

Part I: Introduction and context

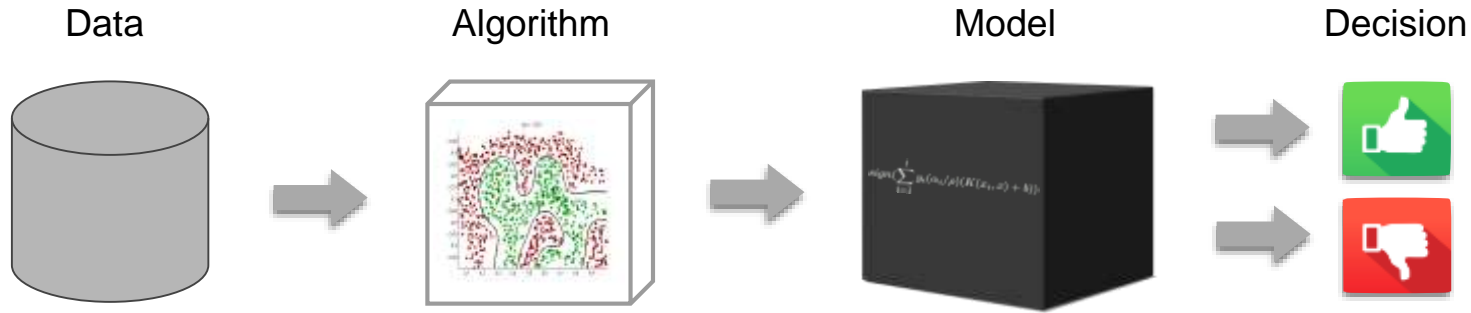
Part II: Discrimination discovery

 **Part III: Fairness-aware data mining**

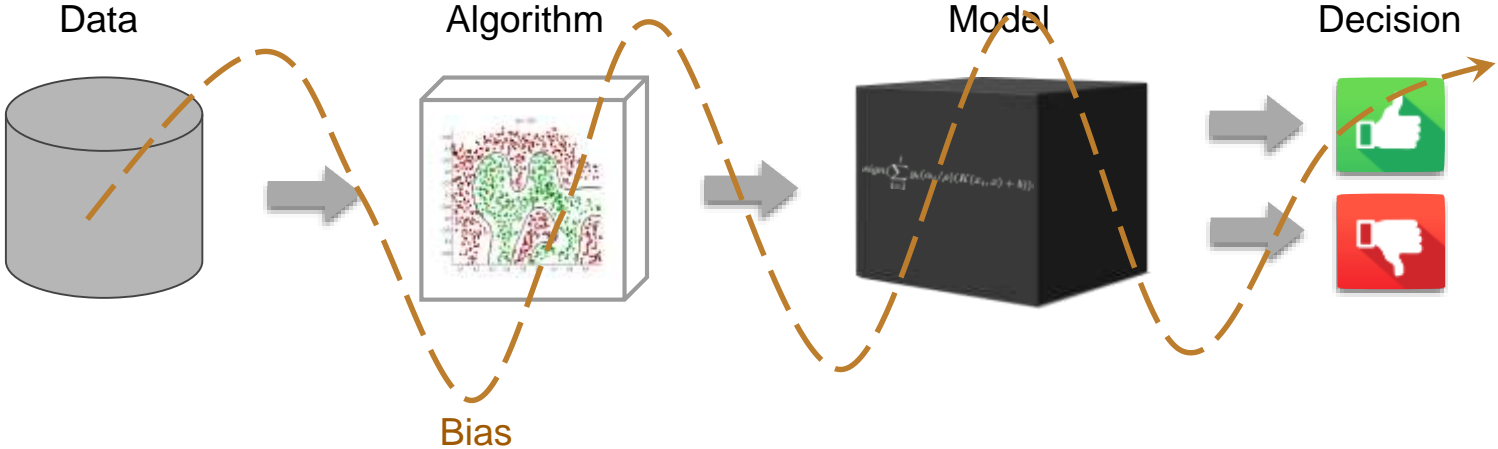
Part IV: Challenges and directions for future research

Discussion and further questions

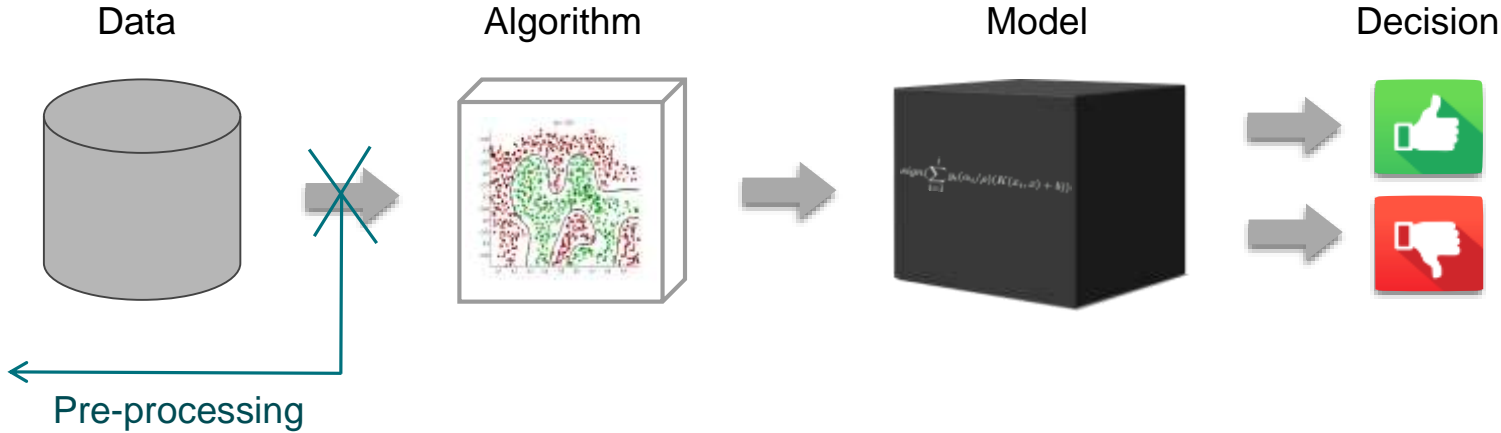
# Non-discriminatory data-driven decision-making



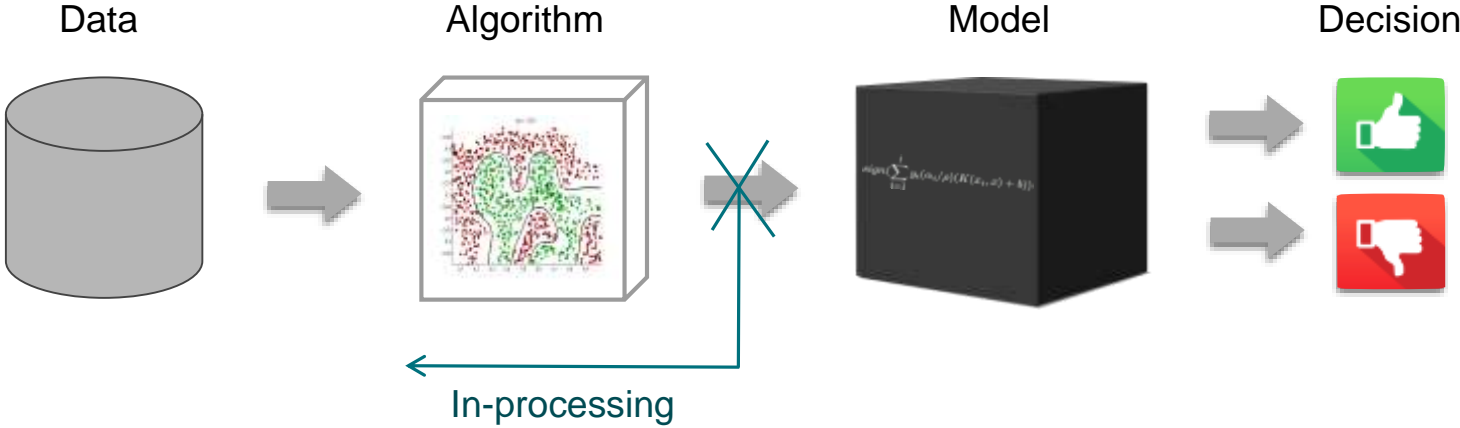
# Non-discriminatory data-driven decision-making



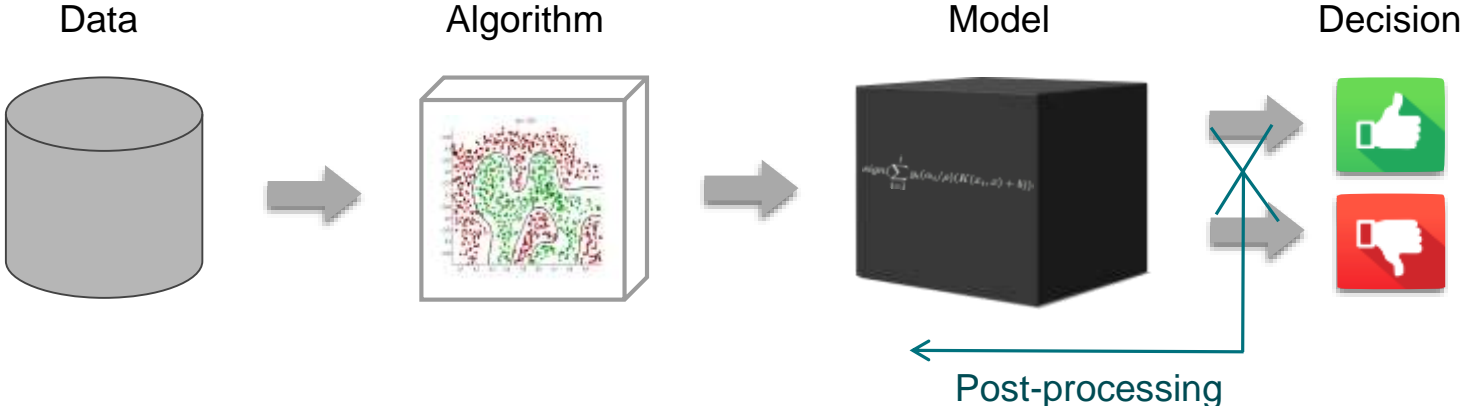
# Non-discriminatory data-driven decision-making



# Non-discriminatory data-driven decision-making



# Non-discriminatory data-driven decision-making



# Fairness-aware data mining: common aspects

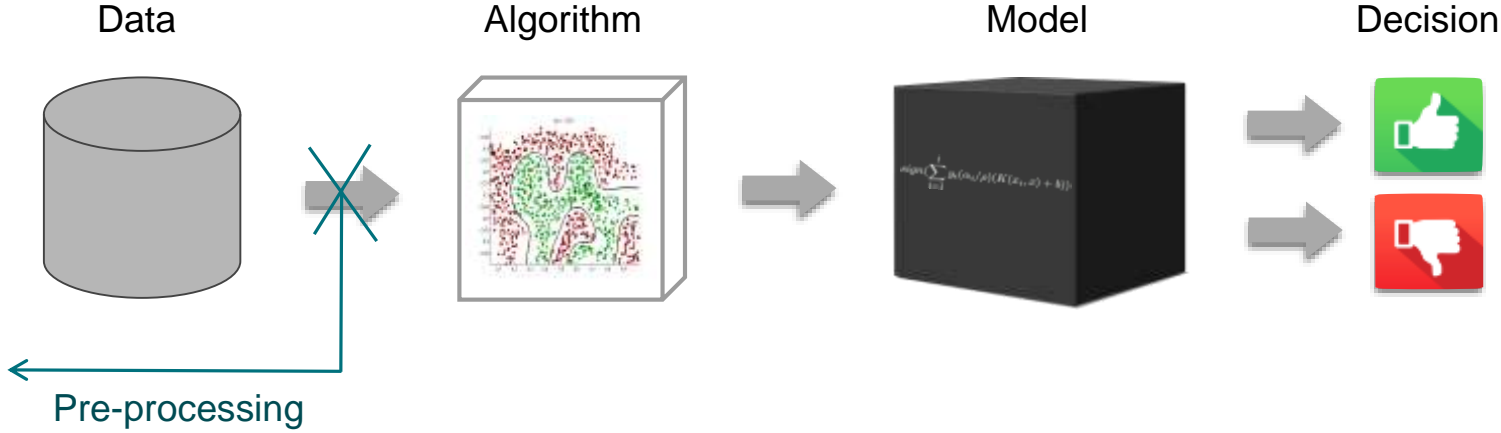
Goal: develop a non-discriminatory decision-making process while preserving as much as possible the quality of the decision.



## Steps:

- (1) Defining anti-discrimination/fairness constraints
- (2) Transforming data/algorithm/model to satisfy the constraints
- (3) Measuring data/model utility

# Non-discriminatory data-driven decision-making





# Fairness-aware data mining

## Pre-processing approaches:

**[Pre\_1]** M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *“Certifying and removing disparate impact”*. In KDD, pp. 259-268, 2015.

**[Pre\_2]** F. Kamiran and T. Calders. *“Data preprocessing techniques for classification without discrimination”*. In Knowledge and Information Systems (KAIS), 33(1), 2012.

**[Pre\_3]** S. Hajian and J. Domingo-Ferrer. *“A methodology for direct and indirect discrimination prevention in data mining”*. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

**[Pre\_4]** I. Zliobaite, F. Kamiran and T. Calders. *“Handling conditional discrimination”*. In ICDM, pp. 992-1001, 2011.

# Fairness-aware data mining

## Pre-processing approaches:

**[Pre\_1]** M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. *“Certifying and removing disparate impact”*. In KDD, pp. 259-268, 2015.

**[Pre\_2]** F. Kamiran and T. Calders. *“Data preprocessing techniques for classification without discrimination”*. In Knowledge and Information Systems (KAIS), 33(1), 2012.

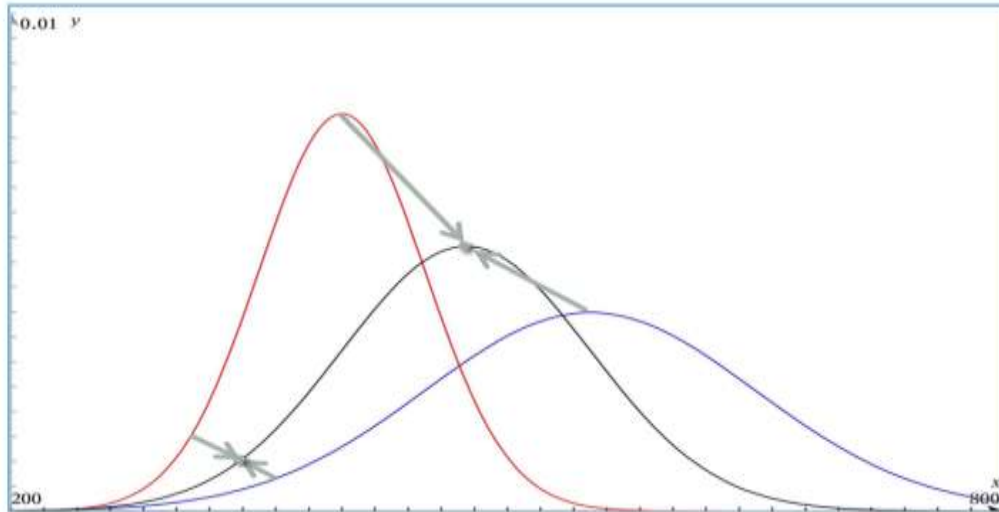
**[Pre\_3]** S. Hajian and J. Domingo-Ferrer. *“A methodology for direct and indirect discrimination prevention in data mining”*. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

**[Pre\_4]** I. Zliobaite, F. Kamiran and T. Calders. *“Handling conditional discrimination”*. In ICDM, pp. 992-1001, 2011.

Given  $D = (X, Y, C)$  which has been certified having disparate impact potential, where  $X$  is protected attribute,  $Y$  the remaining attributes, and  $C$  is the decision class.

Generate a dataset  $D_{\text{new}} = (X, Y_{\text{new}}, C)$  with no disparate impact.

Utility goal: to preserve rank within each marginal distribution  $P(Y | X = x)$



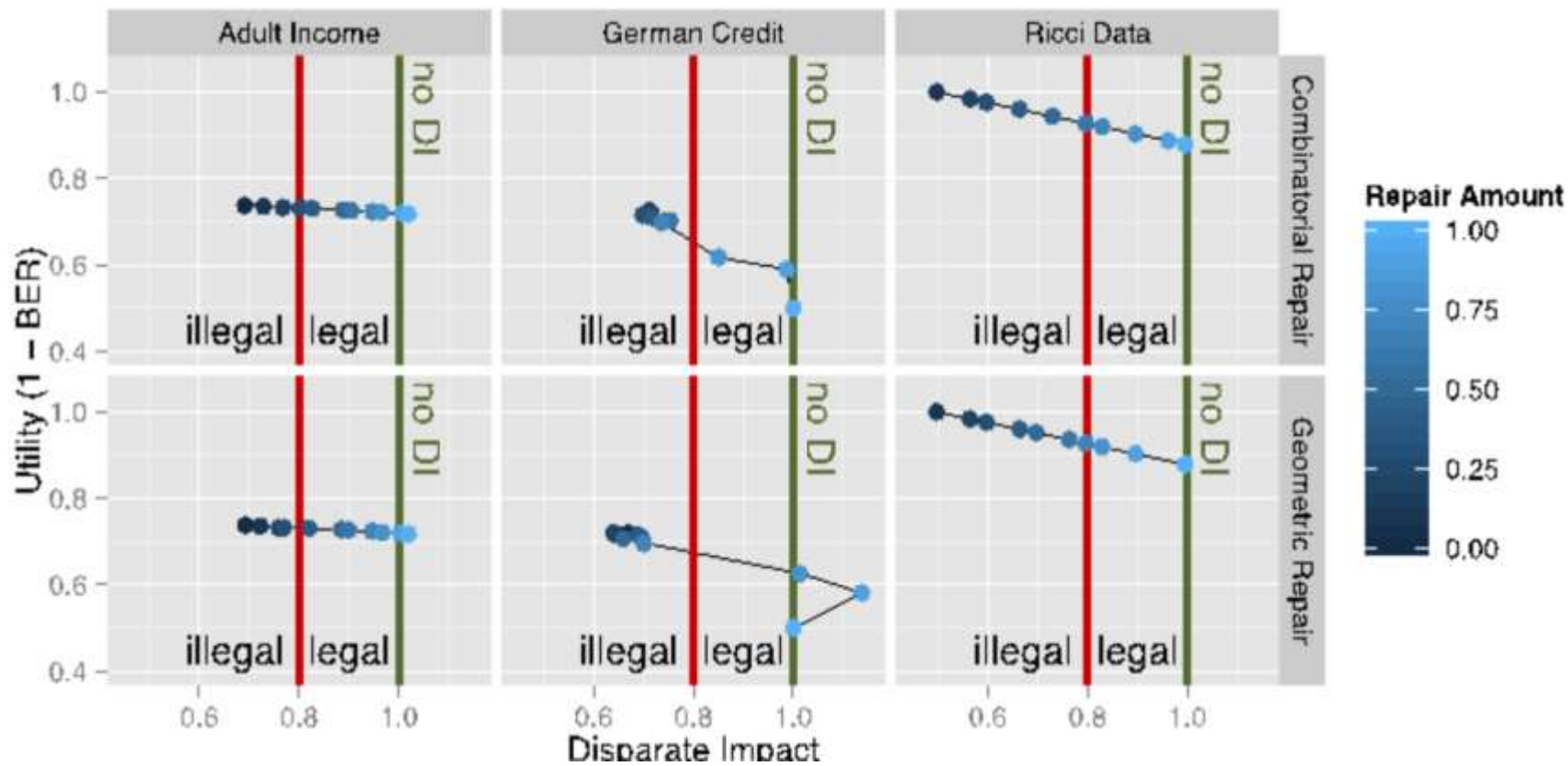
Using the earthmover distance

$$\text{Let } P_i = \Pr(Y = y | X = i)$$
$$F_i = \text{cdf of } P_i$$

$$P_* = \arg \min \sum_i d_{EM}(P, P_i)$$

$$F_*^{-1}(\lambda) = \text{median } F_i^{-1}(\lambda)$$

We find a new distribution that is “close” to all conditional distributions.



# Fairness-aware data mining

## Pre-processing approaches:

[Pre\_1] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. “*Certifying and removing disparate impact*”. In KDD, pp. 259-268, 2015.

[Pre\_2] F. Kamiran and T. Calders. “*Data preprocessing techniques for classification without discrimination*”. In Knowledge and Information Systems (KAIS), 33(1), 2012.

[Pre\_3] S. Hajian and J. Domingo-Ferrer. “*A methodology for direct and indirect discrimination prevention in data mining*”. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

[Pre\_4] I. Zliobaite, F. Kamiran and T. Calders. “*Handling conditional discrimination*”. In ICDM, pp. 992-1001, 2011.

# “Data preprocessing techniques for classification without discrimination” (Kamiran and Calders, 2012)

## Discriminatrion measure:

$$P(C = + | B = b_1 ) - P(C = + | B = b_2 )$$

Goal: minimize discrimination, while maximazing accuracy

## Techniques for removing dependencies from the input data:

1. Suppression (baseline, just remove B and the top-k attributes most correlated with B)
2. Massaging
3. Reweighing
4. Sampling

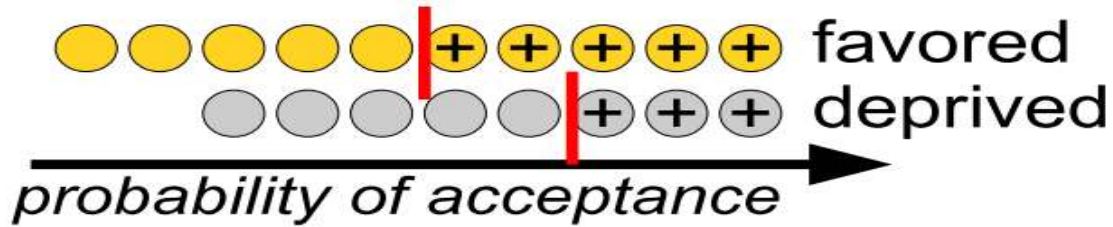
## Job application example

Sex	Ethnicity	Highest Degree	Job Type	Class
m	native	h. school	board	+
m	native	univ.	board	+
m	native	h. school	board	+
m	non-nat.	h. school	healthcare	+
m	non-nat.	univ.	healthcare	-
f	non-nat.	univ.	education	-
f	native	h. school	education	-
f	native	none	healthcare	+
f	non-nat.	univ.	education	-
f	native	h. school	board	+

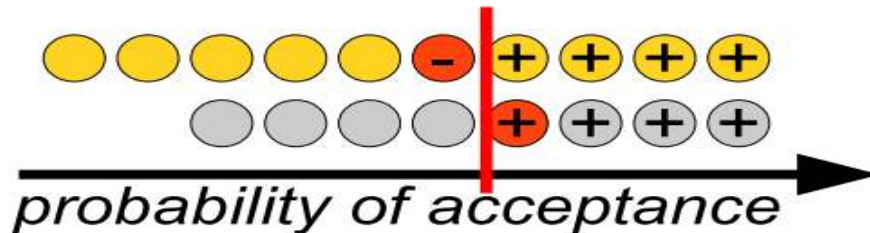


# Massaging

a) rank individuals



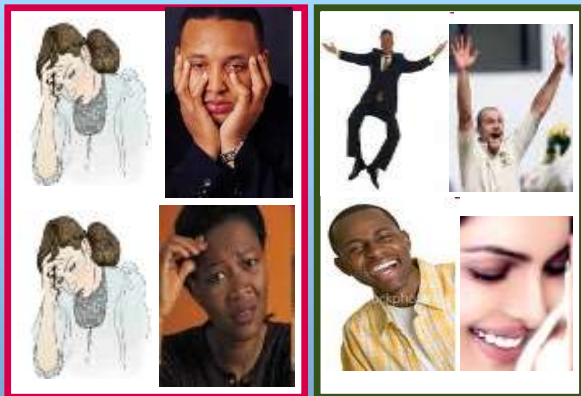
b) change the labels



# Input dataset

Job=No

Job=Yes

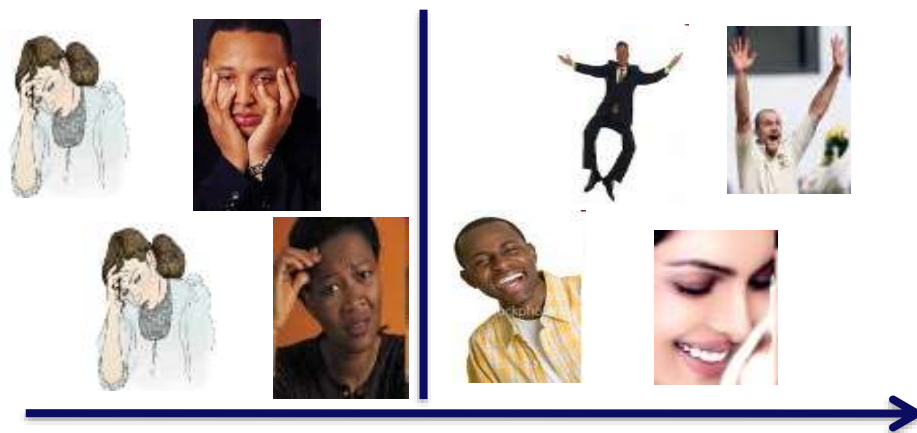


# Decision boundary

Learn a



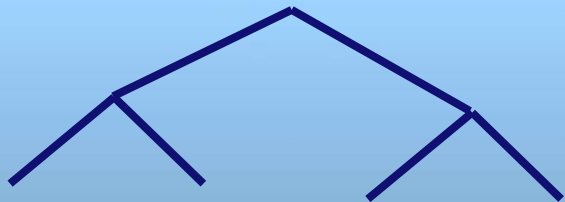
ranker



Relabel



# Final Model



Learn a



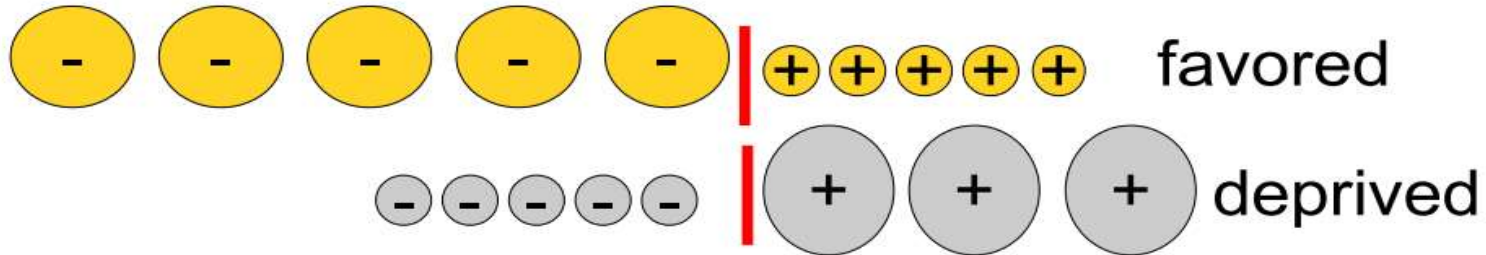
Classifier



# Reweighting

a) calculate weights for the objects to neutralize the discriminatory effects from data

b) assign weights to make the data impartial



# Reweighting

$$W(x(B) = b \mid x(Class) = +) := \frac{P_{exp}(b \wedge +)}{P_{act}(b \wedge +)}$$

$$W(x(B) = b \mid x(Class) = -) := \frac{P_{exp}(b \wedge -)}{P_{act}(b \wedge -)}$$

$$W(x(B) = \bar{b} \mid x(Class) = +) := \frac{P_{exp}(\bar{b} \wedge +)}{P_{act}(\bar{b} \wedge +)}$$

$$W(x(B) = \bar{b} \mid x(Class) = -) := \frac{P_{exp}(\bar{b} \wedge -)}{P_{act}(\bar{b} \wedge -)}$$

Sex	Ethnicity	Highest Degree	Job Type	Cl.	Weight
m	native	h. school	board	+	0.75
m	native	univ.	board	+	0.75
m	native	h. school	board	+	0.75
m	non-nat.	h. school	healthcare	+	0.75
m	non-nat.	univ.	healthcare	-	2
f	non-nat.	univ.	education	-	0.67
f	native	h. school	education	-	0.67
f	native	none	healthcare	+	1.5
f	non-nat.	univ.	education	-	0.67
f	native	h. school	board	+	1.5

$$P_{exp}(Sex = f \mid x(Class) = +) = 0.5 \times 0.6$$

$$W(Sex = f \mid x(Class) = +) = \frac{0.5 \times 0.6}{0.2} = 1.5$$

$$W(Sex = f \mid x(Class) = -) = 0.67$$

$$W(Sex = m \mid x(Class) = +) = 0.75$$

$$W(Sex = m \mid x(Class) = -) = 2 .$$

# Sampling

Similarly to reweighing, compare the expected size of a group with its actual size, to define a sampling probability.

$$DP := \{x \in D \mid x(B) = b \wedge x(Class) = +\}$$

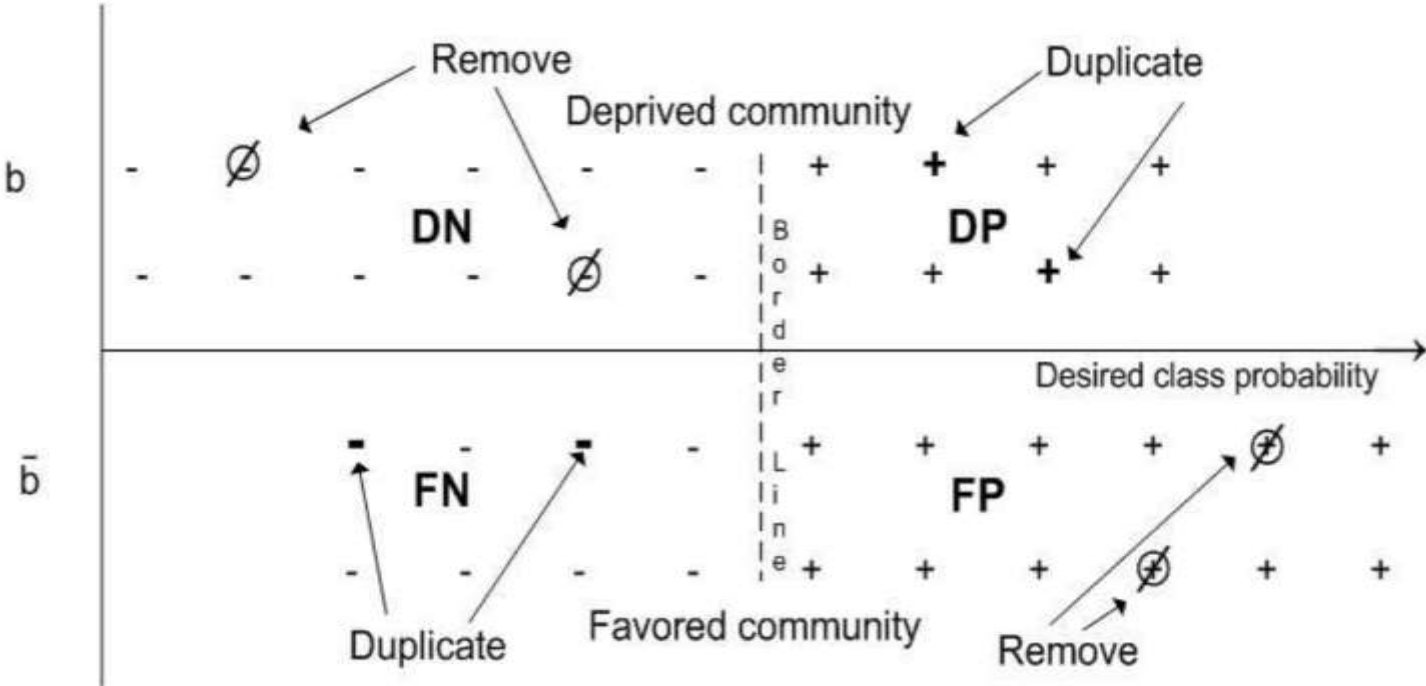
$$DN := \{x \in D \mid x(B) = b \wedge x(Class) = -\}$$

$$FP := \{x \in D \mid x(B) = \bar{b} \wedge x(Class) = +\}$$

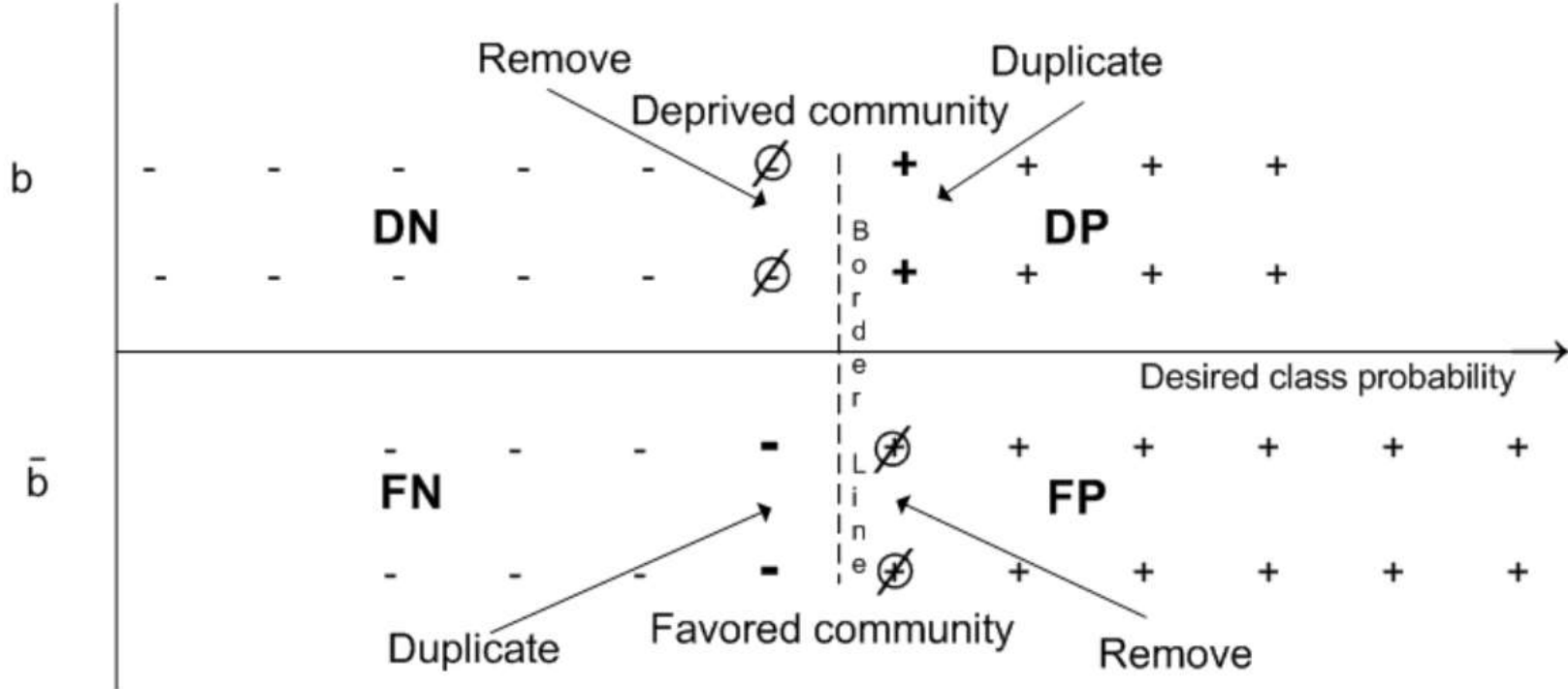
$$FN := \{x \in D \mid x(B) = \bar{b} \wedge x(Class) = -\}$$

Then sample accordingly, possibly duplicating data points.

# Uniform Sampling



# Preferential Sampling



# Performance

Preprocess method	Disc (%)	Acc (%)
No	$16.4 \pm 1.31$	$86.05 \pm 0.29$
No_SA	$16.6 \pm 1.43$	$86.01 \pm 0.31$
RW	$7.97 \pm 1.02$	$85.62 \pm 0.30$
US	$7.91 \pm 2.05$	$85.35 \pm 0.36$
PS	$3.08 \pm 0.79$	$84.30 \pm 0.25$
M_NBS	$1.77 \pm 1.16$	$83.65 \pm 0.24$
M_J48	$2.49 \pm 1.92$	$83.49 \pm 0.47$
M_IBk1	$7.67 \pm 0.86$	$85.35 \pm 0.46$
M_IBk2	$3.62 \pm 0.61$	$84.44 \pm 0.27$
M_IBk3	$2.40 \pm 0.51$	$83.78 \pm 0.43$



# Fairness-aware data mining

## Pre-processing approaches:

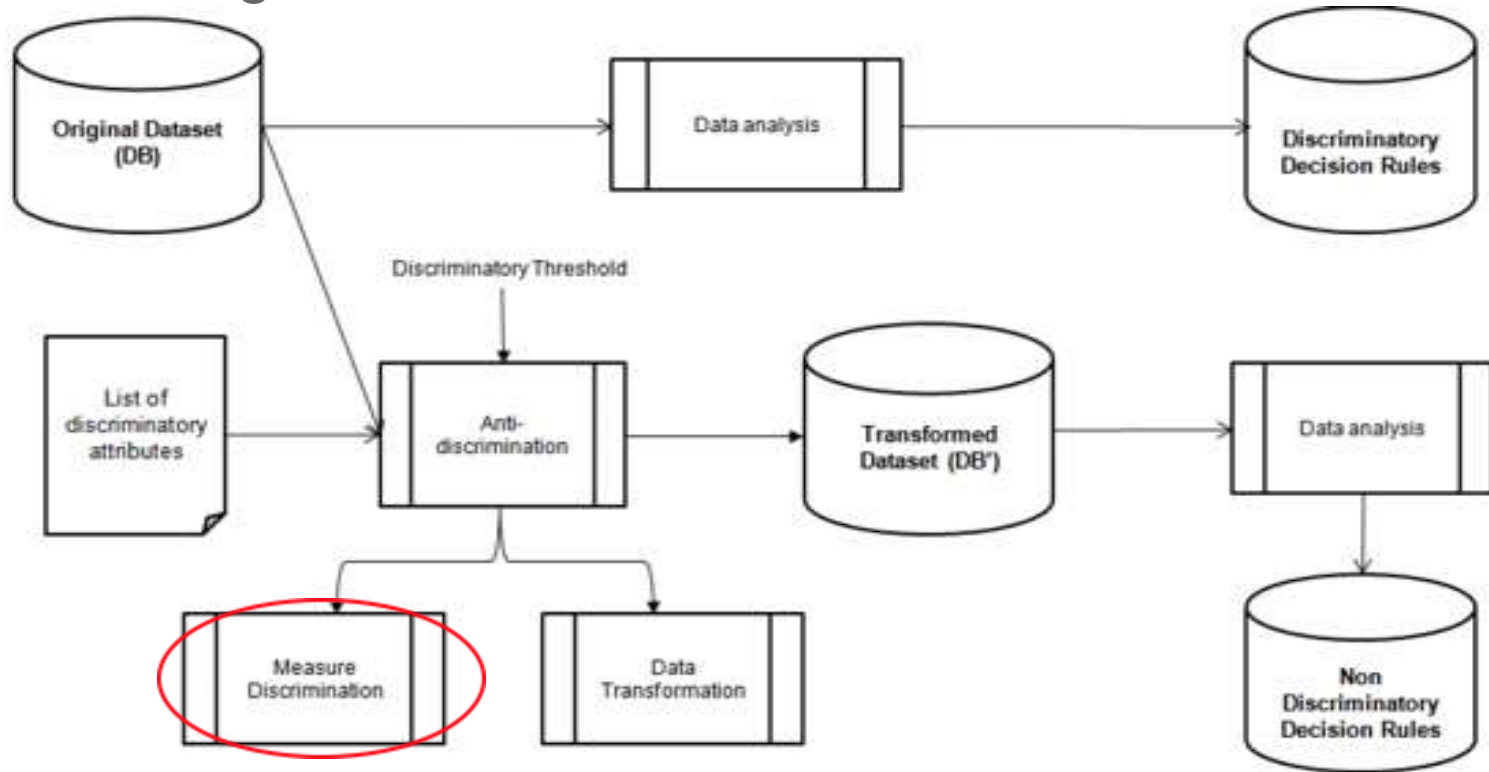
[Pre\_1] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. “*Certifying and removing disparate impact*”. In KDD, pp. 259-268, 2015.

[Pre\_2] F. Kamiran and T. Calders. “*Data preprocessing techniques for classification without discrimination*”. In Knowledge and Information Systems (KAIS), 33(1), 2012.

[Pre\_3] S. Hajian and J. Domingo-Ferrer. “*A methodology for direct and indirect discrimination prevention in data mining*”. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

[Pre\_4] I. Zliobaite, F. Kamiran and T. Calders. “*Handling conditional discrimination*”. In ICDM, pp. 992-1001, 2011.

# A framework for direct and indirect discrimination prevention in data mining



Classification rule:  $r = A, B \rightarrow C$

$B$	$C$	$\neg C$	
$A$	$a_1$	$n_1 - a_1$	$n_1$
$\neg A$	$a_2$	$n_2 - a_2$	$n_2$

$a_1 = \text{supp}(A, B, C)$   
 $a_2 = \text{supp}(\neg A, B, C)$   
 $n_1 = \text{supp}(A, B)$   
 $n_2 = \text{supp}(\neg A, B)$

$$p_1 = a_1/n_1 \quad p_2 = a_2/n_2 \quad p = (a_1 + a_2)/(n_1 + n_2)$$

$$\text{elift}(r) = \frac{p_1}{p}, \quad \text{elift}_d(r) = p_1 - p, \quad \text{elift}_c(r) = \frac{1 - p_1}{1 - p}$$

$$\text{slift}(r) = \frac{p_1}{p_2}, \quad \text{slift}_d(r) = p_1 - p_2, \quad \text{slift}_c(r) = \frac{1 - p_1}{1 - p_2}$$

Based on direct discriminatory measures  $f \in \{\text{elift}, \text{slift}, \dots\}$ , a PD classification rule  $r: A, B \rightarrow C$  is:

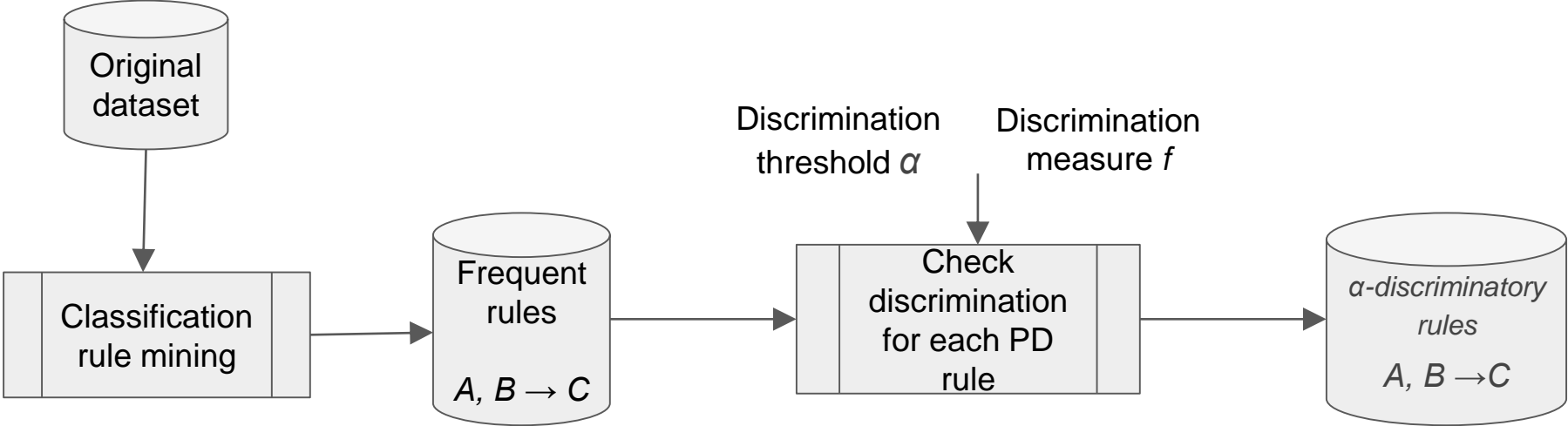
$\alpha$ -discriminatory if  $f(r) \geq \alpha$ ; or

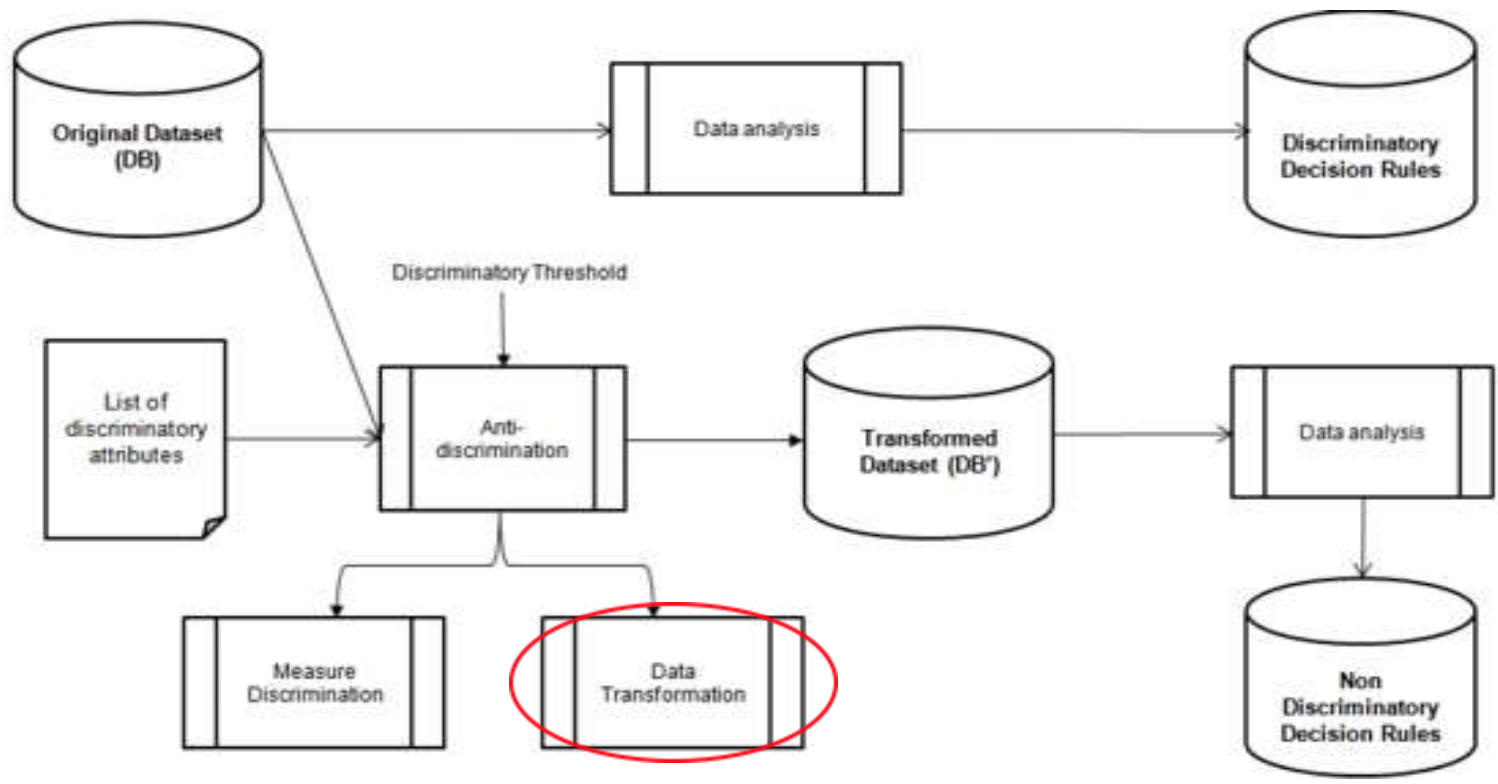
$\alpha$ -protective if  $f(r) < \alpha$

$\alpha$  states an acceptable level of discrimination according to laws and regulations

e.g. U.S. Equal Pay Act: This amounts to using *slift* with  $\alpha = 1.25$ .

# Measure discrimination





# Data transformation

The purpose is transform the original data  $D$  in such a way to remove direct and/or indirect discriminatory biases, with **minimum impact**

On the data, and

On legitimate decision rules

Different metrics and algorithms have been developed to specify

**Which records** (and in **which order**) should be changed?

**How many records** should be changed?

**How those records** should be changed during data transformation?

Metrics for measuring data utility and discrimination removal

# Which records should be change and how?

We need to enforce the following inequality for each  $\alpha$ -discriminatory rule  $r$

$$f(r: A, B \rightarrow C) < \alpha, \text{ where } f \in \{\text{elift, slift, ...}\}$$

Data transformation method to enforce the above inequality where  $f=\text{elift}$

DTM1: **Changes the discriminatory itemset** e.g., gender changed from male to female in the records with granted credits

DTM 2: **Changes the class item** e.g., from grant credit to deny credit in the records with male gender

Data transformation methods for direct rule protection

Direct Rule Protection	
DTM 1	$\neg A, B \rightarrow \neg C \Rightarrow A, B \rightarrow \neg C$
DTM 2	$\neg A, B \rightarrow \neg C \Rightarrow \neg A, B \rightarrow C$

## Which records should be change and how?

A suitable data transformation with minimum information loss to make each  $\alpha$ -discriminatory rule  $\alpha$ -protective.

we should enforce the following inequality for each  $\alpha$ -discriminatory rule  $r$

$$f(r: A, B \rightarrow C) < \alpha, \text{ Where } f \in \{\text{elift, slift, ...}\}$$

**Theorem:** DTM1 and DTM2 methods for making each  $\alpha$ -discriminatory rule  $r$   $\alpha$ -protective w.r.t.  $f$  do not generate new  $\alpha$ -discriminatory rules as a result of their transformations.



## How many records should be changed?

A suitable data transformation with minimum information loss to make each  $\alpha$ -discriminatory rule  $\alpha$ -protective.

we should enforce the following inequality for each  $\alpha$ -discriminatory rule  $r$

$$f(r: A, B \rightarrow C) < \alpha, \text{ where } f = \text{elift}$$

DTM1: Taking  $\Delta_{\text{elift}}$  equal to the ceiling of the right-hand side of Equation (below) suffices to

$$\Delta_{\text{elift}} > \frac{\alpha \times \text{supp}(A, B) \times \text{supp}(B, C) - \text{supp}(A, B, C) \times \text{supp}(B)}{\text{supp}(A, B, C) - \alpha \times \text{supp}(A, B)},$$

In which order records should be changed?

DTM1: perturb the discriminatory itemset from  $\sim A$  (male) to  $A$  (female) in the subset  $\mathcal{D}_c$  of all records of the original data set which completely support the rule  $\sim A, B \rightarrow \sim C$  and have **minimum impact on other rules**

$\mathcal{D}_c \leftarrow$  All records completely supporting  $\neg A, B \rightarrow \neg C$

**for** each  $db_c \in \mathcal{D}_c$  **do**

    Compute  $impact(db_c) = |\{r_a \in \mathcal{FR} | db_c \text{ supports the premise of } r_a\}|$

**end for**

Sort  $\mathcal{D}_c$  by ascending impact

# Fairness-aware data mining

## Pre-processing approaches:

[Pre\_1] M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. “*Certifying and removing disparate impact*”. In KDD, pp. 259-268, 2015.

[Pre\_2] F. Kamiran and T. Calders. “*Data preprocessing techniques for classification without discrimination*”. In Knowledge and Information Systems (KAIS), 33(1), 2012.

[Pre\_3] S. Hajian and J. Domingo-Ferrer. “*A methodology for direct and indirect discrimination prevention in data mining*”. In IEEE Transactions on Knowledge and Data Engineering (TKDE), 25(7), 2013.

[Pre\_4] I. Zliobaite, F. Kamiran and T. Calders. “*Handling conditional discrimination*”. In ICDM, pp. 992-1001, 2011.

# Handling conditional discrimination

(Zliobaite et al., 2011)

Previous pre-processing techniques aimed at removing all discrimination

However:

- Some parts may be *explainable*;
- Leads to *reverse discrimination*

# Example of fully explainable discrimination

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	20%	20%	40%	40%
accepted (+)	160	40	80	320

- 36% of males accepted, 24% of females accepted
- However, the difference is fully explainable by the fact that females applied to the more competitive program (medicine).
- Similar to the famous University of California, Berkeley 1973 case.

# Some explainable + some bad discrimination

	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	15%	25%	35%	45%
accepted (+)	120	50	70	360

$$= (20\% \times 25\% + 80\% \times 45\%)$$

$$- (80\% \times 15\% + 20\% \times 35\%)$$

$$= 41\% - 19\% = 22\%$$

Part of this discrimination can be explained, although not all of it.

# Analysis of explainable discrimination

How much discrimination can be explained?

What *should have been* the acceptance rate  $P^*(+|Fac)$  for faculty Fac?

(1)  $P^*(+|Fac) = P_{obs}(+ | Fac) \rightarrow$  leads to *redlining*

(2)  $P^*(+ | Fac) = [P_{obs}(+ | Fac, m) + P_{obs}(+ | Fac, f)] / 2$

$D_{expl}$  = discrimination when it would be true that:

$$P(+ | m, Fac) = P(+ | f, Fac) = P^*(+ | Fac)$$

$$D_{bad} = D_{all} - D_{expl}$$

# Analysis of explainable discrimination

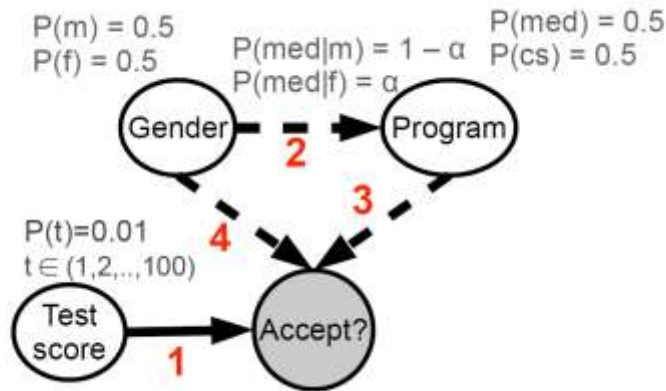
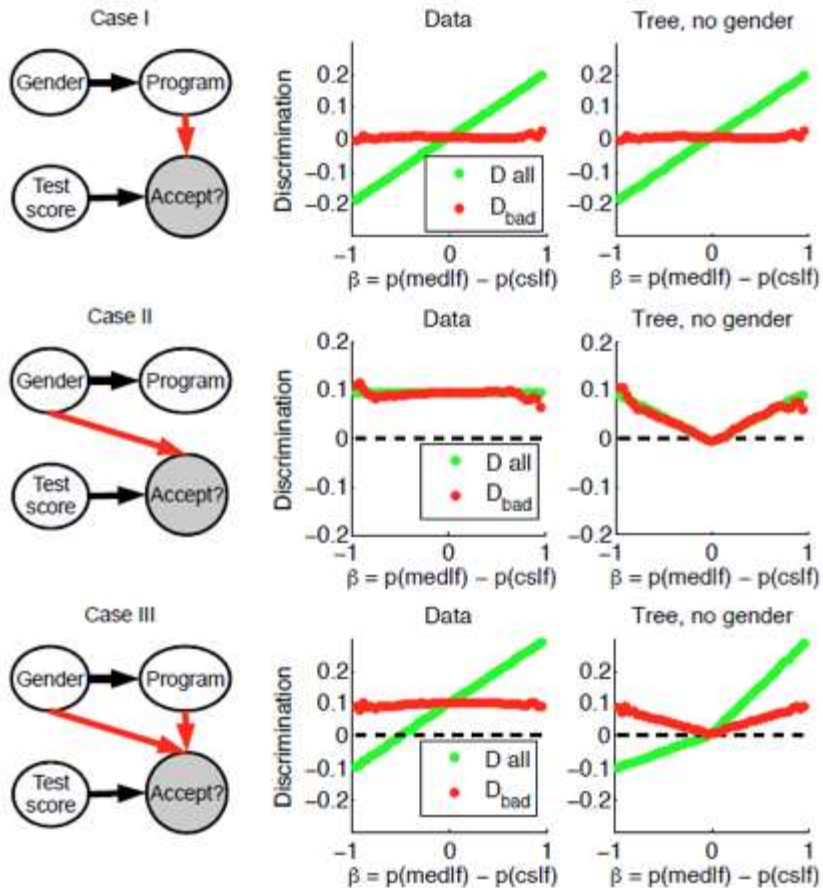
	medicine		computer	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate (Example 2)	15%	25%	35%	45%
corrected acceptance rate	20%		40%	
accepted explainable	160	40	80	320

$$D_{\text{expl}} = (20\% \times 20\% + 80\% \times 40\%) - (80\% \times 20\% + 20\% \times 40\%) = 12\%$$

$$D_{\text{bad}} = D_{\text{all}} - 12\% = 22\% - 12\% = 10\%$$



# Simulation Experiments



$$y = \delta \left[ \left( t + a(-1)^{\delta[\text{med}]} + b(-1)^{\delta[f]} \right) > 70 \right]$$

- $t$  = test score (integer in  $[1, 100]$  uniform at random);
- $a$  = effect on acceptance decision due to program;
- $b$  = effect on acceptance decision due to gender bias

	$P(t)$	$a$	$b$	$P(\text{med} f)$
Case I, only explainable	0.01	10	0	$\alpha$
Case II, only bad	0.01	0	5	$\alpha$
Case III, explainable and bad	0.01	10	5	$\alpha$

# Solution: Locally change input data

1. Divide the dataset according to the explanatory attribute(s)

2. Estimate  $P^*(+|e_i)$  for all partitions  $e_i$

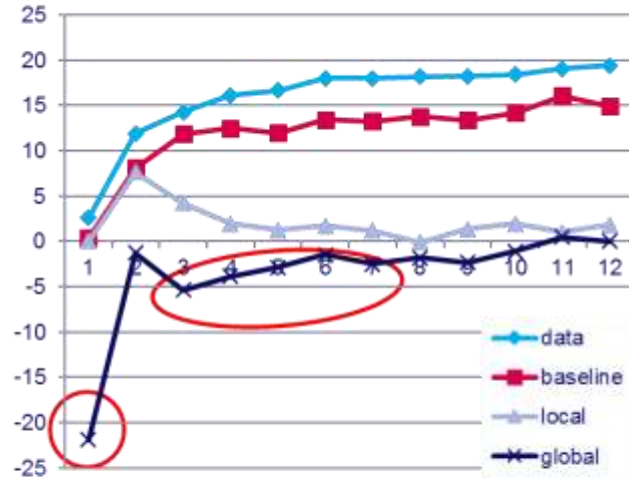
3. Apply local techniques on partition  $e_i$  so that

$P(+|e_i, f) = P(+|e_i, m) = P^*(+|e_i)$  becomes true

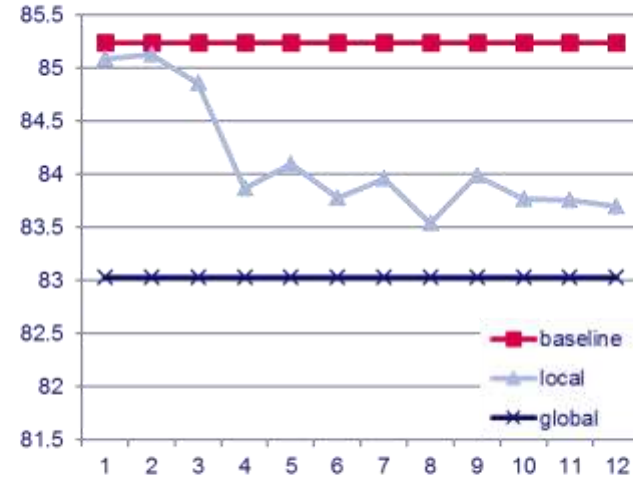
- Local massaging
- Local preferential sampling

# Experiments: Discrimination after Massaging

## Bad Discrimination



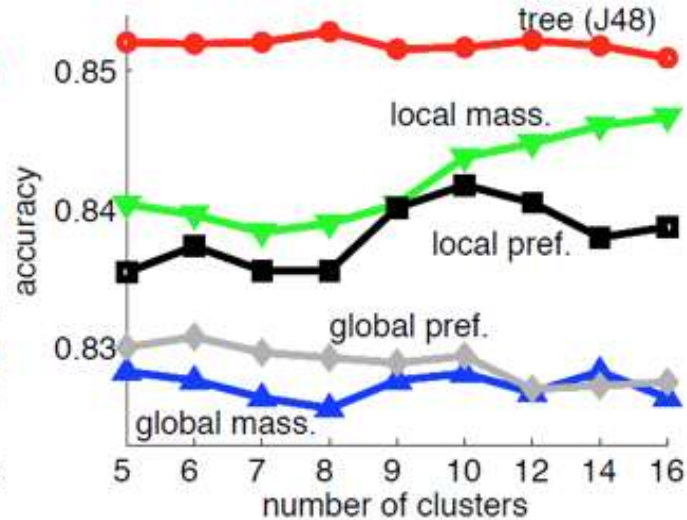
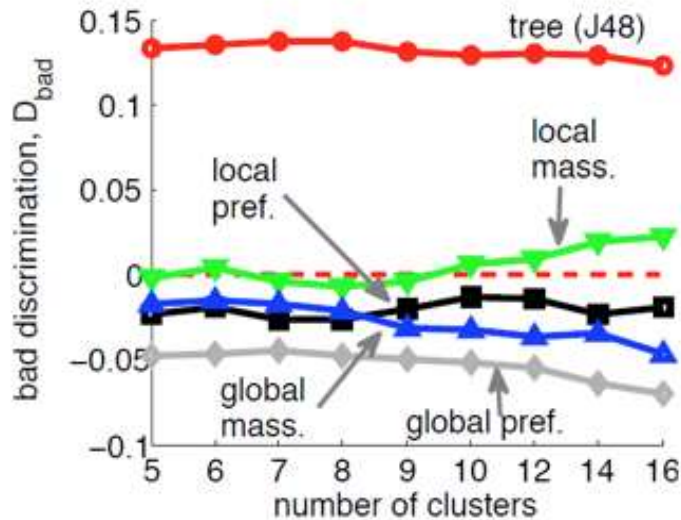
## Accuracy



- **Global techniques tend to overshoot when large part of the discrimination can be explained**

# Experiments with multiple explanatory attributes

If there are multiple explanatory attributes: create groups of individuals by clustering based upon explanatory attributes (e.g., working hours and experience when determining salary).



# Fairness-aware data mining

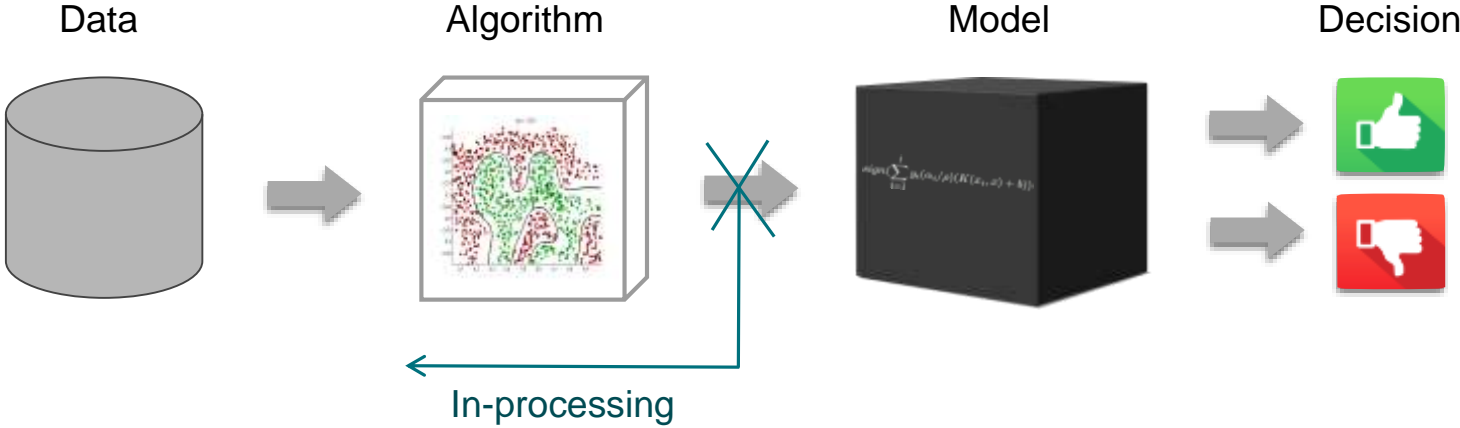
Pre-processing approaches: (not covered here)

**[Pre\_5]** S. Ruggieri. *“Using  $t$ -closeness anonymity to control for non-discrimination”*. Transactions on Data Privacy, 7(2), pp.99-129, 2014.

**[Pre\_6]** S. Hajian, J. Domingo-Ferrer, and O. Farras. *“Generalization-based privacy preservation and discrimination prevention in data publishing and mining”*. Data Mining and Knowledge Discovery, 28(5-6), pp.1158-1188, 2014.

These two papers (together with others) deal with simultaneously with privacy and anti-discrimination. This new promising family of approaches will be discussed in Part 4 of the tutorial.

# Non-discriminatory data-driven decision-making



# Fairness-aware data mining

## In-processing approaches:

[In\_1] F. Kamiran, T. Calders and M. Pechenizkiy. *“Discrimination aware decision tree learning”*. In ICDM, pp. 869-874, 2010.

[In\_2] T. Calders and S. Verwer. *“Three Naïve Bayes Approaches for Discrimination-Free Classification”*. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In\_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. *“Fairness Constraints: A Mechanism for Fair Classification”*. arXiv preprint arXiv:1507.05259, 2015.

[In\_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *“Fairness through awareness”*. In ITCS 2012, pp. 214-226, 2012.

[In\_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. *“Learning fair representations”*. In ICML, pp. 325-333, 2013.

# Fairness-aware data mining

In-processing approaches: (not covered here)

**[In\_6]** T. Kamishima, S. Akaho, H. Asoh and J. Sakuma. *“Fairness-aware classifier with prejudice remover regularizer”*. In PKDD, pp. 35-50, 2012.

**[In\_7]** K. Fukuchi, J. Sakuma, T. Kamishima. *“Prediction with Model-Based Neutrality”*. ECML/PKDD (2) 2013: 499-514.

**[In\_8]** B. Fish, J. Kun and A.D. Lelkes. *“A Confidence-Based Approach for Balancing Fairness and Accuracy”* arXiv preprint:1601.05764, 2015.

many more... (probably)



# Fairness-aware data mining

## In-processing approaches:

**[In\_1]** F. Kamiran, T. Calders and M. Pechenizkiy. *“Discrimination aware decision tree learning”*. In ICDM, pp. 869-874, 2010.

**[In\_2]** T. Calders and S. Verwer. *“Three Naïve Bayes Approaches for Discrimination-Free Classification”*. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

**[In\_3]** M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. *“Fairness Constraints: A Mechanism for Fair Classification”*. arXiv preprint arXiv:1507.05259, 2015.

**[In\_4]** C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *“Fairness through awareness”*. In ITCS 2012, pp. 214-226, 2012.

**[In\_5]** R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. *“Learning fair representations”*. In ICML, pp. 325-333, 2013.

# Problem definition

- Given: dataset  $D$ , an attribute  $B$ , a value  $b \in \text{dom}(B)$
- Find a classifier  $M$  that:
  - Minimizes discrimination w.r.t.  $B=b$
  - Maximizes predictive accuracy

First attempt: decision tree

- Change split criterion
- Leaf Relabeling

# Change split criterion

Purity with respect to *Class* attribute

Impurity with respect to sensitive attribute *B*

Guarantee over resultant discrimination level on training data; e.g., not more than 3%

E.g.: Information gain maximal w.r.t. class and minimal w.r.t. B

Objective:  $GINI_{split}(Class) / GINI_{split}(B)$

Objective:  $GINI_{split}(Class) - GINI_{split}(B)$

**Input:** Dataset D

**Output:** Decision tree t

**Induce(D):**

If all tuples t in D have label + then return 

If all tuples t in D have label - then return 

For all split criteria C:

$$D_{1,C} = \{ t \text{ in } D \mid t \text{ satisfies } C \}$$

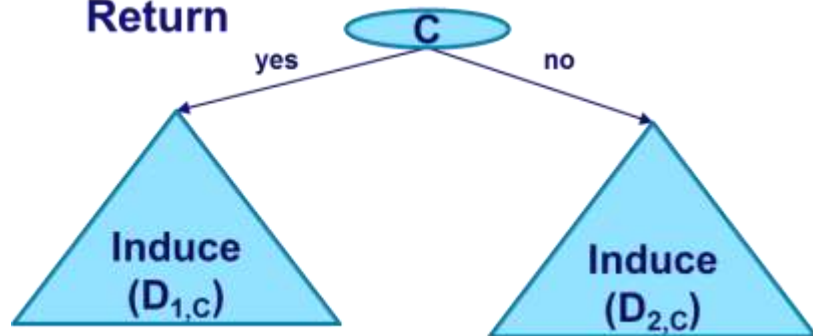
$$D_{2,C} = D - D_1$$

**Measure Quality( $D_1, D_2$ )**

$$\text{GINI}_{\text{split}}(\text{Class}) / \text{GINI}_{\text{split}}(G)$$

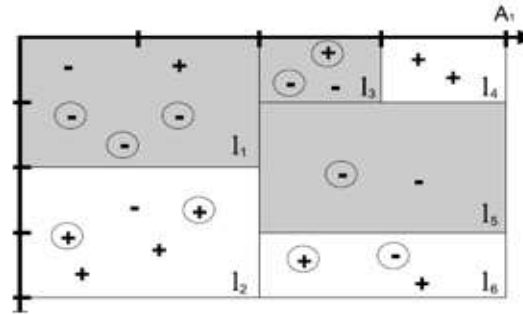
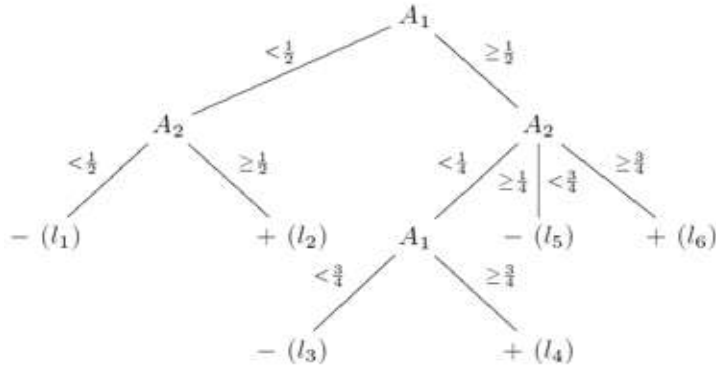
Let C be the best split

Return



# Leaf relabeling

Decision trees divide up the decision space



Labels are assigned according to the majority class

$$\text{Disc}_T = p(M = + | B \neq b) - p(M = + | B = b) = 6/10 - 4/10 = 0.2 \text{ or } 20 \%$$

Relabel some leaves to reduce the discrimination

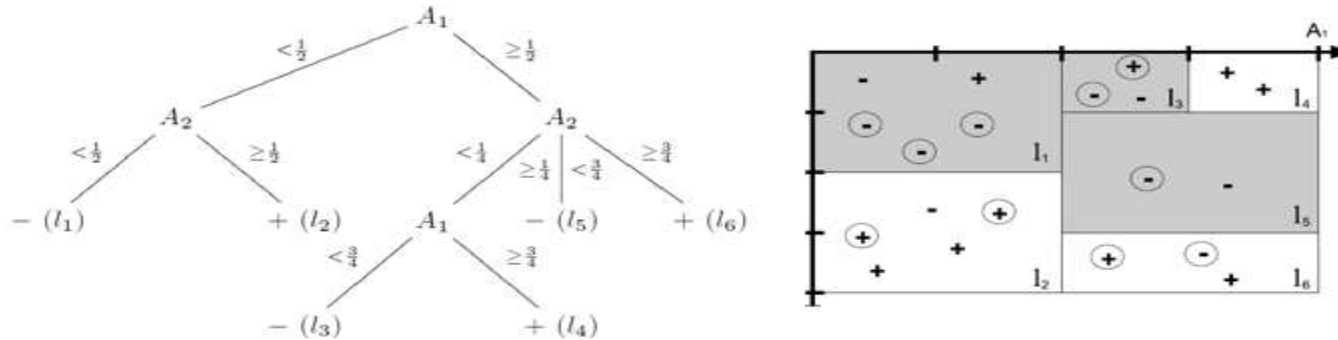
# Leaf relabeling

E.g.: Relabel node  $l_1$  from  $-$  to  $+$

Influence on accuracy:  $-15\%$

Influence on discrimination:  $20\% - 30\% = -10\%$

Change in accuracy and discrimination independent of changes in other leaves



Task: find the optimal relabeling of the nodes

# Leaf relabeling

Optimal Leaf Relabeling is equivalent to the Knapsack problem

Given:

- A knapsack of size  $K$

- A set of objects  $O$

- A weight and a size for every object

Find:

- A subset of objects that fits in the knapsack and maximizes the weight

This problem is known to be **NP-complete**

Yet it has good approximations; e.g., the **greedy algorithm**

# Leaf Relabeling = Knapsack

Do not consider relabelings:

that reduce accuracy

without lowering discrimination

Current discrimination = 20%

Relabeling all: -50%

Hence, 30% can stay

Knapsack problem:

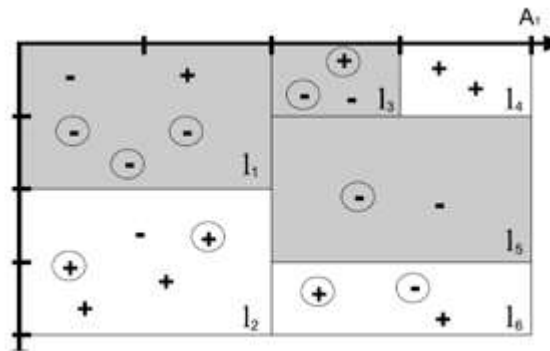
Select nodes **NOT** relabeled

Dacc: weight

Ddisc: size

K = 30% (that can stay)

Outcome: relabel  $l_4$



<i>Node</i>	$\Delta acc$	$\Delta disc$
$l_1$	-15%	-10%
$l_2$	-15%	-10%
$l_3$	-5%	-10%
$l_4$	-10%	-20%
$l_5$	-10%	0%
$l_6$	-5%	10%



# Fairness-aware data mining

## In-processing approaches:

[In\_1] F. Kamiran, T. Calders and M. Pechenizkiy. “*Discrimination aware decision tree learning*”. In ICDM, pp. 869-874, 2010.

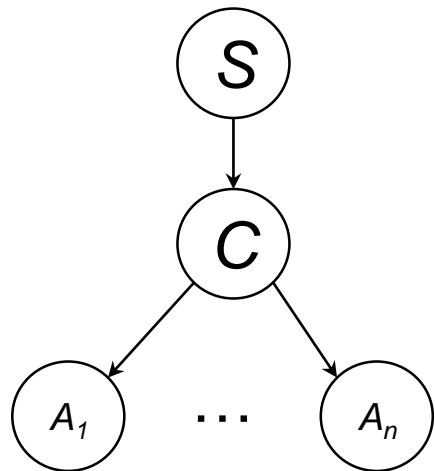
[In\_2] T. Calders and S. Verwer. “*Three Naïve Bayes Approaches for Discrimination-Free Classification*”. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In\_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. “*Fairness Constraints: A Mechanism for Fair Classification*”. arXiv preprint arXiv:1507.05259, 2015.

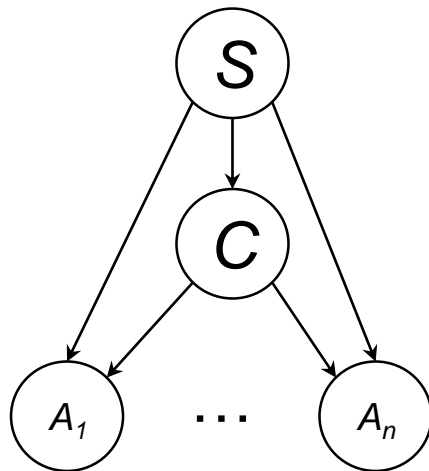
[In\_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. “*Fairness through awareness*”. In ITCS 2012, pp. 214-226, 2012.

[In\_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. “*Learning fair representations*”. In ICML, pp. 325-333, 2013.

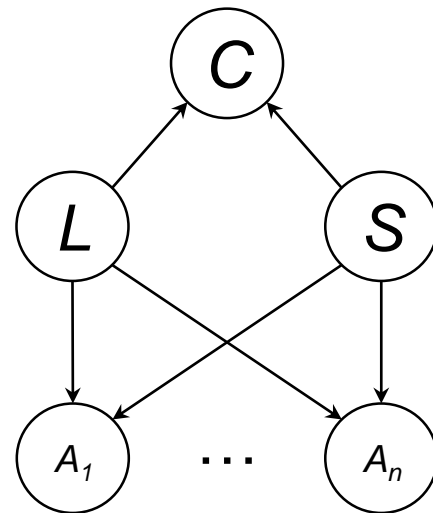
# Three Naive Bayes Approaches for Discrimination-Free Classification



Approach 1:  
Modified Naive Bayes

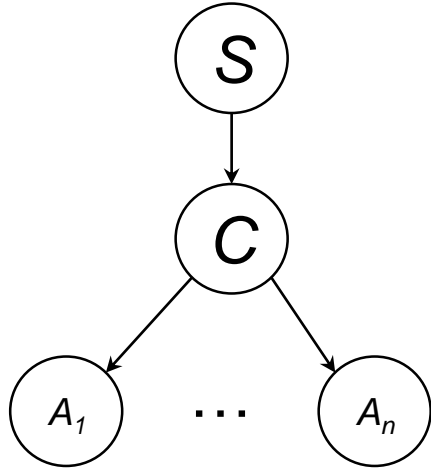


Approach 2:  
Two Naive Bayes models



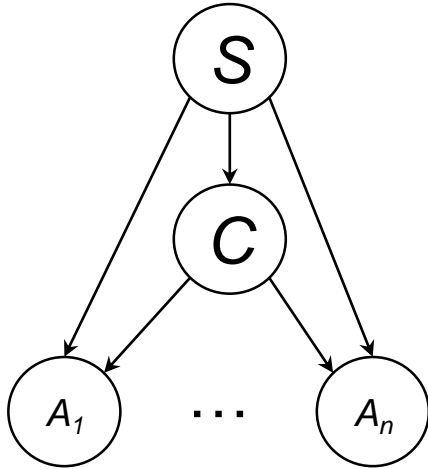
Approach 3:  
Latent variable model

## Approach 1: Modified Naive Bayes



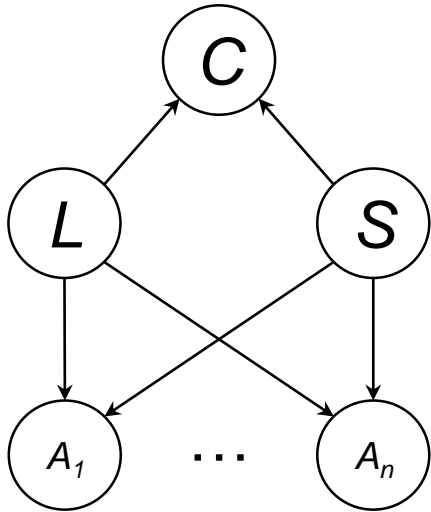
- Use  $P(C, S, A_1, \dots, A_n) = P(S)P(C|S)P(A_1|C) \dots P(A_n|C)$  instead of  $P(C, S, A_1, \dots, A_n) = P(C)P(S|C)P(A_1|C) \dots P(A_n|C)$
- Alterate distribution  $P(C|S)$  until there is no more discrimination.
- It creates a discrimination free Naive Bayes classifier but does not avoid red-lining effect due to attributes  $A_s$  correlated with  $S$ .

## Approach 2: Two Naive Bayes models



- How to remove correlation between attributes  $A_s$  and  $S$ ?
- Simply remove attributes  $A_s \rightarrow$  big loss in accuracy!
- Remove the fact that attributes  $A_s$  can be used to decide  $S$ , by splitting the learning in two, w.r.t. the value of  $S$ . For instance if  $S$  is gender, build one model for male and one model for female.

## Approach 3: Latent variable model



- Try to discover the actual class labels that the dataset should have had if it was discrimination-free.
- This is modeled by a **latent variable L**.
- Assumptions:
  1. L is independent from S  $\rightarrow$  L is discrimination-free;
  2. C is determined by discriminating L using S uniformly at random.
- Fit L by means of Expectation-Maximization (EM)

# Fairness-aware data mining

## In-processing approaches:

[In\_1] F. Kamiran, T. Calders and M. Pechenizkiy. “*Discrimination aware decision tree learning*”. In ICDM, pp. 869-874, 2010.

[In\_2] T. Calders and S. Verwer. “*Three Naïve Bayes Approaches for Discrimination-Free Classification*”. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In\_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. “*Fairness Constraints: A Mechanism for Fair Classification*”. arXiv preprint arXiv:1507.05259, 2015.

[In\_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. “*Fairness through awareness*”. In ITCS 2012, pp. 214-226, 2012.

[In\_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. “*Learning fair representations*”. In ICML, pp. 325-333, 2013.

# Defining fairness

Applying doctrine of disparate impact: 80% rule

If 50% of male applicants get selected for the job, at least 40% of females should also get selected

A fair system might not always be 80:100

In certain scenarios, the prescribed proportion could be 50:10

The goal is to enable a range of "fair" proportions

# Fair classifier

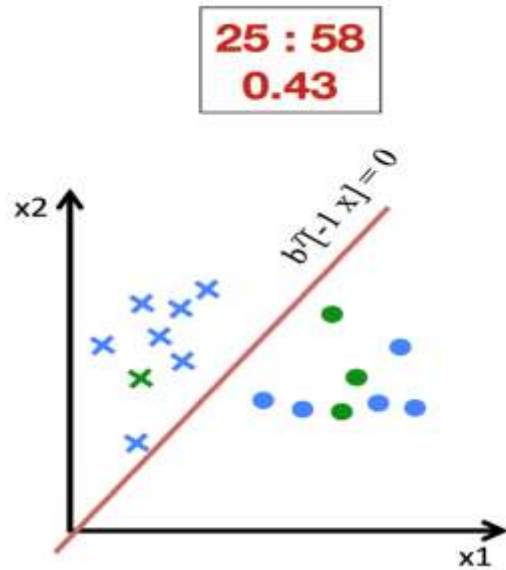
A classifier whose output achieves a given proportion of items (in positive class) with different values of sensitive feature



## Fairness constraint

**Key Idea:** Limit the cross-covariance between **sensitive feature value** and **distance from decision boundary**

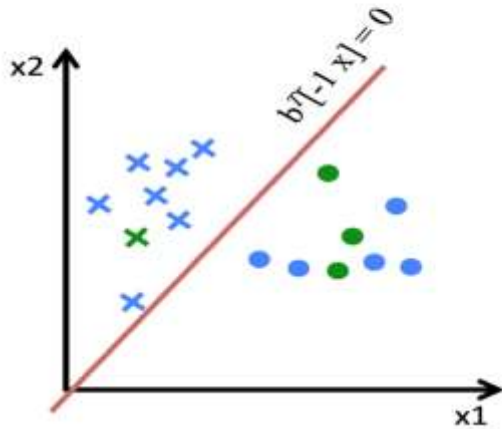
# Fairness constraint



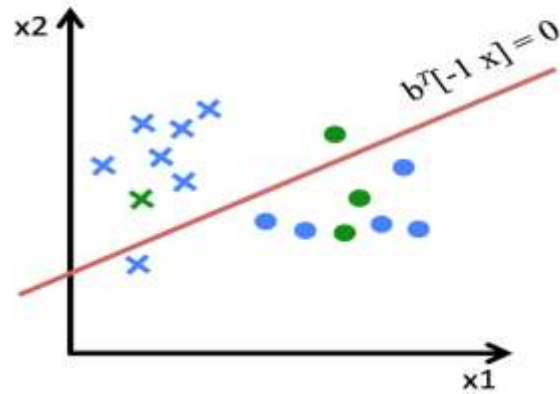
# Fairness constraint

$$\left| \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \mathbf{b}^T [-1 \ \mathbf{x}_i] \right| \leq \mathbf{c}$$

**25 : 58**  
**0.43**



**50 : 50**  
**1.0**



## Modifying the logistic regression classifier

$$p(y_i = 1 | \mathbf{x}_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}}$$

$$\text{maximize } \sum_{i=1}^N \log p(y_i | \mathbf{x}_i)$$

# Modifying the logistic regression classifier

$$p(y_i = 1 | \mathbf{x}_i) = \frac{1}{1 + e^{-b_0 + \sum_j b_j x_{ij}}}$$

$$\begin{aligned} &\text{maximize} && \sum_{i=1}^N \log p(y_i | \mathbf{x}_i) \\ &\text{subject to} && \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \mathbf{b}^T [-1 \ \mathbf{x}_i] \leq \mathbf{c}, \\ &&& \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \mathbf{b}^T [-1 \ \mathbf{x}_i] \geq -\mathbf{c} \end{aligned}$$

**Key point:** possible to solve this problem efficiently

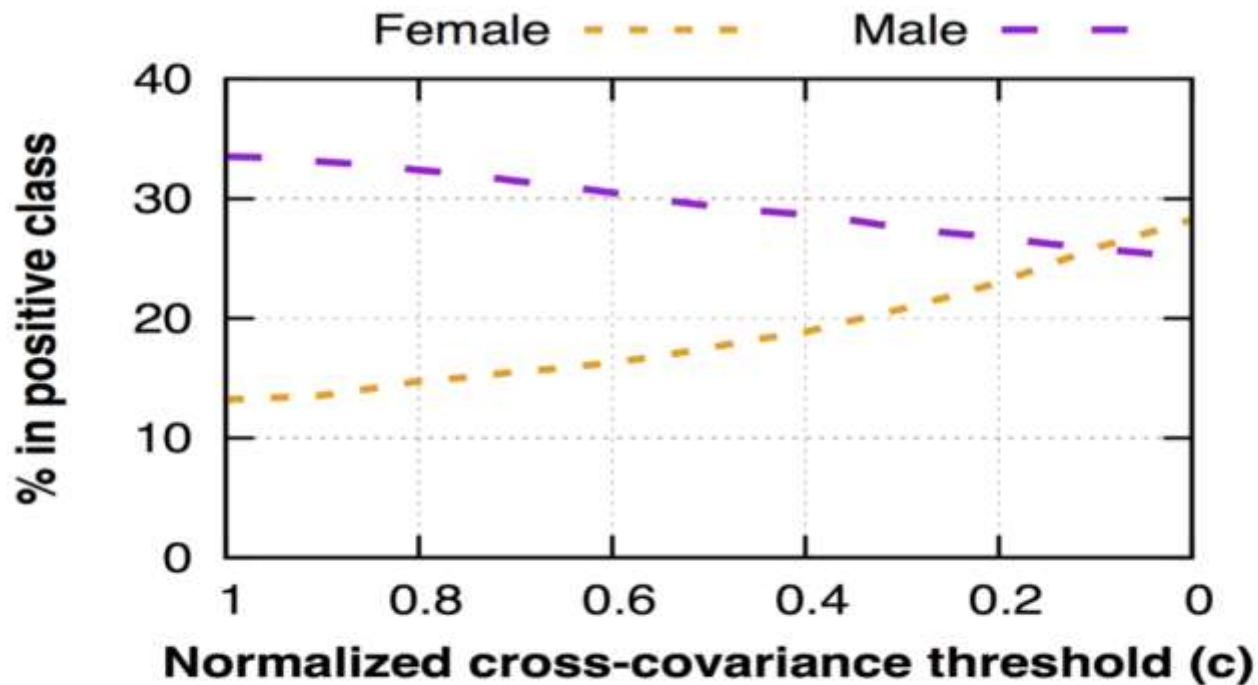
## Modifying the Hinge loss classifier

$$\begin{aligned} &\text{minimize} && \sum_{i=1}^N \max(0, y_i(\mathbf{b}^T[-1 \ \mathbf{x}_i])) \\ &\text{subject to} && \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \mathbf{b}^T[-1 \ \mathbf{x}_i] \leq \mathbf{c}, \\ &&& \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \mathbf{b}^T[-1 \ \mathbf{x}_i] \geq -\mathbf{c}, \end{aligned}$$

## Modifying the SVM classifier

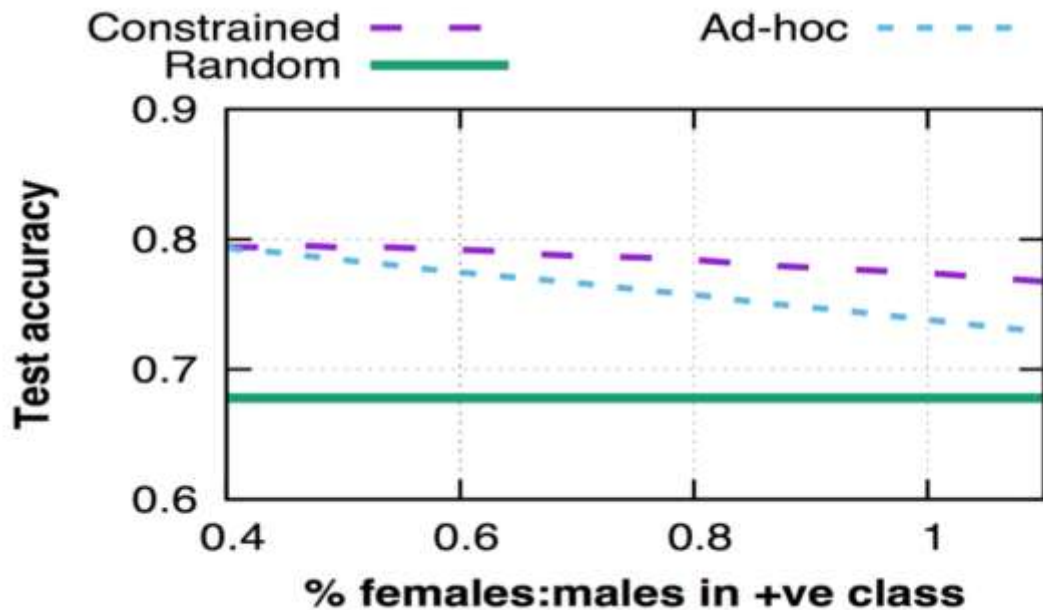
$$\begin{aligned} \text{minimize} \quad & \|\mathbf{b}\|^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & y_i (\mathbf{b}^T [-1 \ \mathbf{x}_i]) \geq 1 - \xi_i, \forall i \in \{1, \dots, n\} \\ & \xi_i \geq 0, \forall i \in \{1, \dots, n\}, \\ & \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \mathbf{b}^T [-1 \ \mathbf{x}_i] \leq \mathbf{c}, \\ & \frac{1}{N} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{z}}) \mathbf{b}^T [-1 \ \mathbf{x}_i] \geq -\mathbf{c}. \end{aligned}$$

# Tightening the constraints increases fairness





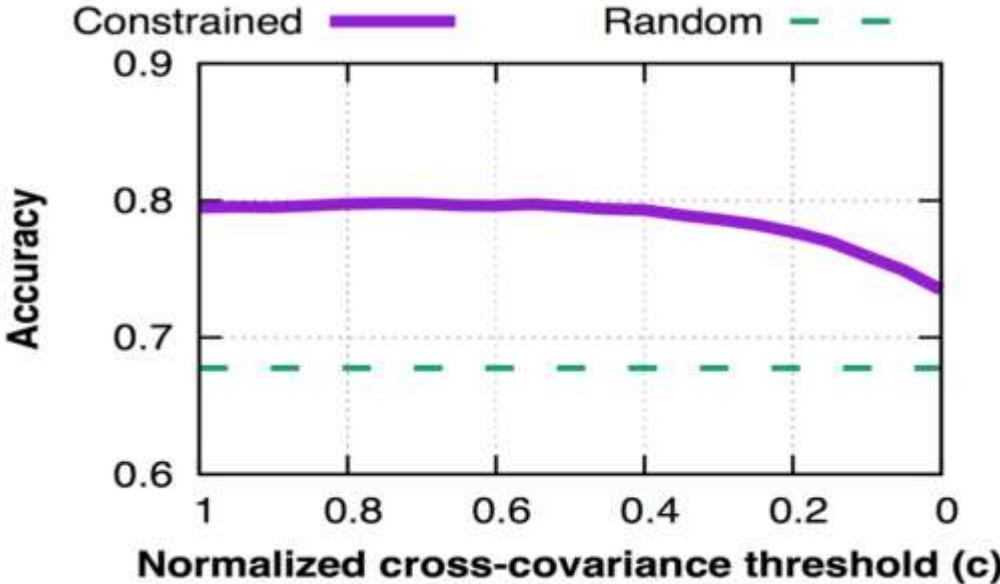
# Fairness vs accuracy trade-off



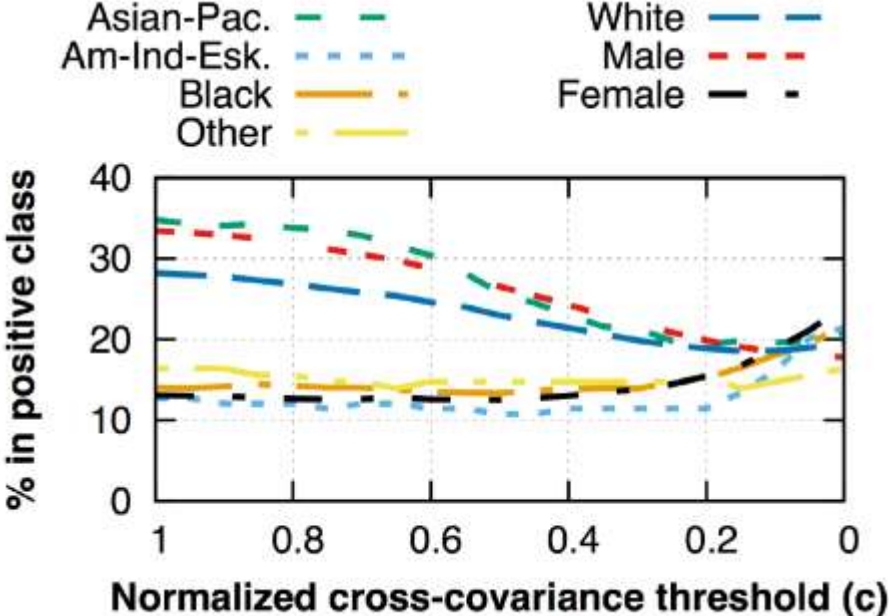
**Random:** takes the output of the unconstrained classifier and shuffles labels randomly until satisfying the given  $c$ .

**Ad-hoc:** takes the output of the unconstrained classifier and change females to +ve until satisfying the given  $c$ .

# Fairness vs accuracy trade-off



# Fairness for Multiple Features



# Fairness-aware data mining

## In-processing approaches:

[In\_1] F. Kamiran, T. Calders and M. Pechenizkiy. “*Discrimination aware decision tree learning*”. In ICDM, pp. 869-874, 2010.

[In\_2] T. Calders and S. Verwer. “*Three Naïve Bayes Approaches for Discrimination-Free Classification*”. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In\_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. “*Fairness Constraints: A Mechanism for Fair Classification*”. arXiv preprint arXiv:1507.05259, 2015.

[In\_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. “*Fairness through awareness*”. In ITCS 2012, pp. 214-226, 2012.

[In\_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. “*Learning fair representations*”. In ICML, pp. 325-333, 2013.

# Credit Application (WSJ 8/4/10)



More miles  
and **no annual fee**

Earn trips faster with VentureOne<sup>SM</sup>

Get Started 

only at  **Card Lab**

Capital One Card Lab  
Platinum Prestige Credit Card

Capital One Card Lab  
VentureOne Card

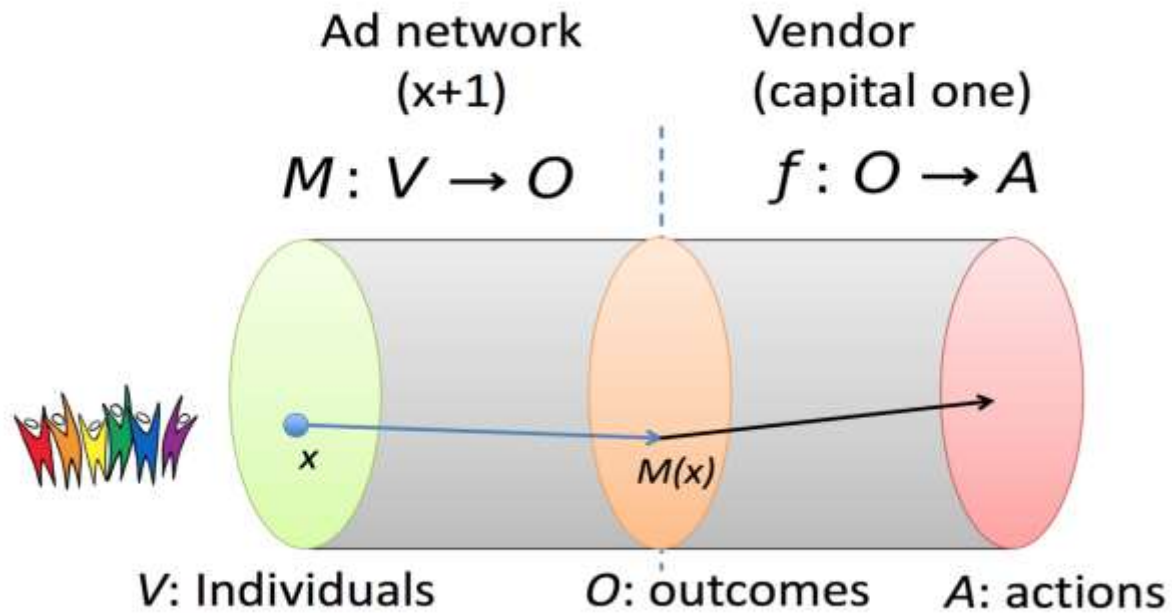
Savings Accounts  
Earn With Great Rates

The advertisement features a yellow Capital One VentureOne Visa Signature credit card floating in the air above a tropical island with palm trees and a blue ocean. The card displays the word 'VENTURE' in large letters, the Visa Signature logo, and a card number. The background is a light blue sky with soft clouds.

User visits [capitalone.com](http://capitalone.com)

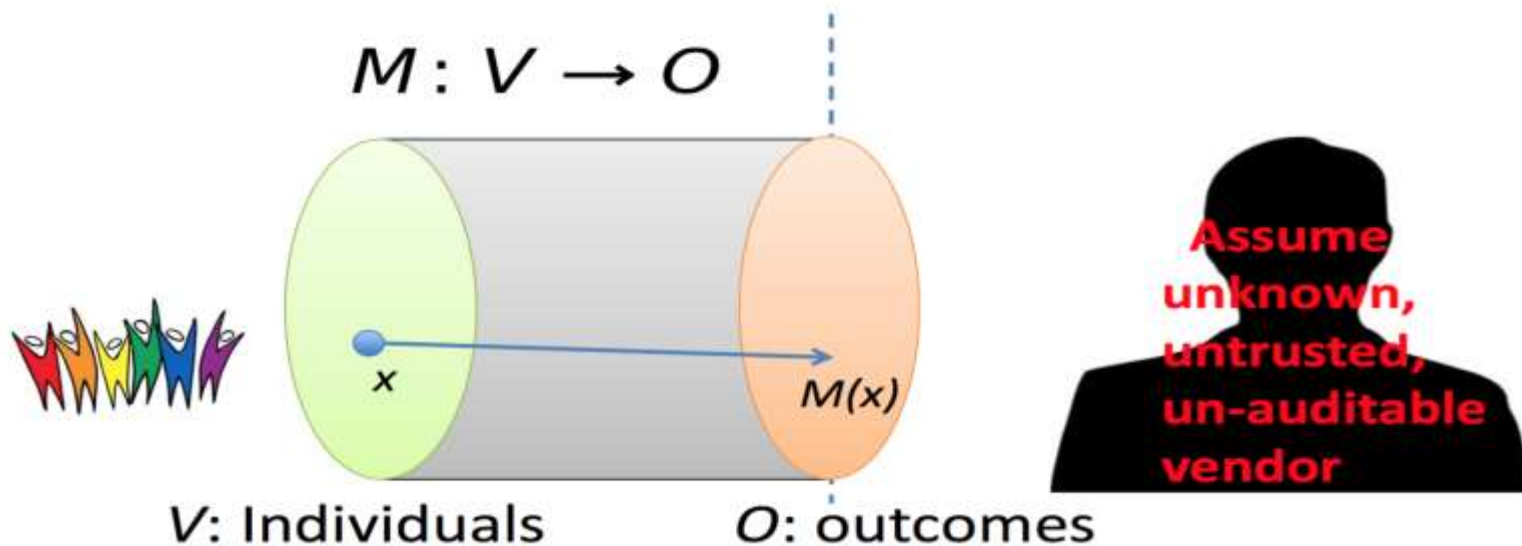
Capital One uses tracking information provided by the tracking network [x+1] to personalize offers

**Concern:** Steering minorities into higher rates (illegal)



# The goal

Achieve fairness in the classification step



# Fairness through Blindness

- Ignore all irrelevant/protected attributes
- Point of **failure**: Redundant encodings
  - Machine learning: You don't need to see the label to be able to predict it



# Group Fairness

- Equalize two groups  $S$ ,  $T$  at the level of outcomes
  - E.g.  $S = \text{minority}$ ,  $T = S^c$
  - $\Pr[\text{outcome } o \mid S] = \Pr[\text{outcome } o \mid T]$
- **Insufficient** as a notion of fairness
  - Has some good properties, but can be abused
  - Example: Advertise burger joint to carnivores in  $T$  and vegans in  $S$ .

# Lesson: Fairness is *task-specific*

- Fairness requires understanding of classification task
  - Cultural understanding of protected groups
  - Awareness



# Individual fairness

Treat *similar* individuals *similarly*



Similar for the purpose of  
the classification task

Similar distribution  
over outcomes

- Assume *task-specific similarity metric*
  - Extent to which two individuals are similar w.r.t. the classification task at hand
- Ideally captures *ground truth*
  - Or, society's best approximation
- Open to public discussion, refinement
  - In the spirit of Rawls
- Typically, does not suggest classification!

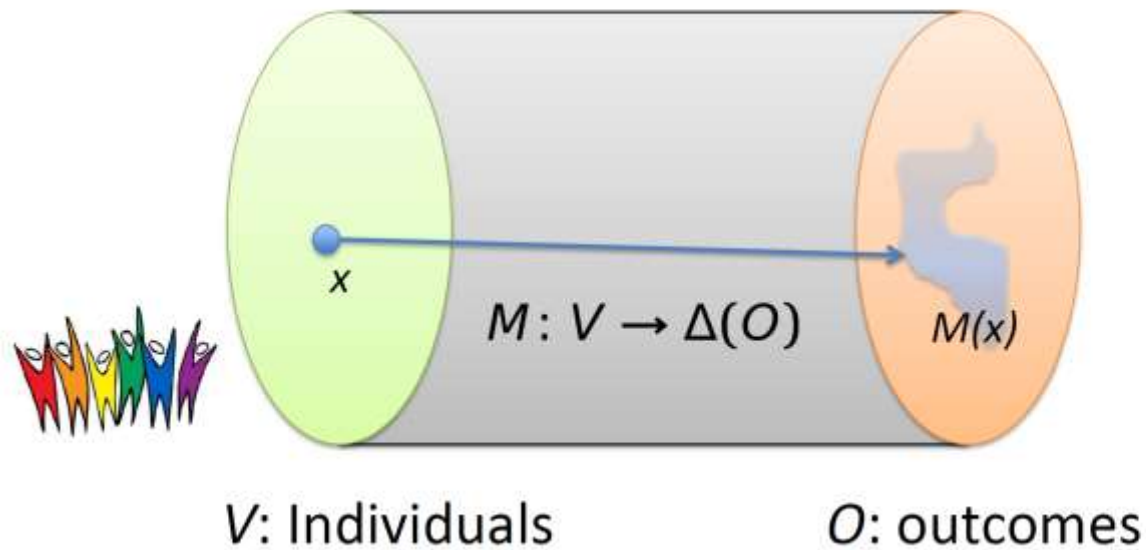
# Examples

- Financial/insurance risk metrics
  - Already widely used (though secret)
- **AALIM health care metric**
  - health metric for treating similar patients similarly
- Roemer's relative effort metric
  - Well-known approach in Economics/Political theory

Maybe not so much science fiction after all...

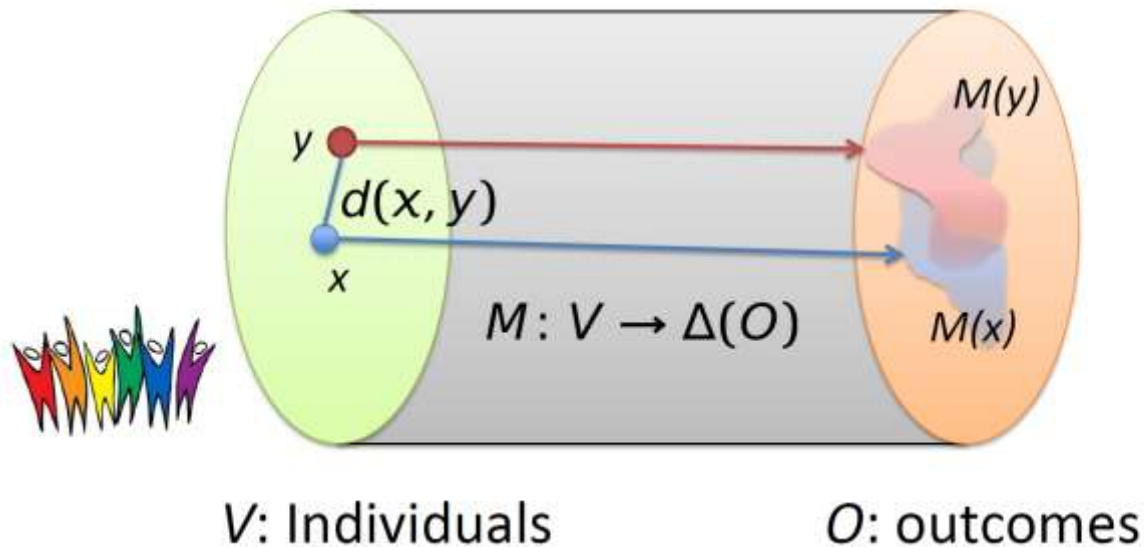
# Formal setup

Classification



Metric  $d: V \times V \rightarrow \mathbb{R}$

Lipschitz condition  $\|M(x) - M(y)\| \leq d(x, y)$



# Utility Maximization

Vendor can specify **arbitrary utility function**

$$U: V \times O \rightarrow \mathbb{R}$$

Can efficiently maximize vendor's expected utility subject to Lipschitz condition

$$\max_{x \in V} \mathbb{E}_{o \sim M(x)} U(x, o)$$

s.t.  $M$  is  $d$ -Lipschitz



# More contributions

- Several examples showing the **inadequacy of group fairness** (or *statistical parity*)
- **Connection between individual and group fairness**: the Lipschitz condition implies statistical parity between two groups if and only if the **Earthmover distance** between two groups is small.
- **Fair affirmative action**. Provide techniques for forcing statistical parity when it is not implied by the Lipschitz condition, while preserving as much fairness for the individuals as possible.
- **Relationship with privacy**: the proposed definition of fairness is a generalization of the notion of **differential privacy**.

# Fairness-aware data mining

## In-processing approaches:

[In\_1] F. Kamiran, T. Calders and M. Pechenizkiy. “*Discrimination aware decision tree learning*”. In ICDM, pp. 869-874, 2010.

[In\_2] T. Calders and S. Verwer. “*Three Naïve Bayes Approaches for Discrimination-Free Classification*”. Data Mining and Knowledge Discovery, 21(2):277-292, 2010.

[In\_3] M.B. Zafar, I. Valera, M.G. Rodriguez, and K.P. Gummadi. “*Fairness Constraints: A Mechanism for Fair Classification*”. arXiv preprint arXiv:1507.05259, 2015.

[In\_4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. “*Fairness through awareness*”. In ITCS 2012, pp. 214-226, 2012.

[In\_5] R. Zemel, Y. Wu, K. Swersky, T. Pitassi and C. Dwork. “*Learning fair representations*”. In ICML, pp. 325-333, 2013.

# Main limitations of “Fairness through awareness”

1. The problem of fairness in classification is reduced to the problem of establishing a fair distance metric. The **distance metric** that defines the similarity between the individuals is **assumed to be given**. This might be unrealistic in certain settings.
2. Their framework is **not formulated as a learning framework**: it gives a mapping for a given set of individuals, but it doesn't provide any mean to generalize to novel unseen data (new individuals).

# “Learning fair representations” (Zemel et al. ICML 2013)

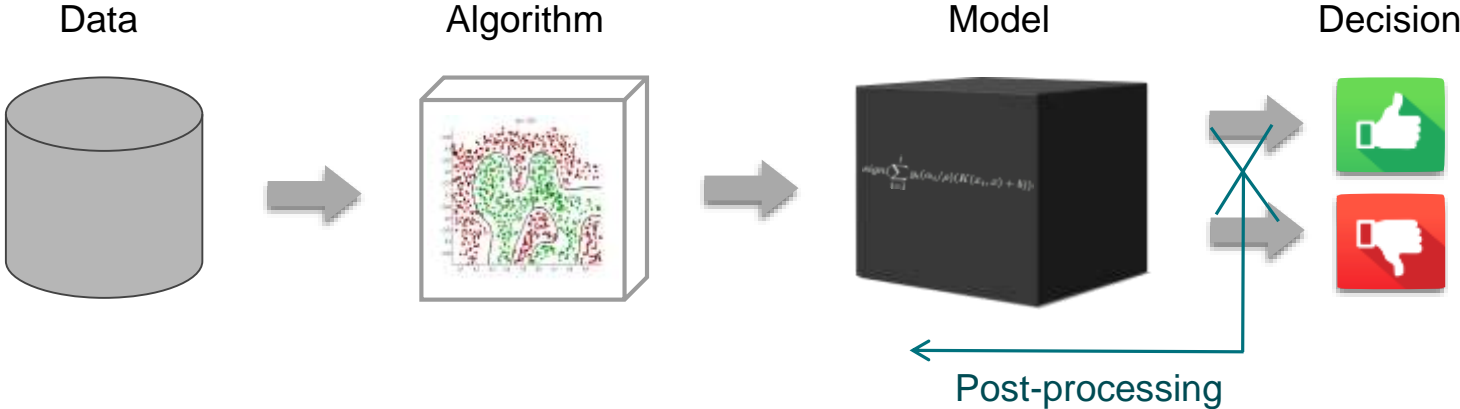
... extends “Fairness through awareness” in several important ways.

1. It develops a learning framework: learn a general mapping, applies to any individual.
2. Learns a restricted form of a distance function as well as the intermediate representation. No longer needed a distance function given a-priori.
3. Achieves both group fairness and individual fairness.
4. The intermediate representation can be used for other classification tasks (i.e., transfer learning is possible).
5. Experimental assessment.

# Main idea [sketch]

- Map each individual (a data point in the input space) to a **probability distribution in a new representation space**.
- The aim of the new representation is to **lose** any information that can reconstruct whether the **individual belongs to the protected subgroups**, while **maintaining** as much **other information** as possible.
- Fairness becomes an optimization problem of finding the intermediate representation that **best encodes the data** while **obfuscating membership** to the protected subgroups.
- **Tool: probabilistic mapping to a set of prototypes** (it can be seen as a form of **discriminative clustering model**). **[Details omitted]**

# Non-discriminatory data-driven decision-making



# Fairness-aware data mining

## Post-processing approaches:

**[Post\_1]** F. Kamiran, T. Calders and M. Pechenizkiy. *“Discrimination aware decision tree learning”*. In ICDM, pp. 869-874, 2010. **(already covered)**

**[Post\_2]** S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. *“Discrimination-and privacy-aware patterns”*. In Data Mining and Knowledge Discovery, 29(6), 2015.

**[Post\_3]** F. Kamiran, A. Karim, and X. Zhang. *“Decision Theory for discrimination-aware classification”*. In ICDM, pp. 924-929, 2012. **(not covered)**

# Fairness-aware data mining

## Post-processing approaches:

**[Post\_1]** F. Kamiran, T. Calders and M. Pechenizkiy. “*Discrimination aware decision tree learning*”. In ICDM, pp. 869-874, 2010. **(already covered)**

**[Post\_2]** S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. “*Discrimination-and privacy-aware patterns*”. In Data Mining and Knowledge Discovery, 29(6), 2015.

**[Post\_3]** F. Kamiran, A. Karim, and X. Zhang. “*Decision Theory for discrimination-aware classification*”. In ICDM, pp. 924-929, 2012. **(not covered)**



# Privacy and anti-discrimination should be addressed together

Suppose to publish frequent pattern (support  $> k$ , for  $k$ -anonymity) extracted from personal data for credit approval decision making.

Sex	Job	Credit_history	Salary	Credit_approved
Male	Writer	No-taken	... €	Yes
Female	Lawyer	Paid-duly	... €	No
Male	Veterinary	Paid-delay	...€	Yes
...	...	...	...	...

## Privacy protection only

sex=female  $\rightarrow$  credit-approved=no (support 126)

## Discrimination protection only

job =veterinarian, salary =low  $\rightarrow$  credit-approved=no (support 40)

job = veterinarian  $\rightarrow$  credit-approved=no (support 41)

# Support > k doesn't imply k-anonymity

Atzori, Bonchi, Giannotti, Pedreschi. *"Anonymity Preserving Pattern Discovery"* VLDB Journal 2008

job = veterinarian, salary = low  $\rightarrow$  credit-approved = no (support 40)

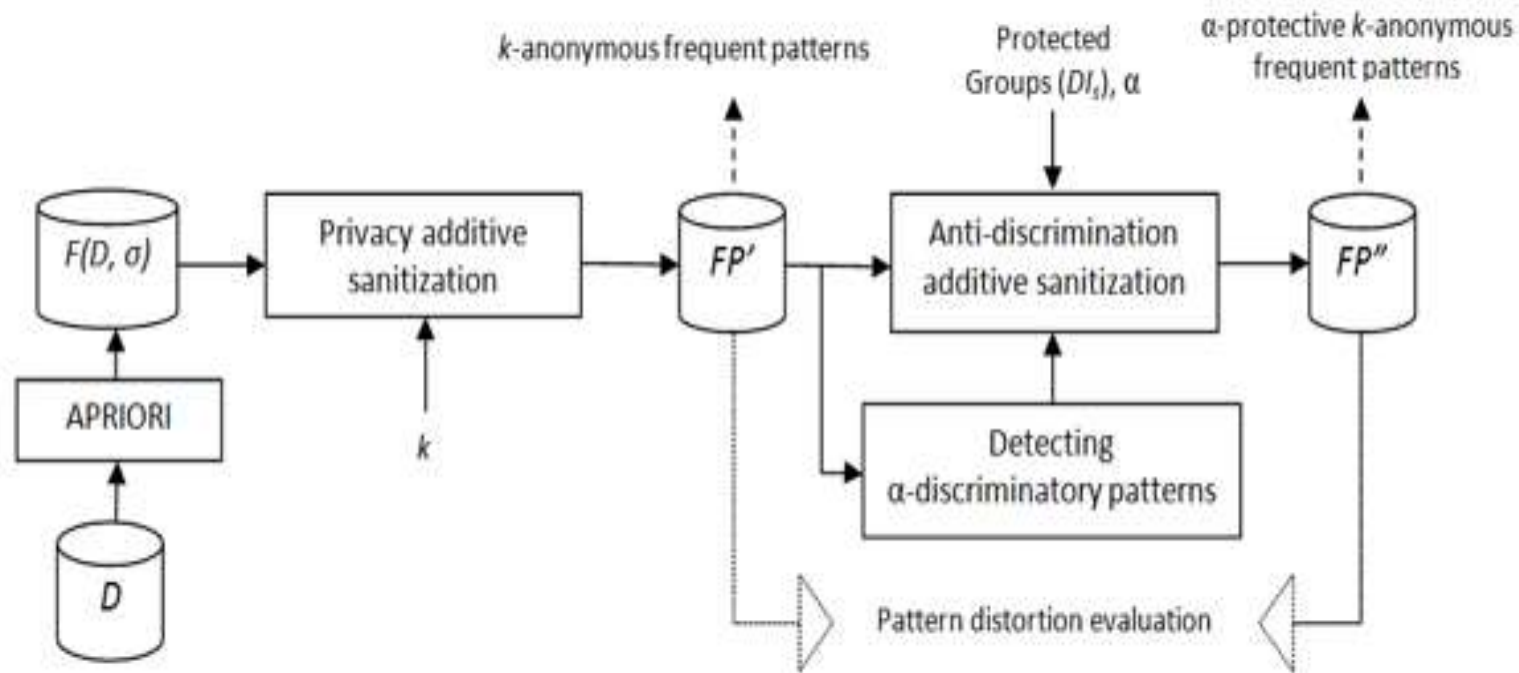
job = veterinarian  $\rightarrow$  credit-approved = no (support 41)

Supp(job = veterinarian, salary = high, credit-approved = no) = 1

In the dataset there is only one veterinarian with high salary.

If somebody knows a veterinarian with high salary, can imply that he/she got credit denied.

# Overall post-processing approach



# Formal results

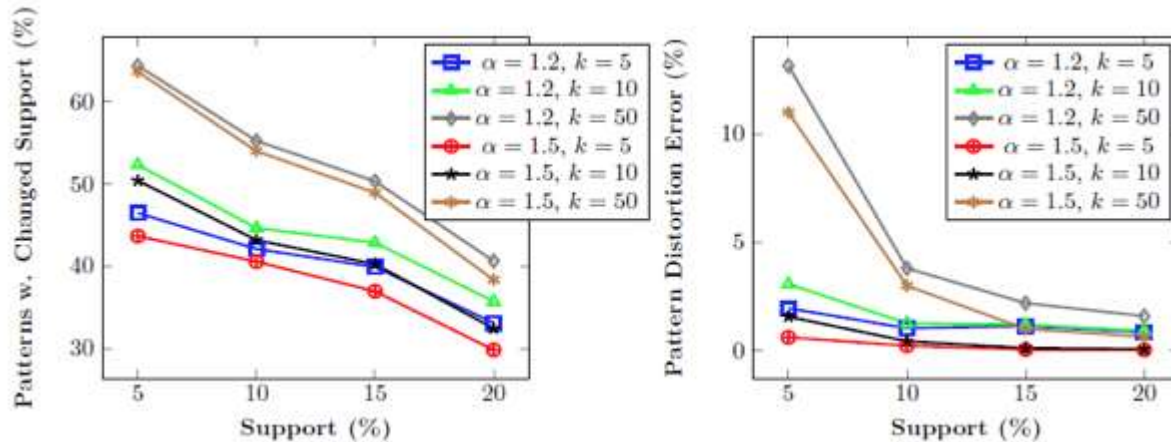
**Theorem:** Anti-discrimination pattern sanitization for making  $F(D,s)$   $a$ -protective **does not** generate **new discrimination** as a result of its transformation.

**Theorem:** Using anti-discrimination pattern sanitization for making  $F(D,s)$   $a$ -protective **cannot** make  $F(D,s)$  non- $k$ -anonymous

**Theorem:** Using privacy pattern sanitization for making  $F(D,s)$   $k$ -anonymous can make  $F(D,s)$  **more or less**  $a$ -protective.

# Evaluation

Pattern distortion scores to make the Adult dataset  $\alpha$ -protective  $k$ -anonymous



## Discrimination-and privacy-aware patterns

Shows that **privacy pattern sanitization methods** based on either  $k$ -anonymity or differential privacy can work **against fairness**.

Proposes new **anti-discrimination pattern sanitization methods** that do not interfere with a privacy-preserving sanitization based on either  $k$ -anonymity or differential privacy.

Shows that the utility loss caused by simultaneous anti-discrimination and privacy protection is only **marginally higher** than the loss caused by each of those protections separately.

# Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining

Part I: Introduction and context

Part II: Discrimination discovery

Part III: Fairness-aware data mining

 **Part IV: Challenges and directions for future research**

Discussion and further questions

# Challenges: the ground-truth problem

- The trade-off between data utility and discrimination avoidance



- Utility based on potentially biased training data!
- Hard to assess the quality of the results
- Lack of datasets and benchmarks



# Challenges: definitions of discrimination

- Unlike for privacy, anti-discrimination legal concepts are diverse and vague
  - Direct vs indirect discrimination
  - Individual vs group fairness
  - Affirmative actions
  - Explainable vs unexplainable discrimination
  - ...
- Current methods in fairness-aware data mining used different definitions of discrimination/fairness
  - No single agreed-upon measure for discrimination/fairness
- How different definitions of fairness affect algorithm design?

# Challenges: interaction with law and policies

- As for research in privacy preservation, there is an interaction between the research on algorithmic fairness and the anti-discrimination regulations:
  - Laws give us the rules of the game: definitions, objective functions, constraints
  - New technical developments need to be taken in consideration by legislators
- However, the communication channel is not clear:
  - Is my data transformation algorithm legal?
  - Can my discrimination-detection algorithm be useful in a real-world case law?
- Wide variety of cases and different interpretations: difficult for a CS to navigate
  - Importance of multidisciplinary
- As usual, many differences between USA and EU regulation

# General Data Protection Regulation (GDPR)

## (Regulation (EU) 2016/679)

- Aims to strengthen and unify data protection for individuals within the EU, as well as setting rules about the export of personal data outside the EU.
- Primary objectives of the GDPR are to **give citizens back the control of their personal data** and to simplify the regulatory environment for international business by unifying the regulation within the EU.
- It deals with concept such as **consent, responsibility, accountability, right to be forgotten**, etc.
- The regulation was adopted on **27 April 2016**. It enters into application **25 May 2018** after a two-year transition period and, unlike a Directive it does not require any enabling legislation to be passed by governments.
- When the GDPR takes effect it will replace the data protection directive (officially Directive 95/46/EC) from 1995.

# Right to explanation (GDPR 2018)

- It will restrict automated decision-making which “significantly affect” individuals.
- An individual can ask for an explanation of an algorithmic decision.
- This law will pose **large challenges for industry**
  - There is a gap between the legislators’ aspirations and technical realities
  - Intentional concealment on the part of corporations or other institutions, where decision making procedures are kept from public scrutiny
  - A “mismatch between the mathematical optimization in high-dimensionality characteristic of machine learning and the demands of human-scale reasoning and styles of interpretation”
- It highlights **opportunities for machine learning researchers** to take the lead in designing algorithms and evaluation frameworks which avoid discrimination.

# Fairness and privacy

- Privacy-preservation
  - How do we prevent sensitive information from being leaked?
- Discrimination-prevention
  - How do we prevent sensitive information from being abused?
- Sensitive features in these two contexts might overlap or not
  - One may not mind other people knowing about their ethnicity, but would strenuously object to be denied a credit or a grant if their ethnicity was part of that decision
- Hiding sensitive information from data due to privacy, might also hide the presence of discriminatory patterns

# A promising direction...

Dealing with privacy-preserving data mining (PPDM) and fairness-aware data mining (FADM) *jointly*...

Share common challenges

Share common techniques

Sometimes one can help the other

PPDM	FADM
Measuring disclosure risk	Measuring potential discrimination
Data, algorithm or model transformation to protect privacy	Data, algorithm or model transformation to prevent discrimination
Measuring data/model utility	Measuring data/model utility
Trade-off between privacy and utility	Trade-off between fairness and utility
...	...

# A promising direction

Dealing with privacy-preserving data mining (PPDM) and fairness-aware data mining (FADM) *jointly*...

[pre-processing] S. Ruggieri. *“Using  $t$ -closeness anonymity to control for non-discrimination”*. Transactions on Data Privacy, 7(2), pp.99-129, 2014.

[pre-processing] S. Hajian, J. Domingo-Ferrer, and O. Farras. *“Generalization-based privacy preservation and discrimination prevention in data publishing and mining”*. Data Mining and Knowledge Discovery, 28(5-6), pp.1158-1188, 2014.

[in-processing] C. Dwork, M. Hardt, T. Pitassi, O. Reingold and R. S. Zemel. *“Fairness through awareness”*. In ITCS 2012, pp. 214-226, 2012.

[post-processing] S. Hajian, J. Domingo-Ferrer, A. Monreale, D. Pedreschi, and F. Giannotti. *“Discrimination-and privacy-aware patterns”*. In Data Mining and Knowledge Discovery, 29(6), 2015.

# Future work: beyond binary classification

So far ...

mostly binary classification problems such as "HIRE" vs "DON'T HIRE"

Future ...

Multi-class and multi-label classification settings

Regression settings

Noisy input data

Multiple protected characteristics

Potentially missing protected characteristics



# Future work: beyond classification

General theory of algorithmic fairness

Fairness in recommender systems and personalization

Fairness in network data (e.g., hire/don't hire based on social network)

Fairness in text data (e.g., automatically detect sexist or racist text)

Tools for discovering discrimination practices in different online settings

E.g, google image search, Airbnb hosts with racist behavior, price discrimination (see e.g., \$heriff tool), ads targeting discrimination (see e.g., Adfisher tool)

# Acknowledgments

- Ricardo Baeza-Yates, Toon Calders, Ciro Cattuto, Faisal Kamran, Salvatore Ruggieri, Suresh Venkatasubramanian and many many other researchers in the area which (more or less consciously) donated their slides to made this tutorial possible.
- This tutorial was partially supported by
  - TYPES: European Union's Horizon 2020 innovation action programme under grant agreement No 653449
  - FARADAY: Partially supported by the Catalonia Trade and Investment Agency (Agència per la competitivitat de l'empresa, ACCIÓ).

# Thank you!

**Sara Hajian**

**Francesco Bonchi** @francescobonchi

**Carlos Castillo** @chatox

Slides will be made available at the tutorial webpage:

[http://francescobonchi.com/algorithmic\\_bias\\_tutorial.html](http://francescobonchi.com/algorithmic_bias_tutorial.html)