>>Please stand by for realtime captions.
>> Welcome and I would like to thank you all for holding. And inform you that your line is in a listen only in today's conference until the question-and-answer session. At that time you may ask a question and press star one. This call is being recorded if you have any objections, please disconnect. Now I would like to turn it over to Karen Geary.
>> Hello, my name is Sam Weber and I am the program director here at the National Science Foundation. I would like to welcome everybody here to the aid in our series of lectures about trustworthy -- and -- we apparently have a quite a few people online as well is in the room. I am really pleased to introduce Vitaly Shmatikov. Our speaker. Associate Professor at the University of Texas in Austin. When I asked them what I should say to introduce is he said that he was really very embarrassed about all the introductions and therefore, I should keep it short. I would just say that he has worked in an impressive number of various computer security ranging from analysis of calls using formal methods to inspecting programs, finding security flaws, Web applications, and of course, privacy which he will talk about today.
>> Thank you.
>> Thank you, thanks and thank you for inviting me. I topic today is going to be -- and privacy, anonymity and privacy in the year of the data. As you all know, we will -- we live in a world where information is collected in a massive scale so no matter what you do, every click you make online and anything you buy goes into a big database or somewhere database somewhere. So pretty much all companies today that sell anything and do anything collect information about their customers and information about peoples -- or Jesus like browsing history, Web churches purchases and medical data and analyzed on a on a massive scale. This come these companies are of social networking companies. Except one very prominent exception, they don't seem to make any money, but -- they promised to make it by telling their investors that they will monetize and make money by analyzing peoples data and data about social relationships. This companies are a little bit less known, they are -- operate behind the scenes you and they specialize in collecting information about people's browsing behavior. Pretty much any click you make anywhere is captured by one of these companies, sometimes by multiple companies and they work together and used to target advertising and monetize in various other ways. Another trend that we are seeing is there is a tremendous amount of aggregation of information going on both online and off-line so information collected about people's online behavior and transactions that they make online is being correlated and connected with off-line databases and things like financial records, credit reports, stuff like that. Social relationships as well. To create a massive database that capture the entire life history. And of course, this massive data collection going on raises an important question. What about privacy?
>> Clearly people understand that a lot of this information is very sensitive and could be used for various nefarious ends. Spam and advertising that you don't want to see. And it could also be useful. What do they do to protect your privacy when they actually share this information and sell it and so on? It is a most excellent question. If you look at what people actually do and they do existing privacy protection method for the most part they focus on the notion of personal identifiable information. The way privacy protection works today for all

intents and purposes as they take the data that they collect about people and anonymize them and removing personal identifiable information. That is a magic -- information once they remove the an the identifiable information data becomes safe and now they it is not linked to a specific individual that could be shared and sold and published and do whatever you want with it. This personal identifying information or just shortly PII, we see all over the place. No matter where we look and no matter whoever asked questions about protecting privacy the answer is always the same. There's nothing to worry about you because by the time we give the data to anyone, Arsenal identifiable information is gone. So for example, this is the previous chief public officer of -- and drawing the distinction critical distinction maintained a use of information and personally identifiable and the sharing of information in non-identifiable form. You can see the built-in assumption that information exists in to form. There is personal identifiable and non- personal identifiable form and when she converted, it becomes a safety share and 60s in many cases shape to publish. You used you see this missing should all over the place. There is -- networking and they say what they say is social engagement data and anonymous information regard regarding people. What does it mean is anonymous, the personally-identifiable information has been removed. If you do a Google search, -- we do not collect personally -- personally-identifiable. Clearly this concept is all over the place. You can even get a certain feel seal of approval. For behavioral advertising. They tell you there's nothing to worry about because this advertising technology does not use PII to deliver relevant advertisements. Of course clearly this notion of PII is so important that even the federal government provide some advice in protecting the personally-identifiable. This is all good and very comforting for us as consumers and users of the Internet and so on. But it seems like an important question -- what exactly is this personally-identifiable that they speak of? because it seems like you see in protection relies in some information -- personally, what is the personally -- personally-identifiable information? That's a good question. And this is what I call the myth. Is this assumption that your dimension data exist in two forms. It is ended -- there is a transformation that you can make to make it personally-identifiable information to non- personal identifiable. They believe that you can take less and by applying something you can turn it into -- the same way that you can take personally-identifiable information and turn it into safe, good, non- personal identifiable information. You remove things like names, e-mail addresses and so on. It's been understood for quite some time that this is not -- there could be remaining information their so-called croissant identifiers. That could still be used to identify people -- like e-mail -- -- had been removed. In this cost site identifier, -- quasi- identifier -- requires see that every -- that occur a certain number of times, but again, the question into a different direction, what is a quasi- identifier what exactly is the identity the information that could identify the individual, that is an important question. It has been shown that things that demographic information, zip codes, gender, could be used to identify people, but is it the only kind of information that could be -- for identification and the same attributes that are identifying quasi identifying every individual and different things that could be used to identify individuals? These are important questions that is fundamental because if your privacy and protection relies on removing identifiers,

what are they? We have known for quite some time that this notion of protecting information and identifying it using specific identifiers is not sufficient. Exhibit A. is AOL search data scandal which you may remember, seven of them years ago when AOL released a data set of complete anonymize Web searches. No IP addresses, nothing that could be considered identifier work quasi identifier. Sure enough -- a newspaper reporters started -- figured out that he can identify certain people like the -- there was some lady looking for businesses and in Georgia, or fair land -- a rare last name and single men over 50 and how to what to do with the dog on piece on everything. As it turns out, that information clearly is not identifying the by any normal definition. Once you put the little bits of information together it is enough to identify the person. If you look up the phone book for Georgia and find some number and last names that match, and there was only one lady who was apparently over 60 and she was recently widowed and called her up and, she said it's me and I have a dog that urinates on everything. A nice older man than me know. What this tells you is there is this information in the data itself that can be used to identify people even in data that is completely anonymize and be identified that contains nothing that could be considered as personally-identifiable information according to the ding to the federal definition. Whatever the definition, this company used when they eat -- when they anonymize the data. And unfortunately as you look at the privacy and adversarial models, no matter what threat you worry about and there are a lot of things to worry about, you could be worried about fishing and ing and nosy employers and -- and people having to worry about surveillance and collection of data by enforcement agencies, no matter what scenario you are worried about, but they all have in common is something that a lot of existing privacy protection mechanisms try to ignore. The notion that anniversary has access and already has some access information on what they are looking for. This is known as background knowledge or auxiliary information in this information is about people that is available for every stare outside of the process. Available for public databases and so on. Like in the AOL search data, that data was correlated with things like public phone books and stuff like this. There's always other sources of information out there and what I am trained to say is that information that is not personally identifiable by itself can become personal identifiable if the adversary has access to a little bit of information about what they are looking for like a little fragment of the record that they are looking for.
>> Of course the first thought that comes to mind when you think about this problem, especially if you are computer scientists is axis control. The way to solve the problem is by axis control. And it makes sense. We know from access control, we have been thinking about this for 40 years. That is a good notion, but once you try to think about it you realize that it is very difficult to apply in a meaningful way to a lot of data that we have to deal with today. For example, social network, information about relationships is shared. I might consider might end of relationship private, the other end they might not consider it private. It seems like genetic data is even worse. What does it mean to exercise access control over genetic data. It is shared with all the blood relatives and huge chunks of the genome constructed with genetic information of the relative and this means that it's some sense access control is meaningful for genetic data because it can be reconstructed from other peoples data. And complete systems I will talk about a little more later, it is even worse

because there is a complicated dependence of the output and input it is impossible to trace dependencies in order to exercise access control. Axis control alone is going to be the problem. And the solution to this problem, but even before we talk about solution, let's try to understand the problem a little more. This is something I have been working on for few years with -- who is a former student of mine at Stanford who looked at several scenarios in which people released anonymized data and tried to figure out, what did you learn from the anonymize identify data?
>> I am going to look at three case studies today one involves anonymize transsexual data and these are applicable to anything that nything that involves purchases, preferences, tastes, stuff like this. I will use a different data from the database and I will look at anonymize that comes up in social networking and graphs of human relationships. The final is going to be what is the what is released as aggregate statistics that involve no rope they -- raw data at all. Start to see what we could learn from that.
>> Let's start with anonymize transactional data.
>> So the data is just a database. For each row is a customer record and columns of transactions preferences and things like that. That infamous net links contains movie viewing records of 500,000 individuals and row is a customer and column has a movie whether they viewed it or not or whether they like it or not. Before talking about online retailer and they may recall some item they might have ought and the record of record of individual customers. The most important thing about the data set from them in the privacy point of view is that that they are highly multi-dimensional. If you can think of the item they bought, like in the case of movies or Netflix or items available for sale online retailer or off-line retailer for that matter, there are hundreds of thousands and millions of them you if you view each record in customer record as a point in the space where the coordinates correspond whether they viewed the movie or not this space is multidimensional and those of you that work in data analysis know there is typically multidimensional spaces characterized by proper known sparsity. Which has tremendous complications for privacy. I am going to show you that the fact that they are so multidimensional and so sparse has very important implications.
>> Let me show you how sparse they are by taking one point from the graph. This corresponds to the Netflix -- there are 500,000 subscribers. Released five years ago to support the data mining competition. In the Netflix data set, if you look at -- whether the customers watched it or not, 90% of the record, the reason that -- similar data is more than 30%. Every record is unique or overwhelming majority is unique in a very strong sense. Not only are they unique, they are highly dissimilar to anybody else's on the database. If you look at the record, they are really standalone. Each record is multidimensional and the space, even if you draw a huge neighborhood in the record, there is nobody that is remotely similar. What this means for privacy is that even if you have a little bit of information about the record and you can roughly pinpoint the neighborhood of the record, you will find the record because there's nobody similar. If there's nobody similar, even a tiny bit of information is enough to identify your record even if it had been completely de-identified. This is a paper that was published a few years ago in Oakland conference on security and privacy where we do analysis both generic analysis trying to show what kind of properties in the data set identification is possible even if the data set -- I should have

mentioned, this is completely anonymous, we do both generic analysis and complete analysis of the Netflix data set. Generically we show that under very mild assumptions we need a small amount of information and to be able to identify the record. Given a little bit of the record, is enough to find the data set. It turns out it is even less. Knowing the data rating of the person is enough to find your record in this huge data set of 500,000 anonymize history. That is how sparsity's. Knowing just a point, is enough to pinpoint the record. How do we know that the matches this -- matches correct, we developed some of physical technology to argue why identification is correct. That is not important, the point I want to me, this question, the idea of asking, how do you know that you found the right record is in many ways the wrong question. Because sometimes -- the question you should be asking is did you find my record, but what did you learn about need by analyzing the data set? In some cases, we cannot quite pinpoint the record, we can only find a cluster of record, a small cluster and proved that the record longs to this cluster. Even though we can pinpoint the individual record, they often have something in common, they have some characteristic in common. But this tells us is that if I find a bunch of records that have the same characteristics and they know that yours is one of them, it is enough to figure out that they have the same characteristics. Identifying the cluster is as good as identifying the record because I learned whatever I wanted to learn by finding the group to which you belong. In this case, we are dealing with very nicely organized transactional data the data in an regional database where record is sparse. A fairly basic technique for identification work. In social network, we have to deal with anonymized graph data which is more complex. Because it is not relational anymore. Here is an important point that I was try to make when they talk about it. When a state social network, people think Facebook and what exactly is the private issue, the whole thing is about that. Exhibitionism. What the privacy issues. Social network have been spotted for a long time before even computer scientist came into the scene and social networks appear. So here's an example of a social network. Taken from our real [audio not understandable] paper. Representing romantic and social network of him among high schoolers in particular Midwestern high school.

>> You can see this data this is anonymize network, the underlying data is extremely sensitive. When I say social network and privacy and social network, try to think about networks like this, not just about Facebook, it's about full graphs, which are outsourced for processing, small companies, small phone companies that don't have the data mining expertise and they have to outsource processing to third parties. Think about this as a social network not just things like Facebook and so on.

>> In the paper that was published in Oakland, we looked at how easy it is to identify people in completely anonymized social networks. In this case our target of records is a pure graph. Unlike the cases of Netflix, there are no attributes, just graphs, unlabeled lines. What are there to identify, it turns out that the topological structure of the neighborhood, that in itself acts as an identifying factor. That could be used to identify people with great success. The fact that people rarely two single social networks, online or off-line, if we have anonymous network and it could have a small overlap with the target network, but as long as there's some overlap, it could be used for identifying people in a completely anonymized graph. So effectively I will not go into details,

but effectively we take a small number of nodes to start mapping and then the use existing known to build the mapping. So effectively, if we see it, if you have to two nodes already mapped and then two more notes and then seems to have a relationship with the map to nodes, we try to track the month mapping and so on. I know it is a little vague, Davey is they did is that it is an iterative process where we use mapping to construct more mapping.
>> Using local social neighborhoods as an identifying factors to help us construct more mapping.
>> The important part of this process and this is something that is often overlooked by people who design technologies, is that this is an iterative process. The de- identification and the anonymous station is -- you use your exhilarating for me and to get to identify the identify the few nodes and then you used those notes as exhilarating information as the next round of identification. You use more information to continue with the DI anonymous station and so on. The technology that exists today is that they will come in and the anonymize. You use that as an input to the second round. Second round as an input to the third round. These are the truly powerful, it has been -- and networks -- -- he continued working on this showing that that you can win data mining competition using this technology a gift that keeps on giving.
>> Okay. So far so good, but at least in these cases, we are given the raw data. Either a anonymize transaction are in this case completely anonymized graph. Let's look at a different setting where we are given no raw data at all where we don't have access to the underlying data set. Let's suppose that all we have is statistics about underlying data. This two sticks, they could be computed in a pretty complex way. -- this to sticks, they could be computed in a pretty complex way.
>> - - the statistics.
>> And Amazon you have seen if you like this big, you might also like this also. They generate the recommendation. And the recommendations updates the correlation between items and they all take old transactions and run some complicated gathering and some rocket science data analysis stuff which computes correlations between items and uses them to produce recommendations. What is very important in the processes that most modern systems, this statistical racial relationship is not between people. People who good-bye X. could also buy why. Is this it is aggregated from thousands of transactions. At this point, when they have the recommendation, all your specific information is gone. All they tell you for the individual item, there are other items that are correlated with this item. This information not only statistics is not only statistics about users, it is about items. It seems completely hopeless to be able to infer anything about individuals by looking at this public statistics. It seems completely safe from privacy perspective right? and people bought the particular book and there could be hundreds of such people, can also buy at some other book, based on some 100 transactions. Clearly see from privacy perspective. -- clearly safe. We also published another paper in Oakland where we show that even though individual recommendations that you get like a snapshot of recommendations over like you might also like a list from the recommended system, but if you track their changes over time, you can figure out what are the individual inputs that resulted in those changes. We can try to illustrate this with a simplified example.

>> Suppose we know some items associated with some users. We know the user is likely to watch certain TV shows. Somehow we came into possession of this information may be from their blog were Facebook to file some other public information. This will be our -- what else can we learn about the user? the outputs of recommended -- by the way I am not just talking about recommended by this user, I am talking about global. With recommended systems do, for each items individually, they will tell us what other items are associated with this item. They will tell us for example, that people who watch house are also likely to watch the office.
>> So the this is what the system gives you for an individual item, they do not give specific user information. It tells us what other items are correlated.
>> How do you know they are not just telling you that [Captionercannot hear, the audio is low or faint.]
>> Yes, they are as a matter of fact doing a complicated model. But at the end of the day, based on computing correlations is on individual transactions, there is a complicated model. It takes a lot of effort to reverse engineer that model and to figure out what they are doing. The important thing, the reason this things appear on the list is because the correlated. By itself that information is harmless because the number of people who washed, watch this -- [audio not understandable]this information doesn't tell you enough. We will watch them over a long time. In our experiments, some ran for two years. We keep collecting the information. We keep watching how this related items list evolved over time. And then one morning, suppose we see the following. There's a particular item that appeared approximately at the same time in all the related items associated with this person. Maybe a bunch [audio not understandable] so what might be the reason for this? Some item would appear on all related items? Like the technical reason. The reason the items on the list is because it is correlated with an item. If the particular item appears on all the list that at approximately the same time, what does this tell us about the correlation with the items associated with [audio not understandable] Yes, what it tells them is that there's a stronger correlation between this new item and all items associated with this user. This means that all of these correlations increase at approximately the same time. Why might this correlation, one in three [audio not understandable] this is simplified. We are looking at 10 or 15 items, why would it increase with all 15 15 items, because they purchase that item and they started watching it and so on. See what is happening question mark not only there is a personally identifiable information, statistics doesn't tell me anything, but how they change over time reveals information. By looking at the dynamics of the statistics and how they change over time I can figure out, but they all increase at approximately the same time. The only explanation for this, of course you need to do some statistical analysis to show, but at an informal level, the only explanation is that something happened. There is underlying transaction that happened like the purchase of this item that caused the correlation of this item to increase with all items previously purchased and viewed and preferred by the user. You can see the effect and in the simultaneous increase of correlation to public recommendation. Based on global information. This means that you see a ripple like a butterfly -- and you can see the ripple in the data set and manifest itself and the subtle changes in the related items that you can use to go back and reconstruct in many cases the individual input that

caused the change. What it tells you, what does this fact that you can infer individual transactions by looking at changes in global statistics, what does it tell you? in my mind, what it tells you is this notion of personal identifiable information is larger -- largely meaningless. Even if the data is gone, you are looking at statistics, you can still learn information about individual [audio not understandable] Let me summarize a few theme of the part of the talk. It is not meaningless, it is hard to capture what people mean by this. Because you can infer information about individual behavior even from completely anonymize transactional data sets. From a completely anonymize graph. Most data sets containing human behavior, characteristics are sparse meaning that even a small amount of ancillary information is sufficient for a re-identification.
>> Also that the identification is an iterative process. Now let me talk a bit about what is to be done.
>> Yes?
>> [Captioner cannot hear, the audio is low or faint.]
>> You could. If I introduce noise in underlying data and it's misleading and so on. Our experience show that extremely hard for people to keep their persona completely separate. For example, in our experiments, and cross correlating social grab data even the small overlapping relationships, even if you have 50 friends and one network and another network, the common subset is three or four, and that turns out to be enough for identification. Creating I and overlapping persona is very hard. Even if people had completely different personas, they tend to reuse the same usernames. There's a lot of factors. It works but hard to do.
>> Of course many real humans don't even bother.
>> Let me talk about where I see -- where I think this field is going or has to be going. What do we need to do as a research community, but also people who work on real-world solutions and what they may be able to get. The first important -- that people really need to understand is that it is fairly futile to think about privacy as a property of calm quotation not a property of data. If you try to find -- and turns it into a non-personally-identifiable information, the public has to shift into uses of data, computations of data and we have to start asking questions what it means for competition to be privacy per serving. How we did design that. Things like that. That think about data, think about the computation. The first area that has seen a lot of action lately is out rhythms for privacy per serving computation. There's still a lot to be done. There's successful algorithmic framework. Differential privacy, but they are his collaborators that have taken the academic world by storm. A lot of papers and techniques. People developing systems. The idea is to give a robust notion of privacy not for the data, but competitions of data. What is differentially private is they say that every algorithm is produced -- whether any given input is included or not. Very informally without symbols, this notion of privacy's, the computation that doesn't attempt to much on any individual data is privacy per serving. Another way to interpret is the -- the risk that's an increase very much if they include a data into the data set. If you take one record, that [audio not understandable] is not too much.
>> It turns out to be a very helpful and powerful mathematical definition, good framework to define algorithms. What we don't know, -- the answer is no. If you talk to people what they mean by privacy, what

they expect, what they want, this notion does not map very cleanly onto people's expectations. Of course you can say, the expectations are not strong, realistic and so on and maybe, but the fact of the matter is often what people are looking for when they are looking for data protection is not captured by the notion. I am happy to talk off-line why this is not. This is pretty much the only notion that we have now and there's been a lot of development of interesting algorithms and we have still have some unsolved -- this is just a framework for doing things. We have a lot of unsolved problems like the -- system, like the recommender system. Without somebody being able to use the outputs to infer the inputs. That has not been solved. That is still an open problem. Another problem is analysis of genetic data and how to analyze data set like genome wide association studies that have huge amounts of input. It results in adding too much noise in the data so they don't work. Many people are interested in noninteractive data release. Finding data representation that could be completed in a private way so you can publish raw data and something that is based on raw data and can be useful.

>> Another line of research I'm interested in is the next step. After you have algorithm for privacy per serving computation, how about system for privacy per serving competition. How about building systems that I love you computing sensitive data. Some of it is implementing all terrific framework, some goes back to fundamental system problems like one problem we were interested in, how to compete on data in a way that guaranteed to leave no trace of data on the machine doing the competition. Ideally you would be able to do it in some cloud computing setting, that would be a holy grail. This way I could give my data to some thing and it comes back and I am sure there's no little bits and pieces of data lying around. This we can solve on a desktop. How to do computation on desktop and be sure no trace of computation remained except that the final output. That is a system research problem. That is not solved. Many problems of the main privacy per serving computation, not even technological problem. They are policy problems and regulation problems. A lot of existing laws that we have including very prominent laws like Hipaa and] -- ferpa they have a deeply ingrained notion of PII.

>> This notion of personally personally-identifiable information is largely meaningless, you can -- -- suppose it goes away and somebody listens to me and decide this is not good for my first laws and regulations and should not be based on personally-identifiable information, how should the write laws that protect consumer data, health data and educational data and so on, what do we actually tell data holders and their lawyers about protecting individual information?in my limited experience, I get a sense of what they like and dislike. What they like is clear rules that -- try to provide in bright lines. And waivers of liability. If you do a BNC and nobody can sue and nobody can, after you because you'd be anonymous the data according to the guidelines. Nevermind that the anonymous station did not work, nevermind that the somebody takes the data set and then shocked chunks and identify people, but it is quick that they want the waiver of liability. Here's what they do not like. And they don't like very strongly, they don't like it being sent to papers, [audio not understandable] they use a lot of Greek symbols and sophisticated math and integrals. What do people want to do, they want to do their job and share data to an extent they can. They don't want to read papers about differential privacy. They don't

like to release and share any of these data. That may be true in some theoretical or academic setting, it is true that there is no interactive data release mechanism that releases untransformed data that is de-identified that nobody can reconstruct. It is not going to work. That is not the answer they want to hear. What they want is a usable mechanism for actually releasing and sharing the data and computing on the data. Certainly for the task they want to do. And medical research they want to do a lot of computation that they want to do. Also telling them they cannot do anything without having a privacy expert in the room and reviewing everything they do, that is a nonstarter. What they need is usable systems that provide guarantees even if they are used by people that are not privacy experts. I think it's important to understand for people who try to work on designing and building privacy protection technologies that eventually this has to be use and they have users. Things that are simple and provide meaningful protection. Another aspect I am interested in is an economics of privacy. Economics of markets for personal information. I started looking at this fairly recently, last year, I was amazed by how little I was able to find out about what happened to this day to data set containing individualized information behind the scenes. The first inclination is to follow the money trail. What actually happened, who buys this information about how is it monetized and so on? That seems like a useful thing to know. To my amazement, I discovered that I did not -- markets for spam, markets for phishing information, I have never seen anything but markets of personal information, know more about what cyber criminals do with the information that is obtained legitimately so what do companies do with the information. I am not implying they are doing anything if areas. All this information about my clicks that's collected and correlated with databases that contain my financial information and driver's license and credit reports and so on, that information seemed to pass through a lot of hands and being marketed to advertisers. But never seen this trail reconstructed. Understanding the market, the economics and the incentives of various players seem important because they are doesn't seem to be any way of building robust privacy protection mechanism otherwise without understanding what drives the main players of the market. Economic models are fine. It's great that we can ride this big formula with parameters. It doesn't help if you don't know what the values for those parameters are. This is an area where we need a lot more empirical research. The same research people have been doing with great success to understand markets for cyber crime. We need to understand markets and economics for personal data. Just to summarize, outdoor them, per serving information systems for privacy per serving information, policies and Rick regulations and economics of privacy. I think the very broad areas is where most of that action will be in the next three years. This is all I have to say today. Thanks and let's open for questions.
>> [APPLAUSE]
>> Thank you, at this time if you would like to ask a question on the audio portion, please press star one on the touch tone phone. You will be prompted to state your name. Star one to ask a question. Start two to withdraw the question. One moment, please.
>> We had a lot of questions from the room, right?
>> You don't have to press star one.
>> [LAUGHTER]
>> I have a question, one moment.

>> Your line is open.
>> Thank you. Here is my question. There's been security breaches yesterday, as Apple has security breach, too. It is not clear how much people in soccer companies are worried about it. At the very end you mentioned there's a market for private data and actually we don't even know about it. There's a lot of export exchange that data miners keep sharing -- Procter & Gamble has huge amount of data about us. And they share with third-party organizations that process data for them. They are making money from data that belongs to the individual in the repository. Why would people get worried about this? they should be, but I don't see -- in other words, my question is privacy is very important problem, but do you see coming to a point when individuals will start suing companies. Companies are worried about it, and so the researchers in the policy, would it ever come to that?
>> I think it's a very valid question. It is true that -- it is not clear how much people are clear -- clear about how worried people are. We will start to see a lot of changes once by a medical information becomes more shared. People may not worry about the record of their purchase and transaction, but once we have online medical records and electronic medical records and those are correlated with other things, I think we will see people being much more aware. -- worried. I think that is a legitimate question. Some data may be more valuable than others. Understanding people's attitude toward Argosy and what they care about is important.
>> Right. In fact, we are willing to help our private data to Stop & Shop to get a few dollars off every time we buy things. For instance, the medical health data, a number of people will hold on to different records and then it does become -- the question -- it becomes a very important issue.
>> This notion of ownership is very important because it is not clear to me that in a legal sense owned by the individual who contributed the data. I am not a lawyer, but I will not be surprised that it is actually owned by people who have the record which made things a little or.
>> At this time I do not show any more questions and audio.
>> Any questions in the room?
>> The European Union has a more stringent privacy protection [Captioner cannot hear, the audio is low or faint.]
>> It is true that there is the EU privacy directive. Which is extremely broad. I am not a lawyer. This is almost a legal question, my impression is anything that has to do with the data is technically in violation of privacy directive strictly construed. It is a difficult question. In Berkeley, it does seem -- and Berkeley that companies in Europe are a little bit more cautious e cautious with sharing data you and a little bit more protective of the data. I have never seen anything any study that tried to measure this. My sense getting something like that directive adopted in this country, that is not going to happen. There's too much money at stake and not doing it their way.
>> In many ways this is not a new problem in some cases [Captioner cannot hear, the audio is low or faint.] in the security areas and information flow, open channels, hard to quantify -- I am wondering if you see any hopes in developing any kind of -- you can almost certainly identify some records if you put enough effort into it. Do you see any hopes in developing any work factors [audio not understandable]

>> Okay. I think one big hope is unlike the [audio not understandable] we can do and furcal research now, we could look how people -- we can look at actual use cases, how people use the data and what kind of computations they could do. I'm not very hopeful about designing a generic mechanism. But you could do sort of use of specific systems and so on. And for those -- data mining and biomedical data, those kinds of things maybe yes there's some hope. And factor analysis, but having actual users and data sets, I think will help us in a way that they did not have those in the 70s.
>> The main -- I'm a believer in the main specificity in the case. Adopting general-purpose computations.
>> [Captioner cannot hear, the audio is low or faint.]
>> .
>> That's right. Instead of the phone book, we have Facebook which is addition to the name and phone number have all your friends your favorite movies, favorite books, what movie and what you watched last night and where you went to dinner last night. You're right, it is all auxiliary information and that train has left the station. We have increasing amount of auxiliary information, one publisher never goes away and with the rise of social network, we have the ultimate auxiliary information now things like Facebook and I don't know what will come up after Facebook, not only do we have the source, they had aggregated in the same place for your convenience. Gold look it up.
>> [Captioner cannot hear, the audio is low or faint.]
>> a lot of people. Yes. A lot of people. Not to mention you can always correlate information in Facebook without [audio not understandable] information.
>> And photographs and [Captioner cannot hear, the audio is low or faint.]
>> Right. That's important, yes. Very nice research done by [audio not understandable] showing that photographs for example, you did not put the photograph there, your buddy did, but that's enough because you can correlate the photograph and complete the chain I think they were constructing Social Security number by pulling photographs put up by friends.
>> You could build an iPhone app.
>> You can find out which [audio not understandable] [Captioner cannot hear, the audio is low or faint.]
>> Sure. This exhilarating for me and is like my sense, any defense based on the assumption I will enumerate -- limit information I think that is [audio not understandable]
>> [Captioner cannot hear, the audio is low or faint.]
>> Unfortunately and now, is also disperse. Yes, it could happen.
>> [Captioner cannot hear, the audio is low or faint.]
>> Right, and there is cloud.
>> [Captioner cannot hear, the audio is low or faint.]
>> Right. Although it is not clear to me at that point, people do this, they really fully understand what they are giving out.
>> It doesn't matter if they understand.
>> I absolutely agree. I found that Article [audio not understandable]
>> I read it.
>> [Captioner cannot hear, the audio is low or faint.]
>> Let me give a short two-part answer. I think in general what we mean by privacy and loss of privacy is very important topic, something I

should have mentioned in the start, but I didn't unfortunately. Privacy is contextual integrity. This is -- the only convincing attempt that I have seen to try to understand from first principle -- principal philosophical and legal, what we mean by privacy. There are certain norms that govern transformation of information and the same information could be violating the norms and not violating and another context, this is solid and substantial discrimination dissemination of privacy. Helen and other people working, as far as divorcing our security from the notion of PII, that is very tricky. That is hard to do because we have infrastructure in place, authentication, things like this about how does the bank know it's you when you call them? That is so -- to the notion of equivalent of PII effectively that is just -- there's so much infrastructure already in place, it's hard to see it happening anytime soon.
>> [Captioner cannot hear, the audio is low or faint.]
>> That's right. That's right. It's hard to even imagine what he would look like without PII.
>> [Captioner cannot hear, the audio is low or faint.]
>> I don't know except I'm sorry I don't have a good answer for you. Live your life, everything you do in the Internet will be public. What else is there to do? You have to assume that all of that information, if you are willing to live with it being public, and if not, maybe you should not be doing that. I know it is not a great answer, but I don't have an answer.
>> Any other questions for our speaker [APPLAUSE]
>> Let me know if you would like to do [audio not understandable] [event concluded]
>>