Algorithmic Discrimination

Carlos Castillo @chatox

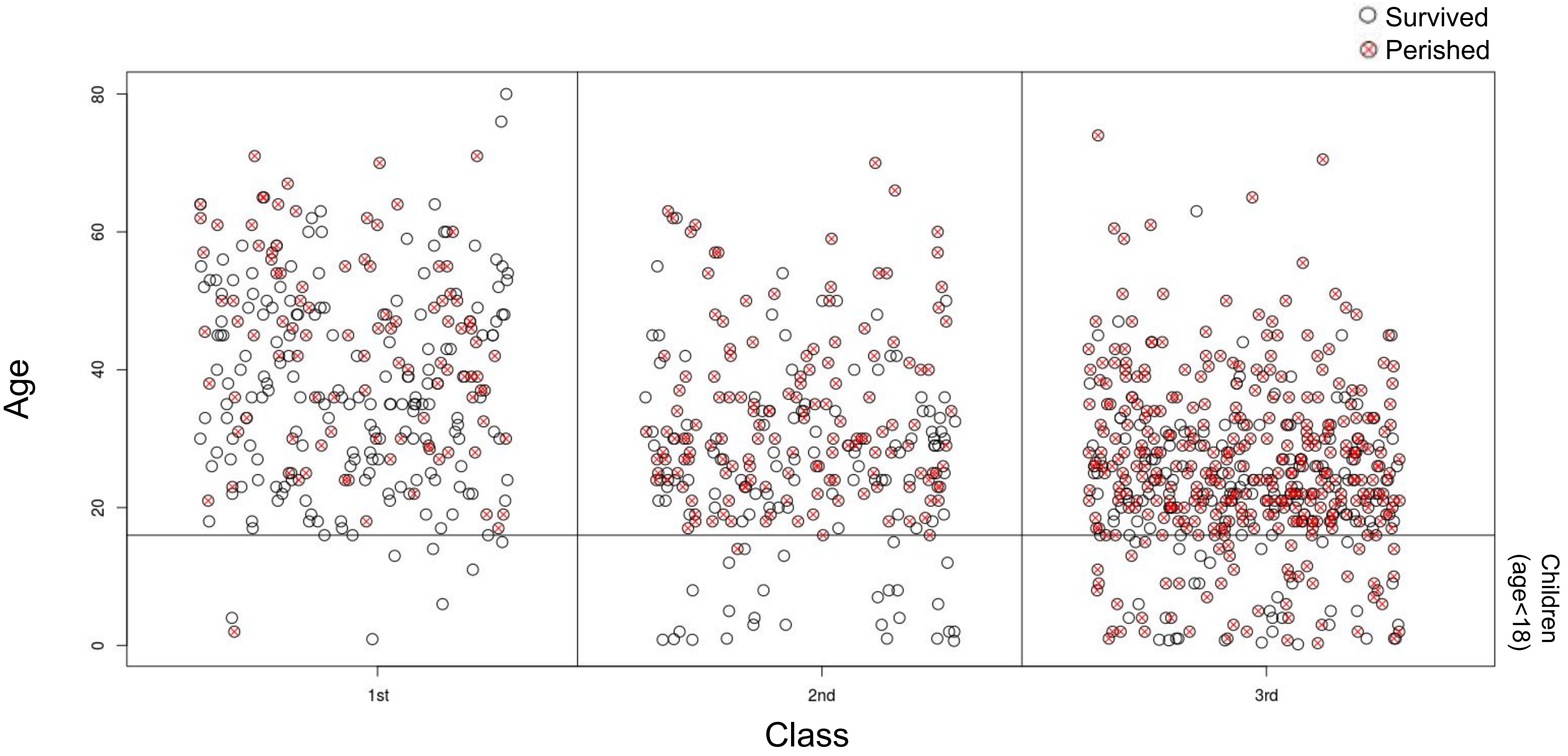
Partially based on KDD 2016 and IC2S2 2017 tutorials with Sara Hajian and Francesco Bonchi

University of Oviedo — February 2019

Part I

Part I	**Motivation**
Part II	Concepts
Part III	Data Mining for Discrimination Discovery
Part IV	Fairness in Classification
Part V	Ideas

Example 0: Titanic (April 1912)



Varian, H. R. (2014). Big data: New tricks for econometrics. The Journal of Economic Perspectives, 28(2), 3-27.

A dangerous reasoning

To discriminate is to treat someone differently

(Unfair) discrimination is based on group membership, not individual merit

People's decisions include objective and subjective elements

Hence, they can be discriminate

Algorithmic inputs include only objective elements

Hence, they cannot discriminate?

Algorithmic discrimination scenarios

Algorithmic systems that generate inputs determining …

- Access to employment
- Access to education
- Access to government benefits
- Access to penitentiary alternatives
- ...

Example 1: COMPAS scores

NORTHPOINTE
Advancing Justice. Embracing Community.

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions): 137-questions questionnaire and predictive model for "risk of recidivism"

Prediction accuracy of recidivism for blacks and whites is about 60%, but ...

- Blacks that did not reoffend
  were classified as high risk twice as much as whites that did not reoffend
- Whites who did reoffend
  were classified as low risk twice as much as blacks who did reoffend

Pro Publica, May 2016. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Example 2: Google Ads

AdFisher: tool to automate the creation of behavioral and demographic profiles.

Used to demonstrate that setting gender = female results in less ads for high-paying jobs.

A. Datta, M. C. Tschantz, and A. Datta (2015). Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies, 2015(1):92–112.

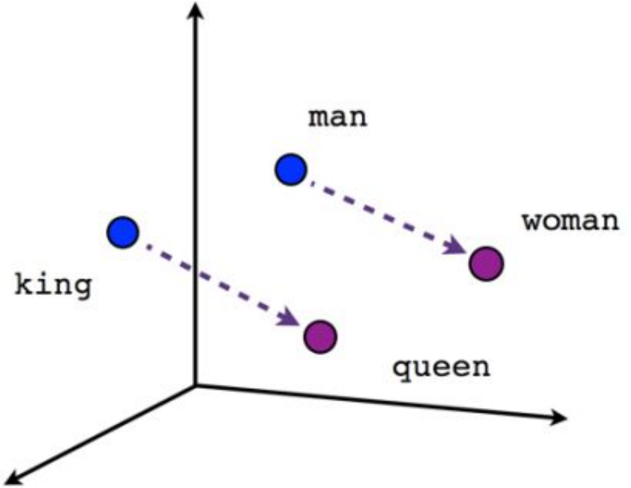
Example 3: Word Embeddings

By difference: he ~ king ⇒ she ~ queen

(arithmetically, he - she ≃ king - queen)

But also …

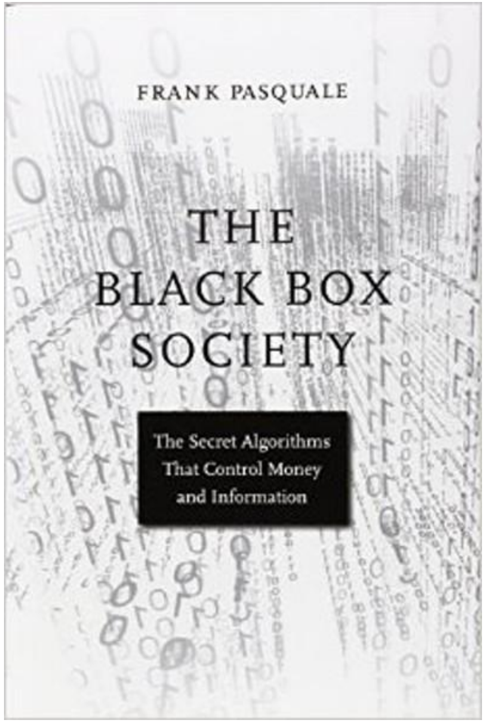
- he ~ pizzas ⇒ she ~ cupcakes
- he ~ architect ⇒ she ~ hairdresser
- he ~ computer programmer ⇒ she ~ homemaker
- he ~ surgeon ⇒ she ~ nurse



Potentially consequences for machine translation

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357).

To complicate things ...



Algorithms are "black boxes" protected by

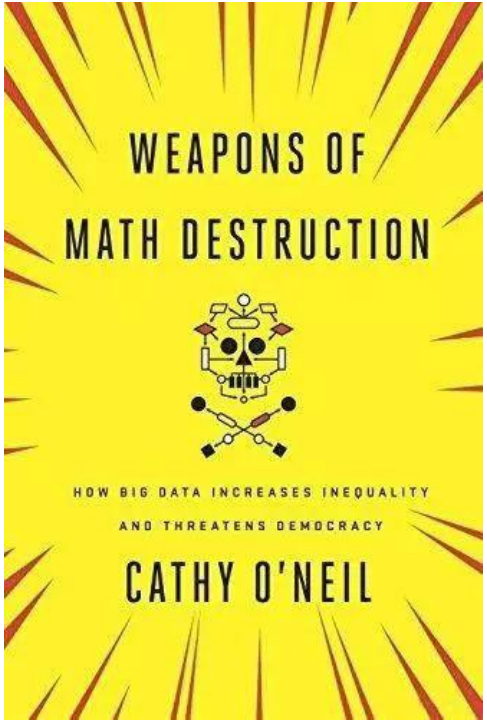
- Industrial secrecy
- Legal protections
- Intentional obfuscation

Discrimination becomes invisible

Mitigation becomes impossible

F. Pasquale (2015): The Black Box Society. Harvard University Press.

Weapons of Math Destruction (WMDs)



"[We] treat mathematical models as a neutral and inevitable force, like the weather or the tides, we abdicate our responsibility" - Cathy O'Neil

Three main ingredients of a "WMD":

- Opacity
- Scale
- Damage

C. O'Neil (2016): Weapons of Math Destruction. Crown Random House. See also TED Talk on April 2016.

Two areas of concern: data and algorithms

Data inputs:

- Poorly selected (e.g., observe only car trips, not bicycle trips)
- Incomplete, incorrect, or outdated
- Selected with bias (e.g., smartphone users)
- Perpetuating and promoting historical biases (e.g., hiring people that "fit the culture")

Algorithmic processing:

- Poorly designed matching systems
- Personalization and recommendation services that narrow instead of expand user options
- Decision making systems that assume correlation implies causation
- Algorithms that do not compensate for datasets that disproportionately represent populations
- Output models that are hard to understand or explain hinder detection and mitigation of bias

Executive Office of the US President (May 2016): "Big Data: A Report on Algorithmic Systems,Opportunity,and Civil Rights"

Part II

Part I	Motivation
Part II	**Concepts**
Part III	Data Mining for Discrimination Discovery
Part IV	Fairness in Classification
Part V	Ideas

Legal concepts

Anti-discrimination legislation typically seeks **equal access** to employment, working conditions, education, social protection, goods, and services

Anti-discrimination legislation is very diverse and includes **many legal concepts:**

Genuine occupational requirement (male actor to portray male character)

Disparate impact and **disparate treatment**

Burden of proof and **situation testing**

Group under-representation principle

Discrimination: treatment vs impact

Modern legal frameworks offer various levels of **protection** for being discriminated by belonging to a particular class of: gender, age, ethnicity, nationality, disability, religious beliefs, and/or sexual orientation

Disparate **treatment** or **direct discrimination**:

Treatment depends on class membership

Disparate **impact** or **indirect discrimination**:

Outcome depends on class membership

Even if (apparently?) people are treated the same way

Europe and the GDPR

The GDPR is the European General Data Protection Regulation

While it could be argued that discrimination is **justified** if the discriminatory model maximizes some form of economic utility …

… algorithmic bias **violates GDPR protections** because it

- infringes on the obligation to **safeguard data subject's rights and freedoms**
- infringes on the principle of **fair data processing**, if the outcome is unfair
- infringes on the principle of **accuracy**, if training data are biased

Individuals access rights to their own data may help them uncover bias

Philipp Hacker, 'Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law' (2018) 55 Common Market Law Review, Issue 4, pp. 1143–1185

Principles for quantifying discrimination

Two basic frameworks for measuring discrimination:

Discrimination at the **individual level**: consistency or individual fairness

Discrimination at the **group level**: statistical parity

I. Žliobaitė (2015): A survey on measuring indirect discrimination in machine learning. arXiv pre-print.

Individual fairness is about consistency

Consistency score

$$C = 1 - \sum_i \sum_{yj \in knn(yi)} |y_i - y_j|$$

Where knn(y_i) = k nearest neighbors of y_i

A consistent or individually fair algorithm is one in which similar people experience similar outcomes … but note that perhaps they are all treated equally badly

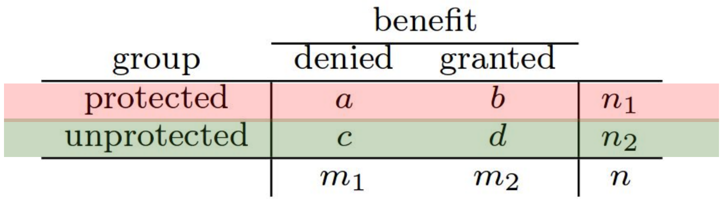
Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In Proc. of the 30th Int. Conf. on Machine Learning. 325–333.

Group fairness is about statistical parity

Example:

"Protected group" = "people with disabilities"

"Benefit granted" = "getting a scholarship"



	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

Intuitively**,** if

a/n_1, the **risk** that **people with disabilities** face of not getting a scholarship
is much larger than

c/n_2, the **risk** that **people without disabilities** face of not getting a scholarship,
then people with disabilities could claim they are being discriminated.

D. Pedreschi, S. Ruggieri, F. Turini: A Study of Top-K Measures for Discrimination Discovery. SAC 2012.

Potentially discriminated (PD) groups

The input to a discrimination discovery task typically includes **potentially discriminated** or **protected groups**:

- Female gender
- Ethnic minority (*racism*) or minority language
- Specific age range (*ageism*)
- Specific sexual orientation (*homophobia*)

These groups are **socially salient** and **disadvantaged**

Part III

Part I	Motivation
Part II	Concepts
Part III	**Data Mining for Discrimination Discovery**
Part IV	Fairness in Classification
Part V	Ideas

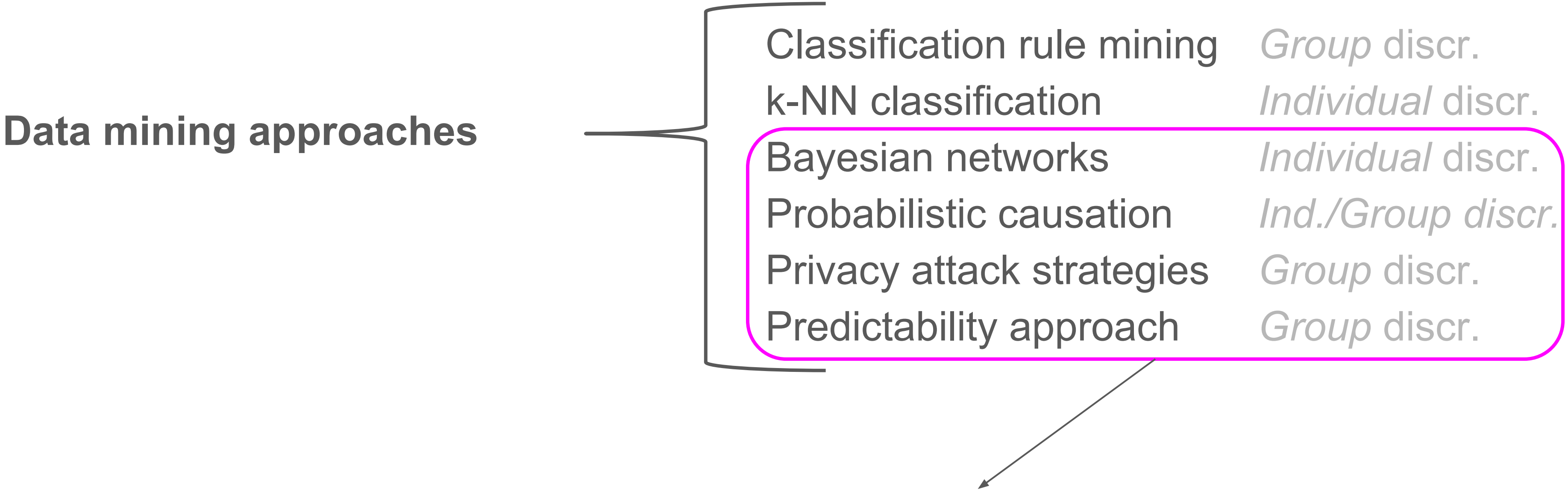
The discrimination discovery task at a glance

Given a large database of historical decision records,

find discriminatory situations and practices.

S. Ruggieri, D. Pedreschi and F. Turini (2010). DCUBE: Discrimination discovery in databases. In SIGMOD, pp. 1127-1130.

Discrimination discovery

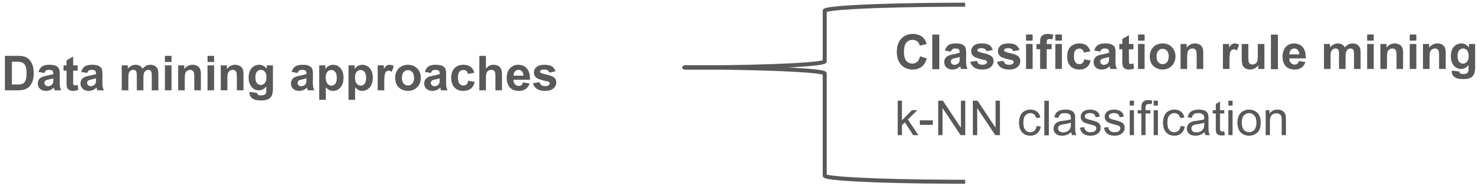


Data mining approaches

Classification rule mining	*Group* discr.
k-NN classification	*Individual* discr.
Bayesian networks	*Individual* discr.
Probabilistic causation	*Ind./Group discr.*
Privacy attack strategies	*Group* discr.
Predictability approach	*Group* discr.

See KDD 2016 tutorial
by S. Hajian, F. Bonchi, C. Castillo

Discrimination discovery



Data mining approaches

Classification rule mining

k-NN classification

B. T. Luong, S. Ruggieri, and F. Turini (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. KDD'11

Direct discrimination

Direct discrimination implies rules or procedures that impose ‘disproportionate burdens’ on minorities

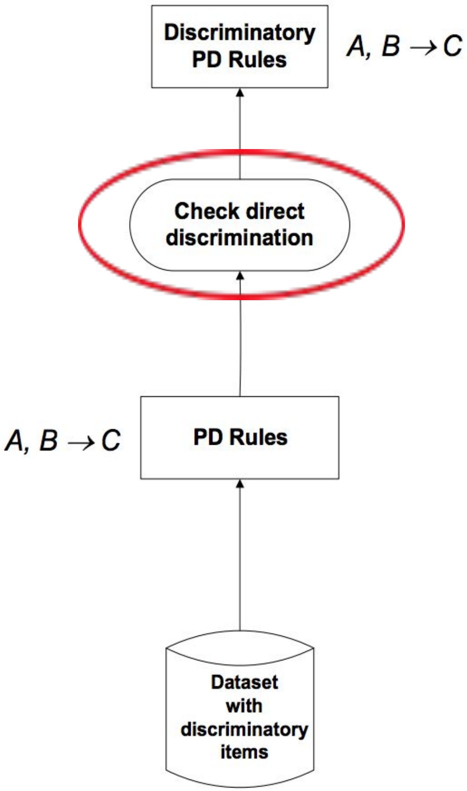
PD rules are any classification rule of the form:

A, B → C

where A is a PD group (B is called a "context")

Example:

gender="female", saving_status="no known savings" → credit=no


Discriminatory PD Rules
A, B → C
Check direct discrimination
A, B → C
PD Rules
Dataset with discriminatory items

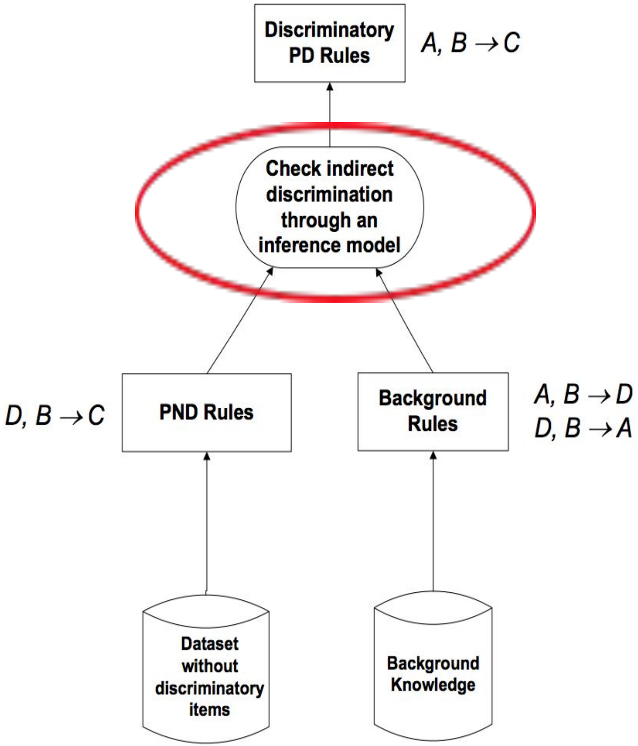
Indirect discrimination

Indirect discrimination implies rules or procedures that impose ‘disproportionate burdens’ on minorities, though not explicitly using discriminatory attributes

Potentially non-discriminatory (PND) rules may unveil discrimination, and are of the form:

D, B → C where D is a PND group

Example:
neighborhood="10451", city="NYC"
→ credit=no



Indirect discrimination example

Suppose we know that with high confidence:

(a) neighborhood=10451, city=NYC → benefit=deny

But we also know that with high confidence:

(b) neighborhood=10451, city=NYC → race=black

Hence:

(c) race=black, neighborhood=10451, city=NYC → benefit=deny

Rule (b) is background knowledge that allows us to infer (c), which shows that **rule (a) is indirectly discriminating against blacks**

Genuine occupational requirements

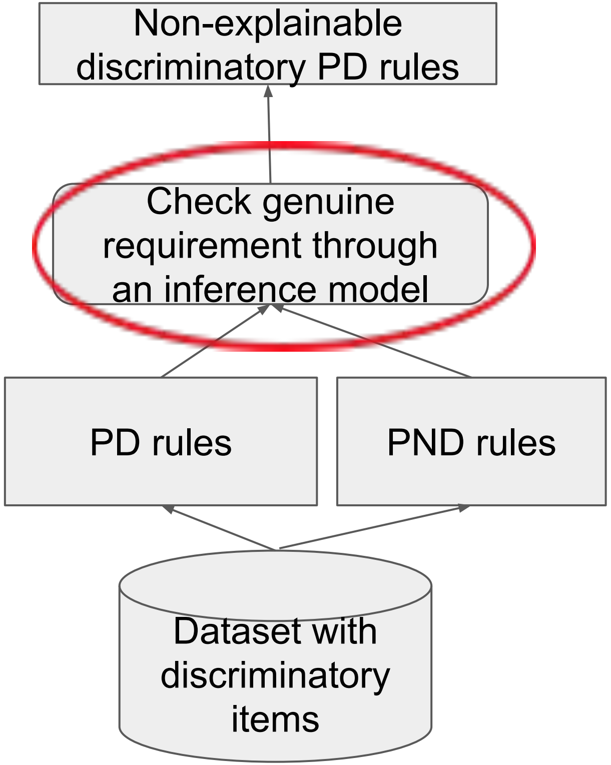
Supported by a PD rule of the form

A, B → C

where C denies some benefit, we search for PND rules of the form

D, B → C

such that D is a legitimate requirement, having the same effects of the PD rule


Non-explainable discriminatory PD rules
Check genuine requirement through an inference model
PD rules
PND rules
Dataset with discriminatory items

D. Pedreschi, S. Ruggieri and F. Turini (2009). Integrating induction and deduction for finding evidence of discrimination. In Proc. of International Conference on Artificial Intelligence and Law (pp. 157-166). ACM.

Example: genuine occupational requirement

(a) [A] gender="female", [B] city="NYC" → [C] hire=no	conf. 0.58
(b) [D] drive_truck="false", [B] city="NYC" → [C] hire=no	conf. 0.81
(c) [A] gender="female", [B] city="NYC" → [D] drive_truck=false	conf. 0.91

Under certain conditions rule (a) is admissible because it is explainable by a genuine occupational requirement (b & c).

D. Pedreschi, S. Ruggieri and F. Turini (2009). Integrating induction and deduction for finding evidence of discrimination. In Proc. of International Conference on Artificial Intelligence and Law (pp. 157-166). ACM.

Discrimination discovery

Data mining approaches

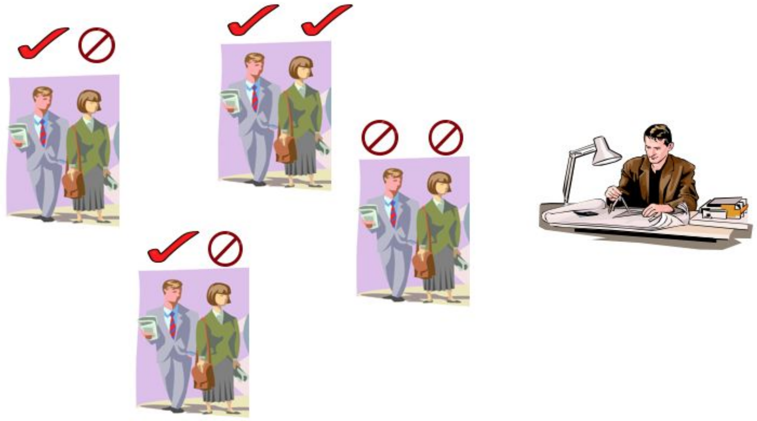
Classification rule mining

k-NN classification

B. T. Luong, S. Ruggieri, and F. Turini (2011). k-NN as an implementation of situation testing for discrimination discovery and prevention. KDD'11

Situation testing

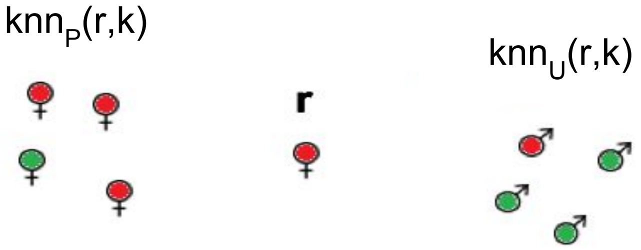
- Legal approach for creating controlled experiments
- Matched pairs undergo the same situation, e.g. apply for a job
    - Same characteristics apart from the discrimination ground



k-NN as situation testing (algorithm)

For r ∈ P(R), look at its k closest neighbors

- ... in the protected set
  - define p_1 = proportion with the same decision as r
- … in the unprotected set
  - define p_2 = proportion with the same decision as r
- measure the degree of discrimination of the decision for r
  - define diff(r) = $p_1 - p_2$ *(think of it as expressed in percentage points of difference)*



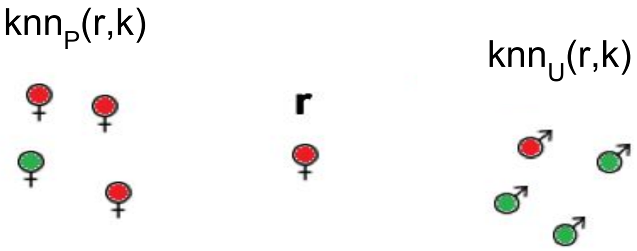
$p_1 = 0.75$

$p_2 = 0.25$

diff(r) = $p_1 - p_2 = 0.50$

Characterizing discrimination using k-NN

- For r ∈ P(R), set a new attribute: "t-discriminated"
  - If dec(r) = deny-benefit and diff(r) ≥ t, t-discriminated(r) := TRUE
    - Otherwise t-discriminated(r) := FALSE
- Example: for t=0.3 the sample r below is classified as t-discriminated



$knn_P(r,k)$

r

$knn_U(r,k)$

$p_1 = 0.75$

$p_2 = 0.25$

$diff(r) = p_1 - p_2 = 0.50$

Characterizing discrimination using k-NN (cont.)

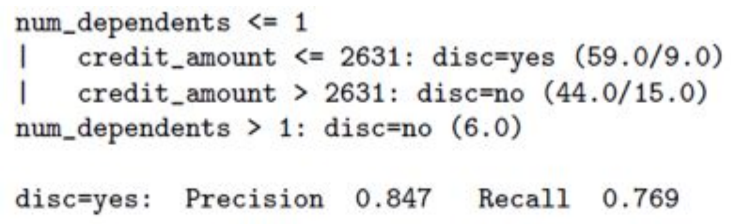
- To answer the question: under which conditions women were t-discriminated?

- We create a classifier with training set P(R), i.e. only protected people, and with class attribute *t-discriminated*

Characterizing discrimination using k-NN (results)

- German credit dataset
  - protected = female non-single
  - 0.10-discriminated cases

- Decision tree model (C4.5)

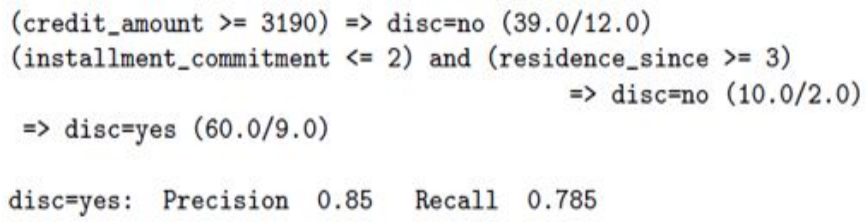


```
num_dependents <= 1
|   credit_amount <= 2631: disc=yes (59.0/9.0)
|   credit_amount > 2631: disc=no (44.0/15.0)
num_dependents > 1: disc=no (6.0)

disc=yes:  Precision  0.847   Recall  0.769
```

Discriminated women had no dependents (children) and were asking for small amounts

- Classification rule model (RIPPER)



```
(credit_amount >= 3190) => disc=no (39.0/12.0)
(installment_commitment <= 2) and (residence_since >= 3)
                                   => disc=no (10.0/2.0)
 => disc=yes (60.0/9.0)

disc=yes:  Precision  0.85   Recall  0.785
```

Discriminated women were asking for small amounts and were either paying in many installments or had been resident for a short time

Part IV

Part I	Motivation
Part II	Concepts
Part III	Data Mining for Discrimination Discovery
Part IV	**Fairness in Classification**
Part V	Ideas

Setting

Let $\boldsymbol{x} \in \mathbb{R}^d$ be a feature vector representing a person

Let $z \in \{ 0, 1 \}$ indicate group labels; with a protected group ($z=1$)

E.g., $z=1$ may indicate minority race, underrepresented gender, disability

Let $y \in \{ 0, 1\}$ indicate labels; with a positive/beneficial label ($y=1$)

E.g., $y=1$ may indicate being hired, getting a credit, receiving a scholarship

Let $\hat{y} \in \{ 0, 1\}$ indicate predictions

Let $X, Z, Y, \hat{Y}$ represent people, protected groups, labels, and inferences

X: attributes
Z: groups
Y: labels
$\hat{Y}$: predictions

Definition 1: "Color blindness"

To satisfy **color blindness**, i.e., **avoid disparate treatment**, the group z should not have any influence on the prediction

$$Pr(\hat{Y}=\hat{y} \mid X=\mathbf{x}, Z=z) = Pr(\hat{Y}=\hat{y} \mid X=\mathbf{x}, Z\neq z) \quad \forall \mathbf{x}, \hat{y}, z$$

This is a problematic definition

X: attributes
Z: groups
Y: labels
$\hat{Y}$: predictions

Definition 2: Avoiding disparate impact

To satisfy **demographic parity**, i.e., **avoid disparate impact**, we want:

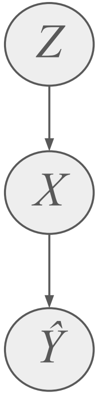
$$Pr(\hat{Y}=\hat{y} \mid Z=1) = Pr(\hat{Y}=\hat{y} \mid Z=0) \quad \forall \hat{y}$$

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

This is in general impossible because:

X usually depends on Z
(e.g., zip code depends on race)

$\hat{Y}$ depends on X



Definition 3: Calibration

X: attributes

Z: groups

Y: labels

$\hat{Y}$: predictions

To satisfy **calibration** or **predictive parity**, we want:

$$Pr(Y=1 \mid \hat{Y}=\hat{y}, Z=1) = Pr(Y=1 \mid \hat{Y}=\hat{y}, Z=0) \quad \forall \hat{y}$$

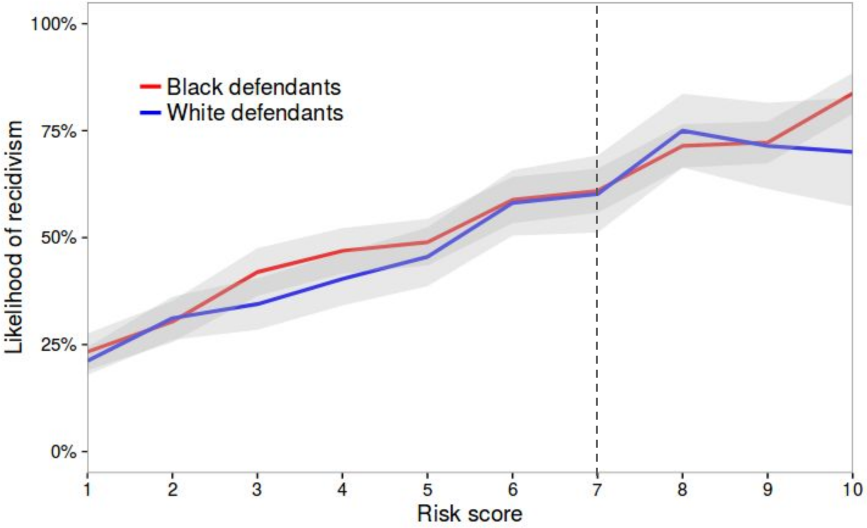
$$Y \perp\!\!\!\perp Z \mid \hat{Y}$$

E.g., among people with a risk score $\hat{y}$=7

60% of blacks reoffended and

61% of whites reoffended

Source: Corbett-Davies et al. 2017



Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In Proc. ITCS.
Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In Proc. of KDD.

X: attributes

Z: groups

Y: labels

$\hat{Y}$: predictions

Definition 4: Equal opportunity

To satisfy **equalized odds**, i.e., **avoiding disparate mistreatment**:

$$Pr(\hat{Y}=1 \mid Y=y, Z=0) = Pr(\hat{Y}=1 \mid Y=y, Z=1) \quad \forall y$$ (Equalized odds)

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

Relaxation: **equal opportunity** refers only to the positive outcome

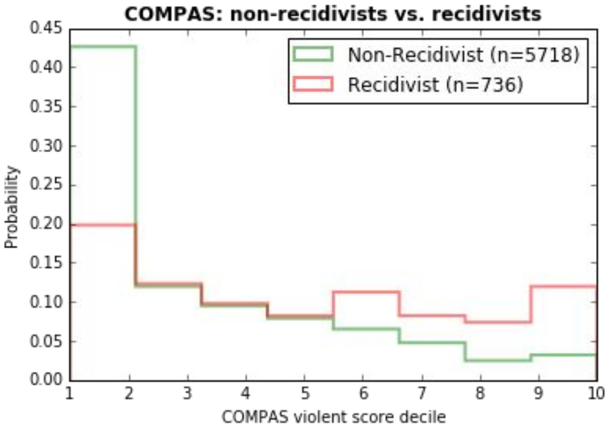
$$\boxed{Pr(\hat{Y}=1 \mid Y=1, Z=0) = Pr(\hat{Y}=1 \mid Y=1, Z=1)}$$

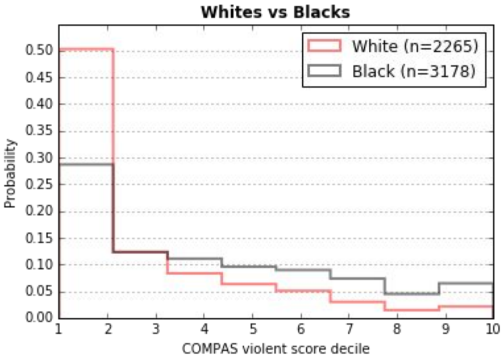
I.e., the true positive rate should be the same for both classes

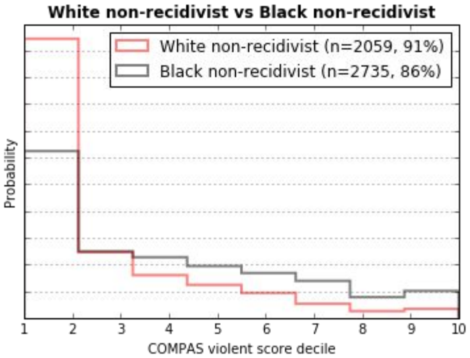
Defined concurrently by Hardt et al. and Zafar et al., who cite each other

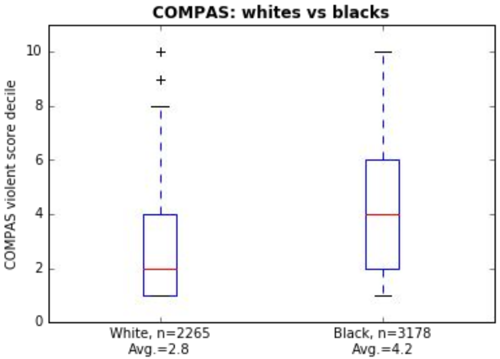
Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Proc. NIPS.
Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact. In Proc. WWW.

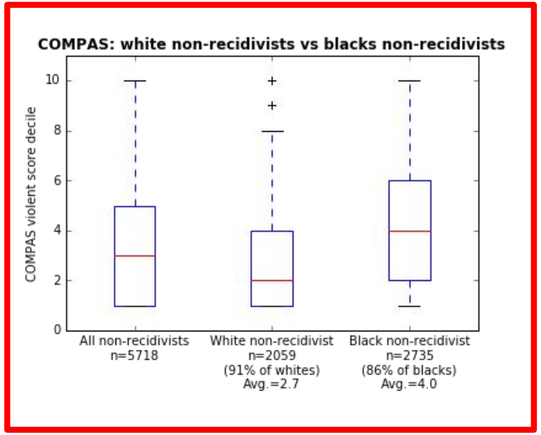
No equal opportunity: COMPAS blacks/whites


COMPAS: non-recidivists vs. recidivists


Whites vs Blacks

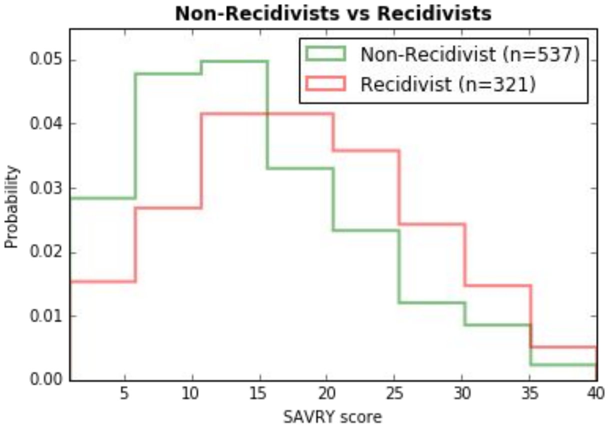

White non-recidivist vs Black non-recidivist

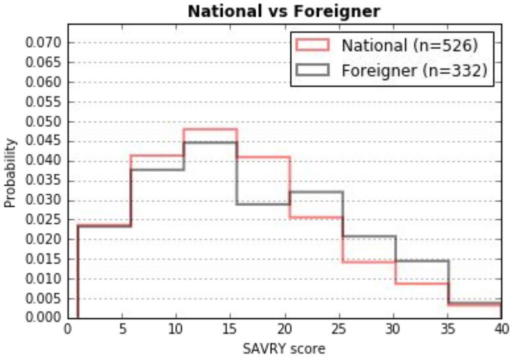

COMPAS: whites vs blacks

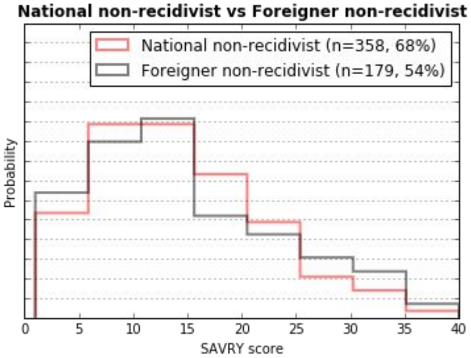

COMPAS: white non-recidivists vs blacks non-recidivists

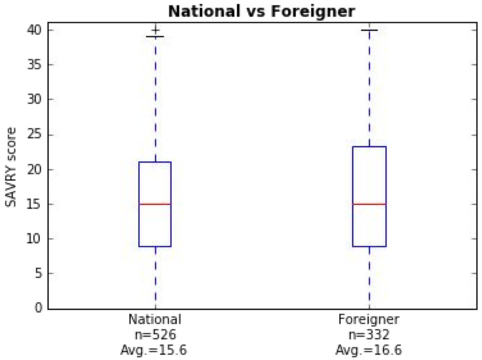
ProPublica 2016 https://github.com/propublica/compas-analysis

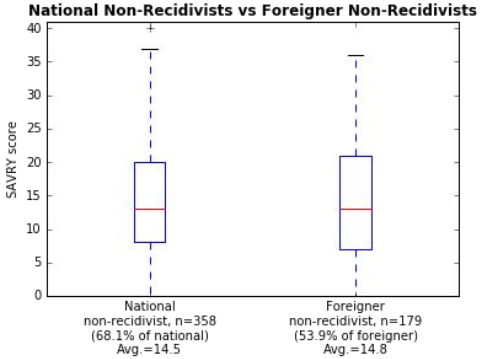
Equal opportunity: SAVRY nationals/foreigners


Non-Recidivists vs Recidivists


National vs Foreigner


National non-recidivist vs Foreigner non-recidivist


National vs Foreigner


National Non-Recidivists vs Foreigner Non-Recidivists

Centre d'Estudis Jurídics i Formació Especialitzada (CEFJE) 2017: http://cejfe.gencat.cat/ca/recerca/cataleg/crono/2017/reincidencia-jj/

Impossibility theorem [Kleinberg et al. 2016]

If $\hat{Y} \neq Y$ (i.e., imperfect classifier) and

$E[Y|Z=1] \neq E[Y|Z=0]$ (i.e., unequal base rates)

Then, the following cannot hold simultaneously:

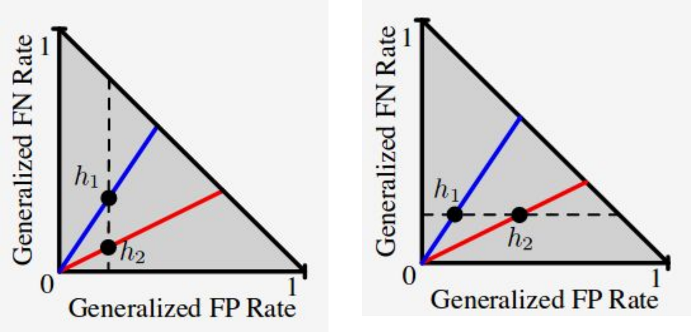
$\hat{Y}$ is *well-calibrated within* each class

$\hat{Y}$ has equal false positive rates in both classes

$\hat{Y}$ has equal false negative rates in both classes

Note: *well-calibrated within* each class

means $Pr(Y=1|\hat{Y}=\hat{y}) = \hat{y}$



The blue and red lines represent calibrated classifiers for each class

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. In Proc. ITCS.
Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. In Proc. NIPS

Part V

Part I	Motivation
Part II	Concepts
Part III	Data Mining for Discrimination Discovery
Part IV	Fairness in Classification
Part V	**Ideas**

Ideas

1. How to apply these to **ranking** settings?
   - We have some results: see our CIKM 2017 paper
   - In general: what are politically acceptable ranking mechanisms that yield equal opportunity?
2. How to develop **simple standardized tests** that are fair?
   - Constrained setting in which each correct question gives one point and the classifier is simply a threshold on the sum of points (simple linear classifier with equal weights, very prevalent)
3. What determines the **cost of being fair**?
   - There are known inherent trade-offs [Kleinberg et al. 2017]

Additional resources

- Presentations/keynotes/book
  - Sara Hajian, Francesco Bonchi, and Carlos Castillo: Algorithmic Bias Tutorial at IC2S2 2017
  - Alexandra Olteanu, Emre Kiciman, Carlos Castillo, Fernando Diaz: Social Data Limits Tutorial at WSDM 2018
  - Workshop on Fairness, Accountability, and Transparency on the Web at WWW 2017
  - Suresh Venkatasubramanian: Keynote at ICWSM 2016
  - Ricardo Baeza: Keynote at WebSci 2016
  - Toon Calders: Keynote at EGC 2016
  - Discrimination and Privacy in the Information Society by Custers et al. 2013
- Groups/workshops/communities
  - Fairness, Accountability, & Transparency in Machine Learning (FATML) workshop and resources
  - Data Transparency Lab - http://dtlconferences.org/