FEDERAL TRADE COMMISSION

COMPETITION AND CONSUMER PROTECTION

IN THE 21ST CENTURY

Tuesday, November 6, 2018

9:00 a.m.

American University

Washington College of Law

4300 Nebraska Avenue, N.W.

Washington, D.C.

FEDERAL TRADE COMMISSION

I N D E X

P R O C E E D I N G S

DR. GILMAN: Good morning, everyone. My name is Dan Gilman. I am at the FTC's Office of Policy Planning. Just a couple of really short announcements before we get to today's program.

First, everyone ought to know that this is a public event, not just for your attendance, but it is being webcast. So you are being recorded. There will also be a transcript of today's proceedings taken and then subsequently made available.

Number two, some of you may have already gotten question cards on the way in. We have them available throughout the day. People will collect them. Staff will read them all, process them all. Some of them will be passed along to panelists during the day, not necessarily all of them, but we will take them. We are going to try and keep a prompt schedule, if we can.

So without spending any more time, I want to introduce -- oh, biographies are available. So we have very, very accomplished people here today. We are not going to recite their accomplishments at you, but the biographies are available.

I just want to introduce Professor Jonathan Baker, an antitrust scholar here at American

University Washington College of Law for welcoming remarks.

WELCOME AND INTRODUCTORY REMARKS

DR. BAKER: Thank you, Dan. I am delighted to welcome the Federal Trade Commission and the antitrust and consumer protection community to my law school. If you have not been here before, I hope you will stay some time to meet some of our terrific students and admire our wonderful facility, where we have now been for about two years.

I have served twice at the Federal Trade Commission, once as an attorney advisor to Commissioner Terry Calvani and then later as the Director of the Bureau of Economics when Bob Pitofsky was Chair.

When Chairman Simons opened these hearings in September, he said he modeled them on the hearings that Chairman Pitofsky held in 1995, when I was at the Federal Trade Commission. The Pitofsky hearings were prompted in part by two ways the economy had changed since the mid-20th Century. First, markets were increasingly globalized. In the four decades since the end of the Second World War, firms across the developed world, particularly in Europe and Japan, had caught up to their U.S. counterparts. And that created more competition for many domestic firms at home and abroad. And antitrust enforcers were increasingly detecting international cartels.

The second change in the economy between the mid-20th Century and 1995 was the growing importance and pace of technological change. You could see that particularly in information technology. This was a decade after Microsoft introduced the Windows Operating System for the IBM PC and we were right at the start of the dot-com boom.

The changes in the economy that we saw in 1995 are still continuing. International trade has continued to increase as a fraction of GDP, and although the overall rate of productivity growth has probably slowed since 1995, many of what are now the largest internet and information technology firms were just being born then. Amazon was only a year old. Facebook and Google were still to come.

The rise of the internet points to new and distinctive challenges for the hearings that the Federal Trade Commission is now conducting, particularly for the ones for this week. The transformation of information technology since 1995, and particularly the growth of online platforms, is at the heart of the novel competition and consumer protection challenges that the FTC must now address.

On the consumer protection side, online

platforms provide a new locus for fraud and deception, and the migration of personal data to online hosts creates new privacy challenges.

On the antitrust side, if you credit the recent economic research that suggests that market power has been on the rise for decades, which is what I talked about last month on the opening day of the hearings, then it is natural to ask whether increasing market power is related to the growth of information technology generally and look closely at the conduct of the internet giants in particular, including the way they develop and use data about their customers and their suppliers.

So the issues that the Federal Trade Commission is concerned with this week are at the center of the new challenges for antitrust and consumer protection that are created by the 21st Century economy.

On behalf of the American University Washington College of Law, I am delighted to welcome everyone to this important two and a half day conversation.

So let me now introduce one of my successors as the Director of the Bureau of Economics, Ginger Jin from the University of Maryland, who will give us an introduction to the economics of big data, privacy, and competition.

(Applause.)

THE ECONOMICS OF BIG DATA, PRIVACY, AND COMPETITION - AN INTRODUCTION

MS. JIN: Thank you so much for having me. I appreciate the opportunity to share my thoughts about big data with you.

As an economic researcher, I had done some research about markets with asymmetric information, but not data or privacy-specific before I joined the Commission in 2015. However, the precious experience at the Commission has exposed me to a lot of cases in data security and privacy, which pushed me to dig deeper into the market and think hard about the potential benefits and risks related to data collection, data use and data sharing.

I remember at that time, when I started this learning process, I felt that I am on a fast-moving train, but I am not sure where it is going. Two years later, even after I had returned to economics, I think the speed of the train has been faster than I thought and the destination is even fuzzier. So, as a result, I have a lot of questions in my mind to which a comprehensive and a satisfactory answer is yet to come.

I hope hearings like this and before and after this would provide opportunity for everyone to think about this issue, to chime in with their own opinion, and really form a collective wisdom. And this collective wisdom, I believe, would have an impact for our policymakers to make informed decisions.

So today, I would just probably organize my thoughts in an economic framework. It probably is not precise to call them thoughts, but just a list of questions, and hopefully that will stir conversation in the two and a half days of this hearing.

So the first question I asked myself is, what is going on in the marketplace? And to begin this question, I want to look at the kind of players in the market. We are all familiar with the role of firms here, but I want to make some comment about consumers, government, and research institutes.

So consumers in the data market are not just consuming products and services backed by data. They are also active data providers and data users. How many of you have, say, a smart watch on you sometime during the day? Some of you.

So you can see from these kind of devices and online apps that we are constantly providing data to the app. We are also consuming data from that. We want to know the statistics, how many steps we have walked today and how many miles we have run, and so forth. So this is a very active data exchange between consumers and firms. So consumers are not passive sort of consumers of the products generated out of data; they are also actively participating in this process.

And to some extent, the Government is similar to consumers. They consume data. They also provide data. However, the Government has the power to make new legislation about this market. They can designate certain law enforcement to enforce the law. So in that sense, the Government is both a player and a referee. So I think that combination probably will make Government's role distinctive from all the other players here.

In terms of research institutes, here I want it to be a broad definition, not only economic institute but also, say, think tanks, consumer groups, even industry associations. And those institutes, we are -- as an economic researcher, I can say that I am always hungry for data to make my research more insightful. But, on the other hand, we also want those research institutes to be kind of a third party to describe the marketplace to us from an objective point of view. So I think that role probably individual consumers cannot play, but will be very important in this marketplace.

In terms of exactly what is going on, I hope this hearing and other hearings would shed more light on who generates most data; who uses which data for what purpose; where and how does data stay, flow and evolve; and how does technology reshape data and data use; who benefits, who loses from certain data practices; and what is the aggregate consequence of data use in the short run and in the long run; and what is known and what is not known, to whom and when.

I really think those questions have to be addressed by probably a multidisciplinary approach, not only from the Commission's own research report, which has been done in 2014 and 2016 about data, but also from, say, computer scientists, economists, law professors, or even psychologists, to really help us understand how each player works in this space. I would encourage all the think tanks and organizations to contribute to this, as well. Of course, firms should give us probably a more intimate view of exactly what they have been using the data and what thoughts they have had when they decide the policies about the data use. So I hope this afternoon's session about the business of big data would really give us more insights on this.

So suppose we sort of understand how the market works, probably we should ask, is there something wrong, and what goes wrong? And as an economist, I often try to think of that question as where does the market fail? We cannot just say this is an issue and then jump directly into intervention. We probably have to ask, to what extent that the market is able to address that question, okay, and then where the market is not able to address that question.

So following that line, I am thinking about the textbook examples of market failures, and there are typically four of them. The first one is well known, market power. There is a long history of antitrust talking about this in monopoly and oligopoly, market structure. The second one is information asymmetry. The third one is externality. The fourth one is bounded rationality.

And I want to push the audience to think exactly whether and how does big data contribute to these market failures, okay? I want to be a little specific. For example, if you think about potential market failure from market power, does data constitute barrier to entry? Does data facilitate collusion between oligopolic firms? Does data facilitate anticompetitive contracting? Does data facilitate perfect price discrimination? And on the other side, data could also generate merger efficiency or contract efficiency.

Based on my experience, I think the potential anticompetitive practice related to data is more often a theoretical possibility than a widespread practice in the real world. I am happy to be corrected by maybe tomorrow’s panel discussion on this, and if there are more evidence towards anticompetitive direction, I will be really happy to be corrected.

So if we identify some contribution of big data to the anticompetitive problem I listed here, I think that still has to be translated into what is the overall impact of that practice on consumer welfare, both short run and long run. That is sort of where the real and tangible harm should be associated with big data before we take antitrust action towards that.

Okay. The second one is information asymmetry. I know not all of you have economic training here. A very textbook example about information asymmetry is prescription drugs. That is, we, as consumers, we do not know exactly what is in that particular pill. The firms probably can do some advertising telling us that, okay, we really have a cancer cure in that tablet. However, even after we consume it, we still cannot tell whether it has really cured our cancer because there are so many other things going on. So this is a very typical information asymmetric problem because the firms know more about the product than individual consumers.

If we sort of borrow that kind of mind set into the data-related issues, then I would say the information asymmetry associated with data is probably even more complicated than prescription drugs, in the sense that we not only have information asymmetry before the focal transaction, consumers do not know how they are going to use that data for the particular transaction, for example. But, also, a lot of asymmetry would arise after that focal transaction. We do not know how the firm is going to store the data, to what extent they are going to change the content and format of the data, and to what extent they are going to sort of link that data with something else, okay?

This is not only just the information set of consumers at the point of focal transaction or after the focal transaction, but, also, sort of, what is the information set of firms as time goes on, right? They may not know exactly what they are going to do with the data, but they will have some say in how they are going to use the data later on. And that question also relates to affiliates or even nonaffiliates of the firm, if they are going to share the data with the firm.

And I would also add black-market players like hackers and the public here because we know in incidents like data breach and other things, that -- maybe this is an unintended data use, but it turns out to be a potential data use in reality.

So coming back to this core question, what is the harm to consumer welfare from the information asymmetry problem of data, and where does it show up and how much is it? Can we really quantify it?

So the third market failure, the potential market failure, is externality. What is the typical example of externality? Let's say air pollution, right? We could have a lot of firms producing harmful gas into the air. We, as, say, the general public or the consumer of air, we sort of probably can tell the air does not smell right, and we can do some lab tests showing that there are some harmful components in the air, but we do not know exactly which firm contributes to that air pollution.

And this negative externality is not taken into account by the firms in their market practice, which generates this negative externality problem. If we bring that mind set to the data issue, there could be questions like, what data practice would generate what spillover? And we know that according to the Bureau of Justice statistics, about 7 percent of American people above the age of 16 is a victim of identity theft, and a lot of identity theft are related to data issues.

However, even if I am a victim of identity theft, I do not know exactly which of the hundreds of firms I interacted with in my past will sort of really contribute to this event of identity theft. In that sense, it is kind of a similar problem of negative externality as the air pollution I just talked about. Okay? So that is just negative externality.

There could also be positive externality in the sense that we know if a lot of data sets pulled together would really help, say, the census or researchers using the census being able to generate research grade outcomes. However, each firm may not have the full incentive to share that data because they are not going to get all the returns from that data use. So in that sense, we could even have positive spillovers which generate an under-incentive to collect and share data.

So I want this hearing -- I am hopeful that this hearing will talk about the externality issues in data and to what extent the parties that generate that spillover have the incentive to internalize that spillover, and how does that spillover affect consumer welfare.

So the last potential market failure is bounded rationality. We know a lot of us have been sophisticated, but we are not as sophisticated as the machine could be or as a rational agent in an economic model would assume. So we always have some level of sort of standard rationality or you can say the rational choice of not paying attention. And this could happen in this area.

And we know, thanks to researchers like Lorrie Cranor that -- we know ten years ago that very few people actually read privacy policy. However, we still have that as one of the main building blocks for today’s data space. So exactly how consumers, how individuals deal with this kind of information presented in front of them when they have very limited attention, but a lot of information to digest. Okay?

On the other hand, firms probably are hungry for data, and they have more resources to deal with the data, and they can employ or even invent technology to process data. So in that sense, my view is the asymmetric information between the consumers and the firms have been magnified by this advance. On one hand, the consumers are driven by inattention, they want quick and straightforward solutions. On the other hand, the firms are really churning up a lot of resources and technology to try to digest as much information as possible.

So that brings a question of who has more bounded rationality in this marketplace? Who suffers from bounded rationality, and whether some parties would have incentive to exploit other people's bounded rationality. And, again, I want this to sort of boil down to exactly how does this bounded rationality affect consumer welfare.

Okay. So that is kind of market failures from the economics point of view. And suppose we identify one or more market failures in this area, then we could talk about a bunch of potential solutions. Here, I am putting kind of a spectrum from free market to having prescriptive regulation from the

Government. Okay? So in the middle, we could have industry self-regulation, some guidance to the industry firms and somehow there is a mechanism for firms to conform with that, or we can sort of strengthen that by more external monitoring, like the consumer education effort, as well as societal monitoring, and all these probably not involve government.

If we could push it a little bit further, we could have government involved in *ex-post* enforcement and that is kind of like, say, nutrition supplements, right? Okay, you can put the nutrition supplements in the market without going through the FDA and clinical trial. But if something goes wrong with that, then law enforcement effort would come in and to try to correct that. So that is probably less aggressive than the FDA approach, say, in food labeling or drug clinical trials.

And that brings me to the *ex-ante* regulation, that we could have heavy-handed regulation like define exactly what you can say, what you cannot say, we are going to find a way to confirm that what you said is correct. We can sort of inspect you saying you have to do A, B, C before you produce a product, because we believe A, B, C is kind of good in ensuring the quality in the final product, or we can even impose a minimum quality standard on the final product you eventually produce, like a clinical trial to make sure that a drug is safe and effective in addressing certain diseases.

We can combine both the ex-ante regulation and *ex-post* enforcement, and sort of having this in a dynamic sense that we can revise our legislation given the new questions coming out and so forth. So I want you to have this spectrum in your mind when you think about what is the potential solution and what is the tradeoff of each solution.

So now, suppose we sort of agreed on which solution we are going to get, and then the question is exactly how we get to the ideal effect of that solution. I have heard people talking about using existing rules, such as competition law and consumer protection law. And I guess the immediate question is, how do they fit in this overall framework I just discussed about market failures and the potential solutions?

And the second question is, what is the relationship between the two poles, okay? They could be sort of -- let’s say on your left-hand side, I put it as a leverage, like the two could be conflicting with each other. Let me give you an example. So antitrust may concern about data not available to a potential entrant into the market and, therefore, push for data access, data portability, and data standardization. However, the consumer protection part may worry about that there might be some unintended use of the data and, therefore, the consumer should have a right to restrict how their data should be used. And that could generate an effect that actually reduces the potential entrant's access to the data and the data portability.

So in that sense, these two may be just sort of contradicting with each other. Is that the world we live in, that we have to find the balance point between the two, or maybe we sort of need the two gears to work together?

Let me give you another example. Say we have a lot of data policy, they are very long, legal language, and hard to understand. If there is no sort of consumer protection enforcement on how clear this policy must be -- and firms may find that the more obscure the language, the better I can get data and really benefit from it, and then promoting competition, actually would push firms to compete in that particular dimension, which means the data available to consumers -- the data policy available to consumers become more and more obscure. So we could talk about like competition in the wrong dimension.

So in that sense, we want the two gears to somehow work together in a complementary way. So I hope the hearing would sort of promote a discussion on exactly what is the relationship between these two existing tours.

Okay. So there are a lot of questions on how to exactly carry out the solution. I would just list some questions here for the base of discussion. For example, should we aim for the legislation to be very comprehensive and detailed or shall we leave the detail to the regulatory and enforcing agencies? There are arguments in both ways.

Who should be this regulatory or enforcement agency? Should that be one or should that be multiple agencies? Should that be, sort of, at the federal level for everything or should that be at both federal and the state level or just the state level? Should we do this industry-specific or should we cover all industries? And there are questions like the degree of enforcement and regulatory freedom, the resources and expertise available to this or these enforcement agencies.

I want to make the extra point here that whatever the agency that the Congress have determined to give power to, assuming that we sort of agree that it is necessary to have such an agency to do their enforcement and regulatory function, I think we should think hard about how do we to limit the agency's power in terms of should we define who this agency should report to, how transparent their practice should be, and how can we make sure that this agency's action is accountable. If they do something over the defined area, how can we correct it and how can we bring external forces to really spot and correct those kind of wrongdoings?

So in that sense, I hope other parties will be able to contribute to that solution, even after we have decided exactly how to carry out that solution. And given how fast technology is moving in this area, I think it is really, really important for all the parties I listed here to continue contributing to that solution on an ongoing basis.

I only have two minutes left so let me make the final comment about international complications. Every country is doing this slightly differently. I think, to me, there are sort of three models at least coming out of this heterogeneity. One is the European model, that they have a comprehensive framework covering all countries in the EU, which is GDPR, and they have DG-comp in the antitrust agency for the EU. But they also have country-specific enforcement, especially for GDPR. Okay? So that is one model.

Another model is sort of the U.S. status quo. We have a patchwork of federal, state, and industry-specific enforcement and they generate some heterogeneity even within the U.S.

And then the third model is the China model. They have nationwide laws in 2017, I think. We do not know exactly how they are going to enforce that yet. But we also know that big data could be an input for government censorship and surveillance there.

So I am not saying that I have a good idea of which model of these three is good or is better than others, but I think it is really important to discuss the pros and cons of these approaches. This is not only because companies are global and they have trouble conforming with all kinds of different regimes, but also because -- I think this is more important -- but also because data, ideas, talents, and the money flow globally. Okay?

So that means if in one corner of the world they have very prescriptive regulation, maybe the money and talent and idea would go somewhere else, okay? And what is the implication of that for the whole economy in terms of consumer welfare, as well as the future innovation and support of the economy? I think that is a very big question. So I am going to stop here.

Thank you very much.

(Applause.)

DR. GILMAN: Thanks very much, Ginger. We have a break scheduled now. I would just ask, you are getting out a little bit early because we started a little bit early, I would ask people to be in their seats promptly at 10:00, so we can start again on time. Thanks very much.

THE ECONOMICS OF BIG DATA AND PERSONAL INFORMATION

DR. SANDFORD: Okay. Good morning to those in the room and those watching on the webcast. This is our panel on the economics of big data and privacy. We have five panelists here to share their views on how markets involving big data and privacy function.

We have Alessandro Acquisti from Carnegie Mellon University. We have Omri Ben-Shahar from the University of Chicago Law School. We have Liad Wagman from the IIT Stuart School of Business in Chicago. We have Florian Zettelmeyer from the Kellogg School of Management at Northwestern University. And we have already heard from Ginger Jin, who is from the University of Maryland.

My name is Jeremy Sandford. I am an economist at the Federal Trade Commission. I work in antitrust, and for the most part, my colleagues in consumer protection at the agency are those that deal with big data and privacy issues. So, hopefully, this mismatch is a feature and not a bug.

The reason we have an antitrust person moderating this panel is, well, there have been calls for increased antitrust enforcement of big data and privacy issues. So, for example, Joe Stiglitz, speaking at an earlier hearing, shared his view that big data and privacy represent one of the biggest challenges to our society and to competition law. So we kind of want to get at the question of should we be doing something different with respect to antitrust when we have, say, a merger or single-firm conduct that involves big data or privacy.

My focus on competition is not a constraint on the panel or their opening statements. You all can talk about whatever you want and we are going to hear from our panel on kind of their views on how these markets work. And then I am going to ask questions that are going to kind of get at are there competition implications for big data and privacy markets that we may not be taking into account with the way we do things now.

Okay. So we are going to proceed as follows. We have already heard from Ginger, so she is not going to speak again. But each of the four remaining panelists will have up to ten minutes for opening remarks and then we will have a Q&A session where I will ask questions and the panel will answer.

If you are in the room here at American University and you would like to ask a question of the panel, we will have people going up and down the aisles with note cards. You can flag one of them down, get the note card, write your question on the note card, and someone will bring it up to me, and I will see what I can do of asking those questions.

So the order of speakers will be alphabetical. So we will have Alessandro, Omri, Florian -- sorry. Alessandro, Omri, Liad and then Florian.

DR. ACQUISTI: So good morning and thank you so much for the invitation. And, more importantly, thank you to the FTC and American University for creating this forum. The quality and diversity of the speakers is -- should I push something?

Thank you so much. So I guess you heard my thanks. And I was adding that the quality and the diversity of the speakers is exactly what we need to bring nuance and some degree of clarity to a complex topic.

And in my remarks, I will focus on two different areas. First, I will go broad and propose some personal framings, some ways to frame the debate over big data and privacy. And I will focus in doing so on two apparent issues, yet common misconceptions, which we, as scholars, are aware of, not often they are properly understood in the public debate over privacy.

Second and next, I will go narrower and I will present some ongoing, yet unpublished, work we are doing on the topic of the allocation of value created by the data economy. Okay?

So starting from the framing of the misconceptions, the first misconception is that privacy and analytics are antithetical. You can have one or the other, but not both. You find echoes of that stance already back in the days in the writings of scholars whom I actually greatly admire and respect because they were the first scholars to bring economics to the field of privacy, Chicago School scholars such as Posner and Stigler, who conceive of privacy as effectively the concealment of information, the blockage of information flows.

Now, we know from the case of work on privacy that a much more nuanced, and I would say, precise view of privacy is in terms of management of information flows, not blockage. It is -- sharing a secret with a friend or posting some information on social media and choosing the visibility setting for the post are sharing behaviors, which are also privacy behaviors. They are privacy behaviors because they encapsulate the ability to manage the boundary between the self and the others, which is far from the notion of privacy as a blockage of data.

Why is this important? It is important because once you realize there is more -- in yourself there is more than one view of privacy as management of this boundary between privacy -- between private and public, then you also realize that it is, in fact, possible to have simultaneous privacy in analytics to protect certain types of data and share certain types of data.

We can do so through truly an actionable, informed consent, something that I do not believe is very common nowadays in the privacy landscape. We can do so through smart regulation. We can do so through privacy-announcing technologies. The best of these technologies do not block data; rather, they try to modulate what data is protected, what data is shared in the interest of increasing welfare of different stakeholders.

The second and a related misconception is that the relationship between data protection and generation of economic value is a monotonic, specifically data protection is always welfare-decreasing and data collection is welfare-increasing. In reality, both in theory papers and empirical ones, we have a much more nuanced view and we realize that the economic impact is very much context-dependent.

For instance, healthcare privacy regulation, if done improperly, could slow down technological innovation in healthcare -- Amalia Miller and Catherine Tucker have important papers in this area -- but if done properly can actually increase innovation, which is something that we found and published in *Management Science* with Idris Adjerid and Rahul Telang. Social media can lead to better matching in labor markets, but can also lead to more discrimination in labor markets. So it is always context-dependent and we should be very, very cautious about taking a one-size-fits-all when we think about the relationship between data and economic value.

I can offer you two further examples of this from scholars who certainly cannot be accused of being against efficiency and against data. The first example is again from scholars I admire from the Chicago School, in particular Posner again, who noticed already in 1981 that privacy is redistributive. The point he was making was that data protection creates economic winners and losers. Now, I believe he is right, but it also turns out that the lack of data protection also creates economic winners and losers. You just cannot avoid this.

And the second example, which is related to the first, is from Hal Varian, who in 1996 pointed out how consumers may rationally want marketers to know their preference so they get offers which are of interest to them. But they also may rationally not want marketers to know their willingness to pay in order to avoid being price-discriminated. The first desire is welfare-increasing for the consumer; the second is to avoid a situation which is welfare-decreasing.

So the lesson here is to be watchful of arguments, such as data protection is monotonically increasing or decreasing value. The reality is much more nuanced and context-dependent, which brings me to the second part of the talk, where I present some ongoing results from studies we have been doing trying to disentangle these nuances.

I will focus in particular on targeted advertising. The reason is that targeted advertising is afflicted by what I was referring to earlier at the beginning of my talk, some of the misconceptions in the public discourse over big data and privacy. There is a sort of magical thinking happening when it comes to targeted advertising, which is reflected in the following words. I am going to cite some words. I am not -- in the privacy spirit of the panel, I am not going to cite the person who wrote them because I do not want to make this an attack on the person. It is a critical argument.

Targeted advertising is not only good for consumers. It is a rare win for anyone. It ensures that ad placements display content that you may be interested in rather than ads that are irrelevant and uninteresting. Advertisers achieve a greater chance of selling the product. Publishers also win because behavior targeting increases the value of the ad placement. So basically, everyone benefits from this.

Now, at first glance, this seems plausible. The problem is that upon further inspection, you realize that there is very little empirical validation in all these claims. I am trying to choose my words carefully. I say there is very little empirical validation. I did not say that there is a disproof. What I am saying is that we actually do not know very well to what extent these claims are true and false. And this is a pretty big problem because so many of these claims are actually accepted unequivocally and they are quite influential in the public debate over privacy.

Why am I claiming that we actually do not know whether these statements are correct? Two reasons. The first reason is that, for all the focus on transparency, the data economy is remarkably an opaque economic black box. For the outsiders -- and outsiders could be maybe the merchant buying online ads or the publishers showing on their websites the ads -- it is very difficult to know what happens inside a black box of the different ad exchanges.

And we have evidence of this from lawsuits and scandals, which have arisen repeatedly in the last few years. *The Guardian* finding out that Rubicon, an advertising firm, retained substantial undisclosed funds, in addition to the fixed percentage fees. We found -- another example of that with Index Exchange, which was using bid caching and gaming auctions for 50 percent of impressions. We find evidence of that in Facebook hiding inflated video ad metrics about ad watching for over a year and these metrics of ad watching were inflated up to 900 percent. So that is worrisome.

The second reason why I claim that we have little validation for one side or the other of the argument is that much of the seminal groundbreaking and high-quality work in this area on targeted advertising from academia focuses, and necessarily so, on very narrow goals, such as what happens if we use targeted advertising rather than untargeted advertising? Are consumers going to click the ads more? And are the merchants going to see a higher commercial rate? And the answer is typically yes and yes. And this is an important, valuable answer.

What that answer misses, however, is the broader picture. What happens in the overall ecosystem? What happens to consumers who do not see those ads or if they see them, what happens if they end up buying something? What would happen, what is the counterfactual if the agency in the ad would have bought a similar good or a higher-priced good or a good with a lesser price, higher quality, lower quality? What happens to the merchants when they start getting engaged in a prisoner's dilemma style dynamics where they have to use targeted advertising because otherwise their competitors will be poaching consumers away from them precisely using target advertising?

So I am referring to more general economic equilibrium kind of analysis. And this is what we will be trying to do recently as well for the past couple years in my research team.

I will end by mentioning very briefly the research we have been doing. One year ago, at PrivacyCon, we presented some critical work suggesting that when you account for the different type of data that ad exchanges can use and share with merchants, you will have varied welfare implications for different stakeholders, consumers, merchants and other exchanges.

Since then, we have been doing empirical work and I will give very brief examples of these studies. In one study, we have done a lab experiment seeing how consumers react in the presence or absence of ads when they search and try to buy products online. We found that actually there was no difference in amount spent and the satisfaction with the products purchased in the presence or absence of ads.

In the second study, we have been gathering data about the prices for goods in organic search results and sponsored search results. We found that prices for goods are, on average, slightly lower in sponsored search results. However, the lowest prices are more likely to be found in organic search results rather than in sponsored search results, so for the ads.

And, finally, we have been doing work with a large American publisher from which we got millions of transactions related to the ads they show on their website. We were trying to see how much more revenues they get from ads which are behaviorally targeted versus those that are not. We can do that because we can see whether the visitor added a cookie or not. In the absence of the cookie, it is not possible to target the ad.

What we found is that, yes, advertising with cookies, so targeted advertising, did increase revenues but by a tiny amount, 4 percent. In absolute terms, the increasing revenues were $0.0008 per advertisement. Simultaneously, we were running a study as merchants buy ads with different degree of targeting, and we found that for the merchants and buying targeted ads over untargeted ads can be 500 -- sorry, 500 percent times as expensive.

So although these -- we have to be careful in comparing the numbers -- nevertheless, I leave with the rhetorical question for all of you to consider, which is how is it possible that for merchants, the cost of targeting ads is so much higher whereas for publishers, the return increased revenues for targeted ads is just 4 percent.

Thank you.

DR. SANDFORD: Thank you, Alessandro.

(Applause.)

DR. SANDFORD: We will now hear from Omri Ben-Shahar.

DR. BEN-SHAHAR: It is always fun and a challenge -- it is not always -- they did not have many opportunities, but it is fun and a challenge to go after my world's all-time favorite privacy researcher, Alessandro, and it sounds fascinating. I should give you my time to tell more about what you are finding because this is really interesting.

I guess, first, I want to apologize. I will speak and participate in the panel, but about half an hour before it ends, I have to run to the airport. I have a 3:30 class that hosts a speaker in Chicago that I cannot miss. But thank you for inviting me to take part in this.

I am not really a privacy expert. I guess I was invited because I circulated this summer a working paper titled "Data Pollution." I thought I was the only person who thought about it until I heard Ginger also discuss the idea of pollution as a metaphor to thinking about what is the problem that we want to address before we identify how we address it. And so

I will briefly discuss what my thinking is in this context.

So data policy is focused on privacy, on harms, potential harms, potential injuries, potential reduction in well-being for the people whose data is being taken, used, shared, lost, and so on. And I suggest that there is an additional perspective that can be used to understand the discomfort that people report that they have with the data economy, and that is that the data that is being collected and used, that databases affect others not in these databases, affect an environment, affect an ecology, affect individuals who are not part necessarily to that data. So there is potential negative externality.

I would also want to save a minute to talk and to think about externality as a problem not just of negative but also positive. Data has immense positive externalities.

What got me to think about this, for a while, I have been kind of -- my area is consumer protection, consumer transactions, consumer contract law. But I have been kind of trying to chime in on debates on privacy, data privacy. I have found that the thing that drives most of what -- of my thinking is what is known as the privacy puzzle, that there are

-- privacy experts and advocates really want to do something about a phenomenon that most users seem to be indifferent about.

They might say in surveys that they want data to be regulated and that there is a problem and -- but they behave as if there is not, and personally, I was very uncomfortable in the aftermath of the Cambridge Analytica and those in the Facebook fiasco. And I asked myself, what is going on? Why is everybody talking here about privacy when the problem is something bigger than the harm to the individuals whose data was used and circulated to make political lies more effective, that the harms were greater than the harm to these individuals.

Namely, there is a problem of -- I thought of it then of pollution, of an entire environment, ecology, being harmed by the practice. Then I started looking and finding many other examples in which this is the -- a year ago there was the Strava fitness app case, in which it turns out that people share where they run and swim and jog and bike, but you can see where there are clusters of users including American troops outside Niger or in Afghanistan or places like this, not good for national security or for the group as a whole. But, again, it is a problem of public good, not of a private good that is affected.

A lot of the -- I also thought that a lot of the data security breaches, Equifax to name one, represent not so much a private harm, but a public good harm. Most people whose data was lost will not be harmed. Those that will be harmed will have -- a lot of it is insured in one way or another. There is -- I do not want to diminish or miscount the important insecurity that is being sensed, but there is an insecurity that is shared by everyone. It is kind of a public -- it is a sense of a degraded environment again.

So if the problem is not a problem of externality, you want to think about it in the way that we have been trained to think about externalities, and there is a great model. Data is just the new -- now, this is a cliche by now, but it is just a new fuel. So let's think about the carbon fuel of the 20th Century and how in the 1960s and '70s and '80s, regulation began to take over private law as the method to curb the problem of externalities from carbon pollution. We realize that tort suits are failing.

And we are realizing now, if you look around, and I can -- you know, many lawyers can attest to that, tort suits in the context of data harms are largely failing, because it is hard to prove causation when Equifax loses your data, how do you know that you are harmed, that your identity theft is related to that and not to something else? The latent effect of the harm and the slow gestation period, exactly the same doctrinal reasons that we had the failure of tort law in the pollution context is failing now. Contracts, of course, are not going to solve the problem of an externality. People are not going to contract for low-emitting products whether they emit carbon or data pollution.

So it is -- part of what I did in my study is look at the case law in the era that led to the emergence of environmental law and the EPA, the private law failure that led to that emergence. And I see fantastic parallels from the analytical point or the conceptual point of view to the situation of private law today in an attempt for lawsuits to take -- to regulate the data economy.

So if private law fails, maybe for the same reason that it failed in the carbon pollution context, maybe the regulatory approach to environmental -- to industrial pollution should enlighten us into thinking about how to deal with data pollution with the external harms that data produces, and this is maybe a little bit similar to how Ginger previously, at the end of her slide, presented it, but I want to say a few things that were not there, although you probably could foresee them.

Environmental law uses three basic regulatory tools, command and control, quantity restrictions. You can only pollute so much. You can only produce so much. Carbon tax, Pigouvian tax, and liability. Now, the GDPR is a type of first -- the first version. Right? Data minimization, data localization, what data you can collect and what you cannot do, this is probably the right way to deal with some of the problems, the problem -- the concern is, of course, that in this area is that it is hard to foresee the problems that will arise and to restrict data only to places where it is harmful and not to also wash out all the potential -- the good effects of data, the immensely good effects of data.

So it is a -- you know, while obviously that is part of the solution, it is a very risky solution. It has high -- some benefits, but could also have high cost on innovation. So I tried to focus instead on solutions that were not yet developed in the privacy context to think about the data public harm context.

So one is data text. Now I know it sounds a little bit crazy. I am just kind of throwing a benchmark idea. What if we could -- if people use data to pay instead of cash, to pay for the services, for search, for social media? Cash is costly. You use it to pay. You cannot buy other private goods. Data, you can keep paying with it and create negative externalities, share the data about your friends, share -- let Gmail collect the data about messages you got from others who are not Gmail users, things like that that affect others. People seem to be largely oblivious to using that and they should not be.

So conceptually -- it is very hard to implement, but conceptually, that problem could be solved by a data text, not a data text that the collectors necessarily pay but that the users that use data as currency have to pay. Now, it really does not matter from an economics point of view who pays for the seller or the buyer. The transaction has to be taxed.

This is not a transfer of payment from one site to another to change the distribution of wealth. It is to solve the problem of negative externality. So that is one idea that I put out in the paper, that I set out in the paper, examine a lot of implementation issues. And I do not propose it. I am just saying that this is one way to think about the social cost of data.

Another aspect is to think about liability. The third form of regulatory -- third regulatory technique is liability. And here I am thinking about -- mostly about nonintentional omission of data, namely data loss, data security breaches. It is very hard to hold these companies liable for -- it for -- I said in private law, but we do think that there is and I think the FTC -- I have seen previous FTC reports about the estimated social cost of these data emissions so why not use something that has been developed in the pollution context, and that is proportional liability.

You do not pay to this victim her actual harm, but when the activity that creates the potential loss, the externality occurs, there should be payment out by the tortfeasor, by the injurer -- it does not matter who it goes to, to the FTC, to the Government -- a fine that represents the expected harm.

So here, too, we have to come up with a measure of what is the average cost to a user, to a consumer whose information Equifax lost. It could be a few hundred dollars. It could be less. It could be $10. But there are 143 million of them. So something has to be borne by Equifax, which currently is very hard to do in private law. So I talked about data tax and proportional liability.

I will end by saying that I think that this framework helps resolve one of the kind of nagging problems in thinking about data policy and that is the well-known privacy puzzle. Why do people say that they care about data security and data privacy and behave as if they do not? Well, my suggestion is that they are saying that they care about something about the ecology as a whole, about the environment. People can be environmentalists and still fly in and out from Chicago to D.C. for every panel and use a lot of carbon.

(Laughter.)

DR. BEN-SHAHAR: The private behavior does not necessarily tell us about the extent in which we all believe that there is a public pollution problem to be dealt with. Thank you.

(Applause.)

DR. SANDFORD: Thank you, Omri.

We will now hear from Liad Wagman.

DR. WAGMAN: Thanks for having me. So I

want to talk a little bit about costs and Omri talked about the costs of data. I want to talk about the costs of privacy.

And I started studying privacy from a modeler's perspective. I modeled consumer surplus as a function of, say, privacy regulation or the cost of privacy. So imagine you could have the strictest regime where everybody has privacy. Everybody is anonymous, say, in front of sellers. Or you could have something in the middle where everybody can choose to become anonymous. Or you could have something on the other far end where everybody is known. Okay?

And the result of this kind of modeling showed that consumer surplus is not necessarily monotonic in the cost of privacy. In fact, it is often not monotonic. And that means that maybe there is some optimal cost of privacy.

That led me to another question. What if we could look at firms that need data in order to service consumers, say, banks, lenders? And with those firms, even in a competitive setting, would they collect an appropriate amount of information or would they collect too much? Even if they had no reason to collect other than to service the consumers, not to offer them other products but just to sell them one product. And the result was that they collect too much, and why do they do that? Well, because they want to offer lower prices. And how do they offer lower prices? By better fitting the consumer to the product. So even in a market where data has no value other than to screen consumers, too much ends up being collected.

And that brought me to the next question. What if firms could -- sorry. Wrong button. Wrong button. It just keeps going. Further back. Okay. I guess these slides are not there. It is okay. The panel slides? That is all right.

The next model was one where those lenders could actually sell the data downstream. They could sell it to, say, insurance sellers. There we go. And in those cases, firms actually collected even more information. Okay? Now, is that good or bad? We took the model to the data and the result was that that could actually benefit consumers. Specifically, we looked at five counties in the San Francisco metropolitan areas. Three of those counties adopted an opt-in approach, where you cannot sell consumer data unless the consumer explicitly gave you the consent do so. And the two other counties, specifically the County of San Francisco and Marin, had to opt-out approach where they could sell consumer data unless the consumer actively opted out.

It turns out most consumers just do not bother. They just go with the default. So if the default is that you need to give consent, you never give consent. And if the default is that you need to actively opt out, you never opt out. Okay? So effectively, these two regimes resulted in a regime of privacy and a regime of no privacy. All right? One where your data could be sold and one where it could not.

Now, when your data could be sold, prices were lower. And in the downstream, there were less foreclosures. So in some sense, consumers were better fitted with financial products. So here we see, sure, we might like that our data cannot be sold without our explicit up-front consent, but there are costs to that. Costs might be we pay more. The other cost might be that we are more poorly matched with products.

So that led me to a bunch of other models where I wanted to see what happens if we cut off firms' access to consumer data. And those are widely spread models. Those are models that I used in antitrust cases, for example. And I looked at the results for each of these in terms of consumer surplus, firm profit, whether some consumers prefer privacy or not, and overall welfare. Now welfare in the sense you pay more, you pay less, welfare from the perspective of prices.

So interestingly enough, in almost all of these models, consumers were actually worse off in an overall sense when their data could not be used to target offers to them. Now, of course, there is no intrinsic benefit to privacy modeled here. This is all about prices. Now, firms actually could benefit because the restriction not to sell data acted as some sort of a solution to this prisoner's dilemma where we are competing on fewer fronts now. It actually led to higher profits.

The next question with this model was what if we are looking at a merger case where, say, we have three firms in the market and two of the three are potentially merging? What would happen to consumer surplus in this case if, on the one hand, firms could access data and on the other they could not? And the result was kind of not what we expected. Okay? Merger policy turned out to be even more lenient when firms could access data. It was easier to approve the merger when firms had access to data.

And the reason, again, was that firms competed on all these fronts when they had data. They could segment the population where that led to more competition and that resulted in lower prices which increased consumer surplus. Okay?

So we tried to extend this. We looked at a variety of market structures. You can think about firms being spread in terms of consumer tastes and some firms may have more customers buying from them. Others not. And if we think about firms A and B merging in this context, then the picture on the left depicts the cases where consumer surplus actually does not suffer much as a result of the merger. Specifically, those areas that are shaded dark basically represent market structures where it would be easy to approve the merger because of the fact that firms have access to data. Okay? So data does influence or should influence merger policy.

So this brings me to the final topic that I will discuss later today, as well. We just recently started looking at the effect of the general data protection regulation in the European Union on investment and technology ventures. So if you look at these two figures, the top one shows the average amount in millions of dollars invested per deal in the European Union and in the U.S. The U.S. is the orange curve. The European Union is the blue curve.

And you can see that they more or less track each other somewhat well up until GDPR takes effect in May of this year, and things start to kind of diverge a little bit. If you look at the second graph, it looks at the total number of deals, venture deals. Think about seed rounds, series A, series B rounds, and so forth. All of those deals were technology ventures and raised money. You can see that again after GDPR, they started to diverge again.

So we could look at this difference and try to quantify it a little bit and see what the impact is on those firms and the result is quite significant, that those firms begin to raise less money. And fewer of those firms come to fruition because there are fewer funding deals. So the regulation has a noticeable impact. Now, of course, we do not know whether this is a long-term impact or whether this is just a short-term reaction. We only have several months of post-GDPR data. But it would be interesting to find out.

At least from the short-term perspective, we can see that there is a significant impact. And this impact can translate into an impact of the products we see, maybe some products do not come to fruition. Maybe those products are developed within established firms entrenching their market power. Maybe some of those products should not come to fruition. Maybe they are bad products, products that abuse our data, and this regulation is helping prevent that. We do not know the answers to that. But what we can see is that less investment has taken place. And we can translate that reduction in investment into an effect on jobs.

And we can see from our calculation that, for firms that are relatively nascent, those are new firms, they are about zero to three years old, the amount of dollars they raise per employee is somewhere between $120,000 and $1 million. Okay? And we can translate that into a very rough preliminary range on the potential number of job losses that they incur as a result of GDPR, somewhere between 3,000 and 30,000 jobs. And as a fraction -- as a percentage of the amount of employment those firms retain at least in our sample, it is substantial. It is between 4 and 11 percent.

So just some overall observations that we have also seen in the literature here, we have theoretical papers that show that identical compliance costs with data regulation tend to disproportionately burden smaller firms. This is something that we saw with the rollout of GDPR. We do not know if it is a long-term effect, but at least in the short term.

Another result shows that compliance costs can push innovation into happening inside established firms. This is also somewhat confirmed by what we see at least in the short term. And some final observations here, it seems that any regulatory approach should embrace nuance. It should be dynamic. It should be market and context-specific. If we just have a blanket approach, we are just likely to burden smaller businesses and maybe entrench market power.

Now, using data regulation, data privacy as kind of a means for data security is intuitive. It is something that makes sense. But we should strike a proper balance. We should not prevent altogether the use of personally identifiable data just because it makes it easier to have data security. Okay?

And then, finally, we should incorporate data considerations into merger review because we see, at least in our models, that they do have an effect.

Thanks very much.

(Applause.)

DR. SANDFORD: Thank you, Liad.

Our final presenter will be Florian Zettelmeyer.

DR. ZETTELMEYER: Thank you. Well, thank you very much for having me here. I appreciate the invitation very much.

I am going to talk about a topic which is quite different than what our prior speakers have done. I am going to sort of take the perspective of what it is that we, as observers, could learn about what is going on. In other words, both as academics but also inside firms. And as a result of that, the basic thesis that I am going to propose to you today is that firms are increasingly adopting machine learning in order to do advertising promotions, inventory optimization, whatever it is to basically run their business.

In many cases, these things now are colloquially interpreted as being AI, a term that you might have heard, which is, in practice, not well-distinguished from machine learning. And the point that I am trying to make is that these high-dimensionally targeting algorithms that exist out there are creating very, very strong selection effects, which make it very difficult to use traditional measurement methods in order to kind of disentangle what happened and what was going on.

And I want to give you an example of a study that I have done and then I will talk to you a little bit through where I think some of these problems are coming from. So I ended up -- for today, the study I want to refer to is the following question, which is that -- so you may be aware of this that there the most overused quote in marketing ever is a quote by a guy called John Wanamaker that says, “I know that half of my advertising is wasted. I just do not know which one, which half.”

And this was something that had a lot do with the way that firms have traditionally been able to track advertising measurement, and the way they did it is that, you know, you basically had maybe a sense of how many people you reached with an ad, so think of TV advertising 40 years ago, and you had kind of a sense of how many people bought. But you could not link at the individual level who bought and who was exposed to any kind of advertising.

So what happened over the last 15 years or so is that this link is now possible. We know in the case of Google, in the case of Facebook, in the case of many of the advertising platforms, we can typically track who ended up being intended to be targeted with an ad, who actually got targeted, did they click and then did they purchasing something as a result?

So the question that we have for us was originally motivated by an industry concern not by a regulatory concern is, does great data with observational nonexperimental methods as are common to user industry allow you to basically accurately measure advertising effects? That was the basic idea.

Now, what we did is we ended up teaming up with Facebook to answer, with a marketing science group at Facebook. And they had just introduced, when we started this project a few years ago, a product called a Facebook "Lift Test" tool, which was a tool to run randomized control trials within the Facebook platform. This turns out to actually be a very difficult thing to do.

You will hear tomorrow from another gentleman, Garrett Johnson, who can tell you how hard it was to implement this at Google as well. There were a lot of technical details about how to make experimentation work in these settings in which algorithms are essentially -- they are sort of machines to break probabilistic equivalents that you need for testing.

And in this case, we looked at 15 large-scale RCTs across a number of different industries. We chose them. They were not supposed to be representative of Facebook advertising. We chose them because they were large enough sample sizes and we had good outcomes we could measure, et cetera. We had about between 2 and 150 million users per experiment, over 1.4 billion ad impressions.

You have to understand that the Facebook data is unusually clean because of the fact that Facebook requires a single-user login which means that you do not have any problems about misidentifying people because their cookies do not match up. And we ended up measuring real outcomes. Most of them were real purchases; in some cases, registrations or website views. But it was mostly actual purchases at online retailers.

Now, you also have to understand that we were able to measure what people did even if they did not click on anything, because of the fact that we could later trace who had been exposed to an ad to that consumer's identity back at the advertiser. Of course, we had no personally identifiable information about any of these people.

So let me give you an example of this study.

So here is a study that was 25.5 million users. Think of this as like an ecommerce website where you can purchase something online. Thirty percent were in the control group; 70 percent were in the test group. The outcome of the measuring was this purchase at a digital retailer. You have what is called a conversion pixel, which the advertiser placed after the checkout page. So this study ran for 17 days, which is a pretty normal duration.

So what we then do is we measure the lift from the randomized control trial sort of to establish a ground truth. And the basic issue here is that in advertising, you cannot guarantee that anybody is exposed to an ad, so these kinds of experiments always intend to treat designs. In other words, you can say, I would like to expose you to an ad, but whether you actually see the ad depends on many things. Like are you trekking in Nepal or are you logging into Facebook today or whatever it is or maybe -- you know, maybe somebody else kind of bid for your ad impression. As a result, you did not get to see the ad.

And so in -- let’s say as an example in our situation, we had about 25 percent exposed user, 75 percent unexposed users and we had a control group that we could guarantee was unexposed. Okay?

So in this particular case, what we did is using sort of traditional average treatment effect on the treated, we observed a conversion outcome of 0.104 percent in the exposed group and then calculated a counterfactual conversion outcome in the control group of 0.059 percent. So these are users who would have been exposed if they had been in the test group.

And what this tells you is that -- and this is the traditional way that a company would express this -- there was a lift of 73 percent. So as a result, sales increased by 73 percent due to the ad. Okay. So think of this as kind of the gold standard truth running through a randomized control trial.

So, in practice, what now happens is that many advertisers do not use control groups. In fact, this is the norm. It is relatively rare to run randomized control trials. So, in our situation, what we basically had is a situation where, since our testing control groups are randomly assigned, we could replicate what you would -- the situation you would find yourself in as an advertiser if we just threw away the control group and just operated with a test group as being our group where we could see that some people were exposed versus unexposed.

In this particular case, it turns out that if you then compared the probability that somebody purchased in the exposed versus the unexposed group, the actual measurement of how well somebody did, in other words, we take people who saw an ad, we took people who did not see an ad, all of which were in the target group, in the test group, the measurement of how well the ad did went up to 316 percent. In other words, a massive overestimate of how well the ad is actually working.

Okay. And so it turns out, of course, that the fact that you get biased measurement because exposure is endogenous in this industry is well known, and as a result of that, a lot of ad measurement companies like, for example, comScore that I have listed here on this example slide from one of their decks, basically says, what we are going to do is we are going to take an ad-exposed group and then we are going to have test and control groups that are matched on demographics and behavioral variables, which gives us a balanced unexposed group, which sometimes is referred to in this industry as a forensic control group. So one that you create exposed using matching methods and things like that.

Okay. So what we did is we said, okay, we have pretty good data, because at Facebook, there is great data about what consumers do. Let us see if we could actually replicate a good balance unexposed group that would allow us to measure what is going on. So we tried. So the basic idea is that we are taking people in the exposed group and then we are taking a subset of the people in the unexposed group that by anything we observe about them should be somehow equivalent to the people in the exposed group.

Good. So in order to do this, we use the best of what exists in industry and academia, at least at the scale that we use, there are more sophisticated methods, but they do not work with 150 million users. So we used exact matching, propensity score matching, stratification, regression, inverse probability-weighed regression adjustment, stratification and regression, and we had really wonderful data because we have data on Facebook characteristics and, moreover, we even have data on -- Facebook ends up having an internal algorithm where you, as an advertiser, give Facebook a set of email addresses and then say, find me other users at Facebook that are like the users that are represented in these email addresses but are not these users.

And what we used is we literally used their algorithm to do this, which is a massive machine learning based algorithm in order to find a balanced unexposed group for the exposed group. Okay. So in other words, we threw at it what is really unusually good data in order to do this.

So let me show you the result. So what you see up here is the following. You see that the benchmark lift is 316 percent. That is what we found from the exposed-unexposed measurement. The benchmark in the RCT is 73 percent, which we take to be the truth. And what you now see here is essentially a sequence of methods that end up -- you notice there is sort of stratification and then propensity score matching and regression, et cetera, that end up becoming better and better as you add more data. So every method is essentially there were three or four variable sets.

And you notice in this case, the world looks hopeful because you can approximate pretty well with the normal observational methods. So you, as an advertiser, could do this or we, as a researcher, could do this. More or less, what happened in the industry. Well, the problem is -- and then so you do this on another method and it looks wonderful. Like, there seems to be a consistent pattern across methods and you start feeling very hopeful about the ability of recovering with the data that we normally observe what our cities do, until you hit one of the other 15 studies, and suddenly it looks like this.

The truth is 2.4 percent. And the closest estimate we have is a 1306 percent lift. So this is a study, by the way, where only 6 percent of consumers actually got exposed to the ad. And what that means is that there was a huge amount of ability for the model of essentially targeting those consumers and making them very different from the unexposed group. You also sometimes find when you get used to the idea that maybe there is always an overestimate, that sometimes these methods actually totally underestimate what is going on.

Good. And so for me -- sorry. I should have warned you about this. Red means massive overestimate. White means more or less okay. Blue means underestimate. And you see that it is all over the map depending on the studies. And so it is very difficult for us ahead of time telling you what is going to happen without knowing more about these particular studies.

Okay. So the basic idea is this, which is that we are in a situation and it is because of the fact that firms are using machine-learning models, where the targeting of consumers is becoming so basically deterministic that a lot of the observational methods that we use, which rely on the idea that there remains random variation after you condition out what we observe of people, start breaking down. And this is quite important because it means that this lack of transparency that Alessandro was talking about earlier is all over the place.

So you have an industry where, for example, many people who spend a ton of money on marketing at the moment simply do not know how well these kinds of interventions are working, because unless you plan ahead big-time and spend lots of money on doing randomized control trials, you literally have no sense of being able to tell whether your expenditures are actually working or not. And this is important both -- so it is this really interesting thing where despite amazing data -- and these algorithms make it very difficult to actually get accurate feedback on what is going on in industry.

And this is not well understood in industry and it creates sort of a level of grayness that I think a lot of people do not expect in this particular industry. Thank you very much.

(Applause.)

DR. SANDFORD: Okay. So once again, if you are in the room here at American University and would like to ask a question, we have people walking up and down the aisles with note cards. So please flag one of them down.

Okay. So my first question is -- I am looking back at Joe Stiglitz's remarks from a prior hearing and he opined that big data provides new tools for price discrimination and those with ability to discriminate better grow. And so the firms that get big and become successful are those with lots of data that can do price discrimination and not necessarily those with the best product.

And Liad's presentation talked about the effect of privacy in a price-discrimination context. I read a survey paper by Alessandro, Liad, and Curtis Taylor and many of the papers there talked about price discrimination, again, as sort of the vector for how privacy affects consumer outcomes. And the question I have here, you know, 20 years ago, I would have said it was obvious that we were headed for an era with individualized pricing. I would go on Amazon and I would get a price that only applied to me. Indeed, I wrote a paper for my intermediate microeconomics class saying as much. The paper received a B for good reason. It was completely wrong. That did not happen. It has not come to pass.

I have a quote from computer scientist, Arvind Narayanan. He wrote “The mystery about online price discrimination is why so little of it seems to be happening.” And so from my perspective, the price discrimination I do see online is the same thing that retailers were doing 100 years ago. It is coupons and sales and starting the price high and then lowering it over time.

So my question is, why don’t we see more price discrimination? And if you agree with my premise that we do not see a lot of price discrimination, should that cause us to update our priors of how we think about privacy given all of this work on the effect of privacy on welfare through price discrimination? So I will just throw that out to the panel, anyone who wants to answer it.

DR. WAGMAN: I would just like to say --

DR. SANDFORD: You have to turn your mic on, by the way.

DR. WAGMAN: I would just like to cite a couple of examples that we are starting to see individualized pricing, at least in the context of the ridesharing apps where the price you see is very, very likely tailored to your record, your history of using the app, for example, on Uber or Lyft. And those efforts are, in my estimation, only intensifying over time.

The other piece of what you mention of offering coupons and discounts, I think that can also be a lot more targeted than it used to be. And so we may not see price discrimination upward from, say, a certain base price or a perceived base price, but we will certainly see it downward with only certain selected individuals being targeted with offers.

DR. SANDFORD: Ginger?

MS. JIN: I just want to add that, from the economic point of view, the word “discrimination” is probably not as loaded as it sounds in plain English. According to this theory, price discrimination is not necessarily welfare-reducing whether that is defined for consumer welfare or total welfare, because when you are comparing with uniform pricing, when you have price discrimination, some people may get a discount from that and some people may have a price higher than the uniform pricing. So the welfare consequence of that is going to be a mixture depending on how many people are getting a discount and how many people are getting a lift.

And in terms of underuse of data in price discrimination, I think there is still some probably preference -- consumer preference about sort of to what extent the firms are using price discrimination. I think that is probably a separate dimension as compared to sort of their willingness to pay for a particular product. And if a firm has a sense that consumers dislike this kind of personalized price discrimination, even if they make a short-term discount on this particular product, I think a firm will take that into account.

I mean, this is just hypothesis. I wonder to what extent that kind of general resistance to personalized price discrimination actually get into firms' sort of choice of how much price discrimination they would use.

DR. SANDFORD: Florian and Alessandro?

DR. ACQUISTI: Thank you. Echoing something Liad was saying and connecting it to what Ginger was saying, I believe that part of the puzzle is that what may be happening is something I call product discrimination. And as Ginger pointed out, I am not using the term "discrimination" with a negative connotation but rather in the economic connotation. By product discrimination, I am referring to the ability of the industry, the advertising industry, to send an ad for a certain product rather than another. In doing so, they may match a consumer to what is maybe a higher-quality or a lower-quality product, higher price, lower price.

So we may not see the very same product being sold at different prices to different consumers. So we may not see first-degree price discrimination, which is most of what the empirical efforts have been trying to do. But we may see basically forms of self-selection, second-degree price discrimination.

By the way, one very small pushback, I would contest the notion that much of the negative welfare consequences of privacy for consumers are related to price discrimination. That is one part of the story, but there are others.

DR. ZETTELMEYER: You know, I think another aspect of price discrimination is that we have -- the question of exactly what is price discrimination, what is intertemporal pricing is actually not very well defined. A nice example of this -- and I will tie this back to online markets in a minute. A nice example of this is I have done a study on pricing at car dealerships, and it turns out you can actually explain a lot of the -- what looks like price discrimination, namely different consumers are paying different amounts of money for the car, simply by the levels of inventory that happen to exist when the consumer is walking into the dealership.

So it looks like the dealer is discriminating against individual consumers, but it is really reflecting the scarcity rents of the inventory that happens to be lying around. So if you have two red Honda Accords on the lot, you are going to price it differently than if you happen to have 53 on the lot. And depending on when you walk in as a consumer, you are going to see different prices.

To us it looks very similar, as if it is price discrimination, but it actually has a very different economic reason for it. So I think similarly in the online context, you do observe a lot more intertemporal price variation and we can think of that as also being at least, you know, fulfilling a similar goal as individual level price discrimination, first-degree price discrimination literally at the same time for the same kinds of consumer.

So I guess there is maybe more price discrimination than meets the eye, which I think was Liad’s point as well.

DR. BEN-SHAHAR: I guess I will add my

perspective. I think that the puzzle is compounded by the fact that we do not see personalizing of other aspects of the product not just the price. Why should everybody get the same right to return the products, the same warranty, the same privacy terms? If we know enough about people, how much they can pay, we probably know a little bit, also, or a lot about what their preferences are.

We do see that, you know, people are sometimes thrown out of Amazon Prime if they are return-aholics or things like that. So it is either zero or one, but we do not use a dimmer and that is kind of puzzling to the same extent that the -- now, I guess one of the problems that jumps to mind, and I have not studied this closely with the data, but is the problem of arbitrage. As long as you are selling products and not services, people can resell them.

I think that once things are done through platforms, apps, and are sold as utilities and services, we might be able to -- we might see the burst of a personalization of various aspects of products.

DR. SANDFORD: Okay, thank you.

Next question, so Omri mentioned the privacy paradox which is, as I understand it, is consumers say overwhelmingly that they prefer greater privacy, yet they do not act in a way consistent with that. So, for example, I think I pulled from Alessandro and Liad's paper, 86 percent of your adults say they do not want targeted advertisements; 93 percent of all adults believe in "being in control of who can get information about them is important." And, yet, it is not clear that consumers behave in a way consistent with the preferences expressed in surveys that ask you, yes or no, do you prefer greater privacy.

So, I mean, my reaction to this is -- well, one, is this actually a paradox? I mean, is this just we are suggesting something that sounds vaguely positive to people and saying, are you in favor of it or not and they say, yeah, sure, I am in favor of animal rights but I like to eat steak. I mean, something like that.

And two, I mean, kind of -- is there in a -- you know, we look at firms in the market. They have different privacy policies. Is there a sense in which consumers have different preferences over these different privacy policies and might go to one firm or another based on their privacy policies? So do consumers have a downward sloping demand for privacy that is -- you know, has meaningful slope across the range of privacy policies we see in the marketplace? Whoever wants to go first.

DR. ACQUISTI: I may start. I feel that there is quite substantial evidence that there is a demand for privacy by consumers and this demand follows, to some extent, canonical, traditional, expectable economic lows. People will exercise their demand for privacy when the price of doing so is small. People close their bathroom door when they are going to the bathroom. People do not post their credit card online because it would be insecure and it would be also probably costly, just the act of doing so.

As you get into more esoteric and costly behavior, consumers engage into that when there is an actual benefit for doing so. So wealthy individuals go to quite extreme measures perhaps sometimes to hide their wealth and use bank accounts which may not be monitored by enforcement agencies, for instance. And they try to have anonymity and they may pay for that because it is very valuable to them. So there is actually a demand for privacy which follows canonical economics lows, but there are also these issues of not always being able to predict what the cost of privacy will be especially online for -- due to the fact that privacy tradeoffs are intertemporal in nature.

So you may reveal information now which may not affect you for a long time, but eventually will affect you. And this, to me, one of the possible explanations, not the only one, for the privacy paradox.

What is very interesting to me and Omri made me think about that through his remarks is that there is another form of paradox which is much less explored but as compelling. The paradox of people who claim that privacy is not important to them, but, in fact, act as it is. And that is really many of us. Even though the people who claim that privacy is not important engage in behaviors every day, both online and offline, which are privacy-seeking behaviors, lowering the tone of the conversation in the restaurant when they are having dinner with their partner when the waiter arrives. That is a form of privacy-seeking behavior in public where you are trying to make your conversation private.

The example I was making earlier of closing the bathroom door when you go to the bathroom; the other example I was making earlier of not sharing your credit card information online. Now, if you ask people about these behaviors, some would probably, in a manner, suggest that it is not about privacy. For instance, it is -- not sharing the credit information online is about security. Closing the bathroom door is about social norms or politeness, not privacy. To me, this suggests that people have very personal definitions about what privacy is, and it is not an intent to disregard other people's definition of privacy in favor of their own. But, in fact, at the end of the day, they are all about the same thing, which is the individual's ability to modulate the degree of public and the private in their lives.

DR. SANDFORD: Ginger?

MS. JIN: Yes. I just want to echo Alessandro that there is a definition problem here. If we think privacy protection or data policy is one product attribute for the product and service I am buying, it is unclear exactly what is that product attribute I am buying. Okay? So you can think of, say, 100 percent protection on one end and zero protection on the other end. I am actually not sure exactly where I am buying in that spectrum because the firm may protect my data very well or run with it. Right? So we do not know exactly. And that fuzziness probably could be one of the explanations for this.

Another related issue I want to echo was

Omri’s point about data pollution. I think from consumers’ point of view, if you view data policy as one product attribute but you just do not have time to track exactly where that product attribute is for every firm, every product you are having, you have this overall impression. Okay? And then when you heard about Equifax or Cambridge Analytica or something, you sort of formed this kind of prior or posterior about exactly where this product attribute is. And that is evolving.

And it could be this firm actually doing a very protective thing about my PII data, but because I heard so many other things that I got sort of afraid. I am afraid you are going to run with my data for some abusive use. So in that sense, you probably get to the second paradox that Alessandro was just talking about, which is how can I convince you that I am actually selling you a product with a very good data policy? It will be very hard to convince given that your prior is sort of polluted by many other firms.

DR. SANDFORD: Florian?

DR. ZETTELMEYER: Yeah, I think to tie some of these things together is simply the link between data and what is done with it is so opaque today, and I think that is what is leading to a lot of the problems. So exactly the same data could be used for ways that absolutely delight you and then for ways that you would find absolutely horrendous. And so I sometimes wonder whether we spend too much time thinking about how to protect the data as opposed to protecting the use of the data. And I think, you know, in some sense, it is the entire promise of this big data enterprise. And if you think about the current advances in machine learning, it is that data can be used in ways that should blow all our minds in order to form predictions that we never thought could reasonably form with data like that.

And as a result, somehow being able to expect that people can have reasonable agency with regards to the protection -- what data they make available in the complete lack of a link between what happens with their data and -- between them giving their data out and what happens with their data is incredibly difficult to accomplish. It is like asking somebody to regulate the electricity usage at home if they have absolutely no idea what the usage of any device is and they cannot measure the outcome of it. How do we expect people to be somehow reactive to how much energy they are using? It is a very similar situation in this realm as well.

DR. SANDFORD: Liad?

DR. WAGMAN: I think there is also a sense of no matter what I do, it is going to be collected. Just to give an anecdotal recent example, GDPR rolled out and a large firm with millions of users put the consent popup on their page. So when users would surf to the page, they would see the consent. And they would have two options. They could say, yes, I am willing to share everything, or, no, I want to choose what I share. 96 percent of users clicked on yes, I will share everything. And 4 percent clicked on, no, I will choose what I share.

And then they clicked on that and they very carefully chose -- they had the option to choose to share nothing. But they very carefully chose to share some and not others. And interestingly enough, based on their choices, they could be easily identified and targeted with ads, because their choices were highly correlated with other information about them. And so there is this sense of inevitability, no matter what I do, it will be collected and I will be identified at least in some sense.

DR. ZETTELMEYER: Or worse, actually, machine-learning algorithms are going to figure out what my preferences are even if I do not state them.

DR. WAGMAN: Right.

DR. BEN-SHAHAR: I would like to touch on two things that the panelists said. I would like to challenge Alessandro's response. He said, you know, people close the bathroom doors. You see there is privacy. You know, but they do not mind the electronic eye that flushes the toilet. Right? Even if there was...

(Laughter.)

DR. BEN-SHAHAR: I mean, that is, I think, the difference between the privacy -- the secrets that we have in the presence of other people and the data privacy, vis-a-vis, the algorithms that are collected. You know, even if the electronic eye was connected to some algorithm and sold me some constipation medication, you know, I think people initially might be alarmed. But, ultimately, I think it would not be out of a -- it would not change their behavior to use these bathrooms pretty comfortably.

So I think that you probably have a lot of evidence that people care about data privacy. I would not use the example of closing bathroom doors to make that -- that seems a little bit like kind of a strawman.

I really like the point that Liad made that, you know, look, four people -- only 4 percent of the people exercised what a lot of privacy advocates and privacy regulators want them to, which is user control. I actually think that 4 percent way, way, way overestimates the prevalence of this phenomenon once the novelty will die out and we will realize that you have to do this not to that one website in that experiment or whatever, but to do it to dozens of places daily and that you really do not know what are the right choices because you do not know what the tradeoffs are. You do not know. It is so complicated.

User control in every aspect -- I have studied that not in the privacy context but in consumer credit, probably a much more fateful decision people make -- user control is kind of a panacea. People cannot make good decisions no matter how well-intentioned regulators are to give them all the aids, decision aids and choice architecture if they do not understand the tradeoffs and they do not have the sophistication to deal with problems that, at the core, are not simple.

DR. SANDFORD: Alessandro, you wanted to make a brief point?

DR. ACQUISTI: Very brief comment. I

actually do not disagree with you, but the contrast between online and offline was intentional. It was to point out that there are situations where individuals take action to protect their privacy, especially when it comes to physical privacy, and there are situations where they may not, especially when it comes to online privacy.

To me, from this to conclude that that implies that people do not care about online privacy, that is, to me, the conclusion that is erroneous, because there are many, many factors which differentiate the offline scenario, the bathroom door, and the online scenario, including intertemporal tradeoffs. You are seen immediately by someone else in the bathroom. If you post something, you may not be seen by someone who with an interest to use your data one year later, five years later. They show information asymmetry.

The issue that Liad was referring to of efficacy, if I close the door, I have control. If I post something on Facebook, even if I use correctly the privacy settings and visibility settings, I still do not have really much control on what happens to that photo after I uploaded it. So it is intentional for me to contrast the online and offline. As a matter of fact, we do have a paper that is about to be submitted about this in particular.

(Laughter.)

DR. SANDFORD: Okay, thank you. So it sounds like obfuscation. It is not clear to me what the privacy policy is and frustration with that is a driver of why consumers do not seem to care about privacy.

I wanted to ask Omri a question before he has to leave. Omri, you wrote a book with Carl Schneider espousing your view that privacy policies are essentially worthless. No one reads them. You said that, in 2008, it would take 76 workdays to read all of the terms of use and privacy policies that one would come across in the course of normal use of the internet, and that was ten years ago. It could be more than 365 workdays now for all we know.

Omri had a picture in the book where Omri is like two inches tall in the photo and the iTunes terms of service are like a foot tall in the photo. I mean, they come down from the second floor and dwarf him. So his point is it is effectively impossible to read everything that you are agreeing to when you use various websites, and so I do not want to put -- my characterization on Omri’s is that these are essentially useless. They provide no bite. They are not helpful to consumers in deciding which websites I should patronize and which I should not.

So I guess my question, Omri -- you can respond to that however you want -- but my question is how many people need to read these for them to be effective? So for example, if a government plaintiff reads a privacy policy and says, hey, you are not behaving in that way, is that meaningful to what kind of privacy policies get promulgated in the marketplace? If a journalist reads one of these policies and says, hey, there is something kind of funny in this policy, would that scare users away and be a check on what goes into the privacy policy? So what do you make of that view?

MR. BEL-SHAHAR: Thank you for raising this. I think the good people at Carnegie Mellon read the privacy policies and grade them for us. I do not think many people go to PrivacyGrades.org. I know occasionally a newspaper, *The New York Times*, calls to ask me questions about the terrible things that Facebook does, and I say, look, your app gets a lower grade then Facebook. But, of course, *The New York Times* is not the problem, maybe Facebook is. And so what are these grades really telling us?

I guess my view about giving people information so that they will make wiser, more prudent choices, is failing everywhere. It is not a privacy problem; it is a disclosure problem. It is a problem with the regulatory technique. It fails miserably and for a long time in consumer credit where it all was invented, truth-in-lending and things like that. It fails all over contract law, because anytime you click “I agree,” people put you through these meaningless rituals of clicking these things, closing boxes because contract law requires consent for all sorts of things that otherwise would be a violation of law, including the privacy terms.

But also all the disclaimers and all the -- yada, yada -- all that stuff, all the consent forms in hospitals that people get, 17 pages of consent forms to participate in human subject research, the evidence is -- the mountains of evidence -- undisputed that nobody reads it. That the people to whom it is given cannot understand it if they were able to read it and the issues, as I mentioned, before are too complex. So I guess in the privacy context, what many -- in many places, the solution that is proposed and in other contexts, too, is to simplify.

Simplification is, I call it in my book that you mentioned, is the *deus ex machina*. It falls from the ceiling and it kind of solves the plot and everything is good afterwards. But it does not. Simplification, in every area that I mentioned, has been tried for decades and failed, again, for the reason -- and now I am saying it for the third time -- that you cannot really simplify the complex. When things are complicated, you cannot just give people red light/green light.

And so I do not know -- I cannot conceptualize in my mind, in response to your question, who will actually read and give consumers the information that will be operational? Ultimately, consumers, if they want to make more prudent choices, should rely on the experience of people like them. So ratings sometimes help them and, in many contexts, they do. They could also be misleading. And it is very important to protect, as a regulator, the integrity of these aids that do not give people information, but give them a good prediction of how content they will be if they actually jump into the experience of this product or service.

DR. SANDFORD: Ginger?

MR. JIN: Just to add on Omri's point, suppose we have a sophisticated machine that government or journalists can use to really squeeze out all the information from those pages and tell a very simplified, but fully informative, story to consumers, I think it does not solve the following problem, which is how can I be sure what you say is exactly what you do and given that what you do is evolving over time with new technology and so forth.

So I think that the second part of this problem is really crucial. Otherwise, you can say anything. Right? So how can we sort of check what you said and then make sure that is consistent with the policy given the amount of data policy and the kind of firms that could use data? I think it will be unfeasible for everyone to be checked in a precise and timely way, and I think that is probably one of the inherent problems in this approach.

DR. SANDFORD: But there is still a deterrent effect if there is a data breach that is very high profile and that might get punishment from the Government or something like that, or -- so enforcement is sporadic, but perhaps severe when it does come. That can still be a check on behavior, could it not, on what goes in privacy policies?

MS. JIN: Well, when we talk about data breach, it is a symptom, right? I mean, the agency, like a doctor trying to come up with a diagnosis. Unfortunately, the link between the symptom and the diagnosis is not that straightforward. If a firm got data-breached, it could be the firm's fault not having enough security, so that it sort of left room for the hackers to come in. Or it could be somehow the hackers have the most cutting-edge technology that will be able to penetrate even the most secure walls.

I mean, you have to tell those two in order to say exactly is that a problem, the hacker's problem, or is that a problem of the firm's problem?

DR. SANDFORD: Okay, thank you. Let's talk now about supply for privacy. How do firms decide what goes into their privacy policy and, in particular, is there a sense in which firms are responding to consumer preferences over privacy? We have talked about how strong those preferences are, whether they are reflected in consumer decisions or not.

Do we see any evidence that firms are just going for the maximalist privacy policy? I am going to just write down everything I want and get you to agree to it, or is there some sense in which firms are responding to consumer preferences, maybe perhaps worried that if I have a maximalist privacy policy, users might shy away from my website and go somewhere else for example?

So is there a sense in which there is a supply curve for privacy that reflects firms either giving a greater level of service in return for more privacy or responding to consumer preferences for privacy?

DR. WAGMAN: I think that is a tough one, because those privacy terms are ever-changing. Right? And if a firm realizes there is way to commercialize, monetize, do something else with data, they will change their terms so they can collect that data as well. They might give some disclosure that, again, nobody will read that they changed their terms. And so I think it adds to that sense of, no matter what I do, I cannot really prevent it being collected. And even if right now the terms are friendly to me and even if the firm actually follows through on those terms, that can change at any time.

DR. ZETTELMEYER: Also, I am not sure to which degree a lot of consumers understand the difference between we have lots of your data and we will keep it safe and we do not collect it in the first place. Right? And so there are very few firms that are using that from a branding point of view at the moment. Apple is a very high-profile one. I mean, as far as I can tell, I am not sure there is any evidence that consumers necessarily care about that.

DR. WAGMAN: I would also add that even those characterizations are sometimes misleading. So even if a firm like Apple says, oh, we do not collect it, they might have partnerships with other firms who do collect it and they benefit from it indirectly.

MS. JIN: I think one anecdote probably does suggest that people care at least about the perception of privacy. I think that example was some years ago Samsung had a TV and the TV has kind of a camera that you can -- or voice recognition that you can sort of give a voice command to the TV. And then there was a kind of public outcry against the possibility that maybe the microphone is always listening, even to the private talks in your living room. And I think in response to that public outcry, Samsung did change their privacy policy. And again, does that exactly reflected what they do in the future is still an open question.

DR. ZETTELMEYER: But, of course, now we have moved on to conversational interfaces like Alexa that listen to everything you do and consumers seem to be fine with it.

DR. ACQUISTI: Your point about a supply curve for privacy is extremely interesting. It makes me think about the -- another question that I find under-explored in the research in this field, which is the relationship between data collection usage of data and the provision of free services and free content, specifically to what extent increasing data collection is necessary for the provision of more and better services.

I know I am maybe about to say something that sounds bold, but once again, I believe that I have some empirical evidence to support the claim. And the claim is that the relationship between data collection and provision of services is more correlational than causal or at least we do not have very strong evidence of it being causal as opposed to correlational.

What I mean is that the provision of free services existed on the internet way back in the days before the more granular techniques of collecting information about users and tracking them across different sites started, which is about 2004 or 2005 with Facebook, et cetera. Even nowadays, there are firms which can do well without data collection. DuckDuckGo is an example.

To me, this brings another question. Once again, I really do not know what the answer is, but the bold claim I am making is that I do not feel many people actually know what the answer is, to what extent the relationship between data collection and provision of free services is correlational, to what extent it is causal.

It goes back to the value allocation question. To what extent when merchants may be paying 500 percent for targeted ads and publishers get 4 percent more for targeted ads. To what extent something gets lost in the middle remains in the realm of the data oligopolies. And this could potentially provide an answer then to the question of causal versus a correlational relationship between provision of free services and data collection.

DR. SANDFORD: Florian?

DR. ZETTELMEYER: Can I ask you a question I thought of, which is related to this issue? I think I agree with you. I wonder, however, whether the one exception to that is the current rise of AI and machine learning in the sense that, if we think, you know, roughly speaking as those being kind of prediction machines that have large effects on the quality of provision of services --

DR. ACQUISTI: And would not be able --

DR. ZETTELMEYER: -- and those cannot work without data.

DR. ACQUISTI: -- exist without data.

DR. ZETTELMEYER: Exactly. So I think that may be the one exception to that. And I am not quite sure how to think through it, but I wonder what would you think.

DR. ACQUISTI: I think you make a good point. And it goes back then to an item I mentioned at the very start of my talk, to what extent for that kind of analysis we always need identified data versus anonymized data, but to a degree of granularity, which is sufficient for the kind of analysis. It goes back to privacy not being monotonic, not being absence or presence of data, but being a modulation of what type of data you use and analyze.

DR. SANDFORD: Omri? Okay, thank you, Omri.

MR. BEL-SHAHAR: Sorry.

DR. SANDFORD: Okay. So the next question I have is, is there a sense in which firms compete in privacy policies or the answer may be no based on the answers -- what we were just discussing. But, I mean, is there a sense in which, you know, say Facebook has a bunch of locked-in users that they can have a more maximalist privacy policy than, like, Walmart that has to go out and compete for every retailer dollar with other online sites? And so is there a sense in which competition matters for privacy? And is there a sense in which, say, removing a competitor, like with a merger, could matter for privacy outcomes?

And your answer can be no, in which case we do not need a long -- it need not be long.

DR. WAGMAN: Sure. I think a couple of examples that were already mentioned of Apple and DuckDuckGo as firms that are trying to market privacy as a feature have been raised. Obviously, there are very few. But those are significant examples.

In terms of mergers and privacy, I mentioned earlier that data does make merger review slightly more favorable because firms are competing on more fronts. So provided there are at least two firms remaining in the market after a merger and data is a component on which they can use to compete with, competition could still be intense because of all the segmentation that can be done and competition over those segments.

DR. SANDFORD: Okay. Does anyone else want to opine yes or no, do firms compete in privacy?

(No response.)

DR. SANDFORD: Okay. So the other potential antitrust issue I might think of with privacy is -- or privacy and data are, do data serve as a barrier to entry? And is that barrier to entry somehow different than just like I own a factory and you do not, so you have a barrier to entering my industry.

So I have a quote here to Darren Tucker and Hill Wellford that states that data are ubiquitous, low-cost and widely available and that an entrant that needs personal data can collect relevant information from its users once a service is operational. Data collected in this manner is free or nearly so.

So the argument is sometimes made that, hey, these firms, like these big tech platforms, have lots and lots of data and that makes it harder to compete with them, that might affect competition in some way. A possible counter to that is you can just go out and buy data, you know. There are lots of places you can go buy data. Firms do buy data on where people live, what their income is, how many people are in their household, maybe some information on what their preferences are. And so is there any sense in which data could be a barrier to entry, in which data that I have, but you do not, is irreproducible and gives me an advantage that you do not have?

Florian?

DR. ZETTELMEYER: So I really disagree with that view. I think that data, in particular back to this discussion of predictions and machine learning and AI, is extremely important. I think what most people do not realize is that the amount of examples that go into being able to train these algorithms is absolutely astronomical. In particular, because in many domains, whether the algorithms get widespread use is very much a function of whether they manage to do predictions in extraordinary ways.

In other words, you know, getting an algorithm for predicting correctly 80 to 90 percent of the time may not be a big deal. But if you are at 98 percent and you get it to predict correctly 99.9 percent, suddenly you have something that is completely usable and creates an enormous change in the way that you can then think of firm strategy of what you compete on, all the services that you produce, it could change the business model that you use.

You know, there is this wonderful example that a book that -- a very nice book that recently came out from Avi Goldfarb, who is going to be here on the panel, and Josh Gans’ book on what he called --

they called “prediction machines,” which I recommend everybody to read. In there, they have this very nice hypothetical example of where they talk about the fact that, at the moment, Amazon has, like, a first shop and then ship model. If you could predict to great accuracy what people are going to buy, you could ship first and then shop. That has an enormous effect on strategy, on how you would operate as a company.

So I think that those advances are only possible with absolutely huge amounts of data. So I think it is true that more and more data, at some point, has sort of slightly fewer returns, but what you can accomplish with the predictions that arise from that data could potentially be a sea change. So the returns to that additional data is huge. And as a result of this, I think that data is very, very important and it is certainly not ubiquitous in this sense.

And we have seen this, by the way, in the search engine wars from a number of years ago, how hard it was for people like Bing to catch up or compete adequately with Google, simply based on the volume of data that they had in order to improve their searches.

DR. SANDFORD: Ginger and Alessandro both wanted to weigh in.

MS. JIN: Just to play devil's advocate here, we have seen entrants disruptively take over the incumbent although the entrant does not have a data advantage. So we think about Google against Yahoo or Facebook against MySpace. But Florian could be right that maybe, at that moment, that data was not used very efficiently or the data scale had not been large enough and granular enough to have sort of the effect that we observe today.

But let's just say, okay, that data is very important today. It is a very valuable asset. It does give an advantage for the incumbent to use that data in a way that has a competitive edge. Okay? Let's say that is true. I think we still need to think hard of how to translate that into, say, antitrust action.

Because you can say, okay, in the oil refinery industry you need a lot of investment to start and that means we need to break up the oil companies. I think there is a leap of logic there when you say sort of the barrier to entry is very high. Whether it is in physical assets or in data assets, there is a question we have to ask about the investment that firms are putting into these kind of algorithms or data collections, and they cost money, they cost efforts, they cost talents. And to what extent that is -- we should think that all that should be available to everybody and how would that undermine the investment incentive for the firms to really improve the algorithms and improve the data collection, I think that is a hard question.

DR. ZETTELMEYER: It is a very hard question. I think it is also very context-specific. I mean, I do not think this applies to every single context, but I think there are contexts in which, you know, going from huge to extra huge does make a difference. I think it is hard to preview at this time, frankly, when that is the case and when it is not.

DR. ACQUISTI: I am not making an antitrust argument because that really is not my field of research or expertise. But it is interesting, I was going in the same direction Ginger was going thinking about examples such as MySpace or Oracle or Yahoo who, notwithstanding having, to some extent, first mover advantage were then replaced by companies like Facebook and Google, and I was thinking what are the differences? To me, there are many. There are many, okay? So it would be simple if it were just one. But an important one is that both Google and Facebook succeeded in creating these two-sided platforms and benefit from network effects on both sides of the platform.

If you are an advertiser, you want to be on the platform that offers you great access to publishers. If you are a publisher, you want great access to advertisers. These dynamics are to be self-reinforcing and they create these very, very strong concentration of power in firms, such as Google and Facebook, which may create this potential issue of antitrust, although I am not getting into the issue of then whether it should be split up or so because that is not my area of expertise.

DR. WAGMAN: I would also add to that that there are examples of firms scooping up other firms that seemingly have different data, for example, Facebook acquiring WhatsApp for 20-some billion dollars. The data seems different. It seems like a different kind of network. And, yet, the data is extremely valuable. It contains, you know, context lists that can be connected with information Facebook already has about users to better pinpoint users, to better identify them.

So this adding up of seemingly disparate graphs or networks or data sets can be extremely beneficial and kind of bring you to that huge point where you can identify people with pinpoint accuracy.

DR. ZETTELMEYER: I should also point out that inside the industry, there is actually a concern about this. I mean, there is this open AI initiative that Elon Musk is involved in, which is precisely about trying to make sure that a lot of the advances in that area are in the public domain somehow in order to be able to be shared across everybody because of the fact that there is a concern that you might get too much of an advantage otherwise.

DR. ACQUISTI: And to Liad's point about WhatsApp was really great and interesting because it also connects, in a way, to a question that I feel bad we did not fully address, the question about competition. We did not have much to say. But the example of WhatsApp and Instagram is quite interesting from a competitive perspective.

Some users started using Facebook less or even migrating away from Facebook to other platforms, such as Instagram also partly, not only, for privacy reasons. And, yet, a powerful company can use the revenues to acquire its competitors -- its more privacy-friendly competitors and reincorporate the data of these users back into their databases. This is an interesting tale about the challenges of privacy-based competition in this market.

DR. SANDFORD: Okay. I want to read a quote. So I want to read from a blog post by the CEO of DuckDuckGo, Gabriel Weinberg. The quote, "It is actually a big myth that search engines need to track your personal search history to make money or deliver quality search results. Almost all of the money search engines make, including Google, is based on the keywords you type in without knowing anything about you, including your search history. The fact is these companies would still be wildly profitable if, for example, they dropped all of these hidden trackers across the web and limited the amount of data they keep only to what is most necessary."

Okay, this is -- I'm guessing Florian is going to say that is not true based on the data he studied. But this sort of raises the question, is he right? I mean, could we drastically scale back the data, say, Google is collecting from us, just sell ads based on keywords and make a little bit less money, but maybe not that much less and maybe we would be better off by having more privacy?

As to the question of the value of targeting ads, I mean, Liad had a -- sorry, Alessandro, in his opening remarks, said that the value of a targeted ad raised revenue by .0008 dollars if I have that right, or maybe there is an extra zero in there.

DR. ACQUISTI: There are four zeros.

DR. SANDFORD: Okay, one extra zero. So, I mean, there is a question of targeted ads raise more revenue, but how much more? And it sort of seemed like Alessandro is saying, by not very much at all, but Florian's Facebook paper is suggesting that maybe the value is quite substantial. So how should I think about the value of ad targeting? Is it big or is it small and what do we think of the DuckDuckGo guy, who obviously is not an unbiased observer? What do we think of his remarks?

DR. ACQUISTI: I am actually curious about what Florian would say about this. I will only comment that the results I was reporting and those found by Florian, they are not contradictory. In fact, they may be very much on the same page. We are looking at what -- at the end of the value chain remains in the hands of publishers. And Florian was, if I understood correctly, looking at how merchants who use certain techniques for advertising can see confluent conversions expand in the presence of targeting.

DR. ZETTELMEYER: So I do not know what -- I think my first approach would be to say that Google is in two ad businesses, one is the keyword search ads -- keyword-based search ads, and the other are display ads and the display ad networks that they run. So those are different from each other. I believe that while it is true that you may only need keywords in order to place search ads, you certainly need information about users in order to participate in the ad networks and display advertising.

So I think maybe that is a little bit lost in that quote. So I do not have, off the top of my head, what percentage of revenue profits in Google depends on one versus the other type of advertising. So I cannot say whether that is correct that, you know, they would still make loads of money if you shut one of the things down or not.

By my sense is that in order to do the targeting, you do do this, and I think the big problem is there would be -- I am just a little concerned to the degree that -- you know, Alessandro, I do not know how generalizable this result is about the benefits of targeting. It is just very difficult to get good measurements in this space, I think even for those who are involved in it. I think a lot of times the firms themselves that target do not know how valuable the targeting is.

That would be certainly a wonderful area for more research because I do not think we have a really great fact base, frankly, to answer -- to question the gentleman or to kind of challenge the statement the gentleman is posing at the moment.

DR. ACQUISTI: I agree.

DR. SANDFORD: Ginger?

MS. JIN: Yeah, I wonder if the observation you quoted will be related to Florian's earlier comment about this huge versus extra huge. I mean, maybe today, we do not see the extra huge effect yet, but who knows. In the future, there will be technology that can much better use the individual identifiable information from Google versus DuckDuckGo and have a huge lift. I mean, we just do not know.

DR. WAGMAN: I would say that from the perspective of economic theory, there is obviously value in knowing more about a consumer. So I could see a consumer, you know, searching for a computer and I know they are predisposed to maybe buying a computer, and then I could maybe know who the consumer is, how much income they have, how much education they have, where they live, whether they have a computer right now or not, and I could use that information to send them to a very different place. Just like firms might steer Mac users to a different list of hotels than PC users.

DR. SANDFORD: Okay. I have a couple questions from the audience I will get to. This one I will direct to florian. Florian, if businesses have no good means to evaluate the impact of their targeted ads, why are they spending so much on such ads?

DR. ZETTELMEYER: That is a wonderful question. I think that there is, in my experience, enormous amounts of information asymmetry as to -- I think a lot of firms or the people in charge of placing ads in many of these firms are not well aware of this problem.

The measurement problem with digital advertising is very pervasive, it is very big. There are a bunch of people who, in academia, have done some amazing work on this, like David Reilly and Garrett Johnson, who is coming tomorrow, and Randall Lewis, et cetera. And you now have an increasing set of people who are very, very sophisticated about thinking about advertising placements and marketing place in general, but the basic problem that exists is that, you know, marketing is a special form of hell when it comes to measurement because of the fact that so much of consumer behavior is highly endogenous and so much of the way that firms target is so endogenous. So measurement, in general, is a very difficult thing.

We used to have an area in marketing that was very well measured, which was the direct mail industry. But somehow the people who went into the digital world are not the old mail order guys. Often, they came out of the advertising industry, which did not have as strong a tradition of very good measurement. So there is just a lot of lack of information.

I would maintain that part of the problem is that there is a little bit of political economy here, as well, which is that beyond the situation where it is not always clear to me that everybody wants to actually know the answer to how well the advertising is actually working. And I will just leave it at that.

DR. SANDFORD: Alessandro?

DR. ACQUISTI: Adding a comment to what Florian so eloquently put out and said. Large companies have troubles in understanding the value of targeted advertising for them. Famously, they see, oh, Unilever made some controversial statements about the benefit of social media advertising to them. And these are large companies with very sophisticated research teams. Think about the challenges for medium and even more so small companies that may not have the know-how and skill set available to run the kind of experiments that Florian has been able to run and the larger companies are running to understand the value that they get from that.

It goes back to the point that we have been discussing. It is kind of like a red line connecting our different comments of this opacity in the very proposition of certain aspects of targeted advertising.

DR. ZETTELMEYER: If I could say one more thing about this, Jeremy --

DR. SANDFORD: Mm-hmm.

DR. ZETTELMEYER: -- which is that I think what is tricky is that a lot of the advances that have been done with analytics and quantitative methods and machine learning, et cetera, they are advances of prediction. The problem is that -- and predictions work incredibly well in many domains. The big problem, however, is that nearly all marketing expenditure is not a traditional prediction problem because it is a problem of causal inference essentially. In other words, what you want to know is what would have happened had I not placed an ad.

And this often does not lend itself very well to sort of organically arising data sets. A lot of people do not understand, in practice, the difference between the fact that something is successful in the sense that it creates a lot of clicks and the idea that what you are really looking for is not whether it creates clicks but whether it creates more clicks than what would have happened had you not done whatever you did. So this deep understanding of causality is surprisingly lacking in a lot of mid to upper-level management areas.

I will make this comment later in the panel on the business side a little bit. But it is a little bit as if we have been given the tools to do great data work and now it means that the people who are directing and engaging in using data like this sort of are lacking a little bit of the training to know how to do great data work. So the importance -- this will be my argument later -- the importance of training sort of the decision-making and managerial class up on how to use quantitative methods in order to derive evidence is really important and it is not sufficiently developed at the moment.

DR. SANDFORD: Okay. Another question from the audience for Liad and all panelists. Liad, your presentation highlighted the differences in mortgage offerings and opt-in and opt-out locales. We know that there are racial disparities in mortgage offerings across the U.S. To what extent might opt-in or opt-out affect racial discriminatory offerings and to what extent can or should noneconomic variables, like reducing racial discrimination, be factored into these types of data-sharing decisions?

DR. WAGMAN: So the analysis did control for race composition. It was done at the census tract level and at the individual loan level. And we did notice the other kind of discrimination in this analysis. For example, certain populations were more likely to be denied a mortgage than others. Now, having said that, the opt-out regime, meaning that by default your information would be traded, had less denials for all groups. Okay?

So if we looked at it that way, you know, there are certain benefits that opt-out has that from that perspective. Now, of course, it is kind of -- less denials can be looked at as a good thing, it could be looked at as a bad thing because maybe you are matching loans with borrowers in a less efficient way that could cause downstream foreclosures. So there are all sorts of tradeoffs here and racial discrimination is just one of them. It is just another factor and we did control for it in the analysis.

DR. SANDFORD: Okay, another audience question. I think I will address this to Ginger since she was the Director of the Bureau of Economics and it is a policy question. Ginger, both in terms of theory and practice, how would you compare *ex-post* punishment following data breaches versus *ex-ante* regulation of data practices to minimize breaches?

MS. JIN: Very good question. I think there are pros and cons in both approaches. I think *ex-post* enforcement would give some flexibility for the market to try out new practices and then the Government would not come in until we see a harm to that practice. On the contrary, I think the *ex-ante* regulatory approach will be really hands-on prescriptive. That is like the Government knows what is going to go on in the near future and you have to do ABC in order to pass whatever threshold I am setting. I think that gives a lot of confidence to the government agency and the employees there to decide exactly what is the right level and how would you define the procedure to reach that.

I do think that tradeoff between *ex-post* enforcement and *ex-ante* regulation is a very important one and should have much a wider debate among different disciplines.

DR. SANDFORD: Okay. Next question. So, you know, if I am an optimist about privacy and sort of big tech companies, I might say something like this. You know, there is a lot more data being collected on me now than there used to be, but it is mostly by companies who give me a product I like for free and the way that they exploit that data is mainly by targeting ads to me. And I do not care that much about targeted ads. It may even be a positive. I get things I am interested instead of random stuff.

I think pessimistic scenarios might involve, like, excessive government surveillance or something like that, but there are curbs about that. If I think about big tech companies kind of gobbling up the economy, well, I think, you know, as Ginger mentioned earlier, that companies like Friendster and MySpace and Yahoo and AOL used to be dominant and now they are not. Upstart competitors are able to come along very quickly and with better products and push them out of the market. So I am not that worried about big companies like Google or Facebook because there is competition out there even if there is no company now as big as Google or Facebook. So that is sort of an optimistic view of tech and privacy.

What does that view miss, if anything, and, you know, what pushback would you like to give that view, if anything? Liad?

DR. WAGMAN: I would say that some products can be made better with data. So for example, if I am on a social network and I see my friends there first, even if we are not connected yesterday, that could be perceived as helpful. In the era of Friendster and MySpace, I do not think data was yet used as part of the product, as part of improving the product quality. Today, it definitely is being used to improve product quality.

So entry in this environment is a little bit harder because anything an entrant makes, an incumbent can make as well and use data to make it better. So in that sense, things have changed.

DR. SANDFORD: Alessandro?

DR. ACQUISTI: I feel that both the optimistic and pessimistic scenarios are both plausible. But I also feel that, going back to something I mentioned at the start of my remarks, we really do not need to choose between the value of analytics and the protection of privacy. We do have tools that go in the direction of trying to achieve both.

Once again, I am trying to use language carefully by saying going the direction of trying to achieve both because when you talk about privacy in nascent technologies, you do have to admit they are still young, that they raise some costs. Every time you degrade quality or granularity of the data, you also lower the utility of the data. The interesting, once again, research question for all of us is, if we do use these technologies and they lower the quality of the data and, therefore, they imply some costs, who is going to bear that cost?

Is it the consumer through not so well-targeted offers? Is it publishers that run out of business because they cannot sell as targeted ads? Is it merchants that cannot target it as well? Is it data intermediary? Is it society as a whole? Once again, I believe that we do not have yet good answers to these questions and this is where we should put lots of attention on.

DR. SANDFORD: Okay. Ginger?

MS. JIN: I think one thing sort of really amazing in this space is kind of idiosyncrasy preference. This is not just to say, okay, we all want a safe drug versus a nonsafe drug. It is amazing that different people may have different preferences. Some may be optimistic, some may be pessimistic. Some may sort of have a strong feeling about sort of not giving away my data, but other people would be exactly the opposite.

I think the challenge is how can you design a framework to accommodate that kind of heterogeneity but still kind of achieve protection for those who care about it, but also innovations for those that care more about the products coming out of the data-intensive practice.

DR. SANDFORD: Okay. This may be a factual question and that is dangerous because you may not know the answer. But going back to the issue of competition between firms and privacy, my factual question is, do firms compete in data security? And the reason I ask that is data security is kind of objectively measurable.

I can look at the hash function you are using for your passwords and tell if yours is better or worse than someone else's. It is objective, whereas privacy policies are, one, hard to quantify, hard to measure in any way and, two, if you offer -- if your website offers a different set of services than mine does, of course, our privacy policies are going to be different to some extent. So it is really fuzzy to compare my privacy policy to yours, okay? But data security, for example, how you encrypt the passwords that are stored on your server, is objectively measurable. There are hashing algorithms that are better than other hashing algorithms, yet both are used in the market.

And it seems to me that I have never seen firms make the claim that we have better data security -- well, okay, never is too strong a word. I do not see firms advertising that I have a better hashing function than this guy so you should come to my website. So, again, it is a factual question. Do firms compete in data security?

Florian?

DR. ZETTELMEYER: I think it does exist in the B2B space, not in the B2C space as much. So I think if you think about some of the cloud services, like Box and Dropbox, et cetera, they definitely sell themselves as having superior security features and compete on that.

DR. ACQUISTI: I agree. There is also potential evidence of some effect in the B2C market in regard to data breach disclosure lows. Sasha Romanosky, who is now with RAND, worked with Rahul Telang and myself on a study on the relationship between data breach disclosure lows and changes in identity theft rates in the United States, across all the states. And there was, indeed, a small, but significant decrease in identify theft.

The main variable did not seem to be that the disclosure allows people to actually take action because, as we know, very few people actually take action after receiving a notification in the mail about their records being compromised. But companies, in order to avoid the significant fees associated with disclosure ex-ante, are investing more in security to avoid the data breaches.

DR. SANDFORD: Okay. So this is really interesting, the point about B2B versus B2C to me. I mean, in fact, when we do merger review at the agencies, we spend, I would say, the majority of my time since I have come here has been spent on talking to businesses as customers of merging parties, say. So it is interesting to me that B2B customers have strong preferences for data security, but, you know, end user customers like myself might not.

Does that suggest that if we think about where antitrust enforcement may need to do something different than it is doing now about data and privacy, would that suggest that it would be mergers that involve businesses as companies? We have one minute left. So that is a good wrap-up question, I guess.

MS. JIN: I think there is an information problem similar to what we have discussed before, maybe this is less in the B2B world. If I claim that my cloud has the best security in the whole world and a business customer may, to some extent, confirm that if they have a sophisticated technician to double-check that, but it is almost impossible for individual consumers to double-check that. If we sort of lack that kind of information look-back, then the firms can all claim that we have the best security and then sort of shirk on that claim.

DR. SANDFORD: Okay. Would anyone like to avail themselves of the remaining 31 seconds?

(No response.)

DR. SANDFORD: All right. Then we will wrap up 27 seconds early. Please join me in thanking the panel. Great job, panel.

(Applause.)

DR. SANDFORD: We will take a -- I believe it is a one-hour break and come back at 1:00 p.m.

THE BUSINESS OF BIG DATA

DR. COOPER: Welcome back from lunch. I am James Cooper. I am with the Bureau of Consumer Protection at the Federal Trade Commission. I will be moderating this panel on the business of big data. So this morning, we heard a lot about some great research in the economics of big data. And so we are going to kick off this afternoon talking about how big data is actually used in a variety of market segments.

So we have a great panel to go over this today. We have Christopher Boone, second to my left. He is the Vice President of Real World Data and Analytics for Pfizer. Liz Heier, right next to him, is Garmin’s Director of Global Data Privacy. Marianela Lopez-Galdos is the Director of Competition and Regulatory Policy for the Computer and Communications Industry Association, right next to Liz. Mark MacCarthy, further down there, is the Senior Vice President for Public Policy at the Software and Information Industry Association.

Morgan Reed is -- three minutes, two minutes ago, you were not there, I just realized that. Morgan Reed, I have not seen him. So Morgan Reed is the President of ACT, The App Association and he also serves as the Executive Director of the organization’s

Connected Health Initiative. Next to Morgan is Andrew Reiskind. He is the Senior Vice President for Data Policy for Mastercard Worldwide.

And then, finally, to my immediate left -- and he is right here because he is going to go first -- is Florian Zettelmeyer. He is the Nancy L. Ertle Professor of Marketing at the Kellogg School of Management at Northwestern University. You have already heard from Florian this morning.

So the way this panel is going to work is we are going to -- each of the panelists has between seven and ten minutes, which will be enforced very, very vigorously. And after that, we will hopefully have a vibrant discussion and we will also be collecting as we did in the morning, collecting questions from the audience as we go.

So without any further delay, let me hand it over to Florian.

DR. ZETTELMEYER: Well, thank you again for having me.

So what I want to talk to you about today is not data, per se, but I think a core complementary asset to data, which is the ability as a firm to manipulate it and to use it. And so what I want to start with is first the observation that I am going to call that complementary asset to actually operate and use data analytics -- the terms are getting slightly muddled. Some people are now interchangeably using AI to mean at least a subset of this. But I am going to call it analytics.

So the first thing to realize is that basically everybody today has pockets of analytics. There are areas where, for example, the airlines have forever had pockets of analytics and revenue management because this was so crucial for their ability of doing business. The oil companies have had pockets of analytics in oil exploration and assessment of geologic formations, et cetera. So everybody really has them.

The trick really is not that they do not exist; the problem is how do you connect them and how do you scale them up at the enterprise level? And that is what a lot of CEOs are worried about, how do I take this expertise and organize in a way that actually allows us to leverage analytics and, therefore, data at scale?

So the point that I want to make today is very simple, which is that I think that companies today are held back by a lack of data science skills at the leadership level. And it is not by the lack of data scientists, that may also be a constraint, but it is a lack of data science skills at the leadership level itself.

So in order to make this point, I am just going to start off with an anecdote that I would like to share and it goes like this. So a little while ago I was invited to a thought leadership retreat in a company that operates in the automotive space. This is a company that is partially responsible for placing ads, and as a result of this, has good visibility on how or what consumers do on the online level. And so I was at the car dealership retreat and I had a senior executive of the company who comes up and basically tells that they are excited because they have been able to do something that nobody has been able to do before, which is to link online ad exposure with offline sales, which is a hard thing to do.

So this executive comes up and says, let me show you what we found. We ended up classifying people who used search engine advertising into four buckets: People who saw no ads for cars, people who saw dealer ads only, people who saw manufacturer ads only, and people who saw both kinds of ads.

And then the exec says, what you see here is the sales conversion rate, the probability that somebody purchases a vehicle after having been exposed to either no ads or dealer ads or manufacturer ads or both ads, and this person says what you can see clearly from here is that the conversion probability goes from 0.7 to 3 to 5, to 14 percent. So this is clear evidence, this person says, that search engine advertising really works and that, in addition to that, it is clear evidence to the fact that dealer and manufacturer ads are complements and not substitutes because 14 percent is more than the sum of 5 plus 3 percent.

So at this point, there is like an excited discussion in the room, people talk for 15 minutes about what this means for industry and how this can be monetized, et cetera. And then there is a person who says, we should put a press release out about this because this is really cool and nobody has seen this so far in the industry. And so at this point, it kind of goes on for 15 minutes and somebody raises their hand in the room and says, let me ask you a question. Why would somebody not see any correlated ads when they are on a search engine like Google? And the answer of course is, they did not search for a car.

And then this person says, so why would somebody see both an ad from a dealer and manufacturer on Google? And the answer of course is, they probably typed in a car name like “Chevy Silverado 1500" and maybe a location that would trigger a deal keyword, like Washington, D.C. And so then this person says, so you are telling me what we have shown -- and he points towards -- you cannot see this here from my -- okay, you see this now.

But he points to this row here, the no ads column, and says, so tell me what we have shown is that if you are not interested in buying a car, you do not buy a car and pointing towards the very right; if you are really interested in buying a car, you buy a car.

So the point about this chart is the following, which is that this data is utterly uninformative about whether advertising works, at all. And the reason is that I do not know whether the difference between 0.7 and 14 percent is driven by the fact that, you know, the people who are on the retail and the manufacturer side and are getting exposed to ads and the people on the left did not or whether it is driven by the fact that they were more interested in buying cars in the first place. Those two things are undistinguishable in this data set. In fact, it is extraordinarily difficult, if not impossible, from this data to say how well search engine advertising works.

And the reason I am bringing this up is because it took the executives in that room 15 minutes and a prompt to realize this was useless data and it should have taken them ten seconds. And if you are trained in causal inference, if you are trained, for example, as an economist or as a social scientist, you see this in ten seconds and start laughing about it.

And this is essentially the problem that I am talking about. I have done this with hundreds of executives and it is the norm that people fall for this inference at the beginning without thinking about it more carefully.

Okay. What is underlying here is that analytics, the typical view of analytics is that analytics is a big data and a technology problem. In other words, that it is something where you, in order to solve it, you need to invest in big data analytics and technology infrastructure, like Hadoop and Hive and R and Python and whatever; that you have to invest in cloud computing, like, you know, Amazon Web Services or whatever other company is doing cloud services, that you have to invest in data scientists.

And I am not saying these things are not important. In fact, they are essential. But the point is they are nowhere close to enough because, at the end of the day, analytics in practice turns out to be mostly a leadership issue. It has to do with things like managerial judgment in which there is nothing wrong with the data I showed you. But what is wrong is how you interpreted this data and many people get that wrong.

Analytics often has the nasty habit of ignoring organizational boundaries. And so, often, data sharing in companies that crosses organizational silos and profit and loss responsibility is very difficult to achieve and it has to be achieved at the top leadership level in order to create those kinds of alignments.

Analytics has to be fundamentally problem-driven. It is really difficult to start with a set of data and say, let me see if I can find something interesting. It virtually never works in practice. But that means that the people who have the problems need to be involved in actually bringing them to bear on analytics issues, and those are decision-makers and executives.

And then the last one is that what a lot of people also do not understand at the executive level is that most of the data that is lying around is actually not particularly useful; that a lot of the data that you need, in particular, as you become more and more sophisticated as a company, needs to be planned and acquired and designed as opposed to collected opportunistically in the normal course of business.

So we think that this means that leaders need what we call a working knowledge of data science, which means judge what good looks like, identify where analytics adds value, and lead with confidence. And the consequence of this is that this working knowledge allows you to make the big managerial decisions, like what tools to invest in, what data you need, what org structure you need, and what people you need because in order to link the problems you want to work on and the C-Suite priorities, it turns out this working knowledge allows you to make that link.

Thank you very much.

(Applause.)

DR. BOONE: So it is ten, right?

DR. COOPER: Yes, seven to ten.

DR. BOONE: So seven to ten, all right. Do not start the clock just yet. Wait one second.

(Laughter.)

DR. BOONE: I want to make sure I reclaim my time like Maxine Waters.

Thanks to the members of the Federal Trade Commission and for the opportunity to provide you with commentary on this very important topic. I would be remiss if I did not acknowledge my distinguished group of fellow panelists on the stage with me here today. But I am going to move on with my comments. I have no slides. So we are just going to talk through this.

When it comes to the topic of big data, no industry has felt the weight of this magnitude like the healthcare industry. As the U.S. healthcare system swiftly evolves into a more consumer-centric model, there is considerable interest in increasing access to medical care and therapies for patients, demonstrating value of care and therapies to patients, and improving clinical outcomes with patients.

Historically, healthcare provider and peer organizations were in the business of providing acute care to patients under a traditional fee-for-service model. However, each has come to recognize and appreciate the need to understand the genetic, behavioral, social, and environmental factors often referred to as the social determinants of health that contribute to delivering positive outcomes and value for patients.

This has, in essence, spawned a new era in healthcare delivery, an era of continual delivery where routinely collected data is continuously fed into a system and ensures we have the information to learn from patient experiences and clinical outcomes. In short, I am referring to the establishment of a learning healthcare system that is built on healthcare informatics, big data, and advanced analytics.

So the $64,000 question is why now? The ubiquity of digital health technologies has served as a key enabler for providing this level of care while generating massive amounts of healthcare data or big data. Big data in healthcare is a direct result of the technological advancements in the industry, advancements that include the accelerated expansion of electronic health record platforms, rapid adoption of smartphones and wearable technologies, penetration of social media in our daily lives, cost reductions and genome sequencing, and the repurposing of nonconventional data sources, such as consumer, social economic, and environmental data sets, along with the sophisticated data, analytical tools and techniques, have created an environment where data is a valuable asset.

In a broader sense big data in healthcare is often referred to as real world data and it holds the potential to significantly increase the efficiency and effectiveness of all process in the development and utilization of medicines from research and development to regulatory decision-making, to pricing and reimbursement decisions, and even clinical practice. Moreover, real world evidence of the output of the analysis of real world data could supplement the evidence generated from randomized clinical trials, which could considerably improve healthcare decision making for all stakeholders.

So what exactly is real world data and why all the excitement? Over the years, the terms “real world data” and “real world evidence” have been used mistakenly as synonymous terms. According to the researchers for the U.S. Food and Drug Administration, the FDA, real world data is defined as data relating to patient health status and/or the delivery of healthcare routinely collected from a variety of sources. These sources typically fall into four major grouping, the first being clinical data, which is patient-level data pulled from electronic health records and/or patient registries that describe treatment in the real world.

The second category is administrative claims data, which is the data that is primarily used for billing purposes by providers to insurers or other payors. The third category is patient-generated data, which is data that describes the patient's experience and is collected and shared by the patient his or herself. And the last category is the nontraditional health-related data sources, such as your behavioral, your social media, environmental, and/or socioeconomic data.

Real world evidence, on the other hand, is defined as clinical evidence regarding the use and potential differences or risks of a medical therapeutic derived from the analysis of real world data. The simplest way to think about it is real world data is any health data not collected in a traditional randomized clinical trial and can also include data from existing secondary sources.

The importance of real world data is critical to all stakeholders across the entire healthcare value chain including physicians, payors, regulatory bodies, patients, and, yes, pharmaceutical and medical device manufacturers. Many are familiar with the use of real world data for informing decisions related to patient treatment options, coverage determinations or even policy options, but some may not be as familiar with how pharma companies actually use real world data. Pharma companies are using real world data and real world evidence across the entire product life cycle to identify targets for the development of new therapies, support regulatory submissions, advance disease understanding and clinical guidelines and support outcomes-based reimbursement decisions.

Real world data analysis has been identified by various regulatory initiatives, including the 21st Century Cures Act and the Prescription Drug User Fee Act, as useful supplements to randomize clinical trials. Specific applications include the acceleration of drug approval pathways and expanded indications for approved medical therapies. When it comes to the process of collecting and analyzing real world data, generally, we think of it in three stages.

The first stage is the study planning, which is where we seek to understand the evidentiary needs of key stakeholder support groups, such as a regulator or a payor. We then formulate a research question that then feeds into a study designed where we identify the appropriate data sources to conduct that study. Now, it is equally important as part of this processing to assess the availability, accessibility, portability, and even quality of the data for that particular study.

The last stage is where we actually communicate and socialize the actual results of that particular study through a scientific publication. From the perspective of Pfizer, we primarily connect deidentified data to use in our real world data study analysis from third-party data aggregators. If there are any data linkage and/or aggregation activities required, we work with these aggregators, who possess the technical expertise and competency, to effectively collect, manage, and link the patient data.

Now, the benefits of analyzing real word data for consumers or patients generally we feel is tremendous. We live in the world where most of the health-related data is collected outside of the walls of a provider organization. For example, consumers now possess apps on their smartphones that allow them to perform tasks such as recording daily vital signs, documenting daily food intake, and even detecting triggers or symptoms for certain clinical events. These real world data sources and studies that are associated with it are vital to documenting and understanding the benefits and risks of medical therapies in a heterogenous population and to determining whether patients in routine clinical practice are achieving positive outcomes.

As is often the case with cutting-edge scientific and technological advancements, a full understanding of the ethical and policy-oriented implications lags behind. There are several key considerations to keep in mind as we think about big data privacy and competition. Quite frankly, I do believe many of the key policy and ethical considerations are pretty much industry-agnostic, which means that we tend to all deal with the same major issues.

At the high level, the issues that are well documented are around informed consent and privacy. Some other concerns that are starting to bubble up are issues around data ownership or the rights to use the data, the appropriateness of methods to analyze the data, the appropriateness of the question being analyzed, and even the legal context for which this analysis takes place.

According to a 2017 consumer voices survey conducted by Consumer Reports, 70 percent of Americans lack confidence that their personal information is private and secure. Ninety-two percent of Americans think companies should have to get permission before sharing or selling their online data and 92 percent of Americans think companies should be required to give consumers a list of all the data they have collected about them.

Privacy concerns related to allowing the access and analysis with large real world data sets have greatly limited its potential. Since pharma manufacturers do not generate real world data directly, data access, data availability, data portability and data quality remain significant barriers to advancing the science.

Other ethical considerations that the FTC should keep in mind are the existence of big data divides, which is created between those who have or lack the necessary resources and infrastructure to effectively analyze these large data sets. The next one is the monetization of data and the potential problems with ownership of intellectual property generated from the analysis of these aggregated data sets.

And lastly, the future of real world data and evidence is in the aggregation of genomic and other “omic” data and the possible dangers of intentional or unintentional group level ethical harms, specifically as it pertains to patients' beliefs about the benefits or harms to a particular racial or ethnic group in studies.

There is considerable high hopes for the use of real world evidence to improve decision-making in the U.S. healthcare system, but all stakeholders have a role to play. Pharma manufacturers have a critical role in driving innovation by using real world evidence to support clinical trial designs and observational studies to generate evidence and new treatment approaches. However, the need to protect personal data, consent, ethics, and data access are equally important and harmonization of public policy and legal frameworks will be necessary to realize the full value of real world evidence.

It is critical that the FTC, as part of its role to protect consumers and promote lawful competition, take affirmative steps to promote ethical use, data ownership and privacy as its pertains to big data and healthcare. These are important considerations to keep in mind as the FTC reviews the state of big data in business and how it affects consumer privacy and industry competition. Pfizer stands ready to discuss the shared responsibility with all interested parties to make this vision a reality.

Thank you.

(Applause.)

MS. HEIER: I am a little bit shorter. Well, first, I want to say thank you to James Cooper and the rest of the FTC staff for inviting me to participate today.

My name is Liz Heier and I am the Director of Global Data Privacy at Garmin. It is a bit of a coincidence that I am following Chris since we are a wearables company.

My 11-year tenure with Garmin did not start in data privacy. My diverse IT experience includes software development, both as an engineer and a manager, incident management, and data security. These roles have given me a unique perspective on the multifaceted issues corporations face in the areas of data protection and privacy.

Garmin was founded in the Kansas City area in 1989 by Gary Burrell and Min Kao, whose belief in the potential of using GPS in avionics and in consumer electronics was not shared by their then current employer. They believed so strongly in the product they were creating that they named the company after themselves by combining their first names, Gary and

Min. This was long before Hollywood came up with Brangelina and Kimye.

(Laughter.)

MS. HEIER: Since its founding, Garmin has grown into a global company of over 12,000 employees spread across 60 offices worldwide. We create products in five market segments, aviation, marine, sports and fitness, outdoor recreation, and automotive. We recently shipped our 200 millionth device.

Over the last three decades, Garmin has grown and thrived through its innovation, ingenuity and diversified product lineup. In the 2000s, a majority of our revenue came from our automotive personal navigation devices which sat on our consumers’ dashboards. By the time that product became saturated and turn-by-turn directions were ubiquitous on mobile phones, Garmin was ready with new, market first products in our other segments.

We have seen phenomenal growth in our sports and fitness segment in recent years with the popularity of our wearables and their companion mobile apps, websites, and services. As I mentioned recently, Garmin recently shipped its 200 millionth device. It was only six years ago that we crossed the

100 million mark. Much of that rapid increase can be attributed to the popularity of our wearable products.

Many of the owners of these wearables choose to provide their data to Garmin through our mobile apps to enhance their user experience. This means that Garmin has been entrusted with the personal data of millions of users from nearly every country in the world. At Garmin, we believe the data that our customers create and upload through our apps and services belong to our customers. We believe that these apps enrich the user experience of our devices and, in turn, enrich the lives of our customers, whether their goal is to become healthier, share their adventures with friends or fans, or travel more safely in the water, in the air, on the road or on the trail.

Garmin makes money selling our devices and we have no need to monetizize our customers' personal data to be profitable. It is not in our business model nor our corporate culture to sell customers' personal data. Today's constantly evolving technology allows our devices to record increasingly detailed and powerful data sets. Through the sensors in our wearables, our customers can monitor their heart rate in real time, as well as view graphs of historical values and averages, all of which could reveal indicators of potential medical issues, such as sleep apnea or atrial fibrillation.

Our devices can detect a bicycle crash and automatically alert a user’s emergency contract with his or her GPS location and our devices can help consumers navigate hostile terrain while sending text messages to their loved ones to let them know all is safe or to call for help if it is not. These are critical services to many of our customers. But the data required to provide them could be harmful if publicized or misused.

We recognize that our customers put their trust in Garmin when they share their personal data with us. We believe that our customers should have the ability to make informed choices when deciding when and how much data to share.

A large majority of our products can be used fully out of the box without ever connecting to the internet. For those customers who do choose to use our apps and services, all sharing options are set to private by default and many individual features can be turned on or off, thereby putting the customer in control of what personal data are processed.

If the customer decides to no longer use our services, he or she could delete their data at any time. We also do not share their data with anyone unless our customers ask us to do so, nor do we constantly track the location of every Garmin device on the planet. So as much as we would like to help your lost or stolen Garmin device, we just cannot.

When the GDPR was approved by the European Parliament in 2016, as was true for many companies, it was Garmin's legal team that began to campaign our leadership and our board of directors that the GDPR issue was big, hairy, and not going away. Our leadership got the message and soon realized that data privacy was not only a legal concern, but something that would have to be integrated into our culture. And that is where I came in.

I am not a lawyer, I am a software engineer. Who better to work with engineers on the GDPR than one of their own? With a strong governance team of key executives, business leaders, and legal counsel supporting me, we used a risk-based approach to create a compliance program that was guided by pragmatism, transparency, and usability. In that spirit, Garmin supports a federal privacy law that would preempt state law and position U.S.-based businesses to better compete in a global economy.

The GDPR is not perfect, but there are many things it gets right, and any U.S. company that does business in Europe has already invested in complying. Garmin alone invested more than 800 person-months of effort to ensure compliance. Consistency and data privacy laws benefit everyone by lowering the cost of implementation, reducing complexity, and allowing for globally recognized and understood paradigms.

One of the things I believe GDPR got right was that it largely harmonized data protection regulations across the EU. Prior to the GDPR, companies that do business across Europe had to navigate the complex data protection regulations of all EU member states. This resulted in confusion, inconsistencies among the various regulations, and a higher cost of compliance. Having a harmonized regulation in the EU, even one that sets a very high bar like the GDPR, brings much-needed certainty to all involved, including the regulators, the businesses, and the consumers.

Without a federal privacy law in the U.S., we would risk going backward to a place like the pre-GDPR European Union where companies could be forced to comply with numerous, possibly inconsistent, state privacy laws. We have seen California recently enact a privacy law and the trend will almost surely expand to other states in the absence of a federal privacy statute that preempts state privacy law.

A federal privacy law would also pave the way for trusted transfers of data between the U.S. and the EU without the uncertainty of yearly assessments and frequent challenges to available transfer mechanisms, like Privacy Shield and standard contractual clauses. Like Garmin's services, today's economy is global and it is cost-prohibitive for companies to maintain localized data centers for every country. We need trusted and stable methods for data transfer that allow personal data to be stored in and managed from locations where resources, both technical and personnel, are available.

In closing, the personal data and associated processing activities, including big data, provide valuable, often life-altering, benefits for our users whether they are taking their first steps towards a healthier lifestyle or are training for next Ironman Triathlon. Adequately securing their data and handling it responsibly and transparently is a duty that we take very seriously. We support federal data privacy legislation that would promote consistency and align with today's global economy.

Thank you.

(Applause.)

MS. LOPEZ-GALDOS: Hi, good afternoon, everyone. My name is Marianela Lopez-Galdos. I am the Director of Competition and Regulatory Policy at the Computer & Communications Industry Association, and we represent big and smaller tech companies from the U.S. and elsewhere. Before I get started, let me thank James Cooper and the FTC for inviting me to be here. It is a great opportunity for us, but also the FTC more broadly for putting together all these hearings. I know there is a lot of effort behind it, so we really commend you for that.

So we are trying to understand how companies use data and I think what I am going to try to do here with my brief remarks is try to explain to you the role that data plays for data-driven companies like the ones that operate in the digital economy. And I bring here today with me three ideas.

First, that data is not essential, that ideas are. Second, that in the digital economy, innovation rather than market positioning is more important. Finally, that as technology progresses, we will see that the need for data will diminish, so therefore we need to be very careful and ensure to preserve the incentives for companies to keep innovating in this industry.

So let me get started with my first idea. What do I mean by saying that data is not essential, that ideas are? What I mean is that similar to the brick-and-mortar world, in the digital economy, companies exist, flourish and compete because they have a good idea and then that idea allows them to bring to the market a product and a service that consumers like. Therefore, it is not access to data, what allows these companies to compete and to exist, but, rather, the initial idea.

So we need to clearly understand that an idea comes first. And this idea that I am -- what I am saying about data being essential seems very obvious, but we sometimes forget when we discuss the role of data and the role that data has for the digital economy that successful winners exist not because they have access to data, but, actually, because they bring to the market something, a product or a service, that consumers lacked. And we have many examples of these in the market if we look at recent history. For examples, you can see how Snapchat or Slack basically became very successful companies without having access to data in the beginning. We also see how Handshake has become a very strong competitor to LinkedIn with more than 14 million users right now among recent graduate students.

And we will have an opportunity to listen to Catherine Tucker, I think, later this afternoon and we have been listening to her during these hearings, also to Professor Lambrecht, and I think in a paper they published recently they have a quote that I would like to share with you because it really summarizes the idea that I bring with me today for you.

The history of the digital economy offers many examples like Airbnb, Uber and Tinder, where a simple insight into consumer needs allowed entry into markets where incumbents already had access to data. So this is how we summarize my idea that data is not essential. But there is something more that I would like to share with you today, which is that the more access to data, it does not bring added value to some companies.

So there is -- Stanford University conducted a study with a set of images from dogs. And they managed to prove that more data gives you better results in data analytics, but to a certain extent. There are limited return for companies when analyzing, for example, images. And I am happy to discuss more about the Stanford study later during our discussion.

But, you know, if you think about our own personal experiences, imagine when you were trying to buy a car and you spent six months looking into cars in the market or looking into different brands as the first speaker explained today. So that data becomes late as soon as you purchase the car. So the value of data is quite limited. And, therefore, we need to be very careful with those who argue that data is an essential input because that rests on a misunderstanding of the concept of data and the role that data represents at least for the digital economy and data-driven companies.

And that leads me to my second idea, which is that innovation rather than market positioning is what drives the digital economy. What do I mean by this? If we accept that data has limited diminishing returns and that it is not essential, then we can actually understand that data cannot be used to drive a competitor out of the market.

So how do companies compete with data? Well, what they do is invest in what I want to call today here, data-driven R&D. They really need to invest and understand data analytics. Because once they have access to data, if they do not have the right analytics and the right decision-making processes for the results that data analytics gives you, that data is basically useless. So that is how companies compete, investing in R&D, investing in innovation.

And basically that leads me to my third and last idea which is that as technology progresses, we see a lot of advances. We have come from the IBM linear computing to quantum computing and now we are talking more about machine learning and more broadly AI, but what we are really talking about is machine learning. And in machine learning, data analytics is fundamental.

If we speak to engineers working in this area, you will learn that they are progressing quite significantly in the last years. And, for example, now, you will hear them talk about synthetic data, where they use kind of artificially-created data that does not track back individuals, so confidentiality and privacy no longer becomes an issue. But, also, you will hear them speak about zero shot learning which is basically a methodology used by machines to recognize objects without having been trained or without having received label training to recognize an object. So, for example, a machine will be able to distinguish a zebra from a horse without having seen a zebra before. So this is what is happening in the digital economy and this is where technologies -- digital companies are investing money and they are advancing quite quickly.

So if we understand that with the progress of this technology, the access to data will diminish over time -- the importance of access to data will diminish over time, we understand how important it is to preserve the incentives to innovate and how important it is for our progress and for the future of AI and machine learning to make sure that we do not intervene in data-driven markets unless there is actual harm to consumer. And by preserving these incentives to innovate, we will make sure that we can keep progressing for our society. And with this idea, I stop here and I look forward to our discussions.

Thank you.

(Applause.)

DR. MACCARTHY: So my name is Mark MacCarthy. I am with the Software and Information Industry Association. And I want to thank the organizers of this workshop, James Cooper and others, for inviting me to be here today to talk about these data analytics issues.

I liked the phrase that you used “analytics” rather than AI or machine learning. It covers a broader range of things.

Let me tell you a word or two about SIIA. We are a technology trade association. We have three groups of members, one group of the traditional technology companies, companies like Adobe and Intuit, Red Hat, although Red Hat just got bought, I think Google and Facebook, and then we have information service companies, companies like LexisNexis, Thomson Reuters, Refinitiv, which used to be part of Thomson Reuters, Dun & Bradstreet, and we have ed tech companies, companies that provide personalized learning services to schools, and that is companies like Pearson and McGraw-Hill and Cengage.

I want to talk to you today a little bit about some of the uses of data and analytics that these companies are involved in, and I want to talk about four specific cases and just remind you at a high level the kinds of things that are being done today with data and analytics.

So the first one I want to talk about is the production of fair and more accurate credit scoring models. The second is the increase in speed and effectiveness of student learning caused by personalized learning technology. The third is the improvement in online personalized ads caused by the new machine learning techniques, and the fourth is the improvement in business risk analytics that is taking place today.

So first, general remarks. There is a new development in the data analytics world, but it is a natural evolution of the older techniques. There is a lot more data that is available. It is different kinds of data and the speed at which the data becomes available is much more rapid. So the techniques used for processing this data are different. And the key thing is that the new techniques allow the detection of patterns that would not be available to human intuition and that are not based on prior hypotheses that are developed by researchers. They emerge, so to speak, from the data itself. While the results are sometimes startling, it turns out that the policy issues that are raised by these newer data analytics technologies are much the same as the older policy issues.

So with that as a general remark, let me get into the discussion of credit scoring. You all are probably familiar with credit scores. The credit-scoring models have been used for generations. They increase the accuracy and fairness of credit-granting decisions, certainly compared to the human judgment of loan officers who often use subjective assessments. But the traditional credit scores have limits. They do not effectively provide scoring for almost 70 million Americans because they rely heavily on data that is from credit reports and that relies mostly on payment information. And this deficit adversely affects, historically, disadvantaged minorities. A study by LexisNexis found that 41 percent of that population could not be scored by traditional credit scores.

So they developed their own credit-scoring model, largely by going to new sources of information, new data sources, educational history, home ownership, court records. And with this new availability of data, they found that they were able to score fully 81 percent of previously unscorable applicants for credit. And this example shows that even just expanding the kind of data being used and not really using dramatically new modes of analysis can dramatically improve outcomes.

In the credit-scoring world, there are also machine-learning models that are being developed by researchers and they will soon be ready for deployment in practice.

The second area I want to talk about is personalized learning. Researchers have shown that many students who eventually drop out of high school can be identified as early as sixth grade. And the basis for this identification is their behavior, their attendance in classes, and their, of course, performance. Even more can be identified by the time the students reach the middle of ninth grade.

Now, early warning indicators based on these data points can be used and can generate risk scores. This knowledge will allow schools and teachers to provide these students at risk more meaningful interventions and support. And when this happens, it increases the number of students that graduate ready for success either in further schooling or in their careers. In one school in 2013, fully one-third of the students who were being flagged for being late at school or missing school got back on track after these remedial programs.

Personalized learning also will help target students according to their learning styles and bring to them the best available learning techniques. In a developmental math program, math courses, used in one community in Chicago, a program called ALEKS, which is produced by McGraw-Hill, uses artificial intelligence to help students progress through the material and it adapts the material to their learning needs. The schools that are using this program report that this new technology gets students through their remedial material much more rapidly than traditional methods.

So let me move on to the third area, improved personalization for online ads. This really takes place at two levels. One is the analysis of website movements, which can aid websites in providing material, content material, and ads, and improved analyses of large customer databases. Now, we are all familiar with this, the movement of website visitors on a website is usually recorded and it contains data, such as which pages are visited, how long you spend on which page, how you shift from one to another, the sequence and so on, and critical patterns of that kind of usage that cannot be identified by human beings or by eyeball inspection of the data that can be inferred through machine-learning programs.

And once these patterns are discovered, website visitors can be segmented into different groups based on the preferences that are inferred about them and the website's content can be personalized to those preferences and the ads that are served to them can be personalized to their interests and needs.

A second way, companies often have large aggregations of their own consumer data or they can obtain them readily from third parties, and they need an effective tool that can detect patterns in the data that will enable them to become better at their marketing campaigns. Now, machine-learning programs can dig through data to find insights that can be used to devise smarter and more effective ad campaigns. They are so good that they can also advise marketers what type of campaign to use, whether it is email or social media engagement or online advertising or recommendations on websites.

In addition, the use of inferred psychological characteristics is often a good mechanism for improving the effectiveness of advertising. The level of extroversion, for example, or openness can be inferred from social media behavior, and if you match the content of advertising to this characteristic, you can improve responses significantly, according to one study, an increase of 40 percent more clicks and up to 50 percent more purchases.

Now, of course, the benefits of these

increasingly effective target ads is the ease and convenience of consumers who are seeing material that is more appropriate to their need. But, also, additional revenue to provide ads supported free or subsidized content.

Let me shift to my last topic, improved business risk management services. Information service companies help their business customers to manage their risks using data sets that they have acquired in various ways. These data sets usually rely on public records and information about people in their business capacity, their status as directors or officers or stockholders of companies, and they also include lots of nonpersonal information, such as the financial and operating characteristics of companies, including how well they have paid back their own debts.

Now, the predictive analytics component of this includes the likelihood of repayment of a business loan or a profitability analysis that would assist a company in a merger analysis. These techniques also help companies make better decisions and manage risks, like identity theft, fraud, money laundering, and terrorism. Regulators also want financial institutions to detect terrorist financing and money laundering using whatever techniques are most effective.

The coming thing in this area is that the same machine learning-techniques that can spot a pattern of bad transactions in the credit card world can also be used to assess the risk that a potential customer would engage in these kinds of suspicious activities.

So that is my quick survey of the areas where big data and analytic techniques are improving things. As I say, one of the major policy takeaways is that while these are new techniques and sometimes produce startling results, I do not think they raise fundamentally new policy issues. And let me put off the discussion of those policy questions for the give and take later on in the discussion.

Thank you for listening to me.

(Applause.)

MR. REED: Well, given the number of people I have been watching slowly move their eyes down to their smartphone, hopefully looking to an app while they are there, I realize that we have started to hear some of the same stories from panelists as we have gone down the line. So I decided to try out some of my notes and kind of try to weave a little bit of a story into what we are up to in the healthcare space, but that maybe can run some threads and even ask some questions for my own panelists for the later session.

So my name is Morgan Reed and I am the President of the App Association. And I hope -- well, a quick show of hands so we can throw out the infidels. Anybody here not have a smartphone? Excellent, thank you all for keeping me fully employed, I love you all. It is great to have you here.

Here is the thing, the technology that I work on and the industry that I help to lead as the President of the App Association is the fastest growing technology in the history of mankind. Full stop. We have successfully put access to the world's connected information into the fingertips of roughly two billion people and we have done it in less than ten years. It is faster than fire, it is faster than the wheel and faster than the next fastest adopted technology which was the microwave, kind of cool, actually, the microwave, the second most fastest adopted technology after the smartphone.

So with all of this access to information and such a life change revolutionary idea, information in the hands of more people, there is a concomitant secondhand to that, which is data. What do those people know? How are they feeding back into this collective system? So when with you all my fellow panelists talk about their kind of segmented chunks of big data and the way that they use it and how well we are protecting it and we are making sure we are being very careful with it, all of that is true. But what we have not really talked about here -- and Mark hit on it and we have touched on it a little bit -- is this is kind of amazing. This is life changing for billions of people in a good way.

Yes, we need to protect it. Yes, we need to be careful with our regulation. Yes, we need to think about how it implies and what it implies when it comes to competition. But let's not lose track of the fact that it is life-changing and life-beneficial to billions of people around the world. So let's dig into some specifics.

I head up the Connected Health Initiative element that we are part of and let me tell you some really depressive things about America that hopefully will not cause people to start drinking until after it is over, but by 2025, the United States will be 90,000 physicians short, 90,000 physicians short. By 2030, we will have 70 million Americans over the age of 65.

And I will give you a little secret, people over the age of 65 are sicker.

Two weeks ago, I testified before the Senate Health Committee, Senator Enzi is the chair. My colleague to the right was the insurance commissioner from the State of Wyoming, who, great guy, former rodeo rider, said in that drawl, the State of Wyoming currently only has 157 physicians for every 100,000 people. If you are sick in the State of Wyoming, leave the state to get care.

So the question that we need to be thinking about when we are asking these questions about big data and the business of big data is, what does it provide to people? To consumers? And I am here to tell you that the demographic numbers are clear. If we do not find a way to engage with digital medicine and big data, we cannot support the number of people in this country who will need quality care. Cannot do it. No amount of money, no amount of change will get the number of physicians that we need to have in practice.

So a little bit more upbeat, right? We can do stuff with data. So what do we do with it and what are some examples of how we move forward? Well, primarily, I thought one of the interesting things my first panelist said was, well, we only need structured data. Anybody in this room, do not show hands because it is medical information, but I am going to assume that everybody in this room has someone that they know that has some kind of autoimmune disorder, whether or not it is one that is related to chronic fatigue syndrome, rheumatoid arthritis, any of the other concomitant diseases that go along with it, impacts from Hashimoto's thyroiditis, all of those cases we do not actually know what is wrong with you. That is the depressing part.

Anybody who has an autoimmune disorder, you go to your rheumatoid arthritis specialist and they say, well, let's try this. And part of the reason why is that physician that you are seeing, so they are at the top of your game -- a physician at the top of their game has seen roughly 29,000 patients by the time that they get to you. Of patients that will have your identical comorbidity, your genetic type, your age, your other key factors, where you live, everything else, you are lucky if your physician has seen 500 people that look like you.

So that means your physician is going to base your treatment off of what they learned 15 years ago in school, that continuing education class they took a booze cruise somewhere and, hopefully, 500 points of data. You should be angry at that. The fact that everybody on this panel has talked about how we are absorbing and utilizing big data and, yet, why is it that your physician is making a treatment decision based on 500 minuscule points of data that you hope will be relevant to your condition?

So as you consider the question of the business of big data, the question that you should be asking is how do we use the business of big data to actually produce a better consumer outcome? And in the healthcare space, I am going to offer a couple of very obvious examples that we are working on right now.

Through the Connected Health Initiative, we work with academic medical centers, businesses, the American Medical Association, patient groups and others. One of the leading areas that we have real difficulty in in this country is, of course, type 2 diabetes. It is an epidemic. It is one that we know how to solve and, yet, people keep continuing the same behavior.

So what can big data provide us in terms of insights? Well, there is a company out of Georgia called Remedi, but spelled with an I. Look at it on

Twitter. You will be able to see it after this session. They are actually using remote patient monitoring data from wearables, like yours and others, to paint a picture of a person. And here is where the analytics and the big data comes in and really makes a difference.

Through something called clinical division support, they actually allow a physician to model the treatment of the patient before prescribing it to them. They actually take in the data from your electronic health record, combine it with wearables information and they create patterns and they say, well, this treatment schedule has about a 60 percent chance of likelihood of success. This one, we see that a person with these similar conditions, you are likely to see this outcome.

The decision is still in the hands of the physician, hence the support part of clinical division support, but ultimately allows that physician to bring in multiple data sets, look at it, overlay it, and instead of going to the doctor and saying, take these pills and three weeks later we will see how you do, they are able to run multiple scenarios prior to your treatment so they get closer to the right answer.

Now, that requires large data sets -- and something Christopher said that I am always a little concerned about is this idea that, well, you know you have got real world data, how do you bring it in and integrate it? I think all health data needs to have that real world element in. Because where we live, what we eat, what our genetic situation is, is all part of figuring out how to be healthy.

And Chris said something else, he talked about consumers. One of the parts that we -- that Christopher and I know in this case is this difference between patient and it has to do with how you are paid. But I realize that a patient is actually a person. None of us want to be a patient, right? If we are sick, we want to get healthy; if we are healthy, we want to stay healthy. And what we need to look at is how does big data get us there. So a product like Remedi helps to get us there.

Earlier on, you said that, you know, we need to structure all that data, but one of the things that we have learned is if we do not know the answer, then I cannot necessarily structure the data the right way to answer the question. But I did agree with that first point, which is it is all about asking the right questions.

So as we go through the rest of the panel and go through the Q&A, we will talk a lot about how do we provide short form notice and what kind of consent mechanisms do we need and what are the regulatory necessity of the GDPR or other elements.

But the primary question you should be asking is, how does big data actually produce an outcome that is good for consumers/mankind, for patients. Because, right now, the medical care you are getting that does not rely enough on big data should not satisfy you. You need to ask more of your data and ask more of the healthcare system that can use that data because we can do better using big data.

Thanks.

(Applause.)

MR. REISKIND: Good afternoon, everyone. Thank you, Morgan, for waking us all up with good news. It is always good to get good news after lunch and keep everybody awake.

So my name is Andrew Reiskind, Senior Vice President for Mastercard. I am responsible for data strategy and innovation. And so who is Mastercard, what is Mastercard? I think most of you are familiar with it as a brand name, but you do not necessarily know what we do.

So we are a network, we are a technology provider. We connect your banks, your consumer banks to the merchants' banks and, therefore, enable cardholders, you who are holding an account, to actually make a purchase with a merchant. But you are not our customers, the merchants are not our customers for the most part for our core network. Instead, you are indirect customers.

So as part of that, we are the pipes that connect everybody to each other. So we do not issue the cards. That is one of the biggest fallacies people have about Mastercard. Instead, see the logo on the front? That says Citi. And if most of you pull out your cards, you will see it has the bank's name who you have the relationship, who you give your personal data to. Instead, you have these things on the back that says "bug," which we call acceptance marks, that says if you go into a store, this will be accepted.

So what does that mean from a data perspective? From a data perspective, I do not have a data relationship with consumers. Instead, what I have is I get enough data to process a payment. What is that? That is an account number, the amount -- the time of the transaction, the total amount of the transaction and the merchant.

So actually, I would like to say 40 years ago, somebody had the foresight to actually do some privacy by design because I do not have your name. I do not need your name to process a transaction. I do not know what you actually buy. I do not need that to process the transaction. Instead, the bank gets the information. The bank says, oh, $50, Yael has $50 in her account, yes, she does, and she is waving her hands and so, therefore, I will approve the transaction. Well, I think Leisl is actually Leisl and I will approve it because I actually think it is her actually making the transaction. Or if I think it is some fraudster, I will not approve the transaction.

So what do we do? So as a result of that, we see 55 billion transactions or so. The number keeps growing exponentially, thank goodness, for our jobs, of transactional data. So what do we do with that data? Well, I will tell you one of the great things that we do with it is we innovate. We are constantly developing new products and solutions, and one of the most important products and solutions that we develop to help all of us, me inclusive as a cardholder, is to protect all of us from fraud.

So what does that mean? So, historically, where all of the data’s been coming from that amount of transaction, time of transaction, that is happening when you are at the cashier in the old days, right, and you would swipe your card. It would come through and we would see it and then the bank would have to authorize that transaction. So you would stand there and hopefully wait only the five milliseconds where you are saying, it is approving, approving.

So during that time, we have tools that enable us to do determinations and to start doing risk scores to say, do we think this is fraud? Do we not think this is fraud? Now, those have evolved over time. In many cases, they used to just be rule-based, simple if/than. Nowadays, we use AI to do it.

So as we have grown our models, as we have grown our technology, we are able to protect people more and more. And another great thing about this, as the rest of the world has moved to adopting cards and payments through accounts like that, then we have enabled protections against fraud for those consumers across the world.

Over time, though, we say, okay, this is our basic data set. How else do we help improve the fight against fraud? Because it is an arms race. There are constantly new players coming in trying to steal data.

There are constantly new players trying to come in and make runs against banks. So we work with the banks to say, hey, how do we help you here?

So in many case, we have worked with the banks to segment you. So we use the data to help determine, hey, here are classes of consumers and this is how they behave, and based upon those classifications, this is what we think fraudsters look like. This is what we think your people look like. So does this help make determinations? Does this help you reduce fraud?

Another service we work with them on is to actually get some information from them or have them get information. So in e-commerce situations, a merchant can pass them the name and they can actually also check that name. Now, Mastercard does not have to get the name. Instead, we are enabling the pipes that allow for passing the data.

Then as technology has evolved, we have evolved to new payment forms. Now, who has Apple Pay, Android Pay, and who has used it? Very nice, simple, easy way. And I am sorry, Garmin, you guys can use it, too, on Garmin. So we helped build the backbone for that, so you can thank the payments industry for enabling you to just put it on your phone.

But when you put it on your phone, what are you doing? Most of you, I think, have gone through an authentication experience. You are providing your name and address so that Apple, in one case, just sends it through to your bank and your bank then confirms that is you. So you are authenticating yourself to your device.

Mastercard just needs that data for a very short time period. We do not really need to retain it. I do not need to continue to authenticate you. So, again, privacy by design, it happens once. But, now, you get to be authenticated to your phone. And so, now, I have an additional way to say hey, this phone actually is Andrew and, therefore, I get that little flag that says, hey, this is great, Andrew just got authenticated to his phone. Mastercard does not know Andrew; Mastercard might know that it is Apple, Apple device or a Garmin device in the case of Garmin that the payment occurred. So that is how authentication might work.

The other way with e-commerce merchants is we work with e-commerce merchants and mobile merchants, m-commerce merchants, to say, hey, guys, if you give us more data or enable some collection of data, we can help you fight fraud even more effectively. So imagine if I'm only seeing the same account number against the same iPhone. Gee, that persistency tells me that there is a reduced risk of fraud here or if I see the IP address as from certain parts of Eastern Europe, that are known for high fraud, I can also say high, high risk of fraud here.

So those are the kinds of things that we are doing to try and help fraud. This is how we -- to fight fraud, not actually help advance it, sorry. And so we are constantly looking at new ways to use data, to look at new data sets, to build on data sets, but as we are doing that, we are trying to minimize the data sets we have. If you do not have the data, you cannot lose it. If you do not have the data, you cannot accidentally abuse it. So, therefore, a lot of privacy by design and data minimization as we are doing product development, but all in furtherance of a good cause to help to protect all of you from fraud.

(Applause.)

DR. COOPER: All right, thank you, Andrew. And, you know, we have about 23 minutes left of discussion here. I heard a nice panoply of the uses of data across different industries.

One of the things I heard a couple of panelists mention and I would like to get others who have not weighed in on this to speak about, is when we think about big data, you know, one of the things that sometimes sets apart big data from just normal data is that you are looking -- the analysis that is performed on it often is looking more for patterns that emerge that you could not see with smaller data sets.

You are looking for associations, as opposed to, when you think about it, sort of normal in economics -- you know, Florian mentioned this in his presentation earlier this morning about the, you know, kind of gold standard of causation -- and you are looking in control groups and figuring out.

So one of the questions I had -- and since we have not heard from Florian in a while, I will start with him, but anyone else can jump in -- is, you know, in general, what is the relative importance of both looking and finding causal versus associations, and sort of related to that, when you think about big data, what is more valuable? And I think I already know the answer you are going to give as seen in your presentation. But what is more valuable having a good team or knowing how to ask the right questions or actually having access to a large and comprehensive data set, actually having access to big data? So what is more important in that?

And I will start with Florian, but I would like anyone else to jump in.

DR. ZETTELMEYER: I think on your first question about kind of what kind of data is the most useful, I would simply say that it is incredibly context-dependent. Roughly speaking, I think of analytics creating three things. It can enable business initiatives. Like if you think about personalization, that is really an enablement function. You are creating something that allows you to achieve an outcome. A lot of the things, for example, that Morgan was talking about I think fit in that area as well, as well as a lot of the things that Liz was talking about, design-abling things.

Then I think the second big use is that it enables you to basically come up with ideas. That is what you were talking about, about large data sets where you can look at correlational patterns and see whether you can come up with ideas from that.

And the third one for me is that data allows you to evaluate whether things that you are doing are reasonable or not and whether they work or not. So, for example, my first talk this morning was really about evaluation. It was like, you know, is this ad working or not? It did not help you come up with the ad, it did not help you necessarily kind of enable the ad. That is what these -- obviously, these targeting mechanisms do.

So I think it just depends completely on what the purposes are. I think one of the mistakes sometimes people do is to think too narrowly about what uses of data exist. And they are very different from each other and you need very different data. Sometimes it has to be causal. In many cases, causality is not at all interesting or required. It is just a matter of what you are looking for.

On the second question of what both you need, I actually think that data and skill teams are complements and not substitutes. So to the degree that you have better data, having the ability of asking great questions suddenly becomes more valuable to a particular firm.

DR. COOPER: Okay, thanks. Morgan, down there?

MR. REED: So it was interesting. You know, I think that it is one of those that are intertwined. But I know that there are some folks in the audience here who are more specialists in this, but I think some of the things that have been revealed through some of the criminal justice reform analysis of big data have been profound and a bit disheartening, but they go to this value of -- what is the old phrase? That quantity is a quality all its own. And sometimes in data the ability to see large shifts or check for some various effectiveness, as Florian talked about, is almost impossible because to separate the signal from the noise is too hard.

And so I think when you say, well, what is the most valuable aspect? Skilled teams, data set, size, those elements of it, I think they are fairly intertwined, but I would recommend that everybody take a look at criminal justice reform questions where big data has been used to show some, like I said, fairly depressing things about if you want to go before a judge, make sure you do it at this time and not after -- you know, before lunch but not when they are hungry. The fact that hunger seems to have more of an impact on whether or not you go to jail as opposed to what you have actually done as a crime.

I do not think you can reveal that without big data sets. And then as you point out, you can reveal it with big data sets, but you have to be able to ask the right question.

DR. COOPER: Mark?

DR. MACCARTHY: So I thought the magic word in the last comment was context-dependent. So do you need large data sets or small data sets? You know, it is like it depends on what you are using it for. Sometimes you need a large data set to get the result. As I think you mentioned earlier, there are studies that show that these effects of size diminish after a certain point and you can add more data to the data set and you do not get anything new. So there are diminishing returns.

And also in a context-dependent sense, whether the information you have is valuable for a long period of time or whether its value decays quickly depends on the context you are operating in. If you have search information, that decays very, very rapidly. You know, someone may be identified as being interested in a vacation in Maine in August, but you better not send him an advertisement for that in December, he probably is not interested.

But on the other hand, medical information might be very valuable years after the data has been collected. The analysis can still be done even though the information is not fresh and insights can be gathered even though data is not last year's data. So I think it does depend on the context and we have to be very, very careful not to make broad generalizations about how valuable is data over time or whether large data sets are better than small data sets. You have to look at the context in which the information is being used.

MR. REED: I want to amend my answer with one thing that Mark brought up that is really important. Mark said something really important. He said, “but medical data.” And here is the thing you heard in what Christopher said and what Liz talked about and what Mark kind of brought up, which is we are not 100 percent sure what is medical data. When we are trying to figure out whether or not there is a cancer cluster, I may need to look at other factors that might not be obvious, that might not have fit into our current understanding of what is medical data in terms of how the FDA judges our product.

So I think, Mark, you were spot on and I think it ties into with what you heard from Christopher and Liz and others. We are not exactingly sure of all of it, but we want treatments that reflect us as a holistic person not merely the data that is contact in our EHR. So I think it is a good point, Mark.

DR. COOPER: Florian, you had a quick followup?

DR. ZETTELMEYER: Yeah, I just wanted to say one more thing about the complementarity of the data and the team. A lot of firms for internal processors are using data to basically improve decision-making. So one of the interesting things about this is that the better decisions get as a result of having used data, the less variation exists in business processes because the data was used in order to optimize those decision processes. This is why, you know, we use the data in the first place.

What that also means is the data is getting less useful over time because now you have less data variation, and as a result of that, the importance of the team is to know when to inject more variation into the data in order to be able to still measure what is going on. In other words, you say do you have experimental design and variation of data and thinking of manipulating or rather designing or varying data as a strategic imperative is incredibly important. That does, at the moment, at least require some teams to set that up.

DR. COOPER: Anyone else like to jump in?

MR. REISKIND: I think I will just reenforce. I have had very personal experiences dealing with geospatial data lately, because a lot of our analytics are based upon where a merchant might be located, as well as your cell phone might be located. And some of the analytics can be worked off of very crude locations, like especially outside the United States, quality of data is kind of limited. There are not postal codes, there is only one city in the entire country, things like that. And you have to work with that as a data quality issue that you cannot overcome, and so it limits some of the things you can do.

But there are things you can do with that data, but they may not be as good as you want to do. So, for example, to tie my cell phone to my physical location, my cell phone to my physical location where I am making a spending purchase would be our nirvana. And in some cases, we can get to that nirvana to prove my iPhone is where I am making an expenditure is a great thing, because them it proves I am not a fraudster. But in many cases, you cannot get there.

So you have to mediate what your innovation is and what you are trying to do based upon the quality of the data that you are dealing with as well as the skill of the data scientist and the tools you have to work with the data. Geospatial data is a very unique data set -- sorry, postal addresses tend to be not very useful for analytical purposes. You need to take 4100 Yuma and actually turn it into a lat-long for analytical purposes to stick it in a model. 4100 Yuma will not work very well in a model, as can you imagine in a mathematical algorithm.

So, therefore, geospatial data sets at least need that level of transformation and, yet, that is only as good as the maps are in Third World countries or underdeveloped countries in many cases. So that is just an example. Like it depends on the data set, it depends on the tool, it depends on the use case. It is all very context-driven.

DR. COOPER: I want to switch gears a bit here. Liz talked about this in her remarks, about regulation that we see, the GDPR and the recent California privacy law, that both -- what I would be interested in hearing from all of you is to what extent do you see either of those types of regulation impacting your use of data and how might that ultimately impact consumers. So anyone who wants to jump in.

Mark, you had your hand up first.

DR. MACCARTHY: So I think it depends a lot on whether you are dealing with a large company or a small company. The compliance burden for both California and for GDPR, for large companies, if it is the kind of thing that they can do, and with sufficient resources, they can find a way to comply, they will be able to do it. I think one of the previous speakers talked about 800 hours of compliance work that was put into getting into compliance.

For larger companies, like many of the companies in my trade association, that is doable. But for many of the smaller companies -- and we have 700 companies in my trade association -- many of whom are very, very small and they would love to operate globally. For them, the choice came down to enormous compliance costs for operating in Europe versus not operating in that market at all, and for them, it was an easy choice.

So I do think we have to pay very, very close attention to the compliance costs that are imposed on businesses. If something is really needed to protect consumers against real harm, then you got to do it and people pay the compliance costs. But if it is just a lot of extra processes, you know, put in there to validate that you are doing the right thing, then there may be less benefits from those compliance costs than we would like.

MS. LOPEZ-GALDOS: I completely agree with you, Mark, and I would like to add just a tiny bit there, which is that resources that are taken to comply with the laws because, obviously, if we adopt regulations, companies are going to comply with them, those are the resources that the smaller companies are going to stop investing in innovation. So we also have to look into the actual effects of the need to comply with the law.

And it is certainly the case that big companies can comply with those new laws much easier than smaller ones. So I think that is a very important point that Mark was making.

MR. REED: And it was worth noting that Liz mentioned not 800 hours, she said, hundreds of person-months. So I want to remind everybody that here is the part that is so cool. Earlier, I talked about two billion people having access. My smallest companies are global players. Our current board president has an app -- kind of a cool app, he has 2.8 million users in about 117 different countries. He is a one-man shop in Oregon.

My example I always use is my literal smallest company, Ann Adair's company that makes kids apps, she is a music teacher that is a part-time coder with her kid and her husband and has a whole slew of really cool kids apps. She is a global player with hundreds and hundreds of thousands of users. So Liz's point about hundreds of person-months to comply with GDPR has a real implication.

And I will dig down to one area of specific that gets into the business of big data. If you are not familiar, right prior to the launch of the GDPR, the Article 29 working party released a letter directed at ICANN specifically about the ability to using the word "including" in your terms of service. And this is always an awkward thing to bring up because everybody is essentially ignoring this letter.

In this letter, ICAN was told, you may not use the word "including" because to use the word "including" means you are not being complete, comprehensive, and explicit. And here is the problem. We are on a panel of the business of big data. How can I cover all the algorithmic learning that I am going to do and be explicit and comprehensive when I quite literally do not know the answer of where the data might take me and back to causal and correlative effects.

So I think there are moments where well-meaning regulators will put language in like that and then the outcome, from a data science perspective, is, well, I do not know what the outcome will be, so how can I be comprehensive and explicit? So I think we need to be cautious about just jumping on board and say that the U.S. version of GDPR needs to plug and play. I think we need to ask real questions about how it will impact good use of big data to solve real problems that people have. So hundreds of person-months plus loose regulatory language will have an impact.

DR. COOPER: Did you want to jump in?

MS. HEIER: Yes. So just to kind of clarify, right? I said 800 person-months of effort and that is really not correlated to the number of users we have or the number of countries we operate in. We have 30 years' worth of devices, services and data that we had to bring up to compliance. So it does not really matter necessarily size of the company. It is really your offerings, right?

So as you said, it could be one person that is operating out of their garage part-time, but operates and has lots of data. Their cost of compliance is going to be much different than ours.

DR. COOPER: That is a good point. Anyone else like to jump in before I get into some questions from the audience?

(No response.)

DR. COOPER: Okay. So this one is directed at Mark, so I will let you take first stab at it, but open it up for everyone else. And it has to do with you talked about credit scoring, how using alternative data and big data methods can actually lead people who do not have credit lines to have lines and be scored or are unscored and be scored.

And this question says, perhaps that makes sense in a credit-scoring situation, but sometimes if you are training a data set -- if you are training these algorithms with historical data, in other contexts, perhaps, they can ingrain bias. So is that something that you should worry about in the context of big data and AI?

DR. MACCARTHY: Yes. Actually, credit scoring is one of the areas where they have had experience with bias and statistical discrimination going back for generations. The credit scoring world is under a legal obligation to avoid the discrimination in lending. The fair lending laws require all of the credit scores that are used in that area to pass a disparate impact test, which means they have to look carefully at whether their algorithms have an adverse effect, a disproportionate adverse effect on minority groups. And if they do, they have to ask themselves, what is the particular purpose they are involved in that makes this disparate impact so important? And if they have a legitimate business need, then they have to also ask themselves is there another model, another credit-scoring model that will achieve the risk reduction that they are looking for with less of a disparate impact?

So all of the credit scoring models have to pass that test if you are in the business of producing one of those models to people who buy it or people who will be examined by federal regulators for compliance with the fair lending laws. Now, if you happen to use machine learning, you know, in that context, that is not a get-out-of-jail-free card for getting rid of discrimination charges. It does not work to just say, well, I used artificial intelligence so I do not have to comply with the fair lending laws anymore. So the new techniques are as much covered under the old laws as the old techniques were, and in that particular case, there really is a regulatory requirement to avoid discrimination.

DR. COOPER: Would anyone else like to weigh in on that in general? I mean, I think related to that, a bigger-picture question is, in general, we think about using big data, using analytic methods or the predictions. Are they more -- we have to look at what the alternative is. Are they more or less discriminatory than what the alternative would be or more or less accurate than what the alternative would be? And I just wonder if -- this is kind of related to the question from the audience that was thrown out to Mark. Does anyone have any thoughts on that?

MS. LOPEZ-GALDOS: I can jump in. I think one of the keys is going to be able to explain, and AI models are going to have to be able to explain how they operate. So definitely the laws are there, the principles that need to be protected are there. The fact that you use an AI or machine-learning methodology is not going to change your obligations, as Mark said, and the difference is that we are going to have to determine what the explainability of those AI models are going to be to be able to prove that we comply with the laws. So I think that is going to be key.

DR. COOPER: All right. We are rapidly running out of time, but here is another question from the audience that says, if we look at data as an asset, how should companies treat this from an ownership perspective? Should it be treated like intellectual property? Should consumers have any sort of ownership interest in this? So how should we think about big data in this context?

MR. REED: Well, there are multiple stages. We have rules governing your health data. Your health data is your data. But the question is once it is manipulated, once the physician has put additional work and information into it, then where does it stand?

The work product of the physician is valuable and valued. So how do we work with that becomes a real question. When it comes to something most people do not know -- we have not talked about HIPAA at all, but the P in HIPAA stands for portability not privacy. So a lot of the questions about big data are very interesting because your health data in particular is something that there is a push to make it portable so you can move it from place to place so the physician is well armed in order to treat your disease.

The question was interesting and you touched on it earlier when Marianela was talking on the intellectual property question. The explainability and transparency of the algorithm also gets very interesting in so much that what you have trained and what you have learned is also a work product of your company and might be protected. So how do you separate the data sets from the work product? If the data set -- if the work product is actually trained off of those data sets, then which thing is the asset?

I think the reality is healthcare is, in a weird way, almost easier because there has been this kind of agreement across the industry that your health and your specific healthy information is yours, the patient's property. But it does get interesting into the question of what is the value of the work product that is created off of that data set and where does that set in the realm of intellectual property.

DR. COOPER: Mark?

DR. MACCARTHY: Yes, I think the ownership lens is the wrong one to bring to bear in this kind of circumstance. I mean, most information is about more than one person. I mean, if I bought something from you, then you sold something to me, and so the question of who owns the data is an attempt to import sort of property law into that circumstance and it just does not help you very much in trying to figure out what the right thing to do is.

If I own the data, does that mean I can destroy any copy of it anywhere, any business record in the world I can sort of destroy because it is mine? Well, that does not make any sense. So I think you might as well go directly to the data protection rules and regulations and the responsibilities on both parties to try to figure out what the right thing to do is, rather than say, I am going to define who owns it and that will end the problem because now I know who owns it.

I think you will not be able to solve the problem of determining the right owner, so I think you just have to go to what are the rules, what kind of consent needs to be given, what kind of access is there, what kind of portability rights are there, and those things really take a lot of careful and hard thought, and you cannot really solve those problems by saying, I fixed the problem, I decided who owns the data.

DR. COOPER: Okay. Florian and then Liz.

DR. ZETTELMEYER: So, Mark, I agree with you that that is true on the regulatory side, but, I mean, in terms of data usage on the company’s side, that is a problem that shows up all the time, and particularly in disintermediated industries like, you know, do you own your data or does the physician own the data or does the retail own the data or does Procter & Gamble own the data and what are you allowed do with it, et cetera. So, I mean, it is an issue that companies have to grapple with. It may not be useful from a regulatory point of view, but it is certainly something that is pretty omnipresent in this data world.

DR. MACCARTHY: I think you have picked on the key, which is what are companies allowed to do with it. That is the question. You do not resolve that by saying, I know who owns it, therefore, I know what use requirements there are. I think you have to go directly to the use restrictions and constraints and who has what right to do what with it.

DR. COOPER: Liz? This will be the last word.

MS. HEIER: Well, just to reiterate what I said in my statement, Garmin believes that the data belongs to the user and the customer. They give it to us to help enhance their experience, to give them new data points they would not have on their own. So we have really formulated, you know, our data privacy program around that user-centric focus.

DR. COOPER: That is perfect, zero, the clock is at 30. So well-timed.

THE IMPACT OF GDPR ON EU TECHNOLOGY VENTURE INVESTMENT

DR. STIVERS: Okay. I think we are going to go ahead and start the afternoon session so we can keep our somewhat amazing track record of staying on time for this hearing. So thank you to OPP and the FTC staff for having kept us on track.

I am Andrew Stivers. I am the Deputy Director for Consumer Protection in the Bureau of Economics, which just means that I am basically in charge of the Consumer Protection economics mission at the FTC. I am delighted to basically just introduce a series of really good speakers this afternoon. So I am going to step out of the way and we are going to start with Liad Wagman, who is a Professor at the Illinois Institute of Technology in the Stuart School of Business.

Liad?

DR. WAGMAN: Thank you again for having me here today. This is joint work that is fresh off the copy machine pretty much with Ginger Jin and Jian Jia. We have been in a mad dash to complete it over the last several weeks.

Basically, we looked at GDPR and we asked ourselves where would we notice an impact right away. And the answer we came back with is that investors are likely to internalize the effects. So we thought the law was passed a couple years ago, back in 2016, maybe we should notice an effect then because investors would form expectations. The thing is, not much was seen and we were wondering why.

Looking through the news events, we saw that as recently as early this year, more than half of mobile applications are not GDPR-ready, and announcements very close to the implementation date, to the enforceability date of May 25th, kept pouring in. The top firms, the top platforms started releasing their rules.

Apple removes apps to share location data without consent, updates their privacy terms. Facebook says that businesses may want to implement code that creates a banner and requires affirmative consent. Each company is responsible for ensuring their own compliance. You are all on your own. Shopify updates its app permissions for merchants/developers. They need to implement them. Google releases consent SDK for developers, these software development kits, just a day before, the eleventh hour before the enforceability date, and then GDPR takes effect. So we kind of understood this all came to this implementation stage of the regulation and so the effect should be noticeable after that or as this was happening.

So this is sort of our motivation, you know, GDPR has a massive overhaul of data regulation in the European Union and anyone who services the European Union. That includes data management; auditing and classification; data risk identification; risk mitigation; interfaces for users to obtain their own data to provide opt-in consent and to request deletion of their personal data. Firms are required to train or hire qualified staff or they face severe penalties that are up to 4 percent of their annual global revenue.

Bloomberg, shortly after, said, 500 biggest corporations are on track to spend a total of $7.8 billion to comply. Now, based on earlier work, we already knew that compliance costs are not incurred equally by firms. Smaller firms tend to take a bigger burden, at least in relative terms. And the other effects we know from theoretical work is that compliance cost will shift some of the innovation activity from smaller firms into the bigger firms.

And the reason, especially for tech, that this happens is because larger firms already have the infrastructure in place for R&D. They have the infrastructure in place for internal innovation. So when entrepreneurs decide to pursue an idea, they have the option of pursuing it internally or pursuing it as a startup externally, as a venture. When they face that choice, they look at the cost, and when the cost of pursuing it on your own increases, your incentive to stay inside and either innovate or not increases. And so the overall, at least, theoretical effect is that innovation is reduced and more innovation happens inside bigger firms.

So the bottom line for us was who is better to assess what really happens than the actual investors who are putting their money where their mouths are, that are actually investing in those firms. So once these policies were rolled out, we figured compliance costs are going to be realized, especially for the smaller ventures because they rely on the larger platforms' policies for compliance, for who bears the liability for violation, and so forth. So that is the general idea.

Now, we wanted to get comprehensive venture data. It is impossible to get it all in one place, but one of the main databases for venture data that is not a complete universe, but it is pretty good, is Crunchbase. So we collected venture data from

Crunchbase from last summer, July 2017, until the end of September, this year. So it is really, really recent.

This data comprises firm information, the firm location, the category it operates in, its founding date, the dates on which it raised money and a range, a lower bound and upper bound, on the number of employees it has. Think 1 to 10, 11 to 50, 51 to 100, something like that.

Now, it also comprises information about each individual financing deal. That includes the size and the date of the deal, which stage, was it a seed deal, a Series A, and so forth, which investors participated, and the dollar amount obviously of the deal.

So just to give you an idea of what the data looks like and to convince you that it is good data, I created some pictures to kind of summarize it. So these first four pictures show the average number of deals per week in the U.S. and in the EU. You can see the U.S. has a larger number by a factor of two or so. The median dollar amount in millions raised per deal is about a million and a half for the EU and three million for the U.S. You can see the average firm age is more or less similar and the average number of investors that participate in a deal is somewhat higher in the U.S.

If we look at the composition of firm ages in our sample, you will notice that about half of them are the very youngest, the zero to three years old ventures. And the rest are distributed more or less similarly between the U.S. and the EU.

So if we dig deeper into these age groups, you can see that the average amount raised per deal is growing the older the firm is. So the youngest group raised the least, they mostly participate in seed rounds and Series A, Series B rounds, and then grows from there. These are averages; they are not medians, so the amounts are a little higher.

Now, if we look at the total number of deals, most of the deals happen for those young firms. They have smaller deals, but they have a lot more of them. And we are talking thousands of deals in just one year of data, a little over a year. And if we look at the median amounts raised per deal, you notice that, again, they grow in the firm's age, and this is kind of indicative that the distribution of those amounts is skewed. The median is smaller than the average.

So if we want to dig deeper into the types of deals that are happening, I hope I have convinced you by now that this data is pretty granular, but it goes further than that. You will see that for those youngest firms, those zero to three year old firms, most of the deals happen on this large circle which is the seed round. Those are the smallest basically rounds that mainly comprise angel investors and amounts of a few hundred thousand dollars a deal.

Then it goes from there. It goes to Series A, Series -- bridge rounds A-B, and others. So on the horizontal axis here, you have the firm age; on the vertical axis, you have the average dollar amount for deals of that type. And then the larger the circle, the more deals we see.

As we move to older firms, you will notice that the bubbles start floating up as the deal amounts increase and there are fewer deals so the bubbles get smaller. We can go to the older group and they keep floating up, the age obviously increases, the bubbles get smaller. And we could go to the oldest group and they keep floating up.

So in terms of where those deals are happening, this is a heat map of U.S. states and the EU member states. We include Britain in the EU because it was still part of the EU as of the time of

GDPR’s rollout. The EU firms are affected by GDPR just as much. In fact, the U.K. adopted its own GDPR-like law.

You will see most of the deals happen in California, happen in the U.K. In terms of the dollar amounts that go in, it is a pretty similar situation. Most of the dollar amounts go to the U.K. and California and Germany picks up some investment dollars as well.

So our observation level here is divided into a state, where a state is either a member state in the EU or a state in the U.S. So we look at least at the aggregate level at states.

In terms of time, we look at weeks. Investment per week, per state, per technology category. I will talk about categories in a second.

At the deal level, we look at individual deals. So I hope this was convincing at least in terms of the granularity of the data we have.

Let me give you some idea of the trends here. This is for the number of deals per week comparing the EU and the U.S. The U.S. is the red line; the EU is the blue line. You notice that they track each other pretty closely. It seems to be a common trend and GDPR takes effect in late May this year, and there seems to be some change going on. Now, you might argue, oh, this is the European summer vacation happening right after GDPR takes effect, but we do not see a similar thing in the summer of 2017.

We dig deeper into the deal per week per state per technology category level. You will notice that this gap becomes easier to spot, this gap that happens between the red line representing the U.S. trend and the blue line representing the EU trend. And there is a drop that happens after GDPR takes hold.

We could look at variations of this of the dollar, for example, raised per week, and see the same thing. We could go further and look at the dollar raised per week per state per technology category, and again, we can see the same thing. And we could look at the dollar amount raised per deal and, again, we see something similar taking shape.

So our next objective here is to quantify this effect, to look empirically at what is going on. Our methodology is what is called difference-in-difference. So what we do is we find the difference in the U.S. from the pre-period, before May 2018, and the post-period, after May 25th, 2018, and we do the same thing for the EU, and then we take the difference of the differences.

So we have a couple specifications. At least at the aggregate level, we use Tobit for the total dollar amount raised per week per state and we use Poisson for the number of deals per week per state. We use macroeconomic controls, like unemployment, consumer price index, GDP. We even included exchange rate. That did not change anything.

And a specification is what you would expect. We are just looking for the effect of the rollout of GDPR. We use time and state, country fixed effects for the EU, and at the deal level, we use a log linear specification because of these outliers that we have where we see the average is much larger than the median and this helps control for that.

At the deal level, we also include the deal-specific controls like the age of the firm, the funding stage of the deal, technology category, things like that. And in terms of technology category, we break it down into two categories. One is healthcare and finance, and the other is everything else. The reason we focus on healthcare and finance is because the U.S. has existing laws in those sectors; specifically, the Gramm-Leach-Bliley Act, GLB, for finance and HIPAA for healthcare. So we would expect maybe to see something different about that category, that grouping of healthcare and finance.

The other reason we divide it into these categories is because it creates a valid sample in the sense that every state has some activity in those categories.

So in terms of results, we see an effect on the dollar amount raised per week per member state per category that is substantial. Across all EU ventures, that dollar effect is $3.38 million per week per state per category. For zero to three-year-old ventures, the effect is almost a million dollars.

Now, in terms of the number of deals, we see a significant drop, a drop of about 17 percent for the number of deals per week per category per state. The figure represents the average amount, just to make it easier to kind of relate to. And we see a similar drop for those youngest ventures, those zero to three-year-old ventures. What this means is that those firms have less of a chance to secure a successful deal which could mean that fewer of them come to fruition.

In terms of the dollar amount per deal, that also drops. Those drops are pretty significant in the overall sense because some of the later deals are very sporadic. When we zoom in on the zero to three-year-old ventures, the drop there is 27 percent.

Overall, we see two effects. We see an effect at the extensive margin in terms of fewer deals taking shape after GDPR takes hold, and at the intensive margin, in terms of fewer dollars invested per average deal.

Let's talk about some of these categories more specifically. So in terms of healthcare and finance, we see a similar drop in the number of deals of 18.8 percent. We see a drop in the aggregate amount raised per week per state of $5 million. The average amount invested per week is $30 million. And we see a huge drop in the amount invested per deal, on average.

Now, interestingly enough, we see similar changes for all other categories. We do not get a significant effect on the aggregate dollar amount invested per week because that pool of categories is just too widely spread. It is too broad. So we are not able to identify that effect, but otherwise it is somewhat similar. This is surprising because you would think that healthcare and finance would be different since the U.S. has existing laws.

Now, what we get out of it is that maybe

GDPR is really transformative in the overall sense across categories. It doesn't matter if there are existing laws; those laws are old. They are outdated. There are systems in place already to handle those laws. Whereas GDPR is new, is fresh, needs new systems, new compliance costs.

Now, zooming back into those zero to three-year-old ventures, those nascent ventures, those startups, the effect there is pronounced. There are 19 percent fewer deals happening. There is a decrease in the aggregate dollar amount invested per week and there is a drop in the dollar amount invested per deal on average. That is, to me, concerning. And at the same time, we do not know if it is a short-term effect or whether it is going to last. We only have four months of post-GDPR data. So that is something to keep in mind. This is at least the short term that we observe -- the short-term effect that we observe.

So in terms of robustness, we looked at the pre-periods before May 25th. At least at the deal level, the number of deals, we did not see an effect before May. At the total dollar amount raised per week, we do see an effect that starts a little bit earlier. It starts in April, late April, kind of crossing over to May, and we see it kicking in in early May, really kicking in. So firms were reacting. They were reacting to those announcements.

So as a robustness, we exclude May from our sample and all the results still go through. As an additional robustness, we exclude the period between summer 2017 and summer 2018 to control for seasonality, and the results still go through.

We top coded observations to reduce the influence of outliers, of those huge deals, and the results still go through. We categorize industries in an unsupervised manner using techniques like K-means or other machine-learning techniques, and the results still go through. And we used other specifications, and the results still go through. So we tried to break the results and they do not break easily.

So what can we do with this? Well, our data set, as I mentioned earlier, has some information about employment numbers, employment ranges, how many employees are employed per firm. And, obviously, we see these dollar amounts decrease in deals, but what does it say about welfare? It does not say much. We cannot draw a welfare implication for this. It could very well be that those less desirable firms are not coming to fruition. We are preventing the next Cambridge Analytica. Who knows.

But we can look at the effect on jobs. So to do that, we got an average for the dollar amount raised per employee by zero to three-year-old firms, and that range is from $123,000 to a million dollars. And we can use this range to see how many jobs are lost because of the less dollars that come into those firms. The fewer dollars that come in terms of the number of investment deals and the dollars per deal.

Just if you are curious, how many dollars are raised on average for a broader swath of firms, say, zero to six-year-old firms, you see that those dollars shrink. And the reason they shrink potentially is because those firms have outside revenue sources. I mean, they have their own revenue sources, perhaps.

So those zero to three-year-old firms are the most susceptible to job losses. They depend on that money in order to hire those people. They depend on those deals coming through in order to operate. So in terms of jobs lost by those firms, based on our back of the envelope, these rough estimates, we see that it is between 3,600 and 30,000 jobs and that amounts to about 4 to 11 percent of the number of employees they employ in our sample.

I want to emphasize that this is the effect we see in the short term. We do not know what is going to happen in the long term. And it could very well be that investors are just pulling out and saying I want to see how this is going to shake up. I will come back later. It could also be that investors are shifting their dollars to the U.S., in which case, our results may be overstated. It could also be that there are investors outside the EU that tend to invest in EU firms that hold their dollars back. We do not see them in our sample because we only focus on the EU and U.S., and so maybe our results are understated.

And the other thing to keep in mind is that these jobs lost are just technology jobs in those zero to three-year-old ventures, at least these rough estimates. There could be more jobs lost. There could be jobs lost by firms that are older. There could be jobs lost by people who would have acted in service positions for these jobs, providing lunch, providing child care, and so forth.

So just to kind of summarize what we see so far is that in the short run, we notice a pronounced negative effect on EU venture financing, both on the number of deals and the dollar amount per deal. Our sample of post-GDPR is relatively short, so more study is definitely needed here. And the reason that investors are holding money back is not crystal clear. It could be a wait-and-see approach. It could be that they are afraid about rising compliance costs. It could be that this regulation is hindering the actual business practices that they want to invest in or the products they want to invest in. It could just be uncertainty.

The other thing to keep in mind is that our sample is a small part of the bigger picture. We do not have a complete universe. We think it is a pretty good sample, but there could, of course, be more.

The other thing we notice here is that GDPR is very transformative. It applies across categories, even those categories we would expect may be less of an effect because of existing laws like HIPAA.

So just one difference between HIPAA and GDPR, one of many, is that HIPAA might require you to provide consent in order to receive service from a healthcare professional, whereas GDPR requires the firm to provide service even if you do not give consent.

In terms of Gramm-Leach-Bliley in financial markets, that regulation provides an opt-out approach. It basically allows customers to opt out of having their data, say, sold to affiliates. Even that is in special circumstances. Whereas GDPR requires an opt-in approach, you have to provide opt-in consent for your data to be used, for your data to be sold.

The penalties are also very different. GDPR has much larger penalties, potentially 4 percent of global revenues.

So aside from the negative effects we see on the number of deals, we also have some conclusions or at least preliminary conclusions for job losses, and, again, it is a rough calculation. Other than that, I would be happy to take some questions.

DR. STIVERS: Thank you, Liad.

So first of all, I would like to say to all of you, hopefully a number of you are researchers in this area, this is the kind of work that is incredibly valuable, both to the FTC and to our sister agencies working in this area, in terms of really trying to understand what the potential effects might be of changing regulation, changing the course in this area. So if you are in this field, I strongly encourage you to -- ah, there we go. I thought I had gotten the button. I guess I had not.

Hopefully, you heard me that I strongly encourage you to do research in this area. I know that a couple of you have some very interesting work coming forward in this area, so we are all eagerly awaiting that.

However, since Liad is here, I get to grill him a little bit. I wonder a little bit about the time period that you are looking at. You look at the time in which GDPR was actually -- the enforcement happened. Did you think about looking at the April 2016 shift? Because you would expect that investors maybe would be -- this was not a surprise, that it was coming, even though I think you point out that perhaps some of the companies were kind of last minute in terms of getting their compliance up and running.

So can you talk a little about why you would not necessarily see most of the effect happening right around April of 2016, before and after, and then what are you actually measuring? Are you measuring the entire effect of GDPR when you look at the May 2018 date or is there something a little more subtle about what you are measuring there in terms of the effect?

DR. WAGMAN: Right. So first, I would like to say that I think both time dates are meaningful. April 2016 is when GDPR passed, came into law, but it was not to kick in until two years later.

Now, the second time period is meaningful because that is the actual implementation stage. A lot of these smaller firms that we focus on, they depend on the policies that are adopted by the larger firms, and those policies were not announced or not adopted until the very few weeks, if not the week of, May 25th, 2018.

So a lot of the realization of those increased compliance costs, those increased liability costs, the actual code that you needed to put in your app in order to be compliant with the app store where your app is published was not available until those few weeks preceding May 2018, at least for the most part. Just to give an extreme example, Google released some code the day before.

Now, we looked at April 2016; in fact, we started with that, and just our early checks did not reveal a significant effect. It could be just, you know, lack of clarity about what was going to happen. Now, we saw that lack of clarity from regulators as well. If you look at their own models for kind of trying to predict what would happen after the regulation, they had their own uncertainties. And those uncertainties, I believe, are still not clear. Until several probably lawsuits settle down, we will not know the full effect.

DR. STIVERS: Okay. Thank you very much,

Liad. If you can thank our speaker.

DR. WAGMAN: Thank you.

(Applause.)

BIG DATA FAILS: RECENT RESEARCH INTO THE SURPRISING INEFFECTIVENESS OF BLACK-BOX AI

DR. STIVERS: All right. We are going to move to a recorded presentation from Catherine Tucker of MIT. And as soon as we move forward in the slides, it is going to start, which is why we still have Liad's last slide up here.

Good, all right.

RECORDING: Good afternoon. My name is Catherine Tucker and I am a Professor at the MIT Sloan School of Management. Today, I am going to be presenting some research I have into the surprising area of big data in the online advertising world.

Before I start, I have two apologies. The first one is obvious, I apologize very much for not being at the hearings in person. I have teaching scheduled on every single day of the hearings from morning to afternoon, and I am very sorry not to be with you. It looks like an amazing program.

The second apology is, unlike many of the presentations you are going to see over these three days, I am going to be presenting a research paper today, and the nature of the research paper, of course, especially an empirical one like this, is it tends to go after a very narrow set of findings, but makes sure that we can really believe in those narrow set of findings. So the second apology is that what you are going to hear is about a very specific set of experiments in a very specific context.

So having said that, perhaps we should actually move to the context. And as I alluded to, this paper is a paper about big data in online advertising. And to set the background, I want to just remind you about how important data can be when we are thinking about showing ads to a pair of eyeballs on a particular website.

I also want to tell you about different types of data that a publisher of the website -- imagine it's a news site and an advertiser could potentially use to make sure they are showing the right ads to the right person. The first thing they could do is they could use something called first party data. And that is data that the website actually has access to because it knows the kind of content that the consumer has browsed at some point in the past. So if that news website knows that whenever I see a cruise story, I read it, then perhaps they could use that data to make sure they show me an ad for an upcoming cruise.

Now, second party data is a little bit more of a narrow pedigree, and this is a capture view of data where a website has data from a partner and they know exactly who that partner is and what kind of data they are getting. So a good example of that I came across recently is that *Rough Guides*, a kind of travel book, shares data, browsing data explicitly, with lastminute.com, which is a travel website.

And you can imagine why they share data and why it might be useful for working out what ad to show. If someone has just booked a cruise to Italy, then if I am *Rough Guides*, I want to show them an ad about my guidebook to Italy, and similarly, if I am lastminute.com and I found out that someone has been buying guidebooks about Italy, it might be time to get those Italy hotel ads up on my website. The key thing, though, about this kind of data is that this is data where everyone knows what it is and where it is coming from.

The last kind of data -- and this is the data I am going to be focusing on in this presentation -- is something called third-party data. And this is data purchased from a third-party source with the aim of identifying what we call in marketing a customer segment or a particular kind of customer you might think is interested in your product or service.

Now, the actual purchase of this data is extremely complex and is subject to a lot of different technologies. I am going to simplify the terminology slightly in this presentation and just talk about data brokers. And you can think of data brokers as being analogous to a data aggregator that comes and collects all the different data sources from browsing behavior across the web -- sometimes offline behavior, too -- into a file which summarizes all of the information that is learned about a particular cookie or a particular pair of eyeballs that is browsing the internet.

Now, as you can imagine, these data brokers have a lot of data. And as aggregate data, just getting the pure data in place does not actually help that much. You need to make inferences about who the customer is, and what they might be interested in, if you want to determine what ad to show them. And this paper is going to be all about how good the algorithms are which use this data to try and make inferences about consumers and what kind of ads they might be interested in.

So to give you an example of what I mean about this third-party data, I thought we would start with a specific example, and I am going to show you how to do this with Twitter. Now, why Twitter? Well, simply because it is actually quite straightforward to get access to this kind of data on the Twitter platform about yourself, and also my gut feeling about the audience of the FTC hearing is many, many people have a Twitter account.

So what you should be doing right now is getting out your mobile, if you are not already playing around with it, and follow along to see how you can find out what data Twitter has about you, which is this kind of third-party data, where people or algorithms have made inferences about your profile as a consumer.

So what you do is you get out your Twitter profile and you go and look at settings and privacy. You can see that I have highlighted it right on the left-hand screen right there. And then after that, you go and choose -- you go to the privacy and safety screen and you scroll down to the bottom where you have the opportunity to see your Twitter data.

Then on the next screen, I would like you to select the second option, which is this third-party data, which is all about inferred interests that Twitter has from third parties who have been collecting data about your browsing of the internet.

Now, if you click on this with me, I will show you what I see. So you are going to see a whole lot of different things that this third-party data and the algorithms have inferred about you. This is what they have inferred about me.

Now, here you can see that they think I have one child. I actually apologize to my other three children, I obviously do not browse enough about you. You can also see that my web-browsing patterns has led Twitter has inferred that I am actually a senior in terms of my age range.

I think probably the thing I worry the most about is how it is that these third-party brokers have inferred that I am a single parent. I think, at this point, I really do have to apologize to my poor husband.

Anyway, the key thing here for the purposes of this talk is that you can see demographics, what they have inferred about your demographics, right, because, in general, a pair of eyeballs browsing on a mobile phone or a desktop, there is no real way of sort of telling, you know, exactly what your background demographics are. So the algorithm is then going to use the data about your browsing to try and infer what your demographics are from your browsing, and that is going to be the focus of this study.

Now, the specific name of the paper I am going to be talking about, if you want to read it in detail and, you know, go into all the nitty-gritty, it is up on SSRN and you can easily find it there and it is called "How Effective is Black-Box Digital Consumer Profiling and Audience Delivery?: Evidence from Field Studies."

I should highlight that this is not work I have done by myself. Instead, I have a wonderful team of coauthors. Nico Neumann is at the University of Melbourne Business School, and he is an amazing very junior professor who really cares about this industry and trying to work out what is going on, and Tim Whitfield, who was actually at one of the large advertising agencies at the time we wrote the paper, and he organized for us to get access to this world to study how well it works. So I owe a huge gratitude to my coauthors.

This paper consists of three separate studies, and in all these studies, we are asking, how well does the big data and online advertising ecosystem do in terms of identifying gender and age? Why gender and age? Well, first and most importantly for us, they are things you can actually, potentially verify.

The second reason -- and this is actually a very popular form of data that advertisers use for targeting the -- if you look at it from industry surveys at least, the age data, gender data tend to be most broadly used types of data for the targeting of ads.

Now, the way we proceeded, as I said, there were three studies and in each study, we actually tried to make the task of identifying whether a particular pair of eyeballs was from a certain gender or a certain age easier and easier. The first study was the most broad-brushed, and as such, I will go through it quickly.

And what we did there was we went to various ad platforms and said, can you show our ad 100,000 times to men between the age of 25 and 54? When we gave them this simple mission, there was a large range of success, but we found they were able to do this, on average, about 59 percent of the time when we compared their performance with our benchmark, which was the Nielsen data that actually reported the age and gender of the eyeballs that were seeing our ads.

Now, in some sense, to be clear, this is an improvement relative to sheer chance. Sheer chance would be below a third given the makeup of the internet compilations. There is an improvement of 184 percent when we use the data ecosystem to try and enhance our advertising. I think, though, the point we are trying to make in the paper is, yes, there is definitely an improvement. But given that advertisers tend to be paying more than 200 percent more to show their ads using these data-targeting tools rather than just showing them by chance to everyone, it was not quite clear to us that the return on investment was there.

Now, as our first study -- and you might say this is somewhat unfair because it was still relying a lot on humans to have discretionary choices perhaps about how they set up the campaign, and that could explain the failure we are seeing. So in our next study, we wanted to try and take out that human element.

What we did for the next study was we tried to make it easier for data brokers to do this. So we sort of tried to take out the human element. And so in our second study, what we did is we said we have this website, please, data brokers, tell us who the audience of the website is. So there was no discretion in finding particular eyeballs; you just have to tell us who the eyeballs at a website is.

Now, when we did this, we did this test with four separate big data brokers. On average, what was just striking is that they told us in terms of proportion of men it is 58 percent, it is 55 percent, 85 percent, 63 percent. I am not sure what we can say about accuracy here. It does not seem great to me. If I got back those numbers, I would still not quite know what the true proportion of men is.

What was also striking to me about this study, and we should see it in the paper, is that never mind getting their gender right, they had no idea when we asked people what the actual number of eyeballs was on these websites. At least those huge -- when I say “no idea” what I mean is there is huge variation in the answers we were given, which ranged all the way from 300,000 to 500,000 eyeballs, which is a large difference if you are an advertiser.

So the second study did not give us much reassurance that we were really getting accurate information here. So what we decided to do in our third study was to just make the task as simple as you could ever possibly imagine. And in this task, what we said to each data broker was, look, you do not have to tell us about a particular website. All you have to do is tell us do you have data or a profile about this particular cookie, and if you do, can you tell us what gender you think this cookie and the set of eyeballs associated with this cookie are.

Now, you might be saying, okay, you keep on saying we know really how many -- you know, what gender people are, how do you know the truth, and what we did in this study to find out the truth, which, you know, I find quite compelling, is that we used a service named Pureprofile to actually verify what the truth is. And what Pureprofile goes out to do is they actually survey people, and so they go out and say, what gender are you, what age are you, and they give you a for answering directly. So we used that as our source of truth about what the true gender and true age is.

And you may, of course, be cynical and say, well, are all people in an online survey really going to be completely honest? And, of course, I am sure there are some people who are not honest when asking these surveys. However, it is said to be our source of truth and at least it is what we can call declared data for what people want me to think about their age and gender. So we are going to use that as our measure of the truth. And the question is, what did we find when we compared this declared data to what the data brokers were telling us.

And you can see here we actually used a lot of different data brokers in this study, and there was a wide range of how many cookies they told us they had information for, and you can see that in the second column.

What I want you to look at, though, is the third column. And we actually asked them the specific task of telling us whether or not that cookie was male. And that is going to be our measure of gender accuracy. And the number you see in the third column is the percentage of times they were able to correctly tell us that a cookie was male.

And I want you to look at those numbers and also register the fact which I always found the most hilarious about this study, and this paper in general is that it is if you sort of take the average of accuracy really pretty close to 50 percent -- in other words, these data brokers, this entire big data ecosystem, seem to be able to tell us the gender of the pair of eyeballs correctly half of the time. And if you have ever taken probability theory, and you have thought about the distribution of men and women, you will see why this is quite funny.

The other thing I want you to look at in this table is the second column, which is the number of cookies. The reason I think this is important for this meeting -- I do not usually emphasize it when presenting the paper, but I think it is interesting -- is as a subset you might think of this as a measure of how much data the data broker is really working, right. Because we asked them, well, how many cookies can you tell us about and so it seems reasonable to infer that if they could tell us about more cookies, they have more data.

Now, the reason this is important is that if you were to try and think about a correlation between the second column and the third column, and look to see is there any relationship between the amount of data these data brokers appear to have access to and how good they are at telling the gender correctly, you know, there is not really enough data points to run a regression, but if you stare at it just see no available correlation really whatsoever. So I think it is important because it suggests that there is a surprising lack of correlation between access to data and how well these data brokers are performing in terms of being able to use an algorithm to infer gender.

So let's just summarize the findings of this research. So back in the big headline news -- and this is going to spill over, I'm sure, into the meetings in two weeks time -- is that, in general, we have often worried about algorithms, big data, AI, and we tend to worry, though, more from an Orwellian privacy intrusive way. However, I am here to tell you we might be worried about these algorithms being too accurate, but I am really worried about the fact that they seem to be surprisingly bad at actually getting something very basic like being able to infer gender from browsing behavior.

Now, it seems very straightforward that, when you think about it, maybe there is a reason these algorithms are doing not bad, but poorly. I mean, I challenge everyone in this room to think about the internet sites you browse and really how informative are they about gender? I can imagine that there are perhaps some particular websites which tell you a lot about gender, maybe a website devoted to the merits of sanitary products or something like that. I do not think there are probably many men browsing those types of websites.

But, in general, if you think about the right browsing behavior, it is talking about gender and I think that is just an overarching problem these algorithms are facing. They are trying to infer something which maybe is just not inferable given how different our -- “browsing behavior” given how different genders really perform -- use the -- how people with the same gender use the internet.

The other reason this is going on actually may be even more simple and, you know, this is not a complete explanation, but it is certainly a partial information, but one of the reasons these algorithms appear to be failing is that we looked to see how does that accuracy vary with household size. And we showed that as your household gets bigger and as you have more than one person potentially using a computer or a device, then the accuracy does appear to fall.

So a simple explanation, we are trying to infer gender potentially from a computer, which in my case is used by my husband, used by me, used by my kids to watch My Little Pony videos. It is going to be very hard to actually work out what gender a pair of eyeballs are when you do not have just one pair of eyeballs.

Now, another point I want to make is not just that this kind of data inference process in the use of algorithms on big data does not seem to provide necessarily insights that we might fear it does in terms of how accurate it is, it is just because these are hearings about competition is that you often hear repeated the mantra, the idea that there is a link between access to data and the ability to compete.

And especially in a world of algorithms, you can see the argument for that and that perhaps if I have a larger data set, I can train my algorithm to perform that much better and be able to outcompete my rivals. However, what I saw in this study, at least in this early -- potentially early and nascent stage in this industry -- is that the size of data did not seem to matter that much, or really at all that I could see in the data, of how well these data brokers were doing in terms of accuracy.

And that suggests perhaps an argument which I think we will probably be hearing about in two weeks, that really the quality of algorithms are going to be potentially more important than the quality of [indiscernible] -- these algorithms may end up being more important than the actual size of data that are used to train these algorithms.

So with that, I will say thank you so much for listening. Apologies again for not being in attendance. It does look wonderful. And if you have any questions, feel free to email me. Thank you very, very much.

DR. STIVERS: Thank you, Catherine, in absentia.

(Applause.)

CORPORATE DATA ETHICS: RISK MANAGEMENT FOR THE BIG DATA ECONOMY

DR. STIVERS: All right. Our next speaker is Dennis Hirsch from The Ohio State University Moritz School of Law.

MR. HIRSCH: Commissioners and FTC staff, thank you for inviting me here today and giving me the opportunity to present my research at this hearing. I am going to discuss one of the hearing's principal topics, whether companies can use improved privacy performance for competitive advantage.

To address this topic, I need first to slightly reframe it. The question should not just be whether companies can use improved privacy performance to achieve competitive advantage, but whether they can use more responsible data practices at large to do so, including issues of bias, procedural fairness, and manipulation. Some companies are doing this, and they have a name for this broader project. They call it data ethics.

I am currently leading an Ohio State research project that is studying corporate data ethics, and today I am going to share with you the preliminary findings from this research and I will address four questions. One, what is data ethics?

Two, why are companies engaging in it? Three, how are companies trying to achieve it? And four, what does this mean for the FTC's regulatory authorities?

I was led to this topic a couple of years ago when, at a roundtable discussion, I heard the chief privacy officer for a large company say that her department was debating what was ethical to do with data and what was not ethical to do with it. And this surprised me. She was a chief privacy officer, why wasn't she worrying about compliance with privacy laws?

And when I began to hear about other companies engaged in the same activity, I thought it would be interesting to study this phenomenon. I put together a terrific team of colleagues, faculty colleagues from the schools of business and computer science, philosophy, and sociology, and together, we decided to use three methods to try to address this question, a literature review, expert interviews, and a broad survey of companies that use big data analytics.

So today, I am going to present our preliminary findings, but first I need to make two caveats. One, we have completed the literature review and we are midway through the interviews, but we have not done our survey as of yet. So this truly is preliminary findings. We are still in the midst of this study.

Second, our interviews focus on corporate managers at large companies. So we are not getting a comprehensive view of Corporate America, nor necessarily are we getting a fully objective view. That said, I think we have been getting some valuable information that I will try and share with you today.

So as told to us by those that we interviewed, the story starts with big data analytics and its sister technologies, machine learning and artificial intelligence. Now, it is well-known that these technologies can create many benefits, some of which we have heard about already today. But what the companies told us is that they also produce important risks.

And they identified four types of risks: Risks of privacy violation, such as when Target used predictive analytics to infer from customer purchasing histories whether its female customers were pregnant; risks of bias, as when Amazon recently discovered that the artificial intelligence application it hoped to use to sort through the thousands of resumes that it received was systematically discriminating against women, and Amazon caught that problem and decided not to use that AI application; risks of procedural unfairness as when black-box algorithms, which are not subject to explanation or appeal, are used to inform decisions whether to grant loans or jobs or housing; and risks of exploitation or manipulation such as when Cambridge Analytica used Facebook users' data to infer the psychological types of those users and target them with political ads that they would find hard to resist.

As the companies see it, these potential harms threaten not just the individuals in question, but also the reputation of the companies themselves, and this creates an urgent issue for these companies, which is how to reduce these risks. As one corporate manager put it to us, if data use has much more impact, then you need a governance structure to help manage the impact of that data use to make sure the organization does not create more risk for itself.

Now, traditionally, companies have mitigated digital risk by complying with privacy laws, but -- and this is a key point -- big data analytics renders that insufficient. And it does so for two main reasons. First of all, the risks that I just mentioned start with privacy, but they go well beyond it to bias, procedural unfairness, and manipulation. So privacy law is not going to be sufficient to address that.

Second, privacy law is premised on the idea that given accurate notice, individuals can make choices about what companies can do with their data. So by making such choices, individuals can protect themselves. But big data analytics changes this. It allows companies to take surface data and infer latent information from it. For example, it allows Target to take customer purchasing histories of its female customers and infer whether they are pregnant.

Given this ability to infer latent data from surface information, people cannot know what they are really revealing when they decide to hand over the surface information. And as a result, they cannot use notice and choice to protect themselves, at least when it comes to big data analytics, machine learning and AI. From the company’s perspective, this means that if they are going to protect individuals against the risks that these technologies pose, and so protect their own reputations, they have to do more than comply with privacy law. They have to ensure that their practices are also ethical.

So here is what one lawyer who advises such companies said to us: Preying on vulnerable populations, treating people unfairly, manipulating people in ways that could harm them, there is some of that stuff that is perfectly legal, but it still may not be a good business decision. I will throw out the word "ethics." It is not the ethical thing to do. Some companies that I work with, they take that stuff very, very seriously. They do not want to do things that feel or could be perceived as unethical.

Now, some, including some of our colleagues in Europe, see data ethics as an attempt to take Kantian or Aristotelian or other ethical philosophies and use them to govern advanced data practices. But that is not what we saw these companies doing. For them, data ethics is beyond-compliance risk mitigation for the big data economy. Hence, the title of my talk today.

So that is what data ethics is. Why do companies seek to achieve it when existing law does not require them do so? We identified three principal motivations: Reputation, employee retention, and the threat of regulation. I have already mentioned reputation, but the companies tell a more nuanced story about it. There is reputation among customers and users. This is essential to preserve the bonds of trust, on which the flow of personal data depends. As one company manager said to us, if you act ethically and ensure that data use is ethical and you are fully accountable for that, then your brand is trustworthy. That is what we are all trying to achieve.

Then there is reputation among regulators and advocates, and a poor reputation among these constituencies can lead to increased scrutiny in litigation. And, finally -- and this is the one that surprised us a bit -- there is reputation among your business partners. A lawyer for one technology company saw this as the most important aspect since other businesses are able to do due diligence in ways that individuals cannot, and will not work with companies that do not pass muster.

Employee retention was a third major driver. Tech companies, in particular, expressed that competition for young engineers is fierce, and is critical to corporate success, and that companies need to align their actions with these young people's values in order retain them.

The third driver we saw was the threat of regulation. Some companies believe that if they took proactive steps to act responsibly, they would reduce the chance of direct regulation: data ethics as a way to preempt direct regulation. Others, with an eye on the GDPR and other rules, saw data ethics not so much as a way to avoid data regulation, but as a way to prepare for it. They felt that if they aligned their products and systems in advance, they would be able to deal with such regulations more effectively and at less cost than their competitors.

So with each of these drivers -- reputation, employee retention, threat of regulation -- companies are seeking a form of competitive advantage. And thus, our research suggests that corporate data ethics represents a new form of competition in the algorithmic society, one that goes beyond just competing on privacy attributes. One leading privacy professional put it this way, “I think that for some of these companies, they have actually seen data stewardship as a competitive differentiator, and that they are more trustworthy and people are more likely to do business with them and, therefore, pay higher prices.”

I should add that several interviewees expressed that their company’s values were also very important in driving their data ethics initiatives, and that was particularly true where a CEO or a founder had instilled those values particularly strongly. So that can also be a motivator.

Now, we have looked at the what and the why of data ethics. The next question is the how. Here, it is helpful to divide this into two areas, process and substance. In terms of process, one of the really interesting developments that we found is the transformation of the privacy officer role into a role that included not only privacy, but also issues of bias and procedural unfairness and manipulation.

Reflecting this, some companies changed the title of the position to include the word “ethics” or “data ethics” in it. This is a new development that has just arisen, we think, within the last year. But it could soon be common to have a chief data ethics officer to go along with your CIO, or your CISO, or your CPO.

Another interesting development was the creation of new committees to advise the companies on ethics. Some created internal committees, sometimes called an ethics review committee, to review data analytics projects that raised ethical risks. Such committees could include representatives from legal, privacy, security, engineering, and the affected business unit, and we saw instances in which such a committee advised against certain projects and the companies turned down significant contracts on this basis.

Other companies ran their ethical questions by external committee, sometimes called external advisory boards, that might include privacy and consumer advocates, or members of civil rights and civil liberties groups, or academics. In contrast to the internal boards, these served in purely an advisory role and helped to sensitize the company to stakeholder concerns.

There was quite a bit of variety in the way the companies managed in this area. For example, they differed on the scope of their ethics management activity. Some focused on the company’s own internal research with customers’ personal information; others expanded the scope to include not only their own activities, but also those of data suppliers, customers, and business partners, anyone whose ethical lapses could be linked to them.

I practiced and taught environmental law before I turned to data and privacy and these programs reminded me of the way in which some companies audit the environmental performance of their entire supply chain, a process they call greening the supply chain.

Companies also diverged in terms of

management structures and reporting systems. Some localize the ethics function in a single person who had a direct line of communication to the C-suite or the CEO. Others had a far more elaborate process in which all data projects had to be submitted for review. As we understood it, the first seemed to produce faster decisions; the second, better quality decisions. So there is a tradeoff here.

Turning from process to substance, we sought to identify the standards that companies employ to assess whether a given data analytics project is ethical or not. The literature suggests that companies employ or should employ formal principles grounded in philosophies of ethics. For example, the Software and Information Industry Association that Mark MacCarthy works with -- and he was here today -- they draw on such ethical traditions and published a report that articulated four core principles -- rights, justice, welfare, and virtue -- that companies should follow when making decisions about ethics.

The companies we talked to were not using any such formal framework. What we saw was far more intuitive. One manager referred to the quote “fairness check,” which the manager described as would my mother think this is okay? Would I want this to happen to my kid? Do I feel good about this personally?

Another employs the ear test. Saying the ear test simply means to me, does that sound right? Does that sound like a bad idea? Do the words coming out of your mouth make sense from both a legal, ethical, and business standpoint?

So these companies are using much more intuitive expectation-based standards rather than formal philosophical ones. Such standards fit with the idea, mentioned earlier, that companies are seeking not to implement an ethical philosophy, but rather to engage in beyond-compliance risk mitigation. In this sense, data ethics is a new dimension of corporate social responsibility. It is CSR for the data-driven business.

Responsibility, appropriateness, trustworthiness, fairness, these seem to be the currency of data ethics. Now, these can be difficult concepts to operationalize, and some companies seem to really struggle with drawing these lines. The hardest question seems to be how to get the balance right, how to determine, considering the potential benefits and risks, what is fair and what is not. As one attorney said to us, when do these lines get crossed? That is not always obvious.

What does all this mean for a regulator, like the FTC? Well, when you step back from what we have learned so far, you really see two things. You see a pretty clear consensus among the larger, more sophisticated companies, at least, that it is important to go beyond compliance and seek to mitigate the risks that big data analytics can pose.

So there is quite a bit of agreement on the what and the why. But the how question is much more murky. Companies are experimenting with many management processes and trying to figure out which will be more effective, and there is some confusion as to how to draw the line between responsible and irresponsible behaviors.

I mentioned that I came to privacy from the environmental field. And this situation reminds me in some ways of that which environmental regulators faced when companies started to compete seriously in terms of their environmental performance, which is known as green business. One thing that environmental regulators did, and that the FTC could do, is to collect and share best practices in this area as a way of getting more companies to adopt them.

Another would be to adopt a leadership

program that recognizes companies that are going above and beyond in this area and so add to the reputational value they derive from doing so.

A third would be to define some standards in this area. Now, I would caution against doing this with respect to process. There seems to be a lot of positive experimentation going on, and regulators may want to let that play out before determining that one approach is preferable to another, but it may be worth giving this further thought with respect to drawing the substantive lines.

Were a regulator to provide some guidance, that could give companies a clearer sense of what the regulator's expectations are and help them to make some of the tough calls. It could also set a floor that all companies have to pay attention to. Right now, we are seeing the larger, more sophisticated companies start to manage data ethics. But other companies that are not paying attention to these issues could do some really bad things that could not only hurt people, but could also turn the public against data analytics, machine learning and AI more generally.

As one attorney said to us, "In this fast-paced world where there is, you know, huge financial opportunity for companies, you can easily see scenarios where someone is going to, quite frankly, bring down the whole house of cards by doing something just totally unethical and totally unfair and screw it for the rest of the industry." Seen in this light, a regulator's decision to set a floor for fair and ethical behavior could potentially support the efforts of the current leaders while still giving them room to distinguish themselves.

If the FTC wanted to develop such substantive guidelines, rules of the road for predictive analytics, it seems to me that it has the power to do so. The line that companies are trying to draw is between advanced analytics that is appropriate, and that which is unappropriate, between that which is responsible and that which is irresponsible, between that which is fair and that which is unfair.

Section 5 of the FTC Act, of course, gives the Commission the power to define unfair business acts or practices and so to draw these lines. The FTC's unfairness authority has some useful features in this regard. Unfairness is an open and flexible standard intended to adapt to emerging and changing technologies and business models. It requires the

Commission to balance benefits and costs, which is important in an area like big data analytics that offers many benefits, as well as many risks. And unfairness is intended precisely for those situations in which individuals cannot protect themselves, where in the language of Section 5(n), the injuries are not reasonably avoidable by the consumers themselves.

That is where we are with respect to advanced analytics. Individuals cannot understand how these technologies work and so cannot use traditional privacy protections, notice and choice, to protect themselves. Some companies are moving proactively to protect them. The FTC could, potentially, use its own unfairness authority to support this corporate data ethics effort.

Thank you for letting me share my thoughts and research with you today.

(Applause.)

FREE SPEECH AND DATA PRIVACY

DR. STIVERS: Next, we have Jane Bambauer from Arizona, Rogers College of Law. Here we go. Thank you.

MS. BAMBAUER: Thanks for having me. I want to do something that will give me a little information and give you a chance to wake up. So please stand up if you can, if you are willing and able, and then if you are not a lawyer, sit back down. I just want to get a sense of what this audience -- if you are not a lawyer -- okay, so that is about half of you.

And then, for those who are still standing, if you do not know what the case *Sorrell vs. IMS* is about, sit down. Oh, good. Okay, this is going to be valuable. Thank you.

Okay, all right. So my goal with this talk is to provide a descriptive account of what is going on in First Amendment law and the ways that it might actually limit some of what the FTC may want to do, even if it has set its sights on what it thinks is the best policy. I am going to be as descriptive as I can without letting my own policy preferences kind of shape that description because I will have a chance tomorrow to talk again about what I think about best policies. So I am going to do my best just to be sort of honest about what is going on in the courts. But the courts are being quite active in this sphere and you all should pay attention.

Okay. So before I dive in, though, let me tell you a little bit about what I am not going to talk about. I won’t talk about restrictions on commercial speech. The commercial speech doctrine, it is a little bit misleading. It is actually a narrow category that covers just marketing messages and advertising. So some people sometimes think that commercial speech, which gives lesser protection to commercial -- it is a doctrine that gives lesser protection, that that would apply any time someone is selling something to someone else, or any time that somebody has a commercially-motivated purpose to say something, and that is not accurate. Like books are sold, right, and obviously books receive full protection.

So commercial speech doctrine is narrow and it is related to potential privacy regulation because privacy laws often have as at least one of their end goals to affect how marketers can craft messages. So in that sense, it may be related.

Also, if companies are giving false assurances, either explicitly or implicitly, to consumers, that their privacy is protected when it is really not, that is related to the commercial speech doctrine. The commercial speech doctrine does not allow any protection, any First Amendment protection to false and misleading commercial claims.

And that is interesting in what it means to be misleading, especially when a company is committing an omission, when they are not saying anything. It is still an open question about whether that is sufficiently misleading to remove the company from the ambit of First Amendment protection, and there are some papers, including one that I have written recently, that kind of have tackled this question of who gets to decide what it is to be false or misleading -- and by the way, I am sorry, I mangled Rebecca Tushnet's article. Hers is actually called "It Depends on What the Meaning of False Is." I was writing these slides late.

But we engage in a little bit of a debate about who should decide and whether the courts need to be involved. And it is interesting, but I am not going to talk about it today.

The other interesting thing that is out of scope for today is the compelled speech doctrine. So that is related to privacy because regulators might be interested in something like just-in-time privacy disclosures that make clear notice about how a company is going it use your data, and we may be interested in forcing companies to actually provide these disclosures.

And the Supreme Court has given its blessing to mandated disclosures that are purely factual and uncontroversial information, so maybe like nutrition labels on food items. I think most people tend to think of that as purely factual. But it is not clear whether a privacy policy would be -- or mandated privacy policies would be purely factual and uncontroversial. And I talk about this at some length in another article. So if you are interested in this topic, you can see my article that tries to map out what courts, especially lower courts, have done to decide whether a factual mandated disclosure is ideological and, therefore, subject to constitutional review, or merely informational and not subject to any amount of review.

Interesting stuff, but we do not have time because I want to get straight to the core of what almost every privacy law is going to wind up potentially coming into conflict with, and that is a restriction on noncommercial speech.

So this usually will happen in the course of privacy regulation through one of two ways. Either a law will put a limit on the transfer of personal data between, say, one company and another, or it will put a limit on the initial collection or maybe even the initial inference based on already-collected data about a person. And, you know, almost every privacy law, if you think about the FIPPs, the Fair Information Practice Principles, they usually involve giving the data subject some amount of control over these two activities. And that control necessarily puts a limit on these data transfers or the data collection.

So much of what I am going to say, but not all of it, is lifted from an earlier article I did called “Is Data Speech?” asking, well, okay, is the First Amendment relevant here? Do we need to worry about potential constitutional review when we are dealing with data privacy?

So let’s start with the -- oh, that is right. To ground the discussion, I would like to have you, in the back of your head, thinking about the California Consumer Privacy Act because I think that -- for many consumers, that seems to be a model privacy law. It seems to tap into what many people want or at least believe that they want.

And the most important rights that are relevant for my discussion is that it gives Californians the right to say no -- this is taken from the website of the designers of the law -- it gives Californians the right to say no to the sale of personal information. It also, by the way, gives them the right to demand the deletion of personal data unless it is required for the service of the company.

And just like with the GDPR, if a Californian does opt out of data sale, for example, they still must be given service on the same terms as somebody who has not opted out. But unlike the GDPR, it is an opt-out regime rather than an opt-in regime.

Okay. So as I tell you about some of the case law, work with this hypo -- law professors love hypos, so ask yourself, okay, how does this affect the constitutionality of California's recently adopted, but not yet implemented, law?

Okay. I am going to start with data transmissions. These little stick figures are meant to be like companies or people who are selling data, and that red thing is data.

So the first question that free speech lawyers generally ask is, well, is the First Amendment even relevant here? Does it cover this kind of activity? Would we call this activity speech? And I am starting with this rather than the initial data collection, even though it usually comes later because I think this question is actually much easier to answer. I think courts are converging on a clear, yes, this is speech, this is covered.

So the Supreme Court itself in earlier cases had found that really dry information, like credit reports or beer ingredients, are speech, and really anything that communicates from one person or entity to another is speech. The lower courts, too, found that even in the context of privacy laws that the privacy laws may survive scrutiny, but that scrutiny must be used.

Then the case of *Sorrell vs. IMS*, which most of you do not know about which delights me because I can tell you about it, really made this even more clear. So this was a case from 2011 or 2012 involving a Vermont statute that banned the sale of prescription data to pharmaceutical companies if the pharmaceutical company was going to use the data to fine-tune the detailing, basically the marketing messages that it made for doctors. So the data did not have the identities of the patients, but the data does have identities of the doctors. So you can see it was justified partly on privacy grounds and partly on public health grounds.

And as a privacy law, this seems rather narrow, but you can see how the implications might affect other types of broader privacy laws because if you think of doctors as standing in for consumers here, the law was trying to give doctors the opportunity -- they could opt-in to these types of marketing messages based on their data if they wanted to, but it was trying to give them some control such that behavioral advertisers basically would not have a lot of detail about their habits.

So the Supreme Court -- by the way, some commenters and even the circuit courts that were hearing similar cases before *Sorrell* was decided thought this type of law would fall outside the First Amendment protection completely because data that is just like sitting in a server and that is just sold for these types of purposes is no different from any other product. I think the First Circuit even said it is like the equivalent of beef jerky -- selling beef jerky. The Supreme Court definitely rejected that.

So in an opinion by Justice Kennedy, he begins the analysis by saying this Court has held that the creation and dissemination of information are speech within the meaning of the First Amendment. Facts, after all, are the beginning point of much of the speech that is essential to advance human knowledge and to conduct human affairs.

In the end, it got a little confusing because the case was ultimately decided on grounds of viewpoint discrimination, because what at least Justice Kennedy thought was the most -- the biggest offense about this law was that it prevented only pharmaceutical companies from using this type of tool to craft their messages to try to persuade doctors to do something, and it left open any other speaker who was trying to persuade a doctor to do anything else, it left access to the data open to them.

So the case was ultimately decided on viewpoint discrimination grounds but the dicta that came earlier seems pretty compelling, and especially because it is consistent with what the Supreme Court has said or at least assumed in the past, that if something communicates, it is speech unless it is in some very narrow special category like fraud, defamation, and a few others, incitement.

Okay. Well, all right, so data privacy law might have to undergo scrutiny or probably will have to undergo scrutiny. What level of scrutiny is going to apply? This question is much harder to answer. So a case called *Dun & Bradstreet vs. Greenmoss Builders* involved a credit report, a credit report that was wrong importantly. And in a defamation action, the Supreme Court decided that only intermediate scrutiny, you know, a lower level of protection applies in this defamation case because credit reports that are given to just a couple potential lenders are matters of purely private concern.

So you can see a line with this case developing, emerging, that separates speech of public concern or general concern from speech of purely private concern. *Dun & Bradstreet* could have been limited to just defamation cases, but it has not been limited to that. So the Supreme Court itself has cited to *Dun & Bradstreet* in cases that have nothing to do with privacy for the proposition that speech of purely private concern is not nearly as protected.

So you might think, okay, well, then privacy laws are going to have to only undergo intermediate scrutiny, but more recently, in *Reed vs. Town of Gilbert*, the Supreme Court decided that strict scrutiny must apply to any law that, on its face, makes a distinction of any sort based on the content of that communication.

And if you think about the California Consumer Privacy Act or many of the regulations that the FTC, in the past at least, has considered or that is included in the GDPR, the linchpin for regulation is personal information and it is defined in certain ways and that is all about the content of the data. So if *Reed* is applied faithfully, it is not clear that courts will be able to do this. But if we are serious about *Reed*, then it looks like strict scrutiny would apply. At this point, I do not have a confident prediction about which level of scrutiny would apply.

But going back to *Sorrell* for a minute, in the end, when the First Amendment is applied to some sort of privacy law, it is possible that courts could distinguish cases like *Sorrell* because even in the opinion itself Justice Kennedy said, well, perhaps the state could have addressed physician confidentiality or privacy through a more coherent policy.

Now, some might object to the idea that the Vermont law was incoherent because it was targeting kind of the most obnoxious form of data sale, then maybe the legislature was right to just pinpoint that particular form of data sale and leave all others, you know, untampered. But if we take this seriously, then perhaps something like the California statute is more likely to survive because it is broad, because it is so comprehensive.

I have some doubts, though, rather that there are at least a few reasons to think that the Government would have to prepare strong arguments and a good base of evidence in order to defend even a broad privacy law that prohibits the transmission of data. For one thing, just in the past, even since the 1960s, the Supreme Court has listened to cases that involve the clash between privacy and the First Amendment, usually in the content of some sort of magazine publication, and has found that the privacy interests are not compelling enough to outweigh the general interest in speech.

The other thing, though -- I am going to skip this for a second in the interest of time. The other thing is there has been a series of Supreme Court cases, none of them directly related to privacy, but each of them showing that the Supreme Court is extremely skeptical now of any attempt by the Government to justify what it is doing based on just kind of common sense ideas of harms or risks.

So *Brown vs. Entertainment Merchants Association*, for example, was a case that involved a

California ban on the sale of violent video games to minors unless the minors had their parents' consent, and the Supreme Court found that the law was unconstitutional, even though the state brought a mountain of social science evidence with it because the Supreme Court -- rightly in my view, but, you know, obviously reasonable minds can differ -- but the Supreme Court thought that social science evidence was actually quite bad. It was poorly done.

So the courts are showing an increasing willingness to even look at the -- probe the quality of the evidence that the Government has and offers in order to justify their restrictions on speech.

Let me spend just a minute talking about the data collection side of things. So for a long time -- so this guy is using his cell phone, I guess, to record someone. So for a long time, the assumption was data collection is not protected by the First Amendment, even though subsequent publication of that information would be.

So in a case called *Deitemann vs. Time*, the Ninth Circuit decided that Time Magazine -- they snuck a couple journalists into a quack's office, like a guy who just was waving wands and turning knobs and pretending to cure diseases, and they did an exposé on him and the Court found that the actual publication was fully protected by the First Amendment. They could not be sued for public disclosure of private facts, irrelevant tort. But the sneaking in of technology to record -- to surreptitiously record what was happening, the secret photographs, that was completely unprotected, the Court said.

And the Supreme Court, in a case, *Bartnicki vs. Vopper*, said something similar. That downstream publications are protected, but actually getting access to information or knowledge is unprotected conduct. It is just conduct; it is not speech. That always seemed weird to me because if you think about the reason to limit data collection, it usually has something to do either with knowledge creation by the person who is collecting the data or with downstream communications that that person intends to have.

And so if we think of both knowledge and communicating as being core to the First Amendment's goals then why should limitations on collecting information in the first place get a free pass and not get any scrutiny at all?

Well, sure enough, in the last couple years -- this is a really recent development, but there have been right-to-record cases that are starting to chip away at this distinction. The first set of cases have to do with recording the police in public. And, now, every circuit that has heard these types of cases has decided that there is a First Amendment right to record police. The Seventh Circuit has gone further and said there is a right to record any time you are in public.

And then, also, there have been successful First Amendment challenges to so-called ag-gag laws that prohibit people from secretly recording at commercial farms. And so that, too, is suggesting that even surreptitious recording, even in private spaces, has been getting increasing First Amendment attention.

All right. So I raise all of these legal limits not to discourage the FTC in any way from crafting responsible privacy policy, but rather in a way to applaud you for doing these types of hearings because it is tempting to do something like what I think the FDA had done, regrettably, a few years ago and to just kind of plan to defend your policy later in court. But it will save you a lot of headache and heartache if you have a good evidence base and a good theory of what type of interest and seclusion or confidentiality you are actually trying to preserve in order to come prepared for a First Amendment defense.

The other option, of course -- and this has come up already -- is to actually prohibit disfavored uses that really are conduct rather than speech. So that is an option as well and then you do not have to defend against the First Amendment at all.

All right, thank you very much.

(Applause.)

FTC EXPERIENCE WITH DATA MARKETS

DR. STIVERS: All right. So we have Haidee Schwartz as our next speaker.

MS. SCHWARTZ: So I am Haidee Schwartz. I am the Acting Deputy Director of the Bureau of Competition at the FTC. First, a disclaimer, these remarks are my own. They are not those of any particular Commissioner or the Commission as a whole. I also want to say that I am looking at this from the Competition side of the FTC. I believe my colleagues in Consumer Protection are talking on other panels. So this is from the Competition side.

When people talk about data, they usually talk about the four Vs of big data: Volume, how much data are we talking about; velocity, how much data is coming through and how quickly, and for us that means how much does it have to be updated and what is the flow; variety, what are the different forms of data and are they complements or substitutes; and veracity, how accurate or inaccurate is the data?

In the FTC context, when we think about data markets, the four Vs are implicitly part of our considerations. And we look at how big data is being used. Is it a product, is it an input, is it a tool? Often, it is two or three of these things. And, of course, we look at whether the data is unique or broadly available. This is particularly important because it affects entry and expansion options of other firms in the market.

So how do these cases often look to us and how do they come to us? In the instance of two database companies merging, they often sell data products. And two of the older examples that I have are these type of cases where it is two merging databases. If we go back to 2001, the FTC challenged the consummated merger of Heart Trust and First DataBank. That involved the merger of two competing providers of integral drug data files.

Then if you go forward to 2010, the FTC challenged Dun & Bradstreet’s acquisition of QED, which was a division of Scholastic that involved K through 12 educational marketing data, such as contact, demographic, and other key information related to teachers, administrators of schools, and school districts. So if you look back, you know, we have a long history of where the database is the product and we are challenge those mergers.

In some of the more recent cases I will discuss in this presentation, data was a key input. It wasn’t the actual product itself, but it was integral and essential to the product. In many cases we will look at, data is also being used as a tool and it can be a tool and a product, a tool and a key input. And in cases involving data markets, we will look at how the data is being used and whether it is a key differentiator as well as other key dynamics.

In these data cases, entry conditions are often critical. What other firms, if any, could replicate the competition lost in their relevant market discussing how data may facilitate or create impediments to that entry.

As I have alluded to, the FTC has a long history of cases involving data markets. The history goes back to at least 1996 when the FTC filed administrative complaints again ADP’s 1995 acquisition of AutoInfo’s assets, charging that the acquisition would raise prices and reduce the quality of service and innovation to the automobile salvage yard information management industry. So these are key tools that the automobile salvage yard used and as well as insurers used. The parties each maintained interchanges which were essentially databases of numbering systems for autoparts and parts assembled that insurers and salvage yard use as sort of an index to determine interchangeability of parts.

The parties also had significant software assets, an electronic communication system that allowed auto salvage yards to actually buy the parts and see automatically and quickly, sort of through a central database, what the inventory was at the other yards that subscribed.

In the end, the case settled with the divestiture of the former AutoInfo's assets as an ongoing business, which included granting the acquirer an unrestricted license to the interchange, which, by that time, had become sort of the default industry standard for a cross-numbering index for parts.

Moving on to 2014, CoreLogic and DataQuick, data as a product. This was a merger the FTC challenged in March. In March 2014, CoreLogic agreed to settle FTC charges that its acquisition of DataQuick would likely substantially lessen competition in the market for national assessor and recorder bulk data.

So what is national assessor and recorder bulk data? It is current and historical data on properties pulled from local public records, like deeds, mortgages, et cetera, that is aggregated and standardized in bulk format that includes information about ownership value and other characteristics of properties. So it is public information, but it is not standardized, it is not easy to collect, and you need both historical and going forward. Customers of this data, so customers of the companies, use the data in various propriety programs for risk and fraud management tools, valuation models, and a lot of other uses.

The complaint alleged that the merger would eliminate one of the three providers of national assessor and recorder bulk data, increasing the risk of coordination between the remaining two firms and the risk that CoreLogic could unilaterally raise prices.

In terms of market structure, there were regional assessor and recorder bulk data firms, but the Commission looked at that and saw that they could not combine or reposition to actually compete in the in the national assessor and recorder bulk data market. They would have gaps, they would not be standardized, and there were other issues there.

At the time of the merger, CoreLogic licensed its current and go-forward data to DataQuick, which DataQuick was permitted to relicense in bulk. So in other word, DataQuick was actually kind of dependent on CoreLogic for the data. DataQuick used the license data, along with its own historical data, to compete head-to-head with CoreLogic.

Importantly, DataQuick was unique in its ability to credibly threaten to enter because it already had historical data. It had acquired a company years before CoreLogic was willing to license to them. Because it had acquired that historical data, CoreLogic viewed it as a potential entrant and, therefore, it sort of got economies of scale and scope by licensing to DataQuick, and it felt that DataQuick would be in there anyway if it did not because it had the historical data. It could have amassed the sort of ongoing data itself. So it was willing to license years ago to DataQuick after it had acquired an historical database.

That said, it was very unlikely that anyone else could enter because the breadth of historical data they would need to be gathered and the ability to continue gathering that data would be prohibitive. So no one else was going to have that unique ability to have the historical data.

The remedy that we constructed was designed to allow a company called RealtyTrac to step into the shoes of DataQuick as CoreLogic’s license. The order required CoreLogic to irrevocably license to

RealtyTrac equivalent data to what DataQuick offered to its customers and update the bulk data for five years. That was then designed -- the five years were designed for RealtyTrac to compete with CoreLogic while developing its own ability to collect national bulk data.

As we implemented this, RealtyTrac realized that CoreLogic was not providing the entire data set that DataQuick had access to and raised concerns that led to a Commission investigation. Just recently, in March of 2018, the Commission modified the order after finding that CoreLogic had not provided RealtyTrac with all the required data on a timely basis. The modification adds three years to the original term of the order and specifically spells out the quality, service levels, and data transfer requirements.

Takeaways from the CoreLogic/DataQuick merger. Here, the data was the product being sold and the breadth, detail, and the complexity of the data created barriers to entry. This matter highlighted the complexities involved in attempting to remedy a lessening of competition when data is the product. You would think it is a database, it is not that hard to transfer, but, here, the buyer's due diligence may not -- what we learned is the buyer's due diligence may not necessarily uncover missing or unnecessary data in a timely fashion, and the Commission had difficulty initially identifying the exact universe of data required to effectively compete and required additional work by the buyer, the monitor, and the Commission to determine what data was missing, how it needed to be delivered, and how it needed to be continuously updated.

Verisk/EagleView, data as an input. So here, it was not -- in 2014, the Commission issued an administrative complaint seeking to block Verisk's proposed acquisition of EagleView in the growing market for rooftop aerial services. A Verisk subsidiary competed with EagleView to provide software that, when combined with the library of aerial images of rooftops, allowed insurance adjustors to effectively and efficiently and safely measure roofs.

As you can imagine, the old-fashioned way they used to do it was adjustors would actually get up -- well, used to get up on the roofs and do the measurements. That has issues with both accuracy and also significant safety issues. The measurements, in turn, allowed insurers to estimate the cost of repair or replacement of insured roofs. Verisk also owns the software that customers used to make other measurements to estimate damage claims.

The Commission alleged a product market of rooftop aerial measurement products, or RAMPs, for insurance purposes. Interestingly, for insurance purposes is key, in terms of if it was actually a targeted customer market because the product was used both by insurers and by adjusters and contractors, but for insurance -- although the software products, you know, functioned somewhat differently, both required the same input, the aerial images, and to carry out the same functions.

That said, insurance companies -- the Commission judged that insurance companies had different needs and requirements than other customers, like the contractors. You know, the contractors generally felt that they could switch to manual measurements. Insurers could not. As I noted, the product here is not the data itself, but the data was a key input to the product.

In terms of the market structure, the merger of these two were judged to create a virtual monopoly. EagleView was the first to develop software using aerial images, and these are actually particular types of aerial images. It is not just any old aerial image. It had to have certain angles, certain types of -- certain types of views, i.e., treeless, leafless. In another few weeks, this will be a good time of year to have aerial images because you can actually see the roof and the measurements. It is not a particular pretty photo, but it does make a difference in terms of aerial photos.

And at the time, EagleView had the first mover advantage, amassing a market share of 90 percent. It also had, by far, the largest aerial image library. Verisk was a relatively new entrant, entering just two years before the proposed acquisition. But it quickly amassed, you know, a not insignificant market share, substantially more than any other competitor, and it was offering discounts and direct competition to EagleView. The Commission alleged that if the transaction was consummated, discounts would disappear and prices would rise.

An important aspect of Verisk and EagleView's competition is their ability to obtain the aerial images that are up-to-date, so the measurements reflected those of current structures, high quality because it allowed adjustors to identify attributes of the insured property, and it also had to be available on a national scale. National insurers wanted to be able to use the software for all of their insured products and it was not worth it to them to sort of have different providers in different areas of the country.

EagleView, as I said, had the most extensive library of aerial images. Importantly, insurers also required the RAMPs integrate seamlessly with claims estimation software, and because Verisk was the leading provider of claims estimation software generally, it was able to overcome and was uniquely positioned to be able to overcome a more limited library of aerial images by capitalizing on its relationship with the insurers and the fact that it had the best software and most sort of commonly-used software.

Verisk and EagleView abandoned the transaction after the Commission issued the complaint. So the case was never considered by a court. But what the Commission considered in the complaint provides us with some insights. In this case, while data was not the product defined in the product market, it was an essential input into the product, and affected a firm's ability to compete and enter the market. The Commission considered the incentives to increase the quality and volume of data as a loss of innovation. So that was also an issue.

Now, I am going to talk a little bit about CCC/Mitchell, which was a challenged merger in 2009 that the FTC challenged. Full disclosure, I actually was in private practice at the time and was working on behalf of Mitchell, but I am basing this entirely on public information.

So access to data as an entry barrier. It was a key input, not the actual product itself. There were two products at issue. One was Estimatics, which is a database used to generate repair estimates for automobiles, this was not particularly the product used for sort of specialized trucks or other things like that, and total loss valuation systems, which were used to determine when a vehicle was totaled and, even more importantly, the value of it.

At the time of the merger, the big three, which were CCC, Audatex and Mitchell, in that order, had about 99 percent of the estimatics market. There were two fringe competitors. Most importantly, that we will talk about later is Web-Est. And for TLV, the big three accounted for 90 percent of the market. Mitchell had entered later and had a significantly smaller share.

There were two types of customers. Insurers and repair facilities for estimatics and primarily insurers for TLV.

Okay, database dynamics. So the primary components of estimatics and TLV were the databases themselves and the software. So how did the firms get the databases? CCC had obtained an exclusive license to the Hearst Business Publishing database called "Motor" in 1998. Audatex and Mitchell each had sort of grown their own system painstakingly over years, and part of the reason why Mitchell was smaller is it had taken them many years to create their own database, and they did so.

Web-Est licensed Mitchell's database, but under very restrictive conditions. It was not allowed to sell to any of the top 50 insurers, it could not have a communicating product, which meant that basically it could only sell to independent repair stations, not those that were part of a particular repair network, and it could not integrate with other third-party apps, you know, vendors, things that other insurers and other service stations used.

So the proposed fix, CCC offered to do two thing in terms of making a database available. One, it offered to relinquish its exclusive rights to the Hearst Motor database. That meant that any new entrant could license that database. And it was fully updated and would continue to be fully updated because Hearst kept that database updated and it was licensed.

And Mitchell would remove restrictions on Web-Est and continue that database license. So, therefore, there would be both Web-Est, with the Mitchell database, and CCC offering to sort of relinquish its exclusive -- anyone else could have access to the Hearst database. Audatex would continue with its proprietary-owned database.

The judge found the availability of databases would reduce the most critical barrier to entry, but she still found that there was significant other barriers. One, customers were sticky, particularly the insurance customers that were critical to success, and you needed to establish a track record and have a lot of sort of support capabilities. Scale mattered.

The judge did note that the Web-Est, which was led by a guy named Eric Seidel, had been in the industry for a while, you know, had good experience, had significantly grown his adjustable market share, which had been really independent service stations, but he only had 10 to 15 employees, and so the sort of growth curve was going to be too long and to steep.

It just would not be sufficient entry in the time required. By comparison to Web-Est, 10 to 15 employees, CCC/Mitchell, after they combined, would have had about 2,000 employees.

Interestingly, the judge actually decided this case as a coordinated effects case, and not as a unilateral effects case. She had found some issues with the FTC's expert's unilateral effects analysis. So it was a PI hearing, not a full trial on the merits, but she decided that the coordinated effects were too likely.

Okay. Microsoft and LinkedIn, and this is the last case I am going to discuss before talking about a few takeaways. So Microsoft is obviously strong in operating systems for personal computers and productivity software. LinkedIn is a professional social network that a lot of us probably use. The U.S. investigated, but did not take any action. The EC concluded that the merger did not raise competitive concerns related to data, but it did find -- so what it found -- what it looked at -- and these are some of the answers to questions that we often ask -- you know, is the data readily available from other sources or similar data. And, yes, we found that other -- you know, the EC found that other sources existed for that data.

They also found that the companies had not particularly provided that data and made it available on the market before. So there was not really going to be a change post-merger. And they had relatively low shares in the market that the EC was concerned about.

The EC did require several commitments, and those are just up there. Those primarily had to do with interoperability and ensuring that others could be competitive on the professional social networks. You can see those there. I am not going to read through them. But they did not have to do with particularly the data possessed by the companies.

Okay, takeaways, and I am going to try and end early. So takeaways, competition analysis, because I am sure you guys have had a long day and I appreciate you all staying. Current antitrust analysis accounts for how firms compete using data. Data markets and sets are highly differentiated. Each investigation looks very closely at the specific facts of the case. We recognize that data markets are dynamic. Quality and innovation effects may be particularly important. They also may be harder to measure than price effects. How data enables or hinders entry or expansion also may be particularly important.

Remedies. In cases that involved data, just like in any other cases, we have a preference for structural remedies. We have seen a number of cases that I have discussed where we look to divest or clone a database versus a license. Sometimes we will allow a license. It depends on the specific facts of the case. There are issues related to how they are going to continue to obtain the data and keep a new data flow that is accurate and is expansive.

What we found in our database cases and what we have learned is there is a lot of complexity to how the data is stored, how it is updated, how it is kept and how it is provided to customers. And it seems simple, but there is actually more due diligence that needs to be done not just by buyers of potential assets, but by the Commission and others during that process.

There are often IP and copyright issues, and while they are not favored, behavioral conditions may be needed. In some of the cases that I talked about, for example, CoreLogic, there were commitments that we required related to allowing customers to break contracts so that the new firm could have contracts going forward over a certain period of time. So sometimes we have to overcome customer's reluctance and, in some cases, ability to switch. We need to give them the ability to switch, to have the new entrant actually be able to have those customers.

There are other types of behavioral conditions, including some support over transition period that we will look at as well. But as noted, structural is always preferred, including in data cases.

Thank you, guys.

(Applause.)

DR. GILMAN: Thanks, Haidee, and thanks all for coming today. We hope we will see many of you back tomorrow morning at 9:00 a.m., and as I think we announced on our website, ultimately there will be a transcript available for these proceedings, as well as the archive webcast. So thanks to all our participants and thanks to all in attendance.

(Hearing adjourned.)

CERTIFICATE OF REPORTER

I, Linda Metcalf, do hereby certify that the foregoing proceedings were digitally recorded by me and reduced to typewriting under my supervision; that I am neither counsel for, related to, nor employed by any of the parties to the action in which these proceedings were transcribed; that I am not a relative or employee of any attorney or counsel employed by the parties hereto, not financially or otherwise interested in the outcome in the action.

LINDA METCALF, CER

Court Reporter