NASCIO Representing Chief Information Officers of the States

Research Brief

Sept. 2004

Think Before You Dig: Privacy Implications of Data Mining & Aggregation

Background

Data mining is generally part of a larger business intelligence or knowledge management initiative. Since state governments are complex organizations that collect and process massive amounts of information, data mining can help provide value to state government operations and taxpayers by extracting useful information out of mountains of collected data. In addition, data mining can be predictive and uncover hidden patterns that states can strategically use to reduce costs, increase business expansion opportunities, and detect fraud, waste and abuse that drains away taxpayer dollars.

With the proliferation of privacy concerns raised by the mere mention of the term "data mining," defining what data mining is and is not has become increasingly important. ***As recently defined by the U.S. Government Accountability Office (formerly the U.S. General Accounting Office) (GAO), data mining is "the application of database technology and techniques—such as statistical analysis and modeling—to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results."***[1] However, data warehousing, ad hoc inquiries/reporting, software agents, online analytical processes (OLAP), and data visualization alone do not constitute data mining.[2] GAO acknowledges that the term "data mining" is ambiguous and, according to some experts in the field, overlaps with other types of analytical activities, such as data profiling, data warehousing, online analytical processing, and enterprise analytical applications.[3] Examples of analytical approaches that fall within the generally accepted definition of data mining are decision trees, nearest neighbor classification, neural networks, rule induction, and k-means clustering.[4]

A common misconception is that data mining and data aggregation are interchangeable terms. ***Data aggregation is considered to be "any process in which information is***

[1] U.S. General Accounting Office (GAO), "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548, May 2004, <http://www.gao.gov/new.items/d04548.pdf>.

[2] Kurt Thearling, Ph.D., "An Introduction to Data Mining," September 2004, <http://www.thearling.com/dmintro/dmintro.htm>.

[3] U.S. General Accounting Office (GAO), "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548, May 2004.

[4] Kurt Thearling, Ph.D., "An Introduction to Data Mining," September 2004.

gathered and expressed in a summary form, for purposes such as statistical analysis."[5] Note, though, that data aggregation may be used to prepare data for subsequent mining. Other important terms related to data mining are:

- **Data Mart:** "An extract of the data warehouse which is established for separate purposes," such as for security, performance or special content.[6]
- **Data Warehouse:** A repository that combines data from existing databases into one database so that the data can be analyzed or mined.
- **Data Augmentation:** The use of information from other sources, such as commercial databases, in order to add to information that an entity already has in its possession.

Purpose and Uses of Data Mining

The purpose of data mining is to identify patterns in order to make predictions from information contained in databases. It allows the user to be proactive in identifying and predicting trends with that information. Common uses of data mining in government include knowledge discovery, fraud detection, analysis of research, decision support, and website personalization. The most common federal government uses of data mining as identified by GAO include:

- Improving service or performance
- Detecting fraud, waste, and abuse
- Analyzing scientific and research information
- Managing human resources
- Detecting criminal activities or patterns
- Analyzing intelligence and detecting terrorist activities.[7]

State government data mining efforts include programs to ensure that the proper beneficiaries of state benefits programs receive the correct amount of benefits. Such uses can save states substantial amounts of money that otherwise would be erroneously paid out in the form of state benefits.[8] Moreover, in a recent report, GAO found that twenty-one states are using data mining software to look for unusual patterns in claims, provider, and beneficiary information stored in data warehouses in order to identify potential provider abuse.[9]

Privacy Implications

As data mining has evolved, its impact on privacy has become increasingly complex and controversial. Data mining technologies initially assisted the user in accessing and

[5] Tech Target Network, <http://searchdatabase.techtarget.com/sDefinition/0,,sid13_gci532310,00.html>.
[6] Wikipedia, September, 2004, <http://en.wikipedia.org/wiki/Data_mart>.
[7] U.S. General Accounting Office (GAO), "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548, May 2004.
[8] National Association of State Auditors, Comptrollers and Treasurers (NASACT), "Using Data Mining to Make Audits More Efficient and Effective," 2004 National State Auditors Association (NSAA) Middle Management Conference Presentations, <http://www.nasact.org/onlineresources/downloads/2004_MM/2004_MM.htm>.
[9] U.S. Government Accountability Office (GAO), "Medicaid Program Integrity: State and Federal Efforts to Prevent and Detect Improper Payments," GAO-04-707, July 2004, <http://www.gao.gov/new.items/d04707.pdf>.

reducing large amounts of information. However, the factors listed below have made addressing privacy in relation to data mining much more difficult:

- Increased availability and decreased cost of data mining tools (for example, the data mining market is expected to grow from $540 million in 2002 to $1.5 billion in 2005)[10]
- Increased digitization of data and consequential increase in the amount of data available and inability of humans to manually process relationships in data without computer assistance
- Increased data aggregation
- Increased ability of data mining tools to extract patterns that go beyond actual data and that attempt to predict repetitive behavior and data value patterns
- Increased use of data warehouses as central repositories for multiple applications.

Personal Information: The privacy implications of data mining technologies tend to be two-fold. First, the mining of *personal* information has raised privacy concerns. For purposes of its data mining study, GAO considered "personal information" to be "all information associated with an individual and includes both identifying and non-identifying information." Examples of identifying information which can be used to locate or identify an individual include an individual's name, aliases, Social Security Number, e-mail address, driver's license number, and agency-assigned case number. Non-identifying personal information includes an individual's age, education, finances, criminal history, physical attributes, and gender.[11] The main concern with aggregating such personal information and mining it is that profiles of individuals can be created using information held in disparate systems located both in the commercial and government sectors.

Identification of Terrorists and Criminals: Another set of privacy issues are raised when data mining is used to identify individuals involved in terrorist or criminal activity or to determine if an already-identified suspect has a pattern of being involved in criminal or terrorist activities. Since data mining can be used to predict which individual might commit a crime, those using data mining for that purpose must be careful to put in place requirements that detail when action may be taken against an individual as the result of data mining activities and what is done with mined information that is subsequently determined not to be relevant to an investigation.

Lessons Learned: The privacy implications of data mining recently have become much more high profile with controversies over TIA (Terrorism Information Awareness) program, CAPPS II (Computer Assisted Passenger Prescreening System), and MATRIX (Multistate Anti-Terrorism Information Exchange). These programs have raised concerns about the collection of personal information by the government and the subsequent mining of that information. The concerns of the privacy advocacy community appear to focus on the following issues:

[10] Kurt Thearling, Ph.D., "An Introduction to Data Mining," September 2004.

[11] U.S. General Accounting Office (GAO), "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548, May 2004.

- Whether there is a clear description of a program's collection of personal information, including how the collected information will serve the program's purpose
- Whether information collected for one purpose will then be used for additional, secondary purposes in the future
- Whether privacy protections are built-in to systems in the developmental stage
- Whether the information will be mined after collection (and possibly combined with other information from government and/or private sector sources) and used to create dossiers on individuals in order to identify potential terrorists or criminals
- What type of action will be taken by the government on the basis of information gleaned from a data mining program
- Whether there is an adequate redress system for individuals to review and correct their personal information that is collected and maintained in order to avoid repeated "false positives" resulting from a data mining program
- Whether there are proper disposal procedures for collected personal information that has been determined to be irrelevant to an investigation.

A lesson learned from TIA, CAPPS II, and MATRIX is that transparency as to a data mining program's purpose, the reason why information is collected, how it will be used, who will have access to the information, how it will be secured, and whether individuals can access and correct their personal information are key. These considerations are based upon "the Fair Information Principles (FIPs)," which are the core underpinnings of information privacy. While public policy concerns, such as national security, may necessitate a lesser level of openness regarding the details of programs that have data mining potential, forthrightness with the public and privacy advocates in the beginning stages of a data mining program (or a program that might appear to the public to have data mining potential) can help to avoid a myriad of public scrutiny once a program is underway. See Appendix A for more on TIA, CAPPS II, and MATRIX.

In addition to the lessons learned from programs like TIA, CAPPS II and MATRIX, an important issue that has been raised within the data mining debate is the existence of an inherent tension between limiting the secondary use of information and mining information that was collected for purposes other than data mining. A common tenet of the FIPs is to limit the use of information to the purposes for which it was originally collected. Because information that is mined may not have been collected with the original purpose of being mined, those conducting data mining activities should examine whether the mining of data is consistent with the purposes for which it was originally collected.

A Reality Check on Current Federal Data Mining Efforts: Examining federal government data mining efforts can assist in putting some privacy concerns in perspective. Initially, data mining was used by the federal government to detect financial fraud and abuse, such as through GAO audits and investigations of federal government purchase and credit card programs.[12] However, based upon GAO's 2004 study of current and planned federal data mining efforts, NASCIO has identified three major points:

- ***Data mining efforts can involve information that is not personal in nature***
- ***There is a wide range of purposes for which data mining can be conducted and not all purposes involve the analysis of intelligence and detection of terrorist activities***
- ***Data mining efforts do not necessarily involve information from multiple sources.***

According to GAO's survey of 128 federal agencies, 199 data mining efforts are either operational or in the planning stages. Of those, 122, or approximately 61%, involve personal information. An example of a data mining effort that not does not involve personal information is a NASA (National Aeronautics and Space Administration) program that mines large earth data sets to find patterns and relationships to detect hidden events.

The most common federal uses of data mining programs identified by GAO are for:

- Improving service or performance (65 data mining programs)
- Detecting fraud, waste and abuse (24)
- Analyzing scientific and research information (23)
- Managing human resources (17)
- Detecting criminal activities or patterns (15)
- Analyzing intelligence and detecting terrorist activities (14).

Of the data mining programs using personal information, the most common purposes of those programs are improving service or performance (33 programs) and detecting fraud, waste and abuse (24). A smaller number of projects were used to manage human resources (15), detect criminal activities or patterns (15), analyze intelligence and detect terrorist activities (10), and increase tax compliance (7).

Regarding the use of information from the private sector or other federal agencies, only 54 of the federal data mining efforts use data from the private sector and only 36 of those projects use personal information. Seventy-seven of the total 199 data mining efforts use data from other federal agencies and 46 of those efforts involve personal information.

The GAO report ultimately concluded that data mining allows agencies to analyze massive volumes of data and that data mining is increasingly being used for a variety of purposes that range from service or performance improvement to the detection of terrorist activities or patterns. GAO stated that it plans to examine selected data mining efforts and their implications.[13]

[12] U.S. General Accounting Office (GAO), "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548, May 2004.

[13] Ibid.

State Government Privacy Implications

Since data mining efforts are emerging at the federal level for a broad range of uses, this may serve as an indicator for how data mining will continue to emerge at the state level. Some states are using data mining to analyze electronic tax filings to discover patterns with the goal of increasing tax compliance. In pursuing these and other similar projects, states should consider the following to avoid privacy problems encountered at the federal level with TIA, CAPPS II and MATRIX:

- ***Be transparent early-on about a data mining or aggregation project's purpose***
- ***Build privacy protections into data mining and aggregation technologies at the beginning.***

NASCIO-Identified Best Practices: The following considerations may assist states in pursuing data mining and aggregation projects that protect individual privacy. See also Appendix B for a 19-item data mining checklist developed by the Technology and Privacy Advisory Committee (TAPAC).[14]

- Clearly state up-front the state business benefits that will be achieved by data mining
- Educate on what data mining and aggregation are and on their privacy implications
- Build-in privacy considerations up-front in a data mining or aggregation project
- Bring in all stakeholders at the beginning, including privacy advocates, to get input
- Clearly state the primary purpose of the project and, if possible, identify possible secondary purposes that might emerge in the future
- Ensure that any new purpose of a project is consistent with the project's original purpose
- Determine whether personal information will be involved in a project
- Provide notice to individuals of the collection and use of their personal information
- Determine whether an individual should have a choice in the collection of information
- Determine whether a project will involve information from other governmental agencies or from the private sector and, if so, provide notice of the combining of the information
- Ensure the accuracy of data entry, cleansing and standardization to improve the quality of inferences from the mined data and make the correction of such data easier
- Implement role-based permissions granting access only to those with a need to know
- Where appropriate, anonymize personal information
- Create audit requirements for each query of mined data and its justification
- Maintain oversight of data mining or aggregation projects

[14] TAPAC, Letter from Newton N. Minnow, Chairman of TAPAC, to the Hon. Donald H. Rumsfeld, Secretary of Defense, *Safeguarding Privacy in the Fight Against Terrorism*, March 2004.

- Limit the actions that may be taken as a result of unverified findings from data mining
- Create a system where individuals can ensure that any incorrect personal information can be corrected to avoid repeated "false positives"
- Destroy or anonymize data that no longer serves the original purpose of the project.

What CIOs Need to Know

Educate agencies, legislators, stakeholders, privacy advocates and the public upfront about any projects that entail data mining or aggregation or that could be perceived as including those types of activities. Remember that not all such projects involve personal information or are for terrorism-related purposes.

Bake privacy into any new data mining or aggregation technologies during their development so that you will be able to clearly state how citizen privacy will be protected. The "Fair Information Principles" provide good guidance.

Need to Dig Deeper? Additional Data Mining and Aggregation Resources

U.S. Government Accountability Office (GAO), "Medicaid Program Integrity: State and Federal Efforts to Prevent and Detect Improper Payments," GAO-04-707, July 2004, <http://www.gao.gov/atext/d04707.txt>.

U.S. General Accounting Office (GAO), "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548, May 2004, <http://www.gao.gov/new.items/d04548.pdf>.

Kurt Thearling, Ph.D., "An Introduction to Data Mining," September 2004, <http://www.thearling.com/dmintro/dmintro.htm>.

National Association of State Auditors, Comptrollers and Treasurers (NASACT), "Using Data Mining to Make Audits More Efficient and Effective," 2004 National State Auditors Association (NSAA) Middle Management Conference Presentations, <http://www.nasact.org/onlineresources/downloads/2004_MM/2004_MM.htm>.

Appendix A: Background on TIA, CAPPS II & MATRIX

The following provides some background on the purpose of these programs and the privacy concerns that they have raised.

TIA (Terrorism Information Awareness) Program: TIA was created under the Defense Advanced Research Projects Agency (DARPA), which is a central research and development organization for the U.S. Department of Defense. After the September 11, 2001 terrorist attacks, DARPA created TIA to research and develop an experimental prototype network in order to combat terrorism through better decision-making. TIA was to include data searching and pattern recognition tools, which raised public concern that TIA would be used to create dossiers on U.S. citizens. In September 2003, Congress terminated funding for TIA with an exception for TIA's "processing, analysis, and collaboration tools for counter-terrorism foreign intelligence." Two factors that may have contributed to TIA's demise were (1) DARPA's failure to clearly articulate TIA, its objectives and the data to which it would apply in a clear, consistent and coherent manner and (2) DARPA's failure to build privacy protections into TIA technologies at the development stage.[15]

CAPPS II (Computer Assisted Passenger Prescreening System): The Department of Homeland Security (DHS) described CAPPS II as "a limited, automated prescreening system authorized by Congress in the wake of the September 11, 2001 terrorist attacks." Its purpose was to "authenticate travelers' identities and perform risk assessments to detect individuals who may pose a terrorist-related threat or who have outstanding federal or state warrants for crimes of violence." CAPPS II would have asked passengers for an expanded amount of reservation information, including name, birth date, home address and phone number. With that information, the system would have verified passenger identity, performed a risk assessment using commercial data and current intelligence, and then would have categorized passengers as "no risk," "unknown," or "elevated or high risk." A passenger's boarding pass would have included an encoded message indicating the appropriate screening level.[16]

However, in 2004, GAO determined that the Transportation Security Administration (TSA) had not addressed a number of key areas of interest to Congress regarding CAPPS II, including privacy concerns that had been raised about the system. Among the concerns was TSA's failure to provide reasons for its proposed rule to exempt the system from the requirements of the Privacy Act of 1974, the law that governs the privacy of personal information held within federal systems of records. Other concerns included the possibility that more personal information than necessary would be collected and maintained, that such information could be used for new purposes in the future, and that passengers could not review all information about them that would have been accessed via the CAPPS II system. Though such concerns did not rise to the level of violations of law, they reflected a limit on the application of some of "the Fair Information Principles."

[15] Technology and Privacy Advisory Committee, (TAPAC),"Safeguarding Privacy in the Fight Against Terrorism," March 2004, <http://www.sainc.com/tapac/TAPAC_Report_Final_5-10-04.pdf>.

[16] U.S. Department of Homeland Security, "Fact Sheet: CAPPS II at a Glance," February 12, 2004, <http://www.dhs.gov/dhspublic/display?theme=43&content=3162&print=true>.

GAO commented that the limited application of "the Fair Information Principles" in those instances resulted from TSA's attempt to balance privacy with other public policy interests, including national security, and that policymakers would have the final determination as to whether TSA's balance was appropriate.[17]

MATRIX (Multistate Anti-Terrorism Information Exchange): MATRIX emerged out of the September 11 attacks as a pilot project to leverage "proven technology to assist criminal investigations by implementing factual data analysis from existing data sources and integrating disparate data from many types of Web-enabled storage systems." The pilot's purpose is to make it more efficient for law enforcement to access information such as criminal history records, drivers' license data, vehicle registrations, corrections records, and other public data and to help law enforcement analyze terrorist activities and other crimes for investigative leads.[18] States participating in MATRIX submit data to the system to be included in the factual data analysis and provide updates to ensure the information is current.[19] The types of information submitted by the participating states are governed by each state's laws on access to public records. The MATRIX pilot emphasizes that the information available through the pilot has been accessible to law enforcement for many years, but that it will make access more efficient. The MATRIX website provides that no new information is collected via the pilot and that no criminal intelligence databases are being connected.[20] Lastly, its website states that MATRIX is not "a data mining application."[21] Privacy advocates, though, have raised concerns that MATRIX is a data mining application[22] and that it creates dossiers on private citizens from government and private sector databases.[23] Originally, thirteen states participated in MATRIX. However, a number have since withdrawn from the pilot that now has five participating states (Connecticut, Florida, Michigan, Ohio and Pennsylvania).[24]

[17] U.S. General Accounting Office (GAO), "Computer-Assisted Passenger Prescreening System Faces Significant Implementation Challenges," GAO-04-385, February 2004, <http://www.gao.gov/new.items/d04385.pdf>.
[18] MATRIX, "MATRIX Defined," September 2004, <http://www.matrix-at.org/matrix_defined.htm>.
[19] MATRIX, "Roles Defined," September 2004, <http://www.matrix-at.org/roles.htm>.
[20] MATRIX, "MATRIX Defined," September 2004.
[21] MARTIX, "MATRIX Misconceptions," September 2004, <http://www.matrix-at.org/misconceptions.htm>
[22] "New Documents Obtained by ACLU Raise Troubling Questions About Matrix Program," American Civil Liberties Union (ACLU), May 2004, <http://www.aclu.org/Privacy/Privacy.cfm?ID=15830&c=130>.
[23] "What is The Matrix? ACLU Seeks Answers on New State-Run Surveillance Program," American Civil Liberties Union (ACLU), October 2003, <http://www.aclu.org/Privacy/Privacy.cfm?ID=14257&c=130>.
[24] "New Documents Obtained by ACLU Raise Troubling Questions About Matrix Program," American Civil Liberties Union (ACLU), May 2004.

Appendix B: TAPAC Data Mining Checklist

TAPAC (Technology and Privacy Advisory Committee) was appointed by the Secretary of the Department of Defense (DoD) in February 2002 to ensure that DoD complies with U.S. law and adheres to "American values related to privacy."[25] You can view the TAPAC report at: http://www.sainc.com/tapac. Please find the checklist below.

The Existence and Purpose of Data Mining

1. Is the proposed activity or system likely to involve the acquisition, use, or sharing of personally identifiable information about U.S. persons?
2. What purpose(s) does the data mining serve? Is it lawful? Is it within the agency's authority? Is it sufficiently important to warrant the risks to informational privacy that data mining poses?
3. Is data mining necessary to accomplish that purpose—i.e., could the purpose be accomplished as well without data mining?
4. Is the data mining tool designed to access, use, retain, and disseminate the least data necessary to serve the purposes for which it is intended?
5. Is the data mining tool designed to use anonymized data whenever possible?

Data Mining Personally Identifiable Information

6. Are there specific and articulable facts that data mining personally identifiable information (or reidentifying previously anonymized information) concerning U.S. persons will be conducted in a manner that otherwise complies with the requirements of applicable laws and recommendations; is reasonably related to identifying or apprehending terrorists, preventing terrorist attacks, or locating or preventing the use of weapons of mass destruction; is likely to yield information relevant to national security; and is not practicable with anonymized data?

The Sources and Nature of Data Concerning U.S. Persons

7. Are the data appropriate for their intended use, taking into account the purpose(s) for which the data were collected, their age, and the conditions under which they have been stored and protected?
8. Are data being accessed or acquired from third parties in violation of the terms and conditions (usually reflected in a privacy policy) under which they were collected?
9. If data are being acquired directly from data subjects, have the individuals been provided with appropriate notice, consistent with the purpose of the data mining activity?
10. Are data being sought in the order provided by Executive Order 12333—i.e., from or with the consent of the data subject, from publicly available sources, through a method requiring authorization less than probable cause (e.g., a pen register or trap and trace device), through a method requiring a warrant, and finally through a method requiring a wire tap?

[25] TAPAC, Letter from Newton N. Minnow, Chairman of TAPAC, to the Hon. Donald H. Rumsfeld, Secretary of Defense, *Safeguarding Privacy in the Fight Against Terrorism*, March 2004.

11. Are personally identifiable data being left in place whenever possible? If such data are being acquired or transferred, is there a system in place for ensuring that they are returned or destroyed as soon as possible?

The Impact of Data Mining

12. What are the likely effect(s) on individuals identified through the data mining—i.e., will they be the subject of further investigation or will they be immediately subject to some adverse action?
13. Does the data mining tool yield a rate of false positives that is acceptable in view of the purpose of the search, the severity of the effect of being identified, and the likelihood of further investigation?
14. Is there an appropriate system in place for dealing with false positives (e.g., reporting false positives to developers to improve the system, correcting incorrect information if possible, etc.), including identifying the frequency and effects of false positives?

Oversight of Data Mining

15. Are data secured against accidental or deliberate unauthorized access, use, or destruction, and access to the data mining tool restricted to persons with a legitimate need and protected by appropriate access controls taking into account the sensitivity of the data?
16. Does the data mining tool generate, to the extent technologically possible, an immutable audit trail showing which data have been accessed or transferred, by what users, and for what purpose?
17. Will the data mining tool be subject to continual oversight to ensure that it is used appropriately and lawfully, and that informational privacy issues raised by new developments or discoveries are identified and addressed promptly?
18. Are all persons engaged in developing or using data mining tools trained in their appropriate use and the laws and regulations applicable to their use?
19. Have determinations as to the efficacy and appropriateness of data mining been made or reviewed by an official other than those intimately involved with the development, acquisition, or use of the data mining tool?