Attacking discrimination with smarter machine learning

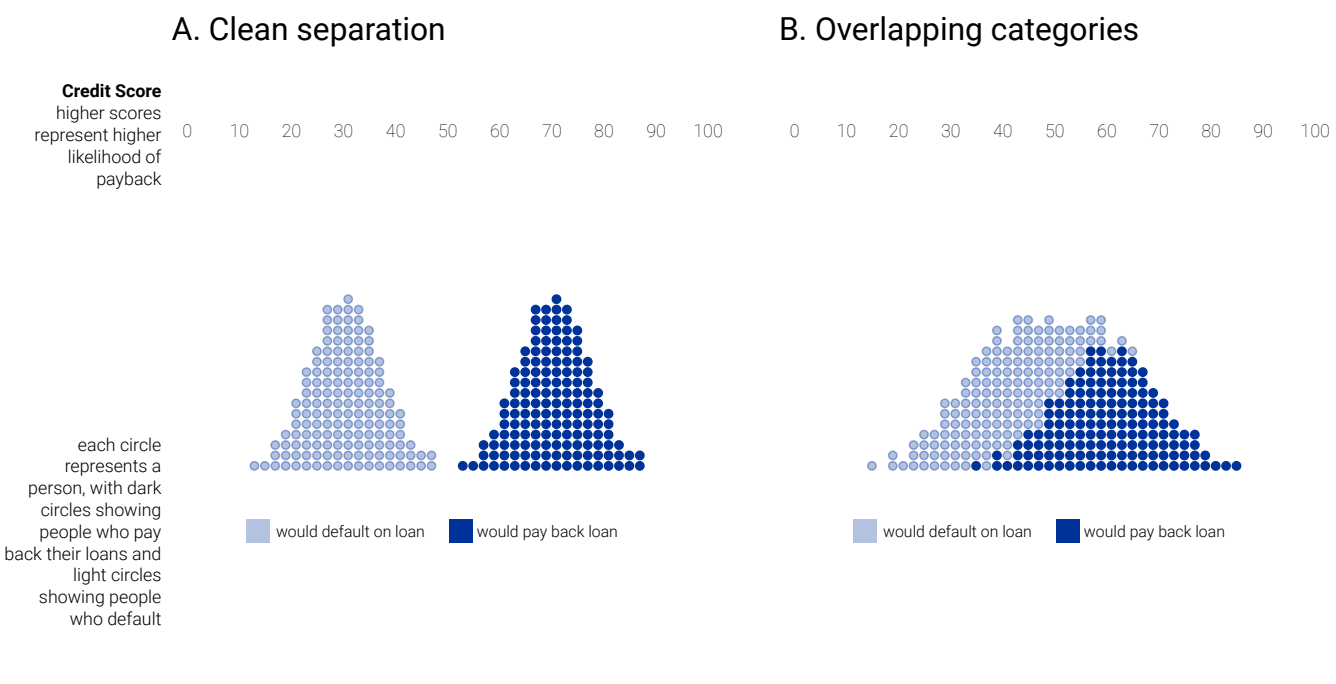
As machine learning is increasingly used to make important decisions across core social domains, the work of ensuring that these decisions aren't discriminatory becomes crucial.

Here we discuss "threshold classifiers," a part of some machine learning systems that is critical to issues of discrimination. A threshold classifier essentially makes a yes/no decision, putting things in one category or another. We look at how these classifiers work, ways they can potentially be unfair, and how you might turn an unfair classifier into a fairer one. As an illustrative example, we focus on loan granting scenarios where a bank may grant or deny a loan based on a single, automatically computed number such as a credit score.

By Martin Wattenberg, Fernanda Viégas, and Moritz Hardt.

This page is a companion to a recent paper by Hardt, Price, Srebro, which discusses ways to define and remove discrimination by improving machine learning systems.

Loan applicants: two scenarios



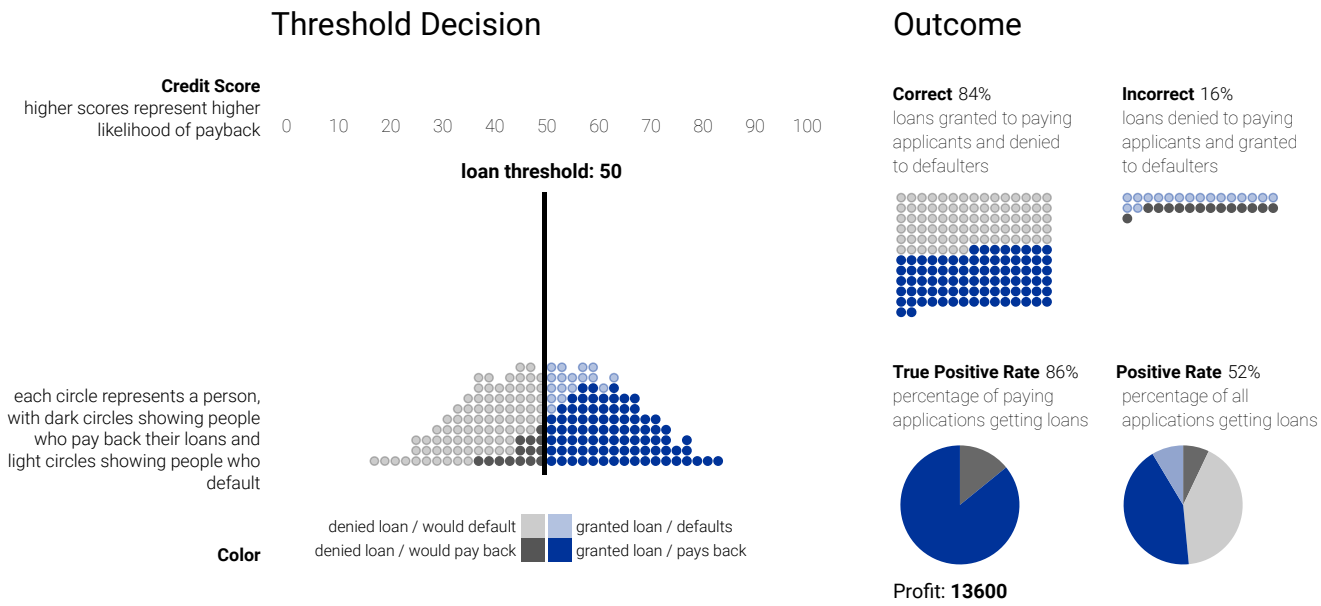
In the diagram above, dark dots represent people who would pay off a loan, and the light dots those who wouldn't. In an ideal world, we would work with statistics that cleanly separate categories as in the left example. Unfortunately, it is far more common to see the situation at the right, where the groups overlap.

A single statistic can stand in for many different variables, boiling them down to one number. In the case of a credit score, which is computed looking at a number of factors, including income, promptness in paying debts, etc., the number might correctly represent the likelihood that a person will pay off a loan, or default. Or it might not. The relationship is usually fuzzy—it's rare to find a statistic that correlates perfectly with real-world outcomes.

This is where the idea of a "threshold classifier" comes in: the bank picks a particular cut-off, or threshold, and people whose credit scores are below it are denied the loan, and people above it are granted the loan. (Obviously real banks have many additional complexities, but this simple model is useful for studying some of the fundamental issues. And just to be clear, Google doesn't use credit scores in its own products.)

Simulating loan thresholds

Drag the black threshold bars left or right to change the cut-offs for loans.



The diagram above uses synthetic data to show how a threshold classifier works. (To simplify the explanation, we're staying away from real-life credit scores or data--what you see shows simulated data with a zero-to-100 based "score".) As you can see, picking a threshold requires some tradeoffs. Too low, and the bank gives loans to many people who default. Too high, and many people who deserve a loan won't get one.

So what is the best threshold? It depends. One goal might be to maximize the number of correct decisions. (What threshold does that in this example?)

Another goal, in a financial situation, could be to maximize profit. At the bottom of the diagram is a readout of a hypothetical "profit", based on a model where a successful loan makes $300, but a default costs the bank $700. What is the most profitable threshold? Does it match the threshold with the most correct decisions?

Classification and Discrimination

The issue of how "the correct" decision is defined, and with sensitivities to which factors, becomes particularly thorny when a statistic like a credit score ends up distributed differently between two groups.

Imagine we have two groups of people, "blue" and "orange." We are interested in making small loans, subject to the following rules:

- A successful loan makes $300
- An unsuccessful loan costs $700
- Everyone has a credit score between 0 and 100

Simulating loan decisions for different groups

Drag the black threshold bars left or right to change the cut-offs for loans.
Click on different preset loan strategies.

Loan Strategy Blue Population Orange Population

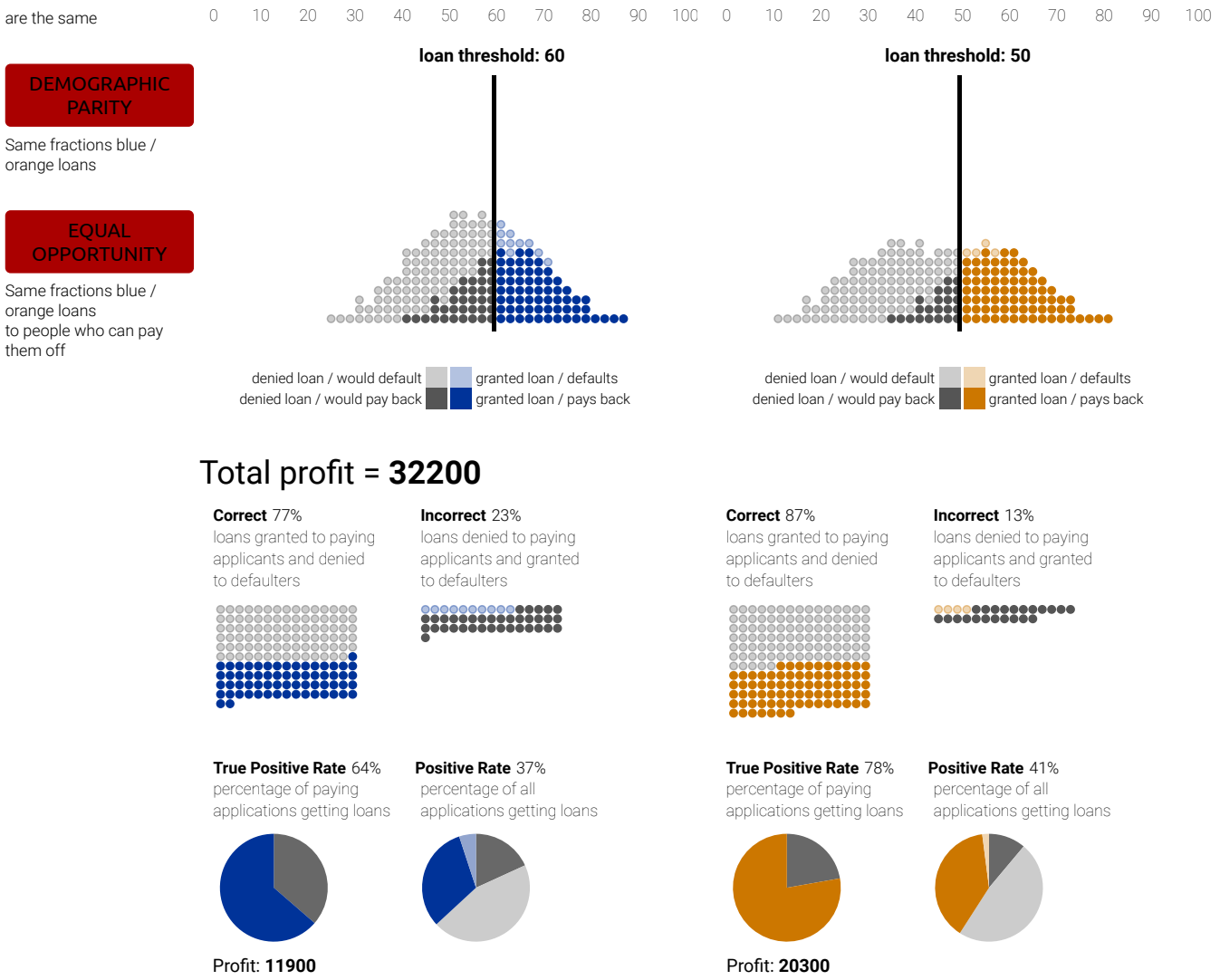
Maximize profit with:

MAX PROFIT

No constraints

GROUP UNAWARE

Blue and orange thresholds



In this case, the distributions of the two groups are slightly different, even though blue and orange people are equally likely to pay off a loan. If you look for pair of thresholds that maximize total profit (or click the "max profit" button) you'll see that the blue group is held to a higher standard than the orange one.

That could be a problem—and one obvious solution for the bank not to just pick thresholds to make as much money as possible. Another approach would be to implement a **group-unaware**, which holds all groups to the same standard. Is this really the right solution, though? For one thing, if there are real differences between two groups, it might not be fair to ignore them—for example, women generally pay less for life insurance than men, since they tend to live longer.

But there are other, mathematical problems with a group-unaware approach even if both groups are equally loan-worthy. In the example above, the differences in score distributions means that the orange group actually gets fewer loans when the bank looks for the most profitable group-unaware threshold.

If the goal is for the two groups to receive the same number of loans, then a natural criterion is **demographic parity**, where the bank uses loan thresholds that yield the same fraction of loans to each group. Or, as a computer scientist might put it, the "positive rate" is the same across both groups.

In some contexts, this might be the right goal. In the situation in the diagram, though, there's still something problematic: a demographic parity constraint only looks at loans given, not rates at which loans are paid back. In this case, the criterion results in fewer qualified people in the blue group being given loans than in the orange group.

To avoid this situation, the paper by Hardt, Price, Srebro defines a concept called **equal opportunity**. Here, the constraint is that of the people who can pay back a loan, the same fraction in each group should actually be granted a loan. Or, in data science jargon, the "true

positive rate" is identical between groups.

Improving machine learning systems

A key result in the paper by Hardt, Price, and Srebro shows that—given essentially any scoring system—it's possible to efficiently find thresholds that meet any of these criteria. In other words, even if you don't have control over the underlying scoring system (a common case) it's still possible to attack the issue of discrimination.

For organizations that do have control over the scoring system, using these definitions can help clarify core issues. If a classifier isn't as effective for some groups as others, it can cause problems for the groups with the most uncertainty. Restricting to equal opportunity thresholds transfers the "burden of uncertainty" away from these groups and onto the creators of the scoring system. Doing so provides an incentive to invest in better classifiers.

This work is just one step in a long chain of research. Optimizing for equal opportunity is just one of many tools that can be used to improve machine learning systems—and mathematics alone is unlikely to lead to the best solutions. Attacking discrimination in machine learning will ultimately require a careful, multidisciplinary approach.

Thanks to Meredith Whittaker, Daniel Smilkov, Jimbo Wilson, and our other colleagues at Google for their help with this piece.