The Privacy Office
Department of Homeland Security
Privacy and Technology Workshop:
Exploring Government Use of Commercial Data for Homeland Security
September 8-9, 2005

OFFICIAL WORKSHOP TRANSCRIPT

Thursday, September 8, 2005
Auditorium
GSA Regional Headquarters Building
7th and D Streets, SW,
Washington, D.C., 20024

PANEL THREE
WHAT ARE THE CURRENT AS WELL AS THE DEVELOPING TECHNOLOGIES TO AID GOVERNMENT IN DATA ANALYSIS FOR HOMELAND SECURITY?

Moderator:
Mr. Peter E. Sand

Panelists:
Mr. Steve Dennis
Mr. Vasant Dhar
Mr. Jack Reis
Mr. Dan Spar

MR. SAND: My name is Peter Sand. I serve as Director of Privacy Technology in the Privacy office of the U.S. Department of Homeland Security. This panel is going to focus on technology.

The first panel was data, the second panel was law, and this panel is going to be about technology. And we're going to start with a brief introduction around the table – ask the panelists to say who they are, what they do, and what they think about this topic in general. Then we're going to walk through a very short outline in terms of structured conversation - talk about the context in which technology sits and then talk about the technology itself and ways to evaluate the effectiveness of the technology. We're going to shoot through that stuff pretty fast and I apologize to the panelists for doing a kind of a bait and switch. We worked out all topics that we were going to talk about, we are still going to cover most of them in the beginning section. But then listening to the other two panels I thought it might be helpful to tie the stuff that we've been thinking about to stuff that you've already heard. So I passed out illegible notes to every body here with the topic and questions and we're going to bounce through that. And I'm hoping that the panelist feel free enough to have an open discussion and challenge each other and have a great dialogue. We're going to start now and we're going to go to 3:30 when we'll stop and invite you folks to ask questions and then continue to 4:00 o'clock.

MR. REIS: My name is Jack Reis, I'm the President of I2, we provide sophisticated visualization and link analysis software to law enforcement agencies, the intelligence community and military organizations around the world. We believe that the use of technology is fundamental to investigative productivity which is key to the issue of National Security.

MR. DHAR: I am Vasant Dhar, I'm a professor at the NYU Business School. I am also Director for the Research Center Digital Economy Research. I've been involved in data mining primary in the financial sector, that has some interesting parallels with the domain that we're discussing today. I wrote a book on methods of data mining, about seven or eight years ago. I look forward to the discussion. Thank you for having me.

MR. DENNIS: My name is Steve Dennis, I'm from the Science and Technology Director the DHS, where I am the Knowledge Technologies Manager in the Threat and Awareness Portfolio. And I have over 20 years of experience in information processing, information analysis, for security applications and I am very interested in today's panel. It brings together a lot of issues that have to do with governance and policy and how those get implemented inside systems and how you do evaluations and common tasks based approaches to developing technology through rapid prototyping.

MR. SPAR: Good afternoon I'm Dan Spar, I'm a solutions architect at Cognos Corporation which is a software vendor of some of the analytical tools that we might touch upon this afternoon. I'm also adjunct faculty at the University of Maryland and I'm looking forward to getting into a discussion with everyone and hopefully make it as lively as possible so please ask any questions if you have them.

MR. SAND: Thank you very much. We're going to talk about context next to frame the technology, then we're going to talk about the technology itself, and then how we evaluate technology.

I was talking with a good friend of mine in industry a couple of years ago about the role that technology developer play in the larger context. And what she said was -- I should preface this by saying I spent too many of my formative years in Brooklyn so everything I say is exaggerated and a bit a sassy, so this isn't exactly what she said, but this is what I took away from it -- She said, technology is a service industry. Technologists work for other people, they build things that serve purposes. So if you asked for it, we'll build it, and it will work. So any problem you have figuring out what you want is your problem. You figure out what you want, you tell us, we'll build it, and it will work. Which I think is an interesting model, but side steps a lot of the issues that we've talked about here today. And I think also what happens in reality is when the heat rises, the pressure builds, a lot of the details in that relationship drop out. A lot of the intermediaries disappear, and what you are left with is the developer and the client. Nobody doing any kind of project management, or any kind of gate keeping or quality assurance, you just have the guy that can build and the guy who wants something. And there's a leap frog model that's built up, where the technology person will build something, show it to the client, the client will go, "yeah that's really cool, can you do this other thing?" And then the person - the developer goes back takes that idea and improves on it, brings it back to the client who takes that idea and improves on it, and pretty soon you have a run away situation where nobody is looking at the details, where nobody is tracking how the final product sits in relationship to the original request and you know success in terms of development ends up being somebody giving you a thumbs up at the end, not whether it matches a specific set of requirements that was articulated at the beginning. Again this is a little exaggerated and a little crass perhaps. But I think it captures a lot of the realities of the technology environment which is that a lot of the stuff is built really fast, can do a lot of stuff and there's a certain amount sex appeal to the stuff that's delivered because you can actually see it. Things show up on a screen in front of you and it's exciting and it spurs your imagination, and it inspires you, and you want to do the next thing, and you want to build and then suddenly things end up over here when they started off going in this direction. And that presents a lot of the problems and I'm wondering if that is in part creating some of the confusion that we were talking about earlier which is, "this is what we said we want to do, but this is what we're actually doing, how did we get there?"

I don't know that there's a real solution to that, I don't know that it's necessarily accurate, but there are pieces of that, there's a flavor of that in the technology world. That's why I think it's important to start talking about context when you talk about technology to understand where it fits.

So I'm throwing that context out to the group here I'm going to ask the folks here to talk about the context that they see for technology. Where they see that technology fits. Let's just start again at this end and work our way around. Jack.

MR. REIS: I feel like you're throwing a hot potato my way. I think technology development is multi dimensional and Pete talked about one dimension which is a user who has some degree of perspective on what they require, can articulate that in some fashion and define some runaway process that forms its way into something deliverable at some point in time which may or may not have matched the original set of requirements. On the other hand there's a great deal of [inaudible] in the vendor community that anticipates requirements, communicates vision, and delivers products that address those, or match that particular vision. We find that often the folks that we talk to aren't always clear about either what they need, or how technology can be applied to addressing that requirement. And as a result the technology community plays a major role in understanding the processes under which our customers, our users function and how we can address technology to those issues in the context of including productivity in the process. So I think it's a multi dimensional approach. And we're seeing today a much more significant movement - wanting to utilize off the shelf techniques or techniques previously tested and proven, rather than building home grown solutions for individual challenges or problems. I don't know what my colleagues think about that, but that's our -

MR. DHAR: Let me just pick up on a couple of things, a couple of questions I see that were raised from the previous panel which I find are fascinating. One of the gentlemen in the previous panel talked about I think the terms were "subject based mining" versus "pattern based mining." There are lots of other similar sort of terms and concepts used and I was sitting in the audience and I sort of was trying to translate that into a problem formulation as a technologist always tries to do. I was trying to sort of visualize data sets in my head and algorithms and so on and so forth. And I think it's important to have some sort of a picture in your head, and I actually made a picture that I'm going to translate into a thousand words. And the picture really has technology on the Y axis in terms of increasing the levels of sophistication. And it had problems on the X axis in terms of increasing level of difficulty, or increasing the level of severity actually is probably more appropriate. Where you know, parking violations and petty theft could be sort of the small problems. Whereas identity theft is a larger problem and arson, and organized crime and narcotics trafficking is more serious and then you have sort of very severe attacks of the right hand side of the -- organized, you know and high impact crime. So you have this two dimensional grid and so it helps to be able to see that increasing level of severity. So I asked myself, well what's a similar ordering if you will on the Y axis. And you know what I found myself thinking of was at the simplest level you have information integrations, you need the database technologies to integrate and there's all kinds of approaches you can use to integrate information whether you do it in real time, whether you do it implicitly, whether it's actually warehoused, so there's that whole problem with collecting and integrating and cleaning up data if you will. Then you have sort of your classic bread and butter, data mining techniques. Which are the classifiers and clustering. As you look at 90 percent of the application of data mining out there they're basically classifiers, or clusters. And all classifiers really do is they order your data in terms of some variable of interest. So in finance the variable interest might be risk or profitability. In Homeland Security it presumably is risk. So you order your universe in terms of some criteria.

In clustering, you basically group your data into some sort of meaningful clusters and of course you don't necessarily tell your system what that basis is, it has to have the smarts to figure that out.

Those are the basic sort of bread and butter methods that are used in data mining. At the next level of sophistication you have text miners. Now you're talking about you know, mining algorithms that can take unstructured data, classifiers and clusters take numeric data, or you know basically structured data, and organize it. Text miners take more unstructured data, and do more sophisticated stuff with it. Like identifying key concepts, or key entities in it. You can also do more sophisticated things like look at the similarity of two unstructured documents. That's you know sort of fairly cutting edge stuff. We can look at two documents and say how similar they are to each other. And similarity here may not be just based on words. You can have two documents that have no words in common that are still very similar and we have the technology now to actually do some of that stuff that that's really interesting. So that's what text planning does.

We've seen sort of more relevant to the problem you're talking about here. Total sophistication would be like network analysis, social network analysis, criminal network analysis, where you're now dealing with data that has some relationship to each other. That is the data elements are marked independent and unordered like they are in many data mining applications but they're now related to each other and so you sort of start talking about learning that is relational or you know takes into account the structure of the network and you talk about that and how that is relevant to this kind of problem.

And finally just to sort of wrap up quickly before I pass this on. We have some other pretty interesting technologies which I will call, one of them is adversarial learning. Which is you're not really learning in a vacuum. I mean I'm sure most of you used the spam filter and found that your spam filter gets progressively less effective over time. Well that's because the people who are generating spam know what your filter is doing and they react. And they start misspelling words, they start putting spaces and so they fool your spam filter, and that's another interesting characteristic of the problem that we're dealing with. We're dealing with an adversary who reacts to what you're learning about. And so it's important to sort of understand that this is also to some extent the gaming problem, the game theory, really comes into the picture as well and finally you have what I call [inaudible] detection algorithms, or concept drift kinds of algorithms that actually can tell you whether there's something - there's a new theme emerging in the text and you know what the implications of that theme are is another question entirely.

But we do have the technologies now that can look at text over time and actually detect whether there's a concept or whether there's a new concept emerging so that's kind of a really high level look at the space of the problems in technologies and that is sort of what I would use I guess for the rest of the discussion to sort of frame some of my thinking about some of the questions that a few people in the earlier panel raised. And the one I find fascinating is the one about subject based, versus pattern based learning, and you know whether pattern based learning makes sense, whether it can be done without infringing on privacy.

MR. DENNIS: I would just like to follow up on your model here of data mining to sort of what I call the cognitive space. And when people talk about data mining often they're talking about it in a general term as though it's possible to apply general data mining algorithms to situations. But it turns out that that's really not true and I think of data mining sometimes as in this analogy to the Dremmel tool, you know the Dremmel tool has all these different bits, and they're good for different types of shaping and certain types of pulling in your data. And it's very important to think about it that way because what happens to the analyst at the back end of all these data mining algorithms is a huge training nightmare, which winds up involving 25, 30 algorithms that only work under certain circumstances that individuals have to recognize. Now why does that occur?

That occurs because there's not a lot of significant problem definition in the first place when people go after using a data mining solution to one of their problems and they don't look at whether humans can agree on whether the answer to a particular task is true or false under certain conditions. And so what is required is the documentation of the problem space, examination of the data space, an understanding as to what [inaudible] agreement might be before humans before you attempt to model these kind of tasks using some kind of technology. And so that takes us through this textural space that you describe where we have these texts that are similar or not similar and they're complete semantic indexing algorithms that help you determine whether topics are similar to one another even independent of words.

But at the same time those techniques have only certain precision of recall under certain circumstances. So the application is all important, the problem application is an extremely important phase of the problem and a lot of times to bring in what Pete was talking about earlier the customer doesn't spend time in the problem definition phase. And often they see their problem as cast in their current business process, in their current context. And there's not a lot of thinking about how to break open, that this is process and innovate, and change the way that you're thinking about the problem.

And so it is very important to spend some time in a lower cost phase of a piloting and problem definition and composition in order to cast the problem correctly, if you work in S&T for very long you learn that problem definition is 50 percent of the solution. So if you can define the problem correctly and you understand the exit criteria. The satisfaction of the customer, what other criteria that caused this problem to be solved, it's time well spent.

In the cognitive space, there is this thing called model drift, that's how I refer to it. Whether you're modeling humans, once you decide that humans can make the system processing decisions or whether you're modeling just the data itself as some kind of indicator of relevance. It is important to revisit the model, the model does drift from its effectiveness if you're not careful especially with changing data sets. And the program management approach that Pete mentioned where customers have what I call creeping featurism.

A lot of times what I see companies doing is not really selling capability, but as you unwrap the offering that they have what you find is, well we have a lot of smart people with really good ideas. They're not really bringing something to the game in terms of capability. But they are looking for that situation where they can have a long ride of creepy featurism just doing what the customer says. And that actually is a good profit motive for many companies. Off the shelf technology, we heard that mentioned, off the shelf technology means that you're only as good as your enemy because your enemy buys off the shelf as well. So if you want to have an information advantage what you have to do is be able to innovate. And I call it an innovation cycle. Fairly quickly putting in the latest ideas and the newest capabilities into an operational space. Through a laboratory who can reflect properly the operational conditions under which systems will be used, can be used properly to train algorithms and paramaterize them in ways that are not available publicly.

I mentioned before it's important to cast problems in terms of business model that many not exist which means that you need metrics that measure the mission function in some independent way, you have to spend time thinking about the system or the component, or the subsystem that you're trying to build or evaluate or insert into the system. In terms of the overall mission goal, not in terms of the current business process. And then there's a second dimension to this metrics which would be to examine the technical performance of those components because just because you have 80 percent recognition at one phase doesn't really spell success at the mission end of this, and you want to be able to look at the air propagation through the system you want to understand how accuracy at one level, effects accuracy at another level. And how that plays into the missions success.

MR. SPAR: We sure covered a lot of topics today so I hope you've gotten all this down so far. But seriously I agree with Steve quite a bit. Really what a lot of this comes down to is starting off with a problem definition. And I have a teacher that used to tell me, a problem well stated is a problem half solved. And that's one of those things that just keeps coming back into your head later in life and you begin to realize a problem well stated, if you state a problem in business or anywhere else and then you start to attach some meaning, some values, some quantitative measurements of prioritization and thresholds that it is already solving a lot of the problem.

You know, we also use the adage that when you have a hammer every problem looks like a nail. Well there's a lot of really go algorithms out there in data mining, I spent a lot of time with them. Mostly in the financial sector on Wall Street. And there's some great stuff there, but what happens is people sometimes forget to take a step back. And look at what the larger problem is that they're trying to solve. Because it doesn't matter how good your algorithms are for data mining, if you're applying them in a framework that's not going to get you the ultimate answer you need to run your business better. One hierarchy that I saw that was kind of interesting was to try to take a look at data organization and actionability and to structure it on a matrix that started you with data then said, now that you've added more meaning to it and more actionability you go from data to information and then you can go from information to knowledge to wisdom. So you can go data, information, knowledge, wisdom. Which shows that as you add more meaning and context to your data, you can move up the chain of being able to do something meaningful with it.

One thing that also came to mind as I heard the discussion today and I read through the notes in advance. Whenever you're approaching a problem related to data mining, or gathering meaning from data, be as resourceful as possible. See other industries that have done quite a bit of this already, and see where the algorithms are the same. The financial sector has done a gigantic amount of this. Walmart has done a gigantic amount of this. We don't need to reinvent a lot of the wheels just because the data has a slightly different business meaning. Much of the ways that we're going to be able to analyze this will be the same.

Finally, for you to assess the value of an algorithm or the value of a technology, or the value of an approach, you have to have already defined business values of getting something done correctly. And then the value of each approach is going to come down to some form of math or statistical modeling to see how well any given approach solved that problem. So many times we say “What's a better vehicle to drive to work? What's a chalkboard technology to solve a certain problem?” Well until you determine the value of this aspect of your business space you can't start to define thresholds and start to give scores. And so I spent a lot of time in the weeds academically and professionally and my conclusion after doing that for 20 plus years is don't forget to take a step back and make sure the problem is well stated.

MR. SAND: Let's take another step closer to the technology itself. When we were talking in preparation for this panel, one of the topics in terms of context was this idea of proactive versus reactive as context. The choices in technology change when you're in proactive versus reactive mode -- I would also like to talk about that for a little bit.

MR. SPAR: Well proactive and reactive is something we can go right to the finance world and see. How many times do you see those infomercials that show you the greatest way to buy stocks going forward. Well one way to test their quality might be to take their models and see how well they did if you worked it backwards over the past couple of years. So a proactive and a reactive aren't quite as different I think as might initially meet the eye. Ultimately what you have to do is say "what is my approach, what is my algorithm. What is my technology being used to execute this algorithm" and then start assigning quantitative values that might say how accurate it's been. How well did a mode predict the past, you can retrofit the data and see. Now how well is that model going to fit the future. You can also see that with data, and it's already been done in several industries. And it can't be done unless it's being done quantitatively.

MR. DHAR: I'll pick up on that one, because that's very close to home for me, because one of the hats I wear is related to the the hedge fund where we test strategies in exactly that way. You know you have a hypothesis and you run it through the data, and you see how well it does. And the question is where did the hypothesis come from, does it come from the machine or does it come from a human being. In one sense it's almost irrelevant. But the thing you really need for that is you need lots and lots of data. You need lots of reliable clean data. You need lots of instances. There are positive instances, negative instances. You can actually train the learner to do that. So to the extent that we have problems that fit that profile we're in great shape. As a standard learner to test the [inaudible] idea. The limitation of that approach is that we're dealing with problems here, at least some of the problems practically don't have a lot of data. You know we don't have a lot of severe attacks so it isn't really feasible to use the same kind of methodology you know that we might use sort of [inaudible] scenario. And I suspect and I'm purely speculating here, that for problems like that which fall more to into the sort of proactive domain I think we're going to have to relax our choice of methods and let's start looking at networks. Or interactions. And you know that will really carry the win. When you start looking at network interactions you are to some extent engaging in exploratory analysis and that really raises some you know, pretty severe privacy issues. And so the question becomes, how can you really do that in a way that preserves privacy.

MR. REIS: I think we probably look at the challenges in similar ways and yet different ways. Unlike some of my colleagues we grew up in the law enforcement intelligence arena, and did not have large data sets or algorithms to apply to a particular analysis. But we have seen an evolution of process if you will from a reactive analysis to a proactive and defined it to in a reactive sense in the law enforcement or intelligence world an undesirable event has occurred, and now we need to try to put the pieces together to determine what transpired and how it occurred and who might have been associated with it eventually, hopefully prosecute them.

In the proactive arena of course we're trying to anticipate some undesirable event and address it from that perspective to prevent it from occurring. The technologies that have been described are consistent in terms of their philosophical operation but different in the way in which we implement them. We look at analysis whether it's reactive or proactive in multiple dimensions.

First, the issue of information acquisitions is fundamental. So how do we go about acquiring information in an effective manner so that we can do the appropriate analysis and evaluation. And information is available from myriad sources today, manually acquired, acquired through various electronic means, or databases and information produced by various folks both classified and unclassified. So how do you get at all of those sources of information? How do you bring all that back to a common place that you can actually function on that information in some useful manner? So that's fundamental to the types of technologies that are key to being successful in the investigative area today.

Second, once you have all this information how do you effectively assimilate it? How do you make sense out of it for anybody who's been involved in the investigative world in the analytical space? You know that that can be overwhelming. You can literally be inundated.

When the Maryland sniper scenario occurred there were more than 10,000 leads to track and follow. How does any human assimilate that? So technology applies itself to making that assimilation simple using things like visualization like analysis technologies. Once then you have assimilated this information you need to sort out what is relevant and what isn't relevant. It's the issue of determining from information what is really intelligence. Intelligence from which informed decisions can be made. And that's where the analytical technologies apply themselves to some degree.

That's where all the rhythmic technology can be effectively utilized and then subsequently how do you present the result of analysis in a way that those that need to execute or make decisions can do so in an informed manner. And those are the types of things that we find ourselves associated with in a very practical manner. And it strikes me, and no disrespect to my colleagues, that many of - much of what we've talked about so far is really academic in nature in many respects.

What we're talking about here is very tangible technologies in use today, designed to improve and are enhancing the investigative process, and there are many examples of how those technologies have been utilized effectively. That you know when you're in this reactive mode, it's because you're proactive process didn't really predict very well what was going to happen. And a lot of times when you're in this reactive mode you have less choices, you are basically engineering a deep solution to a problem that wasn't foreseen and you'll spend a lot more money and time doing that.

I think we are in the space that Vasant described of hairy problems where we have lots and lots of disparate data sources that can be reached and touched under certain conditions once there is an event or something that wasn't properly analyzed in advance, you're going to start connecting data sources that haven't been foreseen before. So having an approach that allows for network effect, and having an approach that allows for information providers to join a network of information is important. At the same time in order to protect privacy, what you have to have are some sort of policy smart devices that are sitting there both at the data source, and the user end that are capable of discerning roles and actions and the ability to focus to do the link analysis and visualization and in order to control their access to information appropriately in accordance with the law and the policy. If we have the right amount of foresight we should be designing that system now, and I think it's being foretold that these information sharing environment actions that are being taken by the ODNI and are internal to Homeland Security. We can reconfigure the whole system, capable of moving in the direction it needs to under the right condition. And having foreseen the need to connect lots of information well in advance of being given the directive, my biggest fear is that we would have an event and we would be order to connect information and someone will say it will take three years and [inaudible] to do that. It's really not where we want to be. And so - and [inaudible] to be very proactive about it.

A lot of the models that we hear about for financial application and from Wal-Mart for example are highly controlled systems. They basically take point of sale information and take information that Wal-Mart designed it gather. A lot of the information that Homeland Security brings together to make decisions and helps drive decision based organizations is not data that is under the direct control of any particular party. And so we can have another data standards, we can describe [inaudible] a lot of times they wind up as unfunded mandates in the government, it's impossible to move the legacy system in that direction. So what we really need is a smart switch in the middle where the thing that connects are systems under the appropriate conditions and scales to handle the size of data problems that we face today.

MR. SPAR: Well it's good that we don't all agree 100 percent that way we can get a more lively discussion of it. My personal view is that whether it's academic or applied, whether it's Wal-Mart, the police department or Homeland Security, when you're looking at solving a problem to try to predict what's going to happen, you have a couple of layers. And these layers are the same and challenge me on this if you don't agree. You've got a business, it could be law enforcement, it could be air travel, it could be anything. You got a business that operates in a specific way. You've got a hypothesis about something that could occur. Maybe if it's air travel you have a hypothesis people with one way tickets and no baggage constitute a greater threat. So you've got - hypothesis is an academic term but the police officer who just walks a beat downtown, who sees a certain behavior develops a hypothesis that he wants to check out as well. Then you've got an algorithm that you follow to check through the data for many some additional information. And the way that you check through the data frequently is through the application of technology, technology is just a lever. The question is there some shift in activity at the aggregate level that gives you an idea that precedes the event that you're anticipating. And that's something that is like I said; it can't be formulated. It's a difficult problem to solve.

MR. REIS: That presumes of course that there is a great deal of data available for some analysis. That isn't always the case particularly in terrorist based activities. Think back to 9-11 for example, or 20 people taking flight lessons concerned about getting the aircraft in the air, and not worrying about landing or taking off constitutes suspicious behavior, and how would we learn about that. We wouldn't learn it by searching any database, or flying any algorithms. We'd learn it by observation. We'd learn it by a suggestion coming from somewhere that would spawn an investigation because it is perceived to be investigated. So while I believe that there are certainly places for activities that relate to mining large amounts of data, and predicting behaviors as a result of that, I don't think that's going to help us materially in the terrorist world that we're living in today. Because these tend to be small cells, small amount of activity. I mean the terrorists that were captured in upstate New York for example, weren't captured as a result of some [inaudible] mining activity or the application of sophisticated algorithm. They were captured because of some observation and the series of connections that were emanating from that. And the application of technology. And a much more primitive level was instrumental in doing that.

MR. DHAR: Just a quick response to that. I think you're assuming that the only thing we can observe is what we are currently observing. And so you know there could be all kinds of cyber activity going on out there that we're just not observing. We are capable of observing it right now, so while I agree with you that and I started off at the outset by saying the most severe events are the ones unfortunately for which we have the least amount of data, I guess I'm not willing to buy right off the bat that there isn't enough data just because we have not observed it in the past.

MR. SAND: I think what's interesting in this conversation is this last part, is the closer connection between technology and data, you find yourself kind of moving back and forth, between the two. The technology is valuable because the data is valuable, and the data becomes usable when you apply the right technology. I think that's an interesting grid to highlight.

MR. DENNIS: I just wanted to jump in this so I'm not left out. Basically we heard the need to have an approach that enables us to look at the problem and define the problem, and we have indeed done that we just initiated the Center. It's a interagency center, for Homeland Security Technology. It's basically gathering different sets of data that someday might need to be somehow similar, but in looking at what it would mean to have systems that would scale to a National level or beyond in order to effect the analytic process.

I would like to say to that I agree with Vasant, I've often thought over the last ten years even that these aggregate data views could be something that's shareable. Something that doesn't violate privacy, that's capable of communication at least in [inaudible] in different parts of the system. And in order to effect a better collection and analysis of data. But at the same time in Homeland Securities perspective I worry about the response to an aggregate view. For example if there's a threat to a single stage, or a threat to the country. How do you respond to that? I mean it's such a massive amount of space to be covered that the response - the threat is very general, the response is also very general. It's hard to take an aggregate point of view and respond to it. But at the same time I think it's good for cueing analysis, and it's good for information sharing. Not a lot of regulation over aggregate points of view that don't get down to the individual point of view. And then we talked a lot about observations and it was observations that led to this law enforcement action but I think it's also very important to notice that the collection of observations is also an interesting point of view. That you're able to solve crimes that expand geographic regions that you connect together through different law enforcement systems and if you're capable of putting together multiple dots, multiple observations they might make sense as a picture, whereas individually they might not have. So that's another point of view for aggregating [inaudible].

MR. DENNIS: I just wanted to add a short comment which is that it may well be that data mining won't be the technique that will predict the next major terrorist event.

It maybe that the best we can get from it is that we now need to direct our attention to an elevated probability of something taking place, because we've discovered certain actions that could be precursors to a large event taking place. And there's a lot of precedence for this in other fields as well. But as far as the comment about observations and connections well, whether an observation is done manually through technology or any other way, what's the meaning of an observation. You're watching a pattern. The data involved in that pattern has certain attribute levels. You see elevated communications. You see someone who looks like their in physical distress. You see congregations of people with certain attributes in certain places. I mean that's what observation is if you break it down to the stated elements. So whether something incurred to any range of the spectrum of sophistication in its analysis it still is something that lends itself to have an hypothesis, [inaudible] and deciding when you can act. And I'm willing to bet that the best we're ever going to get is an elevated need to investigate something more closely.

This isn't the financial problem, or the terrorist problem. But if you're credit card activity triples three months in a row, even your credit card company elevates the degree to which they observe your future behavior because they feel you're greater risk for either fraud or default. So they didn't necessarily predict that you defaulted but maybe they headed that off by having some sort of algorithm in place that that observation fed into.

MR. SPAR: I just wanted to make a small point in response to something that Steve said, and it is a modest point. The concept of data aggregation frightens some people and there has been in fact examples of significant investments made that - the term itself implies putting a lot of data in the same place, and that worries a lot of folks for a lot reasons. But defining data, doesn't necessarily mean putting it all in the same place. Technologies are available today to reach out to information sources wherever they may exist and extract the relevant bits and bring them back for analysis under the appropriate set of circumstances.

MR. DENNIS: That's exactly what I was referring to and I have already moved beyond ever thinking about pooling the data and one place. It seems like it's a unreasonable idea. And more and more we talked about the network of information that's available and making that highly dynamic in your typical network, that has privacy and security protection.

MR. SAND: I'd like to touch on one last piece. On context. This woman I was talking about earlier her favorite question was, “what is that?” Her motto was “Data is the Center of the Universe.” And that if you understand the data, all the other things work themselves out. If fact when I was leaving working with her to move to DHS, I got her a bunch of pens, that said “Data is the Center of the Universe” so she'd never have to say it again - she'd just hand the pen to somebody. But her question was, “what is that?” And that's also - I have a two year old son - and that's his favorite question too. “What is that?” I think it's an interesting parallel.

But I think it's a very powerful question. Because it forces people to really define what they're talking about. And her focus was on data but I think asking that one question over and over - and she asked it over and over and over again. But what she was able to come up with at the end was a picture. That said: this is what you do, and this is how you do it, and this is what you're talking about. We can draw a big circle around it, and then from there really understand what the business issues are and what the technology issues are, and all that kind of stuff. So she always started with that question and asked it so many times that people basically left the room in frustration. But when she was done she had this marvelous thing. I'm wondering if just off the top of your head do you have a favorite question that you ask that you find gives you the most return in terms of identifying what the context is or placing what you're customer, client, or people you're working with, understanding what they really need, what they're really talking about. Any favorite questions.

MR. SPAR: I always ask people off the bat, “what would you consider a success?” When somebody brings me in to help solve a problem with technology I always back it up, what would be success. Someone might say well we need to discover this relationship to that relationship and I would say “To what degree do we need to discover that? To what percentage of accuracy, how quickly?” Because a lot of people say, “well I want a real time system.” Well what does real time mean to you. Real time could be defined as a transaction speed that's faster, than the transaction speed of the unit of work in the business that you care about. The stock market it's a second. Maybe in some other businesses it's an hour. So I always try to start off by saying, what do you consider success today. Because if you don't know that, it makes it pretty tough then to take a next step from that direction.

MR. SAND: Any other favorite questions?

MR. DENNIS: I am very much process oriented. I like to know what comes into an organization and what goes out. How do the people who provide information care about what they provide. And then how do people perceive the products that go out of that organization. And that mapping of input, output is a pretty good starting point, but it's not the final view at all. Basically from the current operation you can start to understand what the customer means when they're giving you information so for example if they have a database, or they have information resources.

Understanding the semantics of the entries in those data sources is extremely important. How do they view suspicious activity report. How do they view suspicious activity and how does it compare against the statistical range of normal activity that might be going on that's captured by that database. How is that important to the overall product. And getting the map from what comes into an organization and what goes out is extremely insightful as people start to unfold what it is that they do. I think having measures of success is good. A lot of times the customers don't see what could be done taken in the context of their current business model and sometimes they're not open to envisioning something completely different. And off the wall. I do like this question, “what is it?” And I've been involved in meetings involving language processing and linguistics, and its gone on for days where that question was asked over and over and people weren't allowed to stand on their egos, and at the end of three days, we had beat everybody down and had an agreement.

Does that mean we solved a problem? No. Does it mean that we really mapped out the answer or the problem space? Not really. But if you ask that question over and over you can burn people down to some consensus a [inaudible] point of view.

I think it's not about the data. I think it's about what the data means. And the data has to be in context before it can be used. I've been to organizations before that have multiple view points of their data and there's not even a consistent view within an organization of what the data means. So getting to that is an important challenge, it's often difficult because you're trying to describe the [inaudible] of humans with data, it's very hard to get to that, which puts an additional leverage against the problem space to say that these systems have to support multiple view points. It's often not true that people agree. And it just puts an additional level of complexity on the system that has to care about individual points of view. Especially when you get tens of thousands of [inaudible].

MR. DHAR: Just a small addition to that, I guess my favorite quote is, "there's nothing better than a good theory." You know just to get primed and that's certainly been my experience in the financial arena, I spent you know many years in banking on data trying to find patterns and sometimes I found patterns that I had no idea why they existed. And you know, so you go back to the theory and you look for reasons why they exist and sometimes you find them and sometimes you don't. But the key point here is the hypothesis space is potentially infinite. And you know the nice thing about human beings is that we do have a intuition. And you know quite often we serve our in the ball park. But we really can't sort of solve the problem by ourselves. But if we just get into the ball park, then key learning does just wonders for you.

MR. SAND: All right. We're going to move to talk more specifically about technology. One of the things I was interested in hearing in the legal section was this apparent philosophy of kind of data mining, love it or leave it. That data mining is the thing, it's the only thing. And either you love it and it works, or you hate it because it doesn't work. And I'd like the panelists to talk specifically about technology that would fit into the general discussion of use of commercial or large scale, or just for data sets. And I'm wondering if the topic is - you know there's data mining and that's how you approach large or disparate data and then underneath that label, data mining all these disparate things, or are there different technologies that actually have nothing to do with data mining.

And as we talk through it, it would be helpful if you label it, describe what it does and then what's important about it. And also I think it would be good for the audience to hear some of these terms. The first panel some of the terms that came out were fuzzy matching, advanced pinning, and I think some of these technologies have really fun names if nothing else it would be fun to say them in public. So if there are things that you guys are familiar with that you work with regularly, or if you just want to break down this big kind of blinding term, "data mining," and just throw out specifics and lets talk about the detail. Anybody.

MR. SPAR: Well I can start with the easy part, which is the very top of the pyramid, and then we can get into more specifics. But basically what a lot of organizations want to start with is something that is a step short of data mining, which is just data analysis. Where you can use some technology just to put together what you might look at as a cross tabular reference. It could be two dimensions, it could be many dimensions, where you start seeing if say you were a police officer if I were to break the day down into blocks of three hours and I were to break crime down into levels of severity let me look at these two dimensions and see okay. Just based on my data not doing any mining, just an analysis what times of day seem to be correlated the most with the most serious crimes. And then you might add another dimension to that, for location. And then you might say, okay cross these three dimensions in what locations, to what times of day do I seem to see where the most serious crimes happen. The good news is there's a lot of technology out there that can bring basic data analysis where data is pulled from many data sources to an end users desk top, with a minimum of training. So step one might be to just get that capability. I've seen a lot of organizations that have two or three super deep in the weeds data miners who are using their own networks, and using very complex algorithms trying to make discoveries and I've seen organizations who don't have that many people doing that and but have the capability for the basic analysis put into thousands of end users. I've seen a lot of those organizations that just put that basic capability out, get more ideas from that. So it's just something to think about, you don't have to be a mathematician to get started you could just start at the basic analysis level.

MR. REIS: I think that's an important point and the question in my mind is what really is data mining when I hear the term, I think of huge data sets and sophisticated algorithms and so forth. And yet, take the Metropolitan Police Department right here in DC, who does precisely what they describe. It's used periodically as a COMSAT Center looking at the frequency of crime. What is the highest frequency event. What day of the week are they more frequently on, what time of day. What are the patterns to that, so that I can allocate my resources in a way to reduce crime. It's a relatively simple application. Just map the number of crimes, and the types of crimes that happened in the district. So, could I track back through maybe 1000's of connections to the place where the virus originally emanated. Now is that data mining? I've got 1000's and 1000's of records. I'm looking for connections between and among those records. It's a process I could not conceivably do manually, just simply because of the number and the nature of the information. But, through technology could do in a matter of seconds. Is that data mining or is that data analysis?

MR. SPAR: You crossed the line once you looked for the correlations in a more automated fashion.

MR. REIS: But, you know what I'm trying to impart is a very practical applications of technology. The very real problems that occur on a daily basis on the law enforcement intelligence community today. There is, in my view there is certainly application, and there are agencies, more in the classified arena working on larger data sets, vastly larger data sets and much more sophisticated analysis or data mining, if you like for predictive purposes. But, there are two ends to the spectrum. One, is the very real and tangible on the street today. And the other is, I hesitate to call it more theoretical because it has a very real and practical purpose. But far more challenging in it's nature.

MR. DHAR: Nothing really to add. I guess I really don't care what we call it. You know, sometimes it's analysis, visualization, automated search, pattern discovery, social network analysis. It really doesn't matter. I think the important thing is to recognize that, you know we have this sort of chest of tools at our disposal. And that collectively it forms a pretty powerful analytic capability. And so, it doesn't really matter what you call it.

MR. DENNIS: So, I think it was 1989 when I first heard the word data mining. And I had a group of mathematicians that were working with me at that point. And I went to them and I said, what is this term data mining and how is it different from the statistics and probability that we normally use when we solve our problems? And basically they said the difference is if you're trying to start a new program or you want to get funding, or you're trying to get business you're going to call it data mining. Because that's basically the marketing term that all these techniques have been put under. Which gave me an interesting fresh perspective as that started up.

And having been through the neural network phrase of 1985 to '87, somewhere in that time frame. And all the funding that was poured into neural networks, you know you can sort of see the pattern coming. That in order to revitalize funding, in order to revitalize science in that area you call it something new and then you change the framework and move on. I think a lot of the techniques that were discussed here at the table are all in that area of data mining. We sort of, in the community been calling it knowledge discovery, there's a new thing called video analytics, you know. There are all kinds of ways to describe these techniques that get used against data. But in the end, how are we really advancing the core science? And how are we making new techniques, new science available in order to attack these large data set problems? And it's certainly a computing problem. It's a problem that demands the basic interest of the National Science Foundation.

We've had cooperative efforts between DHS and the National Science Foundation to look at the core science issues. How do we do basic science to move this problem forward? Is there any need to continue to duplicate our funding against those core problems? Maybe not. A lot of this gets down to what companies, the academic community try to sell to DHS. It comes in with new labels, new descriptions for things that turn out, once you decompose them and absolve them to be the same old techniques that have been around for a long time. That being said, if we look at the problems that we're working there still is no one good approach to solving the pattern recognition problems that we have to solve. We're still down to having this chest of tools that Vasant described. And having individuals have to understand those tools in the variety of contexts that they apply those tools to data.

MR. SAND: Well it kind of sounds like all we're basically saying is we'll find a bunch of data and do things to it, and stuff happens. And I'm wondering if the - Dan talked earlier about this progression from data to wisdom - And I'm wondering if there are more steps than just that one kind of wholesale? Well, we look at it, we take all the stuff and we throw it at the data and suddenly all this great stuff comes out. And I'm wondering if we can break that down a little bit. And then talk about some of the specific things that people do to this data.

MR. SPAR: Sure. Well there are definitely a few steps in between looking at the data and great stuff happening. It depends on which consulting firm you talk to. But, there are probably a lot of steps that go on in between. For starters, as Vasant said at the very beginning, if the data's not clean, if the data's not reliable that's going to throw everything off, right off the bat. And that's a surprisingly large problem in a lot of larger organizations. And it happens in law enforcement if you get a bad tip. It's the same kind of idea.

So, step one is make sure you have a clean set of data to work with. Step two is look at the most basic level of analysis to try to refine you hypotheses, assuming that you're looking for data to try to validate or invalidate a hypothesis. So, start simple, try to hone in a little bit more. The third thing is see if you can't limit the number of variables. Because a lot of these problems have just gigantic lists of variables. Now I'm not saying you should artificially limit it. But, the more variables you have the more it starts to become an unsolvable problem. And if it turns out that we have to create sub-problems within the larger problems so we can solve them, we might have to do that. If it's not realistic to limit the variables, it may not be realistic that this technique would get you the answer. So, those are a few of the preliminary steps. I don't want to go on for an hour on every step you could do to data mine. A few people in the audience have heard me go on for more than an hour on that. So, I will have mercy on everybody and let us take turns talking about what you do. But that's a starting point.

MR. DENNIS: When we look at this spectrum from data to wisdom. I think we can go from data to information fairly #NAME? approach. And we have a constrained problem and a well defined problem. Getting that into the point where you call it knowledge, that's really hard to define what you mean by knowledge, or a knowledge based system. And there are lots of people who talk about knowledge based systems or knowledge bases. And they turn out to be relational databases of information. And so, getting to a definition to what knowledge is, and certainly wisdom is, is probably beyond the scope of what we're doing at DHS at this point. Although we are trying to reach that level of capability by involving humans in networks. That involve expert opinion and having people who can cast context around data is really important. So, we're using humans for that. But when we start looking at bringing together data sources and to do this pattern recognition and mining. We do look at the data, we look at the cleanliness of the data.

It's important to spend time with the users of the data, with the data itself in order to understand what's in there. It's when you get the sample data that you start to understand what are the restrictions on this data. What are the privacy recommendations for proceeding with the examination of that data. It's not until you get the data until all that comes out. And it comes out fairly quickly at that point. Because you're actually trying to take possession of some piece of it. Then once you examine this data and it's structure. And if it has no structure you've got to worry about that. So, we talk about data ingestion. How do we move the data into an organization, into a space, into taxonomy that makes sense for processing? And that has to do with understanding where it needs to go analytically.

But, being able to automate the process of moving data, re- representing data. It's really important not to have human operators or analysts involved in that process to the greatest extent possible. You want to move them more to the cognitive end of the spectrum. And then once we have the organized information, once we've taken unstructured data and we've put structure in it. We've actually indexed information, we've tried to bring it out from it's native form into something that's a little bit more recognizable and useful. And then the transformations that are necessary to do that, that's when we start looking at this whole area of knowledge discovery. How we would apply algorithms that actually perform analytic processes that propose relevant information without flooding the user with false positives.

And it's very important to understand what can be done at that analytic level. And so that's what we call knowledge discovery. Then we worry about how information sharing and collaboration works around that analytic space, around the data space. And have to affect the exchange of information, either information that was in some raw format and has been transformed, or in the form of a hypothesis that has been evolved by a human. And their thinking on the problem. If someone has done a lot of thinking on the problem, it would be nice to pick that up in a wholesale form and reuse it. And the biggest questions there are how to take analytic products that look like that and use them without creating group think or bad collaboration behaviors that could actually lead you to a bad space if you start doing collaboration.

MR. DHAR: To add to that, I just want to [inaudible] great things that the KDD community - that the community of Knowledge, Discoveries, and Databases did about 10 years ago was it established a competition called the KDD Cup. And it's basically a standard data set, every year that's made available to the entire community. And everyone goes off and takes a crack at it. And those data sets have been quite varied from life long data to the Human Genome Project. And the insides from doing that have been pretty remarkable. Especially when you look at the winners, and they come up and describe their approach. I think the whole community has learned a lot from making, sort of some standard data sets available to people. And one of the things I was thinking about on the way down here was, what are the problems that we're dealing with here? Can we define the problems and the appropriate data representation to the point where it can be anonymized, and made available to people who can then try different methods on it? And I think if that's something that's possible, then that will go along way to actually determining the effectiveness of the method that we have around - off the shelf, even cutting edge techniques, whatever. I think that would give us a good sort of evaluation data points on how the various methods work on various types of data.

MR. REIS: I think to a certain extend the steps in between are a function of the problem that you're addressing. These are larger scale challenges that my colleagues talk about. But if you start at a very basic level, perhaps the law enforcement officer pulling over someone for a DUI. Because he was speeding, driving recklessly, or what have you. And then launching a query to determine if any of the local sources of information have data about this individual. Perhaps it's a known or suspected criminal, perhaps it's a stolen car, or whatever the case may be. That can lead to links and connections to other activities that can drive investigations into - in one case that we're familiar with, busting and entire and very major drug ring in the Northeast or Northwest rather. So, it depends to a significant degree on where you start. It can start at a very simple place and lead you into more complex places. If you have a suspect who has a cell phone, for example. Subpoenaing call records from the cell phone may lead to a lot of data that could belying for pattern detection purposes, that could lead you to other potential suspects and connections, and so forth. So, there's very real and practical applications to these kinds of technologies in the investigative world. Both in the law enforcement and intelligence perspective that we see routinely on a daily basis.

MR. SAND: We have about 10 - 15 minutes left of our discussion. Then there will be half an hour of open Q&A. So I encourage you to think of the great questions you want to ask. That microphone over there works really well, if you like talking in public, I encourage you to use it. Ask us questions. I'm going to run through some of the questions that I thought of as the previous panels were going forward. I'm going to just kind of run through those. We also talked, in preparation for this meeting about a bunch of other questions. So, if there are questions there that you guys really want to talk about, I encourage you to rejoin the conversation that way, so we can hit all the good stuff.

One of the interesting from the last two panels was this idea of data having to be clean. And that overlaps with commercial versus government space. The suggestion was that while the government only has messy, dirty, faulty data. So, relies on the commercial sector to give it clean, healthy, perfect data. Which it perhaps it originally got from the government. And we also talked here a little bit about the requirement about good data to do good technology. And I'm wondering if that's always true? Is there something inherent to this tool that we have in this emerging technology that can get us over that hump, and actually move us forward without having perfect data? Can you use the technology against the data has to compensate for problems with the data itself or something like that? Is there a way to work around this problem that we talked about earlier?

MR. SPAR: Well that's one of the toughest parts. As someone who was a DBA for a number of years. One of the toughest things, both in the commercial sector and in the government sector is to have clean data. And there really are two parts to the puzzle.

One part is the data having values that are correct. But, one of the other areas, especially when you're pulling together disparage databases, is to have the proper data definitions. To have the proper normalized structure. So you know that what's called the person name in one database means the same thing in the other database. Even within single large databases sometimes what's known as data value or a table column has morphed to meaning something different than what it originally meant. But, yet the values that were put in there at the original time of meaning are still there. So, I would say there's just a huge amount of work to define what the data is. And unfortunately in a lot of large organizations that knowledge has never been documented. And it may not even be possessed by the people who are still there. So, it's a huge task. It's one that proactively as new systems are designed you can start leading things off by putting together a data model and putting together the right definitions. But, ultimately garbage in, garbage out. Tools can overcome it, but it's kind of like a TV set trying to overcome static. You get your best results if you could clean it up at the source.

MR. REIS: There's clearly no substitute for having clean, ideal data. But, to the point that Dan was making about valid data existing in multiple places, but being labeled differently. There are technologies available today, and in use today to do the Symantec analysis. To determine that someone who's called a person in one database and a suspect in another database is actually the same individual. And there are techniques available to evaluate how close the match is. Of course at the end of the day it comes down to the analytical mind to make that absolute determination. But, there's no way to replace faulty data with technology. And no way for technology to adjust faulty data to correct data. At least not that we've figured out yet.

MR. DENNIS: So, I think it's - we talked earlier about the questions we'd like to ask. And this is primarily one of the reasons I like to ask the input output question. Is where did this data come from? I have been tasked as a researcher before to fix noise problems or to try and come up with methods for cleaning data. Only to find out it was introduced by the customer in the first place. And so I'm doing the process that they were doing to themselves. You know, we want to discover that early. Early on in the game. So, it's important to find and classify these types of error to characterize them, to understand where they come from, and what can be done about them. And if they are indeed introduced through a legitimate process.

In the past one thing I've done is been able to create new applications. These are not applications that the customer was thinking of. They were thinking of the 100 percent accurate approach - you know I need 100 percent accuracy on this dirty data. Whereas with 30 percent accuracy or 40 percent accuracy in a particular processing module, it can be possible to diagnose the data and to put it into binds. And to segment the problem space in order to treat various components of that data separately with different techniques. And so you can drive up the overall accuracy for the process by introducing a diagnosis or some kind of triage process early in the game that allows you to refine problems over the entire data set.

MR. DHAR: I guess alternatively you can also get a sense of what proportion of the false positives to false negatives your willing to accept. You know depending on how much human input you want to put into the process. And this is - this comes from the credit card fraud arena where you can have a classifier that can be very aggressive or it can be very passive. And the question is how aggressive you want it to be. Well that depends on what kinds of resources you have available in terms of people looking at these cases. And how many cases can they process per hour. Which really determines how many case you do per hour that, you know that systems use for [inaudible] people. So, the whole question of like how accurate does a system need to be? Or what kinds of false positives are you willing to tolerate, is also a function of how much human input are you actually willing to put into the process to deal with the false positives that come your way.

MR. DENNIS: I think this is an extremely important point and it needs to be raised early on in the process of trying to solve problems to manage the expectations of a customer who demands 100 percent accuracy. When that's not even technically feasible. And to remind them it's going to take a human process - and even a human process is not 100 percent accurate. If we look at information extraction, the accuracy for a human is about 96 percent over trying to find proper names and locations. And so you can't even get 100 percent with humans, then what are you trying to demand of the technology?

MR. REIS: I think also there's a significantly relevant point there with respect to false positives or false negatives. And that depends to a degree to the problem you're working on. I mean if it's whether you deny somebody credit, or whether you shut down their credit card because of some sequence of events or what have you is one thing. It's another thing if you're issuing a warrant to arrest somebody. So, you need to be far more precise in certainty as they do in others, because of the consequences associated with it.

MR. SAND: We're getting close to open question time. We're going to keep talking until somebody stands up and asks a question. So, I encourage you to ask questions if you have them, otherwise I'll keep asking questions.

I'm wondering if - we're talking 100 percent accuracy versus something less. And I'm wondering if non-purely accurate results are always worse than pure accurate results. Again, going back to this concept of a fuzzy match. Is there anything that technology gives us that we wouldn't have otherwise that's actually better than a perfect answer? Is there something that's better by being fuzzy than it would be if it were complete clear or crisp? Or is it really just leading up to the perfect answer if we can hit this point, it's good enough. But, we'd always really get to 100 percent.

MR. REIS: Well, the way we look at that is that if you're looking for 100 percent match. Then the likely response will be small. And you are likely to miss a legitimate target as a result. So, what we tend to try to focus on with the technology is to produce results that grade the response relative to the probability of a match. And then leave the resulting analysis to human interpretation. So, if I've got a 100 percent match I will tell you that. I'll tell you the next matches maybe in my judgment 80 percent or 70 percent or what have you. But, even as 40 percent match maybe perfect from an analytical point of view. So, if you're looking for 100 percent you're going to miss a lot of data in my view.

MR. SPAR: I think Vasant had it right when he talked about the level of aggression that you put in your algorithm. And the amount of false positives that you're willing to endure. And it's very much supported by what Jack said. You can tune these algorithms to create false positives. You can also tune them to say only bring to me the things that have very high confidence in terms of the match. And what happens if you're in the situation where you say I only want the good stuff.

You miss a lot of things. And most analysts that I work with are will to endure some level of false positives in order to make sure they don't miss as much. Now you can do an analysis of what humans are doing with their data. You can look at data queues. You can look at results from queries, or whatever it is. And you can perform a statistical analysis of where the good data was when people do retrievals. And you can show folks how far down - who many false positives they had to endure before they got to the things that were important. Especially if you instrument the system, which is a very important piece of the pie. You have to - in order to be able to characterize these operational problems you have to be able to understand what's going on as humans move data from one point to the next. As humans produce those outputs from their organization, what was the data that actually led to those outputs? And how much time did it take? And how much time did they spend moving data around? And all that can be had if you can instrument they system and you can get folks who use systems to agree to be watched in that kind of manner. It also help you to understand how they're using the technology. Because often times users will tell you they're using the technology one way, when they're indeed not.

MR. SAND: Sir, you have a question?

AUDIENCE: Sure. The names Jim Waldo from Sun Microsystems. I one of those engineers who has to answer when the Program Manager comes back from the customer and says - after the Program Manager has told the customer that they can do anything. And I'm the one that's told do it. So, I think that there is - in light of this I think there's a category that you're missing in your distension between the stuff that can be done now fairly straight forwardly - stuff that is rather hard. There's also a category which is currently computationally infeasible. And a number, for example of the worried about privacy violations in the TIA program, I think fell into that category. And so, the question that I have is what do you think are now the current upper limits on scaling for the techniques that you know how to do, or think you'll be able to do without say a major breakthrough in data mining technology or a change in the nature of mathematics? Was that a clear question, or not?

MR. SPAR: Yeah. I thought it was clear. My own experience is that I really - unless I resort to doing exhaustive search - I mean in a huge space, I really haven't run into computational limits as being sort of the bottle neck. At least in the kinds of problems I've look at. That's usually been the least of my problems. At least in the last few years. The problems I think have to do with all the other things we've been talking about. Which is where do you start? How do you formulate the problem? What does the data representation look like? What kinds of accuracy are you willing to accept? So, I really haven't run into computational bottlenecks as being the drawback to what we're trying to achieve. I don't know what the other peoples experiences are.

MR. SPAR: Well you're from Sun, can you throw more boxes at the problem? (Laughter)

MR. SPAR: But seriously a lot of the computational limits are functions of what processing capabilities you can throw at it as well. So, you do have some control over that. And I think as Vasant said what you really need to do is to reign in the expectation of what is expected to be searched. What is the meaning to the problem space that you're looking into. Maybe break it into subsets as well. Also, sometimes when I've found a Program Manager has set me up to fail, I try to alter the description a little bit, where I say sure I will get you your working data mining system within sixty days of having clean data. And that kind of gets you off the hook for awhile.

MR. DENNIS: I think there are a number of areas where we have computational infeasibility. And I think that basically some of the problems that have been posed to us in terms of provide us with 100 percent accuracy for information extraction like qualify to that point. There certainly is the agglomeration of large connected spaces from multidimensional data sets that pose a significant challenge. And being able to efficiently conduct search and pattern recognition over large disparate networks of information, where you have no central memory, because of various constraints are certainly computationally intractable at this point.

MR. SAND: Next question, sir.

AUDIENCE: Thank you. Stanley Campbell with Eagle Force Associates. A lot of what we talked about or what we've heard is kind of - I think I agree with everybody, and kind of disagree with everybody. So, and that agreement or disagreement becomes the issue of the question. The answers and questions are all contextual. For example, if we look at the question associated with scaling. If we assume somebody like Goggle has a large data set. And then we look at the reality that they only have about two percent of the worlds content. And that other reality that Wal-Mart has about two and a half time the content of Goggle. And then the reality that the Defense Department probably collects about 50 percent of the worlds content. And so - and that equates to about a petabyte of new data every month. All the answers could be right if they were scaled to a low law enforcement or a high unstructured content. So, then we go to the context of a content. Is it structured? So, a lot of the answers are right, if we're talking about structured data. They then get dismissed when we're talking about unstructured data.

When we look at intelligence we can ill assume that Bin Laden is going to be doing his communications in a structured data environment, for which we control. However, when we have a kid sent us an e-mail that says I put a box cutter on an airplane. That's an environment we control. So, the issue then becomes - let's look at a different case where Marsh McClendin issue, where the context of the content geospatially is different when you ask the question in New York, associated with contingent commission fraud. Or, if you ask the same question in Georgia where the law is different. Or, if you ask the same question in a temporal relationship from a time line standpoint.

When we talked about and used the words that were criminal in 2004 on contingent commission, but those same words in use and actions in 2002 were not criminal. Then the issue of temporal, geospatial, structure, unstructured all become issues of why every answer was correct for a certain consideration. And why every answer might have been wrong for a different contextual presentation. So, my question then becomes, given all of those paradigms, and then when we then look at commercial data ranging from the Lexus Nexis, Equifax, to AP feeds, or web based content.

When are we - kind of define what is large and what is medium and what is small? When do we define how we deal with the relationship with unstructured and structured and semi-structured e-mail traffic. Because that's were the Marsh McClendin case fell, as opposed to where the law enforcement link analysis fell? And when do we look at what is real fact versus supposition versus specific falsehood? BTK Killer case, the issue of the guy giving a lot information, mostly all false. So, how do we rank relevance? Do we ignore Al-Jazeera when they get to take first? Or do we relevance rank it with the same tape coming from the NSA or the CIA? So, how we structure data, how we rank the content from the source. How we then scale, and how we come up with some standards for which those of us who use the data, those of us who try to solve the problems can actually work from a same set of rules for which we then implement towards, and then we then answer the question to.

MR. SAND: So there you go. Next question. (Laughter)

MR. SPAR: Could you paraphrase your question? What is it you want to be able to do? I'm trying to make sure I understand your question correctly.

MR. CAMPBELL: The eventually - is what I think we want to do is - I'll put it a different way. All of the algorithms that we're talking about, Microf, Bazien, Boolean, these squares fit. All of those are 16th century math, 1800's math, turn of the century math. We're just looking at the same math a different way. All the conceptual structured data, semi-structured data, unstructured data and how we use it. It's all basic fundamentals of linguistics. Basic, 200 years old. If we go to the Amazon we communicate exactly the same way as we do here today. So, if we're talking about data mining from the standpoint of use of algorithms to mimic human thought, how then do we set the baseline and rules to say large numbers of - large amounts of content is a petabyte. Medium is a terabyte, and small is a gig. How do we then go to, okay, we're going to be able to use structured, semi-structured, and unstructured collectively. How we then take that and then put a relationship to time? How do we then take that content and put it in relationship to geospatial or space?

So, how we bring this data alive and I'm saying - we've got a lot of individual tools. We group them together, but we don't have a baseline for which we can now start the conversation, to start the solution by having a baseline of what are our definitions in those areas. Does that make sense?

MR. SAND: Why don't we spread those questions amongst the group here.

MR. DENNIS: I'll take a crack at it.

Basically, I think what you're asking is how do we make these systems work on the knowledge and wisdom scale that we talked about earlier. And I think that's very much an open problem. And it's been tried before in the past. You know we had artificial intelligence and it went for some number of years under that guise. And how will we make systems like human. How will we make systems understand and process time and location. I mean the context of our problem. And I think it's a very difficult, and still remains a science based problem.

And I was recently at MIT where they decided that - I don't know how many people are familiar with the Big Psych Project? Where you try to encode all this encyclopedic knowledge and bring it to bare on your problems. And in several instances I was forced to work with that kind of data in order to make a determination about it really was possible to use that kind of information to do information retrieval, or some task that had to do with determining relevance of data. The answer at that point was no. At MIT their answer was we think that the problem has been chewed on and been thought about for many years. That it's time to go back to first principles. And to think about smaller problem spaces in which we might learn. Clearly Psych gathered lots and lots of information. But, it's not clear to us what we can do with that kind of sophistication, especially when the inference could go on forever and never achieve satisfaction in a human successful way.

So, what they did was go back to first principles. And in the area of robotics they're looking a just small, simple worlds and how to reason over blocks that are different shapes. And have robots use visual queues to understand how to move shapes around. And to understand and discover their world, and what it means to have time order in a very small world. So, I think a lot of scientists have gone back to revisit that problem. And I think it's hard. If I look at government investment, it's moved into this thing it calls the hypotheses space. It wants to really take that next step. And it's investigating what is the actual connection between the data space and it's representation and a human thought? The problem is that you have many humans who don't agree on the hypotheses. You have many competing hypotheses. You have a credit assignment problem. It's a motivator for humans participating in an environment like that. And how do you give credit to sign it back to the algorithm, but also back to the humans who led to the observation. Dealing with time and text. Dealing with time and data. Is a very, very difficult and challenging problem.

I recently - Art has spent some funds trying to figure out how to even code and represent time in an [inaudible] way. And I led to time [inaudible] standards. So, it's just now true that we're starting to get to representation that allow us to look at those problems. But, it's also true that I don't think science has taken us to a place where we're ready to tackle that time limit. And that falls in the area of computationally intractable and something that we're just going to have to recognize that we're not going to do it until we have some basic science break through.

MR. DHAR: I just want to quickly disagree with something you said. Which is that this stuff's been around forever. It actually hasn't. If 10 years ago someone had told me you could type anything in English to a computer and come out - give you an answer.

AUDIENCE: Math.

MR. DHAR: Well, okay. Yeah, sure. Arithmetic has been around forever. But, if you look at the kind of progress we made in text finding and some of the newer stuff. It's pretty impressive, in what we've actually just done in the last few years.

MR. SPAR: Well, Vasant you just shaved five minutes off of my 10 minute comment. But, that's okay. I was going to make that exact comment that if - a lot of the basic mathematics has been around for awhile. But, the ability to implement that - those concepts hasn't been. I mean didn't Leonardo DiVinci come up with a prototype picture of a helicopter 500 years before we had a helicopter? I mean there's a difference between having a concept and being able to do anything with the concept. And what is new, is that we've been able to apply those concepts and then have a lot of refinement techniques in the way that they can actually be applied to distributed data sources, both structured and unstructured. It's not a weakness for something to be based on older mathematical concept.

So, I just wanted to bring up that small point. But ultimately part of your question was kind of hitting me as - the world of data is gigantic. It's structured and unstructured. We need to uncover a lot of information. The world of our data is growing too. What can we do? And what can we do within the limits of what we can scale to do? And remember what I said earlier that quote, if a problem well stated is a problem half solved. We need to cut that into subset problems. Because we can't do anything when the list of variables that are unknown is longer than our ability to put together meaning to some of them, so we have to look at a smaller piece.

MS.LEVIN: I hadn't really expected to ask any questions, but the last question was - it was mind expanding for me anyway. So, I'm driven to ask this question. Which is, a lot of legislators seem to think that - or at least thinking that technology, maybe not data mining exactly. But, technology, and maybe not off the shelf. But, that technology can be obtained by DHS and - well let's say DHS. To address it's mission with regards to making the homeland safer, more secure. Identifying terrorist possibly - as one example of implementation. And there's a lot of pressure from outside the department to use technology as an answer here. And I think the last gentleman was sort of asking well what's the baseline, or the measures that we can evaluate technologies to see whether then can - using any of those different modalities. Whether geospatial or time can answer these questions.

Given what your discussion this afternoon, I'm want to find out whether - where is the technology in terms of being able to answer the question that the Hill or DHS is posing with regards to helping fulfill the identification of wrong doers?

MR. SPAR: I can just kick off a short comment, which is you can only evaluate technology or anything against an objective. And when compared with other alternatives. And some of your alternatives might be different levels of technologies. But ultimately technology is going to give you some alternatives you might not have had before. Those alternatives should have a level of performance you can quantify. Then you can decide whether or not that level of performance was worth the incremental time, effort, and cost. And whether it incorporated some aspect of a privacy intrusion that was acceptable or unacceptable. In a vacuum you can't answer that. But against that type of framework you can.

MR. DENNIS: I think a lot of the questions that you're posing there can be answered by data that is going to be [inaudible] coming from our [inaudible] for vital and current technologies. Especially with respect to information technology. We have a very big focus on metrics in DHS and how to evaluate and understand not just the technical performance, but also the contribution commission.

If you look at the information analysis structure of DHS, it's been recently elevated by two SR's. Basically become it's own entity. To try and sort of coalesce the energy that we have in DHS for gathering information, making use of that information, and then leveraging it. There's no doubt in my mind that there's a lot of work that can be brought to bear to automate the processing of that information. To bring it to a point where action can be taken and to bring the accuracy to a level where you're not completely flooding folks with false positives or information against which you took action and it turned out to be wrong. But, that requires a certain about of time in order to figure out exactly how these algorithms work in the DHS context. It requires time preparing the problem statements. The problems here are vastly different than in other agencies.

Bringing together the state and local information to a widely disparate, largely geographically dispersed information data sets to bear. Both for the advantage of DHS and it's mission for homeland security. And the extended homeland security community. But also to give federal information back to those who are on the streets or front line. And that's their operation.

So, S&T is all about metrics-based evaluation of it's programs and systems. And very much also about risk based investment strategy. Trying to put those investments - you might think that the money DHS has it a lot, but when you look at the problem set against which it's placed, it's quite vast. And so having that rear space assessment is extremely important in order to derive what it is that has to be done.

MR. REIS: I think it's fair to say that in the space, and certainly as it relates to the application of technology to national security. We would view that we're in an emerging environment. While the questions being asked are looking for grand solutions to massive problems. Generally speaking when we're in an emerging environment in the technology we tend to see the development of what I would call point solutions. Which means a smaller technology solution to a part of a much larger problem. So, you see the application of technology to enhancing productivity in a particular area for example. That only may be a subset of the total challenge confronting an agency or an organization.

Over time you'll see more and more point solutions developed to address specific problems in a continuum of challenges. And then eventually the amalgamation of these point solutions into more complete and total solutions. We're coming out of the emerging state, if you like. And we're starting to see, at least from our point of view a requirement to amalgamate multiple point solutions into more complete solutions from a users perspective. That is to say a single solution that provides or addresses multiple problems.

MR. SAND: Do you have another question?

MR. DENNIS: Actually, if I can throw in just one more comment to go with what Jack said. Basically these point solutions are indeed the evidence of basically the pilot approach, the understanding of the problem. Taking bites that can actually be chewed and consumed from the problems space in a meaningful way. And as those point solutions - and if you deal with customers inside of DHS, they're tough customers. They basically say we don't care about your five year plan, we want to know what we get in six months. What do we get in one year, what do we get a year and a half? So in order to engage those customers you have to produce these point solutions that are not just deep engineering, throw them over the fence types of systems. But basically are informing both the customer and the science community in DHS about what the problem statement is and what accuracy levels are required in order to solve those problems, while giving value to the customer. Not ignoring the customer.

MR. REIS: Yeah, I think just corollary to that, you know trying to define a complete solution to a major problem is a bit like trying to plan the beltway. You know what you can be absolutely certain is they'll be a 12 to 15 year development project. And by the time it's implemented it'll be obsolete. So, you need to tackle these problems in small bites. And address realistic problems today piece by piece. And eventually those blocks will come together in the form of a larger solution.

AUDIENCE: I'm Linda Ackerman, I'm the Staff Counsel with Privacy Activism. We've heard a lot in this discussion about the technology, and very little about the privacy. Starting with Larry Lutz's assumption that architecture is policy. And drawing from what you've said about being unable to do effective data mining when you start with dirty data. Isn't it essential to build in the privacy to the technology from the beginning. I mean just as you can't start with dirty data, you can't really retrofit a technology for privacy.

MR. SPAR: There are a host of techniques you can undertake to try to implement some aspects of privacy. If you were to, for example replace someone's name with an ID number, then have a separate file that's not available to people who were doing the analysis, where they have the ID numbers and the names. Then you successfully remove that aspect. Just like the way they do drug testing in the Olympics or for Lance Armstrong. So, there are a lot of ways you can separate out the ultimate things that link you to people. Also, a lot of data mining is done initially by scanning individual granular pieces of data. But then you start to discover trends and relationships. And those are a lot more based on the aggregate data, which is a lot less privacy intensive, as individual units of data might be. So, there's - it would really have to be a very clear definition of what is it that you don't want an end user or a database person or a end user analyst to be able to see. And then you can take steps to remove their viability of that, even in a system that's already been built.

MR. DENNIS: So, I think starting around 1999 there were some experiments going on in terms of how we actually implement privacy in a context of an information sharing network. And some experiments were done with analysts trying to achieve jobs. Over which the data that was used in those experiments had actually lots of restraints about how data could be used. And different agencies accessing different parts of that data were allowed to see different pieces. And so that architecture has evolved to serve our research community in a way where we can run experiments and try to implement what have been evolved to be rule based filters that implement policy in terms of privacy. And so what happens in those cases is that an individual user asks a question. And the question is actually analyzed. And you determine whether this question is appropriate for this person to be asking. Is this person in this role suppose to be asking a question like that? If they're not they're question is rejected. The system doesn't even try to pursue an answer.

If that first level filter, which is implemented by their own organization is passed. Then the receiving organization which has data and information. It could be within a single department or across agencies. The question is again scrutinized. And that data provider says is this question appropriate for my data? The next part of that process is the questions allowed in and some answer is formulated based on the match of data behind that veil if you will. And then as the data is going back across the transom it again is scrutinized. So the data itself that's going back to that user is again scrutinized and cleaned up in terms of what that person in that role, in an authenticated role in that organization is allowed to see. So, there's been a lot of mock up and implementation and some experimentation with that kind of technology.

What we'd really like to get to at some point, and we've worked with the advanced scientific computing community of DHS to try and develop good requirements for privacy at every level of the system. But until we get to the point where can embed it throughout the entire system, having these gatekeepers, these policy devices if you will at the edges, is going to serve that view. And conducting experiments that tell us what is the rate of failure for such a device. What is the acceptable level of privacy at those points is extremely important. And getting to people to think about what human performance is also in my mind. You know how are humans protect privacy?

I think about algorithms and their ability to carry out a function. And algorithms can't be co-opted, they don't look up their neighbors, they don't do all these really bad things that have caused this concern. And maybe it's possible to embed algorithms to prove that they're not actually violating someone's privacy. And yet operate at the right level to let people see only the data that they're allowed to see. Yet consider larger connections. So, I think this is a large space to explore. We've made some strides and headway. There are a variety of techniques that can be employed at various levels. And the biggest problem that I see for the technological solution is getting agreement on the governments of such systems, as they interoperate one agency to another. Another thing that this kind of technology does to the policy folks, is it makes them very anxious. Because they're suddenly are codifying their policy. They're not just writing memos that say this is our privacy policy. They're actually codifying it into the system. And that makes people very nervous about the decisions that they're making as they do that. So, it's going to require business transformation as well as technology.

MR. REIS: I think an important corollary point though is privacy is a policy issue. Not a technology issue. Technology has to support the privacy policy. But from a technology point of view, technology is data agnostic. It assumes once the data is presented that the user has access to the data. So, I believe that the technologies have to be developed in such a way to support the policy. But the technology itself will not create the policy.

MR. DENNIS: I agree with that point of view. Except that technology is definitely agnostic to policy. But a lot of people who design systems do so under the context of a policy. So, for example when I ask to do research using certain data sets, there are rules given to me about those data sets. How they can be used. Now if you build you system to those rules you're codifying the policy. It's not as direct. Certainly not fungible, there no way to change it. So that leads us to the scenario where I'm afraid we'll be in the future which is we have systems that are deeply engineered to some specific policy. Tomorrow the policy changes, and then we're looking at three to five years to implement that new policy. So, dialing the policy into the system, allowing for that to be codified at the time of use, or during the agreement. And I've seen standards for electronic brokering of data exchange agreements is a good direction to be heading.

MR. REIS: Yeah, but on top of that though. Think about this from a national perspective where we want to involve the states and local communities into a national process. Where at each level there will be different rules and different guidelines about what assessable and what's not assessable. And under what set of circumstances. So, a single base of technology is not going to work in that kind of environment. It's going to be increasingly challenging to develop technology that's universally able to support full policy. At the end that data policy has to drive technology.

MR. SAND: I'd like to close this session with a little bit of preaching, by way of summary.

We talk a lot in Privacy Office about insuring that you embed privacy into technology and privacy in technology and policy [inaudible]. But I think what I've gotten most out of this discussion was an awareness of what the questions that come out of the technology are. And how they can inform the policy first. And from being on the receiving end of a lot poorly crafted laws in response, we'll implement them through technology in the past. I can tell you that there's great gain to be had by talking to people who have their heads deep in the technology while you're writing a policy. Because what you'll find is that a lot of the questions that come out of the technology community will actually make your policies more precise. More easily implemented, and more easily changed as the policy change. Because it's informed at a much deeper level as to what information it is, how it works, what you can do with it.

A lot of people gloss over that and say well there is right and there is wrong. And you should do right. And so go and do right. When you go back to your machine and try to code that you end up creating your own policy. Because the statement never reaches a sufficient level of detail.

So, just to exploit the situation and to advocate that the people that are writing policy, and people who are making law type decisions, I advocate that those people come to people that are wearing the collar shirt - the short sleeve collar shirts and 57 different kinds of pencils in their pocket. And they're easily identifiable. But go to those people and say, well what about this idea? How does this look to you? What you'll find is that a lot of questions coming back as a result of that dialog being a much more detailed version of what the policy maker wanted to do in the first place.

You get the details, you understand what the variables are. What the relationships are - which by the way is what's going to have to happen next anyway. Somebody’s going to have to figure that out. And the more you can [inaudible] that process and integrate technology awareness into policy and decision making. The better the policy will be when it comes back to people that have to do the work. I want to thank the panelists here and you for sticking around.

I hope you enjoy tomorrow's session.