Do we have to sacrifice accuracy in order to be fair?

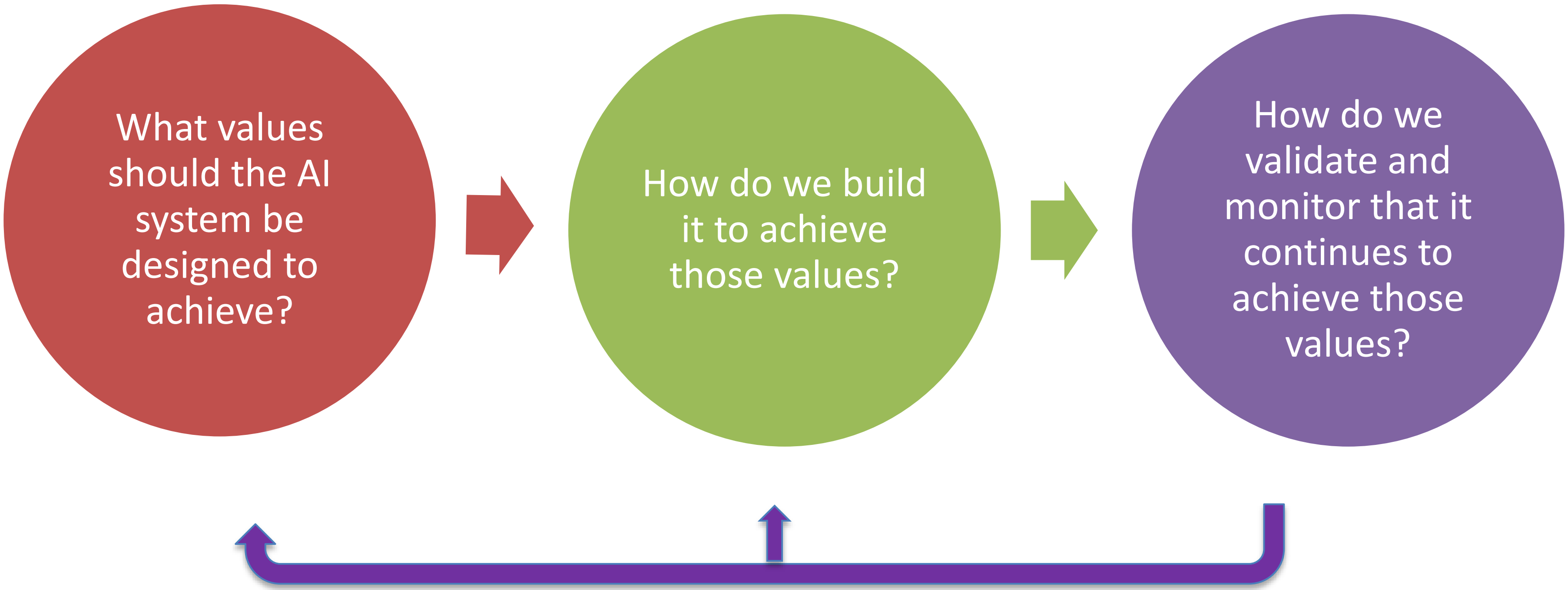
Rayid Ghani
(work with Kit Rodolfa and Hemank Lamba)

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

How should we design AI systems?



What values should we design for?

- Fairness
- Explainability
- Robustness
- Privacy
- Transparency
- Inclusiveness
- Accountability

The focus is not just on making the ML model fair but rather on **making the overall system and outcomes fair**

Define (desired fairness or equity goals) → Measure and Detect (bias) → Understand (root causes of bias) → Improve (fairness of AI systems) → Mitigate (the impact through adjusting interventions) → Monitor & Evaluate

Goals of this work

Empirically explore trade-offs between "accuracy" and "fairness" across a variety of policy problems

Compare Performance of Several Fairness-Enhancing Methods

Policy Settings

Increasing Educational Outcomes in Schools (10+ school districts across the US and with Department of Education, El Salvador

Matching interventions to students in need of extra support

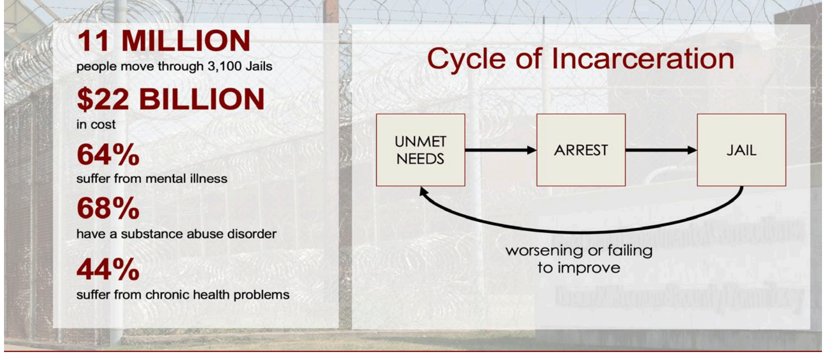
A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcomes. *Lakkaraju et al. KDD 2015*

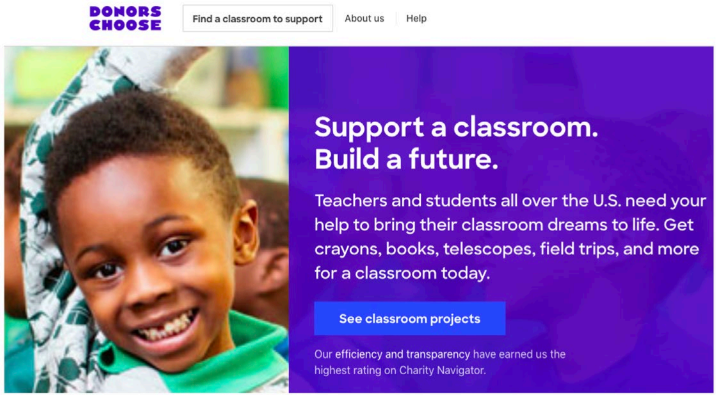
CITY OF SAN JOSE
CAPITAL OF SILICON VALLEY

Reducing Health and Safety Issues in Rental Housing

Reducing number of people going to Jail (Johnson County, KS)

Reducing Incarceration through Prioritized Interventions. Bauman et. Al. ACM COMPASS 2018

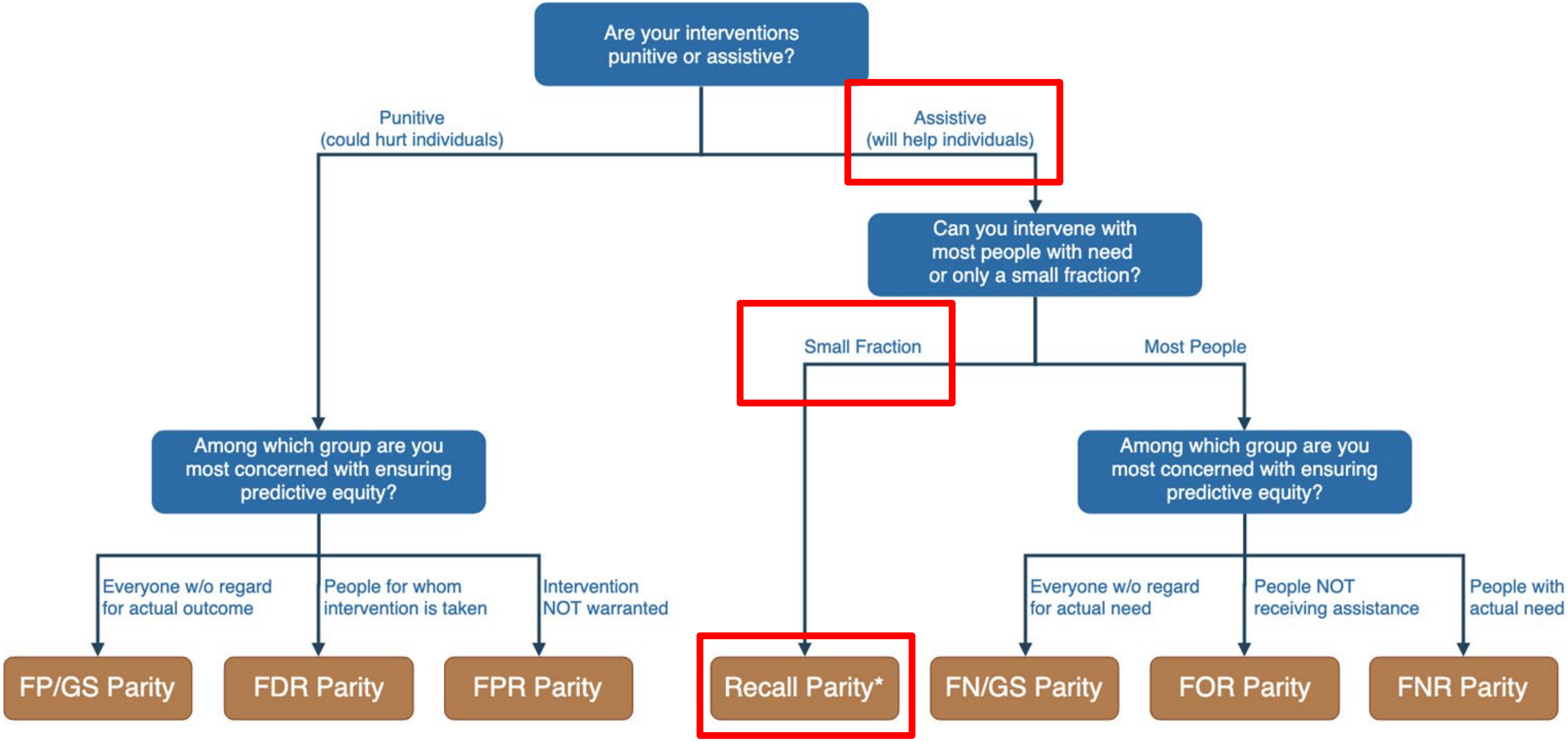


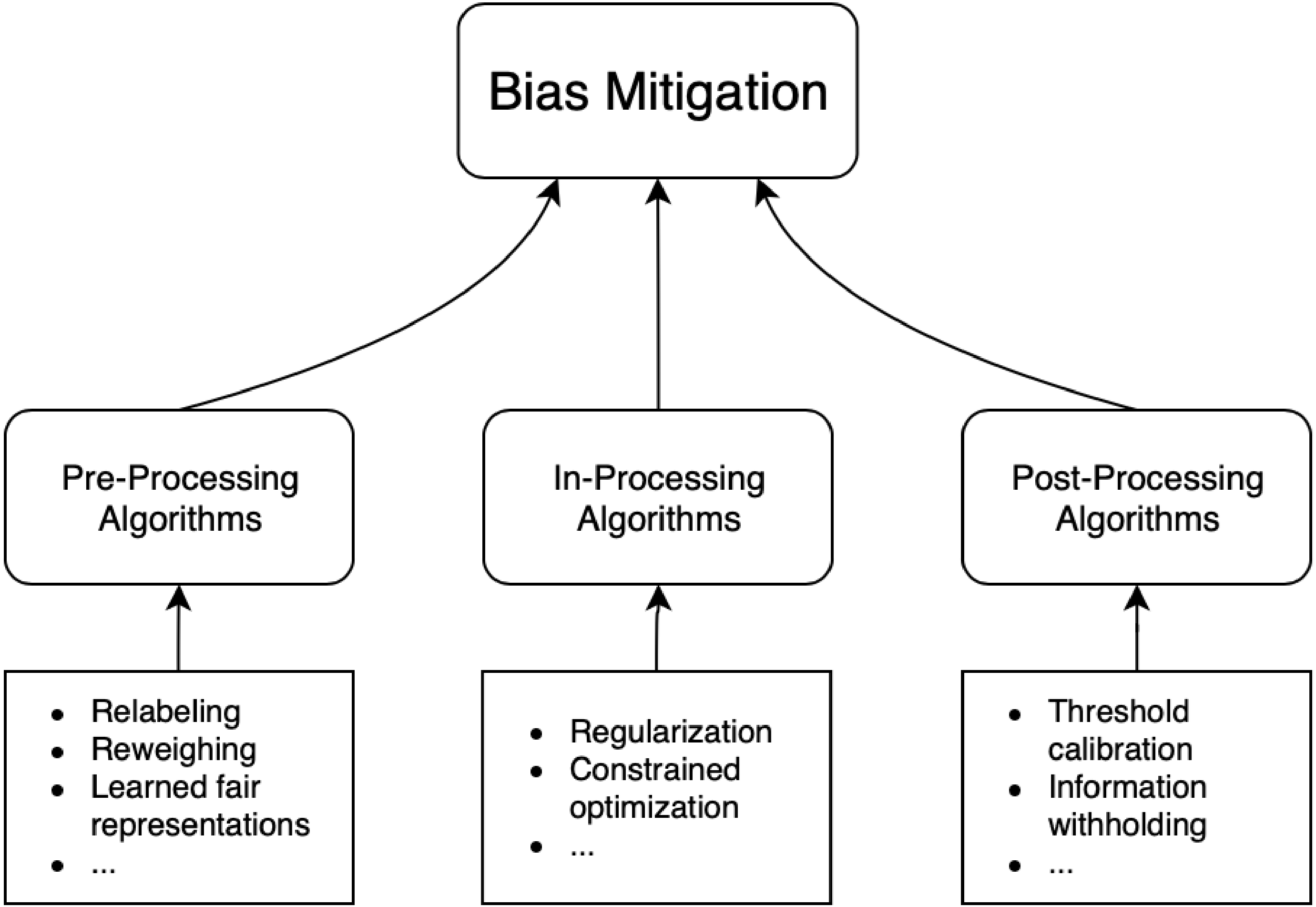


http://www.donorschoose.com/

Fairness metrics are specific to policy context and values

Define (desired fairness or equity goals) → Measure and Detect (bias) → Understand (root causes of bias) → Improve (fairness of AI systems) → Mitigate (the impact through adjusting interventions) → Monitor & Evaluate

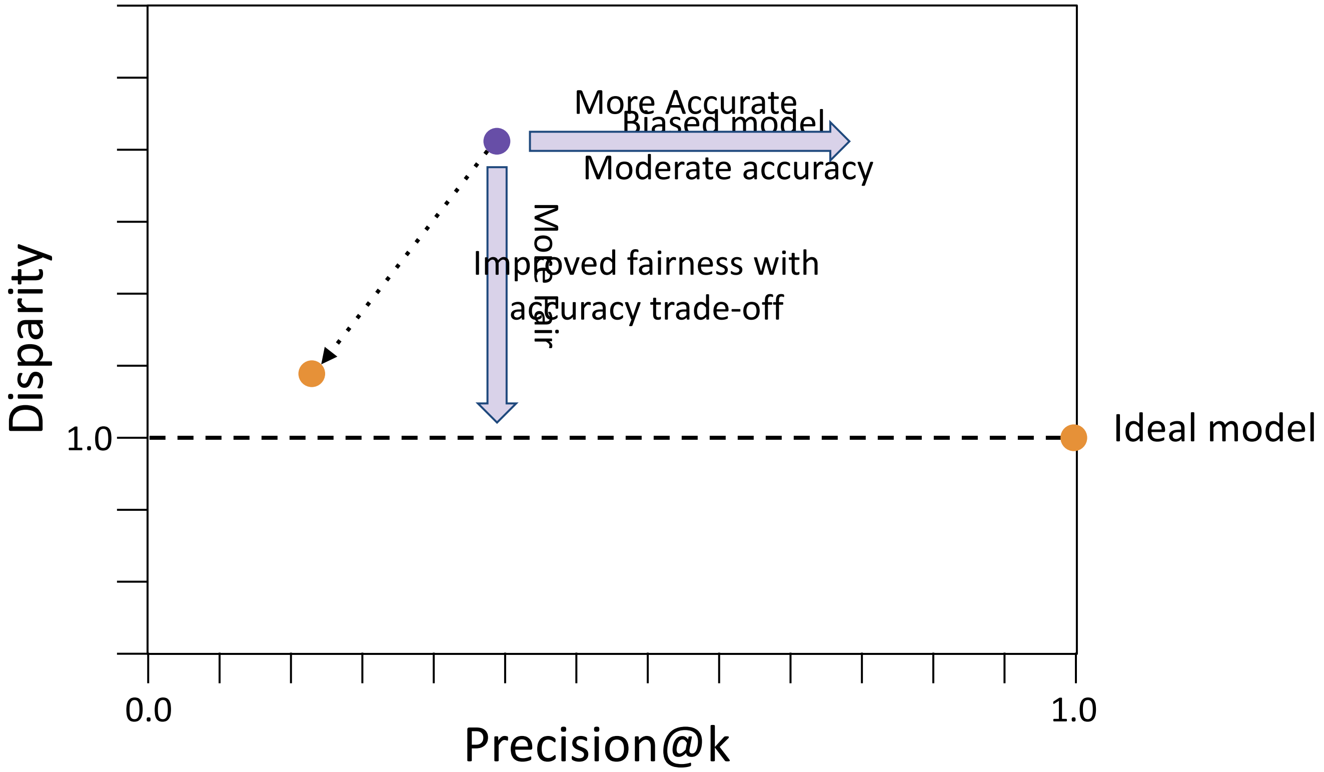

Are your interventions punitive or assistive?
Punitive (could hurt individuals)
Assistive (will help individuals)
Can you intervene with most people with need or only a small fraction?
Small Fraction
Most People
Among which group are you most concerned with ensuring predictive equity?
Everyone w/o regard for actual outcome
People for whom intervention is taken
Intervention NOT warranted
Among which group are you most concerned with ensuring predictive equity?
Everyone w/o regard for actual need
People NOT receiving assistance
People with actual need
FP/GS Parity
FDR Parity
FPR Parity
Recall Parity*
FN/GS Parity
FOR Parity
FNR Parity


Bias Mitigation
Pre-Processing Algorithms
In-Processing Algorithms
Post-Processing Algorithms
• Relabeling
• Reweighing
• Learned fair representations
• ...
• Regularization
• Constrained optimization
• ...
• Threshold calibration
• Information withholding
• ...

Methods Considered

- Pre-Processing
  - Remove sensitive attribute
  - Sampling (several approaches)
- In-Processing
  - Zafar's constrained optimization
- Post-Processing
  - Model selection
  - Post-hoc adjustments
  - Composite models (Dwork)

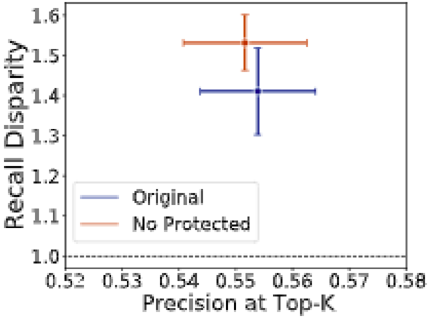
Evaluation Setup



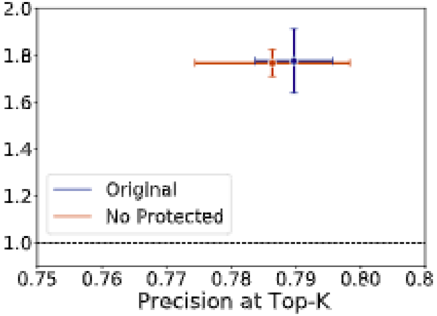
Methods Considered

- **Pre-Processing**
  - **Remove sensitive attribute**
  - Sampling (several approaches)
- In-Processing
  - Zafar's constrained optimization
- Post-Processing
  - Model selection
  - Post-hoc adjustments
  - Composite models (Dwork)

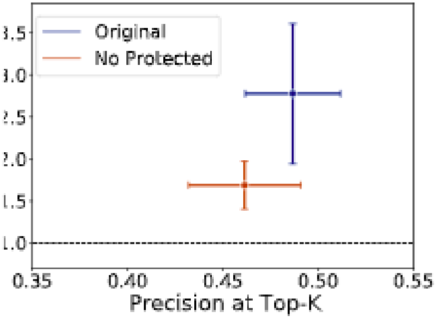
Removing Protected Attribute



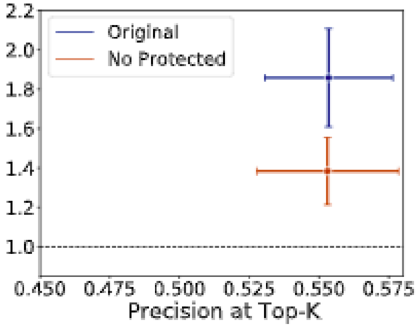
(a) Inmate Mental Health



(b) Housing Safety



(c) Student Outcomes

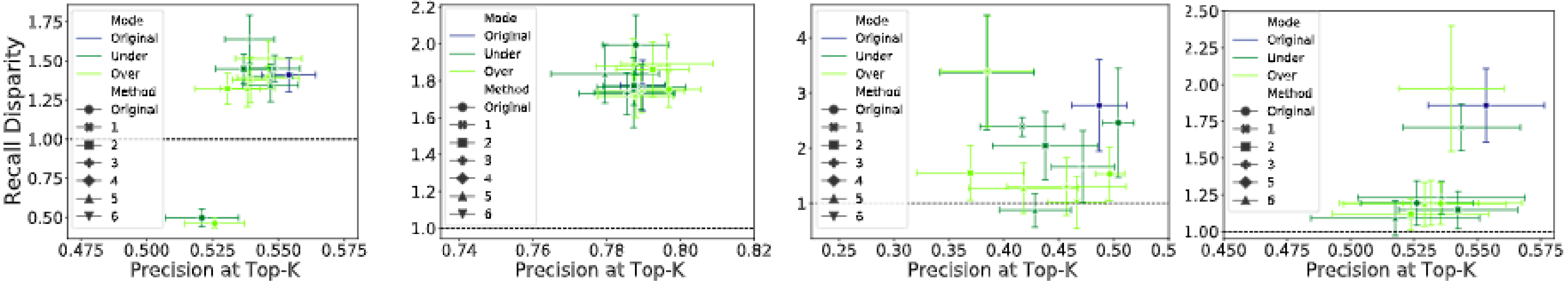


(d) Education Crowdfunding

Methods Considered

- **Pre-Processing**
  - Remove sensitive attribute
  - **Sampling (several approaches)**
- In-Processing
  - Zafar’s constrained optimization
- Post-Processing
  - Model selection
  - Post-hoc adjustments
  - Composite models (Dwork)

Sampling

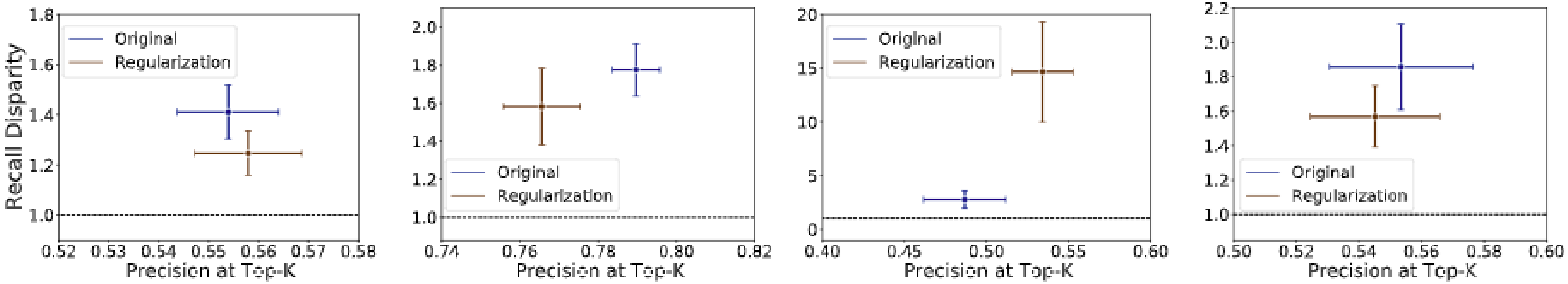


(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Methods Considered

- Pre-Processing
  - Remove sensitive attribute
  - Sampling (several approaches)
- **In-Processing**
  - **Zafar's constrained optimization**
- Post-Processing
  - Model selection
  - Post-hoc adjustments
  - Composite models (Dwork)

Constrained Optimization (Zafar)



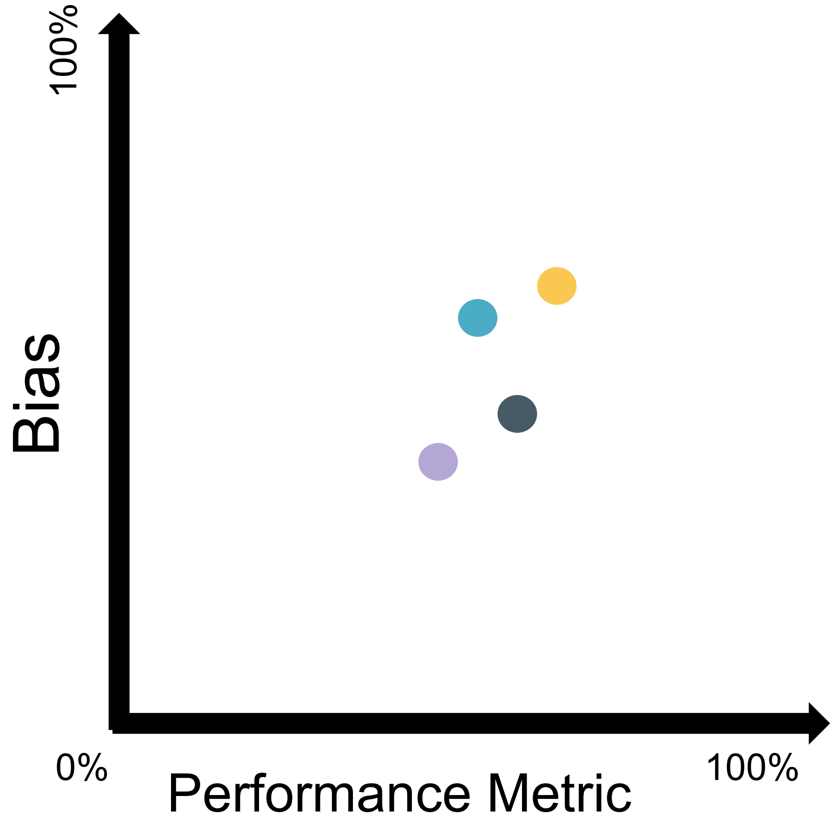
(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Methods Considered

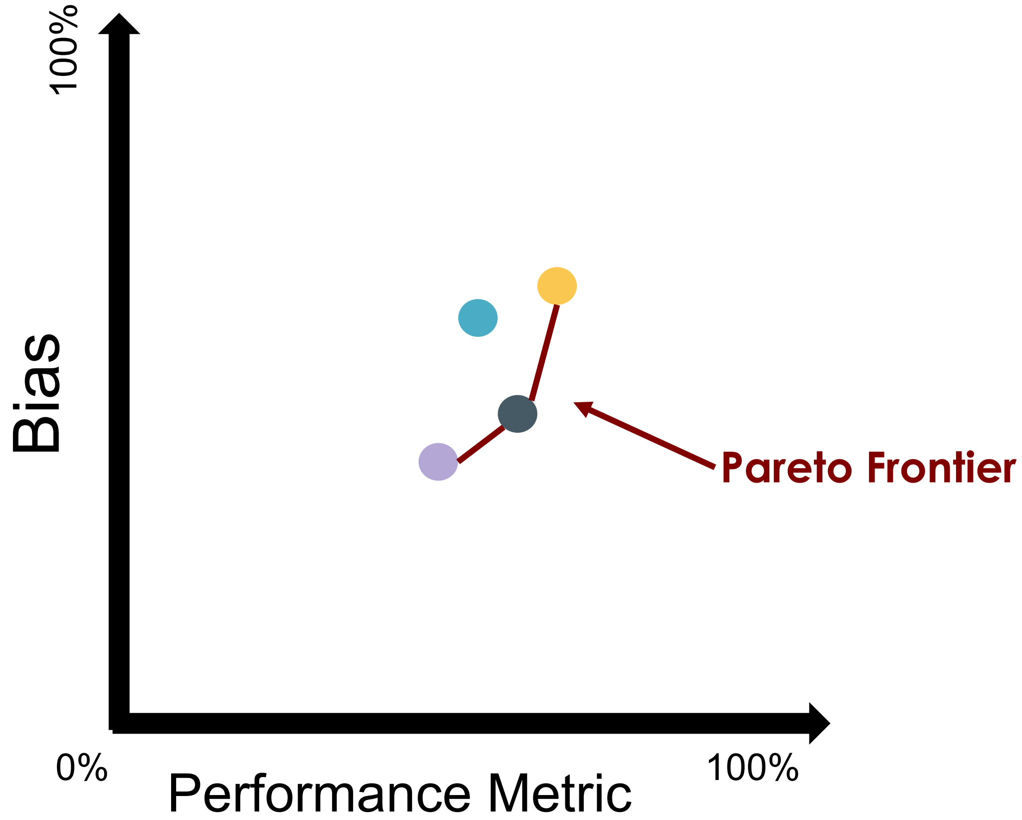
- Pre-Processing
  - Remove sensitive attribute
  - Sampling (several approaches)
- In-Processing
  - Zafar's constrained optimization
- **Post-Processing**
  - **Model selection**
  - Post-hoc adjustments
  - Composite models (Dwork)

Model Selection

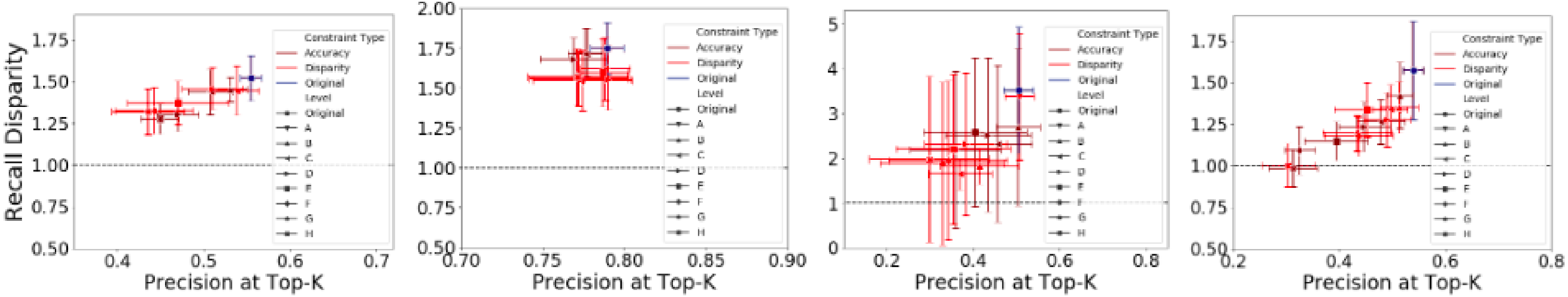
Model Selection



Model Selection



Model Selection

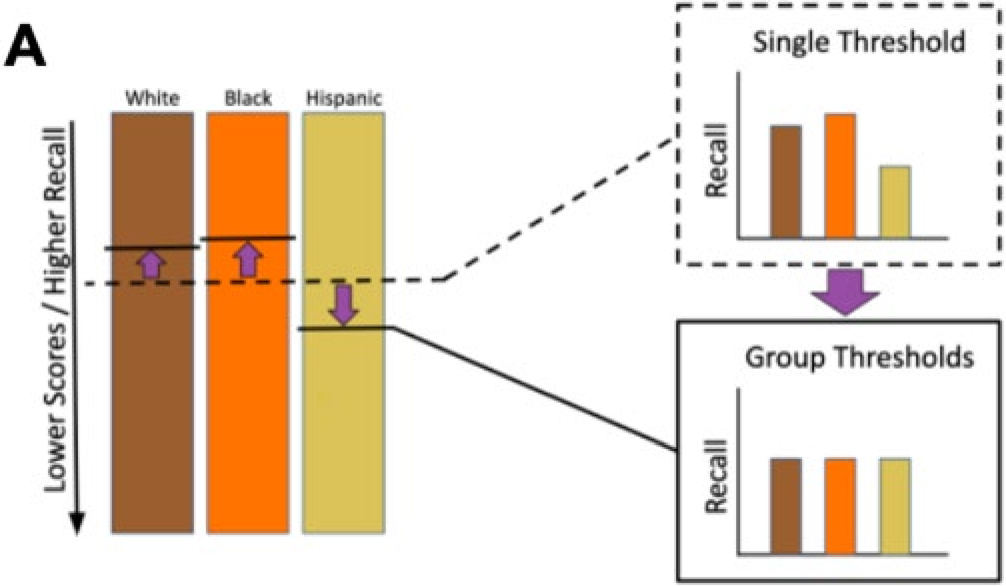


(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

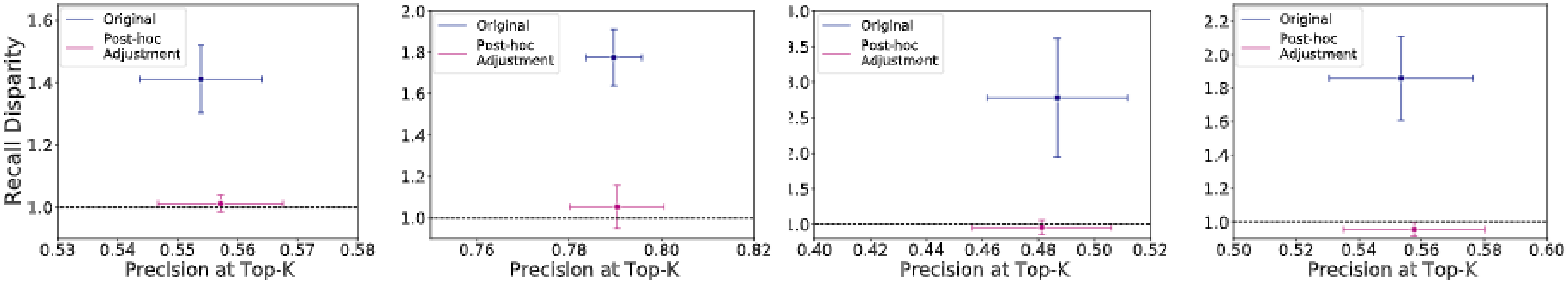
Methods Considered

- Pre-Processing
  - Remove sensitive attribute
  - Sampling (several approaches)
- In-Processing
  - Zafar's constrained optimization
- **Post-Processing**
  - Model selection
  - **Post-hoc adjustments**
  - Composite models (Dwork)

Post-Hoc Adjustments



Post-Hoc Adjustments



(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

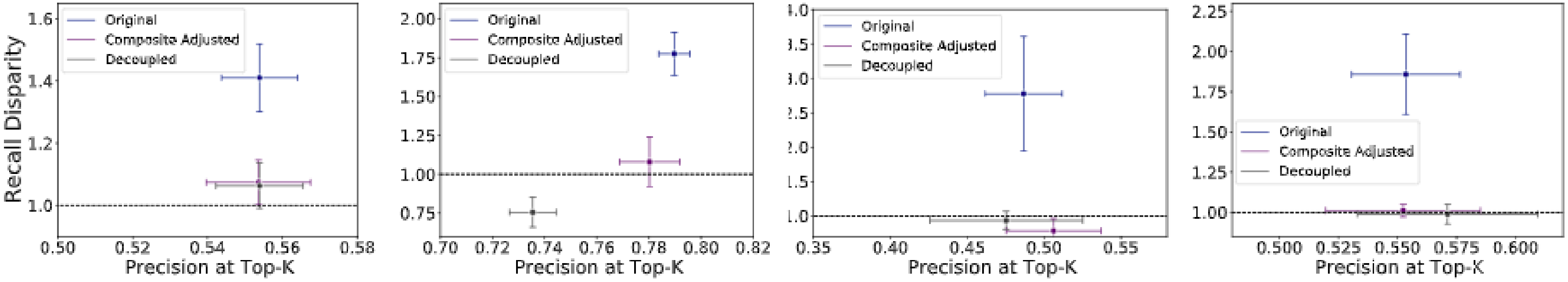
Methods Considered

- Pre-Processing
  - Remove sensitive attribute
  - Sampling (several approaches)

- In-Processing
  - Zafar’s constrained optimization

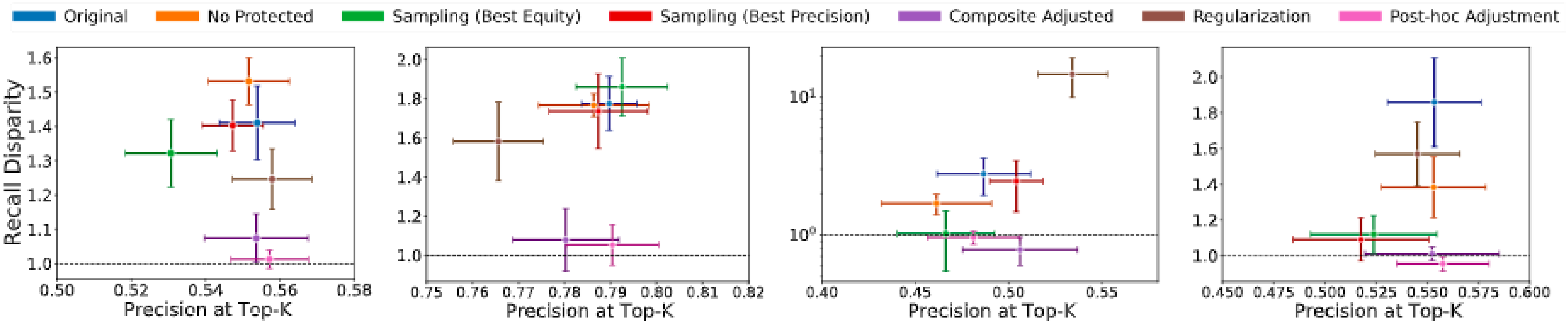
- **Post-Processing**
  - Model selection
  - Post-hoc adjustments
  - **Composite models (Dwork)**

Composite Models



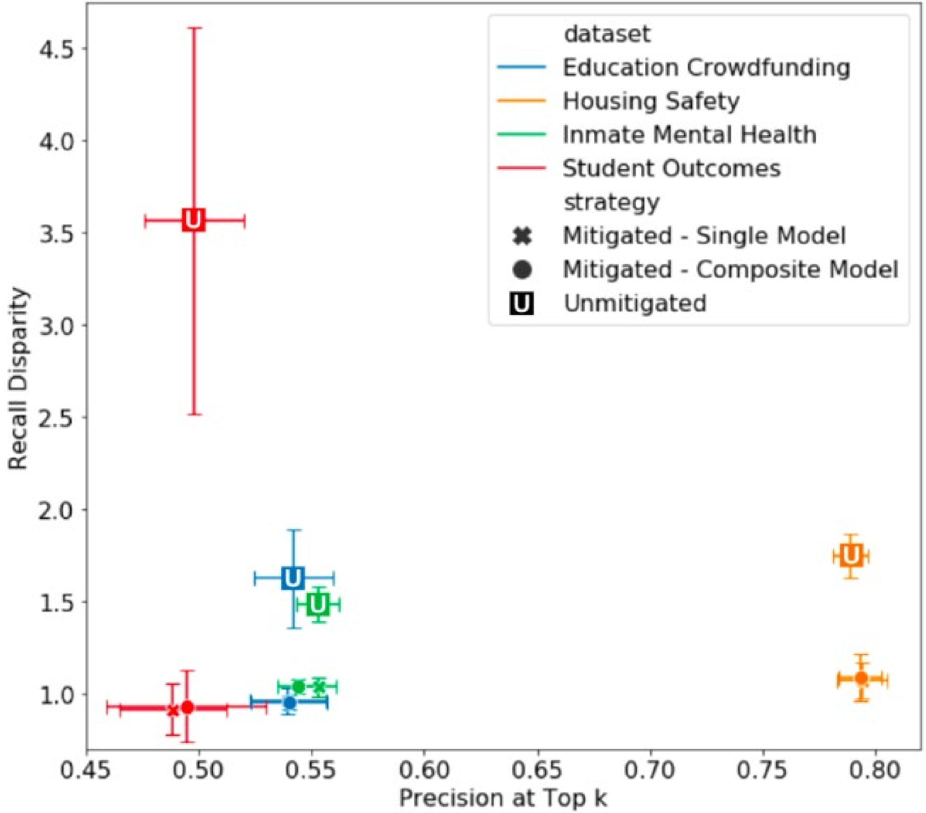
(a) Inmate Mental Health (b) Housing Safety (c) Student Outcomes (d) Education Crowdfunding

Putting It All Together...



(a) Inmate Mental Health (b) House Safety (c) Student Outcomes (d) Education Crowdfunding

Putting It All Together...



Summary

- High variability in performance across methods and contexts
- Post-hoc adjustment was only method that consistently improved fairness of model predictions
- In these contexts, the adjustments were able to improve fairness without cost to accuracy

We may not have to sacrifice accuracy in order to get fairness

but

we do have to deliberately and explicitly design our ML/AI systems for equity and fairness

Limitations / Future Work

- Focusing here on contexts where recall disparities are appropriate fairness metric
- Current in-processing methods not well-suited for top k settings, may be scope for new work here
- Understand how results generalize to settings where sensitive attribute isn’t known exactly but can be estimated

Useful Resources

- Data Science Project Scoping Guide
- Open Source Data Science Tools
  - Triage: ML Toolkit
  - Aequitas: Bias Audit Tool
  - Code for all projects: www.github.com/dssg
- Hands-on Fairness and Bias Tutorial with interactive Jupyter Notebooks
- Data Science for Social Good Fellowship

Rayid Ghani

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

rayid@cmu.edu

Appendix

Diagnosing the root causes of bias in AI/ML Systems

Bias and disparity (in outcomes) can come from any of these four components

World → Data → AI Pipeline → Actions → Disparity in Outcomes

Data:
Sample Bias
Measurement Bias
Label Bias

AI Pipeline:
System Developers
Complexity or flaws
Design Choices

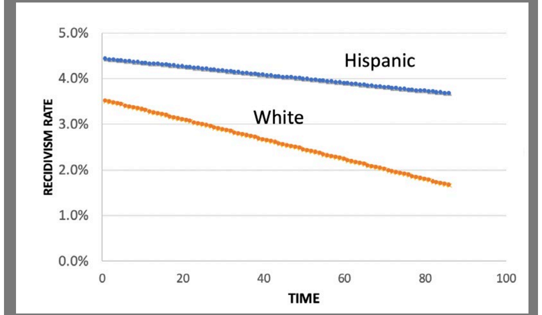
Some common practitioner perceptions

- There is always a tradeoff between fairness and accuracy
- I have to satisfy all measures of bias in order to be fair
- I have to eliminate all bias in order to use/deploy an ML system
- Not using race in my models makes by models not racist
- Using race in my models makes my models racist
- A fair ML model = Fair and equitable outcomes
- Bias comes from and can be fixed by “fixing” the data
- Bias reduction methods actually reduce bias

Policy Menu

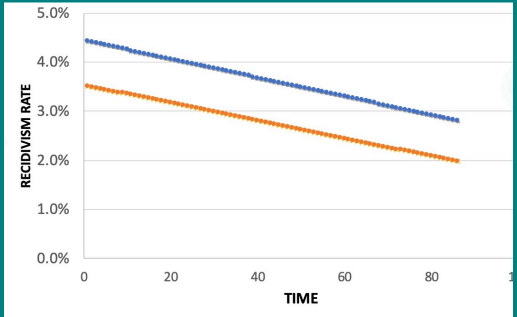
Designing for Efficiency

72.7% Efficient



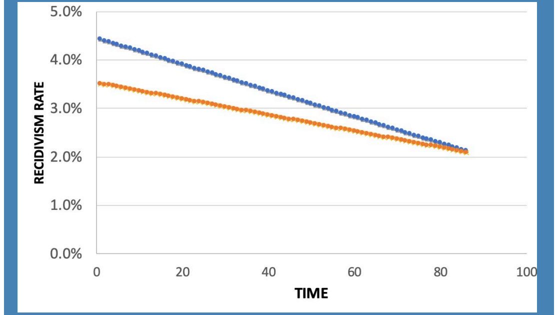
Equality

Additional Cost: 2%

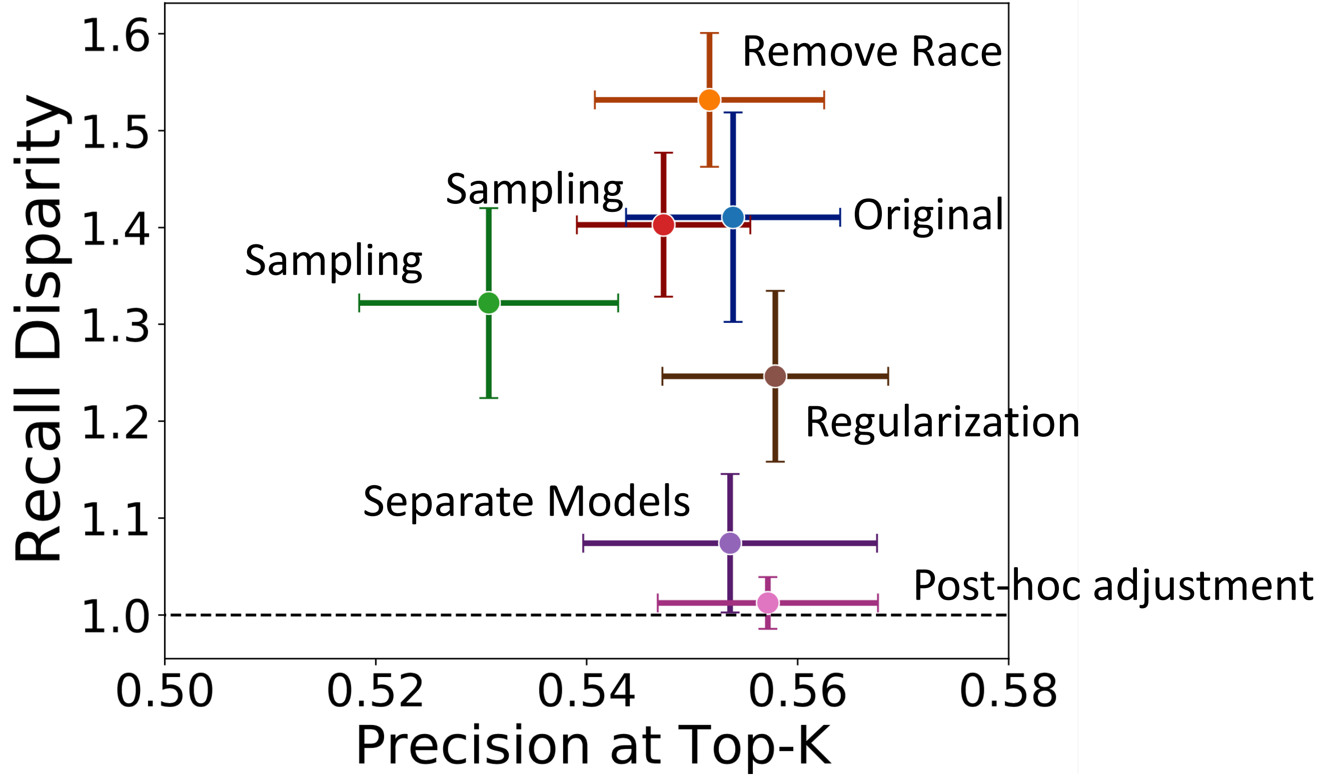


Equity

Additional Cost: 2%



But we do have to deliberately design for fairness



An Empirical Comparison of Bias Reduction Methods on Real-World Problems in High-Stakes Policy Settings. Lamba et al. KDD Explorations 2021 www.dssgfellowship.org

Data Science For Social Good

Summer Fellowship

Carnegie Mellon University

We're training data scientists to tackle problems that really matter.