Trustworthy AI Systems

-- Fairness of AI

Instructor: Guangjing Wang

guangjingwang@usf.edu

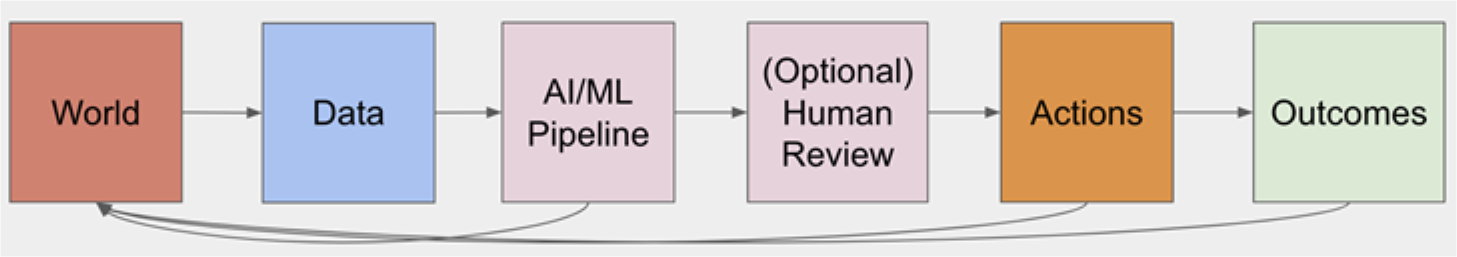
Last Lecture

- Uncertainty and Robustness
- Source of Uncertainty
- Measure the Quality of Uncertainty
- Reduce Uncertainty and Enhance Robustness

This Lecture

- Bias in Data Sources
- Bias Measures
- Fairness Tree
- Hands-on Tutorial

Case Study: Loans



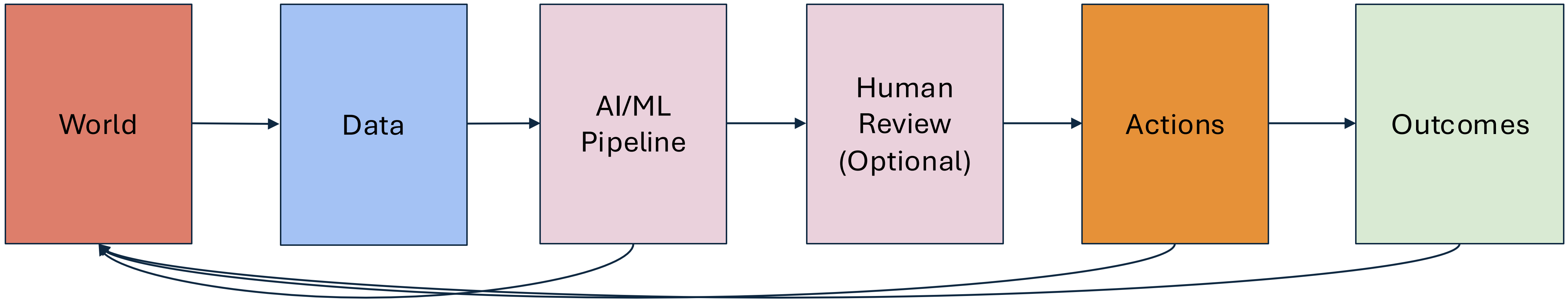
Goal: Provide loans while balancing repayment rates for bank loans

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

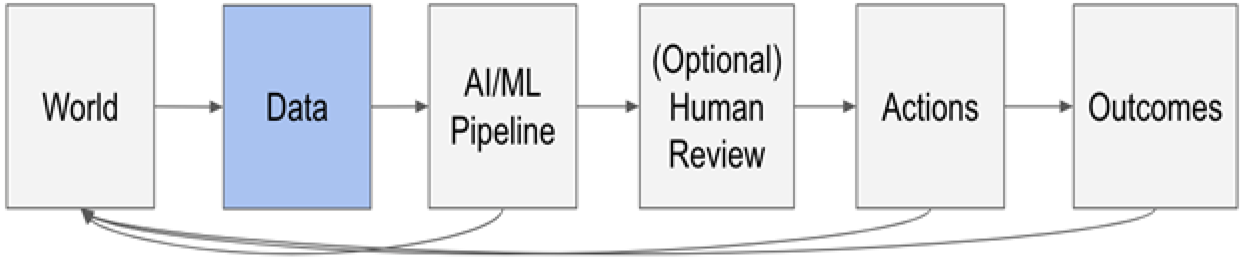
Bias in AI Systems



There are (unfortunately) many sources of bias

Bias in Data Sources

- Choice of Data Sources
- Sample Bias
- Measurement Bias
- 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.

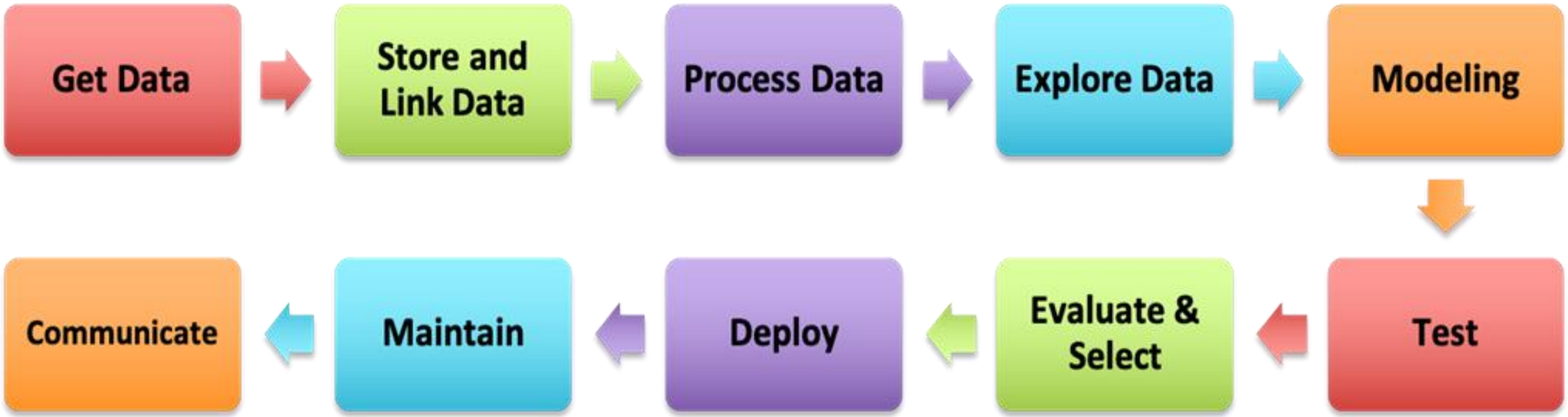
Bias in Data Sources: Label Bias

- 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 the target variable within the majority group but negatively with other groups.

Bias Can Be Introduced in Every Step



Bias Measures

		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 $= \frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}$
		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 = $\frac{1}{\frac{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}{2}}$
		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−) $= \frac{\text{FNR}}{\text{TNR}}$		

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

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.

Why Audit ML models for Bias

"If you don't measure it, you can't improve it."

- Creating awareness among stakeholders helps promote bias and fairness as the main KPI.
- By measuring it, we can improve the system and evaluate bias mitigation approaches.

Aequitas

Bias and Fairness Audit Toolkit

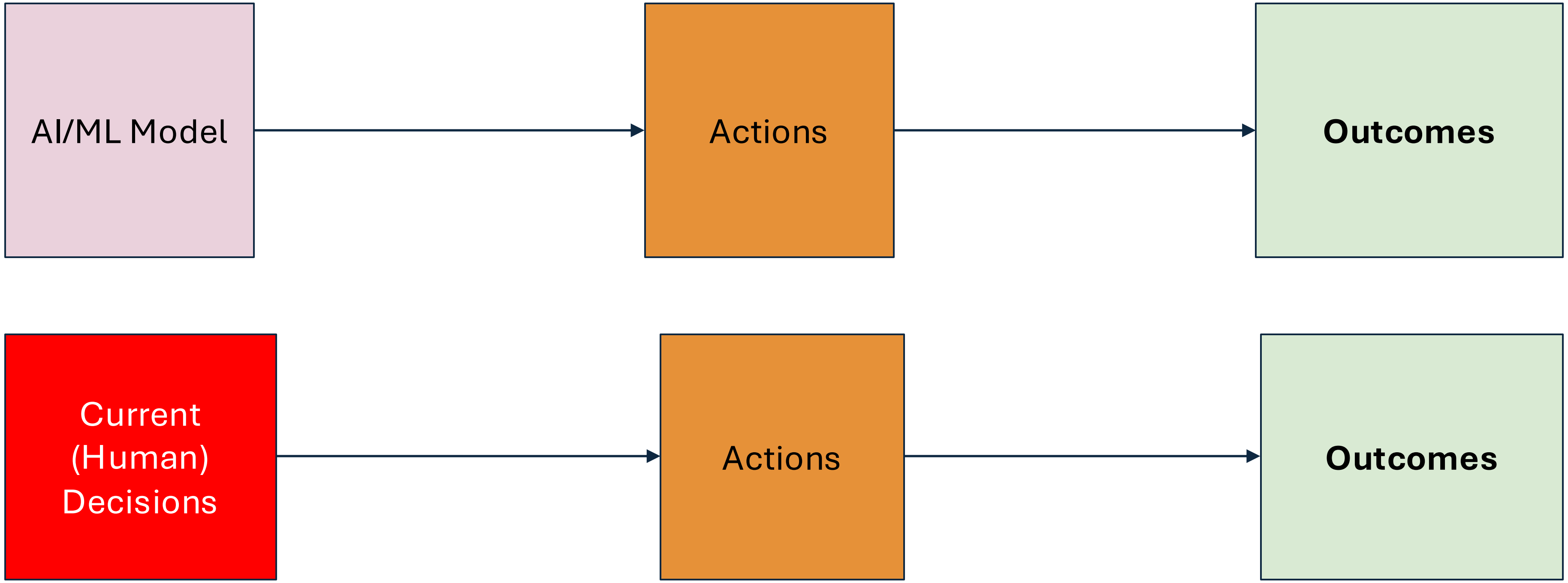
http://www.datasciencepublicpolicy.org/aequitas/

How can we reduce bias in ML models?

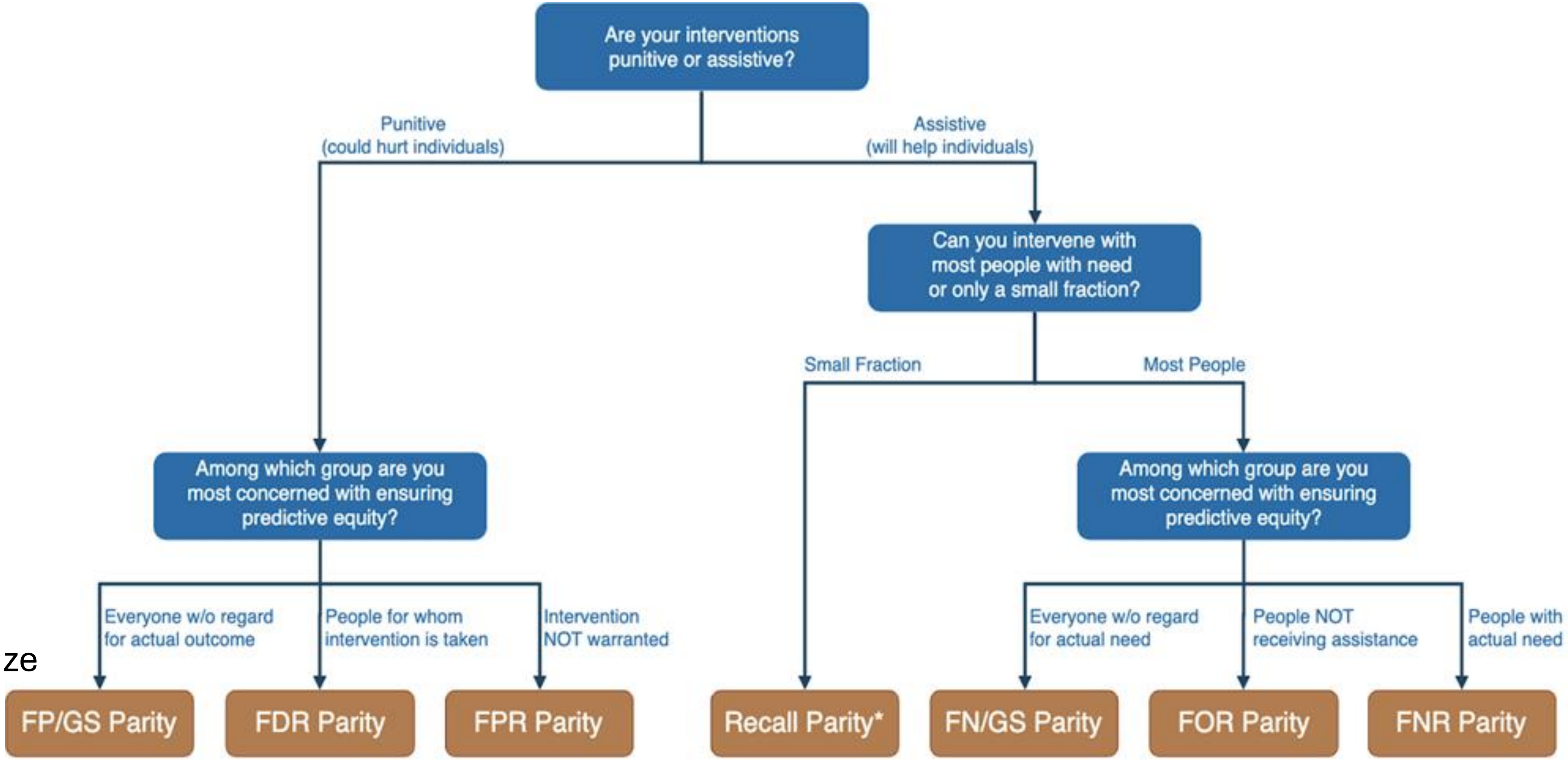
- Fix the world
- Fix the input data
  - ~~Remove sensitive attributes~~
  - Resample and/or reweight protected groups
- **Choose fair models during model selection**
- Optimize for fairness in model training
- Post-hoc adjustments to de-bias model scores

Fairness in AI Systems

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



Fairness Tree



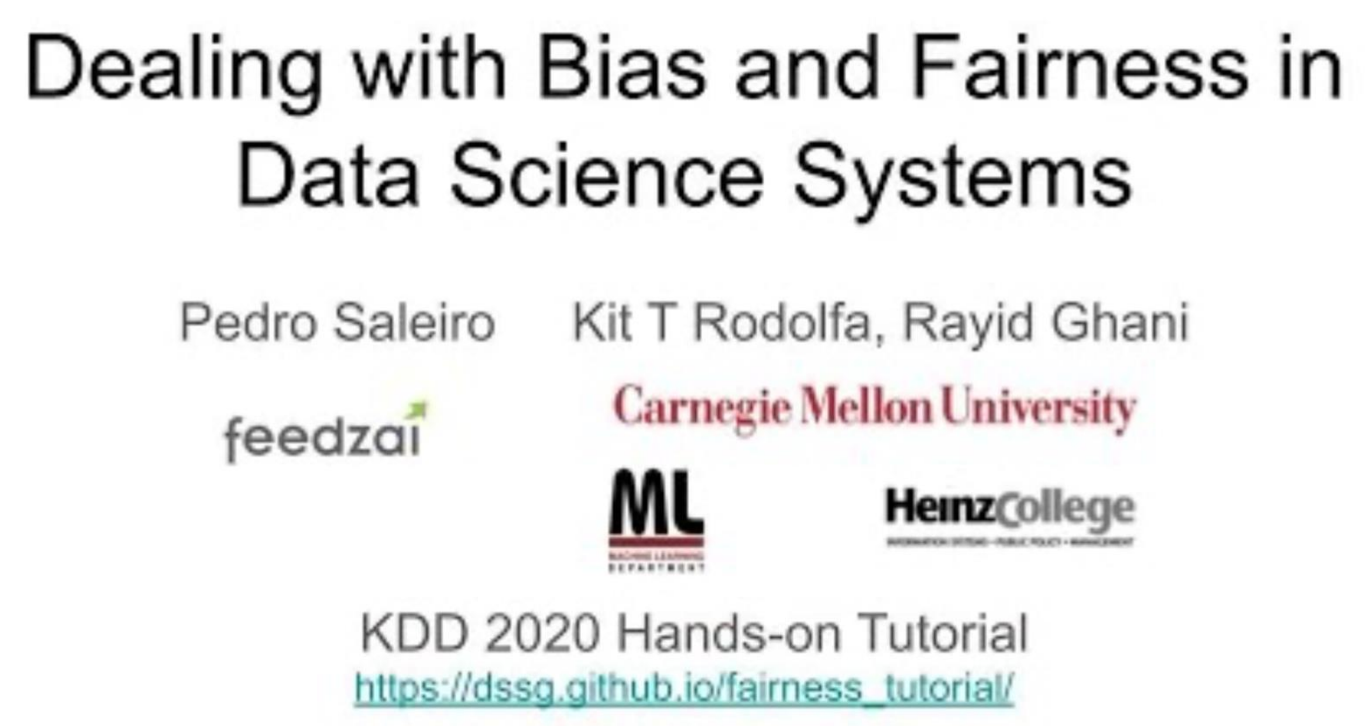
GS: group size

Is the fairness tree “the answer”?

No… but it is 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.

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



https://www.youtube.com/watch?v=N67pE1AF5cM

References

- https://dssg.github.io/fairness_tutorial
- http://github.com/dssg/fairness_tutorial
- https://dssg.github.io/fairness_tutorial/notebooks/