International Journal of Intelligence and CounterIntelligence

ISSN: 0885-0607 (Print) 1521-0561 (Online) Journal homepage: http://www.tandfonline.com/loi/ujic20

IC Data Mining in the Post-Snowden Era

William J. Lahneman

To cite this article: William J. Lahneman (2016) IC Data Mining in the Post-Snowden Era, International Journal of Intelligence and CounterIntelligence, 29:4, 700-723, DOI: 10.1080/08850607.2016.1148488

To link to this article: http://dx.doi.org/10.1080/08850607.2016.1148488

Published online: 13 Jun 2016.

International Journal of Intelligence and CounterIntelligence, 29: 700–723, 2016

ISSN: 0885-0607 print/1521-0561 online
DOI: 10.1080/08850607.2016.1148488

WILLIAM J. LAHNEMAN

IC Data Mining in the Post-Snowden Era

"*Gentlemen do not read other gentlemen's mail.*"
—Henry Stimson, U.S. Secretary of State, 1929[1]

Henry Stimson spoke those now famous words to explain his reason for eliminating the Cipher Bureau, the code breaking organization that had been operating under joint State Department—Department of War sponsorship since the end of World War I. The Cipher Bureau, a legacy organization of the U.S. Army's Cryptologic Service, had distinguished itself during the war in providing important intelligence obtained by breaking foreign diplomatic and military codes and ciphers. During the

Dr. William J. Lahneman, Associate Professor of Security Studies and International Affairs at Embry-Riddle University, Daytona Beach, Florida, is also a Senior Research Scholar at the Center for International and Security Studies at Maryland (CISSM) at the University of Maryland's School of Public Policy. A retired U.S. Navy officer, he was a surface warfare officer specializing in strategic planning, international negotiations, and nuclear propulsion. A graduate of the U.S. Naval Academy, Annapolis, Maryland, he earned his M.A. in National Security Affairs at the Naval Postgraduate School, Monterey, California, and his Ph.D. in International Relations at the Johns Hopkins University's School of Advanced International Studies. Dr. Lahneman is the author of Keeping U.S. Intelligence Effective: The Need for a Revolution in Intelligence Affairs *(Lanham, MD: Scarecrow Press, 2011); co-editor with Ruben Arcos of* The Art of Intelligence: Simulations, Exercises, and Games *(Lanham, MD: Rowman and Littlefield, 2014); and co-editor with Joseph Rudolph of* From Mediation to Nation Building: Third Parties and the Management of Communal Conflict *(Lanham, MD: Lexington Books, 2013).*

interwar years, the Cipher Bureau was able to break the diplomatic codes of several nations. Incidentally, that ability to intercept foreign communications was possible only through the secret and illegal cooperation of the entire American cable industry, which operated the transoceanic underwater cables that constituted the leading edge long-range communications technology of that era.

While budgetary constraints probably played some role in Stimson's decision to abolish the Cipher Bureau, he notably chose to invoke ethical rather than budgetary reasons when asked. Moreover, Herbert Yardley,[2] chief of the Cipher Bureau and a legendary figure in forging America's communications intelligence establishment, asserted that Stimson's primary motive for closing down the unit was indeed his view that it was morally and ethically reprehensible to intercept and decode foreign communications.[3]

The employees of the Cipher Bureau were not civil service employees but were instead paid using secret funds. When the Bureau closed they were given three months' salary and separated. This happened just as the Great Depression was gaining momentum. Yardley unexpectedly found himself in an unenviable situation. ''Out of work, with a family to support, and few prospects as the Great Depression worsened, Yardley decided to write his story. It appeared first in serial form in the *Saturday Evening Post*, then as a book, *The American Black Chamber*,''[4] in which he described many of the capabilities of American cryptanalysis, including specific cases of successful interception and decoding of foreign communications during both World War I and the 1920s. The book was a hit, achieving best-seller status in 1932, domestically and abroad. Understandably, Yardley was criticized roundly for revealing the United States's communications intelligence activities and capabilities. He responded to these charges by stating that his motives were purely altruistic. According to a National Security Agency historical assessment, ''His only motive had been to alert the United States to the weakness of its own systems and to the power of cryptanalysts. What he could do, he said, people in other nations could also do.''[5]

Stimson's comment is now used as an example of just how wrong someone can be, both in terms of interpreting the value of an intelligence activity and the application of situational ethics to intelligence activities. The value of effective clandestine communications interception and codebreaking of encrypted signals from foreign powers had already been proven time and again during World War I, as well as during the Washington Naval Conference, when the Black Chamber's communications intelligence (COMINT) products provided American negotiators with advance notice of Japan's negotiating position.[6] Yet Stimson was no fool. On the contrary, a fool would not have been appointed Secretary of State under Republican President Herbert Hoover and then Secretary of War in Franklin D. Roosevelt's Democratic administration during World War II.

Rather, Stimson's value system apparently clashed with the emerging capabilities and needs of twentieth century intelligence operations. He was out of step with the times.

What about Yardley's unauthorized disclosures? If his reasoning is taken at face value, he was concerned that the United States was making a serious mistake by downgrading its communications intelligence and cryptanalysis activities, and felt so strongly about it that he was willing to risk public censure and perhaps even legal prosecution by disclosing classified information. In fact, many branded him a traitor to his country for making such disclosures. The government then successfully suppressed publication of his follow-on work, *Japanese Diplomatic Secrets*. That book about the Black Chamber's interception and decoding of Japanese diplomatic messages in the early 1920s has the dubious distinction of being "the first and only manuscript in American literary history to be seized and impounded for national security reasons by the United States government."[7] Moreover, the Roosevelt administration sought to criminalize Yardley's actions by pressuring Congress to pass the hotly debated but ultimately enacted Public Law 37, which states in part that:

> Whoever, by virtue of his employment by the United States, shall obtain from another or shall have custody of or access to, any official diplomatic code or any matter prepared in such code, and shall willfully, without authorization or competent authority, publish or furnish to another any such code or matter, or any matter which was obtained while in the process of transmission between any foreign government and its diplomatic mission in the United States, shall be fined not more than $10,000 or imprisoned not more than ten years, or both.[8]

While Yardley was never prosecuted, he nevertheless paid a price for his publication of *The American Black Chamber*. He was repeatedly denied employment in any U.S. cryptologic organization, even after the United States began to realize the vital necessity of dramatically expanding these capabilities when the storm clouds of World War II began to gather.

Fortunately, Stimson's closing of the Black Chamber did not cause the dire consequences that could have resulted from such an action. Instead, the Army's Signal Intelligence Service (SIS) and the Navy's OP-20-G organization continued to refine U.S. cryptologic capabilities to intercept, break, and translate foreign diplomatic and military codes after the Black Chamber was closed down. These organizations, although small during the 1930s, accomplished much good work during that period.[9] Had these offices in the War and Navy Departments not continued operating, the United States would have been at a decisive disadvantage at its entry into World War II and throughout the war until such time as a rapidly

assembled U.S. cryptologic enterprise could have succeeded in cracking Japanese and other codes.

Immediately following the attack on Pearl Harbor, evidence indicated that both the SIS and OP-20-G had decoded and translated Japanese diplomatic messages that clearly warned that an attack was imminent. But delays in transmission due to several administrative mistakes prevented this warning from reaching Pearl Harbor prior to the attack. These circumstances created a firestorm of calls to reform U.S. cryptologic functions to make them more effective. The job of directing this effort fell to the Secretaries of War and the Navy. As James Bamford notes in his book on the NSA:

> It is one of cryptology's supreme ironies that the man who now believed that too little attention had been shown the intercepted diplomatic traffic was the same man who a dozen years earlier had slammed closed Herbert Yardley's Black Chamber with the statement "Gentlemen do not read each other's mail": Henry L. Stimson.[10]

When pressed about how he could switch positions so completely, Stimson commented that the situation had changed. Again, according to Bamford,

> "In 1929," Stimson later wrote, "the world was striving with good will for lasting peace, and in this effort all the nations were parties." Then, speaking in the third person, he continued, "Stimson, Secretary of State, was dealing as a gentleman with the gentlemen sent as ambassadors and ministers from friendly nations." Now, as Secretary of War, the gentleman had turned warrior.[11]

Today, most consider Yardley to be someone who clearly spoke truth to power, even though his disclosures arguably broke the law. He is respected, and regarded as the father of American cryptology, because he was *right*—closing the Black Chamber *was* a grievous mistake that would have seriously compromised America's ability to fight and win World War II had the Army and Navy not continued to develop their own cryptologic organizations.

EDWARD SNOWDEN AND PROJECT PRISM

In 2013, Edward Snowden, a contract employee of the National Security Agency (NSA), leaked information to the press about an NSA Project codenamed PRISM, as well as a large number of classified documents dealing with NSA eavesdropping on foreign government leaders and other matters. Project PRISM paid or otherwise required telecommunications firms to provide the NSA with bulk records of their customers' telephone and email records. The NSA data mined this information in search of hidden patterns that might indicate that an individual was engaging in terrorism or conspiracy to commit terrorism. Project PRISM's pattern

recognition software used telecommunications metadata rather than simple data in its attempt to identify likely terrorists. The term ''Metadata'' is most commonly explained as being ''data about the data.'' In other words, metadata includes the time and place of a phone call, the originating phone number and the number called, the length of the call, and so forth. The use of metadata does not allow the data mining entity to know the identity of the caller or the content of the call.

Advocates for the use of this type of metadata argue that Project PRISM did not violate Americans' privacy rights because it did not involve reading the content of personal communications, meaning the identities of the individuals and the topics that they discussed. If the project's data mining software detected a potentially troublesome pattern of calls or emails in the metadata, law enforcement personnel would then be asked to examine this pattern more closely. If they decided that a credible case could be made that the person engaging in this pattern could be a terrorist, they would then seek to obtain a search warrant in keeping with the provisions of the Fourth Amendment to the U.S. Constitution, which states

> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

If a judge were to grant such a warrant, then law enforcement personnel would gain access to the content of the conversations and emails in question. All subsequent investigation—for instance, the initiation of wiretaps, ''sneak and peak'' searches of an individual's home, and other surveillance activities—would be conducted in accordance with standard U.S. law enforcement protocols designed to, among other things, protect Americans' civil liberties. Incidentally, these protocols also allow law enforcement officials to develop evidence and proceed with the arrest and trial of a suspect. Failure to observe these measures makes it relatively easy for a defendant to gain release on a ''technicality,'' that his/her legal rights have been violated because the person had not been afforded due process under the law. The right to due process is enshrined in both the Fifth and Fourteenth Amendments to the Constitution.

Opponents of Project PRISM's data mining activities have repeatedly charged that the use of metadata violates Americans' privacy rights. Although not explicitly stated in the U.S. Constitution, a general acceptance has developed that the Fourth Amendment gives Americans an implicit right to privacy, which is why law enforcement personnel can obtain search warrants only after demonstrating probable cause to suspect that someone is involved in criminal activity. Since Project PRISM's

implementers did not obtain warrants before searching (data mining) the metadata, the project's activities were said to be illegal.

This objection seems to create a conundrum. To perform a search of the personal information of an individual suspected of wrongdoing—for example being a terrorist—law enforcement officials must first obtain a search warrant. Obtaining that warrant, in turn, requires law enforcement personnel to demonstrate probable cause that this individual has committed a crime (or conspired to commit a crime). Project PRISM clearly cannot achieve this prerequisite because demonstrating criminal intent by an individual before mining the metadata to identify suspicious patterns of behavior is not possible. Arguably, blanket permission could be obtained—that is, a warrant authorizing the bulk collection of metadata for this purpose—and this was in fact how the project was periodically authorized by various judges of the secret Federal Intelligence Surveillance Act (FISA) Court.[12] Critics, however, maintained that such permission constituted a ''general warrant,'' a legal device used by British officials during their colonial period in America that gave magistrates the legal authority to search anywhere that they chose while retaining the patina of British legal legitimacy. The Fourth Amendment was specifically designed to prevent the use of general warrants in the fledgling United States of America by specifying that warrants must list the place to be searched and the persons or things to be seized.

Critics also charge that data mining by the Intelligence Community (IC) creates a slippery slope that leads gradually to a total surveillance state in which any claim to the existence of individual privacy is a mirage. In Snowden's words, ''I don't want to live in a world where there's no privacy and therefore no room for intellectual exploration and creativity. What [Project PRISM is] doing poses an existential threat to democracy.''[13] Certainly a case can be made that extensive government data mining activities *do* run the risk of becoming ever more intrusive into citizens' private lives. But is this activity so seductive and inexorable that avoiding the slippery slope is impossible?

Regardless of individual opinions about these issues, Snowden's disclosure of Project PRISM placed American policymakers and the Intelligence Community on the horns of a dilemma. The (meta)data mining capabilities used by Project PRISM were not available until fairly recently. They appear to offer a new way to identify individuals who exhibit a suspicious pattern of behavior before traditional criminal investigative techniques are capable of so doing, which would provide American intelligence and law enforcement organizations with a new advantage in the fight against transnational terrorism. But the novelty of data mining for this purpose means that current laws do not yet reflect to what degree government data mining activities are appropriate within the framework of Americans' right

to privacy. This learning and adjustment process has been playing out since Snowden's 2013 revelations.

After initially defending Project PRISM as essential in the fight against transnational terrorism, the Obama administration backed down and announced that it was reducing the project's activities following the emergence of a loud and strident public reaction to the project based on assertions that it violated Americans' privacy rights. According to the *Washington Post*,

> President Obama made a forceful call to narrow the government's access to millions of Americans' phone records as part of an overhaul of surveillance activities that have raised concerns about official overreach. The president said he no longer wants the National Security Agency to maintain a database of such records. But he left the creation of a new system to subordinates and lawmakers, many of whom are divided on the need for reform. In a speech at the Justice Department, Obama ordered several immediate steps to limit the NSA program that collects domestic phone records, one of the surveillance practices that were exposed last year by former intelligence contractor Edward Snowden.[14]

Project PRISM's bulk metadata collection originally had been approved in April 2013 by the Foreign Intelligence Surveillance Court (FISC).[15] However, by late 2013, legal challenges to Project PRISM were in full swing, with plaintiffs challenging the project on the grounds that it was either unconstitutional (in violation of privacy rights) and/or illegal (that the government's claim that Section 215 of the USA PATRIOT Act provided the legal basis for the Project's collection activities was unjustified).[16] Court rulings differed on these points. Some ruled that the project was constitutional and legal. However, several courts ruled that the project was either unconstitutional, illegal, or both. In November 2013, the Supreme Court declined to rule on an appeal of a decision that the project was in fact constitutional and legal.[17] During the same period, several telecommunications firms responded to the public outcry about invasion of privacy by adopting policies to assure their customers that they would not share customers' personal information—whether actual content or only metadata—with government intelligence agencies.

In December 2013, U.S. District judge Richard Leon ruled that the NSA had violated the privacy rights of the five plaintiffs involved. In his ruling, Judge Leon stated that "I cannot imagine a more 'indiscriminate' and 'arbitrary invasion' than this systematic and high tech collection and retention of personal data on virtually every citizen for purposes of querying and analyzing it without prior judicial approval."[18] Significantly, Leon stated that the government "does not cite a single instance in which analysis of the NSA's bulk metadata collection actually stopped an imminent attack, or otherwise aided the government in achieving any

objective that was time-sensitive in nature."[19] Leon put enforcement of his decision on hold pending an appeal by the government.

In May 2015, the United States Court of Appeals for the Second Circuit ruled that Project PRISM was illegal, stating that "the once-secret National Security Agency program that is systematically collecting Americans' phone records in bulk is illegal . . . [because] a provision of the USA PATRIOT Act known as Section 215 cannot be legitimately interpreted to allow the systematic bulk collection of domestic calling records."[20] The court reasoned that

> [T]he government takes the position that the metadata collected—a vast amount of which does not contain directly "relevant" information, as the government concedes—are nevertheless "relevant" because they may allow the NSA, at some unknown time in the future, utilizing its ability to sift through the trove of irrelevant data it has collected up to that point, to identify information that is relevant. We agree with appellants that such an expansive concept of "relevance" is unprecedented and unwarranted.

This case occurred shortly before Section 215 of the USA PATRIOT Act was due to expire unless Congress voted to make it permanent. If Section 215 were not extended, then there would be no legal basis for Project PRISM's bulk collection activities, regardless of where the courts stood on the issue. The U.S. Senate was so divided over what course of action to take that it could not reach agreement before the deadline of 12:01 AM on 1 June 2015, which caused the authority for NSA's bulk collection activities to lapse for about a day until the Senate passed and President Barack Obama signed the USA Freedom Act on 2 June (the House of Representatives had already passed the bill).[21] The new law did not immediately terminate the program, but instead "calls for a grace period of six months to 'transition' the program so the phone companies remain the repositories of metadata they generate."[22] About two months later, the Obama administration confirmed that phone records held by the government would no longer be examined in terrorism investigations after 28 November 2015 and would ultimately be destroyed. This answered the open question of whether the NSA would be allowed to data mine phone records that it already held in its databases at the time when its bulk collection authority expired. But the NSA is still permitted to request phone records from telecommunications firms as needed in terrorism investigations pursuant to issuance of a warrant.[23]

SNOWDEN'S FATE AND THE FUTURE OF DATA MINING IN THE IC

Many consider the shutdown of Project PRISM to be a vindication of Edward Snowden's 2013 decision to blow the whistle on the project's

activities. He had warned that the IC was violating Americans' privacy rights and, after a period of discussion and review, the Congress, the courts, and a large portion of Americans agreed with him. The USA Freedom Act underscored this view by protecting Americans' privacy rights by prohibiting the bulk collection of their phone records by the IC.

However, the Obama administration made it clear that it was not letting Snowden off the hook—he is still charged with breaking the law. As *The Guardian* reported:

> The White House has rejected a petition to pardon NSA whistleblower Edward Snowden, dubbing the former contractor's revelations about the U.S. government's surveillance apparatus as ''dangerous'' and compromising to national security. ... ''Mr. Snowden's dangerous decision to steal and disclose classified information had severe consequences for the security of our country and the people who work day in and day out to protect it,'' Lisa Monaco, Obama's adviser on homeland security and counterterrorism, said in a statement. ... ''If he felt his actions were consistent with civil disobedience, then he should do what those who have taken issue with their own government do: challenge it, speak out, engage in a constructive act of protest, and—importantly—accept the consequences of his actions.''[24]

How can things be otherwise? As a former officer in the U.S. Navy[25] who has signed non-disclosure agreements, it seems to me like a straightforward case: Snowden violated his agreement, which clearly states that unauthorized disclosure of classified material is a criminal offense. Thus, like Yardley, Snowden stands to pay a price for his whistleblowing.

Nevertheless, the issue of Snowden's legacy is perhaps more important than the legal penalties he might pay for his whistleblowing. In short, could Snowden achieve Yardley's stature? Snowden's disclosures drew attention to a program that has come to be viewed both as violating Americans' privacy rights and illegal by operating outside the traditional standards of judicial review. Snowden has stated repeatedly that he leaked information about the existence of Project PRISM out of a desire to prompt a national debate on this topic to preserve Americans' civil liberties against the onslaught of excessive government surveillance activities.[26] Should his explanation gain wide acceptance, Snowden could be viewed in much the same light as Yardley is now viewed—as someone who did the wrong thing (broke the law) for the right reason (to preserve Americans' civil liberties and the integrity of the U.S. Constitution).

Snowden's legacy—traitor or defender of the Constitution—arguably hinges on two factors rather than the single issue of privacy rights. Another consideration is whether IC activities such as Project PRISM are effective, or hold the promise of becoming effective in the near future.

Yardley is today respected because he was right. His warning about the value of COMINT was absolutely correct. Now, the vast majority of the American people regard ''reading other gentlemen's mail'' as a *sine qua* non of good intelligence work. They consider the IC's need to perform this activity as a necessary exception to Americans' otherwise cherished right to privacy.

Accordingly, what if Snowden should be correct about the reduction in privacy resulting from projects such as PRISM but his actions nevertheless ignore the effectiveness of these kinds of projects? He then might well be remembered as another Stimson rather than a Yardley, as someone who made a colossal misjudgment and jeopardized U.S. national security at a crucial juncture in the country's history by severely reducing an important component of the nation's intelligence apparatus. In other words, is Snowden, like Stimson, out of step with the times?

This is the most important aspect of the Project PRISM controversy: whether the large-scale mining of metadata from numerous government and private sources is an important intelligence tool for fighting terrorism?[27] If not, then the termination of Project PRISM does not harm U.S. national security. If, however, activities like Project PRISM are important for fighting terrorism, then their elimination of such activities is dysfunctional. But if data mining proves effective and the United States does not develop such programs out of concern for privacy issues, is the government once again deciding not to read other gentlemen's mail when everyone else is doing it? Other states' intelligence organizations are likely to be running ''Project PRISMs'' of their own.

Both privacy rights and countering terrorism are important. Just as in Yardley's day, if projects such as PRISM are effective, the American people and their government are likely to find some way to ensure that the mining of metadata—for purposes of fighting terrorism, at least—becomes an unambiguously legal process that is viewed as an essential element of good intelligence tradecraft, even though it intrudes to some degree upon Americans' current conception of privacy rights.

DOES DATA MINING WORK?

Since the Snowden disclosures, President Obama and many high-ranking members of the IC have stated that Project PRISM was an essential asset in the fight against terrorism. For example, John Brennan, Director of the Central Intelligence Agency (CIA), stated in mid-2015 that the surveillance programs were ''important to American lives.''[28] At the same time, other IC members countered that the programs had not produced any valuable results. And a 2014 ''independent analysis of hundreds of terrorism cases in the US concluded that the NSA's collection of phone records has had no distinguishable impact on preventing acts of terrorism.''[29] An

examination of data mining's history, current capabilities, and future potential can help make sense of such contradictory viewpoints.

Data Mining Defined

The term "data mining" refers to a "set of mechanisms and techniques, realized in software, to extract hidden information from data. The word *hidden* in this definition is important."[30] Most of the available information about data mining in general, and the mining of big data in particular, resides in the business and computer science literature. Data mining appeared in the late 1980s as a sub-process within a larger process called knowledge discovery in databases (or KDD), which is "the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data."[31] Data mining is the part of the KDD process concerned with the discovery of "hidden information" contained in data patterns. As Frans Coenen explains, "Other sub-processes that form part of the KDD process are data preparation (warehousing, data cleaning, pre-processing, etc.) and the analysis/visualization of results."[32] Practically speaking, however, many who use the term data mining are actually referring to the entire KDD process. These distinctions are not significant for this analysis.

Data mining utilizes data warehouses for storing and categorizing data to facilitate data mining activities, although newer approaches are not as reliant on stored data.

> A data warehouse is a copy of transaction data specifically structured for query and analysis. A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. A data warehouse is a centralized repository that stores data from multiple information sources and transforms them into a common, multi-dimensional data model for efficient querying and analysis.[33]

Data mining is distinct from traditional information retrieval from databases, regardless of how large or small those data bases are. A University of North Carolina Webpage explains: "In traditional DBMS [database management systems], database records are returned in response to a query; while in knowledge discovery, what is retrieved is not explicit in the database. Rather, it is implicit patterns. The process of discovering such patterns is termed data mining."[34] Patterns detected by data mining indicate that certain of the different variables involved in the data mining operation are *correlated*. Data mining results do not show *causation*. This is an important distinction because it warns analysts that correlations among data indicate only that a particular pattern displayed by a person

(probably only identified at the metadata level at this point) looks like what a terrorist would display, but it does not mean that the individual is in fact a terrorist.

Data mining is a well-established and growing industry in the private sector. ''Data mining applications are marketed by leading software industry heavyweights, including IBM, Oracle, SPSS, and SAS Institute, Inc. These companies market their offerings in a larger industry that achieved sales of approximately $170 billion during the middle years of the twenty-first century's first decade.''[35] The IC has been involved with data mining research and development since at least the early 2000s, when the public learned about the existence of the Pentagon's Total Information Awareness (TIA) Program, a data mining program that Congress cancelled in 2003 amidst privacy concerns.[36] The IC's development of data mining capabilities nevertheless continued following the TIA Program's cancellation.

> According to the Massachusetts Institute of Technology's *Technology Review*, in early 2006 it became known that the Advanced Research and Development Activity (ARDA) office ... had acquired key components of the former TIA program. One example was the Information Awareness Prototype System. Renamed ''Basketball,'' the publication described this component as ''the core architecture that would have integrated all the information extraction, analysis, and dissemination tools developed under TIA.'' *Technology Review* further explained that ARDA had obtained technology called Genoa II, which ''used information technologies to help analysts and decision makers anticipate and preempt terrorist attacks,'' and renamed it ''Topsail.''[37]

Project PRISM is apparently one of the legacy programs derived from these earlier projects.

Mining Big Data

Before further examining data mining, defining ''big data'' is necessary. Big data, used by the IC in any large-scale data mining project, refers to the huge amounts of information that exists in digital databases and warehouses. As Kenneth Cukier and Viktor Mayer-Schoenberger explain:

> In addition to [big data's] sheer magnitude, most of this information is now in digital form. This explosion in data is relatively new. As recently as the year 2000, only one-quarter of all the world's stored information was digital. The rest was preserved on paper, film, and other analog media. But because the amount of digital data expands so quickly—doubling around every three years—that situation was swiftly inverted. Today, less than two percent of all stored information is nondigital.[38]

As other scholars have noted, ''Every day, 2.5 quintillion bytes of data are created and 90 percent of the data in the world today were produced within the past two years.''[39]

For a visual, Cukier and Viktor Mayer-Schoenberger offer: ''If all this information were placed on CDs and they were stacked up, the CDs would form five separate piles that would all reach to the moon.''[40]

The value of big data stems from the ''hidden information'' that it contains, from which ''. . . we can learn from a large body of information things that we could not comprehend when we used only smaller amounts.''[41] The volume of big data is growing at large rates, and not only because more and more people and organizations are digitizing their information. Two other factors play significant roles. First, the amount of data is growing because formerly restricted information is being re-categorized as open source information, and because different kinds of information are being digitized. The U.S. government's Data.gov Website is an excellent example of the first driver: various government agencies have provided previously restricted information to the public in a digital, searchable format. The Department of Energy's Energy Data Initiative is a case in point:

> To help harness the power of American ingenuity to solve our pressing energy challenges, the Obama Administration has launched the Energy Data Initiative. The Energy Data Initiative is an Administration-wide effort to ''liberate'' government data and voluntarily contributed non-government data as fuel to spur entrepreneurship, create value, and create jobs in the transition to a clean energy economy. The goal of the Energy Data Initiative is to fuel entrepreneurs with newly available and previously untapped data—both government and non-government data—to spur new products and services that help American families and businesses save money on utility bills and at the pump, protect the environment, and ensure a safe and reliable energy future.[42]

Many other countries have followed the U.S. lead and released similar types of information.[43] All of this information is available to be data mined. A similar development is making formerly restricted information also available to a larger number of government personnel and agencies, thereby facilitating data mining of certain kinds of government-held proprietary information.[44] In 2010, the Obama administration issued Executive Order 13556, which created a category called ''Controlled Unclassified Information'' (CUI) to handle information that is sensitive but not classified and to create uniform standards for handling unclassified information.''[45]

> [E.O. 13556] establishes an open and uniform program for managing information that requires safeguarding or dissemination controls pursuant to and consistent with law, regulations, and Government-wide policies, *excluding information that is classified* under Executive Order 13526 of December 29, 2009, or the Atomic Energy Act, as amended. ... At present, executive departments and agencies employ ad hoc, agency-specific policies, procedures, and markings to safeguard and control this information, such as information that involves privacy, security, proprietary business interests, and law enforcement investigations. This inefficient, confusing patchwork has resulted in inconsistent marking and safeguarding of documents, *led to unclear or unnecessarily restrictive dissemination policies, and created impediments to authorized information sharing*. The fact that these agency-specific policies are often hidden from public view has only aggravated these issues. [Italics added][46]

If efforts to establish a working interagency method for sharing CUI succeed, data mining of these extensive databases by the IC will be greatly facilitated.

With respect to the third driver of digital data expansion—the emergence of new kinds of information that is being digitalized,

> Big data is also characterized by the ability to render into data many aspects of the world that have never been quantified before; call it ''datafication.'' For example, location has been datafied, first with the invention of latitude and longitude, and more recently with GPS satellite systems. Words are treated as data when computers mine centuries' worth of books. Even friendships and ''likes'' are datafied, via Facebook.[47]

The emergence of large amounts of digital data on mobile devices appears to be taking data mining to a new level. According to Mirco Musolesi, ''[F]or the first time in history, mining data from smart phones and the cellular infrastructure lets us access detailed knowledge about people, things, and events that are happening right now, potentially everywhere in the world.''[48] These processes have been labeled ''big *mobile* data mining.'' Such new capabilities ride on the global expansion of computer processing capability over the past decades. Interestingly, data storage capability has not increased at the same rate. Where big data is concerned, the ability to process data is not matched by storage capability, so the mining of extremely large data sets is increasingly done in near real time.

As the size of the data sets capable of being mined continues to increase, and near real time processing continues to advance, big data mining

programs have demonstrated the ability to not only detect patterns in stored data but to predict future events. For example,

> Researchers in Canada are developing a big-data approach to spot infections in premature babies before overt symptoms appear. By converting 16 vital signs . . . into an information flow of more than 1,000 data points per second, they have been able to find correlations between very minor changes and more serious problems. Eventually, this technique will enable doctors to act earlier to save lives.[49]

When these kinds of advances are included, three potential types of big data mining activities emerge. First, programs can detect useful general patterns of behavior that do not require the identification of individual actors. For example, a supermarket retail chain uses data mining to identify interesting patterns of customer shopping behavior. The chain can track individual customers' buying habits as long as they use their ''savings club'' cards when they buy something. The chain also can identify what each of these customers purchased because all items' barcodes are scanned during each purchase. When customers apply for a savings club card, they usually are required to list several items of personal information, including their gender, on their applications, and this information is entered into the chain's data warehouse. For marketing purposes, the store does not care as much about individual buying habits as it wants to detect useful patterns of behavior that are hidden in the data. Use of metadata is all that is needed. In one interesting case, a high correlation was noted between purchases of baby formula and beer by male customers during the same visit to the chain's stores. Further analysis showed that fathers of young babies often visited the supermarket to buy baby formula. But as long as they were at the store, they used the opportunity to purchase beer on a fairly consistent basis. The store did not care which particular fathers were buying the beer; rather, knowledge about this hidden pattern of activity by male customers was sufficient. The chain of supermarkets then began to locate supplies of baby formula near supplies of beer as a reminder to fathers. Sales of beer went up.[50]

A second form of data mining uses archived data, as with the supermarket example, but those conducting the data mining here want to know individual identities. This appears to be the kind of data mining activity that such projects as TIA and, perhaps, Project PRISM, sought to conduct. Using metadata, data mining programs look for hidden patterns in archived data—travel records, financial transactions, phone calls, emails, etc.—that might point to someone being a terrorist. Positive results in such programs

might lead to possible identification of a suspect when the pattern is convincing enough to convince a judge to issue a search warrant.

The third and most advanced form of data mining can be used to detect infections in premature infants. This type of data mining uses both archival and real time data to detect patterns in near real time that point to problems in the future. Like the second type of data mining, it identifies specific individuals of interest, not just general patterns of behavior.

CHALLENGES TO SUCCESSFUL DATA MINING

Challenges to successful data mining are based on both statistical principles and the complex nature of data. The larger the data sets involved, the more significant these challenges can become.

Bonferroni's Principle

In trying to detect a particular hidden pattern of behavior in a large amount of data, the problem of false positive cases arises. For instance, a data mining program might be searching for hidden patterns that correlate to someone being a terrorist. Statistically speaking, this pattern will arise not only because someone might be a terrorist but also because the rules of probability dictate that a certain number of innocent peoples' activity patterns will randomly align into the sought after pattern. These latter cases will be false positives: the persons involved are not terrorists, but the data mining program has identified their pattern of behavior as one that terrorists would display. Data mining programs must be able to discriminate between these false positives and the persons who are in fact terrorists.

Bonferroni's principle provides a benchmark for preventing analysts from treating false positive cases as if they were the real thing. Bonferroni's principle is based on calculating

> ...the expected number of occurrences of the events you are looking for, on the assumption that data is random. If this number is significantly larger than the number of real instances you hope to find, then you must expect almost anything you find to be bogus, i.e., a statistical artifact rather than evidence of what you are looking for. This observation is the informal statement of Bonferroni's principle.[51]

Figure 1 provides an example of the application of this principle.

Suppose there are believed to be some "evil-doers" out there, and we want to detect them. Suppose further that we have reason to believe that periodically, evil-doers gather at a hotel to plot their evil. Let us make the following assumptions about the size of the problem:

1. There are one billion people who might be evil-doers.
2. Everyone goes to a hotel one day in 100.
3. A hotel holds 100 people. Hence, there are 100,000 hotels—enough to hold the 1% of a billion people who visit a hotel on any given day.
4. We shall examine hotel records for 1,000 days.

To find evil-doers in this data, we shall look for people who, on two different days, were both at the same hotel. Suppose, however, that there really are no evil-doers. That is, everyone behaves at random, deciding with probability 0.01 to visit a hotel on any given day, and if so, choosing one of the 10^5 hotels at random. Would we find any pairs of people who appear to be evil-doers? We can do a simple approximate calculation as follows. The probability of any two people both deciding to visit a hotel on any given day is 0.0001. The chance that they will visit the same hotel is this probability divided by 10^5, the number of hotels. Thus, the chance that they will visit the same hotel on one given day is 10^{-9}. The chance that they will visit the same hotel on two different given days is the square of this number, 10^{-18}. Note that the hotels can be different on the two days.

Now, we must consider how many events will indicate evil-doing. An "event" in this sense is a pair of people and a pair of days, such that the two people were at the same hotel on each of the two days. To simplify the arithmetic, note that for large n, the number of possible pairs of 2 persons in the population is about $n^2/2$. We shall use this approximation in what follows. Thus, the number of pairs of people is $(10^9)^2/2 = 5 \times 10^{17}$. The number of pairs of days is $1{,}000^2/2 = 5 \times 10^5$. The expected number of events that look like evil-doing is the product of the number of pairs of people, the number of pairs of days, and the probability that any one pair of people and pair of days is an instance of the behavior we are looking for. That number is

$$5 \times 10^{17} \times 5 \times 10^5 \times 10^{-18} = 250{,}000$$

That is, there will be a quarter of a million pairs of people who look like evildoers, even though they are not. Now, suppose there really are 10 pairs of evil-doers out there. The police will need to investigate a quarter of a million other pairs in order to find the real evil-doers. In addition to the intrusion on the lives of half a million innocent people, the work involved is sufficiently great that this approach to finding evil-doers is probably not feasible.

Figure 1. An example of Bonferroni's Principle.[52]

In addition to pointing out the false positive problem that all encounter when dealing with massive datasets, Bonferroni's Principle also indicates ways to reduce the number of random occurrences and thus help identify actual cases of interest. The equation used in the calculation example in Figure 1, when expressed in words, becomes the possible numbers of

$$(\text{Pairs of people}) \times (\text{pairs of days}) \times \begin{pmatrix} \text{pairs of people who visit the} \\ \text{same hotel on two different days} \end{pmatrix}$$
$$= \text{number of random occurrences}$$

The first two numbers are very large because the amount of data is massive. The third number is very small because the random chance of two people visiting the same hotel on the same two days is very remote, again because the dataset is so large. Thus, to bring down the number of false positives, data miners can:

- Reduce the amount of data that is mined. This option can be discounted because it is contrary to the very purpose of mining massive data sets. In fact, the size of data sets will only continue to grow, which will exert upward pressure on false positive rates.
- Increase the number of contacts that the program is searching for. For example, instead of pairs of people, look for occasions when 10 individuals meet together. This will reduce the magnitude of the first term and, all things equal, decrease the number of false positives.
- Increase the number of days that groups of people meet. The larger this number, the lower the probability of random occurrences, which will lower the second term and thus lower the overall false positive numbers. For example, let's set the program to look for ten days rather than pairs of days.
- Increase the variables in the last term. Instead of looking for the (now) ten individuals who visit the same hotel on (now) the same ten days, let's add additional variables that must also be met at the same time. For example, add that individuals also have questionable bank transactions, exhibit suspicious travel patterns, and so on. The probability of all of these factors aligning at the same time will make the third term even smaller and reduce the number of false positives.

Data miners must also remember that variables might be related; changing one variable might affect others. For example, an increase in the amount of data being examined will almost surely raise the number of people involved, which will raise the numbers of pairs (or tens or twenties) of people and also increase the chance of random meetings at the same hotels.

This discussion demonstrates the kinds of considerations that will arise when trying to establish effective data mining activities to detect groups of persons—terrorists for instance—which exist in extremely small numbers

within a massive population. Bonferroni's Principle is definitely a consideration in these kinds of cases. Too many false positives render results meaningless; even a relatively small number would likely overwhelm the investigative capabilities of law enforcement personnel. Judges would soon be reluctant to grant warrants based on untrustworthy information derived from data mining. And complaints about violations of personal privacy by those unjustly accused would soon cause such programs to shut down, at least until improvements could be made and greater effectiveness established.

The HACE Theorem

In addition to the statistical considerations that must be taken into account when performing data mining of massive datasets, the very nature of the data causes other problems. The "HACE Theorem" reminds data miners of the nature of the space in which they operate. It states that "Big Data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data."[53]

The HACE Theorem reminds data miners that big data is not only "big," but also that various datasets reside in many different locations/data warehouses and use different software. Specifically, the theorem states that big data:

- is derived from *heterogeneous sources*—each data set and/or data warehouse might use different kinds of software, and these might be incompatible, making the use of middleware necessary to allow the entity performing the data mining to interface with this source of data. The greater the number of individual datasets being mined in a project, the more significant this consideration becomes.
- is derived from *autonomous sources* with distributed and decentralized control. In most cases, access to data requires permission of the owner. Sometimes this permission is given to all who want to use it. On other occasions, the dataset is proprietary. In such cases, the IC must either (1) receive the owner's permission to use the data (this might include paying for it); (2) use legal means to compel the owner to allow the IC to use it (which works with U.S. companies but not with foreign entities that are not subject to U.S. law); or (3) steal it. Data.gov is an example of an open source of data. The phone and email records that the National Security Agency required several telecommunications companies to provide for Project PRISM are examples of the second way of obtaining data. And the wiretapping of the conversations and emails of foreign leaders and other foreign government officials, also part of Project PRISM, is an example of the third way. In all of these cases, the fact that different organizational entities have possession of needed datasets presents a formidable coordination problem to large-scale data mining activities.

- explores *complex relationships*—As noted in the description of Bonferroni's principle, designers of programs for mining big data must account for the complex nature of their undertaking. This includes the fact that variables might be interrelated and that random chance will produce a number of false positive cases that must be designed out of the program as much as possible. In the real world, access to certain datasets might not be possible, thereby precluding certain avenues for increasing effectiveness. In addition to countering false positives, designers must try to minimize false negatives, i.e., cases where the data mining program fails to detect an actual terrorist. A program characterized by too many false negatives will face difficulty in continuing its operations because these erroneous results will ultimately cause the program to be judged ineffective.
- explores *evolving relationships*—As if complexity were not a difficult enough factor to manage, datasets exhibit dynamic qualities. This is certainly true in the case of identifying terrorists, who will change their operating practices whenever they suspect they are under suspicion. Fortunately, the evolution of mobile big data mining and other data mining techniques that operate in near real time seem to offer some ways to cope with the virtual certainty that target populations such as terrorists will change their patterns of behavior.

DATA MINING AND THE INTELLIGENCE COMMUNITY

Given the history of data mining in the IC, and Project PRISM in particular, the U.S. IC is clearly improving these capabilities slowly but surely. Most likely, other powers, such as Russia and China, which do not have to be responsive to public opinion concerning such matters, are similarly developing large-scale data mining operations. Yet, the Snowden disclosures caused the Obama administration to stop the NSA's bulk collection of telecommunications metadata due to public and media outcries about privacy concerns—even while stating how essential data mining activities are to U.S. national security. Thus, assessing whether this might be a ''Henry Stimson'' moment is critical in order to prevent U.S. policy about IC data mining from being on the wrong side of history in the same way that Stimson's decision to abolish the Cipher Bureau was a huge mistake.

Businesses have committed to data mining because it improves the performance of their firms. Industry revenues are growing at a robust rate. Not all of this data mining is relevant to the IC. The IC's use of data mining centers on detecting small numbers of terrorists hiding in plain sight among the large American population. This is truly looking for a needle in a haystack, which makes minimizing the number of false positive detections very relevant. While challenging, solving the problem appears possible. But if the IC's access to datasets is too limited, its ability to design data mining programs that use a sufficiently large number of

variables to reduce false positive results to manageable levels will be significantly restricted.

Access to sufficiently large amounts of data remains problematic for two reasons, one technical, and the other legal. On the technical side, datasets are owned by various entities; datasets use different software, which makes achieving compatibility among the multiple databases needed for data mining an issue. Designing effective data mining programs is difficult at best, and the hidden patterns that data mining programs look for can change over time without notice. On the legal side, strong concerns abound that the large-scale mining of big datasets violates Americans' privacy rights.

Solving the technical problems associated with the mining of massive datasets remains a challenge, but the IC's relentless pursuit of data mining technology since the discovery of the Total Information Awareness Program seems to indicate a steadfast belief that further research and development will yield valuable results. The continued expansion of data mining applications in the private sector also points to the fact that data mining will become an increasingly important—even commonplace—activity that produces results.

Solving the legal challenges to large-scale data mining will be difficult. Continued secret IC development of such programs is seductive; not only does it bypass public debate of the privacy issue, but secrecy is in keeping with the IC's organizational culture. Such an approach might be effective as long as these programs remain small scale. But large-scale data mining operations by the IC carried out in secret are bound to be disclosed (again) sooner or later, and these disclosures will prompt a strong public condemnation if the Snowden scandal is any indication. If data mining is as important as claimed, such an arguably justifiable backlash will harm U.S. data mining efforts and deprive the IC of a critical capability. However, the Stimson–Yardley controversy of the early 1930s seems to indicate that Americans will accept intrusions into their privacy for good enough reasons. But they need to hear the reasons. If the IC feels that data mining is critical to U.S. national security, then the time has come for an open policy debate on the issue that will include how to redefine the boundaries of privacy in America.

REFERENCES

[1] An alternate version states that Stimson actually said "Gentlemen do not read each other's mail."

[2] For further reading about Herbert Yardley, see Herbert O. Yardley, *The American Black Chamber* (Annapolis, MD: Bluejacket Books, Naval Institute

Press, 1931); David Kahn, *The Reader of Gentlemen's Mail* (New Haven, CT: Yale University Press, 2004); and James Bamford, *The Puzzle Palace: Inside the National Security Agency, America's Most Secret Intelligence Organization* (New York: Penguin Books, 1983), pp. 20–81.

[3] *Pearl Harbor Review—The Black Chamber*, National Security Agency/Central Security Service website, at https://www.nsa.gov/about/cryptologic_heritage/center_crypt_history/pearl_harbor_review/black_chamber.shtml. Accessed 11 September 2015.

[4] National Security Agency, *Pearl Harbor Review—The Black Chamber*.

[5] *Ibid.*

[6] James Bamford, *The Puzzle Palace*, pp. 26–27.

[7] *Ibid.,* p. 43.

[8] *Ibid.,* pp. 45–46.

[9] *Ibid.*, pp. 46–62.

[10] *Ibid.*, p. 62.

[11] *Ibid.*

[12] Doug Mataconis, "Appeals Court Rules N.S.A. Data Mining Illegal," *Outside the Beltway*, 7 May 2015, p. 2, at http://www.outsidethebeltway.com/appeals-court-rules-n-s-a-data-mining-illegal/. Accessed 11 September 2015.

[13] Glenn Greenwald, Ewen MacAskill, and Laura Poitras, "Edward Snowden: The Whistleblower Behind the NSA Surveillance Revelations," *The Guardian*, 11 June 2013.

[14] Ellen Nakashima and Greg Miller, "Obama Calls for Significant Changes in Collection of Phone Records of U.S. Citizens," *The Washington Post*, 17 January 2014.

[15] Bill Mears and Evan Perez, "Judge: NSA Domestic Phone Data-Mining Unconstitutional," CNN.Com, 16 December 2013, p. 3, at http://www.cnn.com/2013/12/16/justice/nsa-surveillance-court-ruling. Accessed 11 September 2015.

[16] Doug Mataconis, "Appeals Court Rules N.S.A. Data Mining Illegal," pp. 2–3.

[17] Bill Mears and Evan Perez, "Judge: NSA Domestic Phone Date-Mining Unconstitutional," p. 3.

[18] *Ibid.*, p. 2.

[19] *Ibid.*

[20] Doug Mataconis, "Appeals Court Rules N.S.A. Data Mining Illegal."

[21] Sabrina Siddiqui, "Congress Passes NSA Surveillance Reform in Vindication for Snowden," *The Guardian*, 3 June 2015, p. 1, at http://theguardian.com/us-news/2015/jun/02/congress-surveillance-reform-edward-snowden. Accessed 11 September 2015.

[22] *Ibid.*

[23] "NSA will Destroy Millions of American Calling Records 'as Soon as Possible,'" *The Guardian*, 27 July 2015, p. 1, at http://www.theguardian.com/us-news/2015/jul/27/nsa-destroy-american-calling-records. Accessed 11 September 2015.

[24] Sabrina Siddiqui, "Congress Passes NSA Surveillance Reform in Vindication for Snowden," p. 1.

[25] Author's active duty as a Surface Warfare Officer from 1974–1995.

[26] "[Snowden] has had 'a very comfortable life' that included a salary of roughly $200,000, a girlfriend with whom he shared a home in Hawaii, a stable career, and a family he loves. 'I'm willing to sacrifice all of that because I can't in good conscience allow the US government to destroy privacy, Internet freedom and basic liberties for people around the world with this massive surveillance machine they're secretly building." Glenn Greenwald, Ewen MacAskill, and Laura Poitras, "Edward Snowden: The Whistleblower Behind the NSA Surveillance Revelations."

[27] If so, data mining activities could turn out to be an effective tool for fighting other transnational problems such as transnational criminal network activity, which includes crimes such as the illegal trafficking in human beings for the sex trade, forced labor, the trade in children, and the harvesting of organs. Data mining might also be effective at targeting the trade in exotic and endangered species, the transnational narcotics trade, and the illegal weapons trade.

[28] Ewen MacAskill, "A Huge Victory on Mass Surveillance for Snowden—and It's Not over Yet," *The Guardian*, 1 June 2015, p. 2, at http://theguardian.com/commentisfree/2015/jun/01/victory-mass-surveillance-snowden-bulk-data-collection-nsa-transparency. Accessed 11 September 2015.

[29] Sabrina Siddiqui, "Congress Passes NSA Surveillance Reform in Vindication for Snowden," p. 3.

[30] Frans Coenen, "Data Mining: Past, Present, and Future," *The Knowledge Engineering Review*, Vol. 26, No. 1, 2011, pp. 25–29.

[31] *Ibid.*, p. 25. Fayyad quote from U. Fayyad, H. Piatetsky-Shapiro, and P. Smyth, "The KDD Process for Extracting Useful Knowledge from Volumes of Data," *Communications of the ACM*, Vol. 39, No. 11, pp. 27–34.

[32] Frans Coenen, "Data Mining: Past, Present, and Future," p. 25.

[33] S. G. Anantwar and R. R. Shelke, "Methodology for Data Mining," *International Journal of Electronics, Communications and Soft Computing Science and Engineering* (IJECSCSE), Vol. 2, No. 4, pp. 6–10, at http://search.proquest.com.ezproxy.libproxy.db.erau.edu/docview/1432078408?.accountid=27203

[34] "What is Data Mining?" Webpage, University of North Carolina, at http://www.unc.edu/~xluan.258.datamining.html

[35] "Data Mining Industry Snapshot," *Encyclopedia of Emerging Industries*, 5th ed., Gale Virtual Reference Library (Detroit: Gale, 2007).

[36] *Ibid.*

[37] "Data Mining Industry Snapshot." See also Mark Williams, "Information Awareness Project Lives On," *Technology Review*, 26 April 2006.

[38] Kenneth Cukier and Viktor Mayer-Schoenberger, "The Rise of Big Data: How It's Changing the Way We Think About the World," *Foreign Affairs*, May/June 2013, pp. 28–40.

[39] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 1, January 2014.

[40] Kenneth Cukier and Viktor Mayer-Schoenberger, "The Rise of Big Data: How It's Changing the Way We Think About the World."

[41] *Ibid.*, p. 28.

[42] Go to http://www.data.gov/energy/energy-data-initiative

[43] Kenneth Cukier and Viktor Mayer-Schoenberger, "The Rise of Big Data and How It's Changing the Way We Think About the World," p. 36.

[44] See William J. Lahneman, "Examining the NGPI Dots," *International Journal of Intelligence and CounterIntelligence*, Vol. 26, No. 4, Winter 2013–2014, pp. 730–743, at pp. 733–734.

[45] Mark M. Lowenthal, *Intelligence: From Secrets to Policy*, 6th Edition (Washington, DC: CQ Press, 2015), p. 93.

[46] Executive Order 13556, *Controlled Unclassified Information* (Washington, DC: The White House, 2010.), at http://www.whitehouse.gov/the-press-office/2010/11/04/executive-order-controlled-unclassified-information. For more information, see William J. Lahneman, "Reshaping Intelligence and Security in the 21st Century," paper presented at the International Conference on Intelligence in the Knowledge Society, Bucharest, Romania, 17 October 2014.

[47] Kenneth Cukier and Viktor Mayer-Schoenberger, "The Rise of Big Data: How It's Changing the Way We Think About the World," p. 29.

[48] Mirco Musolesi, "Big Mobile Data Mining: Good or Evil?" *IEEE Internet Computing*, January/February 2014, p. 78, at www.computer.org/internet.

[49] Kenneth Cukier and Viktor Mayer-Schoenberger, "The Rise of Big Data: How It's Changing the Way We Think About the World," p. 32.

[50] "Data Mining, Industry Snapshot," *Encyclopedia of Emerging Industries*, p. 2.

[51] Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman, *Mining of Massive Datasets* (Palo Alto, CA: Stanford University, 2014). Self-published volume, p. 5, at infolab.stanford.edu/~ullman/mmds/book.pdf.

[52] Extracted from *Ibid.*, p. 6.

[53] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data Mining with Big Data," p. 98.