Welcome! Ask Qs or say hi in chat before/during/after class

CSE 312

Foundations of Computing II

Lecture 29: Algorithmic Fairness

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Rachel Lin, Hunter Schafer

Slide Credit: Based on Jamie Morgenstern's Slides for CSE 446, 20au

Music: Tom Misch

Agenda

- Why Fairness?
- What is “fair”?
- No fair lunch

Machine Learning (ML) in a Slide

A common task in ML is a prediction task:

- Gather data from past about task [training data]
- Use an algorithm to find a model that (hopefully) makes accurate predictions [training / learning]
- Deploy the model to the "real world" and have it make predictions about new data [prediction / inference]

Traditionally, model that is "most accurate" is the one selected.

- Very difficult to find a perfect model in practice, mistakes are expected!

Examples of ML Systems Gone Wrong

INDEPENDENT

GOOGLE'S ALGORITHM SHOWS PRESTIGIOUS JOB ADS TO MEN, BUT NOT TO WOMEN

REUTERS

Business Markets World Politics TV More

BUSINESS NEWS OCTOBER 9, 2018 / 8:12 PM / 5 MONTHS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

The New York Times

Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says

By Katie Benner, Glenn Thrush and Mike Isaac

March 28, 2019

168

MIT Technology Review

Intelligent Machines

How to Fix Silicon Valley's Sexist Algorithms

Computers are inheriting gender bias implanted in language data sets—and not everyone thinks we should correct it.

PROPUBLICA

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

Common theme: Rates of predictions $\{+, -\}$ vary across race, gender, differing abilities, age, …

In other words, an individual is more likely to receive a different (usually more harmful) outcome due to a factor outside of their control.

In many settings (e.g., credit scores, housing, employment) have clear legal protections against differential treatment based on some ‘protected’ attributes.

Why Different Treatment Happens?

More often than not, programmers aren't *intentionally* coding up racism or sexism into their models. Where does this difference in treatment come from?

Most often: the data

- Training a model on biased data will likely result in a biased model
- Where does biased data come from?
  - A world with deeply entrenched structural barriers that make equality a challenging goal to achieve

Goal: A mathematical definition of fairness

Can we come up with a mathematical definition for what it means for the outcomes of a model to be fair/unfair?

– If we have a mathematical definition to spot when a model will be unfair, we can avoid deploying it before it can potentially harm people.

Key Idea: There is not going to be one definition of fairness. Every definition encodes its own set of values for what fairness means. There is not one universal definition, but which one you choose is an important value statement.

Agenda

- Why Fairness?
- What is "fair"?
- No fair lunch

Group Fairness

Today, we are going to be limiting our discussions to notions of **group fairness** which is also called **non-discrimination**.

- Goal: Avoid differential treatment based on membership of some protected group.

A **group** is some defined shared attribute between individuals that we want to ensure non-discrimination.

- Examples: race, gender, different abled, age, etc.
- Can also consider defining groups as intersectional identities

Running Example: College Admissions

We are going to use an example of admitting students to a college.

- Not an endorsement of such a system. Use it since it is simple to describe and lets us better understand definitions of fairness.
- Will make MANY simplifying assumptions in our discussion to focus on understanding.

Assumptions (unrealistic):

- There is a definition of "success" for college applicants, and the goal of an admissions decision is a prediction of "success"
- The only thing we will use as part of our decision is SAT Score
- To talk about group fairness, will assume everyone belongs to exactly one of two races: Circles (66%) or Squares (33%).

Notation

Example: College admissions only using SAT score

X input about a person for prediction
- Example: X = SAT Score

A variable indicating which group X belongs in
- Example: $A = \square$ or $A = \bigcirc$

Y the "true label"
- Example: $Y = +$ if truly successful in college, $Y = -$ if not

$\hat{Y} = f(X)$ is our prediction for Y using a learned model f
- Example: $\hat{Y} = +$ if predicted successful, $\hat{Y} = -$ otherwise

First Attempt: “Shape-blind”

First attempt: To avoid unfair decisions, prevent the model from ever seeing the protected attribute (e.g., if an applicant is Circle/Square).

Doesn’t work: In the real world, many things are correlated with race. Protected attribute can be unintentionally inferred from many other sources.

Fairness Definition 1: Statistical Parity

Idea: "Admit decisions are equivalent across groups."

$$\Pr(\hat{Y} = +|\, A = \square) = P(\hat{Y} = +|\, A = \bigcirc)$$

Pros:

- Aligns with certain legal definitions of equity

Cons:

- Rather weak in requirements (sell-fulfilling prophecy)

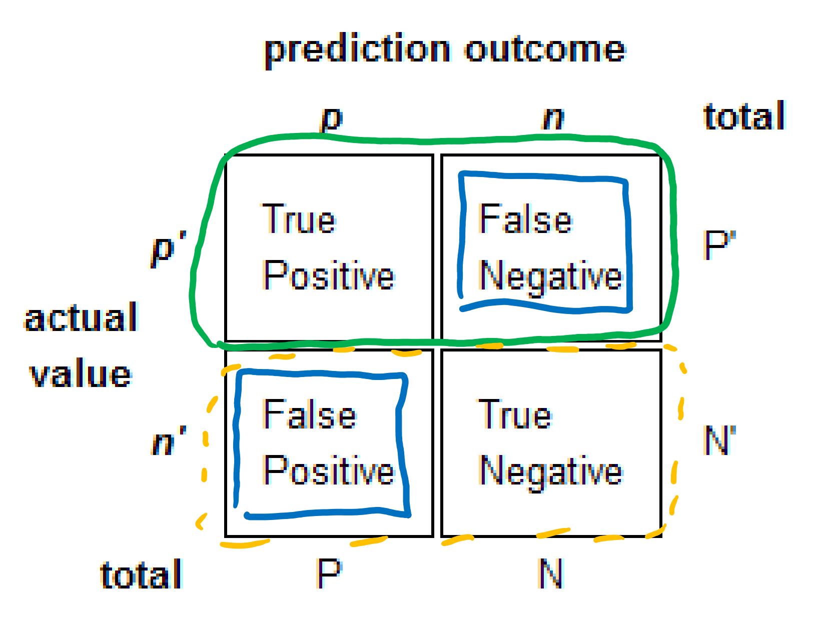
Types of Mistakes

The weakness of Statistical Parity comes from the fact it doesn't care about the true labels.

A stronger definition of fairness could require that the types of mistakes we make across groups are equal.

$$TPR = \frac{TP}{FN + TP} = 1 - FNR$$

$$TNR = \frac{TN}{FP + TN} = 1 - FPR$$



Fairness Definition 2: Equal Opportunity

Idea: The true positive rate should be equivalent across groups.

$$\Pr\left(\hat{Y} = +\mid A = \square, \underbrace{Y = +}\right) = \left(\hat{Y} = +\mid A = \bigcirc, \underbrace{Y = +}\right)$$

Pros:

- Better controls for true outcome

Cons:

- More complex to explain to non-experts
- Really only works in scenarios where there is a well defined $Y = +$

And many, many more

List of demographic fairness criteria

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

Which definition to use?

We can’t tell you! Each definition makes its own statement on what fairness means. Choosing a fairness measure is an explicit statement of what values we hold when thinking about fairness.

Takeaway: Discrimination in ML models is a crucial problem we need to work on. It’s not a problem that will only be solved algorithmically. We need people (e.g., policymakers, regulators, philosophers, developers) to be in the loop to determine the values we want to encode in the system.

Brain Break

Agenda

- Why Fairness?
- What is “fair”?
- No fair lunch

(Im)possibility of Fairness

Four reasonable conditions we want in a real world ML Model:

1. Statistical Parity
2. Equality across false negative rates
3. Equality across false positive rates
4. Good accuracy of the model across subgroups

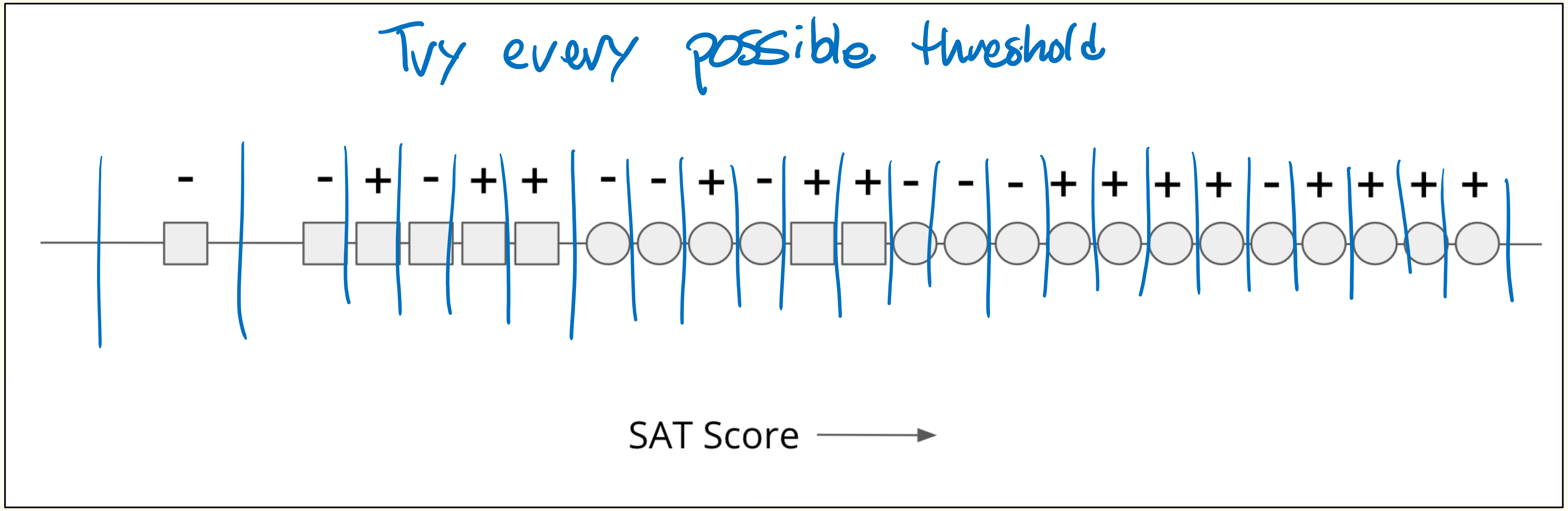
In general, can't satisfy all 4 simultaneously *unless* groups have the exact same underlying distribution.

– This condition is rarely met in practice as we mentioned earlier when there are so many places for bias to enter our data collection.

Example

Continuing overly simplistic college admissions example, with a fake dataset.

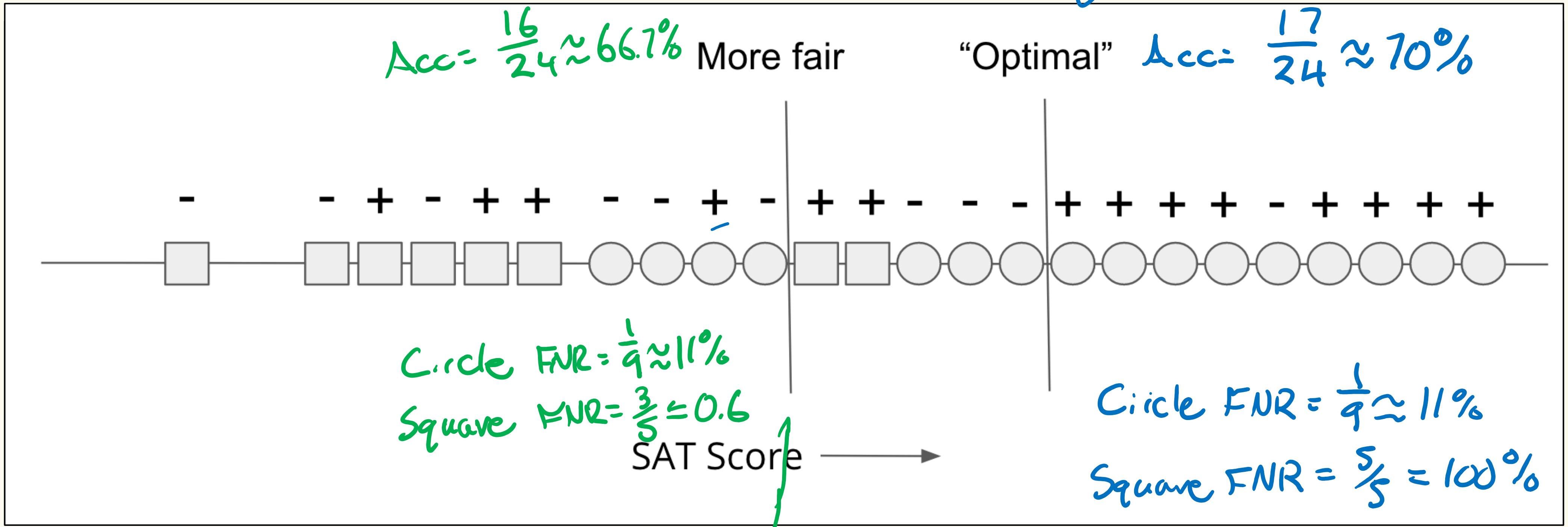
- Majority (2/3) are Circle, the remaining 1/3 are Square
- SAT score for Circles tends to be inflated when compared to Squares. Possibility: Systematic barriers and access to SAT Prep
- Even though we see statistical differences between groups in our data, the rate in which they are actually successful is the same.



Fairness-Accuracy Tradeoff

In general, we find there is a tradeoff between accurate models and fair models. Making a model more fair tends to decrease accuracy by some amount.

Accurate, but not fair



More fair, but less accurate

Notes of Tradeoff

Might argue that my example is overly simplistic (it is!), but I'll claim this is a proof of concept. We saw lots of examples of "accurate" models that were unfair.

This is not a statement that a tradeoff must exist, it just generally happens in real-world datasets.

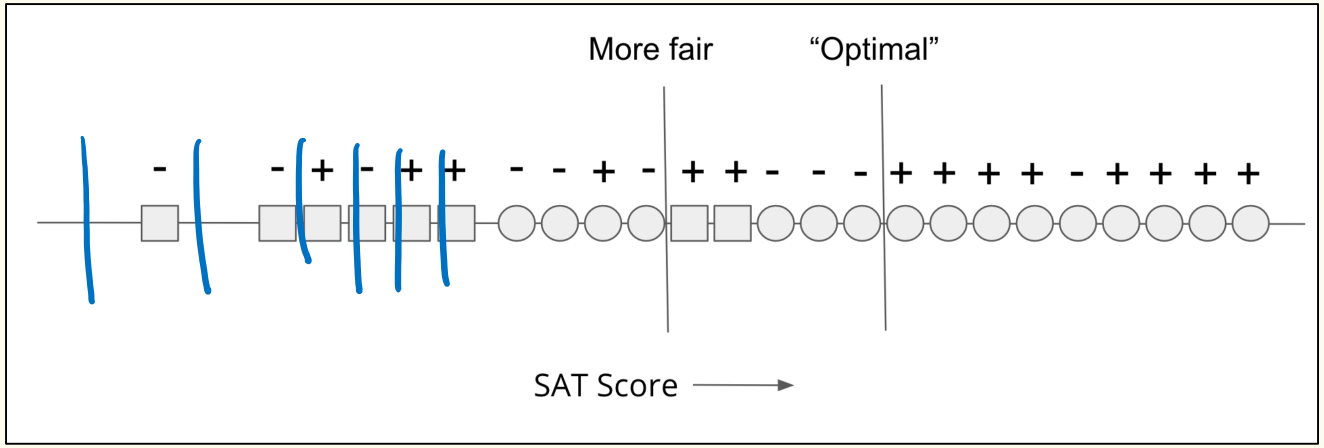
– Originally just cared about finding the most accurate model, saw unfairness as a byproduct. Controlling for fairness will yield a different model than you found before.
– If we recognize data can encode biases and accuracy is determined in terms of that data, trying to achieve fairness will likely hurt accuracy.
  - In the example before, the artificial difference in SAT scores caused the problem.

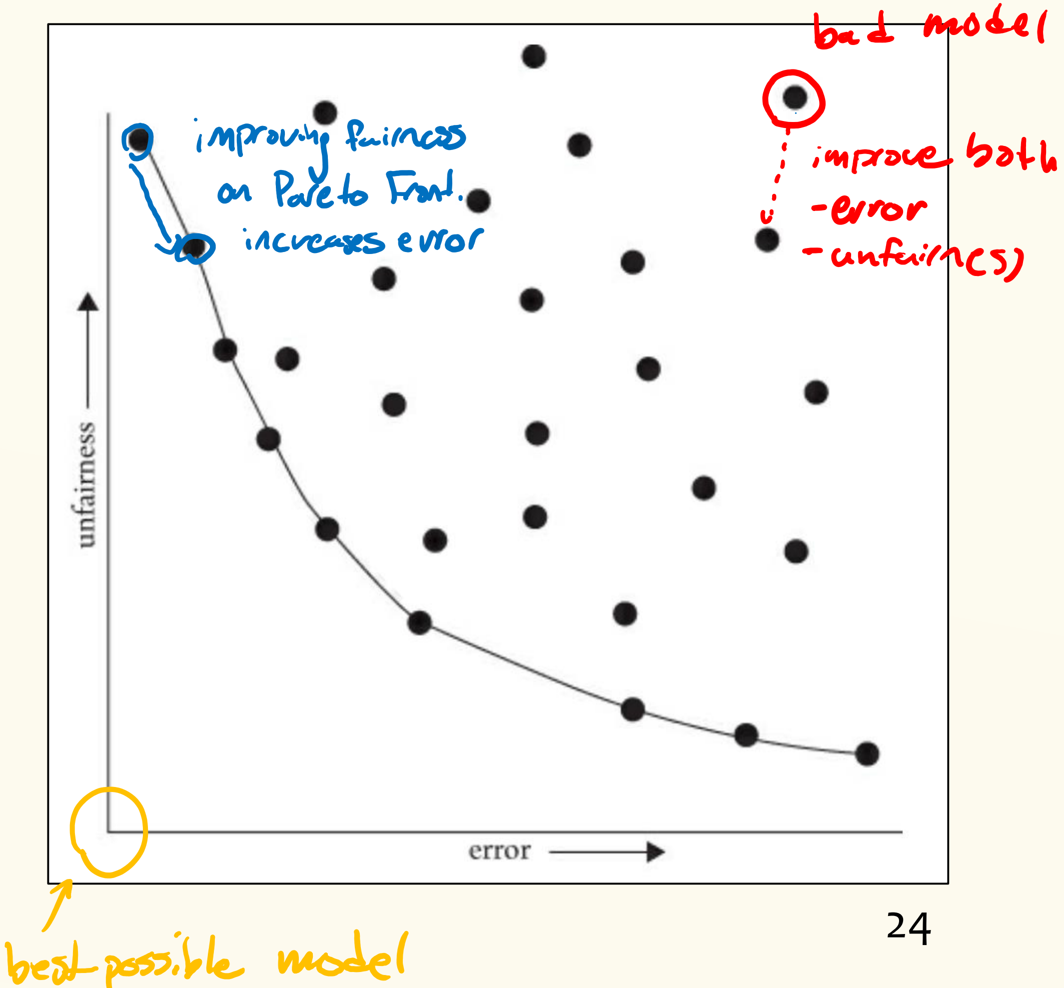
Note: Can sometimes be solved w/ diff threshold for each group. Illegal in some applications to do that!

Pareto Frontier

Visualizing the tradeoff between fairness and accuracy

– **Does not** tell you which tradeoff is appropriate!





On the Pareto Frontier

This feels a bit cold-hearted, it's okay to like this is weird. Michael Kearns and Aaron Roth write in *The Ethical Algorithm*

> While the idea of considering cold, quantitative trade-offs between accuracy and fairness might make you uncomfortable, the point is that there is simply no escaping the Pareto frontier. Machine learning engineers and policymakers alike can be ignorant of it or refuse to look at it. But once we pick a decision-making model (which might in fact be a human decision-maker), there are only two possibilities. Either that model is not on the Pareto frontier, in which case it's a "bad" model (since it could be improved in at least one measure without harm in the other), or it is on the frontier, in which case it implicitly commits to a numerical weighting of the relative importance of error and unfairness. Thinking about fairness in less quantitative ways does nothing to change these realities—it only obscures them.
>
> Making the trade-off between accuracy and fairness quantitative does ***not*** remove the importance of human judgment, policy, and ethics—it simply focuses them where they are most crucial and useful, which is in deciding exactly which model on the Pareto frontier is best (in addition to choosing the notion of fairness in the first place, and which group or groups merit protection under it, […]). Such decisions should be informed by many factors that cannot be made quantitative, including what the societal goal of protecting a particular group is and what is at stake. Most of us would agree that while both racial bias in the ads users are shown online and racial bias in lending decisions are undesirable, the potential harms to individuals in the latter far exceed those in the former. So in choosing a point on the Pareto frontier for a lending algorithm, we might prefer to err strongly on the side of fairness—for example, insisting that the false rejection rate across different racial groups be very nearly equal, even at the cost of reducing bank profits. We'll make more mistakes this way—both false rejections of creditworthy applicants and loans granted to parties who will default—but those mistakes will not be disproportionately concentrated in any one racial group.

Recap

- ML (and human) systems can exhibit unfair behavior, that can have tangible harms on people.
- There are tools to define mathematically how to spot if a model is unfair. There are many definitions, and which you choose is a statement of values.
- In practice, there is generally a tradeoff between the fairness and accuracy of an ML model. Viewing the Pareto Frontier can help you visualize what this tradeoff looks like, but deciding which tradeoff is appropriate is yet another statement of values.

Throughout: This is not something that will be solved "magically" by technology. Unfairness in ML is a reflection of societal injustices, and we need people (and potentially also algorithms) in the process of achieving equity.