Attendance:

Toby Milgrom Levin, Senior Advisor, Privacy Office, U.S. Department of Homeland Security

Hugo Teufel, III, Chief Privacy Officer, U.S. Department of Homeland Security

Jay M. Cohen, Under Secretary, Science and Technology Directorate, U.S. Department of Homeland Security

David Jensen, Associate Professor of Computer Science, University of Massachusetts

Martha Landesberg, Senior Privacy Analyst, Privacy Office, U.S. Department of Homeland Security

Fred H. Cate, Distinguished Professor and Director of the Center for Applied Cybersecurity Research, Indiana University

Greg Nojeim, Senior Counsel and Director, Project on Freedom, Security and Technology, Center for Democracy and Technology

Christopher Slobogin, Milton Underwood Professor of Law, Vanderbilt University Law School

Barry Steinhardt, Director, ACLU Program on Technology and Liberty

Peter Swire, C. William O’Neill Professor of Law, Moritz College of Law, Ohio State University

John Hoyt, Chief, Knowledge Management Tools Branch, Command, Control, and Interoperability Division, DHS Science and Technology Directorate

Stephen Coggeshall, Chief Technology Officer, ID Analytics, Inc.

Stephen Dennis, Technical Director, Homeland Security Advanced Research Projects Agency, DHS Science and Technology Directorate

Chris Clifton, Associate Professor of Computer Science, Purdue University

Dr. Anant Jhingran, Vice President and Chief Technology Officer, IBM Information Management Division

Rebecca Wright, Associate Professor of Computer Science and Deputy Director of DIMACS, Rutgers University

P R O C E E D I N G S

[Convened at 8:39 a.m.]

Ms. Levin: Good morning. The Department of Homeland Security Privacy Office is pleased to welcome you to our workshop, “Implementing Privacy Protections in Government Data Mining.” I especially want to thank all of you who have traveled from far, and I understand a number of you had delays yesterday as the result of the storm, but hopefully everyone who wanted to attend has been able to make it.

My name is Toby Levin, I’m Senior Advisor in the DHS Privacy Office, and I’m co-coordinator with my colleague, Martha Landesberg, who you’ll meet shortly, for this workshop.

Before I introduce our welcoming speakers, I have just a few housekeeping announcements to make. First, you should have a packet for the workshop which includes the agenda and the bios -- we will not be doing biographical introductions -- as well as copies of some of the key slides from the presentations that you’ll be seeing for today and tomorrow. We will post a transcript of the workshop on our workshop website at www.dhs.gov/privacy, hopefully by mid-August. In order to enable additional comments and so that you can perhaps include responses, reactions to what you hear throughout the workshop, we are going to be extending the comment deadline to August the Fifteenth; comments instructions are on our website and we look forward to your additional input.

I want to apologize that we're not able to provide refreshments, but due to our ethics rules, we're not allowed to use your tax dollars to fund refreshments for the workshop. But there are coffee and other refreshments across from the auditorium. Feel free to use those during the breaks but please return promptly.

We'll break for lunch about 11:45 and resume at 1. In addition to the dining options that are located on this floor and upstairs in the hotel, you have a list of eateries in your packet.

After our welcoming speakers we'll move directly to our program; we've set aside the last fifteen minutes of each panel for you to ask questions, and there is a mic up front where you can line up when you're told, queued to line up so that we can hear from you and your questions and any input that you would like to provide. Make sure that you identify yourself by name and affiliation, if any, so that we can have an accurate transcript.

Martha Landesberg and I want to thank our privacy team who helped in preparation of the workshop, particularly Sandra Debnam, Sandra Hawkins, Rachel Drucker, Richard Moore, and the rest of our Privacy staff who are here today.

And finally, if you would please silence your cell phones so that we won’t have interruptions, I think we’re ready to begin. It’s my pleasure to introduce my leader, Hugo Teufel, Chief Privacy Officer of the Department of Homeland Security.

PRESENTATION OF HUGO TEUFEL, CHIEF PRIVACY OFFICER, U.S. DEPARTMENT OF HOMELAND SECURITY.

Mr. Teufel: Good morning. I’m Hugo Teufel, Chief Privacy Officer at the Department of Homeland Security; and I have a few remarks before we have our guest speaker who will be joining me up here in a minute. And I see our colleagues from the Government Accountability Office are here, so, yes of course we comply with the ethics requirements and appropriations laws. There will be no free lunches or snacks or coffee or tea.

Well, it's my pleasure to welcome you all to this, our fifth, in a series of workshops over the existence of the Privacy Office at the Department. Our goal for our workshops has been to educate the public, educate our office, educate the Department and others in government on cutting-edge privacy issues, and today's workshop, "Implementing Privacy Protections in Government Data Mining" should be no exception.

We're fortunate to have with us today and tomorrow some of the most prominent experts in the field, both with respect to privacy as well as with respect to technology, coming to talk to us and to you about the subject matter of this workshop. And I'm really excited about it; I've got to tell you, though, that I'll be popping in and out today and tomorrow because of some unexpected meetings up at the Nebraska Avenue Complex.

We're particularly pleased that the Under Secretary for the Science and Technology Directorate -- my friend Jay Cohen -- will be here to help open this workshop. And then also, computer scientists from the DHS Science and Technology Directorate who are actively engaged in learning how data mining can further the Department's counter-terrorism mission have joined with my staff to make this workshop not only a possibility but hopefully a success.

So beyond that, why are we doing this? Well, if you have followed developments up on the Hill, you are aware of the various annual data mining reports that my office has issued and certainly you're familiar with Section 804 of the 9/11 Commission Report Act, which requires of all agencies data mining reports, and our Department is no exception. Earlier this year we issued a letter report in which we advised Congress that we would be doing some further work, among which would be this workshop and that we would be reporting back to Congress on what we found, and so we are convening the workshop in part because of Section 804 of the 9/11 Commission Report Act.

So I think at this point it's appropriate and necessary to remind everyone here that it's Section 201 of the Homeland Security Act of 2002; the Department has a Congressional mandate to conduct data mining activities in furtherance of its mission. So we looked into this because of course we read the plain language of the statute, and certainly it says that, and we agree with that. And what we wanted to understand a little bit better, what is it that Congress was thinking? And at the time, then House Majority Leader Dick Armey responded to those who were concerned about that provision of the Homeland Security Act and referred approvingly to the new Privacy Office that was to be stood up, the office that Toby and I are in right now, and said that that office would be there at the Department to make sure that there were not abuses of data mining.

So for us, the question isn't then, whether the Department should be conducting data mining? It is, rather, how DHS should use data mining and what ways can it do so that both respect privacy and also support the integrity and effectiveness of the Department's Homeland Security initiatives. So in the interest of brevity, and because we've gotten started a little bit late, I want to wrap up my remarks.

And I want to introduce our guest speaker, DHS Under Secretary Jay Cohen, who heads the Department's Science and Technology Directorate. When Jay joined the Department in August of 2006, a month after I moved over to the Privacy Office, he immediately tackled the challenge of the Science and Technology Directorate so that it could foster the development of vital technologies for protecting the nation. Jay deserves tremendous credit for his efforts to transform the Directorate into an efficient and responsible organization that makes vital technical contributions to the DHS mission to protect against and respond to catastrophic events. The S&T Director provides technology solutions to help the men and women who face risk every day on the front lines of Homeland Security to do their jobs more quickly and safely with greater accuracy. And with that, I'll stop. Jay Cohen, Under Secretary of Science and Technology Directorate. Thank you all very much.

[APPLAUSE]

PRESENTATION OF JAY COHEN, UNDER SECRETARY OF SCIENCE AND TECHNOLOGY DIRECTORATE, U.S. DEPARTMENT OF HOMELAND SECURITY.

Mr. Cohen: Well, good morning. And thank you so much for sharing your most valuable asset with us, and that is your time and also your thoughts, at this workshop. It's a real pleasure to work with Hugo and his team. I don't know anybody who has a tougher job in Homeland

did it to eliminate or minimize seams because terrorists and criminals will always take advantage of seams. And anything that eliminates or minimizes those is good for security and bad for those who would do us harm. So that's my little shtick here; I'm not a Mac person, but I'll do my best with this computer.

So what are the goals in law of the Science and Technology Directorate? And I can tell you, as Chief of Naval Research for six years of a three-year tour and the Office of Naval Research was established in 1946. Half a page in Title 10, it says there will be an Office of Naval Research, it'll be led by a Navy Admiral, report to Secretary of the Navy, and it'll do good research. In 2003, of the 183 pages creating the Department of Homeland Security, 17 pages describe the S&T Directorate. You know, a camel was that animal created by committee, so we could have ended up with a camel. We didn't. It was very, very thoughtfully done. And so half a page in 1946, 17 pages in 2003; it shows you the impact of word processing on the legislative process.

But to synopsize in the law what are the goals and what do I follow, number one, is to accelerate the delivery of enhanced technological capabilities to my customers. Who are my customers? In law, they are the 22 components: TSA, Border Patrol, Coast Guard, Secret Service; and in law, first responders -- the police, fire, emergency, medical, bomb disposal -- our heroes. And I had no appreciation for the scale of our first responders in America. We have 35,000 fire departments in America -- 35,000 fire departments, of which 80 percent are volunteer. When I go and visit them and I say, ‘Hi, I’m from Washington. I’m here to help.’ They say, ‘Great. Buy a raffle ticket or a muffin because we need a new pumper.’ I mean, this is America. So it is a federal goal with a local execution; I can tell you it’s a great challenge.

Second is to establish -- in my words -- a lean and agile government service -- world-class S&T management team. Ladies and gentlemen, I don’t do S&T and my people don’t do S&T; we are a venture capital fund, we are a mutual fund, we invest in S&T to de-risk it to give capabilities to our customers. And when I say government service -- because some political appointees -- people like me come and go, but the half-life of Science and Technology is such that there must be a continuum, and so that’s where government service is so critically important. And in my experience in Navy and in Homeland Security, is that Science and Technology -- unless I do something stupid and Hugo works very hard to help me from doing something stupid -- is bi-partisan, non-partisan, and I believe that that is how it should be.

And then finally -- and this is a labor of love for me -- is to provide the leadership and opportunities for the next generation of our workforce. This is STEM, Science, Technology, Engineering and Math. Ladies and gentlemen, we’re in crisis in this country today. In fact, we’re in crisis in most of the western countries. People in middle school, young people, are turning away from science and math, and when you ask them why, they tell you the truth -- it’s too hard. They’re the Playstation generation; they want instant gratification. If we don’t turn this around, ladies and gentlemen, in my opinion, in fifteen or 20 years we will not be a first-world economy. So that’s a little bit of the background.

Now, what are the threats that we face? This is a PowerPoint presentation, we’ll leave copies, you can move the boxes around however you want. I view the threats from terror -- and oh, by the way, DHS is responsible for all threats. In the law, it's not just terror threats, it's also natural disasters, like earthquakes and fire and flooding, tsunamis, et cetera. But I view the threats as bombs, borders, bugs, and business -- those are the original four b's. It turns out I've got six divisions; two of them didn't have b's originally. I think last spring they saw the Bee Movie, but the division directors came to me and they said, 'Hey, we're without b's; we're b-less.' So I added two b's and that's bodies -- that's human factors, and buildings, which is infrastructure protection. You understand bombs, you understand borders, you understand bugs; what's business? Business is the underlying cyber-backbone that enables everything we do, and it is a very new area, and very threatening and scary area, of warfare.

So if you look across the bottom left to right, you see consequence of occurrence low to high, and then likelihood of occurrence. We're always going to have physical attacks; that's the reality of the world that we live in. If you look in nuclear, that's a nuclear device -- that's a nuclear bomb. The consequence of occurrence of that going off are unimaginable; it’s far off the scale to the right. But today, today a terrorist would have to either buy or build a bomb, and I would tell you -- you can disagree -- that I think the probability of that is somewhat low. Maybe not tomorrow, but today. But the day after 9/11, ladies and gentlemen, we were delivering death by 37-cent stamps in the U.S. mail -- anthrax, biological attack. And so you can see while it may not be as much of a weapon of mass destruction as nuclear, its occurrence is more likely. We have seen it, we will see it again. Biological warfare is the poor man’s weapon of mass destruction. Because today, with the internet, with genomics, all it takes is a brain, a basement, a microscope, and you can create a pathogen that will give you a pandemic.

IED’s -- they’re weapons of mass influence, not weapons of mass destruction. Tom Friedman said IED’s are coming to a theatre near us, and I believe that.

But the tactics, techniques, and procedures that we use so well overseas, many of them don’t apply -- don’t apply in the United States because the Constitution, because of the Fourth Amendment -- many of the things that you're going to be discussing here. Before a bomb squad can actively jam a bomb and its trigger device, they have to get a license from the Federal Communications Commission. It's a very interesting challenge; not what you're going to be addressing today.

But what you are going to be addressing today is up in the upper-right, high and to the right, and that's cyber, because every three seconds someone's losing their identity. And you have Estonia, and you understand if your background, the challenges of what a cyber-attack could do. Those of you who have children or grandchildren in college, you understand they live from ATM swipe to ATM swipe. And if we can't do that, in my opinion, there will be panic in the streets. So you can agree or disagree, but that's sort how I see life.

So Hugo has already talked about the enabling legislation, I think very well thought out, well debated; it has been modified, we've had a change in the Congress in the ensuing years. We get to testify a lot. Everything I do -- I'd contend 99.9 percent of what I do is unclassified. We invite the Congress to our processes, we invite the Inspector General; and Hugo has workshops like this, which I know will be the first of many to come. So the authorizing legislation for me, I have summarized it, in the first, telling you what my goals were. I think I -- I’m too fast.

So as we look at data and we look at the threats, and I looked at what is unique in Homeland Security, I settled really on two things. Because the enabling legislation is very thoughtful, it tells me not to recreate the National Institutes of Health and not to recreate the Center for Disease Control and not to recreate the Department of Energy or Department of Defense labs -- and I think that was very thoughtful -- but in exchange, it allows me to leverage everything they do. I can’t tell them how to invest their billions of dollars in research, but they give me full disclosure. And it really does work. And then I take my precious dollars, our precious dollars, and apply it to the things that are unique in Homeland Security and the missions that we have.

So from my perspective, as I looked around at all of the areas of Science and Technology, all the different disciplines, the two that I felt -- and I still feel that way after two years on the job -- that were unique, was number one, the psychology of terrorism. Why do terrorists do what they do? I mean, you can view them as criminals, you can view them as armies, et cetera, but why do they do what they do? It was not clear to me any other component of government was investing in that.

And the second area is hostile intent, and we're going to talk a little bit about that. Are there ways of knowing that someone is about to do something bad to our society? And so these are focus areas that we are looking at. This is new science. We've gone to the National Academies of Science to help us define those sciences. You know, after World War II, the Battle of the Atlantic, strategic bombing, the science of operations, research operations analysis, was born. And after Sputnik aerospace, you get the idea. As time moves on, challenges change; new areas, new disciplines develop. But how do we know that what we think is appropriate research, even vetted by the Privacy Office, even briefed to the Congress; and of course, the press is very interested in this, as they should be. I mean, at the end of the day, ladies and gentlemen, I am a citizen, I value my privacy, I respect and value your privacy, and when I'm done with government service, I will again be a citizen. I think I’m a citizen while I’m still in government service, but you get the idea.

So Dr. Sharla Rausch and her people are represented here today. She’s head of my Human Factors; this is a division that I set up. There’s a great ad by Dow Chemical, it talks about the human element. I love that ad because it’s the human element that creates terrorism and it’s the human element that will solve the challenges that we have. It really is all about humans.

But Sharla went ahead and worked with the Privacy Office and others, established on her own, the Community Perception of Technologies Panel. And so these are just average people from a wide cross-section -- they have a picture of them here -- and we go ahead and we brief to them. This is our initiative, what we’re looking at, what our research areas are, how we’re approaching it. They are not necessarily experts in privacy; we go to Hugo and his team for that, and I’ve got Jen Schiller on my staff. And I can tell you, she is very tough on me. This is an area where an ounce of prevention is worth pounds if not tons of cure.

And it’s very interesting to sit down, and I sit down with this panel, and get their feedback on their perception on what we are doing, and then we modify as appropriate.

So let’s talk a little bit about the areas of research that we are doing, and then I’ll conclude because I know Hugo does want to get you back on track into panels and the discussions are so important. So I’ll go through this very quickly.

And I must tell you that personal identifier information was a new concept to me when I came on board, and so in the last two years I’ve had a steep learning curve. And I also understand that we can be looking, you know, at totally unclassified, totally public information, but perception of how that is analyzed, et cetera, becomes an issue on its own. And I know you’re going to address all those things. The Congress enabled the S&T Directorate with Centers of Excellence. I have two pillars of basic research: universities and laboratories. And so at the University of Maryland, one of our earliest Centers of Excellence was the Study of Terrorism and Responses to Terrorism (START). In Washington if you’ve got a good acronym, everything else follows. So I salute the University of Maryland on getting started with this.

But as you can see, this is the largest terrorist event database; more than 80,000 events. Basically, this is all out of public venue, public information; and you can see incidents versus fatalities by area, et cetera. It is unclassified, it’s kept up to date, it’s available for researchers, et cetera.

The next area is Biodefense Knowledge Center. I talked to you about my concerns for the poor man’s weapon of mass destruction. This is a 24 by 7 secure website; it uses data fusion, and basically it’s talking about capabilities, because as you know, a bio and genomics are moving at the speed of heat. And so it’s available for subject matter experts, et cetera.

Suspicious behavior detection. The goal here is to identify deception and hostile intent in real time using non-invasive sensors. We’re going to talk a little bit more about this when what we call the FAST program, FAST is Future Attribute Screening Technology. But the goal here is to develop a prototype to detect deception and hostile intent in real time. I must tell you, almost everything we do as we look at, for example, transportation security, is to maximize the throughput of primary screening so the lines are as short as possible, and then only focus on secondary screening which can be question and answers. Those of you who fly overseas, you know they do it a little bit different than we do it. You start out with the questions, and then you go through the metal detector. We put you through the metal detector, and then after there's suspicious activity, we then go into the secondary screening. Secondary screening is very expensive, intensive, and it interferes with our lives.

So what is FAST? Aviation in large measure is a closed transportation system. We put up with the lines because we believe that if we keep bad people and bad things off of aircraft -- and oh, by the way, aircraft is a fixation by some of our terrorists, enemies with aviation -- if we keep bad people and bad things off of planes, the plane will take off and land safely. It's a closed system. The only challenge is the shoulder-fired weapons, and we're working on that independently.

But when you get into Metro, you get into Amtrak, you get into buses, you get into mass transit, where you have thousands of people, we can’t use the same procedures; those are open. And if we kept a bomb from getting on at a Metro or an Amtrak station, you still have miles of unsecured railroad. So what is the balance? And so what we’re looking at here -- and let me give you an example -- during the SARS epidemic overseas, several Asian countries used infrared cameras. As you got off the plane and you walked into Customs, these cameras didn’t care if you were tall or short, male or female, they didn’t care about ethnicity, they were just looking at your forehead. They’re looking at your forehead. And if on infrared your forehead was warmer than everyone else’s forehead, you most likely had a fever, and that’s a precursor or an indicator of SARS, and they didn’t want to have the spread of SARS, and so you went into secondary screening. That is the level of screening that we’re talking about. So if you’re a terrorist, you want to get to your target, you may be nervous, you may be perspiring, your forehead may have evaporative cooling, your heart rate may be raised, your eyes may be flashing, your gait may be different. There are micro-facial features that give away -- and this is a brand new science that we’re learning about today. Are you telling the truth or are you deceptive? And so the goal here is in a public event, like the Super Bowl or the Olympics, to go ahead and see if, can we do this non-invasive screening that will give us indication of hostile intent so that we can take an individual to secondary screening? Now look, your parent may have just died, you may have been late getting to the event, you may have just run; I mean, there are a lot of reasons why you can have all these indicators, so we’re looking at getting to the secondary screening. That’s the thrust of what we do.

Violent Intent Modeling and Simulation. Again, this looks at the systematic collection and analysis of information that is related to understanding terrorist group intent. So we talked about the individual terrorists -- why do they do what they do -- now, what about the group as they come together?

So that’s a summary of what we’re doing; everything we’re doing is fully vetted with the Congress, with the Privacy Office, et cetera. But at the end of the day -- as I’ve told you with my basic mission -- product is job one. Getting those tools to those that would make us safe or keep us safe is what Science and Technology is all about.

So I thank you so much for spending your time here. I wish I could spend a day-and-a-half with you; I think is going to be one of the most fascinating panels that have occurred in the short history of DHS. Remember, we're only five years old. Some of you have 5-year-old grandchildren or children; you know how mature 5-year-olds are, but all the vectors are in the right direction. And the only question I ask myself and I ask my people, and I hope this never happens, I hope there's not another attack, I hope there's peace and happiness in the world. But if you listen to most of the experts on both sides of the aisle, they will tell you, there will be another attack. Our terrorist enemies want to make it even more devastating than that of 9/11. And the question is not if, it is when. And so the question I ask myself every night is, under my tenure will we have done enough with the resources and tools that I have, consistent with the laws and our culture, to make us as safe as we can be? So with that thought, I'll leave you. Hugo, thank you so much for giving me this opportunity, and I look forward to the results of the workshop. Have a great day. Thank you.

[APPLAUSE]

Ms. Landesberg: Thank you, Under Secretary Cohen. I’m Martha Landesberg from the Privacy Office, and it’s my pleasure this morning to introduce our next speaker to you. He is Professor David Jensen who is an Associate Professor of Computer Science and Director of the Knowledge Discovery Laboratory at the University of Massachusetts Amherst. Professor Jensen currently serves on DARPA’s Information, Science, and Technology Group, and he was an analyst in the Office of Technology Assessment from 1991 to 1995. I give you Professor Jensen.

[APPLAUSE]

PRESENTATION OF DAVID JENSEN, ASSOCIATE PROFESSOR OF COMPUTER SCIENCE, UNIVERSITY OF MASSACHUSETTS.

Mr. Jensen: Thank you. Thank you very much. Under Secretary Cohen is a difficult speaker to follow, and so I hope I can keep this as interesting and relevant to today’s conversations. So what I’m going to talk today about is at some level somewhat boring in that it is about definitions. But as many people have said, words mean what we want them to mean. And I think in this particular case, data mining means many things to many different people.

And so I’m going to try today to talk about the range of definitions, and the ways in which we can come to a definition that is both consistent with what the technical community is doing, which is my community, and also consistent with what we mean in a policy context.

So what I’ll talk today about are, first, I’m going to give you some very simple definitions, frequently used definitions of data mining. Then I’m going to give a fairly extended example of some work that I’ve done recently in detecting securities fraud because I think it’s a good example of what modern technology is doing in the area of data mining, and gives us some concrete things to refer back to to try and expand and make more realistic the definitions of data mining that we’re going to be talking about. Then I’ll present some revised definitions, and finally try to answer the question, why we should care about definitions, and talk about how on some sort of expanded understanding of data mining can reframe some existing issues that are often brought up about the technology and potentially raise interesting new issues -- new policy issues.

By the way, if you have a question that is specific to some slide or some comment I’ve just made, please feel free to raise your hand; I’d be happy to take the question then. If I don’t see you, give me a shout. And also, there’ll be a period at the end where we’ll take more questions of the more general kind.

So the major points I’m going to be talking about today are first, that there are simple definitions that portray data mining as a process of filtering or extraction. That these definitions are very easy to state, and in some ways, very vivid, but they are very easy to misinterpret. They’re not really wrong, but they’re easy to misinterpret, and I’ll explain specific reasons why that’s the case. More useful definitions of data mining portray it as an iterative process where you are both learning and doing probabilistic inference, and you’re doing that over interconnected data records, not data records that are independent from each other. Finally, I’ll say that these definitions identify different issues for policy discussions, and I would argue, more interesting and useful ones.

So let’s look at some of the simple definitions. The first is the one that I think has brought us to today’s meeting, from the Federal Agency Data Mining Reporting Act of 2007 Secretary Cohen referred to, in which the -- well, the definition says, it is a “program involving pattern-based queries, searches, or other analyses of one or more electronic databases.” And then there are a series of caveats that I think are really very specific to the Act, saying, well, this has to be done by a federal agency or an agent of a federal agency, it has to be about identifying terrorism or criminal activity instead of other things. But the key thing here is to focus on this question of pattern-based query searches or other analyses.

Now, there are a variety of other definitions of data mining. Let me give you some from the more technical end of the spectrum. “The science of extracting useful knowledge from data repositories,” this is from the Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining, our Curriculum Committee that came up with this definition.

There’s also a very well-known definition from some of the founders of the field, “The non-trivial extraction of implicit, previously unknown and potentially useful information from data.” That’s from an article about knowledge discovery and data mining.

Now, I tend to use the term knowledge discovery because I think it is intrinsically more meaningful and less easy to mistakenly understand than data mining is. I think data mining has a clear and obvious meaning which is wrong; the clear and obvious meaning is that you are mining for data, and that's not actually what data mining is doing. If you say gold mining, that means you're mining for gold. If data mining should be mining for data, you're not. You're mining for knowledge, and knowledge discovery gets at that. Although, it did confuse my Dean greatly when I was introduced to him as doing knowledge discovery, he looked and he said, "Isn't everyone at a university doing that?" And I said, "Yes, yes. But we're doing it with computers." He said, "Oh, well, that's very interesting," and we went on to have a pretty good conversation. There are other terms, as well -- predictive analytics, advanced statistical modeling, machine learning.

So, well, I'll stick with the term data mining even though it's not my preferred term because it is the term that stuck. So let me give you an example of this sort of work -- this sort of technology, and it's about detecting securities fraud. We've been working for about five years now with the National Association of Securities Dealers. This is the non-governmental, private organization in the United States that regulates all stock brokers. They came to us about five years ago and they said, 'We hear you're doing work in analyzing the kind of data that we need to analyze, wonder if we might do some work with you,' and we've been doing joint projects with them ever since. By the way, NASD is now referred to as the Financial Industry Regulatory Authority. They changed their name recently, but I'll be using NASD because it's what sticks in my head and also it's because what's relevant to the work we did over the past five years.

NASD is the parent of the NASDAQ Stock Exchange -- stock market, but they spun that off because their central focus is really regulatory. They monitor a large number of securities firms, branches, and individual people who sell securities to the public. Those are referred to as registered representatives or reps. And one of their responsibilities -- they have several --is to prevent and discover serious misconduct among brokers -- I'll use the term fraud. They incur fines and they can even ban individuals or entire firms from the industry and say, 'You cannot work anymore in the securities industry.' Now, they have a data set which they collected for their regulatory function, not to do analysis on, but for their regulatory function. That data set is called the Central Registration Depository, or CRD database. It consists of data about individual reps -- individual people -- about the branches that they work for, the actual physical organizations that they work, as well as the larger firms that those branches belong to. And finally, a set of event reports, which they call disclosures, where reps abide by the policies of NASD, which they agree to when they become a registered representative, they have to disclose certain events in their lives, including simple things like if a customer complains, but also including things such as liens against them, major issues in their financial history, or if they commit a felony, for instance, that's also a disclosable event. So there are a set of those disclosures that are in this data set.

Now, importantly, this data set is a large set of interconnected records. As you might expect, we know what reps work for what branches, what branches -- what firms

-- own those branches, and what disclosures have been filed on individual reps.

There are about 3.6 million reps in the data set, about 750,000 branches, about 25,000 firms, and about 625,000 disclosures, so a moderately large data set. And that covers a period of over 20 years. And we tend to focus on the smaller subset about over the past ten years or so.

Now, fortunately, the kind of conduct that NASD is trying to discover is relatively rare. Now, fraud among reps is quite rare. If you look at the stats, it's less than 1 percent of reps commit any kind of serious misconduct in a given year. In general, I think it's about 1/10th of 1 percent, so very small incidents of the kind of serious misconduct they're looking for. But it's very important to the public that that be discovered, and very important to the integrity of the industry. So their task is to take data from the past where they know that certain reps or branches were engaged in serious misconduct, and to take that data and to then try to come up with some sort of method which they can use to guide their future activities. So, particularly, they wan to do examinations and they want to do education and enforcement activities that will either prevent fraud from occurring or catch it early. And so they want to use the data they have, which they collected for other reasons, but they came to us saying, ‘We think we can do more with the data; is that the case?’

So what we did with them was to construct statistical models that try to predict the probability -- or estimate the probability that an individual rep will commit some kind of serious misconduct in the next year, the next 12 months.

And so one of the kinds of statistical models that we devised is a kind of probabilistic or statistical model which is tree-structured, and I’m showing you the whole structure of a tree here. And by the way, there are details of these models that are not included in these slides, at the request of NASD for obvious reasons. They’d rather not release exactly how they might be detecting fraud. But this is the structure of the model and it’s structured like a tree. You could think of it as a virtual Pachinko machine where you take an individual rep and their surrounding context -- the disclosures, the branches they’ve worked for, the firms they work in, the other reps that they work with -- and you drop it in the top of this tree. And then you answer a series of yes/no questions, such that it rattles down to a leaf node, a thing at the bottom which gives you a probability distribution -- their probability of committing fraud in the next 12 months. Let’s zoom in on a portion of it. So at the top node we say, ‘How many disclosures have been filed on this rep?’ If it’s greater than a certain number it goes down one branch, if it’s less than that it goes down another branch. And so on. And we ask questions here in this model about the number of the disclosures that were customer complaints, whether that rep has been designated as high-risk in previous years, other kinds of things about the current branch they work at, et cetera. And eventually we come down to a node where we say, ‘Everyone who reaches this point has a particular probability -- estimated probability of committing fraud in the next 12 months.’

Now, importantly, we construct these models automatically, or more accurately, the data mining algorithms we have devised construct these models automatically. They do that by searching a very large space of possible trees. Now, the number of those possible trees is vast. Here just for a five-level tree with the kinds of data that we have, we're talking about 10 to the 106th, an extraordinarily large number of possible trees that are out there. But, fortunately, in the technical work of our field, we've devised a fair number of efficient, approximate search methods to look at that space and not have to examine it exhaustively but still find the trees that are particularly useful or valuable in that space. And we evaluate how well those trees work by comparing them to the data for which we know the right answer; that is, we know at least we have good estimates of the -- which reps have committed fraud in the past. At least those reps that have been identified, so they are probably some -- many, in fact, that have not been identified but we know a large number of reps that have committed fraud in the past, and we can use that past data, that retrospective data, to compare the accuracy of different types of models -- different types of trees in this case.

What the models then do is to infer the values of an unobserved variable. The unobserved variable, the thing we're trying to estimate here, is the risk that a rep will commit fraud in the next 12 months. And there are also some kinds of models I won’t talk about that will simultaneously infer the value of many unobserved variables. But for the new data, for the data we want to apply the model to, we don’t know what reps are committing fraud and thus we want to estimate the probability of those.

The performance of these models has been evaluated in a variety of ways, but one of the ways that we used was we took a bunch of predictions from the model, we took some predictions from NASD’s current method of doing initial screening, and we took reps that showed up on only the list that our model created, only the list that NASD created, neither list and both lists. And then we scrambled those up and put them in front of trained NASD field examiners and we made the estimates, for example, for the previous year, for 2007. We didn’t have data about 2007 about who had actually been found to be committing fraud in that year. But NASD did have information about that, and we asked the examiners the following question, we said, ‘If we had given you this list at the beginning of 2007, how useful would it have been given that you now know what the right answers are?' And they rated each rep that we had given them on a five-point scale. One is, it would have wasted my time to know about this individual; five is, I absolutely would have wanted to know about this. When the reps showed up on neither list -- it's a little difficult to see there -- but when they showed up on neither list, the ratings were almost all one. When they showed up on NASD's current list but not ours, the ratings were roughly on average a three. When they showed up on only our list and not NASD's list, again the average was about three. And if they showed up on both lists -- combined list -- they had an average rating of about four. So showing that we are doing -- the statistical model is doing almost essentially as well as NASD's current rules for doing screening to say, which reps deserve some additional scrutiny to look and see if they're committing fraud. And if you combine the statistical model with the current expert derived rules, we can do even better.

We also got a little bit of anecdotal feedback; one of the field examiners sent us an unsolicited note along with his ratings, and he said, 'One of these reps I was very confident in rating a five,' he said. He had had the pleasure of meeting him at a shady warehouse location during what we think is a sting operation. He said he’d negotiated this rep’s bar from the industry because among other things, he’d actually used fraudulently obtained funds to attend an NASD compliance conference -- conference about how to comply with NASD rules. The examiner said, ‘If you predicted this person, you’d be right on target.’ And in fact we, with some trepidation, we went to NASD’s list, the rep was not on NASD’s list, we went to our list, he was very high up our list. So a nice anecdote to support the idea that this statistical model is a useful one.

All right. With that background and that kind of concrete reference, let’s go back to our definitions of data mining. So again, to recap the simple definitions, we’ve got from the Data Mining Reporting Act, pattern-based queries and searches or other analyses; extracting useful knowledge from data repositories; extracting implicit previously unknown knowledge. So one way of thinking about these definitions, one simple kind of visual to get is the idea of a filter. Where you say the system takes in data, there is some mining or filtering process that’s done on the data, and then out pop predictions out the other side. So that’s what we’ve got, this kind of filtering process. Now, this filtering process -- this idea of a filtering process has been encouraged by some of the most powerful people on the planet, some of the most powerful image makers on the planet. Those people reside in Hollywood, mostly. For those of you who have seen Minority Report, this is a very persuasive image. This idea that there is a black box out there that will be producing predictions, and if the predictions are certain, they are crisp, there is no doubt in them, and they put them out and that’s what then we go act on as a law enforcement agency. For those of you who watch television also, there was a short-lived show called Threat Matrix, which had some similar ideas that were frequently propounded in the show about data mining. And as you might expect, these media images are somewhat simple. They’re simple because it’s very easy to misinterpret the definitions which I’ve given you previously, which can be interpreted accurately but it’s very easy to misinterpret them. Let me explain some reasons why. The first is -- and I’ll explain more about each of these in the next set of slides -- the first is that there is only once process. The misperception is, there's only one process that encompasses what I'll refer to both as learning and inference. The second is that the records that come in the left side are disconnected from each other. Here I'm showing just individual records about reps. Third, that the inferences out the other side are deterministic. Essentially we spit out a set of reps that are bad and a set of reps that are good. Fourth, is that this is only done once, this single stage, it's a once-through process. And finally, that this process of data mining is what I'll call institutionally isolating. That is, it just sits off by itself in this little box and does its job.

Let me explain why each of these I think are not accurate, and what is a more accurate picture. The first is that the processes of learning and inference are distinct. That is, there's not just one process, but actually two. The learning phase takes in data for which we know the correct answer, or we have good estimates of the correct answer, and that puts out a statistical model. That model is then used in an inference process to take in data for which we do not know the correct answer, and put out some kind of prediction.

Importantly, the learning phase is the part of this overall process that is unique, that is the essential component of data mining. In fact, many people in the field would say that the inference part is really almost an afterthought. The goal is to put out a good statistical model. Now I will make one caveat, which is that there is a lot of study in the field about some kinds of techniques which do not immediately appear to fit into this, although I think many of them actually do. So for instance, there is some study of clustering. They’re trying to look at data and find homogeneous regions in it, and while there does not appear to be a statistical model underlying that there often can be and many of the better methods for clustering do that fairly well. So some caveats; this is a little bit simple to say that all of data mining has a statistical model underlying it. But it’s a good -- absolutely a good first pass.

So there’s this learning phase and this inference phase, and they are more or less separate. Learning is what makes data mining unique. It’s also important to point out that the inference taking data for which we don’t know the correct answer and making an inference does not require a statistical model. In fact, people do it all the time. At NASD for instance, they had a set of rules that they had sat down and worked with their experts to derive, and that was what produced an initial list that then field examiners went out and did additional investigation on. And that was not derived from data mining, that was derived from just sitting down and thinking.

Now, an example of the kind of misinterpretation -- and I don’t want to unfairly characterize GAO here in an otherwise excellent report -- they had a graphic which -- this is 2005 report -- which starts out and says, “There’s input to the process, there is an analysis process, and there’s output.” It is a slightly more complex version of this filtering that I’ve talked about and doesn’t clearly distinguish between any kind of learning phase and an inference phase. Instead, what we have, the idea here is that data mining is really complex set of database queries. It’s a complex way of filtering a database to put out matches. And I think that is a misinterpretation which is easy to make, but actually dangerous in terms of public policy. Let me emphasize again, though, that both this report and several earlier reports from GAO are really quite good and have a lot of useful information about data mining.

Second issue, data records are often interconnected, they're not sets of individual records. So I show here these individual reps, but actually what we have are a case often of a network of different types of records that are interconnected. So think back to the NASD example, we have this set of reps, branches, firms and disclosures, this set of interconnected records and those records are the -- provide us a lot more information than just having records about individual reps.

This sort of approach, often called relational learning or relational knowledge discovery or relational data mining, has become increasingly prevalent both in the technical community and now starting to make its way into applications because this can improve both the accuracy of the process and allow us to address entirely new types of tasks, for instance, predicting a link or a connection between two or more records.

Third issue, inferences that come out of the inference process are not a kind of yes/no labeling. Instead they are probabilistic. So rather than having a, these are bad brokers -- these are bad reps and these are good reps, we come out with a probability associated with each of those reps. And almost all, I think, really modern applications of data mining are giving probability distributions on variables rather than kind of yes/no classification.

What is important is that this allows us then to do -- to have a lot more information about the inferences that are being made. So for instance in the case of NASD, you could imagine if we have probabilities we could look at that last and say, it may be that there are a few high probability reps, and then immediately drops to very low. And then we would say, ‘Maybe we should only look at those high probability ones.’ Conversely, there might be a very long list, far longer than NASD would have originally thought they needed to look at, that are very high probability of committing fraud and they might say, ‘Maybe we should expand our screening program to look at a larger set of individuals, if we believe that this is an accurate assessment of probability.’

Finally, it allows you to assess accuracy in new ways because you have these probability judgments and it allows a much finer grained kind of evaluation of how well the model is doing.

So an example of these kinds of probabilistic models is the NASD model. We don't say that everyone who reaches a particular leaf node is going to commit fraud. Instead we say, there is a probability associated with committing fraud.

Fourth issue that I've talked about is inference is done in many real systems in multiple stages. So if you look at the inference process, it's not just a once through process, but instead there's feedback once you have an inference, additional things can be done with that inference in other rounds of inference about either new problems, or in fact, in some cases about the same problem.

So a really good example of this is the way in which screening for many diseases is done. So for instance, AIDS screening is done with an initially very inexpensive test which has a high false positive rate. It is of course disturbing to individuals who get a positive, but doctors are quick to point out, 'Look, this test has a high positive. And even if you get a positive on this test -- high false positive rate -- even if you get a positive, the vast majority of people are actually negative.’ So now we’re going to do the more expensive and more accurate test. So this kind of two-stage screening is a way of cutting down costs and increasing accuracy. And that’s the same way in which data mining can be done in order to do those things, to improve accuracy and to allow a wider range of types of inferences.

So it turns out actually that this is what NASD does, is that this initial set of rules they have, or now the kind of statistical model we’ve given them, gives them an initial set of reps that get enhanced scrutiny from their examination process. It’s not that other reps are not examined, and it’s not that those reps in any way are immediately considered to have committed fraud. Instead it says we should look more closely. And then a human examiner goes out and initially looks at records that are just held centrally, and then often goes out into the field and will examine records that are only held at the firm. That larger set of records both centrally and out in the field are more expensive to access and also more sensitive. And so the question is, should we actually go to the -- should NASD go to the expense and the potential security and privacy issues of examining those additional records? Well, only if they have some initial sense that it would be useful to look at those records.

Final issue. Data mining is used in a larger institutional context than it might appear at first. So if we think about data mining as -- the entire process I've described as a box; we say, well, there's obviously some kind of data gathering that has gone on ahead of time. And once we get inferences out, there's some kind of decision-making process. Those inferences do not immediately indicate what we should do, what any organization should do with that information.

And finally of course, there's some feedback. With decision-making, you may say, actually it's useful to gather additional data and perhaps do additional sorts of analysis on the data. Importantly, many of the really big public policy issues about privacy and utility are about data gathering and about decision-making, not so much about what happens inside that data mining box. The other issue I think is that the use of data mining algorithms actually imposes relatively few constraints on data gathering or decision-making. That is, just because you have maybe in advance decided to do data mining, does not mean necessarily you will collect more data. NASD is a wonderful example of this; they had already collected every last bit of data, which we’ve used over the past five years. They collected it for other reasons, but we’ve gone ahead and used those data sets to do additional kinds of analysis. And also, the output of data mining does not imply necessarily anything about what you should then do. It is input to a decision-making process that of course should take into account a large number of factors.

All right. So those are some enhancements, I hope, and some additional explanation about data mining. And now the question I think may come up, why all of this work? Why care about these definitions? And the basic point I hope to make is that this gives us I think some new perspectives, some new ways of looking at what is important about privacy and questions of utility.

So one large issue that often comes up in discussions about data mining is an issue about false positives, particularly in cases such as counter-terrorism applications or law enforcement or fraud detection applications where the prevalence of the activity, the frequency with which it happens is very low. And the critique goes something like this, if the prevalence of true positives is low, that is there are very few cases of fraud in the case of NASD, then the vast majority of inferred positives will be false positives. So even if you have a very low error rate, if you have 100,000 people who haven’t done something and 1,000 people who have, and you are 99 percent accurate, well then, you’re going to have 10 people who actually did the thing that will show up as positives. And, what did I say, 100 times that number that will show up as false positives. And so this is a simple critique, a relatively easy critique to get across, but it unfortunately presumes this kind of filtered model. So instead, if we think about these more accurate -- what I hope are more accurate definitions -- the first is that probabilistic inference can really help us here because it allows you to control the types of errors which any particular threshold that you put on that probability, we’re going to look at everyone with a probability over .95. You can change the error characteristics of any sort of screening that you do, so probabilistic inference helps us a great deal. You can also use that to account for --

in addition to the expected -- what is sometimes referred to as the expected class distribution of the data. It also allows you to adjust for the relative costs of errors. So if errors of false positives are very expensive or false negatives are very expensive, you can modify those. It's a great deal of work and what's called cost-sensitive classification or cost-sensitive inference.

The second issue is that as I mentioned about disease screening, multi-stage inference, and also it turns out interconnected data records can help you greatly reduce the false positive rate overall. There's some work that several of my students and I did in 2003 showing ways in which interconnected data records and multi-stage inference can dramatically drop your rate of false positives overall so you just end up with a better, more accurate classifier to begin with.

It's not that the issue of false positive goes away, it doesn't. But it's that the simple idea that merely because prevalence is low, data mining methods will utterly fail is incorrect. And the simple definition seems to support it, more accurate definition does not.

Another very frequent issue which has come up, particularly in the past several years is this idea of subject-based versus pattern-based queries. So some people have proposed limitations on data mining under the idea that you want to differentiate between inferences that are based on individuals, that is, I suspect this individual has committed a crime, I'm going to go look at data about them, versus pattern-based queries which says, I think there is an indicator of some kind of misconduct, I'm going to go look for everyone in my data set that has those characteristics.

The first, subject-based queries, is thought to be better because we have some initial suspicion. And pattern-based queries in the worst possible case are seen as some kind of dragnet; we're going to go out there and we're just going to filter and we are going to end up probably with a lot of false positives. So subject-based queries tend to be in this formulation preferred over pattern-based queries. In fact, some have gone so far as to suggest only subject-based queries should be allowed. Now, frankly, I have an enormous difficulty understanding even what this idea means in a realistic scenario. Because if you come from the technical world and you think about how we do probabilistic inference, there is no fundamental distinction whatsoever between inference based on things we observe, that is, I suspect that this individual or set of individuals is engaged in stock fraud, let’s say, securities fraud, and unobserved variables which is what might be loosely matched up with pattern-based queries. All that having initial suspicion is, is evidence to do a better job of inference overall. And so from my perspective, from the technical perspective, there is no essential difference between pattern-based and subject-based queries; it’s all inference and we use what evidence we have available to us.

Another way in which this is very difficult to understand in a technical sense, is that in a multi-stage process of doing inference, pattern-based at one stage -- if we can even formulate in an interesting way -- becomes subject-based at another. Because, for instance, if we have some process that identifies some individuals, let’s say, as having a higher probability of committing securities fraud, then suddenly we are now subject-based in the next phase of inference.

Finally, relational data -- the idea that we are making simultaneous inferences about many interconnected records, again makes this distinction between subject-based and pattern-based queries more or less disappear. Even the name, queries, I think, is showing this filter-based idea of definition of data mining versus a more accurate technical definition.

Another very frequent issue that is raised that I think has a lot of validity in one sense is a concern about having large, centralized databases. So if you have an extremely large centralized database, it is a single point of failure. And computer scientists for a variety of reasons would say it's a bad idea to have a large, centralized database. It's a single point of failure. It also means that if one institution, one agency controls that database, there's a higher probability of what's often referred to as mission creep. That is, the data set is collected for one reason and suddenly people start to say, 'Hmm, we could use it for other reasons,' which may not be strictly in keeping with the statutes behind that organization.

Now there are a variety of legal protections you could put in place to make that not happen, but there are also technical ways in which I think this critique does not fit with the way that I think many really modern applications would be done.

The first is to say real applications -- and certainly this is the case for NASD, but also the case I know for the U.S. Treasury Department -- multi-stage inference means that you don't have to have one large centralized database. In fact, there are good reasons to say that you want to distribute them among either many agencies or at least may different parts of your organization. Because the idea is you have one data set that you do one sort of analysis in, that gives you some potentially initial inferences, and then you say, 'Well, we have now a smaller set of individuals or of records that we want to go look at in more detail.' So now we go out and we've got some additional data because to do that additional examination at NASD for instance, they need to get some additional data and that's from another database. Not a problem. And in fact, a benefit because you don't have a large single point of failure, and in an organizational sense, if these data sets are distributed among different agencies, then you have at least a kind of

students and I have done recently, we published at a conference last month, about how to both understand the privacy implications of this and at least candidate algorithm for protecting the privacy of these kinds of data sets and allowing analysis, and not being able to re-identify people. But preserving privacy in this context is a new problem and a difficult one.

Finally, I think that we need to look at and think more about how to combine statistical and human inferences. Chris?

Christopher Clifton: (Speaking off microphone.)

Mr. Jensen: Excuse me. Thank you, Chris.

It’s actually whether you emailed someone. Sorry, not the number, but whether you -- the links in the data set are merely, I emailed at least five messages to this person, so it’s a very, sort of, stripped down sense of what the social network is -- or the professional organizational network in the Enron Corporation. Thank you.

Last issue is combining statistical and human inferences. We have the statistical models, we also often in many real applications have real experts. In NASD for

with something that is illegal that I need investigate. Granted, the number of individuals in his or her field of view may only be 100 or 1,000 or 10,000 if he or she is at a, you know, a large public event. But nonetheless, there is an inference process occurring there based on knowledge about what is frequently associated with something illegal. I don’t see any difference in kind between that activity and the activity that we -- the kind of inference that would happen in a database. It produces a subject-based concern based on knowledge about statistical associations. I also think that there’s an advantage and, really, we can, you know, go into extreme measures here, but I think there’s an advantage in being able to look at the statistical model which is used in data mining inference because it’s sitting there in front of you; you can examine it, it’s possible to do judicial review on that particular thing. It’s possible to audit it. In contrast, if it’s in somebody’s head exclusively, it’s much more difficult to access and understand. There are special challenges, no question, in understanding the mathematics of the probabilistic models. But in terms of that subject-based/pattern-based distinction, I don’t understand it.

thought to be terrorists. And that list is used airports and other venues to determine if a person matches that list.

And finally, there is pattern-based, or what I call event-based data mining. This is when the government does not have a suspect but has an event, either one that's already occurred or one that they're fearful will occur, and has a profile of a potential perpetrator of that past event or future event, and uses that profile and a data mining endeavor to figure out who the perpetrators might be.

Now, what are the harms of these various kinds of data mining programs? I guess if I were going to be a counter-advocate, a person who would dismiss the privacy harms of data mining, I would compare data mining to a search of a house, which is of course the classic police investigative technique. Data mining is covert. All right. People don't know it's happening most of the time. Suspects are not aware they're suspects, unlike with a house search where the police are in your bedroom, in your living room, going through your belongings. It's not physically intrusive. The physical intrusion concept is one

I tend to think for a pure analytical convenience of three categories of harms, which I think will be very intuitive. There's nothing scholarly about this. One, we might think of the harm -- the impact on individuals or what, my guess we on the panel would largely think of as privacy harms. So the individual is detained; the individual's privacy is invaded; the individual suffers inconvenience; the individual might suffer incarceration. Whatever the individual impact, you can imagine this wide range depending upon, again, the specific type of activity involved -- type of data mining involved.

The second type of harm, which is frankly the one that concerns me the most, or what I think of as efficacy harms; namely, we waste resources worrying about the wrong people, or we fail to worry about the right people. And again, this could be the result of any number of errors that creep into the system which Christopher was describing earlier. It may be the data were inaccurate; it may be they were linked to the wrong person. For example, I doubt if Ted Kennedy actually ever was on the watch list. I suspect that he had a name similar to someone on the watch list, and because we have such lousy identification in this country somebody comparing the identification with the watch list made what can only be considered to be a stupid mistake. But the issue there is not really that Ted Kennedy was detained briefly, it’s that that agent and those dollars were spent dealing with a non-threat when real threats were going unaddressed. So if in fact as Professor Jensen mentioned at the outset, we use data mining as a way of narrowing our focus of probabilistic of saying, where do we spend our scarce resources? To the extent it is in error, it is causing us to waste those scarce resources. We’re putting them in the wrong place.

And then the third type of harm is actually one which you suggested in your opening remarks, Martha, and that is what we might think of as the political or the public support risks. In other words, how many good ideas have been killed because of the public outcry about them? How often do we see hesitancy among extremely talented and well-meaning government officials because of the controversy surrounding this use of a term? I remember in the time that I served as Of Counsel to the Technology and Privacy Advisory Committee that was investigating TIA, an extremely anguished Air Force General testifying before the

know a fair amount about the way in which fraud takes place in the securities industry, we know what to look for. Then you look at terrorists, right, we have something like 60 million unique passengers every year in the United States who fly commercially. If there are, I don't know, 6,000 terrorists -- I don’t believe that, but if there were 6,000 terrorists -- we’re not talking about 1 percent of those people being terrorists, we’re not talking about 1/10th of 1 percent of those people being terrorists, we’re talking about something less than 1/100th of a percent of those people being terrorists. The chances that the data mining is going to be able to identify them are vanishingly small but the results of the enormous number of false positives are significant.

Ms. Landesberg: Thank you, Barry. Peter, let me turn to you. I know you wanted to comment a little more on the subject-based/pattern-based distinction. And then I’d like to turn us to talk a little bit about the sources of data. So why don’t you go ahead.

Mr. Swire: I’m not sure I have much more to say on subject versus pattern-based beyond what Chris Slobogin said. But I know one of the questions was, sources of data

-- some things come from open source intelligence, you get it on the internet from surfing, some things come from private databases, some from government, so can I turn to that? Is that --

Ms. Landesberg: Absolutely. Certainly. Certainly.

Mr. Swire: So I wanted to highlight some things that -- I agree with a very many of the things that have been said, but I wanted to highlight some things that haven't been said yet. And preliminary remark is, my view is as we get better information technology, there are so many wonderful uses of it and so the Homeland Security, law enforcement, everybody else in society should be doing many, many new projects of knowledge discovery, et cetera. And in some ways, this panel and this workshop is saying if there are 100 possible new uses, are there one or two or ten or whatever where maybe there should be questions asked, maybe there are certain kinds of problems. But overwhelmingly we should -- when IT gets better -- we should use it, and then we should also build it in ways that watch out for civil liberties, watch out for the other problems.

In terms of sources of data, one of the debates around data mining is how much should the government get private-sector data? The Total Information Awareness version of that was, always as much as possible. All right. Get the medical data, get the financial data; we'll get command and control of the information battle space. And there were various kinds of push-back on that.

But I want to highlight some history of reasons to be cautious about thinking everything in the private sector should go to the public sector. And this is private sector of the sort that's in your bank or in your hospital records or in your communication records. So one thing is -- to go back to the mid-1980's when this new tool called email was first coming on the horizon, and in the lobbying around what became the Electronic Communications Privacy Act, the big heavy-hitter in the room was IBM because IBM made the corporate decision that to make email grow, there had to be strong statutory protections for privacy because they didn't think that if email were open sesame to everybody including government, that people would trust email. And so IBM put a bunch of lobbying muscle in and working together with civil society

meetings, says, ‘But of course we don't want the government to get this. Of course this would be of no interest to the government. You should not even look behind that curtain and think the government would ever even be interested in any of this data. This is just about advertising and whether we’re going to have cooler content on the internet.’ And then when you talk to the companies privately -- and I’ve talked to people who are working these issues for some of the business companies that do this -- they say, ‘The elephant in the room, the thing we really know is the government could come and ask for all of this.’ And if that happens -- if it’s seen that the online advertising market is basically a front for a full government database, the whole thing is going to collapse around them. And so that debate over there, which is, how are we going to do ads on the internet, is vitally affected by, how much is the government really going to get all that? Or how much are consumers going to think the government gets all that? And so I’d, you know, to the extent you want that robust-free content on the internet, those are important reasons why that industry might do what IBM did, you know, in the 1980’s and say, ‘Government hands

Mr. Swire: Yeah, I think just a moment on the government side because when I worked at OMB we oversaw federal agency use of data under the Privacy Act. The Privacy Act has this very antiquated idea that there's A thing called an agency and everything you do inside one federal agency is okay, but if somehow it crosses a border into other agencies then you need a routine use or an exception.

Just a couple observations. One observation is there is this new agency -- happy fifth birthday Homeland Security -- that's huge, right? But that's a really big agency and everything can go everywhere in Homeland Security and we laugh at the Privacy Act because it's just one agency so there's no problem. So that doesn't seem like exactly what the 1974 Congress had in mind.

And the second thing is, if you talk about the information sharing environment and the sort of presumption of sharing in federal government, those agency boundaries don't seem to match very well. So I think that within the government side the Privacy Act doesn't match up with privacy risks today. It's been very hard to figure out what alternatives to do about that, but I think that's a big challenge on the government side.

Ms. Landesberg: Thanks very much. I’d like to turn to first Chris, and then Fred. Are there types of data mining that pose little or no risk to privacy?

Mr. Slobogin: I will answer that question yes right now, but I reserve the option of changing my answer a little bit later.

Mr. Cate: I think there are -- this -- my answer, I guess, will overlap a little bit with the panel that’s going to take place tomorrow in terms of best practices. I think one type of data mining that would avoid privacy concerns is data mining that’s justified. And what do I mean by that? Well, a good answer would take me an hour; I think the panel tomorrow will talk about it quite a bit. But the bottom line is that the government shouldn’t be able to engage in target-based, match-based, or event-driven surveillance -- data mining surveillance, unless it has good reason to suspect the people who will be pinpointed by that surveillance. What would good reason be? I think if it’s very intrusive kind of data mining that is going into medical records and personal financial records, it should require probable cause or the (inaudible) thereof. If it’s -- if the records involved are less personal, perhaps only reasonable suspicion -- those of you who know Fourth Amendment law know these phrases that I’m throwing out. The bottom line is that there should be very good justification. If there is justification, then I don’t think there is a privacy harm because I think if the government has a good reason for needing it, then privacy can be relinquished.

Another kind of data mining that I think probably does not impinge on privacy -- though, again, I might change my answer later -- is data mining that goes after purely public records, it uses only purely public records as its data source. The reason I’m hesitant in saying that is that even public records, if aggregated, can produce an awful lot of information about an individual. There’s quite a bit in the literature about the fact that one record by itself isn’t very privacy invasive, but aggregating records, even if they’re all from public sources, can take a whole lot about a person, they can be very invasive. Because after all, public records include real estate records, voting patterns, employment records, licenses, and so on. You can get an awful lot of information just from public records.

Finally, I guess I would say -- and this picks up from what Dr. Jensen was talking about earlier -- if we can anonymize data mining, if that’s technologically feasible, then I could see the kind of multi-stage process he was advocating being pretty protective of privacy. But right now I think it’s fair to say we don’t have very good anonymization techniques. So I think that’s more of a hypothetical situation where data mining would not invade privacy in any significant way.

Ms. Landesberg: Thank you. Fred?

Mr. Cate: I think the answer is yes and no, and I’m willing to stand behind that. I think our discussion -- and frankly, I think this goes back again to one of Professor Jensen’s slides -- has really highlighted the fact in some way we’re talking about three pieces of a system. We’re talking about the data, we’re talking about the analysis of the data, and then we’re talking about the consequences what’s done with the data. And so you could imagine depending upon where you are in those three that there would be types of data mining that would not raise significant privacy issues. So, for example, if you use public record data, you do very good analysis with it, and it has very minimal consequences, I think, you know, we might be able probably not to agree on this panel, but a reasonable group of people might agree that that does not raise significant privacy issues. On the other hand, even in that situation, if you used highly sensitive data, you did really bad analysis, whether or not the consequences were serious, we might think of it as nevertheless raising serious privacy issues. I think the way in which probably the public -- with whom I feel like I have the most in common here -- tends to think about these issues is focused on the consequences side. In other words, what is the consequence of this data mining? Is it, you ask me another question? Is it, you use that little swab on my suitcase? Is it, you shoot me dead in the airport? Depending upon the answer, I’m going to have a much greater sense of how this burdens me and burdens society. But I think actually the discussion’s been helpful and we should not lose sight of the fact that even if the consequences are themselves trivial, if the analysis isn’t good or the data are not accurate or are mismatched or not relevant for the purpose for which they’re being analyzed, we may still be wasting very scarce resources, we may still be distracting ourselves from the mission at hand, and so we should not be focused just on consequences when thinking about what are the implications or the impact of the data mining.

Ms. Landesberg: Thank you. Did you want to say something, Greg?

Mr. Nojeim: Not on that.

Ms. Landesberg: Okay. Very good. And I'd now like to turn to Peter again if (inaudible) -- thank you. What do we need to take into account to determine whether a data mining project is effective in identifying future terrorist or criminal activity? How do we weigh privacy risks against potential benefits in counter-terrorism research?

Mr. Swire: I'm going to take that and do something just slightly different than that. But I was at a conference last week where a military person was back from Iraq, and he was talking about IFF - Identify Friend and Foe, and I wanted to talk about how all of the data mining looks if you're in the middle of a battlefield and how it looks if you're, let's say, sitting in a nice hotel in Washington, D.C. -- which you think is not a battlefield except getting coffee today.

So IFF, the traditional Identify Friend and Foe thing is you’re a naval ship, you’re floating along in the ocean and there’s an airplane coming towards you. And at that moment it’s very relevant thing to know, is that an attacking airplane from an enemy or is that a friendly plane that’s just coming by to, you know, to visit you. And during the Cold War, for instance, there was elaborate technology to try to figure out if that was a Soviet airplane coming. And you needed to do that if you were on alert or if you were in Iraq recently, if the convoy of trucks is coming at you and it’s an attacker or else it’s not; you need to do that on a hair-trigger because if they get too close they can blow you up, and that’s a very disappointing outcome. So you really need to figure out, is this a friend or foe, and there’s high stakes and you have to do it quickly. And so if you talk to defense intelligence agency folks, if you talk to people in the middle of a battlefield, this idea of really identifying high-risk or low-risk and doing it immediately is life or death and we want our military people to have really great analysis, ability to answer quickly, all of that. And listening even to the science and technology lead early this morning, I think his model in a lot of ways was an IFF kind of model, which is, you remember that sort of increasing likelihood of bad outcomes and how likely is it -- how big is the magnitude? And it’s, like, as people come into the borders, as people do various things, we want a risk score on each moment, on each activity, on each person that’s coming to our realm. And if -- when you’re at war, that’s a highly relevant way to think about things. And we are at war in Iraq right now, we’re at war in Afghanistan right now; people’s words about how much we’re at war at home in the United States vary. More often, my experience in Homeland Security, we say, ‘We’re at war right now,’ and if you walk up to people at the mall and you say, ‘Are we at war right now here in the United States?’ they don’t tend to act and feel like they’re at war right now. So people’s war analogy varies. But here’s what I want to say. What’s different maybe about IFF when you’re at the train station at Union Station or even the airport or walking down the street in Washington, D.C.? There’s a lot of things that are different, and one is, most of the people are not on the edge of attacking you, right? So most of the time we walk down the street, it’s not that hair-trigger, is this a bomber? Is this a convoy that’s going to blow us up? Another thing is the scale of how many Soviet aircraft types or how many different categories of convoys -- that the scale is in dozens or hundreds in the war zone. There are 300 million Americans, and so the scale is really different and we don’t have anywhere near the same ability to go from -- this is Barry’s point about -- you know, from very low likelihood of harm to, out of 300 million it is. So the hair-trigger is different and the scale and magnitude’s different. And then the other thing is, there’s a lot of reasons not to do Identify Friend or Foe in civil society. So we don’t want to sort of get a complete profile of every political thing you guys have read on the internet and come up with a risk score about how likely you are to be in opposition to the government. That’s a different society than I want to live in. We don’t want to have somebody who has a jaywalking ticket or a marijuana bust 35 years ago; I’m not sure how much we want to have them treated entirely differently as they walk through society. But in a risk-score world all of those things might be in bounds.

And so if we want to have a risk-score on every moment in society as we go in the non-war zone, as we go through the United States, I think that's just an entirely different society than the sort of presumption of freedom society, presumption of openness, presumption of allowed to critique, presumption that we're not (inaudible) on First Amendment grounds. And so I think that in a lot of ways the data mining question is, how much are we in that warzone where we want our Naval ships to know if it's a bomber, and how much are we in a peaceful zone where there's a background risk but our hair-trigger is so different that lots of the same mechanisms we use in the warzone we don't use in safe, mostly peaceful society?

Mr. Cate: May I just add one comment? I think that's just an excellent analogy and it reminds me of another distinction, which is, in the Identify Friend or Foe in a military environment, it's pretty clear what the harm is you are protecting against and it's pretty clear how that harm is threatened. So, for example, in the example of the airplane approaching the ship, we only ask if it's an airplane and not a fish. We're, you know, we have a pretty focused way of applying it. In, of course, our daily lives that’s not true. I mean, you know, it’s not at all clear what the harm is we are guarding against, and it’s not at all clear what data are predictive of that harm. So again, in the example of, for example, knowing reading patterns or browsing patterns or what have you, it would be interesting to know, you know, are terrorists well known for speaking out in open public events before they engage in a terrorist act? Is that predictive in any way to know what they’re protest habits are? If not, why are we bothering collecting that data? Right? If it’s irrelevant data, why bother with it? So I think part of the concern here is that we are investing in collecting data, or we are making decisions based on data, that in fact has no probability whatever of predicting the thing it is we are worried about. And so, again, we’re left with this sense of, we are not only talking about what affects or harms the individual, but what affects or harms all of us much more broadly. And that that is -- that should be of major concern. I mean, that should be a daily concern.

Mr. Nojeim: May I just add one -- I think it’s even -- that’s a very good way to put it. And I think the problem is even bigger than that, so you could figure out what the data was, what data sets would be useful about the very small number of terrorist activities that you could point to today. You still wouldn't know what the next terrorist act would look like or what the people who would do it, what kind of data set might be relevant to them. So it's -- I think it's actually an even tougher problem than you describe because you can't fight yesterday's war.

Ms. Landesberg: Go ahead, Chris.

Mr. Slobogin: Let me play devil's advocate for a second. Let's assume that everything that was just said is correct. We still are talking about the possibility of one terrorist wiping out a major American city, and given that threat, why wouldn't it be okay to let designated agency, say, the Department of Homeland Security, have all the information it wants about anything with the caveat that that information be retained within the agency -- a small group within the agency -- and that it only be used to prevent clear terrorist acts? What's the problem?

Mr. Cate: I'm so glad you asked that. Well, first of all, nobody would ever buy into the conditions that you just said. Not a person in this room would believe that for an instant that they would only be used

intelligence out of the data. This is a little bit like U.S. News rankings of law schools, you know, they still count the volumes in the library. Like having more data is a great thing. Whereas we all know that the goal here is, can you extract useful intelligence from data fast enough to make it practically useful.

Ms. Landesberg: Okay. If I might -- thank you. And I can hear considerable skepticism about whether data mining can be effective for the purposes outlined here, but I am interested in knowing what you think it would take to determine whether a project is effective or not -- what the analysis ought to be. Anybody want to tackle that?

Mr. Cate: I'll take the easy ones and then leave my colleagues the hard ones. I think one thing we would like to see is a stated purpose in advance, and then testing against that purpose. Because one of the most common things we see -- we see it also in PhD dissertations as well -- is, you do the research, you didn't at all come out with what you thought you were so then you change what was the topic that you were researching. And therefore, before we invest public dollars intended to fight

they think it’s really, really great. And then before the deal actually happens, you have to have due diligence; you have to have other people go in and say, ‘Wait a second, don’t you realize most of these assets have already been foreclosed on? You know, maybe that’s not such a good thing to buy.’ So due diligence is the process of having smart analysis before the -- you have the enthusiasts who are trying to go forward with a new thing and then you have other people saying, ‘Wait a second. Let’s see if this’ -- and so without going through the -- reading all ten items because it’s in they’re in your chart and it’s based on a longer article. The first part is something this panel’s stressed a lot which is do we have some reason to think it’s going to improve security? Is it -- even if the project worked out, is it going to lead to some payoff? Is it going to be doing it cost effectively? Is the program going to hurt security by spreading information to the bad guys? And Dr. Jensen this morning didn’t want to tell us exactly where Fraud Alley was; I bet a lot of us were sitting here thinking, ‘Hey, I wonder if that’s’ -- I don’t know, I thought of Miami, I thought of some parts of New Jersey. You know, I don’t -- and maybe you all thought of different places. I used to live in New Jersey; I'm not against New Jersey, but I knew some things there. You know, so -- but he didn't want to tell us because then the next fraudster won't set up in Fraud Alley. They'll eschew those three-digit codes and they'll set up somewhere else. And so there's all this cat and mouse kind of thing. So in my article and in the ten-point list, there's some attempt to try to say, 'What are the problems?' One thing I'll highlight is number six, "Do fairness and anti-discrimination concerns kick in?" Here's a current example - there's a hearing recently in the House on a data mining thing done by the insurance agency. It turns out, in the insurance business, I can do a better job predicting your insurance risk based on your credit score. And so the question has been, is it a good idea/bad idea for credit scores to be used for your car insurance. In the hearing there was discussion that there's a correlation between race and credit scores. So if this started to be used, certain racial groups would pay more on average for car insurance. So you'd have a benefit, which is maybe more accurate person-by-person decisions about how much to charge for insurance; we'd have a more efficient insurance market. And you’d have a sort of fairness question of, if predictably this is going to raise premiums for certain racial groups, is it okay or not? And there was -- it was pretty heated debated in the House Committee -- Financial Services Committee about what to do on this. But it illustrates something far away from terrorism where you get results from data and then you have to work through, ‘Okay, what are we going to do with this?’ and a due diligence process is at least one way to try to head at that.

Mr. Nojeim: (Inaudible). Can I chime in here? It’s really to ask Barry a question ‘cause it goes to, what do you do with data that might be relevant, might be useful? Say it’s not Fred’s example where the NSA is getting hundreds of millions of bits of data, say it’s this example - you’ve got a different NSA, it’s the smart NSA, it’s the focused NSA, it’s the targeted NSA. It has three terrorist phone numbers abroad, that’s all. It’s been watching them, it knows that they’re bad guys, and it also can collect information about who those terrorists call and who call them. And all three of these terrorists talk to somebody in the United States; what should the cops do with that information? Should they show up on that guy’s doorstep and say, ‘What are you up to?’ Or, what do they do with it? That’s --

Mr. Steinhardt: Well, now I’m glad you asked that question. You know, actually I think that’s a relatively easy question, right? Which is to say you have your -- you made it a little more complicated by mixing in, you know, Foreign Intelligence Surveillance Act or domestic eavesdropping laws, but, you know, but basically that question is one we know the answer to, right? Which is that, you know, the government has a lawful right to obtain the numbers that were called by a particular individual, or they have a pen register or whatever it is. And they know who that individual calls; can they do some follow up investigation of the individuals that were called? I think the answer to that is, you know, usually is yes. That’s not exactly data mining, right? I mean, you know, except in the sort of broadest sense of the word. You know, it’s so the way that, you know, that law enforcement follows leads generally. And, I mean, I’m a little, you know, I’m a little less troubled with that than the notion that we are going to intercept everybody’s telephone calls and try to make fairly attenuated connections between individuals based on their pattern of calling, as opposed to this fairly, you know, discreet set of facts that you described.

Ms. Landesberg: Okay. If I might just ask -- Greg and Barry, are there -- we've gotten a good record I think from this workshop already, but are there more specific harms that either of the two of you would like to address before we turn it to the audience for questions?

Mr. Nojeim: I wanted to talk a little bit about commercial data for just a sec. And increasingly the government is using commercial data in its data mining activities, and there's nothing inherently evil about commercial data as opposed to data that's been generated by the government.

But I did want to mention a couple of concerns about it because I think that using commercial data should be done very cautiously. The first is that the data is collected for a particular commercial purpose, and it might be the case that in pursuing that commercial purpose, some problems with the data would be ones that you wouldn't want to expend the necessary resources to correct. And there might be accuracy issues within the data -- just say for example, its' credit data -- it might be the case that for you to fix all the problems in your credit database -- and I saw one estimate that 70 percent of credit reports have an inaccuracy -- but to fix all that it might be prohibitively expensive, and therefore it might be an appropriate model for you to sit back and wait a little bit until a person contacts you and complains about that data and then fix it after you receive that input that there's a particular inaccuracy.

That model, that kind of data might not be appropriate to be using to predict who might be a terrorist, or to match with other data about terrorism because it doesn't have the accuracy level that you would need for that data to be effective.

And the second thing I wanted to stress besides this use issue, was that data in the private sector isn't subject to Privacy Act restrictions. And Peter has outlined some of the problems with the Privacy Act, and they're substantial, but it does provide some protection for people who are subjects of that data. And so one of the issues -- one of the Privacy Act protections is -- goes to accuracy, errant inference, and those kinds of things are things that I think an agency relying on commercial data would need to account for.

Ms. Landesberg: Thank you. Barry?

Mr. Steinhardt: Yeah. I actually wanted to pick up on something that Peter spoke about, which is, sort of, what is the harm or what is the consequence of living in a society where we are all risk-scored? Which is a society that we are increasingly moving toward, and I think that the consequences of that are fairly profound in people's daily lives. It's not simply that there's the risk that you're going to be arrested or interrogated, it's the risk that you are not, for example, going to be able to obtain a mortgage or a bank loan, that you're not going to be able to get a job. I mean, all those things are increasingly becoming real for people. You know, if you go in now to a bank to open a new account, your name is checked against a government list to determine whether or not you, you know, you might be one who is engaged in, you know, criminal activities, terrorists, et cetera.

One of the things we know about that list is that it, you know, essentially it's a new form of risk scoring, right, or a risk not that you -- that, you know, that you're going to be a deadbeat or something, but rather that

program everybody involved in terrorism detection and everybody in Congress, and, you know, immediately have them investigating each other instead of investigating real problems.

Ms. Schiller: My name is Jennifer Schiller; I work for Under Secretary Cohen as his Privacy Liaison. So we are focused exclusively on research development, testing, and evaluation activities.

Mr. Steinhardt: Could you just get a little closer to the microphone, please.

Ms. Schiller: Sure. None of our current programs meet the Congressional definition of data mining, but we do want to move forward with that vein of research, you know, looking at our long-term planning. And as we enter the testing and evaluation phase of developing new technologies, we do need to use real data to test the technology before we can, in good faith, transition it to an operational component that would then go and use it. So my first question is, what factors should we consider in evaluating the impact of privacy in that type of research where we’re not making determinations about individuals, we’re testing the operation of a technology prior to transitioning it to an operational unit?

And the second question I have is that we seem to have two broad categories for data analysis and data mining activities. The first would be what I just described where we're developing a new technology that would eventually transfer to a customer; the second would be where we're looking at data in a social science type of way -- and we do have one of our social science researchers here, I hope she'll ask some questions at some point -- but we're looking at a broad set of data, for example, on terrorist events or on terrorist groups and trying to draw inferences from that data. An increase in rhetoric might be a signifying factor before an event. And Professor Cate, you said, 'Why collect the data if it's not relevant?' Well, we don't know if it's relevant until we go in, collect the data and do the research. So my second question would be, how can you handle collection of data in a research environment where you're not sure what data is relevant?

Mr. Cate: Well, let me say I think you raise a phenomenally important issue, and that is the need for data on which to do research. And it's an issue on which, to be perfectly frank, Congress has been completely tone-deaf, not to mention ignorant, and that is, it is not appropriate to use the same types of privacy protections for data that will be acted upon, as opposed to data that’s being used in a research environment. Having said that, because of the policy of law in this area and the fact that most of the law in this area is, I mean, has no real restraining effect, there’s no possibility to make a promise like, we’re going to have this data but not act on it, because there’s no legal requirement that would bind you to that. You would have to enter into a, I guess, a contract with the American people that said, this is what we’re going to do. So in some ways, I hate to say it given that I think Congress has been a major source of the problem, but I think they’re also going to be an essential part of the solution, which is to create a category of data analysis or data mining or data aggregation for research that has to meet certain conditions and would be subject to certain oversight and so forth. One of the things we haven’t talked about on this panel at all just for lack of time and because I think other panels will, are the procedural and process protections that can diminish the potential harm caused by data mining. But I think, you know, I mean, you would know those as well as any of us, and that what’s needed is a way to get those -- if you will lock those into place so that once a program is declared a research program, and I would continue to disagree with Peter on this. I think when you spend of hundreds of billions of federal dollars you better know what your goal is before you start, rather than the, let’s hope we find it does something once we’ve spent this money. So you say, ‘This is research. Period.’ And then you are under that protective regime; I think that’s going to have to be what the solution is going to ultimately look like, and it’s going to mean going to Congress.

Mr. Swire: I actually think we know quite a bit about research from the medical side of things. So for medical -- I worked in HIPAA and the research parts of the HIPAA Medical Privacy Rule, and so there’s things in HIPAA about limited data sets, about data use agreements, about what kind of audit and oversight there’s supposed to be before the research is approved. In most circumstances it goes to an IRB, an institutional review board. I’m not sure how much all of that exists yet in research at DHS, but there’s been a lot of decades of work on research from the medical side, and that at least gives you some institutional things to look at as you're trying to figure out.

And then in terms of how you do it legally, a statute would be better, but I think DHS could say, 'When we do research we're going to do it under this sort of IRB medical approach when you're working with real people's real data. And you could say, 'We're making a promise to follow the following guidelines,' and then you could say, 'And we're going to have our IG come in on a regular basis, or GAL or whatever,' to make sure it's being followed, and that would -- well, even without a statute, give a pretty decent institutional basis for the world to believe you're actually doing it.

Ms. Hahn: Thank you. Again, my name is Katherine Hahn with SAS. I appreciate your comments. But I want to go back to the question that I asked Professor Jensen this morning. You all have talked a lot about data gathering, data collection, problem articulation; is the privacy concern raised by the statistical model or is it raised by the human intervention where you can't test the bias and the assumptions that people are bringing to bear?

And as a follow-up to that, what is it about data mining as a statistical modeling activity, as a research methodology, that merits heightened privacy scrutiny over other types of research methodologies? Thank you.

Mr. Steinhardt: Let me take a crack at that if I can, in a couple ways. First, I think that the answer to your question is one I said earlier, it's both yes and no. The, you know, the issue here is not only whether or not the collection of the data and the analysis of the data represents a problem, but also whether or not the automated data presents a problem and how that data is used by individuals. I do think it's important to recognize, though, that when it comes to the use of data by the government, the government is not in the same position, really, as for example, the security dealers examples that we were given earlier this morning, where the security dealer, you know, the NSD has this sort of unique ability to collect data about the individuals that it governs. It already has that data; it can use that data to mine into it. But the government on the other hand, we found this, for example, in the area of airline passenger profiling. The government didn't in fact have very much in the data that it wanted to use to mine, whether it was an automated process or a process driven by personnel. That was really the great -- that has always been the great debate about the various versions of airline passenger profiling -- how much data would the airline industry need to collect? Often data that it does now collect from passengers in order to give the government the ability to in some way mine that data or make use of that data. That's really, I think, the important thing to look at here as we're looking at government data mining, government use of data, which is, where does the government collect that data from? How much data does it need to collect? And then ask the question of what it's going to do with it, but to recognize that generally speaking, there is the necessity to go out and collect data in all -- and usually it's the private-sector collecting that data for the government -- but to go out and collect data that is not now collected and not now analyzed.

Ms. Landesberg: Okay. Did you want to respond? Sure.

Mr. Nojeim: I just wanted to say a couple of things. One is that I don't think that you can neatly divide up the two functions between the data mining and the consequences that follow from the data mining. You don't do the data mining in the first place unless you're looking at what to do with the data that you get, with the results that you get. So I just don't think that you can isolate that kind of technical activity from the consequences because that's the whole purpose, is to decide -- to make decisions about people. And the second thing is that people who are involved in the actual data mining activity can build in some of the protections that we've been talking about here. For example, audit trails -- I mean, there is -- you do want to have that capability built into the system that you're coming up with so that you can find out whether the data was misused and whether it was appropriately used.

Ms. Landesberg: Thank you. And if I could ask those of you who are waiting to ask questions to just be very concise in the question so we can get you an answer and then we'll adjourn for lunch when Dr. Jensen's had his chance weigh in.

Mr. Schneiderman: I'm Ben Schneiderman from the University of Maryland. I'm troubled by the narrowness of the definition of data mining, exemplified maybe by Christopher Slobogin's question of what's wrong with collecting data, and the last few answers did get to the question of the socio-technical system that's imbedded in the cost of collecting it as opposed to the benefits that might come from other things, the distraction that's brought by it. But then the -- I guess it's Greg last, you know, comment about the end-game, also what happens once you get it. So what are the -- how do we expand the definition of the systemic view that will give us a socio-technical analysis that will give, for example, citizens whose privacy was violated, recourse and compensation, which is not part of the TSA's current No Fly List. If you're prevented from flying, you don't get compensation. So if you put in the true costs to all the parties that are harmed, you have a better chance of understanding what the payoffs and the negatives are more clearly. So I'm looking for -- the questions about what are the broader aspects that you see to the socio-technical system?

Mr. Slobogin: Well, this is really follow-up on the last answer, but it seems to me we should get away from using the word data mining if that's your major concern.

If you wanted to define data mining the way Professor Jensen did, fine. Then we've got problems with the data collection and with who gets the results of the statistical model and what's done with it? You can assign labels to those different stages of the process. I have to admit, I'm more concerned about the data collection, who gets to see the results of the data -- of the statistical modeling, and what's done with it. Those are my major concerns. The actual technical aspect of statistical model is not a major concern of mine. I think most people apply the word data mining to all those stages; it might be better to break them out into the three, four, five stages, and then focus in on the legal and social consequences of those stages.

Mr. Steinhardt: Can I -- let me take a crack at that, too. I think that's an important question, which is what recourse do you have if you are harmed? One of the real drawbacks to the current, you know, airline passenger profiling system and the watch lists, et cetera, is that there is no real recourse for those people who find themselves on the list by mistake, find themselves harmed. There is a sort of a, you know, kind of a Kafkaesque system of going through the Department of Homeland Security puts you on the list in the first place, and I can take you off and you'd never know if in fact you're off and how you got on and all those things. But, you know, there's a fairly simple solution to that problem, which is to say that we're going to have an independent body out there, whether it's a judge or -- but some other mutual arbiter, right, who is going to take a look at the data and is going to say, you know, that is or is not someone that we -- to be worried about, -- if not, we're taking them off the list and we're ordering all the people that maintain that list to take the person off the list, or to the extent to which they have a name that is, for example, is the same as somebody who should be on the list. We're going to create a white list or put this person on the white list so, you know, you're not -- we have none of that. And I'm not sure, monetary compensation might be nice, but I'm not sure that that's the solution. I think the solution is we've got to have processes in place that allow people to appeal to an independent arbiter, decisions that are being made about them. But the first thing we need people to do is to indicate to that person that, yes, you are on the list or yes you are affected.

Mr. Schneiderman: My point is that true costs are only visible when you have the larger socio-technical context. Thank you.

Ms. Landesberg: Thank you. And, sir?

Mr. Lempert: Yeah. Rick Lempert, I'm with DHS. Two very quick points. One, I was very happy to hear mentioned the commercial issues at the end. And as you're speaking personally, in some ways I'm more concerned about the commercial invasions than I am about government. We saw that Admiral Poindexter's plan -- what happened politically -- there are likely to be political limits on what the government can do. And we see with the IRS data that the government can be very, very protective of confidential information even when it could be used for other governmental purposes. Doesn't mean there aren't governmental concerns.

The other issue is -- in thinking about this, and I agree the costs are very important -- the question is, what is the alternative in the non-data mined universe as we think about it? So, for example, if -- I may not like any broad-based surveillance techniques at the border, perhaps, but if I had the choice between these lists of

here.

But, you know, I -- one of the things we haven't really talked a lot about this morning -- talked at all about this morning -- that I want to sort of emphasize is that all the security, you know, is a zero-sum game. We only have so much money here to spend on our security, and what we have not talked about this morning are some of the alternatives to all of this data analysis -- whatever -- however we want to characterize it and all this list making and all this. I mean, we know, for example, that, you know, that often the most effective weapons that we have against terrorism -- terrorist attacks, are physical security. I mean, we know, for example, that hardening the cockpit doors after 9/11 made a tremendous difference. You cannot replicate what happened on 9/11; you can't get into the cockpit now. Can't use the plane as a missile. We know, for example -- take the example -- the U.K. example of the terrorists who were the London Underground -- tried to set off a bomb, turned out they weren't very good at it, but the -- you know, but there was all this sort of, you know, hyped up surveillance equipment around them, all this -- all the video cameras and whatnot; that didn't stop them. But they went up to the airport in Scotland -- now they decided that they were going to drive a car into the passenger area -- passenger terminal -- what stopped them? The concrete barrier out front of the passenger -- out front of that terminal; same concrete barriers you can find all over Washington in front of government buildings.

So we have to ask ourselves here, before we continue to spend all this money and all these resources and risk our liberties and our privacy on these systems, is this really the most effective way for us to be fighting the so-called War on Terrorism?

Ms. Landesberg: Thanks very much. And, Dr. Jensen, I think we should allow you to have the last question here and then we'll adjourn for lunch.

Mr. Jensen: So it may surprise some of you that I think we almost totally agree, but I want to do a find/replace on all of your comments, which is I think possible because of the court reporter -- which is, to do a find/replace and replace data mining with data collection. So here's the question -- the question is, like you, I am against government incompetence, I am against violation of civil liberties -- I'm an ACLU member, by the way, dues paid --

Unknown Male: We'll check our lists later.

Mr. Jensen: -- I'm against government power grabs, I'm against prejudice, and I'm against authoritarianism. But would those harms stop or significantly lessen if we completely gave up data analysis but kept doing data collection just like we're doing now?

Mr. Swire: I've gotten to be in a lot of different privacy-related meetings over the last bunch of years, and the previous question was -- but I'm going to summarize it, maybe a little unfairly -- I work in the government, the real problem is in the corporate sector. And when I talk to the corporate folks, they all say, 'We're good. I know all of our people are really good; I worry about the government.' We have a data mining statistical person saying, 'Data mining's good but those collection people are nuts.' And then you'd go talk to the police at the collection point and say, 'Look, we collect things. It's a world of cheap sensors, we have to get the data we can get, but it's what they do with it once they collect that's the real problem.' And I'll just -- having -- I'll just observe as a sociological phenomenon that people tend to --

there's aversion to taxes, right? Don't tax me, don't tax, you know, you -- tax that person behind the tree -- and the aversion in privacy is, the part I'm in, we really don't want to have these intrusive rules stopping what's important to do, but it's those folks over there, that's what you have to watch.

Mr. Jensen: Yeah. But governance is about making choices, okay, and ascribing causality correctly. So I do think it really matters.

Mr. Swire: Oh -- it matters to do a sensible analysis on security; what's the tradeoff between physical security and intelligence gathering ahead of time? It makes sense to figure out what are the real risks on collection versus analysis versus actionable. All of that is a logical part, but I -- just observing that the previous two questions ago is the socio-technical system and the political system in the broad view, it turns out we don't have these neat compartments where we can say, 'This is a risk-free zone and the problems are over here.' You look at each part of the system and you keep having the meetings because it turns out these pieces are interrelated.

Mr. Cate: I'd just like to be clear, I'm not against authoritarianism. Your question kind of makes me wonder how badly we've done for the past hour-and-a-half up here, because I don't think there was any suggestion that data mining is inherently bad or data mining is inherently -- should not be used or should be avoided in favor of some other technique. It's that data mining, like all of the other tools we use in fighting terrorism, should be subjected to the same type of scrutiny. And that the more we can break it down into its constituent parts, the more we can be clear about the data analysis and the data collection and aggregation, the data mining tools that are used and the consequences of what's done with those, the more frank we can be in that type of analysis, the better the results are likely to be. And I think in response to the prior comment as well, there are many instances where data mining is by far the preferred tool. I mean, it is the equivalent of putting up the concrete barriers, in some instances. It makes perfect sense, it's cost-effective, and if done well may have little negative impact on individuals. I think part of the problem is that the dialogue about this wide range of data analysis activities

the decision could be a manual process, it could be an automated process, it can be anything. But if you are going to analyze -- I mean, I'm an engineer; I like to take data, test any system whether it's human or otherwise, and produce results that come out of the area of debate. They are scientific engineering results. This is one way to display information about any binary decision process. It came out of designing communications devices. So the basic way was, if you're sending a 1 or a 0, how is your receiver detecting that 1 or 0? It can be generalized to any binary type decision, so is something good/bad, is it, et cetera?

And just to set the stage to understand what this graph is telling you, if someone asked me to design a system that will always detect whatever it is you're trying to detect, I can very easily do that. If you see that point up there at the 1,1 position, the way that you read that graph on the bottom -- the x-axis -- that's the probability of a false alarm. The y-axis is the probability of a correct detection. So if I just always say, 'Yup, that's it; that's what I'm looking for,' then I'm at that 1,1 point. I will never miss any of the things

I’m trying to detect; however, I will always make a false detection any time that it’s not what I’m trying to detect. So that’s one extreme of the detection space. Well, let’s say you say, ‘Well, no, I don’t want to do that; I want to never make a false alarm.’ Well, that’s easy too, down at the 0,0 point. I always say it’s not what I’m looking for, so I’ll never make a false alarm. Well, that’s also kind of useless.

Now, if I just flip a coin every time, I can be along that dotted line between those two points. So, if I’m going to analyze a system, I want to determine, am I on that line -- in other words, I’m doing any better than chance -- am I below that line, where you see the red graph. Because believe me, you can spend a lot of money to design a very complicated system and be somewhere in that red space where you’re doing worse than chance. Or am I somewhere in the green area where I’m doing better than chance?

The other little point to point out there, we’re talking about probability, so if probability of 1 means that absolutely always will take place with certainty -- that means for all time, for all data -- there are very few things that are that certain. Probability 0 means it will never take place for all times, all data; and that's also very difficult in any complex situation to ever do. So, in reality, the best you can do is approach those two extremes. You will never be absolutely sure that you're detecting everything and having absolutely no false alarms. So, just think about that.

The other little way to look at data mining -- and I think one way that that term came about -- it's like refining ore. If I take raw ore and I'm trying to refine it, each stage I go through the ore that I come out with has more of what I'm looking for, more gold or what have you. But that means I'm leaving little flecks of gold behind as I refine it, because again, you can never be absolutely certain.

So having said that, the order is next that David will be giving you an -- we're intending this to sort of be a tutorial of what the technology can and cannot do. We will want to basically -- if we're going to have a debate about the policy, we as technologists would at least like to present you with what that technology can do, and present you with things like this that if we are allowed to test systems, there is a way of testing them, but we need to have real data to do the test for it to be meaningful. Okay.

Mr. Jensen: Thank you, John. So, I’ve already been at this podium for a long time and you’re probably tired of hearing from me, so I thought I’d just give a fairly short presentation.

And one of the basic ideas that I want to make sure that we get across is that we’re trying to compare performance of different ways of doing a task. In many cases, I think in most cases, when we’re dealing with national security issues, with domestic security, Homeland Security, there really isn’t a question of if you will attempt to do some task; the question is, what is the approach you’re going to take to it? And so I think it’s important to compare alternatives and to say, for instance, think about alternative data mining systems, think about alternatives that do not use data mining. And to say, how well do each of those alternatives work? I think in many cases, we imagine that doing nothing is the status quo, when actually there is some existing system. We’re just comfortable with it because we’ve had it for a long time rather than we're comfortable with it because we know its error characteristics and we know them to be good. So this graphic, just from the NASD securities fraud example, we ended up comparing to expert-derived rules which, frankly, NASD had never really thought of as a system before. And we found out that we could, by just analyzing data, come up with as good a set of screening rules as they had, and by combining with that we could produce a better system.

But importantly, I think it's important to think broadly when you're saying, 'We're validating results.' Really what we're doing is evaluating a system and evaluating it in a relative sense of how it compares to existing systems and to other prospective systems.

I also think, as I've said, that it's very important to perform evaluation in context, in the context of the data that you might be gathering or already have, in the context of the decision-making, not view it in isolation. So validation often needs to take into account this kind of larger context, larger institutional context, larger process context. And then it's not just the technical characteristics of the system that matter; it's where it fits. So a good example of this is this question of screening, that if you are doing initial screening for some disease, having a high false positive rate may not be a problem as long as you are putting that in a context where you follow up that first test with a more accurate test, even though it may be more expensive.

Final point which you haven't heard me make before but which is an unusual one -- one I wanted to make sure that we talked about -- is that there's a long history of development of technology of these algorithms, and the history has been benefited by the fact that we've, as a community, have developed algorithms and released them publicly. We don't make an algorithm and say, 'Oh, no, I'm going to keep it secret and just tell you how well it does.' We actually release code, release detailed descriptions of these things in the technical community so that other people can build them themselves and try them out and understand their characteristics.

One of the things we found over the course of 20 or 30 years of research in this area is that it is very frequent to come up with a new technique and only years later -- sometimes ten or fifteen years later -- find out some places that it breaks down that we didn't understand.

And that’s only possible because the algorithms are public.

One of my nightmare scenarios is that someday I will be called in to some windowless room some place and not asked about my own activities, but said -- asked to repair some data mining algorithm that I’m going to find out was used for a long period and no one in the technical community was really told -- or not many people, at least -- and I’ll say, ‘Well, don’t you know, we know that systems like this fail in some horrible way, but because the algorithm wasn’t out there, we didn’t have the ability to raise those issues, talk about them and identify it.’

So there’s a real benefit to using what I term here, public algorithm -- publicly released, described algorithms because that encourages wide scrutiny from the technical community and you can remedy errors quickly.

Now, there are examples of this in the non -- outside of data mining -- Linux and other open-source software operating systems are widely thought to be secure, partially because -- or more secure than they might be otherwise -- partially because errors can be identified quickly and easily, and fixed.

The internet protocol, the basic protocol underlying the internet is a public protocol, and errors and problems with it have been fixed over the years and it’s been improved. The public key encryption is a wonderful example of this, as well. A known public algorithm that is used to encrypt data -- and just because it’s public doesn’t mean it doesn’t work and doesn’t work for very, you know, important, secure applications.

So these are nice examples of public algorithms that we have in other domains; I would argue that we need them in the area of data mining. These algorithms should be public even if the data that they’re operating on or their conclusions, their models are not public. The algorithm can be public even though the data and models are not. And that’s it.

Mr. Coggeshall: Thanks, David. My name is Steve Coggeshall, and I work at a company called ID Analytics. We’re essentially an identity intelligence provider, and we do analytics around very large identity networks -- connectivity of individuals, primarily for identity risk and for authentication -- remote authentication, data breach analysis, things like that. My background: I’m a scientist; I came from academia; I worked for ten years in a national lab doing fusion research; and the last 15 years working in industry doing research. I've spent the last 20 of my years building data mining models in many industries, both in the public and private sector, for governments and for business in many applied contexts.

We're going to talk very briefly about -- I'm going to give you my quick tutorial on what a data mining model is, tell you a little bit about how to build a model, and then a subject that's pertinent in this aspect -- in Homeland Security is, what do you do if you don't have known bads? If we don't have a lot of examples of known terrorists, how do we build and evaluate a model? And then next is how to evaluate a model when you don't have those. And then, finally, I will just talk a little bit about what are the benefits of using models.

So first of all, what is a data mining model? In its simplest form, a model is an algorithm; it's a functional formula that takes inputs and provides an output. And that's what that little box in the center here is; it's just this mathematical functional formula with output y and a set of inputs x. The inputs are typically characteristics about a person or event, and we can denote the string of characteristics in some notation, x-1, x-2, x-3; and the output y is the likelihood that it’s something of interest -- could be the probability that it’s fraud, the probability that it’s bad credit, the probability that a consumer is going to buy a product, the probability that it’s a terrorist. It’s something of interest. And then the model itself is just a mathematical formula. And here’s a very simple example that’s actually used frequently in practice; it’s just a linear combination, a weighted combination, a weight a-1 times the characteristic x-1, plus a weight a-2, times a characteristic x-2, and so on. And when you get done with that, the y, if it’s scaled properly, is a score and the score can be -- represent a probability. The a’s -- the parameter’s a’s are a set of constants that are learned from data, and that’s the -- what we talk about training a model, is showing the model data and then statistically finding the best set of a’s, the best set of parameters that matches your data and does the best value of -- the best prediction for that set of data.

So how do we build a data mining model? Well, first of all, we use lots of data, in general, to build a model. And, in general, the more the better. A data record, you can think of it looking like this, it's just a vector, it's just a string of characteristics, x-1, x-2, x-3, and so on, followed by the outcome -- whether or not he's a terrorist or whether or not this person's bought a product or whether or not he went bad or this credit or whether or not it's a fraud. That record, that string of information is a single data record, and we build and use many millions of data records when we're building models, typically. And again, I want to point out, it's very important to clean the data; if the data is not cleaned and scaled and represented correctly, then you just have, you know, garbage in, garbage out. The model will not train well and will not -- you'll never be able to build a successful model unless you're very careful about how you clean your data.

So what does our data look like now? It's just this arrangement of these many, perhaps millions, of data records. And then what we do is we split them into two sets, a training data set and a testing data set. We use the training data, along with some statistical and machine learning algorithms, a whole field of science has evolved in the past 15, 20 years around this -- around very efficient and very well-built algorithms to do this best functional fit, to find those best parameters, a, in this functional relationship; y is a function of the inputs and those parameters. Once you've built your model, now you put it in place and you have to evaluate how well it works. So you do that by using this testing data. The testing data is holdout data that the model has never seen before, and you evaluate how well the model performs on a whole new set of data that it's never seen before. And you can statistically look at how well your predictions match the real outcomes.

So that's the usual methodology in practice of building models. And this whole process is called supervised training, because you know the outputs, so you're supervising -- your model is learning in a supervised way. But sometimes you don't know the output, so what do you do there? And I think that's frequently the case in Homeland Security; we don't have a lot of examples of terrorists, for example. So if you don't know the outcome, you don't know who's good or bad, so now our data records look like this. It’s just the string of characteristics -- could be their age and their weight and their height or their -- how many times they’ve flown, whatever -- but you don’t know whether or not they’re terrorists. You don’t have a y. So in this case, you can be successful in building unsupervised models. Unsupervised models approach the problem differently: rather than finding the patterns of the relationships between x and y, it just looks in the x space, and the characteristics space, and it looks for things that are unusual -- anomalies, outliers. So I drew a picture of that here. Let’s say, for example, I only have two characteristics to worry about, and in consumer modeling it might be age and income, that’s a very frequently used set of characteristics that describe a lot about how people behave. So this might be their age down here, and this might be their income here; and every person has an age and an income; they have those two numbers. So every person is a point in this space. And so we put all your points in there and you see how the data naturally groups. This is what David was talking about earlier - clustering; this is an example of clustering analysis. I see how my data

there are flags that go off sometimes, maybe it’s Social Security Number matching, maybe it’s past record matching, but there are rules that fire -- that cause certain people to be looked at more closely. So those rules in the -- so there is some kind of an existing process today; we can call that a control process. Events go in, and you go into some set of rules, and most of the people come out as not interesting, and that’s good. But there will be a small subset of people that are flagged as maybe bad, and those go into almost always a human investigation process. Some physical human has to look at this and make a decision about whether or not this a true bad or not. So -- and again, usually as a result of the investigation they are okay, and those are the false positives; those are the ones that were flagged by the model -- by the rules -- but turned out to be okay. And then these are the true bads here, the ones that turn out to be really bads. So in any process, you can start -- you can instruct some important metrics that measure the efficacy of the process. And here are two that I wrote down here; these are two that are very commonly used. The false positive rate, we’ve heard a lot about that; that’s the ratio of how many false -- it’s how many false positives you have. It’s how many people you bothered that you shouldn’t have, so it’s the number of false positives compared to some baseline. And there are different ways of doing the baseline; one way is by dividing it by the number of true bads; another popular way is by dividing it by all the number you investigated. It doesn’t really matter which one you use, it just needs some metric that measures your false positives.

And then another metric is your bad rate, your bad detection rate. How many real-bads do I find, divided by how many I have to investigate to find those bads. So those are two very objective metrics that you can use to measure whatever existing processes you have. Once you set that up and put your metrics in place and see how well you’re doing today, the next thing you do is you put your test process in place. In this case what you do is you send some of your data through a model; you build your unsupervised model or whatever, for whatever methodology. You’ve got a model -- a candidate model -- and you want to evaluate, how does that work? In particular, how well does that work compare to what I’m doing today? So again, you put some of your records through here, and the model will

records go through that. You might find out of 100, I find 60 false positives and 40 bads, so your false positive rate might be better; it's lower. Whereas your bad detection rate is 40 percent; it's twice as high. A bad detection rate twice as high means that I can find twice as many bads with the same amount of work, or I can find the same number of bads with only half the effort or half the intrusion. And that's the -- one of the keys -- one of the key uses of using a data mining model, is that I can reduce the effort and reduce the intrusion.

So I do think data mining in this case is finding a needle in the haystack. I think it's more similar to the -- I think a real problem here is that the needle looks a lot like a piece of hay. And I think that's our real fundamental problem here. And it's a highly non-trivial problem; this is a hard problem. It involves lots of data, lots of -- it will involve a lot of clever data and coding, a lot of understanding of the domain, and domain-expert knowledge. But I do think it's a problem that can be improved a lot with data mining.

And another point is, humans should never be taken out of the loop in this. The point of data mining in

them alongside of existing processes in a double-blind test, measured the efficacy, and it turns out, in those and in every other case I've done, the data mining models provide a lot of benefit.

So, in summary, data mining models work; they're in wide use. I think this discussion is less about whether or not a data mining model can be effective, but it should be more about how would one do it and how would one protect privacy? And there's a lot of questions around that. But I have high confidence that data mining processes would help and would be better than existing expert-driven processes today.

You can build supervised or unsupervised models, and there are ways of testing each one in either case. Another important point is, data mining models can discover patterns that your experts have never even thought of looking for. And that's very frequent; what happens in these data mining model processes, is finding relationships that you never even thought of looking for. And again, the point of this is to minimize the review population, either therefore allowing you to reduce your effort in your investigations, but probably more importantly, reducing intrusiveness, really focusing your investigations where they will have the greatest benefit. That’s it.

Mr. Dennis: Okay. Hi, I’m Steve Dennis, I’m from the Homeland Security Advanced Research Projects Agency, and we typically are trying to work on new revolutionary ideas, and so, certainly, an area of consideration would be privacy protection technology since I don’t think we’ve solved that problem at all, hence we’re having this meeting.

I remember back about fifteen years ago when I first heard the term data mining and it was told to me by a group of mathematicians: ‘Well, you label your program data mining if you want to get funding.’ And today, I think the opposite is true. If you’d like to be de-funded, you might label your program data mining.

So, you know, what are the essential elements that are required if we’re going to validate data mining models? Certainly, this panel is all about the scientific investigation of what works. What the Science and Technology Directorate of DHS is about is discovering what can work. It’s not necessarily about deploying that technology immediately, but understanding what our options are. If faced with a particular situation and we need to do more data analysis, what are the techniques that are on the shelf and immediately available for use? We need to have a cadre of those. And hence, Jennifer Schiller's remarks earlier about, you know, S&T would like to in the future, start moving back into this domain of data analysis.

As we do this work, these are some important performance considerations. And the first, I think, has been discussed already, which is, you know, can the data be prepared? Is the right data available in the appropriate form? And there's a lot of work that goes on to understand what's happening in a data set. And if you think you just plug data sets in to data mining algorithms and they magically start producing something, they don't. It takes a lot of considered preparation in order to make those algorithms start to produce and produce in a productive way. So what are some of the considerations there? We're worried about the speed of a process; can it actually keep up with a data rate or with a large volume of data if it's needed? What is the accuracy of that mechanism? Is there error being introduced as you start to process data? We're worried about storage overhead; if the original data took a terabyte and the process data takes two more terabytes, this is an issue. We’re worried about portability; if I get a solution to one problem in one domain for $100 million, does it take me another $100 million to solve a problem in another domain? And certainly that goes to scalability, but cost is also a factor here throughout each one of these steps that I mention.

If you have structured data, those are the simple cases. If you have unstructured data, there’s a lot more work that has to be done. And so there’s been a history over the last 20 years of trying to automate this process of information extraction. How do we make data available in the right forms and what kinds of errors are introduced during that process? And as you get this drift, error propagation can happen throughout the system so you have to worry about the performance of each one of these components, not just an individual piece. So, if I’m extracting information, I’d be worried about the linguistic features; you know, how good am I at getting verbs that might imply events? How good am I at getting proper nouns that imply people, places, and other things? These are all factors, right? And you’re starting to get a sense for the complexity of this kind of research; it’s not very simple. There’s a manpower factor; if I create a process that’s heavily knowledge-engineering oriented, would a customer ever have a knowledge-engineering branch that can make that work? You know, that might be a consideration at the early stages that make you say, ‘This model’s never going to work because it’s just too manpower intensive.’ And I might trade off the efficiencies that I gain at the back-end having to replace those people that I save at the back-end with people in the front-end. So I might gain nothing by doing that. And certainly, accuracy and speed again.

Once you have a well-prepared data set and you fully understand the error characteristics of that data, then you can move into a pattern-matching function. And whether that happens to be learning patterns or not, the results of those kinds of algorithms are generally either binary, you know, where they either tell you they found it or they didn’t. Sometimes you get rank-ordered lists that say, ‘Okay, here are the top 20 choices that match that pattern;’ or you could even wind up with weights and tables, and it’s even more and more complex to understand the performance of such systems.

Then, I move on to an area that's called policy filtering. You know, once I'm able to understand patterns and I'm able to have data sets that can be processed, I worry about, can I automate the application of policies over the top of the use of that data? So, if I have a privacy policy, can that actually be codified and made part of the system? Would the policy folks be in a position to write their policies, not in English, but in some sort of coded form? And then I worry about things like the receiver operator curve performance of that; you know, we saw the graphs before, whether it's precision and recall or false alarm and missed detection; there are many ways to talk about it. And then I worry about leakage. You know, how much of this information that I have is leaking over the boundary at any one point in time. And I can do an in-depth analysis of that and start to look at tradeoffs, and even start to look at what it might mean to compare human performance in that case, you know, if a human's making a decision to a machine performance, and that's really important. If you can get inner-annotator agreement, and what that means is, if humans can agree on a task, then it probably can be automated. If you have a group of humans who can't agree on a task, then perhaps it can never be automated and you should save your money. So you worry about data retention, audit, traceability, and the policies, the overall effectiveness. And you start to see that you can trace now the use of a policy and how it’s impacting systems and performance.

Above that level -- and these are all architectural issues -- is a system-level concern. And that is, is this system usable? We heard earlier today that, you know, if you generate a number of leads and they all lead to a lot of overtime and there’s no productive result, then that’s not a good thing. And you need to have mission metrics that tell you that a system can actually perform in an efficient manner. Efficiency also goes to cost. If I deploy a very large data mining system, and it takes $100 million a year to keep it going, perhaps I’ve gained nothing.

And we talked about traceability; that’s all throughout the system, not just in the audit of the policy, but, you know, who touched data when and where is a very important factor of each one of these systems. And then information assurance. Have you authenticated the users of the system? Are they in the proper role? And all these factors have to be considered. So now you can understand why the performance evaluation, the test and engineering might be very expensive. And, as a matter of fact, getting involved in some of these efforts, sometimes the data collection, the evaluation, and the preparation for the research can outweigh the budget that you have for the research. So, you know, we're trying to make this as easy as possible, but there are many layers to consider.

One idea that may help us all -- and I think it was alluded to earlier -- is the development of some sort of common research and development framework that allows us to reuse components. If I'm really good at preparing data, do I really have to engineer an entire system around my effort in order to understand the system effects? If there were such a common framework, you know, that was freely available and software could be traded around the community with the kinds of visibility that we heard about earlier, perhaps that would help us make progress. And it would also save us a lot of time at DHS. We are approached by many vendors with many ideas and many universities with many ideas, and often they come in selling us a brain in a box, but no one has done any sort of evaluation that tells us what the true performance of that system is. You can spend two hours unwrapping a package to find out that it’s the same as a system from 1960. So you have to be very careful there. And this kind of framework may help us understand better performance.

We talked about data collection, and it’s very, very important for research. What you’d like to have is a sustainable data set that represents a hard problem that can last for 20 years. And dare I say, I don’t think there’s anybody in this room who will approve of a data set containing private information that could be retained by the research community for 20 years in order to do repeatable experiments. So that sort of works against our normal R&D methodology at that point.

So, folks will often say, ‘Why don’t you use synthetic data sets?’ Well, you can use synthetic data sets early in the process to help debug, perhaps to understand the implications for scalability, but if you have a language problem and perhaps a name-matching problem -- for some of the lists that we heard about before -- you can't really work that name-matching problem with unreal or made-up names. You really need a set of real names and real situations so that you can model the actual condition and understand how to improve the performance of such algorithms.

If you do have real data sets, you can do risk mitigation for those data sets, although it becomes highly complex, if a series of questions comes up, when you start to consider those. Thank God for the Enron data set and things like that that just happen to be out there that we can use among this node, but, you know, those data sets don’t necessarily represent real problems either. So we have issues there.

I think about the problem of doing data mining and doing this kind of research, you know, there’s a chicken and an egg, and if you’re approaching it from the egg, you know, often the questions are, you know, ‘What sort of chicken is this going to be?’ and, ‘What color are his feathers?’ and, you know, ‘Is it going to have a mole?’ You know, I don’t know, you know, until the egg hatches. If you come at it from the chicken end, you know, basically they want to know what color is the egg going to be, and is it going to be speckled or brown and will it contain double yolks? We don’t know.

So at the end of the day, what happens typically -- at least now, before we move through some more policy changes -- is that both wind up fried, and, you know, never -- we’re never allowed to find out what happens.

If we had a common infrastructure that would enable system-level investigations, it would allow us to do more -- get more return on the investment for our research. And we talked about making that code freely available and Reusable, and it might also lead then to common evaluation methodology. And I don’t think we have a really good evaluation methodology that centers around privacy. And I think those kind of tradeoff studies would be very interesting, if we were allowed to do them.

I wanted to leave you with this thought, and it’s a very, very simple cartoon that talks about the points at which we would feel friction doing this kind of research. If you’re talking about doing just a normal data mining system with fixed data sets, you know, that’s in the top left-hand corner there with the pattern-based algorithms running over some set of data that everybody agreed was okay. But if you look at the intradepartmental situation -

- as was mentioned before -- you know, DHS is a collection of a lot of operational elements and each of them have their own rules and their own lawyers, and so it’s not so easy to just put things together and make them happen. There are lots of discussions around that. So clearly, there’s a policy filtering need at that edge of the graph, you know, as we go across the department. Each component, each data set has its rules and charters and implications.

If we look at cross-departmental access, we’re hoping that, you know, it might be possible to somehow design a common analytic space that would allow the government to make use of what it knows in the right circumstances. But that would imply a lot more policy filtering, a lot more comfort with implementing our systems in different ways. And so, just a thought for you to think about as we continue the panel.

Mr. Hoyt: Okay. To raise some questions for the panel, since we have -- and we have time for questions from the audience as well, but one of the areas that I’ve seen in the literature is this perception that somehow data mining can take all types of data and magically combine it and come out with useful results. And I have my own biases about that, but I'd like to open that up for our panel of experts here to at least comment on that perception that I can throw everything in there and somehow it'll make, you know, a gourmet meal out of all the hash I've thrown in.

Mr. Jensen: So why don't I start. One of the things that I often try to tell people, the things I try to tell people is that, if you compare data mining to aircraft design, that we're just out of the Wright brothers' stage. We tend to think of this as a high-performance technology, and certainly, many technologists want to say, hey, we've got these really wonderful algorithms. But the truth of the matter, I think, is that we're actually very early in our understanding of this technology and development of new technologies. And one of the consequences of that is that there are many types of data that we don't know how to deal with effectively, at least nowhere near as effectively as somebody could whose an expert, who is a human who can just look at it and interpret data.

One good example of that is that, until about ten years ago, we didn't actually have methods that could look at interconnected records and make use of those interconnections. So we know if you look at -- if you think about how a doctor does medical diagnosis, you come in with a fever or something into an emergency room, and the doctor starts thinking, ‘Okay, maybe it’s communicable.’ So he or she asks you about who you have come into contact with, if your family members have this disease, et cetera. They also know that, ‘Well, maybe it’s genetic -- maybe they are genetic components to this somehow.’ So, ‘Gee, has your mother or father or children ever suffered from this?’ It also might be occupational, you know; maybe it has to do with where you work. So we think about, naturally, all of these relations, but we didn’t have methods that could think like that, that could develop models that looked at those kinds of relations until quite recently.

And another big area is people say, we have tons and tons of data, but what they really mean is we have tons and tons of textual reports that each of us could sit down and read and extract information from, but as several people pointed out, we don’t actually have good automated methods that can reliably look at lots of unstructured text and pull out the sort of meaning that is anywhere close to the meaning that a person can.

So we are -- I think we are fairly limited to numeric and symbolic data that is connected up in interconnected records, and a very limited kind of extraction from large text documents.

Mr. Coggeshall: I'd just like to add to that I think one of the biggest challenges in data mining these days is unstructured data. We have a huge proliferation - explosion of data, primarily in the unstructured space -- text, audio, voice, image, video -- and it's getting more and more important for all these different needs to be able to use that kind of information. And it's an extremely complex problem, a lot of research going on -- a lot of successful research -- all those categories I mentioned are being used today in data mining, but we're -- I would say we're just at the infancy of building -- of learning how to efficiently encode that unstructured data into numerical representation for algorithms to operate on. So it's still -- we're still at the very beginning stages of that. And it's a hard problem. And I think for Homeland Security it's going to be one of the key areas; it's going to be very difficult.

Mr. Dennis: I too think that we are at early days on data mining and there has been a lot -- even though there has been a lot of investment, we don’t fully understand how to approach this problem-solving in a pro-forma way. If you think about throwing together lots and lots of data, it just reminds me of the Wal-Mart example of putting the beer next to the diapers, you know, sort of where the young father comes in to get the diapers because the wife said, ‘Go get diapers.’ And the correlation happened that, you know, the young father also picks up a six-pack of beer to go with those diapers on the way home. Those kinds of associations are found because you start to look at these patterns in large data sets and you get discoveries that you wouldn’t otherwise get if you didn’t do that.

But very early in the process, it’s sort of a triage stage where you’re looking for those sort of things to happen and to make those discoveries. But very, very soon after you spend time in that discovery phase, you start to worry about what is the contribution of data to my observation or to my inference? And you try to trim the data and get rid of all the things that don't matter, so I think there is a value to doing some of that kind of investigation in a triage mode.

Another thing that I don't think has been mentioned yet is that even though you discover these patterns and they work today, you know, they may not work next month. And so there is certainly a lack of understanding of model lifetime, and models drift, and so, you know, how often you have to reinvestigate these kinds of relationships is probably unknown for a lot of data sets.

Mr. Jensen: Another comment, if I could. When I worked for Congress at the Office of Technology Assessment, we did a study which ended up saying, don't collect more data because that's not necessary and it's not going to help. And I think it's an interesting example of where analysis -- not data analysis -- but careful analysis of the overall task can actually tell you, you don't want to do this. We were asked by Congress to look at the question of whether additional data on wire transfers in the United States -- large money transfers -- would assist in the detection of money laundering -- criminal money laundering by large organized crime groups, particularly. And 18 months of work, of really talking to large numbers of experts, and really understanding the analytical tools that are available at the time, and the conclusion after this intensive 18-month study was, no, don’t collect data on wire transfers because, one, the amount of information it contains is so weak that it’s very unlikely to yield anything; and secondly, it was going to increase an order of magnitude by ten times the amount of data that a small treasury agency had to sift through, and that would have ended up actually swamping them. And they said, ‘Please don’t give it to us.’ But also, we had good, you know, quantitative reasons to say that. So in the end, we came back to Congress and said, ‘No, not a good idea to collect additional data.’ And I think that’s the kind of technical conclusion you can often come to, is that, don’t add more hay to the haystack if you’re looking for a needle.

Mr. Hoyt: Another topic that we get approached with is, given the difficulty of dealing with personal data, can we deal with synthetic data? And at least I’ll give my bias and let the panelists kick in. But my bias is that synthetic data is useful at almost the -- I’ll call

practical problems by using synthetic data.

Mr. Jensen: I guess I would ask whether you would like your -- the pharmaceuticals that you might take in ten years to be tested on simulated humans or on real humans. Yes, there are certain kinds of things you might be able to figure out by looking at simulations of the human metabolism, but it’s not what you want for anything except the very early part of the process.

Mr. Hoyt: There’s another class which actually maybe falls out of the range that this forum is really interested in, but DHS does have problem sets where we care about patterns that have nothing to do with personally identifiable information. We care about pandemics, both for people and for animals and food crops. Obviously, there are other parts of the government that we partner with in that, but in that case we need no personal information. In fact, that’s noise as far as we’re concerned.

The other -- several of our panelists have worked on systems for industry and for other agencies. There is at least some perception that I’m getting out there that people think that we’re just sort of starting off from ground zero, and I think it's been touched on several times in this conference, that there are existing processes that are in place. And I'm assuming if I'm industry funding something, I have a profit motive and I want a system that does it better than the existing system.

Mr. Coggeshall: In my experience, I've built a lot of data mining algorithms for a couple decades now in lots of different industries, installed them all over the world -- and I've done some really tough problems, some pretty basic ones, but some pretty tough problems, too -- and I have never done one where I haven't been able to outperform an existing system. It doesn't mean it's tremendous, but we've been able to beat whatever the existing process is. I think that's just -- and I -- I have high confidence that that could be done here in this case also.

Mr. Hoyt: Having said that, could I open it up for questions from the audience? And if you'd please come up to the microphone.

Ms. Schiller: Hi, Jennifer Schiller again, Science and Technology Directorate. And Steve Coggeshall said in your presentation that when you are building a data mining model, the more data you have the better. But in the government, we have Fair Information Practice Principles that require us to use the minimum possible amount of data, so there seems to be a real tension between our legal and privacy policies and the technical requirements of building a data mining model that will work. So I was wondering if you could speak to that for the whole panel a little bit more.

Mr. Coggeshall: Sure. This is a common problem and we face it in industry all the time, too. And a very good example is in credit scoring. There are certain fields that we know are useful that we cannot use for regulatory reasons in credit scoring. So the reason we know they're useful is because we tried them and they work, but then you go through an iteration process with legal systems and policy, and for a variety of reasons you're forced to remove those pieces of information. And so, philosophically, the more data you have, the more varied and disparate data you have, the more -- the better your models will perform. I mean, obviously, if I have a certain universe of data and my model works to a certain level, if I add more data I never get worse. If you do things right, you only get better. So, at some point, when you add more data, you don't get any better, and that's what scientists are all about -- applied model builders are looking for where that tradeoff is between adding more data and not -- and the amount of performance you gain is not -- is disproportionate to the amount of effort it takes. So, in this case, I think it would -- it might make sense to have an environment where you can, on a trial basis, try lots of data. But, in the end, you will find that a subset of that is what's needed, and then from there on that's what you need certainly for implementation of the model when the data that's streaming into the model only looks at that subset of the data.

Mr. Dennis: I think it's important to have access to a lot of data in order to figure out what the minimum set is. It's not possible to discover that minimum set without some experiments, and so if we have to guess what the minimum set is at the outset, we're likely to spend a lot of money as we continually enlarge that circle until something starts working.

Mr. Jensen: I think one of the things we need -- and I think this was referred to by one of the previous

expert way that presents very efficiently information to an algorithm? And that's how one frequently finds patterns that you've never discovered before.

So it's an iterative process: you do the best you can in your first step; almost always you'll beat the existing process, and then you continue to get better from that and inventing new variables, getting feedback from the experts, trying new auxiliary data sets, and it's just a continuous improvement process.

Ms. Szarfman: The more data you analyze, the better you understand the data. And the technology's evolving very quickly, so what we cannot do today we will be able to do tomorrow. Then we are limiting the amount of data you'd analyze, which will restrict you in ways that you cannot foresee. The more you analyze, the better you can find the outliers; you will understand if your methodologies and appropriate method can find things in higher dimensions, like, in my case -- in our case at the FDA, is to find drug interactions in specific (inaudible) population that may be at risk because they are elderly or they are preemies or restricting. But you have a different problem, you are trying to find unexpected things. It's a very difficult problem because if you knew where to find it, you would not be having this meeting. Then we are -- which data you should analyze is also, you know, you don't know which data you should analyze. Then you have a difficult task.

Mr. Coggeshall: That's all completely correct. This is not an easy problem; this is a hard problem. My point of view is algorithms can be built that will improve upon the existing processes, but by no means is this an easy problem

Ms. Szarfman: No.

Mr. Coggeshall: It's very hard.

Ms. Szarfman: It's very, very hard.

Mr. Coggeshall: Yeah.

Ms. Szarfman: Our problem is also very hard because it was considered impossible in the past that, you know, you could not get anything out of secondary data that you don't have in hypotheses. You need to have an hypothesis, then you set up the data to analyze the problem, but the primary reason I don't understand the data because you have this -- a way of understanding the data. Then you are looking for interactions with alcohol, but if

Ms. Gregory: Hi. I’m Michelle Gregory from the Pacific Northwest National Lab. And I’m a researcher in data mining, and this whole session is about how you can validate data mining models and results, but I’ve been trying to rethink the problem into -- instead of, as you mentioned, overlaying policy on top of the data that you have for the analyses, can you include the policy at the data-collection phase? In other words, how much can you glean from data and patterns can you find that are useful without having all the data available? So in it’s most simplistic form -- from the talk this morning -- it would be, anonymize all the names and maybe places and locations; you find the connections that are interesting, then you reveal them under certain policy conditions. So I just wanted to hear your comments on that.

Mr. Dennis: I think you can certainly distribute the policy control throughout the entire process. And so as you’re ingesting data, certainly as you’re doing information extraction, there’s a lot of anonymization that can be done and I know of programs that do that now for HIPAA applications. It’s absolutely true that you can distribute it at any stage in the game, but it’s important

unfortunate. But there's another example, which is that we do medical research right now on new drugs, on new treatments, and there's a very small number of people -- maybe 100 or 200, you know, who will volunteer for some clinical trial, and they'll participate in it and they get compensated or they get free treatment or something else. And then, as a society, we benefit from that. We say that's a great thing; they did that and now we know whether this drug works. And in the same kind of way, you don't need to implement an enormous data-collection procedure in order to find out if there's some signal there, some statistical regularity that you can catch. You could go to extremely focused, careful -- maybe random samples, but very small amounts of data, find out whether it -- there's something there to model, and then you could say we've got some reason to believe it would be good to do the more general data collection. That's a different kind of idea of clinical trials, but I think it's essential and something that's really outside of the kind of research or complete program that we do right now. We need to have some trial runs at constructing systems.

Mr. Coggeshall: I'd just like to make a comment on anonymization. There’s some tremendous work going on that field -- in both industry, IBM and academia. One needs to always be careful about this. For example, some of the work we do in our company, requires us to do fuzzy matching across multi-dimensional spaces with name, address, Social Security Number, phone, date of birth, things like that that are -- that if you anonymize first before you try to do fuzzy matching -- multi-dimensional fuzzy matching, not just one at a time, but all simultaneously -- it can destroy some of the connectivity, so you need to be very cognizant about where and when you do your anonymization.

Ms. Schiller: That was actually my question - is what are the implications of using anonymized data to validate a model, as opposed to -- early in this day you talked about synthetic data could be useful, but taking real data, anonymizing it and then using it in the validation process -- how would that impact your ability to -- from an S&T perspective to say to a customer, ‘This works; I’m confident it works?’

Mr. Dennis: I think anonymization poses several challenges for the kinds of research that we want to do, and that is if we were looking at the name match application and you anonymize away the name, then that's putting us out of business. If you -- I would ask, though, if you consider lack of human access to the underlying content as an anonymization. You know, if the algorithm gives a chance to see the personal data and yet the human doesn't, is that an effective anonymization? I often wonder about this because machines doesn't have malintent and machines do what they're told, you know, they're not spying on their neighbors. But people, you know, are often accused of that, especially if they work in government. So it seems like to me one way to mask and anonymize the data is to allow for the machine to have access but not for the human.

Mr. Coggeshall: I just have to say, that's an excellent point and that's something we do. All our data in our company is encrypted, so your socials, names, and addresses are encrypted, but they are unencrypted in the algorithm phase itself for the matching. But when we look at the data -- unless we're doing case studies or something like that where we have to explicitly unencrypt the fields -- we interact with it in a completely encrypted format.

CERTIFICATE OF TRANSCRIBER

I, Renee Braun, hereby certify that I am the transcriber who transcribed the audio recording provided by Alderson Reporting Company to the best of my ability and reduced to typewriting the indicated portions of provided audio in this matter.

____

Transcriber