Types of Fairness, An Incomplete List

- Unawareness
- Group prediction parity
- Group error parity
- Individual similarity-based fairness
- Individual counterfactual fairness
- Envy-free fairness

Definitions

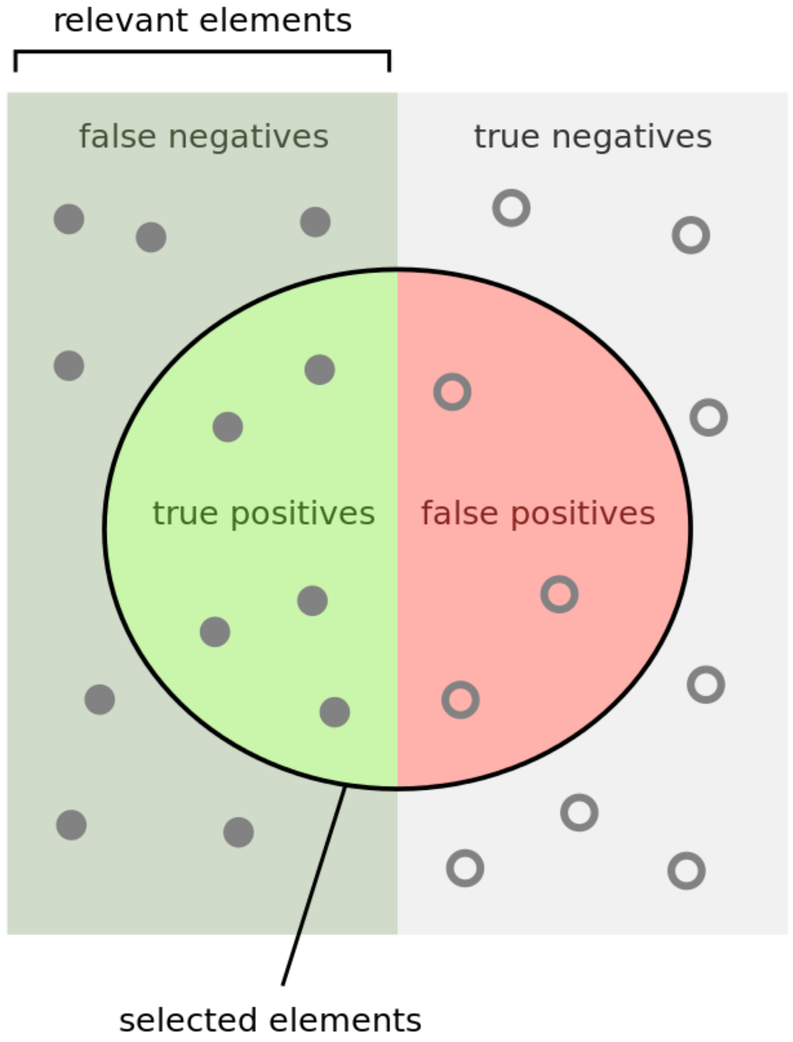
- Data $X = \{x_1, \ldots, x_n\}$
- Target $Y = \{y_1, \ldots, y_n\}$
- Predictor f(x) outputs a guess of y
- Types of accuracy measures:
  - (Raw) accuracy
  - Precision, recall, (sensitivity, specificity)

Accuracy Types

$$\text{Accuracy} \quad \frac{TP + TN}{TP + TN + FP + FN} \quad P(f(x) = y)$$

$$\text{Precision} \quad \frac{TP}{TP + FP} \quad P(y = T \mid f(x) = T)$$

$$\text{Recall} \quad \frac{TP}{TP + FN} \quad P(f(x) = T \mid y = T)$$



Group Fairness vs. Individual Fairness

- Group fairness considers statistics across subpopulations
  - Usually requires defining sensitive features or groups
- Individual fairness applies to individual people
  - Usually stronger requirements to satisfy

Unawareness

- Data $X = \{x_1, \ldots, x_n\}$
- Target $Y = \{y_1, \ldots, y_n\}$
- Sensitive feature $S = \{s_1, \ldots, s_n\}$
- Concern that f(x, s) would use s, so only train f(x)
- Usually fails because some features in x are correlated with s
  http://www.justicemap.org

Group Prediction Parity

- Treat two sub-populations the same
  - Learn f(x, s) such that $E_{s=1}[f(x, s)] \approx E_{s=0}[f(x, s)]$
  - Prediction probability has similar statistics for groups with or without sensitive feature

Group Error Parity

- Treat two sub-populations equally well
  - Learn f(x, s) such that $E_{s=1}[\text{error}(f(x, s), y)] \approx E_{s=0}[\text{error}(f(x, s), y)]$
- Prediction error is *independent* of sensitive feature s
- Defining error as **true-positive rate**, we get equal opportunity
  - Individuals who deserve loans are equally likely to be offered

Forms of Error

- Positive = good result (get loan, accepted to college, etc.)
- Equalized odds: f(x) and s are independent given y
  - $P(f(x) \mid s = T, y = T) = P(f(x) \mid s = F, y = T)$ and $P(f(x) \mid s = T, y = F) = P(f(x) \mid s = F, y = F)$
  - Recall and false-positive rate equal across groups
- Equal opportunity: f(x) and s are independent given y = true
  - Recall is equal across groups

Error Parity Discussion

- What if we equalize precision?
- What if we equalize accuracy?
- What if training labels are biased?

Individual Similarity-Based Fairness

- Treat all similar individual equitably
  - $\| f(x_a) - f(x_b) \| < S(x_a, x_b)$, for some **similarity metric** S
  - Metric determines what makes individuals similar
  - Can also be extended to error-based similarity fairness
    $\| error(x_a) - error(x_b) \| < S(x_a, x_b)$

Individual Counterfactual Fairness

- Treat **each** individual the same regardless of sensitive features
  - Learn $f(x, s)$ such that $f(x, s = 0) \approx f(x, s = 1)$
- Prediction probability is *independent* of sensitive feature s for each individual

Envy-Free Fairness

- In resource allocation, an envy-free assignment is one where each individual would not prefer to receive the assignment of another
  - E.g., cake cutting, chore assignments, ad allocation

Types of Fairness, An Incomplete List

- Unawareness
- Group prediction parity
- Group error parity
- Individual similarity-based fairness
- Individual counterfactual fairness
- Envy-free fairness

Causes of Unfairness, An Incomplete List

- ML mimics data from unfair systems
- Definition of ML tasks is unfair
- Underrepresentation of minority groups
- Feedback loops in deployed ML

Data From Unfair Systems

- Academic/professional performance, salary, crime
- Society is working on making these things more fair
- Learning to replicate old data could be a step back

Unfair ML Problem Definitions

- Predicting race, gender, native language, income level, criminality, religion, sexual orientation
- Some of these ideas don’t even have clear definitions
- And they often have little or nothing to do with input data
- ML will happily learn correlations

Unfairness from Underrepresentation

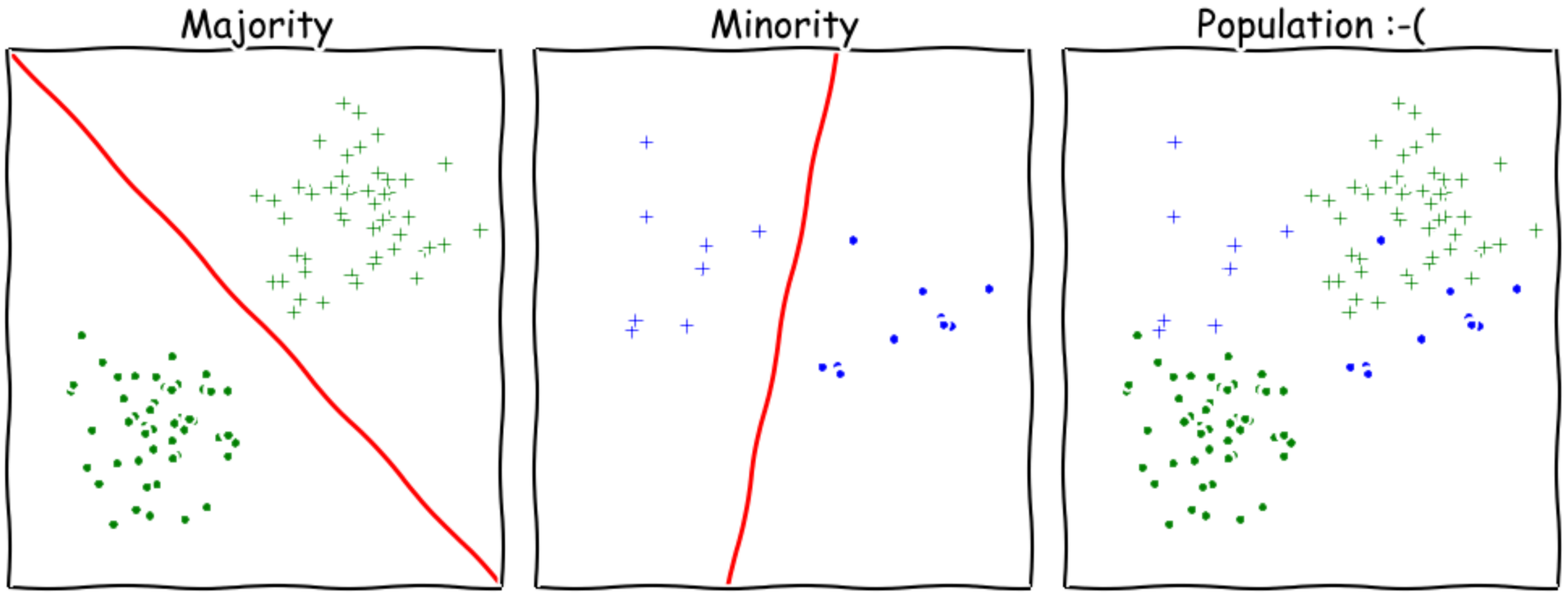
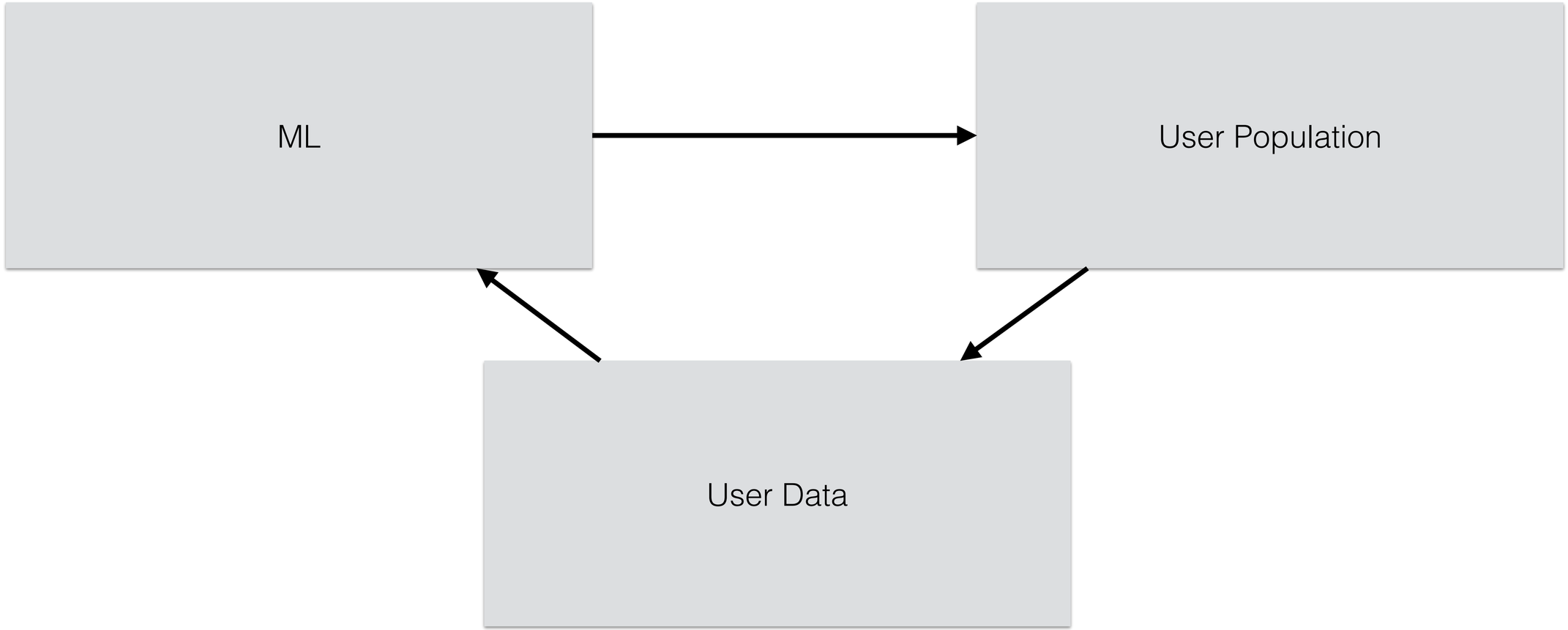


Illustration by Moritz Hardt (https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de)

Feedback Loops



Discussion

- In predictive policing:
  - What types of unfairness do you expect to see?
  - What causes of unfairness do you expect to see?