Testimony of
Kim Taipale, Executive Director
Center for Advanced Studies in Science and Technology Policy
www.advancedstudies.com

Before the
United States Senate Committee on the Judiciary
January 10, 2007

The Privacy Implications of Government Data Mining Programs

Mr. Chairman Leahy, Ranking Member Specter, and Members of the Committee: Thank you for the opportunity to testify today on the Privacy Implications of Government Data Mining Programs.

Official U.S. Government policy calls for the research, development, and implementation of advanced information technologies for analyzing data, including data mining, in the effort to help protect national and domestic security. Civil libertarians and libertarians alike have decried and opposed these efforts as an unprecedented invasion of privacy and a fundamental threat to our freedoms.

While it is true that data mining technologies raise significant policy and privacy issues, the public debate on both sides suffers from a lack of clarity. Technical and policy misunderstandings have lead to the presentation of a false dichotomy—a choice between security or privacy.

In particular, many critics have asserted that data mining is an ineffectual tool for counterterrorism not likely to uncover any terrorist plots and that the number of false positives will waste resources and will impact too many innocent people. Unfortunately, many of these critics fundamentally misunderstand data mining and how it can be used in counterterrorism applications. My testimony today is intended to address some of these misunderstandings.

Introduction.

My name is Kim Taipale. I am the founder and executive director of the Center for Advanced Studies in Science and Technology Policy, an independent, non-partisan research organization focused on information, technology, and national security issues. I am the author of numerous law review articles, academic papers, and book chapters on issues involving technology, national security, and privacy, including several that address data mining in particular. [1]

[1] *See, e.g., Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data*, 5 COLUMBIA SCI. & TECH. L. REV. 2 (Dec. 2003) [hereinafter "*Connecting the Dots*"]; *Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy, and the Lessons of King Ludd*, 7 YALE J.

By way of further identification, I am also a senior fellow at the World Policy Institute at the New School and an adjunct professor of law at New York Law School. I also serve on the Markle Task Force on National Security in the Information Age, the Science and Engineering for National Security Advisory Board at the Heritage Foundation, and the Steering Committee of the American Law Institute project on government access to personal data. Of course, the opinions expressed here today are my own and do not represent the views of any of these organizations.

My testimony is founded on several axiomatic beliefs:

- First, security and privacy are not dichotomous rivals to be "balanced" but rather vital interests to be reconciled (that is, they are dual obligations of a liberal republic, each to be maximized within the constraints of the other—there is no fulcrum point at which the "right" amount of either security or privacy can be achieved);

- Second, while technology development is not deterministic, it is inevitable (that is, we face a certain future of more data availability and more sophisticated analytic tools);

- Third, political strategies premised on simply outlawing particular technologies or techniques are ultimately futile strategies that will result in little security and brittle privacy protections (that is, simply seeking to deny security services widely available tools is not feasible nor good security policy, and simply applying rigid prohibitions that may not survive if there were to be another catastrophic event is not good privacy policy); and

- Fourth, and most importantly, while data mining (or any other) technology cannot provide security on its own, it can, if properly employed, improve intelligence gain and help better allocate scarce security resources, and, if properly designed, do so while still protecting privacy.

I should note that my testimony today is not intended either as critique or endorsement of any particular government data mining program or application, nor is it intended to make any specific policy or legal recommendation for any particular implementation. Rather, it seeks simply to elucidate certain issues at the intersection of technology and policy that

---

L. & TECH. 123 (Mar. 2004) [hereinafter "*Frankenstein*"]; *The Trusted System Problem: Security Envelopes, Statistical Threat Analysis, and the Presumption of Innocence*, IEEE INTELLIGENT SYSTEMS, V.20 No.5, (Sep./Oct. 2005); *Designing Technical Systems to Support Policy: Enterprise Architecture, Policy Appliances, and Civil Liberties*, in EMERGENT INFORMATION TECHNOLOGIES AND ENABLING POLICIES FOR COUNTER TERRORISM (Robert Popp and John Yen, eds., Wiley-IEEE, Jun. 2006); *Whispering Wires and Warrantless Wiretaps: Data Mining and Foreign Intelligence Surveillance*, NYU REV. L. & SECURITY, NO. VII SUPL. (Spring 2006); *Why Can't We All Get Along? How Technology, Security and Privacy Can Co-exist in a Digital World*, in CYBERCRIME AND DIGITAL LAW ENFORCEMENT (Ex Machina: Law, Technology, and Society Book Series) (Jack Balkin, *et al., eds*., NYU Press, forthcoming Spring 2007); and *The Ear of Dionysus: Rethinking Foreign Intelligence Surveillance,* 9 YALE J. L. & TECH. (forthcoming Spring 2007).

are critical, in my view, to a reasoned debate and democratic resolution of these issues and that are widely misunderstood or misrepresented.

Nevertheless, before I begin, I proffer certain overriding policy principles that I believe should govern any development and implementation of these technologies in order to help reconcile security and privacy needs. These principles are:

- First, that these technologies only be used as investigative, not evidentiary, tools (that is, used only as a predicate for further screening or investigation, but not for proof of guilt or otherwise to invoke significant adverse consequences automatically) and only for investigations or analysis of activities about which there is a political consensus that aggressive preventative strategies are appropriate or required (for example, the preemption of terrorist attacks or other threats to national security).

- Second, that specific implementations be subject to strict congressional oversight and review, be subject to appropriate administrative procedures within executive agencies where they are to be employed, and be subject to appropriate judicial review in accordance with existing due process doctrines.

- And, third, that specific technical features be developed and built into systems employing data mining technologies (including rule-based processing, selective revelation, and secure credentialing and tamper-proof audit functions) that, together with complimentary policy implementations (and appropriate systems architecture), can enable familiar, existing privacy protecting oversight and control mechanisms, procedures and doctrines (or their analogues) to function.

My testimony today is in four parts: the first deals with definitions; the second with the need to employ predictive tools in counterterrorism applications; the third answers in part the popular arguments against data mining; and the fourth offers a view in which technology and policy can be designed to conciliate privacy and security needs.

I. Parsing definitions: data mining and pornography.

In a recent policy brief [2] (released by way of a press release headlined: *Data Mining Doesn't Catch Terrorists: New Cato Study Argues it Threatens Liberty*), [3] the authors argue that "data mining" is a "fairly loaded term that means different things to different people" and that "discussions of data mining have probably been hampered by lack of clarity about its meaning," going on to postulate that "[i]ndeed, collective failure to get to the root of the term 'data mining' may have preserved disagreements among people who may be in substantial agreement." The authors then proceed to define data mining extremely narrowly by overdrawing a popular but generally false dichotomy between

---

[2] Jeff Jonas & Jim Harper, *Effective Counterterrorism and the Limited Role of Predictive Data Mining,* Cato Institute (December 11, 2006) at p. 5.

[3] Press Release, *Data Mining Doesn't Catch Terrorists: New Cato Study Argues it Threatens Liberty* (Dec. 11, 2006) available at http://www.cato.org/new/pressrelease.php?id=73

subject-based and pattern-based analysis [4] that allows them to conclude "that [predictive, pattern-based] data mining is costly, ineffective, and a violation of fundamental liberty" [5] while still concluding that other "data analysis"—including "bringing together more information from more diverse sources and correlating the data … to create new knowledge"— is not.[6]

In another recent paper, [7] the former director and deputy director of DARPA's Information Awareness Office describe "a vision for countering terrorism through information and privacy-protection technologies [that] was initially imagined as part of … the Total Information Awareness (TIA) program." "[W]e believe two basic types of queries are necessary: subject-based queries … and pattern-based queries … . Pattern-based queries let analysts take a predictive model and create specific patterns that correspond to anticipated terrorist plots." However, "[w]e call our technique for counterterrorism activity data analysis, not data mining," they write.

It is thus sometimes hard to find the disagreement among the opponents and proponents as data mining seems somewhat like pornography—everyone can be against it (or not engaged in it), as long as they get to define it. [8] Since further parsing of definitions is unlikely to advance the debate let us simply assume instead that there is some form of data analysis based on using patterns and predication that raises novel and challenging policy and privacy issues. The policy concern, it seems to me, is how those issues might be managed to improve security while still protecting privacy.

---

[4] Sophisticated data mining applications use both known (observed) and unknown (queried) variables and use both specific facts (i.e., relating to subjects or entities) and general knowledge (i.e., patterns) to draw inferences. Thus, subject-based and pattern-based are just two ends of spectrum.

[5] Press Release, *supra* note 3.

[6] Jonas & Harper, *supra* note 2 at 4-6. Compare, however, one of the author's previous conclusion that "[w]hen a government is faced with an overwhelming number of predicates (i.e., subjects of investigative interest), data mining can be quite useful for triaging (prioritizing) which subjects should be pursued first. One example: the hundreds of thousands of people currently in the United States with expired visas. The student studying virology from Saudi Arabia holding an expired visa might be more interesting than the holder of an expired work visa from Japan writing game software." jeffjonas. typepad.com (Mar. 12, 2006). Thus highlighting again that even predictive pattern-based data mining can be both "ineffective" and "quite useful" for counterterrorism applications depending seemingly only on the felicitousness of the definition applied.

[7] Robert Popp & John Poindexter, *Countering Terrorism through Information and Privacy Protection Technologies*, IEEE Security & Privacy, Vol.4, No.6 (Nov./Dec. 2006) pp. 18-27.

[8] *Cf.,* Jacobellis v. Ohio, 378 U.S. 184 (1964) (Stewart, J., concurring) in which Justice Potter Stewart famously declared that although he could not define hard-core pornography, "he knows it when he sees it." Note that definitions of data mining in public policy range from the seemingly limitless, for example, the DoD Technology and Privacy Advisory Committee (TAPAC) Report defines "data mining" to mean "searches of one or more electronic databases of information concerning U.S. person by or on behalf of an agency or employee of the government," to the non-existent, for example, The Data-Mining Moratorium Act of 2003, S. 188, 108th Cong. (2003), which does not even define "data-mining."

II. The Need for Predictive Tools.

Security and privacy today both function within a changing context. The potential to initiate catastrophic outcomes that can actually threaten national security is devolving from other nation states (the traditional target of national security power) to organized but stateless groups (the traditional target of law enforcement power) blurring the previously clear demarcation between reactive law enforcement policies and preemptive national security strategies. Thus, there has emerged a political consensus—at least with regard to certain threats—to take a preemptive rather than reactive approach. "Terrorism [simply] cannot be treated as a reactive law enforcement issue, in which we wait until after the bad guys pull the trigger before we stop them." [9] The policy debate is no longer about preemption itself—even the most strident civil libertarians concede the need to identify and stop terrorists before they act—but instead revolves around what methods are to be properly employed in this endeavor. [10]

However, preemption of attacks that can occur at any place and any time requires information useful to anticipate and counter future events—that is, it requires actionable intelligence based on predictions of future behavior. Unfortunately, except in the case of the particularly clairvoyant, prediction of future behavior can only be assessed by examining and analyzing indicia derived from evidence of current or past behavior or from associations. Fortunately, terrorist attacks at scales that can actually endanger national security generally still require some form of organization. [11] Thus, effective counterterrorism strategies in part require analysis to uncover evidence of organization, relationships, or other relevant indicia indicative or predictive of potential threats—that is, actionable intelligence—so that additional law enforcement or security resources can then be allocated to such threats preemptively to prevent attacks.

Thus, the application of data mining technologies in this context is merely the computational automation of necessary and traditional intelligence and investigative techniques, in which, for example, investigators may use pattern recognition strategies to develop modus operandi ("MO") or behavioral profiles, which in turn may lead either to specific suspects (profiling as identifying pattern) or to attack-prevention strategies (profiling as predictor of future attacks, resulting, for example, in focusing additional security resources on particular places, likely targets, or potential perpetrators—that is, to allocate security resources to counter perceived threats). Such intelligence-based policing or resource allocation is a routine investigative and risk-management practice.

---

[9] Editorial, *The Limits of Hindsight*, WALL ST. J. (Jul. 28, 2003) at A10. *See also* U.S. Department of Justice, *Fact Sheet: Shifting from Prosecution to Prevention*, *Redesigning the Justice Department to Prevent Future Acts of Terrorism* (May 29, 2002).

[10] *See generally* Alan Dershowitz, PREEMPTION: A KNIFE THAT CUTS BOTH WAYS (W.W. Norton & Company 2006).

[11] For example, highly coordinated conventional attacks, multidimensional assaults calculated to magnify the disruption, or the use of chemical, biological, or nuclear (CBN) weapons, are all still likely require some coordination of actions or resources.

The application of data mining technologies in the context of counterterrorism is intended to automate certain analytic tasks to allow for better and more timely analysis of existing data in order to help prevent terrorist acts by identifying and cataloging various threads and pieces of information that may already exist but remain unnoticed using traditional manual means of investigation. [12] Further, it attempts to develop predictive models based on known or unknown patterns to identify additional people, objects, or actions that are deserving of further resource commitment or attention. Data mining is simply a productivity tool that when properly employed can increase human analytic capacity and make better use of limited security resources.

(Policy issues relating specifically to the use of data mining tools for analysis must be distinguished from issues relating more generally to data collection, aggregation, access, or fusion, each of which has its own privacy concerns unrelated to data mining itself and which may or may not be implicated by the use of data mining depending on its particular application. [13] The relationship between scope of access, sensitivity of data, and method of query is a complex calculus, a detailed discussion of which is beyond the scope of my formal testimony today. [14] Also to be distinguished for policy purposes, is decision-making, the process of determining thresholds and consequences of a match. [15])

III. Answering the "case" against data mining.

The popular arguments made against employing data mining technologies in counterterrorism applications generally take two forms: the pseudo-technical argument,

---

[12] Data mining is intended to turn low-level data, usually too voluminous to understand, into higher forms (information or knowledge) that might be more compact (for example, a summary), more abstract (for example, a descriptive model), or more useful (for example, a predictive model). *See also* Jensen, *infra* note 28, at slide 22 ("A key problem [for using data mining for counter-terrorism] is to identify high-level things – organizations and activities – based on low-level data – people, places, things and events."). Data mining can allow human analysts to focus on higher-level analytic tasks by identifying obscure relationships and connections among low-level data.

[13] The question of what data should be available for analysis, under what procedure, and by what agency is a related but genuinely separate policy issue from that presented by whether automated analytic tools such as data mining should be used. For a discussion of issues relating to data access and sharing, see the Second Report of the Markle Taskforce on National Security in the Information Age, *Creating a Trusted Information Sharing Network for Homeland Security* (2003). For a discussion of government access to information from the private sector and a proposed data-classification structure providing for different levels of process based on data sensitivity, see p. 66 of that report. For a discussion of the legal and policy issues of data aggregation generally, see *Connecting the Dots*, *supra* note 1 at 58-60; *Frankenstein*, *supra* note 1 at 171-182.

[14] For a detailed discussion of these issues, including a lengthy analysis of the interaction among scope of access, sensitivity of data, and method of query in determining reasonableness, see *Towards a Calculus of Reasonableness*, in *Frankenstein*, *supra* note 1 at 202-217.

[15] For a discussion of how the "reasonableness" of decision thresholds should vary with threat environment and security needs, see *Frankenstein*, *supra* note 1 at 215-217 ("No system … should be … constantly at ease or constantly at general quarters.")

and the subjective-legal argument. Both appear specious, exhibiting different forms of inductive fallacies. [16]

The pseudo-technical argument contends that the benefits to security of predictive data mining are minimal by concluding that "predictive data mining is not useful for counterterrorism" [17] and the cost to privacy and civil liberties is too high. This view is generally supported through erecting a "straw man argument" using commercial data mining as a false analogy and applying a naïve understanding of how data mining applications are actually deployed in the counterterrorism context.

The subjective-legal argument contends that predictive pattern-matching is simply unconstitutional. This view is based on a sophistic reading of legal precedent.

Although much of the concern behind these arguments is legitimate—that is, there are significant policy and privacy issues to be addressed—there are important insights and subtleties missing from the critics' technical and legal analysis that misdirect the public debate.

A. The Pseudo-technical Arguments Against Data Mining.

The pseudo-technical arguments are exemplified in the recent Cato brief referred to earlier, [18] which proceeds in the main like this: predictive data mining is not useful for counterterrorism applications because (1) its use in commercial applications only generates slight improvements in target marketing response rates, (2) terrorist events are rare and so no useful patterns can be gleaned (the "training set" problem), and (3) the combination of (1) and (2) lead to such a high number of false positives so as to overwhelm or waste security resources and impose an impossibly high cost in terms of privacy and civil liberties.

---

[16] In addition, these arguments are not unique to data mining. The problems of efficacy, "training sets", and false positives (as discussed below) are problems common to all methods of intelligence in the counterterrorism context. So, too, the issue of probabilistic predicate and non-particularized suspicion (also discussed below) are common to any preventative or preemptive policing strategy.

[17] *See, e.g.,* Jonas & Harper, *supra* note 2 at 7.

[18] The use of the Cato brief as exemplar of the pseudo-technical argument is not intended as an attack on the authors, both of whom are well-respected and knowledgeable in their respective fields. Indeed, it is precisely the point that even relatively knowledgeable people perpetuate popular misunderstanding regarding the use of data mining in counterterrorism applications. Even within the technical community there is significant divergence in understanding about what these technologies can do, what particular government research programs entail, and the potential impact on privacy and civil liberties of these technologies and programs. *Compare, e.g.*, the Letter from Public Policy Committee of the Association for Computing Machinery (ACM) to Senators John Warner and Carl Levin (Jan. 23, 2003) (expressing reservations about the TIA program) *with* the view of the Executive Committee of the Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD) of the of the ACM, *Data Mining is NOT Against Civil Liberties* (June 30, *rev'd* July 28, 2003) (defending data mining technology and expressing concern that the public debate has been ill-informed and misleading).

While seemingly intuitive and logical on their face, these arguments fall flat upon analysis:

1. The False Analogy and the Base Rate Fallacy

Commercial data mining is propositional (uses statistically independent individual records) but counterterrorism data mining combines propositional with relational data mining. Commercial data mining techniques are generally applied against large transaction databases in order to classify people according to transaction characteristics and extract patterns of widespread applicability. They are most used in the area of consumer direct marketing and this is the example most used by critics.

In counterterrorism applications, however, the focus is on a smaller number of subjects within a large background population that may exhibit links and relationships, or related behaviors, within a far wider variety of activities. Thus, for example, a shared frequent flyer account number may or may not be suspicious alone, but sharing a frequent flyer number with a known or suspected terrorist is and should be investigated. And, to find the latter, you may need to screen the former. [19]

Commercial data mining is focused on classifying propositional data from homogeneous databases (of like-transactions, for example, book sales), while counterterrorism applications seek to detect rare but significant relational links between heterogeneous data (representing a variety of activity or relations) among risk-adjusted populations. In general, commercial users have been concerned with identifying patterns among unrelated subjects based on their transactions in order to make predictions about other unrelated subjects doing the same. Intelligence analysts are interested in identifying patterns that evidence organization or activity among related subjects (or subjects pursuing related goals) in order to expose additional related or like subjects or activities. It is the network itself that must be identified, analyzed, and acted upon. [20]

---

[19] The relevant risk-adjusted population to be screened initially in this example might be all frequent flyer accounts, which would then be subject to two subsequent stages of classification: the first to screen for shared accounts, and the second to screen for shared accounts where one entity or attribute had some suspected terrorist "connection," for example a phone number known to have been used previously by suspected terrorists). Such analyses simply cannot be done manually. More intrusive investigation or analysis would be conducted only against the latter in subsequent stages (and further investigation, data access, or analysis, could be subject to any appropriate legal controls required by the context, for example a FISA warrant to target communications, etc.). See the discussion of multi-pass screening in subsection *False Positives, infra*, for a discussion of how such architecture reduces false positives and provides opportunities to minimize privacy intrusions by controlling access and revelation at each stage.

[20] Covert social networks exhibit certain characteristics that can be identified. *Post-hoc* analysis of the September 11 terror network shows that these relational networks exist and can be identified, at least after the fact. Vladis E. Krebs, *Uncloaking Terrorist Networks*, FIRST MONDAY (mapping and analyzing the relational network among the September 11 hijackers). Research on mafia and drug smuggling networks show characteristics particular to each kind of organization, and current social network research in counterterrorism is focused on identifying unique characteristics of terror networks. *See generally* Philip Vos Fellman & Roxana Wright, *Modeling Terrorist Networks: Complex Systems at the Mid-Range*, presented at Complexity, Ethics and Creativity Conference, LSE, Sept. 17-18, 2003; Joerg Raab & H. Briton Milward, *Dark networks as problems*, J. OF PUB. ADMIN. RES. & THEORY, Vol.13 No.4 at 413-439

Thus, the low incremental improvement rates exhibited in commercial direct marketing applications are simply irrelevant to assessing counterterrorism applications because the analogy fails to consider the implications of relational versus propositional data, and, as discussed below in *False Positives*, ranking versus binary classification, and multi-pass versus single-pass inference. [21]

However, even if the analogy was valid, the proponents of this argument fundamentally misinterpret the outcome of commercial data mining by failing to account for base rates in their examples. [22] For instance, in the Cato brief the authors describe how the Acme Discount retailer might use "data mining" to target market the opening of a new store. [23] In their example, Acme targets a particular consumer demographic in its new market based on a "data mining" analysis of their existing customers. Citing direct marketing industry average response rates in the low to mid single digits, the authors then conclude that the "false positives in marketers' searches for new customers are typically in excess of 90 percent."

The fallacy in this analysis is not accounting for the base rate of the observation in the general population of the old market when assessing the success in the new market. For simple example, suppose that an analysis of Acme's existing customers in the old market showed that all of their current customers "live in a home worth $150,000-$200,000." [24] Acme then targets the same homeowners in the new market but only gets a 5 percent response rate, implying for the authors of the Cato brief a ninety-five percent false positive rate. But, if the number of their customers in the old market was only equal to 5 percent of the demographic in that general population (in other words, 100% of their customers fit the profile but their total number of customers was just 5 percent of homeowners in that demographic within the old market), then the 5 percent response rate in the new market is actually a 100% "success" rate, as they had 5 percent of the target market in their old market, and have captured 5 percent in the new market.

(2003); Matthew Dombroski *et al, Estimating the Shape of Covert Networks*, PROCEEDINGS OF THE 8TH INT'L COMMAND AND CONTROL RES. AND TECH. SYMPOSIUM (2003); H. Brinton Milward & Joerg Raab, *Dark Networks as Problems Revisited: Adaptation and Transformation of Islamic Terror Organizations since 9/11*, presented at the 8th Publ. Mgt. Res. Conference at the School of Policy, Planning and Development at University of Southern California, Los Angeles (Sept. 29-Oct. 1, 2005); D. B. Skillicorn, *Social Network Analysis Via Matrix Decomposition*, in EMERGENT INFORMATION TECHNOLOGIES AND ENABLING POLICIES FOR COUNTER TERRORISM (Robert Popp and John Yen, *eds*., Wiley-IEEE, Jun. 2006).

[21] *See* David Jensen, Matthew Rattigan & Hannah Blau, *Information Awareness: A Prospective Technical Assessment*, Proceedings of the 9th ACM SIGKDD '03 International Conference on Knowledge Discovery and Data Mining (Aug. 2003).

[22] The "base rate fallacy," also called "base rate neglect," is a well-known logical fallacy in statistical and probability analysis in which base rates are ignored in favor of individuating results. *See, e.g.,* Maya Bar-Hillel, *The base-rate fallacy in probability judgments,* ACTA PSYCHOLOGICA Vol.44 No.3 (1980).

[23] Jonas & Harper, *supra* note 2 at 7.

[24] *Cf., id*.

The use of propositional data mining simply allows Acme to reduce the cost of marketing to only those likely to respond, and is not intended to infer or assume that 100 percent of those targeted would respond. If the target demographic in the new market was half the general population, then Acme has improved its potential response rate 100 percent—from 2.5 percent (if they had had to target the entire population) to 5 percent (by targeting only the appropriate demographic) thus, reducing their marketing costs by half. In data mining terms, this is the "lift"—the increased response rate in the targeted population over that that would be expected in the general population. In the context of counterterrorism, any appreciable "lift" results in a better allocation of limited analytic or security resources. [25]

2. The "Training Set" Problem.

Another common argument opposing the use of data mining in counterterrorism applications is that the relatively small number of actual terrorist events implies that there are no meaningful patterns to extract. Because propositional data mining in the commercial sector generally requires training patterns derived from millions of transactions in order to profile the typical or ideal customer or to make inferences about what an unrelated party may or may not do, proponents of this argument leap to the conclusion that the relative dearth of actual terrorist events undermines the use of data mining or pattern-analysis in counterterrorism applications. [26]

Again, the Cato brief advances this argument: "Unlike consumers' shopping habits and financial fraud, terrorism does not occur with enough frequency to enable creation of valid predictive models." [27] However, in counterterrorism applications patterns can be inferred from lower-level precursor activity—for example, illegal immigration, identity theft, money transfers, front businesses, weapons acquisition, attendance at training camps, targeting and surveillance activity, and recruiting activity, among others. [28]

By combining multiple independent models aimed at identifying each of these lower level activities in what is commonly called an ensemble classifier, the ability to make inferences about (and potentially disrupt) the higher level, but rare, activity—the terror attack—is greatly improved. [29]

---

[25] Thus, even a nominal lift, say the equivalent of that in the direct marketing example, would be significant for purposes of allocating analytic resources in counterterrorism in the pre-first stage selection of a risk-adjusted population to be classified (as described in the discussion of multi-stage architectures in, *False Positives, infra*).

[26] The statistical significance of correlating behavior among unrelated entities is highly dependent on the number of observations, however, the correlation of behaviors among related parties may only require a single observation.

[27] Jonas & Harper, *supra* note 2 at 8.

[28] *See*, *e.g.*, David Jensen, *Data Mining in Networks*, Presentation to the Roundtable on Social and Behavior Sciences and Terrorism of the National Research Council, Division of Behavioral and Social Sciences and Education, Committee on Law and Justice (Dec. 1, 2002)

[29] Also, because of the relational nature of the analysis, using ensemble classifiers actually reduces false positives because false positives flagged through a single relationship with a "terrorist identifier" will

Additionally, patterns can be derived from "red-teaming" potential terrorist activity or attributes. Critics of data mining are quick to attack such methods as based on "movie plot" scenarios that are unlikely to uncover real terrorist activity. [30] But, this view is based on a misunderstanding of how terrorist red teaming works. Red teams do not operate in a vacuum without knowledge of how real terrorists are likely to act.

For example, many Jihadist web sites provide training material based on experience gained from previous attacks. In Iraq, for instance, insurgent web sites explain in great detail the use of Improvised Explosive Devices (IEDs) and how to stage attacks. Other sites aimed at global jihad and not tied to the conflict in Iraq describe more generally how to stage attacks on rail lines, airplanes, or other infrastructure, and how to take advantage of Western security practices. So-called "tradecraft" web sites provide analysis of how other plots were uncovered and provide countermeasure training. [31] All of these, combined with detailed review of previous attacks and methods as well as current intelligence reports, provide insight into how terrorist activity is likely to be carried out in the future, particularly by loosely affiliated groups or local "copycat" cells who may get much of their operational training through the Internet.

Another criticism leveled at pattern-analysis and matching is that terrorists will "adapt" to screening algorithms by adopting countermeasures or engaging in other avoidance behavior. [32] However, it is a well-known adage of counterterrorism strategy that increasing the "cost" of terrorist activity by forcing countermeasures or avoidance behavior increases the risk of detection by creating more opportunities for error as well as opportunities to spot avoidance behavior that itself may exhibit an observable signature.

---

be quickly eliminated from further investigation since a true positive is likely to exhibit multiple relationships to a variety of independent identifiers. *Id.* and see discussion in *False Positives, infra.* The use of ensemble classifiers also conforms to the governing legal analysis for determining reasonable suspicion that requires reasonableness to be judged on the "totality of the circumstances" and allows for officers "to make inferences from and deductions about the cumulative information available." *See, e.g.,* U.S. v. Arvizu, 534 U.S. 266 (2002).

[30] *See, e.g.,* Bruce Schneier, *Terrorists Don't Do Movie Plots*, WIRED (Sep. 8, 2005). *See also* Citizens' Protection in Federal Database Act of 2003, seeking to prohibit the "search or other analysis for national security, intelligence, or law enforcement purposes of a database based solely on a hypothetical scenario or hypothetical supposition of who may commit a crime or pose a threat to national security." S. 1484, 108th Cong. §4(a) (2003).

[31] Following the arrest warrants issued in 2005 by an Italian judge for 13 alleged Central Intelligence Agency operatives for activity related to extraordinary renditions, several Jihadist websites posted an analysis of tradecraft errors outlined in news reports and the indictment and alleged to have been committed by the CIA agents. These tradecraft errors included the use of traceable cell phones that allowed Italian authorities to track the agents, and the Jihadist websites supplied countermeasure advice.

[32] *See, e.g.*, the oft-cited but rarely read student paper Samidh Chakrabarti & Aaron Strauss, *Carnival Booth: An Algorithm for Defeating the Computer-assisted Passenger Screening System* (2003). Obviously, if this simplistic critique was taken too seriously on its face it would support the conclusion that locks should not be used on homes because locksmiths (or burglars with locksmithing knowledge) can defeat them. No single layer of defense can be effective against all attacks, thus, effective security strategies are based on defense in depth. In a layered system, the very strategy suggested by the paper is likely to lead to discovery of some members of the group, which through relational analysis is likely to lead to the others.

For instance, in IRA-counterterror operations the British would often watch secondary roads when manning a roadblock at a major intersection to try to spot avoidance behavior. So too, at Israeli checkpoints and border crossings, secondary observation teams are often assigned to watch for avoidance behavior in crowds or surrounding areas. Certain avoidance behavior and countermeasures detailed on Jihadist websites can be spotted through electronic surveillance, as well as potentially through more general data analysis. [33] Indeed, it is an effective counterterrorism tactic to "force" observable avoidance behavior by engaging in activity that elicits known countermeasures and then searching for those signatures.

3. False Positives.

It is commonly agreed that the use of classifiers to detect extremely rare events—even with a highly accurate classifier—is likely to produce mostly false positives. For example, assuming a classifier with a 99.9% accuracy rate applied to the U.S. population of approximately 300 million, and assuming only 3000 true positives (.001%), then some 299,997 false positives and 2997 true positives would be identified through screening—meaning over 100 times more false positives than true positives were selected and 3 true positives would be missed (i.e., there would be 3 false negatives). However, generalizing this simple example to oppose the use of data mining applications in counterterrorism is based on a naïve view of how actual detection systems function and is falsely premised on the assumption that a single classifier operating on a single database would be used and that all entities classified "positive" in that single pass would suffer unacceptable consequences. [34]

In contrast, real detection systems employ ensemble and multiple stage classifiers to carefully selected databases, with the results of each stage providing the predicate for the next. [35] At each stage only those entities with positive classifications are considered for the next and thus subject to additional data collection, access, or analysis at subsequent stages. This architecture significantly improves both the accuracy and privacy impact [36]

---

[33] It would be inappropriate to speculate in detail in open session how certain avoidance behavior or countermeasures can be detected in information systems.

[34] *See* Ted Senator, *Multi-stage Classification*, Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM '05) pp. 386-393 (2005) and *see* Jensen, *supra* note 21. Among the faulty assumptions that have been identified in the use of simplistic models to support the false positive critique are: (1) assuming the statistical independence of data (appropriate for propositional analysis but not for relational analysis), (2) using binary (rather than ranking) classifiers, and (3) applying those classifiers in a single pass (instead of using an iterative, multi-pass process). An enhanced model correcting for these assumptions has been shown to greatly increase accuracy (as well as reduce aggregate data utilization). *Id.*

[35] *See* Senator, *supra* note 34 and Jensen, *supra* note 21, for a detailed discussion of how ensemble classifiers, rankings, multi-pass inference, known facts, relations among records, and probabilistic modeling can be used to significantly reduce false positives.

[36] In multi-stage iterative architectures privacy concerns can be mitigated through selective access and selective revelation strategies applied at each stage (for example, early stage screening can be done on anonymized or de-identified data with disclosure of underlying data requiring some legal or policy procedure). Most entities are dismissed at early stages where privacy intrusions may be minimal.

of systems, reduces false positives, and significantly reduces data requirements. [37] On first glance, such an architecture might also suggest the potential for additional false negatives since only entities scored positive at earlier stages are screened at the next stage, however, in relational systems where classification is coupled with link analysis, true positives identified at each subsequent stage provide the opportunity to reclaim false negatives from earlier stages by following relationship linkages back. [38]

Research using model architectures incorporating an initial risk-adjusted population selection, two subsequent stages of classification, and one group (link) detection calculation has shown greatly reduced false positive selection with virtually no false negatives. [39] A simplistic description of such a system includes the initial selection of a risk-adjusted group in which there is "lift" from the general population, that is, where the frequency of true positives in the selected group exceeds that in the background population. First stage screening of this population then occurs with high *selectivity* (that is, with a bias towards more false positives and fewer false negatives). Positives from the first stage are then screened with high *sensitivity* in the second stage (that is, with more accurate but costly [40] classifiers creating a bias towards only true positives). In each case, link analyses from true positives are used at each stage to recover false negatives from prior stages. Comparison of this architecture with other models has shown it to be especially advantageous for detecting extremely rare phenomena. [41]

Thus, early research has shown that multi-stage classification is a feasible design for investigation and detection of rare events, especially where there are strong group linkages that can compensate for false negatives. These multi-stage classification techniques can significantly reduce—perhaps to acceptable levels—the otherwise unacceptably large number of false positives that can result from even highly accurate single stage screening for rare phenomena. Such architecture can also eliminate most entities from suspicion early in the process at relatively low privacy costs. [42] Obviously, at each subsequent stage additional privacy and screening costs are incurred. Additional research in real world detection systems is required to determine if these costs can be reduced to acceptable levels for wide-spread use. The point is not that all privacy risks

---

[37] The Cato brief perpetuates another common fallacy in stating that "predictive data mining requires lots of data" (p.8). In fact, multi-stage classifier systems actually reduce the overall data requirement by incrementally accessing more data only in subsequent stages for fewer entities. In addition, data mining reduces the need to collect collateral data by focusing analysis on only relevant data. *See* Jensen, *supra* note 21.

[38] Thus, in actual practice, counterterrorism applications combine both "predictive data mining" (as defined and criticized in the Cato brief) with "pulling the strings" (as defined and lauded in the Cato brief).

[39] Senator, *supra* note 34.

[40] "Costly" in this context may mean with greater data collection, access, or analysis requirements with attendant increases in privacy concerns.

[41] Senator, supra note 34.

[42] Initial selection and early stage screening might be done on anonymized or de-identified data to help protect privacy interests. Additional disclosure or more intrusive subsequent analysis could be subject to any legal or other due process procedure appropriate for the circumstance in the particular application.

can be eliminated—they cannot be—only that these technologies can improve intelligence gain by helping better allocate limited analytic resources and that effective system design together with appropriate policies can mitigate many privacy concerns.

Recognizing that no system—technical or other [43]—can provide absolute security or absolute privacy also means that no technical system or technology ought to be burdened with meeting an impossible standard for perfection, especially prior to research and development for its particular use. Technology is a tool and as such it should be evaluated by its ability to either improve a process over existing or alternative means or not. Opposition to research programs on the basis that the technologies "might not work" is an example of what has been called the "zero defect" culture of punishing failure, a policy that stifles bold and creative ideas. [44]

B. The Subjective-legal Arguments Against Data Mining.

To some observers, predictive data mining and pattern-matching also raise Constitutional issues. In particular, it is argued that probability-based suspicion is inherently unreasonable and that pattern-matching does not satisfy the particularity requirements of the Fourth Amendment. [45]

However, for a particular method to be categorically Constitutionally suspect as unreasonable, its probative value—that is, the confidence interval for its particular use—is the relevant criterion. Thus, for example, racial profiling may not be the sole basis for a reasonable suspicion for law enforcement purposes because race has been determined to not be a reliable predictor of criminality. [46]

However, to assert that automated pattern analysis based on behavior or data profiles is *inherently* unreasonable or suspect without determining its efficacy in the circumstances of a particular use seems analytically unsound. The Supreme Court has specifically held that the determination of whether particular criteria are sufficient to meet the reasonable

---

[43] It needs to be recognized that "false positives" are not unique to data mining. All investigative methods begin with more suspects than perpetrators—indeed, the point of the investigative process is to narrow the suspects down until the perpetrator is identified. Nevertheless, the problem of false positives is more acute when contemplating preemptive strategies, however, it is not inherently more problematic when automated. Again, these are legitimate concerns that need to be controlled for through policy development and system design.

[44] *See, e.g.,* David Ignatius, *Back in the Safe Zone*, WASH. POST (Aug. 1, 2003) at A:19.

[45] These and other related legal arguments are discussed in greater detail in *Data Mining and Domestic Security*, supra note 1 at 60-67; *The Fear of Frankenstein*, supra note 1 at 143-159, 176-183, 202-217; and on pp. 7-10 of my testimony to the U.S. House of Representatives Permanent Select Committee on Intelligence (HPSCI) (July 19, 2006).

[46] *See* United States v. Brignoni-Ponce, 422 U.S. 873, 886 (1975). The Court has never ruled explicitly on whether race or ethnicity can be a *relevant* factor for reasonable suspicion under the fourth amendment. *See id.* at 885-887 (implying that race could be a relevant, but not sole, factor). *See also* Whren v. United States, 517 U.S. 806, 813 (1996); Michelle Malkin, IN DEFENSE OF INTERNMENT: THE CASE FOR RACIAL PROFILING IN WORLD WAR II AND THE WAR ON TERROR (2004).

suspicion standard does not turn on the *probabilistic* nature of the criteria but on their *probative* weight:

> The process [of determining reasonable suspicion] does not deal with hard certainties, but with probabilities. Long before the law of probabilities was articulated as such, practical people formulated certain common-sense conclusions about human behavior; jurors as factfinders are permitted to do the same—and so are law enforcement officers. [47]

The fact that patterns of relevant indicia of suspicion may be generated by automated analysis (data-mined) or matched through automated means (computerized pattern-matching) should not change the analysis—the reasonableness of suspicion should be judged on the probative value of the predicate in the particular circumstances of its use—not on its probabilistic nature or whether it is technically mediated.

The point is not that there is no privacy issue involved but that the issue is the traditional one—what subjective and objective expectations of privacy should reasonably apply to the data being analyzed or observed in relation to the government's need for that data in a particular context [48]—not a categorical dismissal of technique based on assertions of "non-particularized suspicion."

Automated pattern-analysis is the electronic equivalent of observing suspicious behavior—the appropriate question is whether the probative weight of any particular set of indicia is reasonable, [49] and what data should be available for analysis. There are legitimate privacy concerns relating to the use of any preemptive policing techniques—but there is not a presumptive Fourth Amendment non-particularized suspicion problem *inherent in the technology or technique* even in the case of automated pattern-matching. Pattern-based queries are reasonable or unreasonable only in the context of their probative value in an intended application—not because they are automated or not.

Further, the particularity requirement of the Fourth Amendment does not impose an irreducible requirement of *individualized* suspicion before a search can be found

---

47 United States v. Cortez, 449 U.S. 411, 418 (1981); and see United States v. Sokolow, 490 U.S. 1, 9-10 (1989) (upholding the use of drug courier profiles).

48 *See* Katz v. United States, 389 U.S. 347, 361 (1967) (Harlan, J., concurring) Setting out the two-part *reasonable expectation of privacy* test, which requires finding both an actual *subjective* expectation of privacy and a *reasonable objective* one:

> My understanding of the rule that has emerged from prior decisions is that there is a twofold requirement, first that a person have exhibited an actual (subjective) expectation of privacy and, second, that the expectation be one that society is prepared to recognize as "reasonable."

49 That is, whether it is a reasonable or rational inference. The Cato brief argues that "reasonable suspicion grows in a mixture of specific facts and rational inferences," supra note 2 at 9, referring to Terry v. Ohio, 392 U.S. 1 (1968) ostensibly to support its position that "predictive, pattern-based data mining" is inappropriate for use because it doesn't meet that standard. But the very point of predictive, pattern-based data mining is to generate support for making rational inferences. *See* Jensen, *supra* note 28.

reasonable, or even to procure a warrant. [50] In at least six cases, the Supreme Court has upheld the use of drug courier profiles as the basis to stop and subject individuals to further investigative actions. [51] More relevant, the court in United States v. Lopez, [52] upheld the validity of hijacker behavior profiling, opining that "in effect ... [the profiling] system itself ... acts as informer" serving as sufficient Constitutional basis for initiating further investigative actions. [53]

Again, although data analysis technologies, including specifically predictive, pattern-based data mining, do raise legitimate and compelling privacy concerns, these concerns are not insurmountable (nor unique to data mining) and can be significantly mitigated by incorporating privacy needs in the technology and policy development and in the system design process itself. By using effective architectures and building in technical features that support policy (including through the use of "policy appliances" [54]) these technologies can be developed and employed in a way that potentially leads to increased security (through more effective intelligence production and better resource allocation) while still protecting privacy interests.

IV. Designing Policy-enabling Architecture and Building in Technical Constraints

Thus, assuming some acceptable baseline efficacy to be determined through research and application experience, I believe that privacy concerns relating to data mining in the context of counterterrorism can be significantly mitigated by developing technologies and

---

[50] An example of a *particular*, but not *individualized*, search follows: In the immediate aftermath of 9/11 the FBI determined that the leaders of the 19 hijackers had made 206 international telephone calls to locations in Saudi Arabia (32 calls), Syria (66), and Germany (29), John Crewdson, *Germany says 9/11 hijackers called Syria, Saudi Arabia*, CHI. TRIB. (Mar. 8, 2006). It is believed that in order to determine whether any other unknown persons—so-called sleeper cells—in the United States might have been in communication with the same pattern of foreign contacts (that is, to uncover others who may not have a direct connection to the 19 known hijackers but who may have exhibited the same or similar *patterns of communication* as the known hijackers) the National Security Agency analyzed Call Data Records (CDRs) of international and domestic phone calls obtained from the major telecommunication companies. (That the NSA obtained these records is alleged in Leslie Cauley, *NSA has massive database of Americans' phone calls*, USA TODAY (May 11, 2006). This is an example of a specific (i.e. likely to meet the Constitutional requirement for particularity)—but not individualized—pattern-based data search.

[51] *See, e.g.,* United States v. Sokolow, *supra* note 47.

[52] 328 F. Supp 1077 (E.D.N.Y. 1971) (although the court in Lopez overturned the conviction in the case, it opined specifically on the Constitutionality of using behavior profiles).

[53] Hijacker profiling was upheld in Lopez despite the 94% false positive rate (that is, only 6% of persons selected for intrusive searches based on profiles were in fact armed). *Id.*

[54] "Policy appliances" are technical control and logging mechanisms to enforce or reconcile policy rules (information access or use rules) and to ensure accountability in information systems and are described in *Designing Technical Systems to Support Policy*, *supra* note 1 at 456. *See also Frankenstein*, *supra* note 1 at 56-58 discussing "privacy appliances." The concept of "privacy appliance" originated with the DARPA TIA project. *See* Presentation by Dr. John Poindexter, Director, Information Awareness Office (IAO), DARPA, at DARPA-Tech 2002 Conference, Anaheim, CA (Aug. 2, 2002); ISAT 2002 Study, Security with Privacy (Dec. 13, 2002); IAO Report to Congress regarding the Terrorism Information Awareness Program at A-13 (May 20, 2003) in response to Consolidated Appropriations Resolution, 2003, No.108-7, Division M, §111(b) [signed Feb. 20, 2003]; and Popp and Poindexter, *supra* note 7.

systems architectures that enable existing legal doctrines and related procedures (or their analogues) to function:

- First, that rule-based processing and a distributed database architecture can significantly ameliorate the general data aggregation problem by limiting or controlling the scope of inquiry and the subsequent processing and use of data within policy guidelines; [55]

- Second, that multi-stage classification architectures and iterative analytic processes together with selective revelation (and selective access) can reduce both the general privacy and the non-particularized suspicion problems, by enabling incremental human process intervention at each stage before additional data collection, access or disclosure (including, in appropriate contexts, judicial intervention or other external due process procedures); [56] and

- Finally, that strong credential and audit features and diversifying authorization and oversight can make misuse and abuse "difficult to achieve and easy to uncover." [57]

Data mining technologies are analytic tools that can help improve intelligence gain from available information thus resulting in better allocation of both scarce human analytic resources as well as security response resources.

Conclusion.

The threat of potential catastrophic outcomes from terrorist attacks raises difficult policy choices for a free society. The need to preempt terrorist acts before they occur challenges traditional law enforcement and policing constructs premised on reacting to events that have already occurred. However, using data mining systems to improve intelligence analysis and help allocate security resources on the basis of risk and threat management may offer significant benefits with manageable harms if policy and system designers take the potential for errors into account during development and control for them in deployment.

Of course, the more reliant we become on probability-based systems, the more likely we are to mistakenly believe in the truth of something that might turn out to be false. That wouldn't necessarily mean that the original conclusions or actions were incorrect. Every decision in which complete information is unavailable requires balancing the cost of false negatives (in this case, not identifying terrorists before they strike) with those of false positives (in this case, the attendant effect on civil liberties and privacy). When mistakes

---

55 *See* Markle Taskforce Second Report*, supra* note 13.

56 *See Connecting the Dots*, *supra* note 1.

57 See Paul Rosenzweig, *Proposals for Implementing the Terrorism Information Awareness System,* 2 Geo. J. L. & Pub. Pol'y 169 (2004); and *Using Immutable Audit Logs to Increase Security, Trust and, Accountability,* Markle Foundation Task Force on National Security Paper (Jeff Jonas & Peter Swire*, lead authors*, Feb. 9, 2006).

are inevitable, prudent policy and design criteria include the need to provide for elegant failures, including robust error control and correction, in both directions.

Thus, any wide-spread implementations of predictive, pattern-based data-mining technologies should be restricted to investigative outcomes (i.e., not automatically trigger significant adverse effects); and should generally be subject to strict congressional oversight and review, be subject to appropriate administrative procedures within executive agencies where they are to be employed, and, to the extent possible in any particular context, be subject to appropriate judicial review in accordance with existing due process doctrines. However, because of the complexity of the interaction among scope of access, sensitivity of data, and method of query, no *a priori* determination that restrictively or rigidly prohibits the use of a particular technology or technique of analysis is possible, or, in my view, desirable. [58] Innovation—whether technical or human—requires the ability to evolve and adapt to the particular circumstance of needs.

Reconciling competing requirements for security and privacy requires an informed debate in which the nature of the problem is better understood in the context of the interests at stake, the technologies at hand for resolution, and the existing resource constraints. Key to resolving these issues is designing a policy and information architecture that can function together to achieve both outcomes, and is flexible and resilient enough to adapt to the rapid pace of technological development and the evolving nature of the threat.

Epilogue

I would again like to thank the Committee for this opportunity to discuss the Privacy Implications of Government Data Mining Programs. These are difficult issues that require a serious and informed public dialogue. Thus, I commend the Chairman and this Committee for holding these hearings and for engaging in this endeavor.

Thank you and I welcome any questions that you may have.

[58] Further, public disclosure of precise authorized procedures or prohibitions will be counterproductive because widespread knowledge of limits enables countermeasures.