Key Points from The Technical Level-Setting Tutorial

Data & Civil Rights
October 30, 2014 – Washington, D.C.
http://www.datacivilrights.org/

What's Machine Learning?

We hear a lot in the civil rights context about new decisions being made by "machine learning," which can sound exotic. But machine learning is all around us. You probably encounter machine learning whenever you open up your email inbox: Spam emails magically go to the spam folder, while the mail you care about ends up in your inbox.

One obvious way to build a spam detection system would be to "hard wire" rules into the system: maybe emails from contacts in your address book are legitimate, while messages in all capital letters are spam. But machine learning takes a different approach.

In a machine learning system, the computer itself (rather than the human being) figures out what rules to use by looking at past examples. The computer then applies those rules in new situations.

How does this work? Let's continue with the spam example. To build a machine learning system that automatically recognizes spam, the system designer would start by assembling some "training data." The training data here would be a selection of example emails that have each been marked by a human as either "spam" or "not spam." The computer program then analyzes this training data, looking for common patterns that distinguish spam from real email. These patterns together make up the "model."

Later, when a new email arrives, the program passes the email through the model. If the email fits the "spammy" patterns it previously observed, the model will classify it as spam and discard it into the spam folder.

In general, machine learning systems become more accurate with more "training data." Patterns that aren't apparent across a small number of examples might become evident when many more examples are available. This is what makes big data so alluring: with more data, these systems are able to detect increasingly faint patterns, which help them build more accurate models about the real world. More accurate models lead to more accurate predictions.

Using methods just like this, Target was able to figure out that a teenage girl was pregnant, based on its data about what pregnant women typically buy.[1] This is also how researchers were able to infer sensitive personal attributes, like someone's race or sexual orientation, simply by examining publicly available Facebook activity.[2] And this is how data brokers today put

---

[1] Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. Times Magazine, Feb. 16, 2012, *available at* www.nytimes.com/2012/02/19/magazine/shopping-habits.html.
[2] Michal Kosinski, David Stillwell, and Thore Graepel, *Private Traits and Attributes Are Predictable From Digital Records of Human Behavior*, 110 Proceedings of the Nat'l Acad. of Sci., 5802, 5805 (Apr. 9, 2013), *available at* http://www.pnas.org/content/110/15/5802.abstract.

consumers into specific marketing segments, like "Fragile Families" or "Ethnic Second-City Strugglers."[3]

Some Machine-Made Rules Aren't Easily Understood

To reliably determine whether an email is spam, computers must "learn" all the many features that distinguish spam from legitimate email. As you might expect, the computer needs to consider an enormous number of factors, so the resulting spam detection model tends to be very complex. In fact, the model is almost always so complex that it defies human interpretation.

This is a common issue in many machine learning applications: There is a trade-off between human interpretability and accuracy. Machine learning is valuable because it can find subtle patterns that human beings might miss. But the consequence is that the patterns it finds—and the predictions made using those patterns—are often too complex for humans to easily understand.

For example, when Target developed its "pregnancy prediction" score for its shoppers, its scoring model considered many factors including whether the shopper "suddenly starts buying lots of scent-free soap and extra-big bags of cotton balls, in addition to hand sanitizers and washcloths."[4] Sometimes, such factors are intuitive, if complex. But, in other cases—and perhaps for some of the other factors in Target's model—the factors may be highly unintuitive. Taken together, the factors may be so numerous that it is practically impossible to explain why the model predicts that a shopper is likely pregnant.

Given the consequential role that machine learning now plays in decision-making in healthcare, employment, policing, and other areas, research efforts are underway to design systems that are not only accurate, but are also more easily understandable by humans.[5]

Minorities and the Burden of Errors

When performing predictions using statistical analysis, it's both desirable and possible to measure *how accurate* the predictions are. A data analyst might find that predictions are accurate 95% of the time, but wrong for the other 5%.

Sometimes, those errors are spread evenly across the population: each person has the same 5% chance of having something wrongly predicted about him. But this may not always be the case: often, the errors will not be evenly spread out, and the burden of those errors will fall disproportionately on certain subgroups.

The concern for the civil rights community is that this burden falls, more often than not, on minority groups. As mentioned above, the accuracy of machine learning typically improves as the number of examples increases. That is, larger social groups will generate more data, and more data means higher accuracy. Because there are, by definition, fewer people in minority populations

---

[3] United States Senate Committee on Commerce, Science, and Transportation, *A Review of the Data Broker Industry: Collection, Use, and Sale of Consumer Data for Marketing Purposes* (2013), http://www.commerce.senate.gov/public/?a=Files.Serve&File_id=0d2b3642-6221-4888-a631-08f2f255b577.
[4] Duhigg *supra* note 1.
[5] Alex A Freitas, "Comprehensible Classification Models - A Position Paper," ACM SIGKDD Explorations Newsletter 15, no. 1 (March 17, 2014): 1–10. doi:10.1145/2594473.2594475.

than in majority ones, predictive models will naturally be less accurate when it comes to minority groups—even when no one has set out to bias the model in this way.

To illustrate the harms that might occur because of this, consider an incident involving Google's "real name policy" from earlier this year.[6] Google created an automated system that would screen the names of new users, to prevent them from using fake names on their service. But, soon after one Native American woman named Elaine Yellow Horse signed up, Google's system automatically flagged her (real) name as fake, and suspended her account. Google's system is likely well-tuned to accurately recognize popular, common names. But on minority names that are less common, the system makes many more mistakes.

Ms. Yellow Horse eventually got Google to reverse its automated decision, but only after significant effort and media attention. For many other minorities, in a range of automated contexts including this one, they will continue to bear disproportionate burdens that stem from systematic errors.

"Garbage In, Garbage Out"

The general idea behind machine learning is that a machine learns from example data. But if the examples are biased in some ways, then the subsequent model that the machine creates will also reflect those same biases.

A recent Harvard research study clearly demonstrated this perpetuating effect.[7] The study looked at the online ads that were displayed next to Google search results. For search results for black-identifying names, it was 25% more likely that an ad suggestive of an arrest record (*e.g.*, "Ebony Jones arrested?") would be appear, as compared to ads accompanying search results for white-identifying names (*e.g.,* "Looking for Jill Jones?"). The difference, the researchers found, was statistically significant.

How and why did this occur? It's unlikely that an engineer at Google *decided* to target ads in this way. Rather, the more likely reason is that the Google's users clicked on ads in a biased way, whether subconsciously or not. The historical data about user clicks was fed into the ad system, which picked up on those same discriminatory habits.

There are technical strategies available to proactively combat the latent biases found in the historical training data. One strategy is to tweak the data that the machine is exposed to, until the predictions begin to match certain desirable outcomes. In the Google example, it may have been possible to *oversample* certain situations—by feeding the machine with more examples of users clicking on non-arrest ads for black name searches—to artificially dampen the discriminatory effect.

The effectiveness of these strategies varies from case-to-case. But while these strategies are available, they're only useful once a bias has been recognized. Detecting the many hidden biases that may be latent in a large dataset can require both significant statistical effort and civil rights

---

[6] Joe Flood, *What Happens When Google Doesn't Think You're A Human*, BuzzFeed (Mar. 6, 2014), http://www.buzzfeed.com/joeflood/what-happens-when-google-doesnt-think-youre-a-human.

[7] Latanya Sweeney, *Discrimination in Online Ad Delivery*, 56 Communications of the ACM 44, 54 (2013), *available at* http://dataprivacylab.org/projects/onlineads.

understanding; it might also depend on paying attention to individual cases where people suffer adverse results, as in the case of Ms. Yellow Horse.

Online Prices Can Be Personalized to Reflect Offline Differences

The *Wall Street Journal* has reported that some stores have begun to adjust prices online to better match the in-store prices available to different consumers.[8] In particular, the *Journal* found that Staples varied the prices on its website according to each shopper's apparent distance to a competing brick-and-mortar office supply store. The further away a shopper lived from an OfficeMax, say, the more Staples would charge for a product, effectively reintroducing the kind of pricing strategy that only seemed possible in physical stores. To figure out where its online customers lived, Staples took advantage of a basic property of the internet protocol—the so-called (internet protocol) IP address. These addresses, a unique string of numbers, identify each piece of equipment connected to the internet. These addresses are often assigned in ways that correspond to specific geographic locations, granting Staples and others the ability to determine the rough location of an online customer based on her IP address alone—and to vary prices accordingly .

This strategy, however rational from a business perspective, can have some unintended and unanticipated effects on historically disadvantaged populations. Communities of color may receive less favorable offers when shopping online because retailers are less likely to open stores in such areas.

Research Frontier: Teaching Machines Not to Discriminate

In many traditional regulatory environments, the strategy for preventing discrimination is to *prohibit* actors from collecting certain protected personal attributes, like information about a person's race or gender.[9] As the reasoning goes, if an actor doesn't know these sensitive details, then they can't discriminate on that basis.

However, in today's technological landscape, not only are such rules increasingly ineffective at preventing discriminatory outcomes, they may actually be *counterproductive* to the development of fair, automated systems. It's evident that even if sensitive attributes aren't directly collected, there are many other ways to indirectly (and accurately) infer the same information. Whether someone buys furniture coasters can be an accurate proxy for whether the person has a high credit score.[10] Web browsing behavior can accurately predict someone's race and socioeconomic status.[11] There are countless other ways to infer sensitive details about a person from seemingly innocuous information.

---

[8] Jennifer Valentino-Devries, Jeremy Singer-Vine and Ashkan Soltani, *Websites Vary Prices, Deals Based on Users' Information*, Wall St. J. (Dec. 24, 2012), *available at* http://online.wsj.com/articles/SB10001424127887323777204578189391813881534.

[9] *See*, *e.g.*, 12 C.F.R. § 202.5(b) ("A creditor shall not inquire about the race, color, religion, national origin, or sex of an applicant or any other person in connection with a credit transaction . . . ").

[10] Steve Henn, *If There's Privacy in the Digital Age, It Has a New Definition*, Nat'l Pub. Radio (Mar. 3, 2014), http://www.npr.org/blogs/alltechconsidered/2014/03/03/285334820/if-theres-privacy-in-the-digital-age-it-has-a-new-definition.

[11] Sara M. Watson, *If Customers Knew How You Use Their Data, Would They Call It Creepy?*, Harv. Bus. Rev. (Apr. 29, 2014), https://hbr.org/2014/04/if-customers-knew-how-you-use-their-data-would-they-call-it-creepy.

Given this problem, computer scientists are researching ways to design systems that can proactively detect and ignore not only the limited set of protected personal attributes, *but also* any proxies that are highly correlated with those attributes.[12] In order to accomplish this, **it may be necessary to actively collect and use the protected personal attributes**—in order to explicitly avoid them. The idea is to create a system that is "aware" that certain sensitive attributes *should not be considered* when developing patterns and making predictions. Only by doing so can we gain confidence that the system's outcomes will not be predicated on race, gender, or other protected attributes.

Research Frontier: Proving How a Decision Was Made

In many situations, a government or corporate decision-maker may be unable to provide full transparency into a decision process, but there may still be a strong need to protect civil rights. In deciding who to pull out of line for enhanced screening at the airport, or whose tax returns to audit, public officials must maintain a level of secrecy to prevent bad actors from gaming of the system. Similarly, in a commercial context, companies may want or need to prove that they followed certain rules without revealing the proprietary "secret sauce" of how particular decisions were reached.

In designing traditional accountability mechanisms, we typically don't need to choose between zero accountability, or absolute, complete transparency. Instead, we can carefully design an accountability mechanism that balances the various interests at stake.

A similar approach is possible for automated, computerized decision-making, thanks to advanced computer science methods.[13] These methods allow system designers to build partial transparency mechanisms that can keep the "secret sauce" secret, while at the same time proving that the rules were followed. The need to withhold some information does not imply that we have to give up entirely on accountability.

Policy, Technology and the Road Ahead

Many traditional policy tools, including requirements for individual consent and transparency, remain powerful ways of making sure that automated systems operate in ways that respect civil rights. At the same time, these existing tools are not a complete tool box. New methods and tools will need to be developed to ensure that we protect civil rights to the best and fullest extent.

Computer scientists and practitioners in the field are starting to map new ways to quantify and operationalize civil rights goals, within automated systems. They are starting to formalize the meaning of fairness and accountability ways that computers can understand and enforce. But they can't do it alone—this process will require a back and forth with the civil rights community.

This new opportunity for quantitative civil rights protections creates new incentives for the evolution of civil rights doctrine. U.S. law generally takes a case-by-case, evolutionary approach

---

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold and Rich Zemel, *Fairness Through Awareness* (Nov. 29, 2011), http://arxiv.org/abs/1104.3913.
[13] *See* Ed Felten, *Accountable Algorithms*, Freedom to Tinker (Sep. 12, 2012), https://freedom-to-tinker.com/blog/felten/accountable-algorithms; *see also* Ed Felten, *Accountable Algorithms: An Example*, Freedom to Tinker (Sep. 13, 2012), https://freedom-to-tinker.com/blog/felten/accountable-algorithms-an-example.

that disfavors bright line rules for what constitutes discrimination. This is particularly true of "disparate impact" cases that find civil rights liability for a pattern or practice that, while not motivated by racial or other bias, nonetheless has a disproportionate adverse impact on racial minorities or other protected status groups. This doctrine is particularly relevant in the context of big data, where automated decisions that reflect past patterns may create new disadvantage for protected status groups, despite a lack of any objectionable motives.[14]

In the face of new quantitative tools that are making potentially life-changing decisions regarding housing, health, criminal justice, education and the like, the potential for disparate impacts from outwardly neutral processes becomes a particularly urgent concern. But computerized fixes to avoid disparate impact can't rely directly on holistic, human assessments of what is or is not fair—they need numeric standards. The benefits of having precise rules about what we expect in fair decision-making are going up. There is work to be done to quantify anti-discrimination goals in ways that computers can understand.

Big data will create a new opportunity to implement those rules, to take account of differences in where people are coming from, to measure and prove the existence of the disparities, and—potentially, with the help of new techniques now being developed—to correct for those disparities.

As businesses become more data intensive and adopt new methodologies, they express a tremendous sense of optimism and opportunity. And we know that if those opportunities are pursued in a naïve fashion, then they can indeed reproduce bias. But that doesn't mean that only business can benefit from these tools and these methods. And it doesn't mean that in the long run, the application of these methods will be naïve.

Data has always played a central role in civil rights protection. The decennial Census, and the American Community Survey are critical diagnostic tools for various kinds of disparities—and they are in a real sense the original big data. They are longstanding yardsticks that can be used to measure disparities and to look at how we're doing in terms of addressing those disparities. From a civil rights point of view, the world of big data could be seen as a world that is newly suffused with measurements that can be used to detect the kinds of disparities that the civil rights community has always aimed to address. From that point, in turn, there may be new opportunities to come up with creative, actionable, computable new ways of addressing those disparities.

[14] Solon Barocas and Andrew D. Selbst, *Big Data's Disparate Impact* (Oct. 19, 2014), *available at* http://ssrn.com/abstract=2477899.