Response to follow up questions of Senator Arlen Specter
by Kim A. Taipale (01/30/07)

"Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs"

U.S. Senate Committee on the Judiciary
January 10, 2007

Question 1. **How do you respond [to] Mr. Barr's statement that "it is absurd for the government to use databases to predict individual's future acts"?**

Answer 1. As I stated in my written testimony:

> [P]reemption 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, … prediction of future behavior can only be [based on] evidence of current or past behavior or from associations. [1]

It is a necessary and increasingly mandated function of government intelligence and law enforcement agencies to make predictions about future events—to provide actionable intelligence—particularly in the context of preempting terrorist attacks. Indeed, it is a cardinal objective of counterterrorism intelligence to make probabilistic predictions about possible future behavior based on available information about current or past behavior or associations. Although there are legitimate privacy and civil liberties concerns that need to be addressed with any preemptive approach to terrorism, there should be no intrinsic difference in the policy analysis merely because drawing appropriate inferences (that is, producing actionable intelligence) is augmented through computational means, including "data mining," or if the information to support the inferences resides in "databases."

The difficulty—as highlighted by question 2 below—is in deciding what information or database is appropriate to use, for what purpose, in what circumstances, and with what consequences; and the problem, unfortunately, is that the relevance and appropriateness of using any particular information (or accessing any particular database) to make

---

[1] Written Testimony of Kim A. Taipale on the *Privacy Implications of Government Data Mining Programs* before the U.S. Senate Committee on the Judiciary at 5 (Jan. 10, 2007).

inferences cannot easily be pre-determined (nor judged in isolation without considering the particular circumstances of its use).

To some extent this is exactly where computational analytic applications such as data mining can help—that is, by identifying previously unknown patterns or relationships among data (by providing the data with relational context) they can help focus human intelligence analysts on relevant information.

It is important to again note that the purpose of data analysis in counterterrorism is not to search randomly for purely statistically significant patterns in the abstract. That is, not to find patterns derived merely from statistical correlations among unrelated individuals in order to make predictions about how other unrelated subjects may act in the future. Rather, the purpose is to find, identify, and search for specific patterns of rare occurrences.

Identifying these patterns—for example, relational or link-based patterns like shared phone numbers, addresses, or frequent flyer accounts; or descriptive or predictive patterns like observed or hypothesized behavior of individuals or groups pursuing like outcomes—is not the same as the often vilified "data dredging" for general patterns of simple correlation (in which data mining is criticized for producing irrelevant correlations like "terrorists tend to order pizza with credit cards"). [2]

There is no silver bullet—no technology that will "find terrorists" on its own and no data that can absolutely predict future behavior. However, in appropriate circumstances, data mining can help shift intelligence or law enforcement resources or attention to more productive outcomes by identifying or matching observed, hypothesized, and, in specific contexts, statistically-derived descriptive or predictive models from information contained in databases.

---

[2] *See* Erik Baard, *Buying Trouble: Your grocery list could spark a terror probe*, VILLAGE VOICE (Jul. 30, 2003) (anecdotally describing a correlation model (attributed to an unidentified source) that supposedly "showed 89.7 percent accuracy 'predicting' [the 9/11 hijackers] from the rest of population, [in which] one of the factors was if you were a person who frequently ordered pizza and paid with a credit card.") This fanciful anecdote (which, in any case, conflates a single correlated attribute with a predictive "factor" supporting an inference) became the single unfounded source of rampant uninformed speculation, commentary and criticism about the government seeking to "find terrorists by searching credit card transactions for pizza purchases." *See, e.g.*, Electronic Frontier Foundation, *Comments on Interim Vessel Security Regulations*, USCG-2003-14749, U.S. Dept. of Transportation (2003) ("Data that has been scooped up … include such activities as … those who like to order pizza via credit card.")

Question 2a. **Would you say that the privacy concerns raised at the hearing are not related to the use of the data mining technology but instead to the use of the underlying data, the government and commercial databases that are being analyzed?**

Answer 2a. Many of the privacy concerns raised at the hearing—for example, problems with watch lists—have little to do with data mining. Thus, focusing only on data mining (that is, solely on the method of query or analysis) as the primary policy problem would be a mistake since it is only one of many factors—and certainly not the most important one—that need to be taken into account in considering privacy matters.

As I noted in my oral testimony, privacy *concerns* are a complex function involving scope of access, sensitivity of data, and method of query. How much data and from what source? How sensitive is the data? And, how specific is the query?

Further, privacy *interests* (that is, those privacy concerns entitled to Constitutional or statutory protection because they are recognized as reasonable) cannot be evaluated independently of the context of use—that is, how is the information to be used and with what consequences? What are the government's needs and the consequences of not acting? What are the alternatives? What are the consequences to the individual? What opportunities are there for error correction or redress?

Thus, for example, with a lot of predicate (say, "probable cause") and a very specific query (say, "subject-based") you can tolerate as reasonable quite severe privacy intrusions and consequences to the individual, even in a free society. However, even ambiguous predicate and a less particular query (say, a hypothesized "predictive pattern") might be reasonable where there are minor consequences to the individual (for example, a simple follow up data match against a watch list), robust error detection and correction for inferences that turn out to be invalid, and where there may be catastrophic consequences in not acting.

The relationship between scope of access, sensitivity of data, and method of query, and how these relate to reasonableness, due process, and threat, is a complex calculus that I have described elsewhere. [3]

As a policy matter, however, issues relating specifically to the use of data mining technologies for analysis should be distinguished both from (i) issues relating more generally to the collection, aggregation, access, or

[3] For a more detailed discussion of these issues, *see Towards a Calculus of Reasonableness* in *Technology, Security and Privacy: The Fear of Frankenstein, the Mythology of Privacy, and the Lessons of King Ludd*, 7 YALE J. L. & TECH. 123 at 202-217 (Mar. 2004) at http://ssrn.com/abstract=601421.

fusion of the underlying data, on the one hand, and (ii) issues relating to decision-making—that is, determining what thresholds trigger what action, and what consequences flow from such triggers, on the other.

Question 2b. **Do you believe that the government's use of <u>commercial</u> databases raises privacy issues?**

Answer 2b. The use of commercial databases certainly raises additional—or at least different—privacy issues than the use of information collected directly under specific authorities for law enforcement or counterterrorism use. [4]

However, it is not the commercial nature of the source alone that is relevant to the analysis. Thus, it may be useful to consider a spectrum of informational databases, for example:

i. Government databases containing lawfully collected intelligence or law enforcement data,
ii. Government databases containing routinely collected government data (that is, data collected in the ordinary course of providing government services) and that is normally subject to the Privacy Act or other statutory protections (for example, tax information or information collected pursuant to various entitlement reporting requirements),
iii. Commercial databases that contain commercially aggregated public data that are either freely available or can be accessed by anyone for a fee (for example, directories or collections of published material),
iv. Commercial databases that contain government data aggregated from "public" sources and that can be accessed by anyone for a fee (for example, court records, property deeds, licensing information),
v. Commercial databases containing proprietary private data that can be accessed by anyone for a fee (for example, marketing data, subscription lists, etc.),
vi. Commercial databases that contain "regulated" private data that can generally be accessed for a fee for legally authorized purposes (for example, credit reports, or medical or insurance data),
vii. Commercial databases containing proprietary private data generally not available to others (for example, account information, transaction history, telecommunication logs).

---

[4] *See generally* Markle Task Force on National Security in the Information Age, *Second Report: Creating a Trusted Network for Homeland Security* at 30-37, 56-67, 150-162 (2003) (discussing the use of private data for national security purposes) available at http://markletaskforce.org/; James X. Dempsey & Lara M. Flint, *Commercial Data and National Security*, 72 GEO. WASH. L. REV. 1459, at 1465-1468 (2004) (providing a detailed discussion of the policy and legal implication relating to the use commercial data for counterterrorism) at http://www.cdt.org/publications/200408dempseyflint.pdf.

So, for example, using routinely collected government information (ii, above) for counterterrorism purposes may raise many of the same issues as using "commercial" information (particularly, v and vi, above) because of the issues discussed below; while using commercial aggregations of truly publicly-available information (for example, iii and iv, above) may only raise incidental issues of increased government efficiency in accessing information that may not be subject to any general expectations of privacy.

A threshold issue, of course, is whether data lawfully acquired from any of these categories for one purpose should be entirely free of constraints for retention or subsequent use for other purposes as is currently generally the case. For example, even "private" data not generally available to third parties (vii, above) may be available to law enforcement for one purpose, for example, counterterrorism through a national security letter; but should it then be retained, shared and made available as law enforcement or intelligence data (i, above) for any subsequent purpose, reuse, or dissemination without any further use restrictions? (See discussion of "authorized uses" in answer to question 3 below).

Subsequent or secondary use of any data (that is, any use unrelated to the purpose of the original collection or disclosure) raises two related concerns: *data quality or reliability* and *expectations of privacy*. I discuss expectations of privacy in my answer to question 2c, below.

The data quality or reliability concern is that data collected for one purpose may not be suitable for another. Thus, data collected for a routine government or commercial purposes where the consequences of using erroneous data are innocuous may not be appropriate for use in a context where outcomes may be consequential. This may be an even greater problem with the use of commercial data since commercial data users tend to deal with error purely as a percentage cost of aggregate benefit (thus, they "invest" in accuracy only on an aggregated basis), whereas use in counterterrorism may have significant individuated consequences.

The commonly expressed example of this is that the consequences of using bad marketing data in the private sector are that someone may receive junk mail that they are not interested in—incurring a slight cost to the commercial data user and a minimal intrusion on the individual. However, the consequences of using that same erroneous data in counterterrorism may be more severe—both for the government user who may rely on the information and to the individual who may become the object of government action.

The problem may be exacerbated when the data is not subject to any mandated quality requirements—for example, when routine government

information becomes exempt from the data accuracy requirements of the Privacy Act through the law enforcement or national security exceptions, or when commercial data subsequently used in law enforcement is never subject to such requirements in the first place. Thus, the data reliability problems associated with data repurposing—especially of commercial data—must be recognized and addressed.

Therefore, as a matter of sound policy and to the extent possible, all data—regardless of where it originates—should be subject to some data quality assessment appropriate to its use in specific counterterrorism applications. Further, the severity of the consequences resulting from its use should generally relate proportionally to its reliability. Thus, for example, a different, and perhaps lower, accuracy standard could be acceptable for information used for general investigative purposes (as long as the potential for error is calibrated) than would be acceptable for information used to deny a particular person a liberty, for example, the "no-fly" list.

These and other issues relating to the use of private sector data are discussed in the Second Report of the Markle Task Force in the more general context of government information sharing. [5] Parts of that analysis may have relevance here.

Question 2c. **Do individuals have an expectation of privacy with respect to information contained in commercial databases?**

Answer 2c. Individuals have varying expectations of privacy in all their personal information, including information contained in commercial databases. The obligatory analysis, however, requires assessing both the *subjective* expectation of privacy and determining a *reasonable objective* one:

> [T]he 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." [6]

Subjective expectations of privacy for information in databases can vary according to the *sensitivity of the data* and the *purpose or intentionality of the original disclosure*. Thus, subjective expectations relating to very personal or sensitive data, such as financial data or medical data in

---

[5] Markle Task Force on National Security in the Information Age, *Second Report: Creating a Trusted Network for Homeland Security* at 30-37, 56-67, 150-162 (2003) available at http://markletaskforce.org/.

[6] Katz v. United States, 389 U.S. 347, 361 (1967) (Harlan, J., concurring).

commercial databases might be high; while those relating to other data, such as general public information in commercial directories, might not. Likewise, information originally disclosed to third parties incidentally in the ordinary course of life—for example, in commercial transaction records that may include personal information for billing purposes—might be subject to higher subjective expectations of privacy than information specifically disclosed for evaluation, for example, on a disclosure form.

Many of these subjective expectations have been recognized through explicit statutory privacy protection that protect particular classes of information deemed sensitive. These statutes generally require that use of these types of information conform to particular procedures. For example, census data, medical records, educational records, tax returns, cable television records, video rental, etc. are all subject to their own statutory protection, usually requiring an elevated level of procedure, for example, a warrant or court order instead of a subpoena, to gain access.

Nevertheless, the general legal rule is well established—in the absence of specific statutory protection information voluntarily given to a third party can be conveyed by that party to government authorities without violating the Fourth Amendment because there can be no reasonable "expectation of privacy" for information that has already been disclosed. [7] Thus, there is likely no Fourth Amendment prohibition to government acquisition of commercially available data (although the "wholesale" acquisition of entire commercial datasets has not been considered directly). Some have questioned whether this blanket rule is still appropriate where vast amounts of personal information is now maintained by third parties in private sector databases; where storage, search and retrieval tools allow such information to be subsequently and regularly reused for other purposes; and where government seeks to acquire complete datasets rather than information specific to any particular subject of interest. [8]

Nevertheless, it seems foregone that appropriately authorized government agencies should, and will ultimately, have access to data that is generally available from commercial databases. It would be an unusual polity that demanded accountability from its representatives to prevent terrorist acts yet denied them access to tools or information widely available in the private sector. For example, it seems politically untenable that a private debt collector or marketing firm could have legal access to data from a commercial database and that a lawfully acting intelligence agency seeking to prevent a terrorist attack with nuclear weapons would not.

---

[7] *See* United States v. Miller, 425 U.S. 435, 441-443 (1976) (holding that there is no reasonable expectation of privacy in banking records held by third party).

[8] *See, e.g.,* Fred H. Cate, *Legal Standards for Data Mining* in EMERGENT INFORMATION TECHNOLOGIES AND ENABLING POLICIES FOR COUNTER TERRORISM (Robert Popp & John Yen, *eds*., 2006).

Thus, it is the procedures under which access to commercial data should be allowed—that is, under what authorities and with what oversight and review should access and use be permitted. These issues are addressed in part in the answer to the next question.

Question 3. **Do you have any concern that the government is using or may use contracts with private industry to evade privacy laws, FOIA rules, [and] constitutional protections that apply to the government?**

Answer 3. Government outsourcing of traditional government functions—which is currently ongoing in many spheres including military operations, intelligence, law enforcement, and corrections—should generally be subject to the same or analogous Constitutional and statutory protections, oversight, and review as if the government were doing them directly.

In the context of this hearing there are two general types of activity of concern: (i) the outsourcing of information *collection* through the acquisition of commercial data or datasets, and (ii) the outsourcing of *intelligence production or security services* through the use of private contractors to provide analysis or surveillance.

As discussed in the preceding answer, it seems both reasonable and inevitable that properly authorized agencies of the government should have access to data that is commercially available to private parties. The problem arises when such data—once initially acquired for a particular and appropriate purpose—is in effect transformed thereafter into law enforcement or intelligence data not subject to any additional reuse or sharing restrictions. This problem is made worse when government acquires or accesses entire datasets.

Existing laws and policies are generally based only on controlling the initial collection or access to data—not the subsequent use or reuse. These rules were adequate when information retention and subsequent reuse was difficult to accomplish due to technical limitations—privacy was protected in part through these inefficiencies. However, these rules are outdated in the present context in which the use or reuse of available information (not its collection) is the primary challenge. Further, maintaining distinctions based on why the data was originally collected, and by whom, are simply unworkable in the present context of widespread data aggregation and commercial availability of datasets composed from diverse sources.

Thus, these outdated rules should be replaced or supplemented by a new, more flexible and dynamic regime based on an *authorized use standard*. An authorized use standard would improve the government's ability to use

information in appropriate circumstances while still protecting privacy and civil liberties.

An authorized use standard would be a mission- or threat-based justification for accessing or using information in a particular context. The concept of an authorized use standard for sharing lawfully acquired intelligence is discussed in the Third Markle Report.[9] The same kind of analysis and standard may also have more general applicability to the use of commercially available data.

Under an authorized use standard, the use of commercially available data (as well as the use of data mining technologies, for that matter) could be authorized, oversighted, and reviewed according to guidelines based on the legal authorities and specific mission of the government agency involved, the sensitivity of the information, and the intended uses and consequences in the peculiar circumstances and needs surrounding its use. Such a standard would be more flexible—allowing appropriate uses but still protecting privacy and civil liberties—than the existing regime based only on binary control of the initial collection or access.

The outsourcing of intelligence production or security services by directly contracting for analysis or surveillance raises additional issues. As a general rule these contracted services should be subject to similar legal protections as if the government were engaged in them directly. However, it may be that in particular circumstances that these rules or requirements will have to be modified to accommodate the specific differences between contracted services and direct action, and to meet the commercial needs of contractors. So, for example, where government contracts for surveillance or analysis services that *but for* the contracting would be provided directly by a government agency, the rules (including oversight) should be more or less the same as if the government had acted directly. However, where government contracts for analysis or surveillance services that are generally available to any private party on a commercial basis, the appropriate disclosure and oversight regime may have to conform to commercial requirements needed to protect proprietary interests.

Question 4. **There are a number of laws on the books already, such as the Privacy Act and E-government Act of 2002, requiring transparency when the government uses personal information, and there are a number of proposals to increase such transparency that are specifically aimed at data mining. Do you believe transparency is important when it comes to government's use of data mining technology, or do you think that it would hamper the government's ability to use technology effectively?**

---

[9] Markle Task Force on National Security in the Information Age, *Third Report: Mobilizing Information to Prevent Terrorism* at 32-41 (2006) at http://markletaskforce.org/.

Answer 4. "Transparency"—generally achieved through reporting and disclosure requirements—is an essential condition for ensuring effective oversight and accountability. However, there are two issues with respect to proposals to increase transparency specifically for data mining: first, can or should specific reporting and disclosure requirements be based on a technology or method of analysis (particularly one with no agreed definition), and, second, how much disclosure is appropriate without hampering effective uses or compromising national security interests.

Because the appropriateness of any particular use of data mining technology ultimately will be highly conditional on the circumstances of its application, including the specific authorities under which an agency is acting and the particular mission or operational needs at the time of use, it would seem unworkable—except perhaps as an interim step to initiate debate—to impose singular or uniform reporting or disclosure requirements simply based on analytic technique. As a general rule, effective oversight and accountability—including reporting and disclosure—could be better achieved using familiar mechanisms that relate oversight and requirements to specific agencies or jurisdictions. (And, an "authorized use" standard as discussed in the previous answer, would enable appropriate government use of commercially available information and data mining technology while still protecting core privacy and civil liberty values by empowering more focused and, thus, effective oversight.)

Another problem with requiring specific disclosures for "data mining" is that there is no universally accepted definition of what data mining is and, for reasons set forth in my written testimony, there is no easy line to draw between "pattern-based" and other queries. Thus, for example, the definition used in the recently introduced Federal Agency Data Mining Reporting Act of 2007—that is, use of a "predictive pattern or anomaly indicative of … criminal activity" to query a database—would seem to encompass (and make no distinction among) long accepted as appropriate uses like Securities and Exchange Commission programs to identify insider trading and rogue brokers from trading records, Internal Revenue Service programs to select returns for audit, Treasury Department efforts to monitor money laundering, certain telecommunication network monitoring to maintain service, on the one hand, and more controverted programs that seem to be the subject of concern, on the other. The utility of detailed, and perhaps onerous, reporting requirements for all "data mining" programs may be an overly broad legislative response to a narrower concern.

Further, appropriate transparency is not the same thing as public disclosure. Thus, care must be taken in any reporting and oversight

structure to avoid hampering effective uses or compromising national security interests. Thus, general disclosure of government-wide limitations or restrictions—for example, declaring that certain information or technologies were "off limits" in all circumstances or that they can be used only under certain delineated and predetermined operational circumstances—would be inappropriate. Public disclosure of limitations or restrictions—even if only broadly outlined—can encourage and facilitate the development of specific avoidance strategies aimed at taking advantage of known limits. [10] Even simple reporting of programs and disclosure of which agencies are using what data and what technologies is likely to impact effectiveness. [11]

Thus, reporting requirements, disclosure and discussion about what information is or should be available for use by lawfully acting security services under what circumstances, and what technical methods of analysis are appropriate for use in counterterrorism, should be decided and overseen through existing mechanisms—including the Congressional judiciary and intelligence committees—using established procedures and practices designed to protect even broad disclosures that may implicate national security.

However, such oversight can only be successful in enabling appropriate uses while protecting against potential abuse or misuse if all participants work together in good faith in executing their responsibilities.

---

[10] For example, following disclosure of the NSA Terrorist Surveillance Program and broad public discussion of how FISA requirements may be applicable to international telephone conversation that terminate in the United States, some *Jihadist* websites specializing in countermeasure tradecraft have suggested acquiring VoIP telephones with domestic U.S. telephone numbers precisely so as to make surveillance more difficult by appearing to be domestic or U.S. person protected communications even when the calls in fact are wholly foreign.

[11] Just as the mere disclosure of the existence of a particular "spy" satellite (much less its capabilities) is likely to undermine its effectiveness. Overseeing data access and data mining for counterrorism applications must be governed as a national security and intelligence matter, not as a routine law enforcement one.