Office of the Director of National Intelligence

Data Mining Report

The Office of the Director of National Intelligence (ODNI) is pleased to provide to Congress its second 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 January 31, 2008 through January 31, 2009. Constituent elements of the Intelligence Community (IC) are reporting their data mining activities to Congress through their own departments or agencies.

Last year's ODNI data mining report detailed a number of efforts within the "Incisive Analysis" Office in the Intelligence Advanced Research Projects Activity (IARPA) that included research of techniques that could be applied to data mining. Two of those programs (Tangram and Paint) have ended, the relevant effort within a third program (Knowledge Discovery and Dissemination) has ended, and the focus of a fourth effort (Reynard) changed in early 2008. Only one program – Video Analysis and Content Extraction (VACE) – is currently funding research that includes the exploration of techniques that might be applied to data mining. As a result, VACE is the program that this report addresses in detail. Information about the disposition of the other efforts is appended to the end of 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 and on one of the discontinued efforts discussed in the appendix at the end of this report has also been prepared and has been transmitted to the appropriate Congressional committees.

Definition of "data mining." The Data Mining Report Act defines "data mining" as "a program involving pattern-based queries, searches or other analyses of 1 or more electronic databases" in order to "discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity . . . ."[2]

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.

[1] Section 804 of the *Implementing the Recommendations of the 9/11 Commission Act of 2007.*
[2] Section 804(b)(1)(A) of the Data Mining Reporting Act.

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **01 MAR 2009**	2. REPORT TYPE	3. DATES COVERED **00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE **Data Mining Report**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of the Director of National Intelligence,Washington,DC**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

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 is one piece of one research program within the Office of Incisive Analysis that is exploring techniques for identifying patterns that may be associated with terrorist activity, as described below. This report details those activities because the act requires a report on any "activity to . . . develop data mining."[3]

IARPA also has a research program aimed at privacy protection which is also described below.

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: Video Analysis and Content Extraction

(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.

- There are no programs within the Incisive Analysis Office that focus on data mining *per se*, but the *Video Analysis and Content Extraction (VACE)* program is researching some technologies that do meet the reporting criteria of an "activity to . . . develop data mining" under the Act. The VACE program seeks to automate what is now a very tedious, generally manual, process of reviewing video for content that is potentially of intelligence value. In general, VACE involves subject-based queries of video databases that do not meet the definition of data mining, but some aspects of the program involve the development of technologies that could possibly be applied to pattern-based data mining.

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

- VACE conducts 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.

- Potential applications of these techniques to pattern-based data-mining could include automated processing of surveillance camera data to determine anomalous behavior, and searching video databases, such as broadcast news archives, to retrieve events such as bombings or beheadings where the query was not subject-based or seeded with a personal identifier.

The VACE program is slated to end in Fall 2009.

(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 programs within the Incisive Analysis Office that focus on data mining *per se*, as explained in Section A above, the VACE program is researching some technologies that do meet the reporting criteria of an "activity to . . . develop data mining" under the Act. Researchers in the VACE program that are developing techniques that could be applied to data mining activities have articulated sound reasons why they believe their technological approaches could ultimately be successful in the act of identifying well established patterns of clearly suspicious behavior in the data. Please refer to Section A for a description of the VACE program and Section D for a description of the approach to validating the research techniques.

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

The video data used by the VACE program consists of lawfully collected data from public places outside the United States and information from public media sources. Additional sources used for testing are the National Institute of Standards and Technology (NIST) Video Retrieval (TRECVID) data, which are simulated video content created through use of volunteers who agreed to appear in such video content specifically for research purposes, consistent with applicable guidelines.[4]

[4] The aspects of the other projects (now ended) that involved data mining and that are described in the appendix used simulated data for testing purposes and one project used lawfully collected foreign data that is described in more detail in the classified annex.

(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.

Researchers in the VACE program have articulated sound reasons for believing that their approaches will be effective in achieving their stated goals. Many of them, during the course of their work, have already taken part in the National Institute of Standards and Technology's TRECVID (the video retrieval evaluation series of NIST's Text REtrieval Conference), which promotes progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation. TRECVID is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations.

TRECVID 2008 tested systems on high-level feature extraction, search, content-based copy detection, and surveillance event detection. Proceedings from these tests will soon be published.

(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 intends to maintain its 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. Through this ongoing relationship, the privacy and civil liberties of individuals will be well preserved with careful oversight and responsible consideration in the decision whether and how to deploy any resulting technologies.

In the fall of 2006, NSA's Disruptive Technology Office (later incorporated into IARPA) and the ODNI CLPO jointly sponsored a series of workshops attended by government experts, private sector experts, and privacy advocates. The attendees at these workshops examined an array of challenges to privacy posed by emerging technologies and government needs for information for intelligence and counterterrorism purposes, and suggested a variety of innovative approaches to applying technology to these problems.

The 2008 ODNI Data Mining Report described a broad range of issues and technologies related to privacy protection inspired by these workshops. During the past year, IARPA evaluated research proposals to explore many of those technologies and address many issues. These proposals and evaluations led to the creation of the IARPA Automatic Privacy Protection (APP) program, which has the following goals:

- Develop and demonstrate practical, 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.

- Specifically, demonstrate a system that permits a client to pose queries to a cooperating database in a manner that the database system cannot in practice infer anything about the query posed or the results returned, but at the same time the database operator can know that only information relevant to the (hidden) query is being returned. Further, demonstrate measures that can be evaluated by a third party to assure that queries submitted to the database conform to established privacy policies. The demonstration must exhibit scalability, performance, and assurance levels relevant to IC applications.

The technologies to be developed in this program are in the areas of Private Information Retrieval, discussed in last year's report, and in the automated analysis of query logs to assure that queries conform to specified policies.

Successful demonstration of these technologies will show that a structured database can be searched for information about a specific topic or individual while assuring that information about other individuals whose information resides in the same database is not returned. Further, the cooperating database owner will not be able to learn anything about the nature of the query posed or the information returned. Current practice for enforcing privacy protection depends largely on manual procedures that are fundamentally unscalable and require placing substantial trust either in humans or large software systems. 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.

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. For example, information that is publicly available or that constitutes foreign intelligence or counterintelligence may be collected, retained or disseminated.

The video data used by the VACE program consists of lawfully collected data from public places outside the United States and from volunteers who have agreed to participate specifically for research purposes in accordance with applicable guidelines. As a result, its use is consistent with 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. However, the IARPA data sources generally consist of non-U.S. person intelligence information and incidentally collected U.S. person information is only 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 VACE program 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, overseen by agency Offices of General Counsel and Offices of Inspector General, 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 law 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, IARPA has committed to a cutting edge research program focused on developing privacy protecting technologies, described in detail above.

The ODNI's Civil Liberties and Privacy Office is working closely with IARPA to develop civil liberties and privacy guidelines for research programs in general, and IARPA will continue this close working relationship to develop additional program-specific policy, privacy and civil liberties guidance if needed in the future. The 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, ensuring that technologies sustain, and do not erode, privacy.

Appendix: Note on Discontinued Programs

Several of the IARPA Incisive Analysis programs mentioned in the 2008 Data Mining Report have been either reformulated or ended, and are no longer relevant to this report. For completeness, we include the updated status of those programs.

1. *Reynard* continues as a seedling effort[5] within IARPA; however, the focus of the effort has changed since the 2008 data mining report was submitted. Reynard is currently exploring the feasibility of understanding and characterizing behavior in virtual worlds by leveraging expertise in the social science research community. As such, the planned program no longer meets the definition of "pattern-based data mining" described in the Act.

2. The *Tangram* program was originally intended to evaluate the efficacy and intelligence value of a terrorist threat surveillance and warning system concept that would (i) report the threat likelihood of known threat entities, and (ii) serve to discover and report the threat likelihood of unexpected threat entities. During FY 2008, the Tangram program conducted elementary experiments on the feasibility of building and maintaining a continuously operating surveillance and warning system from a computer science perspective. The program has ended, the results will be archived, and further research is not planned at this time.

3. The *ProActive Intelligence (PAINT)* program sought to study the dynamics of complex intelligence targets (inclusive of but not solely terrorist organizations) by using a model-based approach to elucidate patterns of causal relationships that are indicative of nefarious activity. The program, which concluded in early January 2009, integrated several modeling technologies in a first-generation proof-of-concept system.

4. *Knowledge Discovery and Dissemination (KDD)*- The goal of the KDD program is to invest in research and technology that will greatly enhance the ability of analysts to collaboratively evaluate and utilize data from multiple, massive data sets in order to generate high quality, accurate, and timely intelligence. The research of KDD is primarily focused on link analysis and graphs as well as techniques to improve and measure collaboration between IC analysts. As such, KDD rarely supports research using pattern based data mining techniques as defined in the Data Mining Reporting Act.

    - In FY06-FY07, there was a KDD sponsored research project that met the reporting criteria of an "activity to . . . develop data mining" in the Act. The project attempted to match known patterns of entity deception in lawfully collected foreign data bases. This project was completed in early 2008 and was included in last year's report. With the completion of this project in 2008, there are currently no research projects supported by KDD that meet the reporting criteria of the Data Mining Reporting Act.

    - BLACKBOOK- the BLACKBOOK capability was developed under the KDD program to provide an infrastructure using a service oriented architecture (SOA) approach for data analysis. The infrastructure does not do any data analysis or

[5] A seedling effort allows an IARPA program manager time and a small amount of money to conduct due diligence on potential research ideas before a commitment to a full IARPA program is made.

data mining per se but rather provides a convenient and organized way for other services[6] to be run that do data analysis.

[6] Services – Typical services include graphical analysis and manipulation, workflow logging, analyst annotation, searching data bases for specific words, names, etc., and formatting data.