

Elements of Machine Learning & Data Science

Responsible Data Science

Lecture 22

Prof. Wil van der Aalst

Marco Pegoraro, M.Sc.

Nina Graves, M.Sc.

Part I: Introduction to RDS

Fairness, accuracy, confidentiality, transparency

Part II: Confidentiality

Risks, encryption, anonymization, quasi-identifiers, K-Anonymity, L-Diversity, and T-Closeness

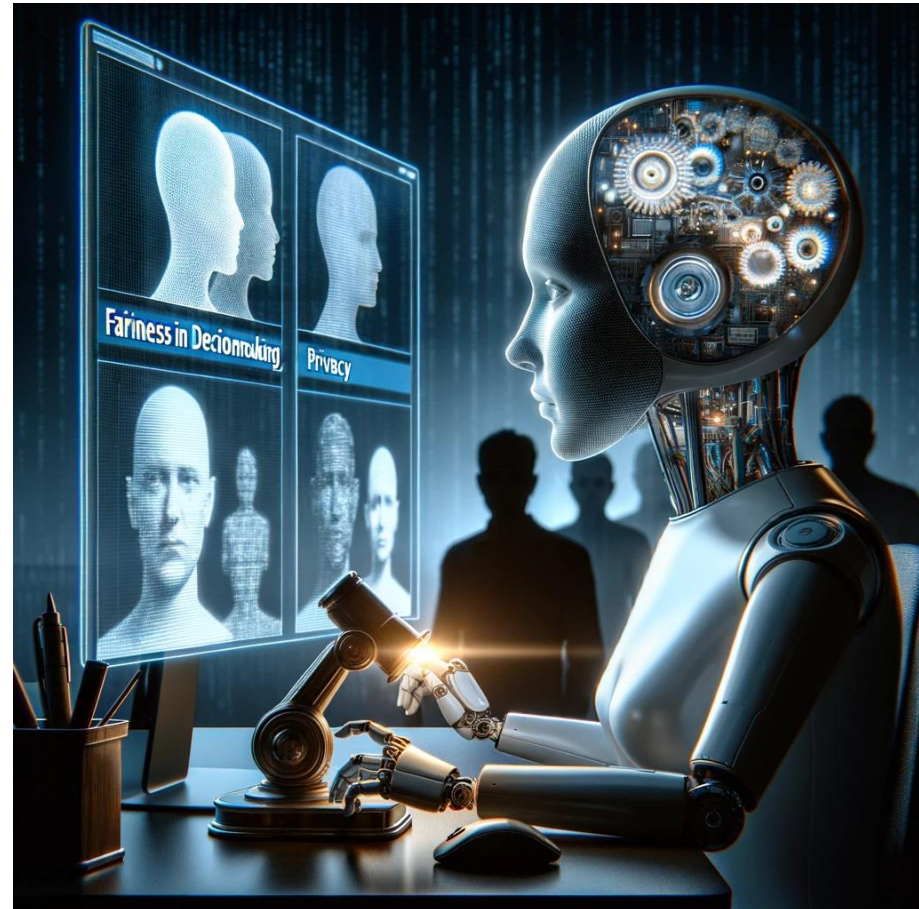
Part III: Fairness

Fairness measures, itemsets/association rules revisited, effect (rule/outcome), making decision trees fair

Part I: Introduction to Responsible Data Science

Responsible Data Science

- **F**airness
- **A**ccuracy
- **C**onfidentiality
- **T**ransparency



Fairness – Data Science Without Prejudice

How to **avoid unfair conclusions** even if they are true?



W. Tingey

Fairness – Data Science Without Prejudice



Banking



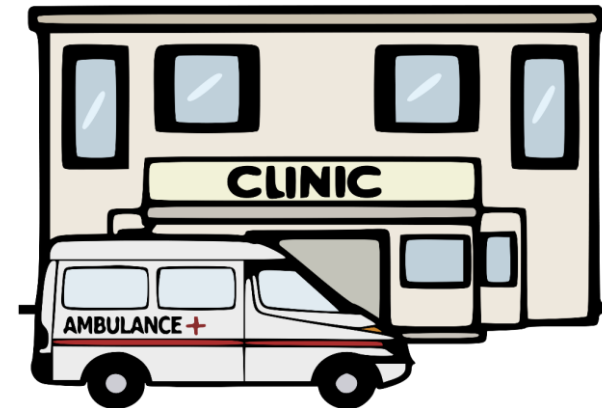
Insurance



Hiring



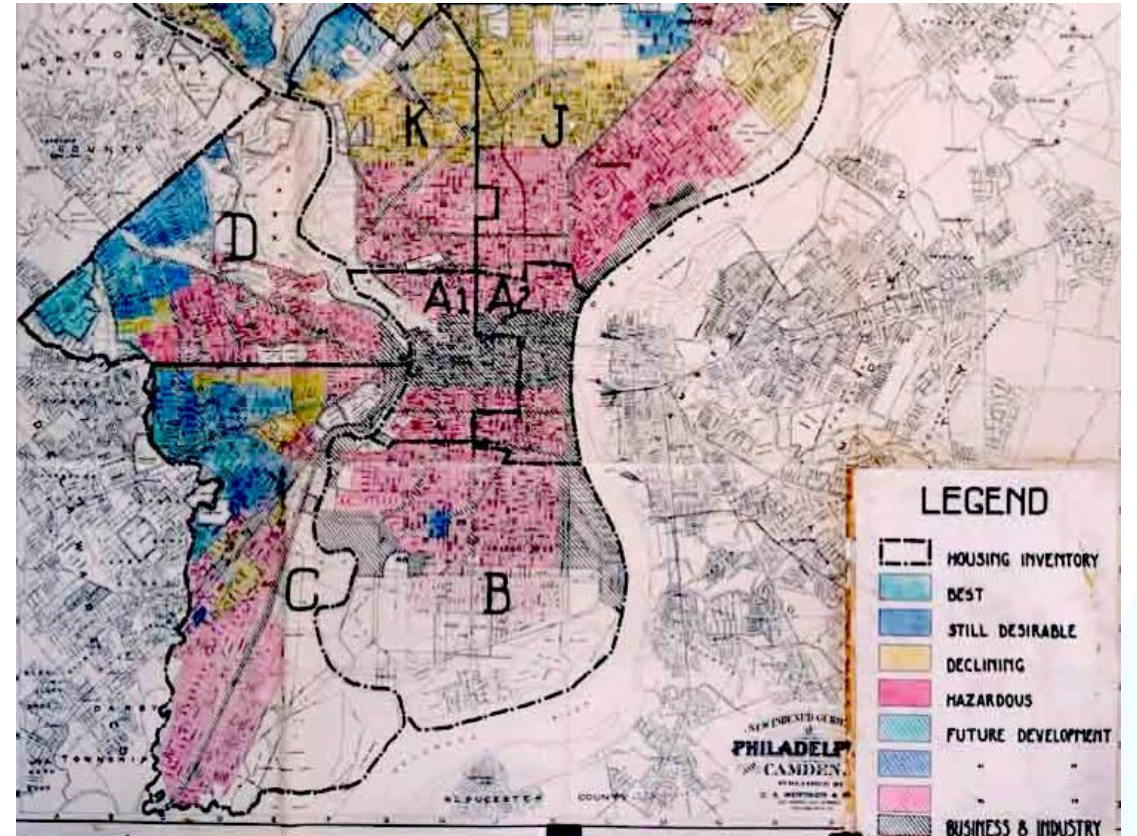
Admission



Health

Not New – Redlining

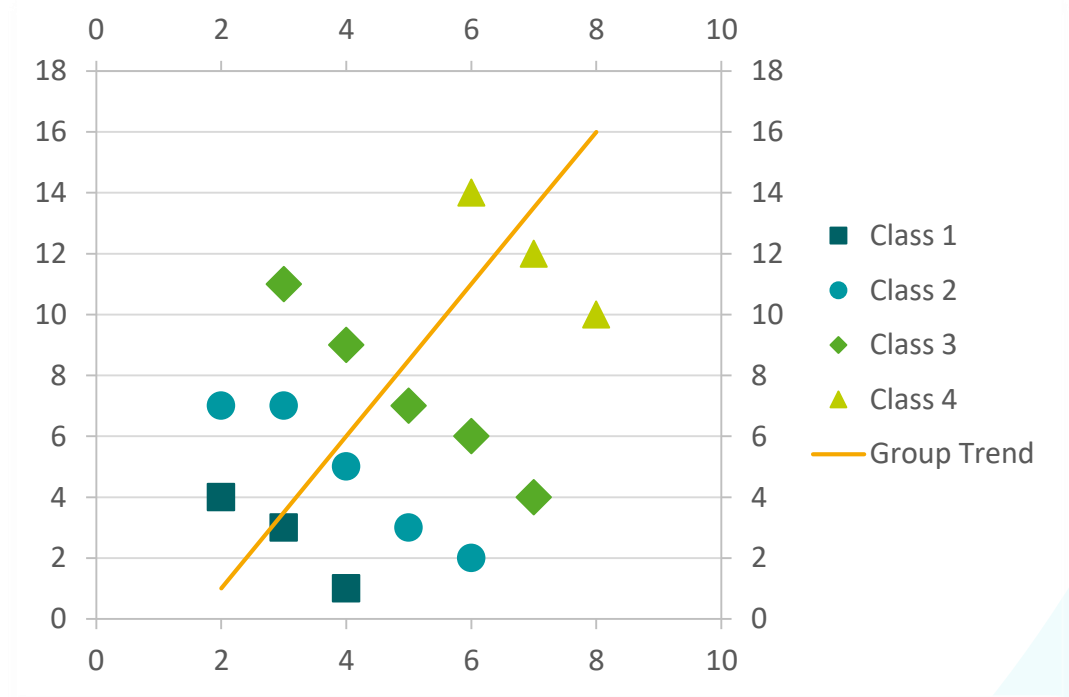
- **Redlining**, a discriminatory practice of denying affordable services based on geographical locations
- You can remove race, gender, age, etc., but if this correlates with your zip code...



1936 security map of Philadelphia showing **redlining of lower income neighborhoods** (estimated risk of mortgage loans)

Remember – Simpson's Paradox

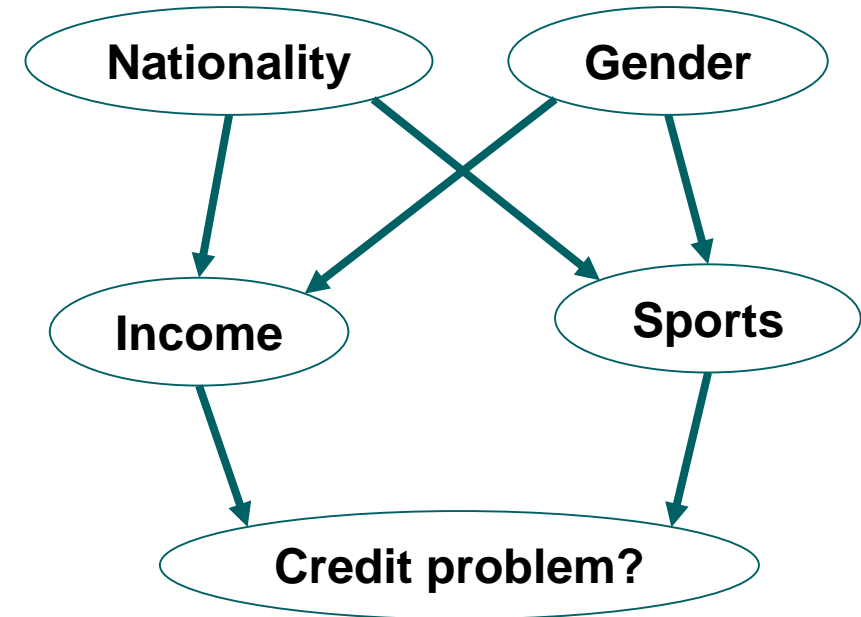
A trend appears in several different groups of data but **disappears** or **reverses** when these groups are combined.



Fairness – How to Avoid Unfair Conclusions?

Even if they are true...

- Removing sensitive features often does not work!
- Enforcing fairness may lead to **less accurate predictions**
- Not so easy to define fairness
 - **Percentage of patients dying** – academic vs regional hospitals, experienced vs inexperienced doctors, etc.
 - **Waiting times for a resource** – busy vs idle resources, part-time vs full-time, etc.



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 the sample may **not** be **representative** (never enough evidence for some groups)
- Even when the data are correct, optimizing for a particular target feature may lead to discrimination (**accuracy for old cases does not imply fairness for new cases !**)

Accuracy – Data Science Without Guesswork

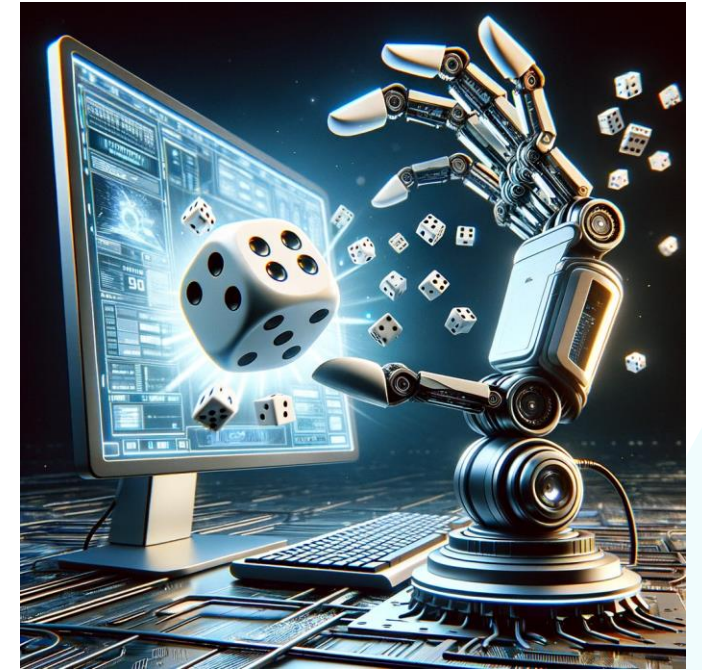
How to answer questions with a **guaranteed level of accuracy**?



See “Computer says no” episode of Little Britain from 2004.
Contrastingly, “ML never says no”.

Accuracy

- ML algorithms **always** return a result, but:
 - The instance may be **close to a decision boundary**
 - There may be **too little training data**
- When testing many null hypotheses, just **by chance**, one will be rejected. **Carlo Emilio Bonferroni** already indicated that one needs to correct for this in 1936.



Generated using DALL·E 3

Accuracy – The Curse of Dimensionality

Find the terrorists



Assumptions:

- 18 million people in NL
- 1800 hotels
- 100 guests per hotel per night
- Hence, on average a person visits a hotel every 100 days

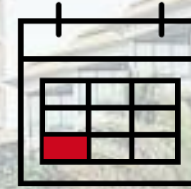
Suspicious event - two persons stay in the same hotel on two different dates

How many suspicious events in a 1000-day period (i.e., less than 3 years)?



Accuracy – Curse of Dimensionality

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



- The probability that two persons $p1$ and $p2$ visit a hotel on a given day (d): $\frac{1}{100} \times \frac{1}{100} = 10^{-4}$
- The probability that $p1$ and $p2$ visit the same hotel on the day (d): $10^{-4} \times \frac{1}{1800} = 5.55 * 10^{-8}$
- The probability that $p1$ and $p2$ visit the same hotel on two different dates: $(5.55 \times 10^{-8})^2$
-

Probability is 0.0000000000000000003086!

Accuracy – Curse of Dimensionality

- The probability that two persons $p1$ and $p2$ visit a hotel on a given day (d): $\frac{1}{100} \times \frac{1}{100} = 10^{-4}$
- The probability that $p1$ and $p2$ visit the same hotel on the day: $10^{-4} \times \frac{1}{1800} = 5.55 \times 10^{-8}$
- The probability that $p1$ and $p2$ visit the same hotel on two different dates $d1$ and $d2$: $(5.55 \times 10^{-8})^2 =$
0.0000000000000000003086

But how many suspicious events in a 1000-day period?

- Number of candidate events $(\{d1,d2\},\{p1,p2\})$: $\binom{1000}{2} \times \binom{18 \times 10^6}{2} = 8.09 \times 10^{19}$
- Hence, the expected number of suspicious events is equal to $(5.55 \times 10^{-8})^2 \times 8.09 \times 10^{19}$
-

These are 249.750 events!

Curse of Dimensionality

- When looking for **patterns** in data (e.g., correlations) and the number of **possible patterns** is as large as the number of data points you have, then, by chance, some of these patterns **will be found!**
- In statistics, the **Bonferroni correction** is a method to counteract the “multiple comparisons problem”.
- Consider **statistical hypothesis testing**, which is based on **rejecting the null hypothesis if the likelihood of the observed data under the null hypothesis** is low (e.g., p-value is below 0.05).
- If many hypotheses are tested, then the probability that the null hypothesis is rejected ($p < 0.05$) increases.
- In other words, **when the number of potential patterns is large compared to the number of instances, then you will find these patterns in the training data.**
- Related to **overfitting**: If the number of weights in the neural network is larger than the number of instances, then one will perfectly fit any training data. However, this does not mean anything.

Confidentiality – Data Science That Ensures Confidentiality

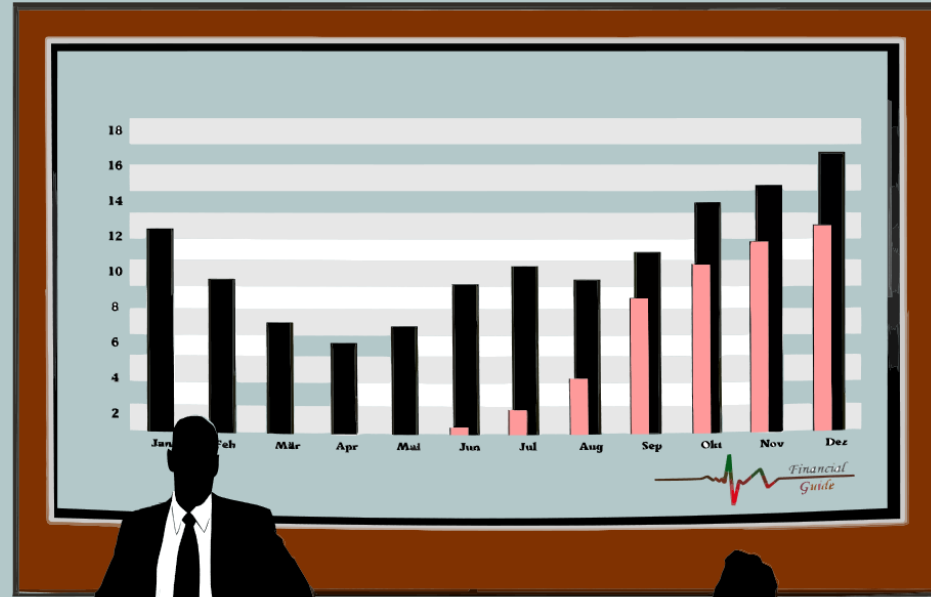


How to answer questions **without revealing secrets?**

If You are Not Paying, You Are the Product!

Social media

Search engines



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

General Data Protection Regulation (GDPR)

- **Communication** – explain why user should leave personal information
- **Consent** – get clear consent to the processing of personal data
- **Data Transfer Outside the EU** – only when adequate level of protection is guaranteed
- **Sensitive Data** – ensure specific safety for sensitive data like race
- **Access** – users should have access to their information
- **Profiling** – individuals have the right to appeal against the decisions when it is based on automated processing
- **Erase Data** – users can request to delete data
- **Warnings** – companies have to notify authorities about data breaches
- **Marketing** – people should be able to give up direct marketing that uses their data



Another AI Regulation: Artificial Intelligence Act (AI Act)

(not yet enforced, more related to fairness)



- It aims to **classify** and **regulate** artificial intelligence applications based on their **risk of causing harm**.
- The classification includes four categories of risk ("unacceptable", "high", "limited" and "minimal") plus one additional category for general-purpose AI (e.g., foundation models like GPT).

unacceptable	prohibited	Social scoring, biometric identification and categorization of people, real-time and remote biometric identification systems, such as facial recognition, etc.
high	conformity assessment	Employment, HRM, education, critical infrastructure, law enforcement, etc.
limited	transparency obligation	Chatbots, deepfakes, etc.
minimal	no restrictions	Spam filters, video games and inventory-management systems, etc.

Transparency – Data Science That Provides Transparency

How to clarify answers such that they become indisputable?

- How was the prediction / decision made?
- What happened in the data pipeline?
- Do people understand the result?
- Can the result be explained?



Generated using DALL·E 3

data selection

visualization

data cleaning

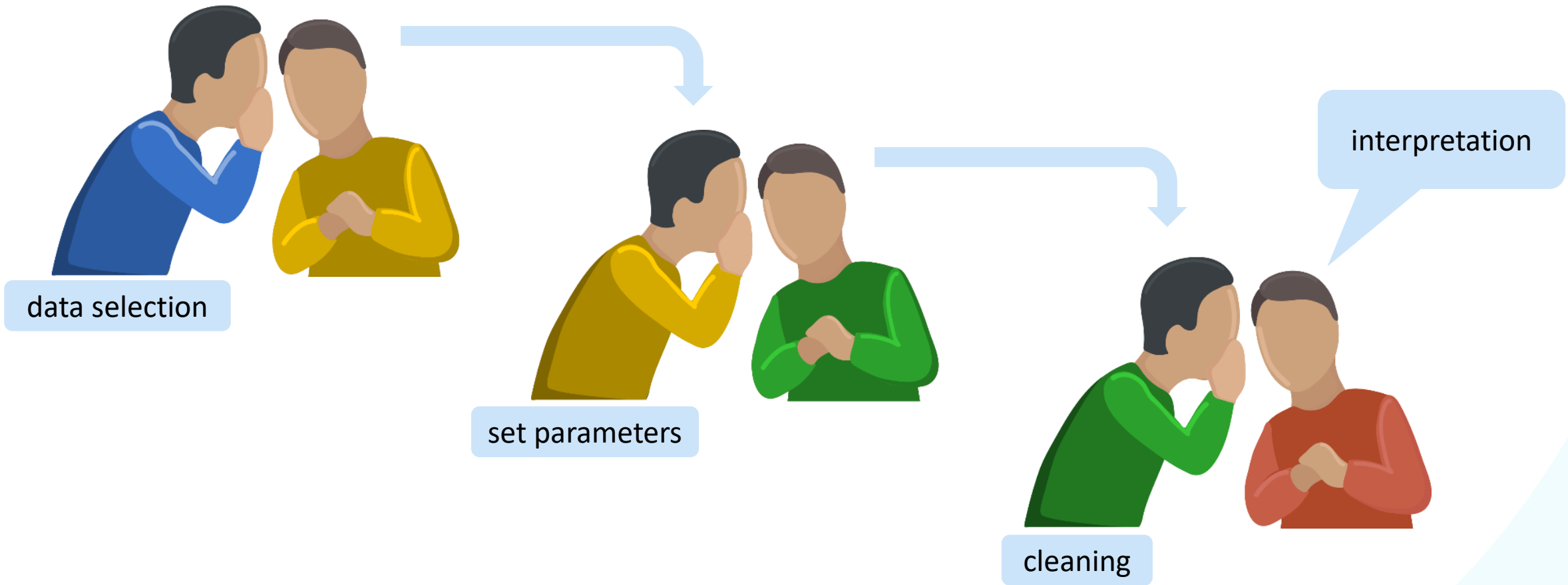
parameter setting

representational bias

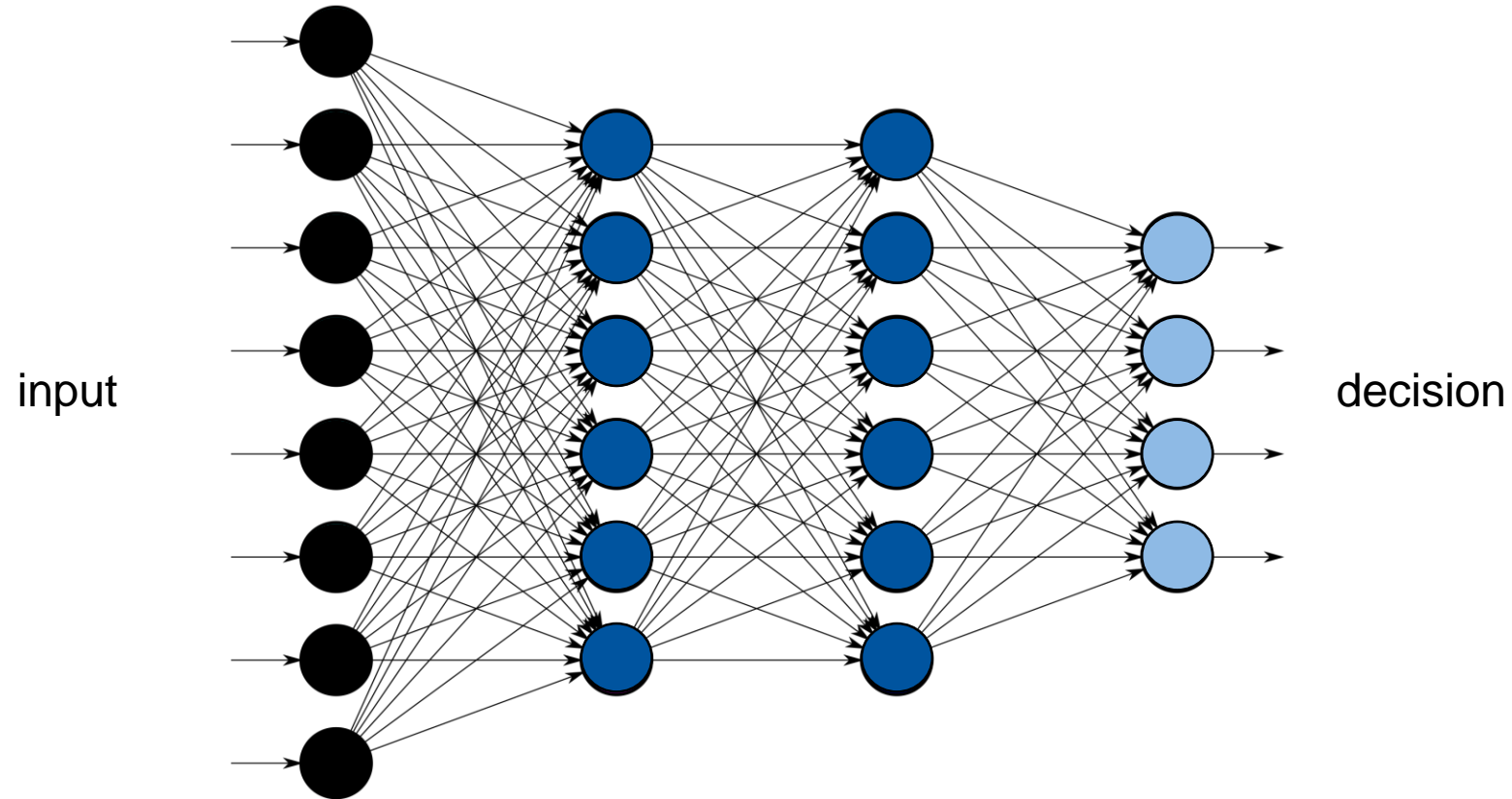
testing hypothesis

storytelling





You Are Guilty, Not Selected, Not Treated, ...

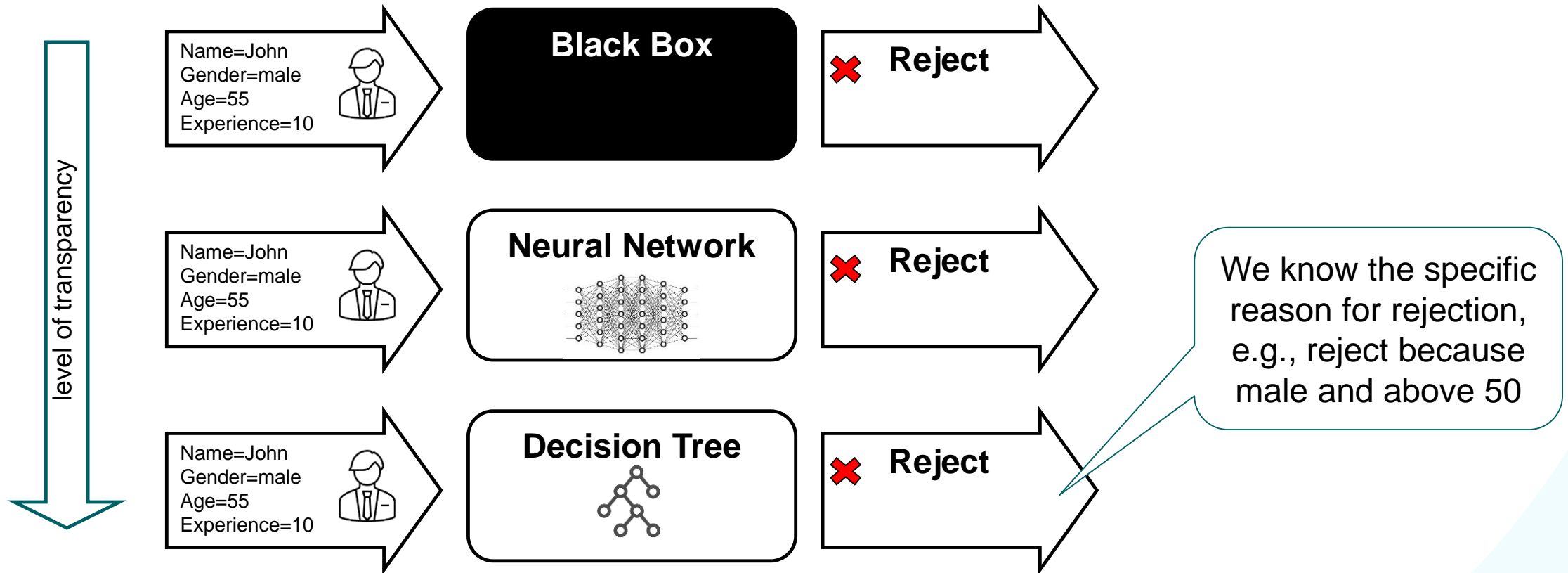


... but we do not know why.



Generated using DALL·E 3

Transparency



Summary

- **Fairness**
 - Ensure data-driven decision-making does not amplify existing biases or discrimination.
- **Accuracy**
 - Aim for data-driven models that make correct predictions.
- **Confidentiality**
 - Protect personal and private information throughout the data lifecycle.
- **Transparency**
 - Provide clear and understandable explanations

Part II: Confidentiality

Responsible Data Science (Confidentiality)

1. **Confidentiality Risks**
2. Using Encryption to Ensure Confidentiality
3. Anonymization Operations
4. K-Anonymity
5. L-Diversity and T-Closeness



4 Types of Features

Name	Age	Gender	ZIP-code	Job	Disease
Smith	27	Male	47577	Engineer	Hepatitis
Johnson	32	Male	47602	Dancer	Hepatitis
Williams	19	Female	47578	Writer	Hepatitis
Brown	55	Male	47905	Engineer	HIV
Jones	31	Male	47609	Dancer	HIV
Garcia	38	Female	47606	Lawyer	HIV
Davis	23	Female	47505	Lawyer	Heart
Martinez	47	Female	47973	Writer	Heart
Taylor	60	Female	47907	Engineer	Heart
Anderson	29	Male	47505	Dancer	Heart

4 Types of Features – Explicit Identifier

An **explicit identifier** is a set of features containing information that explicitly identifies the instance owner

Name	Age	Gender	ZIP-code	Job	Disease
Smith	27	Male	47677	Engineer	Hepatitis
Johnson	42	Male	47502	Dancer	Hepatitis
Williams	19	Female	47678	Writer	Hepatitis
Brown	55	Male	47905	Engineer	HIV
Jones	31	Male	47909	Dancer	HIV
Garcia	38	Female	47906	Lawyer	HIV
Davis	23	Female	47605	Lawyer	Heart
Martinez	47	Female	47673	Writer	Heart
Taylor	60	Female	47507	Engineer	Heart
Anderson	29	Male	47505	Dancer	Heart

4 Types of Features – Quasi-Identifiers

A **quasi-identifier** is a set of features containing information that potentially identifies the instance owner

Name	Age	Gender	ZIP-code	Job	Disease
Smith	27	Male	47677	Engineer	Hepatitis
Johnson	42	Male	47502	Dancer	Hepatitis
Williams	19	Female	47678	Writer	Hepatitis
Brown	55	Male	47905	Engineer	HIV
Jones	31	Male	47909	Dancer	HIV
Garcia	38	Female	47906	Lawyer	HIV
Davis	23	Female	47605	Lawyer	Heart
Martinez	47	Female	47673	Writer	Heart
Taylor	60	Female	47507	Engineer	Heart
Anderson	29	Male	47505	Dancer	Heart

4 Types of Features – Sensitive Features

Sensitive features are sensitive person-specific information about the instance owner

Name	Age	Gender	ZIP-code	Job	Disease
Smith	27	Male	47677	Engineer	Hepatitis
Johnson	42	Male	47502	Dancer	Hepatitis
Williams	19	Female	47678	Writer	Hepatitis
Brown	55	Male	47905	Engineer	HIV
Jones	31	Male	47909	Dancer	HIV
Garcia	38	Female	47906	Lawyer	HIV
Davis	23	Female	47605	Lawyer	Heart
Martinez	47	Female	47673	Writer	Heart
Taylor	60	Female	47507	Engineer	Heart
Anderson	29	Male	47505	Dancer	Heart

4 Types of Features

**explicit
identifier**

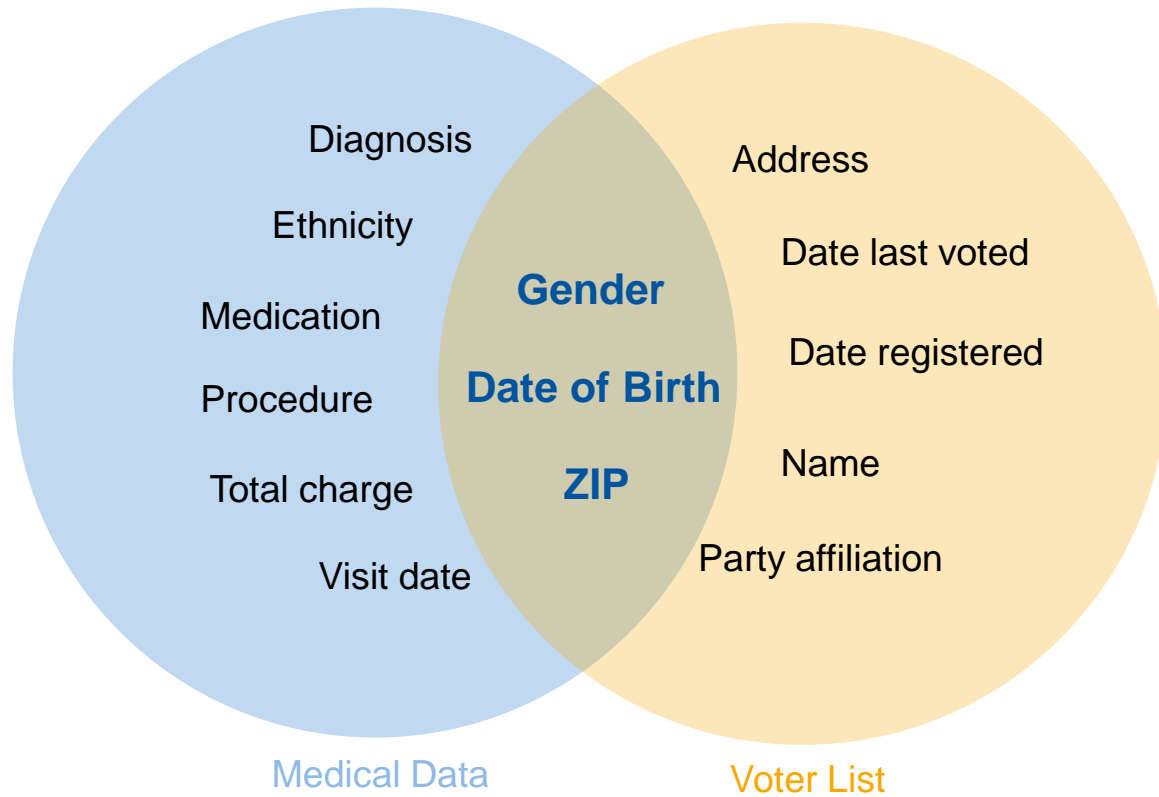
quasi-identifiers

**sensitive
feature**

Name	Age	Gender	ZIP-code	Job	Disease
Smith	27	Male	47577	Engineer	Hepatitis
Johnson	32	Male	47602	Dancer	Hepatitis
Williams	19	Female	47578	Writer	Hepatitis
Brown	55	Male	47905	Engineer	HIV
Jones	31	Male	47609	Dancer	HIV
Garcia	38	Female	47606	Lawyer	HIV
Davis	23	Female	47505	Lawyer	Heart
Martinez	47	Female	47973	Writer	Heart
Taylor	60	Female	47907	Engineer	Heart
Anderson	29	Male	47505	Dancer	Heart

other features (left out)

Problem – Example



87% of the U.S. population had reported characteristics that made them unique based on only such quasi-identifiers

Risks Related to Confidentiality and Privacy

Name	Age	Gender	ZIP-code	Job	Disease
	27	Male	47577	Engineer	Hepatitis
	32	Male	47602	Dancer	Hepatitis
	19	Female	47578	Writer	Hepatitis
	55	Male	47905	Engineer	HIV
	31	Male	47609	Dancer	HIV
	38	Female	47606	Lawyer	HIV
	23	Female	47505	Lawyer	Heart
	47	Female	47973	Writer	Heart
	60	Female	47907	Engineer	Heart
	29	Male	47505	Dancer	Heart



I know my 38 year old female employee is in the data set, what disease does she have?

I know my friend is in the data set, and she started in May 2023, what is her salary?

Responsible Data Science (Confidentiality)

1. Confidentiality Risks
2. **Using Encryption to Ensure Confidentiality**
3. Anonymization Operations
4. K-Anonymity
5. L-Diversity and T-Closeness



Cryptosystem

- **Cryptosystem:** Can be used to ensure **confidentiality** when **storing** or **exchanging sensitive data**
- There is a **wide variety** of cryptosystems:
 - *Symmetric* cryptosystem
 - *Asymmetric* cryptosystem
 - *Deterministic* cryptosystem
 - *Probabilistic* cryptosystem
 - *Homomorphic* cryptosystem
 - Etc

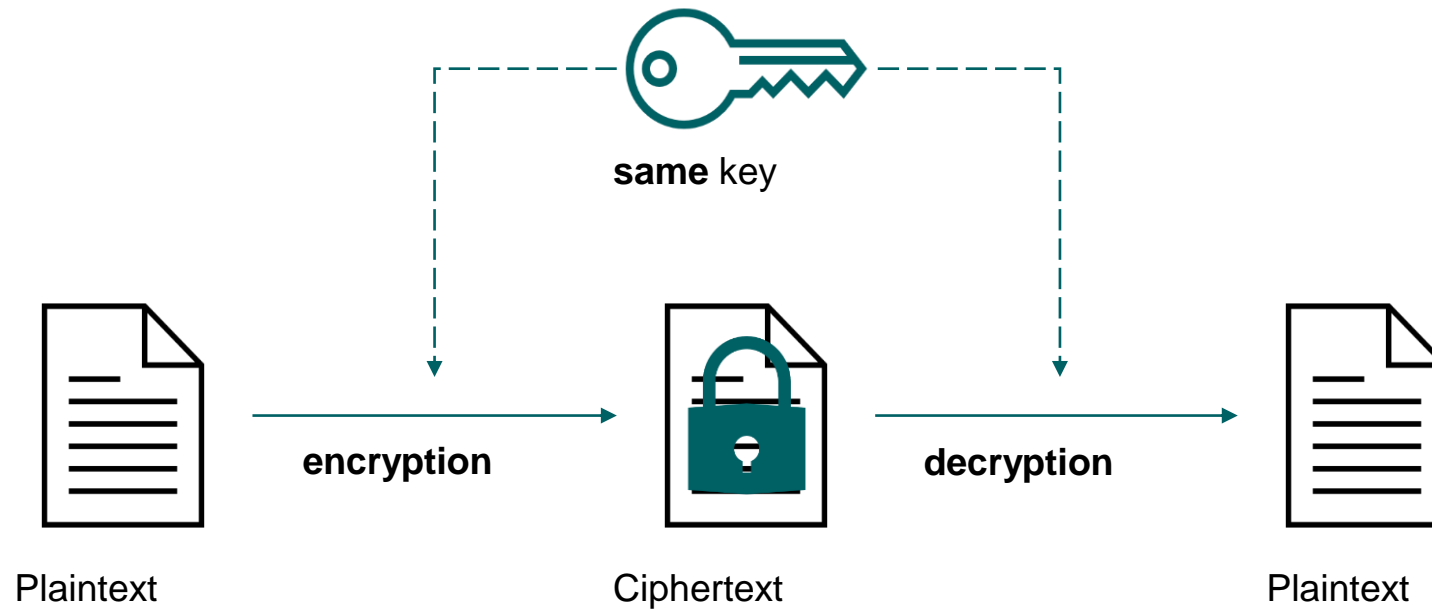
Name	Age	Gender	ZIP-code	Job	Disease
Smith	27	Male	47577	Engineer	Hepatitis
Johnson	32	Male	47602	Dancer	Hepatitis
Williams	19	Female	47578	Writer	Hepatitis
Brown	55	Male	47905	Engineer	HIV
Jones	31	Male	47609	Dancer	HIV
Garcia	38	Female	47606	Lawyer	HIV
Davis	23	Female	47505	Lawyer	Heart
Martinez	47	Female	47973	Writer	Heart
Taylor	60	Female	47907	Engineer	Heart
Anderson	29	Male	47505	Dancer	Heart



```
0110100101110011001000000110000100100000011001100111010101101100
0110110000100000011100000111001001101111011001100110010101110011
0111001101101111011100100010000001100001011101000010000001010010
010101110101010001001000001000000100000101100001011000110110100
0011001010110111000100000010101010110111001101001011101110110010
1011100100111001101101001011101000111100100101100001000000110110
00110010101100001011001000110100101101110 01100111 00100000 ...
```

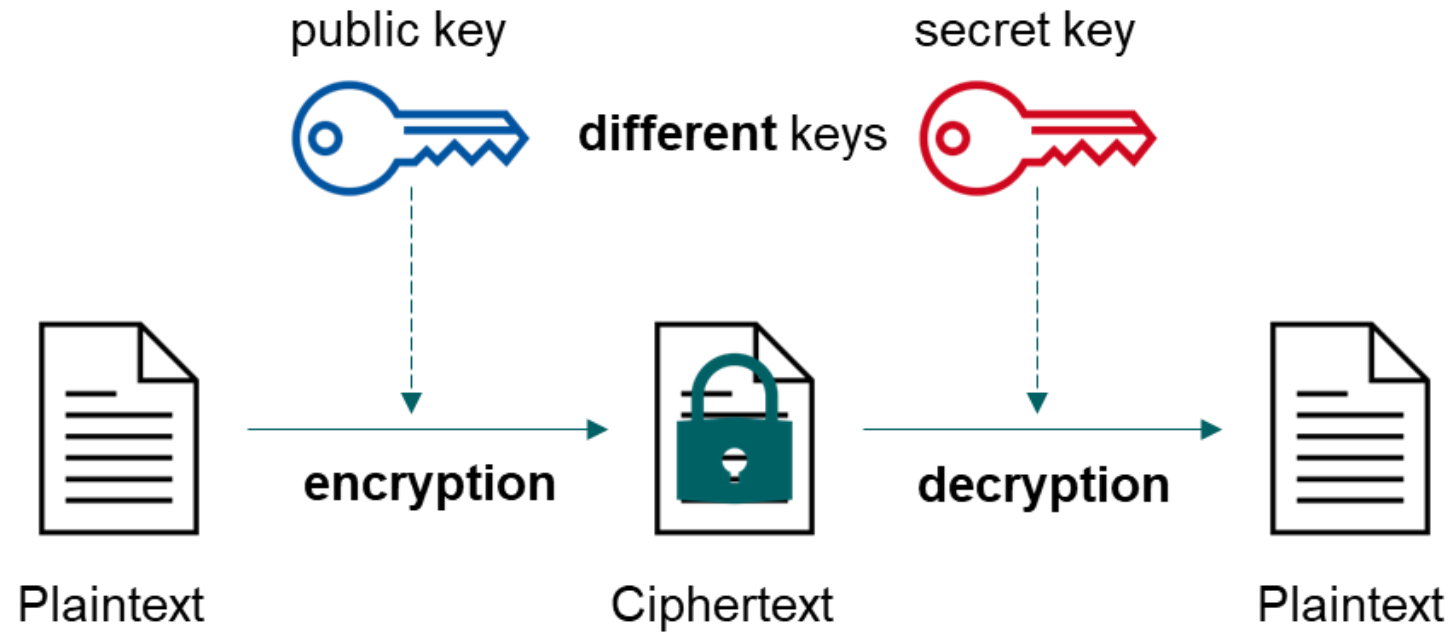
Symmetric Cryptosystem

e.g., AES



Asymmetric Cryptosystem

e.g., RSA

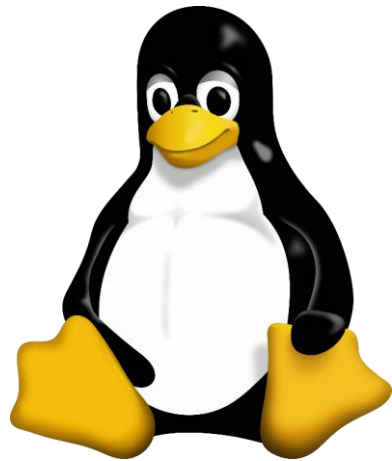


Deterministic Cryptosystem

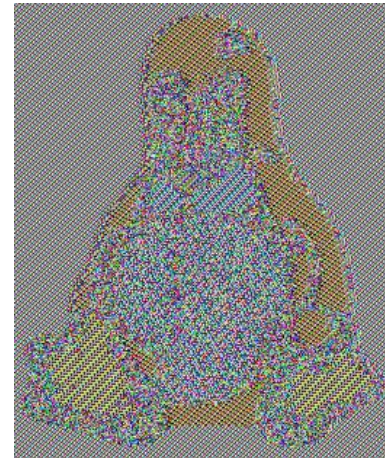
e.g., AES-ECB

A **deterministic** cryptosystem always produces the **same ciphertext** for a given plaintext and key (even over separate executions of the encryption algorithm)

- If we know that certain patterns (e.g., words, phrases, etc.) happen often, we can recognize them due to repeating patterns.
- Most common letters, e.g., “e” (13%) and “a” (8%), N-grams, or words (“the”, “of”, “and”, “to”, etc.)



Original image

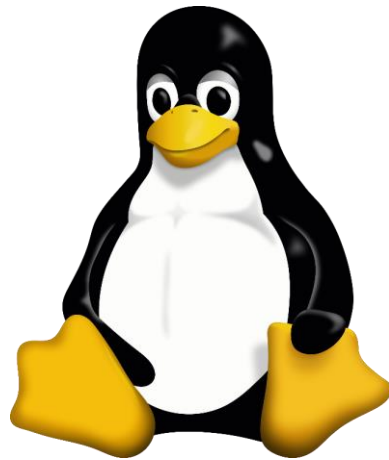


AES-ECB

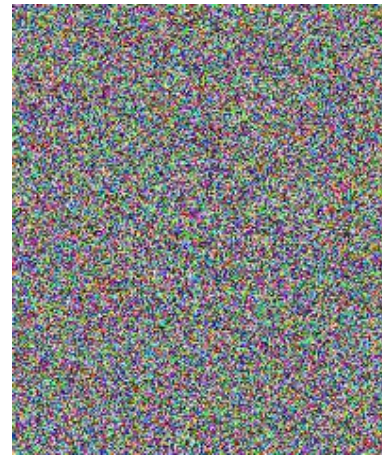
Probabilistic Cryptosystem

e.g., AES-CTR

A **probabilistic** cryptosystem as opposed to deterministic cryptosystem uses **randomness** in an encryption algorithm: when **encrypting the same plaintext** several times it produces **different ciphertexts**.



Original image

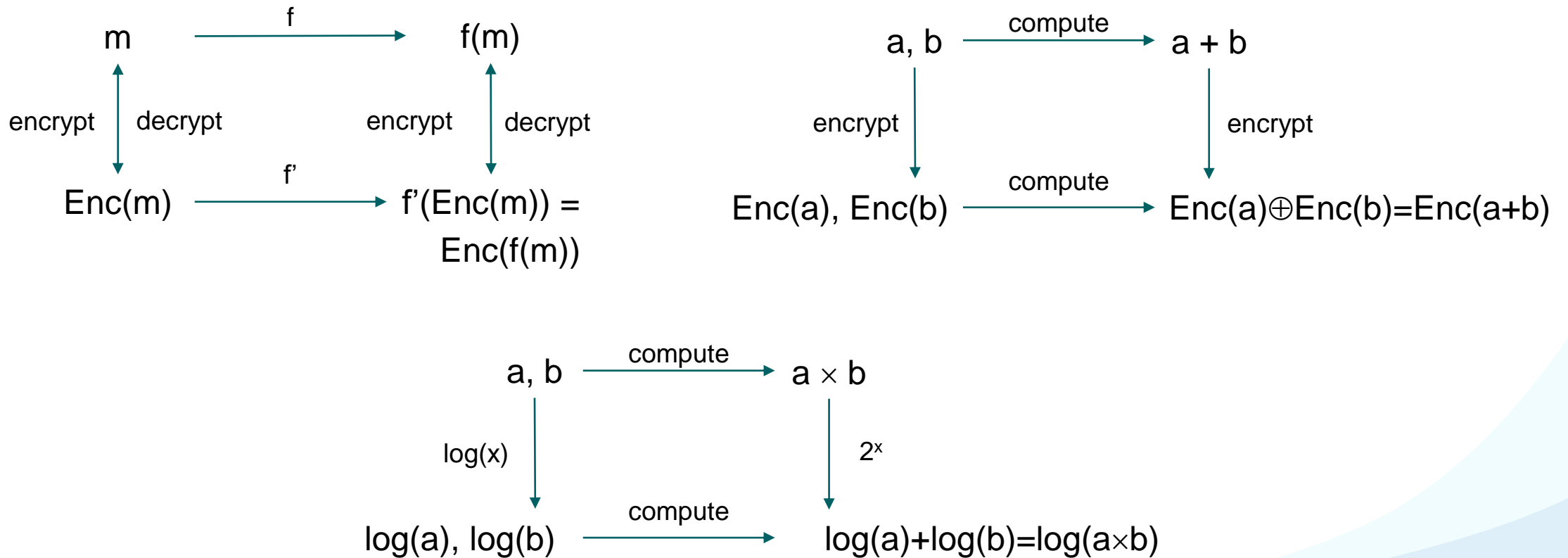


AES-CTR

Homomorphic Cryptosystem

e.g., Paillier

Homomorphic encryption allows computation on ciphertexts without decryption



Cryptosystem

- **Cryptosystem:** Protects sensitive data from unauthorized access when stored or transmitted
- Types of cryptosystems:
 - *Symmetric*
 - Keys are shared between parties
 - *Asymmetric*
 - Public and private keys
 - *Deterministic*
 - Always produces the same ciphertext
 - *Probabilistic*
 - Produces different ciphertexts (randomness)
 - *Homomorphic*
 - Computes on ciphertexts without decryption
 - ...



Responsible Data Science (Confidentiality)

1. Confidentiality Risks
2. Using Encryption to Ensure Confidentiality
3. **Anonymization Operations**
4. K-Anonymity
5. L-Diversity and T-Closeness



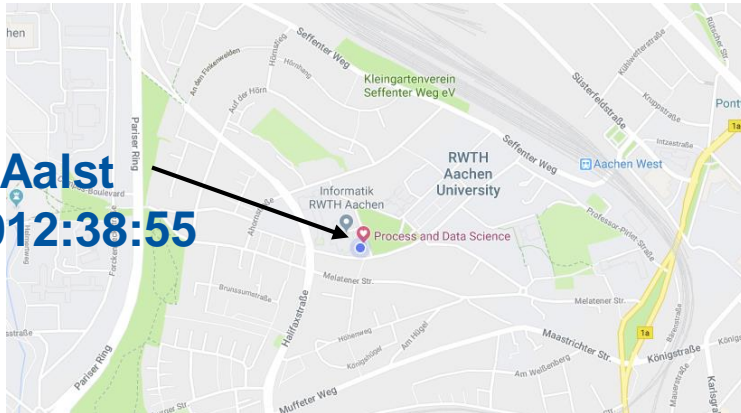
Anonymization Operations

- Provide privacy requirements:
Modify the data by applying a **sequence of anonymization operations**
- Anonymization operations:
 - Generalization
 - Suppression
 - Data swapping
 - Adding noise
 - Anatomization

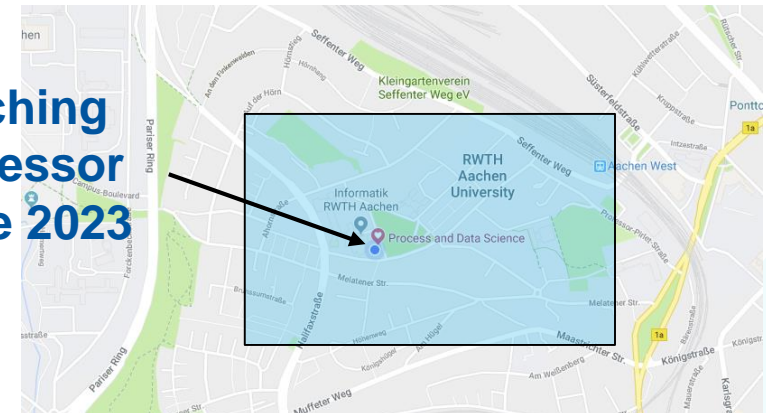
Generalization

- Reduce the **granularity** of representation to increase the privacy
- Can cause some loss of effectiveness of data management or mining algorithms

Lecture
Wil van der Aalst
23-6-2023T012:38:55

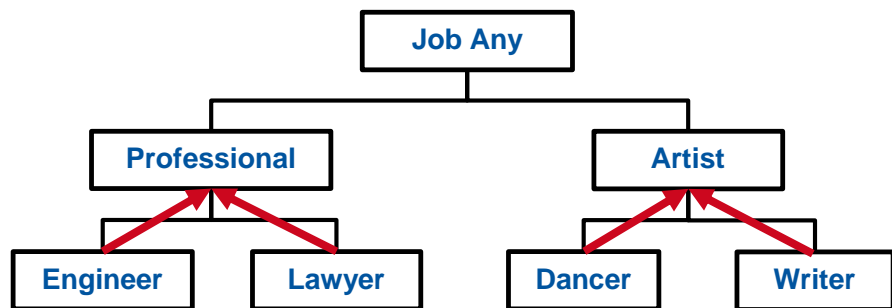


Teaching
Professor
June 2023



Generalization

Example: values of the job feature are generalized to a higher level of abstraction




Name	Gender	Job
Smith	Male	Engineer
Johnson	Male	Dancer
Williams	Female	Writer
Brown	Male	Engineer
Jones	Male	Dancer
Garcia	Female	Lawyer
Davis	Female	Lawyer
Martinez	Female	Writer
Taylor	Female	Engineer
Anderson	Male	Dancer

↷

Name	Gender	Job
Smith	Male	Professional
Johnson	Male	Artist
Williams	Female	Artist
Brown	Male	Professional
Jones	Male	Artist
Garcia	Female	Professional
Davis	Female	Professional
Martinez	Female	Artist
Taylor	Female	Professional
Anderson	Male	Artist

Suppression

Replace some **values with a placeholder value**, indicating that the replaced values are not disclosed




Name	Gender	Job	Disease
Smith	Male	Engineer	Hepatitis
Johnson	Male	Dancer	Hepatitis
Williams	Female	Writer	Hepatitis
Brown	Male	Engineer	HIV
Jones	Male	Dancer	HIV
Garcia	Female	Lawyer	HIV
Davis	Female	Lawyer	Heart
Martinez	Female	Writer	Heart
Taylor	Female	Engineer	Heart
Anderson	Male	Dancer	Heart

Name	Gender	Job	Disease
*	*	*	*
Johnson	Male	Dancer	Hepatitis
*	*	*	*
Brown	Male	Engineer	HIV
Jones	Male	Dancer	HIV
Garcia	Female	Lawyer	HIV
Davis	Female	Lawyer	Heart
Martinez	Female	Writer	Heart
Taylor	Female	Engineer	Heart
Anderson	Male	Dancer	Heart

Record suppression – refers to suppressing an entire instance (i.e., row)

Suppression

Replace some **values with a placeholder value**, indicating that the replaced values are not disclosed



Name	Gender	Job	Disease
Smith	Male	Engineer	Hepatitis
Johnson	Male	Dancer	Hepatitis
Williams	Female	Writer	Hepatitis
Brown	Male	Engineer	HIV
Jones	Male	Dancer	HIV
Garcia	Female	Lawyer	HIV
Davis	Female	Lawyer	Heart
Martinez	Female	Writer	Heart
Taylor	Female	Engineer	Heart
Anderson	Male	Dancer	Heart

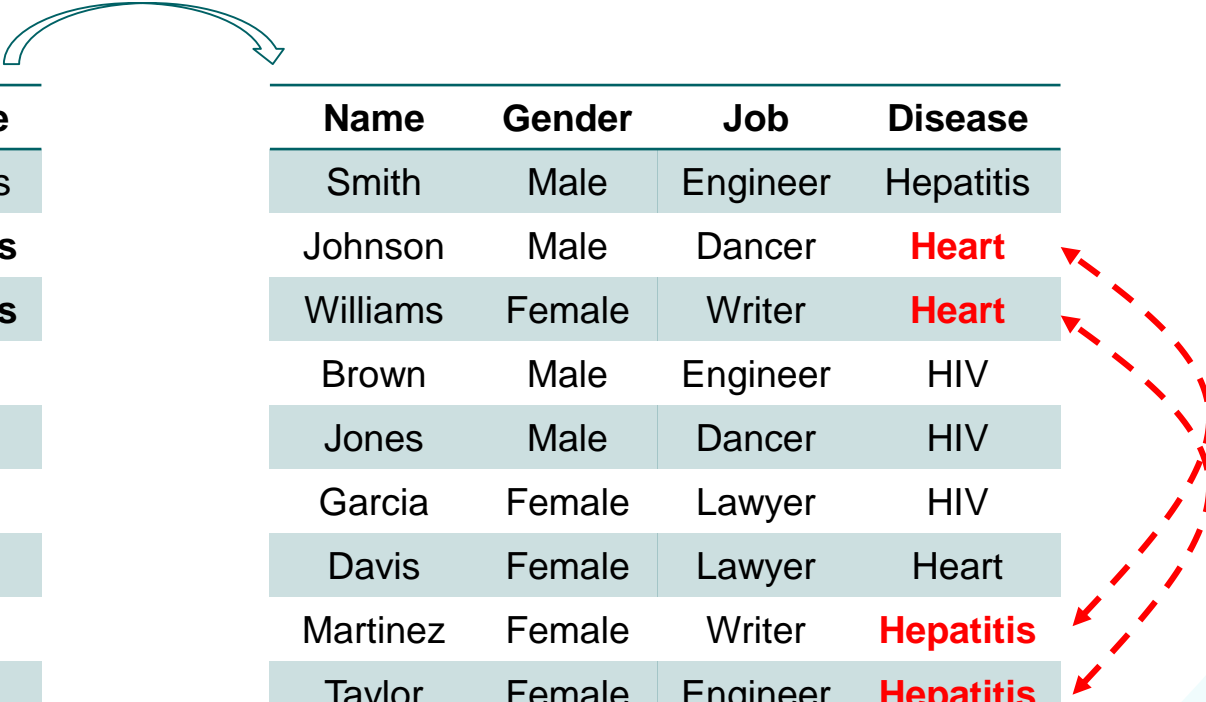
Name	Gender	Job	Disease
Smith	*	Engineer	Hepatitis
Johnson	*	Dancer	Hepatitis
Williams	*	Writer	Hepatitis
Brown	*	Engineer	HIV
Jones	*	Dancer	HIV
Garcia	*	Lawyer	HIV
Davis	*	Lawyer	Heart
Martinez	*	Writer	Heart
Taylor	*	Engineer	Heart
Anderson	*	Dancer	Heart

Column suppression – refers to suppressing a feature (i.e., column)

Many more possibilities!

Data Swapping

Anonymize the data by exchanging values of sensitive features among individuals



Name	Gender	Job	Disease
Smith	Male	Engineer	Hepatitis
Johnson	Male	Dancer	Hepatitis
Williams	Female	Writer	Hepatitis
Brown	Male	Engineer	HIV
Jones	Male	Dancer	HIV
Garcia	Female	Lawyer	HIV
Davis	Female	Lawyer	Heart
Martinez	Female	Writer	Heart
Taylor	Female	Engineer	Heart
Anderson	Male	Dancer	Heart

Name	Gender	Job	Disease
Smith	Male	Engineer	Hepatitis
Johnson	Male	Dancer	Heart
Williams	Female	Writer	Heart
Brown	Male	Engineer	HIV
Jones	Male	Dancer	HIV
Garcia	Female	Lawyer	HIV
Davis	Female	Lawyer	Heart
Martinez	Female	Writer	Hepatitis
Taylor	Female	Engineer	Hepatitis
Anderson	Male	Dancer	Heart

Frequencies remain unchanged

Additive Noise

Replace the original sensitive value s with $s+r$ where r is a random variable drawn from some distribution

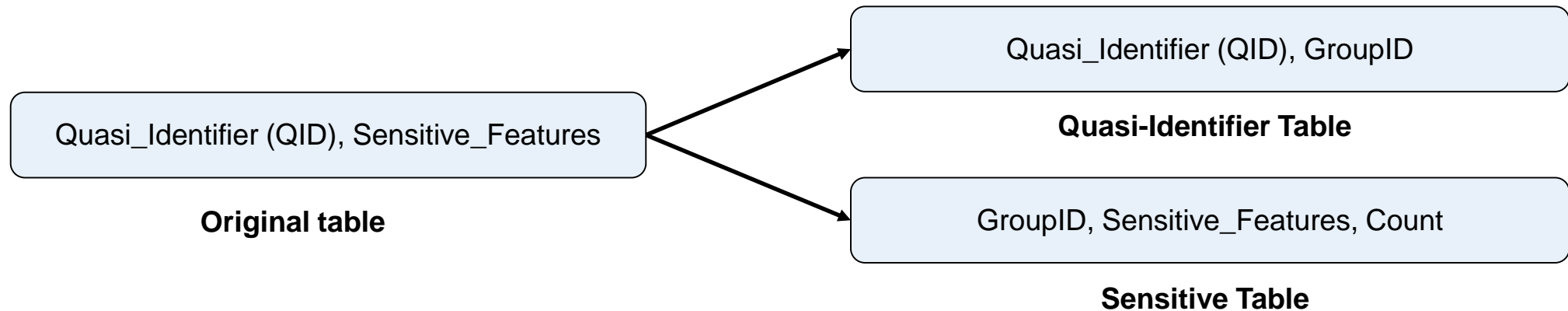
$$\text{Salary} = \text{Salary} + \text{Gaussian}(10000, 4000)$$

Name	Gender	Job	Salary
Smith	Male	Engineer	10000
Johnson	Male	Dancer	5000
Williams	Female	Writer	9500
Brown	Male	Engineer	12000
Jones	Male	Dancer	9000
Garcia	Female	Lawyer	12500
Davis	Female	Lawyer	6800
Martinez	Female	Writer	11000
Taylor	Female	Engineer	8000
Anderson	Male	Dancer	8500

Name	Gender	Job	Salary
Smith	Male	Engineer	22112
Johnson	Male	Dancer	10177
Williams	Female	Writer	13708
Brown	Male	Engineer	23365
Jones	Male	Dancer	13519
Garcia	Female	Lawyer	23954
Davis	Female	Lawyer	13451
Martinez	Female	Writer	24410
Taylor	Female	Engineer	14898
Anderson	Male	Dancer	21154

Anatomization: Decouple instead of delete, change, or swap

- Does not modify the quasi-identifier or the sensitive feature
- Instead, **de-associates the relationship between the two**



Anatomization – Example

Apply generalization to decrease the number of equivalence classes

Gender	Job	Disease
Male	Engineer	Hepatitis
Male	Dancer	Hepatitis
Female	Writer	Hepatitis
Male	Engineer	HIV
Male	Dancer	HIV
Female	Lawyer	HIV
Female	Lawyer	Heart
Female	Writer	Heart
Female	Engineer	Heart
Male	Dancer	Heart

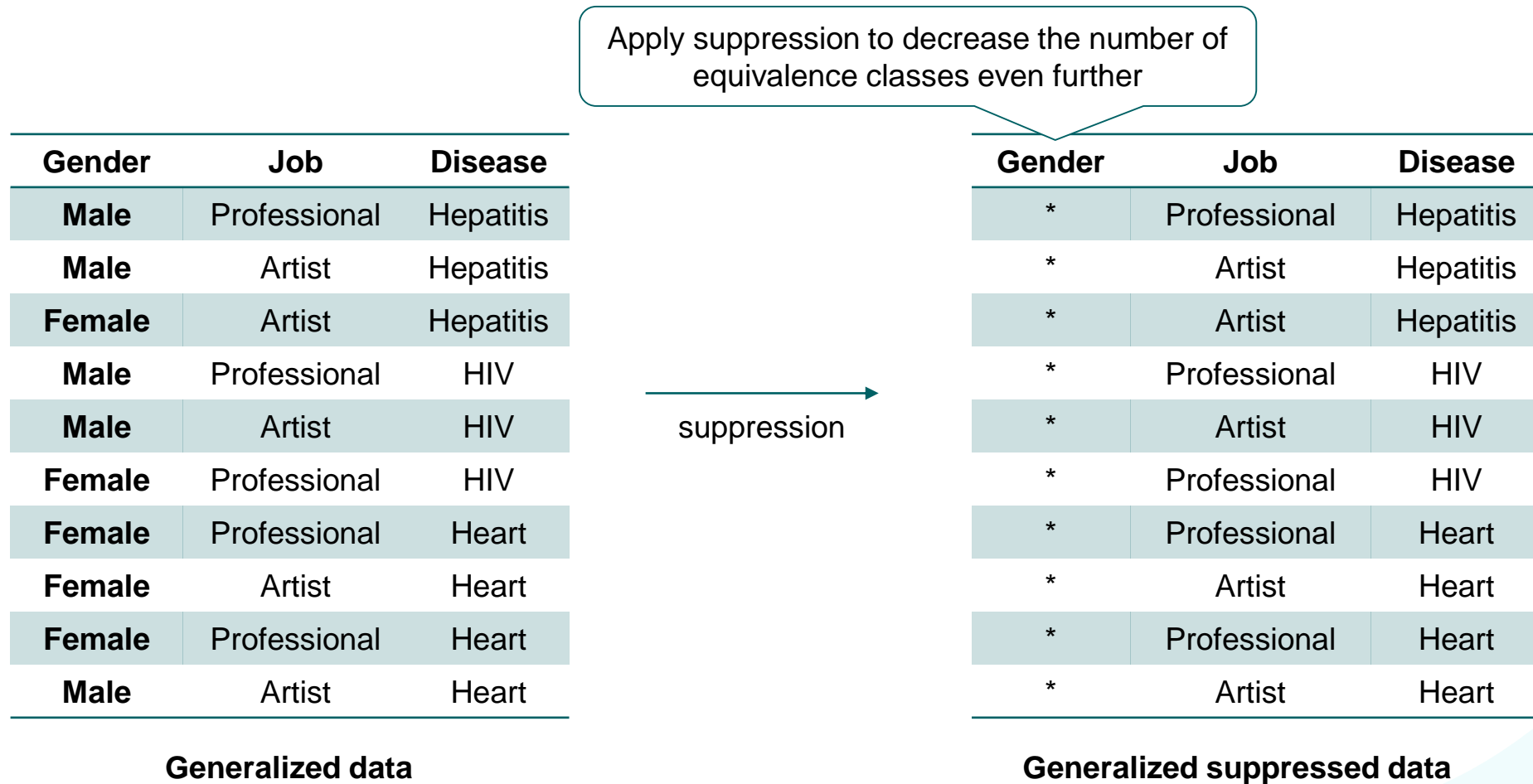
Original table

generalization →

Gender	Job	Disease
Male	Professional	Hepatitis
Male	Artist	Hepatitis
Female	Artist	Hepatitis
Male	Professional	HIV
Male	Artist	HIV
Female	Professional	HIV
Female	Professional	Heart
Female	Artist	Heart
Female	Professional	Heart
Male	Artist	Heart

Generalized data

Anatomization – Example



These suppression/generalization steps may be driven by k-anonymity or l-diversity (see later)

Anatomization – Example

Gender	Job	Disease
*	Professional	Hepatitis
*	Artist	Hepatitis
*	Artist	Hepatitis
*	Professional	HIV
*	Artist	HIV
*	Professional	HIV
*	Professional	Heart
*	Artist	Heart
*	Professional	Heart
*	Artist	Heart

Generalized suppressed data

→
anatomization

Group 1 are all the professionals

Only one professional with hepatitis

Gender	Job	GroupID
*	Professional	1
*	Artist	2
*	Artist	2
*	Professional	1
*	Artist	2
*	Professional	1
*	Professional	1
*	Artist	2
*	Professional	1
*	Artist	2

Quasi identifier table

GroupID	Disease	Count
1	Hepatitis	1
1	HIV	2
1	Heart	2
2	Hepatitis	2
2	HIV	1
2	Heart	2

Sensitive table

Anatomization – Example

Gender	Job	GroupID
*	Professional	1
*	Artist	2
*	Artist	2
*	Professional	1
*	Artist	2
*	Professional	1
*	Professional	1
*	Artist	2
*	Professional	1
*	Artist	2

reinsert the
original quasi
identifiers

Output

Gender	Job	GroupID
Male	Engineer	1
Male	Dancer	2
Female	Writer	2
Male	Engineer	1
Male	Dancer	2
Female	Lawyer	1
Female	Lawyer	1
Female	Writer	2
Female	Engineer	1
Male	Dancer	2

Quasi identifier table

GroupID	Disease	Count
1	Hepatitis	1
1	HIV	2
1	Heart	2
2	Hepatitis	2
2	HIV	1
2	Heart	2

Sensitive table

Now only group-
level probabilities

Anonymization Operations

Anonymization operations:

- Generalization: Replacing specific values with a more general category.
- Suppression: Replacing some values with a placeholder.
- Data swapping: Exchanging data points or attributes between different entities.
- Adding noise: Injecting random noise.
- Anatomization: Decoupling data by grouping.



Responsible Data Science (Confidentiality)

1. Confidentiality Risks
2. Using Encryption to Ensure Confidentiality
3. Anonymization Operations
4. **K-Anonymity**
5. L-Diversity and T-Closeness



Equivalence Class

- **Equivalence class** of an anonymized data table:
a set of instances with the same values for the **quasi-identifiers**
- k-anonymity requires that each equivalence class contains **at least k instances**

Gender	Job	Disease
Male	Professional	Hepatitis
Male	Artist	Hepatitis
Female	Artist	Hepatitis
Male	Professional	HIV
Male	Artist	HIV
Female	Professional	HIV
Female	Professional	Heart
Female	Artist	Heart
Female	Professional	Heart
Male	Artist	Heart

1-Anonymity

Each instance is a separate equivalence class

	quasi-identifiers		sensitive feature
Name	Age	ZIP-code	Disease
*	27	47577	Hepatitis
*	32	47602	Hepatitis
*	19	47578	Hepatitis
*	55	47905	HIV
*	31	47609	HIV
*	38	47606	HIV
*	23	47505	Heart
*	47	47973	Heart
*	60	47907	Heart
*	29	47505	Heart

1-Anonymity to 3-Anonymity

1-anonymity

Name	quasi-identifiers		sensitive feature
	Age	ZIP-code	Disease
*	27	47577	Hepatitis
*	32	47602	Hepatitis
*	19	47578	Hepatitis
*	55	47905	HIV
*	31	47609	HIV
*	38	47606	HIV
*	23	47505	Heart
*	47	47973	Heart
*	60	47907	Heart
*	29	47505	Heart

sort by Age

Name	quasi-identifiers		sensitive feature
	Age	ZIP-code	Disease
*	19	47578	Hepatitis
*	23	47505	Heart
*	27	47577	Hepatitis
*	29	47505	Heart
*	31	47609	HIV
*	32	47602	Hepatitis
*	38	47606	HIV
*	47	47973	Heart
*	55	47905	HIV
*	60	47907	Heart

1-Anonymity to 3-Anonymity

sorted by Age

Name	quasi-identifiers		sensitive feature
	Age	ZIP-code	Disease
*	19	47578	Hepatitis
*	23	47505	Heart
*	27	47577	Hepatitis
*	29	47505	Heart
*	31	47609	HIV
*	32	47602	Hepatitis
*	38	47606	HIV
*	47	47973	Heart
*	55	47905	HIV
*	60	47907	Heart

generalization →

Name	quasi-identifiers		sensitive feature
	Age	ZIP-code	Disease
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	3*	476**	HIV
*	3*	476**	Hepatitis
*	3*	476**	HIV
*	≥40	479**	Heart
*	≥40	479**	HIV
*	≥40	479**	Heart

1-Anonymity vs 3-Anonymity

1-anonymity

Name	quasi-identifiers		sensitive feature
	Age	ZIP-code	Disease
*	27	47577	Hepatitis
*	32	47602	Hepatitis
*	19	47578	Hepatitis
*	55	47905	HIV
*	31	47609	HIV
*	38	47606	HIV
*	23	47505	Heart
*	47	47973	Heart
*	60	47907	Heart
*	29	47505	Heart

3-anonymity (each equivalence class has at least 3 instances)

Name	quasi-identifiers		sensitive feature
	Age	ZIP-code	Disease
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	3*	476**	HIV
*	3*	476**	Hepatitis
*	3*	476**	HIV
*	≥40	479**	Heart
*	≥40	479**	HIV
*	≥40	479**	Heart

The Three Equivalence Classes

	quasi-identifiers		sensitive feature	
Name	Age	ZIP-code	Disease	
*	<30	475**	Hepatitis	Equivalence Class 1
*	<30	475**	Heart	
*	<30	475**	Hepatitis	
*	<30	475**	Heart	
*	3*	476**	HIV	Equivalence Class 2
*	3*	476**	Hepatitis	
*	3*	476**	HIV	
*	≥40	479**	Heart	Equivalence Class 3
*	≥40	479**	HIV	
*	≥40	479**	Heart	

Confidence of Identifying Instances

Consider

- a table that satisfies k-anonymity for some value k
 - an adversarial who knows the quasi-identifier values of one individual
- The adversarial cannot identify the instance corresponding to that individual with confidence greater than $1/k$

Example: confidence in guessing in the given example is not greater than $1/3$

	quasi-identifiers		sensitive feature
Name	Age	ZIP-code	Disease
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	3*	476**	HIV
*	3*	476**	Hepatitis
*	3*	476**	HIV
*	≥40	479**	Heart
*	≥40	479**	HIV
*	≥40	479**	Heart

Problems

- K-anonymity protects against identity disclosure
- is insufficient to prevent feature disclosure
- K-anonymity focuses on **quasi-identifiers (QID)** such that each QID tuple occurs in at least k instances
 → **Sensitive features are not considered!**

	quasi-identifiers		sensitive feature
Name	Age	ZIP-code	Disease
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	3*	476**	HIV
*	3*	476**	Hepatitis
*	3*	476**	HIV
*	≥40	479**	Heart
*	≥40	479**	HIV
*	≥40	479**	Heart

Possible Attacks

Name	quasi-identifiers		sensitive feature
	Age	ZIP-code	Disease
*	<30	475**	Hepatitis
*	<30	475**	Hepatitis
*	<30	475**	Hepatitis
*	<30	475**	Hepatitis
*	3*	476**	HIV
*	3*	476**	Hepatitis
*	3*	476**	HIV
*	≥40	479**	Heart
*	≥40	479**	HIV
*	≥40	479**	Heart

Homogeneity attack:

The sensitive feature value for all the instances of this equivalence class is the same

Background knowledge attack:

- When knowing Mr. Brown's age and zip code, one can conclude that he corresponds to an instance in this equivalence class.
- Also knowing that he has a very low risk for heart disease, one can conclude that he likely has HIV.

Responsible Data Science (Confidentiality)

1. Confidentiality Risks
2. Using Encryption to Ensure Confidentiality
3. Anonymization Operations
4. K-Anonymity
5. **L-Diversity and T-Closeness**



L-Diversity

- Addresses the issue of homogeneity attacks and background knowledge attacks, e.g., all instances in an equivalence class have the same sensitive feature
- An equivalence class has l -diversity, if there are **at least l “well-represented” values** for the sensitive feature
- A table has l -diversity if **every equivalence class** in the table has l -diversity
- Different interpretations of the term “well-represented”:
 - **Distinct l -diversity**
 - **Entropy l -diversity**

Distinct L-Diversity

- The simplest understanding of “well represented”: Ensure **there are at least l distinct values** for the sensitive feature **in each equivalence class**
- Distinct l -diversity does not prevent a background knowledge attack

	quasi-identifiers		sensitive feature	
Name	Age	ZIP-code	Disease	
*	<30	475**	Hepatitis	Two distinct values
*	<30	475**	Heart	
*	<30	475**	Hepatitis	
*	<30	475**	Heart	
*	3*	476**	HIV	Two distinct values
*	3*	476**	Hepatitis	
*	3*	476**	HIV	
*	≥40	479**	Heart	Two distinct values
*	≥40	479**	HIV	
*	≥40	479**	Heart	

Distinct 2-diverse table

Problem Distinct L-Diversity

quasi-identifiers			sensitive feature
Name	Age	ZIP-code	Disease
*	<30	475**	Hepatitis
*	<30	475**	Hepatitis
...
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	3*	476**	Hepatitis
*	3*	476**	HIV
*	3*	476**	HIV
...
*	3*	476**	HIV
*	≥40	479**	Heart
*	≥40	479**	HIV
*	≥40	479**	Heart

99 times Hepatitis

99 times HIV

100 instances and having two distinct values but 99% Hepatitis

100 instances and having two distinct values but 99% HIV

Entropy L-Diversity

- The entropy of an equivalence class E is defined as:

$$H(E) = - \sum_{s \in S} P(E, s) \cdot \log_2(P(E, s))$$



- S : the domain of the sensitive feature
- $P(E, s)$: the fraction of instances in E that have the sensitive value s
- Example: $H(E) = -(\frac{7}{14} \cdot \log_2(\frac{7}{14}) + \frac{3}{14} \cdot \log_2(\frac{3}{14}) + \frac{4}{14} \cdot \log_2(\frac{4}{14})) = 1.49261$
- A data table has **entropy l-diversity** if for every equivalence class E :

$$H(E) \geq -\log_2(\frac{1}{l}) = \log_2(l)$$

- This corresponds to a higher entropy than l equally distributed sensitive values
- Higher entropy is good because it is harder to guess the actual value!

Entropy L-Diversity – Example

	quasi-identifiers		sensitive feature	
Name	Age	ZIP-code	Disease	
*	<30	475**	Hepatitis	} $H(E_1) = 1$
*	<30	475**	Heart	
*	<30	475**	Hepatitis	
*	<30	475**	Heart	
*	3*	476**	HIV	} $H(E_2) = 0.92$
*	3*	476**	Hepatitis	
*	3*	476**	HIV	
*	≥40	479**	Heart	} $H(E_3) = 0.92$
*	≥40	479**	HIV	
*	≥40	479**	Heart	

What is the maximal value of l for entropy l -diversity?

$$H(E) \geq \log_2(l)$$

$$\log_2(l) = 0.92$$

$$l = 2^{0.92} = 1.89$$

Only 1-diversity!

Entropy L-Diversity – Example

quasi-identifiers			sensitive feature
Name	Age	ZIP-code	Disease
*	<30	475**	Hepatitis
*	<30	475**	Hepatitis
...
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	3*	476**	Hepatitis
*	3*	476**	HIV
*	3*	476**	HIV
...
*	3*	476**	HIV
*	≥40	479**	Heart
*	≥40	479**	HIV
*	≥40	479**	Heart

99 times

$$H(E_1) = -\left(\frac{99}{100} \cdot \log_2\left(\frac{99}{100}\right) + \frac{1}{100} \cdot \log_2\left(\frac{1}{100}\right)\right) = 0.02432$$

99 times

$$H(E_2) = -\left(\frac{1}{100} \cdot \log_2\left(\frac{1}{100}\right) + \frac{99}{100} \cdot \log_2\left(\frac{99}{100}\right)\right) = 0.02432$$

$$H(E_i) \geq \log_2(l)$$

$$\log_2(l) = 0.02432$$

$$l = 2^{0.02432} = 1.017$$

(Close to 1, so one can guess the value with high confidence)

Entropy L-Diversity

- A table has **entropy l-diversity** if for every equivalence class E :
 $H(E) \geq \log_2(l)$
- Example:
If $H(E) = 2.9$, then entropy 7-diversity holds, but entropy 8-diversity doesn't hold
- To have entropy l-diversity for each equivalence class, the entropy of the entire table has to be at least $\log_2(l)$
- Can be too restrictive: the entropy of the entire table may be low if a few values are very common

$$\log_2(7) = 2.807 \quad \log_2(8) = 3$$

T-Closeness

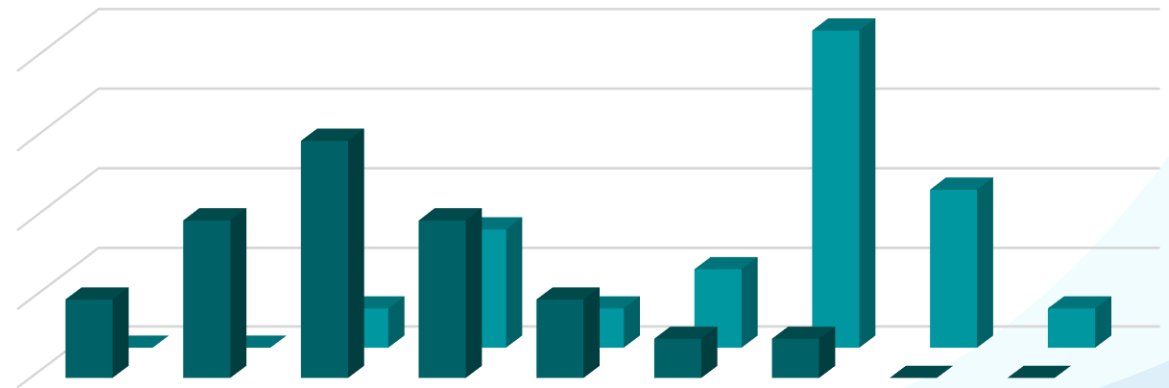
- An equivalence class has t-closeness if the **distance** between the distribution of a sensitive feature **in this class** and the distribution of it **in the whole table** is no more than a **threshold t**

$$Distance(DE, DT) \leq t$$

distribution of the sensitive feature in the equivalence class

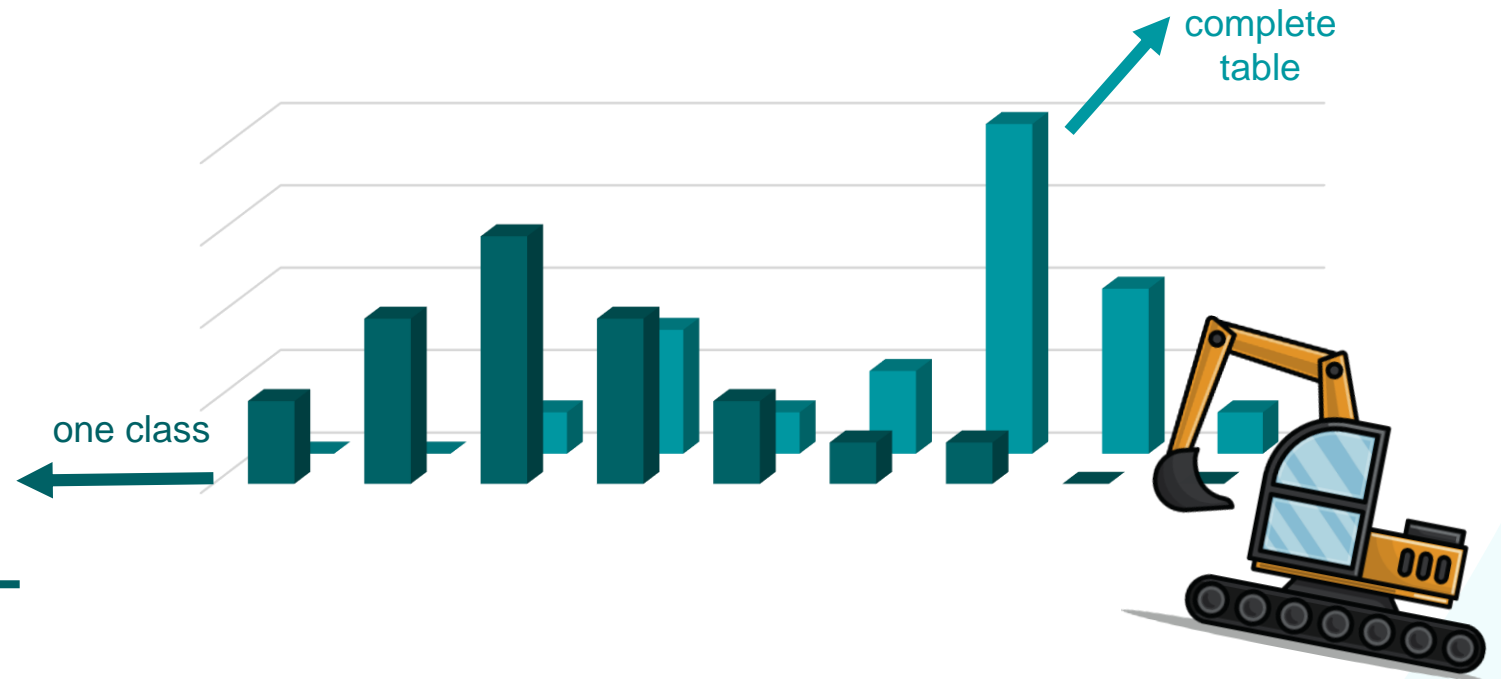
distribution of the sensitive feature in the whole table

- A table has t-closeness if **all** equivalence classes have t-closeness
- Distance measure should reflect the semantic distance among values
→ **Earth Mover's Distance**



T-Closeness

	quasi-identifiers		sensitive feature
Name	Age	ZIP-code	Disease
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	<30	475**	Hepatitis
*	<30	475**	Heart
*	3*	476**	HIV
*	3*	476**	Hepatitis
*	3*	476**	HIV
*	≥40	479**	Heart
*	≥40	479**	HIV
*	≥40	479**	Heart



Confidentiality

- **Important!** If not addressed properly, people will resist data science applications
- Even after removing explicit identifiers, there may be (un)intentional information sharing:
 - Through quasi-identifiers, it may be possible to **uniquely identify instances**
 - Sensitive features in an equivalence class may **not be diverse enough**

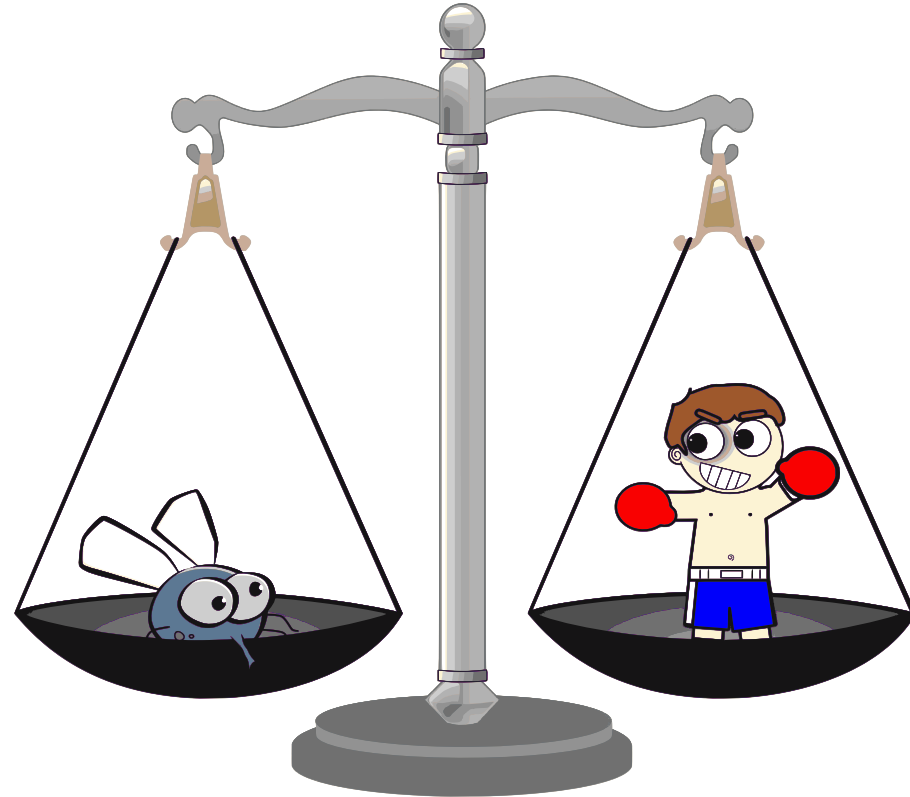
Confidentiality

- **Important!** If not addressed properly, people will resist data science applications
- Can be tackled through **encryption** and **anonymization**
- We can **measure confidentiality**, examples:
 - **K-Anonymity** focusing on instances having the same quasi-identifiers
 - **L-Diversity** and **T-Closeness** focusing on the sensitive feature
- **Tradeoffs** between **utility** and **confidentiality**

Part III: Fairness

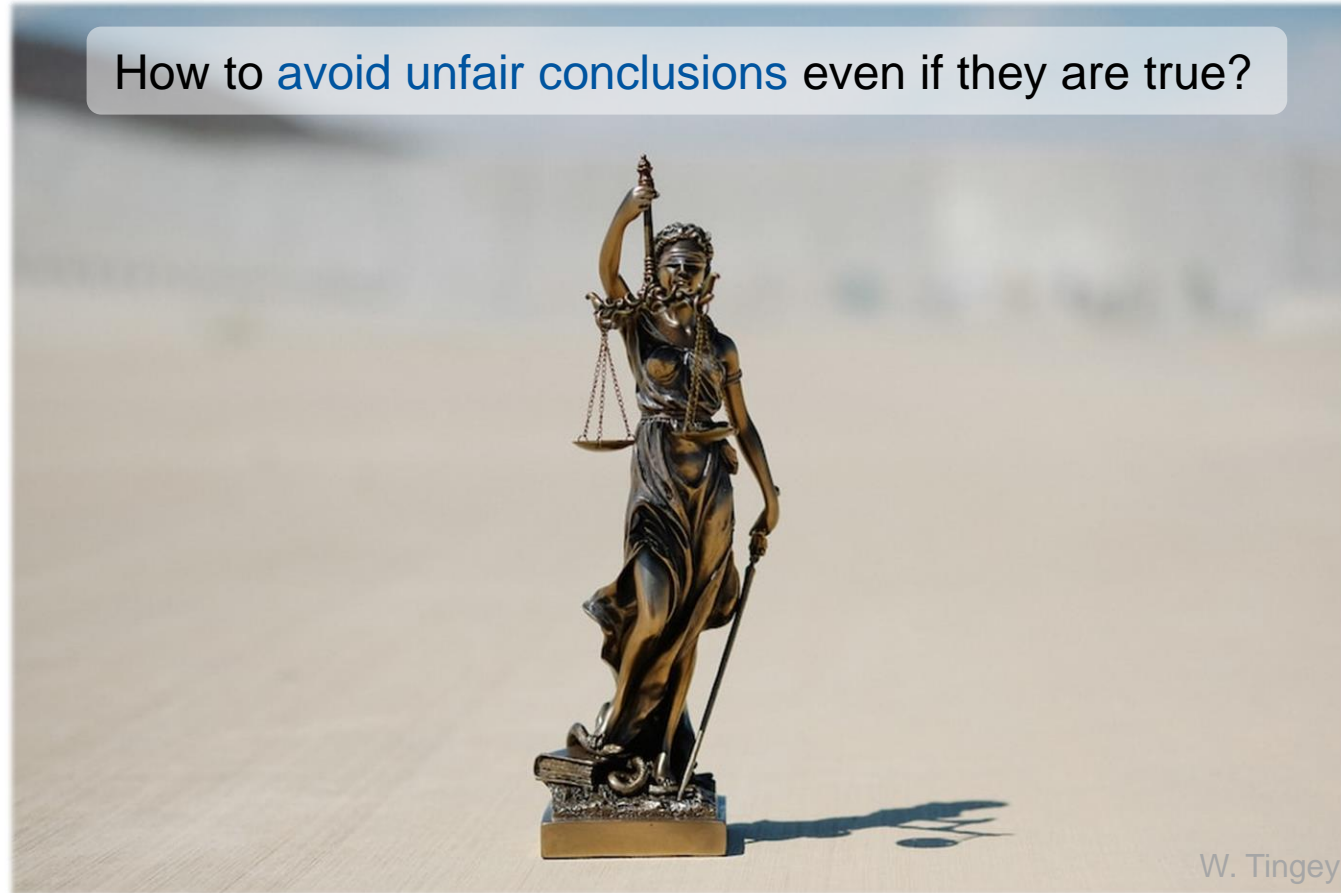
Responsible Data Science (Fairness)

1. **Motivation**
2. Preliminaries
3. Fairness Measures
4. Fair Decision Trees



Fairness – Data Science Without Prejudice

How to **avoid unfair conclusions** even if they are true?



W. Tingey

It Is Hard To Define Fairness

Correct Does Not Imply Fair



Is a company that optimizes its profits by excluding minority neighborhoods from its services fair?

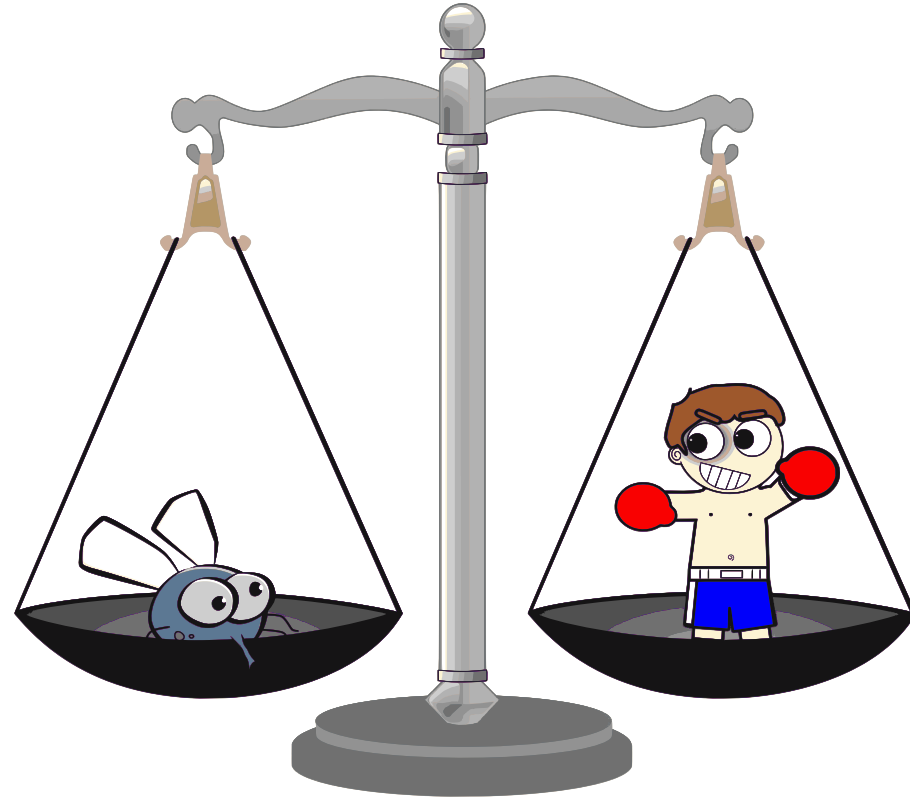
Sometimes we do not want the model with the highest accuracy.

There is often a tradeoff between (1) maximizing the accuracy of a prediction based on training data and (2) fairness incorporating contextual factors.

How to avoid self-fulfilling prophecies?

Responsible Data Science (Fairness)

1. Motivation
2. **Preliminaries**
3. Fairness Measures
4. Fair Decision Trees



Normal Itemsets

Bread	Butter	Chips	Beer
✓	✓		
		✓	✓
✓	✓	✓	✓
		✓	✓
...



{Bread, Butter}
 {Chips, Beer}
 {Bread, Butter, Chips, Beer}
 {Chips, Beer}
 ...

Note: here we ignore quantities

- Any dataset having instances and features can be converted into a multiset of transactions $\mathcal{X} \in \mathbb{M}(\mathbb{P}(\mathcal{I}))$
- Support of an itemset is defined as

$$\text{support}(\mathcal{A}) = \frac{|\{T \in \mathcal{X} \mid \mathcal{A} \subseteq T\}|}{|\mathcal{X}|}$$

$$\begin{aligned} \text{support}(\{Bread\}) &= \frac{2}{4} \\ \text{support}(\{Chips, Beer\}) &= \frac{3}{4} \\ \text{support}(\{Bread, Beer\}) &= \frac{1}{4} \end{aligned}$$

Itemsets Encoding Quantities

Bread	Butter	Chips	Beer
5	2	0	0
0	0	1	2
2	1	2	1
0	0	2	2
...



{Bread=5, Butter=2, Chips=0, Beer=0}
 {Bread=0, Butter=0, Chips=1, Beer=2}
 {Bread=2, Butter=1, Chips=2, Beer=1}
 {Bread=0, Butter=0, Chips=2, Beer=2}

...

$$\text{support}(\mathcal{A}) = \frac{|\{T \in \mathcal{X} \mid \mathcal{A} \subseteq T\}|}{|\mathcal{X}|}$$

$$\text{support}(\{Bread = 5\}) = \frac{1}{4}$$

$$\text{support}(\{Bread = 0\}) = \frac{2}{4}$$

$$\text{support}(\{Bread = 2, Beer = 1\}) = \frac{1}{4}$$

Itemsets Encoding Any Value

Age	City	Income	Gender
34	Bonn	2400	Male
45	Köln	1200	Male
39	Aachen	4200	Female
41	Bonn	2500	Female
...



{Age=34, City=Bonn, Income=2400, Gender=Male}
 {Age=45, City=Köln, Income=1200, Gender=Male}
 {Age=39, City=Aachen, Income=4200, Gender=Female}
 {Age=41, City=Bonn, Income=2500, Gender=Female}

...

$$\text{support}(\mathcal{A}) = \frac{|\{T \in \mathcal{X} \mid \mathcal{A} \subseteq T\}|}{|\mathcal{X}|}$$

$$\text{support}(\{City = Bonn\}) = \frac{2}{4}$$

$$\text{support}(\{Age = 34, Income = 0\}) = 0$$

$$\text{support}(\{Age = 34, Gender = Male\}) = \frac{1}{4}$$

Itemsets Encoding Ranges of Values

Age	City	Income	Gender
34	Bonn	2400	Male
45	Köln	1200	Male
39	Aachen	4200	Female
41	Bonn	2500	Female
...



{Age<40, City=Bonn, Income<3000, Gender=Male}
 {Age≥40, City=Köln, Income<3000, Gender=Male}
 {Age<40, City=Aachen, Income≥3000, Gender=Female}
 {Age≥40, City=Bonn, Income<3000, Gender=Female}

...

$$\text{support}(\mathcal{A}) = \frac{|\{T \in \mathcal{X} \mid \mathcal{A} \subseteq T\}|}{|\mathcal{X}|}$$

$$\text{support}(\{Age < 40\}) = \frac{2}{4}$$

$$\text{support}(\{Age \geq 40, Income < 3000\}) = \frac{2}{4}$$

$$\text{support}(\{Age < 40, Gender = Male\}) = \frac{1}{4}$$

Itemsets Encoding Real Values

Age	City	Income (pred)	Income (real)	Gender
34	Bonn	2400	1667	Male
45	Köln	1200	1456	Male
39	Aachen	4200	3987	Female
41	Bonn	2500	2420	Female
...



{Age=34, City=Bonn, Income=1667, Gender=Male}
{Age=45, City=Köln, Income=1456, Gender=Male}
{Age=39, City=Aachen, Income=3987, Gender=Female}
{Age=41, City=Bonn, Income=2420, Gender=Female}

...

- The itemsets can be partly based on predicted values
- Here, the **real values** are used for the income

By using real data values,
we can check for
data bias

Itemsets Encoding Predicted Values

Age	City	Income (pred)	Income (real)	Gender
34	Bonn	2400	1667	Male
45	Köln	1200	1456	Male
39	Aachen	4200	3987	Female
41	Bonn	2500	2420	Female
...



{Age=34, City=Bonn, **Income=2400**, Gender=Male}
 {Age=45, City=Köln, **Income=1200**, Gender=Male}
 {Age=39, City=Aachen, **Income=4200**, Gender=Female}
 {Age=41, City=Bonn, **Income=2500**, Gender=Female}

...

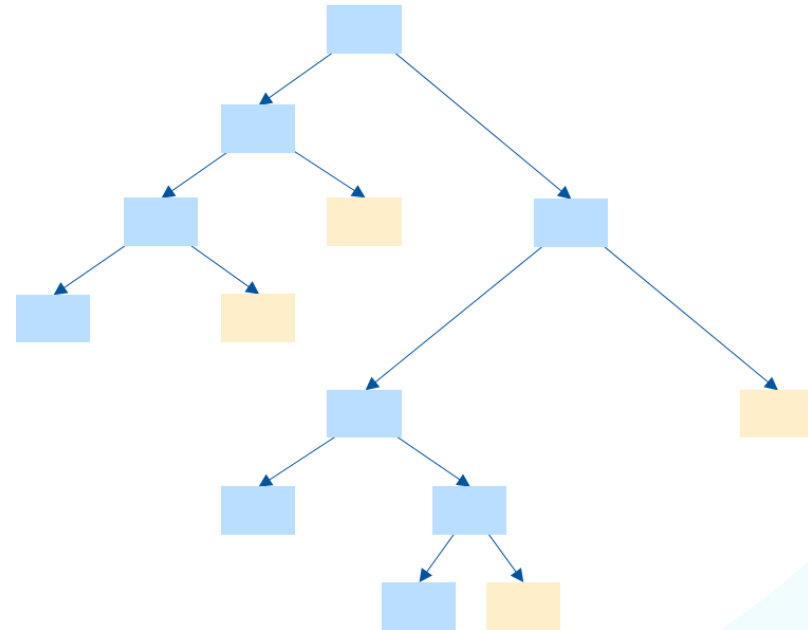
- The itemsets can be partly based on predicted values
- Here, the **predicted values** are used for the income

By using predicted values, we can check for **algorithmic (model) bias**

Itemsets Encoding Predicted Values

Age	City	Income (pred)	Income (real)	Gender
34	Bonn	2400	1667	Male
45	Köln	1200	1456	Male
39	Aachen	4200	3987	Female
41	Bonn	2500	2420	Female
...

Unfairness can be in the data and/or in the model. We cannot change the past, but we can change the model!



Association Rules

Bread	Butter	Chips	Beer
✓	✓		
		✓	✓
✓	✓	✓	✓
		✓	✓
...



{Bread, Butter}
 {Chips, Beer}
 {Bread, Butter, Chips, Beer}
 {Chips, Beer}
 ...

- Dataset $\mathcal{X} \in \mathbb{M}(\mathbb{P}(\mathcal{I}))$
- Association rules are of the form $\mathcal{A} \Rightarrow \mathcal{B}$ with $\mathcal{A} \subseteq \mathcal{I}, \mathcal{B} \subseteq \mathcal{I}$ and $\mathcal{A} \cap \mathcal{B} = \emptyset$
- For any rule, we can calculate its **confidence** as follows:

$$\text{conf}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\text{support}(\mathcal{A} \cup \mathcal{B})}{\text{support}(\mathcal{A})}$$

$$\text{conf}(\{Bread\} \Rightarrow \{Butter\}) = 1$$

$$\text{conf}(\{Chips\} \Rightarrow \{Beer\}) = 1$$

$$\text{conf}(\{Chips, Beer\} \Rightarrow \{Bread\}) = \frac{1}{3}$$

$$\text{conf}(\{Bread\} \Rightarrow \{Chips, Beer\}) = \frac{1}{2}$$

Association Rules Using Feature Values

Age	City	Income	Gender
34	Bonn	2400	Male
45	Köln	1200	Male
39	Aachen	4200	Female
41	Bonn	2500	Female
...



{Age=34, City=Bonn, Income=2400, Gender=Male}
 {Age=45, City=Köln, Income=1200, Gender=Male}
 {Age=39, City=Aachen, Income=4200, Gender=Female}
 {Age=41, City=Bonn, Income=2500, Gender=Female}

...

$$\text{conf}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\text{support}(\mathcal{A} \cup \mathcal{B})}{\text{support}(\mathcal{A})}$$

$$\text{conf}(\{Age = 34\} \Rightarrow \{City = Bonn\}) = 1$$

$$\text{conf}(\{City = Bonn\} \Rightarrow \{Age = 34\}) = \frac{1}{2}$$

Association Rules Using Value Ranges

Age	City	Income	Gender
34	Bonn	2400	Male
45	Köln	1200	Male
39	Aachen	4200	Female
41	Bonn	2500	Female
...



{Age<40, City=Bonn, Income<3000, Gender=Male}
 {Age≥40, City=Köln, Income<3000, Gender=Male}
 {Age<40, City=Aachen, Income≥3000, Gender=Female}
 {Age≥40, City=Bonn, Income<3000, Gender=Female}

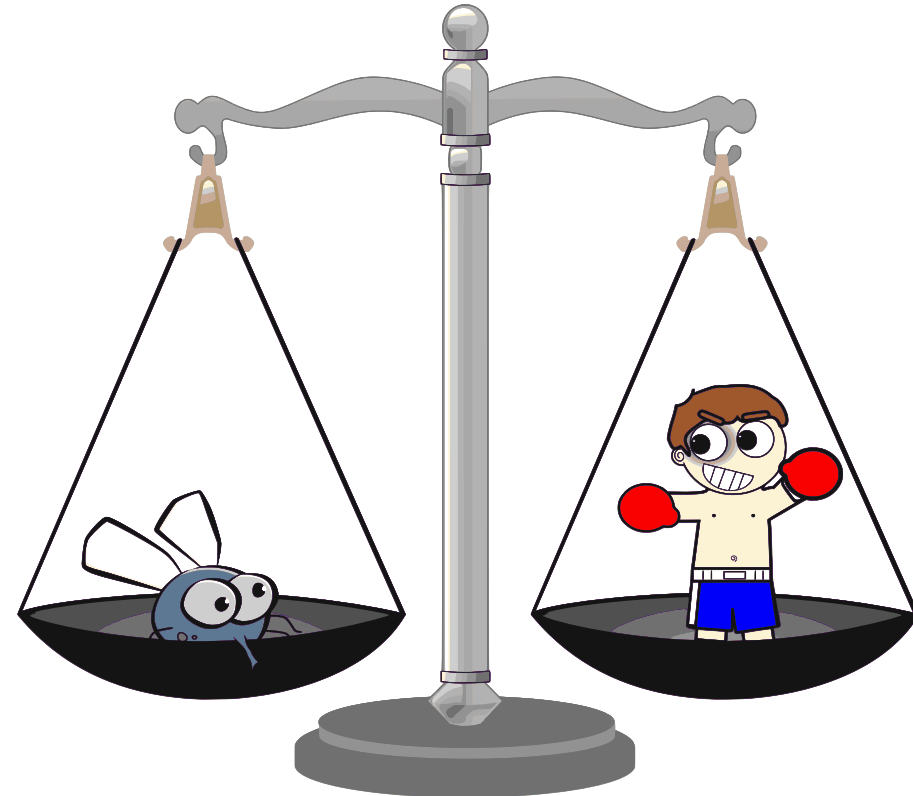
...

$$\text{conf}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\text{support}(\mathcal{A} \cup \mathcal{B})}{\text{support}(\mathcal{A})}$$

$$\begin{aligned} \text{conf}(\{Age < 40\} \Rightarrow \{Income < 3000\}) &= \frac{1}{2} \\ \text{conf}(\{Age \geq 40\} \Rightarrow \{Income < 3000\}) &= \frac{2}{4} \\ \text{conf}(\{City = Bonn\} \Rightarrow \{Income < 3000\}) &= 1 \end{aligned}$$

Responsible Data Science (Fairness)

1. Motivation
2. Preliminaries
3. **Fairness Measures**
4. Fair Decision Trees



Effect: Quantifying The Influence of a Potentially Discriminating Itemset

- A dataset with instances and features can be converted into a multiset of transactions $\mathcal{X} \in \mathbb{M}(\mathbb{P}(\mathcal{I}))$ (see previous video)
- A **potentially discriminating itemset** $\mathcal{D} \subseteq \mathcal{I}$ is an itemset that we do **not** want to have an effect on our results (rules, predictions, outcomes, etc.).
- The **effect** of a potentially discriminating itemset \mathcal{D} on an **association rule** $\mathcal{A} \Rightarrow \mathcal{B}$ is defined as
$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\text{conf}(\mathcal{A} \cup \mathcal{D} \Rightarrow \mathcal{B})}{\text{conf}(\mathcal{A} \Rightarrow \mathcal{B})}$$
- If adding the potentially discriminating itemset has **no effect** on the confidence of the rule, then
$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) \approx 1$$

Effect: Interpretation In The Context Of An Association Rule

- The **effect** of a potentially discriminating itemset \mathcal{D} on an association rule $\mathcal{A} \Rightarrow \mathcal{B}$ in the context of some data set $\mathcal{X} \in \mathbb{M}(\mathbb{P}(\mathcal{I}))$ is defined as

$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\text{conf}(\mathcal{A} \cup \mathcal{D} \Rightarrow \mathcal{B})}{\text{conf}(\mathcal{A} \Rightarrow \mathcal{B})}$$

- If adding the potentially discriminating itemset has **a positive effect** on the confidence of the rule, then

$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) > 1$$

- If adding the potentially discriminating itemset, has **a negative effect** on the confidence of the rule, then

$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) < 1$$

Effect – Example

Age	City	Income	Gender	Health Status
34	Bonn	2400	Male	Good
45	Köln	4800	Male	Medium
39	Aachen	4200	Female	Medium
41	Bonn	2400	Female	Bad
25	Aachen	1000	Male	Good
55	Bonn	5000	Female	Good
34	Aachen	2200	Female	Good
22	Köln	1500	Male	Bad
29	Bonn	2300	Male	Medium
44	Aachen	2600	Male	Medium
...

$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\text{conf}(\mathcal{A} \cup \mathcal{D} \Rightarrow \mathcal{B})}{\text{conf}(\mathcal{A} \Rightarrow \mathcal{B})}$$

$\mathcal{A} = \{City = Bonn\}$
 $\mathcal{B} = \{Health\ Status = Good\}$
 $\mathcal{D} = \{Income < 2500\}$

$$\text{conf}(\mathcal{A} \cup \mathcal{D} \Rightarrow \mathcal{B}) = \frac{\frac{1}{10}}{\frac{3}{10}} = \frac{1}{3}$$

$$\text{conf}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\frac{2}{10}}{\frac{4}{10}} = \frac{1}{2}$$

$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3} < 1$$

Adding the potentially discriminating itemset has a **negative effect** on the confidence of the rule

Effect – Example

Age	City	Income	Gender	Health Status
34	Bonn	2400	Male	Good
45	Köln	4800	Male	Medium
39	Aachen	4200	Female	Medium
41	Bonn	2400	Female	Bad
25	Aachen	1000	Male	Good
55	Bonn	5000	Female	Good
34	Aachen	2200	Female	Good
22	Köln	1500	Male	Bad
29	Bonn	2300	Male	Medium
44	Aachen	2600	Male	Medium
...

$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\text{conf}(\mathcal{A} \cup \mathcal{D} \Rightarrow \mathcal{B})}{\text{conf}(\mathcal{A} \Rightarrow \mathcal{B})}$$

$$\mathcal{A} = \{City = Aachen\}$$

$$\mathcal{B} = \{Health\ Status = Good\}$$

$$\mathcal{D} = \{Income < 2500\}$$

$$\text{conf}(\mathcal{A} \cup \mathcal{D} \Rightarrow \mathcal{B}) = \frac{\frac{2}{10}}{\frac{2}{10}} = 1$$

$$\text{conf}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{\frac{2}{10}}{\frac{4}{10}} = \frac{1}{2}$$

$$\text{effect}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) = \frac{1}{\frac{1}{2}} = 2 > 1$$

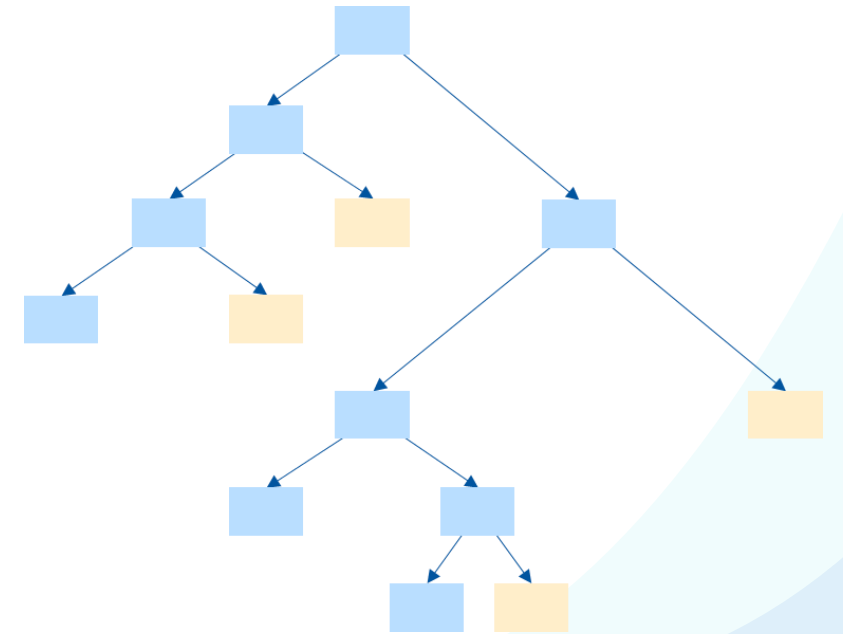
Adding the potentially discriminating itemset has a **positive effect** on the confidence of the rule

Effect – Outcome (e.g., from a decision tree)

- Assume a data set $\mathcal{X} \in \mathbb{M}(\mathbb{P}(\mathcal{I}))$, but now we consider the effect of potentially discriminating itemset $\mathcal{D} \subseteq \mathcal{I}$ on some outcome $\mathcal{B} \subseteq \mathcal{I}$ (e.g., decision to hire someone)
- The effect of a potentially discriminating itemset \mathcal{D} on some outcome \mathcal{B} is defined as:

$$\text{effect}_{\mathcal{D}}(\mathcal{B}) = \frac{\text{support}(\mathcal{B} \cup \mathcal{D})}{\text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D})}$$

- If there is **no effect**, we expect to see $\text{effect}_{\mathcal{D}}(\mathcal{B}) \approx 1$
- If there is **a positive effect** on the likelihood of outcome \mathcal{B} , then $\text{effect}_{\mathcal{D}}(\mathcal{B}) > 1$
- If there is **a negative effect** on the likelihood of outcome \mathcal{B} , then $\text{effect}_{\mathcal{D}}(\mathcal{B}) < 1$



Recall, we can change the predicted value by tweaking our model!

Effect – Outcome (Example)

Age	City	Income	Gender	Health Status
34	Bonn	2400	Male	Good
45	Köln	4800	Male	Medium
39	Aachen	4200	Female	Medium
41	Bonn	2400	Female	Bad
25	Aachen	1000	Male	Good
55	Bonn	5000	Female	Good
34	Aachen	2200	Female	Good
22	Köln	1500	Male	Bad
29	Bonn	2300	Male	Medium
44	Aachen	2600	Male	Medium
...

$$\text{effect}_{\mathcal{D}}(\mathcal{B}) = \frac{\text{support}(\mathcal{B} \cup \mathcal{D})}{\text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D})}$$

$$\mathcal{B} = \{\text{Health Status} = \text{Good}\}$$

$$\mathcal{D} = \{\text{Income} > 4000\}$$

$$\text{support}(\mathcal{B} \cup \mathcal{D}) = \frac{1}{10}$$

$$\text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D}) = \frac{4}{10} \cdot \frac{3}{10} = \frac{12}{100}$$

$$\text{effect}_{\mathcal{D}}(\mathcal{B}) = \frac{\frac{1}{10}}{\frac{12}{100}} = \frac{5}{6} < 1$$

There is a **slightly negative effect** on the likelihood of the outcome (the value is close to one meaning that there is nearly no effect)

Effect – Outcome (Example)

Age	City	Income	Gender	Health Status
34	Bonn	2400	Male	Good
45	Köln	4800	Male	Medium
39	Aachen	4200	Female	Medium
41	Bonn	2400	Female	Bad
25	Aachen	1000	Male	Good
55	Bonn	5000	Female	Good
34	Aachen	2200	Female	Good
22	Köln	1500	Male	Bad
29	Bonn	2300	Male	Medium
44	Aachen	2600	Male	Medium
...

$$\text{effect}_{\mathcal{D}}(\mathcal{B}) = \frac{\text{support}(\mathcal{B} \cup \mathcal{D})}{\text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D})}$$

$$\mathcal{B} = \{\text{Health Status} = \text{Good}\}$$

$$\mathcal{D} = \{\text{Age} < 35\}$$

$$\text{support}(\mathcal{B} \cup \mathcal{D}) = \frac{3}{10}$$

$$\text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D}) = \frac{4}{10} \cdot \frac{5}{10} = \frac{20}{100}$$

$$\text{effect}_{\mathcal{D}}(\mathcal{B}) = \frac{\frac{3}{10}}{\frac{20}{100}} = \frac{3}{2} > 1$$

There is a **positive effect** on the likelihood of the outcome

Discrimination – Association Rules

- Assume a data set $\mathcal{X} \in \mathbb{M}(\mathbb{P}(\mathcal{I}))$
- The level of discrimination given a potentially discriminating itemset $\mathcal{D} \subseteq \mathcal{I}$ on some **association rule** $\mathcal{A} \Rightarrow \mathcal{B}$ is defined as

$$\text{disc}_{\mathcal{D}}(\mathcal{A} \Rightarrow \mathcal{B}) = |\text{conf}(\mathcal{A} \cup \mathcal{D} \Rightarrow \mathcal{B}) - \text{conf}(\mathcal{A} \Rightarrow \mathcal{B})|$$

- This yields a value between 0 and 1 where:
 - 0 – no discrimination
 - 1 – maximal discrimination

Discrimination – Outcome

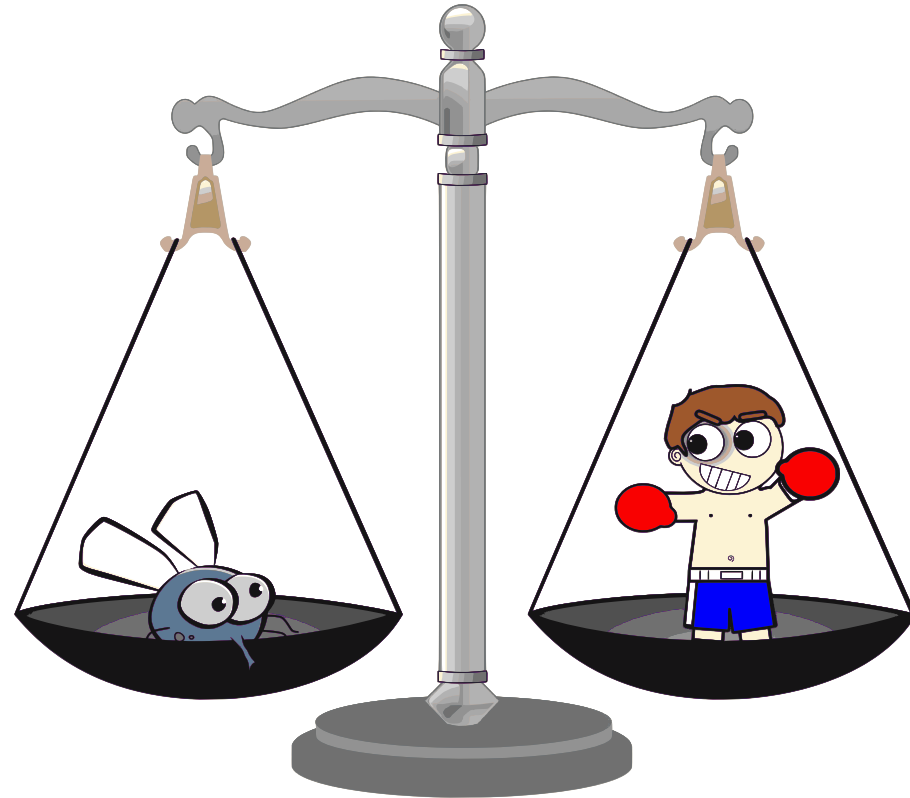
- Assume a data set $\mathcal{X} \in \mathbb{M}(\mathbb{P}(\mathcal{I}))$
- The level of discrimination given a potentially discriminating itemset $\mathcal{D} \subseteq \mathcal{I}$ on some outcome $\mathcal{B} \subseteq \mathcal{I}$ is defined as

$$\text{disc}_{\mathcal{D}}(\mathcal{B}) = |\text{support}(\mathcal{B} \cup \mathcal{D}) - \text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D})|$$

- This yields a value between 0 and 1 where:
 - 0 – no discrimination
 - 1 – maximal discrimination

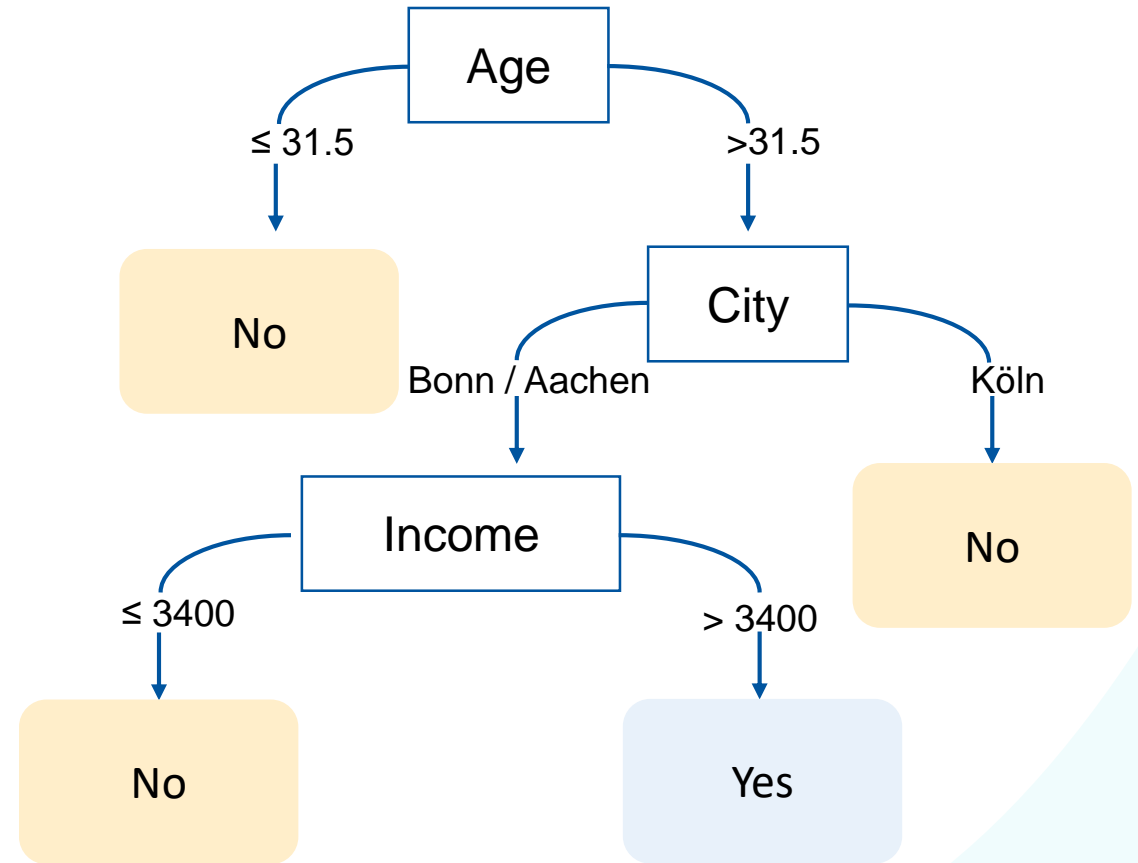
Responsible Data Science (Fairness)

1. Motivation
2. Preliminaries
3. Fairness Measures
4. **Fair Decision Trees**



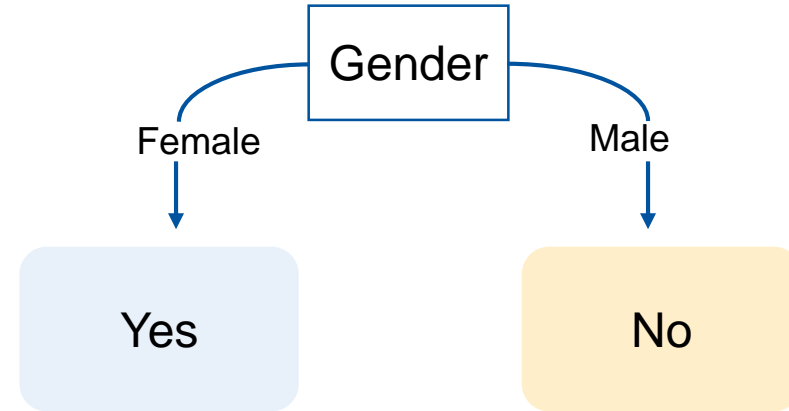
Biased Historic Data May Affect Decisions

Age	City	Income	Flat Ownership
34	Bonn	2400	No
45	Köln	4800	No
39	Aachen	4200	Yes
41	Bonn	2400	Yes
25	Aachen	1000	No
55	Bonn	5000	Yes
34	Aachen	3500	Yes
22	Köln	1500	No
29	Bonn	2300	No
44	Aachen	2600	No
...



Biased Historic Data May Affect Decisions

Age	City	Income	Gender	Flat Ownership
34	Bonn	2400	Male	No
45	Köln	4800	Male	No
39	Aachen	4200	Female	Yes
41	Bonn	2400	Female	Yes
25	Aachen	1000	Male	No
55	Bonn	5000	Female	Yes
34	Aachen	3500	Female	Yes
22	Köln	1500	Male	No
29	Bonn	2300	Male	No
44	Aachen	2600	Male	No
...



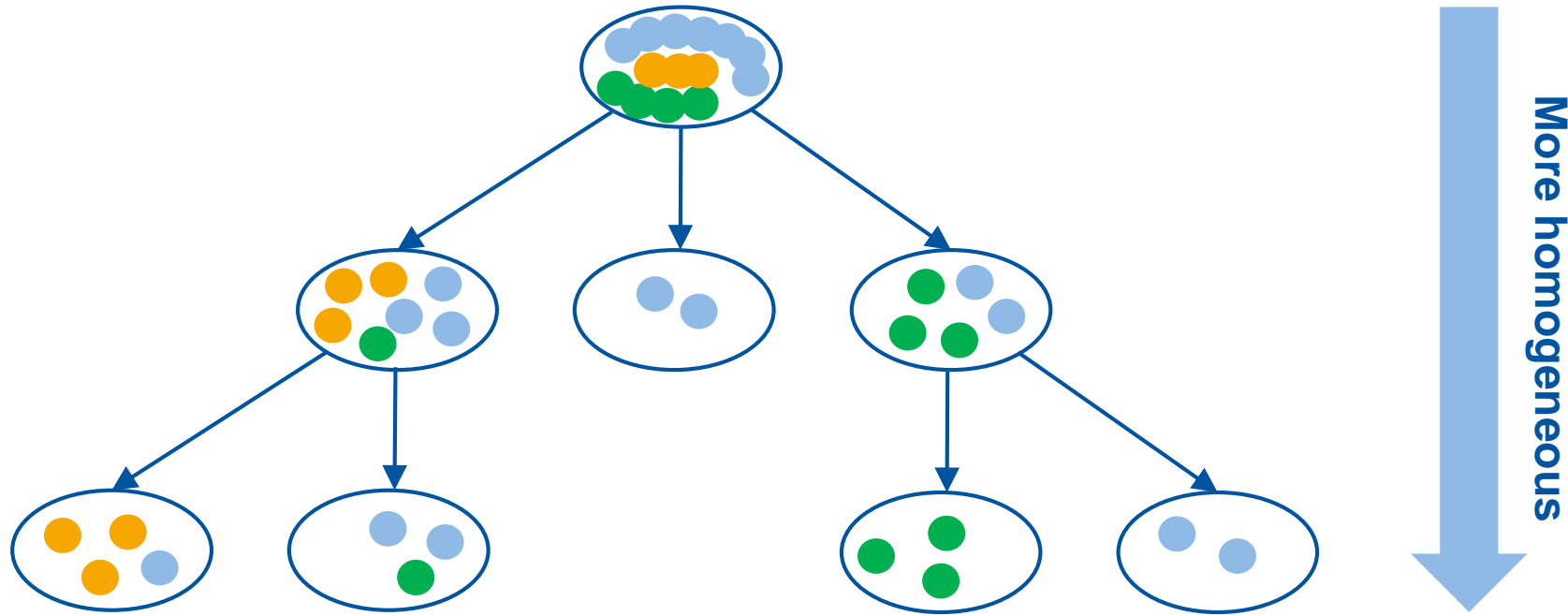
Attribute we do not want to have impact (due to biased data or fairness reasons)

Making Decision Trees Fair – Three Approaches

- **Pre-processing**
 - Removing discriminatory features from data
 - Removing or duplicating instances
 - Problem: indirect discrimination
- **In-processing**
 - Considering a dependency on discriminatory features and the accuracy of the split while making a decision tree
- **Post-processing**
 - Relabeling leaves in a way that discrimination is lowered
 - Problem: loss in the accuracy

Always a trade-off
between fairness and
accuracy!

Traditional Decision Tree Learning Using Information Gain



Check slides on decision trees for details.

Information gain = improvement in knowledge

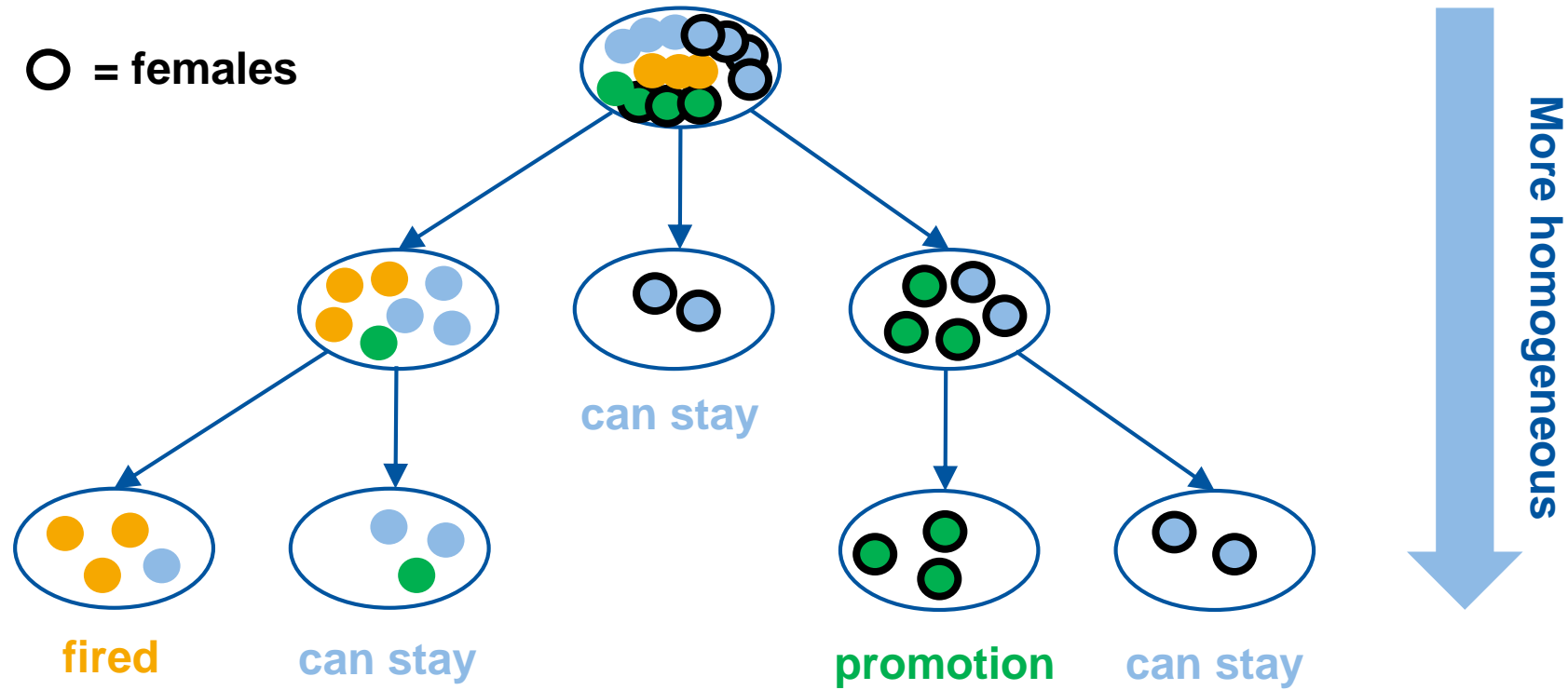
$$IG(d) = H(t) - H_W^d(t)$$

$$H(t) = - \sum_{k=1}^K (P(t = k) \cdot \log_s(P(t = k)))$$

$$H_W(t) = \sum_{node \in nodes(d)} \left(\frac{|node|}{N} \cdot H_{node}(t) \right)$$

What If

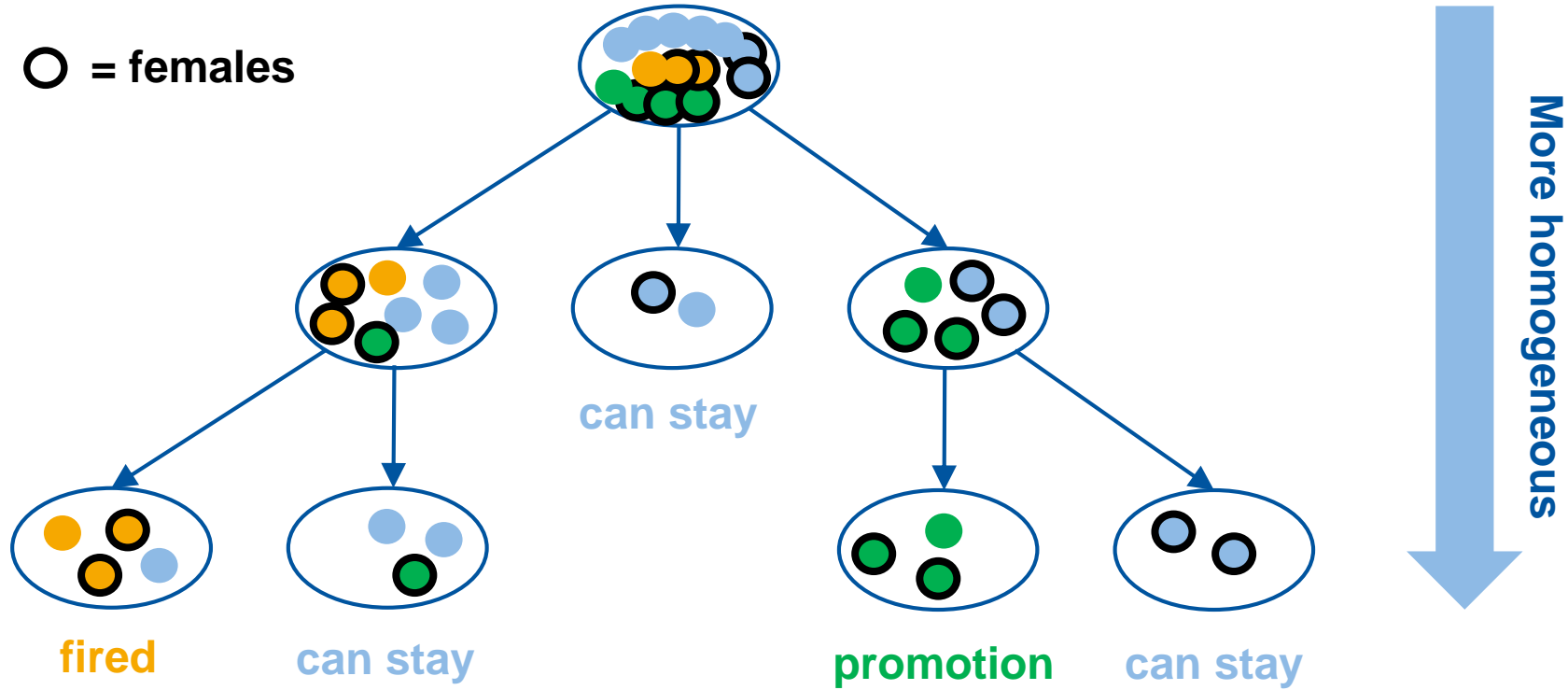
○ = females



Unfair?

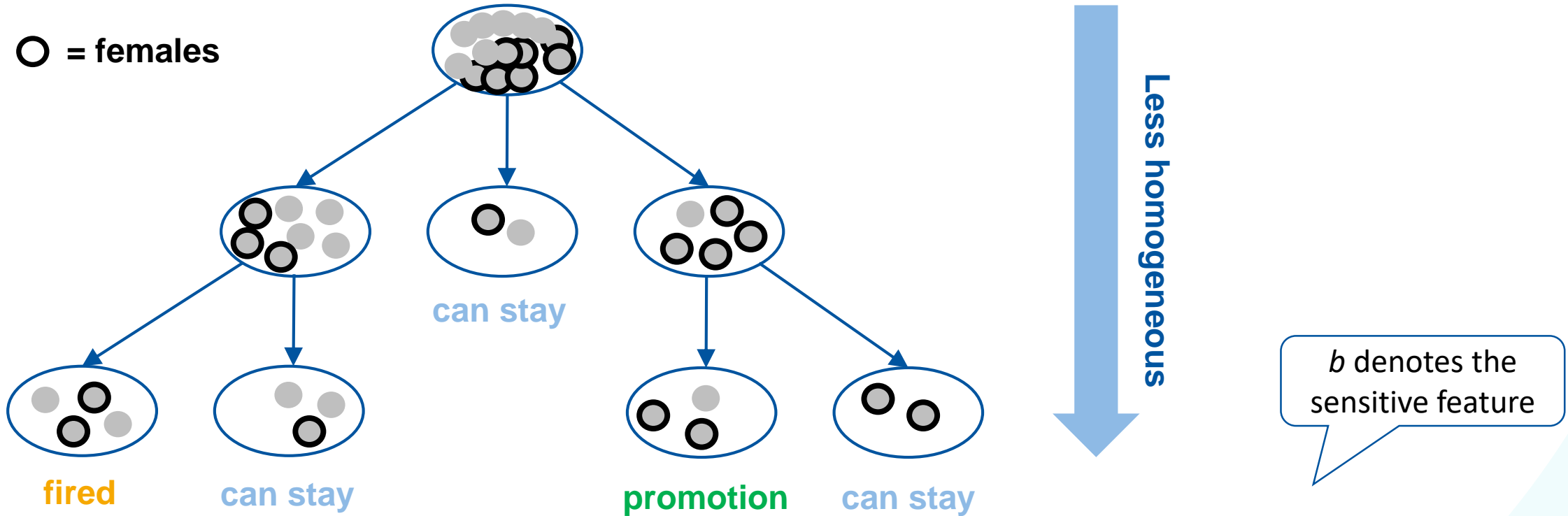
What If

○ = females



Fair?

Solution (In-Processing) – Information Gain in Sensitivity



$$IGS(d) = H(b) - H_W^d(b)$$

(lower is better)

$$H(b) = - \sum_{k=1}^K (P(b = k) \cdot \log_s(P(b = k)))$$

$$H_W(b) = \sum_{node \in nodes(d)} \left(\frac{|node|}{N} \cdot H_{node}(b) \right)$$

Solution (In-Processing) – Two Forces When Splitting

IGC = classical information gain

IGS = gain in sensitivity

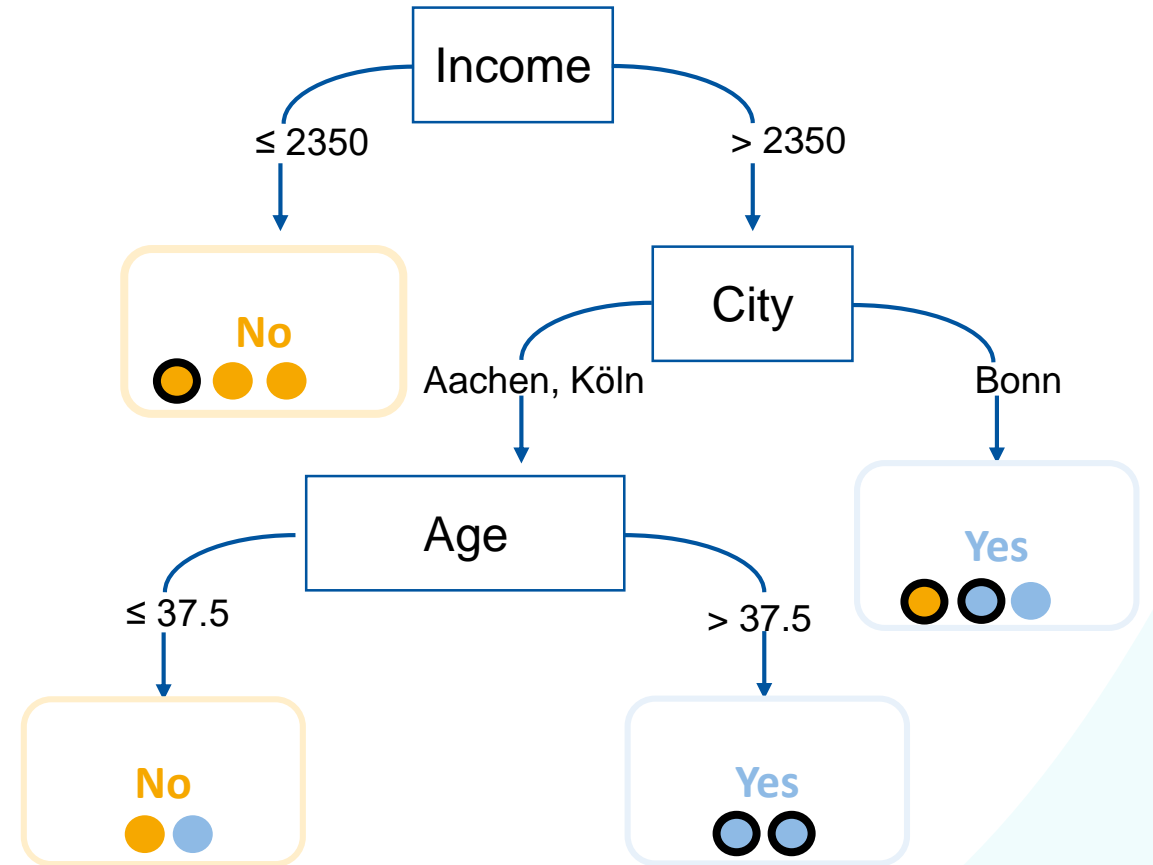
$$IGC(d) = H(t) - H_W^d(t) \quad \longleftrightarrow \quad IGS(d) = H(b) - H_W^d(b)$$

maximize **minimize**

Combine both!

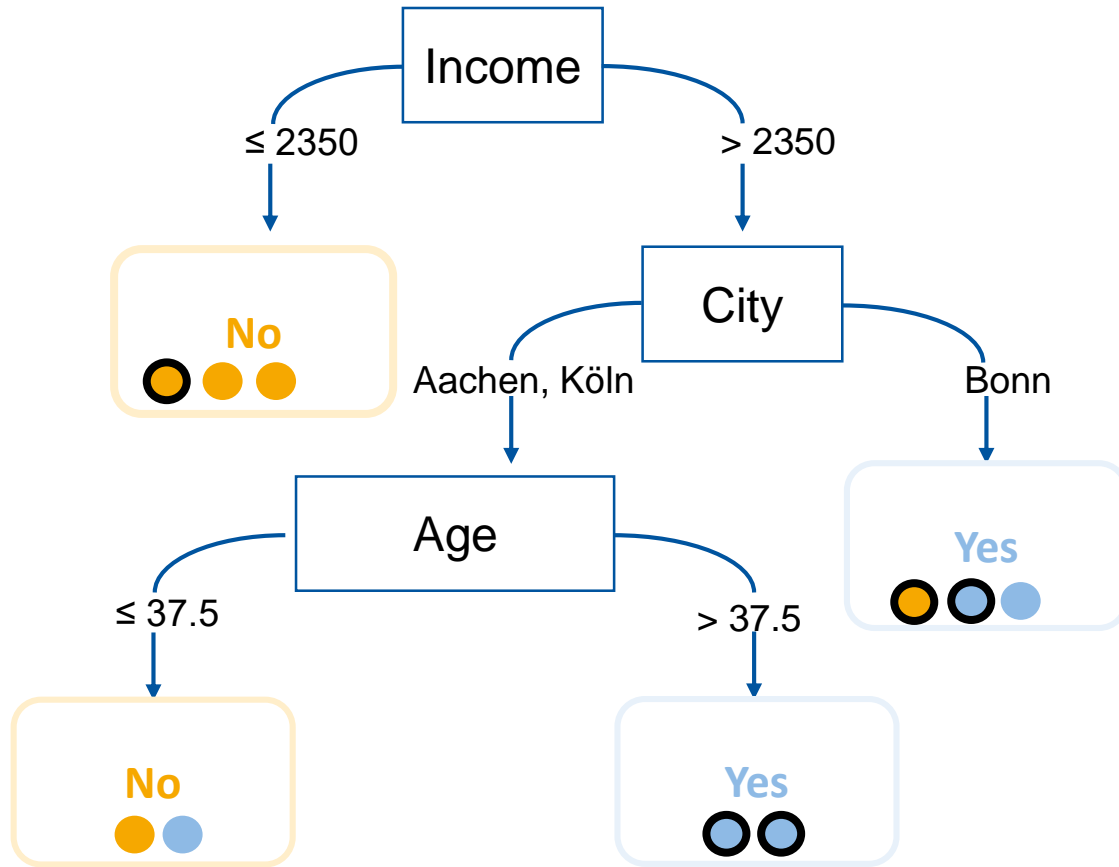
Is This Decision Tree Fair?

Age	City	Income	Gender (disc feature)	Flat Ownership (target)
34	Bonn	2400	Male	Yes
36	Bonn	4800	Female	Yes
39	Aachen	4200	Female	Yes
41	Köln	2400	Female	Yes
25	Aachen	2600	Male	Yes
55	Bonn	5000	Female	No
34	Aachen	3500	Male	No
22	Köln	1500	Male	No
29	Bonn	2300	Male	No
39	Aachen	2200	Female	No
...



$$Accuracy = \frac{8}{10} = 80\%$$

Solution (Post-Processing) – Measuring Discrimination



$$Accuracy = \frac{8}{10} = 80\%$$

Compute discrimination (outcome)

$$disc_{\mathcal{D}}(\mathcal{B}) = |\text{support}(\mathcal{B} \cup \mathcal{D}) - \text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D})|$$

\mathcal{B} : outcome

(target itemset, e.g., Flat Ownership = Yes)

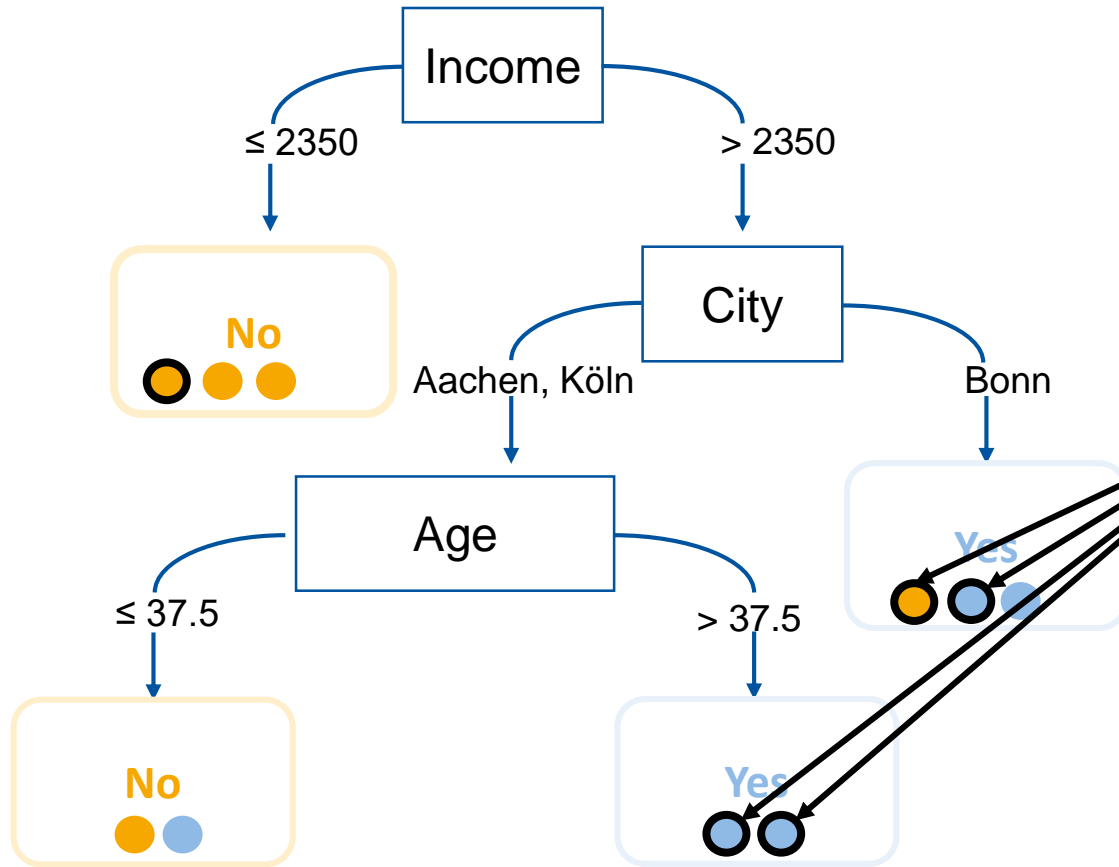
\mathcal{D} : potentially discriminating itemset

(e.g., Gender = Female)

→ A discrimination close to 0 means no discrimination

→ A discrimination close to 1 means maximal discrimination

Solution (Post-Processing) – Measuring Discrimination



Compute discrimination (outcome)

\mathcal{B} : Flat Ownership = Yes

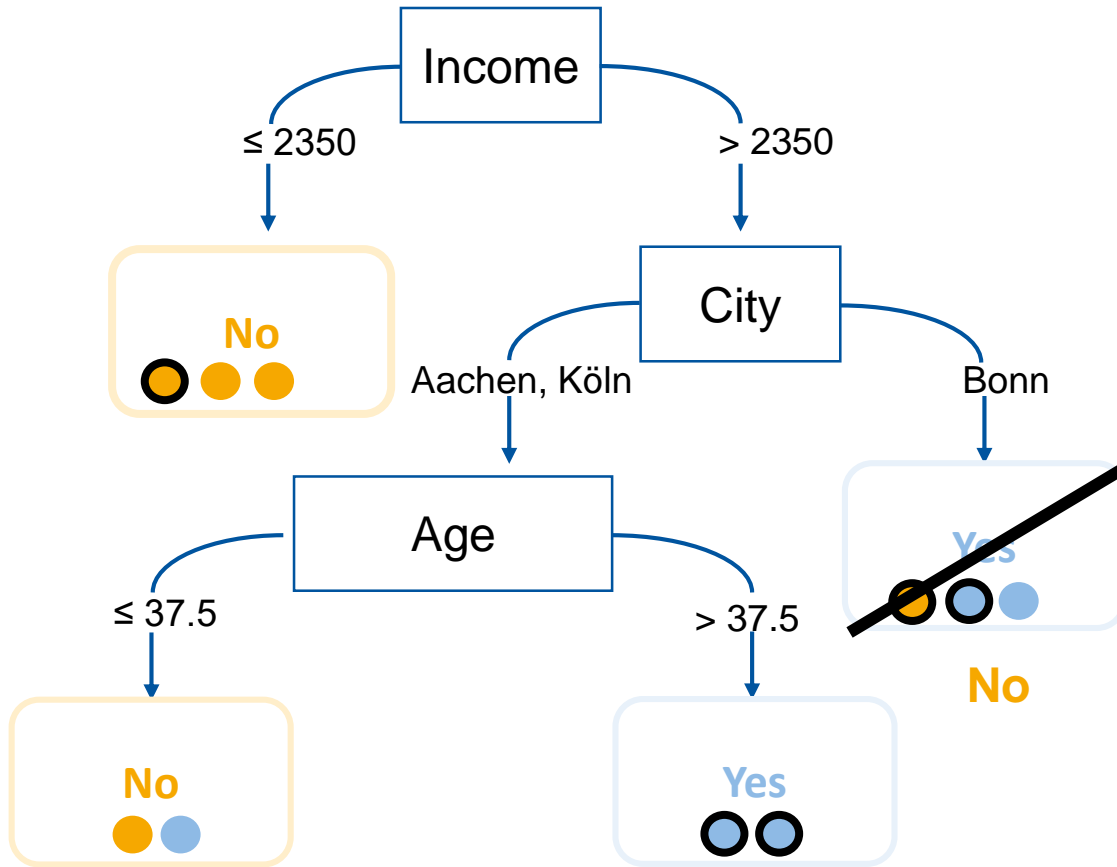
\mathcal{D} : Gender = Female

$$\text{disc}_{\mathcal{D}}(\mathcal{B}) = |\text{support}(\mathcal{B} \cup \mathcal{D}) - \text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D})|$$

$$= \left| \frac{4}{10} - \frac{5}{10} \cdot \frac{5}{10} \right| = 0.15$$

$$\text{Accuracy} = \frac{8}{10} = 80\%$$

Solution (Post-Processing) – Relabeling Leaves



Compute discrimination (outcome)

$$\text{disc}_{\mathcal{D}}(\mathcal{B}) = |\text{support}(\mathcal{B} \cup \mathcal{D}) - \text{support}(\mathcal{B}) \cdot \text{support}(\mathcal{D})|$$

$$= \left| \frac{2}{10} - \frac{2}{10} \cdot \frac{5}{10} \right| = 0.1 \quad \downarrow$$

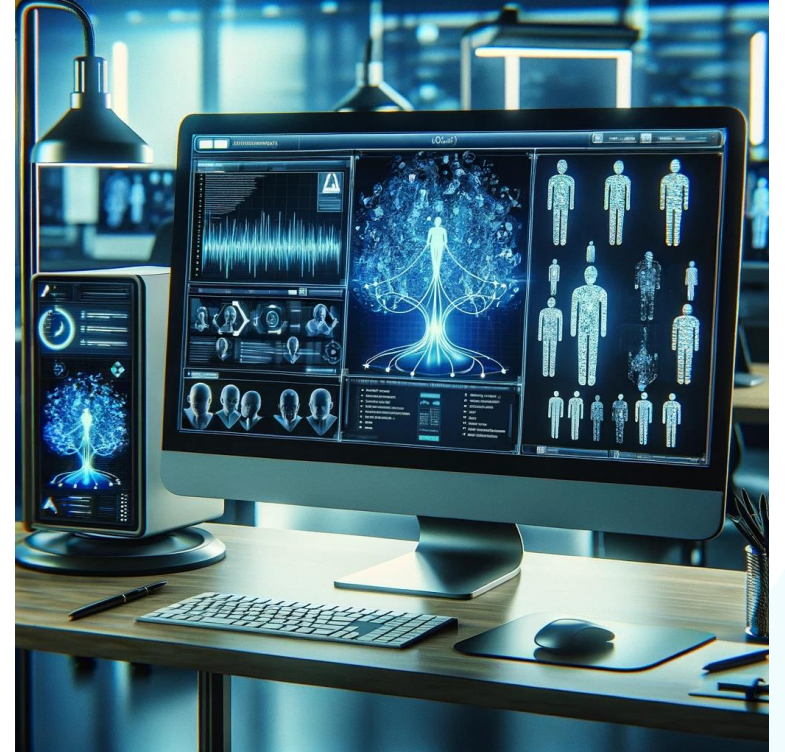
$$\text{Accuracy} = \frac{7}{10} = 70\% \quad \downarrow$$

→ Reduced discrimination

→ Reduced accuracy

Conclusion: Responsible Data Science

- Four **key concerns**: Fairness, Accuracy, Confidentiality, and Transparency (FACT, not FAIR).
- **Confidentiality**: Encryption, anonymization, K-Anonymity, L-Diversity, and T-Closeness
- Measuring **fairness** (effect) and making models fair.



Generated using DALL·E 3

With great power comes great responsibility!



Generated using DALL·E 3



Chair of Process
and Data Science

RWTHAACHEN
UNIVERSITY

Learn more? Visit: www.vdaalst.com & www.pads.rwth-aachen.de

Preparation for class on Friday, 19 January 2024 (mandatory):

Read the following research paper:

Oded Maron and Andrew Moore: Hoeffding Races: Accelerating Model Selection Search for Classification and Function Approximation. Advances in Neural Information Processing Systems 6 (NIPS 1993): 59-66, 1993.
(The paper is available online at <https://proceedings.neurips.cc>.)

Focus on the following questions (which will be further explored in TPS exercises in class):

- (1) What is the fundamental problem when using cross-validation (or performance on a validation set) to select between different ML models?
- (2) What is the key idea behind Hoeffding races and how does it address the problem identified in (1)?
- (3) What is the role of the parameters Δ and δ , respectively?

Bring your answers to these questions (which can be in the form of bullet points) to class; they will be the basis for TSP exercises).

NB: Full understanding of the proof in Section 3 is desirable but not essential.