Order Code RL31798

CRS Report for Congress
Received through the CRS Web

Data Mining and Homeland Security: An Overview

Updated January 27, 2006

Jeffrey W. Seifert
Specialist in Information Science and Technology Policy
Resources, Science, and Industry Division

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **27 JAN 2006**	2. REPORT TYPE **N/A**	3. DATES COVERED **-**

4. TITLE AND SUBTITLE
Data Mining and Homeland Security: An Overview

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Federation of American Scientists 1717 K St, NW Suite 209 Washington, DC 20036

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**SAR**	**27**	

Data Mining and Homeland Security: An Overview

Summary

Data mining has become one of the key features of many homeland security initiatives. Often used as a means for detecting fraud, assessing risk, and product retailing, data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. In the context of homeland security, data mining can be a potential means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records.

While data mining represents a significant advance in the type of analytical tools currently available, there are limitations to its capability. One limitation is that although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. A second limitation is that while data mining can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. Successful data mining still requires skilled technical and analytical specialists who can structure the analysis and interpret the output.

Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance. However, some of the homeland security data mining applications represent a significant expansion in the quantity and scope of data to be analyzed. Some efforts that have attracted a higher level of congressional interest include the Terrorism Information Awareness (TIA) project (now-discontinued) and the Computer-Assisted Passenger Prescreening System II (CAPPS II) project (now-canceled and replaced by Secure Flight). Other initiatives that have been the subject of recent congressional interest include the Multi-State Anti-Terrorism Information Exchange (MATRIX), the Able Danger program and data collection and analysis projects being conducted by the National Security Agency (NSA).

As with other aspects of data mining, while technological capabilities are important, there are other implementation and oversight issues that can influence the success of a project's outcome. One issue is data quality, which refers to the accuracy and completeness of the data being analyzed. A second issue is the interoperability of the data mining software and databases being used by different agencies. A third issue is mission creep, or the use of data for purposes other than for which the data were originally collected. A fourth issue is privacy. Questions that may be considered include the degree to which government agencies should use and mix commercial data with government data, whether data sources are being used for purposes other than those for which they were originally designed, and possible application of the Privacy Act to these initiatives. It is anticipated that congressional oversight of data mining projects will grow as data mining efforts continue to evolve. This report will be updated as events warrant.

Contents

Data Mining and Homeland Security: An Overview

What is Data Mining?

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets.[1] 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.

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can use a variety of parameters to examine the data. They include association (patterns where one event is connected to another event, such as purchasing a pen and purchasing paper), sequence or path analysis (patterns where one event leads to another event, such as the birth of a child and purchasing diapers), classification (identification of new patterns, such as coincidences between duct tape purchases and plastic sheeting purchases), clustering (finding and visually documenting groups of previously unknown facts, such as geographic location and brand preferences), and forecasting (discovering patterns from which one can make reasonable predictions regarding future activities, such as the prediction that people who join an athletic club may take exercise classes).[2]

As an application, compared to other data analysis applications, such as structured queries (used in many commercial databases) or statistical analysis software, data mining represents a *difference of kind rather than degree*. Many simpler analytical tools utilize a verification-based approach, where the user develops a hypothesis and then tests the data to prove or disprove the hypothesis. For example, a user might hypothesize that a customer who buys a hammer, will also buy a box of nails. The effectiveness of this approach can be limited by the creativity of the user to develop various hypotheses, as well as the structure of the software being used. In contrast, data mining utilizes a discovery approach, in which algorithms can be used to examine several multidimensional data relationships simultaneously, identifying those that are unique or frequently represented. For example, a hardware store may compare their customers' tool purchases with home ownership, type of

[1] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999); Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996).

[2] For a more technically-oriented definition of data mining, see [http://searchcrm .techtarget.com/gDefinition/0,294236,sid11_gci211901,00.html].

automobile driven, age, occupation, income, and/or distance between residence and the store. As a result of its complex capabilities, two precursors are important for a successful data mining exercise; a clear formulation of the problem to be solved, and access to the relevant data.[3]

Reflecting this conceptualization of data mining, some observers consider data mining to be just one step in a larger process known as knowledge discovery in databases (KDD). Other steps in the KDD process, in progressive order, include data cleaning, data integration, data selection, data transformation, (data mining), pattern evaluation, and knowledge presentation.[4]

A number of advances in technology and business processes have contributed to a growing interest in data mining in both the public and private sectors. Some of these changes include the growth of computer networks, which can be used to connect databases; the development of enhanced search-related techniques such as neural networks and advanced algorithms; the spread of the client/server computing model, allowing users to access centralized data resources from the desktop; and an increased ability to combine data from disparate sources into a single searchable source.[5]

In addition to these improved data management tools, the increased availability of information and the decreasing costs of storing it have also played a role. Over the past several years there has been a rapid increase in the volume of information collected and stored, with some observers suggesting that the quantity of the world's data approximately doubles every year.[6] At the same time, the costs of data storage have decreased significantly from dollars per megabyte to pennies per megabyte. Similarly, computing power has continued to double every 18-24 months, while the relative cost of computing power has continued to decrease.[7]

Data mining has become increasingly common in both the public and private sectors. Organizations use data mining as a tool to survey customer information, reduce fraud and waste, and assist in medical research. However, the proliferation of data mining has raised some implementation and oversight issues as well. These include concerns about the quality of the data being analyzed, the interoperability of the databases and software between agencies, and potential infringements on privacy. Also, there are some concerns that the limitations of data mining are being overlooked as agencies work to emphasize their homeland security initiatives.

[3] John Makulowich, "Government Data Mining Systems Defy Definition," *Washington Technology*, 22 February 1999, [http://www.washingtontechnology.com/news/13_22/tech_features/393-3.html].

[4] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques* (New York: Morgan Kaufmann Publishers, 2001), p. 7.

[5] Pieter Adriaans and Dolf Zantinge, *Data Mining* (New York: Addison Wesley, 1996), pp. 5-6.

[6] Ibid., p. 2.

[7] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999), p. 4.

Limitations of Data Mining

While data mining products can be very powerful tools, they are not self-sufficient applications. To be successful, data mining requires skilled technical and analytical specialists who can structure the analysis and interpret the output that is created. Consequently, the limitations of data mining are primarily data or personnel-related, rather than technology-related.[8]

Although data mining can help reveal patterns and relationships, it does not tell the user the value or significance of these patterns. These types of determinations must be made by the user. Similarly, the validity of the patterns discovered is dependent on how they compare to "real world" circumstances. For example, to assess the validity of a data mining application designed to identify potential terrorist suspects in a large pool of individuals, the user may test the model using data that includes information about known terrorists. However, while possibly re-affirming a particular profile, it does not necessarily mean that the application will identify a suspect whose behavior significantly deviates from the original model.

Another limitation of data mining is that while it can identify connections between behaviors and/or variables, it does not necessarily identify a causal relationship. For example, an application may identify that a pattern of behavior, such as the propensity to purchase airline tickets just shortly before the flight is scheduled to depart, is related to characteristics such as income, level of education, and Internet use. However, that does not necessarily indicate that the ticket purchasing behavior is caused by one or more of these variables. In fact, the individual's behavior could be affected by some additional variable(s) such as occupation (the need to make trips on short notice), family status (a sick relative needing care), or a hobby (taking advantage of last minute discounts to visit new destinations).[9]

Data Mining Uses

Data mining is used for a variety of purposes in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For example, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. The medical community sometimes uses data mining to help predict the effectiveness of a procedure or medicine. Pharmaceutical firms use data mining of chemical compounds and genetic material to help guide research on new treatments for diseases. Retailers can use information collected through affinity programs (e.g., shoppers' club cards, frequent flyer points, contests) to assess the effectiveness of product selection and placement decisions, coupon offers, and which

[8] Ibid., p. 2.

[9] Ibid., p. 1.

products are often purchased together. Companies such as telephone service providers and music clubs can use data mining to create a "churn analysis," to assess which customers are likely to remain as subscribers and which ones are likely to switch to a competitor.[10]

In the public sector, data mining applications were initially used as a means to detect fraud and waste, but they have grown also to be used for purposes such as measuring and improving program performance. It has been reported that data mining has helped the federal government recover millions of dollars in fraudulent Medicare payments.[11] The Justice Department has been able to use data mining to assess crime patterns and adjust resource allotments accordingly. Similarly, the Department of Veterans Affairs has used data mining to help predict demographic changes in the constituency it serves so that it can better estimate its budgetary needs. Another example is the Federal Aviation Administration, which uses data mining to review plane crash data to recognize common defects and recommend precautionary measures.[12]

Recently, data mining has been increasingly cited as an important tool for homeland security efforts. Some observers suggest that data mining should be used as a means to identify terrorist activities, such as money transfers and communications, and to identify and track individual terrorists themselves, such as through travel and immigration records. Initiatives that have attracted significant attention include the now-discontinued Terrorism Information Awareness (TIA) project[13] conducted by the Defense Advanced Research Projects Agency (DARPA), and the now-canceled Computer-Assisted Passenger Prescreening System II (CAPPS II) that was being developed by the Transportation Security Administration (TSA). CAPPS II is being replaced by a new program called Secure Flight. Other initiatives that have recently been the subject of congressional interest include the Able Danger

[10] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery, Third Edition* (Potomac, MD: Two Crows Corporation, 1999), p. 5; Patrick Dillon, *Data Mining: Transforming Business Data Into Competitive Advantage and Intellectual Capital* (Atlanta GA: The Information Management Forum, 1998), pp. 5-6.

[11] George Cahlink, "Data Mining Taps the Trends," *Government Executive Magazine*, October 1, 2000, [http://www.govexec.com/tech/articles/1000managetech.htm].

[12] Ibid.; for a more detailed review of the purpose for data mining conducted by federal departments and agencies, see U.S. General Accounting Office, *Data Mining: Federal Efforts Cover a Wide Range of Uses,* GAO Report GAO-04-548 (Washington: May 2004).

[13] This project was originally identified as the Total Information Awareness project until DARPA publicly renamed it the Terrorism Information Awareness project in May 2003.

Section 8131 of the FY2004 Department of Defense Appropriations Act (P.L. 108-87) prohibited further funding of TIA as a whole, while allowing unspecified subcomponents of the TIA initiative to be funded as part of DOD's classified budget, subject to the provisions of the National Foreign Intelligence Program, which restricts the processing and analysis of information on U.S. citizens. For further details regarding this provision, see CRS Report RL31805 *Authorization and Appropriations for FY2004: Defense*, by Amy Belasco and Stephen Daggett.

program and data collection and analysis projects being conducted by the National Security Agency (NSA).

Terrorism Information Awareness (TIA) Program

In the immediate aftermath of the September 11, 2001, terrorist attacks, many questions were raised about the country's intelligence tools and capabilities, as well as the government's ability to detect other so-called "sleeper cells," if, indeed, they existed. One response to these concerns was the creation of the Information Awareness Office (IAO) at the Defense Advanced Research Projects Agency (DARPA)[14] in January 2002. The role of IAO was "in part to bring together, under the leadership of one technical office director, several existing DARPA programs focused on applying information technology to combat terrorist threats."[15] The mission statement for IAO suggested that the emphasis on these technology programs was to "counter asymmetric threats by achieving *total information awareness* useful for preemption, national security warning, and national security decision making."[16] To that end, the TIA project was to focus on three specific areas of research, anticipated to be conducted over five years, to develop technologies that would assist in the detection of terrorist groups planning attacks against American interests, both inside and outside the country. The three areas of research and their purposes were described in a DOD Inspector General report as:

> "… language translation, data search with pattern recognition and privacy protection, and advanced collaborative and decision support tools. Language translation technology would enable the rapid analysis of foreign languages, both spoken and written, and allow analysts to quickly search the translated materials for clues about emerging threats. The data search, pattern recognition, and privacy protection technologies would permit analysts to search vast quantities of data for patterns that suggest terrorist activity while at the same time controlling access to the data, enforcing laws and policies, and ensuring detection of misuse of the information obtained. The collaborative reasoning and decision support technologies would allow analysts from different agencies to share data."[17]

[14] DARPA "is the central research and development organization for the Department of Defense (DOD)" that engages in basic and applied research, with a specific focus on "research and technology where risk and payoff are both very high and where success may provide dramatic advances for traditional military roles and missions." [http://www.darpa.mil/]

[15] Department of Defense. 20 May 2003. *Report to Congress Regarding the Terrorism Information Awareness Program, Executive Summary*, p.2.

[16] Department of Defense. 20 May 2003. *Report to Congress Regarding the Terrorism Information Awareness Program, Detailed Information*, p.1 (emphasis added).

[17] Department of Defense, Office of the Inspector General. 12 December 2003. *Information Technology Management: Terrorism Information Awareness Project (D2004-033)*. P. 7.

Each part had the potential to improve the data mining capabilities of agencies that adopt the technology.[18] Automated rapid language translation could allow analysts to search and monitor foreign language documents and transmissions more quickly than currently possible. Improved search and pattern recognition technologies may enable more comprehensive and thorough mining of transactional data, such as passport and visa applications, car rentals, driver license renewals, criminal records, and airline ticket purchases. Improved collaboration and decision support tools might facilitate the search and coordination activities being conducted by different agencies and levels of government.[19]

In public statements DARPA frequently referred to the TIA program as a research and development project designed to create experimental prototype tools, and that the research agency would only use "data that is legally available and obtainable by the U.S. Government."[20] DARPA further emphasized that these tools could be adopted and used by *other* agencies, and that DARPA itself would not be engaging in any actual-use data mining applications, although it could "support production of a scalable leave-behind system prototype."[21] In addition, some of the technology projects being carried out in association with the TIA program did not involve data mining.[22] However, the TIA program's overall emphasis on collecting, tracking, and analyzing data trails left by individuals served to generate significant and vocal opposition soon after John Poindexter made a presentation on TIA at the DARPATech 2002 Conference in August 2002.[23]

Critics of the TIA program were further incensed by two administrative aspects of the project. The first involved the Director of IAO, Dr. John M. Poindexter. Poindexter, a retired Admiral, was, until that time, perhaps most well-known for his alleged role in the Iran-contra scandal during the Reagan Administration. His

[18] It is important to note that while DARPA's mission is to conduct research and development on technologies that can be used to address national-level problems, it would not be responsible for the operation of TIA, if it were to be adopted.

[19] For more details about the Terrorism Information Awareness program and related information and privacy laws, see CRS Report RL31730, *Privacy: Total Information Awareness Programs and Related Information Access, Collection, and Protection Laws*, by Gina Marie Stevens, and CRS Report RL31786, *Total Information Awareness Programs: Funding, Composition, and Oversight Issues*, by Amy Belasco.

[20] Department of Defense, DARPA, "Defense Advanced Research Project Agency's Information Awareness Office and Total Information Awareness Project," p. 1, [http://www.iwar.org.uk/news-archive/tia/iaotia.pdf].

[21] Ibid., p. 2.

[22] Although most of the TIA-related projects did involve some form of data collection, the primary purposes of some of these projects, such as war gaming, language translation, and biological agent detection, were less connected to data mining activities. For a description of these projects, see [http://www.fas.org/irp/agency/dod/poindexter.html].

[23] The text of Poindexter's presentation is available at [http://www.darpa.mil/DARPATech2002/presentations/iao_pdf/speeches/POINDEXT.pdf]. The slide presentation of Poindexter's presentation is available at [http://www.darpa.mil/DARPATech2002/presentations/iao_pdf/slides/PoindexterIAO.pdf].

involvement with the program caused many in the civil liberties community to question the true motives behind TIA.[24] The second source of contention involved TIA's original logo, which depicted an "all-seeing" eye atop of a pyramid looking down over the globe, accompanied by the Latin phrase *scientia est potentia* (knowledge is power).[25] Although DARPA eventually removed the logo from its website, it left a lasting impression.

The continued negative publicity surrounding the TIA program contributed to the introduction of a number of bills in Congress that eventually led to the program's dissolution. Among these bills was S. 188, the Data-Mining Moratorium Act of 2003, which, if passed, would have imposed a moratorium on the implementation of data mining under the TIA program by the Department of Defense, as well as any similar program by the Department of Homeland Security. An amendment included in the Omnibus Appropriations Act for Fiscal Year 2003 (P.L. 108-7) required the Director of Central Intelligence, the Secretary of Defense, and the Attorney General to submit a joint report to Congress within 90 days providing details about the TIA program.[26] Funding for TIA as a whole was prohibited with the passage of the FY2004 Department of Defense Appropriations Act (P.L. 108-87) in September 2003. However, Section 8131 of the law allowed unspecified subcomponents of the TIA initiative to be funded as part of DOD's classified budget, subject to the provisions of the National Foreign Intelligence Program, which restricts the processing and analysis of information on U.S. citizens.[27]

Computer-Assisted Passenger Prescreening System (CAPPS II)

Similar to TIA, the CAPPS II project represented a direct response to the September 11, 2001, terrorist attacks. With the images of airliners flying into buildings fresh in people's minds, air travel was now widely viewed not only as a critically vulnerable terrorist target, but also as a weapon for inflicting larger harm. The CAPPS II initiative was intended to replace the original CAPPS, currently being used. Spurred, in part, by the growing number of airplane bombings, the existing CAPPS (originally called CAPS) was developed through a grant provided by the

[24] Shane Harris, "Counterterrorism Project Assailed By Lawmakers, Privacy Advocates," *Government Executive Magazine,* 25 November 2002, [http://www.govexec.com /dailyfed/1102/112502h1.htm].

[25] The original logo can be found at [http://www.thememoryhole.org/policestate/iao-logo.htm].

[26] The report is available at [http://www.eff.org/Privacy/TIA/TIA-report.pdf]. Some of the information required includes spending schedules, likely effectiveness of the program, likely impact on privacy and civil liberties, and any laws and regulations that may need to be changed to fully deploy TIA. If the report had not submitted within 90 days, funding for the TIA program could have been discontinued. For more details regarding this amendment, see CRS Report RL31786, *Total Information Awareness Programs: Funding, Composition, and Oversight Issues,* by Amy Belasco.

[27] For further details regarding this provision, see CRS Report RL31805 *Authorization and Appropriations for FY2004: Defense,* by Amy Belasco and Stephen Daggett.

Federal Aviation Administration (FAA) to Northwest Airlines, with a prototype system tested in 1996. In 1997, other major carriers also began work on screening systems, and, by 1998, most of the U.S.-based airlines had voluntarily implemented CAPS, with the remaining few working toward implementation.[28] Also, during this time, the White House Commission on Aviation Safety and Security (sometimes referred to as the Gore Commission) released its final report in February 1997.[29] Included in the commission's report was a recommendation that the United States implement automated passenger profiling for its airports.[30] On April 19, 1999, the FAA issued a notice of proposed rulemaking (NPRM) regarding the security of checked baggage on flights within the United States (docket no. FAA-1999-5536).[31] As part of this still-pending rule, domestic flights would be required to utilize "the FAA-approved computer-assisted passenger screening (CAPS) system to select passengers whose checked baggage must be subjected to additional security measures."[32]

The current CAPPS system is a rule-based system that uses the information provided by the passenger when purchasing the ticket to determine if the passenger fits into one of two categories; "selectees" requiring additional security screening, and those who do not. CAPPS also compares the passenger name to those on a list of known or suspected terrorists.[33] CAPPS II was described by TSA as "an enhanced system to confirm the identities of passengers and to identify foreign terrorists or persons with terrorist connections before they can board U.S. aircraft."[34] CAPPS II would have sent information provided by the passenger in the passengers name record (PNR), including full name, address, phone number, and date of birth, to commercial data providers for comparison to authenticate the identity of the passenger. The commercial data provider would have then transmitted a numerical score back to TSA indicating a particular risk level.[35] Passengers with a "green" score would have undergone "normal screening," while passengers with a "yellow" score would have undergone additional screening. Passengers with a "red" score

[28] Department of Transportation, *White House Commission on Aviation and Security: The DOT Status Report*, February 1998, [http://www.dot.gov/affairs/whcoasas.htm].

[29] The Gore Commission was established by Executive Order 13015 on August 22, 1996, following the crash of TWA flight 800 in July 1996.

[30] White House Commission on Aviation Safety and Security: Final Report to President Clinton. 12 February 1997. [http://www.fas.org/irp/threat/212fin~1.html].

[31] The docket can be found online at [http://dms.dot.gov/search/document.cfm?documentid=57279&docketid=5536].

[32] *Federal Register*, 64 (April 19,1999): 19220.

[33] U.S. General Accounting Office, *Aviation Security: Computer-Assisted Passenger Prescreening System Faces Significant Implementation Challenges*, GAO Report GAO-04-385, February 2004, pp. 5-6.

[34] Transportation Security Administration, "TSA's CAPPS II Gives Equal Weight to Privacy, Security," Press Release, 11 March 2003, [http://www.tsa.gov/public/display?theme=44&content=535].

[35] Robert O'Harrow, Jr., "Aviation ID System Stirs Doubt," *Washington Post*, 14 March 2003, p. A16.

would not have been allowed to board the flight, and would have received "the attention of law enforcement."[36] While drawing on information from commercial databases, TSA had stated that it would not see the actual information used to calculate the scores, and that it would not retain the traveler's information.

TSA had planned to test the system at selected airports during spring 2004.[37] However, CAPPS II encountered a number of obstacles to implementation. One obstacle involved obtaining the required data to test the system. Several high-profile debacles resulting in class-action lawsuits have made the U.S.-based airlines very wary of voluntarily providing passenger information. In early 2003, Delta Airlines was to begin testing CAPPS II using its customers' passenger data at three airports across the country. However, Delta became the target of a vociferous boycott campaign, raising further concerns about CAPPS II generally.[38] In September 2003, it was revealed that JetBlue shared private passenger information in September 2002 with Torch Concepts, a defense contractor, which was testing a data mining application for the U.S. Army. The information shared reportedly included itineraries, names, addresses, and phone numbers for 1.5 million passengers.[39] In January 2004, it was reported that Northwest Airlines provided personal information on millions of its passengers to the National Aeronautics and Space Administration (NASA) from October to December 2001 for an airline security-related data mining experiment.[40] In April 2004, it was revealed that American Airlines agreed to provide private passenger data on 1.2 million of its customers to TSA in June 2002, although the information was sent instead to four companies competing to win a contract with TSA.[41] Further instances of data being provided for the purpose of testing CAPPS II were brought to light during a Senate Committee on Government Affairs confirmation hearing on June 23, 2004. In his answers to the committee, the acting director of TSA, David M. Stone, stated that during 2002 and 2003 four airlines; Delta, Continental, America West, and Frontier, and two travel reservation companies; Galileo International and Sabre Holdings, provided passenger records to TSA and/or its contractors.[42]

[36] Transportation Security Administration, "TSA's CAPPS II Gives Equal Weight to Privacy, Security," Press Release, 11 March 2003, [http://www.tsa.gov/public/display?theme=44&content=535].

[37] Sara Kehaulani Goo, "U.S. to Push Airlines for Passenger Records," *Washington Post*, 12 January 2004, p. A1.

[38] The Boycott Delta website is available at [http://www.boycottdelta.org].

[39] Don Phillips, "JetBlue Apologizes for Use of Passenger Records," *The Washington Post*, 20 September 2003, p. E1; Sara Kehaulani Goo, "TSA Helped JetBlue Share Data, Report Says," *Washington Post*, 21 February 2004, p. E1.

[40] Sara Kehaulani Goo, "Northwest Gave U.S. Data on Passengers," *Washington Post*,18 January 2004, p. A1.

[41] Sara Kehaulani Goo, "American Airlines Revealed Passenger Data," *Washington Post*, 10 April 2004, p. D12.

[42] For the written responses to the committee's questions, see [http://www.epic.org/privacy/airtravel/stone_answers.pdf]; Sara Kehaulani Goo, "Agency Got More Airline Records,"*Washington Post*, 24 June 2004, p. A16.

Concerns about privacy protections had also dissuaded the European Union (EU) from providing any data to TSA to test CAPPS II. However, in May 2004, the EU signed an agreement with the United States that would have allowed PNR data for flights originating from the EU to be used in testing CAPPS II, but only after TSA was authorized to use domestic data as well. As part of the agreement, the EU data was to be retained for only three-and-a-half years (unless it is part of a law enforcement action), only 34 of the 39 elements of the PNR were to be accessed by authorities,[43] and there were to be yearly joint DHS-EU reviews of the implementation of the agreement.[44]

Another obstacle was the perception of mission creep. CAPPS II was originally intended to just screen for high-risk passengers who may pose a threat to safe air travel. However, in an August 1, 2003, *Federal Register* notice, TSA stated that CAPPS II could also be used to identify individuals with outstanding state or federal arrest warrants, as well as identify both foreign *and* domestic terrorists (not just foreign terrorists). The notice also states that CAPPS II could be "linked with the U.S. Visitor and Immigrant Status Indicator Technology (US-VISIT) program" to identify individuals who are in the country illegally (e.g., individuals with expired visas, illegal aliens, etc.).[45] In response to critics who cited these possible uses as examples of mission creep, TSA claimed that the suggested uses were consistent with the goals of improving aviation security.[46]

Several other concerns had also been raised, including the length of time passenger information was to be retained, who would have access to the information, the accuracy of the commercial data being used to authenticate a passenger's identity, the creation of procedures to allow passengers the opportunity to correct data errors in their records, and the ability of the system to detect attempts by individuals to use identity theft to board a plane undetected.

In August 2004, TSA announced that the CAPPS II program was being canceled and would be replaced with a new system called Secure Flight. In the Department of Homeland Security Appropriations Act, 2005 (P.L. 108-334), Congress included a provision (Sec. 522) prohibiting the use of appropriated funds for "deployment or implementation, on other than a test basis," of CAPPS II, Secure Flight, "or other follow on/successor programs," until GAO has certified that such a system has met

[43] Some information, such as meal preferences, which could be used to infer religious affiliation, and health considerations will not be made available. Goo, Sara Kehaulani, "U.S., EU Will Share Passenger Records," *Washington Post*, 29 May 2004, p. A2.

[44] Department of Homeland Security, "Fact Sheet: US-EU Passenger Name Record Agreement Signed," 28 May 2004, [http://www.dhs.gov/dhspublic/display?content=3651].

[45] *Federal Register*. Vol. 68 No. 148 Friday August 1, 2003. P. 45266; U.S. General Accounting Office, *Aviation Security: Challenges Delay Implementation of Computer-Assisted Passenger Prescreening System*, GAO Testimony GAO-04-504T, 17 March 2004, p. 17

[46] U.S. General Accounting Office, *Aviation Security: Challenges Delay Implementation of Computer-Assisted Passenger Prescreening System*, GAO Testimony GAO-04-504T, 17 March 2004, p. 17

all of the privacy requirements enumerated in a February 2004 GAO report,[47] can accommodate any unique air transportation needs as it relates to interstate transportation, and that "appropriate life-cycle cost estimates, and expenditure and program plans exist." GAO's certification report[48] was delivered to Congress in March 2005. In its report, GAO found that while "TSA is making progress in addressing key areas of congressional interest ... TSA has not yet completed these efforts or fully addressed these areas, due largely to the current stage of the program's development."[49]

Multistate Anti-Terrorism Information Exchange (MATRIX) Pilot Project

Similar to TIA and CAPPS II, which were born out of an initial reaction to concerns about terrorism, the impetus and initial work on MATRIX grew out of the September 11, 2001, terrorist attacks. MATRIX was initially developed by Seisint, a Florida-based information products company, in an effort to facilitate collaborative information sharing and factual data analysis. At the outset of the project, MATRIX included a component Seisint called the High Terrorist Factor (HTF). Within days of the terrorist attacks, based on an analysis of information that included "age and gender, what they did with their drivers license, either pilots or associations to pilots, proximity to 'dirty' addresses/phone numbers, investigational data, how they shipped; how they received, social security number anomalies, credit history, and ethnicity," Seisint generated a list of 120,000 names with high HTF scores, or so-called terrorism quotients. Seisint provided this list to the Federal Bureau of Investigation (FBI), the Immigration and Naturalization Service (INS), the United States Secret Service (USSS), and the Florida Department of Law Enforcement (FDLE), which, according to a January 2003 presentation, made by the company, led to "several arrests within one week" and "scores of other arrests."[50] Although the HTF scoring system appeared to attract the interest of officials, this feature was reportedly dropped from MATRIX because it relied on intelligence data not normally available to the law enforcement community and concerns about privacy abuses.

[47] The eight issues included establishing an oversight board, ensuring the accuracy of the data used, conducting stress testing, instituting abuse prevention practices, preventing unauthorized access, establishing clear policies for the operation and use of the system, satisfying privacy concerns, and created a redress process. U.S. General Accounting Office, *Aviation Security: Computer-Assisted Passenger Prescreening System Faces Significant Implementation Challenges*, GAO Report GAO-04-385, February 2004.

[48] U.S. Government Accountability Office, *Aviation Security: Secure Flight Development and Testing Under Way, but Risks Should Be Managed as System is Further Developed*, GAO Report GAO-05-356, March 2005.

[49] Ibid., p. 4; for a more detailed analysis of the Secure Flight program, see CRS Report RL32802 *Homeland Security: Air Passenger Screening and Counterterroism*, by Bart Elias and William Krouse.

[50] A copy of the presentation is available at [http://www.aclu.org/Files/OpenFile.cfm?id=15813].

However, some critics of MATRIX continued to raise questions about HTF, citing the lack of any publicly available official documentation verifying such a decision.[51]

As a pilot project, MATRIX was administered through a collaborative effort between Seisint, the FDLE,[52] and the Institute for Intergovernmental Research (IIR), a "Florida-based nonprofit research and training organization, [that] specializes in law enforcement, juvenile justice, and criminal justice issues."[53] The Florida Department of Law Enforcement (FDLE) served as the "Security Agent" for MATRIX, administering control over which agencies and individuals had access to the system. FDLE was also a participant state in MATRIX. IIR was responsible for administrative support, and was the grantee for federal funds received for MATRIX.[54]

The analytical core of the MATRIX pilot project was an application called Factual Analysis Criminal Threat Solution (FACTS). FACTS was described as a "technological, investigative tool allowing query-based searches of available state and public records in the data reference repository."[55] The FACTS application allowed an authorized user to search "dynamically combined records from disparate datasets" based on partial information, and will "assemble" the results.[56] The data reference repository used with FACTS represented the amalgamation of over 3.9 billion public records collected from thousands of sources.[57] Some of the data contained in FACTS included FAA pilot licenses and aircraft ownership records, property ownership records, information on vessels registered with the Coast Guard, state sexual offenders lists, federal terrorist watch lists, corporation filings, Uniform Commercial Code filings, bankruptcy filings, state-issued professional licenses, criminal history information, department of corrections information and photo images, driver's license information and photo images, motor vehicle registration information, and information from commercial sources that "are generally available to the public or legally permissible under federal law."[58] The data reference repository purportedly excluded data such as telemarketing call lists, direct mail mailing lists, airline reservations or travel records, frequent flyer/hotel stay program membership or activity, magazine subscriptions, information about purchases made at retailers or over the Internet, telephone calling logs or records, credit or debit card

[51] Brian Bergstein, "Database Firm Tagged 120,000 Terrorism 'Suspects' for Feds," *The SunHerald*, 20 May 2004, [http://www.sunherald.com/mld/sunherald/business/technology/8715327.htm].

[52] The FDLE website is available at [http://www.fdle.state.fl.us/].

[53] The IIR website is available at [http://www.iir.com/].

[54] See [http://www.matrix-at.org/roles.htm].

[55] For a more detailed description of FACTS, see [http://www.matrix-at.org/FACTS_defined.htm].

[56] Ibid.

[57] See [http://www.matrix-at.org/newsletter.pdf].

[58] For more information about data included and excluded from the data reference repository, see [http://www.matrix-at.org/data_sources.htm].

numbers, mortgage or car payment information, bank account numbers or balance information, birth certificates, marriage licenses, divorce decrees, or utility bill payment information.

Participating law enforcement agencies utilized this information sharing and data mining resource over the Regional Information Sharing Systems (RISS) secure intranet (RISSNET). The RISS Program is an established system of six regional centers that are used to "share intelligence and coordinate efforts against criminal networks that operate in many locations across jurisdictional lines."[59] The RISS Program is used to combat traditional law enforcement targets, such as drug trafficking and violent crime, as well as other activities, such as terrorism and cybercrime. According to its website, RISS has been in operation for nearly 25 years, and has "member agencies in all 50 states, the District of Columbia, U.S. territories, Australia, Canada, and England."[60]

Some critics of MATRIX suggested that the original intentions and design of the pilot project echoed those of DARPA's highly criticized TIA program.[61] However, while it is difficult to ascribe intention, an ongoing series of problems did appear to have affected the trajectory of the project. In August 2003, Hank Asher, the founder of Seisint, resigned from the company's board of directors after questions about his criminal history were raised during contract negotiations between Seisint and the Florida Department of Law Enforcement. In the 1980s, Asher was allegedly a pilot in several drug smuggling cases. However, he was reportedly never charged in the cases in exchange for his testimony at state and federal trials. Similar concerns had surfaced in 1999 when the FBI and the U.S. Drug Enforcement Agency (DEA) reportedly cancelled contracts with an earlier company Asher founded, DBT Online, Inc.[62]

Some civil liberties organizations also raised concerns about law enforcement actions being taken based on algorithms and analytical criteria developed by a private corporation, in this case Seisint, without any public or legislative input.[63] Questions also were raised about the level of involvement of the federal government, particularly the Department of Homeland Security and the Department of Justice, in

[59] For a detailed description of RISS, see [http://www.iir.com/riss/] and [http://www.rissinfo.com/].

[60] [http://www.rissinfo.com/overview2.htm].

[61] John Schwartz, "Privacy Fears Erode Support for a Network to Fight Crime," *New York Times*, 15 March 2004, [http://www.nytimes.com/2004/03/15/technology/15matrix.html].

[62] Cynthia L. Webb, "Total Information Dilemma," *Washington Post*, 27 May 2004, [http://www.washingtonpost.com/ac2/wp-dyn/A60986-2004May27?language=printer]; Lucy Morgan, "Ex-drug Runner Steps Aside," *St.Petersburg Times*, 30 August 2003, [http://www.sptimes.com/2003/08/30/State/Ex_drug_runner_steps_.shtml]; Bill Cotterell, and Nancy Cook Lauer, "Bush Defends Pick of Computer Firm, Former Leader's Background Raises Questions," *Tallahassee Democrat*, 22 May 2004, [http://www.tallahassee.com/mld/tallahassee/news/local/8728776.htm].

[63] Welsh, William Welsh, "Feds Offer to Mend Matrix," *Washington Technology*, 24 May 2004, [http://www.washingtontechnology.com/news/19_4/egov/23597-1.html].

a project that is ostensibly focused on supporting state-based information sharing.[64] It has been reported that the MATRIX pilot project has received a total of $12 million in federal funding - $8 million from the Office of Domestic Preparedness (ODP) at the Department of Homeland Security (DHS), and $4 million from the Bureau of Justice Assistance (BJA) at the Department of Justice (DOJ).[65]

The MATRIX pilot project also suffered some setbacks in recruiting states to participate. The lack of participation can be especially troubling for a networked information sharing project such as MATRIX because, as Metcalfe's Law suggests, "the power of the network increases exponentially by the number of computers connected to it."[66] While as many as 16 states were reported to have either participated or seriously considered participating in MATRIX, several chose to withdraw, leaving a total of four states (Connecticut, Florida, Ohio, and Pennsylvania) at the conclusion of the pilot on April 15, 2005. State officials cited a variety of reasons for not participating in MATRIX, including costs, concerns about violating state privacy laws, and duplication of existing resources.[67]

In its news release announcing the conclusion of the pilot, the FDLE stated that as a proof-of-concept pilot study from July 2003 to April 2005, MATRIX had achieved many "operational successes." Among the statistics cited, the news release stated that

- Between July 2003 and April 2005, there have been 1,866,202 queries to the FACTS application.
- As of April 8, 2005, there were 963 law enforcement users accessing FACTS.
- FACTS assisted a variety of investigations. On average, cases pertained to the following:
  - Fraud — 22.6%
  - Robbery — 18.8%
  - Sex Crime Investigations — 8.6%

[64] O'Harrow, Jr., Robert O'Harrow, Jr., "Anti-Terror Database Got Show at White House," *Washington Post*, 21 May 2004, p. A12.

[65] John Schwartz, "Privacy Fears Erode Support for a Network to Fight Crime," *New York Times*, 15 March 2004, [http://www.nytimes.com/2004/03/15/technology/15matrix.html]; see also: [http://www.matrix-at.org/faq.htm].

[66] For a more detailed discussion of Metcalfe's Law, see [http://searchnetworking.techtarget.com/sDefinition/0,,sid7_gci214115,00.html].

[67] The states that have reportedly decided to withdraw from the pilot project include Alabama, California, Georgia, Kentucky, Louisiana, New York, Oregon, South Carolina, Texas, Utah, and Wisconsin. Larry Greenemeier, "Two More States Withdraw From Database," *InformationWeek*, 12 March 2004, [http://www.informationweek.com/story/showArticle.jhtml?articleID=18312112]; Diane Frank, "Utah No Longer Part of MATRIX," *Federal Computer Week*, 5 April 2004, p. 14; Associated Press, "Two More States Withdraw From Controversial Database Program," *Star-Telegram*, 12 March 2004, [http://www.dfw.com/mld/dfw/business/8170978.htm?1c]; Associated Press, "Matrix Plan Fuels Privacy Fears," *Wired News*, 2 February 2004, [http://www.wired.com/news/business/0,1367,62141,00.html].

- Larceny and Theft — 8.3%
- Extortion/Blackmail — 7.0%
- Burglary/Breaking and Entering — 6.8%
- Stolen Property — 6.2%
- Terrorism/National Security — 2.6%
- Other — 19.1% (e.g., assault, arson, narcotics, homicide)

It was also announced that while the pilot study would not be continued, due to a lack of additional federal funding, that Florida and other participating states were "independently negotiating the continued use of the FACTS application for use within their individual state[s]."[68]

Other Data Mining Initiatives

Able Danger. In summer 2005, news reports began to appear regarding a data mining initiative that had been carried out by the U.S. Army's Land Information Warfare Agency (LIWA) in 1999-2000. The initiative, referred to as Able Danger, had reportedly been requested by the U.S. Special Operations Command (SOCOM) as part of larger effort to develop a plan to combat transnational terrorism. Because the details of Able Danger remain classified, little is known about the program. However, in a briefing to reporters, the Department of Defense characterized Able Danger as a demonstration project to test analytical methods and technology on very large amounts of data.[69] The project involved using link analysis to identify underlying connections and associations between individuals who otherwise appear to have no outward connection with one another. The link analysis used both classified and open source data, totaling a reported 2.5 terabytes.[70] All of this data, which included information on U.S. persons, was reportedly deleted in April 2000 due to U.S. Army regulations requiring information on U.S. persons be destroyed after a project ends or becomes inactive.[71]

Interest in Able Danger was largely driven by controversy over allegations that the data mining analysis had resulted in the identification Mohammed Atta, one of the 9/11 hijackers, as a terrorist suspect before the attacks took place. While some individuals who had been involved in Able Danger were reportedly prepared to testify that they had seen either his name and/or picture on a chart prior to the attacks, the identification claim was strongly disputed by others.

[68] Florida Department of Law Enforcement (FDLE). "News Release: MATRIX Pilot Project Concludes," 15 April 2005, [http://www.fdle.state.fl.us/press_releases/expired/2005/20050415_matrix_project.html].

[69] Department of Defense, Special Defense Department Briefing, 1 September 2005, [http://www.defenselink.mil/transcripts/2005/tr20050901-3844.html].

[70] Shane Harris, "Homeland Security - Intelligence Designs," *National Journal*, 3 December 2005, [http://www.govexec.com/dailyfed/1205/120705nj1.htm].

[71] Erik Kleinsmith, Testimony before the Senate Committee on the Judiciary, *Able Danger and Intelligence Information Sharing*, 21 September 2005, [http://judiciary.senate.gov/testimony.cfm?id=1606&wit_id=4669].

On September 21, 2005, the Senate Committee on the Judiciary held a hearing on Able Danger to consider how the data could or should have been shared with other agencies, and whether the destruction of the data was in fact required by the relevant regulations. While the Department of Defense directed the individuals involved in Able Danger not to testify at the hearing, testimony was taken from the attorney of one of the individuals, as well as others not directly involved with the project.

National Security Agency (NSA) and the Novel Intelligence from Massive Data (NIDM) Program. In December 2005 news reports appeared for the first time revealing the existence of of a classified NSA terrorist surveillance program, dating back to at least 2002, involving the domestic collection, analysis, and sharing of information.[72] Although the details of program, such as what information is being collected and how it is being analyzed, are not publicly known, the controversy over the program has raised congressional concerns about both the prevalence of homeland security data mining and the capacity of the country's intelligence and law enforcement agencies to adequately analyze and share counterterrorism information. The Senate Committee on the Judiciary has announced it plans to hold a hearing regarding the issue on February 6, 2006.

As part of its efforts to better utilize the overwhelming flow of information it collects, NSA has reportedly been supporting the development of new technology and data management techniques by funding grants given by the Advanced Research Development Activity (ARDA). ARDA is an intelligence community (IC) organization whose mission is described as "to sponsor high-risk, high-payoff research designed to leverage leading edge technology to solve some of the most critical problems facing the Intelligence Community (IC)."[73] ARDA's research support is organized into various technology "thrusts" representing the most critical areas of development. Some of ARDA's current research thrusts include Information Exploitation, Quantum Information Science, Global Infosystems Access, Novel Intelligence from Massive Data, and Advanced Information Assurance.

The Novel Intelligence from Massive Data (NIMD) program focuses on the development of data mining and analysis tools to be used in working with massive data.[74] Novel intelligence refers to "actionable information not previously known." Massive data refers to data that has characteristics that are especially challenging to common data analysis tools and methods. These characteristics can include unusual volume, breadth (heterogeneity), and complexity. Data sets that are one petabyte (one quadrillion bytes) or larger are considered to be "massive." Smaller data sets that contain items in a wide variety of formats, or are very heterogeneous (i.e.,

[72] Peter Baker, "President Says He Ordered NSA Domestic Spying," *The Washington Post*, 18 December 2005, p. A1; Walter Pincus, "NSA Gave Other U.S. Agencies Information From Surveillance," *The Washington Post*, 1 January 2006, p. A8.

[73] [https://rrc.mitre.org/cfp06.pdf].

[74] Shane Harris, "NSA Spy Program Hinges on State-of-the-Art Technology," *Government Executive Magazine*, 20 January 2006, [http://www.govexec.com/dailyfed/0106/012006nj1.htm]; Wilson P. Dizard III, "NSA Searches for Novel Intel Answers in the Glass Box," *Government Computer News*, 20 June 2005, [http://www.gcn.com/24_15/news/36139-1.html].

unstructured text, spoken text, audio, video, graphs, diagrams, images, maps, equations, chemical formulas, tables, etc.) can also be considered "massive." According to ARDA's website (no longer available)[75] "some intelligence data sources grow at a rate of four petabytes per month now, and the rate of growth is increasing." With the continued proliferation of both the means and volume of electronic communications, it is expected that the need for more sophisticated tools will intensify. Whereas NSA once predicted it was in danger of becoming proverbially deaf due to the spreading use of encrypted communications, it appears that NSA may now be at greater risk of being "drowned" in information.

Data Mining Issues

As data mining initiatives continue to evolve, there are several issues Congress may decide to consider related to implementation and oversight. These issues include, but are not limited to, data quality, interoperability, mission creep, and privacy. As with other aspects of data mining, while technological capabilities are important, other factors also influence the success of a project's outcome.

Data Quality

Data quality is a multifaceted issue that represents one of the biggest challenges for data mining. Data quality refers to the accuracy and completeness of the data. Data quality can also be affected by the structure and consistency of the data being analyzed. The presence of duplicate records, the lack of data standards, the timeliness of updates, and human error can significantly impact the effectiveness of the more complex data mining techniques, which are sensitive to subtle differences that may exist in the data. To improve data quality, it is sometimes necessary to "clean" the data, which can involve the removal of duplicate records, normalizing the values used to represent information in the database (e.g., ensuring that "no" is represented as a 0 throughout the database, and not sometimes as a 0, sometimes as a N, etc.), accounting for missing data points, removing unneeded data fields, identifying anomalous data points (e.g., an individual whose age is shown as 142 years), and standardizing data formats (e.g., changing dates so they all include MM/DD/YYYY).

Interoperability

Related to data quality, is the issue of interoperability of different databases and data mining software. Interoperability refers to the ability of a computer system and/or data to work with other systems or data using common standards or processes. Interoperability is a critical part of the larger efforts to improve interagency collaboration and information sharing through e-government and homeland security initiatives. For data mining, interoperability of databases and software is important to enable the search and analysis of multiple databases simultaneously, and to help ensure the compatibility of data mining activities of different agencies. Data mining

[75] ARDA's website was previously available at [http://www.ic-arda.org].

projects that are trying to take advantage of existing legacy databases or that are initiating first-time collaborative efforts with other agencies or levels of government (e.g., police departments in different states) may experience interoperability problems. Similarly, as agencies move forward with the creation of new databases and information sharing efforts, they will need to address interoperability issues during their planning stages to better ensure the effectiveness of their data mining projects.

Mission Creep

Mission creep is one of the leading risks of data mining cited by civil libertarians, and represents how control over one's information can be a tenuous proposition. Mission creep refers to the use of data for purposes other than that for which the data was originally collected. This can occur regardless of whether the data was provided voluntarily by the individual or was collected through other means.

Efforts to fight terrorism can, at times, take on an acute sense of urgency. This urgency can create pressure on both data holders and officials who access the data. To leave an available resource unused may appear to some as being negligent. Data holders may feel obligated to make any information available that could be used to prevent a future attack or track a known terrorist. Similarly, government officials responsible for ensuring the safety of others may be pressured to use and/or combine existing databases to identify potential threats. Unlike physical searches, or the detention of individuals, accessing information for purposes other than originally intended may appear to be a victimless or harmless exercise. However, such information use can lead to unintended outcomes and produce misleading results.

One of the primary reasons for misleading results is inaccurate data. All data collection efforts suffer accuracy concerns to some degree. Ensuring the accuracy of information can require costly protocols that may not be cost effective if the data is not of inherently high economic value. In well-managed data mining projects, the original data collecting organization is likely to be aware of the data's limitations and account for these limitations accordingly. However, such awareness may not be communicated or heeded when data is used for other purposes. For example, the accuracy of information collected through a shopper's club card may suffer for a variety of reasons, including the lack of identity authentication when a card is issued, cashiers using their own cards for customers who do not have one, and/or customers who use multiple cards.[76] For the purposes of marketing to consumers, the impact of these inaccuracies is negligible to the individual. If a government agency were to use that information to target individuals based on food purchases associated with particular religious observances though, an outcome based on inaccurate information could be, at the least, a waste of resources by the government agency, and an unpleasant experience for the misidentified individual. As the March 2004 TAPAC report observes, the potential wide reuse of data suggests that concerns about mission creep can extend beyond privacy to the protection of civil rights in the event that

[76] Technology and Privacy Advisory Committee, Department of Defense. *Safeguarding Privacy in the Fight Against Terrorism*, March 2004, p. 40.

information is used for "targeting an individual solely on the basis of religion or expression, or using information in a way that would violate the constitutional guarantee against self-incrimination."[77]

Privacy

As additional information sharing and data mining initiatives have been announced, increased attention has focused on the implications for privacy. Concerns about privacy focus both on actual projects proposed, as well as concerns about the potential for data mining applications to be expanded beyond their original purposes (mission creep). For example, some experts suggest that anti-terrorism data mining applications might also be useful for combating other types of crime as well.[78] So far there has been little consensus about how data mining should be carried out, with several competing points of view being debated. Some observers contend that tradeoffs may need to be made regarding privacy to ensure security. Other observers suggest that existing laws and regulations regarding privacy protections are adequate, and that these initiatives do not pose any threats to privacy. Still other observers argue that not enough is known about how data mining projects will be carried out, and that greater oversight is needed. There is also some disagreement over how privacy concerns should be addressed. Some observers suggest that technical solutions are adequate. In contrast, some privacy advocates argue in favor of creating clearer policies and exercising stronger oversight. As data mining efforts move forward, Congress may consider a variety of questions including, the degree to which government agencies should use and mix commercial data with government data, whether data sources are being used for purposes other than those for which they were originally designed, and the possible application of the Privacy Act to these initiatives.

Legislation in the 108th Congress

During the 108th Congress, a number of legislative proposals were introduced that would restrict data mining activities by some parts of the federal government, and/or increase the reporting requirements of such projects to Congress. For example, on January 16, 2003, Senator Feingold introduced S. 188 the Data-Mining Moratorium Act of 2003, which would have imposed a moratorium on the implementation of data mining under the Total Information Awareness program (now referred to as the Terrorism Information Awareness project) by the Department of Defense, as well as any similar program by the Department of Homeland Security. S. 188 was referred to the Committee on the Judiciary.

On January 23, 2003, Senator Wyden introduced S.Amdt. 59, an amendment to H.J.Res. 2, the Omnibus Appropriations Act for Fiscal Year 2003. As passed in its final form as part of the omnibus spending bill (P.L. 108-7) on February 13, 2003,

[77] Ibid., p. 39.

[78] Drew Clark, "Privacy Experts Differ on Merits of Passenger-Screening Program," *Government Executive Magazine*, November 21, 2003, [http://www.govexec.com/dailyfed/1103/112103td2.htm].

and signed by the President on February 20, 2003, the amendment requires the Director of Central Intelligence, the Secretary of Defense, and the Attorney General to submit a joint report to Congress within 90 days providing details about the TIA program.[79] Some of the information required includes spending schedules, likely effectiveness of the program, likely impact on privacy and civil liberties, and any laws and regulations that may need to be changed to fully deploy TIA. If the report had not submitted within 90 days, funding for the TIA program could have been discontinued.[80] Funding for TIA was later discontinued in Section 8131 of the FY2004 Department of Defense Appropriations Act (P.L. 108-87), signed into law on September 30, 2003.[81]

On March 13, 2003, Senator Wyden introduced an amendment to S. 165 the Air Cargo Security Act, requiring the Secretary of Homeland Security to submit a report to Congress within 90 days providing information about the impact of CAPPS II on privacy and civil liberties. The amendment was passed by the Committee on Commerce, Science, and Transportation, and the bill was forwarded for consideration by the full Senate (S.Rept. 108-38). In May 2003, S. 165 was passed by the Senate with the Wyden amendment included and was sent to the House where it was referred to the Committee on Transportation and Infrastructure.

Funding restrictions on CAPPSII were included in section 519 of the FY2004 Department of Homeland Security Appropriations Act (P.L. 108-90), signed into law October 1, 2003. This provision included restrictions on the "deployment or implementation, on other than a test basis, of the Computer-Assisted Passenger Prescreening System (CAPPSII)," pending the completion of a GAO report regarding the efficacy, accuracy, and security of CAPPSII, as well as the existence of a system of an appeals process for individuals identified as a potential threat by the system.[82] In its report delivered to Congress in February 2004, GAO reported that "As of January 1, 2004, TSA has not fully addressed seven of the eight CAPPSII issues identified by the Congress as key areas of interest."[83] The one issue GAO determined that TSA had addressed is the establishment of an internal oversight board. GAO

[79] The report is available at [http://www.eff.org/Privacy/TIA/TIA-report.pdf].

[80] For more details regarding this amendment, see CRS Report RL31786, *Total Information Awareness Programs: Funding, Composition, and Oversight Issues*, by Amy Belasco.

[81] For further details regarding this provision, see CRS Report RL31805 *Authorization and Appropriations for FY2004: Defense*, by Amy Belasco and Stephen Daggett.

[82] Section 519 of P.L. 108-90 specifically identifies eight issues that TSA must address before it can spend funds to deploy or implement CAPPSII on other than a test basis. These include 1. establishing a system of due process for passengers to correct erroneous information; 2. assess the accuracy of the databases being used; 3. stress test the system and demonstrate the efficiency and accuracy of the search tools; 4. establish and internal oversight board; 5. install operational safeguards to prevent abuse; 6. install security measures to protect against unauthorized access by hackers or other intruders; 7. establish policies for effective oversight of system use and operation; and 8. address any privacy concerns related to the system.

[83] General Accounting Office, *Aviation Security: Computer-Assisted Passenger Prescreening System Faces Significant Implementation Challenges*, GAO-04-385, February 2004, p. 4.

attributed the incomplete progress on these issues partly to the "early stage of the system's development."[84]

On March 25, 2003, the House Committee on Government Reform Subcommittee on Technology, Information Policy, Intergovernmental Relations, and the Census held a hearing on the current and future possibilities of data mining. The witnesses, drawn from federal and state government, industry, and academia, highlighted a number of perceived strengths and weaknesses of data mining, as well as the still-evolving nature of the technology and practices behind data mining.[85] While data mining was alternatively described by some witnesses as a process, and by other witnesses as a productivity tool, there appeared to be a general consensus that the challenges facing the future development and success of government data mining applications were related less to technological concerns than to other issues such as data integrity, security, and privacy. On May 6 and May 20, 2003 the Subcommittee also held hearings on the potential opportunities and challenges for using factual data analysis for national security purposes.

On July 29, 2003, Senator Wyden introduced S. 1484 The Citizens' Protection in Federal Databases Act, which was referred to the Committee on the Judiciary. Among its provisions, S. 1484 would have required the Attorney General, the Secretary of Defense, the Secretary of Homeland Security, the Secretary of the Treasury, the Director of Central Intelligence, and the Director of the Federal Bureau of Investigation to submit to Congress a report containing information regarding the purposes, type of data, costs, contract durations, research methodologies, and other details before obligating or spending any funds on commercially available databases. S. 1484 would also have set restrictions on the conduct of searches or analysis of databases "based solely on a hypothetical scenario or hypothetical supposition of who may commit a crime or pose a threat to national security."

On July 31, 2003, Senator Feingold introduced S. 1544 the Data-Mining Reporting Act of 2003, which was referred to the Committee on the Judiciary. Among its provisions, S. 1544 would have required any department or agency engaged in data mining to submit a public report to Congress regarding these activities. These reports would have been required to include a variety of details about the data mining project, including a description of the technology and data to be used, a discussion of how the technology will be used and when it will be deployed, an assessment of the expected efficacy of the data mining project, a privacy impact assessment, an analysis of the relevant laws and regulations that would govern the project, and a discussion of procedures for informing individuals their personal information will be used and allowing them to opt out, or an explanation of why such procedures are not in place.

[84] Ibid.

[85] Witnesses testifying at the hearing included Florida State Senator Paula Dockery, Dr. Jen Que Louie representing Nautilus Systems, Inc., Mark Forman representing OMB, Gregory Kutz representing GAO, and Jeffrey Rosen, an Associate Professor at George Washington University Law School.

Also on July 31, 2003, Senator Murkowski introduced S. 1552 the Protecting the Rights of Individuals Act, which was referred to the Committee on the Judiciary. Among its provisions, section 7 of S. 1552 would have imposed a moratorium on data mining by any federal department or agency "except pursuant to a law specifically authorizing such data-mining program or activity by such department or agency." It also would have required

> The head of each department or agency of the Federal Government that engages or plans to engage in any activities relating to the development or use of a data-mining program or activity shall submit to Congress, and make available to the public, a report on such activities.

On May 5, 2004, Representative McDermott introduced H.R. 4290 the Data-Mining Reporting Act of 2004, which was referred to the House Committee on Government Reform Subcommittee on Technology, Information Policy, Intergovernmental Relations, and the Census. H.R. 4290 would have required

> each department or agency of the Federal Government that is engaged in any activity or use or develop data-mining technology shall each submit a public report to Congress on all such activities of the department or agency under the jurisdiction of that official.

A similar provision was included in H.R. 4591/S. 2528 the Civil Liberties Restoration Act of 2004. S. 2528 was introduced by Senator Kennedy on June 16, 2004 and referred to the Committee on the Judiciary. H.R. 4591 was introduced by Representative Berman on June 16, 2004 and referred to the Committee on the Judiciary and the Permanent Select Committee on Intelligence.

Legislation in the 109th Congress

Data mining has continued to be a subject of interest to Congress in the 109th Congress. On April 6, 2005, H.R. 1502 the Civil Liberties Restoration Act of 2005 was introduced by Representative Berman and was referred to the Committee on the Judiciary[86], the Permanent Select Committee on Intelligence, and the Committee on Homeland Security. Section 402, Data-Mining Report, of H.R. 1502 would require that

> The Head of each department or agency of the Federal Government that is engaged in any activity or use or develop data-mining technology shall each submit a public report to Congress on all such activities of the department or agency under the jurisdiction of that official.

As part of their content, these reports would be required to provide, for each data mining activity covered by H.R. 1502, information regarding the technology and data being used; information on how the technology would be used and the target dates for deployment; an assessment of the likely efficacy of the data mining technology; an assessment of the likely impact of the activity on privacy and civil

[86] H.R. 1502 was referred to the Subcommittee on Immigration, Border Security, and Claims on May 10, 2005, and later discharged by the subcommittee on July 8, 2005.

liberties; a list and analysis of the laws and regulations that would apply to the data mining activity and whether these laws and regulations would need to be modified to allow the data mining activity to be implemented; information on the policies, procedures, and guidelines that would be developed and applied to protect the privacy and due process rights of individuals, and ensure that only accurate information is collected and used; and information on how individuals whose information is being used in the data mining activity will be notified of the use of their information, and, if applicable, what options will be available for individual to opt-out of the activity. These reports would be due to Congress no later than 90 days after the enactment of H.R. 1502, and would be required to be updated annually to include "any new data-mining technologies."

On June 6, 2005, S. 1169, the Federal Agency Data-Mining Reporting Act of 2005 was introduced by Senator Feingold, and was referred to the Senate Committee on the Judiciary. Among its provisions, S. 1169 would require any department or agency engaged in data mining to submit a public report to Congress regarding these activities. These reports would have been required to include a variety of details about the data mining project, including a description of the technology and data to be used, a discussion of the plans and goals for using the technology when it will be deployed, an assessment of the expected efficacy of the data mining project, a privacy impact assessment, an analysis of the relevant laws and regulations that would govern the project, and a discussion of procedures for informing individuals their personal information will be used and allowing them to opt out, or an explanation of why such procedures are not in place.

On October 6, 2005, H.R. 4009, the Department of Homeland Security Reform Act of 2005, was introduced by Representative Thompson, and was referred to the Committee on Homeland Security, the Permanent Select Committee on Intelligence, and the Committee on Transportation and Infrastructure. Section 203(c)(16) would direct the Chief Intelligence Officer, as established in Section 203(a):

> To establish and utilize, in conjunction with the Chief Information Officer of the Department, a secure communications and information technology infrastructure, including data-mining and other advanced analytical tools, in order to access, receive, and analyze data and information in furtherance of the responsibilities under this section, and to disseminate information acquired and analyzed by the Department, as appropriate.

On December 6, 2005, H.R. 4437, the Border Protection, Antiterrorism, and Illegal Immigration Control Act of 2005 was introduced by Representative Sensenbrenner and was referred to the Committee on the Judiciary and the Committee on Homeland Security. On December 8, 2005, the Committee on the Judiciary held a markup session and ordered an amended version of H.R. 4437 to be reported. On December 13, 2005, the Committee on Homeland Security discharged the bill, which was subsequently referred to and discharged from the Committee on Education and the Workforce and the Committee on Ways and Means. On December 16, 2005, H.R. 4437 was passed by the House and sent to the Senate.

Section 1305, Authority of the Office of Security and Investigations to Detect and Investigate Immigration Benefits Fraud, of H.R. 4437 would grant the Office of

Security and Investigations of the United States Citizenship and Immigration Services at the Department of Homeland Security the authority to:

> (1) to conduct fraud detection operations, including data mining and analysis;
> (2) to investigate any criminal or noncriminal allegations of violations of the Immigration and Nationality Act or title 18, United States Code, that Immigration and Customs Enforcement declines to investigate;
> (3) to turn over to a United States Attorney for prosecution evidence that tends to establish such violations; and
> (4) to engage in information sharing, partnerships, and other collaborative efforts with any —
> (A) Federal, State, or local law enforcement entity;
> (B) foreign partners; or
> (C) entity within the intelligence community (as defined in section 3(4) of the National Security Act of 1947 (50 U.S.C. 401a(4)).

For Further Reading

CRS Report RL32802, *Homeland Security: Air Passenger Prescreening and Counterterrorism*, by Bart Elias and William Krouse.

CRS Report RL31408, *Internet Privacy: Overview and Pending Legislation*, by Marcia S. Smith.

CRS Report RL32536, *Multi-State Anti-Terrorism Information Exchange (MATRIX)Pilot Project*, by William J. Krouse.

CRS Report RL30671, *Personal Privacy Protection: The Legislative Response*, by Harold C. Relyea. Out of print; available from author (7-8679).

CRS Report RL31730, *Privacy: Total Information Awareness Programs and Related Information Access, Collection, and Protection Laws*, by Gina Marie Stevens.

CRS Report RL31786, *Total Information Awareness Programs: Funding, Composition, and Oversight Issues*, by Amy Belasco.

DARPA, *Report to Congress Regarding the Terrorism Information Awareness Program*, May 20, 2003, [http://www.eff.org/Privacy/TIA/TIA-report.pdf].

Department of Defense, Office of the Inspector General, *Information Technology Management: Terrorism Information Awareness Program (D-2004-033)*, Dec. 12, 2003 [http://www.dodig.osd.mil/audit/reports/FY04/04-033.pdf].