Testimony of

Mr. Jim Harper

Director of Information Policy Studies
CATO Institute
January 10, 2007

Testimony of Jim Harper
Director of Information Policy Studies, The Cato Institute
to the Senate Judiciary Committee Hearing Entitled
"Balancing Privacy and Security: The Privacy Implications of
Government Data Mining Programs"
January 10, 2007

Chairman Leahy, Members of the Committee --

It is a pleasure and an honor to be with you today to speak about the privacy implications of government data mining. You have chosen a very important issue to lead off what I know will be an aggressive docket of hearings and oversight in the Senate Judiciary Committee during the 110th Congress.

We all want the government to secure the country using methods that work. And we all want the government to cast aside security methods that do not work. The time and energy of the men and women working in national security is too important to be wasted, and law-abiding American citizens should not give up their privacy to government programs and practices that do not materially improve their security.

For the reasons I will articulate below, data mining is not, and cannot be, a useful tool in the anti-terror arsenal. The incidence of terrorism and terrorism planning is too low for there to be statistically sound modeling of terrorist activity.

The use of predictive data mining in an attempt to find terrorists or terrorism planning among Americans can only be premised on using massive amounts of data about Americans' lifestyles, purchases, communications, travels, and many other facets of their lives. This raises a variety of privacy concerns. And the high false-positive rates that would be produced by predictive data mining for terrorism would subject law-abiding Americans to scrutiny and investigation based on entirely lawful and innocent behavior.

I am director of information policy studies at the Cato Institute, a non-profit research foundation dedicated to preserving the traditional American principles of limited government, individual liberty, free markets, and peace. In that role, I study the unique problems in adapting law and policy to the information age. I also serve as a member of the Department of Homeland Security's Data Privacy and Integrity Advisory Committee, which advises the DHS Privacy Office and the Secretary of Homeland Security.

My most recent book is entitled Identity Crisis: How Identification Is Overused and Misunderstood. I am editor of Privacilla.org, a Web-based think tank devoted exclusively to privacy, and I maintain an online resource about federal legislation and spending called WashingtonWatch.com. At Hastings College of the Law, I was editor-in-chief of the Hastings Constitutional Law Quarterly. I speak only for myself today and not for any of the organizations with which I am affiliated or for any colleague.

There are many facets to data mining and privacy issues, of course, and I will discuss them below, but it is important to start with terminology. The words used to describe these information age issues tend to have fluid definitions. It would be unfortunate if semantics preserved disagreement when common ground is within reach.

What is Privacy?

Everyone agrees that privacy is important, but people often mean different things when they talk about it. There are many dimensions to "privacy" as the term is used in common parlance.

One dimension is the interest in control of information. In his seminal 1967 book Privacy and Freedom, Alan Westin characterized privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others." I use and promote a more precise, legalistic definition of privacy: the subjective condition people experience when they have power to control information about themselves and when they have exercised that power consistent with their interests and values. The "control" dimension of privacy alone has many nuances, but there are other dimensions.

The Department of Homeland Security's Data Privacy and Integrity Advisory Committee has produced a privacy "framework" document that usefully lists the dimensions of privacy, including control, fairness, liberty, and data security, as well as sub-dimensions of these values. This "framework" document helps our committee analyze homeland security programs, technologies, and applications in light of their effects on privacy. I recommend it to you and have attached a copy of it to my testimony.

Fairness is an important value that is highly relevant here. People should be treated fairly when decisions are made about them using stores of data. This requires consideration of both the accuracy and integrity of data, and the legitimacy of the decision-making tool or algorithm.

Privacy is sometimes used to refer to liberty interests, as well. When freedom of movement or action is conditioned on revealing personal information, such as when there is comprehensive surveillance, this is also a privacy problem. "Dataveillance" -- surveillance of data about people's actions -- is equivalent to video camera surveillance. The information it collects is not visual, but the consequences and concerns are tightly in parallel.

Data security and personal security are also important dimensions of "privacy" in its general sense. People are rightly concerned that information collected about them may be used to harm them in some way. We are all familiar with the information age crime of identity fraud, in which people's identifiers are used in remote transactions to impersonate them, debts are run up in their names, and their credit histories are polluted with inaccurate information. The Drivers Privacy

Protection Act, Pub. L. No. 103-322, was passed by Congress in part due to concerns that public records about drivers could be used by stalkers, killers, and other malefactors to locate them.

Privacy Issues in Terms Familiar to the Judiciary Committee

I have spoken about privacy in general terms, but these concepts can be translated into language that is more familiar to the Judiciary Committee. For example, if government data mining will affect individuals' life, liberty, or property -- including the recognized liberty interest in travel -- the questions whether information is accurate and whether an algorithm is legitimate go to Fifth Amendment Due Process. Using inaccurate information or unsound algorithms may violate individuals' Due Process rights if they cannot contest decisions that government officials make about them.

If officials search or seize someone's person, house, papers, or effects because he or she has been made a suspect by data mining, there are Fourth Amendment questions. A search or seizure premised on bad data or lousy math is unlikely to be reasonable and thus will fail to meet the crucial standard set by the Fourth Amendment.

I hasten to add that the Supreme Court's Fourth Amendment doctrine has rapidly fallen out of step with modern life. Information that people create, transmit, or store in online and digital environments is just as sensitive as the letters, writings, and records that the Framers sought protection for through the Fourth Amendment, yet a number of Supreme Court precedents suggest that such information falls outside of the Fourth Amendment because of the mechanics of its creation and transmission, or its remote storage with third parties.

A bad algorithm may also violate Equal Protection by treating people differently or making them suspects based on characteristics the Equal Protection doctrine has ruled out.

There are a number of different concerns that the American people rightly have with government data mining. The protections of our constitution are meant to provide them security against threats to privacy and related interests. But before we draw conclusions about data mining, it is important to work on a common terminology to describe this field.

What is Data Mining?

There is little doubt that public debate about data mining has been hampered by the fact that people often do not use common terms to describe the concepts under consideration. Let me offer the way I think about these issues, first by dividing the field of "data analysis" or "information analysis" into two subsets: link analysis (also called subjectbased analysis) and pattern analysis.

Link Analysis

Link analysis is a relatively unremarkable use of databases. It involves following known information to other information. For example, a phone number associated with terrorist activity might be compared against lists of phone numbers to see who has called that number, who has been called by that number, who has reported that number as their own, and so on. When the

number is found in another database, a link has been made. It is a lead to follow, wherever it goes.

This is all subject to common sense and (often) Fourth Amendment limitations: The suspiciousness or importance of the originating information and of the new information dictates what is appropriate to do with, or based on, the new information. Following links is what law enforcement and national security personnel have done for hundreds of years. We expect them to do it, and we want them to do it. The exciting thing about link analysis in the information age is that observations made by different people at different times, collected in databases, can now readily be combined. As Jeff Jonas and I wrote in our recent paper on data mining:

"Data analysis adds to the investigatory arsenal of national security and law enforcement by bringing together more information from more diverse sources and correlating the data. Finding previously unknown financial or communications links between criminal gangs, for example, can give investigators more insight into their activities and culture, strengthening the hand of law enforcement."

Jonas is distinguished engineer and chief scientist with IBM's Entity Analytic Solutions Group. I have attached our paper, Effective Counterterrorism and the Limited Role of Predictive Data Mining to my testimony.

Following links from known information to new information is distinct from patternbased analysis, which is where the concerns about "data mining" are most merited.

Pattern Analysis

Pattern analysis is looking for a pattern in data that has two characteristics: 1) It is consistent with bad behavior, such as terrorism planning or crime; and 2) it is inconsistent with innocent behavior.

In our paper, Jonas and I wrote about the classic Fourth Amendment case, Terry v. Ohio, where a police officer saw Terry walking past a store multiple times, looking in furtively. This was 1) consistent with criminal planning ("casing" the store for robbery) and 2) inconsistent with innocent behavior - it didn't look like shopping, curiosity, or unrequited love of a store clerk. The officer's "hunch" in Terry can be described as a successful use of pattern analysis before the age of databases.

There are three ways that seem to be used (or, at least, have been proposed) to develop similar "hunches" -- or suitable patterns in data: 1) historical information; 2) redteaming; and 3) anomaly.

Historical Patterns

As Jonas and I discuss in our paper, marketers use historical information to find the patterns that they use as their basis for action. They try to figure out which combinations of variables among current customers make them customers. When the combinations of variables are found again, this points them to potential new customers, and it merits them sending a mailer to the prospects'

homes, for example. Credit issuers do the same things, and there is a fascinating array of different ways that they slice and dice information seeking after good credit risks that other credit issuers have not found. Historical data is widely accepted in these areas as a tool for finding patterns, and consumers enjoy economic benefits from these processes.

Historical patterns can also form the basis for discovery of relatively common crimes, such as credit card fraud. With many thousands of examples per year, credit card networks are in a position to develop patterns of fraud based on historical evidence. Finding these patterns in current data, they are justified in calling their customers to ask whether certain charges are theirs. Jonas and I call this "predictive data mining" because the historical pattern predicts with suitable accuracy that a certain activity or condition (credit card fraud, a willing buyer, etc.) will be found when the pattern is found.

However, the terrorism context has a distinct lack of historical patterns to go on. In our paper, Jonas and I write:

"With a relatively small number of attempts every year and only one or two major terrorist incidents every few years--each one distinct in terms of planning and execution--there are no meaningful patterns that show what behavior indicates planning or preparation for terrorism."

The lack of historical patterns is just half of the problem with finding terrorists using pattern analysis.

False Positives

The rarity of terrorists and terrorist acts is good news, to be sure, but it further compounds the problem of data mining to find them: When a condition is rare, even a very accurate test for it will result in a high number of false positives. Even a highly accurate test is often inappropriate to use in searching for a rare condition among a large group.

In our paper, Jonas and I illustrate this using a hypothetical test for disease that would accurately detect it 99% of the time and yield a false positive only 1 percent of the time. If the test indicated the disease, the protocol would call for a doctor to perform a biopsy on the patient to confirm or falsify the test result.

If 0.1 percent of the U.S. population had the disease, 297,000 of the 300,000 victims would be identified by running the test on the entire population. But doing so would falsely identify 3 million people as having the disease and subject them to an unnecessary biopsy. Running the test multiple times would drive false positives even higher.

The rarity of terrorists and terrorism planning in the U.S. means that even a highly accurate test for terrorists would have very high false positives. This, we conclude, would render predictive data mining for terrorism more harmful than beneficial. It would cost too much money, occupy too much investigator time, and do more to threaten civil liberties than is justified by any improvement in security it would bring.

"Red-Teaming"

A second way to create patterns is "red-teaming." This is the idea that one can create patterns to look for by planning an attack and then watching what data is produced in that planning process, or in preliminaries to carrying out the attack. That pattern, found again in data, would indicate planning or preparation for that type of attack.

This technique was not a subject of our paper, but many of the same problems apply. The pattern developed by red-teaming will match terrorism planning -- it is, after all, synthesized planning. But, to work, it must also not fit a pattern of innocent behavior.

Recall that after 9/11 people were questioned and even arrested for taking pictures of bridges, monuments, and buildings. To common knowledge, photographing landmarks fits a pattern of terrorism planning. After all, terrorists need to case their targets. But photographing landmarks fits many patterns of innocent behavior also, such as tourism, photography as a hobby, architecture, and so on. This clumsy, improvised 'red-teaming' failed the second test of pattern development.

Formal red-teaming would surely be more finely tuned, but it still would have to overcome the false positive problem. Given an extremely small number of terrorists or terrorist activities in a large population, near perfection would be required in the pattern, or it would yield massive error rates, invite waste of investigative energy, and threaten privacy and civil liberties.

It seems doubtful that red teams would be able to devise an attack with a data profile so narrow that it does not create excessive false positives, yet so broad that it matches some group's plan for a terror attack. To me, using red-teaming this way has all the plausibility of stopping a fired bullet with another bullet.

Red-teaming can be useful, it seems, but not for data analysis. If red-teaming were to come up with a viable attack, the means of carrying out that attack should be foreclosed directly with new security measures applied to the tool or target of the attack -- never mind who might carry it out. It would be gross malpractice for anyone in our national security services to conceive of an attack on our infrastructure or people, and then fail to secure against the vulnerability directly while watching for the attack's pattern in data.

Anomaly

Without historical or red-team patterns, some have suggested that anomaly should be the basis of suspicion. Given the patterns in data of "normal" behavior, things deviating from that might be regarded as suspicious. (This is actually a version of historical patterning, but the idea is to find deviation from a pattern rather than matching to a pattern.)

It is downright un-American to think that acting differently could make a person a suspect. On a practical level, one-in-a-million things happen a million times a day. Looking for anomalies will turn up lots of things, but none relevant. And terrorists could avoid this technique by acting as normally as possible. In short, anomaly is not a legitimate basis for forming suspicion.

Historical-pattern-based data analysis -- what Jeff Jonas and I call "predictive data mining" -- has many uses in things such as medical research, marketing and credit scoring, many forms of

scientific inquiry, and other searches for knowledge. It is not useful in the terrorist discovery problem. Searching for "red-teamed" patterns and for anomalies has many of the same flaws.

Data Mining for Terrorists Does Not Work

The conclusion whether a type of data analysis "works" turns on the most important question in the data-analysis analysis: What action does a "match" create a predicate for? When a link, pattern, or deviation from a pattern has been established, and then it is found in the data, what action will be taken?

When marketers use a historical pattern to determine who will receive a promotional flyer, this predictive data mining "works" even if it is wrong 95% of the time. The cost of being wrong may be 50 cents for mailing it, and a few moments of time for the person wrongly identified as a potential customer.

Predictive data mining is appropriate for seeking credit card fraud. A call to a customer from the credit issuer will reassure the customer whether he or she is correctly targeted or not.

Predictive data mining and other forms of pattern analysis might be used to send beat cops to a certain part of town. The harm from being wrong is some wasted resources -- which nobody wants, of course -- but there is no threat to individual rights.

If, on the other hand, government officials are using data mining to pull U.S. citizen travelers out of line, if they are using patterns to determine that phones in the United States should be tapped, and so on, data mining does not "work" unless it is quite a bit more accurate.

The question whether data mining works is not a technical one. It is not a question for computer or database experts to answer. It is a question of reasonableness under the Fourth Amendment, to be determined by the courts, by Congress, and, broadly speaking, by the society as a whole.

Because of the near statistical impossibility of catching terrorists through data mining, and because of its high costs in investigator time, taxpayer dollars, lost privacy, and threatened liberty, I conclude that data mining does not work in the area of terrorism.

But my conclusion should not be determinative. Rather, it should be an early part of a national conversation about government data analysis, the applications in which data analysis and data mining "work," and those in which it does not.

Fairness, Reasonableness, and Transparency

One of the most important places for that conversation to happen is in Congress -- here in this Committee -- and in the courts. This hearing begins to shed light on the questions involved in data mining.

But government data mining programs must also be subjected to the legal controls imposed by the Constitution. The question whether a data analysis program affecting individuals meets constitutional muster brings us to the final important question: whether the program provides redress.

"Redress" is data-analysis jargon for Due Process. If a data mining or other data analysis system is going to affect individuals' rights or liberty, Due Process requires that the person should be able to appeal or contest the decision made using the system, ultimately -- if not originally -- in a court of law.

This requires two things, I think: access to the data that was analyzed in determining that the person should be singled out, and access to the pattern or link information that was used to determine that the person should be singled out.

Access to data is like asking the police officer in Terry v. Ohio what he saw when he determined that he should pat down the defendant. Was the officer entitled to look where he looked? Was he paying sufficient attention to the defendants' actions? We would not deny defendants the chance to explore these questions in a criminal court, and should not let data mining that affects individuals' liberties escape similar scrutiny.

Access to the pattern/algorithm allows review analogous to determining whether the officer's decision to pat down Terry was, as required by the Fourth Amendment, reasonable. Was the pattern of behavior he saw so consistent with wrongful behavior, and so inconsistent with innocent behavior, that it justifies having law enforcement intervene in the privacy and repose of the presumed innocent? This question can and should be asked of data mining programs.

Government data mining and data analysis may seem to involve highly technical issues, reserved for computer and database experts. But, again, the most important questions are routinely addressed by this Committee, by Congress, by the press, and by the American people. The questions are embedded in the Constitution's Fourth and Fifth Amendments and the Supreme Court's precedents. They are about simple fairness: Do these systems use accurate information? Do they draw sensible conclusions? And do their findings justify the actions officialdom takes because of them?

Citizens must have full redress/Due Process when their rights or liberties are affected by government data mining or other data analysis programs, just as when their rights or liberties are affected by any program. This requires transparency, which to date has not been forthcoming.

Many data-intensive programs in the federal government -- data mining or not -- have been obscured from the vision of the press, the public, and Congress. Often, these programs are hidden by thick jargon and inadequate disclosure.

This hearing, and your continued oversight, will help clear the fog. Proponents of these programs should make the case for them, forthrightly and openly.

In some cases, data-intensive programs have been obscured by direct claims to secrecy. These claims would deny the courts, Congress, and the public from determining whether they are fair and reasonable.

The secrecy claims suggest that these systems are poorly designed. It is well known that "security by obscurity" is a weak security practice. It amounts to hiding weaknesses, rather than repairing them, in the hopes that your attacker does not find them. Data intensive systems that

require secrecy to function -- that do not allow people to see the data used or review the algorithm -- are premised on security by obscurity.

These systems have weaknesses. We just do not know what they are. Because people on our side in the press, the public, Congress, and elsewhere cannot probe these systems and look for their flaws, they will tend to have more flaws than systems that are transparent, and subject to criticism and testing. We will not know when an attacker has discovered a flaw and is preparing to exploit it.

The best security systems are available for examination and testing -- by good people and bad people alike -- and they still work to secure. Locks on doors are a good, familiar example. Anyone can study locks and learn how to break them, yet they serve the purpose they are designed for, and we know enough not to use them for things they will not protect.

As long as we are unable to examine government data analysis systems the same way we examine locks and other security tools, these systems will not provide reliable security. But they will manifest an ongoing threat to privacy and civil liberties.

Conclusion

I have devoted my testimony to the question whether government data mining can work to discover terrorism. The security issues are paramount. I feel it clear that data mining does not work for this purpose.

Government data mining relies on access to large stores of data about Americans -- from federal government files, state public records, telecommunications company databases, from banks and payment processors, from health care providers, and so on. Predictive data mining, in particular, hungers for Americans' personal information because it uses data both in the development of patterns and in the search for those patterns.

There is a growing industry that collects consumer data for useful purposes like marketing and consumer credit. But this industry also appears to see the government as a lucrative customer. Most Americans are probably still unaware that a good deal of information about them in the data-stream of commerce may be used by their government to make decisions that coercively affect their lives, liberty, and property.

Here, again, the answer is transparency. Along with the transparency that will give this Committee the ability to do effective oversight into programs and practices, there should be transparency of the type that empowers individuals.

The data used in government data mining programs should be subject to the protections of the Privacy Act, no matter where the data is housed or by whom it is processed. Data in these programs cannot be exempted from the Privacy Act under national security or law enforcement exemptions without them treating all citizens like suspects.

The data sources should be made known, especially when data or analyses are provided to the government by private providers. This would allow the public to better understand where the information economy may work against their interests.

Many things must be done to capture the privacy implications of government data mining. This hearing provides an important first start by commencing a needed conversation on the issues. Transparency and much more examination of government
data mining is the first, most important step toward making sure that this information age practice is used to the maximum benefit of the American people.