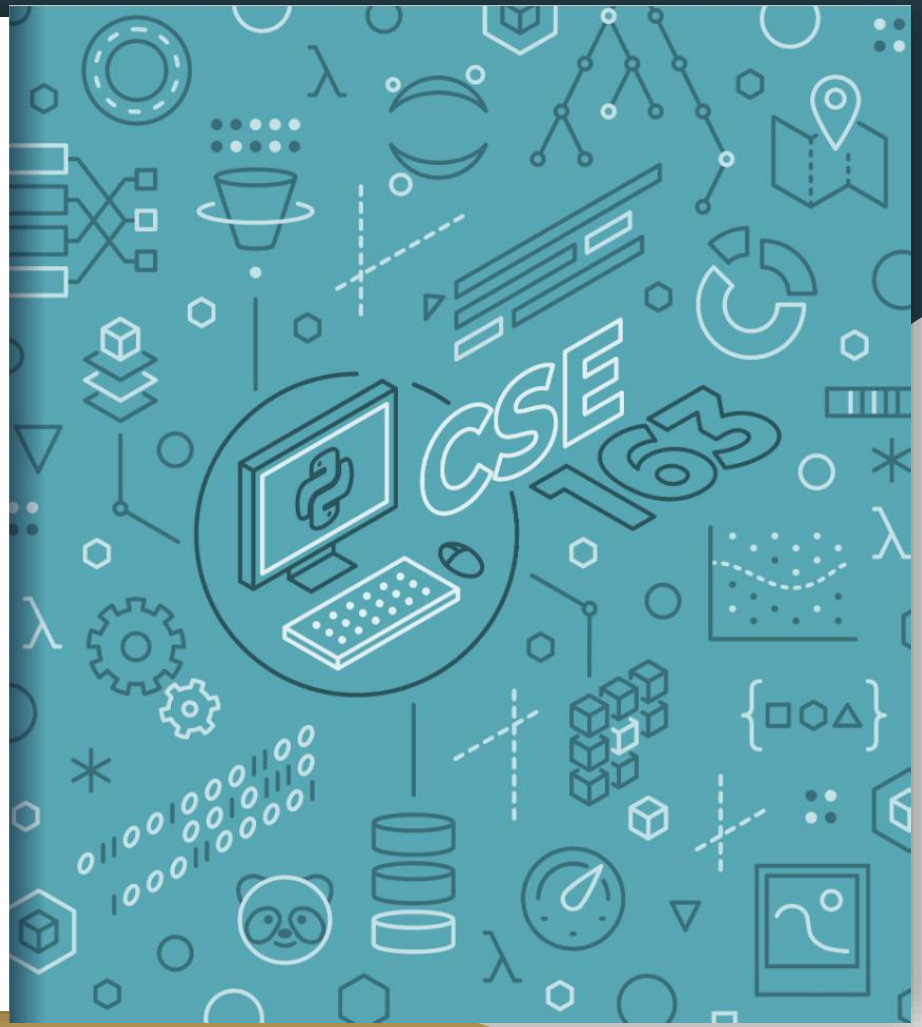
CSE 163
Fairness

Adrian Salguero
Spring 2026

💭 Icebreaker (discuss with neighbors):
Fiction or non-fiction books?
Add to our Slido!

slido.com
#cse163



Announcements

- **Take Home Assessment 5: Mapping** due Tuesday May 26th at 11:59pm
- **Reading Assignment 5** due Tuesday, May 26th at 11:59pm!
- **Lesson 22 Canvas Quiz** due tonight at 11:59pm!
- **Project Part 3** due June 1st at 11:59pm on Gradescope!

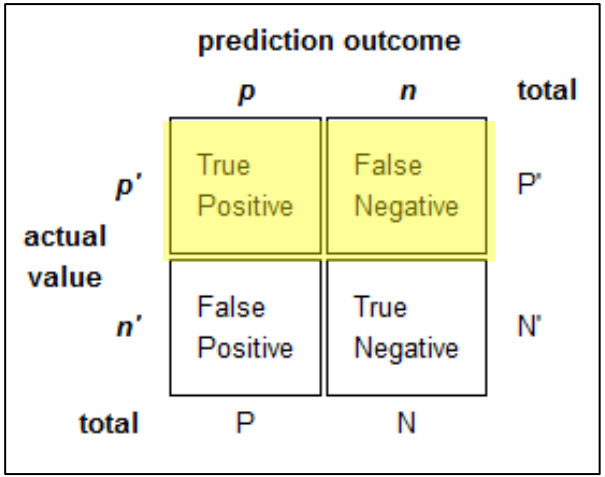
Group Fairness

- ***Intent***: Avoid discrimination against a particular group, as to avoid membership in the group negatively impact outcomes for people in that group.
  - Does not say which groups to protect, that's a decision of policy and societal norms
  - Can be extended to notions of belonging to multiple identities (e.g. intersectionality), but we focus on protecting a single property at this time.
- Usually defined in terms of the ***mistakes*** the system might make.

Definition of Fairness*

- ***Equality of False Negatives (equal opportunity)***: False negative rate should be similar across groups.

$$FNR = \frac{FN}{TP + FN}$$



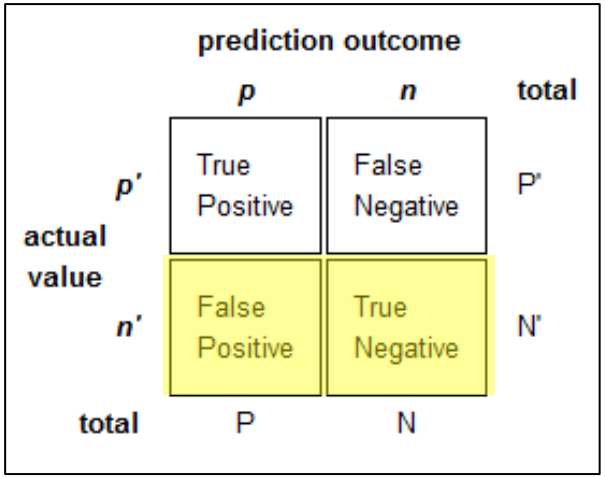
College admission example: P = Successful in college, N = Not successful in college

* Many others exist, many are in form of equations on this confusion matrix. There are other notions of fairness too!

Definition of Fairness*

- ***Equality of False Positives (predictive equality)***: False positive rate should be similar across groups.

$$FPR = \frac{FP}{TN + FP}$$



* Many others exist, many are in form of equations on this confusion matrix. There are other notions of fairness too!

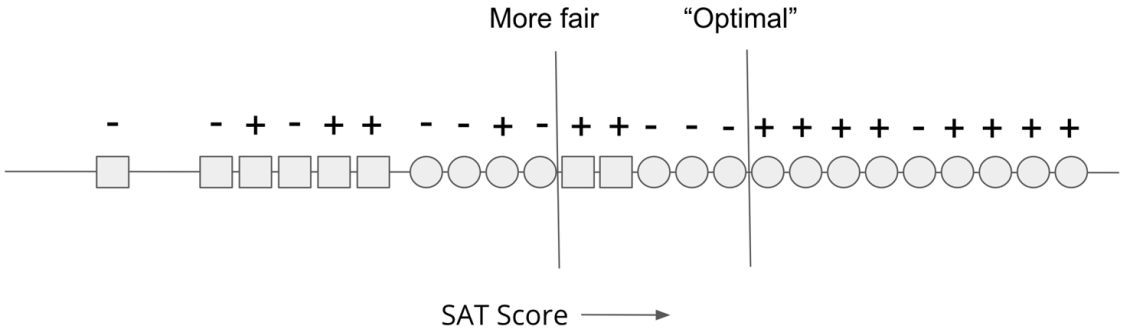
College admission example: P = Successful in college, N = Not successful in college

Human Choice

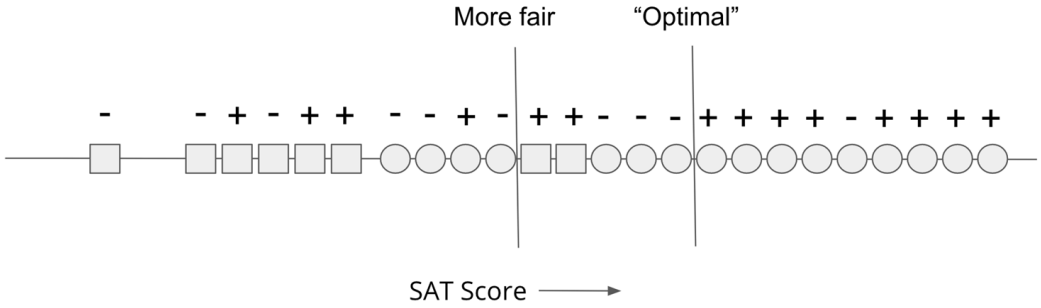
- There is no one “right” definition of fairness. They are all valid and are simply statements of what you believe fairness means in your system.
- It’s possible for definitions of fairness to contradict each other, so it’s important that you pick the one that reflects your values.
- ***Emphasizes the role of people in the process of fixing bias in ML algorithms***
  - Algorithms will do their best to optimize what is asked of them
  - This may lead to unforeseen consequences or behavior

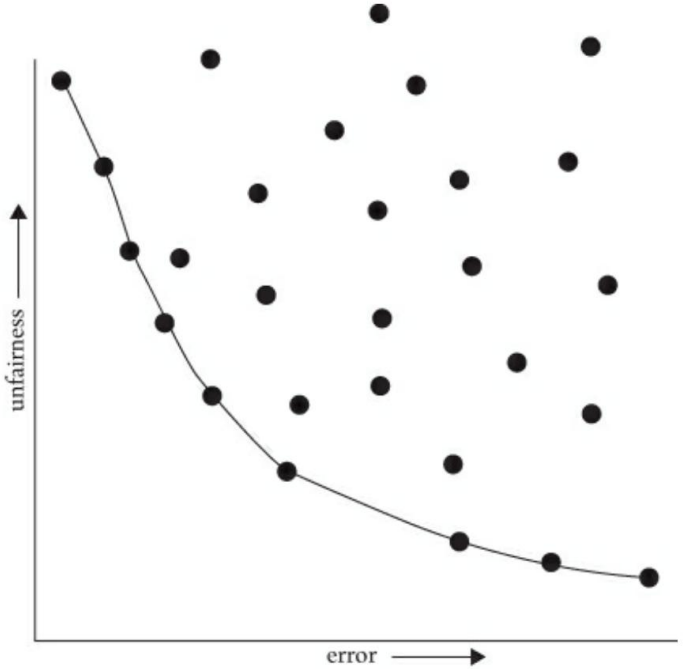
Tradeoff between Fairness and Accuracy

- We can’t get fairness for free, generally finding a more fair model will yield to one that is less accurate.
  - Intuition: We saw lots of examples where bias was a byproduct of an “accurate” model since that model was not trained with fairness in mind.



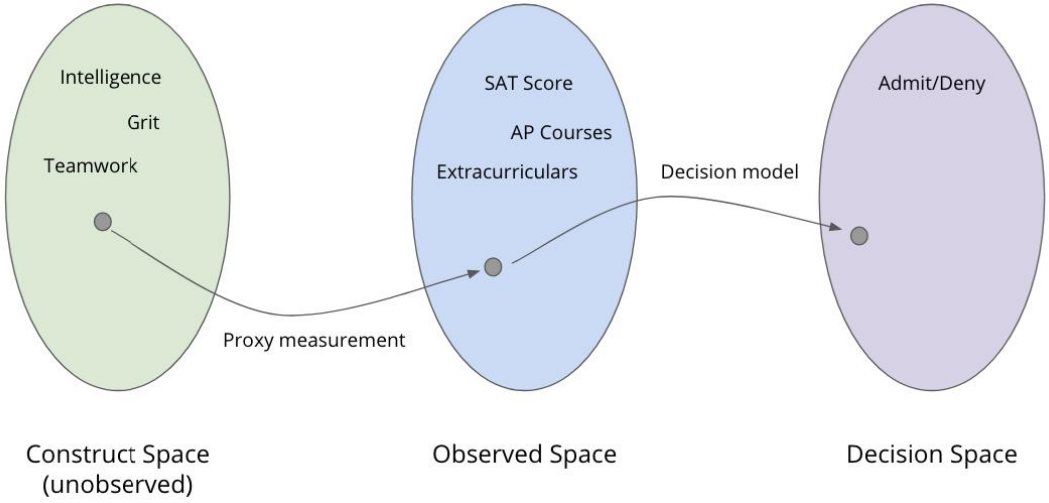
Pareto Frontiers



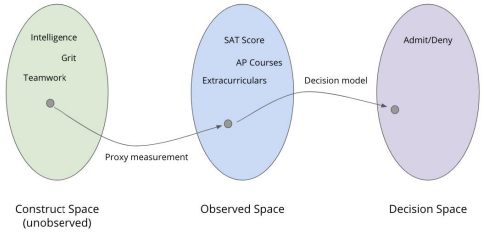


Fairness Worldviews - College Admissions

- We want to measure abstract qualities about a person (e.g. intelligence, grit), but real life measurements may not measure abstract qualities well.
- Only have access to ***Observed Space*** and we hope it's a good representation of ***Construct Space***

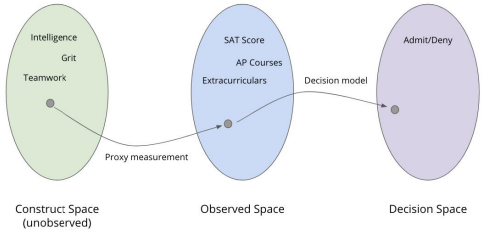


Worldview 1: WYSIWYG



- Worldview 1: What You See is What You Get (WYSIWYG)
  - Assumes the Observed Space is a good representation of the Construct Space
- Under this worldview, can guarantee ***individual fairness***. Individual fairness says if you people are close in the Construct Space, they should receive similar outcomes
  - Easy to verify under this worldview since you can use the Observed Space as a good representation of the Construct Space
- What are some drawbacks of this worldview?

Worldview 2: Bias + WAE



- Worldview 2: Structural Bias and We're All Equal (WAE)
  - Assumes systematic or social systems make different groups that look similar in the Construct Space look more different in the Observed Space
  - Example: SAT scores for one group may be artificially high due to better ability to afford SAT prep. Factors outside of qualities of interest now affect our measurements. So we assume any observed differences between groups are systematic factors rather than inherent factors since WAE.

- Goal in this worldview is to ensure ***non-discrimination*** so that someone isn't negatively impacted by simply being a member of a particular group
  - This is the implicit assumption we were making when discussion notions of group fairness earlier

Contrasting Worldviews

- Unfortunately, there is no way to tell which worldview is right for a given problem (no access to Construct Space). ***The worldview is a statement of beliefs.***
- ***WYSIWYG*** can promise individual fairness but methods of non-discrimination will be individually unfair under this worldview.
- ***Structural Bias + WAE*** can promise non-discrimination. Methods of individual fairness will lead to discrimination (since using biased data as our proxy for closeness will lead to a skewed notion of individually fair).

Case Studies

- With people around you, choose the COMPAS or the “Predicting Criminlity” case study from the lesson.
- Create a slide in the Google Slides document here
- Share your responses to the “Food for Thought” questions in the lesson.

Other discussion questions you may discuss and respond to:

1. Which worldview do you think your case study assumes? (WYSIWYG or WAE, or neither?) Why or why not?
2. What is the meaning of a false positive or false negative in your case study? What are the impacts?