August 15, 2014

Mr. Donald S. Clark
Federal Trade Commission
Office of the Secretary
Room H-113 (Annex J)
600 Pennsylvania Avenue, NW
Washington, DC 20580

Re: Big Data: A Tool for Inclusion or Exclusion? Workshop, Project No. P145406

Dear Mr. Clark,

On behalf of the Center for Data Innovation (www.datainnovation.org), I am pleased to submit these comments in response to the Federal Trade Commission's (FTC) request for public comment on big data and inclusivity in advance of its upcoming workshop.[1]

The Center for Data Innovation is a non-profit, non-partisan, Washington, D.C.-based think tank focusing on the impact of the increased use of information on the economy and society. In addition, the Center formulates and promotes pragmatic public policies designed to enable data-driven innovation in the public and private sectors, create new economic opportunities, and improve quality of life. The Center is affiliated with the Information Technology and Innovation Foundation (ITIF).

As we explain in these comments, organizations have always used predictive models; however, until recently these models were often quite rudimentary or relied on limited data sets.[2] These limitations result in inefficiencies for businesses, higher costs for consumers, and lower overall consumer experiences. While no predictive model is perfect and models vary in the levels of accuracy, precision, bias, and confidence they offer, big data offers many opportunities to create much improved predictive models that allow organizations to better tailor their products and services to meet individual needs.

[1] "Big Data: A Tool for Inclusion or Exclusion," Federal Trade Commission, Accessed August 14, 2014, http://www.ftc.gov/es/node/313371.
[2] Daniel Castro, "Predictive Analytics and How to Decide Who Should Receive Organ Transplants," Center for Data Innovation, July 8, 2013, http://www.datainnovation.org/2013/07/predictive-analytics-and-how-to-decide-who-should-receive-organ-transplants/.

HOW ARE ORGANIZATIONS USING BIG DATA TO CATEGORIZE CONSUMERS?

Organizations have always used methods to categorize consumers (e.g. "new customer" versus "returning customer"); however, the advent of big data allows them to do this on a much greater scale. Organizations may use analytical models to predict either categorical values (e.g. "qualifies for a loan") or continuous values (e.g. "credit score of 650"). Although the statistical methods behind classification algorithms differ from predictive algorithms, organizations across almost every industry—including health care, financial services, and insurance—use both types of analytical tools to customize their interactions with consumers and offer personalized products and services.

In health care, predictive models are used routinely. For example, there are well-established predictive models for life expectancy, as well as more modern inventions such as an individual's genetic predisposition to disease, likelihood of hospital readmission, and likelihood of adhering to a treatment plan.[3] These predictions allow doctors to tailor interventions according to individual patients' characteristics, improving outcomes by dispensing with one-size-fits-all treatments and lowering overall health care costs by reducing readmissions.

In education, standardized test scores are often used by schools as a predictor of students' likelihood of performing well in subsequent grades or education levels.[4] These scores allow educators to prioritize the scarce resources of supplementary tutoring or other remedial programs to the students who most need help. In recent years, higher education officials have begun to use predictive models for students' likelihood of succeeding in particular courses or graduating college. For example, Wichita State University implemented a predictive analytics program to reduce the dropout rate for its students and better identify high-quality applicants. The university found that predictive scores were a significantly more reliable indicator of student success than recommendations made by human reviewers.[5] This means that by using classification algorithms, a university might be able to not only create a fairer admissions system that is firmly rooted in merit, but also improve the quality of its student body.

Insurance providers quantify risk for an extremely broad array of events, including the likelihood that a homeowner will sell, the likelihood a farmer's crop will be wiped out by flood, and the likelihood

---

[3] Alex Guazzelli, "Predictive analytics in healthcare," IBM DeveloperWorks, November 29, 2011, https://www.ibm.com/developerworks/library/ba-ind-PMML3/.
[4] Nathan Kuncel and Sarah Hezlett, "Standardized Tests Predict Graduate Students' Success." *Science* 315 (2007): 1080, http://www.unc.edu/~nielsen/soci708/cdocs/Kuncel_Hezlett_2007_1080.pdf.
[5] Neal Ungerlieder, "Colleges Are Using Big Data To Predict Which Students Will Do Well--Before They Accept Them," Fast Company, October 21, 2013, http://www.fastcoexist.com/3019859/futurist-forum/colleges-are-using-big-data-to-predict-which-students-will-do-well-before-the.

that a particular stock will increase in price.[6] Predictive models in the insurance industry can ensure that safer drivers really do get lower rates and that potential home buyers do not get denied a loan for the wrong reasons. The Climate Corporation, a crop insurance company, creates such precise models based on the likelihood of extreme weather events that it can offer payouts automatically to farmers without requiring them to prove losses. This high level of efficiency allows farmers who grow crops with narrow planting windows to replant before it is too late for that season.[7]

Finally, predictive scores are used in a variety of online recommendation systems, with dating sites modeling how likely it is that two people are romantically compatible, Amazon determining which books a user might like to read, and LinkedIn determining what jobs a user might be interested in.[8] These scores allow companies to offer personalized services, saving consumers' time by obviating the need to slog through enormous numbers of listings to find relevant information and compare alternatives.[9]

WHAT BENEFITS DO CONSUMERS GAIN FROM THESE PRACTICES? DO THESE PRACTICES RAISE CONSUMER PROTECTION CONCERNS? WHAT ARE THE SOCIAL AND ECONOMIC IMPACTS, BOTH POSITIVE AND NEGATIVE, FROM THE USE OF BIG DATA TO CATEGORIZE CONSUMERS?

Predictive models can be used to improve consumer welfare by personalizing products and services, including many applications in fields like health care where these predictive techniques are saving lives. For example, using public data, including news reports, social media, and government sources, a group of researchers from Children's Hospital Boston and Harvard Medical School predicted the current Ebola outbreak more than a week before it was recognized by the World Health Organization.[10] At Carnegie Mellon University, researchers have created a machine-learning algorithm that uses detailed patient data to predict up to four hours in advance patients who are

---

[6] Susanne Sclafane. "Use of Predictive Models Widespread in P/C Insurance: Survey," Insurance Journal, November 7, 2013, http://www.insurancejournal.com/news/national/2013/11/07/310601.htm.
[7] "Agriculture and algorithms," The Economist, November 19, 2012, http://www.economist.com/blogs/schumpeter/2012/11/weather-insurance.
[8] Janet Wagner, "Machine Learning and Predictive Analytics Foster Growth," ProgrammableWeb. February 21, 2014, http://www.programmableweb.com/news/machine-learning-and-predictive-analytics-foster-growth/2014/02/21 and Ryan Neal, "What Is Bright.com? Why LinkedIn Purchased The Job-Matching Website For $120 Million," International Business Times, February 6, 2014, http://www.ibtimes.com/what-brightcom-why-linkedin-purchased-job-matching-website-120-million-1553869.
[9] JP Mangalindan, "Amazon's recommendation secret," CNN Money, July 30, 2012, http://tech.fortune.cnn.com/2012/07/30/amazon-5/.
[10] "Online Tool Nailed Ebola Epidemic," *Politico*, August 9, 2014, http://www.politico.com/story/2014/08/healthmap-ebola-outbreak-109881.html

likely to have a heart attack.[11] In addition, predictive models can often benefit consumers in other ways, such as by increasing privacy, freeing up resources to handle exceptional situations, and rewarding positive behaviors.

First, predictive scores can be used to trigger automated responses, preventing the need to disclose an individual's information to another person. For example, an online store may generate personalized product suggestions automatically, enabling shoppers to avoid discussing their preferences with a storekeeper. Or a health insurance company may send automated reminders to help patients adhere to their doctor's orders, thereby reducing the need for future medical exams. In general, predictive models reduce the need to communicate sensitive information to people who might divulge it later.

Second, using automated classification techniques can help make businesses more productive by reducing the need to manually review data about customers. For example, instead of a loan officer or admissions officer manually processing thousands of applications where the decision is clear, they can instead spend more time reviewing applications for individuals whose characteristics make them statistical outliers. Higher productivity results in savings that are generally passed on to consumers in competitive markets. This shift should also move workers into higher-paying, higher-productivity jobs. In addition, by automating many tasks, organizations can free up limited resources to help individuals most in need. For example, doctors can use a predictive model to learn which of their patients are most likely to benefit from a follow-up call.

Third, predictive modeling makes it easier to reward positive behavior. Individuals may alter their behavior if they know a company is using predictive modeling. While this is sometimes cited as a public policy concern, it is worth noting that many of these changes may be desirable.[12] For example, drivers may better adhere to the speed limit if they know their roadway habits will affect their insurance premiums, or consumers may make more responsible financial decisions in order not to negatively affect their credit rating.

Finally, data-driven automated systems can and should be used to reduce discrimination. For example, a properly tuned home loan evaluation algorithm could be made much fairer than a human loan officer; by definition, a successful loan evaluation algorithm should award loans to the most deserving and ignore unrelated factors like race, religion, or sexual orientation that cloud human

---

[11] Aviva Rutkin, "Machine predicts heart attacks 4 hours before doctors," *New Scientist*, August 11, 2014, ww.newscientist.com/article/mg22329814.400-machine-predicts-heart-attacks-4-hours-before-doctors.html.
[12] Daniel Solove, "'I've Got Nothing to Hide' and other Misunderstandings of Privacy," *San Diego Law Review* 44 (2007): 745.

judgment. Even in areas where automated decisions are the norm today, such as demographic assessment in insurance pricing, data-driven systems could offer improvement; pricing insurance based on the output of a car-bound sensor that measures driving ability is undeniably less discriminatory than pricing it based on the risk associated with a particular demographic profile.

As we discussed in a recent article, the pre-"big data" algorithms used to decide people's fate can often be improved with big data.[13] For example, some organ donation programs are based on how long individuals have been on a waiting list, a problem for disadvantaged populations who receive inadequate information about transplants may not be added to the waitlist until long after their diagnosis. These individuals are at a disadvantage when organs are allocated on a first-come, first-served basis. A better approach might use more data to more fairly allocate organs to individuals based on a number of factors, including their viability.

HOW DO EXISTING LAWS APPLY TO SUCH PRACTICES? ARE THERE GAPS IN THE LEGAL FRAMEWORK?

Numerous existing laws address discrimination of individuals based on various characteristics, and these laws should be applied to discrimination of those characteristics as represented in data. Title VII of the Civil Rights Act of 1964, the Equal Pay Act, the Fair Credit Reporting Act, the Age Discrimination Employment Act, the Americans with Disabilities Act, and the Genetic Information Nondiscrimination Act are just a few examples of laws that contain antidiscrimination provisions. These laws are, in essence, use restrictions on information, so they are naturally applicable to discrimination based on data analytics. They also extend naturally to algorithmically generated models; if a particular variable is prohibited from informing a human's employment decision, it should also be excluded from models that do the same job. To prevent algorithms from using variables that may be proxies for a prohibited variable, such as using an individual's zip code as a proxy for race, enforcement agencies could ensure that organizations not use models that make predictions or recommendations correlating above a certain threshold with membership in a protected class. To the extent that new types of discrimination are identified that are not covered under existing laws, they can be addressed through use restrictions modeled after the earlier legislative efforts.

This is not to say that every problem should be addressed with anti-discrimination laws. A frequently cited example of discrimination from big data comes from a 2013 study by researcher Latanya Sweeney that found the search engine ads for the background check website Instantcheckmate.com

[13] Daniel Castro, "Predictive Analytics and How to Decide Who Should Receive Organ Transplants," Center for Data Innovation, July 8, 2013, http://www.datainnovation.org/2013/07/predictive-analytics-and-how-to-decide-who-should-receive-organ-transplants/.

were disproportionately shown to users searching for "black-sounding names" than "white-sounding names."[14] The recent White House report on big data describes the problem as follows:

> *Unfortunately, "perfect personalization" also leaves room for subtle and not-so-subtle forms of discrimination in pricing, services, and opportunities. For example, one study found web searches involving black-identifying names (e.g., "Jermaine") were more likely to display ads with the word "arrest" in them than searches with white-identifying names (e.g., "Geoffrey"). Outcomes like these, by serving up different kinds of information to different groups, have the potential to cause real harm to individuals.*[15]

Clearly, it is harmful for advertisements to reinforce negative stereotypes about marginalized groups. But given that advertisers frequently come under criticism for accusations of racism, sexism, and ageism, this concern may have less to do with big data than it does with the conduct of advertisers.[16] In addition, changes in privacy laws would not address this harm since the root of the problem in this example was negative stereotypes in advertisements being linked to certain search terms. In this case, the issue was really about the ethics of the advertiser. Although it is unlikely that Congress would pass a law banning this type of activity (since, for better or worse, the 1st Amendment gives advertisers wide latitude in crafting public messages), the advertising industry, as well as individual advertisers and ad platforms, can adopt their own self-regulatory guidelines and company policies banning the negative portrayal of a particular race or ethnicity.

ARE COMPANIES APPROPRIATELY ASSESSING THE IMPACT OF BIG DATA PRACTICES ON LOW INCOME AND UNDERSERVED POPULATIONS? SHOULD ADDITIONAL MEASURES BE CONSIDERED?

Organizations should consider the potential discriminatory impact that the use of predictive tools has on protected classes. In some cases, such provisions may already exist. For example, the Equal Opportunity Employment Commission's (EEOC) Uniform Guidelines on Employee Selection Procedures state that employers should maintain information on whether their employee selection procedures for a particular job has an adverse impact on any protected class and update this

---

[14] Latanya Sweeney, "Discrimination in Online Ad Delivery," Communications of the ACM, May 2013, http://privacytools.seas.harvard.edu/files/privacytools/files/p44-sweeney.pdf.
[15] John Podesta et al., "Big Data: Seizing Opportunities, Preserving Values," Executive Office of the President, May 2014, http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.
[16] Emily Tess Katz, "Feed A Child Apologizes For Allegedly Racist Ad," Huffington Post, July 10, 2014, http://www.huffingtonpost.com/2014/07/10/feed-a-child-ad_n_5573892.html, Stephanie Hayes, "#Letshoesbeshoes Encourages Sexist Clarks Shoe Ads to Enter This Century" June 2014, http://www.bustle.com/articles/28277-letshoesbeshoes-encourages-sexist-clarks-shoe-ads-to-enter-this-century, and Brent Green, "Taco Bell: Serving Up Ageism for Super Bowl Sunday?" Huffington Post, February 4, 2013, http://www.huffingtonpost.com/brent-green/taco-bell-super-bowl-ad_b_2602647.html.

information at least annually for groups that represent a significant portion of the workforce.[17] This can be straightforwardly applied to employee selection procedures that use data. If an employer takes online influence metrics—such as those offered by social media analytics company Klout—into account, and this process has an adverse impact on certain groups because it is correlated with access to the Internet and is not otherwise central to the job requirements, the employer would need to reassess its use of this method.

Industry groups, such as the Digital Advertising Alliance, have begun to develop self-regulatory guidelines specifying how data should be used in advertising, and these types of efforts can serve as a model for future initiatives to establish self-regulatory guidelines in other industries.[18] Self-regulation is generally more responsive to technological innovation than government regulation, thereby better serving rapidly evolving industries.[19] In addition, industry groups can often produce and update codes of conduct faster than government, thereby protecting consumers sooner. For example, the Network Advertising Initiative's code of conduct states that online advertising companies should disclose their data collection, purchase, and use practices in an understandable and prominent manner. In cases where self-regulatory guidelines govern data usage, it is critical to hold companies accountable when they violate their commitments or deceive consumers. Government oversight and enforcement by agencies such as the FTC can supplement these efforts to ensure their effectiveness and accountability.

Policymakers should ensure that predictive models based on positive consumer behaviors do not result in negative feedback to consumers. For example, since it is socially desirable to have people seek treatment for mental health issues, regulators should make sure this type of information is not used to deny someone housing, employment, or life insurance. Fortunately, many laws, including the Civil Rights Act, Americans with Disabilities Act, the Fair Housing Act, the Fair Credit Reporting, and the Genetic Information Nondiscrimination Act, already include restrictions on how data can be used for precisely this reason. To the extent that policymakers find gaps in these laws, they should work to ensure individuals are treated fairly.

Ultimately, predictive models offer substantial benefits to consumers in a broad range of fields, and using predictive analytics is superior to other, less-precise, non-computational forms of prediction

---

[17] Uniform Guidelines on Employee Selection Procedure, Section 15-A, EEOC, 1978, http://www.uniformguidelines.com/uniformguidelines.html.
[18] "Application of Self-Regulatory Principles to the Mobile Environment," Digital Advertising Alliance, July 2013, http://ww.aboutads.info/DAA_Mobile_Guidance.pdf.
[19] Daniel Castro, "Benefits and Limits of Industry Self-Regulation for Online Behavioral Advertising," Information Technology and Innovation Foundation, December 2011, http://www.itif.org/files/2011-self-regulation-online-behavioral-advertising.pdf.

commonly used today. To ensure consumers continue to benefit from this technology, the FTC should work with the private sector to encourage more widespread use of predictive analytics and highlight best practices that protect disadvantaged populations. It should also work to promote laws and regulations that support the use and reuse of data, which is a fundamental enabler of these types of tools. Finally, policymakers should promote the development of tools and techniques to de-identify data, as well as support research on topics like privacy-preserving data mining, so that these tools can be applied to a broader set of data while protecting consumer privacy.

CONCLUSION

In short, as organizations increasingly use advanced data analytics to make decisions, we should remember two things. First, prediction is not something to fear—we have always done it, it will always be imperfect, but technology can help us do it more accurately and in ways that greatly benefits consumers. Second, data analytics should not be a black box—we need to examine its underlying assumptions and values to ensure that protected groups are not treated unfairly. Doing so will enable us to gain the benefits of data while avoiding the tyranny of the algorithm.

Sincerely,

Daniel Castro

Director, Center for Data Innovation
1101 K Street NW, Suite 610
Washington, DC 20005

dcastro@datainnovation.org