Dealing with Bias and Fairness in AI/ML/Data Science Systems

Rayid Ghani, Kit T Rodolfa

Carnegie Mellon University

ML MACHINE LEARNING DEPARTMENT

HeinzCollege INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Pedro Saleiro

feedzai

AAAI 2021 Hands-on Tutorial

https://dssg.github.io/fairness_tutorial/

Before we start

Website: https://dssg.github.io/fairness_tutorial

Github Repo: http://github.com/dssg/fairness_tutorial

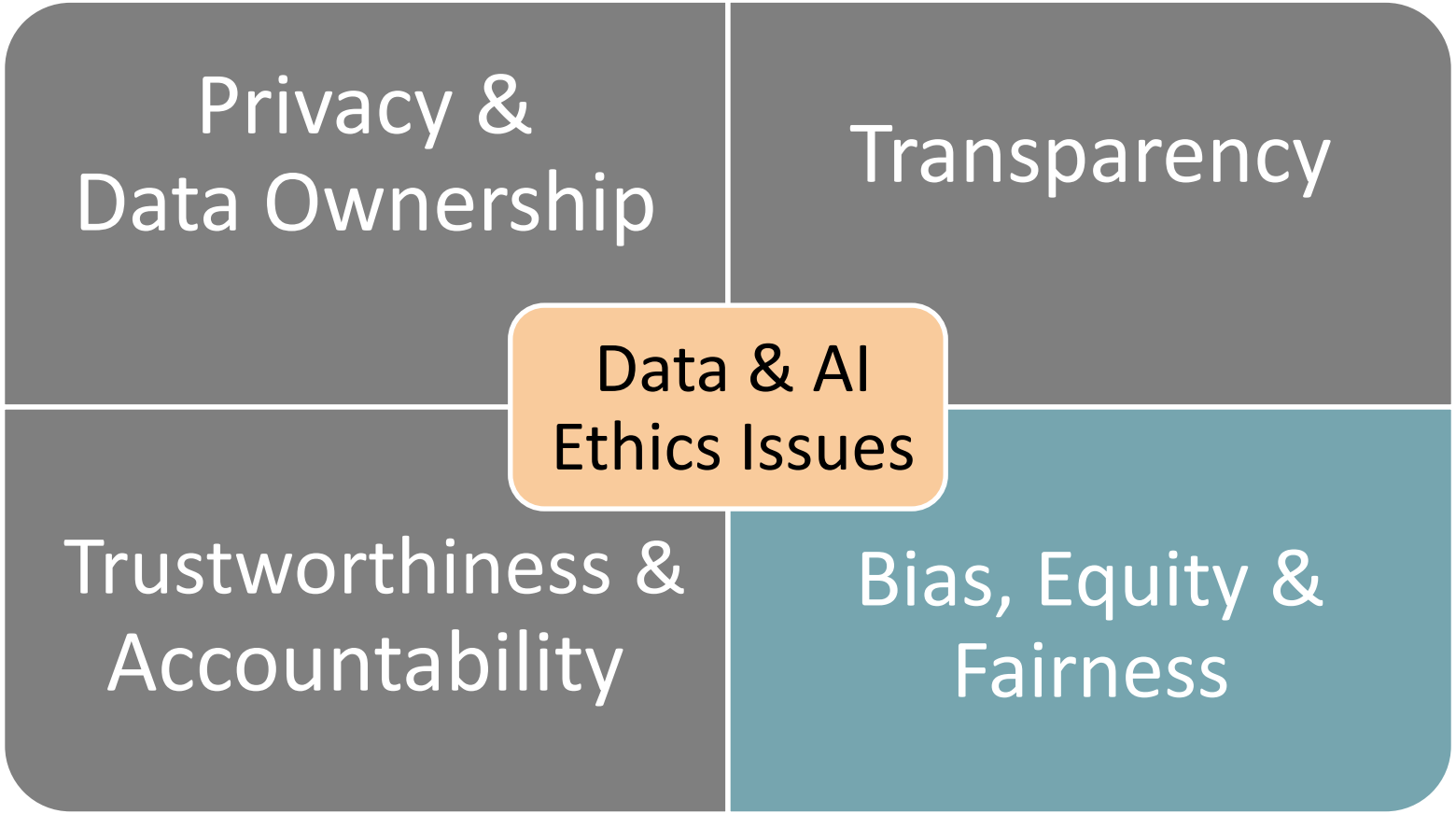
Interactive Colab (Python) Notebooks: https://dssg.github.io/fairness_tutorial/notebooks/

Agenda (8:30am-11:45am pacific)

8:30am	Introduction and Goals (All)	
8:40am	Fairness and Equity at a Systems and Outcomes Level (Rayid)	
8:55am	Sources of Bias: Case Studies (All)	Case Study Writeups
9:15am	From Societal Goals to ML Fairness Metrics (Kit)	Fairness Tree
9:35am	Possible Breakout Exercise: From Societal Goals to ML Fairness Metrics	Case Study Worksheets
10:00am	Break	
10:15am	Auditing Models for Bias (Pedro)	Aequitas
10:30am	Hands on: Bias Audit	Python Notebook
11:00am	Reducing Bias in ML Models (All)	
11:15am	Hands-on: Explore Bias Reduction Strategies (Hands-on)	Python Notebook
11:45am	Wrap-up: Things to Remember and Additional Resources (All)	

About us

More details on projects at http://dssgfellowship.org/projects


Privacy &
Data Ownership
Transparency
Data & AI
Ethics Issues
Trustworthiness &
Accountability
Bias, Equity &
Fairness

Our policies and the systems we build need to reflect our values

How do we develop AI/ML/DS systems that **help make decisions** leading to **fair** and **equitable outcomes**?

Objectives of this Tutorial: Learn how to...

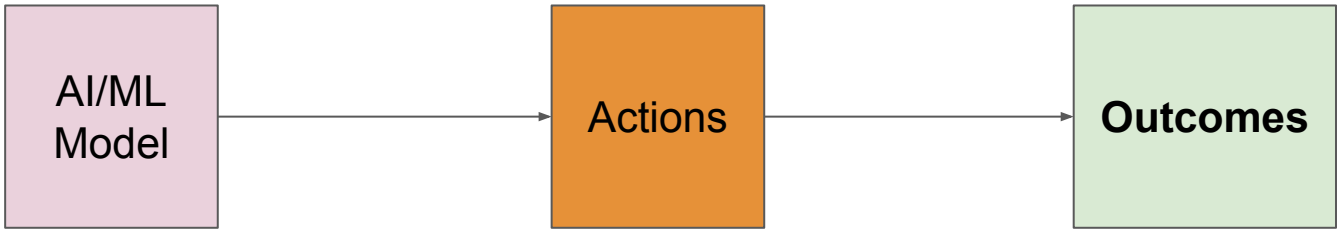
1. Think about overall fairness and equity when building Data Science/ML systems
2. Go from **social goals to fairness goals to ML fairness metrics**
3. **Audit bias and fairness** of a decision-making system
4. Explore **bias reduction strategies**

Part 1

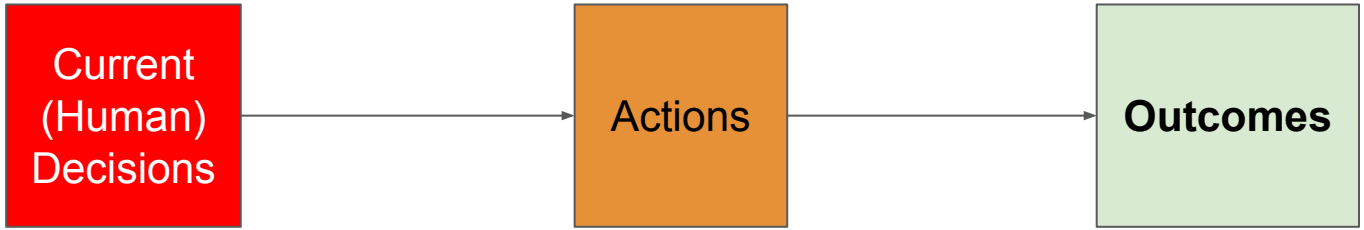
Think about overall fairness and equity when building Data Science/ML systems

The goal is not to make the ML model fair but to **make the overall system and outcomes fair**

The goal is not to make the ML model fair but to **make the overall system and outcomes fair**



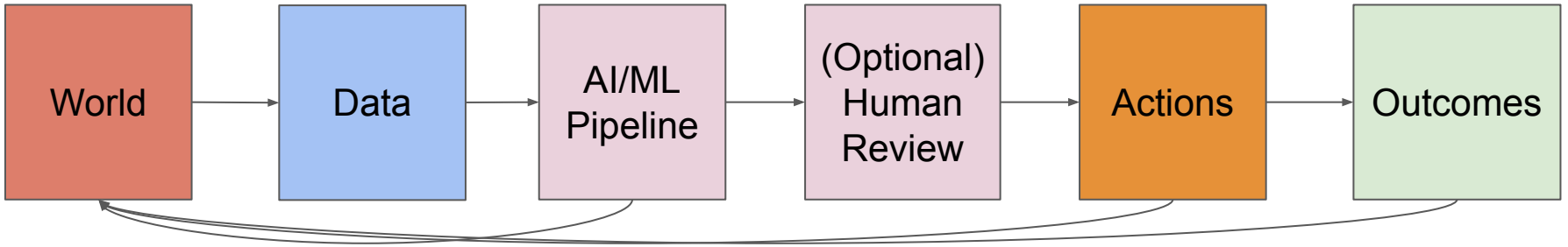
Compared to what?



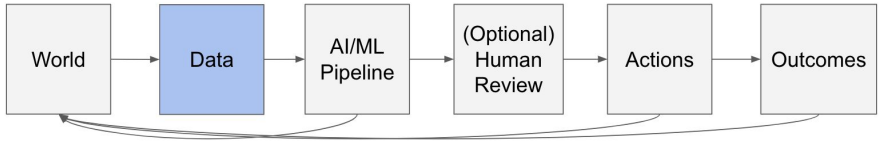
Does the new system need to be perfect or can it be better than the status quo and still worth implementing?

There are (unfortunately) many sources of bias

...it's not (just) the data



Bias in Data Sources



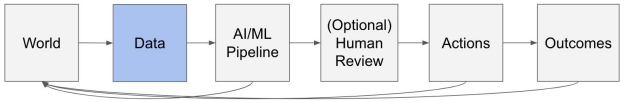
Choice of Data Sources

Sample Bias (Rows)

Measurement Bias (Cells)

Label Bias

Bias in Data Sources: Sample Bias

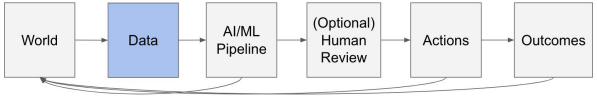


What is the relevant population for the project and how might some individuals be (incorrectly) excluded or included from the data available for modeling?

Are their underlying systemic biases involved in defining that population in general?

Data quality might not be uniform across groups.

Bias in Data Sources: Label Bias



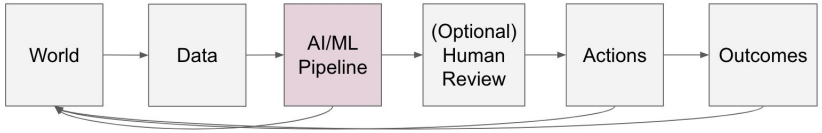
The way the target variable/label is defined and each data point is labeled might represent disparities between groups.

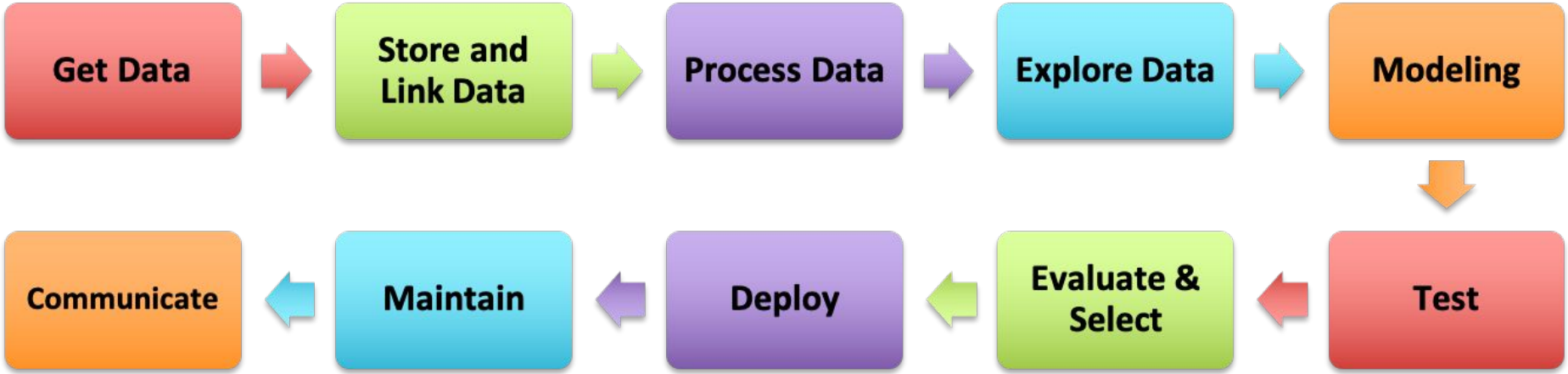
Differential measurement accuracy across groups (labeling quality).

A variable can be positively correlated with target variable within the majority group but negatively on other groups.

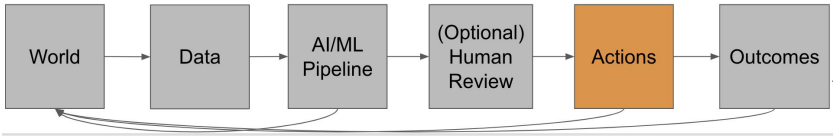
Police Internal Investigations for example

Even within the ML Pipeline, bias can be introduced in every step





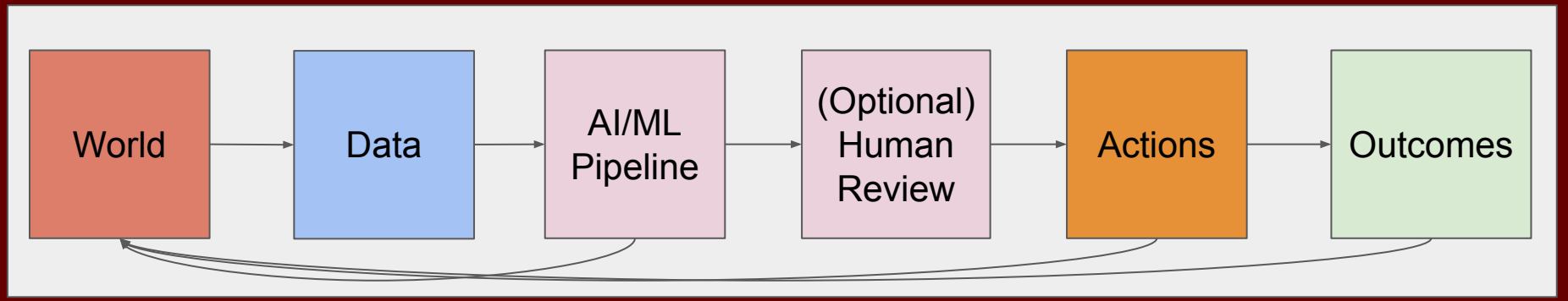
Action/Intervention Bias

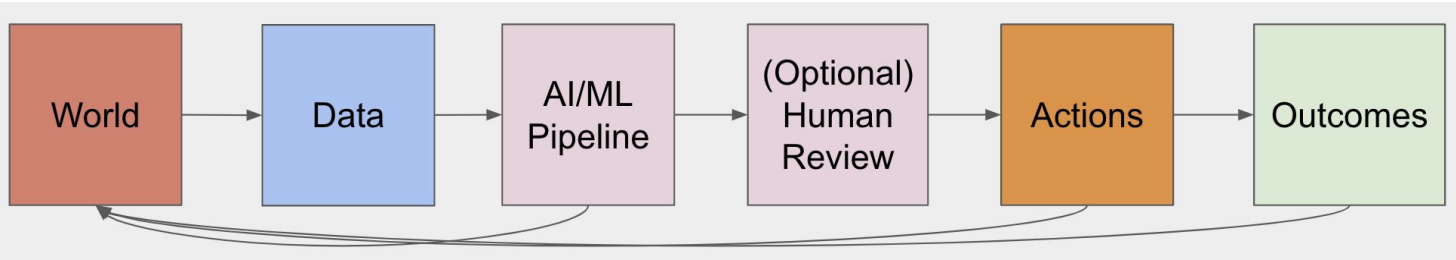


Heterogeneity in the effectiveness of an intervention across groups.

Discriminatory ‘overrides’ by the actors conducting the interventions.

Sources of Bias: Case Studies





Case Study 1: Student Support Programs

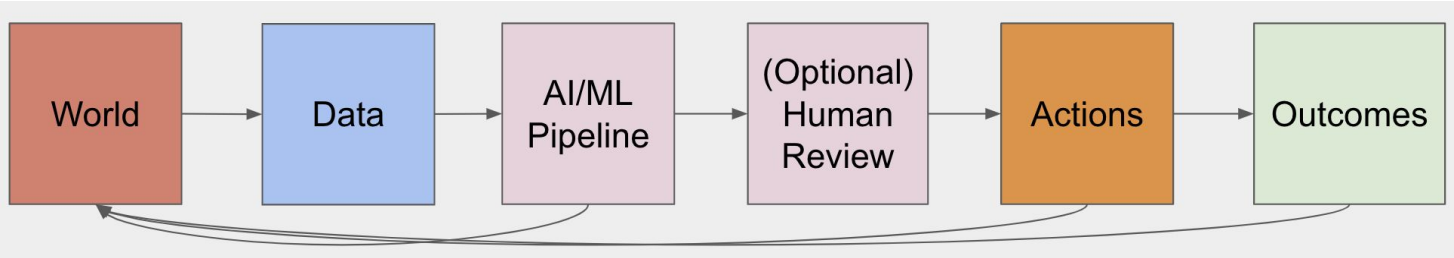
Goal: Improve graduation rates for students

Data: Student records from different school districts and states, national student clearinghouse data (which gives us information about college outcomes)

Analysis: Predict risk of not graduating on time

Actions: Assign after-school programs to most at-risk students

Constraints: Resources are available to target additional tutoring to 10% of students



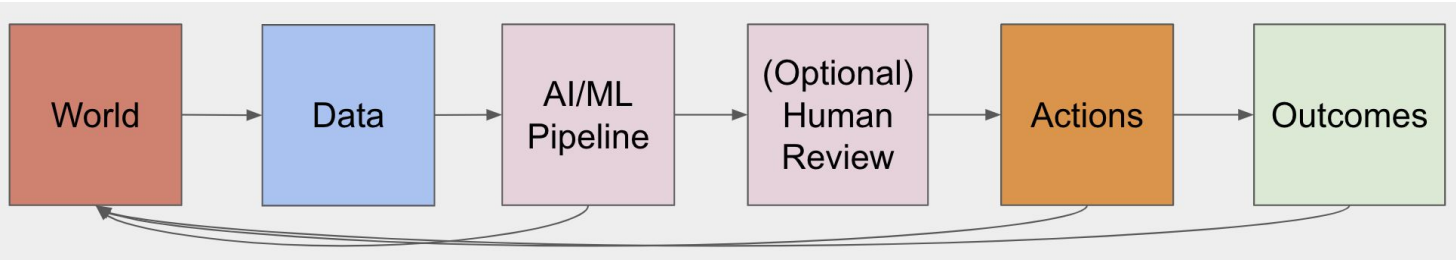
Case Study 2: Loans

Goal: Provide loans while balancing repayment rates for bank loans

Data: Historical loans and payments, credit reporting data, background checks

Analysis: Build model to predict risk of not repaying on time

Actions: Deny loan or increase interest rate/penalties



Case Study 3: Disaster Relief

Goal: Accurately assess damage and send appropriate relief resources

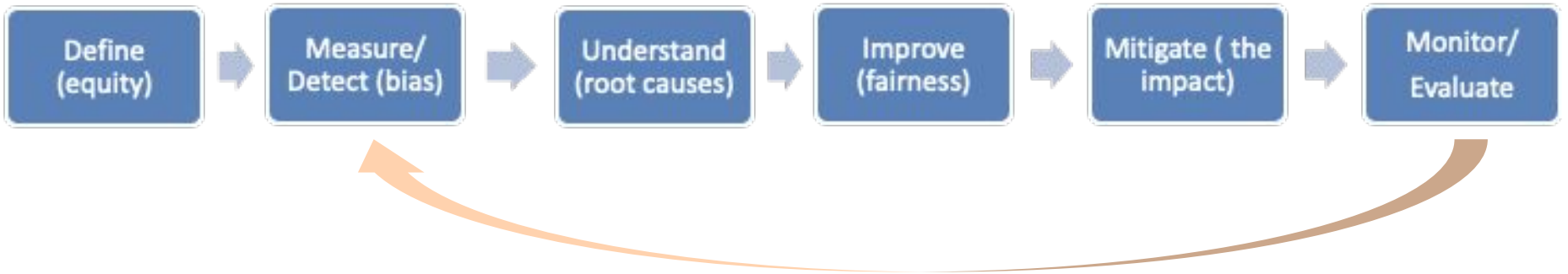
Data: Twitter posts, facebook posts geocoded with lat-long within disaster area and keywords/hashtags related to the storm

Analysis: Intensity and type of damage by neighborhood

Actions: Assessment and allocate relief effort (type and amount)

Constraints: Limited resources for relief efforts

How do we make the overall system and outcomes fair ?



Objectives of this Tutorial: Learn how to...

1. Think about overall fairness and equity when building Data Science/ML systems
2. **ML fairness metrics**
3. **Audit bias and fairness** of a decision-making system
4. Explore **bias reduction strategies**

Objectives of this Tutorial: Learn how to...

1. Think about overall fairness and equity when building Data Science/ML systems
2. **Go from social goals to fairness goals to ML fairness metrics**
3. **Audit bias and fairness** of a decision-making system
4. Explore **bias reduction strategies**

Many Bias Measures: How do we select what we care about?

- Statistical/Demographic Parity
- Impact Parity
- False Discovery Rate (1 - Precision) Parity
- False Omission Rate Parity
- False Positive Rate Parity
- False Negative Rate (1 - Recall) Parity
- ...

Many Bias Measures: How do we select what we care about?

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

Incompatibility Between Fairness Metrics

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{1}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

Incompatibility Between Fairness Metrics

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$

False Positive Rate
Among all actual 0's, fraction predicted to be 1

Prevalence
Fraction of actual 1's in population

False Discovery Rate
Among all predicted 1's, fraction that are actual 0's =(1 – precision)

False Negative Rate
Among all actual 1's, fraction predicted to be 0

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, *5*(2), 153-163.

Does that mean we cannot achieve fairness in ML models?

Punitive Action Example

A model being used to make bail determinations (keeping people in jail)

My fairness definition or yours?

Different people might consider it “fair” if:

It makes mistakes about denying bail to an equal number of white and black individuals.

Equal count of False Positives

$$P(\text{wrongly jailed, group } i) = C \quad \forall\, i$$

Different people might consider it “fair” if:

The chances that a given black or white person will be wrongly denied bail is equal, regardless of race.

Equal Group Size-Adjusted False Positives

$$P(\text{wrongly jailed} \mid \text{group } i) = C \quad \forall\, i$$

Different people might consider it “fair” if:

Among the jailed population, the probability of having been wrongly denied bail is independent of race.

Equal False Discovery Rate

$$P(\text{wrongly jailed} \mid \text{jailed, group } i) = C \quad \forall\, i$$

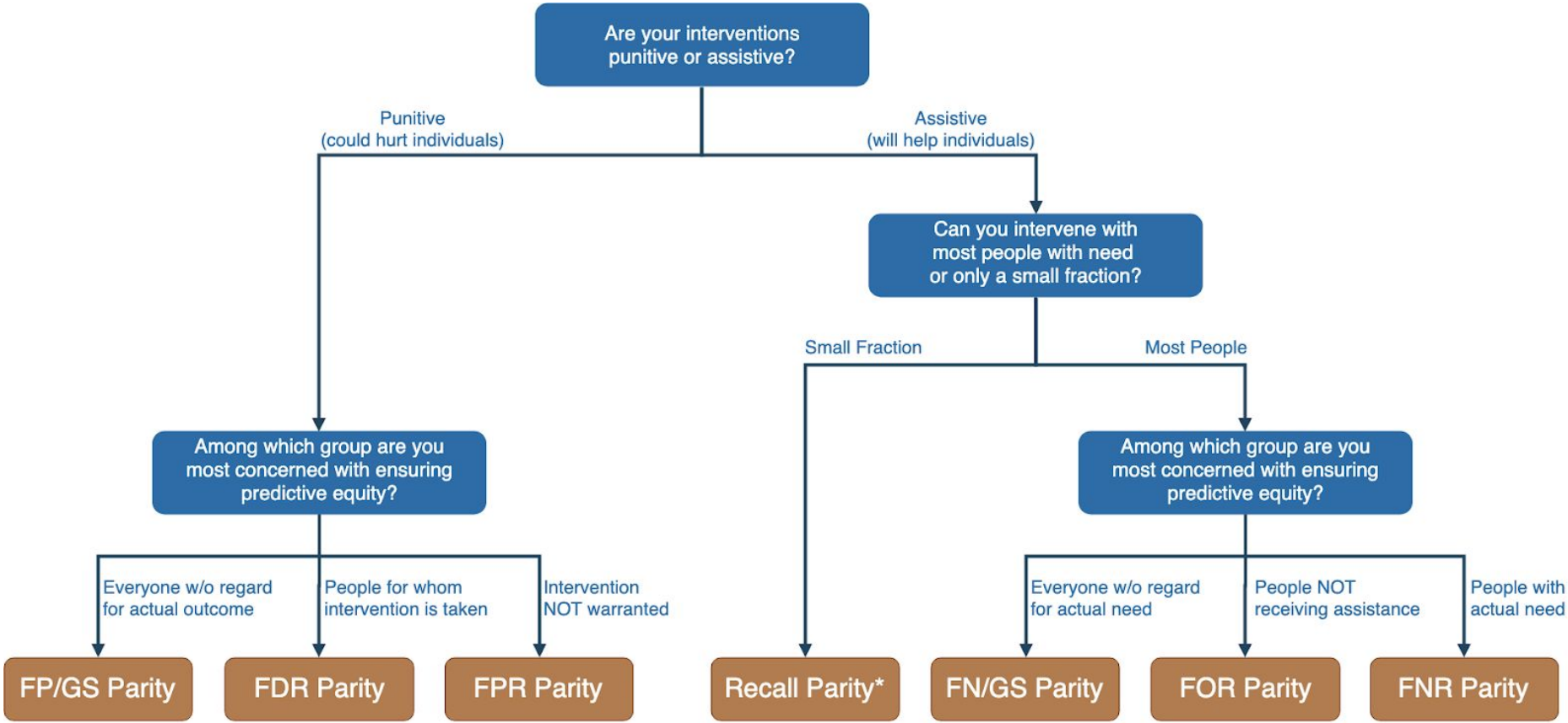
Different people might consider it “fair” if:

For people who should be released, the chances that a given black or white person will be denied bail is equal

Equal False Positive Rate

$$P(\text{wrongly jailed} \mid \text{innocent, group } i) = C \quad \forall\, i$$

Fairness Tree



Is the fairness tree “the answer”?

Is the fairness tree “the answer”?

No… but it’s intended as a starting point to help guide a conversation between ML experts, policy makers, and those affected by the decisions.

Ultimately, the choice of fairness metric(s) is highly dependent on context and stakeholder values.

Legal and Social Principles

Choices about fairness and equity are *implicit* in any decision process but being made *explicit* by the growing use of algorithmic decision support.

Legal and Social Principles

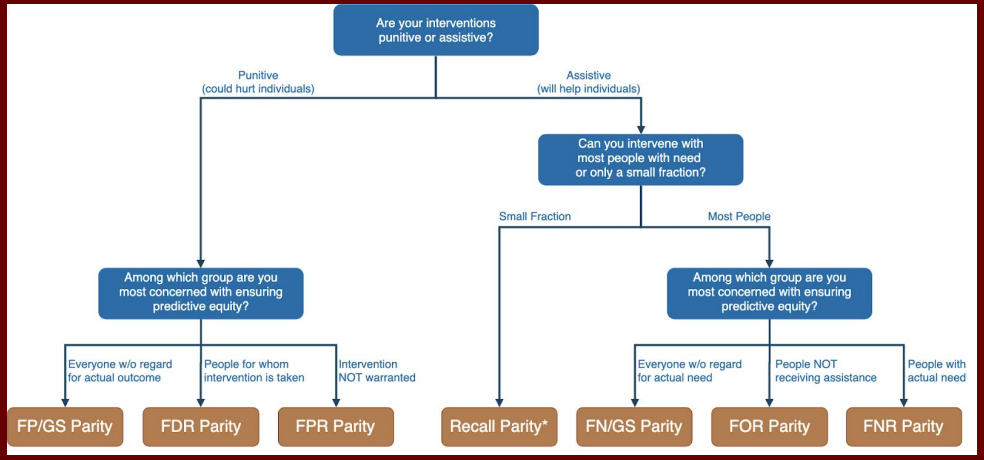
Choices about fairness and equity are *implicit* in any decision process but being made *explicit* by the growing use of algorithmic decision support.

What implicit choices about fairness are encoded when we say…

- Beyond a reasonable doubt? (e.g., criminal court)
- The preponderance of evidence? (e.g., civil court)
- Probable cause? (e.g., indictment by grand jury)
- Innocent until proven guilty?

BREAKOUT SESSION 1:

Fairness Metrics



Case Studies

Student Support

Loans

Disaster Relief

Objectives of this Tutorial: Learn how to...

1. Think about overall fairness and equity when building Data Science/ML systems
2. Go from social goals to fairness goals to ML fairness metrics
3. **Audit bias and fairness of a decision-making system**
4. Explore **bias reduction strategies**

Why Audit ML models for Bias

“If you don’t measure it, you can’t improve it.”

Creating awareness among stakeholders helps promoting bias and fairness as main KPI.

By measuring it, we can improve the system and also evaluate bias mitigation approaches.

http://www.datasciencepublicpolicy.org/aequitas/

What do you need to audit predictions of a model?

Predictions
(either binary predicted labels or scores along with a desired “top-k” list size)

Attributes that define protected groups
(e.g. race, sex, age)

Labels*
(if interested in disparate errors)

Bias Audit flow

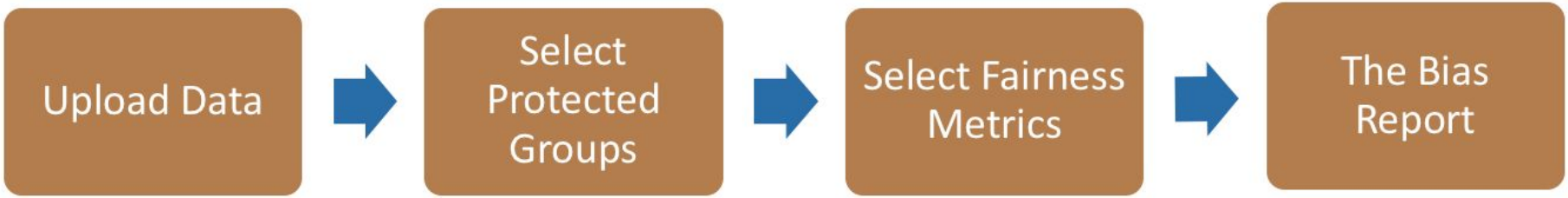
Predicted Scores and Labels

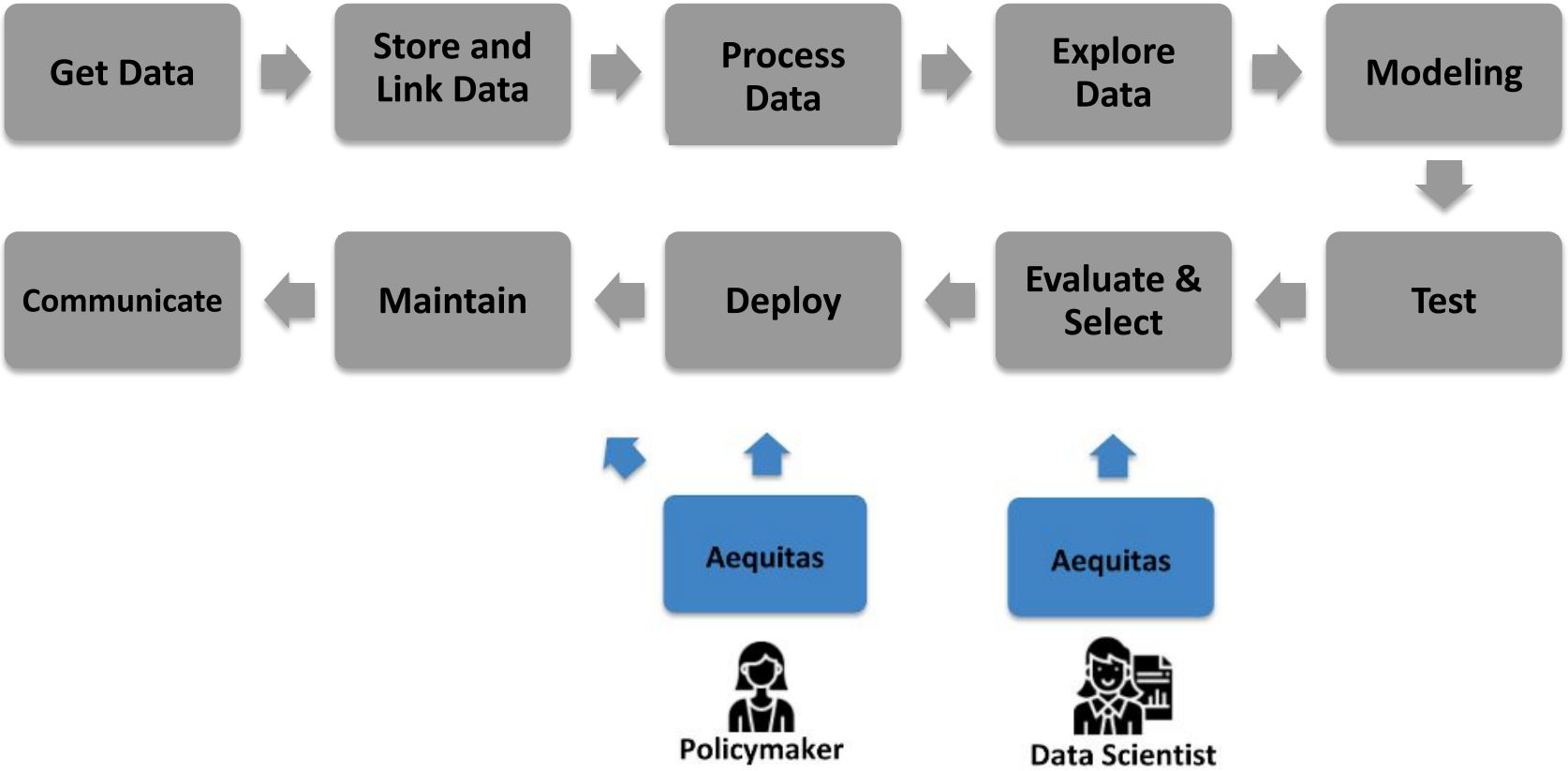
Race

Sex

Age

Equal True Positive Rate




Get Data
Store and Link Data
Process Data
Explore Data
Modeling
Test
Evaluate & Select
Deploy
Maintain
Communicate
Aequitas
Policymaker
Aequitas
Data Scientist

Aequitas - Bias and Fairness Audit Toolkit

How can you use Aequitas?

Web Audit Tool

Try our Audit Tool to generate a Bias Report

1. Upload Data (or use pre-loaded sample data)
2. Configure (bias metrics of interest and reference groups)
3. Generate the Bias Report

Try it out! ›

Python Library

Use our python code library to generate bias and fairness metrics on your data and predictions.

Python Code ›

Command Line Tool

Use our command line tool to generate a report using your own data and predictions.

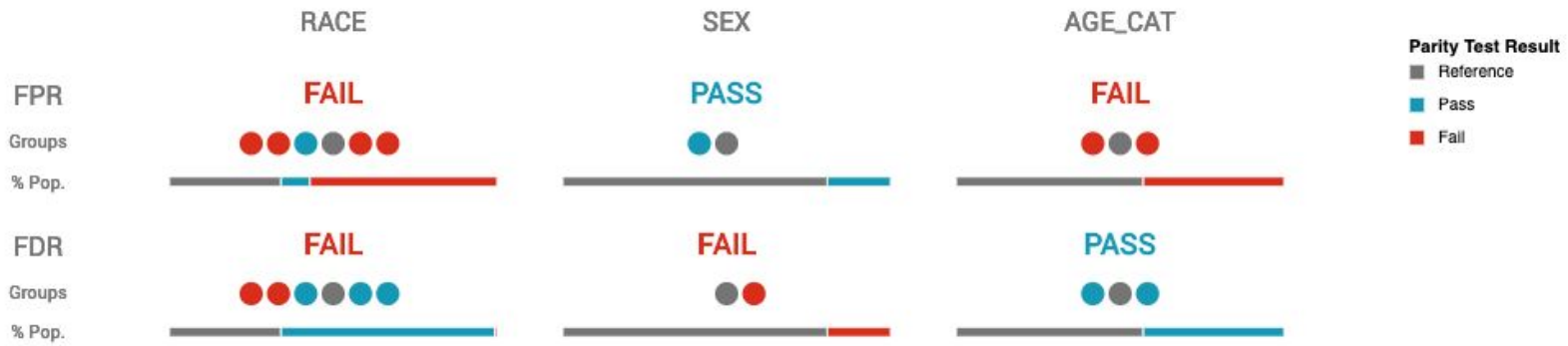
Audit COMPAS using Aequitas

Get acquainted with the tool by running the COMPAS demo notebook.
https://github.com/dssg/aequitas

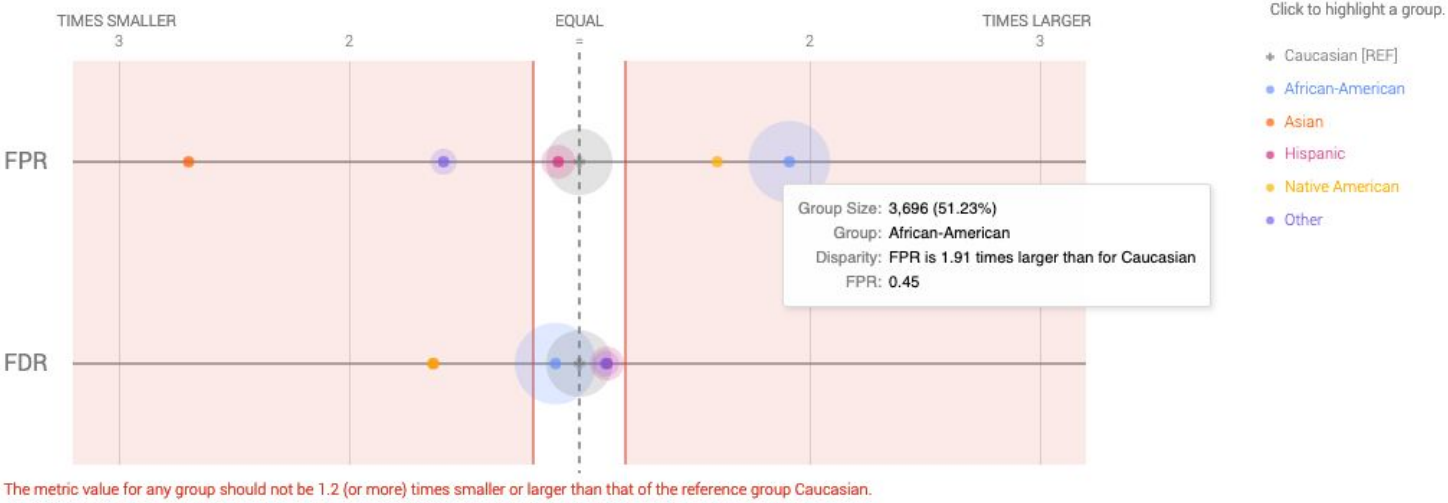
“There’s software used across the country to predict future criminals. And it’s biased against blacks.”

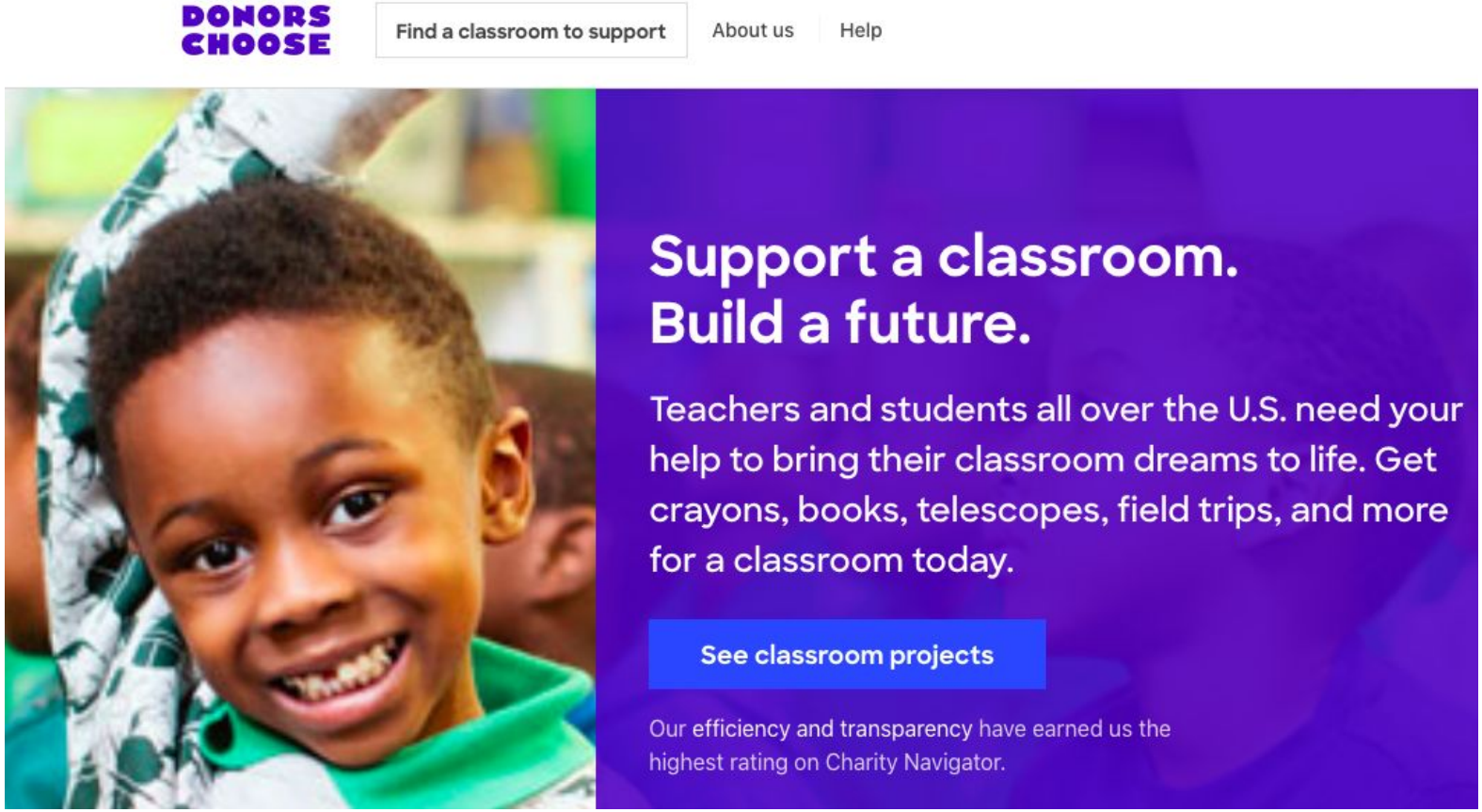
ProPublica, 2016

Aequitas Summary for COMPAS



Disparity Chart





http://www.donorschoose.com/

Shoo Germs!

"Help me give my students the necessary supplies to stay safe in the classroom. To reduce the spread of germs, sanitizing products and storage containers need to be used to keep students healthy."

Mrs. Ng
PS 336 @ 474 • Ozone Park, NY

8 DONORS SO FAR

$68 STILL NEEDED

First Grade Here We Come!

"Help me give my students literature to help facilitate conversations about race, privilege, diversity, and equality as well as the option of flexible lap desk seating amid Covid-19 social distance protocols."

Ms. Harvey
Thiells Elementary School • Thiells, NY

8 DONORS SO FAR

$55 STILL NEEDED

Math and Reading Tools for Success!

"Help me give my students tactile math and reading tools, personal dry-erase boards, books, and other key learning tools so that they can learn successfully in school and at home this year."

Ms. Rodosky
Great Lakes Academy Charter School • Chicago, IL

6 DONORS SO FAR

$40 STILL NEEDED

- Crowdfunding platform seeking to fill funding gaps faced by disadvantaged schools
- Has facilitated $970 million in donations to projects affecting 40 million students in the United States
- About one-third of projects fail to meet their funding goal

Case Study Setting

Goal: Increase the fraction of projects that get funded

Data: Project details (resources, description, ask, etc); teacher, class, and school information; project donations

Analysis: Predict the risk projects will fail to achieve their funding ask within 4 months, prioritizing 1,000 projects for intervention every 2 months

Actions: Provide identified projects with expert review and tailored suggestions for improving their prospects

Data is publicly available on Kaggle from the KDD Cup 2014

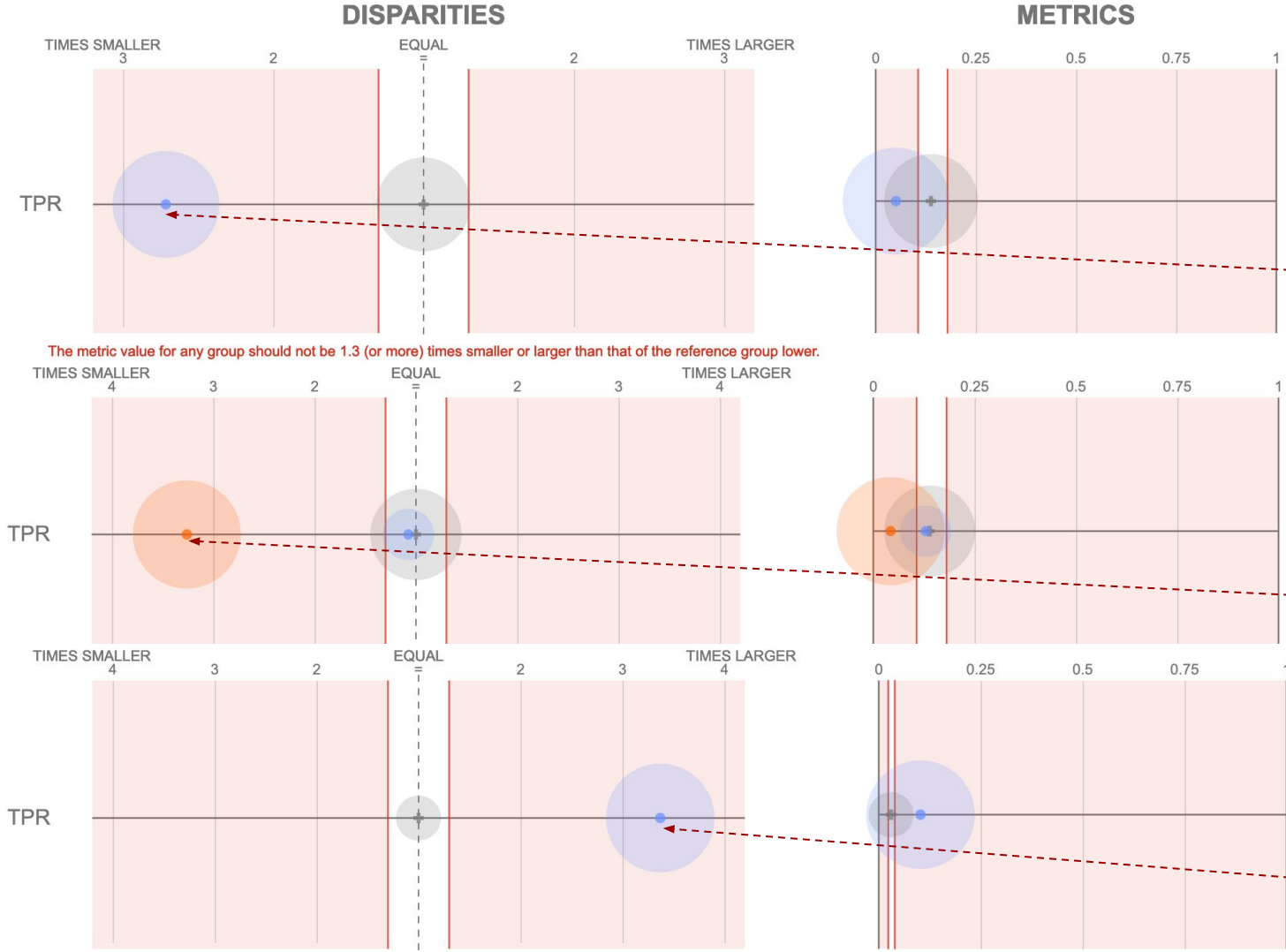
HANDS-ON:

Auditing a Model for Bias

(click on link to load notebook and run setup)

Auditing a Model for Bias

What did we find?



Poverty Level of the School

- lower [REF]
- highest

TPR for high poverty schools is 2.6x lower than low poverty schools

Location Type of the School

- suburban_rural [REF]
- unknown
- urban

TPR for urban schools is 3.3x lower than suburban/rural schools

Sex of the Teacher

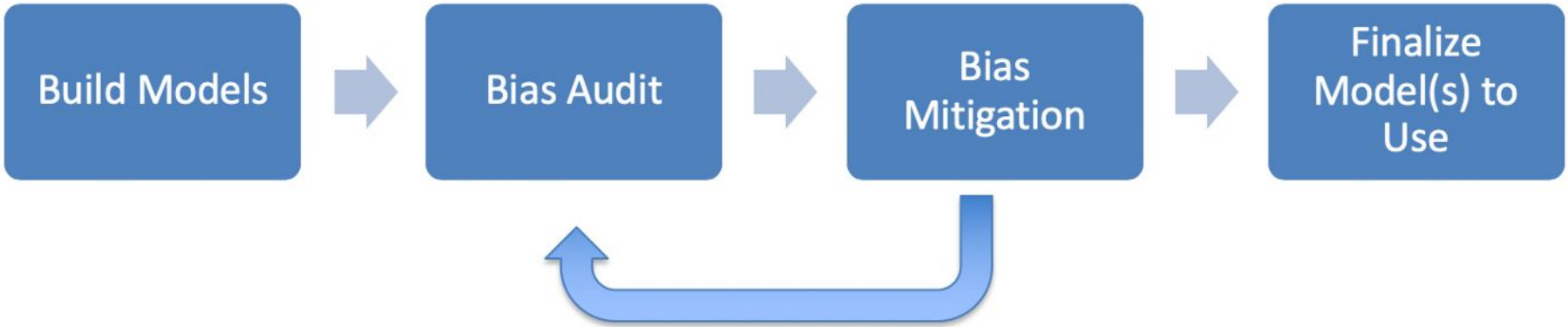
- male [REF]
- female

TPR for male teacher projects is 3.5x lower than female teacher projects

Objectives of this Tutorial: Learn how to...

1. Think about overall fairness and equity when building Data Science/ML systems
2. Go from social goals to fairness goals to ML fairness metrics
3. Audit bias and fairness of a decision-making system
4. **Explore bias reduction strategies**

Workflow


Build Models
Bias Audit
Bias Mitigation
Finalize Model(s) to Use

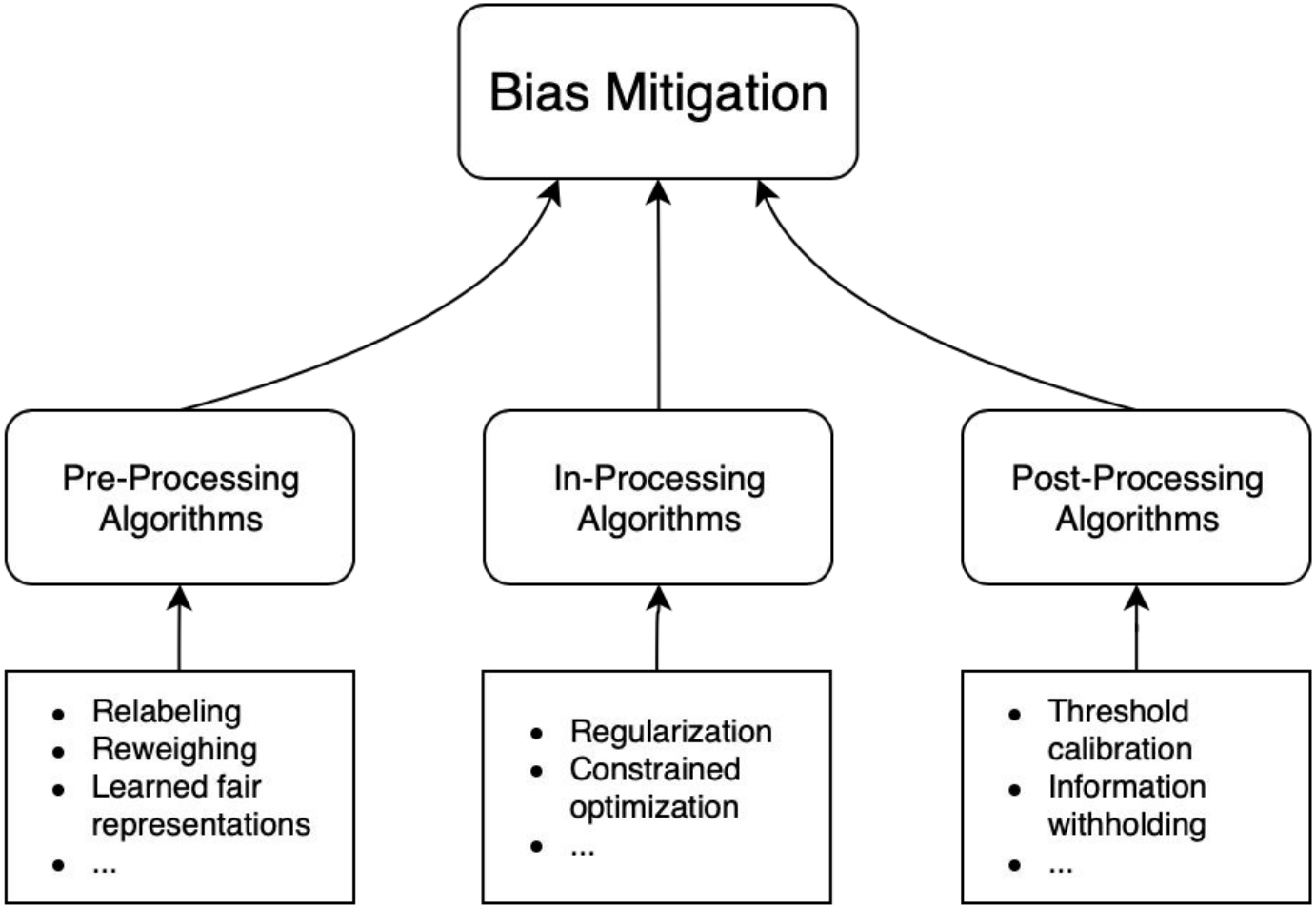
How can we reduce bias in ML models?

- Fix the world
- Fix the input data
  - ~~Remove sensitive attributes~~
  - Resample and/or reweight protected groups
- Choose fair models during model selection
- Optimize for fairness in model training
- Post-hoc adjustments to ‘de-bias’ model scores

If the model doesn’t know the race/sex/... how can it discriminate?

There is no fairness through unawareness.
A “race/sex/... blind” model can discriminate.

"There's no gender bias in our process for extending credit," Solomon said in an interview with **Bloomberg TV** late Thursday. "We don't ask whether — when someone applies — if they're a man or a woman. We don't ask if they're married."


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

How can we reduce bias in ML models?

- Fix the world
- Fix the input data
  - ~~Remove sensitive attributes~~
  - **Resample and/or reweight protected groups**
- Choose fair models during model selection
- Optimize for fairness in model training
- Post-hoc adjustments to de-bias model scores

Relabeling / Massaging

Algorithm 1 *Classification with No Discrimination (CND)*

Input $(D, s, SA, +)$

Output Classifier CND learnt on D without discrimination

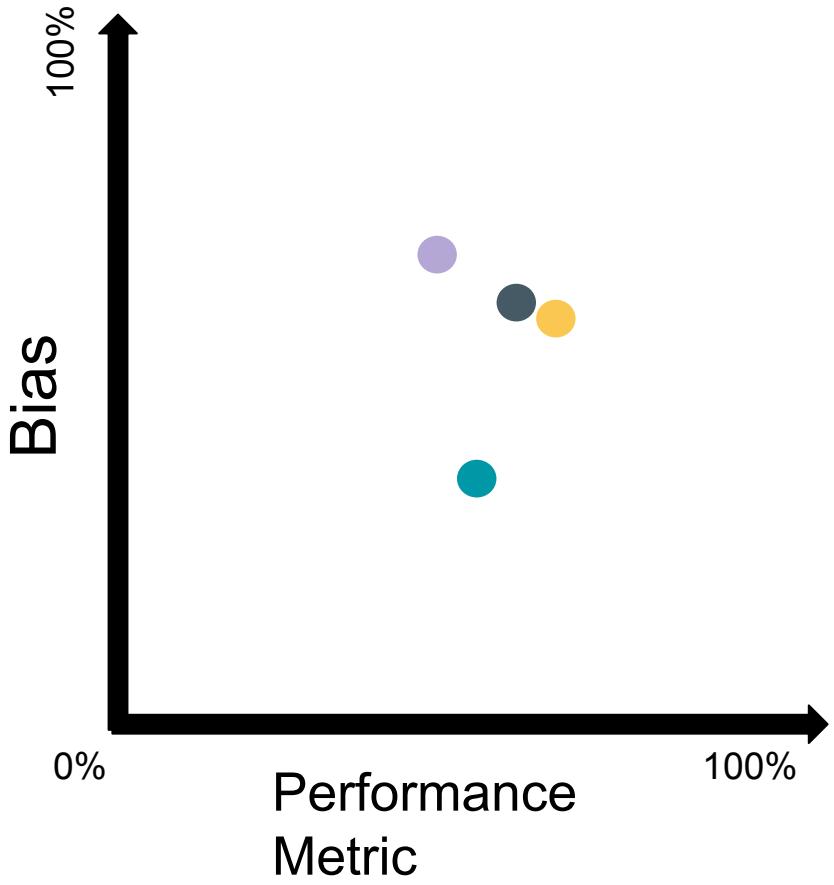
1: $(pr, dem) := Rank(D, SA, s, +)$
2: $existDisc := Disc(D, SA, s, +)$
3: Calculate M, the number of necessary modifications based on $existDisc$
4: **for** M times **do**
5: Select the data object from the top of pr
6: Change the class label of the selected object in D
7: Select the data object from the top of dem
8: Change the class label of the selected object in D
9: Remove the top element both of pr and dem
10: **end for**
11: Train a classifier CND on the modified D
12: **return** CND

[Kamiran and Calders, Classifying without discriminating, ICCCC 2009]

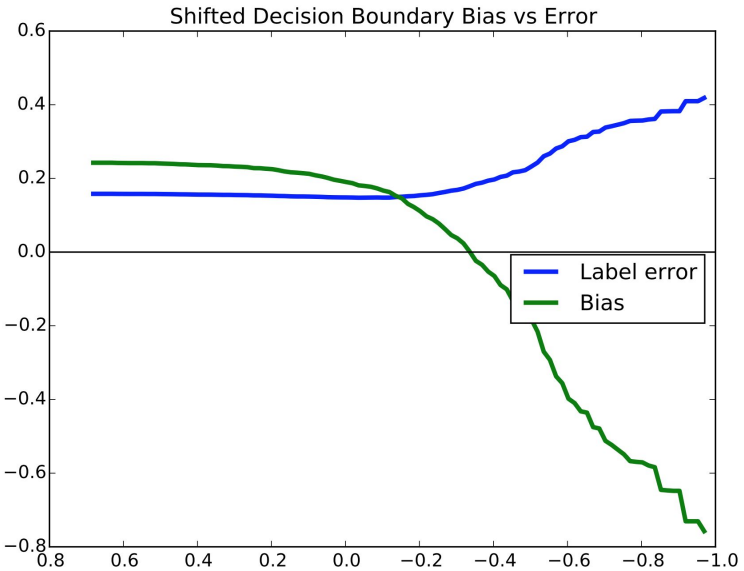
How can we reduce bias in ML models?

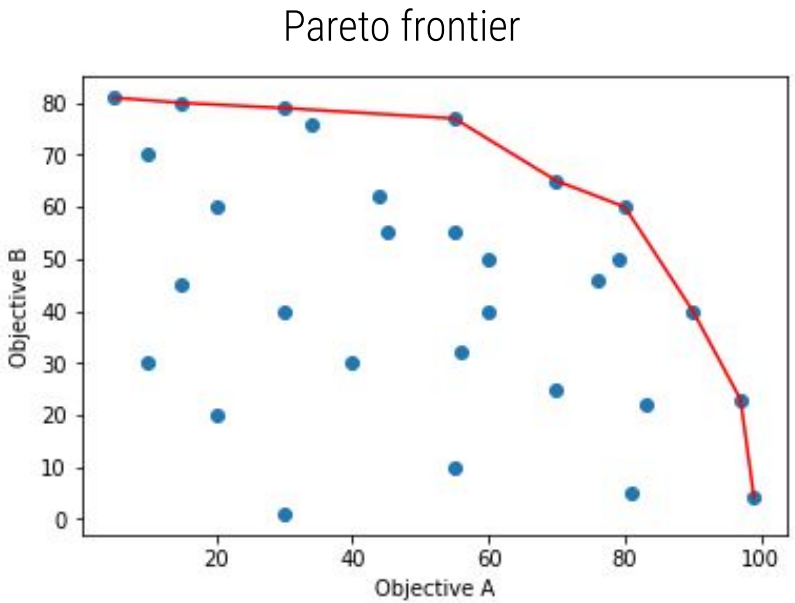
- Fix the world
- Fix the input data
  - ~~Remove sensitive attributes~~
  - Resample and/or reweight protected groups
- **Choose fair models during model selection**
- Optimize for fairness in model training
- Post-hoc adjustments to de-bias model scores


0%
100%
Performance
Metric


100%
Bias
0%
100%
Performance
Metric

Fairness-Accuracy Tradeoff



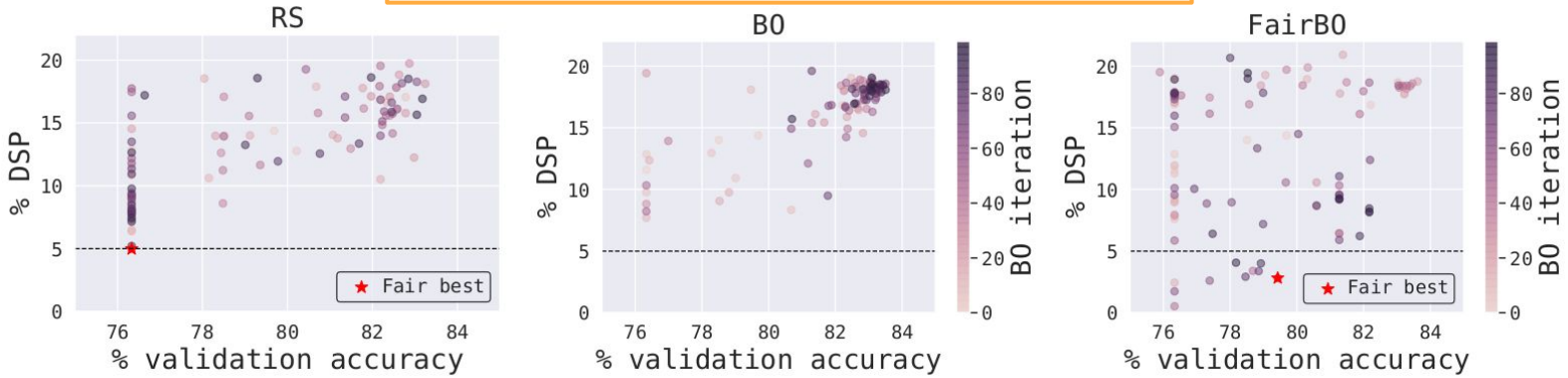


[Fish et al., A Confidence-Based Approach for Balancing Fairness and Accuracy, ICDM 2016]

Fair Bayesian Hyperparameter Optimization
(June 2020)

- Bayesian optimization approach
  - Blind to resource usage
  - Constraint may not be possible to fulfill

$$cEI(\mathbf{x}) = EI(\mathbf{x})P(c(\mathbf{x}) \leq \epsilon)$$



[ref. link]

How can we reduce bias in ML models?

- Fix the world
- Fix the input data
  - ~~Remove sensitive attributes~~
  - Resample and/or reweight protected groups
- Choose fair models during model selection
- **Optimize for fairness in model training**
- Post-hoc adjustments to de-bias model scores

Constrained Optimization

$$\min_{\theta \in \Theta} \mathbb{E}_{x \sim \mathcal{D}} \left[\ell_0 \left(x; \theta \right) \right] \quad \underset{\forall i \in [m]}{\text{s.t.}} \quad \mathbb{E}_{x \sim \mathcal{D}} \left[\ell_i \left(x; \theta \right) \right] \leq 0$$

Minimize loss l0 subject to n data-dependent constraints li.

A Reductions Approach to Fair Classification

Alekh Agarwal [1] Alina Beygelzimer [2] Miroslav Dudík [1] John Langford [1] Hanna Wallach [1]

Algorithm 1 Exp. gradient reduction for fair classification

Input: training examples $\{(X_i, Y_i, A_i)\}_{i=1}^n$
fairness constraints specified by g_j, $\mathcal{E}_j$, $\mathbf{M}$, $\widehat{\mathbf{c}}$
bound B, accuracy ν, learning rate η

Set $\boldsymbol{\theta}_1 = \mathbf{0} \in \mathbb{R}^{|\mathcal{K}|}$

for $t = 1, 2, \ldots$ **do**

Set $\lambda_{t,k} = B \frac{\exp\{\theta_k\}}{1+\sum_{k' \in \mathcal{K}} \exp\{\theta_{k'}\}}$ for all $k \in \mathcal{K}$

$h_t \leftarrow \textsc{Best}_h(\boldsymbol{\lambda}_t)$

$\widehat{Q}_t \leftarrow \frac{1}{t} \sum_{t'=1}^{t} h_{t'}, \quad \overline{L} \leftarrow L\left(\widehat{Q}_t, \textsc{Best}_{\boldsymbol{\lambda}}(\widehat{Q}_t)\right)$

$\widehat{\boldsymbol{\lambda}}_t \leftarrow \frac{1}{t} \sum_{t'=1}^{t} \boldsymbol{\lambda}_{t'}, \quad \underline{L} \leftarrow L\left(\textsc{Best}_h(\widehat{\boldsymbol{\lambda}}_t), \widehat{\boldsymbol{\lambda}}_t\right)$

$\nu_t \leftarrow \max\left\{L(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t) - \underline{L}, \quad \overline{L} - L(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t)\right\}$

if $\nu_t \leq \nu$ **then**

Return $(\widehat{Q}_t, \widehat{\boldsymbol{\lambda}}_t)$

end if

Set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta\,(\mathbf{M}\widehat{\boldsymbol{\mu}}(h_t) - \widehat{\mathbf{c}})$

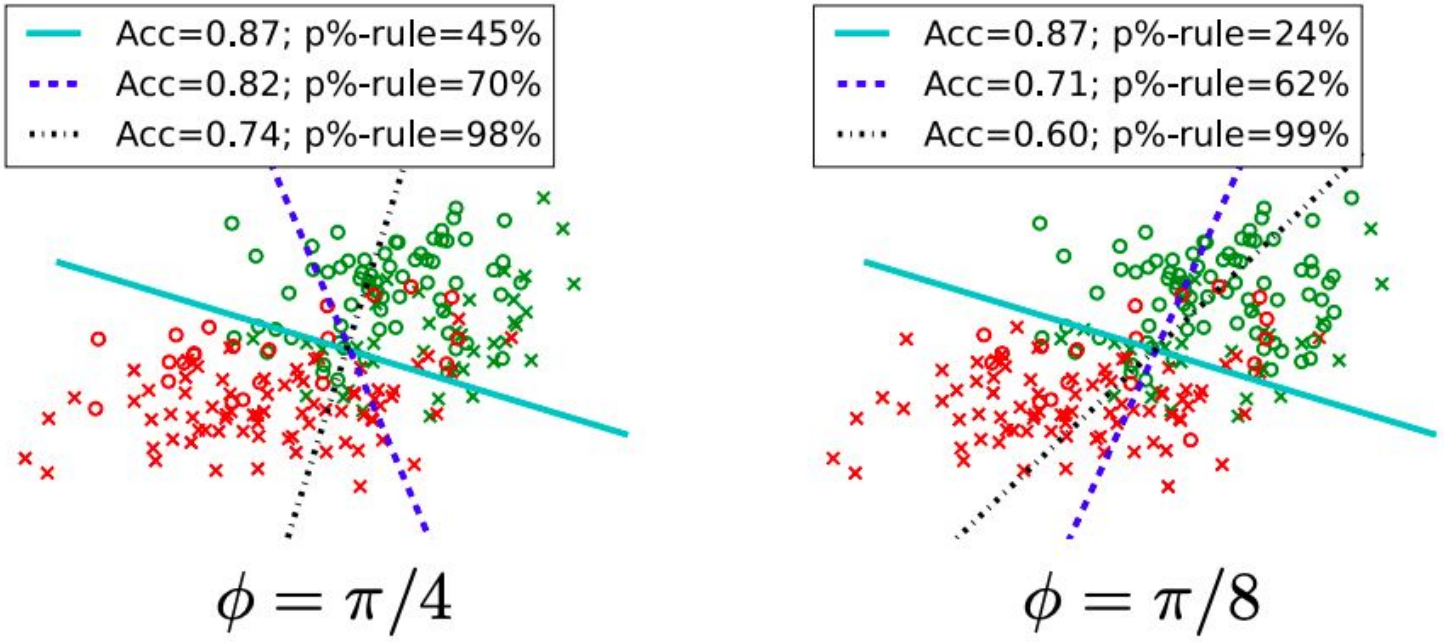
end for

Optimize for Fairness in Model Training

Fairness Constraints: Mechanisms for Fair Classification

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi
Max Planck Institute for Software Systems (MPI-SWS), Germany

Optimize for Fairness in Model Training



(a) Maximizing accuracy under fairness constraints

Training Well-Generalizing Classifiers for Fairness Metrics and Other Data-Dependent Constraints

Andrew Cotter [1] Maya Gupta [1] Heinrich Jiang [1] Nathan Srebro [2] Karthik Sridharan [3] Serena Wang [1]
Blake Woodworth [2] Seungil You [4]

Lagrangian Formulation

$$\mathcal{L}(\theta, \lambda) = l_0(\theta) + \sum_{i=1}^{m} \lambda_i \cdot l_i(\theta)$$

How can we reduce bias in ML models?

- Fix the world
- Fix the input data
  - ~~Remove sensitive attributes~~
  - Resample and/or reweight protected groups
- Choose fair models during model selection
- Optimize for fairness in model training
- **Post-hoc adjustments to de-bias model scores**

Post-hoc Adjustments

Equality of Opportunity in Supervised Learning

Moritz Hardt
Google
m@mrtz.org

Eric Price*
UT Austin
ecprice@cs.utexas.edu

Nathan Srebro
TTI-Chicago
nati@ttic.edu

Post-hoc Adjustments

Definition 2.2 (Equal opportunity)**.** We say that a binary predictor $\widehat{Y}$ satisfies *equal opportunity* with respect to A and Y if $\Pr\{\widehat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\widehat{Y} = 1 \mid A = 1, Y = 1\}$.

aka recall or TPR

Post-hoc Adjustments

Definition 2.2 (Equal opportunity)**.** We say that a binary predictor $\widehat{Y}$ satisfies *equal opportunity* with respect to A and Y if $\Pr\{\widehat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\widehat{Y} = 1 \mid A = 1, Y = 1\}$.

> *That is, to require that people who pay back their loan have an equal opportunity of getting a loan in the first place.*

Post-hoc Adjustments

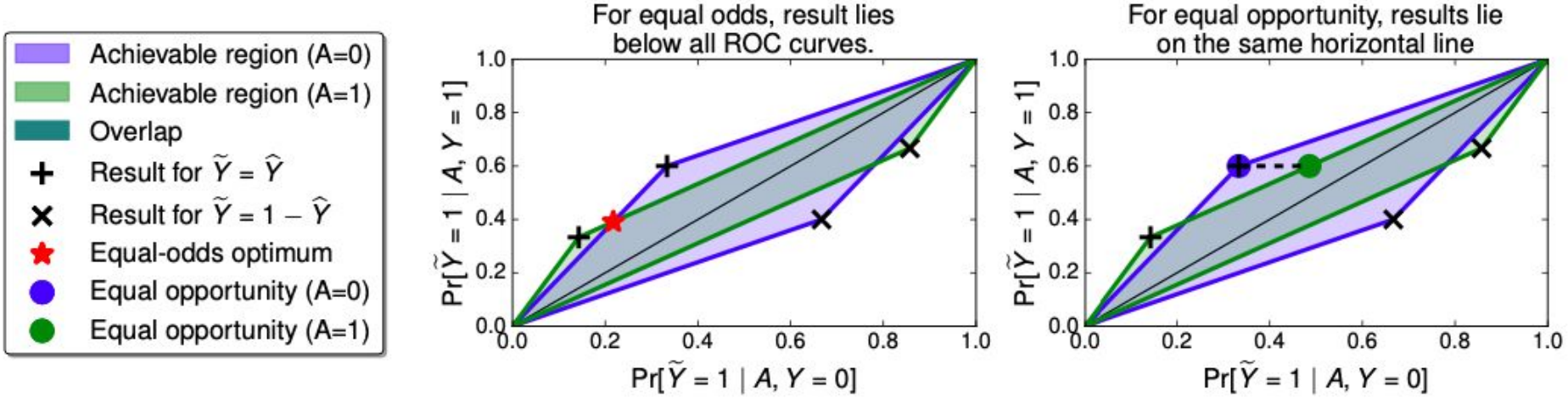
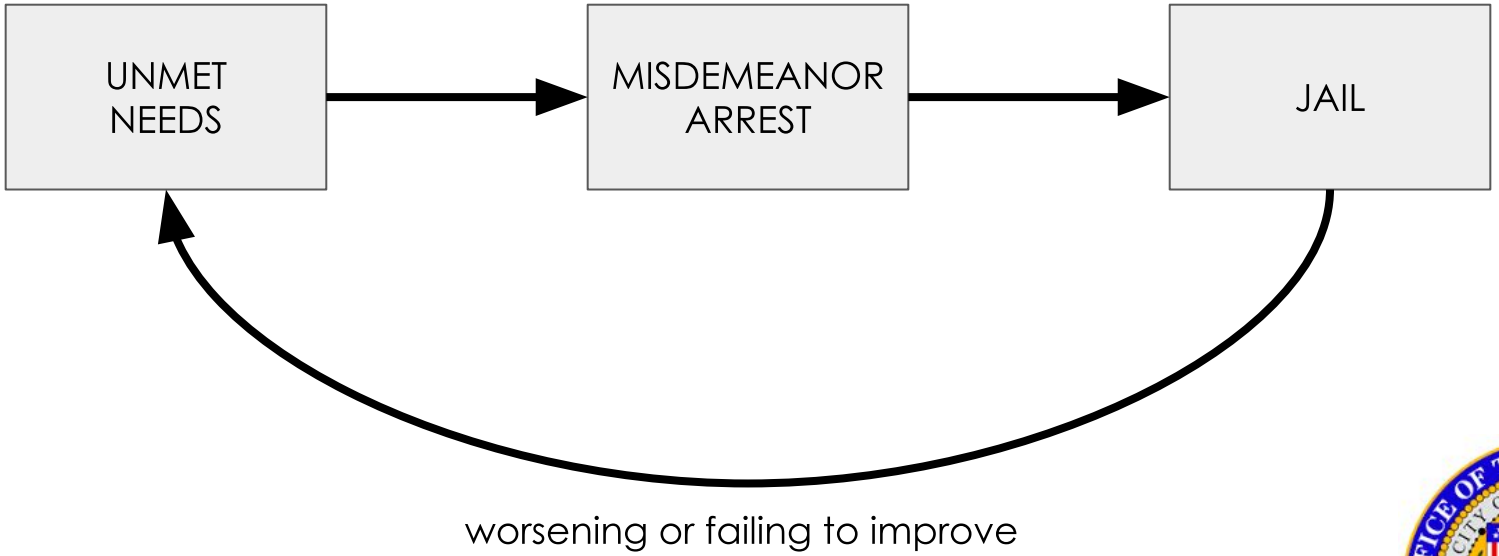


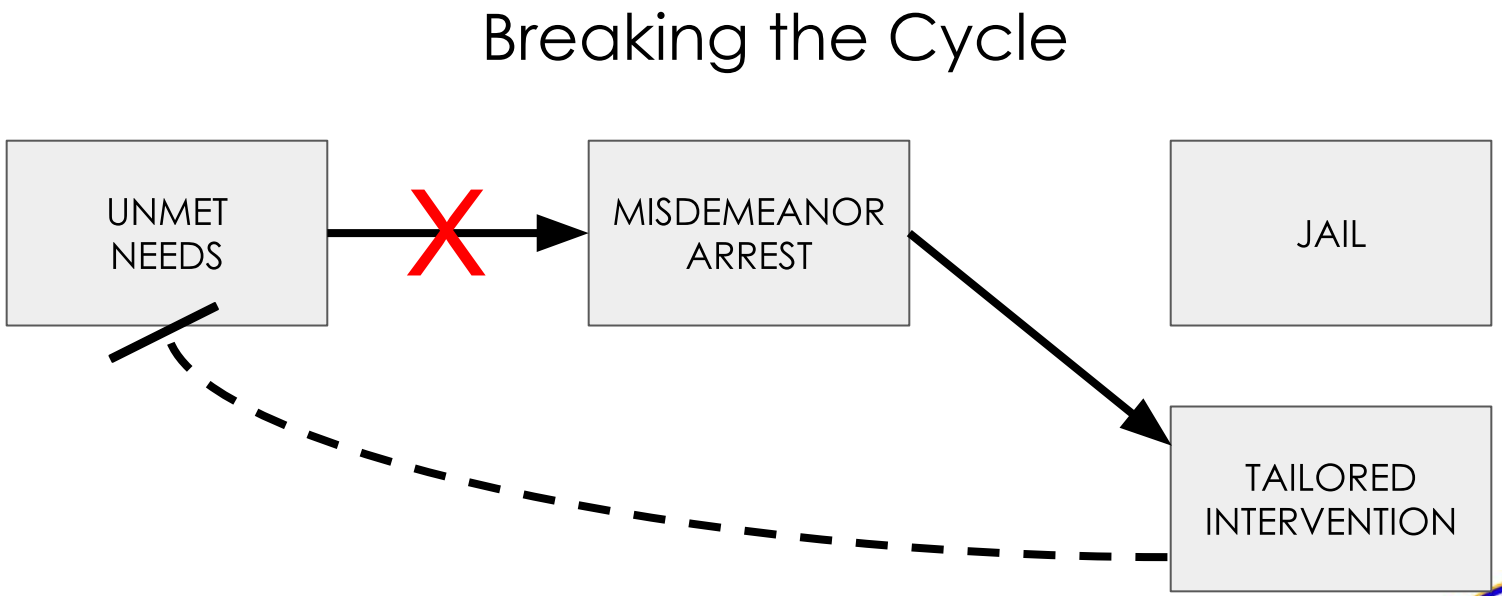
Figure 1: Finding the optimal equalized odds predictor (left), and equal opportunity predictor (right).

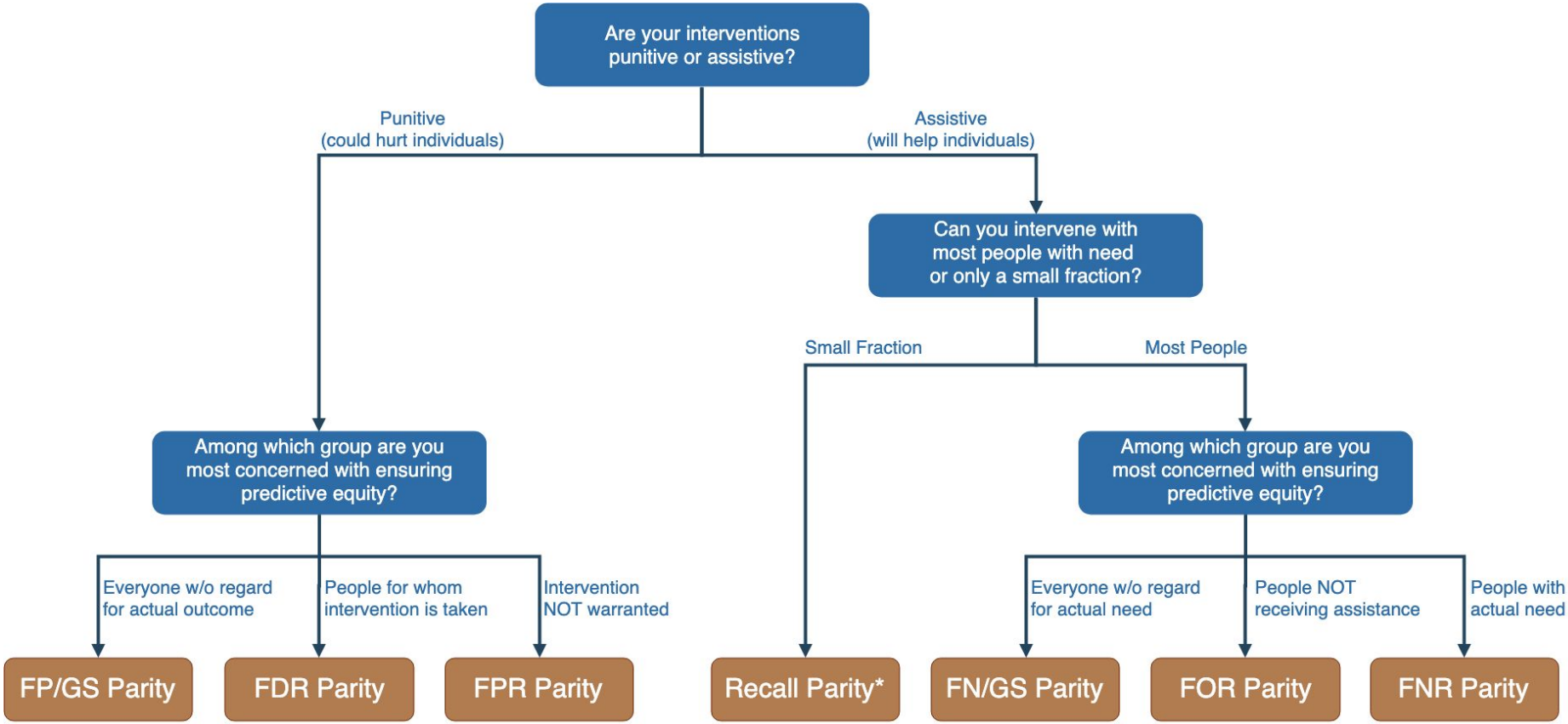
Case Study: Post-hoc Adjustments and Policy Goals

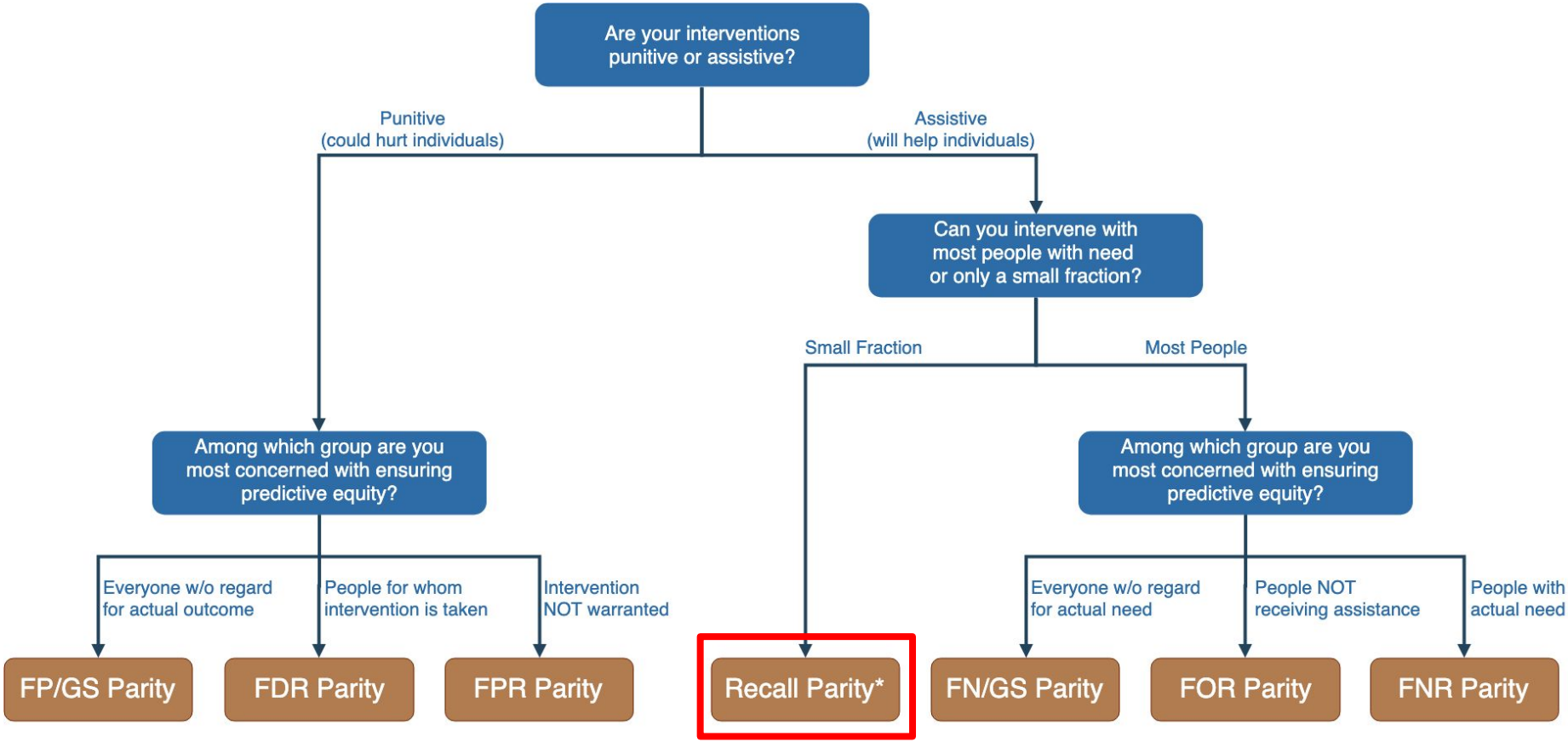
Cycle of Incarceration

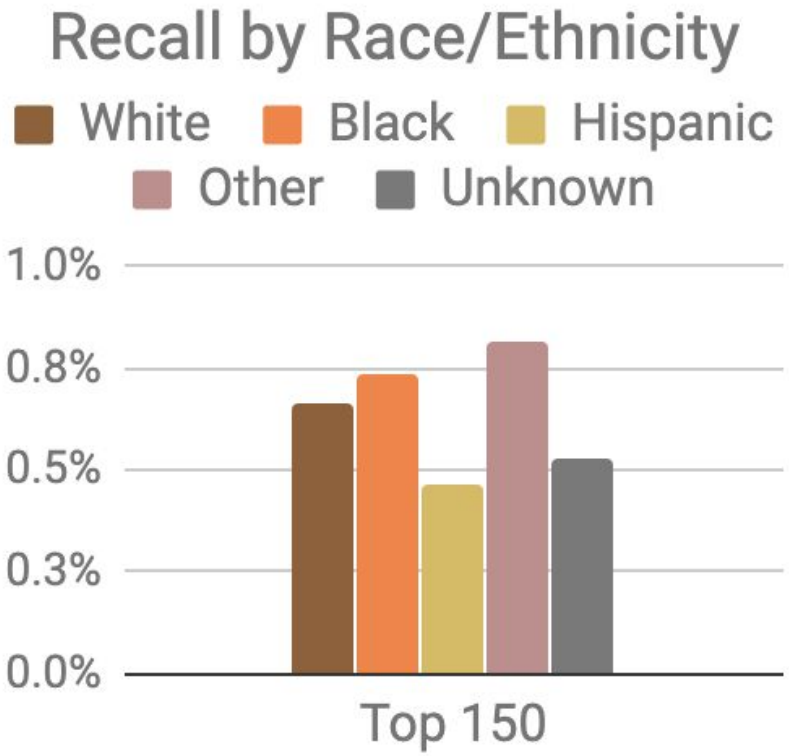


Case Study: Post-hoc Adjustments and Policy Goals



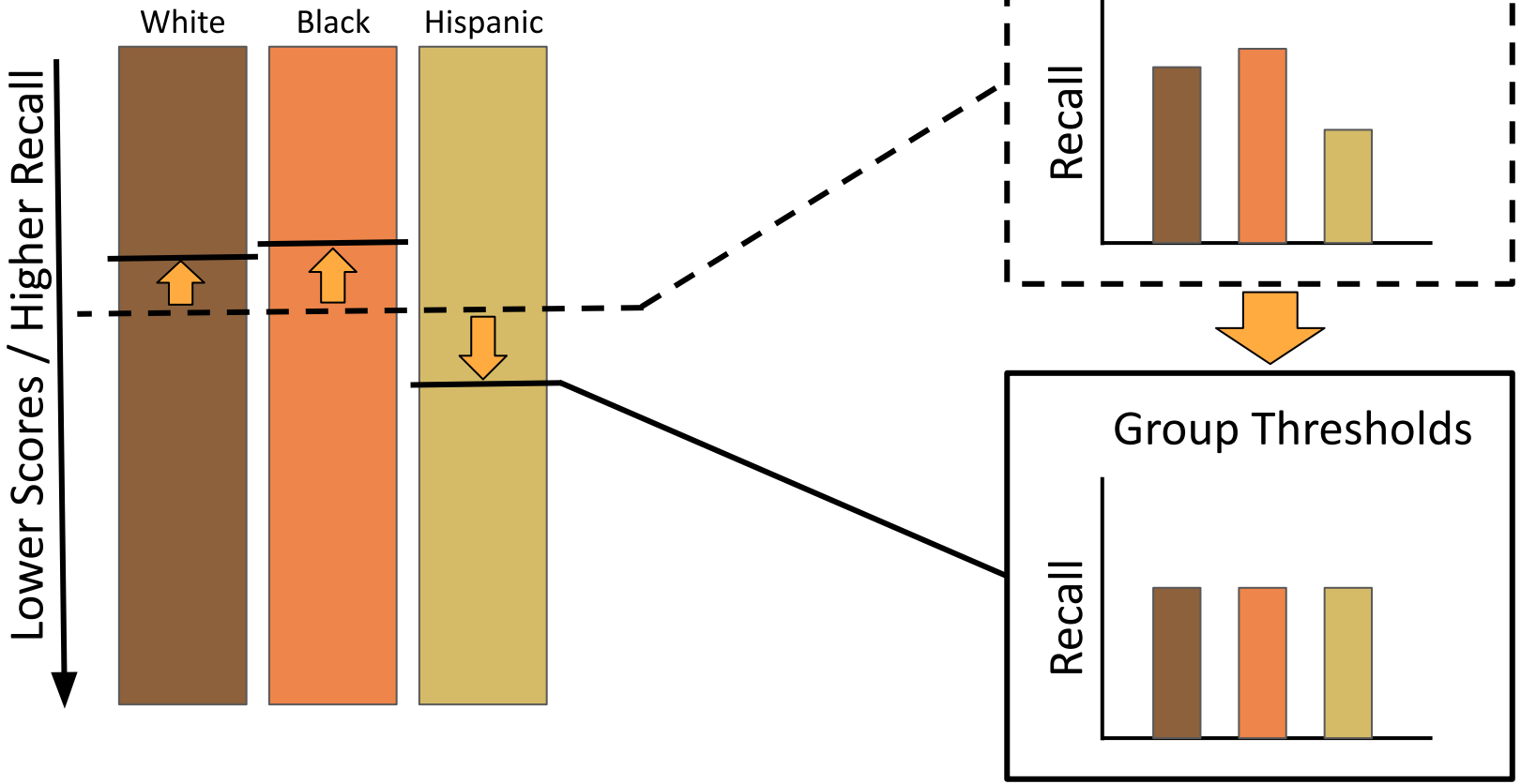

Are your interventions punitive or assistive?
Punitive (could hurt individuals)
Assistive (will help individuals)
Can you intervene with most people with need or only a small fraction?
Small Fraction
Most People
Among which group are you most concerned with ensuring predictive equity?
Among which group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


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?
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
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

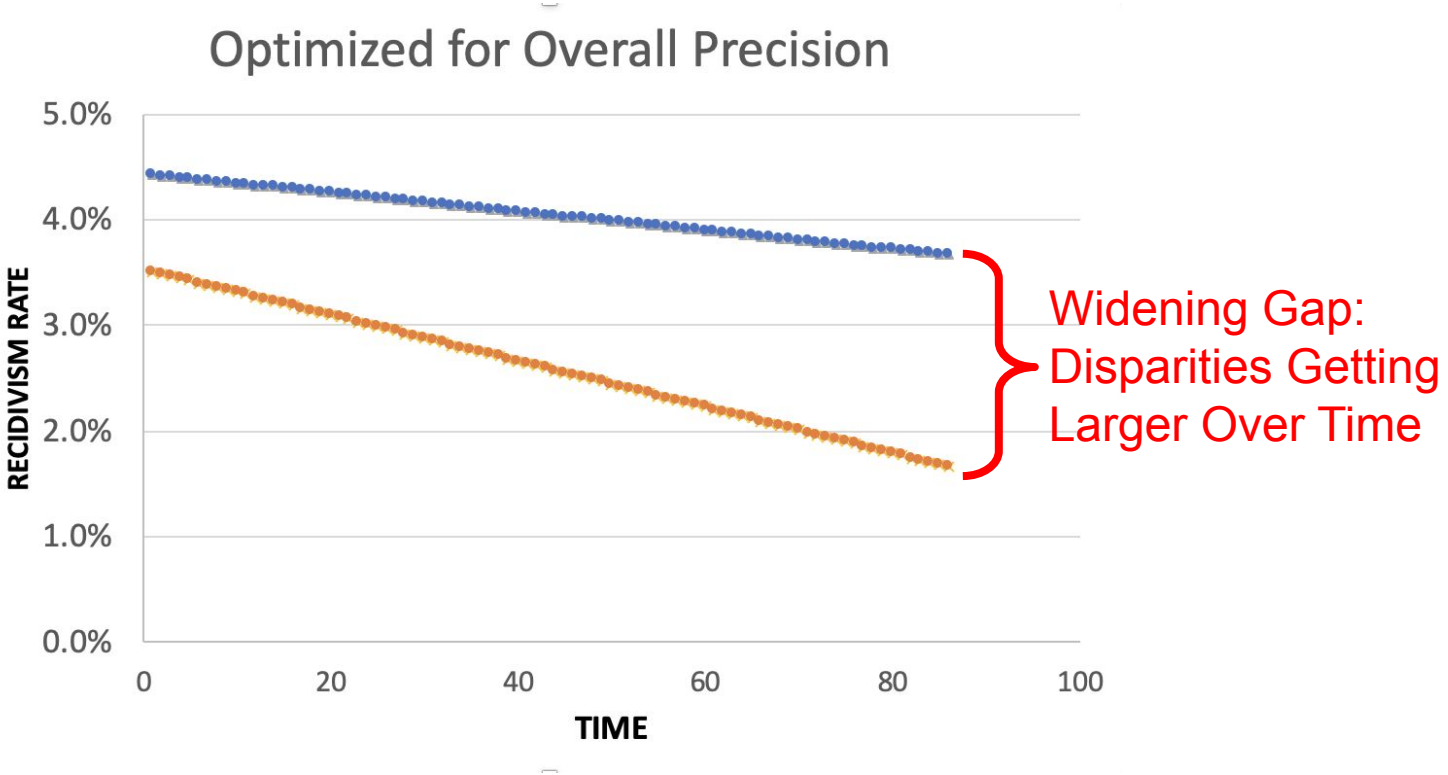


- Model was optimized for efficiency, not equity
- Top 150 highest risk reasonably balanced between black and white individuals
- However, hispanic and unknown race/ethnicity groups very underrepresented

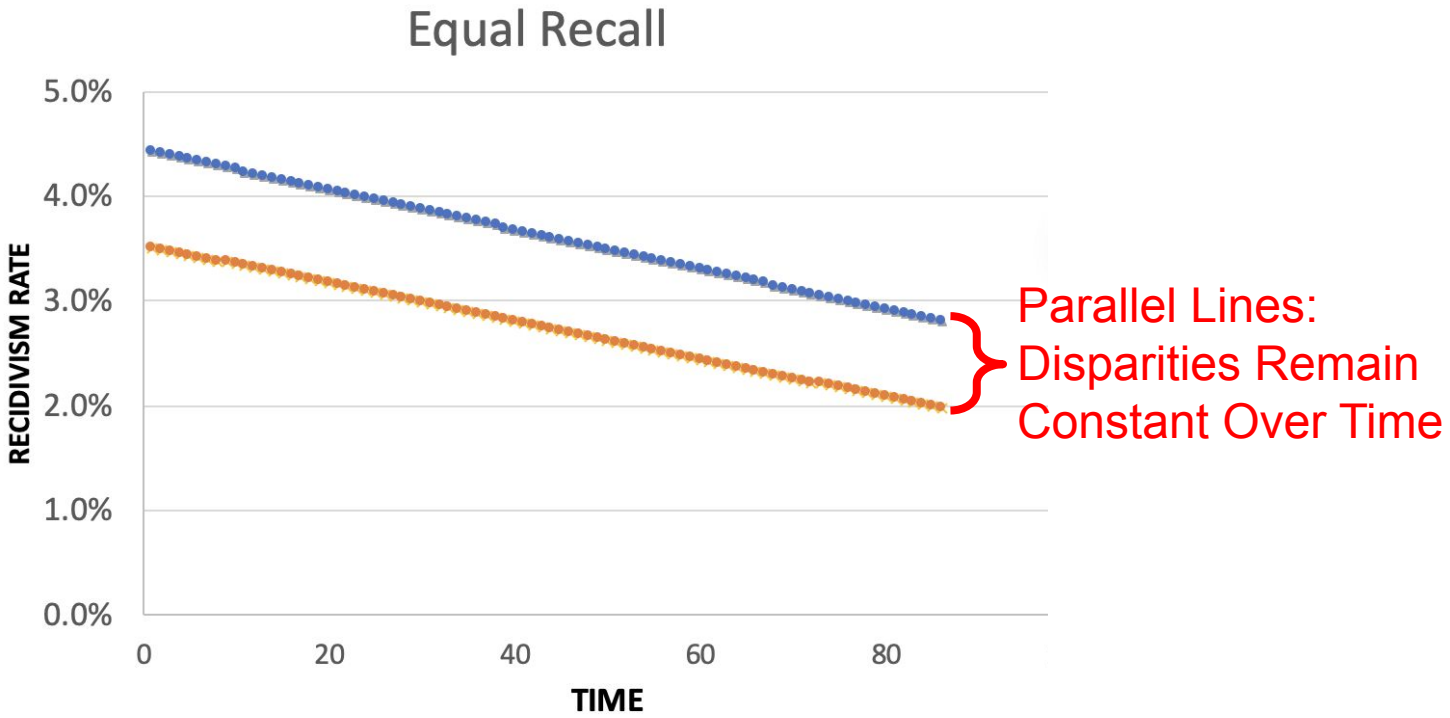
Mitigating Disparities



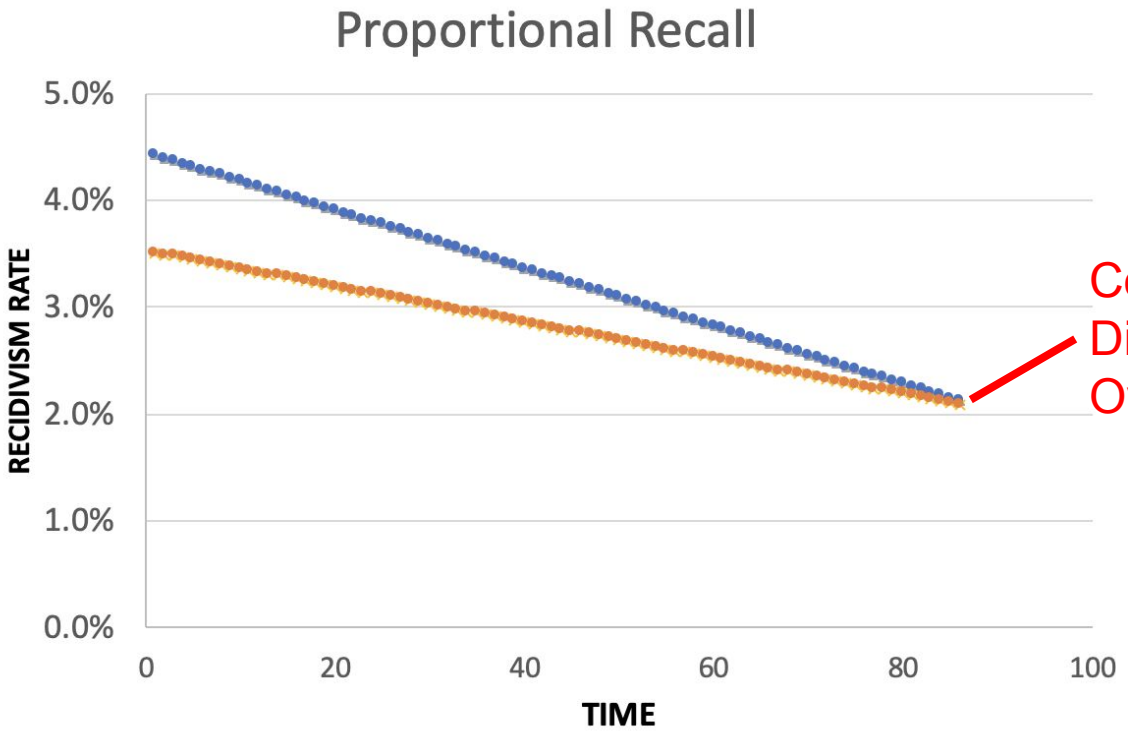
Equality of Predictions vs Underlying Disparities

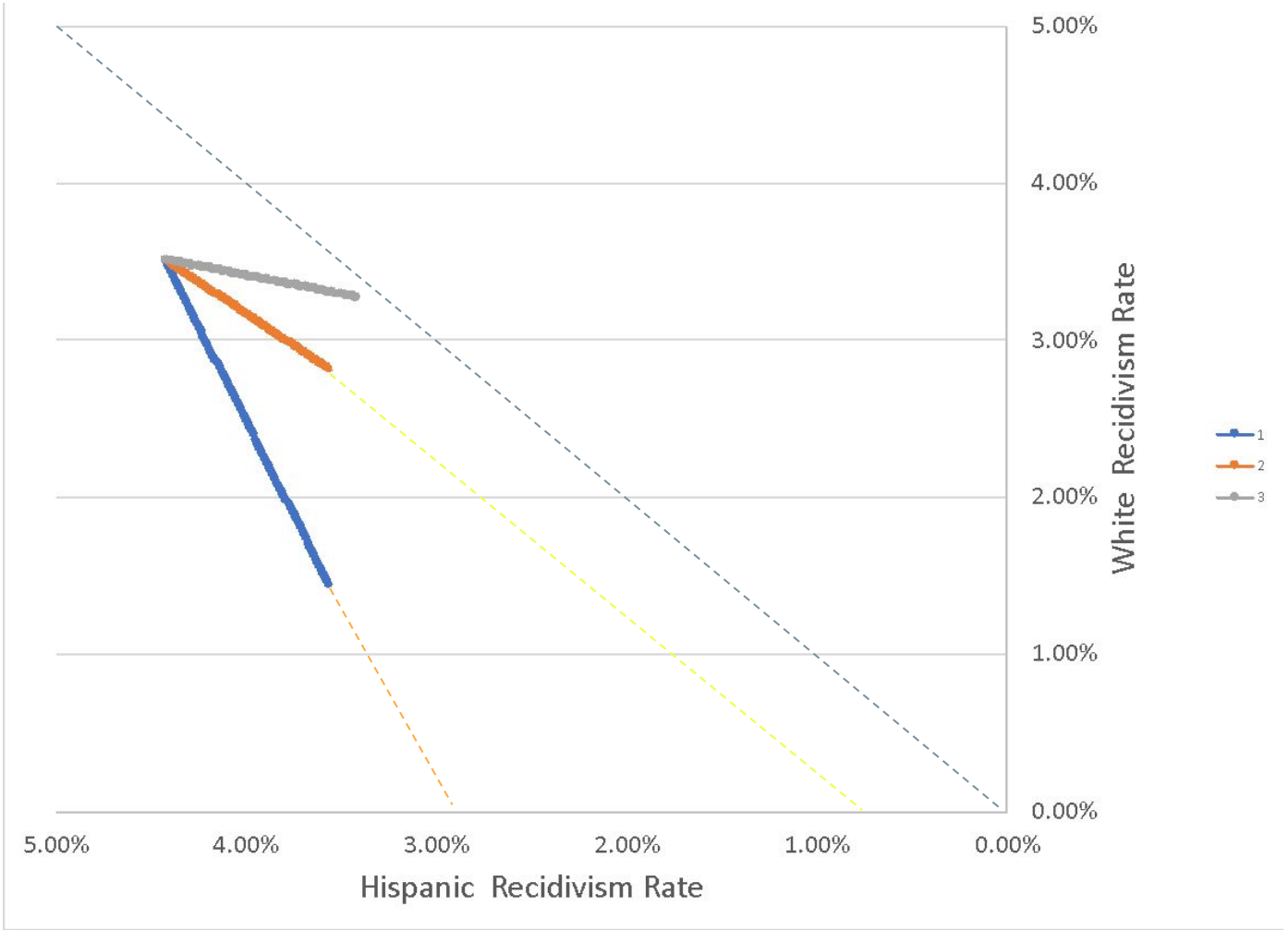


Equality of Predictions vs Underlying Disparities



Equality of Predictions vs Underlying Disparities




5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
White Recidivism Rate
1
2
3
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
Hispanic Recidivism Rate

Menu of Options

	Current Scale	Expanded Scale
No Constraint		
Equalize Recall		
Reduce Disparities		

Menu of Options

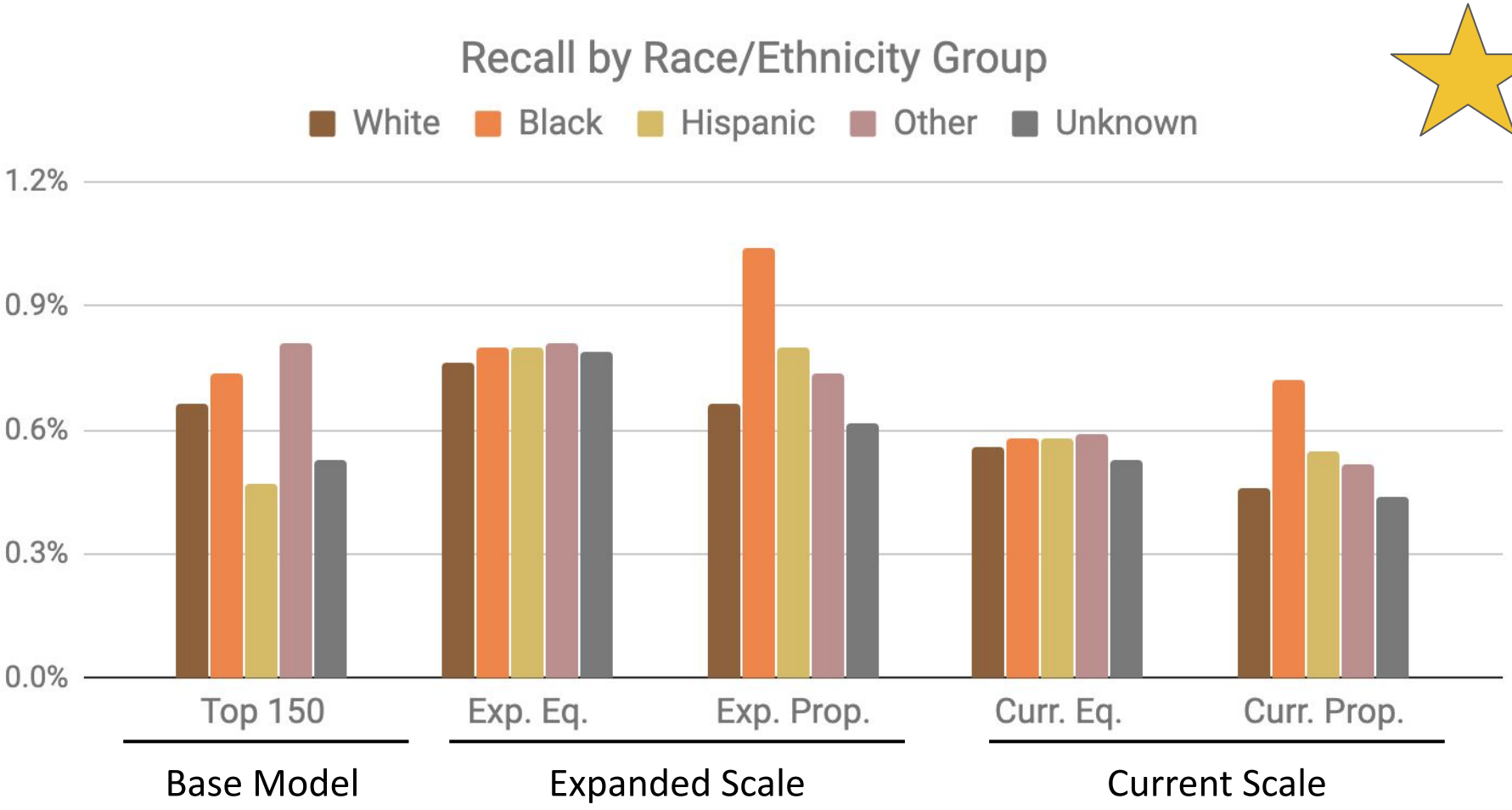
	Current Scale	Expanded Scale
No Constraint	INITIAL MODEL	
Equalize Recall		
Reduce Disparities		

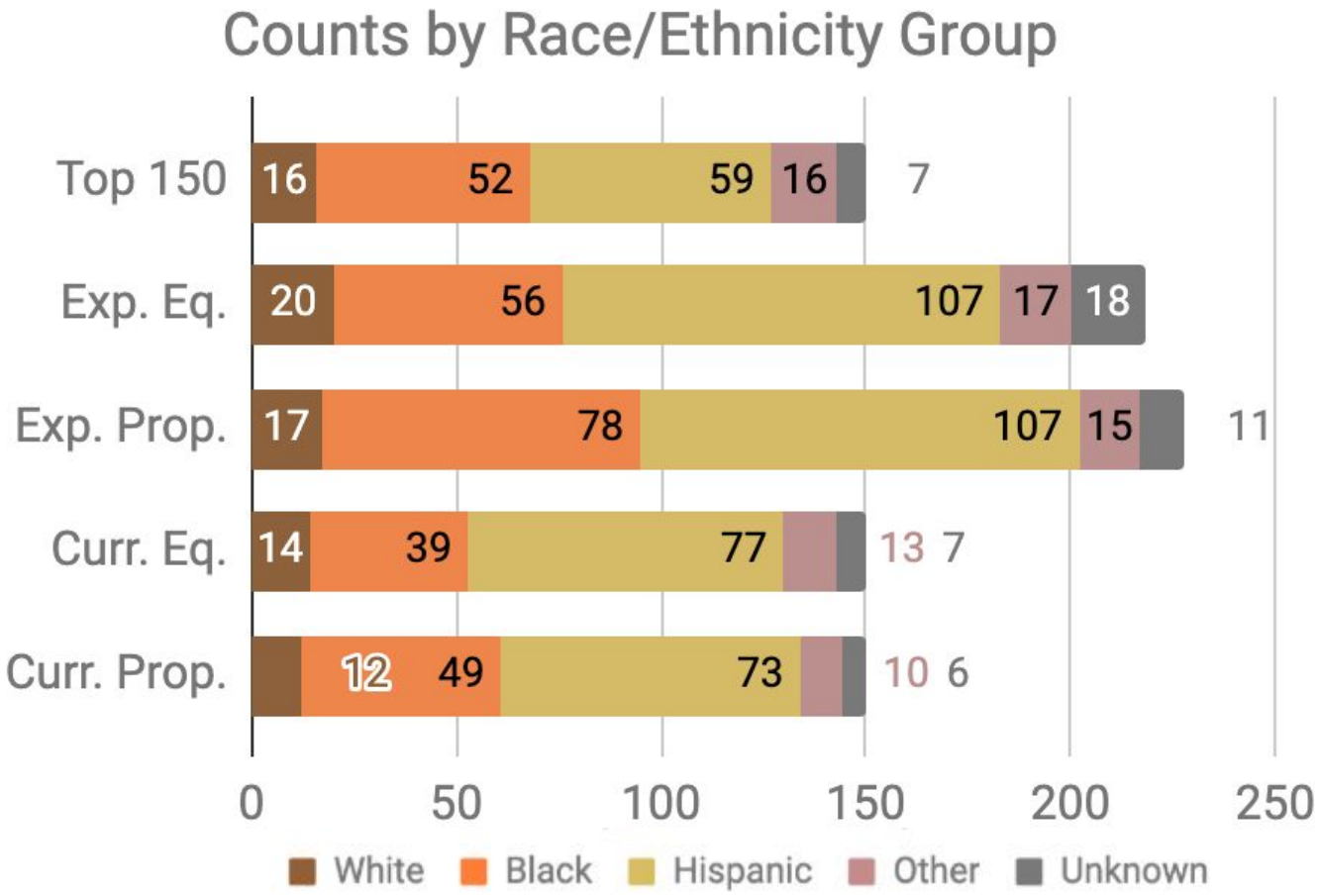
Menu of Options

	Current Scale	Expanded Scale
No Constraint		
Equalize Recall Reduce Disparities	EXPLICIT EQUITY / EFFICIENCY TRADE-OFF	"COST OF EQUITY"

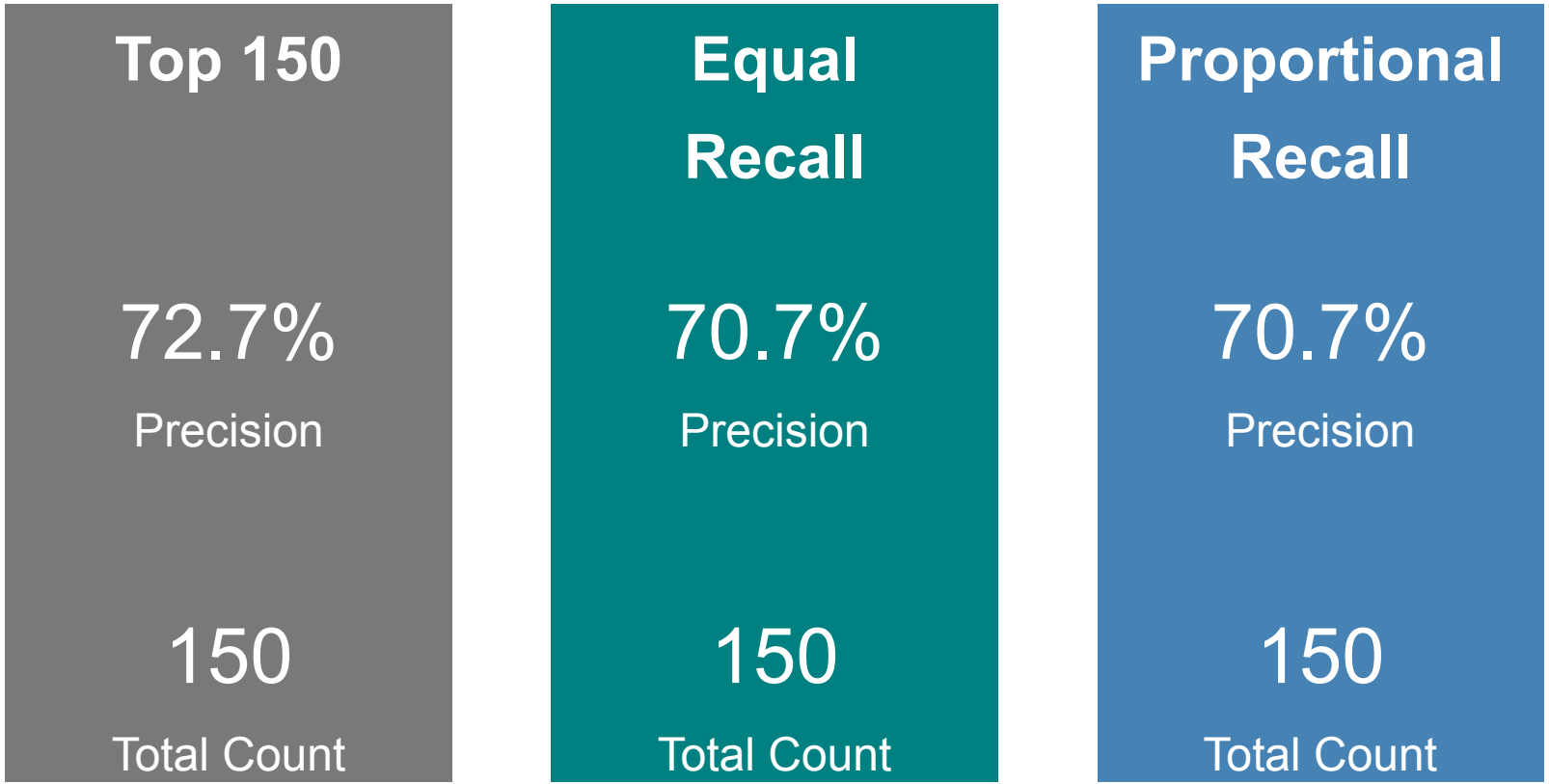
Menu of Options

	Current Scale	Expanded Scale
No Constraint		
Equalize Recall	IMPROVE OUTCOMES AT SAME RATE ACROSS GROUPS	
Reduce Disparities	IMPROVE OUTCOMES FASTER FOR GROUPS WITH HIGHER INCARCERATION RATES	


Recall by Race/Ethnicity Group
White
Black
Hispanic
Other
Unknown
1.2%
0.9%
0.6%
0.3%
0.0%
Top 150
Exp. Eq.
Exp. Prop.
Curr. Eq.
Curr. Prop.
Base Model
Expanded Scale
Current Scale


Counts by Race/Ethnicity Group
Top 150
16
52
59
16
7
Exp. Eq.
20
56
107
17
18
Exp. Prop.
17
78
107
15
11
Curr. Eq.
14
39
77
13
7
Curr. Prop.
12
49
73
10
6
0
50
100
150
200
250
White
Black
Hispanic
Other
Unknown

Little Equity/Efficiency Trade-Off at Current Scale



Delayed Impact of Fair Machine Learning

Lydia T. Liu* Sarah Dean* Esther Rolf* Max Simchowitz* Moritz Hardt*

HANDS-ON:

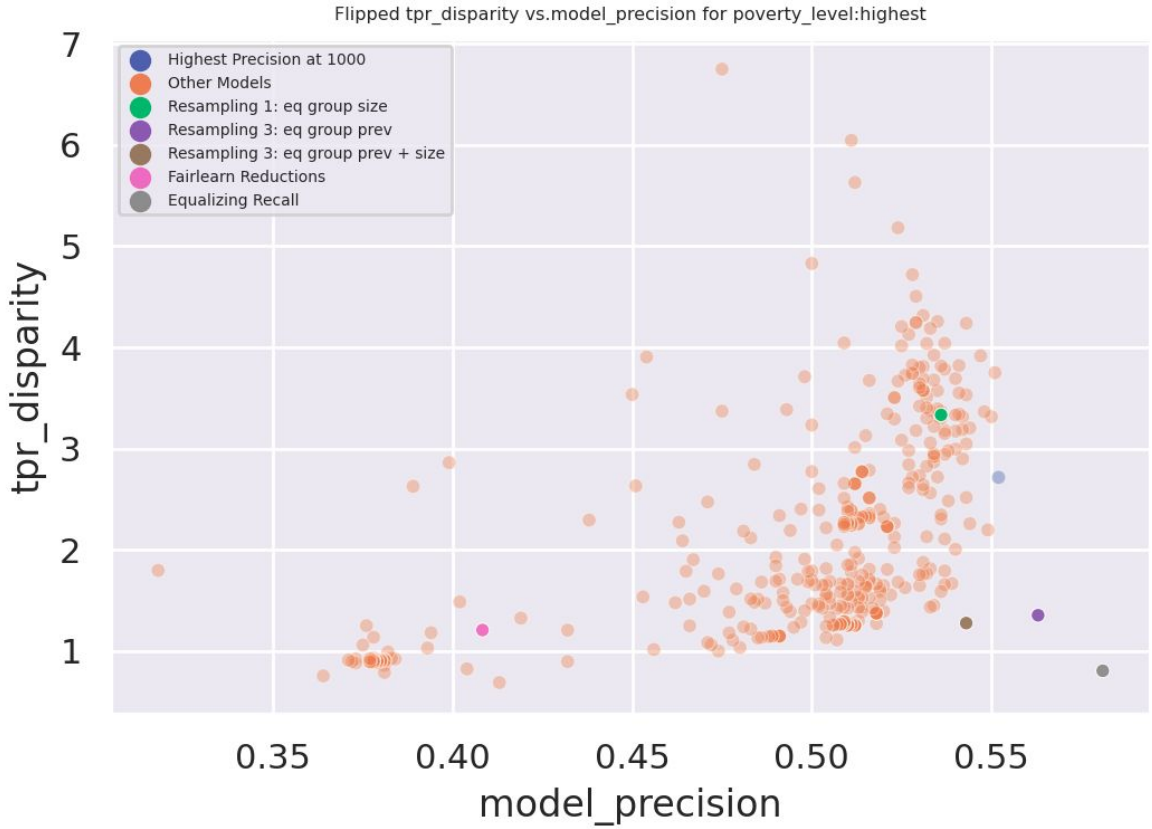
Improving Fairness

(click on link to open notebook)

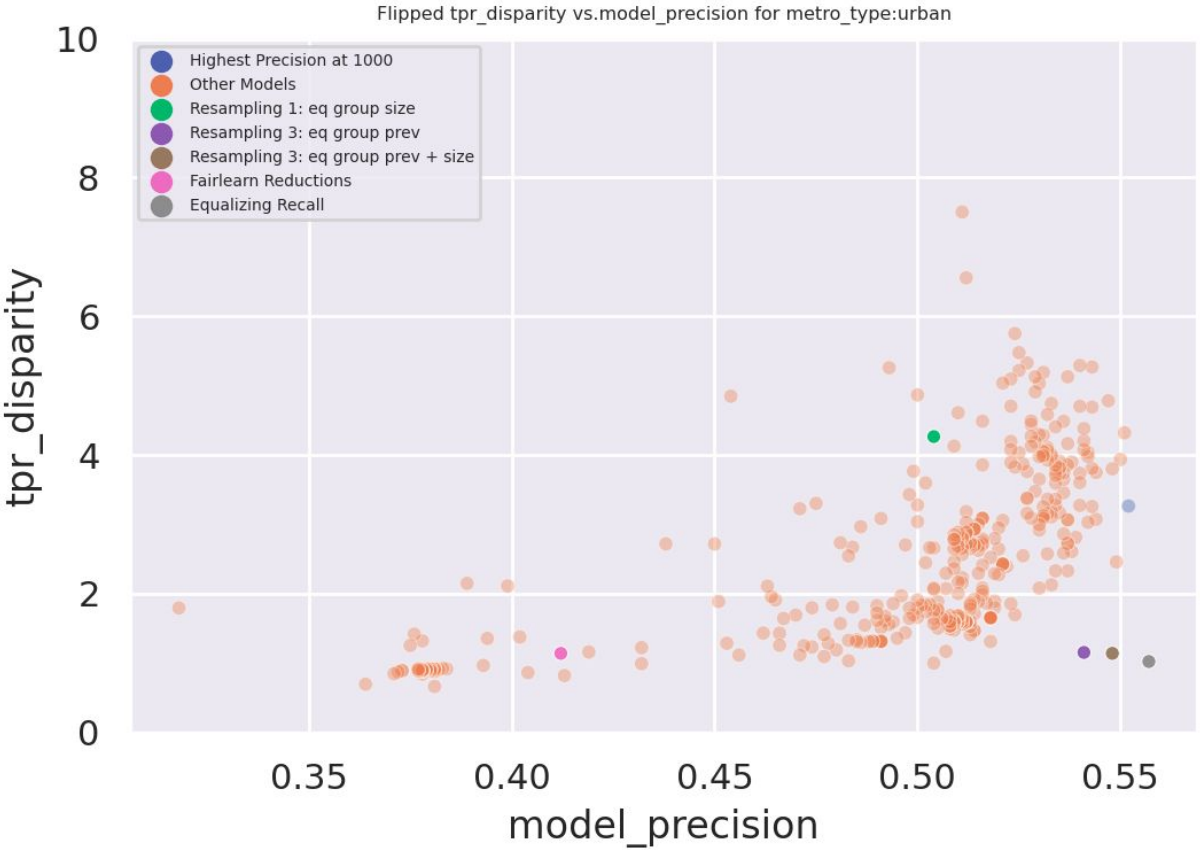
Regroup: Improving Fairness

What did we find?

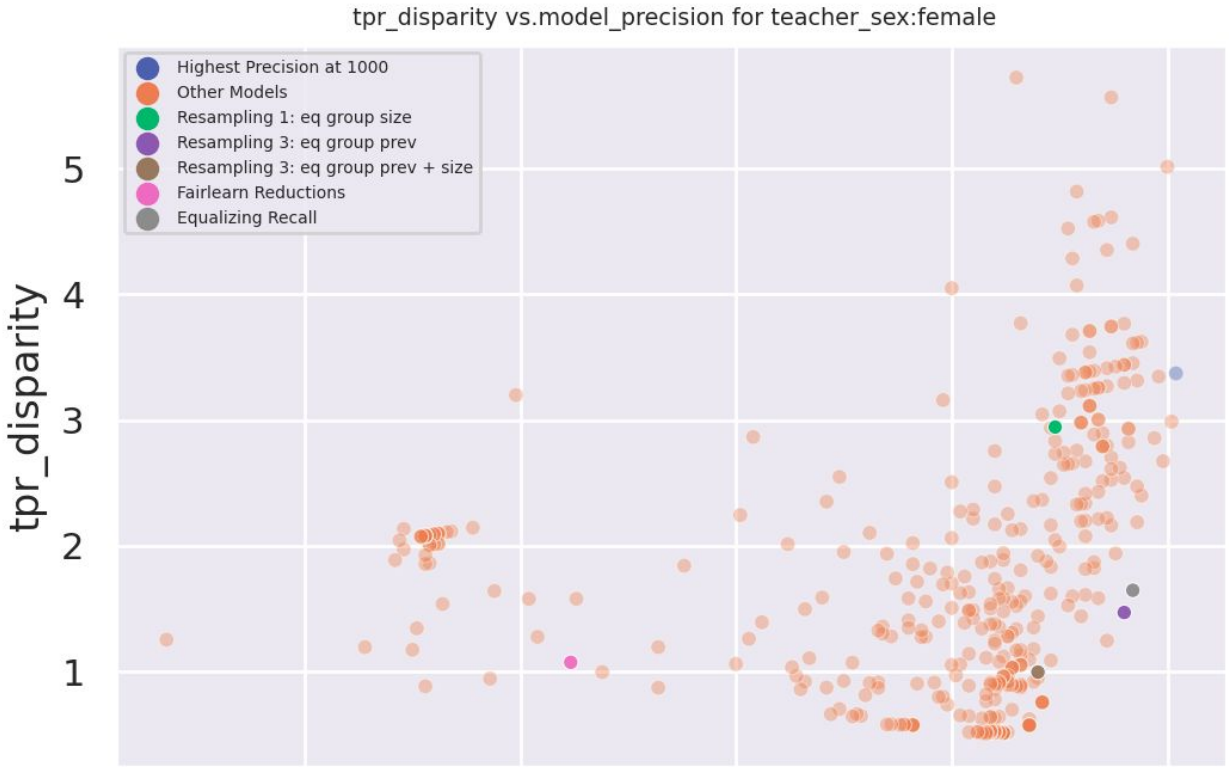
Change in disparity for poverty level

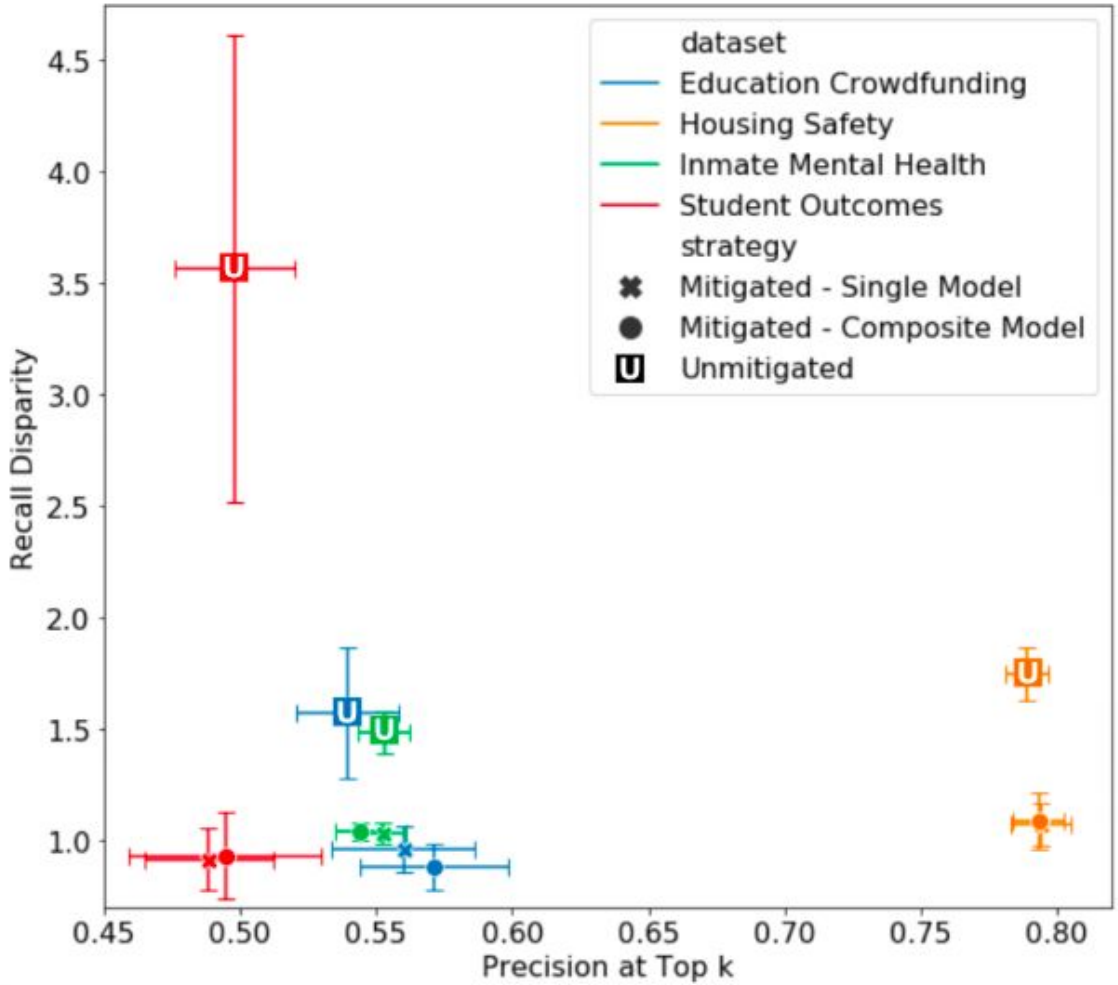


Change in disparity for metro_type



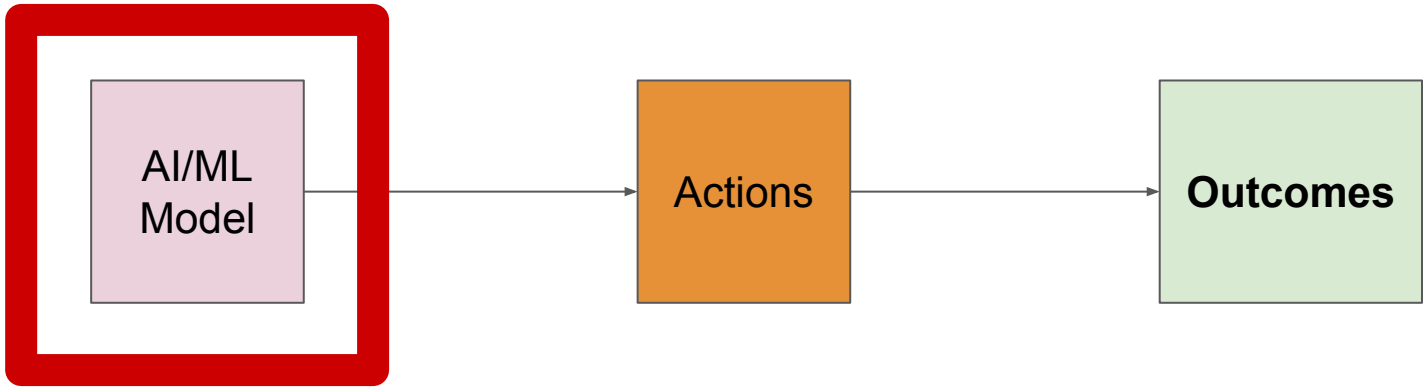
Change in disparity for teacher sex




dataset
Education Crowdfunding
Housing Safety
Inmate Mental Health
Student Outcomes
strategy
Mitigated - Single Model
Mitigated - Composite Model
Unmitigated
Recall Disparity
Precision at Top k
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80

Wrap-Up

The goal is not to make the ML model fair but to **make the overall system and outcomes fair**



Things to remember

Make bias, fairness, and equity an integral part of every project: Scoping, community engagement, metrics, validation, monitoring outcomes

Understand how different phases of the project could lead to downstream bias

All bias metrics are not created equal - use the Fairness Tree to understand your problem/use case and select appropriate metrics

Audit and Explore bias reduction strategies

A perfectly fairn model does not mean fair outcomes. Think about the entire system (including actions) and measure outcomes

Compared to what?

Some useful practices

Create an environment where informed ethical discussions can take place

Talk through ethical issues at each stage of the project (instead of waiting till the end of stopping after the initial setup)

Consider the entire chain of data - collection to analysis to action

Consider how it affects people throughout the chain – especially the people being affected (and include them in these discussions)

Embed ethics into both technical processes as well as people processes

Links

Website: https://dssg.github.io/fairness_tutorial

Github Repo: http://github.com/dssg/fairness_tutorial

Interactive Colab (Python) Notebooks: https://dssg.github.io/fairness_tutorial/notebooks/

Aequitas (Audit Tool): http://www.datasciencepublicpolicy.org/aequitas/

Fairness Tree

https://github.com/dssg/fairness_tutorial

Resources

How do we scope data science projects?

More details at http://www.datasciencepublicpolicy.org/resources/data-science-project-scoping-guide/

Goals: Define the goal(s) of the project (**equity**, efficiency, effectiveness, etc.)

Actions: What actions/interventions will you inform?

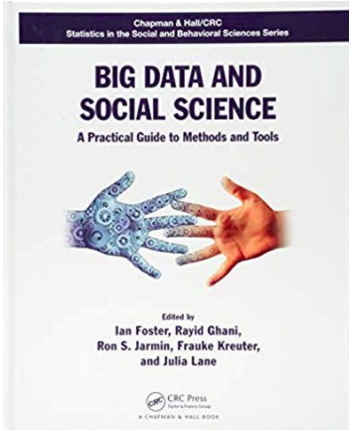
Data: What data do you have internally?

What data do you need?

What can you augment from external and public sources?

Analysis: What analysis needs to be done? How will it be validated? How will the analysis achieve the goals defined above?

http://www.datasciencepublicpolicy.org/aequitas/



11

Bias and Fairness

Kit T. Rodolfa, Pedro Saleiro, and Rayid Ghani

Interest in algorithmic fairness and bias has been growing recently, but it's easy to get lost in the large number of definitions and metrics. There are many different, often competing, ways to measure whether a given model is "fair". In this chapter, we provide an overview of these metrics along with some concrete examples to help navigate these concepts and understand the trade-offs involved in choosing to optimize to one metric over others, focusing on the metrics relevant to binary classification methods used frequently in risk-based models for policy settings.

11.1 Introduction

In Chapter Machine Learning, you learned about several of the concepts, tools, and approaches used in the field of machine learning and how they can be used in the social sciences. In chapter Machine Learning, we focused

https://textbook.coleridgeinitiative.org/chap-bias.html

Algorithmic Impact Assessment (Canada)

Risk Profile

Is the project within an area of intense public scrutiny (e.g. because of privacy concerns) and/or frequent litigation?

- Yes
- No

Are clients in this line of business particularly vulnerable?

- Yes
- No

Are stakes of the decisions very high?

- Yes
- No

Will this project have major impacts on staff, either in terms of their numbers or their roles?

- Yes
- No

https://canada-ca.github.io/aia-eia-js/

No. The data was crawled from public web sources, and the individuals appeared in news stories. But there was no explicit informing of these individuals that their images were being assembled into a dataset.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis)been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown

Any other comments?

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The following steps were taken to process the data:

1. **Gathering raw images:** First the raw images for this dataset were obtained from the Faces in the Wild dataset consisting of images and associated captions gathered from news articles found on the web.
2. **Running the Viola-Jones face detector**[4] The OpenCV version 1.0.0 release 1 implementation of Viola-Jones face detector was used to detect faces in each of these images, using the function cvHaarDetectObjects, with the provided Haar classifier—cascadehaarcascadefrontalfacedefault.xml. The scale factor was set to 1.2, min neighbors was set to 2, and the flag was set to CV HAAR DO CANNY PRUNING.
3. **Manually eliminating false positives:** If a face was detected and the specified region was determined not to be a face (by the operator), or the name of the person with the detected face could not be identified (using step 5 below), the face was omitted from the dataset.
4. **Eliminating duplicate images:** If images were determined to have a common original source photograph, they are defined to be duplicates of each other. An attempt was made to remove all duplicates but a very small number (that were not initially found) might still exist in the dataset. The number of remaining duplicates should be small enough so as not to significantly impact training/testing. The dataset contains distinct images that are not defined to be duplicates but are extremely similar. For example, there are pictures of celebrities that appear to be taken almost at the same time by different photographers from slightly different angles. These images were not removed.
5. **Labeling (naming) the detected people:** The name associated with each person was extracted from the associated news caption. This can be a source of error if the original news caption was incorrect. Photos of the same person were combined into a single group associated with one name. This was a challenging process as photos of some people were associated with multiple names in the news captions (e.g."Bob McNamara" and "Robert McNamara"). In this scenario, an attempt was made to use the most common name. Some people have a single name (e.g. "Madonna" or "Abdullah"). For Chinese and some other Asian names, the common Chinese ordering (family name followed by given name) was used (e.g. "Hu Jintao").
6. **Cropping and rescaling the detected faces:** Each detected region denoting a face was first expanded by 2.2 in each dimension. If the expanded region falls outside of the image, a new image was created by padding the original pixels with black pixels to fill the area outside of the original image. This expanded region was then resized to 250 pixels by 250 pixels using the function cvResize, and cvSetImageROI as necessary. Images were saved in JPEG 2.0 format.
7. **Forming pairs of training and testing pairs for View 1 and View 2 of the dataset:** Each person in the dataset was randomly assigned to a set (with 0.7 probability of being in a training set in View 1 and uniform probability of being in any set in View 2). Matched pairs were formed by picking a person uniformly at random from the set of people who had two or more images in the dataset. Then, two images were drawn uniformly at random from the set of images of each chosen person, repeating the process if the images are identical or if they were already chosen as a matched pair). Mismatched pairs were formed by first choosing two people uniformly at random, repeating the sampling process if the same person was chosen twice. For each chosen person, one image was picked uniformly at random from their set of images. The process is repeated if both images are already contained in a mismatched pair.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The raw unprocessed data (consisting of images of faces and names of the corresponding people in the images) is saved.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

While a script running a sequence of commands is not available, all software used to process the data is open source and has been specified above.

Any other comments?

[4]Paul Viola and Michael Jones. *Robust real-time face detection*. IJCV, 2004

Dataset audit card - ImageNet

Census audit statistics

- 83436 images with 101070 – 132201 persons (Models: DEX ([69]), InsightFace ([38]))
- Mean-age (male): 33.24 (Female):25.58 (RetinaFace [23], ArcFace [22])
- Confirmed misogynistic images: 62. Number of classes with infants: 30
- $(\mu_c^{(A)}$ and $\sigma_c^{(A)}$: Mean and standard-deviation of the gender-estimate of images in class c estimated by algorithm (A).)

Metrics: Class-level mean count ($\eta_c^{(A)}$), mean gender skewness ($\xi_c^{(A)}$) and mean-age ($\alpha_c^{(A)}$):

$$\eta_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i], \alpha_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i]a_i^{(A)} \text{ and}$$

$$\xi_c^{(A)} = \frac{1}{N_c}\sum_{i=1}^{N_c} I[\phi_i]\left(\frac{g_i^{(A)} - \mu_c^{(A)}}{\sigma_c^{(A)}}\right)^3$$

$$\phi_i = \begin{cases} 1 & \text{if face present} \\ 0 & \text{otherwise} \end{cases} \text{ in } i^{th} \text{ image.}$$

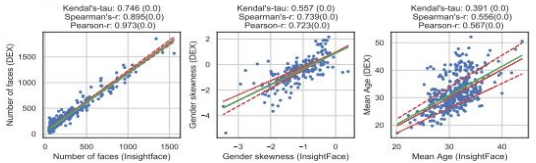


Figure 2: Class-wise cross-categorical scatter-plots across the cardinality, age and gender scores

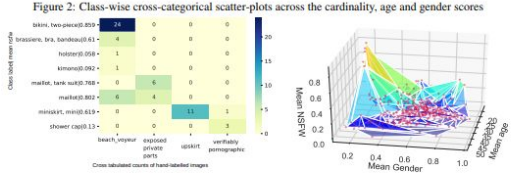


Figure 3: Statistics and locationing of the hand-labelled images

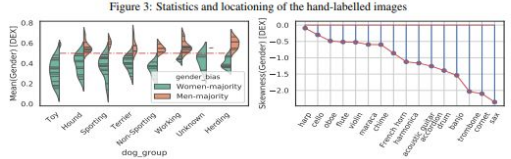


Figure 4: Known *human co-occurrence* based gender-bias analysis

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Fig. 4. Example datasheet for Labeled Faces in the Wild [14], page 4.

https://facctconference.org/

Fairness and machine learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

This online textbook is an incomplete work in progress. Essential chapters are still missing. In the spirit of open review, we solicit broad feedback that will influence existing chapters, as well as the development of later material.

CONTENTS

About this book

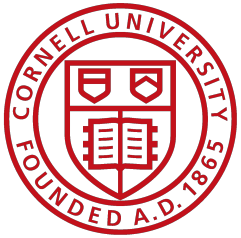
1 Introduction PDF

https://fairmlbook.org/

Course Materials

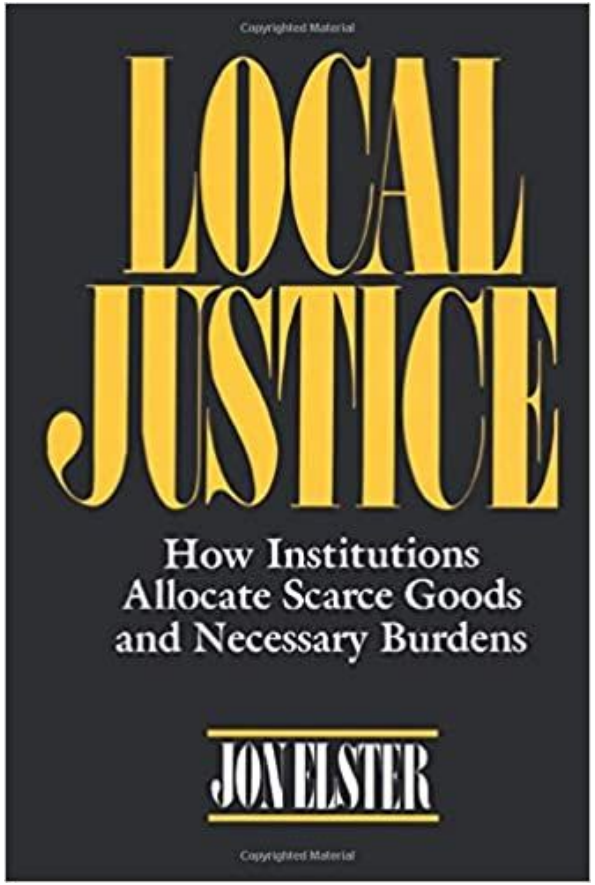
Berkeley CS 294:
Fairness in machine learning

Princeton COS 597E:
Fairness in machine learning

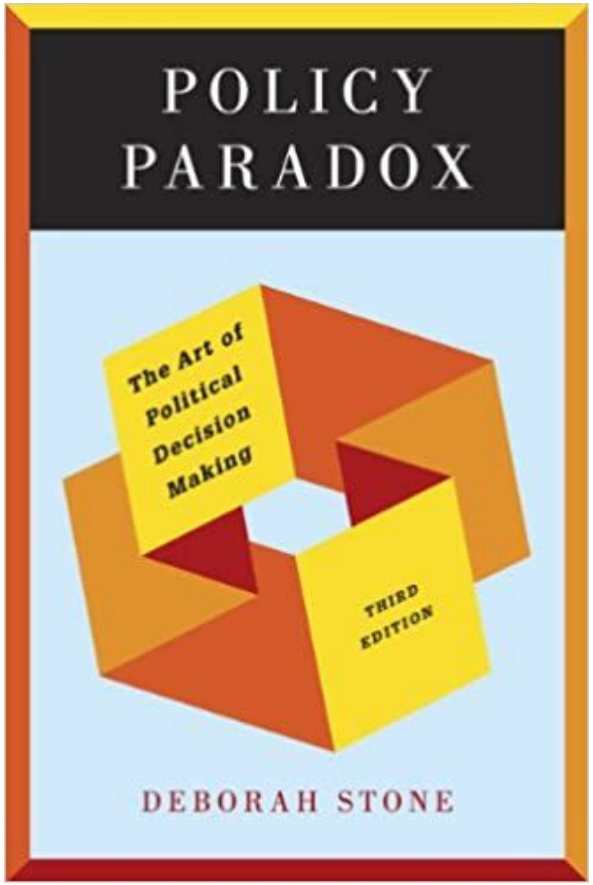


Cornell INFO 4270:
Ethics and policy in data science

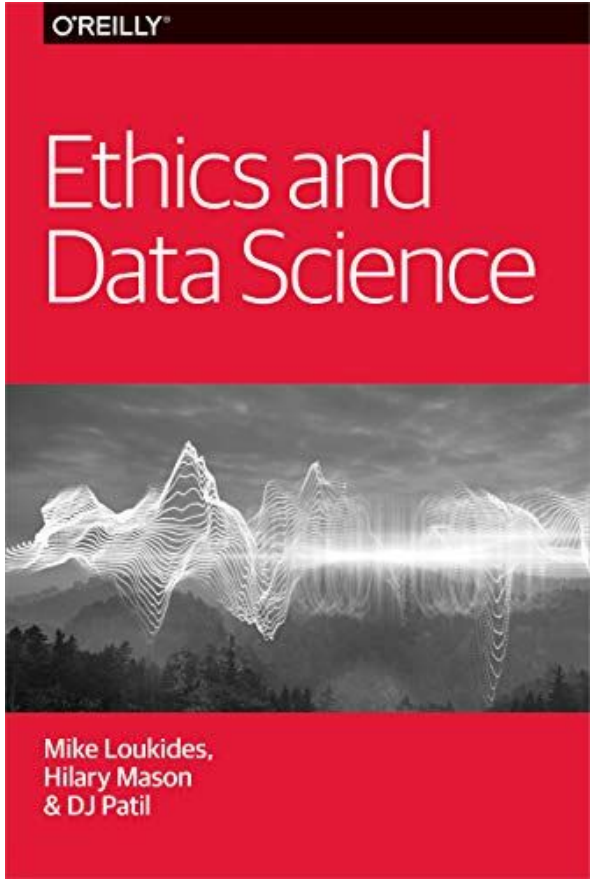
CMU 10718/94889:
ML for Public Policy Lab



Amazon



Amazon



Amazon (Free)

https://www.solveforgood.org/

Thank You!

rayid@cmu.edu
Rayid Ghani

krodolfa@cmu.edu
Kit Rodolfa

pedro.saleiro@feedzai.com
Pedro Saliero