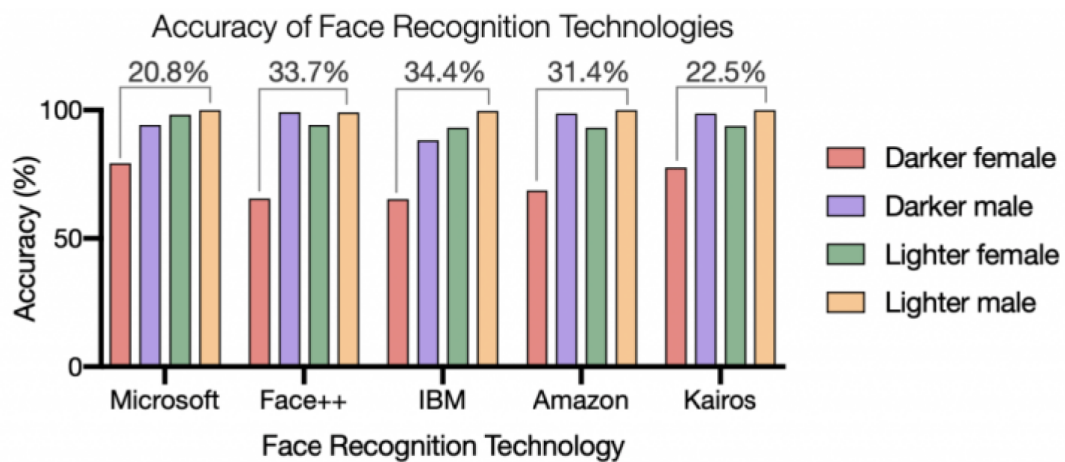# Fairness and Bias
MIE223
Winter 2025

## 1 Introduction to Fairness and Bias

### 1.1 MIT Study on Racial/Gender Discrimination in Image Classifiers

- MIT Project known as Gender Shades

  - Examined facial recognition algorithms from Microsoft and IBM:
  - Did poorly on underrepresented groups (e.g. black females)



### 1.2 Data is Full of Biases

Representation and Collection Bias

- The data is not representative of the population distribution it is intended to serve

  - Critical in medicine

- The data has insufficient representation of minority groups

Measurement and Historical Bias

- What you measure is misaligned with use case, or biased from historical data

  - E.g., using historical loan decisions to inform future decisions

Aggregation bias

- Effect for a group can be reversed when looking at data in aggregate
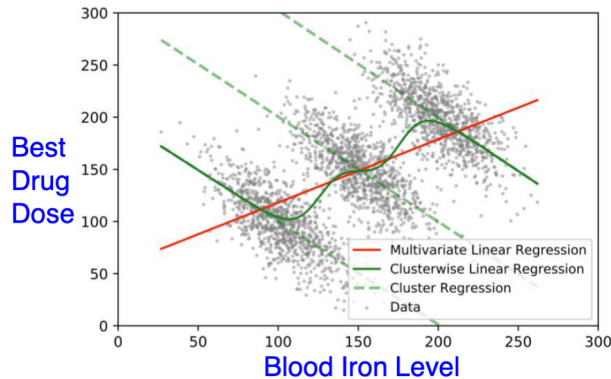
This data is multimodal. Gaussian means unimodal, so be careful applying unimodal tools to multimodal data. If you ignore modes, you get the red line rather than the green line.

## 1.3   Group and Individual Fairness

Much research on fairness aims to ensure "fair treatment" = parity
Group Fairness

- Achieve "parity" (equality) across protected and advantaged groups

- Groups usually determined by demographic attribute

  - e.g., Caucasian, non-Caucasian, Male, non-Male

Individual Fairness

- Similar individuals should have similar predictions / outcomes

- Requires measure of similarity between individuals

  - e.g., Euclidean or cosine similarity between demographic feature vectors

## 1.4   Group and Individual Fairness for Classification

Classification: predicting a discrete label, e.g., parole, no parole

- Group Fairness:

  - Fairness in Treatment: should not consider sensitive/protected attribute (e.g., gender, race, age)
    * But other attributes may be correlated with sensitive ones!
  - Fairness in Impact: classification outcome should be balanced across groups
    * E.g., false positive rate for each group should be the same

- Individual Fairness:

  - Enforce similar individuals to receive similar outputs (or distributions)
    * Requires a similarity function over sensitive attributes

## 1.5 On the chalkboard

Given your grades (vectorized) in previous courses, what letter grade will you get in this course? Your past history helps predict your future performance. You can regress directly to the grade, whereas classification goes to letter grades A,B,C,D,F Focusing on binary classification P, NP:

- True Positive (TP): predicted positive and actually positive

- True Negative (TN): predicted negative and actually negative

- False Positive (FP): predicted positive and actually negative

- False Negative (FN): predicted negative and actually positive

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \tag{3}$$

$$\text{False Discovery Rate (FDR)} = \frac{FP}{TP + FP} \tag{4}$$

$$\text{True Negative Rate (TNR)} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{False Negative Rate (FNR)} = \frac{FN}{TP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

$$\text{F2 Score} = (1 + 2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + 2 \cdot \text{Recall}} \tag{10}$$

$$\text{F0.5 Score} = (1 + 0.5) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + 0.5 \cdot \text{Recall}} \tag{11}$$

Contingency table is a 2x2 table that summarizes the performance of a classification model Multiclass classification uses a confusion matrix

## 1.6 Achieving Parity

What types of parity in binary classification can we aim for?



What types of parity can we aim for in impact fairness?

- Demographic / statistical parity (predicted outcome only)
    - Different groups get same positive classification rates
        * E.g., same admission rate regardless of economic status
- Parity in errors (consider predicted and actual outcome)
    - Different groups have different classification rates
    - But we constrain the error rates (Accuracy, TPR, FPR, FDR, ...)

# 2 Can we have all desired Parities at once?

No

## 2.1 Incompatibility Between Fairness Metrics

| Metric | Caucasian | African American |
|---|---|---|
| False Positive Rate ($FPR$) | 23% | 45% |
| False Negative Rate ($FNR$) | 48% | 28% |
| False Discovery Rate ($FDR$) | 41% | 37% |

ProPublica identified considerable disparities:

- The algorithm was twice as likely to incorrectly predict black individuals were at high risk of recidivating (FPR)

- It was also nearly twice as likely to incorrectly predict white individuals were at low risk of recidivating (FNR)

However, the creator of the algorithm pointed out that the algorithm is well- balanced across races on precision (equivalently, FDR), claiming this is the correct measure of fairness in this context.

- Who is right?

- Is the COMPAS algorithm biased?

- Can't the algorithm achieve both measures of fairness at the same time?



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

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

**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

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

If prevalence is unequal across groups...

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

If prevalence is unequal across groups...

...and FDR (or precision) is equal across groups...

$$FPR = \frac{p}{1-p}\left(\frac{FDR}{1-FDR}\right)(1-FNR)$$
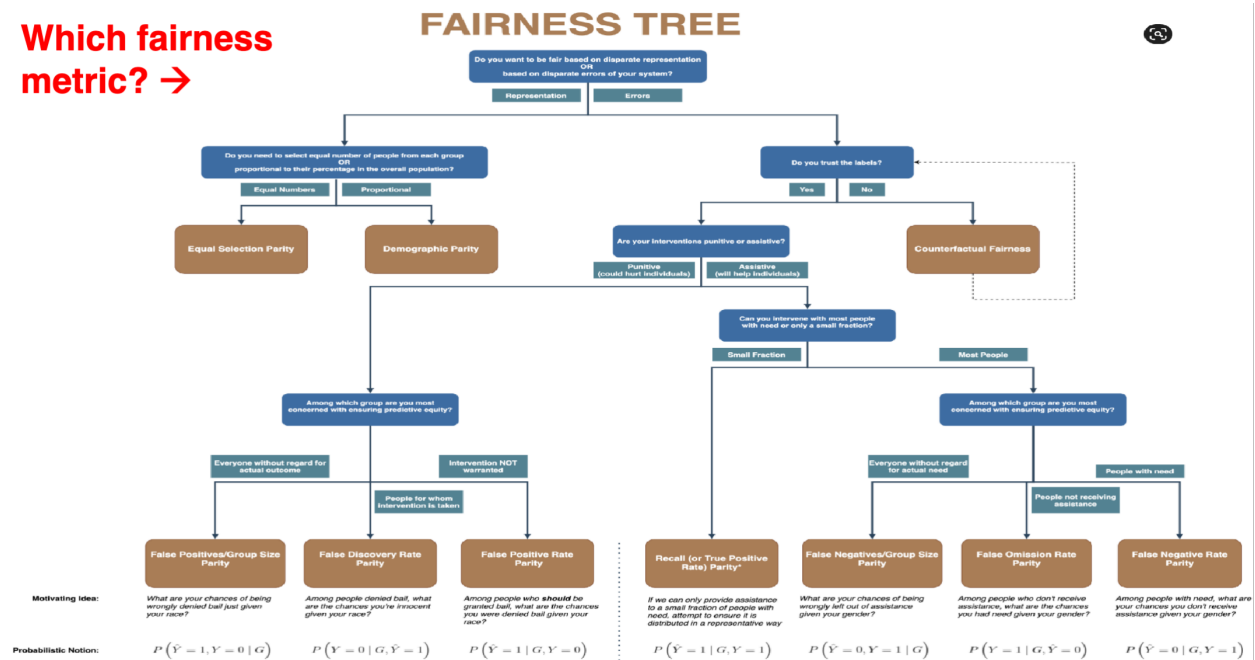
If prevalence is unequal across groups...

...and FDR (or precision) is equal across groups...

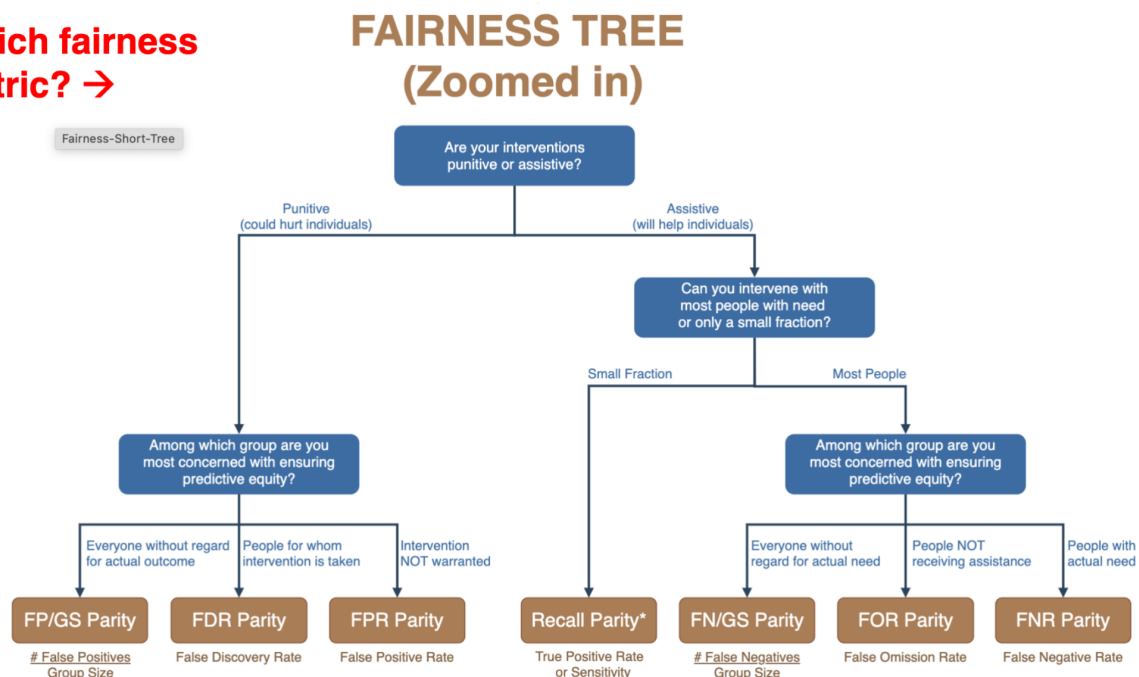…then either FPR or FNR can be equal across groups, **but not both**

## 2.2 So... what does this mean?

Sadly, no... we can't have it all! But this also doesn't mean fairness is unachievable or we shouldn't bother. Instead, it means we have to carefully consider and choose the appropriate fairness metrics for our context

**Which fairness metric? →**

**FAIRNESS TREE**

Do you want to be fair based on disparate representation OR based on disparate errors of your system?

- Representation
- Errors

Do you need to select equal number of people from each group OR proportional to their percentage in the overall population?
- Equal Numbers → **Equal Selection Parity**
- Proportional → **Demographic Parity**

Do you trust the labels?
- Yes
- No → **Counterfactual Fairness**

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 without regard for actual outcome
- Intervention NOT warranted
- People for whom intervention is taken

Among which group are you most concerned with ensuring predictive equity?
- Everyone without regard for actual need
- People not receiving assistance
- People with need

| | False Positives/Group Size Parity | False Discovery Rate Parity | False Positive Rate Parity | Recall (or True Positive Rate) Parity* | False Negatives/Group Size Parity | False Omission Rate Parity | False Negative Rate Parity |
|---|---|---|---|---|---|---|---|
| Motivating Idea: | What are your chances of being wrongly denied bail just given your race? | Among people denied bail, what are the chances you're innocent given your race? | Among people who **should** be granted bail, what are the chances you were denied bail given your race? | If we can only provide assistance to a small fraction of people with need, attempt to ensure it is distributed in a representative way | What are your chances of being wrongly left out of assistance given your gender? | Among people who don't receive assistance, what are the chances you had need given your gender? | Among people with need, what are your chances you don't receive assistance given your gender? |
| Probabilistic Notion: | $P\left(\hat{Y}=1, Y=0 \mid G\right)$ | $P\left(Y=0 \mid G, \hat{Y}=1\right)$ | $P\left(\hat{Y}=1 \mid G, Y=0\right)$ | $P\left(\hat{Y}=1 \mid G, Y=1\right)$ | $P\left(\hat{Y}=0, Y=1 \mid G\right)$ | $P\left(Y=1 \mid G, \hat{Y}=0\right)$ | $P\left(\hat{Y}=0 \mid G, Y=1\right)$ |

**Which fairness metric? →**

**FAIRNESS TREE (Zoomed in)**

Fairness-Short-Tree

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 without regard for actual outcome → **FP/GS Parity** (# False Positives / Group Size)
- People for whom intervention is taken → **FDR Parity** (False Discovery Rate)
- Intervention NOT warranted → **FPR Parity** (False Positive Rate)

- Small Fraction → **Recall Parity*** (True Positive Rate or Sensitivity)

Among which group are you most concerned with ensuring predictive equity?
- Everyone without regard for actual need → **FN/GS Parity** (# False Negatives / Group Size)
- People NOT receiving assistance → **FOR Parity** (False Omission Rate)
- People with actual need → **FNR Parity** (False Negative Rate)

# 3 Working with Data Collected from Humans (e.g., Surveys)

Something you will likely do in your career Beware of many biases that can occur! (with potentially global impact: 2016 US election

## 3.1 Survey Methodology and Bias Prevention

- Survey Sampling and Bias:

- Question Design and Cognitive Biases: