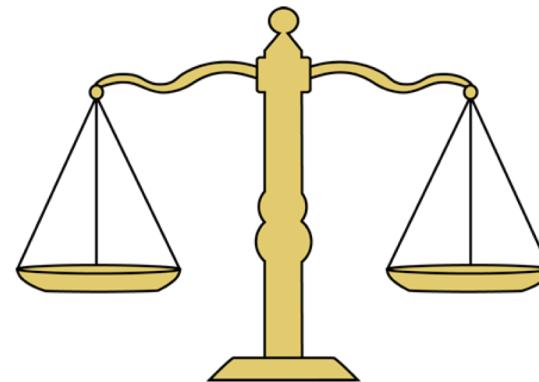


# Fairness in Machine Learning

Isabel Valera

MPI for Intelligent Systems



# Machine Learning can be unfair to many

PROPUBLICA | MACHINE BIAS

## Machine Bias

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



Amazon scraps secret AI recruiting tool that showed bias against women



Rating systems may discriminate against Uber drivers  
December 16, 2016 by Leslie Morris

The New York Times

HIDDEN BIAS

When Algorithms Discriminate

TIME

Google Has a Striking History of Bias Against Black Girls

PROPUBLICA | MACHINE BIAS

Minority Neighborhoods Pay Higher Car Insurance Premiums Than White Areas With the Same Risk

# Example I: Recidivism Predictions

PROPUBLICA | MACHINE BIAS

## Machine Bias

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

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

*Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)*

Fig2: The bias in COMPAS. (from [Larson et al. ProPublica, 2016](#))

# Example II: Face Recognition

TIME

Google Has a Striking History of Bias  
Against Black Girls

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	<b>100</b>
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	<b>20.8</b>	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	<b>100</b>	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	<b>16.3</b>	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	<b>99.3</b>	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	<b>34.5</b>	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	<b>98.9</b>	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	<b>23.4</b>	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	<b>99.7</b>
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	<b>34.7</b>	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	<b>99.6</b>	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	<b>25.2</b>	17.7	5.20	0.4

Fig4: The bias in commercial face recognition services(Buolamwini and Gebru, 2018). DF, DM, LF, LM stand for: darker skin female, darker skin male, lighter skin female and lighter skin male. PPV, TPR, FPR stand for predictive positive value, true positive rate and false positive rate.

# Fairness in ML

## A specific type of unfairness: Discrimination

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

[Altam'16]

### Challenge # 1: From Definitions to Measures

#### How to measure fairness?

- Formalizing or **interpreting** a fuzzy definition to make it **measurable** for empirical observations
- Existing measures may be insufficient or even unsuitable

# Fairness in ML

## A specific type of unfairness: Discrimination

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

[Altam'16]

### Challenge # 2: Mechanisms

#### How to incorporate fairness into algorithmic decision making systems?

- Trade-off fairness-accuracy and between fairness notions
- Efficient training, finite samples and feedback loops.

# Why Discrimination as notion of (un) fairness?

## Legally recognized ‘protected social groups’:

**Race** (Civil Rights Act of 1964); **Color** (Civil Rights Act of 1964); **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964); **Religion** (Civil Rights Act of 1964); **National origin** (Civil Rights Act of 1964); **Citizenship** (Immigration Reform and Control Act); **Age** (Age Discrimination in Employment Act of 1967); **Pregnancy** (Pregnancy Discrimination Act); **Familial status** (Civil Rights Act of 1968); **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990); **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act); **Genetic information** (Genetic Information Nondiscrimination Act)

# **Part I**

# **Sources of Unfairness**

# Sources of Unfairness

## 1. Human biases in historical data

*A father and his son are involved in a horrific car crash and the man died at the scene. But when the child arrived at the hospital and was rushed into the operating theatre, the surgeon pulled away and said: “I can’t operate on this boy, he’s my son”.*

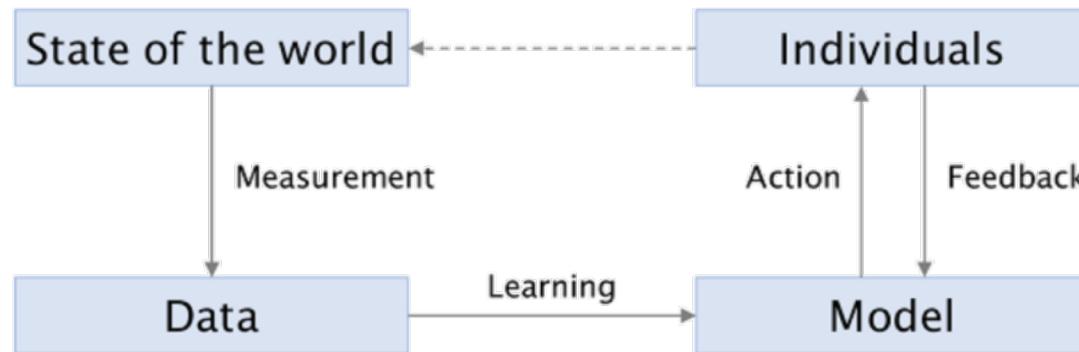
**Historical data** contains human biases and stereotypes

# Sources of Unfairness

## 2. Limited features

Summarizing the world  
(& individuals)  
with a **finite set of features**

Features may be  
**less informative** or reliably  
collected for **minority group(s)**



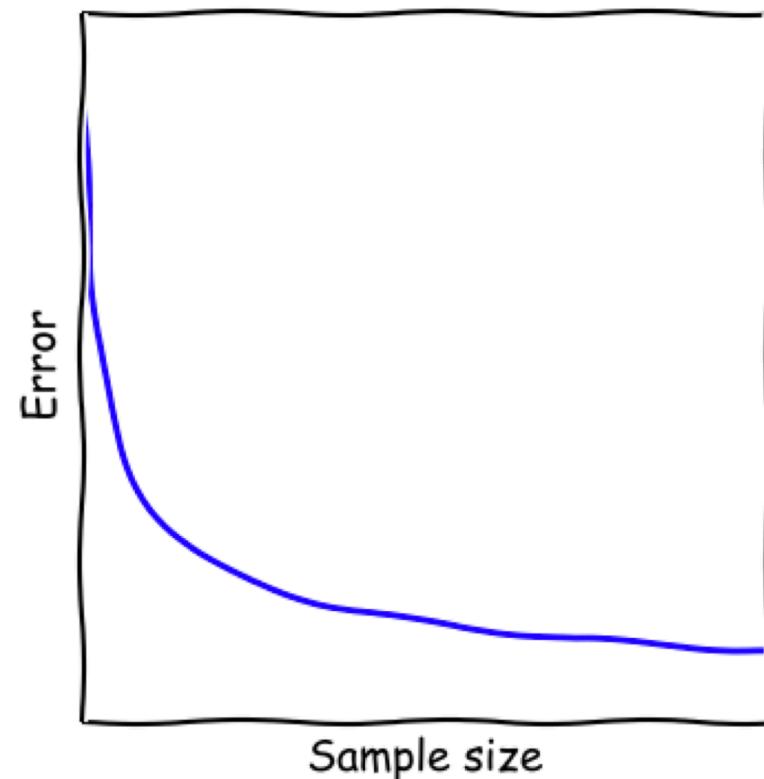
# Sources of Unfairness

## 3. Sample size disparity

**Less data** from  
minority groups

**ML methods** perform  
worse with less data

**ML bias** is a side effect  
of maximizing accuracy

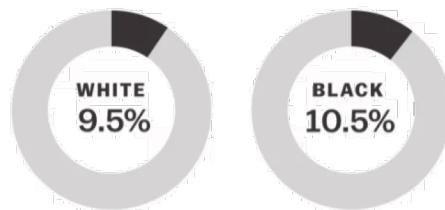


# Sources of Unfairness

## 5. Skewed samples

### Past-month illicit drug use

2013 National Survey on Drug Use and Health

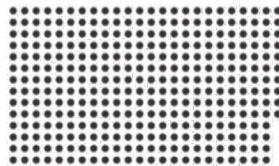


Effect of interventions  
and feedback loops

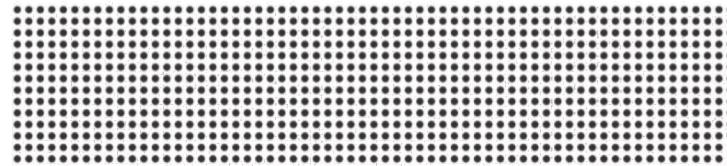
### Drug-related arrests per 100,000 residents of each race

2013 FBI Uniform Crime Reports / US Census Bureau

WHITE 332



BLACK 879



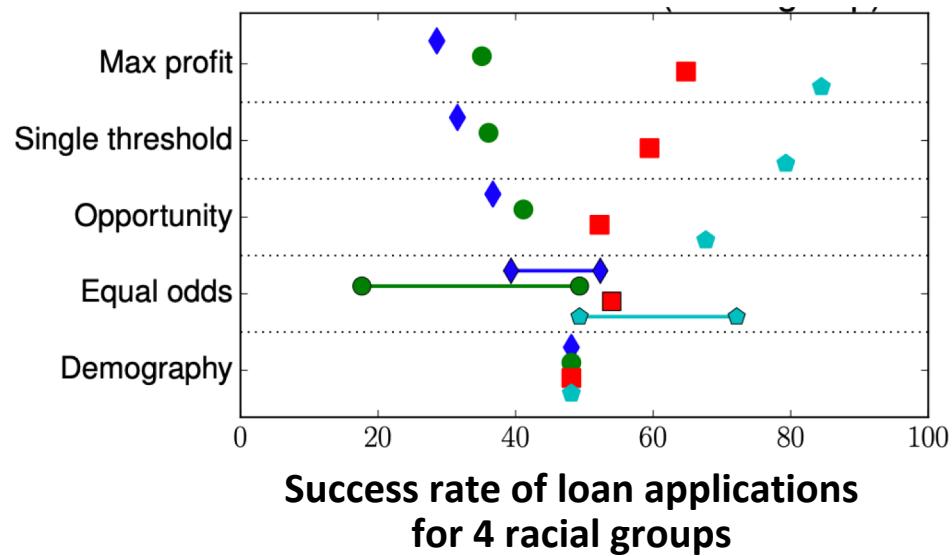
Vox

# Sources of Unfairness

## 4. Proxies

Removing sensitive information from the data is not enough

ML good at picking proxies in the data



# **Part II**

# **Fairness Definitions**

# Group vs. Individual Fairness

A specific type of unfairness: **Discrimination**

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

[Altam'16]

## 1. Individual fairness:

Are **individuals** treated by a decision making system consistently independently of the **social salient groups** they belong to?

## 2. Group fairness:

Do the **outcomes** of a decision making system systematically differ **between social salient groups**?

# 1. Individual Fairness

Are **individuals** treated by a decision making system consistently independently of the **social salient groups** they belong to?

*“similar individuals should be treated similarly”*

[“Fairness Through Awareness”, Dwork et al. in 2012]

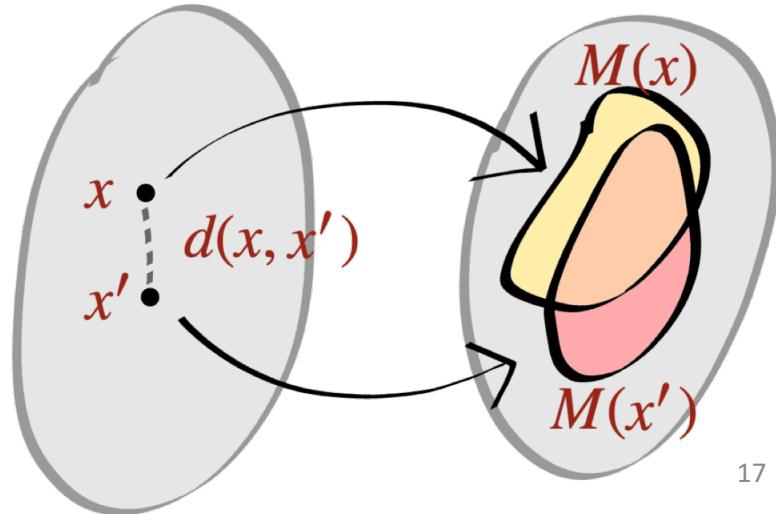
# Ensuring Individual Fairness (I)

*“similar individuals should be treated similarly”*

Assume task specific dissimilarity measure  $d(x, x')$

Require similar individuals map to  
similar distributions over outcomes  
via map  $M: \mathcal{X} \rightarrow \Delta(\mathcal{O})$ :

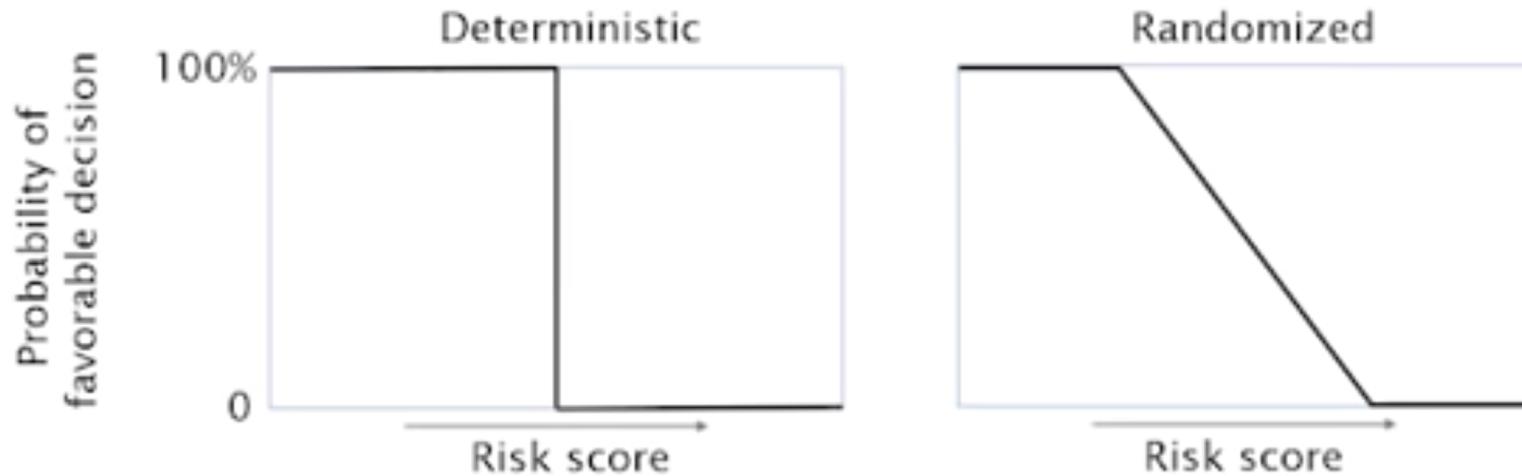
$$D(M(x), M(x')) \leq d(x, x')$$



[Dwork et al., 2012]

[Image from Hardt's tutorial on Fairness in ML, NeurIPS 2017]

# Ensuring Individual Fairness (II)



**Deterministic decision making system do not satisfy individual fairness.  
Individual fairness may be achieve with randomized systems.**

## 2. Group Fairness

### Discrimination

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

[Altam'16]

Do the **outcomes** of a decision making system systematically differ **between social salient groups?**

# How do we define *wrongfully*, formally?

It depends on whether we have access to a groundtruth:

One can *tell* whether a (historical) decision was **right** or **wrong**

Example (groundtruth):

loan request



historical decision



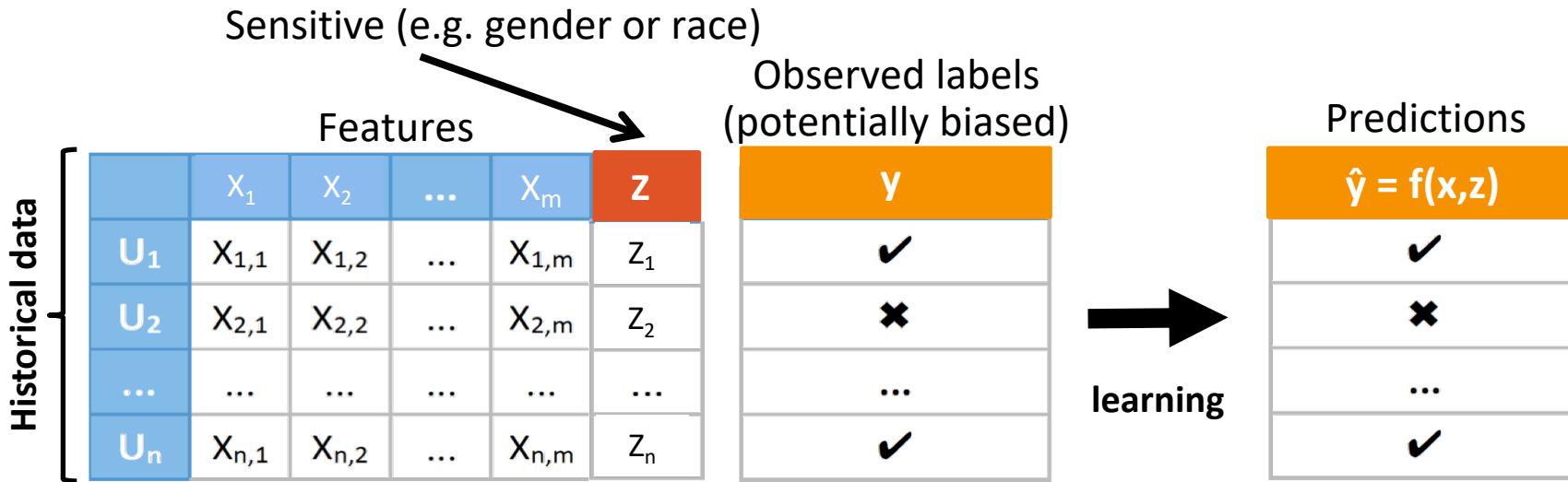
decision was **right**



decision was **wrong**



# Fairness without groundtruth



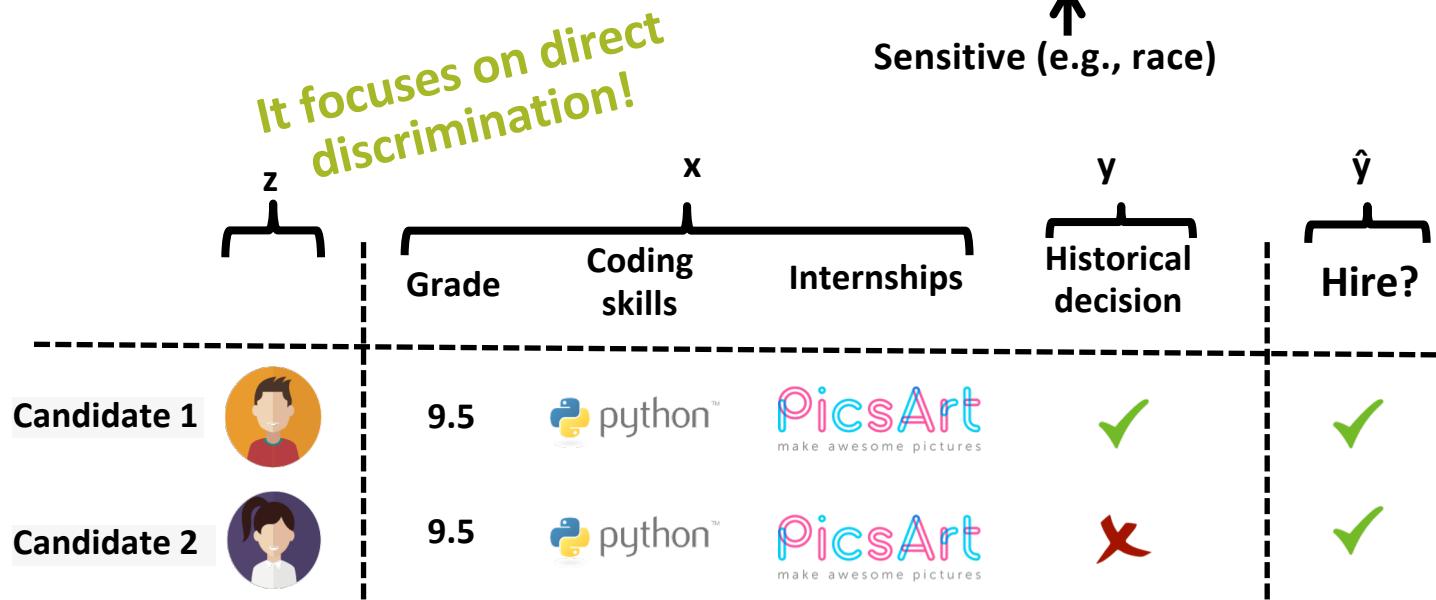
Challenge: we do not know whether a decision was **right** or **wrong**

Fairness: **parity in treatment** or **impact**

# Fairness without groundtruth (I)

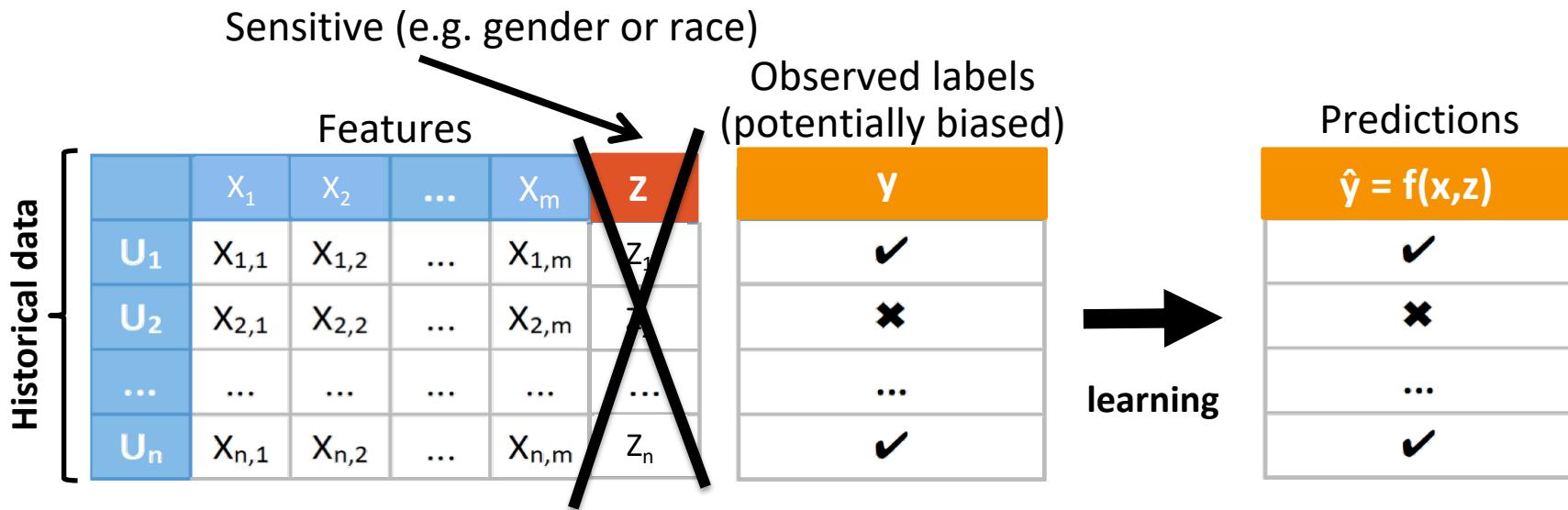
## Parity (or equality) in treatment (*unawareness*)

- Outcomes should not change with change in sensitive feature



# Achieving parity in treatment

Just remove the sensitive features:

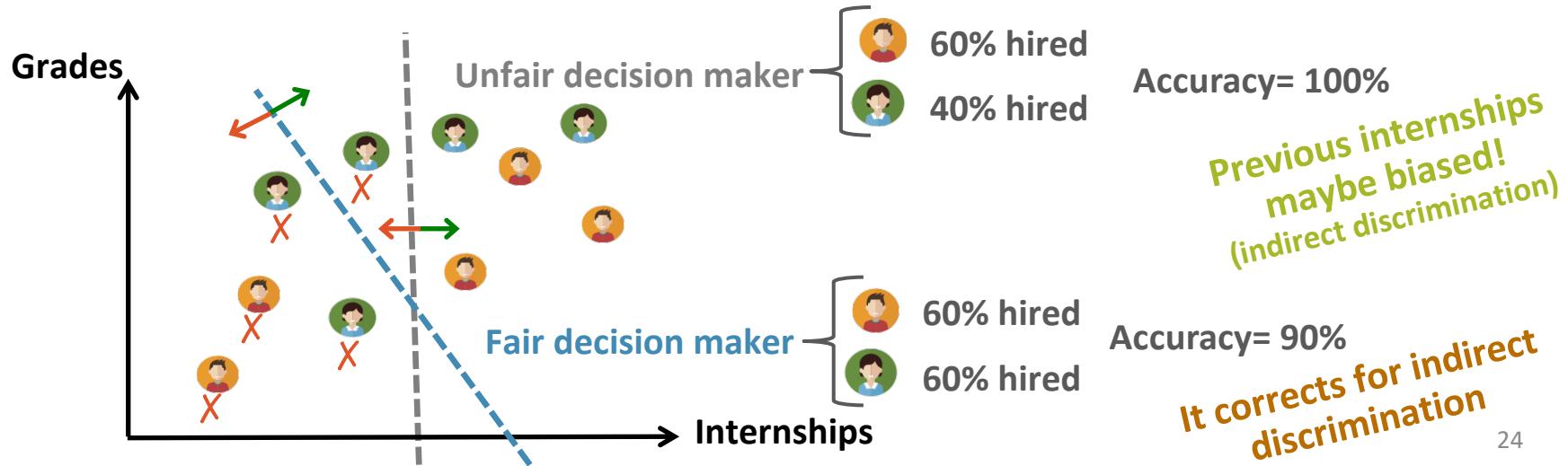


# Fairness without groundtruth (II)

## Parity (or equality) in **impact** (or *demographic parity*)

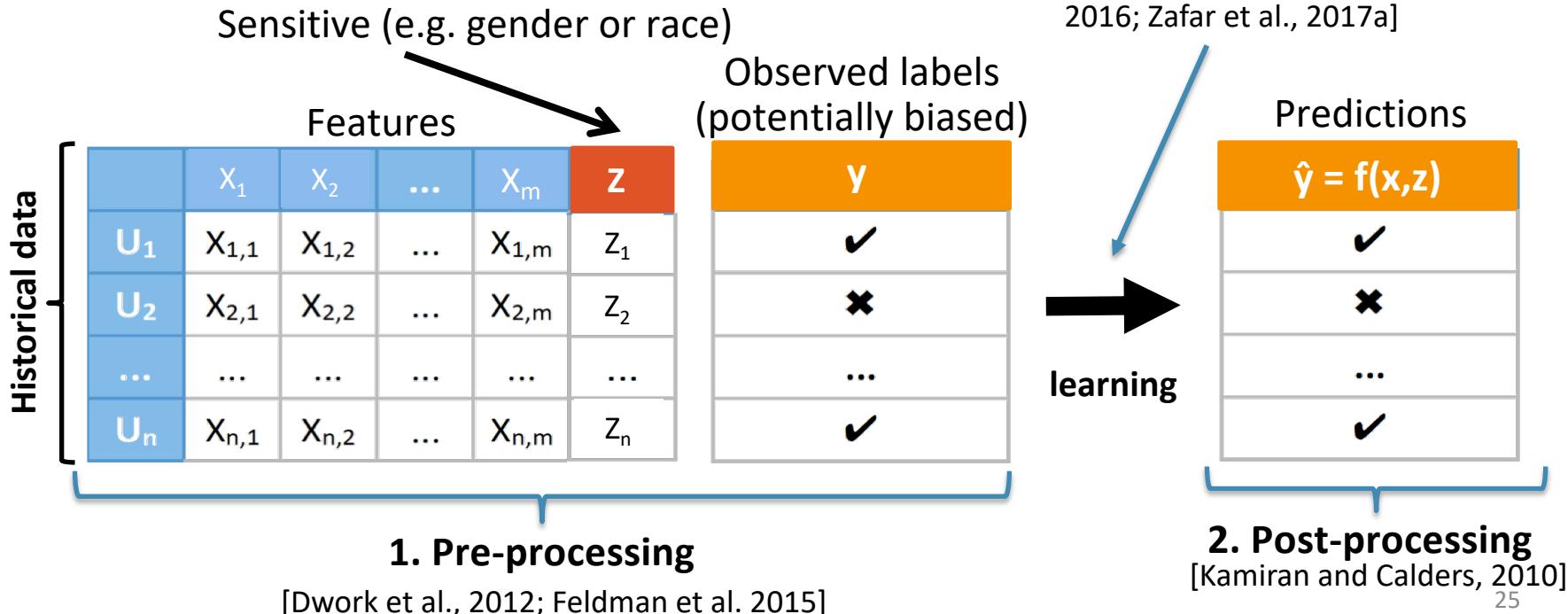
→ Decisions should be the same for all sensitive feature groups

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

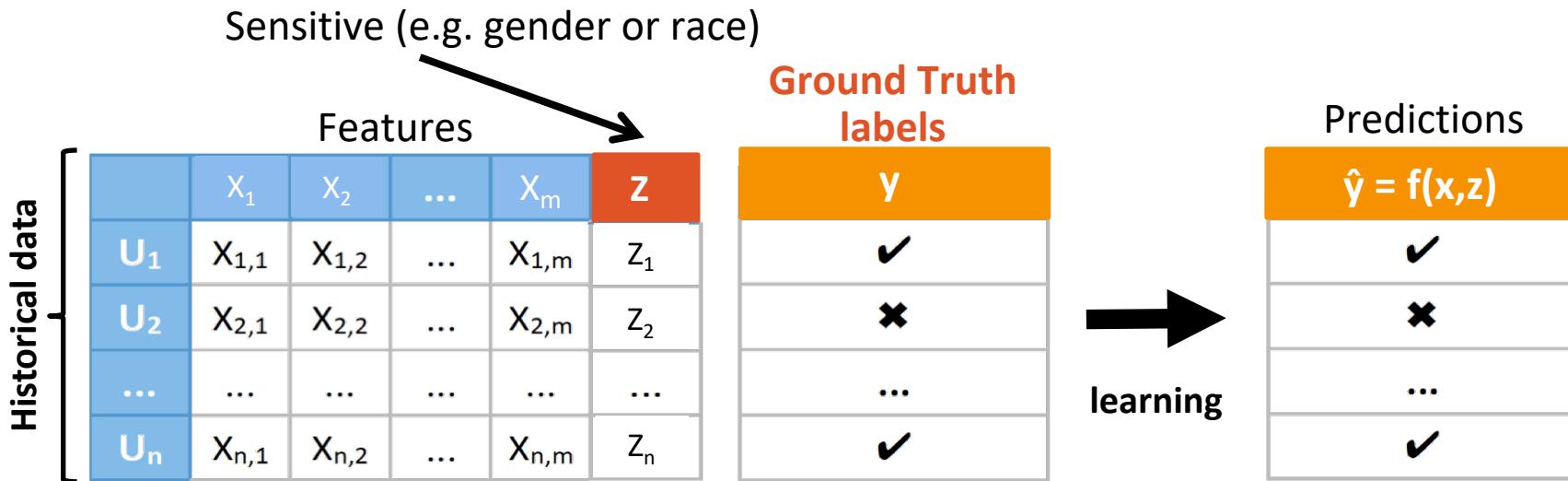


# Achieving parity in impact

## Three types of approaches:



# Fairness with groundtruth



**Groundtruth:** we know whether a decision was **right** or **wrong**

**Fairness:** parity in error/predictive rates

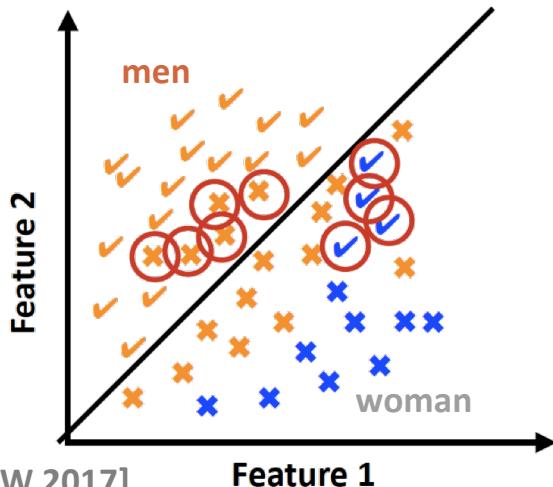
# Fairness with groundtruth (I)

## Parity in mistreatment

→ Errors should be the same for all sensitive feature groups

$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1)$$

Predicted label      True label



Errors for **men** are false positives  
(e.g., wrongly **granted** loan)

Errors for **women** are false negatives  
(e.g., wrongly **denied** loan)

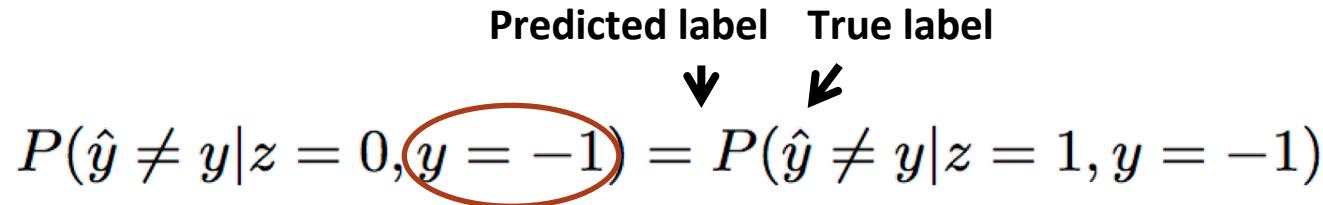
**Disproportionate advantage  
to men!**

# Fairness with groundtruth (II)

## Parity in false positive rates

$$P(\hat{y} \neq y | z = 0, y = -1) = P(\hat{y} \neq y | z = 1, y = -1)$$

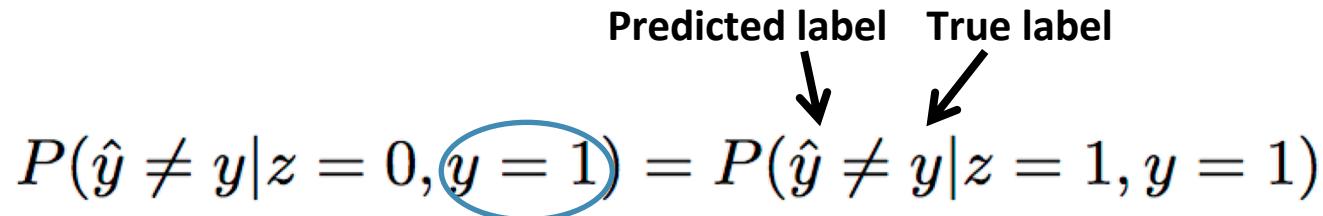
Predicted label    True label  
                        ↓      ↘



## Parity in false negative rates

$$P(\hat{y} \neq y | z = 0, y = 1) = P(\hat{y} \neq y | z = 1, y = 1)$$

Predicted label    True label  
                        ↓      ↘



# Fairness with groundtruth (III)

## Parity in Positive Rate (or *Equalized Odds*)

→ Positive rates should be the same for all sensitive feature groups

$$p(\hat{y} = 1 | z = 0, y) = p(\hat{y} = 1 | z = 1, y)$$

↑  
Predicted label      ↑  
True label

## Parity in True Positive Rate (or *Equal Opportunity*)

→ Positive rates should be the same for all sensitive feature groups

$$p(\hat{y} = 1 | z = 0, y = 1) = p(\hat{y} = 1 | z = 1, y = 1)$$

↑  
Predicted label      ↑  
True label

# Achieving parity in mistreatment

Two types of approaches:

Sensitive (e.g. gender or race)

Features

	$X_1$	$X_2$	...	$X_m$	$Z$
$U_1$	$X_{1,1}$	$X_{1,2}$	...	$X_{1,m}$	$Z_1$
$U_2$	$X_{2,1}$	$X_{2,2}$	...	$X_{2,m}$	$Z_2$
...	...	...	...	...	...
$U_n$	$X_{n,1}$	$X_{n,2}$	...	$X_{n,m}$	$Z_n$

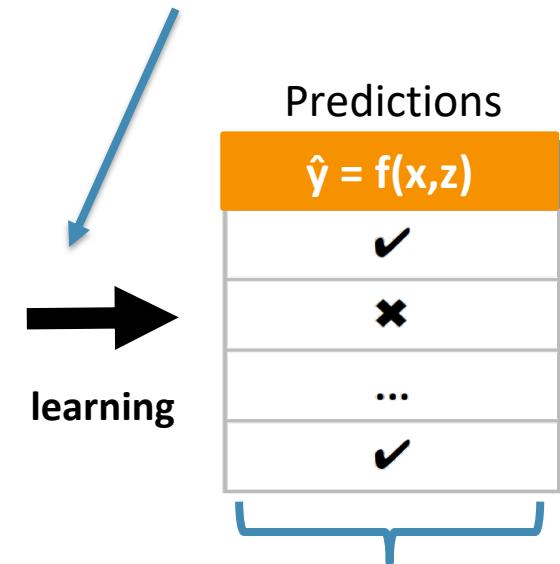
Training true labels

$y$

✓
✗
...
✓

## 2. Modify training

[Zafar et al., 2017b, Woodworth et al., 2017]



## 1. Post-processing

[Hardt et al., 2016]

# More definitions of (group) fairness...

		Predicted Label		
		$\hat{y} = 1$	$\hat{y} = -1$	
True Label	$y = 1$	True positive	False negative	$P(\hat{y} \neq y   y = 1)$ False Negative Rate
	$y = -1$	False positive	True negative	$P(\hat{y} \neq y   y = -1)$ False Positive Rate
		$P(\hat{y} \neq y   \hat{y} = 1)$ False Discovery Rate	$P(\hat{y} \neq y   \hat{y} = -1)$ False Omission Rate	$P(\hat{y} \neq y)$ Overall Misclass. Rate

Conditioning on the prediction

Conditioning on the label

# No universal definition

Different **real-world scenarios**...

Historical decisions  
(potentially biased)

... require different **definitions of fairness**...

Parity (equality) in impact

... translates into into different **measurements**.

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

# No universal definition

Different **real-world scenarios**...

Ground Truth

... require different **definitions of fairness**...

Parity (equality) in mistreatment or Equal Opportunity

... translates into different **measurements**.

$$P(\hat{y} \neq y | z = 0) = P(\hat{y} \neq y | z = 1)$$

# Exercise!!

**Disparate treatment:**  $P(\hat{y}|\mathbf{x}, z) = P(\hat{y}|\mathbf{x})$

**Disparate impact:**  $P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$

**Disparate mistreatment:**  $P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1)$

User Attributes			Ground Truth (Has Weapon)	Classifier's Decision to Stop				Disp. Treat.	Disp. Imp.	Disp. Mist.	
Sensitive	Non-sensitive			C <sub>1</sub>	C <sub>2</sub>	C <sub>3</sub>					
Gender	Clothing Bulge	Prox. Crime	✓	✓	1	1	1	0	1	0	1
Male 1	1	1	✓	✓	1	1	1	0	1	0	1
Male 2	1	0	✓	✓	1	1	0	1	0	1	0
Male 3	0	1	✗	✗	1	0	1	1	0	1	0
Female 1	1	1	✓	✓	1	0	1	1	0	1	0
Female 2	1	0	✗	✗	1	1	1	0	1	0	1
Female 3	0	0	✓	✓	0	1	0	1	0	1	0

# Fairness as Statistical (conditional) Independence

**Independence:** Require the prediction and the sensitive features to be independent:

$$\hat{Y} \perp Z$$

$Z$	Sensitive feature
$Y$	Target (label)
$\hat{Y}$	Prediction

**Separation:** Require the prediction and the sensitive features to be independent conditional on the target:

$$\hat{Y} \perp Z|Y$$

**Sufficiency:** Require the target and the sensitive features to be independent conditional on the prediction:

$$Y \perp Z|\hat{Y}$$

# Independence

$$\hat{Y} \perp Z$$

- Equivalent to demographic parity, statistical parity, disparate impact...
- Ignores dependencies between true label and sensitive feature
- Accepts qualified people in one group, random in other
- Allows to trade FN for FP
- Still usefull when there is no ground truth (label and decision coincide) to correct historical bias (internship example)

$Z$	Sensitive feature
$Y$	Target (label)
$\hat{Y}$	Prediction

# Separation

$$\hat{Y} \perp Z|Y$$

- Accounts for parity in mistreatment, equalized odds, equal opportunity...
- Allows for dependencies between the label and the sensitive attribute to the extent that it is justified by the target.
- Ensures that under imperfect predictors, the errors are equal across groups.
- Suitable when there is available ground truth labels, e.g., in loan application domain (decision=approval and label=repay).

$Z$	Sensitive feature
$Y$	Target (label)
$\hat{Y}$	Prediction

# Sufficiency

$$Y \perp Z | \hat{Y}$$

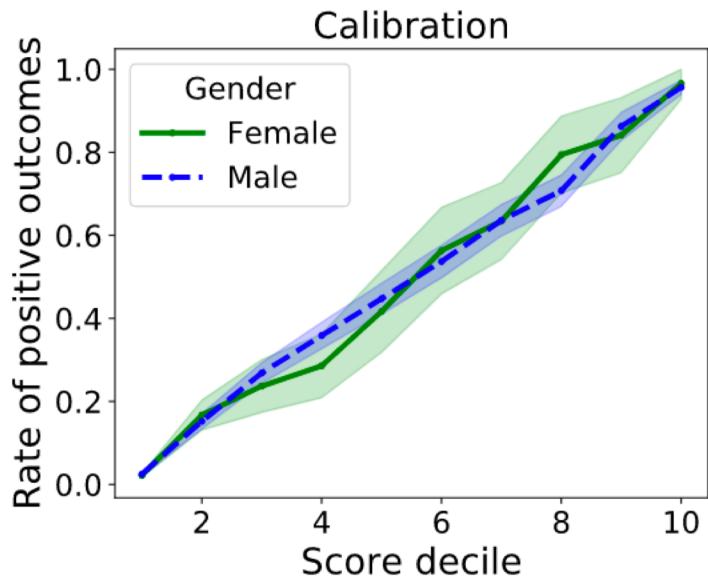
- Accounts for parity in discovery rates, predictive parity, calibration by groups...
- Implies equal chance of success given positive prediction:  $p(y = 1|z = 0, \hat{y}) = p(y = 1|z = 0, \hat{y})$
- Equivalent to *calibration (by group)*, which implies that:  
 $p(y = 1|r) = r$ , given the predicted score  $r = E[\hat{y}|x]$
- Standard ML classifiers are often calibrated (e.g., Bayes optimal score). Imposing sufficiency may not result in a substantial change in current ML practices.

$Z$	Sensitive feature
$Y$	Target (label)
$\hat{Y}$	Prediction

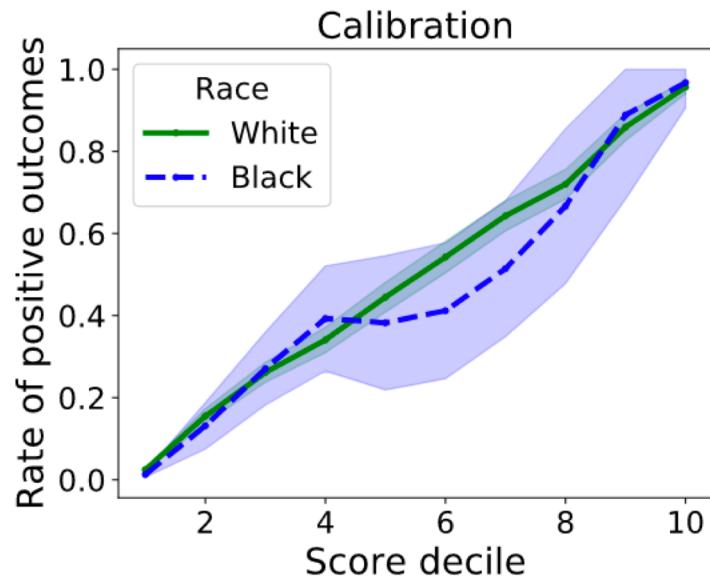
# Sufficiency

$$Y \perp Z | \hat{Y}$$

Example: Logistic regression in UCI Adult data



Well calibrated for gender



Dataset contains only 34 example  
with race equal ,Black'

# Impossibility Results & Trade-offs

- Between different measures of group fairness
- Between fairness and accuracy

# Between measures of fairness

→ Independence vs. Sufficiency

(a.k.a. Demographic parity vs. Predictive Rate)

Proof: If  $Z \not\perp Y$  and  $Z \perp Y|\hat{Y}$ , then  $Z \not\perp \hat{Y}$

→ Independence vs. Separation

(a.k.a. Demographic parity vs. Equalized Odds)

Proof:  $Z \not\perp \hat{Y}$  and  $Z \perp \hat{Y}|Y$ , then either  $Z \perp Y$  or  $\hat{Y} \perp Y$

→ Separation vs. Sufficiency

(a.k.a. Equalized Odds vs. Predictive Rate)

Proof:  $Z \perp \hat{Y}$  and  $Z \perp Y|\hat{Y}$  implies  $Z \perp Y, \hat{Y}$  and, thus  $Z \perp Y$

# Example: Equalized Odds vs. Predictive Rate

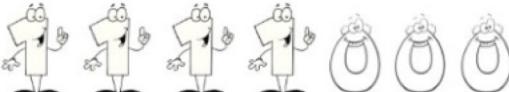
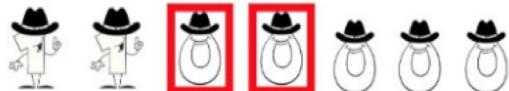
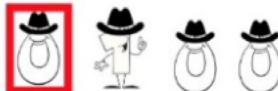
Group	a	b	
Outcome			Unequal base rates
Predictor			

Fig7: illustration of impossibility theorem(original)

False Positive

$$p(\hat{y} = 1|z = 0, y) \neq p(\hat{y} = 1|z = 1, y)$$

# Example: Equalized Odds vs. Predictive Rate

Group	a	b	
Outcome			Unequal base rates
Predictor			
TPR-TNR	1-1/2	1-1/2	

$$p(\hat{y} = 1 | z = 0, y) = p(\hat{y} = 1 | z = 1, y)$$

# Example: Equalized Odds vs. Predictive Rate

Group	a	b	
Outcome			Unequal base rates
Predictor			
NPV	$2/5$	$1/3$	

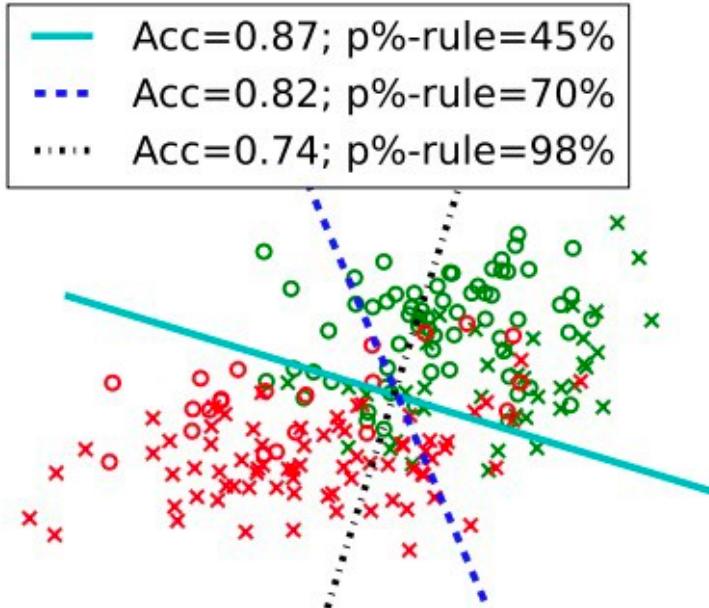
$$p(y|z=0, \hat{y}) \neq p(y|z=1, \hat{y})$$

# Between fairness and accuracy

maximize performance

subject to unfairness  $\leq \alpha$  ←

Trade-off  
fairness-accuracy

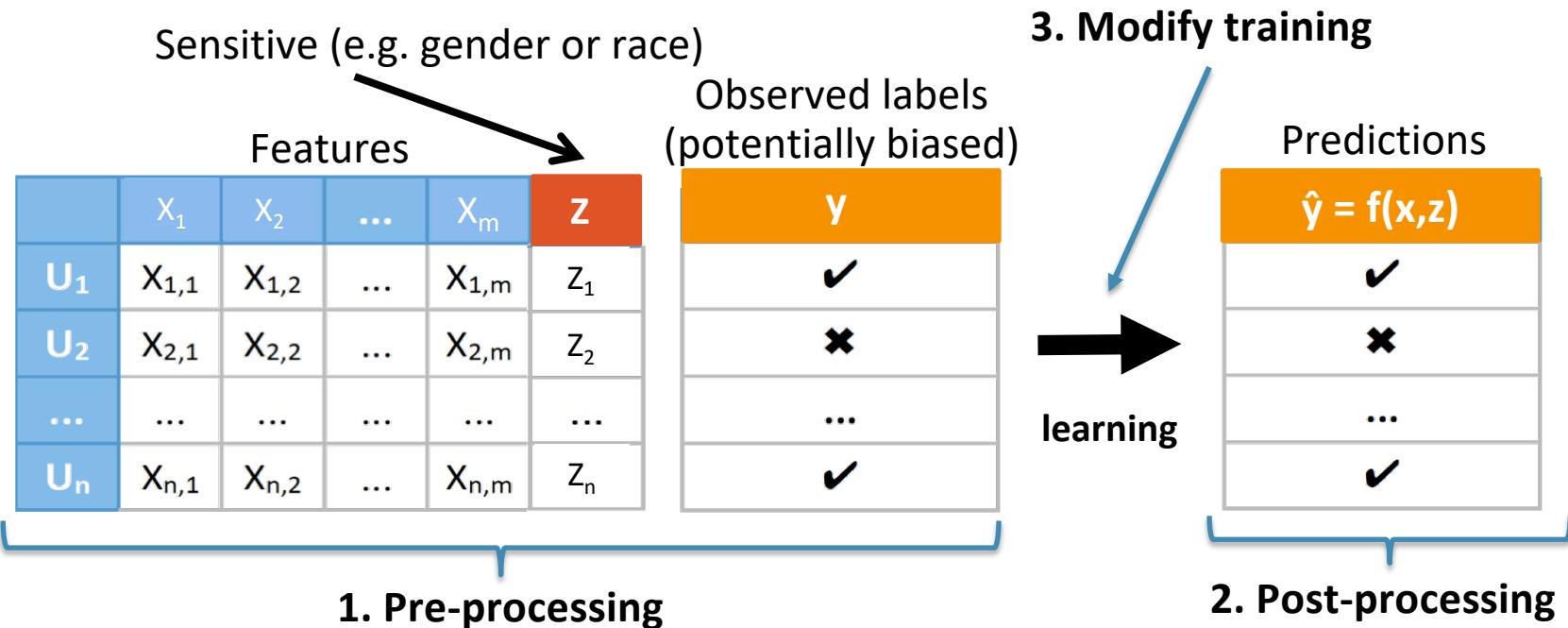


- The impact of imposing fairness constraints depends on the dataset, the fairness definition and the algorithm.
- In general, fairness hurts accuracy because it diverts the objective from accuracy only to both accuracy and fairness.

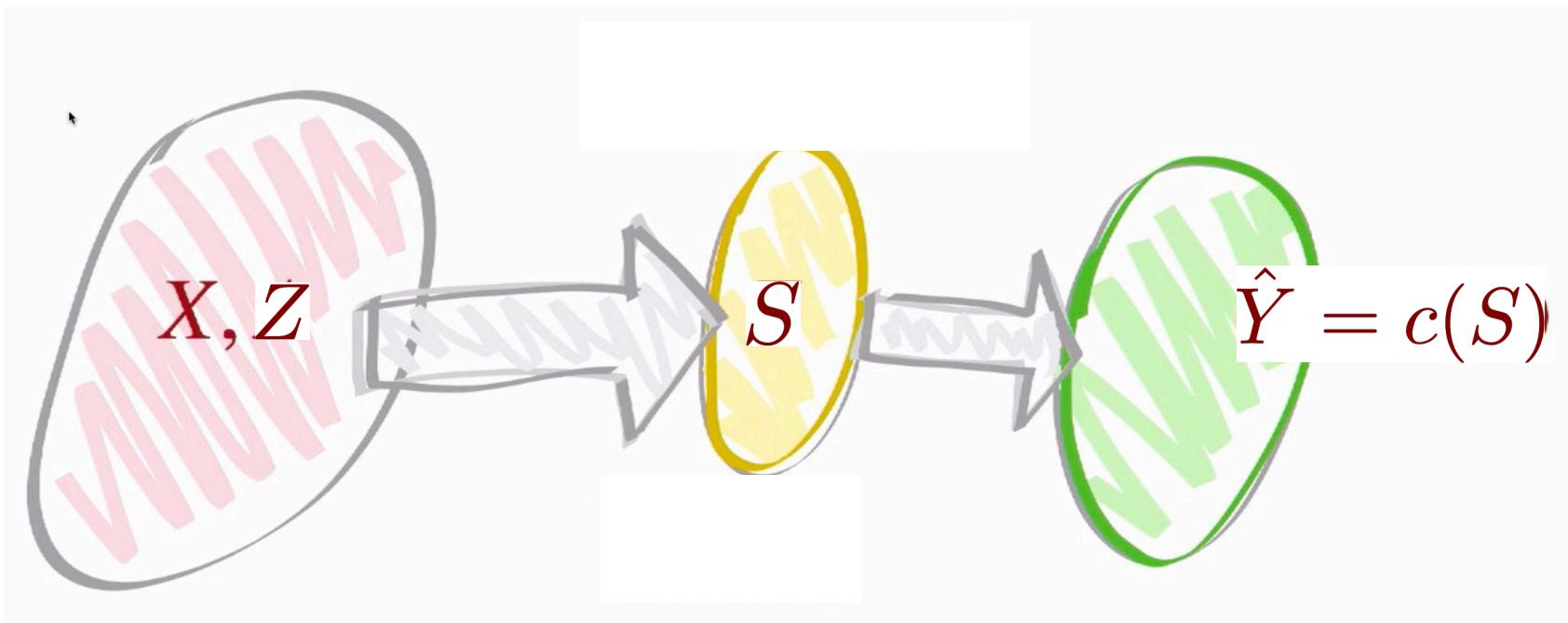
# **Part III**

# **Mechanisms**

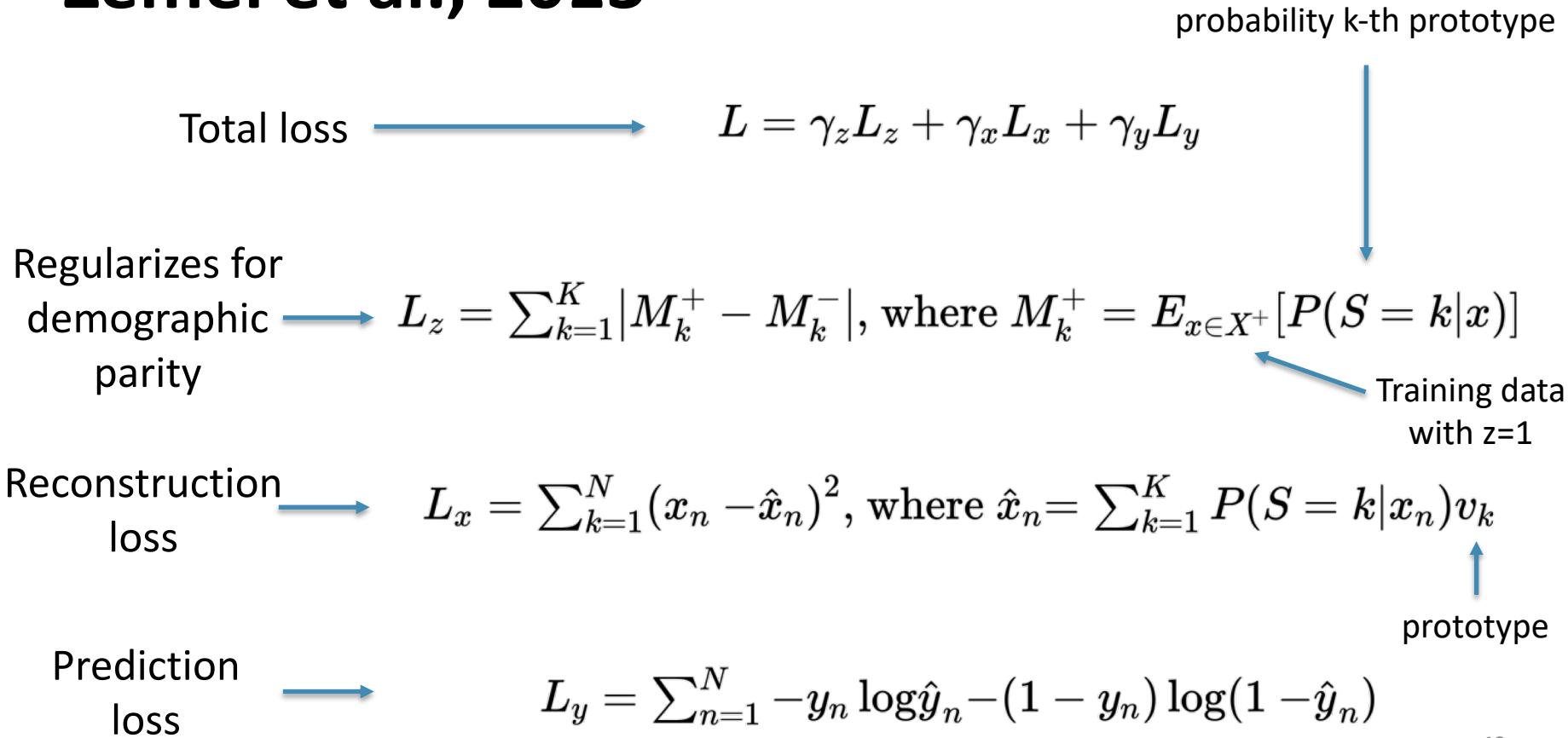
# Mechanisms to achieve group fairness



# 1. Pre-processing



# Zemel et al., 2013



# Pros & Cons

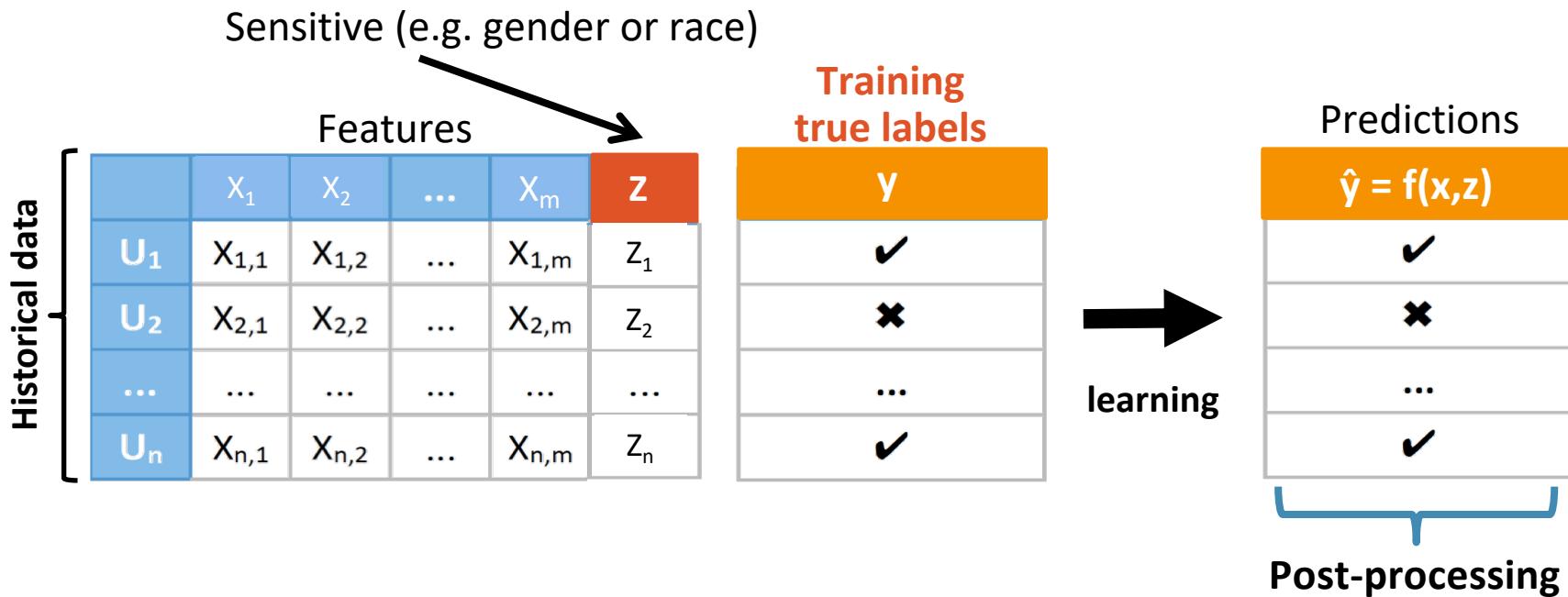
## Pros:

- Preprocessed data can be used for any downstream task.
- No need to modify classifier.
- No need to access sensitive attributes at test time.

## Cons:

- Suitable only for Demopgrapich Parity or Individual Fairness (if a distance metric is given).
- Lack of control of the trade-off between accuracy and fairness. No guarantees of performance in terms of accuracy and fairness.

## 2. Post-processing



# Hardt et al., 2016

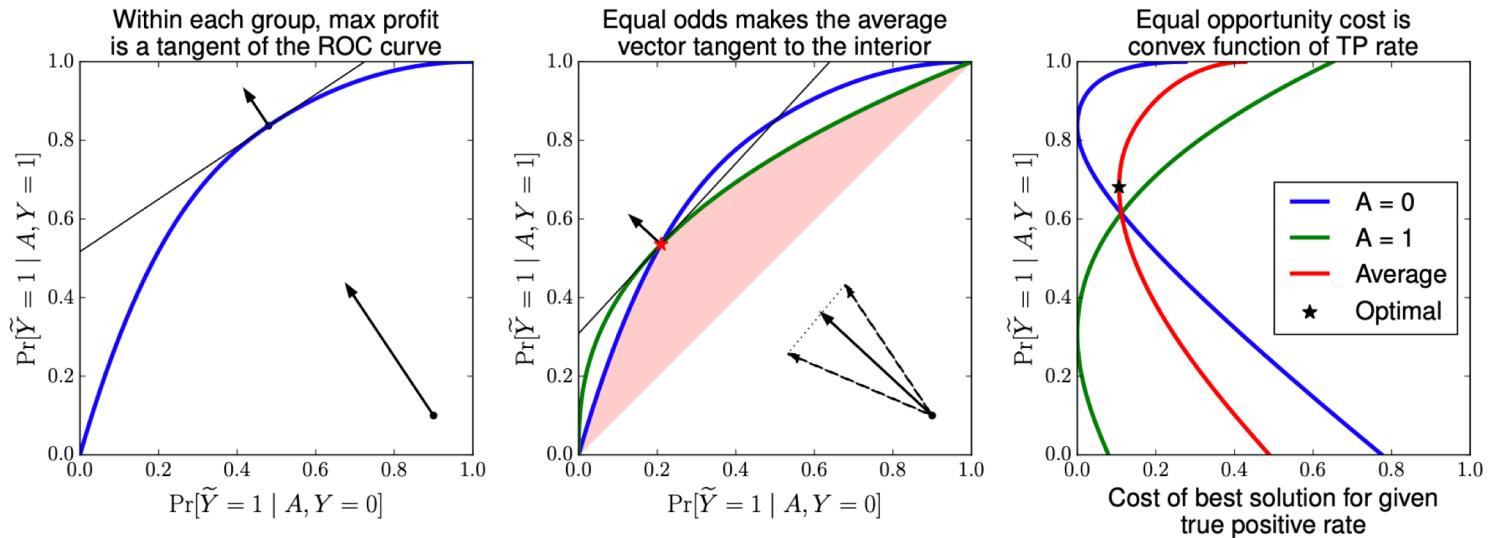


Figure 2: Finding the optimal equalized odds threshold predictor (middle), and equal opportunity threshold predictor (right). For the equal opportunity predictor, within each group the cost for a given true positive rate is proportional to the horizontal gap between the ROC curve and the profit-maximizing tangent line (i.e., the two curves on the left plot), so it is a convex function of the true positive rate (right). This lets us optimize it efficiently with ternary search.

# Pros & Cons

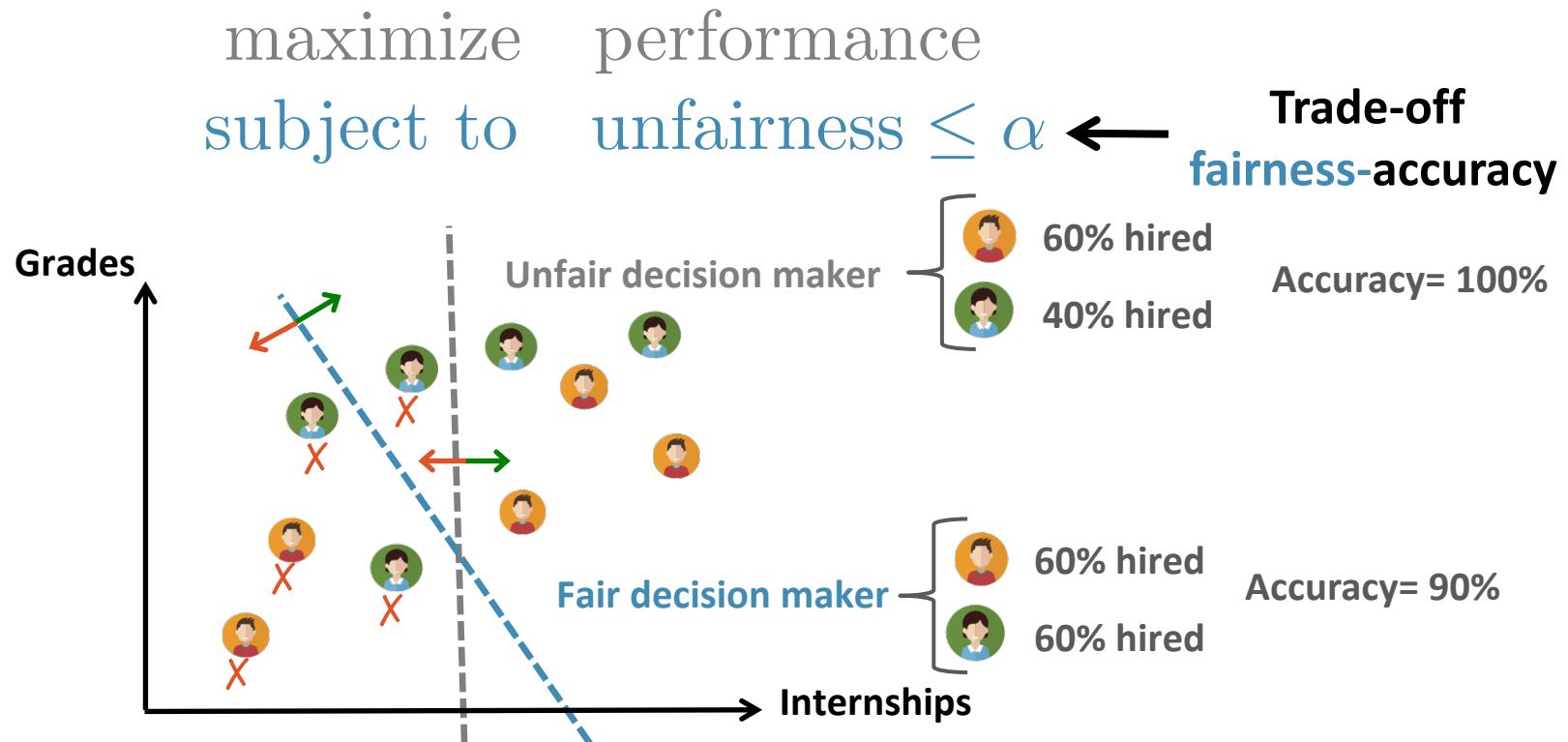
## Pros:

- Can be applied after any classifiers.
- Relatively good performance especially fairness measures.
- No need to modify classifier.

## Cons:

- Require test-time access to the protected attribute.
- Lack the flexibility of picking any accuracy–fairness tradeoff.
- Woodworth et al. (2017) proved that it leads to suboptimal results compared to changing the classifier.

### 3. Modify training



# Training fair margin-based classifiers

**Key idea:** add constraints during training

minimize  $L(\theta)$

subject to  $\underbrace{P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)}$

Free of disparate  
impact

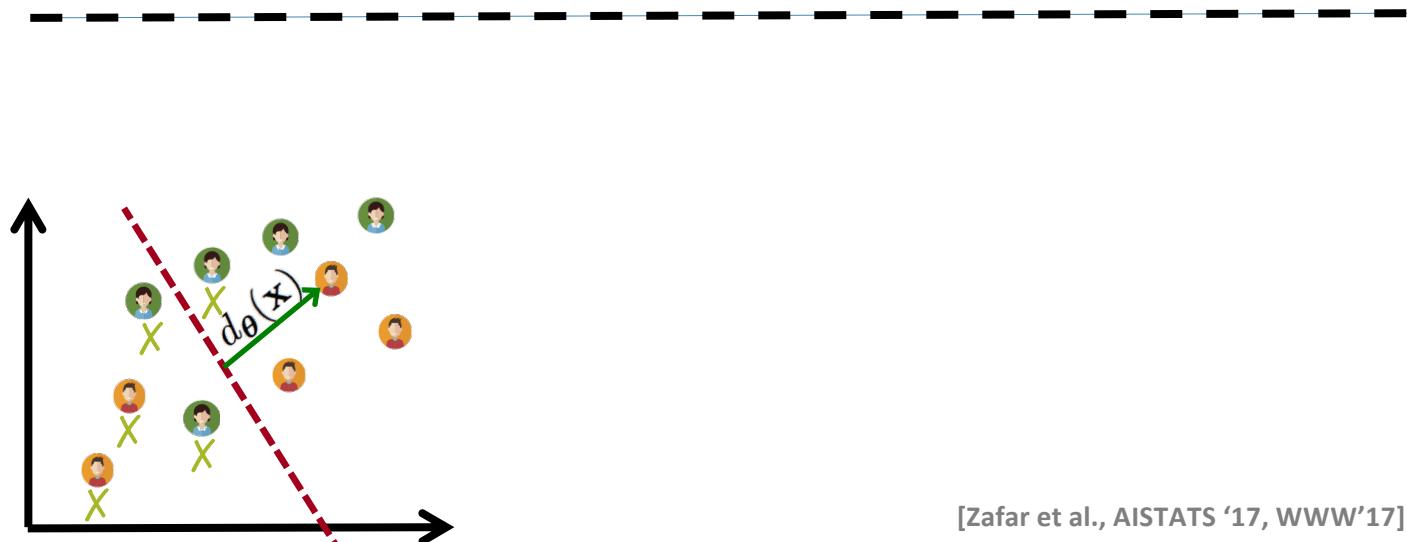
**Challenge:**

1. Difficult to estimate and enforce for many well-known classifiers.
2. It would increase significantly the complexity of training.

# Decision boundary covariance

Disparate  
impact

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$



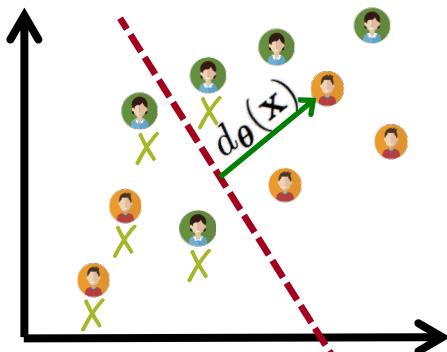
# Decision boundary covariance

Disparate impact

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

$d_{\theta}(\mathbf{x}) \geq 0$      $\text{Cov}(\mathbf{z}, d_{\theta}(\mathbf{x})) = 0$

signed distance



# Decision boundary covariance

Disparate impact

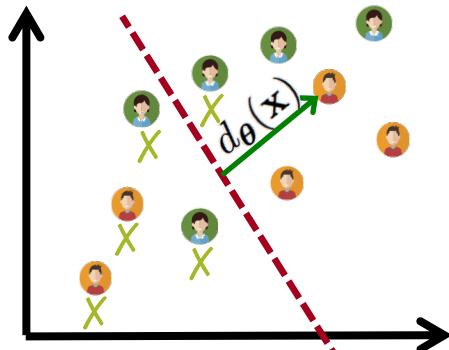
$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

$d_{\theta}(\mathbf{x}) \geq 0$      $\text{Cov}(\mathbf{z}, d_{\theta}(\mathbf{x})) = 0$

signed distance

Disparate mistreatment

$$P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1)$$



# Decision boundary covariance

## Disparate impact

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

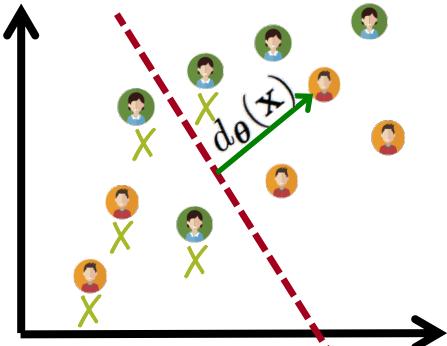
$\overbrace{\quad\quad\quad}$   
 $d_{\theta}(\mathbf{x}) \geq 0 \quad \text{Cov}(z, d_{\theta}(\mathbf{x})) = 0$   
signed distance

## Disparate mistreatment

$$P(\hat{y} \neq y|z = 0) = P(\hat{y} \neq y|z = 1)$$

$\overbrace{\quad\quad\quad}$   
 $y d_{\theta}(\mathbf{x}) \geq 0 \quad \text{Cov}(z, g_{\theta}(y, \mathbf{x})) = 0$   
signed distance

of misclassified points  $\rightarrow \min(0, y d_{\theta}(\mathbf{x}))$



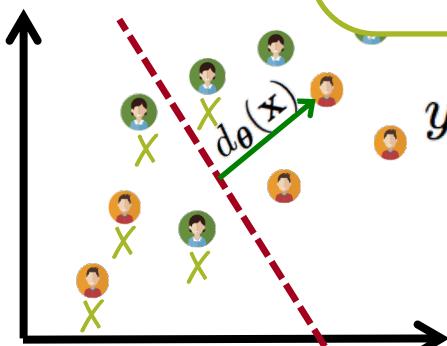
# Decision boundary covariance

Disparate impact

$$P(\hat{y} = 1|z = 0) = P(\hat{y} = 1|z = 1)$$

Disparate mistreatment

Covariances are convex or convex-concave for many well-known classifiers!



$$d_{\theta}(\mathbf{x}) \geq 0$$

$$\text{Cov}(z, g_{\theta}(y, \mathbf{x})) = 0$$

signed distance of misclassified points →  $\min(0, yd_{\theta}(\mathbf{x}))$

# Example (I): Disparate Impact in Logistic Regression

$$p(y_i = 1 | \mathbf{x}_i, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^T \mathbf{x}_i}}$$

↓

minimize  $- \sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \boldsymbol{\theta})$

subject to  $\frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \boldsymbol{\theta}^T \mathbf{x}_i \leq \mathbf{c},$   
 $\frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \boldsymbol{\theta}^T \mathbf{x}_i \geq -\mathbf{c}$

**Boundary covariance  
constraints  
(Linear on  $\boldsymbol{\theta}$  !)**

*It is a convex  
program!*

# Example (II): Disparate Mistreatment in LR

$$p(y_i = 1 | \mathbf{x}_i, \theta) = \frac{1}{1 + e^{-\theta^T \mathbf{x}_i}}$$

minimize  $-\sum_{i=1}^N \log p(y_i | \mathbf{x}_i, \theta)$

subject to

$$\begin{aligned} & \frac{-N_1}{N} \sum_{(\mathbf{x}, y) \in \mathcal{D}_0} g_\theta(y, \mathbf{x}) \\ & + \frac{N_0}{N} \sum_{(\mathbf{x}, y) \in \mathcal{D}_1} g_\theta(y, \mathbf{x}) \leq c \\ & \frac{-N_1}{N} \sum_{(\mathbf{x}, y) \in \mathcal{D}_0} g_\theta(y, \mathbf{x}) \\ & + \frac{N_0}{N} \sum_{(\mathbf{x}, y) \in \mathcal{D}_1} g_\theta(y, \mathbf{x}) \geq -c \end{aligned}$$

Only misclassifications

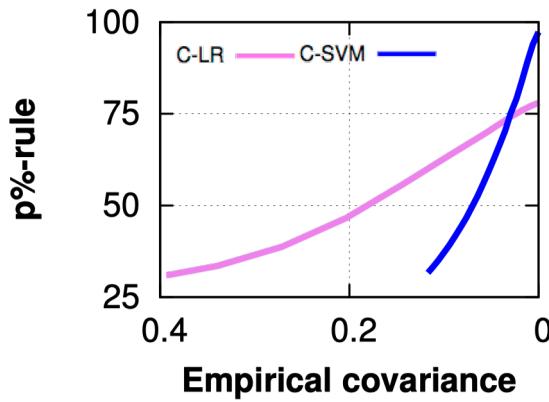
$$\min(0, y d_\theta(\mathbf{x}))$$

(Convex on  $\theta$  !)

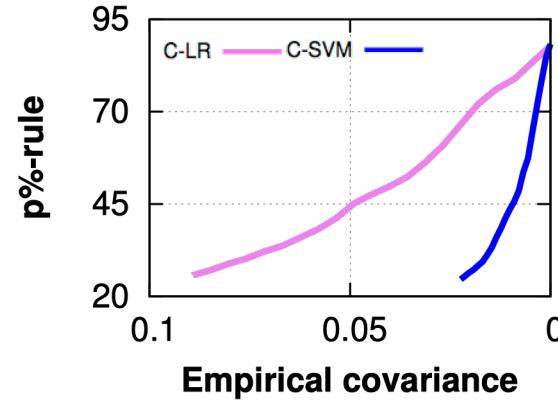
Boundary covariance constraints  
(Convex-concave on  $\theta$ )

It is a convex-concave program!

# Is the covariance a good relaxation?



Adult dataset



Bank dataset

The lower  
the covariance



$$\text{Cov}(z, g_{\theta}(y, \mathbf{x}))$$

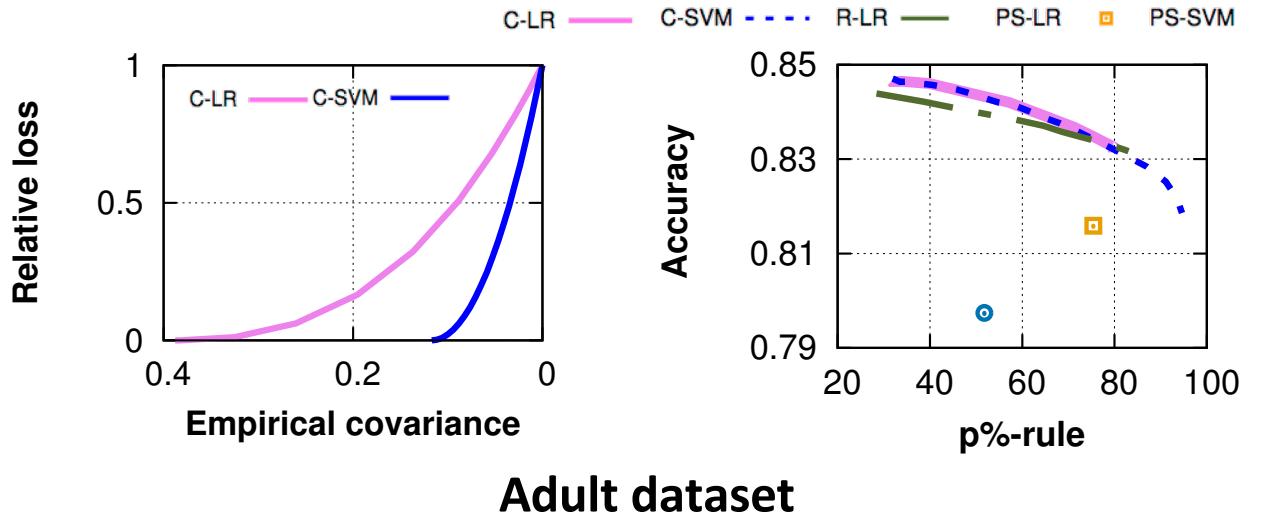


The higher the p%-rule  
(the higher the parity in impact)



$$P(\hat{y} \neq y | z = 0, y = -1) - P(\hat{y} \neq y | z = 1, y = -1)$$

# How much does fairness cost?



1. Achieving parity in impact comes often at small cost in terms of accuracy
2. The tradeoff between loss-covariance is (Pareto) optimal

# Pros & Cons

## Pros:

- Good performance on accuracy and fairness measures.
- Higher flexibility to choose the trade-off between accuracy and fairness measures (depending on specific algorithm).
- No need to access sensitive attributes at test time.

## Cons:

- Method in this category is task and algorithm specific.
- Need to modify classifier.
- Rely on relaxations of fairness notions (e.g., decision boundary covariance)

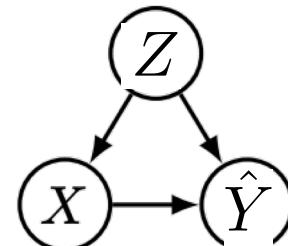
# Part IV

# Fairness Beyond Predictions

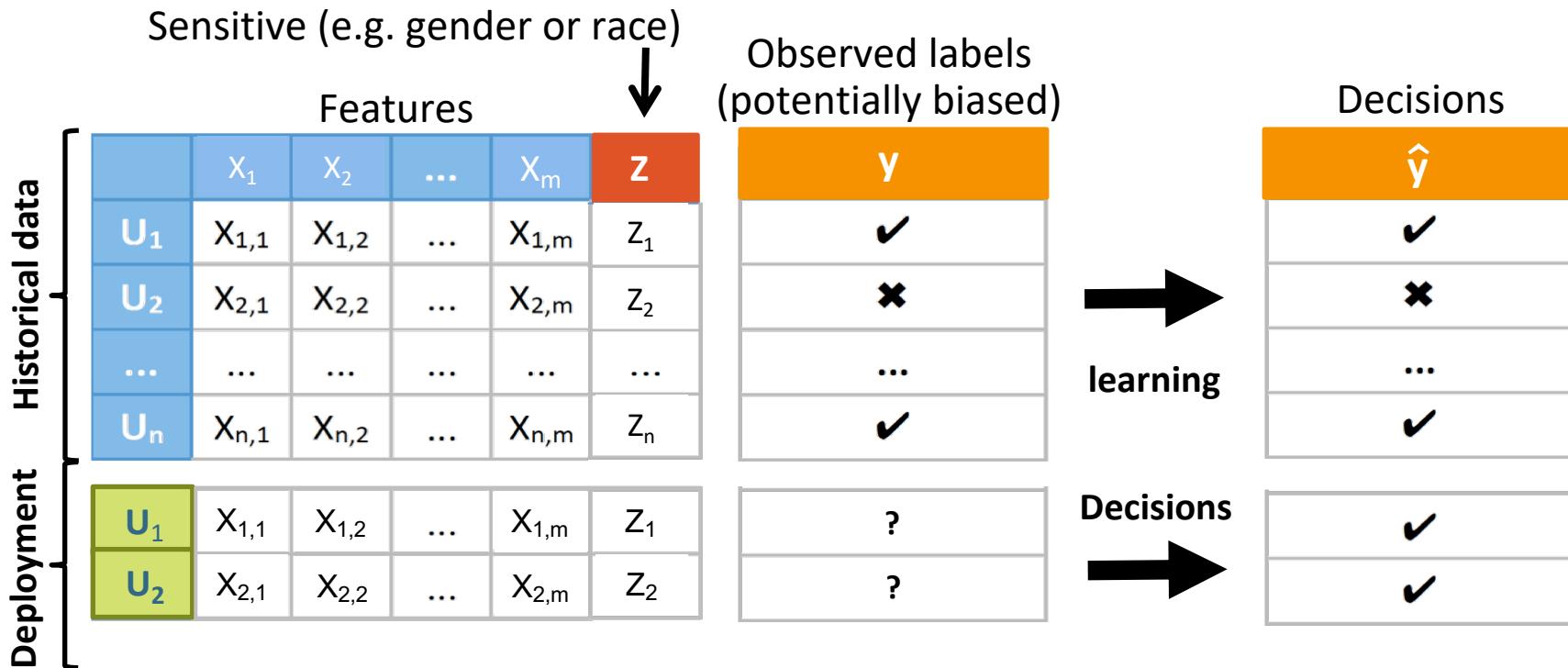
# So far all fairness criteria are observational

- Passive observations
- All criteria based on the joint distribution between features  $X, Z$ , predictions  $\hat{Y}$  and labels  $Y$ .
- Causal reasoning:
  - Paths in causal graphs (Kilbertus et al., 2017)
  - Impact of interventions (Kusner et al., 2017)

Example (Bickel et al., 1975):  
Berkeley college admissions

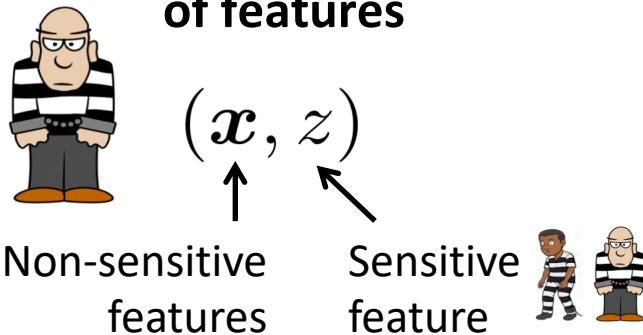


# Fair Machine Learning



# Decisions may not match labels

Each defendant has a set of features



Bail decisions based on defendant features

$$d(x, z) = \begin{cases} 1 & \text{Defendant is released} \\ 0 & \text{Defendant remains in jail} \end{cases}$$

which determines whether the label  $y$  is *observed*. Here,  $y = 1$  if defendant *does not reoffend* and  $y = 0$ , if *reoffends*.

Benefits (for group  $z$ ):

$$b_z(d) = \mathbb{E}_X [1 - d(X, Z = z)]$$



Proportional to  
# people released

Utility (for the justice system):

$$u(d) = \mathbb{E}_{Y, X, Z} [Y d(X, Z) - c d(X, Z)]$$

Proportional to  
# crimes prevented

Proportional to  
# people detained

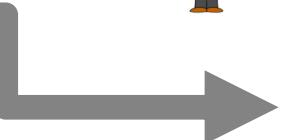
# Optimal (deterministic) decision rules

$$\text{maximize } u(d) = \mathbb{E}_{X,Z}[d(X, Z)(P(Y = 1|X, Z) - c)]$$

Without  
Fairness


$$d^*(X, Z) = \begin{cases} 1 & \text{Defendant is released if } P(Y = 1|X, Z) \geq c \\ 0 & \text{otherwise.} \end{cases}$$

With  
Fairness


$$\text{maximize } u(d) : |b_0(d) - b_1(d)| \leq \alpha$$

known

It depends on the fairness notion

$$d^*(X, Z) = \begin{cases} 1 & \text{if } P(Y = 1|X, Z) \geq \theta_Z \\ 0 & \text{otherwise,} \end{cases}$$

# Algorithmic (fair) decision making

	$X_1$	$X_2$	...	$X_m$	$Z$
$U_1$	$X_{1,1}$	$X_{1,2}$	...	$X_{1,m}$	$Z_1$
$U_2$	$X_{2,1}$	$X_{2,2}$	...	$X_{2,m}$	$Z_2$
...	...	...	...	...	...
$U_n$					

LEARNING

$$p(\hat{y}|\mathbf{x}, z)$$

DEPLOYMENT

$$\pi(d = 1|\mathbf{x}, z) = \mathbf{1}[p(\hat{y}|\mathbf{x}, z) \geq c]$$

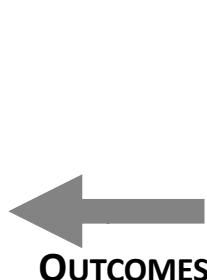
*p(y,*

**Data are not collected from the true population  
but from a distribution induced by the policy.**

$X_{1,m}$	$Z_1$
$X_{2,m}$	$Z_2$



y
✓
✗
?
?



d
✓
✓
✗
✗



# Theoretical results

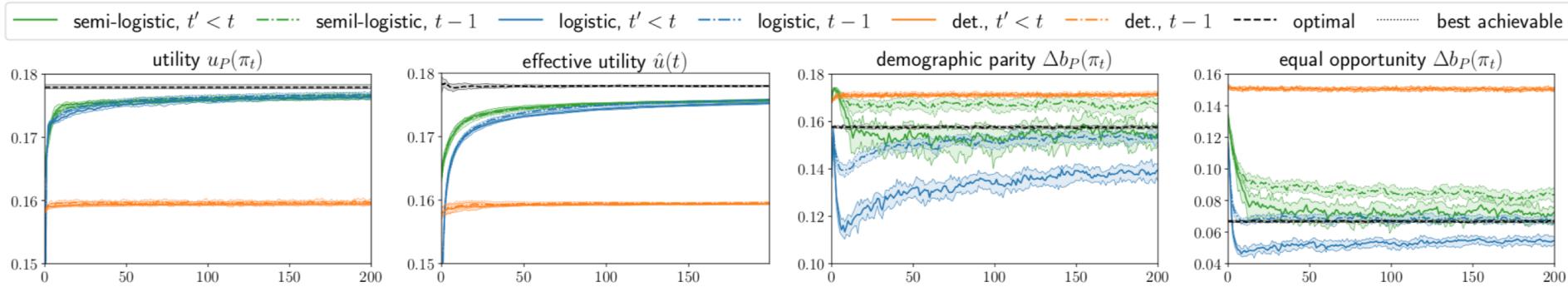
**Theorem 1.** New policies learned from data collected from **deterministic polices\***, such as a decision threshold rule, will be **suboptimal** in both utility and fairness.

*\*More strict than the optimal policy, i.e., provides a negative decision where the optimal policy decides 1.*

**Theorem 2.** New policies learned from data collected from **stochastic exploring policies\*\***, will be **optimal** in both utility and fairness.

*\*A exploring policy satisfy that  $\pi(d = 1 | x, z) > 0$  on any measurable subset of  $X \times Z$  with positive probability under the true population distribution.*

# Empirical results



- Deterministic policies do not only lead to suboptimal utility but also amplify initial biases (stereotypes) in the dataset (original policy).
- Exploring policies lead to optimal utility and fairness level.

# Some take-home messages

- **Observational criteria and approaches** useful but not enough.
- **Causal reasoning** may help but it is a problem itself.
- Observed features are **indirect, noisy and potentially biased** measurements of the “state of the world”. How these features have been **measured?** (Grgić-Hlača et al., 2016)
- We should think **beyond predictions and account for our interventions (decisions).**