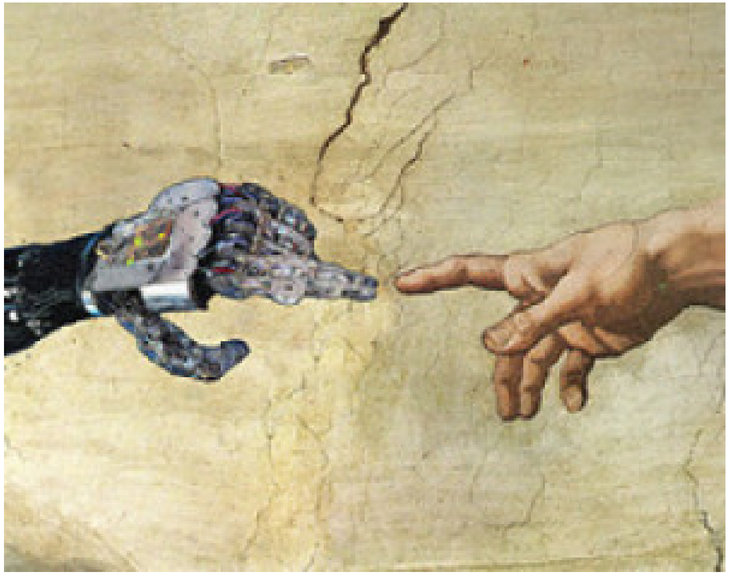
Engineering a Fair Future: Why We Need to Train Unbiased AI



Krishna P. Gummadi
Max Planck Institute for Software Systems

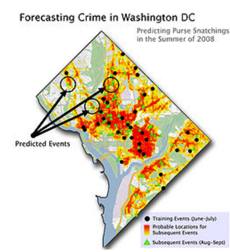
Algorithmic decision making

- Refers to data-driven decision making
  - By learning over data about past decision outcomes
- Increasingly influences every aspect of our life

Search, Recommender, Reputation Algorithms

Match / Market-Making Algorithms

Risk Prediction Algorithms



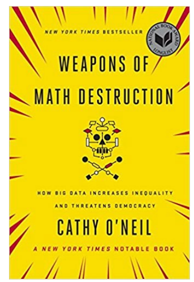
Concerns about their fairness

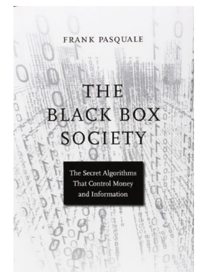
- Discrimination in predictive risk analytics

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

- Opacity of algorithmic (data-driven) decision making





- Implicit biases in

As Germans Seek News, YouTube Delivers Far-Right Tirades

A researcher found the platform's recommendation system had steered viewers to fringe and conspiracy videos on a neo-Nazi demonstration in Chemnitz.

Focus on discrimination

- Discrimination is a specific type of unfairness
- Well-studied in social sciences
  - Political science
  - Moral philosophy
  - Economics
  - Law
    - Majority of countries have anti-discrimination laws
    - Discrimination recognized in several international human rights laws

- But, less-studied from a computational perspective

What is a computational perspective? Why is it needed?

Defining discrimination

- A first approximate normative / moralized definition:

  wrongfully impose a **relative disadvantage** on persons **based on** their membership in some **salient social group** e.g., race or gender

- Challenge: How to operationalize the definition?
  - How to make it clearly distinguishable, measurable, & understandable in terms of empirical observations

Need to operationalize 4 fuzzy notions

1. What constitutes a relative disadvantage?
2. What constitutes a wrongful imposition?
3. What constitutes based on?
4. What constitutes a salient social group?
    1. Defined by **anti-discrimination laws**: Race, Gender

Case study: Recidivism risk prediction

- **COMPAS** recidivism prediction tool
  - Built by a commercial company, Northpointe, Inc.

- Estimates likelihood of criminals re-offending in future
  - Inputs: Based on a long questionnaire
  - Outputs: Used across US by judges and parole officers

- Trained over big historical recidivism data across US
  - Excluding sensitive feature info like gender and race

COMPAS Goal: Criminal justice reform

- Many studies show racial biases in human judgments

- **Idea:** Nudge subjective human decision makers with objective algorithmic predictions
  - Algorithms have no pre-existing biases
  - They simply process information in a consistent manner

- Learn to make accurate predictions without race info.
  - Blacks & whites with same features get same outcomes
  - No disparate treatment & so non-discriminatory!

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: 805 / (805 + 990) = 0.45

349 / (349 + 1139) = 0.23

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: 805 / (805 + 990) = 0.45 349 / (349 + 1139) = 0.23

False Negative Rate: 532 / (532 + 1369) = 0.29 461 / (461 + 505) = 0.48

Is COMPAS non-discriminatory?

	Black Defendants		White Defendants	
	High Risk	Low Risk	High Risk	Low Risk
Recidivated	1369	532	505	461
Stayed Clean	805	990	349	1139

False Positive Rate: 805 / (805 + 990) = 0.45 >> 349 / (349 + 1139) = 0.23

False Negative Rate: 532 / (532 + 1369) = 0.29 << 461 / (461 + 505) = 0.48

- ProPublica: False positive & negative rates are considerably worse for blacks than whites!
  - Constitutes discriminatory **disparate mistreatment**

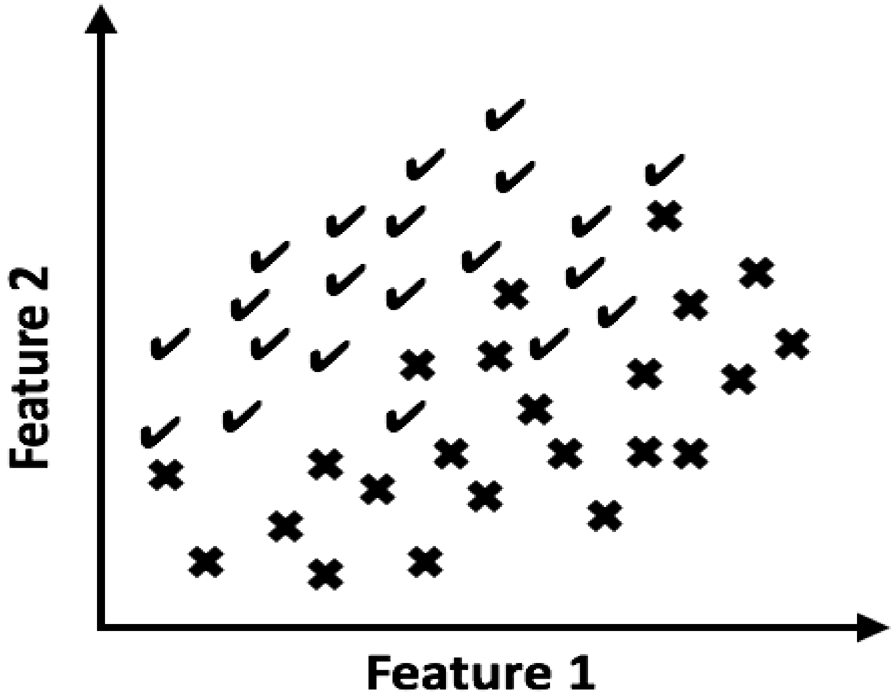
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

COMPASS study raises many questions

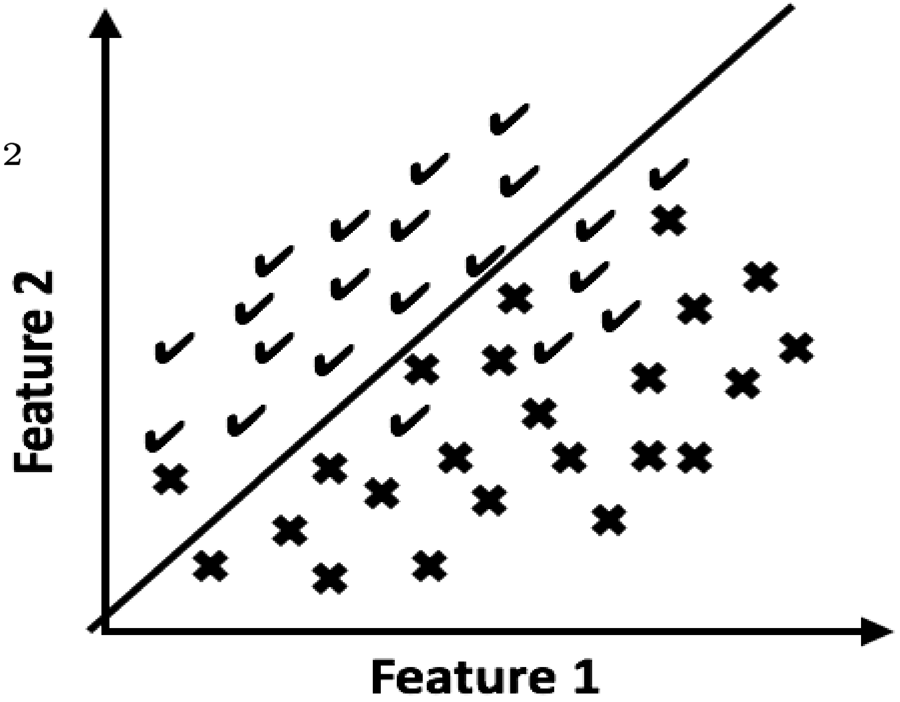
- Why does COMPAS show high racial FPR/FNR disparity?
  - Despite being trained without race information
- Can we train COMPAS to lower racial FPR/FNR disparity?

How COMPAS learns who recidivates



How COMPAS learns who recidivates

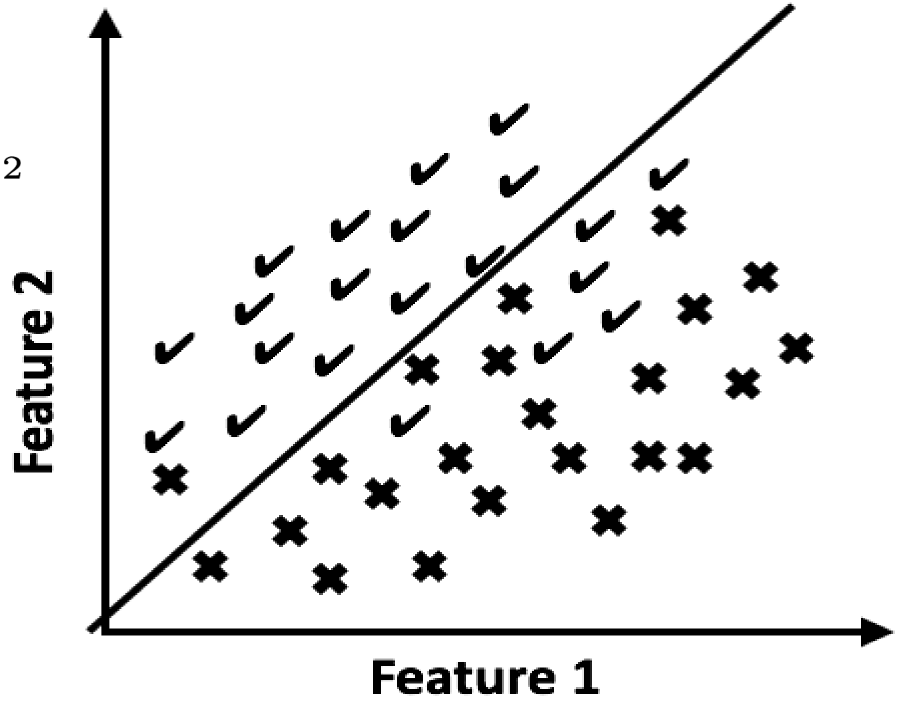
$$\min \quad \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$



- By finding the optimal (most accurate / least loss) linear boundary separating the two classes

How COMPAS learns to discriminate

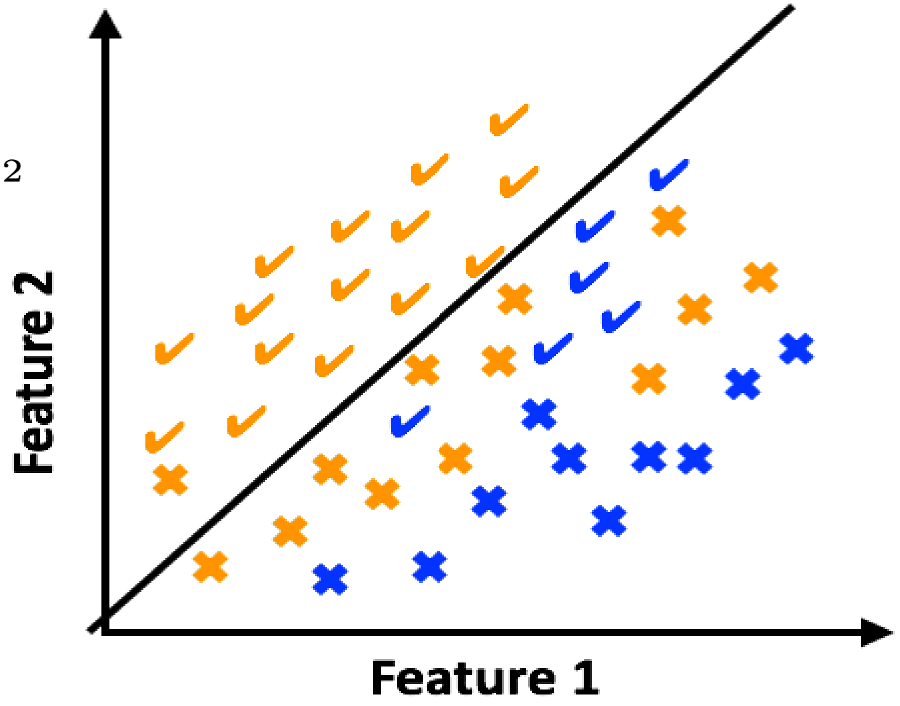
$$\min \quad \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$



- Observe the most accurate linear boundary

How COMPAS learns to discriminate

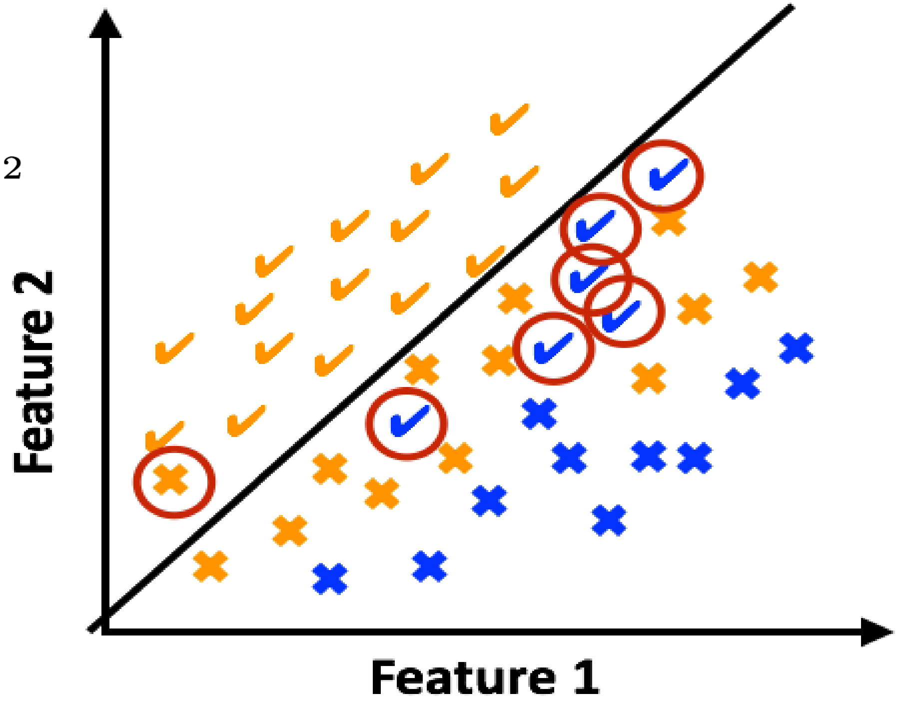
$$\min \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$



- Observe the most accurate linear boundary

How COMPAS learns to discriminate

$$\min \sum_{i=1}^{N} (y_i - d_{\mathbf{w}}(\mathbf{x}_i))^2$$



- Observe the most accurate linear boundary
- Makes few errors for yellow, lots of errors for blue!
  - Causes disparate mistreatment – inequality in error rates

Synthesis:

How to train non-discriminatory classifiers? [WWW '17]

How to learn to avoid discrimination

- Specify discrimination measures as learning constraints
- Optimize for accuracy under those constraints

$$\min \quad P(y_{pred} \neq y_{true})$$

$$\text{s.t.} \quad P(y_{pred} \neq y_{true} \mid race{=}B) = P(y_{pred} \neq y_{true} \mid race{=}W)$$

- The constraints embed ethics & values when learning

- No free lunch: Additional constraints lower accuracy!
- Need race info in training to avoid disp. mistreatment!

Evaluation: Do our constraints work?

- Gathered a recidivism history dataset
  - Broward Country, FL for 2013-14
  - Features: arrest charge, #prior offenses, age,...
  - Class label: 2-year recidivism

- Traditional classifiers without constraints
  - Acc.: **67%** FPR Disparity: **+0.20** FNR Disparity: **-0.30**

- Training classifiers with fairness constraints
  - Acc.: **66%** FPR Disparity: **+0.03** FNR Disparity: **-0.11**

Lessons from the COMPAS story

Take-aways for ethical machine learning

High-level insight: Ethics & Learning

- Learning objectives implicitly embody ethics
  - By how they explicitly define trade-offs in decision errors

- Traditional objective accuracy reflects utilitarian ethics
  - The rightness of decisions is a function of individual outcomes
  - The desired function is maximizing sum of individual utilities

- Lots of scenarios where utilitarian ethics fall short
  - Change learning objectives for other ethical considerations
    - E.g., non-discrimination requires equalizing group-level errors

Three challenges with ethical learning

- Operationalization:
  - How to formally interpret fairness principles in different algorithmic decision making scenarios?
- Synthesis:
  - How to design efficient learning mechanisms for different fairness interpretations?
- Analysis:
  - What are the trade-offs between the learning objectives?

Ongoing work:

From Algorithmic Decision Making To Algorithm-Aided Decision Making

[CSCW '20]

Algorithm-aided Decision Making

- Algorithmic systems are rarely autonomous in practice
- There is often human oversight
  - In recidivism risk prediction, they advice human judges
- Does fair algo. advice lead to fair human decisions?
  - Advice taking is affected by
    - Perceptions of risks and responsibilities for decisions
    - Structure of advice, i.e., timing, framing, representation
    - Trust between algorithmic advisor and human advisee
- Should algo. advice be personalized for human biases?

Looking Forward:

From Non-Discrimination To Fair Algorithmic Decision Making

Social Welfare Theory

Moral Philosophy

Social Choice Theory

Law

Behavioral Economics

Communication & Media Studies

Learning Non-Discriminatory Classification

Regression

Set Selection

Ranking

Matching

Clustering