July 17, 2008

VIA EMAIL
Ms. Toby Milgrom Levin
Senior Advisor
Privacy Office
Department of Homeland Security
Washington, DC 20528

Re: Docket Number (DHS-2008-0061), Data Mining Workshop

Dear Ms. Levin:

Thank you for this opportunity to provide comment on the proposed Data Mining Workshop. We welcome the opportunity to help the Department of Homeland Security better understand the technology and the privacy implications. Our goal in participating is to try to illuminate and demystify the technology of data mining (while discussing inherent issues), as well as providing insights into how it can be used responsibly and respectfully.

Who We Are.
As we understand the Federal Register notice, the purpose of the workshop is to help inform the Department as it prepares the 2008 report required by the Implementing Recommendations of the 9/11 Commission Act of 2007, by bringing together academicians, policy makers, and technology experts. By way of background, SAS is a leading provider of business analytics software and the largest independent software vendor in the business intelligence market. Our software is used in 44,000 sites worldwide, including most of the top Fortune 100 companies, all 15 of the departments in the federal government (including the Department of Homeland Security) as well as most of the federal agencies and independent agencies, all 50 state governments, and governments worldwide. Besides government, we have developed vertical industry expertise and products for the financial services, health care/life sciences, education, and retail sectors, to name just a few. SAS software has been awarded numerous honors, including Gartner's Leader Quadrant for Business Intelligence, Gartner's Leader Quadrant for Customer Data Mining Applications, ranked by IDC Magazine as a leading provider of advanced analytics, and *Technology & Learning Magazine's* Award of Excellence.

The SAS platform has three distinct components:

- data integration and cleansing—providing the tools and architecture for the consistent access and delivery of clean data across the enterprise;
- business intelligence—the ability to provide decision makers with the information that they need, in the form that they need, in the time that they need it; and
- predictive analytics—the use of a variety of analytic techniques and processes, including but not limited to, data mining, to help in the collection, classification, analysis and interpretation of data to reveal patterns, anomalies, key variables and relationships.

Our goal is twofold: To be the first place customers turn when they need help solving problems, and, more fundamentally, to help our customers make better decisions, faster—based on data, rather than instinct.

Consequently, we have coined the trademark phrase "SAS gives you the POWER TO KNOW" to encapsulate and articulate our core value proposition.

What is Data Mining?

The Federal Register notice sets out a number of very specific questions posed by the Privacy Office to help it begin to compile the 2008 Privacy Report. We note that the Federal Register notice is silent as to what the Privacy Office means by "data mining", and further note that in its 2008 preliminary response to Congress, it reflects that it is operating under a number of different definitions. We also submit that while the Privacy Office has hosted a number of workshops, some of which have tangentially dealt with the issue of data mining, we believe that before the dialogue can proceed and the questions answered, participants to this dialogue should reach consensus about what constitutes "data mining".

SAS defines "data mining" as the following:

> *An iterative process of creating predictive and descriptive models to identify trends and patterns in vast amounts of both structured and unstructured data from across the enterprise, in order to support decision making.*

Structured data, of which the government has vast amounts, consists of rows and columns of data in tables. In general, most of the problems the government is attempting to resolve are based on the use of structured data. Unstructured data, on the other hand, is data that is not susceptible to easy categorization, generally consisting of textual content. For example, hand written notes by a consular field office or a memo field constitutes unstructured data.

There is much value in being able to go beyond merely searching for terms and instead, transform textual data into a usable, intelligible format that facilitates classifying documents, finding explicit relationships or associations between textual content. By including textual content in the data mining process, the government will have a 360 degree, or holistic view, of the entire problem. Stated differently, excluding the ability to engage in "text mining" severely limits the government's ability to make decisions of the highest quality because it is looking at an incomplete picture. Increasingly, unstructured data is the type of information that our intelligence community and our homeland security community will need to decipher.

Beyond the need to capture both structured and unstructured data, our definition encompasses several other key points:

- First, that data mining is not a one-time operation, but an iterative process that involves people, processes and technology. The results often become inputs to further refine the techniques used in preparing the data or providing new clues for adjusting the modeling or discovery techniques. As data changes over time, iterations of the predictive and descriptive models are necessary so that the outcomes remain relevant.

- Second, to undertake meaningful data mining, it is *critical* to understand what the problem is for which a solution is sought. That, in turn, will drive the types of variables that will need to be modeled to reach any sort of meaningful conclusion.

- Third, data mining models are not black boxes and can be evaluated objectively through standard statistical diagnostics to measure goodness of fit and overall applicability. Data mining also incorporates validation data to help prevent "overfitting", as well as test tables to evaluate how the model generalizes on new data.

- Fourth, data needs to be well understood and appropriate to undertake the modeling exercise. In fact, data preparation is probably the most fundamental aspect of data mining, involving not just having "clean", relevant data, but having the domain expertise resident to understand the data. Besides the data preparation, it is critical that the enterprise have the human capital on hand to both create the models and to appropriately understand and use the outputs of the process. This requires access to people that have a statistical background and significant training in the use of data mining.

- Fifth, data mining does not exist independently of the problem that is to be solved. It is not a series of random queries on databases, as is suggested by certain legislative formulations of the definition.

In short, for SAS, data mining is simply the use of statistical analysis to reach conclusions with a high degree of probability or confidence. Once champion models have been developed that provide accurate "predictions" of outcomes, new data can be scored and assigned a "percent likelihood" or "probability" of occurrence. Typically, a data mining exercise will involve the comparison of many different models side by side so that the "lift", or increased gain in accurately identifying an outcome, can be achieved mathematically. This eliminates the risk of falsely identifying an outcome as true. We find this approach to be a much better approach to decision making because it introduces much less bias into the process than non-mathematical approaches to problem solving.

Addressing Specific Questions Posed by the Privacy Office

With this background, we now shift our attention to several questions raised by the Federal Register Notice:

2. Are there privacy issues that are unique to government data mining?
4. What should be the criteria for validating government data mining models and rules?
5. Are anonymization techniques or tools currently available that could be used in conjunction with government data mining? How effective are these techniques or tools? What are their costs and benefit? What degree of de-identification do they make possible?
8. Data quality plays an important role in the ability of government data mining techniques to produce accurate results. What data quality standards should DHS adopt for data mining?

Question 2: Are there privacy issues that are unique to government data mining?

Privacy concerns encountered by private and public sector entities vary depending upon a number of factors, including the purposes of processing and the use, disclosure and retention of personal data. Data collected about individuals for national security or law enforcement purposes raises important privacy issues whether the data is collected by the private or public sector. In both instances, the collector/user must ensure that the data is appropriate and "clean". Similarly, both sectors need to engage trained individuals to prepare and validate the relevant predictive and descriptive models, and interpret the results of such models. If done properly in either context, the results of data mining require further investigation before the correct interpretation of the results can be made, irrespective of whether the data mining is being done by the government or private sector. For example, DHS might use data mining to help it identify a pattern of activity that suggests that the activity supports terrorism. It would be incorrect, however, merely on the basis of the data mining, to say that the

results of the data mining definitively identify terrorists. That conclusion can only be reached after further, in-depth and probably manual investigation. Similarly, a bank might use data mining to identify patterns of behavior that suggest money laundering, but the results themselves are not conclusive that any single transaction is, in fact, money laundering without further investigation. In either instance, it may be improper and dangerous to conclude that a person is a terrorist or a money launderer simply on the basis of the data mining. Human intervention is a necessary component in effective data mining and it is the human capital involvement in data mining, and the subsequent activities (or lack thereof), that trigger privacy implications. We perceive the privacy issues to be substantially similar whether the personal data is collected and analyzed for national security or law enforcement purposes directly by governmental entities or indirectly by private entities as stewards or otherwise on behalf or for the benefit of governmental agencies.

Question 4: What should be the criteria for validating government data mining models and rules?

We cannot conclusively provide a "one size fits all" test to determine whether a model is valid. The process of model development is far too complex, with many different variables to consider. We can, however, provide a process and rules that might be employed to assess model validity.

Dorian Pyle, a noted expert on building models, suggests ten "golden rules" for model building in his book "Data Preparation for Data Mining." These rules are:

- Select clearly defined problems that will yield tangible benefits.
- Specify the required solution.
- Define how the solution delivered is going to be used.
- Understand as much as possible about the problem and the data set.
- Let the problem drive the modeling.
- Stipulate assumptions.
- Refine the model iteratively.
- Make the model as simple as possible—but no simpler.
- Define instability in the model (critical areas where change in output is drastically different for a small change in inputs.
- Define uncertainty in the model (critical areas and ranges in the data set where the model produces low confidence predictions or insights).

These steps offer a good perspective of what takes place during the modeling process, and is essentially the process we follow when we help our customers construct models. We would suggest that these rules serve as questions to help outside reviewers understand how to validate the specific modeling process that led to the results. For example, turn the first rule into the following question: what problem is driving the modeling? What assumptions are in place for how the data was prepared? Validation should also consider the metadata, or the data about what happens to the data, from the data preparation and cleansing process. To that end, there is always an "extraction from source data, transformation and load" process that is a precursor to the modeling process. An examination of that metadata would be a necessary component in determining the validity of the data given the problem that one is attempting to solve.

Question 5: "Are anonymization techniques or tools currently available that could be used in conjunction with government data mining? How effective are these techniques or tools? What are their costs and benefit? What degree of de-identification do they make possible?"

Data can be easily "anonymized" so that information is not personally identifiable. These techniques are readily available in the commercial market and are generally not hard to utilize. The techniques are varied in utility and cost. Besides de-identification, there are ways to control row and column access within an architecture that support authentication against metadata. The architecture is one that scales well as both users and data increase, and can be controlled by a single point.

The question implicitly suggests that the data to be studied is personally identifiable information ("pii"). As an obvious starting point, if the information is not PII, then anonymization of the data is not necessary. From our practice, we generally do not want our clients sending us PII—in many cases, the solutions that we are seeking do not require the use of PII. For example, we are currently doing work for a county government health organization seeking to better understand the types of services that are being demanded, where those services are being demanded, and how frequently certain services are being utilized. That type of analysis can be done without the use of names or Social Security Numbers. In this case, anonymizing the variables that are used in the analysis does not serve a privacy function because PII is not critical to the analysis.

Where we do need PII, as in certain types of financial analysis, we ask that our customers provide the PII in a hashed or encrypted format. They retain the key to the hash or encryption. The data is sent to us in hashed format during transit, and since we do not have the algorithm to unlock the data, it stays in hashed form during the analysis. It is absolutely possible to undertake high quality analytics with data in a scrambled form. Once the analysis is complete, we return the scrambled data and analysis to the customer.

We also believe that the use of metadata, as described above, provides both privacy assurances as well as an audit log to track who is accessing the data and when it is being accessed.

Question 8: "Data quality plays an important role in the ability of government data mining techniques to produce accurate results. What data quality standards should DHS adopt for data mining?"

There are many different tasks that are performed during a data quality exercise. The starting point for performing data quality begins with gathering people together who can help answer fundamental questions about the problem, the data, the analysis and the solution. Since tracking and quantifying these discussions is nearly impossible, metadata captured as data moves from its operational data source to the point where it becomes "analytically ready," becomes a surrogate so that the process of what takes place can be understood.

If DHS is interested in understanding the process behind the steps undertaken during data preparation and cleansing, we would recommend they set up a template for what metadata is necessary and can be expected to be captured, and what information falls outside of the metadata that needs to be filled in, so that they can understand what tasks have been conducted during the process. This would provide DHS with an understanding of what has been conducted and then judgment could be made by those who are skilled in the field of data mining and modeling to determine if the appropriate rigor has been applied to the data prior to the interpretation of the results.

Again, we thank you for this opportunity to provide comment, and would welcome participation in the forthcoming Data Mining Workshop. We are also happy to provide any additional information or insights that you may require. We are based in Rosslyn, Virginia, and can be reached at 571-227-7000. My extension is 5208; my colleague, Katherine Hahn, can be reached at extension 5212.

Sincerely,

/s/ John Stultz

John Stultz,
Manager, Federal Systems Engineers

DHS Data Mining Workshop
Comments from the Electronic Frontier Foundation (EFF)
DHS Docket Number DHS-2008-0061

To:
Toby Milgrom Levin
Senior Advisor, Privacy Office
Department of Homeland Security
Washington, DC 20528

From:
Electronic Frontier Foundation
454 Shotwell St.
San Franciso, CA 94110-1914
(415) 436-9333
http://www.eff.org/

Please direct inquiries and responses to:
Jennifer Granick
EFF Civil Liberties Director
(415) 436-9333 x134
jennifer@eff.org

The questions posed by the Department of Homeland Security (DHS) raise important Constitutional privacy issues, and EFF is glad that these issues are being seriously considered in a public forum. The Bill of Rights includes many privacy protections for individuals, but courts and legislators are struggling to maintain them in the face of new governmental powers, such as those created by powerful data mining technologies. It is nonetheless the responsibility of DHS to maintain these historically important Constitutional protections in implementing these technologies. Instead of answering each question separately, our comments will first outline our concerns over specific privacy problems caused by data mining, then provide some guidelines on how these can be addressed in the design of a data mining system.

EFF's overall position is that because government use of data mining technology creates enormous risks to civil liberties, proponents of government data mining must bear the burden of proving before implementation that any such program is effective without compromising our civil liberties. Furthermore, EFF is concerned that DHS appears to be focusing on technical protection of civil liberties. EFF believes that technical protections will fail without strong political, legal, and institutional accountability mechanisms.

Privacy and Data Mining Defined

Privacy is not merely about the confidentiality of personally identifiable information possessed by individuals. "Rather, as is well established by United States Supreme Court cases, the Privacy Act, and privacy laws governing the private sector, the concept of privacy extends to information that an individual has disclosed to another in the course of a commercial or governmental transaction and even to data that is publicly available. In these various contexts, privacy is about control, fairness, and

consequences, rather than simply keeping information confidential." [1] Even information that is available publicly or legitimately obtained through a commercial database retains certain privacy protections. [2] But these statutory protections are riddled with exceptions for law enforcement and national security.

The electronic aggregation of this data also affects privacy. There is a "distinction, in terms of personal privacy, between scattered disclosure of the bits of information . . . and revelation of the [information] as a whole. . . Plainly there is a vast difference between the public records that might be found after a diligent search of courthouse files, county archives, and local police stations throughout the country and a computerized summary located in a single clearinghouse of information." [3]

The Government Accountability Office defines data mining as "the application of database technology and techniques – such as statistical analysis and modeling – to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results.'" [4] This is substantially different from traditional information retrieval and reporting: here, expansive *automated* data analysis attempts to identify *non-obvious* patterns and connections within data. Different types of data mining exist, with correspondingly different privacy implications.

One kind is pattern-based searching, which refers to "searches of large databases when the query does not name a specific individual, address, identification number, or other personally identifiable data element, but instead seeks information that matches or departs from a pattern." [5] In other words, pattern-based searches look for unique signatures that indicate a particular type of activity or risk. Another type is subject-based searching, which looks for "information about a particular subject already under suspicion based on information derived from traditional investigative means, whether that subject is represented by a name, a telephone number, or a bank account number." [6] This type of search can also uncover connections between the subject and other individuals, establishing networks of contacts, or compare subjects to determine if they are connected (or even the same person).

Data mining in government today

Data mining programs have a dubious historyl; many have failed to survive public and governmental scrutiny. An expert panel organized under the Department of Defense identified privacy as a major issue inherent in large-scale data mining.[7] Since the emergence of the Terrorism Information Awareness project, which was shut down by Congress[8] and condemned by groups from across the

---

1 James X. Dempsey & Lara M. Flint, *The Future of Internet Surveillance Law: A Symposium to Discuss Internet Surveillance, Privacy & the USA PATRIOT Act: Commercial Data and National Security*. 72 GEO. WASH. L. REV. 1459, 1462 (2004).

Note: Internal citations and footnotes for quotations in this document have been omitted.

2 *Id.* at 1463 ("Arrest records, for example, are publicly available governmental records, but they cannot be used for employment purposes unless they include disposition data. Driver's license data is available for some purposes and not for others. Bankruptcy records are publicly available, but cannot be included in credit reports if they are more than ten years old. Private compilations of publicly available data used for certain commercial purposes are subject to data quality requirements. Individuals are legally entitled to access their credit reports and insist upon corrections, even though none of the data in the reports is confidential and some of it is publicly available.").

3 U.S. Dep't of Justice v. Reporters Comm., 489 U.S 749, 764 (1989).

4 Lee Tien, *Law Review Symposium on Privacy and Surveillance: Emerging Legal Issues - Privacy, Technology and Data Mining*. 30 OHIO N.U.L. REV. 389, 393 (2004).

5 DEMPSEY & FLINT, *supra* note 1, at 1464.

6 *Id.*

7 INFORMATION SYSTEMS ADVANCED TECHNOLOGY STUDY, SECURITY WITH PRIVACY (2002), http://epic.org/privacy/profiling/tia/isat_study.pdf.

8 *Pentagon Terror Spy Lab Closed*, CBS NEWS, Sept. 25, 2003, http://www.cbsnews.com/stories/2003/07/31/attack/main566133.shtml.

political spectrum,[9] many other similar programs have been attempted. Some are now defunct for the same reasons that TIA was dismantled (CAPPS II[10], ADVISE[11]). But more are being started by various executive agencies[12]: those made public include Secure Flight[13] (a replacement for CAPPS II), the Terrorist Surveillance Program, [14] and the DHS' own Automated Targeting System. [15] Due to statutory grants of power under the PATRIOT ACT, the FBI has also taken on a massive program of collecting private data through National Security Letters. [16] Undoubtedly, many secret data mining programs exist as well.

Notwithstanding the concern about terrorism that politically justifies their creation, protection for constitutional rights must be incorporated into any data mining scheme. [17] DHS is taking a step in the right direction by opening this issue to public discussion.

Risks of data mining

Some security experts are concerned about the effectiveness of data mining in specific contexts, such as identifying terrorists. Data mining may uncover patterns that are useful in traditional law enforcement, [18] but appears to be poorly suited to terrorism detection. The effectiveness of data mining depends greatly on the prevalence and typicality of the targeted activity. [19] It is one thing to use data mining to look for common behavior such as credit card fraud or, in the private sector, purchasing preferences; it is quite another to use data mining to identify rare or idiosyncratic behavior like terrorism.). This has been referred to as the "base rate fallacy," and it means that the logical difficulties in creating an effective detection program are vast, if not insurmountable. It is also very difficult to identify a profile that uniquely locates terrorists based on patterns of activity without also misidentifying large numbers of innocent individuals who also match the pattern. [20] As a result, any data mining program targeted at terrorists or other groups with similar characteristics faces substantial problems in positive

---

[9] These groups include the Electronic Frontier Foundation and the Cato Institute. S*ee Total/Terrorism Information Awareness (TIA): Is It Truly Dead?*, http://w2.eff.org/Privacy/TIA/20031003_comments.php; *Information Awareness Office Makes Us a Nation of Suspects*, http://www.cato.org/research/articles/pena-021122.html.

[10] Ryan Singel, *Congress Puts Brakes on CAPPS II*, WIRED NEWS, Sept. 26, 2003, http://www.wired.com/politics/law/news/2003/09/60600?currentPage=all.

[11] *DHS Scraps ADVISE Data-mining software*, ELECTRONIC FRONTIER FOUNDATION, Sept. 10, 2007, http://www.eff.org/deeplinks/2007/09/dhs-scraps-advise-data-mining-software.

[12] *See e.g.* DHS PRIVACY OFFICE RESPONSE TO THE HOUSE (2006), http://www.dhs.gov/xlibrary/assets/privacy/privacy_data_%20mining_%20report.pdf; DHS SURVEY OF DATA MINING ACTIVITIES (2006), http://www.dhs.gov/xoig/assets/mgmtrpts/OIG_06-56_Aug06.pdf; GAO REPORT ON DATA MINING (2004), http://www.gao.gov/new.items/d04548.pdf.

[13] Ryan Singel. *Life After Death for CAPPS II?*, WIRED NEWS, July 16, 2004, http://www.wired.com/politics/security/news/2004/07/64240; *'Secure Flight' Replaces CAPPS II*, WIRED NEWS, Aug. 26, 2004, http://www.wired.com/politics/security/news/2004/08/64737.

[14] James Risen & Eric Lichtblau, *Bush Lets U.S. Spy on Callers Without Courts*, N.Y. TIMES, Dec. 16, 2005, http://www.nytimes.com/2005/12/16/politics/16program.html?_r=1&oref=slogin.

[15] Department of Homeland Security Notice of Privacy Act system of records, 73 Fed. Reg. 123 (2008), http://edocket.access.gpo.gov/2006/06-9026.htm; *See also* EFF COMMENTS TO DHS (2006), http://w2.eff.org/Privacy/ats/ats_comments.pdf.

[16] Eric Lichtblau, *F.B.I. Data Mining Reached Beyond Initial Targets*, N.Y. TIMES, Sept. 9, 2007, http://www.nytimes.com/2007/09/09/washington/09fbi.html.

[17] Senator Leahy expressed such concerns in a letter to Attorney General Ashcroft, available at http://w2.eff.org/Privacy/TIA/leahy-letter.php.

[18] *See e.g.*,Bruce Schneier, *Police Data Mining Done Right*, SCHNEIER ON SECURITY, Aug. 10, 2007, http://www.schneier.com/blog/archives/2007/08/police_data_min.html.

[19] Bruce Schneier, *Data Mining for Terrorists*, SCHNEIER ON SECURITY, Mar. 9, 2006, http://www.schneier.com/blog/archives/2006/03/data_mining_for.html.

[20] Bruce Schneier, *The Problems with Data Mining*, SCHNEIER ON SECURITY, May 24, 2006, http://www.schneier.com/blog/archives/2006/05/the_problems_wi.html.

identification and the avoidance of false positives. [21] Events in the news prove that this risk of misidentification is substantial. [22]

Other problems plague data mining across all its potential uses. As Professor Slobogin explains:

> Most fundamentally, the information in the records accessed through data mining can be inaccurate. The government's no-fly list, for instance, is notorious for including people who should not be blacklisted. Even more prosaic records are astonishingly inaccurate. Approximately one in four credit reports contain errors serious enough to result in a denial of credit, employment, or housing. According to one study, 54 percent of the reports contain personal demographic information that is misspelled, long outdated, belongs to a stranger, or is otherwise incorrect. Even if the information is accurate, integrating disparate databases may lead to distortions in the information obtained, and computers or analysts can misconstrue it.
>
> With event-driven data mining, inaccuracy is heightened by the difficulty of producing useful algorithms. Even when the base rate for the activity in question is relatively high (for example, credit card fraud) and the profile used is highly sophisticated, data mining will generate more "false positives" (innocent people identified as criminals) than true positives. When the base rate of the criminal activity is low (for example, potential terrorists) and the algorithm less precise (as is probably true of any "terrorist profile"), the ratio of false positives to true positives is likely to be extremely high. In fact, what little we know suggests the government's event-driven antiterrorist data mining efforts have been singularly unsuccessful.
>
> The use of algorithms that produce a high false positive rate exacerbates two other phenomena: invidious profiling and what data mining aficionados call "mission creep." Match- and event-driven data mining can be, and probably have been, heavily dependent on ethnic, religious, and political profiling; while such discrimination is a possibility during traditional investigations as well, it is vastly facilitated by computers. And match- or event-driven data mining designed to ferret out terrorists can easily transform into a campaign to grab illegal immigrants, deadbeat dads, and welfare scammers. The CAPPS II program, for instance, appears to have been used to identify any individual who is in the country illegally. The terrorist watchlist has now grown to over one-half million subjects, suggesting a very broad definition of terrorism. These are not necessarily unmitigated harms, of course, but they should be recognized as a likely byproduct of data mining operations.
>
> Erroneous or inappropriate government actions are not the only costs of data mining. Another problem is the threat large databases pose to innocent people's property and livelihood from entities other than the government. The desire for efficient data mining creates pres sure to accumulate all information in one central repository. As Larry Ellison, the head of Oracle, stated, "The biggest problem today is that we have too many [databases]. The single thing we could do to make life tougher for terrorists would be to ensure that all the information in myriad government databases was integrated into a single, comprehensive national security file." That may be true. But a single database makes it all that much easier for identity thieves and mischief-makers (inside as well as outside the government) to do their dirty work because accessing records is that much easier.
>
> . . . Many of those whose records are accessed through data mining don't know it is happening, and if nothing incriminating is found, may never find out. But we still know that data mining allows the government to accumulate and analyze vast amounts of information about us, sufficient perhaps to create what some have called personality or psychological "mosaics" of its subjects. That capacity for data

---

[21] Bruce Schneier, *Terrorists, Data Mining, and the Base Rate Fallacy,* SCHNEIER ON SECURITY, July 10, 2006, http://www.schneier.com/blog/archives/2006/07/terrorists_data.html; *See also* Jeff Jonas & Jim Harper, *Effective Counterterrorism and the Limited Role of Predictive Data Mining*," CATO INSTITUTE, Dec. 11, 2006, http://www.cato.org/pub_display.php?pub_id=6784.

[22] *See e.g.*,:Lara Jakes Jordan, *Prosecutor Flagged by US Terror Watchlist,* ASSOCIATED PRESS, July 14, 2008, http://ap.google.com/article/ALeqM5iBtqltI8L9tX_XoTheJc_XZxtfSgD91TREMO0; *Documents Show Errors in TSA's 'No Fly' and 'Selectee' Watch Lists,* ELECTRONIC PRIVACY INFORMATION CENTER, http://epic.org/privacy/airtravel/foia/watchlist_foia_analysis.html; Bruce Schneier, *U.S. 'No Fly' Curtails Liberties: Intended as a counterterrorism tool, it doesn't work and tramples on travelers' rights,* SCHNEIER ON SECURITY (published in Newsday), Aug. 25, 2004, http://www.schneier.com/essay-052.html.

> aggregation may be a cost in itself. As Daniel Solove has argued, one result of government's entry into the information age is that faceless bureaucrats will be able to compile dossiers on anyone and everyone, for any reason or for no reason at all. The possibility, even if slim, that this information could somehow be used to our detriment or simply revealed to others can create a chilling effect on all activity. It may have been some vague sense of this possibility that led Congress, however ineffectually, to declare its opposition to the concept of Total Information Awareness, with its epithet "knowledge is power."
>
> . . . [W]hen [privately compiled information] information ends up in the hands of the government, with its enormous power to deprive people of liberty and property and the wide range of behavior that can be considered grounds for such deprivation, the calculus arguably changes even for the completely innocent. Knowing that the government is obsessed with fighting terrorism (as perhaps it should be) and that it views data mining as an essential tool in that fight, one could be forgiven for feeling inhibited about making certain calls (to a Muslim acquaintance?), traveling to certain locations (the Middle East?), and buying certain items (Halal meat, literature criticizing the war?).. [23]

Current Statutory and Constitutional case law inadequately address the problems of data mining

There is currently no comprehensive framework for regulating data privacy. The private sector is governed by a piecemeal collection of subject-specific privacy laws. "Ironically, while some proponents of governmental use of commercial data have argued that the government should have the same access to consumer data that the private sector has, the private sector is actually subject to clearer and stricter rules for the use of data under current law than government counterterrorism agencies. Governmental officials defending plans to use commercial databases for counterterrorism purposes have argued that all such uses will be in strict compliance with applicable privacy laws. Such assurances are misleading, however, because there are very few privacy laws applicable to the government's acquisition and use of commercially compiled data for counterterrorism purposes." [24]

These privacy laws include the Fair Credit Reporting Act, Right to Financial Privacy Act, Health Insurance Portability and Accountability Act, Family Educational Rights and Privacy Act, and Electronic Communications Privacy Act. However, all these privacy laws include exceptions for information under a grand jury subpoena, and most are full of other exceptions. These include FBI National Security Letters and special exemptions for intelligence agencies. [25]

Furthermore, the application of the Privacy Act of 1974 (establishing rules for federal governmental records such as notice and consent for information collection and sharing) to "governmental use of commercial databases is limited, and there exist major exceptions for law enforcement and intelligence agencies. . . The Privacy Act does include a provision that extends its coverage to databases created under government contract, but it seems that the provision does not include governmental searches of private sector databases already compiled and maintained for other purposes." [26]

---

[23] Christopher Slobogin, *Government Data Mining and the Fourth Amendment*, 75 U. CHI. L. REV. 317, 327 (2008).

[24] DEMPSEY & FLINT, *supra* note 1, at 1471.

[25] *Id.* at 1472. Also, "even when information is pulled into governmental databases, law enforcement and intelligence agencies are exempt from many key provisions of the Privacy Act. Simply by publishing a notice in the Federal Register, law enforcement agencies and the Central Intelligence Agency ("CIA") can exempt their records from the Act's requirements that records be maintained accurately and that individuals be permitted to access and correct their records. Any agency can share its records with any other agency if the sharing is a "routine use" and has been noticed in the Federal Register. . .Certainly, this allows all agencies involved in counterterrorism to share information. The definition of "computer matching" excludes matches performed for foreign counterintelligence purposes. Finally, any agency can disclose records to any other federal, state, or local agency for any law enforcement activity upon written request specifying the particular portion desired and the law enforcement activity for which the record is sought." *Id.* at 1475.

[26] *Id.* at 1471. Also, note that originally "Congress passed the Privacy Act in response to concerns about the creation of large, centralized governmental databanks of personal information. The Act's protections apply only where the government is creating a 'system of records.' Counterterrorism uses of commercial information may not involve the

The biggest problem for privacy, however, may be the fact that no statutory protection exists for vast amounts of data that the government can obtain from third parties if they voluntarily provide or sell the data:

> Because the United States has no comprehensive privacy law applicable to commercial databases, the analysis of rules concerning commercial data must start with a presumption of access - so long as no law prohibits it, the government can purchase or request voluntary disclosure of any commercially held records. Especially since September 11, the FBI has obtained commercial databases from private entities, from grocery store frequent-shopper records to scuba diving certification records, without having to exercise any compulsory authority. So long as no statute prohibits government access to the information, a voluntary request is entirely legal. Third parties that hold consumer information often comply with such requests because they want to be helpful to the government or because compliance seems to be the path of least resistance. Categories of information for which there is no applicable privacy law include, inter alia: travel records, retail purchases - online and offline - of anything ranging from books to groceries, "Easy Pass" toll records, real estate and mortgage information, magazine subscriptions, club memberships, and utility bills.
>
> In addition to the ability to compel disclosure of records compiled by businesses for business purposes, the government's power to require businesses to create and report data regularly is expanding. For example, anti-money laundering laws, including the Bank Secrecy Act (enacted in 1970 and expanded by the USA PATRIOT Act), impose extensive reporting requirements on the financial industry, resulting in the availability of far more financial information to law enforcement entities. The Bank Secrecy Act requires financial institutions to report various types of transactions to the government and also imposes certain recordkeeping and record retention requirements on financial institutions. Similar reporting requirements have been imposed on universities with respect to foreign students. [27]

Once these reams of information are in government hands, there are few restrictions on usage or sharing. The USA PATRIOT Act has greatly facilitated information sharing among law enforcement and intelligence agencies. [28] But the Act did not mandate meaningful protections against the enormous privacy risks such information sharing creates. This is a significant departure from the past, when both criminal justice and intelligence organizations had to follow strict rules for obtaining information (which was often directly obtained from the subject of the investigation, pursuant to the relatively high standard of a search warrant) and sharing it. [29] The scope of such practices is now is much wider, permitting “the sharing of 'foreign intelligence,' 'counterintelligence,' or 'foreign intelligence information' obtained in criminal investigations with 'any other Federal law enforcement, intelligence, protective, immigration, national defense, or national security official.'” [30]

It is disturbing that the PATRIOT Act permits the “sharing of a vast array of information that is not related to international terrorism, without regard to whether that information concerns legal or illegal activities. Considering that criminal investigative techniques, especially wiretaps and grand jury subpoenas often produce substantial amounts of private and sensitive information on persons who are not subjects of an investigation and are not involved in any illegal activity, this provision represents an unprecedented expansion of the authority of intelligence agencies to obtain information on American

---

creation of governmental databases covered by the Act because searches and data analysis can be conducted in such a way that the data never leaves private hands. Government agencies can secure (by contract or otherwise) various scans of data held by private corporations without pulling that data into a centralized governmental database. If the government is simply accessing databases created by commercial entities for their own reasons, there may be no system of records subject to Privacy Act requirements.” *Id.* at 1474.

27 *Id.* at 1482-83.

28 *Id.* at 1468.

29 *Id.* at 1483.

30 *Id.* at 1484.

citizens and permanent residents within the United States." [31] In fact, one provision of the Act requires the Attorney General and the heads of law enforcement agencies to automatically turn over any foreign intelligence information they encounter (not limited to terrorism, and including information about lawful activities). [32]

It is clear that government agencies have unprecedented access to information with very little statutory supervision. Due to this, and in light of the myriad privacy risks inherent in data mining activities, agencies must take upon themselves the responsibility for protecting individuals' civil liberties without waiting for Congress or courts to take action.

What rights to information privacy does the Constitution protect?

"Data mining possibly implicates at least three Constitutional provisions: the Due Process Clause's guarantee of fair process, the First Amendment's protection of speech and association, and the Fourth Amendment's prohibition on unreasonable searches. The Due Process Clause might require that government make a good faith effort to secure its databases and that it provide some sort of procedure for challenging erroneous inclusion on no-fly lists and other databases used in match-driven surveillance when such surveillance results in deprivations of liberty or property. The First Amendment's application to data mining is more complicated. It has been argued, on the one hand, that commercial data brokers' speech rights are infringed by rules inhibiting disclosure of the information they acquire and, on the other, that the First Amendment provides special protection for any personal information that evidences one's political views or associations." [33]

In theory, the strongest protection for privacy is enshrined in the Fourth Amendment's prohibition on unreasonable search and seizure. Current constitutional jurisprudence has evolved with respect to the search clause, which "is sensitive to modern privacy concerns by extending constitutional protection to situations that satisfy the reasonable expectation of privacy test." [34] But the seizure of data is a different matter, creating severe problems for privacy. Professor Paul Ohm summarizes the Supreme Court's current standards for search and seizure of intangible property such as electronically stored personal information:

> While imperfect, the evolved Search clause has kept the protections of the Fourth Amendment relevant in an age of digital evidence, ubiquitous communication networks, and increasingly sophisticated and invasive surveillance capabilities. . . In contrast, the Seizure clause [has been]. . . consistently interpreted to protect only physical property rights and to regulate only the deprivation of tangible things . . . Under the modern interpretation of the Fourth Amendment, the government seizes property only when it 'meaningfully interfere[s]' with a 'possessory interest.' Courts have not articulated precisely what is meant by this phrase, but cases like those cited above suggest a physical property-centric model of dispossession. [35]

This inconsistency creates significant problems for privacy interests, because:

> . . .we live in an atoms-before-bits world. At least historically, the government has had to deprive a person of physical property (even if momentarily) before it could make a copy of his intangible

---

[31] *Id.*

[32] *Id.* at 1485.

[33] SLOBOGIN, *supra* note 23, at 328. For more discussion of First Amendment protections in data mining, especially concerning dangers to freedom of association, *see* Katherine J. Strandburg, *Freedom of Association in a Networked World: First Amendment Regulation of Relational Surveillance,* 49 B.C. L. REV. 471 (2008).

[34] Paul Ohm, *The Olmsteadian Seizure Clause: The Fourth Amendment and the Seizure of Intangible Property,* 2008 STAN. TECH. L. REV. AT 1 (2008).

[35] OHM, *supra* note 34, at 1.

property. . . [I]n order to copy the bits from a hard drive, the government must open the plastic or metal case of the computer containing the hard drive. . .[E]ven in purely virtual settings, private "spaces" are delineated by virtual walls and doors, things like passwords and computer dialog boxes. . . The reasonable expectation of privacy test recognizes virtual as well as physical doors and locks.

. . .[Also], a lot of intangible property is held by third party intermediaries . . . Intangible data tends not to be left lying around in publicly available spaces. Of course, this reliance on third party intermediaries makes it much easier for the police to access my data, and to do so without my knowledge. . . [But] profound shifts in technology will end the happy confluence of search and seizure. Because of these shifts, the police will much more frequently grab intangible data without having to handle atoms, pass through virtual walls, or deal with intermediaries. When they do so, courts may reason that the police aren't intruding on a reasonable expectation of privacy.

Search has proved the bulwark protecting us from needing to decide whether intangible copying is seizure. Developments in technology are putting pressure on this bulwark. . . First, intangible data is increasingly stored in virtual spaces that one can access without intruding on any physical space. No longer will the police need to lift the record player to read the serial number; the device itself will report its serial number when asked. Second, getting access to intangible information like a serial number often can be done without even needing to move through a virtual door. Tools can get at data directly from storage, ignoring any passwords that may have been in place.

. . . The problem is that once intangible property loses the protection of the search trio, the government may be tempted to engage in a form of Constitutional gamesmanship: The reasonable expectation of privacy test turns, as a threshold matter, on an intrusion or invasion. What if a packet sniffer is configured to seal its information away from police review until it is unlocked after judicial authorization? What if the image of the hard drive is stored on media that is locked inside a cabinet at police headquarters? Under the expectation of privacy test, the government has a reasonable argument that these actions are not yet invasions or intrusions of any privacy. As long as none of the collected information is reviewed immediately, and so long as review requires an additional, tamper-evident step, the police can claim that they are not conducting a Fourth Amendment search. Unless and until the stored information is "exposed" in some way to a human being, the police will argue, the search is merely contingent, not completed. [36]

Professor Ohm raises serious questions about the meaning of information privacy in a world where so much of what people consider private is either easily copied and retained by law enforcement (thus evading the application of current Fourth Amendment standards with respect to seizure) or stored by third parties (which usually means that government has access, since courts tend to consider such information voluntarily disclosed). [37]

The Fourth Amendment needs to be reconsidered in light of the new ways in which information is stored and communicated. Among other things, it should "protect [the] 'right to destroy' [property that belongs to you] or, in the computer context, 'right to delete' by its terms through its prohibition on unreasonable seizure. . .There is a long tradition of recognizing the right to destroy in property law. As Lior Strahilevitz has discussed, at various times in legal history courts have identified the right to destroy property as one of the 'bundle of rights' intrinsic to physical possession." [38] Seizure has

---

[36] OHM, *supra* note 34, at 8-12.

[37] For more on the problems of third-party information storage and court views on voluntary disclosure, *see* Catherine Crump, *Data Retention: Privacy, Anonymity, and Accountability Online*, 56 STAN. L. REV. 191 (2003)(her primary argument is that courts will consider transactional information voluntarily disclosed rather than private if it meets certain conditions: third-party possession, voluntary transfer, and awareness that it can be recorded; these are unavoidable in most modern transactions). *See also* SLOBOGIN, *supra* note 23, at 330 ("Since virtually all information obtained through data mining comes from third party record holders—either the government itself, commercial data brokers, or a commercial entity like a bank—its acquisition does not implicate the Fourth Amendment.").

[38] OHM, *supra* note 34, at 14. He also discusses the social importance of such protections: "the right to delete assures computer users that their words can be in some sense undone. This provides a sense of privacy that may lead to more candor in discussing sensitive matters electronically, and the increased candor benefits all of society, not only the owners

historically ended when an item is returned to its owner. Since technology makes this no longer necessarily true, law enforcement can potentially continue violating one's expectation of privacy and right to control one's property by keeping your information after the original reason to obtain it expires. It is a much safer policy to prevent the potential for such abuse through robust data destruction policies.

The Fourth Amendment, whichis rooted in the Framers' experiences with practically limitless British general warrants, also raises direct problems with the data mining methodology:

> . . .[D]espite the apparent force of the "knowing exposure" doctrine, the mere fact that things or information have been exposed to others does not automatically mean that the Fourth Amendment becomes irrelevant. One's words in a conversation are by definition exposed to the listener, but such private communications receive full Fourth Amendment protection. . . Moreover, data mining does not only challenge substantive values like privacy; it also challenges procedural values like particularized suspicion. As [a] TAPAC Report noted, '[t]he common feature of all of these programs is that they involve sifting through data about identifiable individuals, even though those individuals have done nothing to warrant government suspicion, in search of useful information.'
>
> . . .The courts have long recognized. . .that there is a "vital relationship between freedom to associate and privacy in one's associations." This concern is especially great in the counter-terrorism context, because "[n]ational security cases . . . often reflect a convergence of First and Fourth Amendment values not present in cases of 'ordinary' crime. . ." [39]

Moreover, data mining can easily circumvent existing privacy protections, depriving individuals of the constitutional protections that have traditionally restrained law enforcement. For example, if a data mining program can use pattern- or subject-searching to track an individual's activities and movements, an agency may be able to find the same information that GPS or wiretap data obtained with a warrant could provide without meeting the accompanying standard of proof and being subjected to independent review. [40]:

> The Fourth Amendment search warrant procedure can be viewed as a more general accountability model with several key default parameters. First and foremost is the need for particularized suspicion: there must be facts that demonstrate a good reason to search this person, place, or thing. Second, this factual justification must meet some standard of certainty or likelihood, e.g. "probable cause." Third, the warrant itself must describe with particularity the scope of the search. Fourth, there must be an independent check, e.g., the requirement of a neutral and detached magistrate, which ensures the objectivity of the determinations that the justification exists, that it meets the requisite certainty standard, and that the scope of the search is objectively defined. Fifth, this independent assessment should take place before the government conducts its search.
>
> . . .The use of patterns discovered through data mining raises similar particularity issues. Imagine a database of a million people and a hypothesis that those who meet certain criteria are highly likely to be terrorists. But you don't know whether any of these million people actually do meet these criteria; if you did, you wouldn't need to run the search. The basic problem is lack of particularized suspicion: data about these persons would be "searched" without any reason to believe either that the database contains evidence of terrorist activity or that any person "in" the database is a terrorist. Like eavesdropping, pattern-oriented data mining (or automated data analysis) by its very nature involves broad intrusions on privacy, and demands careful attention to particularity.
>
> Even when automated data analysis is subject-oriented-as when the government is investigating a particular suspect or incident-particularized suspicion remains a problem. If the government has reason

---

of the data. . . A Fourth Amendment right to delete explains the reasoning and conclusions of the wiretapping courts. Although a wiretap does not dispossess me of my words, once it records my private conversation, my words have been in a sense taken from me — the wiretap deprives me of the ability to conceal or otherwise destroy those words." *Id.*

39 TIEN, *supra* note 4, at 399.

40 *Id.* at 400.

to believe that "John Smith" is a terrorist, it has particularized suspicion as to him. If the government reasonably believes that someone who uses a particular phone number or email address is a terrorist, again there is some particularized suspicion.

But how far does that particularized suspicion get you? Link analysis, for instance, focuses on the transactional connections between the subject and other people: who lives with John Smith, who corresponds with johnsmith@aol.com, and so on. Does the mere fact that Jane Doe has certain connections to John Smith mean that there is particularized suspicion as to her? Under current law, that seems highly unlikely.

. . .[Current case law would not allow you to conduct a] simple link analysis that begins with a person for whom particularized suspicion exists and then follows links to other persons who merely have associated with the original suspect. Accordingly, the TAPAC Report recognized that "the power of data mining technology and the range of data to which the government has access have contributed to blurring the line between the subject- and pattern-based searches . . . [e]ven when a subject-based search starts with a known suspect, it can be transformed into a pattern-based search as investigators target individuals for investigation solely because of their connection with the suspect."

The particularity problem doesn't end with particularized suspicion, of course. Even with particularized suspicion, the search itself must have a particularized scope; the searcher's discretion must be limited and defined. Under the warrant requirement, the places or persons to be searched must be particularly described so that the magistrate can determine on the record that there is an objective justification to search those places or persons. This requirement does not make automated data analysis impossible, but it does require considerable safeguards.

. . .The inherent particularity problem is that the government is running queries, whether subject- or pattern-oriented, over a database or set of databases containing personal information about many people who are not suspected of anything. Indeed, DARPA eventually admitted that the data mining originally contemplated by TIA "must inevitably lead to 'fishing expeditions' through massive amounts of personal data and a wholesale invasion of Americans' privacy that yields, basically, nothing in terms of finding terrorists." [41]

Current Fourth Amendment jurisprudence also uses too narrow a concept of "search." [42] If privacy is to be protected against new technologies, our understanding of the Fourth Amendment must evolve to control the novel risks they pose . In the famous wiretapping case of *Katz v. U.S.*, [43] the Supreme Court recognized this very point in acknowledging that privacy must extend to conversations in a phone booth. They noted that the Fourth Amendment protects people, not just places, rejecting the antiquated standard under *Olmstead*. [44] In order to preserve these Fourth Amendment principles today, when complex patterns of private information undetectable by humans can be revealed by computers engaged in data mining, we need a more expansive definition of what constitutes a search:

. . .data mining is a search, even when the government has lawfully acquired the individual facts in the database being mined, because the patterns or inferences discovered via data mining often deserve to be private and go beyond the information that can fairly be said to be "knowingly exposed" to others. This claim may sound preposterous. Isn't all the information in the database already completely exposed?

Yes and no. As noted earlier, the data-mining literature distinguishes between traditional "query and report tools," which describe what is in a database, and "true" data mining, which identifies "valid, novel, potentially useful and ultimately understandable patterns in data." Because "such patterns are themselves knowledge," it is completely reasonable to say that finding such patterns or relationships exposes information that is not "in" the database. After all, since 9/11 we have been engaged in a public debate about needing to do a better job of "connecting the dots." Merely having or knowing individual

---

41 *Id*. at 404.
42 *Id*. at 408.
43 389 U.S. 347 (1967).
44 Olmstead v. United States, 277 U.S. 438 (1928).

facts is one thing; discerning patterns or relationships within or among facts is another.

. . .Consider, for instance, a database of purchases created by aggregating bank and credit-card transactional records. Inspecting this database reveals the following patterns: Jane Doe had bought birth control pills every 3 months for years; during this period, she began buying bridal magazines and then changed her name and established a joint checking account with someone who is apparently her husband; at some point, she stopped buying birth control pills; after about a year, she began buying pregnancy test kits; and then at some point she stopped buying pregnancy test kits and began buying baby clothes and other things at "Babies 'R' Us."

The story that these patterns reveal is obvious: she got married, intended to have a child, and is now expecting. Other inferences are possible, but considerably less plausible. For my purposes, the question is whether she can be said to have "knowingly exposed" the "fact" that she is pregnant. It's not as though the transactional records that make up the database include the results of her home pregnancy test or her doctor's notes confirming her pregnancy. The single most probative transaction in the database-that she bought a home pregnancy test kit-is perfectly consistent with the alternative hypothesis that she was simply buying it for a friend or relative.

My point is simple: because the fact "Jane Doe is pregnant" is not actually in the database, we must question whether she knowingly exposed the fact of her pregnancy to anyone purely by virtue of her having made these purchases. And if she did not "knowingly expose" the fact of her pregnancy, then this fact should remain the subject of Fourth Amendment protection.

. . .[Case law indicates that] it impossible to say that it cannot be a search to discover "hidden" information "inside" something legitimately possessed. Any impediment to perception, from deliberately placed containers to the mere physical placement of a turntable to the need for specialized equipment in order to extract information, can in the right circumstances transform government acquisition of information into a search." [45]

. . .This reasoning suggests a more general point about the Fourth Amendment treatment of new information technologies that raise privacy concerns. We typically focus on how technology enhances our senses to enable more efficient or extensive collection of information, either by amplifying the range or scope of our natural senses or creating new, somewhat analogous senses. . . But sensation or information collection isn't everything when it comes to privacy invasion. We should think of these kinds of privacy invasions, as well as scientific or device-enabled analysis – whether of bodily fluids or of databases – as "cognition-enhanced" searches that deserve their own Fourth Amendment jurisprudence.[46]

The purpose of data mining itself is to uncover connections and patterns within information sources that are *not* obvious or otherwise detectable. In *U.S. Department of Justice v. Reporters Comm. for Freedom of the Press*, the Supreme Court recognized "a privacy interest in the practical obscurity that we have reasonably come to expect in scattered bits of information."[47] The government should not be able to destroy this veil of privacy without meeting Fourth Amendment standards. Otherwise,

---

[45] TIEN, *supra* note 4, at 411. Also, "the most pertinent line of cases, however, involves scientific testing or analysis of substances. In [certain recent] drug-testing cases [citations omitted], the Court has found that urine testing involves two separate searches: the initial collection of the urine, and the subsequent chemical analysis of the urine. The urinalysis is a search because "chemical analysis of urine, like that of blood, can reveal a host of private medical facts about an employee, including whether she is epileptic, pregnant, or diabetic." By contrast, the chemical analysis that identified an unknown powder as cocaine in United States v. Jacobsen was not a search. The crucial point for the Court was that the test "could disclose only one fact previously unknown to the agent-whether or not a suspicious white powder was cocaine. . ." Thus, the test "compromises no legitimate privacy interest" because a positive finding would only reveal that the substance was cocaine – contraband in which no person can have a legitimate Fourth Amendment interest – while a negative finding would "merely disclos[e] that the substance is something other than cocaine," but "no other arguably 'private' fact . . . ." [F]rom this perspective, data mining is a lot like urinalysis. Data mining that discovers novel patterns discovers knowledge that is not, strictly speaking, "in" the database. Or that knowledge is "in" the database in the same way that private medical facts about a person's being epileptic, pregnant or diabetic are "in" her urine. They are not on the surface or exposed to view; more is required to get at them." *Id.* at 413.

[46] *Id.* at 413.

[47] *Id.*

“cheap surveillance 'simply makes it too easy, without the loss of a lot of shoe leather and the other costs police traditionally have had to take into account in determining the realistic limits upon their enforcement activities, to engage in random and wholesale snooping. Posnerian cost-benefit balancing is thus no longer a sufficient deterrent to such enlarged investigative strategies, and this is precisely why this activity needs to be brought within the purview of the [F]ourth [A]mendment.'” [48]

Recommendations

We have illustrated the great risks that incorrectly designed and circumscribed data mining programs pose to everyone's constitutional rights. Legislative and judicial engagement with the meaning of our rights in the context of new technologies and social information practices is still inadequate. It is thus the responsibility of each agency that develops data mining tools to seriously consider the potential dangers it creates, and to design systems to proactively protect privacy. Many of the following suggestions are based on long-standing information-handling principles, such as the “Fair Information Practices” underlying many privacy reforms dating back to the 1970s.[49]

Control over information and access

The first line of defense for privacy is regulating the kind and quality of the information the government is allowed to receive or mine. Decision-making about the sources of information used for data mining purposes should be as public as possible. A major reason for this is quality control. As mentioned earlier, many databases (even commercially managed ones, which are in widespread use by the government) contain enormous numbers of errors. “As the president of one data mining company said, 'the quality of the prediction is directly proportional to the quality' of the data. . .[D]ata quality has Constitutional implications where information systems are used by law enforcement officials to arrest or detain individuals and in other situations where the government makes decisions adverse to individuals.” [50] Data sources subject to public scrutiny, independent testing and robust procedures for individual correction requests should be strongly preferred. For data sources obtained from other government agencies or which are otherwise confidential, similar standards should be adhered to internally to ensure their trustworthiness. It is especially crucial that individuals be allowed access to the raw data about themselves for accuracy-checking and accountability. .

Second, no more information should be collected than is needed to accomplish the task of data mining. In other words, the amount of data placed into the system should be the narrowest slice of information with which the system can be effective. Statutory authorization should exist for using any potential data source. Special standards should be established for categories of especially sensitive information, such as political or religious affiliation, speech activities, and medical data. [51] Any statutory limitations must also be carefully followed; in many cases, Congressional authorization for collection of private data is only permitted for narrow kinds of use (usually counterterrorism).

“One way to implement the collection limitation principle is by leaving commercial information in the hands of data aggregators and having them respond only to specific queries. Under such an approach, the government would not be acquiring the databases. For any given search, the government

---

48 *Id.* at 415.

49 DEMPSEY & FLINT, *supra* note 1, at 1489.

50 *Id.* at 1492.

51 “One possible source of appropriate guidelines is the Criminal Intelligence Systems Operating Policies promulgated by the DOJ, which state: [Law enforcement] shall not collect or maintain criminal intelligence information about the political, religious or social views, associations, or activities of any individual or any group, association, corporation, business, partnership, or other organization unless such information directly relates to criminal conduct or activity and there is a reasonable suspicion that the subject of the information is or may be involved in criminal conduct or activity.” *Id.* at 1491.

would acquire only the data that matches its search terms (the 'hits').” [52] Anonymization and encryption techniques can protect sensitive government query terms from being disclosed to private companies during this process.

Modern law enforcement creates a special problem because personal data is often transferred to the government, where it is usually copied to ensure security and integrity. That data was initially acquired in a lawful manner cannot automatically justify its indefinite retention; its continued possession by the government should be presumed to be a continuing harm to privacy. In any case, to feed it into a data mining system is equivalent to granting an unconstitutional general search warrant, allowing continued search of individual property without probable cause. As a general matter, information should be retained for “no longer than is necessary for the purpose for which it was collected. . . Purging data significantly reduces the opportunity for abuse.”[53]

Information sharing between enforcement and intelligence agencies poses another challenging problem. Without appropriate limitation, such sharing can violate statutory protections and increase the risks to individual privacy. This is especially important considering the sensitive nature of information agencies often share. Rather than wholesale information sharing, information should be shared only when certain criteria are met by the requesting party.

Most importantly, any usage limitations on data must be carefully preserved as a condition of transfer. If an agency, for instance, was Congressionally authorized to obtain certain private information for counterterrorism purposes and is permitted to share it with another agency, the receiving agency must also only use it for that purpose. Agencies requesting information should also demonstrate a legitimate need, and be provided only as broad a set of information as is necessary to their legitimate purpose; in other words, the same requirements should apply that agencies must satisfy in originally obtaining information. When sharing results of data mining, additional restrictions should take account of their often speculative nature. Tenuous inferences should not be used as solid facts in another search, which would greatly increase the risk that innocent people will be wrongly targeted.

Appropriately strict standards of review for different kinds of data mining

Data mining is in tension “with 'the Constitutional presumption of innocence[,] the Fourth Amendment principle that the government must have individual suspicion before it can conduct a search'. . .[, and potential] psychic harms caused by government scrutiny of innocent people.” [54] Furthermore, there is the well-publicized danger of false positives. Two types of technical safeguards may help address these tensions.

> **Rule-based processing.** Rules should be built into data search queries to ensure that results are tailored to the analyst's authorization. For example, search queries could carry information about the type of permission that the analyst has been granted, or the system could ask an analyst for additional proof or authorization before sharing certain kinds of results. . . Additionally, "data labeling" may be used to describe how data should be accessed. Metadata may be included that summarizes the information, its source, and even the reliability and age of the source. Information might be accessed differently if an analyst is advised that the data relate to a US citizen, rather than a foreign person, and it might be treated as more or less reliable depending on where the information came from and how recently it has been verified. This final point raises the important issue of data quality. As the DHS Privacy Office has argued, strong data quality standards should be adopted for all information used in data mining.

---

[52] *Id.* at 1490.
[53] *Id.* at 1492.
[54] Ira S. Rubinstein, Ronald D. Lee, & Paul M. Schwartz, *Data Mining and Internet Profiling: Emerging Regulatory and Technological Approaches,* 75 U. CHI. L. REV. 261, 263 (2008).

Anonymization and selective revelation. With the goal of minimizing the amount of personal information revealed in the course of running pattern-based searches, the anonymization of data (such as names, addresses, and social security numbers) is essential. The disclosure of personal information would occur only after the "sanitized" pattern-query results establish a reason to pursue further investigations of a subset of the original pool of individuals. Even then, identifying information would only be selectively revealed. Access to additional personal details would require even further narrowing of the searches, independent authorization, or a combination of the two. The Markle Foundation, for example, proposes that "personally identifiable data can be anonymized so that personal data is not seen unless and until the requisite showing (specified in guidelines) is made." [55]

These measures may help ensure compliance with the suggested substantive requirements concerning data usage and sharing, and should be tested publicly. Rule-based processing may create a technical barrier to stop agencies from unauthorized usage of data sources by permitting fine-grained control of standards for access to sensitive information. This is one potential solution, for instance, to the danger of improper use of copied information discussed above.

Selective revelation may be even more useful because of its role in protecting the Fourth Amendment requirement of particularity. First, particularized suspicion must be required before broad searching data on innocent individuals is permitted.[56] Such standards must also limit the scope of data mining even when some particular suspicion exists, given the ease with which a pattern search can identify targets for subject searches, who in turn through link analysis provide additional targets for investigation (testing the limits of the particularity requirement). Here again, however, it cannot be assume that such schemes will in practice protect civil liberties; research and public testing are necessary.

Technical privacy protections may help address some of the civil liberties risks of data mining. A system where privacy is the default and exception rules can be individually identified creates a better environment for Congress, other regulatory forces, and to a limited extent the public to evaluate whether data mining oversteps statutory or Constitutional rights.

Making data mining systems and operators accountable

A central concern with respect to data mining, as with other secretive or invisible government activities, is that there appear to be no effective checks on the power to use it. The substantive and technical protections discussed above cannot be guaranteed if all discretion and control remains with the analysts who will be using the system. [57] Controls and accountability mechanisms must exist both

---

[55] *Id.* at 268. *See also* Bruce Schneier, *Privacy-Enhanced Data Mining*, SCHNEIER ON SECURITY, June 20, 2006, http://www.schneier.com/blog/archives/2006/06/privacyenhanced.html.

[56] One important standard for determining the necessary level of proof is the severity of the potential consequence from using that information. If potential next steps include deportation or arrest, the revelation standards ought to be higher. Another standard to incorporate is the sensitivity and privacy interest in a piece of information. Professor Slobogin provides an interesting analysis on the issue of this privacy interest, derived from an evaluation of people's reasonable expectations of privacy with respect to different kinds of data. He presents a “social networking”-based model of privacy, where expectations are defined by social practices relating to how far information actually travels. This is a much more sensible approach than the current standard in courts (the unrealistic “voluntary disclosure” standard). *See* SLOBOGIN, *supra* note 23, at 331.

[57] This is because "traditionally, the act of and consequences of searching created their own public accountability. Knowledge of searches is relatively public when government physically searches homes or persons. . . Moreover, searches have traditionally been used to combat ordinary crime, meaning that prior judicial review is often augmented by after-the-fact judicial review on motions to suppress. . .Visibility is a powerful regulatory tool.' But modern "search" activity is far less visible to us; electronic surveillance easily operates without a target's knowledge. We will only know about wiretapping – or data mining and automated data analysis – if the government decides to tell us about it.” Furthermore, the technology of data mining itself raises even greater transparency problems, especially without proper recordkeeping (or using neural

inside and outside agencies with data mining programs. “Internal controls need to be developed and adhered to, as well as steps taken to promote a culture of professionalism among the analysts. The inspectors general of agencies that engage in data mining could play an important role in oversight of internal controls,” [58] and there should be a high-level official positions regulating privacy and information security.

To ensure effective oversight, systems such as audit trails (which record what actions are taken on the system, by whom, pursuant to what authorization, etc.) are absolutely necessary. They are the building blocks with which effectiveness, as well as abuses of discretion, of data mining programs can be monitored and reported on. This tracking provides corollary benefits: it is undoubtedly useful to intelligence analysts seeking to understand, among other things, what kinds of queries are effective.

Data mining programs must also be accountable to other governmental bodies. Congress must receive regular reports from agency heads on effectiveness and problem areas in a given program. Congress should also grant review powers for these programs to impartial experts, so that multiple groups provide input and review. “In addition, there should be regular public reports describing the nonclassified aspects of any data mining program. Finally, the government should develop standards for the validation of models used in data modeling and of the results of these programs. As the DHS Privacy Office observes, 'Just because a pattern exists in the data does not mean that the pattern is meaningful or valid.' The need is for independent validation of the model's predictive accuracy.” [59] Open reporting and the development of standards helps both citizens and intelligence agencies by allowing the input of numerous experts in data mining who can raise issues of efficacy or privacy.

Another target for regulation helpful to both the public and the intelligence community is data quality and standardization. As already discussed, the quality of input data has serious effects on the functionality and arguably even the legality of a data mining program whose data can be used as justification for adverse government action. Presumably, the diverse sources with which data mining systems must cope create problems of compatibility and accuracy. Open discussions involving government officials, academics, and technology experts on information storage frameworks could provide solutions to these problems, and present another opportunity to ensure adequate security and privacy measures.

Agencies involved in data mining should also strive to provide notice to citizens. This is another tenet of Fair Information Practices, the parameters of which are described by Professor Slobogin:

> A fundamental principle of fair information practices is that individuals should have notice both of the fact that information is being collected about them and of the purpose for which it is being collected. Notification prior to data collection affords individuals the opportunity to choose not to disclose the information (albeit often at the cost of foregoing the opportunity to engage in the transaction that is made conditional upon disclosure of the information) and is a precondition to the individual's ability to ensure the collected data is accurate. . . Notice is an element of Fourth Amendment searches, but when law enforcement and intelligence agencies are collecting information from third parties in the context of an investigation focused on specific individuals, it may be desirable to delay notice to those individuals that they are the subjects of an investigation. However, it is not automatically apparent that the same constraints must apply when the government seeks to use commercial and governmental databases for the newer screening uses. When the government uses data for screening purposes, all individuals are subject to the same scrutiny. The government can thus give notice of the types of information that it is collecting or accessing and the ways in which it plans to use the information without telling suspected terrorists what is known about them or even that they are suspects. Notice may actually improve the effectiveness of the data collection effort. For example, if an individual is notified when she books a plane ticket that the airline is collecting particular pieces of information to pass along

---

network technologies, where meaningful reporting or explanations are nearly impossible). TIEN, *supra* note 4, at 406.

[58] RUBENSTEIN ET AL, *supra* note 54, at 270.

[59] *Id.*

to the government to review for security purposes, she may be more likely to provide accurate responses or volunteer information to explain discrepancies. A terrorist may try to evade the system by giving false information, but that is a problem that the system must be designed to address regardless of whether notice is given.

Even when the government uses commercial data in criminal justice or intelligence investigations, it can give the public general notice of the types of commercial information it uses and the ways it uses them without compromising specific investigations. The notice principle also can be partly implemented by notice to Congress of data analysis practices.

Conclusion

These comments outline the main risks of data mining to privacy and other civil liberties: some are inherent in mining methods, and some may be avoided through careful regulation and public accountability. Ultimately, however, we doubt that the civil liberties dangers of data mining can be controlled, at least in today's political environment. The fundamental problem is that the civil liberties threat comes from the government itself — which holds the data, operates the hardware and software, and employs the human users and operators of the systems.

Sent: Wednesday, July 16, 2008 8:56 AM
To: Privacy Workshop
Subject: Data Mining Workshop, Docket No. DHS-2008-0061

Advances in computer technology, particularly those that track people's use of the Internet, raise serious public concerns about personal privacy. These concerns are finding their way into the national culture, including popular fiction. For example, in the novel *On the Lip*, http://www.websurferusa.com, a gradulate student invents software that not only tracks where websurfers have been, it also predicts where they will go on the Web and in the brick-and-mortar world. Fearing there is no one else he can trust, the inventor brings his college surfing buddy to Washinngton to help prevent use of his invention for domestic spying.

Tom Rey
Pacific Beach, CA