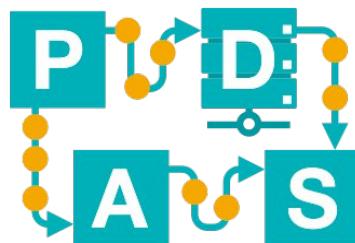


# Responsible Data Science (1/2)

Lecture 20

IDS-L20



Chair of Process  
and Data Science

RWTH AACHEN  
UNIVERSITY

# Outline of Today's Lecture

- **What is RDS?**
  - Introducing FACT
- Tradeoffs between different challenges
- Discrimination-aware data mining
  - Modeling Discrimination
- Discrimination-aware decision trees
  - Relabeling Method

# What is RDS?

Introducing FACT



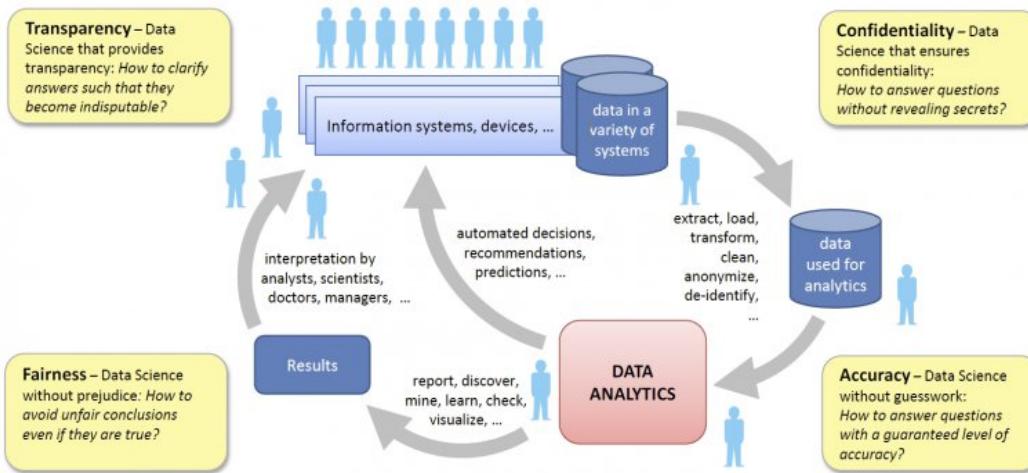
# What is RDS?

- Responsible Data Science (RDS) advocates taking the third part of the data science pipeline (**effect**) as leading when designing or using the first two parts (**infrastructure and analysis**).



# What is RDS?

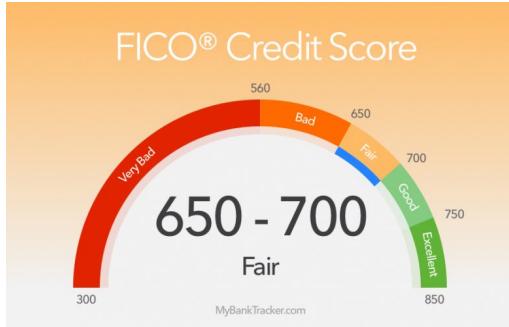
- Responsible Data Science (RDS) advocates the development of techniques, algorithms, tools, laws, ethical/social principles for ensuring **Fairness**, **Accuracy**, **Confidentiality**, and **Transparency** covering the whole Data Science pipeline.



# Fairness: Data Science without prejudice: How to avoid unfair conclusions even if they are true?



# Data-driven decisions are everywhere!



Banking



Insurance



Hiring



Admission



Health



Chair of Process  
and Data Science

## How white engineers built racist code - and why it's dangerous for black people

As facial recognition tools play a bigger role in fighting crime, inbuilt racial biases raise troubling questions about the systems that create them



▲ A protest over police violence against black communities. Photograph: Alamy Stock Photo

**“Y**ou good?” a man asked two narcotics detectives late in the summer of 2015.

The detectives had just finished an undercover drug deal in Brentwood, a predominately black neighborhood in Jacksonville, Florida, that is among the poorest in the country, when the man unexpectedly

**“If you’re black, you’re more likely to be subjected to this technology and the technology is more likely to be wrong,”**

This report found that black individuals, as with so many aspects of the justice system, were the most likely to be scrutinized by facial recognition software in cases. It also suggested that software was most likely to be incorrect when used on black individuals – a finding corroborated by the FBI’s own research.

Experts think that facial recognition software has problems recognizing black faces because its algorithms are usually written by white engineers who dominate the technology sector. These engineers build on pre-existing code libraries, typically written by other white engineers.

## Rise of the racist robots - how AI is learning all our worst impulses

There is a saying in computer science: garbage in, garbage out. When we feed machines data that reflects our prejudices, they mimic them – from antisemitic chatbots to racially biased software. Does a horrifying future await people forced to live at the mercy of algorithms?



▲ Current laws 'largely fail to address discrimination' when it comes to big data. Photograph: artpartner-images/Getty Images

In May last year, a stunning report claimed that a computer program used by a US court for risk assessment was biased against black prisoners. The program, Correctional Offender Management Profiling for Alternative Sanctions (Compas), was much more prone to mistakenly label black defendants as likely to reoffend – wrongly flagging them at almost twice the rate as white people (45% to 24%), according to the investigative journalism organisation ProPublica.

**There is a saying in computer science: garbage in, garbage out. When we feed machines data that reflects our prejudices, they mimic them – from antisemitic chatbots to racially biased software. Does a horrifying future await people forced to live at the mercy of algorithms?**

In May last year, a stunning report claimed that a computer program used by a US court for risk assessment was biased against black prisoners. The program, Correctional Offender Management Profiling for Alternative Sanctions (Compas), was much more prone to mistakenly label black defendants as likely to reoffend – wrongly flagging them at almost twice the rate as white people (45% to 24%), according to the investigative journalism organization ProPublica.

## Rise of the racist robots - how AI is learning all our worst impulses

There is a saying in computer science: garbage in, garbage out. When we feed machines data that reflects our prejudices, they mimic them - from antisemitic chatbots to racially biased software. Does a horrifying future await people forced to live at the mercy of algorithms?



▲ Current laws 'largely fail to address discrimination' when it comes to big data. Photograph: artpartner-images/Getty Images

In May last year, a stunning report claimed that a computer program used by a US court for risk assessment was biased against black prisoners. The program, Correctional Offender Management Profiling for Alternative Sanctions (Compas), was much more prone to mistakenly label black defendants as likely to reoffend - wrongly flagging them at almost twice the rate as white people (45% to 24%), according to the investigative journalism organisation ProPublica.

**How do we prevent these programs from amplifying the inequalities of our past and affecting the most vulnerable members of our society. When the data we feed the machines reflects the history of our own unequal society, we are, in effect, asking the program to learn our own biases.**

**Tech giants Google and Microsoft have recently taken steps to investigate: that as our computational tools have become more advanced, they have become more opaque. The data they rely on – arrest records, postcodes, social affiliations, income – can reflect, and further ingrain, human prejudice.**

Home > Internet

OPINION BY PRESTON GRALLA

## Amazon Prime and the racist algorithms

The company's algorithms told it where to offer its Prime Free Same-Day Delivery service, but an algorithm that uses data tainted by racism will be racist in its outcomes



By Preston Gralla

Contributing Editor, Computerworld | MAY 11, 2016 5:17 AM PT



**The revelation in late April that Amazon has excluded minority neighborhoods in Boston, Atlanta, Chicago, Dallas, New York City, and Washington, D.C., from its Prime Free Same-Day Delivery service while extending the service to white neighborhoods probably shocked few minorities.**

**But to Amazon, and likely others in the tech world, the decision had nothing to do with racism and everything to do with the facts. Amazon argued that it wasn't acting on prejudice when it excluded those neighborhoods. Instead, algorithms and the underlying data on which those algorithms were based made it clear that Amazon couldn't make a profit in them. And, given that Amazon is profit-driven, the company excluded them. Race, Amazon said, had nothing to do with it.**

Home > Internet

OPINION BY PRESTON GRALLA

## Amazon Prime and the racist algorithms

The company's algorithms told it where to offer its Prime Free Same-Day Delivery service, but an algorithm that uses data tainted by racism will be racist in its outcomes



By Preston Gralla

Contributing Editor, Computerworld | MAY 11, 2016 5:17 AM PT

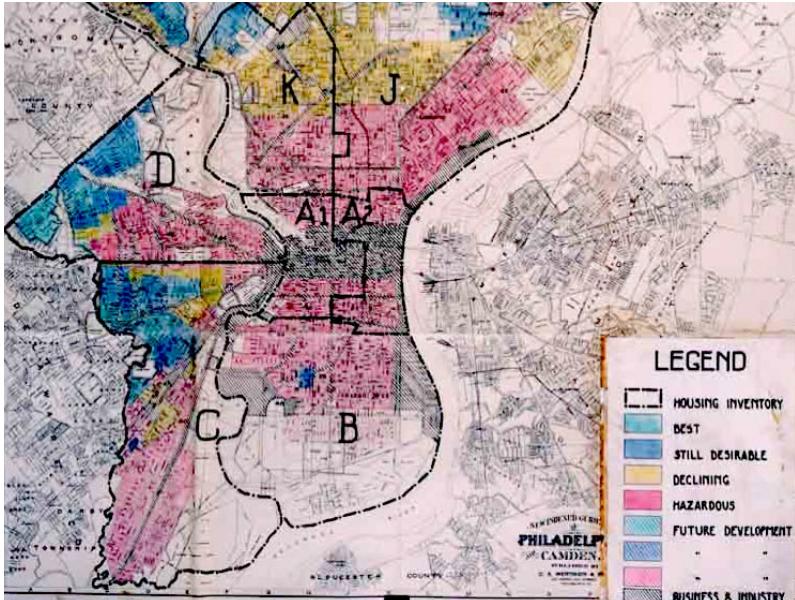


**The tech industry likes to portray itself as data-driven and progressive, leading us to a better future by a reliance on a clear-eyed analysis of facts, while cutting through blind spots and prejudice. But as the delivery issue shows, the way the industry chooses its data and the algorithms it creates can mean supporting an unfair status quo.**



Chair of Process  
and Data Science

# Not new: Redlining



1936 security map of Philadelphia showing redlining of lower income neighborhoods

- Redlining, a discriminatory practice of denying affordable services based on geographical locations.
- You can remove race, gender, age, etc., but if this correlates with your zip code, ...

# The New York Times

HIDDEN BIAS

## *When Algorithms Discriminate*

By Claire Cain Miller

July 9, 2015



The online world is shaped by forces beyond our control, determining the stories we read on Facebook, the people we meet on OkCupid and the search results we see on Google. Big data is used to make decisions about health care, employment, housing, education and policing.

But can computer programs be discriminatory?

There is a widespread belief that software and algorithms that rely on data [are objective](#). But software is not free of human influence. Algorithms are written and maintained by people, and machine learning algorithms adjust what they do based on people's behavior. As a result, say researchers in computer science, ethics and law, algorithms can [reinforce human prejudices](#).

<https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html>

**Machine learning algorithms learn and evolve based on what people do online. The autocomplete feature on Google and Bing is an example. A recent Google search for “Are transgender,” for instance, suggested, “Are transgenders going to hell.”**

**It would be impossible for humans to oversee every decision an algorithm makes. But companies can regularly run simulations to test the results of their algorithms. Algorithms should “be designed from scratch to be aware of values and not discriminate.”**



Are computer scientists|



- are computer scientists **software engineers**
- are computer scientists **engineers**
- are computer scientists **rich**
- are computer scientists **scientists**
- are computer scientists **considered scientists**
- are computer scientists **smarter than engineers**
- are computer scientists **in demand**
- are computer scientists **actually scientists**
- are computer scientists **happy**
- are computer scientists **mathematicians**

Unangemessene Vervollständigungen melden  
[Weitere Informationen](#)



Are females|



- are females **more agreeable than males**
- are females **born with eggs**
- are females **allowed to play in the nfl**
- are females **born with all of their primary oocytes**
- are females **allowed to be navy seals**
- are females **better at multitasking than males**
- are females **allowed in the marines**
- are females **allowed in the nba**
- are females **allowed in nda**
- are females **more health conscious than males**

Unangemessene Vervollständigungen melden  
[Weitere Informationen](#)



Chair of Process  
and Data Science

# Example

**Table 2** No illegal discrimination (Example 1).

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

Example taken from Kamiran, F., Žliobaitė, I. & Calders, T. Knowl Inf Syst (2013) 35: 613. <https://doi.org/10.1007/s10115-012-0584-8>

# Example

**Table 3** Illegal discrimination is present (**Example 2**).

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

# Example

What if ...

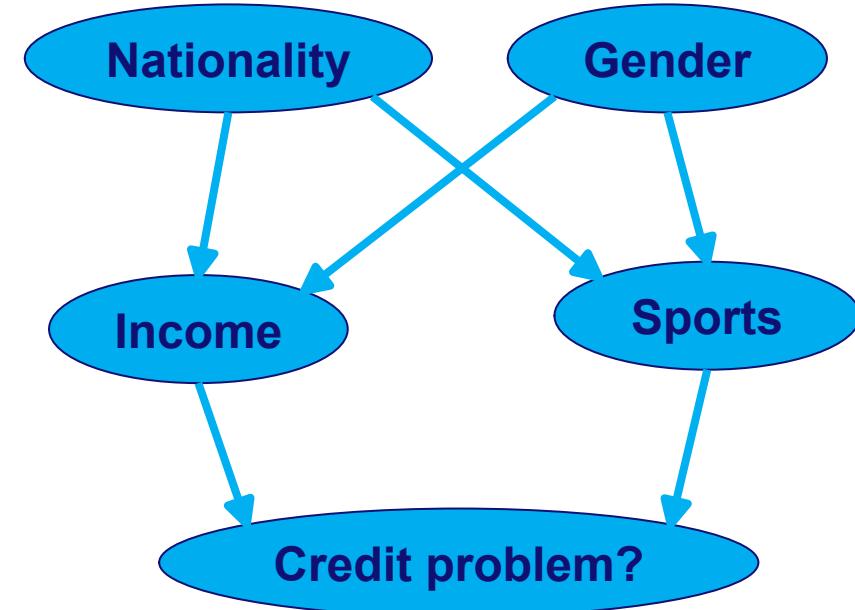
	Non-German $\geq 50$		German $<50$	
	<i>Rothe Erde</i>		<i>Südviertel</i>	
	female	male	female	male
number of applicants	800	200	200	800
acceptance rate	20%	20%	40%	40%
accepted (+)	160	40	80	320

# Fairness: How to avoid unfair conclusions even if they are true?

- There may be **intentional** or **unintentional** discrimination when making data-driven decisions.
- Training data may be **biased** (wrong or outdated) or sample may **not be representative** (never enough evidence for some groups).
- Even when the data are **correct**, optimizing for a particular target variable may lead to discrimination (**correct does not imply fair!**).

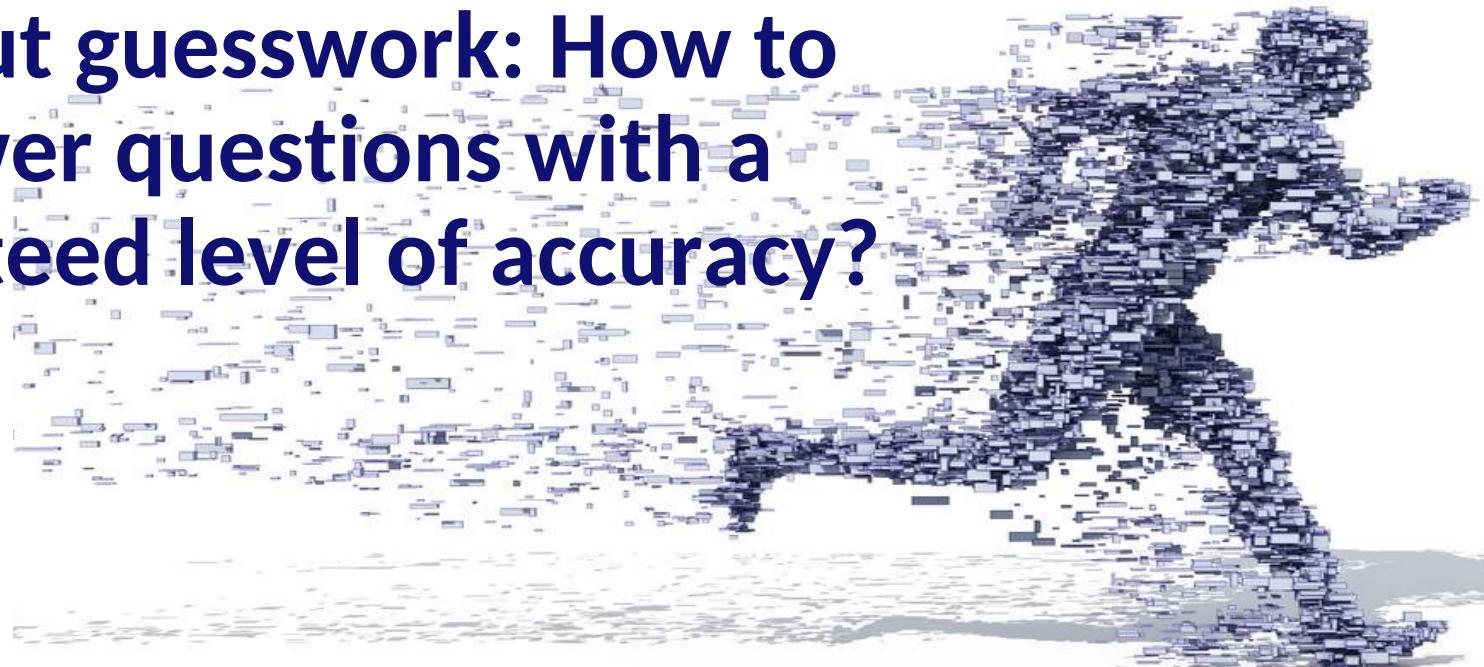
# Fairness: How to avoid unfair conclusions even if they are true?

- **Removing sensitive attributes often does not work!**
- **Enforcing fairness will lead to more incorrect predictions.**
- **Not so easy to define fairness.**
  - Survival ratio: academic vs regional hospitals, experience vs unexperienced doctors, etc.
  - Delays: busy vs idle resources, part-time vs full-time, etc.



Remember Simson's paradox

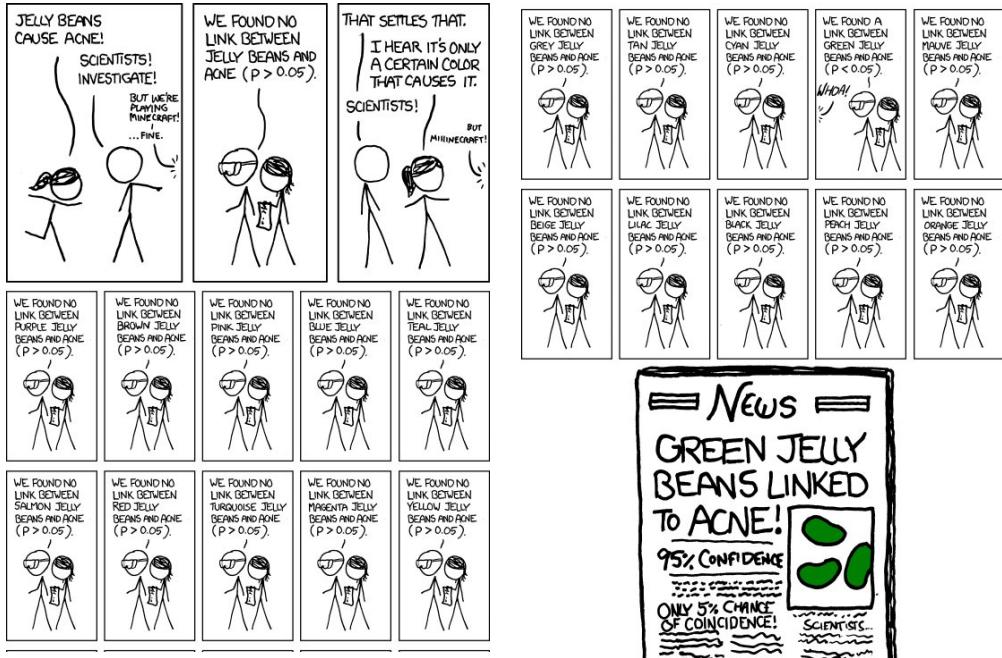
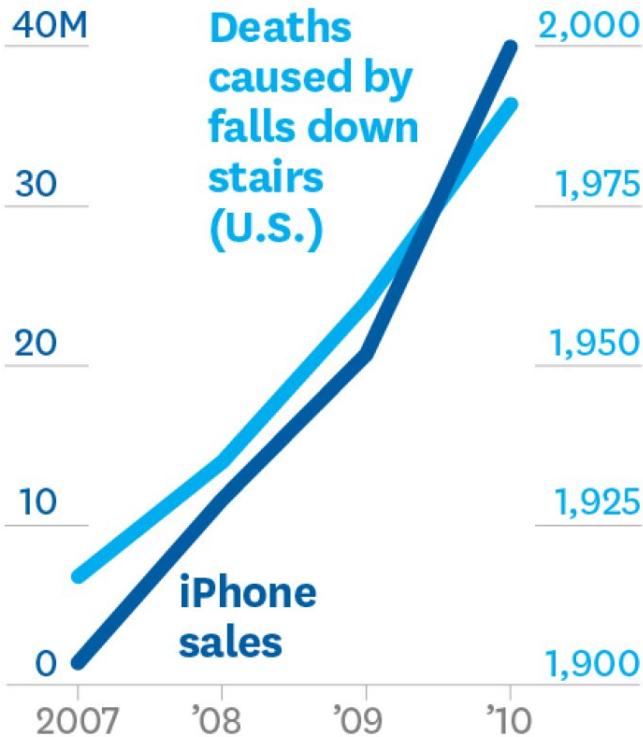
# **Accuracy: Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?**



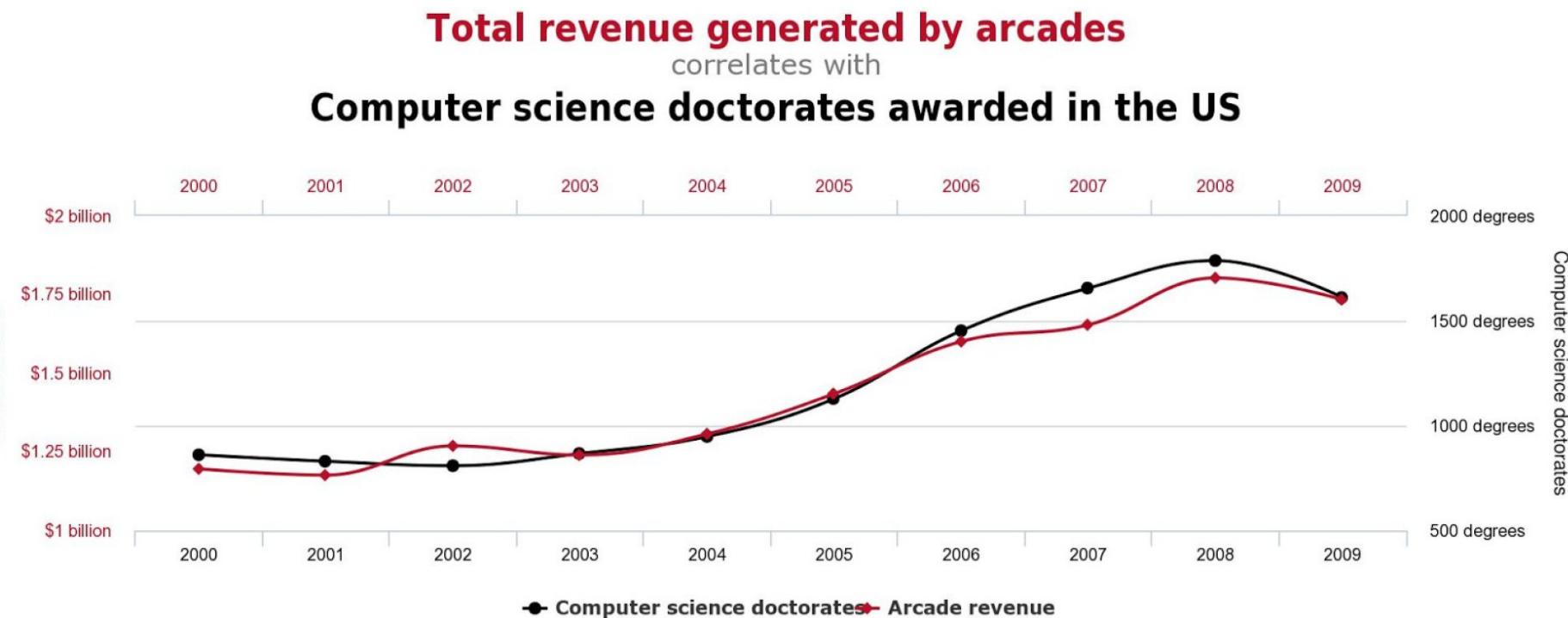
# Accuracy

- Data science approaches should not just present results or make predictions, but also explicitly provide meta-information on the accuracy of the output.
- Analysis results need to be accurate and should not deceive end-users and decision makers. However, there are several factors endangering accuracy.
  - Spurious correlations
  - Curse of dimensionality
- Data science without guesswork – How to answer questions with a guaranteed level of accuracy?

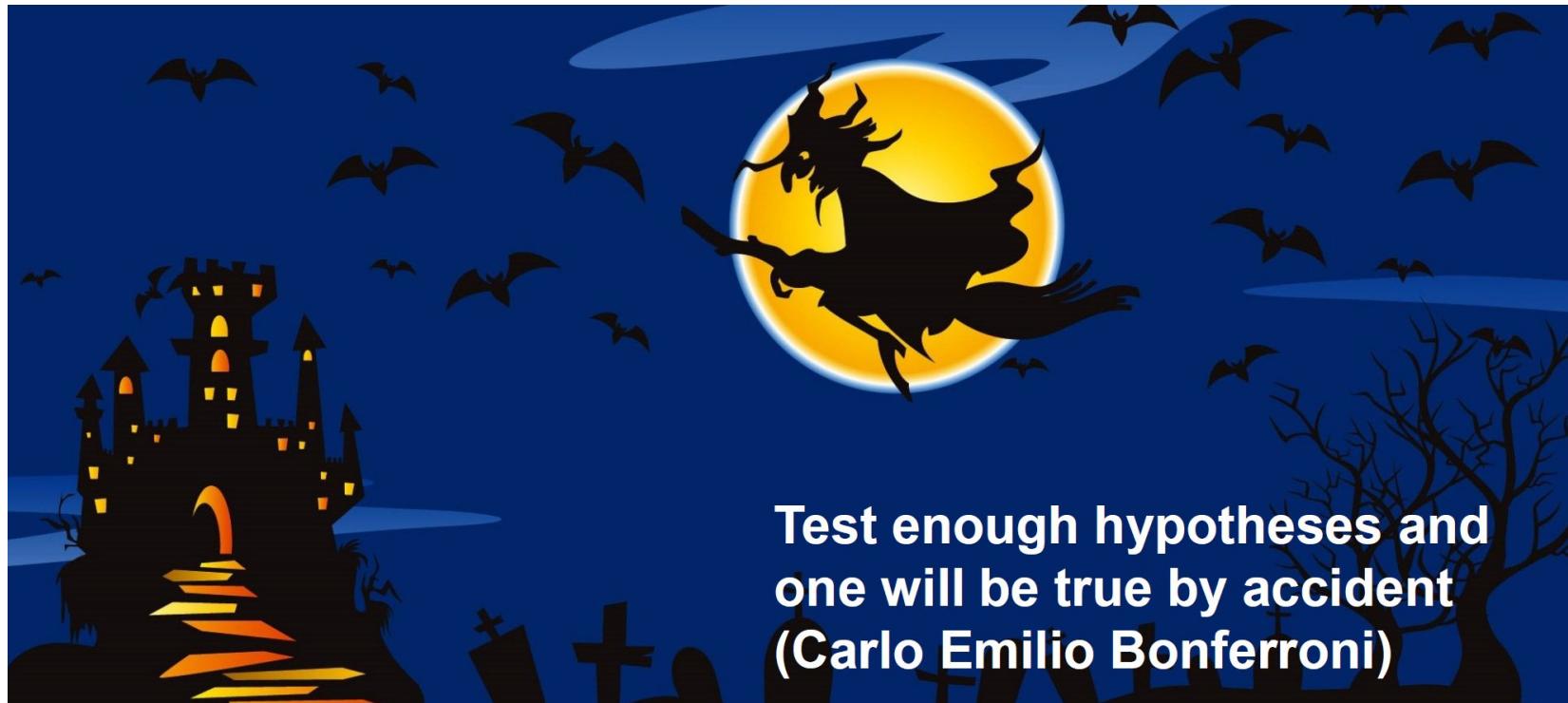
# Accuracy spurious correlations



# Accuracy spurious correlations



# Accuracy Curse of dimensionality



**Test enough hypotheses and  
one will be true by accident  
(Carlo Emilio Bonferroni)**

# Accuracy Curse of dimensionality

Find the terrorists



## Assumptions:

- 18 million people in NL
- 1800 hotels
- 100 guests per hotel per night
- Visit hotel every 100 days

# Accuracy Curse of dimensionality



Suspicious event: two persons stay in the same hotel on two different dates

How many suspicious events in a 1000 day period?

# Accuracy Curse of dimensionality



The probability that two persons visit a hotel on a given day d:  $\frac{1}{100} \times \frac{1}{100} = 10^{-4}$

The probability that two persons visit the same hotel on day d:  $10^{-4} \times \frac{1}{1800} = 5.55 \times 10^{-8}$

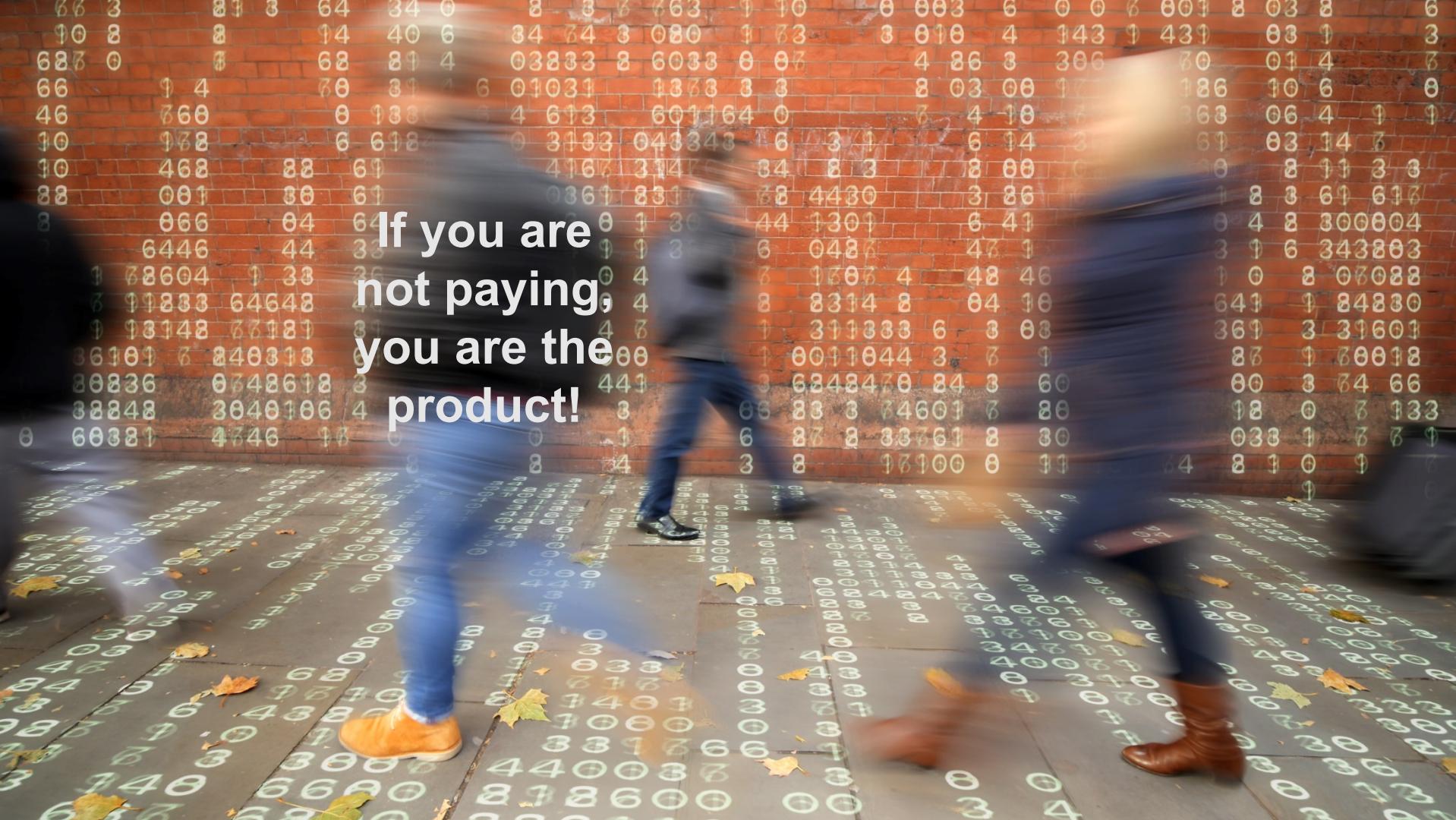
The probability that two persons visit the same hotel on two different days:  $(5.55 \times 10^{-8})^2$



# **Confidentiality:** Data Science that ensures confidentiality: How to answer questions without revealing secrets?



If you are  
not paying,  
you are the  
product!



G0d @\_Orbit

Das 24. Türchen: CDU/CSU

Translate Tweet



G0d @\_Orbit · 23



Das 23. Türchen:

Translate Tweet



G0d @\_Orbit · 23

Das 24. Türchen: CDU/CSU



Today's news



WELCHE MÖGLICHKEITEN HABEN  
WIR, DIE DATENSICHERHEIT  
WIRKSAM ZU BESCHLEUNIGEN ?

DEN  
NETZAUSBAU  
BREMSEN !



IT-Experten.

KARIKATUR: TOMIECK  
Yesterday's newspaper

# Confidentiality Facebook data leakage scandal

# Facebook data leakage scandal



[JESSIE LAPOWSKY](#) BUSINESS 12.20.18 07:00 AM

# THE 21 (AND COUNTING) BIGGEST FACEBOOK SCANDALS OF 2018

## SHARE

 SHARE  
647

TWEET

COMMENT

[EMAIL](#)

A photograph of the front page of The Times newspaper from March 19, 2010. The page features a large, prominent image of Wayne Rooney in a Manchester United kit, looking towards the camera. Above the image, the word 'TIMES' is written in large, bold letters. Below the image, the headline 'Best for football' is displayed in a large, bold font. A sub-headline reads 'An squeeze past Leicester to set up FA Cup semi-final against Southampton'. At the bottom of the page, there is a box containing the text 'Beware the joke sexist joke' and 'By Michael Moran'.

## Facebook told to come clean over leaked data of 50m users

used information from the work. A whistleblower said Cambridge Analytica, which goes "psychographic profiling" after political campaigns, obtained data of millions of Facebook users from a British firm which it was buying "personality tests". Mr Collins also said that he had raised MPs by failing to note that Cambridge Analytica

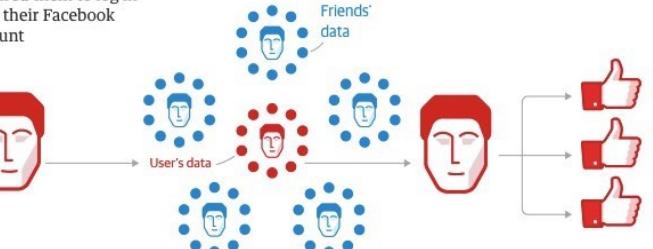
## **Cambridge Analytica: how 50m Facebook records were hijacked**

**2** The success info talk

pp also collected data as likes and personal information from the test's Facebook account ...

The personality quiz results were paired with their Facebook data - such as likes - to seek out psychological patterns

**4** Algorithms combined the data with other sources such as voter records to **create a superior set of records** (initially 2m people in **11 key states\***), with hundreds of data points per person



as well their friends' data, amounting to over 50m people's Facebook data



These individuals could then be targeted with highly personalised advertising based on their personality data



LINE NEWS STANDARD.CO.UK

**WE HAVE OUR STRUGGLES**

**JAY-Z PAYNE ON HIS RELATIONSHIP WITH CHERYL » EXCLUSIVE INTERVIEW PAGE 3**

5: PAYBACK TIME FOR LONDON'S YOUNG BUYERS » HOMES & PROPERTY 1

Digitized by srujanika@gmail.com

**FACEBOOK  
DATA GRAB  
IS JUST TIP  
OF ICEBERG'**



## **WELCOME HOME, MUM! LONDON'S \$1M TEACHER**

ardian graphic. \*Arkansas, Colorado, Florida, Iowa, Louisiana, Nevada, New Hampshire, North Carolina, Oregon, South Carolina, West Virginia

# Confidentiality

The last month (December 2018) News

## Home Addresses Are Up for Sale. Time to Take Back Your Privacy.

Home addresses have always been public information. But now they're too easy to search.



The New York Times

Dec. 16, 2018

## Facebook Says Bug Opened Access to Private Photos



The New York Times

Dec. 14, 2018

MARKETS

## U.K.'s Tesco Bank Fined \$21.4 Million Over Cyberbreach

The Financial Conduct Authority said the cyberattack was 'largely avoidable' had the bank been more diligent

By Mara Lemos Stein

Oct. 1, 2018 12:52 p.m. ET

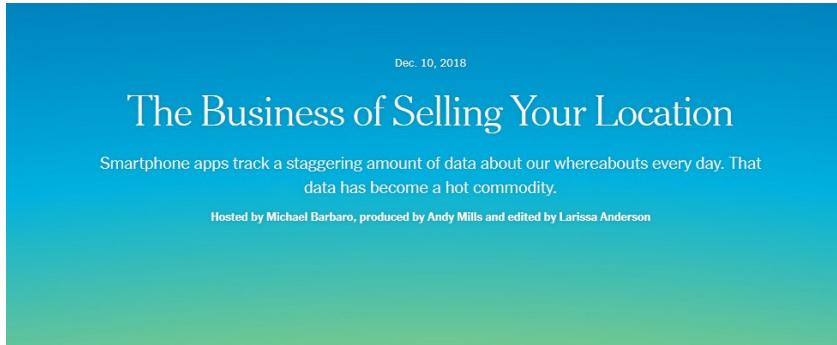
The U.K.'s Financial Conduct Authority issued a £16.4 million (\$21.4 million) penalty to Tesco Bank for failing to protect clients from a cyberattack in November 2016.

This is the first time the FCA has taken enforcement action related to a cyberattack, revealing the regulator's willingness to address lapses of risk management by financial institutions.

Banks...

# Confidentiality

## The last month (December 2018) News



The New York Times

## Sundar Pichai, Google's C.E.O., Testifies on Capitol Hill

By DAISUKE WAKABAYASHI and CECILIA KANG DEC. 11, 2018



Chair of Process  
and Data Science

# Confidentiality

The last month (December 2018) News

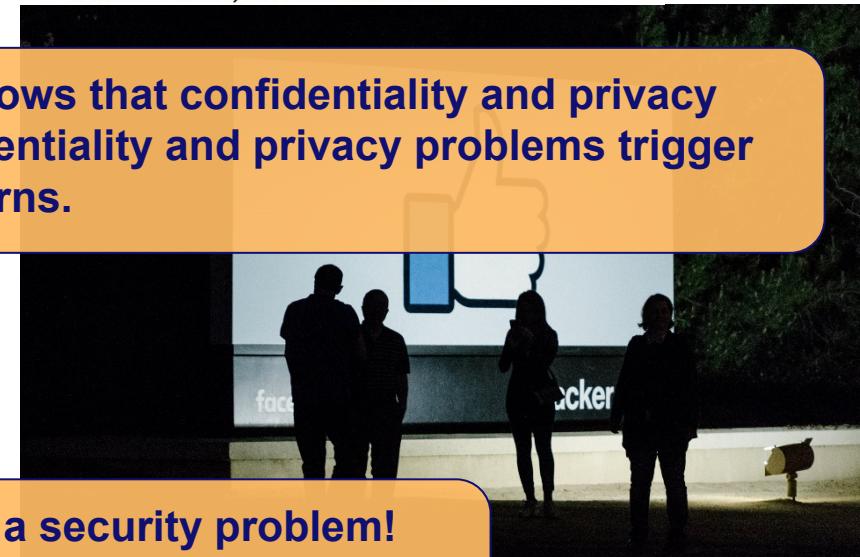
*Oath Agrees to \$5 Million Settlement Over Children's Privacy Online*

*Facebook Used People's Data to Favor Certain Partners and Punish Rivals, Documents Show*

This small selection of news items shows that confidentiality and privacy have become a great concern. Confidentiality and privacy problems trigger social, ethical, and economical concerns.



The penalty that the Verizon-owned Oath agreed to pay is the largest a company has to Coppa. Kristoffer Tripplaar/Sipa, via Associated Press



**Not (just) a security problem!  
Not limited to tech giants!  
New competitive angle.**

Dec. 5, 2018



Chair of Process  
and Data Science

# General Data Protection Regulation (GDPR)

- The General Data Protection Regulation (GDPR) applies to member states of the European Union.
- It came into effect on 25 May, 2018.
- Companies that are found guilty of misusing data can be fined up to €20 million or 4% of the company's annual turnover.
- GDPR states that controllers must make sure it's the case that personal data is processed lawfully, transparently, and for a specific purpose.
- This implies that people must understand why their data is being processed and how it is being processed.



## Communication

Explain in simple language why the user should leave personal information. How the information will be used and how long it will be stored.



## Consent

Get clear consent to the processing of personal data.



## Access

Users should have access to their information and the ability to pass it on to other companies.

## Communication

Explain in simple language why the user should leave personal information. How the information will be used and how long it will be stored.



## Consent

Get clear consent to the processing of personal data.



## Access

Users should have access to their information and the ability to pass it on to other companies.



## Erase data

Any company that processes data, need to remove someone's personal information on request if it is not contrary to the public interest or other fundamental rights of Europeans.



## Sensitive data

Ensure the specific safety for information on health, race, sexual orientation, religion and political view.



## Marketing

People should be able to give up direct marketing that uses their data.



## Warnings

Companies are required to notify regulatory authorities (and in some cases data subjects) of any breach of personal data within 72 hours of the discovery of such breach.

## Profiling

Individuals have the right to appeal against the decision when it is based on automated processing and produces a legal effect or similarly significant effect on the individual.

## Data transfer outside the EU

Personal data can only be transferred to countries outside the EU and the EEA when an adequate level of protection is guaranteed.

# Confidentiality

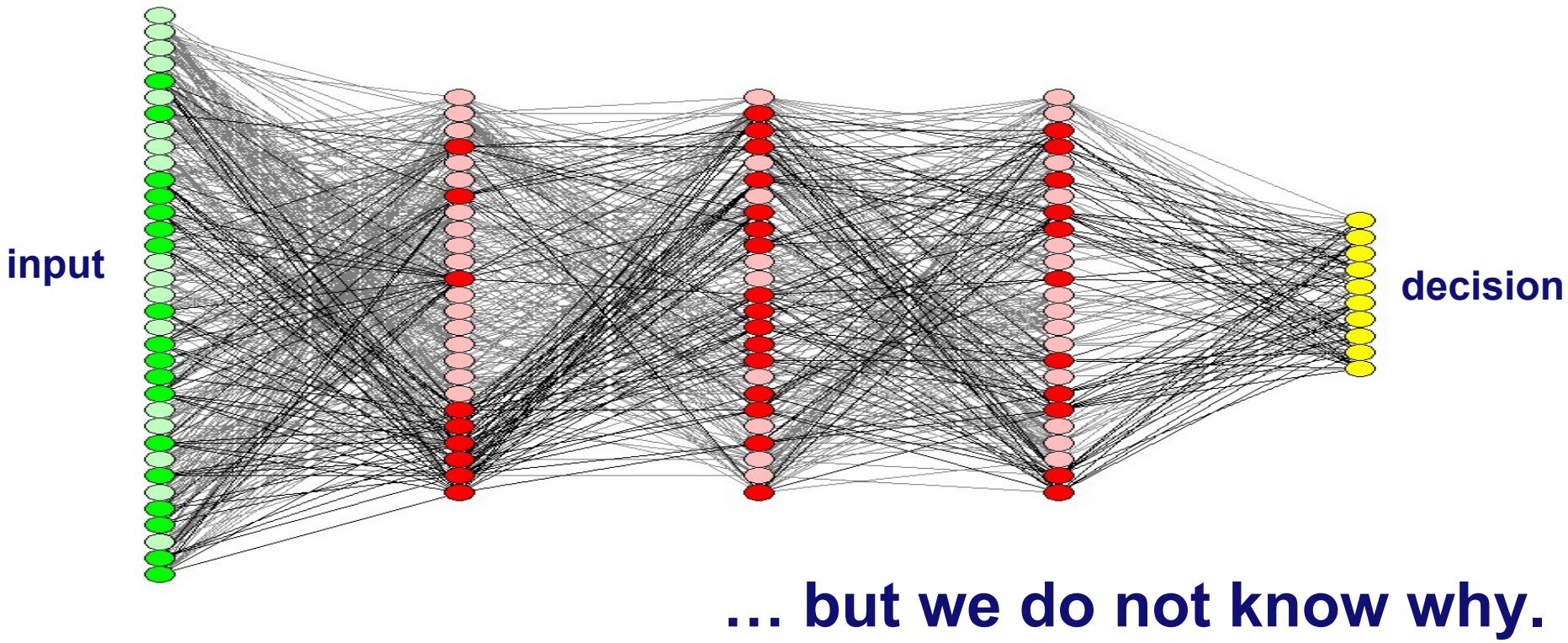
- The goal should **not** be to prevent data from being distributed and gathered, but to use data in a **safe** and **controlled** manner.
- Data science does not aim to find a **guilty person**, but aims to find **inefficiencies, bottlenecks, lack of service, etc.** and provide a solution for **improving processes** and products without revealing sensitive data.
- Data/process science techniques should not reveal the owner's **confidential** information.
- Confidentiality: data science that ensures confidentiality - how to answer questions without revealing secrets?



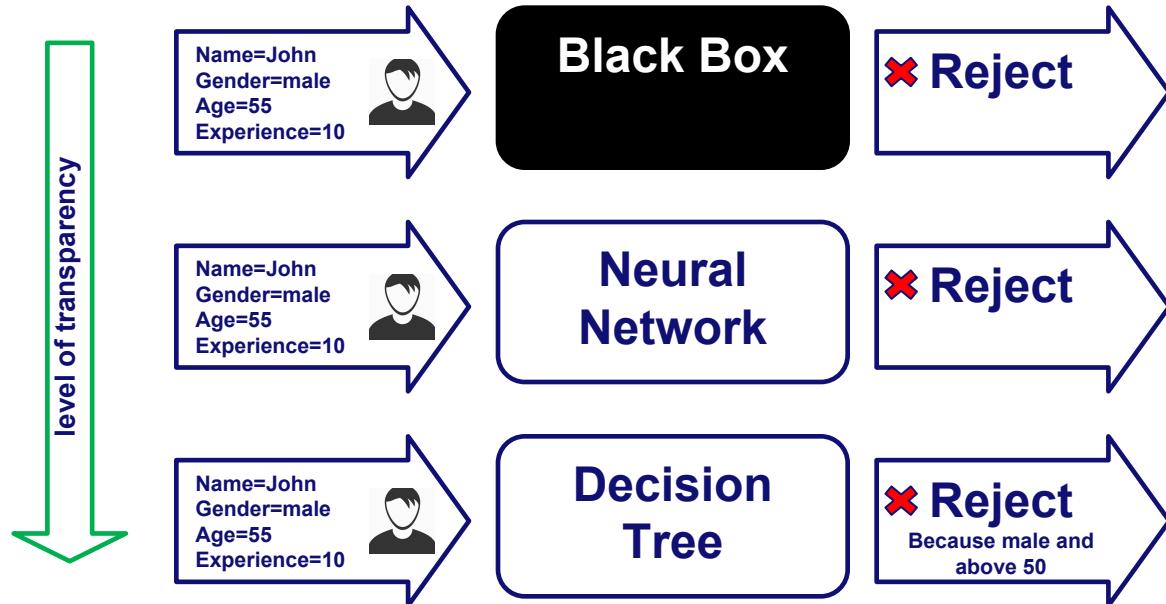
# **Transparency:** Data Science that provides transparency: How to clarify answers such that they become indisputable?

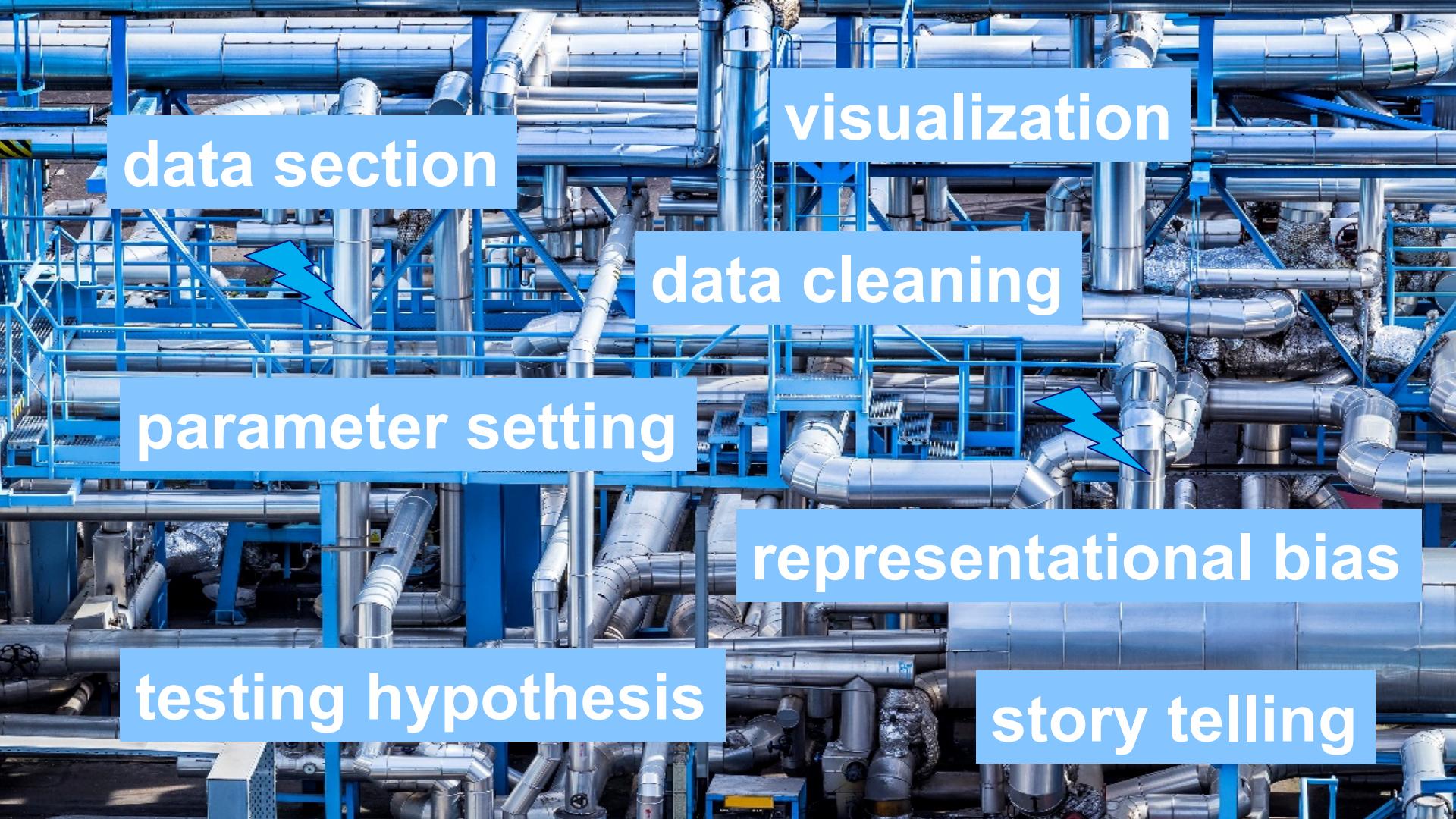


# You are guilty, not selected, not treated, ...



# Transparency





**data section**

**visualization**

**data cleaning**

**parameter setting**

**representational bias**

**testing hypothesis**

**story telling**

A photograph of four young adults (two men and two women) in a white studio setting. They are all holding their hands up to their faces, with their mouths near the other person's ear, as if whispering. The man on the far left is wearing a red t-shirt and looking upwards. The man next to him is wearing a light pink button-down shirt and also looking upwards. The woman in the center is wearing a purple top and looking slightly to the right. The man on the far right is wearing a blue t-shirt and looking towards the center. The background is plain white.

**data selection**

**cleaning**

**set parameters**

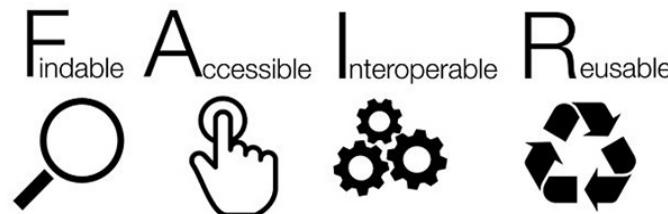
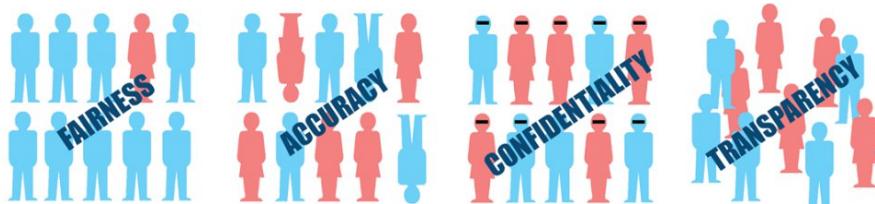
**interpretation**

# Transparency

- Data science can only be effective if people **trust** the results and are able to correctly **interpret** the outcomes.
- Data science should not be viewed as a **black box** that magically transforms data into value.
  - The journey from raw data to meaningful inferences involves multiple steps and actors, thus accountability and comprehensibility are essential for transparency.
- Transparency: data science that provides transparency - how to clarify answers so that they become indisputable?

# Summary FACT

- Responsible data science centers around four challenging questions which called **FACT**
  - Q1 fairness: data science without prejudice - how to avoid unfair conclusions even if they are true?
  - Q2 accuracy: data science without guesswork - how to answer questions with a guaranteed level of accuracy?
  - Q3 confidentiality: data science that ensures confidentiality - how to answer questions without revealing secrets?
  - Q4 transparency: data science that provides transparency - how to clarify answers so that they become indisputable?
- **FACT** is related, but should not be confused with the well-known **FAIR** principles (**Findable, Accessible, Interoperable, and Re-usable**).



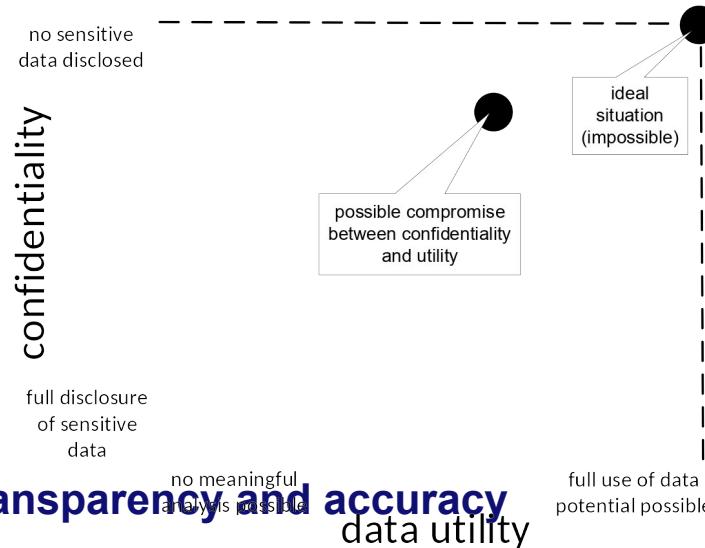
Chair of Process  
and Data Science

# Tradeoffs Between Different Challenges



# Between Confidentiality and Data Utility

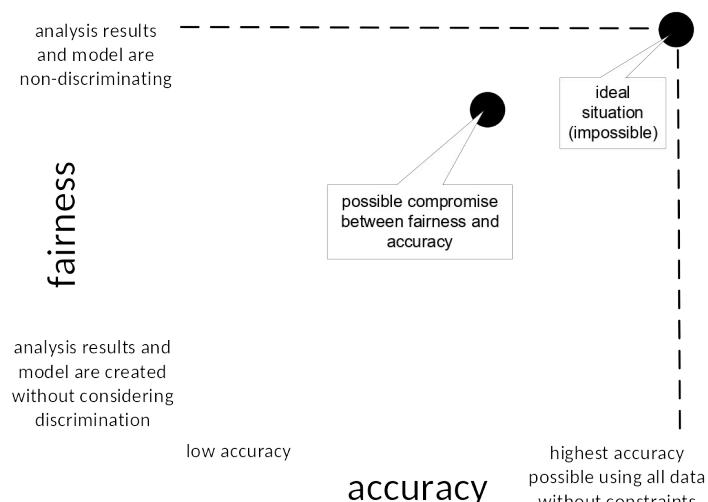
- Removing variables, aggregation, and adding noise can make it hard to produce meaningful results.



- Confidentiality may also affect transparency and accuracy

# Between Fairness and Accuracy

- One of the main approaches to deal with unfairness data is adding constraints.
  - For example, the constraint that “males and females should have the same chance to be selected in an interview”.
- Adding these constraints will decrease the accuracy



# Discrimination-Aware Data Mining



# Discrimination-aware Data Mining

- First paper by Dino Pedreschi, Salvatore Ruggieri, Franco Turini. Discrimination-Aware Data Mining. In KDD 2008. (focus on association rules).
- Followed by work from Faisal Kamiran, Indre Zliobaite, Toon Calders (mostly on decision trees/classification).
- Dedicated workshops on the topic (e.g., FAT/ML).
- Addresses the **Fairness** aspect in FACT.

## Discrimination-aware Data Mining

Dino Pedreschi Salvatore Ruggieri Franco Turini

Dipartimento di Informatica, Università di Pisa  
L.go B. Pontecorvo 3, 56127 Pisa, Italy  
{pedre,ruggieri,turini}@di.unipi.it

### ABSTRACT

In the context of civil rights law, discrimination refers to unfair treatment of individuals based on membership in a category or a minority without regard to individual merit. Rules extracted from databases by data mining techniques, such as classification or association rules, when used for decision tasks such as bundle or credit approval, can be discriminatory in the above sense. In this paper, the notion of discriminatory classification rules is introduced and a methodology for their detection is proposed. It is shown to be a non-trivial task. A naive approach, like taking away all discriminatory attributes, is shown to be not enough when other background knowledge is available. Our approach leads to a precise formulation of the redlining problem along with a formal result relating discriminatory rules with apparently safe ones by means of background knowledge. An empirical assessment of the results on the German credit dataset is also provided.

### Categories and Subject Descriptors

H.2.8 [Database Applications]: Data Mining

### General Terms

Algorithms, Economics, Legal Aspects

### Keywords

Discrimination, Classification Rules

### 1. INTRODUCTION

The word *discrimination* originates from the Latin *discriminatio*, which means “separating, however”. In social sciences, however, discrimination refers specifically to an action based on prejudices resulting in unfair treatment of people on the basis of their membership in a category, without regard to individual merit. As an example, U.S. federal laws [17] prohibit discrimination on the bases of race, color, religion, nationality, sex, marital status, age and pregnancy.

Function to make digital or hard copies of all or part of this work for personal or classroom use is granted below for persons that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.  
Copyright 2008 ACM 978-1-60558-93-4/08/08 ...\$15.00.

In a variety of settings, including: credit-score scoring (Equal Credit Opportunity Act), job, rental, and financing of housing (Fair Housing Act), personnel selection and wage (Intentional Employment Act, Equal Pay Act, Pregnancy Discrimination Act).

Concerning the research side, the issue of discrimination in credit, mortgage, insurance, labor market, education and other social activities has started to interest many researchers in economics and human sciences since '90s, when a theory on the economics of discrimination was proposed [3]. The literature has given evidence of unfair treatment in social profiling and redlining [14], mortgage discrimination [9], personnel selection discrimination [6, 7], wage discrimination [8].

In data mining and machine learning, classification models are constructed on the basis of historical data exactly with the purpose of discrimination in the original Latin sense: i.e. distinguishing between elements of different classes, in order to reveal the reasons of class membership, or to predict it for new elements. However, the same classification models can be adopted as a support to decision making, clearly also in socially sensitive tasks such as the access of applicants to benefits, to public services, to credit. Now the question that naturally arises is the following. While classification models are designed to distinguish between general or less arbitrary decisions, can they be discriminatory in the so-called “negative sense”? The answer is clearly yes: it is evident that relying on mixed models for decision making does not put ourselves on the safe side. Rather dangerously, learning from historical data may mean to discover traditional prejudices that are difficult to identify. It is thus important to provide the user of general rules, namely, the classifier, with those rules can be deeply hidden within a classifier.

Surprisingly, despite the risk of discrimination poses clear ethical and legal obstacles to the practical application of data mining in socially sensitive decision making, to the best of our knowledge, there is no previous work on the topic. In this paper, we tackle the problem of discrimination in data mining in a rule-based setting, by introducing the notion of discriminatory classification rules, as a criterion to identify the potential risks of discrimination.

### 2. CONTROLLING DISCRIMINATION

The first natural approach to formally tackle the problem is to specify a set of selected attribute values (or, at an extreme, an attribute as a whole) as potentially discriminatory; examples include female gender, ethnic minority, low-level job, specific age range. However, this simple ap-

# What is Discrimination-Aware Data Mining?

- In the context of civil rights law, discrimination refers to **unfair** or **unequal** treatment of people based on membership to a category or a minority, without regard to individual merit.
- But, ...  
**we want to discriminate between:**
  - people playing their mortgages and those that do not,
  - students that will graduate and students that will drop out, and
  - patients that are likely to benefit from treatment and those that do not.

Definition of *discrimination* in English:

**discrimination** 



**NOUN**

[mass noun]

1 The unjust or prejudicial treatment of different categories of people, especially on the grounds of race, age, or sex.  
‘victims of racial discrimination’  
‘discrimination against homosexuals’

[+ More example sentences](#) [+ Synonyms](#)

2 Recognition and understanding of the difference between one thing and another.  
‘discrimination between right and wrong’

[count noun] ‘young children have difficulties in making fine discriminations’

[+ More example sentences](#) [+ Synonyms](#)

2.1 The ability to judge what is of high quality; good judgement or taste.  
‘those who could afford to buy showed little taste or discrimination’

[+ More example sentences](#) [+ Synonyms](#)

2.2 *Psychology* The ability to distinguish between different stimuli.  
[as modifier] ‘discrimination learning’

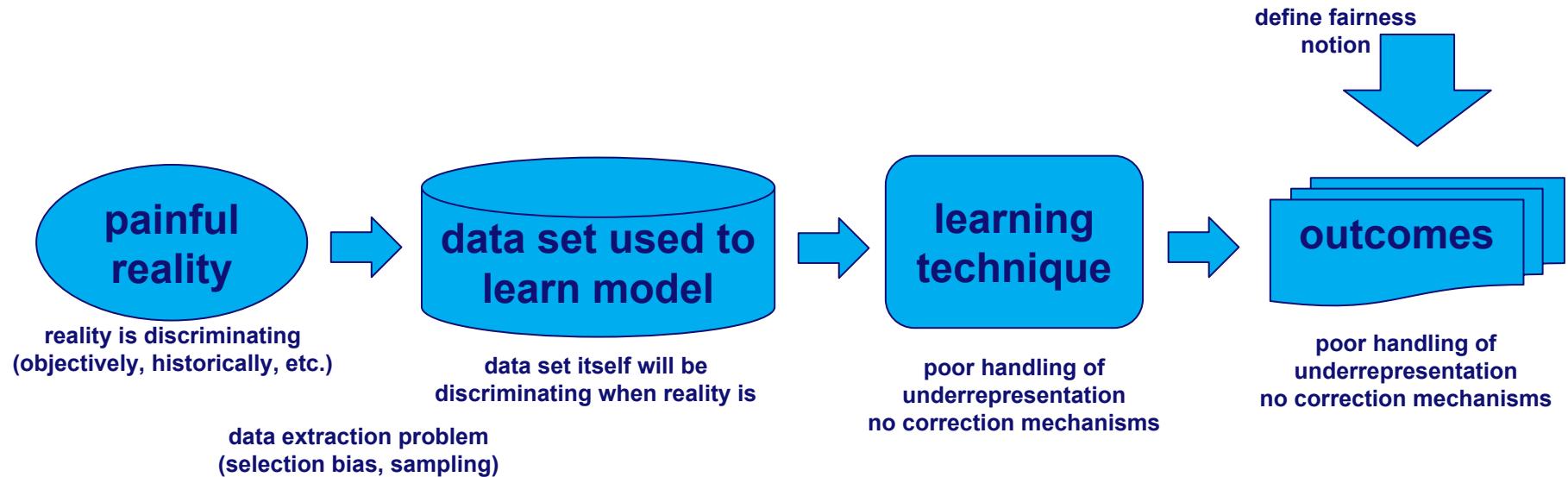
[+ More example sentences](#)

3 *Electronics*

The selection of a signal having a required characteristic, such as frequency or amplitude, by means of a discriminator.

# What is Discrimination-Aware Data Mining?

- Note so easy to measure or define.



# Confidentiality and Discrimination Prevention

- There is a strong relation between **confidentiality** and **fairness/discrimination**.
- Confidentiality may prevent discrimination (e.g., applications do not show age / gender), but this is not guaranteed.

# Two questions

- How to measure discrimination?
- How to prevent discrimination?

# Measuring Discrimination

Based on

**Discrimination-aware Data Mining**

Dino Pedreschi Salvatore Ruggieri Franco Turini

Dipartimento di Informatica, Università di Pisa  
L.go B. Pontecorvo 3, 56127 Pisa, Italy  
[{pedre,ruggieri,turini}@di.unipi.it](mailto:{pedre,ruggieri,turini}@di.unipi.it)



# Some Notations (Association and Classification Rules)

- $a_1, a_2, \dots, a_n$  are **attributes** (i.e., features).
- $a = v$  is an  **$a$ -item** which assigns value  $v$  to  $a$ .
- $I = \{a_1 = v_{11}, a_1 = v_{12}, \dots, a_8 = a_{81}, \dots, a_8 = a_{87}, \dots, a_n = v_{n1}, \dots, a_n = v_{nm}\}$  is the **set of all items**.
- $2^I = \{X \mid X \subseteq I\}$  is the **set of all itemsets**.
- **Transaction**  $T \in 2^I$  is a subset of items assigning a value to each attribute.  $T = \{a_1 = v_1, a_2 = v_2, \dots, a_n = v_n\}$ .
- A **transaction database**  $D \subseteq 2^I$  is a set of transactions.

# Some Notations (Association and Classification Rules)

- $X, Y, A, B, C, D \subseteq I$  represent **itemsets**, e.g.,  $\{gender = male, age = old\}$ .
- $X, Y$  means  $X \cup Y$  (**conjunction**).
- $T \in D \subseteq 2^I$  represents a **transaction**.
- $T$  **verifies**  $X$  if  $X \subseteq T$ , e.g.,  $T = \{gender = male, age = old, city = Aachen\}$  **verifies**  $X = \{gender = male, age = old\}$ .
- **Support:**  $supp_D(X) = \frac{|\{T \in D | X \subseteq T\}|}{|D|}$  (i.e., the fraction of transactions verifying itemset  $X$ ).

# Some Notations (Association and Classification Rules)

- $X \rightarrow Y$  is an **association rule**, e.g.,  $\{gender = male, age = old\} \rightarrow \{city = Aachen, work = RWTH\}$ .
- $X \rightarrow Y$  is a **classification rule** if  $Y = \{c = v\}$  is a **class item** and  $X$  does not contain a class item ( $c$  is the **class attribute**)
- **Support:**  $supp_D(X \rightarrow Y) = supp_D(X, Y) = \frac{|\{T \in D | X \cup Y \subseteq T\}|}{|D|}$ , i.e., fraction of transactions verifying both  $X$  and  $Y$ .
- **Confidence:**  $conf_D(X \rightarrow Y) = \frac{supp_D(X, Y)}{supp_D(X)}$ .

# Some Notations (Association and Classification Rules)

- If  $X = \{gender = male, age = old\}$  and  $Y = \{city = Aachen, work = RWTH\}$ , then:

$$conf_D(X \rightarrow Y) = \frac{supp_D(\{gender=male,age=old,city=Aachen,work=RWTH\})}{supp_D(\{gender=male,age=old\})}$$

- If  $X = \{gender = male, age = old\}$  and  $C = \{c = reject\}$ , then:

$$conf_D(X \rightarrow C) = \frac{supp_D(\{gender=male,age=old,c=reject\})}{supp_D(\{gender=male,age=old\})}$$

- So far nothing new, but note that **association rules** and **classification rules** are unified.

ps. We will drop the subscripts when clear, i.e.,  $conf(X \rightarrow Y) = conf_D(X \rightarrow Y)$ .

# Extended lift (elift)

- Let  $A, B \rightarrow C$  be an association rule such that  $\text{conf}(B \rightarrow C) > 0$ .
- Extended lift:  $\text{elift}(A, B \rightarrow C) = \frac{\text{conf}(A, B \rightarrow C)}{\text{conf}(B \rightarrow C)}$ .
- $B \rightarrow C$  is called the **base-rule**,  $B$  is the **context**.
- $\text{elift}(A, B \rightarrow C)$  quantifies how the confidence changes when adding  $A$ .

# Potentially Discriminating (PD) Itemsets

- $I_d \subseteq 2^I$  is the predefined set of **Potentially Discriminating (PD) itemsets**.
- The set is assumed to be **downward closed**, i.e.,  $X \in I_d$  and  $Y \in I_d$  implies  $X, Y = X \cup Y \in I_d$ .
- Example  $\{gender = male, country = NL\} \in I_d$  and  $\{age = 50 +\} \in I_d$  implies  $\{gender = male, country = NL, age = 50 +\} \in I_d$ .

# Potentially Discriminating (PD) Itemsets

- $I_d \subseteq 2^I$  is the predefined set of **Potentially Discriminating (PD) itemsets**.
- Note that it does **not** suffice to look at **individual items**:
  - $\{animal = cat, color = black, activity = crossing\}$ ,
  - $\{gender = female, condition = pregnant\}$ , and
  - $\{gender = male, job = hairdresser\}$ , etc.
- **Maybe:**  $\{gender = male\} \notin I_d$  and  $\{job = hairdresser\} \notin I_d$ , but  $\{gender = male, job = hairdresser\} \in I_d$ .

# Potentially Non-Discriminating (PND)

- Potentially Non-Discriminating (PND) itemsets are all other itemsets, i.e.,  $2^I \setminus I_d$ .
- It is assumed that all PND itemsets are not directly discriminating (i.e., sensitive). However, they may correlate with Potentially Discriminating (PD) itemsets.
- Any itemset  $X$  can be split into a maximal PD part  $A$  and a remaining PND part  $B$ , i.e.,  $X = A, B$ ,  $A \in I_d$ ,  $B \notin I_d$ , and  $A$  is maximal.
- Whenever we write  $X = A, B$  we refer to an itemset split in this manner.

# PD/PND classification rules

- Consider a classification rule  $X \rightarrow C$ .
- $X \rightarrow C$  is **Potentially Non-Discriminating (PND)** if  $X$  is a PND itemset ( $X \in 2^I \setminus I_d$ ).
- $X \rightarrow C$  is **Potentially Discriminating (PD)** if  $X = A, B$  with  $A \in I_d$  and  $B \in 2^I \setminus I_d$ .
- PND rules will be considered for indirect discrimination and PD rules for direct discrimination.

# Direct discrimination (PD rules)

- Consider a **PD classification rule**  $A, B \rightarrow C$  with  $A \in I_d$  and  $B \in 2^I \setminus I_d$ .
- **Base rule**  $B \rightarrow C$  is the rule we would like to use.
- $\text{elift}(A, B \rightarrow C) = \frac{\text{conf}(A, B \rightarrow C)}{\text{conf}(B \rightarrow C)}$  measures the relative change in confidence of the base rule  $B \rightarrow C$  when  $A$  is added
- $A, B \rightarrow C$  is  **$\alpha$ -discriminatory** if  $\text{elift}(A, B \rightarrow C) \geq \alpha$ .

# Example

- Consider a PD classification rule  $A, B \rightarrow C$  with  $A = \{gender = male\}$ ,  $B = \{work = RWTH\}$ ,  $C = \{rating = poor\}$ .
- Assume:  $conf(A, B \rightarrow C) = 0.8$ ,  $conf(B \rightarrow C) = 0.2$ , and  $\alpha = 3$
- $elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} = \frac{0.8}{0.2} = 4 \geq \alpha$ .
- Therefore,  $A, B \rightarrow C$  is  $\alpha$ -discriminatory.
- It is better not to use the base rule  $B \rightarrow C$ .

# Example

- Consider a PD classification rule  $A, B \rightarrow C$  with  $A = \{age = 50 +\}$ ,  $B = \{work = RWTH\}$ ,  $C = \{rating = poor\}$ .
- Assume:  $conf(A, B \rightarrow C) = 0.4$ ,  $conf(B \rightarrow C) = 0.2$ , and  $\alpha = 3$
- $elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} = \frac{0.4}{0.2} = 2 < \alpha$ .
- Therefore,  $A, B \rightarrow C$  is not  $\alpha$ -discriminatory.
- It is OK to use the base rule  $B \rightarrow C$  (given  $\alpha = 3$ ).

# Visualization

$A \in I_d$   
Potentially  
Discriminating  
(PD) itemset

{gender = female,  
condition = pregnant}

$B \in 2^I \setminus I_d$   
Potentially Non-  
Discriminating (PND)  
itemset

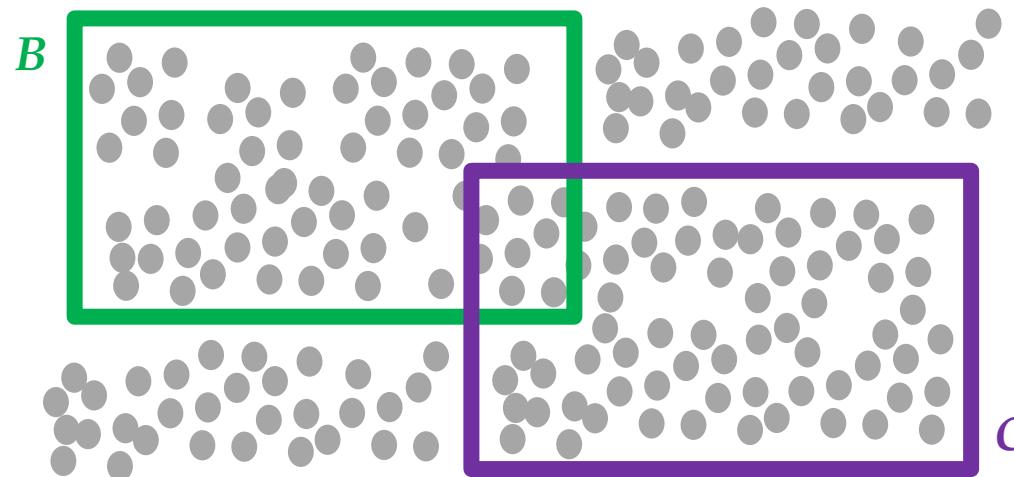
{work hours = < 10}

$C$   
Class Item

{get promoted = No}

Can we use the base rule

{work hours = < 10}  $\rightarrow$  {get promoted = No} ?



$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)}$$

$$conf(B \rightarrow C) = 0.1$$



Chair of Process  
and Data Science

# Visualization

$A \in I_d$   
Potentially  
Discriminating  
(PD) itemset

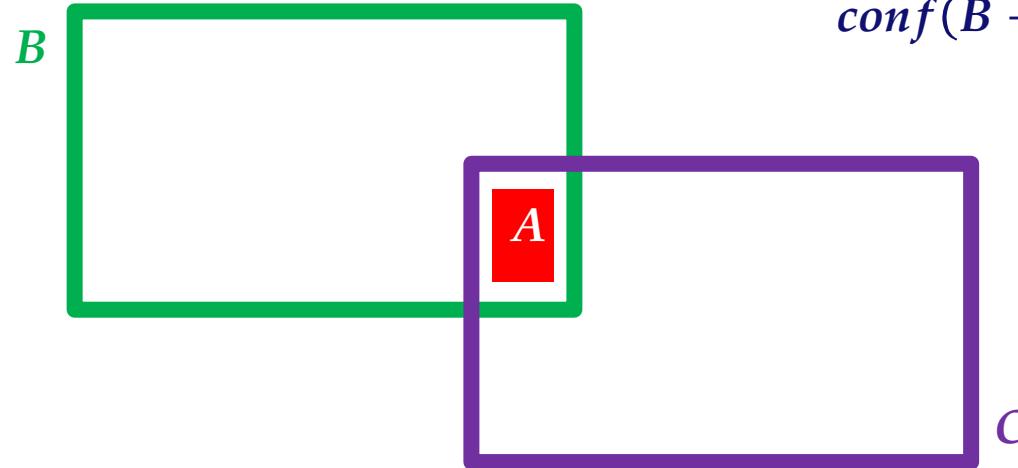
{*gender = female,*  
*condition = pregnant*}

$B \in 2^I \setminus I_d$   
Potentially Non-  
Discriminating (PND)  
itemset

{*work hours = < 10*}

$C$   
Class Item

{*get promoted = No*}



$$conf(A, B \rightarrow C) = 1.0$$

$$conf(B \rightarrow C) = 0.1$$

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} = \frac{1.0}{0.1} = 10$$



Chair of Process  
and Data Science

# Visualization

$A \in I_d$   
Potentially  
Discriminating  
(PD) itemset

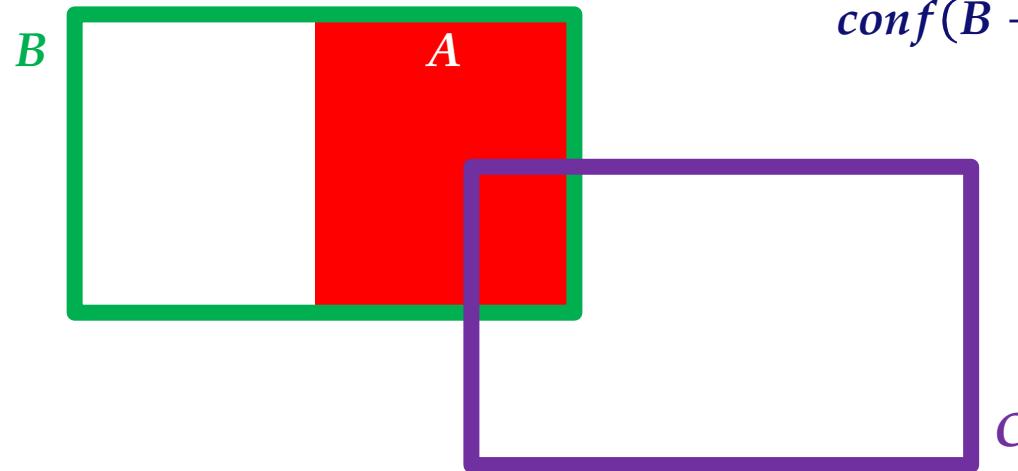
{*gender = female,*  
*condition = pregnant*}

$B \in 2^I \setminus I_d$   
Potentially Non-  
Discriminating (PND)  
itemset

{*work hours = < 10*}

$C$   
Class Item

{*get promoted = No*}



$$conf(A, B \rightarrow C) = 0.2$$

$$conf(B \rightarrow C) = 0.1$$

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} = \frac{0.2}{0.1} = 2$$

# Visualization

$A \in I_d$   
Potentially Discriminating (PD) itemset

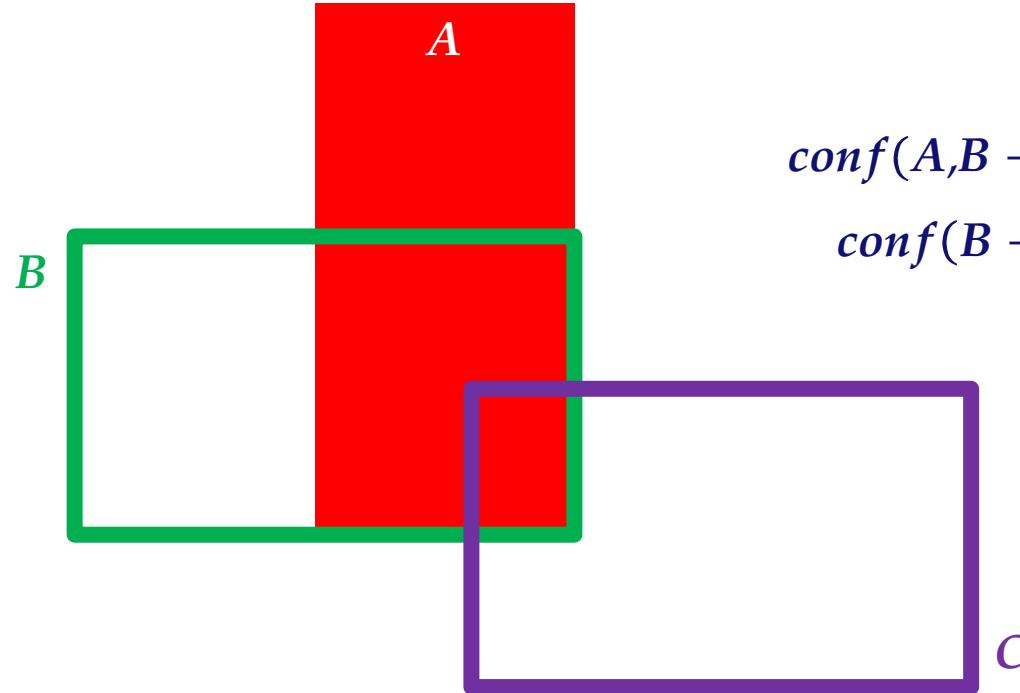
{*gender = female,*  
*condition = pregnant*}

$B \in 2^I \setminus I_d$   
Potentially Non-Discriminating (PND) itemset

{*work hours = < 10*}

$C$   
Class Item

{*get promoted = No*}



$$conf(A, B \rightarrow C) = 0.2$$

$$conf(B \rightarrow C) = 0.1$$

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} = \frac{0.2}{0.1} = 2$$

# Visualization

$A \in I_d$   
Potentially  
Discriminating  
(PD) itemset

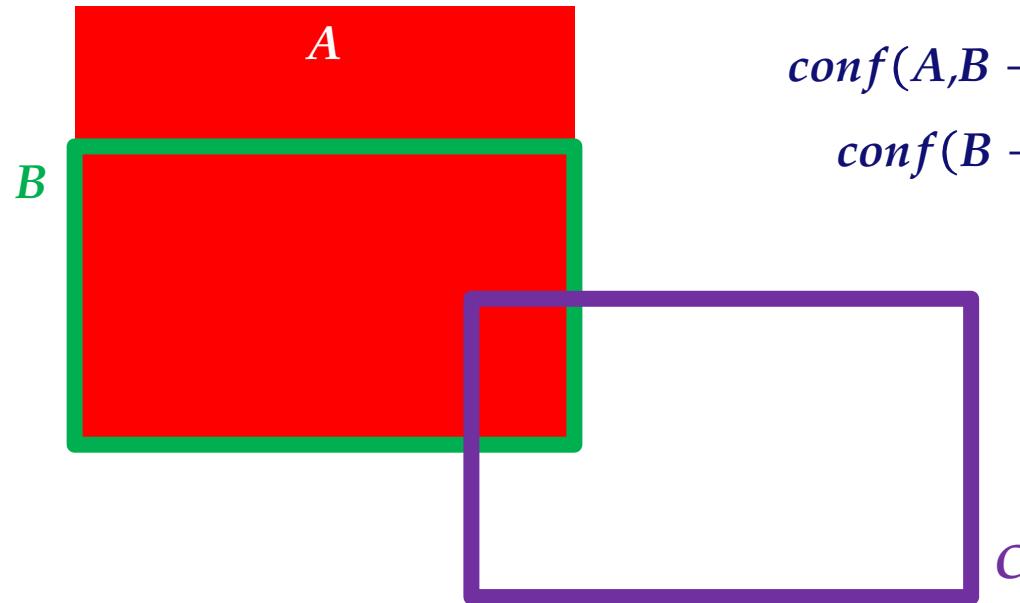
{*gender = female,*  
*condition = pregnant*}

$B \in 2^I \setminus I_d$   
Potentially Non-  
Discriminating (PND)  
itemset

{*work hours = <10*}

$C$   
Class Item

{*get promoted = No*}



$$conf(A, B \rightarrow C) = 0.1$$

$$conf(B \rightarrow C) = 0.1$$

$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} = \frac{0.1}{0.1} = 1$$



Chair of Process  
and Data Science

# Visualization

$A \in I_d$   
Potentially  
Discriminating  
(PD) itemset

{*gender = female,*  
*condition = pregnant*}

$B \in 2^I \setminus I_d$   
Potentially Non-  
Discriminating (PND)  
itemset

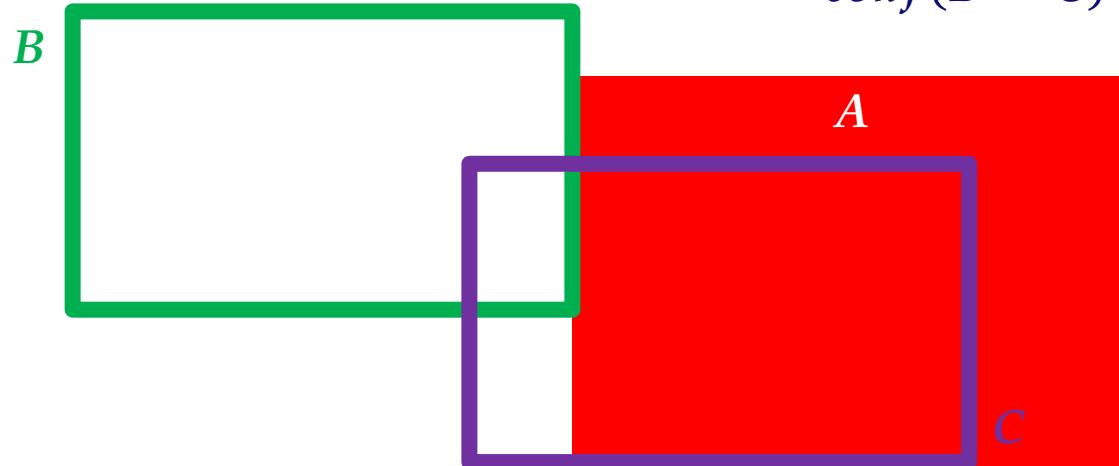
{*work hours = <10*}

$C$   
Class Item

{*get promoted = No*}

$$conf(A, B \rightarrow C) = 0.0$$

$$conf(B \rightarrow C) = 0.1$$



$$elift(A, B \rightarrow C) = \frac{conf(A, B \rightarrow C)}{conf(B \rightarrow C)} = \frac{0.0}{0.1} = 0$$

# Binary Classification and Negation

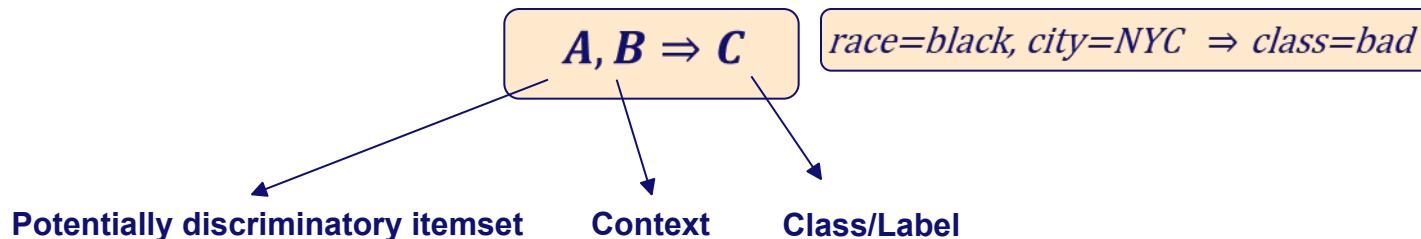
- The paper considers discrimination from two sides, i.e.,  $A, B \rightarrow C$  and  $A, B \rightarrow \neg C$ , in one single metric.
- Left out here (trivial, but leads to more notation).

# Direct and Indirect Discrimination

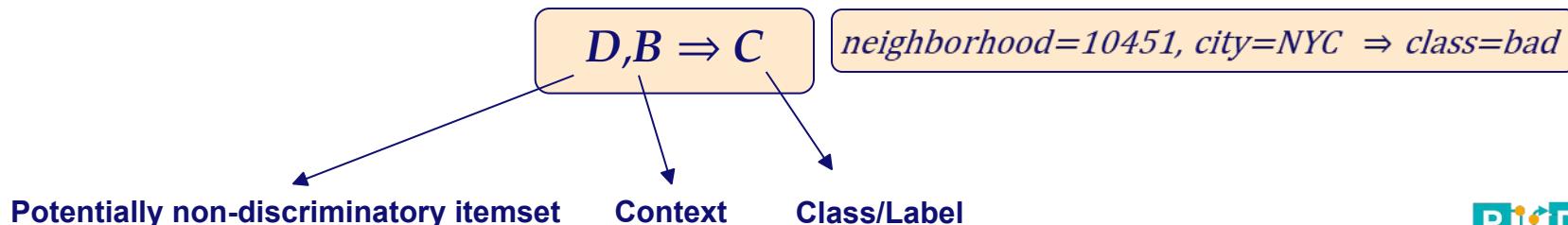
- There are two types of discrimination:
  - Direct
    - Direct discrimination consists of rules or procedures that explicitly impose “disproportionate burdens” on minority or disadvantaged groups.
    - Defined based on potentially discriminatory itemsets
  - Indirect
    - Indirect discrimination consists of rules or procedures that, while not explicitly mentioning discriminatory attributes, intentionally or not impose the same disproportionate burdens.
    - Defined based on potentially non-discriminatory itemsets

# Direct and Indirect Discrimination

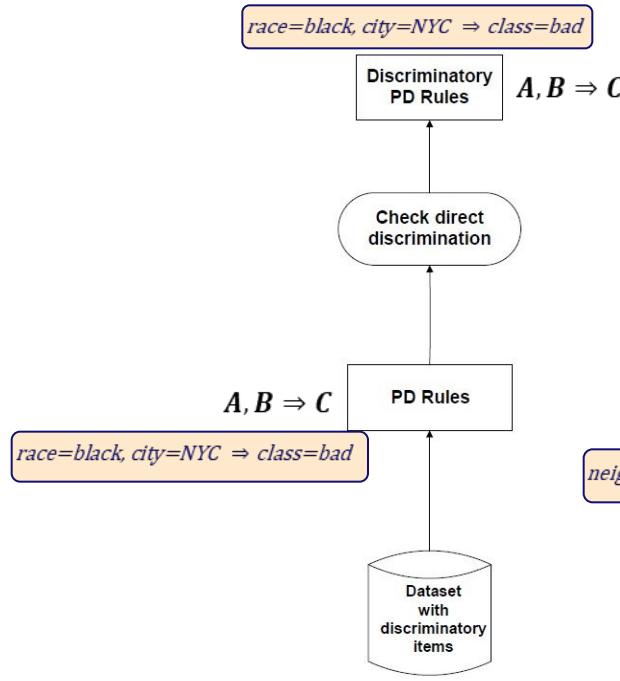
- Direct discrimination is modeled through Potentially Discriminatory (PD) rules:



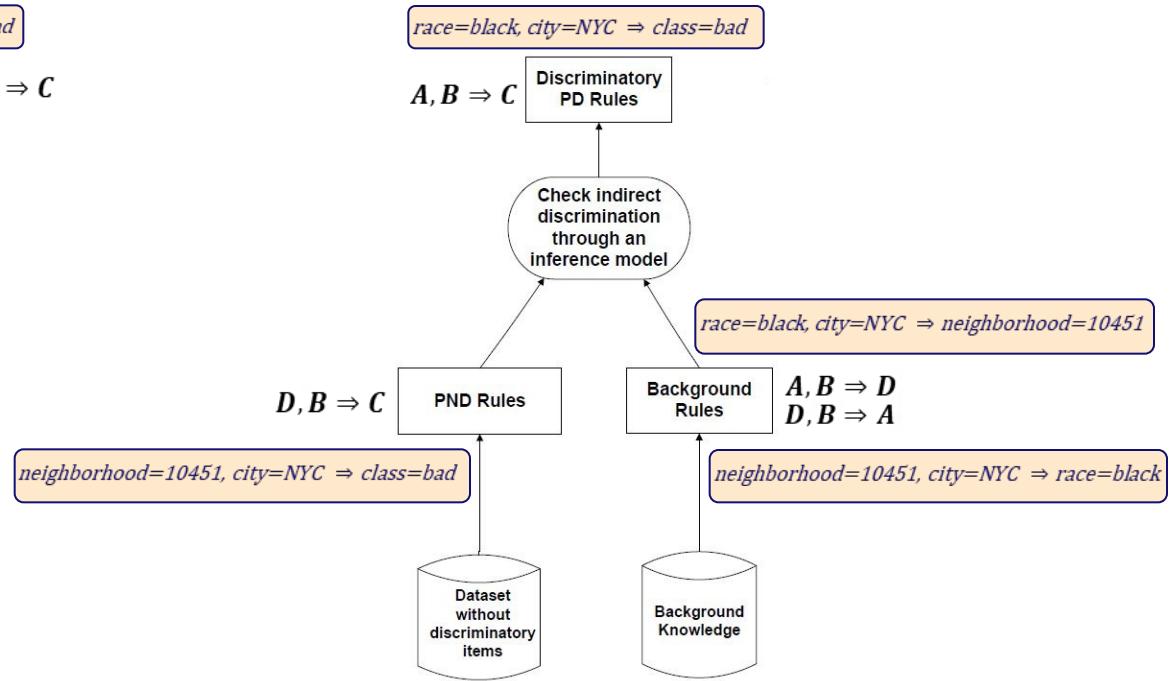
- Indirect discrimination is modeled through Potentially Non-Discriminatory (PND) rules:



# Direct and Indirect Discrimination



Direct Discrimination



Indirect Discrimination

# Example (from paper)

If we consider  $\alpha = 3$ , then:

**Base Rule**

$city=NYC \Rightarrow class=bad$        $B \Rightarrow C$       Confidence: 0.25

**PD Rule**

$race=black, city=NYC \Rightarrow class=bad$        $A, B \Rightarrow C$       Confidence: 0.75

$elift = 0.75/0.25 = 3$        $elift \geq 3$ , the rule is discriminatory

**PND Rule**

$neighborhood=10451, city=NYC \Rightarrow class=bad$        $D, B \Rightarrow C$       Confidence: 0.95

**Background knowledge**

$neighborhood=10451, city=NYC \Rightarrow race=black$        $D, B \Rightarrow A$       Confidence: 0.80

**Inferred Rule**

$race=black, neighborhood=10451, city=NYC \Rightarrow class=bad$        $A, D, B \Rightarrow C$       Lower bound Confidence: 0.94

$elift = 0.94/0.25 = 3.37$        $elift \geq 3$ , the rule is discriminatory



There are some methods to calculate lower bound (see later).

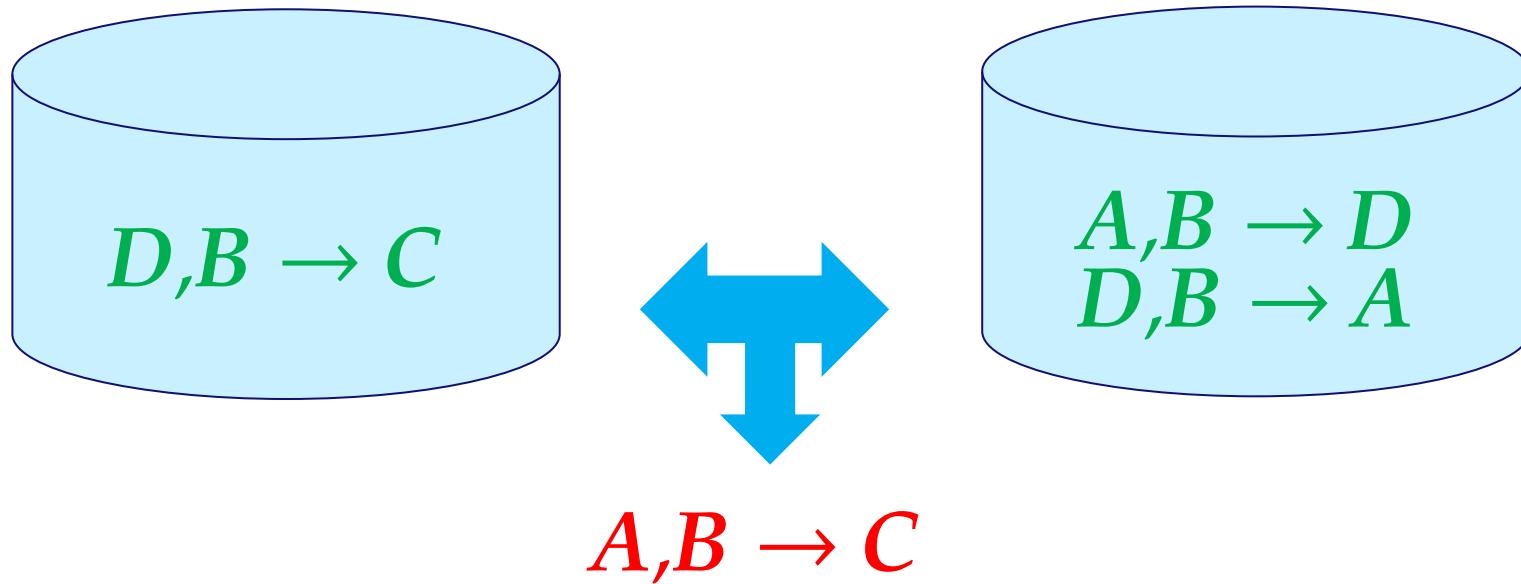


Chair of Process  
and Data Science

# Recall: Direct discrimination

- Consider a **PD classification rule**  $A, B \rightarrow C$  with  $A \in I_d$  and  $B \in 2^I \setminus I_d$ .
- $\text{elift}(A, B \rightarrow C) = \frac{\text{conf}(A, B \rightarrow C)}{\text{conf}(B \rightarrow C)}$  measures the relative change in confidence of the base rule  $B \rightarrow C$  when  $A$  is added
- $A, B \rightarrow C$  is  **$\alpha$ -discriminatory** if  $\text{elift}(A, B \rightarrow C) \geq \alpha$ .
- If so, avoid using  $B \rightarrow C$ .

# Indirect Discrimination (PND rules)



# Indirect Discrimination (PND rules)

- Consider a PND classification rule  $D, B \rightarrow C$  with  $D \cup B \in 2^I \setminus I_d$  and  $\text{conf}(B \rightarrow C) > 0$ .
- Assume:  $\text{conf}(A, B \rightarrow D) = \beta_1$  and  $\text{conf}(D, B \rightarrow A) = \beta_2$
- Then we can derive that:  $A, B \rightarrow C$  is  $\alpha$ -discriminatory if  $\frac{\beta_1(\beta_2 + \text{conf}(D, B \rightarrow C) - 1)}{\beta_2 \cdot \text{conf}(B \rightarrow C)} \geq \alpha$ .
- If so, avoid using  $B \rightarrow C$  (although  $A$  was not in the database).

# Indirect Discrimination (PND rules)

- Consider a PND classification rule  $D, B \rightarrow C$  with  $D \cup B \in 2^I \setminus I_d$  and  $\text{conf}(B \rightarrow C) > 0$ .
- Suppose that “ $A \approx D$ ” in  $B$ :
  - $\text{conf}(A, B \rightarrow D) = 1$  and  $\text{conf}(D, B \rightarrow A) = 1$
- Then we can derive that:  $A, B \rightarrow C$  is  $\alpha$ -discriminatory if  $\frac{1(1+\text{conf}(D, B \rightarrow C)-1)}{1 \cdot \text{conf}(B \rightarrow C)} = \frac{\text{conf}(D, B \rightarrow C)}{\text{conf}(B \rightarrow C)} \geq \alpha$ .
- If so, avoid using  $B \rightarrow C$  (although  $A$  was not in the database).

# Visualization

$A \in I_d$   
Potentially  
Discriminating  
(PD) itemset

{*gender = female,*  
*condition = pregnant*}

$B \in 2^I \setminus I_d$   
Potentially Non-  
Discriminating (PND)  
itemset

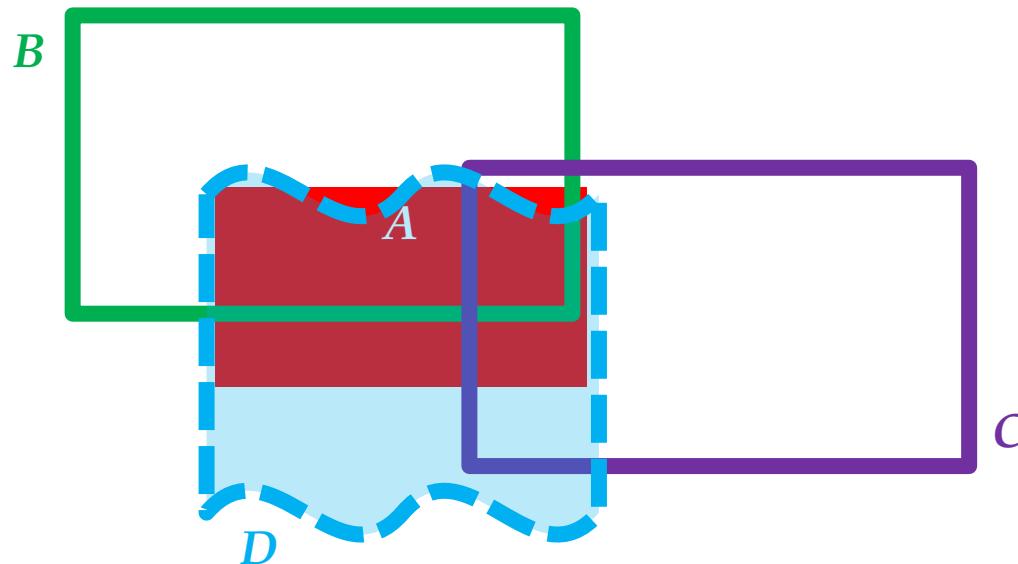
{*work hours = <10*}

$D \in 2^I \setminus I_d$   
(*age = 50 +*)

{*get promoted = No*}

**C**  
Class Item

“ $A \approx D$ ” in  $B$



# Discrimination-aware Decision Tree

based on

## Discrimination Aware Decision Tree Learning

Faisal Kamiran, Toon Calders and Mykola Pechenizkiy  
Email: {f.kamiran,t.calders,m.pechenizkiy}@tue.nl  
Eindhoven University of Technology, The Netherlands



# How biased historic data can affect decisions

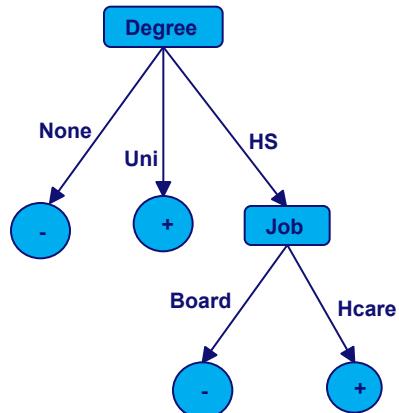
When historic data is suspected to be biased (to contain discrimination)

The decision making model based on this data would be biased

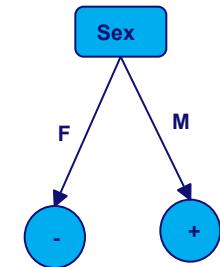
Finally decisions would be discriminatory

# How biased historic data can affect decisions

Exp	Degree	Job	Class
Exp >10	HS	Board	-
5< Exp <10	Uni	Board	+
Exp >10	HS	Board	-
5< Exp <10	HS	Hcare	+
Exp < 5	HS	Hcare	+
Exp < 5	HS	Board	-
Exp < 5	None	Edu	-
Exp >10	None	Hcare	-
Exp < 5	Uni	Edu	+
Exp >10	Uni	Board	+



Sex	Exp	Degree	Job	Class
F	Exp >10	HS	Board	-
M	5< Exp <10	Uni	Board	+
F	Exp >10	HS	Board	-
M	5< Exp <10	HS	Hcare	+
M	Exp < 5	HS	Hcare	+
F	Exp < 5	HS	Board	-
F	Exp < 5	None	Edu	-
F	Exp >10	None	Hcare	-
M	Exp < 5	Uni	Edu	+
M	Exp >10	Uni	Board	+



historical bias

# Three Approaches to make Decision Trees fair

- **Pre-processing**
  - Removing discriminatory attributes from the historic data.
  - Selectively removing/duplicating instances.
  - We still have to deal with indirect discrimination.
- **In-processing**
  - Considering dependency to the discriminatory attribute/attributes and the accuracy of the split simultaneously while making a decision tree.
- **Post-processing**
  - Relabeling leaves in such a way that discrimination is lowered with a minimal loss in accuracy.

**Always a tradeoff between fairness and accuracy!**

# Defining Discrimination

- A classifier  $C$  is a function from  $\prod_{i=1}^n \text{dom}(A_i)$  to  $\{+, -\}$ .
- $B$  is a binary sensitive attribute with  $\text{dom}(B) = \{0, 1\}$ .
  - Discrimination means that instances with  $B=1$  are less likely to be classified as positive by the classifier  $C$ .
  - For example for a job application example we suppose that for men  $B=0$  and for women  $B=1$ .
- Discrimination of  $C$  with respect to  $B$  in dataset  $D$  is defined as:

Probability of getting a positive label for an instance with  $B = 0$

$$\text{disc}_B(C, D) := \frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|}$$

Probability of getting a positive label for an instance with  $B = 1$

We want to protect  $B=1$ .



Chair of Process  
and Data Science

# Defining Discrimination

A higher discrimination means that tuples with  $B = 1$  are less likely to be classified as positive by the classifier  $C$  than others.

## data view

$$disc_B(D) := \frac{\frac{|\{x \in D \mid x.B = 0, x.Class = +\}|}{|\{x \in D \mid x.B = 0\}|}}{|\{x \in D \mid x.B = 1, x.Class = +\}|} - \frac{|\{x \in D \mid x.B = 1, x.Class = +\}|}{|\{x \in D \mid x.B = 1\}|}$$

Annotations:

- database (points to the first denominator)
- real class (points to the first numerator)
- positive outcome (points to the second numerator)
- predicted class (points to the second denominator)

## classifier view

$$disc_B(C, D) := \frac{\frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|}}{|\{x \in D \mid x.B = 1, C(x) = +\}|} - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|}$$

Annotations:

- classifier (points to the first denominator)
- positive outcome (points to the first numerator)
- positive outcome (points to the second numerator)

We want to protect  $B=1$ .



Chair of Process  
and Data Science

# Example

	Sex	Ethnicity	Highest Degree	Job Type	Class
B=0	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	-
B=1	f	non-nat.	univ.	education	-
	f	native	h. school	education	-
	f	native	none	healthcare	+
	f	non-nat.	univ.	education	-
	f	native	h. school	board	+

A higher discrimination means that tuples with  $B = 1$  are less likely to be classified as positive by the classifier  $C$  than others.

$$disc_B(D) := \frac{|\{x \in D \mid x.B = 0, x.Class = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, x.Class = +\}|}{|\{x \in D \mid x.B = 1\}|}$$

$$\frac{4}{5} - \frac{2}{5} = 40\%$$

We want to protect B=1.

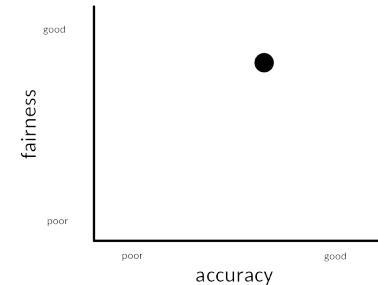
# Trade-off

- The discrimination aware classification problem is to learn a classifier such that
  - The accuracy of C is high, and
  - The discrimination of C with respect to B is low.

So we consider discrimination aware classification as a **multi-objective optimization problem**.

- DA-optimal Classifier** (like Pareto optimality): We call a classifier C optimal w.r.t. Discrimination and Accuracy (DA-optimal) in set of classifiers if for every other classifier C' in set of classifiers either  $\text{disc}(C) < \text{disc}(C')$ , or  $\text{acc}(C') < \text{acc}(C)$ .

$$\begin{aligned} \text{disc}_B(C, D) := & \frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|} \\ & - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|} \end{aligned}$$



# Change Decision Tree Learning (in process)

based on

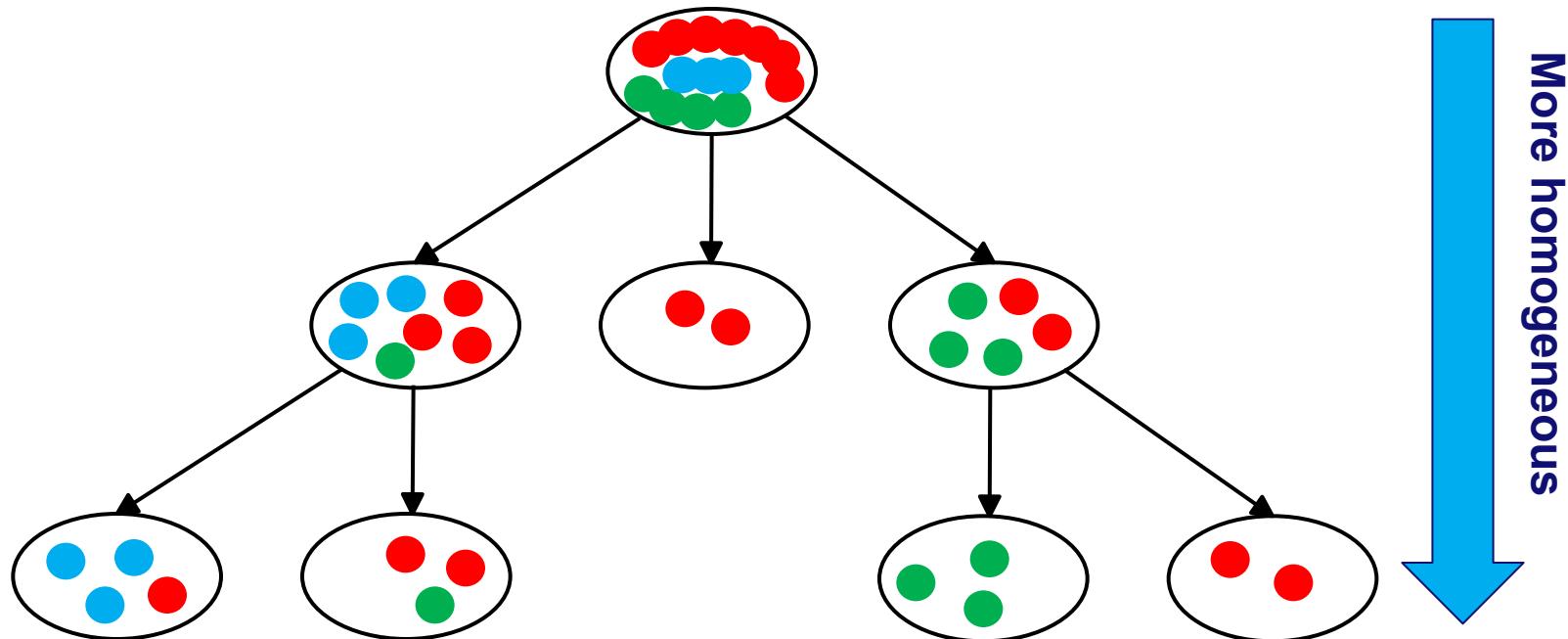
## Discrimination Aware Decision Tree Learning

Faisal Kamiran, Toon Calders and Mykola Pechenizkiy  
Email: {f.kamiran,t.calders,m.pechenizkiy}@tue.nl  
Eindhoven University of Technology, The Netherlands



# Traditional Decision Tree Learning Using Information Gain

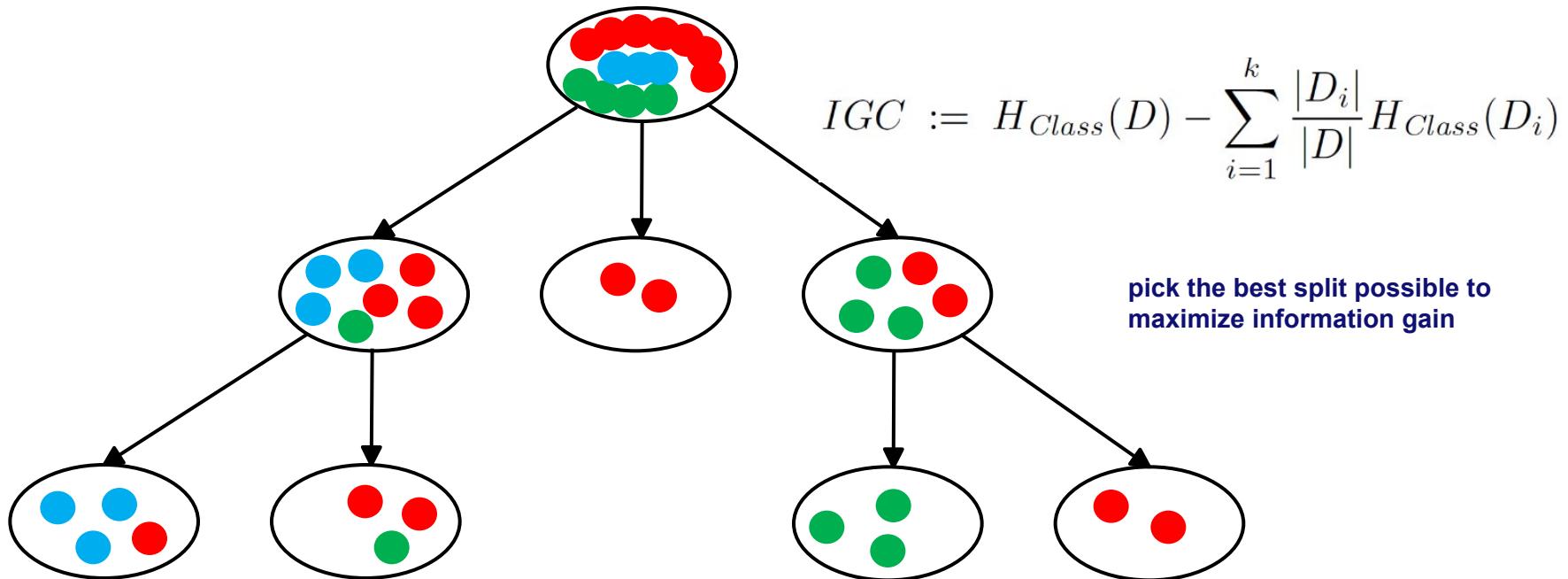
(see Lecture 4)



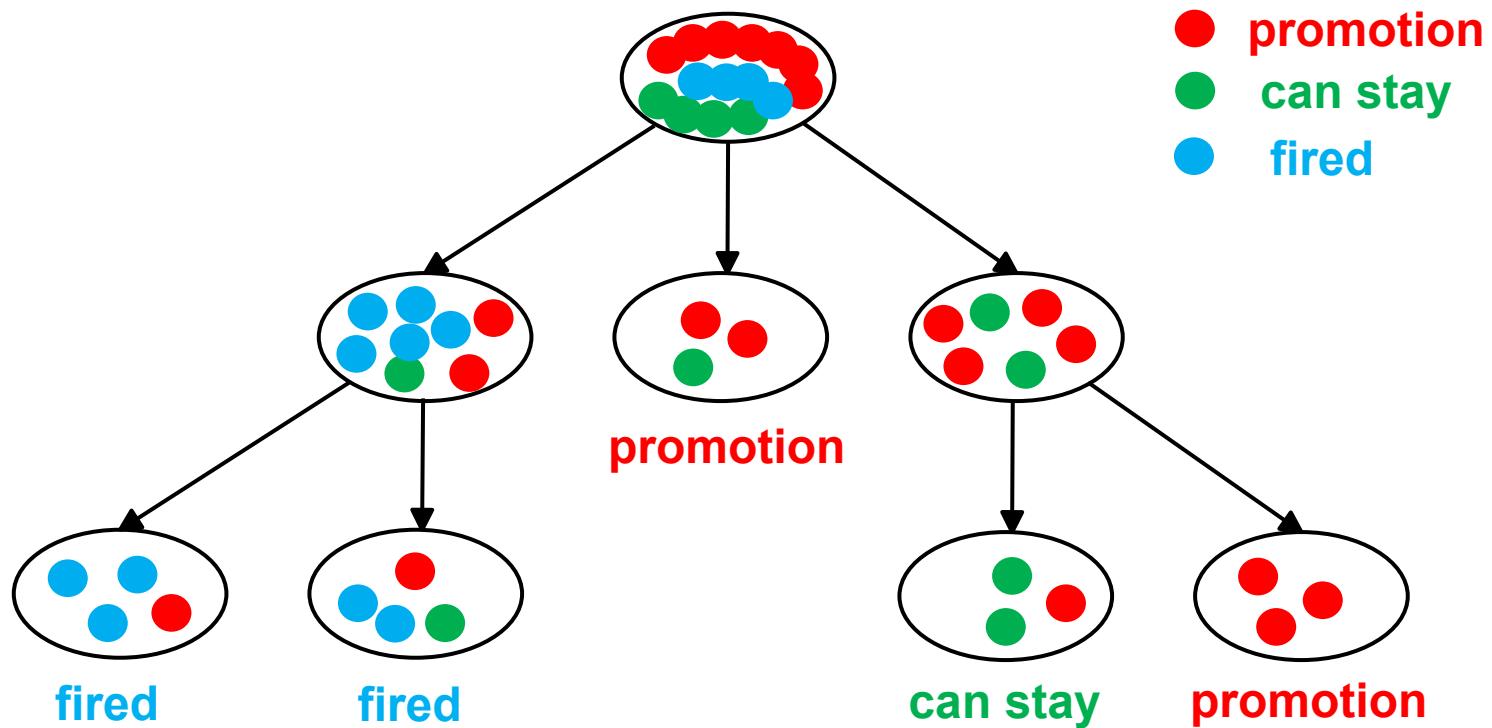
Information gain = improvement in knowledge  
(predictability of class label in nodes)

# Traditional Decision Tree Learning Using Information Gain

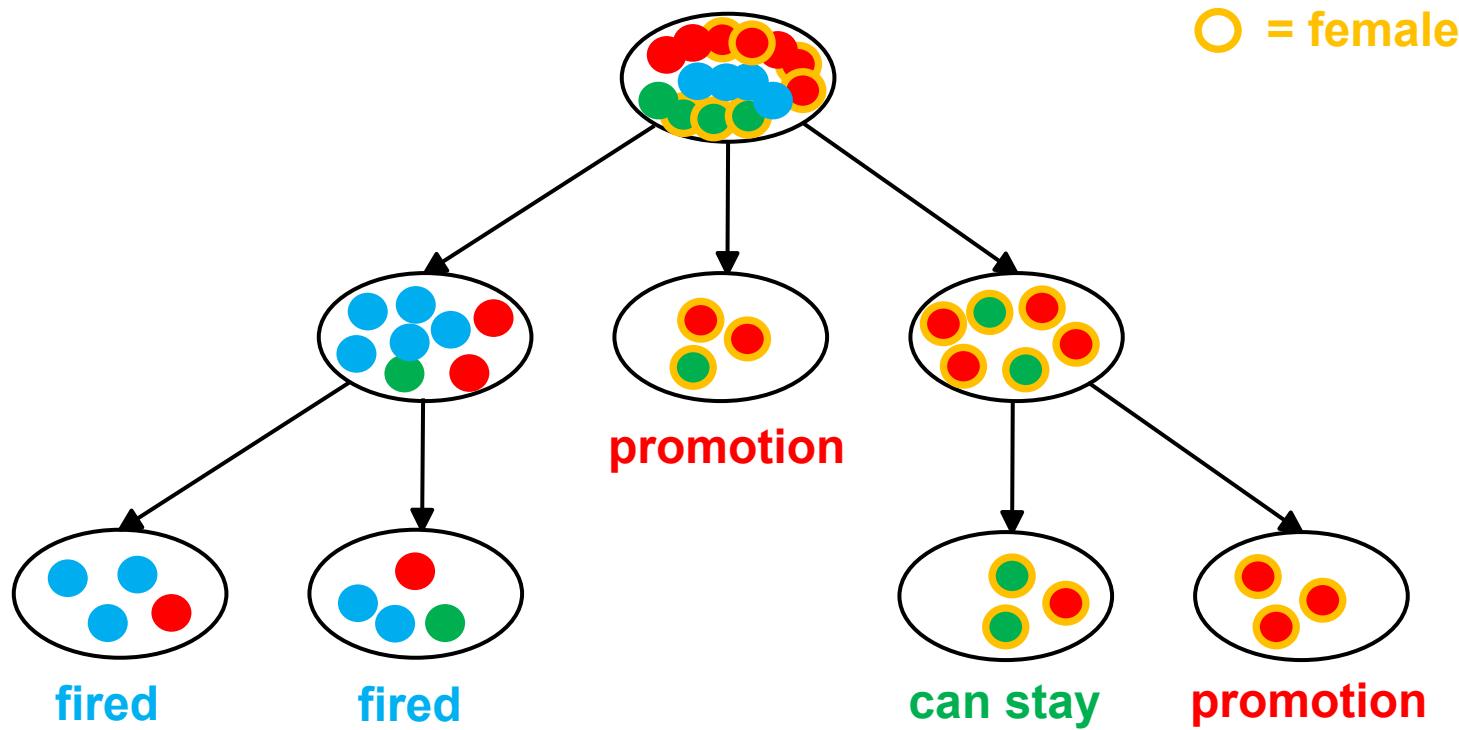
(see Lecture 4)



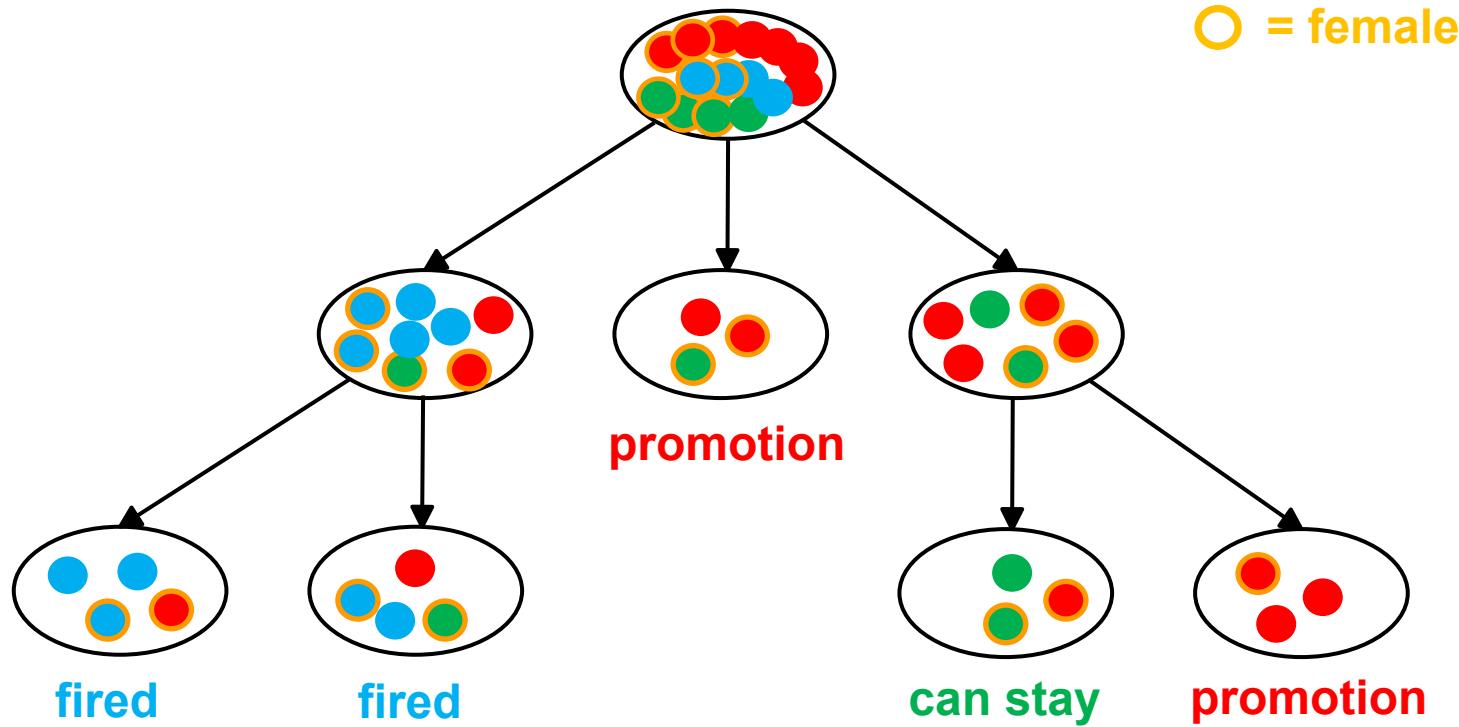
# What if



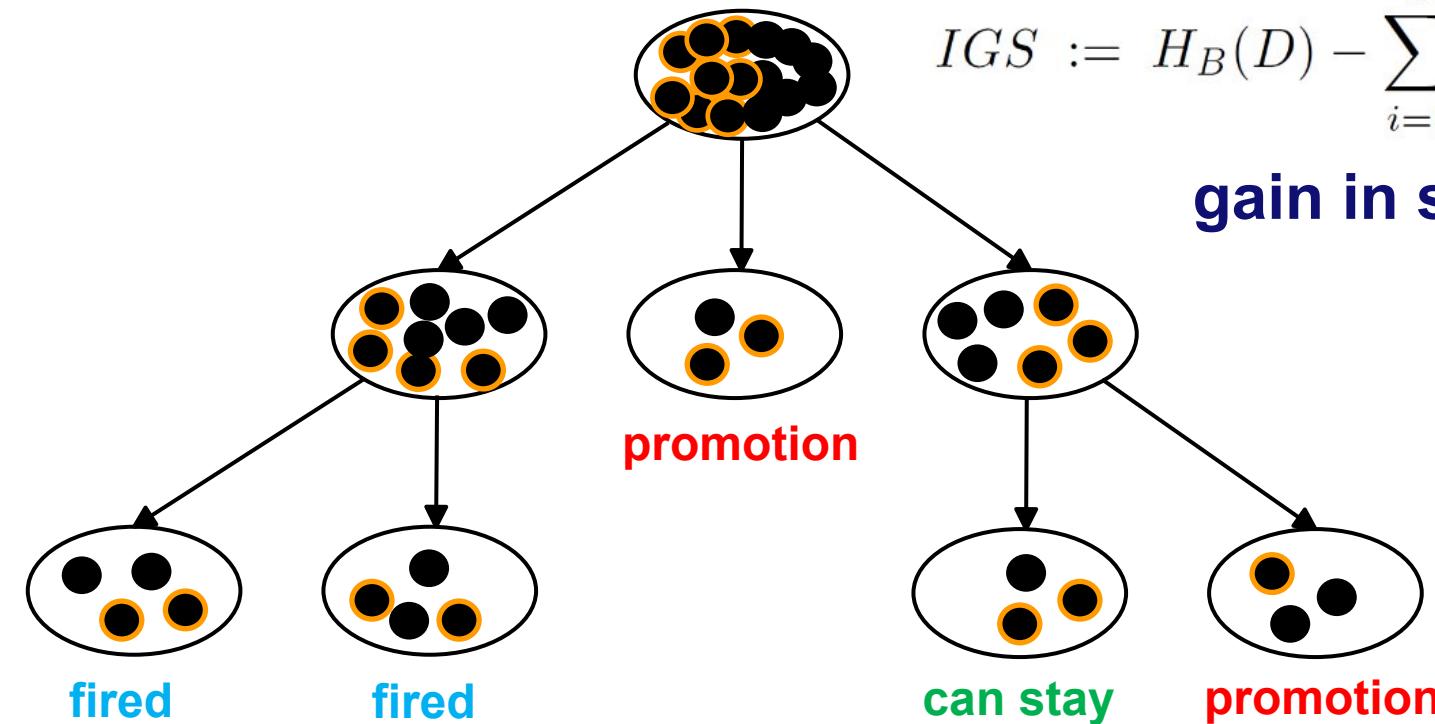
# What if



# What if



# Solution



$$IGS := H_B(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} H_B(D_i)$$

**gain in sensitivity**

# Two forces when splitting

**IGC** = classical information gain

$$IGC := H_{Class}(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} H_{Class}(D_i)$$

maximize

**IGS** = gain in sensitivity

$$IGS := H_B(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} H_B(D_i)$$

minimize

**Combine both!**

# Two forces when splitting

**IGC = classical information gain**

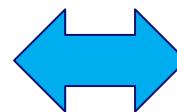
**IGS = gain in sensitivity**

$$IGC := H_{Class}(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} H_{Class}(D_i)$$

maximize

$$IGS := H_B(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} H_B(D_i)$$

minimize



**Paper describes three approaches:**

1. Maximize IGC – IGS
2. Maximize IGC/IGS
3. Maximize IGC + IGS (only with relabeling)

# Dependency-Aware Tree Construction In-processing

- Normally we have just one information gain value which shows that how much a node is good for separating instances into positive and negative labeled.

$$IGC := H_{Class}(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} H_{Class}(D_i)$$

- But here we will add another gain which shows that how much a node is good for separating instances into discriminatory and non-discriminatory.

$$IGS := H_B(D) - \sum_{i=1}^k \frac{|D_i|}{|D|} H_B(D_i)$$

- Then the new criteria for it could be:

$$\frac{IGC}{IGS}$$

$$IGC - IGS$$



# Relabel Tree (post processing)

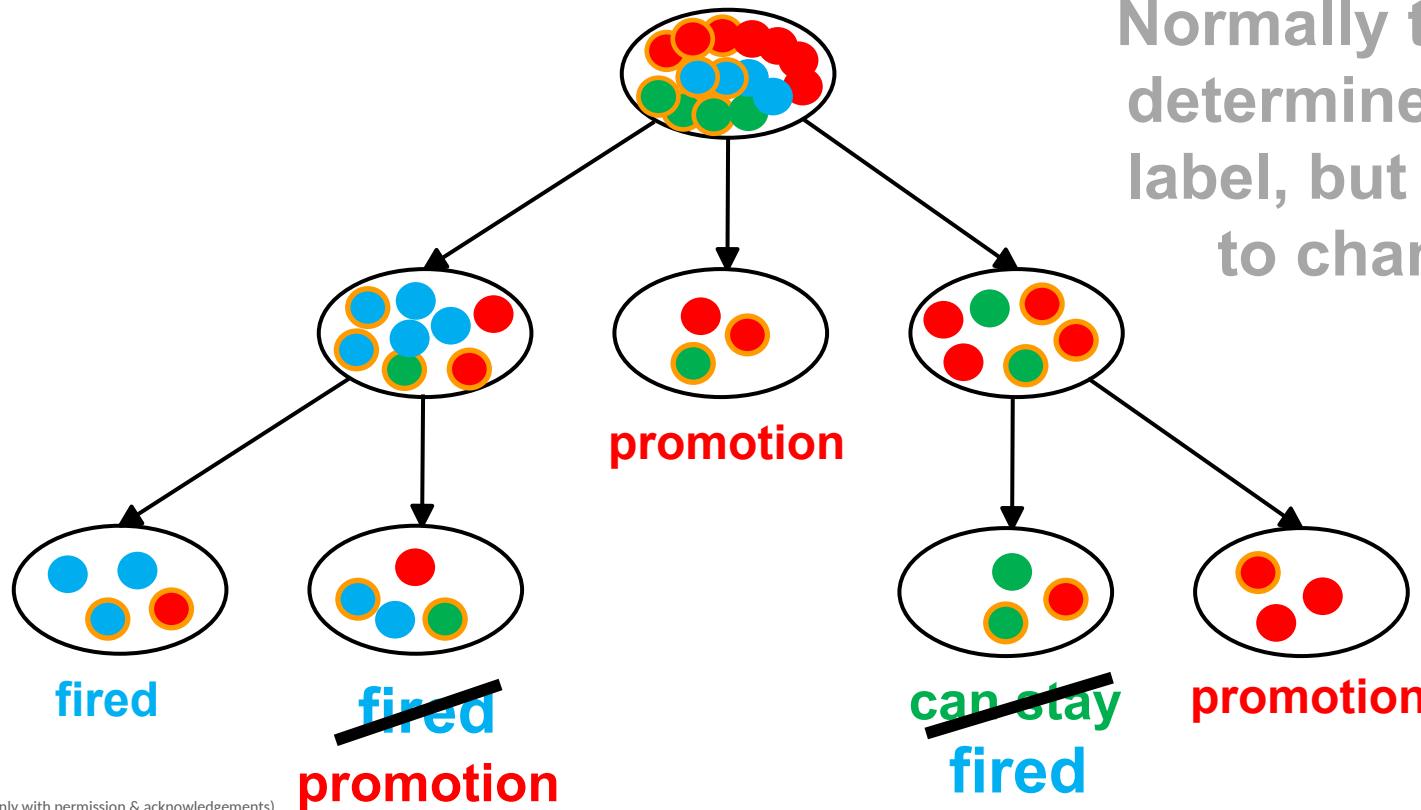
based on

## Discrimination Aware Decision Tree Learning

Faisal Kamiran, Toon Calders and Mykola Pechenizkiy  
Email: {f.kamiran,t.calders,m.pechenizkiy}@tue.nl  
Eindhoven University of Technology, The Netherlands



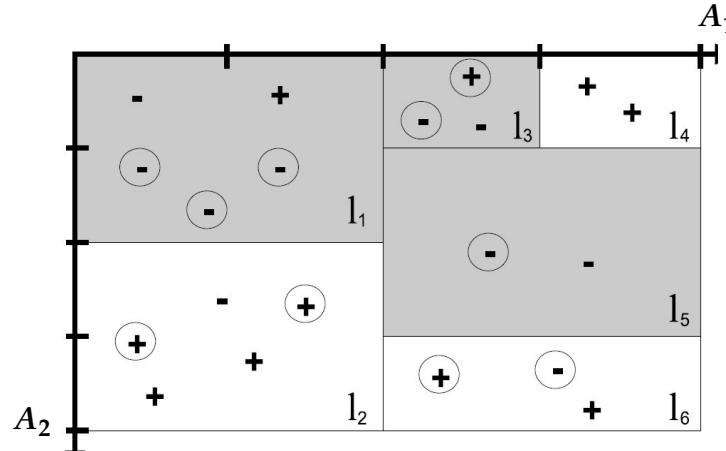
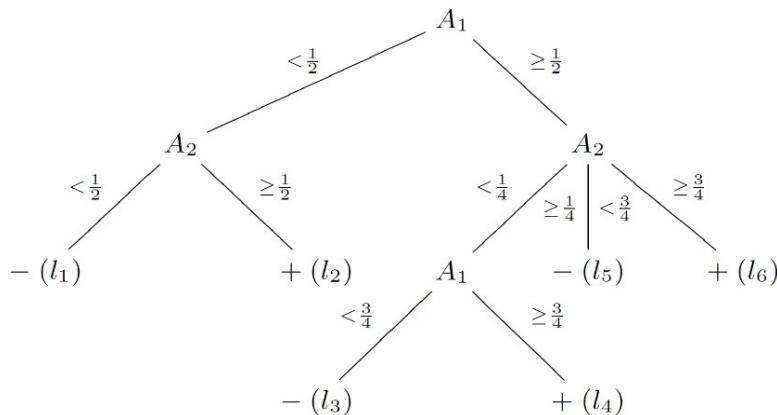
# Idea: Lower Accuracy to Improve Fairness



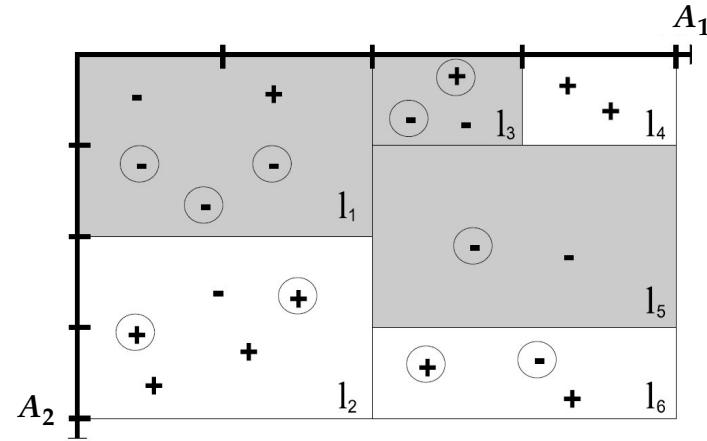
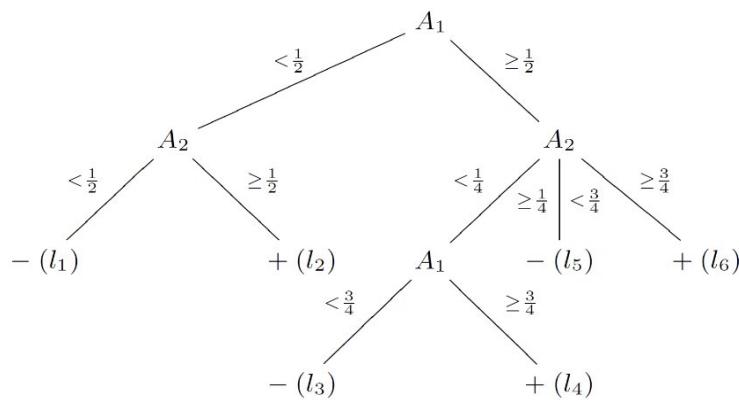
Normally the majority determines the class label, but we are free to change this.

# The Concept of Relabeling

- In this section we assume that a tree is already given and the goal is to reduce the discrimination of the tree by changing the class labels of some of the leaves.
- Let  $T$  be a decision tree with  $n$  leaves. Such a decision tree partitions the example space into  $n$  non-overlapping regions.



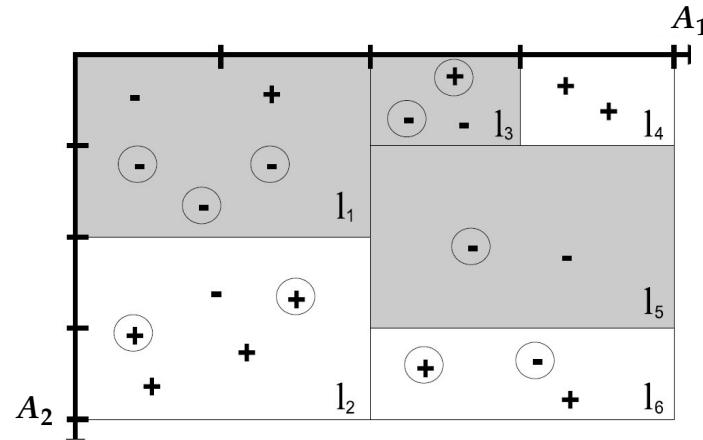
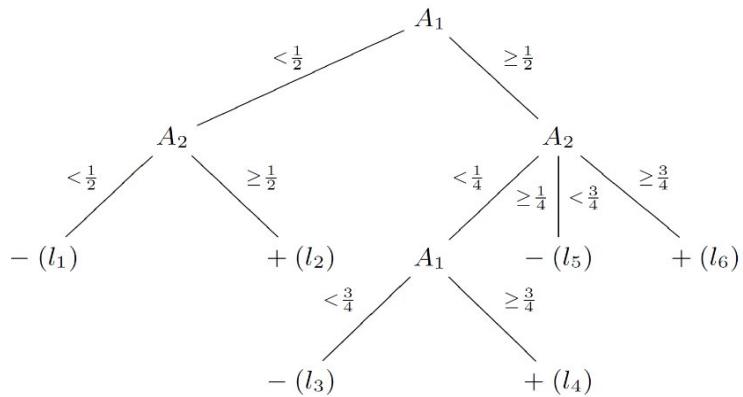
# The Concept of Relabeling



- Regions are classified into white and gray color regions. When the majority class is **-** the color is **gray** and when the majority class in the region is **+** the color is **white**.
- Encircled examples have  $B = 1$  (e.g., minority group).

We want to protect  $B=1$ .

# The Concept of Relabeling



- Actually leaves are labeled with the majority class of their corresponding region. In the relabeling technique we are going to relabel the leaves of the decision tree in such a way that the discrimination decreases with minimal loss in accuracy.
- Since decision tree is made based on the best values for accuracy (because of information gain concept), making any change will decrease the accuracy and we are looking for minimum loss in accuracy and maximum reduction in discrimination.

# Accuracy and Discrimination of Tree

- We will use an extended version of confusion matrix in such a way that probabilities would be based on the probability of being a discriminatory attribute. This matrix is called the **contingency table**.

Class →	−	+	
Pred. →	−/+	−/+	
$B = 1$	$U_1/U_2$	$V_1/V_2$	$b$
$B = 0$	$W_1/W_2$	$X_1/X_2$	$\bar{b}$
	$N_1/N_2$	$P_1/P_2$	1

$$acc_T = N_1 + P_2$$
$$disc_T = \frac{W_2 + X_2}{\bar{b}} - \frac{U_2 + V_2}{b}$$

We want to protect  $B=1$ .

# Accuracy and Discrimination of Tree

- We will use an extended version of confusion matrix in such a way that probabilities would be based on the probability of being a discriminatory attribute. This matrix is called the **contingency table**.

Class →	−	+	
Pred. →	−/+	−/+	
$B = 1$	$U_1/U_2$	$V_1/V_2$	$b$
$B = 0$	$W_1/W_2$	$X_1/X_2$	$\bar{b}$
	$N_1/N_2$	$P_1/P_2$	1

$$acc_T = \frac{N_1 + P_2}{N_1 + P_2 + W_2 + X_2} \quad \text{correctly classified}$$
$$disc_T = \frac{W_2 + X_2}{\bar{b}} - \frac{U_2 + V_2}{b}$$

We want to protect  $B=1$ .

# Accuracy and Discrimination of Tree

$$disc_B(C, D) := \frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|}$$

$$acc_T = \frac{N_1 + P_2}{W_2 + X_2} - \frac{U_2 + V_2}{b}$$

$$disc_T = \frac{W_2 + X_2}{\bar{b}} - \frac{U_2 + V_2}{b}$$

We want to protect B=1.

e.g. fraction of males  
getting positive prediction

e.g. fraction of females  
getting positive prediction

Class →	−	+	
Pred. →	−/+	−/+	
B = 1	$U_1/U_2$	$V_1/V_2$	b
B = 0	$W_1/W_2$	$X_1/X_2$	$\bar{b}$
	$N_1/N_2$	$P_1/P_2$	1

# Accuracy and Discrimination of Tree

$$disc_B(C, D) := \frac{|\{x \in D \mid x.B = 0, C(x) = +\}|}{|\{x \in D \mid x.B = 0\}|} - \frac{|\{x \in D \mid x.B = 1, C(x) = +\}|}{|\{x \in D \mid x.B = 1\}|}$$

$$acc_T = \frac{N_1 + P_2}{W_2 + X_2}$$

$$disc_T = \frac{W_2 + X_2}{\bar{b}} - \frac{U_2 + V_2}{b}$$

We want to protect B=1.

e.g. fraction of males  
getting positive prediction

e.g. fraction of females  
getting positive prediction

Class →	−	+	
Pred. →	−/+	−/+	
B = 1	$U_1/U_2$	$V_1/V_2$	b
B = 0	$W_1/W_2$	$X_1/X_2$	$\bar{b}$
	$N_1/N_2$	$P_1/P_2$	1

# Accuracy and Discrimination of Tree (Example)

Dataset		
Class →	-	+
Pred. →	-/+	-/+
$B = 1$	$U_1/U_2$	$V_1/V_2$
$B = 0$	$W_1/W_2$	$X_1/X_2$
	$N_1/N_2$	$P_1/P_2$
	1	

Leaf 1		
	-	+
$B = 1$	$u$	$v$
$B = 0$	$w$	$x$
	$n$	$p$
		$a$

Dataset		
Class →	-	+
Pred. →	-/+	-/+
$B = 1$	$\frac{5}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{3}{20}$
$B = 0$	$\frac{3}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{5}{20}$
	$\frac{8}{20}/\frac{2}{20}$	$\frac{2}{20}/\frac{8}{20}$
		1

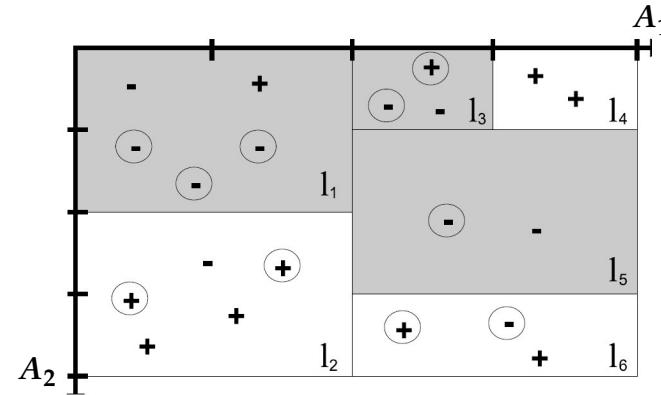
Leaf $l_3$		
	-	+
$B = 1$	$\frac{1}{20}$	$\frac{1}{20}$
$B = 0$	$\frac{1}{20}$	$0$
	$\frac{2}{20}$	$\frac{3}{20}$

$$acc_T = N_1 + P_2$$

$$disc_T = \frac{W_2 + X_2}{\bar{b}} - \frac{U_2 + V_2}{b}$$

$$\text{Accuracy} = \frac{8}{20} + \frac{8}{20} = 0.8$$

$$\text{Discrimination} = \frac{\frac{1}{20} + \frac{5}{20}}{\frac{1}{2}} - \frac{\frac{1}{20} + \frac{3}{20}}{\frac{1}{2}} = 0.2$$



# Effect of Relabeling on Discrimination and Accuracy

- The effect of relabeling the leaf now depends on the majority class of the leaf
- If  $p > n$ , the label of the leaf changes from + to - and the effect on accuracy and discrimination is expressed by:

$$\Delta acc_l = n - p$$

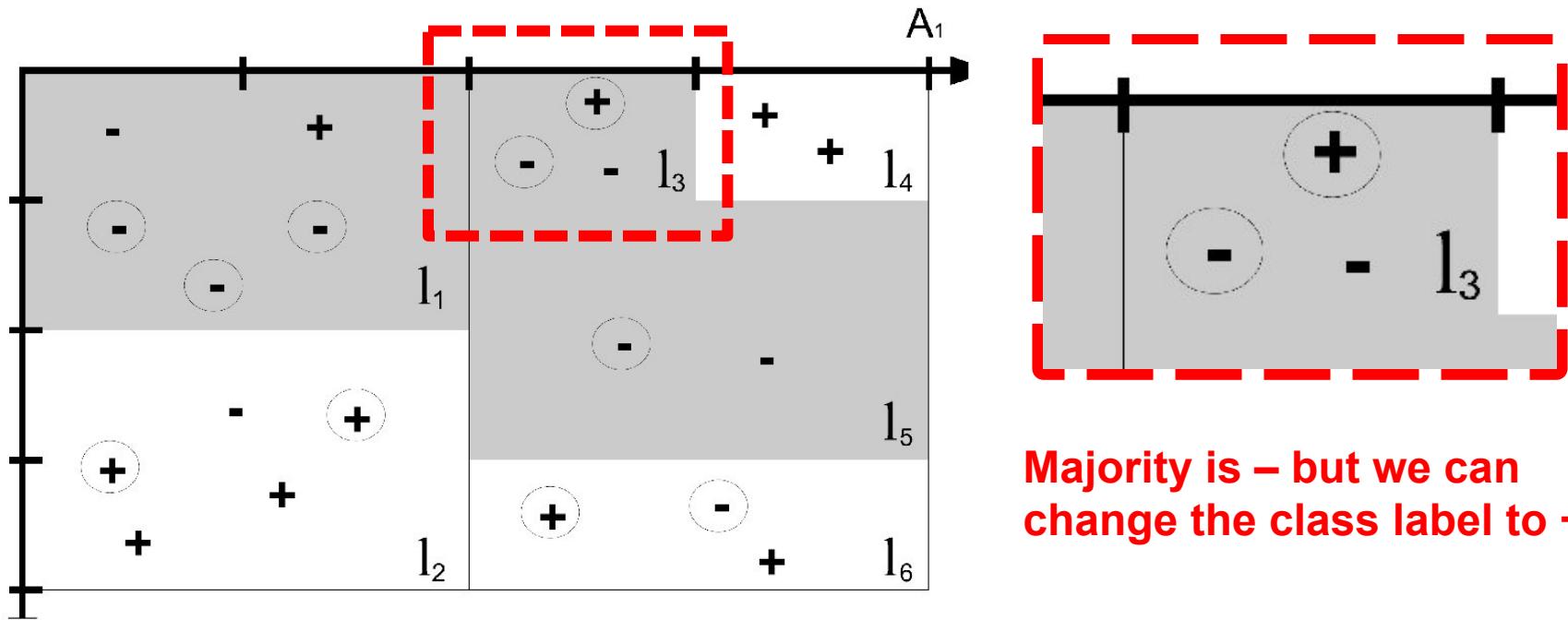
$$\Delta disc_l = \frac{u + v}{b} - \frac{w + x}{\bar{b}}$$

- If  $p < n$ , the label of the leaf changes from - to + and the effect on accuracy and discrimination is expressed by:

$$\Delta acc_l = p - n$$

$$\Delta disc_l = -\frac{u + v}{b} + \frac{w + x}{\bar{b}}$$

# Example



# Effect of Relabeling on Discrimination and Accuracy

Dataset		
Class →	–	+
Pred. →	–/+	–/+
$B = 1$	$U_1/U_2$	$V_1/V_2$
$B = 0$	$W_1/W_2$	$X_1/X_2$
	$N_1/N_2$	$P_1/P_2$
		1

Leaf 1		
	–	+
$B = 1$	$u$	$v$
$B = 0$	$w$	$x$
	$n$	$p$
		a

$$\Delta acc_l = p - n$$

$$\Delta disc_l = -\frac{u+v}{b} + \frac{w+x}{\bar{b}}$$

Dataset		
Class →	–	+
Pred. →	–/+	–/+
$B = 1$	$\frac{5}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{3}{20}$
$B = 0$	$\frac{3}{20}/\frac{1}{20}$	$\frac{1}{20}/\frac{5}{20}$
	$\frac{8}{20}/\frac{2}{20}$	$\frac{2}{20}/\frac{8}{20}$
		1

Leaf $l_3$			
	–	+	
$B = 1$	$\frac{1}{20}$	$\frac{1}{20}$	$\frac{2}{20}$
$B = 0$	$\frac{1}{20}$	0	$\frac{1}{20}$
	$\frac{2}{20}$	$\frac{1}{20}$	$\frac{3}{20}$

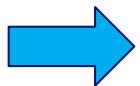
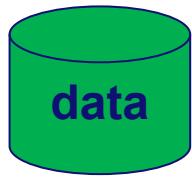
$n > p$

$$\text{Effect on accuracy} = \frac{1}{20} - \frac{2}{20} = -\frac{1}{20}$$

$$\text{Effect on discrimination} = -\frac{\frac{1}{20} + \frac{1}{20}}{\frac{1}{2}} + \frac{\frac{1}{20} + 0}{\frac{1}{2}} = -\frac{1}{10}$$

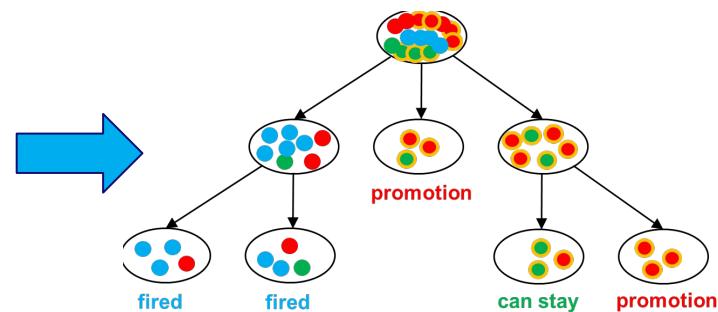
But the most challenging part of this solution is to select exactly the set of leaves (for relabeling) that is optimal with respect to reducing the discrimination with minimal loss in accuracy, and this problem is NP-complete.

# Summary



“massage”  
input data

change DT  
learning



post-process  
DT

# Conclusion



# Summary

- **Responsible Data Science**
- **FACT: Fairness, Accuracy, Confidentiality and Transparency.**
- Today's focus on Fairness.
- How to qualify Fairness?
- How to intervene (before, during, or after learning a classifier)?
- Next lecture focus on Confidentiality.

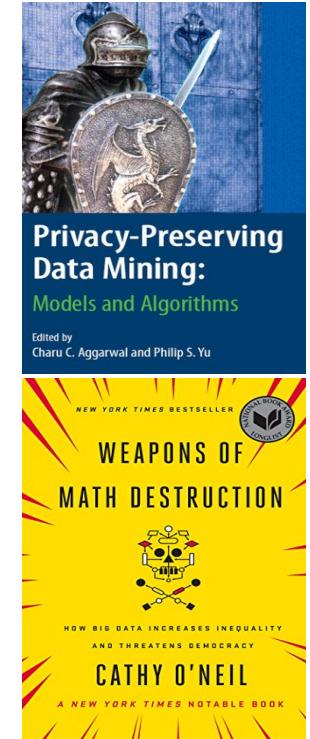


**Making Data Science Green!**



# Relevant Literature

- van der Aalst, Wil MP. "Responsible data science: using event data in a “people friendly” manner." *International Conference on Enterprise Information Systems*. Springer, Cham, 2016.
- Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini. "Discrimination-aware data mining." *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008.
- Kamiran, Faisal, Toon Calders, and Mykola Pechenizkiy. "Discrimination aware decision tree learning." *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010.
- van der Aalst, Wil MP. "Responsible data science." *Business and Information Systems Engineering*. Springer, Cham, 2017.
- Kashid, Asmita, Vrushali Kulkarni, and Ruhi Patankar. "Discrimination-aware data mining: a survey." *International Journal of Data Science* 2.1 (2017): 70-84.
- Sweeney, Latanya. "k-anonymity: A model for protecting privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002): 557-570.
- Machanavajjhala, Ashwin, et al. "\ell-Diversity: Privacy Beyond k-Anonymity." *null*. IEEE, 2006.
- Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." *Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on*. IEEE, 2007.
- Aggarwal, Charu C., and S. Yu Philip. "A general survey of privacy-preserving data mining models and algorithms." *Privacy-preserving data mining*. Springer, Boston, MA, 2008. 11-52.
- Agrawal, Rakesh, and Ramakrishnan Srikant. *Privacy-preserving data mining*. Vol. 29. No. 2. ACM, 2000.
- Saygin, Yücel, Vassilios S. Verykios, and Chris Clifton. "Using unknowns to prevent discovery of association rules." *ACM Sigmod Record* 30.4 (2001): 45-54.
- Evfimievski, Alexandre, et al. "Privacy preserving mining of association rules." *Information Systems* 29.4 (2004): 343-364.
- Oliveira, Stanley RM, Osmar R. Zaiane, and Yücel Saygin. "Secure association rule sharing." *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Berlin, Heidelberg, 2004.
- Lindell, Yehuda, and Benny Pinkas. "Privacy preserving data mining." *Annual International Cryptology Conference*. Springer, Berlin, Heidelberg, 2000.
- Verykios, Vassilios S., et al. "State-of-the-art in privacy preserving data mining." *ACM Sigmod Record* 33.1 (2004): 50-57.



#	Lecture	date	day
	<b>Lecture 1</b> Introduction	10/10/2018	Wednesday
	<b>Lecture 2</b> Crash Course in Python	11/10/2018	Thursday
Instruction 1	Python	12/10/2018	Friday
	<b>Lecture 3</b> Basic data visualisation/exploration	<b>Lecture 20</b>	Responsible data science (1/2)
	<b>Lecture 4</b> Decision trees		Responsible data science (2/2)
Instruction 2	Decision trees and data visualization	<b>Lecture 21</b>	Responsible data science
	Regression		11/01/2019 Friday
	Support vector machines	<b>Lecture 22</b>	16/01/2019 Wednesday
Instruction 3	Regression and support vector machines		17/01/2019 Thursday
	<b>Lecture 7</b> Neural networks (1/2)	<b>Lecture 23</b>	18/01/2019 Friday
Instruction 4	Neural networks and supervised learning		23/01/2019 Wednesday
	<b>Lecture 8</b> Neural networks (2/2)	<b>Lecture 24</b>	24/01/2019 Thursday
	Evaluation of supervised learning problems		25/01/2018 Friday
Instruction 5	Neural networks and supervised learning	backup	30/01/2019 Wednesday
	<b>Lecture 10</b> Clustering	<b>Instruction 12</b>	31/01/2019 Thursday
	Frequent items sets		01/02/2019 Friday
	Association rules	backup	02/02/2019 Saturday
Instruction 6	Sequence mining	backup	03/02/2019 Sunday
	<b>Lecture 14</b> Process mining (unsupervised)	<b>extra</b>	04/02/2019 Monday
	<b>Lecture 15</b> Process mining (supervised)		05/02/2019 Tuesday
Instruction 7	Process mining and sequence mining	30/11/2018	Friday
	<b>Lecture 16</b> Text mining (1/2)	05/12/2018	Wednesday
Instruction 8	Text mining and process mining	06/12/2018	Thursday !!
	<b>Lecture 17</b> Text mining (2/2)	12/12/2018	Wednesday
	<b>Lecture 18</b> Data preprocessing, data quality, binning, etc.	13/12/2018	Thursday
	<b>Lecture 19</b> Visual analytics & information visualization	19/12/2018	Wednesday
	backup	20/12/2018	Thursday
Instruction 9	Text mining, preprocessing and visualization	21/12/2018	Friday
	<b>Lecture 20</b> Responsible data science (1/2)	09/01/2019	Wednesday
	<b>Lecture 21</b> Responsible data science (2/2)	10/01/2019	Thursday
Instruction 10	Responsible data science	11/01/2019	Friday
	<b>Lecture 22</b> Big data (1/2)	16/01/2019	Wednesday
	<b>Lecture 23</b> Big data (2/2)	17/01/2019	Thursday
Instruction 11	Big data	18/01/2019	Friday
	<b>Lecture 24</b> Closing	23/01/2019	Wednesday
	backup	24/01/2019	Thursday
Instruction 12	Example exam questions	25/01/2018	Friday
	backup	30/01/2019	Wednesday
	backup	31/01/2019	Thursday
	extra	Question hour	01/02/2019 Friday