An overview of fairness methods

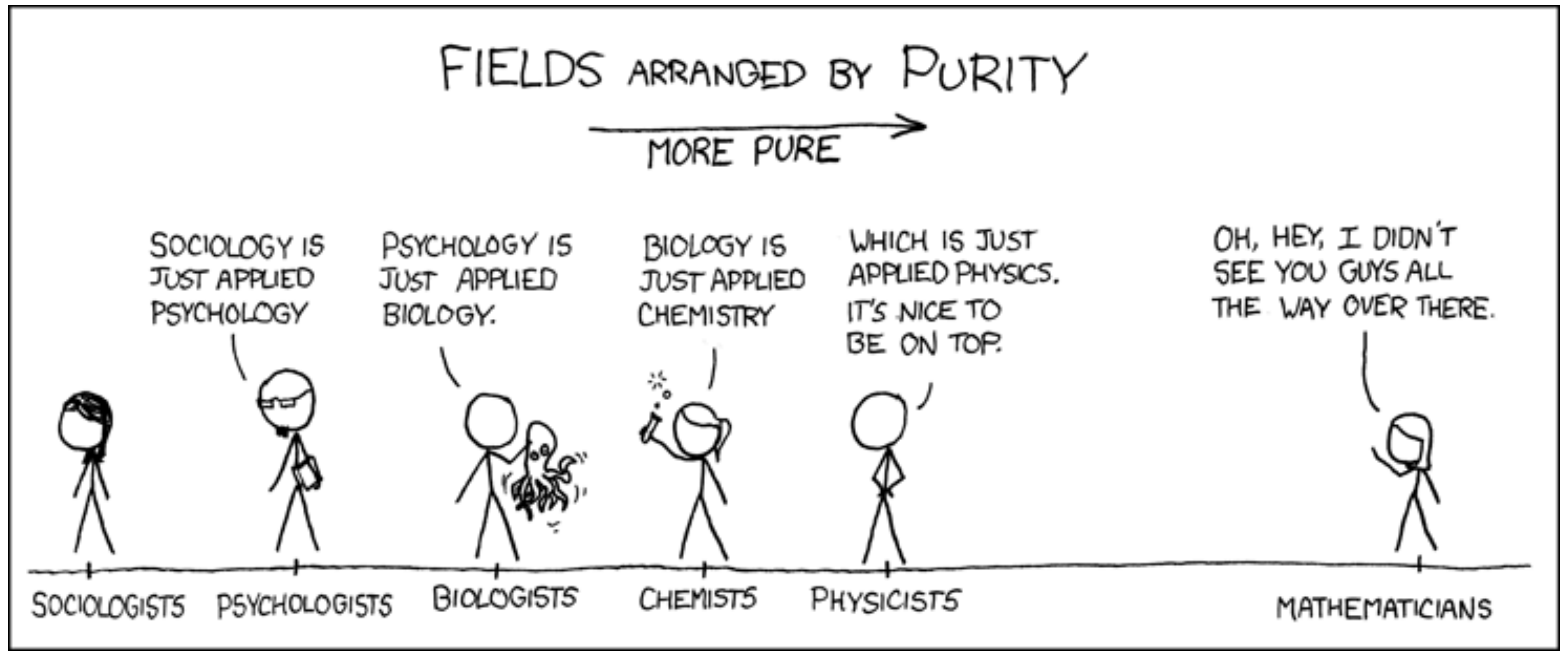
STAT GR5243 Applied Data Science

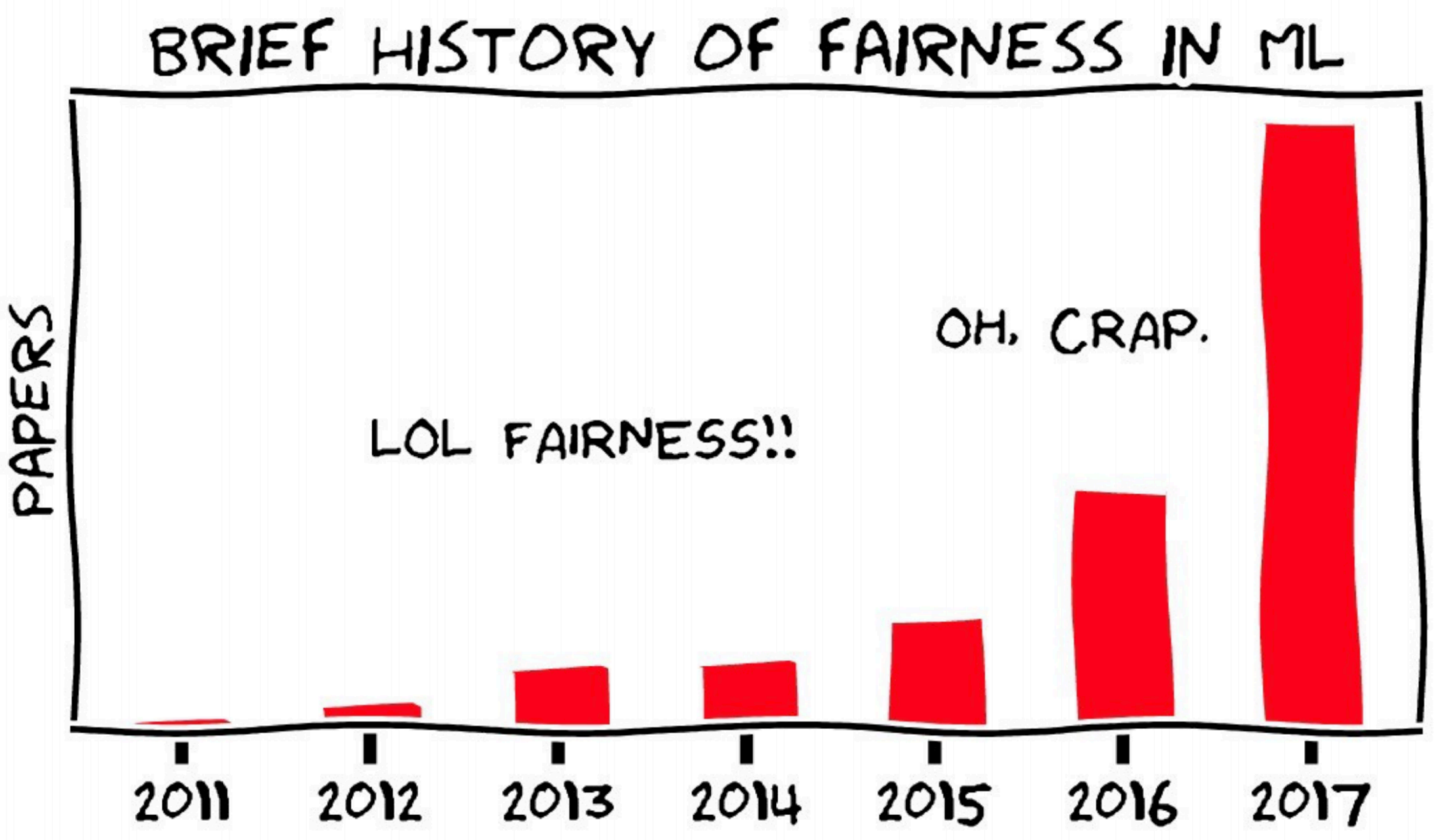
Claire He, Fall 2023

Motivation

Why should we care about fairness in ML?

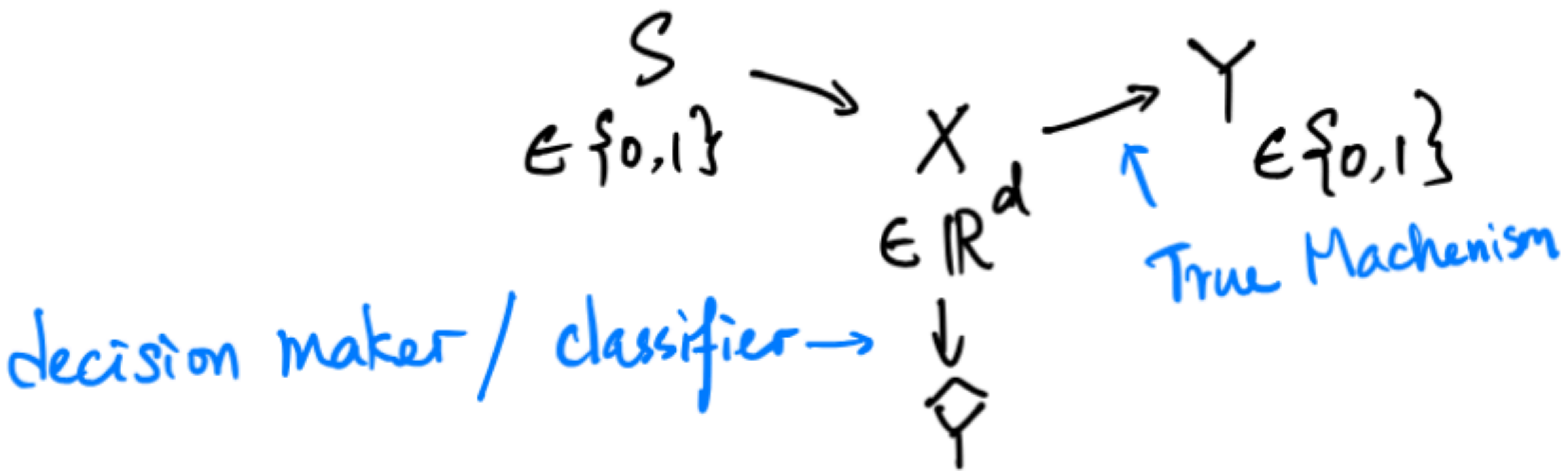
- Philosophical paradigm: science -> objectivity and fairness
- In reality: AI is often a decision-making aiding tool *informed* by domain knowledge/data/engineers/statisticians/data scientist (…)
- Where do we introduce/reproduce **bias, discrimination, …** **« unfairness »**?
- Need for a less confusing definition.


FIELDS ARRANGED BY PURITY
MORE PURE
SOCIOLOGY IS JUST APPLIED PSYCHOLOGY
PSYCHOLOGY IS JUST APPLIED BIOLOGY.
BIOLOGY IS JUST APPLIED CHEMISTRY
WHICH IS JUST APPLIED PHYSICS. IT'S NICE TO BE ON TOP.
OH, HEY, I DIDN'T SEE YOU GUYS ALL THE WAY OVER THERE.
SOCIOLOGISTS
PSYCHOLOGISTS
BIOLOGISTS
CHEMISTS
PHYSICISTS
MATHEMATICIANS



Introduction

What is machine learning fairness? Fish example (classification task)



- Let $Y \in \{0,1\}$ for Bad/Good,

- $X \in \mathbb{R}^d$ our set of features, for Bad/Good classification we can imagine it includes « qualities » of the fish (aggressiveness in the tank to other fishes? Social fish? Small tank fish/big tank fish? …).

- $S \in \{0,1\}$ for blue/red color of the fish.

- We want to predict $\hat{Y}$ by learning a classifier to be as reflective of the true mechanism given features X that we can observe.

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y} = 1 \,|\, S = 0) = P(\hat{Y} = 1 \,|\, S = 1)$

B. Equality of odds: $P(\hat{Y} = 1 \,|\, S = 0, Y = y) = P(\hat{Y} = 1 \,|\, S = 1, Y = y), \; \forall y \in \{0,1\}$

C. Explainable discrimination: $P(\hat{Y} = 1 \,|\, S = 0, X = x) = P(\hat{Y} = 1 \,|\, S = 1, X = x), \; \forall x \in \mathbb{R}^d$

D. Calibration: $P(\hat{Y} = Y \,|\, S = 0) = P(\hat{Y} = Y \,|\, S = 1)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)$ the probability of predicting the fish as good is the same regardless of its color

B. Equality of odds: $P(\hat{Y} = 1 \mid S = 0, Y = y) = P(\hat{Y} = 1 \mid S = 1, Y = y), \ \forall y \in \{0,1\}$

C. Explainable discrimination: $P(\hat{Y} = 1 \mid S = 0, X = x) = P(\hat{Y} = 1 \mid S = 1, X = x), \ \forall x \in \mathbb{R}^d$

D. Calibration: $P(\hat{Y} = Y \mid S = 0) = P(\hat{Y} = Y \mid S = 1)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)$ the probability of predicting the fish as good is the same regardless of its color

B. Equality of odds: $P(\hat{Y} = 1 \mid S = 0, Y = y) = P(\hat{Y} = 1 \mid S = 1, Y = y), \ \forall y \in \{0,1\}$ given the fish is truly good/bad, the probability of prediction is the same regardless of the fish color

C. Explainable discrimination: $P(\hat{Y} = 1 \mid S = 0, X = x) = P(\hat{Y} = 1 \mid S = 1, X = x), \ \forall x \in \mathbb{R}^d$

D. Calibration: $P(\hat{Y} = Y \mid S = 0) = P(\hat{Y} = Y \mid S = 1)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y}=1\,|\,S=0)=P(\hat{Y}=1\,|\,S=1)$ the probability of predicting the fish as good is the same regardless of its color

B. Equality of odds: $P(\hat{Y}=1\,|\,S=0,Y=y)=P(\hat{Y}=1\,|\,S=1,Y=y),\ \forall y\in\{0,1\}$ given the fish is truly good/bad, the probability of prediction is the same regardless of the fish color

C. Explainable discrimination: $P(\hat{Y}=1\,|\,S=0,X=x)=P(\hat{Y}=1\,|\,S=1,X=x),\ \forall x\in\mathbb{R}^d$ the probability of predicting the fish as good is the same regardless of color given the same observed features

D. Calibration: $P(\hat{Y}=Y\,|\,S=0)=P(\hat{Y}=Y\,|\,S=1)$

Introduction

Fairness metrics for classification

A. Parity: $P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1)$ the probability of predicting the fish as good is the same regardless of its color

B. Equality of odds: $P(\hat{Y} = 1 \mid S = 0, Y = y) = P(\hat{Y} = 1 \mid S = 1, Y = y), \forall y \in \{0,1\}$ given the fish is truly good/bad, the probability of prediction is the same regardless of the fish color

C. Explainable discrimination: $P(\hat{Y} = 1 \mid S = 0, X = x) = P(\hat{Y} = 1 \mid S = 1, X = x), \forall x \in \mathbb{R}^d$ the probability of predicting the fish as good is the same regardless of color given the same observed features

D. Calibration: $P(\hat{Y} = Y \mid S = 0) = P(\hat{Y} = Y \mid S = 1)$ the probability of correct classification is the same regardless of the color

Introduction

The Impossibility Theorem

Kleinberg et al. (2016) showed that A, B and D (parity, equalized odds and calibration) can **not** be jointly optimized.

This means we will have to carefully choose and specify our metrics of fairness and that any AI system we build will necessarily violate some notion of fairness.

Our 4 papers introduce frameworks that aim for ensuring some level of **fairness** in ML tasks through different layers of ML workflow.

1. What is the fairness framework?

2. Where is the fairness introduced in the workflow?

ML fairness methods

An overview of some approaches to fairness

1. Pre-processing methods: modify training data
    - A. Local massaging: relabeling points near the boundary
    - B. Local preferential resampling: resample points close to the boundary
2. In-processing methods: modify the learning algorithm
    - C. Through cost functions/constraints (regularization)
    - D. Through the pipeline : adding a latent representation
    - E. Through feature selection
3. Post-processing methods: modify the prediction outcome
4. Causal reasoning

Learning Fair Representations

Paper 1

- Fairness framework: **group/individual fairness**
  - Group: the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole
  - Individual: similar individuals should be treated similarly
  - Fairness metric: $P(Z = k \,|\, X, S = 0) = P(Z = k \,|\, X, S = 1)$
- Method: (2D) learning a latent representation (think dimension reduction methods like PCA)
  - X features, $S \in \{0,1\}$ protected set
  - $Z \sim Mult(n, v)$: with K « prototypes » associated to $(v_k)_{k=1,\ldots,K}$
  - $X \in \mathscr{X} \longrightarrow Z \in \{1,\ldots,K\} \longrightarrow Y \in \{0,1\}$

Learning Fair Representations

Paper 1

$X \in \mathscr{X} \longrightarrow Z \in \{1,\ldots,K\} \longrightarrow Y \in \{0,1\}$

Idea:

- X informative but correlated with S —> discrimination
- Find an intermediate Z that keeps information, but is less correlated with S by adding an unfairness loss that ensures « parity » $P(Z = k \,|\, X, S = 0) = P(Z = k \,|\, X, S = 1)$
- Minimize simultaneously **reconstruction loss** $L_X = ||X - \hat{X}||_2$ where $\hat{X} = f(Z)$**, cross entropy (classification) and unfairness loss** $L_Z = \sum_k |P(Z = k \,|\, S = 0) - P(Z = k \,|\, S = 1)|$

Fairness constraints

Paper 2

- Fairness framework: **Disparate treatment**
  - The decisions are (partly) based on the individual's sensitive attribute.
  - Procedural unfairness, inequal opportunity
  - Resulting in direct discrimination Solution: don't use the sensitive attribute when making decisions.
- Method: (2C) modify the cost functions by adding a penalty term for being « unfair »
  - $D = (X, Y, S)$ dataset
  - $L_\theta(D)$ classification loss (cross entropy f.e.)
  - $R_\theta(D)$ a measure of unfairness

Fairness constraints

Paper 2

Define $R_\theta(D) = |\,Cov(s, d_\theta(x))\,|$ with the signed distance to decision boundary to quantify unfairness.

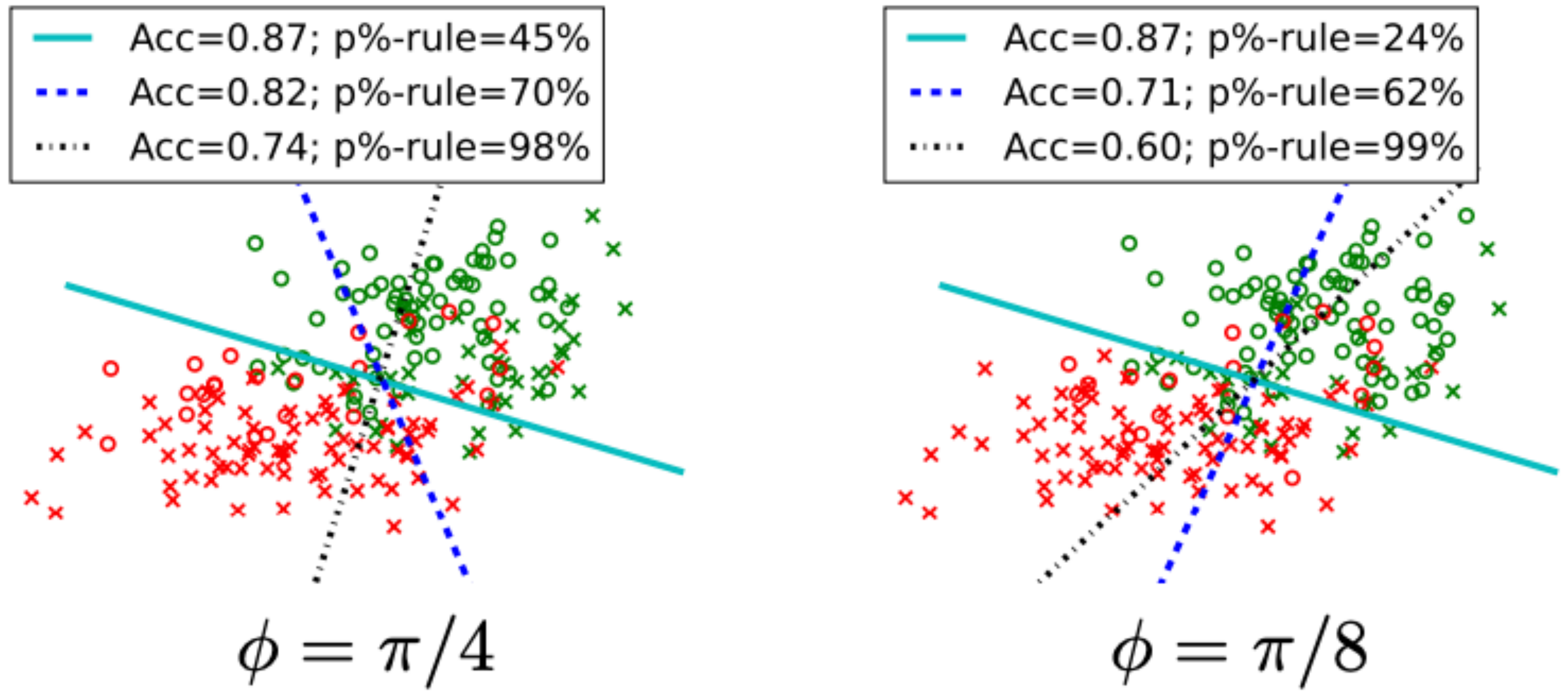
Maximise accuracy under fairness constraint:

$$\min_\theta L_\theta(D) \text{ s.t. } R_\theta(D) \le \tau$$

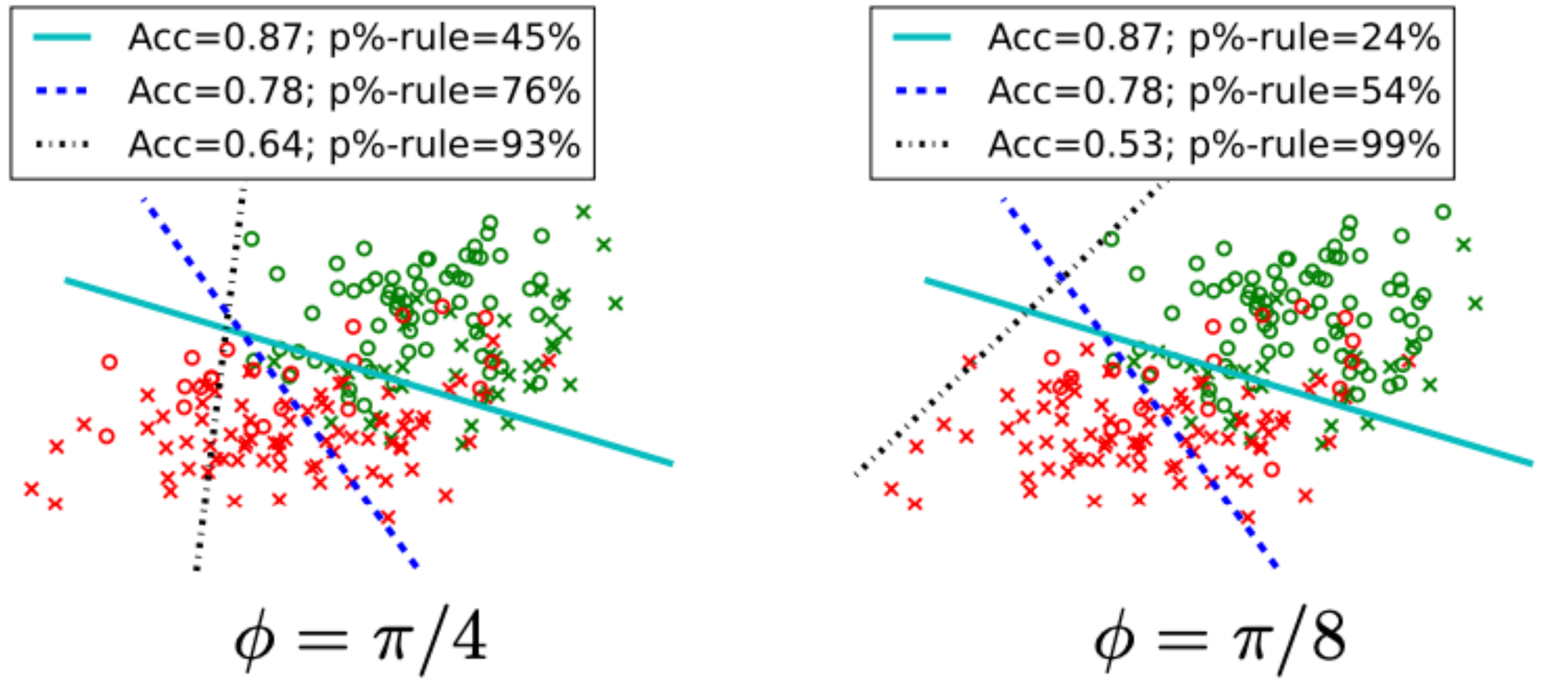
Maximise fairness under accuracy constraint:

$$\min R_\theta(D) \text{ s.t. } L(\theta) \le (1+\gamma)L(\theta^*)$$

Method applied to Logistic Regression and SVM (appendix)



(a) Maximizing accuracy under fairness constraints



(b) Maximizing fairness under accuracy constraints

Learning without Disparate Mistreatment

Paper 3

- Fairness framework: **disparate treatment, mistreatment, impact**
    - No disparate treatment: $P(\hat{y} \,|\, x, s)$
    - No disparate impact: $P(\hat{y} = 1 \,|\, s = 0) = P(\hat{y} = 1 \,|\, s = 1)$
    - No disparate **mistreatment:** if the misclassification rates for different groups of people having different values of the sensitive feature s are the same.
- Methods: (2C)
- Extension builds on the framework from the previous model, we use a continuous version of $Cov(S, \hat{Y}) \rightarrow Cov(s, g_\theta(y, X))$ where we choose g_θ to be some signed distance between misclassified users' feature vectors to the boundary.

Learning without Disparate Mistreatment

Paper 3

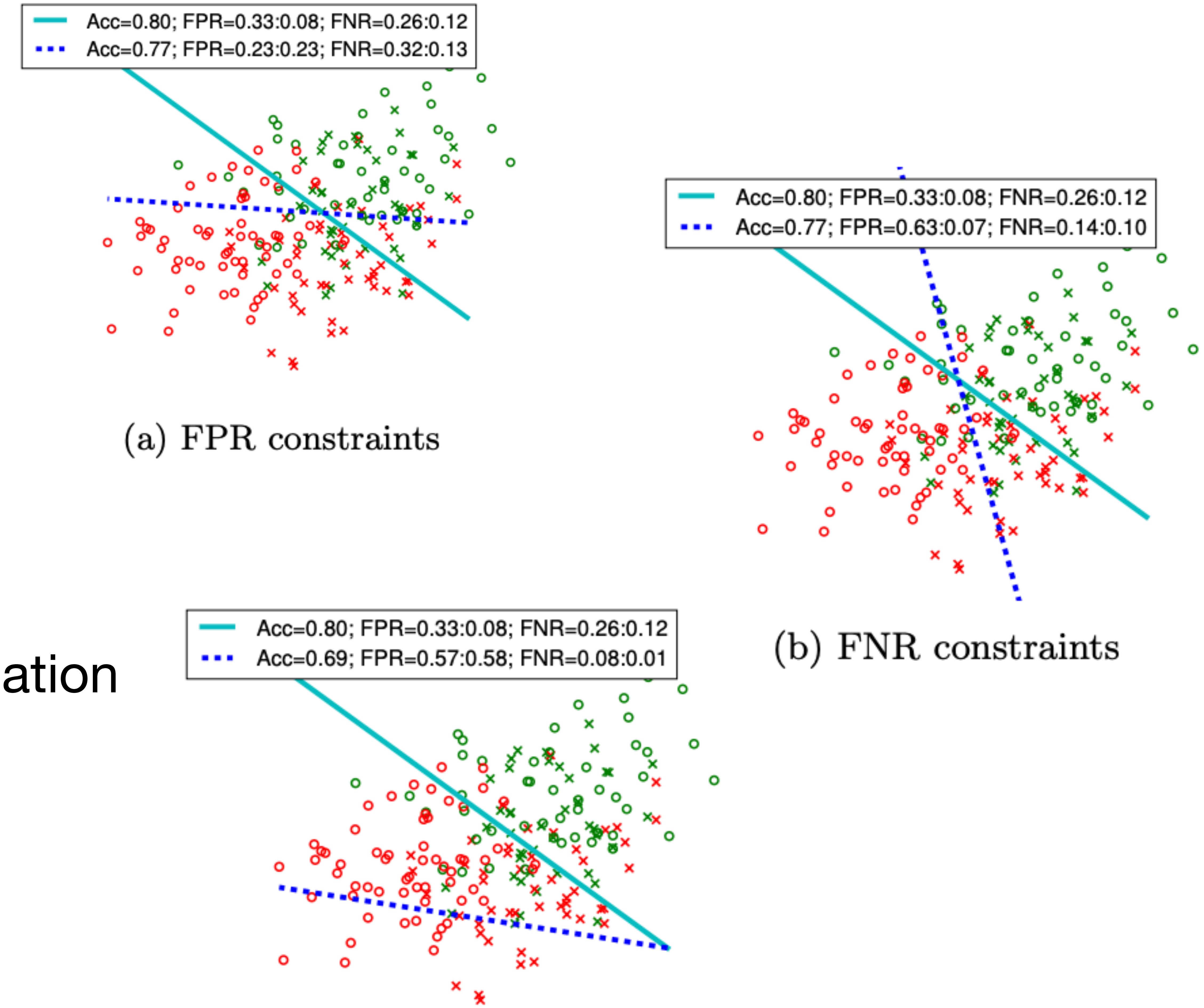
Optimization based classification method:

$$\min L_\theta(D) \;\; \text{s.t.} \;\; M(D) < \epsilon$$

where $M(D) = Cov(s, g_\theta(y, X))$ is some metric of misclassification that brings unfairness.

For example

$$g_\theta(y,x) = \begin{cases} 0 \wedge y d_\theta(x) & \text{if control overall missclassification} \\ \frac{1-y}{2} y d_\theta(x) & \text{if control FPR} \\ \frac{1+y}{2} y d_\theta(x) & \text{if control FNR} \end{cases}$$



(a) FPR constraints

(b) FNR constraints

(c) Both constraints

Fairness-aware feature selection

Paper 4

- Fairness framework:
    - **group fairness**
    - **individual fairness**
- Methods: (1A+ 1B+ 2E + 4)
    - Information theoretical metrics for feature selection: fairness utility score
    - Conditional discrimination through pre-processing via local massaging/local preferential sampling
    - Regularization via prejudice index