Office of the Director of National Intelligence

Data Mining Report for Calendar Year 2013

Office of the Director of National Intelligence
Data Mining Report for Calendar Year 2013

I. Introduction

The Office of the Director of National Intelligence (ODNI) provides this report pursuant to Section 804 of the *Implementing Recommendations of the 9/11 Commission Act of 2007*, entitled *The Federal Agency Data Mining Reporting Act of 2007* (Public Law 110–53).

A. Scope

This report covers the activities of all ODNI components from January 1, 2013 through December 31, 2013. Other elements of the Intelligence Community (IC) are reporting their activities to Congress through their own departments or agencies.

B. Reporting Requirement

The Federal Agency Data Mining Reporting Act of 2007 (hereafter referred to as the "Data Mining Reporting Act") requires departments and agencies of the Federal Government engaged in data mining activities to submit an annual report to Congress. Under this law, "data mining" is defined as:

"… a program involving pattern–based queries, searches or other analyses of one or more electronic databases, where —

(A) a department or agency of the Federal Government, or a non–Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals;

(B) the queries, searches, or other analyses are not subject–based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases;[1] 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.[2]"

[1] As stated in prior reports, certain analytic tools and techniques, such as link-analysis tools, rely on "personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals," such as a known or suspected terrorist, or other subject of foreign intelligence interest, and use various methods to uncover links or relationships between the known subject and potential associates or other persons with whom that subject has a "link" (a contact or relationship). Such tools and techniques are not considered to meet the "data mining" definition of the Act.

[2] Section 804(b)(1)(A) of Public Law 110-53.

C. Report Content

We continue to look for opportunities to rework the format of this report for clarity and readability. In prior reports, we have used Part II to describe those ODNI programs, if any, that meet reporting requirements of the Data Mining Reporting Act, and Part III to describe other programs in the interest of transparency. This year, consistent with the emphasis on transparency in the IC, we will use Part II to more broadly describe programs in the interest of transparency, which may only meet some of the criteria defining "data mining." We will use Part III of this report to provide updates on programs included in the prior year's report. We have added a new Part IV of this report, to provide an overview of the Privacy and Civil Liberties infrastructure within which ODNI conducts it activities.

II. Newly Reported Activities

This section describes activities that are responsive to the Data Mining Reporting Act, and errs on the side of reporting activities in the interest of transparency. This report includes one newly–reported activity, involving an analytic technique used by ODNI's National Counterterrorism Center (NCTC) to narrow the pool of information within NCTC databases that analysts will assess in response to specific threat reports. As noted above, this technique does not meet all of the statutorily–defined criteria for data mining under the Act.

A. NCTC and Threat Reporting.

As the Federal Government's lead in providing the counterterrorism community with 24/7 counterterrorism intelligence monitoring, assessments and situational awareness, NCTC receives intelligence reports relating to terrorism threats, which it analyzes in order to develop lead information for operational partners in the counterterrorism community. To support that effort, NCTC also has access to other agencies' datasets pursuant to applicable laws, executive orders, guidelines, and policies. In analyzing such threat reporting and government data, NCTC analysts employ analytic techniques tailored to the level of detail known to the government about the threat.

In situations where the government is aware of specifics about a threat, the corresponding threat reports received by NCTC will likely include details about the individuals, relevant dates and locations, modes and routes of travel, etc. NCTC uses those details to identify specific information about the threat, including possible actors. For example, NCTC relies to a great degree on the terrorism information in its Terrorism Identities Datamart Environment (TIDE), which is its centralized repository of information about known or suspected international terrorists (KSTs). It also has access to other government information about terrorists and terrorist organizations. NCTC uses threat report details to help it identify the terrorism information most responsive to the threat. It then correlates that terrorism information with other pertinent datasets to which NCTC has access.[3] NCTC examines the results in order to identify, analyze, and provide leads to its partners.

For example, if the threat stream involves travel by a particular KST to carry out an attack at a particular location, NCTC identifies terrorism information from its holdings, and then uses that information to assess travel–related datasets to identify information about that KST's travel to the location in question. NCTC then prepares intelligence reports for appropriate partners based on what it discovers.

Where the government is aware of a threat, but lacks details, NCTC is less able to tap into existing terrorism information that might provide leads for analysis and follow–up. For example, a threat report might warn of a terrorist plot by a particular group involving travel to a particular region during a particular timeframe, but lack any specifics on the individuals involved. In such a case, NCTC could theoretically correlate all accessible travel–related data with all accessible terrorism information. However, assuming this was feasible, this approach would be time–consuming and resource–intensive, and would generate over–inclusive results, i.e., all instances in which any terrorism–related record matches any travel–related record. While the result of that correlation might serve other counterterrorism purposes, it is not responsive to the specific threat at issue.

B. Narrowing the Data to be Correlated.

When responding to generalized threat reporting of this sort, NCTC narrows the data to be correlated with NCTC's terrorism information holdings in order to generate appropriately–focused results. NCTC does this by deriving limiting parameters, based in part on analytic assumptions derived from experience and knowledge about the characteristics of the group or individuals historically involved in such threats, and about general terrorist tradecraft (e.g., communications, travel and counterintelligence). NCTC then applies those parameters to the data at hand.

For example, a threat report may lack specifics about how unidentified terrorist plotters will travel to the identified geographical area to carry out the threatened attack. In such a case, NCTC narrows the possibilities by filtering the accessible travel data based upon previously–utilized modes of travel, or route of travel, or group or individual characteristics of the potential plotters. The resulting body of data is then correlated with TIDE and other terrorism information. The resulting matches, if any, are more likely to yield focused results that can be analyzed for lead information (i.e., focused on those KSTs with travel matching the parameters.)

C. Procedures for Protecting Privacy and Civil Liberties.

When these kinds of analytic parameters are applied to non–terrorism information, it is important to ensure that the results themselves are not used as analytic conclusions. In

[3] "Overview of NCTC's Data Access as Authorized by the 2012 Attorney General Guidelines," published by the National Counterterrorism Center, available at www.nctc.gov/transparency.html

other words, if information from a travel dataset is identified based on a travel pattern that fits certain parameters, that "narrowed–down" travel information should not be used, by itself, to "predict" whether particular individuals who fit that travel pattern might be engaged in terrorist activity. NCTC does not use this technique in that way. Rather, NCTC directly and immediately correlates such subsets with its existing terrorism holdings.

Analytic determinations are made by trained NCTC analysts based on the resulting "matches" with terrorism information already accessible to NCTC – such as TIDE records. Through this technique, analysts focus only on terrorism information that NCTC has already identified through other means, as highlighted, supplemented and prioritized for NCTC analytic review by virtue of this focused correlation. Authorized and trained analysts then analyze the results to identify leads, including information about the identity of individuals with an apparent nexus to terrorism, to report to counterterrorism partners in response to the threat reporting. NCTC does not otherwise make use of the information that is "narrowed down" through the use of these parameters.

If this technique is applied to U.S. government datasets obtained or accessed by NCTC pursuant to the 2012 DNI–Attorney General guidelines, NCTC must apply the baseline safeguards under Section III.C.3(d) of the guidelines in order to protect the privacy and civil liberties of U.S. Persons (USP) whose personal information is contained within this data.[4] Under the baseline safeguards, assessment of information in these datasets must be designed solely to identify information that is reasonably believed to constitute terrorism information, and to minimize the review of USP information that does not constitute terrorism information. Pattern–based assessment is permitted, subject to appropriate reporting under the Data Mining Reporting Act, but must comply with the baseline safeguards. Because the narrowing technique described above is designed solely to narrow the data to be correlated with NCTC's terrorism information holdings, such as KST information in TIDE, it is designed to identify terrorism information. Similarly, because the narrowed–down information is used only for the purposes of such correlation, with NCTC only analyzing the resulting matches with existing terrorism information, the technique is designed to minimize the review of non–terrorism information by analysts.

NCTC has applied the 2012 DNI–Attorney General guidelines to non–terrorism and non–U.S. Person datasets, such as the Department of Homeland Security's Electronic System for Travel Authorization data. For example, in searching these non–U.S. Persons datasets (and in accordance with the established baseline safeguard), analysts are trained to narrowly tailor their queries to "identify information that is reasonably believed to constitute terrorism information" and to "minimize the review" of information about persons that does not constitute terrorism information. This has both an important

[4] Upon certain findings (for which every data set acquired by NCTC is reviewed) – such as data that contains especially sensitive personal information – "Enhanced Safeguards" may also be applied. NCTC reports annually to the ODNI CLPO, the ODNI General Counsel and the IC IG on the measures that NCTC is taking to ensure that USP information in its possession is being handled appropriately.

privacy protective impact—by reducing the number of non–terrorist non–U.S. Persons scrutinized in the analytical process—as well as a practical benefit to the individual analyst, in that it minimizes the need to review unresponsive or irrelevant search results.

In addition, as required by the 2012 DNI–Attorney General guidelines, only specially trained, authorized personnel are permitted to access the information involved in this process, and their analytic activities on NCTC systems are monitored, recorded, and audited. If erroneous or outdated data is identified, it must be corrected, updated, or removed from NCTC systems as appropriate, and the data provider must be notified of the error. Determinations regarding permanent retention, use, and dissemination of USP information are predicated upon an appropriate assessment that the USP information is reasonably believed to constitute terrorism information. Disseminations must satisfy the dissemination requirements of the 2012 guidelines (including any requirements established by the agency that originally provided the data to NCTC), as well as the Privacy Act. Once information has been disseminated by NCTC to its counterterrorism partners, the information is protected by applicable laws and policies, including the Privacy Act (for all Federal agencies) and Executive Order (EO) 12333 (for IC elements).

These measures are subject to compliance and oversight measures at NCTC, as implemented by the NCTC Civil Liberties and Privacy Officer, NCTC legal counsel, and NCTC management.

III. Previously Reported Activities

This section provides updates on programs that were described in last year's report. In the interest of transparency, certain research programs of the Intelligence Advanced Research Project Activity (IARPA) were discussed in last year's report. The mission of IARPA is to invest in high-risk/high-payoff research programs that have the potential to provide the United States with an overwhelming intelligence advantage over its future adversaries. It does not have an operational mission and it does not deploy technologies directly to the field. As a scientific research funding organization, IARPA does not use, nor does it expect to make use of, data mining technology. IARPA programs are by nature experimental and are designed to produce new capabilities. The end goal of an IARPA program is typically a proof–of–concept experiment or prototype of an entirely new capability. Due to their high–risk research nature, IARPA programs do not always achieve their end goals, and when they do, further steps are required to transform the results into real world applications. Any results from IARPA research programs that do get incorporated into future operational programs within the IC, or other parts of the Federal Government, will be subject to appropriate legal, privacy, civil liberties and policy safeguards.

A. Knowledge Discovery and Dissemination (KDD) Program.

The KDD scientific research program is an IARPA program begun in 2009. A Broad Agency Announcement (BAA) for KDD was released on December 22, 2009 and KDD research contracts were awarded in September 2010. The KDD program completed its third period in November 2013.

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

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

B. Automated Low–level Analysis and Description of Diverse Intelligence Video (ALADDIN Video) Program.

The ALADDIN Video scientific research program released a BAA in June 2010, and research contracts were awarded in February 2011. The ALADDIN program completed its third round of testing in the Fall of 2013.

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

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

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

C. Security and Privacy Assurance (SPAR) Program.

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

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

The SPAR program was launched in 2011 to build on the successes of APP and explore additional applications of PIR to realistic IC scenarios. The program completed its first phase of research in March 2013 and started its second phase of research in April 2013. SPAR includes research projects in three technical areas. The first technical area protects security and privacy for database access. Unlike the simple queries and static databases of APP, SPAR will investigate protocols that handle multiple types of complex queries and databases whose records are frequently created, deleted, or updated. In addition, the protocols must integrate policy compliance checking with the security and privacy assurances so that the Server can verify that a query is compliant with a policy even though the query is never learned. The second technical area will build on advances in fully-homomorphic encryption (FHE) schemes to implement PIR without relying on any third parties. FHE is a recent breakthrough result of thirty years of cryptographic research, but current schemes are impractical due to high costs in time and memory. SPAR will attempt to explore gains in performance by modified FHE schemes that support only the computations necessary for information retrieval. The third technical area will investigate applications of PIR to the specialized information sharing architectures of publish/subscribe, email/message queues, and outsourced data storage systems.

If successful, the SPAR protocols will enable the IC to meet the need to access data for classified or sensitive purposes with strong civil liberties and privacy protections. SPAR allows the IC to access specific records without having to disclose classified data and without accessing, learning, ingesting, or retaining any private information about non–

relevant persons. The technology may enhance cooperative information sharing with the IC and other parts of the Federal Government, and with the private sector, by expanding policy options for satisfying security and privacy concerns when information is shared.

IV. Protection of Privacy and Civil Liberties.

The ODNI Civil Liberties and Privacy Office (CLPO) works closely with the ODNI Office of General Counsel, other ODNI components and with IC elements to ensure appropriate legal, privacy, and civil liberties safeguards are incorporated into policies, processes and procedures that support the intelligence mission. The CLPO is led by the Civil Liberties Protection Officer, a position established by the *Implementing Recommendations of the 9/11 Commission Act of 2007*. The duties of this position are set forth in this Act, and include: "ensuring that the protection of civil liberties and privacy is appropriately incorporated in the policies of the ODNI and the IC; overseeing compliance by the ODNI with legal requirements relating to civil liberties and privacy; reviewing complaints about potential abuses of privacy and civil liberties in ODNI programs and activities; and ensuring that technologies sustain, and do not erode, privacy protections relating to the use, collection, and other disclosure of personal information."[5] Before any tool or technology could be used in an operational setting, the use of the tool or technology would need to be examined pursuant to EO 12333, the Privacy Act, and other applicable requirements to determine how the tool could be used consistent with the framework for protecting USP information.

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

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

Going forward, the IC will also conform to policies and procedures relating to protections for all personal information contained in SIGINT, which are required to be put in place by Presidential Policy Directive 28 (issued on January 17, 2014).

[5] National Security Act of 1947 [50 U.S.C. 3029].