Fairness and Wrap-Up

CS 344 – 2022-04-21

Automating High-Stakes Decisions

- What credit score should be required to get a certain loan?
- Which defendants should be freed until their trial?
- Which candidates should get a certain job?
- Which students should get into a university?

Think of some examples of *unfair* decisions.

How can you tell that the decision is unfair?

Main Point

- Every policy is “unfair” by some definition.
- Different stakeholders may care about different definitions
- Different worldviews underlie different definitions

Recidivism Prediction

- Bail is archaic (the rich can go free)! Is there an objective alternative?
- **Idea**: release people unlikely to commit a crime before their trial

Northpointe COMPAS machine learning system

- **Data**: criminal records, demographics
- **Prediction**: will the defendant be arrested again in ≤ 2 years?

What could possibly go wrong?

PROPUBLICA

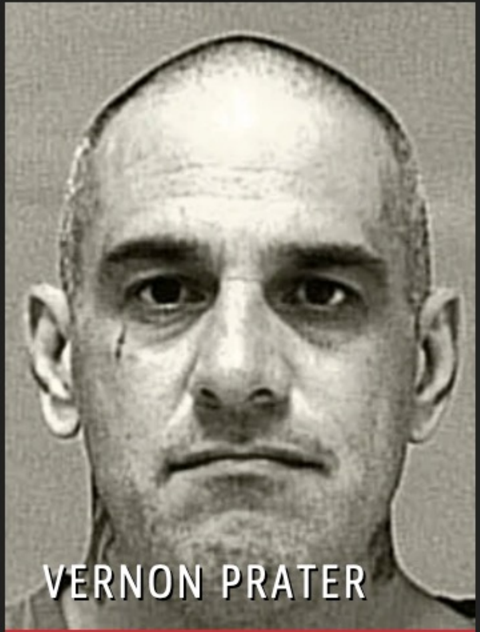
Machine Bias

the country to p
against bla

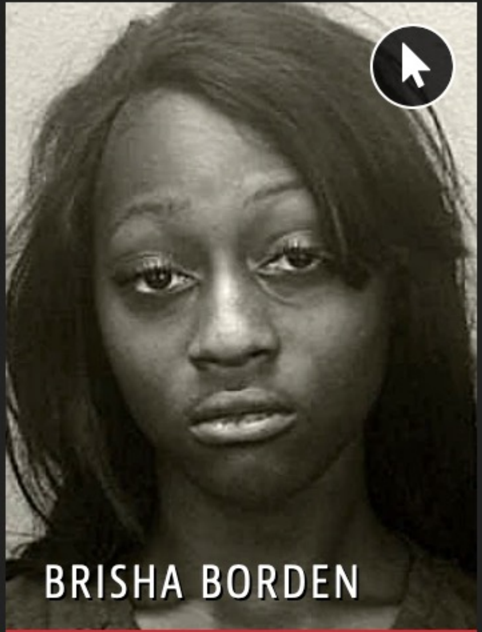
f Larson, Surya Mattu

May 23, 201

Two Petty Theft Arrests



LOW RISK 3

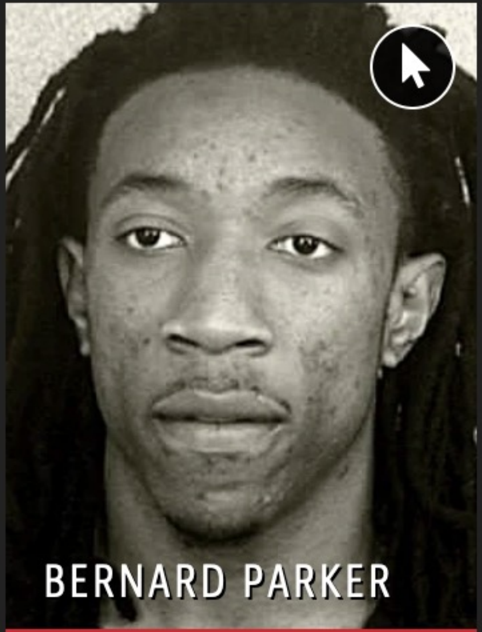


HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests

LOW RISK 3



HIGH RISK 10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

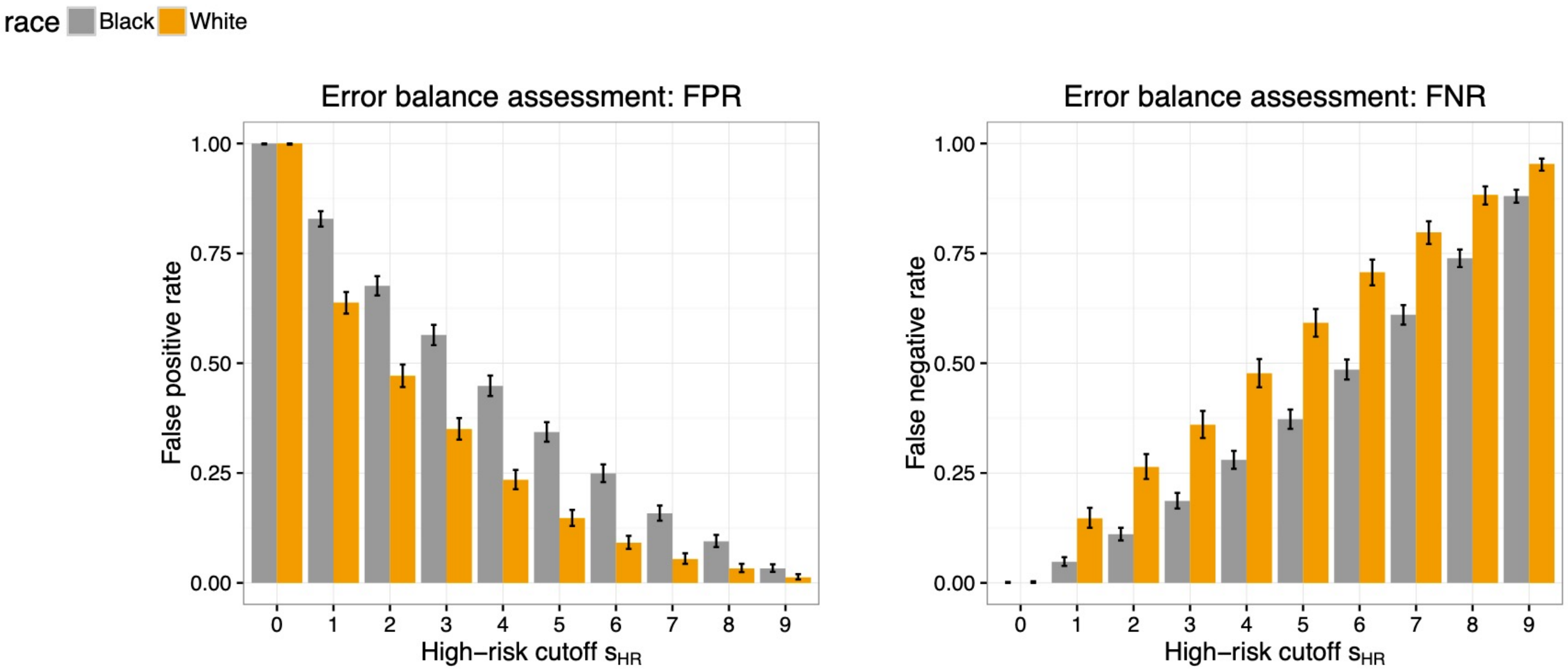
Confusion Matrix

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

https://en.wikipedia.org/wiki/Sensitivity_and_specificity

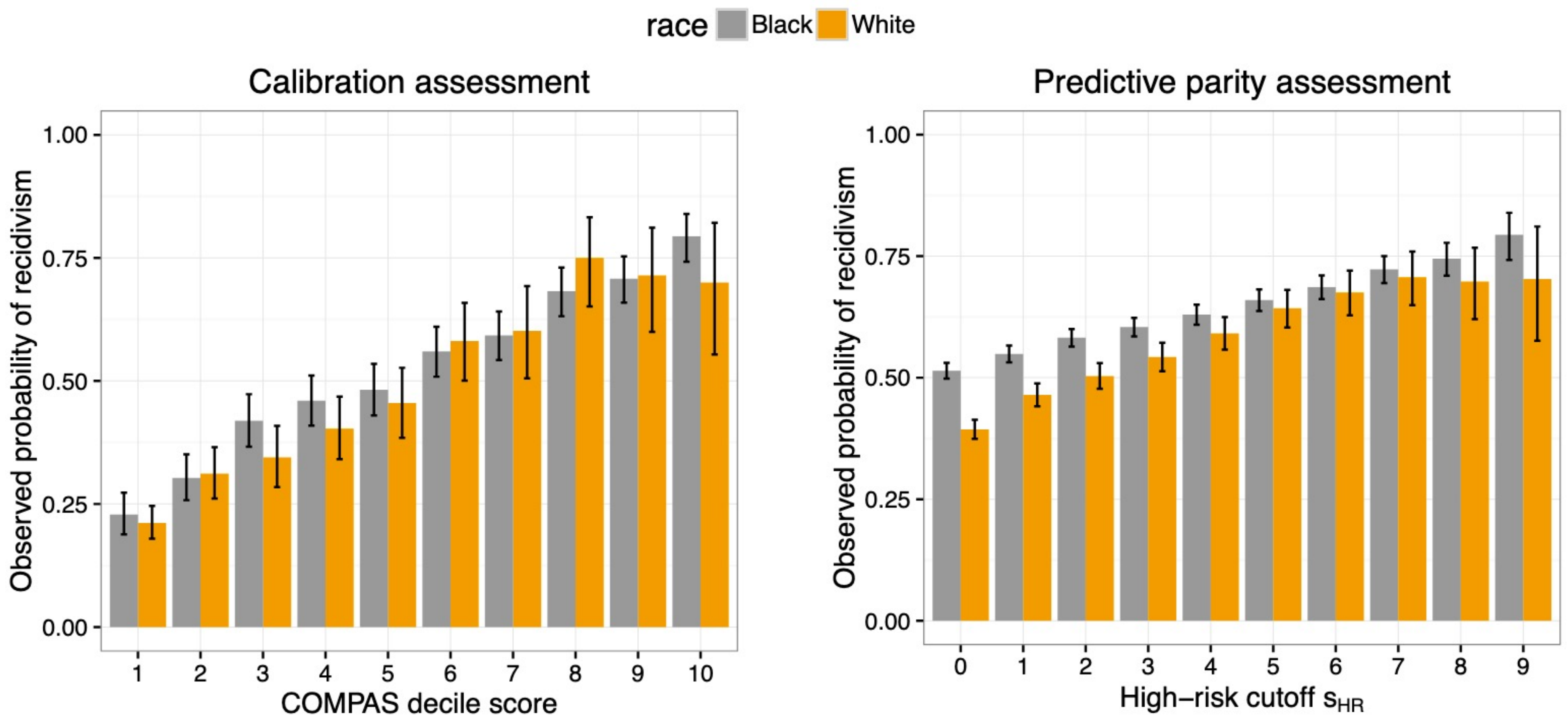
Confusion Matrix

		True condition		
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$
Predicted condition	Predicted condition positive	**True positive**	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection, Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$



(c) Bars represent observed false positive rates, which are empirical estimates of the expressions in (2.3): $\mathbb{P}(S > s_{\mathrm{HR}} \mid Y = 0, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \ldots, 9\}$

(d) Bars represent observed false negative rates, which are empirical estimates of the expressions in (2.4): $\mathbb{P}(S \leq s_{\mathrm{HR}} \mid Y = 1, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \ldots, 9\}$



(a) Bars represent empirical estimates of the expressions in (2.1): $\mathbb{P}(Y = 1 \mid S = s, R = r)$ for decile scores $s \in \{1, \dots, 10\}$.

(b) Bars represent empirical estimates of the expressions in (2.2): $\mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = r)$ for values of the high-risk cutoff $s_{\mathrm{HR}} \in \{0, \dots, 9\}$

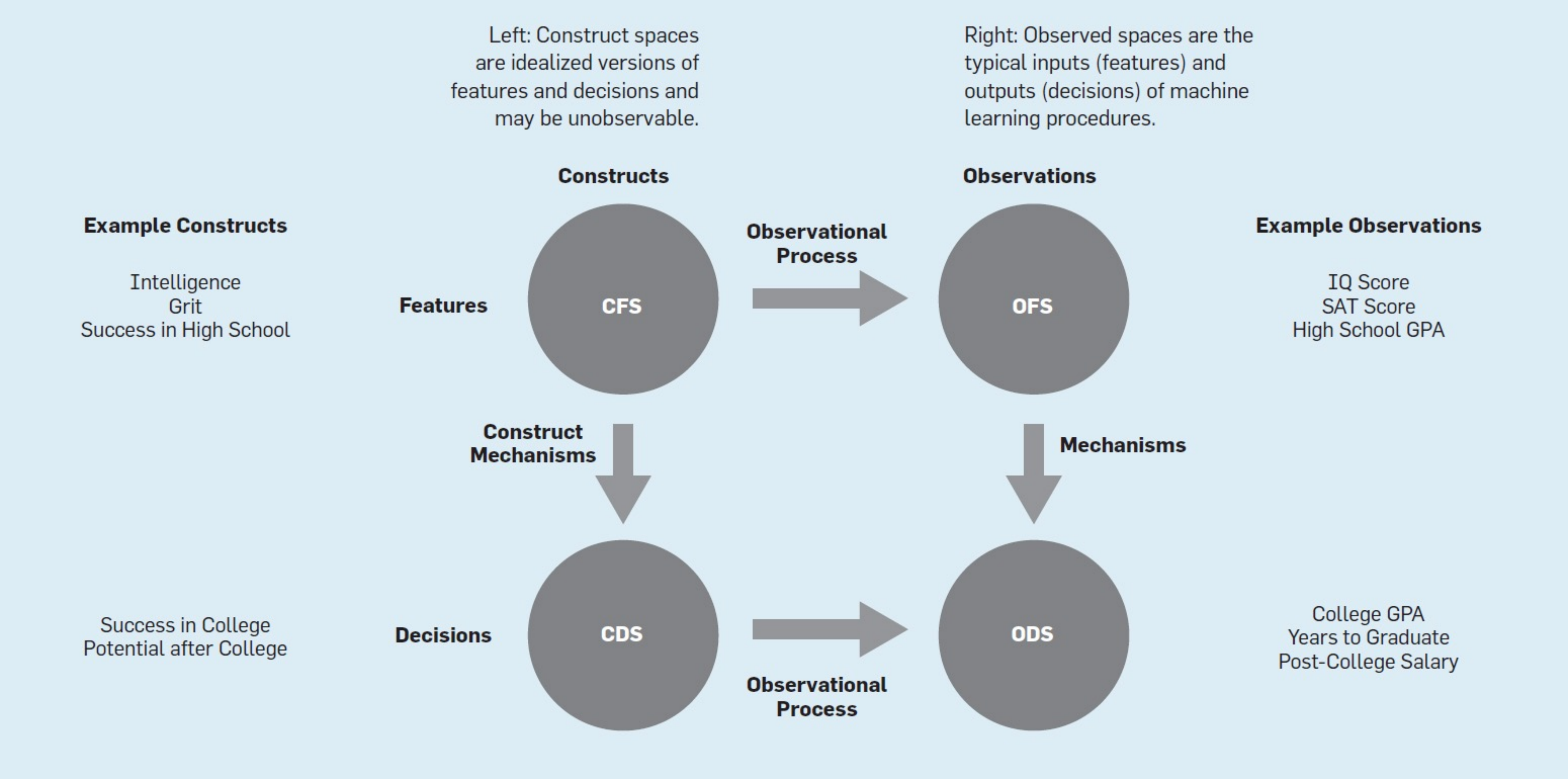
Chouldechova 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments

Automating High-Stakes Decisions

- What credit score should be required to get a certain loan?
- Which defendants should be freed until their trial?
- Which candidates should get a certain job?
- Which students should get into a university?

Characteristics of these decisions

- Decisions try to predict the future
  Will they repay the loan? Commit another crime? Succeed at the job?
- Quantifiable measures used
  Income, past arrests, …
- Measures may not reflect reality
  Arrests ≠ crimes, current income ≠ ability to pay
- Measures may not reflect our values
  Historical redlining -> lower current income
  Incarceration weakens families, …
- AI systems encode a policy



Friedler et al., 2021. [The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making]

Goals

- Individual fairness: similar people get similar decisions
- Non-discrimination: similar groups get similar decisions

- Algorithm (or policy) should aim to equalize, across groups:
    - Error rate
    - False positive rate
    - False negative rate
    - % classified positive
    - …

Impossibility of Fairness(?)

Pick no more than 2 of:

- Predictive parity (if it says 40% will recidivate, about 40% recidivate)
- Demographic parity (offer jobs to Black / White at same rate)
- Equal false positive rates
- Equal false negative rates
- Equal accuracy
- …

Chouldechova 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments; extended based on slides by Arvind Narayanan

Are there true differences between groups?

WYSIWYG

- **Individual fairness**: “Treat similar individuals similarly”

- Measures are good enough (even if groups differ)

WAE

- **Group fairness**: “Equalize the outcomes between groups”

- Each group has equal merit
  - so any differences in measures is a flaw to correct

Friedler et al., 2021. [The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making]()

Tell my people their transgression and the house of Jacob their sins.
They seek me day after day and delight to know my ways,
like a nation that does what is right and does not abandon the
justice of their God. They ask me for righteous judgments; they
delight in the nearness of God." …
Isn't this the fast I choose:
To break the chains of wickedness, to untie the ropes of the yoke,
to set the oppressed free, and to tear off every yoke?
Is it not to share your bread with the hungry,
to bring the poor and homeless into your house,
to clothe the naked when you see him,
and not to ignore your own flesh and blood?
…
If you get rid of the yoke among you,
the finger-pointing and malicious speaking,
and if you offer yourself to the hungry, and satisfy the afflicted one,
then your light will shine in the darkness,
and your night will be like noonday.

Isaiah 58 (CSB)

What sorts of automated decision policies would satisfy Isaiah's demands?

What fairness worldview is Isaiah using?

Are human lives predictable?

Hundreds of researchers attempted to predict six life outcomes, such as a child's grade point average and whether a family would be evicted from their home. These researchers used machine-learning methods optimized for prediction, and they drew on a vast dataset that was painstakingly collected by social scientists over 15 y. **However, no one made very accurate predictions.** For policymakers considering using predictive models in settings such as criminal justice and child-protective services, these results raise a number of concerns. Additionally, researchers must reconcile the idea that they understand life trajectories with the fact that none of the predictions were very accurate.

Salganik et al. 2020. [Measuring the predictability of life outcomes with a scientific mass collaboration]

Review of Christian perspectives

God made a data-rich world

- A rich environment for us to explore and learn
- Our senses should lead us to worship
  - Romans 1
  - Psalm 19
  - Romans 10

God expects us to use our intelligence

- Part of the **image of God**
- We’re commanded to **see**, **hear**, **remember**, use our **minds**, …

But we have misused our intelligence

- Selfish accumulation of data, power, …
- Designing for engagement over thoughtfulness, love
- Surveillance replaces relationships
- Over-quantification
- Exploitation
- How do we treat those who can’t repay us?

Jesus redeems our technology

- Serve others, hold organizations accountable, protect environment
- Care about other people and cultures

How else?

Final Discussions

Personal Impacts

- How AI has impacted my life in the past few years. For better? For worse?
- How AI has impacted the lives of people unlike me.
- How AI might impact our lives in the next 5 years.

Development

- Something useful or cool that has recently become possible thanks to AI.
- What are some things that AI systems are already better than humans at?
- What are some things that humans are still much better at than AI systems?

Broader Impacts

- Earth Day is tomorrow. Is AI good for the environment? Bad?
- Is AI good for society? Bad?
- Is AI good for human creativity? is it bad?

Christian Perspectives

- Something that Christians should consider as people who *consume* AI-powered products?
- …As people who *use* AI in their organizations?
- …as people who *develop* AI?