Discrimination by Database

Author : James Grimmelmann

Date : November 4, 2014

Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, available at SSRN (2014).

I have previously written about an NYU School of Internet scholars, led by the philosopher Helen Nissenbaum, whose work is "philosophically careful, intellectually critical, rich in detail, and humanely empathetic." There is also a Princeton School, which orbits around the computer scientist Ed Felten, and which is committed to technical rigor, clear exposition, social impact, and creative problem-solving. These traditions converge in *Big Data's Disparate Impact* by Solon Barocas and Andrew Selbst. The article is an attempt to map Title VII employment discrimination doctrine on to data mining, and it is one of the most interesting discussions of algorithmic prediction I have read.

The pairing—anti-discrimination law and data mining—is ideal. They are both centrally concerned with how overall patterns emerge from individual acts; they shift back and forth between the micro and the macro, the stones and the mosaic. Moreover, they are both centrally concerned with making good decisions: each in its own way aspires to replace crude stereotypes with nuanced reason. It would seem then, that Big Data ought to be an ideal ally in Title VII's anti-discrimination mission. But Barocas and Selbst give reasons to think that the opposite may be true: that data mining will introduce new forms of bias that Title VII is ill-equipped to remedy.

In any interesting decision problem, there is a gap between the evidence available to a decision-maker and her goals. A recruiter would like to avoid hiring candidates who will stab customers, but the candidates who end up stabbing customers never seem to list that fact on their resumes. Thus, the decision will be mediated through a rule: a prediction about how the observable evidence correlates with a goal. In this context, then, data mining is a discipline of using large datasets, sophisticated statistics, and raw computational power to formulate better, more predictive rules.

The resulting rules are at once intensely automated and intensely human. On the one hand, data mining algorithms can discover surprising rules that human rules would not have thought to look for or complicated rules that humans would not have been able to formulate. In this sense, the algorithmic turn allows the use of rules that really are supported by the data, rather than the biased rules we flawed humans would think to try.

At the same time, as Barocas and Selbst deftly show, data mining requires human craftwork at every step. Humans pick the datasets to use, and they massage that data to make it usable for the learning algorithms (e.g., by imputing ZIP codes for customers who haven't listed them). Humans do the same thing on the other end, both approximating and constructing the characteristics they wish to select for. To learn who is a good employee, an algorithm needs to train on a dataset in which a human has flagged employees as "good" or "bad," but that flagging process in a very real sense defines what it means to be a "good" employee. In the gap between evidence and goals, humans specify the set of possible rules the algorithm will choose among, and the algorithm that will choose among them.

Barocas and Selbst circle over this ground three times, each time at a higher level of abstraction: technical, doctrinal, prescriptive. On the first pass, they survey the ways that invidious biases can enter into the automated algorithmic judgments. On the second, they show that Title VII doctrine often fails to

catch these biases, even when they would result in serious and unjustified mistreatment. And on the third, they show that it will not be easy to patch Title VII—that the challenges of Big Data go to the heart of the American project of equality.

Injecting algorithms into what was formerly a human decision-making process can undermine accountability by diffusing responsibility. For one thing, the data intensitivity of data mining makes it easier for bad actors to hide their fingerprints. Take the deeply uncool process of collecting, cleaning, and merging datasets to prepare them for mining. If a data broker redlines a tenant database that is then used as an input to an employment-applicant screening algorithm, the resulting hiring decisions will in a very real sense be racially motivated, but it will be almost impossible for anyone to reconstruct why. Proof problems abound in the land of Big Data, and *Big Data's Disparate Impact* is replete with examples. Ring of Gyges, anyone?

It gets worse. Big Data optimists have argued that employers and other decision-makers rely on race as a crude proxy for the characteristics they really care about, so that with better data they will be less racist. Perhaps. But if Bert is a proxy for Ernie, then Ernie can also be a proxy for Bert. In a world where everything predicts everything else, as Paul Ohm has half-jokingly hypothesized, a data-mining algorithm does not need direct access to forbidden criteria like religion or race to make decisions on the basis of them. Indeed, it can find far subtler ones than humans are capable of: perhaps birth year plus car color plus favorite potato chip brand equals national origin. Put another way, data mining can be just as efficient at optimizing discrimination as at avoiding it.

Moreover, on closer inspection, almost every interesting dataset is tainted by the effects of past discrimination. In a classic example, St. George's Hospital trained an algorithm to replicate its admission's staff's evaluations of medical-school applicants with 90% to 95% fidelity. Unfortunately, the staff's past decisions had been racist and sexist, so "the program was not introducing new bias but merely reflecting that already in the system." That last phrase should be alarming to anyone who has worried about the divide between disparate treatment and disparate impact. "In certain contexts, data miners will never be able to disentangle legitimate and proscribed criteria," Barocas and Selbst write, because the "legitimate" criteria redundantly encode the "proscribed" ones. But if "the computer did it," and these patterns seem to emerge from the data as if by magic, Title VII has a hard time explaining who if anyone has done something culpably wrong in relying on them.

In other words, as Barocas and Selbst observe, data mining brings civil rights law face to face with the unresolved tension between its nondiscrimination and antisubordination missions. On the one hand, individual acts of invidious discrimination dissolve into the dataset; on the other, the dataset itself is permeated by past discrimination. This would be a familiar enough observation about the limits of strictly race-neutral analysis in a world of self-perpetuating patterns of exclusion, but the algorithmic angle makes it new and urgent. Algorithms are not neutral; they make fraught decisions for complex reasons. In all of this, perhaps, Big Data is surprisingly human.

Cite as: James Grimmelmann, *Discrimination by Database*, JOTWELL (November 4, 2014) (reviewing Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, available at SSRN (2014)), https://cyber.jotwell.com/discrimination-by-database/.