DATA MINING AND THE DISCOURSE ON DISCRIMINATION

Solon Barocas
Center for Information Technology Policy
Princeton University
306 Sherrerd Hall
Princeton, NJ 08544
+1-609-258-2278
sbarocas@princeton.edu

ABSTRACT

This paper surveys and brings some order to the broad set of charges that commentators have begun to levy against data mining, all expressed in the language of discrimination. It maps the myriad kinds of discrimination ascribed to data mining, clarifies the precise mechanisms the commentators see as giving rise to these objectionable forms of discrimination, and specifies the principles or policies that such discrimination seems to contravene.

Categories and Subject Descriptors

K.4.1 [**Computers and Society**]: Public Policy Issues – *Ethics, Privacy, Regulation, Use/abuse of power*

General Terms

Algorithms, Management, Legal Aspects.

Keywords

Discrimination, Equity, Fairness

1. INTRODUCTION

There is an enormous diversity of wrongs that data mining seems to be committing when described as discriminatory. This paper presents and parses a selection of comments by industry professionals, regulators, advocates, scholars, and journalists that all hint at the various and more specific offenses attributed to data mining. In working through these examples, the paper will unpack what commentators mean by discrimination, how they see data mining as giving rise to that discrimination, and why they view it as objectionable. In so doing, it will reveal striking inconsistencies in the anxieties provoked by data mining, each expressed as fears of discrimination, but also useful points of contrast that can help structure future debate.

2. DRAWING DISTINCTIONS

2.1 Inferring Membership in Protected Class

To start, consider a lengthy and well-cited passage from a blog post by Alistair Croll, chair of O'Reilly's Strata Conferences, the first major series of industry gatherings devoted to 'big data': "*We're great at using taste to predict things about people. OKcupid's 2010 blog post "The Real Stuff White People Like"* [1] *showed just how easily we can use information to guess at race. It's a real eye-opener (and the guys who wrote it didn't include everything they learned—some of it was a bit too controversial.) They simply looked at the words one group used which others didn't often use. The result was a list of "trigger" words for a particular race or gender […] Now run this backwards. If I know you like these things, or see you mention them in blog posts, on Facebook, or in tweets, then there's a good chance I know your gender and your race, and maybe even your religion and your sexual orientation. And that I can personalize my marketing efforts towards you […] That makes it a civil rights issue […] If I collect information on the music you listen to, you might assume I will use that data in order to suggest new songs, or share it with your friends. But instead, I could use it to guess at your racial background. And then I could use that data to deny you a loan*" [2].

For all the novelty of this scenario, Croll is worried about a very traditional form of discrimination. Indeed, Croll describes a scenario in which firms infer consumers' membership in a protected class[1] from correlated preferences and discriminate against them on that basis. In other words, he fears that data mining will allow firms to consciously and purposefully disadvantage members of a protected class. Were this to take the form of steering these members, through marketing, to financial products with less favorable terms or of denying them a loan, the firms' actions could be illegal under existing anti-discrimination law [3]. Croll's concern, then, is not with the absence of laws registering the objectionableness of these actions, but with the fact that data mining seems to help circumvent the protections that these laws offer. Because the existing ways of enforcing the laws rest on restricting access to information about membership in protected classes or prohibiting consideration of that information explicitly in decision-making, the ability to obtain and draw on this information indirectly and furtively, through inference, poses a threat to civil rights. Thus, for Croll, data mining is objectionable, as a form of discrimination, because it enables *and* masks the purposeful disadvantaging of members of a protected class.

While these are legitimately worrisome possibilities, they describe a situation in which data mining is not itself discriminatory. Rather, data mining here serves as a tool for those who purposefully seek out new ways to discriminate. Conscious prejudice motivates both the decision to infer whether an individual is a member of a protected class and the decision to disadvantage individuals on that basis. This seems to suggest that, even if data mining can more effectively realize discriminatory intent, no such prejudice inheres in the data mining process itself.

2.2 Statistical Bias

As Kate Crawford has warned, however, discrimination need not be intentional. Disproportionately adverse determinations for members of protected classes may be the result of *"'signal problems' in big-data sets—dark zones or shadows where some*

[1] 'Protected classes' are groups entitled to special legal protections against discrimination on the basis of certain characteristics. Established by legal fiat and first enumerated in the Civil Rights Act of 1964, these characteristics include race, color, religion, sex, and national origin. Subsequent federal and state laws have also established protections against discrimination on the basis of age, pregnancy, disability, genetic information, and sexual orientation.

citizens and communities are overlooked or underrepresented" [4]. To explain how this might come to pass, Crawford describes Boston's experience with Street Bump, an application that turns residents' smart phones into passive sensors for potholes. The application relies on built-in accelerometers to detect when drivers happen upon particularly uneven road and then automatically reports the location to the city. While Crawford recognizes the potential value of Street Bump, she warms that the uneven rates of smartphone ownership across different parts of the city would likely result in reports that under-represent the incidence of potholes in poorer areas—areas in which members of protected classes reside in relatively larger numbers. Were the city to rely on this data to direct its repair efforts or make predictions about future road problems, it would underserve exactly those citizens already in a position of relative disadvantage. Indeed, the city would likely address far less of the infrastructural decay in communities composed disproportionately of members of protected classes. As Crawford elsewhere concludes, "*[a]s we move into an era in which personal devices are seen as proxies for public needs, we run the risk that already existing inequities will be further entrenched*" [5]. Where economic inequality creates a reporting bias that decision-makers fail to recognize or address, even data-driven decisions can effectively discriminate against the poor and legally protected populations. And while discrimination is an artifact of a sampling bias in this instance, rather than the conscious prejudice of any particular decision-maker, its adverse effects could be equally or even more severe than those feared by Croll.

2.3 Faulty Inferences

The discrimination that worries David Vladeck, former Director of the Bureau of Consumer Protection at the Federal Trade Commission (FTC), has to do with the conclusions that firms draw from the limited set of activities that they can observe. Writing for *The New York Times,* Steve Lohr paraphrases Vladeck, who argues that "*[d]iscrimination by statistical inference is a real risk in the Big Data world, as some personal data trails suggest a correlation that may be wrong […] Imagine spending a few hours looking online for information on deep fat fryers. You could be looking for a gift for a friend or researching a report for cooking school. But to a data miner, tracking your online viewing, this hunt could be read as a telltale sign of an unhealthy habit — a data-based prediction that could make its way to a health insurer or potential employer*" [6].

Here data mining constitutes a form of objectionable discrimination because it draws an incorrect inference from certain behavior, affecting the way firms will subsequently view and treat that individual. This line of reasoning seems to suggest that data mining is invidious because it makes mistakes; it is morally wrong when it draws the wrong conclusions and discriminates on that basis. At issue is the simple fact that certain individuals may be subject to erroneous inferences. In stark contrast to Croll and Crawford whose concerns with discrimination hinge on the treatment of members of protected classes or historically disadvantaged groups, Vladeck sees data mining as unfairly discriminatory because it can subject individuals—any individual—to decisions informed by faulty inferences.

FTC Chairwoman Edith Ramirez has expressed the exact same sentiment, explaining that "*big data [can] be used to make determinations about individuals, not based on concrete facts, but on inferences or correlations that may be unwarranted*" [7]. Here, again, the perceived legitimacy of decisions that involve data mining seems to rest rather narrowly on their apparent accuracy; inferences are "unwarranted" when they fail to correspond to the "concrete facts".

This line of reasoning is exceedingly common among privacy and consumer advocates, in part because it is much easier to object to decisions that rest on faulty inferences. Note how Peter Eckersley of the Electronic Frontier Foundation, for instance, starts by saying that "*[t]racking data can be used to figure out your political bent, religious beliefs, sexuality preferences, health issues or the fact that you're looking for a new job*", but concludes that "*[t]here are all sorts of ways to form wrong judgments about people*" [8].

In another piece for *The New York Times*, Lohr stakes out a similar position of his own: "*These models, like metaphors in literature, are explanatory simplifications. They are useful for understanding, but they have their limits. A model might spot a correlation and draw a statistical inference that is unfair or discriminatory, based on online searches, affecting the products, bank loans and health insurance a person is offered*" [9].

According to Lohr, data mining can give rise to objectionable discrimination when, drawing on the insufficiently rich information communicated by online searches, for example, it incorrectly lumps individuals into groups to which they don't actually belong. Here, again, data mining is invidious because its overly simple models result in individuals receiving the wrong offers. Note, however, that Lohr distinguishes between inferences that are unfair and those that are discriminatory. This seems to suggest that Lohr views *any* decision informed by erroneous inference as unfair. Which, in turn, suggests that Lohr's definition of discrimination, in distinction to his understanding of unfairness, hinges on some additional criterion. Presumably, Lohr views any decision that somehow turns on whether a person is a member of a legally defined protected class as uniquely offensive, but also necessarily erroneous because membership in a protected class should hold no justifiable relevance to decisions involving consumer products, financial services, or healthcare—or at least far less relevance than alternative attributes. If so, Lohr's objections would seem to suggest that Vladeck, Ramirez, and Eckersley's concerns with erroneous inferences simply subsume Croll's concerns with decisions that consider membership in protected classes. They are all objectionable because they all involve spurious reasoning.

2.4 Overly Precise Inferences

Julie Brill, one of the acting Commissioners of the FTC, is likewise concerned with the prospect of firms denying offers or opportunities to individuals because of their health or financial status. In an article for *The New York Times* describing Brill's position, Natasha Singer notes that "*federal regulators have long warned about the potential for such data-mining to discriminate against consumers based on sensitive details like financial or health information*" [10]. This is a significant variant of the concerns documented in the previous section, wherein the discrimination enabled by access to financial or health information is objectionable, not because that information is wrong, but because that information is right. Brill's objections, like those described immediately above, do not rest on the narrower requirement that consumers' membership in protected classes affects their access to goods and services. Both imply that *anyone* can be subject to discrimination. But, unlike Vladeck, Ramirez, Eckersley, and Lohr, Brill thinks that there is reason to worry about *correctly* informed decisions to treat consumers differently in the marketplace.[2] In other words, Brill would object even when firms rightly infer from the purchase of a deep fat fryer that a specific individual does indeed have unhealthy eating habits.

Much of this worry has been couched in the language of consumer awareness and control, suggesting that Brill is most concerned with the fact that consumers rarely know if and how such information affects their access to goods and services, even if they could not contest those decisions on the grounds of accuracy or relevance.[3] But there is something more substantive at stake in Brill's concerns as well, even if these are not issues that fall within the remit of the FTC: access to especially revealing information may encourage firms to discriminate at a level of granularity that begins to threaten other public policy goals, namely risk-pooling.

In another blog post, Croll contemplates this very same situation: "*Perhaps the biggest threat that a data-driven world presents is an ethical one. Our social safety net is woven on uncertainty. We have welfare, insurance, and other institutions precisely because we can't tell what's going to happen—so we amortize that risk across shared resources. The better we are at predicting the future, the less we'll be willing to share our fates with others*" [13].

This is an entirely different reason to object to data mining than the one laid out in the earlier passage from Croll. The "biggest threat" posed by data mining is not illegal and invidious discrimination, as first imagined, but rather the overly precise capacity to discriminate between individuals who may place a greater or less demand on shared resources, regardless of their membership in a protected class. In dividing populations into ever-smaller groups, defined by apparent differences in the risk that they pose, data mining will affect the willingness for differently situated individuals to share common cause. These divisions effect a kind of discrimination that is suspect, Croll argues, not because they depend on the attributes that define protected classes, but because they may facilitate the pursuit of rational self-interest that is corrosive to the social fabric of welfare states.

2.5 Shifting the Sample Frame

Jason Schultz has likewise warned of the potentially perverse consequences of rational decision-making driven by the results of data mining. Responding to questions about the efficacy and impartiality of predictive policing, in particular, Shultz argues that "*[i]t comes with inherent biases and prejudices that can be worse than the help it offers […] It kind of reinforces its own data by redirecting resources to those areas*" [14]. This uneasiness with predictive policing belies a number of interrelated concerns. In deploying police to areas predicted to have higher crime rates, data mining first shifts a greater proportion of the police's attention to those areas and away from others. Second, because the police will invariably find crime at a higher rate in those areas that they happen to subject to greater scrutiny, those who commit a crime in these areas will face a higher likelihood of being caught than those in other areas. Third, where this has the effect of progressively distorting where police observe crime and where they have corresponding cause to continue to redirect the focus of their attention, the rational decision-making informed by data mining may begin to subject specific sub-populations to disproportionate scrutiny—a degree of scrutiny that may no longer correspond to the actual, and not just the observed, rates at which crimes occur in different areas.

For some, the simple fact of disproportionality would seem to qualify this form of data-driven policing as discriminatory [15]. But this disproportionality can take on a more traditional discriminatory quality when the predictions generated by data mining lead to more patrolling (and thus more vigorous law enforcement) in areas where members of protected classes reside in disproportionate numbers. That is, data mining will constitute a form of discrimination if it leads members of protected classes to have disproportionate contact with the police. The specter of discrimination will become especially acute if members of these groups find that they have a greater chance of being caught when committing the same crime as others.

As Schultz warns, this can have the effect of reifying and reinforcing the tendencies that made certain areas stand out from others in the first place. Bizarrely, acting on predictions may help to confirm what had been predicted [16]. The concern, then, is that the rational allocation of attention and resources will skew the mechanism (the sample frame, in the language of statistics) by which evidence is obtained, leading to results that confirm *and* responses that compound apparent differences between groups. While many of the earlier charges of discrimination rested on whether data mining resulted in erroneous inferences and thus inappropriate (i.e., incorrect) treatment, the issue here is that the use of data mining may result in self-validating decisions that place groups under increasingly disproportionate and thus inappropriate amounts of scrutiny. This differs from all the other claims of discrimination because rationally choosing to act on the strategies suggested by data mining ultimately biases the allocation of attention and resources; the *reliance* on data mining introduces the bias that results in discriminatory acts.

3. UNFAIRNESS IN ITS MANY FORMS

Inconsistencies in charges of discrimination have been the cause of recent teeth gnashing; scholars have begun to express alarm at what they see as chronically underspecified claims [17]. While it would be easy to explain away use of the term as little more than a strategy to bring to bear both the legal force and rhetorical weight of charges of discrimination when expressing apprehensions about data mining, the fact that discrimination offers a way to express all of these concerns is significant.

This brief survey reveals that commentators see at least three rather different ways through which data mining gives rise to discrimination. The first involves conscious intentions to disadvantage members of protected class in ways that would be difficult to detect; the second focuses on problems with the data mining process itself that result in seemingly avoidable errors; and the third concerns the unwelcome effects when data mining

[2] Strictly speaking, this scenario does not involve data mining; gaining direct access to this kind of information obviates the need to infer financial and health status from other indicators. But Brill's worry should carry over to these efforts as well, so long as accurate inferences, like access to accurate information, endanger the solidarity upon which important social institutions rest. Brill's recent comments certainly seem to suggest thinking along these lines [11].

[3] Generally speaking, the FTC has taken issue with mistakes because it sees them as uniquely pernicious, given that most consumers are likely unaware that erroneous conclusions have been drawn from their behavior and lack ways to rectify the error [12].

significantly enhances certain decision-makers' powers of discernment. Describing each of these as discrimination can be confusing because each raises different concerns. Objections to the first rest on concerns with prejudicial decision-making and its masking; the second on concerns with bias and error and the distribution of those errors across different social groups, and the third on concerns with the perpetuation of inequality, even in the absence of prejudice, bias, and error. These objections, in turn, appeal to different legal and normative principles. The first and second sets of objections appeal to principles of procedural fairness, which plays to the dominant notion of justice enshrined in antidiscrimination law. The third set insists that fair procedures may nevertheless result in unfair (i.e., systematically unequal) outcomes, and thus calls for a commitment to distributive justice to compensate for the limits of procedural fairness. While this second principle also has an analogue in the law in the form of the disparate impact doctrine, many contests its potential function as a mechanism for redistribution [18].

The popular discourse is helpful because it shows that the current debate suffers from many of the same conceptual challenges that have characterized discrimination from its very inception as a formal notion in the law. The fiercest debates about discrimination have turned almost entirely on which principles of fairness the law should seek to enforce. Indeed, at issue in many of the long-standing debates about discrimination is what even qualifies as a problem to which antidiscrimination law is the appropriate response [19]. Dismissing the popular discourse as insufficiently precise fails to recognize that these claims reflect the decades-long dispute over the appropriate way to conceptualize and respond to discrimination, generally.

So far, commentators have tended to pivot between the two rather different notions of fairness as the particular details of the case demand. While there has been a tendency to place greater emphasis on threats to procedural fairness, commentators are increasingly dismissing as inadequate even the most rigorously enforced procedural remedies where data mining nevertheless perpetuates inequality. Oscillating between these two positions will become progressively more difficult, as it becomes more apparent that they actually call for very different policy responses.

While greater clarity on the particular principle at stake may aid efforts to devise remedial mechanisms, it will also compel critics of data mining to confront difficult problems with the concept of discrimination, generally, and shortcomings of the specific legal instruments available to address it [18]. The debate will eventually have to grapple with basic questions about what it means to describe data mining as fair and the appropriate mechanisms to achieve fairness.

4. REFERENCES

[1] Rudder, C., 2010. The REAL 'Stuff White People Like'. *oktrends: Dating Research from OkCupid*.

[2] Croll, A., 2012. Big Data Is Our Generation's Civil Rights Issue, and We Don't Know It. *Solve for Interesting*.

[3] Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R., 2012. Fairness through Awareness. In the 3rd Innovations in Theoretical Computer Science Conference. New York, NY: ACM, pp. 214–226.

[4] Crawford, K., 2013. Think Again: Big Data. *Foreign Policy*.

[5] Crawford, K., 2013. The Hidden Biases in Big Data. *Harvard Business Review*.

[6] Lohr, S., 2013. Sizing Up Big Data, Broadening Beyond the Internet. *The New York Times*.

[7] Ramirez, E., 2013. The Privacy Challenges of Big Data: A View From the Lifeguard's Chair. In Technology Policy Institute Aspen Forum. Aspen, CO.

[8] Quoted in Acohido, B., 2011. Facebook Tracking Is Under Scrutiny. *USA Today*.

[9] Lohr, S., 2012. The Age of Big Data. *The New York Times*.

[10] Singer, N., 2013. F.T.C. Member Starts "Reclaim Your Name" Campaign for Personal Data. *The New York Times*.

[11] Brill, J., 2012. Big Data, Big Issues. In Sixth Annual Law & Information Society Symposium: Big Data, Big Issues. New York, NY.

[12] Brill, J., 2013. Brill Reclaim Your Name. In Computers, Freedom, and Privacy Conference. Washington, DC.

[13] Croll, A., 2012. New Ethics for a New World. *O'Reilly Radar*.

[14] Quoted in Sengupta, S., 2013. In Hot Pursuit of Numbers to Ward Off Crime. *The New York Times*.

[15] Harcourt, B., 2007. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*, Chicago, IL: University of Chicago Press.

[16] Donald MacKenzie, 2006. *An Engine, Not a Camera: How Financial Models Shape Markets*, Cambridge, MA: MIT Press.

[17] Polonetsky, J. and Tene, O., 2013. Privacy and Big Data: Making Ends Meet. *Stanford Law Review Online*, 66, pp.25–33.

[18] Barocas, S. and Selbst, A., 2014. Big Data's Disparate Impact.

[19] Altman, A., 2011. Discrimination E. N. Zalta, ed. *Stanford Encyclopedia of Philosophy*.