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 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 ...."

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[1]) 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.

---

[1] IARPA invests in cutting-edge research projects that have the potential to result in revolutionary, game-changing capabilities for the Ie. 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 I) 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.[2]

  - 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 warnmgs..

- *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.

---

2 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 willenable 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[3], 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

3 A unique research effort within the intelligence community.

private information flows only where policy permits. The issues grow more complex when one enterprise deals with another either on behalf of a customer or on its own account. When the enterprise outsources or subcontracts, for example, the enterprise must assure that any PH conveyed to the subcontractor continues to be handled according to the applicable policies.

The data collected and held by the intelligence community and used to alert national leaders and to understand global affairs must be handled in accordance with many governing laws, regulations, policies and procedures that continue to evolve in response to public concerns both about privacy preservation and national security. Technologies are needed that can help assure that publicly agreed-upon policies are applied effectively and routinely. This assurance needs to be in place in order to facilitate legitimate, controlled sharing of information among the government organizations, including federal, state, local, and tribal entities, and also sharing with other countries and governments. Since any mechanism may be abused by authorized users, the technology also needs to resist insider threats and to help establish accountability for violations.

IARPA is evaluating proposals for research in these general areas of concern:

- *Accuracy.* The ability to review records and enter corrections or at least contest the entries is a conventional component of fair information practice. Like the data in a credit report, intelligence data can be incorrect. Unlike credit data, however, it may be impractical or infeasible to permit the subject of the intelligence data to review it and enter corrections. For example, these cases include mistaken identification of an individual with a person who is on a watch list, or cases where an individual is correctly identified but has been mistakenly entered into a database or placed on a watch list. Not only facts, but the deductions and analyses based upon them, are at issue. One might imagine a system which, once an error of fact is detected, could more easily identify where that fact had been used and flag or roll back the reports needing correction. Methods for extracting meaning from content are not the concern here; the focus is on how to retract erroneous data and the inferences based on those erroneous data. Can new technologies help solve these problems without revealing sensitive information about sources and methods?

- *Access.* Just as sensitive corporate information may be revealed to litigants as part of the discovery process when a lawsuit is underway, information about an individual that would ordinarily be private may legitimately be disclosed to investigators if the individual is the subject of an authorized investigation. If that information is in a database with the information of many other individuals, it may be inappropriate to (a) search all of the other data and (b) to reveal the search criteria (i.e., aliases for the individual or information that distinguishes the individual from other individuals with similar identifiers) to the holder of the database.

- *Accountability.* Accountability is a key tool to help assure that those entrusted with special privileges for data access use those privileges responsibly and also to assure that they can be credited with doing so. In some cases, a key part of the control regime may involve assuring that those accessing data can be held accountable for complying with the applicable rules and restrictions. Effective accountability measures can serve as a crucial deterrent to misbehavior. Audit controls can be effective in this context; completeness and integrity of the audit trail will be primary concerns.

Technology areas of particular interest include (but are not limited to) the following:

- *Secure multi-party function evaluation.* While the mathematics of this technology has been studied for some time, practical applications have been lacking. Projects that can demonstrate how this technology could be applied to problems of realistic scale and complexity will be of interest. For example, agencies at different levels of the U.S. government, as well as selected foreign government and private sector entities, are all interested in comparing intelligence information concerning terrorist financing, yet these entities may be unwilling or unable to disclose their own detailed information for fear of violating privacy rules or compromising sources and methods. Secure multi-party function evaluation might provide a way for such entities to cooperate in computing the results regarding such financial flows without either sharing the information with each other or resorting to a trusted third party to compute it for them.

- *Entity disambiguation methods.* Innovative techniques that improve the precision with which one individual (perhaps with the same name or other identifier) can be distinguished from another and can be correctly associated with some set of data are of high interest. For example, during the course of an individual's security adjudication process, an investigator finds records containing names that are variations of the subject's name *(e.g.*, John, Jon, Johnny, Jack). The adjudicator must ensure that all records pertaining to the subject are examined while records containing the same name variants, but unrelated to the subject, are excluded.

- *Applications of non-monotonic logic.* Applications of non-monotonic logic to databases of significant size and complexity may be useful. For example, such applications could aid in locating and purging data that -- some time after collection -- are determined to be incorrect. Further, techniques that can automatically locate (and possibly retract) information or reports inferred or derived from incorrect data will be of interest. For example, a source deemed credible is later determined to be unreliable. Enough time has passed that intelligence from this source has been used in numerous analyses and cited in many reports and briefings. Technologies have been developed that can link documents, and even portions of documents, within a limited and controlled context (for example a single unified system using a common document processing system). A challenge is to use the power of non-monotonic logic to

facilitate the automatic or semi-automatic retraction of inferences that have been made across a wide range of document types and formats.

- *Rules-based access enforcement.* Many rules, including rules designed to protect the privacy of U.S. persons, govern the accesses that a member of the Intelligence Community may make to various databases. Enforcement of these rules by requiring prior human review of every record before granting some level of access may be impracticable, but auditing of accesses can be used to complement enforcement. Automated checking of access rules against the audited data can be used to highlight potential abuses. Techniques for effectively checking realistic access sets against realistic rule bases will be of interest. For example, consider a financial records system that normally enforces specific access rules for access to customer data, but also supports special emergency access to sensitive financial data by law enforcement or intelligence personnel. Automated validation of the emergency access patterns after the fact could provide a strong deterrent for abuse.

- *Private information retrieval (PIR).* The ability to query a database without revealing to the operator of the database the search criteria or the data retrieved can prevent a database operator from making potentially injurious inferences about the subject of the search terms. PIR has been much researched but so far has not been reduced to practice. Techniques that hold promise for making PIR applicable in a realistic context will be of interest. For example, suppose a mining company wishes to outsource the storage of voluminous mineral exploration data. The company also wishes to make queries against the data to look for promising areas to develop, yet the queries themselves can reveal the geographic areas of interest, and this information would be valuable to a competitor. So the company is only willing to outsource the storage to a firm that can guarantee the queries will not be revealed. The outsourcing company could potentially use PIR as a means of assuring its customers that it will not (cannot) compromise the data or the queries. Can private information retrieval technologies scale to deal with a global database of images?

- *Anonymous matching.* This technique, sometimes called *private matching,* and considered a special case of secure multiparty function evaluation, permits comparisons of multiple sets of data under the control of different parties while revealing selected data only (or no data at all} to the parties involved. It can protect privacy by shielding the identities of persons in the searched database from human review until, with a high probability, they match specific criteria. At this point, and with appropriate legal authority, some or all related data might be individually reviewed. Some products now incorporate techniques for exact matching, but matching that supports a "closeness" metric rather than a simple Boolean match is a research topic at present. Techniques that significantly advance the state of the art in this area would be of interest. For example, watch lists for suspected terrorists are clearly sensitive information, so governments are unwilling to release them publicly. The airline companies are similarly reluctant

to provide unfettered access to their databases of passenger lists to governments. Effective techniques for anonymous matching could support the interests of both parties.

Assessing the effectiveness of a particular privacy protecting technology in a particular context can be a difficult problem in itself. Research that supports evaluation of the strength of a particular mechanism - how easily it may be bypassed or under what conditions it may leak information - will be of particular interest.

(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.**

Most of the information contained in the RDEC network and other lawfully collected intelligence information used by IARPA for test purposes consists of information about non-United States persons, *i.e.*, foreign citizens who are not permanent United States residents. Any information about U.S. persons contained within intelligence databases is governed by Executive Order 12,333, "United States Intelligence Activities," which governs the collection, retention and dissemination of information by the Intelligence Community that may include U.S. person information.

E.O. 12,333 requires each element of the Intelligence Community 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 E.O. 12,333. For example, information that is publicly available or that constitutes foreign intelligence or counterintelligence may be collected, retained or disseminated.

The RDEC network databases generally consist of information collected and retained by elements of the Intelligence Community within the Department of Defense. As a result, the information within the RDEC network complies with Department of Defense regulation 5240.1, which is unclassified and governs all DOD intelligence activities. Implementing regulations and directives serve to protect the privacy of U.S. persons whose information may have been incidentally obtained in the course of foreign intelligence collection. For example, if the identity of a U.S. person is not needed to assess or understand foreign intelligence, that identity will be replaced with a generic term such as "U.S. person."

In addition to classified data sources, IARPA projects make some use of publicly available information. As described above, such information may be collected, retained or disseminated under E.O. 12,333 for lawful purposes related to IARPA's mission. Even in the case of such information, however, IARPA researchers have carefully coordinated their activities with the ODNI's Civil Liberties and Privacy Office to assess whether to use open source information that may pose real or perceived privacy or civil liberties risks.

The video data used by the VACE program consists of lawfully collected data from public places outside the United States and from volunteers specifically for research purposes. As a result, its use is consistent with E.O. 12,333.

In addition to E.O. 12,333, 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 E.O. 12,333. 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 projects result in deployable technologies, it is difficult to assess the real and practical impact of the data mining activity on actual privacy and civil liberties interests. While one could imagine possible implications of successful research in each of the Incisive Analysis projects, the actual impact could be quite different depending on what technology actually results from the research and the context in which a given technology is deployed.

The intelligence community has in place a robust protective infrastructure. It consists of a core set of U.S. person rules derived from E.O. 12,333, 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 Foreign Intelligence Advisory Board.

Before any technology that might be developed by IARPA could be used in an operational setting, the use of the analytic tool would need to be examined pursuant to E.O. 12,333 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, because IARPA anticipates that capabilities derived from some of its research efforts could potentially have some impact on privacy or civil liberties, it has undertaken a separate, but related, research effort into privacy protecting technologies, described in detail above.

Tailored policy, privacy and civil liberties guidance has also been developed to address specific problems posed by particular IARPA projects in coordination with ODNI's

Privacy and Civil Liberties Office. For example, KDD has developed policy guidance, which is awaiting review and approval, to govern its planned joint activities with law enforcement and the Department of Homeland Security. Likewise, the Tangram project has implemented specific restrictions on the data sources to which its researchers have access, in response to guidance from CLPO.

IARPA will continue its close working relationship with the ODNI Civil Liberties and Privacy Office to develop additional policy, privacy and civil liberties guidance as needed. 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), and reports directly to the DNI.

The goal of the CLPO is to help the intelligence community accomplish its national security mission in a way that remains true to the Constitution and protects privacy and civil liberties. Under the IRTPA, the CLPO's duties include ensuring that the protection of civil liberties and privacy is appropriately incorporated in the policies of the ODNI and the intelligence community, 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.

Perhaps most relevant to IARPA, the CLPO is also charged with ensuring that technologies sustain, and do not erode, privacy. As the intelligence community seeks to use new tools, techniques, and approaches to keep the country safe, so too the ODNI must develop and use new tools, techniques, and approaches to protect civil liberties and privacy.

The CLPO works through the intelligence community's civil liberties protection infrastructure -- and recommends improvements where needed -- to ensure that privacy and civil liberties issues are identified and addressed as early as feasible, and that safeguards are formulated and implemented to protect privacy and civil liberties. The CLPO enhances, and does not replace, the civil liberties protection functions of existing offices and mechanisms, and seeks to ensure that new protective functions and mechanisms -- such as civil liberties positions as they are created in other agencies, and the Privacy and Civil Liberties Oversight Board -- interact effectively with the intelligence community to further strengthen privacy and civil liberties safeguards.