Algorithmic Bias

from measures of discrimination to methods of discovery

Sara Hajian @eurecat.org

NIPS Symposium ML and the Law,
December 8th, 2016. Barcelona, Spain

Introduction and context

Measures of discrimination

Methods of discrimination discovery

Conclusion

- Introduction and context
- Measures of discrimination
- Methods of discrimination discovery
- Conclusion

Introduction and context

- **Examples of algorithmic bias**
- Sources of algorithmic bias
- Legal definitions and principles of discrimination
- Discrimination and privacy
- Solutions

Decision making: humans versus algorithms

- **People's decisions** include objective and subjective elements

- **Algorithmic** inputs include only objective elements

Google image search: gender stereotypes

- Google image search for "**C.E.O.**" produced 11 percent women, even though 27 percent of United States chief executives are women.

M. Kay, C. Matuszek, S. Munson (2015): Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. CHI'15.

Gender bias in Google's Ad-targeting system

- Google's algorithm shows **prestigious job ads** to men, but not to women.

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

Racism

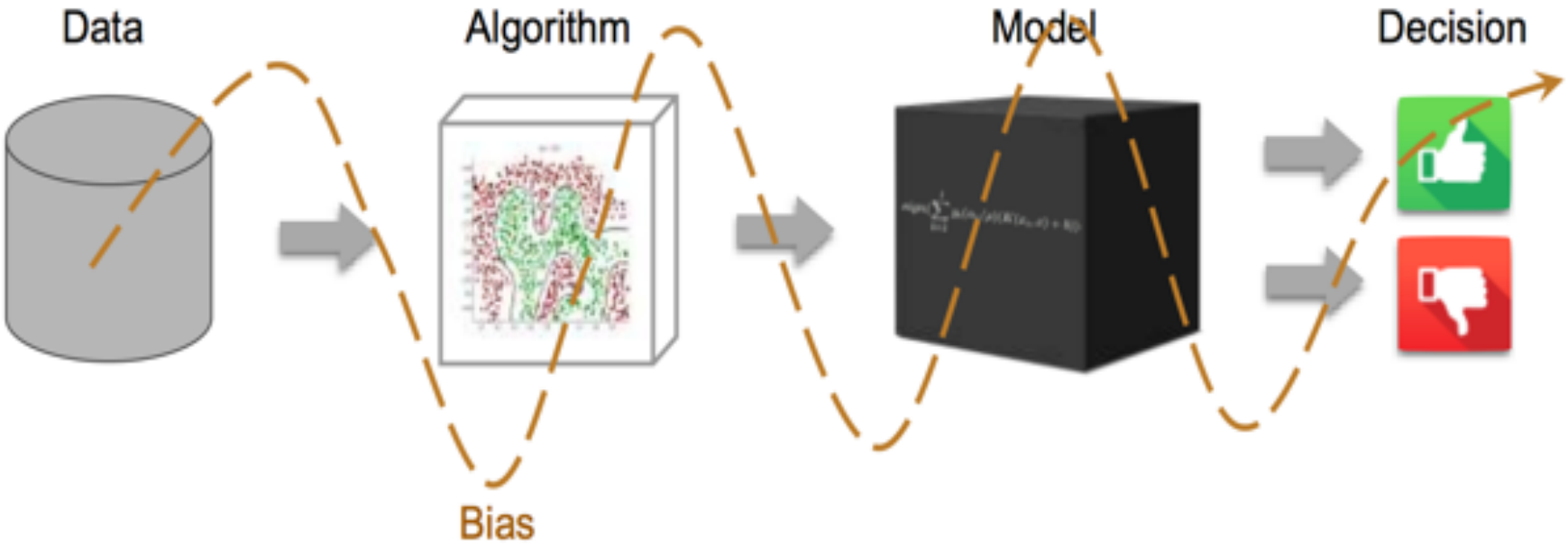
- The importance of being Latanya
- Names used predominantly by black men and women are much more likely to **generate ads related to arrest records**, than names used predominantly by white men and women.

L. Sweeney (2013). Discrimination in online ad delivery. Queue, 11(3). See also N. Newman (2011) in Huffington Post.

Introduction and context

- Examples of algorithmic bias
- **Sources of algorithmic bias**
- Legal definitions and principles of discrimination
- Discrimination and privacy
- Solutions

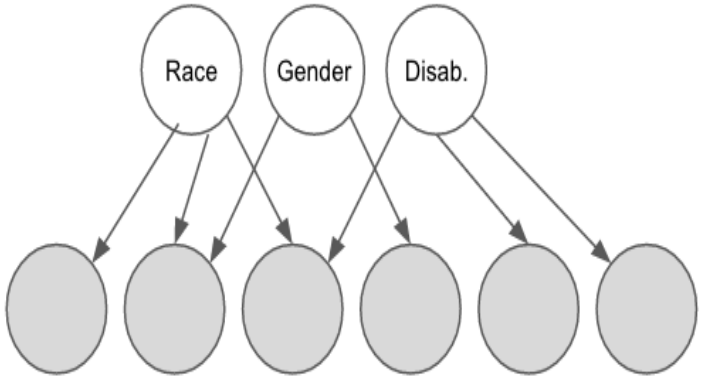
Data-driven decision making process

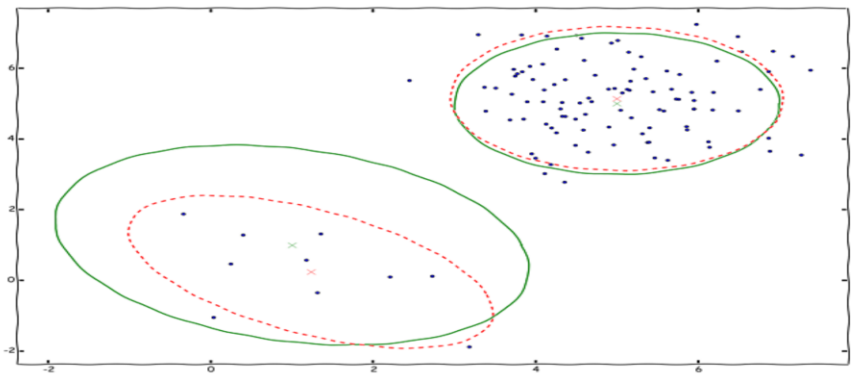


S. Hajian, F. Bonchi and C. Castillo (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. In KDD, pp. 2125-2126.

Sources of algorithmic bias: Data

- Data as a social mirror
- Sample size disparity
- Cultural differences
- Incomplete, incorrect, or outdated data

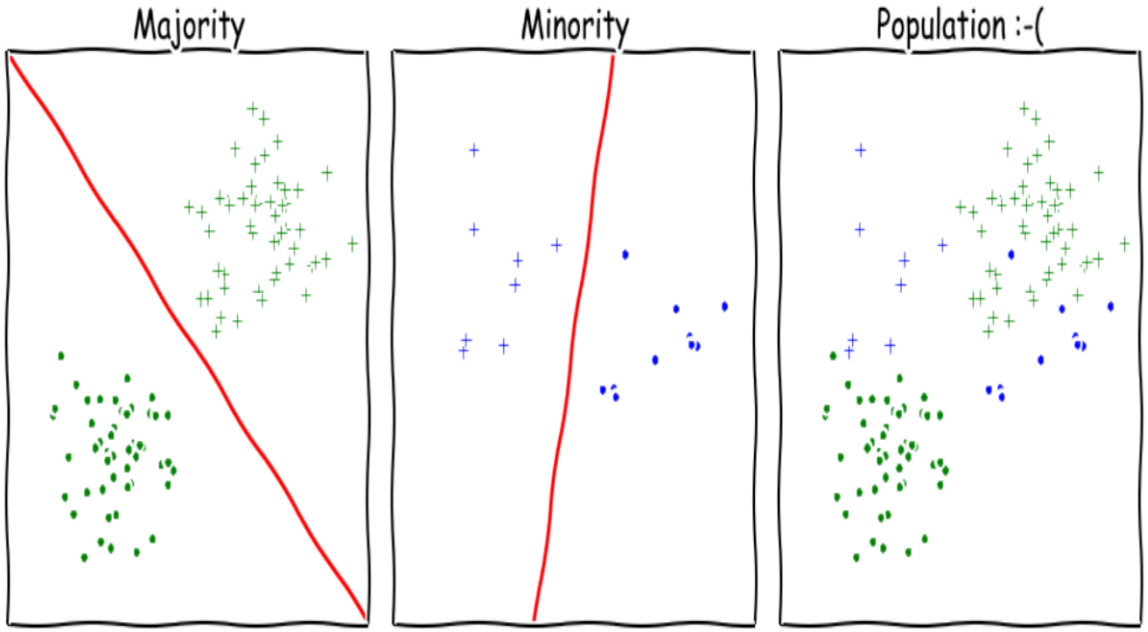




M. Hardt (2014): "How big data is unfair". Medium.

Sources of algorithmic bias: Algorithm and model

- Undesired complexity
- Noise and meaning of 5% error



M. Hardt (2014): "How big data is unfair". Medium.

Introduction and context

- Examples of algorithmic bias
- Sources of algorithmic bias
- **Legal definitions and principles of discrimination**
- Discrimination and privacy
- Solutions

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**:

Treatment depends on class membership

Disparate **impact**:

Outcome depends on class membership

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

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. California Law Review 104.

Introduction and context

- Examples of algorithmic bias
- Sources of algorithmic bias
- Legal definitions and principles of discrimination
- **Discrimination and privacy**
- Solutions

Privacy and data protection legislation differ

Privacy legislation cares about **one action** (storage of personal data) independently of the consequences

Discrimination legislation cares about **one consequence** (unfair treatment) independently of the mechanism

A connection between privacy and discrimination

Finding if people having attribute X were discriminated is like inferring attribute X from a database in which:

- the attribute X was removed
- a new attribute (the decision), which is based on X, was added

This is similar to trying to reconstruct a column from a privacy-scrubbed dataset

Introduction and context

- Examples of algorithmic bias
- Sources of algorithmic bias
- Legal definitions and principles of discrimination
- Discrimination and privacy
- **Solutions**

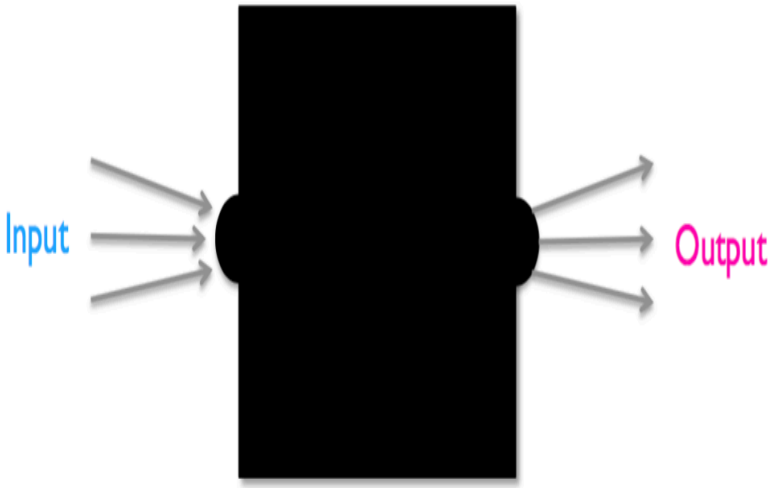
Algorithmic bias: solutions

- **Legal:**
  - Anti-discrimination regulations
    - Give us the rules of the game: **definitions**, objective functions, constraints
    - E.g., General Data Protection Regulation (2018): **Right to explanation**

B. Goodman and S. Flaxman (2016): EU regulations on algorithmic decision-making and a" right to explanation". arXiv preprint arXiv:1606.08813.

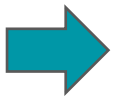
Algorithmic bias: solutions

- **Technical:**
  - Tools for discrimination **risk evaluation**
  - Tools for discrimination **risk mitigation**
  - Tools for **algorithmic auditing**
  - **Explainable** data mining models and user interfaces



Anti-discrimination by design

Introduction and context



Measures of discrimination

Methods of discrimination discovery

Conclusion

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.

Consistency or individual fairness

Consistency score

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

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

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.

Statistical parity focuses on proportions

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 fraction of people with disabilities that **does not** get a scholarship is much **larger** than

c/n_1, the fraction of people without disabilities that **does not** get 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.

Simple discrimination measures

These measures compare the protected group against the unprotected group:

Risk difference = RD = $p_1 - p_2$

Risk ratio or relative risk = RR = p_1 / p_2

Relative chance = RC = $(1-p_1) / (1-p_2)$

Odds ratio = RR/RC

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

$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$

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

Simple discrimination measures

These measures compare the protected group against the unprotected group:

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

Risk difference = RD = $p_1 - p_2$ ← Mentioned in UK law

Risk ratio or relative risk = RR = p_1 / p_2 ← Mentioned by EU Court of Justice

Relative chance = RC = $(1-p_1) / (1-p_2)$

Odds ratio = RR/RC

US courts focus on selection rates: $(1-p_1)$ and $(1-p_2)$

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

Extended discrimination measures

These measures compare the **protected** group against the **entire population**:

Extended risk difference = $p_1 - p$

Extended risk ratio or extended lift = p_1 / p

Extended chance = $(1-p_1) / (1-p)$

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

$$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$$

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

Other measures of discrimination

Differences of mean

Difference of regression coefficients

Rank tests

Mutual information (between outcome and protected attribute)

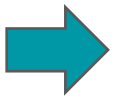
Unexplained difference (residuals of predictions built with non-protected attributes)

Consistency (comparison of prediction with nearest neighbors)

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

Introduction and context

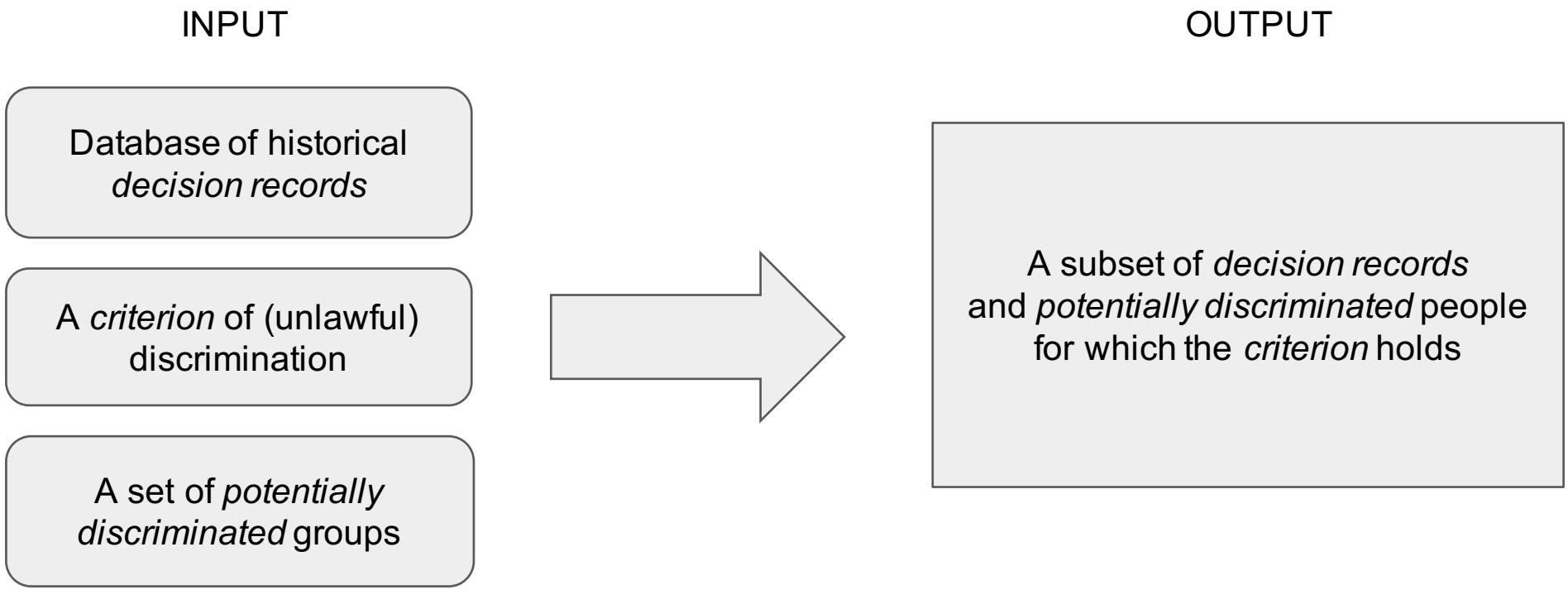
Measures of discrimination



Methods of discrimination discovery

Conclusion

The discrimination discovery task at a glance



D. Pedreschi, S. Ruggieri and F. Turini (2008). Discrimination-Aware Data Mining. In KDD, pp. 560-568.

Why is discrimination discovery hard?

Many different concepts regarding discrimination

- Including all the ones we mention in Part I of this talk

High dimensionality

- There are a huge number of possible contexts that may, or may not, be theater for discrimination.

Hidden indirect discrimination

- The features that may be the object of discrimination may not be directly recorded in the data
- The original data may have been pre-processed due to privacy constraints

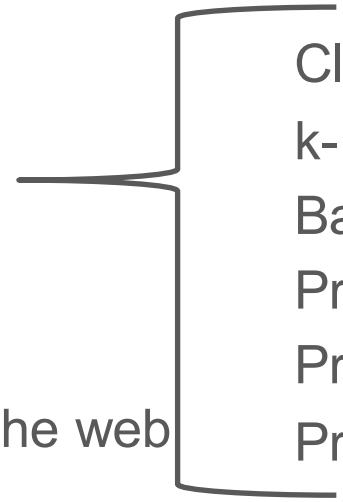
Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web



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.

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- **Classification rule mining**
- k-NN classification
- Bayesian networks
- Probabilistic causation
- Privacy attack strategies
- Predictability approach

D. Pedreschi, S. Ruggieri and F. Turini (2008). Discrimination-aware data mining. In KDD'08.
D. Pedreschi, S. Ruggieri, and F. Turini (2009). Measuring discrimination in socially-sensitive decision records. In SDM'09.
S. Ruggieri, D. Pedreschi, and F. Turini (2010). Data mining for discrimination discovery. In TKDD 4(2).

Defining potentially discriminated (PD) groups

A subset of attribute values are perceived as potentially discriminatory based on background knowledge. Potentially discriminated groups are people with those attribute values.

Examples:

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

Direct discrimination

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

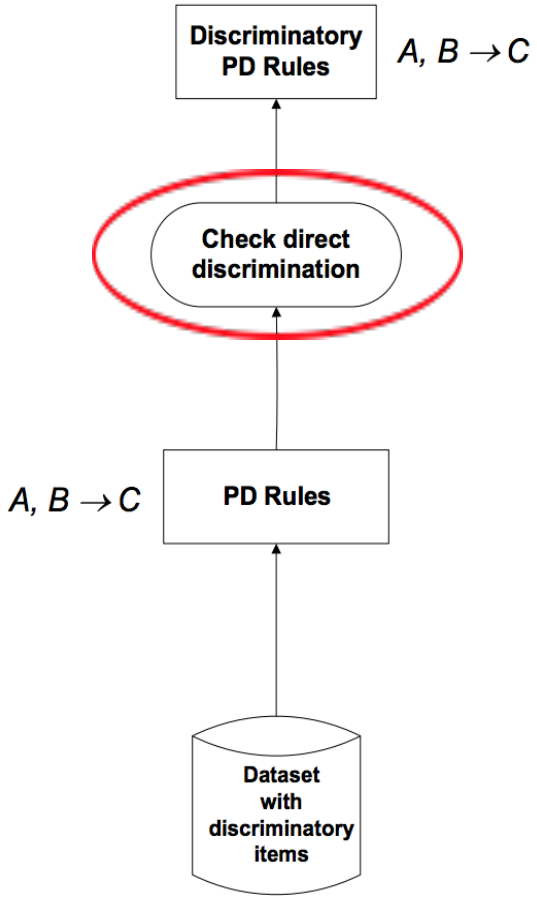
PD rules are any classification rule of the form:

$$A, B \rightarrow C$$

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

Example:

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



The concept of α-protection

For a given threshold α, we say that PD rule A, B → C, involving a PD group A in a context B for an outcome C, is α-protective if:

discrimination measure $f(A, B \rightarrow C) \leq \alpha$

e.g., *f= elift:* $\text{elift}_B(A, B \rightarrow C) = \text{conf}(A, B \rightarrow C) / \text{conf}(B \rightarrow C)$

Otherwise, we say that A, B → C is an α-discriminatory rule

Direct discrimination example

Rule (a):

city="NYC"
→ benefit=deny
with confidence 0.25

Rule (b):

race="black", city="NYC"
→ benefit=deny
with confidence 0.75 **elift 3.0**

Additional (discriminatory) element increases the rule confidence up to 3 times.

According to α-protection method, if the threshold α=3 is fixed then the rule (b) is classified as discriminatory

Indirect discrimination

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

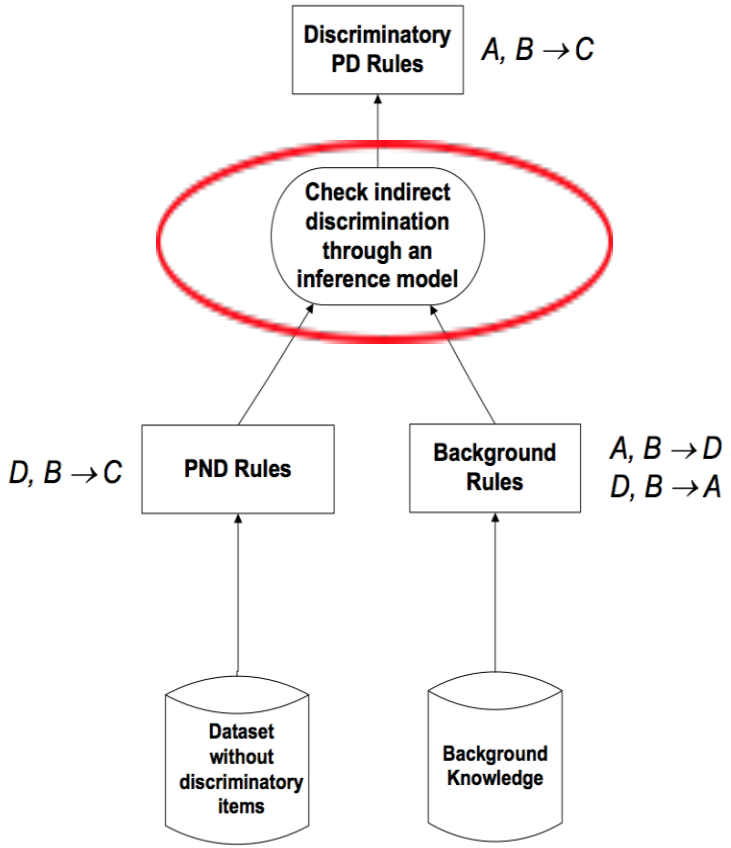
Potentially non-discriminatory (PND) rules may unveil discrimination, and are of the form:

D, B → C where D is a PND group

Example:

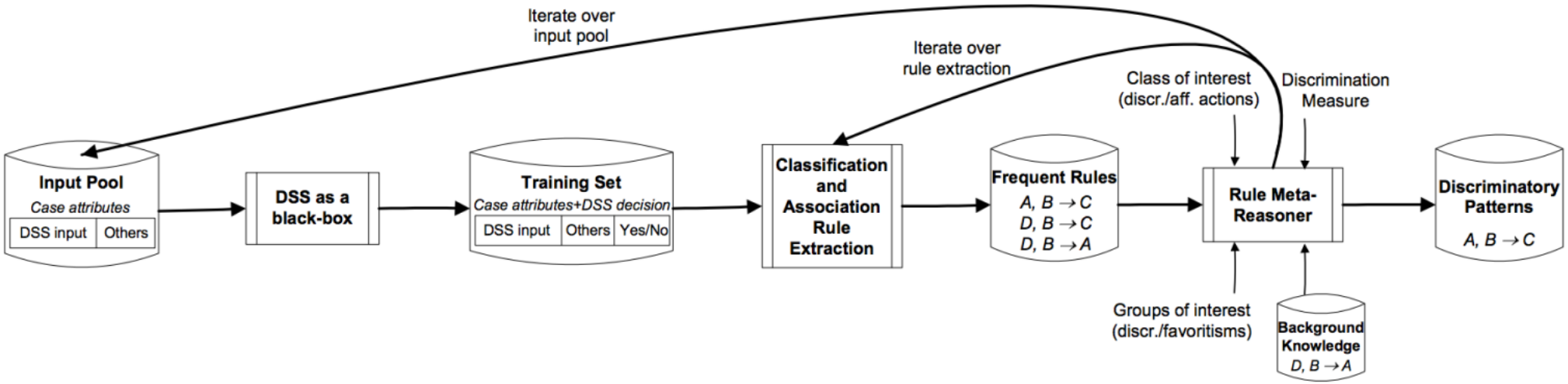
a) neighborhood="10451", city="NYC"
→ credit=no

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



Pipeline for analyzing discrimination

Reference model for analysing and reasoning on discrimination



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

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- **k-NN classification**
- Bayesian networks
- Probabilistic causation
- Privacy attack strategies
- Predictability approach

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

Limitations of classification rules approach

Interpretational limitations

Local contexts, possibly overlapping

No global description of who is discriminated and who is not

Technical limitations

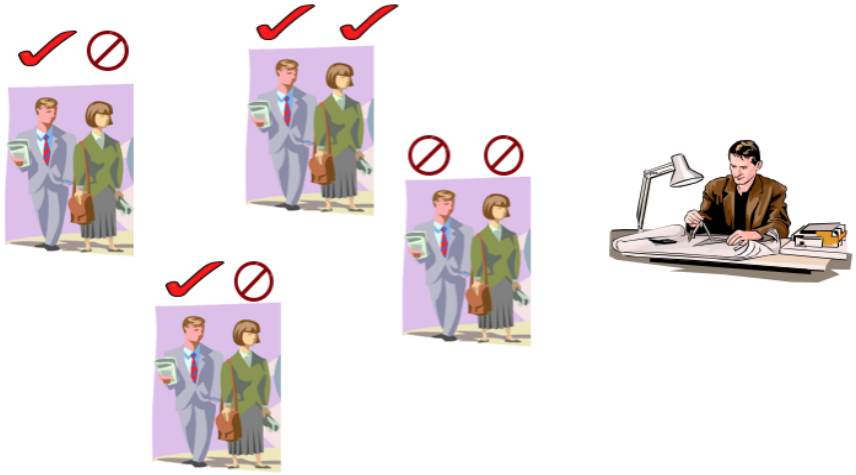
Due to the use of frequent itemset mining (nominal attributes, nominal decisions) it requires discretization

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 $r \in P(R)$, look at its k closest neighbors

... in the protected set

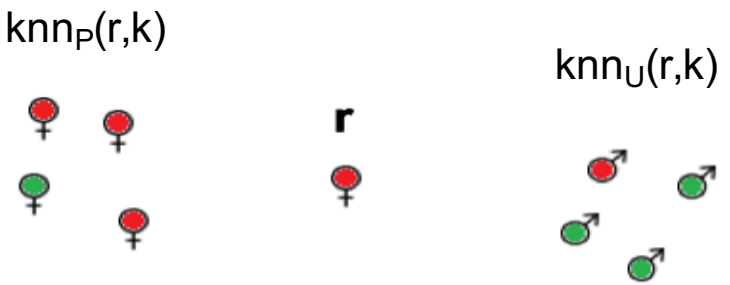
define p_1 = proportion with the same decision as r

… 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)*



$p_1 = 0.75$

$p_2 = 0.25$

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

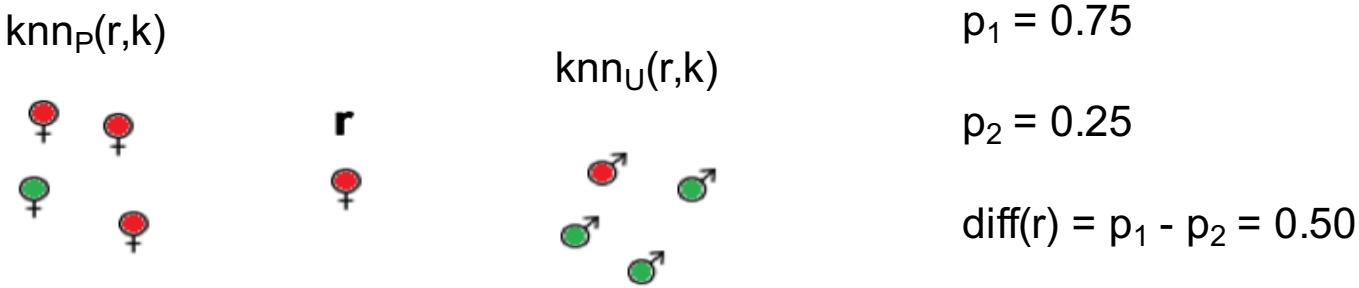
Characterizing discrimination using k-NN

For $r \in P(R)$, 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



$p_1 = 0.75$

$p_2 = 0.25$

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

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

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- k-NN classification
- **Bayesian networks**
- Probabilistic causation
- Privacy attack strategies
- Predictability approach

K. Mancuhan and C. Clifton (2014). Combating discrimination using bayesian networks. In Artificial Intelligence and Law, 22(2).

Let's go back to the classification-based approach

It hinges on computing $\text{elift}_B(A, B \rightarrow C) = \text{conf}(A, B \rightarrow C) / \text{conf}(B \rightarrow C)$

What are these quantities?

$\text{conf}(A, B \rightarrow C)$

is the confidence on an outcome given a protected attribute and a context

$\text{conf}(B \rightarrow C)$

is the confidence on an outcome given just the context

Bayesian networks

- Bayesian networks estimate the probability *P(A, B, C)* by capturing the conditional dependencies between the attributes within the sets *A* and *B*.
  - Bayesian networks can be used to estimate *P(A, B,C)* probability and the *P(C|A,B)* class probability can be derived from the Bayes theorem
  - Bayesian networks capture correlations between attributes
  - Bayesian networks are appropriate to define a decision process
- The *elift* can be extended to *belift* by calculating the numerator and the denominator probabilities with Bayesian networks

$$belift = \frac{P(C \mid a_1, a_2, \ldots, a_l, b_1, b_2, \ldots, b_m, r_1, r_2, \ldots, r_n)}{P(C \mid b_1, b_2, \ldots, b_m)}$$

Bayesian *elift* (*belift*)

$$belift = \frac{P(C \mid a_1, a_2, \ldots, a_l, b_1, b_2, \ldots, b_m, r_1, r_2, \ldots, r_n)}{P(C \mid b_1, b_2, \ldots, b_m)}$$

This indicates how many times the usage of protected attributes (*A*) and the redlining attributes (*R*) increase the class probability for a given instance with respect to its probability using only non-protected attributes (B)

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- k-NN classification
- Bayesian networks
- **Probabilistic causation**
- Privacy attack strategies
- Predictability approach

F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti (2015). Exposing the probabilistic causal structure of discrimination. arXiv:1510.00552.

Previous approaches

- Legal limitations
  - Any legally-valid proof of discrimination requires evidence of causality [Foster, 2004]
- Technical limitations
  - The state-of-the-art methods are essentially correlation-based
  - Spurious correlations can lead to false negatives and false positives
  - A Bayesian network would not be able to disentangle the direction of any causal relationship

Sheila R. Foster (2004). Causation in Antidiscrimination Law: Beyond Intent Versus Impact. Houston Law Review 41, p. 1469

To prove discrimination:

We need to assess discrimination as a causal inference problem from a database of past decisions, where causality can be inferred probabilistically

Suppes probabilistic causation theory (constraints)

Let *h* denote cause, *e* denote effect

Temporal priority

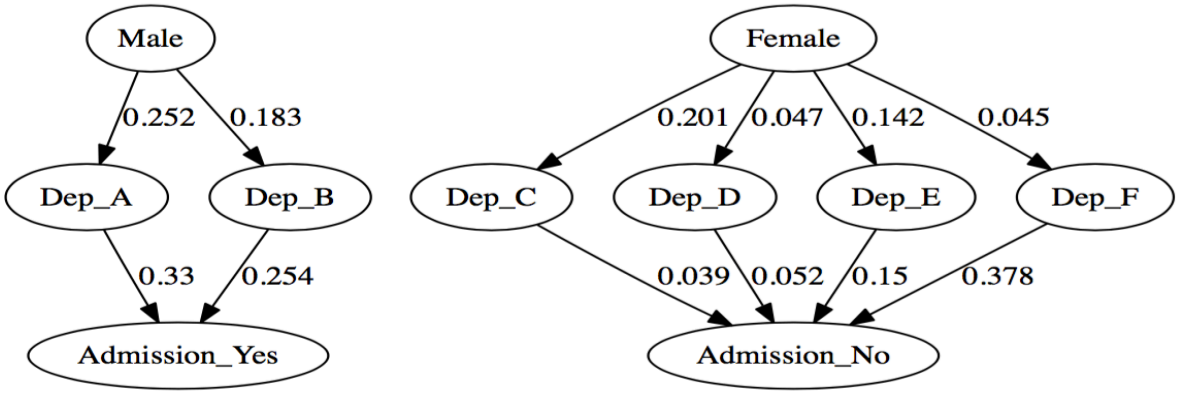
Any cause must happen before its effect: $t_h < t_e$

Probability raising

Cause must raise the probability of observing the effect: $P(e \mid h) > P(e \mid \neg h)$

The Suppes-Bayes Causal Network (SBCN)

- Represents the causal structures existing among the attributes in the data by a constrained Bayesian network:
  - each node represents an assignment attribute=value
  - each arc (u,v) represents the existence of a relation between u and v satisfying Suppes' constraints (temporal priority and probability raising)
  - each arc is labeled with a positive weight: p(u|v) - p(u|¬v)

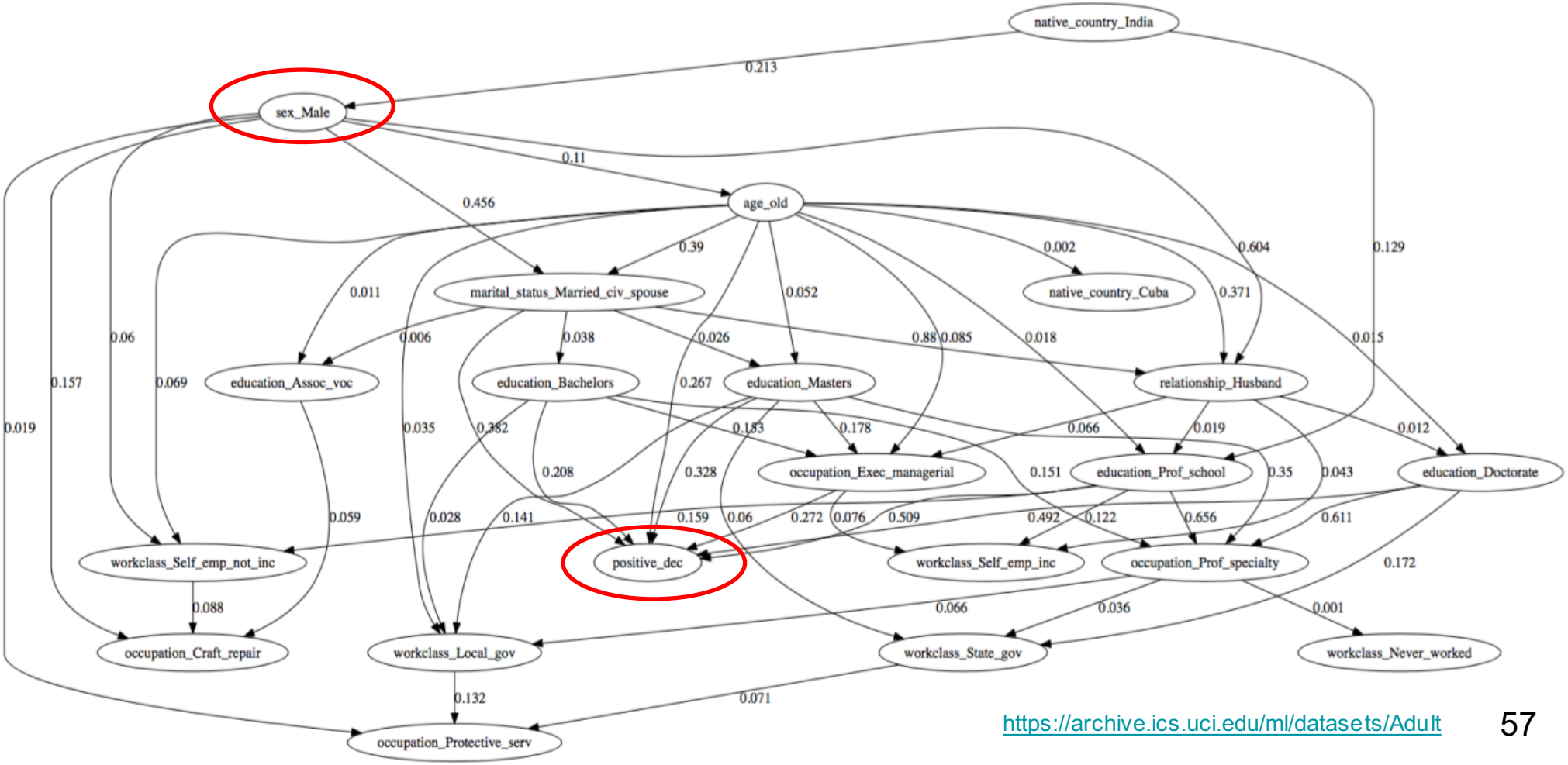


The Suppes-Bayes Causal Network (SBCN)

- Reconstructed from the data using **maximum likelihood estimation (MLE)**, where we force the conditional probability distributions induced by the reconstructed graph to obey Suppes' constraints
  - *Network simplification* by extracting a minimal set of edges which best explain the data. This **regularization** is done by means of the Bayesian Information Criterion (BIC)

$$score_{\mathrm{BIC}}(D', G') = LL(D'|G') - \frac{\log s}{2}\dim(G').$$

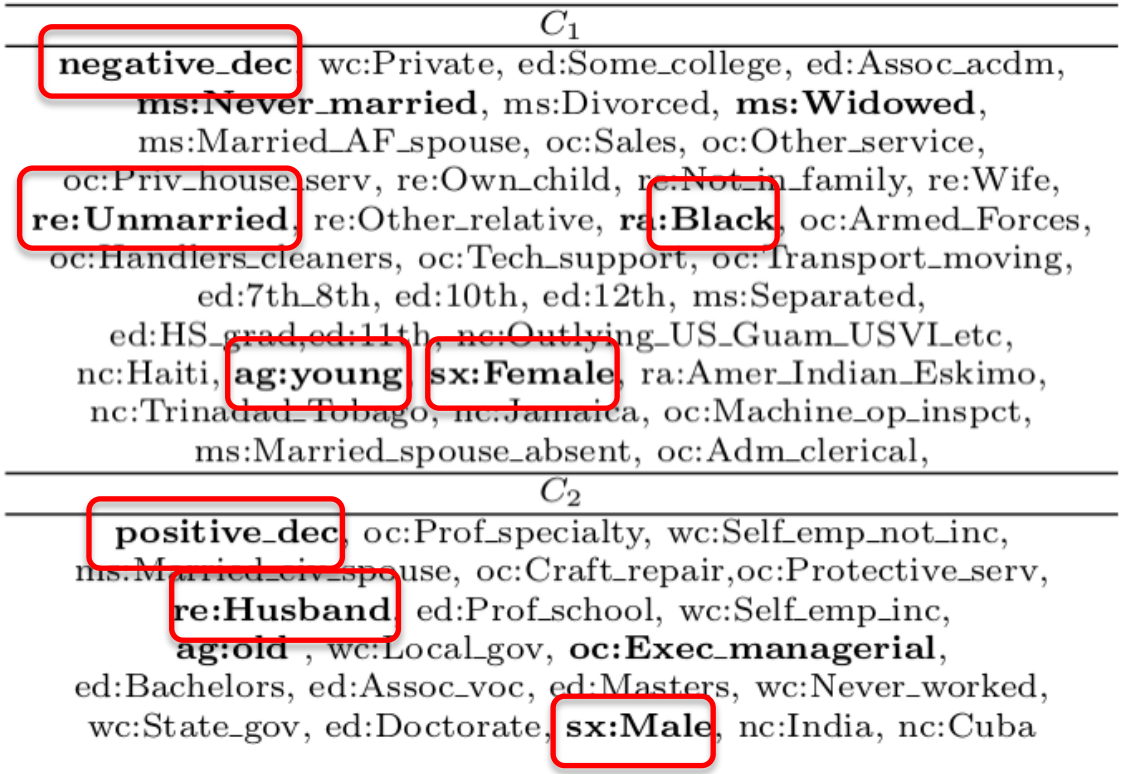
SBCN example: Census Income dataset



https://archive.ics.uci.edu/ml/datasets/Adult

Graph representation allows interesting applications

Two largest communities detected using the Walktrap algorithm [Pons and Latapy 2006]



C_1
negative_dec, wc:Private, ed:Some_college, ed:Assoc_acdm, **ms:Never_married**, ms:Divorced, **ms:Widowed**, ms:Married_AF_spouse, oc:Sales, oc:Other_service, oc:Priv_house_serv, re:Own_child, re:Not_in_family, re:Wife, **re:Unmarried**, re:Other_relative, **ra:Black**, oc:Armed_Forces, oc:Handlers_cleaners, oc:Tech_support, oc:Transport_moving, ed:7th_8th, ed:10th, ed:12th, ms:Separated, ed:HS_grad,ed:11th, nc:Outlying_US_Guam_USVI_etc, nc:Haiti, **ag:young**, **sx:Female**, ra:Amer_Indian_Eskimo, nc:Trinadad_Tobago, nc:Jamaica, oc:Machine_op_inspct, ms:Married_spouse_absent, oc:Adm_clerical,
C_2
positive_dec, oc:Prof_specialty, wc:Self_emp_not_inc, ms:Married_civ_spouse, oc:Craft_repair,oc:Protective_serv, **re:Husband**, ed:Prof_school, wc:Self_emp_inc, **ag:old**, wc:Local_gov, **oc:Exec_managerial**, ed:Bachelors, ed:Assoc_voc, ed:Masters, wc:Never_worked, wc:State_gov, ed:Doctorate, **sx:Male**, nc:India, nc:Cuba

age (ag), education (ed), marital status (ms), native country (nc), occupation (oc), race(ra), relationship (re), sex (sx), workclass (wc)

P. Pons and M. Latapy. (2006). "Computing communities in large networks using random walks." J. Graph Algorithms Appl. 10.2: 191-218.

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- k-NN classification
- Bayesian networks
- Probabilistic causation
- **Privacy attack strategies**
- Predictability approach

S. Ruggieri, S. Hajian, F. Kamiran, X. Zhang (2014). Anti-discrimination analysis using privacy attack strategies. In PKDD'14.

Privacy attack strategies

Methods for direct discrimination discovery require that:

- The dataset explicitly contains an attribute denoting the PD group
  - Otherwise, we are in an indirect discrimination discovery setting
- The dataset has not been pre-processed prior to discrimination discovery
  - Otherwise, we are in a privacy-aware discrimination discovery setting

Discrimination discovery scenarios

Indirect discrimination discovery

Privacy-aware discrimination discovery
(e.g., original data with all attributes is no longer available)

Discrimination data recovery
(e.g., original data with all attributes has been hidden from authorities)

Discrimination discovery scenarios: intuition

Indirect discrimination discovery

Privacy-aware discrimination discovery

Discrimination data recovery

There is an interesting parallel between the role of the anti-discrimination authority in these three scenarios and the role of an attacker in private data publishing

Privacy attack strategies

Combinatorial attacks based on Frèchet bounds inference

Attribute inference attacks

Minimality attacks

We will use risk difference (RD)

group	decision -	decision +	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1$$
$$p_2 = c/n_2$$
$$p = m_1/n$$
$$RD = p_1 - p_2$$

A discrimination table is α-protective (w.r.t. the *risk difference* measure RD) if $RD \leq \alpha$. Otherwise, it is α-discriminatory.

Indirect discrimination discovery

Think of g_1 and g_2 as *redlining* attributes (i.e., correlated with being protected)

group	decision -	decision +	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

unknown contingency table

rel. group	decision -	decision +	
g1	$\hat{a}$	$\hat{b}$	$\hat{n}_1$
g2	$\hat{c}$	$\hat{d}$	$\hat{n}_2$
	m_1	m_2	n

known contingency table

group	rel. group g1	rel. group g2	
protected	e	f	n_1
unprotected	g	h	n_2
	$\hat{n}_1$	$\hat{n}_2$	n

background knowledge contingency table

We do the same with a_2, and a similar decomposition for c, obtaining:

$$min\{\hat{a}, e\} + min\{\hat{c}, f\} \geq a \geq max\{e - \hat{b}, 0\} + max\{f - \hat{d}, 0\}$$
$$min\{\hat{a}, g\} + min\{\hat{c}, h\} \geq c \geq max\{g - \hat{b}, 0\} + max\{h - \hat{d}, 0\}$$

$$RD \geq RDlb = \frac{max\{e - \hat{b}, 0\} + max\{f - \hat{d}, 0\}}{n_1} - \frac{min\{\hat{a}, g\} + min\{\hat{c}, h\}}{n_2}$$

Lower bound

Uses only known and background data

Discrimination discovery

Definition

Data mining approaches

Case studies

Discrimination discovery on the web

- Classification rule mining
- k-NN classification
- Bayesian networks
- Probabilistic causation
- Privacy attack strategies
- **Predictability approach**

M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian (2015). Certifying and removing disparate impact. In KDD'15.

Disparate impact

Measured using risk ratio p_1/p_2

$$\frac{\Pr(\text{decision=deny\_benefit} \mid \text{group=protected})}{\Pr(\text{decision=deny\_benefit} \mid \text{group=unprotected})}$$

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

$$p_1 = a/n_1 \quad p_2 = c/n_2 \quad p = m_1/n$$

Feldman et al. note this is:

1/LR+ = (1-specificity)/sensitivity

LR+ is the likelihood ratio of the positive class

US Equal Employment Opportunity Commision (EEOC) says $p_1/p_2 \leq 0.8$ means disparate impact

M. Feldman, S.A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian (2015). Certifying and removing disparate impact. In KDD'15.

Disparate impact as predictability

Alice runs an algorithm on D = (X, Y), with X protected, Y unprotected

Bob receives (D, C), with C the outcomes

If Bob can predict X based on D\X, C,

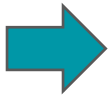
then the algorithm was discriminatory

Essentially, we want to know if the decisions are leaking information about the protected attributes

Introduction and context

Measures of discrimination

Methods of discrimination discovery


Conclusion

Conclusion

- **Bad news:** The algorithms and big data are not just **mirroring** the existing bias but also they are **reinforcing** that bias and amplifying **inequality**

- **Good news:** Algorithmic discrimination: Despite its challenges, it brings a lot **opportunities** for machine learning researchers to build tools for addressing different aspects of this problem

Additional resources

Presentations/keynotes/book

- Sara Hajian, Francesco Bonchi, and Carlos Castillo: Algorithmic Bias Tutorial at KDD 2016
- 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, and Transparency in Machine Learning (FATML) workshop and resources
- Data Transparency Lab - http://dtlconferences.org/