Navigating the “Trackless Ocean”:
Privacy and Fairness in Big Data Research and Decision Making
Keynote Address at the Columbia University Data Science Institute
Symposium on “Data on a Mission: Transforming Privacy, Cities, and Finance”
April 1, 2015

Good morning. Thank you, Steve, for your very kind introduction. And thanks to Columbia University and the Data Science Institute for inviting me to speak with you today. The Data Science Institute is poised to examine some of the most challenging and exciting problems in a way that combines scientific, social, economic, and business perspectives. We need all of these perspectives to understand the role that big data is playing, and will continue to play, in our society. Law and policy need to be part of this equation, and I am pleased to have the opportunity to share my thoughts about how policy and legal concerns relating to privacy and data security can be integrated into your work.

You in the audience enjoy a commanding view of these issues. You are the faculty and students at one of the world’s leading centers of research and experimentation on the fundamental questions of data science and its application in areas ranging from genetics to planning cities of the future. You work and study in a city that is a leader in civic uses of big data. New York City publishes more than 1200 data sets on a seemingly endless variety of topics, from pothole complaints to school-level SAT results, and makes them freely available to the public.[1] New York is also encouraging engineers, developers, and designers to turn this data into apps that make life better for its residents and, perhaps, less intimidating for tourists. And in the financial world, where New York plays such a pivotal role, big data is an ingredient in everything from enterprise risk analysis to generating new ideas for retail banking products[2] and making consumer lending decisions.[3] You have a vista into what’s possible today and a vision for where big data can take us.

I’m right there with you. Well, maybe not at a neighboring keyboard or whiteboard, but like you I believe the benefits we can realize from a data-driven society could be tremendous. Healthcare,[4] education,[5] transportation,[6] agriculture,[7] environmental conservation,[8] and the

[1] NYC Open Data, *available at* https://data.cityofnewyork.us/data (last visited Mar. 30, 2015).

[2] PricewaterhouseCoopers LLP, *Where Have You Been All My Life? How the Financial Services Industry Can Unlock the Value in Big Data* 11 (Oct. 2013), *available at* http://www.pwc.com/en_US/us/financial-services/publications/viewpoints/assets/pwc-unlocking-big-data-value.pdf.

[3] *See* Steve Lohr, *Banking Start-Ups Adopt New Tools for Lending*, N.Y. TIMES (Jan. 18, 2015), *available at* http://www.nytimes.com/2015/01/19/technology/banking-start-ups-adopt-new-tools-for-lending.html.

[4] *See, e.g.*, Public Health Watch, *How A Computer Algorithm Predicted West Africa’s Ebola Outbreak Before It Was Announced*, PUBLIC HEALTH WATCH (Aug. 10, 2014), http://publichealthwatch.wordpress.com/2014/08/10/how-a-computer-algorithm-predicted-west-africas-ebola-outbreak-before-it-was-announced/.

[5] *See* Doug Guthrie, *The Coming Big Data Education Revolution*, U.S. NEWS & WORLD REPORT (Aug. 15, 2013 3:53 PM), *available at* http://www.usnews.com/opinion/articles/2013/08/15/why-big-data-not-moocs-will-revolutionize-education.

planning and delivery of government services[9] are all areas being transformed by data analytics. These efforts truly represent data "on a mission." And we are only at the very beginning of these developments.

But with these opportunities come challenges. Big data challenges some of our notions of privacy and creates data security risks on a large scale. And some uses of big data could undermine principles about the fair treatment of individuals. These are issues that I would like to discuss with you today. Concepts like privacy and fairness are hardly uncontested. I can't summarize them in a tidy checklist or set of rules that you can immediately translate into code. So asking you to make privacy and fairness part of what you do day to day in your labs and companies may seem like a big request.

One of Columbia's most illustrious alumni, Supreme Court Justice Benjamin Cardozo, encountered a similar feeling when he first became a judge in New York State.[10] In most cases, Cardozo said, the law was well settled and the facts were clear, and the decision in those cases was obvious.[11] In a smaller number of cases, the law was clear, but the facts were less so. In those cases, Cardozo took comfort from knowing that the general rules were clear, even if he had to take a deep dive into the facts to reach a decision. He also believed that these cases could lead reasonable people to disagree, even if they agreed on the general rules that governed the case.[12]

Cardozo had a third category for the small minority of cases that get all the attention. These are the cases in which the very rule of law that the judge ought to apply was in doubt. It was in these cases that Cardozo described himself as "much troubled in spirit, . . . to find how trackless was the ocean on which [he] had embarked."[13] Never one for understatement, he went on: "I sought for certainty. I was oppressed and disheartened when I found that the quest was futile. I was trying to reach land, the solid land of fixed and settled rules, the paradise of a

[6] *See, e.g.*, Nat'l Highway Transp. Safety Admin., Advance Notice of Proposed Rulemaking Regarding Federal Motor Vehicle Safety Standards: Vehicle-to-Vehicle (V2V) Communications, 79 Fed. Reg. 49,270 (Aug. 20, 2014).

[7] *See* Jacob Bunge, *Big Data Comes to the Farm, Sowing Seeds of Mistrust*, WALL ST. J. (Feb. 25, 2014 10:38 PM), *available at* http://www.wsj.com/articles/SB10001424052702304450904579369283869192124.

[8] *See* Meg Whitman, *Harnessing Big Data Helps to Drive Environmental Progress: HP Earth Insights Develops Early Warning System to Support Conservation Efforts*, HP NEXT (Dec. 10, 2013), *available at* http://www8.hp.com/hpnext/posts/harnessing-big-data-drive-environmental-progress-hp-earth-insights-develops-early-warning#.VRm02GO1_7Q.

[9] *See* Ben Casselman, *Big Government is Getting in the Way of Big Data*, FIVETHIRTYEIGHT ECONOMICS (Mar. 9, 2015), *available at* http://fivethirtyeight.com/features/big-government-is-getting-in-the-way-of-big-data/.

[10] *See* BENJAMIN N. CARDOZO, THE NATURE OF THE JUDICIAL PROCESS 164-66 (Yale Univ. Press 1921), *available at* http://www.constitution.org/cmt/cardozo/jud_proc.htm.

[11] *Id.* at 164.

[12] *See id* at 164. ("Often these cases and others like them provoke difference of opinion among judges. Jurisprudence remains untouched, however, regardless of the outcome.)"

[13] *Id.* at 166.

justice that would declare itself by tokens plainer and more commanding than its pale and glimmering reflections in my own vacillating mind and conscience."[14]

I'm guessing that some of you data scientists may feel the same way about privacy questions. Or maybe you are now wondering whether you ought to. That is not my intention; I don't mean to trouble your spirits. In fact, what I want to suggest to you is that you're not on a trackless ocean. There are some well settled principles of privacy and fairness that you can turn to as you develop data driven research projects here at Columbia or other research institutions, or within your companies. This isn't to say that most big data privacy questions fall into Cardozo's first category, in which there's only one reasonable answer. Nor is it to say that there are no disagreements about what basic values we should seek to uphold in the era of big data.

But you should not resign yourself to drifting on a trackless ocean. There is a wide expanse of solid land for applying principles of privacy and data security protections to your work, and being familiar with this broad landscape will serve you well. You may need to take a close look at the specifics of a research proposal or algorithm to spot and address privacy issues. The answers may not be unique or self-evident; reasonable people may disagree about what specific steps best address the issues that you find. But whether you're running a company or doing research supervised by an institutional review board, your work depends on the trust of those who provide data to you. Thinking carefully through the issues and putting reasonable protections in place is critical to building this trust.

To help you on your journey, I'd like to show you how to navigate the trackless ocean of privacy and data security for big data research and decision making.

Solid Land: Section 5 of the FTC Act

Let me start by explaining the role that my agency, the Federal Trade Commission (FTC), plays in protecting consumers' privacy and data security, which are among our highest priorities. We enforce several sector-specific privacy and data security laws, such as those dealing with financial information, children's information, and credit reporting. We also enforce Section 5 of the FTC Act, which prohibits unfair or deceptive acts or practices, to address practices that these more specific laws do not cover. Over the past 15 years or so, we have brought nearly 100 actions under Section 5 protecting millions of consumers – in the United States, Europe, and elsewhere – from deceptive and unfair data practices. We have used this authority to bring enforcement actions against well-known companies like Google, Facebook, Twitter and Snapchat.[15] We have also brought cases against companies that are not household names, but

[14] *Id.*

[15] *See, e.g.*, Snapchat, Inc., No. C-4501 (F.T.C. Dec. 23, 2014), (decision and order), *available at* http://www.ftc.gov/system/files/documents/cases/141231snapchatdo.pdf; Facebook, Inc., C-4365 (F.T.C. July 27, 2012) (decision and order), available at http://www.ftc.gov/sites/default/files/documents/cases/2012/08/120810facebookdo.pdf; Google, Inc., C-4336 (F.T.C. Oct. 13, 2011) (decision and order), *available at* http://www.ftc.gov/sites/default/files/documents/cases/2011/10/111024googlebuzzdo.pdf; Twitter, Inc. C-4316

which we believed violated the law by spamming consumers,[16] installing spyware on their computers,[17] failing to secure consumers' personal information,[18] deceptively tracking consumers online,[19] violating children's privacy,[20] and inappropriately collecting information on consumers' mobile devices.[21] Most importantly, the broad reach and remedial focus of Section 5 allows the FTC to protect consumers from harm as new technologies and business practices emerge. I'd like to spend a moment or two explaining how my agency has done this, because Section 5 enforcement is an important part of the "solid land" that you should know about when considering privacy and security aspects of data science.

Some of our cases in this arena have been pretty straightforward. A company *said* in its privacy policy or elsewhere, that it would do one thing; but it actually *did* something else.[22] We have also brought cases against companies that deceptively *omit* information about their data collection and use practices. One example is the FTC's action against the vendor of an app that turned the LED on a mobile phone into a flashlight. But we believed the flashlight app was collecting precise geolocation information, along with a number that uniquely identified consumers' phones. The company's privacy policy disclosed that the app collected data for "product support" and similar purposes, but inappropriately failed to mention the collection of this more sensitive information.[23]

(F.T.C. Mar. 2, 2011) (decision and order), *available at* http://www.ftc.gov/sites/default/files/documents/cases/2011/03/110311twitterdo.pdf.

[16] *See, e.g.*, FTC v. Flora, 2011 U.S. Dist. LEXIS 121712 (C.D. Cal. Aug. 12, 2011), *available at* http://www.ftc.gov/os/caselist/1023005/110929loanmodorder.pdf.

[17] *See, e.g.*, FTC v. CyberSpy Software, LLC, *et al.*, No. 08-CV-01872 (M.D. Fla. Apr. 22, 2010), (stipulated final order), *available at* http://www.ftc.gov/os/caselist/0823160/100602cyberspystip.pdf.

[18] *See* FTC v. Bayview Solutions, LLC, Case 1:14-cv-01830-RC (D.D.C. Aug. 27, 2014), *available at* http://www.ftc.gov/system/files/documents/cases/111014bayviewcmp.pdf *and* FTC v. Cornerstone and Co., LLC, Case 1:14-cv-01479-RC (D.D.C. Aug. 27, 2014), *available at* http://www.ftc.gov/system/files/documents/cases/141001cornerstonecmpt.pdf. The courts in both cases have entered preliminary injunctions against the defendants.

[19] *See, e.g.*, Epic Marketplace, Docket No. C-4389 (F.T.C. Mar. 19, 2013), *available at* http://www.ftc.gov/sites/default/files/documents/cases/2013/03/130315epicmarketplacedo.pdf

[20] *See, e.g.*, United States v. Artist Arena, LLC, No. 12-CV-7386 (S.D.N.Y. Oct. 3, 2012) (stipulated final order), *available at* http://www.ftc.gov/os/caselist/1123167/121003artistarenadecree.pdf.

[21] *See* United States v. Path, Inc., No. 13-CV-0448 (N.D. Cal. Feb. 8, 2013) (consent decree and order), *available at* http://www.ftc.gov/os/caselist/1223158/130201pathincdo.pdf; HTC America, Inc., C-4406 (F.T.C. June 25, 2013) (decision and order), *available at* http://www.ftc.gov/sites/default/files/documents/cases/2013/07/130702htcdo.pdf.

[22] *See, e.g.*, *In re GeoCities, Inc.*, 127 F.T.C. 94 (1999) (consent order) (settling charges that website had misrepresented the purposes for which it was collecting personally identifiable information from children and adults); *FTC v. Toysmart.com, LLC*, No. 00-11341-RGS, 2000 WL 34016434 (D. Mass. July 21, 2000) (consent order) (challenging website's attempts to sell children's personal information, despite a promise in its privacy policy that such information would never be disclosed).

[23] Goldenshores Techs., LLC, C-4466 (F.T.C. Mar. 31, 2014) ¶¶ 11-12 (complaint), *available at* http://www.ftc.gov/system/files/documents/cases/140409goldenshorescmpt.pdf.

The FTC has also used Section 5 to address data collection irrespective of specific representations to consumers. In 2013, for example, the FTC brought an action against a firm that developed software for rent to own companies to install on computers they offered to consumers, to disable the computer if the consumer failed to make timely payments, or the computer was stolen. An add-on feature for the software, called "Detective Mode," allowed the rent-to-own companies to log keystrokes and capture screenshots of confidential and personal information such as user names and passwords, social media interactions and transactions with financial institutions. It also allowed the rent to own companies to take pictures of anyone within view of the computer's webcam, all without even alerting consumers to the existence of the software.[24] We believed that collecting this deeply personal information was harmful to consumers, and therefore unfair.[25]

Data security is a large part of our enforcement program. Over the past 13 years, we have brought 55 cases involving companies that we believed failed to engage in reasonable data security practices. The FTC's initial data security enforcement efforts focused on the financial harms that consumers could suffer when their Social Security numbers or information about their credit cards or bank accounts fell into the wrong hands.[26] But we also focus on security lapses that expose other types of sensitive personal information,[27] including medical information,[28] pharmaceutical records,[29] and our social contacts.[30]

[24] DesignerWare, LLC, C-4390 (F.T.C. Apr. 11, 2013), at ¶ 14 (complaint), *available at* http://www.ftc.gov/sites/default/files/documents/cases/2013/04/130415designerwarecmpt.pdf . The Commission also settled an action against the rent-to-own company that used the software and its franchisees.

[25] An unfair act or practice is one that "causes or is likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition." 15 U.S.C. § 45(n).

[26] *See, e.g.*, The TJX Cos., Inc., No. C-4227 (F.T.C. July 29, 2008) (consent order), *available at* http://www.ftc.gov/enforcement/cases-and-proceedings/cases/2008/08/tjx-companies-inc-matter; Dave & Buster's, Inc., No. C-4291 (F.T.C. May 20, 2010) (consent order), *available at* http://www.ftc.gov/enforcement/cases-and-proceedings/cases/2010/06/dave-busters-incin-matter; DSW, Inc., No. C-4157 (F.T.C. Mar. 7, 2006) (consent order), *available at* http://www.ftc.gov/enforcement/cases-and-proceedings/cases/2006/03/dsw-incin-matter; *BJ's Wholesale Club, Inc.*, No. C-4148 (F.T.C. Sept. 20, 2005) (consent order), *available at* http://www.ftc.gov/enforcement/cases-and-proceedings/cases/2005/09/bjs-wholesale-club-inc-matter.

[27] *See* HTC America, Inc., C-4406 (F.T.C. June 25, 2013) (decision and order), *available at* http://www.ftc.gov/sites/default/files/documents/cases/2013/07/130702htcdo.pdf.

[28] *See* GMR Transcription Servs., No. C-4482 (F.T.C. Aug.14, 2014) (consent order), *available at* http://www.ftc.gov/system/files/documents/cases/140821gmrdo.pdf.

[29] *See* FTC, Press Release, Rite Aid Settles FTC Charges That It Failed to Protect Medical and Financial Privacy of Customers and Employees (July 27, 2010), *available at* http://www.ftc.gov/news-events/press-releases/2010/07/rite-aid-settles-ftc-charges-it-failed-protect-medical-and; FTC, Press Release, CVS Caremark Settles FTC Charges: Failed to Protect Medical and Financial Privacy of Customers and Employees; CVS Pharmacy Also Pays $2.25 Million to Settle Allegations of HIPAA Violations (Feb. 18, 2009), *available at* http://www.ftc.gov/news-events/press-releases/2009/02/cvs-caremark-settles-ftc-chargesfailed-protect-medical-financial.

[30] *See* Snapchat, Inc., No. C-4501 (F.T.C. Dec. 23, 2014), at ¶¶ 34-45 (complaint), *available at* http://www.ftc.gov/system/files/documents/cases/141231snapchatcmpt.pdf.

So the "solid land" of privacy and data security under Section 5 extends pretty far. The FTC's enforcement actions make it clear that the law's prohibitions on deception and unfairness, which have been in force since 1938, apply to personal data collected offline and online, from "old" technologies like PCs and laptops as well as from apps, smartphones, and connected devices.

Line of Sight Navigation: Fair Information Practice Principles

As researchers and members of industry working on cutting-edge problems, you are probably thinking beyond the bare minimum of what you should do to stay on the right side of the law. You might wonder, for example, whether you're dealing with sensitive information; or, if you know that you are, what privacy safeguards you ought to put in place as you work with sensitive data.

A set of Fair Information Practice Principles (FIPPs), which have been part of the privacy landscape for at least four decades, are invaluable in analyzing the challenges that come up in big data analytics today. The question we need to ask is not *whether* they apply, *how* but they apply.

In 2012, the FTC set out a framework for thinking about how these principles can be applied in our data-intensive, highly connected world. The FTC's framework has three basic elements: privacy by design, effective transparency, and simplified consumer choice. These three elements incorporate many of the individual principles that are part of the FIPPs, including data minimization, data security, access, and accuracy. More recently, the FTC also recommended practicing *security* by design, which includes making security part of the design of products and services; testing products for vulnerabilities before shipping or deploying them; setting secure defaults on technologies that consumers buy or use; training personnel to handle personal information properly; employing a range of security measures to establish defense-in-depth; securing device functionality as well as data; and monitoring for vulnerabilities to devices throughout their lifecycles.[31] These principles are scalable. Whether you're a start-up working in the consumer financial analytics space or a global corporation, the FTC's framework is flexible enough to help guide your data practices.

And the framework we recommend focuses on an additional important concept – sensitive information. Some types of data are more deeply personal or potentially more damaging than others. Information about health, finances, precise location, and children are some of the types of information that we have identified as sensitive.[32] Sensitive data deserves

[31] FTC, INTERNET OF THINGS: PRIVACY & SECURITY IN A CONNECTED WORLD 19-22 (2015) (staff report), *available at* http://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf (discussing views of workshop participants) [IOT REPORT].

[32] FTC, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICYMAKERS 58-60 (2012), *available at* http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf (noting "general consensus" about the sensitivity of these four categories of personal information but also noting that other categories may be sensitive) [2012 PRIVACY REPORT].

stronger protections than other kinds of personal data, including more rigorous security and more robust notice and choice before collection. For instance, the FTC recommends that companies obtain affirmative express consent before collecting sensitive information. This means, at minimum, giving consumers a clear and prominent disclosure about what information the entity plans to collect and how it plans to use the information, and only collecting the information after a consumer has opted in.[33] This level of consent helps consumers understand what data is being collected and what a company or researcher will do with it.

Consider health information. Health information is inherently sensitive. Just having it fall into the wrong hands can cause profound embarrassment, harm an individual's job and other economic prospects, or reveal information about family members. Health information can also say a lot about what we do, how we live, and who we are socially and genetically. Our friends, family members, and employers – current or prospective – might look at us quite differently if they knew our blood pressure, cholesterol, glucose levels, risks for long-term disease, or other facts about our health.

HIPAA, the federal law that protects personal health information, is a long-standing piece of our landscape of privacy and data security protection.[34] But HIPAA mainly covers traditional health care providers and insurers, and their business associates.[35] Health information is now flowing in many more places than doctors' offices and hospitals. Wearable devices can measure, record, and transmit our heart rates and how much oxygen we have in our bloodstreams. Apps encourage us to record details about what we eat and how long (and hard) we worked out. Some of the most exciting prospects for society-changing innovations come from wearable devices and mobile apps that encourage consumers to collect and store their own health data.[36] This information may be just as detailed and sensitive as what your doctor collects, but it generally falls outside the current boundaries of HIPAA because of *who* collects it and *where* it is collected.

Some companies are using health information from apps and wearables in surprising ways that are at odds with consumer trust. FTC staff recently reviewed twelve health-related

[33] *See id.* at 57 n.274. *See also, e.g.*, DesignerWare, LLC, No. C-4390 (F.T.C. Apr. 11, 2013) (decision and order), at § II.A, *available at* https://www.ftc.gov/sites/default/files/documents/cases/2013/04/130415designerwaredo.pdf (requiring the respondent to give consumers "an equally clear and prominent choice to either agree or not agree to any geophysical location tracking technology [which was at issue in this action], and neither option may be highlighted or preselected as a default setting" and prohibiting the respondent from using "any geophysical location tracking technology" until such consent is obtained).

[34] Health Insurance Portability and Accountability Act, Pub. L. No.104-191, 110 Stat. 1936 (1996) (codified in scattered sections of 18, 26, 29, and 42 U.S.C.).

[35] *See* Dept. of Health and Human Svcs., Health Information Privacy – For Covered Entities and Business Associates, *available at* http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/ (last visited Mar. 30, 2015).

[36] *See, e.g.*, Elizabeth Whitman, *Apple ResearchKit: Is New Open-Source Software for Sales or the Greater Good of Health Care*, INTL. BUS. TIMES (Mar. 16, 2015 3:51 PM), *available at* http://www.ibtimes.com/apple-researchkit-new-open-source-software-sales-or-greater-good-health-care-1848612.

mobile apps and found that they transmitted information – some of it relating to sensitive health conditions – to seventy-six third parties, including ad networks and analytics firms.[37] For example, one app transmitted health-related search terms, such as "ovulation" and "pregnancy," to third parties. In many instances, third parties received information about consumers' workouts, meals, or diets identified by a real name, email address, or other unique and persistent identifiers.[38]

When it comes to wearables and health apps, one question I am often asked is whether it is too difficult to put privacy safeguards in place. Wearable fitness devices, for example, might not even have a user interface to serve as a means to present consumers with a choice about data collection. Connected devices will become too numerous for consumers to manage their information. And the information generated from these devices is too valuable to consider deleting. In other words, some are arguing that the principles of transparency, individual choice, and data minimization, which have helped organizations navigate their uses of personal data over many years, should not apply to the new world of big data.

I disagree. I urge companies to recognize that individual control and transparency for personal information is an enduring expectation and a much broader concept than simply permitting or refusing information collection at one point in time.[39] Connected devices are no different, but providing transparency and control will require some creative thinking. Immersive apps and websites should be employed to describe to consumers in meaningful and relatively simple terms the nature of the information being collected, and to provide consumers with choices about whether any of this information can be used by entities or persons who fall outside the context in which the consumer is employing the device, and in which the consumer expects her information to remain private. "Command center" technologies that are being rolled out to help consumers manage their many connected household devices[40] could also be employed to help them understand and navigate the collection and use of their data. I believe that if companies use more creativity, we can make meaningful consumer interfaces like these a reality.

Moreover, I believe that embracing transparency and data subject control will benefit big data analytics and research efforts, particularly where sensitive data are concerned. Transparency and control help to build trust.[41] This is especially important with big data, where

[37] *See* Jared Ho, Comments at Federal Trade Commission Consumer Generated and Controlled Health Data Seminar 26–27 (May 7, 2014), *available at* http://www.ftc.gov/system/files/documents/public_events/195411/2014_05_07_consumer-generated-controlled-health-data-final-transcript.pdf.

[38] *Id.* at 26.

[39] *See, e.g.*, Julie Brill, Commissioner, It's Getting Real: Privacy, Security, and Fairness in the Internet of Things, Keynote Address at Carnegie Mellon University Data Privacy Day (Jan. 28, 2015), available at https://www.ftc.gov/system/files/documents/public_statements/621381/150128dataprivacyday.pdf.

[40] *See* Don Clark, *The Race to Build Command Centers for Smart Homes*, WALL ST. J. (Jan. 5, 2015), *available at* http://www.wsj.com/articles/the-race-to-build-command-centers-for-smart-homes-1420399511.

[41] *See* Yaniv Erlich, James B. Williams, David Glazer, Kenneth Yocum, Nita Farahany, Maynard Olson, Arvind Narayanan, Lincoln D. Stein, Jan A. Witkowski, and Robert C. Kain, *Redefining Genomic Privacy: Trust and Empowerment*, PLOS BIOLOGY (Nov. 4, 2014), *available at*

the amount of linkable data in the hands of researchers can allow them to personally identify their data subjects. For example, a study performed by Yaniv Erlich, a computer scientist who is now on the faculty of Columbia, and his fellow researchers, showed that you can identify a man based on part of the DNA sequence from his Y chromosome, his age, and his state of residence.[42] Studies involving credit card transactions,[43] geolocation,[44] hospital discharge data,[45] and other types of data have reached similarly jarring conclusions. This is why appropriate technical measures are only one part of the FTC's recommendations governing deidentification. We also recommend that the companies that control deidentified data publicly commit to refrain from attempting to reidentify the data, and that companies ensure any downstream data recipients also agree to not reidentify the data. Together, these technical and accountability measures will reinforce the trust that consenting data subjects place in companies and researchers.[46]

Now let me turn to data minimization. Data minimization is often presented as an obstacle to some of the most tantalizing portions of the big data program. Big data sets can allow us to see interesting correlations, or make significant predictions from seemingly humdrum data. Critics of data minimization argue that data scientists won't be able to find correlations or make predictions unless they have free rein to explore data. They argue that just collecting data without using it any further is not harmful at all. Furthermore, in the critics' view, data analysts will be able to distinguish between beneficial and harmful uses, so there is no need to worry about how much data companies collect. Their answer to data minimization is a "use-based framework" that limits high-risk data uses, rather than scaling back data collection.

Certainly, it is quite helpful for companies to examine how they use personal data, identify specific risks to individuals, and try to minimize them. But I don't believe that we can rely entirely on such "use-based" frameworks to protect consumers' privacy. A lack of transparency is one drawback to a use-based model. Unless permissible uses have been specified in a way that's accessible to everyone – in legislation, for example – it will be difficult or impossible for consumers to understand what's happening with their data.[47] In other words,

http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1001983 (recommending three principles to govern data sharing in genetic research: (1) transparency creates trust; (2) increased control enhances trust; and (3) reciprocity maintains trust).

[42] Melissa Gymrek, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich, *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321 (2013).

[43] *See* Yves-Alexandre de Montoyje, Laura Radelli, Vivek Kumar, and Sandy Pentland, *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*, 347 SCIENCE 536 (2015).

[44] Richard Becker et al., *Human Mobility Characterization from Cellular Network Data*, 56. COMM, OF THE ACM 74 (2013).

[45] *See, e.g.*, theDataMap, *available at* http://thedatamap.org/ (last visited Mar. 30, 2015) (depicting health data flows among different types of entities).

[46] *See* Jennifer Couzin-Frankel, *Trust Me, I'm a Medical Researcher*, 347 SCIENCE 501, 503 (quoting law professor Mark Rothstein: "[Individuals are not hung up on privacy so much as autonomy. Let's assume that you've de-identified, anonymized, and nobody can figure out who it is – but if people think you've used that information without their permission, they're still going to be very angry.").

[47] *See* IOT REPORT, *supra* note 32, at 42.

giving up on data minimization could start to chip away at other privacy protections, undermine consumers' trust, and ultimately hinder big data analysis efforts.

Focusing solely on the risks and benefits of data uses could also lead companies to ignore the risks created by data collection and retention on their own.[48] One such risk is that the vast amounts of data that companies collect will become an attractive target for hackers, and the risk of harm to consumers from a security breach increases along with the amount of data that companies store.[49] Just as we don't know what benefits might lie undiscovered in big data sets, so too we cannot realistically say that we understand the harms that may occur when the same data is in the hands of a determined adversary. But we do know that you can't lose what you don't have, and so you can't have a security breach of data that you don't collect in the first place.

Another risk is that companies will collect lots of sensitive information about consumers, or infer it from other data that they collect.[50] For example, the *Wall Street Journal* reported back in 2013 that some companies are using data from consumer marketing profiles, such as the type of car a consumer drives and his cable TV subscriptions, to infer whether the consumer is obese or potentially interested in a clinical trial of a drug for treating diabetes.[51] Even if companies don't make further use of such sensitive data, I believe that its collection or creation through inference is something that consumers will want to know about and be able to control.

Dead Reckoning: Big Data and Fairness

I don't want to leave you with the impression that a tried and tested framework is ready to address every consumer protection challenge that will arise with big data analytics. There are some emerging areas where the basic principles are incomplete or unclear. The questions in these areas go beyond privacy to encompass broader questions of fairness. In these areas, we find ourselves, if not on the trackless ocean, then perhaps reckoning our position without as much navigational detail as we would like.

Companies are reaching further for data that could shed light on individual traits and characteristics. Much of this individual-level analysis is done in the context of "marketing," but that label underplays some of what's at stake. For example, in our May 2014 report on data brokers, we detailed how the vast amounts of data that are available about each of us can be used to create alarmingly detailed profiles.[52] These profiles can tell marketers a great deal about where we live, where we work, how much we earn – as well as our daily activities (both offline

[48] *Id.* at 43.

[49] *See id.*

[50] *See id.*

[51] *See* Joseph Walker, *Data Mining to Recruit Sick People*, Wall St. J. (Dec. 17, 2013), *available at* http://online.wsj.com/news/articles/SB10001424052702303722104579240140554518458.

[52] *See generally* FTC, Data Brokers: A Call for Transparency and Accountability (2014), *available at* http://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf (defining "data broker") [Data Broker Report].

and online), and our interests. But they can also contain inferences about more sensitive attributes, such our race, our health conditions, and our financial status. Data brokers may describe us as "Financially Challenged" or perhaps having a "Bible Lifestyle."[53] They may place us in a category of "Diabetes Interest" or "Smoker in Household."[54] Some of them sell marketing lists that identify consumers with addictions or AIDS. Others focus on ethnicity and finances, creating consumer lists such as "Metro Parents" (single parents who are "primarily high school or vocationally educated" and are handling the "stresses of urban life on a small budget") and "Timeless Traditions" (immigrants who "speak[] some English, but generally prefer[] Spanish").[55]

Marketing based on such profiles could benefit consumers. For example, banks might target "Financially Challenged" consumers with offers for safe, low-cost banking products as an alternative to high-cost options like check cashing services and payday loans. But those high-cost lenders could just as easily use the same data. This is, in some sense, "just marketing," but it involves a combination of precision and financial impact that could harm low-income and other vulnerable consumers by encouraging them to take on high-interest debt that can deepen their financial distress.

The same data that fuels marketing based on individual consumer profiles can also be used for more substantive decisions about consumers. An increasing range of algorithmic scores and decisions are part of so-called "risk mitigation" services and other potentially significant decisions about consumers. These services answer questions like "Is this consumer who she claims to be?" and "Is the purchase that this consumer is attempting to make likely to be fraudulent?" While some uses of these "risk mitigation" scores may fall under existing consumer protection statutes, such as the Fair Credit Reporting Act (FCRA), an important set of them does not.[56]

In these circumstances, consumers do not have the right to know when their profiles are being used to reach adverse decisions about them or to dispute and correct inaccurate information in those profiles. In addition, a lot remains unknown about how big data-driven decisions may or may not use factors that are proxies for race, sex, or other traits that U.S. laws generally prohibit from being used in a wide range of commercial decisions.

This spectrum of data-driven decision-making will be driven by the availability of even more data as the Internet of Things and other technologies develop, and they will certainly take advantage of advances in algorithms to make more precise predictions and inferences. What can be done to make sure these products and services –and the companies that use them – treat consumers fairly and ethically?

[53] *Id.* at 20 n.52, 21.

[54] *Id.* at 46, 55.

[55] *Id.* at 20 n.52.

[56] *E.g.*, the Fair Credit Reporting Act, 15 U.S.C. § 1681 *et seq.*

I believe that we need to begin with more transparency and accountability, and everyone who participates in the very broad range of settings that I have discussed – companies, technologists, consumers, and policymakers – has a role to play.

Let me begin with consumers. Consumers should be able to exercise appropriate control over information that goes into the pipelines that feed the algorithms that end up having an effect on their lives, particularly where the pipelines are not visible to consumers. I have long urged data brokers and similar firms to give consumers tools so they can tell consumers that they do not want to have their information used for marketing purposes. Consumers should also have the ability to correct information that is used for risk mitigation and other comparably substantive decisions. And these tools should be immersive, with intuitive UIs, so consumers can easily exercise this control. The FTC's data broker report, as well as the White House's big data review, included my recommendations along these lines.[57]

Some in industry are taking steps to provide greater transparency and control to consumers, but we have a very long way to go here. Ultimately, I believe we need legislation to address these issues, but industry can and should do more right now to make these tools available to consumers.

Of course, transparency is not the whole answer, because consumers cannot navigate this complex ecosystem themselves. Responsible data brokers and analytics firms should recognize that they are the Pole Stars in this ecosystem, and they should strive to ensure accountability throughout their data supply chain. They should examine carefully their own practices, as well as the practices of the companies that feed into their pipeline, and those customers who use the output at the end of the pipeline.

Companies should also do more to determine whether their own data analytics result in unfair, unethical, or discriminatory effects on consumers. For example, what if a company analyzing its own data, in an effort to identify "good" versus "troublesome" customers, ends up tracking individuals along racial or ethnic lines. A *Harvard Business Review* article argues that this kind of result isn't just possible, but inevitable.[58] I believe legislation will be necessary to ensure that all companies – the most scrupulous ones and those that would rather remain in the shadows – are held to the same standards. But companies should not wait for legislation to start tackling these issues.

Of course, you data scientists and technologists in the audience have a key role to play. You have the technical insights that are necessary to determine whether specific analytics practices pose risks of excluding, or otherwise placing at a disadvantage, groups defined according to sensitive traits. You also have the skills to make data access tools that are easy for

[57] *See* DATA BROKER REPORT, *supra* note 45, at 49-54; EXEC. OFFICE OF THE PRESIDENT, BIG DATA: SEIZING OPPORTUNITIES (May 2014), *available at* http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[58] *See* Michael Schrage, *Big Data's Dangerous New Era of Discrimination*, HARVARD BUSINESS REVIEW BLOG NETWORK (Jan. 29, 2014, 8:00 a.m.), http://blogs.hbr.org/2014/01/big-datas-dangerous-new-era-of-discrimination/.

consumers to use. Moreover, you have perspectives that need to be part of policy discussions. I suspect that many of the goals that I am setting for fair treatment of consumers cannot be reduced to decision-making algorithms or automated monitoring of the outputs of those algorithms. Your knowledge of what is and is not feasible from a technical perspective is crucially important to sound policymaking. Bringing your perspectives to the policy arena may not come naturally, but it's critically important that policymakers hear from you, and work with you.

* * * * * *

Examining personal data intensively and on a large scale is integral to many of your research and commercial ventures. Like any efforts that expand the boundaries of our knowledge, your research will raise lots of difficult questions. The questions may vary depending on whether you're developing decision-making algorithms on Wall Street, trying to make city services more efficient, or analyzing data from wearables to gain insights into certain diseases. To answer these questions, we might need new guidelines, principles, or best practices. I am committed to working with academics, industry, and others to help develop them when the need arises. In the meantime, I hope you'll recognize the value of applying the privacy and fairness principles that have served pretty well over decades. These principles will strengthen the trust of individuals who contribute data and, in doing so, will advance the missions that you are all undertaking. Thank you.