Bias and Fairness in ML

Bias in ML

Fair ML

Fairness and algorithmic decision making

- ML models are increasingly used in high-stakes decisions
    - Loan applications, hiring, court decisions, predictive policing
- ML systems increase effectiveness and consistency?
- Various forms of (data) biases can be fed into the system
    - Models trained on biased data learn to reproduce biases

> "*In the context of decision-making, fairness is the absence of any prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics.*" (Mehrabi et al. 2019)

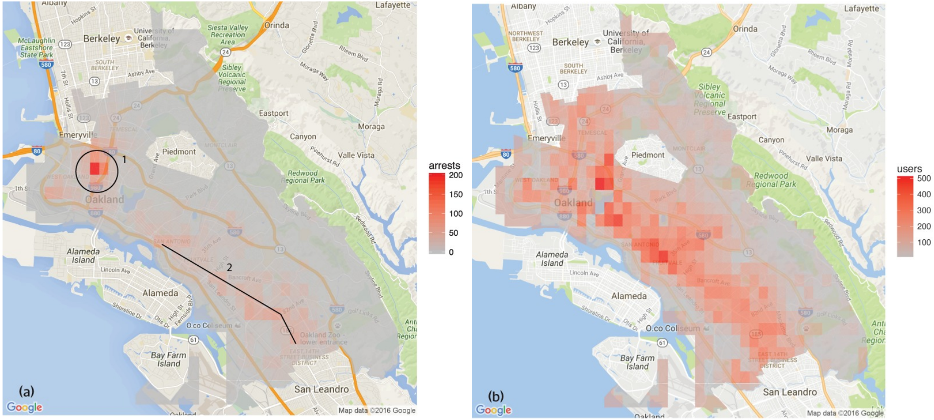
→ Protected attributes in US context: sex, race, age, marital status, ...

Bias in Data

Figure: Predictive policing example (Lum and Isaac 2016)

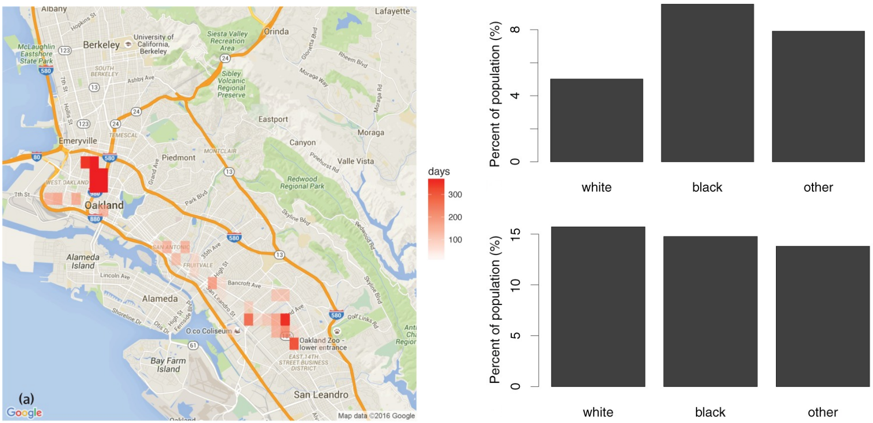
(a) Drug arrests (training data)

(b) Drug crimes



Bias in Data

Figure: Predictive policing example (Lum and Isaac 2016)



(a) Areas targeted by PredPol

(b) Targeted policing and drug use by race

Bias in Data

Types of data biases (Mehrabi et al. 2019)

- Historical (Label) Bias
  - "*Historical bias is a normative concern with the world as it is; it is a fundamental, structural issue with the first step of the data generation process and can exist even given perfect sampling and feature selection*"
- Representation (Sample) Bias
  - Representation bias arises when defining and sampling from a population
- Measurement Bias
  - Measurement bias arises when choosing and measuring the particular features of interest
- Many more...