DATA MINING: CURRENT APPLICATIONS AND FUTURE POSSIBILITIES

HEARING

BEFORE THE

SUBCOMMITTEE ON TECHNOLOGY, INFORMATION POLICY, INTERGOVERNMENTAL RELATIONS AND THE CENSUS

OF THE

COMMITTEE ON GOVERNMENT REFORM

HOUSE OF REPRESENTATIVES

ONE HUNDRED EIGHTH CONGRESS

FIRST SESSION

MARCH 25, 2003

Serial No. 108–11

Printed for the use of the Committee on Government Reform

Available via the World Wide Web: http://www.gpo.gov/congress/house
http://www.house.gov/reform

U.S. GOVERNMENT PRINTING OFFICE

87–229 PDF WASHINGTON : 2003

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

CONTENTS

DATA MINING: CURRENT APPLICATIONS AND FUTURE POSSIBILITIES

TUESDAY, MARCH 25, 2003

HOUSE OF REPRESENTATIVES,
SUBCOMMITTEE ON TECHNOLOGY, INFORMATION POLICY,
INTERGOVERNMENTAL RELATIONS AND THE CENSUS,
COMMITTEE ON GOVERNMENT REFORM,
Washington, DC.

The subcommittee met, pursuant to notice, at 9:30 a.m., in room 2154, Rayburn House Office Building, Hon. Adam Putnam (chairman of the subcommittee) presiding.

Present: Representatives Putnam, Miller, Turner, and Clay.

Staff present: Bob Dix, staff director; John Hambel, senior counsel; Chip Walker and Lori Martin, professional staff members; Ursula Wojciechowski, clerk; David McMillen, minority professional staff member; Jean Gosa, minority clerk; and Earley Green, minority chief clerk.

Mr. PUTNAM. A quorum being present, the Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census will come to order.

Good morning and welcome to the first in a planned series of hearings addressing the important subject of data mining technology or "factual data analysis," as some might refer to it.

Before we get into my opening statement, considering the events of the world today and the enormous pressures that this Congress and our President are under, I would ask that we pause for a moment of silence.

[Moment of silence.]

Mr. PUTNAM. Thank you.

There are a number of proven uses for this data mining technology which has played a prominent role in many arenas, public and private, for years. This morning we will work to define the technology itself and examine the parameters of its application. There is no secret that some have expressed concerns about the role of data mining, particularly in the context of privacy intrusions. We will attempt to explore the manner in which this technology will continue to be a valuable tool in a variety of governmental uses, not just those of national security, while also acknowledging the public interest in protecting the privacy of personal information. Data mining is a technology that facilitates the ability to sort through large amounts of information through data base exploration, extract specific information in accordance with defined criteria, and identify patterns of interest to its user.

As I understand the technology, the user has the ability to tailor a data mining program to a particular purpose by selecting a number of different data bases to search and setting the criteria for that search. Data mining technology has been utilized successfully for many years in both public and private sectors to identify and analyze data that might otherwise be overlooked or inaccessible. Examples of the variety of commercial or governmental uses associated with data mining software would include businesses being able to develop a targeted marketing campaign in an effort to identify prospective customers; government agencies expanding opportunities to track down tax evaders; detection of Medicaid or Medicare fraud; and corporations using this tool to estimate spending in revenue more accurately, just to name a few.

For example, a mortgage refinancing lender may seek to determine potential candidates for their services by attempting to identify mortgage holders who have lived in their homes for a certain period of time in a particular geographic location with a market value range of property at a certain level in order to target a special refinancing rate offer. As you can imagine, this type of technology is invaluable to a number of institutions. Because it is such a vast and evolving field, the subcommittee is very interested in exploring the uses and effects of this technology in subsequent follow-up hearings to address more particular applications.

While data mining may have many legitimate and worthwhile uses, we must always be vigilant of any potential encroachment on the privacy of the American public. We have great responsibilities as elected officials. We must protect the American ideals of life, liberty, and freedom. At times these ideals would seem to come into conflict with one another, and it's our job to ensure that we do all we can to protect the public while maintaining the faith entrusted to us by the Founding Fathers to protect the right of the people to privacy and freedom. Ben Franklin once said, "Those who would give up freedom for security deserve neither."

I would like to welcome the following witnesses who are offering their expert testimony before us today: The Honorable Paula Dockery, Florida State Senator; Dr. Jen Que Louie, president of Nautilus Systems, Inc.; Mark Forman, Associate Director of Information Technology and Electronic Government, Office of Management and Budget, our Nation's CIO; Gregory Kutz, Director of Financial Management and Assurance, General Accounting Office; and Jeffrey Rosen, associate professor of the George Washington University Law School, legal affairs editor of the New Republic. Mr. Armey was unable to be with us today.

Interest in expanding the use of this technology at the Federal level of government has become more widespread as we look to use modern technology to improve intergovernmental communications and national security. From our oversight perspective as the subcommittee, we have a special interest in learning the pros and cons to data mining technology as well as how its use could be or is being expanded at the Federal level.

We appreciate the participation of today's witnesses as they provide tremendous information to the subcommittee on this important topic, and we thank you again for taking the time out of your busy schedules. Today's hearing can be viewed live via WebCast by

going to reform.house.gov and clicking on the link under "Live Committee Broadcast."

As we await the ranking member from Missouri, I want to recognize our vice chair, Candace Miller from Michigan, for her opening statement. Gentlelady from Michigan.

[The prepared statement of Hon. Adam H. Putnam follows:]

TOM DAVIS, VIRGINIA
CHAIRMAN

HENRY A. WAXMAN, CALIFORN
RANKING MINORITY MEMBER

ONE HUNDRED EIGHTH CONGRESS
Congress of the United States
House of Representatives
COMMITTEE ON GOVERNMENT REFORM
2157 Rayburn House Office Building
Washington, DC 20515-6143

SUBCOMMITTEE ON TECHNOLOGY, INFORMATION POLICY, INTERGOVERNMENTAL RELATIONS AND THE CENSUS
Oversight Hearing

Hearing topic: ***"Data Mining: Current Applications and Future Possibilities"***

Tuesday, March 25, 2003
9:30 a.m.
Room 2154 Rayburn House Office Building

OPENING STATEMENT

Good morning and welcome to the first in a planned series of hearings addressing the important subject of data mining technology…or "factual data analysis" as some might refer to it. There are a number of proven uses for this technology, which has played a prominent role in many arenas, public and private, for years. This morning we will seek to define the technology itself and examine the parameters of its application. It is no secret that some have expressed concerns about the role of data mining, particularly in the context of potential privacy intrusions.

We will attempt to explore the manner in which this technology will continue to be a valuable tool in a variety of governmental uses…not just those of national security while also acknowledging the public interest in protecting the privacy of personal information. Data mining is a technology that facilitates the ability to sort through large amounts of information through database exploration, extract specific information in accordance with defined criteria, and then identify patterns of interest to its user.

As I understand the technology, a user has the ability to tailor a data mining program to a particular purpose by selecting a number of different databases to search, and setting the criteria for the search. Data mining technology has been utilized successfully for many years in both the private and public sectors to identify and analyze useful data that might otherwise be overlooked or inaccessible. Examples of the variety of commercial or governmental uses associated with data mining software would include; businesses being able to develop a targeted marketing campaign in an effort to identify prospective customers; government agencies expanding opportunities to track down tax evaders; detection of Medicaid and Medicare fraud and corporations utilizing this tool to estimate spending and revenue more accurately, just to name a few.

For instance, a mortgage refinancing lender may seek to determine potential candidates for their services by attempting to identify mortgage holders who have lived in their homes for a certain period of

time, in a particular geographic location, with a market value range of property at a certain level, in order to target a special refinancing rate offer.

As you can imagine, this kind of technology is invaluable to a number of institutions. Because it is such a vast and evolving field, the Subcommittee is interested in exploring the various uses and effects of this technology and in subsequent follow up hearings addressing more particular applications.

While data mining may have many legitimate and worthwhile uses, we must always be vigilant of any potential encroachment on the privacy of the American public. We have great responsibilities as elected officials. We must protect the American ideals of life, liberty and freedom. At times these ideals would seem to come into conflict with each other and it's our job to ensure that we do all we can to protect the public while maintaining the faith entrusted to us by the Founding Fathers -- to protect the right of the people to privacy and freedom. Ben Franklin once said, those that would give up freedom for security, deserve neither.

Today, we have a number of expert witnesses on data mining that will provide us with their professional insight. I'd like to welcome:

- The Honorable Paula Dockery, Florida State Senator;
- Dr. Jen Que Louie, President, Nautilus Systems, Inc.;
- Mark Forman, Associate Director, Information Technology and Electronic Government, Office of Management and Budget;
- Mr. Gregory D. Kutz, Director Financial Management and Assurance, U.S. GAO; and,
- Jeffrey Rosen, Associate Professor at the George Washington University Law School, Legal Affairs Editor of The New Republic.

Interest in expanding the use of this technology at the Federal level of government has become more widespread as we look to use modern technology to improve intergovernmental communications and national security. From our oversight perspective as the Subcommittee on Technology and Information Policy, we have a special interest in learning the pros and cons to data mining technology as well as how its use could be...or is being... expanded at the Federal level.

We appreciate the participation of today's witnesses as they provide valuable information to the Subcommittee on this important topic. Thank you again for taking time out of your busy schedules.

Today's hearing can be viewed live via WebCast by going to **http://reform.house.gov** and then clicking on the link under "Live Committee Broadcast".

Mrs. MILLER. Thank you, Mr. Chairman.

I want to thank the witnesses for coming today, and Mr. Forman, good to see you again. I'm sure this committee will be seeing certainly a lot of you.

As I mentioned at the last committee hearing, I am so particularly interested in the subjects, and this data mining is a fascinating one. I had been the Secretary of State in Michigan where not only did I have the elections there with all the registered voters, I also did the motor vehicle administrative kinds of things. We had a big data base in our State with everybody who had a boat, a snowmobile, and a trailer and a car and a truck and everything, and there was always a lot of consternation about what was government doing with this information; who had the information; for what purposes. If you wanted to get licensed in Michigan, you had to give me certain amounts of information. But what was government doing with it and what was the citizens' expectation of what we would do with all of that data?

There was a time when our State—and I know many States still do this—sell the information. It is a huge revenue source, of course. But I don't think citizens are normally expecting that the government will be selling their personal and private information. And so there is a consternation about who can access the information, how will it be massaged, how will it be utilized, and certainly on the part of the citizens, invasion of personal privacy by "Big Brother," by government.

As we march down the information highway, sometimes there is a slippery slope there that I think all of us in government at the Federal level, the State level, the county level, anyone that has interaction with these various data, that we always keep that uppermost in our mind about invasion of personal privacy.

With that being said, the technology is certainly out there and it can be utilized to make huge advances in society, and there are so many things in every layer of government that could be done so much better if we were able to use the technology properly. So I am very pleased to see you all today. Thank you for coming. I certainly look forward to hearing your testimony this morning. Thank you.

Mr. PUTNAM. I thank the gentlelady. She brings tremendous experience from her days as Secretary of State and work in bringing that office into the Information Age.

We are joined by a former mayor, the gentleman from Ohio, Mr. Turner. For your opening statement you are recognized.

Mr. TURNER. Thank you, Mr. Chairman. I am particularly interested in this area. NCR is located in Dayton, OH, which is a leading technology company in this issue of data mining for the private sector. And recently they hosted a forum on the issue of data mining applications, taking them from the private sector and applying them to government issues. And it was an interesting discussion because they began in telling us that Wal-Mart, at the end of the day, can tell us how many socks they have sold; but we are not necessarily able to tell ourselves, in reference to foreign visitors, how many visas have expired today and who they are.

So the possible applications of data mining on very simple tasks that clearly do not violate issues of privacy is a wide open field which we need to pursue vigorously.

Also the issue that was fascinating to me in their discussion is how you look at the process of data mining, not looking first at what data that you have, but looking at what questions do you want answered, and that the issue of technology is there. The issue of the application of technology is demonstrated in the private sector; the issue before us in government is to begin the process of asking what questions do we need to know answers to and then turning to the experts in data mining that have applied it in the private sector to assist us so we can have those answers in the public sector.

Thank you.

Mr. PUTNAM. I thank the gentleman.

We will now take the testimony from the witnesses. Each has been very gracious to prepare written testimony which will be included in the record of this hearing. And I have asked each of you to summarize your presentation into 5 minutes, if you could, to leave ample time for questions and answers. Witnesses will notice that there is a timer with a light on the witness table. Green light means you begin your remarks, the yellow light means it's time to wrap up, and the red light means that we hit the ejection seat.

In order to be sensitive to everyone's time schedule, we ask that you cooperate with us in our time schedule. As is the policy of the Committee on Government Reform, all witnesses will be sworn in. So I'll ask you to rise, please, and raise your right hands.

[Witnesses sworn.]

Mr. PUTNAM. All witnesses responded in the affirmative. Thank you.

I would like to introduce our witnesses first and then call on them for their testimony, followed by questions. We begin our panel with an old colleague of mine and a very dear friend from Florida, State Senator Paula Dockery. Florida is one of the States where data mining techniques have been used in several areas, and quite successfully. Senator Dockery's experience will lend a very helpful perspective to us today. She serves as majority whip in the Senate as well as chairman of the Committee on Homeland Security and Seaports. Senator Dockery, welcome to the committee and we look forward to your testimony, please.

STATEMENT OF STATE SENATOR PAULA DOCKERY, MAJORITY WHIP, FLORIDA STATE SENATE

Ms. DOCKERY. Thank you, Mr. Chairman, and good morning, Mr. Chairman and members of the committee. Thank you very much for the opportunity to be here today not only to share with you what we think we are doing right in the State of Florida, but also to be part of this distinguished panel and to learn from the experts to my left. I apologize in advance. I'm going to be reading so I can make my time limit, and I'm going to probably have to read pretty fast because I timed it at 7 minutes. But I would like to get started with that.

The issue of enhanced information sharing by our law enforcement and public safety professionals is at the forefront in our war

against terrorism in our efforts to keep America safe. Florida, I believe, has taken a strong leadership role in this effort, one that can serve as a model for other States. This model and its reliance on data mining is the focus of our discussion today.

Florida uses the term "factual data analysis" to describe this information processing system. This process includes the collection of information from multiple sources. Once this information is processed, analyzed, and evaluated, the resulting products represents the intelligence needed to assist law enforcement. Intelligence can then can be used in a proactive and preventive approach to detect criminal patterns, crime trends, modus operandi, financial criminal activity and criminal organizations.

Data collection is much different today than in years past. The number of data bases and the information contained there is immense, as is the ability to effectively and efficiently analyze available data in a timely manner. The results can be overwhelming. Factual data analysis plays a crucial role in filtering the vast quantity of information by separating the significant data from the insignificant data. Some individuals and groups voice concern for perceived loss of privacy and a perceived attempt to foster the examination of private information.

Florida's law enforcement efforts are aimed at utilizing only that specific data which law enforcement already has a legal right to use, while doing so in a proficient, professional, and expeditious manner. Many safeguards have been implemented to ensure appropriate use of information. These include user name and password protection, user training, agency user agreements, system audits, quality control reviews and established purge criteria.

Florida's intelligence criminal systems are operated in compliance with standards established by 28 Code of Federal Regulations, Part 23. This regulation was written to protect the privacy rights of individuals and to encourage and expedite the exchange of criminal intelligence information between and among law enforcement agencies. The regulation provides operational guidance for law enforcement agencies in five primary areas.

Prior to the September 11th attacks, Florida utilized factual data analysis on criminal investigations through the Financial Crime Analysis Center at the Florida Department of Law Enforcement. The Center integrates and analyzes financial data in partnership with local and Federal criminal justice agencies to identify and combat financial crimes.

The Center has developed a "data warehouse" which contains information from various sources already available to law enforcement. As part of the analytical process, the Center utilizes specialized software to identify anomalies associated with financial transactions. Analytical personnel and investigators then examine the results to determine if the information is related to a crime. The software currently used by law enforcement agencies provides a graphical representation of suspicious activity identified by financial services companies. This method ensures that the user does not see individual records, only the result, a safeguard that we believe is very important.

The pattern of behavior is a key element of the decision process of whether to investigate further. Users of this system are trained

to identify behaviors of known criminal activity during all stages of money laundering. It is important to note that by FDLE guidelines, reasonable suspicion is necessary before initiating an investigation.

When reasonable suspicion is developed, analyzed data are supplied to local State and Federal law enforcement agencies as well as to other States for possible investigation. This proactive approach results in increased team work amongst law enforcement entities as well as a force multiplier effect for the investigative process. FDLE agents regularly travel to other States to investigate common targets.

Arizona and Florida are known as the two most effective States in conducting these types of proactive investigations.

After the September 11th attacks, FDLE integrated this process and applied it toward the fight against terrorism. FDLE employed the assistance of public corporations that have access to civil data records. In certain domestic security related situations, FDLE has contracted with nationally recognized public search businesses to analyze the records based on criteria supplied by law enforcement. After the data is processed, the results are provided to law enforcement for further review. To ensure that the results are as indicative as possible, a mathematical analysis is used and includes as many as 14 criteria, producing a probability score for criminal behavior. Prior to additional investigation or dissemination, intelligence analysts and investigators examine only the results with the highest scores. This information can be used to identify, locate, target and monitor terrorists and other criminals. This ability is essential if future terrorist events are to be prevented.

Florida has partnered with a vendor, Seisint Technologies, to provide the data analysis tools using both public and private data. Over several years, Seisint Technologies has acquired technology and data for multiple sources useful to law enforcement. Following the terrorist attacks of September 11th, Seisint focused on helping local State and Federal law enforcement agencies locate and track individuals who might be a threat to the United States. As a result of their partnership with Florida law enforcement, a customized investigative tool was developed. This system has already proven useful in that a review of the known information intelligence and reported activities of the 19 hijackers associated with the terrorist events of September 11th identified several common and associated variables. This system has proven useful in Florida, but the need for timely sharing and exchange of information nationwide remains a critical need.

Mr. PUTNAM. Thank you Senator Dockery.

[The prepared statement of Ms. Dockery follows:]

TESTIMONY OF THE HONORABLE PAULA B. DOCKERY, FLORIDA STATE SENATOR, BEFORE THE SUBCOMMITTEE ON TECHNOLOGY, INFORMATION POLICY, INTERGOVERNMENTAL RELATIONS AND THE CENSUS COMMITTEE ON GOVERNMENT REFORM

THE HONORABLE PAULA B. DOCKERY
Chairman, Senate Committee on Home Defense, Public Security and Ports
The Florida Senate, District 15

Tuesday, March 25, 2003
Room 2154, Rayburn House Office Building
Washington, D.C.

Good Morning, Mr. Chairman and Members of the Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census. It is an honor to be here to testify before you this morning at a time when our nation faces such a challenging situation. As a member of the Florida Senate, I hope my comments here today will give you insight into the important work being performed in states around this nation to support the war on terrorism. As you will see, Florida has seized the initiative in this fight for the protection of our homeland.

The issue of enhanced information sharing by our law enforcement and public safety professionals is at the forefront in our war against terrorism and our efforts to keep America safe. Florida has taken a strong leadership role in this effort, one that can serve as a model for other states. This model and its reliance on data mining is the focus of our discussion today.

Factual Data Analysis

Data mining is a technological process, which provides the ability to sort through and analyze massive amounts of data in a systematic and logical manner. The core elements of this powerful analytical tool are data systems, decision trees, deviation detection, algorithms, and image analysis.

Florida uses the term Factual Data Analysis to describe this information processing system. This process includes the collection of information from multiple sources.

Once this information is collected, it is then processed, analyzed and evaluated resulting in the *intelligence* needed to assist law enforcement. This intelligence can then be used in a proactive and preventative approach to detect criminal patterns, crime trends, modus operandi, financial criminal activity, and criminal organizations.

Data collection is much different today than in years past. The number of databases and the information contained therein is immense. Factual Data Analysis plays a critical role in filtering the vast quantity of information by separating significant data from insignificant data. This analysis is crucial to effectively and efficiently analyze available data in a timely manner. Without a defined analysis method, the quantity of potential information results could be overwhelming.

Privacy Safeguards

Some individuals and groups voiced concern for a perceived loss of privacy and a perceived attempt to foster the examination of private information. Florida's law enforcement efforts are aimed at utilizing only that specific data which law enforcement already has a legal right to use, while doing so in a proficient, professional, and expeditious manner. Many safeguards have been implemented to ensure appropriate use of information. These include user name and password protection, user training, agency user agreements, system audits, quality control reviews and established purge criteria.

Florida's criminal intelligence systems are operated in compliance with standards established by 28 Code of Federal Regulations (CFR) Part 23. This regulation was written to protect the privacy rights of individuals and to encourage and expedite the exchange of criminal intelligence information between and among law enforcement agencies. The regulation provides operational guidance for law enforcement agencies in five primary areas: submission and entry of criminal intelligence information; system security; inquiry; dissemination; and, review and purge process.

Application of Factual Data Analysis: Financial Crime Analysis Center

Prior to the September 11th attacks, Florida utilized Factual Data Analysis on criminal investigations through the Financial Crime Analysis Center (FCAC) at the Florida Department of Law enforcement (FDLE). FCAC integrates and analyzes financial data in partnership with local and federal criminal justice agencies to identify and combat financial crimes.

FDLE's FCAC has developed a "data-warehouse" which contains information from various sources already available to law enforcement. As part of the analytical process, the Center utilizes specialized software to identify anomalies associated with financial transactions. Analytical personnel and investigators then examine the results to determine if the information is related to a crime. The software currently used by law enforcement agencies provides a graphical representation of suspicious activity identified by financial services companies. This method ensures that the user does not see individual records, only the result - a safeguard we believe important. The pattern of behavior is a key element of the decision process of whether to investigative further. Users of this system are trained to identify behaviors of known criminal activity during all stages of money laundering. It is important to note that, by FDLE guidelines, reasonable suspicion is necessary before initiating an investigation.

When reasonable suspicion is developed, analyzed data are supplied to local, state and federal law enforcement agencies, as well as to other states, for possible investigation. This proactive approach results in increased teamwork amongst law enforcement entities as well as a force multiplier effect for the investigative process. FDLE agents regularly travel to other states to investigate common targets. Arizona and Florida are known as the two most effective states in conducting these types of proactive investigations.

Arizona has identified tens of millions of dollars in laundering by illegal migrant smuggling groups. Florida has discovered major international narcotics smuggling

rings: Jamaican networks that move millions in narcotics; international organizations trafficking in heroin; and even suspicious money transactions that identified dozens of victims of Nigerian fraud scams.

Application of Factual Data Analysis: Terrorist Investigations

After the September 11th attacks, FDLE integrated this process and applied it toward the fight against terrorism. FDLE employed the assistance of public corporations that have access to civil data records. In certain domestic security related situations, FDLE has contracted with nationally recognized public search businesses to analyze their records based on criteria supplied by law enforcement. After the data is processed, the results are provided to law enforcement for further review. To ensure that the results are as indicative as possible, a mathematical analysis is used and includes as many as 14 criteria, producing a probability score for criminal behavior. Prior to additional investigation or dissemination, intelligence analysts and investigators examine only the results with the highest scores. This information can be used to identify, locate, target and monitor terrorists and other criminals. This ability is essential if future terrorist events are to be prevented.

Florida has partnered with a vendor, Seisint Technologies, to provide the data analysis tools using both public and private data. Over several years, Seisint Technologies has acquired technology and data from multiple sources useful to law enforcement. Following the terrorist attacks of September 11th, Seisint focused on helping local, state, and federal law enforcement agencies locate and track individuals who might be a threat to the United States. As a result of Seisint's partnership with Florida law enforcement, a customized investigative tool was developed. This system has already proven useful in that a review of the known information, intelligence and reported activities of the 19 hijackers associated with the terrorist events of September 11 identified several common and associated variables. This system has proven useful in Florida, but the need for timely sharing and exchange of information nationwide remains a critical need.

Project MATRIX

This critical need for timely sharing and exchange of information nationwide is being addressed with a pilot project: the Multistate Anti-Terrorism Information Exchange (MATRIX). This effort, which is partially funded through a grant from the Department of Justice, is a thirteen-state pilot project that utilizes Factual Data Analysis to increase and enhance the exchange of sensitive terrorist and criminal intelligence information. The project maximizes and integrates existing and proven technology while appropriately disseminating information nationwide in a secure, efficient and timely manner. The ulitmate goal is to expand this system to all states. Implementation of this pilot represents a critical component of a nationwide prevention plan. While some skepticism exists, the results of data analysis are made available only to law enforcement agencies, and then only on a need-to-know and right-to-know basis.

It is imperative that our law enforcement agencies have access to appropriate information. We have demonstrated that prior to September 11th, Factual Data Analysis was a successful tool for developing associations between people and organizations, tracking and identifying financial inconsistencies, and proactively partnering with other states and organizations. After the attacks, Florida joined forces with Seisint Technologies to create a system that analyzes diverse information in minutes. Such analyses would have taken hours, days or weeks prior to the utilization of this new "Factual Data Analysis" tool.

Finally, through MATRIX, law enforcement agencies nationwide will be able to maximize efforts and better prepare for future activities. The ultimate goal is to ensure the safety of our citizens. To this end, we should employ all the necessary tools to detect, prevent and respond to criminal and terrorist activity.

Thank you for your time. I look forward to working with you on these complex issues.

Mr. PUTNAM. I would like to introduce our next witness, Dr. Jen Que Louie. He has spent over 25 years working with data analysis systems, specifically with large data base systems, data warehousing and data mining. Some of his projects include designing, developing, and refining military logistics and C3I capability models for the Department of Defense. He has designed and implemented medical system diagnostic and analysis programs, knowledge- and rules-based business systems, work flow process and analysis systems, image management storage and retrieval systems, and emergency management information systems. Dr. Louie is president of Nautilus Systems, which is located in Fairfax, VA. We look forward to your testimony. Welcome to the subcommittee.

STATEMENT OF JEN QUE LOUIE, PRESIDENT, NAUTILUS SYSTEMS, INC.

Dr. LOUIE. Good morning, Mr. Chairman and distinguished members of the subcommittee. Thank you for the opportunity to testify today on data mining current applications and future possibilities. Other than my prepared statement, this is a quick summarization of data mining in general.

It is difficult to come up with a universal definition for data mining. One consistent focus of data mining has been basically that it is an analytic process with an ultimate goal of prediction. You are looking to find something that is going to be actionable, that is going to get you somewhere. In a nutshell, data mining is an extraction of knowledge or information from data. And at first glance, this may not seem like a very powerful utility, but unlike mere data, knowledge leads to incisive decisions and previously unknown relationships that could have a bearing on your decision process.

Data mining, unfortunately, like artificial intelligence of the early eighties, is getting a lot of media hype and we will call it slightly exaggerated benefits or feasibility of it. And what I usually tell my clients is the first fallacy is data mining tools. Data mining is a process. It is not a specific tool, and the process will generally raise more questions than it does produce answers. And while data mining does have the ability to uncover patterns that can be remarkable, it still requires a human with skills, analytical skills, to interpret the meaning of what patterns you are looking at.

And my usual examples are a Dilbert cartoon where the marketing person is telling the CEO, "Our product is always seen with people who have flu-like systems." And the product development team is the reason they have flu-like systems; it is because they are taking the product. So how you interpret the data, how you apply it is an important part of how you apply data mining.

Data mining is sometimes advertised and portrayed as being an autonomous process; that once you have these rules that you don't require analysts, and that is another fallacy. Another fallacy is that it will pay for itself very rapidly. While there is sometimes, we will call it articles, portraying very high returns for the investment in data mining, those are not very common. And yes, you can achieve a lot of return on your investment with data mining. Credit card fraud is one. Tax evasion is another. Money laundering. There are several tools that are out in the market that require a lot of exten-

sive capabilities. Our company has worked with FinCEN on clearing a lot of their caseloads. Those, I would say, are great paybacks for the amount of money invested in those areas.

Data mining also sometimes raises the question about missing data. Sometimes the data that's missing is more interesting than the data that is there, and that provides some other insights. Meeting your data mining expectations, planning is the single most important step in any data mining effort. You have to know and understand what the consumers of your information product need and basically deliver it. Once you determine what that is, the next thing in your investment in your data mining effort is the environment that you run it in. It should be what we call the best you can get, the fastest you can get, the most storage you can get, and always allow yourself plenty of time to review and analyze the data and look at all the facets that are there in order to determine that you are delivering the right message, and it is actionable in the direction that user needs that information to be.

So, my quick summation: Data analysis is concerned with the discovery and examination of patterns and associations found with data. There are various ways to achieve this objective, but all share the same fundamental notion that patterns examined are present in the data. Also remember that what is not in data can be just as interesting in certain situations, and more useful to know.

Data mining is a process that involves multiple analytical tools, methodologies driven by the needs of the information product's consumer. The quality of information is directly proportional to the trustworthiness and quality of that data. The confidence of the prediction is dependent upon the data mining practitioner's subject matter expertise and insight to deliver actionable results. The data mining process is highly computational, takes time; therefore, planning the approach and selection of tools is influenced by the needs of the consumer. Thank you.

Mr. PUTNAM. Thank you very much, Dr. Louie.

[The prepared statement of Dr. Louie follows:]

Data Mining: Current Applications and Future Possibilities

Testimony before the Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census, March 25, 2003

Jen Que Louie, President, Nautilus Systems, Inc.

Thank you Mr. Chairman, Mr. Ranking Member, and other members of the Committee on Government Reform for the opportunity to testify today on the subject of "Data Mining: Current Applications and Future Possibilities." I will summarize my thoughts briefly in the first pages of my prepared statement and opening remarks, and include more detailed explanation about what data mining is, dispel some of the fallacies about data mining, and finally address what is required for successful data mining analysis and meeting your data mining expectations.

What Is Data Mining?

Depending upon whom you ask, a universal definition of what data mining exactly is can be next to impossible. While the definition seems to be in constant metamorphosis, data mining is an analytic process, whose goal is prediction. The data mining process applies one or more algorithms (computations, queries, links, or sorts) to explore extremely large volumes of data in the hope of discovering patterns and identifying relationships, that were previously unknown, and ultimately make a prediction.

In a nutshell, data mining is the extraction of knowledge or information from data. Apparent as it may seem at first glance, this concept is a deceptively powerful one. Unlike mere data, knowledge can (1) lead to incisive decisions and (2) reveal previously unknown relationships.

Data Mining Fallacies

The first fallacy is that there are "data mining tools." In fact, data mining is a process. With data mining, you do not just turn loose a plethora of analytic tools, and thus find answers. In reality, data mining will raise more questions. While data mining's ability to uncover data patterns can be remarkable, it requires human skills to interpret the results accurately.

A second fallacy is that the whole data mining process can be autonomous and does not require an analyst once a pattern or set of rules has been identified. This is only true for the specific instance, which the rules were generated from. For example, if a minimum purchase is made with a credit card and that transaction is followed by the purchase of expensive items, then there is a high probability that the credit card is stolen. This rule is only applicable to identifying possible credit card fraud.

A third fallacy is that the savings realized using data mining pays for itself very rapidly. That depends on what "rapidly" means, along with the cost of the tools, the computational engine, the analyst's time, and your business operation model. Generally

speaking, data mining is computationally intensive and returns a lift of less than one percent to the bottom line.

A fourth fallacy is that advertised data mining packages are "easy to use and intuitive." Very unlikely, but if you understand the problem you are trying to answer and the tool or tools meet those needs – you are very lucky. Chances are you will require someone with subject matter expertise who is intuitive, analytical, and mathematically inclined to view the overall process, analyze the results, and then make those results actionable events.

Successful Data Mining Analysis

"Data mining is more about letting the data speak for itself," Linoff[1] says.

Data mining differs from other traditional analytical processes by the way data is queried. An analyst using traditional analytical processes usually approaches the problem by constructing a hypothesis or identifying the specific needs to be addressed and using the data available to prove or disprove the hypothesis. Data mining, by comparison, involves targeting a specific problem and using algorithms to form general hypotheses that may expose patterns and relationships that were previously unseen. On the whole, data mining is more predictive in nature than traditional tools that tend to either support or disprove a hypothesis.

An example of how data mining differs from a traditional analysis approach of querying available data may be useful. Let us say that a school district administrator's student information database contains historical data about the students enrolled in that district's schools. The administrator wants to know how test scores vary among students from different economic backgrounds. The administrator uses the available data and formulates a query.

A query using a traditional approach may be structured something like: "High school students from low-income households tend to score lower on tests than students from high-income households. Is this true?" The analyst would then generate the appropriate query language for the student information database and generate a report that either supports the hypothesis as correct or wrong. The data might show that students from low-income households do score lower on tests, but overall, the results do not provide much more related information.

Applying the data mining process to the same student information system may identify related information that provides more insight and value. The results might show students from low-income households do tend to score lower on tests; but at the same time, it may also point to other reasons contributing to this pattern. The data mining process might group students who have part-time jobs, come from single-parent households, are not enrolled in a tutorial program, have a learning disability, have recently moved to that school district, or are frequently absent, as factors that contribute to low test scores. Data mining identifies relationships and interdependencies affecting an objective – subject course grades.

Although my example is simple, it illustrates how data mining can unearth unseen and often overlooked pieces of information. This identified information is now actionable information that the administrator can use to apply solutions to the problem.

Putting Data Mining to Use

Over the past few decades we have collected more and more data, to the point that we have no idea what we have. However, with the availability of affordable fast computational processing capabilities in recent years, data mining can make sense of this data for specific business, educational, intelligence or other purposes.

The one significant shortfall of data mining is that it requires massive quantities of data to be effective. The quality of the prediction is directly proportional to the quality (trustworthiness) and quantity of the data, and the final value of the prediction is dependent on the data mining practitioner's subject matter expertise and insight to deliver actionable results.

When What Is Missing Is What Is Interesting

Sometimes while mining data, the data mining process will kick out an anomaly or flag trends and patterns that should be in the data. The data mining practitioner, on review, will either ignore it or follow the thread. Why does it matter?

Case study: The state of California appears to have as many as half a million residents that fail to file state income taxes. Through the use of data mining, the California Franchise Tax Board is able to identify the fraudulent practice of not filing taxes. This was accomplished by examining past tax returns (historical data) and third-party data (Federal W-2 forms). The state of California was able to determine who should have filed a return. [2] The use of historical and third party data makes it relatively easy to determine expected trends and patterns and then detect the absence of them.

The goal here was to identify the absence or lack of a pattern in data, and it is this absence that is flagged as truly interesting. This approach may prove to be especially valuable.

Technology Used By Data Mining Practitioners

The technologies used by data mining practitioners are primarily based on statistical methods such as linear regression, factor analysis, and distribution analysis. The data mining process has extended these foundation algorithms to include more complex and innovative tools that can identify frequencies, associations, temporal events, and patterns from data being mined.

Tools that are used in connection with data mining are often categorized by their origins, and usually consist of methods (processes) involving neural networks, clustering, decision trees, classifications, linked lists, correlation, and other numeric methods.

- *Neural networks* use artificial intelligence (AI) or machine learning processes, which use deductive reasoning, make intelligent estimates, and learn by example.
- *Decision trees*, originally developed for operations research, provides best-fit logical path solutions.
- *Classification and clustering algorithms* provide methods for explicit data segmentation. One of the most publicized instances of clustering is in the Geographic Information System (GIS) field, where analysts can, with startling accuracy, identify relative consumption of products within counties and ZIP codes.
- *Linked lists* map the relation to any data point; for example, parents to their children, children to their spouses, resulting in another cycle of parents to their children. This is applicable to following money, identify potential laundering, or other types of fraud.
- *Correlation matrixes* match the same data elements on the X-axis and Y-axis. For example, we list phone numbers on the X and Y-axis, and at every intersecting point we enter the number of times (frequency) that the other number calls a particular number. When we graph the matrix, you will be able to visualize the communication relationships. In other words, we might see that groups of subscribers by geographic areas call the same pharmacy, transportation schedule recording, or weather report. This information could be utilized to schedule preventative maintenance or upgrades, or evaluate communication equipment utilization by area.
- *Data cubes* or *Multi-Dimensional Database* (MDDB), are often categorized as a data mining product. An MDDB is a repository holding aggregations of data in cells which are the intersection of multiple dimensions (time, geography, product, customer) of the data.
- *OLAP* (On-Line Analytical Processing) is usually associated with MDDB and has the ability to analyze data across multiple dimensions in a timely manner, in order to support critical decision making.
- *Numerical methods* in a broad generalization encompass all data mining processes and applications.
- Hybrid Tools
    - *CART* is a proprietary algorithm developed by Salford Systems, Inc. that uses a nontraditional decision tree methodology. It has a high degree of automation (requires only moderate supervision by the analyst), and has the ability to handle arbitrarily complex data structures. Salford Systems claims that a novice-generated first iteration CART model is often as good as a neural net model developed by an expert.
    - *Eigen analysis* is a multivariable statistical procedure that may be either a prediction or classification technique. It is also capable of discrimination analysis,

partial correlation, multiple regression, principal component analysis, and factor analysis.

- *Origami*[3] is a hybrid link analysis application developed on the concept of data-cartography that maps the relationship of data points to each other in a visual representation.
- *Information mapping*[4] is based on research into how the human mind actually reads processes, remembers, and retrieves information. Nautilus Systems' hyperbolic directory applies information-mapping principles in breaking complex information into its most basic elements and then presents those elements optimally for users. The result is a set of precisely defined information modules that are consistent from author to author and document to document.
- *Intelligent agents* are autonomous software computer programs which can dig through data repositories unsupervised and returned with the requested information, monitor for changes in data, or even track who is requesting the data element.
- *Taxonomy-directed intelligent agents* are capable of human-like understanding of text, and can modify the agent's behavior and responses accordingly. This software application, when combined with TRW's FDF® 4 processor, is the fastest, most accurate adaptive information filtering system in the world. It is designed to search, filter and categorize massive quantities of free (i.e., unindexed) text and distribute the search results to multiple users. A single FDF® 4 chip contains 96 parallel processors, with more than a million transistors on each chip. Paracel[5] TextFinder™ uses the FDF® 4 chip, is commercially available, each unit contains more than 12,000 processors, and multiple TextFinder™ units can be easily clustered together to create a seamless, integrated system that provides virtually unlimited scalability. Paracel's TextFinder™ has a proven track record as the undisputed leader in large-scale, text-filtering applications.

Meeting Your Data Mining Expectations

Planning is the single most important step in any data mining endeavor. Know and understand what the consumers of your information product need. Then get the best you can afford hardware and software to enhance the environment that will meet your analyst's performance expectations from the outset.

Data mining environments grow more complex and demanding, and sometimes in a short span of time. Design your computational environment with scalability in mind. If your system is not easily scalable, you will have serious performance bottlenecks and major upgrading costs later.

Understand your consumer's operational and information needs. The success of your data mining efforts depends on how well you respond to the dynamics of your consumer's environment.

Time is your worst enemy and faster computers do not necessarily translate into faster insight. Remember, it takes a woman nine months to produce a baby, and no matter how hard you try, you cannot get nine women to make a baby in one month. Allow time for quality assurance and review before delivery of the information product.

Do not underestimate the need for training. Even the brightest science and international law graduates can be shockingly unprepared to take advantage of the tools you are providing them. Do not assume a level of expertise they may not have. Be prepared to provide a substantial amount of training, especially in the area of turning strategic questions into structured queries.

Summation

Data analysis is concerned with the discovery and examination of patterns and associations found in the data. There are various ways to achieve this objective, but all share the fundamental notion that patterns to be examined are present in the data. Also remember that what is not in the data can be just as interesting and in certain situations more useful to know.

Data mining is a process that involves multiple analytic tools and methodologies, driven by the needs of that information product's consumer.

The quality of the information product is directly proportional to the trustworthiness and quantity of the data available.

The confidence of the prediction is dependent on the data mining practitioner's subject matter expertise and insight to deliver actionable results.

The data mining process is highly computational and takes time. Therefore, planning the approach and the selection of tools is influenced by the needs of the consumer.

---

[1] Michael Berry and Gordon Linoff are co-authors of several books on data mining, including "Mastering Data Mining" and "Data Mining Techniques For Marketing, Sales, and Customer Support." They are also the founders of Data Miners<http://www.data-miners.com>, a consultant agency specializing in data mining training and planning.

[2] Round table discussion at Salashan '99 High Performance Computing Conference, statement made by Dr. Inderpal Bhandari, founder and CEO of Virtual Gold, Inc., <http://www.virtualgold.com/>and internationally recognized expert in data mining.

[3] Presentation by Jen Que Louie at: KDD-99: The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, entitled "Origami: A New Data Visualization Tool", by Jen Que Louie and Tom Kraay, Mission Valley Marriott Hotel, San Diego, California, August 15-August 18, 1999.

[4] Robert E. Horn, while a student at Harvard University and Columbia University, conducted research about how readers deal with large amounts of information. This resulted in a standard approach for communicating information, which is based on learning theory, human factors engineering, and cognitive science.

[5] Paracel Inc. <http://www.paracel.com/>

Mr. PUTNAM. Our next witness is Mark Forman. He served as Associate Director for Information Technology in E-Government for the Office of Management and Budget, a position he has held since June 2001. He is effectively in charge of information technology oversight for the entire Federal Government. And his—he has a background in the private sector from Unysis and IBM as well as work at the Senate Governmental Affairs Committee staff. He is an invaluable resource on all of our IT issues, and we believe his insight from the Federal perspective will be enlightening to us as well. So with that, Mr. Forman, you are recognized.

STATEMENT OF MARK A. FORMAN, ASSOCIATE DIRECTOR, INFORMATION TECHNOLOGY AND ELECTRONIC GOVERNMENT, OFFICE OF MANAGEMENT AND BUDGET

Mr. FORMAN. Thank you, Mr. Chairman, and members of the subcommittee. Thank you for the opportunity to appear and to discuss the administration's views on data mining. And I also want to thank you for taking a very rational, well-balanced approach in exploring data mining issues and opportunities. While there are many definitions of data mining, the committee's definition is generally accepted and we believe helpful in defining the issues and its challenges.

I would like to start by talking about private sector uses how we are using it in the Federal Government, and then the challenges and opportunities. The private sector uses data mining to make sense of a wide breadth of data. Some examples are customer relationship management. Applied to customer relationship management, data mining is used to analyze disparate customer data and provide insights into customer needs and wants. Companies that use data mining shorten response time to market changes, which allows for better alignment of their products with the customer needs. They do this to increase revenue performance and allocate investment to products that meet customer demand effectively.

Fraud detection. Companies use software that provide comprehensive transaction-level financial reporting and analysis to support automatic fraud detection and proactive alerting.

Retail analysis and supply chain analysis. Companies such as Wal-Mart are broadly recognized for analyzing sales trends. Retail analysis and supply chain analysis can be used to predict the effectiveness of promotions, decide which products to stock in each store, and help managers understand cost and revenue trends in order to adjust pricing and promotion in anticipation of changes in marketplace conditions.

Medical analysis and diagnostics. The health care industry uses analysis to predict the effectiveness of surgical procedures, medical tests and medications. High-risk segments of the population can be identified and targeted for proactive treatment. The result is improved quality of life for patients, reduced stress on hospitals and insurance providers using such activities as proactive approaches to healing, I think it is fair to say, and I have many more examples of the commercial use of data mining. All of them deal with how fast we can understand what customers need, and the Federal Government would be well advanced to be able to respond more quickly to what our citizens need.

So I will turn now to the government applications of data mining and go through some of the examples and more of the effects, both the way we deal with the citizens and how we manage the government.

The Federal Government analyzes data that has been collected from the public for several purposes, including determining the eligibility of applicants for Federal benefits, detecting potential instances of fraud, waste, and abuse in Federal programs and for law enforcement activities. Some of this analysis is facilitated by data mining.

So let us talk through a few of the examples. First, financial management. Poor management practices create opportunities for a wide range of fraud and abuse in the use of government travel and purchase cards. Several agency inspector general investigations have used data mining-type tools to document inappropriate purchases and misuse of cards. OMB is taking and will continue to take substantive affirmative steps to ensure agencies improve their internal control systems to monitor expenditures appropriately.

Human resource management. One of the 24 E-Government initiatives, which we call the Enterprise H.R. Integration, and which is managed by the Office of Personnel Management, is leading the effort to provide a governmentwide data warehouse of H.R. information to minimize the workload as employees move from one department to another. A key component of this is the E-Clearance project. OPM and its partner agencies on the E-Clearance project are using data mining to more quickly access information which speeds up the overall security clearance investigation process.

Reducing erroneous payments and fraud detection. Data analysis accomplished by the matching of electronic data bases between government agencies has been an important and successful tool for identifying improper payments under Federal benefit and loan programs, as well as detecting potential instances of fraud, waste, and abuse in the Federal programs. As highlighted in the President's 2004 budget, agencies are now required to report the extent of erroneous payments made in the major benefit program. Through the President's Management Agenda Initiative for improving financial performance, we are getting a hand on the problem of erroneous payments. Furthermore, the administration has proposed several pieces of legislation regarding the administration's authority to share data that will greatly improve efforts erroneous payments.

Policy analysis. The quality of policy decisions is a function of our ability to correctly analyze enormous amounts of data that describe a problem faced by modern society. For example, the Department of Education mines data from a variety of student financial aid systems, permitting professionals to analyze Federal education programs quickly and easily without the time expense and burden on citizens.

Law enforcement and homeland security. Federal agencies have found data mining techniques to be an important tool for assisting law enforcement in combating terrorism. For example, a system such as the Department of Homeland Security's Bureau of Customs and Border Protection operates the Automated Commercial Envi-

ronment which utilizes a series of data mining tools to strengthen border security efforts.

Benefits and pitfalls. While the use of data mining to access timely data and to identify relationships that were previously known as powerful tools for identifying errors, fraud, threats, etc., the application of such techniques to personal information raises serious questions about privacy and how it should be protected. In my written statement I focused on two areas. First, the data analysis must be consistent with law. We monitor that with business cases. Second, the Federal Information Security Management Act further requires protection of the data under security processes and techniques. Mr. Chairman, thank you.

Mr. PUTNAM. Thank you very much.

[The prepared statement of Mr. Forman follows:]

MARK A. FORMAN
ASSOCIATE DIRECTOR FOR INFORMATION TECHNOLOGY AND E-GOVERNMENT
OFFICE OF MANAGEMENT AND BUDGET
BEFORE THE
SUBCOMMITTEE ON TECHNOLOGY, INFORMATION POLICY, INTERGOVERNMENTAL RELATIONS, AND THE CENSUS
COMMITTEE ON GOVERNMENT REFORM
UNITED STATES HOUSE OF REPRESENTATIVES

MARCH 25, 2003

Mr. Chairman and Members of the Subcommittee,

Thank you for the opportunity to appear before the Subcommittee to discuss the Administration's views on data mining.

This committee has defined "data mining" as a "technology that facilitates the ability to sort through masses of information through database exploration, extract specific information in accordance with defined criteria, and then identify patterns of interest to its user." While there are many definitions of "data mining", the Committee's definition is generally accepted and helpful in defining the issue and its challenges. Additionally, data warehouses are being used as the source of data for many data mining applications. A data warehouse is a managed data repository of integrated, cleansed data whose source is mainly transactional data. Data is aggregated from various sources and structured for the use of analysis and reporting.

Commercial Types and Uses of Data Mining

The private sector uses data mining to make sense of the wide breadth of data that companies and industries have available. Some examples of these uses:

- **Customer Relationship Management/ Segmentation Analysis**-- Applied to Customer relationship management (CRM), data mining is used to analyze disparate customer data and provide insight into customer needs and wants. Data mining is used to analyze and segment customer buying patterns and to identify potential goods and services that are in demand. Companies that use data mining shorten response time to market changes, which allows for better alignment of their products with their customers' needs. They do this to increase revenue performance and allocate investment to products that meet consumer demand effectively.

- **Fraud Detection** – Companies use software that provides comprehensive, transaction-level financial reporting and analysis to support automatic fraud detection and proactive alerting. Software packages can also be used to detect

anomalies, variances, and patterns in databases. For example, BlueCross/BlueShield and other health care payers use data mining tools to catch and prevent fraudulent and abusive billing practices. BlueCross/BlueShield's solution can quickly search through millions of medical claims and detect inappropriate billing practices with a high degree of reliability.

- **Retail Analysis and Supply Chain Analysis** – Companies such as Wal-mart are broadly recognized for analyzing sales trends. Retail analysis and supply chain analysis can be used to predict the effectiveness of promotions, decide which products to stock in each store, and help managers understand cost and revenue trends in order to adjust pricing and promotions in anticipation of changes in marketplace conditions. Data mining also allows supply chain tools to monitor and analyze inventory trends, forecast product demand for replenishment, track vendor performance and identify problems, analyze distribution network efficiency, and understand supply chain costs and inefficiencies.

- **Medical Analysis/Diagnostics** – The health care industry uses analysis to predict the effectiveness of surgical procedures, medical tests, and medications. High-risk segments of the population can be identified and targeted for proactive treatment. For example, American Healthways relies on predictive modeling to identify patient types who trend toward high-risk conditions, giving care coordinators a proactive approach to healing. The result is improved quality of life for the patients and reduced stress on hospitals and insurance providers.

- **Document Analysis (Text Mining)** – Documents can be searched for information and insights in a fraction of the time an individual will spend locating one document. Document analysis involves analysis of text and structured and unstructured data, organized by categories, to determine trends, pattern and relationships and organized by categories. This can be highly effective in survey analysis. Content management systems and software packages perform analyzes on an organization's information products to help companies control information flows and work products. For example, Autonomy at BAE Systems aggregates content from many sources in many different formats, structured or unstructured, including their intranet and 10,000 news feeds per day. The goal is to personalize the delivery of that information to each user, and to eliminate work duplication and time-consuming searches. Autonomy automatically alerts BAE Systems employees to documents in the system that relate to what they're doing, or to other employees in the company whose interests and expertise match their own.

- **Use of Decision Support Systems (DSS) --** Decision Support Systems may use data mining to identify trends and present the information in intuitively useful ways -- supporting more informed and effective decisions for business and organizational activities. For example, one DSS solution for HR management is now providing essential insights into The Bank of Scotland Group's HR activities

worldwide, giving managers personnel and staffing information needed to make hiring and placement decisions. Managers can determine if job turnover in a particular area or occupation classification is higher than expected and investigate influences on loyalty such as the physical working environment.

- **Financial Analysis --** The insurance industry uses and data mining algorithms to conduct risk analysis, such as evaluating actuarial experience studies for mortality, withdrawal and disability, dynamically calculating exposures and expectations for period ranges. For example, Canada Life performs timely and accurate actuarial studies using a data warehouse and advanced data analysis methods; the Generali Group uses data mining tools to manage financial market risk and customer credit risk via a common analytical framework for rapid and flexible analysis and reporting of risk exposure.

Government Applications of Data Mining

The Federal government analyzes data that has been collected from the public for several purposes, including determining the eligibility of applicants for Federal benefits, detecting potential instances of fraud, waste, and abuse in Federal programs, and for law enforcement activities. Some of this analysis is facilitated by data mining. Here are a few examples of agency uses of data analysis techniques and software:

- **Financial management --** Poor management practices have created opportunities for a wide range of fraud and abuse in the use of government travel and purchase cards. Several agency inspector general (IG) investigations have used statistical sampling processes to document inappropriate purchases and misuse of these cards. OMB is taking and will continue to take substantive, affirmative steps to ensure agencies improve their internal control systems to monitor expenditures properly.

- **Human Resources Management** - One of the 24 E-government initiatives, the Enterprise HR Integration under the Office of Personnel Management, is leading the effort to provide a government wide data warehouse of HR information to minimize the workload as employees move from one department to another. A key component of this is the E-Clearance project – OPM and its partner agencies on the E-clearance project are using data mining to more quickly access information which speeds up the overall security clearance investigation process. Given the backlog in clearances, this use of data mining is critical to our ability to get staff for effectively and rapidly through the human resources management processes.

- **Reducing Erroneous Payments and Fraud Detection** - Data analysis accomplished via the matching of electronic databases between government agencies has been an important and successful tool for identifying improper payments under federal benefit and loan programs, as well as detecting potential instances of fraud, waste, and abuse in Federal programs. As highlighted in the FY 2004 President's Budget, agencies are now required to report the extent of erroneous payments made in their major benefit programs. In addition, the last decade has shown an increased reliance and increased spending on non-discretionary social services, such as Medicare and Medicaid. These expenditures -- and therefore the potential for improper payments -- are likely to increase unless appropriate steps are taken to protect against errors and fraud. Through the President's Management Agenda initiative for improving financial performance, we are getting a handle on the problem of erroneous payments. For example, Medicare's erroneous payment rate has fallen from 6.8 percent to 6.3 percent and the Food Stamp program reduced its national error rate from 8.9 percent to 8.7 percent. Just these small rate reductions prevented the waste of almost $1 billion. Furthermore, the Administration has proposed several pieces of legislation regarding the Administration's authority to share data that will greatly improve efforts to reduce erroneous payments.

- **Policy Analysis** – The quality of policy decisions is a function of our ability to correctly analyze enormous amounts of data that describe a problem faced by modern society. For example, the Department of Education mines data from a variety of its student financial aid systems, including the Central Processing System, Pell Grant Payment System and National Student Loan Data System, permitting professionals to analyze Federal education programs quickly and easily, without the time, expense, and burden on citizens of paper-driven surveys.

- **Law enforcement and Homeland Security** – Federal agencies have found data mining techniques to be an important tool for assisting law enforcement combating terrorism. For example, system such as the Department of Homeland Security's Bureau of Customs and Border Protection operates the Automated Commercial Environment (ACE) can utilize a series of data mining tools to strengthen border security efforts. ACE will provide the IT mechanisms for making quick evaluations on whether particular people or goods should be deemed high-risk or low-risk. Also, ACE will enable the Department of Homeland Security and other Federal agencies to more precisely target for inspection or investigation the highest risk people and cargo crossing the border. Through tools such as ACE, agencies have the ability to instantaneously analyze vast amounts of data and intelligence to see links among businesses and people, thus revealing security threats that might otherwise have gone unnoticed.

- **Citizen access to government data** -- Search sites such as the one available at the FirstGov website provide a facility for searching vast amounts of unstructured data across the Federal government by using publicly available search engines. In addition, the Federal government conducts its own data analyses for statistical purposes and facilitates data user access to statistical data. For example, the Census Bureau's "American FactFinder System (Advanced Query)" uses a data mining tool to allow users to query Census 2000 detailed data files. The tool provides simplified access to and extraction of data.

Benefits and Pitfalls

As outlined above, the government has found a number of ways to use collected information to improve program effectiveness and to reduce misuse of taxpayer dollars. While the use of data mining techniques to access useful, timely data and to identify relationships that were previously unknown is a powerful tool for identifying errors, fraud, threats, etc., the application of such techniques to personal information raises serious questions about privacy and how it should be protected. In order for this to be accomplished, the government must continue to act in several areas:

1. Federal data analyses must be consistent with law

In the federal arena, data mining activities must be implemented consistent with the protections of the Privacy Act of 1974, as amended by the Computer Matching and Privacy Protection Act of 1988, and other privacy statutes. These statutes do not address data-mining per se, but they outline privacy principles the government must follow in data collection, including: notice and reasonable disclosure; use and purpose limitations; choice; access to government-held information, information security; redress; and oversight. Agencies are well-versed in the legal, policy, and technical requirements governing access to and sharing of personal data. Agencies may aggregate information by analyzing data across databases, a concept known as "virtual data warehousing"; however, when information can be accessed or exchanged at numerous locations by many users, a potential exists for inadvertent disclosure of personal information or misuse of personal information, by alteration or for unauthorized purposes. Agencies that adhere to the existing legal and policy structure including OMB and NIST policy guidance can protect personal information in their possession even as they participate in data-mining activities. Furthermore, the E-Government Act of 2002 requires that an agency conduct a Privacy Impact Assessment (PIA) when agencies develop or procure information technology to initiate a new online collection of information that involves personally identifiable information changing hands, such as in the case of matching.

2. Ensuring the Security of Federal IT Systems

The Federal Information Security Management Act (FISMA) provides a comprehensive framework for ensuring the effectiveness of information security controls over federal

information resources, including resources that result from data mining. FISMA requires the head of each agency to periodically assess the risk and magnitude of harm that could result from unauthorized access, use, modification, or disclosure of information. The agency must then provide information security protections that are commensurate with the stated risk. Agencies are required to periodically test their information security controls and techniques to ensure that they are effectively implemented. The results of this testing are reported to OMB on an annual basis.

Conclusion

"Data mining" can have many uses. The Administration is strongly committed to using available technologies like data mining to serve citizens and protect citizens from other threats, while the Administration is also strongly committed to protecting the privacy of citizens when such tools are used. Through data analysis and data mining, the private sector has improved customer service and customer needs, and has been able to help customers take proactive approaches to health care. The federal government has reduced the number of erroneous payments, and has been able to determine patterns in databases that help predict both weather patterns and the spread of deadly viruses.

We need to use modern analytic tools, such as data mining, to improve government performance, from policy analysis to fraud to homeland security. We can maintain privacy and security while improving government productivity, but we must employ tools like data mining appropriately. We hope to work with this Committee to ensure that the benefits of data analysis continue to help Federal agencies to perform their missions, while protecting against the problems that aggressive and abusive data mining can cause.

Mr. PUTNAM. For insight from a Federal agency that uses data pattern analysis, we have Gregory Kutz, Director of Financial Management and Assurance at the General Accounting Office. As a Director in the Financial Management Assurance Team, Mr. Kutz is responsible for financial management issues relating to the Department of Defense, NASA, the State Department, and AID. He has also been recently involved in preparation of reports issued by GAO and testimony relating to credit card fraud and abuse at DOD, financial and operational management issues at the IRS, financial condition and cost recovery practices of the Department of Energy's Power Marketing Administration, the Tennessee Valley Authority, and AMTRAK.

You have been very busy. We look forward to your testimony.

STATEMENT OF GREGORY KUTZ, DIRECTOR, FINANCIAL MANAGEMENT AND ASSURANCE, U.S. GENERAL ACCOUNTING OFFICE

Mr. KUTZ. Thank you, Mr. Chairman, and members of the subcommittee. I'm here to talk about our use of data mining in audits of Federal programs. To date we have used data mining primarily as an integral part of our audits of credit card programs.

My testimony has two parts: First, the use of data mining in our audits and investigations; and second, future uses of data mining and related challenges.

First, our strategy is to use data mining to put a face on issues of breakdowns in internal controls. It allows us to go beyond simply saying that a program is vulnerable. For example, data mining allowed us to report that government credit cards were used for escort services, women's lingerie, prostitution, gambling, cruises, and Los Angeles Lakers tickets.

Our data mining has helped us to identify specific instances of fraud, waste, and abuse. The posterboard shows several examples of government travel card abuse that we identified through data mining, including the purchase of a used car from Budget Rental Car; adult entertainment charges, including gentlemen's clubs; Internet and casino gambling, including an individual who charged $14,000 to pay for his blackjack gambling habit and reimbursed travel money used to pay for closing costs on a home purchase. For each of these examples, we used various data mining inquiries to identify the transactions and completed the case with auditor and investigator followup.

The second posterboard is an excerpt from a government purchase card statement. As you can see, somebody went on a Christmas shopping spree. This bill, which includes nearly $12,000 of fraudulent charges, was identified using data mining. We identified these fraudulent transactions because of the suspicious vendors and because of the timing of the transactions. We used these findings in conjunction with systematic internal control testing to make recommendations to Federal agencies to develop effective systems and controls that provide reasonable assurance that fraud, waste, and abuse are minimized.

An important element of our success with data mining is the synergy of auditors and investigators working together. Our auditors have expertise in financial systems, data manipulation, and evalu-

ating internal control systems. Our investigators bring a much different perspective. For example, Special Agent Ryan, who is with me today, has several decades of experience working on financial crimes for the Secret Service. Investigators and auditors work together to assess system vulnerabilities and develop our data mining strategies.

Moving on to my second point, our data mining work for the Congress is expanding. Currently, we have a number of audits underway that use data mining, including nine that I am directly responsible for. Some examples of our expanded data mining audits include DOD vendor payments, Army military pay systems, HUD housing programs and Department of Energy national laboratories. As we move forward, challenges will include data reliability and security issues.

For the credit card work to date, we have used commercial bank data bases to do our data mining, which we found to be highly reliable. However, as we move beyond the credit cards, one major challenge is the poor quality of Federal Government data bases. In most cases, data base quality issues can be overcome, but they result in less productive data mining and a greater cost to our work.

Data security and privacy protection is another challenge. For example, in handling large data bases of credit card transactions, we developed strict protocols to protect this sensitive data. We were especially concerned with protecting credit card account numbers and individuals' Social Security numbers. Data security issues must be addressed before embarking on audits involving data mining.

In summary, data mining is a powerful tool that has increased our ability to effectively audit Federal programs. We are just beginning to make full use of data mining strategies. With the right mix of technology, human capital expertise, and data security measures, we believe that data mining will continue to improve our audit and investigative work for the Congress. Mr. Chairman, that ends my statement.

Mr. PUTNAM. Thank you Mr. Kutz. And I want to thank all the witnesses for being so gracious and complying with our time limitations.

[The prepared statement of Mr. Kutz follows:]

United States General Accounting Office

GAO

Testimony

Before the Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census, Committee on Government Reform, House of Representatives

For Release on Delivery
Expected at time 9:30 a.m. EST
Tuesday, March 25, 2003

DATA MINING

Results and Challenges for Government Program Audits and Investigations

Statement of Gregory D. Kutz, Director
Financial Management and Assurance

GAO-03-591T

March 25, 2003

GAO
Accountability · Integrity · Reliability
Highlights

Highlights of GAO-03-591T, a report to the Subcommittee on Technology, Information Policy, Intergovernmental Relations and the Census, Committee on Government Reform, House of Representatives

DATA MINING

Results and Challenges for Government Program Audits and Investigations

Why GAO Did This Study

The Subcommittee asked GAO to testify on its experiences with the use of data mining as part of its audits and investigations of various government programs. GAO's testimony focused on (1) examples and benefits of the use of data mining in audits and investigations and (2) some of the future uses and challenges in expanding the use of data mining in audits of federal programs. Much of GAO's experience with data mining to date relates to its audits of the Department of Defense's (DOD) credit card programs.

What GAO Found

GAO's data mining work related to audits and investigations of federal government credit card and other programs has identified fraud, waste, and abuse resulting from breakdowns in internal controls. We used these data mining techniques, in conjunction with systematic internal control testing, to make recommendations to federal agencies to develop effective systems and controls that provide reasonable assurance that fraud, waste, and abuse in these credit card and other programs are minimized. For these programs, GAO's data mining often involves extracting information on credit card users or vendors using a set of defined criteria (e.g., vendors that the federal government would not typically do business with) and then having auditors and investigators follow-up on selected transactions or vendors.

Data mining alone is generally not sufficient to identify systemic breakdowns in controls and to provide management with recommendations to improve systems of internal controls. Systemic breakdowns can best be demonstrated using statistical tests of key controls along with a thorough assessment of the overall control environment. Data mining results serve to "put a face" on the control breakdowns and provide managers with examples of the real and costly consequences of failing to properly control these large programs.

Recent GAO audits using data mining of DOD purchase and travel card programs have identified numerous prohibited purchases of goods and services from vendors such as restaurants, grocery stores, casinos, toy stores, clothing or luggage stores, electronics stores, gentlemen's clubs, legalized brothels, automobile dealers, and gasoline service stations.

GAO's use of data mining has expanded beyond the government credit card programs. At the request of several congressional committees and Members, we currently have underway a number of audits and investigations that will utilize data mining, including.

- DOD vendor pay systems
- Army military pay systems
- Department of Housing and Urban Development housing programs
- Department of Energy national laboratories

Challenges to expanding the use of data mining in the federal arena include data integrity and security issues. For example, DOD has long-standing problems with financial systems that are fundamentally deficient and are unable to provide timely and reliable data. Data security issues related to the use of large, detailed databases are another issue that must be considered before undertaking a data mining project. With the right mix of technology, human capital expertise, and data security measures, GAO believes that data mining will prove to be an important tool to help it to continue to improve the efficiency and effectiveness of its audit and investigative work for the Congress.

www.gao.gov/cgi-bin/getrpt?GAO-03-591T.

To view the full report, including the scope and methodology, click on the link above. For more information, contact Gregory D. Kutz at (202) 512-9095 or kutzg@gao.gov.

Mr. Chairman and Members of the Subcommittee:

Thank you for the opportunity to discuss current applications and future possibilities for the use of data mining. We use the term "data mining" to mean analyzing diverse data to identify relationships that indicate possible instances of previously undetected fraud, waste, and abuse. Auditors can use data mining to extract individual, or a series of, questionable transactions from large data files for follow up by auditors or investigators. Data mining can also help serve as a deterrent to those who believe they can get away with fraud because of weak or nonexistent internal control systems.

To date, GAO has used data mining as an integral part of our audits and investigations of federal government credit card programs. For these programs, our data mining work has identified fraud, waste, and abuse resulting from breakdowns in internal controls. We used these findings, in conjunction with systemic internal control testing, to make recommendations to federal agencies on actions needed to develop effective systems and controls that provide reasonable assurance that fraud, waste, and abuse in these credit card programs are minimized. My testimony will (1) discuss examples and benefits of the use of data mining in our audits and investigations and (2) some of the possible future uses and challenges to expanding our data mining beyond federal government credit card programs.

Use of Data Mining in Federal Government Audits and Investigations

Data mining has been an integral part of our audits and investigations of federal government purchase and travel card programs. For these programs, data mining has involved obtaining large databases of credit card transactions and related activity and using software to search or "mine" data looking for suspicious vendors, transactions, or patterns of activity. Our data mining often involves extracting information on credit card users or vendors using a set of defined criteria (e.g., vendors that the federal government would not typically do business with) and then having auditors and investigators follow-up on selected transactions or vendors. (See attachment 1 for a list of related GAO products resulting from our data mining.)

We have used data mining for credit card audits in conjunction with our evaluation of the design and effectiveness of internal controls intended to prevent fraud, waste, and abuse in these programs. Our methodology for performing these audits included the following four basic steps:

- gain an understanding of the credit card program;
- make a preliminary assessment of the adequacy of internal controls;
- test the effectiveness of internal controls; and
- identify, using data mining, case studies demonstrating the cause and real life effect of the control breakdowns.

An important element of success in our audits is the integration of our audit and investigative functions. Our auditors and investigators work together on a daily basis on all four steps of the process. In developing effective data mining strategies, we found that it is critical for the auditors and investigators to have a thorough understanding of the program and the related processes and internal controls. Once the process and controls are understood, we then assessed the adequacy of key internal control activities and the overall control environment. For example, in making this assessment for the Department of Defense (DOD) purchase card program, we identified a weak overall internal control environment, including a proliferation of credit cards, which left the program vulnerable to fraud, waste, and abuse. In addition, once vulnerabilities are identified, investigators and auditors work together to identify various schemes that could be used to abuse the program including committing fraud. Our understanding of the program and its vulnerabilities is then used to develop our data mining strategy.

We used data mining and follow on audit and investigative work to demonstrate the effect of systemic breakdowns in internal controls. Data mining alone is generally not sufficient to identify systemic breakdowns in controls and to provide management with recommendations to improve systems of internal controls. Systemic breakdowns can best be demonstrated using statistical tests of key controls along with a thorough assessment of the overall control environment, including existing policies and procedures that govern control activities.

Data Mining Criteria and Techniques Used in DOD Purchase and Travel Card Program Audits

The use of purchase cards has dramatically increased in past years as agencies have sought to lower transaction processing costs and eliminate the lengthy processes and paperwork long associated with making small purchases. DOD is promoting department wide use of purchase cards for obtaining goods and services. It reported that for the year ended September 30, 2002, purchase cards were used by about 214,000 cardholders to make about 11 million transactions valued at over $6.8 billion. Purchase cards may be used for acquisitions at or below the $2,500 micropurchase threshold, and for payment of items costing over $2,500 from contracts or other purchase agreements. DOD estimated that in fiscal year 2001, about 95 percent of its transactions of $2,500 or less were made by purchase card.

In 1983, the General Services Administration (GSA) awarded a governmentwide master contract with a private company to provide government-sponsored, contractor-issued travel cards to be used by federal employees to pay for costs incurred on official business travel. The intent of the travel card program was to provide increased convenience to the traveler and to reduce the government's cost of administering travel by reducing the need for cash advances to the traveler and the administrative workload associated with processing and reconciling travel advances. Our audits of DOD's travel card program focused on individually billed accounts, which are held and paid by individual cardholders. According to GSA, as of September 30, 2002, DOD had over 1.3

million individually billed travel cardholders who charged $2.4 billion during the fiscal year.

We assessed controls over the Army, Navy, and Air Force purchase and travel card programs. In each case, we found that a weak overall control environment and breakdowns in key internal control activities left the military services vulnerable to fraud, waste, and abuse. We looked for indications of potential fraud, waste, and abuse as part of our statistical sampling and through nonrepresentative selections of transactions using data mining. Because DOD's purchase and travel card programs involved different key control activities and vulnerabilities, we tailored our data mining techniques to address the unique characteristics of each program. However, we did not look at all potential abuses of either the purchase and travel card and our work was not designed to identify, and we did not attempt to determine, the full extent of potential fraud, waste, and abuse related to the purchase and travel card programs.

For our purchase card audits, we obtained transaction databases for our study period from the purchase card contract banks—U.S. Bank for the Army and Air Force and Citibank for the Navy. For our travel card audits, we obtained transaction databases for the three military services from DOD's travel card contractor—Bank of America. In all cases, control totals from these databases were reconciled to bank or GSA reports to ensure we had a complete and accurate database for our sampling and data mining. Using several database manipulation software tools, we selected transactions or patterns of activity that appeared to represent potential fraud, waste, or abuse. We then conducted additional audit and investigative follow-up based on the nature, amount, timing, and other characteristics of the transactions. In some instances, we also compared ("bumped") data from different databases to identify anomalies. Our data mining criteria included the following.

Nature of the transaction

- Prohibited merchant category codes[1] that should have been blocked, such as jewelry stores, pawn shops, and gambling establishments.
- Personal use, including food, clothing, luggage and accessories, such as sunglasses, purses, and totes.
- Travel related transactions, such as airfare, hotels, and restaurants (for purchase card audits).

[1] Merchant category codes (MCC) are established by the banking industry for commercial and consumer reporting purposes. Currently, about 800 category codes are used to identify the nature of the merchants' businesses or trades, such as airlines, hotels, ATMs, jewelry stores, casinos, gentleman's clubs, and theaters.

Merchants

- Specialty stores, such as hobby shops, sporting goods stores, Victoria's Secret, L.L. Bean and toy stores (e.g., Toys 'R' Us).
- "Dot com" vendors, such as REI, SkyMall, Internet gambling sites, and pornography sites.
- High-end stores, such as Dooney & Bourke, Coach, and Louis Vuitton.
- Department stores, such as Nordstrom and Macy's.
- Other personal use vendors, such as Ticketmaster, Mary Kay Cosmetics, and Avon.
- Gentlemen's clubs and legalized brothels.
- Cruise lines, sporting events, casinos, taxidermy services, and theaters.

Dollar Amount of Transaction

- Transactions having unusually high dollar amounts (for travel card audits).
- Convenience checks over $2,500 (for purchase card audits).
- Numerous recurring transactions with the same vendor indicating the need for a contract (for purchase card audits).
- Transactions in round dollar amounts, such as $330, $440, etc., indicating possible fee for cash schemes (for travel card audits).
- Multiple, recurring small ATM transactions, indicating possible personal use (for travel card audits).

Timing of Transactions

- Holiday and weekend transactions.
- End of fiscal year transactions.
- Transactions that were made late at night.
- Multiple transactions on the same day, at same vendor, totaling more than $2,500, indicating split purchases (for purchase card audits).

Other Characteristics

- Out of state purchases, when similar items have been purchased locally (for purchase card audits).
- Transaction in which the cardholder and merchant had the same name.
- Cardholders who wrote nonsufficient funds checks (for travel card audits).
- Charged-off accounts, and accounts in salary offset or fixed payment plans (for travel card audits).

To fully develop the case study examples that we included in our reports required extensive collaboration on the part of auditors and investigators. It is clear that data mining techniques, although a powerful tool by themselves, are best used in combination with strategies that create a synergy between teams of auditors and investigators to identify and develop case studies on the causes and effects of any control breakdowns. Our auditors have expertise in financial systems, data manipulation, and evaluating internal controls. Our investigators are federal agents with years of law enforcement experience, particularly in the area of detecting financial crimes. Further, we found that the experience gained with each successive audit increased the knowledge base of our auditors and investigators and improved the overall data mining results.

Data Mining Results in DOD
Purchase and Travel Card Program Audits

Data mining "puts a face" on the control breakdowns and provides managers with examples of the real and costly consequences of failing to properly control these large programs. Recent GAO audits using data mining of DOD purchase and travel card programs have identified numerous prohibited abusive or questionable purchases of goods and services from vendors such as restaurants, grocery stores, casinos, toy stores, clothing or luggage stores, electronics stores, gentlemen's clubs, legalized brothels, automobile dealers, and gasoline service stations.

Specific examples of abusive and questionable activity identified as a result of the previously discussed data mining criteria and techniques include

- Nature of the transaction: blocked merchant category code (MCC) – As part of our audit of the Army purchase card program, we identified a cardholder transaction for $630 that was coded as being from an escort service, which should have been a blocked MCC code. As part of our investigation we determined that this was an unauthorized, potentially fraudulent transaction, and that the cardholder was also being investigated for possible theft of chapel funds.
- Merchants – Gentlemen's Clubs and Brothels – We found that DOD cardholders used their government travel cards at legalized brothels in Nevada and at gentlemen's clubs that provide adult entertainment. We initially identified this

abusive use of the travel card based on our interviews with cardholders. Subsequently, we used this information to refine our data mining and identify a substantial number of these transactions.

- Merchants – Taxidermy Services – An Air Force cardholder used the purchase card to prepare a shoulder mount of a mule deer head. The deer was a "road kill" that was found on the roadside by an approving official who approved the purchase of taxidermy services. The deer head was hung on the wall in the Natural Resources Office. The cardholder, approving official, and two other employees occupy the office where the deer head currently hangs.

- Dollar Amount of Transaction: High Dollar Purchases – For the Army travel program, we found that a cardholder's spouse used his government travel card to make two payments of $2,050 each to Budget Rent-A-Car for the purchase of a used automobile.

- Dollar Amount of Transaction: Recurring Purchases – During fiscal year 2001, the Navy purchased over $1 million from 122 different vendors using the purchase card. In total, these vendors were paid about $330 million. However, despite this heavy sales volume, the Navy had not negotiated reduced-price contracts with any of the vendors.

- Timing of Transaction – In an audit of the Navy purchase card program, we identified about $12,000 in potentially fraudulent fiscal year 2000 transactions. These purchases occurred primarily between December 20 and December 26, 1999, and included an Amana range, Compaq computers, gift certificates, groceries, and clothes.

In addition, we used data mining techniques to identify 220 cardholders that abused their travel card or had been involved in potentially fraudulent activity and who had severe financial problems. We compared records for these cardholders with DOD databases that included security clearance information. Based on this analysis, we found that 97 of 220 individuals with severe financial problems continued to maintain secret or top-secret security clearances at the end of our respective audits.

Data Mining Results at Other Federal Agencies

We have used data mining techniques to help assess the controls over various programs at the Departments of Housing and Urban Development (HUD) and Education and the Federal Aviation Administration, among others. Further, our October 2001 Executive Guide entitled, *Strategies to Manage Improper Payments: Learning From Public and Private Sector Organizations* (GAO-02-69G), discusses the use of data mining techniques by various state and federal programs as part of a research-based approach to fraud prevention and detection. For example, the Illinois Department of Public Aid used data mining techniques to identify health care providers that were billing for services provided in excess of 24 hours in a single day. Their analysis identified 18 providers that had billed over 25 hours for at least 1 day during the 6 months ended December 31, 1999.

As a result, the Illinois Department of Public Aid Office of Inspector General planned to refer serious cases to appropriate law enforcement agencies and take administrative action against the less serious violators.

Additional examples of the results of our data mining at other agencies include the following:

- At the Department of Education, we performed a variety of data mining queries and found that three schools fraudulently disbursed about $2 million in Pell Grants to ineligible students and another school improperly disbursed about $1.4 million in Pell Grants to ineligible students.

- At the Department of Housing and Urban Development (HUD), we identified a scheme where only one-third of the work paid for by HUD to replace a concrete sidewalk was actually performed. As a result, more than $164,000 of the $227,500 billed and paid for appeared to be fraudulent.

Future Use of Data Mining and Related Challenges

Our use of data mining has expanded beyond government credit card programs. This expansion provides opportunities for significant impact and improvements in other programs but also presents other challenges. At the request of several congressional committees and Members, we currently have a number of audits, which will utilize data mining. These audits include the following.

- DOD Vendor Pay Systems – This effort is an evaluation of the adequacy and effectiveness of DOD's controls over its vendor pay processes. With reported annual vendor payments in excess of $77 billion, this program entails most of DOD's disbursements for items (excluding major weapons systems).
- Army Military Pay Systems – This effort is an evaluation of the Army's controls over the payroll payments to military members. For fiscal year 2002, Army's reported payroll was about $32 billion.
- Centrally-billed travel accounts – These accounts are used primarily to purchase transportation including airline tickets. This activity was about $1.5 billion for fiscal year 2002.
- Governmentwide purchase card program – We are evaluating whether the federal government is effectively managing its procurements of $15 billion in goods and services using purchase cards.
- HUD single and multifamily properties - As a follow-on to previous work, we are evaluating the propriety of payments made related to HUD-owned single and multifamily properties.
- Department of Energy contractor-managed national laboratories - In response to allegations of improprieties at the Los Alamos national laboratory, we are assessing internal controls over disbursements and whether purchases made are a valid use of government funds at selected other laboratories.

For each of these audits, we are in the process of developing and/or executing data mining strategies to assist with the identification of breakdowns in controls or the inefficient use of federal funds. In addition, in response to a congressional request, we are preparing a guide to assist federal agencies in their efforts to audit internal controls of government purchase card programs. We have found that as government purchase card use grows, federal and state and local government auditors are increasingly being asked to do more audits of these programs. Building on the lessons learned from our purchase card work, our guide is intended to provide a blueprint for other auditors to use when auditing purchase card programs. This guide will include a section on data mining and related follow-up.

For the credit card work to date, we have used databases provided by the contractor banks. We found that the data quality is high, thus allowing us to do efficient and effective data mining. However, a challenge with federal government databases is that the quality and availability of information from which to mine data is often poor. For example, we have previously reported that DOD's financial systems are fundamentally deficient and are unable to provide data in a timely and reliable manner for decisionmaking. These data problems result in the following challenges for future data mining.

- For DOD, data needed for effective data mining may not be available in any one system. Consequently, obtaining and reconciling data from numerous databases is necessary to develop populations from which to data mine. In addition, because of the large volume of transactions involved in many DOD program areas, storing and conducting data mining queries of such large files may present a significant challenge.

- Because databases do not reconcile to independent, reliable sources, the completeness of databases used for data mining is questionable.

- Many agencies have known problems with data reliability.

In most cases these issues can be overcome, but they result in less productive data mining, and increase the cost of doing the work.

Other challenges lie in the area of data security and privacy protection. For example, as part of our extensive use of many detailed databases to assess the controls over DOD's credit card programs, we developed strict protocols to protect the sensitive data included in the databases. We were especially concerned with protecting active credit card account numbers and individual social security numbers. Data security issues must be addressed before embarking on audits involving data mining.

Conclusions

The use of data mining is a critical component of the audit and investigation of certain federal programs. The results of data mining show real consequences or effect of breakdowns in internal controls. In addition, data mining results contribute greatly to

the development and implementation of recommendations to management on improvements in controls that can provide assurance that fraud, waste, and abuse is minimized. We are in the process of moving beyond the use of data mining for government credit card programs to other areas of interest to the Congress. We are just beginning to make full use of data mining strategies. With the right mix of technology, human capital expertise, and data security measures, we believe that data mining will prove to be an important tool to help us to continue to improve the efficiency and effectiveness of our audit and investigative work for the Congress.

Contacts and Acknowledgments

For future contacts regarding this testimony, please contact Gregory D. Kutz at (202) 512-9095. Individuals making key contributions to this testimony included Francine DelVecchio, Steve Donahue, Gayle Fischer, Geoffrey Frank, John Kelly, Mai Nguyen, John Ryan, Kara Scott, and Scott Wrightson.

Attachment 1

Related GAO Products

Travel Cards: Control Weaknesses Leave Navy Vulnerable to Fraud and Abuse. GAO-03-147. Washington, D.C.: December 23, 2002.

Travel Cards: Air Force Management Focus Has Reduced Delinquencies, but Improvements in Controls Are Needed. GAO-03-298. Washington, D.C.: December 20, 2002.

Purchase Cards: Control Weaknesses Leave the Air Force Vulnerable to Fraud, Waste, and Abuse. GAO-03-292. Washington, D.C.: December 20, 2002.

Travel Cards: Control Weaknesses Leave Army Vulnerable to Potential Fraud and Abuse. GAO-03-169. Washington, D.C.: October 11, 2002.

Travel Cards: Control Weaknesses Leave Navy Vulnerable to Fraud and Abuse. GAO-03-148T. Washington, D.C.: October 8, 2002.

Financial Management: Strategies to Address Improper Payments at HUD, Education, and Other Federal Agencies. GAO-03-167T. Washington, D.C.: October 3, 2002.

Purchase Cards: Navy Is Vulnerable to Fraud and Abuse but Is Taking Action to Resolve Control Weaknesses. GAO-02-1041. Washington, D.C.: September 27, 2002.

Travel Cards: Control Weaknesses Leave Army Vulnerable to Potential Fraud and Abuse. GAO-02-863T. Washington, D.C.: July 17, 2002.

Purchase Cards: Control Weaknesses Leave Army Vulnerable to Fraud, Waste, and Abuse. GAO-02-844T. Washington, D.C.: July 17, 2002.

Purchase Cards: Control Weaknesses Leave Army Vulnerable to Fraud, Waste, and Abuse. GAO-02-732. Washington, D.C.: June 27, 2002.

FAA Alaska: Weak Controls Resulted in Improper and Wasteful Purchases. GAO-02-606. Washington, D.C.: May 30, 2002.

Government Purchase Cards: Control Weaknesses Expose Agencies to Fraud and Abuse. GAO-02-676T. Washington, D.C.: May 1, 2002.

Education Financial Management: Weak Internal Controls Led to Instances of Fraud and Other Improper Payments. GAO-02-406. Washington, D.C.: March 28, 2002.

Purchase Cards: Continued Control Weaknesses Leave Two Navy Units Vulnerable to Fraud and Abuse. GAO-02-506T. Washington, D.C.: March 13, 2002.

Purchase Cards: Control Weaknesses Leave Two Navy Units Vulnerable to Fraud and Abuse. GAO-02-32. Washington, D.C.: November 30, 2001.

Purchase Cards: Control Weaknesses Leave Two Navy Units Vulnerable to Fraud and Abuse. GAO-01-995T. Washington, D.C.: July 30, 2001.

(192095)

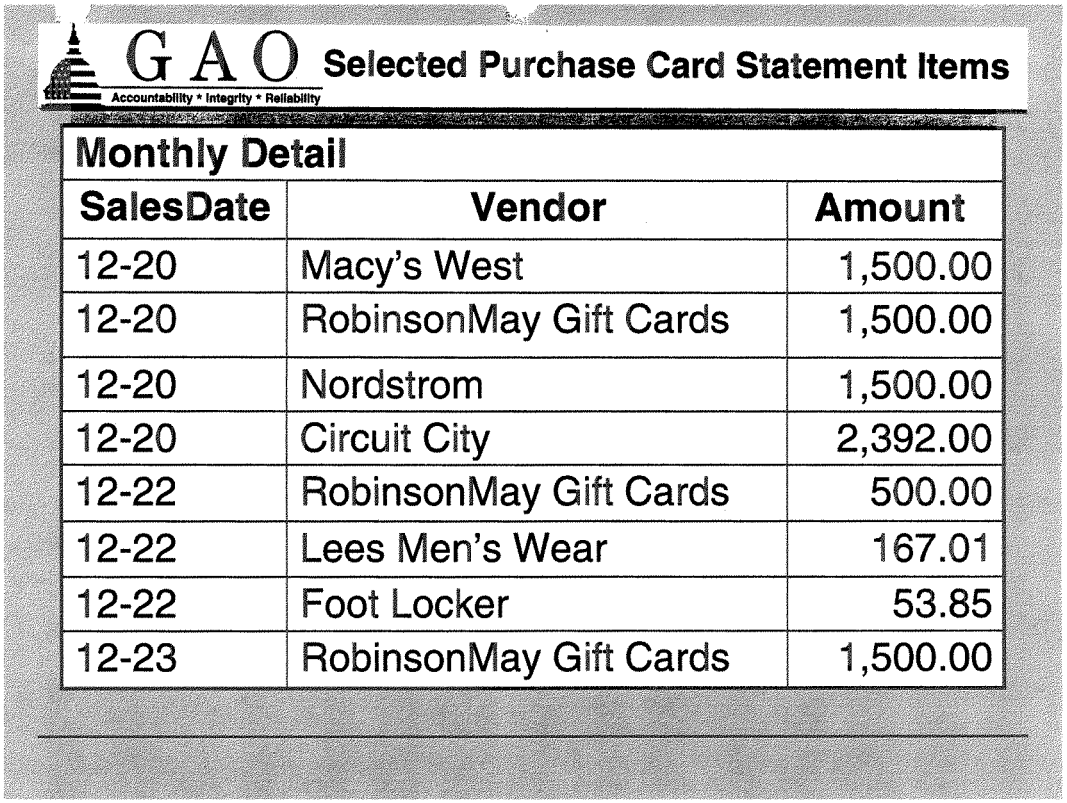
GAO

Accountability * Integrity * Reliability

Abusive Travel Card Activity

Vendor	Description
Budget®	Purchase used car
Spearmint Rhino Adult Cabaret	Adult entertainment
Cryptologic Inc. GCA*Turning Stone	Internet and casino gambling
Unknown	Down payment on home

Source: GAO analysis.



Monthly Detail		
SalesDate	**Vendor**	**Amount**
12-20	Macy's West	1,500.00
12-20	RobinsonMay Gift Cards	1,500.00
12-20	Nordstrom	1,500.00
12-20	Circuit City	2,392.00
12-22	RobinsonMay Gift Cards	500.00
12-22	Lees Men's Wear	167.01
12-22	Foot Locker	53.85
12-23	RobinsonMay Gift Cards	1,500.00

Data Mining and the Federal Government

Testimony of Mark Forman, Associate Director of IT and E-gov, OMB

House Government Reform Subcommittee on Technology, Information Policy, Intergovernmental Relations, and Census

March 25, 2003

Overview of Testimony

- Introduction of Data Mining
- Private Sector Applications of Data Mining
- Public Section Applications of Data Mining
- Benefits and Pitfalls

Private Sector Applications

- Customer Relationship Management/Segmentation Analysis
- Fraud Detection
- Retail Analysis and Supply Chain Analysis
- Medical Analysis/Diagnostics
- Document Analysis (Text Mining)
- Use of Decision Support Systems (DSS)
- Financial Analysis

Public Section Applications

- Financial Management
- Human Resources Management
- Reducing Erroneous Payments/Fraud Detection
- Policy Analysis
- Computer Matching of Databases
- Law Enforcement and Homeland Security
- Citizen Access to Government Data

Benefits and Pitfalls

- Federal data analyzes must be consistent with law
- Ensuring the Security of Federal IT Systems

Mr. PUTNAM. Our final witness is Jeffrey Rosen, a law professor at George Washington Law School. Mr. Rosen's area of expertise is in privacy and technology issues. He has written dozens of articles on the subject as well as a book. His testimony will be valuable as we look to the legal and ethical questions surrounding the use of data mining technology. Welcome.

STATEMENT OF JEFFREY ROSEN, GEORGE WASHINGTON UNIVERSITY LAW SCHOOL, LEGAL AFFAIRS EDITOR OF THE NEW REPUBLIC

Mr. ROSEN. Thank you, Mr. Chairman, and members of the subcommittee. It is an honor to be here. I am delighted that you are holding this hearing because the effort to strike a balance between privacy and security is a bipartisan issue and I am delighted that you are informing yourself about the complicated legal and technological choices that you face as these technologies are implemented.

My thesis this morning is simple: It's possible through law and technology to design data mining systems that strike better rather than worse balances between privacy and security. But there is no guarantee that the executive branch will demand them or the technologist will provide them on their own. You therefore, ladies and gentlemen of the Congress, have a special responsibility to provide legal and technological oversight to ensure that the technologies are developed and deployed in ways that strike a good rather than a bad balance between privacy and security.

Let me give you an example of the kind of design choice that I have in mind. And I want to focus just for the sake of argument on the Total Information Awareness Program that Congress has recently decided, at least for the foreseeable future, to block. Total information awareness provides a model for the kind of mass dataveillance that we have been discussing this morning and is being proposed in other contexts. Now, just a question of definition, "mass dataveillance" refers to the suspicionless surveillance of large groups of people. And that is different from personal dataveillance of the kind that Senator Dockery described which involves targeted surveillance of individuals who have been identified in advance as being unusually suspicious. Mass dataveillance poses special dangers. In some ways it poses some of the same dangers of the general warrants that the framers of the fourth amendment to the Constitution were especially concerned about prohibiting.

When the government engages in mass dataveillance without individualized suspicion, there is a danger of unlimited discretion, as the government searches through masses of personal information and searches suspicious activity without specifying in advance the people, places, or things it expects to find. Both general warrants and mass dataveillance run the risk of allowing fishing expeditions in which the government is trolling for crimes rather than particular criminals, violating the privacy of millions of innocent people in the hope of finding a handful of unknown and unidentified terrorists. At the same time there is an important question of effectiveness.

And I want you to think pragmatically about these technologies. Will they work in the national security arena? Unlike people who commit credit card fraud of the kind that Mr. Kutz described, cred-

it card fraud is a form of systematic, repetitive, and predictable behavior that fits a consistent profile identified by millions of transactions. There is no special reason to believe that terrorists in the future will resemble those in the past. By trying to pick 11 out of 300 million people out of a computer profile, you may be looking for a needle in a haystack, but the shape and the color of the needle keep changing and, as a result, the profiles may produce great numbers of false positives: those people wrongly identified as terrorists.

I want you to think about the privacy issues and the effectiveness issues. Does the technology that works in a credit card arena make sense to apply in the national security arena? Assuming that these technologies will be deployed in different spheres, I urge you to recognize that they can be designed in better or worse ways. The Total Information Awareness Office itself recognized this and proposed technology that it called "selective revelation," which proposed to minimize personally identifiable information while allowing data mining and analysis on a large scale. The insight of selective revelation is useful and may provide models for ways privacy and liberty could be protected at the same time.

The Total Information Awareness Office had a project called Ginisys that was exploring ways of separating identifying information from personal transactions and only allowing the link to be recreated when there is legal authority to do so. This might allow, for example, the Centers for Disease Control to have access to medical information while other groups do not.

Using this model of selective revelation, Congress could think about creating laws and technology that separate identifying information from the data itself.

And Mr. Forman talked about the searches in existence with current law. My strong belief is current law is not adequate, the kind of complicated regulation that faces us, and you need to think creatively about rising to this new challenge by developing new oversight bodies and new technologies to ensure the protection of privacy. But just hypothetically we could imagine what those regulations would look like. Congress could create a special oversight court with the authority to decide when identifying data obtained during mass dataveillance may be connected to transactional information. After intelligence analysts have identified a series of transactions that they think might be evidence of a terrorist plan or suggest that a particular individual is unusually suspicious, they could petition the oversight body for authorization to identify the individuals concerned. In deciding whether or not to grant the request, Congress could direct the court to satisfy itself that the crime for which the evidence has been presented is a serious threat of force or violence rather than a low-level or trivial crime, and that the evidence suggests a link between the suspects and terrorists. If the court granted the order, then the analyst could link the identifying information and they could share the information with State and local bodies and so forth.

And there are other needs for regulation. You might have to create standards for citizen oversights. Citizens should be able to correct their data if it's incorrect or misused. And fair information practices would give citizens the right to know the information that

the government has collected. So, you see the general model. The search is anonymous unless there is cause to believe that a particular individual is suspicious, and then there is oversight to make sure that the individuals are identified in connection with serious crimes. Merely to describe the complexity of this regulation is to raise legitimate questions about whether Congress is ready to adopt them.

But Congress has met its oversight responsibilities in the past. The most important checks on poorly designed technologies of surveillance since September 11th have come from Congress ranging from the decision to block total information awareness in its current form to the insistence on creating oversight mechanisms for the Carnivore e-mail program. I urge Congress to accept the task of learning about the design choices inherent in these technologies. You have it in your power to strike a balance between liberty and security, and all you need now is the will. Thank you very much.

Mr. PUTNAM. Thank you Mr. Rosen.

[The prepared statement of Mr. Rosen follows:]

TESTIMONY BEFORE THE HOUSE COMMITTEE ON GOVERNMENT REFORM

SUBCOMMITTEE ON TECHNOLOGY, INFORMATION POLICY, INTERGOVERNMENTAL RELATIONS, AND THE CENSUS

BY JEFFREY ROSEN

MARCH 25, 2003

My name is Jeffrey Rosen. I am an associate professor at the George Washington University Law School and legal affairs editor of The New Republic. It is an honor to submit to the Subcommittee on Technology, Information Policy, Intergovernmental Relations, and the Census this prepared testimony on "Data Mining: Current Applications and Future Possibilities," which is adapted from my forthcoming book *The Naked Crowd: Liberty and Security in an Age of Terror* (Random House.)

My thesis is simple: it's possible to design data mining technologies in ways that strike better or worse balances between liberty and security. But there is no guarantee that the executive branch or the technologists, left to their own devices, will demand and provide technologies that strike the balance in a reasonable way. Congress, therefore, has a special responsibility to provide technological and legal oversight of data mining, to ensure that the most invasive searches are focused on the most serious crimes.

For an example of the kind of design choices and oversight I have in mind, I want to focus on the most controversial data mining technology, the so called Total Information Awareness Program designed by the Department of Defense. Although the House and Senate have voted to block funding for TIA in the foreseeable future, its architecture remains a model for other data mining programs that are currently being evaluated by the federal government.

The TIA program is an example of what Roger Clarke has called "mass dataveillance" – that is, the suspicionless surveillance of large groups of people – as opposed to "personal dataveillance," which Clarke defines as the targeted surveillance of individuals who have been identified in advance as suspicious or dangerous. By analyzing financial records, educational records, travel records, and medical records, as well as criminal and other governmental records, TIA proposed to develop technologies that could create risk profiles of millions of Americans citizens and visitors, looking for suspicious patterns of behavior.

In its unregulated form, as Congress recognized, mass dataveillance along the TIA model poses great threats to privacy as well as promising dubious benefits in increased security. When the government engages in mass dataveillance to conduct general searches of millions of citizens without cause to believe that a crime has been committed, the searches arguably raise the same

dangers in the twenty-first century as the general warrants that the Framers of the Fourth Amendment feared in the eighteenth century. Dataveillance, like a general warrant, gives the government essentially unlimited discretion to search through masses of personal information in search of suspicious activity, without specifying in advance the people, places, or things it expects to find. Both general warrants and dataveillance allow fishing expeditions in which the government is trolling for crimes rather than particular criminals, violating the privacy of millions of innocent people in the hope of finding a handful of unknown and unidentified terrorists.

At the same time, mass dataveillance along the TIA model may not be effective in identifying terrorists and picking them out of the crowd. Unlike people who commit credit card fraud – a form of systematic, repetitive and predictable behavior that fits a consistent profile identified by millions of transaction – there is no reason to believe that terrorists in the future will resemble those in the past. There were only 11 hijackers on 9/11, and those who followed them during the following year weren't Saudi Arabians who went to flight school in Florida: they included Richard Reeves, the English citizen who hid a bomb in his shoe, and had a Jamaican father and an English mother. By trying to identify people who look like the 9/11 hijackers, the profiling scheme is looking for a needle in a haystack, but the color and the shape of the needle keep changing. And because the sample of known terrorists is so small, the profiles are bound to be produce a prohibitive number of "false positives" – that is, passengers whom the system wrongly identifies as a likely terrorist. A profiling system that was has a 50% accuracy rate in identifying terrorists would mean that one out of every two passengers would be wrongly singled out for special searches.

Nevertheless, it's possible to design mass dataveillance systems in ways that strike a better balance between privacy and security. The Defense Advanced Research Project Agency, or DARPA, which brought us the Total Information Awareness Program, has also been studying technologies of "selective revelation," which minimize personally identifiable information while allowing data mining and analysis on a grand scale. "The idea of selective revelation is that initially we reveal information to the analyst only in sanitized form, that is, in terms of statistics and categories that do not reveal (directly or indirectly anyone's private information," writes a DARPA report called "Security With Privacy." "If the analyst sees reason for concern he or she can follow up by seeking permission to get more precise information. This permission would be granted if the initial information provides sufficient cause to allow the revelation of more information, under appropriate legal and policy guidelines."[1] But without careful oversight of these secret searches – including audit trails that are reviewed by independent agencies – it's easy to imagine opportunities for abuse.

One way to protect innocent citizens is to ensure that general data searches are constructed in ways that make the data traceable but not easily identifiable – in other words generally anonymous unless officials receive permission to link the data with a particular individual. The Total Information Awareness office has a project called Ginisys that is exploring this kind of architecture for general data searches. According to the program director, Ginisys could protect privacy by separating identifying information from the personal transactions, only recreating the association when there is evidence and legal authority to do so. This might allow, for example, the Center for Disease control to have access to medical information while other groups do not.

The Ginisys staff also plans to develop information privacy filters to keep information that is not relevant out of the repository – encouraging the government to adopt laws that limit the types of data that can be recorded about specific people or transactions. Finally, Ginisys plans to use software agents to mine the information in the repository to expunge information that is unrelated to terrorism.[2]

Although TIA in its current incarnation promises questionable security benefits and grave threats to privacy, the selective revelation architecture poses fewer threats to privacy than the unregulated architecture. How, then, can Congress and the other branches of government encourage the development of mass dataveillance technologies that protect privacy rather than threatening it? Congressional oversight provides the most promising path for America in the twenty-first century. Since the 1960s, Congress has passed more than a dozen important laws protecting privacy, ranging from the Privacy Act of 1974, passed in the wake of Watergate, to the Video Privacy Protection Act of 1988, passed in the wake of journalistic snooping into the video rental records of Robert Bork, the rejected Supreme Court nominee.[3]

As Marc Rotenberg of the Electronic Privacy Information Center has argued, one way for Congress to understand the challenge of balancing liberty and security after 9/11 is the model of checks and balances in the U.S. Constitution. That means that if Congress grants the president new authority engage in foreign intelligence surveillance, it should also create new means of public oversight, or if the Department of Homeland Security proposes a trusted traveler program, it should be subject to open government standards. And if it allows mass dataveillance – whether through the TIA program or the risk profiling systems now being proposed for airports, it should insist on congressional oversight.

Congress could create a special oversight court with the authority to decide when identifying data obtained during mass dataveillance may be connected to transactional information. After intelligence analysts have identified a series of transactions that they believe might be evidence of a terrorist plan, they could petition the special court for authorization to identify the individuals concerned. In considering whether to grant the request, Congress could direct the court to satisfy itself that the crime for which evidence has been presented is a serious threat of force or violence, and that the evidence suggests a links between the suspects and organized terrorists. If the court granted the order, the analysts could link the identifying information with the transaction data, and they could contact federal state and local law enforcement officials to inform them of the threat. In addition to creating this oversight body, and determining legal standards to guide its operation, Congress might also have create standards for federal and citizen oversight, along with penalties for abuse; a dispute resolution process that would give citizens recourse when their data is incorrect or misused; and a series of fair information practices that would give citizens the right to know what personal information the government has collected, and to correct any inaccuracies.

Merely to describe the complexity of these regulations is to raise legitimate questions about whether Congress is ready to adopt them. But Congress has met its oversight responsibilities before. The most important checks on poorly designed technologies of surveillance since 9/11 have come from the Congress – ranging from the decision to block TIA in its current form to the insistence on creating oversight mechanisms for the Carnivore e-mail search program. Rather

than accepting the extreme views of luddites, who believe that all surveillance technology should be resisted, or technopositivists, who believe that no surveillance technology should be regulated, I urge Congress to accept the task of learning about the design choices inherent in these technologies. By evaluating their effectiveness, their necessity, and their impact on privacy Congress can ensure that these technologies are designed in ways that strike reasonable, rather than unreasonable, balances between liberty and security. You have it in your power to strike a thoughtful balance; all you need now is the will.

1."Security with Privacy, ISAT 2002 Study, December 13, 2002, p 10, available at epic.org.

2.Doug Dyer, Address at DAPRATech 2002 Conference, August, 2002.

3.For a comprehensive list of congressional privacy legislation and argument for the superiority of legislative to judicial regulation, see Orin S. Kerr, *The Fourth Amendment in New Technologies: Constitutional Myths and the Case for Restraint* (unpublished draft on file with the author.)

Mr. PUTNAM. I certainly believe our witnesses have set the table and created an environment for some outstanding dialog.

The gentlelady from Michigan has another appointment so I will recognize her to lead off with our questions.

Mrs. MILLER. Thank you, Mr. Chairman. I think my question is for Mr. Kutz.

As I heard you talk about some of the various audits that your agency is currently engaged in, you talked about nine different audits that you are getting involved with, Energy labs and DOD, etc., and certainly the testimony you gave about the credit card fraud is startling. It is sickening. Those are the kinds of things I think make people crazy about what is happening at the Federal level. But you know, last week the Congress had a very exhaustive debate about a budget resolution and there was a lot of talk about waste, fraud, and abuse and the kinds of problems in large numbers numerically that we could get at to look at some reduction in our budgeting process.

And I heard a lot of conversation last week—and I don't know if this is one of your nine universes or not—but in the area of Social Security, that there is as much as 10 percent of the Social Security payments that are going to people who are either deceased or for some reason do not qualify. And I don't know if that is an area that you are auditing in your universe there; and, if so, what kind of numbers are we talking about and how would you do a construct to do the data mining? Do you have any idea of how you might begin to proceed to take a look at that type of waste, fraud, and abuse?

Mr. KUTZ. Social Security is not one that we have on our plate right now. We typically do our work at the request of various Members of Congress or committees or subcommittees, and that is not one we have been asked to do at this point.

Some of the ways you can use the technology for that, for example, have been used by the Inspector General to look for people who are receiving benefits that are over 90 or 100 years old, and those are potential indicators of a family that might be keeping the checks and didn't report the death to Social Security and therefore received improper payments.

There are certainly lots of different queries and methods you could use. And I believe the Inspector General has done a lot of that, and I believe it has been used extensively there.

Also for Medicare, there has been extensive use of data mining technologies to find fraud, waste, and abuse and also to project the amount. Annually, the various agencies project how much is going out the door in improper payments and, as you know, there are tens of billions of dollars. And we are talking about real money here, which is why we need good internal control systems to minimize this waste, fraud, and abuse.

Mr. FORMAN. If I may, let me point out two projects in particular. One is 1 of the 24 E-Government Initiatives that is called the E-Vital Project. And so much of this is tied to, for example, the Social Security Administration getting timely notification when a person has passed on. That is explicitly the target of the E-Vital Project that continues to have good traction in the States that have been moving the death records and other medical records on-line. It is

a slow process. And as you may recall, Michigan may have been one of the States. The State has charged the agency to provide that information to them. So there is some negotiation, because the cost should be reduced when we put in place that as a computer system.

The other project is called PARIS, the Public Assistance Reporting Information System, and that is a joint Federal/State information network that was set up explicitly to allow for data matching and mining on interagency-related benefits program. So that would cover things like Supplemental Security Income, the TANF program, Medicaid, Food Stamps, and Veterans Affairs Program.

Mrs. MILLER. In regards to the Social Security link that the States have as they interact with the Federal Government, isn't it true now—because I think every State is required to solicit the Social Security number of every licensed driver—that is something new in the last several years, and all of the States are required to link to the Social Security Administration because of that? Has that been helpful in information sharing?

Mr. FORMAN. You know, to be quite honest, I think ultimately, while there is a requirement to share information, the reality is a big chunk of the benefit here in terms of identifying people who are getting Social Security income but have passed on comes back to the ability of States to share information on the death certificates in a timely manner. And some of the States and local county offices where that information initially starts just haven't been electrified yet.

Mrs. MILLER. My experience had been with the Social Security link that we had in Michigan—I know some of the other States were mentioning this as well—there was no way to verify the Social Security number, so someone could give you any digits that they wanted to. There was no way for the States to verify that the Social Security number was in fact a valid Social Security number. That is a problem, I think.

Mr. FORMAN. There has been some progress made on that, and I know we looked at this a month ago when we did a review. I would ask, if it is OK with the chairman, that we get back to you on the Social Security Administration progress on that.

Mr. PUTNAM. We have been joined by the big Chair, the chairman of the full committee. Mr. Davis, do you have any comments or questions?

Mr. DAVIS. I will be very brief. I think data mining is critical. If you go back 100 years, a visionary at the start of the 20th century might have said, what is going to guide the economy in the 20th century? The visionary might have said, oil. And in fact, it was your entrepreneurs and your visionaries who figured out how you get the oil, identified where the oil was, how you get it out of the ground, how you refine it, how you get it to markets, dominated much of the economic activity of the 20th century.

Here we are at the start of the 21st. What would a visionary say now? Really, the oil today is information. How would we get that information and get it out of the ground, so to speak; how do we refine it; how do we distribute it; what uses does it have? And it is those entrepreneurs that are going to in large part be the economic wunderkinds of the 21st century. Had we had the EPA and

all of the regulations on oil in 1900, this stuff would still be in the ground. We never would not have gotten it out.

My theory is we need to be slow about it coming in and overregulating. You let the marketplace and let the public and let the industry come up with its own protocols before the government comes in and starts imposing a regulatory and taxing regime that could stifle the growth and the potential for this. That is kind of the way I look at it. Certainly there is going to be a role for government down the way, and maybe in ways we don't even envision today, because I think we are just at the very beginning of a whole revolution. But that is kind of the way I have looked at it.

And I don't know if you have any reaction. Mark Forman has been working with us on a number of issues. I don't know if anyone wants to react with that or disagree. Obviously, the professor is here and has his own view.

Mr. ROSEN. I guess I would just urge the chairman to ask whether the kind of data mining that is appropriate in the private sphere can be brought into the national security arena. Much of the history of our privacy laws for the past 50 years has been based on the idea that completely unregulated information sharing is not consistent with the values of the Constitution or of American citizens. We don't want every low-level information officer in the field to know that I had a youthful indiscretion or I am late in my child support payments before I go onto an airplane, or that I am late on my credit card or maybe I have some IRS issues against me.

Complete transparency of information, total unregulated use, which is what many Silicon Valley people are urging, wouldn't be consistent with the value of the fourth amendment. It wouldn't be consistent with current privacy laws which prohibit privacy sharing without good cause, and it also—and I want to urge the chairman to think about it—would it be effective? Is there any reason to believe that centralizing all of our public and private data bases and allowing for a risk prediction to be made would identify terrorists?

It is not like credit card fraud. Credit card fraud is something you have 10 million examples of it and it takes predictable patterns. People who steal credit cards test them at service stations and then buy clothes at a mall. And because it happens so often, you can use the technology to predict credit card fraud.

We have no reason to believe that the next terrorist attack is going to take the place of people who lived in Florida and went to flight schools. It could take many forms. I respect your libertarian instincts and the desire to use this technology as effectively as possible. I just would say that if you, the Congress, doesn't stand up for Constitutional values to ensure inefficiencies as well as centralization, I don't think the technologists of the executive branch will either.

Mr. DAVIS. Most of this information has been public. It has just never been able to get collated and so rapidly deployed and disseminated. That's what scares people. It is something in the old days that could have taken 10 private detectives 6 months going through records to find you can get like that.

And as you spoke of in your testimony, it is a balance issue; and I don't know what that right balance is, but I am on the go-slow side rather than the overregulation side. We know, for example,

that the terrorists on September 11th—the information that was out there between flight schools and arrests and Immigration. Had we been able to collate that information and get it in one place, we could have prevented it from happening.

And some of you view this as an infringement on privacy, but I don't know what you say to the victims and the families of over 3,000 people that died that day. I don't know what the right balance is, and I agree, and that is why we need to hear from you and keep you at the table as we work our way through this brand-new territory. And that is why we appreciate you being here.

And I am not sure we have that right balance today. And I am not sure, given the technologies that we have today, that we can even start writing rules, because who knows what technologies will be deployed and invented tomorrow that we may not be able to have any idea what their application could be? And I appreciate everybody's input and I appreciate you holding this very important hearing.

Mr. PUTNAM. I believe the Senator had a response.

Ms. DOCKERY. Thank you, Mr. Chairman, and I just wanted to comment that I agree very much with the Congressman, Congressman Davis, and to comment to the professor, we in Florida believe that the factual data analysis that we are using now is appropriate for tracking down terrorists, and we also believe that it led to the arrest recently of—a national news story you may have heard about of a professor at University of South Florida. And that was done through collection of information that was all part of our public records in the State of Florida that showed some connections.

So we think that this is a valuable tool and we think we have shown in Florida its criminal possibilities. I will say that in Florida, we have one of the most open record laws in the country. We call it "Government in the Sunshine," and it is kind of interesting that the people in Florida just in the past election voted a Constitutional amendment to require that anytime we provide an exception to the open records law, it would now require a two-thirds vote of both the House and the Senate to make that exemption. The open public records law actually helps law enforcement in Florida by making more and more records available for us to use in our factual data analysis.

So to that extent I wholeheartedly support Congressman Davis's comments and would tell you that we probably need some regulation to prevent us from going overboard and to protect the forth amendment rights, but we should err on the side of allowing the technologies to prove themselves out before we overregulate an industry that is just beginning.

Mr. PUTNAM. For the professor and anyone else who would like to respond, how would you compare data mining technology to the emerging technology of DNA as a law enforcement tool 25 years ago?

Mr. ROSEN. I think DNA offers greater security benefits and fewer privacy threats for this reason. DNA is usually used in the kind of focused investigation of the kind that Senator Dockery was just suggesting: You have a clue and you can plug it into a data base and it can be used to exonerate or inculpate. And as long as there are restrictions on the use of DNA for secondary purposes,

the government can't turn it over to insurance companies to deny me a job or make predictions about my future health, I don't have privacy concerns about it.

Data mining, by contrast, of the kind that Roger Clark calls "mass dataveillance" rather than "personal dataveillance," poses very different privacy issues. And I want to distinguish the two, because Senator Dockery just talked about how useful it is once you know something about an individual. This USF professor, you can plug him into a data base and draw connections. That is the same thing that was done with the sniper. When you have the tip in Alabama and plug it into the data bases and establish connections, that is useful and that doesn't raise grave privacy concerns because the individual has been identified in advance as suspicious.

My concern is the kind of mass dataveillance, not only the total information awareness level, but the profiling systems that are being proposed at airports. And the reason I am concerned about them, this is the surveillance of the data of millions of innocent citizens. And it's just not a little bit of data. If the projects go forward, there are credit card records, phone calls, tax records, all public and private data; mass risk predictions based on this that could be used to prosecute people not for terrorism—which I'm all for—but for very low-level crimes.

It is that kind of fishing expedition—it is the example of an unconstitutional search. At the time of the fourth amendment, what the framers were most concerned about was breaking into everyone's house looking for enemies of the government, reading their private diaries, looking at innocent information, in the course of seeing whether or not they were a critic of the king, and then arresting them for whatever you found in their House. That was a general search and it was unconstitutional because it exposed a lot of innocent information while looking at guilty information. That is what mass dataveillance does. And that's why, without Constitutional restrictions, I don't see how we could deny that there are privacy concerns.

Mr. PUTNAM. A recent New York Times article, a Dr. Gilman Louie, CEO of InQTel, outlined in a recent speech two different approaches, one which he identified as the data mining approach which results in what he calls watch lists and what he indicated was too blunt an instrument; the second being data analysis which begins with some type of investigative lead and then uses software to scan for links between a person under investigation and known terrorists. I presume that is an approach you are advocating?

Mr. ROSEN. I like that approach and I respect Mr. Louie, who is sensitive to these issues, and he is distinguishing between focused data mining based on individualized suspicion and mass dataveillance.

And the same model interestingly has been taken by the Foreign Intelligence Surveillance Court. Just yesterday the Supreme Court decided not to review that decision of the Foreign Intelligence Surveillance Court that said we don't have to worry about broad surveillance of people who have been identified in advance as agents of foreign powers because we suspect that they're bad guys. And if we then find that they're guilty of lower level crimes it's good to get them off the streets because we're pretty sure that they're sus-

picious. That's different, said the Foreign Intelligence Surveillance Court, from using this mass dataveillance to look at everyone without any cause to suspect them and going after them for lower level crimes.

So I'm glad that Mr. Louie, who is at the forefront of the government's effort to merge technologies that have been developed in the private sector and apply them in the national security area, is sensitive to that distinction, too.

Mr. PUTNAM. Let me direct that to our witness, Dr. Louie, who is not the person I was just quoting. You indicated in your testimony that data mining is a process, not a tool. Please elaborate on that in the context of Mr. Rosen's comments.

Dr. LOUIE. Data mining goes—some of the focus that I keep hearing is the emphasis going back to patterns. Data mining deals with patterns, but I think the term "patterns" needs to be expanded a little bit to understand in terms of other ways of interpreting a pattern. A pattern can also be a series of events. A led to B, B led to C, and on down the line. If we are planning a—we'll call it a filtering mechanism to look at everybody, you have to establish some parameters of saying if we are looking for people who buy large quantities of potassium nitrate fertilizer and they are not in agriculture or landscaping and the like, maybe that should raise a flag. But all it does is just put up a flag, says this is of interest. And then if other events or other ties go back to it, then that should, we'll call it, raise a level of suspicion that maybe forwards it to somebody else to review. I think that's the way, we will call it, data mining in general can be applied in terms of looking for potential terrorists, whether it be something like Oklahoma City or something like September 11th.

In terms of September 11th here we have another potentially interesting, we will call it, information exchange of Immigration's data base or when they applied for visas was, we'll call it, a little bit more broader in their perception of how they looked at the information coming in for, let's say, applications of visas. We have, we'll call it, the linguistic issue of how do you spell the name, what are the variations of the name, variations being, let's say, diminutive form of the name or a, we'll call it, a common substitution, Robert for Bob, John for Jack, you know, and down the line. If we had a way to compare that and also previous visas, abbreviations of the names, transposing of the name that would have identified, had these people come through our visa process before, where did they go, did that raise any suspicions.

That's the way I see data mining being applied in terms of broad, we'll call it, filtering of information. Not tracking somebody necessarily, but raising, we'll call it, levels of questionable flags or activities that may lead to something. That way you are not tracking an individual, you're just tracking recent events. If that event tracks out and says all these events lead up to a suspicious activity, then we can go back and say, OK, where did all these names come in or what is the relationship of that. And that's up for the analysts. It's the same way we track money laundering, we track bank accounts. The banks are required to report any transaction of $10,000 or greater. So if I deposit $ 9,999 it's not going to trip the flag. But if, let's say, at the bank level they consolidate the end of

the day receipts and they see that account exceeded that $10,000 maybe it should just raise a flag and make FINCEN aware that there was a transaction, didn't meet the criteria but it's just something maybe to watch. Either the bank watches it or FINCEN watches it.

But that's the way I see you apply data mining. And in terms of—I believe that was Gilman Louie from In-Q-Tel.

Mr. PUTNAM. Yes.

Dr. LOUIE. I agree with his prospect and the way he outlines the way we should look at it. Data mining is an inert tool. You can take very thin slices and basically create a sandwich of a nice depth in order to act upon. And that's where we use the term "actionable information." And one slice of information in itself, it may be totally insignificant and of no value. But it's the cumulative process of all the associations associated with that data point that become interesting. And you don't have to store it. You just have to essentially flag it. And when we have enough flags that trip, we'll call it, your suspicion level, then you look at it. You don't necessarily take an action on it, but evaluate it. And that's where the human aspect or the analysts and subject matter experts in that area can say this does look suspicious or this should be maybe questioned.

Mr. PUTNAM. Mr. Forman.

Mr. FORMAN. I think it's incredibly important to keep in mind that data mining is a productivity tool. Yes, it's part of a process, but at the end of the day our decision has to be is that a process that we want to have that is a more productive process. And that's, I think, one of the big differences to understand about the Total Information Awareness Initiative. That's an R&D project. That is not a Federal IT program. And when it hits the stage where somebody says, geez, we ought to buy something, it falls into the process by which we put out the standards associated with the business case. Are we going to get any productivity out of it?

I have always kept in mind early in my years when I did a lot of data analysis and operations research this notion of garbage in, garbage out that Dr. Louie raised. I am very, very mindful, especially in this area of homeland security, where we have got dozens of data bases, merely hooking them together and applying an algorithm is not going to make the data there any better. Even so, merely allowing those islands of automation to exist and the business process that run off of those islands of automation aren't going to give us any greater homeland security. The core and the issue here is to find out do we have a better way, as we see in Florida, for the investigators to do their work. And are we happy that this is appropriate, given the Privacy Act, given the other laws that cover that. And there is a policy decision to be made there. That now is clearly required to be addressed in the business case process under the E-Government Act, and under OMB guidance we are updating it to comply with that.

Mr. PUTNAM. Anyone else wish to comment on that? With regard to the private sector, is there an industry standard out there that is being used to guard privacy and security of the information in the data mining process? Solely in the private sector. Is there a single industry standard?

Dr. LOUIE. There are no unified business industry guidelines as far as, we'll call it, protecting the privacy of the data. I think that most of our clients have relied on us to devise a, we'll call it, a privacy statement of how we are going to handle data, how we are going to handle the physical storage as well as dissemination of the information and how—who will actually get to see and touch it. That's something that we have devised as being the consultants or the practitioners to different companies. But there are no formal guidelines. We have adapted the, we'll call it, guidelines as specified by the Society of Competitive Intelligence Professionals in terms of saying, OK, this is how we will handle the data. This is how we will ensure our clients' privacy and we will try to abide by that as a form of ethics.

Mr. FORMAN. I would say from the standpoint of what we have seen, there are two standards that have existed over the last couple of years. Opt in and opt out. And I know we have looked an awful lot at those standards to see what would be appropriate for the Federal Government. Opt out being a company tells you you have got this data: If you want to continue with this on-line service or continue as a customer with us, we are going to show the data unless you tell us not to. And opt in is essentially like we see with the little cards at the Giant grocery store chains. If you get this card you get a lot of discounts; in return you give us information about your buying habits. And those discounts give you better products and so forth. And so, how the data is used and how the option is available to the consumer, I think they still have a couple of common standards that have been around for a couple of years.

Mr. PUTNAM. Mr. Rosen.

Mr. ROSEN. But opt in and opt out wouldn't begin to be adequate to the challenge of the regulation you're thinking about now because much of this is data that you can't opt out of sharing. It's data such as credit card purchases that goes automatically to warehouses like TRW or telephone calls that go to the telephone company and that the court has held are not legally protected because of the circular reasoning that you voluntarily turned the information over for one purpose and can't withhold it for another. So I'd gather the kind of regulations that you want to be thinking about are the patchwork of laws that do currently regulate information sharing in the private sector, such as the Fair Credit Reporting Act that would prohibit the kind of personally identifiable financial information that can be shared. As I understand several of the data mining proposals, such as the Total Information Awareness Program, in its original form there was a suggestion that those laws should be relaxed and that the government should have access to data that's currently restricted by law, such as personally identifiable credit card information that can ordinarily be shared and the records of international telephone calls that are regulated by other statutes. So I wouldn't—with respect to the effort of using private sector regulations as a model to guide you in the new world that you face in Federal data mining, I don't think that a simple opt in standard which is based on this voluntariness notion would begin to do the trick. And that's why I think at some point you may down the line have to think about comprehensive reform at the level of

the Privacy Act, which has proved inadequate for regulating the kind of things we are talking about now.

Mr. PUTNAM. Speaking now about the public sector, what level of information sharing is currently allowable by law within and between all government agencies without a special or a specific warrant or request for that information? In other words, how much information sharing is there between HUD, VA, HHS, INS now from a technical potential and from a legal potential.

Mr. FORMAN. There's very little information sharing. This issue came up about a year ago with the concept after program that was called gov.net, and there was a fear for cyber security purposes that we had to protect the sharing of information between agencies, and we found out there was virtually no sharing of information between agencies. There generally, it gets back to this issue that each agency built its own data base, it's own data store, if you want to use the parlance of today's hearing, to support its own mission. And the question is, when can you look across the agencies, when is there a need? Going back several years, two decades almost in the scientific community, there was sharing probably most extensive as it relates to what we now call geospatial information or geographic information systems. There are generally requirements associated to that that we handle via the computer security rules and models and the business case practices. Where we have seen a ramp-up of sharing between agencies has been in the data management area that I've alluded to in my testimony, and that happens to be with these major Welfare programs and it is generally by the PARIS Project. There's been explicit congressional authorization, literally laws authorizing that. We have asked for some additional legal authorities or additional data sharing, a creation of the matching data base that has current job data, but even that is only updated quarterly. We probably could do better than that.

Mr. PUTNAM. So would a successful data mining or factual data analysis project that was attempting to identify a particular profile of a terrorist, for example, would they be able to access any and all Federal Governmental data bases without a specific change in the law? Or would they be able to do that as a result of the law's silence on the topic? First part of the question. The second part of the question is, as a technical matter, could it actually be done?

Dr. LOUIE. On the technical side I say we could do that. We have for several government agencies, but the technical side of making it happen is not really the problem. The problem is the quality and trustworthiness of the information that's in those data bases, is I would say poor to—you know, it is amazing that they can conduct business.

Mr. PUTNAM. Senator Dockery.

Ms. DOCKERY. Thank you, Mr. Chairman. In Florida we require reasonable suspicion to be developed before we use factual data analysis, and then we abide by the standards established in 28 Code of Federal Regulations. To answer your question about sharing intelligence information, Florida deals well with sharing information with other States. In fact, there's a pilot project, the Multistate Antiterrorism Information Exchange, called MATRIX, which is going to consist of 13 States in this pilot project. Our problem has been to share information with the Federal Government,

both in terms of us willingly giving you information and you not being able to receive it and us trying to receive information from the Federal Government.

One case in point, Florida has 16 million residents, but 60 million tourists. We have a lot of people moving through the State and it would be very helpful to us if we could access the visa data base, particularly if we could have access to anyone who may be in Florida who has overstayed their visa and that could lead to a lot of useful information in making these connections. We do not keep dossiers on individuals. We look for linkages based on reasonable suspicion in assorted events and then we look for those linkages. Then just as soon as we see them they're gone. So it is not a matter of starting a file on an individual. It's looking at an activity and trying to find who had some access to something involved within that activity. But it would be very helpful to us and to other States if there was a better cooperation of sharing information.

We have now linked almost everything in Florida together so we can access various agencies' data, but we cannot access anything from the Federal Government nor can they for us because the information that the State has is their possession. But we are willing to share it. We just don't have the technology to do so.

Mr. PUTNAM. Mr. Forman.

Mr. FORMAN. From a legal perspective, I believe there's a pretty broad coverage, let me refer to three laws in particular, the Privacy Act of 1974, the Computer Matching and Privacy Protection Act of 1988 and the E-Government Act of 2002, all of which lay out the principles and the areas that must be addressed, ultimately leading up to what we would look for in the business case of privacy impact assessment. There is a policy decision that will have to be made. There's guidance from both OMB and the National Institute for Standards and Technology on that for Federal information systems to ensure appropriate protections of personal information. I think it's fair to review some of those cases and how that's being done. But the legal framework exists. This does not have to be built from the ground up, per say.

I guess I'm more concerned about this on the technology side. These data bases were largely poorly crafted to start with. The business processes generally are nonexistent and when we try to share information which have different embedded rules in the data bases into a data warehouse and mine that data, I keep in the back of my head garbage in, garbage out, because I think that's the reality that we'll be forever patching together in the Federal arena. I believe that this at the end of the day is not so much a technology issue as we know. The technology exists. It's been used in many governments, including the U.S. Government, for years. The question comes down to can we figure out what's the right business process and who should be in charge or how we want to oversee that, pulling that information together and the person who says I've got a terrorist threat. The best framework for that so far as it links to terrorism is the Department of Homeland Security Act.

Mr. PUTNAM. Mr. Rosen, do you have a comment?

Mr. ROSEN. It's an interesting question whether there are meaningful legal regulations on the sharing of data in the case of individualized suspicion. The Privacy Act has a broad law enforcement

exception and a national security exception, so I'd imagine that when we're talking about personal dataveillance, focused on suspicious individuals, there wouldn't be meaningful legal restrictions on sharing. Mass dataveillance is a different question. And I think that the people who have analyzed this are divided about whether dataveillance along the total information awareness model would violate the Privacy Act. It's not clear whether the information that is being accessed would count as a system of records according to the Privacy Act, and the mere phrase itself shows how outdated that 1970's idea, which presumes that information stored in different file cabinets is for regulating data sharing in the 21st century. So—and then there's also the case that much of this data is already held in the private sector and law enforcement has a long history of piggybacking on the grand data warehouses like TRW, and so forth, in order to get information that it couldn't get on its own.

All this is to say that if you're in any way concerned about restrictions on information sharing, as I hope that you will be to the degree that the PATRIOT Act and the homeland security bill create new provisions for information sharing and the interest of national security, you're going to have to think about this issue afresh and try to craft sensible regulations for these new technologies.

Mr. PUTNAM. Do you presume then that under the current law, particularly the Privacy Act, that authorization of personal information that can be held by the IRS, for example, under the current law would not be eligible to be transferred to Homeland Security or INS or a different agency?

Mr. ROSEN. As I understand it. I'm not an expert on the IRS. The IRS has a series of complicated regulations that have ensured that it especially doesn't lightly share information with law enforcement. So both by practice and regulation, I am not sure that there'd be easy access to that data. But the mere—but you're right to focus on precisely that question and then extrapolate from there to other sensitive information that you might not want to be shared without cause, and then you will get a sense of the degree of the challenge that you face.

Mr. PUTNAM. Well, Chairman Davis pointed out something that in many of these cases data mining is the collation of previously existing, perhaps even public data bases and collections of information and that the amalgamation of that data is what allows you to get a more useful outcome than the time and effort and energy involved in searching each one discretely. The blowup over TIA, characterizing it, I think, has been over this presumption of the next step of data collection between public and private and even into the more personal side of things in terms of habits and patterns based on purchases or travel destinations and things like that. But is there anything—is there any effort currently underway other than what had been a research and development project? Is there any active program in the Federal Government that is doing that type of surveillance or data mining?

Mr. ROSEN. I understand that the CAPPS II program, which is Computer Assisted Passenger Profiling Act—I think I have got the acronym right—is based on very much of a TIA model and is also trying to collate information which is already in the public's sphere

and make risk predictions for particular passengers at airports. So that's why I think the TIA model is one that you will have to think about hard, and I think that the chairman's notion that all this information is already in the private domain and therefore is not of concern and can be analyzed perhaps misses the fact that once the analysis becomes granular there is a difference between having me watched on the street when I walk from door to door by a cop or a neighbor and the government planting a camera on my back that follows me from door to door and records each of my activities throughout the day. That reality, the fact that a level of instrusiveness is inconsistent with the values of a free society is one that our law is not well set up to deal with. The Supreme Court's test for invasion of privacy, as you know, Congressman, says the question there is a subjective expectation of privacy that society is prepared to accept as reasonable and as the invasions become more invasive people's expectations are lowered with a lowering of Constitutional protections. So I would resist the chairman's notion that as long as the information is out there, that any degree of collation and technical analysis is fair game because there is a point at which as you have said when very intimate personal information becomes available to the government on a massive scale that's quite different from some reporter going down to the courthouse and rummaging through a couple of paper records 50 years ago.

Mr. PUTNAM. Mr. Forman.

Mr. FORMAN. Well, in preparation for this hearing, I did a run on our major IT investments of the Federal Government. I did actually two runs, to identify all the data mining and then to identify all the data warehouses because why do a data warehouse if you're not going to mine the data. And zero projects showed up. So I didn't believe that. We don't have anything go on with regards to this. So I used a data mining tool, the search engine on first.gov and got well over 1,000 hits. There's an awful lot of activity going on. Now the question that seems to me comes down to is do we have anything going on as an official IT investment that relates to kind of these random searches. And I'm not aware of any that Dr. Rosen is so concerned about. It doesn't mean that it's not out there. I really need to go back and dig deeper. I just have not found any yet. On the other hand, is there—are there some data mining applications that are similar to that and I think, yeah, you'd have to say that the credit card fraud is very similar. You know the pattern. Same thing on Medicare, Medicaid, mischarging. We know that we should be spending, for example, a certain amount for a certain type of procedure. If we see a company that is routinely overcharging us, we know that it's not an error, it's a systematic overcharging. And so that's a very similar type issue and I think in the areas of government accounts payables, where we know some tolerances and we can use data mining to identify people who are overcharging or fraudulently charging us. You do see that and that has gone through the privacy impact assessment reviews generally.

Mr. PUTNAM. Senator Dockery, hasn't the State of Florida for some time used a data analysis, data sharing, data mining type technology to compare and even correlate employment records with

child support payments to develop a list of folks who are behind in that and whether or not they are cheating the system?

Ms. DOCKERY. Yes, that's one of many areas that Florida has used the technology. Also, in smuggling rings, money laundering, child molestations, so we—after September 11th it was the technology was already there and it was just a matter of adapting it to now apply it to homeland security.

Mr. PUTNAM. So there's a history of civil uses as well as the criminal uses, at least in the State of Florida.

Ms. DOCKERY. Exactly.

Mr. PUTNAM. We have been joined by our ranking member, gentleman from Missouri, Mr. Clay, and I'd ask unanimous consent that he be able to enter his statement into the record. And without objection, show it done, and now recognize him for his statement and questions.

Mr. CLAY. Thank you very much, Mr. Chairman. Let me say, for Mr. Rosen, the Transportation Security Administration plans to use data mining to develop terrorist profiling for anyone who flies. And if Congress goes along with this proposal, what safeguard should be established at the same time to assure public rights similar to those provided in the Privacy Act? Let me also say that—do you believe that airlines are now using profiles when you go to the kiosk to get your boarding pass, and you put your card through the kiosk, don't you think that they examine some of your recent credit activity now and is profiling occurring now by the airlines?

Mr. ROSEN. I do, Congressman. As I understand CAPPS I, or the computer assisted profiling system that's now in use, it does indeed analyze publicly available information from the private and public sector and make risk predictions that can lead people to be taken aside for different searches. As I understand, CAPPS II would only increase this profiling by adding information to the data base. It's difficult to answer your question adequately, because the Transportation Security Administration is not forthcoming about exactly what information it's analyzing and how it's using it, and I think a crucial part of your oversight role should be to ensure that the data in the data base is transparent, not the algorithms. The transportation authority says, well, we can't tell you what algorithms we're using or the terrorists can beat the system. What Congress needs to know is not what the algorithms are, but is this data that the Federal Government is entitled to analyze.

So when you think about how to regulate this new system, and this will be a pressing concern, even more so than total information awareness because that's been tabled for the moment, think about transparency, accountability. Citizens should be able to correct errors in their data base. We have heard a lot this morning about the poor quality of the data. Imagine being stopped repeatedly on the basis of inaccurate information and having no remedy, not even being told why you've been stopped. The application of fair information practices to the transportation arena is something that Congress urgently needs to think about because the Privacy Act in its incarnation is not adequate to the task.

So I think that this should be a good model for you as you think about regulation.

Mr. CLAY. Thank you very much.

Mr. Forman, along those same lines, airline security has had a troubled history of racial profiling, even before the attack on the World Trade Towers. During the 1991 Gulf war individuals with Middle Eastern names were forced off their flights despite the fact they were American citizens. Last year the ACLU testified before Congress of dozens of such incidents, individuals discriminated against in airports or on airplanes based on race and heritage. The same people who oversaw the private contractors who provided discriminatory security are now designing new systems. What is OMB doing to prevent racial profiling from continuing in air transportation?

Mr. FORMAN. Well, let me put this into the context of the CAPPS II program. The CAPPS II program was not approved by OMB to proceed at the pace that they seem to want to proceed. I have a huge spotlight on that project right now. They're late in getting back to me the information that they need to proceed. So the issues that we're talking about, the issues that concern me essentially, CAPPS II could quickly become the 80th watchlist. And I have to take a step back in my job and say, what value added do we get by yet another island of automation coming up with something farther away from something that's going to give us the productivity and effectiveness we're looking for. You know, the argument that I have heard in favor of CAPPS and CAPPS II essentially went back to the question of do you want this random? Because my father, my grandmother was pulled out of line. And it just didn't seem to make sense. So there has to be something better. And I think, and I allude to this in my testimony in the customs arena, in the package movement, we seem to figure out this risk paradigm. Now, I think that's what we are looking for. We're clearly not looking for a racial profiling. We are looking for a risk profiling. And there the data that I'm asking for, it's got to be in the business case, would give us both the technical programmatic reviews as well as the policy review. We don't have it yet.

Mr. CLAY. In this process you're looking for random, random profiling and not racial profiling or heritage?

Mr. FORMAN. We are looking for risk based—.

Mr. CLAY. Risk based.

Mr. FORMAN. Reduction. So not random profiling.

Mr. CLAY. So the 9-year-old little girl that goes through, you may not want to search her, through TSA. You may not want to search her?

Mr. FORMAN. As a random selection, that would be correct.

Mr. CLAY. Or the 85-year-old grandmother?

Mr. FORMAN. As a random selection, that would be correct. We are looking for clear documentation that they have actually figured out an approach that's going to improve the productivity. You know, we can spend hundreds of millions of dollars on a terrific IT system with very pretty screens or very fruitful data mining techniques. But at the end of the day, if it somehow does not lower the risk, to me, I would have to say that is not a good IT investment for the Federal Government and would recommend against that.

Mr. CLAY. OK. All right. Thank you.

Mr. Kutz, does data mining need individual identities in order to detect patterns of unusual activity? And can the government de-

velop profiles of unusual activity and then followup on the specifics with appropriate oversight?

Mr. KUTZ. Again, what—most of what we have done so far relates to credit card data bases, but we have gone beyond that certainly for the credit card data bases and these were government credit cards, ones issued by the—on behalf of the Federal Government to use for government purposes. We did have that information to basically analyze and put together patterns of activity, etc. But we have also gone beyond, I was going to mention an example last year. We testified before Representative Shays on the JS List suit, which is the current chem-bio suits that are being used in the Middle East. And what we identified there was that they were excessing and selling those goods on the Internet at the same time they were buying them. And so in that instance, we tried to identify who was buying these suits and whether or not they might be using them for something that would be against the government. So we try to identify, where it is appropriate, individual identities to followup for investigative purposes.

Mr. CLAY. Let me ask you a followup on the question I asked Mr. Rosen. What exactly do the airlines look for when we go to the kiosk and put our credit card through? What kind of financial activity are they looking at? Just out of curiosity.

Mr. KUTZ. I couldn't answer that question.

Mr. CLAY. You don't know. Does anyone on the panel know what they're looking at? I mean, is it one purchasing one-way tickets or what exactly.

Mr. ROSEN. We know from criminal procedure cases that there's certainly public information that they look for, one-way tickets, certain points of origin passengers and the addresses and phone numbers that you check in with and the people that you also are traveling with, and information neuro network analysis can be done on that. But we are assuming that they're respecting legal limitations on, for example, looking at personally identifiable phone calls or personally identifiable credit card information. But finding out the precise answer to that, I know there are groups like some of the privacy groups in town have Freedom of Information Act requests to find out exactly what information is being used and they haven't found the TSA terribly forthcoming, as I understand it.

Mr. CLAY. Do you think they also look at recent purchases in retail outlets?

Mr. ROSEN. As I understand it, they would be restricted from doing that by the Federal Credit Reporting Act, but you need a closer parsing of the statute than I can give you for that.

Mr. CLAY. OK. Thank you very much.

[The prepared statement of Hon. Wm. Lacy Clay follows:]

STATEMENT OF THE HONORABLE WM. LACY CLAY
AT THE HEARING ON
DATA MINING

MARCH 25, 2003

Thank you Mr. Chairman. I would like to join you in welcoming the witnesses to today's hearing, and thank them for taking the time to share with us their knowledge on this subject. I am sorry that former Majority Leader Armey cannot be with us today. His defense of individual privacy during his career in the House is admirable. I am sure he would have added an important voice to this discussion.

I was pleased to read Mr. Rosen's testimony because it reflects by basic reaction to the issue -- data mining can be used well or badly, but it is all in how it is used. The more openness and oversight to the process, the less likely serious violation of citizen rights.

Let's be clear from the beginning. Data mining is profiling using computers and statistical models. We have all seen TV shows where the police have contacted the psychologists at Quantico and gotten a profile of the criminal that leads to his capture. We are also aware of individuals who have been wrongly arrested because they fit some profile. Indeed, innocent people arrested on a profile, have been wrongly convicted. One of the problems

with profiles is that they too often create the presumption of guilt. We have that same problem with data mining.

When credit card companies use data mining to track our purchases and then try to quickly stop fraudulent use of our cards, few people object. However, the government must be much more careful in using these techniques. First, much of what has been proposed for government use of data mining violates the basic principles of the Privacy Act. Second, when government uses these techniques, we have to be much more concerned with the cost of being wrong. If the credit card company is wrong, it often means nothing more than answering a phone call. When the government is wrong, the consequences are far greater.

Let me give you a simple example. One of the companies that produce face recognition software claims that they have an accuracy of 99.32%. Let's stop for a moment and think about what that means. About 20 million passengers pass through Dulles Airport each year. If we used this face recognition software to identify suspected terrorists, and no terrorists passed through Dulles at all, then 165,000 people would be stopped. Those people would be stopped and treated as terrorists, and the officials would be saying to themselves -- this guy has to be a bad guy. After all, this system is accurate more than 99% of the time. How would you feel if you were stopped as a terrorist, denied your rights, and subjected to the kind of interrogation we reserve for this kind of criminal?

The Transportation Security Administration (TSA) wants to create a system that uses data mining to give a terrorist score to every person who buys an airplane ticket. Those with high scores would be searched carefully. Those with low scores would go through the system with minimal screening.

To make matters worse, the TSA wants to keep the information on its data mining a secret. You won't know what information was used to create your terrorist score, and you will have no right to examine that information and correct errors. If you get a high score because of some mistake in the data or the computer program, you are stuck with it. If that makes traveling more difficult for you, you are out of luck.

Mr. Rosen proposes an oversight system for these kinds of security systems. I look forward to discussing that proposal. However, I would like to close with a thought from the world of cryptography -- the science of securing messages. In the 19th century, the cryptographer Auguste Kerckhoffs set down a principle that guide the most advanced work in cryptography today -- in good systems, the system should not depend on secrecy, and it should be able to fall into enemy's hands without disadvantage. In other words, the system should keep messages secret even if the enemy knows how the system works. That is the basic principle that underlies today's public key infrastructure. Unfortunately, that is not the principle that guides the systems being set up by agencies like TSA.

Mr. PUTNAM. The gentleman raises an interesting point. Immediately after September 11th I was pulled every single time I flew because I was not in a frequent flier program, we bought our tickets at the last minute because of the Congressional schedule and it was always one-way. And so I got the body cavity search just about every time I flew. And it's terribly frustrating and it begs some better type of profiling, particularly based on risk. And while some Members of Congress can be shady characters at times, hopefully we wouldn't fit the risk profile.

Mr. CLAY. Hopefully we wouldn't get stopped as often.

Mr. PUTNAM. Well, hopefully, at least not quite as often. Every time got a little old.

But let's get back to the people component of this, because I think everyone has agreed that at the end of the day, no matter what type of process there is and no matter what type of information or data is out there, at the end of the day it is going to require some analysis by a human being. And everyone in general has seemed to stress the need for quality data as well as those high quality analytical skills in the personnel.

Can you expand on that a little bit and talk about where we are in terms of our human capital and the role that they play in obtaining acceptable results through this process?

Mr. FORMAN. I think there are some very, very good examples of the training and culture change that has to take place here. When you move from a paper based—technically we call knowledge management environment—to an on-line you're going to use different interfaces. To do—to have that tool kit, if you will, generally, people have to become computer literate and willing to use computers. And that's where we see, especially in the law enforcement arena, a cultural, maybe generational change that we are working through. Certainly you'll see that at the FBI if you look at their use of the TRILOGY program and the culture of change that the Director is bringing. From my perspective, in the business case itself I look at that. I look to see are we investing in training and process reengineering, change management projects. And when I see generally data mining or tools that use these knowledge management systems and support systems tools without any training, that is a flag to us that this should go on the high risk list. Unfortunately, that has been the pattern of government. Somebody in the technology side invests in these tools and then they get ready to deploy and they find out culturally or from an education standpoint people don't want to use them. And as in the case of the INS, then we go on a binge of buying training services. So I'd say right now, training or the education part has been an afterthought and it's one that needs a lot more attention and funding from the upfront. We are trying to put that discipline in the process.

Mr. KUTZ. Mr. Chairman, I would add to that the software that we had to do the data mining that we have done in the fraud, waste and abuse type applications which is fantastic. It's flexible. We certainly train our people, etc. But the real element that makes it work is the people and the continuous learning that goes on with even using that software and the various programs. So we've kind of got a process where as we look at a system and a program, we understand the program, understand the controls, understand the

vulnerabilities, and we use that too as a feedback into the actual data mining strategy, combining auditors and investigators again.

I mentioned Mr. Ryan, who's with me today, who worked for the Secret Service doing money laundering and credit card crimes for decades. People with that kind of experience teaching younger people some of the things that they know really provides a great atmosphere for learning and developing all those human capital skills.

Mr. PUTNAM. Have you an estimate of the savings that have been derived from that type of data sharing initiative?

Mr. KUTZ. From the data mining with respect to the fraud, waste and abuse?

Mr. PUTNAM. From the financial management side, yes.

Mr. KUTZ. If you go back to the improper payments reporting that's gone on in Federal Government for years, I think that areas like Medicare have shown large decreases in estimated improper payments, and that's I think in part due to the data mining that's gone on there. Another program that's had a great deal of oversight in that area is the earned income tax credit, which had estimates of as much as $8 billion of improper or fraudulent type payments over the years. So there's certainly been savings. I don't think it's been quantified necessarily, but the focus of data mining and the focus on improper payments going out the door has led to better controls in the government and probably saved billions of dollars.

Mr. PUTNAM. Senator Dockery.

Ms. DOCKERY. Thank you, Mr. Chairman. You bring up a good point and one that piggybacks on to Congressman Davis. The information that we are using in tracking criminal activity and potential terrorist events takes into consideration what used to be information in various locations. By putting that all together, it cuts the time down from weeks or months to a matter of minutes. Once that information has identified a risk, that's when the investigations begin. So it still comes down to our human investigators, but instead of spending all their time digging through paper to find out where to start, they now have a starting point and spend their time more wisely looking at those individuals who have come up as a potential risk. So it does involve a lot of training. We do—the success of what we do with that information lies within our law enforcement, but this allows them to spend their time in the investigation and not in trying to put together a pattern.

Mr. PUTNAM. How reliable is that data? How often is it maintained? How often is it upgraded? And we have certainly learned in our experience with the election that sometimes our data bases are a little old with respect to eligible voters and convicted felons and things like that. How good a job does the State do in maintaining that data base that they depend on?

Ms. DOCKERY. Well, I am not an expert in that area, but I would say that we do have systems put in place to purge information. We have systems put into place to check information. And the sharing of the information allows us to hear from other sources in the law enforcement community that some information may be suspect. So I think our information is good. Keep in mind that when it lists people with risk factors, that doesn't point to that person as being

guilty of anything. It points to that person as coming up as maybe a place to start the investigation.

Mr. PUTNAM. Mr. Forman, you had referred to geospatial information earlier in your testimony. In my understanding that is 1 of the 24 E-Government initiatives, and that would involve an overlay of information from a variety of sources with regard to identifying the geography of data. In essence, you overlay the census data with USGS data and we can look at, you know, where the population threats are to sensitive estuaries or any of a million combinations of things by combining all the data that's collected and stacking it in a meaningful way to derive answers about what's going on. Isn't that data mining?

Mr. FORMAN. Yeah. That very definitely will have to require data mining. There are two approaches to leveraging the redundant data sources. One is the concept of buy once and use many. We are definitely proceeding with that. But then where do you put that data? Is it some is maintained at National Weather Service, for example, or NOAA and some is maintained at the U.S. Geological Survey, some is maintained at Environmental Protection Agency? That kind of pier to pier computing model is the emerging concept of a virtual data warehouse in which case probably at that program office you would have the meditative description of where do I go to find this data, what is the standard, and access that. Regardless of whether it is a physical data warehouse or this virtual data warehouse to get access to that data, to make sense of it, data mining techniques will be used. They have been used, you know, for example, probably the best example today, if you go to the Census Web site, American Fact Finder, you can find out supposedly, I haven't done this, but the theory was you could find out how many kids of soccer age for second grade soccer teams, second and third grade soccer teams are in your track, you know, in your soccer league area. That wouldn't tell you by house, but that would tell you maybe by block or by subdivision.

Mr. PUTNAM. The opportunities for the beneficial use strike me as endless. When you compare weather patterns with farm payments, with crop insurance, perils and things like that, then maybe we start raising the risk premiums for that area or maybe we adjust our farm payments so we don't let people plant in that area until El Nino clears up. I mean the opportunities are endless to derive information. The Federal Government spends a fortune collecting information and the fact that it is for the large part underutilized is distressing from a taxpayer perspective.

Mr. Rosen, you mentioned earlier that perhaps we should consider the creation of a special court to consider these types of requests for specific searches, I believe.

Mr. ROSEN. I did. And, Congressman, I would distinguish the need for a special court when we are talking about the mass dataveillance of personally identifiable data with the kind of syndromic surveillance that you and Mr. Forman have just been talking about. This is indeed a wonderful resource, and there are no privacy issues when you're making general statements about weather patterns or census information that's not personally identifiable or the Centers for Disease Control using data mining to figure out when people are checking in in one area with an epidemic

or, to give another example that I am very impressed by, the city of Chicago using data mining to figure out when crime patterns correspond with particular weather patterns and sports events and then they can deploy the cops to that area of town when there is a particular game on and that's really hot and then they can stop crime. These are wonderful things that don't raise any privacy issues at all. That's very different though from, and again if the jargon isn't helpful let's come up with another term, but mass dataveillance, suspicionless searches at airports, the total information awareness model, this is something that needs regulations.

So my message has been this stuff isn't all good or all bad and the technology isn't evil, just be especially attuned to the privacy dangers of suspicionless searches that allow personal information to be collected in ways that are not currently available. And for that I think you do need—it doesn't have to be a special court. You could have a magistrate. You could have a congressional oversight body. There are all sorts of ways to do it. But you have to separate the model as the data is traceable but not identifiable. You can do those sort of general predictions and risk profiles that Mr. Forman is talking about, but you can't actually identify me as the person who's been buying fertilizer unless it really looks like I'm a terrorist because I've done some other things that are suspicious, too.

Mr. PUTNAM. Well, I would remind you and the rest of the panel and the audience that on May 6th we will convene our next oversight hearing on this topic, specifically to address TIA CAPPS II and some other similar programs.

With that, I will yield back to the gentleman from Missouri for any questions.

Mr. CLAY. Thank you, Mr. Chairman. Senator Dockery, I'd be interested to know what Florida does to protect individual rights. Does an individual have a right to know what information about them is included in the data analyzed in the factual data analysis? Does the individual have a right to correct the information in those data bases that is wrong? And what happens if an individual is singled out because of incorrect information in one of these data bases? Can you kind of expound on that for me?

Ms. DOCKERY. Yes. Thank you. All the information that is in the data bases are part of Florida's open public records. So any individual is at any time able to check out those records and to clarify any misinformation on those records. We don't keep particular files on any individuals. We look for events, and risk factors may make somebody come up. Then it goes to a human being, an investigator to investigate that and they may find that just because the individual was identified as being—fitting those risk profile that person was nowhere near the event. So there are a lot of safeguards built in. And of course, we abide by the Federal Code that I mentioned earlier.

Mr. CLAY. So the safeguards are there and they're helpful and people can followup and correct them?

Ms. DOCKERY. Yes.

Mr. CLAY. That sounds like a pretty foolproof system. Thank you.

Mr. Kutz, what would you recommend Congress do to stop the racial profiling that is going on in today's airline security? Do you have any recommendations?

Mr. KUTZ. No, that's not an area that I deal with so I can't comment on that.

Mr. CLAY. OK. Well, let me also ask you, you recently did some work for Congress where you identified several people getting treatment at veterans hospitals who were listed as deceased on Social Security records. With further investigation, you showed that the problem was errors in the Social Security records. Now, if TSA had those Social Security records in their data base, those people would be stopped from flying and they would have no way of knowing why or correcting the incorrect information. Would you agree that any system used by TSA has to allow for the public to know what information is being used to rate them and what other safeguards should be in place?

Mr. KUTZ. Your question gets back to the issue I think Mr. Forman talked about, about data quality in the Federal Government, and we did indeed find, and this was from military treatment facilities, we had compared people who were served at some military treatment facilities with a Social Security death file and there were some hits that came out of people that appeared to be dead that were not really dead. And so there were errors in the Social Security death file, and that certainly raises issues about what that file is used for. That file is certainly shared with others. It's sold to others. And the Social Security Inspector General has reported other examples of errors with that.

So this issue of Federal Government data base reliability is a major challenge here in all applications of data mining going forward. And I had some experiences I was going to share with you on the IRS, where I used to be responsible for the IRS financial audit, and we found lots of instances there with the errors in the system there were people who were being pursued and having taxes collected from them but didn't owe any taxes. At the same time we were issuing lots of refunds to people who weren't due refunds.

So, again you've got lots of issues with data quality and I would say that the Federal Government is decades behind the private sector in that area. I got to go to Bentonville, AR within the last year to visit the Wal-Mart headquarters and it was quite fascinating to see the technology that they use in their inventory supply chain management, and when I compare that to where the Federal Government is with its inventory management again it's just decades behind. And they were able to tell us at Wal-Mart headquarters how many tubes of toothpaste there were at the Fairfax Wal-Mart here in 1 minute. And not only that, but how many they had actually stocked in the last week, how many had been bought in the last week, just tremendous technology, whereas again in the Federal Government I'll go back to the JS List, the chem-bio suits used by our troops. Once those left the defense warehouses into the military services, complete visibility was lost and we were unable to determine where these chem-bio suits were, some from prior years that had been defective through a fraud scheme by a private sector company.

Mr. CLAY. You do make recommendations to the different agencies how to correct the errors that you all find?

Mr. KUTZ. Right. That's the value of data mining. It helps us to make valuable recommendations to Federal agencies to improve their control systems, etc., to try to minimize the risk of these things happening that I've just described.

Mr. CLAY. What was your recommendation to the Social Security Administration?

Mr. KUTZ. We didn't make any recommendations to them because the Inspector General had already made recommendations to them, and they are working to clean up that data base.

Mr. CLAY. I see. Thank you very much.

Mr. Forman, would you support legislation that prohibited the TSA from using any system that used profiles based on race, religion, national origin, gender, sexual orientation or proxies for those characteristics?

Mr. FORMAN. I forever remember my time on the Hill and a good staffer on detail from GAO who has been a staffer to this committee before, the devil's in the details. I'd have to see the specifics.

Mr. CLAY. See the specifics. OK. Thank you very much. And thank you, Mr. Chairman.

Mr. PUTNAM. Thank you, Mr. Clay. And Mr. Kutz, when Mr. Forman gets done with the Federal Government, Bentonville, AR is going to be sending executives up here to tour the Federal Government to see how efficient we are. Isn't that right?

Mr. FORMAN. Absolutely.

Mr. PUTNAM. I want to thank the witnesses for their outstanding testimony and for the questions of the subcommittee. We will be focusing very, very directly on this topic throughout the 108th Congress. Our next hearing on the topic is May 6th to look at some of the specific issues that have been raised. But this is very clearly on my radar screen and something that we will continue to monitor very closely. It is an important issue. It holds the promise of tremendous potential benefits to our taxpayers in eliminating waste, fraud and abuse and bringing better financial management practice, and frankly it raises some red flags in terms of protecting those very same taxpayers' privacy and personal information. So we will do what we can to determine where that fine line is and attempt to walk it.

So I understand Mr. Rosen has to be out to teach his class, but do any of you have one last question that you wish we had asked you that you want to answer?

Senator Dockery.

Ms. DOCKERY. It's not a question. But, Mr. Chairman, if I could just take this minute since I don't have the opportunity to speak to a congressional committee every day, I want to thank you on behalf of the States for what you do in Congress, to send money down to the States to allow us to do the job of protecting the residents in our State against any threat to our homeland security, and I would ask that in the future when moneys are coming down from the Federal Government, the more flexibility you could give us in spending those moneys and if you could have those moneys go through the State rather than directly to the local governments so that we can have a better feel for what's coming down and avoid duplication of effort. But thank you for all that you do for us and thank you for letting me participate today.

Mr. PUTNAM. Thank you, Senator.

Dr. Louie.

Dr. LOUIE. Yeah. This is on-line data collection. The point about individual data elements are not necessarily very important in themselves, but you should also look at how this data is used as if it were classified material. Individual elements in themselves are not necessarily important. It's the combination of multiple elements that make it an interesting issue as far as questionable invasion of privacy or whether it raises flags about how that data is being used in the case of are we really profiling or are we looking at a risk assessment. Should we look at race and national origin? Probably yes. In themselves they are not necessarily the most important item, but in combination with other data elements they may raise a level of risk, and it needs to be considered in that manner. It needs to be viewed not as an individual component, but the sum of all the components looked at in terms of evaluating whether this information is something that warrants looking into or not looking into.

So does it make it actionable? That's the way you need to look at the collection of data, not the individual elements necessarily.

Thank you for the opportunity.

Mr. PUTNAM. My pleasure. Thank you. Anyone else?

Mr. KUTZ. Yeah, I would just say I appreciate you inviting us to the hearing today. Since we work for Congress, we certainly believe data mining is a tool that's going to be able to help us better serve you and to do better audits and investigations on your behalf. So I appreciate that.

Mr. PUTNAM. Thank you. Mr. Rosen. Mr. Forman. We appreciate your efforts. I'm reminded that in the event there are additional questions the record will remain open for 2 weeks for submitted answers. And with that, the meeting is adjourned.

[Whereupon, at 11:30 a.m., the subcommittee was adjourned.]

[Additional information submitted for the hearing record follows:]

ELECTRONIC PRIVACY INFORMATION CENTER

epic.org

1718 Connecticut Ave NW
Suite 200
Washington DC 20009
USA
+1 202 483 1140 [tel]
+1 202 483 1248 [fax]
www.epic.org

March 25, 2003

Representative Adam Putnam
Chair, House Government Reform Subcommittee
on Technology, Information Policy, Intergovernmental
Relations, and the Census
B-349-A RHOB
Washington, DC 20515

Representative William Clay
Ranking Member, House Government Reform Subcommittee
on Technology, Information Policy, Intergovernmental
Relations, and the Census
B-349-A RHOB
Washington, DC 20515

Re: Hearing on Data Mining: Current Applications and Future Possibilities

Dear Chairman Putnam and Ranking Member Clay,

The Electronic Privacy Information Center (EPIC) submits this letter for inclusion in the hearing record for the March 25, 2003 Oversight Hearing on Data Mining. EPIC is a not-for-profit research center based in Washington, D.C. It was established in 1994 to focus public attention on emerging civil liberties issues and to protect privacy, the First Amendment, and constitutional values. We appreciate the Committee's attention to data mining and its civil liberties implications.

We write to call your attention to the growing practice of federal agencies purchasing commercial databases for law enforcement purposes. It is our view that these activities violate the intent of the Privacy Act and should be suspended.

EPIC initiated a Freedom of Information Act (FOIA) request to seven federal law enforcement agencies in July 2001 to obtain agency records relating to government purchase of personal data from commercial information brokers. The documents obtained from the request and subsequent litigation show that a number of information brokerage companies provide law enforcement agencies with information ranging from Social Security Numbers to professional licenses.

As Congress considers the impact of data mining on privacy and civil liberties, it should focus attention on the risks created by relationships between federal agencies and their private-sector information broker partners. Some of these private-sector information brokers sell detailed consumer purchasing data that exists in no public records system. The information sold by these commercial data brokers could be used for data mining.

We recommend that Congress act now to limit private-sector collection of information, because information collected by private entities is regularly sold to federal law enforcement agencies. This practice contravenes the clear intent of the Privacy Act of 1974.

Data Mining

Data mining is "the process of finding patterns in information contained in large databases."[1] Data mining is employed in different contexts in order to achieve different goals. For instance, data mining is commonly used to detect fraudulent use of credit cards. It has also been employed by companies to detect defective parts in a manufacturing line.

When employed for the limited purposes of fraud detection or product quality, data mining poses little risk to privacy and civil liberties. However, when these systems are employed to evaluate future intent or action, data mining presents serious risks to a distinctly American value: "the right to be let alone." For instance, the Transportation Security Administration is currently developing a data mining system called CAPPS II, the Enhanced Computer Assisted Passenger Profiling System. CAPPS II would sift through credit report header information and over one hundred unnamed commercial and government databases to attempt to assess a passenger's risk to a transportation system.

Retired Admiral John Poindexter leads a research project at the Defense Advanced Research Project Agency that is developing a data mining system similar to CAPPS II. The system, Total Information Awareness, purports to capture the "information signature" of people so that the government can track suspicious persons. The project calls for the development of "revolutionary technology for ultra-large all-source information repositories," which would contain information from multiple sources to create a "virtual, centralized, grand database." This database would be populated by transaction data contained in current databases such as financial records, medical records, communication records, and travel records as well as new sources of information. Also fed into the database would be intelligence data.

These two systems are highly invasive because they are operated by federal agencies, are conducted in secret, draw upon a wide array of data sources, and attempt to predict future human action. These two systems have been made possible by private sector information sources, which have voraciously collected information on individuals.

The Privacy Act

In 1974, the Congress, with broad bipartisan support, enacted comprehensive legislation to prevent precisely the type of data profiling that is now under consideration by several federal agencies.

---

[1] Usama Fayyad, Data Mining, *in* ENCYCLOPEDIA OF COMPUTER SCIENCE (A. Ralston, E. Reilly, & D. Hemmendinger eds., 4th ed. 2000).

In passing the Privacy Act, Congress found that:

> (1) the privacy of an individual is directly affected by the collection, maintenance, use, and dissemination of personal information by Federal agencies;
>
> (2) the increasing use of computers and sophisticated information technology, while essential to the efficient operations of the Government, has greatly magnified the harm to individual privacy that can occur from any collection, maintenance, use, or dissemination of personal information;
>
> (3) the opportunities for an individual to secure employment, insurance, and credit, and his right to due process, and other legal protections are endangered by the misuse of certain information systems;
>
> (4) the right to privacy is a personal and fundamental right protected by the Constitution of the United States; and
>
> (5) in order to protect the privacy of individuals identified in information systems maintained by Federal agencies, it is necessary and proper for the Congress to regulate the collection, maintenance, use, and dissemination of information by such agencies.

The Privacy Act set out several specific purposes. These are to:

> (1) permit an individual to determine what records pertaining to him are collected, maintained, used, or disseminated by such agencies;
>
> (2) permit an individual to prevent records pertaining to him obtained by such agencies for a particular purpose from being used or made available for another purpose without his consent;
>
> (3) permit an individual to gain access to information pertaining to him in Federal agency records, to have a copy made of all or any portion thereof, and to correct or amend such records;
>
> (4) collect, maintain, use, or disseminate any record of identifiable personal information in a manner that assures that such action is for a necessary and lawful purpose, that the information is current and accurate for its intended use, and that adequate safeguards are provided to prevent misuse of such information;
>
> (5) permit exemptions from the requirements with respect to records provided in this Act only in those cases where there is an important public policy need for such exemption as has been determined by specific statutory authority; and
>
> (6) be subject to civil suit for any damages which occur as a result of willful or intentional action which violates any individual's rights under this Act.

The Privacy Protection Study Commission created by the Privacy Act recommended that these protections be extended to private-sector collection of information.[2] However, Congress did not act to extend protections to private-sector information collectors.

Now that private sector entities are engaging in practices that enable federal agencies to violate the purposes of the federal Privacy Act, we believe that Congress should regulate these businesses.

Private-Public Sector Partnerships Create New Data Mining Risks

EPIC initiated a FOIA request to seven federal law enforcement agencies in July 2001. Documents obtained from the request and subsequent litigation show that a number of companies provide law enforcement with personal information. There is a risk that this information could be used for wide-scale data mining.

The documents obtained by EPIC under the FOIA demonstrate that commercial database vendors sell volumes of personal information to federal investigative agencies. These companies possess multi-million dollar contracts with federal agencies to provide desktop computer access to personal information. If these databases of information were used for data mining, it would represent a serious threat to First and Fourth Amendment Constitutional values.

The documents obtained by EPIC show that a number of companies are selling personal data to the government:[3]

1. The Department of Justice obtained a $11,000,000 contract for access to ChoicePoint databases in fiscal year 2002.[4] ChoicePoint is a large provider of credit header and public records information.[5] A credit header lists the name, address, previous address, place of employment, spouse's name, and the Social Security Number of an individual. The company's databases include financial reports, education and employment verification, criminal records checks, and motor vehicle records.[6] ChoicePoint also sells personal information on citizens of Argentina, Brazil, Columbia, Costa Rica, Mexico, Honduras, Nicaragua, Guatemala, and Venezuela.[7]
2. Several agencies have contracts with Dun and Bradstreet in order to obtain personal information of business owners.
3. Lexis Nexis sells a broad array of information to government, including access to its "Nationwide Person Tracker," a database of 324 million individuals along with their

[2] The Commission recommended that Privacy Act protections extend to the consumer credit, insurance, banking, and medical care industries. U.S. Privacy Protection Study Commission, *Personal Privacy in an Information Society* (Washington: GPO, 1977), available at http://aspe.hhs.gov/datacncl/1977privacy/toc.htm

[3] Sample documents are enclosed as attachments "A-D." An entire collection of documents obtained from the Justice Management Division are online at http://www.epic.org/privacy/publicrecords/jmdchoicepoint.pdf.

[4] *See* attachment A.

[5] *See generally* ChoicePoint Online List of Services, available at http://www.epic.org/privacy/profiling/choicepointlistofservices.pdf;

[6] *See* attachment B.

[7] *See* attachment C.

Social Security Numbers. Lexis Nexis also sells motor vehicle records, flight license records, professional license records, and a military personnel location service.[8]

4. One document obtained from the Internal Revenue Service shows that the agency wished to obtain 25,000 credit headers a month from private databases. Experian, one of the largest credit bureaus, is listed as the source for credit headers and full reports for IRS access.
5. One document obtained from the Immigration and Naturalization Service shows that the agency queries private sector databases 20,000 times a month.
6. Although some documents reference regulations that prohibit personal use of these information services, none indicates that the agencies audit or otherwise monitor agency use or misuse of the records systems.

The federal agencies that purchase this information are circumventing privacy protections passed by Congress in the Privacy Act. In effect, federal agencies are able to access detailed personal information, maintained by the private sector, while technically side-stepping obligations under the Privacy Act. Simply put, since the federal government is prohibited from building a general national data center, agencies have privatized this function, and can now obtain information on anyone from their desktop computers.

Now that commercial-sector brokers regularly sell information to the federal government, thereby allowing the government to have access to detailed dossiers without actually maintaining the database, Congress should revisit this issue, and apply Privacy Act protections to the private sector.

Future Possibilities: Employment of Consumer Data for Government Data Mining

A future data mining risk flows from private-sector collection of consumer habit information. Some of the same companies that are engaged in private-public sector partnerships also maintain databases of consumer information that could be sold to the government. Experian, for instance, sells marketing databases with the names, addresses, and other personal details of racial and ethnic minorities.[9] The company also sells medical information for marketing. Its medical marketing databases, for instance, include a list of people believed to be suffering from bladder control problems.[10]

Collectors of consumer information are willing to categorize, compile, and sell virtually any tidbit of information. For instance, the Medical Marketing Service sells lists of persons suffering from various ailments. These lists are cross-referenced with information regarding age, educational level, family dwelling size, gender, income, lifestyle, marital status, and presence of children. The list of ailments includes: diabetes, breast cancer, and heart disease.[11] Other

---

[8] *See* attachment D.

[9] Experian List Services Catalog (on file with EPIC), excerpts available at http://www.epic.org/privacy/profiling/experianlistservices.pdf.

[10] *Id.*

[11] Consumers By Ailment, Medical Marketing Service (on file with author). This list has been removed from the Internet, but is still available via the Google Cache: http://216.239.53.100/search?q=cache:kKDlOrzU2Q4C:www.mmslists.com/consumers_by_ailment_counts.htm+&hl=en&ie=UTF-8.

companies sell databases of information relating to individuals lifestyle habits, reading preferences, and even religion.[12]

Another consumer profiling company divides individuals into fifteen different groups, which are in turn categorized into various subgroups. These include "Pools & Patios," "Big Fish Small Pond," "Shotguns and Pickups," and "Urban Cores."[13] The assumptions drawn on these categories of people often can be racially-charged and objectionable. They also can catalog populations of people who are at-risk for hate crimes or other stigmatization. For instance, PlanetOut.com sells lists of consumers identified as homosexual.[14]

Consumer collection of information occurs through aggregating information from online and offline purchase data, supermarket savings cards, white pages, surveys, sweepstakes and contest entries, financial records, property records, U.S. Census records, motor vehicle data, automatic number information, credit card transactions, phone records (Customer Proprietary Network Information or "CPNI"), credit records, product warranty cards, the sale of magazine and catalog subscriptions, and public records.[15]

There are no standards for the collection of consumer data, and it widely known in the industry that consumer information databases are riddled with errors. There is a serious and credible risk that this consumer information may be employed for data mining purposes related to risk assessment. Congress should act now to prevent this improper, secondary use of personal information.

Recommendations

Congress should take action to ensure that the government does not use commercial data sources for data mining.

1. Congress should begin oversight hearings on the information brokers' practices.
2. Agencies should be asked to routinely report on the private-sector databases that they have purchased, including the number of records obtained, and the specific characteristics of the data.
3. Congress should determine whether Privacy Act obligations should be applied to the entire information broker industry, as these businesses are now engaged in the practice of building government profiles of individuals that would be regulated under the Privacy Act.

---

[12] A number of companies sell religious affiliation information, including the Post-Newsweek company's "Catholic Subscriber" database, which is described online at http://dmipublic.directmedia.com/datacard/dmicards/dmi/47/dm47610.stm.

[13] The Claritas Prizm and MicroVision clustering services are online at http://cluster2.claritas.com/YAWYL/Default.wjsp?System=WL.

[14] Meet Your Best Customer, PlanetOut Partners, at http://www.planetoutpartners.com/sales.html (last visited Jan 20, 2003).

[15] *See generally* Experian Insource Enhancement, available at http://www.epic.org/privacy/profiling/experianinsourceenhancement.pdf

We appreciate this opportunity to share with the Committee information about the risks inherent in certain types of data mining. Please contact us if we can be of more assistance in this debate.

Sincerely,

Marc Rotenberg
Executive Director

Chris Jay Hoofnagle
Deputy Counsel

AUG-14-2002 12:46 DOJ/JMD/PSS 202 307 1933 P.06/41

PIC ATTACHMENT A

ORDER FOR SUPPLIES OR SERVICES

PAGE 1 OF 2 PAGES

IMPORTANT: Mark all packages and papers with contract and/or order numbers.

1. DATE OF ORDER: SEP 25 2001
2. CONTRACT NO. (If any): 02-F-0464
3. ORDER NO.: 02-F-0464 DO 001
4. REQUISITION/REFERENCE NO.:
5. ISSUING OFFICE (Address correspondence to): U.S. Department of Justice, Procurement Services Staff

6. SHIP TO:
a. NAME OF CONSIGNEE
b. STREET ADDRESS
c. CITY
d. STATE e. ZIP CODE
f. SHIP VIA

7. TO:
a. NAME OF CONTRACTOR
b. COMPANY NAME: Choicepoint, Inc.
c. STREET ADDRESS: 11350 Random Hills Road, Suite 240
d. CITY: Fairfax
e. STATE: VA
f. ZIP CODE: 22030

8. TYPE OF ORDER
— a. PURCHASE
REFERENCE YOUR: Please furnish the following on the terms and conditions specified on both sides of this order and on the attached sheet, if any, including delivery as indicated.
— b. DELIVERY -- Except for billing instructions on the reverse, this delivery order is subject to instructions contained on this side only of this form and is issued subject to the terms and conditions of the above-numbered contract

9. ACCOUNTING AND APPROPRIATION DATA: 2T959-1507 0T0940100704 DC 1507 O/C 2533
10. REQUISITIONING OFFICE
11. BUSINESS CLASSIFICATION (Check appropriate box(es)): — a. SMALL — b. OTHER THAN SMALL — c. DISADVANTAGED — d. WOMEN-OWNED
12. F.O.B. POINT: Destination
13. PLACE OF: a. INSPECTION b. ACCEPTANCE
14. GOVERMENT B/L NO.
15. DELIVER TO F.O.B. POINT ON OR BEFORE (Date)
16. DISCOUNT TERMS

17. SCHEDULE (See reverse for Rejections)

ITEM NO. (a)	SUPPLIES OR SERVICES (b)	QUANTITY ORDERED (c)	UNIT (d)	UNIT PRICE (e)	AMOUNT (f)	QUANTITY ACCEPTED (g)
1.	Obligation of funds for access to Choicepoint databases under BPA 02-F-0464. Eleven million dollars is obligated for fiscal year 2002. Submit monthly invoices IAW terms of the BPA.	1.00		11,000,000.00	11,000,000.00	

18. SHIPPING POINT
19. GROSS SHIPPING WEIGHT
20. INVOICE NO.
17(h) TOT. (Cont. pages)

SEE BILLING INSTRUCTIONS ON REVERSE

21. MAIL INVOICE TO:
a. NAME
b. STREET ADDRESS (or P.O. Box)
c. CITY
d. STATE e. ZIP CODE

11,000,000.00 17(i) GRAND TOTAL

22. UNITED STATES OF AMERICA BY (Signature): Tim Spencer

23. NAME (Typed): Tim Spencer
TITLE: CONTRACTING/ORDERING OFFICER

NSN 7540-01-152-8083
Previous edition not usable

WHITE

OPTIONAL FORM 347 (REV. 6/95)
Prescribed by GSA/FAR 48 CFR 53.213(c)

PIC ATTCHMENT B

02-F-0464
Modification 0001

Pricing Schedule D

AutoTrackXP Online Public Record Data

AutoTrackXP offers Federal customers an additional interface for access to ChoicePoint's online public record information. ChoicePoint proposes adding AutoTrackXP to the DOJ order at the following per-unit prices:

Product	Price
Search	$2.00
Link - Other Address	$1.00
Link - Neighbor	$1.00
Link - Other SSN	$1.00
Smart Search	$5.00
RT Phone Search	$0.75
Basic Report	$5.00
Natl Comp Report	$10.00
Basic+Assoc Report	$7.00
Natl Comp+Assoc Report	$12.00
Business Comprehensive Report	$15.00
DE Corps Business Search	$10.00
DE Corps File Search	$19.00
DE Corps Abstract Detail	$9.00
Link IT	$2.00
Provider Verification Search	$1.50
Provider Business Report	$7.00
Provider Individual Report	$7.00
ABI Search	$1.00
ABI Detail	$1.00
ABI Detail Full	$3.50
D&B Search	$1.00
D&B Detail	$1.10
D&B Detail Full	$3.50
RT Vehicles	$2.00
RT Drivers	$2.00
Boats of the Nation	$2.00
Corporations of the Nation	$2.00
DEA Controlled Substance Licenses	$2.00
Deed Transfers of the Nation	$2.00
Drivers of the Nation	$2.00
FAA Pilots and Aircraft	$2.00
Faces of the Nation	$2.00
FCC Marine Radio Licenses	$2.00
Federal Employer Identification Numbers	$2.00
Federal Firearms and Explosives Licenses	$2.00
Liens, Judgements, and Bankruptcies	$2.00
Professional Licenses of the Nation	$2.00
Properties of the Nation	$2.00
Significant Shareholder Records	$2.00
Social Security Death Master Filings	$2.00
UCC Liens of the Nation	$2.00
US Military Personnel	$2.00

02-F-0464
Modification 0001

Product	Price
USCG Documented Vessels	$2.00
Vehicles of the Nation	$2.00
Boat Manufacturers	$2.00
Trademarks	$2.00
Address Profiles	$2.00
Broward Cty FL Felonies and Misdemeanors	$2.00
Broward Cty FL Warrants	$2.00
Broward Cty FL Traffic Citations	$2.00
FL Accidents	$2.00
FL Attorneys	$2.00
FL Banking Licenses	$2.00
FL Beverage Licenses	$2.00
FL Boat Registrations	$2.00
FL Boating Citations	$2.00
FL Closed Claims	$2.00
FL Concealed Weapons	$2.00
FL Condos and Co-ops	$2.00
FL Convicted Felony Offenders	$2.00
FL Day Care Licenses	$2.00
FL Department of Education	$2.00
FL Divorces	$2.00
FL Driver Licenses	$2.00
FL Handicapped Parking Permits	$2.00
FL Hotel and Restaurant Licenses	$2.00
FL Insurance Agents	$2.00
FL Lab Licenses	$2.00
FL Marriages	$2.00
FL Money Transmitters	$2.00
FL Notary Licenses	$2.00
FL Nursing Licenses	$2.00
FL Real Estate Licenses	$2.00
FL Salt Water Product Licenses	$2.00
FL Securities Dealers	$2.00
FL Sexual Predators	$2.00
FL Sweepstakes	$2.00
FL Tangible Property	$2.00
FL Tobacco Licenses	$2.00
FL Unclaimed Property	$2.00
FL Vehicle Registrations	$2.00
FL Worker Compensation	$2.00
FL Real Property	$2.00
FL Medical Malpractice	$2.00
Miami-Dade Cty FL Warrants	$2.00
FL Statutes	$0.00
Telephone Listings	$2.00
RT Telephone Listings	$2.00
Dallas Cty TX Criminal Histories	$2.00
TX Beverage Licenses	$2.00
TX Criminal Histories	$2.00
TX Divorces	$2.00
TX Hunting and Fishing Licenses	$2.00
TX Marriages	$2.00
TX Trademarks	$2.00
TX Voter Registrations	$2.00

GA Residents	$2.00
NY Residents	$2.00
OH Residents	$2.00
OR Beverage Licenses	$2.00

Additonally, ChoicePoint will offer DOJ customers flat-rate pricing for access to AutoTrackXP under a usage-based schedule. Customers wishing to obtain flat rate pricing for AutoTrackXP will be billed on a transactional basis for three months, under the pricing terms described above. After this three-month period, ChoicePoint will compute the average usage over the three-month period and apply a 10-30% discount (dependent on total volume) to this average to establish a monthly flat rate going forward. For example, if a customer agency averages $10,000 in usage over the initial three-month period, ChoicePoint will charge a flat rate of $7,000 per month for the duration of the customer's fiscal year. Flat rates will be renegotiated on an annual basis, dependent upon customer usage.

EPIC ATTACHMENT C

02-F-0464
Modification 0001

Pricing Schedule B

Description **On Demand Searches** **International Searches**	Price
Argentina Citizen	$30.00
Argentina Reverse Telephone	$15.00
Argentina Telephone-Other	$15.00
Argentina Ex/IM	$75.00
Argentina Co. Details	$40.00
Brazil Reverse Telephone	$15.00
Brazil Telephone-Other	$15.00
Brazil Ex/IM	$75.00
Brazil Investor Profile	$100.00
Brazil Co. Ownership	$100.00
Brazil Company Staff	$50.00
Brazil Company Details	$40.00
Columbia Citizen	$90.00
Columbia Co. Details	$80.00
Costa Rica Citizen	$30.00
Mexico Citizen	$30.00
Mexico Driver's License	$20.00
Mexico Vehicle ID	$20.00
Mexico Reverse Telephone	$15.00
Mexico Telephone-Other	$15.00
Mexico Company Details	$40.00
Multi-Nation Aircraft	$10.00
Honduras Citizen Search	$90.00
Nicaragua Citizen Search	$90.00
Guatemala Citizen Search	$90.00
Venezuela Citizen Search	$90.00

LEXIS·NEXIS

EPIC Attachment D

Find Elusive Facts That Strengthen Investigations ...

The LEXIS®-NEXIS® services, the world's leading full-text electronic legal, news and business information services, puts a suspect's paper trail -- property records, current business affiliations and other hard-to-find evidence links -- at the your fingertips.

- *No special equipment needed.* Access LEXIS-NEXIS services through the same personal computers and modems you also use for word processing.
- *Easy to use.* Finding information through the LEXIS-NEXIS services can be as simple as typing a name or an address. To ensure your users are comfortable using the LEXIS-NEXIS services, training is included in your subscription. Plus, all users have toll-free access to our Customer Service research and technical specialists. They're available virtually around the clock, even on weekends and holidays.
- *Predictable pricing.* One subscription can provide access to all your investigators and analysts -- a complete information and customer support package. And the price can be customized to fit your specific needs and budget.

Here's just a few examples of the information you can find within the LEXIS-NEXIS services:

Build Basic Profiles of People, Property and Events

The LEXIS-NEXIS services offer references to over 300 million names, addresses and phone numbers from sources such as property records, phone directories, deed records, census bureau data, news stories, etc. Find current and historical data:

- complete names
- dates of birth
- addresses of homes and businesses
- home and business phone numbers
- addresses and phone numbers of residences or businesses on the same street
- value of real property and/or deed transfers
- watercraft and aircraft ownership
- motor vehicle registrations and licenses
- tax liens and judgments
- registered DBA names
- professional licensing records
- Social Security numbers and names for which the Social Security Administration has paid death benefits since 1962 -- an easy way to determine if a SSN is being used fraudulently

Verify Business Connections and Relationships

Access the ABI Business Directory of U.S. and Canadian public and private companies -- phone numbers, business classifications and ownership information. Get details on businesses of all sizes -- from the corner gas station to multi-national corporations. Find:

- secretaries of state records on active and inactive corporations and partnerships
- business names and addresses
- registered DBA names
- registered agents, officers and directors
- business assets and liabilities, including tax liens and judgments listing debtors and secured parties
- companies, subsidiaries, numbers of employees, etc.
- news coverage of company events, product announcements, lawsuits and more
- banking relationships as divulged in Uniform Commercial Code filings

Check Past Litigation and Criminal Activities

The LEXIS service provides a comprehensive archive of full-text federal and state court decisions. U.S. Supreme Court decisions are available within an hour of release. And it's easy to do an intensive analysis of all available federal and state opinions -- with one search request. The LEXIS-NEXIS services also include full-text administrative opinions and actions, news coverage, details on lawsuit settlements and more. Find:

- parties in local, state and federal lawsuits and criminal cases, including RICO cases
- how the judge and/or jury ruled -- and why
- state attorneys general opinions
- federal and state agency actions
- National Financial Institutions sanctions and legal actions
- federal bankruptcy cases and filings
- judgments and liens
- civil and/or criminal docket information
- nationwide case verdict and lawsuit settlement details
- local/regional news coverage -- past arrests and arraignments, plus statements of parties, witnesses, judges, etc.

Expand Your Investigation -- Build On The Facts

The LEXIS-NEXIS services can help you broaden your research beyond your local resources. Find:

- names and credentials of expert witnesses in hundreds of fields
- verdicts or settlements in similar cases nationwide
- The *National Directory of Law Enforcement Administrators, Correctional Institutions and Related Agencies* for fast access to other agencies and prosecutors -- municipal, county, state and federal

Find Background for Major Policy Decisions

The LEXIS-NEXIS services put you in touch with the background information you need to make informed decisions and recommendations. Locate hard-to-find facts -- everything from the latest grant announcements in the *Federal Register* to the full text of pending bills in your state legislature or in the U.S. Congress. Find:

- details on how other agencies have incorporated legislation such as the *Americans with Disabilities Act* and the *Family and Medical Leave Act* into department policy.
- state-of-the-art news on manufacturers of cruiser video cameras, laser radar guns, pepper spray and more
- check product claims and comparisons
- suggestions from law enforcement administrators for dealing with vital personnel issues such as peer counseling
- details on new federal grants, including specific requirements for grant proposals
- news on how other cities and counties have addressed issues such as teen curfews, gun ordinances, etc.
- legislation pending at the state and federal levels that could affect your department -- and which legislators support and oppose it
- decisions from the major labor-management agencies, including the National Labor Relations Board, Equal Employment Opportunity Commission, Occupational Safety and Health Administration, National Mediation Board, Federal Services Impasse Panel, Merit Systems Protection Board, and other federal and state labor agencies

For more information about the LEXIS-NEXIS services for law enforcement, call us at 1-800-985-8766 extension J-400

TOTAL P.09