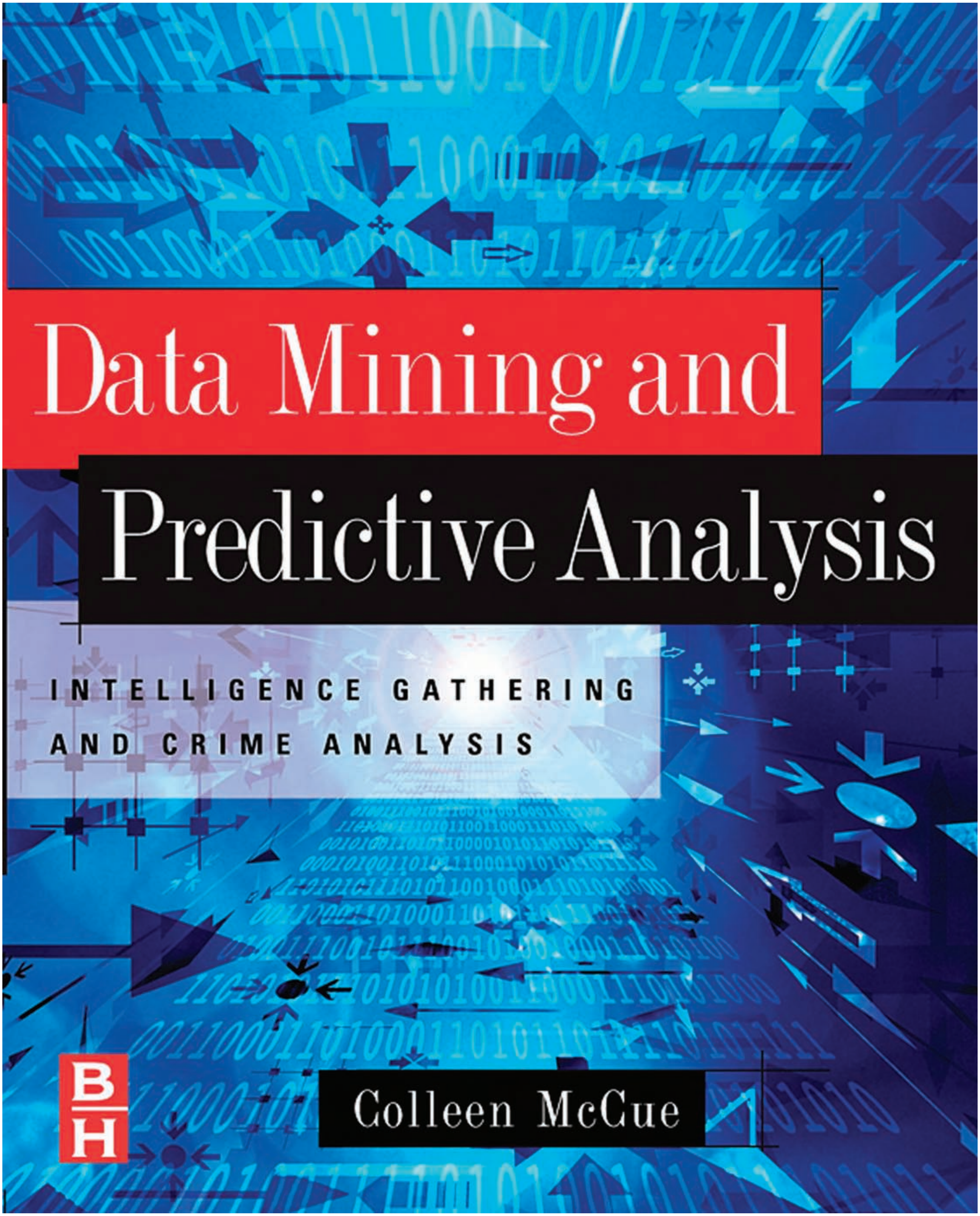


Data Mining and Predictive Analysis

Intelligence Gathering and Crime Analysis

Colleen McCue

This Page Intentionally Left Blank

Data Mining and Predictive Analysis

Praise for Data Mining and Predictive Analysis

"Dr. Colleen McCue pairs an educational background in neuroscience and psychology with extensive experience in the fields of behavioral science, cirme analysis, and intelligence gathering to create *Data Mining and Predictive Analysis*, a must-read for all law enforcement professionals. Within the ever-growing fields of criminal justice and crime analysis, Dr. McCue combines all facets of the public safety community, effortlessly examining techniques in which law enforcement, analysts, and researchers are able to delve deeper through her accessible explanations of relative degrees of data quality, validity and reliability; all essential tools in this modern, technological era."

Arthur E. Westveer (Associate Professor, L. Douglas Wilder School of Government and Public Affairs, Virginia Commonwealth University)

"[Data Mining and Predictive Analysis] is a must-read . . ., blending analytical horsepower with real-life operational examples. Operators owe it to themselves to dig in and make tactical decisions more efficiently, and learn the language that sells good tactics to leadership. Analysts, intell support, and leaders owe it to themselves to learn a new way to attack the problem in support of law enforcement, security, and intelligence operations. Not just a dilettante academic, Dr. McCue is passionate about getting the best tactical solution in the most efficient way—and she uses data mining to do it. Understandable yet detailed, [*Data Mining and Predictive Analysis*] puts forth a solid argument for integrating predictive analytics into action. Not just for analysts!"

Tim King (Director, Special Programs and Global Business Development, ArmorGroup International Training)

"Dr. McCue's clear and brilliant guide to attacking society's greatest threats reveals how to best combine the powers of statistical computation and the experience of domain experts. Her emphasis on understanding the essential data through fieldwork and close partnership with the end users of the information is vital to making the discovered patterns "actionable". Anyone seeking to harness the power of data mining to "connect the dots" or "find needles in a haystack" will benefit from this lively and reliable book packed with practical techniques proven effective on tough real-world problems."

Dr. John Elder (Chief Scientist of Elder Research, Inc., www.dataminingla b.com)

"[Data mining] is a hot area—not just for Hollywood any more—but real people and real situations are benefiting from these analytical investigations. "

Mary Grace Crissey (Technology Marketing Manager, SAS Institute)

Data Mining and Predictive Analysis

Intelligence Gathering and Crime Analysis

Colleen McCue

ELSEVIER

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

Butterworth-Heinemann is an imprint of Elsevier

Butterworth-Heinemann is an imprint of Elsevier
30 Corporate Drive, Suite 400, Burlington, MA 01803, USA
Linacre House, Jordan Hill, Oxford OX2 8DP, UK

Recognizing the importance of preserving what has been written, Elsevier prints its books on acid-free paper whenever possible.

Library of Congress Cataloging-in-Publication Data
McCue, Colleen.
Data mining and predictive analysis: intelligence gathering and crime analysis/ Colleen McCue
p. cm.
Includes bibliographical references and index.
ISBN 0-7506-7796-1 (alk. paper)
1. Crime analysis. 2. Data mining. 3. Law enforcement–Data processing. 4. Criminal behavior, Prediction of. I. Title.

HV7936.C88M37 2006
63.25′6–dc23
2006040568

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

ISBN 13: 978-0-7506-7796-7
ISBN 10: 0-7506-7796-1

For information on all Butterworth-Heinemann publications
visit our Web site at www.books.elsevier.com

Printed in the United States of America
06 07 08 09 10 10 9 8 7 6 5 4 3 2 1

This book is dedicated to Patrick Michael McLaughlin,
the first miner in our family.

This Page Intentionally Left Blank

Contents

Case Examples 237

Advanced Concepts and Future Trends 299

Foreword

We all know crime doesn't pay. But did you know there is "prophet" in policing? Thanks to the fine work of Dr. Colleen McCue of the Richmond Police Department, Crime Analysis Unit, it is now possible to predict the future when it comes to crime, such as identifying crime trends, anticipating hotspots in the community, refining resource deployment decisions, and ensuring the greatest protection for citizens in the most efficient manner.

A number of years ago, the United States Attorney's Office for the Eastern District of Virginia formed a partnership with the Richmond Police Department to address the pressing problem of gun violence in the city. In 2002, we renewed that relationship and formed a new commitment as part of President George W. Bush's antigun crime initiative, Project Safe Neighborhoods (PSN). At that time, Dr. McCue was selected as our research partner to assist our efforts in evaluating the outcomes of our districtwide PSN initiatives. In light of the work Dr. McCue was already doing for the Richmond Police Department, we wanted to apply the innovative tools she had used so effectively in Richmond to support our efforts targeting gun crime in other hot spots around eastern Virginia.

Dr. McCue has done pioneering work in the practical application of data-mining techniques to the administration of a police department. In this book, she describes her use of "off-the-shelf" software to correlate data on gun violence with data on other violent crimes in order to graphically depict crime trends in a most compelling way and to predict where future crimes are likely to occur. Armed with such analyses, the police executive is thus enabled to develop "risk-based deployment strategies," permitting the executive to make informed and cost-efficient staffing decisions based on the likelihood of specific criminal activity.

The application of Dr. McCue's techniques has paid off in Richmond, where the police department used them to deploy resources during the period surrounding the New Year's Eve holiday—December 31, 2003, through January 1, 2004. The results of that effort were dramatic. Not only were gunfire complaints reduced by almost 50% on New Year's Eve, but the number of seized illegal weapons increased by an impressive 246% from the previous year. These statistics represent compelling evidence that these techniques are adding value to the work of fighting gun crime. But there is more. This accomplishment was realized using fewer street officers than originally planned. In other words, risk-based deployment enabled the Richmond Police Department to deploy fewer officers strategically, while at the same time obtaining better results.

In writing this book, Dr. McCue was mindful of the need to convey sophisticated analyses in practical terms and, accordingly, she prepared her text in a very user-friendly manner. As United States Attorney, I am proud to be associated with such a dedicated partner in our shared mission. I am confident that you, too, will benefit from Dr. McCue's exceptional contribution to the field of police science.

Paul J. McNulty

Preface

Like many kids growing up in America, I always had a love of science. I also happened to be blessed with two incredibly supportive and involved parents. My mother was always there with words of encouragement. Her typing skills got me through high school and most of college. She also led by example, balancing her work as a probation and parole officer with her role as wife and mother. My father, on the other hand, would try to learn as much as he could about what we were interested in so that he could participate in the activities with us. When I started graduate school, however, there was something of a dilemma. What do an engineer and a budding neuroscientist have in common, particularly when the engineer is not big on things like rats and brains? Fortunately, it was during this time that cognitive neuroscience and artificial intelligence systems started becoming accessible to the mainstream. So, throughout graduate school and my subsequent career, my father would send me books and articles on topics such as neural nets, case-based reasoning, machine learning, and cognitive neuroscience. It provided for interesting conversation and some common ground for two professionals in relatively disparate fields.

As time went on and life changed, I found myself working as a behavioral scientist in the criminal justice field. In this environment, I was able to bring my training as a scientist to the study of human criminal behavior. I found that I was able to apply much of what I had learned about psychology, behavioral science, and, perhaps most importantly, multivariate statistics and computer modeling to my new field. I was in an interesting position, working in a local police department and receiving first-hand training in a variety of topics, from death investigation to CompStat. While I did not realize it at the time, I also was acquiring a tremendous amount of domain expertise, something absolutely essential to competent data mining, which would distinguish my work from many others trying to gain entry into a rather closed professional world. I also began to understand the relative degrees of data quality, validity, and reliability associated with law enforcement and intelligence data. Although I was familiar with the work regarding the often questionable reliability of eyewitness testimony, it was not until I had read many offense reports that trends and patterns to the witness statements began to emerge and make sense.

I became profoundly intrigued by how many of the seasoned detectives who 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 meetings 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 this same time period, I 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, something that I continue to be involved in. Similarly, it 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. Unfortunately, the transfer of this powerful technology to the public safety arena has not advanced nearly as quickly.

While I did not realize it at the time, a relatively new approach to marketing and business 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. Contrasting the state of public safety analytical capacity to that of the business community only serves 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. 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 Tenent, the amount of information almost continuously 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 analysis and sharing of information resulted in tragic consequences. While information sharing will require culture change and a paradigm shift in the larger public safety community, advanced analytical techniques are available now. The same tools that were being used to prevent people from switching their cellular 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 used widely in the business community. The key is to apply them to questions or challenges in public safety, law enforcement, and intelligence analysis. Adapting existing technologies and analytics to the public safety domain will keep many of us busy for years to come. If the past is any indicator, however, by the time we have completed this initial technology transfer and have caught up to where the business community is today, there should be other new and exciting technologies to appropriate from the private sector. In all seriousness, the public safety community has become extremely adept at developing and adapting new and advanced technologies for operational capacity and support. The battlefields have changed, though. To achieve dominance in the war on terrorism, the war on drugs, and the war on crime, we need to devote additional attention to our ability to manage, analyze, and utilize the incredible amounts of information available. Ultimately, data mining and predictive analytics offer the promise of allowing data and information to serve as a transparent, fluid interface between analytical and operational personnel, rather than the vast ideological divide that frequently is encountered today.

Although I say "I" quite a bit in this book, the book 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 would like to thank Dave Dunn from Advizor Solutions, Inc. Dave first suggested that I write this book, and it never would have occurred to me that this was possible without his feedback and support. Mark Listewnik at Elsevier has the patience of a saint. His ongoing support and encouragement, not to mention the very nice Christmas cards that I continued to receive despite the fact that I was horrendously late on my rewrite and edits, kept me going if for no other reason than I felt very guilty putting things off even further in the face of his ongoing kindness. Finally, Kayla Gray at RTI International edited the manuscript and helped create something far more readable than what I originally wrote. Her attention to detail and thoughtful comments are reflected throughout the text.

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 have learned a tremendous amount from his lectures, which are punctuated with his dry sense of humor and wonderful anecdotes from a very successful career with the Baltimore Police Department. Rich Weaver and Tim King, president and vice president, respectively, at International Training, Inc. graciously allowed me to attend their lectures and training 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.

While 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 recent 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 want to recognize Frank Shults and Brian Whisler, who are blessed with both excellent writing skills and unbridled humility. To them, I am grateful.

I also would like to thank Dr. Harvey Sugerman. 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 companies that have provided me with some wonderful toys—I mean software. Without their support, I would still be performing unnatural acts with multivariate statistics and trying to convey the results to operational personnel with a lot of hand waving. In particular, Dr. Tom Khabaza and Bill Haffey with SPSS, Tracye Giles at SAS, and Dave Dunn with Advizor Solutions trusted me enough to give me the tools to do a lot of the work outlined in this book.

My family at the Richmond, Virginia Police Department has 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 Colonel Andre Parker and Lieutenant Colonel Teresa Gooch for their ongoing support of my work. I also 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. Other colleagues in the Department include Captain David Martin, Majors Peggy Horn and Dave McCoy, the late Major Rick Hicks and Captain Donnie Robinson, and my friend Alicia Zatcoff, Esq. I also owe a tremendous debt to the Virginia Homicide Investigators Association, where I have received some outstanding training in death investigation and was fortunate enough to meet my husband, who is a member of their board of directors. My colleagues in law enforcement have taught me as much, if not more, about life in the many years that I worked with them.

Underscoring the length of time that it took me to complete the text, I changed employment during the writing of this book. After several years in the applied setting, I joined RTI International, a nonprofit research organization with an international reputation for excellence in criminal justice research. The ability to work with other like-minded researchers in an effort to advance the science and practice of public safety and security has been energizing. In particular, Dr. Victoria Franchetti Haynes, president and CEO of RTI International, has created an environment that fosters creativity and the opportunity to improve the human condition by turning knowledge into practice. Adam Saffer and Brent Ward have helped me translate my work into something tangible that can be shared with other public safety and security organizations through the creation of technology and the provision of professional services. Other colleagues at RTI include Drs. Al Miedema and Jim Trudeau, and MG (Ret) Lon "Bert" Maggart, as well as the other members of my research team, which includes Dr. Kevin Strom and Mark Pope. Confucius said that if you love your job you will never work a day in your life, something that I am blessed to live.

I also would like to thank Mike Sullivan, USMC Staff Sergeant Tom Ferguson, and Special Agent BJ Kang 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. Joey Vail from SAS, Bill Haffey from SPSS, Eric Greisdorf from Information Builders, and Kurt Rivard from Advizor Solutions all provided screen shots that illustrated the value that their software can bring to applied public safety and security analysis.

Perhaps most importantly, I would like to acknowledge my family. My parents, Phil and Lucy McLaughlin, always expected the best from me and my siblings, Michele and Tim, giving us the tools necessary to achieve that and more. This included loving words and kind gestures, as well as giving us permission to find our own way in life. My path has not always been direct or easy, but they always loved me enough to allow me to find my own way, having faith in me even when I did not. Some of the most challenging lectures that I have ever given professionally were the ones where they were in the audience. To look out and see their faces filled with pride was at once humbling, heart-warming, and also terrifying. Who would have known that the girl from

Downers Grove, Illinois, who started out as an aerospace engineering student would have taken the career path that I did? It still seems amazing to me at times, but I know that I am a far more successful person because of it. Unfortunately, I think that my parents went prematurely grey in the process. Hopefully, it was worth it.

In many ways, my husband, Special Agent Rick McCue, has contributed more than enough to have earned the right to be a coauthor. Through him, I have 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 also would like to thank the United States government for sending him out of the country so much during the writing of this book. I always looked for projects to occupy my time when he was out of pocket; we could not afford any more redecorating, so this book seemed like a good alternative. In all seriousness, though, 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 with Operations Noble Eagle and Iraqi Freedom, I began to truly understand the value that good intelligence and analysis will bring to the war on terrorism.

Our children, Paul, Alexandra, Elaine, Patraic, and Gabriel, keep me humble. Although Rick and I lead very exciting lives professionally, our kids still think that we are the biggest dorks in the whole world, clueless and goofy. That fact alone keeps me anchored in reality and reminds 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, 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. In all humility, it is my sincere wish that the techniques and approaches outlined in this book will help us increase 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
Senior Research Scientist
RTI International

This Page Intentionally Left Blank

Introduction

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

Whether it is called data mining, predictive analytics, sense making, or knowledge discovery, 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 security, crime prevention, and crime analysis. 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 some of the categorization, linking, and predictive analytics associated with data mining and predictive analytics.

While the relationship between the two areas seems to be natural, the law enforcement community in particular has not enjoyed many of the analytical benefits coming 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 more important than trying to create these associations, though, will be addressing gaps in information and information sharing. 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 prevented. The emphasis needs to shift from describing the past to predicting the future. 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.

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 have been adapted to run 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,[ii] even within smaller agencies with comparatively limited personnel resources.

The more that 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 recent security expo, 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."[iii] In my opinion, 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. While most analysts probably do not need to fear for their jobs just yet, increasingly friendly and intuitive computer systems 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.

“Agent/Analysts” and Future Trends

I see a day in the not-too-distant future when analysis will be available without immediate access to an analyst. Information from operations will feed analysis, while the analysis will concomitantly drive the operations, thereby creating a feedback loop of ever-increasing information and actionable intelligence. I see a day when a patrol officer will come back to work after several days off and, at the beginning of the tour, will be able to review recent patterns and trends within the context of historical data and accumulated knowledge from the mobile data terminal in his cruiser. After responding to his first call, he will be able to enter the incident information directly into the department’s computerized records management system (RMS) using direct voice commands. This information then will be used to create the computerized offense report. Any digital images captured from the incident will be quickly uploaded and linked directly to the offense report, as well as any associated or linked information already stored in the RMS. During the data entry process, this new information will pass through an analytical filter prepared earlier in the week by the analytical staff, who are home asleep at this hour. The algorithm running in the background will quickly link this most recent incident to a recent series and prompt the patrol officer to consider several possible alternatives. With this real time, value-added analysis, the officer can make quick, information-based operational decisions that result in a rapid apprehension of the criminal.

This is handled similarly when an agent in a remote location is debriefing a suspected terrorist. The verbal information is recorded and transcribed directly into a free format text file using voice recognition software. The file is then uploaded to an analytical fusion center a thousand miles away. An analyst there uses sophisticated text mining technology to probe and characterize the results of the interview. Several key phrases are identified and compared to an existing database generated from earlier interviews with members of the same terrorist cell being held in other locations around the world. Based on the analysis of the current interview and its comparison to the existing models, areas of possible deception and truth are identified and highlighted, as are promising interviewing strategies. This information, including the interviewing strategies and approaches, is sent back to the agent in the field, further informing and guiding the ongoing interview process, while concomitantly enhancing the existing intelligence on the operations, practices, and strategies of this particular terrorist group.

Are these extravagant predictions? Absolutely not. Both scenarios outlined above are based on existing technologies and resources. In many ways, approaches and methodologies similar to information management have been used in the business community for years. All that is required to implement these strategies is a commitment to take advantage of the currently existing analytical tools and incorporate them into our world. Unfortunately, a paradigm shift in how we view information, analysis, and the relationship between analytical and operational personnel also will be required. That probably will be the most difficult task. Once we overcome that hurdle, however, adapting these new technologies promises to be one of the most exciting adventures in public safety in our lifetime.

How To Use This Book

All of the examples included in this book come 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 methods. To be sure, though, while the names might have been changed to protect the "not so innocent," the examples are based on real experiences.

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. While 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 data mining products. While 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 merely provides 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 using data mining and predictive analytics, which each law enforcement or intelligence professional can tailor to their own unique situation and responsibilities. While 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

i. Casey, E. (2002). Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and internet crime investigation. Knowledge Solutions. www.corpus-delicti.com/case_based.html

ii. McCue, C. and Parker, A. (2004). Web-based data mining and predictive analytics: 24/7 crime analysis. *Law Enforcement Technology*, **31**: 92–99.

iii. Fisher, D. (2003). Clancy urges CIOs: seek out the "smart people." eWeek, www.eweek.com.

This Page Intentionally Left Blank

Introductory Section

This Page Intentionally Left Blank

1

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 below are some simple statistical terms and concepts that are relevant to data mining and analysis, 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 Inferential versus Descriptive Statistics and Data Mining

Descriptive statistics, as the name implies, is the process of categorizing and describing the 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 relationships in the information or models, particularly those that can be used to anticipate or predict future events. Therefore, data mining more closely resembles descriptive statistics.

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.

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 that we know about crime and criminals is based on the unsuccessful ones, those that got caught. Most criminal justice research is based on correctional populations, or offenders that have some sort of relationship with the criminal justice system. 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.

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.

Infrequent Events

When dealing with violent crime, 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.[1] 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.

One way to evaluate the specific nature of the errors is to create something called a confusion or confidence matrix. Basically, 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 it 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.

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 1-1).

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 1-2, 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

Figure 1-1 *Link charts can not only depict relationships between individuals or events, but also relative strength of the relationship based on relative differences in the amount of contact.*

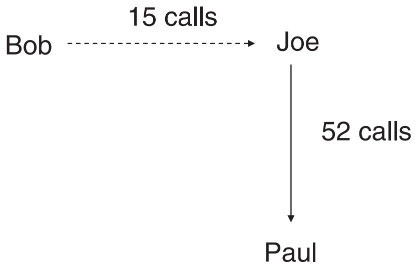
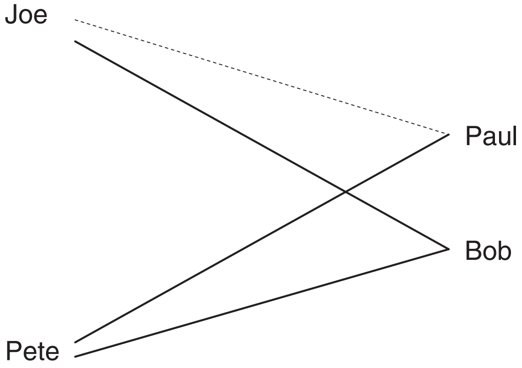


Figure 1-2 *Examination of this link chart suggests that Paul has a stronger relationship with Pete compared to his relationship with Joe, while Bob appears to have relatively similar relationships with both Joe and Pete, based on relatively equal levels of contact. These apparent differences in the relationships are based on differences in the strength of the association illustrated by relative differences in the lines in the link chart.*



the related association matrix, however, indicates that this might not be true (Figure 1-3). 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 1-1) 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 10 associations or less, while strong links require 20 associations or more. Unfortunately, unequal

Figure 1-3 *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 1-2.*

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

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, but 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 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.

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, one-third 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.

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.[2] 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.[3] When adjusted to per-capita rates, the FBI Uniform Crime Reports indicate that blacks are eight times more likely to commit homicides than whites.[4] 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.

1.6 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 it 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-minute 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-minute 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 a "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.[5] 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.[6] 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.

1.9 Bibliography

1. McCue, C. and McNulty, P.J. (2003). Gazing into the crystal ball: Data mining and risk-based deployment. Violent Crime Newsletter, U.S. Department of Justice, September, 1–2.
2. U.S. Census Bureau. (2000). 2000 Census. www.census.gov
3. 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
4. Ibid.
5. McLaughlin, C.R., Daniel, J., and Joost, T.F. (2000). The relationship between substance use, drug selling and lethal violence in 25 juvenile murderers. *Journal of Forensic Sciences*, **45,** 349–53.
6. McCue, C. and McNulty, P.J. (2004). Guns, drugs and violence: Breaking the nexus with data mining. *Law and Order*, **51**, 34–36.

2

Domain Expertise

"When all else fails, ask your father; he knows practically everything. And those things of which he is ignorant, are not worth knowing anyway."

Phillip McLaughlin.

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

Analysts that 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 two o'clock in the morning 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 two o'clock in the morning in a 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 Compromise

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

Figure 2-1 *In the traditional model, analysts prepare reports and other analytical output with little input or feedback from the operational end users.*

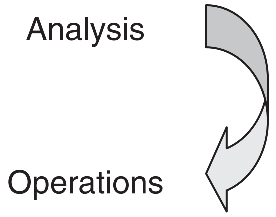
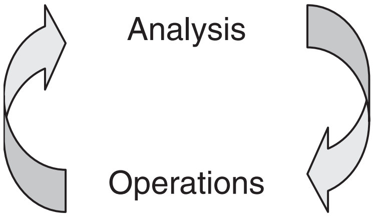


Figure 2-2 *By establishing information as a fluid interface between operational and analytical domains, it is analyzed within an operational process. Sworn personnel are able to guide the process, including the nature, structure, and format of analytical output, which increases the likelihood that analytical 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.*



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

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.

Think back to one of the scenarios outlined in the Introduction: An agent is debriefing a suspected terrorist in some remote location. The verbal information is automatically recorded, transcribed, and then uploaded to an analytical fusion center a thousand miles away. There, an analyst, using sophisticated text and data mining technology, characterizes and models the results of the interview based on additional information gathered from similar operations throughout the world. Based on this analysis, the agent in the field receives timely information and analysis that enhances his current interview process, while concomitantly increasing the existing intelligence on the operations, practices, and strategies of this particular group. Through this innovative use of expert systems, the intuitive nature of data mining and predictive analytics affords new opportunities for collaboration between analytical and operational personnel that ultimately will enhance our awareness for the war on terrorism, the war on drugs, or the war on crime.

2.4 Analyze Your Own Data

The folks working on the job frequently are in the best position to analyze their own data. Whether the analytical staff, the operational personnel, or a combination of the two, these individuals clearly have the requisite domain expertise to engage in a thoughtful, meaningful analysis of the information and create actionable analytical output that will have value in the field. While it is tempting to pass this task along to outside consultants or statistical gurus, professionals in the public safety arena have a solemn responsibility to those they serve to ensure that they receive the best service possible. Given the domain expertise and experience within the organization, it does not make sense to abdicate this function to someone who might not have the same level of knowledge or understanding just because they have some skills with analytical software. One of the most exciting aspects associated with the newer generation of data mining software applications is that they are intuitive enough to enable even “mere mortals” access to these powerful crime-fighting techniques, which will allow law enforcement and intelligence personnel the opportunity to analyze their own data.

2.5 Bibliography

1. Casey, E. (2002). Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and Internet crime investigation. *Knowledge Solutions*. www.corpus-delicti.com/case_based.html

3

Data Mining

Revealing its origins and widespread use in business, data mining goes by many names, including knowledge management, knowledge discovery, and sense making.[1] Data mining is "[a]n information extraction activity whose goal is to discover hidden facts contained in databases."[2] 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. 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.

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 is 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 the Defense Advanced Research Projects Agency (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 9mm, 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. To relinquish the responsibility for analysis to outside organizations or consultants should be viewed in the same way as a suggestion to entirely contract patrol services to a private security corporation: an unacceptable abdication of an essential responsibility.

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.

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 twelve hours, 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.[3] 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. Post-mortem 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.[4] 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

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.[5] 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"[6]

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.[7] 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.[8] Given the nature of these crimes 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.[9] 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, data mining 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 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 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.

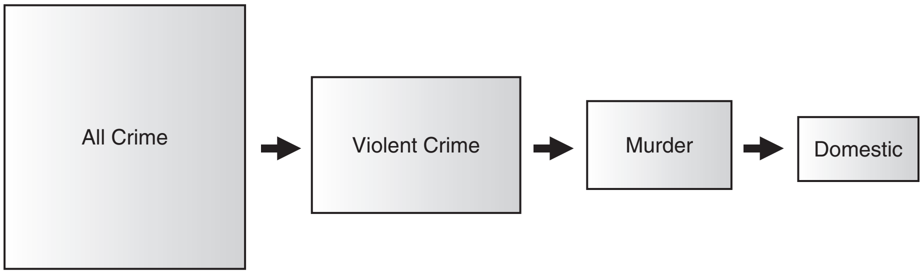
One particularly interesting market forecast indicates that an area of growth in the data mining field includes specialty niche markets.[10] Products developed for these niche markets will be tailored toward domain experts. These analytical tools will require less technical expertise and training, relying 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 promises to increase even further the availability of data mining and predictive analytics in the applied setting.

The first question might be: Do I need power tools for this? The answer is a most definite "maybe." Exploring the data to identify unique 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.

Figure 3-1 illustrates 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.

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

Figure 3-1 *Narrowing the focus on the data frequently can reveal smaller groups that are relatively similar in their attributes, which facilitates subsequent analysis.*



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.

Why is this important? In order to accurately characterize data, so as to reveal important associations and create accurate and reliable models, it is important 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.[11] 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 Analysis[12]

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 over-interpret 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.

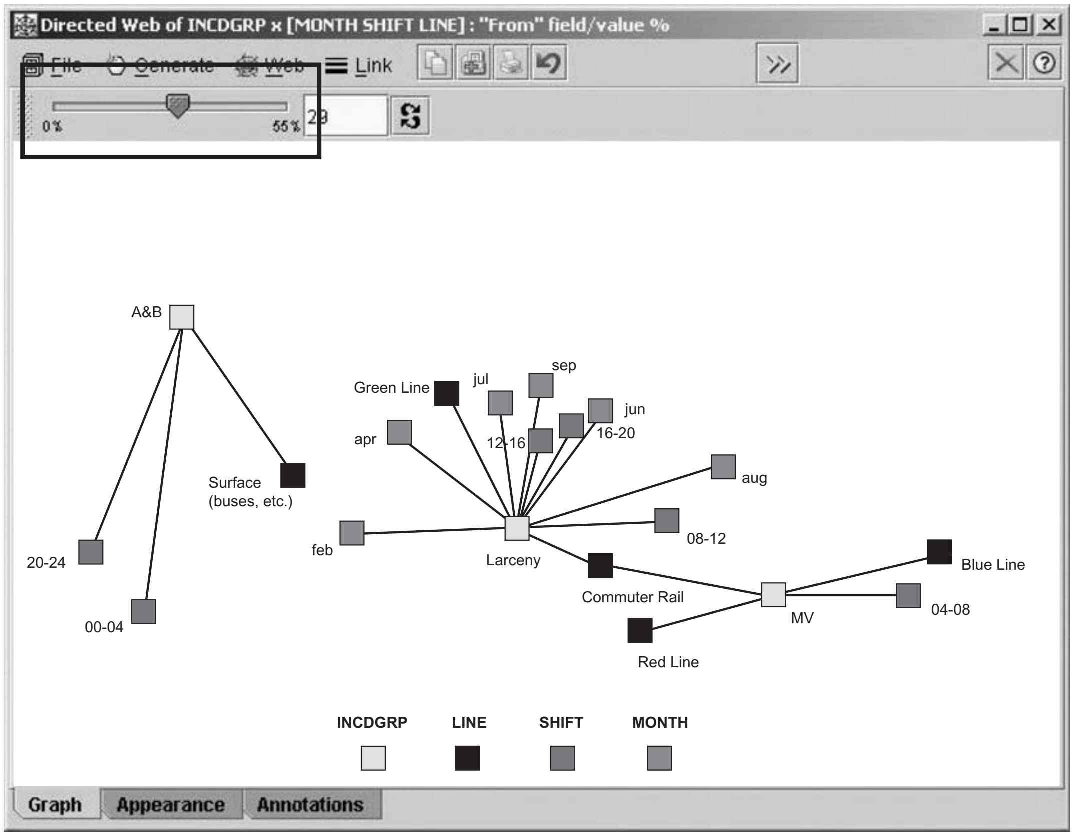
Many software tools provide a toggle option or sliding scale that gives the analyst the opportunity to adjust the thresholds used to determine the relative strengths of multiple relationships (Figure 3-2). This can be a tremendous tool, particularly during the exploration process, as it allows the analyst 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.

3.8 Nonobvious Relationship Analysis (NORA)[13]

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 name or attempt to vary her identity slightly in an effort to avoid detection. Similarly, Richard, Dick, Rick, Rich, Ricky, etc., 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

Figure 3-2 *This figure depicts a web graph. The box highlights the tool used to fluidly adjust the thresholds in an effort to reveal or mask certain differences. (B. Haffey, SPSS, Inc.; used with permission.)*



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.

Recently, however, automated techniques referred to as Nonobvious 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. While not available as "off-the-shelf" software products at the time of this writing, they can 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 identify 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 Future Trends

Many of the new data mining programs are very intuitive and also incredibly fast. This will facilitate their use in both the planning and the operational environments. Policy decisions now can become information-based through the use of these tools. This capability was tested at the 2003 International Association of Chiefs of Police annual conference, where data mining tools were used during a workshop on police pursuits. Initial findings were augmented and enhanced on scene by analysis in this live environment. Questions regarding the data were answered as quickly as they were raised through the use of these extremely quick and intuitive tools. The use of these analytical approaches in a live environment promises to put more science and less fiction into law enforcement policy and planning. In addition, the potential value of these tools in real-time, operational planning is tremendous. The ability to model possible scenarios and outcomes will not only enhance operational strategy and deployment but also can save lives as potential risks are identified and characterized.

3.11 Bibliography

1. Helberg, C. (2002). Data mining with confidence, 2nd ed. SPSS, Inc., Chicago, IL.
2. Definition from Two Crows (www.twocrows.com), which is an excellent source of accurate yet easy to understand information on data mining and predictive analytics.
3. Casey, E. (2002). Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and Internet crime investigation. *Knowledge Solutions.* www.corpus-delicti.com/case_based.html
4. McLaughlin, C.R., Reiner, S.M., Smith, B.W., Waite, D.E., Reams, P.N., Joost, T.F., and Gervin, A.S. (1996). Firearm injuries among Virginia juvenile drug traffickers, 1992 through 1994 (Letter). *American Journal of Public Health*, **86**, 751–752; McLaughlin, C.R., Smith, B.W., Reiner, S.M., Waite, D.E., and Glover, A.W. (1996). Juvenile drug traffickers: Characterization and substance use patterns. *Free Inquiry in Creative Sociology*, **24**, 3–10; McLaughlin, C.R., Reiner, S.M., Smith, B.W., Waite, D.E.,

Reams, P.N., Joost, T.F., and Gervin, A.S. (1996). Factors associated with a history of firearm injuries in juvenile drug traffickers and violent juvenile offenders. *Free Inquiry in Creative Sociology, Special Issue: Gangs, Drugs and Violence*, **24**, 157–165.

5. McCue, C., Smith, G.L., Diehl, R.L., Dabbs, D.F., McDonough, J.J., and Ferrara, P.B. (2001). Why DNA databases should include all felons. *Police Chief*, **68**, 94–100.
6. Tabussum, Z. (2003). CIA turns to data mining; www.parallaxresearch.com/news/2001/0309/cia_turns_to.html
7. www.KXTV10.com (2003). Despite avalanche of tips, police stymied in Laci Peterson case. January 9.
8. Eastham, T. (2002). Washington sniper kills 8, truck sketch released. October 12. www.sunherald.com
9. For a current review and comparison of specific data mining products, go to Elder Research, Inc. (www.dataminininglab.com).
10. METAspectrum^{SM} Market Summary (2004). Data mining tools: METAspectrum^{SM} evaluation. META Group, Inc.
11. McLaughlin, C.R., Daniel, J., Reiner, S.M., Waite, D.E., Reams, P.N., Joost, T.F., Anderson, J.L., and Gervin, A.S. (2000). Factors associated with assault-related firearms injuries in male adolescents. *Journal of Adolescent Health*, **27**, 195–201.
12. Helberg, C. (2002).
13. Franklin, D. (2002). Data miners: New software instantly connects key bits of data that once eluded teams of researchers. *Time,* December 23.

This Page Intentionally Left Blank

Methods

This Page Intentionally Left Blank

4

Process Models for Data Mining and Analysis

This chapter includes an overview of two complementary analytical process models: the Central Intelligence Agency (CIA) Intelligence Process[1] and the CRoss Industry Standard Process for Data Mining (CRISP-DM), [2] 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 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 repetition can serve to validate successful analysis and information-based operations. 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, predict, 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 is also 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.

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 homicide that can be defined by their motives (e.g., domestic, drug related, sexual).[3] 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 global violence injury prevention programs might be less than adequate if they do not address the unique constellation of risk associated with specific victim groups.[4] 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 is as much analytical process as it is math and statistics. In fact, the general rule is that the data mining process is 80:20—80% preparation and 20% analysis. The specific elements or tasks associated with the data mining process will be addressed separately and in greater detail in subsequent chapters; however, a general overview of these models is provided below 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 Process

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: Needs, Collection, Processing and Exploitation, Analysis and Production, Dissemination, and Feedback.[5]

Needs

During the needs or "requirements" phase,[6] 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.

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 modalities:[7]

1. 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).
2. Imagery Intelligence (IMINT)—IMINT includes information obtained through satellite, aerial, and ground-based collection methods.
3. 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. 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.
5. Open-Source Information (OSINT)—OSINT includes information available publicly and can include but is not limited to newspapers, radio, television, and the Internet.

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.

Analysis and Production

It is during the analysis and production phase of the process that raw data and information are converted into 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, which affords a greater depth of analysis and insight.

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 may be regular reports for events such as as the President's Daily Briefing; or they may reflect the results of a long-term study or analysis, such as the National Intelligence Estimates.

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.

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. 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 CRISP-DM

What the CIA model brings in terms of specificity to intelligence, and by extension applied public safety and security analysis, the 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.[8] The result of this task was the CRoss-Industry Standard Process for Data Mining (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. Data mining is as much analytical process as it is 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.[9]

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.

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.

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.

Modeling

During the modeling phase of the project, specific modeling algorithms are selected and run on the data. Selection of the specific algorithms employed in the data mining process is based on the nature of the question and outputs desired. For example, scoring algorithms or decision tree models are used to create decision rules based on known categories or relationships that can be applied unknown data. Unsupervised learning or clustering techniques are used to uncover natural patterns or relationships in the data when group membership or category has not been identified previously. These algorithms can be categorized into two general groups: rule induction models or decision trees, and unsupervised learning or clustering techniques. Additional considerations in model selection and creation include the ability to balance accuracy and comprehensibility. Some extremely powerful models, although very accurate, can be very difficult to interpret and thus validate. On the other hand, models that generate output that can be understood and validated frequently compromise overall accuracy in order to achieve this.

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?

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.

Summary

This model has worked very well for many business applications;[10] 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

Table 4-1 *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

nontechnical end users while being directly actionable in the applied setting in almost all cases. Finally, unlike in the business community, 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 both somewhat limited in directly addressing the unique challenges and needs related to the direct application of data mining and predictive analytics in the public safety and security arena. Therefore, an integrated process model specific to public safety and security data mining and predictive analytics is outlined below. 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.

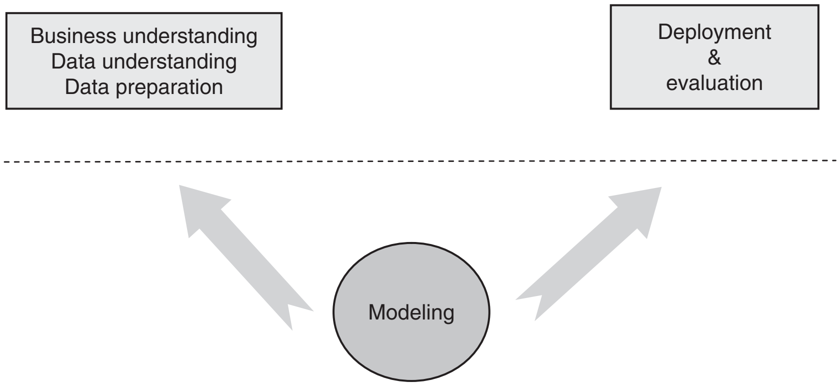


Figure 4-1 *In the CRISP-DM process model, the steps preceding and following the modeling phase require significant domain expertise and understanding of the operational requirements.*

4.3 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 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

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."[11] 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.

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 above, 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. 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.[12] 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.

Operationally Relevant Preprocessing

As mentioned above, 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.

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. Further, 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.

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.

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, which 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,[13] 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.

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 this limits 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 with 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.

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. 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.

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 only can 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.

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 above, 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.

Summary

The Actionable Mining and Predictive Analysis model presented above 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 4-2 compares the three analytical process models covered in this chapter.

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

Table 4-2 *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

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 Integrated 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 Integrated 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 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.

4.4 Bibliography

1. Office of Public Affairs, Central Intelligence Agency. (1993). A consumer's guide to intelligence. National Technical Information Service, Springfield, VA.
2. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (1999). CRISP-DM 1.0: Step-by-step data mining guide. http://www.crisp-dm.org
3. Douglas, J.E., Burgess, A.W., Burgess, A.G., and Ressler, R.K. (1997). Crime classification manual: A standard system for investigating and classifying violent crimes. Jossey-Bass, Hoboken, NJ.
4. Lord, W.D., Boudreaux, M.C., and Lanning, K.V. (2001). Investigation of potential child abduction cases: A developmental perspective. *FBI Law Enforcement Bulletin*, April.
5. 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 Process (http://www.fbi.gov/intelligence/process.htm), which is very similar to the

CIA model. For additional reading in this area, see: Lowenthal, M.M. (2003). Intelligence: From secrets to policy. CQ Press, Washington, D.C.

6. Lowenthal (2003).
7. CIA Handbook, and Air War College, Gateway to Intelligence (http://www.au.af.mil/au/awc/awcgate/awc-ntel.htm#humint).
8. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (1999). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium (www.crisp-dm.org).
9. Ibid.
10. For review, see, Piatetsky-Shapiro, G. (1999). CRISP-DM: A proposed global standard for data mining. *DS Star*, **3**, no. 15;www.taborcommunications.com/dsstar/99/0413/990413.html; KDnuggets (2002). What main methodology are you using for data mining? July; http://www.kdnuggets.com/polls/2002/methodology.htm; and Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, **5**, 13–22.
11. Chapman et al. (1999).
12. Faggiani, D., and McLaughlin, C.R. (1999). A discussion on the use of NIBRS data for tactical crime analysis. *Journal of Quantitative Criminology*, **15**, 181–191.
13. Goldstein, P.J. (1985). The drugs/violence nexus: A tripartite conceptual framework. *J Drug Issues*, **15**, 493–506.

5

Data

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. 79[1]

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. Until the Vulcan mind meld becomes a reality, it seems that the only way to acquire domain expertise is through experience and a hunger for a deeper understanding of "why."

As indicated in the above 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. Building a better mousetrap 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" and "how," 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.

In many ways, statistics and formal analytics are just a mathematical manifestation of how analysts think and function in their native state. As data mining tools become increasingly more accessible and easier to use, the need to teach the statistician will diminish, as those with intuitive analytical savvy and domain expertise will gain a controlling interest in the field. Data mining tools are highly intuitive and very visual. Good analysts, as well as many operational personnel, seem to gravitate toward them naturally, embracing the intuitive processing and actionable output. Data mining tools represent the hammers and chisels that these information artisans can use to reveal the underlying form and structure in the data. It is my expectation that the new data mining tools will change dramatically what analytical positions in the future look like and how they function, as well as the role that operational personnel will play in analysis and how they will work with the analytical staff. As data and information assume the role of a seamless interface between analytical and operational personnel, I anticipate the emergence of an "agent analyst" class of personnel who can function in both worlds, utilizing data and intelligence as integral tools in the operational environment, while guiding the analytical process from collection to output.

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 in the lab as a scientist, one of the most important lessons that I learned was to know your subjects. The same is true for 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, or observing suspect interviews, the more that the analyst can understand about where the data and information came from, 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 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 Data[2]

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 or 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 degrees Fahrenheit is the same as the difference between 180 and 190 degrees: 10 degrees. Ordinal scales imply some ranking in the information. Though 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.

5.4 Types of Data Resources

Working with the Operators to Get Good Data

The business community provides some guidance on 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. 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.

This same principle can be applied in crime analysis data collection. One frequently voiced frustration from sworn personnel, particularly those working in patrol, 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, but they frequently confer benefits only to the 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 frequently 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.

It is also 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,[3] 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 fifty 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.

Records Management Systems

Most departments maintain large records management systems 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 return. 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.

Ad Hoc or Self-Generated Databases

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.[4] Similarly, at the height of the Washington D.C. sniper investigation, authorities received as many as 1000 tips an hour.[5]

The sheer volume of information associated with these cases requires some sort of automated system of information management.

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 requires 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.

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.

Additional Data Resources

The analyst is likely to encounter two types of data that have their own sets of issues and challenges: relational data, like those associated with incident-based reporting and police calls for service, and dispatch data. These data are used frequently in public safety analysis; therefore, their unique features and challenges are worth addressing.

Relational Data

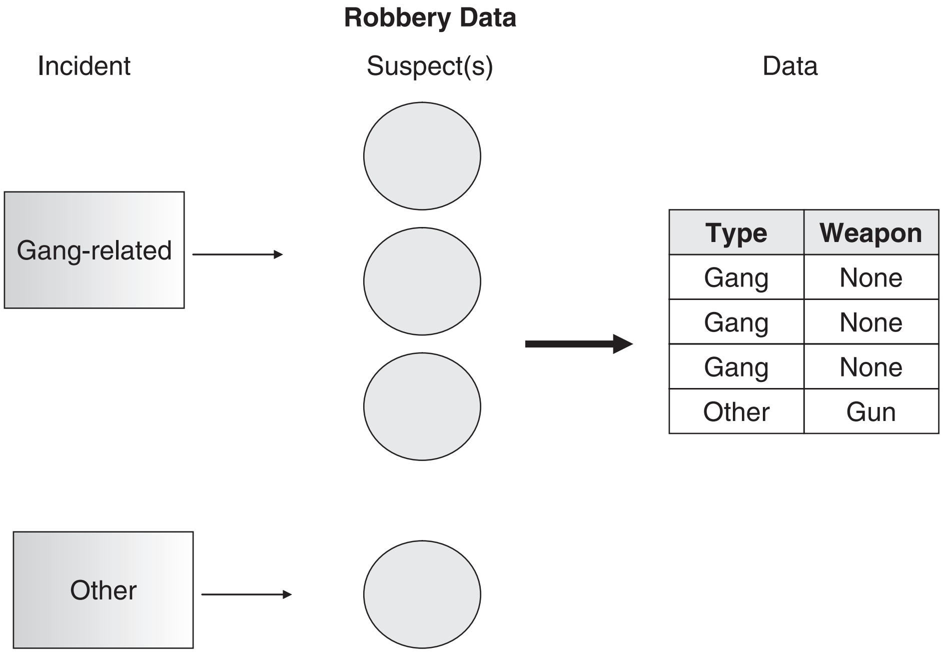
Relational databases are comprised of a series of associated or linked tables.[6] This facilitates the inclusion of a greatly expanded amount of information; however, it can create some challenges in analysis and interpretation of data, as outlined below. An example of a relational database that is frequently encountered in law enforcement and public safety is the National Incident-Based Reporting System, or NIBRS.[7] 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 5-1, a gang robbery 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

Figure 5-1 *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 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.*



Type	Weapon
Gang	None
Gang	None
Gang	None
Other	Gun

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.

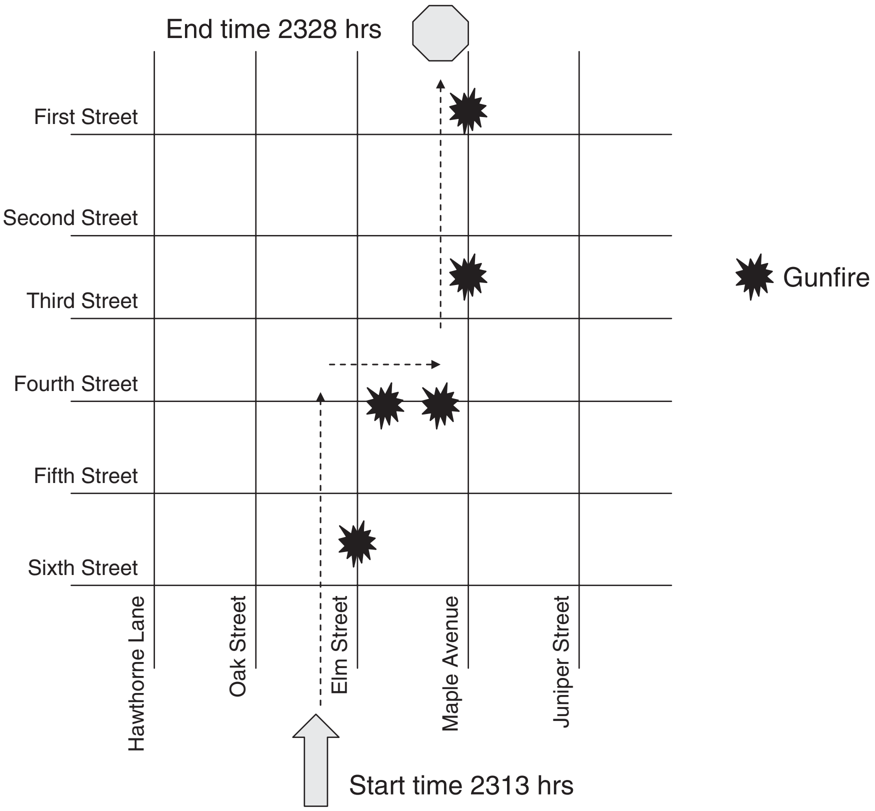
Dispatch Data

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 5-2 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.

Figure 5-2 *This street map illustrates a vehicle driving through a neighborhood engaging in random gunfire.*



As can been seen on the spreadsheet in Table 5-1, the calls start coming into the dispatch center almost immediately. Three separate calls come in within 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 eleven calls from citizens reporting random gunfire (Table 5-2), which are aggregated into seven distinct calls,

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
02-27-2004	2313	545 Elm St	Shots fired	258
02-27-2004	2314	545 Elm St	Shots fired	258
02-27-2004	2315	545 Elm St	Shots fired	258
02-27-2004	2315	200 W 6th St	Shots fired	259
02-27-2004	2318	150 W 4th St	Shots fired	260
02-27-2004	2320	148 W 4th St	Shots fired	260
02-27-2004	2321	50 W 4th St	Shots fired	260
02-27-2004	2326	242 Maple Ave	Shots fired	261
02-27-2004	2328	242 Juniper St	Shots fired	262
02-27-2004	2329	105 Maple Ave	Shots fired	263
02-27-2004	2345	545 Elm St	Shots fired	264

Table 5-2 *In this table, the eleven citizen complaints from a single series of random gunfire have been aggregated into seven 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
02-27-2004	2313	545 Elm St	Shots fired	258
02-27-2004	2315	200 W 6th St	Shots fired	259
02-27-2004	2318	150 W 4th St	Shots fired	260
02-27-2004	2326	242 Maple Ave	Shots fired	261
02-27-2004	2328	242 Juniper St	Shots fired	262
02-27-2004	2329	105 Maple Ave	Shots fired	263
02-27-2004	2345	545 Elm St	Shots fired	264

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 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.[8]

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.

Other Considerations

Another important feature in this discussion is that not all crime is 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 most open-air drug markets are more likely to be located in higher-risk areas. 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 that expired city decal or burned-out headlight. 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.

5.5 Data Challenges

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.

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.

Intentional Misrepresentation (aka Criminals Often Lie)

Criminals often misrepresent the facts, both when it matters and when it does not. For that matter, so do witnesses, informants, and many other folks that we use on a regular basis to help us to gather information and data. This is not unique to law enforcement or intelligence data. Responders to marketing surveys lie as well. Consider the last time that you were completely honest and completed every question on a survey, if you even responded to it in the first place. 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? Similarly, other people withhold information because they have something to hide, are concerned that they might implicate themselves or someone else, or just because. All of this affects the reliability and particularly the validity of the data that we encounter.

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.

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.

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.

5.6 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 can be extremely valuable, particularly regarding commonly occurring patterns associated with false allegations.

But knowing certain data is false does not mean it 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 side effects. Generally those that you are not aware of create the most trouble in the end.

5.7 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 one-half by removing the unnecessary duplication in the data.

5.8 Merging Data Resources

The emergence of regional fusion centers throughout the country represents an analytical windfall for crime analysts. Patterns that transcend jurisdictional boundaries can be identified and analyzed by merging data from neighboring or otherwise related communities. This approach can increase the number of observations, which can be particularly important with rare events. It also can identify other patterns of offending and criminal groups that exploit jurisdictional boundaries. Examples of this may include traveling scam artists, large criminal gangs, and even serial killers like Ted Bundy with his expansive geographic range. By spreading their crimes across a large geographic region that transcends local jurisdictional boundaries, these criminals can elude detection and capture. Groups like the Mid-Atlantic Regional Gang Investigators Network (MARGIN) and the FBI's VIolent Crime Apprehension Program (VICAP) have been developed in an effort to identify patterns that cross jurisdictions and address this issue.

Data also can be aggregated across collection modalities, reflecting the fact that crime frequently does not exist in one domain. For example, through the analysis of telephone records one might be able to characterize the calling patterns of a drug dealer as he sets up buys and deliveries, while an analysis of banking records and deposits might be necessary to extrapolate cash flow and economic value of the products that he is moving. Being able to analyze all of this information in the same analytical environment starts to develop the big picture, which represents a powerful tool available to the analyst. The following sections provide a few examples that highlight the value of this approach.

5.9 Public Health Data

Be creative. One thing that I realized several years ago through the Cops & Docs program[9] was that when new drugs enter a market, people generally get into trouble medically before they get into trouble legally. Analyzing health care and law enforcement-related drug overdose data in the same environment provides a more comprehensive view of a pattern of behavior that transcends the health care and legal systems. Clearly, there are both legal and ethical issues associated with this type of analytical endeavor, but they are not insurmountable. Public health and public safety can be complementary functions, particularly if data and information are shared in a meaningful fashion.

One particular example where this worked extremely well was with trauma patients admitted after being involved in car wrecks. During informal conversation, the health care providers noted that several of the recent admissions for car wrecks were noteworthy in that the patients tested positive for illegal narcotics. A quick tutorial on the finding that many heroin addicts use their drugs almost immediately after purchasing them enlightened the health care providers regarding substance use patterns and their direct link to trauma. This resulted in a comprehensive analysis of health care and law enforcement-related narcotics data in the same analytical environment. Although the health care information was zip code–based and relatively nonspecific due to medical confidentiality requirements and investigative sensitivity, the resulting value-added analysis of the multidisciplinary narcotics data highlighted a population that had been underrepresented in other analyses. Ultimately, both the health care and law enforcement professionals involved benefited from this information exchange, as did the communities experiencing these increases in illegal drug use.

5.10 Weather and Crime Data

Normal seasonal weather trends, as well as dramatic changes, can have an impact on crime trends and patterns. 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. Analyzing crime and weather data in the same analytical environment can add significant value to police operational planning and crime prevention.

Data mining and predictive analytics tools are perfectly designed to fully exploit the possibilities associated with this type of novel analysis.

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. While largely anecdotal, this highlights the value that nontraditional crime measures 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 some 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.

5.11 Bibliography

1. Lowenthal, M.M. (2000). Intelligence: From secrets to policy. CQ Press, Washington, D.C.
2. Howell, D. (1992). Statistical methods for psychology, 3rd ed. Duxbury Press, Belmont, CA. 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.
3. McCue, C., Parker, A., McNulty, P.J., and 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.
4. www.KXTV10.com. (2003). Despite avalanche of tips, police stymied in Laci Peterson case, January 9.
5. Eastham, T. (2002). Washington sniper kills 8, truck sketch released. October 12. www.sunherald.com
6. Agosta, L. (2000). The essential guide to data warehousing. Prentice Hall, Upper Saddle River, NJ.

7. 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.

8. Sampson, R.J., Raudenbush, S.W., and Earls, F. (1997). Neighborhoods and violent crime: A multi-level study of collective efficacy. Science, **277**, 918–924.

9. McCue, C. (2001). Cops and Docs program brings police and ED staff together to address the cycle of violence. *Journal of Emergency Nursing*, **27**, 578–580.

6

Operationally Relevant Preprocessing

This chapter covers operationally relevant preprocessing, including a specialized analytical framework called Trinity Sight, which supports integrated analysis of time, space, and behavioral indicators or threats. 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 it.[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 Trinity Sight

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. In response to the unique importance of these three factors, our team at RTI International has developed the Trinity Sight analytical model. Trinity Sight is a special application of the Actionable Mining and Predictive Analysis model presented in Chapter 4 and includes an integrated analysis of time, space, and the nature of the incident or potential threat. As always, the emphasis is on operationally meaningful, actionable output, which frequently begins with operationally relevant preprocessing.

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 hours than 0258 hours +/− 53 minutes. Four- to eight-hour time blocks work well for deployment analysis. Personnel generally do not work four-hour shifts, but using a four-hour level of analysis does afford some flexibility, as eight- or twelve-hour shifts can be overlapped to provide additional coverage for time periods associated with greater anticipated workload. Shorter than four-hour 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, 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 eight hours 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 four-hour period that spanned from 10:00 pm on New Year's Eve to 2:00 am on New Year's Day. While it might be attractive from a cost standpoint to craft a four-hour 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 four hours. 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 eight-hour shifts, twelve-hour 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. Similarly, events such as concerts and sporting events 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.

Space

The advancement of 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. 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 analyst 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 an analysis of unusual or suspicious behavior on a moving location like a commuter train or airplane. For example, several of the 9/11 highjackers 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.

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. 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 rather than replace. In some situations, the focus will change over time as the suspect refines his 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.

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.[3] 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.[4] 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 its 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.

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.

6.3 Duplication

As mentioned in the previous chapter, 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 is 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 6-1, 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 6-1, information can be added through recoding, as shown in Table 6-2.

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

Table 6-1 *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 6-2 *Example of additional telephone number recoding.*

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

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 above, 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.

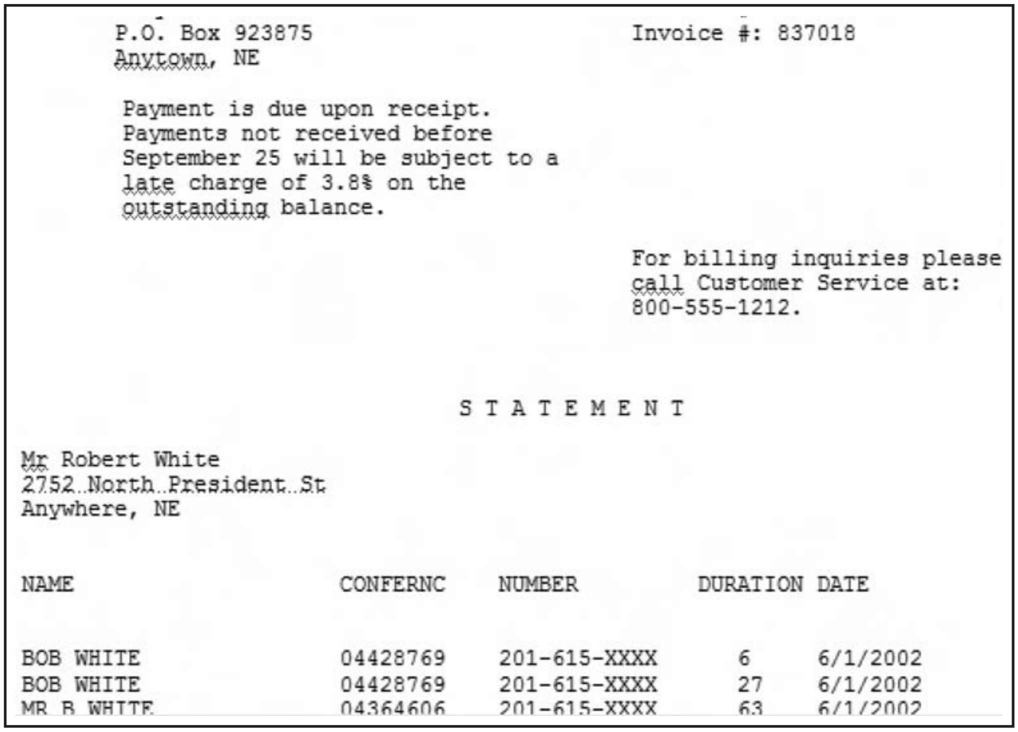
The following example is based on real case materials. While many of the details have been changed, the analytical approach and techniques are identical.

The local police department received a thirty-seven-page invoice that was associated with a large, unpaid bill (Figure 6-1). 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 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 6-2). After this information had been removed, the resulting document included the conference IDs (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

Figure 6-1 *Billing invoice that was associated with a large, unpaid bill.*



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.

STATEMENT

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 B WHITE	04364606	201-615-XXXX	63	6/1/2002

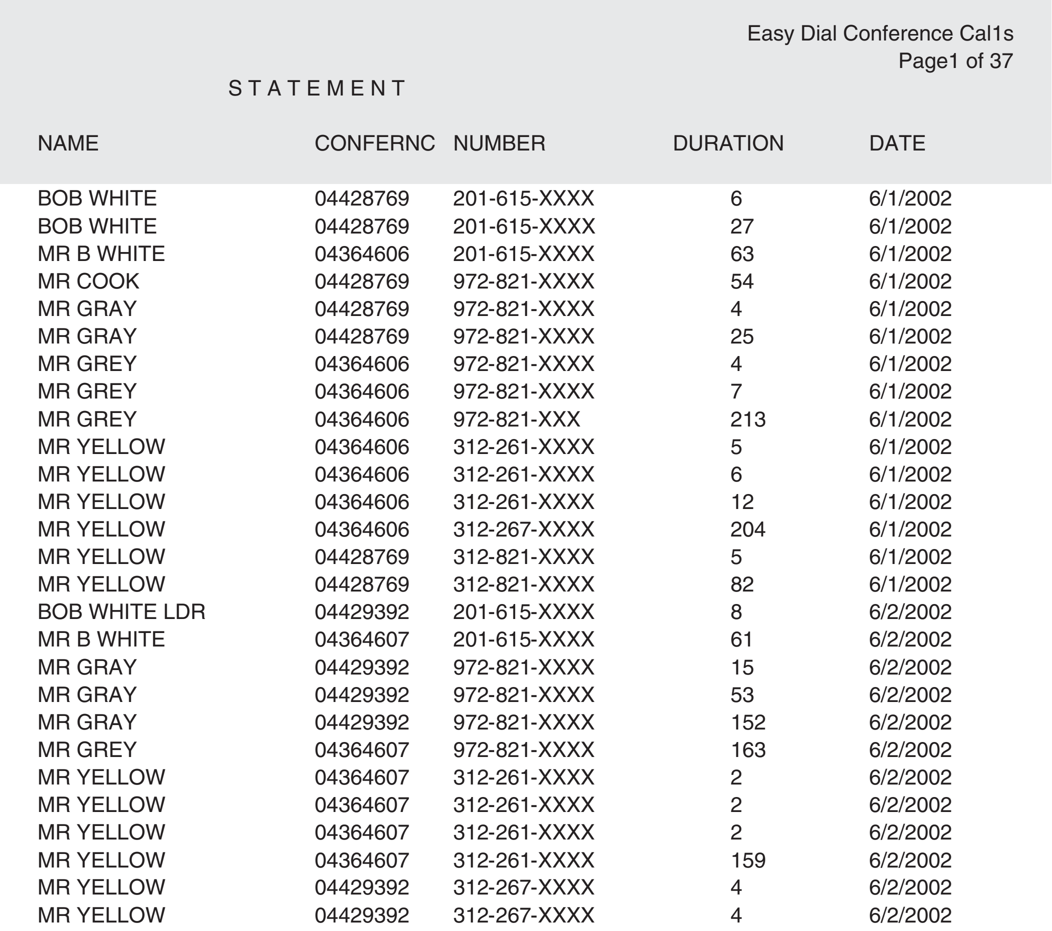
were fraudulent, they were retained in the data because they could be used for additional linking.

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 6-3).

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 timelines and sequencing, while the day of the week can reveal other patterns (Figure 6-4).

An initial review of the data indicated 2,017 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.

Figure 6-2 *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.*



Easy Dial Conference Cal1s
Page1 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-XXX	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

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.

Figure 6-3 *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	NUMBER	DURATION	DATE
BOB WHITE	04428769	201	615XXXX	6	06/01/2002
BOB WHITE	04428769	201	615XXXX	27	06/01/2002
MR B WHITE	04364606	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 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
BOB WHITE LDR	04429392	201	615XXXX	8	06/02/2002
MR B WHITE	04364607	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 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	04429392	312	267XXXX	4	06/02/2002
MR YELLOW	04429392	312	267XXXX	4	06/02/2002
MR YELLOW	04429392	312	267XXXX	5	06/02/2002
MR YELLOW	04429392	312	267XXXX	55	06/02/2002

As can be seen in Figure 6-5, duplicate numbers within a unique conference call were deleted, while duplicated numbers across different conference calls were retained. Culling the duplicates revealed 1,042 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.

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

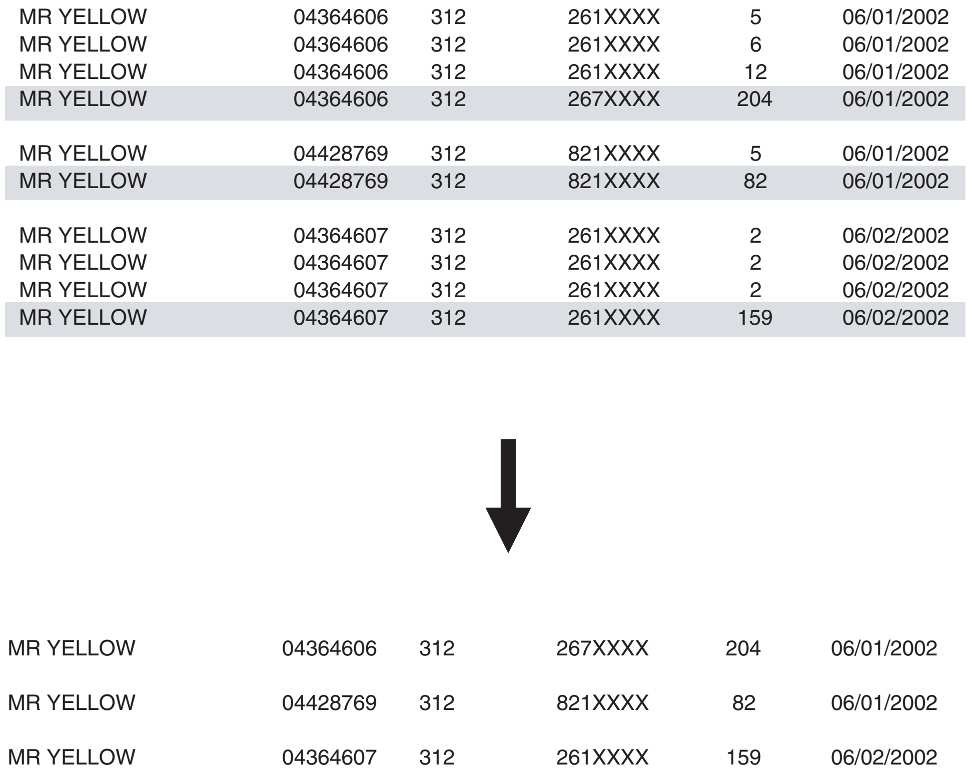
Figure 6-4 *An additional variable was re-created, which indicated the country or general geographic area associated with the recoded country and area codes. The date also was converted into day of the week.*

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

the conference occurred and the number of participants involved in a particular call (Figure 6-6). 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 continued until the end of the month, at which time it ceased abruptly. Termination of

Figure 6-5 *The data were culled to remove some duplicative calls that were determined to be unnecessary for subsequent analyses.*



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

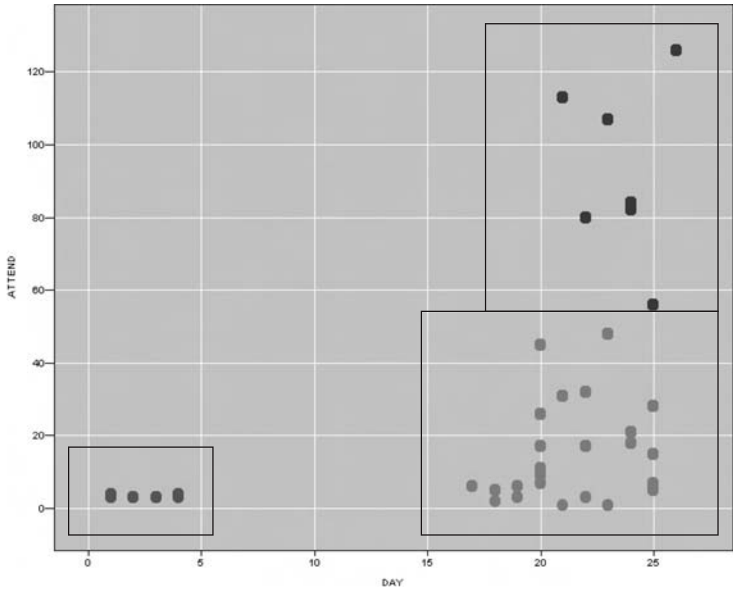
MR YELLOW	04364606	312	267XXXX	204	06/01/2002
MR YELLOW	04428769	312	821XXXX	82	06/01/2002
MR YELLOW	04364607	312	261XXXX	159	06/02/2002

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,

Figure 6-6 *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. (Screenshot of Two-Step output taken by the author is from Clementine 8.5; SPSS, Inc.)*



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 compete 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 IP address or browsing patterns. Although not perfect, it can represent a viable option for further analysis of suspicious activity.

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 affect 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.[5] 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. 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.

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 an analysis of drug-related violence, 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.

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. 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.[6] Knowledge of the suspect's substance use patterns and criminal history would require knowing the suspect's identity. If we knew who did it, then we could just ask them why. Generally, identification of a specific motive is used to direct the investigation toward a possible suspect, rather than the other way around. 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. Variable selection will be revisited during the modeling process.

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.

Variable selection 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 also come into play. In the movie *Apollo 13,* the engineers are tasked with creating a system that will make two incompatible carbon dioxide scrubbing filters work together. The only catch is that they can only use parts already aboard the spacecraft. The team leader walks into a conference room with a box of assorted items that they might use to create this system, and later walks out with an ugly contraption that utilizes miscellaneous junk and a fair amount of duct tape—but it works. It certainly is not elegant or state of the art, but it gets the job done. Analysts are faced with a similar task, particularly when analyzing information associated with the investigative process. Frequently, we are given a box of seemingly useless information and are asked to create something that will assist in the investigative process. At times, I have discussed my analytical work within the operational environment with academic researchers and felt like the proverbial country bumpkin, with my shoddy little models cobbled together with the equivalent of spare parts and duct tape. But that frequently is what is required in the applied setting. 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.

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. Although 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.[7] As always, common sense and judgment is an excellent partner to probable cause and ethics in the analytical process.

6.9 Bibliography

1. Helberg, C. (2002). Data mining with confidence, 2nd ed. SPSS, Inc., Chicago, IL.

2. Howell, D. (1992). Statistical methods for psychology, 3rd ed. Duxbury Press, Belmont, CA.

3. McCue, C., Smith, G.L., Diehl, R.L., Dabbs, D.F., McDonough, J.J., and Ferrara, P.B. (2001). Why DNA databases should include all felons. *Police Chief*, **68**, 94–100.

4. McCue, C. (2004). Lecture presented to Diplomatic Security Service personnel at U.S. Department of State (ArmorGroup, International Training), Rosslyn, VA, May 14, June 25.

5. Wellford, C. and Cronin, J. (2000). Clearing up homicide clearance rates. *National Institute of Justice Journal*, **243**, 1–7.

6. McLaughlin, C.R., Daniel, J., and Joost, T.F. (2000). The relationship between substance use, drug selling and lethal violence in 25 juvenile murderers. *Journal of Forensic Sciences*, **45**, 349–353.

7. See Gould, S.J. (1981). The mismeasure of man. WW Norton & Company, New York; and Lewontin, R.C., Rose, S., and Kamin, L.J. (1984). Not in our genes. Pantheon Books, New York.

This Page Intentionally Left Blank

7

Predictive Analytics

In their Worldwide End-User Business Analytics Forecast, IDC, a global provider of market intelligence, divided the market and distinguished between "core" and "predictive" analytics.[1] Core analytics are described as those tools that "analyze a current or past state," generally focusing on descriptive statistics, query, and reporting. Predictive analytics, on the other hand, "are used to determine the probable future outcome of an event or the likelihood of a current state where it is unknown." These tools incorporate more sophisticated analytical strategies, and include data mining and modeling.

By using data mining, we can begin to characterize and describe trends and patterns that reside in data and information. While this might help us gain a better understanding of crime or intelligence data, it is limited in terms of the actionable intelligence or predictive models that can be obtained. To be able to anticipate, predict, and ultimately prevent bad things from happening in the world, we need to be able to develop accurate and reliable models. Predictive analytics encompasses a variety of model making tools or algorithms that can be employed to characterize historical information, which then can be used to predict the nature and likelihood of future events or occurrences. In other words, although the focus has been on "connecting the dots," that is only part of the task. What we really need to do is to connect the dots and use the resulting picture to predict the next image or scenario so that we will have a chance to alter the future and prevent something bad.

7.1 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, 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 be 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. But confining analysis to only what is available *because* it is available is probably just as inappropriate. Relegating analysis to last place 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 Chapter 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. Therefore, the savings associated with information-based decisions and shared resources can be used to expand analytical capacity.

7.2 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 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 telephone call data. As can be seen in Figure 7-1, which illustrates an analysis of the conference call data covered in Chapter 6, associations or links can be depicted visually to help illuminate particular relationships in the data. As shown in Figures 7-2 and 7-3, expert settings in some packages allow the analyst to highlight different relationships or examine some associations in greater detail. 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.

7.4 Supervised versus Unsupervised Learning Techniques[2]

While this is somewhat simplistic, modeling algorithms can be divided generally into supervised and unsupervised learning techniques. 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

Figure 7-1 *Example of a sophisticated link analysis tool (Advizor Solutions, Inc.).*

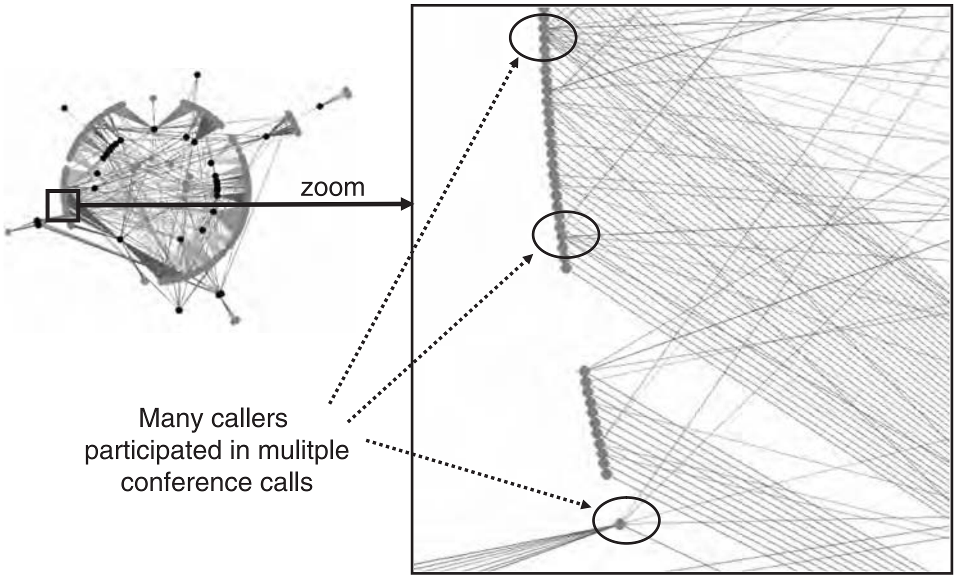
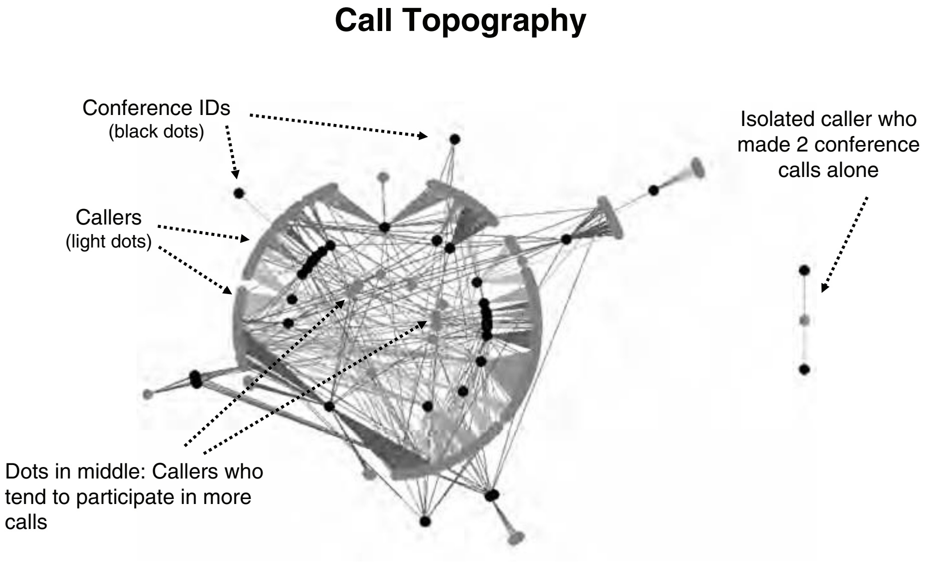
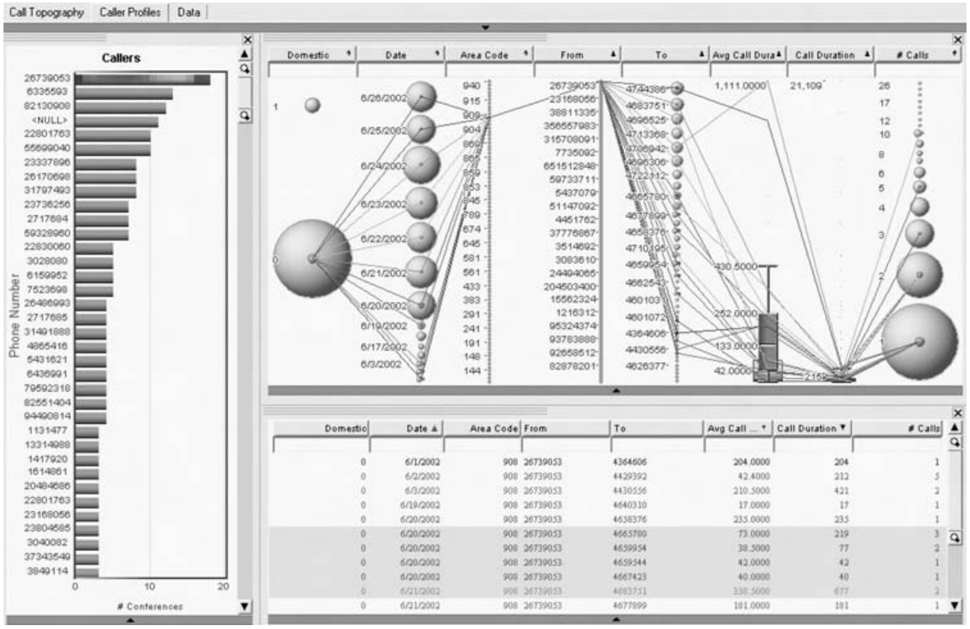


Figure 7-2 *This example illustrates the ability to dissect specific attributes of a link analysis and examine specific associations and relationships in greater detail (Advizor Solutions, Inc.).*


Call Topography
Conference IDs (black dots)
Isolated caller who made 2 conference calls alone
Callers (light dots)
Dots in middle: Callers who tend to participate in more calls

Links:
- Represent participation in a conference call by a caller *(Note: Duplicates removed)*
- Lighter shading ⇒ more recent conference calls

Figure 7-3 *This figure illustrates a "profile" of the top caller in context of all other callers. This caller participated in more conference calls throughout the period (almost daily), and accounted for some of the highest duration calls (Advizor Solutions, Inc.).*



Domestic	Date ▲	Area Code	From	To	Avg Call ... ▼	Call Duration ▼	# Calls
0	6/1/2002	908	26739053	4364606	204.0000	204	1
0	6/2/2002	908	26739053	4429392	42.4000	212	5
0	6/3/2002	908	26739053	4430556	210.5000	421	2
0	6/19/2002	908	26739053	4640310	17.0000	17	1
0	6/20/2002	908	26739053	4658376	235.0000	235	1
0	6/20/2002	908	26739053	4665780	73.0000	219	3
0	6/20/2002	908	26739053	4659954	38.5000	77	2
0	6/20/2002	908	26739053	4659544	42.0000	42	1
0	6/20/2002	908	26739053	4667423	40.0000	40	1
0	6/21/2002	908	26739053	4683751	338.5000	677	2
0	6/21/2002	908	26739053	4677899	181.0000	181	1

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 numeric as well as symbolic data. Rule induction models that can accommodate continuous data still end up parsing them into categories by identifying breaks, or establishing "cut points," in the range.

7.5 Discriminant Analysis

Discriminant analysis is covered in more detail in Chapter 11. Briefly, one of the assumptions of this model is that the data are categorical. This assumption can 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, and it is almost always better for the analysis to come up empty than to identify a spurious or false relationship in the data.[3] 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 carry a far greater risk to public safety in most situations.

7.6 Unsupervised Learning Algorithms

Unsupervised learning algorithms are used to group cases based on similar attributes. These models also are referred to as self-organizing maps. Unsupervised models include clustering techniques and neural networks. 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 difficult to interpret.

Cognitive Neuroscience and Neural Nets

> *"My religion consists of a humble admiration of the illimitable superior spirit who reveals himself in the slight details we are able to perceive with our frail and feeble mind."*
>
> Albert Einstein

For as long as I can remember, I have been fascinated by science and the wonders of the universe. From stargazing in the backyard with the homemade telescope that my father and I built to the absolute awe that I experience when contemplating the vastness of the cosmos and the subtle elegance of nature, I have been hooked on science from the start.

During college I began to focus my interest on neuroscience and the brain. What an incredible machine! As I sit here now I can recall the muffled quiet of my first snowfall at Dartmouth, the sound of a lawnmower running on a Saturday morning from my childhood in Downers Grove, Illinois, and the smell of fresh cut grass. I can see the windows steam up in our kitchen on Thanksgiving, and smell my mother's turkey, which I never have been able to replicate. The truly amazing thing about all of this, though, is that all of these memories, including their associated sights, smells, and sounds, reside in a mass of biological material sitting between my ears that basically has two settings: on and off. Some might argue that neuromodulators and other similar entities complicate the situation somewhat, but the bottom line is that neurons, the basic components of our brains, are either on or they are off. Like a computer, it is this combination of "on" and "off," the interconnectedness of these simple elements and the associated parallel processing, that gives us the complexity of what we know to be brain function.

While I do not necessarily hold the conviction with Descartes that the seat of my soul resides somewhere at the base of my brain, I do know that everything from unconscious activities like breathing to my preference for the color green sits up there with rarely a conscious thought from me. More to the point, I know that the individual differences that make the world so interesting, as well as the similarities both between and within humans and their behavior that allow me to do my job as a behavioral scientist, also reside in this neural computer.

Analysts spend a considerable amount of time trying to categorize and model the complexities of human behavior. This practice is complicated even further for crime analysts because the behavior being modeled differs in some way from "normal" behavior, if only for the reason that it is illegal. In additional, criminal behavior tends to be relatively infrequent and is something that most folks have limited experience with outside of the public safety and security worlds. The ability to reduce these behaviors to patterns and trends that can be not only described but even anticipated or predicted in some situations still amazes me because it says as much about human nature as it does about analysis. In many ways, predictive analytics and artificial intelligence are fascinating in their power and complexity, but perhaps the real wonder is the fact that human behavior can even be modeled and predicted at all.

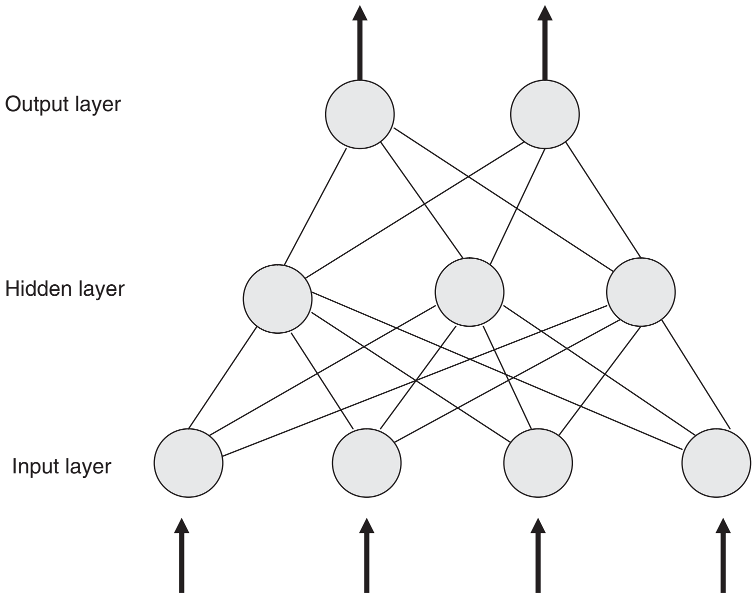
The most incomprehensible thing about the world is that it is comprehensible.
Albert Einstein

7.7 Neural Networks

One of the most fantastic things about nature is her ability to create complexity with unique combinations of a few simple elements. With a few exceptions and some variation, the entire complexity of 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, 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 "Six Degrees of Kevin Bacon" game. 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 7-4 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 increases the possible numbers of connections and associated complexity of the model. This ability to layer

Figure 7-4
Simplified neural net model.



connections adds a tremendous degree of additional complexity 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 basic elements, these sophisticated algorithms can be used to model extremely complex associations and relationships.

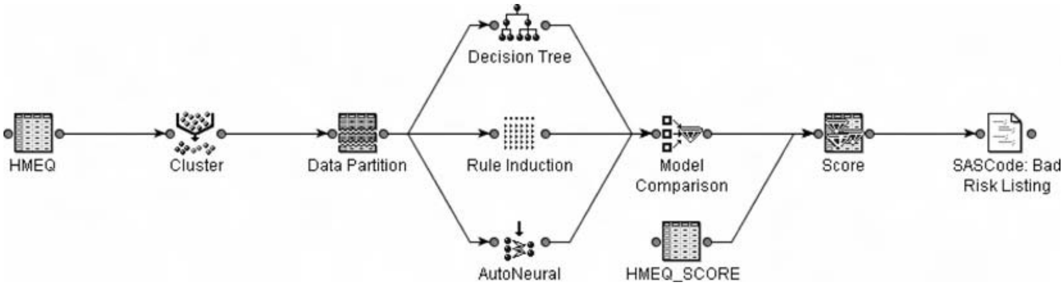
7.8 Kohonan Network Models

Kohenan 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 preexisting 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.

7.9 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 are 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

Figure 7-5 *This figure highlights a model comparison feature that can be used to compare directly different modeling approaches or strategies. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



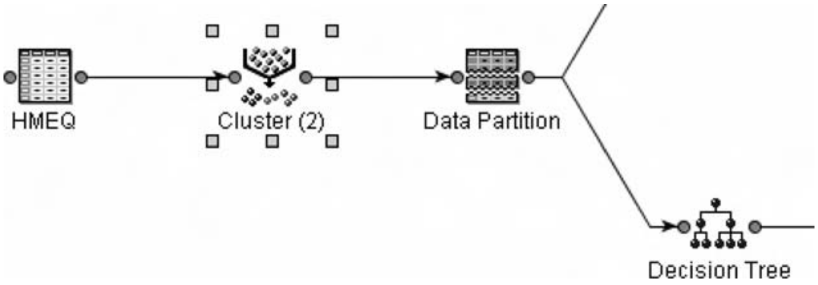
one way in an effort to find the analytical approach or algorithm that works best. It is unlikely that one particular modeling technique will emerge as a runaway leader, but subtle differences are not only possible, they are expected. As can be seen in Figure 7-5, the Enterprise Miner includes a feature 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 above.

7.10 Combining Algorithms

Different modeling algorithms also can be used in sequence. For example, the analyst can use unsupervised approaches to explore the data. If an interesting group or relationship is identified, then a supervised learning technique can be developed and used to identify new cases. An example of this approach can be seen in fraud detection.

A similar combined approach includes the direct, sequential use of unsupervised and supervised learning algorithms in the same analysis. Using this approach, the unsupervised learning approach is run against the data. The group assignment generated by the clustering algorithm then becomes a variable that is used in the decision tree algorithm. A sample of this approach is shown in Figure 7-6.

Figure 7-6 *This figure illustrates the sequential use of unsupervised and supervised algorithms in the same analysis. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



7.11 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.

7.12 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 anomaly detector has gone off, although they probably will use less delicate terminology.

Sometimes an investigator's internal anomaly detector will point out that things are just outright strange. 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.[4] By accumulating a veritable 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 bear further scrutiny.

7.13 Defining "Normal"

People often get tripped up and caught when they try to behave normally or "fly under the radar." 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. It is difficult to completely understand what normal looks like until we characterize it and then analyze it in some detail. 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.

How It Works

Briefly, there is a variety of clustering algorithms that group cases based on similarities between them.[5] 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 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 7-7, 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.

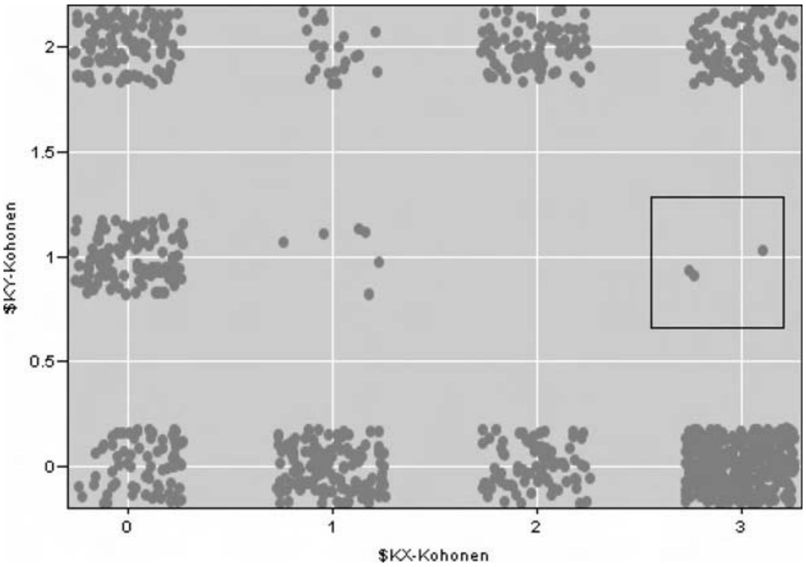


Figure 7-7 *This figure illustrates the results of a clustering algorithm. The box highlights three anomalous cases in the data set, which do not fit into any of the other clusters. (Screenshot taken by the author is from Clementine 8.5; SPSS, Inc.)*

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.

7.14 Deviations from Normal Patterns

It is almost always easier and more effective to respond to an emerging trend or pattern than to one that is established. The Regional Crime Analysis Program (RECAP) developed by researchers at the University of Virginia includes automated control charting.[6] Briefly, this control-charting function plots average frequencies of crime. Thresholds for unacceptable deviation from average frequencies can be preset. The analyst is alerted if the frequency transcends this preset threshold, which indicates that further inquiry into the change in frequency is warranted. This program can be set to run during the evening, providing any necessary alerts to the analytical personnel each morning.

The Cops & Docs program, developed by the Richmond, Virginia, Police Department in collaboration with local health care providers, has created a similar program for drug-related incidents.[7] Although it is not automated at this point, local health care providers and law enforcement personnel share drug-related incident information in recognition of the fact that this type of information frequently transcends both professional domains. Some drug overdoses result in a telephone call to the local poison control center; others end up in the emergency department; others go directly to the morgue. Consequently, health care providers and law enforcement personnel have only limited access to the entire array of drug-related incidents. By sharing information within appropriate legal and ethical guidelines, even through informal communication routes, the ability to compile and monitor drug-related incidents in an effort to detect deviations from normal is increased significantly. Although involving brute force techniques, this method has proven to be relatively effective in identifying increases in drug-related incidents of potential concern to both health care and law enforcement providers.

7.15 Deviations from Normal Behavior

The relationship between property crimes and stranger rapes has been discussed in other chapters. The salient feature for anomaly detection, however, 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. Upon subsequent examination, there frequently is additional secondary gain associated with these crimes that indicates the potential for significant escalation into violent crime, particularly sexually violent crime.[8]

Staged Crimes

The subject of "normal" crime is addressed in Chapter 10. 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 7-8, 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. 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.

Further review of the crime scene was consistent with the story. There was a knife on the right side of the victim's body, next to her left hand. Left-handedness is relatively infrequent within the population, and a relatively easy mistake in

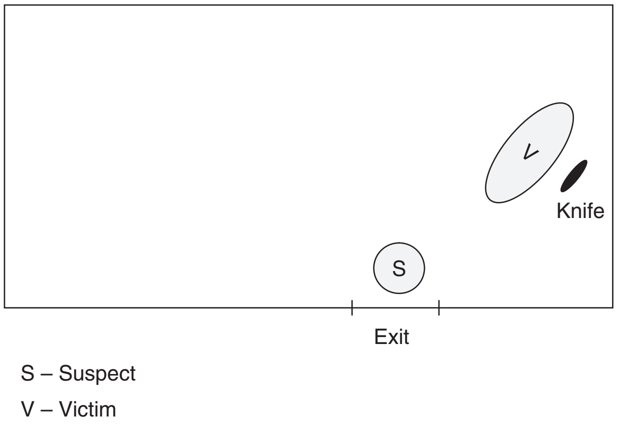


Figure 7-8
Scene diagram of a staged crime.

staging a crime is 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. One tipoff to an embezzlement case 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 $100 was unusually high, particularly compared to what we would expect for usual expenses.

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 twice within such a short period of time—particularly abductions associated with unusual circumstances. In another example, we were able to trace several individuals involved in suspicious behavior around a critical facility to addresses that were within walking distance from one another. While not definitive indicators of wrongdoing, most investigators and analysts tend to become extremely suspicious when encountering coincidences like these.

7.16 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.

7.17 A Perfect World Scenario

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,[9] while the Iraqis were researching psyops and related topics in the days prior to the most recent Gulf War.[10] 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 7-9). Clearly, this

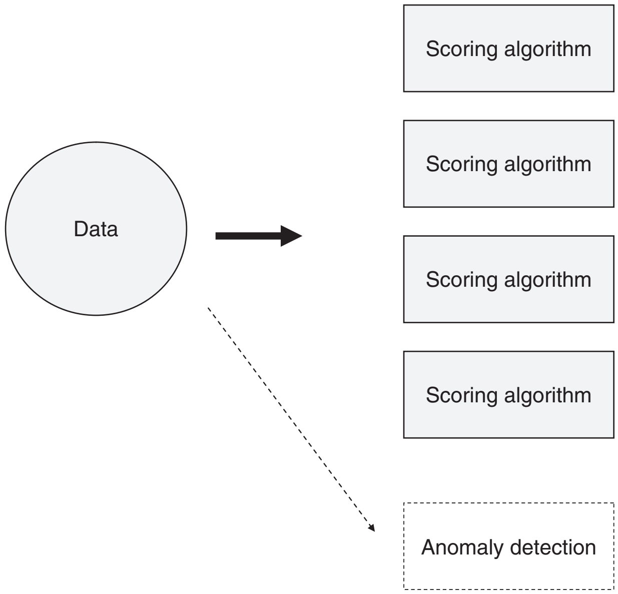


Figure 7-9 *This figure illustrates the concept of running anomaly detection in parallel with traditional scoring algorithms in an effort to further increase the likelihood that criminal or suspicious behavior will be detected.*

system would not catch everything, but it would represent an analytical safety net for those patterns of activities or behaviors that we do not know about currently or have not anticipated.

Which Tools???

The META Group market analysis of data mining tools[11] recently has identified market leaders and prepared a forecast of future market trends in data mining technology. The report advised that leadership positions for data mining vendors have been established by SAS, SPSS, and Oracle, and that market growth in this area is expected to occur in the development of tools specifically designed for use by less technical users in niche markets. This is consistent with the IDC analysis,[12] which recommends that "users of analytics should look for individualized relevancy . . . and the ability to easily move from analytics to decision to action."

Market share and success in other professional domains is an important consideration; however, other issues including flexibility, power, and a user-friendly interface can significantly affect the analysts' ability to effectively use these tools and realize their potential in the applied setting. Even the most powerful tool will have limited value if the analysts cannot get their data into it, use it, and create operationally actionable output. There are several good reviews of data mining tools that cover these issues,[13] and more products are being released and reviewed almost daily.

Another issue for consideration is what other agencies are using. As discussed in this chapter, the analytical processes can be captured, saved, and shared. This represents a unique opportunity for the applied public safety and security community to share knowledge without needing to share data. Criminals do not respect jurisdictional boundaries and even exploit them in many patterns of offending. Data sharing has been an issue fraught with technical limitations related to incompatibility between legacy systems and, perhaps more challenging, a general unwillingness or legal inability to effectively share data. By using analytical tools that can capture, save, and share analytical strategies, separate jurisdictions and agencies can share knowledge without needing to share data.

As of this writing, the "niche market" of less technical subject matter experts described by the META Group is already being exploited. SPSS and Information Builders recently teamed in the development of a law enforcement data mining application specifically created for use in the applied setting. Based on the research and analytical framework developed by RTI International, this tool was specifically created to support deployment decisions. Given the current interest in public safety, security, and intelligence analysis, other companies are sure to follow this example.

Regarding cost, there is no way around the issue; this is expensive stuff. Analytical software is not inexpensive and data mining software falls into the high end of the range. It is possible to do data mining without sophisticated software, although it really helps. Predictive analytics requires specialized software resources. This investment is what separates the men from the boys, so to speak. In contrast to even a few years ago, these products now are available for purchase "off the shelf" and can be used by mere mortals. Many in law enforcement will experience sticker shock when considering the purchase of these products. While it is true that the software purchase price will not compare favorably with the cost of ballistic vests, radios, or even cruisers, the amount of money saved through effective deployment of personnel resources can pay for these products quickly. The increased public safety achieved through effective deployment and enhanced investigative efficacy can be difficult to quantify. Suffice it to say that a safe community attracts business, residents, and visitors, which can greatly enhance the economic health of any location and pay for additional software upgrades.

Ultimately, it is important to remember that these tools support the data mining process and incorporate the mathematical algorithms necessary but not sufficient for some of the data preparation and modeling. As outlined in Chapter 4, it is the domain expertise and ability to create operationally actionable output that is the priceless element in the applied public safety and security analytical process.

7.18 Tools of the Trade

Some specific modeling algorithms and approaches were covered earlier in this chapter, and specific strategies also are suggested in each chapter. The following section will go through the general layout and use of two popular data mining tools: Enterprise Miner (SAS) and Clementine (SPSS).

Figure 7-10 depicts a sample analytical "stream" that was used to analyze the conference call data reviewed in Chapter 6. This example was generated with the Clementine suite of predictive analytics (SPSS Inc.). Figure 7-11 depicts a

Figure 7-10 *Sample Clementine Stream. Not all analytical packages use the same tools or format for cleaning and recoding data; this illustration has been provided to depict the analytical process. (Screenshot taken by the author is from Clementine 8.5; SPSS, Inc.)*

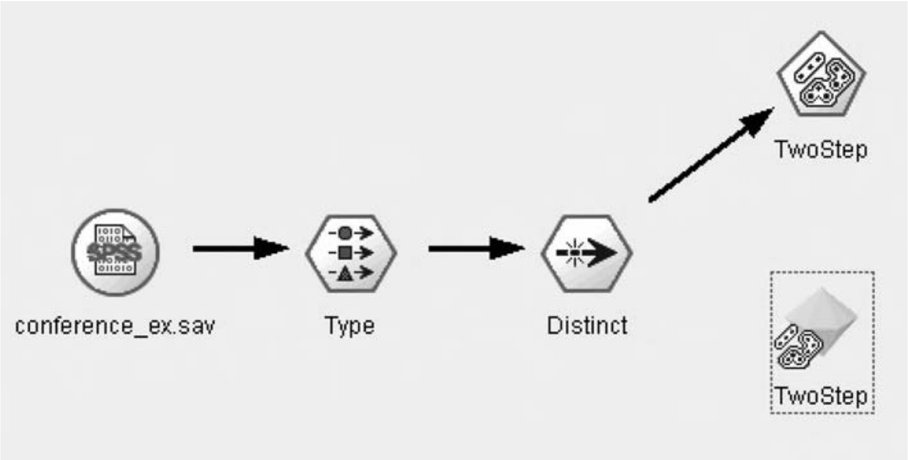
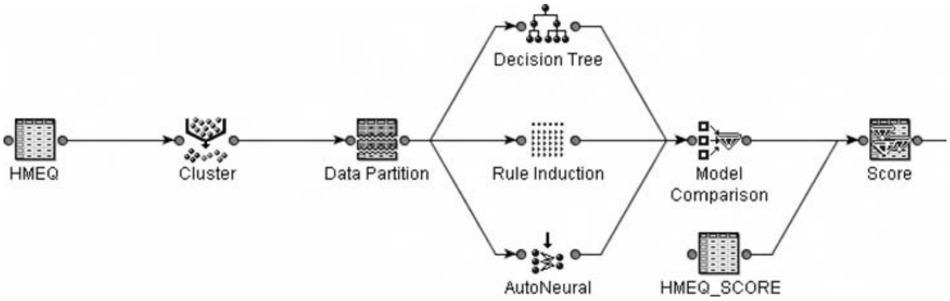


Figure 7-11 *Sample Enterprise Miner analytical process. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



similar analytical pathway that was generated using the SAS Enterprise Miner™ data mining solution. Enterprise Miner and Clementine are increasing in popularity among law enforcement and intelligence professionals, particularly in light of the powerful analytics and ease of use.

Moving from left to right, 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 sources, which would require the inclusion of more than one source node.

The next node in the Clementine stream is the "type" node, 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.

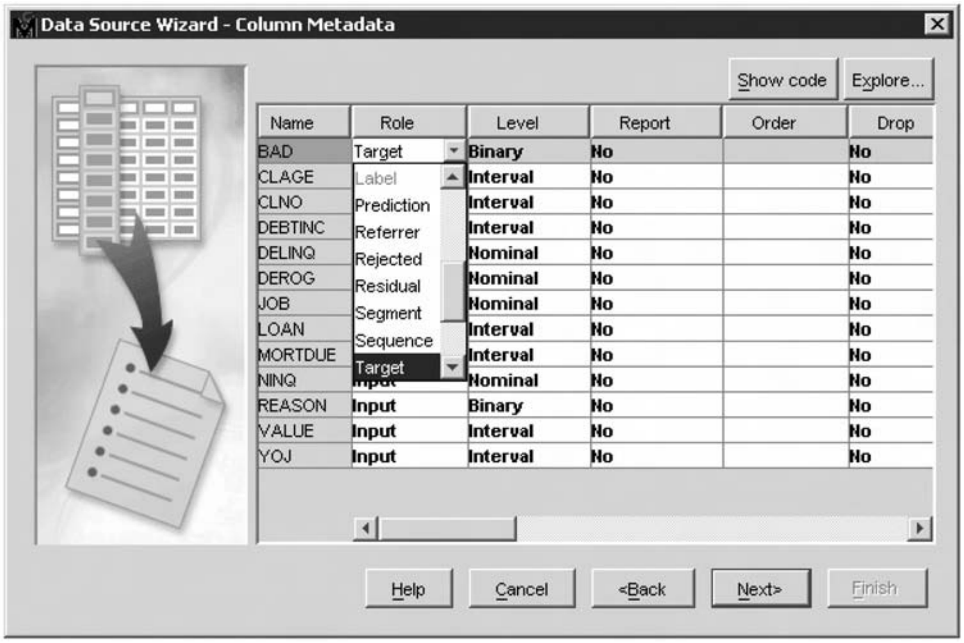
The next node in the stream is the modeling node. In this particular example, a clustering algorithm was used to create groups of similar cases. This also is referred to as an unsupervised learning approach because it does not start with a predetermined outcome or classification system.

The Enterprise Miner process includes a segmentation icon, which randomly assigns the data into training and test samples. This analytical process also incorporates the use of multiple modeling algorithms, and includes an icon that later will be used to evaluate and compare these different modeling tools.

7.19 General Considerations and Some Expert Options

Some packages will "read" the data and identify preliminary numeric properties. Additional steps include selection of the target variables and the identification of which variables should be included for further consideration and modeling. While it might seem foolish to leave anything out, most data sets generally include information that would be inappropriate or irrelevant for modeling (e.g., case ID numbers). Figure 7-12 illustrates the Enterprise Miner dialog box that allows the analyst to select variables for inclusion in subsequent analyses.

Figure 7-12 *Sample dialog box illustrating the selection of variables. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



Name	Role	Level	Report	Order	Drop
BAD	Target	Binary	No		No
CLAGE		Interval	No		No
CLNO		Interval	No		No
DEBTINC		Interval	No		No
DELINQ		Nominal	No		No
DEROG		Nominal	No		No
JOB		Nominal	No		No
LOAN		Interval	No		No
MORTDUE		Interval	No		No
NINQ	Input	Nominal	No		No
REASON	Input	Binary	No		No
VALUE	Input	Interval	No		No
YOJ	Input	Interval	No		No

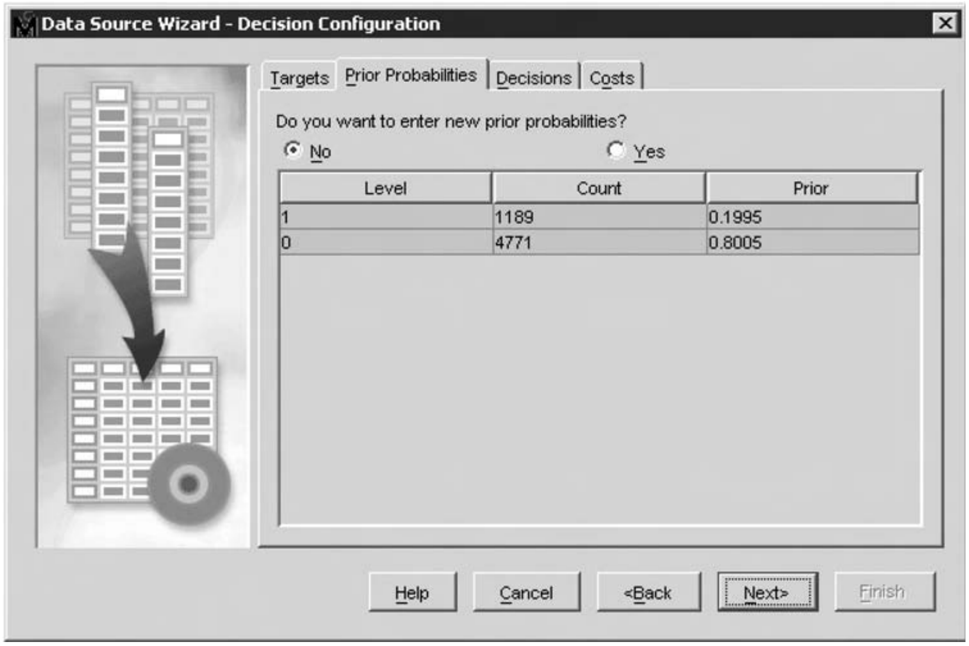
7.20 Variable Entry

Stepwise entry of data allows the inclusion of only those variables contributing to an increase in accuracy, in order of importance. Once the preset accuracy is achieved, 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.

7.21 Prior Probabilities

The issue of prior probabilities and its particular relevance in modeling rare or infrequent events was mentioned in Chapter 1. Figure 7-13 illustrates a sample dialog box that supports this particular function. As can be seen, the program automatically determines the prior probabilities of the target and accordingly

Figure 7-13 *Sample dialog box illustrating the automatic determination prior probabilities. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



Do you want to enter new prior probabilities?

● No ○ Yes

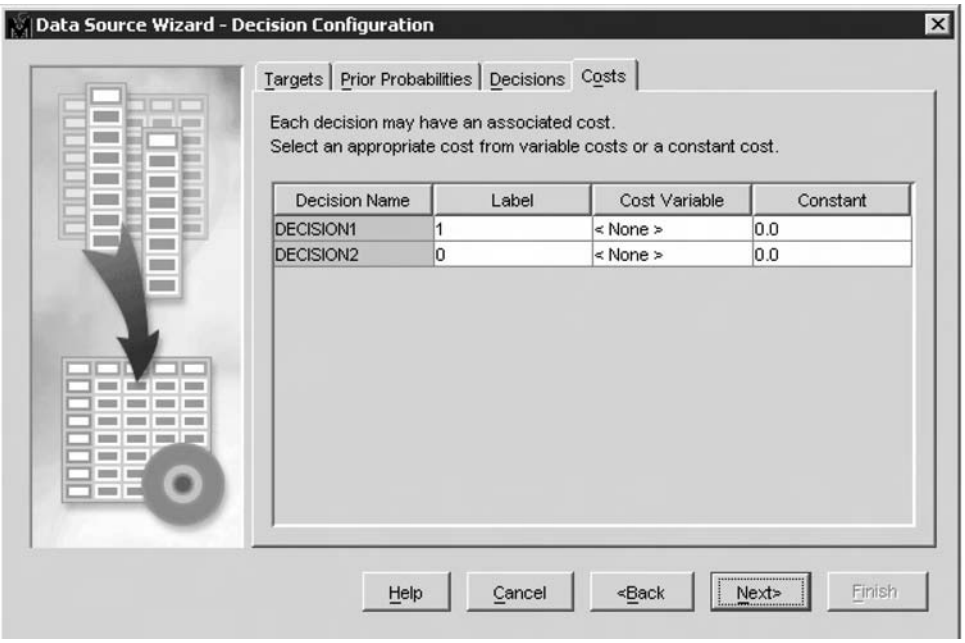
Level	Count	Prior
1	1189	0.1995
0	4771	0.8005

sets the expected probabilities in the model. The analyst can adjust these probabilities in this dialog box in an effort to alter the predicted occurrence of the target; however, this usually is not necessary. 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 these packages, the prior probabilities are automatically determined and used in the modeling and prediction.

7.22 Costs

Costs were also mentioned in Chapter 1. Figure 7-14 illustrates the sample dialog box. As much art as technology, the analyst simply adjusts the costs of certain types of errors and then reviews the accuracy achieved to determine the best trade-off for the particular analysis. It frequently takes a

Figure 7-14 *Sample dialog box illustrating options for adjusting the cost of errors. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



Each decision may have an associated cost.
Select an appropriate cost from variable costs or a constant cost.

Decision Name	Label	Cost Variable	Constant
DECISION1	1	< None >	0.0
DECISION2	0	< None >	0.0

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 above, 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 suite includes greater functionality and capacity packaged in a more intuitive interface. All that being said, 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 save something without solid domain expertise.

7.23 Bibliography

1. IDC (2004). Worldwide end-user business analytics 2004-2008 forecast: Core versus predictive; http://www.marketresearch.com/product/display.asp?productid=1073540
2. Helberg, C. (2002). Data mining with confidence, 2nd ed. SPSS, Inc., Chicago, IL; and Two Crows (www.twocrows.com), which is an excellent source of accurate, yet easy to understand information on data mining and predictive analytics.
3. Klecka, W.R. (1980). Discriminant analysis. *Quantitative Applications in the Social Sciences.*
4. Casey, E. (2002). Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and Internet crime investigation. *Knowledge Solutions*; www.corpus-delicti.com/case_based.html
5. Helberg (2002).
6. Brown, D.E. (1998) The Regional Crime Analysis Program (RECAP): A framework for mining data to catch criminals. University of Virginia.
7. McCue, C. (2001). Cops and Docs program brings police and ED staff together to address the cycle of violence. *Journal of Emergency Nursing*, **27**, 578–580.
8. McCue, C., Smith, G.L., Diehl, R.L., Dabbs, D.F., McDonough, J.J., and Ferrara, P.B. (2001). Why DNA databases should include all felons. *Police Chief*, **68**, 94–100.
9. 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/listinfo/papyrus-news
10. McWilliams, B. (2003). Iraq's crash course in cyberwar. *Wired News*, May 22.
11. METAspectrum Market Summary (2004). Data mining tools: METAsprectrum evaluation.
12. IDC (2004).
13. For specific product reviews, see Elder Research, Inc. at http://www.dataminininglab.com, or KDnuggets.com.

This Page Intentionally Left Blank

8

Public Safety–Specific Evaluation

One question that comes up with increasing frequency is: How effective is a particular deployment strategy or operational plan at reducing crime? What 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. And, in many cases, the challenge involves the documentation of the absence of crime. So, if nothing happened, how do we show that nothing happened? 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.

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.

In reality, 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

The time to consider outcome measures is before the plan is implemented, in fact, 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. The outcome measures should be relatively specific in time, space, and nature, and relatively easy to assess.

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 initiative[1] 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.

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.

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.

Specific Measure

It is very important to select the specific outcome measure with thoughtful consideration. 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.

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. 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.

8.2 Think Big

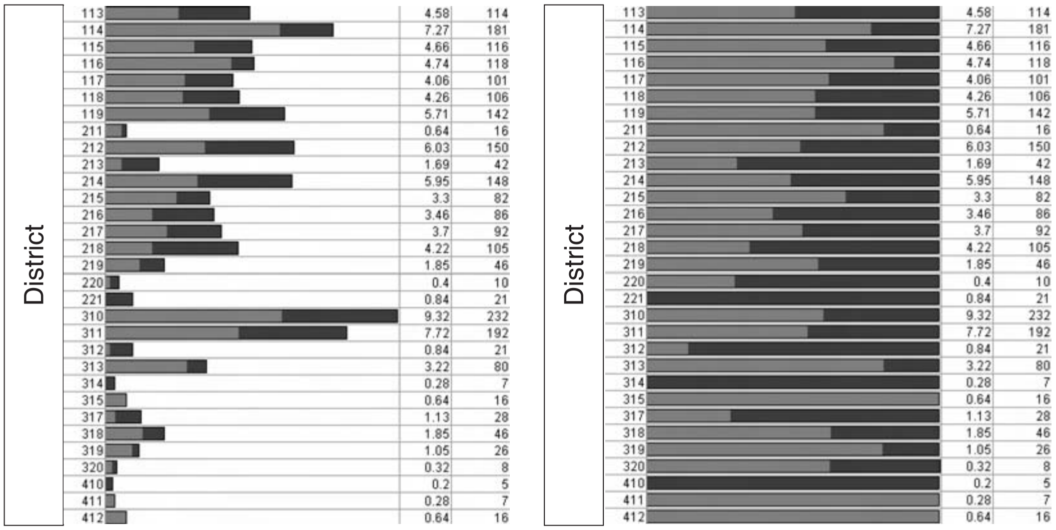
It is unfortunate, but violent crime often is evaluated based on one measure: the homicide rate. Generally, these tend to be low-frequency events compared to other types of violent crime. While this is not a bad thing, it can seriously hamper evaluation efforts. Murders are committed for a variety of 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? 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 8-1 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 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

Figure 8-1 *This figure illustrates hypothetical random gunfire complaints in several distinct geographic locations within a particular community.*



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 post-intervention 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 initiative[2] 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 eight hours associated with the initiative. This figure does not include the fact that data mining was used to confine the initiative to an eight-hour 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.

This example highlights 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.

The New Year's Eve initiative represents only one approach to documenting the value of data mining 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 this example that were directly linked to the successful evaluation of this strategy 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 conversation 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 a readily 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 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.

One additional feature of this New Year's Eve initiative evaluation is that the data were analyzed in aggregate form, and then parsed to identify specific areas of improvement. This was extremely important because only a few areas were targeted. The target areas were evaluated specifically to ensure that they were consistent with the overall pattern of reduced random gunfire complaints observed globally, which they were. In this case, the overall numbers for the entire city improved so significantly that it highlights further that the appropriate choices were made in selection of the target areas.

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 return on investment (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. Now 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.

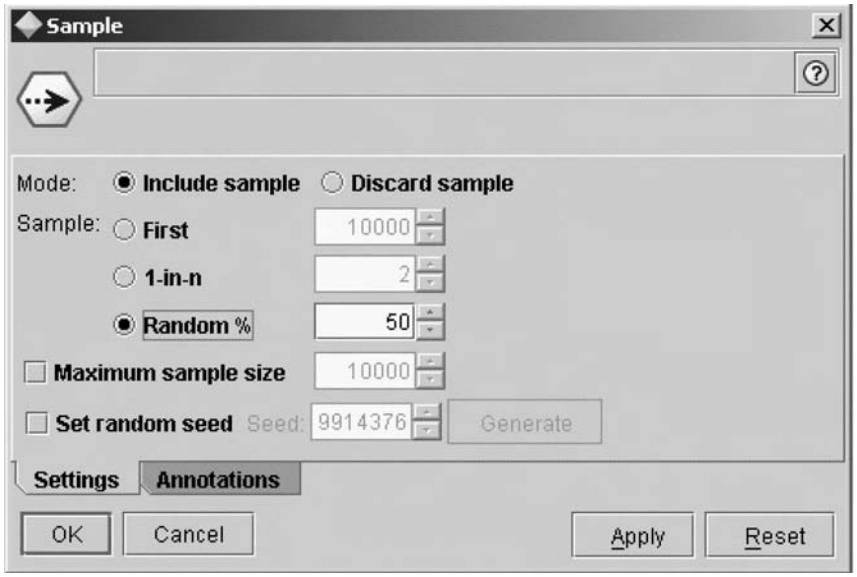
8.3 Training and Test Samples

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 skews the results. When a model is over-trained, 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 over-training 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 8-2, 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



Sample

Mode: ● Include sample ○ Discard sample

Sample: ○ First 10000

○ 1-in-n 2

● Random % 50

□ Maximum sample size 10000

□ Set random seed Seed: 9914376

Figure 8-2 *This dialog box illustrates the random assignment of data into training and test samples. (Screenshot taken by the author is from Clementine 8.5.)*

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, 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 is 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

Table 8-1 *A small sample of age data that has been divided into training and test samples. It is important to ensure that training and test samples are similar, particularly when they are relatively small.*

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

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 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 also is 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, which increases 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. 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 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

Table 8-2 *This figure illustrates a typical confidence matrix.*

		Predictions	
		False	**True**
Actual	False	33	5
	True	1	14

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 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

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

		Predictions	
		False	**True**
Actual	False	23	15
	True	1	14

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

		Predictions	
		False	**True**
Actual	False	148	38
	True	5	5

Table 8-5 *This confidence matrix depicts a revised model for Table 8-2. The overall accuracy has not been increased substantially but the "true" predictions or classifications have been improved.*

		Predictions	
		False	**True**
Actual	False	151	33
	True	3	6

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 one-third 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.

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 places, 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 Caveat Emptor

There is one area in which to be particularly cautious when using some outside consultants. As mentioned earlier, it is relatively easy to develop a model that describes existing data very well but that is over-trained and therefore relatively inaccurate with new data. In other words, 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. Unfortunately when this happens, the outside "expert" generally is gone, and the locals left behind might mistakenly assume the blame for the poor outcome. This can be particularly troubling in situations where accurately evaluating the performance of the model is beyond the technical expertise of the customer. In this situation, the same consultant is rehired and then "fixes" the model, generally by overfitting it to a new set of sample data.

The best way to evaluate the performance of a model is to see how well it can predict or classify a second sample, whether a "test" sample or future data. Be particularly leery of any consultant who advises you that he or she will "adjust" the model until it reaches a certain level of performance, which frequently is very high, but warns you that in your hands the accuracy of the model is likely to decrease significantly. Generally, what these folks are doing is "overfitting" the model to increase the apparent accuracy. The performance of these models often will decrease significantly when used on new data, and frequently will perform far worse than a more general, less accurate model. At a minimum, the informed consumer will require that any outside consultant employ both training and test samples to ensure a minimal level of stability in the performance of the model. An even better scenario results when the agency analyzes its own data and creates its own models, perhaps with some outside assistance or analytical support and guidance. It is the folks on the inside that have the best understanding of where the data came from and how the model will be used. These individuals also have the requisite domain expertise, or knowledge of crime and criminals, that is necessary to generate meaningful data mining solutions. Generally, for law enforcement personnel to use predictive analytics effectively, some guidance is required regarding a few "rules of the road" and best practices, like the use of training and test samples, and perhaps some analytical support. But in my experience, this is not an overwhelming challenge.

Finally, it is important to remember that the cost of errors in the applied public safety setting frequently is life itself. 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 loss of life.

8.7 Bibliography

1. McCue, C., Parker, A., McNulty, P.J., and McCoy, D. (in press). Doing more with less: Data mining in police deployment decisions. *Violent Crime Newsletter*, U.S. Department of Justice, Spring 2004, 1, 4–5.
2. Ibid.

This Page Intentionally Left Blank

9

Operationally Actionable Output

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." Everything that comes before 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, generally does not translate well into the applied setting, which significantly limits its operational value. 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 that can significantly enhance the knowledge discovery while guiding operational strategy and tactics. Ultimately, effective visual output should say to the end user, "Go here now and look for this." "This" should be immediately obvious, allowing the end users to build on their tacit knowledge, while enhancing their ability to "know it when they see it."

9.1 Actionable Output

It is important to remember that an analyst may see things not readily apparent to others, given their experience with the data. Graphical representations can help them share their vision of the patterns and trends in the data with others. What follows is a brief overview of some novel visualization techniques.

Heat Maps and Schedules

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. These can be used to depict simple frequencies or risk. In Figure 9-1, different types of crime (INCDGRP) are depicted over different four-hour 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 1600–2000 hours time blocks. Similarly, assault and battery (A & B) offenses were more frequent during the evening time blocks.

Recalling that deployment represents the allocation of resources across time and space, a very simple deployment schedule was created to represent the occurrence of crime in five different areas during six four-hour time blocks (Figure 9-2). We can see opportunities for fluid deployment strategies, in which

Figure 9-1 *"Heat map" example (B. Haffey, SPSS, Inc.; used with permission).*

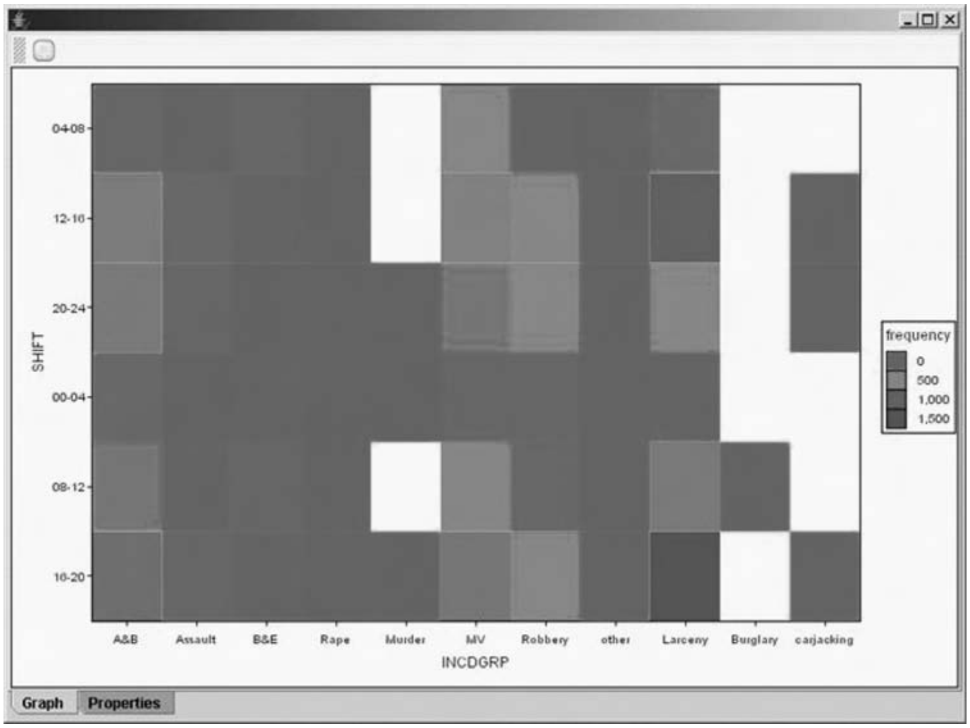
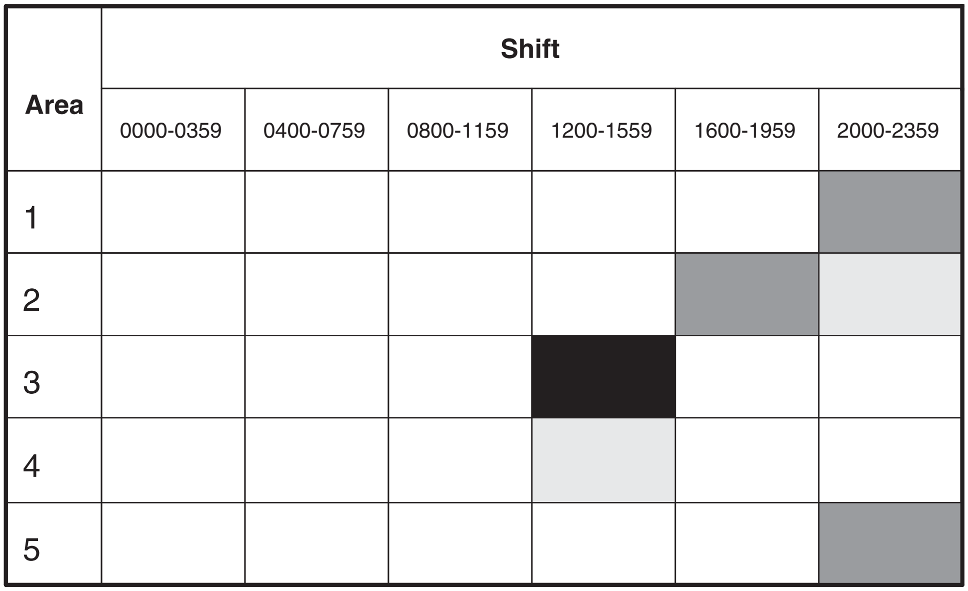


Figure 9-2 *Sample deployment "schedule."*

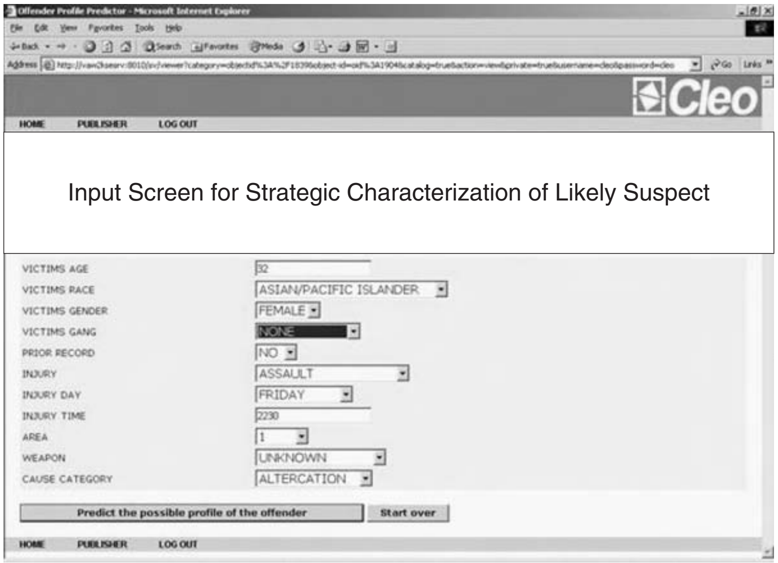


Area	Shift					
	0000-0359	0400-0759	0800-1159	1200-1559	1600-1959	2000-2359
1						▓
2					▓	░
3				■		
4				░		
5						▓

resources are proactively deployed to a particular location and then moved or flexed to accommodate anticipated changes in activity. In this particular example, we can take advantage of the sequential increases and fluctuation in activity. For example, we can anticipate the greatest amount of activity in Area Three during the 1200–1559 hours time block and respond by deploying most of our resources to that location before the anticipated increase in activity. Review of a map could provide additional insight regarding Area Four, helping us determine whether this activity is contiguous with Area Three and represents possible risk for displacement in response to the planned suppression in Area Three. Anticipating an increase in activity in Area Two at 1600 hours, we could then plan to flex our resources to that area.

Identifying ways to translate data mining output into a format with direct relevance to operational personnel has greatly increased optimization of analytical resources. As illustrated in Figure 9-3, 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 9-4 illustrates the relevant, easy-to-understand output generated in response to the input illustrated in the previous figure.

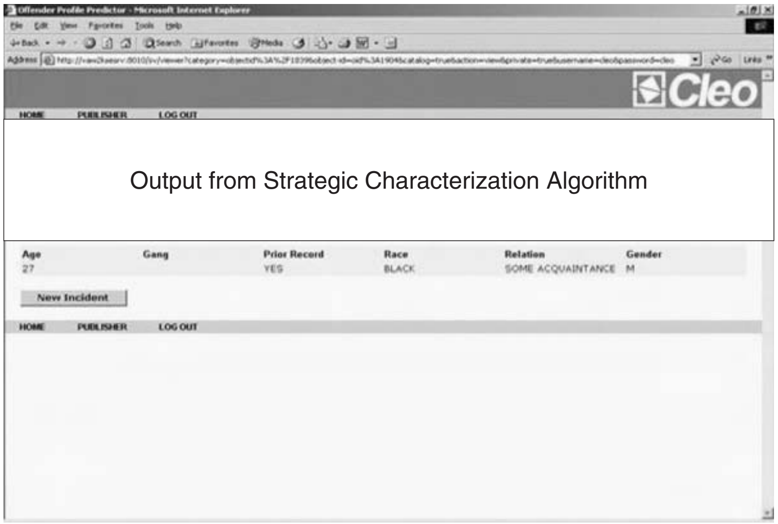
Figure 9-3 *By using web-based deployment of data mining models, the investigator can benefit from 24/7 crime analysis, which means that analytical capacity will be available when and where it is needed, even at the crime scene. (Cleo, SPSS, Inc.)*


Offender Profile Predictor - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address Go Links
Cleo
HOME PUBLISHER LOG OUT
Predict the possible profile of the offender
Start over
HOME PUBLISHER LOG OUT

Input Screen for Strategic Characterization of Likely Suspect

VICTIMS AGE 32
VICTIMS RACE ASIAN/PACIFIC ISLANDER
VICTIMS GENDER FEMALE
VICTIMS GANG NONE
PRIOR RECORD NO
INJURY ASSAULT
INJURY DAY FRIDAY
INJURY TIME 2230
AREA 1
WEAPON UNKNOWN
CAUSE CATEGORY ALTERCATION

Figure 9-4 *Sample output generated by the tool depicted in Figure 9-3. (Cleo, SPSS, Inc.)*


Offender Profile Predictor - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address Go Links
Cleo
HOME PUBLISHER LOG OUT
New Incident
HOME PUBLISHER LOG OUT

Output from Strategic Characterization Algorithm

Age	Gang	Prior Record	Race	Relation	Gender
27		YES	BLACK	SOME ACQUAINTANCE	M

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 in an officer's car during a tour of duty (Figure 9-5) or at a crime scene (Figure 9-6), analytical support is just a keystroke away.

The particular example in Figure 9-7 is covered in Chapters 12 and 14; however, the novel strategy of depicting the results of an unsupervised 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 immediately actionable.

Figures 9-8 through 9-11 illustrate the domain-specific tool developed by SPSS and Information Builders, which was based on the RTI International analytical framework. Figure 9-8 depicts the probability of crime by dispatch zone 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 within the square. Figure 9-9 includes the captured area and provides additional detail through the use of an orthophotography layer. This image gives additional detail about the area, allowing the end users to incorporate their own knowledge to identify particular environmental features,

Figure 9-5 *Scoring algorithms can be pushed out into the field through web-based interfaces through mobile computers . . .*

Figure 9-6
. . . or through PDAs, providing analytical support when and where it is needed.

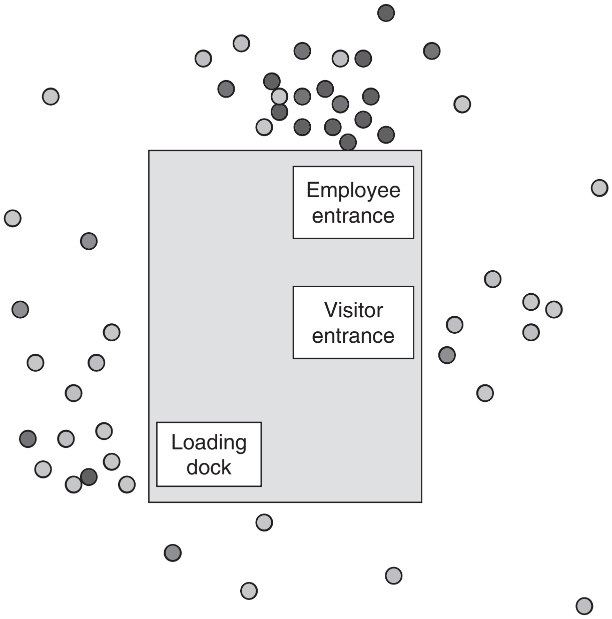
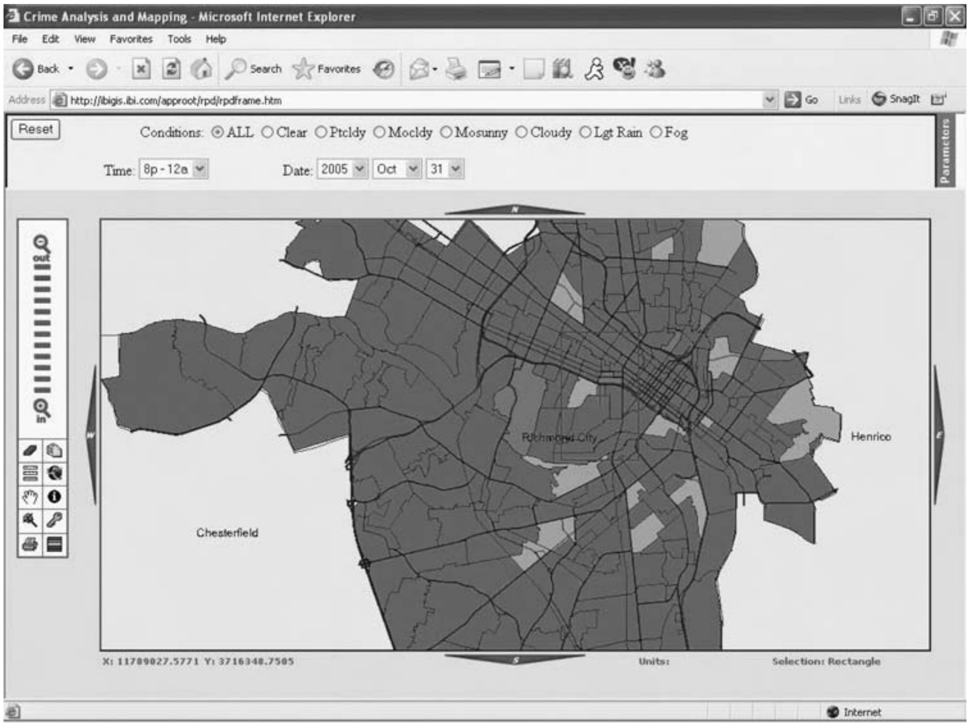


Figure 9-7
Facility map with overlay of unsupervised learning output.

Figure 9-8 *Map depicting probability of crime by dispatch zone. (Screen images of the Information Builders' Law Enforcement Application.)*



Crime Analysis and Mapping - Microsoft Internet Explorer

Address http://ibigis.ibi.com/approot/rpd/rpdframe.htm

Conditions: ◉ ALL ○ Clear ○ Ptcldy ○ Mocldy ○ Mosunny ○ Cloudy ○ Lgt Rain ○ Fog

Time: 8p - 12a

Date: 2005 Oct 31

X: 11789027.5771 Y: 3716348.7505

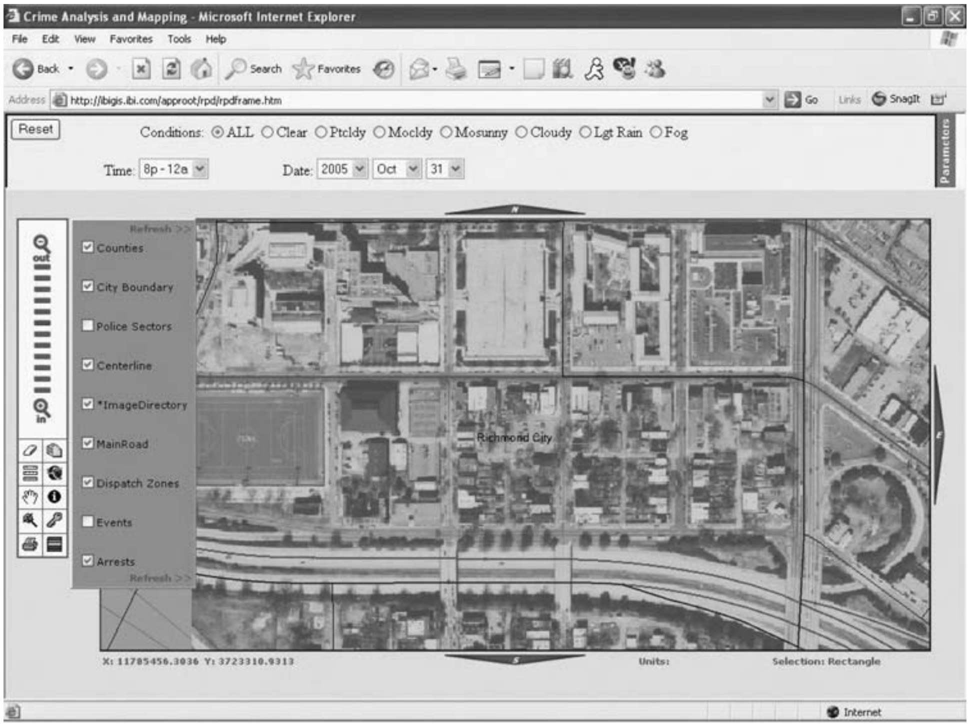
Units:

Selection: Rectangle

landmarks, or physical attributes in the interpretation of the results, as well as any operational planning.

Figure 9-10 adds specific information pertaining to drug arrests through the use of icons overlaid on the map. Again, the end users can exercise their knowledge of the area and domain expertise as they add value to the interpretation and operational use of these images. Finally, Figure 9-11 illustrates the direct link between the data and the image. By using this feature, the end user is able to call up additional information about a specific incident portrayed on the map, providing even more value to the analysis.

Figure 9-9 *Orthophotography detail of area highlighted in box from previous figure. (Screen images of the Information Builders' Law Enforcement Application.)*



Crime Analysis and Mapping - Microsoft Internet Explorer

Address http://ibigis.ibi.com/approot/rpd/rpdframe.htm

Conditions: ● ALL ○ Clear ○ Ptcldy ○ Mocldy ○ Mosunny ○ Cloudy ○ Lgt Rain ○ Fog

Time: 8p - 12a

Date: 2005 Oct 31

☑ Counties

☑ City Boundary

□ Police Sectors

☑ Centerline

☑ *ImageDirectory

☑ MainRoad

☑ Dispatch Zones

□ Events

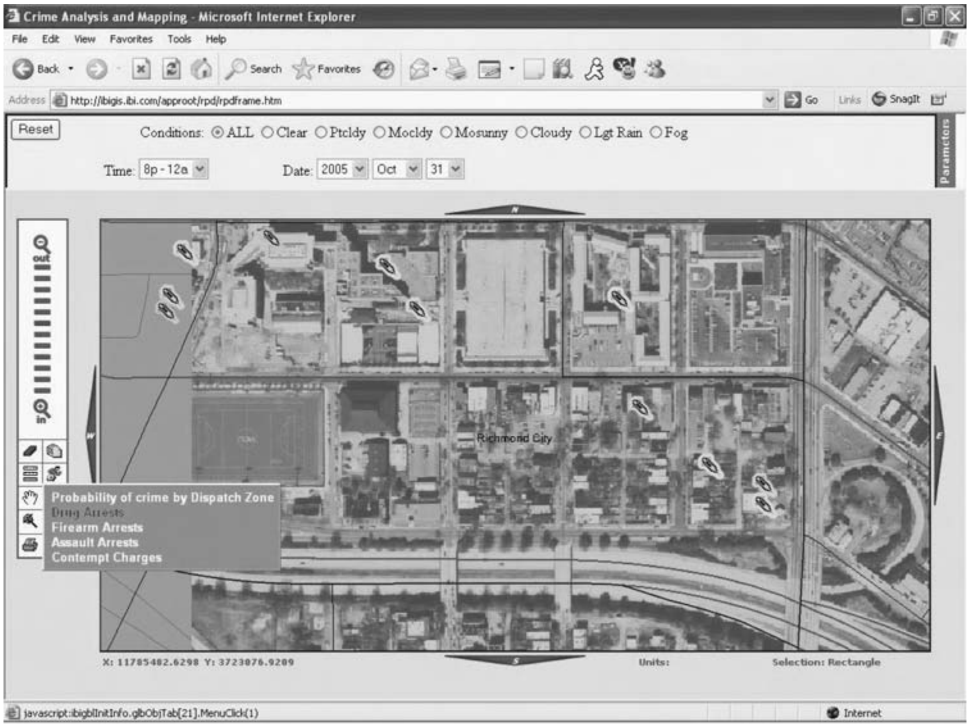
☑ Arrests

X: 11785456.3036 Y: 3723310.9313

Units:

Selection: Rectangle

Figure 9-10 *Overlay illustrating the direct link between the visual image and the underlying data. "Needle" icons represent locations associated with drug arrests. (Screen images of the Information Builders' Law Enforcement Application.)*



Crime Analysis and Mapping - Microsoft Internet Explorer

Address http://ibigis.ibi.com/approot/rpd/rpdframe.htm

Conditions: ◉ ALL ○ Clear ○ Ptcldy ○ Mocldy ○ Mosunny ○ Cloudy ○ Lgt Rain ○ Fog

Time: 8p - 12a

Date: 2005 Oct 31

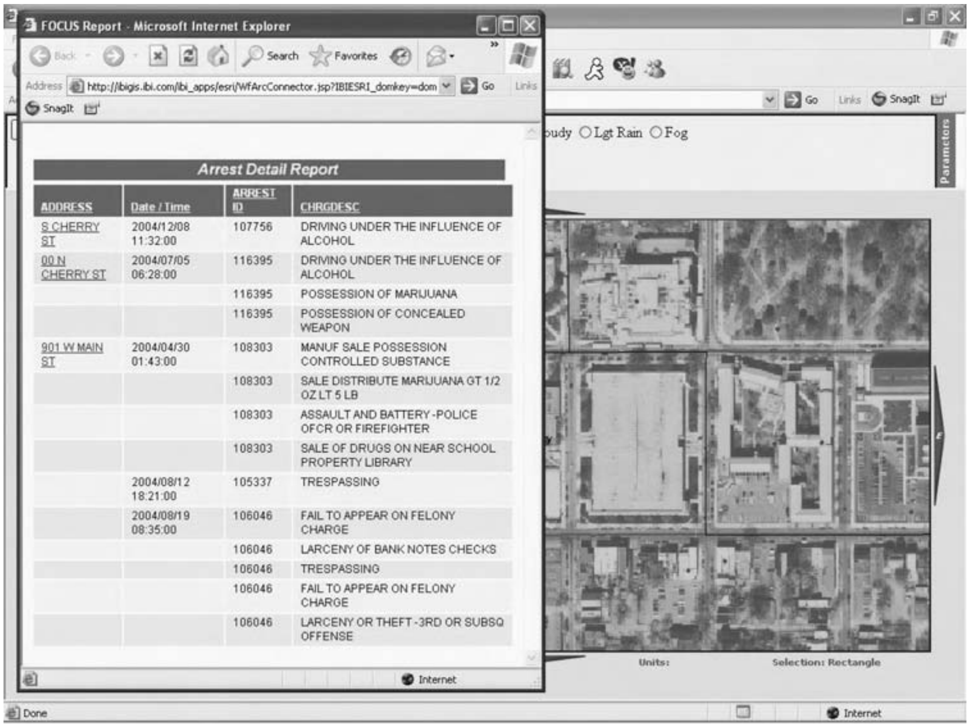
X: 11785482.6298 Y: 3723076.9209

Units:

Selection: Rectangle

javascript:ibigblInitInfo.glbObjTab[21].MenuClick(1)

Figure 9-11 *Pop-up screen illustrating the link between the incident data and the image.*



FOCUS Report - Microsoft Internet Explorer
Back
Search
Favorites
Address
http://ibigis.ibi.com/ibi_apps/esri/WfArcConnector.jsp?IBIESRI_domkey=dom
Go
Links
SnagIt
oudy ○ Lgt Rain ○ Fog
Parameters
Units:
Selection: Rectangle
Internet
Done

Arrest Detail Report

ADDRESS	Date / Time	ARREST ID	CHRGDESC
S CHERRY ST	2004/12/08 11:32:00	107756	DRIVING UNDER THE INFLUENCE OF ALCOHOL
00 N CHERRY ST	2004/07/05 06:28:00	116395	DRIVING UNDER THE INFLUENCE OF ALCOHOL
		116395	POSSESSION OF MARIJUANA
		116395	POSSESSION OF CONCEALED WEAPON
901 W MAIN ST	2004/04/30 01:43:00	108303	MANUF SALE POSSESSION CONTROLLED SUBSTANCE
		108303	SALE DISTRIBUTE MARIJUANA GT 1/2 OZ LT 5 LB
		108303	ASSAULT AND BATTERY -POLICE OFCR OR FIREFIGHTER
		108303	SALE OF DRUGS ON NEAR SCHOOL PROPERTY LIBRARY
	2004/08/12 18:21:00	105337	TRESPASSING
	2004/08/19 08:35:00	106046	FAIL TO APPEAR ON FELONY CHARGE
		106046	LARCENY OF BANK NOTES CHECKS
		106046	TRESPASSING
		106046	FAIL TO APPEAR ON FELONY CHARGE
		106046	LARCENY OR THEFT -3RD OR SUBSQ OFFENSE

Applications

This Page Intentionally Left Blank

10

Normal Crime

"When people are free to do as they please, they usually imitate each other."

Eric Hoffer

Crime is behavior. Like many other patterns of behavior, even the most serious violent crime frequently can be characterized, categorized, anticipated, and hopefully even predicted. 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 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. Moreover, in our experience, deviation from normal generally indicates the potential for escalation or the presence of something “bad.” “Normal” burglaries generally require that the criminal get in and out without being detected, and that the criminals 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 addressed in greater detail in Chapter 11, but suffice it to say 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 Knowing Normal

“Science is nothing but trained and organized common sense.”

Thomas Henry Huxley

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 been 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 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

Table 10-1 *Transaction records from a suspicious bank account.*

	date	amount	chkno	bal
38	10/26/20XX	10.50	543	1640.46
39	09/29/20XX	120.90	544	3041.58
40	03/16/20XX	165.00	545	2406.63
41	10/20/20XX	39.12	546	1900.96
42	11/17/20XX	421.20	548	2536.81
43	10/19/20XX	42.00	549	1940.08
44	10/30/20XX	250.00	550	1390.46
45	11/20/20XX	75.58	551	1705.23
46	11/21/20XX	54.96	552	1584.82
47	11/20/20XX	756.00	553	1780.81
48	11/17/20XX	500.00	554	2958.01
49	11/21/20XX	65.45	555	1639.78
50	12/26/20XX	157.86	556	2068.47
51	11/28/20XX	39.49	557	1499.33
52	11/24/20XX	46.00	558	1538.82
53	01/10/20XX	11.40	559	1677.26
54	12/26/20XX	118.25	560	1950.22
55	01/03/20XX	71.56	561	1878.66
56	01/16/20XX	378.00	562	1299.26
57	01/31/20XX	8.00	563	1291.26
58	02/28/20XX	197.87	564	1620.39

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.

Table 10-2 *Frequency distribution of the withdrawals associated with the bank account illustrated in Table 10-1. An unusual distribution of expenditures in round amounts was suspicious and warranted additional investigation.*

80.05	1	.4	.4
81.00	1	.4	.4
87.86	1	.4	.4
88.78	1	.4	.4
89.00	1	.4	.4
89.50	1	.4	.4
95.70	1	.4	.4
96.00	2	.8	.8
97.50	1	.4	.4
100.00	23	9.3	9.3
105.00	1	.4	.4
110.00	1	.4	.4
117.74	1	.4	.4
118.25	1	.4	.4
120.00	1	.4	.4
120.90	1	.4	.4
125.00	2	.8	.8
125.25	1	.4	.4
126.43	1	.4	.4
126.50	1	.4	.4
128.00	1	.4	.4
130.00	3	1.2	1.2

- 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.

10.2 "Normal" Criminal Behavior

In the next example, a more rigorous statistical approach was employed to determine the relationship between different patterns of criminal offending, particularly as they relate to the potential for significant escalation.

Through a casual conversation, 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.[1] Rather, many of these individual 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. 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. By drilling down, 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 "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.

10.3 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. An example of this is the review of drug-related homicides to identify the relationship between the home address and crime location. 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.[2]

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 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 would allow 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 Staged Crime

One group of individuals that frequently gets it wrong with “normal” are 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.

A recent example of a staged crime that deviated significantly from “normal” involved a college coed who went missing from her residence late one night. After she was found a few days later, several items supposedly belonging to her abductor also were recovered. Several unusual aspects of this crime aroused suspicions that ultimately proved to be correct.

Most 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. 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.[3] 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.

10.5 Bibliography

1. McCue, C., Smith, G.L., Diehl, R.L., Dabbs, D.F., McDonough, J.J., and Ferrara, P.B. (2001). Why DNA databases should include all felons. *Police Chief*, **68**, 94–100.

2. Brown, D.E. (1998). The Regional Crime Analysis Program (RECAP): A framework for mining data to catch criminals. University of Virginia.

3. 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/listinfo/papyrus-news

This Page Intentionally Left Blank

11

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 predicted. 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

Figure 11-1 *Homicide rates can alter perceptions of the relative safety and associated desirability of a community.*

thesis of this discussion was that the behavioral analysis of violent crime or criminal “profiling” works because violent crime is relatively homogeneous and predictable.

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 concept is explored further in the chapter on modeling.

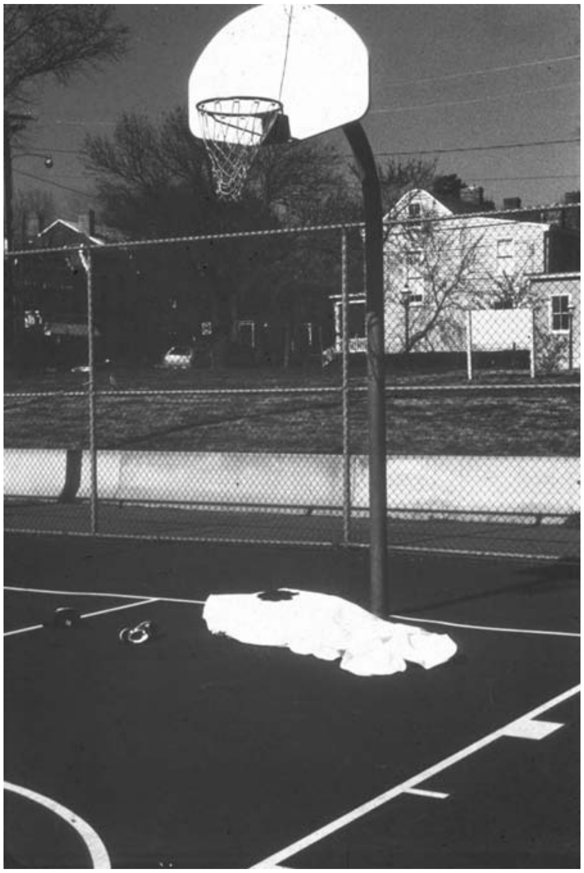


Figure 11-2 *Many communities inundated with drugs and violence resemble urban combat zones.*

Animal Models

> *"Innocence about Science is the worst crime today."*
>
> Sir Charles Percy Snow

It may not be obvious, but animal research can help us understand violent crime and enhance data mining strategies. An understanding of the research on nonhuman animal behavior frequently holds some insight for students of human behavior that can be applied to criminal or violent behavior in some cases. While it is unlikely that there are any studies in scientific literature entitled "When Mice Go Bad," some excellent comparison data can be identified with a creative approach to the literature. For example, the theory of species-specific defense reactions (SSDR), which was developed using animals models, provides a good explanation for situations in which people fail to react in life-threatening situations and perish as a result. Briefly, this theory suggests that there are certain behaviors that have been selected for during the evolutionary process that are preprogrammed, or hardwired into the brain, because they previously have been integral to survival. The fight or flight response falls into this category, as well as another behavior called "freezing," in which an individual "freezes," or otherwise is incapable of movement. Why is an understanding of animal models of behavior important to the analysis of human public safety data? First, it provides a general outline of existing behavioral categorization that can be adapted and used to describe and understand human behavior. For example, we know based on the literature that animal responses to threatening situations can be classified into three mutually exclusive and distinct categories: fight, flight, or freeze. Within a context of conservation of the behavioral mechanism, this same behavioral categorization can be applied to human responses to threatening situations.

Second, by identifying and linking common behavioral elements found in animal research literature to questions regarding human behavior, we often can enrich our understanding of the human behavior in question. There are some "experiments" that would be illegal or unethical to conduct on human subjects. Characterizing and reducing human behavior to common elements and themes, however, frequently can facilitate the identification of a similar animal model that might further illuminate other aspects or outcomes related to the behavior not considered previously. For example, models describing predator selection and acquisition of prey can be used to describe victim selection in some patterns of violent crime.

Returning to the example of SSDR, we know from animal studies that it can be extremely difficult to elicit a behavior other than the natural response in a threatening situation. By further reviewing the literature, specific examples can be identified and applied to human situations. For example, animal research information can be used to enhance escape planning in protective operations. If particular situations or scenarios are likely to elicit a freeze or fight response when escape would be the best choice, then operational planning can be modified to incorporate our understanding of the natural or likely response to this particular situation or combination of operational elements. Alternatively, an adaptive response can be taught in an effort to overcome the natural response in the scenario.

Why does this have value to data mining and predictive analytics? In many ways, the behavioral analysis of violent crime works because behavior is relatively homogeneous and predictable when reduced to its simplest form. Therefore, examination of the nonhuman behavioral literature provides two benefits. First, it forces the analyst to reduce behavior to common elements. That is, it forces the development of operational definitions for the behavior of interest. For example, overkill can be extremely upsetting for responders who encounter it but, operationally defined, it just means that the amount of force used was greater than that required to kill the victim. So overkill is either present or it is not, which reduces it to a yes/no question. It does not matter whether the victim was stabbed 50 times or 500 times—overkill is overkill. By reducing this question to a yes/no choice, a binary selection, the data and related questions have been categorized in a way that is appropriate for data mining and predictive analytics.

Behavioral literature often can provide a level of insight and understanding that can be used to structure and guide data mining analysis of criminal behavior while illuminating likely outcomes or responses to the behavior in question. For example, optimal foraging models and predatory behavior in animals can be used to describe the nature and function of some illegal drug markets. Therefore, by looking to these examples, drug enforcement also can be understood within a context of changing the behavioral cost of the criminal behavior, similar to the approach used in the very effective Project Exile.

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 activities theory of crime.[2] 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 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. Mapping should not be confined to pin maps showing what happened. Maps should be used to convey additional details or information regarding crime trends and patterns of interest.

11.1 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 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? Overkill, as discussed above, and many other factors considered during investigation can be broken down into "yes/no" or dichotomous decisions. Using this type of approach, decision trees can be developed that guide the investigative process, which ultimately can be used in automated motive determination algorithms and even strategic characterization models of likely suspect attributes. These and other data management techniques can add more structure and meaning to 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,[4] 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 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 subtley 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).

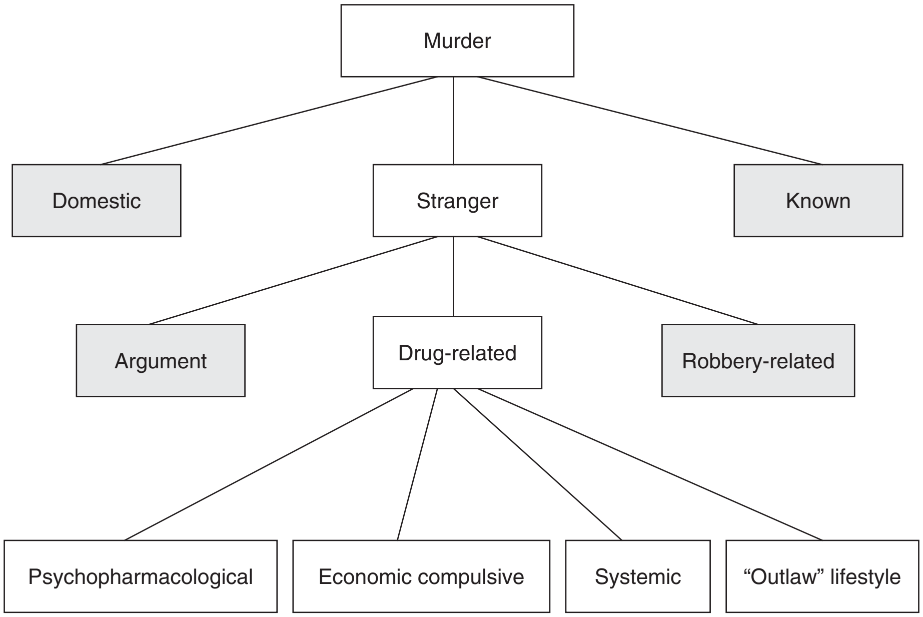
Does this mean that we should discard detectives in favor of computer modeling programs and expert systems? Absolutely not! Expert systems, no matter how "smart," lack one element critical to effective and meaningful data mining: domain expertise. Without an "expert" in the field to evaluate the nature 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 minimally 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.

11.2 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.[5] Figure 11-3 depicts a possible decision tree for murder. The first branch of

Figure 11-3 *Possible decision tree for categorizing murder.*



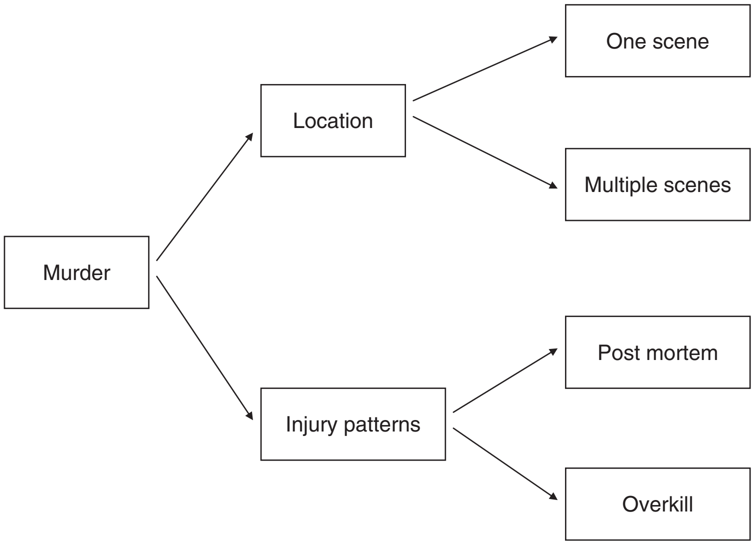
the tree is divided by three possible victim-perpetrator relationships: familial or domestic, acquaintance or known, and stranger. The second division in the branches includes possible motives associated with stranger murder. Finally, within the various motives, it is possible to characterize the motive even further. In this case, the drug-related homicides have been divided using the Tripartite Model of Drug-Related violence developed by Goldstein.[6] Briefly, this model describes three types of drug-related violence: Psychopharmacological, economic compulsive, and systemic violence, which are related to the different role that drugs and/or the drug-related lifestyle may play in violent crime.

While the Goldstein model might be helpful for counting crime or determining the proper focus for needed crime prevention strategies, it frequently adds little value to the investigative process because the nature of the victim-perpetrator relationship and true motive generally are not confirmed until a suspect has been identified and apprehended.

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 11-4, 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

Figure 11-4
Possible decision tree for categorizing behavioral and forensic evidence.



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 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., race)—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 often is greater than a sum of the parts. It also is important to identify all of the crimes associated with a particular individual for prosecutorial purposes. Presenting a linked series can be helpful in jogging the memory of a reluctant 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. 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).

11.3 Strategic Characterization

Where Do Murderers Come From?

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 example 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 crime scene behavior and 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, DiIulio, and Walters wrote a book titled *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 cognitive set, 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.

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. 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, given the lack of contact between the individuals and the relatively limited public knowledge and understanding of this behavior even a few years ago. 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?"

Even more intriguing are the young people who go through a similar developmental process as they begin to explore violence. Recent increases in the prevalence and availability of reference materials regarding serial murder and other types of violence seem to have increased the sophistication of the methods and desires associated with this group. It is for this reason, perhaps, that this group can be a particular challenge in terms of the relative degree of sophistication associated with their pattern and methods. Through the use of educational materials or facilitators, their crimes can be almost indistinguishable from those committed by much older individuals, which presents a significant challenge to those charged with investigating their crimes. Frequently, the question is not, "Could a child do this?" The answer to that question is almost always, "Yes." Rather, the question now has become, "Is the juvenile suspect that I am considering capable of committing this crime?"

Again, by using data mining and predictive analytics, it is possible to transcend human bias and opinion, particularly regarding juvenile suspects, 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.

Frequently in an actual investigation, the nature of the victim-perpetrator relationship is unknown. It is the crime scene characteristics and victim lifestyle factors that suggest a possible motive, which then is used to identify possible suspects. In fact, it is this type of strategic 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.[11] 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 modeling algorithm. These tools are 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. As a modeling tool, however, discriminant analysis is robust enough to accommodate a violation of this assumption. Basically, the type of error most likely to occur with a violation of this type is that a model will not emerge, even though there might be a relationship in the data.[12] The issue of errors and how they should be evaluated is addressed in Chapter 4. In a motive determination model, any error that would misdirect an investigation could potentially be catastrophic, as it would waste investigative resources. Moreover, delay in the investigative process could seriously compromise the ultimate ability to solve the case. 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 ultimate role for any potential models.

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.[13] 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 suspect, similar to a division of labor within drug distribution networks.[14] 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 little more 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, 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.

11.4 Automated Motive Determination

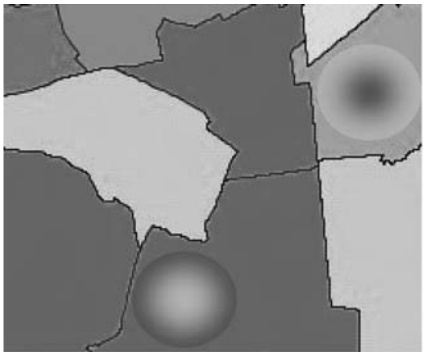
More recent work in this area has been confined exclusively to the information that is available early on in an investigation. Again, drug-related violence represents a good area of study for several reasons. First, drug-related violence frequently drives the overall violent crime trends in many communities plagued by serious increases in violent crime;[15] 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 "part 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 automated classification techniques similar to those used 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. 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 11-5).

The results were distilled and deployed through maps, 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

Figure 11-5 *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.*



Drug-related violence Characterized by differential victim risk

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.

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 increasingly has become a requirement as resources are increasingly limited.

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, these results were deployed through a web-based package, which provides for 24/7 analytical capacity.[16] 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 have the benefit of analytical help when they most need it: early in an investigation, when the case is progressing most rapidly.

11.5 Drug-Related Violence

Like homicide in general, drug-related violence can be categorized and divided further, based on the specific motive for the violence. 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.

In his tripartite model of drug-related violence, Dr. Paul Goldstein[17] has proposed three categories of drug-related violence. A fourth type of drug-related violence also should be considered that has been described as the "outlaw lifestyle."[18] 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.

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.

11.6 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 good thing for a community, it also is a very good thing for the analyst, as there generally are more incidents available for analysis and modeling.

11.7 Sexual Assault

Stranger rapists can be some 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 scientists 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.[19]

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,[20] 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. 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.[21]

This work is particularly encouraging because many of the category clusters identified in this study match classifications described previously in the United States.[22] 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.

11.8 Victimology

In some ways, victimology, or the study of the relationship between certain individual attributes 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." If that was the case, there would be very little that we could do to reduce violent crime, other than identify the perpetrators and get 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.

We have come to realize that certain victim attributes 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," we are limited in our ability to prevent violent crime. 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 try and change those behaviors, which might concomitantly reduce the risk for violent crime. By reducing the risk associated with some of these individuals, the amount of lead flying through the air also will be reduced, which will reduce the "bad luck" associated with those individuals who truly were in the wrong place at the wrong time. By drilling down into the data, it is possible to identify certain groups that are at greater risk for violence than others.

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 offending. 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. Certain attributes like promiscuity were noted among most of the categories, but others differed significantly. 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 ultimately had implications for officer safety in the field.

Analysis of aggregate statewide injury rates really did not reveal many differences among the different patterns of offending. When the data were categorized by community, however, 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.

This finding was revisited several years later when it was noted that certain victim characteristics were related to geography in determining drug-related homicides associated with a particular drug market or area. During the development of the risk-based deployment strategies for drug-related violence, for example, it was noted that certain areas were associated with more victims who were employed, while other areas were associated with victims who were less likely to be employed. While correlation certainly does not imply causality, this particular finding had value in terms of creating meaningful enforcement strategies that could be targeted specifically to the likely victims.

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.[23] 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.

11.9 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 crimes[24] 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 could be 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.

11.10 Bibliography

1. McLaughlin, C.R., Yelon, J.A., Ivatury, R., and Sugerman, H.J. (2000). Youth violence: A tripartite examination of putative causes, consequences and correlates. *Trauma, Violence & Abuse*, **1**, 115–127.
2. Cohen, L.E. and Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, **44**, 588–608.
3. McLaughlin et al. (2000).

4. Casey, E. (2002). Using case-based reasoning and cognitive apprenticeship to teach criminal profiling and internet crime investigation. *Knowledge Solutions;* www.corpus-delicti.com/case_based.html

5. See Westveer, A.E., ed. (2002). Managing death investigation. Washington, D.C. U.S. Department of Justice, Federal Bureau of Investigation; or Geberth, V.J. (1996). Practical homicide investigation: Tactics, procedures, and forensic techniques, 3rd ed. CRC Press: New York.

6. Goldstein, P.J. (1985). The drugs/violence nexus: A tripartite conceptual framework. *J Drug Issues*, **15**, 493–506.

7. McCue, C. (2002). Juvenile murderers. *In* Managing death investigation (Arthur E. Westveer, ed.), pp. 481–489. Washington, D.C., U.S. Department of Justice, Federal Bureau of Investigation.

8. Goldstein (1985).

9. Bennett, W.J., DiIulio, J.J., and Walters, J.P. (1996). Body count. Simon & Schuster, New York.

10. McLaughlin et al. (2000).

11. McLaughlin, C.R., Daniel, J., and Joost, T.F. (2000). The relationship between substance use, drug selling and lethal violence in 25 juvenile murderers. *Journal of Forensic Sciences*, **45**, 349–353.

12. Klecka, W.R. (1980). Discriminant analysis, Sage: Beverly Hills.

13. McLaughlin, C.R., Reiner, S.M., Smith, B.W., Waite, D.E., Reams, P.N., Joost, T.F., and Gervin, A.S. (1996). Firearm injuries among Virginia juvenile drug traffickers, 1992 through 1994 (Letter). *American Journal of Public Health*, **86**, 751–752.

    McLaughlin, C.R., Smith, B.W., Reiner, S.M., Waite, D.E., and Glover, A.W. (1996). Juvenile drug traffickers: Characterization and substance use patterns. *Free Inquiry in Creative Sociology*, **24**, 3–10.

    McLaughlin, C.R., Reiner, S.M., Smith, B.W., Waite, D.E., Reams, P.N., Joost, T.F., and Gervin, A.S. (1996). Factors associated with a history of firearm injuries in juvenile drug traffickers and violent juvenile offenders. *Free Inquiry in Creative Sociology, Special Issue: Gangs, Drugs and Violence*, **24**, 157–165.

14. Goldstein (1985).

15. McLaughlin, C.R., Robinson, D.W., and Faggiani, D. (1998). Declining homicide rates in the 1990s: Not everywhere! *Academy of Criminal Justice Sciences*.

16. McCue, C. and Parker, A. (2004). Web-based data mining and predictive analytics: 24/7 crime analysis. *Law Enforcement Technology*, **31**, 92–99.

17. Goldstein (1985).

18. McCue (2002).

19. McCue, C., Smith, G.L., Diehl, R.L., Dabbs, D.F., McDonough, J.J., and Ferrara, P.B. (2001). Why DNA databases should include all felons. *Police Chief*, **68**, 94–100.

20. Klecka (1980).

21. Adderly, R. and Musgrove, P.B. (2001). Data mining case study: Modeling the behavior of offenders who commit serious sexual assault. ACM Special Interest Group on Knowledge Discovery and Data Mining.

22. Ressler, R.K., Burgess, A.W., and Douglas, J.E. (1988). Sexual homicide: Patterns and motives. Lexington Books, New York.

23. Lord, W.D., Boudreaux, M.C., and Lanning, K.V. (2001). Investigation of potential child abduction cases: A developmental perspective. *FBI Law Enforcement Bulletin*, April.

24. Sims, D.W., Bivins, B.A., OBeid, F.N., Horst, H.M., Sorenson, V.J., and Fath, J.J. (1989). Urban trauma: A chronic recurring disease. *Journal of Trauma*, **29**, 940–947.

This Page Intentionally Left Blank

12

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. This likelihood or risk, however, 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.

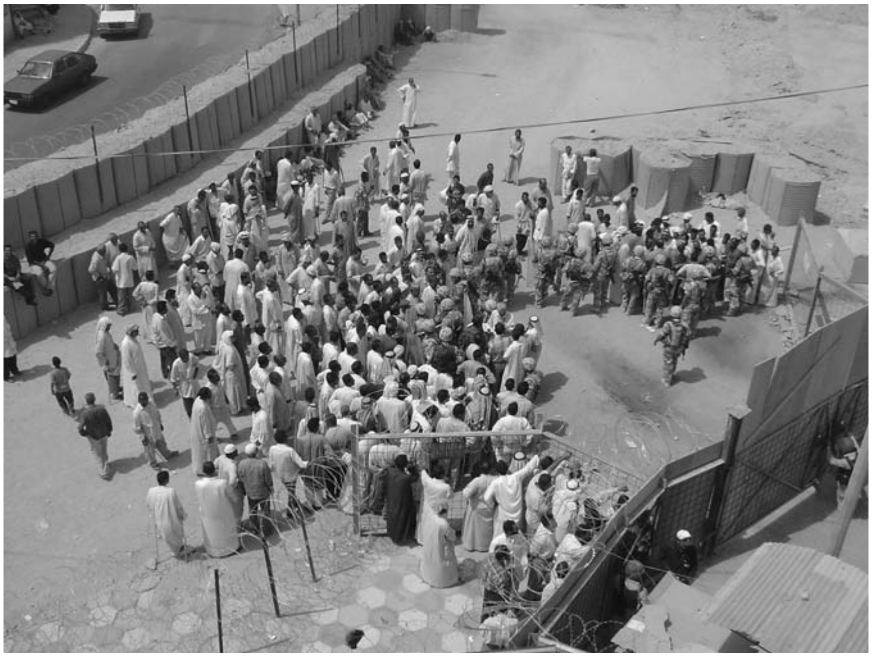


Figure 12-1 *Potentially hostile crowd outside the Coalition Provisional Authority headquarters in Basra, Iraq (Courtesy of Staff Sergeant Tom Ferguson, USMC).*

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 U.S. 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.

Risk and threat changed forever after the terrorist attacks of September 11. Increasingly, local law enforcement finds itself on the front lines of the war on terrorism. Knowledge of data mining and predictive analytical techniques is only part of the requirements for the development of reliable and accurate risk and threat assessment models that have value to the public safety community. Domain expertise is absolutely critical in the development and thoughtful creation and evaluation of risk models. Domain experts understand which data are available, what types of models are needed, and which ones are actionable or have value. Just as it would be difficult for someone without a background or training in meteorology to develop weather models, domain expertise in public safety should be a prerequisite for work in this area.

12.1 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,[2] 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 six 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 six 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.

12.2 Experts versus Expert Systems

If data mining and predictive analytics truly are game changing, why haven't they 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.[3]

I have direct personal experience with this phenomenon. Several years ago, I attended a scientific meeting that included a lively debate over expert 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.[4] 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.

12.3 "Normal" Crime

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.

12.4 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 previous 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.

12.5 Strategic Characterization

Generic profiles of suspected suicide bombers have been developed, although the area of strategic characterization of possible or likely offenders has been associated with considerable controversy. These methods of strategic characterization have been used to guide security screening processes, as well as mandatory registration and law enforcement interviewing protocols. A further examination of these models starts to reveal the challenge associated with these general profiles.

While no one would dispute the fact that all of the 9/11 hijackers were of Middle Eastern descent, it is not equally true that all individuals of Middle Eastern descent are probable terrorists. A significant increase in false alarms is a common problem associated with the challenge of modeling infrequent events. Worldwide, the number of likely terrorists is extremely small, something that is magnified further when compared to the total world population. Attempts to further refine and characterize suspected terrorist profiles have included tightening the possible age range and focusing on males. While significantly limiting the relative percentage of possible suspects, these decision rules still catch a very large number of individuals in their net. Two problems emerge from this inability to further refine the model.

First, by creating such a general model, the likelihood of a false positive is increased significantly. Accurate prediction of very low-frequency events can be extremely challenging. The nature of the errors must be considered when evaluating the accuracy of the model. For example, model predicting an event that occurs only 1% of the time would be 99% accurate if it always predicted that nothing would happen. Clearly, this is not acceptable, particularly when public safety and human lives are in the balance. Therefore, a concomitant increase in the number of false alarms generally is associated with acceptable prediction rates for these infrequent events. Given the very low frequency of potential terrorists within the general population and the limited knowledge regarding this comparatively hidden group of individuals, these models are even more challenging to develop, particularly with acceptable levels of accuracy and false alarms.

Second, the numbers identified by these decision rules are so large as to be almost useless for enforcement strategies. Even if the decision rules were refined considerably, the selection process is still likely to include large numbers of individuals. To accommodate this volume challenge, random selection protocols have been implemented. In light of the extremely low frequency associated with the subjects of interest, however, the likelihood that anyone of interest will be identified through this layered process of generic strategic characterization, which is further diluted by a random sampling procedure, is extremely low.

While refinements to these models are under development, the technique of tactical characterization of possible terrorist behavior also is being explored. This approach generally has been accepted by the public, perhaps because there is a sense that if someone is acting suspiciously or strangely, then they should expect to be stopped.

Characteristics of possible pre-incident surveillance or behavior have been compiled. These include various types of surveillance activity and security probes. The Air Force Office of Special Investigations has further categorized and characterized suspicious behaviors, which certainly represents an excellent foundation for further analytical work in this area.[5] Similarly, possible pre-attack behaviors have been characterized. These include nervous behavior and bulky clothing, which might be hiding a bomb vest or other weapon.[6]

These behavioral models also represent a good start, but still lack a certain degree of specificity. For example, while the description of an individual who is overdressed for the prevailing weather conditions make sense to most, it is not a perfect discriminator. In fact, it was not too long ago that it was possible to drive through the streets of Richmond, Virginia, and find groups of young men similarly dressed. Looking like a group of climbers ready to summit Mount Everest, these individuals would be dressed in heavy coats and parkas. They frequently were sweating and looked nervous, not because they were wearing a bomb vest, but because they were on the corner selling drugs and it was really hot. Their reason for wearing the heavy coat was to avoid detection of drugs and weapons during a pat and frisk, but their potential threat to the community was very different than that of a potential suicide bomber.

Therefore, one missing piece in many of these models is a control group, which is a shortcoming in many studies of risk. For example, in medical studies, people are much more likely report a side effect or another type of adverse outcome than they are to let the drug company know that everything worked out fine. While some side effects are extremely rare and do not emerge until large numbers of individual have experience with a particular treatment or medication, this can result in a pattern of bias in outcome studies. Similarly, in law enforcement and intelligence analysis, people are much more likely to report cases when something actually happened, as opposed to suspicious situations that turned out to be benign. This control group, however, is extremely important to the development of meaningful models that will have significant predictive value with future, unknown samples of behavior. Therefore, it can be just as important to include the false alarms in the analysis as it is ensure that the accurate hits are represented. How can we predict that something is likely to happen if we do not know how true incidents differ from false alarms? These comparison data can have tremendous value in that they significantly increase our ability to create models that classify reliably and accurately.

12.6 Vulnerable 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.

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, D.C. 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.

Sheep, Wolves, and Sheepdogs

Lieutenant Colonel (Ret.) Dave Grossman, a global expert on violence and terrorism, is a hero to many in the operational worlds, 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.[7] 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 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.

12.7 Schools

Colonel Grossman has studied the potential vulnerability associated with schools. He has suggested that schools are a particularly desirable target for

Islamic fundamentalists and has cited several examples underscoring the number of international terrorist attacks specifically targeting schools.[8]

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 Belsan, 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 hours later, 300 to 400 were dead.[11]

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.[12] 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 Belsan siege,[13] 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. Colonel Grossman highlights the terrorists' use of 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 Belsan 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 preattack 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. The longer the planning cycle, theoretically the more opportunities for identification, which ultimately supports proactive, information-based prevention, deterrence, and response efforts.

Prior to the attack on the school in Belsan, 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 information gathering had included preoperational surveillance of the school. In his review of the incident, John Giduck noted that the Belsan operation reflected not only the lessons learned from the Nord-Ost Theater attack, but also al Qaeda tactical training, particularly the sections that addressed how to deal with hostages.[14] This preattack planning activity represents opportunities for identification, characterization, and proactive responses to potential threats, including prevention.

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 Belsan 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 particular facility because 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.

Specific issues to consider in information-based risk and threat assessment include the following.

12.8 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.[15] "As predictable as water coming to a boil,"[16] 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.[17] 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. 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. Ultimately, this increases the level of natural surveillance in and around a facility, as well as the willingness to report it.

12.9 Accuracy versus Generalizability

The issue of model accuracy versus generalizability and error types have been previously 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 and been identified and 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.

12.10 "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. A "confusion" matrix can be generated to determine the nature of 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 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.

12.11 Evaluation

Colonel Grossman has discussed U.S. regional responses to mass killings in schools that include mandated "lockdown" drills and a requirement for emergency response plans in schools.[18] 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.

This combined approach highlights the dual metrics that can be used to evaluate effective risk and threat assessment: 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.

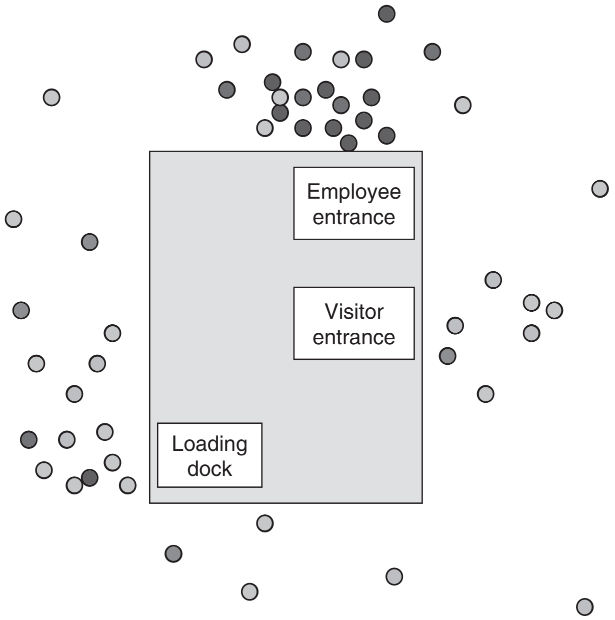
Another excellent example of effective response planning occurred in this country on September 11. Rick Rescorla, Vice President for Security at Morgan-Stanley/Dean-Witter, had the prescience to know that the World Trade Center was at risk for a terrorist attack. While his foreknowledge of the likely method for the attack was startling in its accuracy, it was his insistence on routine evacuation drills that was credited for saving the lives of 2700 of his colleagues in the South Tower.[19]

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 murders) was in the hundreds, yet fire drills and response plans are mandated while similar planning for violence (the greater threat) generally does not exist.[20]

12.12 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 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.

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.*



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.

12.13 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.[21] 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.[22] 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 U.S. 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.[23] 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.*[24] 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.

Interestingly, this concept still exists. The website 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.

12.14 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.
2. McCue, C. and McNulty, P.J. (2003). Gazing into the crystal ball: Data mining and risk-based deployment. *Violent Crime Newsletter*, September, 1–2.
3. Colvin, G. (2006). Ditch the 'experts.' *Fortune*, February 6, 44.
4. Ibid.
5. United States Air Force, Office of Special Investigations (2003). Eagle eyes: Categories of suspicious activities. http://www.dtic.mil/afosi/eagle/suspicious_ behavior.html
6. The International Association of Chiefs of Police (2005). Suicide (Homicide) Bombers, Part I. Training Key #581. The International Association of Chiefs of Police, Alexandria, VA. http://www.theiacp.org/pubinfo/IACP581SuicideBombersPart1.pdf.
7. http://www.blackwaterusa.com/btw2004/articles/0726sheep.html
8. Grossman, D. (2005). Lecture for ArmorGroup, International Training, Richmond, VA, October 31.
9. Dorn, M. and Dorn, C. (2005). Innocent targets: When terrorism comes to school. Safe Havens International, Macon, GA.
10. See Dorn, M. and Dorn, C.; and Giduck, J. (2005). Terror at Belsan. Archangel Group, Royersford, PA.

11. Dorn and Dorn.
12. Grossman.
13. Giduck.
14. Ibid.
15. De Becker, G. (1997). The gift of fear. Dell, New York.
16. Ibid.
17. Grossman.
18. Ibid.
19. Stewart, J. and Stewart, J.B. (2003). Heart of a soldier. Simon & Schuster, New York.
20. Grossman.
21. De Becker.
22. Shermer, M. (2004). 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&colID=13
23. DARPA - FutureMAP Program. Policy analysis market (PAM) cancelled. IWS - The Information Warfare Site; http://www.iwar.org.uk/news-archive/tia/futuremap-program.htm, July 29, 2003.
24. Sontag, S. and Drew, C. (1999). Blind man's bluff: The untold story of American submarine espionage. HarperCollins, New York.

This Page Intentionally Left Blank

Case Examples

This Page Intentionally Left Blank

13

Deployment

In many ways, the goal of force deployment is similar to most resource allocation decisions: try to do the most work possible with existing, or even fewer, resources, and try to have staff in place when and where they are needed. The deployment challenges facing police executives and command staff include maintaining public safety, efficiently allocating increasingly scarce personnel resources, and responding to citizen needs for appropriate police presence in their community.

Effective deployment of police resources can address all three challenges. By placing police units when and where they are likely to be needed, public safety can be increased. This can be achieved by the deterrent effect associated with police presence, as well as by the ability to respond quickly and rapidly apprehend criminals. Concomitantly, deploying police personnel when and where they are likely to be needed decreases the likelihood that they will be deployed frivolously or inefficiently. The inappropriate allocation of public safety resources can have tragic consequences if poor deployment decisions delay a timely response to an emergency. When crime is increasing, the first response can be calls for increased personnel resources and heavy deployment; however, many public safety agencies do not have the luxury of addressing rising crime with the indiscriminate use of resources, particularly given shrinking law enforcement budgets and diminishing personnel resources.

Unfortunately, police deployment decisions often are made based on historical precedence, gut instincts, citizen requests, or decisions made by uniformed policy makers. But addressing the need for police presence can be achieved with effective deployment strategies. Frequently when crime is escalating, the citizen outcry is for increased visibility. The quickest solution often seems to be increased deployment in those areas, increased foot patrols and a reemphasis on "community policing" in an effort to ensure that the citizens know that the police are there for them. 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. Therefore, reducing crime frequently addresses police visibility issues by reducing the perceived need for an additional police presence. Ultimately, this results in an actual decreased need for resource deployment in these areas, which further supports an overall plan of cost-effective resource allocation and savings.

Deployment can be divided grossly into patrol and tactical deployment, depending on the desired goal and associated tasks for these specific units.

13.1 Patrol Services

The general goals of police patrol deployment are to place officers in locations where they can respond rapidly to citizen-initiated work, or calls for service, while engaging in proactive policing. This could include everything from crime deterrence through police presence to routine patrol and participation in community policing initiatives. So, the ideal patrol deployment plan would place resources when and where they are likely to be needed, while simultaneously reducing deployment where the need is more limited. Because personnel resources, especially patrol resources, are the single largest expenditure in almost any public safety agency, appropriate allocation of this resource through effective, proactive patrol deployment can reap numerous benefits from both a personnel and fiscal standpoint.

13.2 Structuring Patrol Deployment

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 call numbers. 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.

13.3 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 numbers 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.

Weather and Climate Data

It was conventional wisdom in one 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 analysts to document this effect and report back to him when we had the numbers. Sure enough, the period during the storm was relatively quiet, save for a few folks trying to buy drugs, while the period immediately after the storm must have looked something like the OK Corral, based on the number of citizen complaints for random gunfire in the community.

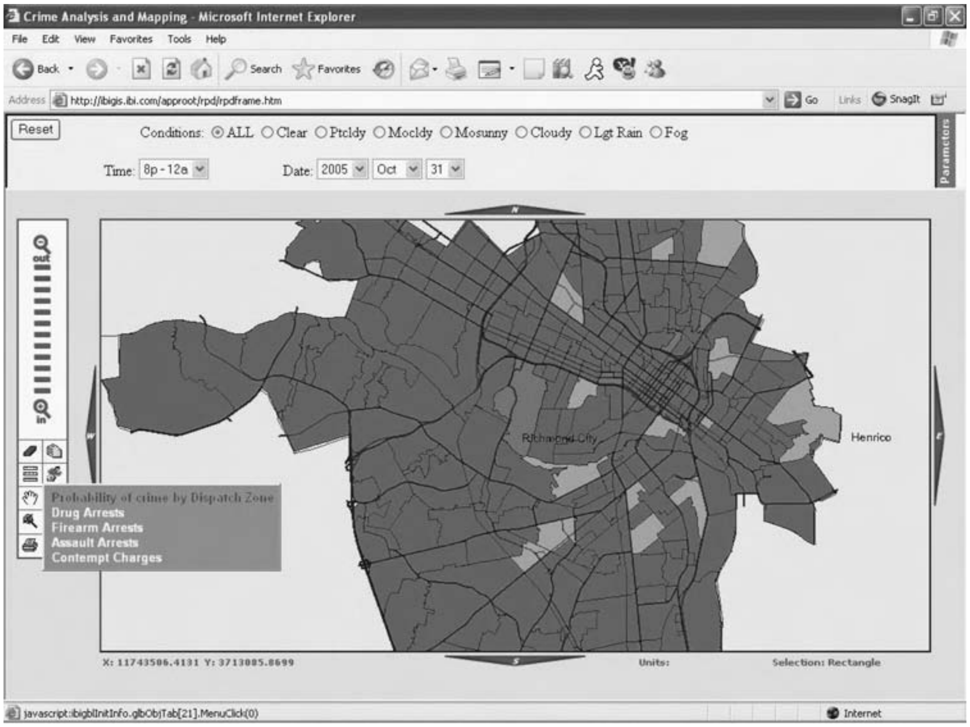
While I could speculate regarding the true meaning behind the association between inclement weather and random gunfire, the truth is that I have absolutely no idea why this occurred in that particular community. The important thing is, however, that the relationship was noted and documented and could be anticipated and effectively responded to in the future. New Orleans, Louisiana, noted similar, unanticipated 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. 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.

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 crunch 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.

Weather data usually are readily available and can add considerable value to a deployment strategy. For example, inclement weather might suppress some types of criminal activity that require that potential victims continue to engage in their routine activities. 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 plan to another, and from routine patrol deployment to a more reactive deployment mode that flexes residential patrol units to major thoroughfares in support of increased traffic workload.

A specific example illustrating how weather can affect crime trends and patterns is shown in Figures 13-1 and 13-2. A software tool that uses past criminal activity to predict areas in which crimes are most likely to occur generated the results in Figure 13-1, which illustrates the probability of crime by dispatch zone. The tool was developed by SPSS and Information Builders and is based on an analytical framework developed by RTI International. Lighter areas are associated with a greater probability of crime, while darker areas reflect a probability of little to no criminal activity. Using operationally actionable analytical

Figure 13-1 *Map visually depicting the probability of crime by dispatch zone in Richmond, Virginia. (Screen image of the Information Builders' Law Enforcement Application.)*



Crime Analysis and Mapping - Microsoft Internet Explorer

Address http://ibigis.ibi.com/approot/rpd/rpdframe.htm

Conditions: ⦿ ALL ○ Clear ○ Ptcldy ○ Mocldy ○ Mosunny ○ Cloudy ○ Lgt Rain ○ Fog

Time: 8p - 12a

Date: 2005 Oct 31

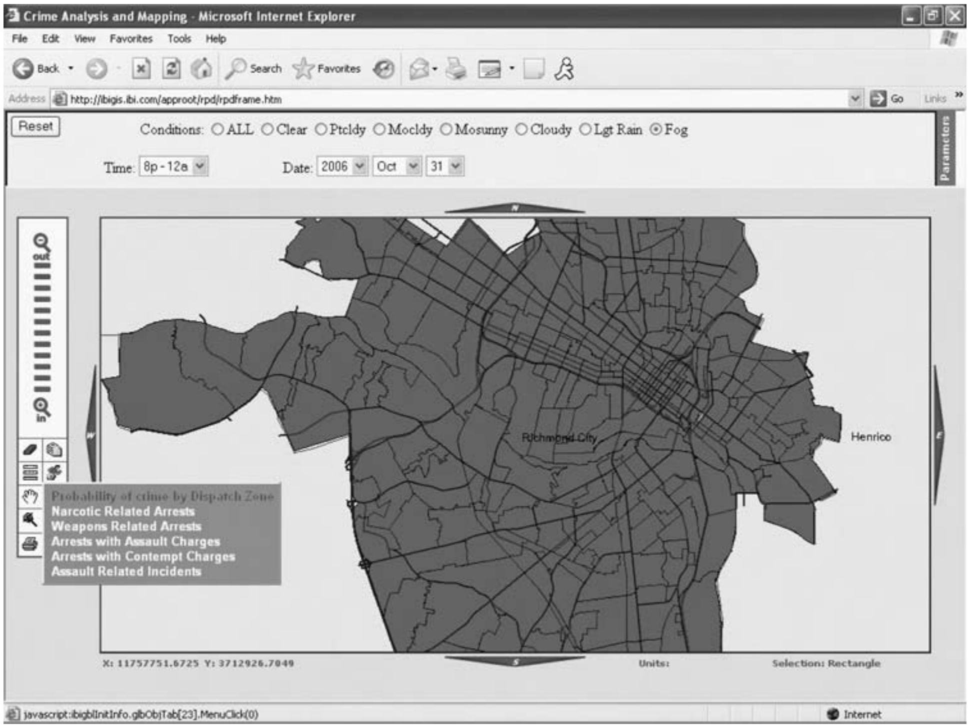
X: 11743506.4131 Y: 3713085.8699

Units:

Selection: Rectangle

javascript:ibigblInitInfo.glbObjTab[21].MenuClick(0)

Figure 13-2 *Map highlighting the reduced probability of crime by dispatch zone, as compared to Figure 13-1, in response to a forecast of fog. (Screen image of the Information Builders' Law Enforcement Application.)*



Crime Analysis and Mapping - Microsoft Internet Explorer

Address http://ibigis.ibi.com/approot/rpd/rpdframe.htm

Conditions: ○ ALL ○ Clear ○ Ptcldy ○ Mocldy ○ Mosunny ○ Cloudy ○ Lgt Rain ● Fog

Time: 8p - 12a

Date: 2006 Oct 31

X: 11757751.6725 Y: 3712926.7049

Units:

Selection: Rectangle

javascript:ibigblInitInfo.glbObjTab[23].MenuClick(0)

products like these, command staff and managers can make information-based decisions regarding the allocation and deployment of patrol resources.

As can be seen in Figure 13-2, adjusting the parameters to change the local weather conditions to "fog" results in a dramatic reduction in the probability of crime throughout the city. Commanders can use images like these to structure patrol deployment and resource allocation. Ultimately, these approaches 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.

Seasonal changes also can be associated with changes in demand in police services. Particularly 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 using data mining and predictive analytics. 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 service provision to the community. The weather department associated with the local news station almost certainly maintains archival weather data in an electronic format that would be suitable for inclusion in deployment modeling and analysis.

Census Data

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 numbers of people living in a particular area. 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 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, or domain expertise. 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.

13.4 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,[1] 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.

Figure 13-3 shows that there were 266 citizen calls for service during the week selected for analysis. More than half of the calls (58%) were confined to the three-day period between Friday and Sunday. This initial analysis already adds

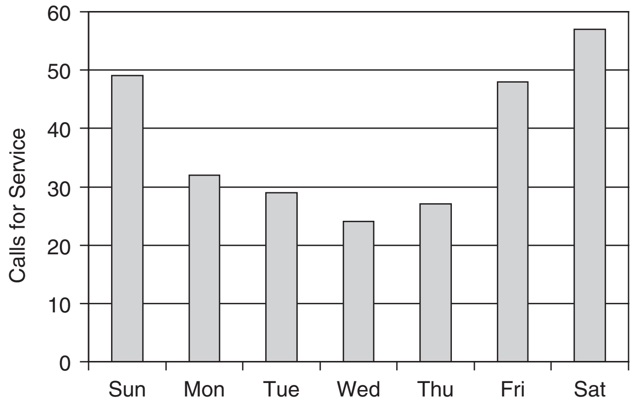
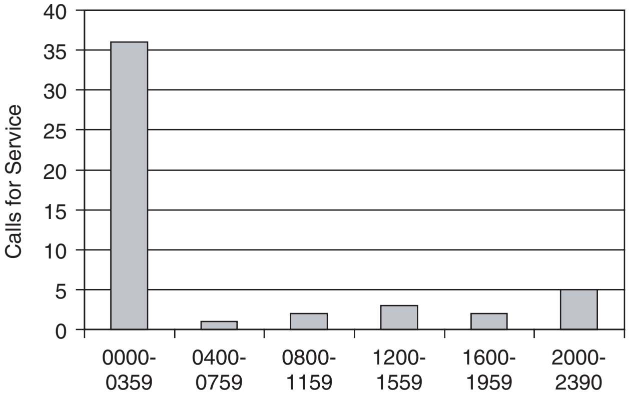


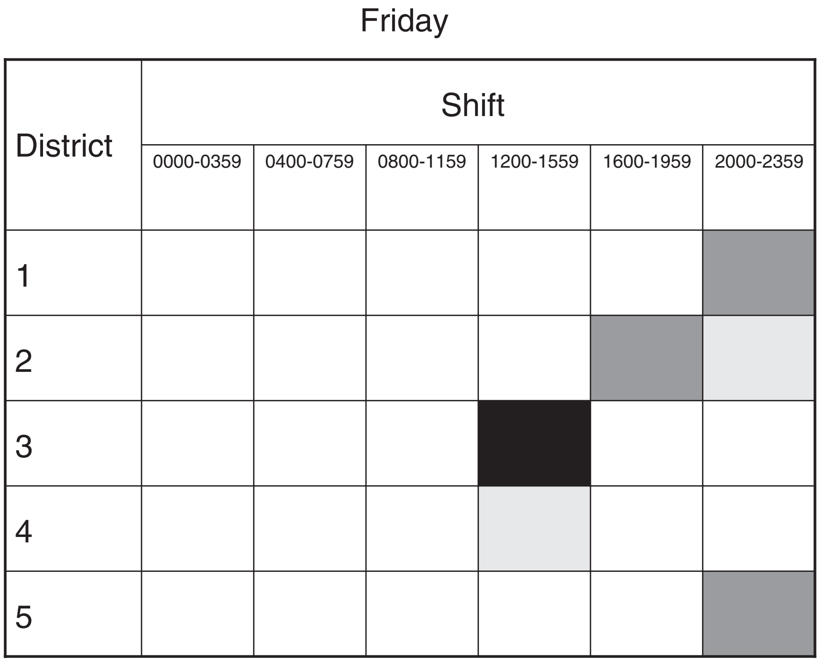
Figure 13-3 *Frequency distribution depicting citizen calls for service by the day of week during a one-week period.*

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 three days out of the week accounted for more than one-half 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 13-4). For example, drilling down into the service calls reported on Sunday reveals that most of them occurred between midnight and 0400 hours. 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.

Figure 13-4
Citizen calls for service from Figure 13-3 by time of day.





Friday

District	Shift					
	0000-0359	0400-0759	0800-1159	1200-1559	1600-1959	2000-2359
1						▒
2					▒	░
3				■		
4				░		
5						▒

Figure 13-5 *This figure illustrates 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 Kohenen network.*

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 13-5 depicts a hypothetical police deployment schedule, which was extracted from a self-organizing network called a Kohenen network, that includes time of day and police district for a particular day of the week, in this case Friday. This model has been 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, Third District is associated with the greatest need on Fridays, from 1200 to 1359 hours. 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.

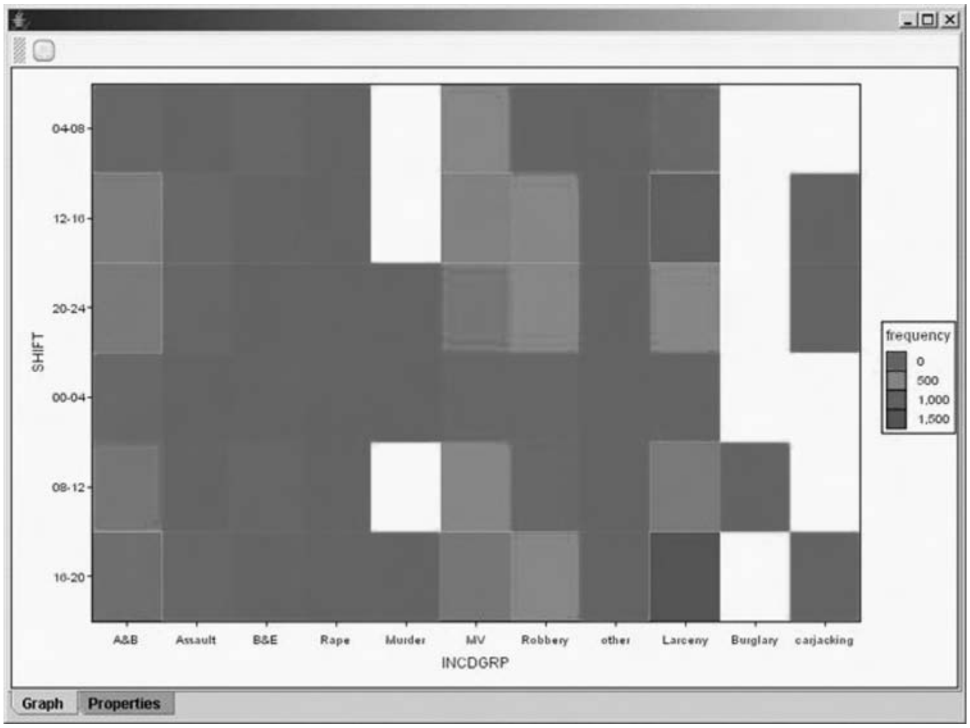
The increased demand in Second District was linked to a regular after-work party each Friday near the business corridor, while the increased demand in

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 13-5, the analytical team was able to deploy the results of their data mining 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.

Figure 13-6 depicts a heat map (see Chapter 9) of different crimes over four-hour time blocks. These results indicate an increased likelihood of larcenies between 1600 and 2000 hours, while burglaries are more likely to occur between 0800 and 1200 hours.

Figure 13-6 *This type of graph is referred to as a heat map, because relative differences in intensity depict relative differences in frequency or probability, similar to the relative differences used in temperature gradients. In this illustration, higher intensity conveys an increased likelihood of the particular type of crime (B. Haffey, SPSS, Inc.; used with permission).*



This example was created to be relatively simple in an effort to highlight specific points. In the applied settling, a sampling frame longer than one 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 previously. 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.

13.5 Tactical Deployment

Many agencies have tactical units that supplement standard patrol deployment. These units are not necessarily anchored to any particular geographic areas, like patrol units, which allows them to be rapidly mobilized and respond to crimes quickly and aggressively. Placing these tactical units when and where they might be needed serves two functions. First, response time is decreased, which increases the possibility of rapid identification and apprehension of the suspect. Second, and perhaps even more preferable, heavy deployment of these units in areas anticipated to be at greater risk for trouble might provide deterrence. As nice as it is to get bad guys off the streets, it is even better in many situations to prevent crime.

Identifying the Problem

One area where these units can be particularly effective is with narcotics-related or violent crime, which are linked in many situations. The basic approach to guiding tactical deployment with data mining and predictive analytics bears many similarities to the approaches used with traditional patrol deployment. The first step in the process generally involves the use of exploratory graphics in an effort to characterize the problem and drilling down in an effort to identify patterns of associated risk and possible solutions.

Several years ago, review of the homicide data from Richmond, Virginia, showed an extremely high homicide rate that placed it repeatedly in the top ten 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.[2] 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.[3]

Why is this important from a tactical 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. Drugs, violence, and drug-related violence can be addressed with relative efficacy through the use of strategies and initiatives that incorporate the use of tactical patrol units. 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.

13.6 Risk-Based Deployment Overview

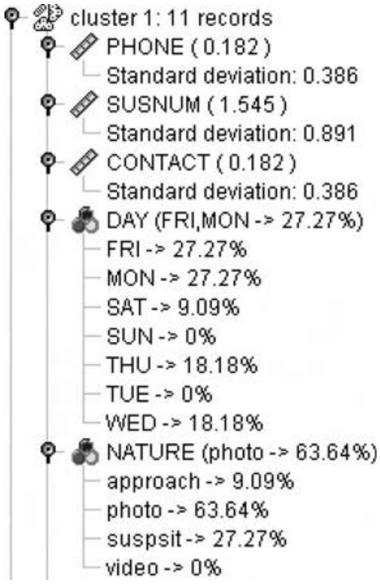
More recently, the Richmond Police Department has been exploring the use of “risk-based” deployment strategies as part of their role in the Project Safe Neighborhoods initiative with the United States Attorney General’s Office in the Eastern District of Virginia.[4] A detailed example of this approach is discussed later in this chapter, but the fundamental idea behind this approach is that if a model of relative risk can be created, personnel resources can be proactively deployed in response to this anticipated risk. In other words, if it is possible to characterize and anticipate violent crimes, tactical units can be deployed proactively and make rapid apprehensions due to their proximity in time and space to the predicted crime locations. Ultimately, increased use of this method provides an opportunity to move into proactive policing, whereby crime can be deterred or even prevented.

13.7 Operationally Actionable Output

As has been stated several times before, any analysis, no matter how elegant, sophisticated, or accurate, has no value in an operational environment if it is not actionable. The output shown in Figure 13-7 has limited value in an operational setting in its current form. In order for it to gain value it needs to be actionable; it needs to be apparent from looking at the analytical output which course of action would be best. This is particularly true with deployment. Outlined in the following sections are two possible mechanisms for deploying analytical output products directly into the operational environment in a format that can be used readily.

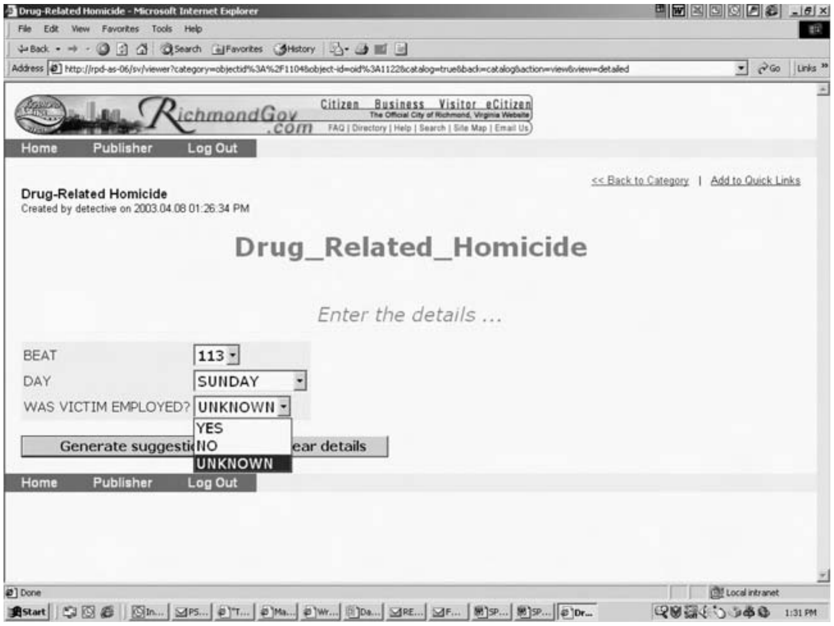
Web-Based Analytics

New improvements in the deployment of predictive analytics outputs and rule sets now allow the use of web-based analytics.[5] With these systems, an analyst is able to analyze data and create rule sets or scoring algorithms that can be



```
cluster 1: 11 records
    PHONE ( 0.182 )
        Standard deviation: 0.386
    SUSNUM ( 1.545 )
        Standard deviation: 0.891
    CONTACT ( 0.182 )
        Standard deviation: 0.386
    DAY (FRI,MON -> 27.27%)
        FRI -> 27.27%
        MON -> 27.27%
        SAT -> 9.09%
        SUN -> 0%
        THU -> 18.18%
        TUE -> 0%
        WED -> 18.18%
    NATURE (photo -> 63.64%)
        approach -> 9.09%
        photo -> 63.64%
        suspsit -> 27.27%
        video -> 0%
```

Figure 13-7 *Some analytical output products, while accurate, can be very difficult to interpret and deploy within an operational environment (B. Haffey, SPSS, Inc.; used with permission).*


Drug-Related Homicide - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://rpd-as-06/sv/viewer?category=objectid%3A%2F1104&object-id=oid%3A1122&catalog=true&back=catalog&action=view&view=detailed
Go Links
RichmondGov.com
Citizen Business Visitor eCitizen
The Official City of Richmond, Virginia Website
FAQ | Directory | Help | Search | Site Map | Email Us
Home Publisher Log Out
<< Back to Category | Add to Quick Links
Generate suggesti
ear details
Home Publisher Log Out
Done
Local intranet
Start
1:31 PM

Drug-Related Homicide
Created by detective on 2003.04.08 01:26:34 PM

Drug_Related_Homicide

Enter the details ...

BEAT	113
DAY	SUNDAY
WAS VICTIM EMPLOYED?	UNKNOWN

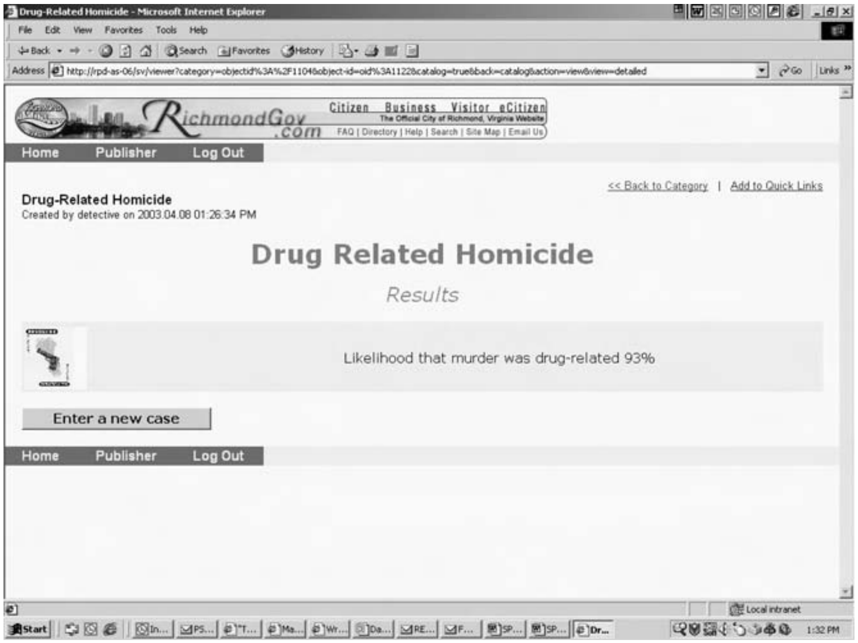
YES
NO
UNKNOWN

Figure 13-8 *Web-based data entry screen that can be used for remote access to scoring algorithms. (Screenshot of web interface taken by the author is from Cleo, SPSS, Inc.)*

deployed remotely. The end users are then able to log on through a secure Internet or intranet connection. As illustrated in Figure 13-8, they are directed to a data entry screen, which prompts them to enter a few relevant details through the use of pull-down menus. For example, a supervisor can enter information relative to the time, day, and assigned location associated with her shift. Based on the calculated risk for each area within her purview (Figure 13-9), the supervisor then can proactively deploy resources in anticipation of likely activity or anticipated risk.

This system is 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 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 situational needs of the end user.

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


Drug-Related Homicide - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://rpd-as-06/sv/viewer?category=objectid%3A%2F1104&object-id=oid%3A1122&catalog=true&back=catalog&action=view&view=detailed
Go Links
RichmondGov.com
Citizen Business Visitor eCitizen
The Official City of Richmond, Virginia Website
FAQ | Directory | Help | Search | Site Map | Email Us
Home Publisher Log Out
<< Back to Category | Add to Quick Links
Enter a new case
Home Publisher Log Out
Local intranet
Start
1:32 PM

Drug-Related Homicide
Created by detective on 2003.04.08 01:26:34 PM

Drug Related Homicide

Results

Likelihood that murder was drug-related 93%

Figure 13-9 *Web-based output generated by a remotely deployed scoring algorithm in response to the input depicted in Figure 13-8. (Screenshot of web-based output taken by the author is from Cleo, SPSS, Inc.)*

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 centers that can be used by multiple operational units. Crime analysis functions can be centralized and available to remote precinct locations and offsite specialized units, 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, as can be seen in Figure 13-10, utilization of centralized analytical capacity and remote deployment of scoring algorithms and models also facilitates the integration of different types of data and information. Crime frequently transcends traditional offense definitions and arbitrary boundaries.[6] Prostitutes use drugs, drug users rob convenience stores, and gang members commit violent crimes. While a certain degree of operational specificity often

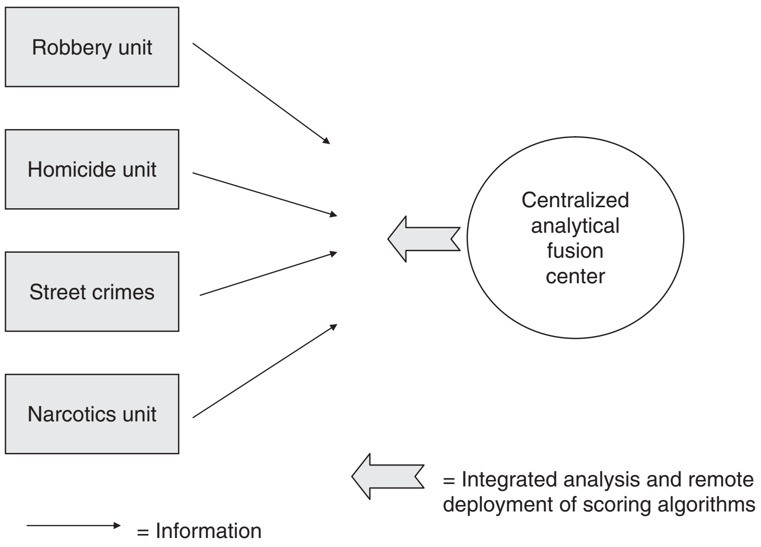

Robbery unit
Homicide unit
Street crimes
Narcotics unit
Centralized analytical fusion center
= Integrated analysis and remote deployment of scoring algorithms
= Information

Figure 13-10 *Centralized analytical capacity and the remote deployment of scoring algorithms and models can facilitate the integration and deployment of different types of data and information.*

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 through the use of these web-based deployment tools.

Web deployment of data mining models also permits the use of more complex scoring algorithms, because the scoring algorithms do 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 the limited amount of information for use in the predictive rule set. 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.

Mapping

The ability to effectively convey the results of data mining can determine whether the results are used operationally or not. If the command staff, operators, or other end users are not able to interpret the results, 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 as actionable information. One excellent tool that can be used to add value to the results of data mining is mapping.

Maps can be thought of as specialized graphs that convey information within an operational context.

Deployment of data mining and predictive analytic products through mapping are highlighted throughout this text. From a deployment standpoint, however, there are few systems that even approach the effectiveness and functional utility associated with mapping. Through the use of color, shape, shading, and other options, a tremendous amount of information can be conveyed through a two-dimensional mapping product.

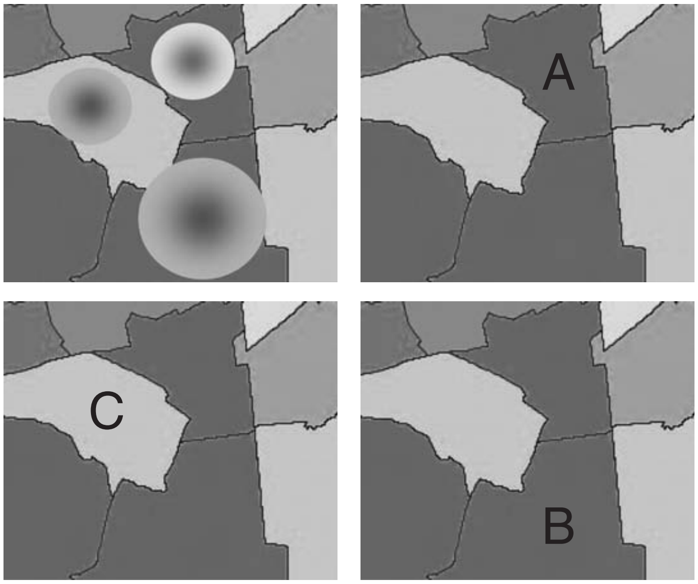
Many patterns of offending tend to be geographically dependent. For example, certain locations are associated with open-air drug markets, while others might be linked to an increased likelihood of robberies. This information has tremendous value from a deployment perspective because, at its most primitive, deployment generally involves placing personnel resources in time and space. Ideally, personnel are deployed in response to anticipated patterns of offending or calls for service, something that data mining and predictive analytics can facilitate. Through the creative use of mapping tools, meaningful differences in time and space can be conveyed in an actionable format to operational personnel.

Data mining and mapping tools are evolving at a rapid pace. At the time of this writing, data mining tools were being developed specifically to deploy generated algorithms directly into a mapping environment. While it is possible to do this with "brute-force" techniques now, this alternative approach will allow for a more detailed depiction of risk, as well as the inclusion of a larger number of parameters. For example, without these enhanced mapping techniques, the analyst would be required to develop one map for each four-hour time block during each day of the week; that amounts to 42 different maps to convey a week's worth of information. This would be a tremendous amount of work and would severely limit the amount of detail in each map. Moreover, the necessary flipping between maps would be operationally cumbersome.

The ability to deploy data mining algorithms directly through mapping products would decrease the amount of work required for each map, shifting the focus to the value of the algorithm and increasing the amount of detail available with each map. Ultimately, using these techniques, it would be possible to create a dynamic map that could depict "clouds" of risk moving throughout the community depending on the time of day and day of week, similar to the weather maps that highlight changing patterns and trends that also change with time and day. Ideally, the command staff and other planners would be able to scroll fluidly through these maps, determining deployment while considering a variety of different scenarios. The use of touch-screen technology that could activate additional embedded models would further enhance this rapidly developing technology. The opportunity to develop the predictive algorithms in a centralized location with access to a range of data and intelligence would increase the value of these tools even more. Deployment of these maps directly into the field would provide actionable analysis directly into the theater of operations, thereby concurrently increasing situational awareness and informed decision making. These dynamic crime deployment maps would require a greater degree of deployment fluidity, but would maximize personnel resources in a manner not currently possible.

Figure 13-11 depicts a temporally compressed version of this type of map. This figure depicts three temporally distinct patterns of crime, associated with unique locations. By using different sizes of risk "clouds," as well as differential shading, differing levels of risk or anticipated frequency can be conveyed within

Figure 13-11 *These maps illustrate the distribution of crime over time and space. Relative differences in intensity and area on the map correspond to relative differences in risk and indicate a need for greater police resources in the residential area, as compared to the business corridor, and finally the nightclub area. Specifically, a greater level of activity is associated in the nightclub area by the beach on weekend nights (A), while a need for increased deployment shifts over to the business corridor during weekdays (B), and the associated residential areas on weekday evenings (C).*



a two-dimensional representation of information. Review of this map highlights relative changes in expected criminal activity or citizen-initiated complaints. Specifically, a greater level of activity is associated with the nightclub area by the beach on weekend nights (A), while a need for increased deployment shifts to the business corridor during weekdays (B), and the associated residential areas on weekday evenings (C). Relative differences in intensity and area on the map correspond to relative differences in risk and indicate a need for greater police resources in the residential area, as compared to the business corridor, and finally the nightclub area.

Ideally, this information would be depicted in a series of maps that would transition smoothly and further illustrate the relatively fluid nature associated with the movement of crime throughout a community across the time of day and day of the week. This kind of direct linking of data mining and mapping software further increases the amount of information that can be conveyed, as compared to the brute-force techniques that require manual entry of the various densities depicting risk.

Through the creative use of mapping and information-based deployment strategies, resources can be allocated to specific locations associated with increased activity or risk. They also can help further characterize the problem and guide additional targeted responses. For example, increased activity linked to a particular location within a multifacility complex might suggest the need for some environmental changes, such as increased lighting, in addition to heavier patrol in the area. On the other hand, activity associated with a single unit in the same building might suggest a different type of intervention, such as counseling or eviction procedures for the identified tenants, which would not necessarily require additional patrol.

Mapping need not be confined exclusively to traditional geographic boundaries. For example, "maps" can be generated 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. This type of map can be extremely useful for personnel deployment or specifically targeted crime prevention strategies. For example, in Figure 13-12, police calls for service are concentrated on one end of the Happyland Apartments building. Additional investigation might indicate the need for increased lighting in this area or perhaps a review of possible problem tenants. Further refining the map by including a time of day or day of week dimension might reveal additional information regarding this citizen-initiated police work that would have direct implications for deployment, as well as possible proactive crime prevention work that could reduce the complaint burden associated with this particular location.

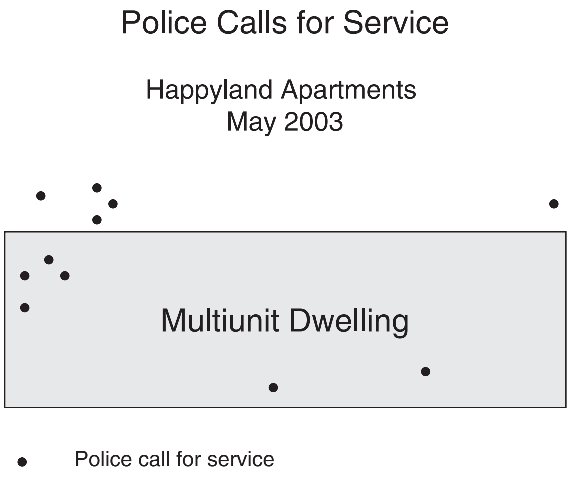


Figure 13-12 *Police calls for service can be mapped in an effort to illuminate possible cause and support related crime prevention strategies.*

Depicting information in a mapping format adds value to the information by placing it within a spatial context and facilitates information-based deployment decisions, as opposed to traditional reactive responses.

Further enhancements with predictive algorithms might provide additional guidance regarding which approaches could be particularly well-suited to a particular challenge. Similar to treatment matching in medicine and psychotherapy, we now know that there is no "one size fits all" approach to crime reduction. These types of analyses are conducted regularly within the public safety environment; however, the combination of data mining and integrated mapping further automates this process, which reduces the amount of time and effort necessary to conduct these types of value-added, operationally actionable analyses. Therefore, in the Happyland Apartments example, evaluation of previous outcomes in similar situations might provide additional guidance regarding the approach or intervention most likely to succeed under this particular set of conditions.

13.8 Risk-Based Deployment Case Studies[7]

"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. 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.

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. As the thinking goes, if armed robberies are bad, then an armed robbery where the victim gets hurt is worse.[8] 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 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, was 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,[9] 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.

Because risk-based deployment maps are investigatively sensitive, the map depicted in Figure 13-13 does not include real data. Rather, it depicts how a possible distribution of robbery-related aggravated assault risk might look throughout a community.

The differential grades of shading on this map highlights areas that are predicted to be at greater risk for a robbery-related firearms assault, based on a review of the armed robbery data for a six-month period. It is important to note, however, that the areas 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

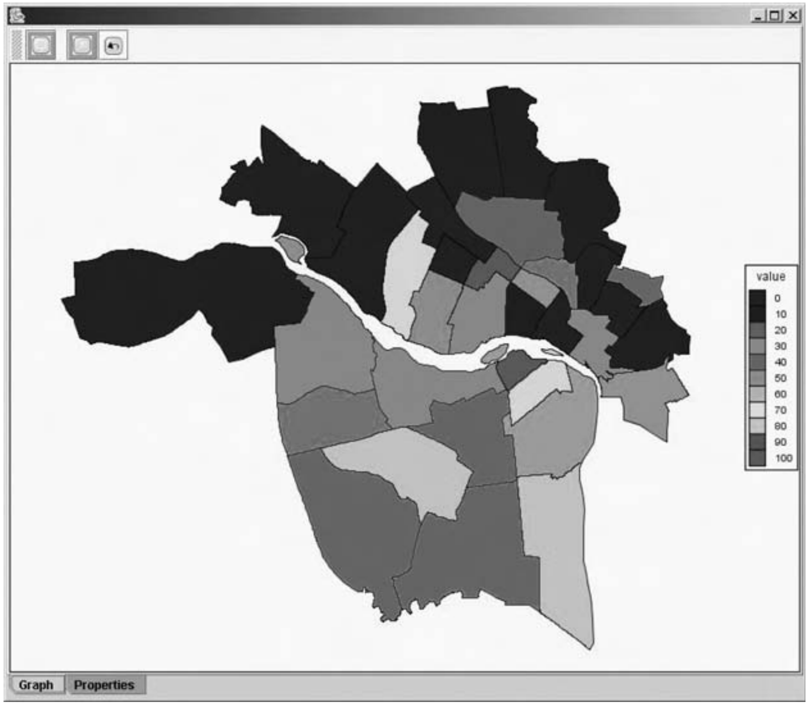


Figure 13-13 *This figure illustrates a fictional risk-based deployment map. The differential grades of shading used on this map highlights areas that are predicted to be at greater risk for a robbery-related firearms assault (B. Haffey, SPSS, Inc.; used with permission).*

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. 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.

With even more advanced technology, time can be added as a dynamic feature in these maps. The supervisor or commander can scan through the various maps in a very efficient manner, easily seeing locations and times where heavier deployment might be required. This affords a tremendous amount of extremely fluid information that can be used for targeted troop deployment in direct response to anticipated risk or workload.

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.

These models evaluated reasonably well using training and test samples; however, it was difficult to generate a measurable outcome, particularly given the nature of this relatively low-frequency event. The question inevitably is asked, "How has this 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.[10]

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. Briefly, Project Safe Neighborhoods is a violence reduction strategy that has been instrumental in supporting the use of data mining and risk-based deployment strategies in the development of innovated approaches to reducing gun violence.

To create this targeted strategy, 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 hours on New Year's Eve and 0200 hours 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 eight-hour 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 elevated 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 two-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 two 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 eight-hour 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 a economic savings of approximately $15,000 in personnel costs and associated holiday pay during an eight-hour period. Further information related to the outcome evaluation associated with this very successful initiative has been addressed in greater detail in Chapters 6 and 8.

Homeland Security and Deployment

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. Moreover, 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.

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. Data mining, predictive analytics, and similar information-based deployment strategies facilitate the provision of more science and less fiction in personnel deployment. Similarly, further enhancements and integration of data mining and mapping software offer additional opportunities for the development of actionable deployment strategies that can move from the analysis unit directly into the operational environment.

13.9 Bibliography

1. Helberg, C. (2002). Data mining with confidence, 2nd ed. SPSS, Inc., Chicago, IL.
2. McLaughlin, C.R., Robinson, D.W., and Faggiani, D. (1998). Declining homicide rates in the 1990s: Not everywhere! *Academy of Criminal Justice Sciences*.
3. McLaughlin, C.R., Yelon, J.A., Ivatury, R., and Sugerman, H.J. (2000). Youth violence: A tripartite examination of putative causes, consequences and correlates, *Trauma, Violence & Abuse*, **1**, 115–127.
4. McCue, C. and McNulty, P.J. (2003). Gazing into the crystal ball: Data mining and risk-based deployment. *Violent Crime Newsletter*, September, 1-2.
5. McCue, C. and Parker, A. (2004). Web-based data mining and predictive analytics: 24/7 crime analysis, *Law Enforcement Technology*, **31**, 92–99.
6. Faggiani, D. and McLaughlin, C.R. (1999). A discussion on the use of NIBRS data for tactical crime analysis. *Journal of Quantitative Criminology*, **15**, 181–191.
7. McCue and McNulty.
8. Ibid.
9. McCue and Parker.
10. McCue, C., Parker, A., McNulty, P.J., and McCoy, D. (2004). Doing more with less: Data mining in police deployment decisions. *Violent Crime Newsletter*, U.S. Department of Justice, Spring, 1, 4-5.

This Page Intentionally Left Blank

14

Surveillance Detection

"We can learn even from our enemies."
Ovid

14.1 Surveillance Detection and Other Suspicious Situations

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. To address 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.

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."

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, 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.

In many ways, surveillance detection techniques can play a significant role in traditional crime analysis. Preoperational surveillance is not unusual for 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. 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 make a rapid apprehension. 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.

Preoperational or hostile surveillance generally is intended to be covert or to appear relatively innocuous to uninformed observers. Frequently, it is only when 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 have suggested increased interest in facilities in northwest Washington state.[1] These reports outline 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.

Terrorist groups, including Al Qaeda, historically have shown a preference for multiple, simultaneous, yet geographically distinct attacks. Examples of this behavior include the African embassy bombings, the 9/11 attacks,[2] and the London and Madrid transportation bombings. Subsequent analysis in each of these incidents revealed extensive, long-term preoperational surveillance of the targets, including a "dry run" in London several weeks before the attacks. Similarly, the casing reports collected on financial institutions within the United States indicate increased activity on Wednesdays.[3] Again, 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 idea of coordinated activities and common planning.

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 data mining strategies can be used to maximize surveillance detection resources.[4] Like patrol deployment, the use of data mining 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 Natural Surveillance

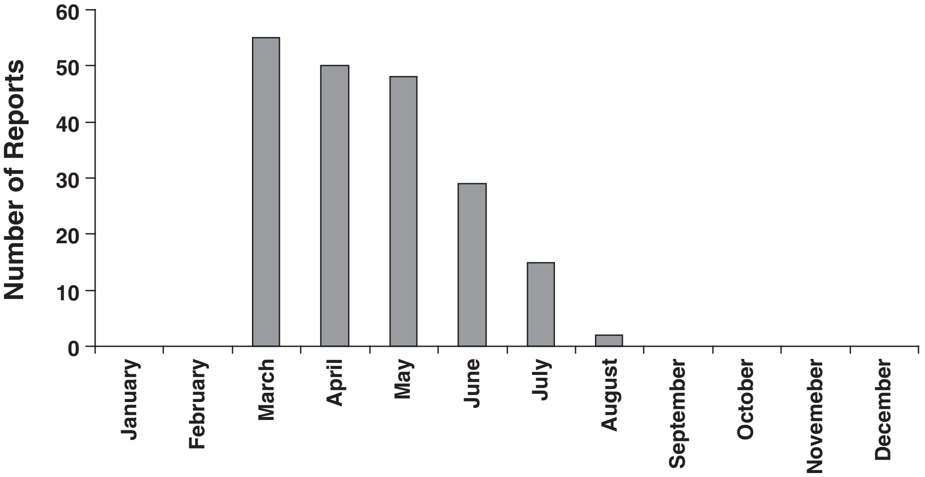
It is not unusual to interview witnesses after a major event and have them recount unreported suspicious behavior that indicated something bad was about to happen. In fact, Gavin de Becker in *The Gift of Fear* recounts cases of workplace violence in which the event was so anticipated that as soon as the shooting started people correctly identified the suspect before the actual nature of the event was even known.[5] Similarly, during the 9/11 inquiries, reports surfaced outlining unusual or troubling pre-incident 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. And 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.

Plotting the total number of suspicious situation reports is a good first step in the process of using analysis to identify possible surveillance behavior. Therefore, it is very important to support consistency in reporting if at all possible. For example, in Figure 14-1, 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 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

Figure 14-1 *Graph of suspicious situation reports over time.*



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.*

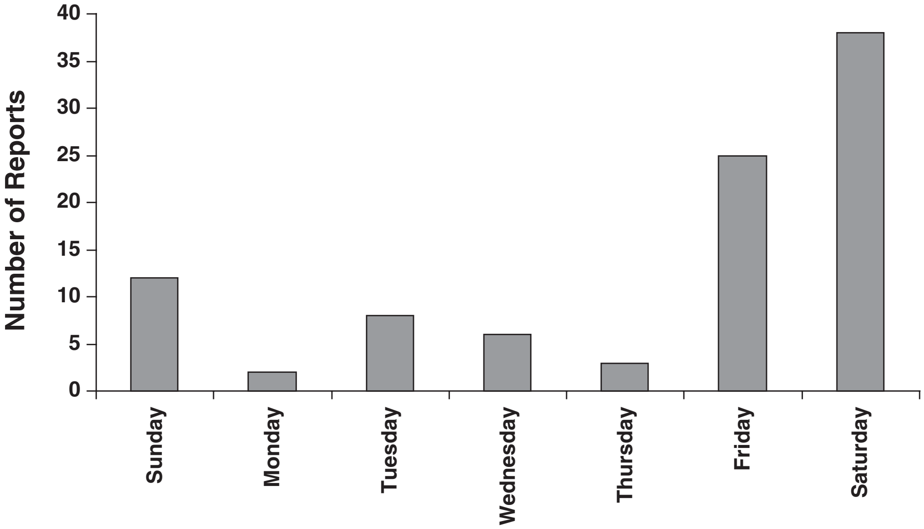
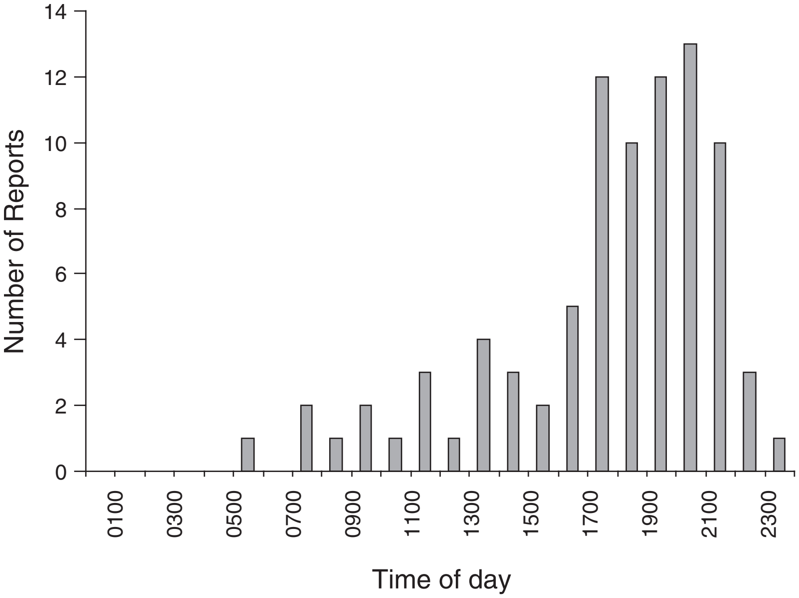


Figure 14-3 *Graph of same suspicious situation reports depicted in Figure 14-2, by time of day.*



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.

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, 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 presence, while concomitantly enhancing the perception of increased public safety in that area.

14.3 Location, Location, Location

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 information (e.g., exact location) in a standardized format, they can be thought of as 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.

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 additional 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.

For example, 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. 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 multi-building facilities or complexes, like the mall in this 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.

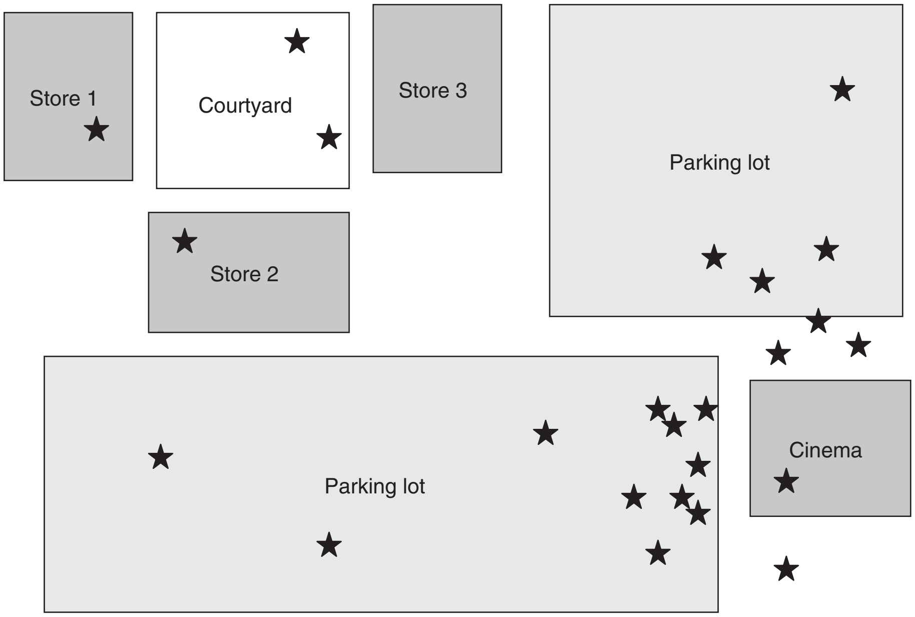
Sophisticated mapping software, although generally very beneficial and frequently used by most public safety agencies, is not entirely necessary for an analysis of this nature. Internet-based mapping tools, orthophotography images, or even line drawings such as the one shown here all convey the necessary 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.

By creating a map of the report locations throughout the mall, 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 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 is 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

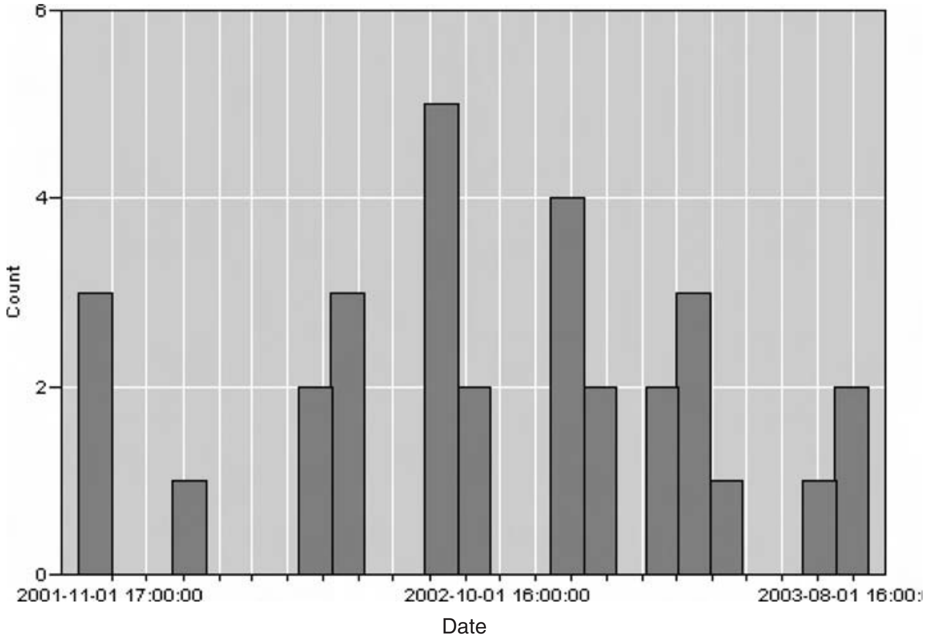
Figure 14-4 *Map illustrating specific locations associated with the situation reports from Figures 14-2 and 14-3.*



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.

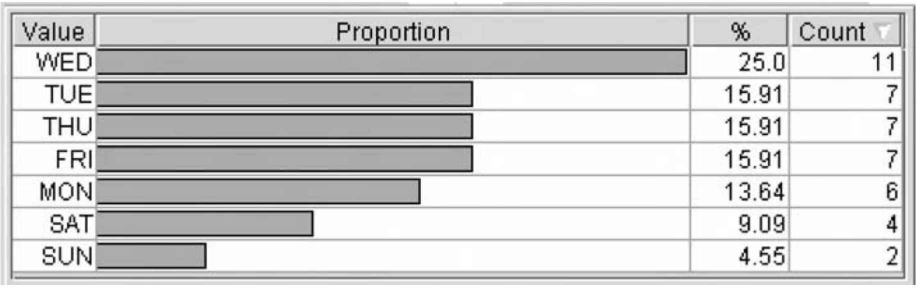
Figure 14-5 *Frequency distribution depicting the relative occurrence of suspicious activity reports associated with a facility of interest. (Screenshot of output taken by the author is from Clementine 8.5, SPSS, Inc.)*



In another situation, data mining and predictive analytics were used to characterize possible surveillance activity associated with a facility of interest. In this example, suspicious situation reports had been investigated and then compiled, although never analyzed. As can be seen in Figure 14-5, a 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 14-6).

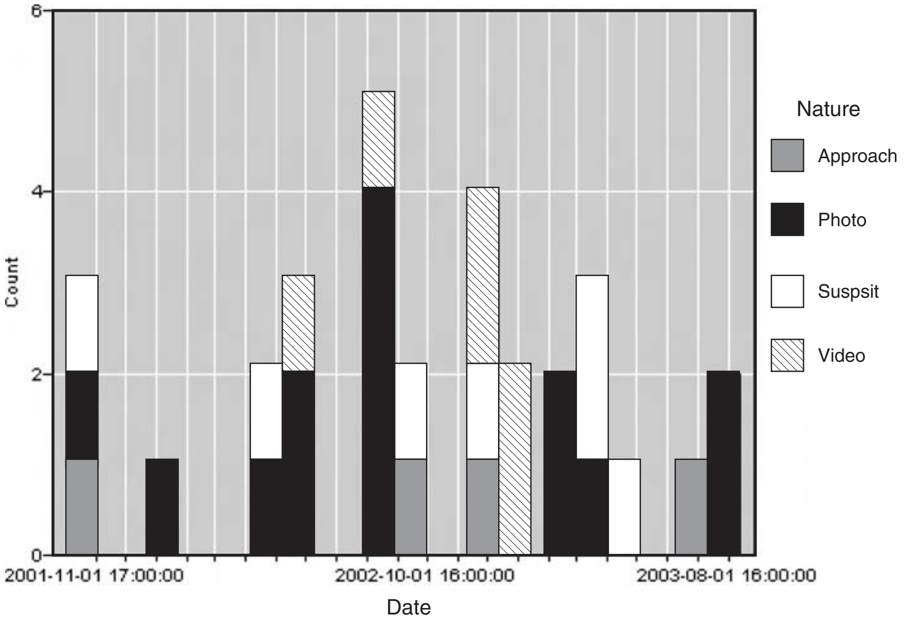
The incidents were recoded into operationally relevant categories that more accurately described the suspect behavior. These included still photography ("photo"); video photography ("video"); any movement toward the facility, attempted interaction with the security personnel, or probing of the perimeter ("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 14-7, any interpretation of these results was limited by the complexity of the graph created.

Figure 14-6 *Distribution of suspicious activity by day of week. (Screenshot of output taken by the author is from Clementine 8.5, SPSS, Inc.)*



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

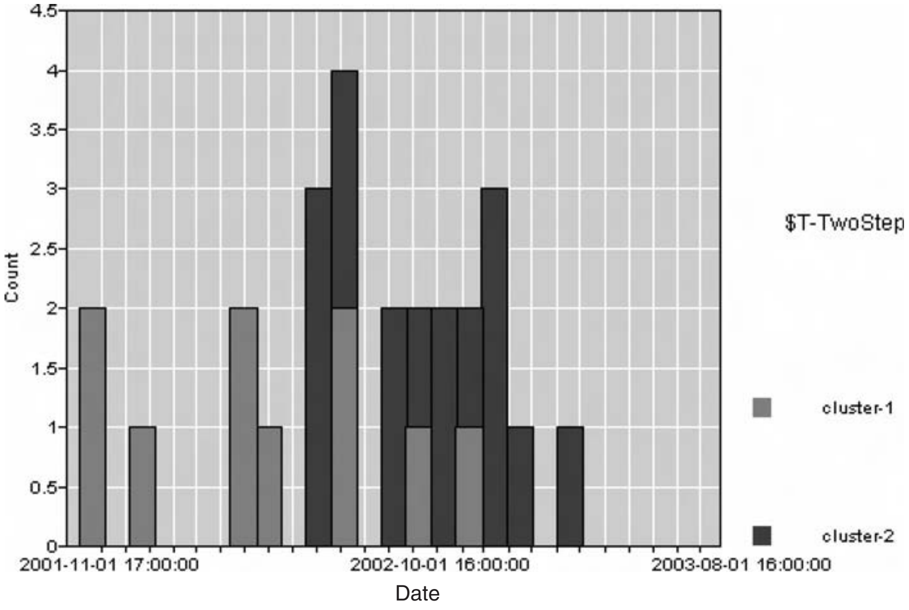
Figure 14-7 *This figure depicts the distribution of suspicious behavior over time. "Approach" indicates that the suspect physically approached the facility or attempted to probe the security features or personnel, "Photo" indicates the suspect use of still photography, "Video" refers to the suspect use of video photography, and "Suspsit" includes all other behavior not included in the previous groups. (Screenshot of output taken by the author is from Clementine 8.5, SPSS, Inc.)*



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. This analysis revealed two different groups of suspicious situation incidents, which generally were associated with different types of observed behavior (e.g., still photography versus video and other operationally oriented surveillance). As can been seen in Figure 14-8, 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.

Using relatively simple techniques, it was possible to generate operationally actionable output from the analysis. As illustrated in Figure 14-9, preparation of a crude facility map highlighted the relative spatial distribution of the incidents.

Figure 14-8 *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. (Screenshot of Two-Step output taken by the author is from Clementine 8.5, SPSS, Inc.)*



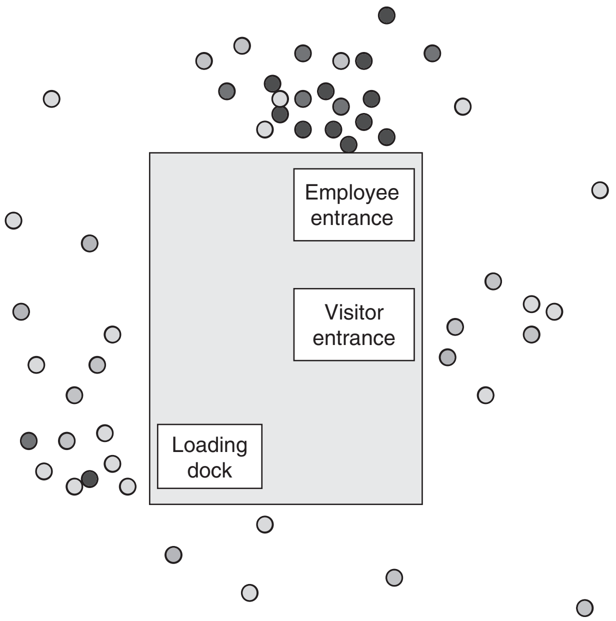


Figure 14-9 *"Map" depicting the spatial distribution of suspicious activity around the facility of interest. Note that the shade of the icon is associated with the cluster membership, which visually highlights the spatial focusing of interest over time.*

Additional value was added to the map through the use of using 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.

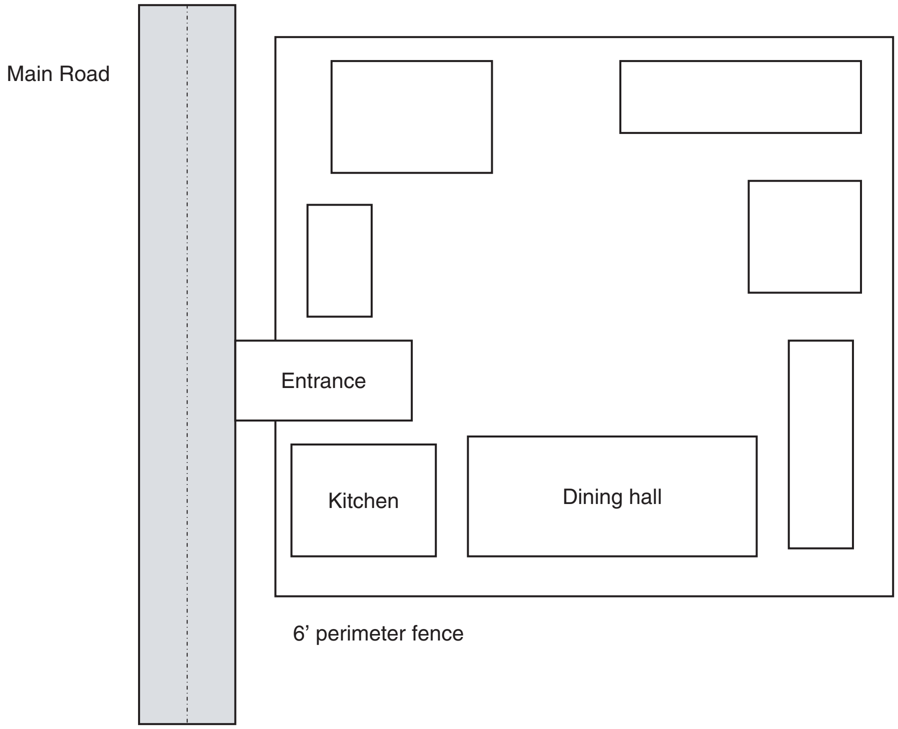
As outlined in this case study, the techniques do not need to be fancy or sophisticated. Rather, the key is to convey analytical output and information in a format that is relevant to the end user and immediately actionable in the applied setting. For example, the use of risk-based deployment maps[6] or "schedules" provides operationally actionable analytical products that can be given directly to personnel in the field. The ability to integrate and analyze data from multiple, disparate locations can further enhance our understanding, particularly regarding those groups and organizations with a historical preference for multiple, simultaneous, geographically distinct attacks. In this situation, 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.

14.4 More Complex Surveillance Detection

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. 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 14-10. The facility is surrounded by a six-foot perimeter fence (Figure 14-11). There is only one point of access to the facility, through

Figure 14-10 *Map of a fictitious multifacility complex associated with suspicious activity.*



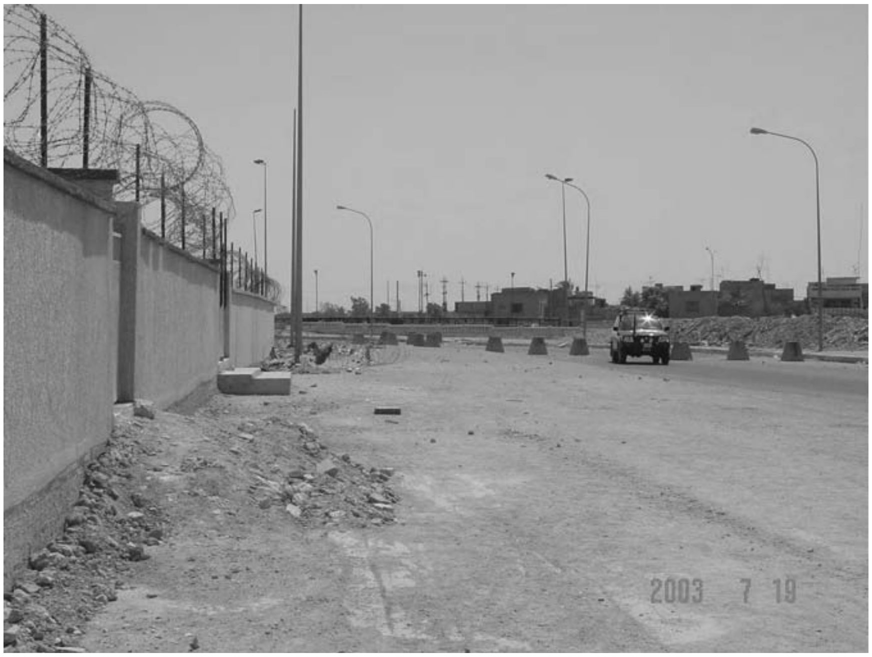
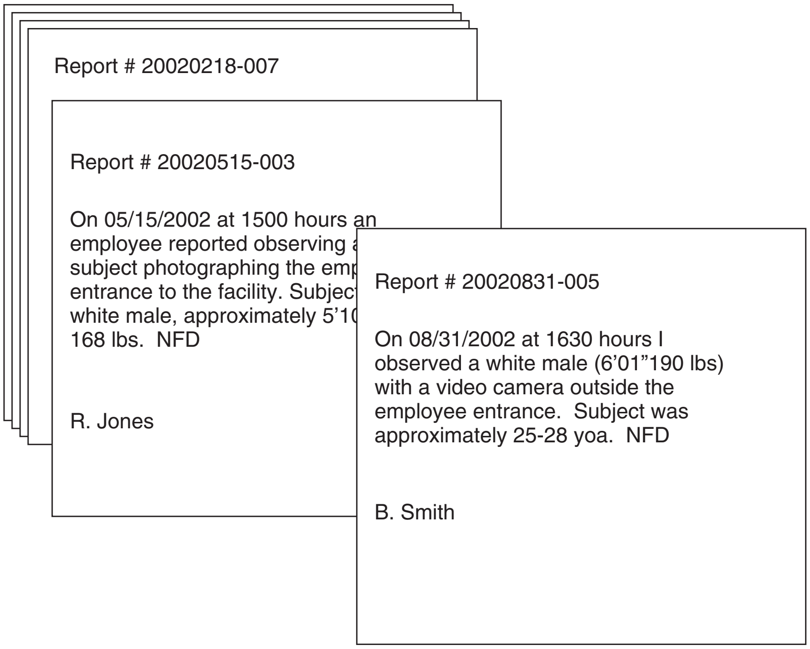


Figure 14-11 *Six-foot fence surrounding the perimeter of the compound. (Staff Sergeant Tom Ferguson, USMC; used with permission.)*

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 14-12).

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 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,



Report # 20020218-007

Report # 20020515-003

On 05/15/2002 at 1500 hours an
employee reported observing a
subject photographing the emp
entrance to the facility. Subjec
white male, approximately 5'1(
168 lbs. NFD

R. Jones

Report # 20020831-005

On 08/31/2002 at 1630 hours I
observed a white male (6'01"190 lbs)
with a video camera outside the
employee entrance. Subject was
approximately 25-28 yoa. NFD

B. Smith

Figure 14-12 *Samples of suspicious situation reports received and logged by security.*

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" on Figure 14-13, was characterized by activity outside the perimeter. This was frequent, as indicated by larger clouds of risk on the diagram. In particular, significant activity was associated with the front gate (Figure 14-14). 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 14-15) in relative proximity 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

Figure 14-13 *Map depicting "clouds" of risk associated with various locations within the compound.*

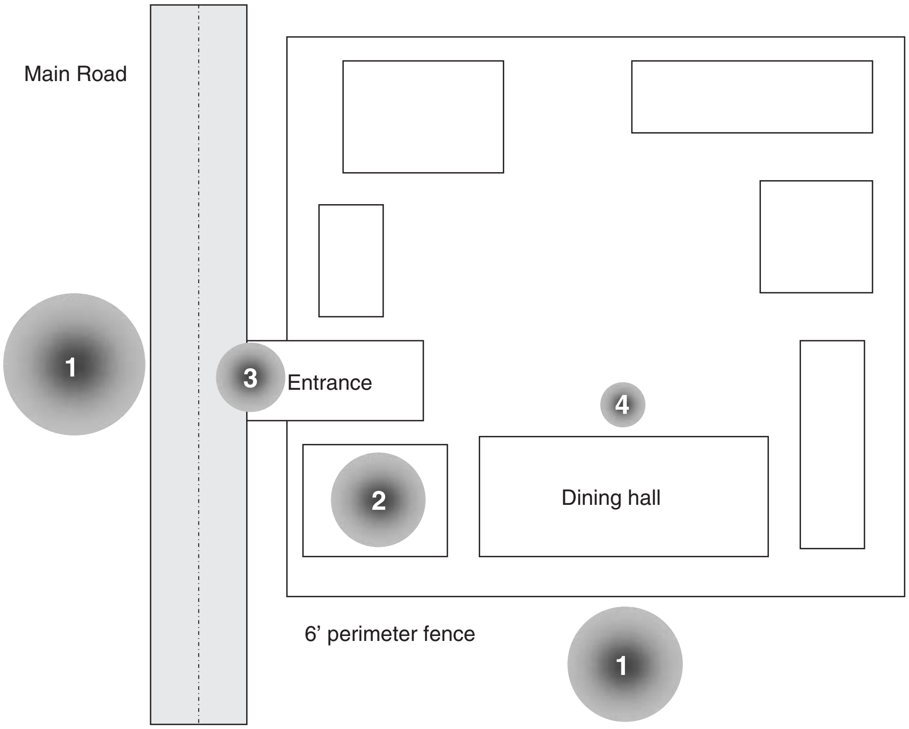
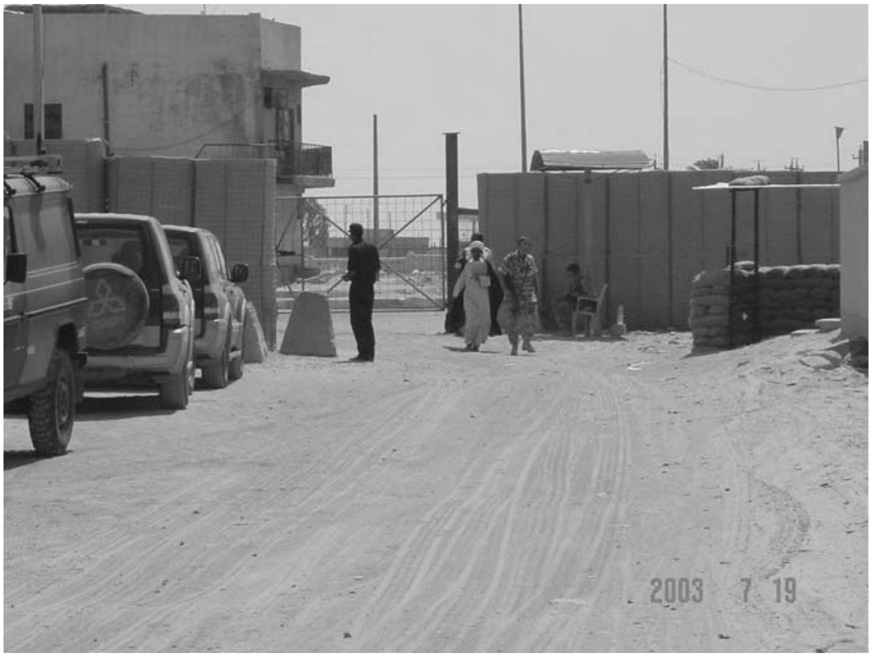


Figure 14-14 *Front gate of compound. (Staff Sergeant Tom Ferguson, USMC; used with permission.)*



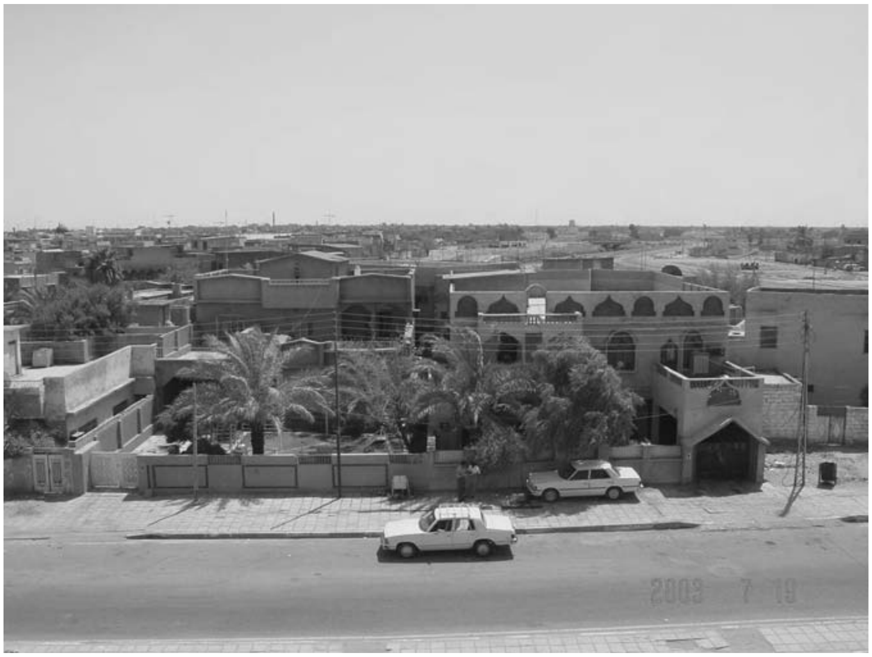


Figure 14-15 *Area outside the fence near the dining hall. (Staff Sergeant Tom Ferguson, USMC; used with permission.)*

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.

Internet Surveillance: To Delete Information or Not

This is a relatively tough question that the analyst generally does not participate in, but there are benefits both ways. If the information truly has the potential to either create or contribute to serious threats to public safety, then prudence might dictate removal of the information. Unfortunately, electronic information that is deleted might still be available through Internet archives and other related sites. On the other hand, leaving the information intact, perhaps with a few modifications, offers unique opportunities for surveillance detection and strategic misdirection.

Of course, now that this subject has been reported in the popular press, the false sense of security that these Internet interlopers might have had, thinking that they were lost in the vastness of municipal weblog data, has been lost. This probably means that they will choose another way to assess our capacity and response systems, which highlights an important point in surveillance detection. Once the watchers realize that they have been discovered, the activity generally stops. Therefore, any surveillance detection, even at the analytical level, should be conducted surreptitiously in an effort not to alert the watchers that they also are being watched.

In their discussion of fourth-generation warfare, Lind et al.[7] 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 orthophotography images of sensitive locations. Similarly, while certain military facilities have been blocked (e.g., Navy facilities in the Tidewater area of Virginia), the surrounding localities deploy detailed information related to military facilities. As early as 2001, the Israelis reported that their adversaries were exploiting the increased availability of orthophotography images freely available over the Internet.[8]

Even information included in contractor solicitations can have value if it outlines direction, internal capacity, abilities, or vulnerabilities. Most local, state, and federal agencies have requirements for a competitive bid process associated with any major purchase or contract. Frequently included in these solicitations is sufficient information regarding the desired product or service specifications as well as anticipated deployment. This is meant to ensure that a potential contractor can generate a bid that is both responsive and competitive. One of the easiest ways to disseminate this information is through the Internet. Unfortunately, detailed bid solicitations for systems, equipment, and services, while essential for a fair and effective bid process, also sends a strong message regarding the current and future capacity of the organization. Moreover, requests for proposals such as these provide unique opportunities for cover, as individuals might request access to secure locations or request additional documents and specifications under the guise of attempting to prepare a competitive bid, something of great potential value to a surveillance operation.

A recent Rand report concluded that the U.S. government generally does not deploy enough information with operational value over the Internet to aid in terrorist planning.[9] The report cautioned, however, that some nongovernmental agencies might. Therefore, the next question should be: Who else is releasing information about my organization or locality, and will they let me analyze the traffic, as well as future behavior? Keep in mind that the information pertaining to your community or location of interest probably is deployed over a variety of websites, few if any of which you have any control over or access to for security or analysis purposes. For example, at the time of this writing, detailed orthophotography images of Washington, D.C. were available through two academic institutions that were located outside of the District. Similarly, shared public safety and response information related to mutual aid agreements might be available through multiple websites. Vendors, consultants, and contractors might deploy information related to customers and projects. Tourism and public interest sites provide relatively detailed information, including photographs of locations of interest. At the time of this writing, the train bombings in Madrid were very recent. Despite the heightened level of alert in the travel industry as a whole after this brutal attack, it was still possible to locate a simulated "webcam" in a major transportation portal in this country. Although the images were not live, it was possible to scan the "camera" and view the entire area in tremendous detail, again from the anonymity of the Internet.

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 in relative anonymity. There is a tremendous degree of anonymity associated with movement throughout the Internet. 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 use the exploitation of readily available 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 superhighway. The amount of information contained in weblogs alone can be staggering, and it is increasing continuously.

Articles and tutorials outlining military thinking, tactics, and strategy are also available over the Internet. Abu 'Ubeid Al-Qurashi, one of Osama Bin Laden's aides, has made specific mention of the concept of fourth-generation warfare when outlining the al Qaeda combat doctrine.[10] Other reports have noted activity on the C4I.org site that was associated with Internet addresses linked to Iraq.[11] A review of the activity suggested that most of their interest appeared to be related to psychological tactics, information warfare, and other military issues.

14.5 Internet Surveillance Detection

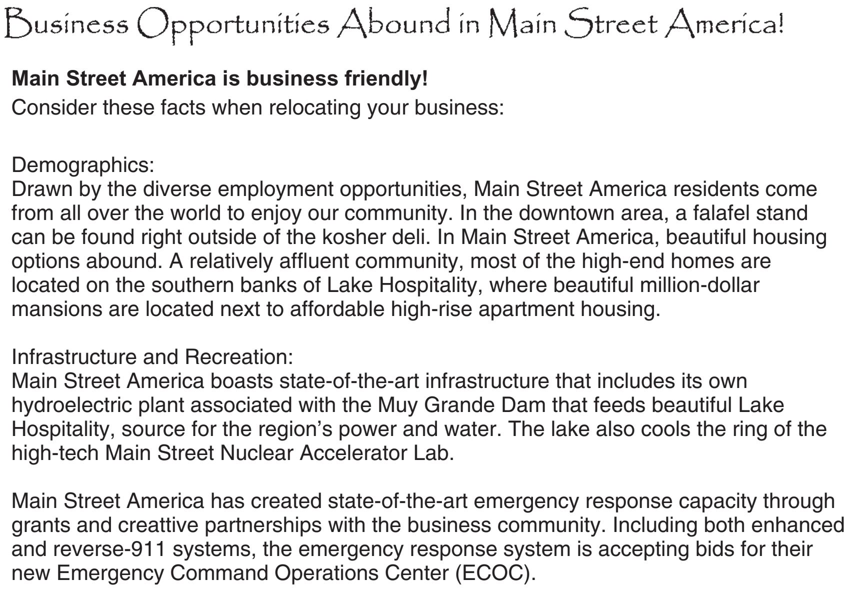
What does this have to do with data mining? Approaches and tools similar to those employed by online retailers to characterize online shopping behavior and create models of potential buyers, so-called web mining tools, can be exploited by analysts in an effort to identify and monitor possible Internet surveillance and protect critical infrastructure. Using these same web mining tools, Internet activity can be analyzed in a timely, comprehensive fashion. Integration of the data across multiple sites or reference points can provide additional value through the revelation of complex patterns of activity or behavior. Moreover, compilation and analysis of data from multiple sites provides the added benefit of comparison between sites. These data then can then be merged with additional surveillance detection information and used to support the integrated surveillance detection process outlined above.

Many localities have reported suspicious or unusual activity on their websites. In particular, IP addresses associated with locations in the Middle East have been noted searching pages related to local infrastructure and public safety.[12] In one particular community, activity associated with the suspicious IP addresses frequently came in and out of the site through one specific page. While this particular web page was not noteworthy or unusual in and of itself, analysis of "normal" traffic patterns revealed that less than one-half of one percent of all page hits on this website were associated with this particular page. In other words, visits to this web page associated with "normal" or routine traffic were extremely rare. Yet it was unusual for a session associated with the suspicious IP addresses to not include at least one hit on this page.

Using this method of brute-force anomaly detection, this particular web page could be used as a screening tool or "magnet" for suspicious or unusual activity. By comparing screened traffic against known patterns of activity previously associated with suspicious IP addresses, additional IP addresses were identified. Ultimately, it was possible to create some simple filters to identify and capture potentially suspicious behavior on this website by screening for known IP addresses, as well as patterns of activity or particular page hits previously associated with suspicious addresses and behavior. Therefore, while many agencies and organizations have focused on intrusion detection, another potential vulnerability includes surveillance or misuse of information available through the Internet and other open-source venues. The ability to identify, characterize, and monitor unusual or suspicious Internet activity can provide additional insight regarding our adversaries' interests and possible intentions, thereby increasing our battlespace awareness. Again, it is possible to gather a tremendous amount of information regarding potential surveillance activity on websites of interest by simply using a good understanding of "normal" behavior and anomaly detection.

Internet Surveillance Example

A fictitious local law enforcement agency identified regular, consistent activity on its website that was associated with an IP address linked to a particular terrorist group. Through data mining, the agency was able to identify one particular page that was visited frequently by the group of interest. By characterizing "normal," the crime analysts were able to determine that this page was associated with limited activity beyond the terrorist group. In fact, less than 1% of all of the page hits on the entire website were associated with this page, and almost all of the visits associated with IP addresses not linked to the terrorist group came from one single referring site, the local chamber of commerce.



Business Opportunities Abound in Main Street America!

Main Street America is business friendly!
Consider these facts when relocating your business:

Demographics:
Drawn by the diverse employment opportunities, Main Street America residents come from all over the world to enjoy our community. In the downtown area, a falafel stand can be found right outside of the kosher deli. In Main Street America, beautiful housing options abound. A relatively affluent community, most of the high-end homes are located on the southern banks of Lake Hospitality, where beautiful million-dollar mansions are located next to affordable high-rise apartment housing.

Infrastructure and Recreation:
Main Street America boasts state-of-the-art infrastructure that includes its own hydroelectric plant associated with the Muy Grande Dam that feeds beautiful Lake Hospitality, source for the region's power and water. The lake also cools the ring of the high-tech Main Street Nuclear Accelerator Lab.

Main Street America has created state-of-the-art emergency response capacity through grants and creattive partnerships with the business community. Including both enhanced and reverse-911 systems, the emergency response system is accepting bids for their new Emergency Command Operations Center (ECOC).

Figure 14-16 *"Business Opportunities" web page associated with suspicious activity.*

Initial review of the page did not offer any clues as to why this group might be so interested in it (Figure 14-16). The page was constructed in an effort to provide additional information for businesses considering relocating to the city, which we will call Main Street America. General information, including regional demographics, existing infrastructure resources, and emergency response capacity were provided. Further examination, however, provided some indication as to why the terrorist group might be interested in it.

The page in question made it clear that damage to the dam would cause loss of power and water to the region, as well as possible flooding to the homes surrounding the lake formed by the dam. The page also indicates that housing on the shores of the lake includes apartments. This high-density housing would increase the potential casualty estimates in the event of a disaster at the dam. The demographic information also provides clues as to who might appear unusual or out of the ordinary and what would be necessary to blend into that community. Information pertaining to the local population provides insight regarding local possibilities for recruitment. The existence of a nuclear laboratory in the region has value in terms of potential resources, including highly trained personnel, equipment, and raw materials, as well as potential target opportunities. In sum, the information provided innocently for potential recruiting, which did not seem critical when viewed simply as a page on a website, has significant potential tactical and strategic value to individuals with less than positive intentions.

This example combined data mining with some "brute force" techniques. Characterizing activity on a website is only the first step in the process. Again, domain expertise is essential to the accurate and meaningful review of weblog data. In most cases, it is extremely important to review the actual content of the pages visited, particularly those that are associated with unusual or repeated activity. In some cases, "brute force" manual review of the pages for key words, concepts, or other valuable information may be necessary. Text mining techniques, however, promise to automate this process further. Additional advantages of text mining include not only speed of analysis but also completeness and accuracy. The brain can only take in and hold a limited number of concepts. Expert systems, on the other hand, have almost unlimited capacity to identify and consider multiple concepts simultaneously without error, bias, or fatigue.

Going back to the "business opportunities" page, we noted that less than 1% of the hits on that particular website were to that page. Many of them involved the IP address that had been associated previously with a suspected terrorist group, and some others were associated with referrals from the local chamber of commerce website. What about the remaining IP addresses associated with hits to this page? In many ways, looking for suspicious activity on a website is like looking for a needle in a haystack. These generally are very infrequent events that can be extremely difficult to identify. In some cases, however, we can find a "magnet" that makes finding the needle a little easier. One obvious magnet is a list of IP addresses associated with known or suspected terrorist groups or locations. Another magnet is IP addresses identified by their involvement in unusual or suspicious activity. In the Main Street America example, the "business opportunities" page represents a possible magnet that could be used to identify other IP addresses worthy of additional review and analysis.

After initial data mining and characterization, browsing patterns can be characterized and modeled using sequence analysis techniques similar to those employed by online retailers for the analysis of consumer behavior. Special attention needs to be paid to prior probabilities, however, because the behavior tends to be extremely infrequent when viewed in light of the total volume of activity of a particular website. Failure to account for this could mask important details and seriously limit the ability to identify any unique trends or patterns.

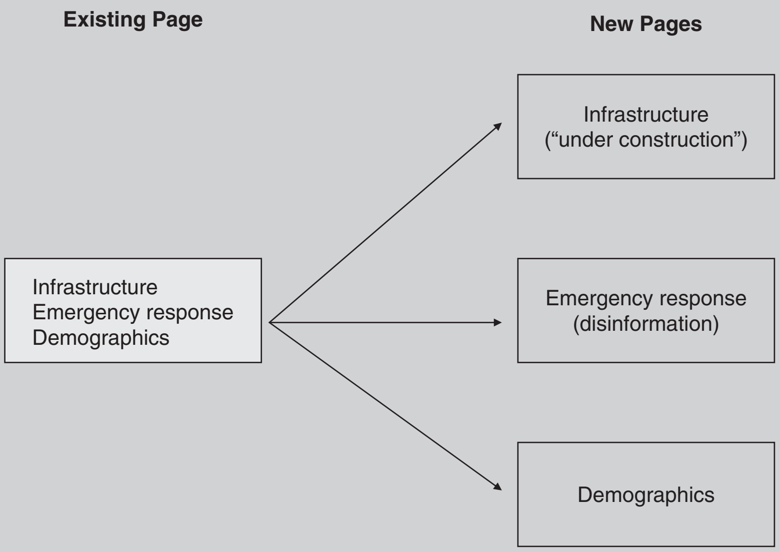
Internet Honeypots

What happens once suspicious activity has been identified and characterized? One solution would be to remove the information from the website; however, this could be comparable to closing the proverbial barn door after all of the horses are gone. In addition, Internet archives exist that can hold information long after it has been removed from active web pages. An alternate solution would be to establish electronic surveillance detection. In many ways, electronic surveillance detection and countersurveillance has value that might extend beyond blocking access to the information.

Once suspicious activity has been detected, there are additional opportunities to evaluate the interest and intentions of the bad guys. For example, a particular pattern of activity might suggest interest in some areas. Deploying additional information provides an opportunity to engage in additional "hypothesis testing" regarding their true intentions. For this approach to work, actual information does not need to be deployed. The pages of interest might deploy generic information or be "under construction" indefinitely. In fact, the opportunities for disinformation abound with this arrangement. Through a specific series of new information and links, additional informational specificity can be identified and determined.

In terms of Main Street America, the analysts noted three potential areas of interest or concern associated with the web page. Figure 14-17 depicts a possible scenario that could help identify the specific area of interest associated with the suspicious activity.

Figure 14-17 *Conceptual illustration of possible Internet honeypot for further characterizing suspicious activity.*



14.6 How To

The first step in this process is to review and assess what information is available and what value it 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 over the Internet, how it is organized, and what it might mean from either a tactical or strategic perspective is critical to understanding the information held in the weblogs. It also is important to consider how this information might be combined with other information on the same website or with other resources to add value. 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.

The second reason that a review or threat assessment of the information deployed over the website is important is because if you are reading this, then you probably have some responsibility for or involvement in public safety. Knowing what information has been deployed publicly and considering how it might be used against you, your agency, or locality can help guide possible surveillance detection efforts and response planning.

The analysis from this point on is not trivial. The amount of information contained in weblogs can be staggering, even for an analyst who has some experience wrestling with large complaint databases or telephone records. In addition, the information itself is ugly and can be very difficult to manipulate without the correct tools. Those choosing to tackle Internet data most certainly will want to bring power tools, rather than a hammer and chisel, to this battle.

Some very detailed materials and excellent software tools are available for the analysis of Internet data, but some tasks can be completed with standard data mining tools. Many public safety agencies are at a disadvantage in mining Internet data, however, because they do not set cookies. Briefly, session cookies are temporary and can be used to track movement through a website during a single session. Persistent cookies, on the other hand, not only track movement through your site, but follow the user out into the Internet, leaving a trail similar to electronic breadcrumbs. Without cookies, it is very difficult to link related page hits into a single session. This issue received some attention a few years ago after it was reported that the Central Intelligence Agency used cookies on its website.[13] After they were criticized for this practice, they, as well as many other public safety agencies, made the decision to discontinue this practice. While this might limit analysis somewhat, it is still possible to presumptively link some activity based on IP address and other unique characteristics, particularly with some of the more sophisticated web mining tools. Additional "brute force" techniques can be used to characterize and analyze some of the activity on a website, particularly if the IP addresses of interest are not prevalent.

The first step generally involves an overview and characterization of routine activity on the website. As always, it is helpful to develop an understanding of "normal" behavior in an effort to create a baseline for future comparison. For example, where do people go, how to they normally enter the site, are there common referring pages, which pages are popular, and which are not? While we would like to think that every detail that is deployed over our website has value and interest, most of it is relatively boring. Thankfully, most of the information that has tactical or strategic value for those with bad intentions generally falls into the boring category, so almost any activity associated with those pages is worthy of note. For example, pages deploying critical infrastructure information generally do not appear on the list of top ten websites, but they have been some of the more popular ones associated with unusual or suspicious activity.

A second activity might include characterizing activity based on the user IP address. There will be some relatively common addresses; particularly those associated with local service providers, but some relatively infrequent addresses might be identified as well. These can be checked through reverse lookup mechanisms, similar to those used for telephone numbers. If suspicious activity or patterns are identified, then modeling algorithms can be created to further characterize the activity and generate rule sets that can be used to screen future behavior. Similarly, specific IP addresses or web pages associated with unusual or suspicious activity can be identified, and any future activity associated with those addresses or pages flagged for further investigation.

Unsupervised learning or clustering techniques also may be used to characterize "normal" patterns of Internet activity in an effort to identify possible surveillance activity. This method of anomaly detection frequently can reveal patterns of unusual behavior or potential misuse of open-source information. 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.

14.7 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 or those from other cultures also might reveal themselves when they fail to blend into the surrounding environment.

Again, identifying suspicious or unusual activity can be compared to finding a needle in the haystack. This is where anomaly detection, which is an 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 above, 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, though, 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.

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.

14.8 Bibliography

1. Carter, M. (2004). Why feds believe terrorists are probing ferry system. *The Seattle Times*, October 10. http://seattletimes.nwsource.com/cgi-bin/PrintStory.pl?document_id=2002058959&zsection_id=2001780260&slug=ferry10m&date=20041010
2. The National Commission on Terrorist Attacks Upon the United States, T.H. Kean, Chair (2004). The 9/11 Commission Report. www.9-11commission.gov/.
3. Joint DHS and FBI Advisory. (2004). Homeland security system increased to orange for financial institutions in specific geographic areas. August 1. www.dhs.gov/interweb/assetlibrary/IAIP_AdvisoryOrangeFinancialInst_080104.pdf
4. McCue, C., Parker, A., McNulty, P.J., and McCoy, D. (2004). Doing more with less: Data mining in police deployment decisions. *Violent Crime Newsletter*, Spring, 1, 4–5; McCue, C. and McNulty, P.J. (2003). Gazing into the crystal ball: Data mining and risk-based deployment. *Violent Crime Newsletter*, September, 1–2; McCue, C. and McNulty, P.J. (2004). Guns, drugs and violence: Breaking the nexus with data mining. *Law and Order*, **51**, 34–36.
5. De Becker, G. (1997). The gift of fear. Little, Brown and Company, New York.
6. McCue, C. and McNulty, P.J. (2003). Gazing into the crystal ball: Data mining and risk-based deployment. *Violent Crime Newsletter*, September, 1–2.
7. 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.
8. Shapira, R. (2001). We are on the Palestinians' map. *Maariv (Tel Aviv)*, May 18.

9. Baker, J.C., Lachman, B., Frelinger, D., O'Connell, K.M., Hou, A.C., Tseng, M.S., Orletsky, D.T., and Yost, C.W. (2004). Mapping the Risks: Assessing the Homeland Security Implications of Publicly Available Geospatial Information, Rand Corporation.

10. 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/listinfo/papyrus-news

11. McWilliams, B. (2003). Iraq's crash course in cyberwar. *Wired News*, May 22.

12. Gellman, B. "Cyber-attacks by Al Qaeda feared." *Washington Post*, June 27; http://www.washingtonpost.com/wp-dyn/articles/A50765-202June26.html

13. CIA caught sneaking cookies. (2002). CBSNews.com, March 20.

Advanced Concepts and Future Trends

This Page Intentionally Left Blank

15

Advanced Topics

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 Intrusion Detection

People tend to move through the world in predictable patterns. Generally when they show up where they should not be, they are either lost or have some other purpose for being there. Intrusion detection is a rapidly growing industry in the information technology world. Within the field of intrusion detection, unacceptable behavior has been further subdivided into intrusion detection, which generally looks for outside intruders, and misuse detection, which evaluates threats from within.[1] It is well beyond the scope of this text to delve into all of the intricacies of intrusion detection systems, which often change rapidly in response to industry requirements and novel ways to access a system without authorization. Briefly, intrusion detection can be used to identify unusual activity or access patterns that might indicate someone has breached or has attempted to breach a system. These can be set to identify common patterns of attempted access or unusual behavior that warrants additional review.

Similarly, misuse detection looks for unusual patterns of behavior or attempted access within a system. For example, someone from the mailroom trying to access executive salaries or a disgruntled employee logging on late at night from a remote location in an effort to secure some intellectual parting gifts prior to resignation would be cause for concern within an organization. These tools work not only on intranet servers but also on automated access control systems. Again, most individuals will move though the physical structure with predictable, readily identifiable patterns associated with their function and routine activities. Deviation from these normal patterns can indicate almost anything from industrial espionage to a budding romance between someone in maintenance and a member of the secretarial pool. Either way, unusual movement throughout or activity within a facility generally warrants additional analysis.

Another pattern of unusual activity that can be detected within a system is vanity checks. While not presumptive evidence of criminal behavior or intent, checking one's internal records and who might have accessed them is certainly unusual behavior that often indicates that one has something to hide. Robert Hansen, the FBI agent involved in espionage, was found to have engaged in periodic vanity checks within the FBI's computer systems.

In some situations, people might have a legitimate reason for browsing a particular system; however, their movement through the system highlights their behavior as unusual. Recently, it was reported that individuals from the Middle East had been accessing municipal websites, apparently in search of public safety infrastructure information. In this case, the origin of the activity as well as the activity itself was unusual and worthy of further investigation. This particular subject was addressed in greater detail in Chapter 14.

15.2 Identify Theft

Identity theft is not a new problem. In fact, my own grandfather, to whom this book is dedicated, 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 highjackers 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.

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 identify 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.

15.3 Syndromic Surveillance

Syndromic surveillance systems have been developed for the detection of disease outbreak and bioterrorism. By using anomaly detection, these automated systems are able to identify unusual clusters of symptoms or unanticipated changes in normal disease rates. 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.[2] By enhancing standard anomaly detection with decision rules, the performance of these screening algorithms can be increased.[3]

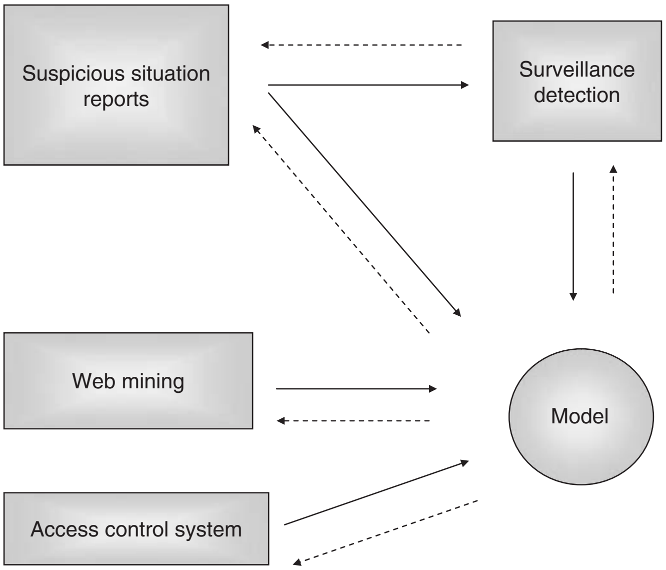
15.4 Data Collection, Fusion and Preprocessing

The topic of surveillance detection is addressed in Chapter 14, but it is worth revisiting it 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, though, 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. Similarly, concurrent surveillance of multiple, geographically isolated targets also can be detected through the use of data mining algorithms on combined data resources (Figure 15-1).

An overview of anomaly detection was provided in Chapter 7, which included a "perfect world" scenario that employed a two-pronged monitoring and detection approach. The monitoring component utilized predetermined decision rules or scoring algorithms based on previously identified, characterized, and modeled bad behavior; the detection part of the process incorporated the use of anomaly detection in an effort to identify behavior that had not been previously observed or anticipated. In many ways, surveillance detection offers a great opportunity for deploying this type of analytical approach. Suspicious situation reports often reflect only a small fraction of the ongoing surveillance activity. Moreover, as the example above highlights, a certain degree

Figure 15-1 *Data mining tools can be used to correlate physical surveillance with Internet surveillance and suspect interviews.*



of preliminary analysis and classification can significantly enhance the analytical products and overall understanding. These initial analyses can represent the first component of a multilevel analysis of unusual or suspicious activity, while concomitantly guiding additional collection efforts.

A possible integrated surveillance analytical model is depicted in Figure 15-2. This model incorporates the use of predetermined scoring algorithms that are based on a known behavior of interest, with anomaly detection running in the background, as described previously, in an effort to identify any unusual or unanticipated activity. Information inputs would include suspicious situation reports as well as other complementary information, including access control systems, sensors, the agency intranet, and even their website. Ongoing analysis of suspicious situation reports would guide additional surveillance detection and information collection efforts, which also would be included in the analysis. In this model, the analytical outputs would be used to further guide and refine additional collection efforts, highlighting the potential intelligence enhancements that can occur when data and analysis are viewed as dynamic components of the analytical–operational continuum, rather than static commodities in the collection process.

Again, 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 may have far greater value when characterizing movement and identifying a potential risk or threat.

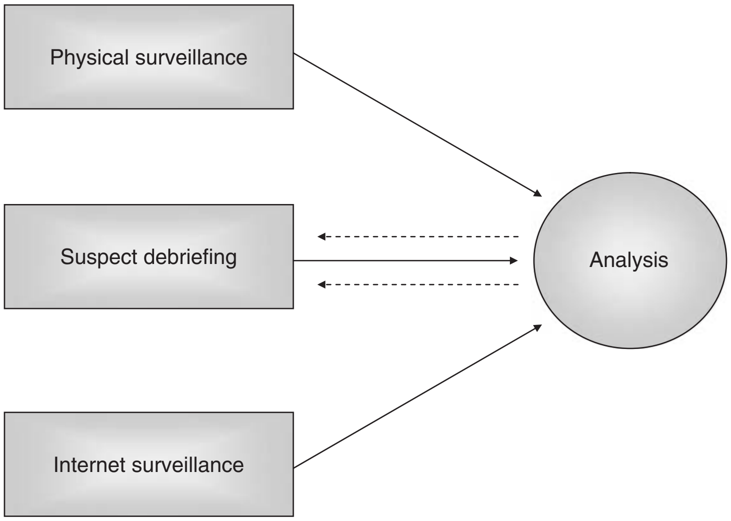


Figure 15-2 *Integrated surveillance analysis model, which incorporated dynamic feedback within the model to further enhance surveillance detection efforts.*

It is also worth considering the fact that the space in and around a facility is not homogenous. For example, unusual or suspicious behavior around a facility frequently is informally weighted. Different types of activity are given greater or lesser value. The value given to 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 subtlely, 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.

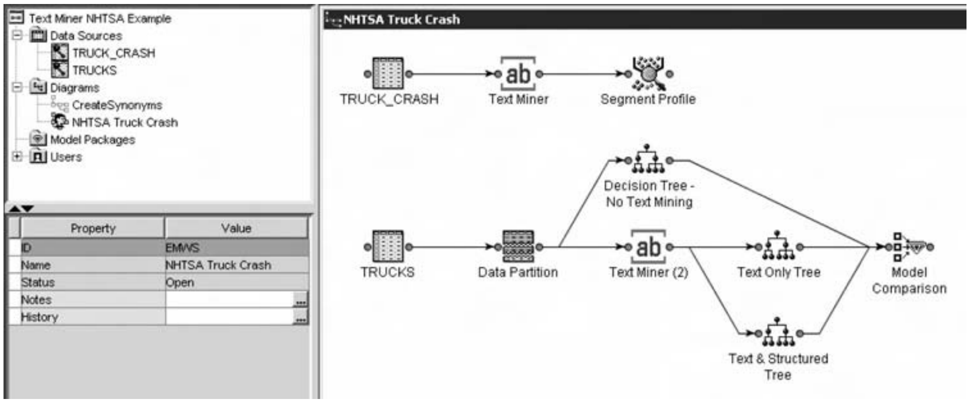
On the other hand, the exact physical location of a possible incident of surveillance generally is not as important as where the person was and what he or she was looking at. Extending this, five different individuals could occupy five slightly different locations, but if they were all observing the 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.

15.5 Text Mining

Text mining tools are increasing in capacity and in value to the analyst on an almost daily basis, particularly as the software developers continue to expand the capacity of these tools.

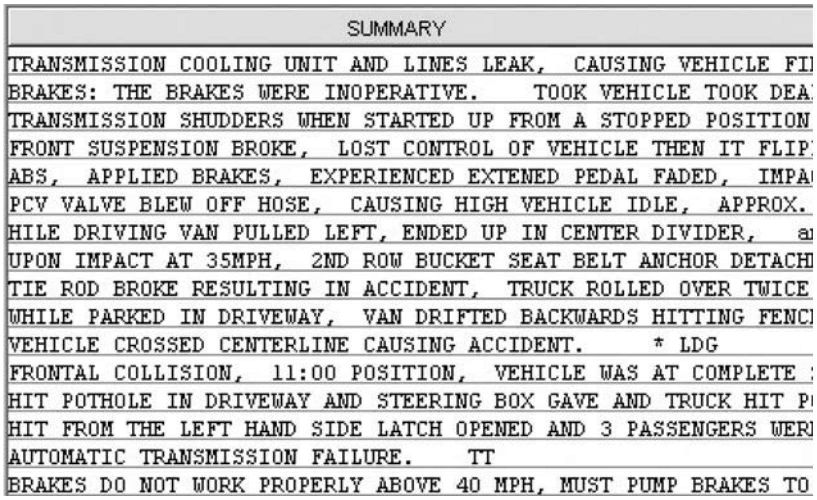
Figure 15-3 illustrates the inclusion of a text mining process in an analysis. In this particular example, related narrative information has been included in an analysis of truck crashes. By using text mining, key concepts can be extracted from this unstructured information and used to enhance the modeling process.

Figure 15-3 *Example of a text mining tool and its use in the analysis of truck crash data. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



Property	Value
ID	EMWS
Name	NHTSA Truck Crash
Status	Open
Notes	
History	

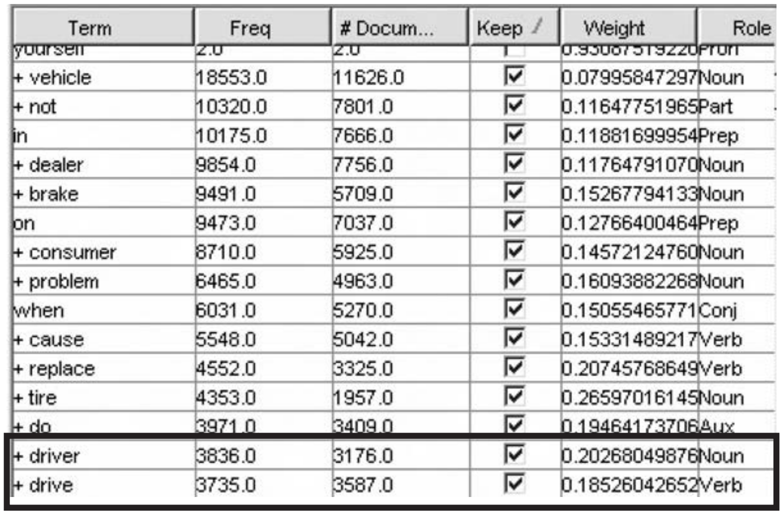
Figure 15-4 *Sample of the narrative portion used in the text mining example. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



SUMMARY
TRANSMISSION COOLING UNIT AND LINES LEAK, CAUSING VEHICLE FI
BRAKES: THE BRAKES WERE INOPERATIVE. TOOK VEHICLE TOOK DEA
TRANSMISSION SHUDDERS WHEN STARTED UP FROM A STOPPED POSITION
FRONT SUSPENSION BROKE, LOST CONTROL OF VEHICLE THEN IT FLIP
ABS, APPLIED BRAKES, EXPERIENCED EXTENED PEDAL FADED, IMPA
PCV VALVE BLEW OFF HOSE, CAUSING HIGH VEHICLE IDLE, APPROX.
HILE DRIVING VAN PULLED LEFT, ENDED UP IN CENTER DIVIDER, a
UPON IMPACT AT 35MPH, 2ND ROW BUCKET SEAT BELT ANCHOR DETACH
TIE ROD BROKE RESULTING IN ACCIDENT, TRUCK ROLLED OVER TWICE
WHILE PARKED IN DRIVEWAY, VAN DRIFTED BACKWARDS HITTING FENC
VEHICLE CROSSED CENTERLINE CAUSING ACCIDENT. * LDG
FRONTAL COLLISION, 11:00 POSITION, VEHICLE WAS AT COMPLETE
HIT POTHOLE IN DRIVEWAY AND STEERING BOX GAVE AND TRUCK HIT P
HIT FROM THE LEFT HAND SIDE LATCH OPENED AND 3 PASSENGERS WER
AUTOMATIC TRANSMISSION FAILURE. TT
BRAKES DO NOT WORK PROPERLY ABOVE 40 MPH, MUST PUMP BRAKES TO

Figure 15-4 illustrates the type of narrative information that was used in the text mining process. Figure 15-5 demonstrates the key concepts extracted, as well as the frequency with which they occur. These frequency estimates can help the analyst identify concepts and terms that will be used in the analysis by identifying very frequent or extremely rare terms and concepts. The parts of speech have been identified and labeled here. As can be seen in the highlighted

Figure 15-5 *Example of the concepts extracted from the narrative. The highlighted section illustrates the ability to identify and differentiate different parts of speech. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



Term	Freq	# Docum...	Keep	Weight	Role
yourself	2.0	2.0	☐	0.93087519220	Pron
+ vehicle	18553.0	11626.0	☑	0.07995847297	Noun
+ not	10320.0	7801.0	☑	0.11647751965	Part
in	10175.0	7666.0	☑	0.11881699954	Prep
+ dealer	9854.0	7756.0	☑	0.11764791070	Noun
+ brake	9491.0	5709.0	☑	0.15267794133	Noun
on	9473.0	7037.0	☑	0.12766400464	Prep
+ consumer	8710.0	5925.0	☑	0.14572124760	Noun
+ problem	6465.0	4963.0	☑	0.16093882268	Noun
when	6031.0	5270.0	☑	0.15055465771	Conj
+ cause	5548.0	5042.0	☑	0.15331489217	Verb
+ replace	4552.0	3325.0	☑	0.20745768649	Verb
+ tire	4353.0	1957.0	☑	0.26597016145	Noun
+ do	3971.0	3409.0	☑	0.19464173706	Aux
+ driver	3836.0	3176.0	☑	0.20268049876	Noun
+ drive	3735.0	3587.0	☑	0.18526042652	Verb

portion, "driver" and "drive" have been correctly identified as noun and verb, further highlighting the capacity of this tool.

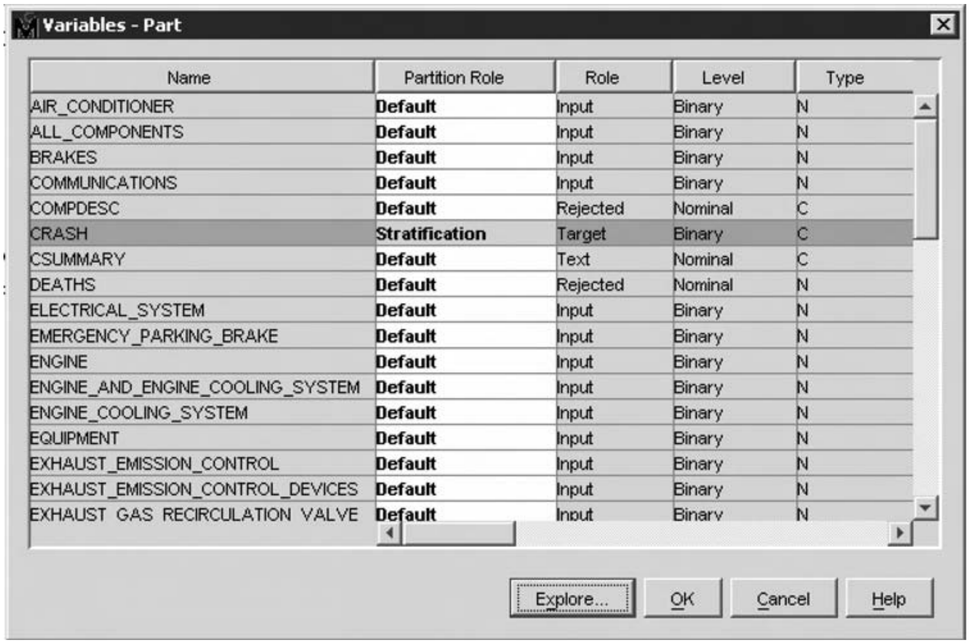
The results of the text mining process then can be used in the creation of the model. As illustrated in Figure 15-6, "crash" has been selected as the target or outcome variable. Those variables selected as inputs are available for inclusion in the created model. The ability to use the terms and concepts identified during the text mining process allows the analyst to tap into the vast potential associated with unstructured narrative data and increase the fidelity of the models created.

Finally, as can be seen in Figure 15-7, additional language options further increase the capacity of these tools to incorporate narrative data captured in languages other than English.

15.6 Fraud Detection

This topic is so large that an entire textbook could be devoted to it; however, it is important to note that data mining and predictive analytics represent the most effective approaches to addressing this pattern of illegal behavior.

Figure 15-6 *This dialog box demonstrates the use of concepts identified using text mining as variables in a modeling algorithm. The highlighted record illustrates the selection of the binary variable "crash" as the target or outcome variable selected. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*

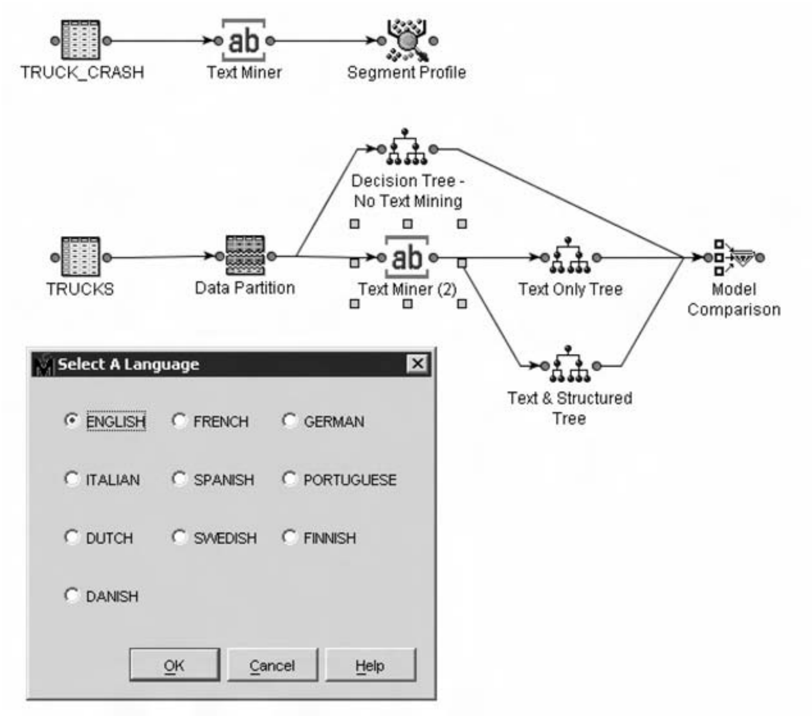


Name	Partition Role	Role	Level	Type
AIR_CONDITIONER	Default	Input	Binary	N
ALL_COMPONENTS	Default	Input	Binary	N
BRAKES	Default	Input	Binary	N
COMMUNICATIONS	Default	Input	Binary	N
COMPDESC	Default	Rejected	Nominal	C
CRASH	Stratification	Target	Binary	C
CSUMMARY	Default	Text	Nominal	C
DEATHS	Default	Rejected	Nominal	N
ELECTRICAL_SYSTEM	Default	Input	Binary	N
EMERGENCY_PARKING_BRAKE	Default	Input	Binary	N
ENGINE	Default	Input	Binary	N
ENGINE_AND_ENGINE_COOLING_SYSTEM	Default	Input	Binary	N
ENGINE_COOLING_SYSTEM	Default	Input	Binary	N
EQUIPMENT	Default	Input	Binary	N
EXHAUST_EMISSION_CONTROL	Default	Input	Binary	N
EXHAUST_EMISSION_CONTROL_DEVICES	Default	Input	Binary	N
EXHAUST GAS RECIRCULATION VALVE	Default	Input	Binary	N

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



◉ ENGLISH ○ FRENCH ○ GERMAN

○ ITALIAN ○ SPANISH ○ PORTUGUESE

○ DUTCH ○ SWEDISH ○ FINNISH

○ DANISH

Figure 15-7 *This dialog box illustrates the enhanced language capacity increasingly available with text mining tools. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*

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.

15.7 Consensus Opinions

Although the 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's theorem, naval scientist John Craven used consensus expert opinions to locate the U.S. nuclear submarine *Scorpion*.[4] 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.

15.8 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.

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.

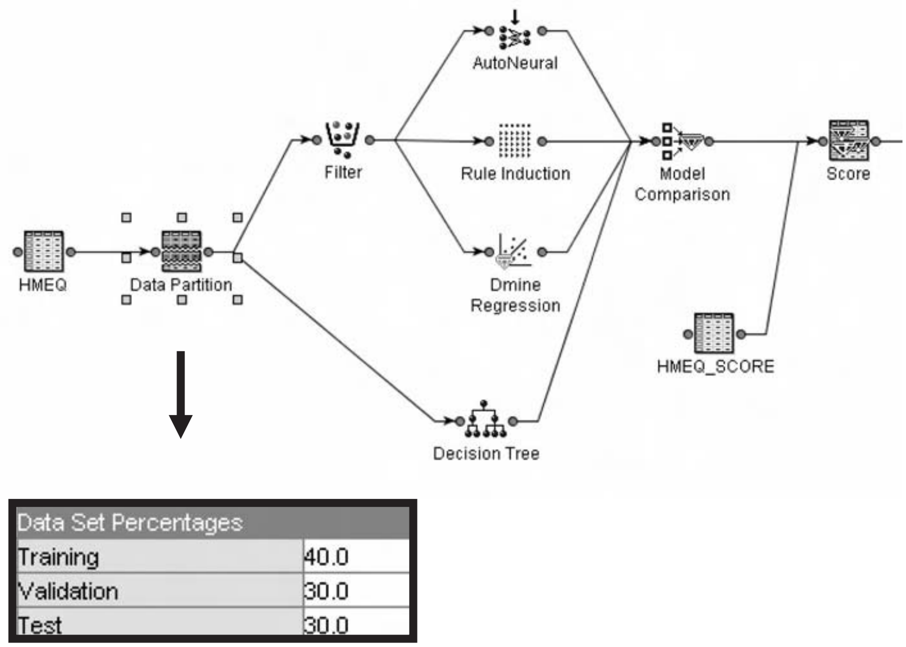
Data Partitioning

The important 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. A three-sample approach to data partitioning is illustrated in Figure 15-8. The three samples include training, validation and test. 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 overrepresenting unusual or spurious findings, which is a limitation with boosting methods.

The topics listed above represent a small sampling of the additional work that has been done using data mining and predictive analytics in the applied

Figure 15-8 *Example of the inclusion of data partitioning in the analytical process. The dialog box illustrates a three-sample method for partitioning data, which includes training, validation, and test samples. (Created with SAS software. Copyright 2006, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC.)*



Data Set Percentages	
Training	40.0
Validation	30.0
Test	30.0

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.

15.9 Bibliography

1. Phung, M. (2000). Data mining in intrusion detection. SANS Institute Resources (10/24/2000).
2. Reis, B.Y. and Mandl, K.D. (2003). Integrating syndromic surveillance data across multiple locations: Effects on outbreak detection performance. Proc AMIA Symp, 549–553.

3. Wong, W.K., Moore, A., Cooper, G., and Wagner, M. (2002). Rule-based anomaly pattern detection for detecting disease outbreaks. American Association for Artificial Intelligence.
4. Sontag, S., Drew, C., and Drew, A. (1999). Blind man's bluff: The untold story of american submarine espionage. HarperCollins, New York.

This Page Intentionally Left Blank

16

Future Trends

"Computers in the future may weigh no more than 1.5 tons." Popular Mechanics *("Forecasting the Relentless March of Science," 1949)*

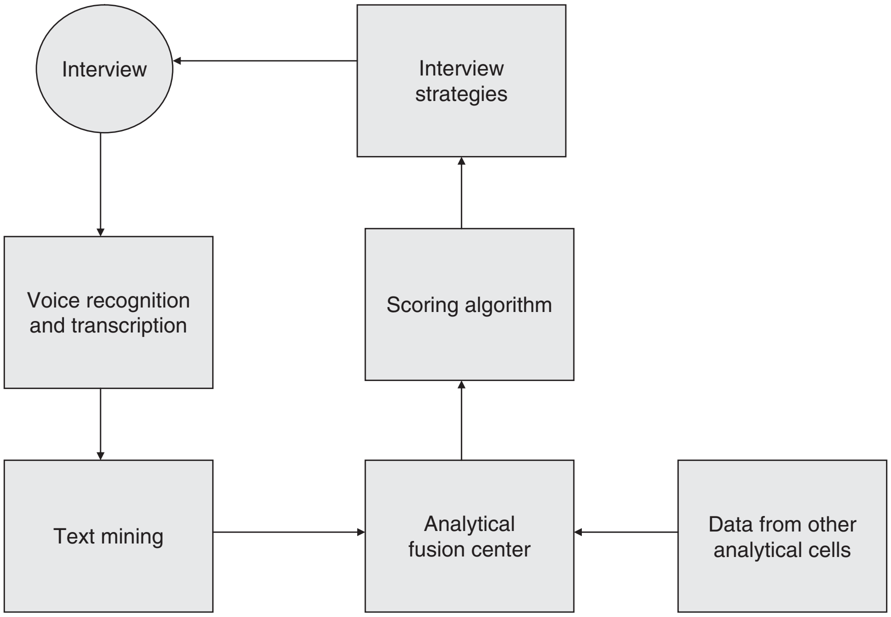
As the above quote illustrates, it is pretty easy to go far astray when trying to project the future of technology. In the cases of data mining and predictive analytics, however, the future is becoming reality at such a rapid pace that almost anything that I write will be outdated before the first copy 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.

16.1 Text Mining

Text mining holds considerable promise for applied public safety and security mining and analysis. The ability to tap directly into and use unstructured narrative data will be 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 experience, 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.

In the very near future, I envision the development of public safety and security-specific glossaries that will incorporate the unique lexicon and terms associated with law enforcement and intelligence analysis. Like the specialized glossaries developed for other professions that represent their own unique language and terms (e.g., medicine), these enhanced capabilities will allow analysts to build on the work of others, creating a critical mass of shared knowledge that will allow all of us to do our jobs better. Also, as outlined in the Introduction and illustrated in Figure 16-1, I envision the development of tools that will link voice recognition and translation software directly to these text mining tools and provide "just-in-time" analytical support to operationans when and where they need it, without the need for cumbersome and time-consuming data collection and entry processes. Integrated tools can be used to integrate and analyze data from a variety of geographically distinct collection locations, further enhancing the depth of knowledge afforded to the operational personnel serving on the front lines. Ultimately, tools like these not only promise to enhance field-based interviewing but also will provide a level of situational awareness not currently available. They will increase the efficacy of the operational personnel while keeping them safer.

Figure 16-1 *This model conceptually illustrates how information and analysis can inform and feed the operational process.*

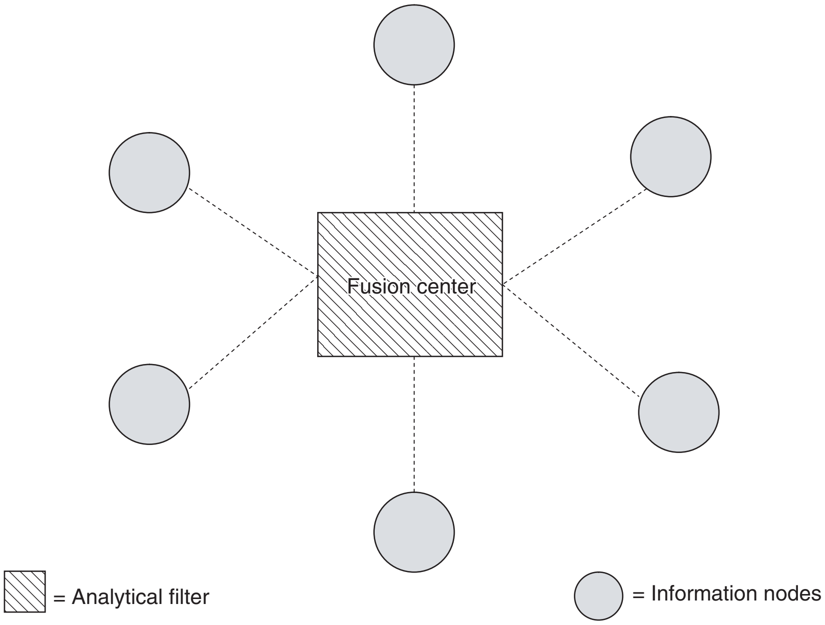


16.2 Fusion Centers

The concept of regional data fusion and the emergence of regional fusion centers supporting this task has been mentioned. While the idea of even more data to analyze secretly thrills most analysts, creation of these fusion centers represents only the first step in the process. Currently, many of these centers focus almost exclusively on the collection aspect of the process and support only limited analytical capacity. The analytical functions offered by these centers frequently include only the capacity to search the data repository and some ability to perform descriptive statistics. The next step in the process should involve moving away from counting crime and toward characterizing and modeling incidents in an effort to anticipate, predict, prevent, and deter.

As we link these regional fusion centers and include options for predictive analysis, we set the stage for the creation of analytical webs that become almost self-perpetuating in their identification of associations and relationships. As illustrated in Figure 16-2, an array of linked collection nodes that are supported by a common analytical fusion center starts to resemble a neural network model. Enhancing the connectivity by integrating predictive analytics would bring this model even closer to a true neural network. The knowledge discovery and predictive analysis that could be supported by a model like this

Figure 16-2 *Conceptual illustration of an information web that includes an analytical core or fusion center.*



example could be game changing in its ability to support information-based decisions in the applied setting.

16.3 "Functional" Interoperability

The concept of interoperability has received considerable attention recently, yet the true success of this approach lies in the ability to achieve "functional interoperability." Functional interoperability involves more than just information access and can be achieved only when the shared resources are used effectively to support joint decisions and complementary responses. By using data mining tools and emphasizing "operationally actionable" output that focuses on the time, location, and nature of potentially threatening behavior, the relevant public safety resources can be deployed efficiently in support of public safety and homeland security objectives.

16.4 "Virtual" Warehouses

Traditional data warehouses can be extremely costly and generally lack the speed and agility required by public safety and security analysis. Tools like Google and the FAST Enterprise Search Platform represent a new approach to information access and management. By indexing large data sets, including the World Wide Web, these tools give analysts the ability to create "virtual" warehouses that embody the speed and flexibility generally not available with traditional data warehouses. These indices can be updated and refreshed, noting changes in the information available as they occur. The accessibility and ease with which these tools can be used are reflected in their ubiquitous use and the fact the "Google" has gained status as a verb (e.g., "Why don't you Google this and see what comes up?" or "I Googled you last night and came up with a number of hits.").

These tools also give analysts an opportunity to tap into open-source material more effectively. I continue to be absolutely amazed at the amount of information available over the Internet. While chat rooms and blogs have represented a valuable source of information for several years, the sheer magnitude of the amount of data freely available and accessible over the World Wide Web is staggering, far beyond what the human mind can monitor and track. In addition, propaganda highlighting terrorist tactics and strategy has been showing up on the Internet, including the Al Qaeda handbook and streaming video demonstrating preoperational surveillance, target selection, and attack planning.[1] This information has tremendous value if identified, monitored, and studied, a task that these indexing tools has made more manageable.

16.5 Domain-Specific Tools

The emergence of tools designed specifically for public safety and security analysis is an exciting trend that I hope continues. While data mining and predictive analytics can be very intuitive, matching an analyst's style of investigation and query, many of the existing tools are somewhat limited in their ability to be used directly in the applied setting. To be sure, almost every example included in this book was generated using standard, off-the-shelf versions of existing data mining software tools, which supports the fact that these products can be used in the applied setting. The development of applications designed specifically for public safety and security analysis, however, makes them even more accessible. Similarly, advances in the visual depiction of complex analytical output are the focus of considerable research effort. Analytical output that builds on the end user's tacit knowledge and that can be transferred directly to the applied setting in support of information-based decisions and operations is an exciting area that is sure to grow considerably in the very near future.

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 Age [2]

If we have learned anything since 9/11, it is that we need to do more than collect data; we need to effectively analyze it in a way that facilitates the direct translation of the analytical output into information-based decisions and operational support.

Most of the data used in crime and intelligence analysis was gathered for some other purpose: billing, case management, and so on. In some situations, however, data sets are generated for the purpose of crime and intelligence analysis. Whenever possible, it is best to use data management systems that were designed for that purpose. There are a variety of database and data management tools commercially available. But be extremely cautious of analytical tools that also have an integrated database function that does not demonstrate clearly the future accessibility of the stored information.

One of the most exciting things in the public safety sector is the increased availability of technology resources for data management and analysis. This rapid increase in technology certainly is related to increased need, as well as to increased demand for new products. Unfortunately, some of the increase in the availability of products also might be related to the lucrative nature of public safety technology. Particularly after 9/11, the availability of funding for homeland security-related technology has increased astronomically. Unfortunately, not all of the products made available have the necessary internal capacity and flexibility to make them part of a long-term plan for meaningful, integrated technology enhancements.

The worst-case scenario associated with this proliferation of new technology is that information can be entered into some type of database or program and, when a need arises in the future, it becomes readily apparent that retrieval of the data is extremely difficult, if not impossible. In other words, the data has been put in an information "lockbox" and is inaccessible to other types of information processing or analytical tools. Unfortunately, this situation seems to arise with increasing frequency as agencies attempt to automate without consideration of future needs.

New programs and techniques are being developed daily, and most agencies have limited funds, particularly for purchasing new technology. Multifunction software can be appealing, particularly those programs that store as well as analyze data. While attractive to some, the use of a "one-stop shopping" approach to data storage and analysis can have a disastrous effect if it is difficult or impossible to extract data. Rather than saving money, these packages often end up costing more in the long run. Double entry of data is not only extremely unpleasant but also costly in terms of duplicative personnel efforts. Perhaps more importantly, though, it significantly increases the possibility of errors. When considering these options, therefore, one question that should always be asked is, "Can I get the data out of here if another software program comes along that I would like to use?" It is critical to ensure that data and information are maintained in a common, readily accessible format, preferably something that was designed to store data and permit access. Whether a simple spreadsheet program or something more elaborate like a data warehouse, the important consideration is whether it will allow the analyst to access the information and exploit new analytical packages as they become available.

If we have learned little else in the past few years, one thing that has become abundantly clear is that there probably is no Rosetta Stone of crime and intelligence information. The sharp organizations on the cutting edge of analytics have acquired and maintain the ability to integrate different data resources and exploit new technologies as soon as they become available. Similarly, one consistent theme throughout this text is the importance of maintaining flexibility in the analytical process. It is unclear what challenges will be present tomorrow, or what new analytical tools are just over the horizon. The paradigm shift associated with incorporating business tools and analysis in crime and intelligence analysis has been absolutely amazing, and analysts are an incredibly resourceful group out of necessity. As a result, it is exciting to consider what will be incorporated into our world tomorrow. Do not get left behind. Ensure that your data are accessible and available as these new technologies come on line.

16.7 Bibliography

1. Diamond, J. (2006). Insurgents give U.S. valuable training tool. *USA Today*, January 26.
2. The Markle Task Force on National Security in the Information Age, including James B. Steinberg, Vice President and Director, Foreign Policy Studies. (2002). *Protecting America's Freedom in the Information Age*, Markle Foundation.

This Page Intentionally Left Blank

Index

NOTE: An italicized *f*, *ff*, and *t* following a page number denotes a figure, multiple figures, and tables, respectively, on that page.

This Page Intentionally Left Blank

This Page Intentionally Left Blank