Office of the Director of National Intelligence

2009 Data Mining Report

For the Period of February 1, 2009 through December 31, 2009

The Office of the Director of National Intelligence (ODNI) is pleased to provide to Congress its third report pursuant to the Data Mining Reporting Act.[1] 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," as defined by the Act, to report annually on such activities to the Congress.

Introduction

Scope. This report covers the data mining activities of all elements of the ODNI from February 1, 2009 through December 31, 2009.[2] Constituent elements of the Intelligence Community (IC) are reporting their data mining activities to Congress through their own departments or agencies.[3]

Last year's ODNI data mining report detailed a single ongoing effort within the "Incisive Analysis" Office in the Intelligence Advanced Research Projects Activity (IARPA) that included research of general-purpose techniques that could, among other things, be applied to data mining. That program, the Video Analysis and Content Extraction (VACE) program, ended in September 2009. Details on the completed VACE program are included in an appendix to this report.

This year's report includes details on two new IARPA programs, the Knowledge Discovery and Dissemination (KDD) program and the Automated Low-level Analysis and Description of Diverse Intelligence Video (ALADDIN Video) program. Both programs were in formulation during 2009. Once again, although neither program is a data mining program, technologies developed by the programs could potentially be used to support data mining and therefore they are being included in this report. We also provide an update on a privacy protection program noted in last year's data mining report. Finally, an appendix provides summary information on two efforts being administered by the ODNI Chief Information Officer (CIO) that do not meet the criteria for full reporting in this report.

This report covering ODNI activities is unclassified and has been made available to the public through the ODNI's website. For completeness, a classified annex containing more detailed information on VACE (discontinued in September 2009) has been prepared and transmitted to the appropriate Congressional committees.

Definition of "data mining." The Data Mining Reporting Act defines "data mining" as:

---

[1] Section 804 of the *Implementing the Recommendations of the 9/11 Commission Act of 2007*.

[2] Last year's Data Mining Report covered the period of January 31, 2008 through January 31, 2009. Going forward, Data Mining Reports will cover activities within a given calendar year (for example, January 1, 2010 through December 31, 2010).

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

a program involving pattern-based queries, searches or other analyses of 1 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.[4]

This definition limits covered activities to predictive, "pattern-based" data mining, which is significant because analysis performed within the ODNI and its constituent elements for counterterrorism and similar purposes is often performed using various types of link analysis tools. Unlike "pattern-based" tools, these link analysis tools start with a known or suspected terrorist or other subject of foreign intelligence interest and use various methods to uncover links between that known subject and potential associates or other persons with whom that subject is or has been in contact.

The Data Mining Reporting Act does not include such analyses within its definition of "data mining" because such analyses are not "pattern-based." Rather, these analyses rely on inputting the "personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals," which is excluded from the definition of "data mining" under the Act.

ODNI is neither involved in, nor does it directly employ, pattern-based data mining programs to discover or locate patterns or anomalies indicative of terrorist or criminal activity in any of its constituent elements, such as the National Counterterrorism Center, National Counterproliferation Center, National Intelligence Council or other offices within ODNI.

However, within the ODNI's Intelligence Advanced Research Projects Activity (IARPA) there are two research programs within the Office of Incisive Analysis that are exploring technologies which could, potentially, be applied to pattern based analysis at some future point in time, as described below. This report details those activities because the Act requires a report on any "activity to . . . develop data mining."[5] Appendix 2 of this report also provides summaries of two efforts being administered by the ODNI/CIO that do not meet the reporting requirements of the Act, but that are included for completeness. Finally, this report also details an ongoing IARPA research program aimed at developing technologies that will enhance privacy protection.

---

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

Background on IARPA. It is IARPA's mission to invest in high-risk/high payoff research programs that have the potential to provide the U.S. with an overwhelming intelligence advantage over its future adversaries. IARPA's time horizon is measured in years, not months. It does not have an operational mission and it does not deploy technologies directly to the field. IARPA programs are by nature highly experimental and pioneering and are designed to produce new capabilities not even imagined by the operational agencies it serves. The end goal of an IARPA program is typically a proof-of-concept experiment or prototype of a never-before-seen capability. Because IARPA programs are on the cutting-edge of research, they do not always achieve their end goals, but even 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.

Report on ODNI Data Mining Activities:

(A) **A thorough description of the data mining activity, its goals, and, where appropriate, the target dates for the deployment of the data mining activity**.

- The Knowledge Discovery and Dissemination (KDD) scientific research program is a new program, begun in 2009.[6] While KDD does not itself constitute data mining, the program does plan to investigate new technologies that, if successful, could potentially be applied to support data mining (amongst other potential uses), as defined by the Act, and is therefore being included in this report. KDD would enable analysts to utilize large, complex, varied and unfamiliar data sets to produce actionable intelligence in a timely manner. The two technical challenge areas addressed by KDD are: (1) how to align the data models of multiple data sets; and (2) how to build advanced analytic algorithms that can work across multiple data sets.

    - If KDD is successful, then it may be possible for a program involving pattern-based queries, searches or other analyses, which was previously only able to use one data set, to more easily use multiple data sets.

KDD held a Proposers' Day in August 2009, and a Broad Agency Announcement (BAA) was released on December 22, 2009. KDD research activities are expected to begin in June 2010.

- The Automated Low-level Analysis and Description of Diverse Intelligence Video (ALADDIN Video) scientific research program is a new program that aims to capitalize on capabilities developed through VACE, referenced above, as well as other similar projects. ALADDIN began in 2009. ALADDIN does not constitute data mining, but the ALADDIN program plans to investigate new technologies that, if successful, could potentially be used to support data mining (in addition to many other uses). Specifically, ALADDIN seeks to enhance the ability of an analyst to quickly and reliably locate video of interest in very large and complex video data sets. Thus, ALADDIN seeks to dramatically increase the level of automation available to support the analysts who already review video data. Technical challenges addressed

[6] As noted in the 2008 Data Mining Report, a program using the KDD name was discontinued in early 2008, and although this new program utilizes the same name, the focus of this new program is substantively different.

by ALADDIN include filtering, metadata generation, content description, detection of events of interest, and recounting of events of interest.

- If ALADDIN is successful, then resulting technologies may augment capabilities for pattern-based searches, queries, or analyses of video data.

ALADDIN held a Proposers' Day in October 2009 and is expected to release a Broad Agency Announcement (BAA) in April 2010. ALADDIN research activities are expected to begin in August 2010.

(B) A thorough description of the data mining technology that is being used or will be used, including the basis for determining whether a particular pattern or anomaly is indicative of terrorist or criminal activity.

While there are no research programs within the Incisive Analysis Office that constitute data mining, as explained in Section A above, the KDD and ALADDIN programs will research technologies that could potentially support data mining programs, as defined by the Act, and are therefore included below. As a scientific research funding organization, IARPA does not use, nor does it expect to make use of, data mining technology.

- The KDD program is intended to enhance our ability to quickly produce actionable intelligence from unanticipated multiple and varied data sets. This requires research advances in two key areas: (1) alignment of data models; and (2) advanced analytic algorithms. While KDD will research techniques to support data model alignment and portability of advanced analytic technologies, neither is "data mining." However, results from KDD research could potentially be applied by operational organizations to support capabilities that involve pattern recognition. Such pattern recognition algorithms would have to be developed independently, by an operational organization.

- The ALADDIN program will research technologies designed to skim through large numbers of video data files and direct an analyst to those video data files that are more likely to contain events of interest. ALADDIN's technologies, if successful, will help to automate a filtering process that is currently performed manually by analysts. This is not "data mining." However, technologies that result from ALADDIN research could, potentially, be applied by operational organizations to support capabilities that involve pattern recognition.

(C) A thorough description of the data sources that are being or will be used.

- In evaluations of research teams' prototypes, the KDD scientific research program will utilize real-world, classified data sets that are large and complex. KDD researchers' work will be evaluated in the context of challenge problems using these data sets. The data sets used by KDD researchers will be highly varied, and may include, for example, regional biographic data, incident reports, translated newspaper articles, etc. All data sets used will be consistent with applicable civil liberties and privacy laws/regulations.

- The ALADDIN program will use video data files in its evaluations that are acquired by the National Institute of Standards and Technology (NIST). The same data sets will be used by NIST in its annual, international, Video Retrieval (TRECVID) research program. The NIST data is only available to support multimedia research, and only after users sign a usage agreement. For more information on NIST's TRECVID, see http://www-nlpir.nist.gov/projects/trecvid.

(D) An assessment of the efficacy or likely efficacy of the data mining activity in providing accurate information consistent with and valuable to the stated goals and plans for the use or development of the data mining activity.

- The goal of the KDD program is to enhance our ability to rapidly align, analyze and integrate data from multiple, varied, data sets. As such, KDD research and testing will not utilize, or promote, pattern based techniques as defined in the Data Mining Reporting Act. The KDD researchers' results will be evaluated on an annual basis using classified data sets. The evaluations involve measuring how well a set of straightforward analysis queries are answered via workflow scripting, and also how well trained analysts do when their toolset is augmented by KDD research prototypes. For more details, see the KDD Broad Agency Announcement, http://www.iarpa.gov/solicitations_kdd.html.

- All ALADDIN performers will be required to participate annually in the National Institute of Standards and Technology's TRECVID (the video retrieval evaluation series of NIST's Text Retrieval Conference), a long-standing, NIST-coordinated, open video extraction/search technology evaluation series with significant worldwide participation. TRECVID is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations. By requiring performers to annually take part in TRECVID, they will be vetted in an ongoing and public manner, measured against their colleagues around the world, thereby providing an ongoing opportunity for IARPA to assess the quality of their work.

(E) An assessment of the impact or likely impact of the implementation of the data mining activity on the privacy and civil liberties of individuals, including a thorough description of the actions that are being taken or will be taken with regard to the property, privacy, or other rights or privileges of any individual or individuals as a result of the implementation of the data mining activity.

IARPA recognizes that data mining techniques explored as part of a research program could, potentially, impact the privacy or civil liberties of individuals if they are successfully transitioned to an operational partner without careful consideration of these issues. To this end, IARPA maintains a longstanding relationship with the ODNI CLPO for the purpose of validating that its research programs are conducted consistent with the protection of individual privacy and civil liberties. IARPA researchers communicate with CLPO personnel multiple times throughout the lifecycle of a program, thereby ensuring timely handling of any potential civil liberties or privacy issues, if and when they arise. Through this ongoing relationship, the privacy and civil liberties of individuals is well preserved, with careful oversight and responsible consideration given to deciding whether and how to deploy any resulting technologies.

As discussed above, neither KDD nor ALADDIN will be conducting any data mining activities, so no assessment of impact on privacy is required in this context. IARPA is, however, conducting research on techniques to enhance privacy of individuals, which we shall briefly describe.

The 2008 ODNI Data Mining Report included a discussion of the IARPA Automatic Privacy Protection (APP) program, an advanced research effort that seeks to build on the state of the art in technology for private information retrieval to solve scalability problems, broaden its applicability, and expand policy options for information sharing. This 3-year program is now underway and is scheduled for its mid-way evaluation in Spring 2010.

As discussed in last year's Data Mining Report, the goals of the APP program include:

- Development and demonstration of practical and sound automated methods for the use of private information retrieval techniques in IC systems, to automatically protect the private data of untargeted individuals, to assure that mandated policies are enforced, and to enable more effective interagency and intergovernmental data sharing for improved security.

- Demonstration of a system that assures the confidentiality of a searcher's query from a database owner, while simultaneously assuring the database owner that only data relevant to an authorized query will be disclosed to the searcher. Specifically, such a system will permit a client to pose queries to a cooperating database in such a way that the database system cannot infer anything about the query posed or the results returned. At the same time, the database operator can be confident that only information relevant to the (hidden) query is returned, even though the database owner is not himself privy to the query submitted.

- Development of audit measures that permit a third party to evaluate whether queries submitted to the database conform to established privacy policies. The demonstration must exhibit scalability, performance, and assurance levels relevant to IC applications.

The APP program does not seek to enhance a searcher's ability to discover or locate predictive patterns or anomalies of any kind. In fact, by prohibiting queries that fail to conform to privacy policies, the program may in fact diminish a searcher's abilities in this area. Moreover, by preventing the database owner from analyzing records of previous queries (which could, theoretically, allow the database owner to discover or locate predictive patterns or anomalies indicative of the searcher's interests), the APP program seeks to protect the searcher's interests from the database owner.

Taken together, these efforts will benefit the Intelligence Community by enabling systems that protect privacy interests of both searchers and database owners, and thus enhance cooperative information sharing. More generally, the underlying technology for private information retrieval holds the potential for broad application and can expand the policy options available for dealing with information sharing, coalition operations, and international cooperation throughout the IC.

(F) A list and analysis of the laws and regulations that govern the information being or to be collected, reviewed, gathered, analyzed, or used in conjunction with the data mining activity, to the extent applicable in the context of the data mining activity.

Executive Order ("EO") 12333 requires each element of the IC 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 EO 12333.

In addition to EO 12333, personal data retrieved by the name or identifier of a U.S. person must comply with the Privacy Act. In general, the IARPA data sources consist of non-U.S. person intelligence information and incidentally collected U.S. person information retained consistent with EO 12333. Because IARPA does not retrieve information from these data sources using any personal identifiers associated with a U.S. person, IARPA does not maintain a system of records under the Privacy Act for these research purposes.

(G) A thorough discussion of the policies, procedures, and guidelines that are in place or that are to be developed and applied in the use of such data mining activity in order to— (i) protect the privacy and due process rights of individuals, such as redress procedures; and (ii) ensure that only accurate and complete information is collected, reviewed, gathered, analyzed, or used, and guard against any harmful consequences of potential inaccuracies.

Until the research results from the KDD and ALADDIN programs transition into deployable technologies, it is difficult to assess the real and practical impact of the data mining activity on actual privacy and civil liberties interests.

The IC has in place a robust protective infrastructure. It consists of a core set of U.S. person rules derived from EO 12333, as interpreted, applied, and overseen by agency Offices of General Counsel and Offices of Inspectors General, in coordination with CLPO, with violations reported to the Intelligence Oversight Board of the President's Intelligence Advisory Board.

Before any IARPA developed 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 and other applicable laws to determine how the tool could be used consistent with the agency's U.S. person guidelines. As discussed above, these guidelines are extensive. For example, the Department of Defense (DOD) guidelines, which are unclassified, consist of sixty-four pages of detailed procedures and rules governing the intelligence activities of DOD components that affect U.S. persons.

In addition to the above, IARPA has committed to a cutting edge research program focused on developing privacy protecting technologies, specifically, the APP program described more fully above, in Section E.

ODNI recognizes that data mining techniques explored as part of a research program could, potentially, impact the privacy or civil liberties of individuals if the techniques are successfully transitioned to an operational partner without careful consideration of these issues. IARPA

works closely with ODNI CLPO to ensure its research programs are conducted consistent with the protection of individual privacy and civil liberties. During program formulation, the IARPA program manager works closely with CLPO to determine whether there are any privacy or civil liberties issues associated with the program. If any such issues are identified, CLPO works with IARPA to develop mitigation strategies, such as implementing proper data handling techniques. This close relationship continues throughout the lifecycle of the program.

The ODNI Civil Liberties and Privacy Office is headed by the Civil Liberties Protection Officer, a position established by the Intelligence Reform and Terrorism Prevention Act of 2004 (IRTPA). The duties of that officer are set forth in Sections 103D and 1062 of that Act, as amended, 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.

In addition to collaborating with the ODNI CLPO, IARPA also works closely with ODNI's Office of General Counsel (OGC) to ensure that programs comply with all applicable legal requirements, including all constitutional protections and EO 12333, as well as all civil liberties and privacy laws.

Appendix 1: Note on Discontinued Programs

The IARPA Video Analysis and Content Extraction (VACE) program mentioned in last year's Data Mining Report has ended, and is no longer relevant to this report. For completeness, we include the updated status of this program below.

1. The VACE program detailed in last year's ODNI Data Mining report concluded as scheduled in September 2009. The VACE program conducted research in computer vision and machine learning topics such as (a) Object detection, tracking, event detection and understanding, (b) scene classification, recognition and modeling, (c) Intelligent content services such as indexing, video browsing, summarization, content browsing, video mining, and change detection. The program integrated several technologies into proof-of-concept software.

   As noted in last year's Data Mining report and Section C (above), some of the VACE researchers took part in the National Institute of Standards and Technology (NIST)'s TRECVID 2008 evaluations. The results of TRECVID 2008 were published as part of the event's proceedings and may be obtained at the following URL: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2008.

Appendix 2: Noteworthy Programs Under Development/ Significant Modifications to Previously Reported Programs/Platforms

1. The CATALYST program, still being developed by ODNI/CIO (in consultation with ODNI's policy, legal, and CLPO offices), was funded during the reporting period. While

CATALYST is not currently being developed to support pattern-matching functionality, the program could potentially be expanded to support such functionality in the future. We therefore include this program in an Appendix, for informational purposes only.

The volumes of data collected by the IC today, stored and processed under different regimes - different technical systems, access controls, legal authorities, data schema etc. - makes large-scale data sharing and correlation within the IC highly problematic. The CATALYST Program seeks to address this challenge. Specifically, the purpose of CATALYST is to provide the IC with a means for codifying what it already knows, collectively, about entities-of-interest, by enabling the IC to: 1) derive entity data from its collective holdings; 2) apply correlation analytics to the resultant data; and 3) do so in a secure manner that takes into account legal, policy, privacy, and security interests. CATALYST will initially focus on development of data disambiguation capabilities, as well as link analysis and data correlation.

2. During the 2009 reporting period, the Research and Development Experimental Collaborative Network (RDEC) -- reported in the 2007 Data Mining Report -- was replaced by the ODNI/CIO's Single Integrated Test Environment (SITE) program. Specifically, SITE became operational on 28 September 2009, and RDEC was discontinued on 1 October 2009. Although SITE itself is not a reportable program (it is a test and evaluation network), and although SITE was not utilized as a testing platform for a reportable program during the 2009 reporting period, we believe it prudent to keep Congress informed about this platform change from RDEC to SITE.