FTC PrivacyCon
January 14, 2016
Segment 3
Transcript

-Next, we're pleased to have commissioner Julie Brill provide a few remarks. Commissioner Brill has long been a zealous advocate for consumers, their privacy, and the security of their data. And we're thrilled to have her here today. Commissioner Brill.

JULIE BRILL: Thank you, Kristen. And thank you everybody who's here as well as all of you out in TV land. Lunch may be over, but the feast of scholarship will continue. It's really my pleasure to open in the afternoon with a few remarks about the research that's on display here at Privacy Con.

But before I do that, I really have to take a moment to do exactly what Chairwoman Ramirez did. And that is to thank the FTC staff who worked incredibly hard and incredibly well to pull this together. Kristen, Dan, Justin, Monesha. I know I'm leaving out probably 25 people, but they all did a really, really wonderful job. So can we just have round of applause for these fabulous people? Great job.

OK, now aside from the quality of projects and presentations, one thing that struck me about today's agenda is that, instead of being organized by discipline-- no computer science here, economists over there-- the day is organized around the key substantive issues in consumer privacy. This thoughtful organization is leading us toward something we need for sound privacy policy development, a cross-disciplinary, richly detailed picture of consumers and how they make decisions about technology use. Lurking behind the main regulatory approaches to privacy, whether it's notice and choice, informational self-determination, or a use-based model, are questions about individual consumers, their goals in exercising their privacy rights, and their ability to do so in the environment around them.

At a high level, I think two principles should guide policy and practice. First, individuals have to be in the loop regarding decisions about what data is collected about them and how it is used. Outside the privacy sphere, companies have excelled at helping consumers manage and use highly complex systems.

We heard a little bit about Chipotle and the burritos. I actually think a much better analogy in this space would be cars. Cars are now computers on wheels. But we can all drive them, because companies have kept the complexity behind user interfaces that are simple to use. I think companies can do the same for privacy, but building the right tools depends on understanding which decisions are important to individuals.

Second, I'm wary of solutions that depend too heavily on any one technical measure. Now just as an example, it's a positive development that companies are offering more services that allow individuals to encrypt their communications, and these services are getting more user friendly. But their ease of use is limited to communications that stay within one particular service. If you want to communicate between services, you may be forced to use tools that only a few select experts can really implement properly at this time.

But these principles leave many questions open and details unspecified. What data do consumers expect companies to collect from them? How do they expect companies to use this data? What do consumers understand about what actually happens to their data? Which aspects of data processing should be under consumers' control? And how effective are the tools that companies offer to consumers to exercise this control? Answering these questions requires a three dimensional approach, so I was excited to hear this morning from researchers who are using structured surveys, qualitative interviews, and looking at human computer interactions to map out what consumers understand about the data practices of the services and devices they use.

Of course, it is just as important to understand more about what happens behind the scenes, outside the view of consumers. Data and device security are incredibly important to consumers, yet assessing security remains well beyond the capabilities of most consumers, Including most of us, but not all of us, in this room.

So I'm thrilled to see researchers doing a deep dive on security vulnerabilities on specific internet of things devices, while others are analyzing data from thousands of vulnerability reports to better understand the kinds of incentives that will spur a virtuous cycle of discovery, reporting, and patching. Also beyond consumers' purview lie the big data analytics that have developed more quickly than have frameworks for specific concrete guidance on legal and ethical issues. Our big data report issued last week is intended as our first step towards providing such guidance. The report recommends that companies review their data sets and algorithms to determine whether they may be having unintended effects, such as treating certain populations disparately and in ways that may potentially violate the law. Our report also recommends that companies bring a broad set of fairness and ethical considerations into their use of big data analytics. The presentations in the next segment of Privacy Con address exactly those issues.

Finally, I want give a shoutout to the institutions that have helped produce the specific pieces of research that we're hearing about today. They are just as important as the research itself. Much of the research presented today comes from universities that have made substantial, long term commitments to examining the relationships among law, technology, and public policy. In addition to generating new research that also contains policy insights, these universities helped train students to become leaders in their fields.

Technology focused centers and clinics have sprouted up at law schools all over the country in the last decade. They expose law students to technology and, probably just as importantly, to the way technologists think. Departments, schools, and even entire campuses that make interdisciplinary work a core mission are doing much the same for students of computer science, engineering, economics, public policy, and social sciences.

Building these programs has not been easy. It's often easier to stick closer to traditional disciplinary lines. So let me offer a word of encouragement. Privacy Con is just one example of the impact that scientists, lawyers, and others can have when they're trained to do groundbreaking research as well as to identify and analyze public policy questions and issues.

This combination of research capability and capacity for action also describes, just coincidentally, the design of the FTC itself. So naturally, we are a ready audience for research the sheds light on the challenges we confront in enforcement and policy development. And I hope the institutions that many of our presenters call home will be lasting platforms for robust exchange of ideas with the public and private sectors for many years to come.

So with that, let's hear what you have. Thank you very much.

[APPLAUSE]

And Dan-- Dan will introduce the next panelist.

DAN SALSBURG: Thank you, Commissioner Brill. Could be the next panel come on up?

Our first session today really looked at what kind of data is being collected about consumers. Our second panel looked at what consumers expect is happening with that data. And now in this session, we're going to look at what actually is happening with the data.

I'm really pleased to have with me researchers who are going to present three outstanding research presentations. And we're then going to discuss them. So why don't we get things started with a presentation from Michael Tschantz and Anupam Datta? Michael is from Berkeley and Anupam is from Carnegie Mellon. They're going to lead things off with a presentation titled Automated Experiments on Ad Privacy Settings.

MICHAEL TSCHANTZ: Thank you. I am Michael Tschantz, and this is going to be a joint presentation with Anupam Datta. We're going to talk about Ad Fisher-- a system for looking at online trackers and determining what information they are using about people to select the ads they show to people. There are two things I want you to take away from this talk.

First, it is possible to do this with scientific rigor, despite not having access to the internals of the system. And second, we can find certain flows of information, but we can't figure out why they happened. So let's get started by motivating the problem. Here's a web page. It's the Times of India. I find it an interesting example because it has a lot of ads from Google on it. Here's two.

Now, Google has little pieces of code across the internet. In fact, this web page has two little pieces of code. And these pieces of code report back to Google about what other web pages you visited. Google can then select the ads it shows on the Times of India based upon this information.

And this is generally true of online behavior trackers. There are many trackers with many little pieces of code all over the place. There's a seemingly endless number of companies doing this kind of thing. But it can be concerning.

Suppose, for example, you want to show a friend a newspaper article, and you see nothing but ads for anti-depressants, which Google will show under certain circumstances. Now Google understands that people have concerns like this, so they and other companies have provided things like the ad privacy settings.

Here's a screenshot of my ad privacy settings. It shows various information inferred about me. Google got my age correct, but got my gender wrong. Google also allows you to go in and edit this information. So if I cared, I could go in there and provide my correct gender. Google doesn't give us a whole lot of information about exactly how this thing is working, however.

So what we have is a situation where we have our web browsing behavior going into an ad ecosystem at one end. You have various things like ad settings sitting in the middle, providing sort of a window into how that ad ecosystem works, providing inferences they create and allowing you to put edits in. And then we see advertisements coming out the other end. But we would like to understand the flows of information in the system better than they currently make clear from their privacy policies and descriptions of how these systems work. And this is a difficult task because the system is opaque. We don't know what's going on in that ad ecosystem. Google and other online behavioral trackers won't share its source code with us. We can't do the traditional forms of program analysis. So we designed Ad Fisher, a system that allows us to run experiments on these kinds of opaque ad ecosystems.

Let me run through quickly how Ad Fisher works. Ad Fisher creates a bunch of fresh Firefox browser instances which simulate users. So these could be simulating people who browse various websites. It randomly assigns them to either a control or an experimental group. These two groups of simulated users will display different behaviors on the internet.

They then interact with the internet in various ways. And we collect measurements about how advertisers change their behavior towards the simulated uses. These measurements go into a test of statistical significance, which reports whether there's a statistically significant systematic difference between the experimental and the control group. If so, we know that whatever information describes the difference between these two groups and how they behave towards the ad ecosystem is information being used by the ad ecosystem to select ads.

So this is our main contribution. We brought the rigor of experimental science to these online black box experiments in such a way that allows us to detect causal effects, which are equivalent to flows of information with the theorem reproved. It does it with statistical significance, without making questionable assumptions about how Google operates. This is important because Google is an extremely complex system. Pretty much any assumption you make about how it operates might not hold. Or perhaps it even holds at one moment in time, but not later when you're running your experiment. And we provide a high degree of automation.

So now I'm going to give you an example of one of the findings we discovered with our system. In this experiment, what we did was we fired up our simulated users, and we had half of them set the gender bit to be male, and the other half to female on the Google ad settings page. We then had them all browse websites related to finding jobs.

We then collected the ads shown to them at the Times of India. And we found a statistically significant difference in the ads shown to the male and female groups. And this, in and of itself, isn't terribly surprising. We know that advertisers show different ads towards men and women. But what's concerning is the nature of this difference, something that Ad Fisher can also share with us.

What we found is that there were a series of ads from a career coaching service that was shown almost only to the male simulated users. In fact, the ratio was so large that it's in violation of the 80% rule often used in employment law to detect disparate impact. That being said, we're not claiming that this is an instance of illegal disparate impact, because this is an ad for a career coaching service. It's not actually for a job. Nevertheless, we find this ad being shown predominantly to men to be concerning.

Now this is just one of the findings. We have another interesting one involving substance abuse. We found that if you visited a website for a rehab center, suddenly Google would start showing you ads for that rehab center across the web, or at least at the Times of India. And this is concerning since it's sort of like medical information being used for determining the ads you see on a newspaper's website.

I've use my time to explain some of the things we know. Anupam is going to now explain some of the questions left open.

ANUPAM DATTA: I'm very excited about where this research area is going in terms of developing rigorous science and useful tools that are beginning to find effects in the ad ecosystem, and more generally in online personalisation systems. At the same time, I'm deeply concerned also about the findings themselves that we and others in this research area are beginning to develop. We'll hear more from the two other speakers shortly about other findings.

These studies are beginning to get a lot of attention in the popular press, indicating that these concerns are shared much more broadly in the community. But there's much more to do in this space. There are questions like, how widespread are instances of discriminatory targeting or targeting that violates privacy expectations of, perhaps, contextual integrity or other notions? And then there's also the question of who is responsible. So I want to take a few minutes to highlight that these questions are incredibly nuanced to answer in the presence of the complexities of data analytics and other pieces of an ad ecosystem.

So I'm going to focus on this question of responsibility, partly because, following up on the conversations from the morning, I think that detection is an important step. But we can't just stop there. We have to go towards accountability, meaning assignment of responsibility and then institution of corrective measures. And this is going to involve collaboration between computer scientists and legal scholars and public policy changes. I want to focus only on the computer science piece of it for now. But we are working on the interaction between computer science and law in collaboration with Deirdre Mulligan.

Let me highlight some of the nuances of assigning responsibility with this concrete instance of discriminatory targeting that we found. Just to remind you, this was an instance where high paying, job related ads were being served in significantly higher numbers to simulated male users rather than female users.

So what are some possibilities here? Which entity could be responsible? So one possibility is that Google's programmers intentionally programmed their targeting system to be discriminatory in this way. We considered that to be highly unlikely, but nevertheless, it's not something we can rule out because we don't have enough visibility or access into the system that they use internally.

Another possibility is that the advertiser, the specific advertiser in this case the Barrett group that was advertising for this career coaching service, might have indicated when they submitted their bid for the ad that Google should show this ad more to male users than to female users. And Google may have honored that request.

A third possibility is that perhaps the Barrett Group indicated that the ad should be shown to high earners. In fact, in response to the questions from journalists at Pittsburgh Post Gadget, the Barrett Group actually said they were targeting users who are over the age of 45, and who earn more than $100,000, because they thought that would be an appropriate group to target, people who want to go one level up and go to the 200k plus jobs. Now it could be these high earners are much more strongly correlated. There's a stronger correlation with the male gender than the female gender. And Google may have inferred that and then decided they should send more impressions of this ad to male users than to female users.

Yet another possibility is that other advertisers might be targeting the female demographic more. And there's some evidence that the female demographic is targeted more by advertisers because they make more purchasing decisions. And those other ads may have come with higher bid amounts, which took up the slots for the female users. And the males just got the ad from this particular service because they were the leftover untargeted. There were just more slots available for the male users.

Yet another possibility, and this would be the case of machine learning introducing discrimination, is that Google's internal systems may have observed that more male users are clicking on this particular ad than female users. And since machine learning systems learn from these kinds of observations, and they're trying to optimize the clickthrough rate, they may have started serving more impressions of this ad to the male users.

So all of these are hypothetical scenarios because we don't have enough visibility into the system to determine which, if any of these possible explanations, is the real explanation. But I wanted to highlight this to explain the nuance of this problem. This is a very complicated problem. And if you want to go towards making systems more accountable in this space, then the researchers will need additional access to the internals of the system.

So being able to work, not just from the outside like we have in this work. And Roxana will talk about shortly about her work as well with the sunlight system. They have a similar model. People on the inside who have more access might, if they are interested in proactively testing their systems, that additional step will be very crucial towards proactive detection of violations as well of identifying responsibility.

So that's something that I urge this community to go towards. And it's an open call to technology companies to work with researchers like us to work on problems of this form that are socially important. So let me stop here with this summary. This body of work, Ad Fisher and the previous result that introduces methodology, brings rigorous experimental design ideas to this research idea, which lets us discover causal effects.

For example, it's really the difference in gender that caused the difference in high paying job related ads being targeted, which is statistically significant. So with confidence that it's not just a fluke observation, but it is really how the system is behaving. And a third contribution here is to bring automation that allows us to discover these kinds of effects at scale. And this combination was the first in our work, and the community has grown and developed it in many different dimensions.

So we found evidence of gender based discrimination. That was one specific highlight. And the other highlight is how browsing health related websites has a significant effect on targeting. In particular, how browsing substance abuse websites results in rehab ads being targeted. And the two big, open questions that I want to open up for discussion-- and these are active areas of research in this area-- is how widespread is this discrimination?

And how do we go, from here, to assigning responsibility? And as a corollary, I would like to emphasize that additional access to the internals of the systems-- and people with additional access to the internals of the system-- working with such people is going to be highly crucial towards achieving these goals. Thank you very much.

DAN SALSBURG: Thank you, Anupam and Michael. Now we're going to hear a presentation by Roxana Geambasu of Columbia University titled Sunlight Fine-grained Targeting Detection at Scale with Statistical Confidence.

ROXANA GEAMBASU: I'm very happy to be here. I will now tell you about some tools that we are building at Columbia to increase the web's transparency at large scale. To explain our work, I'll start with an example that shows just how opaque today's web is. You probably already know that Gmail uses emails in order to target ads.

But did you know how the keywords or inferences drawn from these emails are being used to target you specifically? I'll test to see how aware you are of how you are being targeted by showing you some examples that we got from an experiment. We created this Gmail account, and populated it with a bunch of very simple, single topic emails. I'm showing here on the left hand side, five of those emails of about 300 that we created.

After that, we retrieved ads that Gmail showed in this account. I'm showing here on the right hand side ads, two ads, out of about 20,000 that we got. So this was a pretty large scale experiment. What I want to do is to challenge you guys to tell me what each ad is targeting.

So for example, what does ad 1 target? Which of the emails? What do you think? Just quickly, whatever comes to mind. Vacation. Well it actually it turns out that ad 1 targets the pregnancy related email. It's pretty hard to tell, right? Nothing in the ad really tells you anything about how it's actually targeted. What about ad 2? It's about a hotel. What does this one target?

AUDIENCE: Homosexuals.

ROXANA GEAMBASU: I'm sorry?

AUDIENCE: Homosexuals.

ROXANA GEAMBASU: You got it right. That's exactly right, the homosexuality related email. Again, it's still pretty hard to tell. And it's not just the targeting of ads on Gmail that's hard to discern. Everything is obscure on the web. For example, data brokers apparently are using-- can tell when you're sick or depressed and actually apparently sell this information.

Or some credit companies, for example, are apparently trying to use Facebook information to decide whether or not to give out a loan. You may have heard of these things from the media, just like I did. But do you know that whether these things are actually happening, to what degree and how those things affect you? I bet not. People don't know too much about these things.

Welcome to the data driven web. Web services and third parties collect huge amounts of information about us. Every location, every site, every site that we visit, every click that we make and so on. And they leverage all of this information for all sorts of purposes. Some are in line with our interests. For example, we all love our Netflix recommendations or Pandora recommendations.

But other uses may not be so beneficial for us. And the problem is that we have absolutely no visibility into what happens with our data in this huge complex web data ecosystem. Who has access to what data? For what purposes are they using it? Is this good or bad for us? How do their uses affect us? And it's not just the end users that don't know how to answer these questions.

But society as a whole has a hard time answering these questions. And I believe the FTC does as well, from my communications with them. And that's very dangerous because obscurity and lack of oversight can lead to abuses, either intentional or not.

So in my group at Columbia we are developing new kinds of tools, which we call transparency infrastructures that shed light into this dark, data driven web. Our goal is to build really large-scale infrastructures that can go on the web and track the flow of information and reveal it.

So on one hand, we can increase users' awareness of what happens with their data online. And on the other hand, empower privacy watchdogs, such as the Federal Trade Commission, to audit what web services are doing with users' data and keep them accountable for their actions.

And over the past several years, we've been building a number of these transparency infrastructures. And we're continuing to do so now. And in this talk, I'll tell you about just one of these infrastructures in the remaining time, the latest, essentially public domain, transparency infrastructure that we've built. Before I do that, I want to acknowledge my students and collaborators without whom I wouldn't be standing here telling you about these systems.

So what's Sunlight? It's a generic and broadly applicable system that detects personal data use for the specific purpose of targeting and personalisation. It detects which specific data about the user, such as email searches or visited websites, are being used to target which service output, such as ads, recommendations, or prices.

The ads I showed you at the beginning of the talk. Their targeting was discovered by Sunlight. Sunlight has three unique properties in its combination compared to everything else that exists. It is precise, scalable, and very broadly applicable.

We've already tried it with great success to review targeting of Gmail ads, ads on arbitrary websites, recommendations on Amazon and YouTube, and prices on various travel websites. Not all of these experiments are actually in open domain yet. In all of these cases, Sunlight works with high precision, about 95%, as well as reasonable recall.

How does it work? Well, the details are pretty complex. But at the high level, the idea is intuitive. Sunlight targets by correlating users' inputs, such as emails, with service outputs like ads by performing experiments on accounts with differentiated user inputs. We can actually make the link from correlation to causation if we control how those inputs are placed in the accounts. Let me show you an example quickly just to illustrate this process.

Remember the ads I showed you at the beginning of the talk? I'll show you how Sunlight might have detected their targeting. But let me first simplify the example a bit. Let's skip three emails and one ad. And let's ditch the contents of the emails and ads.

So what we have is a main account that consists of emails E1, E2, and E3. These accounts are ad 1. What we want to do is explain the targeting of ad 1 on one or a combination of these three emails. What we'll do is three things.

First, we'll create a set of extra accounts. We call these shadow accounts, say, three accounts. We populate them different subsets of the emails. We do this randomly, so the placement of the emails into the accounts is random, is done randomly, independent of any other variable. Second, we collect ads from the shadow accounts. Say, for example, that shadow accounts two and three observe ad 1, but shadow account one doesn't.

Third, we analyzed these observations and yielded a targeting prediction. In this case, the most natural prediction we would reach is that ad 1 targets email three, because the ad appears in all accounts with email three, but never in accounts without email three. So that's kind of how Sunlight works.

And there is an important distinction that I'd like to make, which is that the first two stages of this process-- populating shadow accounts with subsets of the emails and collecting ads from them-- are server-specific. And in particular, in Sunlight, the emails are kind of mindless, pretty simple, simplistic. We just do some browser automations. The last stage, however-- the analysis of these observations to yield a targeting prediction-- is intellectually challenging. And that's what Sunlight actually provides.

Specifically, the example I showed you here is trivial. In reality, the scale is much larger. There are a lot more emails to consider. A lot more ads to explain. There's a lot more noise and so on. So all of these things make targeting prediction challenging. And Sunlight addresses these challenges by designing a rigorous methodology that leverages well known methods from statistics to provide precise targeting predictions at scale. And it does so, very importantly and quite uniquely, in a service agnostic way, so that we can be re-use the analysis across many different services like I said before.

Let me show you some of these challenges just to exemplify the kinds of mechanisms that we use to address them. Let's look at that simple example that we had with the three emails. Look at what we did. We used three shadow accounts to explain targeting on three emails. That's a lot of shadow accounts that we needed to create. What if we were trying to explain targeting on a more realistic user account with thousands of emails and potentially other online activity too that compounds together with the emails to produce the ads?

Would we have needed to create all combinations of a number of accounts that's equal to all combinations of these inputs? That's a huge scaling challenge that I think is tremendously important. And it turns out that we don't need as many extra accounts. We can get away with a lot fewer on a logarithmic number in terms of the number of inputs that we're trying to explain targeting on. And my theoretician collaborator, Augustin Chaintreau, proved this aspect theoretically. And we validated it experimentally.

And the intuition is that if we can assume that an ad targets only a small subset of the many inputs that we have in a main account, then we can leverage sparsity properties, the same concept underlying compressed sensing, which say that you don't need a whole lot of observations to reconstruct, accurately, a sparse signal. For those of you who are familiar with machine learning, that's what sparse regressions is, and that's what we use in Sunlight.

However, these particular methods only guarantee asymptotic correctness. They do not guarantee the correctness of any individual prediction. And what we want is a correctness assessment of individual targeting association, so that we can trust the results we get from Sunlight. For that we used hypothesis testing, just like in Ad Fisher, a well-known method that provided quantification of the statistical significance of each prediction.

Sunlight puts all of these things and other mechanisms together in a particular architecture that provides the unique properties that I mentioned before-- genericity, broad applicability, scalability, and precision. I won't go into the details of this. Instead, in the remaining two minutes, I'll tell you how Sunlight can be used. Specifically, Sunlight is a transparency infrastructure which provides some valuable primitives for targeting prediction. And on top of it, we and others build transparency tools for studying specific services. And we've built a bunch of these tools. And it's actually extremely convenient to build on top of Sunlight.

I'll tell you about just one of these tools that we've built, which we call the Gmail Ad Observatory. It's an online service that enables studies of targeting of Gmail ads on user's inboxes. Here's how it works. A researcher or journalist supplies a set of emails on which they want to detect targeting. The Gmail Ad Observatory uses Gmail accounts to send emails to a separate set of Gmail accounts that become, then, the shadow accounts from which we extract the observations or collect the ads and infer the targeting.

The Gmail Ad Observatory then collects the ads periodically and supplies them to Sunlight to infer their targeting. This is the tool we built, and we used this tool to run a 33 day study of ad targeting in Gmail, a pretty large scale study. We got, overall, about 20 million impressions of ads and about 20,000 unique ads.

We found a bunch of things. I'll share just one result, which is a contradiction of one particular policy or statement that Gmail makes in one of their FAQs. Specifically, they say they don't target ads based on sensitive information such as religion, sexual orientation, health, or sensitive financial categories. Well guess what? We actually found examples-- a lot of examples-- that target each and every of these specific topics. I've already shown you, for example, that targets homosexuals. Let me show you another example from the health category specifically.

There are some senior related-- a lot, actually-- of senior assisted living ads that target Alzheimer's. There were many ads, actually, that target Alzheimer's in general. There's also an interesting ad that you can see there that targets depression related keywords. The "Is He a Cheater?" ad for a cheating spouse search site, apparently. There are a number of ads in our example that target the keyword cancer. I'm showing just one of them. We found a number of other ones.

DAN SALSBURG: Just take a sentence to wrap up.

ROXANA GEAMBASU: Yeah that's right. So to wrap it up, I told you about our agenda of building generic and broadly applicable transparency tools, which enable oversight at scale. These tools can be used to study targeting phenomena of various kinds, like ad targeting for example. But also price targeting, and I actually have a demo of that if you guys would like to see it later. Thank you.

DAN SALSBURG: Thank you, Roxana.

[APPLAUSE]

Our final big data and algorithm research presentation will be from Daniel Hsu of Columbia University. It's titled Discovering Unwarranted Associations in Data-Driven Applications with the FairTest Testing Toolkit.

DANIEL HSU: I'm going to tell you another tool we've been developing at Columbia but also at EPFL and Cornell Tech-- a lot of collaborators on this project. I should preface this by saying that I'm an outsider in this community. I mostly do research in machine learning on the algorithms that are used by Google, by Yahoo, by Microsoft for doing data analysis or maybe doing the targeting. And so this is kind of a different perspective. I'm going to give a different perspective on this problem.

You're all well aware of the kinds of issues that come up with a lot of data driven applications. You've probably heard of the study that was done about detecting differences in prices from Staples' online store based on where you live. This turned out to have some kind of correlation with the income of potential customers. And this was sort of an interesting finding. But what's more interesting from our perspective is that this was an unintended consequence of the pricing mechanism that Staples was using.

Here's another example of a data driven application that may have unintended consequences. This was in the case of Google's image tagging application where, if you were to upload photos onto Google social network services, Google would try to automatically tag your images with various things like. There's a car here. Here are your friends. There was this very unfortunate incident where people found that African American users' pictures were being tagged as gorillas. This was definitely not what Google was intending. This is not something that they wanted to happen.

These are the problems that arise when you are creating these data driven applications. What we want to argue in this work is that these are bugs. Developers should be testing for these kinds of bugs and trying to debug them to correct issues the same way they would try to correct or debug to find functionality bugs, performance bugs and so on.

This is where our work comes in. We know that this is not easy. This is not an easy problem to solve. These bugs are pretty nefarious. They're pretty hard to detect. So what people might suggest is that, OK, take some preventive measures. But these, we know, also have a lot of limitations.

One thing you might suggest to do is say-- OK, maybe we should just completely ignore certain attributes about the data when we're designing these data driven applications, so that we do not create these unwarranted associations in the service outputs. But we know it just doesn't work because there are always other attributes that may be associated or correlated with sensitive attributes like income level or race. This is indeed what maybe happened with the Staples pricing application, where location just happened to be correlated with income level. OK so that might not work.

Another thing you might try to do is to apply some kind of sanity checks to see if there's some kind of statistical parity in your outputs to make sure that, if you look at race, you're at parity across different race attributes. We know this can be insufficient as well just because there could be smaller sub-populations with a particular attribute that end up having a strong association with the service output.

These are really hard problems for developers to solve. What we're trying to argue here is that developers really do need new tools to help them find these kinds of bugs. So detecting these kinds of unwarranted associations is already a hard task for them to do. This is where our research comes in.

We've been developing this toolkit that we call FairTest. We call it a testing suite for data driven applications for a developer to integrate into their tool chain to try to check the application, to do debugging, to run every time they compile to make sure that their application is working as they want it it. The way we caricature data driven application is that, in a data driven application, it somehow takes user data as inputs. And then there's some kind of output that the application provides, maybe it's service prices, image tags, recommendations, and so on. These are some kinds of functions of these outputs. Things like the user inputs might be locations of the user and their profiles, whether they click on various things on a website. And like you said, the outputs are like the prices or image tags.

So FairTest comes in as something that you could strap onto your development tool chain. It would look at these kinds of user inputs and the application outputs and try to check for various kinds of unwarranted associations between the output and protected attributes you wouldn't want to have some kind of strong association there. So FairTest is a tool for automatically doing this. It does this with some kind of data. The hope is that it will, at the end, produce some kind of bug report that the developer will be able to look at.

So what the developer would have to do is specify which of the user inputs are the ones that we want to check for a strong association. These are what we call protected variables or protected attributes. These might be things like the gender or race of the user. There are many other attributes that are used by the application. These are things that we're going to use to try to define or search various kinds of context in which there might be some kind of unwarranted association. The last one I'll talk about in a bit.

The goal of FairTest is to find these kinds of context specific associations between protected attributes and the application output. The bug reports is something that we'll apply some statistics or machine learning to produce something the developer can understand in terms of which kind of context, what kinds of associations were found by FairTest, and to rank them by the association, the statistical significance. So that is something that the developer can actually look at and understand.

Let me say a little bit about how FairTest works. At its core, it's a machine learning algorithm or machine learning application. FairTest itself is a kind of data driven application. The way that it works is that it starts by collecting-- you start by providing it some kind of source of data. This is where it's really important for the developer to have some kind of source of data that is representative of a population or of their user base. this is where it's difficult for other parties to have access to this. But a developer, presumably, is working at Google or at Microsoft, so they have access to this kind of data already. When they have this kind of data, they can really check the application on the real user population to really discover the effects that have some meaning in terms of the actual users.

FairTest relies on this kind of data. So we'll do something very similar to how Ad Fisher and Sunlight operate. t will split this data into two parts. One we call the training data. The other part we call the test data. We use the training data, part of the data set, to find these kinds of associations through a clever machine learning algorithm. Once we find these kinds of associations between protected attributes and application outputs, we'll use the remaining data as separate data to actually validate these things, measure their effect sizes, and to check if these things are harming a large segment of a population. Is it very significant and so on. This is where there's a lot of technical machinery coming from machine learning. At the end-- actually, a lot of the work here is to make these findings consumable by the application developer. So something that's interpretable and that they can actually use to help them debug their application.

Let me give you an example. We've actually applied this tool to a couple applications, real applications that are data driven applications. One of them, the first one I want to tell you about, is this health care application. This is actually something that was produced by one of these machine learning contests-- er, data science contests. Some company, in this case, Heritage Health company, ran this competition where they provided some data about patients going to hospitals, a description of the patient records, how many times they've been to the hospital before, what were their symptoms, and things like this.

The task was to use this information to predict whether or not --whether or not or how many times the patient would visit the hospital in the following year. So there's a readmission rate prediction. What we did we do-- we looked at the winning entry to this competition. It was a pretty good entry. It was a certain application that was able to correctly predict with some pretty high accuracy, I think around 85% accuracy, whether or not the patient would be readmitted to a hospital in the following year.

This was the data driven application. It takes these kinds of inputs-- age, gender, number of times they've been to the hospital, and so on. And then it tries to predict whether they'll be readmitted to the hospital. What did we find by applying FairTest here? We found that there are some specific contexts where there's association between the age of the patient and how bad the predictions were, the rate of error or the size of the error in the prediction.

This was a contextual association that we discovered. It was not for the entire population, but for some well defined segment of the population. I think it was something like male patients who have been to the hospital-- who had been to the ER at least twice in the past year. But within this sub-population there was a really strong effect and a really strong association between age and the error in the prediction. This is an interesting finding.

We think that this is important in this social sense because this is something that could potentially lead to actual harms. If this application was actually going to be used for insurance purposes to adjust your insurance premiums and so on. So these are associations that can really have some impact on the users of the system.

I want to tell you about another application. This is not a real application. It's sort of a historical application. But I thought that would illustrate a different capability of FairTest. So this is a very well known data set. You can think of it as graduate school admissions application. It takes people who apply to Berkeley graduate school, and then decides whether to admit them or not. This is a well known data set from the 70s.

If you don't know about this data set, what happened was that they discovered there was gender bias in the admission rate at Berkeley. Men were being admitted at higher rates than women.

Indeed, FairTest can be used to discover this kind of association. We can try to explain where this association comes from. And indeed, this paper by Bickel et al in 1975 discovered that once you condition which department the applicant wants to get into, then the effect either goes away or the impact reverses. Women in specific departments would be admitted at higher rates than men.

What we want to do is illustrate how FairTest can be used to help a developer debug their system and try to explain what's going on, what's going wrong in their system. There's this other capability in FairTest for doing this. We call it providing some kind of explanatory variables. And this will really make this a real system, a real tool for developers to use to debug their applications.

Let me just make a few closing remarks. We also apply FairTest in a couple of other applications. You can read about in our preprint, which is available on the web. I already mentioned this other feature of explanatory variables. There's another big issue out there in data analysis, which is that of adaptive data analysis, where you want to be able to reuse a data set many times. This is something that we're starting to look at and integrate in FairTest. This is open source software that can be used by developers right now.

[BEEPING]

Really, what we're trying to advocate here is that we need to empower developers with better statistical trainings, better statistical tools to make these data driven applications more fair, more socially conscious, and so on. We think that FairTest is a good way to start here. Thank you.

[APPLAUSE]

DAN SALSBURG: Joining me on the stage now are discussors James Cooper of George Mason University Law School and Deirdre Mulligan of UC Berkeley. So we just heard three presentations about tools that are designed to shed some light on how data is collected from consumers, how it results in them receiving targeted ads, web content, or it can result in discrimination. Let me turn first to James and Deirdre. What are the common themes you see running through these three presentations?

DIERDRE MULLIGAN: I teach at the School of Information at Berkeley, and I spend-- one of the departments, one of the programs in which I teach-- is a Master's in Data Science. And we teach about privacy, we teach about security. These are people who are going to be doing data analytics. And one of the areas we've been lacking, both methodologies and tools, is to deal with issues of fairness. How do we think about the biases in our data? How do we think about the biases in our algorithms? Most importantly, I think in particular-- I'm most deeply engaged with Anupam and Michael's work because we have some collaborative work that we're doing.

How do we think about bias in systems where there are multiple inputs? It's very difficult to track an output back to a single actor's decisions. As somebody who's working in that sort of program, one of the things I think is most important about these tools is, on the one hand, we have our last presentation, FairTest, which is actually trying to empower people who want to avoid-- All algorithms have biases. If you design an algorithm without a bias, it has no purpose in the world.

Let's be clear. It has a bias, it's just that we want to avoid certain bad outcomes. The question about how we empower people who are designing systems to proactively avoid those outcomes is something that we need research on technical systems. People have called for, oh, we need access to the algorithm. We need access to the data. As though, if they can look at it, they're going to understand it.

And that just isn't the case in many instances. We actually need technical systems. We need the use of statistical machine learning techniques to police machine learning systems. This is particularly important because, I think what all of them are highlighting and really focusing on is-- I mean, we're concerned about intentional discrimination. But what I think many of us are worried about exploding is disparate impact. It's that nobody is intending for a bad thing to happen.

But because what machine learning enables, what makes it different from what's gone before is that, the meaning of information emerges. It turns out that these three pieces of data add up to some particular protected tree. As machine learning techniques continue to uncover the way in which we have correlations that equate to these different things, we have this ongoing need to try to figure it out proactively, how to avoid those sort of problematic correlations.

So I think they're all working on this shared problem from two different sides. There's a long history of testing. We think about discrimination, housing discrimination, sending people out in the world. I think Ad Fisher and Sunlight are working on that side.

Can we test from the outside? I think Daniel's work is really nice because it's saying, for the people who are trying to do good, trying to avoid bad outcomes-- can we empower them with tools that are based on the same sorts of statistical techniques that we need to police machine learning? So I think they're really powerful in that way.

DAN SALSBURG: James, what do you see as the common themes?

JAMES COOPER: I would agree with what Deidre said. They're obviously-- many of the common themes are pretty self evident. They're co-authors back and forth on two papers. And their papers describe algorithms that do very similar work. Valuable work, as Deidre pointed out. I don't really have much to add beyond that.

DAN SALSBURG: Dierdre pointed out that, in the real world, there are lots of inputs. A consumer profile consists of a million data points or more. How can your tools account for that? When you're creating user profiles, is there any way to really mimic what would really be happening to a consumer?

ROXANA GEAMBASU: So this is a real problem, a very big problem. I would quote it as the biggest problem in web transparency work today in my opinion. Which is to actually emulate real users with controlled experiments. Both Ad Fisher and Sunlight rely on controlled experiments with fake accounts that are assigned fake input sets or inputs. That results in some targeting. We are seeing, all of us, some targeting.

But it's not necessarily true that it's the realistic kind of targeting that real users would actually see. We may be losing a lot of the targeting that real users see. We may actually have targeting that real users never see, and so on. I think that's a big problem. I think we need research in designing tools that leverage direct data from real users to achieve some of the goals we have in our system, transparency goals we have in our systems.

That said, because I've been working and have invested so much in scalability, building scalable systems that can take many inputs. But the millions, you know, not the size that real users produce. We've been focusing on that. And sunlight does scale pretty well with respect to trying many inputs and discovering effects on many of these inputs. But there are big limitations still even there. I also wanted to point out, because maybe the audience didn't realize-- FairTest and Sunlight-- we're actually both collaborators on both. We just split the talks so that we wouldn't have to both talk about each one.

ANUPAM DATTA: One quick thing I would add is that there are two ways to go about getting access to real data. One is to actually work with the technology companies who have that data. And so we have an ongoing collaboration now with Microsoft research, where we are actually beginning to get started working with the internal data they have about their users. The other way to do it, or at least one other way to do it, is to try to get data from real users through crowdsourcing. There is a recent interesting paper from AT&T research and collaborators elsewhere which tried to do that. So the way they do their experiments is to crowdsource and collect data from users about their browsing profiles. Then they compare it against the same user without the history, some amount of the history, and then see if there's a differential treatment. That's beginning to get towards experimental findings that have some amount of real user data.

DAN SALSBURG: James, do you have a question you want to throw out?

JAMES COOPER: Sure. It's sort of a question and a comment. I'm an academic, so of course I'm going to spend most of this question saying what I want to say and then ask you. So one of the issues, and I think this applies more to Michael and Anupam's paper, but I think to all papers is--

If we think about the transmission of your findings into policy-- I think one of the touchstones of policy, in my view, should be harm. I think about the finding of the job search ad different for men and different for women. If you look at the statistics. Let's assume the data's there, and there's a statistical difference. And we can even say it's causal. Digging down deeper, what's the real world impact of that?

Clickthrough rates are what? Maybe one out of 1,000 if you're lucky, right? That's the average I think, right? One out of 1,000. So let's say one out of 1,000 people who visit this website click on that. And these are people whose profiles had visited other job searching websites. My point there would be, to what extent-- they're not going to be limited. This isn't really necessarily, hey, I've gone to 1,000 job websites. But now I've gone to the Times of India. I'm just going to take a job. I'm going to follow my career based on this ad that's served to me. I think that's probably not likely.

I visited both those websites. I'm sure you have. I don't know how many have. The one's a head hunter website. I'm not sure. It's got the nice banner-- 200k plus. It's a head hunter. I am not saying it's-- I'm sure it's legit. I'm not suggesting the FTC look into it or anything. But compared to the other one, where the women were served more often-- that was, I think, Jobs Near You. Go on that. And the first page click down menu, they're not like blue collar jobs. They're accountant, lawyer, bio.

If you look at what would be the real world impact. If you could imagine the two random people, the man and the woman. The woman says, well I didn't see the head hunter ad. So I'm just going to go with Jobs For Me. I think about the real world impact. You did find statistically significant difference between men and women, but at the end of the day, before we get into issues of harm, which I think should be the touchstone of any policy, especially here at the FTC-- do you need to find more? Is there actually some sort of evidence of harm here?

MICHAEL TSCHANTZ: Well, there's a saying amongst advertisers which is, I waste half of my budget. I just wish I knew which half. I really don't think anyone can look at any one ad and necessarily know what its entire impact is. But we do know the advertisers. You don't see Coke ads on the TV because they expect you to stop watching TV and run out and buy a Coke, right? These ads can be functioning in a similar way. It's about creating an impact upon people that lasts when they see something over and over again or don't see something over and over again.

So we're concerned about the women not being exposed to the encouragement to seek high paying ads, just as much as we're concerned about whether any one person clicks on that ad or not. I do think you raise an interesting point about this firm putting up this ad. I looked up some customer reviews, and it didn't really have the highest customer reviews. So if we look at just the lack of women developing a business relationship with them, then it might be actually in their favor that they're not seeing this ad. So I don't know. You are correct. We can't pinpoint and measure the exact amount of harm. But we do know that men and women are being--

JAMES COOPER: Or any harm. I would go that far.

DIERDRE MULLIGAN: I think there are a few things to highlight. One, there was another example brought out about proximity to work. I don't remember whose paper it was in. Was it in Daniel's?

ROXANA GEAMBASU: Proximity to the location of a store.

DIERDRE MULLIGAN: Yeah, no. No, the proximity to work. It may have been in the FTC's big data report that just came out. It's an example that's been used before. If you were looking for a potential employee that you wanted to advertise to. And you said, oh well, people who live closer tend to be better employees. Then you might find out that has a lot to do with income. It could be a proxy for something else.

And when we're thinking about employment, equal access to, not just employment opportunities, but also we think about the advertising of those employment opportunities as something where we're concerned about racial disparities and gender disparities in how we're making information about opportunities available. As a legal matter, we're concerned about that. So let me finish. Hold on.

Setting aside this particular example, which we agree is problematic for many reasons, I think one of the most interesting things this particular example of the headhunter ad brought out, which Anupam noted is that-- the most likely, we think, or at least a highly likely reason, that men were seeing this more than women is that people were willing to pay more to show women advertisements for hair care products and other things.

The point being that, if you were a company and you were trying to use this to make information available about employment opportunities, you don't have complete control over who sees them full stop. And when we're thinking about anything that requires-- where you as an advertiser want to be attentive to who's getting access to your ads, because you're interested in making sure that they are equally available to the population to find whatever you want.

And you realize that there are other people whose bidding and decisions are interfering with your ability to know whether or not they're going equally to men and women or they're going equally to people of different races or whatever. You begin to say, well how do we think about causality? And how do we think about the relationship between bad outcomes and infrastructure? Because it becomes an infrastructure issue.

In the Staples example, Staples had access to their data. They were making decisions. They had access to lots of stuff. And they weren't seeking to have a particular bad outcome, from your description, Daniel. Yet they didn't do enough work or they didn't think through what was going to happen. So again, it's about how do we create an infrastructure and tools.

JAMES COOPER: Well, two things. My only point was using findings like this to inject into policy and potential enforcement actions. Because that seems to be an undercurrent in the papers, at least two of them. Well here's a Google privacy policy, and my ad suggested there's tracking, which could lay the predicate for-- So my point is, there seems to be a lack of harm. In the Staples example, to say it's an unintended outcome-- I think it's completely intended. That's just channel conflict mediation. That's just the idea that I've got a brick and mortar store, and I don't-- so that has nothing to do--

DIERDRE MULLIGAN: Their intent wasn't to disempower people who were poor.

JAMES COOPER: No, absolutely. But I guess when you said they didn't intend the bad outcome-- to them, it's the correct outcome. Because it's the correct outcome based on, that's the local pricing, I'm not going to undercut. That has everything to do with competition. It has nothing to do with, well-- there's no model that would set a price discrimination and say, let's charge the poor people more than the rich people. When I go to the movies, and I hold up my George Mason ID, I try to cover up the faculty part, right? Because they charge students less. Oh, you're faculty. Sorry, you pay full price.

ANUPAM DATTA: I have a brief comment on the question. For the job related advertising example. I think this is where I was positioning that open problem of examining how widespread this phenomenon is. This one particular ad is not enough for us to change how public policy works. Part of what Roxana is doing is building these infrastructures that allow examination of the entire internet, possibly. Much more broader variety of sites at scale over many, many months.

If then she finds that there are many, many instances of these kinds of ads-- maybe not this particular questionable ad, but from legitimate services that are showing up repeatedly in a differential treatment form, differential and disparate impact-- then the establishment of harm comment that you were saying is absolutely valid. That additional layer of analysis will not come from the kinds of tools that we are building. That has to come from people like you and the regulatory agencies who will look deeper, into is this really a legitimate disparate impact. The additional harm consideration. I'm absolutely on board with you on that in addition to the other comments.

ROXANA GEAMBASU: So I just wanted something very, very brief. I completely agree with Anupam. I wanted to note that this research is at the beginning. This kind of research into building infrastructures that can tell what's happening is at the beginning. And as a result, we know very little. We have a bunch of examples. That's pretty much what we have. I have great hope for this field, especially because more and more people are coming into it. We'll develop the kinds of infrastructures that we'll need to actually make impact in the legal domain. But right now, I think we know too little to do that.

ANUPAM DATTA: Having proof of existence is useful as a starting point. We don't have evidence that it's widespread. That's ongoing work.

DAN SALSBURG: The good company that wants to ensure it's not discriminating can use Daniel's tool. And the others can get caught by the two other tools.

ROXANA GEAMBASU: That's exactly the way we're thinking and why we've been developing both from the exterior tools from the exterior for auditing. And tools for the developers to actually help them figure out what to do when the pressure is on from the exterior.

DAN SALSBURG: So we have 50 seconds. That gives each of you about 10 seconds to give a final thought.

MICHAEL TSCHANTZ: Just to complete my thought. We've decided in employment, men and women should be treated the same. So to me, the fact that they're not being treated the same is, in and of itself, a harm. Maybe it's not to you. But that's my opinion.

ANUPAM DATTA: So I would say we need a complete accountability tool chain that goes from detection to responsibility assignment to correction mechanisms. And there is an emerging body of work on each of these pieces of the puzzle. Our focus here has primarily been on detection. There's a small amount of explanations in the last talk. But there's a huge set of open questions related to responsibility assignment and corrective measures.

[BEEPING]

ROXANA GEAMBASU: It's OK.

DAN SALSBURG: Well, with that, we will wrap up the session. So thank you all so much. The cafeteria will be open during this break. So you can get coffee without standing in a long, long line. We'll be back in about 15 minutes.

[APPLAUSE]

[MUSIC PLAYING]