Fairness in Machine Learning

CS5824/ECE4424
Bert Huang

CNBC Report on ML in Law Enforcement

https://youtu.be/ZMsSc_utZ40

Plan

- Different forms and causes of fairness in machine learning
- Case studies of recent solutions for fairer ML
  - Post processing predictions for equal opportunity
  - Fair representation learning
  - Fixing feedback loops

Types of Fairness, An Incomplete List

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

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

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

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

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

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

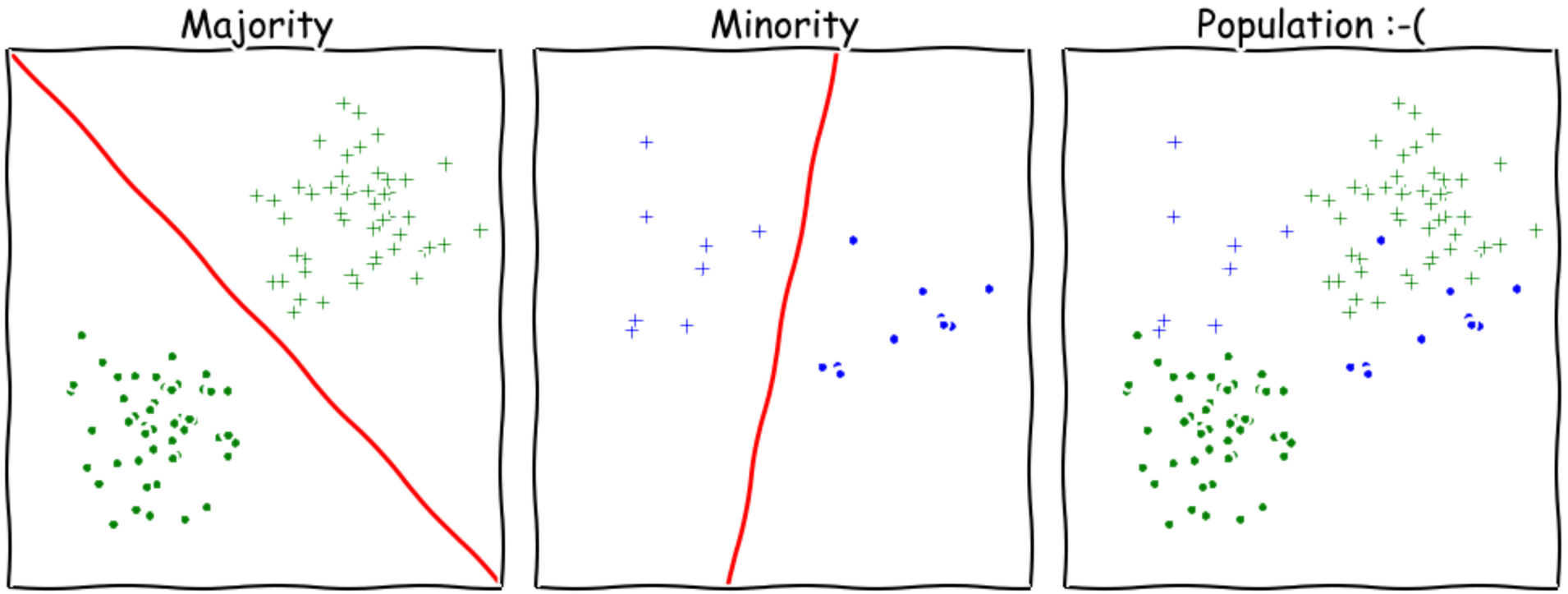
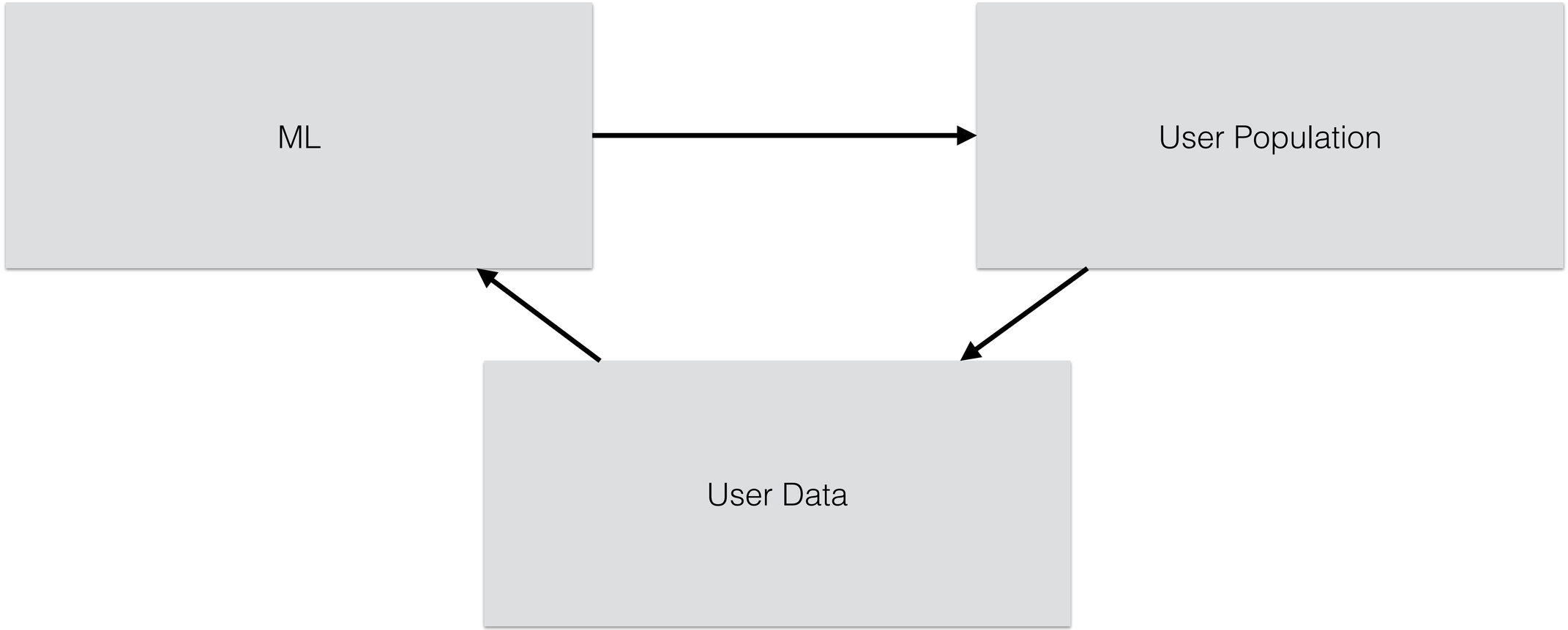


Illustration by Moritz Hardt (https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de)

Feedback Loops



Plan

- Different forms and causes of fairness in machine learning
- Case studies of recent solutions for fairer ML
  - Post processing predictions for equal opportunity
  - Fair representation learning
  - Fixing feedback loops