DEPARTMENT OF HOMELAND SECURITY

Office of Inspector General

Survey of DHS Data Mining Activities

Office of Information Technology

OIG-06-56

August 2006

Office of Inspector General

U.S. Department of Homeland Security
Washington, DC 20528

August 15, 2006

Preface

The Department of Homeland Security (DHS) Office of Inspector General (OIG) was established by the *Homeland Security Act of 2002* (Public Law 107-296) by amendment to the *Inspector General Act of 1978*. This is one of a series of audit, inspection, and special reports prepared as part of our oversight responsibilities to promote economy, effectiveness, and efficiency within the Department.

This report identifies and describes a selection of the Department's data mining and advanced analytics that contribute toward counterterrorism efforts. It is based on direct observations, review of applicable documents, and interviews with Department officials, program managers, and technical staff.

It is our hope that this report will result in more effective, efficient, and economical operations. We express our appreciation to all of those who contributed to the preparation of this report.

Richard L. Skinner
Inspector General

Appendices

Abbreviations

ACE S1	Automated Commercial Environment Screening and Targeting Release S1
ADVISE	Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement
ATS	Automated Targeting System
CBP	United States Customs and Border Protection
CIO	Chief Information Officer
CIS	United States Citizenship and Immigration Services
CVS	Crew Vetting System
DARTTS	Data Analysis and Research for Trade Transparency System
DHS	Department of Homeland Security
DOJ	U.S. Department of Justice
FAS	Freight Assessment System
FEMA	Federal Emergency Management Agency
FLETC	Federal Law Enforcement Training Center
I2F	Intelligence and Information Fusion
ICE	Immigration and Customs Enforcement
ICEPIC	Immigration and Customs Enforcement Pattern Analysis and Information Collection System
IT	Information Technology
NETLEADS	Law Enforcement Analysis Data System
NIPS	Numerical Integrated Processing System
OCIO	Office of the Chief Information Officer
OIA	Office of Intelligence and Analysis
OIG	Office of Inspector General
OLAP	On-Line Analytical Processing
QID	Questioned Identification Documents

RMRS	Risk Management Reporting System
S&T	Science and Technology
TISS	Tactical Information Sharing System
TSA	Transportation Security Administration
USSS	United States Secret Service
US VISIT	United States Visitor and Immigrant Status Indicator Technology
VISAT	Vulnerability Identification Self-Assessment Tool

Tables

Executive Summary

We surveyed the Department of Homeland Security (DHS) to identify and describe data mining activities used to support the counterterrorism mission. Data mining and advanced analytics are evolving technologies that assist in the discovery of patterns and relationships from vast quantities of data. Data mining employs techniques from statistics, machine learning, database management, and visualization. These techniques aid the work of analysts, agents, and investigators and provide knowledge in a manner that aids and informs decision-makers. While various definitions of data mining exist, for the purpose of our survey we defined data mining in a manner to broadly illustrate the range of applications and tools that the Department uses to assist DHS personnel with knowledge discovery, predictive modeling, and analytics.

We identified 12 systems and capabilities that DHS personnel use to perform data mining activities to support DHS' mission of counter-terrorism. Nine systems are operational and three systems are under development. While these data mining activities may perform various processes, we categorized and arranged our descriptions in a way that describes selected data mining processes and tools ranging from basic to advanced analytical tasks. The categories include expert systems, association processes, threat and risk assessment tools, collaboration and visualization processes, and advanced analytics.

Background

While various definitions of data mining exist, we defined data mining to be the process of knowledge discovery, predictive modeling, and analytics. Traditionally, this involves the discovery of patterns and relationships from structured databases of historical occurrences.[1] However, data mining technology has expanded to include different processes, technologies, and methodologies.[2]

Since the early 1900s, prediction has been a central goal of traditional statistics. The tools used for prediction have matured and evolved over time. For example, during the 1980s, analysts in the field of artificial intelligence took advantage of increased computing power to surpass statistical techniques by introducing new methods of prediction,

[1] Traditional data mining infers rules or codes to predict future results via classification or segmentation processes.
[2] Related data mining processes include name matching, and entity, event, and expression extraction from unstructured content such as text, images, audio and video; the clustering of observations or events; aberration or anomaly detection; information matching and sharing via link analysis; visualization; and, the generation of alerts to personnel or other software agents.

classification, and clustering using neural networks, self-organizing maps, genetic and machine learning algorithms capable of pattern recognition in extremely large databases with precise accuracy.[3] Today, data mining activities have been incorporated into sophisticated analytical, modeling, and predictive systems to perform pattern recognition analysis of structured and unstructured data.[4,5]

Analytics and modeling have become so pervasive and essential to private industry that they are the drivers of business intelligence for many different enterprises. Incorporating data mining tools into analytical processes provides benefits to organizations as well as their analysts, such as expanding and entering into new opportunities for business; identifying and maintaining best customer prospects; quickly adapting operations for changes in supply or demand; identifying parameters that influence trends in sales; and, optimizing business operations and performance. Although data mining does not replace the expertise that an analyst provides it automates some of the laborious tasks that an analyst performs, as well as aids in summarizing large quantities of data into meaningful information with which the analyst can work. The roots of these sophisticated intelligence systems are in traditional statistics, machine learning, Internet standards, software agents, and computational linguistics.[6,7]

Some key goals of data mining are: to understand behaviors; to forecast trends and demands; to track performance; and, to transform seemingly unrelated data into meaningful information. Today, private industry and government use data mining as part of their normal course of business, as illustrated in Table 1.

[3] Algorithms provide step-by-step details for particular ways of implementing data mining techniques, such as neural networks, decision trees, self-organizing maps, Bayesian networks, and machine learning.
[4] Structured data refers to sources, which represent a collection of records stored in a computer in a systematic way, with each record organized in a definitive schema, describing the objects that are represented in the database and the relationships among them.
[5] Unstructured data refers to computerized information, which does not have a data structure. This may include audio, video and unstructured text such as e-mails or documents
[6] Software agent is a program that can exercise an individual's or organization's authority, work autonomously toward a goal, and meet and interact with other agents.
[7] Computational linguistics is an interdisciplinary field dealing with the statistical and logical modeling of natural language from a computational perspective. Computational linguistics originated with efforts in the United States in the 1950s to have computers automatically translate foreign languages into English.

Table 1: Common Data Mining Uses

COMMERCIAL USES
• To analyze and segment customer buying patterns and identify potential goods and services that are in demand. • To identify and prevent fraudulent and abusive billing practices. • To analyze sales trends and predict the effectiveness of promotions. • To predict the effectiveness of surgical procedures, medical tests, and medications. • To search information from a number of documents and written sources on a particular topic (text mining). • To identify trends and present statistics in ways that are easily understood and useful.
GOVERNMENT USES
• To monitor expenditures of employee travel and purchase cards. • To quickly access information that speeds up the overall security clearance investigation process for employees. • To identify improper payments under federal benefit and loan programs and help detect instances of fraud, waste, and abuse. • To rank programs quickly by using established performance indicators. • To assist law enforcement in combating terrorism.

Results of Survey

Description of Identified Data Mining Activities

The *Homeland Security Act of 2002* requires DHS to use data mining tools and other advanced analytics to access, receive and analyze law enforcement and intelligence information for the purpose of identifying potential threats of terrorism within the United States.[8] While serving as Assistant Attorney General of the Criminal Division of the U.S. Department of Justice (DOJ), the Secretary of DHS stated that data mining is a promising tool in thwarting terrorism, too.[9] DHS is using data mining to achieve its strategic goals of awareness and prevention. Under the strategic goal of awareness, it is the Department's duty to identify and understand threats, assess vulnerabilities, determine potential impacts and disseminate timely information to its homeland security partners and the public.[10] DHS is also committed to the prevention of terrorism by implementing the technologies and capabilities to detect and prevent terrorist attacks.[11] Advances in pattern recognition, networking, and encryption technologies provide DHS a means by which it can more

[8] Homeland Security Act of 2002, 6 U.S.C. § 121(d)(14) (2002).
[9] *The Financial War on Terrorism and the Administration's Implementation of the Anti-Money Laundering Provisions of the USA Patriot Act: Hearing Before the Senate Comm. on Banking, Housing, and Urban Affairs*, 107th Cong. (2002) (statement of Michael Chertoff, Assistant Att'y Gen., Criminal Div. U.S. D.O.J.).
[10] *Securing Our Homeland: U.S. Department of Homeland Security Strategic Plan*, 9 (February 24, 2004).
[11] *Securing Our Homeland: U.S. Department of Homeland Security Strategic Plan*, 16 (February 24, 2004).

efficiently "connect the dots" to combat terrorism and secure the United States.[12]

We identified 12 systems and capabilities within DHS that support data mining activities. They reside within United States Customs and Border Protection (CBP), Immigration and Customs Enforcement (ICE), Office of Intelligence and Analysis (OIA), United States Secret Service (USSS), and Transportation Security Administration (TSA). Those systems and capabilities perform a variety of functions that contribute toward the counterterrorism effort. Nine systems are operational and three systems are under development.

In the following section, we describe the 12 data mining activities identified during our review. While these activities perform various analytical processes, we grouped them to illustrate five types of analytics performed at DHS: expert systems; association processes; threat and risk assessment tools; collaboration and visualization processes; as well as advanced analytics.

Expert Systems

An expert system is a class of computer programs first developed by researchers in artificial intelligence during the 1970s. In essence, these programs were made up of a set of human-developed rules that analyze information about a specific class of problems, as well as provide analysis of the problem(s), and - depending upon their design - recommend a course of action for the user.

We identified two systems in the Department that are considered expert systems: the Automated Commercial Environment Screening and Targeting Release S1 (ACE S1); and, the Freight Assessment System (FAS). Table 2 summarizes information regarding their respective purposes.

[12] Encryption is the process of obscuring information to make it unreadable.

Table 2: Expert Systems

Expert Systems A group of computer programs comprised of a set of human-developed rules that analyze information about a specific class of problems.		
Data Mining Activity	**Purpose**	**Directorate Mission**
Automated Commercial Environment Screening and Targeting Release S1 (ACE S1)	Identifies the highest risk cargo shipments for further detailed examination by agents and inspectors.	CBP – Prevent terrorists and terrorist weapons from entering the United States by eliminating potential threats before they arrive at our borders and ports.
Freight Assessment System (FAS)	Currently under development. Pre-screens and identifies cargo that has an elevated risk, enabling agents to use efficient inspection methods for cargo.	TSA – Protect the Nation's transportation systems to ensure freedom of movement for people and commerce.

CBP's ACE S1 release is part of a long-term plan for modernizing the screening and targeting of high-risk shipments to assist agents and inspectors at our borders. It employs an expert system, the Automated Targeting System (ATS), that uses electronic shipment data to search criteria that could indicate high-risk cargo. ACE S1 primarily uses custom-built software to perform basic data mining and includes features such as the ability to establish a centralized database to store all screening and targeting criteria and results. Future releases include plans for extracting knowledge out of unstructured data and integrating disparate data and observations and prototypes, such as geospatial event mapping.

While ACE S1 automates the use of information to sort high-risk cargo entering the U.S. and targets it for further examination or inspection, there are some limitations to the system. For example, it uses a business rules engine to enhance screening capabilities for manifest and entry transactions.[13] These business rules have values associated with an individual such as an importer, manufacturer, or broker. Based on these values, the business rules engine makes a request to take some action regarding a transaction, such as perform a document review, conduct an examination, or stop an individual. Business rules can be modified through user committees, which can reinforce or introduce new biases into the rules engine. Currently, ACE S1 does not include an automated mechanism to enable CBP to objectively measure or assess the accuracy, performance, or error rates of the rules. As a result, CBP might not be able to determine if the system does what it was originally developed to do. The system performs analysis based on limited data types, although capable of processing data from other sources.

[13] Business rules engine refers to a set of rules for entering data in a database that are specific to an enterprise's methods of conducting its operations.

The second system we identified as using expert system technology is FAS. TSA is developing FAS to pre-screen cargo before it enters our nation's transportation system. FAS will identify cargo that has an elevated risk. The identified cargo will be flagged and set aside for further inspection by air carriers. To reduce the current reliance on random inspections, FAS plans on using a (human-developed) risk rules engine.

Currently, performance of the risk model and the targeting effort are evaluated through the use of a series of statistical analysis and data quality reports. Without taking into account anomalies, TSA staff will not know if the system and the rules are performing as intended.[14] TSA's future plans include incorporating automated analysis (machine-based rules) and using additional data sources to identify and assess high-risk cargo.

TSA is developing FAS by using historical data on past shipments. Designers are trying to identify unique information elements for pattern recognition. A test of this approach during a pilot phase of FAS revealed the need to use a larger population for testing. Since TSA does not have a large historical database to help it develop indicators to build the rules, TSA plans to develop and incorporate predictive indicators in the future.

One unique problem that TSA faces in using FAS is that it does not regulate shippers.[15] Therefore, the information that TSA has on shippers is provided on a voluntary basis, which limits the amount and type of data that it receives.[16] TSA plans to use additional sources of information, which will aid TSA by verifying the shipping company's identity and legitimacy. Additionally, vetting shippers through additional sources, such as using information from reports of shipper violations, will help TSA identify known shippers.

Association Processes

Association refers to the process of discovering two or more variables that are related. Association, however, does not imply a direct causal connection between the associated variables. This process operates across a variety of platforms and can quickly search through vast sources of data to identify co-existence. It employs algorithms to perform link analysis

[14] Anomaly detection compares a profile of allowed or expected attributes against a population, with any deviation from that profile flagged as a potential risk.

[15] Known shippers are entities that have routine business dealings with freight forwarders or air carriers and are considered trusted shippers. In contrast, unknown shippers are entities that have conducted limited or no prior business with a freight forwarder or air carrier.

[16] According to a TSA official, TSA has issued a Final Rule that will require all air carriers and indirect air carriers to provide all known shipper information to the TSA-managed Known Shipper Management System by December 1, 2006.

and to uncover associations that are normally difficult to detect. For example, association processes can show that persons A, B, and C are at the same location at the same time. Link analysis is used to uncover, interpret, and display relationships between persons, places, and events in a visual format.

We identified four data mining activities that use association processes to perform analysis. These are Data Analysis and Research for Trade Transparency System (DARTTS), Immigration and Customs Enforcement Pattern Analysis and Information Collection System (ICEPIC), Law Enforcement Analysis Data System (NETLEADS), and Crew Vetting System (CVS). Table 3 summarizes information regarding their respective purposes.

Table 3: Association Processes

Association Processes **The process of discovering two or more variables that are related.**		
Data Mining Activity	**Purpose**	**Directorate Mission**
Data Analysis and Research for Trade Transparency System (DARTTS)	Assists agents in identifying and detecting money laundering, drug trafficking, and other illegal activities through financial transactions.	ICE – Protect the United States and uphold public safety by identifying criminal activities and eliminating vulnerabilities that pose a threat to our nation's border, as well as economic, transportation, and infrastructure security.
Immigration and Customs Enforcement Pattern Analysis and Information Collection System (ICEPIC)	Enables investigators to conduct targeted checks of non-resident aliens and provides leads for the disruption of potential terrorist activities.	
Law Enforcement Analysis Data System (NETLEADS)	Supports agents in identifying criminal activity patterns and trends and associations among criminal organizations.	
Crew Vetting System (CVS)	Assists analysts in screening air carrier personnel to ensure the security of air transportation.	TSA – Protect the Nation's transportation systems to ensure freedom of movement for people and commerce.

DARTTS is a legacy system of the former U.S. Customs Service. It was developed to assist agents in identifying and detecting money laundering, drug trafficking, and other illegal activities. This system now resides under ICE. This small-scale, stand-alone system uses commercial off-the-shelf software to aid in the analysis of structured data found in databases. System owners from the ICE Financial and Trade Investigation's Trade Transparency Unit are collaborating with the ICE Office of the Chief Information Officer and Information Systems Security Manager to provide DARTTS to users in a web environment.

DARTTS allows for data mining and analysis that is not available in other systems. For instance, this system allows the user to produce aggregate

totals for importation of currency, and then sort based on any number of variables, such as country of origin, party name, or total currency value. The ICE Financial and Trade Investigations Division's Trade Transparency Unit also uses DARTTS as its platform for sharing and analyzing U.S. and foreign trade data, pursuant to Customs Mutual Assistance agreements. This allows the user to see the "bigger picture"--permitting investigators to identify anomalies that can be indicative of trade-based money laundering or other import-export crimes.

The second association process, ICEPIC, was developed to assist investigators in meeting the goal of disrupting and preventing terrorism activities. ICEPIC is a small-scale system that employs a variety of commercial off-the-shelf software and government off-the-shelf software to support criminal investigators. ICEPIC employs matching to integrate and confirm information from structured data sources in DHS databases. It uses associations for discovery of patterns and relationships. The system connects to ICE's network to aid investigators in generating leads, conducting batch queries of names, and reporting.

ICE uses a third system, NETLEADS, in the area of criminal investigations. It is a web-enabled intelligence and investigations analysis database repository. NETLEADS tools give users the capability to rapidly search and conduct analysis. It is designed to support agents in identifying criminal activity patterns, trends and associations among criminal organizations. It uses commercial off-the-shelf products to discover relationships within data.

The system consists of fifty million indexed names, intelligence, subject records, investigation reports, and global intelligence information on topics such as smuggling, terrorism, and transnational trends. NETLEADS provides an integrated common interface for querying intelligence and enforcement applications and data.

NETLEADS includes hosted data marts of intelligence and investigative information.[17] The system was built based on proprietary design and uses associations and name matching, and queries data. Investigators and analysts use it for identifying connections between persons of interest or persons under investigation. The users can access multiple government and commercial databases to discover patterns and relationships. Benefits of this system include providing the capability to search multiple

[17] A data mart is a specialized version of a data warehouse. Like data warehouses, data marts contain a snapshot of operational data that helps people to strategize based on analyses of past trends and experiences. In contrast to a data warehouse, the creation of a data mart is predicated on a specific, predefined need for a certain grouping and configuration of selected data.

databases and assemble information into a common analytical environment.

The last system we identified as using an association process is CVS. TSA uses this system to screen commercial air carrier personnel to help ensure the security of air transportation. CVS encompasses the Flight Crew Manifest and Master Crew List vetting. These two sub-programs ensure that 100% of the crew members of commercial air carriers flying into, out of, over, and through the U.S. are vetted on a recurring basis. The Flight Crew Manifest process ensures that the actual people on the aircraft are vetted each time they fly. The Master Crew List population is subject to perpetual vetting, therefore whenever there is any change in data or placement on or removal from a watch list, the system triggers a notification that new data has resulted in a match, which is then referred for appropriate action.

CVS uses proprietary software to perform data element matching of names, social security numbers, passports, and other pertinent information. It also uses commercial-off-the shelf products for error checking and message processing and logging.

Threat and Risk Assessment Tools

Risk management is the process of identifying risk, assessing risk, and taking steps to reduce risk to an acceptable level. Threat and risk assessments are widely recognized as decision support tools to establish and prioritize security program requirements. A risk-based approach allows organizations to make better judgments about where to deploy resources and where to prioritize protection efforts.

We identified one tool, the Risk Management Reporting System (RMRS), which uses a threat and risk-based approach to analyze results and conduct data mining activities. RMRS generates risk-assessed scores for assets based on advanced analytics. Table 4 summarizes information regarding its purpose.

Table 4: Threat and Risk Assessment Tools

Threat and Risk Assessment Tools **Decision support tools that enable the deliberate, analytical approach to identify which threats can exploit vulnerabilities in an organization's specific assets.**		
Data Mining Activity	**Purpose**	**Directorate Mission**
Risk Management Reporting System (RMRS)	Collects information and generates a score based on level of risk for assets comprising our Nation's critical infrastructures.	TSA – Protect the Nation's transportation systems to ensure freedom of movement for people and commerce.

RMRS is a stand-alone system that TSA uses to store and analyze risk. Because RMRS is flexible and modular, it has the capability to store and analyze risk for assets including, but not limited to, maritime facilities and vessels, airports, mass transit, and public assembly facilities such as stadiums and arenas. RMRS generates a score based on the level of risk associated with a particular asset. The information that RMRS has in its database is reported by facility managers, security personnel, and law enforcement agents and entered into one of the available tools associated with RMRS, such as TSA's Vulnerability Identification Self-Assessment Tool (VISAT).

VISAT is a voluntary, on-line assessment tool that helps transportation asset owners and operators enhance the security for assets including maritime vessels, heavy railways, subways, rail stations, and highway bridges. The goal of VISAT is to raise the level of security awareness in public assembly facilities nationwide and to establish a common baseline of security awareness from which these facilities can build their protection plans. In addition to VISAT, RMRS captures asset information from tools that TSA field inspectors use when conducting independent inspections. RMRS processes the information captured from these tools using algorithms to generate a level of impact score to assess the likelihood of a terrorism attack, attractiveness of the target, and the consequences of an act of terrorism. RMRS also captures and analyzes criticality information, including potential life-threatening, economic, and psychological impacts from threat scenarios when applied to particular assets.

Collaboration and Visualization Processes

Collaboration and visualization processes assist agents, analysts, and investigators in collecting and analyzing information. Collaboration is the strategic management process of collecting, tagging, classifying, organizing, and applying an organization's internal content and expertise. Visualization processes aid in analyzing data sets by defining views, highlighting findings, navigating on specific features to find trends, and pinpointing exceptions. Data visualization simplifies the presentation of information while maintaining the integrity and depth of the information. These technologies can also easily deal with very large and highly non-homogeneous amounts of data.

We identified four processes that primarily use collaboration or visualization to perform data mining activities. These are Intelligence and Information Fusion (I2F), Numerical Integrated Processing System (NIPS), Questioned Identification Documents (QID), and Tactical

Information Sharing System (TISS). Table 5 summarizes information regarding their purposes in relation to the respective missions.

Table 5: Collaboration and Visualization Processes

Collaboration and Visualization Processes **The strategic management process of collecting, tagging, classifying, organizing, and applying an organization's internal content and expertise Presents the data sets by defining views, highlighting findings, flagging exceptions, etc in large data sets.**		
Data Mining Activity	**Purpose**	**Directorate or Critical Agency Mission**
Intelligence and Information Fusion (I2F)	Currently under development. Will provide information analysts with state-of-the-art analysis tools that aid in the discovery and tracking of terrorism threats to the U.S. population and infrastructure.	OIA – Gather, analyze, and disseminate information in a mission-oriented manner.
Numerical Integrated Processing System (NIPS)	Assists agents in identifying anomalies indicative of criminal activity, such as immigration violations, customs fraud, export violations, drug smuggling, and terrorism.	ICE – Protect the United States and uphold public safety by identifying criminal activities and eliminating vulnerabilities that pose a threat to our nation's border, as well as economic, transportation, and infrastructure security.
Questioned Identification Documents (QID)	Allows analysts to compare questionable documents against genuine documents such as passports and drivers licenses.	USSS – Protect key individuals and investigate crimes related to counterfeiting and financial sector, including identity theft, computer fraud, and cyber attacks.
Tactical Information Sharing System (TISS)	Captures observations of suspicious activities in aviation and provides law enforcement officials with information for examining long-term trends and patterns.	TSA – Protect the Nation's transportation systems to ensure freedom of movement for people and commerce.

The purpose of the I2F is to make operational an integrated intelligence and information capability for DHS. This capability will enable intelligence analysts to understand relationships that would otherwise not be readily apparent. I2F is in early development and is primarily dependent on the analyst manually processing, compiling, and analyzing data. The next version of the system will be a set of tools and technologies integrated to support the intelligence analyst.

I2F provides intelligence analysts with tools that aid in the discovery and tracking of terrorism threats to the United States population and infrastructure. I2F is principally made up of commercial off-the-shelf software, but also integrates government off-the-shelf programs. These

programs are used for entity extraction, search capabilities, and link analysis.[18]

The second system, NIPS, is a web-based strategic analytical tool used by DHS agents and analysts to manipulate, compare, and analyze large data sets of commercial, passenger, financial, and enforcement data. Stakeholders in this project include the ICE Intelligence Division, ICE Field Intelligence Units, ICE Office of Investigations, CBP Intelligence Division, CBP Office of Field Operations, CBP National Targeting Center, Container Security Initiative, and Customs Trade Partnership against Terrorism.

NIPS enables users to identify anomalies that are indicative of possible criminal activity, including illicit actions in support of terrorism, money laundering, tax evasion, weapons proliferation, immigration violations, and drug smuggling. Using an On-Line Analytical Processing (OLAP) tool, NIPS leverages a manual approach to quickly respond to ad hoc queries across multiple databases.[19]

Functional capabilities include link analysis, rule based intelligence, power search, summary reports, and geospatial integration. According to a senior ICE official, NIPS greatly enhances the overall capability of DHS to identify, target, and disrupt potential acts of terrorism.

The third system that uses collaboration is QID. QID assists analysts in evaluating whether a document is counterfeit. Security personnel at airports, seaports, and borders use QID as a validation tool to view samples of internationally-issued identity documents. Specifically, it allows analysts to compare questionable documents against genuine documents, such as passports and drivers licenses. QID has Intranet connectivity as well as the capability to identify associations and patterns to determine if documents are genuine. This process provides suggestions to the analyst on what to look for in making a determination about the authenticity of a questioned document.

The fourth system that uses collaboration is TISS. This system facilitates the collection of tactical information related to suspicious activity. It uses a combination of technologies and domain awareness, so there is greater capability to observe and report suspicious activities that may indicate

[18] Extraction is a technique for reducing the number of attributes used in the processes of classification, segmentation, and pattern recognition. Within a document, entity extraction can also be customized to recognize pattern-based, list-based entities, events, and relationships.

[19] OLAP enables users to analyze data across multiple dimensions, usually reserved to key business metrics such as products, departments, regions, and time segments.

terrorist planning. TISS captures Federal Air Marshal Service observations of suspicious activity in aviation. This system enables Federal Air Marshals and other law enforcement officers in the field to report these observations into the database for analysis using the TISS Analytic Tool.

In addition, TISS allows Federal Air Marshals to prepare and immediately submit reports of suspicious activity to the Federal Air Marshal Service Investigations Division. For example, an individual taking pictures of a building or videotaping the operations of a maritime port may raise suspicions compared to others who normally work or visit such locations. This suspicious activity can be noted by security or law enforcement officers and immediately entered into the system. As another example, if an individual attempted to pass through airport security screening with a handgun at two separate airports and on different occasions, the surveillance reports that were entered into TISS could be used by the Federal Air Marshal Service to establish a link between the incidents.

Advanced Analytics

Science and Technology (S&T) is developing an advanced analytics capability called Analysis, Dissemination, Visualization, Insight and Semantic Enhancement (ADVISE), as described in Table 6. ADVISE is an advanced information technology that can integrate information and facts from many different types of data. Since ADVISE is a "technology framework," it can be tailored and deployed for specific purposes and areas of interest. For example, it is being developed to incorporate chemical, biological, radiological, nuclear, and explosive threat and effects data. It is intended to ingest data from a variety of sources, ranging from highly structured content, such as database records, to unstructured content, such as message traffic. Still in development, ADVISE will connect information extracted from text and images, databases, and simulation and modeling tools to provide a watch-and-warning system for analysts.

ADVISE employs semantic graphs to determine relationships and patterns among data and multiple visualization techniques to display the resulting information.[20] The Department seeks to predict threat and vulnerabilities, such as through the detection of relationships between seemingly disjointed entities. Semantic graphs organize data entities regarding threats and vulnerabilities and link their relationships. Thus, hidden relationships in the data are uncovered by examining the structure and properties of the

[20] A semantic graph is a network of heterogeneous nodes (a point at which two lines or systems meet or cross) and links. Because these graphs are ideal for representing relationship and linkage information, they have emerged as a key technology for organizing DHS data.

semantic graph. For example, a simple semantic graph can link people, workplaces, and towns as well as indicate a relationship with various friends. Studying the links can assist in understanding the relationships between entities, and help identify threats and vulnerabilities. S&T expects ADVISE's ability to apply the capabilities of semantic data fusion, link analysis, and unstructured text analysis will be a powerful capability that will allow analysts to find the expected and discover the unexpected.

Table 6: Advanced Analytics

Advanced Analytics Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement		
Data Mining Activity	**Purpose**	**Component Mission**
Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement (ADVISE)	Currently under development. Integrates the various information analysis and synthesis, visualization, and knowledge discovery component capabilities. Will incorporate comprehensive chemical, biological, radiological, nuclear, and explosive threat and effects data.	S&T- Protect the homeland by providing Federal and local officials with state-of-the-art technology and resources.

Management Comments

We provided a draft version of this report to DHS' OCIO and requested written comments from this office. In response, the OCIO indicated that it had no comments on the draft report.

Our objective was to identify and describe DHS' data-mining activities. We researched various definitions and prepared a definition of data mining that focused on applications and tools used for knowledge discovery and analytics. We shared this definition with interviewees during our survey.

To obtain a selection of data mining activities for review, we extracted an initial list of systems using key word searches on the DHS' Trusted Agent FISMA database. We refined the list by removing systems relating to administrative functions or efforts that did not contribute toward DHS counterterrorism efforts. We sent components lists to verify whether the specified systems performed data mining activities and to add applicable systems that were not on the list. Officials from FEMA, FLETC, and US VISIT responded that they do not conduct any data mining activities.

We interviewed officials from the OCIO and the Privacy Office. We conducted interviews with CIOs, program managers, and officials representing CBP, CIS, FEMA, ICE, OIA, S&T, TSA, and USSS. The purposes of our interviews were to clarify the objective of the system, describe data mining activities, and obtain documentation. We collected and reviewed technical information and documentation through a data call to database administrators regarding the system, including tools and techniques used to conduct data mining activities, too.

We conducted fieldwork in the Washington, DC metropolitan area. Our analysis is based upon direct observation, review of applicable documentation, and interviews. We conducted our survey from November 2005 through April 2006 under the authority of the *Inspector General Act of 1978*, as amended, and according to generally accepted government auditing standards.

The principal OIG points of contact for the survey are Frank Deffer, Assistant Inspector General for Information Technology Audit (202) 254-4100 and Marj Leaming, Director Special Projects Audit Division (202) 254-4172. Major OIG contributors to the survey are identified in Appendix B.

Special Projects Division

Marj Leaming, Director
Barbara Ferris, Audit Manager
Audilia Wedderburn, Auditor
Juliana Meek, Student Temporary Employment Program
Scott Binder, IT Auditor

Technical Consultants

Jesus Mena, Data Mining Consultant
Michael Pridgen, Database Consultant
Richard Streeter, Database Consultant

Advanced Technology Division

Michael Goodman, Security Engineer

Department of Homeland Security

Secretary
Deputy Secretary
Chief of Staff
Deputy Chief of Staff
General Counsel
Executive Secretary
Assistant Secretary for Policy
Assistant Secretary for Public
DHS Legislative and Intergovernmental Affairs
DHS GAO OIG Audit Liaison
Director, Operations Directorate
Chief Privacy Officer
Chief Information Officer
Component Chief Information Officers

Office of Management and Budget

Chief, Homeland Security Branch
DHS OIG Budget Examiner

Congress

Congressional Oversight and Appropriations Committees, as appropriate

Additional Information and Copies

To obtain additional copies of this report, call the Office of Inspector General (OIG) at (202) 254-4100, fax your request to (202) 254-4285, or visit the OIG web site at www.dhs.gov/oig.

OIG Hotline

To report alleged fraud, waste, abuse or mismanagement, or any other kind of criminal or noncriminal misconduct relative to department programs or operations, call the OIG Hotline at 1-800-323-8603; write to DHS Office of Inspector General/MAIL STOP 2600, Attention: Office of Investigations–Hotline, 245 Murray Drive, SW, Building 410, Washington, DC 20528; fax the complaint to (202) 254-4292; or email DHSOIGHOTLINE@dhs.gov. The OIG seeks to protect the identity of each writer and caller.