Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Executive Office of the President

May 2016

Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights

Contents

PREFACE 4

The Assumption: Big Data is Objective 6

Challenge 1: Inputs to an Algorithm 7

Challenge 2: The Design of Algorithmic Systems and Machine Learning 8

CASE STUDIES IN THE USE OF BIG DATA 10

Big Data and Access to Credit 11

The Problem: Many Americans lack access to affordable credit due to thin or non-existent credit files. 11

The Big Data Opportunity: Use of big data in lending can increase access to credit for the financially underserved. 11

The Big Data Challenge: Expanding access to affordable credit while preserving consumer rights that protect against discrimination in credit eligibility decisions. 12

Big Data and Employment 13

The Problem: Traditional hiring practices may unnecessarily filter out applicants whose skills match the job opening. 13

The Big Data Opportunity: Big data can be used to uncover or possibly reduce employment discrimination. 14

The Big Data Challenge: Promoting fairness, ethics, and mechanisms for mitigating discrimination in employment opportunity. 15

Big Data and Higher Education 16

The Problem: Students often face challenges accessing higher education, finding information to help choose the right college, and staying enrolled. 16

The Big Data Opportunity: Using big data can increase educational opportunities for the students who most need them. 17

The Big Data Challenge: Administrators must be careful to address the possibility of discrimination in higher education admissions decisions. 18

Big Data and Criminal Justice 19

The Problem: In a rapidly evolving world, law enforcement officials are looking for smart ways to use new technologies to increase community safety and trust. 19

The Big Data Opportunity: Data and algorithms can potentially help law enforcement become more transparent, effective, and efficient. 19

The Big Data Challenge: The law enforcement community can use new technologies to enhance trust and public safety in the community, especially through measures that promote transparency and accountability and mitigate risks of disparities in treatment and outcomes based on individual characteristics. 21

LOOKING TO THE FUTURE 22

PREFACE

Big data and associated technologies have enormous potential for positive impact in the United States, from augmenting the sophistication of online interactions to enhancing understanding of climate change to making advances in healthcare. These efforts, as well as the technological trends of always-on networked devices, ubiquitous data collection, cheap storage, sensors, and computing power, will spur broader use of big data. Our challenge is to support growth in the beneficial use of big data while ensuring that it does not create unintended discriminatory consequences.

The Obama Administration's Big Data Working Group released reports on May 1, 2014[1] and February 5, 2015.[2] These reports surveyed the use of data in the public and private sectors and analyzed opportunities for technological innovation as well as privacy challenges. One important social justice concern the 2014 report highlighted was "the potential of encoding discrimination in automated decisions"—that is, that discrimination may "be the inadvertent outcome of the way big data technologies are structured and used." Building on these prior reports and the 2014 study conducted by the President's Council of Advisors on Science and Technology (PCAST), the Administration is further examining how big data is used in the public and private sectors.[3] Specifically, we are examining case studies involving credit, employment, education, and criminal justice to shed light on how using big data to expand opportunity has the potential to introduce bias inadvertently that could affect individuals or groups. As discussed in a report released by the Federal Trade Commission (FTC) earlier this year, big data provides opportunities for innovations that reduce discrimination and promote fairness and opportunity, including expanding access to credit in low-income communities, removing subconscious human bias from hiring decisions and classrooms, and providing extra resources to at-risk students.[4] However, the FTC also emphasized the need to prevent such technologies from being used to deny low-income communities credit, perpetuate long-standing biases in employment, or exclude underserved communities from other benefits and opportunities.[5]

This report examines several case studies from the spheres of credit and lending, hiring and employment, higher education, and criminal justice to provide snapshots of opportunities and dangers, as well as ways that government policies can work to harness the power of big data and avoid discriminatory outcomes. These are issues that strike at the heart of American values, which we must work to advance in the face of emerging, innovative technologies.

CECILIA MUÑOZ
Director
Domestic Policy Council

MEGAN SMITH
U.S. Chief Technology Officer
Office of Science and Technology Policy

DJ PATIL
Deputy Chief Technology Officer for Data Policy and Chief Data Scientist
Office of Science and Technology Policy

THE BIG DATA REVIEW AND RECENT PROGRESS

Civil rights legislation of the last century responded to the reality that some Americans were being denied access to fundamental building blocks of opportunity and security, such as employment, housing, access to financial products, and admission to universities, on the basis of race, color, national origin, religion, sex, gender, sexual orientation, disability, or family status. Today, in the United States, anti-discrimination laws help enforce the tenet that all people are to be treated equally. These safeguards are important to protect all Americans against discrimination. Big data techniques have the potential to enhance our ability to detect and prevent discriminatory harm. But, if these technologies are not implemented with care, they can also perpetuate, exacerbate, or mask harmful discrimination. [6]

In May 2014, a federal working group led by then-Counselor to the President John Podesta released a report on the emerging role of big data in our changing world and the opportunities and challenges that it presents. [7] This was followed by an update on February 5, 2015.[8] These reports highlight how big data tools are already being used to improve lives, make the economy work better, and save taxpayer dollars. In our increasingly networked world, the building blocks of big data are everywhere. We upload messages and photos over social media to stay connected to our friends; our phones transmit our specific locations to transportation apps; and information about who we are and what we are interested in is collected by a wide variety of retail, advertising, and analytics companies. Supplying data to these services enables a greater degree of improvement and customization, but this sharing also creates opportunities for additional uses of our data that may be unexpected, invasive, or discriminatory. As data-driven services become increasingly ubiquitous, and as we come to depend on them more and more, we must address concerns about intentional or implicit biases that may emerge from both the data and the algorithms used as well as the impact they may have on the user and society. Questions of transparency arise when companies, institutions, and organizations use algorithmic systems and automated processes to inform decisions that affect our lives, such as whether or not we qualify for credit or employment opportunities, or which financial, employment and housing advertisements we see. [9]

Ideally, data systems will contribute to removing inappropriate human bias where it has previously existed. We must pay ongoing and careful attention to ensure that the use of big data does not contribute to systematically disadvantaging certain groups. To avoid exacerbating biases by encoding them into technological systems, we need to develop a principle of “equal opportunity by design”—designing data systems that promote fairness and safeguard against

discrimination from the first step of the engineering process and continuing throughout their lifespan.

The 2014 report focused on the emerging role of big data in our changing world and the opportunities and challenges posed by such technology.[10] The report made a series of recommendations to the Federal government, among them to advance several pieces of key legislation that would further individual privacy in the digital economy. The Administration, Congress, and the public have made progress in acting on these recommendations; in 2015 we issued an Interim Progress Report illustrating the steps that have been taken thus far.[11] Further, the Council of Economic Advisers issued a report on price discrimination, exploring the processes that companies use to harvest information to in essence charge different consumers different prices for products and services rendered, and the Federal Trade Commission (FTC) included price discrimination in its analysis of issues relating to big data.[12]

The purpose of this report is to advance our understanding of the problems that we can address with big data and algorithmic systems, along with the challenges that exist in deploying them. This new set of practices has great potential to increase access to opportunity and help overcome discrimination, if fairness in computational systems and ethical approaches to data analytics are the norm. At the same time, there are great risks that the very same innovations could perpetuate discrimination and unequal access to opportunity as the use of data expands. This report examines instances where big data methods and systems are being used in the public and private sectors in order to illustrate the potential for positive and negative outcomes and the extent to which "equal opportunity by design" safeguards may help address harms. These examples are meant to be a snapshot of the problems that data analytics can help to solve and the potential issues that its use might create, rather than an exhaustive look or set of recommendations on avoiding discrimination as big data becomes more central to the work of government and business.

OPPORTUNITIES AND CHALLENGES IN BIG DATA

The Assumption: Big Data is Objective

It is often assumed that big data techniques are unbiased because of the scale of the data and because the techniques are implemented through algorithmic systems. However, it is a mistake to assume they are objective simply because they are data-driven.[13]

The challenges of promoting fairness and overcoming the discriminatory effects of data can be grouped into the following two categories:

1) Challenges relating to ***data used as inputs*** to an algorithm; and

2) Challenges related to ***the inner workings of the algorithm itself***.

Challenge 1: Inputs to an Algorithm

The algorithmic systems of big data employ sophisticated processes. These processes need inputs. Consider how you might use a smart phone or GPS device to get the fastest route to a particular destination. To begin, the device needs at least three inputs: (1) where you are, (2) where you want to go, and (3) a map of the area. It then employs algorithms to calculate the fastest route from (1) to (2) and generate directions for how to proceed. At its most basic, this algorithmic system might simply draw upon map data for its inputs, such as the names, locations, and length of the roads in a given city. To compute the fastest route, the algorithmic system would then calculate which path on the map involved the fewest number of roads with the shortest road length. A more advanced routing system could include maximum speed limits for each road or even information about traffic congestion such as data on the speed of other drivers collected from those drivers' mobile devices. For these more complex algorithmic systems, the inputs extend beyond a simple street map to potentially include many others, such as weather updates, historical traffic patterns, and the presence of disruptive events occurring nearby.

The decision to use certain inputs and not others can result in discriminatory outputs. Some of the technical themes that can cause discriminatory outputs include:

- *Poorly selected data*, where the designers of the algorithmic system decide that certain data are important to the decision but not others. In the "fastest route" example, the architect of the system might only include information about roads but not public transportation schedules or bike routes, thereby disadvantaging individuals who do not own a vehicle. Such issues can be regarded as qualitative errors, where human choices in the selection of certain datasets as algorithmic inputs over others are ill-advised. Careless choices of input might lead to biased results—in the "fastest route" example, results that might favor routes for cars, discourage use of public transport, and create transit deserts. Relatedly, designers might select data that is of too much or too little granularity, resulting in potentially discriminatory effects.

- *Incomplete, incorrect, or outdated data*, where there may be a lack of technical rigor and comprehensiveness to data collection, or where inaccuracies or gaps may exist in the data collected. In the "fastest route" example, this could occur if, for instance, the algorithmic system does not update bus or train schedules regularly. Even if the system works perfectly in other respects, the resulting directions could again discourage use of

public transportation and disadvantage those who have no viable alternatives, such as many lower-income commuters and residents.

- *Selection bias*, where the set of data inputs to a model is not representative of a population and thereby results in conclusions that could favor certain groups over others. In the "fastest route" example, if speed data is collected only from the individuals that own smartphones, then the system's results may be more accurate for wealthier populations with higher concentrations of smart phones and less accurate in poorer areas where smart-phone concentrations are lower.[14]

- *Unintentional perpetuation and promotion of historical biases*, where a feedback loop causes bias in inputs or results of the past to replicate itself in the outputs of an algorithmic system. For instance, when companies emphasize "hiring for culture fit" in their employment practices, they may inadvertently perpetuate past hiring patterns if their current workplace culture is primarily based on a specific and narrow set of experiences. In a workplace populated primarily with young white men, for example, an algorithmic system designed primarily to hire for culture fit (without taking into account other hiring goals, such as diversity of experience and perspective) might disproportionally recommend hiring more white men because they score best on fitting in with the culture.

Each of these issues is critical to take into account in designing systems to deliver services effectively, fairly, and ethically to consumers and community members, or to influence processes like credit-granting, hiring, housing allocation, and admissions.[15] Transparency, accountability, and due process mechanisms are important components of ensuring that the inputs to an algorithmic system are accurate and appropriate.

Challenge 2: The Design of Algorithmic Systems and Machine Learning

For those who are not directly involved in the technical development of algorithms for large scale data systems, the end product of such as system can feel like a "black box"—an opaque machine that takes inputs, carries out some inscrutable process, and delivers unexplained outputs based on that process.[16] The technical processes involved in algorithmic systems are typically unknown to a consumer, potential student, job candidate, defendant, or the public as they are often treated as confidential or proprietary to the entities that use them.[17] Some systems even "passively" pre-screen potential candidates without notice as a preemptive effort to streamline decision-making processes at a later date.[18] This lack of transparency means that

affected individuals—such as those who receive word that they will not receive a job offer, were denied admission to their college of choice, or will be denied a line of credit or lease—have limited ability to learn the reasons why such decisions were made and limited ability to detect and seek correction of any errors or bias if they do occur. It may even mean that certain individuals will be entirely excluded from certain opportunities—for instance, seeing particular advertisements for jobs, financial products, or educational opportunities and never discover that they were denied these opportunities.[19] Such situations can be complex and difficult to address, especially if the outputs are relied upon again in subsequent determinations.[20] At a minimum, it is important to encourage transparency, accountability, and due process mechanisms wherever possible in the use of big data. Without these safeguards, hard-to-detect flaws could proliferate. Such flaws include:

- *Poorly designed matching systems,* which are intended to help find information, resources, or services. For example, search engines, social media platforms, and applications rely on matching systems to determine search results, what advertisements to display, and which businesses to recommend. These matching systems may result in discriminatory outcomes if the system designs are not kept current or do not take into account historical biases or blind spots within the data or the algorithms used.[21]
- *Personalization and recommendation services that narrow instead of expand user options*, where detailed information about individual users might be collected and analyzed to infer their preferences, interests, and beliefs in order to point them to opportunities such as new music to download, videos to watch, price discounts, or products to purchase. Academic studies have shown that the algorithms used to recommend such content may inadvertently restrict the flow of information to certain groups, leaving them without the same opportunities for economic access and inclusion as others.[22]
- *Decision-making systems that assume correlation necessarily implies causation*, whereby a programmer or the algorithmic system itself may assume that because two factors frequently occur together (e.g., having a certain income level and being of a particular ethnicity), there is necessarily a causal relationship between the two. Assuming a causal relationship in these circumstances can lead to discrimination.
- *Data sets that lack information or disproportionately represent certain populations,* resulting in skewed algorithmic systems that effectively encode discrimination because of the flawed nature of the initial inputs. Data availability, access to technology, and participation in the digital ecosystem vary considerably, due to economic, linguistic, structural or socioeconomic barriers, among others. Unaddressed, this systemic flaw can

reinforce existing patterns of discrimination by over-representing some populations and under representing others.

An additional area that presents challenges for further study is a genre of computer science known as machine learning—the "science of getting computers to act without being explicitly programmed."[23] Complex and often inscrutable even at times to their programmers, machine learning models are starting to be used in areas such as credit offers, entrepreneurial funding, or hiring. As these methods continue to advance, it may become more difficult to explain or account for the decisions machines make through this process unless mechanisms are built into their designs to ensure accountability.[24] Using the principle of "equal opportunity by design" and grounding engineering with sound ethical and professional best practices will also help mitigate discriminatory results over time and increase inclusion.

Just as in other areas, programmers and data scientists may inadvertently or unconsciously design, train, or deploy big data systems with biases. Therefore, an important factor in implementing the "equal opportunity by design" principle is engaging with the field of "bias mitigation" to avoid building in the designers' biases that are an inevitable product of their own culture and experiences in life.[25] Research-based methods are emerging that can help reduce biases in decision-making around hiring, promotions, classroom grading, funding, social engagement, and more.[26] Use of these methods can help stop biased big data systems from becoming the norm, instead of the exception.

As improvements in the uses of big data and machine learning continue, it will remain important not to place too much reliance on these new systems without questioning and continuously testing the inputs and mechanics behind them and the results they produce. "Data fundamentalism"—the belief that numbers cannot lie and always represent objective truth—can present serious and obfuscated bias problems that negatively impact people's lives.[27]

As we work to address challenges related to data inputs and the inner workings of algorithms, we must also pay attention to how to the products of these algorithmic systems are used, with an eye for ensuring that information about places, people, preferences, and more is used legally, ethically, and to advance democratic principles, such as equality and opportunity.

CASE STUDIES IN THE USE OF BIG DATA

In this section, we present four case studies in the use of big data analytics: (1) access to credit, (2) higher education, (3) employment, and (4) criminal justice. We describe the opportunities each case study presents for algorithmic systems to support personal, commercial, and

organizational missions, as well as the challenges that arise in utilizing the data without adverse impacts.

Big Data and Access to Credit

The Problem: Many Americans lack access to affordable credit due to thin or non-existent credit files.

Access to fairly-priced and affordable credit is an important factor in enabling Americans to thrive economically, especially those working to enter the middle class. For decades, lenders have largely relied on credit scores, such as the FICO score, to decide whether and on what terms to make a loan. Credit scores represent a prediction of the likelihood that someone will have a negative financial event, such as defaulting on a loan, within a specific period of time. Traditionally, this prediction is made based on actual data about a particular person's credit history, and turned into a score using algorithms developed from past lending experiences. While traditional credit scores serve many Americans well, according to a study by the Consumer Financial Protection Bureau (CFPB), as many as 11 percent of consumers are "credit invisible"—they simply do not have enough up-to-date credit repayment history for the algorithm to produce a credit score.[28] In addition, the CFPB found a strong relationship between income and a scorable credit record—30 percent of consumers in low-income neighborhoods are "credit invisible" and the credit records of another 15 percent are unscorable.[29] According to the CFPB, African-Americans and Latinos are more likely to be credit invisible, at rates of around 15 percent in comparison to 9 percent for whites.[30] The CFPB also found that an additional 13 percent of African-Americans and 12 percent of Latinos are unscorable, compared to 7 percent for whites.[31]

The Big Data Opportunity: Use of big data in lending can increase access to credit for the financially underserved.

One possible approach to this problem is to use data analytics drawing on multiple sources of information to create more opportunity for consumers to gain access to better credit. As companies collect information and score individuals, especially those without sufficient or updated credit information, data may be useful in assessing credit risk. Some companies look at previously untapped data, such as phone bills, public records, previous addresses, educational background, and tax records, while others may consider less conventional sources, such as location data derived from use of cellphones, information gleaned from social media platforms,

purchasing preferences via online shopping histories, and even the speeds at which applicants scroll through personal finance websites.

New tools designed with big data have potential to create alternative scoring mechanisms and new opportunities for access to credit for the tens of millions of Americans who do not have enough information in their credit files to receive a traditional credit score or who have an undeservedly low score. For example, many Americans regularly pay their phone and utility bills—payment records that predict creditworthiness. But phone and utility payment information is generally only reported as part of a credit history if a customer falls far behind, and therefore it only ever serves to penalize a consumer's perceived creditworthiness. One study by the Policy and Economic Research Council looked at more than four million credit files and found that if both positive and negative utility and telecom payments were included, over 70 percent of the unscorable files would become scorable and 64 percent of the "thin files" (files with very little other credit history) would see improved scores.[32] The study also found that this change especially benefits low-income borrowers.[33]

The Big Data Challenge: Expanding access to affordable credit while preserving consumer rights that protect against discrimination in credit eligibility decisions.

While big data has the ability to increase American's access to affordable credit, if not used with care, it also has the potential to perpetuate, exacerbate, or mask discrimination. For example, consider a technology that would glean information from an individual's social media connections and use social analytic systems to create an alternative credit score. While such a tool might expand access to credit for those underserved by the traditional market, it could also function to reinforce disparities that already exist among those whose social networks are, like them, largely disconnected from everyday lending.[34] Such tools could also raise questions about the ability of consumers to dispute an adverse decision or to correct inaccurate information. When such decisions are made within computationally-driven 'black box' systems, traditional notions of transparency may fail to fully capture and disclose the information consumers need to understand the basis of such decisions and the role that various data played in determining their credit eligibility.

The right to be informed about and to dispute the accuracy of the underlying data used to create a credit score is particularly important because credit bureaus have significant data accuracy issues, which are likely to be exacerbated by the use of new, fast-changing data sources. An FTC study found that 21 percent of its sample of consumers had a confirmed error on at least one of their three credit bureau reports.[35] Expanding the data sources for credit scoring systems from long-collected items like collections notices and credit card payments to fast-changing, large-volume data like social media usage and GPS location information would

most likely increase the presence of factually inaccurate data, leading to scores that are not based on a consumers' likelihood of delinquency.[36] Additionally, as the number of data sources increases, the relationship to creditworthiness becomes more complex and dynamic, and therefore consumers may have more difficulty interpreting the notices required under FRCA and ECOA and identifying problems.[37] Consumers with less experience dealing with large institutions or complex data products may be particularly vulnerable to these data accuracy and transparency challenges.

These concerns are not necessarily unique to the emerging data-analytics approaches to credit scoring. For example, in 2007 the FTC released a study of credit-based insurances scores finding that although there are substantial score disparities among ethnic groups, the scores are effective predictors of risk under automobile-insurance policies and are not simply proxies for race, ethnicity, sex, or other prohibited bases.[38] As algorithms develop to measure creditworthiness in new ways, it will be critical to design and test them with similar concerns in mind and to guard against unintentionally using information that is a proxy for race, gender, or other protected characteristics .[39] The limited research that does publicly exist has looked at whether the scores are effective predictors of delinquency, it has not examined whether new ways of evaluating credit worthiness adequately avoid considering proxies for traits like race or ethnicity.[40]

The shortage of studies on these new scoring products is a potential cause for concern because of the complexity and proprietary nature of these new products. If poorly implemented, algorithmic systems that utilize new scoring products to connect targeted marketing of credit opportunities with individual credit determinations could produce discriminatory harms. This is particularly concerning because the rapid pace of evolution in the credit sector, especially combined with ongoing advances in data science, makes it difficult for researchers and consumers alike to identify discrimination and take steps to prevent it.[41]

Big Data and Employment

The Problem: Traditional hiring practices may unnecessarily filter out applicants whose skills match the job opening.

Beginning in the 1990s, a growing number of companies realized there was a new way to access and analyze a larger pool of applicants for a job opening rather than simply reviewing paper files.[42] Resume-database websites provided a place where individuals and companies could gain access to opportunities and talent. To deal with the sudden influx of candidates, companies looking to hire also turned to new ways of rating applicants, using analytical tools to

automatically sort and identify the preferred candidates to move forward in a hiring process. With this change, the task of identifying and scoring applicants began to shift from industrial psychologists and recruiting specialists to computer scientists, through the use of algorithms and large data sets.[43]

Yet even as recruiting and hiring managers look to make greater use of algorithmic systems and automation, the inclination remains for individuals to hire someone similar to themselves, an unconscious phenomenon often referred to as "like me" bias, which can impede diversity.[44] Algorithmic systems can be designed to help prevent this bias and increase diversity in the hiring process. Yet despite these goals, because they are built by humans and rely on imperfect data, these algorithmic systems may also be based on flawed judgments and assumptions that perpetuate bias as well. Because these technologies are new, rapidly changing, difficult to decipher, and often subject to proprietary protections, their determinations can be even more difficult to challenge.

The Big Data Opportunity: Big data can be used to uncover or possibly reduce employment discrimination.

Just as with credit scoring, data analytics can be beneficial to the workplace in helping match people with the right jobs. As discussed above, research has documented a "like me bias" or "affinity bias" in hiring; even well-intentioned hiring managers often choose candidates with whom they share characteristics.[45] By contrast, algorithmically-driven processes have the potential to avoid individual biases and identify candidates who possess the skills that fit the particular job.[46]

Companies can use data-driven approaches to find potential employees who otherwise might have been overlooked based on traditional educational or workplace-experience requirements. Data-analytics systems allow companies to objectively consider experiences and skill sets that have a proven correlation with success. By looking at the skills that have made previous employees successful, a human-resources data system can "pattern match" in order to recognize the characteristics the next generation of hires should have.[47] When fairness, ethics, and opportunity are a core part of the original design, large-scale data systems can help combat the implicit and explicit bias often seen in traditional hiring practices that can lead to problematic discrimination.[48] Beyond hiring decisions, properly deployed, advanced algorithmic systems present the possibility of tackling age-old employment discrimination challenges, such as the wage gap or occupational segregation.[49]

The Big Data Challenge: Promoting fairness, ethics, and mechanisms for mitigating discrimination in employment opportunity.

Data-analytics companies are creating new kinds of "candidate scores" by using diverse and novel sources of information on job candidates. These sources, and the algorithms used to develop them, sometimes use factors that could closely align with race or other protected characteristics, or may be unreliable in predicting success of an individual at a job. For example, workers who were unemployed for long periods during the recent economic downturn may have a harder time re-entering the workforce because candidate-scoring systems that consider "length of time since last job" can generate scores that send negative signals to potential employers that are unrelated to job performance. Similarly, one employment research firm found commuting distance to be one of the strongest predictors of how long a customer service employee will stay with a job.[50] If algorithmic systems were trained to rely heavily on this factor without further consideration, they could end up discriminating against the candidates who, while otherwise qualified, happen to live in areas that are further away from the job than other candidates. While the factor of commuting distance was ultimately disregarded in this particular study out of concern for how highly it might correlate with race,[51] other employers might overlook such important factors. Other common hiring criteria, such as credit-worthiness (also the work of algorithms) and criminal records, compromise the validity of these tools if they inaccurately or inadequately reflect an individual's qualifications. Implementation of such systems with an eye to their broader effects on fairness and equal opportunity is therefore essential.

Finally, as described earlier, machine-learning algorithms can help determine what kinds of employees are likely to be successful by reviewing the past performance of existing employees or by analyzing the preferences of hiring managers as shown by their past decisions.[52] But if those sources themselves contain historical biases, the scores may well replicate those same biases. For example, if machine-learning algorithms emphasize the age that a candidate became interested in computing compared to their peers, cultural messages and assumptions that associate computing with boys more often than with girls could promote environments where more boys than girls are exposed to computers at an earlier age, thereby skewing later hiring patterns toward more male hires, even though a company's hiring goals may be focused on gender equality. [53] Similar concerns could emerge regarding age discrimination, since older workers may be less likely to have grown up with home computers. Further, hiring algorithms that emphasize the need for a four-year college degree, or even a particular field of study or degree can leave out highly qualified, talented individuals who might not have those specific qualifications and could instead come into the job opportunity through on-the-job training or emerging alternative training and apprenticeship models—or who might have a four-year degree but in a different field than the ones sought by the algorithmic systems. These are the

types of factors that engineers and managers need to consider at the inception of designing analytical hiring systems and incorporate into the machine learning design work. "Equal opportunity by design" approaches are one way to promote these considerations.

Companies have begun to filter their applicant pools for job openings using various human resources analytics platforms. It is critical to the fairness of American workplaces that all companies continue to promote fairness and ethical approaches to the use of data tools and ensure against the perpetuation of biases that could disfavor certain groups. Businesses also stand to benefit, because those that do not look beyond historical hiring patterns (even as mediated by an algorithm) will miss great candidates for important jobs.

Big Data and Higher Education

The Problem: Students often face challenges accessing higher education, finding information to help choose the right college, and staying enrolled.

Prospective students and their families must grapple with assessing which of the many institutions of higher education available will best prepare them to achieve their goals. The decisions to pursue a degree, at which degree level, and at which institution all have a lasting impact on students and their futures. For example, obtaining a bachelor's degree can increase total earnings by 84 percent over a lifetime relative to expected earnings for someone with a high school diploma.[54] Similarly, differences in the price of attendance across institutions affect financial returns, and may lead to differences in the amount that students have to borrow, which may also affect their career decisions and personal lives in meaningful ways. Despite the importance of this decision, there is a surprising lack of clear, easy to use, and accessible information available to guide the students making these choices.

At the same time, institutions of higher education collect and analyze tremendous amounts of data about their students and applicants. Before students arrive on campus, colleges use student information in recruitment, admissions, and financial aid decisions. After students enroll, some schools are using the data collected through the application process and in the classroom to tailor their students' educational experiences.[55] The opportunities to use big data in higher education can either produce or prevent discrimination—the same technology that can help identify and serve students who are more likely to be in need of extra help can also be used to deny admissions or other opportunities based on the very same characteristics.

The Big Data Opportunity: Using big data can increase educational opportunities for the students who most need them.

To address the lack of information about college quality and costs, the Administration has created a new College Scorecard to provide reliable information about college performance.[56] The College Scorecard is a large step toward helping students and their families evaluate college choices. Never-before-released national data about post-college outcomes—including the most comparable and reliable data on the earnings of colleges' alumni and new data on student debt—and student-loan repayment provides students, families, and their advisers with a more accurate picture of college cost and value. The data also encourages colleges to strengthen support that helps students persist in and complete college, and to provide increased opportunities for disadvantaged students to get a college education. Armed with this information, students and their families can make more informed decisions and better understand the opportunities and tradeoffs of their choices.

In addition to data that the Department of Education has made available through the College Scorecard, institutions of higher education also present a unique environment to utilize innovations in big data for students once enrolled. Schools can use data they are already collecting to help track student progress. The opportunity for innovation lies in how schools use that existing data to create a tailored learning experience. Big data techniques are already helping students learn more effectively through tailored instruction, which can help overcome continuing disparities in learning outcomes, and providing extra help for those more likely to drop out or fail.[57]

Georgia State University is one example of a college using big data to drive student success. Its Graduation and Progression Success (GPS) Advising program, which started in 2013, is designed to keep the school's more than 32,000 students on track for graduation. It tracks eight hundred different risk factors for each student on a daily basis. When a problem is detected, the university deploys proactive advising and timely interventions to provide the support that students need. At times the interventions are as simple—and essential—as ensuring the student has registered for the right courses; at other times, the system uses predictive analytics to make sure that the student's performance in a prerequisite course makes success likely at the next level. Since the GPS Advising initiative began in 2013, there have been nearly 100,000 proactive interventions with Georgia State students based on the analytics-based alerts coming from the system. Over the past three years, Georgia State's graduation rate has increased by 6 percentage points, from 48 percent to 54 percent, when compared to the baseline year. The biggest gains have been enjoyed by at-risk populations. This year for the first time in Georgia State's history, first-generation, black, and Latino students as well as those on federally-funded Pell grants all graduated at rates at or above that of the student body overall, and Georgia State

now awards more Bachelor's degrees to black students than any non-profit college or university in the United States. Over the last two-years alone, Georgia State has reduced time-to-degree by an average of half a semester per student, saving students more than $10 million in tuition and fees.[58]

The Big Data Challenge: Administrators must be careful to address the possibility of discrimination in higher education admissions decisions.

Though data can help high school students choose the right college, there are several challenges involved in accurately estimating the extent to which the specific school a student attends makes causal contributions to student success. One important limitation of Federal data sources is the lack of individual student-level data indicating academic preparation for college, such as their high-school GPA or college admissions test scores (e.g., SAT or ACT scores). Since academic preparation is an important element that adds context to measures of college quality, omitting this variable may bias estimates of college quality. As the College Scorecard continues to be refined and developed, these are challenges that the Department of Education will continue to face.

In making admissions decisions, institutions of higher education may use big data techniques to try to predict the likelihood that an applicant will graduate before they ever set foot on campus.[59] Using these types of data practices, some students could face barriers to admission because they are statistically less likely to graduate. Institutions could also deny students from low-income families, or other students who face unique challenges in graduating, the financial support that they deserve or need to afford college. This, in turn, creates a concern that as schools rush to cut costs, some applicants might face greater barriers to admission if they are considered unworthy of the extra resources it would take to keep them enrolled.[60] One significant predictor of whether or not a student will graduate from college is family income, and the use of big data in this case may discriminate against students from lower-income families.[61] The same data used to help students succeed can also be used to discourage low-income students from enrolling, and while there may be ways to mitigate this, such as financial incentives, especially at less selective institutions, it remains a cautionary example of using data to perpetuate discrimination.

On the other hand, some schools and states are actively using data to promote access and success, and to prevent discrimination. For example, the State of Tennessee's outcomes-based funding formula for four-year institutions offers an illustration of how data can promote both student success and access.[62] Tennessee's model places extra value on the "credit accumulation" and "degree attainment" outcomes of both students eligible for Pell grant funding and adult students (those over the age of 24).[63] In particular, these outcomes are valued 40 percent more than the same outcomes for non-Pell grant eligible traditional-age

students.[64] By doing this, institutions have an incentive to enroll and promote the success of lower-income and adult students, who are traditionally under-represented in higher education.[65]

There are valuable opportunities for the use of big data in higher education, but they must be implemented with care. As learning itself is a process of trial and error, it is particularly important to use data in a manner that allows the benefits of those innovations, but still allows a safe space for students to explore, make mistakes, and learn without concern that there will be long term consequences for errors that are part of the learning process.

Big Data and Criminal Justice

The Problem: In a rapidly evolving world, law enforcement officials are looking for smart ways to use new technologies to increase community safety and trust.

Local, state, and federal law enforcement agencies are increasingly drawing on data analytics and algorithmic systems to further their mission of protecting America. Using information gathered from the field and through the use of new technologies, law enforcement officials are analyzing situations in order to determine the appropriate response. At the same time, law enforcement agencies are expected to be accountable at all times to the communities they serve and will continue to be so in the digital age. Similarly, the technologies that assist law enforcement's decisions and actions should also be accountable to ensure they are used in a thoughtful manner that considers the impact on communities and promotes successful community partnerships built on trust.

The Big Data Opportunity: Data and algorithms can potentially help law enforcement become more transparent, effective, and efficient.

Law enforcement agencies have long attempted to identify patterns in criminal activity in order to allocate scarce resources more efficiently. New technologies are replacing manual techniques, and many police departments now use sophisticated computer modeling systems to refine their understanding of crime hot spots, linking offense data to patterns in temperature, time of day, proximity to other structures and facilities, and other variables. The President's Task Force on 21st Century Policing recommended, among many other steps, that law enforcement agencies adopt model policies and best practices for technology-based engagement that increases community trust and access; work toward national standards on the issue of technology's impact on privacy concerns; and develop best practices that can be

adopted by state legislative bodies to govern the acquisition, use, retention, and dissemination of auditory, visual, and biometric data by law enforcement.[66]

Since the Task Force released its recommendations, the White House and the Department of Justice have been engaged in several initiatives to ensure that the report's recommendations are put into practice across the United States. As part of these efforts, the White House launched the Police Data Initiative to make policing data more transparent and improve community trust.[67] More than 50 police departments throughout the nation have joined in this work to realize the benefits of better technology. Commitments from participating jurisdictions include: increased use of open policing data to build community trust and increase departmental transparency, and use of data to more effectively identify policies that could be improved or officers who may contribute to adverse public interactions so they can be linked with effective training and interventions.[68]

Consistent with these goals, several police departments in the United States have developed and deployed "early warning systems" to identify officers who may benefit from additional training, resources, or counseling to prevent excessive uses of force, citizen complaints and other problems.[69] Using de-identified police data, as well as contextual data about local crime and demographics, these systems are designed to detect the factors most indicative of future problems by attempting to determine behavioral patterns that predict a higher risk of future adverse incidents. Detecting these patterns opens new opportunities to develop targeted interventions for officers to protect their safety and improve police/community interactions.

Separately, some of the newest analytical modeling techniques, often called "predictive policing," might provide greater precision in predicting locations and times at which criminal activity is likely to occur. Research demonstrates that a neighborhood that has recently been victimized by one or more burglaries is likely to be targeted for additional property crimes in the coming days. An analytical method known as "near-repeat modeling" attempts to predict crimes based on this insight.[70] Similarly, a technique known as "risk terrain modeling" can identify specific locations where criminal activity often clusters, such as bars, motels or convenience stores, and can predict the specific social and physical factors that attract would-be offenders and create conditions ripe for criminal activity.[71] Current Los Angeles Police Department (LAPD) Chief of Police, Charlie Beck, described predictive policing as enabling "directed, information-based patrol; rapid response supported by fact-based prepositioning of assets; and proactive, intelligence-based tactics, strategy, and policy."[72] In some instances these systems have shown significant promise. In experiments conducted by the LAPD's Foothill Division in which large sets of policing data were analyzed to predict occurrences of crime, the Division experienced a larger reduction in reported crime than any other division in the Department.[73]

The Big Data Challenge: The law enforcement community can use new technologies to enhance trust and public safety in the community, especially through measures that promote transparency and accountability and mitigate risks of disparities in treatment and outcomes based on individual characteristics.

When designed and deployed carefully, data-based methodologies can help law enforcement make decisions based on factors and variables that empirically correlate with risk, rather than on flawed human instincts and prejudices. However, it is important that data and algorithmic systems not be used in ways that exacerbate unwarranted disparities in the criminal justice system. For example, unadjusted data could entrench rather than ameliorate documented racial disparities where they already exist, such as in traffic stops and drug arrest rates.[74]

Those leading efforts to use data analytics to create and implement predictive tools must work hard to ensure that such algorithms are not dependent on factors that disproportionately single out particular communities based on characteristics such as race, religion, income level, education, or other data inputs that may serve as proxies for characteristics with little or no bearing on an individual's likelihood of association with criminal activity. For instance, when historical information is used with predictive algorithms to direct patrols, prior arrest data could be used to advise beat officers to patrol certain areas with greater frequency or intensity. If feedback loops are not thoughtfully constructed, a predictive algorithmic system built in this manner could perpetuate policing practices that are not sufficiently attuned to community needs and potentially impede efforts to improve community trust and safety. For example, machine learning systems that take into account past arrests could indicate that certain communities require more policing and oversight, when in fact the communities may be changing for the better over time. Moving forward, law enforcement agencies could work to account for these issues: transparency and accountability on data input and processes, a focus on eliminating data that could serve as proxies for race or poverty, and ensuring that bias is not replicated through these tools are key steps.

It is also important to note that criminal justice data is notoriously poor.[75] This is in part because one of the major data repositories, the Federal Bureau of Investigation's Uniform Crime Report (UCR), is in need of modernization and relies on voluntary contributions that often do not capture data with the degree of richness and completeness needed for in-depth analysis. FBI Director James Comey has prioritized improving data collection and repeatedly called on communities across the United States to increase participation in the UCR's National Incident-Based Reporting System (NIBRS), explaining that this method of collecting data enhances our understanding of crime because "[it] doesn't just include statistics. It gives the full picture–the circumstances and the context involving each incident. It asks: What happened?

Where did it happen? What time did it occur? Who was there and what is their demographic information? What is the relationship between the perpetrator and the victim?"[76]

Even if crime reporting is improved, there will remain reasons to approach any crime dataset with care and caution. Many criminal-justice data inputs are inherently subjective. Officers use discretion in enforcement decisions (e.g., deciding whom to stop, search, question, and arrest) just as police officers and prosecutors use discretion in charging (e.g., simple assault vs. felonious assault). The underlying data reflects these judgement calls.

Policymakers should also continue to look for ways to better use the increasing amount of data that law enforcement agencies now have in order to improve public safety and accountability. For example, the number of agencies presently using body cameras is expanding exponentially. With this expansion comes thousands of hours of video and audio. As part of the Police Data Initiative, the White House has engaged with academics and technologists to determine if there is a machine-readable way to review this video and audio to identify both beneficial and problematic interactions between law enforcement and civilian community members. On December 8, 2015, the White House Office of Science and Technology Policy participated in a workshop co-hosted by Stanford University and the City of Oakland Police Department focused on accelerating research and development to make body-worn camera data more searchable and interoperable with other systems, and on automating processes to reduce reporting burdens.

More broadly, the conversation about ways to effectively use predictive analytics in law enforcement should continue, building on the work that has already begun between key stakeholders—ranging from law enforcement agencies to academics, community leaders and civil society groups.

LOOKING TO THE FUTURE

The use of big data can create great value for the American people, but as these technologies expand in reach throughout society, we must uphold our fundamental values so these systems are neither destructive nor opportunity limiting. Moving forward, it is essential that the public and private sectors continue to have collaborative conversations about how to achieve the most out of big data technologies while deliberately applying these tools to avoid—and when appropriate, address—discrimination. In order to ensure growth in the use of data analytics is matched with equal innovation to protect the rights of Americans, it will be important to:

- ***Support research into mitigating algorithmic discrimination, building systems that support fairness and accountability, and developing strong data ethics frameworks.*** The Networking and Information Technology Research and Development Program and

the National Science Foundation (NSF) are developing research strategy proposals that will incorporate these elements and encourage researchers to continue to look at these issues. Through its support of the Council for Big Data, Ethics, and Society, as well as other efforts, NSF will continue to work with scientific, technical, and academic leaders to encourage the inclusion of data ethics within both research projects and student coursework and to develop interdisciplinary frameworks to help researchers, practitioners, and the public understand the complex issues surrounding big data, including discrimination, disparate impact, and associated issues of transparency and accountability. In particular, it will be important to bring together computer scientists, social scientists, and those studying the humanities in order to understand these issues in their historical, social, and technological contexts.[77]

- ***Encourage market participants to design the best algorithmic systems, including transparency and accountability mechanisms such as the ability for subjects to correct inaccurate data and appeal algorithmic-based decisions***. Big data technologies can support the success of public and private institutions, but to do so, they must be implemented in a responsible and ethical manner. Organizations, institutions, and companies should be held accountable for the decisions they make with the aid of computerized decision-making systems and technology. The FTC's recent big data report included considerations for companies using big data techniques, discussed potentially applicable laws, and suggested questions for legal compliance.[78] Private companies using data analytics to expand opportunity should take these considerations into account in order to ensure that they treat consumers, students, job candidates, and the public fairly. Both private and public entities should also consider improved methods of providing individuals and communities with the means to access and correct their data, as well as better ways of providing notice about how their information is being used to inform decisions, such as those described in the case studies of this report. Recognizing the research issues outlined, experts from the data science and social science communities, among others, should continue to develop additional best practices for fair and ethical use of big data techniques and machine learning in the public and private sectors.

- ***Promote academic research and industry development of algorithmic auditing and external testing of big data systems to ensure that people are being treated fairly.*** One way these issues can be tackled is through the emerging field of algorithmic systems accountability, where stakeholders and designers of technology "investigate normatively significant instances of discrimination involving computer algorithms" and use nascent tools and approaches to proactively avoid discrimination through the use of new technologies employing research-based behavior science.[79] These efforts should

also include an analysis identifying the constituent elements of transparency and accountability to better inform the ethical and policy considerations of big data technologies. There are other promising avenues for research and development that could address fairness and discrimination in algorithmic systems, such as those that would enable the design of machine learning systems that constrain disparate impact or construction of algorithms that incorporate fairness properties into their design and execution.

- ***Broaden participation in computer science and data science, including opportunities to improve basic fluencies and capabilities of all Americans.*** Consistent with the goals of the President's Computer Science for All and TechHire initiatives, educational institutions and employers can strive to broaden participation in these fields.[80] In particular, they should look for ways to provide more Americans with opportunities to have greater fluency and awareness of how these issues impact them and to influence how these fields evolve going forward.
- ***Consider the roles of the government and private sector in setting the rules of the road for how data is used.*** As use of big data moves from new and novel to mainstream, the private sector, citizens, institutions, and the public sector are establishing expectations, norms, and standards that will serve as guides for the future. How big data is used ethically to reduce discrimination and advance opportunity, fairness, and inclusion should inform the development of both private sector standards and public policy making in this space.

[1] The White House. "Big Data: Seizing Opportunities, Preserving Values." May 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.
[2] The White House. "Big Data: Seizing Opportunities and Preserving Values: Interim Progress Report." February 2015. https://www.whitehouse.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf.
[3] Executive Office of the President. President's Council of Advisors on Science and Technology. "Report to the President: Big Data and Privacy: A Technological Perspective." May 2014. https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.
[4] The Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues." January 2016. https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.
[5] Ibid.
[6] This report uses the term "discrimination" in a very broad sense to refer to outsized harmful impacts—whether intended or otherwise—that the design, implementation, and utilization of algorithmic systems can have on discrete communities and other groups that share certain characteristics. This report does not intend to reach a legal conclusion about discrimination as defined under federal or other law. Some instances of discrimination may be unintentional and even unforeseen. Others may be the result of a deliberate policy decision to concentrate services or assistance on those who are most in need. Still others may create adverse consequences for particular populations that create or exacerbate inequality of opportunity for those already at a disadvantage. We discuss all of these types of discrimination throughout the report—our aim being to draw attention to the fact that, as big data algorithmic systems play an increasing role in shaping our experiences and environment, those who design and implement them must be increasingly thoughtful about the potential for discriminatory effects that can have long-term or systemic consequences.
[7] The White House. "Big Data: Seizing Opportunities, Preserving Values." May 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf
[8] The White House. "Big Data: Seizing Opportunities and Preserving Values: Interim Progress Report." February 2015. https://www.whitehouse.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf.
[9] Ibid. "Digital Ad Revenues Surge 19% to $27.5 Billion in First Half of 2015." IAB.com. 2015. January 27, 2016. http://www.iab.com/news/digital-ad-revenues-surge-19-climbing-to-27-5-billion-in-first-half-of-2015-according-to-iab-internet-advertising-revenue-report/.
[10] The White House. "Big Data: Seizing Opportunities and Preserving Values: Interim Progress Report." February 2015. https://www.whitehouse.gov/sites/default/files/docs/20150204_Big_Data_Seizing_Opportunities_Preserving_Values_Memo.pdf
[11] Ibid.
[12] The White House Council of Economic Advisers. "Big Data and Differential Pricing." February 2015. https://www.whitehouse.gov/sites/default/files/whitehouse_files/docs/Big_Data_Report_Nonembargo_v2.pdf; The Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion?: Understanding the Issues." January 2016. https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.
[13] Crawford, K. "The Hidden Biases of Big Data." *Harvard Business Review*. April 1, 2013, https://hbr.org/2013/04/the-hidden-biases-in-big-data.
[14] "U.S. Smartphone Use in 2015." Pew Research Center. March 31, 2015. http://www.pewinternet.org/2015/04/01/u-s-smartphone-use-in-2015/pi_2015-04-01_smartphones_07/.

[15] Citron, D., and Pasquale, F. "The Scored Society: Due Process for Automated Predictions." 89 *Washington L. Rev.* 1, 2014. https://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1318/89WLR0001.pdf?sequence=1.
[16] Pasquale, F. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard University Press. 2015.
[17] Dwork, C. and Mulligan, D. "It's Not Privacy, and It's Not Fair." 66 *Stan. L. Rev. Online* 35. September 3, 2013. http://www.stanfordlawreview.org/online/privacy-and-big-data/its-not-privacy-and-its-not-fair.
[18] Taylor, F. "Hiring in the Digital Age: What's Next for Business Recruiting?" *Business News Daily*. January 11, 2016. http://www.businessnewsdaily.com/6975-future-of-recruiting.html.
[19] The Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion?: Understanding the Issues." January 2016. https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.
[20] Kraemer, F., van Overveld, K., and Peterson, M. 2011. "Is There an Ethics of Algorithms?" *Ethics and Information Technology* 13 (3): 251–60.
[21] Miller, C. "When Algorithms Discriminate." *The New York Times*. July 9, 2015. http://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html.
[22] Sweeney, L. "Discrimination in Online Ad Delivery". Harvard University. January 28, 2013 http://ssrn.com/abstract=2208240 or http://dx.doi.org/10.2139/ssrn.2208240; Datta, A., Tschantz, M. C., and Datta, A. "Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination." In *Proceedings on Privacy Enhancing Technologies (PoPETs)*. 2015. https://www.andrew.cmu.edu/user/danupam/dtd-pets15.pdf.
[23] Ng, A. Machine Learning course. Stanford University. 2016. https://www.coursera.org/learn/machine-learning.
[24] Citron, D. "Big Data Should Be Regulated by 'Technological Due Process'." *The New York Times*. August 6, 2014. http://www.nytimes.com/roomfordebate/2014/08/06/is-big-data-spreading-inequality/big-data-should-be-regulated-by-technological-due-process.
[25] Nelson, B. "The Data on Diversity." *Communications of the ACM*. Vol 57. No 11, Pages 86-95. http://cacm.acm.org/magazines/2014/11/179827-the-data-on-diversity/fulltext; Welle, B. Smith, M. "Time to Raise the Profile of Women and Minorities in Science." *Scientific American*. October 1, 2014. http://www.scientificamerican.com/article/time-to-raise-the-profile-of-women-and-minorities-in-science/.
[26] Self W., Mitchell G., Mellers B., Tetlock P., and Hildreth J. "Balancing Fairness and Efficiency: The Impact of Identity-Blind and Identity-Conscious Accountability on Applicant Screening." *PLoS ONE*. December 14, 2015. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0145208.
[27] Hardy, Q. "Why Data is Not the Truth." *The New York Times*, June 1, 2013. http://bits.blogs.nytimes.com/2013/06/01/why-big-data-is-not-truth/.
[28] The Consumer Financial Protection Bureau Office of Research. "Data Point: Credit Invisibles." May 2015. http://files.consumerfinance.gov/f/201505_cfpb_data-point-credit-invisibles.pdf.
[29] Ibid.
[30] Ibid.
[31] Ibid.
[32] Turner, M. Walker, P., Chaudhuri, S., and Varghese, R. "A New Pathway to Financial Inclusion: Alternative Data, Credit Building, and Responsible Lending in the Wake of the Great Recession." Policy and Economic Research Council. 2012. http://www.perc.net/wp-content/uploads/2013/09/WEB-file-ADI5-layout1.pdf
[33] Ibid.
[34] Wei, Y., Yildirim, P., Van den Bulte, C. and Dellarocas, C., "Credit Scoring with Social Network Data." Marketing Science. 2015. http://pubsonline.informs.org/doi/abs/10.1287/mksc.2015.0949.
[35] The Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues." January 2016. https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.
[36] Javelin Strategy and Research. "Evaluating the Viability of Alternative Credit Decisioning Tools." LexisNexis. May 2013. http://www.lexisnexis.com/risk/downloads/whitepaper/alternative-credit-decisioning.pdf.

[37] Talbot, D. "Data Discrimination Means the Poor May Experience a Different Internet." *MIT Technology Review.* October 9, 2013. http://www.technologyreview.com/news/520131/data-discrimination-means-the-poor-may-experience-a-different-internet/.
[38] The Federal Trade Commission. "Credit-Based Insurance Scores: Impacts on Consumers of Automobile Insurance: A Report to Congress." July 2007. https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf; The Federal Reserve Board. "Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit." August 2007. http://www.federalreserve.gov/boarddocs/rptcongress/creditscore/.
[39] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. "Fairness Through Awareness." *Proceedings of the 3rd ACM Innovations in Theoretical Computer Science Conference*, 214-226. 2012. http://arxiv.org/pdf/1104.3913.pdf.
[40] The Federal Trade Commission. "Credit-Based Insurance Scores: Impacts on Consumers of Automobile Insurance: A Report to Congress." July 2007. https://www.ftc.gov/sites/default/files/documents/reports/credit-based-insurance-scores-impacts-consumers-automobile-insurance-report-congress-federal-trade/p044804facta_report_credit-based_insurance_scores.pdf; The Federal Reserve Board. "Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit." August 2007. http://www.federalreserve.gov/boarddocs/rptcongress/creditscore/.
[41] Robinson, D., and Yu, H. "Knowing the Score: New Data, Underwriting, and Marketing in the Consumer Credit Marketplace. A Guide for Financial Inclusion Stakeholders." October 2014. https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf.
[42] Capelli, P. "Making the Most of Online Recruiting." *Harvard Business Review*. March 2001. https://hbr.org/2001/03/making-the-most-of-on-line-recruiting.
[43] Ibid.
[44] Goldberg, C. "Relational Demography and Similarity-Attraction in Interview Assessments and Subsequent Offer Decisions." *Group and Organization Management* 30.6: 597-624. 2005. http://gom.sagepub.com/content/30/6/597.short; Davidson, J. "Feds urged to fight 'unconscious bias' in hiring and promotions." *The Washington* Post, April 14, 2016. https://www.washingtonpost.com/news/powerpost/wp/2016/04/14/feds-urged-to-fight-unconscious-bias-in-hiring-and-promotions/.
[45] Goldberg, C. "Relational Demography and Similarity-Attraction in Interview Assessments and Subsequent Offer Decisions." *Group and Organization Management* 30.6: 597-624, 2005. http://gom.sagepub.com/content/30/6/597.short.
[46] "Robots Are Color Blind: How Big Data Is Removing Biases from the Hiring Process." *Michaelhousman.com*. http://michaelhousman.com/robots-are-color-blind/.
[47] "Why Machines Discrimination and How to Fix Them." *Science Friday*. November 20, 2015. http://www.sciencefriday.com/segments/why-machines-discriminate-and-how-to-fix-them/.
[48] The Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues." January 2016. https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report.
[49] U.S. Equal Employment Opportunity Commission. Notice of Proposed Changes to the EEO-1 to Collect Pay Data from Certain Employers. https://www.eeoc.gov/employers/eeo1survey/2016_eeo-1_proposed_changes_qa.cfm
[50] "Robot Recruiters: Big Data and Hiring." *The Economist*. April 6, 2013. http://www.economist.com/news/business/21575820-how-software-helps-firms-hire-workers-more-efficiently-robot-recruiters.
[51] Peck, D. "They're Watching You at Work." *The Atlantic*. December 2013. http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/.
[52] Barocas, S. and Selbst, A. "Big Data's Disparate Impact." 104 *Calif. L. Rev.* 2016. http://ssrn.com/abstract=2477899.
[53] "Why Machines Discrimination and How to Fix Them." *Science Friday*. November 20, 2015. http://www.sciencefriday.com/segments/why-machines-discriminate-and-how-to-fix-them/.

[54] Carnevale, A., Rose, S., and Cheah, B. "The College Payoff." *Georgetown Center on Education and the Workforce*. August 5, 2011. https://cew.georgetown.edu/report/the-college-payoff/.
[55] The Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues." January 2016. https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report.
[56] U.S. Department of Education. College Scorecard. https://collegescorecard.ed.gov/.
[57] Nelson, L. "Big Data 101." *Vox*. July 14, 2014. http://www.vox.com/2014/7/14/5890403/colleges-are-hoping-predictive-analytics-can-fix-their-graduation-rates.
[58] Kurweil, M. and Wu, D. "Building a Pathway to Student Success at Georgia State University." *Ithaka S&R Case Study*. 2015. http://www.sr.ithaka.org/publications/building-a-pathway-to-student-success-at-georgia-state-university/; Marcus, J. "Colleges Use Data to Predict Grades and Graduation." *The Hechinger Report*. December 10, 2014. http://hechingerreport.org/like-retailers-tracking-trends-colleges-use-data-predict-grades-graduations/.
[59] Ungerleider, N. "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.
[60] Nelson, L. "Big Data 101." *Vox*. July 14, 2014. http://www.vox.com/2014/7/14/5890403/colleges-are-hoping-predictive-analytics-can-fix-their-graduation-rates.
[61] Ibid.
[62] Tennessee Higher Education Commission. Outcomes Based Funding Formula Resources. https://www.tn.gov/thec/topic/funding-formula-resources.
[63] Ibid.
[64] Tennessee Higher Education Commission. 2015-20 Outcomes Based Funding Formula Overview. https://www.tn.gov/thec/article/2015-20-funding-formula.
[65] Ness, E., Deupree, M., and Gándara M. "Campus Responses to Outcomes-Based Funding in Tennessee: Robust, Aligned, and Contested." https://www.tn.gov/assets/entities/thec/attachments/FordFoundationPaper.pdf.
[66] The U.S. Department of Justice's Office of Community Oriented Policing Services. "Final Report of the President's Task Force on 21st Century Policing." May 2015. http://www.cops.usdoj.gov/pdf/taskforce/taskforce_finalreport.pdf.
[67] "The Police Data Initiative: A 5-Month Update." October 27, 2015. https://www.whitehouse.gov/blog/2015/10/27/police-data-initiative-5-month-update; White House Police Data Initiative Highlights New Commitments. April 21, 2016. https://www.whitehouse.gov/the-press-office/2016/04/22/fact-sheet-white-house-police-data-initiative-highlights-new-commitments.
[68] Ibid.
[69] Abdollah, T. "'Early Warning Systems' aim to ID Troubled Policy Officers." *Los Angeles Daily News*. September 7, 2014. http://www.dailynews.com/government-and-politics/20140907/early-warning-systems-aim-to-id-troubled-police-officers.
[70] Haberman, C. and Ratcliffe, J. "The Predictive Policing Challenges of Near Repeat Armed Street Robberies." *Policing*, Vol 6, No. 2, 151-166. May 2, 2012. http://www.jratcliffe.net/wp-content/uploads/Haberman-Ratcliffe-2012-The-predictive-policing-challenges-of-near-repeat-armed-street-robberies.pdf.
[71] Kennedy, L., Caplan, J., and Piza, E. "Risk Clusters, Hotspots, and Spatial Intelligence: Risk Terrain Modeling as an Algorithm for Policy Resource Allocation Strategies." *Journal of Quantitative Criminology*. September 2010. https://www.researchgate.net/profile/Eric_Piza/publication/226431177_Risk_Clusters_Hotspots_and_Spatial_Intelligence_Risk_Terrain_Modeling_as_an_Algorithm_for_Police_Resource_Allocation_Strategies/links/543827f90cf24a6ddb92a4b9.pdf.
[72] Beck, C. and McCue, C. "Predictive Policing: What Can We Learn from Wal-Mart and Amazon about Fighting Crime in a Recession?" *Police Chief*, Vol. 76, No. 11. November 2009. http://www.policechiefmagazine.org/magazine/index.cfm?fuseaction=display_arch&article_id=1942&issue_id=112009.
[73] Friend, Z. "Predictive Policing: Using Technology to Reduce Crime." *FBI Law Enforcement Bulletin*. October 4, 2013. https://leb.fbi.gov/2013/april/predictive-policing-using-technology-to-reduce-crime.

[74] The U.S. Department of Justice, Office of Justice Programs, National Institute of Justice. "Racial Profiling and Traffic Stops." January 10, 2013. http://www.nij.gov/topics/law-enforcement/legitimacy/pages/traffic-stops.aspx; Department of Justice, Bureau of Justice Statistics. "Racial Disparity in US Drug Arrests." October 1, 1995. http://www.bjs.gov/content/pub/pdf/rdusda.pdf.

[75] The U.S. Department of Justice, Bureau of Justice Statistics. "Bridging Gaps in Police Crime Data." September 1999. http://www.bjs.gov/content/pub/pdf/bgpcd.pdf.

[76] Comey, J. "Uniform Crime Reporting Program Message from the Director." Fall 2015. https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/resource-pages/message-from-director.

[77] "Big Data, Big Questions." *International Journal of Communication*, Vol. 8. 2014. http://ijoc.org/index.php/ijoc/article/view/2169/1162.

[78] The Federal Trade Commission. "Big Data: A Tool for Inclusion or Exclusion? Understanding the Issues." January 2016. https://www.ftc.gov/reports/big-data-tool-inclusion-or-exclusion-understanding-issues-ftc-report.

[79] Sandvig, C., Hamilton, K., Karahalios, K., and Langbort, C. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." *Center for People and Infrastructures*, May 22, 2014 http://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf; Pope, D. and Syndor, J. "Implementing Anti-Discrimination Policies in Statistical Profiling Models." *American Economic Journal: Economic Policy*, Vol. 3, No. 3, pp. 206-231. 2011. http://faculty.chicagobooth.edu/devin.pope/research/pdf/website_antidiscrimination%20models.pdf.

[80] The White House. Computer Science for All. January 30, 2016. https://www.whitehouse.gov/blog/2016/01/30/computer-science-all; The White House. Tech Hire Initiative. https://www.whitehouse.gov/issues/technology/techhire.