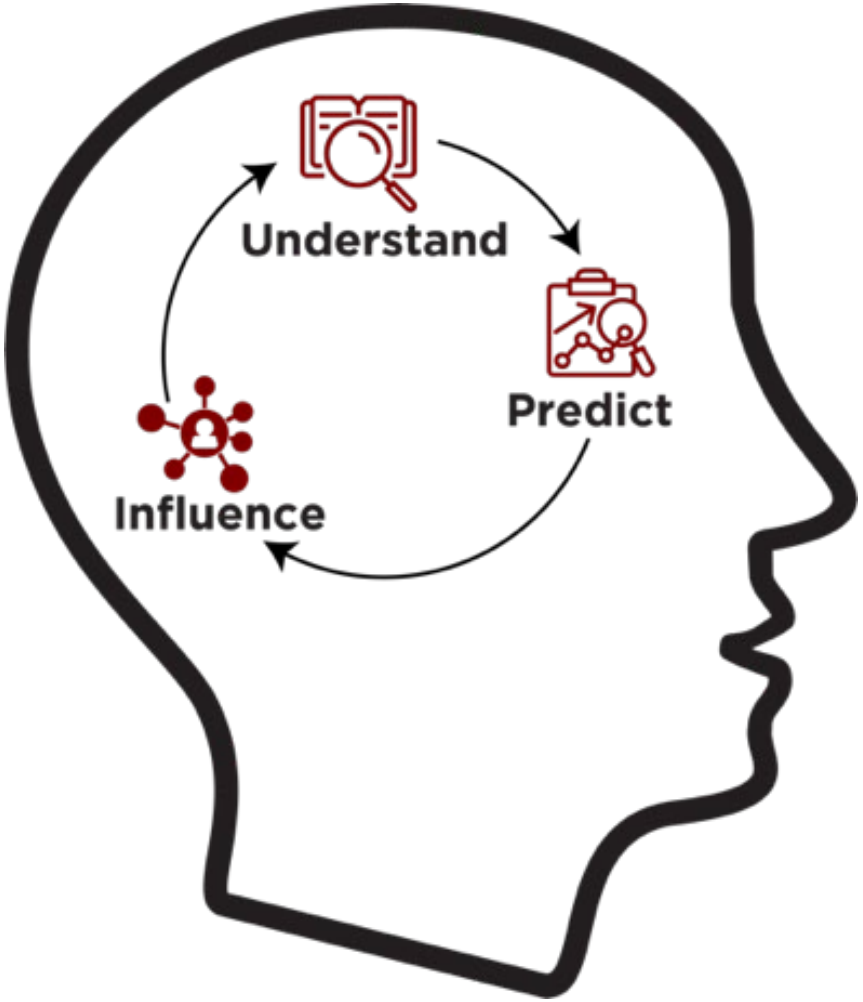
Practical Lessons and Challenges in Building Fair and Equitable Suicide Risk Assessment Systems

Rayid Ghani

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

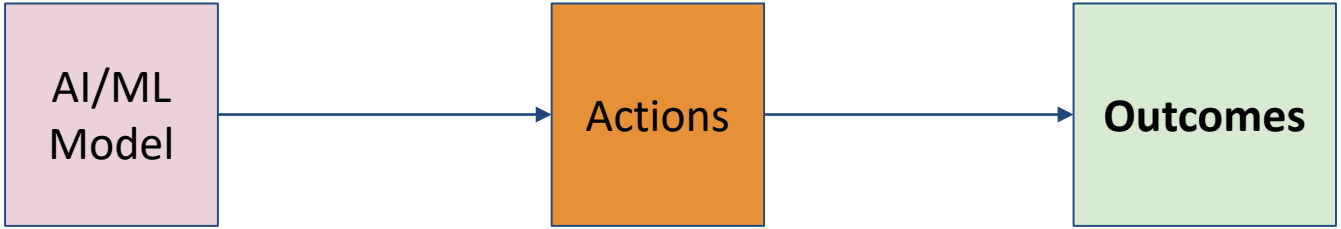

Understand
Predict
Influence

Human-Machine Collaborative Systems

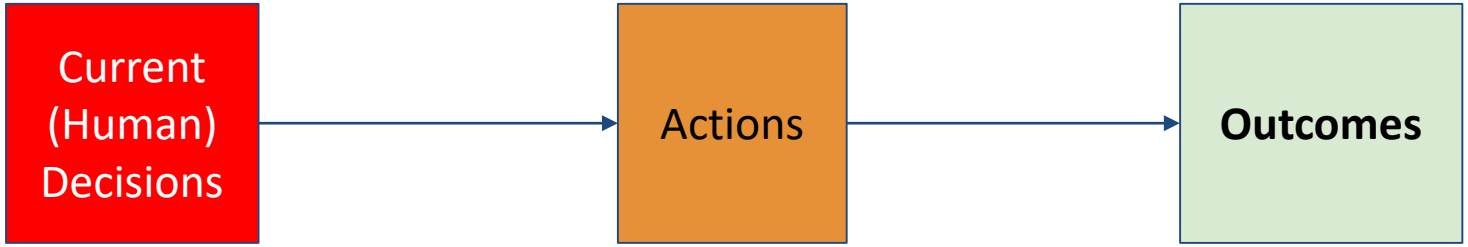
Allocation of Limited Resources

Balancing goals of equity, efficiency, and effectiveness

The goal is not to make the data unbiased or ML/AI 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?

We need to understand and discuss…

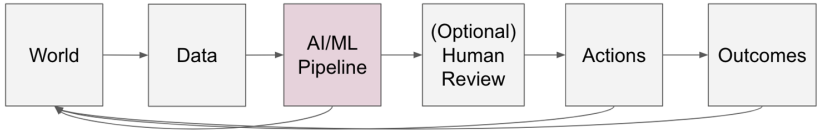
- **Where does bias come from?**
- How do we determine what type of bias(es) to care about?
- How can we detect the bias(es)?
- How can we reduce the bias(es)?

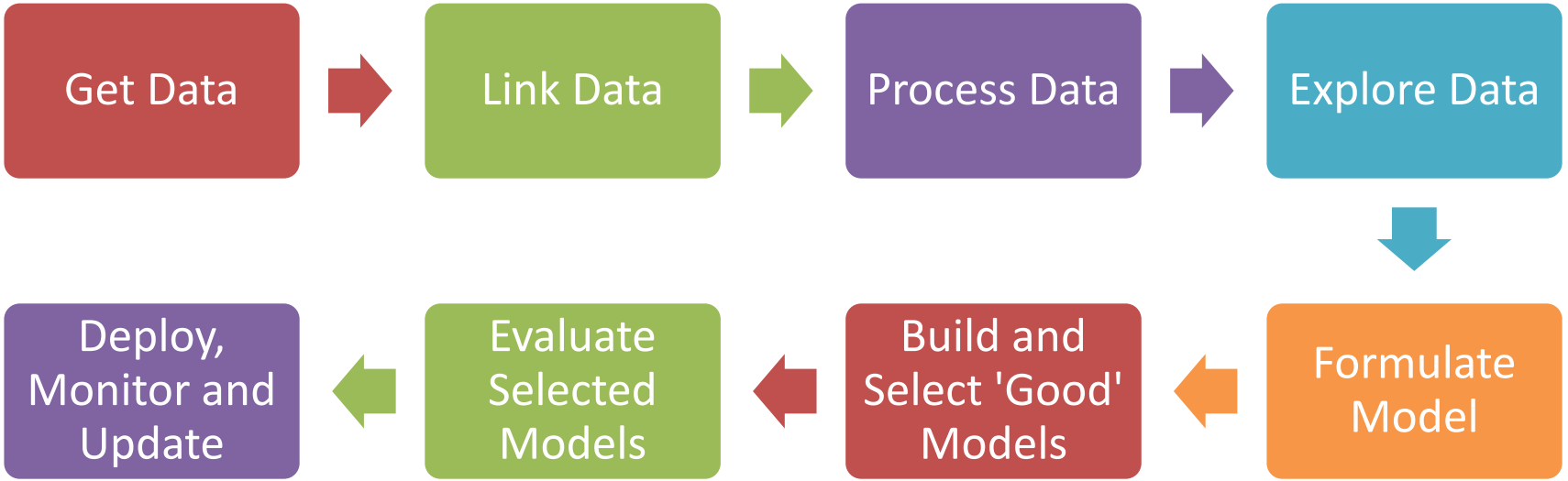
There are (unfortunately) many sources of bias

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

Choice of Data
Sample Bias
Measurement Bias
Label Bias

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



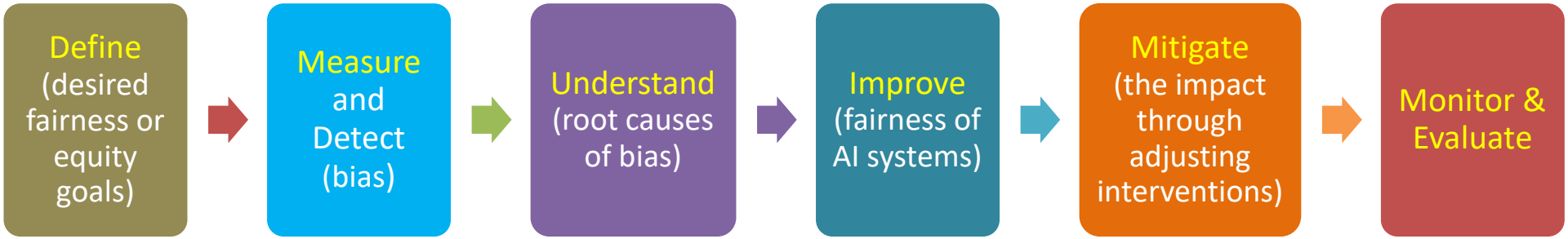


There are (unfortunately) many sources of bias

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

Choice of Data
Sample Bias
Measurement Bias
Label Bias

How do we make the overall system and outcomes fair ?



We need to understand and discuss…

- Where does bias come from?
- **How do we determine what type of bias to care about?**
- How can we detect it?
- How can we reduce it?
- Wrap-up and Practical Tips

Types of Biases

- Based on who is selected for an intervention
  - Only/disproportionately selecting people from a certain background/age/race/gender/…

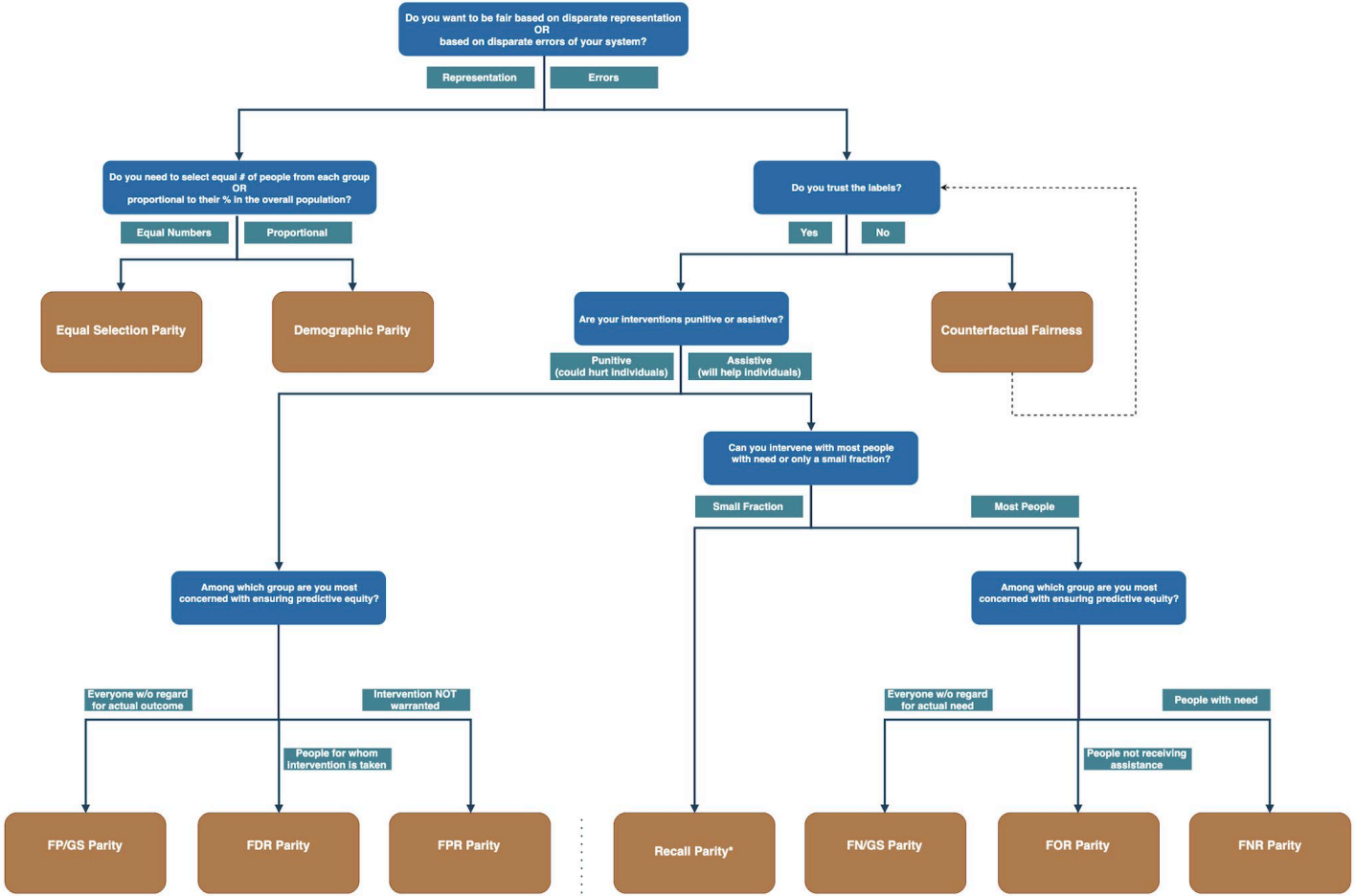
- Based on the types of mistakes in the selection/allocation
  - Selecting people who are not at risk
  - Missing people who are at risk

Many Bias Measures: How do we select what we design for?

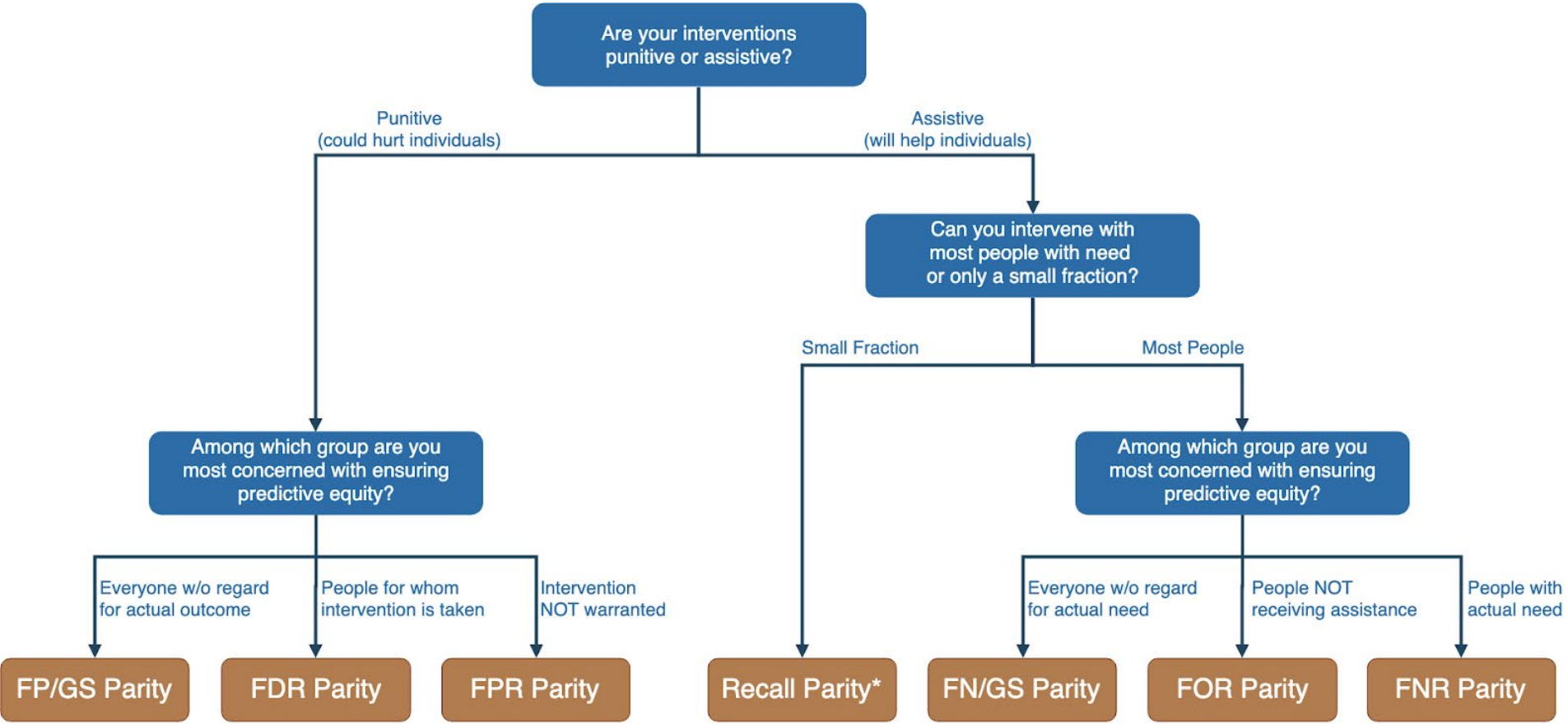
- Do we allocate equal resources to every group?
- Do we allocate resources proportional to need?
- Do we make sure we don't miss people with of a certain group disproportionately?
- Do we make sure we don't have disproportionate false positives from a particular group?
- ...

- Statistical/Demographic Parity
- Impact Parity
- False Discovery Rate Parity
- False Omission Rate Parity
- False Positive Rate Parity
- False Negative Rate Parity
- ...

Fairness Tree



Zoomed in Version



We need to understand and discuss…

- Where does bias come from?
- How do we determine what type of bias to care about?
- **How can we detect it?**
- How can we reduce it?
- Wrap-up and Practical Tips

Aequitas

Open Source Bias & Fairness Audit Tool

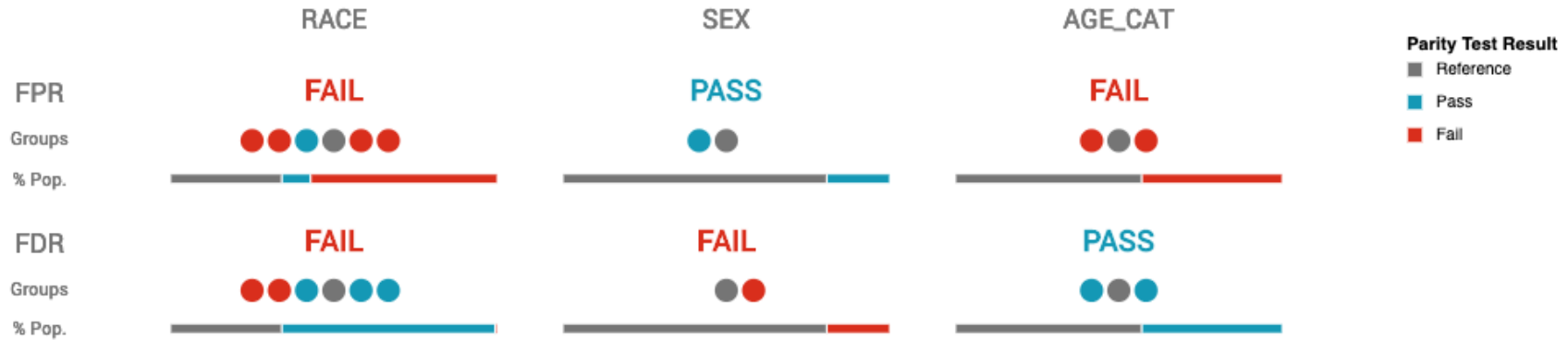
http://www.datasciencepublicpolicy.org/aequitas/

Aequitas: Bias Audit Tool
http://datasciencepublicpolicy.org/aequitas

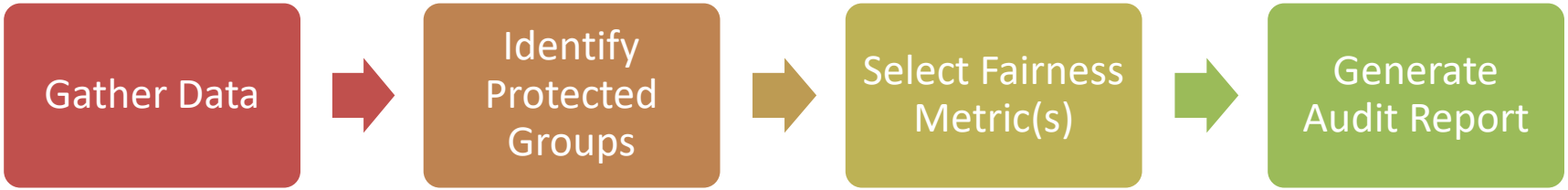
Bias and Fairness Audit Report

Generated by Aequitas for [Large US City] Criminal Justice Project
January 29, 2018

Project Goal: Identify individuals likely to get booked/charged by police in the near future
Performance Metric: Accuracy (Precision) in the top 150 identified individuals
Bias Metrics Considered: Demographic Disparity, Impact Disparity, FPR Disparity, FNR Disparity, FOR Disparity, FDR Disparity
Reference Groups: Race/Ethnicity – White, Gender: Male, Age: None



Bias Audit Flow



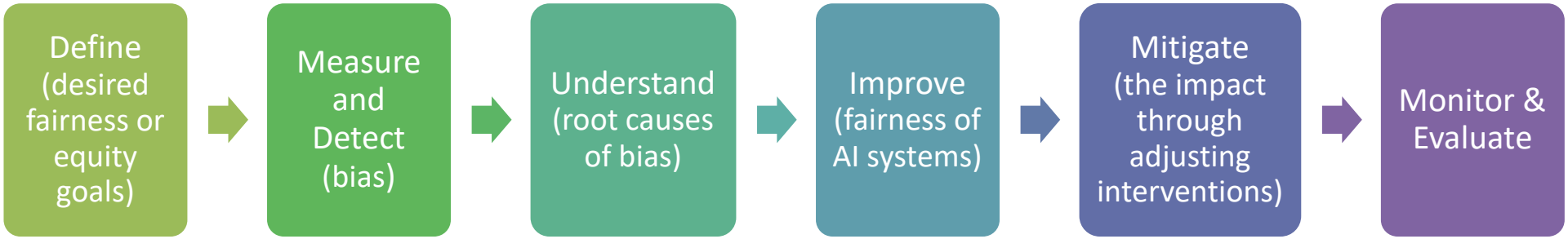
What data do we need to audit the predictions of an AI model?

1. **Predictions (or classifications)**
2. **Attributes that define protected groups** (e.g. race, sex, age)
3. **(True) Labels/Outcomes** (if interested in disparate errors)

Outline for this module

- Where does bias come from?
- How do we determine what type of bias to care about?
- How can we detect it?
- **How can we reduce it?**
- Wrap-up and Practical Tips

Embedding Fairness in the Entire Process



How can we reduce bias in ML models?

- Fix the world
- Fix the input data
- Fix the AI Pipeline
- (Post-hoc) Fix model predictions

Some common practitioner perceptions

- Not using race in my models makes by models not racist
- Using race in my models makes my models racist
- Bias comes from and can be fixed by “fixing” the data
- 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
- A fair ML model = Fair and equitable outcomes

Summary

- Machine Learning and AI are giving us ways to design more “personalized” risk assessment systems that are more effective and more efficient, and have the potential to be more equitable

- Fairness and Equity need to be treated as primary goals/metrics in ML systems and treated as an integral part of every project: Scoping, community and stakeholder engagement, metrics, validation, monitoring outcomes

- Our focus should not just be on making the data unbiased and the ML model fair but rather on making the overall system and outcomes fair

- Dealing with fairness in data science systems is a new and rapidly changing area and practitioners need to be careful about methods and tools that may not have been fully validated

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

ML MACHINE LEARNING DEPARTMENT

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

rayid@cmu.edu