Department of Homeland Security Meeting

Implementing Privacy Protections in

Government Data Mining

International Ballroom East

Hilton, Washington

July 25, 2008

Attendance: Peter Sand, Michael A. Aisenberg, Brian Tretick, Daniel Weitzner, K. Waterman, Toby Milgrom Levin, Fred H. Cate, Thomas Oscherwitz, Barry Steinhardt, Peter Swire.

Transcribed by:

Dana A. Cohen

For

Alderson Reporting

[Convened at 8:38 a.m.]

Mr. Peter Sand: Thank you very much. Brian, do you want to start us off?

STATEMENT OF BRIAN TRETICK, EXECUTIVE DIRECTOR FOR PRIVACY ADVISORY SERVICES, ERNST & YOUNG

Mr. Brian Tretick: Good morning, again. I'm going to read from some prepared remarks, so I make sure I cover the points that we had discussed in preparing for this panel. I'm going to take off my glasses so I can see my paper. I can't see you --

[Laughter.]

Mr. Brian Tretick: -- if anybody will interrupt, you need to speak up.

So, again, I want to continue the conversation with a discussion of basically this thing called the "audit" and I'm using air quotes around it, and a lower-cased A for audit, because it means a lot of different things to a lot of different people, and most of those people don't use it -- the term audit -- with any kind of precision.

So, let's focus on the definition from the American Institute of Certified Public Accountants, I don't

know that we have any other members of that august body in this room, but me, but the AICPA publishes -- one of the things it does, it publishes professional standards for auditing and accounting. AU 100 is the foundation for the financial audit, the legacy from which audits over other subject matter, non-financial subject matter, are based.

So, I'm going to paraphrase from its opening to just describe, again, the philosophy behind an audit, okay? The objective of an ordinary audit, and this is of financial statements, by the independent auditor, is the expression of an opinion -- so the objective is the expression of an opinion, in this case of the fairness with which the financial statements presents in material respects, the position of the company financial, its operations, its cash flow.

So, again, let me summarize that. The opinion -- the objective of the audit is the expression of an opinion about the reasonableness in material ways, okay? So, that's what we want to get at. So, it's not absolutism, and it's not on the immaterial side. So, again, a company may have had some errors in its tallying of its financials, but as long as those aren't

significant and material, we look for issues of the fairness of the representation of the financial position of the company.

So, that's a philosophical aspect of it. So, the standards -- so the auditor looks at the subject matter and applies something called generally accepted auditing standards. You know, this term "generally accepted," we're going to hear over and over again.

And really, what that means is, a bunch of auditors got together, and floated a standard, we know the standard process in a number of areas, and then at some point it gets pushed out for scrutiny -- these accounting practices, and auditing standards have been around for decades, and evolved over those decades. But -- so, that's where we get the "generally accepted" part.

The auditor has the responsibility to plan and perform the audit, to obtain that reasonable level of assurance, about whether -- in this case the financial statements are free and material, and the statement, whether they're caused by error or fraud, it doesn't matter.

So, the important point here is that the audit is

designed and executed, again, to obtain reasonable assurance -- not absolute assurance -- typically in the financial community, and this is what's acceptable, focused on material misstatements. This reasonable assurance, then, is the basis of an auditor's opinion. Okay, that's a simplification -- there are hundreds of pages in this AU 100, just in the audit guide itself -- but there are AU series documents, or AT series documents, more than you'll ever want to imagine.

Traditional audit standards have been augmented more recently in the last couple of decades, with standards for more general attestations. But they carry forward the principles of independence, objectivity, completeness from the auditing standards.

Under ATIS Standards -- and these are general non-financial audit standards. An auditor may express an opinion about things in other subject matters, such as measurable characteristics. One of the things we might do is the physical size of a factory, or real estate or something, or the number of bricks in a brickyard, you know, there's some physical aspects.

Historical events, such as the price I offered to

customers over certain dates, might be something somebody wants to know. Or even compliance with laws and regulations, and more importantly for this subject, internal controls over some sort of process.

So, the last two -- compliance with laws, regulations and other obligations, and internal controls are the most important for us, here.

So, let's come back to the topic, audit over an organization engaged in data mining, might be designed to provide insurance, okay, independently and objectively, so that the so-called internal controls over that process -- data mining, data sharing, whatever -- whatever you want to call it. The controls are in place, designed appropriately, so not missing, and they're pretty good -- they're good at the -- they're the right level of rigor. Okay, in place, designed appropriately, and operating effectively over time, okay? So, that's the controls part: in place, designed appropriately, and operating effectively over time.

The other component to this that we'll want to consider here is that it complies with the laws, regulations, or other obligations that are imposed upon it.

The rules, for example, data usage scenario that you've discussed. So, these -- this is the basis that we want to put forth -- an internal controls audit, and a compliance-type audit could apply here.

So, who audits is a second question that we're asked to discuss. The answer depends on the purpose of the audit, the nature of the assurance that is required. In some cases, the management -- the organization itself -- may use its own organization, perhaps its internal audit function, or its own business team, its own group to conduct the audit. And others, you may like to even have or be imposed an external auditor, or Inspector General, for example, over that process to conduct the audit.

The assurance that you can derive from the audit often depends on the selection of the auditor. Internal self-checking may be appropriate when, for example, the audience is internal, okay? So -- and you can rely upon that auditor.

An external auditor, an independent external auditor, would be most appropriate for developing assurance when it needs to be conveyed to the outside parties. Again, the reason you bring in an auditor is to develop a

position of assurance, reasonable assurance over something. But you have to be able to believe the messenger. Again, my mother -- my mother loves me, she's not objective. So, if you're looking for a character reference for me, don't ask her. But, you'll want to -- if you really need to rely upon it, you'll want to get character references from external parties. So, it's the same kind of argument in this case.

So, independent auditor is, again, a common condition of SEC-mandated financial audits, and even the FISMA audits the government agencies go through.

So, the AICPA, the American Institute of Certified Public Accountants, got together in 2001, and by 2003 came out of a conference room and published what has become known as generally accepted privacy principles, there's a reference on my last slide, to where that is located. This body of work translated the ATIS Standards for the subject of controls over the use of personal information, so privacy audits can be done, and basically addressed the common principles for privacy and data protection that were prevailing, and would have included U.S. Federal government rule -- common U.S. Federal government rules at the time.

It's like the garlic in the Ragu sauce, "It's in there." It's baked into that process.

So, basically, the result of examination procedures, planned and conducted by the independent auditor, in this case, the audit, would be an opinion over what the organization claimed it did, its policies, its privacy policies and practices, and over the organization's ability to meet those criteria. So, the generally accepted privacy principles gave us a foundation, a common foundation, in which we as an industry could conduct objective, relevant, complete, and measurable privacy audits.

More recently, as an example, the Transportation Security Administration established its own requirements for security and privacy, within the commercial Registered Traveler program, and requires participating organizations to submit reports of their independent auditors, according to those requirements on an annual basis -- the reports convey the opinion of the auditor, that it, the auditor, has developed the position of reasonable assurance that personal information was used, stored and retained, in conformity with the program's requirements. So, the TSA

took a bunch of the NIST requirements, and then added things that were specific to the Registered Traveler Program, to create the security and privacy standards that it has to meet.

One of the Registered Traveler Programs, in particular, Verified Identity Pass, which operates the CLEAR program, you see it at Reagan National and Dulles as an example, they also had their system audited against the AICPA standards, so they can issue a public report to their members. So, they're not only providing assurance to the TSA as required by assurance to their members.

Finally, what gets done in audit, remember, it's up to the auditor, normally, to determine the procedures to be executed, and the auditor is the one who issues the opinion, so we're responsible for coming up with the right procedures to develop that position of reasonable assurance, because it's our signature at the bottom of the report, and our reputation, and in many cases, our liability.

The generally accepted privacy principles offers an auditor illustrative controls, things to look at. Additional considerations, plus, again I want to refer to

you to the thousands of pages of audit and ATIS Standards that come along in the backpack of any auditor who graces your door, so it's not a crap shoot, in any way.

I think I want to flip on -- so basically to cut this short, observation, inquiry, inspection and test -- these are all common test procedures, these are the tools of the auditor, as well. I want to test the -- whether the -- the rule that the organization has set up, I want to test whether there's a controlling place, it's designed appropriately, and it's operating effectively over time, and this is what gives me the basis for my reasonable assurance.

The aspects of generally accepted privacy principles include things like the management of privacy and security policies, procedures, programs and controls, so the administrative part of it. The things affecting the individual; the notice, the choice and consent, the redress programs and other processes that are focused on the individual, the subject of the data, if that's applicable in this case.

Controls over the life cycle of the personal information, that's the collection, use, retention and

disclosure I mentioned earlier; disclosure to other parties involved.

Security, integrity, quality and other similar requirements, and then finally, the monitoring of controls, because an auditor can't have confidence that controls are in place and operating effectively just by looking at themselves, they need to also see that the organization has an ongoing program of monitoring the effectiveness of their own controls. Without that, it's difficult for me to form a basis of reasonable assurance over your operations.

So, in most organizations, use of personal information is complex, and not only that, especially in the case of the subject of the hands, you know, these arrangements for data sharing, data mining, are going to cross organizational borders internally, and externally. So, it complicates this thing called the audit.

There are legacy approaches to auditing such complex environments. First thing we do is we take a system like this and we decompose it, okay? Where the requirements, then, map out all of my requirements on my claims that I want to make, or the rules of the law, regulations or other agreements, and say, "Where are they

applicable?" Okay? Because, they're not uniformly applicable in all entities and all processes, okay? So, the first part is the decomposition to the processes that handle personal information, and the supporting, as we'll hear later, technologies and manual procedures, okay? So, some of that requirements that we have at the highest level aren't applicable in some of these areas, so we want to rule those out, and that's okay to do.

The decomposition shown here includes multiple parties under different controls and governance, a common audit strategy may involve either the auditor going to each one of these boxes on this chart, and conducting examination procedures. Or as another option, I can look at the third party on the right side, and say, "I'm not going to come and audit you, but you need to provide me a reasonable basis of trust and confidence in yourself."

And this is, again, for example, looking back at the TSA Registered Traveler Program, the TSA is not going out and auditing everybody. But they're asking for, again, under a certain program, an audit report from these organizations, so that you can play within its cloud, okay?

This is the same kind of basis of assurance that

we have in the financial community where, if I'm a corporation, I have outside service providers -- payroll providers, other people processing my financial information -- I don't want to send my auditors everywhere to audit every one, and these service providers don't want to be audited by all of their customers.

So there's a construct, in this case, a SAS 70 Reports -- Statement of Auditing Standards number 70, again, they go up into the hundreds, so beware -- but a SAS 70 Report allows an organization to be audited once, and that the audit report is an auditor-to-auditor report, it's really big and ugly, but what it tells any other people who come knocking on the door, it says, "Listen, we just got -- we just got audited, here are the results," and an auditor, again, if that's done under generally accepted auditing standards, I can rely on somebody else's work in that case, okay? So, we look at the reliance on other parties' work.

Okay, so I think what I'd like to do is kind of cut to the chase. When looking -- when you are looking at the philosophy, the objectives, the methodology and practices for privacy audits in data mining, the Department

and the other parties here should take into consideration the legacy of this financial audit; more importantly, the interpretation by the AICPA in the generally accepted privacy principles.

Where audits are needed, the Department can look at the recent example of the Transportation Security Administration as a, again, a worked example in this case, although, again, it's not data mining.

And with that, I'd like to thank you and turn it back over to Peter.

[Applause.]

Mr. Peter Sand: Thank you very much, Brian.

STATEMENT OF DANIEL WEITZNER, CO-DIRECTOR, COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY (CSAIL) DECENTRALIZED INFORMATION GROUP, MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Mr. Daniel Weitzner: Hi, everyone.

Thanks very much to Pete and Toby and Martha for giving us the opportunity to come and talk with you. As I said, my name is Danny Weitzner, and I run a research group at MIT, where we look, in general, at advanced web technologies and the public policy implications of many of

those technologies.

So, we have a project that we've been working on for about the last two and a half, three years, called Information Accountability, and our interest, really, is to understand how to design large-scale information systems -- whether they're data mining systems, or the web, or large enterprise systems -- how to design these very large-scale systems in a way that they have a property that we call information accountability.

What we mean by information accountability is that it's possible to keep track of how information is used in an information system, and then where there are misuses, to detect that misuse, and hold accountable either the individual or the institution, as appropriate, for that misuse.

I want to talk, quickly, about some of our larger motivations for doing this work, and I'm -- we're going to come back to the specific application to data mining context, but I want to talk about the bigger picture.

As we've looked at, and been involved with, the design of much of the next-generation web technology, what we've seen -- both in the technology and in the way people

use it -- is that we're really living in an increasingly transparent environment. It's not a value judgment, it's just a statement of how we all interact with personal information, and information, in general.

The picture you have here -- I'm going to tell you just a little story to illustrate the kinds of problems we're interested in solving -- the picture you have there is a diagram of a part of a social network -- it's actually the social network at MIT, and some students of mine collected some information from that social network, and set out to test the hypothesis whether they could discover in this social network who was gay, and who wasn't, based on a very small amount of information. They happened to get this data from Facebook, and we can talk separately about how that worked.

[Laughter.]

Mr. Daniel Weitzner: But, the important thing is that, simply by looking at information that people routinely disclose in the course of their interaction in that part of the web, in that kind of system, it ended up that people were revealing, really, quite a large amount of information, in some cases information they didn't quite

intend.

And this is kind of a -- I think just the -- sort of one indication that, in some respects, we're really all data miners, we really all -- have the ability in one way or the other to learn quite a lot about people out on the web. You'll know, certainly, about the -- the more controversial data mining programs, I think you probably know that the person who designed this logo was asked to retire, at some point, but I think that what we do know is that there really are many socially important uses -- whether it's -- of data mining -- whether it's for national security purposes, or for learning more about our health status, for getting more personalized treatment of our health conditions, for sharing information. We're all in a situation where we're going to be making more and more use and sharing information more and more about ourselves.

And so the question that we've come to is to ask, once lots of information is disclosed and available about people, what can we do about privacy protection? Do we have to just say, "Forget it, get over it, it's out there, you have no more privacy," or do we have a way to reintroduce some set of controls on how that information

can be used?

So, again, our research is really focused on the question of -- once information is flowing rather freely in any given system, can we still maintain control over how that's used, and have some rules about what are acceptable and unacceptable limits on personal information.

In the context of the technical projects that we're developing, looking at data mining models, we have really three very particular goals, and my colleague, K. Waterman, will show you some of how we've implemented them.

First and foremost, what we're trying to do is -- while we want to enable data mining to go on, we want to prevent mission creep where possible, we want to make sure that once information is collected and analyzed, that it's used for the purposes that were intended, that it follows whatever the rules are about how that information can be used.

At the same time, we want to enable more seamless sharing of information. Our sense -- and I think you probably know this far better than we do -- is that in many cases information sharing and analysis that actually would be permissible under the rules, doesn't happen because it's

hard to tell whether that analysis is actually allowed, or whether that usage is actually allowed or not. So, we want to put a wall around mission creep, but enable the kind of sharing that's actually possible.

The way that we approach this is to implement a system-wide and very fine-grained audit, both of what information is accessed by whom, and most importantly, how that information is used, and whether those uses conform to rules.

Just very quickly, this is research that's going on with a team that we've put together that includes colleagues at MIT, Computer Sciences Program at Yale and RPI, and I'll -- we're going to try to stay away from the technical details here, but what I do want to do is, I'm going to turn the floor over to K. Waterman, whose going to talk about one particular application that we've developed to illustrate how this technology works. And then I'll be back.

Ms. K. Waterman: Great. So, we like to work with hypotheticals that are based in fact, so in this case, we're looking at a public health investigation. The CDC is investigating a case of extra drug-resistant tuberculosis

which is a real problem and, actually, came up after we started working on this scenario, but basically it's a form of tuberculosis that's incredibly hard to treat, and so it is important for them to quickly figure out who the patient has been in contact with, and begin to work with those people to figure out if they also have the disease, or to do what they can to stop the spread.

In our hypothetical, so we could get to data mining, our patient is in a coma, so they can't ask that person any question, and they have to use data mining to try to figure out who this person has been in contact with.

They use traditional data mining techniques, and one of them is to go to the phone company, find out who this person has been calling, and then look at those phone records to try to figure out how close the association is.

Later, a person who was one of those people whose phone records was checked, goes to get a phone company to send out a service guy, a repair guy, and the phone company refuses.

So, this -- that's our hypothetical, we're trying to figure out whether that refusal was permissible.

So, the way that we do that is, right now we're

working with the user interface where, on the left-hand side -- I'm not sure I can -- so, on the left-hand side we have a little sidebar, and basically what it does is it says, first, what transaction log do you want to look at? What -- where is the file where the system recorded what happened? Somebody entered one, pressed the button, and on the right-hand side what you see is our representation of what those events were that would have been recorded in a systems log -- somebody pressed this, somebody requested a copy of that, that sort of thing.

The next thing that happens -- let's see, is the system says, "And what policy do you want to know about?" Basically, what rule are you asking if this series of events complied with? So, a little further down in that same thing is another little box like that where they put in where the file is that has the representation of the policy, they press the button, and the right-hand side fills with our form of representation. Understandably, this is a little more to the programmer's side than to the end-user's side, but sort of pretty far down there, it's breaking the rule into a pattern, and at the end there's a place where it's showing that refusing a request for public

service based on health information is not permissible.

So, then it actually puts those two things together -- it reasons, using the rule, over the data it has, and it comes to a conclusion. And the first one, again, is sort of a little programmer-y for most of our folks, so we said, "Okay, let's do something that's human readable."

And one of my favorite things, because I was a five-year-old who said the question, "why?" probably more than anything else, is we have the little "why" button. So, the system says to you that this isn't compliant, there's a little sentence down here where it says to you that the -- Betty, who's our operator, rejecting Bob's request -- is not compliant with the Massachusetts Disability Discrimination Policy, and then you get to press the little "why" button, and ask it, "Well, how come?" And it begins to give you more information and you can just keep pressing, why, why, why, and it will give you more information.

But, we still said, that might not make people completely happy, so we're now playing with a sort of lawyer interface, to show you what the system is really

doing underneath. Because it has the transactions, and the rules, it can actually put together what those of you who are lawyers may remember from first year of law school -- a traditional legal reasoning. It will basically say to you, "What's the issue?"

And this issue -- it's putting this sentence together based on the data that it has. It's always asking whether some transaction log is compliant with some rule, so here it's the transaction log, is it complying with the Massachusetts Discrimination Policy, because that's what somebody asked.

And then it says, "Well, what rule?" Well, the policy that you told it to look at is one that says that to be compliant, denial -- the Denial-of-Service rule part of the Massachusetts Disability Rule says that the reason for the event has to be something other than a disability, it can't be looking for a disability.

It goes on, and it says, "Well, what do I know about what happened here? What are the facts of my case?" And it says Bob Same, who's our phone company customer, made the request, and it was refused because of a particular record, that the system shows it's one record

where the refusal took place, and that this man is a resident of Massachusetts, and we know that he's, in fact, covered by this law, so then it does the analysis.

And it says, "That request was denied based on health information," the category of the information was health information, and that's what's contained in that phone record -- the phone company record they were using, and so it ultimately concludes that the transaction of Betty rejecting Bob's request, isn't compliant, is non-compliant with the Disability Discrimination policy.

So, what it's doing, just to repeat, because that's a lot, fast, is it's taking a system log of transactions that have occurred over time, and it's running a particular rule or policy that you asked it about, over that transaction log to see if it was compliant or non-compliant with the rule. And, of course, as you can imagine, you could do that with lots of other data mining questions.

And with that, I'm going to turn it back to Danny for, perhaps, a little more technical explanation.

Mr. Daniel Weitzner: Thanks. I'm going to talk really briefly about how we do this. I should say, first

of all, as to some of these user interfaces, we think we've done a lot of interesting work here, technically, we don't think the user interfaces reflect the state of the art of anything in particular, so this is a -- just as a matter of mechanics, this is a little extension that we've written for the Firefox browser, at the end of the slides you'll see there is a place you can actually download this stuff and try it yourself, but an operational system would have to do a somewhat better job.

What's -- what we've done that I think will have some value, as K. said, we have a general mechanism by which we can look at records of events, records of events of information usage and access, and then a way to record what the policies are that govern that usage and that access, and come up with results like this. So, this -- we spent a lot of time putting together -- mocking up the data that reflects this example, that's reflected in this example, but the underlying mechanism is, we believe, relatively generic, so it could be used in lots of different contexts, and in fact, could be used for rules that even have nothing to do with privacy. You might use it for rules that assess the reliability of information or

assess who's allowed to use different information based on its classification status, or anything else.

Very quickly, the system that we've built -- and I'll stress that this is a prototype, proof-of-concept system, has three significant components. First of all, there is a mechanism for doing relatively traditional access control, through rules-based access control techniques. That's not really new, but we've integrated that into the existing rules and policy description language that we have.

All of the access events and the usage events in the system are then logged in the system. What we're working on right now are techniques by which we could analyze logs that are generated by the systems that you might use -- the enterprise platforms that are commonly used in government and private sector, and transform those logs, of which as you know, there are many -- transform those logs into data that we can actually feed into a system like this, and get some accountability answers.

And finally, we've designed all this, really, as a set of components that can be used in the context of the worldwide web, so these are components that can be added

onto browsers and web servers. What that means, we believe, is that this is an architecture that can be relatively easily implemented in the context of any reasonably open architecture -- we're looking now at ways to integrate these tools into enterprise architectures.

A little more about the approach that we've taken, and this is where I will try not to fall off a technical cliff -- I'd say that what's important about what we've done here, the technologies that we've used -- as distinct from other kinds of access control systems and production rule systems that are widely used, is that we've implemented a technique which is known as dependency tracking, in the reasoning. That is, the computer program that takes the logs that you've heard about, and the rules that you've heard about and puts them together to try to get answers about compliance, uses a particular style of reasoning.

And the important thing that it does is that, it not only gives you an answer, a yes or no answer, "Are you compliant?" or "Are you not compliant?" but it also gives you an explanation for that answer -- it allows you to actually look through the logical process that the reasoner

went through to help you understand why you either are compliant, or are not compliant.

We didn't show it, here, but one of the useful features of this kind of dependency tracking reasoning, is that in many cases, as you well know, you don't get a simply yes or no answer to a question about whether a particular information usage, or information is compliant. What our reasoner will allow you to do, is to get as much information as the program can glean about whether a particular event is or is not compliant, and will then say, this is dependent upon a certain set of conditions or on a certain set of facts, and if you believe those facts or believe those conditions, then you're compliant. And if you don't, you don't.

So, what we're really trying to do is to go for a kind of a hybrid system that allows for integration of as much machine judgment as possible, but in the end, application of human analysis of what these results are.

I want to just close by giving my opinion about how this -- how we ought to think about these kinds of information accountability tools in the larger privacy context.

We think that these tools will be necessary in order to have a reasonable kind of compliance in complex data mining environments. We think that it is going to be, because of the scale of information analysis that goes on, because of the variety of different rules that may go into any given analytic effort, we think that some amount of machine assistance is needed, in order to supplement human judgment about compliance. But we don't think that these kinds of systems replace either procedural rules, or substantive legal rules about how information can and cannot be used.

Having lived through the, kind of, first generation of the interaction between the internet and the web and lots of legal issues like privacy and freedom of expression, and security, there's a tendency sometimes to think that one can wave a technical wand over any given problem and have a result. That is, decidedly, not what we're saying. We think that these tools can help with compliance, but don't avoid the question of what rules one is expected to comply with. And hopefully, this kind of technology will give a view towards the -- the sort of, the structure and nature rules that we actually have to be able

to put into place.

Let me just say a bit about where we're going from here. As I said, and I'll say again, this is very much proof-of-concept technology, it's not available off the shelf or off our -- it is available off of our website, but it won't get you very far, unless you just like to play around with these things.

Our next step is to try to test some of these approaches in more real-world pilot project contexts. We're very interested in the question of how -- how these systems can meet the needs of large enterprises, whether government enterprises, or private sector enterprises, so to that end, we're looking for test cases, we're looking for pilot project partners, and it is that way that we'll understand whether the systems that we've built work efficiently enough for the needs of an enterprise, and probably most importantly, understand whether the policy languages that we've written are actually able to express to the kinds of rules that any given organization needs to be accountable to.

So, thanks very much, and I look forward to questions.

[Applause.]

Mr. Daniel Weitzner: The slides didn't make it into the packet, my fault, but they are on the back table, and there is an article, as well, about our work. If you look at the slides, you'll find some links to some of our papers, and also some of the technology if you feel like playing around with it yourself.

So, thanks very much.

Mr. Peter Sand: Thank you very much, Danny.

Let's just take a few moments to talk about some of the things that have been raised so far, and then go into the last presentation.

One of the -- I'll just start off with a question, but I invite the rest of the panelists to ask each other questions, or raise issues that have been -- bubbled up from the presentations, so far.

One of the comments from a previous panel that really got my attention was that it -- I think it was Steve Dennis from DHS S&T, who talked about rules in a slightly different context. What he basically said is, if you can get two people to agree on something, then that becomes a policy than you can actually implement in technology.

And I'm wondering if there are things about the technology that drive the nature of the rules themselves -- when we talk about being able to go back and audit compliance with certain rules, or when we talk about building rules into technology, is there something inherent about the structure of technology that requires these rules to be stated in certain ways? Are there limits that you have to place on the rules that can be built into technology, that are driven by the nature of technology itself?

Mr. Brian Tretick: I guess I'd like to start out and say the first thing -- from an auditability perspective is, we like the rules to be objective, measurable, complete and relevant. And so, in some cases you might find a situation where we set one rule or two rules when there might be 20 that are relevant to make a complete set of rules, again, just to make sure you're doing all that you should be doing, not just focusing on one thing.

The other piece of that is, the measurable and objective piece. But the measurability is important, because you could have a rule that requires you to -- you can't prove something, again, that's immeasurable, and I'll

try to come up with an example as we keep going, because one won't come to mind right now, sir.

Mr. Daniel Weitzner: That's a good question -- I guess I'd say, overall, rules that can be expressed as specific prohibitions are a whole lot easier to instrument than rules that reflect, kind of, subjective judgment calls. So, a rule that says -- as you had in the example, that K. showed -- you can not use health information to deny service, is a rule that we can describe in a way that a computer can do something with it, and where we can detect a violation, because we understand what health information is, and we understand what a service denial is.

A rule that says -- that has reference, for example, to reasonable expectation of privacy, is one that our system is just going to kind of say, "Well, you're on your own on that one." And that's not necessarily a disaster, but it -- so some rules may kind of end with that question, but you know, specific prohibitions are better.

The second point I would make is that a lot of the time that we've spent on this work has actually not been on designing new algorithms or spending lots of time programming, it's time spent with people like K., and

engineers, who look at how to describe an environment -- both the information in an environment and the rules, in a kind of a consistent framework.

So, I think that going forward, and the technical term for that is ontology development, but we don't have to go into that -- what I would say is that, I think that to the extent that there is a commitment to be able to have this kind of accountability, it's going to be important to have an effort to have a consistent view of -- a consistent description of the information in the environment that you're trying to control, and make sure that that's built on.

If we have 17 definitions, for example, of what an agency is, we're going to have a hard time understanding how to do Privacy Act compliance of FOIA compliance, or name your legal area.

So, a view towards gradually building up a set of consistent categorizations for information and consistent styles of expressing rules, I think, is going to be important.

STATEMENT OF MICHAEL AISENBERG, PRESIDENT, INTERNATIONAL SYSTEMS SECURITY ENGINEERING ASSOCIATION

Mr. Michael A. Aisenberg: Peter, your question, for me, really previews the opening thought that I was going to use to tee up my remarks which, I think is, in part, reflected in this slide, with apologies to our friend and colleague, Tony Rikowski, is a map of the standards bodies which, as of January of 2007, were engaged in projects addressing identity management that incorporated information security, information assurance and privacy issues, and that number -- the individual ovals, to find -- individual committees that had those projects at that time, and actually a more recent update of that slide is in preparation now, with roughly twice the number of standards bodies, going from roughly 50 to roughly 100 now, that are addressing that issue.

Those are places where the rules that Danny has just talked about are attempting to be developed by communities of experts, or those who describe themselves as experts, and they will promulgate them for applicability and contexts, ranging from government agencies around the planet, to individual users, in particular market stovepipes -- whether it's banking and finance and securities, or healthcare, or other particular areas of

application.

So, the question that I think is begged by Peter's, question to the panel is, which rules are you talking about? We are at the ever-widening end of a cornucopia of continuing rule creation.

Let me just read from a document that has a date of May 7, 2008, it's from the Office of the Press Secretary, Crawford, Texas, "Subject: Designation and sharing of controlled unclassified information," interesting acronym, CUI. Paragraph two, "The global nature of the threats facing the United States requires that, a) our nation's entire network of defenders be able to share information more rapidly, so that those who must act have the information they need, and, b) the United States government protects sensitive information, information privacy, and other legal rights of Americans." Definitions, 3(a), "Controlled unclassified information is a categorical designation that refers to unclassified information that does not meet the standards for National Security Classification, under EO-12958, as amended, but is, 1) pertinent to the national interests of the U.S., 2) under law or policy, requires protection from unauthorized

disclosure, special handling safeguards or prescribed limits on exchange or dissemination." "(f), enhanced safeguarding, is a handling requirement that means those information so designated is subject to measures more stringent than those normally required, because inadvertent or unauthorized disclosure would create a risk of substantial harm, this requirement is indicated by the marking 'controlled enhanced.'"

Section headed, "Policy, the CUI Framework," paragraph 7, "All CUI shall be, a) categorized into one of three combinations of safeguarding procedures and dissemination controls, and be so indicated through the use of the following corresponding markings." And I won't read the paragraphs under them, but the three are controlled, with standard dissemination, controlled with specified dissemination, or controlled-enhanced with specified dissemination.

[Laughter.]

Mr. Michael A. Aisenberg: That's a press release describing an Executive Order, that has not been passed by the Congress of the United States, it does not impact the classification system that is limited to agencies that have

classified data origination authority, which are, actually, very few, but it does intend, by its other nine pages, to apply to every agency.

So, the question really being begged for me, is what do agency heads, and organization managers, need to be thinking about in determining what controls to employ and deploy? What will be the rules that will be audited? The universe of candidates for those rules is not simple, it's not precise, it's not easily understood, it's growing day by day, and the challenge of picking the rules, and understanding what those rules mean, has to be addressed even before reaching a question of auditing against the rule that you've picked, or in Danny's architecture, being able to test whether there's some validity to the deployment of the rules, and the auditability.

So, I think we have a series of problems that are -- it’s almost monumental -- I won't say insurmountable, but certainly monumental -- to reach, in terms of the selection of, what are we going to do to protect individual, personally identifiable information both in the government context and in the general marketplace, and what is going to be mined?

Ms. K. Waterman: I want to take the opportunity to respond to that, because I think in a way, it actually goes directly to what Danny was saying about driving towards consistency. The CUI project that Michael's referring to came about because it was discovered that there were many different labels people were putting on things that they wanted protected, but that wasn't classified, that could be private -- privacy, which is not part of the classification system. It could be things that were going to be bid out for contract before they became public knowledge, it could be employee -- well, that would still be privacy promotion records, things like that, but there were a large number of categories of things that had nothing to do with classification, and they had labels. And every agency made up their own label.

So, some of them used "Sensitive but Unclassified," some of them used, "For Official Use Only," and it was ultimately determined that there were over 400 of these. And as we began sharing with State and local agencies, they complained bitterly that they were getting too many labels, they didn't know what to do with them, and could somebody, please, make this more understandable.

So, exactly what Danny was talking about in the consistency standard -- while I agree that, perhaps, the titles are a mouthful -- what's been done is take in over 400 labels, and reduce them to three, so that those people getting them could understand they had one of three obligations for how to handle what was being passed to them.

So, that kind of consistency, if it actually gets implemented, would be fabulously helpful to the ability to do something at the technical level.

Mr. Peter Sand: Another issue which came up in the previous panels, I'll categorize it as a scope issue, which is -- there was a general framework that was laid out in terms of data mining that was basically the data collection, the analysis, and then some kind of decision making at the end. And as you looked at auditing complex systems like data mining systems, are you looking, or would you look just at the internal operation of the system itself, or would you look at a broader context, in terms of what comes out of it?

One of the issues that we talked about a little bit in the previous panel was kind of the output of a data

mining process. Does it deliver a specific answer? Yes, no? That's the guy? Or is it delivering a probability, like, well, maybe this category of things is interesting and you should look at it.

And I would think that as somebody's looking kind of from a distance at one of these systems and where it fits in an organization, the issue would be what message is coming out of the data mining system for the person that has to make the decision.

In terms of your reviews or as you look at these instructionally, do you look at a larger scope, do you look at where this system fits in a broader context? Or, how do you decide what's within the scope of your audit, and what's left for the people to work out?

Mr. Brian Tretick: I'd like to take a first cut at that, one, you generally audit something that happens, so it's more than just the technology, but it's the entire business process around that. So, from a financial auditing perspective, it's the -- you audit the integrity of the financial reporting process. For, under Sarbanes-Oxley Act, section 404, part of this accountability, you audit the effectiveness of internal

controls over financial reporting.

In a situation like this, we would look to the rules that are applied, try to map them to some framework, generally accepted privacy principles is one of those, and we would look, then, at the technical, procedural, physical controls, and even the legal controls that are in place to determine whether they were designed in place, designed effectively, and operating -- designed and adequately operating effectively over time -- to determine whether or not we could develop, again, a position of reasonable assurance that there was a controlled environment in place.

That way, if you have that kind of controlled environment, you don't need to do the other thing we call, the other end of the spectrum, which is substantive testing, where you use Danny's tool, K.'s tool for every transaction that occurred to see whether it was good, bad, or ugly. One would, in a financial audit, if there were no technical controls, one would have to hull some huge, statistically relevant number of transactions to develop the position of reasonable assurance, and not absolute assurance, that you'd look at all of the transactions.

If I could trust the process from a physical, procedural, technical and legal perspective, I can rely -- and I can see that it was designed, in place, designed appropriately, and operating effectively over time, I wouldn't have to do that kind of substantial audit, and I wouldn't even have to know what went on, because I was in a trustworthy environment as it is.

So, you really do need to look at everything. The short answer is yes, I guess.

[Laughter.]

Mr. Daniel Weitzner: So, part of what our -- I think this is a -- it's a very useful distinction that Brian makes between the kind of audit that is traditionally done in a financial setting which says, "Is the system such that you can trust that the right thing happened?" I think what we're trying to do is to -- so we're not really trying to build an audit system. We're trying to build a system which helps make sure that the right things happen as they do, and -- as they are happening.

So, the analysis that K. showed can be applied either if, for example, that hypothetical customer had a complaint and said, "Well, I was denied service for an

illegal reason," or, hopefully to save everyone trouble, the moment that that customer service agent said, "Well, I'm about to -- I'm going to deny this guy service, because I just got a look at his health information." Right then, we ought to be able to tell the user of the system, "You can't do that." Or, at least, "If you do that, you will have broken some law."

So, we're really about visibility into the information flow as it happens, and then providing the ability to go back and check, whether either for audit purposes or for other forensic purposes.

But, I think what that takes, to come back to Pete's question, it really does take maintaining, a sense of the context of the information. People in computer science tend to call that provenance, where did the data come from? We're interested in kind of a policy provenance, what are the rules that come with the data and how do we keep track of that? So that the person who uses the results of some data mining activity, and then takes that somewhere else and communicates it to somewhere else, is able to easily maintain the context from which it arose, and whatever restrictions would have applied.

Mr. Michael A. Aisenberg: I think what's interesting about this question is that it begs the interest of whose interest is being vindicated in the process. Typically, the reason an audit is imposed is because the institution is seeking validation of its own behavior in some community. You're trying to assure that the managers of an organization have followed some rule that they're supposed to follow. And it's their own interests that's typically being vindicated.

What Danny's tool adds as an interesting aspect to the audit and conformity community's toolkit, is now you have an additional tool that, perhaps, is available to third parties, like the subjects of the PII, to determine whether an appropriate or inappropriate use has been made of the data -- it expands the universe of remedies in the community of those who are beneficiaries of remedies in the data life cycle community. Now, the PII subject has a potential tool that can demonstrate that there has been an abuse, or a non-conformity with the rule as to how that data is supposed to be handled.

To me that's an interesting -- a very interesting development, but it still begs the further question that I

asked, or addressed in response to the previous question which is, which rule? And, to me, that still is a much more difficult question in today's environment, because we are -- in this era of massive rule transformation -- coming from an enormous number of sources, from industry-developed internal best practices, even from individual company practices, and industry practices, sector-wide practices, standards bodies, individual nation-states, global standards bodies -- the sources of rules are enormous. And even in the face of a statute, looking for conformity across 86 major U.S. government agencies and the way they will apply those rules is going to be, you know, just complying with the CUI standards -- I guarantee you, you will have a wide diversity of deployment across the agencies as to how they select which aspects of the CUI memorandum they are going to emphasize, between DoD and the intelligence agents. He's at one end of the spectrum, and the more human-oriented agencies like HHS at the other end of the spectrum.

Ms. K. Waterman: I agree with everything that Michael has to say here, and I just wanted to add something that didn't come up in our demo. In trying to be flexible

with this project, you can apply a rule at any point. So, you can have the system deploying a rule set, which is what people who own the data think is involved, but if, at a later point, you want to audit because somebody's raised the question of a different one, you can actually deploy a different rule.

Mr. Peter Sand: Thank you very much.

Let's go into our last presentation, and then hopefully we'll have some time for some more panel discussion, and then open up for your questions.

Michael?

Mr. Michael A. Aisenberg: Thank you, Peter.

I had not intended to work from the slides, but as the discussion has gone on, because I would have the propensity to wander and fill most of the rest of the day with observations, I'll use the slides to just to try and impose some self-discipline, here.

And the reason I had wanted to lead off with this slide about the standards community was to make a point we've just been talking about -- that the sources of rules today in this space are enormous. And we are hardly at a punctuation point, where any definitive understanding can

be asserted. There's nothing to lock in. And even now as we approach a national Presidential transition point, the likelihood of rule and policy changes in the area of data management and data safeguarding privacy across the United States Federal government is likely to be upon us for the next year or 18 months, irrespective of party. It's partially a cultural reality that the new guys like to do a little housecleaning, both at the personnel level and at the policy and practice level, and I think we can anticipate policy and practices regarding data and data collection, and data storage, and data management, and data security to be upon us for the next 18 months, and it won't come just from statute and regulation, it will come from individual agency behavior, individual manager behavior, and in the technology space, it will also come from these emerging standards bodies.

I guess my point here is that context is important. And data mining and data exfiltration are perhaps understood as two faces of the same cube, that what may stand for a perfectly permissible practice and behavior by one entity may, in fact, be viewed as something akin to an act of war through a different lens.

We have to remember that the networks that we're talking about where these pieces of data are stored are global, and that the operators of the networks are not necessarily U.S. citizens, or U.S. actors are not necessarily acting exclusively in the United States, they are acting globally, and the threats to those data sets and networks are global.

So, policies that we make, and policies that we follow, and policies that we observe being followed elsewhere, have to be understood in that global context. And rules that we write may well, as consumed us in the late nineties with the years of negotiation with the E.U. over their privacy principles, may require extensive, and perhaps even unsatisfying -- agreeing to disagree -- kind of outcomes. I think it is important that we understand this.

A further contextual hand grenade -- someone simply described it benignly as a landmark, I think it's more of a landmine -- is the Administration's cyber initiative. Finally, after 6 years, there was a public recognition that massive amounts of classified data are being exfiltrated by our adversaries, and the Director of

National Intelligence has, since October, been engaged in a program with an investment tag, estimated somewhere between six and $30 billion to do a number of things. To fund the R&D necessary to make our networks more secure, to enhance the relationship between government network managers and their private sector stewards and contractors, and to take a hard look at the use of personally identifiable information that exists across networks, and understand what rules are being applied by agencies, including defense and intelligence agencies, may act as either enhancements or deterrents to that condition of exfiltration and sensitivity and vulnerability that the initiative is designed to stop. So, I think that initiative will certainly transcend the election and into the next Administration, and it's a major artifact in the landscape of data security and data privacy. And I think it bears watching.

Another point that I would suggest is that we should not become too enamored with technology as the exclusive solution. I think you've heard both Brian, and Danny, now describe the importance of the other two legs of the triad -- it's people, processes and technology, and I

think we've under emphasized the people and processes part of the triad, in favor of the technology. The tools are, essentially, in most cases, neutral. One person's tool can become another person's weapon, depending on the intent and how it's used.

It's very important that we not overstate the capacity of any particular tool to do the job. Right now, cryptography is considered to be the tool of choice for protection, and it appears, specifically, in all kinds of statutes and regulations. However, a recent paper published out of Princeton, and you see the URL for it up there, and the exact title of it, if you will bear with me for a minute, is Less We Remember, Cold Boot Attacks on Encryption Keys. And basically, it's a 12-page paper that describes the human behavior that is necessary in order to make strong encryption truly effective, to recognize the vulnerabilities that exist when data -- even encrypted data -- is stored on a live memory chip, on a live board, in an active processor that happens not to be turned off. The simple benefit of rebooting, and clearing the cache, has an enormous impact on the security of the system, and the ability of that cryptographic system to achieve its

intended benefit. And those who fail to turn the system off suffer the penalty of opening up certain means of access, of that key to those who have the technical tools to decipher and attack the data in storage.

The point being not that crypto isn't a good tool, but that every good tool has to be deployed in an effective manner, and this year's good tool may be defeated by next year's better tool cutter. One of the first lessons I learned in security was that you go back to basic principles from the 12th Century -- for every chain that you deploy in the castle keep to lock up the prisoners, there's going to come along a strong chain cutter to cut that chain. And so it's a constant effort of mutually assured destruction, if you will, nuclear arms race kind of environment of constant pursuit of the encryption or other protective tools, and the constant battle by the bad guys -- and in some cases the good guys -- to develop the tools to defeat the last generation of protection with the next generation of cracking. And unless we are aware of and sensitive to that constant evolution in the marketplace, we run the risk of writing particular technologies into practice, into regulation, into statute, in a manner that

leaves us very vulnerable.

Flexibility, and the ability to deploy the most appropriate approaches to protections -- and that includes not just the technology, but good practice, and good inquiring about the individuals who are the practitioners.

One of the major areas of vulnerability in the Administration's Cyber Security Initiative, is taking a real hard look at the people with the highest levels of clearances, in the highest levels of our own national security and cyber-security community. Because the insider threat -- or, as some in the industry like to describe it, the threat that exists in the 18 inches between the screen and the back of the chair -- is often overlooked, and very often where the worst offense is occurring.

One last point, the last quote on this page, and this is an old homily, but built-in is typically cheaper and often easier than bolted-on. As you design system, baking the security, and baking the privacy protection in, is typically a better approach than waiting until the system has been developed, and then coming back after the fact to bolt it on. That's a historical lesson that goes back -- for the engineering and technical wonks in the

room, there was lots of good security in the X-500 architecture, but when we developed X-509 TCPIP, the basic framework that now modern desktop computing is based on today, an awful lot of the security was left off because of the scalability issues, now we're facing the consequences of that by trying to bolt on security to the TCPIP environment, at a cost of billions of dollars that, perhaps, could have been avoided if we had thought about this or recognized the risk 20 years ago.

Finally, I would suggest that the path forward ought to be one -- and I think this is sort of implicit in a lot of the work that Danny and the folks at MIT have done -- there's a reason why PGP is such a fascinating name. The P doesn't stand for Perfect Privacy, it stands for Pretty Good Privacy. Eighty percent solutions may be not only the best we can do, but more than enough to overcome the threats that we're facing, to deter the bad guys, to send them off looking for other avenues of achieving whatever their ends are. And the relentless pursuit of perfection may be an interesting advertising slogan, and appropriate to auto engines, but it may not be necessary in this environment.

What is more appropriate, I would argue, is the relentless pursuit of progress, that we should be trying to do better than we have been in the past. That includes policies that don't focus on locking the barn door after the horses have escaped, which is what I think breach notification is all about, but end-to-end data custody practices that instill both in the subjects of the PII, and all of those who are depository institutions, whether they are those who are able to compel deposit, like government agencies, or obligatory ones like universities, or voluntary ones, the people that you transact, give up credit card data with, as you're doing online transactions, whether it be eBay or whoever.

They have varying levels of obligation, I would contend, to protect that data when they hold it. The greatest burden, I would argue, falls on government who -- because they can compel the deposit of data, and the gratuitous third-party folks, the data brokers with whom individuals have no relationship, no legal privity, who perhaps don't even know that those parties hold their PII, and yet those parties are making money off of the use.

When it's abused, those folks, I would contend,

ought to be pilloried and given the worst that our legal remedy system allows us to impose, but for my money, it's insufficient for us to develop those structures unilaterally as a nation-state. We have to do that in collaboration with our allies, and those who may be not be -- necessarily be our allies. These systems, in the face of the global networks I described at the outset, have to be addressed on a global basis.

That's a challenge. Getting the Chinese to come to a table for a diplomatic discussion about common remedies for privacy abuses is not going to be an easy thing to do, but neither is discussion of global warming or nuclear nonproliferation -- the important tasks facing us as a global community going forward, and this is sort of a transition, I guess, to the next panel, and the discussion of policy.

But the transition to that has to be, in my view, done on a global basis, both with our close friends and those who are not our friends, and we have to be willing to invest in a long-term negotiation, in order to achieve the benefits that are commensurate with the enormous economic and national security value that these technologies now

provide for us.

My friend, Steve Malfrus, at the Fed, likes to talk about -- we're a $12 trillion, going to $13 trillion economy, well guess what? $3.5 trillion of economic activity happens over the network every day. Most of that is initially handled on secure EDI networks that are Electronic Data Interchange networks that are managed by the Federal banking regulators, but without bankofamerica.com and NYSE.net and treasury.gov and barkleys.uk -- without a secure, safe, and available internet, to receive that data -- never mind nuclear command and control codes and the national security stuff -- the fundamental economic juice of the United States and the Western economy will begin to grind to a halt at the rate of $135 billion an hour. If that's not a number that catches the attention of CEOs and Congressmen and agency heads, I don't know what number will.

So, there's a lot at stake here, we have a lot of work left to do. But as you can tell from the discussion over the last day and a half, an awful lot of important work has been done -- now it remains for the policy community to make choices from among the candidates for

rules, to figure out what the right rules are to deploy, and help people understand how to use them in an effective way.

Thanks.

[Applause.]

Mr. Peter Sand: Thank you very much, Michael. Again, going back to one of the themes that was raised in an earlier panel, which I think your talk really pressed on here is, whether data mining systems or technology itself present unique challenges, or one of the earlier panelists talked about data mining tools as more, like a power tool. It does what you've been doing, but it does it better, it does it faster, it's easier for you, it's more efficient, more effective. And I'm wondering if there, if that's what we're looking at when we're talking about data mining, is it just doing the traditional things, but better, faster, stronger, cheaper, or is there something qualitatively new about using technology to do these same kinds of operations and rules-based behaviors that raise new challenges that we need to address separately from the underlying processes that would happen without the computers.

Mr. Daniel Weitzner: I think that what's

qualitatively new about data mining, which I gather yesterday you decided you were not going to define. But, I think what's qualitatively new is the scale. It's the scale of information that's available and the analytic power over that information, the social network example I gave, it's just one kind of little toy indication of that relative to what I suspect is available in places that the public can't see.

And I think that, you know, to Michael's point, when the problem that we had was securing, you know, a small number of very high-value assets, like nuclear launch codes, or even when the problem was securing relatively well-defined networks amongst a defined number of financial institutions, we could take one approach, and I think that approach in many cases was characterized by, you know, very, very innovative use of cryptographic techniques to keep information out of the hands of people who shouldn't have it.

I think that what we're learning, even in network security, but I would say even more so in the kind of information accountability that we now need is that, as Michael said, to close the barn door after the horse is out

is not going to be a satisfying result any more.

So, I think that as we get to a kind of web-scale analysis of information, and by web-scale I just mean big, a lot of it, I think our ability to control, up front, every single possible use of that information, is going to decline, so what we have to do is supplement the traditional security techniques -- we don't throw the traditional security techniques out, we don't just say, well, everything's now free, and public and available. But, I think what we have to recognize that there are limits to our ability to control information flow and information use up front, and we have to supplement them with the kind of auditing and analysis techniques that detect behavior or harm that we want to prevent against. And this is where the 80 percent rule is going to become awfully important.

Most of the rules that we live with in the world, particularly in the financial world, most of the rules are not enforced up front. Most of them are enforced by virtue of the fact that people know they will get caught if they break them -- the tens of thousands of publicly traded companies that have to tell the Securities and Exchange

Commission what they're doing, file huge amounts of information with those regulatory agencies, and no one in those agencies look at that data, no one checks whether it's correct or not.

Other people may check -- do check -- when there's some reason to look, and then the penalty for misreporting, or false reporting, is very high, if you get caught, but mostly people comply with those rules because they believe they would get caught and there's a, sort of a credible threat along the lines of, you know, sort of a credible, trusted system that you could audit, that you could detect and audit, that will catch the violations. And I think we have to stop treating personal information as this sort of special secret commodity, it is special, but it's just less and less secret. And we have to, I think, start thinking more about the kind of regulatory models that I think are largely successful in very large-scale institutions, like the financial markets, that recognize widespread flow, but still have expectations about ultimate compliance.

Mr. Michael A. Aisenberg: Two quick comments on what you just said, Danny, one on the issue of scale, lest

anyone think that it's not getting huge -- the daily number of resolutions on the network that -- resolutions means either an email address being sought to direct a piece of mail, or a URL that you type into a browser, trying to direct you to a website, has been hovering in the 500 billion resolutions per day number for the last 12 to 18 months -- 500 billion resolutions per day.

It grew from about 250 billion at the start of this decade, but it has plateaued. Some of that is capacity related, but as we go from a billion users outside of North America, to 2 billion outside of North America, it's expected that that scale will rise, and on the SOX point, in terms of cost.

I think it's public information that when I was at Verisign, the first year that we had a SOX-compliance audit, Stratton Sclavos was CEO, complained bitterly in a public forum that it had cost him an additional $16 million to pay his auditors for that SOX conformity. And less than a year later, an HR employee's laptop was stolen, which resulted in a roughly $5 million cost to pay all the dismissed employees whose PII was on that laptop, to get them all three credit report free-year-of-service remedies -

- that seems to be the prevailing way that you deal with stolen PII from laptops from employers these days.

So, you can spend a lot of money on Sarbanes-Oxley conformity, and still not necessarily get a cast-iron defense against bad things happening.

Mr. Brian Tretick: Excellent.

I've got a couple of points. First, to go back and say that, again, I would contend that the legacy of the financial audit analog works here with data mining, works with whatever you're doing, on privacy. And as we explain it, the financial audit is developing assurance over the collection, use, storage and retention of financial information -- disclosure of financial information -- integrity of the process. There's confidentiality requirements before public release, so everything is in there.

It's the same deal with personal information, or other confidential information -- trade secrets or government-classified data - have controls over the collection, use, retention, and disclosure of that data.

The AICPA generally accepted privacy principles can be seen -- not just as a testable set of standards, and

auditable standards -- but also can be seen as an API, like an Application Program Interface, where we have the myriad of requirements that are at the agency level, or the department level, or the business level, or the legal or regulatory level. We believe these can be mapped to this so that you can build programs of controls and assurance that see all those myriad of requirements through the lens of generally accepted privacy principles.

Like a medical examination, a privacy audit may not always just tell you you're in good health. It could tell you that you've got to drink less, eat less fatty foods, or that you're near death.

So, in some cases that'll tell you -- a privacy audit will tell you that a control isn't in place; it isn't designed adequately. It's weak or it can't be shown to be operating effectively over time, or in some cases, it's just not auditable; it's just not knowable.

What this does -- these qualified opinions, everything's good by these 6 things - these qualified opinions point to risk, and would allow management of the organization, or the partners in a data-sharing environment, to point to somebody and say, "You need to fix

that."

Okay, so that's the two -- so, the -- and the last point doesn't quite fit neatly with all of these others. We look at three types of controls: we look at the prevent controls -- stop bad things from happening, encourage good things to happen, okay? That's the first level.

The second is the detect controls -- if something bad does happen, or starts to happen, or whatever, tell the right person so they can be reacted to.

And the third are things that correct the situation. And your rules may not be prevent, prevent, prevent; it could be, here's a little prevention, a little detection, and a little correction, and that may be your rule set, and you can operate it imperfectly -- if you have a forest fire, again, preventing forest fires is a good thing. But, if you have a forest fire, you have a process to contain it. And, in many ways, that acknowledgement that there could still be a forest fire, that you can't prevent them all, is probably the most healthy way of going about your business.

Mr. Michael A. Aisenberg: I think what Brian has

just described is what Toby certainly will remember from the FTC workshops on privacy and security, the homily that's been around for some time, that you can have security without privacy, but you can't have privacy without security. Brian's three points are the essence of systems security deterring, preventing and defending against incursions. You do your best to keep them from hitting you when they are there; you try and make them bounce off of your defensive tools, and when they finally make it through, you try to mitigate and minimize the amount of damage they do to your systems. That's the nature of the security, and without those security tools in places, any sort of privacy architecture is going to be incapable of being deployed and achieved effectively.

Mr. Peter Sand: Well, thank you very much for the presentations, it's been really helpful, thank you.

[Applause.]

Mr. Peter Sand: And I'd like to open up the floor to your questions, if you have any, please come to the microphone and ask us what you like.

Mr. Bob Burns: Good morning, Bob Burns, HSR, DHS Science and Technology Director. I really enjoyed the

presentation on your system. It was amazing. Yesterday we talked about the complexity of data, showing all its trees, and if you touch this point, you go here; you're showing the converse with the rules and the regulations and the impact to the world we live in.

The question has to deal with not the person who was refused service; the question kind of deals with the person who started it all: the patient in the coma. Everybody acknowledges, okay, this person had tuberculosis, and we needed to find out who he had been in contact with -- we did a privacy search, I gather, in your theoretical model, but it would happen in the real world, too.

So, in that case, a diagnosis was made which validated us doing data mining, and therefore, taking an action. And everybody says, “good,” because the threat to the populace, I guess, exceeded the need for the person's privacy.

So, the question is -- you mentioned this earlier in your brief, and it kind of opened it up to everybody -- assessing data, using the data, in the data mining process. Would it be reasonable - or does it get too far into the danger Zone - can we use the assessment of the data, the analysis

of the data, as the diagnosis phase?

Then we have to make the judgment call how to use the data, and is that taking it too close to where you might have mission creep, or the danger zone, so to speak, because what we're finding, or what I see in many cases, is we're not able to make the diagnosis, so we can't decide what we need to do.

And in a lot of the cases, the things we want to diagnose either haven't fully occurred yet, such as a TB case, or we're trying to prevent them because of the nature of what would occur. How do we balance that? The assessing, putting in the safeguards, so that I can make the diagnosis and make the decision to use it, or am I getting too close to that danger zone?

Ms. K. Waterman: If I understood your question correctly, I think from our -- from the process of a system like ours, at each step where data is moving, or data is being used, you can apply the exact same analysis. So, if the question is, at the very front, can the CDC use the information it has to conduct an investigation, to go out and seek more information? There are rules about that, as well, and you do exactly the same process ; you would take

the information it had at that time, and its attempt to go out and collect, and run the rule over to see whether the collection was permissible.

Mr. Daniel Weitzner: Could I just -- let me just see if I understand your question. I think you're asking whether it's acceptable -- according to some set of standards, and I'm not sure what the standards are here -- but is it acceptable, according to some notion of privacy, to generate what you might call the reasonable suspicion or the probable cause from the data? So, is it okay to look first, in order to find the justification that we would normally expect to continue with an investigation?

And I'm going to dodge -- if that's your question, I'm dodging it, and I'll tell you why I'm going to dodge it. Because I think, as you know, the answer is, that it depends. That there are times when, you know, when traditional Fourth Amendment law, when we say there are exigent circumstances, and go back and prove that, you know, prove that you had probable cause later, or get the order later. There are times - we picked the public health example because it's what people tend to come up with as the break-the-glass, sort of, example -- it's an emergency.

You don't have time to ask permission of the patient who's in the coma, or the -- whoever else it is. You go; you need to find the answers.

Now, I don't -- I guess my belief is that amongst the range of, sort of, analytic activities that are possible, there are going to be times when we say, "Yeah, break the glass, and prove that you were right in doing it later." And there are going to be times when we say, "No, the information is too sensitive, the activity is too intrusive, and we're not" -- and simply the fact that we could find some evidence of illegal activity is not sufficient to say, "Let's just sort of sweep up every piece of information about every person, and see what we find." That's the extreme. And, I'm not --

So, I guess what I would say is that we've tried to design a system that allows us to operate across that spectrum, so that if justification is required before doing the analysis, we can represent that. If what you want to be able to do is retroactively establish justification at some later point, we could keep track of what happened, so that the actions that were taken can then be assessed to see if they would have been justifiable.

And I -- so, from a technical standpoint, I don't actually think it's our job to build a system that is biased towards one end or the other of that spectrum, because I think all conditions are going to happen. What's important from our standpoint is to make sure that in -- wherever you are on that spectrum -- that you can establish either in advance or after the fact that you actually followed the rules that were in place.

Mr. Bob Burns: And your dodging -- to use your term -- is fine, because it drives home a point that we came up with yesterday, as well. When trying to ascertain what is acceptable, what we can look for, how we can use the data, I understand completely why you go with the tuberculosis patient example -- everybody rogers it; it's locked in law; you know what you can work with. We're having the problem -- how we come up with what is back to Congress, perhaps, or legislation, or rules, one of the things we can look for, and what are those paradigms that we can work within, like the tuberculosis patient, how do we make that happen? How do we drive that?

Ms. K. Waterman: One thing I just want to loop back, and then -- Michael, if he wants to answer -- what

Danny talked about before with the dependency tracker is, we have the ability, if somebody's established criteria, which, in most cases they have, for what is probable cause or what is a reasonable belief. If we can define those things, the system can at least pop up what facts it has, and some person could make a judgment about whether that rose to the level of enough or not enough, but you could at least percolate the facts back up.

Mr. Michael A. Aisenberg: I think this is an area where we have to be careful again, and I personally believe that the technologies are, in virtually all cases, from a public policy, rule-of-law standpoint, neutral. And the question is, how are the technologies used?

And at the end of the day, for me, this goes to the issue that Peter Swire and a number of us were talking about in the back of the room, which is the excellent - I guess it's operating code of the privacy group at DHS - which is Ben Franklin's statement, which is one of my operating rules in this space: "Those who would sacrifice a moment's liberty for the sake of temporary security soon will find themselves having neither."

We've heard any number of cabinet secretaries and

committee chairs since 9/11 justify what they consider to be compromises and exigent deployments of tools and behaviors and practices in law enforcement and elsewhere over the last 7 years, based on the terrorism threat, the national security compulsion or some other rationalization. And I say, "Wrong."

For me, as an attorney, for me as I understand our constitution, for me as someone who is hoping that the world that my grandchildren and great-grandchildren grow up in is not all that different from the one that we've understood, at least in terms of legal remedies, the most important thing is not letting that balancing act become the way we behave in this space and become the excuse for deploying what can become draconian applications of technology.

The technology is very seductive, especially when it's poorly understood. We've got two communities: we've got the rule-of-law guys over here, and we've got the Ph.D. computer scientists over here, and there are a couple of folks in the middle who are both, and they try to translate back and forth, but the technology can very quickly get out of the barn and become the devil.

And it's the responsibility, I would argue, of those who are part of that rule-of-law community to remind us over and over again of what Franklin said, and make sure that what we try to do is use our genius as Americans to maintain and maximize both liberty and security at the same time, and come up with the tools and with the application of those tools that are consistent with that value set.

That's a challenge, but I think we're up to it.

Mr. Peter Sand: Thank you very much.

Let's move on to the next question.

Mr. Chris Clifton: Yeah, I -- almost a follow on, but just to step back, there's -- this panel, I get the feeling that there's -- we're starting to move towards thinking that, or equating privacy with access control. And I think privacy is much more complex, and it's starting to sound like it's a form of access control.

As an example, in this scenario we've been discussing, I think there's a very big difference between general use of the information to discover pandemics and this particular scenario, which is saying, "Hey, we're trying to discover people who need our help." If you look at privacy laws, they get it.

Use of my data to do something that I -- you know, that is for my benefit, or that I have requested, is very different from using it for some general use. There's also the risk of harm. Using, you know, using my private data for something which may result in my loss of phone service is very different from using my private data for something which may result in my incarceration.

And I just want to make sure that we keep in mind that privacy is not about binary -- you see the data or you don't. There's a lot more to it than that. And I'm just wondering, do you have any comments on that?

Mr. Brian Tretick: I'll take that as, not to be seen as the mouthpiece of the American Institute of Certified Public Accountants, I am a card-carrying member --

Mr. Daniel Weitzner: Can I see the card?

Mr. Brian Tretick: Actually, I don't know where I keep it. Nobody's ever asked me for it before.

[Laughter.]

Mr. Brian Tretick: We try to define privacy in the -- for the -- what became the generally accepted privacy principles; we ended up describing it, and we

described it using these words: privacy is described as the, this kind of tension between the rights of the individual, and the obligations of the organization, over the life cycle of personal information.

So, the rights of the individual, obligation to the organization over the collection, use, disclosure, and retention of the personal information, but the use is the big deal. The way I describe it is, privacy is -- it's the wrong word. It doesn't mean anything, but we're stuck with it. Privacy is protect the data and govern its use. So, it's security and govern its use, whatever -- by whatever the rules are.

Going back to the first question, as an auditor, I'd look to have control over governing the use. If you have the data and you don't know what the use is, perhaps the governance is some sort of rules of engagement over the data, which could include, for example, an independent review board, or like your FISA court or something like that.

So, what we would look for is not necessarily specific controls over the use of the data inherent in a system, but a procedural process and a governance, a data

governance overlay, over top of that that would say that, when data is used, it has gone through the right level of a review and approval before it happens. That would be my definition.

Mr. Daniel Weitzner: Could I just say, really -- Chris, I think that was about the most -- you said, I think, about the most important thing that one could say about privacy, which is that it's more than just access. And, I'm not sure exactly to whom you addressed the question, but I can say that the motivation for our work is to be able to do a better job of detecting uses that are contrary to whatever the rules are. Because I think, as you know, systems are reasonably good at preventing and detecting access, because that's an event that, kind of, computers are better at representing. I think it's harder to represent uses both legally and computationally. So, I can only agree 100 percent.

Mr. Peter Sand: Let's move on to the last question.

Mr. Dave Weitzel: Dave Weitzel from Mitre.

The question’s focus is to Danny Weitzner, but I wanted to build on Brian. In the audit world, and as a

former Director of PWC, I think I understand a little bit of your business, you have the GAAP standards that go back a long, long time, and they really are generally accepted. And financial controls - we can go through 150 years of audit history to see how that stuff's evolved.

One of the challenges of GAAP standards is, they're really not generally accepted -- yet, right? We have it by the folks that Michael works with in the security controls industry, so my question to Danny is, in information security space, we have ISO-1799, and its progeny; we have ITL; we have COBIT; we have NIST, FISMA 800-Series standards -- how do we make that happen so that people know that we're talking about pick your standard? Maybe we should be starting with the GAAP standards here, but how do we do that for the tools you're building, and build it in a way it can be adopted broadly?

Mr. Daniel Weitzner: That's a great question. I think that it has to start, Dave, with an understanding of what the rules are we're trying to enforce and have accountability against. So, I guess I'd poke a little bit at the definition that Brian offered of privacy as this balance between institutional and individual interest --

well, that's true, obviously. I'm not sure how much it helps us.

It certainly doesn't help me, at all, in thinking about how to design a system; it just tells me that there's a problem, and I think that's -- and I don't think that's the fault of the auditors or accountants. I, frankly, think it's the fault of policymakers who have to say what's allowed and what's not.

We, you know, whether you look at the various TSA Passenger Risk Assessment programs, and, in all candor, I see the debate over the last, I don't know, 5 or more years of those kind of programs, as a failure to come to consensus about what the underlying substantive rules ought to be. Once we know those, then there are a whole bunch of procedural rules we obviously have to put in place to make sure that we can enforce those, but I think we've got it a little backwards, frankly.

I think we know a lot about good privacy process, and I think that is, in many ways, inspired by the experience of -- well, certainly the experience of fair information practices, and the experience of applying, sort of, auditing and accounting techniques, but I just don't

think we have enough clarity about what the actual rules are. And I can say that, in some part, because we've kind of gone around and tried to find scenarios that we can model, what we find over and over again is it's really not clear what you could do with a lot of this data, one way or the other.

And I think that the problem is that, a) we can't build systems, and b) I think citizens who, to follow Michael's point, I think have -- really have a right to know what's going on with their personal information, especially as to the government -- simply don't know. We just don't know. And that's because we don't know what's on which side of the line.

Mr. Michael A. Aisenberg: I want to vigorously disagree with part of what Danny just said, and vigorously agree with another part of it.

The disagreement is with the notion that what we -- and maybe you didn't mean to say it this way, maybe we actually agree across the board -- I don't think the issue is selecting which rules we want to follow because of the rule's inherent importance. I think the issue is, what's the outcome that we ultimately seek?

What's the behavior that we're seeking to have? And the condition that we're seeking, in terms of -- you know, we can talk about macro conditions -- we want to relieve ourselves, as a nation, from the threat of terrorism? Or, we want to see that all personally identifiable information is only used in accordance with its originator's intended uses and never abused.

I think you have to define certain endpoint conditions before you can select which rules sets in operation will get you there. I think that's one of the great deficiencies that the industry that I've spent 30 years in has. The computer and technology industry constantly wants to have satisfied and happy customers, and so it doesn't much care what the customer's end use is. They just want to make sure that the tools they have to offer will meet the procuring agent's demands and the specifications for whatever that particular customer wants to use the tool for, and they don't care what the output of that process is.

We've got to start caring, as a community, about the output. We've got to start caring about what it is that we're developing that may be much more suitable for

enabling abuses than for enabling protection.

Mr. Peter Sand: I'd like to end with one last thing, and that is, if there's one last thing you could tell a developer or a program manager going into data mining, specifically, if there's one thing you could tell them, based on your experience that they should do, before they start or in the beginning, what would it be? Just a quick one-liner, thanks.

Mr. Michael A. Aisenberg: Well, I'll just follow on what I just said -- have a clear idea of what your actual intended purpose of your program or project is before you start designing it. If you don't know what you want to do with this technology and this data, then you ought not to be building a system in the first place.

Ms. K. Waterman: Going back to something that Michael talked about at the very beginning, I think that one of the really important things to tell a system designer is to make sure that the system has enough available information about what took place or what is taking place.

Mr. Daniel Weitzner: I think -- I'm just building on K.'s point -- that the transparency of

system operations, both to be able to assess effectiveness, and to be able to measure compliance against whatever rules you feel you're accountable to, are really -- really the important thing. Where it seems to me we're kind of in early days of understanding how well these systems work, what impact they have, both on their national security or law enforcement goals, and on their privacy impact, and we really need as much as possible to have both of those measures available to policymakers in order to chart a course forward.

Mr. Brian Tretick: What I'll say is kind of embedded in all of your design -- your systems and the supporting processes, the legal relationships between entities, with the control -- controllable environment and auditable environment -- in mind.

Mr. Peter Sand: Well, thank you very much for a great panel, and thank you for participating.

[Applause.]

Mr. Peter Sand: And we have a break now until 10:30, so please be back at 10:30. Thank you.

[Recessed 10:22 a.m.]

STATEMENT OF TOBY MILGROM LEVIN, SENIOR ADVISOR,

DHS PRIVACY OFFICE, MODERATOR

Ms. Toby Levin: Welcome back to the last panel of the workshop and a significant -- all of the panels are significant, but I like to think of this panel as the most significant among the significant panels, because it will be our effort to pull together much of what you've been hearing over the last day and a half, and really set us off in what, hopefully will be a very -- give us a roadmap for the direction that will lead our office in the coming weeks.

Again, my name is Toby Levin, with the DHS Privacy Office, and I want to introduce you to our panel, many of whom you saw in earlier panels.

To my right, Peter Swire, who is the William O'Neill Professor of Law at the College of Law at Ohio State University, Danny Weitzner, who is Co-Director of the Computer Science and Artificial Intelligence Laboratory, CSAIL, Decentralized Information Group, at MIT.

To my left, Tom Oscherwitz, who is Vice President of Government Affairs and Chief Privacy Officer for ID Analytics, and then Fred Cate, Distinguished Professor and Director, Center for Applied Cybersecurity Research, at

Indiana University, and Barry Steinhardt, Director of the ACLU Program on Technology and Liberty.

The end of the last panel really teed up our panel quite well. It ended with the challenge in defining the rules -- what's allowed, and what's not and how do we do that? We're not going to necessarily be able to define the rules, but hopefully we'll be able to discuss a process for moving forward on data mining in a privacy-sensitive manner at the Department of Homeland Security.

And we can do that at a 30,000-foot level, 20,000, ten, on the ground -- obviously we won't be able to accomplish all of that today, but we will be looking at how government agencies can work with a set of principles -- work toward developing a set of principles -- that can guide data mining research.

We'll begin with an overview by Professor Fred Cate, on what are the principles that may define best practices based on prior efforts, because we're not starting from a blank slate, and then we'll move on to a panel discussion.

So, Fred, we appreciate you starting us off.

STATEMENT OF FRED H. CATE, DISTINGUISHED

PROFESSOR AND DIRECTOR, CENTER FOR APPLIED CYBERSECURITY RESEARCH, INDIANA UNIVERSITY

Mr. Fred Cate: Thank you very much, Toby.

And, a special thanks for recognizing that, in fact, this is not the first time the issue of data mining frameworks and legal standards have been discussed in the nation's Capitol.

There have been many efforts before, many scholarly efforts, many political efforts, many efforts by industry groups, and others, to try to identify standards for data mining -- standards that would allow data mining to proceed in a way that's useful to protect against terrorism, and to enhance national security, at the same time as to protect privacy.

And so, it seems a useful place to start to identify some of the -- what I refer to -- as sort of the low-hanging fruit. These are the places where there has actually been a fair amount of consensus, even if the details have not, in every case, been worked out, of the steps necessary.

But, before starting there, it might be useful just to identify three fairly broad conclusions that, I

think, virtually all of the groups that have looked at these issues have agreed upon.

One is that privacy is important, and we talked about this, at length, yesterday, I'm certainly not going to belabor this now, but that privacy is important for many reasons, and in part, because it is a fundamental condition for most other civil rights to be exercised, as well. And therefore, because of that intrinsic linkage, privacy is worth protecting.

Second of all, that data mining does affect privacy -- that, inevitably, when the government collects or uses data about individuals, it affects privacy -- it may affect it to different degrees, depending upon the type of data mining, the consequences will obviously impact that, but that it does so goes without saying, despite the fact that this has certainly been questioned by some leaders in Washington -- that when the government engages in data mining, privacy is affected.

And third, that it is possible to both do data mining and protect privacy -- that this is not an insolvable, or unsolvable problem -- rather, this is a case of doing the things that have already been identified as

worth of undertaking.

So, in your packets, you have a number of summaries of documents, you have one of the Technology and Privacy Advisory Committee, you have one of Peter Swire's excellent article, which I assume Peter may be touching on later. I wanted to direct you to the one-page, two-sided one which seeks to collect together a number of the prior efforts, and I'm going to use this as an opportunity to focus on some, but not all, in these brief oral comments.

The first -- it would appear, now, that there's broad consensus on, is the need for some sort of standards framework to be enacted in some form of law. Now, that could be a statute, that could be a regulation, that could be an Executive Order -- lots of disagreement about the specific form that it would need to take, but that there needs to be some broad level of authorization by somebody.

And that the purpose of this is, of course, not merely to ensure that privacy is protected, but to ensure that data mining resources are well-invested, and to ensure that the individuals who were involved in these programs are protected. So that they are given clear guidance, and

they know when they stay within that guidance, they are going to be protected by the resources of the government.

And the second on this list of 9 points, of course, is that, then compliance with that law or regulation or set of standard or Executive Order matters. That there is really never a justification for saying, "We're going to act outside of the law." And in fact that's true, certainly, of the law we have today, which already -- includes provisions for emergencies, for obtaining authorization after the fact. That to say we're simply going to cast aside the law, is not an option.

Some of these, let me go through a little more quickly -- it's important to evaluate effectiveness, there should be ways of doing that systematically, there should be ways of doing that continually, so, repeatedly, not just while the system's being developed, but after it's been deployed.

There should be limits on who can use the system, and for what purposes. And therefore, particularly to deal with this subject of mission creep, which we've heard of in many, many settings, but especially over the past two days. There should be clear limits on who and for what purposes

the data and the data mining are used.

Some form of external authorization for specific data mining programs -- and this is the area, probably, of greatest controversy of the many documents on which this draws. Some would say you have to go to a court, you have to go to either the FISA court, or to an Article III judge. Some would say Congress has to authorize it, or perhaps a special committee of Congress. The Technology and Privacy Advisory Committee used this phrase, "An official confirmed by the Senate," in order, I think, to highlight the point that somebody who was both of a high enough level, and of enough, sort of, political sensitivity, so that the -- we did not have the situation in which the same person or group that proposed the system, also authorized it. That we're trying to separate out those two functions.

Working down this list, six, the use of some of the technologies that were talked about in the last panel, that we've talked about yesterday, as well. Data minimization, anonymization -- other technological tools to try to make the data mining compliance with the standards or the framework as efficient and as consistent as possible.

The use of audit tools, both to allow for greater

compliance during the data mining, but I think, in particular, to create a record for going back to resolve incidents, after they have arisen.

Eighth, a system of redress. And the redress point is worth spending just a moment on -- redress here really implies -- includes two separate concepts. One is that the individuals who were affected have a chance to have that impact assessed, and to be -- if not compensated, which would suggest financial compensation -- at least to have whatever benefit or service they've been denied, to have that now provided. So, if you're denied boarding the aircraft, so that the next time you can board the aircraft.

Redress to ensure that we don't continue to affect the same individual in the same way, but also redress from the other side, so that the system learns from its mistakes. So that process of constant iteration that we talked about yesterday is built into the system. And this is one place where, I think, frankly the public is astonished and confused in the ways in which we interact with most systems today, having to deal with security, that, you know, we pull a bag because something set it off, but do we learn anything from that experience? Do we

actually improve our systems, or do we just say, "We're going to continue to make this mistake," every day, and that redress is intended to help deal with that, as well.

And then finally, and in some ways, perhaps most important, the notion of some form of serious independent and rigorous oversight of data mining. And this, again, would be accomplished at many levels -- at the agency level, it would be accomplished through agency Inspector General investigations, which to date, have probably been one of the most revealing forms of oversight we've had -- and also by Congress.

And, in fairness, I think it is appropriate to say that all of these various reports and documents that have looked at the type of oversight that exists today, have tended to be very critical of Congress for its failure to provide either systematic or useful oversight. For example, to identify a single committee in each House to exercise the oversight function, or to exercise consistent oversight, so that it's not within the same month compelling -- commanding DHS to engage in data mining, but then prohibiting the Department of Defense from engaging in data mining. It's very difficult to figure out this type

of inconsistency. And frankly, for the people engaged in data mining, it creates an untenable situation.

So, I think these 9 by no means establish the full range of things that might be undertaken, but they establish 9 broad standards where there is, I think, quite a large consensus that these are important -- I think many people would say these are necessary -- and in this case, it's not a question of needing to invent something new, it's needing to do what we already know needs to be done.

And let me stop there.

Ms. Toby Levin: Okay, one could say that those were a high level, those are high level principles for consideration. So, now I want to ask the panel, are there other considerations that were not included in those nine? Or can we now bring it down a few thousand feet to articulate some more specifics that might be helpful to guide the work of this Department?

Peter?

STATEMENT OF PETER SWIRE, C. WILLIAM O'NEILL PROFESSOR OF LAW, MORITZ COLLEGE OF LAW, OHIO STATE UNIVERSITY

Mr. Peter Swire: So, this is the last panel, and

I've really enjoyed being here all day yesterday, and so far today, and feel like there's been a bunch of learning that might help inform what I'm going to say.

Yesterday, David Jensen and I, in the question session, had disagreements about whether subject-based versus pattern-based was a useful distinction. And we got to continue the discussion at dinner last night, and I think we came to an agreement.

So, at least, I'm going to try to describe it, or at least a way to -- he's at least nodding somewhat, and willing to let me go on for a few sentences.

[Laughter.]

Mr. Peter Swire: The very excellent point that David made yesterday was that at some level, subject-based and pattern-based are both just different data points, different sorts of evidence that get used for your overall conclusion, and so really, this distinction that's been used often about data mining isn't so helpful.

But, I think at dinner we came to, at least -- or at least I'll say it my way and then in the question session, he can say it later -- so there are some things that are very much subject-based, which is the way we've

traditionally done law enforcement.

I get a hot tip that Danny Weitzner downloaded a song on the internet, that's an important social thing to stop at all costs, and so we're going to then go after that individual subject, develop probable cause to grab his laptop, and do the important things that --

[Laughter.]

Mr. Peter Swire: I know, you don't want to lose your laptop, that's a different meeting, we're not doing that one today -- anyways, but that's a subject-based moment. I had a very identified thing about Danny, and that led to probable cause, reasonable suspicion, something like that.

On the other hand, other end of the spectrum, you have pattern-based. I have no names at all, I just have for downloading songs, somebody who has a Facebook membership, somebody who has six other attributes, it turns out, when you do the data mining, that certain people at the end have a high likelihood of having downloaded songs. No names. Those are ends of the spectrum, those are quite different.

You end up, though -- there's some parts in the

middle that are a little different. So, it could be that -- to pick on Danny just a bit more -- that he's had emails and other web traffic with certain people, and those people have a heightened level of suspicion, so that looks sort of subject-based, because it has to do with Danny's network. But it fits some overall pattern of when we get suspicion, so that looks pattern-based. And so we have not the end of the spectrum, but some things in the middle that have elements of both.

And this is similar to another famous distinction in privacy, between personally identifiable information -- Danny's name, address, phone number -- and things that are de-identified -- no idea who Danny is -- and right now, there's a lot of things in between. Like, your IP address is in between and there's big policy debates -- is that identified or not?

So, again, you have from identified to de-identified, you have from subject-based to pattern-based, and a bunch of cases in the middle of the spectrum where it's a little fuzzy.

But we can usefully say, at the end of that, that some things are subject-based -- and that's what law

enforcement traditionally focused on a lot -- and then some things are pattern-based, and I think when you look at the discussions about data mining, having lots and lots of pattern-based -- unless we have a bunch of framework in place, has led to problems, and suspicions and concerns. And so this distinction that -- yesterday there was an argument, "Is this even a useful distinction?" Instead, I think, can be understood as this spectrum between very subject-based, very pattern-based, and we do know that there has been at least some extra level of concerns about the dragnet quality and other problems that come up with pure pattern-based.

So, I'll stop there.

Mr. Daniel Weitzner: So, since -- oh, go ahead, Barry.

STATEMENT OF BARRY STEINHARDT, DIRECTOR, ACLU PROGRAM ON TECHNOLOGY AND LIBERTY

Mr. Barry Steinhardt: I hate to interrupt you, Danny, especially since your name has been taken in vain, here.

[Laughter.]

Mr. Peter Swire: And his computer.

Mr. Barry Steinhardt: And your computer is just hanging by a thread.

Mr. Daniel Weitzner: You'll protect me if my computer is actually --

Mr. Barry Steinhardt: You know where to come if your computer is actually seized, you see me.

Let me say that , also, that I have greatly enjoyed this meeting, I thought it was very useful. I just wanted to add a couple of things.

I thought Fred's summary was very, very helpful, I want to add a couple of things to the summary that I think we need to talk about.

One is the Privacy Act, which I think is in desperate need of modernization. And I know that there are some members of Congress who -- some of them fairly high ranking -- who feel the same way, and I hope the next Congress will take a look at that.

There are two things, in particular, that I would note here. One is -- and a special resonance on this subject of data mining is the third-party doctrine, and this notion that if the government collects data from a third party, it doesn't collect it itself, but it simply --

goes out, for example, to the free market and buys it -- that somehow the Privacy Act does not apply. Even the procedural, or the notice requirements of the Privacy Act do not apply. That needs to get dealt with.

And secondly, we have to come to grips with what's a routine use under the Privacy Act, both for the transfer of data outside of agencies, but I think also, the Privacy Act needs to reflect the fact that we now have agencies like the Department of Homeland Security which has lots and lots of functions, and lots and lots of subdivisions, and the Privacy Act needs to reflect that.

Thirdly, I do think that we need to examine the question of who is going to be the watchdog here, perhaps who enforces the Privacy Act. We need to come to an international consensus on this issue, and have some form of independent official -- a Privacy Commissioner, or a Privacy Commission, or perhaps a much-enlarged Privacy and Civil Liberties Oversight Board, if it ever gets populated and exists, that looks at those issues. So, I do think it's useful to talk about that. I'm not sure whether that's at, you know, 5,000 feet or 20,000 feet, but I would just add those to the mix.

The other issue which I -- which perhaps is a little closer to the ground, here, that I would just highlight is this question of research. You know, one of the things I think that I came away from the last couple of days is that we need to come to grips with the issue of research in data mining and how that can go forward in ways that don't infringe on -- get us the results that we need to get in order to make judgments about whether this is data mining, that any particular data mining proposal is, in fact, useful, has some use in crime fighting, or anti-terrorism, before we move forward with it.

But, at the same time, it's cabined enough that it's understood that it's, in fact, research and not operation, and we certainly -- I can't speak for the ACLU at large, or the civil liberties community at large, but certainly would be interested in talking to DHS and others, about how to make sure that we can cabin the research in a way that's understood, it's that there's an opportunity to both get the fruits of the research and to have a public policy discussion before it becomes operational, but to allow the research to go forward.

As I said yesterday, I think that the first

question that we always need to ask when one of these proposals is made is, will it, in fact, make us safer? Will we get any significant security gain from the proposal, because if we don't get any significant security gain, any enhancement of our safety out of the proposal, then there's no point in even debating the question about whether or not it intrudes on our civil liberties -- why go forward at all?

There are times when research is necessary to try to begin to come to grips with that question, and we need to figure out a way to do it.

One model that was mentioned yesterday that's probably worth exploring here, was the Institutional Review Boards, the IRBs that exist in the medical context. That may be a -- it's a fairly formal, rigorous process, and it may be a model that we want to look at in this field, as well.

Mr. Daniel Weitzner: So, I also like Fred's 9 points, I want to just amplify on one, which is this question of system improvement, feedback, learning.

I think that one of the baseline privacy problems that we have here is that we don't have a good handle,

either from a, sort of expert public policy perspective, from a technical perspective, or from a perspective of public perception, on just how powerful these data mining tools are. We just don't have good ways, technically, to characterize the intrusive or revealing power.

And I think most importantly, we -- if you kind of line up our current experience with data mining with the historical experience of wiretap and electronic surveillance, which is another arena in which we had to have a long-running debate about how much intrusive power we want the government to have, and for what purposes -- we were dealing there in the area of electronic surveillance with really a fantastically simply technology. It was just, you know, "Give me a copy of what you already have."

I think that -- that this really goes to Peter's question about this spectrum between subject, or -- versus pattern-based data mining. I think the reason we get concerned when we leave that hard edge of the subject-based research is that we start to talk about the application of intrusive power that we don't know how to quantify, we don't know how to understand what it is.

And so to the -- people raised questions earlier

about the fact that, you know, we want to have the most analytic power available, of course, for public health research, because we want to find the disease carrier, or we want to find the terrorist who's getting onto the airplane.

But what we don't really know -- we don't have a, I think, an intuitive, or a public policy sense of just how much power we're putting in the hands of the investigators, so as soon as we depart from the -- the sort of traditional model of an investigation that's going after a single person, or a set of -- a limited number of years -- as soon as we amplify the power of investigation with this data mining technology we get, I think, appropriately worried.

We worry, partly, about effectiveness, but I think more importantly, we worry about power. We worry about what power we're putting in the hands of government.

And I think that -- I think we're going to have to feel our way along this spectrum, because I don't think this technology is going away, it's been used in the private sector, as I think was probably discussed, for 20, 30 years. I thought one of the neatest tricks of the TIA program was to pretend it was something new, as opposed to

just something that was going to be applied in a new context -- and I would say that because of that, I think it, it really puts a lot of -- it requires that we put a lot of stress on this, kind of, system improvement and feedback tab, and we have to work as hard as we can to make the uses of data mining as transparent as possible. There are, obviously, security limitations on being able to do that.

But, again, to come back to the wiretap example, we reached some sort of reasonable, you can argue, but we reached some social middle ground, some policy middle ground, about what kind of electronic surveillance was going to be acceptable, and what not, by going through, you know, hundreds and hundreds of cases in open court, and sort of teasing out the question of the balance of power between individuals and government.

And I think we have to figure out how to have a similarly open process for the application of this new kind of intrusive technology, otherwise, we're never going to be able to answer the sort of basic questions about who ought to have what kind of rights, and who ought to have what kind of power.

STATEMENT OF THOMAS OSCHERWITZ, VICE PRESIDENT OF GOVERNMENT AFFAIRS, AND CHIEF PRIVACY OFFICER, ID ANALYTICS, INC.

Mr. Thomas Oscherwitz: You know, I agree with Fred's 9 principles so much that I'm actually sharing his mic --

[Laughter.]

Mr. Thomas Oscherwitz: But, I think one way to actually approach this problem, I think is to actually unpack the word data mining. If you ask yourself the question, what should we do about employing protections against data mining or for data mining, you get certain reactions, I think, often because it's conflated with surveillance.

But taking a definition that Professor Jensen mentioned yesterday, what's your response if the question is, "What are your feelings about an iterative process of learning and probabilistic inference about interconnected data records?" It's just a different connotation when you think about it in that perspective.

And the reason why I say that is, when we're talking about data mining, we're actually packing several

things together. And one of the points that was made earlier in the conference, or workshop, was that there's a collection process, there's an analytical process, there's what happens after you do that analysis?

And so, from the perspective of talking about data mining we need to recognize that it's a technology, and some of the concerns we have are independent of the analytical process.

I'd also like to point out -- going back to the Professor's definition, some of the words, although I can't pronounce them -- iterative. The implication there is that when you're evaluating issues like efficacy, it's iterative, so you have to be able to look at efficacy over time. Efficacy is going to change. So, when you make an analysis of a program, we have to recognize how the technology works.

The second point is probabilistic inference. That deals with issues about what you do when the model is implemented. If you -- this is not a yes or no answer, so when you're evaluating what to do with this information, how are people who are going to get this information going to use that information, and what are the consequences

based on the probabilistic inference?

The other -- I would point out is, not all data mining is the same. If we're talking about an analytical process, even the definition used by Congress gives an exception for fraud. You're using data mining, but you're using it for fraud. And the question about whether or not external authorization and other types of controls -- I think it really goes back to, what are you using the information for? Is it a new use of information? What are the consequences of that information?

So, I'm not sure if it's a one-size-fits-all legislative or regulatory solution for data mining, it also depends on what the information is being used for.

Ms. Toby Levin: Well, I think building on the elements that were on Fred's list, and some of the things that have been added, what we're really looking to do is develop a set of considerations, a set of principles to consider when we're conducting data mining research.

And let me just add some complexity to that challenge is -- several I guess, Danny mentioned specifically, the need for making the data mining project, the activity, more transparent, as transparent as

possible -- well, how do we do that when some of the activities may be in a classified setting? How do we take these principles that we're working on drafting, and deal with the fact that some projects may be classified?

Mr. Daniel Weitzner: I think that -- I want to just connect that question to Tom's point. I -- as you would have heard from the last panel -- I very much agree with the view that we do have to focus more attention on the ultimate uses of this analysis. That's where, I think, some of the important new legislative and rulemaking activity is going to have to come.

As to the question of how we look into what will often be a classified, or at least a sensitive, environment, I think that there are two ways to approach that question.

Number one, Fred mentioned the role of Inspectors General, particularly, the DOJ Inspector General has been able to uncover quite a bit of information about what's been going on, I think that process ought to be able to continue.

So that even if any -- if a given data mining activity, whether it's for research or for operational

purposes, even if that's happening in a classified environment, there ought to be enough information preserved in that classified environment that enables some third party to go in and evaluate both the effectiveness and the intrusive power of those systems.

And that information -- the results of that information, ought to be made visible, to regulators and legislators, and ultimately to the public.

Ms. Toby Levin: But that's after the fact, Danny. What about before? What about the early stages?

Mr. Daniel Weitzner: Well, I think that can be ongoing. I think that tests of systems can be subject to that same kind of analysis. If systems are going to be tested with synthetic data, I think it's actually a great place to start testing what kind of oversight process could work. And it may be oversight over the privacy interests of dummy data subjects, but that's not a bad way to start.

I think that there is a lot that can be analyzed in a way that doesn't reveal sensitive information that still tells us things that we need to know about effectiveness questions, and about ultimate intrusive power.

But again, if we don't start by expecting a high level of transparency of these systems, even if the transparency is only to people who have the appropriate clearances, then we're never going to get past this kind of shadow boxing debate, where we don't even know what we're talking about, we don't even know what kind of intrusion we're actually trying to regulate, or told we shouldn't worry about.

So, I think that's where we have to start.

Ms. Toby Levin: Okay, Barry?

Mr. Barry Steinhardt: I think the real problem here is the overuse of various secret classifications, which has become, over the last 7 years, sort of endemic. Where, virtually everything is now characterized as "sensitive security information."

I've had some experiences that border, kind of on the serial comic. There was a meeting, I don't know, about a year ago to talk about the issue of the fairly aggressive pat-down searches that the TSA was then doing, particularly on women. And, you know, in a moment when I thought I was trying to be helpful, you know, I asked the question, well, you know, "When are you going to move towards some form of

technology that allows you to detect things other than metal? After all, you know, we all go through the metal detector now," and I was stopped and I was told, "Well, no, we can't tell you whether or not the metal detector only detects metal. That is sensitive security information." This, by a high-ranking member of DHS.

And we went back and forth on this for about 10 minutes, how, sort of Alice in Wonderland all of this was. But, the point was, every time we, you know, we try to deal with these issues, we run up against a claim that something is sensitive security information, and therefore cannot be discussed -- whether it's how many people are on the watch list, or whether the metal detector can detect things other than metal. And, you know, we can't even have this discussion until we have a more transparent government. And until we have a more transparent government, it's not possible to say that we're going to have a real public policy debate about whether technology is worth using or not, or data mining program is worth using.

Mr. Peter Swire: So, Toby's question was, in response to the demand for transparency, what about the fact that so many of these things are classified or

sensitive? I had three quick points about that.

The first is, if it's going to be a system that's in the classified space, you still want to have, what I've called, due diligence, by somebody who's not the proponent. So, you have the enthusiast for every project in government, or in a company, and then you want to have someone else who doesn't have a stake in having the program win, at least look at it.

And that may mean something that I think was briefly mentioned in some of the materials here -- there might be a supplementary Privacy Impact Assessment when it comes to data mining. You know, you might have your standard PIA, and then it might be there's an additional set of questions that get built up over time.

The second point is, what would those additional set of questions be, and on this, I'm a little bit of an optimist. So, I think that there's been a social learning curve about data mining. The list of footnotes in Fred's document shows just a huge outpouring things, compared to what we had three or four years ago, about what the criteria are for data mining -- what are the issues, what are the legal structures. And then, the fact

that this room is as full as it is today on privacy and data mining, when people sat here for 2 days -- there's a lot more people who know a lot more than they used to about these issues, and the Privacy Office at DHS and all of the other organizations you represent can come into play, here. So, I think we have a much bigger set of things we know than we did.

The third point is that audits -- as Brian Tretick, I think, said earlier today -- are not just for after-the-fact detection. The knowledge that everything's going to be audited is supposed to prevent misuse -- it sends a signal to people that you're supposed to follow the rules, and it's supposed to, going forward, have closer approximation to what we're supposed to do.

Mr. Fred H. Cate: Toby, I just wanted to add the somewhat mundane point to your question, and that is, in some ways, the classification issue has become entirely a way of, I think, misdirecting the debate. Because we have so many tools for dealing with oversight in classified environments, and it's not like we're using all of them, and then saying, "Now what?" We're not using them.

So, in other words, do we have a Privacy and Civil

Liberties Oversight Board -- which is supposed to have a security classification -- you know, Congress won't even consider the nominations for that. I mean, we don't even have it. It's not -- it doesn't take a genius to say that would be a useful place to start for oversight -- we have internal Inspector Generals who, in many instances, have, you know, been extremely important in bringing about effective oversight -- they operate in a classified environment, we have the Privacy Office within DHS, which I assume, also operates in a classified environment.

We have members of Congress and committees of Congress with clearances who operate in a classified environment. I think on this, the public is really quite understanding. You know, I understand that you're not going to be able to tell me, or publish in the newspaper, a lot of the details about the activities in which the Department's engaged.

It is the apparent unwillingness to tell anyone about it that is the agonizing issue. And I think, exploiting those very, already widely available tools would then make it much easier to say, "But we're not telling you."

You know, one of the recommendations of the TAPAC

was to have external advisors who would have clearances, so that you would take the year-long process of getting them their clearance, they would be ready, and then when you're developing the program, and you need to bring Danny in, or you need to bring somebody in to get a read on this, they would already have gone through that process and be available.

You know, I know that that's done in limited ways today, it ought to be done, in some ways, in more visible -- you know, the ways that corporations advertise, we have Consumer Advisory Panels, you would think agencies would be advertising, "We have the very best privacy advice that we can get," even if we had to get it in a classified environment.

Mr. Thomas Oscherwitz: Just to, sort of echo Fred's point, I do think there's --

Mr. Fred H. Cate: This is my microphone --

Mr. Thomas Oscherwitz: Thank you, Fred, that's -- yeah.

Just to echo what Fred said, I do think some of the benefits that in a public forum you have in terms of expert advisors, or IRBs, that was mentioned by Barry

before, or advisory committees -- some sort of expert process would also be beneficial here, especially in terms of issues like efficacy. You know, is this program actually going to provide any valuable, or useful, or possibly useful results?

Having external advisors not only on privacy, but on whether data mining, in this context, is actually effective, I think would be quite useful, regardless of whether it's a public or private context.

Ms. Toby Levin: Okay, so I think all of that went very effectively to addressing the question I posed.

Let me ask now what we can learn from private sector experiences here with regards to other regulatory models that may be useful in looking for protections, or processes or redresses.

Tom?

Mr. Thomas Oscherwitz: Sure. I think I'd like to make a couple of points here. Going back to that definition that I couldn't pronounce, about iterative processes, in terms of making probabilistic inferences -- one of the things, you know, the private sector has had a lot of experience in is that when you're building a data

mining model, the first model may not necessarily be the most effective. There is a process where R&D is crucial.

And in terms of going forward, there does need to be some space, and it's recognized in some statutes, for research and development. So, one point I would make is that there needs to be some sort of separate process for allowing folks to build the models.

Along those lines, data acquisition and model building is different than data use and model implementation. When you're looking at building a model, you need to have access to a broad array of resources, because you may not know, necessarily, what works and what doesn't.

So, one insight that, I think, we should consider is restrictions on access to information, building the model and the research process may actually limit the effectiveness of the model.

On the other hand, I would also point out that if you look at various statutes like the Equal Credit Opportunity Act, there are societal considerations, even in the model-building process itself, which is, there's a determination made that certain variables are simply off

limits, because the incremental advantage of using those variables is outweighed by the detrimental impact and other aspects of society, for example, invidious discrimination based on race or religion.

And so, there may be a couple of variables that prior to actually building the model, even though I'm generally advocating for broad access to information, where there may be some limits, and there could be some advice there.

So, one thing we have to think about, you know, from the private sector experience is that there may be some variables that need to be off limits, but generally keep it broader.

The other thing I would say is that rules actually do work in the private sector. There are fraud exceptions in a variety of statutes, which allow data mining in that context, and you said data mining in that context has been quite effective, but it's also limited in how you can use that data. So, the private sector, I think, has a lot of experience with purpose limitation, and that's been, I think, a very effective way to deal with some of -- what I consider the biggest risk of data mining,

which is mission creep. Which is, you start with one purpose, and then the purpose expands, and expands and expands.

There's been a number of discussions about building design specs in the model that respect privacy -- I think that's very critical. One of the things that we often talk about in the private sector is something called investment risk, where you build a model, and then the rules change in the middle of the game, and all of a sudden the model that you built is no longer applicable, because certain data elements are no longer available.

So, to some degree, we may not necessarily know what the privacy concerns are going to be 5 years from now, but you want to have a system that can adjust and say, wait a second, this variable is no longer appropriate. I'm thinking here, for example, of the Driver's Privacy Protection Act, and how folks were using that data at one point in the late nineties, and no longer being used. So, that would be the other point I would make on access to information.

I guess the last point I would make is, going back again to the technology itself, which is that, you

know, data mining is policy neutral. And one of the things that you have to do when you're evaluating data mining is compare it to the current system that's being used. You know, airline scanners are already screening passengers -- don't we want them to improve their criteria for screening? Cargo is currently being screened, don't we want to improve that process?

And so, one of the points I would make about the private sector is, we've used data mining in a lot of different contexts to improve current processes, and that's a lot of what this data mining is about.

Mr. Daniel Weitzner: Could I make one quick response? I think it's a mistake to say that data mining is policy neutral. I think that it is -- and by that, I don't mean that it's bad -- but I think data mining brings a qualitatively different set of regulatory requirements to the table. We can't expect to have accountability to whatever set of privacy and civil liberties rules in data mining by using the same techniques that we have for the cop investigating, you know, a murder on the streets, or a drug dealer. And it's because of the scale problem.

And I think you'd probably agree with this, and I

may be taking your definition of neutrality differently than you mean it, but I think that we can't -- we have to have kind of a parity between the intrusive power of the tools, and the accountability power of tools on the other side.

That we -- in that the whole idea of data mining is, in some sense, to enable investigators to discover things they wouldn't otherwise discover by application of their brain, we need enforcement tools that enable enforcers and regulators to discover violations that they couldn't discover just with their brain. That's sort of -- the problem here is the, sort of, scaling up, and I think we have to make sure that the scaling happens both with the analytic power, and with the enforcement power. It's got to go to together, somehow.

Mr. Thomas Oscherwitz: My only response, I mean, I think I -- there isn't that much disagreement at all. I guess I would try to unpack the analytical process from the data collection process where, if you have vast more quantities of information, that definitely changes the calculus.

I certainly agree that the ability to provide

greater insight does have some implications, but again, is data mining that different from a lot of other tools out there that increase -- I mean, the internet, searching -- are we having a workshop on DHS's use of the internet and using search terms? I mean --

Ms. Toby Levin: Not yet.

Mr. Daniel Weitzner: And I would put those, in many ways, in the same category, so I agree --

-- I don't think that -- I don't think what I'm saying applies narrowly to data mining, but I think it applies to this question of the scale of investigators' powers has got to be matched by the scale of regulatory response.

Ms. Toby Levin: Okay.

Mr. Thomas Oscherwitz: I want to go back to your question for a moment, Toby, sort of what can we learn from the private sector -- one of the things I think that we've learned from the private sector over the past few years, is the degree to which the private sector has been, in a sense, compromised by the government. By which I mean that the private sector collects all sorts of data for all sorts of purposes, sometimes under fairly heavy regulation,

restrictions of what can be done with that data, what can be collected, et cetera.

But, all of the laws -- virtually all of the laws -- that govern the collection of that data have these huge exceptions for law enforcement or national security exceptions. And I have to be looking at Peter, here, the very first conversation Peter Swire and I ever had was about the law enforcement exceptions in the HIPAA regulations.

Mr. Peter Swire: A highly specialized conversation.

Mr. Thomas Oscherwitz: Highly specialized, but foreshadowed many, many future conversations with Peter.

But, in general, we need to go back, and I think we need to go back to kind of first principles here, which is to say that, you know, even to the extent to which we're going to regulate the private sector and say to the private sector, "These are the rules that you must follow to protect privacy," we cannot have these huge holes for law enforcement, or national security acts, as we have to go back to the principle that there's got to be a predicate to get that information, in most circumstances, you've got to

go to a neutral arbiter or magistrate to get that information -- we've got to go back to that, or all of the rules we can create for the private sector get swallowed up by this exception.

Mr. Peter Swire: I'm not going to talk about HIPAA and national security, but I have some comments on research, because I think what Tom said and -- it's been interesting -- so Barry, earlier in the panel today said we've had some interesting promise here that maybe we can have a research approach with Institutional Review Boards, IRBs and I think that it may be that a follow-on process or workshop a year from now might be on research and privacy.

But, we've seen interest from S&T the last two days, and I have a few comments on it. One is something that Tom said that makes total sense analytically, but I don't think most people have focused on, which is at the research level, the researcher has reason to have more data, rather than less. It's not a little tiny test bed, instead, if you're going to do data mining research, the logic of the research is you want lots and lots and lots of data so you can throw out 90 percent of it and find the 10 percent that's most useful.

Well, I'll just note that if there's a government program that's collecting lots and lots and lots of data for the first time, that raises a whole bunch of privacy issues, and governance issues, and how's it going to work out. When Jet Blue handed over its research data, it led to a certain level of controversy, so the fact that it's called research doesn't immunize it from controversy.

But I think that if we want to have research going forward, and we want to have better tech going forward which, you know, probably all of us do, it suggests that we're probably going to need to develop some institutional things around that, that won't be exactly like the IRBs for HHS. And so, I just offer a few thoughts.

One is that the research side will be under enormous pressure from the operational side. And just having been around conferences, around DHS over the last several years, you can imagine a conversation that goes something like this: the researchers are doing the research, they have all of this data and somebody says, "Don't you realize the next attack could come any day? It's urgent, we need to get answers now, we can't wait

until this thing's 100 percent done, we have to act now. So, give us what you have now, because we need it, because it might be the only way to stop the next attack." Right? I think a lot of us recognize some variation of that as an important part of the urgency to have Homeland Security to stop the next attack that we've been living under.

If that's the case, then -- when we create, as we create the research side, we're going to need to think about what the research folks say to that urgency request.

One possibility is that there's a huge firewall, and the answer is never, ever under any circumstance -- that will help us do more research, if we say that. But there will be enormous pressure at certain moments to start to get preliminary reports out, or something like that. So, there's going to have to be a whole conversation about when does the research, you know, get accelerated, and move over to the urgency side -- that happens in medical research, right? If you're doing a double-blind study, and it turns out one half of the study people are dying, and the other half people are living, you cut off the study -- you don't watch to see how many more are going to die, right? We're getting a nod from our medical researcher.

So, that's something that I don't think has been highlighted up till now, which is when does research cross over into operations, and there's going to have to be a bunch of learning and discussion about that. And the reason is that, the people who don't want the government to have lots and lots and lots and lots of new data, who are scared about mission creep and all the rest, are going to be very worried that that leakage is going to happen.

And so, I just sort of flag that as a big issue in the research area.

Another thing to point out is that in research and privacy -- something I've been trying to think about over the last period of time -- there's several other legal regimes besides IRBs that are relevant, that I'll just sort of throw out for people to start thinking about.

One of them, which was discussed yesterday, briefly, is at the FDA, the Food and Drug Administration, when you go from Phase I to Phase II to Phase III, there are these gatekeepers, there's people who say, "Yes, we have enough data that it's able to go to the next stage." So, there are some decisions made about what's promising enough for effectiveness and safety, so that you get to do

the next step, so that's FDA.

There's also been a recent report from Berkeley, from the Samuelson Center, about doing research on network behavior -- when people surf the internet and do things, researchers want to say, "Hey, what's going on with people's surfing?" That's sometimes called research, and sometimes that's call wiretapping, right?

[Laughter.]

Mr. Peter Swire: So -- or advertising, as Danny said -- so that's something to think through, so you can look at the Samuelson Center's research.

Another area for research in American law is under the DMCA, the Digital Millennium Copyright Act, there's a security researcher exception under DMCA, which has been often criticized as not a very good one, but it's another place where research has been built into the American legal structure when you're trying to figure things out.

And just one more comment and then I'll stop -- so we have these different legal regimes to look at for research, we see that separating research and operations is

going to be hard, and I'm looking at my notes, here, trying to find the next wonderful point I meant to say, I apologize here, we've all had this happen -- I'm sorry, I can't remember it. So, I'll stop.

Mr. Fred H. Cate: Just two comments, one on the research point -- I absolutely agree, and I think Peter's contribution here is extremely important -- this is one of the critical issues, about how do you do research in data mining technologies, and how do you do it in a way that doesn't, either, get you into political trouble, or into legal trouble.

I think, frankly, all of us who live in the research environment face this all the time, I mean, most -- much of the research we do involves deception, or it involves wiretapping -- it involves something that brings us into a legal conundrum, and frankly, our own research infrastructure is not that good at dealing with those issues. You know, if it's wiretapping, it's wiretapping, you know, there's no large research exception for, you can conduct surveillance if your purpose is research.

I do think the research discussion suggests, though, again why the type of framework that we've talked

about earlier, and that many of these prior groups that have looked at this and talked about, still works pretty well. Because you still need some sort of regulatory framework, you know, for most of us it's IRBs, required by Congress, that's why we all use them, not because we dream them up, but because we're required by law to use them.

We need some sort of limits on, you know who gets access -- these are the types of questions IRBs ask. We need some sort of authorization before we can go into action, whether that's the FDA as a gatekeeper -- so the same types of things you see in a framework for data mining generally are likely to work in the research environment, as well.

The other point I wanted to make, really going back to Danny and Tom's interaction over the question of the scope or tools for dealing with the scope of data mining, you know, one interesting parallel which does not come from the private sector, but the public sector is the IRS. So, the IRS is probably the ultimate data mining organization; it gets this massive amount of data, some of which is, if you will, volunteered by the taxpayer, the vast majority of which is collected from third parties.

And then it combines it all together, and it sends out notices saying what you owe that you didn't think you owed.

It is required by law when it does that to send out a copy of the Taxpayer Bill of Rights, to give you access, it must have a -- not really a Privacy Office, but it has a Taxpayer Advocates Office, it has to give you information on how to contact that office, so that once -- so you get the bad news, you get the, "You've been caught in our data mining framework," and then you get a bunch of other things along with that.

And, I'm not even saying it works particularly well, but it is a step -- I mean, these are tools in the right direction, so that at least there's some response for saying, if I think it's a computer error, or if I think I've been misidentified, if I think the rules are being applied unfairly, there is some recourse set out in front of me, even before I make the payment.

Mr. Peter Swire: Can I just follow-up? I remember the last point, and Toby was gracious enough to allow me 30 seconds.

So, the last point about research is, if we create a research exception, then everyone else will try to

have their stuff look like its research. And this is a familiar bureaucratic thing -- if it turns out that being called a turnip gets you money, then you'll suddenly be a turnip lover.

And so here's the way it played out for research for HIPAA, when we were writing the HIPAA research rules. So, research -- when you think of medical research, you tend to think about, you know, solving leukemia, somehow, you know, saving the -- coming up with the clinical trial that's going to solve the next disease. When some people -- let's call them pharmaceutical companies -- think of research, they might think of that as, "Let's be -- let's call lots and lots of people at home and see which of our sale pitches gets the most new drug purchases." And that's research -- that's marketing research, but it's research.

And so, when you want to have this, it's wonderful, we love it side of research for medical things, a lot of people didn't want to have that enabling a whole new round of data mining by pharmaceutical companies to annoy people at home, and so there was this effort, then, you have to draw this line between what's the good research, or really research, and what are people trying to

shoehorn their way into the definition for other purposes.

So, just because somebody puts a label "research" on it, doesn't mean it's going to be something everyone decides is a good thing, something else you have to watch for as you define your research rules.

Mr. Daniel Weitzner: Just quickly, I think there's a pretty important strategic choice that, in many ways, is presented to the national security community, and those in the government who would use these technologies.

There's a way to approach this research question in a kind of a defensive mode that says, "Well, we'll sort of do enough to satisfy people who are grumbling at the edges, and we'll come up with some studies that establish effectiveness and, that's it."

And then there's another approach that says, it's actually important to develop a sense of public understanding and trust, here. Not only is it important, presumably, for the users of these technologies to have them be effective, I think that's going to be kind of a self-executing, self-actuating sort of process, but it's also important for people to actually feel good about it. I think that's what's, to a large extent, happened in

medical research, people have a sense of confidence that double-blind studies don't go on killing people when they can stop.

And though there are arguments about that around the edges, there's the basic trust in the -- at least in the kind of medically driven research community, pharmaceutical company-driven research may be slightly different -- but there's a basic sense of trust that's been established, so that we feel good about public health research, and the public health process, in general, and we give that process, as we discussed earlier, phenomenal latitude, from a privacy perspective.

And I think the question here really, is how are we going to generate that overall sense of trust? So that the public, regulators, others are going to accord the latitude that I think is obviously going to be necessary, because this is all -- these are new investigative techniques, it's new technology -- at least this new application of technology, and so I think it's great that you've raised the question. And I hope that it somehow gets raised beyond the level of just what's the minimum requirement here to satisfy those who are pounding the

table, but more, what will actually persuade the public that this is worthy of trust.

Ms. Toby Levin: Well, that's the area I really wanted to move into next, because those of us who work in the area of privacy have said for years that, you know, "Don't view," when we talk to our customers, whether it's in the private sector or on the government side, or the agency within which we work, "Don't view privacy as an obstacle, don't view it as a barrier, don't view it as a pain in the you-know-what, rather view privacy as a way of gaining trust."

So, to what extent can we demonstrate to the researchers that going through the processes that we're discussing this morning, which clearly will be some -- time-consuming process. It's not merely a checklist, at least that's not how we want to do things at the Department of Homeland Security, do we want to do something meaningful, how do we explain this to the research community within the Department, and to the senior leadership at the Department that this is worth doing?

Mr. Barry Steinhardt: I'm not sure if this is one where you appeal to people's best instincts, or their

most self-protective instincts.

You know, I mean, in the latter vein, one answer to that is total information awareness program, right? Which is to say, if you don't do it, it's going to come back to bite you in the rear end. And, you know, the Congress is going to get involved, the advocacy community is going to get involved, the press is going to get involved, you're never going to be able to explain what you were doing, and why you were doing it. That may be kind of the real politic answer here.

On the -- I would hope, I'll sort of, just tell you what I think is sort of the better instinct here, which is to say, "Look, we all want to do something that's going to be effective, we all want to protect the homeland, so it's important that we do the research, a) to demonstrate effectiveness, and b) to find out if it's effective, how we can do it in the least intrusive way to preserve our way of life, our liberties," now it's known as the good angel, but I don't know how often we get to think about our good angels.

Mr. Daniel Weitzner: You know, there's a lesson from human -- from computer interface design, it's kind of

an aphorism from Ben Schneiderman, who's really the founder of the field that looked at what are -- how do you build good interfaces to computer systems that people can actually use and understand, and -- Ben has a bunch of principles, just like Fred's principles, but -- the first principle that Ben has is he says, in any computer system, the interface to any computer system, if people can't see a function, they won't use it, and they won't understand it.

And I think that, it's a real challenge here, but we talk about issues like redress and feedback and correction in systems, and I think someone was saying earlier, the problem that we have with a lot of these threat detection systems is that people -- the average person's interaction with these systems is really bordering on the absurd. And I think that the challenge that these systems are going to face, is that people, overall, are used to a higher and higher degree of information access, and information interactivity -- we're used to that from our banks, we're going to get used to that, I think, more and more from the medical system, we're used to having a much -- a dynamic relationship with information systems

that we depend on -- you can pull them up on the web, you can -- in the best case, you can correct things, you can fix things -- sometimes they're frustrating, but there's an interactivity.

The problem with these systems is they seem to be buried behind many layers of protection, and there are some good reasons for that, but I have a feeling that in the long run in order to build a sense of public trust, that's going to have to change. Because it's going to become more and more -- these systems are going to become -- are going to seem more and more anomalous, and frankly, more and more threatening to the extent that you can't even go to an airport and find out why you keep getting stopped for a secondary screening.

I don't want to minimize the challenge of doing this -- the IRS is a great example, I think. The IRS, I think, finds over and over again, that they get increased compliance and more effective processing of claims when they're more helpful, provide more data to taxpayers, and make the process easier. And that ultimately makes people feel better about a system they're part of, as opposed to alienated from it.

And I think that in and of itself, is a subject for research to ask, how can you accomplish these goals in a way that still provides people some reasonable amount of feedback?

Ms. Toby Levin: Well, yesterday we really started to get into the complex issue of data collection, and the many different sources -- private sector, public sector data, data that was collected for one purpose, but now is being used for another purpose, and the issues, then, of building trust when you have such complex data inputs. And it isn't possible, really, to go back to the individual and say, "You know, the information you gave me, you know, 3 years ago for a particular purpose," maybe it was FEMA, you're a Katrina victim, and we collected information in order to provide you with a benefit, you know, do we need to go back to individuals in order to let them know, "Now, we're repurposing your information and we're using it for, in order to develop data mining models, or to develop -- or to be used in another context." It's quite complicated.

Mr. Peter Swire: Well, that's where HIPAA is -- offers you some hope, right? So, one way you can do

research under HIPAA is by consent of the patient. Another way is you can get the IRB to approve it, and you don't need individual consent. And so -- and the reason is, that there's this process that exists where -- with what you hope are expert people -- and that goes back to the question a moment earlier, which is, how do you explain to researchers why they should go through the pain of filling out these forms, or whatever it is?

And I think part of the answer is what Barry said, which is, if you don't you'll get your whole program blown up and defunded by Congress.

But, I think there are somewhat subtler, and additional, points to make to that -- that's a good one to get people's attention. Basically, you say, how are you doing at getting the really, really big databases today? And the researchers will say, "Well, we've proposed 5 and none of them have been approved," right? And so the path forward for research databases, is probably to have a process with guarantees around it, and then once you build those processes with guarantees, you hope that the IRB or its equivalent becomes a source of expertise -- it helps to shape what the researchers ask for, so they know if they do

it this way instead of that way, they're likely to get approved, so you start to build up some experience of what works. And if they see, over a period of time that this is a path to getting things approved, then that's a much happier outcome than having this risk that the whole thing will suddenly become a firestorm of criticism.

Ms. Toby Levin: And I think you probably would agree that once you lose the trust, it's very hard to regain it. So if you've got, you know, a lineup of projects and you lose the trust on the first one out of the gate, it's going to make it hard to get the other ones out.

Mr. Thomas Oscherwitz: And just to sort of follow-up on what Danny said, I do think there is a value in both the implementers and the public having a better understanding of the program. And I'm just thinking of an example here, where you have a -- predict a model where people are flying because on one airplane 5 people bought one-way tickets. And what that suggests, if you understand the model is that, you know, you can have a false positive that doesn't mean the model's not working, it just identified an outlier. So, the person who's implementing that model would say, "Well, I have no idea whether you

have a perfectly legitimate reason for being an outlier, or you're a bad guy, but I'm going to check you out," and the person on the other sides needs to understand, "Well, there's nothing about my innocence or guilt that's causing me to be checked, it's just a model."

So, I mean, there has to be some understanding of what a probabilistic inference means -- it doesn't mean that you're more likely to be a bad guy or not, it just happens to be this is the way we're checking. And I do think that without transparency and people understanding what being checked means, there's going to be a lot of discomfort and anger.

So, we really do have to give people some understanding of how the models work.

Ms. Toby Levin: Let me -- that's probably a terrific example that bridges us into the next topic, I want to get into a little more detail, on the redress process. Because if people are -- if people's information is being used in a data mining project, how do we address redress when they may be very, either very removed, the data collection was very removed from the activity that's being -- it's underway. Or to use the example in a

screening program, where there may be a model -- maybe it's random screening that's occurring, but people are being caught up, and then there are a lot of assumptions made that, you know, "What's wrong, why am I being singled out?" And so, developing some understanding about a well-designed redress program to help address that. But in a data mining context, it's very difficult.

Does anyone want to take on the redress issue?

Mr. Fred H. Cate: Not really.

[Laughter.]

Mr. Fred H. Cate: But first of all, let me see, I don't think we can in any way overstress how important redress is. In part, for the reasons that Danny and I have already talked about, that it serves at least two functions. I mean, one is redress to the individual, and one is feedback to the system.

But also because redress is an opportunity for some of the transparency we've talked about before. You know, one of the things I think we see from privacy in the private sector, from privacy issues, is that the public almost never uses the rights that they have, but it's the knowledge that they have those rights that's very important

to building their comfort level or their trust.

We know this about privacy consent notices, we know this about breach notices, we know this about all sorts of opt-in and opt-out opportunities, that it hardly matters what the mechanism is, they're not going to respond to any of them, it's that they have had the chance to respond.

So, one I think really clear lesson from this is that no program ever should be rolled out without having an incredibly visible redress mechanism, even if ultimately, the redress -- because of classification issues, or because of the data mining challenges -- has to be thinner than we might like on this panel, you at least want to have it. You know, you at least want to start with it as, you know, that would be a leading point to have.

Second, and this goes back, partly to our discussion about the classified information -- often it seems when involved in redress in areas like airline security, that there's a certain amount of unnecessarily hiding the ball. That really, national security is not going to be injured by the fact of giving me a general description of where the data came from. Did it come from

the private sector or not? I mean, if that's going to key the bad guys in, so that they can then get on planes in the future, it's completely lost on me on how that is.

So that, the more that we can provide some information, even if we can't provide all information, we can avoid that sort of unfortunate circumstance of, "I can't tell you. Sorry, too bad. Go away." You know, we can engage in some sort of dialogue, even if the end result is still, "I can't give you the specific piece of data you would like to have."

So, again, it's one of the things I would say, and maybe taking from the prior panel, we shouldn't let the perfect be the enemy of the good; and, frankly, here, the failure to do at least the good is one of the certain ways to bring down even a fabulous system.

Mr. Thomas Oscherwitz: I think that's a really critical point. And from -- as far as I can understand, I think a lot of the -- I think there are a lot of untested assumptions about what amount of redress is possible and what isn't, given the constraints, you maintain a sense of national security. And these are really technical questions, these are, in a sense that these are data mining

questions, also, these are questions about information flow -- how much information can you reveal to a particular subject about the operation of a system, before you compromise the whole system?

I think that unless there's a whole lot of classified research I'm not aware of, I don't think this is a question that's been studied much in its application to redress systems, and it really ought to be. People who -- not me -- but people who do cryptographic research, and do research in information flow, can answer these questions in relatively formal ways, and we could figure out what kind of redress is actually possible, rather than just saying, "Oh, no. If we tell you, we'd reveal X, Y and Z."

And so we should get serious about this -- this is, you know, it's probably a little subspecialty of various parts of information theory, but it's one that I think really could be worked on quite profitably. And rather than having this kind of knee-jerk reaction that says, "Sorry, we can't tell you," there would be a more systematic and trusted way of saying, "Well, this is what we can tell you, and then there's stuff we can't." But we ought to have a better basis for that.

Mr. Barry Steinhardt: I think the first thing you said about redress is we need to have it. And it needs to be real and it needs to be substantive.

The -- you know, we have of course the one example, sort of glaring, staring us in the face, here, which is the airline passenger system, where you have, however many people there are on that list, whether it's 400,000 or a million, have lots of people who are impacted by that. But there is no real redress, here, for someone to get off that list, and whether they were put on the list as a result of data mining or data matching, it's a little hard to know.

One of the things that's missing that's essential to any redress system, is an independent arbiter who's making the final decision. You cannot have redress systems that rely on the same people who made the operative decision in the first place, doing the review and making the final decision, you need to have an independent official.

And we have to be careful about these claims, that somehow, you know, if we have an independent official that you know, the national security secrets are going to

be revealed, or that the bad guys are going to take advantage of it, as if, you know, Osama bin Laden is going to go into Federal Court in the United States, demanding to have his name taken off the terrorist watch list -- those things don't happen.

But, until there is some degree of genuine due process, here, I don't think there's going to be a whole lot of public confidence in the outcomes of these data mining or data matching experiments.

Mr. Peter Swire: So, the question is, what about redress, and the problem is that we might reveal our secrets, and help the bad guys if we tell too much, that's the way it's usually framed.

There's a saying in computer security circles for, especially for the internet and open source, that there's no security through obscurity. And this is, for computer security folks, this is sort of a well-known cliché -- there's no security through obscurity. The thought is, if you publish your code, or show what crypto system you're using, that you let everybody else sort of whack away at it, you find the flaws, you fix the flaws, you improve the system.

And there's a somewhat different saying in the military side, which is "loose lips sink ships." If we tell you anything, the U-boats will get the convoy and kill everybody. And that's saying that secrecy's going to help security.

So, I've written a couple of articles about when each of those is correct -- when openness helps security and when openness is going to hurt security, so to some extent, you can go look at that. But there's a lot of situations in a networked world, where people get to probe a system over and over again, where if you think it's a secret, you're wrong. Because the bad guys get to probe over and over again, and they get to find out, and it's often that you're keeping the secret from your fellow defenders who aren't able to know what the flaws are, and you don't end up fixing it.

So, the more that people -- if it's a distributed system -- the more that there's lots of different people probing it and attacking it, the more likely the bad guys you care about already know what the flaws are, and so you're sort of fooling yourself if you think secrecy is the answer.

A second point is, that in defensive systems, randomness is important. And so some fraction of the 2-year-olds in the car seats and the grandparents should get searched, because sometimes they'll be the ones carrying the bad stuff, even though our data mining doesn't think that's where the highest risk is. And if you have randomness built in, occasionally people are going to get searched and they're going to say, "Why did I get searched on Thursday?" And the answer is, "We don't really have much to say, except everyone's subject to search, sometimes." You'd expect randomness to be part of the answer, so you're not going to have full redress for the occasional search, but there's a lot more room under the sort of theory of when openness helps security, there's probably a lot more room than we've seen to give redress without revealing sources and methods, without compromising the security that has been the instinct, I think, in the last few years.

Ms. Toby Levin: We'd be remiss, I think, if we didn't talk about one of the points on Fred's list, which is the use of data minimization, and anonymization and other tools -- basically, technology to help limit

information that's revealed or perhaps techniques that might provide another approach to protecting privacy. Do any of you have some thoughts on how we can better use these tools? We had some introduction to it yesterday; frankly it's very challenging for lay people to understand these tools, so -- but I would, I definitely want to make sure we've included that in our discussion this morning.

So, I'm going to -- I'm going to start with Danny, who's been working with tools and technology for decades.

Mr. Daniel Weitzner: Well, I might be the wrong person to start with, because I have to say I come to this with a certain amount of skepticism. I mean, these are -- in general, I think these are the -- these are a class of tools, they're a set of cryptographic techniques that, under controlled circumstances, allow you, for example, to query databases and only get certain -- get answers back, but only to a certain level of accuracy or answers about the overall patterns in the database, without revealing individual identity or without revealing information that can be tagged to an individual person.

And these are exciting, they're sort of

tantalizing technologies, and they have, in particular arenas, particular applications, and I'll let someone else defend them. I'm going to tell you what I think their limitations are.

I think that the limitation of these sorts of tools is exactly what Peter said -- that it's hard to use these tools in open, distributed, network environments. When you can control, very tightly, the universe of people Who are asking a limited set of questions to a limited amount of data, then you can have good privacy guarantees, but when you have lots of people asking lots of questions, and when those same people can go ask questions of other databases that you don't control, then you have a very hard time making any sort of guarantee about how much information will leak out.

So, I think there's a place for these, but they sort of travel under what I think are somewhat -- they're sort of grand-sounding names like "privacy-preserving data mining," and I frankly think that's not quite truth in advertising.

So, I think that there is reason to apply these where you can, but the labels may be a little bit

misleading.

Ms. Toby Levin: Let me ask you, then, in the information-sharing environment where agencies are -- the paradigm has changed from "need to know" to "need to provide" and you have information sources across the Federal government, many different databases, and agencies would like to share information appropriately across these databases; there may be some rules that apply to the data coming from one agency, but does the other receiving agency, then, have to respect those rules -- it gets quite complicated.

Mr. Daniel Weitzner: Right, oh, and I hope someone on the panel will defend these technologies, because I could but I'm just not going to. I will if pressed. But --

Mr. Peter Swire: Aren't you being a little broad in your condemnation of privacy-protecting technologies, here?

Mr. Daniel Weitzner: I'm not condemning privacy-protecting technologies. I'm --

Mr. Peter Swire: Which large sub-set of them are you condemning?

Ms. Toby Levin: Because you demo'd one.

Mr. Daniel Weitzner: No, I didn't use the term "privacy-protecting," I used the term "privacy-preserving data mining," which presumes that you can ask questions of a set of data and only -- and be guaranteed that personal information will not be revealed. And my -- it's a very -- you have to be very careful about what we're talking about here, we're talking about this hope that, in particular, before you get legal authorization to go do a subject-based search, that you could go do a sort of a pattern analysis of a bunch of data and get back something that's useful for your investigation, but isn't privacy invasive -- that doesn't uniquely identify anyone in that database. And my assertion is that there are real limits to the ability to make that guarantee.

It's a useful technology, particularly in -- there are circumstances in which it may be useful.

But, to Toby's question -- what do you do when that doesn't work? So, that to me is the interesting question. Sometimes, if these technologies work, then you almost have no privacy problem, because you haven't revealed any personal information, you haven't learned

anything about any individual, everyone's happy. What happens when everyone isn't happy?

Well, then, I think you have to do just what you suggest, Toby, you have to make sure that as information moves around in a complex, analytic environment, that the users of that data understand what the restrictions are on how it can be used further on. That, if it came from the airline passenger screening system, it better not end up in the hands of the IRS for the purpose of catching someone who is getting off an airplane, who is late on their taxes. And that you do it, not by restricting the flow of the data, but you do it by making sure that you understand the user restrictions on that information as it goes through the process.

Ms. Toby Levin: Let me ask just a clarification, Danny. Is your point, though, about the technologies that you can't ensure that they can't be -- that the identities can't be revealed, by going back and re-identifying, or somehow, actually figuring out who the individuals are?

Mr. Daniel Weitzner: Yes, yes. That you get a little bit -- so, there are, you know, it's a very carefully studied area of cryptography, and I will say, I'm

not an expert in this area, I just read the papers. But the -- again, the guarantees about how much personal information get revealed are very, very carefully scoped, by people like Rebecca Wright who work in this area,

-

Ms. Toby Levin: But operationally, it would mean that identities would not be part of the operational -- you know, people using the data would not have those identities revealed to them, not to say that they -- someone with expertise could not go back in and deduce who that, the identities of the individuals, but in fact, while this data is being used, it would provide some obscurity.

Mr. Daniel Weitzner: Only if it's used in isolation. And I just don't think investigations tend to work that way. Investigators don't tend to ask one question and say, "Oh, good. I now know that there are 17 people who are age 12 in this group, and I know I'm looking at a 12-year-old, but I'm not going to look any further, because I know there -- " the problem is, then I go and say, "Well, what else can I find out about 12-year-olds," and then I find out are only five 12-year-old girls, I don't know why I picked 12 -- 35 -- not to get into

conflict issues --

[Laughter.]

Mr. Daniel Weitzner: You know, I gradually piece together information that -- where each piece of the information is not directly attached to a person, but if I ask enough questions of enough different sources, each of those privacy-preserving answers will all of a sudden become revealing, personally revealing, once I've asked enough questions. And I think that's what investigators tend to do, is they tent to want to -- they don't want to just find interesting statistics, they want to find people.

Ms. Toby Levin: But from a security perspective, the information would be less accessible.

Mr. Daniel Weitzner: On a moment-by-moment basis, yes.

Mr. Barry Steinhardt: But that's not necessarily -- that's not -- I think we're making the mistake here, of you're using, you're referring to data minimization and anonymization as examples of privacy-preserving technologies, in which case I completely agree with you, I think it's been dramatically oversold, and we should be cautious about that.

That's not what I said, and that's not what this Says. It says, using tools such as data minimization and anonymization, so think about the airline security example. We started out thinking we needed all of this information from passengers to identify them. We wanted their social, we wanted everything. Then, we later discovered we only needed three elements to identify them. So, that was data minimization, in action, it worked extremely well, we can still identify the vast majority of passengers with great ease, and now we're not burdening the system with a bunch of other data. That's the type of use of this tool.

So, similarly, in the investigatory context, of course, you want all the data, you want to get your hands around it, and anything that denies you access to some piece of relevant data, presumably raises -- or runs the potential of raising a security issue -- but of course we use data mining in all sorts of investigatory contexts. Where we're not investigating something, we're just trying to assess who should have access to this building, this plane, this facility, what have you, and again there anonymization, or tools to reduce who sees how much data -- even moving the screen that you see the backscatter x-ray

result, so that it's in a private place, rather than in the middle of the airport terminal, I think these are extremely important tools, and frankly, I think they're important not just to the reality of privacy, but they're also important to the public confidence, the public trust in the system.

Mr. Peter Swire: So, Danny was referring, I think, to a literature about how hard it is to de-identify someone, when you make queries into a closed database, and you can also make queries into an open database. And it turns out that a lot of the hopes for de-identification for those people doing repeated queries, it's much harder to de-identify than I used to dream that it would be. There's a lot of results that show it's hard to stay de-identified, when you can keep doing these queries. Is that roughly the -- ?

Mr. Daniel Weitzner: Yes.

Mr. Peter Swire: Okay. So, that -- so, I agree with Danny on that.

Let me be a little more optimistic, though, picking up on Fred's approach on a variety of privacy-related things you can do in the technology, and I have an idea I'd never thought of before, I don't think it's been

proposed for -- putting on my old OMB hat.

I think that there might -- you know how there's veteran's preferences for hiring in the Federal government? What about privacy preference for procurement and research contracts for the DHS? You know, you get some bonus -- maybe there would be a pot of money at S&T, maybe there'd be a pot of money -- very large pot of money, right in the Privacy Office, to do more workshops like this -- but also a pot of money so that there's bonuses in the process to encourage researchers and people proposing these systems to have a clearly defined, privacy-enhancing aspects to their proposals to the government. And if you had this pot of money or Challenge Grants, or whatever, you need some process to figure out who deserves it, and you'd have to have criteria and avoid graft and corruption and all the rest, but the point of having -- we have veteran's preferences, why not have privacy preferences? Because there are so many pressures the other way, that dangling some dollars this way might have a good effect.

Mr. Thomas Oscherwitz: Maybe my vision isn't so grand, but in the process, in building a data mining and implementing a data mining model, there's a lot of ways

that you can minimize data access, whether it's the person who's actually getting the results of the model, how the people are managing it, how the result is actually -- is it a score, is it a bunch of personally identifiable information, there's a lot of ways that you can build processes, where people work on it in an encrypted fashion -- whether or not you can get to the perfect of having no access, you certainly can minimize.

Mr. Daniel Weitzner: So that I don't end on a pessimistic note, I want to --

[Laughter.]

Mr. Daniel Weitzner: I want to nominate my current favorite privacy-preserving technology. I recently heard about this. There is a hospital in Boston that was one of the first hospitals to have a pretty comprehensive electronic patient record system, so, you know, all the tests, all the drugs, everything's all in there. And there are terminals, you know, spaced around the hospital so that everyone can get the data they need, this is great.

It happened that a very prominent doctor at the hospital had a relationship with someone else in the hospital, which ended in an unfortunate way, and it -- for

some reason, then, after that, that doctor was found looking at this other person's medical records. And this was a publicly reported case, the doctor was actually -- probably should have been fired, but he was demoted, substantially, from his august position for doing this.

And the way that he was caught was really interesting, and this is my nomination for privacy-preserving technologies. It happens that the way this system was designed -- this is an old system, it's not a new system -- this system is designed in such a way that when you pull up a patient record, what you see on it, on the first screen, is you see a list of the last 10 people who accessed that record. That's it. It doesn't tell you whether they were -- it doesn't tell you why they -- why they accessed it, it doesn't do anything fancy at all. It just says, here are the last 10 people.

So, this woman's doctor, when he looked at her record -- this is her actual doctor, not the doctor with whom she had the relationship -- her actual doctor saw this record and said, "Well, why did that guy look at her record? He's not her doctor, he doesn't treat her," you know, maybe he knew some other things, and so this is just

a really, really simple little piece of technology again, that sort of makes information usage more visible, and allows for the application of human intelligence and integrity, and really caught what was a very serious violation of privacy rules.

Ms. Toby Levin: Well, that's a good segue to my last question which -- that type of log-in control is something that we look for very frequently in our -- when we do our Privacy Impact Assessments, which are the bread and butter operation of the Privacy Office, in terms of making sure that privacy is part of the activities at the Department.

So, Peter, I think, made a reference to PIAs, and possibly thinking in terms of additional questions or elements that could be added to PIAs for, with regard to data mining, and so I'd like to sort of close our panel today with just -- if there are any other suggestions with respect to the PIA content or process that you think would be helpful, in light of the issues that we've been discussing.

Mr. Peter Swire: One part -- this is a longer conversation is, there's a Civil Liberties Office at DHS, and

there's civil liberties issues about, that have to do with sometimes with effects on different ethnic groups or a variety of other things.

It seems to me, at least worth considering, whether Privacy Impact Assessments should be combined with any civil liberties assessments.

A second point is, now that we've had -- the eGov Act of 2002 said these Privacy Impact Assessments should happen for new systems. I don't think they yet apply routinely -- certainly not at the Federal level, to regulations, and I think that they should, and so there might be some wider range of things that should trigger the PIAs.

A third point is, if you have a PIA, maybe with supplementary questions for data mining, there might be other categories you get experience with, for instance, I think biometric systems would be a good candidate for having supplementary questions for PIAs. And as you develop experience in various sectors, as the Privacy Office, developing supplementary things that don't apply to everybody, but apply when there are certain areas that have certain kinds of issues that recur, I think that that can

be sort of a next generation, that maybe you're on the path for already. I don't know all of the details, but I think that could deepen what we got from 2002.

Ms. Toby Levin: I'd like to thank the panel, and we'll have some closing remarks, so thanks very much.

[Applause.]

Ms. Toby Levin: Oh, I'm sorry, I forgot our question period, I'm jumping the gun.

Okay, so if we have some -- anyone who would like to pose a question?

Ms. Jennifer Schiller: Hello, Jennifer Schiller from S&T. I have about 50, but I'll limit myself to 2, because I'm sure there are others who have them.

Toby had asked what you would say to the researchers to convince them to implement these privacy processes, and I would just say, our researchers are subjected to a variety of mixed messages. From Congress, we get, "Focus on domestic radicalization, focus on suspicious behavior detection, but don't use any personally identifiable information." We get, "Be fast, be nimble, be effective, but jump through 5,000 bureaucratic hoops between your idea and an implementable technology."

So, I would ask, my first question is what would you say to our researchers and to our Privacy Office, to help us work together to navigate that tension?

And my second question is for Professor Cate -- you had talked about standards. And I have a question about standards when it comes to the sources of information. Is information that's Google-able, for lack of a better word, you know, we create research databases, for instance, newspaper articles about terrorist events, that might have the name of like, a U.S. citizen in it. When we create those types of collections of information, is there a lower standard of privacy protection, given that this information is in the public arena, these are public figures, public events?

So, these are my questions.

Mr. Fred H. Cate: May I start with the second question?

I think the answer is yes, although I wouldn't phrase it that way. In other words, I wouldn't say there's a lower standard, I would say the standard is more easily met. And so, if it's publicly available information, obviously for efficacy reasons, you still care about the

provenance of the information, is it useful, is it accurate, is it going to lead you to an outcome that actually enhances security?

I wouldn't, in that sense, think of lowering any of those standards. But, I think in terms of thinking about the privacy impact on the individual of accessing or using that information is much easier -- it's going to be much easier to satisfy whoever's going to authorize that program, that, "Well, this is widely available, it's publicly accessible, it's accessible without charge," that would diminish whatever sort of privacy interests we might think -- and hear, in that information.

If I could just also add, with response to your first question, first of all, it doesn't help a bit, but I could not be more sympathetic. I think the lack of consistency and clarity -- from Congress, in particular, here -- has made everybody's job in the national security arena, much more difficult.

But, having said that, I also think that there are things that can be done to, you know, you've got to operate, even with an inconsistent, somewhat ineffective Congress. And I don't live in a world of sort of enhanced

bureaucracy. In other words, I'm not looking to give you five new forms to fill out before the researchers can get on with their job.

I, in particular, think of privacy and security -- especially in the research environment -- as almost always intrinsically linked. That you're going to get a better outcome in terms of detecting terrorists, if you are also privacy sensitive. Because the same types of questions -- the accuracy of the data, the efficacy of the system, the targetedness of the inquiry -- all of those, I think, are going to help in both.

So I would hope, and part of what we're doing is accentuating steps that you would think the researchers are already going through, but maybe haven't thought of in these privacy-enhancing terms. And that is going to be a practical step for the researchers and the Privacy Office in the future, to sort of re-characterize some of what's already going on, without just adding levels of bureaucracy, to add more.

Ms. Toby Levin: Yes, go ahead.

Chris?

Mr. Chris Clifton: Believe it or not, I'm not

here to debate Danny about privacy-preserving data mining --

Ms. Toby Levin: Right, this is Chris Clifton, for purposes of the transcript.

Mr. Chris Clifton: That's my hobby horse, but I'm here to go after the whole panel, actually. Fred made a point, very briefly, about a second advantage to redress mechanisms, and that is feedback. And then there was a lot of discussion about, well, feedback, or redress mechanisms help you build trust, and it's not even important that they be used. I think the Privacy Office really needs to sell redress mechanisms as a feedback mechanism -- this is the way you make your systems better, is you find out where you're failing, and this is how you do it. And make it easy for people to provide that feedback.

Ms. Toby Levin: I would like to say that our email address is probably used by more members of the public than any other email address at the Department. We do have a very public email address, privacy@dhs.gov, so we, I'm sure we'll be hearing more about that, but I certainly agree with your recommendation.

Next?

Mr. Daniel Weitzner: Could I just quickly respond to Chris -- I think that, I think that is an important lesson from the private sector, that's an important lesson from credit reporting systems, that what we know is that people pay attention to data quality when it hurts them.

What we also know, as Fred was saying, people pay no attention to privacy or data quality when they're just being asked for their name for the 15th time that day, and they need to give it anyway to get the book they want from the website, or whatever it is.

And so, I think really looking closely at how those kind of large-scale systems work and maintain, obviously, not perfect data quality, but do better with the interaction, I think, is quite valuable.

Mr. Fred H. Cate: Just to echo -- and that's not happenstance. In other words, that's a calculated -- you know, Congress made a choice back in 1971 that credit information furnishers only had to make efforts to be reasonably accurate. They would get the rest of the accuracy out of letting, if you will, victims of inaccuracy get free credit reports, get a chance to respond.

Over the past 30 years, we've seen more and more rights added on, you know, mandatory dispute resolution, opportunity to correct -- that's not a bad model, frankly, here, as well. To say, you know, you can only do so much up front, the cost of making it perfectly accurate would sink the ship, but that only works if you then have a meaningful redress system to come back around, to make it more accurate as you discover inaccuracies.

Ms. Toby Levin: Okay, next question, please?

Ms. Jennifer O'Connor: I'm Jennifer O'Connor, S&T, and I'm one of the research program managers that Jennifer keeps referring to, that's had problems with this whole issue. And I'll just kind of make you aware, from our perspective, of some of the issues that have come along.

The issue of, when do you classify something -- I have a program, the program I'm managing, that deals quite a bit with social-behavioral science. We have actually stepped back -- but it's using data mining tools, depending upon how you define data mining in terms of modeling techniques --

Ms. Toby Levin: We're having a little trouble

hearing you, could you raise the mic, please?

Ms. Jennifer O'Connor: We're actually using data mining tools, but we're using social and behavioral sciences. And we're actually doing something that most of the data mining folks aren't doing. We're actually taking the social science and trying to show how you can take theories from a semantic standpoint, you know, based in social science, and apply them to why you would use certain variables in a data mining enterprise. Which, those are two -- that's actual research. I mean, for those of you, I mean, you guys are familiar with data mining and a lot of cases, when you're training an algorithm, there's no theory there, you're just letting whatever comes out come out.

What we ran into, though, unfortunately, was when we went to the Privacy Office with this, we didn't have answers, we had a bunch of theories in terms of what we wanted to test. A year and a half later, after our PTA went through -- meanwhile, I'm trying to keep this funding alive, which our division doesn't have very much of to begin with, because it's social-behavioral science, we're not predicting -- you know, we're not protecting

critical infrastructure, we're not stopping people at the border, we're trying to stop radicalization, basically.

My question to you would be, a) when you're dealing specifically with social-behavioral scientists, unfortunately anybody that works in government seems, for some reason, to think they are a social scientist of some sort --

[Laughter.]

Ms. Jennifer O'Connor: We get asked these questions that it would be great to be able to take to an IRB, instead of going back and forth, and I -- Jennifer you can correct me -- we went back and forth probably 6 times with your Privacy Office clarifying scientific principles in social science that I guarantee you wouldn't have asked of somebody doing critical infrastructure or specific data mining techniques, or et cetera. And there's got to be some balance there --

Ms. Toby Levin: Well, let me explain, though, part of that is because most PIAs are public documents, and they need to be written in a way that the public can understand them, so we look at it, "Can we understand the principles? Can we understand the responses

to questions?" If we can't, then how can the public understand? So, part of it is that.

Ms. Jennifer O'Connor: Then explain the medical community, because the medical community -- if I was a doctor trying to explain why Chemical X and Chemical Y work together to do that, they have a real simple answer, "It stops this." We should be able to give you the same answer. It stops this.

But, we're not. You guys dig down further, and unfortunately some of what we have is fragile data, the stuff that really works, yeah, a law protects it. You know, you cannot stop giving that information because the law says, "You will get this information as a United States citizen." There's some stuff that, we don't know whether it works yet, or not, that's fragile. If it does work, and it gets out in the public information through a PIA, for whatever reason, it's dead, it no longer works. The whole research effort is no longer useful.

And that's the classified, not because -- I have worked very hard to keep my program unclassified, because I strongly believe that people have the right to know what we're doing. And it would be -- it would have been a lot

easier if I went into the classified realm, and didn't have to deal with some of these issues.

And I guess, there needs to be an education, there needs to be an education, both in terms of the Privacy Office's need to understand that they don't necessarily need to understand -- specifically in the social sciences, and some of the human factors areas -- exactly what's going on, because there is science there. There needs to be some trust from the Privacy Office's standpoint that when we tell you we're using social science to actually data model, that that is science.

Ms. Toby Levin: Right.

Ms. Jennifer O'Connor: So, I guess --

Ms. Toby Levin: Well, we have a complex role, because we're also acting as, in a sense, as an auditor at the early stages of scientific research. So, I guess the question is, when you've got projects at very early stages, as she described, what are the ways that the privacy issues can be best addressed?

Ms. Jennifer O'Connor: Like an IRB, when -- I mean, can the Privacy Office have an IRB on call, that actually is an expert that we can go to and talk to, maybe

even at the classified level?

Ms. Toby Levin: All right, Danny? Thanks.

Mr. Daniel Weitzner: I have to be a little pessimistic here again, for a moment, Jennifer. There's an interesting divide as to IRBs in the academic community. The medical community may argue about whether they're effective or not or what they're -- but they're sort of a well-worked out process.

Most people in social sciences, at least, that I know, hate IRBs. And that's because to a very large extent, you have social scientists who are often studying what is, essentially, public behavior, they want to look at people who are just out there in the world. And, you know, you look at IRBs for medical research -- you know, they have to make sure that people are not going to get hurt. And, where hurt is a relatively well-understood quantity.

When you start applying IRB mechanisms to social science -- and I fear this would apply in a lot of cases in which personal information is involved that's not medical -- you have a hard time putting a, sort of getting your head around what the mission of the IRB is supposed to be.

If it's supposed to enforce the privacy rules, and we've already said, "Well, it's a little, you know, we're trying to establish what those are," what is it that we're going to tell the IRB to do?

So, I think that it's quite reasonable to expect an expert group to be able to look at a research program. To expect that the group that's evaluating the research has expertise in the area. That makes a lot of sense. I wouldn't want to get too enthusiastic about just, throw some IRBs in there, because there's the beginning of, I think, a real rebellion, actually, amongst social scientists -- particularly anthropologists and sociologists -- in having their research subject to IRBs, at all, so --

Ms. Toby Levin: So, the question is, how can we respond to help in this situation?

Mr. Fred H. Cate: But there is one, you know, silver lining to your dark cloud there, Danny --

Mr. Daniel Weitzner: I'm usually a pretty optimistic person, I don't know what's wrong with me.

[Laughter.]

Mr. Fred H. Cate: And that is, places where IRBs have worked better for social science researchers, it's

because it reflects a real expertise of the membership of the IRB, including social scientists.

And so one of the things that is clever in the theory -- and occasionally in the practice of IRBs -- is usually they are composed of a pretty broad variety of members, who bring expertise in different areas of the types of proposals they'll be looking at, the theory being that that way there is always somebody who can do the sort of translation thing, that, you know, bring the context of the background of research norms from that field. And then the body, as a whole, can help do the -- how do, do we need to explain this differently, do we need to think about a mechanism for protecting the human interest here? That aspect, I think, would still be quite useful, even if not a formally structured IRB.

Ms. Toby Levin: And maybe Privacy Office staff will be able to sit down with you and learn a little bit more after this session, as to how we can better serve -- or best serve -- your operations and understanding how to do these PIAs better. Happy to commit to doing that.

Michael?

Mr. Michael A. Aisenberg: Three comments from

the -- hopefully private sector perspective on some things that were just said. The first will be on Peter's incentives and preferences point, the second on a possible tool, based on aging of the data and the third on, just an observation about public records.

There has been proposed, a preference for industry, in contracting with the Department and other agencies of government, for industry members from the critical infrastructures of the industry ISACs. So -- the Information Sharing and Analysis Centers across the various critical infrastructures, so the notion of preferences for those who can demonstrate a -- some sort of a privacy-conforming practice, as well, is not off the reservation very far, and I think it makes a lot of sense.

The second point is, in terms of industry data retention practices right now -- even though the cost of storage has crashed dramatically over the last generation, storage is still expensive, and secure storage -- reliably secure storage -- is expensive, and so a lot of people age their data, or obliterate data, or let caches expire through TTL, Time to Live tools that wipe out data at a certain point. And I think for certain classes of PII that

are being collected, meta-tagging data with a time-stamp of its date of origin, and a time to live that would cause that data to blow up at a certain point, where the argument for its use has evaporated could readily make a lot of sense.

The third point is simply one about public record compilations. So, we had an experience in the President's NSTAC a number of years ago -- this is very similar to the project that came out of the Geographer from George Mason University -- a consultant to the President's NSTAC was asked to compile a list of the 20 most important telecommunications and 20 most important information technology physical installations. That research project was conducted entirely from zoning records and public -- other public documents. The compilation -- that list, along with Google maps, in full color, showing exactly where these buildings were and how to get to them -- were put together in a report that was left on the table of every member of a classified session of the President's NSTAC.

Needless to say, those reports were picked up, individually and burnt afterwards, because the compilation

had turned into a very valuable, and necessarily classifiable tool. So, the same thing, in terms of -- this goes back to the point I was making about the different faces of the Rubik's cube -- one party's data mining exercise can become another party's act of war, and I think we have to be sensitive to the fact that the sources of that data can be very benign public records, can be, for example, as simple as a telephone book.

Ms. Toby Levin: Okay, thank you.

Next?

Mr. Andrew Feinberg: Hi, Andrew Feinberg from Washington Internet Daily.

I'm --

Ms. Toby Levin: I'm sorry, I couldn't hear you?

Mr. Andrew Feinberg: Hi, Andrew Feinberg from Washington Internet Daily.

I wanted to respond to -- or, let me ask a follow-up to the previous question on Google-able data. And I believe Mr. Cate had said that there could be a lower standard for using publicly available data in that sort of sense, but when you're looking at records collected by private companies, like Google, a lot of these records,

people aren't aware that they're being made public, or that they're even available to these spiders, these crawlers that gather the information.

So, how do you make that distinction? Because when I give my records to a credit agency or to someone who is checking, you know, for an apartment or for anything, that's -- I know I'm giving it to somebody. I don't know everything that is in that cloud of data that Google and other companies like that have on me. So, how can you -- how do you reconcile that?

Ms. Toby Levin: I think we keep coming back to this question of, you know, of how we deal with the fact that so much information is being collected and used in ways that the public just has no awareness.

Mr. Fred H. Cate: Well, first of all, I just want to be clear, I don't think I said a lower standard, I think I said the standard would be easier to meet, if it were publicly available data.

And, also the question had to do with Google-able data, not data held by Google. So, there's a huge difference there. In other words, if the data is out there where anybody can get it, at any time, for free -- so, the

phone book, the internet, places like that, I think it's harder to make a really strong privacy argument surrounding that data.

If we take the data that Google is collecting about my searches, which is not accessible to anybody anytime for free, I wouldn't make that argument at all. In fact, I would make the opposite argument, that this is data that is not disclosed to the government, this is data that the government would be getting from the private sector, it's not data that's publicly available, and so we would expect the privacy standards to be even harder to meet.

And one of the things that pretty much all of the prior groups to look at these issues, and I think, most of the conversation on the past two days has suggested, is that where data is completely repurposed, so for example, it moves from the private sector where it was being used to complete a transaction, to the public sector where it's being used for some other purpose, the privacy issues become, frankly, at their most acute, and -- for lots of reasons. Because, we have all sorts of questions about the accuracy of the data for this new purpose, because of the sort of the transparency issues, because of the privacy

impact issues.

So, again, it's not at all to be nit-picky, it's to say that the terms make a huge difference here. So, really, publicly available data, to me, would be at one end of the spectrum of, say, concern, and data held by private industry -- I don't mean to point to you, Tom -- and that is not publicly available, and is not normally held by the government, would be at the far other end of the spectrum of concern, in my way of thinking.

Mr. Andrew Feinberg: Can I follow-up, though? When you said "publicly available," and she said "Google-able," to me that means almost the same thing. I know what you mean by publicly available, but to me, if this private company, using some kind of algorithm that we don't know what they're using can find it somehow, and therefore, it's out in the public sphere. And associated with, let's say, my name, someone just searched for my name and this document comes up. Obviously, that is publicly available; we don't know how it was made publicly available, though.

Ms. Toby Levin: Let me just interject a statement with -- in defense of the Privacy Act. I know there's been points made today about its limitations, and

questioning whether it's up to the challenge of the information world that we're living in today, but -- if information, even if it's publicly available -- is brought into a Department of Homeland Security or other Federal agency system of record, into a database where it's used as part of our government operations, that information -- regardless of the fact that it came from a Choice Point, or Axiom, or an information broker, or came from a public website -- if that information is brought into the Federal government, into a database, into a system -- that information then becomes subject to the Privacy Act, and the protections that it affords.

So, I think there's sometimes confusion about that, so once we bring it into the government operations and into our systems, and we're actually making it part of our operations, that is subject to the Privacy Act.

Mr. Andrew Feinberg: And that's any information?

Ms. Toby Levin: Excuse me?

Mr. Andrew Feinberg: That's anything?

Ms. Toby Levin: Right, it doesn't matter what the source is of the information. The fact that it's brought under our umbrella, yes.

Mr. Andrew Feinberg: Thank you.

Mr. Thomas Oscherwitz: Just a brief comment, I think that you do raise a good point, which is that I think there's an incredible trend right now that more and more personal information is going into the "cloud" which is under -- not necessarily under control by, you know, any party, but it's just out there. And I think we're going to see that going on more and more, and I think that has to be addressed.

Ms. Toby Levin: Barry?

Mr. Barry Steinhardt: Yeah, let me just make several disconnected points, here, to respond to various things that have been suggested.

One is, that my understanding of the Privacy Act, Toby, is that it -- the information that comes into the Department of Homeland Security or any other Federal agency is not subject to the Privacy Act unless it's as part of a system of records.

Ms. Toby Levin: I don't understand. You're saying it's not part --

Mr. Barry Steinhardt: I'm sorry?

Ms. Toby Levin: You're saying it's not part of

the -- ?

Mr. Barry Steinhardt: That, the kind of notice requirements that are attendant to the Privacy Act do not kick in until you do something beyond just collect a single piece of data, which is what you're suggesting, here.

Ms. Toby Levin: Well, the ad hoc, there's -- The ad hoc collection of pieces of information has been interpreted to be outside of the Privacy Act.

Mr. Barry Steinhardt: Right.

Ms. Toby Levin: But, when we're systematically going out and collecting information and putting it into our records, that are used either for transportation, or for border protection, or any of our systems – if that’s systematically the way we operate, then that would be included.

Mr. Barry Steinhardt: I think we're saying the same thing.

The other point I wanted to make, though, was about Google. You know, Google is also sort of cited as the reason why we have no more privacy.

[Laughter.]

Mr. Barry Steinhardt: Poor Google. And, but you

know, I think that that's an exaggeration, at this point, for a couple of reasons. A lot of the most sensitive data that people are concerned about -- their medical records, their financial records, are not available -- at least until Google takes over the Medical Information Bureau -- but, you know, they're not available right now by Googling -- there may be little episodic portions of it, you know, you may be able to go and determine that somebody subscribes to a cancer awareness bulletin board or something -- but overall, that information is not yet available for that kind of public searching.

And the second thing is the notion that somehow Google, and some of these similar enterprises, are somehow exempt from, you know, the rule of law. We had a really interesting development recently, involving Google Earth, not Google Earth, I'm sorry, Google Street View, that has these very detailed photographs, right? You know, of street views that show personally identifiable people, show license plates, other personally identifiable information.

There was a discussion about the possibility of that technology being introduced into Canada. And the

Canadian Privacy Commission said -- whether it was going to be or not is another matter -- but the Canadian Privacy Commission said to Google, "Well, you know, we have some laws here about personally identifiable information, and you just can't come into Canada and offer to everybody a view of a street in Montreal that reveals this personally identifiable information." To which Google -- they've had the exact same experience in Europe, more recently -- to which Google's response was, "Okay," eventually, "We understand that. Here's what we're going to do. We are going to blur -- or use the available technology -- to blur the images so that they're no longer personally identifiable."

Now, is that a sort of perfect solution? Can you get around that? No. But it's a pretty good solution. It was sort of -- I took that as an important object lesson which is that, if you have laws that require that privacy be protected, that the technologists will find the technical solutions to accomplish that. And that's largely what we are lacking in the U.S., is that we don't have those laws that require all of these technologies that we're talking about be in place, at least with respect to

the Federal government.

Mr. Andrew Feinberg: That was my next question, if you thought that that would bring a need for the U.S. to have similar privacy legislation, but it seems you answered it.

Mr. Barry Steinhardt: My answer is -- I'm sorry, I'm having a little trouble hearing.

Mr. Andrew Feinberg: I think you answered my next question, which was going to be if you thought the U.S. needed similar privacy legislation.

Mr. Barry Steinhardt: Yes, I think the U.S. has to get with the rest of the developed world.

[Laughter.]

Mr. Barry Steinhardt: And both have overarching privacy laws, and independent privacy officials. It would be nice if we joined the 27 or the 29 OECD nations that have those rules.

Ms. Toby Levin: Okay.

Mr. Andrew Feinberg: Thank you.

Ms. Toby Levin: That will be the last question -- thanks very much.

Well, I hope all of you feel as optimistic as I

do now that we've set things in motion. I think we've -- the Privacy Office is very appreciative of your participation and the cooperation that we've had with the Science and Technology Directorate to move forward in actually coming up with principles that can provide a roadmap for the Department, so that it can do what it's really authorized to do, which is to do data mining to help protect the homeland, and which is also -- which it should do in a way that protects and preserves privacy, because that's, I think, ultimately what we all want when we say "preserve the homeland," it is a homeland that is respectful of privacy.

We will extend the period for written comments to August 15th, so hopefully if any of you have additional information you'd like for us to consider, you'll submit it.

I particularly want to, again, thank our team that helped pull this workshop together -- Sandra Debnam, Sandra Hawkins, Tamara Baker, and Rachel Drucker, who were specifically very helpful in supporting our work here.

And we hope you'll come again to these workshops, I think they're informative for all of us as policymakers,

and as people with an interest in these highly -- sometimes controversial -- but important issues, and we look forward to hearing from you in the future, by whatever vehicle -- whether it be email, or by your attendance.

So, thank you, again.

[Applause.]

[Whereupon, at 12:43 p.m., the meeting was adjourned.]