Big Data, Big Issues
Fordham University School of Law
March 2, 2012

Thank you, Joel, for that kind introduction. I'm so pleased to be part of this important discussion about the benefits and concerns surrounding collection, use and retention of Big Data.

As we heard this morning, there is no question that collecting, culling, dissecting and cataloguing vast quantities of consumer data, from such sources as social media, online behavior, geolocation data, and the like, has important beneficial uses. In the past several months we've heard reports about how health care costs can be reduced through large scale analyses made possible by big data. Other researchers have reported how sophisticated analyses of traffic patterns and congestion can be analyzed for "smart routing," which could be designed to save consumers' time. And this morning, we heard about uses of Big Data that really make a difference in people's lives, such as predicting infections in newborn babies – where having this information in real time can save lives.

Among the most intriguing uses of Big Data have been efforts to tease out, from social media services, how consumers in the aggregate feel about a product or brand. Internet entrepreneurs are forming companies to provide so-called "Sentiment Analysis" to investors, with insights gleaned from social networks about how consumers feel about certain brands.[1] To institutional investors, this information could be extremely valuable in building strong portfolios.

Sentiment analysis has other uses, including identifying public health concerns and other areas of need. A new initiative by the United Nations analyzes social media public information and text messages to predict job losses, spending reductions or disease outbreaks in the developing world. [2]

Used this way, sentiment analysis can tease out early warning signs to aid better planning and target assistance programs in a region on the brink of possible crisis.

The New York Times notes how much of a growth industry this is: in a recent article it pointed out that statisticians are now in high demand in the retail industry.[3] "Mathematicians", the Times said, "are suddenly sexy." As an economics major myself, I've always found those with a facility to crunch numbers to be appealing. I'm glad others now agree with me.

Sentiment analysis is a form of Big Data that uses large amounts of anonymous data. At least that is what we are told. The individual source of the data feeding into the sentiment conclusions is beside the point.

---

[1] Interview by Pimm Fox, Bloomberg Television, with Jack Hidary, Chairman, RealTimeMonitor.com. (Feb. 6, 2012) *available at* http://www.bloomberg.com/video/85825558/.
[2] *See* Global Pulse (Feb. 29, 2012) http://www.unglobalpulse.org/.
[3] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. Times, Feb 19, 2012, *available at* http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all.

So if this information indeed is – and remains – truly anonymous, and can be put to such creative and beneficial uses, I'm not going to lose sleep over it. Rather, I'll be intrigued to see how beneficial sentiment analysis proves to be in the coming years.

But there are certain Big Data uses and practices that are on my radar screen and get between me and a good night's sleep. As a Commissioner at the Federal Trade Commission, it is my job to protect consumer privacy—a business that the FTC has been in for quite some time. And a business that I have also put many years into—even before I started at the Commission in 2010.

I spent more than 20 years working on consumer privacy issues at the state level – leading the State Attorneys General as chair of their Privacy Working Group, and engaging in enforcement actions on behalf of the states of Vermont and North Carolina. So, privacy is not new to me.

But Big Data's impact on privacy is requiring some new and hard thinking by all of us.

Today I'd like to talk to you about four issues relating to Big Data and privacy that I have been thinking about, and that I pose as issues for you to consider today.

First, turning back to sentiment analysis and other uses of data that claim to be deidentified: it is critical that we drill down and determine whether the information amassed for such analysis is in fact truly anonymous. Researchers – including some here today – have demonstrated that it can be relatively easy to take some types of deidentified data and reassociate it with specific consumers. I am concerned when I hear other researchers claim that information is "deidentified" when it is merely stripped of a name and address. Much of this information may instead by linked to a specific smartphone or laptop. Given how closely these devices are now associated with each of us — many of us sleep more closely to our cell phones than we do our spouses! — data that is linked to specific devices through UDIDs, IP addresses, "fingerprinting" and other means are, for all intents and purposes, linked to individuals.

Second, in the vast collection of data about consumers, we must be careful to insure that collection and use of sensitive information – such as information related to health, finances, or sexual orientation – triggers the heightened protection it deserves. There are important questions about how to implement this principle in the context of Big Data. I am concerned that the vast collection of data about consumers can unintentionally – or even intentionally – include sensitive information, and that the necessary heightened protections are not being provided.

As everyone here knows, the New York Times recently reported on Target's efforts to develop, through analysis of various online and offline data points, a "pregnancy prediction" score.[4] The reason Target developed this analysis was its belief that major life changes, such as a pregnancy, create perfect marketing opportunities, because at such times consumers are the most receptive to changing their shopping habits. The analysis was designed to predict not only

[4] *Ibid.*

whether a consumer was pregnant, but also when her baby was due, so that Target could tailor its offers depending on her stage of pregnancy.

Now let's suppose that Target didn't use any health information in creating its pregnancy prediction score. Let's simply suppose that it used "innocuous" data – such as the purchase of lotions and then several months later the purchase of newborn-size diapers – to determine the kind of purchases and other customer habits that indicate a shopper is pregnant. That is, it used non-sensitive information to create a prediction about health status. The same type of innocuous data could be used to make other predictions of a sensitive nature, like sexual orientation, financial status, and the like.

As Steve Bellovin pointed out, even Target came to understand that its customers were "creeped out" by getting coupons and other offers that clearly indicated that Target "knew" they were pregnant. So Target "disguised" its knowledge by including among the coupons aimed at expectant moms some coupons for other items, making it less obvious that Target was targeting the women with pregnancy- and baby-related items.

We need to address whether heightened protections should be required in this type of situation. Of course I want to hear and weigh the thoughts and opinions of experts like those in the room here today. But at first blush, it seems that some form of heightened protections are in order.

Third, we are all very familiar with the harms that can occur when there is a data breach. The collection and retention of vast amounts of identifiable data creates a greater risk when a data breach occurs. Holding on to vast stores of data flies in the face of one of the fundamental principles of "privacy by design" – data minimization.

In some areas, industry has made progress in providing consumers with certain choices to limit the information collected about them. In connection with behavioral advertising, industry heard the Federal Trade Commission's call for the development of Do Not Track mechanisms that would enable consumers to make choices in connection with targeted ads. Industry has developed both browser-based solutions, and an opt-out cookie-based solution. I am closely watching these developments. For me, one of the most critical points is that Do Not Track is not just Do Not Target, but also, when the consumer so chooses, Do Not Collect.

I know some of you have heard, over the past half-year, my call for industry to "play well in the sandbox": for the cookie-based Do Not Track mechanism—the Digital Advertising Alliance About Ads Program—to work collaboratively with the browser-based solutions so that consumers' choices would be honored no matter how they were initially exercised.

I was pleased to participate in last week's announcement by the Digital Advertising Alliance that it will begin work to incorporate consistent, easy-to-use browser-based tools within

the DAA program.[5] As usual, the details on how this will be implemented will be critically important. I will closely watch industry's progress in creating a robust solution to effectuating consumers' choices, including whether choices provided to consumers address the collection of their information in the first instance, and not just the receipt of targeted ads.

The fourth Big Data privacy issue that I am concerned about is the extent to which the analysis of vast amounts of data results in consumer profiles that will be used to deny consumers important benefits. As Joel Reidenberg mentioned before lunch, concerns about the collection of vast amounts of information about consumers, the accuracy of that information, and the appropriate use of that information are not new to this country. These concerns led to the passage – over 40 years ago – of the Fair Credit Reporting Act in 1970 (FCRA). [6]

But the world is a very different place today, and traditional credit reports are not the only source for information about consumers that can impact their ability to secure benefits and opportunities such as employment, housing, insurance, or credit.

Think about the young woman from Target. Is Target the only one that knows, or has predicted with some certainty through other pieces of information, that she is in fact pregnant? Can that information be used by her employer? Will it impact her chances to get a promotion? Is the fact, or possibility that she is pregnant, being considered by a potential employer?

We've seen press reports about how life insurers use consumer consumption patterns to predict life expectancy, and they use that information to set the rates and coverage they offer. Social media habits can similarly be analyzed as an indicator of future behavior to determine whether someone might be a trusted employee, or a credit risk.

Information can – and will – be scraped from here, there, and everywhere, and then sold to those who are evaluating consumers for jobs, credit, insurance, housing, and other benefits.

And now, as with nearly everything else, "there's an app for that." Just a few weeks ago, the Federal Trade Commission warned several marketers of apps that provide background screening on individuals that they must comply with the FCRA – giving consumers notice, access and correction rights – if they have reason to believe that their background reports are being used for employment screening, housing, credit, or insurance.[7]

To consumers, the practices of data brokers are unknown. Indeed, consumers are often unaware of the existence of data brokers. This must change. I believe that consumers should have access to the information that many data brokers hold about them.

---

[5] Press Release, Digital Advertising Alliance, White House, DOC and FTC Commend DAA's Self-Regulatory Program to Protect Consumer Online Privacy (Feb. 23, 2012) *available at* http://www.aboutads.info/resource/download/DAA%20White%20House%20Event.pdf.

[6] 15 U.S.C. § 1681s(a)(2)(A).

[7] Press Release, Federal Trade Commission, FTC Warns Marketers That Mobile Apps May Violate Fair Credit Reporting Act (Feb. 7, 2012) *available at* http://www.ftc.gov/opa/2012/02/mobileapps.shtm.

Just six weeks ago, I called on the data broker industry to develop a user-friendly, one-stop shop that will give consumers information about who the data brokers are, and provide access to information that data brokers have amassed about them. If a consumer learns that the data broker sells her information for marketing purposes, she should be able to opt-out.

And with respect to information about consumers used for substantive decisions – like credit, insurance, employment, and other benefits – consumers should have the ability to access this information and correct it. And correct such errors wherever they occur. I call on the data broker industry to develop a system where a consumer's corrections to one data broker's files will automatically correct the same information held by other data brokers. It is critical that all data brokers come to the table to develop this mechanism—including those in the mobile space.

We've spent the morning talking about the possibilities – both positive and negative – of Big Data.

While we learned quite a bit this morning, I'll sum up the main points in just a few words. The potential benefits of Big Data are many, consumer understanding is lacking, and the potential risks are considerable.

We need to pay attention to these issues so that the promise of Big Data is realized, and the risks are kept to a minimum. Industry has a strong and justifiable need to continue to innovate. But in order for industry to thrive, it must engage in an honest discussion about its collection and use practices in order to instill consumer trust in the online and mobile marketplace.

Thank you.