2006

Lost in Translation? Data Mining, National Security and the Adverse Inference Problem

Anita Ramasastry
University of Washington School of Law

LOST IN TRANSLATION? DATA MINING, NATIONAL SECURITY AND THE "ADVERSE INFERENCE" PROBLEM

Anita Ramasastry†

I. INTRODUCTION

Imagine the childhood game of telephone: children sit in a circle and one child whispers a secret in the ear of a child sitting next to him. That "secret" is then relayed to the next child through a whispered remark. Eventually, the secret is relayed through the entire chain of children, as if multiple telephone calls had been placed, and a message relayed to numerous parties. Very often, the message that is relayed to the last child in the chain is very different from the original secret conveyed. The message has gotten distorted and lost in translation. Why? As data gets shared with a wider circle of people and multiple human actors have been asked to interpret and relay a message, as they understood it, to new people, the original message gets lost in translation.[1]

Today, a giant game of telephone is going on with our personal data. Our personal data, which we might disclose to a bank clerk over a counter, or provide to a customer service representative over the telephone, ends up far away from where it first started. Through a process known as data aggregation, messages about who we are and records containing our personal identifiers are being compiled from many sources, and this aggregated data is being transferred to third parties at a rapid speed. During this process of travel and aggregation, a seemingly inconsequential disclosure of a phone number to a store clerk can snowball into an "enhanced" file. Such

---

† Associate Professor and Co-Director, Shidler Center for Law, Commerce & Technology, University of Washington School of Law. The author would like to thank Chris Hoofnagle of the Electronic Privacy Information Center for his comments on this article.

1. Erick Breck & Clare Cardie, *Playing the Telephone Game: Determining the Hierarchical Structure of Perspective and Speech Expressions*, CORNELL UNIV. DEP'T OF COMPUTER SCI. (2004), http://www.cs.cornell.edu/~ebreck/publications/docs/breck04playing.pdf.

files are reformatted, edited, analyzed, interpreted, and changed to include volumes of information about a household, personal interests, and even medical conditions.

Often, we have little control over the messages that are being conveyed because we are not privy to this process — it happens behind the scenes. It is only recently that the American public has begun to realize that they have little control over their personal information once it is shared with a third party. And this has become of even greater concern now that we know that the federal government may be at the end of a "telephone" chain, reviewing personal data in order to combat terrorism.

Since September 11, 2001, the federal government has tried to connect more "dots" (data points) to prevent terrorism — by piecing together pieces of information and data to uncover possible plots and patterns. As part of this effort, the Executive Branch has introduced various proposals to "mine" private sector commercial databases and public records (as well as public databases) for information on everything from consumer addresses to financial and credit profiles. Such information, when fed into computers and analyzed, is meant to help the U.S. government predict who might be involved in terrorist activity.[2]

Data mining is a technique that uses information technology to identify previously undisclosed patterns and connections between different points of existing data, often with the goal of predicting future behavior.[3] In the world of commerce, this is done to maximize profit and to improve consumer experience. In recent years, the business of data collection has expanded with the rise of commercial data brokers — companies that aggregate consumer data from a wide variety of records, both public and private. Such data is then combined to create robust and detailed profiles on consumers. Data brokers may aggregate everything from cell phone records to travel reservations.[4] The government often purchases consumer data from

---

For a comprehensive overview, see JEFFREY W. SEIFERT, CONG. RESEARCH SERV., CRS Report RL31798, DATA MINING AND HOMELAND SECURITY: AN OVERVIEW (2006), http://www.fas.org/sgp/crs/intel/RL31798.pdf.

3. *Id.* at 1.

4. MARKLE FOUND. TASK FORCE ON NAT'L SEC. IN THE INFO. AGE, SECOND REPORT, CREATING A TRUSTED INFORMATION NETWORK FOR HOMELAND SECURITY 57-58 (2003), http://www.markletaskforce.org/reports/TFNS_Report2_Master.pdf [hereinafter MARKLE TASK FORCE REPORT].

such data brokers for varied purposes.[5] Law enforcement, for example, might want to try and locate the addresses of persons with outstanding arrest warrants.

Civil rights advocates have declared that governmental use of private sector data is a serious infringement on the right to privacy. One of the reasons for this declaration is that government now has access to vast "digital dossiers" maintained by commercial data brokers.[6] These private sector dossiers are available for purchase and reveal a great deal about our habits, patterns and daily activities. If combined with data from other sources, the government may be able to peek into our personal lives without having to obtain search warrants. Individual data points may not reveal much about our personality, but when compiled and stored over a longer period of time, our personal profiles become more revealing.

In essence, digital dossiers may enable the government to keep track of all of us through centralized databases. Our movements can be reconstructed by examining our credit card and debit card purchases; traffic cameras may scan our license plate; even our cell phone can help the government locate our travel path. Due to the volume of data that the government can now access with just the click of a button, data aggregation has changed the nature of the debate about if, when, and how government should have access to private sector data.

While civil libertarians may worry about erosions of individual privacy, many Americans who are concerned about fighting terrorism express the belief that if people are honest and law abiding, they have nothing to fear from the government knowing more about them. This can be referred to as the "I have nothing to hide" sentiment; the idea there should be no reason to fear government surveillance if one is a law-abiding person.[7]

This paper examines the use of data mining by the federal government to achieve national security goals and prevent terrorism,

---

5. *See* Robert Pear, *Survey Finds U.S. Agencies Engaged in 'Data Mining'*, N.Y. TIMES, May 27, 2004, at A24, *available at* http://www.nytimes.com/2004/05/27/national/27privacy.html?ex=1401076800.

6. Professor Daniel Solove first used the term "digital dossiers" in this context. Daniel Solove, *Digital Dossiers and the Dissipation of Fourth Amendment Privacy*, 75 S. CAL. L. REV 1084 (2002).

7. For a useful article on American opinions on issues relating to privacy versus security, see Darren W. Davis & Brian D. Silver, *Civil Liberties vs. Security: Public Opinion in the Context of the Terrorist Attacks on America*, 48 AM. J. POL. SCI. 28 (2004), *available at* http://www.msu.edu/~bsilver/AJPSCivLib.pdf.

and in particular attempts to rebut this notion that governmental analysis of commercial data and other personal data has little or no adverse consequences to the American population. To the contrary, governmental data mining can lead to the problem of "adverse inferences." Specifically, where a person has committed no crime or has nothing to fear, the use of data mining by the government can lead to the drawing of erroneous and adverse inferences, including the inference that this person poses a potential security risk. Adverse inferences such as these could lead to citizens being wrongly denied certain rights, such as the right to travel or the right to access governmental services.[8] When data mining generates an incorrect inference, it is referred to as a "false positive": "false" because the person is not truly exhibiting behavior that the data mining system is meant to uncover — such as terrorist activity, and "positive" because the system has flagged that person as a possible risk, thus triggering further scrutiny of that person or of their data.

Being identified as a security risk may mean public stigma and a fear of loss of control over one's life. For example, if mistakenly labeled a threat or possible danger to national security, how does one clear her name or remove herself from a government watch list? When a person is singled out based on unknown data retrieved from an unknown database, there is no clear-cut answer to this question. These fears become more palpable when one realizes that commercial data varies greatly in quality, may contain numerous errors, and is not subject to any mandated standards or control. The crucial issue with respect to data mining by the federal government is not whether or not we have anything to hide, but whether we can achieve national security goals while safeguarding the population from adverse inferences and their consequences.

The more data the government "mines," the larger the number of possible adverse inferences that may be drawn or false positives generated. In terms of a signal to noise ratio, proponents of data mining need to demonstrate that the success of such mining efforts will outweigh the frequency of false positives that will be generated.[9]

---

8. DANIEL J, WIETNZER, ET AL., MIT COMPUTER SCI. AND ARTIFICIAL INTELLIGENCE LAB., TECHNICAL REPORT MIT-CSAIL-TR-2006-007, TRANSPARENT ACCOUNTABLE DATA MINING: NEW STRATEGIES FOR PRIVACY PROTECTION 4 (2006), http://www.w3.org/2006/01/tami-privacy-strategies-aaai.pdf (discussing the fact that adverse actions can sometimes be taken based on factually incorrect antecedents).

9. *See, e.g.*, ASSOC. FOR COMPUTING MACHINERY, U.S. PUB. POLICY COMMITTEE, LETTER TO THE SENATE ARMED SERVICES COMMITTEE (Jan. 23, 2003), *available at* http://www.eff.org/Privacy/TIA/acm-letter.php. For an alternative view, which critiques some of

While the adverse inference problem should not be a reason for prohibiting governmental data mining, it should certainly cause legislators and others to examine data mining proposals with increased vigilance. While focusing on the problem of adverse (and incorrect) inferences, one also needs to examine the level of effectiveness of any data-mining project. The risk of "false negatives" — missing crucial people or events — is another risk that needs to be weighed when deciding what types of governmental programs should be authorized.[10]

This paper is not a polemic against the use of data mining as a means of preventing terrorism. Indeed, experts have lined up on both sides of the debate over data mining for national security purposes.[11] Rather, it attempts to highlight the harm that may arise if there are not proper safeguards built into project design and implementation. In the long run, governmental use of databases to authenticate identity may be beneficial, as it allows a greater degree of precision in determining who a person is and whether persons with similar names or characteristics are in fact different people.

The use of the term "lost in translation" as the title of this essay is an apt one because as data migrates from the private sector — where it was used for specific purposes[12] — into the government's

---

the assumptions used by opponents of data mining, see David Jensen, Mathew Rattigan & Hannah Blau, *Information Awareness: A Prospective Technical Assessment* (Assoc. for Computing Machinery 2003), *available at* http://kdl.cs.umass.edu/papers/jensen-et-al-kdd2003.pdf. The authors contend that current critiques of government data mining or information awareness systems use a "simplistic model of data mining" which does not factor in many of the techniques available to develop more sophisticated data analysis. *Id.* at 3. The authors also note that using relational data, a ranking classifier and multi-passing reference as part of a program design, a system can be configured to greatly reduce the number of false positives. *Id.* at 6.

10. Terrance A. Maxwell, *Information Policy, Data Mining, and National Security: False Positives and Unidentified Negatives*, *in* PROCEEDINGS OF THE 38TH HAWAII INT'L CONFERENCE ON SYS. SCIS. 7 (2005), *available at* http://csdl2.computer.org/comp/proceedings/hicss/2005/2268/05/22680134c.pdf.

11. For a thoughtful article on federal initiatives relating to data mining and terrorism prevention, and whether they should be allowed to proceed with appropriate safeguards, see K. A. Tapiale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 COLUM. SCI. & TECH. L. REV. 2 (2003). The Markle Task Force on National Security in the Information Age recommends that the government should have increased access to data in order to try and prevent terrorism, but also recommends revising current standards for governmental access to private sector data. MARKLE TASK FORCE REPORT, supra note 4, at 1 (Recommendation 6).

12. *See, e.g.*, David M. Lawson, *Mining External Data Sources: Making Sure Nothing is Lost in Translation*, CONNECTIONS, Spring 2004, at 8, *available at*

domain, the data is being used in new contexts and for purposes not originally anticipated. The data collected by commercial data brokers was not initially gathered with terrorism prevention in mind, and can thus be misinterpreted by government entities when used in the new context of terrorism risk assessment.

Subsidiary questions arise from the fact that once government agencies have access to our digital dossiers, they may be tempted to keep the data and use it for purposes other than that for which it was initially obtained. This is sometimes referred to as "mission creep" — where data collected in the name of national security will be used by the government to do things like chase tax evaders, monitor protected First Amendment activity, or build profiles on people to use for the prosecution of other crimes. [13]

While Congressional regulation and oversight is not always appropriate, in the case of data mining, Congress needs to maintain a healthy skepticism about the power of data mining as a tool in terrorist prevention. Government can more effectively authenticate a person by using more data points — and surely this is a laudable goal. But there are risks inherent in the systems that have been proposed by the government, and those risks need to be more openly examined and debated.[14]

To the extent that we permit data mining programs to proceed, they must provide adequate due process and redress mechanisms that permit individuals to clear their names. A crucial criteria for such a mechanism is to allow access to information that was used to make adverse assessments so that errors may be corrected. While some information may have to be kept secret for national security purposes, a degree of transparency is needed when individuals are trying to

---

http://www.aprahome.org/pdfs/Connections/LawsonMiningSpring04.pdf (example of how data needs to be carefully interpreted when being used in a new or different context from the one for which it was originally collected).

13. SEIFERT, *supra* note 2, at 18-19.

14. *See, e.g.*, Statement of Senator Patrick Leahy On The Data-Mining Reporting Act of 2003, http://leahy.senate.gov/press/200404/042204g.html (Apr. 22, 2004). Leahy notes:

> We have a responsibility to remain vigilant in our scrutiny of these programs. The Data-Mining Reporting Act simply provides us the necessary oversight information to carry out this responsibility. It offers a dose of sunshine to help Congress' oversight of the unprecedented mixture of these powerful new technologies that have great potential to help make us more secure, but also to make us less free.

protect their right to travel or access government services free from suspicion.[15]

Part II of this essay briefly outlines the government's ability to gain access to private sector data held by commercial entities or "third parties." Part III of this essay examines data mining and some of the problems inherent in using data analysis as a predictive tool for terrorism prevention. Part IV of this paper focuses on the specific problem of adverse inferences. This section examines the recent efforts of the federal Transportation Security Administration (TSA) to use data mining in airline passenger profiling. The Computer Assisted Passenger Profiling and Prescreening System II (CAPPS II) as mapped out by the TSA, and the most recent initiative, Secure Flight, illustrate some of the perceived risks inherent in the use of data mining to try and predict whether individuals are a security risk. Part V of this paper explores what efforts Congress and policymakers can make to address the risk of false positives and adverse influences, and the rise of commercial data mining as a favored tool for combating terrorism.

II: GOVERNMENT ACCESS TO DATA HELD BY THIRD PARTIES

Before analyzing government efforts with respect to data mining, one needs to first understand how and when the federal government may obtain access to data held by third parties. Americans often think of the Fourth Amendment of the U.S. Constitution as providing blanket protection against unreasonable governmental "searches and seizures" of their personal belongings and information, regardless of whether such information is stored in their own homes or with a third party. But, as Fourth Amendment scholars are quick to point out, the need to obtain a search warrant before gaining access to our possessions or our information is dependent upon whether we have a "reasonable expectation of privacy" in a particular context. If no reasonable expectation of privacy exists, the government is not

15. Certain government efforts, including passenger screening, have been challenged on Constitutional grounds as violating a citizen's right to travel freely, as well as their right to due process. These issues are not addressed in this article, which focuses more on the privacy interests that should be protected, even when the federal government is obtaining data in a manner that does not violate the U.S. Constitution. For a useful overview of recent Constitutional challenges to transportation security regulations, see TODD TATELMAN, CONG. RESEARCH SERV., CRS REPORT RL32664, INTERSTATE TRAVEL: CONSTITUTIONAL CHALLENGES TO THE IDENTIFICATION REQUIREMENT AND OTHER TRANSPORTATION SECURITY REGULATIONS (2004), http://www.fas.org/sgp/crs/RL32664.pdf.

required to obtain a warrant before effecting a search or seizure of our belongings.

When we subscribe to a magazine, check out library books, or apply for a credit card, we voluntarily relinquish some of our personal data (including sensitive data, such as a Social Security Number) to a third party. While to some extent we give up this data freely in order to gain access to a desired service, like getting access to credit, this notion of voluntariness is a bit false, as it is not possible to transact and survive in modern society without giving up some information to third parties on a daily basis. We could, for example, refuse to have a telephone and never give up data to a telecommunications company. A phone, however, is a vital lifeline for many; it provides access to essential government services such as the police, hospitals and public utilities. And for those who live in rural areas or who are housebound, it would be hard to say that they can reasonably choose not to maintain telephone service.[16]

Does the government need to obtain a warrant supported by "probable cause" before they are allowed access to the data held by third parties such as telecommunication companies or other businesses?[17] In other words, do members of the public have a reasonable expectation of privacy when third parties retain sensitive information about them? Two United States Supreme Court cases, *United States v. Miller*[18] and *Smith v. Maryland*,[19] establish a general rule that if information is in the hands of third parties, an individual lacks a reasonable expectation of privacy in regards to that information. In such a case, the Fourth Amendment warrant requirement would not apply.[20]

---

16. Solove, *supra* note 6, at 1157. Solove notes that "entities that maintain systems of records collect data in a power dynamic where information disclosure is often not consensual." Solove mentions that employers and landlords have a substantial amount of power to gather information. He contrasts employers and landlords with business actors, and notes that business merchants have less coercive relationships with individuals than entities that govern our livelihood and dwellings. At the same time, however, he acknowledges that "[e]ven if people are informed, that have little choice but to hand over information to third parties. Life in the Information Age depends upon sharing information with a host of third party entities including phone companies, ISPs. . . ." *Id.* at 1158.

17. *Id. See also* Stephen E. Henderson, *Nothing New Under the Sun? A Technologically Rational Doctrine of Fourth Amendment Search*, 56 MERCER L. REV. 507, 509-10 (2005).

18. United States v. Miller, 425 U.S. 45 (1976).

19. Smith v. Maryland, 442 U.S. 435 (1979).

20. DANIEL SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 201 (NYU Press 2004). *See* Michael J. Woods, *Counterintelligence and Access to Transactional Records: A Practical History of USA Patriot Act Section 215*, 1 J. NAT'L SEC. L. & POL'Y 37, 42 (discussing *U.S. v. Miller*). As Woods notes:

This concept that information held by third parties has, in essence, already been disclosed to someone and is therefore no longer private, is also referred to as the third party doctrine. However, there may be some limits to this doctrine. For example, individuals must relinquish sensitive data when seeking access to medical services; they provide this information to a doctor or a hospital because it is helpful for the health care provider to know about their medical history when providing necessary and, at times, lifesaving treatment. But to say that this is a voluntary bargained-for relinquishment would be to stretch the notion of voluntariness.

Medical history might be one of a limited number of classes of information held by third parties that retains its private nature when retained by a third party. So what about individual profiles maintained by a data broker, which may include a social security number, names of relatives, or previous employers? As Professor Daniel Solove has noted "[t]he government's harvesting of information from these extensive dossiers being assembled with modern computer technology poses one of the most significant threats to the privacy of our times."[21] Thus, scholars such as Solove have asked us to rethink our notions of the Fourth Amendment and when government may have unfettered access to data held by third parties.[22]

> In *United States v. Miller*, the Supreme Court held that the government can use a grand jury subpoena to obtain a defendant's financial records from a bank without intruding into an area protected by the Fourth Amendment. The Court pointed out that "'no interest legitimately protected by the Fourth Amendment' is implicated by governmental investigative activities unless there is an intrusion into a zone of privacy, into 'the security a man relies upon when he places himself or his property within a constitutionally protected area.'" The checks, deposit slips, and bank statements produced in response to the subpoena were not the defendant's "private papers," the Court held; rather, they contained "only information voluntarily conveyed to the banks and exposed to their employees in the ordinary course of business." By handing over this information to a third party, the defendant took the risk that it would be conveyed to the government by that third party.

See also Fred H. Cate, *Legal Standards for Data Mining*, at 7, *in* 21ST CENTURY ENABLING TECHNOLOGIES AND POLICIES FOR COUNTER TERRORISM (Robert Popp & John Yee eds., IEEE/Wiley & Sons 2006), *available at* http://www.hunton.com/files/tbl_s47Details/FileUpload265/1250/Cate_Fourth_Amendment.pdf. Cate discusses *Smith v. Maryland*, where the U.S. Supreme Court held that information about telephone calls such as the number dialed and time and duration of the call were not protected by the Fourth Amendment because such data was observable by third parties.

21. SOLOVE, *supra* note 20, at 202.

22. Solove, *supra* note 6, at 1151-56 (asking us to rethink nature of privacy architecture to take into account relationships between parties). *See also* Joseph T. Thai, *Is Data Mining Ever a Search Under Justice Stevens's Fourth Amendment?*, 74 FORDHAM L. REV. 1731, 1735 (2006)

Though data held by third parties is not constitutionally protected, most Americans are still likely to assume that the government cannot gain access to it. This may relate to notions of privacy and the context in which we originally provide information. For example, when we relinquish information to a credit card company, we probably do not expect this data will be sold to commercial data brokers. And Americans may not expect that information handed over to a grocer will end up with the government or a data broker.[23] This relates more to our own beliefs and ways of perceiving privacy and the flow of information. When a consumer shops at Safeway, he does not think that the federal government might someday be privy to whether he prefers Charmin toilet paper over a generic brand. Of course, if the government knew what groceries he bought, the consumer might not mind. But he would be more concerned if the government knew of every store that he has ever shopped at, or of every magazine that he has ever purchased.

That being said, Congress has stepped in to provide legislative privacy protection to various types of personal data including health and medical records, video rental records, and email messages. In these instances, there are specific statutes that govern how law enforcement may gain access to individual records. However, the standards vary greatly.[24] While at times the government needs a court order, a showing of probable cause is not always required. The Markle Foundation Task Force on National Security in the Information Age has produced a useful table which lists the relevant

---

(noting that although Justice Stevens' jurisprudence supports the third party doctrine, other opinions help "discern critical principles for limiting the reach of the third-party doctrine and its application to data mining").

23. Helen Nissenbaum argues for a rethinking of privacy rules by focusing on contextual integrity. She also notes that in the case of commercial data mining, there are situations where merchants may breach rules of appropriateness and norms of information flow when selling customer data to brokers and others. Helen Nissenbaum, *Privacy as Contextual Integrity*, 74 WASH. L. REV. 119, 152-53 (2004).

24. *See, e.g.*, MARKLE TASK FORCE REPORT, *supra* note 4, at 58:

> Government agencies already have access to certain kinds of privately held information. However, the rules governing access to it have evolved haphazardly and are confusing and sometimes contradictory. Moreover, the rules and practices fail to take into account the dramatic evolution of information technologies that can substantially increase the value of such data in haloing to prevent acts of terror.

See also Cate, *supra* note 20, at 16 (mentioning sectoral statutes imposing "modest limits" on the governments ability to seize data from third parties).

statutes governing access to information in the private sector by law enforcement and intelligence agencies.[25]

There are standards for specific sectors of information, but there is currently no legislation that regulates the ways in which the government may access data maintained by data brokers. While the government already has significant access to a variety of data sources, the new debate over data mining focuses to a large extent on commercial data brokers who have aggregated and warehoused substantial digital dossiers on individual Americans.[26] This paper focuses mainly on this aspect of data mining.

III. DATA MINING: HOW DOES IT WORK AND WHAT ARE ITS LIMITATIONS?

A. Background

Data mining refers to a process of using data analysis tools to discover or infer previously unknown patterns and relationships from large data sets.[27] Data analysis tools may include statistical models, mathematical algorithms and machine learning methods, where machines improve their performance over time as they are "taught" to identify and distinguish correct patterns from incorrect patterns. Thus, data mining involves various key steps, including data collection, aggregation, processing and analysis.[28] This will ultimately lead to decision making. With pattern-based data analysis,

---

25. *See* MARKLE FOUND. TASK FORCE ON NAT'L SEC. IN THE INFO. AGE, RULES GOVERNING ACCESS TO PRIVATE SECTOR DATA, http://www.markletaskforce.org/guidelines/government_matrix.shtml (last visited Apr. 6, 2006).

26. Chris Hoofnagle, *Big Brother's Little Helpers: How Choicepoint and Other Commercial Data Brokers Collect, Process and Package Your Data for Law Enforcement*, 29 N.C. J. INT'L L. & COM. REG. 595, 622-23 (2004) (arguing that the federal Privacy Act should apply to commercial data brokers when they sell private sector data to the government).

27. GEN. ACCOUNTING OFFICE, REPORT NO. GAO-04-548, DATA MINING: FEDERAL EFFORTS COVER A WIDE RANGE OF USES 1 (2004):

> The term data mining" has a number of meanings. For purposes of this work, we define 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.

28. MARY DEROSA, CTR. FOR STRATEGIC & INT'L STUD., DATA MINING AND DATA ANALYSIS FOR COUNTERTERRORSM 9-13 (2004), http://www.csis.org/media/csis/pubs/040301_data_mining_report.pdf; *See also* Taipale, *supra* note 11, at 24-28.

this may also lead to a prediction about whether certain data are indicative of a possible future event occurring.[29]

In the private sector, data mining has been used to detect fraud and assess risk, as well as to discover future patterns of consumer behavior. For instance, when Amazon.com recommends new books to its customers, it does so by mining its customer database to try and identify consumer preferences.[30] Banks and insurers use credit scoring as a way to decide whether to grant us credit, and at what rates or premiums.[31] Factors such as our credit score are meant to be predictors of how likely we are to repay our loans; likewise, our health, age and other physical characterereristics are meant to be predictors of what our life expectancy may be. These are all different applications of data mining.

B. Data Mining as a Counter-Terrorism Strategy

Recently, data mining has also been touted as a potential means to identify terrorist attacks. This is not to say that law enforcement has previously avoided data mining. The U.S Treasury Department collects data from banks on certain types of financial transactions pursuant to the federal Bank Secrecy Act, and these transactions are then analyzed to try and find instances of money laundering.[32] Such methods are also used to try and detect terrorist activity, including

---

29. DEROSA, *supra* note 28, at 12.

30. Charu C. Aggarwal & Philip S. Yu, *Data Mining Techniques for Personalization*, BULLETIN OF THE IEEE TECHNICAL COMMITTEE ON DATA ENGINEERING, at 5-6 (2002), *available at* http://cs.chungnam.ac.kr/~ykim/courses/grad-dbs2000/papers/DataMining.pdf.

31. Lyn C. Thomas, *A Survey of Credit and Behavioral Scoring: Forecasting Financial Risk of Lending to Consumers*, 16 INTL' J. OF FORECASTING 149, 150-52 (2000), *available at* http://socsci2.ucsd.edu/~aronatas/project/academic/A%20survey%20of%20credit%20and%20behavioural%20scoring%20Forecasting%20fina.pdf. *See, e.g.*, BRENT KABLER, MO. DEP'T OF INS., INSURANCE-BASED CREDIT SCORES: IMPACT ON MINORITY AND LOW INCOME POPULATIONS IN MISSOURI (2004), http://insurance.mo.gov/reports/credscore.pdf.

32. The Financial Crimes Enforcement Network of the U.S. Department of Treasury states on its mission page:

> Technology such as data extraction, data mining and leading edge analytical tools enhance the analysis and manipulation of BSA data—identifying and linking related information to add value to what is already known by investigators, enabling analysts to identify trends and patterns in money laundering and other financial crimes, and linking BSA data with other information to identify suspicious activity tied to organized crime groups.

The Financial Crimes Enforcement Network of the U.S. Dep't of Treasury, Mission Page, http://www.fincen.gov/fin_mission.html.

terrorist finance flows.[33] What is new with respect to governmental data mining is: (1) the announcement of specific anti-terrorism initiatives focused on the general public, and (2) the plan to glean large amounts of data from commercial data brokers as part of new counter-terrorism initiatives.[34]

Perhaps the most ambitious of federal data mining initiatives was the poorly named "Total Information Awareness" (TIA) Program,[35] which was renamed "Terrorism Information Awareness," perhaps in response to public criticism of the original moniker.[36] The program was to be run by the Information Awareness Office, which at the time was part of the federal Defense Advanced Research Projects Agency (DARPA). Retired Admiral John Poindexter, National Security Adviser for former President Ronald Reagan, returned to the Pentagon in February 2002 to run the Information Awareness Office and TIA.[37]

TIA envisioned the creation of a huge, centralized national database of information, gathered from existing government and commercial data banks. The records that were intended to be compiled and consolidated included bank records, tax returns, driver's license data, credit card purchases, airline tickets, gun purchases, work permits, and more. [38] TIA was to then use computer algorithms to detect suspicious patterns and try and identify possible terrorist activity. However, in September 2003, Congress eliminated funding for the controversial project and closed the Pentagon's Information Awareness Office, which had created TIA.[39]

---

33. MARTIN A. WEISS, CONG. RESEARCH SERVICE, CRS REPORT RL31798, TERRORIST FINANCING: U.S. AGENCY EFFORTS AND INTER-AGENCY COORDINATION 38 (2005), http://www.fas.org/sgp/crs/terror/RL33020.pdf.

34. *See* Cate, *supra* note 20, at 2.

35. Anita Ramasastry, *Why We Should Care About "Total Information Awareness" and Other Anti-Terrorism Strategies for the Internet*, FINDLAW, Dec. 31, 2002, http://writ.news.findlaw.com/ramasastry/20021231.html. The original logo for the TIA web site was a pyramid with a giant eye beaming its rays onto the entire planet. This logo was eventually removed from the site. TIA Logo, http://www.thememoryhole.org/policestate/iao-logo.htm (last visited Apr. 26, 2006).

36. For a detailed description of "Terrorism Information Awareness" see REPORT TO CONGRESS REGARDING THE TERRORISM INFORMATION AWARENESS PROGRAM (May 20, 2003), *available at* http://www.epic.org/privacy/profiling/tia/may03_report.pdf. *See also* EEF REVIEW OF MAY 20 REPORT IN TOTAL INFO AWARENESS EXECUTIVE SUMMARY, http://www.eff.org/Privacy/TIA/20030523_tia_report_review.php (last visited, Apr. 5, 2006).

37. Ramasastry, *supra* note 35.

38. *Id.*

39. CONFERENCE REPORT ON H.R. 2658, DEP'T OF DEF. APPROPRIATIONS ACT 2004, H.R. REP. NO. 108-283 (2004), http://www.fas.org/sgp/congress/2003/tia.html.

The TIA concept is an example of a new and broad scale experiment with governmental data mining of private sector databases; an attempt to classify and predict terrorist populations through an analysis of the general population. It is the depth and breadth of such proposed initiatives that may have sparked Congressional debates over data mining. TIA captured the public's imagination, and eventually led to the program's demise.

C. Limitations of Data Mining For Terrorism Prevention

There are some inherent limits to data mining. Data mining can help reveal patterns or relationships, but it may not tell a user the significance of those patterns or relations; the user herself must determine the significance or draw inferences.[40] Also, the effectiveness of data analysis tools will be dependent upon the type of criteria or assumptions that are built into any model or algorithm. And, data mining can identify connections between variables, but it cannot necessarily identify a causal relationship.[41]

For example, a data-mining program might flag airline passengers who purchased their tickets within a short time prior to the flight's departure. Perhaps this has been determined to be a common feature of terrorist or criminal travel. This, of course, does not mean that everyone who buys an airplane ticket at short notice is a criminal or terrorist.[42] The decision to place persons on a list based on the nature of their ticket purchase is an adverse inference. The passenger will have to subject himself to additional screening in order to prove that he is not a security risk. People may purchase tickets on short notice for multiple reasons. For example, a journalist has to travel on short notice because of the nature of her profession. Other travelers may need to travel on short notice due to an unexpected personal event, such as a death or illness in the family. Thus, data mining can identify behaviors, but not the underlying cause or motive for such a behavior or pattern.

Successful data mining requires skilled experts who can analyze the data and draw causal relationships and inferences from the patterns identified by machines. The success of identifying behaviors is dependent upon how relevant such patterns are to the underlying behavior that a user is trying to find — such as terrorist activity. If there are more non-terrorist than terrorist travelers who fly on short

40. SEIFERT, *supra* note 2, at 3.

41. *Id.*

42. *Id.*

notice, resources may be wasted in the investigation of "false positives" — innocent persons identified as possible risks by a data analysis program. The risk of these false positives is significant, if a government actor chooses to proceed on the basis of data analysis which includes possibly innocent persons.[43]

Patterns uncovered by data mining also need to be compared to real world circumstances in order to confirm their validity. Law enforcement agencies could use data analysis tools to identify persons whose behaviors or actions matched profiles of known terrorists. The question remains, however, of what to do when a person's behavior deviates significantly? Critics of the use of data mining as a counter terrorism measure have noted that once terrorists ascertain the types of behaviors being sought out via data analysis, they can change their patterns or activities to evade detection. Terrorists can ascertain what sort of patterns trigger enhanced security screening, for example, by sending testers on flights just to see if they are singled out. By testing a system repeatedly, terrorists can determine what factors are important and how someone can fly undetected. For example, if persons who purchase one-way airplane tickets are seen as security risks, a terrorist could buy a round trip ticket instead to evade detection.[44]

---

43. DEROSA, *supra* note 28, at 15. As DeRosa notes:

> But the stakes are so high when fighting catastrophic terrorism that there will be great temptation for the government to use these [data mining] techniques as more than an analytical tool. Government actors will want to take action based on the results of data-analysis queries alone. This action could include detention, arrest, or denial of a benefit. Even if the government later corrects its mistake, the damage to reputation could already be done, with longer-term negative consequences for the individual. Even when the error is identified, there may be difficulties correcting it. . . . Although the technology exists to follow inaccurate data and correct cascading occurrences, it has not been a priority, and its implementation lags far behinds the technology for collecting and analyzing data.

44. *See, e.g.*, Samidh Chakrabati & Aaron Strauss, *Carnival Booth: An Algorithm for Defeating the Computer-Assisted Passenger Screening System* (May 16, 2002), *available at* http://swissnet.ai.mit.edu/6805/student-papers/spring02-papers/caps.htm. (Paper prepared for MIT Seminar on Law and Ethics on the Electronic Frontier). Two MIT students noted:

> Any CAPS like airport security system that uses profiles to select passengers for increased scrutiny is bound to be less secure than systems that randomly select passengers through inspection. Using mathematical models and computer simulation, we show how a terrorist cell can increase their chances of mounting a successful attack under the CAPS system as opposed to a security system that uses only random searches.

Id at 5. As the authors point out, the fact that one knows if one has been flagged as a security risk is the

> Achilles heel of CAPS: the fact that individuals know their CAPS status enables the system to be reverse engineered. . . . You know if you've been questioned.

1. Infrequent Events May Lead To Poorly Corroborated Profiles

Security experts note that data mining works best when there is a well-defined profile. An example of a more successful use of data mining relates to the prevention of credit card fraud.[45] The frequency of credit card theft gives credit card companies many opportunities to test their data analysis algorithms and programs, and to refine and improve their fraud detection indicators.

Credit card issuers mine their transaction database and look for suspicious transactions or unusual spending patterns that may indicate that a card has been stolen and is being used for unauthorized transactions.[46] Apparently, many credit card thieves exhibit a pattern of purchasing expensive luxury goods and consumer goods that can be easily fenced.[47] A credit card issuer can then either notify the cardholder to try and verify the transaction, or in some cases shut down the card. Cardholders typically do not mind these security measures. Usually, a phone call to the cardholder resolves the problem — either the card is indeed stolen or the cardholder can verify the purchases as his or her own.[48] In some instances, the cardholder may be traveling overseas and is unable to verify or deny

---

> You know if you're asked to stand in a special line. You know if you've been frisked. All of this open scrutiny makes it possible to learn an anti-profile to death CAPS, even if the profile itself is always kept secret.
>
> Terrorists can send operatives on flights, without any explosives or intention to do harm. This will allow terrorists to test the system to see if they are detected or searched, as a way of figuring out what anti-profile will be successful in evading detection.

Id. at 8-9. *See also* ACM LETTER, *supra* note 9.

45. Schneier on Security, *Data Mining for Terrorists*, http://www.schneier.com/blog/archives/2006/03/data_mining_for.html (Mar. 9, 2006). For an overview of data mining techniques in credit card fraud detection, see PHILIP K. CHAN ET AL., DISTRIBUTED DATA MINING IN CREDIT CARD FRAUD DETECTION 14(6):67-74 (IEEE Intelligent Systems 1999). *See also* DEROSA, *supra* note 28, at 4.

46. *See, e.g.*, American Express Fraud Protection Guarantee, https://www124.americanexpress.com/cards/cda/dynamic.jsp?name=FraudProtectionGuarantee_SharedDetailsALL&type=intBenefitDetail (last visited Apr. 5, 2006), which states: "Our Fraud Detection system watches your account for uncharacteristic or unusually high charges. If we detect unusual activity on your Card, we may suspend further charges until we can confirm the purchases were authorized."

47. Schneier, *supra* note 45.

48. *Id.* MasterCard is offering to notify cardholders using SMS (text messaging) to cardholder cell phones. Laura Rohde, *MasterCard Offers SMS to Detect Credit Card Fraud*, INFOWORLD, Feb. 9, 2005, *available at* http://www.infoworld.com/article/05/02/09/HNmastercardfraud_1.html.

the purchases, in which case his card may be blocked.[49] In these few cases, there may be mere annoyance when a cardholder finds himself unexpectedly unable to complete a purchase, but he can often resolve the issue and get the card unblocked. The few cases where this is not possible leads to more annoyance to the customer, but the cost to cardholders is marginal compared to the benefits of fraud detection,[50] and ultimately, the possibility of corroborating the activity profile of a credit card thief.

Terrorism prevention is a different matter, as security experts point out. There is no well-defined and reinforced profile for terrorists. Further, attacks are relatively infrequent, making it harder to reinforce any profiles that do exist. Security expert and computer scientist Bruce Schneier does a good job of explaining some of the problems with data mining in the terrorism context. There are 900 million credit cards in circulation in the United States, and according to an FTC study from September 2003, about one percent (one million) of these cards are stolen and used for unauthorized purposes each year.[51]. While the total number of credit card transactions is large, Visa and MasterCard are only analyzing one type of credit card transaction — purchases and sales. This is a more limited class of data to analyze, as contrasted with data mining for terrorism prevention, which might involve many different types of data. Furthermore, there are enough instances of confirmed credit card fraud that data analysis techniques can constantly test any assumptions against a large number of real world cases. This helps the credit card industry to update its models and spot trends as they emerge.

Terrorism prevention presents a much greater magnitude of data than credit card transactions. As Schneier indicates, while there are only a very few number of known terrorist plots at any given time, there are "trillions of connections between people and events — things that the data mining system will have to 'look at'— and very few plots. This rarity makes even accurate identification systems useless."[52] The government will have to look at a much larger amount of data in order to try and thwart a future terrorist event.

---

49. Carol Pucci, *You could be out of the Country and out of Luck with your ATM Card*, SEATTLE TIMES, Mar. 18, 2006, *available at* http://seattletimes.nwsource.com/html/travel/2002867484_pucci19.html.

50. Schneier, *supra* note 45.

51. *Id.*

52. *Id.*

Schneier elaborates by explaining that even if one was to assume the existence of a very "accurate " data mining system with only a 1 in 100 false positive rate (99% accurate) and a 1 in 1,000 false negative rate (99.9% accurate), there would be one trillion indicators to sift through daily. This amounts to about ten events — emails, phone calls, purchases, etc. — per person in the U.S. per day. Of these one trillion events, perhaps ten in any one day are actual events connected to a terrorist plot. This "unrealistically" accurate system would still generate one billion false alarms for each terrorist plot it uncovers. The police may have to investigate 27 million potential plots every day of every year in order to find one real plot per month.[53] A recent example of a real data mining system generating false alarms involves the National Security Agency's (NSA) eavesdropping programs. According to a New York Times article, NSA computers generated thousand of tips in the months following 9/11.[54] Every one turned out to be a false alarm.

With government investigations, the cost of false positives is clearly high, both in terms of government resources and invasion of personal privacy. Of course, Schneier's example is assuming a data-mining program more akin to TIA, rather than a more limited type of data analysis, such as the examination of financial flows to try and detect terrorist financing. If one limits the data analyzed, for example to a certain region and to certain types of transactions, the risk of errors is likely to decrease.

2. Erroneous and Inaccurate Records Impair the Reliability of Third Party Data

As just demonstrated, one significant concern about the governmental use of commercial data relates to the possibility of error.[55] One factor in error rates is data quality, which refers to the accuracy and completeness of data used to draw inferences. Duplicate records, the inconsistent or complete lack of data standards, the timeliness of updates, and human error (e.g. incorrect data entry) can all impact how effective data analysis will be. Even something as slight as different date formats can cause records to be inconsistent with one another. Data needs to be reliable and kept up to date if it is

53. *Id.*

54. Lowell Bergman, Eric Lichtblau, Scott Shane & Don Van Natta, Jr., *Domestic Surveillance: The Program; Spy Agency Data After Sept. 11 led FBI to Dead Ends*, N.Y. TIMES, Jan. 17, 2006, at A1.

55. SEIFERT, *supra* note 2, at 17.

to be used on a frequent basis for identity verification.[56]In many data mining or processing initiatives, data is "cleaned" to correct errors before it is used and analyzed.[57]

At present, data integration methods vary and may be based solely on name and address, with additional information such as social security numbers, date of birth, or driver's license data added in when available. A report prepared by the Markle Foundation's Task Force on National Security in the Information Age notes that name and address data is currently captured in multiple formats, allowing for the introduction of errors. Also, the information is often out of date. According to Markle, twenty percent of the U.S. population moves every year; five percent have second homes. Five million marriages and two million divorces occur each year, resulting in name and address changes.[58]

Commercial data brokers do have techniques to resolve these issues and to correct some errors, such as recognizing whether someone listed as "Robert" in one record and "Bob" in another are in fact the same person. These techniques, at present, have reduced error rates by one to two percent, [59] which can still be significant when it comes to the governmental use of such data.

However, errors do frequently exist in consumer credit reports.[60] This is important to remember, because many of the existing government data mining proposals relate to the use of credit profiles or consumer profiles retained by data brokers. A 2004 survey of the U.S Public Interest Research group revealed some startling statistics. Approximately one in four credit reports contain errors serious

---

56. As the Markle Task Force on National Security in the Information Age noted: "False or incomplete data will accentuate the problem of both false positives and false negatives. There are even broader implications if the government can access this faulty data and attach consequences to it (for example, restricting the right of an individual to board an airplane." MARKLE TASK FORCE REPORT, *supra* note 4, at 58. *See also* REPORT OF THE SECURE FLIGHT WORKING GROUP, PRESENTED TO THE TSA 10 (Sept. 19, 2005) (discussing problems with false positives and negatives in identity matching) [hereinafter REPORT OF THE SECURE FLIGHT WORKING GROUP].

57. DEROSA, *supra* note 28, at 10 (describing high quality practices for cleansing and standardizing identity data including name standardization and address hygiene).

58. MARKLE TASK FORCE REPORT, *supra* note 4, at 60.

59. *Id.*

60. PUB. INTEREST RESEARCH GROUP, MISTAKES DO HAPPEN: A LOOK AT ERRORS IN CONSUMER CREDIT REPORTS (2004). *See also* CONSUMER FEDERATION OF AMERICA AND NATIONAL CREDIT REPORTING ASSOCIATION, CREDIT SCORE ACCURACY AND IMPLICATIONS FOR CONSUMERS 37-40 (2002) (cataloging nature of errors on consumer credit reports and consequences of inaccuracies in terms of cost of credit), http://www.consumerfed.org/pdfs/121702CFA_NCRA_Credit_Score_Report_Final.pdf.

enough to result in a denial of credit, employment, or housing. Fifty-four percent of the reports analyzed contained personal demographic information that was misspelled, long outdated, belonged to a stranger, or was otherwise incorrect.[61] This is a pertinent figure when one realizes that governmental data mining is often premised upon the use of such data when verifying a person's identity.

In well-managed data mining projects, an original data-collecting organization is likely to be aware of the data's limitations and will account for those limitations or errors accordingly, either by manually correcting the errors, using data analysis that can correct the errors, or by simply analyzing the data with these errors in mind. But when data is used in another context or by another entity, such safeguards may not exist.

Moreover, data may be interpreted out of context when reviewed by a new user. For example, data collected from a shopper's club card may be reflective not of a single purchaser, but of a household or multiple groups of purchasers. A person may use a spouse's shopper card when buying groceries for the family. The data is useful to the grocery store because it shows patterns of household consumption. However, that same data, if used to confirm identity, may produce inaccurate results for a number of reasons. Shoppers may use multiple cards, provide false names when they do obtain cards, or swap cards within households or living units. Cashiers even swipe their own cards or a "store card" when a shopper does not have one at his disposal.

For marketing purposes these inaccuracies are negligible. If a wife uses a husband's card, and he typically buys shaving supplies, she might receive a coupon for razors when she next uses his shopping card. While the grocery store's database made an incorrect inference about who she is, the only consequence to her is that she got a coupon that she might discard.

If, however, a government agency were to target individuals based on the purchase of foods associated with religious observances — Halal meat, for example — the consequences of inaccurate identification change significantly. Similarly, if shopper card data were used to infer where someone shopped on a particular day, the risk of false positives increases. Re-use of data or data mining with repurposed data thus presents larger problems with respect to data

61. PUB. INTEREST RESEARCH GROUP, *supra* note 60, at 4.

quality and the drawing of incorrect inferences from data compiled by other sources.

Another problem with data integrity relates to identity theft.[62] To the extent that the government wishes to use data from commercial brokers, federal agencies may be relying on credit histories and data that are created by thieves or imposters who have assumed another person's identity. Thus, data mismatches may occur because the victim of identity theft is unaware that a thief has, in essence, rewritten the victim's personal history. This makes the data less reliable.

The context in which individuals provide information to third parties may also compromise data integrity. We may have more incentive to be truthful when disclosing information to a government agency, but to the extent that a consumer provides information in various contexts in the private sector, he may have reasons for being more or less candid about something — depending on who is asking for the information and for what purpose.[63] For instance, one might lie or provide only partial information when applying for credit, or even making a purchase. Embarrassment might make a person provide false data when purchasing pornography. When asked about income, people may overstate or understate their earnings for a variety of reasons. There are also problems of interpretation; persons do not expect that information handed over in one context will wind up being examined in a different one.

Another problem is that discrimination may result from the use of commercial databases — disadvantage may be misinterpreted by the TSA as shiftiness. On average, minority populations tend to have lower credit scores, and certain categories of persons may have no credit record at all.[64] Certain populations, such as low-income

---

62. REPORT OF THE SECURE FLIGHT WORKING GROUP, *supra* note 56, at 12.

63. *See, e.g.*, *Web Lies and Privacy Studies*, SAN JOSE BUS. J., May 30, 2002, http://sanjose.bizjournals.com/sanjose/stories/2002/05/27/daily46.html; Ann Bednarz, *IBM Unveils Web Privacy Work*, NETWORK WORLD, May 31, 2002, http://www.networkworld.com/news/2002/0531ibmprivacy.html; Amy Bruckman, *Gender Swapping on the Internet*, PROC. OF THE INTERNET SOC'Y (1993), http://citeseer.ist.psu.edu/bruckman93gender.html (discussing pattern of women identifying themselves as male online to access information more freely); *See generally* Nissenbaum, *supra* note 23; *See also* Solove, *supra* note 6, at 49.

64. *See, e.g.*, Kabler, *supra* note 31, at 1-2 (concluding that insurance credit scoring was lower for residents in low income and high minority populated zip codes; minority and low-income individuals were significantly likely to have worse credit scores than wealthy individuals and non-minorities). The report noted that "the relationship between minority concentration in a zip code and credit scores remained after eliminating a broad array of

individuals, students, immigrants and homeless persons, may also move more frequently than others, so their data may be missing, inaccurate, or contain suspicious gaps. There is also a fear that a person may be excluded from certain types of commercial transactions because of associations made or inferences drawn based on limited classes of data such as a zip code.[65] Thus, the absence of data could lead to businesses and government entities making adverse inferences that a person is suspicious or has something to hide.

These are all examples of how data can get "lost" in translation. First, data may not be accurate when provided initially by a consumer or may only reflect a partial truth. Second, data may get mistranslated due to human error (e.g. typing in a birth date incorrectly) when placed into a database. Third, when data is used in a new context, it may not be interpreted in the same way as previously used, because the new party using the data may not understand how the data was originally classified. For example, racial or ethnic classifications in one database may be different than in a new database. Fourth, when data from different sources is combined into a larger database, it may be incorrectly integrated. In other words, data from different people who share the same surname might be incorrectly merged, creating a new profile that is incorrect. Thus, there are multiple ways in which data may be erroneous. Where human agents are involved in compiling or aggregating different data, data sources can be mistranslated.[66]

---

socioeconomic variables, such as income, educational attainment, marital status and unemployment rates as possible causes. Indeed, minority concentration provided to be the single most reliable predictor of credit scores." *Id.* at 2. While there may be specific reasons why certain populations might have lower credit scores, there is a fear that location or address might be used to correlate risk or lack of creditworthiness.

65. *See, e.g.*, Claire Cherry v. Amoco Oil Co, 490 F. Supp. 1026, 1030 (N.D. Ga. 1980) (White plaintiff brought suit alleging that she had been denied a credit card solely because her zip code placed her in a neighborhood with a high proportion of African Americans. Court held that plaintiff's proof failed to show that the zip code ratings used by the creditor tended to adversely affect black applicants disproportionately. Rather, it showed only that the 38-criteria computerized grading system scheme, taken as a whole, tended to reject a disproportionate number of persons living in predominately black areas).

66. Data, when combined from different sources, can create new records that are inaccurate. The author's husband, for example, happens to share the same name with another person who lives in our neighborhood. The neighbor has filed for Chapter 13 bankruptcy. Data brokers and merchants have combined a list of persons who have filed for Chapter 13 bankruptcy, which includes the neighbor, with a list of property records (which includes the author's husband's name) to create a list of property owners who have filed for Chapter 13 bankruptcy. Alas, my husband receives mailings and calls from collection agencies, real estate agents and others, who are either chasing the like-named neighbor, or trying to sell him products and services aimed at individuals with financial troubles. The new record, which merges data

IV. AIRLINE PASSENGER PROFILING: AN EXAMPLE OF THE ADVERSE INFLUENCE PROBLEM

The recent efforts of the Department of Homeland Security (DHS) and TSA with respect to airline passenger profiling provide a useful illustration of how data mining can create erroneous or adverse inferences that have significant consequences for the American population. These same illustrations also highlight the need for clearer policies on (1) what information may be collected by the government, (2) what processes will be in place for error correction, and (3) how erroneously targeted citizens may clear their name

The program currently in existence provides a palpable example of the problems innocent travelers may encounter when errors arise due to data mining. The current passenger screening system, which tries to locate passengers who are on a "no-fly" list, has made the process of boarding a plane a burden for numerous Americans, in come cases causing persons to lose the right to travel altogether.

Two proposed systems have attempted to use commercial data as part of a process to better identify possible terrorist threats. Both the Computer Assisted Passenger Prescreening System II (CAPPS II) and Secure Flight have been criticized by Congress and the General Accounting Office (GAO) for not being able to more fully safeguard against errors in commercial databases and in the data analysis process. They were also criticized for having vague or ambiguous redress procedures. The two proposed programs highlight problems that may exist with data mining, as well as the problem of categorizing persons as potential security risks based on commercial data.

A. The "No-Fly" and "Selectee" Lists

The administration of the "no-fly" and "selectee" lists, currently maintained by the TSA in cooperation with the DHS, provide an example of the harm caused by adverse inferences.[67] The no-fly list was originally created in 1990 for the FBI (and was later maintained by the Federal Aviation Administration), and lists individuals who

---

on two men, but which links the data to our home address, is erroneous and is used to draw adverse inferences about my husband's financial situation.

67. ELEC. PRIVACY INFO. CTR. (EPIC), DOCUMENTS SHOW ERRORS IN TSA'S "NO FLY" AND "SELECTED" WATCH LISTS,
http://www.epic.org/privacy/airtravel/foia/watchlist_foia_analysis.html
[hereinafter EPIC WATCH LIST].

have been "determined to pose a direct threat to U.S. civil aviation."[68] The selectee list is an outgrowth of a system known as Computer Assisted Passenger Profiling, which commenced in 1996. While the "no fly" and "selectee" lists do not involve commercial data mining, the screening for these lists reveals the inherent problems in trying to match data from different sources to identify which persons are security risks.

Those who find their name on the lists suffer a variety of consequences: they are typically denied the use of curbside check-in and electronic ticket kiosks, are sometimes subjected to questioning in full view of fellow passengers, and are often delayed for hours before being placed on a different flight.[69] In many cases, a TSA or airline agent takes the passenger's identification and travel documents, further adding to the feeling that the passenger is not free to leave the airport.[70]

Despite all these penalties, those on the lists are not given the right to know why they were listed, or an avenue to try to clear their name. According to some estimates, the no-fly and selectee lists alone may affect thousands of innocent people.[71] Moreover, according to a recently issued GAO report, the no-fly and selectee lists are just two of 12 terrorist and criminal watch lists maintained by the federal government.

In November 2001, the TSA assumed full responsibility for the No-Fly and Selectee lists, and the Transportation Security Intelligence Service (TSIS) currently serves as the clearinghouse for the addition of names to the lists. The lists are distributed to all airlines, with instructions to stop or conduct extra searches of people on the lists who are suspected of being terror threats. The lists also appear to have been shared widely among U.S. law enforcement agencies, international agencies, and the U.S. military.

---

68. *Id.*

69. Am. Civil Liberties Union, Frequently Asked Questions About the "No Fly List" (Oct. 26, 2005), http://www.aclu.org/safefree/general/21164res20051026.html; *See also* Preliminary Statement, Class Action Complaint in Green v. TSA, 6-9 (Wash. Dist. Ct. 2004), *available at* http://files.findlaw.com/news.findlaw.com/hdocs/docs/aclu/greenvtsa40604cmp.pdf.

70. Preliminary Statement, Class Action Complaint in Green v. TSA, *supra* note 69, at 6; *See also* Anita Ramasastry, *A New ACLU Lawsuit Challenges "No Fly" and "Selectee List" Procedures. Do these Government Watch Lists Violate Due Process?*, FINDLAW, Apr. 13, 2004, http://writ.news.findlaw.com/ramasastry/20040413.html.

71. Preliminary Statement, Class Action Complaint in Green v. TSA, *supra* note 69, at 1.

1. Getting on the Lists: Selection Criteria Unclear

How do the lists work? In October 2002, the Electronic Privacy Information Center (EPIC) submitted a Freedom of Information Act request to answer that very question, but the TSA failed to respond. Initially, the TSA had denied that the lists even existed, but went on to acknowledge their existence only weeks later.[72] In December 2002, EPIC filed suit to compel disclosure of the information being sought — including TSA's criteria for putting people on the lists.[73] The suit also included complaints from passengers who felt that they had been mistakenly placed on these lists. EPIC finally received the documents it had requested, but found that they only revealed limited details on how the lists are administered.

The documents did indicate that those on the no-fly list are not allowed to fly, while those on the selectee list must go through additional security measures, and that both lists are stored on air carriers' computer systems. When a customer requests a boarding pass, his or her name is compared to the two lists. If the name is on the no-fly list, the airline must call a law enforcement officer to detain and interview the passenger. If a name is on the selectee list, an "S" or special mark is printed on the passenger's boarding pass and the person is subjected to enhanced screening at security.[74]

According to the Oakland Airport, the government's no-fly list is eighty-eight pages long.[75] The documents received by EPIC provide scant information on how one gets placed on the list, for the government insists on keeping this procedure secret.[76] This, of course, may be legitimate for security reasons. But in the event criteria may not be revealed, the TSA needs to articulate enough measures that a person may take to get removed from the list. In other words, how can one effectively prove she is not a security risk? There is no indication in the disclosed government documents of any review by an independent third party to verify the lists' accuracy. Nor is there reference in the documents to any formal procedure by

72. EPIC WATCH LIST, *supra* note 67.

73. Complaint for Injunctive Relief, EPIC v. TSA and Dept. of Transportion (D.C. Cir. 2002), *available at* http://www.epic.org/privacy/airtravel/tsa_foia_suit.pdf.

74. EPIC WATCH LIST, *supra* note 67.

75. JAYASHRI SRIKANTIAH, ACLU OF NORTHERN CALIFORNIA, THE PUBLIC STILL LACKS BASIC INFORMATION ABOUT THE NO-FLY LIST: AN ANALYSIS OF TSA'S FOIA RESPONSE 2 (2003), *available at* http://www.aclu.org/FilesPDFs/072103-noflysummary.pdf.

76. *Id.*

which individuals might clear their names — other than calling their local FBI office.[77]

Documents received in an April 2003 ACLU lawsuit, triggered by the December 2002 detention of two female anti-war protesters attempting to fly from San Francisco to Boston, are no more encouraging. They reveal a confused process in which the government expressed uncertainty about how the lists should be shared. And once again, the documents failed to answer basic questions about how the lists are generated and administered.[78]

2. Getting off the Lists: A Bureaucratic Maze

As indicated by the complaint letters received by EPIC, passengers who try to get off the lists face a bureaucratic maze. This process, moreover, increasingly affects a larger number of innocent Americans. According to ACLU reports, between September 11, 2001, and April 2003, at least 339 innocent air passengers were stopped and questioned by police at the San Francisco Airport, because their names were believed to match names on the No-Fly/Selectee lists.[79] Indeed, the TSA has also admitted that it receives thirty calls per day regarding false positives — which amounts to well over a thousand calls a year.[80] This, of course, does not factor in passengers who may have been flagged but who have not reported their difficulties.

Currently, the TSA has an ombudsman for handling passenger complaints, but not everyone is made aware of that fact.[81] According to a second ACLU lawsuit, some passengers were told instead to

77. EPIC WATCH LIST, *supra* note 67.

78. Am. Civil Liberties Union, ACLU Seeks Government Accountability For No-Fly List, Rebecca Gordon et al. v. FBI et al. (feature on ACLU's No-Fly List FOIA Lawsuit and links to litigation documents), http://www.aclu.org/safefree/general/17463res20030606.html.

79. SRIKANTIATH, *supra* note 75.

80. *Id. See also* ELEC. PRIVACY INFO. CTR. (EPIC), FREEDOM OF INFORMATION ACT NOTES, COMPLAINT LOG, http://www.epic.org/foia_notes/complaint_log.pdf.

81. Preliminary Statement, Class Action Complaint in Green v. TSA, *supra* note 69, at ¶ 30-34. The plaintiffs in the Green case included persons who allege that their names were wrongly included on the No-Fly list. Plaintiffs include a member of the military, a retired Presbyterian minister, and a college student A federal court dismissed the lawsuit in January 2005, citing a lesser-known jurisdictional statute. In December 2004, Congress enacted a new statute that requires TSA to enact new rules for the No Fly List. Rather than pursue an appeal based on allegations that predate the new rules, the ACLU stated it would wait and see how the new rules were implemented. *See* Am. Civil Liberties Union, Background of Green v. TSA, http://www.aclu-wa.org/inthecourts/detail.cfm?id=319&CFID=4638238&CFTOKEN=85637108 (last visited Apr. 26, 2006).

contact the airlines or the FBI to be removed from the lists. Those who do write to the ombudsman are not assured of getting a response; the TSA states that the ombudsman will respond "only if circumstances warrant it."[82] Furthermore, those who do get a reply are asked to complete and return a Passenger Identity Verification Form (PIVF) as well as submit copies of various identification documents. Compliance with these requests will not necessarily generate a response, and those who do receive a response may only get a letter from the TSA that can be shown to the airlines in an effort to clarify the mistake. In some cases, this letter does no good — would-be passengers are still subject to search and scrutiny.[83]

In December 2004, Congress enacted legislation aimed at providing due process mechanisms for those wrongly placed on the No-Fly list. The law requires the TSA to create a system for travelers to correct the inaccurate information that caused their placement on the no-fly list. DHS was also required to create a Privacy and Civil Liberties Oversight Board to review various government actions that may have civil liberties implications for individuals.[84] In response, the TSA has now published "Watch List Clearance Procedures." Here, the TSA explains:

> The TSA clearance process will not remove a name from the Watch Lists. Instead this process distinguishes passengers from persons who are in fact on the Watch Lists by placing their names and identifying information in a cleared portion of the Lists. Airline personnel can then more quickly determine when implementing TSA-required identity verification procedures that these passengers are not the person of interest whose name is actually on the Watch Lists.

The TSA web site goes on to state:

> Clearance by TSA may not eliminate the need to go to the ticket counter in order to check-in. While TSA cannot ensure that this procedure will relieve all delays, we hope it will facilitate a more efficient check-in process for you. Additionally, TSA has issued guidance to the airlines to clarify further the Watch List clearance protocol.[85]

---

82. Class Action Complaint in Green v. TSA, *supra* note 69, at ¶ 30-34.

83. *Id.*

84. Sara Kehaulani Goo, *Law Lets Passengers Appeal No Fly List*, WASH. POST, Dec. 18, 2004, at A21, *available at* http://www.washingtonpost.com/wp-dyn/articles/A8986-2004Dec17.html.

85. TSA, Watch List Clearance Procedures,

B. CAPPS II

The Computer Assisted Passenger Pre-Screening System II (CAPPS II) has been one of the most controversial initiatives to propose the use of commercial data mining for the assessment of terror risk. With CAPPS II, there was a concern that data quality and integrity might be compromised by integrating commercial data from different sources — with different meanings for different sets of data.[86] The reason for the controversy may relate to the concept that the TSA would peek into private sector records in order to "code" people as different risk levels. CAPPS II was an attempt to use data mining of commercial and government data not just to verify identity, but to actually try and predict whether someone will be a security risk based on commercial data.[87]

The TSA was the agency tasked with implementing CAPPS II, which was an outgrowth of an earlier system known as the Computer Assisted Passenger Screening (CAPS) system, commenced in 1996 by Northwest Airlines.[88] As other airlines worked to implement CAPS, the White House Commission on Aviation Safety and Security (established by Executive Order following the crash of TWA Flight 800 in July of 1996, and sometimes referred to as the Gore Commission) recommended in its final report to President Clinton that the United States implement automated passenger profiling for its airports.[89] CAPS identified certain passengers as posing a security risk based on assumptions of how terrorists travel. Passengers were flagged for additional screening if they did things like purchase a one-way ticket or pay for a ticket with cash. CAPPS II was portrayed by

---

http://www.tsa.gov/public/display?theme=157&content=09000519800fb8af (last visited Apr. 6, 2006); *See also* TSA, How the Process Works, http://www.tsa.gov/public/interapp/editorial/editorial_1829.xml (for further explanation by the TSA of how the clearance process works) (last visited Apr. 6, 2006).

86. Valerie Alberto & Dominique Bogatz, *Computer Assisted Passenger Prescreening System ("CAPPS II"): National Security v. Civil Liberties* 4 (May 19, 2004), http://insct.syr.edu/Information%20Sharing%20Conf/AlbertoBogatz.pdf (Unpublished background paper prepared for Conference on Information Sharing and Homeland Security Program at the Institute for National Security and Counterterrorism, Syracuse University).

87. At least one commentator notes that CAPPS II is not data mining but rather "involves data matching against a watch list or data aggregation to confirm identity." Taipale, *supra* note 11, at 38.

88. SEIFERT, *supra* note 2, at 7-8. The program was piloted by Northwest Airlines and then was implemented by all of the airlines in 1998. *See* GEN. ACCOUNTING OFFICE, REPORT NO. GAO-04-385, AVIATION SECURITY COMPUTER-ASSISTED PASSENGER PRESCREENING SYSTEM FACES SIGNIFICANT IMPLEMENTEATION CHALLENGES 5 (2004) [hereinafter GAO, AVIATION SECURITY].

89. SEIFERT, *supra* note 2, at 8.

the TSA as an "enhanced system to confirm the identities of passengers and to identify foreign terrorists or persons with terrorist connections before they can board a U.S. aircraft." It was unveiled on August 1, 2003.

1. The CAPPS II Design

CAPPS II was designed to send the information provided by a passenger in the passenger's name record (referred to as a PNR, this included the passenger's name, full address, phone number and date of birth) to a commercial data provider so as to authenticate the identity of the passenger.[90] The commercial data broker would then send a numerical score back to the TSA to indicate the passenger's perceived risk level. Passengers with a "green" score would undergo normal screening, while passengers with a "yellow" score would be subject to enhanced screening. Passengers with a "red" score would not be allowed to board the flight, and would be interviewed or put in contact with law enforcement. The scoring and "coding" of passengers would be based on the percentage match between the commercial data and the data held by the TSA.

According to an August 2003 notice in the Federal Register, CAPPS II was meant to proceed through four steps:[91]

(1) Data Collection. Airlines were required to collect certain data from every passenger, and to pass it along to the TSA. Upon purchasing an airplane ticket, passengers were to provide four pieces of information: their name, address, telephone number, and date of birth.

(2) Identity Authentication. The TSA would send that airline information to commercial data services, which would then send back an "authentication score" intended to indicate "a confidence level in that passenger's identity."

(3) Risk Assessment. The TSA would then perform a risk assessment for each passenger, drawing upon law enforcement, intelligence, or other government databases. Each person would be scored as either an "acceptable," "unknown," or "unacceptable" risk.

90. *Id.* at 8-9.

91. GAO, AVIATION SECURITY, *supra* note 88, at 6-7.

(4) Enforcement of "Unacceptable" and "Unknown" Risk Assessments. Each passenger's risk score would then be forwarded to airport security personnel. Those who scored as an "unknown" risk would be subjected to heightened scrutiny. Those who received an "unacceptable" risk assessment would be denied boarding passes, and interviewed by law enforcement authorities to determine if they could board the flight.

CAPPS II was initially intended to detect terrorists and keep them off airplanes. In August 2003, however, TSA announced that it would also serve as a law enforcement tool to identify individuals wanted for violent crimes.[92] TSA had hoped to test the system at selected airports during the spring of 2004.

2. Shortcomings of CAPPS II

The CAPPS II screening procedure raised questions with respect to the accuracy of data mining and its reliability as a predictor of terrorist risk. What if a person's information was incorrect, or his name was similar to that of a known terrorist or criminal? What if someone else had assumed the passenger's identity? What are tolerable false positive and false negative rates when it comes to verifying identity, and how can mistakes be corrected? Or, what if a person, due to her age or lack of income or credit history, was not present in the databases, and so could not have her identity verified? It was unclear, for example, whether someone having a cell phone rather then a home phone number would be permitted to fly.[93] The TSA CAPPS II notice did not make clear the criteria for such assessments, and much of the data relied upon would have been confidential data, so that a passenger may never have known why he or she had been deemed an "unacceptable" risk.

---

92. Privacy Act of 1974: System of Records, 68 Fed. Reg. 45266 (Aug. 1, 2003) (outlining the CAPPS II System).

93. EFF Comments on CAPPS II, Letter to Privacy Office of the Department of Homeland Security, RE: Docket No. DHS/TSA-2003-1,
http://www.eff.org/Privacy/cappsii/20030930_comments.php.
Privacy advocates were concerned about the possible number of false positives should CAPPS II be fully implemented. The American Civil Liberties Union (ACLU) has pointed out that even a tiny error rate in the identification of suspected terrorists would create huge problems. With CAPPS II checking an estimated billion transactions, the ACLU points out, "[e]ven if we assume an unrealistic accuracy rate of 99.9%, mistakes will be made on approximately one million transactions, and *100,000 separate individuals*." (emphasis added). AM. CIVIL LIBERTIES UNION, THE FIVE PROBLEMS WITH CAPPS II (2003), *available at*
http://www.aclu.org/safefree/general/16769leg20030825.html.

Based on privacy concerns, Congress voted to block funding for CAPPS II unless the TSA could satisfy eight criteria relating to privacy, security, accuracy and oversight. In addition, Congress asked the Government Accounting Office (GAO) to conduct a review of CAPPS II to determine whether it met the relevant criteria.[94]

In February 2004, the GAO published its report on CAPPS II and concluded that CAPPS II had numerous problems.[95] The GAO report found that as of January 2004, the TSA had not adequately addressed seven of Congress' eight concerns. Why did the TSA fail so spectacularly? In part, the GAO noted, because it failed to timely test the CAPSS II program; according to the report, the TSA had not effectively managed and monitored CAPPS II's development and operation. [96] In addition, the TSA had also failed to protect passenger privacy; address the accuracy of the data relied upon; create a system to address erroneous labeling of passengers; prevent abuse; or create security procedures. [97] Such procedures are necessary to prevent hackers from compromising the data used in the screening process.

Further, TSA has also failed to adequately "stress test" CAPPS II to see if it even worked.[98] Did it really spot "high risk" passengers? Did it waste resources with false high-risk assessments? Based on the GAO findings, the answer was inconclusive. The report contains some troubling conclusions. For example, the GAO reported:

> TSA is also developing a redress procedure where passengers can attempt to get erroneous data corrected. However, it is unclear what access passengers will have to information found in either the governmental or commercial databases, or who is ultimately responsible for making corrections. Additionally, if errors are identified during the redress process, TSA does not have the authority to correct erroneous data in commercial or government databases. TSA officials said they plan to address this issue by establishing protocols with commercial data providers and other federal agencies to assist in the process of getting erroneous data corrected.[99]

---

94. GAO, AVIATION SECURITY, *supra* note 88, at 3; *See also* Dept. of Homeland Security Appropriations Act, Pub. L. No. 108-90, 117 Stat. 1137 (2003).

95. *See* GAO, AVIATION SECURITY, *supra* note 88 at 31, Appendix I for a list of the eight factors that the GAO was asked to evaluate. Many of the factors focused on due process, error rate, and privacy concerns.

96. *Id.* at 11-13.

97. *Id.* at 21-25.

98. *Id.* at 16.

99. *Id.*

The GAO also pointed out that CAPPS II might not be effective in detecting cases of identity theft.[100] As CAPPS II was ultimately cancelled, these shortcomings were never fully addressed.

3. The JetBlue/Torch Concepts Pilot Test

The TSA abandoned CAPPS II in August 2004. Since it was never fully implemented, there is much speculation as to how such a program would have worked in practice. There is one pilot test, however, which gives us a glimpse into potential methodology. In late 2003, JetBlue Airways confirmed reports that in 2002 it had provided five million passenger itineraries to a private defense contractor, Torch Concepts — *without* passenger consent.[101] The JetBlue controversy provides a useful window into how commercial data sets would be used to identify possible terrorist suspects.

As the *New York Times* reported, Torch Concepts had been hired by the Army "to determine how information from public and private records might be analyzed to help defend military bases from attack by terrorists and other adversaries."[102] In connection with this study, Torch Concepts contacted the TSA. And, according to a TSA spokesperson, the TSA facilitated the transfer of the JetBlue passenger data to Torch Concepts.

Torch Concepts separately purchased demographic data from Acxiom, a large data aggregating company. The data related to about 40% of the passengers who held JetBlue itineraries. The demographic data Acxiom provider for each passenger included: gender; whether the passenger owned or rented his or her residence: length of residence; economic status, including income; number of children; social security number; occupation; and vehicle information.[103] Torch Concepts matched the itineraries from JetBlue with the new data from Acxiom, and used the data as part of a study

---

100. *Id.* at 30 (noting that TSA officials "acknowledge that some identity theft is difficult to spot, particularly if the identity theft involves collusion, where someone permits his or her identity to be assumed by another person, is involved").

101. Anita Ramasastry, *Airline Passenger Profiling Based on Private Sector Data: Why It Raises Serious Privacy Concerns*, FINDLAW, Oct. 1, 2003, http://writ.news.findlaw.com/ramasastry/20031001.html. *See also* DEP'T OF HOMELAND SEC. PRIVACY OFFICE, REPORT TO THE PUBLIC ON EVENTS SURROUNDING JETBLUE DATA TRANSFER (Feb. 2004), http://www.epic.org/privacy/airtravel/jetblue/dhs_report.pdf.

102. Philip Shenon, *Airline Gave Defense Firm Passenger Files*, N.Y. TIMES, Sept. 20, 2003, at A1. *See also* Noah Schactman & Ryan Singel, *Army Admits Using JetBlue Data*, WIRED, Sept. 23, 2003, http://www.wired.com/news/privacy/0,1848,60540,00.html.

103. TORCH CONCEPTS, INC., HOMELAND SECURITY – AIRLINE PASSENGER RISK ASSESSMENT (2003), http://www.abditum.com/~rabbi/S3B3_Roark.pdf.

entitled "Homeland Security – Airline Passenger Risk Assessment."[104] That study was presented in February 2003 at a conference sponsored by the Tennessee Valley Chapter of the National Defense Industries Association. The Association then posted the presentation on its website, where it remained available until September 16, 2003.

In the study, Torch Concepts concluded that analysis of passenger demographics indicated distribution within three groupings, (with most travelers falling into the first two groups): (1) young middle income home owners with short length-of-residence; (2) older upper income home owners with longer length-of-residence; and (3) passengers with "anomalous records."[105] The third category, by definition, might potentially include renters, students with both home and school addresses, older persons who have moved recently, and persons with low incomes. Of course, such persons are in some senses the norm in America. Yet the program may have deemed them "anomalous" — and thus a risk from a security standpoint.

This should make those in the third group, in particular, nervous. People may face increased screening simply because they cannot afford to own their home, because they have recently relocated, or because, as students, they maintain both home and dorm addresses. The Torch Concepts study provides concrete example of how data mining can lead to adverse inferences and consequences for certain groups of people, not because of any criminal activity, but because their patterns replicate what has been identified as "suspicious" behavior.

Finally, there is a possibility that the data transfer may have violated the federal Privacy Act of 1974. However, while the Act governs databases that the U.S. government compiles, it does not regulate how government agencies and their contractors access private sector databases.

C. Secure Flight

After CAPPS II was scrapped, the TSA introduced another potential system for passenger screening, known as Secure Flight.[106]

---

104. *Id.*

105. *Id.*

106. OFFICE OF THE INSPECTOR GEN., U.S. DEP'T OF JUSTICE, AUDIT REPORT 05-34, REVIEW OF THE TERRORIST SCREENING CENTER'S EFFORTS TO SUPPORT THE SECURE FLIGHT PROGRAM, at 3-4 (Aug. 2005), *available at*

This program was meant to provide an enhanced system for verifying the identity of passengers, and to see if they matched up against a number of government terrorist/watch lists.[107] TSA had initially hoped to launch the program with two airlines in August 2005, but announced in February 2006 that it had suspended the program due to security concerns for passenger data.[108] More generally, Secure Flight was not to be launched until the GAO certified the program, which it has yet to do.

Secure Flight was meant to include the following four steps, each of which raises unique problems.[109] First, commercial airlines would be required to collect additional information from every passenger in PNRs, including date of birth, and transfer it to the TSA. Second, the TSA would send the passenger's PNRs (including names, dates of birth, and whatever other personal information is collected) to commercial data brokers or aggregators for authentication. These commercial brokers would report back to the TSA as to whether the information provided by the passenger matches the information in their records. Third, the TSA would run the passenger's name through watch lists maintained by the government's Terrorist Screening Center (TSC), an entity that is supposed to aggregate the many dispersed watch lists that the government was maintaining after 9/11. However, this step contains a problem: the evidence indicates that in the years since 9/11, these watch lists have not been properly consolidated or purged of erroneous names. An August 2004 report by the DHS' own Inspector General found continued problems with attempts to create a unified watch list.[110] In the final step, law enforcement must make a decision about how to proceed with a person identified as a possible risk.

---

http://www.usdoj.gov/oig/reports/FBI/a0534/final.pdf [hereinafter AUDIT REPORT 05-34]. *See also* Anita Ramasastry, *Secure Flight Set to take Off but Will our Data be Secure?*, FINDLAW, July 26, 2005, http://writ.news.findlaw.com/ramasastry/20050726.html.

107. *Id.*

108. Associated Press, *TSA's Secure Flight Suspended*, ASSOCIATED PRESS, Feb. 2, 2006, http://msnbc.msn.com/id/11254968.

109. For a useful overview, see GEN. ACCOUNTING OFFICE, REPORT NO. GAO-05-356, AVIATION SECURITY, SECURE FLIGHT DEVELOPMENT AND TESTING UNDER WAY, BUT RISKS SHOULD BE MANAGED AS SYSTEM IS FURTHER DEVELOPED 18-20 (2005), *available at* http://www.gao.gov/new.items/d05356.pdf [hereinafter GAO, SECURE FLIGHT DEVELOPMENT].

110. OFFICE OF THE INSPECTOR GEN., U.S. DEP'T OF HOMELAND SEC., OIG-04-31, DHS CHALLENGES IN CONSOLIDATING TERRORIST WATCH LIST INFORMATION 12-32 (2004), *available at* http://www.dhs.gov/interweb/assetlibrary/OIG-04-31_Watch_List.pdf.

In February 2005, the GAO issued a report on the use of commercial data in the Secure Flight Process.[111] The February GAO report focused solely on Secure Flight's use of commercial databases such as Choice Point, and found that the TSA had not developed successful measures by which to judge the performance of those commercial databases. Then, in March 2005, the GAO issued a second report, in response to a October 2004 request from Congress to evaluate Secure Flight in ten different areas. This second report was entitled "Initial Secure Flight Test Results Show Improvements over Current Passenger Prescreening, but Key Issues Regarding How Data Will Be Obtained and Transmitted Have Not Yet Been Resolved."[112]

The title says a great deal — and the report says more:

> [T]he ability of Secure Flight to make accurate matches between PNR data and data contained in the terrorist screening database is dependent on the type and quality of data contained in the database as well as in PNRs. While TSC and TSA have taken, or plan to take, a number of actions to improve the quality of the data in the terrorist screening database, *the accuracy of the database has not been determined.* The effectiveness of data matches will also be dependent on the accuracy of commercial data used to augment the matching, should TSA decide to use commercial data for Secure Flight. However, *the accuracy of commercial data is undetermined* because there are no industry standards for processes or requirements to ensure accuracy.[113]

GAO also noted that before Secure Flight goes forward, there must be a procedure for correcting erroneous data in the hands of data brokers: the TSA "will need to reach specific agreements with commercial data aggregators on a process for correcting erroneous information."[114]

The TSA also appointed a body of external experts to constitute the Secure Flight Working Group (SFWG). This group noted that the use of commercial data from identity matching, if done appropriately, could actually help to verify or authenticate passenger identity. However, in its report to the TSA, the SFWG noted that it had not received enough information to assess whether: (1) the information

---

111. GEN. ACCOUNTING OFFICE, REPORT NO. GAO-05-324, AVIATION SECURITY: MEASURE FOR TESTING THE IMPACT OF USING DATA FOR SECURE FLIGHT (2005), *available at* http://www.gao.gov/new.items/d05324.pdf [hereinafter GAO, DATA FOR SECURE FLIGHT].

112. GAO, SECURE FLIGHT DEVELOPMENT, *supra* note 109.

113. *Id.* at 5 (emphases added).

114. *Id.*

that would be provided from the airlines would be sufficient for data matching, and (2) whether commercial databases have the kind of data necessary to resolve the challenge of comparing passenger lists against terrorist watch lists.[115] The SFWG also did not receive information from the TSA indicating which hatching algorithms worked best. The SFWG, in fact, noted that it saw no evidence that the TSA had even compared different products and competing solutions. [116]

Even Secure Flight, which is not a predictive system, but a way to improve accuracy in matching passengers against consolidated watch lists, was found to be problematic. In essence, it appeared that the TSA had not made its case as to the accuracy of its program or the error resolution mechanisms it had proposed to use in the event of a false positive.

V. IMPROVING GOVERNMENT DATA MINING PRACTICES IN THE CONTEXT OF TERRORISM PREVENTION

The No-Fly List demonstrates that the use of data mining to draw inferences can have adverse consequences for Americans. This is a large reason why the TSA's recent forays into commercial data mining have been viewed with suspicion. The larger issue remains, however, about what steps the TSA will take to ensure that any future programs are properly devised and implemented. Government audits of CAPPS II, Secure Flight and other data mining initiatives indicate quite clearly that the TSA and other agencies have not paid sufficient attention to concerns of individual privacy, error correction and redress.[117]

Data mining, without adequate privacy safeguards, has the potential to be used as a tool to spy on American citizens without the judicial or procedural constraints that limit traditional surveillance techniques. To the extent that future government initiatives will link commercial databases to terrorism prevention, it is incumbent upon policymakers to rethink the issue on a larger scale. Rather than examining individual programs, it is time to envision what sort of privacy rules and practices should be used by both partners in the

115. REPORT OF THE SECURE FLIGHT WORKING GROUP, *supra* note 56, at 11-12.

116. *Id.* at 6.

117. MARKLE TASK FORCE REPORT, *supra* note 4, at 11 (stating that government data mining efforts "continue to provoke controversy because of the lack of systematic effort to consider the privacy implications of the proposed programs or to develop an overall policy framework that would govern the deployment of new technologies").

system — the commercial entities acting as agents of the state, and the government, which is purchasing commercial data for law enforcement purposes.

As Christopher Hoofnagle has pointed out, there is currently ambiguity as to whether the federal Privacy Act of 1974 applies to commercial data brokers who supply information to the government. The Privacy Act embodies so-called fair information practices, and requires the government to apply these practices to records that it maintains which contain personal information. It also prohibits the government from collecting personal information unless it has a "proper purpose" for doing so.[118]

If the government does collect data, such collection is subject to a series of rights: the federal government must give notice to the public of all of the databases it maintains; it must allow people to access and correct their data; and the data collecting must be limited only to information necessary to fulfill a specific government function. The Privacy Act, however, applies only to the federal government and companies administering a system of records for the government. At present, it is not clear whether a database which originates in the private sector, and is then used by the government, is subject to the Act. In fact, "credit-reporting agencies are specifically exempted from being considered a federal contractor for systems of records."[119] Thus, one important legislative fix would be to require that data brokers who supply information to the government for data mining be subject to the Privacy Act when serving a government function. Unless this is clarified, if the data is maintained in the private sector and only "accessed" by the government, it is unclear what types of access and error correction will be required of data brokers.

The Markle Foundation convened a Task Force on National Security in the Information Age, which examined the issue of government access to private sector data and attempted to come up with rules to help government decide when it is appropriate to examine commercial data, and how such data should be handled. One recommendation was that if a false positive imposes significant consequences on a person, the requirement for data accuracy needs to

118. Hoofnagle, *supra* note 26, at 622-23. *See also* JAMES DEMPSEY, CENTER FOR DEMOCRACY AND TECHNOLOGY, PRIVACY'S GAP: THE LARGELY NON-EXISTENT LEGAL FRAMEWORK FOR GOVERNMENT MINING OF COMMERCIAL DATA (2003), http://www.cdt.org/security/usapatriot/030528cdt.pdf.

119. Hoofnagle, *supra* note 26, at 623.

be stringent.[120] The Markle Task Force noted, for example, that "we believe that the government should not have routine access to personally identifiable information even if that information is widely available to the public."[121] A higher bar of relevance to a legitimate purpose applies to government acquisition, and that bar should be even higher with greater degrees of sensitivity. The Task Force proposes a three-tier standard for data classification and acquisition — low, medium and high. This is a useful first step.

But perhaps the best way to begin to imagine how we can safeguard privacy in the wake of data mining is to require the government to provide robust data-mining privacy impact assessments. To date, the TSA and other federal agencies have not provided transparent and meaningful assessments of existing or proposed data mining initiatives.[122] What should a privacy impact assessment for a data-mining program (whether it is CAPPS II or Secure Flight) look like? It should begin with the basics — information that, surprisingly, the government has been unwilling to disclose in the context of TIA and CAPPS.

First, information about the databases themselves should be provided. What kind of data will be compiled, and, from what sources? How will the data be merged? Will it include information individuals tend to consider highly private, such as medical data, data on children, and financial data? Second, information about access to the data and databases should be provided. With whom can the data be shared, and upon what showing, if any? Which authorities, in particular, will have access? Can only federal authorities access the data, or can state authorities access it as well? What about foreign governments who are U.S. allies? What about private companies involved in defense or security as government contractors, or private domestic and foreign airline companies? Third, how will limits on data sharing be enforced? What are the legal and technological guarantees that those in the private sector, or those in the government who lack authorization, will *not* be able to access the data and databases? How can hackers be prevented from accessing the data?

---

120. MARKLE TASK FORCE REPORT, *supra* note 4, at 60.

121. *Id* at 33.

122. GEN. ACCOUNTING OFFICE, REPORT. NO. GAO-05-866, DATA MINING: AGENCIES HAVE TAKEN KEY STEPS TO PROTECT PRIVACY IN SELECTED EFFORTS, BUT SIGNIFICANT COMPLIANCE ISSUES REMAIN 33 (2005), http://www.gao.gov/new.items/d05866.pdf. *See also* Aliya Sternstein, *GAO: Federal Data Mining not Obeying Federal Privacy Rules*, FED. COMPUTER WEEK, Aug. 29, 2005, http://www.fcw.com/article90517-08-29-05-Web.

After these basics are out of the way, a cost-benefit analysis must begin. For instance, is it truly worthwhile to pay the privacy cost of including very sensitive information, such as medical data, or to allow wide access to the databases? Or is the benefit of including these types of information modest enough that its collection should be eliminated or curtailed? The chance of catching a terrorist due to the inclusion of individuals' medical histories may be slight, and not worth the privacy cost. Similarly, the chance of catching terrorists as a result of allowing local government officials to access the databases may be slight, and not worth the risk that employees may try to check up on their neighbors.

A cost-benefit analysis ought also to be made with respect to particular technologies. A search system may be advantageous in certain ways, but may also turn up a large number of false positives. Are the false positives — and resulting wrongful accusations — too great a cost? Perhaps a different system, one that is more accurate but casts a narrower net, is preferable. Once these decisions about the nature of the searching system, the databases, and the data are made, they must be made public. If there are procedures for protecting privacy, they should be made public so that people will know if they have been victimized by a violation.

If the government uses data compiled in the private sector, such data may very well contain errors. The government must consider how errors in its databases — errors that may prevent innocent persons from flying, or have other consequences — will be corrected. A related question is how often the databases will be updated and revised. Procedures for error correction are needed and should be well publicized, so that no one is forced to live with a mistake in his or her data that limits the freedom to travel – or any other freedoms. If a citizen is falsely identified as a risk, he or she needs to be able to clear his or her name from any security or watch lists, for example.

Finally, any data-mining program needs to have a redress mechanism. Arguably, Americans must not only be able to correct database errors, but also, if they are harmed by privacy violations, they must be able to take specific legal recourse against the government. First, passengers should know why they are on the lists. For instance, they should know what person or organization reported them — unless that interferes with national security. Some passengers on the No-Fly/Selectee lists have expressed concern that they may have been singled out because of their ethnicity, religion, or political activity. Second, there should be a fair, uniform process to

be used by passengers who seek to get off the lists. The failures of the TSA ombudsman are obvious, and must be corrected.

Everyone must be told, through notices posted at airports, how to contact the relevant decision maker and what the procedures are for attempting to get removed from the list. There need to be standards set for removal. And everyone who complains must get a response — a *reasoned* response. In deciding what the response should be, the decision maker must adhere to rules, treating like cases alike. The decision maker must give would-be passengers as much information as possible, consistent with national security, as to how they were placed on the lists. Negative decisions should be subject to appeal. Finally, when a person is cleared, he or she ought to be taken off the list — not just given a letter to show to airlines and airport security.

The lack of due process provides a gap in which discrimination and other abuses can become exacerbated. Blacklisting has drastic consequences for people, and can cause lasting harm to a person. Yet until the government watch-lists comply with due process, blacklists are what they will be.

VI. CONCLUSION

The "No-Fly" list provides a telling example of the problems created by incorrect inferences. This does not mean that government use of data mining should be prohibited. In fact, a government's effort to tap into commercial databases was meant to make identity verification more robust and ultimately more reliable. In other words, in a best-case scenario, governmental data mining could identify and distinguish, from among a group of persons who share a common name, one who truly poses a security risk from one who does not.

What is currently missing from the equation is effective debate and discussion on how to make data mining processes more protective of individual privacy, and how to safeguard the rights of individuals who are wrongly targeted for adverse government action as a result of adverse inferences. The public needs to understand how data is being translated from the private sector into the public sector, and what sort of interpretations are being made by government decision makers with that data.