United States General Accounting Office

GAO

Report to the Ranking Minority Member, Subcommittee on Financial Management, the Budget, and International Security, Committee on Governmental Affairs, U.S. Senate

May 2004

DATA MINING

Federal Efforts Cover a Wide Range of Uses

GAO-04-548

May 2004

DATA MINING

Federal Efforts Cover a Wide Range of Uses

GAO Accountability • Integrity • Reliability

Highlights

Highlights of GAO-04-548, a report to the Ranking Minority Member, Subcommittee on Financial Management, the Budget, and International Security, Committee on Governmental Affairs, U.S. Senate

Why GAO Did This Study

Both the government and the private sector are increasingly using "data mining"—that 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. As has been widely reported, many federal data mining efforts involve the use of personal information that is mined from databases maintained by public as well as private sector organizations.

GAO was asked to survey data mining systems and activities in federal agencies. Specifically, GAO was asked to identify planned and operational federal data mining efforts and describe their characteristics.

www.gao.gov/cgi-bin/getrpt?GAO-04-548

To view the full product, including the scope and methodology, click on the link above. For more information, contact Linda Koontz at (202) 512-6240 or koontzl@gao.gov.

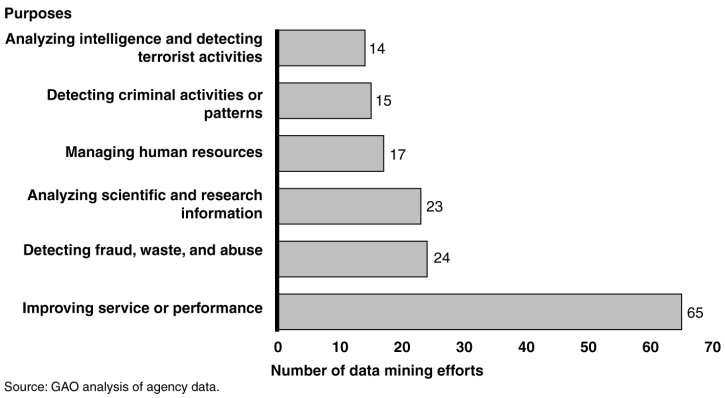
What GAO Found

Federal agencies are using data mining for a variety of purposes, ranging from improving service or performance to analyzing and detecting terrorist patterns and activities. Our survey of 128 federal departments and agencies on their use of data mining shows that 52 agencies are using or are planning to use data mining. These departments and agencies reported 199 data mining efforts, of which 68 are planned and 131 are operational. The figure here shows the most common uses of data mining efforts as described by agencies. Of these uses, the Department of Defense reported the largest number of efforts aimed at improving service or performance, managing human resources, and analyzing intelligence and detecting terrorist activities. The Department of Education reported the largest number of efforts aimed at detecting fraud, waste, and abuse. The National Aeronautics and Space Administration reported the largest number of efforts aimed at analyzing scientific and research information. For detecting criminal activities or patterns, however, efforts are spread relatively evenly among the agencies that reported having such efforts.

In addition, out of all 199 data mining efforts identified, 122 used personal information. For these efforts, the primary purposes were improving service or performance; detecting fraud, waste, and abuse; analyzing scientific and research information; managing human resources; detecting criminal activities or patterns; and analyzing intelligence and detecting terrorist activities.

Agencies also identified efforts to mine data from the private sector and data from other federal agencies, both of which could include personal information. Of 54 efforts to mine data from the private sector (such as credit reports or credit card transactions), 36 involve personal information. Of 77 efforts to mine data from other federal agencies, 46 involve personal information (including student loan application data, bank account numbers, credit card information, and taxpayer identification numbers).

Top Six Purposes of Data Mining Efforts in Departments and Agencies



Source: GAO analysis of agency data.

Contents

Figures

Abbreviations

CARDS	Counterintelligence Analytical Research Data System
CG	Coast Guard
CI-AIMS	Counterintelligence Automated Investigative Management System
DHHS	Department of Health and Human Services
DOD	Department of Defense
DOE	Department of Energy
DOT	Department of Transportation
EFTPS	Electronic Federal Tax Payment System
EOS	Earth Observing System
FARS	Fatality Analysis Reporting System
FDA	Food and Drug Administration
GENESIS	Global Environmental and Earth Science Information System
GSFC	Goddard Space Federal Center
HR	Human Resources
HRSA	Health Resources and Services Administration
MATRIX	Multistate Anti-terrorism Information Exchange System
NASA	National Aeronautics and Space Administration
NVO	National Virtual Observatory
OIG	Office of Inspector General
OLAP	On-line Analytical Processing
RSST	Real Estate Stress Test
SAA	Spectral Analysis Automation
SAS	Safety Automated System
SMARTS	Statistical Management Analysis and Reporting Tool System
SWC	Space Warfare Center
TIMS	Technical Information Management System
TOP	Treasury Offset Program
VA	Veterans Affairs
VHA	Veterans Health Administration
VISN	Veterans Integrated Service Network

G A O
Accountability * Integrity * Reliability

United States General Accounting Office
Washington, D.C. 20548

May 4, 2004

The Honorable Daniel K. Akaka
Ranking Minority Member
Subcommittee on Financial Management, the Budget,
and International Security
Committee on Governmental Affairs
United States Senate

Dear Senator Akaka:

Data mining—a technique for extracting knowledge from large volumes of data—is increasingly being used by government and by the private sector. As has been widely reported, many federal data mining efforts involve the use of personal information[1] that is mined from public as well as private sector organizations.

This report responds to your request that we identify and describe operational and planned data mining systems and activities in federal agencies. In a follow-up report, we plan to perform an in-depth review of selected federal data mining efforts.

The term "data mining" has a number of meanings. For purposes of this work, we define data mining as 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. We based this definition on the most commonly used terms found in a survey of the technical literature. In our initial survey of chief information officers, these officials found the definition sufficient to identify agency data mining efforts.

[1]As used in this report, personal information is all information associated with an individual and includes both identifying information and nonidentifying information. Identifying information, which can be used to locate or identify an individual, includes name, aliases, Social Security number, e-mail address, driver's license number, and agency-assigned case number. Nonidentifying personal information includes age, education, finances, criminal history, physical attributes, and gender.

To address our objective to identify and describe operational and planned data mining systems and activities in federal agencies, we surveyed chief information officers or comparable officials at 128 federal departments and agencies to determine whether the agencies had operational and planned data mining systems or activities.[2] We then conducted telephone interviews with the reported system managers to obtain information on the characteristics of the identified data mining efforts. To verify the information we received, we sent follow-up letters to agencies that responded as well as to those that did not respond, we asked responsible officials to verify the information, and we performed random assessments of the means that these officials used to verify the information.

In addition, we conducted a search of technical literature and periodicals to develop a comprehensive list of federal government data mining efforts and then compared these efforts with data mining efforts reported by federal agencies. If the data mining efforts on our lists were not reported on the survey, we contacted the appropriate chief information officers and, with their concurrence, added the efforts.

We performed our work from May 2003 to April 2004 in accordance with generally accepted government auditing standards. Additional details on our scope and methodology are provided in appendix I.

Results in Brief

Federal agencies are using data mining for a variety of purposes, ranging from improving service or performance to analyzing and detecting terrorist patterns and activities. Our survey of 128 federal departments and agencies on their use of data mining shows that 52 agencies are using or are planning to use data mining. These departments and agencies reported 199 data mining efforts, of which 68 were planned and 131 were operational. The most common uses of data mining efforts were described by agencies as

- improving service or performance;
- detecting fraud, waste, and abuse;
- analyzing scientific and research information;

[2]That is, we asked about both systems explicitly dedicated to data mining and activities using automated tools to "mine" databases that are part of other systems. In this report, we use the word "efforts" to refer to both systems and activities, unless otherwise specified.

- managing human resources;
- detecting criminal activities or patterns; and
- analyzing intelligence and detecting terrorist activities.

The Department of Defense reported having the largest number of data mining efforts aimed at improving service or performance and at managing human resources. Defense was also the most frequent user of efforts aimed at analyzing intelligence and detecting terrorist activities, followed by the Departments of Homeland Security, Justice, and Education.

The Department of Education reported the largest number of efforts aimed at detecting fraud, waste, and abuse, while the National Aeronautics and Space Administration targets most of their data mining efforts (21 out of 23) toward analyzing scientific and research information. Data mining efforts for detecting criminal activities or patterns, however, were spread relatively evenly among the reporting agencies.

In addition, out of all 199 data mining efforts identified, 122 used personal information. For these efforts, the primary purposes were detecting fraud, waste, and abuse; detecting criminal activities or patterns; analyzing intelligence and detecting terrorist activities; and increasing tax compliance.

Agencies also identified efforts to mine data from the private sector and data from other federal agencies, both of which could include personal information. Of 54 efforts to mine data from the private sector (such as credit reports or credit card transactions), 36 involve personal information. Of 77 efforts to mine data from other federal agencies, 46 involve personal information (including student loan application data, bank account numbers, credit card information, and taxpayer identification numbers).

Background

Data mining enables corporations and government agencies to analyze massive volumes of data quickly and relatively inexpensively. The use of this type of information retrieval has been driven by the exponential growth in the volumes and availability of information collected by the public and private sectors, as well as by advances in computing and data storage capabilities. In response to these trends, generic data mining tools are increasingly available for—or built into—major commercial database applications. Today, mining can be performed on many types of data,

including those in structured, textual, spatial, Web, or multimedia forms. Data mining is becoming a big business; Forrester Research has estimated that the data mining market is passing the billion dollar mark.

Although the use and sophistication of data mining have increased in both the government and the private sector, data mining remains an ambiguous term. According to some experts, data mining overlaps a wide range of analytical activities, including data profiling, data warehousing, online analytical processing, and enterprise analytical applications.[3] Some of the terms used to describe data mining or similar analytical activities include "factual data analysis" and "predictive analytics." We surveyed technical literature and developed a definition of data mining based on the most commonly used terms found in this literature. Based on this search, we define data mining as 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. We used this definition in our initial survey of chief information officers; these officials found the definition sufficient to identify agency data mining efforts.

Data mining has been used successfully for a number of years in the private and public sectors in a broad range of applications. In the private sector, these applications include customer relationship management, market research, retail and supply chain analysis, medical analysis and diagnostics, financial analysis, and fraud detection. In the government, data mining was initially used to detect financial fraud and abuse. For example, data mining has been an integral part of GAO audits and investigations of federal government purchase and credit card programs.[4] Data mining and related technologies are also emerging as key tools in Department of Homeland Security initiatives.

[3]Lou Agosta, "Data Mining Is Dead—Long Live Predictive Analytics!" (Forrester Research, Oct. 30, 2003), http://www.forrester.com/Research/LegacyIT/0,7208,33030,00.html (downloaded Jan. 26, 2004).

[4]For more information on the uses of data mining in GAO audits, see U.S. General Accounting Office, *Data Mining: Results and Challenges for Government Programs, Audits, and Investigations*, GAO-03-591T (Washington, D.C: Mar. 25, 2003).

Data Mining Poses Privacy Challenge

Since the terrorist attacks of September 11, 2001, data mining has been seen increasingly as a useful tool to help detect terrorist threats by improving the collection and analysis of public and private sector data. In a recent report on information sharing and analysis to address the challenges of homeland security, it was noted that agencies at all levels of government are now interested in collecting and mining large amounts of data from commercial sources.[5] The report noted that agencies may use such data not only for investigations of known terrorists, but also to perform large-scale data analysis and pattern discovery in order to discern potential terrorist activity by unknown individuals. Such use of data mining by federal agencies has raised public and congressional concerns regarding privacy.

One example of a large-scale development effort launched in the wake of the September 11 attacks is the Multistate Anti-terrorism Information Exchange System, known as MATRIX. MATRIX, currently used in five states,[6] provides the capability to store, analyze, and exchange sensitive terrorism-related and other criminal intelligence data among agencies within a state, among states, and between state and federal agencies. Information in MATRIX databases includes criminal history records, driver's license data, vehicle registration records, incarceration records, and digitized photographs. Public awareness of MATRIX and of similar large-scale data mining or data mining-like projects has led to concerns about the government's use of data mining to conduct a mass "dataveillance"[7]—a surveillance of large groups of people—to sift through vast amounts of personally identifying data to find individuals who might fit a terrorist profile.

[5]*Creating a Trusted Information Network for Homeland Security* (New York City: The Markle Foundation, December 2003), http://www.markletaskforce.org/Report2_Full_Report.pdf (downloaded Mar. 8, 2004).

[6]Five states are currently participating in the MATRIX pilot project: Connecticut, Florida, Michigan, Ohio, and Pennsylvania.

[7]Roger Clarke, "*Information Technology and Dataveillance*," *Communications of the ACM*, vol. 31, issue 5 (New York City: ACM Press, May 1988), http://www.anu.edu.au/people/Roger.Clarke/DV/CACM88.html (downloaded Mar. 5, 2004). Clarke defines mass dataveillance as the systematic use of personal data systems in the investigation or monitoring of the actions or communications of groups of people.

Mining government and private databases containing personal information creates a range of privacy concerns. Through data mining, agencies can quickly and efficiently obtain information on individuals or groups by exploiting large databases containing personal information aggregated from public and private records. Information can be developed about a specific individual or about unknown individuals whose behavior or characteristics fit a specific pattern. Before data aggregation and data mining came into use, personal information contained in paper records stored at widely dispersed locations, such as courthouses or other government offices, was relatively difficult to gather and analyze. As one expert noted, data mining technologies that provide for easy access and analysis of aggregated data challenge the concept of privacy protection afforded to individuals through the inherent inefficiency of government agencies analyzing paper, rather than aggregated, computer records.[8]

Privacy concerns about mined or analyzed personal data also include concerns about the quality and accuracy of the mined data; the use of the data for other than the original purpose for which the data were collected without the consent of the individual; the protection of the data against unauthorized access, modification, or disclosure; and the right of individuals to know about the collection of personal information, how to access that information, and how to request a correction of inaccurate information.[9]

[8]K.A. Taipale, "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data," *The Columbia Science and Technology Law Review*, vol. V, 2003-2004 (New York City: Columbia Law School, 2004), http://www.stlr.org/cite.cgi?volume=5&article=2 (downloaded Mar. 18, 2004).

[9]These privacy concerns are reflected in the Fair Information Practices proposed in 1980 by the Organization for Economic Cooperation and Development and endorsed by the U.S. Department of Commerce in 1981. These practices govern collection limitation, purpose specification, use limitation, data quality, security safeguards, openness, individual participation, and accountability.

Agencies Identified Numerous Data Mining Efforts with Various Aims

Of 128 federal departments and agencies surveyed for information on their planned and operational data mining efforts (listed in app. II), 52 agencies reported 199 data mining efforts, and 69 agencies reported that they were not engaged in data mining and were not planning such efforts (listed in app. III). Of the 199 data mining efforts, 68 were planned and 131 were operational. Seven agencies did not respond to our survey.[10] Appendix IV lists the 199 data mining efforts reported, along with key characteristics.

Agencies described the most common purposes of data mining efforts as

- improving service or performance;
- detecting fraud, waste, and abuse;
- analyzing scientific and research information;
- managing human resources;
- detecting criminal activities or patterns; and
- analyzing intelligence and detecting terrorist activities.

As shown in table 1, the Department of Defense reported the largest number of efforts aimed at improving service or performance (with 19 out of 65 reported efforts) and at managing human resources (with 14 out of 17 efforts). Defense was also the most frequent user of efforts aimed at analyzing intelligence and detecting terrorist activities, with 5 of 14 efforts, followed by the Departments of Homeland Security and Justice, with 4 and 3 efforts, respectively. The Department of Education has the largest number of efforts aimed at detecting fraud, waste, and abuse (9 out of 24 efforts reported). The National Aeronautics and Space Administration accounts for 21 of the 23 identified efforts for analyzing scientific and research information. Efforts are spread relatively evenly among the agencies that reported using data mining efforts for detecting criminal

[10]Agencies that did not respond to our survey are (1) the Central Intelligence Agency; (2) the Corporation for National and Community Services; (3) the Department of Army, Department of Defense; (4) the Equal Employment Opportunity Commission; (5) the National Park Service, Department of the Interior; (6) the National Security Agency, Department of Defense; and (7) the Rural Utilities Service, Department of Agriculture.

activities or patterns. Table 1 summarizes the top six uses of data mining efforts among the responding agencies.

Table 1: Top Six Purposes of Data Mining Efforts in Departments and Agencies and Number of Efforts Reported

Department or agency	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
Department of Agriculture	8	1				
Department of Commerce						
Department of Defense	19	1	1	14	1	5
Department of Education	6	9			3	1
Department of Energy					3	
Department of Health and Human Services	4		1			1
Department of Homeland Security	5			2	2	4
Department of the Interior	1					
Department of Justice	1			1	3	3
Department of Labor	3	1				
Department of State		2				
Department of Transportation		1				
Department of the Treasury	4	1			2	
Department of Veterans Affairs	5	5			1	
Environmental Protection Agency		1				
Export-Import Bank of the United States	1					
Federal Deposit Insurance Corporation	1					
Federal Reserve System		1				
National Aeronautics and Space Administration	1	1	21			
Nuclear Regulatory Commission	1					
Office of Personnel Management	1					
Pension Benefit Guaranty Corporation	2					
Railroad Retirement Board	1					
Small Business Administration	1					
Total	**65**	**24**	**23**	**17**	**15**	**14**

Source: GAO analysis of agency-provided data.

Some data mining purposes focus on human activities and therefore are inherently likely to involve personal information; examples of these purposes are detecting fraud, waste, and abuse; detecting criminal activities or patterns; managing human resources; and analyzing intelligence. The following are examples of data mining efforts for each of these purposes:

- *Detecting fraud, waste, and abuse.* The Veterans Benefits Administration's C & P Payment Data Analysis effort mines veterans' compensation and pension data for evidence of fraud.

- *Detecting criminal activities or patterns.* The Department of Education's Title IV Identity Theft Initiative effort focuses on identity theft cases involving education loans.

- *Managing human resources.* The U.S. Air Force's Oracle HR (Human Resources) uses data mining to provide information on promotions, pay grades, clearances, and other information relevant to human resources planning.

- *Analyzing intelligence and detecting terrorist activities.* The Defense Intelligence Agency's Verity K2 Enterprise mines data from the intelligence community and Internet sources to identify foreign terrorists or U.S. citizens connected to foreign terrorism activities.

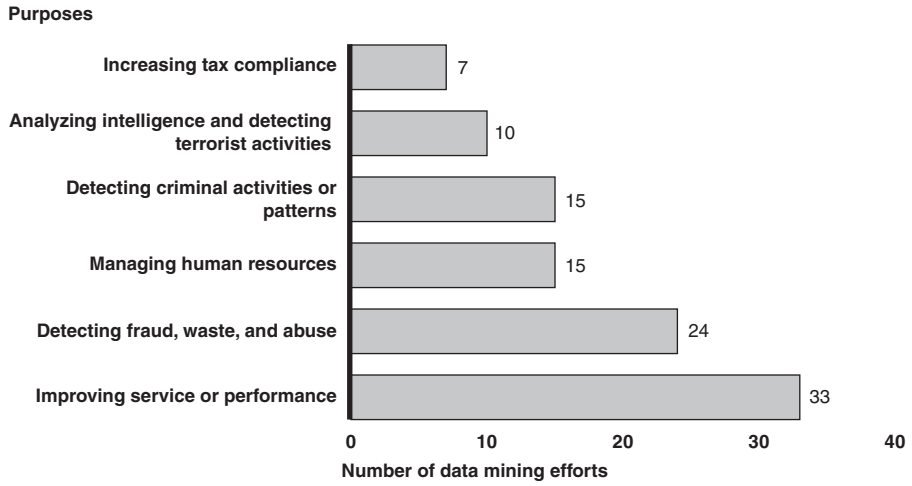
On the other hand, other categories of efforts do not necessarily focus on human activities or involve personal information, such as many of the efforts aimed at analyzing scientific and research information. The National Aeronautics and Space Administration, for example, mines large, complex earth science data sets to find patterns and relationships to detect hidden events (the system is called Machine Learning and Data Mining for Improved Data Understanding of High Dimensional Earth Sensed Data).

Similarly, many efforts aimed at improving service or performance (the most frequently cited purpose of data mining efforts) do not involve personal information. For example, the Department of the Navy's Supply Management System Multidimensional Cubes system includes a data warehouse containing data on every ship part that has been ordered since the 1980s, with multidimensional information on each part. The Navy uses data mining to calculate failure rates and identify needed improvements; according to the Navy, this system reduces downtime on ships by improving parts replacement.

However, some efforts aimed at improving service or performance do involve personal information. For example, the Veterans Administration's VISN (Veterans Integrated Service Network) 16 Data Warehouse is mined for a variety of information, including patient visits, laboratory tests, and pharmacy records, to provide management with health care system performance information.

Overall, 122 of the 199 data mining efforts involve personal information. Figure 1 shows the top six purposes of these efforts, as well as their distribution.

Figure 1: Top Six Purposes of Data Mining Efforts That Involve Personal Information



Source: GAO analysis of agency data.

Of the 199 data mining efforts, 54 use or plan to use data from the private sector. Of these, 36 involve personal information. The personal information from the private sector included credit reports and credit card transaction records. Figure 2 shows the distribution of the top six purposes of the 54 efforts involving data from the private sector.

Figure 2: Top Six Purposes of Data Mining Efforts That Involve Private Sector Data

Source: GAO analysis of agency data.

Of the 199 data mining efforts, 77 efforts use or plan to use data from other federal agencies. Of the 77 efforts, 46 involve personal information. The personal information from other federal agencies included student loan application data, bank account numbers, credit card information, and taxpayer identification numbers. Figure 3 shows the top six uses for the 77 efforts involving data from other federal agencies and their distribution.

Figure 3: Top Six Purposes of Data Mining Efforts That Involve Data from Other Federal Agencies

Source: GAO analysis of agency data.

Summary

Driven by advances in computing and data storage capabilities and by growth in the volumes and availability of information collected by the public and private sectors, data mining enables government agencies to analyze massive volumes of data. Our survey shows that data mining is increasingly being used by government for a variety of purposes, ranging from improving service or performance to analyzing and detecting terrorist patterns and activities.

Although this survey provides a broad overview of the emerging uses of data mining in the federal government, more work is needed to shed light on the privacy implications of these efforts. In future work, we plan to examine selected federal data mining efforts and their implications.

As agreed with your office, unless you publicly announce the contents of the report earlier, we plan no further distribution until 30 days from the report date. At that time, we will send copies of this report to the Chairmen and Ranking Minority Members of the House Committee on Government Reform; Subcommittee on Civil Service and Agency Organization, House Committee on Government Reform; Select Committee on Homeland Security, House of Representatives; Senate Committee on Governmental

Affairs; and the Subcommittee on Oversight of Government Management, the Federal Workforce and the District of Columbia, Senate Committee on Governmental Affairs. We will also make copies available to others on request. In addition, this report will be available at no charge on the GAO Web site at http://www.gao.gov.

If you have any questions concerning this report, please call me at (202) 512-6240 or Mirko J. Dolak, Assistant Director, at (202) 512-6362. We can also be reached by e-mail at koontzl@gao.gov and dolakm@gao.gov, respectively. Key contributors to this report were Camille M. Chaires, Barbara S. Collier, Orlando O. Copeland, Nancy E. Glover, Stuart M. Kaufman, Lori D. Martinez, Morgan F. Walts, and Marcia C. Washington.

Sincerely yours,

Linda D. Koontz
Director, Information Management Issues

Objective, Scope, and Methodology

Our objective was to identify and describe planned and operational federal data mining efforts. As a first step in addressing this objective, we developed a definition of "data mining." Because this expression has a range of meanings, we surveyed the technical literature to develop a definition based on the most commonly used terms found in this literature. We defined data mining as 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. In our initial survey of chief information officers, these officials found the definition sufficient to identify agency data mining efforts.

We then surveyed chief information officers or comparable officials at 128 federal departments and agencies (see app. II) and asked them to identify whether their agency had operational and planned data mining efforts. We achieved a 95 percent response rate. Of the 121 agencies that responded, 69 reported that they did not have any data mining efforts (see app. III). We followed up with these 69 agencies and gave them another opportunity to report data mining efforts.

To obtain information on the characteristics of the identified operational or planned data mining efforts, we conducted structured telephone interviews[1] with the identified system owners or activity managers. The interviews were designed to obtain detailed information about each data mining system, including the purpose and size, the use of personal information, and the use of data from the private sector or other federal organizations. We pretested the structured interview to ensure relevance and clarity.

We aggregated these data by agency and sent them back to the chief information officer, comparable official, or their designee and asked that they review the characteristics for completeness and accuracy. One of the 52 departments and agencies that reported data mining systems—the Department of Homeland Security—has not responded to our request to review the reported data for completeness and accuracy.

[1]In a structured interview, the interviewer asks the same questions of numerous individuals or individuals representing numerous organizations in a precise manner, offering each interviewee the same set of possible responses.

We performed random assessments of the means that these officials used to verify the information. Based on these assessments, we concluded that the agencies' verification methods were reasonable and that as a result, we could rely on the accuracy of the reported data. We also conducted a search of technical literature and periodicals to develop a list of federal government data mining efforts and then compared the efforts on this list with the data mining efforts reported by federal agencies. If the data mining efforts on our list were not reported on the survey, we contacted the chief information officer or comparable official to determine whether that data mining effort should be included in our survey.

Because this was not a sample survey, there are no sampling errors. However, the practical difficulties of conducting any survey may introduce errors, commonly referred to as nonsampling errors. For example, difficulties in how a particular question is interpreted, in the sources of information that are available to respondents, or in how the data are entered into a database or were analyzed can introduce unwanted variability into the survey results. We took steps in the development of the structured interview, the data collection, and the data analysis to minimize these nonsampling errors. Among these steps, we pretested the structured interview instrument, contacted nonresponding agencies as well as agencies not identifying data mining efforts, and sent the aggregated data to the agency chief information officer for review.

We conducted our work from May 2003 to April 2004 in accordance with generally accepted government auditing standards.

Surveyed Departments and Agencies

Department of Agriculture

- Agricultural Marketing Service
- Agricultural Research Service
- Animal and Plant Health Inspection Service
- Cooperative State Research, Education, and Extension Service
- Farm Service Agency
- Food and Nutrition Service
- Food Safety and Inspection Service
- Foreign Agricultural Service
- Forest Service
- National Agricultural Statistics Service
- Natural Resources Conservation Service
- Risk Management Agency
- Rural Utilities Service

Department of Commerce

- Bureau of the Census
- Economic Development Administration
- International Trade Administration
- National Oceanic and Atmospheric Administration
- U.S. Patent and Trademark Office

Department of Defense

- Missile Defense Agency
- Defense Advanced Research Projects Agency
- Defense Commissary Agency
- Defense Contract Audit Agency
- Defense Contract Management Agency
- Defense Information Systems Agency
- Defense Intelligence Agency
- Defense Legal Services Agency
- Defense Logistics Agency
- Defense Security Cooperation Agency
- Defense Security Service
- Defense Threat Reduction Agency
- Department of the Air Force
- Department of the Army
- Department of the Navy
- National Geospatial-Intelligence Agency
- National Security Agency
- U.S. Marine Corps

Department of Education

Department of Energy

- Bonneville Power Administration
- Southeastern Power Administration
- Southwestern Power Administration
- Western Area Power Administration

Department of Health and Human Services

- Administration for Children and Families
- Agency for Healthcare Research and Quality
- Centers for Disease Control and Prevention
- Centers for Medicare and Medicaid Services
- Food and Drug Administration
- Health Resources and Services Administration
- Indian Health Service
- National Institutes of Health
- Program Support Center

Department of Homeland Security

- Border and Transportation Security Directorate
- Bureau of Citizenship and Immigration Services
- Emergency Preparedness and Response Directorate
- Information Analysis and Infrastructure Protection Directorate
- Management Directorate

- Science and Technology Directorate
- U.S. Coast Guard
- U.S. Secret Service

Department of Housing and Urban Development

Department of the Interior

- Bureau of Indian Affairs
- Bureau of Land Management
- Bureau of Reclamation
- Minerals Management Service
- National Park Service
- Office of Surface Mining Reclamation and Enforcement
- U.S. Fish and Wildlife Service
- U.S. Geological Survey

Department of Justice

- Bureau of Alcohol, Tobacco, Firearms, and Explosives
- Drug Enforcement Administration
- Federal Bureau of Investigation
- Federal Bureau of Prisons
- U.S. Marshals Service

Department of Labor

Department of State

Department of Transportation

- Federal Aviation Administration
- Federal Highway Administration
- Federal Motor Carrier Safety Administration
- Federal Railroad Administration
- Federal Transit Administration
- National Highway Traffic Safety Administration

Department of the Treasury

- Bureau of Engraving and Printing
- Bureau of the Public Debt
- Financial Management Service
- Internal Revenue Service
- Office of the Comptroller of the Currency
- Office of Thrift Supervision
- U.S. Mint

Department of Veterans Affairs

- Veterans Benefits Administration
- Veterans Health Administration

Agency for International Development

Central Intelligence Agency

Corporation for National and Community Service

Environmental Protection Agency

Equal Employment Opportunity Commission

Executive Office of the President

Export-Import Bank of the United States

Federal Deposit Insurance Corporation

Federal Energy Regulatory Commission

Federal Reserve System

Federal Retirement Thrift Investment Board

General Services Administration

Legal Services Corporation

National Aeronautics and Space Administration

National Credit Union Administration

National Labor Relations Board

National Science Foundation

Nuclear Regulatory Commission

Office of Management and Budget

Office of Personnel Management

Peace Corps

Pension Benefit Guaranty Corporation

Railroad Retirement Board

Securities and Exchange Commission

Small Business Administration

Smithsonian Institution

Social Security Administration

U.S. Postal Service

Departments and Agencies Reporting No Data Mining Efforts

The following 69 departments and agencies reported that they have no operational or planned data mining efforts:

Department of Agriculture

- Agricultural Marketing Service
- Agricultural Research Service
- Animal and Plant Health Inspection Service
- Cooperative State Research, Education, and Extension Service
- Farm Service Agency
- Foreign Agricultural Service
- Forest Service
- National Agricultural Statistics Service
- Food Safety and Inspection Service

Department of Commerce

- Economic Development Administration
- Bureau of the Census
- International Trade Administration
- Department of Commerce Headquarters
- National Oceanic and Atmospheric Administration

Department of Defense

- Defense Contract Audit Agency
- Missile Defense Agency
- Defense Legal Services Agency

- Defense Security Service
- Defense Threat Reduction Agency
- Defense Logistics Agency
- Defense Advanced Research Projects Agency
- Defense Contract Management Agency
- Defense Security Cooperation Agency

Department of Energy

- Bonneville Power Administration
- Southeastern Power Administration
- Southwestern Power Administration
- Western Area Power Administration

Department of Health and Human Services

- Centers for Medicare and Medicaid Services
- Administration for Children and Families
- National Institutes of Health
- Indian Health Service

Department of Homeland Security

- Science and Technology Directorate
- Management Directorate
- Bureau of Citizenship and Immigration Services
- Department of Homeland Security Headquarters

Department of Housing and Urban Development

Department of the Interior

- Bureau of Reclamation
- Bureau of Land Management
- U.S. Geological Survey
- Fish and Wildlife Service
- Office of Surface Mining Reclamation and Enforcement
- Bureau of Indian Affairs
- Department of the Interior Headquarters

Department of Justice

- Bureau of Alcohol, Tobacco, Firearms, and Explosives

Department of Transportation

- Federal Aviation Administration
- Federal Transit Administration
- Federal Railroad Administration
- Federal Motor Carrier Safety Administration
- Federal Highway Administration

Department of the Treasury

- Comptroller of the Currency
- Bureau of the Public Debt
- Office of Thrift Supervision

- Department of the Treasury Headquarters
- Bureau of Engraving and Printing

Agency for International Development

Executive Office of the President

Federal Energy Regulatory Commission

Federal Retirement Thrift Investment Board

General Services Administration

Legal Services Corporation

National Credit Union Administration

National Labor Relations Board

National Science Foundation

Office of Management and Budget

Peace Corps

Security and Exchange Commission

Smithsonian Institution

Social Security Administration

U.S. Postal service

Inventories of Efforts

The following tables present selected information from our survey of 128 major federal departments and agencies on their use of data mining. The tables list the purpose of each data mining effort, whether the system is planned or operational, and whether the system uses personal information, data from the private sector, or data from other federal agencies. The survey shows that 52 departments and agencies are using or are planning to use data mining. These departments and agencies reported 199 data mining efforts, of which 68 were planned and 131 were operational.

Table 2: Department of Agriculture's Inventory of Data Mining Efforts

				Features		
Organization/ system name	**Description**	**Purpose**	**Status**	**Personal information**	**Private sector data**	**Other agency data**
Department of Agriculture Headquarters						
Travel Data Mart	Will consolidate employee travel information from financial and travel systems. Will allow for a governmentwide e-travel system and provide the department with information on the financial ramifications of its travel.	Improving service or performance	Planned	Yes	No	No
Financial Statements Data Warehouse	Is used in the production of consolidated financial statements. Provides information for products that are used to satisfy external reporting requirements, such as Office of Management and Budget and Department of the Treasury requirements.	Financial management	Operational	No	No	No
Financial Data Warehouse	Is the department's internal financial management reporting system. Data mining is done for ad hoc and on-demand reports.	Financial management	Operational	Yes	No	No
Food and Nutrition Service						
Grantee Monitoring Activities—Southeast Regional Office	Assists in monitoring the financial status of grant holders. Grantees are required to provide expenditure reports, and analysis is performed quarterly that matches stated draws to the actual draws from the U.S. Treasury.	Improving service or performance	Operational	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Grantee Monitoring Activities—Mountain Plains Regional Office	Assists in monitoring the management and distribution of Indian funds for major food benefit programs, such as food stamps, in 10 grantee states.	Improving service or performance	Operational	Yes	No	No
Grantee Monitoring Activities—Southwest Regional Office	Maximizes on-site monitoring efforts by confirming the accuracy of grantee accounting. Reduces on-site time, maximizes time to complete reviews, and has achieved a 50 percent travel savings.	Improving service or performance	Operational	Yes	No	No
Grantee Monitoring Activities—Midwest Regional Office	Will be a reporting system to provide reports and automate the audit process. Plans are to acquire data mining tools to review and compare budgets, reports, and plans.	Improving service or performance	Planned	No	No	Yes
Grantee Monitoring Activities—Northeast Regional Office	Supports on-site reviews of analyses to confirm financial report information.	Improving service or performance	Operational	Yes	Yes	No
Integrated Program Accounting System Data Integrity	Will create ad-hoc reporting centers to validate accounting information.	Improving service or performance	Planned	No	No	No
Natural Resources Conservation Service						
National Resource Inventory Used for Statistical Analysis of Past Soil Survey Databases.	Is a trending database that tracks more than 200 resource issues such as monitoring erosion. Also processes statistical technology.	Improving service or performance	Operational	No	No	No
Risk Management Agency						
CAE	Is part of a congressionally mandated project to assist the Risk Management Agency in controlling fraud, waste, and abuse in the Federal Crop Insurance Corporation program.	Detecting fraud, waste, and abuse	Operational	Yes	Yes	Yes

Source: Department of Agriculture.

Table 3: Department of Commerce's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
U.S. Patent and Trademark Office						
Compensation Projection Model in the Enterprise Data Warehouse	Generates and makes available compensation projection data, both salary and benefits, on current employees and on planned hires. It also accounts for planned attritions.	Managing human resources	Operational	Yes	No	Yes

Source: Department of Commerce.

Table 4: Department of Defense's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Defense Commissary Agency						
DeCA Electronic Records Management and Archive System	Will be a corporate information system for managing unstructured data. It will allow for electronic record keeping, document management, and automated receipt processes.	Improving service or performance	Planned	Yes	Yes	Yes
Corporate Decision Support System/ Commissary Operations Management System	Mines data to produce analytical data on commissary operations. Provides information such as what items stores are selling and helps determine whether cashiers are being honest.	Improving service or performance	Operational	No	No	No
Defense Information Systems Agency						
Enterprise Business Intelligence System	Will replace the current management information environment, which includes operations, reporting, billing, statistics, and other management information activities.	Improving service or performance	Planned	No	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Defense Intelligence Agency						
Insight Smart Discovery	Will be a data mining knowledge discovery tool to work against unstructured text. Will categorize nouns (names, locations, events) and present information in images.	Analyzing intelligence and detecting terrorist activities	Planned	Yes	No	Yes
Verity K2 Enterprise	Mines data from the intelligence community and Internet searches to identify foreign terrorists or U.S. citizens connected to foreign terrorism activities.	Analyzing intelligence and detecting terrorist activities	Operational	Yes	Yes	Yes
PATHFINDER	Is a data mining tool developed for analysts that provides the ability to analyze government and private sector databases rapidly. It can compare and search multiple large databases quickly.	Analyzing intelligence and detecting terrorist activities	Operational	Yes	No	Yes
Autonomy	Is a large search engine tool that is used to search hundreds of thousands of word documents. Is used for the organization and knowledge discovery of intelligence.	Analyzing intelligence and detecting terrorist activities	Operational	No	No	Yes
Department of the Air Force						
ANG Data Warehouse—Guardian	Will be used to measure military readiness. It incorporates information on all disciplines to provide management information needed to assess military readiness.	Measuring military readiness	Planned	Yes	No	No
Integrated Space Warfare Center (SWC) Information System	Will be an internal database containing information on all development/execution activities within the SWC. Will be used by all management and analyst personnel to track and align the center's activities to warfighter needs, report on execution status, financial status, schedule status, and performance measurements.	Improving service or performance	Planned	Yes	No	No
Safety Automated System (SAS)	Will query databases to find automation mishaps. Governed by Directive 920124 and will allow for the investigation and reporting of identified automation mishaps.	Improving safety	Planned	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Enterprise Business System	Will support strategic planning, assist in building scientific and technical budgets for the Air Force, and serve as a launch point for all new programs. Research and development case files will be maintained for 75 years; the activity indexes, catalogs, and tracks these files.	Improving service or performance	Planned	No	No	Yes
Genomic and Proteomic Results Analysis	Analyzes National Institutes of Health's genetic data.	Analyzing scientific and research information	Operational	No	No	Yes
IG Corporate Information System	Enhances combat readiness and mission capabilities for Air Combat Command units and commanders. It assists in preparing for and conducting inspections.	Improving service or performance	Operational	Yes	No	No
Computer Network Defense System	Evaluates network activities to create rules for intrusion detection system signature sets.	Improving information security	Operational	No	No	No
FAME	Will serve as a central repository for Air Force manpower information. Will track manpower and unit authorization funding.	Managing human resources	Planned	No	No	Yes
Resource Wizard	Serves as a manpower tracking system. Tracks positions and captures data for specific funding purposes.	Improving service or performance	Operational	No	No	No
Government Purchase Card	Is used in overseeing purchases made by Air Force personnel with government-provided credit cards.	Detecting fraud, waste, and abuse	Operational	Yes	Yes	No
Ambulatory Data System Queries	Tracks the initial diagnosis of patients with the results of further testing and diagnosis. Allows for early notification of diseases and injuries.	Monitoring public health	Operational	Yes	No	No
Modus Operandi Database	Is an investigative tool used to identify and track trends in criminal behavior. It links characteristics of crimes and provides details on crime scenes and other crime factors.	Detecting criminal activities or patterns	Operational	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Executive Decision Support System	Takes data from all functional metric balances. Processes charts and graphs to identify trends and to make sure goals are accomplished.	Improving service or performance	Operational	No	No	No
Inspire	Is a tool that assists in providing a narrative description of all research and development that is being conducted within the Air Force. Provides cost and milestone information on research and development projects.	Performing strategic planning	Operational	Yes	No	Yes
Discoverer	Is used to manage personnel records, including individual aliases and histories.	Managing human resources	Operational	Yes	No	No
Requirements and Concepts System	Will serve as a repository for new system projects and system requirements. It will be available for consultation for information on all project requests and identified requirements.	Improving service or performance	Planned	No	No	No
Business Objects	Is a commercial off-the-shelf tool that is used to analyze and report on human resources activities.	Managing human resources	Operational	Yes	No	Yes
THRMIS	Uses commercial off-the-shelf software to maintain a data warehouse of integrated inventory and manpower data for the Total Force: active duty (officer and enlisted), Air Force Reserve, Air National Guard, and civilians. Is used to assess and analyze the health of the Air Force.	Managing human resources	Operational	Yes	No	No
SAS	Is a Web-enabled personnel data system that gives authorized users worldwide the ability to tabulate demographic data on recruitment, promotion, and retention.	Managing human resources	Operational	Yes	No	No
Oracle HR	Is a personnel management system that manages information for promotions, pay grades, clearances, and other information relevant to human resources.	Managing human resources	Operational	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Health Modeling and Informatics Division Data Mart	Provides information and decision support to the Air Force headquarters' surgeon general for decision making, policy development, and resource allocation. It also provides performance information and analysis to medical field units in support of performance measurement objectives.	Improving service or performance	Operational	Yes	No	No
FIRST EDV (BRIO)	Will deal with Air Force budgets and other components of its financial environment. Historical analyses and trend analyses will be performed on the budget process.	Improving service or performance	Planned	No	Yes	No
IG World	Is used to store and track data and requirements, such as lodging and augmentee requirements, for the PAC inspector general.	Improving service or performance	Operational	Yes	No	No
Department of Defense Headquarters						
Automated Continuing Evaluation System	Will be used to improve personnel security continuing evaluation efforts within Department of Defense (DOD) by identifying issues of security concern between the normal reinvestigation cycle for those who hold DOD security clearances and have signed a consent form that is still in effect.	Managing human resources	Operational	Yes	Yes	Yes
Department of the Navy						
Human Resource Trend Analysis	Is used to improve Navy readiness. Data on personnel manning levels are mined to ensure that each Navy unit has the correct number of training personnel aboard.	Managing human resources	Operational	No	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
U.S. Naval Academy	Allows for the assessment of academic performance of midshipmen. It includes demographic information, information on grades, participation in sports, leadership positions, etc. It is an extension of the registrar's system and is mined for comparisons and trends.	Managing human resources	Operational	Yes	No	No
Navy Training Master Planning System	Provides overall Navy training information to assist in delivering Navy training in the most efficient manner. Pertinent data from multiple databases are consolidated into a single database that is mined.	Managing human resources	Operational	Yes	Yes	No
DHAMS Multidimensional Cubes	Is a database that contains information on the time and attendance of 3,000 mariners across 120 ships. Allows managers to look at what people were doing at a particular time and to look across the fleet as a whole and compare ship activities.	Improving service or performance	Operational	No	No	No
National Cargo Tracking Plan Cargo Tracking Division	Is used to conduct predictive analysis for counterterrorism, small weapons of mass destruction proliferation, narcotics, alien smuggling, and other high-interest activities involving container shipping activity.	Analyzing intelligence and detecting terrorist activities	Operational	No	Yes	No
Supply Management System Multidimensional Cubes	Reduces downtime on ships by allowing for the analysis of ship parts information. The data warehouse contains data on every part that has been ordered since the 1980s, and has multidimensional information on each part. Failure rates can be calculated and improvements can be identified.	Improving service or performance	Operational	No	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Type Commanders Readiness Management System	Is designed to provide a fully integrated environment for online analytical processing of readiness indicators. Examples of readiness indicators include status of supplies available, equipment in operation, health status, and capabilities of the crew.	Measuring military readiness	Operational	No	No	Yes
FATHOM (APMC—Human Resources)	Will be an internal program and project tool used to improve staffing, recruiting, and managing day-to-day operations.	Managing human resources	Planned	Yes	No	No
Navy Training Quota Management System	Is used for planning and forecasting training needs based on skill requirements.	Improving service or performance	Operational	No	No	Yes
National Geospatial-Intelligence Agency						
OLAP (On-Line Analytical Processing)	Will provide aggregations of imagery system performance data for management officers and senior source decision makers to characterize system performance and contribution to intelligence issues of national priority.	Improving service or performance	Planned	No	No	No
CITO Data Mining	Will evaluate and identify imagery system performance trends for optimization, monitoring, or reengineering.	Improving service or performance	Planned	No	No	No
Information Relevance Prototype	Will establish an information relevancy prototype to serve as a framework for community evaluation of commercial information relevance approaches, methods, and technology. The term information relevance refers to the ability of users to receive or extract, then display and describe, information with measurable satisfaction according to their need.	Improving service or performance	Planned	No	No	No
U.S. Marine Corps						
Operational Data Store Enterprise	Is used for workforce planning.	Managing human resources	Operational	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Global Combat Support Systems—Marine Corps	Will be a physical implementation of the IT enterprise architecture designed to support both improved and enhanced marine air/ground task force combat service support functions and commander and combatant commander joint task force combatant support information requirements. Data mining will allow for interoperability with legacy Marine Corps systems and allow for a shared data environment.	Improving service or performance	Planned	No	Yes	No
Total Force Data Warehouse	Is a system whose primary purpose is workforce planning and workforce policy decision making. It contains current (after 30 days) and historical workforce data.	Managing human resources	Operational	Yes	No	No
Marine Corps Recruiting Information Support System	Is a Web-based information system used for managing assets and tracking enlisted and officer accessions into the Marine Corps.	Managing human resources	Operational	Yes	No	No

Source: Department of Defense.

Table 5: Department of Education's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Citizenship of PLUS Loan Borrowers—National Student Loan Data Systems	Looks for issues regarding citizenship among its PLUS loan borrowers. Flags records based on selected criteria and requests additional information from schools.	Improving service or performance	Operational	Yes	Yes	Yes
Foreign Schools Initiatives National Student Loan Data System/Central Processing	Is a proactive investigation effort that looks at whether financial aid was granted individuals attending foreign institutions during periods of nonenrollment.	Detecting criminal activities or patterns	Operational	Yes	No	Yes
Professional Judgment Practices: Title IV Pell Grants, National Student Loan Data	Used to determine when professional judgment has been exercised for "special" situations where families cannot afford college expenses.	Improving service or performance	Operational	Yes	Yes	Yes
Title IV Applicant—Death Database Match	Compares Department of Education data with the Social Security Administration's death database to detect fraud or criminal activity.	Detecting fraud, waste, and abuse	Operational	Yes	No	Yes
Title IV Loans with No Applications	Will compare information from the Free Application for Federal Student Aid Program with the Federal Family Education Loan Program to identify fraud.	Detecting fraud, waste, and abuse	Planned	Yes	No	No
OIG—Project Strikeback	Compares Department of Education and Federal Bureau of Investigation data for anomalies. Also verifies personal identifiers.	Analyzing intelligence and detecting terrorist activities	Operational	Yes	No	Yes
Accuracy of U.S. Department of Education Personal Data	Audits and verifies personal information that is contained in the Department of Education's personal data system.	Detecting fraud, waste, and abuse	Operational	Yes	No	Yes
Impact of Cohort Default Rate Redefinition—National Student Loan Data System	Audits data to determine the impact of legislation that extended the college loan repayment default period from 180 to 270 days.	Legislative impact	Operational	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
CheckFree Software/Purchase Card Program	Takes monthly billing information from the Bank of America to create reports on purchases, purchase quantity, and frequency of purchases. Data are mined for instances of fraud or abuse.	Detecting fraud, waste, and abuse	Operational	Yes	Yes	No
Improper Pell Grant Payment Activity	Will compare Pell Grants issued with the amounts received and look at the eligibility of grant recipients.	Detecting fraud, waste, and abuse	Planned	Yes	No	No
Title IV Identity Theft Initiative	Helps identify patterns and trends in identity theft cases involving loans for education. Provides an investigative resource for victims of identity theft.	Detecting criminal activities or patterns	Operational	Yes	No	No
Title IV Applicant—Use of Multiple Addresses/Central Processing System	Reviews addresses listed on Title IV applications to see if they are valid. For example, jails or employment addresses are not considered valid addresses.	Improving service or performance	Operational	Yes	No	Yes
Lapsed Funds/Improper Draw of Federal Grant Proceeds	Identifies funds that remain in the grants and payment processing system beyond the time period for allocating the funds.	Improving service or performance	Operational	No	No	No
Decision Support System with Online Analytical Processing Query	Will support the department's performance-based initiative. Will allow custom queries of schools from state and local databases for demographics and test scores.	Improving service or performance	Planned	No	No	No
Grant Administration and Payment System	Assists in managing grant activities and aids in detecting instances of fraud or abuse in grant activities.	Detecting fraud, waste, and abuse	Operational	Yes	Yes	Yes
Budget Execution Support	Uses information in the National Student Loan Data System and a sample drawn from it to estimate cohort distributions for financial activities related to the Federal Family Education Loan Program pursuant to the Credit Reform Act.	Financial management	Operational	Yes	No	No
Pell Grant Model Assumptions	Provides estimates on the total cost of the Pell Grant program. It uses data from previous years and makes assumptions for future years.	Financial management	Operational	No	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
National Student Loan Data System	Compiles student loan information from the guaranteeing agencies. Is used for eligibility tracking and to calculate default rates.	Detecting fraud, waste, and abuse	Operational	Yes	No	Yes
Loan Model Assumptions	Estimates the cost of loan programs. Also analyzes loan default behavior.	Financial management	Operational	Yes	No	Yes
Office of the Inspector General (OIG) Projects: Tumbleweed/ Snowball	Is part of an OIG investigation to determine potential fraud of financial aid grants primarily in New Hampshire.	Detecting criminal activities or patterns	Operational	Yes	No	Yes
Central Processing System	Processes applications for student aid. Contains data on more than 13 million applications. Data are mined for demographic trends.	Detecting fraud, waste, and abuse	Operational	Yes	No	No
Direct Loan Services System	Is used to track the life of student direct loans and to monitor loan repayments.	Improving service or performance	Operational	Yes	Yes	Yes
CheckFree Software/Travel Card Program	Uses monthly billing information from Bank of America to create reports on travel expenditures to look for improper use of travel cards.	Detecting fraud, waste, and abuse	Operational	Yes	Yes	No

Source: Department of Education.

Table 6: Department of Energy's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Counterintelligence Automated Investigative Management System (CI-AIMS)	Is an investigative management system used by Department of Energy (DOE) field sites to track investigative cases on individuals or countries that threaten DOE assets. Information stored in this database is also used to support federal and state law enforcement agencies in support of national security.	Detecting criminal activities or patterns	Operational	Yes	No	No
Autonomy	Will be used to mine a myriad intelligence-related databases within the intelligence community to uncover criminal or terrorist activities relating to DOE assets.	Detecting criminal activities or patterns	Planned	Yes	No	No
Counterintelligence Analytical Research Data System (CARDS)	Is used to log briefings and debriefings given to DOE employees who travel to foreign countries or interact with foreign visitors to DOE facilities. Data are mined to identify potential threats to DOE assets.	Detecting criminal activities or patterns	Operational	Yes	No	Yes

Source: Department of Energy.

Table 7: Department of Health and Human Services' Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Agency for Healthcare Research and Quality						
National Patient Safety Network	Will contain reports on adverse medical events that are filed by hospitals. The planned network's purpose is to take out patient personal identifiers and other items that may violate certain rules and create a warehouse that can be used by registered and unregistered users to evaluate and implement patient safety and quality measures. The network will be used to create tools that hospitals can use for making quality improvements.	Improving service or performance	Planned	No	No	No
Centers for Disease Control and Prevention						
BioSense	Enhances the nation's capability to rapidly detect bioterrorism events.	Analyzing intelligence and detecting terrorist activities	Operational	No	Yes	Yes
Department of Health and Human Services Headquarters						
DHHS Blood Monitoring Program	Monitors the country's blood supply by keeping an inventory on red blood cells and platelets and monitors blood supply shortages, the nature of the shortage, and size of the shortages.	Monitoring public health	Operational	No	Yes	No
Food and Drug Administration						
Mission Accomplishment and Regulatory Compliance Services System	Is a comprehensive redesign and reengineering of two core mission-critical legacy systems at Food and Drug Administration (FDA) that support the regulatory functions that primarily take place in FDA's field offices.	Monitoring food or drug safety	Operational	No	Yes	Yes

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Turbo Establishment Inspection Report	Provides a standardized database of citations of regulations and statutes, and help investigators in preparing reports. It will collect data on specific observations uncovered during inspections and provide a more uniform format nationwide that will allow for electronic searches and statistical analysis to be performed by citation.	Improving safety	Operational	No	Yes	No
Phonetic Orthographic Computer Analysis	Is a search engine that provides results indicating how similar two drug names are on a phonetic and orthographic basis. Its purpose is to help in the safety evaluation of proposed proprietary names to reduce drug name confusion after an application is approved by the FDA.	Improving safety	Operational	No	Yes	No
MPRIS Data Warehouse	Will provide data to support end user ad-hoc query analysis and standard reporting needs. It will provide the foundation for a central reporting repository that can be used to populate business-specific data marts.	Improving service or performance	Planned	No	No	No
Development and Deployment of Advanced Analytical Tools for Drug Safety Risk Assessment	Will develop advanced software tools for quantitative analysis of drug safety data. Medical officers and safety evaluators will use these advances in software tools.	Analyzing scientific and research information	Planned	Yes	Yes	Yes
Add data mining capability to CFSAN Adverse Event Reporting System	Is a comprehensive system for tracking, reviewing, and reporting adverse event incidences involving foods, cosmetics, and dietary supplements. Integrating and centralizing the system and eliminating patchwork systems make information on these adverse events available to federal, state, and local governments as well as to industry and the public in a more timely and efficient manner.	Monitoring food or drug safety	Planned	Yes	Yes	Yes

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Health Resources and Services Administration						
HRSA Geospatial Data Warehouse	Data warehouse that primarily collects programmatic, demographic, and statistical data.	Improving service or performance	Operational	No	Yes	Yes
Program Support Center						
Employee Assistance Program Analysis	Uses information from a database of employee assistance program case information that does not contain client personal identifiers. Data are mined for quality assurance and program management information that is used to enhance the quality and cost effectiveness of services.	Improving service or performance	Operational	No	No	No

Source: Department of Health and Human Services.

Table 8: Department of Homeland Security's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Border and Transportation Security Directorate						
Workforce Profile Data Mart	Contains payroll and personnel data and is mined for workforce trends.	Managing human resources	Operational	Yes	No	Yes
Customs Integrated Personnel Payroll System Data Mart	Is a Customs data mart contained within Department of Homeland Security's workforce profile data mart. Personnel and payroll data are mined for workforce trends.	Managing human resources	Operational	Yes	No	Yes
Internal Affairs Treasury Enforcement Communications System Audit Data Mart	Assists the Internal Affairs group by mining criminal activity data to ascertain how Customs' employees are using the Treasury Enforcement System.	Detecting criminal activities or patterns	Operational	Yes	No	Yes
Operations Management Reports Data Mart	Assists in managing the operation of all ports of entry for incoming carriers, people, and cargo. Helps in making resource (people and equipment) allocation and operational improvement decisions.	Improving service or performance	Operational	No	No	Yes

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Automated Export System Data Mart	Mines data on export trade in the U.S. and produces reports on historical shipping and receiving trends.	Improving service or performance	Operational	No	Yes	Yes
Seized Property/ Forfeitures, Penalties, and Fines Case Management Data Mart	Mines data to ensure data quality and review work assignments. System has two components: one that processes legal cases like a law firm, and a second that serves as property and inventory control by tracking property seized.	Improving service or performance	Operational	Yes	No	No
Incident Data Mart	Will look through incident logs for patterns of events. An incident is an event involving a law enforcement or government agency for which a log was created (e.g., traffic ticket, drug arrest, or firearm possession). The system may look at crimes in a particular geographic location, particular types of arrests, or any type of unusual activity.	Analyzing intelligence and detecting terrorist activities	Planned	Yes	Yes	Yes
Case Management Data Mart	Assists in managing law enforcement cases, including Customs cases. Reviews case loads, status, and relationships among cases.	Analyzing intelligence and detecting terrorist activities	Operational	Yes	Yes	Yes
Emergency Preparedness and Response Directorate						
Enterprise Data Warehouse	Will take data from multiple, disparate systems and integrate the data into one reporting environment. The objective of the effort is to allow for the reduction of data within the agency and to provide an enterprise view of information necessary to drive critical business processes and decisions. Data on internal human resources, all aspects of disaster management, infrastructure, equipment location, etc., will be used.	Disaster response and recovery	Planned	Yes	Yes	Yes
Information Analysis and Infrastructure Protection Directorate						
Analyst Notebook I2	Correlates events and people to specific information	Analyzing intelligence and detecting terrorist activities	Operational	Yes	Yes	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Automatic Message Handling System (Verity)	Automatically takes messages from external agencies and routes them to appropriate recipients	Analyzing intelligence and detecting terrorist activities	Planned	No	No	Yes
U.S. Coast Guard						
Readiness Management System	Assists in ensuring readiness for all Coast Guard missions.	Improving service or performance	Operational	Yes	No	No
CG Info	Provides one-stop shopping for Coast Guard information. It is the central location and common interface for the entire Coast Guard to gain near real-time access to data from multiple, disparate Coast Guard information systems. It provides a single interface for users to view mission-critical support data.	Improving service or performance	Operational	Yes	No	Yes
U.S. Secret Service						
Criminal Investigation Division Data Mining	Mines data in suspicious activity reports received from banks to find commonalities in data to assist in strategically allocating resources.	Detecting criminal activities or patterns	Operational	Yes	No	Yes

Source: Department of Homeland Security.

Table 9: Department of the Interior's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Minerals Management Service						
Data Mining of the Technical Information Management System (TIMS) Database	Is a corporate database for oil and gas leases. The database is mined in support of policy development. One area of data mining is identification of leases that will be abandoned in the near future. Data mining has shown that leases with six or more producing wells in 1 year are almost never abandoned in the next year. Another application of data mining is the safety of oil and gas operations. For example, data mining has shown that accidents have a peak rate on Thursday mornings.	Improving service or performance	Operational	Yes	Yes	No

Source: Department of the Interior.

Table 10: Department of Justice's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Department of Justice Headquarters						
Drug/Financial Fusion Center	Will contain data from, and be used by, Organized Crime and Drug Enforcement Task Force agencies. The system will permit the collection and cross case analysis of all drug and related financial investigative data.	Detecting criminal activities or patterns	Planned	Yes	Yes	Yes
Drug Enforcement Administration						
Statistical Management Analysis and Reporting Tool System (SMARTS) /SPSS	Is a query analysis and reporting tool that pulls data from many systems. It allows for statistical analyses of drug cases Drug Enforcement Administration's statistical reporting.	Detecting criminal activities or patterns	Operational	Yes	No	Yes
TOLLS	Is a database of telephone calls from court ordered and approved wiretaps and Title III investigations. Information such as telephone numbers, time and date of calls, and call duration is captured. Data are mined for patterns to give leads in investigations of drug trafficking.	Detecting criminal activities or patterns	Operational	Yes	No	No
Federal Bureau of Investigation						
Secure Collaborative Operational Prototype Environment/ Investigative Data Warehouse	Allows the FBI to search multiple data sources through one interface to uncover terrorist and criminal activities and relationships. Data sources are a combination of structured and unstructured text.	Analyzing intelligence and detecting terrorist activities	Operational	Yes	No	Yes
Foreign Terrorist Tracking Task Force Activity	Supports the Foreign Terrorist Tracking Task Force that seeks to prevent foreign terrorists from gaining access to the United States. Data from the Department of Homeland Security, Federal Bureau of Investigation, and public data sources are put into a data mart and mined to determine unlawful entry and to support deportations and prosecutions.	Analyzing intelligence and detecting terrorist activities	Operational	Yes	Yes	Yes

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
FBI Intelligence Community Data Marts	Is intended to take a subset of approved data from a data warehouse and make it available to the intelligence community.	Analyzing intelligence and detecting terrorist activities	Planned	Yes	No	Yes
Federal Bureau of Prisons						
Business Information Warehouse	Will be a warehouse designed to provide information on manufacturing by Federal Prison Industries, which runs 100 factories in various prisons. Data will be mined for information on the manufacturing environment (such as information on material on hand, scheduling, and the production process) and financial activities.	Improving service or performance	Planned	No	No	Yes
U.S. Marshals Service						
USMS Workload Modeling	Will seek to develop a workforce model that will support budget formulation, execution, and resource analysis. Will be a planning and execution activity that will be used to help determine the quantity and location of required resources.	Managing human resources	Planned	Yes	No	No

Source: Department of Justice.

Table 11: Department of Labor's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Dashboard Display	Provides links to programs throughout the Department of Labor's Employment Training Administration to provide reports or information on financial activities.	Improving service or performance	Operational	Yes	No	No
Enforcement Management System, Case Opening, and Results Analysis	Is used to track investigations of violations of Title I and other criminal laws pertaining to pension and welfare rights.	Improving service or performance	Operational	Yes	Yes	No
Employee Retirement Income Security Act Data System	Is used to monitor compliance with Title I of the Employee Retirement Income Security Act.	Detecting fraud, waste, and abuse	Operational	Yes	No	No
Mine Safety and Health Administration Teradata Data Store	Mines data from a data store of information on safety and health enforcement and demographic data for mine operations, along with miner accidents, injury, and illness data.	Improving safety	Operational	Yes	No	Yes
Mathematical Statistics Research Center	Will look at data from economic surveys to compare rates of nonresponse for Bureau of Labor Statistics.	Improving service or performance	Planned	No	No	No

Source: Department of Labor.

Table 12: Department of State's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Citibank's Ad Hoc Reporting System	Enables purchase card managers to track trends related to the usage of credit cards by employees in purchasing supplies and services for official use. Purchase card program is worldwide, and spending patterns and purchases are monitored for potential misuse or fraud.	Detecting fraud, waste, and abuse	Operational	Yes	Yes	No
Purchase Card Management System	Will involve the automation of internal workflow processes (system is in the early phases of development). Will use internal data and bank data to track trends and anomalies in the Department of State's worldwide purchase card program.	Detecting fraud, waste, and abuse	Planned	Yes	Yes	No

Source: Department of State.

Table 13: Department of Transportation's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Department of Transportation Headquarters						
DOT IT Security Management System	Will collect information to allow management to assess its IT security infrastructure.	Detecting fraud, waste, and abuse	Planned	Yes	No	No
National Highway Traffic Safety Administration						
State Data System	Analyzes, mines, and researches automotive crash data, such as statistics from rollovers of SUVs, from 22 states to improve highway safety and lessen fatalities. Policies can be set based on the data.	Improving safety	Operational	No	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Fatality Analysis Reporting System (FARS)	Helps to evaluate the effectiveness of motor vehicle safety standards and highway safety programs. Data are collected from all 50 states, the District of Columbia, and Puerto Rico and are used to evaluate and support highway safety.	Improving safety	Operational	Yes	Yes	Yes
National Automotive Sampling System	Collects and mines information on automotive crashes. System is related to the Federal Motor Vehicle Safety Standards that regulate vehicle compliance items such as seat belts, air bags, and the stopping distance of brakes.	Improving safety	Operational	Yes	Yes	No

Source: Department of Transportation.

Table 14: Department of the Treasury's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Financial Management Service						
Treasury Offset Program (TOP) Cleanup	Mines data to reduce the number of debts listed in TOP.	Improving service or performance	Operational	Yes	No	Yes
Electronic Federal Tax Payment System (EFTPS) Marketing	Is a free service offered by the Department of the Treasury for individuals and business taxpayers who pay their federal taxes electronically. Mining activity tracks enrollment, tax payment history, and usage trends.	Increasing tax compliance	Operational	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Internal Revenue Service						
Planning, Analysis, and Decision Support System	Will be a component of the Custodial Accounting Program, which is the warehouse that is used to query transactional data and produce reports. This activity is meant to improve reporting and use decision support tools.	Improving service or performance	Planned	Yes	No	Yes
Abusive Corporate Tax Shelter Detection Model	Will model characteristics of corporate tax shelters and use models to predict corporate tax shelter abuse and to assess compliance risk in the corporate taxpayer population.	Increasing tax compliance	Planned	Yes	Yes	No
K-1 Link Analysis	Will be used to detect potential tax evasion.	Increasing tax compliance	Planned	Yes	No	No
Research on the Population of Taxpayers Who Receive Earned Income Tax Credit	Will be used to research data on taxpayers who receive the EITC.	Detecting fraud, waste, and abuse	Planned	Yes	No	No
Issue Based Management Information System	Will provide access to a variety of data sources within IRS. Will assist in research and case work.	Increasing tax compliance	Planned	No	Yes	No
Electronic Fraud Detection System	Mines data to evaluate and rate potentially fraudulent individual tax returns.	Improving service or performance	Operational	Yes	No	No
Reveal	Will be used to detect financial criminal activity such as tax evasion.	Detecting criminal activities or patterns	Planned	Yes	Yes	No
Oracle Model 22 Partnership Return Scoring System	Takes information from individual tax returns and attempts to replicate judgments made by taxpayers to detect the likelihood of material errors.	Increasing tax compliance	Operational	Yes	No	No
SPSS Form 1120-S Return Scoring System	Will automate the classification of certain corporate tax returns.	Increasing tax compliance	Planned	Yes	No	No
Oracle Model 33 Partnership Scoring Model	Will identify noncompliance in partnership returns.	Increasing tax compliance	Planned	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Compliance Laboratory	Will identify taxpayer noncompliance by looking at groups of returns.	Increasing tax compliance	Planned	Yes	Yes	Yes
U.S. Mint						
Information Technology Intrusion Detection System	Collects information on potential intrusions to U.S. Mint systems. Looks for trends in information reported by sensors to determine if illicit activity has occurred. Minimizes false positives.	Improving information security	Operational	No	No	No
E-Commerce Fraud Analysis Activity	Attempts to identify and stop fraudulent activity involving stolen credit cards to order products over the Internet or via telephone. Fraud rating identifiers are used to identify areas where fraud has occurred and to determine the likelihood of fraud. Allows for orders to be stopped or for orders over a certain dollar limit to be stopped.	Detecting criminal activities or patterns	Operational	Yes	Yes	Yes
Data Warehouse	Will be an integrated, scalable, expandable data warehouse that will support business functions by grouping the data in subject-oriented data marts. Each warehouse data mart will be defined to integrate both internal and external data to provide the necessary information to perform both historical and predictive analysis and support numerous calculations.	Improving service or performance	Planned	No	No	No

Source: Department of the Treasury.

Table 15: Department of Veterans Affairs' Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Department of Veterans Affairs Headquarters						
Veterans Affairs Central Incident Response Center	Is used to monitor and manage intrusion detection and firewalls. Scripts are written for forensic analysis to go through data collected from system and network logs.	Detecting criminal activities or patterns	Operational	Yes	Yes	No
Purchase Card Data Mining (SAS) Reports	Will identify patterns in purchase card use to identify fraud and misuse and to maintain good internal controls.	Detecting fraud, waste, and abuse	Planned	Yes	Yes	No
Travel Card Data Mining (SAS) Reports	Will be used to look for patterns in the use of travel credit cards that indicate misuse or fraud and to maintain good internal controls.	Detecting fraud, waste, and abuse	Planned	Yes	Yes	No
Office of Inspector General (OIG)	Analyzes and matches (within the guidelines of the law) Veterans Affairs (VA) files, pertaining to both VA-provided benefits and health care services to detect patterns of waste, fraud, or abuse.	Detecting fraud, waste, and abuse	Operational	Yes	No	No
Veterans Benefits Administration						
C & P Payment Data Analysis	Analyzes compensation and pension data to detect fraud, waste, and abuse.	Detecting fraud, waste, and abuse	Operational	Yes	No	Yes
C & P Large Payment Verification Process	Serves as an internal control intended to make sure that payments over a certain dollar threshold are reviewed to detect potential fraud or abuse.	Detecting fraud, waste, and abuse	Operational	Yes	No	No
Veterans Health Administration						
Primary Analysis and Classification	Is used mainly to discover trends, incidents/events, and vulnerabilities that may exist in VA hospitals.	Improving safety	Operational	No	No	No
Allocation Resource Center Database	Is used in making resource allocation decisions based on the analysis of patient workload and cost data.	Improving service or performance	Operational	Yes	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Veteran's Health Administration (VHA) Financial and Clinical Data Mart	Integrates patient, clinical, and financial data to present a unified management perspective and enable consistent reporting.	Improving service or performance	Operational	Yes	No	No
Decision Support System	Is used to identify patterns of care and patient outcomes linked to resource consumption and costs associated with each patient encounter.	Improving service or performance	Operational	Yes	No	No
Top 50 Standardization Listing/Managed Inventory System	Is used to standardize medical and hospital supplies and equipment to (1) improve VHA's bargaining position when soliciting bids and (2) facilitate the ability to move doctors among hospitals.	Improving service or performance	Operational	No	Yes	No
VISN 16 Data Warehouse	Provides unified view of the VISN 16 VA region, composed of 10 medical centers and 30 outpatient clinics. The system gives a view of the enterprise for management purposes. It is mined for a variety of types of information such as patient encounters, lab tests, pharmacy records, etc.	Improving service or performance	Operational	Yes	No	No

Source: Department of Veterans Affairs.

Table 16: Environmental Protection Agency's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Conceptual Plans to Design an Approach and System to Review Financial Data	Will regularly review financial data systems for contracts, bank cards, and small purchases and other financial databases for misuse or fraud of Environmental Protection Agency's assets.	Detecting fraud, waste, and abuse	Planned	Yes	No	No
Drinking Water Data Warehouse	Integrates and analyzes drinking water information from state, regional, and headquarters sources. Includes data on water systems, compliance, sample analytical results, and audit data.	Monitoring public health	Operational	Yes	No	Yes

Source: Environmental Protection Agency.

Table 17: Export-Import Bank of the United States' Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Integrated Information System Data Warehouse Mining for Financial Risk Information	Is used to generate reports that describe bank lending activities and exposure trends.	Improving service or performance	Operational	Yes	No	No

Source: Export-Import Bank of the United States.

Table 18: Federal Deposit Insurance Corporation's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Real Estate Stress Test (RSST)	Is used to measure real estate risk. Bank examiners use data from the system data as part of a pre-examination planning process to assist in identifying risk concentrations.	Detecting risk in financial systems	Operational	No	No	Yes
Determination of Insured Deposits	Will support the development of a new system for implementing the deposit insurance claims.	Improving service or performance	Planned	Yes	No	No
Statistical CAMELS Offsite Review	Is used to rate financial institutions' performance and risk management practices.	Detecting risk in financial systems	Operational	No	No	Yes
Growth Monitoring System	Is used to identify financial institutions that have experienced significant growth. Serves as an early warning system for detecting financial institutions that might pose financial risk to FDIC.	Detecting risk in financial systems	Operational	No	No	Yes

Source: Federal Deposit Insurance Corporation.

Table 19: Federal Reserve System's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Office of the Inspector General (OIG), Audit Services	Will support audits and evaluations. Using ACL, queries will be run against the board's financial and personnel systems to detect fraud, waste, and abuse, or to provide information supporting any aspect of an OIG project.	Detecting fraud, waste, and abuse	Planned	Yes	No	No

Source: Federal Reserve System.

Table 20: National Aeronautics and Space Administration's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Archiving of Web Information at National Aeronautics and Space Administration (NASA) and Goddard Space Federal Center (GSFC)	Will gather metadata on the GSFC Web site at NASA to preserve NASA legacy information.	Analyzing scientific and research information	Planned	No	No	Yes
My Goddard Search—Mining of Goddard's Web environment	Will allow Web mining of scientific data at Goddard Space Center. It is referred to as "Google for Goddard."	Analyzing scientific and research information	Planned	No	Yes	No
NetContext	Will monitor network traffic for the purpose of identifying bandwidth use, fraud, abuse, and IT security-related activities.	Detecting fraud, waste and abuse	Planned	Yes	No	No
Geophysics Time Series Analysis	Will develop a set of algorithms to identify patterns within temporal activities. The data will be trajectories of objects and movement of objects within images.	Analyzing scientific and research information	Planned	No	No	Yes
"Simmarizer" (Simulation-Based Summary/ Discovery of Knowledge)	Uses data mining techniques to extract knowledge from simulators to understand conditions and scenarios regarding space missions.	Analyzing scientific and research information	Operational	No	No	No
Global Environmental and Earth Science Information System (GENESIS)	Is used to obtain information about global climate changes.	Analyzing scientific and research information	Operational	No	No	Yes
Machine Learning and Data Mining for Improved Intelligent Data Understanding of High Dimensional Earth Sensed Data	Will find patterns and relationships in large, complex earth science data sets, specifically for rare and small events hidden in larger data signals. Will build new capabilities to understand NASA science data.	Analyzing scientific and research information	Planned	No	No	Yes

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Distributed Data Mining Techniques for Object Discovery in the National Virtual Observatory (NVO)	Involves using a data mining tool set for space science research. Incorporates a small number of targeted data mining techniques in order to address specific NASA space science research programs. In particular, the data mining environment will be used to explore NASA's large space science data collections. These techniques are being applied to astronomical object discovery, identification, classification, and interpretation across large multiple distributed astronomy data collections.	Analyzing scientific and research information	Operational	No	No	Yes
Diamond Eye (System for Mining Images)	Analyzes large sets of images looking for specific features.	Analyzing scientific and research information	Operational	No	Yes	Yes
Data Mining of 3-D Numerical Model Forecast Output and Its Application to Atmospheric Research	Will automate the analysis of weather model output, observation, and satellite data to allow for a better understanding of the science of weather dynamics and to predict future weather events.	Analyzing scientific and research information	Planned	No	No	Yes
Ecological Forecasting	Will develop an adaptable system that can be used to mine large volumes of scientific data, identify novel causal relationships in the data about earth system processes, and rapidly incorporate discoveries with biospheric models to generate now-casts and forecasts of biospheric events and conditions.	Analyzing scientific and research information	Planned	No	No	Yes

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Distributed Data Mining for Large NASA Databases (Earth Science Earth Observing System Data)	Will research changes, trends, and relationships in Earth Observing System (EOS) data. The major feature of this activity is that it will allow for different data to be mined in parts and then merged. The capability is needed for instances when scientific data are at different locations. A research quality software will be used to allow for a communication system and run-time environment for applying a collective data analysis approach not bound to any specific platform, learning algorithm, or representation of knowledge.	Analyzing scientific and research information	Planned	No	No	Yes
Discovery of Changes from the Global Carbon Cycle and Climate System Using Data Mining Activity	Will detect patterns in scientific data that are geospatial and dynamic and represented as raster data (gridded cells of surfaces such as the sun's or earth's surfaces). Mining capabilities are being developed for future NASA-relevant data and science.	Analyzing scientific and research information	Planned	No	No	Yes
"AutoSciProd" (Automatic Generation of Science Products from Large Image Data Sets)	Uses statistical and image data to determine and improve science products.	Analyzing scientific and research information	Operational	No	No	No
Near Archive Data Mining of Earth Science Data	Pulls data from an archive of earth science data and applies scientists' analyses and algorithms to the data.	Analyzing scientific and research information	Operational	No	No	No
Spectral Analysis Automation (SAA) System	Will improve the collection, identification, and evaluation of spectral data to better meet scientists' requirements.	Analyzing scientific and research information	Planned	No	No	No
Multiple Sensor Image Registration, Image Fusion and Dimension Reduction Using Wavelets	Will be used for collaborative preprocessing of data and research on wavelets. Will comprise research software that looks at different technologies such as image processing and dimensions.	Analyzing scientific and research information	Planned	No	No	No

(Continued From Previous Page)

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
GMSEC Event Message Data Mining Task	Will be used to determine health of and reasons for problems with satellite systems.	Analyzing scientific and research information	Planned	No	No	No
Intrusion Detection System	Looks at all traffic that traverses NASA's networks' borders.	Improving information security	Operational	Yes	No	No
AvSP/ASMM Foreign Object Detection Toolset	Is used with simulations to identify foreign object damage indicators for commercial jet engines.	Analyzing scientific and research information	Operational	No	Yes	No
Mission and Science Measurement and Discovery Systems	Will be a basic technology research program that will also support infusion of resulting technologies into NASA missions. Purpose of the program is to solve the research challenge in extracting the most scientific knowledge from NASA's space missions and data archives.	Analyzing scientific and research information	Planned	No	No	No
StarTool: Solar Active Region Detection	Is used for recognition of solar activity in sequences of multiband solar images.	Analyzing scientific and research information	Operational	No	No	No
"Toogle" (Times-Series Search Engine)	Searches for time-series data. Is similar to a Google search engine.	Improving safety	Operational	No	No	No
Use of Data Mining, Remote Sensing, and Geographic Information Systems for Wildfire Detection and Prediction	Will help the National Oceanic and Atmospheric Administration automate its fire detection systems and improve the accuracy of fire detection systems.	Improving service or performance	Planned	No	No	Yes
Knowledge Discovery and Data Mining Based on Hierarchical Image Segmentation	Will mine data using software that has been developed to exploit information from a hierarchical image segmentation process.	Analyzing scientific and research information	Planned	No	Yes	Yes

Source: National Aeronautics and Space Administration.

Table 21: Nuclear Regulatory Commission's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Licensee Event Report Data	Identifies nuclear safety trends and patterns in commercial nuclear power events.	Improving safety	Operational	No	Yes	No
Centralized Information Delivery	Will consolidate and standardize reporting for nuclear reactor regulations.	Improving service or performance	Planned	Yes	No	No

Source: Nuclear Regulatory Commission.

Table 22: Office of Personnel Management's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
CRIS Retirement Data Mining Activity	Mines federal employee benefits data such as information on retirement and life insurance to assist in managing federal employee eligibilities and entitlements.	Improving service or performance	Operational	Yes	No	Yes

Source: Office of Personnel Management.

Table 23: Pension Benefit Guaranty Corporation's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Corporate Performance Indicators and Analytics	Will streamline access to management and operational performance measures and permit the correlation of performance and output measures.	Improving service or performance	Planned	No	No	Yes
Corporate Policy and Research Department's Forecasting System	Is a stochastic simulation model that incorporates historic equity and interest rates and bankruptcy possibilities to forecast scenarios for more than 300 pension plans and their related corporate sponsors.	Improving service or performance	Operational	No	Yes	Yes

Source: Pension Benefit Guaranty Corporation.

Table 24: Railroad Retirement Board's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Railroad Retirement Board Data Stores	Consists of two major databases (payment and entitlement history and employment data maintenance) that are mined by actuaries to produce annual actuarial reports and for audit support and quality control.	Improving service or performance	Operational	Yes	No	Yes

Source: Railroad Retirement Board.

Table 25: Small Business Administration's Inventory of Data Mining Efforts

Organization/ system name	Description	Purpose	Status	Features		
				Personal information	Private sector data	Other agency data
Loan Monitoring System	Helps to identify, measure, and manage the risk of Small Business Administration's portfolio. Business credit scores are used but individual credit scores are not.	Improving service or performance	Operational	Yes	Yes	No
MONSTER and Econometric Models	Mines data from database that includes all transactions for each loan that affects SBA subsidy costs, to assist in determining credit subsidy rates for SBA's various credit programs.	Financial management	Operational	Yes	No	No

Source: Small Business Administration.

GAO's Mission

The General Accounting Office, the audit, evaluation and investigative arm of Congress, exists to support Congress in meeting its constitutional responsibilities and to help improve the performance and accountability of the federal government for the American people. GAO examines the use of public funds; evaluates federal programs and policies; and provides analyses, recommendations, and other assistance to help Congress make informed oversight, policy, and funding decisions. GAO's commitment to good government is reflected in its core values of accountability, integrity, and reliability.

Obtaining Copies of GAO Reports and Testimony

The fastest and easiest way to obtain copies of GAO documents at no cost is through the Internet. GAO's Web site (www.gao.gov) contains abstracts and full-text files of current reports and testimony and an expanding archive of older products. The Web site features a search engine to help you locate documents using key words and phrases. You can print these documents in their entirety, including charts and other graphics.

Each day, GAO issues a list of newly released reports, testimony, and correspondence. GAO posts this list, known as "Today's Reports," on its Web site daily. The list contains links to the full-text document files. To have GAO e-mail this list to you every afternoon, go to www.gao.gov and select "Subscribe to e-mail alerts" under the "Order GAO Products" heading.

Order by Mail or Phone

The first copy of each printed report is free. Additional copies are $2 each. A check or money order should be made out to the Superintendent of Documents. GAO also accepts VISA and Mastercard. Orders for 100 or more copies mailed to a single address are discounted 25 percent. Orders should be sent to:

U.S. General Accounting Office
441 G Street NW, Room LM
Washington, D.C. 20548

To order by Phone: Voice: (202) 512-6000
TDD: (202) 512-2537
Fax: (202) 512-6061

To Report Fraud, Waste, and Abuse in Federal Programs

Contact:

Web site: www.gao.gov/fraudnet/fraudnet.htm
E-mail: fraudnet@gao.gov
Automated answering system: (800) 424-5454 or (202) 512-7470

Public Affairs

Jeff Nelligan, Managing Director, NelliganJ@gao.gov (202) 512-4800
U.S. General Accounting Office, 441 G Street NW, Room 7149
Washington, D.C. 20548

United States
General Accounting Office
Washington, D.C. 20548-0001

Official Business
Penalty for Private Use $300

Address Service Requested

Presorted Standard
Postage & Fees Paid
GAO
Permit No. GI00