

**Office of the Director of National Intelligence****Data Mining Report**

15 February 2008

The Office of the Director of National Intelligence (ODNI) is pleased to provide to the Congress 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"). The Data Mining Reporting Act requires "the head of each department or agency of the Federal Government" that is engaged in activities defined as "data mining" to report on such activities to the Congress.

Scope. This report covers the data mining activities of all elements of the ODNI. Constituent elements of the Intelligence Community are reporting their data mining activities through their own departments or agencies. This report covering ODNI activities is unclassified and has been made available to the public through the ODNI's website. A classified annex has also been prepared and has been transmitted to the appropriate Congressional committees.

Other Intelligence Community elements. The ODNI's Civil Liberties and Privacy Office (CLPO) has requested that civil liberties and privacy officers within the Intelligence Community disclose their data mining activities to the ODNI and provide copies of any reports submitted in response to the Data Mining Reporting Act.

Definition of "data mining." The Data Mining Reporting 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 . . . ."

The limitation to predictive, "pattern-based" data mining 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. These 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 Data Mining Activities**

The ODNI's Office of Science and Technology's Intelligence Advanced Research Projects Activity (IARPA<sup>1</sup>) has a portfolio of research projects, some of which include the exploration of techniques that could be applied to data mining. These projects are research projects and their activities are being conducted for research purposes. They have not been deployed for use in any operational or other real life environments.

Results from such research projects may in the future be incorporated into operational programs employing data mining technology within the ODNI, the Intelligence Community, or other parts of the United States government, subject to appropriate legal, privacy, civil liberties and policy safeguards. Because of the potential privacy and civil liberties impact of the development of these technologies, IARPA is also specifically researching the use of technology to better protect privacy in light of these challenges.

### **Overview of Incisive Analysis Area**

The Intelligence Advanced Research Projects Activity (IARPA) funds cutting edge research to address difficult and complex intelligence community challenges. Today's intelligence analysts must wade through an exponentially increasing amount of data (both classified and open source) to uncover potentially key nuggets of valuable information in time for them to be transformed into timely, actionable intelligence.

IARPA's Incisive Analysis efforts are comprised of research projects designed to address this challenge by harnessing advanced analytic tools to aid the human analyst in looking through large volumes of data to uncover the information that is most relevant in a timely fashion. Some of these projects include elements that meet the Data Mining Reporting Act's definition of data mining.

IARPA activities are by nature highly experimental and pioneering and are designed to produce new capabilities not even imagined by the operational agencies it serves. The typical context behind an IARPA project involves the demonstration of a never-before-seen capability, the establishment of relevant technical metrics for this capability, a baseline comparison (if possible) with existing legacy solution(s), and the tracking of such performance metrics throughout the life of the project. Because IARPA pursues high-risk, high-payoff solutions, its projects do not necessarily result in deployable technologies. When they do, additional steps are needed to transform the results of IARPA research into real world applications, which may be different from what was originally envisioned.

---

<sup>1</sup> IARPA invests in cutting-edge research projects that have the potential to result in revolutionary, game-changing capabilities for the IC. Should such advanced research projects prove feasible, they may subsequently be transitioned into settings involving operational use.

As a result, the activities within IARPA's Incisive Analysis projects that meet the "data mining" definition of the Data Mining Reporting Act will generally not be at a level of technological readiness that permits an accurate judgment as to their efficacy and generally will not have an existing basis for determining that a pattern or anomaly is indicative of terrorist activity. Indeed, the very purpose of the Incisive Analysis projects, as with all of IARPA's research, is to answer these questions in order to determine whether such technology is promising enough to warrant additional investment to develop tools that could be deployed within the Intelligence Community.

### **Incisive Analysis Projects: Detailed Response**

The Data Mining Reporting Act requires detailed information regarding data mining activities. Some information is classified and a classified annex to this report has been prepared and made available to appropriate Congressional committees.

**(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.**

In order to deal with the challenge of addressing the exponential growth of information faced by the intelligence analysts, IARPA has created the Incisive Analysis portfolio, an ensemble of advanced research projects dedicated to achieving the DNI goals of creating a culture of collaboration, fostering collection and analytic transformation, and accelerating information sharing.

The Incisive Analysis portfolio does not focus on data mining *per se*, but certain projects within that portfolio are researching technologies that do meet the definition of data mining. These are:

- *Knowledge Discovery and Dissemination (KDD)*. KDD is seeking ways to coordinate access to and effectively exploit multiple, lawfully-collected data sources across the disparate intelligence community agencies.
  - KDD research is not operationally engaged in discovering patterns of behavior in data that are indicative of criminal or terrorist groups. At a pure research level, one effort attempts to match known patterns of deception as provided by subject matter experts in foreign intelligence data.
  - A few of the tools being developed by KDD may be used in the future for the conduct of data mining as defined by the Act. These include network tomography, predictive analysis, and hypothesis generation and validation tools.
  - Some of the KDD tools are installed in a number of Intelligence Community Science and Technology (IC S&T) offices for testing and

evaluation. Tools may be transitioned for operational use in accordance with KDD's plan to develop alliances with other IC S&T programs within the next year. Any transition resulting in operational use will coordinate with the civil liberties and privacy protection mechanisms of the receiving agency prior to implementation.

- KDD is planning additional collaboration with the Department of Homeland Security (DHS) and law enforcement organizations by using a test and evaluation center to test analytic tools.
- The *Tangram* project is intended to evaluate the efficacy and intelligence value of a terrorism threat surveillance and warning system concept. Tangram explores the viability of a "surveillance and warning" system that will (i) report the threat likelihood of *known* threat entities, and (ii) serve to discover and report the threat likelihood of *unexpected* threat entities.
  - The experimental methodology includes continuously assessing the information we have about known threat entities. This assessment function would not necessarily involve data mining because known entities would be the subjects of the assessment. However, Tangram is evaluating the viability of pattern-based detection methods that could reveal a change in the threat likelihood of a known individual. We envision three areas where data mining techniques could potential improve Tangram performance: (a) overcoming common data problems, such as sparseness, incompleteness or incorrectness; (b) assessing multiple entity threat hypotheses; and (c) providing warning of unexpected threat entities. Areas (a) and (b) are scientifically proven techniques, while area (c) is an important research area of the project.
- The *Video Analysis and Content Extraction (VACE)* project seeks to automate what is now a very tedious, generally human-powered process of reviewing video for content that is potentially of intelligence value. In general, VACE will involve subject-based queries of video databases that do not meet the definition of data mining. However, two aspects of the VACE program involve possible use of pattern-based data mining technologies.
  - 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.
  - Application of these techniques to pattern-based problems includes (1) an effort seeking to automate processing of surveillance cameras, such as might be found in public transit terminals, to determine anomalous behavior, and (2) an effort that searches video databases, such as broadcast news video archives, to retrieve events such as bombings or beheadings

where the query was not subject-based or seeded with a personal identifier.

- The *ProActive Intelligence (PAINT)* project seeks to study the dynamics of complex intelligence targets (inclusive of terrorist organizations) by examining patterns of causal relationships that are indicative of nefarious activity.
  - PAINT does not specifically aim to uncover patterns or anomalies suggestive of terrorist or criminal activities directly; therefore it is not related to the definition of data mining under the Act. However, future applications may have a tangential connection, so for completeness, it has been included.
- *Reynard* is a seedling effort to study the emerging phenomenon of social (particularly terrorist) dynamics in virtual worlds and large-scale online games and their implications for the Intelligence Community.
  - The cultural and behavioral norms of virtual worlds and gaming are generally unstudied. Therefore, Reynard will seek to identify the emerging social, behavioral and cultural norms in virtual worlds and gaming environments. The project would then apply the lessons learned to determine the feasibility of automatically detecting suspicious behavior and actions in the virtual world.
  - If it shows early promise, this small seedling effort may increase its scope to a full project.

Because application of results from these research projects may ultimately have implications for privacy and civil liberties, IARPA is also investing in projects that develop privacy protecting technologies (*cf.*, Section E - *Focus Area: Privacy Protecting Technologies*)

**(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.**

IARPA conducts advanced research projects that explore new concepts and technologies. Although researchers in each of the Incisive Analysis projects that involve data mining activities have articulated sound reasons why they believe their technological approaches could be successful in identifying patterns or anomalies that could be useful to the Intelligence Community in discovering terrorist, criminal or other activities of interest, they usually do not have a specific, documented basis for determining whether a particular pattern or anomaly is indicative of specific activity (*e.g.*, terrorism, criminal acts, *etc.*) In fact, one of the goals of these projects, as with all IARPA programs, is to create a basis for quantitative measurements.

- *Knowledge Discovery and Dissemination (KDD)*.

- Most tools that researchers are developing in the KDD program do not involve pattern-based data mining. However, some tools have been developed to discover patterns associated with deceptive behavior in groups using an analytic technique called network tomography. This tool looks for deception patterns in large databases. Other tools have been designed for predictive analysis, attempting to identify the next step in an emerging pattern, and hypothesis generation, seeking to provide possible explanations for observed anomalous patterns.
- *Tangram* is seeking to demonstrate the feasibility and intelligence value of a semi-autonomous terrorist threat assessment system concept. Its most immediate objective is to assess the threat likelihood of known threat entities. The simplest of methods would be initiated by a search for information about the specific entity. However, a surveillance and warning system must also provide warnings where 1) the data are sparse, incomplete or erroneous, and 2) the threats are assessed across multiple lines of inquiry that individually would not reveal an entity's threat likelihood. Pattern-based data mining methods have proven effective at compensating for common data issues and fusing multi-sensor data to produce warnings. *Tangram* hopes to capitalize on these methods to improve its overall intelligence value as a function of true positives and false positives. Under these conditions *Tangram* will evaluate the efficacy of data mining methods.
  - *Tangram* will take advantage of research from IARPA's Privacy Protecting Technologies and KDD projects in addition to the tools that have been tested on the Research and Development Experimental Collaboration (RDEC) testing platform.<sup>2</sup>
  - A significant aspect of *Tangram*'s research is discovering highly reliable threat patterns and statistics that will provide reliable warnings. Consequently, *Tangram* has no pre-existing patterns that have been applied to real intelligence data.
  - Because most of *Tangram*'s computational methods have been successfully used for niche applications within the Intelligence Community and in commercial applications, *Tangram* researchers believe there is a proven basis for continuing to explore whether aggregations of subject-based and pattern-based methods could provide highly reliable warnings..
- *Video Analysis and Content Extraction (VACE)* seeks to dramatically increase analyst efficiency in processing video content. While VACE is not a data mining project *per se*, two aspects of VACE could involve pattern-based searches of video content for indications of possible terrorist or other criminal activity.

---

<sup>2</sup> RDEC is a testing environment that contains the data sources used by these and other projects, and is described in more detail below in response to the Data Mining Reporting Act's question regarding data sources.

- VACE is developing advanced video searching capabilities that could involve looking for particular patterns that might indicate a broadcast of terrorist events (*e.g.*, bombings, beheadings).
- VACE has developed a Video Event Manager that permits analysts to find a particular event within video, such as an event that has possible security significance – for example, a person is observed entering a restricted area or leaves a bag in a public place.
- *ProActive Intelligence (PAINT)* studies the dynamics of complex intelligence targets, such as terrorist organizations, and employs models of causal relationships that are designed to increase analyst efficiency.
  - As noted in section (A), PAINT does not specifically aim to uncover patterns or anomalies suggestive of terrorist or criminal activities directly; therefore it is not related to the definition of data mining under the Act. However, future applications may have a tangential connection, so for completeness, it has been included.
  - *Reynard* is a small and highly exploratory seedling project which seeks to identify the emerging social, behavioral, and cultural norms in virtual worlds and gaming environments. If successful, Reynard may be expanded to a full-scale project.

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

Most of the projects within the Incisive Analysis area that involve data mining make use of the unique testing and evaluation capabilities and structured databases of the Research and Development Experimental Collaboration (RDEC) network. The RDEC network provides access to data from a number of classified databases containing lawfully collected foreign intelligence information. They have been copied and selected for inclusion within the RDEC network to permit testing and evaluation of new and promising analytical tools without any danger that the tools would damage or corrupt the data.

A listing of these databases is found in the classified appendix to this report.

KDD relies on research partners who develop analytic tools in accordance with research protocols that do not make data available in bulk form to IARPA. IARPA requires its partners to take steps to minimize the information provided in research results, even though many of these data sets are publicly available. This may require, for example, that partners report their findings in aggregate form (*e.g.*, without reporting any personally identifying information). Within the Intelligence Community, validated KDD tools are being tested with data sources associated with the counter-terrorism and counter-WMD missions to ensure robustness and ease of use.

All data sources used by the Tangram program since its inception have been synthetic, *i.e.*, fabricated simulations of real intelligence data. Tangram researchers anticipate using RDEC data in the future.

The video data used by the VACE program consists of lawfully collected data from public places outside the United States. Additional data sources used for testing are the National Institute of Standards and Technology (NIST) Video Retrieval (TRECVID) data, which are simulated video content created by volunteers specifically for research purposes.

PAINT uses lawfully collected foreign intelligence information for research purposes.

Reynard will conduct unclassified research in a public virtual world environment. The research will use publicly available data and will begin with observational studies to establish baseline normative behaviors.

**(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 have sound reasons for believing that their approaches have the potential to develop real world applications that will be effective in achieving their stated goals. Because Incisive Analysis programs are ongoing research, the projects are largely designed precisely to determine whether and how effective each of the analytical tools, including pattern-based tools, may be.

After a measurement system is developed based upon the capability in question, each data mining-related activity within IARPA will be compared with the systems in use by the Intelligence Community today to determine whether the capabilities developed by IARPA provide faster and more accurate information than traditional approaches. Processes will be carefully measured throughout the life of each project.

Even as researchers study how effective their proposed approaches are, they will coordinate with the IARPA's privacy protecting technologies project. This coordination will help researchers to determine how to incorporate additional privacy protecting technologies within the analytic tools they are developing. Coordination with the privacy protecting technologies project will enable researchers to assess whether such tools can be incorporated into their projects while preserving or even enhancing the efficacy of those tools in achieving their mission of enhancing intelligence analysis.

**(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.**

The data mining activities that are part of the research projects within IARPA's Incisive Analysis portfolio 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. IARPA's privacy protecting technologies initiative<sup>3</sup>, and the related work being sponsored by some individual projects, will ensure that the technologies developed by IARPA to serve national security ends do so through means that are both constitutionally valid and privacy enabling.

IARPA's privacy protecting technologies initiative is based in part on a unique collaboration of government experts, private sector experts, and privacy advocates at a series of workshops jointly sponsored by IARPA and the ODNI Civil Liberties and Privacy Office (CLPO) in the fall of 2006. These workshops examined an array of challenges to privacy posed by emerging technologies and government needs for information for intelligence and counterterrorism purposes. Experts suggested a variety of path breaking and innovative approaches to applying technology to these problems, and IARPA developed the privacy protecting technologies initiative to jump start research in the most promising areas.

In short, IARPA believes that the privacy and civil liberties of individuals will be well preserved with careful oversight of the projects and responsible consideration of privacy and civil liberties in the decision whether and how to deploy any resulting technologies. IARPA intends on maintaining its long-term relationship with the ODNI CLPO for the purpose of validating that its research agenda is consistent with the protection of individual privacy and civil liberties.

*Focus Area: Privacy Protecting Technologies*

This IARPA focus area seeks innovative technologies that can advance both the security and privacy of information collected, processed, and held for intelligence community purposes. In many cases, these two properties are well aligned. Good privacy policy calls for limiting use or retention of data beyond the purposes for which it is collected. This policy is also good for security, in that it limits what can be compromised. Technologies that promote the responsible handling of private data can also help build public confidence in the process.

Accountability, privacy, information sharing, and the technologies that promote them often overlap and display complex interrelationships in the private sector. For example, many enterprises collect information that can be used to identify individuals (personally identifiable information, or PII) in order to be able to conduct business with their customers. The customer must be willing to share some information (PII, in particular) with the enterprise in order to obtain desired services. As the enterprise wants to hold the customer accountable for the cost of services it provides, the customer wants to hold the enterprise accountable for providing the service and also for protecting the private information the enterprise holds. The enterprise may provide this accountability to the customer by using privacy protecting technologies, including technologies to assure that

---

<sup>3</sup> A unique research effort within the intelligence community.