Written Questions for Jim Harper
Hearing on "Balancing Privacy and Security:
The Privacy Implications of Government Data Mining Programs"
Submitted by Chairman Patrick Leahy
January 17, 2007

1. At last week's hearing, Senator Specter questioned whether the Government's use of data mining programs have stopped dangerous persons from entering the country, and prevented terrorist attacks. What is your evaluation of the efficacy of data mining programs? What evidence is there that the Government's use of data mining technology has been an effective tool for preventing terrorist attacks?

Given the obscurity and secrecy surrounding the many security programs being conducted in various government departments, it is difficult to determine what programs may use data mining, in what ways they may use it, and, of course, whether or not it is effective. In the paper I co-wrote, "Effective Counterterrorism and the Limited Role of Predictive Data Mining," my co-author and I made the case that predictive data mining is unlikely, as a statistical matter, to find terrorists. The burden of proof is on proponents of data mining for terrorist discovery that they can do this successfully and consistently with American law and values.

There is some evidence of data mining's effectiveness in this task, but it is not very good. First, there are theoretical arguments that data mining can be used to catch terrorists.

Another hearing witness, Mr. Kim Taipale, is a proponent of data mining for terrorism discovery who could have made the theoretical case to the committee. Instead of making that affirmative case — instead of telling the committee how data mining works and how it can work in this special case — he sought only to refute the arguments against data mining. Not having made the case, I do not believe he carried the burden of proof, and I do not know that anyone has.

Then there is anecdotal evidence. At the hearing, Senator Specter briefly cited the case of a Jordanian man turned away from the U.S. border in 2003, in part because of the Automated Targeting System (ATS). Eighteen months later, DHS officials say, his arm and hand were found handcuffed to the steering wheel of a car bomb that had been detonated in Baghdad.

This story is provocative and exciting — and it is evidence — but it is not good evidence that data mining can discover terrorists and prevent terror attacks. (He would not have had access to the support system that exists in Iraq for car bombings, of course, so it would be error to assume that a similar incident would have happened in the U.S. had he not been excluded from the country.)

First, it is unclear whether or not this man was turned away as a result of data mining, or any kind of pattern-based analysis or risk scoring. It may have been a review of records about him — also apparently a part of the ATS program — that caused him to get additional scrutiny. This is link analysis, and it has fewer privacy and civil liberties concerns (though many remain).

As Senator Specter noted, it is hard to find out the facts in cases like this. For example, the man was handcuffed to the car bomb, which suggests he may not have been a volunteer for the attack he was involved in. It is plausible that he may have driven the car bomb to secure the release of his family from terrorist captors, for example. This undercuts the suggestion that he had attempted to enter the U.S. with evil intent. Indeed, he may have been entering our country to escape the violence and terror of the Middle East. Selectively released anecdotes make this kind of speculation necessary.

Using this anecdote in a speech at the Center for Strategic and International Studies, however, Department of Homeland Security Assistant Secretary for Policy Stewart Baker said that a Customs and Border Patrol officer turned him away because he "wasn't confident that this guy was going to live up to the obligations that we imposed under [his] visa." This suggests that he was not flagged as a potential terrorist, but perhaps as an intending immigrant. If this man *was* a terrorist, it is not evidence of ATS' or data mining's effectiveness to point out that, during their use, we stopped someone thanks to serendipity.

Evidence of the utility of data mining for catching terrorists is weak. It is up to proponents to come forward with evidence that it can work for this purpose.

2. Some have questioned whether data mining programs have harmed residents of the United States. In your estimation, how have United States residents been harmed by data mining programs? Do the data mining programs implicate constitutional protections against unreasonable search and seizures?

Tangible "harm" is not necessarily the threshold for determining whether data mining programs might violate American law or values, which is the relevant question. Americans' Fourth Amendment right to freedom from unreasonable searches and seizures, for example, is not conditional on whether or not an unreasonable search caused harm. An action under 42 U.S.C. § 1983 — the leading legal tool Congress established to help enforce civil rights and liberties — is not conditional on harm or damages either.

There may be examples of people harmed in tangible ways, in that they are unable or less able to travel freely, for example. In the current environment of obscurity and secrecy, it is difficult to determine who may have been harmed, by what program, and in what ways, through data mining or other data analysis. Oversight from Congress and policies that grant full redress and due process to Americans affected by these types of programs will help expose the nature and scope of what harms may be done by any data mining.

It is important to recognize that programs using large amounts of personal information about Americans, uncontrolled and untested by sufficient congressional and public oversight, threaten future abuses of various kinds. It does not impugn the beneficent motives of most public servants to note that some do abuse their station and power. A large store of information about people in the hands of government, originally used for good purposes, stands as a threat to privacy and civil liberties nonetheless. This "cost" of data mining is not an immediate or tangible harm, but it is an equally important consideration for formulating public policy about data mining.

3. Without proper safeguards, I am concerned that data mining technology could be used to erode the bedrock Fourth Amendment principle of individualized suspicion in our criminal justice system, by permitting the Government to simply vacuum up large amounts of sensitive personal information about ordinary, law-abiding Americans, without first obtaining a warrant or establishing the legitimate need for this information.

A. Is the movement away from the principle of individualized suspicion constitutionally justifiable?

There is no justification, constitutional or otherwise, for moving away from the principle of individualized suspicion. Important a touchstone as individualized suspicion is, however, I do not believe that the Fourth Amendment restricts searches and seizures to only those based on individualized suspicion. It can be reasonable to investigate, briefly and unobtrusively, those about whom there is only general suspicion.

For example, where a law enforcement officer has learned that a woman with blond hair has just stolen a stuffed animal from a carnival booth, it may be reasonable within the next few minutes to ever-so-briefly detain and question any woman with blond hair near the scene of the crime. The sharpness of the suspicion will tend to control the scope of the search or seizure: Immediately handcuffing and frisking any blond woman in the vicinity would be unreasonable because the mere fact that it was a blond woman is not sufficiently precise to justify this level of search and seizure.

I think the concern you express in your prefatory remarks to this question goes to the complex tangle of issues created by data mining, and other data-intensive security programs, that rely on collection and maintenance of records about law-abiding Americans.

Typically, these systems must be cloaked in secrecy so that they cannot be reverse-engineered and defeated by wrongdoers. To hide the databases, they are typically exempted from the Privacy Act's protections under the law enforcement exception. This exception was created to prevent criminals from getting access to investigatory files about themselves and their cases. The result is that these systems treat all Americans like criminals, subject to general surveillance.

In my written testimony to the Committee, I discussed how these systems, premised as they are on "security by obscurity," are fundamentally flawed. The fact that they must treat each citizen as a suspect is one result of their poor design as security systems.

B. In his written testimony, Kim Taipale testified that data mining programs do not raise Fourth Amendment concerns. In particular, he testified that the particularity requirement of the Fourth Amendment does not impose a requirement of individualized suspicion before a search can be reasonable. Do you agree?

I do not read Mr. Taipale's testimony as denying the existence of Fourth Amendment concerns, though he gives them very short shrift. He does appear to argue against the case, made by someone, somewhere, that "pattern-matching does not satisfy the particularity requirements of the Fourth Amendment."

The particularity requirement of the Fourth Amendment goes to the scope and nature of a lawful warrant. (". . . and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and *particularly* describing the place to be searched, and the persons or things to be seized" (emphasis added).) This has at best a tangential relationship to the issues in data mining.

His argument appears to be that automated pattern analysis based on behavior or data profiles is not inherently unreasonable. This is true. The argument I put forth in my paper and my testimony is that predictive data mining will not work to catch terrorists. It flows from this that the investigation of a person based on a predictive-data-mining search for terrorists is unreasonable. This is not inherent to data mining, only to data mining for terrorists.

Questions of Senator Edward M. Kennedy
"Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs"
Submitted by Chairman Patrick Leahy
January 17, 2007

Questions for Jim Harper

1. In his written testimony, Kim Taipale critiqued the article you co-authored, "Effective Counterterrorism and the Limited Role of Predictive Data Mining," and your conclusion that data mining is not an effective tool for predicting and preventing terrorism. Please respond to Mr. Taipale.

Mr. Taipale's testimony leveled many criticisms at the paper Jeff Jonas and I wrote about the inutility of data mining to the problem of catching terrorists. Our paper has been well-received and persuasive, so it is not a surprise — and not wrong — that it should invite criticism.

Jeff Jonas and I did not pen any startling new insight in our paper. We relied on many other thinkers and authors, cited and uncited. Most likely, the paper was so well-received because we did our best to talk about this complicated subject in clear, natural language, and we stuck as much as we could to the best terminology for our subject.

On terminology, we wrote, "[D]iscussions of data mining have probably been hampered by lack of clarity about its meaning. Indeed, collective failure to get to the root of the term 'data mining' may have preserved disagreements among people who may be in substantial agreement."

We went on to define our terms, breaking data analysis into discrete sub-parts that are distinct in relevant ways. The result was a persuasive paper.

In his testimony, Mr. Taipale specifically declined to settle on common terminology, saying, "[F]urther parsing of definitions is unlikely to advance the debate[, so] let us simply assume instead that there is some form of data analysis based on using patterns and prediction that raises novel and challenging policy and privacy issues." His refusal to adopt any stable terminology made his testimony very difficult to follow, and sapped it of persuasiveness.

There may be weaknesses in my paper, my arguments, and my testimony — all of which should be explored — but Mr. Taipale did not expose them in any useful sense. Rather than attempt to respond point by point, I will merely reiterate what I said in my testimony: The burden of proof remains with the proponents of predictive data mining to show that it can work to catch terrorists.

2. What is the potential that data mining programs will target particular groups, such as Muslims and those of Arab descent? What safeguards—administrative and statutory—might be used to prevent discriminatory profiling?

Only a very, very badly designed data mining program would use national origin, ethnic background, or religion to seek after terrorists. Terrorists spring up in many nations, including countries well outside the Middle East. They can be from any ethnic group. They can be members of any religion, or no religion at all. There is so little correlation between these factors and terrorist activity that any data mining program that is designed with even a half-measure of care will not include them. Simple oversight and transparency will ensure against such things.

People who argue that there is a correlation between religion or ethnicity and terrorism, perhaps based on the September 11, 2001 attacks, are neither students of terrorism nor serious about securing the country against it.

3. In response to Chairman Leahy, you testified that you are not aware of any comprehensive, scientific study of the effectiveness of data mining as a tool in preventing terrorism. What would such a study entail?

As fields of study, information policy, information quality, data mining, and related areas are very immature. It will take several years, and perhaps decades, for there to be a reliable, organized way to study data mining for uses like the search for terrorists, which has high consequences when it elicits either false positives or true positives.

In the meantime, it is worthwhile to rely on common experience with data mining. In our paper, Jeff Jonas and I used the extensive experience that marketers have with data mining to reveal its inutility for catching terrorists.

Similar natural experiments exist in things like professional sports: Baseball and football teams study each other very carefully, using highly refined statistical methods. Yet, when they arrive on the field, neither team knows what the other will do on the first play, much less the tenth. When a football team is able to beat every one of its opponents using predictive data mining, data mining may be provably useful for catching terrorists — that is, until their techniques are revealed, inviting counter-measures.

Witness Questions

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

A Hearing before the
Senate Judiciary Committee
January 17, 2007
Questions of Senator Arlen Specter

Jim Harper

1. **In your testimony, you state that if the government seizes your person, house, papers, or effects because you have been made a suspect by data mining, that it raises Fourth Amendment concerns. However, it is my understanding that the vast bulk of the data analyzed by data mining technology is either information already in the possession of the government or information that an individual has relinquished to a third party. Do I have a reasonable expectation of privacy when I give information to Amazon or Google? I may hope that they keep that information private, but as a constitutional matter can I reasonably expect that information will be protected?**

There are two senses in which a search or seizure based on data mining raises Fourth Amendment concerns. The first, which your question focuses on, is whether the information used in data mining might be something in which people have a Fourth Amendment privacy interest.

There are two ways to respond to this question. One is to ask whether current Supreme Court doctrine allows it to be used under the "reasonable expectation of privacy" formulation. The Court's cases, the "third-party doctrine" in particular, are increasingly unsatisfactory. As I wrote in my testimony:

> [T]he 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.

The second approach to this question is to inquire about real Americans' actual expectations as to information that they create through, or entrust to, third parties. Last August, AOL publicly released 685,000 users' search queries. The uproar was immense and AOL immediately called the release a "screw-up" and apologized. The reason? AOL users expect the company to keep this information to itself, and this expectation is

widely held. If you are like most Americans, you have more than a vain "hope" that service providers will maintain information about you in confidence. You expect it, and you are being reasonable in doing so.

There is a second sense in which a search or seizure based on data mining raises Fourth Amendment concerns. If data mining is going to be used to focus investigative attention or direct suspicion at people, it must do so somewhat accurately and fairly. If data mining were used to help develop a predicate for searching a home or tapping a phone, for instance, this would implicate the protections of the Fourth Amendment. The data used in the operation must be sufficiently accurate, and the algorithm must be well drawn. Otherwise, the search or seizure will lack the reasonableness that is required by the Fourth Amendment. Due process (or "redress") requires that people should be able to explore these issues in their particular cases.

2. I agree with your argument that there should be some kind of redress process available for individuals who are negatively impacted when the government acts on information obtained using data mining technology. A newspaper story on this hearing discussed a pilot for a major U.S. airline who was stopped dozens of times after he returned from flying an overseas route because information obtained using data mining technology kept turning up his name. It seems to me someone like that should have some sort of redress. Can you provide other examples of when information from data mining is used that there may be a need for a redress process?

Given the obscurity and secrecy surrounding the many security programs being conducted in various government departments, it is difficult to determine what programs may use data mining, in what ways they may use it, whether or not they are effective, and what consequences they have for law-abiding Americans.

Rather than seeking after anecdotes — favoring data mining or opposing it — the Committee should ensure that all programs are designed consistent with constitutional law and values. That means that people negatively affected by a program have resort to "redress" — ultimately in a court of law — and that this redress allows them the ability to access information about the program, the information used in it, and whether the program was consistent with law.

It is worth mentioning that tangible "harm" is not necessarily the threshold for determining whether data mining programs might violate American law or values. Americans' Fourth Amendment right to freedom from unreasonable searches and seizures, for example, is not conditional on whether or not an unreasonable search caused harm. An action under 42 U.S.C. § 1983 — the leading legal tool Congress established to help enforce civil rights and liberties — is not conditional on harm or damages either.

3. Mr. Taipale has argued that data mining technology should only by used for investigative purposes – as a predicate for further screening or investigation. If that were the case, would that alleviate some of your concerns about using data mining technology?

The concerns I have raised about data mining are premised on its use for investigative purposes only. Predictive data mining does not produce evidence, and no one has ever plausibly argued that the information produced by data mining could be used as proof of guilt.

4. Can you please respond to what Mr. Taipale has said about applying multiple factors to identify patterns or relationships? He seems to be arguing that this would significantly reduce false positives when using data mining technology.

The question reveals how deeply Mr. Taipale has obscured his argument in jargon. He does seem to argue that some techniques would reduce false positives. I do not believe that my responses can make his case more clear. Here is how he makes it:

> [R]eal detection systems employ ensemble and multiple stage classifiers to carefully selected databases, with the results of each stage providing the predicate for the next. 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 of systems, reduces false positives, and significantly reduces data requirements. 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.
>
> 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. 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 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.
>
> 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. 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 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.

(footnotes omitted)

This dense discussion is very difficult to parse, but weaknesses in his argument may include the following:

- The use of "carefully selected databases" appears to mean that researchers are giving themselves a boost by granting themselves advance knowledge of which databases to look at. They would not have this advantage in looking for terrorists, unless they are strictly fighting the last battle.

- "[E]ntities with positive classifications" are made the subject of data collection and data access after the first pass. Investigation of people after this first pass may have the consequences for privacy and civil liberties that a "one pass" or "one factor" system.

- Using "true positives" to "reclaim false negatives" appears to be another form of cheating. In a real test, one would not know the true positives and thus would not be able to add back false negatives based on links to them. (The alternative interpretation of this jargon is that entities/suspects would be retained as suspects based on links to suspects. This, though, would likely make all entities suspects, under the Kevin Bacon principle.)

- The use of an "initial risk-adjusted population" (perhaps the same thing as "carefully selected databases") "where the frequency of true positives in the

selected group exceeds that in the background population" appears to be another example of cheating. Researchers cannot assume knowledge about what population terrorists are in then congratulate themselves for discovering what population terrorists are in.

- The need for "strong group linkages that can compensate for false negatives" is an important concession. Perhaps link analysis "saves" pattern-based data mining, but more likely, as Jeff Jonas and I argued in our paper, link analysis is what actually works.

- Another concession"[M]ulti-stage classification techniques can significantly reduce—*perhaps to acceptable levels*—the otherwise unacceptably large number of false positives . . . ."

With great confidence, Mr. Taipale claims all this *might* work. He asks for continued "research in real world detection systems" — meaning he wants taxpayer dollars spent on sifting through personal data about law-abiding citizens. But what he promises for all the costs in dollars and privacy is to "better allocate limited analytic resources." He cannot quite bring himself to say that all this will actually help catch terrorists.