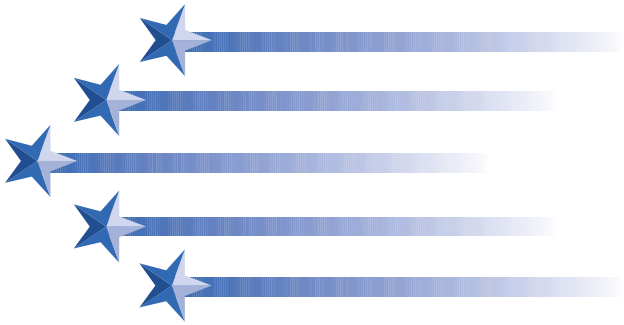
SAVER Reprint



Data Mining and Analysis Tools Operational Needs and Software Requirements Analysis

Original report published by
Space and Naval Warfare Systems Center, Charleston

System Assessment and Validation for Emergency Responders
Sponsored by the U.S. Department of Homeland Security
Office of State and Local Government Coordination and Preparedness

SAVER Reprint
Data Mining and Analysis Tools Operational Needs and Software Requirements

September 2005

This document was published by the Space and Naval Warfare Systems Center, Charleston, and reprinted by the SAVER Program Support Office. Questions about the content should be directed to the SAVER Program Support Office.

SAVER Program Support Office
E-Mail
Web

July 2005

SPAWAR

Systems Center Charleston

Data Mining and Analysis Tools

Operational Needs and Software Requirements Analysis

Version 1.0

SPAWAR Systems Center, Charleston
P.O. Box 190022
North Charleston, SC 29419-9022

July 2005

Data Mining and Analysis Tools

Operational Needs and Software Requirements Analysis

Version 1.0

SPAWAR Systems Center, Charleston
P.O. Box 190022
North Charleston, SC 29419-9022

This Project was funded under Interagency Agreement #2003-TK-R-040, from the U.S. Department of Homeland Security, Office of State and Local Government Coordination and Preparedness, Systems Support Division.

Points of Contact

National Urban Security Technology Laboratory
U.S. Department of Homeland Security
Science and Technology Directorate
201 Varick Street
New York, NY 10014

SPAWAR Systems Center Charleston
Law Enforcement Advanced Technology Branch
P.O. Box 190022
North Charleston, SC 29419-9022

Foreword

The U.S. Department of Homeland Security's, Office of State and Local Government Coordination and Preparedness (SLGCP) established the System Assessment and Validation for Emergency Responders (SAVER) Program to assist emergency responders in performing their duties. The mission of the SAVER Program is to:

- Provide impartial, relevant, and cost-effective evaluation and validation of equipment and software
- Enable decision makers and responders to better select, procure, use, and maintain equipment and software
- Evaluate and validate the interoperability of products within a system, as well as systems within systems
- Provide feedback to the user community through a well-maintained, Web-based database.

As a SAVER Program partner, the Space and Naval Warfare Systems Center (SPAWARSYSCEN), Charleston has been tasked by SLGCP to provide expertise and analysis on key subject areas including communications, sensors, perimeter security, weapon detection, and surveillance, among others. In support of this tasking, SPAWARSYSCEN Charleston conducted a study on Data Mining and Analysis tools. This study will help first responders make informed decisions regarding acquisition and utilization of Data Mining and Analysis tools. The following Operational Needs and Software Requirements Analysis presents a portion of the study's findings. Visit the SAVER Web site for more information on this and other studies.

The SAVER Program is focused on evaluating processes and procedures for components as well as establishing system-level interoperability. The SAVER Program databases, processes, and expertise are a resource available to emergency responders at a national level. SAVER is the place to find answers to critical emergency responder needs. This sharing of information will be a life-saving and cost-saving asset to the U.S. Department of Homeland Security, as well as to federal, state, local, and tribal users of emergency response equipment.

TABLE OF CONTENTS

1. Introduction

This report is intended for Homeland Security Professionals (HSPs) at state, regional, or specialty fusion centers, or federal agencies, who are involved in data mining and analysis. We believe that these HSPs must have access to specific information housed in regional, federal, or disparate databases in order to perform data mining and analysis activities that will produce actionable information.

But, what are data mining and analysis, and why are these processes important to HSPs? *Data analysis* is the process of examining individual elements, or relationships between elements, in order to draw conclusions. *Data mining* is the process of using intelligent methods to extract knowledge from structured or unstructured data. The data mining process uses rules based methods to uncover information, patterns, and trends that can be used to predict future events, find new associations between events, or organize related data in new ways. HSPs, using link analysis and visualization tools on mined data, can expose suspect or terrorist associations and suspected connections. This is one example of the many applications of data mining and analysis.

Realizing the importance of data mining and analysis activities, the Office of State and Local Government Coordination and Preparedness (SLGCP) tasked the Space and Naval Warfare Systems Center (SPAWARSYSCEN), Charleston to assess and evaluate data mining and analysis tools that support homeland security and public safety organizations. SPAWARSYSCEN Charleston responded to this task by forming the Data Mining and Analysis Tools (DMA) study. This study provides information that enables HSP decision makers and practitioners to better identify and select data mining and analysis tools that meet their requirements.

This report, the *Operational Needs and Software Requirements Analysis*, is a product of the Data Mining and Analysis Tools study. It sets the foundation for the remainder of the study in that it establishes the audience, their operational needs, software requirements, and data source considerations based on inputs gathered directly from the HSP community. This report is most effective when used in conjunction with the study's other publications. The compilation of documents will help HSPs understand the need for and uses of data mining and analysis tools, what software tools are currently available, and aid in selecting the appropriate software for their specific needs.

2. Methodology

The study's first step was to understand the HSP community's data mining and analysis needs by reaching out to the community itself. A data-gathering and analysis process involving HSPs produced the operational needs and software requirements that appear in this report. The process involved: 1) identifying Subject Matter Experts (SMEs) in the HSP community, 2) generating a list of operational needs, 3) validating the operational needs, 4) reverse-engineering the software functional requirements based on the HSP's operational needs, and 5) validating the software functional requirements.

A SME is defined as an experienced technical and operational practitioner that has a broad understanding and application of data mining and analysis processes and tools. SLGCP identified several members from the InterAgency Board (IAB) for Equipment Standardization and InterOperability Working Group as experts in Homeland Security. The IAB is organized into six subgroups, which are staffed with SMEs in that subgroup's area of interest. Other data mining practitioners were also identified as SMEs to broaden the input base.

After identifying the SMEs, project members conducted surveys with the SMEs via e-mail and phone to collect information on operational needs. The completed surveys described the SMEs' agency functions and data mining and analysis requirements. Extensive online research was then conducted to support the identified operational needs. Further clarification of agency needs was made through telephone conversations with SMEs. These efforts culminated in an operational needs list.

Project members met with study participants, who included members of the Interoperable Communications & Information Systems (ICIS) subgroup of the IAB, to validate the operational needs list in a focus group meeting conducted on October 7, 2004, in Oklahoma City. A number of new operational needs were identified and existing operational needs were confirmed during the focus group meeting. The updated operational needs list was sent to select SMEs for validation following the focus group meeting.

Extensive research identified the data mining and analysis tools currently available, both commercial off-the-shelf and government off-the-shelf. From this research, a list of vendors and available software functions and features was identified. In some cases the operational needs directly related to the functions and features offered by vendors. The remaining operational needs were analyzed and several additional software requirements were derived.

The remainder of this document describes the findings from these efforts. Initial findings indicated that data mining and analysis tools are most useful at state, regional, or specialty fusion centers or at the federal level. Public safety officers use the results of data mining and analysis. A detailed list of operational needs in Section 3.2 was derived from interactions with SMEs. While there were many operational needs, they map to general software requirements. These derived software requirements can be found in Section 3.3. Finally, a list of possible data sources is in Section 3.4.

3. Operational Needs and Software Requirements

This section describes the levels of operations, the operational needs, derived software requirements, and data sources. This information is important because it provides a basis for the remainder of the DMA study. The information was used to select data mining and analysis software tools included in the *Data Mining and Analysis Tools: Product Catalog* and in developing a method for choosing a tool that best meets the Homeland Security organizations operational needs.

3.1 Levels of Operations

The knowledge gained from the survey responses, phone conversations, and focus group discussion resulted in a better understanding of HSPs' needs for data mining and analysis tools. An overall theme that emerged from this analysis is that *needs differ between operational levels*.

Generally, public safety officers at the operations level do not need to perform data mining and analysis as it is beyond the range of their authority or responsibility. However, they need accurate and reliable information from data mining and analysis activities to make the best on-scene decisions. Investigators need the capability to mine data for associations between people, aliases, places, organizations, social networks, behavioral and cultural patterns, and events. State and local agencies are developing data sharing and analysis systems to function in regional fusion centers, to broaden the scope of mining capabilities. At the federal level, data analysis and data mining are routine.

The following table shows how HSPs data mining and analysis needs may vary depending upon the operational level and function or role.

Levels of Operation	Function/Role	Data Mining and Analysis Need
Operations	Public Safety Officers (e.g., law enforcement officers, firefighters, EMS personnel)	Provide analysis results to operations personnel.

Levels of Operation	Function/Role	Data Mining and Analysis Need
Investigations	Intelligence Officers (persons who gather and analyze technical intelligence)	Ability to associate people, places, and events to link related data.
State, Regional and Specialty Fusion Centers	Intelligence Analysts, Criminologists	Ability to associate people, places, and events across state lines.
Federal	Specialized Analysts	Ability to access global data to protect against international terrorism.

3.1.1 Operations

Operations personnel, or "operators," such as law enforcement officers and firefighters, are front-line public safety officers. These operators both contribute to the databases being mined, as well as receive the actionable results from data mining and analysis. For example, law enforcement officers typically use information contained in regional databases or multi-state databases to positively identify suspects and find known associates. Public safety officers may also require information from these data sources as an incident is in-progress. Information that identifies unexpected hazards (e.g., chemicals stored within a building, buried utilities), obstacles (e.g., the number of people in a building, the building's structure), or other public infrastructure (e.g., hospitals, schools) is critical in deciding on the proper response. In these cases, the operators need reliable, accurate, and credible information from data mining and analysis tools to augment their current capabilities.

Currently, there is a trend in the homeland security community to develop regional clearinghouses for data and analysis in support of nearby, smaller jurisdictions with limited manpower. In these circumstances, agencies with regional and area databases access information through Web browsers that lead users through pre-defined queries. These pre-defined data analysis queries do not provide the ability to perform custom, often times content based, queries. The homeland security community would prefer to possess the flexibility to access, structure, and tailor easy-to-use queries for the specific and possibly unique aspects and circumstances of the emergency based on "content analysis versus data analysis".

There are various methods by which operations personnel can contribute to the larger scope of data mining, even if their own organization does not have the tools or skills required to access and mine data. One way is for local agencies to "pool" their data with others for regional data sharing, analysis, and mining. Public safety officers contribute to databases through their records management systems, which often contain information such as identities of suspects, locations of fires, and suspicious activities in their area. There are also several options available for small agencies that do not employ analysts, but that need data analysis. These agencies can hire analysts with the skills to perform more complicated queries, or they can work through a "trusted" agent, meaning that a group of agencies pool

their resources to share an analyst who performs the more complicated queries and analysis.

3.1.2 Investigations

Intelligence analysts, such as detectives or firefighters, realize the volume of data they access can answer many more questions than pre-defined queries allow, and consequently, build ad hoc queries to satisfy those questions. The intelligence community has initiated programs that direct intelligence analysts to work collaboratively with analysts to build broader menu-driven and user-friendly ad hoc queries. This effort promotes flexible investigative data mining and analysis for discovering associations and actionable information. We envision this level of collaboration to become normal and routine, especially when agencies understand the breadth and wealth of information potentially available to them by following the process. The benefits will be greater information sharing, collaboration, and trust between cultures and communities that have only limited historical experience working together. Moreover, the net effect will be increased effectiveness from the acquisition and transport of data and information to more efficient and safer operations, and thus, a win-win formula for the successful quest for actionable information, intelligence, and investigative material. This becomes increasingly important in today's environment where investigators and analysts are being called upon to not only understand and explain the environment and circumstances, but to predict future events by applying pattern and link analysis to existing data.

3.1.3 State, Regional, and Specialty Fusion Centers

Recently, more emphasis has been placed on mining information across multiple states. The ability to do this across state lines provides a means to proactively respond to threats or criminal incidents quickly. Generally, state-level Public Safety Officers fulfill their mining and analysis needs by accessing regional, multi-state, or federal systems. They use these systems to link suspects with known associates, events, or locations. Frequently, they input information to and retrieve information from the Department of Motor Vehicles (DMV), local case investigations, and Department of Justice databases.

In one scenario, a suspect is pulled over in Ohio for driving under the influence of alcohol. The suspect's vehicle has no plates, but the driver has a Pennsylvania driver's license. In a multi-state system, the officer would enter the driver's license number and the system would access both Pennsylvania's DMV to positively identify the driver, and the state's criminal history records to determine if there is an outstanding warrant. In addition, a multi-state system has the potential to query all participating states' DMV and criminal history records, possibly leading to the capture of a wanted person.

In that regard, the Multistate Anti-Terrorism Information Exchange (MATRIX) is an existing pilot program for multi-state data sharing and analysis, specifically aimed at increasing and enhancing the exchange of criminal activity information

among local, state, and federal law enforcement agencies. The MATRIX project provides data analysis capabilities using existing data sources and integrates disparate data from many types of storage systems.[1] MATRIX data sources include criminal history records, Department of Corrections records, sexual offender records, driver's licenses database, and motor vehicle registration.

One example of a regional fusion center is the Regional Information Sharing Systems (RISS) program, which shares criminal justice data among several states. The RISS program is composed of six regional centers that share intelligence and coordinate efforts against criminal networks that operate in many locations, and across jurisdictional lines. RISS provides the ability to analyze case data by connecting subjects to criminal events. Local, county, state, and federal agencies use RISS to identify gangs, firearms trafficking, and violent criminal activities across the United States.

Specialty centers discover links between people, places, and events. The Terrorism Early Warning Group (TEWG) network is an example of a specialty center. TEWG combines criminal and operational intelligence information and aims to provide terrorist warnings. State and local counterterrorism offices use the TEWG network. The TEWG searches for patterns and associations using commercial data mining tools on structured and unstructured data. Suspicious patterns and associations are indicators or warnings that the information warrants further investigation. When the TEWG discovers suspicious persons and/or associations or patterns, the Joint Terrorism Task Force (JTTF), an FBI program, is notified of the findings in order to continue the investigation.

3.1.4 Federal

The Federal Government uses data mining and analysis tools to support one of the country's top priority efforts – fighting terrorism. By introducing the Patriot Act, the President and Congress enhanced the ability of the FBI and other law enforcement agencies to investigate and share information with other agencies. During an investigation, the agencies conduct threat analyses and create data collection strategies using the latest data mining and analysis tools available.

The Federal Government is in charge of predicting when and where the next terrorist attack will occur, who the terrorists will be, and how they will attack. In order to prevent terrorist attacks, they must be able to positively identify potential terrorists. One means of identifying suspected terrorists is by monitoring the whereabouts of people on watch lists. This task includes constant data mining and analysis of individuals' travel patterns through transactional airline data, the National Security Entry-Exit Registration System, and U.S. Visitor & Immigrant Status Indication technology. Link analysis tools are useful in finding common sets of associations between watch list individuals and clues for how the plot may occur. One possible scenario is that several individuals on a watch list have crop duster pilot licenses and have been purchasing large quantities of fertilizer in the United States. Knowing the travel plans, "the where and when" of the associated suspects, may be enough to put the plot together. Once the suspected plot is

known, government officials can notify the appropriate airline authorities, such as Transportation Security Agency, immigration, and airline personnel. The airline authorities are expected to take action based on this actionable information.

The Terrorist Threat Integration Center (TTIC) houses a group of federal agencies (FBI, CIA, DHS, DOD, NSA, NIMA, and OMB) that are tasked to lead the war on terror. The TTIC task directly relates to data mining and analysis, and the distribution of new actionable information. These agencies have unfettered access to intelligence information. One of their tasks is to maintain an up-to-date database of known and suspected terrorists that appropriate officials at all levels of government can access. TTIC analysts use the latest tools for a wide range of analysis needs.[2] These data mining and analysis tools help analysts to collaborate, produce reports, perform link analysis, and apply geospatial techniques. TTIC uses Web technology, supported by information exchange standards, to search across multiple databases and create knowledge from existing information.[3] Specifically, they use the World Wide Web, XML, SOAP and a host of other Web tools to access a wide range of open sources and other data.

3.2 Operational Needs

The following table describes the data mining needs and capabilities generated and validated by the SMEs involved in this study. These needs were derived from user surveys; a focus group; the Markle report on *Creating a Trusted Network for Homeland Security*[4]; and the Memorial Institute for the Prevention of Terrorism, *Project Responder Interim Report: Emergency Responders' Needs, Goals, and Priorities*.[5] The shaded numbers (entries highlighted in green and hash-marked) in the table below were identified during the SME validation process as the priority needs. The numbers map the operational needs to the software requirements in Section 3.3. These numbers do not indicate the rank or order of importance.

No.	Operational Needs
1	Credible threat reports (within 4 hours) gained through continual mining of resource and infrastructure data, including private sector data relating to the status of that threat.
2	Common operational picture of the critical infrastructure status.
3	Locate (within 5 minutes) critical infrastructure nodes, such as pipelines, power-generation plants and transmission lines, and communications facilities, in the vicinity of an attack.
4	Real-time credible threat information dissemination to all relevant jurisdictions, levels of government, and disciplines (technical data as well as observations from law enforcement officers, firefighters, epidemiologists, and other responders).
5	Information sharing across all levels of security from "Top Secret/Code Word" to "Sensitive But Unclassified" between those involved in the incident.
6	Controlled access to classified sources of threat information and updates, including electronic clearinghouses.
7	Access to open-source data for queries.

No.	Operational Needs
8	Access to U.S. and international financial records associated with a suspect within 30 seconds (for counterterrorism).
9	Selectively request and receive data sets of specific interest associated with a threat.
10	Make the data anonymous, whenever possible. However, analysts should be able to perform link analysis, queries, and entity resolution on the data.
11	Scoring mechanisms that order data by relevance to aid in identifying people who warrant closer examination.
12	Credibility ratings of the data source, analyst, and analysis associated with the mined data.
13	Pointers to the data source, so that analysts are able to contact the source for additional information as required.
14	Identify the agency responsible for collecting and analyzing certain types of intelligence.
15	Analyze, validate, and assess threats for elevating threat levels credibly.
16	Perform redundant data mining to compare results of different data mining models.
17	Data verification when different data mining methods produce differing results.
18	Validate data credibility through cross-analysis of disparate information.
19	Sharing of actionable and relevant information between agencies.
20	Capability to share classified information (or a subset of the information) with state and local agencies, as well as the capability for these agencies to contribute information.
21	Access to an "all hazards" view that provides the status of common hazards.
22	Common and readily understood alert system that automatically monitors the status of common hazards and disseminates this information to critical personnel.
23	A system that will pre-position instructions for responders, elected officials, and media for "all hazard" responses. The instructions should be easy to read and linked to common threat levels or, possibly, multimedia.
24	Establish minimum data sharing requirements for each level of operation and type of incident.
25	Customized report formats for data delivery.
26	Share data locally after it has been selected through data mining.
27	Cache often-used data results locally with expiration dates for data sharing purposes.
28	Reduce data in an organized way to produce actionable information at an appropriate level.
29	Base information technology, developed for intelligence fusion, on a geospatial foundation.
30	Access geospatial data specific to a credible location threat.
31	Check for false identities.

No.	Operational Needs
32	Positively identify foreign students by tracking each student's status and location (for counterterrorism agencies).
33	Fuse all intelligence information disciplines, including human intelligence, signals intelligence, and electronic intelligence, into one location (database or clearinghouse).
34	Take information, in any form, perform quality assurance, and publish it on a network for authorized users.
35	Ensure that all data on the network is digital.
36	Allow users to move large amounts of data easily and in any format.
37	Adhere to industry-standard data-exchange practices.
38	Regional clearinghouses for data mining and analysis to support nearby smaller jurisdictions without the manpower for all-source situational understanding capability.
39	Execute queries to identify a person through multiple sources at one time. If associated people are found, perform a search on each associate to determine additional associations via link analysis.
40	Identify known associates of a terrorist suspect within 30 seconds.
41	Rapidly assess a threat and search for other corroborating evidence or activities, including potential relationships to any watch list suspects.
42	Ability for agencies to create actionable information.
43	Integrate data mining tools for data exploitation through clustering, linking, normalizing, and interpreting extracted data, including statistical analysis to support situational understanding.
44	Model or mining operation results available within the time frame of the phenomenon it is predicting.
45	A network that enables participants to distinguish signals of potential terrorist activity from useless noise.
46	Create systems and processes that support "thinking outside the box," rather than forcing analysts to consider limited or uncreative hypotheses, options, etc.
47	Integrate tools for data mining and analysis of both structured and unstructured information.
48	Integrate tools for data mining and analysis geared toward detecting alert situations, including anomaly detection via normalization.
49	Integrate tools to automate mining foreign language translation, including using cultural intelligence to understand the context of conversations and text.
50	Integrate mining tools for processing images, video, audio, and signal data with automated text and image extraction.
51	Capability to examine temporal data, based on the assumption that all data will have a date-time stamp.
52	Intelligent agents (i.e., systems/programs capable of learning) that can perform automated searches. Results should be coded to indicate the importance of the data and show need for further queries.
53	Automated queries that produce visual cues that indicate when a situation or pattern changes.

No.	Operational Needs
54	Threat assessment toolkit that aids in pattern recognition.
55	Detect terrorist-sponsored money laundering activities.
56	User-friendly querying tools for public safety officers.
57	Customizable searches of any data source.
58	Access to a trusted agent for complicated searches.
59	Support on-demand and ad hoc information queries.
60	Strong data protection for the network, including restricted access privileges and time out capability.
61	Read-only audit logs for counterterrorism systems. The audit logs include immediate and permanent records of authorship, editing, access, queries, results, and information sharing.
62	Establish tamper-resistant controls and audit log inspection procedures.
63	Establish safeguards that include pattern reviews of searches to protect civil liberties.
64	Establish a consolidated list of terrorist suspects assembled by different agencies.
65	Update watch lists based on credible information on a real-time basis.
66	Establish accountability standards for sharing watch lists on a real-time basis.
67	Integrate local law enforcement identity checks with the federal suspect watch list.
68	Alert local law enforcement officers automatically when a license or license plate of a terrorist suspect is checked.
69	Automated network queries that continually screen for new patterns related to watch lists.
70	Access data about people in response to a credible methodology threat.

3.3 Derived Software Requirements

The operational needs as defined in Section 3.2 were translated into the software requirements tabulated in the table below. The software requirements listed in this table are based on SME input and are not meant to be an all-inclusive list of data mining and analysis software requirements. The requirements are broken into six areas, data sources, data preparation, accessing and using data, data mining and analysis, rank results, and results visualization. The *Mapping to Operational Needs* column in the table below corresponds with the number in the *Operational Needs* table in Section 3.2.

Note: In order to perform the operational needs, a tool should provide most, if not all, of the basic data mining and analysis operations highlighted in the table below.

Derived Software Requirements	**Mapping to Operational Needs**
Provide access to the following types of data sources:	
Public (open source) and private data sources including those related to the critical infrastructure and financial transactions both domestic and foreign.	1-4, 6-18, 21-23, 25, 27-59, 64-70
File types including: text files, audio format files, spreadsheet files, databases.	
Data that has been converted from paper to digital format, including satellite images and scanned data.	
Support legacy and proprietary data sources.	
Prepare data for sharing, analysis, and mining through the following types of operations:	
Convert actual data values to anonymous values prior to data sharing.	1-4, 6-18, 21-23, 25, 27-59, 64-70
Manage missing values, delete duplicate entries, and handle entry errors and inconsistencies, such as misspellings.	
Database and free text search and refinement capability.	
Perform linguistic analysis of foreign language text.	
Convert audio tracks to text for keyword queries.	
Support the following manipulations: aggregate, append, filter rows, partition, sample, shuffle, sort, split, stack, bin, create column, filter column, join, modify column, transpose, normalize, summarize, and compare.	
Support the following operations: mathematical, logical, comparisons, statistical (comparison of data values).	
Automatically read and filter material and assign the material to specific categories.	
Provide the following capabilities for accessing and using the data:	
Provide a centralized access point for multi-source analysis of disparate sources.	All
Concurrent access to distributed sources.	

Derived Software Requirements	Mapping to Operational Needs
Pointing system that searches and matches data across disparate database systems.	
Store data in a variety of formats, including database, Web page, Web exchange, image, video, and text files.	
Data warehouse the mined information in a single location.	
Provide pointers back to original data source.	
Cache frequently accessed data for rapid retrieval and look up tables.	
Control access to data through users' groups, user authentication, automatic file encryption, security certificates, and secure socket layer.	
Pass-through security payload to allow connectors to directly use the security of underlying systems.	
Tracking log capability that includes data access, querying, and data sharing with limited access to the log file.	
Provide the following types of data query, data analysis, and data mining:	
Single search capability to retrieve information from any data source available.	1-4, 6-18, 21-23, 25, 27-59, 64-70
Compare the data against various data sources both public and private.	
Query capabilities, including: query on key words or phrases, wild cards, and numeric expressions; query within a query; context based querying; graphical querying (both source selection and query building); and complex, ad-hoc querying.	
Locate and track movement of a person by comparing data values such as names, IDs and addresses. If necessary, provide continuous updates when database is updated.	
Automatic updates when database is updated (including mapping information).	
Graphically represent relationships (both direct and indirect) between people and addresses, places, events, associates, bank accounts, financial transactions, and criminal organizations.	
Cluster related data.	
Identify gaps between associations that require further research.	
Anomaly detection in data.	
Use historical patterns based on date and time to predict future incidents.	
Intelligent agents (programs or systems capable of learning) that can perform automated searches.	
Database and free text search and refinement capability.	
Perform linguistic analysis of foreign language text.	
Redundant data mining with different models on the same set of data (including data verification functionality).	
Transform data on the fly to perform advanced analysis on search results.	
Identify unusual or pre-operational activities through data mining of select data sources.	

Derived Software Requirements	Mapping to Operational Needs
Receive partial results even if the query failed.	
Rank Results:	
Web-based real-time scoring of individual observations, select records, or entire database.	1-4, 6-18, 21-23, 25, 27-59, 64-70
Rank credibility of source.	
Provide the following types of results visualization:	
Alert system capable of sending alerts to a variety of devices. If possible, automate alert system.	All
Notification system capable of sending notices to a variety of devices.	
Distribute data in a customizable format to a variety of devices, including fax, e-mail, beeper, voice mail, and PDAs.	
Track locations of deployed resources.	
Automatic updates of GIS maps when the database is modified or updated.	
Mapping capabilities including drawing toolset, toggle-selection for displays, multiple displays, full area map images, point and click enlargement or reduction of area on a map, and printing.	
Integrate plume models to predict where environmental hazards will be dispersed.	
Integrate environmental models for nowcast and forecast of weather conditions.	
Integrate environmental models to GIS models.	
Export results to other tools via XML, HTML, and text files.	
Collaborative visualization environment.	
Visual display of model predictions and observations with automated updates when data changes, and time series of events.	
Provide graph types including density, histogram, bar, dot, pie, scatter, box, strip, contour, level, surface, and scatter plot.	

3.4 Data Sources

Satisfying the HSP's data mining and analysis needs includes having or gaining access to pertinent information. HSPs should consider the following as possible data sources. The table below provides the name of the data source and describes the information it contains.

Data sources are categorized by:

- Secure or non-secure data access
- Structured or unstructured data
- Incident response, prevention, or both.

The *Non-secure and Secure* column shows the data's accessibility. Secure access generally includes private, proprietary, or classified data. Non-secure data is considered open source or public data. A database is considered structured data. Unstructured data can be found in documents, Web pages, or e-mail.

Data Sources

Source	Information Gained	Secure/ Non-Secure	Structured/ Unstructured	Response/ Prevention
Hazardous Substance Release/Health Effects Database (HAZDAT)	Provides access to information on the release of hazardous substances and the effects of hazardous substances on the health of human populations.	Non-Secure	Structured	Response
911 call data	Open source information; public safety answering point.	Non-Secure	Structured	Response
The National Map from the US Geological Survey (USGS)	Alternate transportation routes, geographic data specific to a credible location threat, critical infrastructure nodes in the vicinity of an attack.	Non-Secure	Structured	Response/ Prevention
NOAA	Real-time, accurate weather information.	Non-Secure	Structured	Response
Navy weather	Real-time, accurate weather information.	Non-Secure	Structured	Response
National Weather Service (NWS)	Real-time, accurate weather information.	Non-Secure	Structured	Response
Social Security Administration	Check identities against death records, which might be used to create a false identity.	Non-Secure	Structured	Response/ Prevention
ChoicePoint	Property deeds and ownership.	Non-Secure	Structured	Response/ Prevention
Material Data Safety	Emergency procedures for transportation incidents involving hazardous materials.	Non-Secure	Structured	Response
Graphical Information Services (GIS) vendor data	Critical infrastructure nodes in the vicinity of an attack.	Secure	Structured	Response

Source	Information Gained	Secure/ Non-Secure	Structured/ Unstructured	Response/ Prevention
ChoicePoint (iMapData)	Real-time, accurate weather information and critical infrastructure nodes in the vicinity of an attack.	Secure	Structured	Response
Qsent & ChoicePoint (phone book records)	Check for false identities. An indicator of a false identity is an individual who has no phone book record.	Secure	Structured	Response
AT&T, Bell Companies, SPRINT, Verizon, & Cingular	Identify known associates of a terrorist suspect using shared addresses and records of phone calls to and from the suspect's phone. Predict future incidents based on historical call patterns; identify groups of phone calls that repeatedly occur together.	Secure	Structured	Response/ Prevention
Department of Motor Vehicles (DMV)	Driver's licenses and automobile ownership tracking.	Secure	Structured	Response
Joint Terrorism Task Force (JTTF)	Classified terrorist information (JTTF Intel reports), foreign agency threat information, and security incident information.	Secure	Structured	Response/ Prevention
Lexis-Nexis	Open source data for queries related to news stories and press releases.	Secure	Structured	Response/ Prevention
National Interagency Fire Center	Past and current U.S. wild land fire stats, advisories, and maps.	Non-Secure	Unstructured	Response/ Prevention
Center for Disease Control (CDC) & Division of Public Health Surveillance and Informatics (DPHSI)	Information to enhance health decisions and promote health, disease prevention, and control.	Non-Secure	Unstructured	Response/ Prevention
AOL, MSN, Yahoo, CompuServe, & EarthLink	Identify known associates of a terrorist suspect through emails to and from the suspect's accounts.	Secure	Unstructured	Response/ Prevention
French-McCay, 2001 (chemical model)	Provides integration of environmental data, GIS data, and chemical models to predict atmospheric plumes from chemical releases.	Secure	Structured	Prevention
Spaulding et al., 1992; Howlett et al., 1993, 1996 (oil model)	Provides integration of environmental data, GIS data, and oil models to predict direction and speed of oil spills.	Secure	Structured	Prevention
Homann, 1999 (nuclear fallout model)	Provides integration of environmental data, GIS data, and nuclear fallout models to predict fallout from nuclear explosion or release.	Secure	Structured	Prevention
Financial Crimes Enforcement	Banks supply U.S. and international financial information for law enforcement investigations.	Secure	Structured	Prevention

Source	Information Gained	Secure/ Non-Secure	Structured/ Unstructured	Response/ Prevention
Network (FinCEN)				
Visa, MasterCard, American Express, & First Data Express	Transactional data, point-of-sale data, and credit card records in real-time makes it possible to track the activities of individuals.	Secure	Structured	Prevention
Equifax & Experian (Credit Report Header)	Check for false identities. An indicator of a false identity is no credit-header files.	Secure	Structured	Prevention
Professional Association of Diving Instructors (PADI) & National Association of Underwater Instructors (NAUI) (SCUBA certification agencies)	Data searches should be conducted on an ongoing basis using special licenses.	Secure	Structured	Prevention
ChoicePoint (licenses)	Data searches should be conducted on an ongoing basis using special licenses.	Secure	Structured	Prevention
Terrorist Threat Integration Center (TTIC)	Provides comprehensive, all-source-based picture of potential terrorist threats to U.S. interests.	Secure	Structured	Prevention
Bureau of Citizenship & Immigration Services (BCIS)	Generate watch lists on various suspicious individuals.	Secure	Structured	Prevention
Galileo, Sabre, WorldSpan, & Amadeus (travel information)	Transactional travel records in real-time to track the activities of individuals (itineraries, reservations, rentals).	Secure	Structured	Prevention
Student & Exchange Visitor Information System (SEVIS)	Status and locations of foreign students, including prospective and former students, research assistants, and teachers in programs.	Secure	Structured	Prevention
National Security Entry-Exit Visitor Registration System (NSEERS)	National registry for temporary foreign visitors arriving from certain countries or who meet a combination of intelligence-based criteria, and are identified as presenting an elevated national security concern.	Secure	Structured	Prevention
U.S. Visitor & Immigrant Status Indication Technology (U.S. VISIT)	Biographic information, index finger scan, and photo of anyone applying for visa; tracks U.S. entry and exit, and land border crossings. Match with watch lists.	Secure	Structured	Prevention
Port Import Export Reporting Service	Identify suspicious patterns of shipping containers through cargo and itinerary historical	Secure	Structured	Prevention

Source	Information Gained	Secure/ Non-Secure	Structured/ Unstructured	Response/ Prevention
(PIERS)	data.			
BCIS I-94s	Border entry and exit records.	Secure	Structured	Prevention
Computer Assisted Passenger Prescreening System (CAPPS II)	Profiles passengers and creates passenger watch lists.	Secure	Structured	Prevention

4. Conclusions

Data mining and analysis tools provide the ability to extract knowledge from structured or unstructured data. This knowledge can be used to predict future events, find new associations between events, or organize related data in new ways. The Operational Needs and Software Requirements Analysis document details data mining and analysis needs based on input from SMEs. These needs were then reverse engineered to the associated software requirements, which were validated by SMEs.

The operational needs analysis found that HSPs at the state, regional, or federal level are most likely to use data mining and analysis tools and that public safety officers use the results of these tools. The analysis also indicates that HSPs require tools that: provide access to data sources, prepare data for mining, store the prepared data, perform data mining analysis, rank results, and provide visualization of the results.

The information in this report is the result of the data mining and analysis study. Additional reports stemming from this study include a listing of software tools which meet the software requirements derived through this study, and a comprehensive report that includes a method for assisting the HSP in determining which tool best fits their software requirements based on their operational needs.

5. References

[1] *Multi-State Anti-Terrorism Information Exchange: A Pilot Information Sharing Project*. 16 March 2005. Institute for Intergovernmental Research. 16 March 2005.

[2] Dizard, Wilson P. (2004, June). *Data Sharing Starts on the Web*. *Government Computer News 23* (15). Retrieved September 14, 2004, from

[3] Homeland Security Demin. (2003). Retrieved October 18, 2004, from Texas A&M University, Institute for Public Policy Research Institute Web site.

[4] Markle Foundation Task Force. (2003). *Creating a Trusted Network for Homeland Security*. New York: Markle Foundation.

[5] United States Department of Justice, National Institute of Justice, Department of Homeland Security, & Oklahoma City National Memorial Institute for the Prevention of Terrorism. (2003, March). *Project Responder Interim Report: Emergency Responders' Needs, Goals, and Priorities*. Virginia: Hicks & Associates, Inc.