Policy Analysis

No. 584 December 11, 2006

Routing

Effective Counterterrorism and the Limited Role of Predictive Data Mining

by Jeff Jonas and Jim Harper

Executive Summary

The terrorist attacks on September 11, 2001, spurred extraordinary efforts intended to protect America from the newly highlighted scourge of international terrorism. Among the efforts was the consideration and possible use of "data mining" as a way to discover planning and preparation for terrorism. Data mining is the process of searching data for previously unknown patterns and using those patterns to predict future outcomes.

Information about key members of the 9/11 plot was available to the U.S. government prior to the attacks, and the 9/11 terrorists were closely connected to one another in a multitude of ways. The National Commission on Terrorist Attacks upon the United States concluded that, by pursuing the leads available to it at the time, the government might have derailed the plan.

Though data mining has many valuable uses, it is not well suited to the terrorist discovery problem. It would be unfortunate if data mining for terrorism discovery had currency within national security, law enforcement, and technology circles because pursuing this use of data mining would waste taxpayer dollars, needlessly infringe on privacy and civil liberties, and misdirect the valuable time and energy of the men and women in the national security community.

What the 9/11 story most clearly calls for is a sharper focus on the part of our national security agencies—their focus had undoubtedly sharpened by the end of the day on September 11, 2001—along with the ability to efficiently locate, access, and aggregate information about specific suspects.

Jeff Jonas is distinguished engineer and chief scientist with IBM's Entity Analytic Solutions Group. Jim Harper is director of information policy studies at the Cato Institute and author of Identity Crisis: How Identification Is Overused and Misunderstood.

CATO INSTITUTE

Though data mining has many valuable uses, it is not well suited to the terrorist discovery problem.

Introduction

The terrorist attacks on September 11, 2001, spurred extraordinary efforts intended to protect America from the newly highlighted scourge of international terrorism. Congress and the president reacted quickly to the attacks, passing the USA-PATRIOT Act,[1] which made substantial changes to laws that govern criminal and national security investigations. In 2004 the report of the National Commission on Terrorist Attacks upon the United States (also known as the 9/11 Commission) provided enormous insight into the lead-up to 9/11 and the events of that day. The report spawned a further round of policy changes, most notably enactment of the Intelligence Reform and Terrorism Prevention Act of 2004.[2]

Information about key members of the 9/11 plot was available to the U.S. government prior to the attacks, and the 9/11 terrorists were closely connected to one another in a multitude of ways. The 9/11 Commission concluded that, by pursuing the leads available to it at the time, the government might have derailed the plan.

What the 9/11 story most clearly calls for is sharper focus on the part of our national security agencies and the ability to efficiently locate, access, and aggregate information about specific suspects. Investigators should use intelligence to identify subjects of interest and then follow specific leads to detect and preempt terrorism. But a significant reaction to 9/11 beyond Congress's amendments to federal law was the consideration and possible use of "data mining" as a way to discover planning and preparation for terrorism.

Data mining is not an effective way to discover incipient terrorism. Though data mining has many valuable uses, it is not well suited to the terrorist discovery problem. It would be unfortunate if data mining for terrorism discovery had currency within national security, law enforcement, and technology circles because pursuing this use of data mining would waste taxpayer dollars, needlessly infringe on privacy and civil liberties, and misdirect the valuable time and energy of the men and women in the national security community.

We must continue to study and analyze the events surrounding the 9/11 attacks so that the most appropriate policies can be used to suppress terror, safeguard Americans, and protect American values. This is all the more important in light of recent controversies about the monitoring of telephone calls and the collection of telephone traffic data by the U.S. National Security Agency, as well as surveillance of international financial transactions by the U.S. Department of the Treasury.

While hindsight is 20/20, the details of the 9/11 story reveal that federal authorities had significant opportunities to unravel the 9/11 terrorist plot and potentially avert that day's tragedies. Two of the terrorists who ultimately hijacked and destroyed American Airlines flight 77 were already considered suspects by federal authorities and known to be in the United States. One of them was known to have associated with what a CIA official called a "major league killer."[3] Finding them and connecting them to other September 11 hijackers would have been possible—indeed, quite feasible—using the legal authority and investigative systems that existed before the attacks.

In the days and months before 9/11, new laws and technologies like predictive data mining were not necessary to connect the dots. What was needed to reveal the remaining 9/11 conspirators was better communication, collaboration, a heightened focus on the two known terrorists, and traditional investigative processes.

This paper is not intended to attack the hard-working and well-intentioned members of our law enforcement and intelligence communities. Rather, it seeks to illustrate that predictive data mining, while well suited to certain endeavors, is problematic and generally counterproductive in national security settings where its use is intended to ferret out the next terrorist.

The Story behind 9/11

Details of the run-up to 9/11 provide tremendous insight into what could have

been done to hamper or even entirely avert the 9/11 attacks. Failing to recognize these details and learn from them could compound the tragedy either by permitting future attacks or by encouraging acquiescence to measures that erode civil liberties without protecting the country.

In early January 2000 covert surveillance revealed a terrorist planning meeting in Kuala Lumpur that included Nawaf al-Hazmi, Khalid al-Mihdhar, and others.[4] In March 2000 the CIA was informed that Nawaf al-Hazmi departed Malaysia on a United Airlines flight for Los Angeles. (Although unreported at the time, al-Mihdhar was on the same flight.) The CIA did not notify the State Department and the FBI.[5] Later to join the 9/11 hijackings, both were known to be linked with al-Qaeda and specifically with the 1998 embassy bombings in Tanzania and Kenya.[6] As the 9/11 Commission reported, the trail was lost without a clear realization that it had been lost, and without much effort to pick it up again.[7]

In January 2001, almost one year after being lost in Bangkok, al-Mihdhar was on the radar screen again after being identified by a joint CIA-FBI investigation of the bombing of the USS *Cole*, the October 2000 attack on a U.S. guided missile destroyer in Yemen's Aden Harbor that killed 17 crew members and injured 39.[8] Even with this new knowledge the CIA did not renew its search for al-Mihdhar and did not make his identity known to the State Department (which presumably would have interfered with his plans to re-enter the United States).[9] Al-Mihdhar flew to New York City on July 4, 2001, on a new visa. As the 9/11 Commission reported, "No one was looking for him."[10]

On August 21, 2001, an FBI analyst who had been detailed to the CIA's Bin Laden unit finally made the connection and "grasped the significance" of Nawaf al-Hazmi and al-Mihdhar's visits to the United States. The Immigration and Naturalization Service was immediately notified. On August 22, 2001, the INS responded with information that caused the FBI analyst to conclude that al-Mihdhar might still be in the country.[11]

With the knowledge that the associate of a "major league killer" was possibly roaming free in the United States, the hunt by the FBI should have been on. The FBI certainly had a valid reason to open a case against these two individuals as they were connected to the ongoing USS *Cole* bombing investigation, the 1998 embassy bombing, and al-Qaeda.[12] On August 24, 2001, Nawaf al-Hazmi and al-Mihdhar were added to the State Department's TIPOFF[13] watchlist.[14]

Efforts to locate Nawaf al-Hazmi and al-Mihdhar initially foundered on confusion within the FBI about the sharing and use of data collected through intelligence versus criminal channels.[15] The search for al-Mihdhar was assigned to one FBI agent, his first ever counterterrorism lead.[16] Because the lead was "routine," he was given 30 days to open an intelligence case and make some effort to locate al-Mihdhar.[17] If more attention had been paid to these subjects, the location and detention of al-Mihdhar and Nawaf al-Hazmi could have derailed the 9/11 attack.[18]

Hiding in Plain Sight

The 9/11 terrorists did not take significant steps to mask their identities or obscure their activities. They were hiding in plain sight. They had P.O. boxes, e-mail accounts, drivers' licenses, bank accounts, and ATM cards.[19] For example, Nawaf al-Hazmi and al-Mihdhar used their true names to obtain California drivers' licenses and to open New Jersey bank accounts.[20] Nawaf al-Hazmi had a car registered, and his name appeared in the San Diego white pages with an address of 6401 Mount Ada Road, San Diego, California.[21] Mohamed Atta registered his red Pontiac Grand Prix car in Florida with the address 4890 Pompano Road, Venice.[22] Ziad Jarrah registered his red 1990 Mitsubishi Eclipse as well.[23] Fourteen of the terrorists got drivers' licenses or ID cards from either Florida or Virginia.[24]

The terrorists not only operated in plain sight, they were interconnected. They lived together, shared P.O. boxes and frequent flyer numbers, used the same credit card

The 9/11 terrorists did not take significant steps to mask their identities or obscure their activities.

numbers to make airline travel reservations, and made reservations using common addresses and contact phone numbers. For example, al-Mihdhar and Nawaf al-Hazmi lived together in San Diego.[25] Hamza al-Ghamdi and Mohand al-Shehri rented Box 260 at a Mail Boxes Etc. for a year in Delray Beach, Florida.[26] Hani Hanjour and Majed Moqed rented an apartment together at 486 Union Avenue, Patterson, New Jersey.[27] Atta stayed with Marwan al-Shehhi at the Hamlet Country Club in Delray Beach, Florida. Later, they checked into the Panther Inn in Deerfield Beach together.[28]

When Ahmed al-Nami applied for his Florida ID card he provided the same address that was used by Nawaf al-Hazmi and Saeed al-Ghamdi.[29] Wail al-Shehri purchased plane tickets using the same address and phone number as Waleed al-Shehri.[30] Nawaf al-Hazmi and Salem al-Hazmi booked tickets through Travelocity.com using the same Fort Lee, New Jersey, address and the same Visa card.[31] Abdulaziz al-Omari purchased his ticket via the American Airlines website and used Atta's frequent flyer number and the same Visa card and address as Atta (the same address used by Marwan al-Shehhi).[32] The phone number al-Omari used on his plane reservation was also the same as that of Atta and Wail and Waleed al-Shehri.[33] Hani Hanjour and Majed Moqed rented room 343 at the Valencia Hotel on Route 1 in Laurel, Maryland; they were joined by al-Mihdhar, Nawaf al-Hazmi, and Salem al-Hazmi.[34] While these are plentiful examples of the 9/11 terrorists' interconnectedness, even more connections existed.

Finding a Few Bad Guys

Interference with and detention of the right subset of the 9/11 terrorists might have derailed the plan.

In late August 2001 the FBI began to search for al-Mihdhar and Nawaf al-Hazmi.[35] The two might have been located easily even by a private investigator (PI). A PI would have performed a public records search using a service such as those provided by ChoicePoint or LexisNexis, perhaps both. These organizations aggregate public record data, assembling them into reports that simplify basic background investigations done by PIs, potential employers, potential landlords, and others. These databases include phone book data, driver's license data, vehicle registration data, credit header data, voter registration, property ownership, felony convictions, and the like. Such a search could have unearthed the driver's license, the car registration, and the telephone listing of Nawaf al-Hazmi and al-Mihdhar.[36]

Given the connections of Nawaf al-Hazmi and al-Mihdhar to terrorist activities overseas, the FBI, of course, could have sought subpoenas for credit card and banking information, travel information, and other business records. It could have conducted intensive surveillance under FISA, the Foreign Intelligence Surveillance Act, because the case involved a foreign power or an agent of a foreign power.[37] The FBI could not only have located these subjects but could have started to unravel their highly interconnected network, had it been pursuing available leads.

It is Monday morning quarterbacking, of course, to suggest that all 19 of the 9/11 hijackers could have been rolled up by the proper investigation. But interference with and detention of the right subset of the 9/11 terrorists might have "derailed the plan," as the 9/11 Commission concluded in its report.[38]

If our federal law enforcement and intelligence agencies needed anything, it was neither new technology nor more laws but simply a sharper focus and perhaps the ability to more efficiently locate, access, and aggregate information about specific suspects. They lacked this focus and capability—with tragic results.

Data Analysis and Data Mining

As we have seen, authorities could have and should have more aggressively hunted some of the 9/11 terrorists. If they had been hunted, they could have been found. Their web of connections would have led suffi-

ciently motivated investigators to information that could have confounded the 9/11 plot. Better interagency information sharing,[39] investigatory legwork in pursuit of genuine leads, and better training are what the 9/11 story most clearly calls for.

A number of policy changes followed the 9/11 attacks. The Intelligence Reform and Terrorism Prevention Act of 2004 revamped the nation's intelligence operations, and the USA-PATRIOT Act eased information sharing between investigators pursuing criminal and national security cases.

Data mining also gained some currency in national security and technology circles as a potential anti-terrorism tool,[40] though whether and to what extent it has been used are unclear. The Total Information Awareness program within the Department of Defense is widely believed to have contemplated using data mining, though the program's documentation is unclear.[41] The documentation discusses research on privacy-protecting technologies,[42] but Congress defunded the program in 2003 because of privacy concerns. However, the *National Journal* reported in February 2006 that research on "predict[ing] terrorist attacks by mining government databases and the personal records of people in the United States" has been moved from the Department of Defense to another group linked to the National Security Agency.[43]

In May 2004 the Government Accountability Office reported the existence of 14 data-mining programs, planned or operational, dedicated to analyzing intelligence and detecting terrorist activity, in the Departments of Defense, Education, Health and Human Services, Homeland Security, and Justice.[44] Ten of them were reported to use personal information. Of those, half use information acquired from the private sector, other agencies, or both.

"Data mining" is a broad and fairly loaded term that means different things to different people. Up to this point, discussions of data mining have probably been hampered by lack of clarity about its meaning. Indeed, collective failure to get to the root of the term "data mining" may have preserved disagreements among people who may be in substantial agreement.

Several authorities have offered definitions or discussions of data mining that are important touchstones, though they still may not be sufficiently precise. In its May 2004 report, for example, the Government Accountability Office surveyed the literature and produced the following definition of data mining: "the application of database technology and techniques—such as statistical analysis and modeling—to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results."[45] In a January 2006 report, the Congressional Research Service said:

> Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.[46]

Data mining is best understood as a subset of the broader practice of data analysis. Data analysis adds to the investigatory arsenal of national security and law enforcement by bringing together more information from more diverse sources and correlating the data. Finding previously unknown financial or communications links between criminal gangs, for example, can give investigators more insight into their activities and culture, strengthening the hand of law enforcement.

The key goal—and challenge—is to produce not just more information but more *useful* information. "Useful information" is information that puts the analyst in a position to act appropriately in a given context. It is the usefulness of the result—the fact that it

Data analysis adds to the investigatory arsenal of national security and law enforcement by bringing together more information from more diverse sources and correlating the data.

can be used effectively for a given purpose—that establishes the value of any given algorithm. The ultimate goal of data analysis is to discover knowledge.[47]

The term "predicate" is often used in law enforcement to refer to a piece of information that warrants further investigation or action. When a police officer sees a person attacking another with a knife, that is a sound basis, or predicate, for intervening by drawing his or her weapon and calling for a stop to the attack. When a police officer observes people appearing to "case" a store, that may be a predicate for making a display of authority or briefly questioning the people about their purposes. In Fourth Amendment law, probable cause to believe that information about a crime can be found in a particular place is a predicate for the issuance of a warrant to search that place.

Here is an example of a potential terrorism-related predicate: The combined facts that a particular person has been identified by an informant as having visited Afghanistan during June 2001 and participated in scuba training some years later, and that al-Qaeda plans to have divers mine cruise ships, may form a predicate for investigating the person or monitoring his or her communications.

Attempting to use predictive data mining to ferret out terrorists before they strike would be a subtle but important misdirection of national security resources.

In the first two examples discussed above—the knife attack and thieves casing a store—all the observations needed to establish a predicate for action were collected at once. Those are simple cases. Other than judging whether the response is proportional to the predicate, there is little need to parse them. But in the terror-suspect example, several observations made by different people at different times are combined to create the predicate. The fact that the person visited Afghan training camps might have come from an informant in Europe. The fact that he took scuba training might have come from business records in Corpus Christi, Texas. And the fact that al-Qaeda contemplated using scuba divers may have come from a computer captured in Pakistan. Because multiple observations are combined, this predicate can be said to result from data analysis. Data analysis brought information from diverse sources together to create new knowledge.

There are two loose categories of data analysis that are relevant to this discussion: subject based and pattern based.[48] Subject-based data analysis seeks to trace links from known individuals or things to others. The example just cited and the opportunities to disrupt the 9/11 plot described further above would have used subject-based data analysis because each of them starts with information about specific suspects, combined with general knowledge.

In pattern-based analysis, investigators use statistical probabilities to seek predicates in large data sets. This type of analysis seeks to find new knowledge, not from the investigative and deductive process of following specific leads, but from statistical, inductive processes. Because it is more characterized by prediction than by the traditional notion of suspicion, we refer to it as "predictive data mining."

The question in predictive data mining is whether and when it comes up with actionable information, with knowledge: suitable predicates for subsequent action. As we will discuss below, there are many instances when it does. But terrorism is not one. Attempting to use predictive data mining to ferret out terrorists before they strike would be a subtle but important misdirection of national security resources.

The possible benefits of predictive data mining for finding planning or preparation for terrorism are minimal. The financial costs, wasted effort, and threats to privacy and civil liberties are potentially vast. Those costs outstrip any conceivable benefits of using predictive data mining for this purpose.

Predictive Data Mining in Action

Predictive data mining has been applied most heavily in the area of consumer direct marketing. Companies have spent hundreds of millions if not billions of dollars imple-

menting and perfecting their direct marketing data-mining initiatives. Data mining certainly gives a "lift" to efforts to find people with certain propensities. In marketing, data mining is used to reduce the expense (to companies) and annoyance (to consumers) of unwanted advertising. And that is valuable to companies despite the fact that response rates to bulk mailings tuned by data mining improve by only single-digit percentages.

Consider how a large retailer such as Acme Discount Retail ("Acme Discount")—a fictional retailer trying to compete with Wal-Mart and Target—might use data mining: Acme Discount wants to promote its new store that just opened in a suburb of Chicago. It has many other stores and thousands of customers. Starting with the names and addresses of the top 1,000 Acme Discount customers, it contracts with a data broker to enhance what it knows about those customers. (This is known in database marketing as an "append" process.) Acme Discount may purchase magazine subscription and warranty card information (just to name a couple of likely data sources). Those sources augment what Acme Discount knows about its customers with such data points as income levels, presence of children, purchasing power, home value, and personal interests, such as a subscription to *Golf Digest*.

Thus, Acme Discount develops a demographic profile of what makes a good Acme Discount customer. For example, the ideal customer might be a family that subscribes to magazines of the *Vanity Fair* genre, that has two to four children, that owns two or fewer cars, and that lives in a home worth $150,000–$225,000. Acme Discount's next objective is to locate noncustomers near its new Chicago store that fit this pattern and market to them in the hope they will do business at the newly opened store. The goal is to predict as accurately as possible who might be swayed to shop at Acme Discount.

Despite all of this information collection and statistical analysis, the percent chance that Acme Discount will target someone willing to transact is in the low to mid single digits.[49] This means that false positives in marketers' searches for new customers are typically in excess of 90 percent.

The "damage" done by an imperfectly aimed direct-mail piece may be a dollar lost to the marketer and a moment's time wasted by the consumer. That is an acceptable loss to most people. The same results in a terror investigation would not be acceptable. Civil liberties violations would be routine and person-years of investigators' precious time would be wasted if investigations, surveillance, or the commitment of people to screening lists were based on algorithms that were wrong the overwhelming majority of the time.

Perhaps, though, more assiduous work by government authorities and contractors—using a great deal more data—could overcome the low precision of data mining and bring false positives from 90+ percent to the low single digits. For at least two related reasons, predictive data mining is not useful for counterterrorism: First, the absence of terrorism patterns means that it would be impossible to develop useful algorithms. Second, the corresponding statistical likelihood of false positives is so high that predictive data mining will inevitably waste resources and threaten civil liberties.

The Absence of Terrorism Patterns

One of the fundamental underpinnings of predictive data mining in the commercial sector is the use of training patterns. Corporations that study consumer behavior have millions of patterns that they can draw upon to profile their typical or ideal consumer. Even when data mining is used to seek out instances of identity and credit card fraud, this relies on models constructed using many thousands of known examples of fraud per year.

Terrorism has no similar indicia. With a relatively small number of attempts every year and only one or two major terrorist incidents every few years—each one distinct in terms of planning and execution—there are no meaningful patterns that show what

The statistical likelihood of false positives is so high that predictive data mining will inevitably waste resources and threaten civil liberties.

Without well-constructed algorithms based on extensive historical patterns, predictive data mining for terrorism will fail.

behavior indicates planning or preparation for terrorism.

Unlike consumers' shopping habits and financial fraud, terrorism does not occur with enough frequency to enable the creation of valid predictive models. Predictive data mining for the purpose of turning up terrorist planning using all available demographic and transactional data points will produce no better results than the highly sophisticated commercial data mining done today. The one thing predictable about predictive data mining for terrorism is that it would be consistently wrong.

Without patterns to use, one fallback for terrorism data mining is the idea that any anomaly may provide the basis for investigation of terrorism planning. Given a "typical" American pattern of Internet use, phone calling, doctor visits, purchases, travel, reading, and so on, perhaps all outliers merit some level of investigation. This theory is offensive to traditional American freedom, because in the United States everyone can and should be an "outlier" in some sense. More concretely, though, using data mining in this way could be worse than searching at random; terrorists could defeat it by acting as normally as possible.

Treating "anomalous" behavior as suspicious may appear scientific, but, without patterns to look for, the design of a search algorithm based on anomaly is no more likely to turn up terrorists than twisting the end of a kaleidoscope is likely to draw an image of the Mona Lisa.

Without well-constructed algorithms based on extensive historical patterns, predictive data mining for terrorism will fail. The result would be to flood the national security system with false positives—suspects who are truly innocent.

False Positives

The concepts of false positive and false negative come from probability theory. They have a great deal of use in health care, where tests for disease have known inaccuracy rates. A false positive, or Type I error, is when a test wrongly reports the presence of disease. A false negative, or Type II error, is when a test wrongly reports the absence of disease. Study of the false positive and false negative rates in particular tests, combined with the incidence of the disease in the population, helps determine when the test should be administered and how test results are used.

Even a test with very high accuracy—low false positives and false negatives—may be inappropriate to use widely if a disease is not terribly common. Suppose, for example, that a test for a particular disease accurately detects the disease (reports a true positive) 99 percent of the time and inaccurately reports the presence of the disease (false positive) 1 percent of the time. Suppose also that only one in a thousand, or 0.1 percent of the population, has that disease. Finally, suppose that if the test indicates the presence of disease the way to confirm it is with a biopsy, or the taking of a tissue sample from the potential victim's body.

It would seem that a test this good should be used on everyone. After all, in a population of 300 million people, 300,000 people have the disease, and running the test on the entire population would reveal the disease in 297,000 of the victims. But it would cause 10 times that number—nearly three million people—to undergo an unnecessary biopsy. If the test were run annually, every 5 years, or every 10 years, the number of people unnecessarily affected would rise accordingly.

In his book *The Naked Crowd*, George Washington University law professor Jeffrey Rosen discusses false positive rates in a system that might have been designed to identify the 19 hijackers involved in the 9/11 attacks.[50] Assuming a 99 percent accuracy rate, searching our population of nearly 300,000,000, some 3,000,000 people would be identified as potential terrorists.

Costs of Predictive Data Mining

Given the assumption that the devastation of the 9/11 attacks can be replicated,

some people may consider the investigation of 1 percent of the population (or whatever the false positive rate) acceptable, just as some might consider it acceptable for 10 people to undergo unnecessary surgery for every 1 person diagnosed with a certain disease. Fewer would consider a 5 percent error rate (or 15,000,000 people) acceptable. And even fewer would consider a 10 percent error rate (or 30,000,000 people) acceptable.

The question is not simply one of medical ethics or Fourth Amendment law but one of resources. The expenditure of resources needed to investigate 3,000,000, 15,000,000, or 30,000,000 fellow citizens is not practical from a budgetary point of view, to say nothing of the risk that millions of innocent people would likely be under the microscope of progressively more invasive surveillance as they were added to suspect lists by successive data-mining operations.

As we have shown, the unfocused, false-positive-laden results of predictive data mining in the terrorism context would waste national resources. Worse yet, the resources expended following those "leads" would detract directly from pursuing genuine leads that have been developed by genuine intelligence.

The corollary would be to threaten the civil liberties of the many Americans deemed suspects by predictive data mining. As Supreme Court precedents show, the agar in which reasonable suspicion grows is a mixture of specific facts and rational inferences. Thus, in *Terry v. Ohio*, the Supreme Court approved a brief interrogation and pat-down of men who appeared to have been "casing" a store for robbery.[51] An experienced officer observed their repeated, furtive passes by a store window; that gave him sufficient cause to approach the men, ask their business, and pat them down for weapons, which he found. The behavior exhibited by the men he frisked fit a pattern of robbery planning and did not fit any common pattern of lawful and innocent behavior. Any less correlation between their behavior and inchoate crime and the Court would likely have struck down the stop-and-frisk as a violation of the Fourth Amendment.

If predictive data mining is used as the basis for investigating specific people, it must meet this test: there must be a pattern that fits terrorism planning—a pattern that is exceedingly unlikely ever to exist—and the actions of investigated persons must fit that pattern while not fitting any common pattern of lawful behavior. Predictive data mining premised on insufficient pattern information could not possibly meet this test. Unless investigators can winnow their investigations down to data sets already known to reflect a high incidence of actual terrorist information, the high number of false positives will render any results essentially useless.

Predictive data mining requires lots of data. Bringing all the data, either physically or logically, into a central system poses a number of challenging problems, including the difficulty of keeping the data current and the difficulty of protecting so much sensitive data from misuse. Large aggregations of data create additional security risks from both insiders and outsiders because such aggregates are so valuable and attractive.

Many Americans already chafe at the large amount and variety of information about them available to marketers and data aggregators. Those data are collected from their many commercial transactions and from public records. Most data-mining efforts would rely on even more collections of transactional and behavioral information, and on centralization of that data, all to examine Americans for criminality or disloyalty to the United States or Western society. That level of surveillance, aimed at the entire citizenry, would be inconsistent with American values.

The Deceptiveness of Predictive Data Mining

Experience with a program that used predictive data mining shows that it is not very helpful in finding terrorists, even when abundant information is available. Using predic-

The unfocused, false-positive-laden results of predictive data mining in the terrorism context would waste national resources.

tive analysis—even in hindsight—the universe of "suspects" generated contains so many irrelevant entries that such analysis is essentially useless.

In his book *No Place to Hide*, *Washington Post* reporter Robert O'Harrow tells the story of how Hank Asher, owner of an information service called Seisint, concocted a way to fight back against terrorists in the days after September 11, 2001.

> Using artificial intelligence software and insights from profiling programs he'd created for marketers over the years, he told Seisint's computers to look for people in America who had certain characteristics that he thought might suggest ties to terrorists. Key elements included ethnicity and religion. In other words, he was using the data to look for certain Muslims. "Boom," he said, "32,000 people came up that looked pretty interesting.". . .
>
> In his darkened bedroom that night, he put the system through its paces over a swift connection to Seisint. "I got down to a list of 419 through an artificial intelligence algorithm that I had written," he recalled later. The list contained names of Muslims with odd ties or living in suspicious-seeming circumstances, at least according to Asher's analysis.[52]

Ultimately, Asher produced a list of 1,200 people he deemed the biggest threats. Of those, five were hijackers on the planes that crashed September 11, 2001.

What seems like a remarkable feat of predictive analysis is more an example of how deceptive hindsight can be. Asher produced a list of 9/11 terror suspects with a greater than 99 percent false positive rate—*after* the attack, its perpetrators, and their modus operandi were known.

The proof provided by the Seisint experience is not that there is a viable method in predictive analysis for finding incipient terrorism but that data mining of this type is almost certain to fail when information about attackers and their plans, associates, and methods is not known.

Data mining is almost certain to fail when information about attackers and their plans, associates, and methods is not known.

Conclusion

So how should one find bad guys? The most efficient, effective approach—and the one that protects civil liberties—is the one suggested by 9/11: pulling the strings that connect bad guys to other plotters.

Searching for terrorists must begin with actionable information, and it must follow logically through the available data toward greater knowledge. Predictive data mining always provides "information," but useful knowledge comes from context and from inferences drawn from known facts about known people and events.

The Fourth Amendment is a help, not a hindrance: It guides the investigator toward specific facts and rational inferences. When they focus on following leads, investigators can avoid the mistaken goal of attempting to "predict" terrorist attacks, an effort certain to flood investigators with false positives, to waste resources, and to open the door to infringements of civil liberties. That approach focuses our national security effort on developing information about terrorism plotters, their plans, and associates. It offers no panacea or technological quick fix to the security dilemmas created by terrorism. But there is no quick fix. Predictive data mining is not a sharp enough sword, and it will never replace traditional investigation and intelligence, because it cannot predict precisely enough who will be the next bad guy.

Since 9/11 there has been a great deal of discussion about whether data mining can prevent acts of terrorism. In fact, the most efficient means of detecting and preempting terrorism have been within our grasp all along. Protecting America requires no predictive-data-mining technologies.

Indeed, if there is a lesson to be learned from 9/11, it is not very groundbreaking. It is this: Enable investigators to efficiently dis-

cover, access, and aggregate relevant information related to actionable suspects. Period. Sufficient dedication of national resources to more precisely "pull the strings" offers the best chance of detecting and preempting future acts of terrorism.

Notes

1. Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001 (USA PATRIOT Act), Pub. L. No. 107-56 (Oct. 12, 2001).

2. Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. No. 108-458 (Dec. 17, 2004).

3. National Commission on Terrorist Attacks upon the United States, *The 9/11 Commission Report*, 2004, p. 268 (hereinafter *9/11 Commission Report*).

4. Ibid., p. 181.

5. Ibid., pp. 181–82.

6. Ibid., p. 181.

7. Ibid., p. 266.

8. Ibid., p. 266.

9. Ibid., pp. 266–67.

10. Ibid., p. 269.

11. Ibid., p. 270.

12. Ibid., p. 271.

13. The TIPOFF database contains a list if foreigners who will be denied a U.S. visa.

14. *9/11 Commission Report*, p. 270.

15. Ibid., p. 271.

16. Ibid., p. 288.

17. Ibid., p. 271.

18. Ibid., p. 272.

19. Tim Golden et al., "A Nation Challenged: The Plot," *New York Times*, September 23, 2001.

20. *9/11 Commission Report*, p. 539, n 85.

21. Ibid.; and Jane Black, "Don't Make Privacy the Next Victim of Terror," *BusinessWeek Online*, http://www.businessweek.com/bwdaily/dnflash/oct2001/nf2001104_7412.htm.

22. Dan Eggen et al., "The Plot: A Web of Connections," *WashingtonPost.com*, October 4, 2001, http://www.washingtonpost.com/wp-srv/nation/graphics/attack/investigation_24.html (hereinafter "Web of Connections").

23. "Web of Connections."

24. Ibid.

25. Ibid.

26. Ibid.

27. Ibid.

28. Ibid.

29. Ibid.

30. Ibid.

31. Ibid.

32. Ibid.

33. Ibid.

34. Ibid.

35. *9/11 Commission Report*, pp. 271–72.

36. Ibid., p. 539, n 85.

37. 50 U.S.C. § 1804.

38. *9/11 Commission Report*, p. 272.

39. Ibid., p. 271.

40. Arshad Mohammed and Sara Kehaulani Goo, "Government Increasingly Turning to Data Mining," *Washington Post*, June 15, 2006, http://www.washingtonpost.com/wp-dyn/content/article/2006/06/14/AR2006061402063.html.

41. See Defense Advanced Research Projects Agency, Information Awareness Office, "Report to Congress Regarding the Terrorism Information Awareness Program, " May 30, 2003, pp. 7–8, 17, A-4, A-14, A-15 (referring variously to "discovery of . . . patterns of activity"; "ability to automatically learn patterns"; "training software algorithms to recognize patterns"; and "developing technology to . . . suggest previously unknown but potentially significant patterns"), http://foi.missouri.edu/totalinfoaware/tia2.pdf.

42. Ibid., pp. 6–7.

43. Shane Harris, "TIA Lives On," *National Journal*, February 23, 2006, http://nationaljournal.com/about/njweekly/stories/2006/0223nj1.htm.

44. Government Accountability Office, "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548, May 3004.

45. Ibid., p. 1.

46. Jeffrey W. Seifert, "Data Mining and Homeland Security: An Overview," Congressional Research Service, updated January 27, 2006 (Order Code RL31798). See also K. A. Taipale, "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data," *Columbia Science and Technology Law Review* 22–23 (2003), http://papers..ssrn.can/abstract=5467827.

47. The major annual conference on data mining is called "KDD," for Knowledge Discovery and Data Mining. See http://www.acm.org/sigs/sigkdd/kdd2006.

48. See Martha Baer et al., *SAFE: The Race to Protect Ourselves in a Newly Dangerous World* (New York: Harper Collins 2005), p. 331; and Mary DeRosa, "Data Mining and Data Analysis for Counterterrorism," Center for Strategic and International Studies, March 2004.

49. Direct marketing results are dependent upon many factors such as industry and offer. For example, offering a consumer a loss-leader discount raises response rates. Despite billions invested and unprecedented access to U.S. consumer behavior data, current direct marketing response rates industrywide range from 5.78 percent for telephone solicitation to 0.04 percent for direct response television. Direct Marketing Association, "DMA Releases New Response Rate Report," news release, October 17, 2004, http://www.the-dma.org/cgi/dispnewsstand?article=2891.

50. Jeffrey Rosen, *The Naked Crowd* (New York: Random House, 2004), pp. 104–7.

51. 392 U.S. 1 (1968).

52. Robert O'Harrow Jr., *No Place to Hide* (New York: Free Press, 2005), pp. 98, 102.

CATO INSTITUTE
1000 Massachusetts Ave., N.W.
Washington, D.C. 20001

Published by the Cato Institute, Policy Analysis is a regular series evaluating government policies and offering proposals for reform. Nothing in Policy Analysis should be construed as necessarily reflecting the views of the Cato Institute or as an attempt to aid or hinder the passage of any bill before Congress. Contact the Cato Institute for reprint permission. Additional copies of Policy Analysis are $6.00 each ($3.00 each for five or more). To order, or for a complete listing of available studies, write the Cato Institute, 1000 Massachusetts Ave., N.W., Washington, D.C. 20001 or call toll free 1-800-767-1241 (8:30-4:30 eastern time).
Fax (202) 842-3490 • www.cato.org

CATO INSTITUTE