ability, if any, to compromise privacy. Privacy is maintained through restricting access to data and information. Data mining and predictive analytics merely analyze the data that are made available; they may be extremely powerful tools, but they are tools nonetheless. With data mining, ensuring privacy should be no different than with any other technique or analytical approach.

Unfortunately, many of these fears were based on a misunderstanding of the Total Information Awareness system (TIA, later changed to the Terrorism Information Awareness system), which promised to combine and integrate wide-ranging data and information systems from both the public and private sectors in an effort to identify possible terrorists. Originally developed by DARPA, this program was ultimately dismantled, due at least in part to the public outcry and concern regarding potential abuses of private information. Subsequent review of the program, however, determined that its main shortcoming was related the failure to conduct a privacy impact study in an effort to ensure the maintenance of individual privacy; this is something that organizations considering these approaches should include in their deployment strategies and use of data-mining tools.

On the other hand, some have suggested that incorporation of data mining and predictive analytics might result in a waste of resources. This underscores a lack of information regarding these analytical tools. Blindly deploying resources based on gut feelings, public pressure, historical precedent, or some other vague notion of crime prevention represents a true waste of resources. One of the greatest potential strengths of data mining is that it gives public safety organizations the ability to allocate increasingly scarce law enforcement and intelligence resources in a more efficient manner while accommodating a concomitant explosion in the available information –

the so-called “volume challenge” that has been cited repeatedly during investigations into law enforcement and intelligence failures associated with 9/11. Data mining and predictive analytics give law enforcement and intelligence professionals the ability to put more evidence-based input into operational decisions and the deployment of scarce resources, thereby limiting the potential waste of resources in a way not available previously.

Regarding the suggestion that data mining has been associated with false leads and law enforcement mistakes, it is important to note that these errors happen already, without data mining. This is why there are so many checks and balances in the system – to protect the innocent. We do not need data mining or technology to make errors; we have been able to do that without the assistance of technology for many years. There is no reason to believe that these same checks and balances would not continue to protect the innocent were data mining to be used extensively. On the other hand, basing our activities on real evidence can only increase the likelihood that we will correctly identify the bad guys while helping to protect the innocent by casting a more targeted net. Like the difference between a shotgun and a laser-sited 9 mm, there is always the possibility of an error, but there is much less collateral damage with the more accurate weapon.

Again, the real issue in the debate comes back to privacy concerns. People do not like law enforcement knowing their business, which is a very reasonable concern, particularly when viewed in light of past abuses. Unfortunately, this attitude confuses process with input issues and places the blame on the tool rather than on the data resources tapped. Data mining can only be used on the data that are made available to it. Data mining is not a vast repository designed to maintain extensive files containing both public and private records on each and every American, as has been suggested by some. It is an analytical tool. If people are concerned about privacy issues, then they should focus on the availability of and access to sensitive data resources, not the analytical tools. Banning an analytical tool because of fear that it will be misused is similar to banning pocket calculators because some people use them to cheat on their taxes.

As with any powerful weapon used in the war on terrorism, the war on drugs, or the war on crime, safety starts with informed public safety consumers and well-trained personnel. As is emphasized throughout this text, domain expertise frequently is the most important component of a well-informed, professional program of data mining and predictive analytics. As such, it should be seen as an essential responsibility of each agency to ensure active participation on the part of those in the know; those professionals from within each organization that know where the data came from and how it will be used.

Unfortunately, serious misinformation regarding this very important tool might limit or somehow curtail its future use when we most need it in our fight against terrorism. As such, it is incumbent upon each organization to ensure absolute integrity and an informed decision-making process regarding the use of these tools and their output in an effort to ensure their ongoing availability and access for public safety applications.

16.6 CLOSING THOUGHTS

“Information analysis is the brain of homeland security. Used well, it can guide strategic, timely moves throughout our country and around the world. Done poorly, even armies of guards and analysts will be useless.” Markle Foundation’s Task Force on National Security in the Information Age³⁴

This statement continues to be true and emphasizes the critical importance of sound crime and intelligence analysis. While new technology, tools, and tradecraft can enhance our ability to effectively ask and answer the “hard questions,” we still need to do more than collect data. As analysts, we need to analyze it in a way that yields the meaningful insight that will translate directly into information-based decisions and operational support.

In closing, “[p]redictive analytics...does not provide guarantees. Instead, it is all about increasing the likelihood that a desired outcome will occur – at the right time, the first time. These concepts of increased likelihood and timeliness are what make applying it to decision making so enticing.”³⁵ These are very worthy, yet attainable goals for operational public safety and security analysis. With that objective in mind, I wish you well, and encourage you to go forward and do good.

Bibliography

- 1 McCue C. Data mining and predictive analysis: intelligence gathering and crime analysis. Burlington, MA: Butterworth-Heinemann (Elsevier); 2007. p. 315.
- 2 Dumbill E. What is big data? O’Reilly Radar. <http://strata.oreilly.com/2012/01/what-is-big-data.html>; 2012 [accessed 11.01.2012]
- 3 Farber D. At 15, Google’s ambitions remain unbridled. CNET, September 27. <http://www.cnet.com/news/at-15-googles-ambitions-remain-unbridled/>; 2013.
- 4 McCue C, Smith GL, Diehl RL, Dabbs DE, McDonough JJ, Ferrara PB. Why DNA databases should include all felons. *Police Chief* 2001; 68: 94–100.
- 5 Collier K. TELCOs and other businesses are preparing to store their biometric data such as voice fingerprint records to identify customers. *Herald Sun*, March 24. <http://m.heraldsun.com.au/news/victoria/shops-and-telcos-collecting-fingerprints-voice-records-of-customers/story-fni0fit3-1226863673700?sv=56d3cd522b852c4520ade640b93eba71&nk=9f770a49e74bac80d5ca091371d1f78d>; 2014.
- 6 Long L. Remarks as prepared, Letitia A. Long, Director, National Geospatial-Intelligence Agency, SPIE 2013 Defense, Security + Sensing Symposium, May 01. <https://www1.nga.mil/MEDIAROOM/SPEECHESREMARKS/Pages/SPIEDSSSymposium.aspx>; 2013.
- 7 Long L. Remarks as prepared, Letitia A. Long, Director, National Geospatial-Intelligence Agency, SPIE 2013 Defense, Security + Sensing Symposium, May 01. <https://www1.nga.mil/MEDIAROOM/SPEECHESREMARKS/Pages/SPIEDSSSymposium.aspx>; 2013.
- 8 Colvin G. (2006). Ditch the ‘experts.’ *Fortune* 2006; February 6, p. 44.
- 9 Ibid.
- 10 Sontag S, Drew C, Drew A. *Blind man’s bluff: the untold story of American submarine espionage*. New York: HarperCollins; 1999.
- 11 Kelley P. Crowdsourcing the cosmos: astronomers welcome all to identify star clusters in Andromeda galaxy. University of Washington, December 04. <http://www.washington.edu/news/2012/12/04/crowdsourcing-the-cosmos-astronomers-welcome-all-to-identify-star-clusters-in-andromeda-galaxy/>; 2012.
- 12 Adams T. Galaxy Zoo and the new dawn of citizen science. *The Guardian*, March 17. <http://www.theguardian.com/science/2012/mar/18/galaxy-zoo-crowdsourcing-citizen-scientists>; 2012.
- 13 Tang JC, Cebrian M, Giacobe NA, Kim H-W, Wickert D. Reflecting on the DARPA Red Balloon Challenge. *Commun ACM* 2011; 54(4): 78–85.

- 14 An interesting parlor game, you can see your Google profile and better understand how it sets the ads that you see. In my case, Google has determined my age correctly, believes that I am a male, and have an interest in astronomy (Read M. How old does Google think you are? Gawker.com, January 27. <http://gawker.com/5879895/how-old-does-google-think-you-are/>; 2010; Stanley C. Find out how old Google thinks you are (among other things). Flavorwire, January 27. <http://flavorwire.com/253616/find-out-how-old-google-thinks-you-are/>; 2012).
- 15 Brustein J. If your phone knows which aisle you're in, will it have deals on groceries? Bloomberg Businessweek, January 06. <http://www.businessweek.com/articles/2014-01-06/apples-ibeacon-helps-marketer-beam-ads-to-grocery-shoppers-phones>; 2014.
- 16 Issenberg S. How President Obama's campaign used big data to rally individual voters. MIT Technol Rev. <http://www.technologyreview.com/featuredstory/509026/how-obamas-team-used-big-data-to-rally-voters/>; 2012 [accessed 19.12.2012].
- 17 Few S. Data analysis at the speed of thought. InformationWeek, 01 March, http://www.informationweek.com/software/information-management/data-analysis-at-the-speed-of-thought/d/d-id/1030748?page_number=2; 2005.
- 18 Cohan P. How you can profit from Watson's 'Jeopardy' win, Daily Finance, 30 March. <http://www.dailyfinance.com/2011/03/30/how-you-can-profit-from-watsons-jeopardy-win/>; 2011.
- 19 Brownlee J. IBM's next big thing: psychic Twitter bots. Fast Company, March 03. <http://www.fastcodesign.com/3025738/ibms-next-big-thing-psycho-twitter-bots>; 2014.
- 20 McCue C, Miller L, Lambert S. The Northern Virginia military shooting series: operational validation of geospatial predictive analytics. Police Chief, February. http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display&article_id=2871&issue_id=22013; 2013.
- 21 Miles D. Ham: Africa presents opportunity, challenges. U.S. Department of Defense. June 19. <http://www.defense.gov/news/newsarticle.aspx?id=116802>; 2012.
- 22 Miklaucic M, Brewer J. Convergence: Illicit networks and national security in the age of globalization. Washington, DC: National Defense University Press; 2013.
- 23 Rittel H., Webber M. Dilemmas in a General Theory of Planning, Policy Sciences 1973, 4, p. 166.
- 24 Braun ML. Data analysis: the hard parts. Marginally Interesting: Machine Learning, Computer Science, Jazz, and All That. <http://blog.mikiobraun.de/2014/02/data-analysis-hard-parts.html>; 2014 [accessed 17.02.2014].
- 25 Omand D, Bartlett J, Miller C. "A balance between security and privacy online must be struck...." DEMOS. http://www.demos.co.uk/files/_Intelligence_-_web.pdf?1335197327; 2012.
- 26 The Centre for Spatial Law and Policy, legal and policy issues associated with geospatial data and technology (<http://spatiallaw.com>).
- 27 Morris T. 3 ways to personalize the customer experience without getting too personal, parature, June 10, <http://www.parature.com/personalize-cx/>; 2013; Cisco. Cisco customer experience research: automotive industry global data. http://www.cisco.com/web/about/ac79/docs/ccer_report_manufacturing.pdf; 2013.
- 28 The Geolocation Privacy and Surveillance (GPS) Act, and related geolocation privacy legislation (<http://www.gps.gov/policy/legislation/gps-act/>).
- 29 Duhigg C. How companies learn your secrets. The New York Times. <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&r=0>; 2012 [accessed 16.02.2012]; also, the talk that triggered the controversy: Pole A. How Target gets the most out of its guest data. Predictive Analytics World, <http://rmportal.performedia.com/node/1373>; 2010.
- 30 Stroud M. The minority report: Chicago's new police computer predicts crimes, but is it racist? The Verge, February 19. <http://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist>; 2014.
- 31 For review of the issue, McKinney JM. Washington State's return to indeterminate sentencing for sex offenses: correcting past sentencing mistakes and preventing future harm. Seattle Law Rev 2002; 26: 309–336, <http://digitalcommons.law.seattleu.edu/cgi/viewcontent.cgi?article=1747&context=sulr>

- 32 Bioethicist Dr. Lachlan Forrow, director of ethics and palliative care programs, Beth Israel Deaconess Medical Center – in e-mail to Sheri Fink (12NOV90), as quoted in: Fink S. Five days at memorial. New York: Crown; 2013. p. 468.
- 33 DETECTER, Detection Technologies, Terrorism, Ethics, and Human Rights (<http://www.detecter.eu>)
- 34 The Markle Task Force on National Security in the Information Age, including James B. Steinberg, Vice President and Director, Foreign Policy Studies. Protecting America's freedom in the information age 2002, Markle Foundation.
- 35 Bernstein D. Big data's greatest power: predictive analytics. SAP, 22 November. <http://blogs.sap.com/innovation/big-data/big-datas-greatest-power-predictive-analytics-01138403>; 2013.

Index

A

Abductions. *See also* Kidnapping
Accuracy, 9, 18, 68, 141, 177, 178, 186
ACLED. *See* Armed Conflict Location & Event Data Project
Actionable Mining and Predictive Analysis for Public Safety and Security model, 59, 60, 62, 72, 185
Actionable Mining and Predictive Analysis model, 51
Actionable output, 186, 189, 206. *See also* Operationally actionable output; Output geospatial, 193
operationally relevant, 191
Activity-based intelligence, 370
Advanced analytics, 139
dissemination of, 302
for preoperational surveillance, 315
Aggravated assaults, 8, 163, 188
patterns, 246
robbery-related, 273, 295, 298
AI. *See* Artificial intelligence
ALPR. *See* Automated license plate recognition
Amazon Mechanical Turk, 145
Analysis. *See also specific analysis types*
future trends in, 369
process, 53
rapid, 94
sequential iterations in, 52
spatial, 162
Analytic competitor, 32, 33
Andromeda Project, 371
Anomaly detection, 148, 217, 344
process, 217
scoring algorithms and, 149

Anticipated events, 110, 260
Anticipation, 254
Anticipatory Black Swans, 10
AOI. *See* Area of interest
Approach to Counterinsurgency Operations, 361
Arab Spring, 5, 91
Area of interest (AOI), 167, 339
Arlington National Cemetery, 308, 309
Armed Conflict Location & Event Data Project (ACLED), 106
Artificial intelligence (AI), 25, 145, 228, 229
Association matrices, 16, 142
Authority Miner, 152, 153, 355, 356
Auto thefts, 192, 194, 195, 298, 299, 301
Automated feature extraction, 352
Automated license plate recognition (ALPR), 369, 376
Availability, 67

B

Beck, Charlie, 31
Behavior
characterization of, 33
fluid changes in, 271
heterogeneous, 177
hostile surveillance and, 322
modeling, 33
normal, 211, 259
normal criminal, 215
predictability of, 137
suspicious, 316, 336–338
Behavioral analysis, 116, 227, 252
pattern recognition in, 228
risk and threat assessment and, 259
of violent crime, 225, 244

Behavioral profiling, 223, 227, 246
Behavioral segmentation, 236, 239, 240
Bennett, William J., 231
Beslan School siege, 103, 260, 263, 271
Bias
cognitive, 71
expert systems and, 229
machine learning and, 71
against nonevents, 11
outcome measures and, 166
reporting, 323
violent crime and, 251
Big data, 79, 80, 367
HP modules and, 147
manipulation of, 204
PII and, 207
real time access to, 204
velocity and, 79
volume and, 79
Big Society, 276, 277
Bin Laden, Osama, 280
Biometrics, 368
Black Swans, 10
Bodycount (Bennett, Dilulio, and Walters), 231
Boosting, 176, 349
Boston Marathon bombing, 164
Brand integrity, 363
Burglary, 36, 215, 247
Business understanding, 56

C

Calls for service (CFS), 79, 84, 158
citizens complaints compared to, 87
frequency distribution of, 290
hot spots based on, 159
by time, 291

- Case clearance rates, 166
- Case-based reasoning, 34, 212, 227
- Categorization, 226
- evidence, 245
 - in homicides, 52, 245
 - of murder, 243
- CCTV. *See* Closed-circuit television
- Central Address systems, 107
- Central Intelligence Agency (CIA), 51
- CFS. *See* Calls for service
- Change detection, 204, 205
- Characterization, 35, 36, 226
- of behavior, 33
 - of drug-related violence, 231
 - in homicides, 245
 - victim, 242
- Cholera, 194
- CIA. *See* Central Intelligence Agency
- CIA Intelligence Process, 51, 53, 54, 55
- comparison, 72
 - CRISP-DM compared to, 59
- CIP. *See* Critical Infrastructure Protection
- Citizen complaints, 84–86, 105, 287
- CFS compared to, 87
 - patterns, 286, 287
 - random gunfire, 170
- Civil liberties, 376
- Classification models, 140, 179, 180
- Clearance metrics, 40
- Clearance rates, 183
- Climate data, 288
- Closed-circuit television (CCTV), 164, 352
- Cloud computing, 373
- Clustering techniques, 67, 140, 217, 218
- IEDs and, 195
 - with mapping, 331
 - sexual assault patterns and, 248
 - in surveillance detection, 328
- Coalition Provisional Authority, 258
- Cold case investigation, 229
- Cold hits, 246, 247
- Collaboration, 61, 198, 343
- COMINT. *See* Communications Intelligence
- Commercial sector, 32, 33
- Communications intelligence (COMINT), 54
- Community-based interventions, 61
- Community-Oriented Policing, 61
- Comprehensibility, 68
- Conference calls, 122
- Confidence matrices, 9, 178, 180
- Confusion matrices, 9, 177, 178
- Consensus opinions, 279, 280, 371
- Consistency, 95
- Context, 26, 46, 198
- Cookies, 129
- Costs, 151
- analysis, 268
 - of crime prevention, 172
 - of false alarms, 269
 - personnel resources and, 311
 - of responses, 268
 - risk-based deployment and, 167
- Counterfeiting, 363
- Counting
- incidents, 87
 - reporting differentiated from, 138
- Craven, John, 279
- Crime, 3
- associated with suspects, 246
 - displacement, 162, 170
 - distribution, 102
 - frequency, 40
 - globalization of, 374
 - heinous, 230
 - infrequent events and, 10
 - interrelated patterns in, 304
 - locations, 115
 - normal, 211, 215
 - primary gain from, 216
 - probability, 192
 - rates, 223
 - routine activity theory of, 225
 - secondary gain from, 216
 - staged, 219, 220
 - unequal, 165
- Crime Classification Manual, 237
- Crime incident reporting forms, 63
- Crime prevention, 223
- anticipated events and, 260
 - costs of, 172
 - move towards, 253
 - prediction and, 33
 - scope of, 165
- Crime reduction strategies, 236, 253
- Crime scene details, 229
- Crime trends, 41, 217, 273
- mapping and, 226
 - normal, 216
 - regional analysis of, 63
 - resource allocation and, 42
- CRISP-DM. *See* Cross Industry Standard Process for Data Mining
- Critical Infrastructure Protection (CIP), 265
- Cross Industry Standard Process for Data Mining (CRISP-DM), 51, 55, 56, 57, 60
- CIA Intelligence Process compared to, 59
 - comparison, 72
 - SEMMA compared to, 57
- Crowd physics, 278
- Crowd sourcing, 371
- Cyber threats, 358
- D**
- DARPA. *See* Defense Advanced Research Projects Agency
- Data. *See also specific data types*
- access to, 103
 - barriers, 104
 - binary, 78
 - breaches, 358
 - challenges, 97
 - collection, 53, 62, 81, 95, 104
 - descriptive, 4
 - design, 97
 - discrete, 78
 - exploration, 38
 - fusion, 62
 - imputation, 120
 - inaccurate, 98
 - instantiation of, 300, 312
 - integration, 63
 - layout, 97
 - limitations of, 26
 - missing, 120
 - narrowing focus on, 39
 - partitioning, 350
 - preoperational surveillance, 316
 - preparation, 56
 - privacy, 369, 376
 - quality, 65, 131
 - reliability, 26, 65
 - resources, 81
 - security, 62
 - sensitivity, 119
 - separation from, 26
 - sources, 14, 82
 - types of, 77
 - understanding, 56
 - unformatted, 78
 - validity, 65
 - verification, 104
 - violent crime, 40

- Data entry, 84, 95
 errors, 98
 nonessential, 298
- Data mining, 3
 definition of, 31
 descriptive statistics and, 4
 divisions of, 34
 privacy and, 380
 ROI and, 172
 in violent crime, 226
- Data Mining Moratorium Act, 378, 380
- Databases. *See also* DNA databases
 ad hoc, 92
 correctional, 247
 relational, 89, 90
 self-generated, 92
 specialized, 92
 of suspicious situation reports, 324
 timeliness and, 93
 tip information, 46, 92, 94
- Davenport, Tom, 33
- D.C. sniper investigations, 37, 38, 94, 262
- De Becker, Gavin, 267, 279, 316
- Deception, 250
- Decision rules, 8, 20, 103, 105
- Decision trees, 301, 302
 for categorizing evidence, 245
 for categorizing murder, 243
 intuitive, 141
 models, 68
 supervised learning algorithms and, 140
- Defense Advanced Research Projects Agency (DARPA), 279, 371
- Department of Defense (DOD), 93, 279, 305
- Department of Homeland Security (DHS), 265, 304
- Deployment, 287. *See also* Risk-based deployment
 of analytic content, 206
 of analytical output, 187
 decisions, 67, 117
 domain expertise and, 290
 effectiveness of, 158
 exploratory graphics and, 290
 fluidity, 302
 force, 285
 general concepts in, 286
 information-based, 285
 methods, 290
 modeling, 18, 293
 nature of incident and, 293
 operational plans, 333
 operationally actionable output and, 294
 outcome measure bias and, 166
 personnel resources and, 305
 remote, 304
 resource allocation and, 190
 schedules, 190, 292, 293
 of scoring algorithms, to end users, 302
 space and, 292
 time and, 290
 web-based, 202
- DHS. *See* Department of Homeland Security
- Diagnostics, screening *versus*, 154
- Digital cameras, 267, 269
- Dilulio, John J., Jr., 231
- Disaster response, 76
- Discovery, 32, 34, 157
 prediction and, 33
 process, 226
 stranger rapists and, 248
- Discriminant analysis, 36, 232, 247
- Dispatch zones, 191, 192
- Dissemination, 55, 57
- Distance functions, 342
- Distributed computing, 147
- DNA databases, 36, 215, 246, 247, 368
- DOD. *See* Department of Defense
- Domain expertise, 25, 37
 for analysts, 26
 complementary, 30
 deployment and, 290
- Domain-specific tools, 152, 372
- Dorn, Chris, 263
- Dorn, Michael, 263
- Drones, 326
- Drug dealers, 35, 43, 67
 juvenile, 242
- Drug enforcement strategies, 164, 225
- Drug-related violence, 36, 117, 244, 294
 characterization of, 231
 locations and, 234
 mapping and, 235
 outcome measures and, 160
 predictive model of, 201
 timeliness and, 234
- Duplication, 65, 101, 102, 119
 combined modeling algorithms and, 152
 recoding and, 127
- E**
- E-commerce, 31, 187, 188
- Elder Research, 172
- Electronic intelligence (ELINT), 54
- E-mail, 360
- Embezzlement, 213
- Emergency response plans, 32
- End users
 deployment of scoring algorithms to, 302
 recoding spatial data and, 115
- Enterprise Miner, 145, 147
- Errors, 8, 10, 182
 confusion matrices and, 9
 data entry, 98
 distribution of, 160
 false positives and, 179
 keystroke, 45
 law enforcement, 381
 logic, 132
 nature of, 177
- Ethical issues, 6
- Ethnographic research, 101
- Evacuations, 269
- Evidence categorization, 245
- Expert options, 349
- Expert systems, 228, 229, 370
 bias and, 229
 crime scene details and, 229
 limitations of, 230
- Exploratory graphics, 38, 42, 290
- Eyewitness testimony, 252
- F**
- Fabrication, 100
- Facebook, 354
- False alarms, 269
- False allegations, 219
- False positives, 179, 296
- Feedback, 55, 189
- Ferrara, Paul, 246, 247
- Fieldwork, 26–28
- Fingerprints, 367
- Fire safety, 270
- Firearms
 carry rate, 166
 illegal, 162
 injuries, 242
- “Fixing Intel” (Flynn), 29
- Flynn, Mike, 29
- FMV. *See* Full-motion video
- Fourth-generation warfare (4GW), 148, 258, 265

Fraud detection, 19, 79, 148, 164, 356
 Frequency distributions, 98, 214
 anomaly detection and, 218
 CFS, 290
 of SARs, 334, 335
 Full-motion video (FMV), 352
 Functional interoperability, 118
 FutureMAP, 279, 371

G

Galaxy Zoo, 371
 Generalizability, 18, 141, 173, 186
 Geographic coordinates, 107
 Geographic information system (GIS), 113
 Geography, 91, 144
 GEOINT. *See* Geospatial intelligence
 Geospatial data, 79, 351
 actionable output and, 193
 high-resolution, 325
 patterns, 240
 in UK riots, 274
 Geospatial intelligence (GEOINT), 54, 91, 202, 339, 340
 Geospatial predictive analysis, 80, 237
 future trends in, 370
 of LRA, 200
 in Northern Virginia military shooting series, 307
 as rule induction model, 306
 supervised learning and, 144
 Giduck, John, 263
The Gift of Fear (De Becker), 267, 279, 316
 GIS. *See* Geographic information system
 Globalization, 374
 Google, 371
 Google Flu Trends, 361
 GoogleEarth, 325
 Gray market distribution, 363
 Grossman, Dave, 263, 267, 269
 Gunfire, random, 84, 85, 86, 170
 with New Year's Eve initiative, 290
 spatial analysis of, 162
 weather and, 96

H

Habits, 226
 HANA, 367
 Heat maps, 88, 199
 Hierarchical organizational strategies, 113

High Performance analytics, 367
 High-Performance (HP) modules, 147
 Historic events, 296
 HITs. *See* Human Intelligence Tasks
 Homicide rates, 223
 aggravated assaults and, 163
 analysis, 33
 causes of, 293
 community impact of, 224
 drug-related, 234
 skewing, 102
 violent crime compared to, 169
 Homicides, 40. *See also* Murder categorization in, 52, 245
 characterization in, 245
 drug-related, 21, 33, 203, 225
 hostile actions and, 14
 incidental, 237, 239
 juvenile, 231
 motives for, 169
 patterns, 243
 race and, 13
 serial sexual, 232
 violent crime modeling and, 20
 in vulnerable populations, 106
 Hostages, 263
 Hostile actions, 14
 Hostile surveillance, 115. *See also* Preoperational surveillance;
 Surveillance detection
 analysis of, 270, 316, 325
 behavior and, 322
 escalating, 334
 identifying, 261
 intention of, 313
 Hot spots, 153, 157
 based on CFS, 159
 based on institutional knowledge, 158
 Hotels, 266, 281
 HP modules. *See* High-Performance modules
 Human Intelligence Tasks (HITs), 145
 Human terrain, 289
 Human-source intelligence (HUMINT), 54, 339, 340

I

IBM, 164
 IBM SPSS Modeler, 145, 146
 IBM Watson, 373
 IBR Resource Center, 106
 Identity theft, 357

IED. *See* Improvised explosive device
 Imagery intelligence (IMINT), 54
 Improvised explosive device (IED), 22, 93, 260
 clustering of, 195
 incident data, 196
 infrastructure information and, 197
 visualization of, 196
 Incidents, 116, 293
 counting, 87
 high frequency, 220
 linked, 133
 nonevents *versus*, 11
 Influence, 254
 Information. *See also* Personally identifiable information; Tip information
 binary, 245
 categorizing, 77
 infrastructure, 197
 offense, 118
 personal, 327
 types, 53
 Information Builders, 172, 191, 193–195
 Infrastructure
 critical, 265
 information, 197
 Infrequent events, 8. *See also* Low-frequency events
 Input, 20
 Institutional knowledge, 158
 Integrated model, 28, 29
 Intelligence, 91. *See also specific intelligence types*
 Intelligence analysis
 baseline data and, 13
 infrequent events and, 10
 operational value in, 18
 Internal rules, 34
 Internally Displaced People (IDP), 236, 237
 Internet
 activity analysis, 326
 anonymity of, 325
 data, 129
 misrepresentation and, 130
 over-sharing on, 327
 security, 202
 Internet protocol (IP) address, 129, 130
 Interoperability, 118
 Interval scales, 78

- Intuition, 230
Investigative efficacy, 166, 241
IP address. *See* Internet protocol address
Islamic State of Iraq and Syria (ISIS), 354
Iterative processes, 51, 52
in modeling, 52
recoding as, 112
- J**
Juvenile delinquency, 110
- K**
Kapow software, 274
Katrina (hurricane), 32, 76, 96
interoperability and, 118
lifeline CIP and, 265
public safety after, 160
KIA. *See* Killed in Action
Kidnapping, 249, 249. *See also* Abductions
Killed in Action (KIA), 11, 12
Knowledge management, 104
Kohonen network models, 143, 292
- L**
Law enforcement
errors, 381
4GW and, 258
operational value in, 18
suspicious situations and, 315
Law Enforcement Analytics, 191, 193–195
“Leaks from the future”, 181
Lifeline CIP, 265
Link analysis, 15, 43, 142
decision rule models and, 20
Link charts, 16, 44, 142
Locations, 107
auto thefts and, 301
complex, 329
crime, 115
drug-related violence and, 234
noncontiguous, 144
of Northern Virginia military shooting series, 306
physical, 320
public, 266
selection, by predators, 272
vulnerable, 262, 266
Lockdown drills, 269
London bombings of 2005, 164
Long, Letitia, 202
- Looting, 274
Lord’s Resistance Army (LRA), 106, 167, 237
behavioral segmentation and, 236, 240
geospatial predictive model of, 200
Los Angeles Police Department, 31
Low-frequency events, 69
boosting of, 176
modeling of, 175
risk-based deployment and, 273
LRA. *See* Lord’s Resistance Army
- M**
Machine learning, 71
Magnified events, 15
Mapping, 193. *See also* Geospatial data; Heat maps; Self-organizing maps
clustering techniques with, 331
crime trends and, 226
drug-related violence and, 235
facility, 205
operationally actionable output and, 294
risk-based deployment, 298
simple, 337
surveillance detection and, 323
of suspicious behavior, 338
of suspicious situation reports, 324
Mapping environment, 33
modeling algorithms in, 302, 303
translation of modeling algorithms for, 190
Marijuana, 233
Marine Corps, 305, 361
Measurement and signature intelligence (MASINT), 54
Medical fraud, 113
Melaku, Yonathan, 308, 309
Metadata, 376
MGRS. *See* Military Grid Reference System
Microtargeting, 371
Middle names, 131
Military applications, 361
Military Grid Reference System (MGRS), 107
Minority Report, 33, 378
Misrepresentation, 100, 130
MO. *See* Modus operandi
Mobile data computers, 83
Modeling, 7. *See also specific modeling techniques*
inaccurate, 19
iterative processes in, 52
of low-frequency events, 175
predictive drug-related, 201
public safety and, 19
risk and threat assessment, 268
suspicious situations reports, 316
updating, 181
violent crime, 20, 21, 234
Modeling algorithms, 139
appropriateness of, 139
combining, 151
comparison, 146
integrating, 148
in mapping environment, 302, 303
predictive efficacy of, 159
refreshing, 152
selecting, 139, 145
translation of, 190
Modus operandi (MO), 83, 133, 227, 271
Morgan Stanley Dean Witter, 76, 260
Motion, 204, 205
Motives, 132
automated determination of, 140
determination, 201, 232
for homicide, 169
homicide patterns and, 243
identifying, 34
victim risk factors and, 241
Mubarak, Hosni, 5, 115
Multifacility complex, 329, 330, 332
MultiNT environment, 198
Murder. *See also* Homicides
categorization of, 243
decision tree for, 243
juvenile, 251, 252
- N**
Narrative data, 350
National Center for the Analysis of Violent Crime, 242
National Geospatial-Intelligence Agency (NGA), 91
National Incident-Based Reporting System (NIBRS), 89
National Security Solutions (NS2), 339, 340
Natural language processing, 46
Nature, as outcome measure, 163
“Nature *versus* nurture” question, 230
Naval Research Laboratory, 131
Navy SEAL team, 12

- Neural networks, 140–142
 algorithms, 142
 simplified, 143
- Neurons, 142
- New Year's Eve initiative, 82, 310
- NGA. *See* National Geospatial-Intelligence Agency
- NIBRS. *See* National Incident-Based Reporting System
- "No free lunch" theorem, 181
- Noise reduction, 44
- Nominal scales, 78
- Nonevents, 22
 bias against, 11
 incidents *versus*, 11
 as outcome measure, 168
- Non-obvious relationship analysis (NORA), 45
- Nontraditional sources, 95
- NORA. *See* Non-obvious relationship analysis
- Nord-Ost Theater hostage siege, 263, 271
- Norms, 34
 internal, 212
 knowing, 213
 violence and, 239
- Northern Virginia military shooting series, 167
 geospatial predictive analysis in, 307
 high-resolution geospatial data in, 325
 locations of, 306
 risk-based deployment in, 304
- Notional data, 152
- NS2. *See* National Security Solutions
- O**
- Obscured events, 15
- O'Dell, Peter, 358
- ODNI. *See* Office of the Director of National Intelligence
- Offender counseling, 36
- Offense information, 118
- Offense reports, 27
- Office of the Director of National Intelligence (ODNI), 29
- Officer safety, 42
- Open-source information (OSINT), 54
- Operation Red Wings, 11, 12
- Operation Smokescreen, 363
- Operational personnel, 26, 81
 analysts relationship with, 28
 collaboration with, 61
 needs of, 27
 web-based analytics and, 303
- Operational security (OPSEC), 328
- Operational value, 18, 66
- Operationally actionable output, 58
 deployment and, 294
 mapping and, 294
 surveillance detection and, 322
- Operations, 28, 29
- Operator-analyst interaction, 29
- OPSEC. *See* Operational security
- Ordinal scales, 78
- Organizational readiness, 33
- OSINT. *See* Open-source information
- Outcome evaluation, 159, 169
- Outcome measures, 160, 164
 bias and, 166
 changes in, 165
 drug-related violence and, 160
 for fraud detection, 164
 nature as, 163
 nonevents as, 168
 ROI and, 171
 space as, 162
 specific, 163
 time as, 161
- Outlaw lifestyle, 244
- Outliers, 17
 clustering algorithms and, 217
 impact of, 173
 significant, 217
 with value, 101
- Output, 20. *See also* Actionable output; Operationally actionable output
 analytical, 185, 187
 risk and threat assessment and, 270
- Overfitting, 17
- Over-sharing, 327, 328
- P**
- Palantir analytic platform, 274
- Patrol deployment schedule, 88
- Pattern recognition, in behavioral analysis, 228
- "Pay and chase" decision rules, 103, 105
- PED cycle. *See* Processing, exploitation, and dissemination cycle
- Performance metrics, 164
- Personal information, 327
- Personalization, 327
- Personally identifiable information (PII), 70, 207, 376
- Personnel resources, 305, 311
- Peterson, Laci, 37
- PII. *See* Personally identifiable information
- Pirate attacks, 8
- Populations
 distribution, 163
 hidden, 4, 12
 measures, 289
 samples *versus*, 4
 transient, 289
 vulnerable, 106
- Praescent Analytics, 274
- Predators, 263, 272
- Prediction
 crime prevention and, 33
 discovery and, 33
 of low-frequency events, 69
 markets, 282
Minority Report and, 378
 terrorism, 280
- Predictive algorithms, 34
- Predictive analytics, 137. *See also* Geospatial predictive analysis
 additional considerations for, 70
 descriptive statistics and, 4
 discovery aspect of, 32
 in fraud detection, 19
 hazards, 71
 security concerns with, 71
 in violent crime, 226
- Preflight safety briefing, 259
- Preoperational surveillance. *See also* Hostile surveillance
 activity, 270
 advanced analytics for, 315
 data, 316
 indicators, 261
 intention of, 313
 reports, 344
 space and, 320
- Preprocessing, 107, 301
 operationally relevant, 64
 space in, 113
 time in, 110
 violent crime and, 107
- Printing devices, 206
- Prior probabilities, 10, 151

- Privacy, 70, 369
 analysis and, 377
 civil liberties and, 376
 data, 369, 376
 data mining and, 380
 Procedure reviews, 61
 Process models, 51, 266
 Processing, exploitation, and dissemination (PED) cycle, 352
 Processing trends, 372
 Profiling. *See* Behavioral profiling
 Project Exile, 161, 162, 165
 Project Safe Neighborhoods, 273, 286, 295
 Property crime, in sexual assault patterns, 247, 248
 Protective orders, 36
 Psypops, 148
 Public safety
 commercial sector and, 32
 data, 62
 after Katrina (hurricane), 160
 missing data and, 120
 modeling and, 19
 predators and, 263
 risk and threat assessment and, 257
 ROI in, 171
 time and, 110
 tip information and, 38
 violent crime modeling and, 21
- Q**
 al Qaeda, 45, 148, 263
 QRF. *See* Quick Reaction Force
 Quality-of-life increases, 82
 Questions, 60
 Quick Reaction Force (QRF), 12
- R**
 Race, homicide and, 13
 Rader, Dennis Lynn, 230
 RADR, 148
 Random assignment software, 175
 Random selection, 174
 Rank, 78
 Rapid Analytic Support and Expeditionary Response (RASER), 29
 Ratio scales, 78
 Recoding, 45, 46, 64
 of area codes, 126
 duplication and, 127
 as iterative process, 112
 offense information, 118
 operationally relevant, 109
 of SARs, 335
 spatial data, 113, 115
 of telephone data, 121, 122
 time and, 110
 variables and, 109
 Records management systems (RMS), 82, 152
 Red Balloon Challenge, 371
 Red teaming, 326
 Red zone, 316
 Regional fusion centers, 63
 Relational data, 89, 90
 Relationships
 confirmation of, 69
 link analysis and, 142
 overinterpretation of, 43
 surfacing, 15
 threshold for, 44
 Reliability
 checks, 99
 data, 26, 65
 validity and, 97
 Reporting. *See also specific report types*
 bias, 294
 counting differentiated from, 138
 streamlining of, 27
 Rescorla, Rick, 76, 254, 260
 Resource allocation
 crime trends and, 42
 deployment and, 190
 force deployment and, 285
 real time management of, 80
 after September 11 terrorist attacks, 304
 Return on investment (ROI), 171–173
 Reversals, 135
 Reverse lookup programs, 120, 121
 Reverse sting operations, 135
 Risk and threat assessment, 257, 267
 basic concepts, 259
 in behavior analysis, 259
 cost analysis in, 268
 evaluation in, 269
 modeling, 268
 normal behavior and, 259
 novel approaches to, 279
 operational aspects of, 271
 output and, 270
 process model considerations in, 266
 risk-based deployment and, 273
 surveillance detection and, 261
 Risk areas, 331
 Risk factors, victim, 241, 242
 Risk-based deployment, 138, 139, 286
 for auto thefts, 298
 based on victim lifestyle information, 160
 case studies, 295
 costs and, 167
 low-frequency events and, 273
 mapping, 298
 with New Year's Eve initiative, 310
 in Northern Virginia military shooting series, 304
 premise behind, 168
 risk and threat assessment and, 273
 robbery-related aggravated assaults and, 295
 ROI and, 172
 strategies, 51
 web-based analytics and, 298
 RMS. *See* Records management systems
 Robberies, 8, 137, 138
 armed, 297
 bank, 112, 113
 of drug dealers, 67
 kidnapping and, 249
 related aggravated assaults, 273, 295, 298
 seasonal variations and, 287
 ROI. *See* Return on investment
 Routine activity theory of crime, 225
 RTI International, 65
 Rule induction models, 68, 140, 301, 306
 Rule sets, 35
 intuitive, 141
 supervised learning algorithms and, 140
- S**
 SaaS. *See* Software as a Service
 Samples, 6, 7
 analysis of, 7
 nonrandom, 6
 populations *versus*, 4
 random, 6
 test, 173, 174
 SAP, 339, 367
 SARs. *See* Suspicious activity reports

- SAS, 57
- SAT. *See* Semantic Analysis Technology
- SCAN. *See* Scientific Content Analysis
- Schools, 262
- Scientific Content Analysis (SCAN), 120, 250
- Scoring algorithms, 148
 - anomaly detection and, 149
 - deployment to end user of, 302
 - e-commerce, 188
 - real-time, 191
 - samples, 302
- Scorpion*, 279
- Screening, diagnostics *versus*, 154
- Search engines, 130
- Seasonal variations, 111, 153, 181, 217, 287
- Security
 - commercial, 362
 - concerns, 71
 - data, 62
 - Internet, 202
 - physical enhancements, 333
 - private, 362
 - time and, 110
- Self-organizing maps, 140, 143, 248. *See also* Kohonen network models
- Self-organizing networks, 292
- Semantic Analysis Technology (SAT), 373
- SEMMA, 51, 57, 58
 - CRISP-DM compared to, 57
- Sense making, 76
- Sentiment, 341
- September 11 terrorist attacks, 37, 76
 - anticipated events and, 260
 - BOLOs after, 314
 - resource allocation after, 304
- Serial killers, 13, 52
- Sexual assault patterns, 246
- al Shabab, 103
- Shermer, Michael, 279
- Shopping centers, 266
- Signals intelligence (SIGINT), 54, 79
- Signal-to-noise issues, 16
- Signature, 227
- Situational awareness, enhanced, 343
- Six degrees of separation, 44
- SNA. *See* Social network analysis
- Snow, John, 194, 195, 198
- Social disorder, 276
- Social learning, 231
- Social media, 340, 353, 355
 - analysis, 353
 - geolocated, 114, 115
 - in Tahrir Square protests, 278
- Social media intelligence (SOCMINT), 91
- Social network analysis (SNA), 152, 153, 355, 356
- SOCMINT. *See* Social media intelligence
- Software as a Service (SaaS), 373
- Southside Strangler, 36
- Space, 113
 - deployment and, 292
 - in deployment schedules, 293
 - as outcome measure, 162
 - personal, 321
 - preoperational surveillance and, 320
 - relative indicators of, 321
- Spatial boundaries, 321
- Spatial data, 113, 116
- Special Operations, 12
- Spencer, Timothy, 36
- Spoofing, 130, 135
- Staffing shortages, 27
- Stalking, 327
- Statement analysis, 104
- Statistics, 75
 - aggregate level, 170
 - basic, 3
 - descriptive, 4
 - preliminary, 300, 301
 - summative content analysis through, 5
 - drilling down in, 170
 - inferential *versus* descriptive, 4
 - summary, 137
- Stranger rapists, 246, 248
- Summative content analysis, 4, 5
- Superpredators, 231
- Supervised learning
 - actionable output and, 186
 - algorithms, 140
 - geospatial predictive analysis and, 144
 - unsupervised learning *versus*, 140
- Surprise, 35
- Surveillance detection, 129, 314. *See also* Hostile surveillance
 - case studies, 328
 - clustering techniques in, 328
 - in complex locations, 329
 - countermeasures, 271
 - general concepts of, 315
 - mapping and, 323
 - operational benefits of, 272
 - operationally actionable output and, 322
 - red teaming and, 326
 - risk and threat assessment and, 261
 - risk areas, 331
 - unsupervised learning in, 328
 - web services model, 339
- Suspects
 - crime associated with, 246
 - recoding spatial data and, 115
- Suspicious activity reports (SARs), 317
 - frequency distribution of, 334, 335
 - recoding of, 335
- Suspicious situation reports, 313, 314, 319, 330
 - analysis, 344
 - databases of, 324
 - by day, 319
 - mapping of, 324
 - modeling, 316
 - over time, 318
- Synapses, 142
- Syndromic surveillance, 360
- ## T
- Tactics, techniques, and procedures (TTP), 259
- Tahrir Square protests, 114, 115, 278, 351
- Technology, 39, 372
 - costs and, 151
 - data entry and, 84
 - incompatible, 118
 - warfare and, 258
- Telecommunications, 123
- Telephone data, 120–122
- Terrorism, 8, 326
 - 4GW and, 258
 - predictions, 280
 - schools and, 263, 264
- Text data, 78
- Text mining, 45, 350, 351
- The Onion Router (Tor), 131
- Threats. *See also* Risk and threat
 - assessment
 - assessment, 259
 - irregular, 361
 - nature of, 116, 293
- TIA. *See* Total Information Awareness Time, 107, 108

- blocks, 111
 - CFS by, 291
 - deployment and, 290
 - in deployment schedules, 293
 - as outcome measure, 161
 - public safety and, 110
 - security and, 110
 - suspicious behavior over, 336
 - suspicious situation reports over, 318
 - as variable, 108
 - Timeliness, 67, 83, 93, 234
 - Tip information, 37, 38, 45, 94
 - databases, 46, 92, 94
 - TomNod, 145, 371
 - Tor. *See* The Onion Router
 - Total Information Awareness (TIA), 380
 - Tradesports.com, 280
 - Traffic patterns, 197
 - Training, 25, 34, 173, 174
 - Transaction records, 213
 - Transparency, 200
 - Treatment matching, 240
 - Trinity Sight, 65
 - TPP. *See* Tactics, techniques, and procedures
 - Twitter, 106, 278, 354
- U**
- UAVs. *See* Unmanned aerial vehicles
 - UCR. *See* Uniform Crime Reports
 - UK riots, 274
 - Uniform Crime Reports (UCR), 89, 183
 - Unintended consequences, 169
 - United States Office of the Director of National Intelligence (US ODNI), 73
 - Unmanned aerial vehicles (UAVs), 326
 - Unsupervised learning, 67, 68, 127
 - algorithms, 128, 140, 204
 - Kohonen network models and, 143
 - supervised learning *versus*, 140
 - in surveillance detection, 328
 - US ODNI. *See* United States Office of the Director of National Intelligence
 - U.S. Postal Service Office of Inspector General, 172
 - Ushahidi, 371
- V**
- Validity, 25, 97, 104
 - Variables
 - actionable, 131, 132
 - continuous, 77
 - discrete, 78
 - entry, 149
 - for instantiation of data, 312
 - operationally relevant, 131
 - recoding and, 109
 - selection, 66, 150
 - challenges, 133
 - manual, 150
 - specifying, 300
 - time as, 108
 - Victimology, 117, 232, 241
 - Victim-operated improvised explosive devices (VOIED), 133
 - Victim-perpetrator relationship, 232, 243
 - Victims, 233, 242, 250, 263
 - characterization, 242
 - education, 36
 - group identification of, 52
 - handedness of, 220
 - lifestyle information, 160
 - risk factors, 241, 242
 - Violence, 225. *See also* Drug-related violence
 - domestic, 36
 - as enforcement tool, 240
 - norms and, 239
 - with secondary objective, 231
 - systemic, 244
 - workplace, 316
 - Violent crime, 8, 250, 250., 253. *See also specific violent crimes*
 - behavioral analysis of, 225, 244
 - bias and, 251
 - common patterns in, 243
 - community impact of, 223, 224
 - data, 40
 - data mining in, 226
 - drug-related, 117
 - frequencies, 41
 - homicide rates compared to, 169
 - homogeneity of, 244
 - index, 165
 - modeling, 20, 21, 234
 - patterns, 41
 - predictive analytics in, 226
 - preprocessing and, 107
 - weather and, 96
 - Visualization, 39, 137
 - analytical output and, 185
 - future trends, 202
 - goal of, 186
 - of IEDs, 196
 - overselling, 207
 - tools, 41, 42
 - VOIED. *See* Victim-operated improvised explosive devices
- W**
- Wal-Mart, 31, 96
 - Walters, John P., 231
 - WAMI. *See* Wide-area motion imagery
 - Warfare, 258, 258. *See also* Fourth-generation warfare
 - Weather, 95, 153, 288
 - Web interface, 187, 189
 - Web logs, 130
 - Web services model, of surveillance detection, 339
 - Web-based analytics, 296
 - dissemination of advanced analytics through, 302
 - interface, 300
 - operational personnel and, 303
 - risk-based deployments and, 298
 - Web-based deployment, 202
 - Web-enabled tools, 189
 - Wide-area motion imagery (WAMI), 352
 - Workflow, 145, 146
 - World Trade Center bombing of 1993, 76, 118
- Z**
- al-Zarqawi, Abu Musad, 280