

CSIS REPORT

DATA MINING AND DATA ANALYSIS FOR COUNTERTERRORISM

Author
Mary DeRosa

March 2004



**DATA MINING
AND
DATA ANALYSIS
FOR
COUNTERTERRORISM**

Author
Mary DeRosa

March 2004



About CSIS

For four decades, the Center for Strategic and International Studies (CSIS) has been dedicated to providing world leaders with strategic insights on—and policy solutions to—current and emerging global issues.

CSIS is led by John J. Hamre, former U.S. deputy secretary of defense. It is guided by a board of trustees chaired by former U.S. senator Sam Nunn and consisting of prominent individuals from both the public and private sectors.

The CSIS staff of 190 researchers and support staff focus primarily on three subject areas. First, CSIS addresses the full spectrum of new challenges to national and international security. Second, it maintains resident experts on all of the world's major geographical regions. Third, it is committed to helping to develop new methods of governance for the global age; to this end, CSIS has programs on technology and public policy, international trade and finance, and energy.

Headquartered in Washington, D.C., CSIS is private, bipartisan, and tax-exempt. CSIS does not take specific policy positions; accordingly, all views expressed herein should be understood to be solely those of the author(s).

Library of Congress Cataloging-in-Publication Data
CIP information available on request.
ISBN 0-89206-443-9

© 2004 by the Center for Strategic and International Studies.
All rights reserved

The CSIS Press

Center for Strategic and International Studies
1800 K Street, N.W., Washington, D.C. 20006
Tel: (202) 887-0200
Fax: (202) 775-3199
E-mail: books@csis.org
Web site: <http://www.csis.org/>

Contents

Acknowledgments.....	iv
Executive Summary	v
Introduction.....	1
I. Background and Some Terminology	3
II. Why Data Mining for Counterterrorism	5
III. The Process	9
Gathering and Processing the Data	
Finding Search Models	
Decisionmaking	
IV. The Risks	13
False Positives	
Inadequate Government Control of Data	
V. Mitigating Privacy Concerns with Technology	16
Resolving False Positives	
Anonymization	
Audit Technology	
Rule-based Processing	
VI. Areas for Policy Development.....	20
Data-mining Research	
Clarity about Use of Data Mining and Data Analysis	
Use of Search Results	
Controls on the Use of Identifying Information	
Conclusion	23
About the Author	24

Acknowledgments

This report would not have been possible without the excellent presentations, expertise, and insights of the speakers at the CSIS Data Mining Roundtables: David Jensen, research assistant professor of computer science and director of the Knowledge Discovery Laboratory, Department of Computer Science, University of Massachusetts; Jeff Jonas, founder and chief scientist, Systems Research & Development; Teresa Lunt, principal scientist, Computer Sciences Laboratory, Xerox Palo Alto Research Center; Farzad Mostashari, assistant coordinator for epidemiology services, New York City Department of Health and Mental Hygiene; Ted Senator, program manager, Defense Advanced Research Projects Agency; Gary Strong, director of behavioral research and biometrics, Science and Technology Directorate, Department of Homeland Security; Latanya Sweeney, assistant professor of computer science, technology, and policy, School of Computer Science, Institute for Software Research International, and director, Laboratory for International Data Privacy, Carnegie Mellon University; and James Zimbardi, vice president, business and government services, ChoicePoint.

In addition, I would like to acknowledge the invaluable assistance of the people who reviewed drafts of this document. Thank you to Gerald Epstein, David Jensen, Jeff Jonas, Jason Keiber, James Lewis, Teresa Lunt, Mary McCarthy, Robert Popp, Steve Rubley, Ted Senator, and William Still for their time, expertise, and excellent comments, which contributed to a stronger product. Any remaining errors are solely my responsibility.

Finally, special thanks to Jason Keiber for his research assistance and to Joelle Laszlo for her tireless substantive, research, and administrative support.

Executive Summary

Defeating terrorism requires a more nimble intelligence apparatus that operates more actively within the United States and makes use of advanced information technology. Data-mining and automated data-analysis techniques are powerful tools for intelligence and law enforcement officials fighting terrorism. But these tools also generate controversy and concern. They make analysis of data—including private data—easier and more powerful. This can make private data more useful and attractive to the government. Data mining and data analysis are simply too valuable to prohibit, but they should not be embraced without guidelines and controls for their use. Policymakers must acquire an understanding of data-mining and automated data-analysis tools so that they can craft policy that encourages responsible use and sets parameters for that use.

This report builds on a series of roundtable discussions held by CSIS. It provides a basic description of how data-mining techniques work, how they can be used for counterterrorism, and their privacy implications. It also identifies where informed policy development is necessary to address privacy and other issues.

One of the first problems with “data mining” is that there are varying understandings of what the term means. “Data mining” actually has a relatively narrow meaning: it is a process that uses algorithms to discover predictive patterns in data sets. “Automated data analysis” applies models to data to predict behavior, assess risk, determine associations, or do other types of analysis. The models used for automated data analysis can be based on patterns (from data mining or discovered by other methods) or subject based, which start with a specific known subject. There are a number of common misconceptions about these techniques. For example, data mining and data analysis do not increase access to private data. Data mining and data analysis certainly can make private data more useful, but they can only operate on data that is already accessible. Another myth is that data mining and data analysis require masses of data in one large database. In fact, data mining and analysis can be conducted using a number of databases of varying sizes.

Although these techniques are powerful, it is a mistake to view data mining and automated data analysis as complete solutions to security problems. Their strength is as tools to assist analysts and investigators. They can automate some functions that analysts would otherwise have to perform manually, they can help prioritize attention and focus an inquiry, and they can even do some early analysis and sorting of masses of data. But in the complex world of counterterrorism, they are not likely to be useful as the only source for a conclusion or decision. When these techniques are used as more than an analytical tool, the potential for harm to individuals is far more significant.

Automated data-analysis techniques can be useful tools for counterterrorism in a number of ways. One initial benefit of the data-analysis process is to assist in

the important task of accurate identification. Technologies that use large collections of identity information can help resolve whether two records represent the same or different people. Accurate identification not only is critical for determining whether a person is of interest for a terrorism-related investigation, it also makes the government better at determining when someone is not of interest, thereby reducing the chance that the government will inconvenience that person.

Subject-based “link analysis” uses public records or other large collections of data to find links between a subject—a suspect, an address, or other piece of relevant information—and other people, places, or things. This technique is already being used for, among other things, background investigations and as an investigatory tool in national security and law enforcement investigations.

Pattern-based analysis may also have potential counterterrorism uses. Pattern-based queries take a predictive model or pattern of behavior and search for that pattern in data sets. If models can be perfected, pattern-based searches could provide clues to “sleeper” cells made up of people who have never engaged in activity that would link them to known terrorists.

The potential benefits for counterterrorism are significant. But when the government can analyze private data so much more effectively, that data could become more attractive, and the government’s power to affect the lives of individuals can increase. There is significant public unease about whether protections for privacy are adequate to address the negative consequences of increased government use of private data. These concerns are heightened because there is so little understanding of how the government might use these data-analysis tools. Nor is there typically much public debate or discussion before these tools are adopted. This lack of transparency not only can make the government’s decisions less informed, but it increases public fear and misunderstanding about uses of these techniques.

Perhaps the most significant concern with data mining and automated data analysis is that the government might get it wrong, and innocent people will be stigmatized and inconvenienced. This is the problem of “false positives”—when a process incorrectly reports that it has found what it is looking for. With these tools, a false positive could mean that because of bad data or imperfect search models a person is incorrectly identified as having a terrorist connection.

But even if results are accurate, government mechanisms are currently inadequate for controlling the use of these results. If they are not controlled, private data can be used improperly. There are no clear guidelines now for who sees private data, for what reasons, how long it is retained, and to whom it is disseminated. A related concern is “mission creep”—the tendency to expand the use of a controversial technique beyond the original purposes. Use of controversial tools may be deemed acceptable given the potential harm of catastrophic terrorism, but there will then be a great temptation to expand their use to address other law enforcement or societal concerns ranging from the serious to the trivial.

One important avenue for addressing many of these challenges to privacy and liberties, at least in part, is technology. Some privacy-protecting technology is already available and much more is being researched. Researchers are looking at methods to perfect search models and cleanse data to reduce false positives: “anonymizing” technology designed to mask or selectively reveal identifying data so that the government can conduct searches and share data without knowing the names and identities of Americans; audit technology to “watch the watchers” by recording activity in databases and networks to provide effective oversight; and rule-processing or permissioning technology that ensures that data can be retrieved only in a manner that is consistent with privacy safeguards and other rules.

Although this technology can address some of the risks with use of data-mining and automated data-analysis techniques, it will not be adequate on its own. Policy action is needed to ensure that controls and protections accompany use of these powerful tools. The policy issues that require attention include:

- ◆ ***Research on data mining and automated data analysis.*** Data-mining and automated data-analysis tools have great potential for counterterrorism, but to realize that potential fully, more research is needed. The government should support this research. A government policy for this research should take into account the context in which these tools may eventually be deployed. This means research on privacy-protecting technology and even some analysis of privacy policy issues should be included.
- ◆ ***Clarity about use of data mining and automated data analysis.*** One of the principal reasons for public concern about these tools is that there appears to be no consistent policy guiding decisions about when and how to use them. Policies for data-mining and automated data-analysis techniques should set forth standards and a process for decisionmaking on the type of data-analysis technique to use—subject based or pattern based, for example—and the data that will be accessed. They should mandate inquiries into data accuracy and the level of errors that the analysis is expected to generate, and they should require government to put in place a mechanism for correcting errors before operations begin.
- ◆ ***Use of search results.*** There should also be a consistent policy on what action can be taken based on search results. When automated data-analysis results are used only to further analysis and investigation, and not as the sole basis for detention or some other government action, there are fewer possible negative consequences for individuals. Therefore, guidance is necessary on the circumstances, if any, under which results can be used as the basis for action.
- ◆ ***Controls on the use of identifying information.*** Currently no clear guidance exists for government entities and employees about how to handle private data, and this lack of direction can lead to mistakes and inconsistent use of data. Perhaps the most important step to address

privacy concerns with the use of data mining and automated data analysis is for the executive branch to implement clear guidelines for federal employees on how they may access, use, retain, and disseminate private data.

Data Mining and Data Analysis for Counterterrorism

Mary DeRosa

Introduction

As almost everyone now recognizes, the fight against terrorism requires the government to find new approaches to intelligence gathering and analysis. At the same time, advances in technology provide new opportunities to collect and use information. “Data mining” is one technique that has significant potential for use in countering terrorism. Data-mining and automated data-analysis techniques are not new; they are already being used effectively in the private sector and in government. They have generated concern and controversy, however, because they allow the government far greater ability to use and analyze private information effectively. This makes private data a more attractive and powerful resource for the government and increases the potential for government intrusion on privacy. Recent high-profile government programs that would explore or employ data-mining and data-analysis techniques for counterterrorism have caused public concern and congressional action, but the debate has not always been fully informed. Resolving this debate intelligently and rationally is critical if we are to move forward in protecting both our security and our liberties.

Legislative action on data mining has had an “all-or-nothing” quality. For example, Congress terminated the controversial Terrorism Information Awareness (earlier called Total Information Awareness) (TIA) research program at the Department of Defense Advanced Research Projects Agency (DARPA), rather than deal with concerns by imposing conditions or controls on the program.¹ Other current legislative proposals would impose a “moratorium” on all “data-

¹ Conference report on H.R. 2658, Department of Defense Appropriations Act, 2004, H.R. Conf. Rep. No. 108-283 (9/24/2003), available at <http://thomas.loc.gov/cgi-bin/query/R?r108:FLD001:H08501>.

mining” activities² and prohibit the use of “hypothetical scenarios” in searching databases for law enforcement, national security, or intelligence purposes.³ That a complex policy issue has been handled with so little nuance is due, at least in part, to a lack of understanding in Congress and the public of what data mining is, its current and potential uses, and how it might be controlled.

Policy on data mining and related techniques that impact privacy should not rely solely on prohibition. Policymakers must make informed decisions about how to oversee and control government use of private information most effectively when using these techniques. To make these decisions, policymakers should ask probing questions:

- ◆ Is the proposed program for research or for application?
- ◆ If research, does the program include research on privacy protection?
- ◆ If application, what type of data analysis will be used?
- ◆ What data will be accessed?
- ◆ What level of errors—false positives and false negatives⁴—is the analysis expected to generate?
- ◆ For what purpose is the analysis being used and how narrowly tailored is it to that purpose?
- ◆ Are there ways to assure that its use will not be expanded beyond this purpose without further debate?
- ◆ Is the data mining or automated analysis to be used only as an analytical or investigatory tool, or will decisions that affect individuals be made based on data-analysis results alone?
- ◆ What controls are being applied to collection, use, retention, and dissemination of identities?
- ◆ Is technology that can assist with privacy protection being used?

CSIS held a series of roundtable discussions to increase understanding of data mining and related techniques. This paper builds on those discussions and looks at what these techniques are, how they can be used for counterterrorism, what the privacy implications of these processes are, and how they might be addressed. Finally, it identifies some of the significant areas where informed policy development is necessary.

² Data-Mining Moratorium Act of 2003 (S.188, introduced in Senate 1/16/2003), available at [http://thomas.loc.gov/cgi-bin/query/z?c108:S.188](http://thomas.loc.gov/cgi-bin/query/z?c108:S.188;); Protecting the Rights of Individuals Act (S.1552, introduced in Senate 7/21/03) available at <http://thomas.loc.gov/cgi-bin/query/z?c108:S.1552>.

³ Citizens’ Protection in Federal Databases Act (S.1484, introduced in Senate 7/29/2003) available at <http://thomas.loc.gov/cgi-bin/query/z?c108:S.1484>.

⁴ A false positive is when a process incorrectly reports that it has found what it is looking for; a false negative is when it incorrectly reports that it has not found what it is looking for.

I. Background and Some Terminology

One of the first problems with the term “data mining” is that it means different things to different audiences; lay use of the term is often much broader than its technical definition. A good description of what data mining does is: “discover useful, previously unknown knowledge by analyzing large and complex” data sets.⁵ Data mining is one step in a broader “knowledge-discovery” process. Data mining itself is a relatively narrow process of using algorithms to discover predictive patterns in data sets.⁶ The process of applying or using those patterns to analyze data and make predictions is not data mining. A more accurate term for those analytical applications is “automated data analysis,” which can include analysis based on pattern queries (the patterns can be developed from data mining or by methods other than data mining) or on less controversial subject-based queries. The term “data mining” is often used casually to refer both to actual data mining and the application of automated data-analysis tools. Both sets of techniques are relevant to counterterrorism, and this paper addresses both.

It is important also to understand what these terms do not include. Data-mining and automated data-analysis tools are not for locating and retrieving pieces of data in databases that might have been hard to find. This Google-type function is important but separate. Automated data-analysis tools find previously unknown knowledge through links, associations, and patterns in data.⁷ Also, these tools are not for discovering just any knowledge; they are used to discover useful knowledge in data. It is possible to find an endless number of patterns and associations in masses of data; many will be statistically significant, but they will not have any real world significance. An essential and sometimes extremely difficult aspect of data mining and automated data analysis is finding the patterns and associations that have value—the ones that actually mean something.⁸

There are two general ways to use automated data analysis: by following subject-based queries or pattern-based queries. Subject-based queries start with a specific and known subject and search for more information. The subject could be an identity—a suspect, an airline passenger, or a name on a watch list, for example—or it could be something else specific, like a place or a telephone number. A subject-based query will seek more information about and a more complete understanding of the subject, such as activities a person has engaged in or links to other people, places, and things. It will also provide leads to other

⁵ David Jensen, “Data Mining in Networks,” presentation at CSIS Data Mining Roundtable, Washington, D.C., July 23, 2003. Presentation slides available at <http://kdl.cs.umass.edu/people/jensen/papers/nrcdbsse02.html>, at slide 10. I have used the phrase “data sets” rather than “databases” as Jensen did to avoid confusion. As Jensen states, data mining does not require one large database but can be conducted on distributed sets of data. This will be discussed in more detail later.

⁶ Jensen, “Data Mining in Networks,” slide 9; K.A. Taipale, “Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data,” *Columbia Science and Technology Law Review* 5 (December 2003): 28, available at <http://stlr.org/cite.cgi?volume=5&article=2>.

⁷ Jensen, “Data Mining in Networks,” slide 10.

⁸ Taipale, “Data Mining and Domestic Security,” pp. 23, 24.

subjects that can be investigated. “Link analysis” is a type of subject-based query that is already in use in the private sector and in government. Subject-based queries are not related to “data mining,” but they do fall into the category of automated data analysis.

An example of subject-based queries used in the private sector is the Non Obvious Relationship Awareness™ or NORA™ software that Systems Research and Development (SRD) has developed, which is used in Las Vegas to prevent fraud, cheating, and theft from casinos. The gaming industry has developed an “excluded persons” watch list with names of individuals who are prohibited from entering casinos. NORA can search through massive databases to find whether there are associations between, for example, a person seeking a job at a casino and a person on a watch list. Maybe the applicant once roomed with, sold a house to, or used as an employment reference, a person who is on a watch list. This is information the casino can use to focus its investigatory resources.⁹

Pattern-based queries involve identifying some predictive model or pattern of behavior and searching for that pattern in data sets. These predictive models can be discovered through data mining, or they can come from outside knowledge—intelligence or expertise about a subject. However the patterns are obtained, the process involves looking for occurrences of these patterns of activity in data.

Probably the most well-known use of pattern-based searching involves credit card fraud. Banks search databases of credit card transactions, some of which are known to be fraudulent, and determine, through data mining or otherwise, the patterns of fraudulent activity. A simple example of such a pattern is use of a stolen credit card for a small purchase at a gas station—done to confirm whether the card is valid—before making a very significant purchase.¹⁰ The banks then use these patterns to identify fraudulent activity in databases of ongoing credit card transactions and take steps to stop that activity. Another long-standing use of pattern-based queries is by the U.S. Treasury Department’s Financial Crimes Enforcement Network (FinCEN) to detect money-laundering activity. FinCEN looks at databases of financial data and identifies patterns of previous known cases of money laundering. For example, money laundering often involves people injecting large amounts of money into the financial system in small increments, under the guise of an existing business, and then using that money to import overpriced goods, so that the money flows out of the United States. None of these steps independently would necessarily be suspicious, but the whole pattern is consistent with money laundering. FinCEN looks for these patterns in data that exists in a variety of databases and uses the information it collects in its enforcement activities.¹¹

⁹ Jeff Jonas, “Using Data to Detect and Preempt Bad Things from Happening,” presentation at CSIS Data Mining Roundtable, Washington, D.C., July 23, 2003.

¹⁰ Jensen, “Data Mining in Networks,” slide 11.

¹¹ Ibid., slides 5–9; U.S. Congress, Office of Technology Assessment, *Information Technologies for the Control of Money Laundering*, OTA-ITC-630 (Washington, D.C.: GPO, September 1995).

Both subject-based and pattern-based queries have the potential to be useful in counterterrorism, but we are currently farther along in our ability to deploy subject-based queries effectively in the counterterrorism realm. Moreover, subject-based queries raise somewhat fewer policy difficulties because they are more like the kinds of inquiries that are common in intelligence and law enforcement practice; that is, they are developed from a particularized suspicion or reason for interest and seek additional information.¹² Pattern-based queries are less familiar in the law enforcement and intelligence worlds in that they do not arise from a particular interest in a person, place, or thing. Instead, they seek information about people, places, and things based on patterns of activity, none of the components of which might on its own arouse suspicion or be in any way improper.

II. Why Data Mining for Counterterrorism?

It is clear now that the threat we face from terrorism is far different from Cold War threats and requires adjustments to our approach to intelligence collection and analysis. Unlike our Cold War adversaries, the terrorists are loosely organized in a diffuse, nonhierarchical structure. We cannot rely to the degree we did in the Cold War on finding a relatively few rich sources of intelligence that will provide insight into capabilities, tactics, and plans. Although all traditional intelligence-collection methods remain important, understanding the terrorists and predicting their actions requires us to rely more on making sense of many small pieces of information.

In the Cold War, much of our edge came from acquiring significant pieces of critical information clandestinely and protecting them from disclosure. Now, that kind of information is far more difficult to find. The September 11, 2001, attacks illustrate this point. Even in hindsight, we can see no single source—other than perhaps an extraordinarily well-placed human asset—that could have provided the full or even a large part of the picture of what was being planned. We have seen a number of clues, however, that if recognized, combined, and analyzed might have given us enough to track down the terrorists and stop their plan. Therefore, although we must still focus on improving our ability to collect human and other traditional sources of intelligence,¹³ our edge now will come more from breadth of access to information and quality analysis.¹⁴ For counterterrorism, we must be able to find a few small dots of data in a sea of information and make a picture out of them.

Data-mining and automated data-analysis techniques are not a complete solution. They are only tools, but they can be powerful tools for this new intelligence requirement. Although intuition and continual hypothesizing remain

¹² Paul Rosenzweig, “Proposals for Implementing the Terrorism Information Awareness System,” Legal Memorandum No. 8, Heritage Foundation (August 7, 2003), p. 6.

¹³ See Robert Bryant et al., “America Needs More Spies,” *The Economist*, vol. 368, no. 8332 (July 12, 2003).

¹⁴ Jensen, “Data Mining in Networks,” slide 17, paraphrasing “DIA analyst, 2002.”

irreplaceable parts of the analytic process, these techniques can assist analysts and investigators by automating some low-level functions that they would otherwise have to perform manually. These techniques can help prioritize attention and provide clues about where to focus, thereby freeing analysts and investigators to engage in the analysis that requires human judgment. In addition, data mining and related techniques are useful tools for some early analysis and sorting tasks that would be impossible for human analysts. They can find links, patterns, and anomalies in masses of data that humans could never detect without this assistance. These can form the basis for further human inquiry and analysis.

One initial potential benefit of the data-analysis process is that the use of large databases containing identifying information assists in the important task of accurate identification. As will be explained in more detail in the next section,¹⁵ more information makes it far easier to resolve whether two or more records represent the same or different people. For example, an investigator might want to determine whether the John Doe boarding a plane is the same person as the Jack Doe on a terrorist watch list or the J.R. Doe that shared a residence with a suspected terrorist. If the government has only names, it is virtually impossible to resolve these identities for certain; if the government has a social security number, a date of birth, or an address, it is easier to make that judgment accurately. The task of identity resolution is far easier to perform when there are large data sets of identifying information to call on. Not incidentally, identity resolution also makes the government better at determining when a person in question is not the one suspected of terrorist ties, thereby potentially reducing inconvenience to that person.

A relatively simple and useful data-analysis tool for counterterrorism is subject-based “link analysis.” This technique uses aggregated public records or other large collections of data to find links between a subject—a suspect, an address, or other piece of relevant information—and other people, places, or things. This can provide additional clues for analysts and investigators to follow. Link analysis is a tool that is available now and is used for, among other things, background checks of applicants for sensitive jobs and as an investigatory tool in national security and law enforcement investigations.

A hindsight analysis of the September 11 attacks provides an example of how simple, subject-based link analysis could be used effectively to assist investigations or analysis of terrorist plans. By using government watch list information, airline reservation records, and aggregated public record data, link analysis could have identified all 19 September 11 terrorists—for follow-up investigation—before September 11.¹⁶ The links can be summarized as follows:

¹⁵ See discussion on pages 10 and 11 herein.

¹⁶ Of course, this kind of analysis will always appear neater and easier with hindsight, but it is a useful demonstration nonetheless.

Direct Links—Watch List Information

- ♦ **Khalid Almihdhar** and **Nawaf Alhazmi**, both hijackers of American Airlines (AA) Flight 77, which crashed into the Pentagon, appeared on a U.S. government terrorist watch list. Both used their real names to reserve their flights.
- ♦ **Ahmed Alghamdi**, who hijacked United Airlines (UA) Flight 175, which crashed into the World Trade Center South Tower, was on an Immigration and Naturalization Service (INS) watch list for illegal or expired visas. He used his real name to reserve his flight.

Link Analysis—One Degree of Separation

- ♦ Two other hijackers used the same contact address for their flight reservations that Khalid Almihdhar listed on his reservation. These were **Mohamed Atta**, who hijacked AA Flight 11, which crashed into the World Trade Center North Tower, and **Marwan Al Shehhi**, who hijacked UA Flight 175.
- ♦ **Salem Alhazmi**, who hijacked AA Flight 77, used the same contact address on his reservation as Nawaf Alhazmi.
- ♦ The frequent flyer number that Khalid Almihdhar used to make his reservation was also used by hijacker **Majed Moqed** to make his reservation on AA Flight 77.
- ♦ **Hamza Alghamdi**, who hijacked UA Flight 175, used the same contact address on his reservation as Ahmed Alghamdi used on his.
- ♦ **Hani Hanjour**, who hijacked AA Flight 77, lived with both Nawaf Alhazmi and Khalid Almihdhar, a fact that searches of public records could have revealed.

Link Analysis—Two Degrees of Separation

- ♦ Mohamed Atta, already tied to Khalid Almihdhar, used a telephone number as a contact number for his reservation that was also used as a contact number by **Waleed Alshehri**, **Wail Alshehri**, and **Abdulaziz Alomari**, all from AA Flight 11, and by **Fayez Ahmed** and **Mohand Alshehri**, both from UA Flight 175.
- ♦ Public records show that Hamza Alghamdi lived with **Saeed Alghamdi**, **Ahmed Al Haznawi**, and **Ahmed Alnami**, all hijackers of UA Flight 93, which crashed in Pennsylvania.

Link Analysis—Three Degrees of Separation

- ♦ Wail Alshehri was roommates with and shared a P.O. Box with **Satam Al Suqami**, an AA Flight 11 hijacker.

- ♦ Ahmed Al Haznawi lived with **Ziad Jarrah**, a UA Flight 93 hijacker.¹⁷

Thus, if the government had started with watch list data and pursued links, it is at least possible that all of the hijackers would have been identified as subjects for further investigation. Of course, this example does not show the false positives—names of people with no connection to the terror attacks that might also have been linked to the watch list subjects.

Pattern-based data analysis also has potential for counterterrorism in the longer term, if research on uses of those techniques continues. As will be discussed in more detail in the next section, data-mining research must find ways to identify useful patterns that can predict an extremely rare activity—terrorist planning and attacks.¹⁸ It must also identify how to separate the “signal” of pattern from the “noise” of innocent activity in the data. One possible advantage of pattern-based searches—if they can be perfected—would be that they could provide clues to “sleeper” activity by unknown terrorists who have never engaged in activity that would link them to known terrorists. Unlike subject-based queries, pattern-based searches do not require a link to a known suspicious subject.

Types of pattern-based searches that could prove useful include searches for particular combinations of lower-level activity that together are predictive of terrorist activity. For example, a pattern of a “sleeper” terrorist might be a person in the country on a student visa who purchases a bomb-making book and 50 medium-sized loads of fertilizer. Or, if the concern is that terrorists will use large trucks for attacks, automated data analysis might be conducted regularly to identify people who have rented large trucks, used hotels or drop boxes as addresses, and fall within certain age ranges or have other qualities that are part of a known terrorist pattern. Significant patterns in e-mail traffic might be discovered that could reveal terrorist activity and terrorist “ringleaders.”¹⁹ Pattern-based searches might also be very useful in response and consequence management. For example, searches of hospital data for reports of certain combinations of symptoms, or of other databases for patterns of behavior, such as pharmaceutical purchases or work absenteeism might provide an early signal of a terrorist attack using a biological weapon.²⁰

¹⁷ Zoë Baird et al., *Protecting America's Freedom in the Information Age*, a report of the Markle Foundation Task Force (New York: Markle Foundation, October 2002), p. 28, available at <http://www.markletaskforce.org>. The Markle report uses information drawn from work done by Systems Research and Development; additional information from Jeff Jonas of Systems Research and Development.

¹⁸ See discussion on pages 12 and 13 herein.

¹⁹ Taipale, “Data Mining and Domestic Security,” p. 33, note 120, quoting Hazel Muir, “Email Traffic Patterns Can Reveal Ringleaders,” *New Scientist* (March 27, 2003), available at <http://www.newscientist.com/news/news.jsp?id=ns99993550>.

²⁰ Farzad Mostashari, “Syndromic Surveillance in Practice: New York City,” presentation at CSIS Data Mining Roundtable, Washington, D.C., October 9, 2003. See also Richard Perez-Pena, “An Early Warning System for Diseases in New York,” *New York Times*, April 4, 2003.

III. The Process

Although there are obvious potential benefits of data-mining and automated data-analysis techniques, it is important to have an understanding of the process used in those practices and the risks of error and intrusions on privacy. This section provides a basic description of how these techniques work. The next section describes some of the risks with these processes.

Gathering and Processing the Data

The first step for data mining and data analysis is identifying, gathering, and processing the data that will be analyzed. To do this requires first identifying what the analysis is intended to discover and the type of data that will be useful. This is not always a simple task. For data mining, researchers have developed techniques for “active learning” that can find data that would be useful to collect.²¹ The data-mining process itself will often assist in identifying kinds of data that are not useful. Dr. David Jensen from the University of Massachusetts uses the example of work he conducted in the early 1990s searching for diagnostic rules for Alzheimer’s disease. The goal of the research was to construct accurate diagnostic tools using the answers of patients diagnosed with Alzheimer’s to a long list of interview questions. In addition to coming up with a useful, relatively simple screening tool, the data mining identified interview questions that did not help distinguish between patients with Alzheimer’s and healthy patients. This kind of finding can focus and streamline future data collection.²²

One common myth about data mining and automated data analysis is that they require data to reside in one large database. Typically, data for data mining have been combined into a single database, called a data warehouse or data mart, for mining. There are advantages to this approach—it allows for more efficient searching and for easier standardization and cleansing of the data—but it is not necessary. Data mining can be conducted over a number of databases of varying sizes, provided that certain very low size thresholds are exceeded to provide statistical validity.²³ The same is true for automated data analysis. The Treasury Department’s Financial Crimes Enforcement Network, for example, has conducted data analysis to uncover money-laundering activity using a primary internal data warehouse and a number of secondary databases. When analysis of the primary database, which contains currency transaction reports, indicates the need for additional data, FinCEN analysts seek access to the secondary databases controlled by other entities.²⁴ A more distributed architecture can have some advantages for privacy and database security because it allows different access and privacy standards to be applied to the different databases and also allows for

²¹ Jensen, “Data Mining in Networks,” slide 19.

²² Ibid.

²³ Ibid., slide 18.

²⁴ Ibid.

distributed control of database access, which reduces the opportunities for misuse.²⁵

The final step in this first phase is transforming the data to make them useful. This is often referred to as “data aggregation.” This step involves gathering the data, “cleansing” them to eliminate redundant and other unusable data, and standardizing them to make searches more accurate. When done well, this process has a significant positive impact on the quality of the data-mining or data-analysis product because it reduces data errors such as false positives and false negatives.

One goal of transforming data for data mining is identity resolution—determining whether disparate identity records all represent one individual or different people. Some high-quality practices for cleansing and standardizing identity data have been developed, including “name standardization” and “address hygiene.”²⁶ Name standardization takes name data and recognizes alternate spellings, misspellings, language variations, and nicknames. Many names, like Mohammed, can be spelled a number of different ways. Richard can be Dick, Rich, Ricardo, Ricky, and many more. Name standardization causes Dick, Ricky, and the other alternatives to be considered as Richard, making it possible to match names that might not otherwise appear to be the same.²⁷ Address hygiene performs a similar function for address data. U.S. Postal Service address base files or other data can be used to correct and validate addresses, thus 310 Oak Street might be adjusted to 310 Oak Avenue.²⁸ The more information that is introduced into the process of cleansing and standardizing identity data, the more effective that process becomes. For example, if all you have is three names that are similar, but not identical, you cannot say for sure that they are the same. If for each name you have additional information—social security number, address, or telephone number—you are more likely to be able to resolve whether the names represent the same person. All of this makes the data set far more accurate, which means later data searches will have fewer false-positive and false-negative results.

Commercial data aggregators, like ChoicePoint, Lexis-Nexis, and Axciom, have compiled and aggregated large databases of identity data. These databases are important tools in the identity resolution process. To generate these databases, data aggregators take vast amounts of disparate data, mostly from public records or publicly available sources, and engage in a process in which the data are “gathered, standardized, cleansed, matched, merged, and expressed in a summary form, and [are] periodically monitored and updated.”²⁹ The data include names, addresses, phone numbers, date of birth, height, weight, and social security numbers drawn from various sources.³⁰ As noted above, the more information that is available, the more accurate the identity resolution process becomes. Therefore,

²⁵ Ibid.

²⁶ Jonas, “Using Data to Detect.”

²⁷ Ibid.

²⁸ Ibid.

²⁹ James Zimbardi, “Data Aggregation vs. Data Mining,” presentation at CSIS Data Mining Roundtable, Washington, D.C., July 23, 2003.

³⁰ Ibid.

the sheer volume of the data these aggregators collect (ChoicePoint collects just under 4,000 data sources a month and houses over 100 terabytes of data³¹) allows for a robust process of identity resolution.

Finding Search Models

To conduct an automated data analysis requires a search model. When we are discussing pattern-based searching, finding and perfecting those models can be a very complex and difficult task. There are several ways to come up with the patterns on which a model is based. Models can be found from data-mining analysis, which is a “bottom-up” approach to finding a model in data. That is, it starts with the data and looks for anomalies or patterns that indicate certain behavior. With data mining, the process begins with researchers developing a data-mining algorithm. The algorithm is then applied to “training sets” of data, for which the correct answers are known, to find a model. For example, from a database containing answers to interview questions by people known to have Alzheimer’s disease and others known not to have the disease, a model for Alzheimer’s diagnosis was constructed that could be applied prospectively.³²

“Top-down” data analysis can also be used to find models. This involves starting with a hypothesis about the model and determining whether it exists in data. The hypothesis for a “top-down” analysis might come from an initial “bottom-up” review or from knowledge acquired elsewhere.³³ Expertise or intelligence can be the source of a predictive model that will later be applied to data;³⁴ that is, experts in relevant fields can develop a pattern to use in data-mining analysis.

Whatever method is used to discover them, models must be useful. That is, they must be predictive when applied in real-world situations. In data-mining research, producing blind or poorly designed models that are meaningless is sometimes referred to as “data dredging” or “overfitting the model.”³⁵ A significant amount of data-mining research involves finding ways to avoid trivial, misleading, or irrelevant models. A major goal in research on data mining for counterterrorism, for example, is not only to identify terrorist “signatures,” but also to find ways to separate those patterns of activity from all other “noise” in databases.

Whether they are obtained from data mining or other processes, validating models is critical, and to do this adequately requires conducting real-world testing or realistic simulations using the models. Also, results should be continually traced and analyzed during use to see that they remain valid.³⁶ An acceptable

³¹ Ibid.

³² Jensen, “Data Mining in Networks,” slide 11.

³³ Taipale, “Data Mining and Domestic Security,” p. 30.

³⁴ Ted Senator, “Some Thoughts and Issues Regarding Data Mining,” presentation at CSIS Data Mining Roundtable, Washington, D.C., October 9, 2003.

³⁵ Jensen, “Data Mining in Networks,” slide 12.

³⁶ Senator, “Some Thoughts and Issues.”

model would have low and acceptable numbers of false negatives, while producing manageable false positives that minimally impact the civil liberties of the innocent.

Although automated data analysis using pattern-based predictive models has become relatively common in the private sector, developing these models for counterterrorism presents new and significant challenges for which additional research is necessary. Common commercial models are designed to find patterns that are broadly applicable among data points that are unrelated.³⁷ For example, a retailer will look for broad patterns from unrelated customer purchase data that will predict future customer behavior. This is “propositional” data—that is, data about unrelated instances—from a homogenous database of purchase information. For counterterrorism, on the other hand, the challenge is to find patterns in “relational” data—data in which the key facts are relationships between people, organizations, and activities—from a variety of different types and sources of data. This is because there are no broad patterns of terrorist activity; that is too rare. Terrorists operate in loose networks, and the effective models must find links among lower-level activities, people, organizations, and events that can allow inferences about higher-level clandestine organizations and activities. The data on these lower-level activities exist in different places, and it is the relationships between them that are important.

Looking for ways to develop predictive models for the relational data relevant to counterterrorism was one of the goals of research conducted on data mining for counterterrorism applications in the Evidence Extraction and Link Detection (EELD) program, which predated September 11, but later became part of DARPA’s Information Awareness Office (IAO) research agenda, which included TIA.³⁸ Among the insights from that research are that it is more productive and less prone to error to follow connections from known starting points. If we know, for example, that “good guys” interact with each other a lot, and “bad guys” interact with each other a lot, but good guys and bad guys interact infrequently, we can use this knowledge to inform the data analysis.³⁹ Another method to improve analysis of relational data is combining low-level pattern instances to provide leads for detecting rare high-level patterns. Terrorist plots are rare and difficult to predict reliably, but preparatory and planning activities in which terrorists engage can be identified. Detecting combinations of these low-level activities—such as illegal immigration, operating front businesses, money transfers, use of drop boxes and hotel addresses for commercial activities, and having multiple identities—could help predict terrorist plots.⁴⁰

³⁷ Jensen, “Data Mining in Networks,” slides 21–24; Senator, “Some Thoughts and Issues.”

³⁸ Senator, “Some Thoughts and Issues.”

³⁹ Ibid.; see also, Jensen, “Data Mining in Networks,” discussion of “relational autocorrelation,” slides 24–25.

⁴⁰ Senator, “Some Thoughts and Issues”; Jensen, “Data Mining in Networks,” slide 35.

Decisionmaking

The final stage of the data-mining and data-analysis process involves conducting the searches, interpreting the results, and making decisions about how to use these results. In the context of government use of these techniques for counterterrorism, there are very significant policy issues that arise at this stage. A key issue is the degree to which decisions are made automatically, based on the results of automated data analysis. These techniques are most useful as tools to inform analysis and decisions made by humans, not to substitute for them.

In the commercial realm, some steps are taken automatically—with little or no human intervention—based on results of automated data analysis. For example, a retailer might apply data-mining models to predict the buying interests of a particular shopper based on his past purchases and those of others in the retail database. In that case, an automatic recommendation might be sent to the shopper, without the intervention of an employee. Patterns developed from data mining are also sometimes used to automatically “trigger” creditworthiness decisions.

In most cases, however, data-mining results will be used as “power tools”⁴¹ for humans engaged in analysis or investigation. Certainly in the government, when the stakes of any action taken can be quite high, the results of automated data analysis are most appropriately used to inform human analysis, focus resources, or inspire additional investigation. Indeed, in the complex world of counterterrorism, application of data-mining models and related techniques are likely to be useful at several stages of a multistage process of developing a complete picture out of many “dots.” Analysts might use these techniques to evaluate the significance of leads or suspicions, to generate those leads, to structure or order an investigation, or to acquire additional information along the way.⁴² But they are not likely to be useful as the only source for a decision or conclusion in investigations or analysis.

The decisionmaking stage is also significant because it is where many legal, policy, procedural, or technical controls on the acquisition or use of private information could be imposed, as will be discussed in more detail in sections V and VI. An example of this type of control would be technology that allows access to private information only for certain individuals or after certain permissions have been obtained. Controls might also include a requirement of approval by a neutral third party, based on a standard, before a government employee may obtain private information.

IV. The Risks

U.S. citizens have never wanted their government to know too much about them. Americans are aware of the many actions their government can take based on personal information that can have negative consequences. The government can investigate and conduct surveillance; it can cause significant inconvenience, such

⁴¹ Jensen, “Data Mining in Networks,” slide 39.

⁴² Senator, “Some Thoughts and Issues.”

as when it detains people at airports; it can deny or rescind privileges and liberties, including by detaining and arresting; and by these and other actions it can cause a stigma that results in other consequences such as loss of reputation or livelihood.

The potential benefits of data-mining and automated data-analysis techniques as tools for counterterrorism are significant. Moreover—and this is often misunderstood—these techniques do not permit the government any greater access to personal data; they can only operate on data that the government already has. But they can make private data more useful, and that is what causes controversy. When the government can analyze private data so much more effectively, accessing that data can become more attractive, and the government's power to affect the lives of individuals can increase. There is significant public unease about whether current protections for privacy are adequate to address the potential consequences of this government use.

A fact that heightens concerns about all of the risks described in this section is that there is currently so little public understanding of how these automated data-analysis techniques are used in the government. Nor is there typically public debate or discussion before they are adopted. There are no government-wide standards or guidelines for adoption or use of these techniques. If the government is making judgments on a case-by-case basis about whether automated analysis applications will be sufficiently accurate, whether their use is narrowly tailored to the need, or whether the potential intrusiveness of the practices is justified by their counterterrorism benefits, these deliberations are subject to little public scrutiny. This lack of transparency not only makes the government less accountable and more likely to adopt ill-considered data-analysis practices, but it increases fear and misunderstanding about well-designed and beneficial uses of these techniques.

False Positives

Perhaps the most significant concern with data mining and automated data analysis is that the government might get it wrong and innocent people will be stigmatized as “terrorists” simply because they engaged in unusual patterns of behavior or have some innocent link to a suspected terrorist. A major challenge in the use of these techniques is addressing the possibility of bad data or imperfect search models that result in “false positives.”

If automated data analysis is conducted on vast sets of data gathered from a variety of sources, data quality is inevitably an issue because many records will contain incorrect or obsolete information. If the data are not corrected or “cleansed” before they become the basis for government data analysis, inaccurate or incomplete identification could result. This means either false negatives—a significant security issue—or false positives that incorrectly identify people as matches or links. Even if the data quality is adequate, there is an additional false-positive problem with pattern-based searches: if the data-mining model cannot

separate the “noise” of innocent behavior from the “signal” of terrorist activities, innocent behavior will be viewed as suspicious.

A critical issue is what the government does with false-positive results. If data mining and automated data analysis are used correctly as a “power tools” for analysts and investigators—a way to conduct low-level tasks that will provide clues to assist analysts and investigators—false positives are less dangerous. Data-mining results will then lead only to more analysis or investigation, and false positives can be discovered before there are significant negative consequences for the individual. But the stakes are so high when fighting catastrophic terrorism that there will be great temptation for the government to use these techniques as more than an analytical tool. Government actors will want to take action based on the results of data-analysis queries alone. This action could include detention, arrest, or denial of a benefit. Even if the government later corrects its mistake, the damage to reputation could already be done, with longer-term negative consequences for the individual.

Even when an error is identified, there may be difficulties correcting it. There are often inadequate procedures for correcting watch lists or other similar information. Systems that provide citizens the chance for redress of this kind of error either do not exist or are extremely difficult to use. In addition, if the false-positive search results have been disseminated to other databases, they will be difficult to locate and correct. Although the technology exists to follow inaccurate data and correct cascading occurrences, it has not been a priority, and its implementation lags far behind the technology for collecting and analyzing data.⁴³

Inadequate Government Control of Data

Even if automated data analysis does not result in errors, it will leave private data in the government’s hands. Unfairness can result if the government exercises inadequate control over this type of information. We do not have to look far back in our history for examples of intrusion on liberties that occurred when the U.S. government collected private information without oversight and control. The 1976 Senate Select Committee to Study Government Operations with Respect to Intelligence Activities, chaired by Senator Frank Church (D-Idaho)—the “Church Committee”—investigated domestic intelligence activities of J. Edgar Hoover’s FBI and other U.S. government agencies from the late 1930s through the early 1970s. The Church Committee found that the agencies collected “vast amounts of information about the intimate details of citizens’ lives and about their participation in legal and peaceful activities”⁴⁴ and used that information to abuse the privacy and liberties of U.S. citizens. Most significantly for the current debate, the Church Committee found that pervasive failures in control of private information and lack of accountability for misuse contributed to abuses.

⁴³ Jonas, “Using Data to Detect.”

⁴⁴ Senate Select Committee to Study Governmental Operations with Respect to Intelligence Activities, *Intelligence Activities and the Rights of Americans* [“Church Committee Report”], book 2, 94th Cong., 2nd sess., 1976, p.7.

Information was collected without guidelines or procedural checks; no one monitored the activities of those who collected the information; and information on individuals was disseminated too freely and retained past any point of relevance for national security purposes.⁴⁵

Data-mining and automated data-analysis techniques do not collect private data; they analyze data that is already available. Nonetheless, the power of these tools could mean that the government will collect more private data in order to use them. And if the government exercises inadequate control over who sees that information, for what reasons, how long it is retained, and to whom it is disseminated, unfairness can result. No matter how legitimate the reason for collection or how careful the initial use, information can take on a life of its own if not controlled, and it can be used by others for reasons unrelated to the initial collection. Currently, no government-wide guidelines exist for collection, use, retention, and dissemination of private data, and oversight of these activities is inconsistent at best.

A related concern is “mission creep.” To some degree, there is always some balancing of privacy risks against potential harm to security when making decisions about implementing new, potentially more intrusive technology. A program that uses data mining or automated data analysis might be adopted because it is deemed acceptable given the potential harm of catastrophic terrorism. But there will be great temptation to expand the use of new tools once they have been implemented for one purpose. At any time, another type of illegal behavior could take on a high profile, and authorities will be under pressure to expand the use of these techniques, for example, to help investigate other violent criminals, immigration law violators, or even “deadbeat dads.” It may be that some of these uses are legitimate, but there will be less opportunity for robust public debate on this expanded use.

V. Mitigating Privacy Concerns with Technology

One important avenue for addressing many of the challenges described in the last section, at least in part, is technology. Some technology is already available, and there is much more on which research is ongoing.⁴⁶ Four examples of promising new categories of technology designed to protect privacy and prevent abuse when the government uses large databases of private information are: (1) technology to address inaccurate data and false positives; (2) technology designed to mask or selectively reveal identifying data; (3) audit technology; and (4) rule-processing or permissioning technology. This section will briefly introduce and describe these categories.

⁴⁵ Ibid., pp. 138, 165, 225, 253, 265, 266; see also Mary DeRosa “Privacy in the Age of Terror,” *Washington Quarterly* 26, no. 3 (Summer 2003): 28–30.

⁴⁶ DARPA’s IAO, in connection with its TIA research, was a significant sponsor of this type of research.

Resolving False Positives

Research to solve the problem of false positives is really about perfecting the data-analysis process itself. One cause of false-positive data-analysis results is “bad” or “dirty” data. Technology exists currently that goes a long way toward resolving the problems of bad or incomplete data leading to faulty identification in large data sets. The key here is that more information improves the fidelity of the data. As described in section III, given enough information, “data-cleansing” techniques like name standardization and address hygiene, identity resolution is highly effective.⁴⁷ There can always be improvement, though, and research on data cleansing continues. One of the goals of DARPA’s TIA research was to find ways to increase the accuracy for analysis of nonconforming data from multiple sources.⁴⁸ In addition, large data aggregators work with algorithms that evaluate the historical accuracy of different data sources and use those to “score” the accuracy of a particular identity or other search result.⁴⁹

Eliminating false positives that are generated by a pattern-based data-mining model requires perfecting the model. A model must look for accurate patterns and be able to separate the “signal” of those patterns from the “noise” of innocent transactions in the data. Research on pattern-based data mining for counterterrorism must include model accuracy as a primary goal.

Anonymization

One major area of research and technology development to protect privacy is finding “anonymization” techniques that mask identifying information so that government analysts can conduct searches of data and share the data without accessing identities. This is also sometimes referred to as “selective revelation” because the research looks for methods to anonymize data and reveal the identifying information only when authorized—and gradually, as suspicion increases. To be useful, these tools must be effective at masking—which involves more than removing names, social security numbers, and other obvious identifiers—while maintaining the usefulness of the search.

The first step in anonymization is identifying what needs to be masked. This includes identifiers, such as names, addresses, social security numbers, credit card numbers, phone numbers, vehicle license numbers, etc. But even when these identifiers are withheld, analysts and systems can infer identity from collections of less-sensitive data.⁵⁰ For example, identities of 87.1 percent of individuals in the United States can be inferred by knowing only a date of birth, gender, and zip

⁴⁷ Jonas, “Using Data to Detect.” Jeff Jonas described it as a “solved problem.”

⁴⁸ DARPA, “Report to Congress regarding the Terrorism Information Awareness Programs,” (May 20, 2003), available at <http://www.darpa.mil/body/tia/TIA%20ES.pdf>.

⁴⁹ Zimbardi, “Data Aggregation.”

⁵⁰ Teresa Lunt, “Protecting Privacy in Terrorist Tracking Applications,” presentation at CSIS Data Mining Roundtable, Washington, D.C., September 16, 2003.

code.⁵¹ A challenge with anonymization research is finding the data that allows an inference of identity and controlling that inference.⁵²

Latanya Sweeney of Carnegie Mellon University has developed a privacy-protection model known as “K-anonymity.” This model ensures that no release of data will allow a person to be distinguished from fewer than $k-1$ other individuals. The value of “ k ” is a policy question. She does this by looking at statistical distribution frequencies for the data set (i.e., how often does someone have the same date of birth, zip code, and gender, in a particular geographic area); then she starts hiding parts of the value (e.g., changing a zip code of 20008 to 2000* or replacing a date of birth digit with *, or both). This is done until the remaining data cannot be used to infer identities or “de-anonymize” less than a designated population of prospects. For example, if $K=1000$, the data will identify no fewer than 1000 possible candidates.⁵³

Teresa Lunt from the Palo Alto Research Center is developing mechanisms for inference control as part of her research into a “privacy appliance” designed to allow authorized analysts to search for terrorist-related activity while providing a realistic degree of privacy protection for the data of ordinary citizens. (The privacy appliance would also include rule-based access control and audit mechanisms.) This privacy appliance would be under the control of the data owners. It would analyze data for potential undesired identifying inferences and store the results so that only authorized users could access them. Who is authorized and what kind of process they must follow to access the identifying information would be a policy question.⁵⁴

Other approaches would allow anonymous searching and matching—such as identity resolution—over a number of separate databases by using cryptology or other methods. Jeff Jonas of SRD is developing an approach to this using “one-way hashes.” With one-way hashes, data are converted at their source to uniform formats, then “hashed”—made completely indistinguishable and irreversible—essentially becoming unique digital signatures. A search using that unique identifier can then determine whether a piece of data matches others in other databases. The hash cannot be reversed, but if a match is discovered steps can be taken—and what those steps are is a policy matter—to reveal the identity to an authorized person. This method allows the release of the actual data to be governed by the party that owns the data; the analyst must make an individual information request to the data owner to see any of his data.⁵⁵

⁵¹ Latanya Sweeney, untitled presentation at CSIS Data Mining Roundtable, Washington, D.C., October 9, 2003.

⁵² Lunt, “Protecting Privacy.”

⁵³ Sweeney, untitled presentation at CSIS, October 9, 2003.

⁵⁴ Lunt, “Protecting Privacy.”

⁵⁵ Jonas, “Using Data to Detect.”

Audit Technology

Access control using anonymization procedures will not protect against authorized users who use legitimate access improperly. Audit technology records activity in databases and on networks and allows the government to see who is conducting searches, what kinds of searches, and how often. Audit technology can be used to “watch the watchers.” It would support a policy of overseeing the actions of government employees who use data mining and other technology tools to access private information.

With strong audit technology that maintains a record of queries, government overseers can be alerted to ongoing improper activity, can track activity after a mistake or abuse has been identified, and can conduct random checks. If audits are to catch ongoing improper activity, someone must be watching the audit trail that is being created. This can be a massive task, and policy and technology must be developed to assist in it.⁵⁶

Care must be taken with audit trails. Audit data can be very sensitive information because they can contain private information and because information about the searches intelligence analysts conduct can reveal sensitive intelligence information. Audit data will no doubt be voluminous, and access to these data must be restricted. Techniques such as separating audit data into “shares” controlled by different entities, none of which alone has any useful information,⁵⁷ or other types of encryption must be developed to protect audit information.

Rule-based Processing

The idea of rule-based processing as a privacy-protection tool is that certain policy rules can be built into search queries so that data can only be retrieved consistent with those rules. Rule-based processing has two elements. First, the query must carry with it information about the types of permission the user has. For example, a query might indicate that it is pursuant to a search warrant, which would allow it to retrieve certain kinds of data that would be unavailable without a warrant.⁵⁸ Second, data must be labeled with information about how they may be accessed. Data items might be labeled with “meta data”—data that summarize or describe the qualities of the data—that indicates how the data can be processed. This meta-data label would travel with the data and guide access to them wherever they reside. The meta data for a particular data item might, for example, indicate whether it identifies an American or a foreign person, and access can be controlled accordingly.⁵⁹

⁵⁶ Information Sciences and Technologies Study Group (ISAT), “Security with Privacy: ISAT 2002 Study,” (December 13, 2002), p. 13, available at http://www.epic.org/privacy/profiling/tia/isat_study.pdf.

⁵⁷ Lunt, “Protecting Privacy.”

⁵⁸ Taipale, “Data Mining and Domestic Security,” pp. 75, 76.

⁵⁹ Ibid.

Data labeling is not a new idea, and it is being used in many new systems, including some digital rights management (DRM) systems. There are many challenges with this kind of technology, however. Accuracy of data labeling is dependent on the accuracy of the data. Handling existing, unlabeled data in “legacy” databases will be a significant problem. And when data are combined to produce “derived data,” it is unclear how they should be labeled.⁶⁰

Finally, just what rules should be enforced through rule-based processing is an extraordinarily difficult policy question. Privacy policy is notoriously confusing, and getting it wrong could have significant consequences. Before this kind of technology can be deployed effectively, some of the policy issues in the next section will need to be resolved.

VI. Areas for Policy Development

Data-mining and automated data-analysis techniques are valuable tools in the fight against terrorism, and the U.S. government must have them at its disposal. Their use also poses risks to individual privacy and due process. Although some of these risks can be addressed with the technology described in the previous section, that technology alone will be inadequate. Therefore, policy action is required to ensure that controls and protections accompany the use of these powerful tools. This complex area has received too little detailed attention from and discussion by policymakers. More debate is necessary; this section identifies issues that the debate must address.

Data-mining Research

In discussing policy on data mining, it is first critical to distinguish between research and application of the technology. These tools have great potential, but to realize that potential fully more research is needed. The government should support research on data mining and related tools for counterterrorism. But even research causes concern if the public and Congress perceive it to be structured in a way that is insensitive to privacy issues. The experience with Congress’s termination of DARPA’s TIA research—which was due in part to an initial lack of clarity about how privacy issues would be addressed—highlights the need for a clear and consistent government policy on data-mining research. That policy must address how private data may be used in research. It should also take into account the context in which these tools may eventually be deployed and the privacy issues they raise.

A government policy for data-mining research should require any research program to identify potential privacy concerns. The policy should support research on privacy-protecting technologies, like the ones described in the previous section, as part of data-mining research. To be sure that research programs are able to address privacy concerns, a government policy on research should require some analysis of these privacy issues as part of a research program.

⁶⁰ ISAT, “Security with Privacy,” p. 16.

This sort of policy analysis, which does not ordinarily take place at the research stage, could in this case provide the kind of context and understanding that is necessary for researchers to move forward in the right direction. Recently, the Congressional Intelligence Committee conferees adopted such an approach.⁶¹ While expressing support for certain data-mining and data-analysis research, the conferees directed the attorney general and the director of central intelligence to produce a report to the Intelligence Committees “regarding the applications of the Constitution, laws, regulations, Executive Orders, and guidelines of the United States to the use of these advanced analytic tools by the Intelligence Community.”⁶²

Clarity about Use of Data Mining and Data Analysis

When it comes to application of these automated data-analysis tools, the first crucial step, again, is for the government to adopt a clearly articulated policy. One of the principal reasons for public concern about these tools is that there appears to be no consistent policy guiding decisions about when and how to use them. Indeed, use of the tools appears to be ad hoc, with each entity making its own decisions about implementation based on its own criteria. There is no place to go—outside or inside of the government—for information about how the techniques are being used. From time to time information comes out about some use of private data for automated data analysis, such as recent revelations about Jet Blue⁶³ and Northwest Airlines⁶⁴ providing passenger data to government agencies or their contractors. The troubled public reaction to these revelations is as much about the concern that this might be the “tip of the iceberg” as it is about the particular privacy invasion. These concerns will continue to generate public and congressional backlash unless they are addressed.

A policy about use of data-mining and automated data-analysis techniques should set forth standards for decisionmaking on the type of data-analysis technique to use and the data that will be accessed. It should require an inquiry into data accuracy and the level of errors—false positives and false negatives—that analysis is expected to generate, and it should provide for some mechanism for correcting errors.

A government-wide policy for use of data-mining and data-analysis techniques must also clarify the process of decisionmaking. Although there will be variations among agencies, decisions about use of these practices should require some kind of senior-level notice and approval. There may even be a place for some more-central decisionmaker on adoption of the more controversial

⁶¹ *Intelligence Authorization Act for Fiscal Year 2004*, HR 2417, 108th Cong., 1st sess., (November 19, 2003), Title V, 108–381, available at <http://thomas.loc.gov/cgi-bin/cpquery/T?&report=hr381&dbname=cp108&>.

⁶² *Ibid.*

⁶³ Philip Shenon, “Airline Gave Defense Firm Passenger Files,” *New York Times*, September 20, 2003.

⁶⁴ Matthew L. Wald, “Airline Gave Government Information on Passengers,” *New York Times*, January 18, 2004.

techniques such as complex pattern-based searching. Some have recommended creation of a high-level group or organization, the function of which would be to develop and implement policy guidelines for data-mining and automated data-analysis tools.⁶⁵

Use of Search Results

Consistent policy is also necessary on what action can be taken based on search results once they are obtained. As discussed in section III, data mining and data analysis should be used as “power tools” for analysts and investigators—a way to conduct low-level tasks that will provide leads to assist analysts and investigators. These techniques are not effective at providing answers to the ultimate questions in an analysis or investigation. A policy that allows automated data-analysis results to be used only to further analysis or investigation, not as the sole basis for government action, would avoid many of the possible negative consequences for individuals from data mining because fewer mistakes would be made based on false-positive results. If there are circumstances under which data-mining results should be used as the basis for action, guidance on those circumstances should be generated and articulated clearly to employees who are making these decisions.

Controls on the Use of Identifying Information

The controversy about data mining and related techniques is really about how the government will control private data, whatever its source. Currently, no clear guidance exists for government entities and employees about how to handle this sensitive data, and this lack of direction can cause mistakes and ad hoc, inconsistent use of the data. Perhaps the most important step to address these concerns is for the executive branch to implement clear guidelines generally for employees on how they may access, use, retain, and disseminate private data.

The most recent report of the Markle Task Force on National Security in the Information Age, *Creating a Trusted Network for Homeland Security*, includes recommendations for guidelines on government use of private information.⁶⁶ The task force advocates guidelines that address acquisition and use of private data, retention, and dissemination for reasons other than terrorism. It suggests the government start by identifying the types of information it will need for counterterrorism purposes, and then decide on the appropriate levels of protection for each type of data and what kinds of standards and procedures will provide that protection.⁶⁷ The answers to all of these questions will inevitably be controversial, and the government will benefit greatly in terms of public understanding and acceptance if there is open debate about the contents of the guidelines.

⁶⁵ Rosenzweig, “Proposals for Implementing the Terrorism Information Awareness System,” pp. 20, 21.

⁶⁶ Zoë Baird et al., *Creating a Trusted Network for Homeland Security*, second report of the Markle Foundation Task Force (New York: Markle Foundation, December 2002), pp. 30–38, available at <http://www.markletaskforce.org>.

⁶⁷ Ibid.

Development of these policies is critical also if the privacy-protection technology discussed in the previous section is to be developed and adopted. As discussed in section V, there is significant research being conducted on anonymization and selective revelation techniques that would mask identities in data and allow government employees access to those data only under certain circumstances.⁶⁸ But before this technology can be used, policymakers must decide what those circumstances are. The answers will no doubt be different depending on the types of searches employed, the reasons for the search, and the legal landscape in which the searches operate (law enforcement versus intelligence, for example). Decisions must be made about the appropriate standards for acquiring different kinds of private information and the level of approval necessary, if any.⁶⁹ Will the line employee decide whether he or she can access an identity? A senior officer? The attorney general? A judge? Before technology can be used to help enforce rules, the rules must be understood.

Finally, guidelines and supporting technology, including audit technology, will not work unless there is a vigorous system of oversight, including regular audits and accountability for wrongdoing. Currently, oversight in the executive branch tends to focus more on after-the-fact investigation than on auditing and ongoing control of the use of private information. With access to private information increasing significantly, there is a need for more systemic, ongoing auditing of employee actions. As discussed in the previous section, new technology can significantly enhance the government's ability to watch those who access private data. But human beings must use these audit tools, and they must do so pursuant to clear policies and practices. There have been some recent innovations in this area, such as the establishment of a privacy officer at the new Department of Homeland Security, but there has been no systematic review or reevaluation, in the Congress or the executive branch, of the process for oversight.

Conclusion

Defeating terrorism requires a more nimble intelligence apparatus that operates more actively within the United States and makes use of advanced information technology to a degree unknown during the Cold War. Data-mining and automated data-analysis techniques are powerful tools for intelligence and law enforcement officials fighting terrorism. But privacy concerns with the use of these tools have generated significant fear and controversy. These tools are too valuable to be rejected outright. On the other hand, embracing them without any guidelines or controls for their use poses a great risk that they, and the private information they analyze, will be misused. Policymakers must acquire a greater understanding of data-mining and automated data-analysis tools and craft policy that encourages responsible use and sets parameters for that use.

⁶⁸ ISAT, "Security with Privacy," pp. 14–18.

⁶⁹ Baird et al., *Creating a Trusted Network*, pp. 30–38.

About the Author

Mary DeRosa joined CSIS in February 2002 as a senior fellow in the Technology and Public Policy Program. Previously, she served on the National Security Council staff (1997–2001) as special assistant to the president and legal adviser and, earlier, as deputy legal adviser. From 1995 to 1997, she was special counsel to the general counsel at the Department of Defense, and in 1994 she was an attorney for the Advisory Board on the Investigative Capability of the Department of Defense. Before joining the government, Ms. DeRosa was a lawyer in private practice in the Washington, D.C., and Los Angeles offices of Arnold & Porter. She was a law clerk to the Honorable Richard J. Cardamone on the U.S. Court of Appeals for the Second Circuit and is a graduate of the George Washington University Law School and the University of Virginia.