Office of the Director of National Intelligence

2011 Data Mining Report

For the Period January 1, 2011 through December 31, 2011

Office of the Director of National Intelligence
2011 Data Mining Report
January 1, 2011 through December 31, 2011

I. Introduction

The Office of the Director of National Intelligence (ODNI) provides this report pursuant to Section 804 of the *Implementing the Recommendations of the 9/11 Commission Act of 2007*, entitled *The Federal Agency Data Mining Reporting Act of 2007* (Data Mining Reporting Act).

A. Scope

This report covers the activities of all ODNI components from January 1, 2011 through December 31, 2011. Constituent elements of the Intelligence Community (IC) will report their activities to Congress through their own departments or agencies.[1]

B. Reporting Requirement

The Data Mining Reporting Act requires "the head of each department or agency of the Federal Government that is engaged in an activity to use or develop data mining shall submit a report to Congress on all such activities of the department or agency."[2] This is an annual requirement. Under the Act, "data mining" is defined as:

> "… a program involving pattern-based queries, searches or other analyses of one or more electronic databases, where —
> (A) a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals;
> (B) the queries, searches, or other analyses are not subject-based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases; and
> (C) the purpose of the queries, searches, or other analyses is not solely— (i) the detection of fraud, waste, or abuse in a Government agency or program; or (ii) the security of a Government computer system.[3]"

When focusing on individuals or groups, the ODNI typically uses analytic tools and techniques that rely on "personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals," such as link-analysis tools. Unlike the predictive, pattern-

[1] Section 804(c)(1) of the Data Mining Reporting Act.
[2] Section 804(c)(1) of the Data Mining Reporting Act.
[3] Section 804(b)(1)(A) of the Data Mining Reporting Act

based technologies envisioned by the Act, these tools and techniques utilized by ODNI components start with a known or suspected terrorist, or other subject of foreign intelligence interest, and use various methods to uncover links or relationships between the known subject and potential associates or other persons with whom that subject has a "link" (a contact or relationship). Thus, such analytic tools and techniques do not fall within the statutory definition of data mining.

C. Report Content

Part II of this report describes those ODNI programs that meet the reporting requirements of the Data Mining Act. For this year's submission, there are no reportable programs.

Part III of this report includes discussion of five programs, all of which were included in last year's report. Although these programs are not "data mining" programs, information is nonetheless provided in the interest of transparency.

Of the five programs discussed, two were discontinued during the reporting period – the IC CIO's CATALYST program and the National Counterterrorism Center's (NCTC's) DataSphere program. Two programs from the Intelligence Advanced Research Projects Activity (IARPA), KDD and ALADDIN Video, were again included in this year's report since technologies investigated by, or later developed from, these programs could be used to support data mining. SPAR, the fifth program, also sponsored by IARPA, was included to provide information on research that may have applicability in enhancing security and privacy protections in data mining activities.

Descriptions of these programs are provided in narrative form, with information generally responsive to the reporting elements of the Data Mining Reporting Act.

D. Protection of Privacy and Civil Liberties.

The ODNI Civil Liberties and Privacy Office (CLPO) works closely with the ODNI Office of General Counsel, ODNI components and the IC elements to ensure appropriate legal, privacy, and civil liberties safeguards are incorporated into policies, processes and procedures that support the intelligence mission. The CLPO is led by the Civil Liberties Protection Officer, a position established by the IRTPA. The duties of this Officer are set forth in that Act, and include: ensuring that the protection of civil liberties and privacy is appropriately incorporated in the policies of the ODNI and the IC; overseeing compliance by the ODNI with legal requirements relating to civil liberties and privacy; reviewing complaints about potential abuses of privacy and civil liberties in ODNI programs and activities; and ensuring that technologies sustain, and do not erode, privacy protections relating to the use, collection, and other disclosure of personal information."[4]

[4] National Security Act of 1947, as amended by the Intelligence Reform and Terrorism Prevention Act of 2004 (IRTPA), 50 § 403-3d.

The IC has in place a protective infrastructure built in principal part on a core set of U.S. Person rules derived from Executive Order (EO) 12333. This EO requires each IC element to maintain procedures, approved by the Attorney General, governing the collection, retention and dissemination of U.S. Person information. These procedures limit the type of information that may be collected, retained or disseminated to the categories listed in part 2.3 of the EO. Each IC element's Attorney General-approved U.S. Persons guidance is interpreted, applied, and overseen by that element's Office of General Counsel, Office of Inspector General, and other compliance offices as appropriate. Violations are reported to the Intelligence Oversight Board of the President's Intelligence Advisory Board. In addition to EO 12333, IC elements are subject to the requirements of the Privacy Act, which protects information about U.S. citizens and permanent resident aliens that a government agency maintains and retrieves by name or unique identifier.

The IC's privacy and civil liberties protective infrastructure is bolstered also by guidance and directives issued by the Office of Management and Budget, including memoranda regarding the reporting of and response to incidents involving personally identifiable information and the minimization of Social Security Numbers.

Before any tool or technology could be used in an operational setting, the use of the tool or technology would need to be examined pursuant to EO 12333, the Privacy Act, and other applicable requirements to determine how the tool could be used consistent with the framework for protecting information about Americans and other U.S. Persons.

CLPO has been considering how advanced technologies, employed in accordance with proper laws and policies, enable sharing and use of information while protecting privacy and civil liberties. Such privacy-enhancing technologies (PETs) also prove useful in providing protections for data mining activities. CLPO is making available the results of its PET-related research to IC elements and other government offices with which CLPO collaborates, and continues to consider how to make use of PETs for intelligence activities.

II. ODNI Data Mining Activities

The ODNI did not engage in any activities to use or develop data mining functionality in the reporting period.

III. Other Programs

The following programs are reported in the interest of transparency.

A. CATALYST

The Catalyst program, managed by the IC CIO, was established as a means to improve inter-agency, multi-intelligence (multi-INT) information sharing by providing IC analysts with capabilities for entity disambiguation, correlation, and co-referencing.

During Fiscal Years 2010 and 2011, program activities remained focused on research activities that did not include development of a pattern-matching functionality. Although such functionality had been contemplated for later development phases, the Catalyst Program was discontinued in 2011.

B. DATASPHERE

The DataSphere program, managed by the National Counterterrorism Center (NCTC), continued as a pilot to develop a means to enhance data fusion and entity resolution, as well as the discovery of unknown relationships.

Although plans for future iterations of DataSphere contemplated the development of a pattern-matching capability, no such functionality was developed and the DataSphere effort was discontinued in 2011.

C. IARPA Research Programs

The mission of IARPA is to invest in high-risk/high payoff research programs that have the potential to provide the United States with an overwhelming intelligence advantage over its future adversaries. It does not have an operational mission and it does not deploy technologies directly to the field. As a scientific research funding organization, IARPA does not use, nor does it expect to make use of, data mining technology. IARPA programs are by nature experimental and are designed to produce new capabilities. The end goal of an IARPA program is typically a proof-of-concept experiment or prototype of an entirely new capability. Due to their high-risk research nature, IARPA programs do not always achieve their end goals, and when they do, further steps are required to transform the results into real world applications. Any results from IARPA research programs that do get incorporated into future operational programs within the IC, or other parts of the United States government, will be subject to appropriate legal, privacy, civil liberties and policy safeguards.

As with last year's report, the KDD and ALADDIN Video programs below are reported in the interest of transparency, due to the potential that technologies explored by, or later developed from, those programs might be ultimately used to support data mining (such technologies would then be subject to reporting under the Data Mining Reporting Act).

1. *Knowledge Discovery and Dissemination (KDD) Program.* The KDD scientific research program is an IARPA program begun in 2009. A Broad Agency Announcement (BAA) for KDD was released on December 22, 2009 and KDD research contracts were awarded in September 2010. The KDD program completed its first period in December 2011.

The objective of the KDD program is to enable an analyst to utilize large, complex and varied data sets that he has not seen before to produce actionable intelligence in a timely manner. KDD tackles two significant technical areas: (1) how to quickly understand the novel data sets so that the contents can be correctly integrated with data sets that are already in use (this is termed "alignment"); (2) how to construct automatic analysis tools that are able to work effectively

across multiple aligned data sets. KDD research results will be evaluated using realistic challenge problems throughout the program.

In evaluations of research teams' prototypes, the KDD scientific research program utilizes real-world, classified data sets that are large and complex. KDD researchers' work is evaluated in the context of challenge problems using these data sets. The challenge problems are not problems that require data mining technology as defined by the Act. The data sets used by researchers are highly varied, including, for example, regional biographic data, incident reports, translated newspaper articles, etc. The use of all data sets is consistent with all U.S. laws and regulations.

2. *Automated Low-level Analysis and Description of Diverse Intelligence Video (ALADDIN Video) Program.* The ALADDIN Video scientific research program released a BAA in June 2010, and research contracts were awarded in February 2011.

The objective of the ALADDIN program is to enable an analyst to query large video data sets to quickly and reliably locate those video clips that show a specific type of event. The ALADDIN program is researching technologies designed to automatically search large numbers of video data files for analyst-defined events of interest and direct the analyst to those video data files that are likely to contain occurrences of those events. ALADDIN's technologies, if successful, will help to automate a triage process that is currently performed largely manually by analysts. Although this is not "data mining," technologies that result from ALADDIN research could, potentially, be applied by operational organizations to support capabilities that involve pattern recognition.

ALADDIN research addresses three significant technical areas: (1) High-speed processing of large amounts of video clips to extract information that can later be used to support queries about each clip's contents; (2) Generation of effective queries from small sets of example video clips and a textual description; (3) Robust query processing that identifies the clips of interest and summarizes the rationale for their selection. ALADDIN research results will be evaluated by IARPA and the National Institute for Standards and Technology (NIST).

The ALADDIN program uses video data files in its research and evaluations that are acquired by NIST for its annual, international video search technology research program (TRECViD). TRECViD sponsors public evaluations of video and multimedia search technologies that are open to worldwide participation. ALADDIN performers will participate in these evaluations to demonstrate objective progress in their research. The data collections used in the TRECViD evaluations are made available to all participants through an evaluation participation agreement that stipulates that the TRECVID data collections are to be used for research purposes only. The TRECViD data is collected using a rigorous process that protects privacy.

3. *Security and Privacy Assurance (SPAR) Program.* The SPAR program is a follow-on to the Automatic Privacy Protection (APP) program discussed in the 2009 and 2010 ODNI Data Mining Reports. Neither the SPAR nor APP programs involve data mining, but the research results from both programs may enhance security and protect privacy in data mining activities.

The APP program ended in 2010 after achieving two goals. First, it developed secure distributed private information retrieval (PIR) protocols that permit an entity (Client) to query a cooperating data provider (Server) and retrieve only the records that match the query without the Server learning what query was posed or what results were returned. These protocols are able to add only minimal overheads in computation and communication for simple queries and databases by using a cooperating third party who has access only to encrypted data. Second, APP demonstrated algorithms to determine automatically if complex queries are in compliance with privacy policies. This allows a Client's auditor with access to the policy and the query history to rapidly verify that only authorized queries have been submitted to the Server.

The SPAR program was launched in 2011 to build on the successes of APP and explore additional applications of PIR to realistic IC scenarios. SPAR includes research projects in three technical areas. The first technical area protects security and privacy for database access. Unlike the simple queries and static databases of APP, SPAR will investigate protocols that handle multiple types of complex queries and databases whose records are frequently created, deleted, or updated. In addition, the protocols must integrate policy compliance checking with the security and privacy assurances so that the Server can verify that a query is compliant with a policy even though the query is never learned. The second technical area will build on advances in fully homomorphic encryption (FHE) schemes to implement PIR without relying on any third parties. FHE is a recent breakthrough result of thirty years of cryptographic research, but current schemes are impractical due to high costs in time and memory. SPAR will attempt to explore gains in performance by modified FHE schemes that support only the computations necessary for information retrieval. The third technical area will investigate applications of PIR to the specialized information sharing architectures of publish/subscribe, email/message queues, and outsourced data storage systems.

If successful, the SPAR program will benefit the IC by securing and protecting the privacy interests of both the custodians and the consumers of data. The technology may enhance cooperative information sharing within the IC, and among Government and the private sector, by expanding policy options for satisfying security and privacy concerns when information is shared.