Data Mining Report

DHS Privacy Office Response to House Report 108-774

July 6, 2006

Homeland Security

Report to Congress on the Impact of Data Mining Technologies on Privacy and Civil Liberties

Respectfully submitted
Maureen Cooney
Acting Chief Privacy Officer
U.S. Department of Homeland Security
Washington, DC

July 6, 2006

TABLE OF CONTENTS

I. EXECUTIVE SUMMARY 1

A. DEFINITION OF DATA MINING 1

B. DATA MINING PROCESS STEPS AND ATTENDANT PRIVACY ISSUES 1

C. DHS DATA MINING ACTIVITIES 2

I. INTRODUCTION 5

II. DESCRIPTION OF DATA MINING TECHNOLOGY 6

A. DEFINITION OF DATA MINING 6

B. THE PROCESS OF DATA MINING 8

1. DEFINITION OF THE PROBLEM TO BE SOLVED 9
2. DATA IDENTIFICATION AND COLLECTION 9
3. DATA QUALITY ASSESSMENT AND DATA CLEANSING 9
4. MODEL BUILDING 10
5. MODEL VALIDATION 10
6. MODEL DEPLOYMENT 10

C. DATA MINING TECHNIQUES 11

III. PRIVACY AND CIVIL LIBERTIES CONCERNS IN THE USE OF DATA MINING TECHNOLOGIES FOR HOMELAND SECURITY 11

A. PURPOSES OF DATA MINING 12

1. INAPPROPRIATE DATA MINING 12
2. FUNCTION OR MISSION “CREEP” 12

B. DATA IDENTIFICATION AND COLLECTION 13

1. INAPPROPRIATE ACCESS TO INFORMATION 13
2. DUPLICATION OF DATA 14
3. DATA RETENTION 15
4. USE OF DATA FOR PURPOSES INCOMPATIBLE WITH PURPOSES OF DATA COLLECTION 15
5. SUBJECT SYNOPSIS 16

C. DATA QUALITY ASSESSMENT AND CLEANSING 16
1. INTRODUCTION OF ERRORS DURING DATA PREPARATION 17

D. MODEL BUILDING AND EVALUATION 17
1. DATA LEAKAGE 17
2. IMPROPER MODEL VALIDATION 17

E. MODEL DEPLOYMENT 18
1. NEW PERSONAL INFORMATION ABOUT INDIVIDUALS 18
2. FALSE POSITIVES 18
3. LACK OF APPROPRIATE REVIEW AND REDRESS 19

F. CONCLUSION 19

IV. DHS DATA MINING ACTIVITIES 19

A. DATA ANALYSIS FOR IMPROVING OPERATIONAL EFFICIENCY U.S. CUSTOMS AND BORDER PROTECTION 20
1. PURPOSES OF THE PROGRAM 20
2. DATA SOURCES 21
3. DEPLOYMENT DATES 21
4. POLICIES, PROCEDURES, AND GUIDANCE 21

B. LAW ENFORCEMENT ANALYTIC DATA SYSTEM (NETLEADS) IMMIGRATION AND CUSTOMS ENFORCEMENT 21
1. PURPOSES OF THE PROGRAM 22
2. DATA SOURCES 22
3. DEPLOYMENT DATES 23
4. POLICIES, PROCEDURES AND GUIDANCE 23

C. ICE PATTERN ANALYSIS AND INFORMATION COLLECTION SYSTEM (ICEPIC) IMMIGRATION AND CUSTOMS ENFORCEMENT 24
1. PURPOSES OF THE PROGRAM 24
2. DATA SOURCES 25
3. DEPLOYMENT DATES 25
4. POLICIES, PROCEDURES, AND GUIDANCE 25

D. INTELLIGENCE AND INFORMATION FUSION (I2F) OFFICE OF INTELLIGENCE AND ANALYSIS 26
1. PURPOSES OF THE PROGRAM 26
2. DATA SOURCES 26
3. DEPLOYMENT DATES 26
4. POLICIES, PROCEDURES AND GUIDANCE 26

E. FRAUD DETECTION AND NATIONAL SECURITY DATA SYSTEM (FDNS-DS) US CITIZENSHIP AND IMMIGRATION SERVICES 27
1. PURPOSES OF THE PROGRAM 27
2. DATA SOURCES 27

3. DEPLOYMENT DATES 27
4. POLICIES, PROCEDURES AND GUIDANCE 27

F. NATIONAL IMMIGRATION INFORMATION SHARING OFFICE (NIISO) US CITIZENSHIP AND IMMIGRATION SERVICES 28
1. PURPOSES OF THE PROGRAM 28
2. DATA SOURCES 28
3. DEPLOYMENT DATES 28
4. POLICIES, PROCEDURES AND GUIDANCE 29

V. CONCLUSIONS AND RECOMMENDATIONS 29

VI. APPENDIX A 31

Data Mining Report
DHS Privacy Office
July 6, 2006

I. Executive Summary

This report is prepared pursuant to the requirements of House Report 108-774 – *Making Appropriations for the Department of Homeland Security for the Fiscal Year ending September 30, 2005, and for Other Purposes*. This report provides information related to the status, issues, and programs related to DHS data mining activities.

A. Definition of Data Mining

There is no agreed-upon definition for the term "data mining." Based on the definitions used by the Congressional Research Service and the Government Accountability Office, data mining is defined in this report as follows:

> ***Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Data mining consists of more than collecting and managing data; it also includes analysis and prediction.***

The application of patterns, relationships, and rules to searches, whether these are derived through data mining, observation, intelligence, or theoretical models, is not addressed in this report.[1]

B. Data Mining Process Steps and Attendant Privacy Issues

Data mining is a process that consists of a series of steps. Privacy and civil liberties issues arise in every step of the data mining process.

The first step in the data mining process is to define the business need that data mining expects to address. As with any activity undertaken by a Federal agency, a data mining project must be performed for a lawful purpose, consistent with the agency's mission. After an agency determines the problem that data mining may be useful in solving, and finds that it has the mission authority to perform the project, it needs to identify and then collect or aggregate the data for analysis. The privacy and civil liberties issues that may arise during this step include inappropriate access to information, duplication of data and the resulting inability of the original data collector to control subsequent uses or maintain quality of the data, inappropriate data retention policies, use of data incompatible with purposes for which it was originally collected, and profiling of individuals.

After data is collected or aggregated, it undergoes a "cleansing" process. Inaccuracy of data is a significant concern in data mining. If data is inaccurate or incomplete, then the

[1] Thus, this report would exclude searches using patterns, relationships, and rules focused on a particular individual, such as used in a threat and risk assessment vetting program.

patterns, relationships, or rules detected in the data may be meaningless or wrong. Worse from a privacy and civil liberties perspective, if patterns, relationships, or rules used for law enforcement or intelligence are determined through mining inaccurate data, such patterns, relationships or rules may implicate innocent individuals. For this reason, data intended for data mining usually undergoes a "cleansing" or validation process prior to the start of analysis. However, the data cleansing process can itself introduce inaccuracies into the data.

After data is cleansed and validated, the model building process begins. This is the step during which patterns in the data are detected and validated and rules for predicting future events or behaviors are created. Potential privacy and civil liberties issues during this step of the process include security risks, such as access to data by unauthorized persons, as well as inappropriate disclosures by authorized users. Additional concerns arise if the model is inappropriately validated before deployment.

The final step in data mining involves the deployment of the model to the field. It is at this step of the data mining process that concerns arise about false positives and appropriate due process for individuals who are flagged by the model. There are also questions about ownership and uses of new information about individuals produced through the use of data mining models.

C. Recommendations for DHS Data Mining Activities

Several components of DHS engage or plan to engage in data mining activities, as defined by this report. Based on our analysis of DHS activities that involve current and projected future uses of data mining, we note that data mining is usually only one part of a larger set of analytic activities and tools. Such analytic activities include searches and traditional analyses.

Although DHS programs that employ data mining tools and technologies also employ traditional privacy and security protections, such as Privacy Impact Assessments, Memoranda of Understanding between agencies that own source data systems, privacy and security training, and role-based access, we recommend additional protections that are aimed specifically at addressing the privacy concerns raised by data mining and we will take steps to implement these recommendations within the Department.

1. Prior to the start of any data mining activity, the authority of the agency to undertake such activity should be determined to be consistent with the purposes of the data mining project or program. The authority to collect or aggregate data required to perform the data mining project should also be ascertained, whether the project involves collection of new data or aggregation of existing data from various sources. While oversight functions exist in different components within

DHS,[2] the Department as a whole could benefit from more centralized oversight with a broader view of DHS activities and data holdings. One such body, which could assist the function of overseeing DHS data mining programs, is the DHS Privacy and Data Integrity Board, an internal privacy board that is charged under the Privacy Act to examine and approve data matching agreements between DHS and other departments and that considers Departmental privacy issues. The Board, which includes representatives from all DHS components, the Office of the Inspector General, the Chief Information Officer, and the Office of General Counsel, and is chaired by the DHS Chief Privacy Officer, could provide oversight and confirmation to ensure responsible application of data mining tools and technologies. The Board assisted with the collection of data for this report concerning current and planned data mining programs within DHS.

2. As discussed in the report, data mining searches for patterns, relationships, and rules in the data without basing this search on observations or a theoretical model. Because the existence of patterns in the data may not reflect cause and effect, data mining tools should be used principally for investigative purposes. DHS components that use data mining tools should have written policies, stating that no decisions may be made automatically regarding individual rights or benefits solely on the basis of the results produced by patterns or rules derived from data mining.

3. Because the patterns, relationships, and rules in the data may not be derived from specific personal identifiers, such as a name or Social Security Number, when a data set includes personally identifiable information, data mining projects should give explicit consideration to using anonymized data in data mining activities. A discussion of the extent to which anonymization was considered should be included in the Privacy Impact Assessment for such a data mining project.

4. Data quality plays an important role in the ability of data mining techniques to produce accurate results. DHS should adopt data quality standards for data used in data mining. Application of these standards, which should affect systems using both data from government and commercial sources, should be ensured prior to the deployment of data mining models or predictive rules for use in the field.

5. In order to ensure that data mining models produce useful and accurate results, DHS should adopt standards for the validation of models or rules derived from data mining. Evaluation of the model validation process and the ability to meet these standards should be reviewed and documented prior to the deployment of the model to the field.

[2] Including the DHS Office of Civil Rights and Civil Liberties.

6. Each DHS component that employs data mining should implement policies and procedures that provide an appropriate level of review and redress for individuals identified for additional investigation by patterns, relationships, and rules derived from data mining. Because data mining algorithms most times produce highly complex patterns, relationships, and rules that may not be fully understandable as to the particular reasons for identifying the individuals, a complete procedure should include a step that a person, acting independently from the data mining process, substantiates the particular identification of individuals prior to any determinative processes and procedures. To ensure a complete understanding of the capabilities and limitations of data mining in this regard, employees who use data mining processes should be required to complete training on these polices and procedures.

7. In order to provide demonstrable accountability, each component that employs data mining should include strong, automatic audit capabilities to record access to source data systems, data marts, and data mining patterns and rules. Programs should conduct random audits at regular intervals, and all employees should be given notice that their activities are subject to such audits. These actions help underline the importance of transparency.

Data Mining Report
DHS Privacy Office
July 6, 2006

I. Introduction

This report is prepared pursuant to the requirements of House Report 108-774 – *Making Appropriations for the Department of Homeland Security for the Fiscal Year ending September 30, 2005, and for Other Purposes*. The report includes the following requirements:

> *The conferees direct the DHS Privacy Officer, in consultation with the head of each Department of Homeland Security agency that is developing or using data-mining technology, to submit a report no later than 90 days after the end of fiscal year 2005 that provides (1) a thorough description of the data-mining technology, the plans for use of such technology, the data that will be used, and the target dates for the deployment of the technology; (2) an assessment of the likely impact of the implementation of the technology on privacy and civil liberties; and (3) a thorough discussion of the policies, procedures, and guidelines that are to be developed and applied in the use of such technology for data-mining in order to protect the privacy and due process rights of individuals and to ensure that only accurate information is collected and used.*

The Department of Homeland Security ("DHS") Privacy Office is the first statutorily required comprehensive privacy office in any U.S. federal agency. It operates under the direction of the Chief Privacy Officer, who is appointed by and reports directly to the Secretary. The DHS Privacy Office serves as a steward of Section 222 of the Homeland Security Act of 2002, and has programmatic responsibilities involving the Privacy Act of 1974, the Freedom of Information Act ("FOIA"), the privacy provisions of the E-Government Act of 2002, and DHS policies that protect the collection, use, and disclosure of personal information. Additionally, the Privacy Office develops privacy policy and oversees certain information disclosure issues. The Office is also statutorily required to evaluate all new technologies used by the Department for their impact on personal privacy.

The Privacy Office wishes to acknowledge the generous assistance it received from U.S. Customs and Border Protection ("CBP"), Immigration and Customs Enforcement ("ICE"), U.S. Citizenship and Immigration Services ("USCIS"), and the Office of Intelligence and Analysis in writing this report. We further wish to acknowledge consultation with other offices within the Department, including Civil Rights and Civil Liberties, the Science and Technology Directorate, and the Policy Office.

The report contains the following sections. Section II describes data mining technologies and how these technologies can be used in homeland security applications. Section III addresses privacy and civil liberties concerns that have been raised with regard to data mining technologies. Section IV discusses current and anticipated DHS data mining activities, including the policies, procedures, and guidelines designed to protect privacy, civil liberties and due process rights when data mining technologies are used. The final section presents the conclusions of the report.

II. Description of Data Mining Technology

This section of the report defines data mining and examines the process and technologies for conducting data mining.

A. Definition of Data Mining

There is no universally agreed-upon definition for the term "data mining." Some definitions of the term are quite broad. For example, the Technology and Privacy Advisory Committee ("TAPAC") of the Department of Defense defined data mining as:

> [S]earches of one or more electronic databases of information concerning U.S. persons, by or on behalf of an agency or employee of the government.[3]

This presents too broad a definition of data mining. While searches, particularly pattern-based searches and searches of multiple databases do raise privacy concerns, data-retrieval via computerized search is, in many cases, a faster and more efficient way to perform an activity that could be performed manually. Additionally, the TAPAC definition covers activities requested by the individual who is a subject of the information, such as searches of a single database in response to a customer service query or a request under FOIA, and it is our conclusion that such simple data retrievals should not be included in the context of a discussion about data mining.

Authors of other reports use narrower definitions. For example, the Congressional Research Service ("CRS") defines data mining as follows:

> Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.[4]

The Government Accountability Office ("GAO") defines data mining similarly, as

[3] Technology and Privacy Advisory Committee, "Safeguarding Privacy In the Fight Against Terrorism," March 2004, p. viii.

[4] J.W. Seifert, *Data Mining: An Overview*, Congressional Research Service, RL31798, June 2005, p. 1.

> [T]he 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.[5]

There are two important components in the definitions used by CRS and GAO. The first is the discovery of hidden patterns in the data and the second is the use of these patterns to predict future results. Only the first of these, the search of databases for hidden, valid patterns, relationships, and rules, is unique to data mining. As such, data mining would not include searches for connections, direct or indirect, between data points focused on a known subject.

Looking for rules[6] that allow prediction of future behavior or results is an important part of many branches of data analysis. For example, probability theory, a branch of mathematics that has been studied since the seventeenth century, is a study of ways to predict future events from past occurrences. The significant difference between data mining and other analytic techniques is in the way the prediction rules are determined. Generally, analytic techniques test hypotheses generated through observation or theory.[7] In data mining, the analysis of the data itself is expected to produce patterns, relationships, and rules that are not known and that are not based on observation or a theoretical model, but are nevertheless valid.

It is important to note that because data mining is not based on a theoretical underpinning, it can only identify patterns in the data; it cannot reveal whether any discovered pattern is meaningful or significant.[8] Only someone who understands the business problem under analysis can determine the significance of a discovered pattern. Most importantly, from a privacy and civil liberties point of view, the patterns, relationships, or rules produced through data mining do not reveal specifically the reason that such a pattern, relationship, or rule exists. That makes it essential that someone familiar with the reason for the analysis reviews and confirms the results.

[5] United States Government Accountability Office, *Data Mining: Agencies Have Taken Key Steps To Protect Privacy in Selected Areas, but Significant Compliance Issues Remain*, GAO-05-866, August 2005, p. 4.

[6] In this report, “rules” specify a set of actions that are expected to follow a particular set of conditions. An example of a rule might be, “If an individual sponsors more than one fiancée for immigration at the same time, there is likelihood of immigration fraud.”

[7] This type of analysis is generally described as the scientific method. See, for example, “Steps of the Scientific Method” at <http://www.cdc.gov/ncbddd/folicacid/excite/Files_in_use/steps_of_the_scientific_method.htm>, last visited December 27, 2005.

[8] Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery*, Third Edition, 1999, p. 1.

Because analysis designed to predict future behavior or results is not unique to data mining, this report focuses on the feature of data mining that is unique—discovering new patterns, relationships, and rules in data. Therefore, in this report data mining is defined as follows:

> Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Data mining consists of more than collecting and managing data; it also includes analysis and prediction.[9]

This means data mining consists of the collection and management of data associated with analysis and prediction of future outcomes. The application of patterns, relationships, and rules to searches, whether these are derived through data mining, observation, intelligence, or theoretical models, is not addressed in this report.

The term "data mining" is often used to describe analysis of numerical or structured data. The term "text mining" is often used to describe analysis of unstructured text. Following the definition of the CRS, the definition in this report includes the analysis of data in all forms: quantitative, textual and digitized images.

B. The Process of Data Mining

Data mining is an analytic process that involves a series of steps.[10]

- Definition of the problem to be solved
- Data identification and collection
- Data quality assessment and data cleansing
- Model building
- Model validation
- Model deployment

The data mining process is iterative, with information learned in later steps leading the analyst back to earlier steps for clarification and adjustment.

[9] Thus, this report would exclude searches using patterns, relationships, and rules focused on a particular individual, such as used in a threat and risk assessment vetting program.

[10] The Appendix contains examples of data modeling processes used in the U.S. and Europe.

1. Definition of the Problem to be Solved

It is essential that data mining begin with the understanding of a business need which will be served by the data mining analysis. Without this understanding, it is not possible to determine what data is needed or whether the patterns detected in the data are useful or meaningful.

2. Data Identification and Collection

Once the business need is understood, the data can be identified, collected and prepared for analysis. These activities can take the majority of the time and effort in the data mining process.[11] The data to be analyzed is generally copied into a separate data base, usually called a data warehouse or data mart, although techniques for distributed data mining[12] and the use of "virtual" data warehouses are being developed. The use of separate data warehouses or data marts can help prevent accidental changes in source data, and allows analysts to work with the data without reducing performance of other applications being run on source databases.

3. Data Quality Assessment and Data Cleansing

Data quality assessment and data cleansing are essential for preparing data for analysis. Aggregating data from different sources into a single database brings with it several potential concerns.

- Individual data fields may have incorrect values. Some of these may be obvious, for example "Age = 200", but others may not be.

- There may be incorrect combinations of data values, such as associating data with an incorrect individual's name. There may also be logically impossible combination of values, such as "City = New York," "State = New York," "Population = 2,500."

- There may be missing data values.

- Different databases may use the same term to describe data values that have different meaning. For example, in one database the field labeled "Address" may refer to home address, but in another database it may refer to shipping address.

[11] Two Crows Corporation, p. 23.

[12] "Distributed data mining" is a technique for doing data mining on databases that reside on different computers or in different organizations. Data mining techniques are now being developed that permit analysis of these databases without first combining data into one large database.

This problem can be particularly severe when data collected for one purpose is used for other purposes.

The issues listed above may be present in a single database, but can be exacerbated when data from different sources are combined for analysis. The data cleansing process is the process of looking for and, when possible, correcting potential errors in the data.

4. Model Building

Often, when people talk about data mining, they mean the step of building the models that correspond to the underlying information in the data. The goal of the data modeling process is to discover valid relationships between data elements. These relationships, sometimes referred to as patterns or rules, can then be used to predict future behavior or search for additional cases where the relationship between variables holds. For example, data mining may indicate that fraudulent applications for benefits have particular characteristics, which would lead to an investigation of future benefits applications with similar characteristics. Techniques used in building data mining models are discussed in the next section of this report.

5. Model Validation

Finally, before a model is deployed, it must be evaluated and validated. As mentioned above, just because a pattern exists in the data does not mean that the pattern is meaningful or valid. Some correlations between data elements can be spurious, such as when two people attend the same university at the same time. Correlations by themselves do not provide any information about cause and effect. For example, data mining may demonstrate a correlation between high average family income and high quality of education in local public schools, but it will not explain whether families with high incomes move to areas with good public schools or public school quality improves because families with high incomes have more resources to devote to education.

To validate patterns discovered through data mining, model builders often divide the data into separate data sets—one set to build the model and the other set to validate the model's predictive ability. If a model cannot make predictions with a pre-specified degree of accuracy, it is generally rejected.

6. Model Deployment

The final step in data mining is the deployment of the model to the field. Models can be used to make recommendations or to analyze new data. In cases where the model becomes part of a set of analytic tools, new users must be trained on appropriate uses and limitations of the model and on the process that must be followed with the results produced by the model. Performance of the model must also be monitored over time as the changes in the external environment affect the patterns of behavior that the model was built to analyze.

C. Data Mining Techniques

Data mining techniques look for various types of patterns, relationships, and rules in the data:

- Association or link analysis (i.e., pattern in which events and/or people are associated with one another)[13]

- Sequence or path analysis (i.e., patterns where one event leads to another event)

- Classification (i.e., looking for events, objects, or people with shared characteristics)

- Clustering (i.e., finding and documenting groups of people or entities whose attributes are similar to each other but different from those in other clusters)

- Forecasting (i.e., discovering patterns from which one can make reasonable predictions regarding future activities or events)

Visualization techniques, while not analytic techniques in themselves, are often used to assist analysts by displaying analytic results in easily comprehensible form. For example, link analysis can be presented as a group of objects connected by lines. By looking at the visual representation of analytic results, an analyst can focus on a particular object, examine the underlying data, or look at ways in which connections between objects evolved over time.

III. Privacy and Civil Liberties Concerns in the Use of Data Mining Technologies for Homeland Security

Data mining provides a set of analytic tools. In conjunction with other tools, data mining can provide the capability to explore and fully exploit enormous quantities of available transaction, operational, and other data. As is true of all tools, data mining can be used appropriately to enhance security, reduce fraud and increase operational efficiency. Data mining can be used as a tool to provide insight and access into information not otherwise available through other means. In particular, if the pattern, relationships, and rules discovered validate other means of making determinations, especially in subject identification, data mining can enhance security.[14]

[13] Again, this does not includes searches predicated upon a known subject.

[14] Deployed appropriately, data mining can provide an effectual means to reduce not only false positives, but also false negatives. In a security setting, any tool and its capabilities should be viewed using a risk assessment model in order to recognize essential protections based upon the risks associated. Nonetheless,

When data mining is used to analyze information about individuals so that decisions can be made about these individuals, there are greater risks to privacy and civil liberties and the data mining activity requires greater controls and oversight.

Privacy and civil liberties issues potentially arise in every phase of the data mining process. Some of these concerns are not unique to data mining. For example, data aggregation during the Data Identification and Collection phase of data mining raises privacy and civil liberties concerns just as it would if data was aggregated for other purposes; however, because data mining involves a search for relationships that is not based either on observation or on a theoretical prototype, either of which can be directly challenged, privacy and civil liberties concerns are more pronounced.

A. Purposes of Data Mining

As noted above, the first step in the data mining process is to define the business need that data mining expects to address. Nonetheless, as with any activity undertaken by a Federal agency, a data mining project must be performed for a lawful purpose, consistent with the agency's mission and improper uses and implementations avoided.

1. Inappropriate Data Mining

If a data mining project conducted by a government agency does not fall within the authority of the agency, this project is inappropriate. If multiple agencies participate in a project, the project may be inappropriate if all agencies participating do not have proper authority or if data sharing agreements, both for underlying data sources and for data mining results, are not in place.

2. Function or Mission "Creep"

When Federal agencies undertake the creation of a system of records that houses personal data and from which individual information can be retrieved by some personal identifier or initiate a project that uses such data, they are mandated to provide notice to the public about their activities by the Privacy Act of 1974.[15] Nonetheless, there are instances when agencies want to use data for a purpose that is different from the purpose for which the data was originally collected. In some cases, these additional purposes may be

strong security controls and procedures along with a robust audit capability are essential to protect against unauthorized access or misuse of data.

[15] The information contained within the system must be retrieved by an individual identifier. 5 U.S.C. §552a(a)(5). The noted requirement comes from the duties created under the Privacy Act of 1974, 5 U.S.C. §552a, which requires agencies to "inform each individual whom it asks to supply information … [of] the authority … which authorizes the solicitation of the information …[,] the principal purpose or purposes for which the information is intended to be used[,]the routine uses which may be made of the information … and[,]the effects on him, if any, of not providing all or any part of the requested information." 5 U.S.C. § 552a(e)(3).

appropriate, if provided for as an exception under the Privacy Act or defined in a routine use in the associated system of records notice. Much of the time, the use purpose is somewhat related to the collection purpose, such as terrorist screening and access adjudication; however, it is possible that the use purpose may not be related to the collection purpose and such a situation must be examined carefully to ensure appropriate privacy protections.

As noted in the TAPAC report, "[e]ven data accessed or used under an explicit guarantee that they are intended only for one purpose are likely to be used for others later."[16] This is often referred to as "function creep" or "mission creep." Once the data is collected or aggregated, there is great temptation to use it for more and more purposes, if for no other reason than to spread the cost of data collection and storage over more programs and projects.

B. Data Identification and Collection

After an agency determines the problem that data mining may be useful in solving and finds that it has the authority to perform the project, it needs to identify and then collect or aggregate the data for analysis. Several privacy and civil liberties issues arise when data is collected or aggregated for any purpose, including data mining.

1. Inappropriate Access to Information

When an agency collects or aggregates data from various sources, it may gain access to data which it does not have direct authority to collect. As described in the TAPAC report,

> Data aggregation creates the risk that the resulting profile provides the government with substitutes for information it is otherwise not allowed to access or act upon. Similarly, the ability to aggregate records held by third parties may provide the government with precisely the same information it previously would have been required to obtain a warrant to access.[17]

This is a risk that must be seriously considered, whether the government is aggregating data from government sources, commercial sources, or a combination of government and commercial sources.

This risk is not diminished by the use of distributed data mining techniques, which do not require aggregation of source data into a single database,[18] because these techniques still

[16] TAPAC Report, p. 39.

[17] TAPAC Report, p. 36.

[18] See footnote 12 for a definition of distributed data mining.

produce patterns, relationships, and rules based on data that the government may not have the authority to access in certain situations.[19]

This issue can be mitigated through the use of anonymized data. Search for patterns, relationships, and rules may not require individual identifiers such as names and Social Security Numbers. Performing data mining on data that has been stripped of individual identifiers, as appropriate, would reduce the chances of creating aggregate, identifiable records outside appropriate access procedures.

Additional mitigation may be provided through the development and implementation of business process rules that control access to the information, possibly based on roles or responsibilities. In this way, appropriate information may be accessed fully providing filters or barriers that manage the right to use information even in an aggregated environment.

2. Duplication of Data

When data is moved into a data warehouse or data mart for data mining, it is copied, creating a duplicate of data that remains in the original database. As a result, multiple copies of the data come into existence. These copies are not necessarily linked with original source systems. Proliferation of data without linkages to sources can make data correction more difficult and error propagation more likely. It also can make it extremely difficult for the original collector or custodian of the data to control what happens to the data subsequent to its transfer to the data warehouse or data mart. As a result, the original collector or custodian cannot ensure that subsequent data uses remain consistent with the purposes of data collection and notices provided to the public at the time of the original data collection. Additionally, the original collector or custodian cannot ensure that any duplicative data sets incorporate any future changes and corrections in the data.

This issue can be mitigated to some extent through appropriate policies, procedures, and technical safeguards which require, as part of data preparation and cleansing, that data continue to be linked to its sources. Duplicate data in a data warehouse or data mart can be linked via metadata[20] to the original source to permit updating as necessary. If procedures require regular refreshing of data in a data warehouse or data mart, such periodic updates should capture any changes to the data in the original source database.

[19] Note that this statement does not discourage distributed data mining, but rather highlights that the obligations on the government entity remain.

[20] "Metadata" is data about the data in a database. Metadata can describe the meaning of data element definitions, as well as where, how and by whom the data was collected and processed. See, for example, <http://www.csc.noaa.gov/metadata/>, last visited on January 9, 2006.

3. Data Retention

Data mining, with its goal of finding patterns in data, provides a significant incentive for data retention by government agencies and private sector organizations. As discussed above, in data mining, no observation or theoretical model provides the basis for pattern detection. Of course, no patterns can be detected if data is not available. Therefore, an organization may wish to collect great amounts of data and keep it for long periods on the possibility that the greater amount of data might demonstrate or present a pattern or link in a data mining analysis that might not have been otherwise observed.

However, data retention poses risks to privacy and civil liberties. As noted in the TAPAC report, long data retention periods can reduce the integrity of the data because data can become outdated if it is not refreshed to reflect changes. Additionally, as noted by TAPAC,

> [i]f data are retained by industry for government's use or by government itself, this raises concerns about the inability to move beyond one's own past or to overcome the effects of erroneous data. … The passage of time can heighten the privacy interest in information, especially when that information has been aggregated from diverse sources.[21]

4. Use of Data For Purposes Incompatible With Purposes of Data Collection

When data from multiple systems is analyzed through data mining, it is often being used for purposes other than ones for which the data was initially collected. Such use raises the question of whether appropriate notice has been given when data was collected, and whether data collected for one purpose is of appropriate quality for a different use.

If the data was originally collected by the government and if it resides in a System of Records under the Privacy Act of 1974, [22] subsequent uses of the data must be consistent with the System of Records Notice associated with the System of Records or fall within one of the exceptions set forth by the Privacy Act. When data from multiple sources and collected for different purposes is combined into a single data warehouse, some of the data may not be covered by notices that include the contemplated data mining projects.

In addition to the issue of appropriate notice, analysis of data for purposes other than ones for which it was originally collected introduces concerns about data quality. Data

[21] TAPAC Report, p. 41.

[22] Government systems may or may not be Systems of Records as defined under the Privacy Act of 1974. If information is derived from Systems of Records, the disclosure of the data and its subsequent uses must be consistent with the published System of Records Notice for such system.

collected for one purpose may not be complete or accurate enough to answer a different, unrelated question. Moreover, depending on the entity collecting data and on stated purpose at the time of data collection, individuals may not have provided truthful information, particularly if they perceive provision of incorrect data as privacy-protective behavior or if they believe consequences of providing incorrect data are not serious. Data provided by consumers to commercial entities often falls into this category. Additionally, data content is often context-dependent, so different answers may be truthful, depending on the context in which data is collected. For example, when providing an address to an online merchant, an individual may provide a business address when asked to provide an address. For the purposes of the online transaction, the information in the “address” field is truthful and accurate, but this would not be the case for an application that required the individual’s home address.

Without a thorough understanding of the meaning and quality of the data being used, data mining may produce invalid or spurious patterns, relationships, or rules.

5. Subject Synopsis

Data aggregation that results from putting all data into a single data warehouse or by accessing data in separate databases via distributed data mining techniques pose the risk of profiling individuals. When information that was collected and stored separately becomes combined, the created data set may provide a more complete picture of individuals’ activities and associations. Because separate data elements were collected for various purposes, and have different degrees of accuracy and verifiability, the created profiles may contain outdated, inaccurate or improperly attributed information. If an agency uses the discovered profiles, individuals may be inappropriately targeted for investigation or may be inappropriately denied rights or benefits.

In certain instances, the data used may be accurate and subject synopsis through its use would be appropriate by governmental officials. Of course, the use of such data must not violate existing policies against improper profiling based on particular categories.

C. Data Quality Assessment and Cleansing

Inaccuracy of data is a significant concern in data mining. If data is inaccurate or incomplete, then the patterns, relationships, or rules detected in the data may be meaningless or wrong. Worse from a privacy and civil liberties perspective, if patterns, relationships, or rules used for law enforcement or intelligence are determined through mining inaccurate data, such patterns, relationships, or rules may implicate innocent individuals. For this reason, data intended for data mining usually undergoes a “cleansing” or validation process prior to the start of analysis. Nonetheless, while the validation process is a necessary and proper step in the entire data mining process, it is important to recognize that data can never be, nor expected to be, completely accurate prior to use.

1. Introduction of Errors During Data Preparation

Like all forms of data processing, the data cleansing process creates the possibility that inaccuracies will be introduced into the data. This can happen if during a data validation processes data inappropriately linked to individuals, if data categories are misunderstood or mislabeled, or if missing or obviously incorrect data values are inappropriately calculated or filled in from other sources.

D. Model Building and Evaluation

The model building and evaluation step is what most people think of as "data mining." This is the step during which patterns in the data are detected and validated and at which time rules for predicting future events or behaviors can be created.

1. Data Leakage

The act of data processing, for data mining or other purposes, raises potential privacy concerns. Some of these concerns are associated with security risks, such as misappropriation of data by unauthorized persons. However, even uses by authorized persons pose potential concerns.

Whenever data is processed or copied from one system to another, a chance of data leakage exists. Data leakage is described by the TAPAC report as "[d]isclosure … by an authorized user who determines there is some public value in disclosure, by an authorized user who simply fails to protect the data's confidentiality, by an authorized user who engages in unauthorized access … or through a security breach."[23]

Additionally, data may be misused, either with malicious intent or because of curiosity.

2. Improper Model Validation

A model is only useful if it produces valid patterns, relationships, and rules. In order to determine a model's validity, the process of building a data mining model generally involves multiple data sets. Often, a large data set is split into two parts, and patterns detected in one part of the data set (sometimes called the training data set) are validated by applying them to the second part (sometimes called the validation data set). A model can never be made to fit the data in the training data set with 100 percent accuracy because at that level it will fit the old data perfectly and will lose its ability to identify patterns in new data. Nevertheless, if the fit is not sufficiently good, as shown by various statistical measures, the model will produce false positives when applied to new data. Finding the appropriate level of fit is an essential step in ensuring that the patterns,

[23] TAPAC, p. 40.

relationships, and rules within the data mining model do not produce an unacceptable number of false positives when deployed to the field.

E. Model Deployment

When an agency deploys a model that results from data mining to the field, the agency generally uses the model in the same way it uses a model from any other source when doing pattern-based searches. However, because a model derived from data mining is not based on observation or theoretical underpinnings, its deployment poses some special concerns.

1. New Personal Information About Individuals

Data mining can create new personal information about individuals.[24] This can be a result of data aggregation or of scoring or evaluation performed by the data mining algorithm. Data mining may also reveal previously unknown or hidden information about individuals, such as associations with others. As such, the revealed associations may have positive consequences, as in revealing the relationship amongst certain terrorists, or it may have adverse consequences, potentially resulting in a chilling effect on activities protected under the First Amendment. If the data mining algorithm is based on data from multiple sources, it is not clear what rights individuals have with respect to this information, particularly because data mining programs generally do not have direct contact with the individuals who are the data subjects.

2. False Positives

False positives are a concern in all data-based systems. Generally, a false positive results when a system cannot distinguish between an innocent individual and a suspicious one because of insufficient data or because of an insufficient understanding about behavioral or transactional patterns.

In the context of data mining for law enforcement or intelligence, false positives result from patterns, relationships, and rules that incorrectly identify and implicate innocent individuals. If a model is not appropriately validated during data mining (i.e., if inappropriate statistical measures are used to measure the way a model fits the data), the model may produce false positives when deployed with new data.

Patterns and rules that result from data mining can produce lists of individuals who are subject to investigation or who are denied rights or benefits. If models are not appropriately validated before being deployed and not periodically re-validated after deployment, individuals may be inappropriately placed or remain on such lists.

[24] D. Loukidelis, “National Security Claims & Transparency Respecting Privacy Practices,” November 3, 2005, p. 6, available at < http://www.cacr.math.uwaterloo.ca/conferences/2005/psw/loukidelis.pdf>.

This qualification does not imply that in order for an agency to deploy data mining that the developed model must be absolutely accurate. In fact, depending on the ability of the model to correctly identify suspicious individuals, reducing the false negatives, and the consequences of failing to deploy the model, increasing the security risk, even though a model may produce false positives, the need to use the model may outweigh the choice not to use it. This aspect of any data tool must be acknowledged, because unqualified modeling accuracy is neither necessary nor to be expected.

3. Lack of Appropriate Review and Redress

Lack of appropriate review and redress procedures for individuals "flagged" through pattern-based, relationship-based, or rules-based data searches pose a special concern when these patterns, relationships, or rules are derived from data mining. Patterns, relationships, and rules produced by data mining algorithms are often highly complex, with many branches and many data elements. Even those who use such patterns may not fully understand how or why particular individuals have been identified by the search algorithm. Because individuals may not have access to data on which data mining patterns are based and because patterns, relationships, and rules derived from data mining may not be accessible or understandable by the users of models derived from data mining, deployment of searches based on such models must be accompanied by appropriate review and redress policies and procedures to evaluate individuals' claims of inappropriate treatment and to provide redress for loss of rights and benefits.

F. Conclusion

Addressing privacy and civil liberties concerns in the use of data mining technologies makes valuable tools available to secure the nation by ensuring that these technologies protect and do not diminish privacy or civil liberties. A well-designed process can take advantage of the analytical tools that organize vast amounts of transactional, operational, and collected data and increase the value of that data as long as appropriate precautions mitigate potential intrusions into privacy and civil liberties.

As noted in the discussion below, adherence to the requirements of the Privacy Act, fulfillment of the Privacy Impact Assessment obligation, and respectful privacy and civil liberties policies and procedures permit the appropriate application of data mining technologies to support the Department of Homeland Security's mission. Done properly, security and privacy can be achieved together.

IV. DHS Data Mining Activities

DHS engages in some activities that meet this report's definition of data mining. The Department is also planning several new activities that involve data mining. In all of these cases, data mining activities and tools are integrated with other analytic activities.

The listing and descriptions below provide an overview of technologies, activities, and uses of data mining within the Department as defined by this report as well as the safeguards employed by the Department.[25]

A. Data Analysis for Improving Operational Efficiency U.S. Customs and Border Protection

Customs and Border Protection ("CBP") has created an Enterprise Data Warehouse ("EDW") that collects data from CBP transactional systems and then subdivides it into data marts for analysis. CBP operating units use the data marts to analyze and improve performance of their operations.

There are six data marts currently in operation. Most of the activities performed with these data marts involve the production of statistical reports and do not fall under this report's definition of data mining. However, data marts are also used to identify and monitor trends and patterns, such as changes in types of seized items, and changes in numbers and characteristics of cases handled by the component and, as such, may include personally identifiable information (PII).

1. Purposes of the Program

Operating units employ the EDW data marts to generate statistical reports and other analytical products that CBP uses to optimize operations. For example, a data mart operated by the Seizures and Penalties unit allows users to look for various patterns associated with different types of seizures (e.g., drugs, cash, etc.) and to determine how CBP should deploy resources based upon changes in distribution or type of seizures.

Analyses generated via data marts identify patterns, cases, or activities that require further investigation. Users of the data marts also have access to transactional systems that serve as sources of the data and to additional information for investigations is obtained directly from these source systems.

[25] Note that this report focuses only on those DHS programs that implement data mining as defined by this report. Thus, this report does not include programs or systems that are tools or technologies, such as the ADVISE tool being developed by the Science and Technology directorate of DHS, because such tools or technologies do not perform data mining, rather a specific implementation of the tools or technologies may incorporate data mining as defined by this report. In addition, this report does not include programs that search or match data, such as the Secure Flight program being developed by the Transportation Security Administration, because such searches or matches are done with a known name or subject, which does not perform data mining as defined by this report.

2. Data Sources

EDW does not house any data collected specifically for inclusion in EDW or data marts. It uses only government data from CBP transactional systems, such as the Treasury Enforcement Communication System ("TECS").

3. Deployment Dates

EDW has been in operation since 2000. It uses commercial off-the-shelf (COTS) analytic software from Informatica and Cognos, and takes advantage of regular software upgrades.

4. Policies, Procedures, and Guidance

Data mining performed with the data in the EDW is subject to the same policies and procedures as other programs at CBP.

For example, when data is pulled into EDW from a transactional system, the requirements of the Privacy Act that applied to data in the original source system also apply to the data once in the EDW and associated data marts. The same is true for the requirement to protect trade secrets in data received from commercial entities such as airlines. CBP employs various techniques to ensure data protection and appropriate operation.

All users with access to the source databases that provide data to the EDW must undergo privacy and security training before being granted access.

All data marts have role-based access controls. Users are given access only to the level of information relevant to their work. Depending on the job and function of the employee, access may be granted to all data, Field Office level data, or service port level data.

All systems associated with the EDW and data marts are certified and accredited in accordance with FISMA requirements.

In order to ensure that data in the EDW is correct, source systems are designated as the authoritative data source. The data in the data warehouse and data marts is refreshed every 24 hours, so any changes or corrections that have taken place in the source system propagate to the data marts and are reflected in the analyses performed with the data. EDW and data marts are read-only systems. Any changes in the data must be made to transactional systems.

B. Law Enforcement Analytic Data System (NETLEADS)
Immigration and Customs Enforcement

The NETLEADS project is designed to facilitate Immigration and Customs Enforcement ("ICE") law enforcement activities and intelligence analysis capabilities by increasing efficiency of multiple data source searches and patterns and trends analyses. This project was established under legacy INS prior to the advent of DHS, facilitating the data sharing of law enforcement sensitive data between the Immigration Inspectors, Criminal Investigators and Border Patrol. These tools are now in use within DHS, ICE, and CBP.

The NETLEADS project is a tool suite designed to provide a means of performing more efficient searches on a combination of structured data, such as Oracle, Microsoft and mainframe databases, and unstructured data, such as textual reports, open source documentation, Web pages, Reports of Investigation narratives from ICE databases, and images such as PDF files. There are two data mining and visualization features in the program—link analysis, which permits analysis of connections between entities, such as individuals and organizations, and trend analysis across cases. In link analysis, a visual display can be created to demonstrate links between entities; the diagram can be stored, and then compared to similar diagrams generated earlier or later with different versions of the data through the Timeline Analysis feature. NETLEADS also permits its users to look for trends across cases. These tools allow disparate data elements to be examined within the analysis constructs of ICE investigations and intelligence operations. All data evaluated is owned by consigned partners within ICE thus ensuring secure access and auditing activities.

NETLEADS includes the ICE nationwide Significant Event Notification (SEN) application providing real time notification to ICE Operations Center incident and intelligence activities. The SEN has an e-mail alert capability that notifies key ICE managers that new incident data is available. NETLEADS uses dynamic AdHoc data query allowing agents and analyst to search any topic, while also possessing a canned search query to allow for routine searches.

1. Purposes of the Program

The NETLEADS project provides the ability to meet the investigative and intelligence community's needs. The tools produce reports of search results and link diagrams, consolidated from all data sources searched. Patterns, relationships, and rules produced through data mining are used as investigative aids in the same way as patterns based on intelligence or observation. The tools within the suit also notify the user when additional information has been received that meets his search criteria.

2. Data Sources

NETLEADS performs simultaneous search and analysis of multiple databases with over 50,000,000 records, documents, and images from fifteen data sources stored in them. NETLEADS tools perform searches on authorized databases maintained by DHS via the protected DHS secure intranet. ICE is engaged in creating and executing Memorandums of Understandings and Service Level Agreements that authorize data sharing efforts with other federal and state government law enforcement and intelligence agencies, such as

Department of State. In addition to government databases, NETLEADS tools access data from commercial sources, such as geographical location data and news feeds.

NETLEADS generates derived records from multiple sources and stores this data for analysis. The records are indexed for searching, analysis and the production of link analysis and new all-source products. NETLEADS includes both enterprise-level and desktop-level tools. These tools are intranet Web enabled, providing access to all authorized users with no additional software or special installations. The use of a web browser allows users to gain access to data and the ability to work in environments such as joint taskforce, forward operations, or offices with inconsistent communications.

Derivative documents can be re-introduced into NETLEADS with an upload capability for authorized users completing the investigative and intelligence life cycle. The cycle would generally be to identify the need, create a collection plan, conduct the collection of the information, assemble a project folder, conduct the analysis, evaluate the analysis and collection and produce a final intelligence or investigative product.

3. Deployment Dates

The NETLEADS project has been in operation for approximately seven years. Over 10,000 ICE Special Agents, Border Patrol Agents, the Border Patrol Field Intelligence Center, Sector Intelligence offices, and ICE Intelligence Communities in ICE and CBP offices in Ports of Entry, Sea Ports, Airports at over 344 worldwide have access to NETLEADS.

4. Policies, Procedures and Guidance

NETLEADS tools search mostly existing government databases. These databases are covered by System of Records Notices under the Privacy Act of 1974, as appropriate. Because several databases predate the E-Government Act of 2002, they have not yet undergone a Privacy Impact Assessment.

The NETLEADS database is an intelligence and investigative database that uses derived information from other data systems. This data at the time NETLEADS was activated was identified as a law enforcement sensitive system. Information that populates NETLEADS is criminal alien information, terrorism, smuggling, and criminal case information.

The information is used only by authorized DHS employees who have a need to know and have been subject to appropriate security background investigations. Users who have access to NETLEADS tools have role-based access to underlying databases. When a user logs in, he or she will only see the databases that are associated with his or her respective job role, and responsibilities. Prior to being given access to these databases, all users receive annual ADP security training that encompasses the appropriate uses of information for law enforcement and intelligence purposes. Within the past year all

NETLEADS user accounts were evaluated for need of access and a more stringent security level was implemented (security level requirements were raised from a T5 to a T6). NETLEADS includes audit logs, which are examined daily to identify anomalous activities.

NETLEADS databases use the same criteria for storage and access as the source databases. The NETLEADS system has successfully passed critical inspection by the DHS ADP security officer for Certification Authorization (C&A), meeting or exceeding all DHS requirements for sensitive data storage and handling.

The NETLEADS applications monitor all activities of its users including reading, writing, and printing of files. Entries into databases are audited by the user account identification and are available for audit. The underlying data for NETLEADS is an Oracle 9.i database that provides the ability to conduct multiple layer auditing. Security is enhanced through this process because each user is subject to account auditing for every document read, printed, or downloaded to desktop workstations for analysis.

C. ICE Pattern Analysis and Information Collection System (ICEPIC) Immigration and Customs Enforcement

ICEPIC is an application designed to facilitate ICE counterterrorism activities by increasing efficiency of data searches and analyses. ICEPIC tools permit analysts to perform searches and analysis of a large number of databases through a greatly simplified query.

ICEPIC uses IBM’s Non-Obvious Relationships Awareness (NORA) technology to identify entities and individuals that may appear under different names and in different circumstances but are, in fact, the same entity or individual, and to identify entities and individuals that are related to each other. While this technology is not used to predict behavior, NORA does use algorithms to perform link analysis to uncover associations between entities and/or individuals.

1. Purposes of the Program

ICEPIC produces reports of search results, consolidated from all data sources searched, to support counterterrorism leads generation and intelligence analysis. Leads generation provides analysts with potential subjects for further investigation or validates existing subjects for better identification. Intelligence analysis provides analysts with better understandings of the connections between objects whether they are entities or individuals. Patterns, relationships, and rules produced through ICEPIC data mining are used as investigative aids in the same way as patterns based on intelligence or observation.

2. Data Sources

ICEPIC uses data from multiple databases. Most of the databases used with ICEPIC tools are maintained by DHS. In addition, ICEPIC searches databases from the Department of State, the Department of Justice, and the Social Security Administration. No data from commercial sources is used by ICEPIC at this time. The use of NORA technology requires duplication and storage of the source databases and data sets.

3. Deployment Dates

ICEPIC is currently in the pilot stage, but is mission-operational with between 12 and 15 users. A full scale (enterprise) roll-out is anticipated in FY06, at which time the number of users will likely increase significantly.

4. Policies, Procedures, and Guidance

Policies, procedures and guidance for ICEPIC are being developed at the same time as its analytic tools. However, security policy, plans and guidance have been developed and maintained since inception and initial deployment of ICEPIC. As such, ICEPIC operates currently under full compliance with DHS IT security policy directives, FISMA requirements, and other applicable IT security requirements. The present ICEPIC deployment was certified as a major application by ICE Office of the Information Systems Security Manager (OISSM) in October 2005 and was granted full approval to operate by the ICE Director of Investigations in November 2005.

All applicable security controls and countermeasures required for “moderate” systems and applications have been implemented, validated, tested, and verified for the ICEPIC steady-state system.

As ICEPIC transitions from the operational pilot stage to an enterprise deployment, it will continue to undergo all reviews, evaluations and certifications required for an operational DHS system including re-certification and accreditation.

Source databases that are searched through ICEPIC tools are covered by System of Records Notices under the Privacy Act of 1974, as appropriate. Because several databases predate the E-Government Act of 2002, they may not have not as of yet have undergone a Privacy Impact Assessment. A Privacy Impact Assessment of ICEPIC is expected to be performed before ICEPIC enterprise deployment.

Individuals who are granted access to ICEPIC must also have approved access to the underlying source databases. Prior to being given access to these databases, all users receive security training and training on appropriate uses of information for law enforcement and intelligence purposes. All ICEPIC users are also required to read, acknowledge through signature, and adhere to “Rules of Behavior” specific to the ICEPIC application.

ICEPIC tools, in their current state of development, require manual extracts of data, and data is updated only when those extracts are received. Data validation is performed by evaluating data for internal consistency, through the evaluation of search results by analysts, and through the additional evaluation of the leads by field agents who are the ultimate consumers of the results. No mechanism exists within ICEPIC to determine whether the underlying data has been modified, although this has been recognized as a potential issue by users and may be addressed through audit logs on deployment.

D. Intelligence and Information Fusion (I2F)
Office of Intelligence and Analysis

1. Purposes of the Program

I2F is designed to provide intelligence analysts with an ability to view, query, and analyze multiple data sources. The program intends to use mostly COTS tools, including tools for search, link analysis, entity resolution, geospatial analysis, and temporal analysis. The tools may include the ability to discover across multiple data sources unpredicated patterns, relationships, and rules.

2. Data Sources

The developers anticipate that I2F can incorporate data from both government and commercial sources. All DHS databases can be included. All data storage media and all digital formats will be included. At this time, “virtual” data repositories are envisioned, i.e., leaving most data in place at the source and bringing into the I2F data repository specific items of interest. Data is expected to be validated for accuracy and internal consistency through comparison of multiple data sources.

3. Deployment Dates

The program is currently in development.

4. Policies, Procedures and Guidance

Policies and procedures are under development at this time. Data will be collected and retained under the authority of Executive Order 12333. It is anticipated that Memoranda of Understanding will be signed between DHS and various agencies that will either contribute databases to I2F or that will receive results based on I2F analysis.

I2F will complete a Privacy Impact Assessment prior to the start of data collection. All persons associated with the development or use of I2F will receive periodic training in their responsibilities relative to applicable privacy and intelligence oversight laws, regulations, and policies. The system will undergo periodic reviews to assess the program with regard to compliance.

E. Fraud Detection and National Security Data System (FDNS-DS) U.S. Citizenship and Immigration Services

The Fraud Detection and National Security Data System (FDNS-DS), formerly known as the Fraud Tracking System, is being developed by U.S. Citizenship and Immigration Services ("USCIS") to track immigration related fraud, public safety referrals to ICE, and national security concerns discovered during the background checks performed on individuals applying for immigration-related benefits. At present, FDNS-DS is a case management system used to track fraud leads, cases where a suspicion of fraud has been articulated, referrals to ICE, and requests for assistance from law enforcement agencies. In the future, FDNS-DS is envisioned to have the capability to identify immigration fraud schemes nation-wide through the use of analytics tools. Currently SCCLAIMS, which is a copy of the data contained in the Computer-Linked Application Information Management System (CLAIMS), is used as a way to search for other receipts that match a known or suspected fraud scheme. For example, if it is known that fraudulent applications are filed using a fake address, SCCLAIMS can be used to search for other applications that use that address. As FDNS-DS develops, it will replace SCCLAIMS.

1. Purposes of the Program

Future data mining activities envisioned for FDNS-DS would look for new fraud patterns and new associations to provide additional investigative leads. These leads would be used in the same way as leads developed from other sources.

2. Data Sources

At present, the underlying data source for FDNS-DS is CLAIMS. Officer's notes and reports are also stored in FDNS-DS. It is expected that case specific data from commercial data aggregators will be stored in future releases of FDNS-DS.

3. Deployment Dates

The initial release of FDNS-DS is operational. The analytics tool is expected to be available in the next year. A release that would include data mining tools is not anticipated to be available within the next two to three years.

4. Policies, Procedures and Guidance

CLAIMS, the underlying database for FDNS-DS, is covered by a System of Records Notice under the Privacy Act of 1974. A Privacy Impact Assessment for the FDNS-DS (using the name Fraud Tracking System) was published on June 24, 2005. The system has been certified and accredited in accordance with FISMA requirements. All relevant documents will be updated as the new tools and data sources are brought online and as new capabilities are developed.

USCIS trains each of its FDNS Immigration Officers and Intelligence Research Specialists extensively and continuously on the proper use of information in its databases and each of its analysis and query tools. This training will be updated and reinforced each time a new tool, such as the pattern analysis module of FDNS-DS, is introduced. In addition, FDNS-DS and SCCLAIMS functional administrators provide regular reviews as part of oversight and auditing of the types of queries being conducted. FDNS-DS and SCCLAIMS have role-based access.

Policies, procedures, and guidelines for protection of privacy and civil rights of those who are suspected or accused of immigration-related fraud are well established, and would apply to potential fraud identified through data mining as they do to potential fraud identified through other means, such as tips or statistical random sampling.

F. National Immigration Information Sharing Office (NIISO) US Citizenship and Immigration Services

NIISO is a program within USCIS FDNS and is responsible for fulfilling requests for information from other DHS components, as well as, external law enforcement and intelligence agencies. USCIS, in conjunction with the Office of Intelligence and Analysis (OI&A) at DHS, is exploring the potential expansion of the NIISO program to facilitate more involvement and additional requests from the law enforcement community and Intelligence Community, including inclusion of data mining tools and techniques.

1. Purposes of the Program

At present, FDNS personnel perform searches of DHS's extensive repository of Alien files and electronic databases containing immigration information.

2. Data Sources

The NIISO program uses USCIS CLAIMS, specifically SCCLAIMS, as a way to identify the relevant immigration information. The FDNS-DS will also be utilized as the system develops. Future plans call for the FDNS-DS to include analytics capabilities that would replace the use of SCCLAIMS. Relevant records that are identified are further investigated and evaluated through use of the paper and electronic immigration files of DHS. Publicly available information, and data from commercial data aggregators, may also be searched.

3. Deployment Dates

NIISO is currently operational. A pilot project of the expanded functionalities is projected for some time in FY06.

4. Policies, Procedures and Guidance

For the expanded NIISO, the concept of operations, policies, procedures, and interagency agreements are currently under development.

The current proposal includes the creation of an audit system that logs all searches and model development efforts conducted by NIISO. This would provide a mechanism for senior management review and oversight of data uses and disclosures.

V. Conclusions and Recommendations

Based on our analysis of DHS activities that involve current and projected future uses of data mining, we note that data mining is usually only one part of a larger set of analytic activities and tools. Such analytic activities include traditional searches and analyses, as well as data mining. Although programs that employ data mining tools and technologies also employ traditional privacy and security protections, such as Privacy Impact Assessments, Memoranda of Understanding between agencies that own source data systems, privacy and security training, and role-based access, we recommend additional protections that are aimed specifically at addressing the privacy concerns raised by data mining and we will take steps to implement these recommendations within the Department.

1. Prior to the start of any data mining activity, the authority of the agency to undertake such activity should be determined to be consistent with the purposes of the data mining project or program. The authority to collect or aggregate data required to perform the data mining project should also be ascertained, whether the project involves collection of new data or aggregation of existing data from various sources. While oversight functions exist in different components within DHS[26], the Department as a whole could benefit from more centralized oversight with a broader view of DHS activities and data holdings. One such body, which could assume the function of overseeing DHS data mining programs, is the DHS Privacy and Data Integrity Board, charged under the Privacy Act to examine and approve data matching agreements between DHS and other departments. The Board, which includes representatives from all DHS components, the Office of the Inspector General, the Chief Information Officer, and the Office of General Counsel, and is chaired by the DHS Chief Privacy Officer, could provide oversight and confirmation to ensure responsible application of data mining tools and technologies. The Board assisted with the collection of data for this report concerning current and planned data mining programs within DHS.

2. As discussed in the report, data mining searches for patterns, relationships, and rules in the data without basing this search on observations or a theoretical model.

[26] Including the DHS Office of Civil Rights and Civil Liberties.

Because the existence of patterns in the data does not reflect cause and effect, data mining tools should be used principally for investigative purposes. DHS components that use data mining tools should have written policies, stating that no decisions may be made automatically regarding individual rights or benefits solely on the basis of the results produced by patterns or rules derived from data mining.

3. Because the patterns, relationships, and rules in the data may not be derived from specific personal identifiers, such as a name or Social Security Number, when a data set includes personally identifiable information, data mining projects should give explicit consideration to using anonymized data in data mining activities. A discussion of the extent to which anonymization was considered should be included in the Privacy Impact Assessment for such a data mining project.

4. Data quality plays an important role in the ability of data mining techniques to produce accurate results. DHS should adopt data quality standards for data used in data mining. Application of these standards, which should affect systems using both data from government and commercial sources, should be ensured prior to the deployment of data mining models or predictive rules for use in the field.

5. In order to ensure that data mining models produce useful and accurate results, DHS should adopt standards for the validation of models or rules derived from data mining. Evaluation of the model validation process and the ability to meet these standards should be reviewed and documented prior to the deployment of the model to the field.

6. Each component that employs data mining should implement policies and procedures that provide an appropriate level of review and redress for individuals identified for additional investigation by patterns, relationships, and rules derived from data mining. Because data mining algorithms most times produce highly complex patterns, relationships, and rules that may not be fully understandable as to the particular reasons for identifying the individuals, a complete procedure should include a step that a person, acting independently from the data mining process, substantiates the particular identification of individuals, prior to any determinative processes and procedures. To ensure a complete understanding of the capabilities and limitations of data mining in this regard, employees who use data mining processes should be required to complete training on these polices and procedures.

7. In order to provide demonstrable accountability, each component that employs data mining should include strong, automatic audit capabilities to record access to source data systems, data marts, and data mining patterns and rules. Programs should conduct random audits at regular intervals, and all employees should be given notice that their activities are subject to such audits. These actions help underline the importance of transparency.

VI. Appendix A

Data Mining Process Models

Different organizations have designed different processes for data mining, although these processes have similarities. Table 1 shows two data mining process models, one from a U.S. organization and the other from a European consortium, Cross-Industry Standard Process for Data Mining (CRISP-DM).

Two Crows Corporation Model[27]	**CRISP-DM Model**[28]
1. Define business problem	1. Business understanding
2. Build data mining database	2. Data understanding
3. Explore data	3. Data preparation
4. Prepare data for modeling	4. Modeling
5. Build model	5. Evaluation
6. Evaluate model	6. Deployment
7. Deploy model and results	

Table 1: Data Mining Process Models

In addition to the process models shown above, there are process models that focus more narrowly on the model-building and validation process.

The model used as a basis for this report is a synthesis of the models presented above and of other publicly available information.

[27] Two Crows Corporation, op. cit., p. 22.

[28] Cross-Industry Standard Process for Data Mining, available at < http://www.crisp-dm.org/Process/index.htm>, last visited December 27, 2005.