Is Big Data Spreading Inequality?

NY Times Room for Debate, August 6, 2014

Social media companies depend on [selling information](http://www.nytimes.com/2014/08/03/technology/how-facebook-sold-you-krill-oil.html) about their users’ clicks and purchases to data brokers who match ads to the most receptive individuals.

But the [Federal Trade Commission](http://www.nytimes.com/2014/05/28/technology/ftc-urges-legislation-to-shed-more-light-on-data-collection.html) and the [White House](http://www.nytimes.com/2014/05/02/us/white-house-report-calls-for-transparency-in-online-data-collection.html) have called for legislation that would inform consumers about the data collected and sold to companies, [warning](http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf) of analytics that have “the potential to eclipse longstanding civil rights protections.”

Does the collection of data by companies threaten consumers’ civil rights?

The Dangers of High-Tech Profiling, Using Big Data



[*Seeta Peña Gangadharan*](http://newamerica.net/people/seeta_gangadharan)*is a senior research fellow at the New America Foundation's*[*Open Technology Institute*](http://oti.newamerica.net/)*and a fellow at the*[*Data & Society Research Institute*](http://www.datasociety.net/)*.*

The rise of commercial data profiling is exacerbating existing inequities in society and could turn de facto discrimination into a high-tech enterprise.

During the subprime lending boom, [mortgage brokers were](http://www.centerfordigitaldemocracy.org/sites/default/files/FTCsupplemental_statement1107_0.pdf) blending cookie data that tracks behavior and user data gathered from sites like Bankrate.com, along with geographical and other demographic data, to easily infer race. This data mining helped brokers – or the marketing companies they hired – target minorities with “ghetto loans,” sometimes couched in the “[language of African-American](https://www.nclc.org/images/pdf/unreported/paschal-decl-balt.pdf).”

While multicultural and targeted marketing is perfectly legal, using race as the basis on which to discriminate against someone in the lending market is not.

Today, commercial data profiling is not only enabling more efficient predatory targeting but also leading to a problem of spurious connections. People can now be rated on risk or creditworthiness based on analysis of hundreds of variables that yield correlations with little explanatory power, but with the potential to have disparate impacts, according to the [National Consumer Law Center](http://www.nclc.org/images/pdf/pr-reports/report-big-data.pdf).

Information on cellphone payments, rental payments and from social media can all be factored into how people are evaluated for loans. For example, social media posts about a car breakdown can [now determine creditworthiness](http://online.wsj.com/news/articles/SB10001424052702304732804579425631517880424?mg=reno64-wsj&url=http://online.wsj.com/article/SB10001424052702304732804579425631517880424.html).

And while all consumers might be affected by these kinds of correlative analyses, members of low-income communities of color stand to lose the most: better access to credit will go a long way in helping a poor person get back on her feet.

Let’s respond to this digital discrimination by making civil rights a core driver of data-powered innovations and getting companies to share best practices in detecting and avoiding discriminatory outcomes.

By connecting people with services, goods, information and institutions they need to improve their lives, new social networks and technologies can serve as an important countervailing force to inequities faced by marginalized groups.

Fifty years after the passage of the Civil Rights Act, it’s clear that data-powered technology is an integral player in the [opportunity equation](http://www.civilrights.org/press/2014/civil-rights-principles-big-data.html). Knowing how and when data profiling creates unfairness will go a long way in ensuring the Internet serves as a democratizing force in society.

The Commercial Use of Big Data Can Help Improve Equality



[*Christopher Wolf*](http://www.futureofprivacy.org/about/about-christopher-wolf/)*is the co-chairman of the*[*Future of Privacy Forum*](http://www.futureofprivacy.org/)*and the national civil rights chairman of the*[*Anti-Defamation League*](http://dc.adl.org/)*. He is on*[*Twitter*](https://twitter.com/PrivacyWolf)*.*

The White House [report](http://www.whitehouse.gov/issues/technology/big-data-review) that sounded an alarm about the misuse of big data in housing, credit, employment and education also highlighted its ability to identify health risks at early stages in communities, create efficiencies in energy distribution and uncover fraud through predictive analyses.

And while the Federal Trade Commission and others have pointed to the potential for discrimination in big data, it is important to remember that it can also advance the interests of minorities and actually fight discrimination.

For example, data can be used to improve diversity recruiting in the workplace. [Entelo Diversity](http://blog.entelo.com/company-news/announcing-entelo-diversity), a recruiting company, uses an algorithm to sift through publicly available data and data pulled from social media platforms to find job candidates from within underrepresented segments of the population.

Big data can also be used to analyze land use for discriminatory effects. The Cedar Grove Institute for Sustainable Development has been able to identify disparities in city services (among other [discriminatory effects of land annexation](http://www.cedargroveinst.org/files/Southern%20Moore%20County%20Case%20Study.pdf) policies) in Moore County, North Carolina, through data.

Big data was responsible for a National School Board Association [report](http://www.nsba.org/sites/default/files/reports/Partnerships_Not_Pushouts_Guide.pdf) that showed school suspensions affect one out of five African-American students and only one out of 20 Caucasian students, which may partly explain the large discrepancy between graduation rates of those two groups. With this information, we identify — and, then, can work to support — student populations with unique needs.

In New York, there is a new campaign to [collect coordinated data](http://www.capitalnewyork.com/article/albany/2014/07/8549536/state-agencies-launch-lgbt-data-collection-effort) to improve health and human services for lesbian, gay, bisexual and transgender individuals, after a 2011 Institute of Medicine [report](http://www.iom.edu/Reports/2011/The-Health-of-Lesbian-Gay-Bisexual-and-Transgender-People/Press-Release.aspx) showed that a “scarcity of research yields an incomplete picture of the L.G.B.T. health status and needs.”

Concerns about potential misuse of big data are fair and do deserve attention, but the newfound flood of information also represents an advance in the use of technology that has real potential for bettering society.

Big Data Should Be Regulated by ‘Technological Due Process’



[*Danielle Keats Citron*](http://cyberlaw.stanford.edu/about/people/danielle-citron)*is a law professor at the University of Maryland and the author of the forthcoming "*[*Hate Crimes in Cyberspace*](http://www.hup.harvard.edu/catalog.php?isbn=9780674368293)*."*

In our increasingly [scored society](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2376209) – where algorithms turn our browsing habits, click patterns, purchases and GPS location data into ratings and predictions of who we are – it is very difficult for those who are mislabeled, or tagged in an undesirable way, to break out of their scoring prisons, in part because they are usually unaware they are being reviewed.

When the government makes important decisions that affect our life, liberty and property, it owes us “due process” – understood as notice of, and a chance to object to, those decisions. Unlike the government, private companies have no obligation to tell us about their scoring systems. Nonetheless, all predictive systems should be subject to fairness requirements that reflect their centrality in people’s lives.

Oversight of scoring determinations – a sort of “[technological due process](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1012360)” – would go a long way to ensure their fairness and accuracy for both government and private systems. In the case of governmental automated systems, their opacity has frustrated due process guarantees. Some systems like the “no-fly” list adjudicate in secret, while others lack record-keeping audit trails, making review of the law and facts supporting a system’s decisions impossible.

When it comes to the scoring systems of private companies, consumers have no inkling that they are being scored as “depression inclined,” “poor candidates” or “potentially pregnant” based on their online activity. They do not know their resumes are excluded from talent lists due to their Internet browsing habits.

The best way to ensure the fairness of scoring systems is through routine auditing by an expert agency. Much like the I.R.S. does with taxes, the Federal Trade Commission could randomly select private scoring systems for review on an annual basis. F.T.C. technologists could run expected and unexpected hypothetical scenarios to assess whether algorithmic predictions are statistical proxies for race, gender, religion and disability – thereby cutting down the possibility that the algorithms infringe on civil rights. The ever-present threat of an audit would encourage the adoption of precautions and, perhaps, encourage entities that are building scoring systems to be more mindful of concerns about discrimination and inaccurate predictions based on polluted data.

Precedent exists for such agency review. In 2008, the F.T.C. [invoked](http://www.ftc.gov/sites/default/files/documents/cases/2008/06/080610compucreditcmptsigned.pdf) its “unfairness authority” against CompuCredit – which marketed credit cards to people with subprime credit – for reducing users’ credit limits based on an undisclosed behavioral scoring model that penalized cardholders for certain transactions, including visits to pool halls and pawn shops, and personal counseling appointments.

As the demand for big data grows, we need to pierce the secrecy behind the systems that determine how we fit into society. Hard and fast rules are not the answer: Predictive systems are built on correlations and algorithms that change dynamically. But procedural regularity is essential to prevent “[arbitrariness by algorithm](http://www.ftc.gov/sites/default/files/documents/public_statements/privacy-challenges-big-data-view-lifeguard%E2%80%99s-chair/130819bigdataaspen.pdf).”

Big Data Can Be Used for Good in the Community



*Maurice Mitchell is the executive director of the*[*New York State Civic Engagement Table*](http://nycet.org/)*.*

An emerging group of investors, developers, engineers and philanthropists are attempting to harness the social impact potential of big data under the rubric of “[civic tech](http://www.knightfoundation.org/media/uploads/publication_pdfs/knight-civic-tech.pdf).” Prescriptions for our most pressing social issues emerge from the patterns found in the bonanza of collected data points.

We all should be concerned that racial and ethnic data might be used to continue to bar minorities from access to economic and political democracy. But this same information helps expose clear examples of racial bias and, that way, can also be used to combat it.

For example, the New York Police Department’s copious tracking of [data](http://www.nyc.gov/html/nypd/html/analysis_and_planning/stop_question_and_frisk_report.shtml) from their stops and arrests led to findings that clearly showed the racially disproportionate nature of stop-and-frisk. These findings confirmed the anecdotal claims of grassroots activists and community members in the city’s neighborhoods of color. Now that police forces are beginning to experiment with [predictive policing](http://www.nij.gov/topics/law-enforcement/strategies/predictive-policing/Pages/welcome.aspx), organizers can employ big data techniques to verify these practices don’t have racially disparate effects or infringe on our basic liberties. We should continue to apply this kind of oversight to the companies that collect and use data, as well.

A positive application of data collection is one used to create algorithms that predict where and when a community may face another wave of gentrification. Civic activists are also using race and ethnicity modeling to improve voter registration and participation of the most historically disenfranchised communities. At the New York State Civic Engagement Table we merge publicly accessible voting history data and data on public housing to help grassroots organizers build the political power of public housing residents by engaging the developments with lowest participation.

Partisans on both sides are buying marketing data to develop constituent-focused models. Socially responsible organizations are beginning to develop their own models based on how people would respond to issues like environmental conservation or whether or not someone is likely to be an active community member. These models are produced in part through mining commercial data.

It is ultimately up to those interested in justice and fairness to wield collected data in ways that provide opportunity and expand our ability to interact within our democracy. The ability to process multiple data points and develop predictive and descriptive algorithms around them helps to anticipate where and when social problems may arise and has the potential to unravel the roots of our current social ills to provide efficient solutions to them.

Losing Out on Employment Because of Big Data Mining

 

*Solon Barocas is a postdoctoral research associate at the Center for Information Technology Policy at Princeton University. Andrew Selbst is the Alan Morrison Supreme Court Assistance fellow at Public Citizen.*

Unfortunately, our civil rights statutes may not be up to the task of addressing the discriminatory effects of commercial data mining. Antidiscrimination law is primarily focused on prejudice, but there is an even greater risk that seemingly innocuous decisions by data miners could result in unintentional discrimination that can be just as severe.

For example, companies may attempt to improve their hiring decisions by [looking for patterns](http://www.theatlantic.com/magazine/archive/2013/12/theyre-watching-you-at-work/354681/) in the online behavior of prior applicants who later performed well as employees. While this has the possibility to suggest new and surprising criteria for job candidates, the company can only draw lessons from the limited set of employees that it has hired in the past.

Worse, data mining can pick up the bad examples set by previous hiring decisions: If past prejudice denied certain classes of candidates the opportunity to demonstrate their talents, data mining may fail to find anything to like in many qualified individuals who happen to fall into that category.

It is also not always evident where predictions are skewed by selection bias or past prejudice. Even where these are recognized as potential problems, the appropriate solution is not always clear. How do employers account for the kinds of candidates they have never hired in the past? What corrective measures can they take to purge the often subtle influence of prejudice or implicit bias from its hiring records? And how much should employers be expected to bear the costs that these solutions would entail?   
  
It is also not clear that richer data sets — even those untainted by bias or prejudice — would be more beneficial to disadvantaged groups. Data mining may still have a disproportionately negative effect on protected classes if the criteria that reliably predict some job-related quality also happen to correlate with class membership. This could further entrench inequality for protected classes.   
  
Liability in these cases is unclear. The disparate impact doctrine — which deals with "[practices that are fair in form, but discriminatory in operation](http://www.naacpldf.org/case/griggs-v-duke-power-co)" — contains an exception that allows employers to make decisions necessary to their businesses. Where data mining ranks candidates outright and involves no errors, but still evidences a disparate impact, it may fall into this exception, because an employer using the system could believe he was merely choosing the best candidate. The exception may apply even though another way of mining the data might rank candidates differently.

How much onus should we place on employers to explore alternatives in these situations? Discrimination in data mining will eventually force us to examine or reexamine several tough, but basic, questions about fairness.

Extending Credit To Sub-Prime Customers Through Big Data



*Jake Rosenberg is the co-founder and chief technology officer of*[*LendUp*](https://www.lendup.com/)*.*

Commercial use of big – or, at least, bigger – data is one of few ways to enable access and choice for subprime customers who need financial services.

Evaluating credit risk is exceptionally difficult. Many customers have damaged credit files that provide only a partial narrative of their current reality. Missed payments, defaults and bankruptcies can be just as prevalent in the files of customers who are on their way to improving their financial situation, as in those who are increasingly mired in debt.

Just as common are credit files that are thin and incomplete. Meaningful data such as on-time rent and bill payments, or even payday loan repayments, do not make it into traditional credit bureau data files.

Advancements in the variety and availability of financial data has allowed us to start a company, LendUp, which uses technology to help people move out of the payday debt cycle. We access large amounts of Fair Credit Reporting Act-compliant data and use supervised machine learning to analyze and make informed decisions about the borrowers we approve. These data are verifiable, related to credit behavior, and have shown empirically that they are predictive of a borrower's ability and willingness to repay.

Our ability to say yes to these customers gives them access to our series of products. As customers engage with us in various ways – by paying back loans or taking our free credit education courses, for example – they move up in credit status with us. Each level up means access to more money over longer periods of time at significantly lower fees. And the top two levels report payments to major credit bureaus, giving customers an opportunity to build their credit history and improve their credit scores.

Accessing and analyzing large amounts of data is key to our strategy, allowing us to make informed decisions and offer customers access and choice.