S. HRG. 110–2

BALANCING PRIVACY AND SECURITY: THE PRIVACY IMPLICATIONS OF GOVERNMENT DATA MINING PROGRAMS

HEARING

BEFORE THE

COMMITTEE ON THE JUDICIARY
UNITED STATES SENATE

ONE HUNDRED TENTH CONGRESS

FIRST SESSION

JANUARY 10, 2007

Serial No. J–110–1

Printed for the use of the Committee on the Judiciary

U.S. GOVERNMENT PRINTING OFFICE

33–226 PDF WASHINGTON : 2007

For sale by the Superintendent of Documents, U.S. Government Printing Office
Internet: bookstore.gpo.gov Phone: toll free (866) 512–1800; DC area (202) 512–1800
Fax: (202) 512–2250 Mail: Stop SSOP, Washington, DC 20402–0001

COMMITTEE ON THE JUDICIARY

PATRICK J. LEAHY, Vermont, *Chairman*

EDWARD M. KENNEDY, Massachusetts	ARLEN SPECTER, Pennsylvania
JOSEPH R. BIDEN, JR., Delaware	ORRIN G. HATCH, Utah
HERB KOHL, Wisconsin	CHARLES E. GRASSLEY, Iowa
DIANNE FEINSTEIN, California	JON KYL, Arizona
RUSSELL D. FEINGOLD, Wisconsin	JEFF SESSIONS, Alabama
CHARLES E. SCHUMER, New York	LINDSEY O. GRAHAM, South Carolina
RICHARD J. DURBIN, Illinois	JOHN CORNYN, Texas
BENJAMIN L. CARDIN, Maryland	SAM BROWNBACK, Kansas
SHELDON WHITEHOUSE, Rhode Island	TOM COBURN, Oklahoma

BRUCE A. COHEN, *Chief Counsel and Staff Director*
MICHAEL O'NEILL, *Republican Chief Counsel and Staff Director*

CONTENTS

STATEMENTS OF COMMITTEE MEMBERS

WITNESSES

QUESTIONS AND ANSWERS

SUBMISSIONS FOR THE RECORD

Page

ADDITIONAL SUBMISSIONS FOR THE RECORD

BALANCING PRIVACY AND SECURITY: THE PRIVACY IMPLICATIONS OF GOVERNMENT DATA MINING PROGRAMS

WEDNESDAY, JANUARY 10, 2007

UNITED STATES SENATE,
COMMITTEE ON THE JUDICIARY,
Washington, DC

The Committee met, pursuant to notice, at 9:31 a.m., in room 226, Dirksen Senate Office Building, Hon. Patrick J. Leahy (chairman of the committee) presiding.

Also present: Senators Specter, Feingold, and Whitehouse.

OPENING STATEMENT OF HON. PATRICK J. LEAHY, A U.S. SENATOR FROM THE STATE OF VERMONT

Chairman LEAHY. The Judiciary Committee will be in order.

Today the Senate Judiciary Committee holds an important hearing on the privacy implications of government data mining programs. This committee has a special stewardship role in protecting our most cherished rights and liberties as Americans, including the right of privacy.

Today's hearing on government data mining programs is our first in the new Congress. This hearing is also the first of what I plan to be a series of hearings on privacy-related issues throughout this Congress.

The Bush administration has dramatically increased its use of data mining technology, namely the collection and monitoring of large volumes of sensitive personal data to identify patterns or relationships.

Indeed, in recent years the Federal Government's use of data mining technology has exploded, without congressional oversight or comprehensive privacy safeguards.

According to a May 2004 report by the Government Accountability Office, at least 52 different Federal agencies are currently using data mining technology. There are at least 199 different government data mining programs.

Think about that just for a moment. One hundred and ninety-nine different programs that are operating or are planned throughout the Federal Government. Of course, advances in technology make data mining and data banks far more powerful than ever before.

Now, these can be valuable tools in our national security arsenal, but I think the Congress has a duty to ensure that there are proper safeguards so they can be most effective.

One of the most common and controversial uses of this technology is to predict whom among our 300 million Americans are likely to be involved in terrorist activities.

According to GAO and a recent study by the CATO Institute, there are at least 14 different government data mining programs within the Departments of Defense, Justice, Homeland Security, and Health. That figure does not include the NSA's programs.

I think Congress is overdue in taking stock of the proliferation of these databases that are increasingly collecting information on Americans.

Now, they are billed, of course, as counterterrorism tools, but you wonder why there have to be so many, in so many different departments. But the overwhelming majority of them use, collect, and analyze personal information about ordinary American citizens.

We have just learned through the media that the Bush administration has used data mining technology secretly to compile files on the travel habits of millions of law-abiding Americans.

Incredibly, through the Department of Homeland Security's Automated Targeting System program, ATS, our government has been collecting information on Americans, just average Americans.

They then share this sensitive, personal information with foreign governments. They are shared with private employers. There is only one group they will not share it with: the American citizens they collected it on.

So if there is a mistake in there and you suddenly find you cannot get into another country, or a mistake in there and you find you do not get a promotion in your job because your employer has it, you never know why and you never even know what the mistake was.

Following years of denial, the Transportation Security Administration, TSA, has finally admitted that its controversial secure flight data mining program, which collects and analyzes airline passenger data obtained from commercial data brokers, violated Federal privacy laws by failing to give notice to U.S. air travelers that their personal data was being collected for government use. I think you find out why they denied they were doing it: because they were breaking the law in doing it.

Last month, the Washington Post reported that the Department of Justice will expand its one-DOJ program, a massive database that would allow State and local law enforcement officials to review and search millions of sensitive criminal files, following the FBI, DEA, and other Federal law enforcement agencies.

That means sensitive information about thousands of individuals, including thousands who have never been charged with a crime, will be available to your local law enforcement agencies no matter what their own system of protection of that data might be.

So you have to have proper safeguards and oversight of these, and other, government data programs, otherwise the American people do not have the assurance that these massive databases are going to make them safer, nor the confidence their privacy rights will be protected.

And, of course, there are some very legitimate questions about whether these data mining programs actually do make us safer. It

becomes almost humorous. Some of the consequences, I have talked about before.

Senator Kennedy has been stopped 10 times going on a plane, a flight he has been taking for 40 years back to Boston, because somehow his name got, by mistake, on one of these databases.

We had a 1-year-old child who was stopped because their name was on as a terrorist. The parents had to go and get a passport to prove this 1-year-old was not really a 44-year-old terrorist.

So the CATO Institute study found that data mining is not an effective tool for predicting or combatting terrorism because of the high risk of false positive results.

We need look no further than the government's own terrorist watch list, which now contains the names of more than 300,000 individuals, including, as I said, Members of Congress, infants, and Catholic nuns, to understand the inefficiencies that can result in data mining and government dragnets.

So let us find out how we can make ourselves safer, but not make ourselves the object of a mistake and ruin our lives that way.

I am joined today by Senator Feingold, Senator Sununu, and others in a bipartisan attempt to provide congressional oversight. We are reintroducing the Federal Agency Data Mining Reporting Act, which we have supported since 2003. It would require Federal agencies to report to Congress about their data mining programs.

We in Congress have to make sure that our government uses technology to detect and deter illegal activity, but do it in a way that protects our basic rights.

I also might say, on a personal note, I want to thank Chairman Specter for scheduling this hearing at my request. At the beginning of every Congress we have to do various reorganizational things, and I understand this is to be completed today or early tomorrow, and allowing me to be Chairman, even though I am not, technically, yet.

So, Chairman Specter, it is up to you. You do whatever you want to do.

STATEMENT OF HON. ARLEN SPECTER, A U.S. SENATOR FROM THE STATE OF PENNSYLVANIA

Senator SPECTER. Well, thank you very much. I hope you will not mind if I address you as "Mr. Chairman", Mr. Chairman.

Chairman LEAHY. I can put up with it.

[Laughter].

Senator SPECTER. The 109th Congress was very productive for the Judiciary Committee because of the close cooperation which Senator Leahy and I have had, which goes back to a period before we were Senators.

The National District Attorneys' Conference was held in 1970 in Philadelphia when I was District Attorney, and District Attorney Leahy from Burlington, Vermont attended. We formed a partnership which has lasted and withstood partisan pressures in Washington, DC.

When Chairman Leahy refers to my scheduling of a hearing at his request, I think there were a number of hearings which were at Senator Leahy's request when he was only Senator Leahy and

not Chairman Leahy. We had a very close, coordinated relationship and I am sure that will continue.

Senator Harkin and I have passed the gavel for many years in the Subcommittee on Appropriations, and we call it a seamless transfer. This is our first transfer of the gavel between Chairman Leahy and myself, and I am looking forward to a seamless operation.

In fact, Senator Leahy and I coordinated with the introduction of the Personal Data Privacy and Security Act of 2005, which we reported out of committee and have coordinated with the Commerce Committee, which dealt with identity theft significantly, but also with data mining.

There are some very important issues which are raised in the collation of all this material. The presence of the material in so many contexts led the Supreme Court to observe, in the case of *U.S. Department of Justice* v. *Reporters' Committee for Freedom of the Press*, that when information is located in so many spots, it is a matter of "practical obscurity", but when it is all brought together, it is a different matter.

The committee focused on one aspect of this last year when we were looking at the telephone company responses to the government's request for collection of data. There may be very important law enforcement activities which utilized this data appropriately, but it is a balancing test of what kind of privacy was invaded, and what is the benefit for law enforcement, what is the benefit for society.

I want to start my tenure as the non-chairman by observing the time limit, so I yield back a balance of 20 seconds. Thank you, Mr. Chairman.

Chairman LEAHY. I thank Chairman Specter. We have tried to work together. We have worked together ever since Senator Specter came here in 1986.

Senator SPECTER. 1980.

Chairman LEAHY. 1980. I am sorry. Time goes by when you are having fun. And we did know each other as former prosecutors. We worked closely together. We have been on the Appropriations Committee together and worked together, and on this committee.

I think we lowered the level of partisanship in this committee during the past 2 years, and I hope to continue that. I am hoping that we are going to reach a point where things can work the way the Senate should.

I do note that Senator Feingold of Wisconsin is here. He is, as I mentioned, the lead sponsor on this bill. I would yield to Senator Feingold if he wished to say anything.

STATEMENT OF HON. RUSSELL D. FEINGOLD, A U.S. SENATOR FROM THE STATE OF WISCONSIN

Senator FEINGOLD. Thank you, Mr. Chairman, and thank the Ranking Member. It is a pleasure working with you in the different capacities, and I look forward to working with both of you again on this committee.

Thanks for holding this hearing. It raises important policy questions about the capabilities of data mining technologies and the privacy and civil liberties implications for ordinary Americans if this

type of technology were to be deployed. These are questions that Congress has to address.

This hearing is a critical first step in the process of understanding, evaluating, and perhaps regulating this type of technology. Many Americans are understandably concerned about the specter of secret government programs analyzing vast quantities of public and private data about the every-day pursuits of mostly innocent people in search of patterns of suspicious activity.

So let me start by reiterating a point that Senator Wyden and I made in a recent letter to Director of National Intelligence Negroponte. Obviously, protecting our national security secrets is essential and the intelligence community would not be doing its job if it did not take advantage of new technologies.

But when it comes to data mining, we must be able to have a public discussion, what one of our witnesses has called a national conversation, about its potential efficacy and privacy implications before our government deploys it domestically.

We can have that public debate about these policy issues without revealing sensitive information that the government has developed. The witnesses here today have for years been debating a variety of issues related to data mining.

It is time to get Congress and the executive branch into that discussion, not just in reaction to the latest news story, which has sort of been the position we have been in in the past, but in a proactive, thoughtful, and collaborative way.

As I have said before, this hearing is an important first step. I hope that the next step will be the enactment of the Federal Data Mining Reporting Act, which I am reintroducing today along with Senator Sununu, Senator Leahy, and others. I thank the Chairman for mentioning it, and for his excellent support of the bill.

The bill requires Federal agencies to report on their development and use of data mining technologies to discover predictive or anomalous patterns indicating criminal or terrorist activity, the types of data analysis that raise the most serious privacy concerns. It would, of course, allow classified information to be provided to Congress separately under appropriate security measures.

Along with this hearing, I hope these reports will help Congress, and to the degree appropriate the public, finally understand what is going on behind the closed doors of the executive branch so we can start to have the policy discussion about data mining that is long overdue. I would urge my colleagues to support the legislation.

Mr. Chairman, I also want to note that last night I received a response from the Director of National Intelligence Negraponte to the letter Senator Wyden and I wrote to him regarding the Tangram Data Mining Program.

In it, ODNI states that Tangram is a research project, and acknowledged that it has "a real risk of failure." It also assured us that no Tangram tools would be deployed without consultation with the DNI's Civil Liberties and Privacy Officer.

I would just add that I would hope that Congress also would be consulted prior to any deployment of the Tangram data mining tool. So, I do thank you, Mr. Chairman, very much for the opportunity to make this opening statement.

Chairman LEAHY. Thank you.

Would the panel please rise and raise your right hand?

[Whereupon, the panel was duly sworn.]

Chairman LEAHY. Following our normal procedure—and I am sure you understand this, Mr. Harper—we have a former Member of Congress and we will recognize him first. Bob Barr represented the Seventh District of Georgia in the U.S. House of Representatives from 1995 to 2003. He was on the Judiciary Committee. He was Vice Chairman of the Government Reform Committee and a member of the Committee on Financial Services.

He occupies the 21st Century Liberties Chair for Freedom and Privacy at the American Conservative Union; serves as a board member of the National Rifle Association; is chairman of Patriots to Restore Checks and Balances; provides advice to several organizations, including—this is interesting—consulting on privacy issues with the ACLU, serving as a chair for youth leadership training at the Leadership Institute in Arlington, Virginia; and is a member of the Constitution Project's Initiative on Liberty and Security based at Georgetown University's Public Policy Institute.

The Congressman served as a member of the Long-Term Strategy Project for Preserving Security and Democratic Norms in the War on Terrorism at the Kennedy School of Government at Harvard University from 2000 to 2005. He was a New York Times columnist, and a close personal friend of mine, Mr. Safire, has called him "Mr. Privacy".

So with all that, Bob, go ahead.

STATEMENT OF ROBERT BARR, CHAIRMAN, PATRIOTS TO RESTORE CHECKS AND BALANCES, WASHINGTON, D.C.

Mr. BARR. Thank you very much, Mr. Chairman. Let me add my personal congratulations to the many I know you have received since your ascendancy to the chairmanship.

Let me also congratulate the fine work that Senator Specter has been involved in in laying the groundwork for the work that I know is coming this Congress with regard to the fundamental right to privacy and other civil liberties, particularly vis-a-vis fighting against acts of terrorism.

I very much appreciate both the former chairman and the current chairman inviting me today to this very important hearing.

I appreciate very much the attendance of at least two other Senators at this time whose presence here today obviously indicates a keen interest on their part in the issues before this committee, Senator Whitehouse and Senator Feingold, who has been a leader in the last Congress, and even before that.

I very much appreciate the committee indicating, I think very clearly, to the American people and to your colleagues here in the Congress that the issue of privacy, particularly as it relates to government data mining and the secrecy surrounding that and the extent thereof, is a top A–1 priority. I think that sends a very important message.

Of course, mindful of the committee's many responsibilities, I would ask that my prepared testimony be included in full in the record.

Chairman LEAHY. It will.

[The prepared statement of Mr. Barr appears as a submission for the record.]

Mr. BARR. What I would like to do, simply, in addition to that, is indicate to the committee, I think that a very appropriate starting point, or at least one of the starting points for the 110th Congress' long-term discussion of these issues, looking at and laying the groundwork for particular pieces of legislation, such as that which the committee has indicated will be introduced today.

I think it is important also to focus on some fundamental questions which have given rise because of the extensive secret data mining by the government and by private industry in conjunction with the government to a culture of suspicion in our society.

Perhaps one of the most fundamental issues, the most fundamental questions that really needs to be addressed, is who owns all of this data, this private data, this private, personal information that is the subject of all of this data mining?

The extent of the data mining, Mr. Chairman, you indicated is the tip of the iceberg. There have been recent disclosures that there are at least some 200 different data mining systems in the government.

You can hardly pick up the paper any day or watch the news any day and yet not walk away with new revelations about new data mining, whatever agency of the government it is, not just the Department of Justice, the Department of Defense, CDC, HUD, Homeland Security, Social Security Administration, IRS, SBA. They all seem to be enamored of, and have this blind interest and faith, in data mining.

The problem is, there has never been a comprehensive look at who owns this data. The fact that over the last several years the administration has been treating that data as its own—that is, information on private citizens—begins us down that slippery slope.

That slippery slope, we are all aware now, leads not only to secret data mining, which includes very personal data on American citizens and others in this country who have rights equal to those of our citizens under the Bill of Rights, First Amendment, Second Amendment, Fourth Amendment, and Fifth Amendment, being maintained in these government databases with no knowledge thereof, with no way to correct errors or improper information.

But it also leads us down that slippery slope to where we now see this administration, and that is viewing private mail that Americans and others have sent through the U.S. Postal Service.

If, in fact, the government can continue to believe or view this data that is the subject of data mining as its own, that it owns it, then everything else that it wants to do follows from that false premise.

Certainly, they can read people's mail, they can read people's e-mails. I think that is really a fundamental question that the committee must look at. There are others on which I would be glad to provide whatever information I have in terms of questions and follow-up.

But I really do think there are fundamental issues regarding the ownership of that data and the extent to which the government already, and should be, engaged in that that provide more than fertile ground for this committee to look into.

Chairman LEAHY. Thank you, Congressman. In fact, those will be among the questions that will be asked of the Attorney General when he comes here next week, the mail opening one. More and more, we hear about these things only because we read about it in the press, and this creates a strong concern for me.

Jim Harper is the Director of Information Policy Studies at the CATO Institute. As Director of Information Policy Studies, he focuses on the difficult problem of adopting law and policy to the unique situation of the information age. He is a member of the Department of Homeland Security's Data Privacy Integrity Advisory Committee.

His work has been cited by USA Today, Associated Press, and Reuter's. He has appeared on Fox News channel, CBS, and MSNBC, and other media. His scholarly articles appear in the Administrative Law Review, the Minnesota Law Review, and the Hastings Constitutional Law Quarterly.

He wrote the book, Identity Crises: How Identification is Overused and Misunderstood. He is the editor of privasilla.org, a web-based think tank devoted exclusively to privacy. He maintains the online Federal spending resource, washingtonwatch.com. He holds a J.D. from Hastings College of Law.

Mr. Harper, it is yours. Again, I apologize. We have to ask you to keep the statement brief—your whole statement will be part of the record—because we want to ask questions.

I should also note that Senator Whitehouse of Rhode Island has joined us here, not only today for the hearing, but Senator Whitehouse is a former attorney general. I had asked him, before he knew all the work that goes on in this committee, if he would join the committee. In a moment of weakness, he said yes. Senator, I am glad to have you here.

Senator Whitehouse. I am glad to be with you, Mr. Chairman. Delighted to be with the Ranking Member. And it was no moment of weakness.

Chairman LEAHY. Thank you.

Mr. Harper?

STATEMENT OF JIM HARPER, DIRECTOR OF INFORMATION POLICY STUDIES, CATO INSTITUTE, WASHINGTON, DC

Mr. HARPER. Thank you, Mr. Chairman.

If I can briefly start with a personal note that extends my biography just a little bit, my first job here on Capitol Hill was working for Senator Biden during the period when he was Chairman of this committee. I was an intern at the time.

It inspired my legal career, including my focus on constitutional law. My first paid job when I returned to the Hill after that was with Senator Hatch as a legal fellow on this committee. So I really appreciate being here before you.

Chairman LEAHY. You covered both sides of the aisle very well.

Mr. HARPER. In the spirit of bipartisanship. This committee has influenced my life and career a great deal and I hope that, in a small way, I will be able to influence you today.

The questions about data mining are complicated. Questions about privacy are complicated. When you combine the two, you have a very complex set of issues to deal with.

So we will obviously start to sort them out, but I think the conversation that you are starting with this hearing and with the oversight you intend to do this year in this Congress is very important.

My resort is to a document that we produced in the Department of Homeland Security Data Privacy Committee, where we created a structure, a framework for thinking about problems like this.

The first step in that framework is to ask how a program or technology serves a homeland security purpose. What risk does it address and how well does it address that risk? Once you determine that, you can make decisions about privacy and decide whether you want to use this technology, and how you want to use it.

I think in the area of data mining we have not gotten past that step yet. What is the theoretical explanation for how data mining can catch terrorists, is the major question that is before us.

The positive case for the use of data mining in this particular area has not yet been made, so I suppose that my colleague, Jeff Jonas, and I laid down something of a marker when we issued our paper on the dis-utility of data mining for the purpose of finding terrorists.

We argue that what we call "predictive data mining", that is, finding a pattern in data and then seeking that pattern again in data sets, predictive data mining, cannot catch terrorists.

Data mining can give a lift. There are many good uses to data mining. It can give a lift to researchers, their study of people, of scientific phenomena. But with the absence of terrorism patterns on which to develop a model, you're going to have a very hard time finding terrorists in data.

The result will be that you will get a lot of false positives. That is, you will find that many people who are not terrorists are suspects. You will waste a lot of resources going after these people. You will follow a lot of dead ends. And, very importantly, you will threaten the privacy and civil liberties of innocent, law-abiding Americans.

Now, I personally think that this applies equally well to developing patterns to search for through red-teaming and in searching for anomalies, though this was not the subject of our paper.

I think it is important to recognize this is not an indictment of data mining in toto. There are many data mining programs that may not even use personal information.

There are data mining programs that use personal information that may successfully ferret out fraud, for example, in health care payments or areas like that, so it is important to be clear about where data mining does not work and where it certainly may work.

I think the proponents of data mining need to make that affirmative case. It is not enough to attack nominal opponents of data mining. The affirmative case, again, has to be made.

You on this committee should be able to say to yourselves, oh, yes, I get it. I understand how data mining works. Then the country will be ready to accept data mining as a law enforcement or national security tool.

Once the benefits of data mining are understood and clear, then you can consider the privacy and other costs. Certainly there are

dollar costs, as there are with any program, and a lot of dollars are going into data mining at this point.

But the privacy costs, which I have articulated, or attempted to articulate, in my paper include the lack of control that people have over personal information about themselves, the questions of fairness, of liberty, and data security.

In this committee, we have referred to some of these things as due process, or the Fourth Amendment right to be free of unreasonable search and seizure, and equal protection. So the thing that I think we need, and the thing that I think we are seeing in the bill that is being introduced today—and I am quite happy about that—is transparency.

Transparency should be seen as an opportunity for the proponents of data mining to make their case, to make the affirmative case for data mining. We need to see how it works, where it is being used, what data is being used, what assures that the data is of high quality, and so on and so forth.

You will run into the problem of secrecy, that is, secrecy being put forward as a reason why not to share this information with you, why not to explain data mining to you. But I think you will have to address that at the right point, and I hope you will.

Thanks very much for the opportunity to present to you today.

Chairman LEAHY. Thank you, Mr. Harper.

[The prepared statement of Mr. Harper appears as a submission for the record.]

Chairman LEAHY. Leslie Harris is the Executive Director for the Center for Democracy and Technology. She joined CDT in the fall of 2005, and became Executive Director at the beginning of 2006. She brings over two decades of experience to CDT as a civil liberties lawyer, a lobbyist, and public policy strategist.

Her areas of expertise include free expression, privacy, and intellectual property. Prior to joining CDT, Ms. Harris was Founder and President of Leslie Harris & Associates, a public interest, public policy, and strategic services firm, representing both corporate and nonprofit clients before Congress and the executive branch on a broad range of Internet- and technology- related issues, including intellectual property, online privacy, telecommunications, and Spectrum.

During that time she was involved in the enactment of many landmark pieces of legislation, including the landmark e-rate amendment to the 1996 Telecommunications Act, the Children's Online Privacy Protection Act, and the 2002 Technology, Education, and Copyright Harmonization Act, or the TEACH Act, which updated copyright law for digital distance learning. I would note that Ms. Harris has appeared before this committee many times, and I appreciate that.

Please go ahead.

STATEMENT OF LESLIE HARRIS, EXECUTIVE DIRECTOR, CENTER FOR DEMOCRACY AND TECHNOLOGY, WASHINGTON, DC

Ms. HARRIS. Thank you so much, Mr. Chairman. I appreciate the opportunity to be here. I want to applaud the Chairman, in particular, for making this data mining question, and privacy in general, a first order of business for this committee.

From the perspective of CDT, we believe that information technology ought to be used to better share and analyze the oceans of information that the government has in the digital age, but both national security and civil liberties require that technology only be used when there is a demonstrable, effective impact, and then only within a framework of accountability, oversight, and most importantly, protection of individual rights.

Data mining, in the abstract, is neither good nor bad, but as Jim Harper has pointed out, there is very little evidence of the effectiveness of at least the protective or patterned data mining. Yet, frankly, the executive branch is bewitched with this technology.

Unless and until a particular data mining technology can be shown to be an effect tool for counterterrorism and appropriate safeguards are in place to protect the privacy and due process rights of Americans, Congress should simply not permit the executive branch to deploy pattern-based data mining tools for any terrorism purposes.

Mr. Chairman, for some time you have sounded the alarm about how the legal context for data collection and analysis has been far outstripped by technology; at the very time that the legal standards for government access to data have been lowered and legal safeguards like the Privacy Act have been bypassed and the Fourth Amendment requirements for probable cause, particularity, and notice have been thrown into doubt, we are moving into this very sophisticated and troubling data mining era.

The impact of this perfect storm of technological innovation, growing government power, and outdated legal protections is well illustrated by the revelation last month that the Automatic Targeting System, which is designed to screen cargo, is now being used to conduct risk assessments on individuals. Those risk assessments, as I read this Privacy Act notice, can be used for a wide variety of uses wholly unrelated to border security.

There is much Congress can do. The first step, of course, is to pierce this veil of secrecy. We strongly endorse the legislation that you, Senators Feingold, Sununu, and others have introduced today. We need vigorous oversight. We need transparency. Ultimately, we need legislation. We cannot do any of that until we are able to get a handle on what is going on.

We believe that Congress ought to go further and not permit any particular data mining applications to be deployed until there is a demonstration of effectiveness. We believe research should continue, but in terms of deploying these technologies, we do not even have to reach the privacy questions until we know whether or not they are working.

While it is the job of the executive branch, in the first instance, to develop serious guidelines for the deployment of data mining for data sharing and analysis, we do not believe that job has been adequately done.

If necessary, this body needs to impose those guidelines. There is much in the Markle recommendations and others to guide you in that regard.

Finally, we have to get our arms around how commercial databases are being used for data mining. Those activites fall entirely outside of the Privacy Act and all other rules.

Last year, Mr. Chairman, Mr. Specter, you introduced the Personal Data Privacy and Security Act. That bill included important to ensure that government use of commercial data bases for data mining was brought under the Privacy Act. We ought to enact that bill and we ought to enact some other protections as well.

I appreciate the opportunity to testify, and am ready for your questions.

Chairman LEAHY. Thank you.

[The prepared statement of Ms. Harris appears as a submission for the record.]

Chairman LEAHY. Our next witness is Kim Taipale. Now, have I pronounced it right?

Mr. TAIPALE. Close enough.

Chairman LEAHY. How do you pronounce it?

Mr. TAIPALE. Taipale.

Chairman LEAHY. Taipale. Mr. Taipale is the Founder and Executive Director of the Center for Advanced Studies in Science and Technology Policy. It is a private, nonpartisan research and advisory organization focused on information technology and global and national security policy.

He is a Senior Fellow at the World Policy Institute, where he serves as Director of the Global Information Society Project, and the Program on Law Enforcement and National Security in the Information Age. He is an Adjunct Professor of Law at New York Law School, where he teaches cyber crime, cyber terrorism, and digital law enforcement.

He serves on the Markle Task Force on National Security in the Information Age, the Science and Engineering for National Security Advisory Board of The Heritage Foundation, the Lexis-Nexis Information Policy Forum, and the Steering Committee of the American Law Institute's Digital Information Privacy Project.

Thank you for joining us here today.

STATEMENT OF KIM TAIPALE, FOUNDER AND EXECUTIVE DIRECTOR, CENTER FOR ADVANCED STUDIES IN SCIENCE AND TECHNOLOGY POLICY, NEW YORK CITY, NEW YORK

Mr. TAIPALE. Thank you, Mr. Chairman. Mr. Chairman, Senator Specter, members of the committee, thank you for the opportunity to testify today on the implications of government data mining.

Data mining technology has raised significant policy and privacy issues, and we have heard a lot of them today. I agree with all of those. But the discussion about data mining suffers from a lot of misunderstandings that have led to a presentation of a false dichotomy, that is, that there is a choice between security and privacy.

My testimony today is founded on several beliefs. First, that privacy and security are not dichotomous rivals, but dual obligations that must be reconciled in a free society. Second, we face a future of more data and more powerful tools, and those tools will be widely available.

Therefore, third, political strategies premised on outlawing particular technologies or techniques are doomed to failure and will result in little security and brittle privacy protection.

Fourth, there is no silver bullet. Everybody is right here. Data mining technologies alone cannot provide security. However, if they

are properly employed they can improve intelligence gain and they can help better allocate intelligence and security resources. If they are properly designed, I believe they can still do that while protecting privacy.

Before getting to my two main points, there are also some general policy principles that I think should govern the use of any of these technologies if they are implemented.

First, they should be used only for investigative purposes. That is, as a predicate for further investigation, not for proof of guilt or to otherwise automatically trigger significant adverse consequences.

Second, any programmatic implementations should be subject to strict oversight and review, both congressional and, to the extent appropriate, judicial review, consistant with existing notions of due process.

Third, specific technology features and architectures should be developed that help enforce these policy rules, protect privacy, and ensure accountability. So let me just make two main points.

The first, is a definitional problem. What is data mining? Data mining is widely misunderstood, but just defining it better is not the solution. If we are talking about some undirected massive computer searching through huge databases of every individual's private information and intimate secrets, and the result of a positive match is that you face a firing squad, I think we will all agree that we are opposed to that.

If, on the other hand, we are talking about uncovering evidence of organizational links among unknown conspirators from within legally collected intelligence databases in order to focus additional analytical resources on those targets, I think we will all agree that we are for it. The question is, can we draw a line between those two?

I doubt it if we start by focusing only on trying to define data mining. That is precisely the mistake that detracts us from the issues we should be focused on, some of which were actually raised in your opening statements. Drawing some false dichotomy between subject-based and pattern-based analysis is sophistry, both technical- and policy-wise.

The privacy issue in a database society, or to put it the other way around, the reasonableness of government access to data or use of any particular data, can only be determined through a complex calculus that includes looking at the due process of a system, the relationships between the particular privacy intrusion and security gain, and the threat level. They simply cannot be judged in isolation.

Even privacy concerns, themselves, are a function of scope, sensitivity of the data, and method: how much data, how sensitive is the data, and how specific is the query? But we really need to separate the access question and the decision-making question—on either side—from the data mining question itself and the use of data mining tools.

More importantly, even the privacy concerns cannot be considered away from due process. Due process is a function of predicate: alternatives, consequences, and error correction.

A lot of predicate and you can tolerate severe consequences even in a free society, but even ambiguous predicate maybe all right if there are minor consequences and there is robust error correction and oversight.

While we are on predicate we should note that there is no blanket prohibition against probablistic predicates, such as using predicate patterns. We do it all the time. Nor is there a requirement for non-individualized suspicion, such as using pattern mining.

My point is not that there are no privacy concerns, only that focusing only on data mining, however you define it, is not terribly useful. It really needs to be looked at more broadly. It is basically the computational automation of the intelligence function as a productivity tool that, when properly employed, can increase human analytical capacity and make better use of limited security resources.

My second and final point, is that you cannot look at data mining in this context through the "it won't work" lens and simply dismiss potential. First, the popular arguments about why it will not work for counterterrorism are simply wrong.

As I explain in my written testimony, the commercial analogy is irrelevant, the training set problem is a red herring, and the false positive problem can be significantly reduced by using appropriate architectures. In any case, it is not unique to data mining. It is fundamental to the intelligence function. The intelligence function deals with uncertainties and ambiguities.

Second, you cannot burden technology development with proving efficacy before the fact. We need R&D and we need real-world implementations and experience, done correctly with oversight, so we can correct errors.

Third, you cannot require perfection. To paraphrase Voltaire, the perfect ought to not be the enemy of the better.

Finally, you need to bear in mind that any human and technological process will fail under some conditions. Some innocent people will be burdened in any preemptive approach to terrorism and, unfortunately, some bad guys will get through. That is reality.

The question is, can we use these data mining tools and improve intelligence analysis and help better allocate security resources on the basis of risk and threat management?

I think we can, and still protect privacy, but only if policy and system designers take the potential for errors into account during development and control for them in deployment.

Chairman LEAHY. Thank you.

[The prepared statement of Mr. Taipale appears as a submission for the record.]

Chairman LEAHY. I would note that a number of the Senators have expressed a great deal of interest in this subject, both on the Republican side and the Democratic side. They are not here this morning simply because we have several major committees meeting at the same time.

One of the problems with the Senate, is you cannot be in more than one place at a time. Senator Feingold, for exmaple, is at the Foreign Relations Committee, and several other Senators have mentioned they wanted to be here.

Dr. Carafano, our next witness, is the Assistant Director for the Kathryn and Shelby Cullom Davis Institute for International Studies. He is a Senior Research Fellow at the Douglas and Sarah Allison Center for Foreign Policy Studies. Dr. Carafano is one of The Heritage Foundation's leading scholars on defense affairs, military operations and strategy, and homeland security.

His research focuses on developing the national security that the Nation needs to secure the long-term interests of the United States, realizing as we all do that terrorism is going to face us for the rest of our lifetimes, and how you protect our citizens and provide for economic growth and preserve civil liberties.

He is an accomplished historian and teacher. He was an Assistant Professor at the U.S. Military Academy at West Point, served as Director of Military Studies at the Army's Center of Military History, taught at Mt. Saint Mary College in New York, served as a Fleet Professor at the U.S. Naval War College. He is a Visiting Professor at the National Defense University at Georgetown University.

I do not want anybody to think that we have this large proliferation of people connected with Georgetown just because I went to Georgetown Law School; it is purely coincidence.

Dr. Carafano, go ahead.

STATEMENT OF JAMES JAY CARAFANO, HERITAGE FOUNDATION, ASSISTANT DIRECTOR, KATHRYN AND SHELBY CULLOM DAVIS INSTITUTE FOR INTERNATIONAL STUDIES, SENIOR RESEARCH FELLOW, DOUGLAS AND SARAH ALLISON CENTER FOR FOREIGN POLICY STUDIES, WASHINGTON, DC

Mr. CARAFANO. Thank you, Mr. Chairman. I also got my Ph.D. from Georgetown.

[Laughter].

I have submitted my statement for the record.

Mr. CARAFANO. I would like to do three things, very quickly: place the issue in context, state what I really think the problem is, and then argue why it is really essential that Congress address the issue and solve it.

First of all, I come at this not as a lawyer, because I am not a lawyer, but as an historian and strategist. One of the fundamentals of good, long war strategy for competing well over the long term is that you have to have security and the preservation of civil liberties, as well as maintaining civil society.

It is not a question of balance. You simply have to do both over the long term. I think there is no issue or no security tool in which this issue is more important than the one we are discussing today.

The problem is simply this. In the good old days when we were kids, technology evolved fairly slowly and policy could always keep up. We could look, we could observe, we could correct—trial and error.

But the fact is, today technologies evolve far more quickly than policies can be developed. Information proliferates, capabilities proliferate, and if the technology evolution has to stop for the policy to catch up, it is never going to happen.

In fact, it will not stop. You cannot stop it. So what you have to do is take a principled approach. You have to have a set of funda-

mental principles at the front end as guidelines to guide the development and implementation of the technology.

Among these, we have argued—some Kim already mentioned—are a clear definition of what data mining really is, addressing the requirements for efficacy, addressing the requirements for the protections, putting in appropriate checks and balances, and most importantly and often forgotten, is addressing the issue of the requirement for human capital and programming investments to actually implement these programs correctly.

The third point that I will make very quickly, is why is this really so important? There are really two aspects to that. The first, is we do not have infinite resources. What we need to do is focus our information and intelligence and law enforcement resources where they are going to do the most good.

And while it is absolutely important that any system protect the rights of everyone, we should also have systems that inconvenience as few people as possible. That is part of keeping a free, open, and healthy civil society. So we should be looking for systems which are directing on us on where we most live.

I would argue, for example, that programs like the Container Security Initiative and the Automated Targeting System—which, by the way, I think you could argue are not data mining systems—are good examples of where we try to focus scarce resources on things that might be problematic. Contrast that, for example, with the bill passed yesterday in the House, which argues that we should strip-search every container and package that comes into the United States (where you look at everything), or the lines that we have at TSA, which look at grandmothers and people coming through absolutely equally.

So we want systems that are going to focus our assets, where we inconvenience the least amount of citizens, friends, and allies of the United States, and we want to use our law enforcement efforts to best effect.

If we can create reporting requirements and a set of principles at the front end that guide the administration in doing that and adapting these new technologies, I think it will be time well spent by the Congress.

Thank you, Mr. Chairman.

[The prepared statement of Mr. Carafano appears as a submission for the record.]

Chairman LEAHY. Thank you. I am going to come back to this question of which things work best, because we are talking about millions of dollars—perhaps billions of dollars—being spent. I worry about a shotgun approach as compared to a rifle approach where you might actually pick what works.

When I see 90-year-old people in walkers take their shoes off to go onto an airplane and then not physically able to even put the shoes back on, I am curious just what happens.

I have been worried about the lack of privacy safeguards. In early 2003, I wrote to former Attorney General Ashcroft to inquire about the data mining operations, practices, and policies within the Department of Justice.

I would ask that a copy of my January 10, 2003 letter be made a part of the record. I would love to be able to put a response in the record too, but of course I never got one.

In 2003, I joined Senator Wyden in a bipartisan coalition of Senators in offering an amendment to the omnibus appropriations bill that ended the funding for the controversial TIA, Total Information Awareness, program because there were no safeguards.

In April of that year I joined with Senator Feingold in introducing the Federal Data Mining Reporting Act, which required all Federal agencies to report back to Congress on their data mining programs in connection with terrorism and law enforcement efforts, and a version of our measure was put on the Department of Homeland Security appropriations bill.

But basically the administration has ignored a lot of the bans that Congress, in a bipartisan way, has put on these things. Just last month, Representative Martin Sabo, one of the leaders in enacting the legal prohibition on developing and testing data mining programs, told the Washington Post that the law clearly prohibits the testing or development of the Department of Homeland Security's ATS data mining program, even though that has been used for years to secretly assign so-called terror scores to law-abiding Americans, I suppose that 90-year-old person in the walker. I will put the Washington Post article in as part of the record.

All I want is the administration to follow the law. They want us to follow the law, they ought to follow the law and let us develop what is best. We all want to stop terrorists, but we do not want to make our own government treat us, all of us, like we are terrorists.

So, Mr. Harper, I read your article on "Effective Counterterrorism and the Limited Role of Predictive Data Mining" with a great deal of interest because data mining becomes more and more a tool to detect terrorist threats.

In May of 2004, 2 years ago, GAO reported that there were at least 14 different government data mining programs in existence today. That was back then.

Now, I favor the use of data mining technology if there are safeguards, but we are talking about millions of dollars—probably billions of dollars by now—in data mining technology in order to predict future terrorist threats. I worry about the huge amount of stuff coming in that does not do a darned thing.

Are you aware of any scientific evidence or empirical data that shows the government data mining programs are an effective tool in predicting future terrorist activity or identifying potential terrorists?

Mr. HARPER. I am not aware of any scientific evidence, of any studies. Unfortunately, the discussion tends to happen in terms of bomb throwing or anecdote, where the ATS system, for example, has been defended based on one anecdote of someone who was turned away from the U.S. border based on ATS and ended up being a bomber in Iraq.

Now, I recently spoke with a reporter who is apparently investigating that story, and it was not necessarily ATS signaling that this was a potential terrorist, but rather that it was a potential immigration over-stayer. So was that an example of the system work-

ing or was it not? That is just an anecdote. We would be much better off with scientific background that justifies this.

Chairman LEAHY. Do you not think we should have a scientific study to find out if we are going to spend millions, even billions, whether this thing actually works?

Mr. HARPER. Absolutely. I think, along with scientific study, allowing technologies like data mining to prove themselves in the private sector will give us much more than allowing government research to happen.

Chairman LEAHY. Dr. Carafano, are you aware of any empirical studies?

Mr. CARAFANO. Well, I think, quite frankly, a review of the scientific literature does not give you a definitive answer of the ultimate potential of data mining technologies to predict behavior. But we should also realize, if you look at the state of behavioral science—

Chairman LEAHY. I am not asking about the potential that someday it may work. Are you aware of any empirical study that these millions of dollars—maybe billions of dollars—we are spending on all these systems seem to be proliferating? Everybody has got to have their own. Are you aware of scientific or empirical studies that say they work?

Mr. CARAFANO. Senator, somebody would have to specifically describe to me the program, then we would have to have a discussion about whether it is actually a data mining program or not. I am not sure that all the systems that GA qualifies is data mining, or ATS, which I do not believe is a data mining system. But the point is, behavioral science modeling is a rapidly developing field.

The combination of computer technology and informatics and behavioral science is producing new advances every day, and so even if I gave you a definitive answer today that said I can guarantee you for a fact that data mining processes cannot predict terrorist behavior, that answer may be totally false 6 months, a year, or 2 years from now. I cannot give you that answer—

Chairman LEAHY. Might we suggest there are some mistakes when Senator Kennedy and Congressman Lewis are told they cannot go on an airplane, or a pilot has to lose a lot of his income because he gets delayed every single time they go through, even though they know it is the wrong guy?

Mr. CARAFANO. Yes, sir. But in all those systems you are doing one-to-one matches. They have got a data point and they are matching a person to that data point. Sometimes those data points are incorrect. That is not data mining.

Chairman LEAHY. I could follow up for a couple of hours on that one, but we will go back to it.

Congressman Barr, in November of 2002, the New York Times reported that DARPA was developing a tracking system, which turned out to be Total Information Awareness.

Privacy concerns were so abhorrent that a Republican-controlled Congress cut the funding for it. But October 31st of last year, an article in the National Journal reported that the Office of the Director of National Intelligence is testing a new computerized system to search very large stores of personal information, including

records of individuals' private communications, financial transactions, and everyday activities that looks very much like TIA.

Are you concerned that a system shut down by the Congress is now reappearing under another form?

Mr. BARR. Very concerned, both as a former Federal prosecutor, certainly as a former Member of this great institution on the House side, and as a citizen concerned about the rule of law.

I think that allowing any administration—and this administration has shown itself to favor this, time again—to do what it wants regardless of what Congress says, either through an appropriations rider or through specific legislation, it breeds contempt for the law, it breeds a lack of credibility that cuts across the board in reducing people's faith in government, and it leads to this further sort of cultural suspicion.

I think it is extremely problematic and I believe that, so long as the Congress allows the administration to do this without either providing an overall architecture such as the Europeans did over a decade ago, and a number of other countries that have shown themselves much more willing than our government to establish a framework within which proper privacy protections can be employed and shall be employed, and yet not harm business at all—the Swiss are a perfect example of that—until Congress addresses this issue, the administration is going to continue to do precisely what you put your finger on, Mr. Chairman, and that is essentially to thumb its nose at the Congress and do what it wants. They just call it something different.

Chairman LEAHY. The concern I have, I mean, you fly on commercial flights, as I do, as most of us do. You have to assume that you have some kind of a terror index score somewhere. You have no way of finding out what that is. I have no way of finding out what that is.

If you are a person working for a bank and you are up for vice president or head of one of the branches or something, and you are suddenly turned down because the bank has found this score, you have no way of knowing what it is, do you?

Mr. BARR. This is the very pernicious nature of what is going on here. You have no way of knowing. You have no way of correcting it.

The particular system that you referred to, Mr. Chairman, that has given rise to the absurd situation of the U.S. Senator and the U.S. Congressman being halted from boarding a plane because their name appears on some list, whether one considers that data mining technically or not, the fact of the matter is, it points out a major problem and a major shortcoming, a fundamental problem in the way we allow government to operate to do this without, as Jim correctly put his finger on, the transparency that at least provides some knowledge and protection for the citizen.

Chairman LEAHY. Thank you. I have further questions of Ms. Harris and others, but my time is virtually up. I will yield to Senator Specter, then we will go, by the early bird rule, to Senator Whitehouse.

Senator SPECTER. Thank you very much, Mr. Chairman.

Congressman Barr, was your privacy violated by the interview in Borat?

Mr. BARR. In what?

Senator SPECTER. Borat.

Mr. BARR. I do not know. Was he an agent of the Federal Government or not? It is a very good question that ought to be proposed to him.

Senator SPECTER. Was your privacy violated?

Mr. BARR. I believe it was. Information was gathered at that interview under false pretenses.

Senator SPECTER. It was an extraordinarily moving interview. Did you have any right to stop its showing or distribution because of the invasion of your privacy?

Mr. BARR. There may be. I know that some legal actions by some other persons involved are being pursued. I elected not to pursue it, believing essentially that the more one wastes time or engages in those sorts of activities, the more publicity you bring to something.

Senator SPECTER. I think that is a valid generalization. If somebody is a Member of Congress with that kind of a high-profile position, you sort of have to take your lumps here and there.

Did you see the movie?

Mr. BARR. I have not. I know folks that have. The movie that revels in nude male wrestling is not something that puts it high on my priority list to see.

[Laughter].

Senator SPECTER. Well, I think the record ought to be clear that you were not featured in any nude male wrestling.

[Laughter].

Mr. BARR. I was going to, but I appreciate the Ranking Member indicating that.

Senator SPECTER. It was a sedate interview in your office somewhere and it was a most extraordinary movie. I do not want to hype it too much or get people to go to see it, but the interview with you was about the only part of the movie worth seeing, Congressman Barr.

Mr. BARR. I will take that as a compliment, Senator.

Senator SPECTER. Well, you should. You should. It is a compliment.

There has been a reference made to the situation where the Automated Targeting System has been credited with the exclusion of an airline passenger. Proponents of ATS point to an incident, purportedly, where ATS was used by the Customs and Border Patrol agent in Chicago's O'Hare Airport to refuse to allow a traveler arriving from Jordan to enter the United States, a man named Riyib Al-Bama, who had a Jordanian visa and a U.S. business visa when he attempted to enter the United States, and 18 months later he reputedly—it is always hard to find out the facts in these matters, but this is the report—killed 125 Iraqis when he drove into a crowd and set off a massive car bomb.

Ms. Harris, are you familiar with that reported incident?

Ms. HARRIS. Well, I am familiar with the allegation. Obviously, there is no way for me to know. But let us assume for the sake of argument that that is true.

Senator SPECTER. Well, now, wait a minute. I am asking you if you are familiar with it.

Ms. HARRIS. Specifically with that case?

Senator SPECTER. Yes.

Ms. HARRIS. Yes. All I know is what I read. I mean, there is no way for me to know.

Senator SPECTER. Well, that is about all any of us could say.

Ms. HARRIS. Right. All I know is what I read.

Senator SPECTER. And when we go to top secret briefings, we walk out with the same conclusion.

Ms. HARRIS. Exactly.

Senator SPECTER. All we know is what we read in the newspapers.

Ms. HARRIS. Right.

Senator SPECTER. In your testimony, you state that unless and until a particular application can be shown to be an effective tool for counterterrorism, the government should not deploy pattern-based data mining as an anti- terrorism tool.

Our hearing today is built on a very, very high level of generalization.

Ms. HARRIS. Right.

Senator SPECTER. And later, if the Chairman has a second round, I want to come back to a question as to, for those who like data mining, what can you point to that it has produced? For those who do not like data mining, what can you point to where there has been an invasion of privacy which has been damaging? I would like to get specifics so we can have some basis to evaluate it.

Because we sit here and listen to high-level generalizations. You talk about oversight. When you pursue oversight—and I am going to be interested in the pursuit of the Attorney General next week—it is a heavy line of pursuit and diligent prosecutors have a hard time catching up.

But before my time goes too much further—

Chairman LEAHY. I should note that I was told that there was an error on the clock before. I thought I was within the time and I went over the time. So, please, take what time you need, then we will go to Senator Whitehouse.

Senator SPECTER. Well, I will just finish up this one question, then yield.

When you talk about proving it to be an effective tool for counterterrorism, how do we make the determination as to what is an effective tool for counter-terrorism?

Ms. HARRIS. Well, I think you have to get the facts. At the moment, Congress does not have the facts. It is not for me to say that a program is corrective because it works once or works ten times. At some point there has to be evaluation criteria, whether it is set in those agencies or Congress sets them.

If the information on the effectiveness has to be secret and is shared only with Congress to make that determination, that is fine. But even if you assume that that program is effective, and I do so only for the sake of argument, there is nothing that exists in that program to protect the rights of the rest of the people, the innocent people.

There is no way that a program like that is designed where we know, because of the level of secrecy, what the impact is. You ask the question, what is the impact? There is a potential that we may

have caught one terrorist, and that would be a good thing. We also do not know what the impact is on the millions of other people who are in that system because they do not know that they are in that system, they have no way to know they are in that system.

So there is no reason for us to deploy these systems and leave us in a situation where there is no due process and no fair information practices. I mean, there are two different questions: one, are they effective and should they deploy it at all?

The second is, if you are going to deploy them, why do we have to deploy them without the traditional procedural protections that this body has imposed, fair information practices, and the Privacy Act, in a variety of other contexts.

So you have to look at both of them. I do not think you address the second, privacy, until you get to the first, efficacy. But if there is, in fact, a person out there in Senator Leahy's example who is trying to figure out why they were fired, that person has no way to know.

Senator SPECTER. Well, you sort of lost me along the way.

Ms. HARRIS. All I am saying is—

Senator SPECTER. Wait a minute.

Ms. HARRIS. Yes.

Senator SPECTER. You sort of lost me along the way.

Ms. HARRIS. All right.

Senator SPECTER. Can you point out any specific instance where data mining has resulted in somebody's demonstrable prejudice?

Ms. HARRIS. Well, of course. I mean, there is demonstrable prejudice. The only ones that we can see visibly at this point are people being searched or people being kept off the plane.

But you have a privacy notice that specifically said, we will share this for any other purpose with the rest of the government, down to the local level. So people are walking around with a risk assessment that they do not know, that is secret, that can be shared all over the government for any other purpose.

If they are prejudiced by that, they do not know because nobody is going to say to them, we have now looked at your risk assessment and that is why you did not get a security clearance, that is why you did not get a job.

Senator SPECTER. If they are kept off the plane though, if they are challenged—

Ms. HARRIS. If they are kept off the plane, they know they have been kept off the plane. But nobody has said to them, we have identified you as a high risk, and here is how you can get out of that. There is no procedure for challenging a risks score.

Senator SPECTER. But until they are kept off the plane, when they have been prejudiced, at that juncture they have a right to challenge it until—

Ms. HARRIS. They have no right to challenge it.

Senator SPECTER. Wait a minute.

Ms. HARRIS. They have no right to challenge it.

Senator SPECTER. Wait a minute. Wait a minute. The question is not posed yet.

Ms. HARRIS. Yes, Senator.

Senator SPECTER. At what point is there prejudice? If they have been kept off the plane, it has been identified, they then have a right to challenge it. But until that time, what is their prejudice?

Ms. HARRIS. Senator, I am not quite sure I agree with you about their right to challenge it. We do not have procedures set up for people to know their risk assessment and to be able to go and challenge it. We do not have those procedures. You can kind of go to TSA or whoever and try to get a response.

I do not mean to be talking past you, but if you are kept off a plane you probably have an idea that perhaps you have a risk assessment that is high. If that is based on data that is inaccurate, I do not know where you go to challenge that data.

We do not have Privacy Rights Act-like privileges. These notices specifically exclude people from those kinds of rights in these programs. All we are arguing is, just putting efficacy aside, that people do have those rights, that we restore them.

Chairman LEAHY. I might use an example, I alluded to it in my opening statement, of an airline pilot. I will identify him. It is Kieran O'Dwyer. Having an Irish surname, I kind of noticed this, notwithstanding my Italian ancestry.

But Kieran O'Dwyer of Pittsboro, North Carolina, an airline pilot for American Airlines. In 2003, he gets off the plane and is detained for 19 minutes on international flight because they told him his name matched one on a government terrorist watch list, apparently somebody from the IRA.

Over the next almost 2 years, he was detained 70 to 80 times. He talked to his Republican Senator and Democratic Congressman and they could not get him off the list. It got so bad, he said, Custom agents came to greet him by his first name. But they still had to detain him because he was on the list and he could not get off it.

So he finally, after missing numerous connecting flights where he has to get to the next flight that he is supposed to fly, having to pay to stay in hotels because he has missed them, he gave up flying internationally, even though he took a five-figure drop in his pay. He just could not do it. That is one example, and I am sure we have many more.

Senator Whitehouse?

Senator WHITEHOUSE. Thank you, Mr. Chairman.

Just a word on my background. Rhode Island is one of those States in which the Attorney General has State-wide criminal law enforcement authority, so like the Senator and the Ranking Member I was, in effect, the DA. I was also the U.S. Attorney for Rhode Island.

I have led and overseen undercover and confidential investigations, so I am well aware of the critical value of that, and also well aware of the civil liberties hazard that that creates. It is very interesting to me to be seeking to apply that balance in this area where there is a new and inevitable technology that has arrived upon our society.

My question to anyone on the panel who would care to answer it, is this. Does it make sense to look at the use of the data mining capability in different ways depending on the different uses of that capability?

And specifically, can we talk about two different uses being one in which a dragnet is run through the data mine based on a profile or based on a formulary, and as a result individual names are surfaced and then further action ensues with respect to those previously unknown or undisclosed names? That would be one category of access to the data mining capability.

The other would be taking a preexisting identified subject of some variety, perhaps a predicated subject of some kind, perhaps not, and running that individual name through the data mining capability to seek for links, contacts, and other things that would be useful in investigating the activities of that individual.

Are those two meaningfully distinct uses of the data mining capability, and in our deliberations should we be considering them separately?

Ms. HARRIS. Mr. Whitehouse, at least from our perspective we do think that those are differing capabilities. I mean, there is a very interesting—I cannot remember if it is a footnote or a page in the Markle report that shows how using sort of existing data and starting with the two terrorists who are on the watch list and looking for links about addresses and a variety of things, that you might have been able to identify all the terrorists. That, to me, is traditional law enforcement.

Now, I understand from Dr. Carafano's view that the line between that as technology advances, and what Mr. Harper and I sort of refer to as predictive or pattern- based, is going to get more muddled as technology advances. But it does offer, I think, a useful place to make a distinction.

First of all, in the suspicion-based, you are sort of engaged in a law enforcement activity. People get identified at some point and action is taken that is, if not public, goes into the law enforcement realm, procedures attach under our laws.

In the predictive realm, we are starting with no predicate. We are starting with no suspect. We may be starting with a set of hypotheticals that are maybe worth testing, but then we are literally moving towards identifying, labeling, perhaps taking actions on people and there never is a procedure that attaches. I think that that is a very big difference.

Senator WHITEHOUSE. Does anyone disagree that this is a meaningful distinction?

Ms. HARRIS. I think these witnesses do.

Senator WHITEHOUSE. Congressman Barr?

Mr. BARR. Thank you, Senator. I do not disagree. I think it is a very important distinction. I think that if, in fact, there is information developed through legitimate intelligence operations, for example, that a particular person is a legitimate suspect, the government certainly needs to follow up on that and run that person's name through in whatever permutations there might be.

But the question or the issue that is the more fundamental one to determine what those distinctions are and how to proceed, is that whatever the system is, it has to pass Fourth Amendment muster.

Data mining, the way I believe it is being used by the government where everybody is a suspect and there is no suspicion, reasonable or otherwise, that a person is or has done something wrong

before evidence is gathered against them, put into, manipulated, retained and disseminated through a data mining base, is not consistent with the Fourth Amendment and it should not hinge, with all due respect to the Ranking Member, on whether or not a person can show that, I have in fact been harmed.

I think the harm is done to society generally where you have a government that can treat all of its citizens and all other persons lawfully in the country as suspects, gather evidence on them, use that data to deny any particular one of them or a group of them, a fundamental right. That, I think, ought to be the starting point for the analysis.

Senator WHITEHOUSE. Would you require a warrant for a government agent to do a Google search?

Mr. BARR. No. The government does not need a warrant to do a search of publicly available information. But in order to be consistent with both existing laws such as the Privacy Act, and consistent with the basic edicts of the Fourth Amendment, if they in fact take it further steps and include information, private information on a person in a database that is to be mined through algorithms manipulated in some way and then potential adverse action taken against a person, I think they do need to consider that, and ought to.

Senator WHITEHOUSE. This will be my last question. So in your view, the privacy barrier that is intruded upon by this is breached when private information goes into the data mine, not when the name emerges from the data mine and the government then begins to take action against an individual.

Mr. BARR. That is correct.

Senator WHITEHOUSE. All right. Thank you, Mr. Chairman.

Mr. TAIPALE. Could I just address it?

Chairman LEAHY. Go ahead. In fact, that is a very good question. If anybody else wants to address, briefly, what Senator Whitehouse asked, go ahead.

Mr. TAIPALE. I think the issue of trying to draw a distinction between link and pattern analysis is very difficult. Again, let me preface all this by saying, I am completely in favor of privacy protection and oversight, and all of those things.

But when you start to get into, actually, the use of these technologies, in context, I mean, we are talking about a lot of different things and going back and forth. So, for instance, in the Ted Kennedy example, that is a one-to-one match. That is a problem with watch lists. If we want to talk about watch lists, that is a problem. There ought to be procedures to deal with that.

Data mining in that case may actually help solve the problem. Here, if Ted Kennedy has stopped because he's on the watch list, but his terror score is very low because he is a U.S. Senator—I do not know if that is true—but if he does have a low score because he is a U.S. Senator, then that ought to be the basis for determining—sort of using independent models to come up with whether that is someplace to spend resources against, as Jim said earlier.

Again, I am not in favor of any particular government program. I am not here endorsing any particular government program. I am merely saying that these are tools that can allocate investigative

and intelligence resources. Going back to the premise of your question about using it in law enforcement, we do this all the time.

The difference between looking for John Smith, or the man in a black suit, or a man in a blue suit, or a person cashing a check under $10,000, or whatever, we do this all the time. We used pattern-based analyses in the IRS to select who gets audited. We do it in the SEC and NSAD to find insider traders. We do it in money laundering.

We do it at the borders with ICE to find drug couriers using drug courier profiles. We use hijacker profiles. All of those have been upheld and, quite frankly, the issue of using a probability-based predicate is something that is not inherently contrary to the Fourth or Fifth Amendment.

Senator SPECTER. Dr. Carafano, you wanted to add something?

Mr. CARAFANO. Yes. I do think that useful distinction in how we address the public policy issues is distinguishing between automating traditional law enforcement activities and the more exotic knowledge management of information to do predictive behaviors.

But the point I would disagree with your division is, not all law enforcement activities begin with a suspect, essentially. I come from a long line of cops. When a cop goes on the street, he is collecting information every second. He is looking for behavior that is out of place. He pulls a car over, and everything else. That leads to a whole thing.

So, no, he is not starting with a suspect, yet he is continually gathering freely accessible information. In a sense, ATS is automating that. I do think that that belongs in a separate discussion because the law there is clear. The question is, are the checks and balances in place? Those are not science experiments. Knowledge management is.

Chairman LEAHY. Ms. Harris, did you want to add to that?

Ms. HARRIS. Well, I wanted to respond to the idea that this is no different than sort of the profiling we do that has been upheld for, for example, stopping a car under a drug profile. That seems to be the basis for this analysis, that this is all right under the Fourth Amendment.

First of all, it is not secret. The police stop you. They know they have stopped you. You have an immediate opportunity to resolve the situation. If you are an innocent person and they have stopped you, and you have consented, there is no long-term use of the data.

Two years later you do not show up for a job with the Federal Government and you get a security clearance denied because somewhere there is now a file that says they stopped you at the Vermont border. That is more like a metal detector.

I really object to this effort to take these cases that involve one-on-one suspicion, one-on-one record analysis from 20 years ago and try to apply them to this complex technical environment we are in. The Supreme Court may have said it is fine to do stops for drug profiling, but it has also said we have to update the Fourth Amendment to take into account technology. That is where we have fallen short. The one thing that I hear from everybody on this committee, is that we all think we have got to do something about the safeguards, whether or not we think predictive data mining works.

Chairman LEAHY. I smiled just briefly. In talking about being stopped down at the Vermont border, I was actually stopped a few years ago. It was a huge stop. They were stopping everybody. I drive back from Vermont about once a year, usually after the August recess, my wife and I. About 100-some-odd miles from the Canadian border, here is this big stop. I had license plate one on the car.

They asked for identification and I was a little bit annoyed and showed them my Senate ID that says I am a U.S. Senator. But they asked, do I have proof of citizenship. I said, you may want to check the Constitution.

[Laughter].

Anyway, I digress. Not that it annoyed me; I still remember it like it was yesterday.

Today, as you said, Ms. Harris, it is something that could be resolved right there. Today we read that the Department of Defense has agreed to alter the uses of a database with information on high school and college students and they have agreed to alter that.

I wish they had done it because of questions being asked by Members of Congress. They did it because they got sued. I will include in the record information on that settlement, including the filing in the Federal Register yesterday amending this government information system.

Senator Specter, did you have anything further? Otherwise I was going to keep the record open so that Senators on both sides could submit anything they wanted to.

Senator SPECTER. Well, thank you, Mr. Chairman. Just one comment. I do not think that I have any disagreement with Congressman Barr with respect to probable cause if there is going to be, as he puts it, an adverse action. I think that is true. But within the range of investigative tools, if there is no adverse action, as Congressman Barr says, and there is no specific prejudice to the individual, then I think there is latitude for law enforcement to look for patterns.

If you put together the 9/11 hijackers, for example, and you have connecting points where they entered about the same time, where they used the same banks, where they go to the same flight schools and do it in a confidential way where there is no disclosure, they have no prejudice and not saying anything adverse about it and doing it in a confidential, discreet way—Congressman Barr used to be a prosecuting attorney. It is a popular background. It gives you a lot of insights into investigative techniques and protection of civil liberties. That is one of the prosecutor's fundamental duties. He is quasi-judicial, to be sure that civil rights are not violated.

But it is a very complex field and it is hard to put your arms around it. It is really hard to figure out exactly where it is going. When we have open sessions, you see on C–SPAN how little we find out, and the sessions you saw which were closed, how little we find out, you would be amazed. Congressman Barr knows. He has been in a lot of them. This 407 on the Senate side, and the House has its own side.

But we have to pursue the matters and we have to keep various Federal agencies on their toes, and give them latitude, but expect them to respect rights. So, thank you, Mr. Chairman.

Chairman LEAHY. Well, thank you. Thank you, Senator Specter. We will have more hearings on it.

I also want to thank the panel. I know that you spent a lot of time preparing for this. It seems like, kind of zip in, zip out. This is important. It is important to this committee.

I worry very much about this privacy matter. We Vermonters just naturally have a sense of privacy, but I think most Americans, too. We want to be secure. But at some point, especially in an interconnected age of the Internet and everything else, when mistakes are made, they are really bad mistakes.

The worst mistakes are those when you do not know a mistake has happened, but it affects everything from your credit rating to your job. It is not what America is about. We talk about connecting the dots with the people in the flight school. Unfortunately, the FBI had all that information. They just chose not to act on it, and we had 9/11.

Thank you all very much. We stand in recess.

[Whereupon, at 10:45 a.m. the hearing was concluded.]

[Questions and answers and submissions for the record follow.]

[Additional material is being retained in the Committee files, see Contents.]

QUESTIONS AND ANSWERS

February 6, 2007

The Honorable Edward M. Kennedy
SR-317 Russell Senate Office Building
Washington, DC 20510-2101

Attn. Nikole Burroughs
Hearing Clerk
Senate Judiciary Committee
224 Dirksen Senate Office Building
Washington, DC 20510

IN RE: Answers to Questions of Senator Edward M. Kennedy, Regarding the Hearing "Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs," January 10, 2007

Dear Senator Kennedy:

In response to your written questions posed after the above-referenced hearing, I am pleased to provide the following answers:

Question 1. Senator Specter asked whether any of the panelists was aware of an instance where data mining had resulted in demonstrable harm? Please expand upon your answer and describe the ways in which ordinary Americans can be harmed by data mining.

Answer. In fact, government data mining has repeatedly led to well-publicized and significant harm to ordinary Americans, including residents of Pennsylvania. A Pittsburgh, Pennsylvania-based example involves data mining queries run against federal employees to determine their suitability to continue working.

As reported by National Public Radio ("NPR") on June 18, 2006 by reporter Pam Fessler, Sodexho, manager of staff for the federal courthouse and prison, placed on administrative leave two long-time employees after the Federal Protective Service (FPS) wrongly determined there were work-disqualifying events in their histories that were uncovered by data mining. Mary Broughton, one of the women erroneously fired, had served as a cook in the cafeteria for 24 years prior to her termination.

Similarly, Judy Miller was employed as a cook and cashier at the federal courthouse for 20 years prior to her termination. Both women were forced to file for unemployment.

Despite hundreds of calls by the women to the U.S. Department of Homeland Security to rectify the situation and return both to their positions, it took the intervention of U.S. Rep. Mike Doyle (D-PA) to obtain an admission by the Department of Homeland Security that its data mining query had identified the wrong women.

In a similar fashion, government data mining can lead to suspicion erroneously being placed on individuals with the same or similar names to known or suspected criminals and terrorists leading to either delays or denials of the right to travel, or the right to work.

Question 2. In your testimony, you explained how data mining infringes upon fundamental constitutional rights. Is it possible to devise safeguards and procedures that would permit the use of data mining against terrorism, without infringing upon constitutional rights? Please explain your answer. In particular, please address the due process and unreasonable search and seizure concerns and how these concerns might be addressed.

Answer. First, no data mining technique has been demonstrated to be able to identify either unknown terrorists or predict when someone might in the future commit an act of terrorism. Until this capability can be scientifically demonstrated, Congress should not pursue or permit federal agencies to pursue data mining.

Second, because of the classified nature of most terrorism investigations, there is unlikely to be any transparency permitting scrutiny of the data mining to determine whether an agency is implementing a program in a manner that is consistent with constitutional rights and privileges. Without transparency and statutory penalties, any purported procedures are meaningless because Congress, much less the public, will never become aware of abuses. Instead, we may only hear when mistakes *were* made. That having been said, any data mining Congress authorizes should only be conducted after a predicate crime or terrorist act has been committed, or where an act towards commission of a conspiracy to commit a crime or act of terrorism has been committed, so that an individual could be charged with conspiracy. Further, any data

mining authorized should only be conducted against suspects and their known associates, and any abuses of this should lead to automatic, statutorily-mandated fines and criminal sanctions.

Finally, without transparency, any due process is unlikely to be provided (or meaningful if provided). However, any due process must be statutorily-guaranteed and provide access to review erroneous information, a meaningful and realistic ability to correct erroneous information and guaranteed government-paid recompense for any injury to the subject of the data mining.

I appreciate the opportunity to provide additional information on this very important topic, if there is anything else you need or have further questions, please do not hesitate to let me know.

With kind regards, I remain,

very truly yours,

BOB BARR
Member of Congress, 1995-2003
Chairman, Patriots to Restore Checks and Balances

Senate Judiciary Committee Hearing
January 10, 2007
"Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs"

Questions re Testimony of James Jay Carafano

Question 1, Bullet 1:

1. **In your written testimony, you suggest several "guidelines" for government data-mining programs. Among other things, you recommend that, "[t]o protect individual privacy, any disclosure of a person's identity should require a judge's approval."**

   - **Would this restriction apply to disclosures from one government agency to another? If, for example, the Department of Defense conducts a data-mining program that yields three or four names of suspected terrorists that might be worthy of follow-up investigation, would you require the DOD to get a court order before sharing that information with the FBI?**

No, I am not recommending that information disclosures among departments or agencies within the federal government require a judge's approval. But at this relatively early stage in the federal government's development and implementation of data-mining technology to predict and prevent terrorist activity, personally identifiable information of an individual who has been identified as a terrorism suspect based solely on data-mining technology should not be disclosed to parties outside of the U.S. government without court involvement. Until the federal government has more experience using data-mining technology for predictive or preventative purposes, the U.S. government should not share such information with, for example, foreign governments to identify persons who should be restricted from travel or private financial institutions that are providing voluntary assistance in terrorist finance investigations.

This restriction against sharing personally identifiable information with parties outside of the U.S. government should not apply, however, if the suspicion of terrorist activity has an independent basis in information obtained using sources or methods other than data-mining technology. Thus, if data-mining technology merely provided the first indication that a person might be involved in terrorist activity, or if the information obtained using data-mining technology merely supports other information obtained using other sources and methods, then the fact that some information was obtained using data-mining technology should not be enough to require court involvement.

Question 1, Bullet 2:

- **Would you likewise require a judge's involvement before the FBI could share a name gleaned through data-mining with state or local law enforcement officers for further investigation?**

At this relatively early stage in the federal government's development and implementation of data-mining technology to predict and prevent terrorist activity, disclosure to state and local law enforcement officials should require some level of court involvement. I recommend that for now many of the same standards should govern disclosures to state and local law enforcement officials in the U.S. as are applied to disclosures to foreign governments and to private parties; however, a more relaxed standard of proof should apply to disclosures to state and local law enforcement officials in the U.S. as long as the person is merely being investigated and not subjected to a Fourth Amendment search or seizure. (See answer to following question.)

Question 1, Bullet 3:

- **If so, what standard of proof would apply? Would you require, for instance, probable cause to believe the identified person is involved in terrorism, even if the person is not going to be arrested or subjected to a Fourth Amendment search?**

I would not recommend that the same standard of proof be required for disclosure to state and local law enforcement officials in the U.S. as should be applied to disclosures to officials of other governments or to private parties. At this time, the judicially created reasonable suspicion standard, or a standard similar to it, is all that should be required for the federal government to disclose to state and local law enforcement officials in the U.S. personally identifiable information about an individual suspected of terrorist activity based solely on predictive data-mining technology.

Question 2:

2. **Your proposed "guidelines" for government data-mining programs include the suggestion that "[t]he federal government's use of data-mining technology should be strictly limited to national security-related investigations." Presumably this restriction would only apply to pattern-based data-mining used for predictive or preventative purposes?**

Correct. My recommendation should have noted that data-mining technology is already in use, with little actual controversy, by organizations in both the public and private sectors to identify patterns of existing fraudulent or otherwise criminal conduct. I am by no means recommending that non-predictive, non-preventative uses of data-mining technology such as these be restricted to national security-related investigations only.

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.

CENTER FOR
DEMOCRACY
&
TECHNOLOGY
Working for Civil Liberties on the Internet

1634 I Street, NW Suite 1100
Washington, DC 20006
202.637.9800
fax 202.637.0968
http://www.cdt.org

February 7, 2007

The Honorable Patrick Leahy
Chairman
Senate Judiciary Committee
Washington, DC 20510

Dear Chairman Leahy,

Thank you again for the opportunity to testify before the Committee on January 10 about the implications of government data mining. We also thank you, Senator Kennedy and Senator Specter for submitting follow-up questions, asking us to elaborate on the important issues raised at the hearing.

All of our answers should be read in the context of the statement in our written testimony about the broad way in which the term "data mining" is used. As we stressed in our testimony, one cannot be either for or against data mining. It is a tool for data analysis. The important questions are: What kind of data mining should the government use, for what purposes, with what consequences for individuals, under what guidelines, and subject to what oversight, auditing and redress?

Answer to Chairman Leahy's question #1.A:

Yes, Congress should consider legislation to place limits on Governmental access to third-party records.

As we stated in our prepared testimony, Congress should make clear that the Privacy Act applies whether the government is creating its own database or acquiring access to a database from a commercial entity. This reform could be accomplished by amending Subsection (m) of the Act to apply to all PII acquired by the government from private sector information services providers. In addition, Congress should require Privacy Impact Assessments for the acquisition of commercial databases. Section 208 of the E-Government Act of 2002 already requires a PIA if the government initiates a new "collection" of information. The same process should apply when the government acquires access to a commercial database containing the same type of information that would be covered if the government itself were collecting it. (In order to improve the utility of PIAs, Congress should require, as a general rule, that they be publicly issued some period of time (such as 60 days) before a program is launched.)

In addition, Congress should require the government to perform an audit of private sector

databases before using them and to publish in the Federal Register a description of the database, the name of the entity from which the agency obtained the database and the amount of the contract for use of the database. Agencies should further be required to adopt regulations that establish fair information practices including a process for redress when it acquires information from the private sector for use in making decisions about individuals.

Congress should require agencies to incorporate provisions into their contracts with commercial entities provisions that provide for penalties when the commercial entity sells information to the agency that the commercial entity knows or should know is inaccurate or when the commercial entity fails to inform the agency of corrections or changes to data in the database.

A number of these ideas are reflected in the Personal Data Privacy and Security Act, introduced in this Congress by Senators Leahy and Specter, which CDT strongly supports.

Additional legislative reforms are needed to address the very low standards for compulsory governmental access to third-party records. In particular, Congress should strengthen the standards for issuance of National Security Letters and orders under Section 215 of the PATRIOT Act. The bi-partisan SAFE Act, S. 737 in the 109th Congress, is an excellent starting point for those reforms; it should be reintroduced and given priority consideration.

Answer to Chairman Leahy's question #1.B:

Yes, the wall between the government and the private sector has been eroded. CDT would not say that data brokers should be considered quasi-governmental, but we do agree that information services companies should be subject to a comprehensive baseline federal privacy law. Unfortunately, Congress has not acted since the Committee's April 2005 hearing, "Securing Electronic Personal Data: Striking a Balance Between Privacy and Commercial and Governmental Use," which examined the roles and responsibilities of information services companies. CDT testified at that hearing and offered five main recommendations:

1. As a first step towards preventing identity theft, entities, including government entities, holding personal data should be required to notify individuals in the event of a security breach.
1. Since notice only kicks in after a breach has occurred, Congress should require entities that electronically store personal information to implement security safeguards, similar to those required by California AB 1950 and the regulations under Gramm-Leach-Bliley.
1. Congress should impose tighter controls on the sale, disclosure and use of Social Security numbers and should seek to break the habit of using the SSN as an authenticator.
1. Congress should address the federal government's growing use of commercial

databases, especially in the law enforcement and national security contexts.

1. Finally, Congress should examinee the "Fair Information Practices" that have helped define privacy in the credit and financial sectors and adapt them as appropriate to the data flows of this new technological and economic landscape.

These ideas are reflected in the Personal Data Privacy and Security Act, introduced in this Congress by Senators Leahy and Specter, which CDT strongly supports.

Answer to Chairman Leahy's question #1.C:

Ye, it is possible to balance the government's legitimate need for information and our most important freedoms. The proposals in the Personal Data Privacy and Security Act (110^{th} Congress) and the SAFE Act (109^{th} Congress) reflect this necessary balance. Those bills would protect privacy and strengthen the national security and law enforcement.

Answer to Chairman Leahy's question #2:

Yes, it is possible to strike a meaningful balance between privacy and security in government data mining programs. In fact, it is necessary if we are to improve security. Privacy protection, checks and balances, accountability and redress are not incompatible with security. To the contrary, clear guidelines and oversight mechanisms are part of the solution. As the 9/11 Commission stated: "The choice between security and liberty is a false choice." The shift in government power and authority that is occurring in response to terrorism, the 9/11 Commission concluded, "calls for an enhanced system of checks and balances to protect the precious liberties that are vital to our way of life."

This conclusion - that privacy protection and accountability must be built into the design and implementation of counterterrorism information sharing systems -- is central to the recommendations of the Markle Task Force on National Security in the Information Age and other bipartisan expert bodies that have carefully studied information technology and its role in fighting terrorism. "We must not sacrifice liberty for security," concluded the Technology and Privacy Advisory Committee (TAPAC) appointed by Secretary of Defense Rumsfeld to study the Total Information Awareness program and related activities. Likewise, the Advisory Panel to Assess Domestic Response Capabilities for Terrorism Involving Weapons of Mass Destruction, chaired by former Virginia Governor James Gilmore, repeatedly stressed that personal freedoms must be at the foundation of the nation's efforts to counter terrorist threats.

Answer to Senator Kennedy's question #1:

We do not believe that the line between "punishing" and merely developing leads for further investigation is as clear as Mr. Taipale suggests. There are many ways in which a government can "punish" a person. Indeed, it is accepted as a matter of First Amendment law that being targeted for investigation can itself have a chilling effect on fundamental

freedoms. And wiretapping is clearly an intrusion on Fourth Amendment rights, so the use of data mining as the trigger for wiretapping would clearly impose a harm on an individual. Moreover, the Executive Branch has been increasing the ways in which it seriously disrupts persons' lives without inflicting punishment in the context of criminal prosecution. Would the use of data mining to generate investigative "leads" which were then pursued by coercive interrogation be on the punishment side of the line or the investigative side of the line? How about something as common as being repeatedly stopped at the airport for secondary screening?

See also our answer to Senator Specter's question #3, where we challenge Mr. Taipale's assumption that the use of data mining results in the courtroom is more objectionable than the use of data mining results for investigative or screening purposes. The use of the results of data mining to punish, assuming such punishment is not extrajudicial, would, in many ways, be more subject to checks and balances than would the use of data mining for screening or investigative purposes.

Answer to Senator Kennedy's question #2.a:

We don't have a clear picture of how the Administration is using data mining, so it is hard to cite concrete examples of demonstrable harm that has resulted from data mining.

We do know, however, that the Administration has relied on seriously erroneous data and faulty analytic tools in some of its key security programs, resulting in demonstrable harm to individuals. Some of the most notorious cases involve the watch lists maintained by the government and their use in screening passengers at airports. Senator Kennedy himself has been incorrectly associated with someone on the watchlist. While the Senator brushed off the inconvenience, others, such as the famous David Nelson of Alaska, has suffered genuine harm in terms of disrupted travel and business plans. One mistake recently revealed involved Sen. Ted Stevens, R-Alaska, whose wife, Catherine, was being identified as "Cat" Stevens and frequently stopped due to confusion with the former name of the folk singer now known as Yusuf Islam, whose name is on the list. The GAO found last year that about half of the tens of thousands of potential matches sent to the Terrorist Screening Center between December 2003 and January 2006 for further research turned out to be misidentifications. Most of these people experienced at least the inconvenience of secondary search.

At the other end of the spectrum is Maher Arar, a Canadian citizen detained in the US on the basis of faulty information and removed to Syria, where he was tortured. After Arar returned to Canada, an investigation was conducted. Last month, Canadian Prime Minister Stephen Harper called on the U.S. government to remove Arar from any of its no-fly or terrorist watchlists, saying "We think the evidence is absolutely clear and that the United States should in good faith remove Mr. Arar from the list."

Answer to Senator Kennedy's question #2.b:

Traditional police stops are subject to a series of protections lacking in the data mining context: The person subjected to the traditional police stop receives immediate notice - the police officer comes up to him and tells him he has been singled out. The scope of the policeman's search is limited: he cannot, for example, look inside a person's luggage - he has to ask for consent or get a warrant. Moreover, the innocent person subject to a traditional police stop has immediate recourse to conclusively clear his name - he opens his luggage and empties his pockets and proves he has no drugs, in which case he is free to go and no adverse record is kept. In the data mining context, the government provides no notice, it denies access to the risk score, so there is no opportunity for a person to clear himself, and the adverse inference may linger for a very long time (40 years in the case of ATS). The protections available in the traditional police stop make it a "reasonable" search, while their absence in the data mining context makes the search unreasonable.

Answer to Senator Kennedy's question #3:

At one level, data mining might be seen as "color blind" or blind to ethnicity and religion. One would hope that government agents would not use overtly ethnic parameters for data analysis in the absence of a specific lead. (The FBI instituted a census of mosques in 2003, and it was reported in December 2005 that FBI agents had been secretly monitoring radiation levels at Islamic mosques, businesses and homes for several years in large cities to determine whether nuclear or chemical bombs were being assembled - no suspicious radiation levels were found.) But it is easy to see how a pattern-based analysis could use factors that are a substitute for ethnicity or religion. For example, a traffic analysis program targeting between the US and an Arab country will inevitably target the calls of Arab-Americans with relatives and legitimate business connections in that country.

CDT has proposed safeguards that could help prevent discriminatory profiling. One approach is what we call "section 215 with teeth." As you know, Section 215 is a provision in the PATRIOT Act giving the government access to commercial data under a very weak standard. An amended section 215 could require a judicial finding, based on facts shown by the government, that there is a reason to believe that terrorist activity is afoot fitting a certain pattern, and that reliable information relevant to the interdiction of that activity would likely be obtained from the search of one or more commercial databases. Under this approach, an agency that had intelligence information about a possible future attack and that wanted to run a pattern-based search to identify potential planners would be required to demonstrate to the court: (1) facts giving reason to believe that a threat existed displaying certain characteristics; (2) a description of the databases that the government wants to search, including an assessment of the sensitivity of the data involved and its accuracy and reliability; (3) an explanation of why other methods of investigation were inadequate; and (4) a statement indicating whether the commercial databases would remain under the control of the commercial source or whether they would be acquired by the government. Among other things, this approach would give the

court the opportunity to determine whether ethnic or religious profiling was an impermissible part of the government's proposed search.

Answer to Senator Specter's question #1:

Mr. Taipale says, "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." If the government has a list of people who attended training camps, that alone gives it the basis for collecting pretty much whatever data it wants about those people and to collect a fair amount of data about those who are closely associated with them. This is not the kind of precursor activity from which one needs to discover some obscure pattern. The same is true of those engaged in "recruiting activity." On the other hand, if the government tries to compile a list of all illegal aliens who transfer money overseas, it is likely to get an undigestable number of leads. However, if the government could run an analysis for all illegal aliens (we're not sure such a list exists) engaged in identity theft who run "front businesses," make money transfers to Pakistan, and possess a lot of weapons, that might in fact be a justifiable "data mining" program. So far, as far as we know, the government has not shown that it has the kind of data that would support such an analysis. Like much of the discussion of data mining, Mr. Taipale's example seems highly speculative.

Answer to Senator Specter's question #2:

In CDT's view, the standard is not perfection. Rather the standard is: does the program materially assist in the pursuit of a mission (keeping terrorists off airplanes, keeping terrorists from entering the country), without high levels of collateral damage to civil liberties, to the extent that in a world of limited resources, the program deserves to made a priority over other efforts that would serve the same mission. That's not a mathematical formula, but we believe it is better than anything the government is applying today to decide which data mining programs to launch. With such a standard, the government would not be precluded from deploying data mining technology. Rather, it would be empowered to deploy data mining technology that meaningfully advances the national security

Answer to Senator Specter's question #3:

Yes, recognizing that "data mining" is a very broad term that may include intuitively uncontroversial data analysis techniques, one should distinguish between using data mining as an evidentiary tool in a court of law as opposed to an investigative or screening tool. Some "data mining" techniques might be perfectly suited to the analysis of evidence for presentation in the courtroom. For example, it might be appropriate to apply data analysis techniques to the large amount of data collected with a court authorized pen register, and to introduce those results in a courtroom to illustrate a chain of events of circumstantial significance.

However, the key point to recognize is that, in the courtroom, use of data mining for evidentiary purposes would be subject to vigorous cross-examination, presentation of contrary evidence and the other due process protections afforded in the trail setting. Among other protections, there is full notice of the use of the technique. If the matter were criminal in nature, the burden of proof would be on the government. The data mining technique itself might be subject to scrutiny under Federal Rule of Evidence 702, which charges the federal courts with the responsibility of acting as gatekeepers for all scientific and expert testimony. The threshold question for introduction of the evidence is reliability (or, as we stressed in our testimony, "efficacy").

None of these protections are available in the screening or investigative contexts, so in some ways data mining is riskier in those contexts. As we said in our written testimony, application of data mining in the investigative or screening contexts must be preceded by an independent assessment of the reliability or effectiveness of the technique. There should also be notice, beginning with the kind of generic notice that would be provided by Senator Feingold's bill. Redress procedures must be adopted so that individuals can challenge false inferences drawn about them and correct faulty information.

The differences between the protections that would be available when data mining is used as an evidentiary tool and the current lack of those protections when it is used for investigative or screening purposes argues for the position CDT took at the hearing: Congress should use the power of the purse to prohibit the use of unauthorized data mining (defined as predictive or pattern-based scans of large sets of data, where the goal is to assign risk scores or find individuals whose behavior matches some pattern believed to be associated with terrorist or criminal behavior). If the Executive Branch thinks it has an effective program, it should come forward and tell Congress, explain the program and get the money for it. Congress has already put that limit on implementation of the risk assessment program "Secure Flight." CDT urges Congress to do the same across the board.

Answer to Senator Specter's question #4(a):

Your question asks whether requiring the government to demonstrate an application's absolute effectiveness before permitting its use would interfere with or prohibit innovation. It might, but we know of no one who has proposed "absolute effectiveness" as the standard for deployment of any technique. That is certainly not CDT's position. The current posture of the Executive Branch is that it need offer no showing of effectiveness before deploying a technique. That approach is dangerous to civil liberties and national security. As we stated above in answer to your question #2, we believe that a workable standard would be whether the program materially assists in the pursuit of a mission (keeping terrorists off airplanes, keeping terrorists from entering the country), without high levels of collateral damage to civil liberties, to the extent that in a world of limited resources, the program deserves to be made a priority over other efforts that would serve the same mission. Ultimately, it would be a judgment call. We believe

Congress, as the appropriator, should have a role in that judgment. Right now, as far as we can tell, that judgment is not made on a systematic basis by the Executive branch and is certainly made without Congressional input.

Answer to Senator Specter's question # 4(b):

The development and publication of authorized procedures or prohibitions for data mining could and should be done without enabling countermeasures and evasion. (Evasion isn't necessarily a bad thing. A lot of our national counterterrorism program is intended to induce evasion, in the sense that airline screening is intended to induce terrorists to avoid airports, and physical protection measures around important sites are intended to compel terrorists to go elsewhere.)

In our testimony, we outlined several elements of guidelines that could be developed without disclosing anything of use to the enemy:

1. Strong data quality standards, including minimum standards for watchlists, and other procedures to ensure that the databases the government uses to establish the identity of individuals or make assessments about individuals are sufficiently accurate and reliable that they will not produce a large number of false positives or unjustified adverse consequences.
2. Corrective mechanisms, including assessments of the reliability of commercial databases and automated mechanisms that can identify and correct errors in shared data, with responsibility on both the originator and the recipient of data.
3. Access controls, security measures and permissioning technologies that can protect against improper access to personal information, including the ability to restrict access privileges so that data can be used only for a particular purpose, for a finite period of time, and by people with the necessary permissions.
4. Automated and tamper-proof audit trails that can protect against misuse of data, improve security, and facilitate oversight.
5. Redress mechanisms that allow individuals to respond when they are about to face adverse consequences based on information. This includes the right to challenge inaccurate information.
6. Effective oversight of the use and operation of the system, including privacy officers with sufficient powers and resources to enforce the guidelines.

CDT has prepared a detailed analysis of guidelines for information sharing issued by the Administration in December 2006. The analysis describes in further detail some fo the issues that should be addressed in guidelines, none of which would jeopardize operational effectiveness. http://www.cdt.org/security/20070205iseanalysis.pdf.

The Center for Democracy and Technology appreciates this opportunity to discuss in greater detail the important questions surrounding the privacy implications of government data mining. We look forward to working with the Committee as you continue your oversight and legislative work in this area, seeking to develop a more balanced approach to the government's use of information.

Sincerely,

Leslie Harris
Executive Director

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.

SUBMISSIONS FOR THE RECORD

OFFICE OF BOB BARR
Member of Congress, 1995-2003

TESTIMONY
BEFORE THE
SENATE JUDICIARY COMMITTEE ON
"BALANCING PRIVACY AND SECURITY: THE PRIVACY IMPLICATIONS OF GOVERNMENT DATA MINING PROGRAMS"
BY
BOB BARR
JANUARY 10, 2007

Thank you for inviting me to this first oversight hearing of the Senate Judiciary Committee in the 110th Congress. I am extremely pleased that Congress is finally asking hard questions about the impact of the administration's security policies on Americans' privacy and civil liberties. This dialogue is long overdue.

As a former member of Congress, I have been disappointed to see the Congress shirk its responsibility to the American people and sit silently by while the Constitution is gutted of meaning.

As chairman of Patriots to Restore Checks and Balances, an alliance of individuals and organizations – conservatives and liberals – committed to upholding the Constitution, I have worked with many Republicans

and Democrats to do what is right for the American people. I appreciate the opportunity to talk today about the constitutional questions raised by the federal government's data-mining practices.

It is unconscionable that ordinary Americans' jobs and finances – their entire lives – are at risk because they do not know what information the government is collecting about them; or what it is doing with that private information; or who government is sharing the information with. And if that information is wrong, they have no way of knowing about it, no way of seeing it, and no way of correcting it.

Data mining presents many serious threats to the First, Second, Fourth and Fifth Amendments to the Constitution. That is nearly half of the Bill or Rights! Where will this end? With the repeal of the Constitution so that the White House won't have to worry about those inconvenient and troublesome laws any more?

The federal government constantly is taking in huge amounts of information on Americans from many sources; some of these databases are known, some are not; some may be lawful, others not. Every month there is a new revelation. Last week we learned that the administration

wants to open our mail at its discretion, in addition to listening to our phone calls and reading our e-mails without court order.

Just weeks earlier the Department of Homeland Security admitted that its Secure Flight program to screen domestic air passengers violated the Privacy Act. Just prior to that, we learned that Customs and Border Patrol was using the Automated Targeting System, designed initially for cargo security, to assign a terror risk score to travelers entering the United States. Anyone in this room who has traveled abroad in recent years is likely in this system. And their records will be kept for 40 years.

States will soon begin to implement the Real ID law, creating a national registry of tens of millions of drivers. Accessible to officials across the nation, this database, currently being finalized for implementation in 2008, is almost certain to contain individuals' fingerprints, photo, Social Security number, immigration status and more (possibly including other biometric data and an RFID chip).

We learned recently the FBI has been using "national security letters" to excess; in one example, using this easy way to demand access to private data, to collect information on nearly 300,000 people who did nothing

more suspicious than that they spent the Christmas holiday in Las Vegas. Who knows how many other instances of mass data collection have occurred in the past few years, all in the name of national security? The government is re-analyzing perfectly lawful behavior through unproven data-mining programs and bringing vast numbers of innocent Americans under suspicion.

Adding insult to injury, there is no scientific proof that data-mining to identify terrorists even works. No scientist has ever demonstrated that the government can predict who will commit an act of terror at some future time. Yet, the government spends tens of billions of taxpayers' dollars on data-mining programs each year --collecting, manipulating, retaining and disseminating the most personal and private information on unknowing American citizens and others.

Chilling effects on ordinary Americans necessarily follow. For example, an individual decides to learn Arabic to help their country fight terrorism. They travel to an Arabic speaking nation such as Egypt, which maintains close and cooperative ties with the U.S., to study the language, but when they come home and apply for a job with the federal government, they can't pass the background check because a database,

perhaps the Automated Targeting System, shows that they traveled to Egypt. This just isn't right, and it may very well be counter-productive. Data-mining, therefore, has the propensity to make us more vulnerable, not safer.

Data-mining undermines the First Amendment guarantees for freedom of association. Using link-analysis data mining, a person can easily be found guilty by association. This means that anyone who comes into contact – even incidental contact – with a person whose name appears on some list as a terrorist suspect, become a suspect themselves. Once a person is linked to a terrorist, it is virtually impossible to clear his or her name – if they even know they have come under suspicion.

The First Amendment also implies a right to travel and to move freely throughout society. However, when the simple fact of traveling puts people under suspicion, then they may very well curtail or stop traveling for business and other purposes to avoid the hassle of extra scrutiny at the airport or being put on a "watch list."

Concerns about data-mining relate to other of our rights guaranteed by the Bill of Rights. For example, I am deeply concerned about data

mining threatening the Second Amendment right to bear arms. Although the government is prohibited by law from creating a national registry of gun owners, it can purchase records from data brokers that in a sense provide this information. This is also a problem under the Real ID Act, which will contain all sorts of data the average applicant for, or holder of, a state drivers license, possibly including information on firearms records. The government will claim it isn't creating a "registry," it is just analyzing data, and they will have circumvented the registry prohibition. Perhaps the nation's farmers who buy nitrate-rich fertilizer will also end up in the data mining programs and come under suspicion without reason.

Data mining is also entirely incompatible with the Fourth Amendment prohibition on unreasonable search and seizure. Our justice system and ability to prosecute suspects is based on crimes that have been committed or planned. It is absurd for the government to use databases to **predict** individuals' future acts. We do not live in the Hollywood movie scenario depicted in *Minority Report,* where law enforcement halts "pre-crimes" before they happen, yet the practice of government data-mining, which collects personal information on citizens and other persons often without any suspicion or evidence they have done

anything wrong, grows exponentially; a practice undermining the very rights supposed to be protected by the Fourth Amendment.

The Fifth Amendment's Due Process Clause requires that the government tell Americans what personal information is collected about us, how it is being used, and to provide a right to challenge and correct erroneous information that wrongly could be used to deny us our rights and privileges. Yet, none of these shadowy data mining programs provides such a process.

I urge this committee and Congress to consider seriously strict laws to regulate data–mining by government and private industry, and provide oversight concerning their government contracts, so that government agencies are not able to evade federal laws that provide at least some protection against abuse; laws such as the Privacy Act and the Freedom of Information Act. The point here is not to unduly restrict or prohibit the accumulation or analysis of commercially-relevant data for legitimate business purposes. Rather the goal should be to ensure the process possesses a necessary degree of transparency, that it provides essential privacy protection for the consumer, and that such databases are not a

tool whereby government can circumvent the law or the requirements of the Bill of Rights.

Funding for the Total Information Awareness system and other discredited programs may have been cut off because of privacy concerns, but other heads of the beast have sprung up in its place with new names. These programs have no greater safeguards for Americans' privacy and should also be ended.

Finally, I urge the committee to re-introduce and pass the Personal Data Privacy and Security Act and the Federal Agency Data-Mining Reporting Act in the 110th Congress.

Thank you again for having me here today. I look forward to working with the committee over the next two years on these important issues.

January 19, 2007

The Honorable Patrick Leahy
United States Senate
Washington, DC 20510

Dear Senator Leahy:

On behalf of Patriots to Restore Checks and Balances and its alliance of conservative and liberal individuals and organizations, I am writing to express strong support for S. 236, the bipartisan "Federal Agency Data-Mining Reporting Act of 2007." Recently reintroduced by Senators Russ Feingold, D-Wis., and John Sununu, R-N.H., the act would require all federal agencies to annually report to Congress—in classified form if needed—on certain data-mining programs and how these programs impact the civil liberties and privacy of Americans.

According to a May 2004 report by the General Accounting Office, there are at least 199 different government data-mining programs operating or planned throughout the federal government, with at least 52 different federal agencies currently using data-mining technology. Most of these data-mining programs have been operating based on the president's unilateral actions, without needed oversight from either the judicial or legislative branches of our government.

The Federal Agency Data-Mining Reporting Act is a small but significant step toward ensuring that intelligence operations comply with the law and safeguard Americans' civil liberties. Most importantly, this legislation makes it clear that the president's data-mining activities must adhere to review by Congress. This means that the government must provide the following information:

- Descriptions of what patterns are searched, how they are developed and why they connect to a criminal or terrorist activity;
- The number and type of searches run and patterns evaluated;
- The number of individuals found to fit those patterns and how many are eventually investigated and/or arrested; and
- What information is being used in the searches.

We must preserve our system of checks and balances, not simply forgo it in the name of "national security." The Federal Agency Data-Mining Reporting Act upholds these principles and would afford at least some protection for ordinary Americans from unlawful invasions of their privacy. Congress should move quickly to pass this modest, but important and responsible piece of legislation.

I look forward to working with you to strike an appropriate balance between privacy and security in the 110th Congress.

Sincerely,

Bob Barr
Member of Congress, 1995-2003
Chairman, Patriots to Restore Checks and Balances

1718 M Street, NW, Mailbox #232, Washington, DC 20036
Phone: 1-800-583-9122 **Web site**: www.checksbalances.org

01/19/2007 2:06PM

STATEMENT OF

DR. JAMES JAY CARAFANO

SENIOR RESEARCH FELLOW
THE HERITAGE FOUNDATION

214 MASSACHUSETTS AVENUE, NE
WASHINGTON, DC 20002

BEFORE THE SENATE JUDICIARY COMMITTEE

PROMOTING SECURITY AND CIVIL LIBERTIES: THE ROLE OF DATA MINING IN COMBATTING TERRORISM

JANUARY 10, 2007

Mr. Chairman and other distinguished Members, I am honored to testify before you today.[1] In my testimony, I would like to: 1) describe the nature of the challenge facing Congress; 2) offer a set of principles for both enhancing counterterrorism programs and protecting civil liberties; and 3) suggest how these principles should be applied to the employment of data mining technologies.

Between Liberty and Order

Even though I appreciate the opportunity to testify before the committee, I must state at the outset that I reject the premise of this hearing. It is wrong to conceptualize the government's task as an effort to "balance" preventing terrorist attacks and protecting the liberties of individual citizens. Such a paradigm implies making trade-offs. Indeed, the late Supreme Court Justice William Rehnquist suggested that in time of war compromises had to be made. He wrote:

---

[1] The title and affiliation are for identification purposes only. Staff of The Heritage Foundation testify as individuals. The views expressed are our own and do not reflect an institutional position for The Heritage Foundation or its board of trustees. The Heritage Foundation is a public policy, research, and educational organization. It is privately supported, receives no funds from government at any level, and performs no government or other contract work. The Heritage Foundation is the most broadly supported think tank in the United States. During the past two years, it had approximately 275,000 individual, foundation, and corporate supporters representing every State in the nation. Its 2005 contributions came from the following sources: individuals (63%), foundations (21%), corporations (4%), investment income (9%), publication sales and other sources (3%).

> In any civilized society the most important task is achieving a proper balance between freedom and order. In wartime, reason and history both suggest that this balance shifts in favor of order—in favor of the government's ability to deal with conditions that threaten the national well-being.[2]

Yet in a long war, where societies must remain secure, free, and prosperous in order to compete and thrive, shifting the balance between liberty and order is fraught with danger.[3] This is particularly true when facing a protracted terrorist threat. One clear advantage for any country facing a determined enemy is a strong civil society. A resilient populace can better resist the fear, doubt, and despair that terrorists try to sow. Paradoxically one of the great fears of fighting terrorism is that civil society will become the first casualty—that efforts to add security and forestall attacks will undermine the liberties that make societies free and strong to begin with. To frame the fight against terrorism as a choice between safety and freedom offers a false choice. The most effective way to wage a war on terrorism is to adopt policies that secure both safety and freedom equally well.

Freedom from Fear

There has, however, been a concerted effort since September 11 to make the case that enhancing security and protecting freedoms are mutually exclusive. There are three factors animating fears about anti-terrorism campaigns.

- First, critics frequently decry the expansion of executive authority in its own right. They generically equate the potential for abuse of executive branch authority with the existence of actual abuse. They argue that the growth in presidential power is a threat, whether or not that power has, in fact, been misused. These critics come from a long tradition of limited government, which fears any expansion of executive authority.

- The second kind of criticism is stimulated by the "Luddite response"—a fear of technology. As the government begins to explore ways of taking advantage of the information age's superior capacity to manage data through new information technologies, there are rising concerns that it will use these means intrude into our personnel lives. Information equals power. With great efficiency comes more effective use of power. And with more power comes more abuse.

- A third theme underlying criticism is more blatantly political. Take, for example, the passage of the first major post-9/11 anti-terrorism law in the Unites States, popularly called the Patriot Act. The Patriot Act, regardless of its true merits or laws, has been a *cause célèbre* for raising money and energizing constituencies

---

[2]William Rehnquist, *All the Laws But One: Civil Liberties in Wartime* (New York: Knopf, 1998), p. 222.
[3]See Chapter 3, "Between Liberty and Order," in James Jay Carafano and Paul Rosenzweig, *Winning the Long War: Lessons from the Cold War for Defeating Terrorism and Preserving Freedom* (Washington, D.C.: The Heritage Foundation, 2005), pp.79–97.

> that are predisposed to be critical of the Bush Administration's response to terrorism. Brand labeling has become a part of the political process.[4]

One key task of understanding how well government policies affirm the dual priorities of liberty and order is distinguishing real conflicts in achieving both from merely rhetorical arguments that are more concerned with advancing ideological and political agendas than adopting security measures to keep people safe, free, and prosperous.

The Reality of Terrorism

Simply arguing against adding security out of the fear that it might encroach on individual liberties might be prudent if there were no real threats to be addressed. That, however, is not the case. The sad truth is that terrorism remains a potent threat to international security. All we know for sure is that no one can say with much certainty how many terrorists with aspirations of waging transnational war there are, where they are, and what they are planning. Virtually every terrorism expert in and out of the government believes there is a significant risk of more attacks.

In addition, we know that an efficacious defense against terrorism will not be accomplished by military power alone. Rather, effective law enforcement and intelligence gathering are essential instruments. Equally important, this is policing of a different form—preventative rather than reactive.

An understanding of the nature of the terrorist threat helps to explain why the traditional law enforcement paradigm needs to be modified and why government can't avoid its obligation to advance both liberty and order. The traditional law enforcement model is highly protective of civil liberty in preference to physical security. All lawyers have heard some form of the maxim "It is better that ten guilty persons go free than that one innocent person be mistakenly punished."[5] This embodies a fundamentally moral judgment that when it comes to enforcing criminal law. This dictum, however, does not suffice when considering matters of national security in which the state has a dual responsibility to protect both the individual and the people.

Principles for Preserving Security and Civil Liberties

---

[4]See MoveOn.org, "The Administration Is Using Fear as a Political Tool," *The New York Times,* November 25, 2003, p. A1. Their Web site offers a full-page ad reprinting excerpts of speeches by former Vice President Al Gore. It is no coincidence that many Democratic presidential aspirants garnered great applause with the "novel" suggestion that, if elected, they would fire Attorney General John Ashcroft. See Carl Matzelle, "Gephardt Talks the Talk Steelworkers Want to Hear," *Cleveland Plain Dealer*, December 7, 2003, p. A24 (includes a promise to fire Ashcroft "within [the] first five seconds" of new Administration); and Greg Pierce, "Inside Politics," *The Washington Times*, September 23, 2003, p. A6 (noting the "frenzy" of "Ashcroft bashing"). To the extent that criticism of the Patriot Act and related activities is purely political, the debate about these truly difficult questions is diminished. Thoughtful criticism recognizes both the new realities of the post-9/11 world and the potential for benefit *and* abuse in governmental activity.

[5]*Furman v. Georgia*, 408 U.S. 238, 367, n.158 (1972).

Although a large portion of the debate about new law enforcement and intelligence measures focuses on perceived intrusions on human liberties, we should keep in mind that good governance weighs heavily on both sides of the debate. Thus, as we assess questions of civil liberty and human rights, we cannot lose sight of the dual purpose of government—protecting personal and national security. So how do we square the circle?

What we need for the war on terrorism is a set of principles that work for this long war, principles that are consistent with good governance that give us the tools we need to get the terrorists before they get us. The "first" principles that I have advocated for include:

- **No fundamental liberty guaranteed by the laws of a sovereign state can be breached or infringed upon.** This should include the protection of human rights guaranteed by international treaties, which when ratified by the state have the force of national law.

- **Any new intrusion must be justified by a demonstration of its effectiveness in diminishing the threat.** If the new system works poorly by, for example, creating a large number of false positives, it is suspect. Conversely, if there is a close "fit" between the technology and the threat (that is, if it is accurate and useful in predicting or thwarting terrorism), the technology should be more willingly embraced.

- **The full extent and nature of the intrusion worked by the system must be understood and appropriately limited.** Not all intrusions are justified simply because they are effective. Strip searches at airports would prevent people from boarding planes with weapons, but at too high a cost.

- **Whatever the justification for the intrusion, if there are less intrusive means of achieving the same end (at a reasonably comparable cost), the less intrusive means ought to be preferred.** There is no reason to erode Americans' privacy when equivalent results can be achieved without doing so.

Any new system developed and implemented must be designed to be tolerable in the long term. The War on Terrorism is one with no immediately foreseeable end. Thus, excessive intrusions may not be justified as emergency measures that will lapse upon the termination of hostilities. Policymakers must be restrained in their actions; Americans might have to live with their consequences for a long time.

Rules for New Technologies

Because technology is going to be an important part of any set of counterterrorism tools, and because our lives in the information age are so dependent on many of the systems and databases in which these technologies will look for information about terrorists, we also need a set of rules to guide how we implement the basic principles of long-war fighting in the electronic world. This is what these principles should look like:

- **No new system should alter or contravene existing legal restrictions on the government's ability to access data about private individuals.** Any new system should mirror and implement existing legal limitations on domestic or foreign activity.

- **Development of new technology is not a basis for authorizing new government powers or new government capabilities.** Any such expansion should be independently justified.

- **No new system that materially affects citizens' privacy should be developed without specific authorization by the people's representatives and without provisions for oversight of the system's operation.**

- **Any new system should be, to the maximum extent practical, tamper proof.** To the extent the prevention of abuse is impossible, any new system should have built-in safeguards to ensure that abuse is both evident and traceable.

- **Any new system should, to the maximum extent practical, be developed in a manner that incorporates technological improvements in the protection of civil liberties.**

Finally, no new system should be implemented without this full panoply of protections against its abuse.

Application to Employing Data Mining Technologies

First, we must always protect liberties guaranteed by the Constitution. From a practical perspective, there are two distinct types of constitutional violations to be avoided. It should go without stating, but we must never countenance intentional or systemic constitutional violations. In other words, we should design every data-mining system so that, if properly used, it will never violate constitutional rights.

Nevertheless, even an information system that is properly designed using state-of-the-art technologies and privacy safeguards can carry the potential for misuse and abuse. Our goal in the second instance must be to remain vigilant to prevent, identify, and appropriately punish such violations. Inadvertent or negligent violations should be punishable by civil penalties. Intentional violations should be punishable by both civil and criminal penalties.

Second, any imposition on a valid privacy interest by a data-mining program must be justified by the severity of the threat. Standards should be developed for assessing and comparing the relative severity of various threats. Federal departments and agencies should adopt and implement these standards widely and uniformly. Standardization poses the risk of a widespread over-estimate or under-estimate of a particular threat's severity, but the alternative is a flying-by-the-seat-of-the-pants approach that cannot be properly vetted or tested.

Similarly, any new intrusion must be justified by a demonstration of the data-mining program's effectiveness in diminishing the terrorist threat. If the new program works poorly by, for example, creating a large number of false positives, it should be considered suspect. Conversely, if there is a close "fit" between the technology and the threat (that is, for example, if it is accurate and useful in predicting or thwarting terrorism), the technology should be more willingly embraced.

Third, we must understand and limit the imposition on privacy interests. The full extent and nature of the intrusion worked by the system must be understood and appropriately limited. Intrusions should not be justified simply because they are effective.

Fourth, we must strive to develop methods and systems for data mining that are—of the reasonable and feasible alternatives—the least intrusive upon privacy rights. There is no reason to erode Americans' privacy when equivalent results can be achieved without doing so.

Moving Forward

There is clearly a roll for Congress in advancing the use of data mining and other information technologies and ensuring they are employed in an appropriate manner. Establishing federal guidelines for the use of these technologies is one way to address the issue. Such guidelines would begin by defining what programs should come under the scope of data-mining programs. The guidelines should also include the following elements:

- **Every deployment of federal data-mining technology should require authorization by Congress;**
- **Agencies should institute internal guidelines for using data analysis technologies, and all systems should be structured to meet existing legal limitations on access to third-party data;**
- **A Senate-confirmed official should authorize any use of data-mining technology to examine terrorist patterns, and the system used should allow only for the initial query of government databases and disaggregate personally identifying information from the pattern analysis results;**
- **To protect individual privacy, any disclosure of a person's identity should require a judge's approval;**
- **A statute or regulation should require that the only consequence of being identified through pattern analysis is further investigation;**
- **A robust legal mechanism should be created to correct false positive identifications;**

- **To prevent abuse, accountability and oversight should be strengthened by including internal policy controls, training, executive and legislative oversight, and civil and criminal penalties for abuse; and**

- **The federal government's use of data-mining technology should be strictly limited to national security-related investigations.** [6]

Congress should also require agencies to report on their intent to establish data-mining programs and require annual reports on their implementation, as well as their compliance with federal guidelines.

Thank you for the opportunity to testify on this important subject.

---

[6]Paul Rosenzweig, "Proposals for Implementing the Terrorism Information Awareness System," Heritage Foundation *Legal Memorandum* No. 8, August 7, 2003, at *www.heritage.org/Research/HomelandDefense/lm8.cfm.*

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 subject-based 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 pattern-based 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) red-teaming; 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.

Policy Analysis

No. 584 December 11, 2006

Routing

Effective Counterterrorism and the Limited Role of Predictive Data Mining

by Jeff Jonas and Jim Harper

Executive Summary

The terrorist attacks on September 11, 2001, spurred extraordinary efforts intended to protect America from the newly highlighted scourge of international terrorism. Among the efforts was the consideration and possible use of "data mining" as a way to discover planning and preparation for terrorism. Data mining is the process of searching data for previously unknown patterns and using those patterns to predict future outcomes.

Information about key members of the 9/11 plot was available to the U.S. government prior to the attacks, and the 9/11 terrorists were closely connected to one another in a multitude of ways. The National Commission on Terrorist Attacks upon the United States concluded that, by pursuing the leads available to it at the time, the government might have derailed the plan.

Though data mining has many valuable uses, it is not well suited to the terrorist discovery problem. It would be unfortunate if data mining for terrorism discovery had currency within national security, law enforcement, and technology circles because pursuing this use of data mining would waste taxpayer dollars, needlessly infringe on privacy and civil liberties, and misdirect the valuable time and energy of the men and women in the national security community.

What the 9/11 story most clearly calls for is a sharper focus on the part of our national security agencies—their focus had undoubtedly sharpened by the end of the day on September 11, 2001—along with the ability to efficiently locate, access, and aggregate information about specific suspects.

Jeff Jonas is distinguished engineer and chief scientist with IBM's Entity Analytic Solutions Group. Jim Harper is director of information policy studies at the Cato Institute and author of Identity Crisis: How Identification Is Overused and Misunderstood.

CATO INSTITUTE

Though data mining has many valuable uses, it is not well suited to the terrorist discovery problem.

Introduction

The terrorist attacks on September 11, 2001, spurred extraordinary efforts intended to protect America from the newly highlighted scourge of international terrorism. Congress and the president reacted quickly to the attacks, passing the USA-PATRIOT Act,[1] which made substantial changes to laws that govern criminal and national security investigations. In 2004 the report of the National Commission on Terrorist Attacks upon the United States (also known as the 9/11 Commission) provided enormous insight into the lead-up to 9/11 and the events of that day. The report spawned a further round of policy changes, most notably enactment of the Intelligence Reform and Terrorism Prevention Act of 2004.[2]

Information about key members of the 9/11 plot was available to the U.S. government prior to the attacks, and the 9/11 terrorists were closely connected to one another in a multitude of ways. The 9/11 Commission concluded that, by pursuing the leads available to it at the time, the government might have derailed the plan.

What the 9/11 story most clearly calls for is sharper focus on the part of our national security agencies and the ability to efficiently locate, access, and aggregate information about specific suspects. Investigators should use intelligence to identify subjects of interest and then follow specific leads to detect and preempt terrorism. But a significant reaction to 9/11 beyond Congress's amendments to federal law was the consideration and possible use of "data mining" as a way to discover planning and preparation for terrorism.

Data mining is not an effective way to discover incipient terrorism. Though data mining has many valuable uses, it is not well suited to the terrorist discovery problem. It would be unfortunate if data mining for terrorism discovery had currency within national security, law enforcement, and technology circles because pursuing this use of data mining would waste taxpayer dollars, needlessly infringe on privacy and civil liberties, and misdirect the valuable time and energy of the men and women in the national security community.

We must continue to study and analyze the events surrounding the 9/11 attacks so that the most appropriate policies can be used to suppress terror, safeguard Americans, and protect American values. This is all the more important in light of recent controversies about the monitoring of telephone calls and the collection of telephone traffic data by the U.S. National Security Agency, as well as surveillance of international financial transactions by the U.S. Department of the Treasury.

While hindsight is 20/20, the details of the 9/11 story reveal that federal authorities had significant opportunities to unravel the 9/11 terrorist plot and potentially avert that day's tragedies. Two of the terrorists who ultimately hijacked and destroyed American Airlines flight 77 were already considered suspects by federal authorities and known to be in the United States. One of them was known to have associated with what a CIA official called a "major league killer."[3] Finding them and connecting them to other September 11 hijackers would have been possible—indeed, quite feasible—using the legal authority and investigative systems that existed before the attacks.

In the days and months before 9/11, new laws and technologies like predictive data mining were not necessary to connect the dots. What was needed to reveal the remaining 9/11 conspirators was better communication, collaboration, a heightened focus on the two known terrorists, and traditional investigative processes.

This paper is not intended to attack the hard-working and well-intentioned members of our law enforcement and intelligence communities. Rather, it seeks to illustrate that predictive data mining, while well suited to certain endeavors, is problematic and generally counterproductive in national security settings where its use is intended to ferret out the next terrorist.

The Story behind 9/11

Details of the run-up to 9/11 provide tremendous insight into what could have

been done to hamper or even entirely avert the 9/11 attacks. Failing to recognize these details and learn from them could compound the tragedy either by permitting future attacks or by encouraging acquiescence to measures that erode civil liberties without protecting the country.

In early January 2000 covert surveillance revealed a terrorist planning meeting in Kuala Lumpur that included Nawaf al-Hazmi, Khalid al-Mihdhar, and others.[4] In March 2000 the CIA was informed that Nawaf al-Hazmi departed Malaysia on a United Airlines flight for Los Angeles. (Although unreported at the time, al-Mihdhar was on the same flight.) The CIA did not notify the State Department and the FBI.[5] Later to join the 9/11 hijackings, both were known to be linked with al-Qaeda and specifically with the 1998 embassy bombings in Tanzania and Kenya.[6] As the 9/11 Commission reported, the trail was lost without a clear realization that it had been lost, and without much effort to pick it up again.[7]

In January 2001, almost one year after being lost in Bangkok, al-Mihdhar was on the radar screen again after being identified by a joint CIA-FBI investigation of the bombing of the USS *Cole*, the October 2000 attack on a U.S. guided missile destroyer in Yemen's Aden Harbor that killed 17 crew members and injured 39.[8] Even with this new knowledge the CIA did not renew its search for al-Mihdhar and did not make his identity known to the State Department (which presumably would have interfered with his plans to re-enter the United States).[9] Al-Mihdhar flew to New York City on July 4, 2001, on a new visa. As the 9/11 Commission reported, "No one was looking for him."[10]

On August 21, 2001, an FBI analyst who had been detailed to the CIA's Bin Laden unit finally made the connection and "grasped the significance" of Nawaf al-Hazmi and al-Mihdhar's visits to the United States. The Immigration and Naturalization Service was immediately notified. On August 22, 2001, the INS responded with information that caused the FBI analyst to conclude that al-Mihdhar might still be in the country.[11]

With the knowledge that the associate of a "major league killer" was possibly roaming free in the United States, the hunt by the FBI should have been on. The FBI certainly had a valid reason to open a case against these two individuals as they were connected to the ongoing USS *Cole* bombing investigation, the 1998 embassy bombing, and al-Qaeda.[12] On August 24, 2001, Nawaf al-Hazmi and al-Mihdhar were added to the State Department's TIPOFF[13] watchlist.[14]

Efforts to locate Nawaf al-Hazmi and al-Mihdhar initially foundered on confusion within the FBI about the sharing and use of data collected through intelligence versus criminal channels.[15] The search for al-Mihdhar was assigned to one FBI agent, his first ever counterterrorism lead.[16] Because the lead was "routine," he was given 30 days to open an intelligence case and make some effort to locate al-Mihdhar.[17] If more attention had been paid to these subjects, the location and detention of al-Mihdhar and Nawaf al-Hazmi could have derailed the 9/11 attack.[18]

Hiding in Plain Sight

The 9/11 terrorists did not take significant steps to mask their identities or obscure their activities. They were hiding in plain sight. They had P.O. boxes, e-mail accounts, drivers' licenses, bank accounts, and ATM cards.[19] For example, Nawaf al-Hazmi and al-Mihdhar used their true names to obtain California drivers' licenses and to open New Jersey bank accounts.[20] Nawaf al-Hazmi had a car registered, and his name appeared in the San Diego white pages with an address of 6401 Mount Ada Road, San Diego, California.[21] Mohamed Atta registered his red Pontiac Grand Prix car in Florida with the address 4890 Pompano Road, Venice.[22] Ziad Jarrah registered his red 1990 Mitsubishi Eclipse as well.[23] Fourteen of the terrorists got drivers' licenses or ID cards from either Florida or Virginia.[24]

The terrorists not only operated in plain sight, they were interconnected. They lived together, shared P.O. boxes and frequent flyer numbers, used the same credit card

The 9/11 terrorists did not take significant steps to mask their identities or obscure their activities.

numbers to make airline travel reservations, and made reservations using common addresses and contact phone numbers. For example, al-Mihdhar and Nawaf al-Hazmi lived together in San Diego.[25] Hamza al-Ghamdi and Mohand al-Shehri rented Box 260 at a Mail Boxes Etc. for a year in Delray Beach, Florida.[26] Hani Hanjour and Majed Moqed rented an apartment together at 486 Union Avenue, Patterson, New Jersey.[27] Atta stayed with Marwan al-Shehhi at the Hamlet Country Club in Delray Beach, Florida. Later, they checked into the Panther Inn in Deerfield Beach together.[28]

When Ahmed al-Nami applied for his Florida ID card he provided the same address that was used by Nawaf al-Hazmi and Saeed al-Ghamdi.[29] Wail al-Shehri purchased plane tickets using the same address and phone number as Waleed al-Shehri.[30] Nawaf al-Hazmi and Salem al-Hazmi booked tickets through Travelocity.com using the same Fort Lee, New Jersey, address and the same Visa card.[31] Abdulaziz al-Omari purchased his ticket via the American Airlines website and used Atta's frequent flyer number and the same Visa card and address as Atta (the same address used by Marwan al-Shehhi).[32] The phone number al-Omari used on his plane reservation was also the same as that of Atta and Wail and Waleed al-Shehri.[33] Hani Hanjour and Majed Moqed rented room 343 at the Valencia Hotel on Route 1 in Laurel, Maryland; they were joined by al-Mihdhar, Nawaf al-Hazmi, and Salem al-Hazmi.[34] While these are plentiful examples of the 9/11 terrorists' interconnectedness, even more connections existed.

Finding a Few Bad Guys

Interference with and detention of the right subset of the 9/11 terrorists might have derailed the plan.

In late August 2001 the FBI began to search for al-Mihdhar and Nawaf al-Hazmi.[35] The two might have been located easily even by a private investigator (PI). A PI would have performed a public records search using a service such as those provided by ChoicePoint or LexisNexis, perhaps both. These organizations aggregate public record data, assembling them into reports that simplify basic background investigations done by PIs, potential employers, potential landlords, and others. These databases include phone book data, driver's license data, vehicle registration data, credit header data, voter registration, property ownership, felony convictions, and the like. Such a search could have unearthed the driver's license, the car registration, and the telephone listing of Nawaf al-Hazmi and al-Mihdhar.[36]

Given the connections of Nawaf al-Hazmi and al-Mihdhar to terrorist activities overseas, the FBI, of course, could have sought subpoenas for credit card and banking information, travel information, and other business records. It could have conducted intensive surveillance under FISA, the Foreign Intelligence Surveillance Act, because the case involved a foreign power or an agent of a foreign power.[37] The FBI could not only have located these subjects but could have started to unravel their highly interconnected network, had it been pursuing available leads.

It is Monday morning quarterbacking, of course, to suggest that all 19 of the 9/11 hijackers could have been rolled up by the proper investigation. But interference with and detention of the right subset of the 9/11 terrorists might have "derailed the plan," as the 9/11 Commission concluded in its report.[38]

If our federal law enforcement and intelligence agencies needed anything, it was neither new technology nor more laws but simply a sharper focus and perhaps the ability to more efficiently locate, access, and aggregate information about specific suspects. They lacked this focus and capability—with tragic results.

Data Analysis and Data Mining

As we have seen, authorities could have and should have more aggressively hunted some of the 9/11 terrorists. If they had been hunted, they could have been found. Their web of connections would have led suffi-

ciently motivated investigators to information that could have confounded the 9/11 plot. Better interagency information sharing,[39] investigatory legwork in pursuit of genuine leads, and better training are what the 9/11 story most clearly calls for.

A number of policy changes followed the 9/11 attacks. The Intelligence Reform and Terrorism Prevention Act of 2004 revamped the nation's intelligence operations, and the USA-PATRIOT Act eased information sharing between investigators pursuing criminal and national security cases.

Data mining also gained some currency in national security and technology circles as a potential anti-terrorism tool,[40] though whether and to what extent it has been used are unclear. The Total Information Awareness program within the Department of Defense is widely believed to have contemplated using data mining, though the program's documentation is unclear.[41] The documentation discusses research on privacy-protecting technologies,[42] but Congress defunded the program in 2003 because of privacy concerns. However, the *National Journal* reported in February 2006 that research on "predict[ing] terrorist attacks by mining government databases and the personal records of people in the United States" has been moved from the Department of Defense to another group linked to the National Security Agency.[43]

In May 2004 the Government Accountability Office reported the existence of 14 data-mining programs, planned or operational, dedicated to analyzing intelligence and detecting terrorist activity, in the Departments of Defense, Education, Health and Human Services, Homeland Security, and Justice.[44] Ten of them were reported to use personal information. Of those, half use information acquired from the private sector, other agencies, or both.

"Data mining" is a broad and fairly loaded term that means different things to different people. Up to this point, discussions 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.

Several authorities have offered definitions or discussions of data mining that are important touchstones, though they still may not be sufficiently precise. In its May 2004 report, for example, the Government Accountability Office surveyed the literature and produced the following definition of data mining: "the application of database technology and techniques—such as statistical analysis and modeling—to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results."[45] In a January 2006 report, the Congressional Research Service said:

> Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees). Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction.[46]

Data mining is best understood as a subset of the broader practice of data analysis. 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.

The key goal—and challenge—is to produce not just more information but more *useful* information. "Useful information" is information that puts the analyst in a position to act appropriately in a given context. It is the usefulness of the result—the fact that it

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.

Attempting to use predictive data mining to ferret out terrorists before they strike would be a subtle but important misdirection of national security resources.

can be used effectively for a given purpose—that establishes the value of any given algorithm. The ultimate goal of data analysis is to discover knowledge.[47]

The term "predicate" is often used in law enforcement to refer to a piece of information that warrants further investigation or action. When a police officer sees a person attacking another with a knife, that is a sound basis, or predicate, for intervening by drawing his or her weapon and calling for a stop to the attack. When a police officer observes people appearing to "case" a store, that may be a predicate for making a display of authority or briefly questioning the people about their purposes. In Fourth Amendment law, probable cause to believe that information about a crime can be found in a particular place is a predicate for the issuance of a warrant to search that place.

Here is an example of a potential terrorism-related predicate: The combined facts that a particular person has been identified by an informant as having visited Afghanistan during June 2001 and participated in scuba training some years later, and that al-Qaeda plans to have divers mine cruise ships, may form a predicate for investigating the person or monitoring his or her communications.

In the first two examples discussed above—the knife attack and thieves casing a store—all the observations needed to establish a predicate for action were collected at once. Those are simple cases. Other than judging whether the response is proportional to the predicate, there is little need to parse them. But in the terror-suspect example, several observations made by different people at different times are combined to create the predicate. The fact that the person visited Afghan training camps might have come from an informant in Europe. The fact that he took scuba training might have come from business records in Corpus Christi, Texas. And the fact that al-Qaeda contemplated using scuba divers may have come from a computer captured in Pakistan. Because multiple observations are combined, this predicate can be said to result from data analysis. Data analysis brought information from diverse sources together to create new knowledge.

There are two loose categories of data analysis that are relevant to this discussion: subject based and pattern based.[48] Subject-based data analysis seeks to trace links from known individuals or things to others. The example just cited and the opportunities to disrupt the 9/11 plot described further above would have used subject-based data analysis because each of them starts with information about specific suspects, combined with general knowledge.

In pattern-based analysis, investigators use statistical probabilities to seek predicates in large data sets. This type of analysis seeks to find new knowledge, not from the investigative and deductive process of following specific leads, but from statistical, inductive processes. Because it is more characterized by prediction than by the traditional notion of suspicion, we refer to it as "predictive data mining."

The question in predictive data mining is whether and when it comes up with actionable information, with knowledge: suitable predicates for subsequent action. As we will discuss below, there are many instances when it does. But terrorism is not one. Attempting to use predictive data mining to ferret out terrorists before they strike would be a subtle but important misdirection of national security resources.

The possible benefits of predictive data mining for finding planning or preparation for terrorism are minimal. The financial costs, wasted effort, and threats to privacy and civil liberties are potentially vast. Those costs outstrip any conceivable benefits of using predictive data mining for this purpose.

Predictive Data Mining in Action

Predictive data mining has been applied most heavily in the area of consumer direct marketing. Companies have spent hundreds of millions if not billions of dollars imple-

menting and perfecting their direct marketing data-mining initiatives. Data mining certainly gives a "lift" to efforts to find people with certain propensities. In marketing, data mining is used to reduce the expense (to companies) and annoyance (to consumers) of unwanted advertising. And that is valuable to companies despite the fact that response rates to bulk mailings tuned by data mining improve by only single-digit percentages.

Consider how a large retailer such as Acme Discount Retail ("Acme Discount")—a fictional retailer trying to compete with Wal-Mart and Target—might use data mining: Acme Discount wants to promote its new store that just opened in a suburb of Chicago. It has many other stores and thousands of customers. Starting with the names and addresses of the top 1,000 Acme Discount customers, it contracts with a data broker to enhance what it knows about those customers. (This is known in database marketing as an "append" process.) Acme Discount may purchase magazine subscription and warranty card information (just to name a couple of likely data sources). Those sources augment what Acme Discount knows about its customers with such data points as income levels, presence of children, purchasing power, home value, and personal interests, such as a subscription to *Golf Digest*.

Thus, Acme Discount develops a demographic profile of what makes a good Acme Discount customer. For example, the ideal customer might be a family that subscribes to magazines of the *Vanity Fair* genre, that has two to four children, that owns two or fewer cars, and that lives in a home worth $150,000–$225,000. Acme Discount's next objective is to locate noncustomers near its new Chicago store that fit this pattern and market to them in the hope they will do business at the newly opened store. The goal is to predict as accurately as possible who might be swayed to shop at Acme Discount.

Despite all of this information collection and statistical analysis, the percent chance that Acme Discount will target someone willing to transact is in the low to mid single digits.[49] This means that false positives in marketers' searches for new customers are typically in excess of 90 percent.

The "damage" done by an imperfectly aimed direct-mail piece may be a dollar lost to the marketer and a moment's time wasted by the consumer. That is an acceptable loss to most people. The same results in a terror investigation would not be acceptable. Civil liberties violations would be routine and person-years of investigators' precious time would be wasted if investigations, surveillance, or the commitment of people to screening lists were based on algorithms that were wrong the overwhelming majority of the time.

Perhaps, though, more assiduous work by government authorities and contractors—using a great deal more data—could overcome the low precision of data mining and bring false positives from 90+ percent to the low single digits. For at least two related reasons, predictive data mining is not useful for counterterrorism: First, the absence of terrorism patterns means that it would be impossible to develop useful algorithms. Second, the corresponding statistical likelihood of false positives is so high that predictive data mining will inevitably waste resources and threaten civil liberties.

The Absence of Terrorism Patterns

One of the fundamental underpinnings of predictive data mining in the commercial sector is the use of training patterns. Corporations that study consumer behavior have millions of patterns that they can draw upon to profile their typical or ideal consumer. Even when data mining is used to seek out instances of identity and credit card fraud, this relies on models constructed using many thousands of known examples of fraud per year.

Terrorism has no similar indicia. 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

The statistical likelihood of false positives is so high that predictive data mining will inevitably waste resources and threaten civil liberties.

Without well-constructed algorithms based on extensive historical patterns, predictive data mining for terrorism will fail.

behavior indicates planning or preparation for terrorism.

Unlike consumers' shopping habits and financial fraud, terrorism does not occur with enough frequency to enable the creation of valid predictive models. Predictive data mining for the purpose of turning up terrorist planning using all available demographic and transactional data points will produce no better results than the highly sophisticated commercial data mining done today. The one thing predictable about predictive data mining for terrorism is that it would be consistently wrong.

Without patterns to use, one fallback for terrorism data mining is the idea that any anomaly may provide the basis for investigation of terrorism planning. Given a "typical" American pattern of Internet use, phone calling, doctor visits, purchases, travel, reading, and so on, perhaps all outliers merit some level of investigation. This theory is offensive to traditional American freedom, because in the United States everyone can and should be an "outlier" in some sense. More concretely, though, using data mining in this way could be worse than searching at random; terrorists could defeat it by acting as normally as possible.

Treating "anomalous" behavior as suspicious may appear scientific, but, without patterns to look for, the design of a search algorithm based on anomaly is no more likely to turn up terrorists than twisting the end of a kaleidoscope is likely to draw an image of the Mona Lisa.

Without well-constructed algorithms based on extensive historical patterns, predictive data mining for terrorism will fail. The result would be to flood the national security system with false positives—suspects who are truly innocent.

False Positives

The concepts of false positive and false negative come from probability theory. They have a great deal of use in health care, where tests for disease have known inaccuracy rates. A false positive, or Type I error, is when a test wrongly reports the presence of disease. A false negative, or Type II error, is when a test wrongly reports the absence of disease. Study of the false positive and false negative rates in particular tests, combined with the incidence of the disease in the population, helps determine when the test should be administered and how test results are used.

Even a test with very high accuracy—low false positives and false negatives—may be inappropriate to use widely if a disease is not terribly common. Suppose, for example, that a test for a particular disease accurately detects the disease (reports a true positive) 99 percent of the time and inaccurately reports the presence of the disease (false positive) 1 percent of the time. Suppose also that only one in a thousand, or 0.1 percent of the population, has that disease. Finally, suppose that if the test indicates the presence of disease the way to confirm it is with a biopsy, or the taking of a tissue sample from the potential victim's body.

It would seem that a test this good should be used on everyone. After all, in a population of 300 million people, 300,000 people have the disease, and running the test on the entire population would reveal the disease in 297,000 of the victims. But it would cause 10 times that number—nearly three million people—to undergo an unnecessary biopsy. If the test were run annually, every 5 years, or every 10 years, the number of people unnecessarily affected would rise accordingly.

In his book *The Naked Crowd*, George Washington University law professor Jeffrey Rosen discusses false positive rates in a system that might have been designed to identify the 19 hijackers involved in the 9/11 attacks.[50] Assuming a 99 percent accuracy rate, searching our population of nearly 300,000,000, some 3,000,000 people would be identified as potential terrorists.

Costs of Predictive Data Mining

Given the assumption that the devastation of the 9/11 attacks can be replicated,

some people may consider the investigation of 1 percent of the population (or whatever the false positive rate) acceptable, just as some might consider it acceptable for 10 people to undergo unnecessary surgery for every 1 person diagnosed with a certain disease. Fewer would consider a 5 percent error rate (or 15,000,000 people) acceptable. And even fewer would consider a 10 percent error rate (or 30,000,000 people) acceptable.

The question is not simply one of medical ethics or Fourth Amendment law but one of resources. The expenditure of resources needed to investigate 3,000,000, 15,000,000, or 30,000,000 fellow citizens is not practical from a budgetary point of view, to say nothing of the risk that millions of innocent people would likely be under the microscope of progressively more invasive surveillance as they were added to suspect lists by successive data-mining operations.

As we have shown, the unfocused, false-positive-laden results of predictive data mining in the terrorism context would waste national resources. Worse yet, the resources expended following those "leads" would detract directly from pursuing genuine leads that have been developed by genuine intelligence.

The corollary would be to threaten the civil liberties of the many Americans deemed suspects by predictive data mining. As Supreme Court precedents show, the agar in which reasonable suspicion grows is a mixture of specific facts and rational inferences. Thus, in *Terry v. Ohio*, the Supreme Court approved a brief interrogation and pat-down of men who appeared to have been "casing" a store for robbery.[51] An experienced officer observed their repeated, furtive passes by a store window; that gave him sufficient cause to approach the men, ask their business, and pat them down for weapons, which he found. The behavior exhibited by the men he frisked fit a pattern of robbery planning and did not fit any common pattern of lawful and innocent behavior. Any less correlation between their behavior and inchoate crime and the Court would likely have struck down the stop-and-frisk as a violation of the Fourth Amendment.

If predictive data mining is used as the basis for investigating specific people, it must meet this test: there must be a pattern that fits terrorism planning—a pattern that is exceedingly unlikely ever to exist—and the actions of investigated persons must fit that pattern while not fitting any common pattern of lawful behavior. Predictive data mining premised on insufficient pattern information could not possibly meet this test. Unless investigators can winnow their investigations down to data sets already known to reflect a high incidence of actual terrorist information, the high number of false positives will render any results essentially useless.

Predictive data mining requires lots of data. Bringing all the data, either physically or logically, into a central system poses a number of challenging problems, including the difficulty of keeping the data current and the difficulty of protecting so much sensitive data from misuse. Large aggregations of data create additional security risks from both insiders and outsiders because such aggregates are so valuable and attractive.

Many Americans already chafe at the large amount and variety of information about them available to marketers and data aggregators. Those data are collected from their many commercial transactions and from public records. Most data-mining efforts would rely on even more collections of transactional and behavioral information, and on centralization of that data, all to examine Americans for criminality or disloyalty to the United States or Western society. That level of surveillance, aimed at the entire citizenry, would be inconsistent with American values.

The Deceptiveness of Predictive Data Mining

Experience with a program that used predictive data mining shows that it is not very helpful in finding terrorists, even when abundant information is available. Using predic-

The unfocused, false-positive-laden results of predictive data mining in the terrorism context would waste national resources.

tive analysis—even in hindsight—the universe of "suspects" generated contains so many irrelevant entries that such analysis is essentially useless.

In his book *No Place to Hide*, *Washington Post* reporter Robert O'Harrow tells the story of how Hank Asher, owner of an information service called Seisint, concocted a way to fight back against terrorists in the days after September 11, 2001.

> Using artificial intelligence software and insights from profiling programs he'd created for marketers over the years, he told Seisint's computers to look for people in America who had certain characteristics that he thought might suggest ties to terrorists. Key elements included ethnicity and religion. In other words, he was using the data to look for certain Muslims. "Boom," he said, "32,000 people came up that looked pretty interesting."...
>
> In his darkened bedroom that night, he put the system through its paces over a swift connection to Seisint. "I got down to a list of 419 through an artificial intelligence algorithm that I had written," he recalled later. The list contained names of Muslims with odd ties or living in suspicious-seeming circumstances, at least according to Asher's analysis.[52]

Ultimately, Asher produced a list of 1,200 people he deemed the biggest threats. Of those, five were hijackers on the planes that crashed September 11, 2001.

Data mining is almost certain to fail when information about attackers and their plans, associates, and methods is not known.

What seems like a remarkable feat of predictive analysis is more an example of how deceptive hindsight can be. Asher produced a list of 9/11 terror suspects with a greater than 99 percent false positive rate—*after* the attack, its perpetrators, and their modus operandi were known.

The proof provided by the Seisint experience is not that there is a viable method in predictive analysis for finding incipient terrorism but that data mining of this type is almost certain to fail when information about attackers and their plans, associates, and methods is not known.

Conclusion

So how should one find bad guys? The most efficient, effective approach—and the one that protects civil liberties—is the one suggested by 9/11: pulling the strings that connect bad guys to other plotters.

Searching for terrorists must begin with actionable information, and it must follow logically through the available data toward greater knowledge. Predictive data mining always provides "information," but useful knowledge comes from context and from inferences drawn from known facts about known people and events.

The Fourth Amendment is a help, not a hindrance: It guides the investigator toward specific facts and rational inferences. When they focus on following leads, investigators can avoid the mistaken goal of attempting to "predict" terrorist attacks, an effort certain to flood investigators with false positives, to waste resources, and to open the door to infringements of civil liberties. That approach focuses our national security effort on developing information about terrorism plotters, their plans, and associates. It offers no panacea or technological quick fix to the security dilemmas created by terrorism. But there is no quick fix. Predictive data mining is not a sharp enough sword, and it will never replace traditional investigation and intelligence, because it cannot predict precisely enough who will be the next bad guy.

Since 9/11 there has been a great deal of discussion about whether data mining can prevent acts of terrorism. In fact, the most efficient means of detecting and preempting terrorism have been within our grasp all along. Protecting America requires no predictive-data-mining technologies.

Indeed, if there is a lesson to be learned from 9/11, it is not very groundbreaking. It is this: Enable investigators to efficiently dis-

cover, access, and aggregate relevant information related to actionable suspects. Period. Sufficient dedication of national resources to more precisely "pull the strings" offers the best chance of detecting and preempting future acts of terrorism.

Notes

1. Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001 (USA PATRIOT Act), Pub. L. No. 107-56 (Oct. 12, 2001).

2. Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. No. 108-458 (Dec. 17, 2004).

3. National Commission on Terrorist Attacks upon the United States, *The 9/11 Commission Report*, 2004, p. 268 (hereinafter *9/11 Commission Report*).

4. Ibid., p. 181.

5. Ibid., pp. 181–82.

6. Ibid., p. 181.

7. Ibid., p. 266.

8. Ibid., p. 266.

9. Ibid., pp. 266–67.

10. Ibid., p. 269.

11. Ibid., p. 270.

12. Ibid., p. 271.

13. The TIPOFF database contains a list if foreigners who will be denied a U.S. visa.

14. *9/11 Commission Report*, p. 270.

15. Ibid., p. 271.

16. Ibid., p. 288.

17. Ibid., p. 271.

18. Ibid., p. 272.

19. Tim Golden et al., "A Nation Challenged: The Plot," *New York Times*, September 23, 2001.

20. *9/11 Commission Report*, p. 539, n 85.

21. Ibid.; and Jane Black, "Don't Make Privacy the Next Victim of Terror," *BusinessWeek Online*, http://www.businessweek.com/bwdaily/dnflash/oct2001/nf2001104_7412.htm.

22. Dan Eggen et al., "The Plot: A Web of Connections," *WashingtonPost.com*, October 4, 2001, http://www.washingtonpost.com/wp-srv/nation/graphics/attack/investigation_24.html (hereinafter "Web of Connections").

23. "Web of Connections."

24. Ibid.

25. Ibid.

26. Ibid.

27. Ibid.

28. Ibid.

29. Ibid.

30. Ibid.

31. Ibid.

32. Ibid.

33. Ibid.

34. Ibid.

35. *9/11 Commission Report*, pp. 271–72.

36. Ibid., p. 539, n 85.

37. 50 U.S.C. § 1804.

38. *9/11 Commission Report*, p. 272.

39. Ibid., p. 271.

40. Arshad Mohammed and Sara Kehaulani Goo, "Government Increasingly Turning to Data Mining," *Washington Post*, June 15, 2006, http://www.washingtonpost.com/wp-dyn/content/article/2006/06/14/AR2006061402063.html.

41. See Defense Advanced Research Projects Agency, Information Awareness Office, "Report to Congress Regarding the Terrorism Information Awareness Program," May 30, 2003, pp. 7–8, 17, A-4, A-14, A-15 (referring variously to "discovery of . . . patterns of activity"; "ability to automatically learn patterns"; "training software algorithms to recognize patterns"; and "developing technology to . . . suggest previously unknown but potentially significant patterns"), http://foi.missouri.edu/totalinfoaware/tia2.pdf.

42. Ibid., pp. 6–7.

43. Shane Harris, "TIA Lives On," *National Journal*, February 23, 2006, http://nationaljournal.com/about/njweekly/stories/2006/0223nj1.htm.

44. Government Accountability Office, "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548, May 3004.

45. Ibid., p. 1.

46. Jeffrey W. Seifert, "Data Mining and Homeland Security: An Overview," Congressional Research Service, updated January 27, 2006 (Order Code RL31798). See also K. A. Taipale, "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data," *Columbia Science and Technology Law Review* 22–23 (2003), http://papers..ssrn.can/abstract=S467827.

47. The major annual conference on data mining is called "KDD," for Knowledge Discovery and Data Mining. See http://www.acm.org/sigs/sigkdd/kdd2006.

48. See Martha Baer et al., *SAFE: The Race to Protect Ourselves in a Newly Dangerous World* (New York: Harper Collins 2005), p. 331; and Mary DeRosa, "Data Mining and Data Analysis for Counterterrorism," Center for Strategic and International Studies, March 2004.

49. Direct marketing results are dependent upon many factors such as industry and offer. For example, offering a consumer a loss-leader discount raises response rates. Despite billions invested and unprecedented access to U.S. consumer behavior data, current direct marketing response rates industrywide range from 5.78 percent for telephone solicitation to 0.04 percent for direct response television. Direct Marketing Association, "DMA Releases New Response Rate Report," news release, October 17, 2004, http://www.the-dma.org/cgi/dispnewsstand?article=2891.

50. Jeffrey Rosen, *The Naked Crowd* (New York: Random House, 2004), pp. 104–7.

51. 392 U.S. 1 (1968).

52. Robert O'Harrow Jr., *No Place to Hide* (New York: Free Press, 2005), pp. 98, 102.

CATO INSTITUTE
1000 Massachusetts Ave., N.W.
Washington, D.C. 20001

Statement of Leslie Harris
Executive Director
Center for Democracy & Technology*

before the
Senate Committee on the Judiciary

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

January 10, 2007

Mr. Chairman, Senator Specter, and Members of the Committee:

Good morning, and thank you for the opportunity to testify at this important hearing. In our testimony this morning, we want to emphasize the following six points:

- Terrorism poses a grave threat to our nation. To prevent further terrorist attacks, the government should use information technology to better share and better analyze the ocean of information at its disposal in this digital age.

- Both national security and the protection of civil liberties require that the technology be used only when it is demonstrably effective and then only within a framework of accountability, oversight and protection of individual rights.

- "Data mining" broadly defined is the use of computer tools to extract useful knowledge from large sets of data. It is in the abstract neither good nor bad. Rather, the questions are: What kind of data mining should the government use, for what purposes, with what consequences for individuals, under what guidelines, and subject to what oversight, auditing and redress?

- There is a vast difference both in terms of proven effectiveness and in terms of risks to privacy and due process between pattern-based data mining, especially when based on hypotheticals, and subject-based data mining, such as where the government is starting with some particularized suspicion.

- The threshold question for the application of any technology is efficacy. So far, there has been no evidence of the effectiveness of the broad forms of predictive data mining that have been proposed and deployed by the government. Unless and

---

* The Center for Democracy & Technology (CDT) is a non-profit public interest organization dedicated to promoting privacy and other democratic values for the new digital communications media. Among other activities, CDT coordinates the Digital Privacy and Security Working Group (DPSWG), a forum for computer, communications, and public interest organizations, companies and associations interested in information privacy and security issues.

until a particular application can be shown to be an effective tool for counterterrorism, and appropriate safeguards to protect the privacy and due process rights of Americans are put in place, the government should not deploy pattern-based data mining as an antiterrorism tool.

- Finally, the technological and legal context for data collection and analysis has changed dramatically in recent years. Technology has far outstripped existing privacy protections at the very time that legal standards for government access to data have been lowered (or ignored by Executive fiat). Core laws like the Privacy Act are inadequate and almost irrelevant to data mining.

In light of these considerations, we offer below a series of recommendations to Congress, focused on oversight, accountability, and due process.

I. The Rights at Stake: Privacy and Due Process

It is very important to start by defining what we mean by "privacy." Information privacy is not merely about keeping personal information confidential. In the context of a function like data mining, privacy is equally about due process: how to make fair decisions about people.

It is well established by U.S. Supreme Court cases, the federal Privacy Act, and other privacy laws like the Fair Credit Reporting Act (FCRA) and the Health Insurance Portability and Accountability Act (HIPAA) that individuals retain a privacy (or due process) interest in information about themselves even after they have disclosed it in the course of a commercial or governmental transaction. Our interest in the fair use of information to make decisions about us extends even to data that is publicly available: if that information is used to make decisions that can have adverse consequences, then we should have a right to know about the use of that information and an opportunity to respond to information that is inaccurate or misleading.

The term "Fair Information Practices" (FIPs) best describes the values at stake with regard to data mining. First articulated in the 1970s, these principles govern not just the initial collection of information, but also its use. The "Fair Information Practices" have been embodied in varying degrees in the Privacy Act, FCRA, and the other "sectoral" federal privacy laws that govern commercial uses of information. The concept of FIPs has remained remarkably relevant despite the dramatic advancements in information technology that have occurred since these principles were first developed.

While applying these principles to the current data landscape and the context of counterterrorism poses challenges, FIPs provide a remarkably sound basis for analyzing the issues associated with data mining: what information is being collected, how long will it be kept, how accurate and reliable is the information, how will an individual be able to correct erroneous information, what are the redress and enforcement mechanisms?

II. What Are the Kinds of Data Mining and What Risks Do They Pose?

Policy discussions about data mining often suffer from a lack of clarity about key terms and concepts. An informed policy discussion requires an understanding of the different ways the term "data mining" is used and the risks to privacy and due process associated with different applications of data analysis tools. For simplicity's sake, it may be useful to identify two different kinds of data mining: subject-based, which seeks information about a particular individual who is already under suspicion; and pattern-based (sometimes referred to as predictive) data mining, which seeks to find a pattern, anomaly or signature among oceans of personal transactional data.[1] As a general matter, the value of subject-based approaches is more readily apparent, and there are fewer privacy concerns associated with data searches that begin with particularized suspicion.[2] Throughout our testimony today, we focus on pattern-based data mining in the counterterrorism context.

Pattern-based data mining does not begin with any particularized suspicion. Rather, it searches large databases containing transactional information on the everyday activities of millions of people in an attempt to determine the level of risk associated with individuals or to find patterns that may indicate terrorist behavior. Some proponents of data mining have suggested that the searches may be based on no more than a hypothetical set of assumptions about how terrorists behave.

In the counterterrorism field, we must be careful in the adoption of any data analysis tool. The consequences to individuals of being mistakenly designated as a possible terrorist or an associate of terrorists can be devastating and can include arrest, deportation, loss of a

---

[1] Strictly speaking, some would say that only the latter is data mining. The Government Accountability Office, after surveying the technical literature, focused specifically on what we call "pattern-based" data mining when it defined data mining as "the application of database technology and techniques—such as statistical analysis and modeling—to uncover hidden patterns and subtle relationships in data and to infer rules that allow for the prediction of future results." GAO, "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548 (May 2004). However, in public policy circles, the term data mining has been broadly used. For more on the varities of data mining from a policy perspective, see Mary DeRosa, "Data Mining and Data Analysis for Counterterrorism," CSIS (March 2004); James X. Dempsey and Lara M. Flint, "Commercial Data and National Security," *The George Washington Law Review*, Vol. 72, No. 6 (August 2004).

[2] As others have noted, "the power of data mining technology and the range of data to which the government has access have contributed to blurring the line between the subject- and pattern-based searches . . . [e]ven when a subject-based search starts with a known suspect, it can be transformed into a pattern-based search as investigators target individuals for investigation solely because of their connection with the suspect." U.S. Department of Defense, Report of the Technology and Privacy Advisory Committee (TAPAC), "Safeguarding Privacy in the Fight Against Terrorism," p. 45 (March 2004) <http://www.cdt.org/security/usapatriot/20040300tapac.pdf>.

job, more intrusive investigation, discrimination, damage to reputation and a lifetime of suspicion, with little or no opportunity for redress or correction of errors. False leads also have serious consequences for national security, diverting resources from true threats.

Currently, there is little evidence of the efficacy of pattern-based data mining in the antiterrorism context. Indeed, there is substantial reason to believe that the technique will not prove useful in identifying terrorists, but will instead lead to significant violations of civil liberties. As experts have explained, the sample of known terrorists whose behavior can be studied is statistically insignificant to identify an unusual or unique pattern of behavior.[3] Any pattern-based search based on characteristics drawn from such a sample will make it difficult to separate the "noise" of innocent behavior from the "signal" of terrorist activities, leading innocent behavior to be viewed as suspicious.[4] When unproven pattern data mining algorithms are applied to the records of millions of people, the false positive rate can be higher than 99%, potentially subjecting large numbers of law abiding citizens to a range of consequences, often with little recourse. The danger of false positives is exacerbated by well-recognized problems with data quality, not only in government databases but also in data drawn from commercial sources. However, under current rules, once the data is collected and analyzed, there are few if any effective controls within the government to prevent inaccurate information from being widely disseminated and used for other purposes.

III. The Changed Legal and Technological Landscape

In the past, the government by and large collected data on one person at a time (i.e., with particularity), either in the course of administering a government program or where there was some suspicion that a person was engaged in criminal conduct, terrorism or intelligence activity. The government was authorized to keep this data for long periods of time, and to retrieve, share and analyze it for compatible purposes without serious controls. However, before it could take action based on that data, the government was bound by procedural due process principles of notice and an opportunity to respond. In the traditional data environment, the greater the consequences for the individual, the greater the due process requirements. For example, the criminal due process standards in the Bill of Rights place the burden of proof on the government and force it to disclose all of its evidence to the accused, for challenge.

Now, in contrast, Section 215 of the PATRIOT Act, the expanded National Security Letter authorities, the growing implications of the Supreme Court's "business records" decisions (which place most commercial data outside the protections of the Fourth Amendment), the President's claims of inherent power, and the nature of technology itself can result in the wholesale collection of data and databases by the government without particularized suspicion. Yet the traditional rules on storage and use remain in

---

[3] Jeff Jonas and Jim Harper, "Effective Counterterrorism and the Limited Role of Predictive Data Mining," Cato Institute (December 11, 2006).

[4] DeRosa, *supra* note 1, at 15.

place, permitting the government to keep that data forever and to go back to it for further analysis (e.g., data mining) with little legal constraint.

Meanwhile, many traditional limits on information sharing have been removed. The wall between intelligence and law enforcement is down. The Executive Branch is moving forward with development of the Information Sharing Environment.[5] State and local information sharing and analysis centers are proliferating. The Justice Department is developing its own information sharing system to make millions of law enforcement investigatory records available to state and local police.[6] The Administration has been expansive in exempting law enforcement and intelligence systems from the Privacy Act.[7]

We must stress that information sharing to prevent terrorism and for other governmental purposes is generally desirable. But, especially in the counterterrorism context, a major shift in the data collection and use landscape is taking place without a suitable privacy and due process framework. The detailed guidelines called for by the Markle Foundation Task Force on National Security in the Information Age, the Defense Secretary's Technology and Privacy Advisory Committee and others have not been issued yet, and existing privacy laws are not up to the task. Yet the government is moving ahead with screening and risk assessment programs. At the same time, the government is claiming the power to make highly consequential decisions about people, cut off from the normal checks and balances: for example, deporting immigrants on the basis of secret evidence, holding individuals for extended periods as "material witnesses," incarcerating hundreds of people at Guantanamo and elsewhere without fundamental due process, and even asserting the power to imprison citizens without the protections of the criminal justice system.

The impact of this "perfect storm" of technological innovation, increased government power and outdated legal protections is well illustrated by the government's recent acknowledgement that it is, through its "Automated Targeting System," collecting travel records on all American citizens entering and leaving the country, assigning risks scores to those citizens, and keeping the records for 40 years. In the current legal environment (we address the failure of the Customs and Border Patrol to comply with the Privacy Act later in our testimony), all those records (including the secret risk score) may be freely shared with other federal agencies engaged in a wide range of activities and also accessed by various state and local law enforcement agencies. In this context, a risk score developed for border screening purposes could easily migrate to other uses (years after

---

[5] The Information Sharing Environment (ISE) was mandated by Section 1016 of the Intelligence Reform and Terrorism Prevention Act of 2004. In November 2006, the ISE Program Manager issued an Implementation Plan for the ISE, marking an important milestone in its development.

[6] Dan Eggen, "Justice Dept. Database Stirs Privacy Fears," *Washington Post* (December 26, 2006).

[7] *See, e.g.*, Transportation Security Administration, Notice of Privacy Act System of Records, 68 Fed. Reg. 2101 (January 15, 2003); Department of Homeland Security, Notice of Privacy Act System of Records, 71 Fed. Reg. 64543 (November 2, 2006).

the citizen was determined not to be a threat) and result in a host of consequences where the individual would find it impossible to respond.

The Privacy Act of 1974 was intended to subject government agencies that collect personally identifiable information to the Fair Information Practices. It was intended to require notice to and consent from individuals when the government collects and shares information about them, give citizens the right to see whatever information the government has about them, and hold government databases to certain accuracy standards. While those practices remain highly relevant today, the Act is increasingly impotent to address the modern data sharing environment. For one, the Act's exemptions for law enforcement and intelligence data have been interpreted in a manner that neuters the Act. Second, the Act's protections only apply to federal "systems of records." That means that the government can bypass the Privacy Act by accessing existing private sector databases, rather than collecting the information itself. Currently, when it accesses commercial databases, the government need not ensure (or even evaluate) the accuracy of the data; it need not allow individuals to review and correct the data; and the government is not limited in how it interprets or characterizes the data.

Finally, and most remarkably, unless the courts and Congress respond, the legal and technological changes of the past decade could spell the effective end of key protections associated with the Fourth Amendment. Traditionally, as we all know, to search the intimate details of one's life the government required a judicial warrant, issued on a finding of probable cause to believe that a specific crime was being committed and naming with particularity the person or place to be searched and the items to be seized. In the 1970s, before the digital revolution and all it has entailed for the creation of electronic databases about our daily lives, the Supreme Court held that the Fourth Amendment does not apply to personal information contained in records held by third parties, with the result that the government could acquire it without meeting Fourth Amendment requirements of probable cause, particularity and notice.[8]

CDT questions the continued viability of these business records cases, for they were decided, by and large, in the context where the government was still collecting information one person at a time, usually in the course of criminal investigative activity where the individual would eventually have a robust set of due process protections. And previously, although the government could keep that data and retrieve it and use it in subsequent investigations, its ability to do so was severely limited by practical realities of incompatible data formats and limited search technology. In today's data mining context, the government is accessing entire buckets of data without a warrant and without particularized suspicion – some by purchase or subscription, some from files generated in the course of other government activities, and some by the forced disclosure of datasets using NSLs or other instruments. If this information on presumptively innocent people, having been acquired without Fourth Amendment protections, can be kept forever and analyzed at will without probable cause or individualized suspicion, what would be left of the Fourth Amendment?

---

[8] *U.S. v. Miller*, 425 U.S. 435 (1976); *Smith v. Maryland*, 442 U.S. 735 (1979).

CDT believes that the business records cases are inapplicable to the modern data environment and we are at the beginning of a long project to urge the courts and legislatures to re-examine them. The Supreme Court itself has made it clear in a related context that the law must advance with the technology to ensure the continued viability of the Fourth Amendment.[9] We also believe that other trends in Supreme Court rulings indicate that the analysis of opaque datasets are "searches" for Fourth Amendment purposes, just the search of a computer lawfully in the hands of the government is itself a separate "search" under the Fourth Amendment. For all of these reasons, we believe that the automated, pattern-based analysis of massive databases should be recognized as a search within the definition of the Constitution. Even if, for the sake of argument, the initial collection of the information as part of a database is not considered a search for Fourth Amendment purposes, once the government applies analytic techniques to extract from it meaning not readily apparent on the surface, the law should, at a minimum, consider that analysis a new search that requires procedural protections.[10] For now, the law is not there. Hence, there are huge gaps in privacy law, and Congress needs to respond.

-- **Comparison With the Commercial Sector**

While there is no comprehensive privacy law controlling the collection and use of personally identifiable information by the private sector, the private sector is still subject to more robust privacy protections than the government. Sector-specific privacy rules such as the Fair Credit Reporting Act place comparatively strict controls on private entities engaging in profiling or risk assessment.[11] When commercial data analysis could have adverse implications for a person's credit, insurance or employment, the private sector is under a legal obligation to use only accurate data, individuals have a right to access and challenge data about them, and individuals must be given notice and an opportunity to respond before adverse action is taken.

Furthermore, while the private sector uses pattern-based data mining to detect fraud, it has a large baseline of known frauds that can be used to develop and constantly refine risk assessment models. In contrast, agencies searching for a terrorist signature have a very small sample set on which to base their predictions.

---

[9] *See Kyllo v United States,* 533 U.S. 27 (2001).

[10] Lee Tien, "Privacy, Technology and Data Mining," 30 Ohio N.U. L. Rev. 389 (2004).

[11] Sectoral privacy laws that apply to personal data held by the private sector include, *inter alia*, the Fair Credit Reporting Act of 1970, 15 U.S.C. § 1681 (sets out rights for consumers with respect to credit information), the Family Educational Rights and Privacy Act of 1974, 20 U.S.C §1232g (governs access to personally identifiable information in educational records held by federally funded educational institutions); the Health Insurance Portability and Accountability Act of 1996, P.L. 104-191§ 264 (requires issuance of a privacy rule for individually identifiable health information); and the Right to Financial Privacy Act of 1978, 12 U.S.C.§ 3414 (sets out procedures for the federal government's access to financial institutions customer records).

The contrast is striking. While the private sector is subject to strict rules for at least some of its data mining activities, we have not yet devised a suitable set of rules for government data mining, where constitutional liberties are at stake and the consequences of error are much higher.

IV. What We Know So Far About Government Data Mining Illustrates the Need for Closer Oversight and Control

Little is known about the full extent of pattern-based data mining for counterterrorism and homeland security. While Congress has broadly authorized collection of data under extraordinarily low standards in the USA Patriot Act and authorized data sharing among law enforcement and intelligence agencies, our understanding of the data mining activities that these changes in the law have encouraged remains limited. Since the disclosure of the existence of the Total Information Awareness Program ("TIA") in 2002, there has been a steady stream of revelations about other data mining programs that raised concerns about privacy and efficacy, including CAPPS II.[12] With each new revelation, Congress has scrambled to respond, ultimately de-funding some elements of the TIA program[13] and postponing the deployment of the CAPPS II or Secure Flight airline passenger screening program until the GAO reported that the system had met certain reliability and privacy requirements.[14]

Yet, notwithstanding public and Congressional discomfort with these programs, they continue to proliferate without any apparent controls. Just last month, the Customs and Border Patrol ("CBP") acknowledged that, without notice and in violation of the Privacy Act, it has been using the Automatic Targeting System (ATS), which was designed to screen shipping cargo, to conduct "risk assessments" on tens of millions of travelers,

---

[12] In 2004, the GAO reported 199 data mining efforts, of which 68 were planned and 131 were operational.[12] The programs spanned 52 agencies and departments. Out of all 199 data mining efforts identified, 122 used personal information. The uses of data mining included improving service or performance, detecting fraud, waste, and abuse, analyzing scientific and research information, managing human resources, detecting criminal activities or patterns, and analyzing intelligence and detecting terrorist activities. GAO, "Data Mining: Federal Efforts Cover a Wide Range of Uses," GAO-04-548 (May 2004). For detailed descriptions and analysis of current antiterrorism and homeland security data mining programs, see "Data Mining and Homeland Security: An Overview." Congressional Research Service (January 27, 2006); "Survey of DHS Data Mining Activities," Department of Homeland Security Office of Inspector General (August 2006); "Privacy: Total Information Awareness Programs and Related Information Access Collection and Protection Laws" Congressional Research Service (February 14, 2003).

[13] Congressional action did not actually end many of the program's components; they moved into classified environments. Shane Harris, "TIA Lives On," National Journal (February 26, 2006) p. 66.

[14] Section 514, Department of Homeland Security Appropriations Act, 2007, Pub. L. 109-295; Section 522, Department of Homeland Security Appropriations Act 2005, Pub. L. 108-334.

including U. S. citizens. These "risk assessments" determine whether individuals will be subject to more invasive searches. No Privacy Act notice was issued before the focus of the massive data program was turned towards individuals nor was a Privacy Impact Assessment conducted before initiating the program as required by the E-Government Act of 2002.[15] An "after the fact" privacy assessment fails utterly to address the risks posed by the system.

But for the recent ATS Privacy Act notice, which boldly asserts unprecedented uses of the "routine use" exception, sweeping exemptions for law enforcement and intelligence investigations, and wide sharing of the data for a wide variety of uses wholly unrelated to border security, it is unclear whether or when Congress would have been made aware of the program. The danger to the rights of Americans under this program is self-evident. It is not hyperbolic to assume that the data---including the secret risk scores--- will find its way through the government and down to the state and local where it can easily be abused.

VI. Recommendations: What Congress and the Executive Can Do to Create a More Balanced Framework for Data Mining

A. Transparency and Congressional Oversight

Non-partisan congressional oversight is one of the pillars of a system of checks and balances. Congress has a critical role to play in ensuring that pattern-based data mining programs are effective and protect civil liberties. The first step, of course, is for Congress to get a comprehensive and accurate picture of the data mining activities of the federal government. While Congress, on its own and through the GAO, has conducted some oversight of data mining and has used that oversight to impose some constraints on particular programs through the budget process, Congress' response has largely been reactive, driven by revelations about excesses in particular programs rather than by facts developed during comprehensive and consistent oversight. Congress and – to the extent possible – the American people need to know what programs are being developed and deployed, whether those programs are likely to be effective, and what risks those programs pose to the rights of the American people. Congress should hold public oversight hearings with testimony from the Executive Branch, and should conduct annual reviews of data mining programs and issue public reports on the effectiveness of data mining in counterterrorism programs and its impact on privacy and other civil liberties.

As a first step toward developing a more balanced framework for data mining, CDT believes that the relevant agencies should be required by law to report in detail on pattern-based data mining programs that are being developed or deployed, and to provide assessments of each program's efficacy and impact on civil liberties.

---

[15] E-Government Act of 2002 [H.R. 2458] Pub. L. 107-347 (December 17, 2002).

B. Prior Congressional Authorization for Data Mining Programs

CDT believes that Congress should go further and expressly limit deployment of pattern-based data mining in law enforcement and antiterrorism contexts, by requiring an authorization based on a showing of effectiveness before a program is launched against U.S. citizens. In essence, we are proposing that the language Congress, on a bi-partisan basis, has adopted and annually renewed since FY 2005 for Secure Flight be applied government-wide. Under the approach we are proposing, research and development would be permitted without express prior authorization, under the careful oversight of Congress. In addition, as Congress mandated for Secure Flight, pattern-based programs should not be authorized until and unless there is in place a set of guidelines for data sharing and mining that protect privacy and ensure due process. While it is the job of the Executive Branch in the first instance to adopt adequate government-wide guidelines, that job has not yet been accomplished. In the absence of detailed and comprehensive Executive Branch guidelines, Congress may need to step in and legislate guidelines.

C. The Elements of Effective Guidelines

The elements of a set of robust and workable guidelines for information sharing and analysis have already been outlined in specific laws adopted by Congress and in leading studies, notably the three reports of the Markle Foundation Task Force on National Security in the Information Age.[16]

Congress has already legislated on some of the elements of a sound framework for data analysis in the limitations it placed on implementation of the Secure Flight passenger screening system[17] and in the rules it established for improvements in the government's terrorist "watch lists."[18] Drawing upon these laws, the Markle Task Force reports, and experiences in the commercial sector, one can develop a detailed set of guidelines that include the following elements:

- A concept of sharing that leaves information with the originator, using directories and search techniques that permit discovery and sharing of relevant information but minimize unnecessary transfers of data to central repositories.
- Strong data quality standards, including minimum standards for watchlists, and other procedures to ensure that the databases the government uses to establish the identity of individuals or make assessments about individuals are sufficiently

---

[16] "Mobilizing Information to Prevent Terrorism: Accelerating Development of a Trusted Information Sharing Environment" (July 13, 2006); "Creating a Trusted Network for Homeland Security (December 2, 2003); "Protecting America's Freedom in the Information Age (October 7, 2002), available at http://www.markletaskforce.org/.

[17] Section 514, Department of Homeland Security Appropriations Act, 2007, Pub. L. 109-295; Section 522, Department of Homeland Security Appropriations Act 2005, Pub. L. 108-334.

[18] Section 4012(c), Intelligence Reform and Terrorism Prevention Act of 2004, Pub. L. 108-548, 118 Stat. 3638, 3718.

accurate and reliable that they will not produce a large number of false positives or unjustified adverse consequences.

- Corrective mechanisms, including assessments of the reliability of commercial databases and automated mechanisms that can identify and correct errors in shared data, with responsibility on both the originator and the recipient of data.
- Access controls, security measures and permissioning technologies that can protect against improper access to personal information, including the ability to restrict access privileges so that data can be used only for a particular purpose, for a finite period of time, and by people with the necessary permissions.
- Automated and tamper-proof audit trails that can protect against misuse of data, improve security, and facilitate oversight.
- Redress mechanisms that allow individuals to respond when they are about to face adverse consequences based on information. This includes the right to challenge inaccurate information.
- Effective oversight of the use and operation of the system, including privacy officers with sufficient powers and resources to enforce the guidelines.

While technology is no substitute for policy, various commercially-available technologies can help implement and enforce these policies. Auditing technology can provide built-in recordation and documentation capabilities to track how information is used and shared. Technologies can help assure that information is up-to-date. Software can ensure that information is updated regularly and that it is unusable after a certain date if not refreshed. Other technology can permit users to track where information came from and who received it and alert users if the original data is subsequently disproved or corrected. Anonymization technologies can minimize unnecessary disclosure of personal information when not needed.

D. Apply Fair Information Practices to Commercial Databases Accessed For Pattern-Based Data Mining

Congress should legislate to ensure that commercial databases accessed for data mining are subject to strong privacy rules. Congress should make clear that the Privacy Act applies whether the government is creating its own database or acquiring access to a database from a commercial entity. In addition, Congress should require Privacy Impact Assessments for the acquisition of commercial databases. Section 208 of the E-Government Act of 2002 already requires a PIA if the government initiates a new "collection" of information. The same process should apply when the government acquires access to a commercial database containing the same type of information that would be covered if the government itself were collecting it.

In addition, Congress should require the government to perform an accounting of private sector databases before using them and to publish in the Federal Register a description of the database, the name of the entity from which the agency obtained the database and the amount of the contract for use of the database. Agencies should further be required to adopt regulations that establish fair information practices including a process for redress. Finally, Congress should require agencies to incorporate provisions into their contracts

with commercial entities provisions that provide for penalties when the commercial entity sells information to the agency that the commercial entity knows or should know is inaccurate or when the commercial entity fails to inform the agency of corrections or changes to data in the database.[19]

These approaches that have been proposed strike a balance between the government's need for information and the privacy interests of individuals. Adapting the Privacy Act and Fair Information Principles to government uses of commercial databases would go a long way toward closing the unintended gap in privacy protection that exists under the current law.

E. Strong Internal Mechanisms for Accountability and Oversight

Congress has created Chief Privacy Officers for the Departments of Homeland Security and Justice and for the office of the Director of National Intelligence. The independence and authority of these officers should be improved. If taken seriously, Privacy Act notices and Privacy Impact Assessments can help in raising and mitigating privacy concerns surrounding the government's use of personal information. Inspectors General should also have a role to play. Inspectors General, in particular, provide a critical internal ability to identify civil liberties violations, and should regularly review agency actions to assess their privacy implications.

VI. Conclusion

The Center for Democracy and Technology appreciates the opportunity to present its views on government data mining. Our nation is at a critical moment on this issue. As the ATS revelations indicate, pattern-based data mining is moving forward in the Executive Branch without a legal framework that will protect the privacy and due process rights of Americans. Congress needs to ensure that the proper legal and policy framework is in place before these programs move forward, and limit their deployment to those with proven effectiveness. Oversight and accountability, done right, will benefit both national security and civil liberties. Checks and balances result in clear lines of responsibility, well-allocated resources, protection against abuse, and the ability to evaluate and correct past mistakes. Appropriate, well-implemented accountability mechanisms will help to ensure that systems are effective as well as protective of due process.

---

[19] A number of bills were proposed in the 108th and 109th Congresses that incorporate many of these concepts. For example, S.1484, the "Citizens Protection in Federal Databases Act," sponsored by Sen. Wyden in the 108th Congress; S.1789, the "Personal Data Privacy and Security Act of 2005," sponsored by Sens. Leahy and Specter; and S.1169, "The Federal Agency Data Mining Reporting Act of 2005," sponsored by Sens. Feingold, Sununu, Leahy, Akaka, Jeffords and Wyden.

U.S. Department of Justice
Office of Legislative Affairs

Office of the Assistant Attorney General

Washington, D.C. 20530

January 12, 2007

The Honorable Patrick J. Leahy
Chairman
Committee on the Judiciary
United States Senate
Washington, D.C. 20510

Dear Mr. Chairman:

This is in reference to the January 10, 2007 Senate Judiciary Committee hearing, entitled "Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs, during which you requested that your January 10, 2003, letter regarding "data mining" operations, practices, and policies at the Department of Justice be made part of the record. You indicated that the Department never responded. The Department did, however, respond in a letter dated June 8, 2004, which is enclosed for your reference. We request that this letter also be made part of the hearing record. Please do not hesitate to contact this office if we may be of assistance on this matter.

Sincerely,

Richard A. Hertling
Acting Assistant Attorney General

Enclosure

cc: The Honorable Arlen Specter
Ranking Minority Member

U.S. Department of Justice
Office of Legislative Affairs

Office of the Assistant Attorney General

Washington, D.C. 20530

June 8, 2004

The Honorable Patrick Leahy
Ranking Minority Member
Committee on the Judiciary
United States Senate
Washington, D.C. 20510

Dear Senator Leahy:

This responds to your letter, dated January 10, 2003, to the Attorney General regarding "data mining" operations, practices, and policies at the Department of Justice. We apologize for the delay in our response to you. An identical letter is being sent to your colleagues who signed your letter.

1. Data-Mining Operations Underway Within the Department of Justice

a. **Please identify any private or proprietary databases obtained or being used by the Department of Justice for data-mining or pattern recognition as well as any databases from government agencies outside DOJ being used for such purposes.**

ANSWER: The FBI does not have a practice of obtaining other parties' databases in whole. To aid in its investigations, FBI employees often have access to outside databases, either with the ability to log on and perform a search or with the ability to request a data extract. For example, an FBI employee may have a Lexis/Nexis userid and the ability to search for information about specific persons, organizations, places, or events. Or, an employee might be working on a project in which it is appropriate to request a CD-ROM with a list of "absconders" from the Bureau of Immigration and Customs Enforcement. The FBI also has access to a number of other databases from non-DOJ components of the intelligence community at the classified level. A listing of all the classified databases that are available through Intelink, Intelink-S, and CT-Link is beyond the scope of this request and should be addressed to the DCI. Additionally, the FBI has access to a number of unclassified sources from non-DOJ components such as the State Department VISA application DB and INS data, as well as unclassified data from the Open Sources Information System (OSIS) as part of the

Intelligence Community. Ultimately, though, the range of sources that an FBI employee might reach out to is as varied as the responsibilities of the FBI and changes with each new investigation. It includes requests for data available for free and for purchase, data accessible by subpoena, data voluntarily provided, and classified and unclassified government data.

The FBI would like to make clear what we mean by "data mining." Broadly speaking, the term simply refers to the ability to work with larger amounts of data, at faster speeds, in ways that were previously not possible computationally due to size or speed limitations. In recent debates, however, some have begun to use the term data mining as a shorthand reference to the specter of abusive searches through vast amounts of publicly available data on innocent private citizens. The term should not be confused as connoting any such abuse.

"Data mining" really means searching. When permissible by law, and useful to a particular work activity, pertinent information gleaned from searching other databases is included in FBI systems. Once there, it may be accessible to another employee conducting a search. In the simplest example, an employee in one case obtains an address or phone number through an outside database search and then enters the information in an FBI system; an employee working on a different case conducts a search for something the two cases have in common and the second employee discovers the information the first employee got from an outside source. In a more complex example, the Foreign Terrorist Tracking Task Force looks for evidence that known terrorists are, or have been, in the United States by searching a whole list of names at the same time.

In responding to the inquiry about "pattern recognition," the FBI also would like to make clear what it means by the term. "Pattern recognition" refers to the ability to search a database or multiple databases for information that appears to be statistically significant. The FBI is exploring the potential of pattern recognition. For example, it would be useful if a "pattern recognition" program could identify anything statistically significant about known terrorists which is distinct from the general population -- this might be an aid in identifying tradecraft. Another example is the concept of using pattern recognition to enhance security; the ability to identify a computer user whose use of the system is statistically anomalous to his/her assigned duties might provide significant assistance to those responsible for internal security. Neither of these examples is currently an active project in the FBI. Rather, these are examples of the sort of discussions underway about the potential

uses of pattern recognition. The FBI is mindful that all such projects require legal scrutiny before implementation.

b. **Have any private sector or proprietary databases referred to in (A) above been aggregated with any data from government agency databases for data mining or pattern-recognition?**

ANSWER: As described above, the FBI does not seek whole databases. As also described above, extract information may be "aggregated" - placed in FBI databases when legal and appropriate.

c. **Is the Department using any data-mining tools to obtain information for law enforcement purposes unrelated to the detection and prosecution of terrorism?**

ANSWER: Yes. As data mining is defined in 1(a), every time the FBI provides a user i.d. to an employee who works crime, cybercrime, and counterintelligence, it is authorizing them to use data mining tools.

d. **To the extent that the Department is using proprietary data provided by private intermediaries, (i) what procedures are you using to preserve the confidentiality policies of these intermediaries? (ii) Is the Department compensating the private intermediaries for assisting in the data-mining? (iii) Has the Department taken any steps to shield the private intermediaries from liability for their cooperation with the government?**

ANSWER: (i) Access to proprietary data is limited in the same way as access to law enforcement data. It is restricted to those with a need to know and is limited to official duties. Access to all data is logged and recorded. (ii) The Department pays for the use of proprietary information as specified in contractual agreements, which have been subjected to standard FBI and/or DOJ procurement requirements. Other data is provided in response to court orders or volunteered at no cost to the government. (iii) DOJ has taken no steps to limit vendor's liability.

e. **What procedures, if any, does the Department follow to ensure the accuracy and reliability of information currently collected and stored in databases used for data mining?**

ANSWER: The FBI is interested in any information that may pertinent to authorized FBI mission activities. In the pursuit of such information, it is often not possible to determine in advance what information is accurate or reliable. With the passage of time seemingly irrelevant or dated information may acquire new significance as further investigation brings new details to light, and even information determined to be unreliable may continue to be of mission interest (e.g., in assessing the reliability of an information source, or in subsequent re-evaluations of the information's accuracy). Trained investigators and analysts exercise due diligence to verify information through links, relationships and other interpretations discovered during data mining and other investigative efforts.

f. **By contrast to the use of private sector or proprietary databases, in the search for proper data-mining tools, to what extent is the Department of Justice developing new tools and to what extent is it making use of existing tools developed in the private sector or used by other government agencies (such as search engines and data-mining software)? What are the pros and cons of these differing approaches?**

ANSWER: The Department has a policy to leverage the tools developed by private industry wherever possible. The advantage of this approach is greatly reduced cost, much faster fielding, reduced technical and schedule risks, and the advantage of constant modernization provided by Commercial-Off-The-Shelf (COTS) products. Also, vendors will often make modifications to a currently deployed product in specific response to government needs. The disadvantage is the cost of tailoring and integrating the products, the inability to unilaterally change commercial products, and some potential problems in security depending on the product. Where no COTS tools is available, DOJ next looks to products and tools developed by DOD, US Laboratories, the Intelligence Community and other government agencies, commonly referred to as Government-Off-The-Shelf (GOTS) products. The advantage of such products is, again, faster fielding and, sometimes, synergy of application. However, modification to a GOTS tools is often not available from the originator. On the occasion when no pre-existing tool can be identified, the Department will arrange to have its own built. This is often the slowest, most expensive option.

2. Foreign Terrorist Tracking Task Force:

a. **Please explain how the Department's FTTTF "lookout list" differs in substance and use from the FBI's Terrorism Watch List and how the FTTTF's "other intelligence-related projects" will differ from the functions of the FBI's JTTF, and IIIA database, and new Office of Intelligence. Please also explain how the FTTTF's "lookout lists" differ from or interface with those used by Customs, INS, and State Department (and successor agencies) for border control purposes and by the Transportation Security Administration?**

ANSWER: The FTTTF was created at the end of October 2001. One of its core functions is to provide information that locates or detects the presence of known or suspected terrorists within the United States by exploiting public and proprietary data sources to find an "electronic footprint" of known and suspected terrorists. In order to fulfill its mission, a top priority was obtaining an inclusive list of known and suspected terrorists. The FTTTF determined that no single list existed. Beginning in December 2001, the FTTTF began to compile the Consolidated Terrorist List (CTL). The CTL resides in a database where it can be checked automatically as part of the vetting mechanism for tasks such as the Alien Flight Training candidates and National Security Entry/Exit Registration (NSEERs) compliance checks.

The FBI's Terrorism Watch List (TWL), not to be confused with an FBI "Watch List" briefly used to located subjects and material witnesses immediately following 9/11, is an extension of the FBI's National Crime Information Center (NCIC) Violent Gang and Terrorist Organization File (VGTOF). VGTOF is designed to provide unclassified identifying information about violent criminal gangs and terrorist organizations, as well as members of those gangs and organizations, to law enforcement personnel on a query basis. VGTOF is the primary mechanism the FBI utilizes to provide other Law Enforcement Agencies, who have access to NCIC, with the names and identities of known and suspected terrorists. The TWL is maintained by FBIHQ personnel and includes classified backup documentation concerning VGTOF entries. When names are added to, removed from or modified in VGTOF, they are also added to, removed from or modified in the TWL. The classified information retained in the TWL database is shared with law enforcement officers and U.S. Intelligence Community personnel with proper security clearance and a need to know.

The FTTTF CTL is a compilation of the FBI's VGTOF and Department of States (DOS) TIPOFF databases that contains the names and identifying data of approximately 40,000 known and suspected terrorists.

The names contained in TIPOFF are of non-U.S. citizens, non Permanent Resident Aliens. TIPOFF is designed to prevent the entry of these individuals into the U.S. The TIPOFF names are shared with The Bureau of Immigration and Customs Enforcement (BICE).

BICE utilizes the Treasury Enforcement Computer System (TECS) database and the Interagency Border Inspection System (IBIS). TECS interfaces with approximately 10 other systems, including NCIC. TECS cross matches all incoming international flight and ship manifests against TECS data to identify subjects of interest. The system is not utilized for travel within the U.S. IBIS is the primary screening tool used by BICE at Ports of Entry.

The Transportation Security Administration (TSA) utilizes the No Fly and Selectee Lists. The No Fly List is designed to prevent individuals from using commercial aviation who are deemed by TSA to be a threat to civil aviation based on information provided by various sources, one being the FBI. The Selectee List consists of individuals who are not known to be a threat to aviation, but an agency such as the FBI and DOS, has determined the individual has a possible connection to terrorism. Additions to the No Fly and Selectee lists are based on recommendations from the U.S. Intelligence Community, primarily the FBI and DOS.

With the establishment of the Terrorist Screening Center the consolidation of "watchlist" information has been centralized with nominations of international terrorists coming through the Terrorist Threat Integration Center (TTIC) and domestic terrorists coming through the FBI. The FTTTF continues to maintain identifying information about known and suspected terrorists to permit computerized cross matching for identification of possible matches.

The FTTTF's "other intelligence-related projects" differ from the functions of the FBI's JTTF and IIIA database in that the FTTTF is narrowly focused to conduct analytical processes that have the objective of preventing the entry into the United States of persons listed by the U.S. Intelligence Community as terrorists, or detecting their presence in the United States irrespective of whether there is any reason to believe these aliens have entered or are attempting to enter the United States. Because the FTTTF is designed to be the component that provides expeditious information relating to the possible location of a known or suspected foreign terrorist, the JTTFs benefit from this

intelligence information, as they are involved in the operational action in the field that follows up on the FTTTF's information.

b. **Since Director Mueller routinely briefs the President with the CIA Director on terrorist threats, please explain why you decided to place the FTTTF in the Deputy Attorney General's office rather than within the FBI as part of its new Office of Intelligence?**

ANSWER: By memorandum dated August 6, 2002, the Attorney General ordered the Director of the FBI to "formally consolidate" the FTTTF within the Counterterrorism Division of the FBI, as part of "Phase II" of the FBI's reorganization. However, consistent with the original Presidential order creating the FTTTF, the Director of the FTTTF reports both to the Director of the FBI and to the Deputy Attorney General, which promotes coordinated information sharing with the highest levels of the Department of Justice. Congressional concurrence to move the FTTTF to the Office of Intelligence as part of a plan to transform intelligence within the FBI was sought, but not approved. Thus, it remains assigned to the FBI Counterterrorism Division. As the capabilities at FTTTF develop, it will continue to strengthen the FBI intelligence apparatus, and will maximize a number of unique core competencies. These include the automated extraction of public source data, visual mapping capabilities, and the development of analytical tools that can greatly enhance the FBI's ability to create optimum intelligence strategies, structure, and procedures to address evolving threats.

c. **Are the investigative restrictions applicable to FBI agents also applicable to employees conducting data-mining and operating the FTTTF under the guidance of the Deputy Attorney General? Conversely, given your guidelines on tracking terrorists are limited to the FBI, what is the source of and what are the guidelines defining the authority of the FTTTF?**

ANSWER: Since the FTTTF is now a part of the Counterterrorism Division of the FBI and operates with many FBI employees, including the FTTTF Director, the same investigative restrictions that apply to the FBI apply to the FTTTF. The source of and guidelines defining the authority of the FTTTF include: the Attorney General's Guidelines on General Crimes, Racketeering Enterprise and Terrorism Enterprise Investigations, Homeland Security Presidential Directive-2 (Oct. 29, 2001) directing the Attorney General to create the FTTTF and outlining its mission; the Attorney General's Guidelines for FBI National Security Investigations and Foreign Intelligence

Collection; the Attorney General's conforming action establishing the Task Force (Oct. 31, 2001); the Attorney General Delegation of Authority [to the FTTTF Director] to Conduct Security Checks of Students Requesting Flight Training Pursuant to Section 113 of the Aviation and Transportation Security Act (Jan. 7, 2002); the Attorney General's Order regarding the Coordination of Information Relating to Terrorism (Apr. 11, 2002); the Attorney General's delegation of authority to the Director and Deputy Directors of the FTTTF pursuant to 15 U.S.C. §1681v(b) (July 3, 2002); the Attorney General's memorandum to FBI Director Mueller (Aug. 6, 2002); and the Attorney General's delegation of authority to the Director and Deputy Directors of the FTTTF pursuant to 12 U.S.C. §3414(a) (Sept. 3, 2002). The applicable regulations of the FTTTF's mission are located in Part 16 and 105 of Title 28 of the Code of Federal Regulations, and in AAG/A Order No. 276-2002.

d. What information is necessary to trigger a data-mining inquiry on a particular individual or targeted activity to ensure that this technique is only being used for purposes relevant to detecting, preventing or punishing terrorism or other criminal activity?

ANSWER: It is important to note that the term "data mining" simply reflects the next stage of technology enhancements to search capabilities provided by the industry. To ensure that FTTTF inquiries are only being used for relevant purposes, inquiries are only accepted and conducted that originate from official governmental channels established with the FTTTF and comply with the Attorney General's Guidelines.

3. Admiral Poindexter's Total Information Awareness Project (TIA)

According to the Department of Defense, the Defense Advanced Research Project Agency (DARPA) has established the Total Information Awareness (TIA) Project to develop technologies for rapid language translation, commercial transaction data-mining, and interagency analysis and decision-making tools.

a. To what extent are you and the Department of Justice consulting or collaborating with Admiral Poindexter or the Department of Defense in designing and implementing TIA surveillance tools and related programs?

ANSWER: No data mining or analytic tools from TIA have been delivered to the Department of Justice. There was limited collaboration or consultation with TIA research and development efforts. The TIA was being developed to fulfill a DOD mission requirement (i.e., countering the international terrorist threat by making better use of existing data as opposed to collecting more data via surveillance). DARPA provided briefings on TIA to the FBI. These briefings were intended to establish a dialogue for evaluating these technologies as a means of satisfying the FBI's own developmental and operational requirements. DARPA has also provided briefings about TIA to other Justice Department personnel.

The FBI does participate in a biometric technologies project in which DARPA has participated. Prior to the creation of TIA, a test was conducted under DARPA's Human Identification at Distance (HumanID) program. The results from the Face Recognition Vendor Test (FRVT) 2002 are being included in a report on biometrics for border security mandated under the Patriot Act and Enhanced Border Security Act, which require the report to be jointly written by DOJ, NIST and State. The FBI, through the Technical Support Working Group (TSWG), is jointly funding efforts under the HumanID program to establish Daubert statistics for face recognition systems for use in court testimony. The scope of the HumanID program is focused on research to advance state-of-the-art biometric technologies and is not involved in data-mining.

The "Consolidated Appropriations Resolution, 2003" (the Act) provided that, unless a prescribed report on the "Total Information Awareness" (TIA) program was submitted to the Congress within 90 days of enactment, "no funds appropriated or otherwise made available to the Department of Defense . . . may be obligated or expended on research and development" of the TIA program. *See* Pub. L. No. 108-7, Div. M, § 111(a). The report required to avoid this funding cut-off "is a report, in writing, of the Secretary of Defense, the Attorney General, and the Director of Central Intelligence, acting jointly," that addresses five specified subjects. Since the passage of the Act, various representatives from the Department of the Defense (DoD), the Department of Justice (DOJ), and the Central Intelligence Agency (CIA) worked together to prepare the required report, which was submitted to the Congress on May 20, 2003. In September, 2003, the Congress, with certain limited exceptions, eliminated funding for TIA in the Fiscal Year 2004 DoD Appropriations Act. See Pub. L. No. 108-87, sect. 8131, 117 Stat. 1054 (Sept. 30, 2003).

b. **Have any TIA generated or developed technologies been delivered to the Department of Justice and, if so, (i) are any being used? (ii) describe the purposes for which they are being used; and (iii) are any of the tools for data-mining and pattern recognition?**

ANSWER: No TIA generated or developed technologies were delivered to the Department of Justice.

c. **TIA has programs called Genoa I and II. Has this program been delivered in whole or in part to the Department of Justice and, if so, (i) is it being used? (ii) Describe the purposes for which it is being used; and (iii) is this a tool for data-mining or pattern recognition?**

ANSWER: Genoa I and II represented separate research programs associated with TIA. Neither Genoa I nor II resulted in a specific hardware or software components capable of being operated by another agency but instead resulted in tools and the refinement of techniques. Genoa I pre-dated and provided a basis for TIA. It did not result in any tools or techniques that were delivered to the Department of Justice. Genoa II further extended the concepts and technology derived from Genoa I. Genoa II developed and evaluated technologies in an experimental setting for possible integration as components of a TIA network. None of the technology associated with Genoa II was delivered to the Department of Justice.

d. **TIA has a program called EELD (Evidence Extraction and Link Discovery). Has this program been delivered in whole or in part to the Department of Justice and, if so, (i) is it being used? (ii) Describe the purposes for which it is being used; and (iii) is this a tool for data-mining or pattern recognition?**

ANSWER: EELD was a research program that pre-dated TIA intended to produce prototype tools and techniques for further development within TIA. It did not result in a software or hardware component capable of being deployed or operated by any agency. The program was to develop and evaluate technologies in an experimental setting for possible integration as components of a TIA network. No EELD components were delivered to the Department of Justice.

e. **TIA has a program called Genisys. Has this program been delivered in whole or in part to the Department of Justice and, if so, (i) is it being used? (ii) Describe the purposes for which it is being used; and (iii) is this a tool for data mining or pattern recognition?**

ANSWER: The program was to develop and evaluate technologies in an experimental setting for possible integration as components of a TIA network. Nothing related to Genisys was delivered to the Department of Justice.

f. **TIA has a program called TIDES (Translingual Information Detection, Extraction and Summarization). Has this program been delivered in whole or in part to the Department of Justice and, if so, (i) is it being used? (ii) Describe the purposes for which it is being used; and (iii) is this a tool for data-mining or pattern recognition?**

ANSWER: TIDES was a research program intended to make it possible for English speakers to find and interpret needed information quickly and effectively, regardless of the language or medium. The TIDES program began in FY 2001. The program was to develop and evaluate technologies in an experimental setting for possible integration as components of a TIA network. No TIDES technology was delivered to the Department of Justice.

g. **Is the FTTTF coordinating its work in any way with the TIA?**

ANSWER: No. The FTTTF is not coordinating its work in any way with the TIA. When the FTTTF was in its start-up phase, it had courtesy briefings with DARPA. DARPA expressed some interest in having FTTTF participate as an experimental "node" in the TIA program, no further discussions were ever held.

h. **What safeguards, if any, do you believe should be included in any data-mining tools developed by TIA to ensure the accuracy and reliability of the information collected and stored in databases? Have you recommended such safeguards to the Department of Defense?**

ANSWER: As noted above, on May 20, 2003, a joint report prepared by DoD, DOJ, and CIA was submitted to Congress on the TIA program, in accordance with Pub. L. No. 108-7, Div. M, § 111(b). Section 111(b) specifies that the report must contain "recommendations, endorsed by the Attorney General, for practices, procedures, regulations, or legislation on the

deployment, implementation, or use" of the TIA program "to eliminate or minimize adverse effects of such program on privacy and other civil liberties".

As set forth in the joint report, the Justice Department endorsed several recommendations in this regard. Specifically, the report concludes that when and if TIA's search tools are developed, any future deployment of those tools with respect to data sources that contain information on U.S. persons would raise significant privacy issues that would require careful and serious examination before any such deployment should be undertaken. In particular, the joint report recommends that the following factors be considered, in advance, in evaluating TIA's suitability for deployment in particular contexts:

- The *efficacy and accuracy* of TIA's search tools must be carefully tested and demonstrated.
- It is critical that there be *built-in operational safeguards* to reduce the opportunities for abuse.
- It is essential to ensure that *substantial security measures* are in place to protect such tools from unauthorized access by hackers or other intruders.
- Any agency contemplating deploying TIA tools for use in particular contexts, particularly deployments with respect to data sources that contain information on U.S. persons, must be required first to conduct a thorough *pre-deployment legal review.*
- Any such agency must also have in place policies establishing *effective oversight* of the actual use and operation of the system before it is deployed in particular contexts.

In addition, the joint report specifically recommends that, as research and development proceed, careful study should be given to whether anything about the particular *technological architecture* of the TIA tools raises specific privacy concerns. In particular, any technological system that would involve the installation of government-developed software code onto privately owned databases would raise significant legal and policy concerns that would require careful scrutiny.

The report also notes that DoD has taken several steps, including the appointment of an oversight board and of a Federal Advisory Committee, to address these issues. In addition, the report properly emphasizes DoD's commitment "to address privacy and civil liberties issues squarely as they arise," and its affirmation that the "protection of privacy and civil liberties is an integral and paramount goal in the development of counterterrorism technologies and in their implementation."

The report does not recommend any changes in statutory law, but instead contemplates that any deployment of TIA's search tools may occur only to the extent that such a deployment is consistent with current law. Accordingly, the report specifically notes that the strictures of current law protecting certain categories and sources of information may well constrain or (as a logistical matter) completely preclude deployment of TIA search tools with respect to such data. This, of course, underscores the importance of the report's recommendation for a careful pre-deployment legal review.

4. **Compliance with the Privacy Act**

a. **Does the Privacy Act impose any restriction on data-mining activities by the Department and, if so, what are those restrictions?**

ANSWER: The Privacy Act does not impose any restrictions on data-mining activities per se. However, to the extent that data-mining involves maintenance, collection, use or dissemination of records that are subject to Privacy Act restrictions, the Department must comply with these restrictions, regardless of the medium involved. In addition, certain data mining scenarios could also implicate the restrictions of the Computer Matching and Privacy Protection Act of 1998 (CMPPA). (See discussion in "g" below).

b. **Does the Department employ any outside contractors to perform data-mining services and, if so, how does the Privacy Act apply, if at all, to the outsourcing of data-mining activities?**

ANSWER: Using the definition for "data mining" provided in response to question 1(a), the FBI may employ outside contractors to perform or assist in performing data-mining services from time to time. The remainder of this response refers to data mining in in-house systems of records. To the extent that a contractor might assist in searches of public source records, the primary concern should be, and is, that the necessary legal predicate exists

to conduct a search. The Privacy Act does not apply until the resulting information is brought within an FBI system of records.

Paragraph (m)(1) of the Privacy Act provides: "When an agency provides by a contract for the operation by or on behalf of the agency of a system of records to accomplish an agency function, the agency shall, consistent with its authority, cause the requirements of this section to be applied to such system. For purposes of [the criminal penalties] of this section any such contractor and any employee of such contractor, if such contract is agreed to on or after the effective date of this section, shall be considered to be an employee of an agency."

The Department has promulgated guidance in 28 CFR § 16.52: "Any approved contract for the operation of a record system will contain the standard contract requirements issued by the General Services Administration to ensure compliance with the requirements of the Privacy Act for that record system. The contracting component will be responsible for ensuring that the contractor complies with these contract requirements."

Case law is split as to whether disclosure of records to a contractor that serves the function of an agency employee is a permissible intra-agency disclosure pursuant to paragraph (b)(1) of the Privacy Act. See Coakley v. United States Dep't of Transportation, No. 93-1420, 1994 U.S. Dist. LEXIS 21402, at **3-4 (D.D.C. Apr. 7, 1994); Hulett v. Dep't of the Navy, No. TH 85-310-C, slip op. at 3-4 (S.D. Ind. Oct. 26, 1987), aff'd, 886 F.2d 432 (7th Cir. 1988); Taylor v. Orr, No. 83-0389, 1983 U.S. Dist. LEXIS 20334, at **7-10 (D.D.C. Dec. 5, 1983) (cited in the Department of Justice, Office of Information and Privacy, Freedom of Information Act Guide and Privacy Act Overview, at 818 (May 2002)).

The Office of Management and Budget (OMB) has published guidance regarding disclosure of records to contractors: "When an agency provides by contract for the operation of a system of records, it should ensure that a system of records notice describing the system has been published. It should also review the notice to ensure that it contains a routine use under section (e)(4)(D) of the Act permitting disclosure to the contractor and his or her personnel." Appendix I to OMB Circular No. A-130 -- Federal Agency Responsibilities for Maintaining Records About Individuals, 61 Federal Register 6428, 6439 (Feb. 20, 1996). Accordingly, the FBI has published a routine use permitting disclosure of records to contractors. See Blanket Routine Uses (BRU) Applicable to More Than One FBI Privacy Act System of Records (JUSTICE/FBI-BRU), 66 Federal Register 33559-33560

(June 22, 2001). The FTTTF has also published such a routine use for alien flight training records. See 67 Federal Register 47571 (July 19, 2002).

c. **The Privacy Act, 5 U.S.C. § 552a(e)(4), requires agencies to "publish in the Federal Register upon establishment or revision a notice of the existence and character of the system of records." Have you promulgated any regulations regarding the FTTTF?**

ANSWER: Yes. Although many FTTTF records relate to persons not covered by the Privacy Act, the FTTTF has published a Privacy Act System Notice, Routine Uses and Privacy Act exemptions for the Flight Training Candidates File System (JUSTICE/FTTTF-001), 67 Federal Register 39839 (June 10, 2002), 67 Federal Register 47570 (July 19, 2002), and 67 Federal Register 51756 (Aug. 9, 2002). Now that the FTTTF has been transferred to the FBI, the remainder of the FTTTF's records are currently covered by the Privacy Act System Notice and Privacy Act exemptions for the FBI's Central Records System (JUSTICE/FBI-002), 63 Federal Register 8671 (February 20, 1998), 28 CFR § 16.96(a), (b) and final rule, 68 Federal Register 14140 (March 24, 2003). See also, Blanket Routine Uses (BRU) Applicable to More Than One FBI Privacy Act System of Records (JUSTICE/FBI-BRU), 66 Federal Register 33559-33560 (June 22, 2001).

d. **The Privacy Act, 5 U.S.C. § 552a(e)(4)(E), requires publication of the policies and practices of the agency regarding storage, retrievability, access, controls, retention and disposal of the records. Have you published such policies and practices regarding the FTTTF?**

ANSWER: Yes. See the answer to "c" above.

e. **Generally, the Privacy Act prohibits governmental agencies from disclosing records to another agency, unless it falls under the "routine use" exception. 5 U.S.C. § 552a(b)(3). Does the Department rely on this "routine use" exception to obtain databases from other agencies for aggregation in the FTTTF and other databases within the Department?**

ANSWER: When the Department obtains records from another agency under the Privacy Act it is the responsibility of the source agency to ensure that the source agency has proper routine uses or other proper authority before the source agency discloses records to the Department.

Nevertheless, to the extent that records involve non-United States citizens or nonresident aliens, or the records are not from systems of records, the Privacy Act does not apply. To the extent that records are subject to the Privacy Act, a number of authorities may support the Department's obtaining the records. These include: disclosure to the Department after consent is obtained from the individual to whom the record pertains; an intra-agency disclosure under paragraph (b)(1) of the Act when the source is another DOJ component; disclosure pursuant to paragraph (b)(7) of the Act, following a written request by the Attorney General or his delegate, for an authorized law enforcement activity; or disclosure pursuant to a routine use that is published by the source agency in the Federal Register.

f. **The Privacy Act, 5 U.S.C. § 552a(e)(4)(D), requires Federal Register publication of "each routine use of the records contained in the system, including the categories of users and the purpose of such use." If the answer to (E) above is affirmative, has the Department published any Federal Register notice required by the Privacy Act? If so, please provide a copy of any such notice and, if not, please explain why.**

ANSWER: The Department is not responsible for publishing routine uses for other agencies to disclose information to us. When DOJ receives records from another agency, it is the responsibility of that other agency to have proper routine uses or other authority in place to disclose records to the Department. The FBI and the FTTTF have published Privacy Act system notices and routine uses, as discussed in "c" above.

g. **The Privacy Act imposes restrictions on "matching" programs conducted by the government or the private sector on behalf of the government, unless the matching is conducted "subsequent to the initiation of a specific criminal or civil law enforcement investigation" or "for foreign counterintelligence purposes." How does the Department ensure that the FTTTF and other Department databases comprised of aggregated data from other agencies are operated within these restrictions?**

ANSWER: The Computer Matching and Privacy Protection Act (CMPPA) applies only to "matching programs," which is a term of art quite narrowly defined in the Privacy Act. A "matching program" is defined as a computerized comparison of two or more automated systems of records or a system of records with non-Federal records for the purpose of (a)

establishing or verifying initial or continuing eligibility for Federal benefit programs, (b) verifying compliance with the requirements of such programs, or (c) recouping payments or delinquent debts under such programs. 5 U.S.C. 552a(a)(8)(A)(i). A "matching program" also includes a computerized comparison of two or more Federal personnel or payroll system of records or a system of Federal personnel or payroll records with non-Federal records. 5 U.S.C. 552a(a)(8)(A)(ii).

DOJ has promulgated guidance within the Department apprizing components of the provisions of the CMPPA and applicable OMB guidelines. This guidance is disseminated to FOIA/Privacy Act officers in each component. As of February 28, 2003, DOJ had 15 computer matching agreements. With INS' transfer to DHS, ten of these agreements transferred to DHS. None of the agreements involve either the FBI or the FTTTF. The Justice Management Division (JMD), in compliance with the requirements of the CMPPA, presents these computer-matching agreements at stipulated intervals to the DOJ Data Integrity Board for the Board's review and approval. In further compliance with the CMPPA, JMD also periodically reports to OMB, Congress, and the public (through publication in the Federal Register) when DOJ either enters into a new computer matching agreement or renews an existing agreement. Finally, JMD periodically reports to the Attorney General on the Department's computer matching activity.

To date, the FTTTF has not found any of its operations to be subject to the CMPPA.

h. **Does the Department believe that any amendments to the Privacy Act would be helpful to facilitate data-mining by the Department and, if so, does the Department intend to transmit to the Congress any amendments to the Privacy Act to clarify the legality of data-mining by Federal agencies?**

ANSWER: Not at this time.

5. **Coordination With the Department of Homeland Security**

a. **The Homeland Security Act expressly authorizes the new department to request, access, receive, analyze and integrate information from government agencies and private sector entities, and to establish and utilize "a secure communications and information technology infrastructure, including data-mining and other advanced analytical**

tools, in order to assess, receive and analyze data and information" [P.L. 107-296, Sections 201(d)(1), (13), (14)]. Does the Department of Justice have any such express statutory authority to conduct data mining? If so, please describe that authority.

ANSWER: FBI authority to conduct such activities is inherent in and derived from the FBI's core mission responsibilities, such as those contained in 28 U.S.C. 533 ("The Attorney General may appoint officials to .... detect and prosecute crimes against the United States, to assist in the protection of the person of the President, to assist in the protection of the person of the Attorney General and to conduct such other investigations regarding official matters under the control of the Department of Justice and the Department of State as may be directed by the Attorney General.")

b. **Do you anticipate the Department of Justice's data-mining operations being transferred to the new Department of Homeland Security? If not, please explain why.**

ANSWER: From the FBI's point of view, the use of automated data analysis techniques will not be transferred to another department because the use of such techniques is authorized as part of the FBI's criminal and national security investigative and intelligence functions under the Attorney General's Guidelines. The FBI has primary jurisdiction for investigations and for the collection, analysis, and production of intelligence to detect and prevent domestic and international terrorism within the United States. To fulfill this mission the FBI must have the ability to use all lawful investigative and analytic techniques with full regard for the protection of constitutional rights.

c. **Do you believe it is valuable to have a coordinated data-mining effort with one agency clearly held accountable for setting guidelines of data uniformity and reliability and, if so, which agency do you believe should take primary position in order to avoid duplication of effort?**

ANSWER: The Department of Justice believes that there should be a coordinated data analysis effort for law enforcement agencies to collect, analyze, share, and exchange information. Data analysis is a rapidly evolving area with new analytical software tools being developed and deployed at a rapid pace. One of the strategic goals and part of various initiatives at the Department is to improve the effectiveness and coordinate efforts within law enforcement agencies to utilize these analytical software

tools and set guidelines for data uniformity and information sharing. The Office of Justice Programs has established guidelines for data uniformity and reliability as part of the Justice Extensible Markup Language (XML) data model that is being used as part of the National Criminal Intelligence Sharing Plan. This model facilitates the exchange and analysis of information.

Sincerely,

William E. Moschella
Assistant Attorney General

cc: The Honorable Orrin G. Hatch
Chairman

Statement of Senator Edward M. Kennedy
Senate Judiciary Committee Hearing on "Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs"
January 10, 2007

Mr. Chairman, I commend you for holding this hearing and for your dedication to ensuring that our citizens' privacy and civil liberties are not unnecessarily or unjustifiably violated in the battle against terrorism.

Terrorism is the greatest challenge we face today as a nation. We all agree on the need for strong powers to investigate terrorism, prevent future attacks, and improve information-sharing by federal, state and local law enforcement. But legitimate concerns about the terrorist threat should not be misused as an excuse to grant extraordinary and unchecked powers to the President.

Modern technology holds great promise for meeting all the challenges we face, but we can't afford to overlook the new encroachments on privacy and our civil liberties that such technology now makes possible. We must not sacrifice core American values in our battle against terrorism, for there can be no victory at the cost of these ideals.

It is the duty of Congress to ensure that the proper balance is achieved between realizing the promise of technology and safeguarding civil liberties and the right to privacy. The Bush Administration is not entitled – nor should it expect – a blank check when it comes to fighting terrorism. We don't question the sincerity of the Administration in wanting to protect the American people against new terrorist attacks. But it is our responsibility to conduct meaningful oversight over the judgments and methods involved.

For these reasons, last Congress I urged my colleagues to adopt an amendment that would have required the Administration to tell Congress what the National Security Agency is doing. My amendment would have merely required that the NSA report to Congress on the legal standards used for electronic surveillance within 60 days after the enactment of this bill. My amendment was identical to a requirement sponsored by one of the Committee's witnesses' today, former Representative Bob Barr, and adopted with unanimous bipartisan support in the 106th Congress. Now, a few years later, the Congress and the American people are still entitled to know what standards the NSA is observing in conducting electronic surveillance.

As part of the 2000 Intelligence Authorization Act, the Congress mandated that the National Security Agency report on the legal standards that it was using to conduct surveillance on U.S. soil. At that time, many in Congress were concerned about rumors that the NSA was engaging in broad eavesdropping, and language requiring the Attorney General and the Director of National Intelligence to report on the standards used for electronic surveillance was adopted – without a single objection in either the House or the Senate. Based on public reports now, the NSA had not begun its so-called "Terrorist Surveillance Program" at the time that it provided its report to Congress on the legal standards being used. The Administration owes us a current answer on how they define the playing field. Unfortunately, the Committee did not adopt my amendment last year but now we have another opportunity to press ahead and today's hearing is a good start.

The Administration has repeatedly betrayed the public trust with its broad interpretation of Executive power and authority. This President thinks it's permissible to listen in on our phone conversations and read our mail – without any Congressional or judicial oversight. At the time the Foreign Intelligence Surveillance Act was enacted in 1978, the President and Congress concluded that our national security laws are toughest when they are clear and meet the test of common sense. Without such guidance, the actions of national security officials and law enforcement officials will be subject to frequent challenge in the courts.

Today, however, we face the unsettling prospect that there are no clear rules for the government's national security actions or data collection.

Our common concern for national security can best be met when the President and Congress work together to approve the means he uses to keep us safe. But instead of uniting us to strengthen our national security, the Administration has taken its own controversial and divisive course. Instead of working with Congress to improve our laws, the President has chosen to ignore them and ride roughshod over basic constitutional principles.

Already, the Administration has had to terminate its data mining programs, after details of their operation came to light. The Pentagon's Terrorism Information Awareness program, formerly known as the Total Information Awareness program, was dropped in the face of legislation aimed at ending the program. The Transportation Security Administration's passenger prescreening program was stopped after lawsuits were filed by passengers challenging the sharing of their personal information between commercial airlines and the TSA. Another concern was TSA's stated intention to use the information it collected for purposes other than fighting terrorism, such as to identify individuals with outstanding warrants or expired visas.

Data mining is a developing technology that may prove effective in fighting terrorism without unduly compromising privacy and civil liberties. But we need more information. It's time for the Administration to give us an accounting of all its data mining programs currently in existence. We need to know the answers to many questions about these programs. Are they effective? Are the costs - in terms of dollars and privacy - worth it? Where is the data coming from? Who is collecting it? Is the data captured accurate and complete? For what purposes is the information used? Who has access to it? What safeguards are in place to ensure that the desire for information does not trample upon our basic rights? Are there safeguards to ensure that groups, such as Arab Americans or Muslims, are not targeted unfairly? Are these safeguards effective? We have a long list of questions and today's hearing is a welcome start to obtaining meaningful answers.

As we look forward to further debate on this important topic during the 110th Congress, I'd like to remind my colleagues that the Administration is required to submit a report to Congress on its current data-mining activities by March 9, 2007. When we reauthorized the PATRIOT Act, Congress established a requirement for the Attorney General to report to Congress regarding the Department of Justice's use or development of data mining technologies – within one year of the date of the enactment of the reauthorized PATRIOT Act. I expect that the Administration will meet the statutory deadline so that we will finally have more detailed information on their current practices and procedures.

I look forward to today's testimony and to new and more vigorous oversight by the Judiciary Committee under the leadership of Chairman Leahy. Across party lines, many of us stand ready to improve our surveillance laws to serve our country's best interests. It would be wrong for Congress to continue to rubber stamp programs that will change the law in far-reaching ways and produce devastating losses of basic constitutional freedoms and protections. Now more than ever, we must be vigilant in our defense of safeguards that limit the President's power to collect and store vast amounts of information on Americans.

Statement
United States Senate Committee on the Judiciary
Balancing Privacy and Security: The Privacy Implications of Government Data Mining Programs
January 10, 2007

The Honorable Patrick Leahy
United States Senator , Vermont

Opening Statement of Senator Patrick Leahy

Senate Judiciary Committee

Hearing on "Balancing Privacy and Security:

The Privacy Implications of Government Data Mining Programs"

January 10, 2007

Today, the Senate Judiciary Committee holds an important hearing on the privacy implications of government data mining programs.

This Committee has a special stewardship role in protecting our most cherished rights and liberties as Americans, including the right to privacy. Today's hearing on government data mining programs is our first in the new Congress. It is the first of what I plan to be a series of hearings on privacy-related issues throughout this Congress.

The Bush Administration has dramatically increased its use of data mining technology -- namely, the collection and monitoring of large volumes of sensitive personal data to identify patterns or relationships. Indeed, in recent years, the federal government's use of data mining technology has exploded, without congressional oversight or comprehensive privacy safeguards. According to a May 2004 report by the General Accounting Office, at least 52 different federal agencies are currently using data mining technology, and there are at least 199 different government data mining programs operating or planned throughout the federal government.

Advances in technologies make data banks and data mining more powerful and more useful than ever before. These can be valuable tools in our national security arsenal, but we need to ensure we use them appropriately and with the proper safeguards so that they can be most effective.

One of the most common – and controversial – uses of this technology is to predict who among our 300 million people are likely to be involved in terrorist activities. According to the GAO and a recent study by the CATO Institute, there are at least 14 different government data mining programs within the Departments of Defense, Justice, Homeland Security and Health. That does not include the NSA's programs.

Congress is overdue in taking stock of the proliferation of these databases that increasingly are collecting and sifting more and more information about each and every American.

Although billed as counterterrorism tools, the overwhelming majority of these data mining programs use, collect, and analyze personal information about ordinary American citizens. Despite their

prevalence, these government data mining programs often lack adequate safeguards to protect privacy and civil liberties.

Just recently, we learned through the media that the Bush Administration has used data mining technology secretly to compile files on the travel habits of millions of law-abiding Americans. Incredibly, under the Department of Homeland Security's Automated Targeting System program ("ATS"), our government has been collecting and sharing this sensitive personal information with foreign governments and even private employers, while refusing to allow U.S. citizens to see or challenge their own so-called "terror scores."

Following years of denial, the Transportation Security Administration ("TSA") has finally admitted that its controversial "Secure Flight" data mining program – which collects and analyzes airline passenger data obtained from commercial data brokers – violated federal privacy laws by failing to give notice to U.S. air travelers that their personal data was being collected for government use.

And last month, The Washington Post reported that the Department of Justice will expand its ONE-DOJ program – a massive data base that will allow state and local law enforcement officials to review and search millions of sensitive criminal files belonging to the FBI, DEA and other federal law enforcement agencies. This will make sensitive investigative information about thousands of individuals – including those who have never been charged with a crime – available to local and state law agencies.

Without the proper safeguards and oversight of these and other government data mining programs, the American people have neither the assurance that these massive data banks will make us safer, nor the confidence that their privacy rights will be protected. In addition, there are legitimate questions about whether data mining technology is actually effective in identifying risks or terrorists.

A recent CATO Institute study also found that data mining is not an effective tool for predicting or combating terrorism, in part because of the high risk of false positive results. A front-page article several months ago included interviews with experts who conceded how ineffective and haphazard these programs have been. We need look no further than the government's own terrorist watch list, which now contains the names of more than 300,000 individuals – including infants, nuns, and even members of Congress – to understand the inefficiencies that can result from data mining and government dragnets. If these databases are being used in ways that create more wheel-spinning that saps critical investigative resources from effective tasks, we need to know that so we can use our tools and our talent more efficiently to get the real results in needed in thwarting terrorism. We also need to understand that a mistake in a government data base could cost a person his or her job, sacrifice their liberty, and wreak havoc on their life and reputation.

Given the many challenges posed by this technology, we in Congress must do our part to examine data mining technology and to ensure that government data mining programs actually do keep Americans safe – not just from enemies abroad, but also from abuses at home.

We begin that important task today. I am joining with Senator Feingold, Senator Sununu and others in a bipartisan attempt to provide congressional oversight to these programs. We are introducing the Federal Agency Data Mining Reporting Act of 2007. This threshold privacy legislation would begin to restore key checks and balances by requiring federal agencies to report to Congress on their data-mining programs and activities. We joined together to introduce a similar bill last Congress. Regrettably, it received no attention. This year, I intend to make sure that we do a better job in considering Americans' privacy, checks and balances, and the proper balance to protect Americans'

privacy rights while fighting smarter and more effectively against security threats.

This legislation takes a crucial first step in addressing these concerns by pulling back the curtain on how this Administration is using this technology. It does not prohibit the use of this technology, but rather provides an oversight mechanism to begin to ensure it is being used appropriately and effectively. Our bill would require federal agencies to report to Congress about its data mining programs. The legislation provides a much-needed check on federal agencies to disclose the steps that they are taking to protect the privacy and due process rights of American citizens when they use these programs.

We need checks and balances to keep government data bases from being misused against the American people. That is what the Constitution and our laws should provide. We in Congress must make sure that when our government uses technology to detect and deter illegal activity, the government does so in ways that also protect our most basic rights and liberties, and in ways that limit opportunities for abuse of these powerful tools. Our bill advances this important goal.

I thank Chairman Specter for scheduling this hearing at my request while the Republican caucus proceeds to deliberate Committee reorganization, and I thank our distinguished panel of witnesses for appearing here today.

January 10, 2003

The Honorable John Ashcroft
Attorney General
United States Department of Justice
Main Justice Building, Room 5137
950 Pennsylvania Avenue, N.W.
Washington, D.C. 20530

Dear Attorney General Ashcroft:

I am writing to inquire about the current "data mining" operations, practices and policies at the Department of Justice. Improved access to and the sharing of information among intelligence and law enforcement agencies at the federal, state and local levels is crucial in promoting our national security interests. These national security interests are most effectively and efficiently served, however, when the information being collected and shared is relevant, reliable, timely and accurate. As one recent expert report observed, "Data mining, like any other government data analysis, should occur where there is a focused and demonstrable need to know, balanced against the dangers to civil liberties. It should be purposeful and responsible." *(Protecting America's Freedom in the Information Age, A Report of the Markle Foundation Task Force*, October, 2002, p. 27.)

Adequate oversight by the Congress, and especially by the appropriate committees of jurisdiction, is essential in helping to ensure that adequate standards are set and met, so that these activities can be both effective and respectful of the constitutional rights of the American people. Accordingly, I am interested in learning the extent to which the Department is relying on data mining to deal with the terrorism threat or other criminal activity, and how this technology is being used.

I raise this inquiry against the backdrop of public concern over the Total Information Awareness System (TIA) being developed under the supervision of Admiral Poindexter within the Defense Advanced Research Project Agency (DARPA). TIA is intended, according to Department of Defense officials, to generate tools for monitoring the daily personal transactions by Americans and others, including tracking the use of passports, driver's licenses, credit cards, airline tickets, and rental cars. The Administration's goal is to turn these tools over to law enforcement agencies. According to press reports, one such tool, a software program called "Genoa," has already been delivered by DARPA to the Department of Justice.

Advances in the technological capability to search, track or "mine" commercial and government databases and Americans' consumer transactions have provided powerful tools that have dramatically changed the ways that companies market their products and services. Collection and use by government law enforcement agencies of such commercial transactional data on law-abiding Americans poses unique issues and

concerns, however. These concerns include the specter of excessive government surveillance that may intrude on important privacy interests and chill the exercise of First Amendment-protected speech and associational rights.

Moreover, as Federal law enforcement agencies obtain public source and proprietary data for mining, the sheer volume of information may make updating the data and checks for reliability and accuracy difficult, if not impossible. Reliance on data mining by law enforcement agencies may produce an increase in false leads and law enforcement mistakes. While the former is a waste of resources, the latter may result in mistaken arrests or surveillance. Such mistakes do occur, even without data-mining.[1] In short, while the only ill effect of business reliance on outdated or incorrect information may be misdirected marketing efforts, data mining mistakes made by a law enforcement agency may result in misdirection or misallocation of limited government resources and devastating consequences for mistakenly targeted Americans.

I am interested in determining the extent to which the Justice Department is relying on data-mining and how the Department is addressing these concerns with appropriate safeguards on the collection, use and dissemination of information obtained through data mining. Specifically, I ask for and would appreciate your responses to the following questions.

1. Data-Mining Operations Underway Within the Department of Justice.

(A) Please identify any private sector or proprietary databases obtained or being used by the Department of Justice for data-mining or pattern-recognition activities.

(B) Have any private sector or proprietary databases referred to in (A) above been aggregated by the Department with any data from government agency databases for data-mining or pattern-recognition activities?

(C) Is the Department using any data-mining tools to obtain information for law enforcement purposes unrelated to the detection and prosecution of terrorism?

(D) To the extent that the Department is using proprietary data provided by private intermediaries, (i) what procedures are you using to preserve the confidentiality policies of these intermediaries? (ii) Is the Department compensating the private intermediaries for assisting in the data mining? (iii) Has the Department taken any steps to shield the private intermediaries from liability for their cooperation with the government?

(E) What procedures, if any, does the Department follow to ensure the accuracy and reliability of information currently collected and stored in databases used for data-mining?

(F) By contrast to the use of private sector or proprietary databases, in the search for proper data mining tools, to what extent is the Department of Justice developing new tools and to what extent is it making use of existing tools developed in the private sector or used by other government agencies (such as search engines and data mining software)? What are the pros and cons of these differing approaches?

2. Foreign Terrorist Tracking Task Force. On October 29, 2001, the President directed the Department to establish the Foreign Terrorist Tracking Task Force (FTTTF) to "ensure that, to the maximum extent permitted by law, Federal agencies coordinate programs to . . . 1) deny entry into the United States of aliens associated with, suspected of being engaged in, or supporting terrorist activity; and 2) locate, detain, prosecute, or deport any such aliens already present in the United States." Your April 11, 2002, order establishing the FTTTF would do more than ensure that agencies "coordinate programs" and requires the FTTTF to have "electronic access to large sets of data, including the most sensitive material from law enforcement and intelligence sources." In response to my request for more detailed description of the mission and activities of the FTTTF, you stated in response to written questions that:

> "The FTTTF has identified a number of specific projects which it can coordinate or run to fill gaps in existing government efforts relating to prevention of terrorist activities. For example, the FTTTF is pursuing projects to: 1) create a unified, cohesive lookout list; 2) identify foreign terrorists and their supporters who have entered or seek to enter the U.S. or its territories; and 3) detect such factors as violations of criminal or immigration law which would permit exclusion, detention or deportation of such individuals. In addition, the FTTTF is in the process of identifying other intelligence-related projects that it can support through its collaborative capability to co-locate data from multiple agency sources."

> (A) Redundancy within government programs can be both expensive and ineffective. The "projects" of the FTTTF appear to overlap other initiatives underway within the Department. For example, the FBI has an Information Sharing Task Force and participates in 47 Joint Terrorism Task Forces (JTTF) to unify all levels and branches of law enforcement in preventing and investigating terrorist activity and helps coordinate the JTTF in Regional Terrorism Task Forces (RTTF). Director Mueller has also created a permanent Terrorism Watch List, a new Office of Intelligence, a new Integrated Intelligence Information Application (IIIA) database, and new hiring and recruiting initiatives. Please explain how the Department's FTTTF "lookout list" differs in substance and use from the FBI's Terrorism Watch List and how the FTTTF's "other intelligence-related projects" will differ from the functions of the FBI's JTTF, and IIIA database, and new Office of Intelligence.

(B) The FBI's new Office of Intelligence is intended to provide strategic analysis and gather information from current and past cases and other agencies, to look for patterns and analyze risks, and to meet the needs of other organizations responsible for homeland security. The separate FTTTF supervised by the Deputy Attorney General is required, with a budget of over $20 million, to conduct its own intelligence analysis projects and create and maintain its own databases and lookout list. Since Director Mueller routinely briefs the President with the CIA Director on terrorist threats, please explain why you decided to place the FTTTP in the Deputy Attorney General's office rather than within the FBI as part of its new Office of Intelligence? [2]

(C) The FBI has traditionally performed the critical intelligence-gathering mission under the supervision of a Director appointed for a ten-year term in a structure designed, in part, to insulate the exercise of Bureau powers from political considerations, and pursuant to formal guidelines and Congressional oversight. Are the investigative restrictions applicable to FBI agents also applicable to employees conducting data mining and operating the FTTTF under the guidance of the Deputy Attorney General?

(D) What information is necessary to trigger a data-mining inquiry on a particular individual or targeted activity to ensure that this technique is only being used for purposes relevant to detecting, preventing or punishing terrorism or other criminal activity?

3. Admiral Poindexter's Total Information Awareness Project (TIA). According to the Department of Defense, the Defense Advanced Research Project Agency (DARPA) has established the Total Information Awareness (TIA) Project to develop technologies for rapid language translation, commercial transaction data mining, and interagency analysis and decision-making tools.

(A) To what extent are you and the Department of Justice consulting or collaborating with Admiral Poindexter or the Department of Defense in designing and implementing TIA surveillance tools and related programs?

(B) Have any TIA generated or developed technologies been delivered to the Department of Justice and, if so, (i) are any being used? (ii) describe the purposes for which they are being used; and (iii) are any of the tools for data mining and pattern recognition?

(C) TIA has programs called Genoa I and II. Has this program been delivered in whole or in part to the Department of Justice and, if so, (i) is it being used? (ii) Describe the purposes for which it is being used; and (iii) is this a tool for data mining or pattern recognition?

(D) TIA has a program called EELD (Evidence Extraction and Link Discovery). Has this program been delivered in whole or in part to the Department of Justice and, if so, (i) is it being used? (ii) Describe the purposes for which it is being used; and (iii) is this a tool for data mining or pattern recognition?

(E) TIA has a program called Genisys. Has this program been delivered in whole or in part to the Department of Justice and, if so, (i) is it being used? (ii) Describe the purposes for which it is being used; and (iii) is this a tool for data mining or pattern recognition?

(F) TIA has a program called TIDES (Translingual Information Detection, Extraction and Summarization. Has this program been delivered in whole or in part to the Department of Justice and, if so, (i) is it being used? (ii) Describe the purposes for which it is being used; and (iii) is this a tool for data mining or pattern recognition?

(G) Is the FTTTF coordinating its work in any way with the TIA?

(H) What safeguards, if any, do you believe should be included in any data mining tools developed by TIA to ensure the accuracy and reliability of the information collected and stored in databases? Have you recommended such safeguards to the Department of Defense?

4. Compliance With The Privacy Act

(A) Does the Privacy Act impose any restriction on data-mining activities by the Department and, if so, what are those restrictions?

(B) Does the Department employ any outside contractors to perform data mining services and, if so, how does the Privacy Act apply, if at all, to the out-sourcing of data mining activities?

(C) The Privacy Act, 5 U.S.C. §552a(e)(4), requires agencies to "publish in the Federal Register upon establishment or revision a notice of the existence and character of the system of records." Have you promulgated any regulations regarding the FTTTF?

(D) The Privacy Act, 5 U.S.C. §552a(e)(4)(E), requires publication of the policies and practices of the agency regarding storage, retrievability, access, controls, retention and disposal of the records. Have you published such policies and practices regarding the FTTTF?

(E) Generally, the Privacy Act prohibits governmental agencies from disclosing records to another agency, unless it falls under the "routine use" exception. 5

U.S.C. §552a(b)(3). Does the Department rely on this "routine use" exception to obtain databases from other agencies for aggregation in the FTTTF and other databases within the Department?

(F) The Privacy Act, 5 U.S.C. §552a(e)(4)(D), requires Federal Register publication of "each routine use of the records contained in the system, including the categories of users and the purpose of such use." If the answer to (E) above is affirmative, has the Department published any Federal Register notice required by the Privacy Act? If so, please provide a copy of any such notice and, if not, please explain why.

(G) The Privacy Act imposes restrictions on "matching" programs conducted by the government or the private sector on behalf of the government, unless the matching is conducted "subsequent to the initiation of a specific criminal or civil law enforcement investigation" or "for foreign counterintelligence purposes." How does the Department ensure that the FTTTF and other Department databases comprised of aggregated data from other agencies are operated within these restrictions?

(H) Does the Department believe that any amendments to the Privacy Act would be helpful to facilitate data mining by the Department and, if so, does the Department intend to transmit to the Congress any amendments to the Privacy Act to clarify the legality of data-mining by Federal agencies?

5. Coordination With the Department of Homeland Security.

(A) The Homeland Security Act expressly authorizes the new department to request, access, receive, analyze and integrate information from government agencies and private sector entities, and to establish and utilize "a secure communications and information technology infrastructure, including data-mining and other advanced analytical tools, in order to assess, receive and analyze data and information. . . ." [P.L. 107-296, Sections 201(d)(1), (13), (14)]. Does the Department of Justice have any such express statutory authority to conduct data mining? If so, please describe that authority.

(B) Do you anticipate the Department of Justice's data mining operations being transferred to the new Department of Homeland Security? If not, please explain why.

(C) Do you believe it is valuable to have a coordinated data mining effort with one agency clearly held accountable for setting guidelines of data uniformity and reliability and, if so, which agency do you believe should take this primary position in order to avoid duplication of effort?

I appreciate your attention to this important matter.

Sincerely,

PATRICK LEAHY

Chairman

---

[1] A recently declassified FBI memorandum, dated April 14, 2000, makes this point with startling details about incidents of mistaken surveillance activity, including a Foreign Intelligence Surveillance Act (FISA) order being improperly implemented with unauthorized videotaping of a meeting; wiretapping a cellular telephone that had been dropped by the target and assigned to an innocent user, who "was therefore the target of unauthorized electronic surveillance for a substantial period of time;" unauthorized monitoring of an e-mail account; and "unauthorized searches, incorrect addresses, incorrect interpretation of a FISA order and overruns of ELSUR [electronic surveillance]."

[2] This question was originally directed to Deputy Attorney General Thompson in May 2002, but no response has been provided.

Posted on Tue, Jan. 09, 2007

`Data mining' may implicate innocent people in search for terrorists

By Greg Gordon

McClatchy Newspapers

▪ Data mining tells government and business a lot about you

WASHINGTON - In his first hearing Wednesday as the chairman of the Senate Judiciary Committee, Democratic Sen. Patrick Leahy of Vermont plans to examine federal "data-mining" programs, the computerized hunt for terrorists that can implicate innocent people.

Consider the case of American Airlines pilot Kieran O'Dwyer of Pittsboro, N.C.

O'Dwyer said Tuesday that U.S. Customs agents detained him for 90 minutes in 2003 when he got off an international flight in New York, telling him his name matched one on a government terrorist watch list.

Over the next 22 months or so, O'Dwyer said, he was temporarily detained 70 to 80 times by authorities who apparently were worried that he was a fugitive member of the Irish Republican Army.

It's not clear how O'Dwyer came under suspicion or how his name wound up on a terrorist watch list, but critics charge that such miscues can occur during the data-mining process, in which computers analyze multiple databases in search of suspicious patterns.

Amy Kudwa, a TSA spokeswoman, said she couldn't comment on O'Dwyer's circumstance, but that an average of 1,500 airline travelers applied each week for redress on the grounds that they'd been mistakenly included on terrorist watch lists. She said 33,000 had applied as of last April.

In a speech last month, Homeland Security Secretary Michael Chertoff said his agency was working to put a system of "one-stop redress" in place in 2007.

Leahy, a longtime congressional champion on privacy issues, plans to make "a signature issue" of protecting civil rights in the face of a "proliferation of government databanks and data mining" in the war on terrorism, said his chief spokesman, David Carle.

"He believes Congress is way overdue in taking stock of the surge in data mining by the government," Carle said.

Carle said the inquiry had gained import because of powerful new technologies, the outsourcing of data mining to private firms and "the Bush administration's lack of cooperation" with Congress' attempts to police these surveillance programs.

For years now, the Bush administration has invested heavily in data mining, viewing it as a valuable intelligence tool that can alert U.S. authorities to terrorist plots in their early stages. The government is reported to have spent tens of millions on such surveillance systems.

Among witnesses summoned to Wednesday's hearing is Jim Harper of the Cato Institute, a libertarian research center. Harper co-authored a paper last year that concluded data mining is a poor use of tax dollars, won't identify terrorists and will lead to "false positives" implicating innocent people.

In the paper, Harper and IBM engineer Jeff Jonas charge that "the one thing predictable about predictive data mining for terrorism is that it would be consistently wrong."

In testimony prepared for delivery Wednesday, Harper says data mining uses "massive amounts of data about Americans' lifestyles, purchases, communications, travels and many other facets of their lives. . . . This raises a variety of privacy concerns."

Coinciding with the hearing, Democratic Sen. Russ Feingold of Wisconsin plans to reintroduce a bill, co-sponsored by Leahy, that would require federal agencies to disclose all data-mining activities to Congress and the public.

In a 2004 report, the General Accountability Office identified at least 10 data-mining programs being used in the hunt for terrorists, and several others have emerged publicly since then.

Senate Democrats also are expected to closely question John Michael McConnell, President Bush's nominee to serve as the new intelligence czar, about 13 Pentagon data-mining contracts that his consulting firm has obtained since 1997.

Recently, the Transportation Security Administration revealed that it has conducted a program known as the Automated Targeting System, which assigns a risk assessment to every international air traveler, for four years.

O'Dwyer, the pilot, said he tried multiple ways to end his detentions but that customs agents dismissed a TSA letter clearing him, saying "it could be a forgery."

Not even persistent calls to the TSA and FBI from aides to Rep. Bob Etheridge, D-N.C., and Sen. Richard Burr, R-N.C., could solve the problem.

O'Dwyer said customs agents came to greet him by his first name, and one joked that "this was profiling against the Irish."

After missing numerous connecting flights home and having to pay to stay in New York hotels, O'Dwyer gave up flying internationally last May, forgoing about $10,000 in annual bonus pay.

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.

Traveler Data Program Defied Ban, Critics Say; Congress Barred Funds for DHS Development

BYLINE: By Spencer S. Hsu and Ellen Nakashima, Washington Post Staff Writers

DATE: December 9, 2006

The Department of Homeland Security violated a congressional funding ban when it continued to develop a computerized program that creates risk assessments of travelers entering and leaving the United States, according to lawmakers and privacy advocates.

Although congressional testimony shows that department officials apparently disclosed some important elements of the controversial Automated Targeting System program to lawmakers in recent months, several key members of Congress said that they were in the dark about the program and that it violated their intentions.

"Clearly the law prohibits testing or development" of such computer programs, said Rep. Martin O. Sabo (D-Minn.), who wrote the three-year-old prohibition into homeland security funding legislation. "And if they are saying that they just took some system, used it and therefore did not test or develop it, they clearly were not upfront about saying it."

Privacy advocates and members of Congress expressed growing skepticism this week about the legality, scope and effectiveness of the massive data-mining program -- particularly the creation of risk assessments on Americans that would be retained for up to 40 years -- whose existence was first disclosed in detail in a Nov. 2 notice in the Federal Register.

The department announced yesterday that after receiving more than 50 objections to the program, it has extended a public comment period for ATS from Dec. 4 to Dec. 29. Sen. Joseph I. Lieberman (D-Conn.) and Rep. Bennie Thompson (D-Miss.), incoming chairmen of the Senate and House homeland security committees, and others have questioned the effort and called for hearings or additional administration briefings.

Developed to help customs inspectors target narcotics and other contraband, ATS began scrutinizing air travelers entering and leaving the United States in the mid-1990s, said Jayson P. Ahern, assistant commissioner of U.S. Customs and Border Protection. After the 2001 terrorist attacks, it was used to assign risk assessments to cargo and passengers, officials' testimony and a February 2005 DHS report to Congress show.

Two years ago, it was expanded again to a limited but growing number of land border crossers, according to the report and Ahern. About 309 million land crossings and 87 million air crossings of U.S. borders are made each year.

Travelers are not allowed to see their risk assessments and must file Freedom of Information Act requests to view the original records on which the assessment is based.

The Center for Democracy and Technology said the program violated the 1974 Privacy Act because customs officials targeted U.S. travelers and shared their data with other agencies without notifying the public. Homeland Security officials say that notice was implicit in an announcement in 2001 about an older program.

"The lack of a notice at all was clearly illegal for however many years they claim this was in operation," said David Sobel, Electronic Frontier Foundation senior counsel.

"This is everybody's worst nightmare," said Kevin Mitchell, chairman of the Business Travel Coalition, who was angered by the revelation that profiles were being kept without travelers' knowledge.

Homeland Security officials said the funding ban applied only to successor programs to its aborted attempt in 2004 to use commercial databases to assign risk to domestic air passengers -- then known as CAPPS II and renamed Secure Flight -- not to preexisting programs.

Homeland Security Secretary Michael Chertoff acknowledged that the November notice was an attempt "to be even more transparent and write, in even clearer English, about what we were going to do." But in an interview with the National Journal, he expressed frustration with critics' surprise.

"Otherwise, why are we collecting the data?" he asked. "Just to have it to sit around?"

DHS leaders have described in speeches and congressional hearings their efforts over the years to process data from manifests and airline passenger records on U.S.-bound international flights to "detect anomalies and 'red flags' " for high-risk individuals.

DHS has been more explicit recently about ATS data mining and risk profiling, saying computer algorithms were used to "produce potential matches" of inbound and outbound travelers with "potential . . . connections to terrorist risk factors."

But senior Homeland Security officials made only a few short references in 2004 and 2005 to using the program to assess land travelers. At the time, they cited it only as a future possibility.

"Funding will allow us to develop and implement a version of ATS that, for the first time, will be able to identify potentially high-risk travelers in passenger vehicles," then-Customs and Border Protection Commissioner Robert C. Bonner told Congress.

ADDITIONAL SUBMISSIONS FOR THE RECORD

Hanson et al. vs. Rumsfeld, 2006; Case No. 06 CV 3118; Judicial Case Files; United States District Court Southern District of New York, Complaint; New York City.

Hanson et al. vs. Rumsfeld, 2007; Case No. 06 CV 3118; Judicial Case Files; United States District Court Southern District of New York, Stipulation of Voluntary Dismissal Pursuant to F.R.C.P. 41(a)(1)(ii); New York City.

Taipale, A. K. "Technology, Security, and Privacy: The Fear of Frankenstein, The Mythology of Privacy, and The Lessons of King Ludd." *Yale Journal of Law and Technology*. 7 Yale J.L. & Tech. 123; 9 INTL. J. Comm. L. & Pol'y 8. (Dec. 2004).

United States. Office of the Secretary of Defense. *The Privacy Act of 1974 Notice to Amend Systems of Records*. January 9, 2007.

United States. Department of Homeland Security. *Report of the Department of Homeland Security Data Privacy And Integrity Advisory Committee: Framework for Privacy Analysis of Programs, Technologies, and Applications Report No. 2006–01*. Adopted March 7, 2006.

○