# Government Data Mining and the Fourth Amendment

*Christopher Slobogin*†

## INTRODUCTION

The government's ability to obtain and analyze recorded information about its citizens through the process known as data mining has expanded enormously over the past decade. Since at least the mid-1990s, the quantity of the world's recorded data has doubled every year.[1] At the same time, the computing power necessary to store, access, and analyze these data has increased geometrically, at increasingly cheaper cost.[2] Governments that want to know about their subjects would be foolish not to take advantage of this situation, and federal and state bodies in this country have done so with alacrity.

For a time, most government data mining in the United States was devoted to ferreting out fraud against the government and monitoring the effectiveness of various programs.[3] In more recent years, however, and especially since 9/11, government agencies have been eager to experiment with the data mining process as a way of nabbing criminals and terrorists. Although details of their operation often remain murky, a number of such programs have come to light since 2001.

The best-known government data mining operation supposedly no longer exists. The federal program formerly known as Total Information Awareness—more recently dubbed Terrorism Information Awareness (TIA)—was sufficiently mysterious that Congress decided to cut off funding for it in 2003.[4] Spurred by rumors that TIA would involve the accumulation and analysis of vast amounts of data about the everyday transactions of American citizens, and probably influ-

† Stephen C. O'Connell Professor of Law, University of Florida Fredric G. Levin College of Law. The author would like to thank participants in workshops at Stanford Law School and Florida Law School for their feedback on the content of this article, and Victoria Ianni for her research assistance. This paper is a version of a talk given at The University of Chicago Law School's Surveillance Symposium, June 15–16, 2007.

[1] Jeffrey W. Seifert, *Data Mining and Homeland Security: An Overview* 2 (Congressional Research Service, Jan 18, 2007), online at http://www.fas.org/sgp/crs/intel/RL31798.pdf (visited Jan 12, 2008).

[2] See id.

[3] Id at 4.

[4] See Consolidated Appropriations Resolution, 2003, Pub L No 108-7, 117 Stat 11, 534 (stating that "no funds appropriated or otherwise made available to the Department of Defense . . . may be obligated or expended on research and development on the Total Information Awareness program, unless [statutory exceptions apply]").

enced as well by TIA's icon—an eye on top of a pyramid looking over the globe, accompanied by the slogan "knowledge is power"—even a majority of Republicans voted in favor of the bill to end the program.[5] Coming within two years of September 11, 2001, and against the background of otherwise unrestrained congressional enthusiasm for expansive government authority to combat terrorism, the anti-TIA vote appeared, on the surface at least, to signal a particular hostility toward computerized data aggregation and dissection.

Yet large-scale data mining by federal agencies devoted to enforcing criminal and counterterrorism laws has continued unabated. The legislation that limited TIA's reach still permitted the Defense Department and other agencies, after "appropriate consultation with Congress," to pursue data mining of records, on American as well as foreign citizens, for the purpose of gathering information relevant to "law enforcement activities" as well as foreign intelligence.[6] The government has taken full advantage of this authority. Beginning soon after the passing of TIA, it spent at least $40 million developing a program called ADVISE (for Analysis, Dissemination, Visualization, Insight, and Semantic Enhancement), which was designed "to troll a vast sea of information, including audio and visual, and extract suspicious people, places and other elements based on their links and behavioral patterns."[7] More recently, it has become entranced with the concept of "fusion centers," which are, as described by one commentator, "an amalgamation of commercial and public sector resources for the purpose of optimizing the collection, analysis, and sharing of information on individuals," designed to gather data about banking and finance, real estate, education, retail sales, social services, transportation, postal and shipping, and hospitality and lodging transactions.[8] As of September 2006, there were thirty-eight state and local government Information Fusion Centers, supported by $380 million in federal funding.[9]

[5] The Senate passed the measure by a voice vote. See 149 Cong Rec S 1379-02, 1416 (Jan 23, 2003); *Senate Rebuffs Domestic Spy Plan*, Reuters (Jan 23, 2003), online at http://www.wired.com/politics/law/news/2003/01/57386 (visited Jan 12, 2008).

[6] Consolidated Appropriations Resolution, 2003, 117 Stat at 536 ("[T]he Total Information Awareness program should not be used to develop technologies for use in conducting intelligence activities or law enforcement activities against United States persons without appropriate consultation with Congress or without clear adherence to principles to protect civil liberties and privacy.").

[7] Ellen Nakashima and Alec Klein, *Profiling Program Raises Privacy Concerns*, Wash Post B1 (Feb 28, 2007). See also generally Shane Harris, *TIA Lives On*, Natl J 66 (Feb 25, 2006) (describing a variety of surveillance programs that survived the formal closure of TIA).

[8] Lillie Coney, Statement to the Department of Homeland Security Data Privacy and Integrity Advisory Committee 1, 4 (Electronic Privacy Information Center, Sept 19, 2007), online at http://www.epic.org/privacy/fusion/fusion-dhs.pdf (visited Jan 12, 2008) (claiming that "[i]t would be very difficult to imagine someone" who would not be included in the system).

[9] Id at 3.

And this program is just the tip of the iceberg. According to a GAO report issued in 2004, just one year after TIA's demise, 52 federal agencies were using or were planning to use data mining, for a total of 199 data mining efforts, 68 planned and 131 operational.[10] Of these programs, at least 122 are designed to access "personal" data.[11]

The Defense Department, the progenitor of TIA, sponsors the largest number of data mining operations.[12] One such program is called Verity K2 Enterprise, which mines data from the intelligence community and internet searches in an effort to identify foreign terrorists or US citizens connected to foreign intelligence activities.[13] Another is known as Pathfinder, which provides the ability to rapidly analyze and compare government and private sector databases.[14] There is also TALON (Threat and Local Observation Notice), a program which has collected information on thousands of American citizens involved in protesting the war in Iraq and other government policies, and made the data accessible to twenty-eight government organizations and over 3,500 government officials.[15] A fourth Defense Department program, apparently not named, has accumulated files on hundreds of Americans suspected of being spies, which contain information from their banks, credit card companies, and other financial institutions. The Pentagon plans to keep these files indefinitely, even though to date apparently no arrests have resulted.[16]

Many other government agencies are also involved in data mining. The fusion center initiative, which appears to be the new TIA, is operated by the Department of Homeland Security. The DOJ, through the FBI, has been collecting telephone logs, banking records, and other personal information regarding thousands of Americans not only in connection with counterterrorism efforts,[17] but also in further-

[10] GAO, *Data Mining: Federal Efforts Cover a Wide Range of Uses*, GAO-04-548, 7 (May 2004), online at http://www.gao.gov/new.items/d04548.pdf (visited Jan 12, 2008).

[11] Id at 10.

[12] See id at 7 (noting, however, that the CIA and NSA did not respond to the audit). See also James Bamford, *Private Lives: The Agency That Could Be Big Brother*, NY Times sec 4 at 1 (Dec 25, 2005).

[13] GAO, *Data Mining* at 30 (cited in note 10).

[14] Id.

[15] Walter Pincus, *Protesters Found in Database; ACLU Is Questioning Entries in Defense Dept. System*, Wash Post A8 (Jan 17, 2007). In April 2007, the Pentagon announced it would be ending the program. Mark Mazzetti, *Pentagon Intelligence Chief Proposes Ending a Database*, NY Times A18 (Apr 25, 2007).

[16] Eric Lichtblau and Mark Mazzetti, *Military Expands Intelligence Role in U.S.*, NY Times sec 1 at 1 (Jan 14, 2007) (describing claims that the documents are useful "even when the initial suspicions are unproven").

[17] The most prominent effort in this regard is the FBI's System-to-Assess-Risk (STAR) program, which makes use of the Foreign Terrorist Tracking Task Force "Data Mart," consisting of a wide array of sources, to acquire more information about suspected terrorists and other

ance of ordinary law enforcement.[18] And it was disclosed in January 2007 that the IRS and the Social Security Administration made more than twelve thousand "emergency disclosures" of personal data to federal intelligence and law enforcement agencies in 2002 and thousands more such disclosures each year since then, often via a program called REVEAL that combines sixteen government databases with databases maintained by private companies.[19]

As this last example illustrates, many of these programs rely in whole or in part on private companies, called commercial data brokers, to provide their input, which is then analyzed by government officials. Companies like Acxiom, Docusearch, ChoicePoint, and Oracle can provide the inquirer with a wide array of data about any of us, including basic demographic information, income, net worth, real property holdings, social security number, current and previous addresses, phone numbers and fax numbers, names of neighbors, driver records, license plate and VIN numbers, bankruptcy and debtor filings, employment, business and criminal records, bank account balances and activity, stock purchases, and credit card activity.[20] The government routinely makes use of these services. Even in the years *before* 9/11, ChoicePoint and similar services ran between fourteen thousand and forty thousand searches per month for the United States Marshals Service alone.[21]

Because of the ubiquity of these private companies, even state governments, which otherwise might not have the resources to engage in data mining, have entered the field. Here the best known commercial data broker is Seisint, a concern now owned by LexisNexis. According to its advertising, Seisint, through its subsidiary Accurint (for accurate intelligence) can, in mere seconds, "search tens of billions of data records on individuals and businesses," armed with no more than a name, address, phone number, or social security number.[22] All of this was for a time made accessible to state law enforcement officials with

"persons of interest." DOJ, *Report on "Data-mining" Activities Pursuant to Section 126 of the USA Patriot Improvement and Reauthorization Act of 2005* 7–10 (July 9, 2007), online at http://www.epic.org/privacy/fusion/doj-dataming.pdf (visited Jan 12, 2008).

[18] See David Johnston and Eric Lipton, *U.S. Report to Fault Wide Use of Special Subpoenas by F.B.I.*, NY Times A1 (Mar 9, 2007).

[19] Dalia Naamani-Goldman, *Anti-terrorism Program Mines IRS' Records; Privacy Advocates Are Concerned That Tax Data and Other Information May Be Used Improperly*, LA Times C1 (Jan 15, 2007).

[20] For a description of the types of information data brokers can produce, see Laura K. Donohue, *Anglo-American Privacy and Surveillance*, 96 J Crim L & Criminol 1059, 1142 (2006).

[21] Chris Jay Hoofnagle, *Big Brother's Little Helpers: How ChoicePoint and Other Commercial Data Brokers Collect and Package Your Data for Law Enforcement*, 29 NC J Intl L & Comm Reg 595, 600 (2004).

[22] *Accurint Overview*, online at http://www.accurint.com/aboutus.html (visited Jan 12, 2008).

the establishment of MATRIX (Multi-state Anti-terrorism Information Exchange), a consortium funded in part by the federal government that allowed state police to use Accurint for investigative purposes.[23] Although today the scope of MATRIX is much reduced, it still flourishes in Florida and elsewhere.[24]

This paper addresses three puzzles about data mining. First, when data mining is undertaken by the government, does it implicate the Fourth Amendment? Second, does the analysis change when data mining is undertaken by private entities that then make the data or data analysis available to the government? Third, if the Fourth Amendment does impose some restrictions on government data mining, what might they look like? Current Fourth Amendment jurisprudence appears to leave data mining completely unregulated, while most commentators have called for stringent regulation or a prohibition on large-scale operations such as TIA.[25] I end up taking an intermediate position on these issues. A careful look at data mining suggests that many versions of it should be only minimally regulated, while other versions ought to be subject to significant constitutionally based restrictions, whether controlled solely by the government or reliant on private entities for information. In aid of this project, I describe a study that investigated lay views on data mining.

Part I describes data mining and its effects in a bit more detail. Part II sketches current Fourth Amendment doctrine. Finally, Part III suggests how that doctrine might be interpreted to require limitations on government data mining. The proposed framework requires attention to the type of records obtained via data mining, the extent to which they can be connected to particular individuals, and the government's goal in obtaining them. Based on proportionality reasoning that I have applied in other contexts, the highest degree of justification for data mining should be required when the data is private in nature and sought in connection with investigation of a particular target. In contrast, data mining that relies on impersonal or anonymized

[23] See Donohue, 96 J Crim L & Criminol at 1151 (cited in note 20).

[24] See id at 1151–52. MATRIX downsized in large part because states involved in the consortium were concerned about both costs and privacy. See id.

[25] See, for example, Anita Ramasastry, *Lost in Translation?: Data Mining, National Security and the "Adverse Inference" Problem*, 22 Santa Clara Computer & High Tech L J 757, 794 (2006) ("[P]erhaps 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."); Jay Stanley and Barry Steinhardt, *Bigger Monster, Weaker Chains: The Growth of an American Surveillance Society* 12 (ACLU, Jan 2003), online at www.aclu.org/FilesPDFs/aclu_report_bigger_monster_weaker_chains.pdf (visited Jan 12, 2008) (asserting that if programs like TIA are allowed to continue we will "have the worst of both worlds: poor security and a supercharged surveillance tool that would destroy Americans' privacy and threaten our freedom").

records, or that is sought in an effort to identify a perpetrator of a past or future event, need not be as strictly regulated.

# I. DATA MINING AND ITS EFFECTS

Sensible regulation of data mining depends on understanding its many variants and its potential harms. Data mining programs all have analogues in traditional investigative techniques involving records. But given its scope, its potential harms can be much more significant than those associated with these traditional practices.

## A. A Typology of Data Mining

Data mining for governmental purposes can be divided into numerous categories. Already mentioned is the fact that it can be either run entirely by the government or largely dependent on private data brokers. Data mining can also differ with respect to the type of data acquired, the degree to which the data are aggregated and associated with an identified person, and the extent to which personal information is knowingly provided to the collecting entity. All of these factors might be relevant in thinking through whether and how data mining by the government should be regulated and thus will be discussed in later parts of this article. But the most fundamentally useful categorization of data mining for legal purposes focuses on its goal. Data mining can either be target-driven, match-driven, or event-driven.

*Target-driven data mining*, sometimes called subject-based data mining, is a search of records to obtain information about an identified target. The REVEAL and FBI programs described earlier are good examples of this type of data mining. Both sift through "personal" records—tax records, bank records, phone and ISP logs—in an effort to find out more about particular individuals who are suspected of engaging in illegal activity.

*Match-driven data mining* programs are designed to determine whether a particular individual has already been identified as a "person of interest." In other words, the goal here is not to find out more about a suspect, but rather to determine whether a particular person is a known suspect. A good example of match-driven data mining is the program once known as the Computer Assisted Passenger Prescreening System (CAPPS II), and then as Secure Flight, a "no-fly list" that supposedly compares airline passengers to lists of known or suspected terrorists and produces a particular risk level with respect to each passenger.[26] Comparison of a suspect's DNA or fingerprints to a national database is another example of match-driven data mining.

*Event-driven data mining*, also called pattern-based surveillance, is data mining designed to discover the perpetrator of a past or future event; in contrast to both target-based and match-based data mining, this type of data mining does not start with an identified suspect. A simple example of such data mining, apparently actually used in an effort to track down a serial rapist, involved a search of residential records to determine the names of males who had lived in Philadelphia, Pennsylvania and Fort Collins, Colorado at the time rapes with a similar modus operandi were committed in those two cities (forty males were identified who were investigated further).[27] More complicated versions, such as TIA and ADVISE, use algorithms that are thought to correlate with a past or future event. For instance, TIA consisted of a number of operations designed to gather vast amounts of information useful to targeting terrorist activity. According to literature created by TIA's progenitors, the program had three articulated goals: (1) to increase access to counterterrorism information "by an order of magnitude" (to be accomplished through the Genisys program); (2) to accumulate "patterns that cover at least 90 percent of all known previous foreign terrorist attacks" and "[a]utomatically cue analysts based on partial pattern matches" (the objective of the Evidence Extraction and Link Discovery program); and (3) to "[s]upport collaboration, analytical reasoning, and information sharing so analysts can hypothesize, test, and propose theories and mitigating strategies about possible futures" (to be implemented through the previous two programs and the Scalable Social Network Analysis algorithms program).[28] Put in plain English, TIA was an attempt to use computers to sift through a large number of databases containing credit card purchases, tax returns, driver's license data, work permits, and travel itineraries to discover or apply patterns predictive of terrorist activity.

## B. The Benefits and Harms of Data Mining

The potential benefits of data mining are clear. Target-based programs such as REVEAL and MATRIX have helped apprehend or

[26] See Seifert, *Data Mining and Homeland Security* at 9, 11 (cited in note 1) (describing how the program would use data provided to the airline and then return a green, yellow, or red indication).

[27] William J. Krouse, *The Multi-state Anti-terrorism Information Exchange (MATRIX) Pilot Project* 9 (Congressional Research Service, Aug 18, 2004), online at http://www.fas.org/irp/crs/RL32536.pdf (visited Jan 12, 2008).

[28] See Defense Advanced Research Projects Agency (DARPA), *Report to Congress regarding the Terrorism Information Awareness Program* 3–9 (May 20, 2003), online at http://www.eff.org/Privacy/TIA/TIA-report.pdf (visited Jan 12, 2008).

develop cases against numerous criminals.[29] Match-based programs like CAPPS II have undoubtedly kept some dangerous individuals off planes and probably deterred others from trying to get on.[30] Event-based data mining has helped the government recover millions of dollars in fraudulent Medicare payments, detect money laundering and immigrant smuggling operations, and solve identity theft cases.[31]

The costs of data mining can be significant as well. A first, obvious cost is that data mining might lead to the wrong people being arrested, kept off airplanes, or subject to further investigation. Unfortunately, those occurrences are routine, for numerous reasons.

Most fundamentally, the information in the records accessed through data mining can be inaccurate. The government's no-fly list, for instance, is notorious for including people who should not be blacklisted.[32] Even more prosaic records are astonishingly inaccurate. Approximately one in four credit reports contain errors serious enough to result in a denial of credit, employment, or housing.[33] According to one study, 54 percent of the reports contain personal demographic information that is misspelled, long outdated, belongs to a stranger, or is otherwise incorrect.[34] Even if the information is accu-

[29] MATRIX is said to have assisted law enforcement officials in almost one thousand cases in a two-year period, primarily in terms of tracking down suspects and victims. See Florida Department of Law Enforcement, News Release, *MATRIX Pilot Project Concludes* (Apr 15, 2005).

[30] Proponents of Secure Flight assert that, at worst, its margin of error is 30 percent and may be as low as 2 percent. See sources cited in Stephen W. Dummer, Comment, *Secure Flight and Dataveillance, A New Type of Civil Liberties Erosion: Stripping Your Rights When You Don't Even Know It*, 75 Miss L J 583, 606 nn 128–29 (2006). These data are highly suspect, however. See GAO, *Aviation Security: Computer-assisted Passenger Prescreening System Faces Significant Implementation Challenges* 15 (Feb 2004), online at http://www.gao.gov/new.items/d04385.pdf (visited Jan 12, 2008) ("[A] senior program official said that TSA has no indication of the accuracy of information contained in government databases."); Dummer, 75 Miss L J at 607 & n 131 (reporting that between 400 and 1,200 innocent people will be flagged per day).

[31] GAO, *Data Mining* at 9 (cited in note 10) (describing the C & P Data Analysis program used by the Veterans Benefits Administration); Hearing before the House Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census, House Committee on Governmental Reform 3–4 (March 25, 2003) (testimony of Florida state Senator Paula B. Dockery) (discussing the use of data mining to investigate money laundering and narcotics smuggling); George Cahlink, *Data Mining Taps the Trends; Data Mining Helps Managers Make Sense and Better Use of Mounds of Government Data*, Gov Exec Mag 85 (Oct 1, 2000) (reporting that tracking down fraud is the most common use of data mining).

[32] Justin Florence, *Making the No Fly List Fly: A Due Process Model for Terrorist Watchlists*, 115 Yale L J 2148, 2153 (2006).

[33] US PIRG, *Mistakes Do Happen: A Look at Errors in Consumer Credit Reports* (June 17, 2004), online at http://www.uspirg.org/home/reports/report-archives/financial-privacy--security/financial-privacy--security/mistakes-do-happen-a-look-at-errors-in-consumer-credit-reports (visited Jan 12, 2008).

[34] Id.

rate, integrating disparate databases may lead to distortions in the information obtained, and computers or analysts can misconstrue it.[35]

With event-driven data mining, inaccuracy is heightened by the difficulty of producing useful algorithms. Even when the base rate for the activity in question is relatively high (for example, credit card fraud) and the profile used is highly sophisticated, data mining will generate more "false positives" (innocent people identified as criminals) than true positives.[36] When the base rate of the criminal activity is low (for example, potential terrorists) and the algorithm less precise (as is probably true of any "terrorist profile"), the ratio of false positives to true positives is likely to be extremely high.[37] In fact, what little we know suggests the government's event-driven antiterrorist data mining efforts have been singularly unsuccessful.[38]

The use of algorithms that produce a high false positive rate exacerbates two other phenomena: invidious profiling and what data mining aficionados call "mission creep." Match- and event-driven data mining can be, and probably have been, heavily dependent on ethnic, religious, and political profiling;[39] while such discrimination is a possibility during traditional investigations as well, it is vastly facilitated by

[35] Seifert, *Data Mining and Homeland Security* at 22 (cited in note 1) (discussing "interoperability" problems associated with searching and analyzing multiple, disparate databases).

[36] Amy Belasco, *Total Information Awareness Programs: Funding, Composition, and Oversight Issues* 15–16 (Congressional Research Service, Mar 21, 2003), online at http://usacm.acm.org/usacm/PDF/CRSTIAReport.pdf (visited Jan 12, 2008) (positing a 2.6:1 false positive rate in credit card fraud investigations).

[37] See id at 16 (providing an example producing a 200:1 false positive rate).

[38] According to the *New York Times*, the NSA program generated thousands of tips in the months following 9/11 but virtually none panned out. Lowell Bergman, et al, *Domestic Surveillance: The Program; Spy Agency Data after Sept. 11 Led F.B.I. to Dead Ends*, NY Times A1 (Jan 17, 2006) (reporting how the NSA flooded the FBI with tips, virtually all of which were "dead ends or innocent Americans"). See also note 16 and accompanying text. Seisint claimed to have generated a list of 120,000 names with "High Terrorist Factor" (HTF) scores and that "scores of arrests" were made based on this information. The validity of these arrests, assuming they occurred, has not been corroborated, and the HTF feature was reportedly dropped because of concerns about privacy abuses. Brian Bergstein, *Database Measured "Terrorism Quotient*,*"* AP (May 23, 2004).

[39] For examples of profiling in domestic spying, see ACLU, *FBI Counterterrorism Unit Spies on Peaceful, Faith-Based Protest Group* (May 4, 2006), online at http://www.aclu.org/safefree/spying/25442prs20060504.html (visited Jan 12, 2008) (describing the results of a FOIA request showing the FBI spied on School of the Americas Watch); William E. Gibson, *Boca Activist Blasts Spying Acts: Anti-Bush Groups Targeted, He Says*, S Fla Sun-Sentinel 3A (Jan 21, 2006) (reporting on the use of a domestic spying program to investigate the Truth Project, a political group adverse to President Bush's politics); Douglas Birch, *NSA Used City Police as Trackers; Activists Monitored on Way to Fort Meade War Protest, Agency Memos Show*, Baltimore Sun 1B (Jan 13, 2006); Matthew Rothschild, *Rumsfeld Spies on Quakers and Grannies*, The Progressive (Dec 16, 2005), online at http://progressive.org/mag_mc121605 (visited Jan 12, 2008) (criticizing Pentagon "political spying" and linking to a partial spreadsheet from the Pentagon listing some of the targeted political groups).

computers. And match- or event-driven data mining designed to ferret out terrorists can easily transform into a campaign to grab illegal immigrants, deadbeat dads, and welfare scammers. The CAPPS II program, for instance, appears to have been used to identify *any* individual who is in the country illegally.[40] The terrorist watchlist has now grown to over one-half million subjects, suggesting a very broad definition of terrorism.[41] These are not necessarily unmitigated harms, of course, but they should be recognized as a likely byproduct of data mining operations.

Erroneous or inappropriate government actions are not the only costs of data mining. Another problem is the threat large databases pose to innocent people's property and livelihood from entities *other* than the government. The desire for efficient data mining creates pressure to accumulate all information in one central repository. As Larry Ellison, the head of Oracle, stated, "The biggest problem today is that we have too many [databases]. The single thing we could do to make life tougher for terrorists would be to ensure that all the information in myriad government databases was integrated into a single, comprehensive national security file."[42] That may be true. But a single database makes it all that much easier for identity thieves and mischief-makers (inside as well as outside the government[43]) to do their dirty work because accessing records is that much easier.

All of these concerns can add up to a sense of unease about data mining. For those innocent people who are kept off airplanes, interviewed, or arrested based on erroneous data, or who lose their identities because of government sloppiness, the unease is palpable. For the rest of us, the harm is admittedly not as obvious. Many of those whose records are accessed through data mining don't know it is happening,

[40] The federal government has admitted as much with respect to immigrants. See *Privacy Act of 1974: System of Records*, 68 Fed Reg 45265-01, 45268 (2003) (describing "[r]outine uses of records maintained in the system . . . [by] Federal, State, local, international, or foreign agencies or authorities, including those concerned with law enforcement, visas and immigration"). See also Lara Jakes Jordan, *Audit: Anti-Terror Case Data Flawed*, AP (Feb 21, 2007).

[41] Justin Rood, *FBI Terror Watch List "Out of Control,"* ABC News: The Blotter (June 13, 2007), online at http://blogs.abcnews.com/theblotter/2007/06/fbi_terror_watc.html (visited Jan 12, 2008).

[42] Larry Ellison, *Digital IDs Can Help Prevent Terrorism*, Wall St J A26 (Oct 8, 2001).

[43] A number of prosecutions have been brought against government officials who have misused databases. See generally, for example, *United States v Stanley*, 2006 WL 2792904 (ND Okla) (upholding an indictment against Tulsa police officers charged with "theft" of confidential information stored in police department computers); *United States v Czubinski*, 106 F3d 1069 (1st Cir 1997) (reversing the conviction of an IRS employee prosecuted for trolling IRS databases for personal enjoyment). Two other cases involving allegedly similar facts are *United States v Fudge* (convicting an FBI analyst of improperly using law enforcement databases) and *United States v Pellicano* (alleging bribery of Los Angeles police to obtain access to law enforcement databases), neither of which are reported. Email from Howard W. Cox, Assistant Deputy Chief, DOJ Computer Crime & Intellectual Property Section, to Christopher Slobogin (Oct 25, 2007).

and if nothing incriminating is found, may never find out. But we still know that data mining allows the government to accumulate and analyze vast amounts of information about us, sufficient perhaps to create what some have called personality or psychological "mosaics" of its subjects.[44] That capacity for data aggregation may be a cost in itself. As Daniel Solove has argued, one result of government's entry into the information age is that faceless bureaucrats will be able to compile dossiers on anyone and everyone, for any reason or for no reason at all.[45] The possibility, even if slim, that this information could somehow be used to our detriment or simply revealed to others can create a chilling effect on all activity. It may have been some vague sense of this possibility that led Congress, however ineffectually, to declare its opposition to the concept of *Total* Information Awareness, with its epithet "knowledge is power."

Commercial data brokers are already constructing dossiers on us, of course. As Larry Ellison of Oracle has stated, his database "is used to keep track of basically everything";[46] a representative of Google has likewise stated that the company's mission "is to organize all the information in the world."[47] But when this information ends up in the hands of the government, with its enormous power to deprive people of liberty and property and the wide range of behavior that can be considered grounds for such deprivation, the calculus arguably changes even for the completely innocent. Knowing that the government is obsessed with fighting terrorism (as perhaps it should be) and that it views data mining as an essential tool in that fight, one could be forgiven for feeling inhibited about making certain calls (to a Muslim acquaintance?), traveling to certain locations (the Middle East?), and buying certain items (Halal meat, literature criticizing the war?).

These potential costs of data mining do not necessarily outweigh its benefits. But they at least suggest that data mining by the government should be subject to some regulation.

[44] Anthony Paul Miller, *Teleinformatics, Transborder Data Flows and the Emerging Struggle for Information: An Introduction to the Arrival of the New Information Age*, 20 Colum J L & Soc Probs 89, 111–12 (1986).

[45] See Daniel J. Solove, *The Digital Person: Technology and Privacy in the Information Age* 177–80 (NYU 2004).

[46] Jeffrey Rosen, *The Naked Crowd: Reclaiming Security and Freedom in an Anxious Age* 113 (Random House 2004).

[47] Barbara Cassin, *The New World According to Google*, Le Nouvel Observateur (Feb 8, 2007), online at http://hebdo.nouvelobs.com/p2205/articles/a332473.html (visited Jan 12, 2008). An unofficial English translation is online at http://www.truthout.org/docs_2006/021307G.shtml (visited Jan 12, 2008).

## II. FOURTH AMENDMENT CONSIDERATIONS

Regulation of data mining could come from many sources. At present, various scattered statutes affect the practice, albeit not to any significant extent.[48] Legislation is ultimately the best means of regulating data mining, given its complexity. But, as with other complicated areas of the law, the Constitution can provide a framework for analysis and the courts can and should prod legislators to guard against threats to constitutional values.

Data mining possibly implicates at least three constitutional provisions: the Due Process Clause's guarantee of fair process, the First Amendment's protection of speech and association, and the Fourth Amendment's prohibition on unreasonable searches.[49] The Due Process Clause might require that government make a good faith effort to secure its databases[50] and that it provide some sort of procedure for challenging erroneous inclusion on no-fly lists and other databases used in match-driven surveillance when such surveillance results in deprivations of liberty or property.[51] The First Amendment's application to data mining is more complicated. It has been argued, on the one hand, that commercial data brokers' speech rights are infringed by rules inhibiting disclosure of the information they acquire[52] and, on the other, that the First Amendment provides special protection for any personal information that evidences one's political views or asso-

[48] For a fairly up-to-date description of the statutes and a critique of them, see Christopher Slobogin, *Transaction Surveillance by the Government*, 75 Miss L J 139, 149–64 (2005). According to the GAO, none of the five best-known data mining efforts aimed at terrorists have complied with federal law requiring an assessment of their impact on privacy. Matthew B. Stannard, *U.S. Phone-Call Database Ignites Privacy Uproar; Data Mining: Commonly Used in Business to Find Patterns, It Rarely Focuses on Individuals*, San Fran Chron A1 (May 12, 2006).

[49] In this discussion, I entirely sidestep the important issue of whether the constitutional analysis changes when the government can make a plausible claim that a wartime enemy is involved. Compare John Yoo, *The Terrorist Surveillance Program and the Constitution*, 14 Geo Mason U L Rev 565, 566 (2007) (arguing that the NSA program, described below, is a constitutional exercise of the president's wartime powers), with David Cole and Mark S. Lederman, *The National Security Agency's Domestic Spying Program: Framing the Debate*, 81 Ind L J 1355, 1359 (2006) (arguing to the contrary in the course of introducing a symposium on "War, Terrorism, and Torture: Limits on Presidential Power in the 21st Century").

[50] See *Whalen v Roe*, 429 US 589, 605 (1977) (stating that "[t]he right to collect and use [medical] data for public purposes is typically accompanied by a concomitant statutory or regulatory duty to avoid unwarranted disclosures" and that such a duty "arguably has its roots in the Constitution").

[51] See Daniel J. Steinbock, *Designating the Dangerous: From Blacklists to Watchlists*, 30 Seattle U L Rev 65, 105–10 (2006) (assessing the value of adversarial proceedings to contest inclusion on a watchlist as being impractical).

[52] See Eugene Volokh, *Freedom of Speech and Information Privacy: The Troubling Implications of a Right to Stop People from Speaking about You*, 52 Stan L Rev 1049, 1051 (2000).

ciations.[53] I will not enter this debate here, because I think Fourth Amendment analysis subsumes it.[54] It is to that analysis that I now turn.

## A. Current Fourth Amendment Law and Data Mining

According to the Supreme Court, a Fourth Amendment search occurs when the government infringes "expectation[s of privacy] that society is prepared to recognize as reasonable," language that originated with Justice Harlan's concurring opinion in *Katz v United States*.[55] If a government action is a search it usually must be based on probable cause (a level of certainty akin to the civil preponderance standard), although sometimes, as is the case with a patdown of the outer clothing, all that is required is reasonable suspicion (which might be quantified at around a 30 percent level of certainty).[56] If a government action is not a search, then the Fourth Amendment is inapplicable and no justification is required.

In *United States v Miller*,[57] the Supreme Court held that a subpoena for information held by Miller's bank was not a search because Miller could not reasonably expect his bank information to remain private. Noting that *Katz* itself had stated that "[w]hat a person knowingly exposes to the public . . . is not a subject of Fourth Amendment protection,"[58] the Court reasoned that a subject

> takes the risk, in revealing his affairs to another, that the information will be conveyed by that person to the Government . . . even if the information is revealed on the assumption that it will

[53] See Daniel J. Solove, *The First Amendment as Criminal Procedure*, 82 NYU L Rev 112, 114–15 (2007).

[54] Despite his First Amendment concerns, Volokh would recognize implicit privacy-protective contracts between citizens and information-gathering entities in situations where privacy is generally expected. See Volokh, 52 Stan L Rev at 1057–60 (cited in note 52). My argument below (based in part on empirical research) is that people expect privacy with respect to the information obtained through data mining and thus implicitly contract for it. Solove argues that the First Amendment is needed to pick up the slack created by the Court's third-party records cases. See Solove, 82 NYU L Rev at 123–28 (cited in note 53). I argue that the Fourth Amendment, properly construed, should lead to reversal of those cases.

[55] 389 US 347, 362 (1967) (Harlan concurring). See also *California v Ciraolo*, 476 US 207, 211 (1986).

[56] For a defense of this quantification of Fourth Amendment standards, see Christopher Slobogin, *Let's Not Bury* Terry*: A Call for Rejuvenation of the Proportionality Principle*, 72 St John's L Rev 1053, 1082–85 (1998) (describing probable cause as a 50 percent level of certainty and reasonable suspicion as 30 percent, and arguing that such percentages should be increased but still on a sliding scale).

[57] 425 US 435 (1976).

[58] Id at 442, quoting *Katz*, 389 US at 351.

> be used only for a limited purpose and the confidence placed in the third party will not be betrayed.[59]

Three years later the Court followed *Miller* in holding, in *Smith v Maryland*,[60] that any expectation we might have that phone company logs are private is unreasonable because we know that phone companies keep records of the numbers we dial.[61]

The implications of *Miller* and *Smith* for data mining are fairly clear. These cases stand for the proposition that the government can obtain information about us from third parties without worrying about the Fourth Amendment. Since virtually all information obtained through data mining comes from third party record holders—either the government itself, commercial data brokers, or a commercial entity like a bank—its acquisition does not implicate the Fourth Amendment.

If one looks more closely at the Court's cases, there may be a few chinks in *Miller*'s armor, but they are very small. *Miller* itself relied to some extent on the fact that Miller had "voluntarily" provided his financial information to the bank,[62] leaving open the possibility that situations involving inadvertent disclosure could produce a different result. The Court has also backed off from *Miller* in two recent cases. In *Ferguson v City of Charleston*,[63] the Court found the Fourth Amendment was implicated by a hospital program that turned the results of pregnant women's drug tests over to the police without their explicit consent.[64] In distinguishing this program from other drug testing programs that it had approved, the Court noted that most patients believe diagnostic results will normally be withheld from nonmedical personnel and concluded that "[i]n none of our prior cases was there any intrusion upon that kind of expectation."[65] And in *Georgia v Randolph*,[66] the Court held that when one occupant of a residence consents to entry but another refuses, police must honor the refusal because "there is no common understanding that one co-tenant generally has a right or authority to prevail over the express wishes of another."[67] What is important about *Randolph* for present purposes is the majority's dismissal of Chief Justice Roberts's assertion in dissent, based on *Miller* and its progeny, that when "an individual shares in-

[59] *Miller*, 425 US at 443.

[60] 442 US 735 (1979).

[61] See id at 742.

[62] See 425 US at 442.

[63] 532 US 67 (2001).

[64] See id at 86.

[65] Id at 78.

[66] 547 US 103 (2006).

[67] Id at 114.

formation, papers, or places with another, he assumes the risk that the other person will in turn share access to that information or those papers or places with the government."[68] Far from agreeing with this statement, the majority chastised the Chief Justice for his "easy assumption that privacy shared with another individual is privacy waived for all purposes including warrantless searches by the police."[69]

*Ferguson* and *Randolph* signal that the Court is willing to consider at least minor exceptions to *Miller*'s dictate that the government does not effect a constitutionally regulated search when it accesses information the subject shared with a third party. If information is disclosed inadvertently or is particularly private (as with medical data), or if we specifically refuse to disclose it to the government, perhaps a reasonable expectation of privacy attaches. Should these exceptions be strengthened? Should they be broadened? If so, what form might they take?

## B. The Case for a Hierarchy of Records

A few lower courts have been willing to resist the broad language in *Miller* and grant Fourth Amendment protection (or protection under the analogous state constitutional provision) to some types of records. Stephen Henderson's survey of the case law identifies more than a dozen factors the courts have considered,[70] principal among them: (1) the extent to which disclosure of the information is necessary to function in society (with one court, for instance, distinguishing between phone numbers maintained by the phone company and information given to a locksmith[71]); (2) the degree to which the information is personal (with one court, for example, evidencing deep disagreement over whether power consumption records are personal[72]); and (3) the amount of information obtained (with some courts distinguishing between multiple records and a record of one transaction[73]).

[68] Id at 128 (Roberts dissenting) (emphasis omitted).

[69] Id at 115 n 4 (majority).

[70] See Stephen E. Henderson, *Beyond the (Current) Fourth Amendment: Protecting Third Party Information, Third Parties, and the Rest of Us Too*, 34 Pepperdine L Rev 975, 985–1018 (2007).

[71] See *People v Abbott*, 208 Cal Rptr 738, 741 (1984).

[72] See generally *In re Maxfield*, 945 P2d 196 (Wash 1997) (four justices holding that electricity records are protected by the state constitution, four justices disagreeing with that holding, and one justice agreeing with the dissent's constitutional analysis but finding a statutory basis for siding with the first group of justices).

[73] See, for example, *Commonwealth v Duncan*, 817 A2d 455, 463 (Pa 2003) ("[A] particular ATM card number is obviously different in kind from the disclosure of substantive bank records . . . . A person's name and address do not, by themselves, reveal anything concerning his personal affairs, opinions, habits or associations. Such innocuous information does not provide or complete a virtual current biography.") (quotation marks omitted). See also *People v Sporleder*, 666

On an abstract level, these are sensible criteria for evaluating Fourth Amendment privacy. But applying them in a judicious manner is another matter. Putting aside the number of variables involved (Henderson himself insists that nine of the thirteen factors he discusses are relevant to Fourth Amendment analysis[74]), the indeterminacy of the three just described should be apparent. The first, which looks at how important a given service is to modern life, triggers real quandaries: using the case noted above as an example, why are locksmiths any less essential to functioning in today's world than phones, given the need for security and the frequency with which people are locked out of home, office, or car? Also daunting is the task of calibrating, in the abstract, the extent to which particular information is "personal." In *Kyllo v United States*[75] (which held that using a thermal imager to measure heat differentials inside a house *is* a search), the Supreme Court explicitly avoided this type of question on the ground it could not be answered coherently,[76] a difficulty brought home by the fact that in the power consumption case noted above, four judges vigorously dissented from the conclusion that electricity usage data is personal.[77] An equally perplexing question, raised by the third factor, is the number of transactions a record must contain before its seizure by the government implicates the Fourth Amendment.

Admittedly, any attempt to assess privacy in a meaningful fashion will run into these types of definitional conundrums (as the proposal I make below with respect to data mining demonstrates). A more fundamental problem is that privacy may not be measurable in the predominately normative terms these courts are applying. Robert Post, for instance, has concluded that the scope of privacy, when conceptualized as a form of dignity, is entirely dependent on everyday social practices, not foundational theory.[78] In an article about expectations of privacy in the tort context, Lior Strahilevitz agrees that, given the highly contestable nature of the concept, any effort to arrive at an objectively neutral take on privacy is "doomed."[79] Instead Strahilevitz

P2d 135, 142 (Colo 1983) (fearing that allowing the government to acquire all of an individual's telephone records would give it the capacity to create a "virtual mosaic of a person's life").

[74] See Henderson, 34 Pepperdine L Rev at 988–89 (cited in note 70).

[75] 533 US 27 (2001).

[76] Id at 37–38 ("The Fourth Amendment's protection of the home has never been tied to measurement of the quality or quantity of information obtained. . . . In the home, our cases show, *all* details are intimate details, because the entire area is held safe from prying government eyes.").

[77] See *In re Maxfield*, 945 P2d at 207 (Guy dissenting) ("Electrical consumption information, unlike telephone or bank records or garbage, does not reveal discrete information about a customer's activities.").

[78] See Robert C. Post, *Three Concepts of Privacy*, 89 Georgetown L J 2087, 2092, 2094 (2001).

[79] See Lior Jacob Strahilevitz, *A Social Networks Theory of Privacy*, 72 U Chi L Rev 919, 932 (2005).

argues that, at least for purposes of defining privacy torts, the law's approach to privacy should derive primarily from empirical investigation of social norms.[80]

The type of empirical work Strahilevitz has in mind for this purpose focuses on how we "network" socially. His reading of the social network literature indicates that unless it is "likely to be regarded as highly interesting, novel, revealing, or entertaining," information that we reveal about ourselves rarely gets past "two degrees of separation"—that is, beyond a friend of a friend.[81] This limited range of disclosure is partly the result of routine inefficiencies in communication. But it would exist even if the internet were to radically reduce these inefficiencies, because people simply don't care about the private affairs of strangers unless the events are dramatic or are somehow economically useful.

The implications of social network theory for data mining are straightforward. Unless it is part of a public record designed for consumption by everyone or describes an activity observed by strangers, the transactional information government seeks through data mining is rarely known outside our families, much less outside our social network (aside from the third-party institutions to which we provide it). Expectations that such information will remain "private" are reasonable from the social network perspective.

Independent empirical support for an enlarged view of privacy in individual records is provided by a study I conducted of a group of jury pool members ($N = 76$). Following a methodology I have used in the past to evaluate other types of policing techniques,[82] the participants in this study were asked to rate on a scale of 1 to 100 the relative intrusiveness of twenty-five scenarios involving investigative actions by law enforcement. Most of these scenarios involved some type of government effort to obtain records, or what I have called in other writing "transaction surveillance."[83] Most of these transaction surveillance scenarios involved target-driven investigations, but five described event-driven data mining. Additionally, to establish a baseline, the survey included five scenarios describing investigative techniques that do *not* involve transaction surveillance and that the Supreme Court has held *do* implicate the Fourth Amendment: searches of bed-

[80] See id at 931–35.

[81] Id at 967.

[82] See Christopher Slobogin, *Public Privacy: Camera Surveillance of Public Places and the Right to Anonymity*, 72 Miss L J 213, 275–76 (2002); Christopher Slobogin and Joseph Schumacher, *Reasonable Expectations of Privacy and Autonomy in Fourth Amendment Cases: An Empirical Look at "Understandings Recognized and Permitted by Society*,*"* 42 Duke L J 727, 735–37 (1993).

[83] See Slobogin, 75 Miss L J at 140 (cited in note 48).

rooms (which require probable cause and, in non-exigent circumstances, a warrant[84]); searches of cars (which require probable cause[85]); patdowns or frisks (which require reasonable suspicion[86]); a brief stop for purposes of obtaining identification (which may under some circumstances require reasonable suspicion, depending on its length[87]); and a stop at a roadblock (which is permitted if the government can demonstrate the roadblock addresses a significant travel-related problem such as illegal immigration or drunk driving[88]).

The results of the survey, showing the average mean intrusiveness rating along with a confidence interval indicating the significance of the finding, are found in the Table. As an initial matter, the most important result of this study is that the participants considered many types of transaction surveillance to be more intrusive than patdowns (which require reasonable suspicion) and searches of cars (which require probable cause). Consistent with the lower court cases described above, the participants distinguished between the types of information obtained (for example, credit card records, $M = 75.3$, as opposed to electricity consumption records, $M = 57.4$), and surveillance that is isolated as opposed to aggregating (compare Scenario 14, obtaining a record of a specific phone call, $M = 59.8$, with Scenario 17, obtaining a person's composite phone records, $M = 74.1$). Participants also distinguished between event-driven data mining and target-driven surveillance of the same types of information (compare Scenarios 2, 3, 5, 10, and 13 to Scenarios 20–24). Such distinctions notwithstanding, all these government actions, as well as searches of corporate and public records, were perceived as more intrusive than a roadblock (see Scenario 1), which is governed by the Fourth Amendment, and many were viewed as more intrusive than a stop and a patdown.

[84] See *Chimel v California*, 395 US 752, 763 (1969).

[85] See *United States v Ross*, 456 US 798, 807–08 (1982).

[86] See *Terry v Ohio*, 392 US 1, 30 (1968).

[87] Compare *Brown v Texas*, 443 US 47, 50 (1979) ("When the officers detained appellant for the purpose of requiring him to identify himself, they performed a seizure subject to the requirements of the Fourth Amendment."), with *INS v Delgado*, 466 US 210, 216–17 (1984) (holding that police questioning is not a seizure unless the person reasonably believes he is not free to leave).

[88] See *City of Indianapolis v Edmond*, 531 US 32, 44, 47 (2000) (declining "to approve a program whose primary purpose is ultimately indistinguishable from the general interest in crime control," but carefully indicating that this ruling did not alter the constitutionality of "sobriety and border checkpoints").

## TABLE

Mean Intrusiveness Ratings of Twenty-five Scenarios

| Scenario | | Mean | Confidence Intervals |
|---|---|---|---|
| 1 | *Roadblock* | 30.2 | ±7.5 |
| 2 | Airplane passenger lists (event-driven) | 32.4 | 8 |
| 3 | Store patron lists (event-driven) | 34.1 | 7.5 |
| 4 | Criminal/traffic records | 36.2 | 7 |
| 5 | Anonymous phone, credit card, and travel records (event-driven) | 38.5 | 7 |
| 6 | Corporate records | 40.6 | 7 |
| 7 | Real estate records | 45.5 | 8 |
| 8 | *ID check and questioning during brief stop* | 49.1 | 8 |
| 9 | Club membership records | 49.5 | 8 |
| 10 | Phone records (event-driven) | 50.0 | 8 |
| 11 | Electricity records | 57.5 | 8 |
| 12 | High school records | 58.3 | 9 |
| 13 | Phone, credit card, and travel records (event-driven) | 59.7 | 8 |
| 14 | Record of specific phone call | 59.8 | 7.5 |
| 15 | List of food purchases | 65.3 | 7.5 |
| 16 | *Patdown* | 71.5 | 7.5 |
| 17 | Phone records | 74.1 | 7.5 |
| 18 | Websites visited | 74.4 | 8 |
| 19 | *Search of car* | 74.6 | 7 |
| 20 | Credit card records | 75.3 | 7.5 |
| 21 | Email addresses sent to and received from | 77.1 | 8 |
| 22 | Pharmacy records | 78.0 | 7.5 |
| 23 | Use of snoopware to target subject | 79.0 | 8 |
| 24 | Bank records | 80.3 | 7.5 |
| 25 | *Bedroom search* | 81.2 | 6.5 |

Note: Scenarios not involving transaction surveillance appear in italics. These findings are based on a survey administered to seventy-six members of the Gainesville, Florida jury pool, randomly selected from a list composed of all residents who have a driver's license or identification card. See Fla Stat Ann § 40.011 (West 2007).

These empirical observations suggest that, contrary to the Supreme Court's insinuation in cases like *Miller* and *Smith*, transferring information to third parties or allowing third parties to accumulate it does not, by itself, lessen the intrusiveness of government efforts to obtain it. To the members of society queried in this survey, the important variable appears to be the nature of the record, not who or what institution possesses it. As *Katz*'s language appears to mandate, Fourth Amendment jurisprudence ought to recognize society's apparent expectation, whether measured directly or through social network research, that this type of information is private.

At the same time, the empirical observations from my study, and to a lesser extent the logic of social network theory, indicate that society does not view all transaction surveillance as equally intrusive. More specifically, the findings summarized in the Table above suggest three broad categories of intrusiveness, divided by Scenario 8 (a police stop demanding identification, which verges on being a Fourth Amendment seizure) and Scenario 16 (a patdown, which requires reasonable suspicion). Into the first category (Scenarios 2–7) fall government acquisition of corporate records, public records, and many types of data mining. These types of transaction surveillance are all ranked lower than the street identification scenario, although still above a roadblock. At the other end of the spectrum (Scenarios 17–25) are government efforts to obtain many types of information maintained by private entities, including records of phone and email correspondence, websites visited, credit card purchases, and pharmacy and bank records. These types of transaction surveillance are all ranked as more intrusive than a patdown and about as intrusive as either a car search (Scenario 19) or a search of a bedroom (Scenario 25), both of which require probable cause. Between the identification check and patdown scenarios are several types of transaction surveillance: (1) acquisition of what might be called "quasi-private" records from clubs, electric companies, high schools, and grocery stores (Scenarios 9, 11, 12, and 15); (2) private records depicting a single event (Scenario 14); and (3) data mining of private records (Scenarios 10 and 13).

## III. Application to Data Mining

Elsewhere I have addressed methodological and relevance issues associated with the type of research summarized in the Table.[89] For now, let us assume that the hierarchy indicated in this research roughly cap-

[89] See Slobogin, 72 Miss L J at 280–85 (cited in note 82) (addressing the relevance of survey findings to Fourth Amendment analysis); Slobogin and Schumacher, 42 Duke L J at 743–51 (cited in note 82) (addressing internal and external validity issues).

tures expectations society has about the relative invasiveness of the indicated police actions. The most significant implication of that assumption is that *Miller* and *Smith* should be overturned, and that transaction surveillance should be subject to Fourth Amendment regulation. A second implication is that different types of transaction surveillance should be subject to different types of constitutional regulation.

The latter proposition is based on what I have called the *proportionality principle*, which states that the level of justification required for a search or seizure should be roughly proportionate to its intrusiveness.[90] This simple idea is not, of course, my invention but rather was endorsed by the Supreme Court as far back as the 1960s, in *Camara v Municipal Court of City and County of San Francisco*[91] and *Terry v Ohio*.[92] In the latter case, the Court stated (quoting *Camara* in part) that "there is 'no ready test for determining reasonableness other than by balancing the need to search [or seize] against the invasion which the search [or seizure] entails.'"[93]

Application of this proportionality principle to *target-driven* data mining might lead to the creation of three tiers of record searches, following the three levels of records described above and set out in the Table. Into the bottom tier, which would merely require the government to show it has a legitimate interest in the records, would fall target-driven efforts to obtain corporate records and most public records (for example, criminal and real estate records). The other end of the spectrum, which would require probable cause, would include target-driven attempts to obtain records containing the most personal information (for example, bank records and phone and ISP logs). The middle tier, which would require reasonable suspicion, would involve target-driven efforts to obtain records that are quasi-private because they contain information considered less personal (for example, power consumption records, high school records).

Regulation of *match-driven* data mining, in contrast, would depend not on the nature of the records searched but on the nature of the action the government contemplates taking when a match occurs. If the consequence of being on a no-fly list is arrest, a person should not appear on the list unless probable cause exists to believe the indi-

[90] See Slobogin, 72 St John's L Rev at 1054 (cited in note 56).

[91] 387 US 523, 539 (1967) (recognizing that using a reasonableness approach to the Fourth Amendment "neither endangers time-honored doctrines applicable to criminal investigations nor makes a nullity of the probable cause requirement . . . [but] merely gives full recognition to the competing public and private interests here at stake").

[92] 392 US 1, 21 (1968).

[93] Id at 21 (alterations in original), quoting *Camara*, 387 US at 536–37.

vidual is a criminal/terrorist. If the consequence is instead merely a prohibition on boarding, reasonable suspicion might be sufficient.

The most interesting application of the proportionality principle occurs in connection with *event-driven* data mining. Some types of event-driven data mining seem relatively unintrusive. Recall the rape investigation example involving accessing residential records from two different cities. A similar example might involve tracking down people who have bought a type of shoe or sweater that has been linked to the scene of a homicide. In these cases, the information sought (residential information and purchases) comes from public or quasi-private records. Furthermore, in contrast to many types of transaction surveillance, the government acquires only one or two bits of information about the persons so identified (for example, that they lived in a certain city during a certain period or bought a particular type of shoe). Finally, the information has not been obtained in single-minded pursuit of a particular person but rather in an effort to determine whom to pursue; any given individual's record is merely one of hundreds or thousands, and will be discarded or at least ignored if it does not prove of interest to investigators. For all these reasons, this investigative technique appears to be a far cry from the creation of personality mosaics through data aggregation, the scenario that has worried those who criticize large-scale transaction surveillance.

Consistent with this intuition, the survey participants rated these types of event-driven data mining, depicted as Scenarios 2 and 3 in the Table, as less intrusive than an ID check. Both of these scenarios involved quasi-private records (airline passenger lists and store patron lists) that recount actions observed by multiple strangers outside one's social network. Proportionality reasoning might permit this type of event-driven data mining whenever the government can demonstrate a legitimate need for the information.

Other event-driven data mining might call for a different approach, however, particularly if it focuses on records perceived to contain highly private information. For instance, media reports indicate that the National Security Agency has accumulated the phone records (revealing the numbers dialed) of millions of Americans so that it can conduct "link analysis," another term for event-driven data mining.[94]

[94] See, for example, Karen Tumulty, *Inside Bush's Secret Spy Net; Your Phone Records Have Been Enlisted in the War on Terrorism. Should That Make You Worry More or Less?*, Time 32, 35 (May 22, 2006) ("The idea is to sift through all that data, using a process called link analysis, searching for patterns—a burst of calls from pay phones in Detroit to cell phones in Pakistan, for instance."); Leslie Cauley, *NSA Has Massive Database of Americans' Phone Calls; 3 Telecoms Help Government Collect Billions of Domestic Records*, USA Today 1A (May 11, 2006) (report-

The NSA, *Time* has alleged, is trying to "whittle down the hundreds of millions of phone numbers harvested to hundreds of thousands that fit certain profiles it finds interesting; those in turn are cross-checked with other intelligence databases to find, perhaps, a few thousand that warrant more investigation."[95] The survey participants were much more leery of this type of data mining, ranking it as more intrusive than an ID check, whether aimed at multiple record sets (see Scenario 13, involving data mining of phone, credit card, and travel records) or only one (see Scenario 10, involving data mining of phone records).

Assuming this finding accurately represents societal views, proportionality reasoning would suggest that event-driven data mining of private records should occur only if reasonable suspicion exists. Note, however, that given the large-scale nature of this type of event-driven mining, "individualized" reasonable suspicion would be impossible to generate. As I have argued elsewhere,[96] in group search situations of this sort, permitting the government to demonstrate "generalized" or group-wide suspicion might make sense. That would mean the government's profile should achieve roughly a 30 percent hit rate—that is, roughly a one out of three chance that data mining of this sort will discover useful evidence.[97]

Proponents of the NSA program would likely resist this type of restriction by claiming that the program is necessary to stem the threat posed by terrorism.[98] It is certainly reasonable to relax the showing required under proportionality analysis when the government can demonstrate that data mining is necessary to detect a significant, imminent threat.[99] Outside of the emergency context, however,

ing that the NSA used telephone records from AT&T, Verizon, and BellSouth while attempting "to create a database of every call ever made" in the US).

[95] Tumulty, *Inside Bush's Secret Spy Net*, Time at 35 (cited in note 94).

[96] See Slobogin, 72 St John's L Rev at 1085–91 (cited in note 56).

[97] How is the government to meet the burden demanded by this proportionality analysis? Sometimes the government's profile may satisfy the requisite certainty level on its face, as in an investigation of purchasing fraud where the profile singles out those individuals who have bought items they are clearly not authorized to buy. Other types of profiles might be tested through hypothetical computer runs, something the government is apparently doing now. See DARPA, *Report* at 17 (cited in note 28) (describing use of "synthetic data" to test the efficacy of data mining processes). As a last resort, an actual data mining program could be carried out on a small sample under secure conditions to determine its efficacy. Finally, if the government can provide a convincing explanation as to why relevant data cannot be obtained, while at the same time suggesting why the relevant hit rate can be met, it might be allowed to proceed.

[98] See, for example, Yoo, 14 Geo Mason L Rev at 577 (cited in note 49) ("Data mining is the best hope for an innovative counterterrorism strategy to detect and prevent future al Qaeda attacks.").

[99] The analogue in traditional Fourth Amendment jurisprudence might be the hot pursuit doctrine, where the courts have struggled to differentiate between hot and lukewarm pursuit, but have refused to adopt exceptions based solely on the seriousness of the crime. See Charles H. Whitebread and Christopher Slobogin, *Criminal Procedure: An Analysis of Cases and Concepts* § 8.03 at 228–32 (Foundation 5th ed 2008). It is worth noting in this regard that Germany, which

proportionality reasoning would more strictly regulate data mining of private records than does current law.

While it thus imposes greater restrictions on data mining than presently apply, the upshot of proportionality reasoning is that event-driven data mining would not be as stringently monitored as target-driven data mining. Event-driven data mining of private records would require only reasonable suspicion and event-driven data mining of quasi-private and public records could be carried out on a relevance showing. But it should also be noted that even the latter requirement would be difficult to meet in many event-driven data mining contexts. For instance, as noted earlier, given the small number of terrorists in the United States, even application of a highly accurate profile is likely to produce a very high ratio of false positives (nonterrorists identified as terrorists) to true positives (actual terrorists) if millions of records have to be sifted to find them.[100] Barring an emergency, then, many of the government's antiterrorism data mining efforts aimed at domestic records might fail to meet the relevant threshold.

Responding to this type of concern, some have suggested that the government could keep the results of its initial data mining passes anonymous—using pseudonyms or nonhuman (computerized) techniques—until it produces a group for which it has the requisite cause.[101] Although this latter type of multistage analysis—sometimes called "selective revelation"—is technologically feasible (and was viewed as relatively unintrusive in the survey, as indicated by the ranking of Scenario 4 involving anonymous acquisition of personal information), it is largely untested in most law enforcement contexts.[102] Furthermore, under a proportionality regime, stringent auditing proce-

has had considerable experience with dragnet information gathering, much of it negative, permits event-driven surveillance *only* in response to a specifically articulated danger. See Francesca Bignami, *European versus American Liberty: A Comparative Privacy Analysis of Antiterrorism Data Mining*, 48 BC L Rev 609, 654–55 (2007) (describing a German court decision finding unconstitutional a post-9/11 data-mining program aimed at identifying people with certain characteristics—male, age 18–40, student or former student, Islamic faith, citizenship or birthplace in a country with a predominantly Islamic population—because there were no facts demonstrating "an imminent and specific endangerment").

[100] See Bruce Schneier, *Beyond Fear: Thinking Sensibly about Security in an Uncertain World* 253–54 (Copernicus 2003) (explaining why it is very difficult to uncover terrorist plots through data mining).

[101] For a description of how selective revelation might work, see K.A. Taipale, *Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 Colum Sci & Tech L Rev 2, 79–80 (2003).

[102] According to one source, the technology has yet to reach the stage at which anonymity can be preserved. See Palo Alto Research Company, *Privacy Appliance*, online at http://www.parc.com/research/projects/privacyappliance (visited Jan 12, 2008) (describing yet-to-be-developed protocols that ensure "inference control," that is, protection against the identification of an individual through combining different pieces of information).

dures would need to be in place to ensure the government didn't cheat during this process by prematurely linking the files with names or hacking into the computerized investigation.

## CONCLUSION

Space limitations have necessitated an abbreviated account of how data mining might be regulated under the Fourth Amendment. In my recent book, I analyze these issues in more detail.[103] The most important conclusion is that the Supreme Court's current hands-off approach to record searches cannot justifiably be applied to data mining if societal views about privacy expectations are taken seriously. At the same time, specific justification rules should differ depending on whether the data mining is target-, match-, or event-driven, and the types of records the data mining accesses.

[103] See generally Christopher Slobogin, *Privacy at Risk: The New Government Surveillance and the Fourth Amendment* (Chicago 2007).