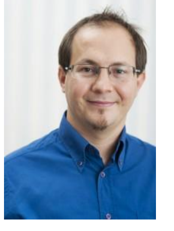
Security, Privacy & Explainability in Machine Learning

Rudolf Mayer

March 10^{th}



mayer@ifs.tuwien.ac.at
https://www.ifs.tuwien.ac.at/~mayer
https://www.sba-research.org/rudolf-mayer/

FACULTY OF !NFORMATICS

- Overview on the lecture topics
  - **Privacy preserving data publishing**
  - Secure computation
  - Adversarial examples
  - Backdoor attacks
  - Explainable AI

- Privacy-preserving data publishing:
  - Pseudonimity
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
  - Synthetic data
  - Differential privacy

- Other concerns in data publishing:
  - Intellectual digital property protection → watermarking & fingerprinting data

Outline

- Privacy: definitions and motivation
- Pseudonimisation
    - Record-Linkage Attack
- Anonymisation
    - *k*-anonymity
    - *l*-diversity
    - *t*-closeness
- Data watermarking and fingerprinting

Outline

- Privacy: definitions and motivation
- Pseudonimisation
  - Record-Linkage Attack
- Anonymisation
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
- Data watermarking and fingerprinting

Privacy definitions

- *“Privacy is the ability of an individual or group to seclude themselves, or* ***information*** *about themselves”*
- *“The challenge of* ***data privacy*** *is to use data while protecting an individual's privacy preferences and their personally identifiable information”*

- ***Pseudonymity*** *is the use of pseudonyms as IDs*
- ***Anonymity*** *is the state of being not identifiable within a set of subjects, the anonymity set*

Anonymity, Unobservability, and Pseudonymity - A Proposal for Terminology. Pfitzmann & Köhntopp. 2001

Privacy-preserving data analysis

- Concerned with **micro-data**
  - Data at the level of an individual

- **Macro** data describes mainly two subtypes of data:
  - Aggregated data
  - System-level data

- **Meso** data: data on collective and cooperative actors
  - Commercial companies, organizations or political parties

Privacy-preserving data analysis

- Large amounts of personal data becomes available
  - Analysis, distribution, sharing often conflicting with data protection laws (GDPR, ...)

  - Especially critical with highly sensitive information
    - E.g. health data, financial data, ..

- *Solutions?*
  - E.g. Data sanitisation to allow privacy-preserving data publishing (PPDP), privacy-preserving computation

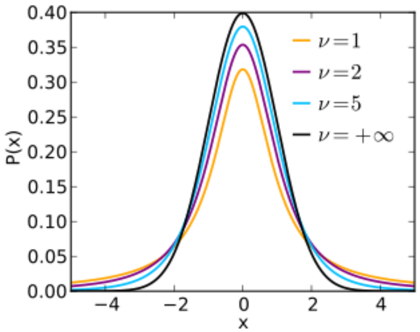
Privacy-preserving data analysis

- Two main approaches
  - Privacy-preserving data publishing
    - De-identification of information: making sure that the data published does not contain personal identifiable information;
    - k-anonymity
    - Differential Privacy
    - Synthetic Data
    - ...
  - Privacy-preserving computation
    - Making sure that computed result doesn't allow inference on the data
    - Secure Multi-Party Computation
    - Homomorphic Encryption
    - Differential Privacy
    - …

Outline

- Privacy: definitions and motivation
- Pseudonimisation
  - Record-Linkage Attack
- Anonymisation
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
- Data watermarking and fingerprinting

Pseudonymisation



- A state of disguised identity
- Pseudonym identifies a holder, that is, one or more human beings who possess but do not disclose their true names (legal identities)
- It enables a consolidation of a persons' data without revealing identities
  - Data can also mean books, paintings, etc...
- Depending on requirements:
  - One-way pseudonymisation
  - Reversible pseudonymisation – trusted third party!

ID	Name	Date of birth	City of residence
1	William Smith	1/2/73	Berkeley, California
2	Anna Williams	23/8/79	Berkeley, CA

ID	**Pseudonym**	Date of birth	City of residence
1	**John Doe**	1/2/73	Berkeley, California
2	**Jane Doe**	23/8/79	Berkeley, CA

Pseudonymisation

- GDPR:
  - *"…personal data … that can no longer be attributed to a specific data subject without the use of additional information"*
  - pseudonymized personal data **remain** personal data nonetheless, provided the controller **or** another party has the means to **reverse** the process

- Thus the same principles for storing, processing, sharing, etc. still apply!
  - *However, potentially changing interpretation of that status*

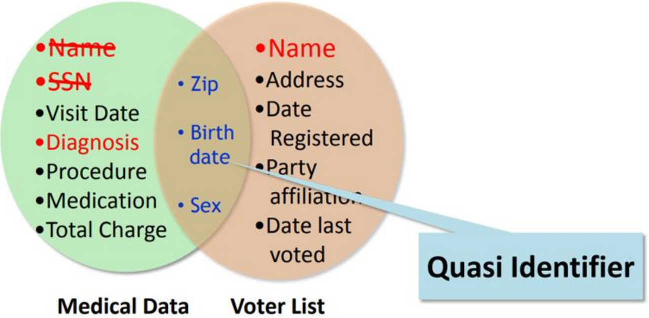
Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Data Sanitisation

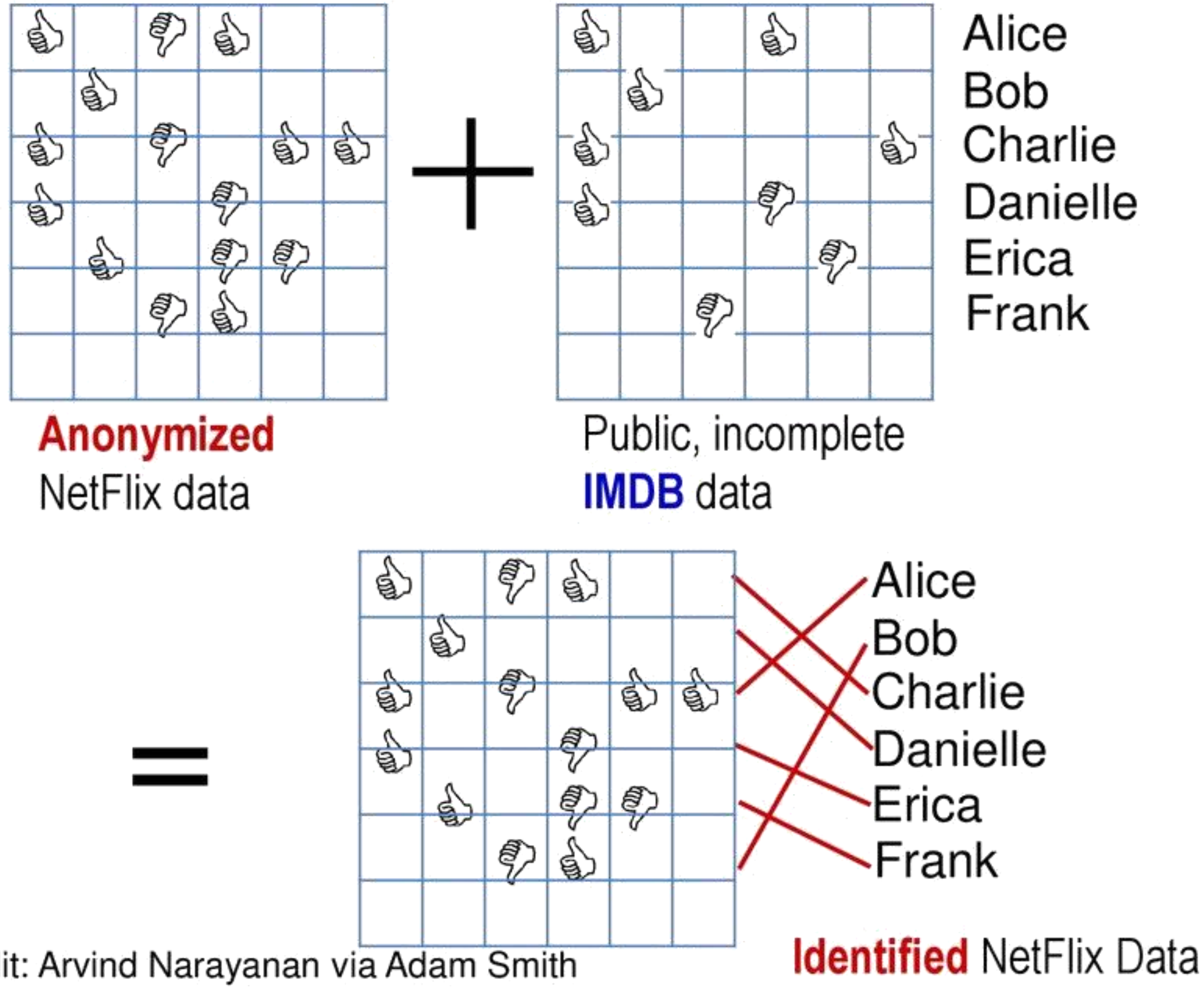
- **Pseudonymisation**: remove directly identifying information
  - *That is often **not** enough!*

- Massachusetts Health records of public employees
  - With the birthdate, ZIP Code, sex: Governor of Massachusetts William Weld uniquely identified
    - Linkage attack with public voting records

  - Other examples: Netflix prize (2006), AOL data, ...

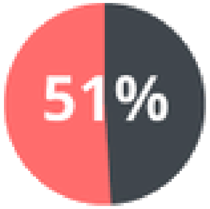


Record Linkage attacks

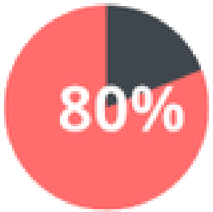


Credit: Arvind Narayanan via Adam Smith

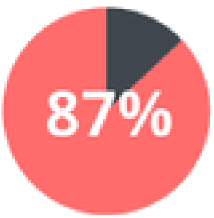
Record linkage attacks



of **mobile phone owners** are re-identified simply by 2 antenna signals, even when coarsened to the hour of the day



of **credit card owners** are re-identified by 3 transactions, even when only merchant and the date of transaction is revealed



of **all people** are re-identified, merely by their date-of-birth, their gender and their ZIP code of residence

L. Sweeney. Simple Demographics Often Identify People Uniquely. 2000

Record linkage

- Finding records that refer to the same entity
  - Across data sets from different sources
  - May or may not share a **common identifier**

- Steps include
  - Preprocessing / normalisation
    - Rule based, hidden Markov models, …
    - Phonetic algorithms, …
  - Some form of identity resolution

DataSet	Name	Date of birth	City of residence
1	William J. Smith	1/2/73	Berkeley, California
2	Smith, W. J.	1973.1.2	Berkeley, CA
3	Bill Smith	Jan 2, 1973	Berkeley, Calif.

Fellegi & Sunter. A Theory for Record Linkage. Journal of the American Statistical Association. 64 (328), 1969

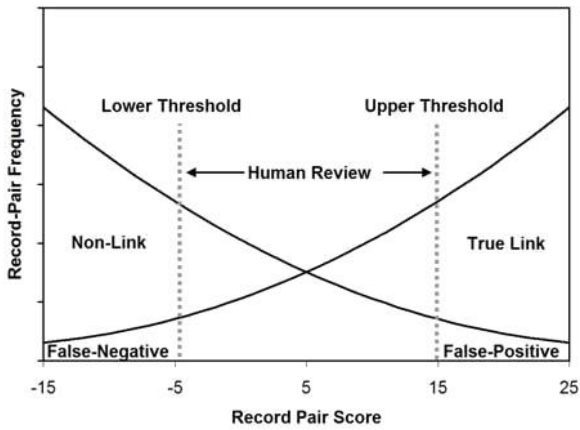
Record linkage

- Deterministic (rules-based) record linkage
  - Links based on the number of individual identifiers that match among the available data sets
  - Records match if all or some identifiers (above a certain threshold) are identical
  - Good option when entities in data sets are identified by a common identifier
    - Or when there are several representative identifiers whose quality of data is relatively high
    - (e.g., name, date of birth, and sex when identifying a person)

Roos & Wajda. Record linkage strategies. Part I: Estimating information and evaluating approaches. Methods of Information in Medicine. 30 (2). 1991

Probabilistic (fuzzy) record linkage

- Takes wider range of potential identifiers into account
- Computes weights for each identifier based on its estimated ability to correctly identify a match/non-match
- Uses weights to calculate probability that two given records refer to the same entity
- Three types of matches
  - Pairs with probabilities above a threshold considered to be matches
  - Pairs with probabilities below another threshold considered to be non-matches
  - Pairs between these thresholds are "possible matches“
    - Can be dealt with e.g., human review



Probabilistic (fuzzy) record linkage

- Algorithms assign match/non-match weights to identifiers by two probabilities *u* and *m*
- *u:* probability that identifier in two non-matching records will agree purely by chance
  - *What is that for the birth month?*
  - 1/12 ≈ 0.083
- *m:* the probability that identifier in matching pairs will agree
  - Or sufficiently similar, e.g. strings with low Levenshtein distance
  - 1.0 in case of perfect data; estimated in practice
    - Based on prior knowledge of the data sets
    - By estimation on a large number of matching and non-matching pairs
    - By iteratively running the algorithm to obtain closer estimations of *m*

Outline

- Privacy: definitions and motivation
- Pseudonimisation
  - Record-Linkage Attack
- Anonymisation
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
- Data watermarking and fingerprinting

Data Sanitisation

- **Anonymisation**: sanitize also quasi-identifiers (QI)
  - Those attributes that can identify when used in combination
  - Birthdate, ZIP Code, sex, occupation, ...
  - *Issues?*
    - List is not complete
    - Case-dependent
      - Adversary's background knowledge!
    - Dependent on the available other data (present **AND** future!)

- Anonymised data is not subject to GDPR regulations anymore!

Data Sanitisation: HIPAA

- The Privacy Rule of the US *Health Insurance Portability and Accountability Act* of 1996 (HIPAA) establishes comprehensive protections for medical privacy *(revised & came into effect 2002)*
- Protected health information (PHI) is “identifiable” health information acquired in the course of serving patients
  - One of the few authoritative sources that lists identifiable attributes
- Sanitisation standard before data sharing in medical domains (research and professional)

Data Sanitisation: 18 HIPAA Identifiers

- Names
- All geographic subdivisions smaller than a State
- All elements of dates (except year)
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- Any other unique identifying number, characteristic, or code

Data Sanitisation: HIPAA

- Massachusetts Health records of public employees
  - With the birthdate, ZIP Code, sex: would the governor be re-identified by applying HIPAA?

- ***k*-anonymity**
  - Each released record should be indistinguishable from at least (k-1) others on its QI attributes
  - Or: cardinality of any query result on released data should be at least k

Outline

- Privacy: definitions and motivation
- Pseudonimisation
  - ➢ Record-Linkage Attack
- Anonymisation
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
- Data watermarking and fingerprinting

k-Anonymity

- Ensures that at least k records have same QI, via
  - Generalisation of values (exact age to a range of values, …)

	QI_1	QI_2	S_1
ID	Age	ZIP	Disease
1	5	15	Flu
2	15	25	Fever
3	28	28	COVID
4	25	15	Fever
5	22	28	Flu
6	32	35	Fever
7	38	32	Flu
8	35	25	COVID

→

Equivalence class

	QI_1	QI_2	S_1
ID	Age	ZIP	Disease
1	0-20	10-30	Flu
2	0-20	10-30	Fever
3	20-30	10-30	COVID
4	20-30	10-30	Fever
5	20-30	10-30	Flu
6	30-40	20-40	Fever
7	30-40	20-40	Flu
8	30-40	20-40	COVID

k-Anonymity

- Ensures that at least k records have same QI, via
  - Generalisation of values (exact age to a range of values, …)
  - Suppression of values

Birthday	Sex	ZIP
21/1/79	M	53715
10/1/79	F	55410
1/10/44	F	90210
21/2/83	M	02274
19/4/82	M	02237

Equivalence class

	Birthday	Sex	ZIP
Group 1	*/1/79	*	5*
	*/1/79	*	5*
suppress	~~1/10/44~~	~~F~~	~~90210~~
Group 2	*/*/8*	M	022*
	//8*	M	022*

k-Anonymity

- Ensures that at least k records have same QI, via
  - Generalisation of values (exact age to a range of values, …)
  - Suppression of values
  - Microaggregation

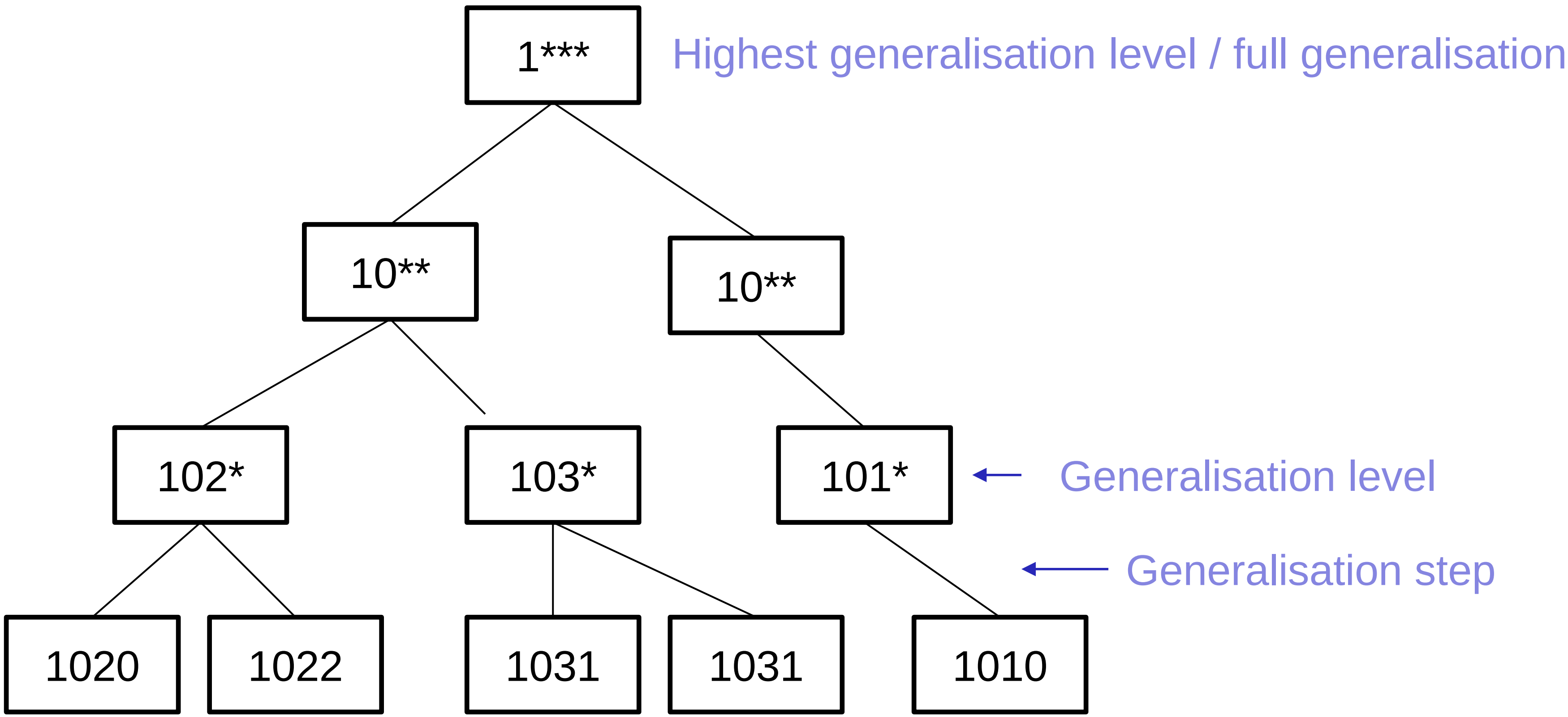
Height	Weight	High Cholestorol
165	72	N
162	74	Y
171	73	N
177	71	N
…	…	…

Equivalence class

	Height	Weight	High Cholestorol
Group 1	**166**	**73**	N
	166	**73**	Y
	166	**73**	N
Group 2	181	70	N
	…	…	…

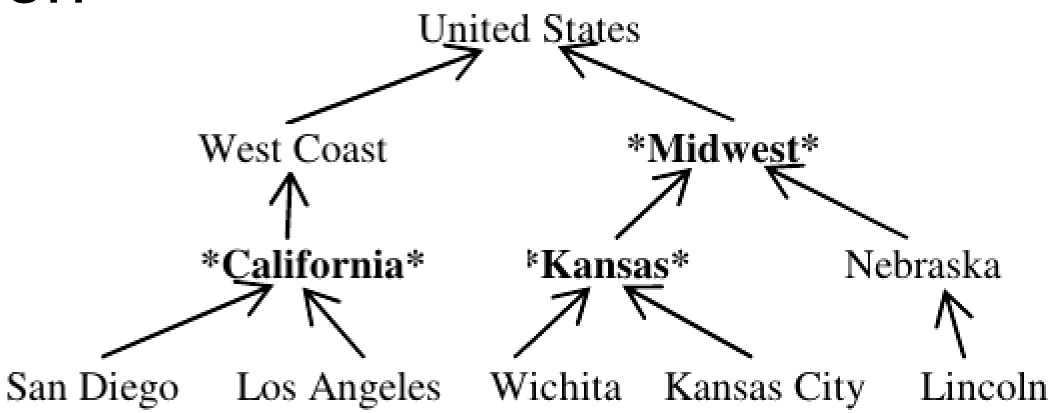
k-Anonymity: hierarchies

- Generalisation is achieved by using a hierarchy
  - Example: ZIP code

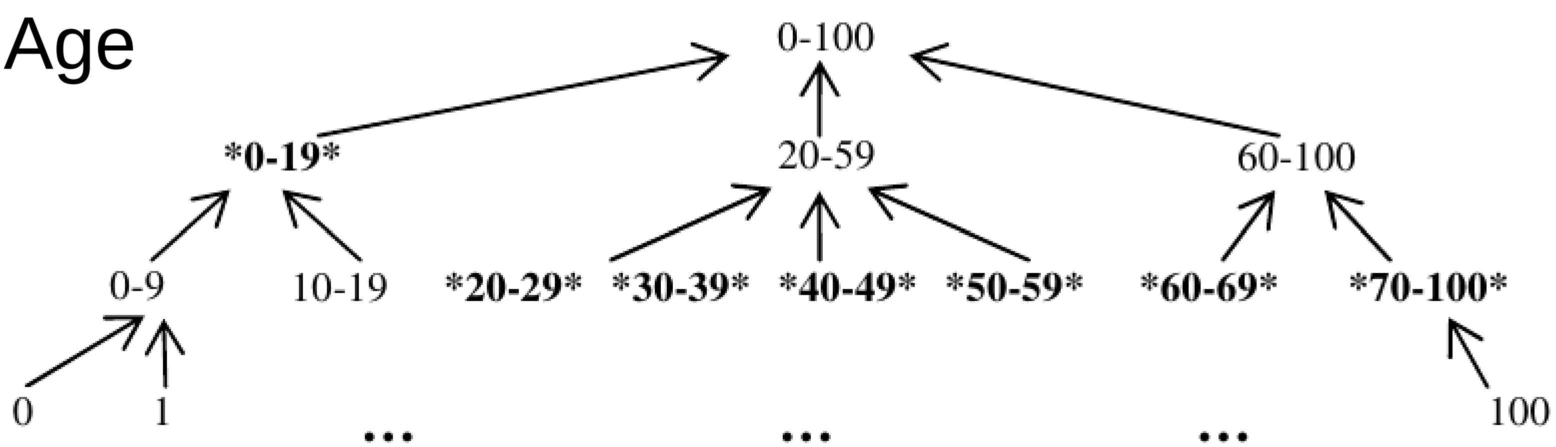


k-Anonymity: hierarchies

– Location



– Age



Global vs Local Transformation

Birthdate	Sex	Zipcode	Disease
*	Male	537**	Flu
*	Male	537**	Broken Arm
*	Male	537**	Bronchitis
*	Female	537**	Hepatitis
*	Female	537**	Sprained Ankle
*	Female	537**	Hang Nail

Global:

All values of the attribute generalized to the same level

Birthdate	Sex	Zipcode	Disease
21.1.'76	Male	537**	Flu
21.1.'76	Male	537**	Broken Arm
*	Female	537**	Hepatitis
*	Female	537**	Hang Nail
*	*	5370*	Bronchitis
*	*	5370*	Sprained Ankle

Local:

Different levels of generalization within a single attribute

Minimal generalisation

Race E_0	ZIP Z_0
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race E_1	ZIP Z_0
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

$\mathbf{GT_{[1,0]}}$

Race E_1	ZIP Z_1
Person	0213*
Person	0213*
Person	0214*
Person	0214*
Person	0213*
Person	0213*
Person	0214*
Person	0214*

$\mathbf{GT_{[1,1]}}$

Race E_0	ZIP Z_2
Black	021**
Black	021**
Black	021**
Black	021**
White	021**
White	021**
White	021**
White	021**

$\mathbf{GT_{[0,2]}}$

Race E_0	ZIP Z_1
Black	0213*
Black	0213*
Black	0214*
Black	0214*
White	0213*
White	0213*
White	0214*
White	0214*

$\mathbf{GT_{[0,1]}}$

Minimal generalisation – generalization (that satisfies k-anonymity) such that it is impossible to lower the anonymity level of any attribute and obtain the same level of anonymity for the database

Methods for k-anonymisation

- **Microaggregation**
  – Data partitioned based on *similarity* of records
  – Aggregation functions applied on data
    - Mean for continuous numerical data
    - Median for categorical data

Age	Sex	Zipcode	Disease
44	Male	53715	Flu
35	Female	53715	Hepatitis
45	Male	53703	Bronchitis
44	Male	53703	Broken Arm
35	Female	53706	Sprained Ankle
45	Female	53706	Hang Nail

Domingo-Ferrer, J., and Vicenç T. "Ordinal, continuous and heterogeneous k-anonymity through microaggregation."

Methods for k-anonymisation

- **Microaggregation**
  - Data partitioned based on *similarity* of records
  - Aggregation functions applied on data
    - Mean for continuous numerical data
    - Median for categorical data

Age	Sex	Zipcode	Disease
44	Male	53703	Flu
38	Female	53706	Hepatitis
44	Male	53703	Bronchitis
44	Male	53703	Broken Arm
38	Female	53706	Sprained Ankle
38	Female	53706	Hang Nail

Domingo-Ferrer, J., and Vicenç T. "Ordinal, continuous and heterogeneous k-anonymity through microaggregation."

k-anonymity: types of attributes

- Direct identifiers
  - SSN, driving licence number, …
- Quasi-identifiers
  - Personal information that can be combined to identify a person
  - Birthdate, zip code, …
- Sensitive attributes
  - Non-identifying sensitive/confidental personal information
  - Health diagnosis, salary, political affiliation …
- Insensitive attributes

k-Anonymity: example results

Record	Name	SSN	Age	Location	Sex	Race	Diagnosis	Income
r_1	Alice	123456789	32	San Diego	M	W	AIDS	17,000
r_2	Bob	323232323	30	Los Angeles	M	W	Asthma	68,000
r_3	Charley	232345656	42	Wichita	M	W	Asthma	80,000
r_4	Dave	333333333	30	Kansas City	M	W	Asthma	55,000
r_5	Eva	666666666	35	Lincoln	F	W	Diabetes	23,000
r_6	John	214365879	20	Lincoln	M	B	Asthma	55,000
r_7	Casey	909090909	25	Wichita	F	B	Diabetes	23,000

Record	Age	Location	Sex	Race
r_1	30-32	California	M	W
r_2	30-32	California	M	W
r_3	30-42	MidWest	*	W
r_4	30-42	MidWest	*	W
r_5	30-42	MidWest	*	W
r_6	20-25	MidWest	*	B
r_7	20-25	MidWest	*	B

Record	Age	Location	Sex	Race
r_1	30-32	California	M	W
r_2	30-32	California	M	W
r_3	25-42	Kansas	*	*
r_4	25-42	Kansas	*	*
r_7	25-42	Kansas	*	*
r_5	20-35	Lincoln	*	*
r_6	20-35	Lincoln	*	*

Solving k-anonymity

- *k*-anonymity problem:
  - Given a dataset R, find a dataset R’ such that:
    - R’ satisfies k-anonymity condition
    - R’ has the maximum utility (minimum information loss)

- Given some data set *R* and a QI *Q*, does *R* satisfy *k*-anonymity over *Q*?
  - Easy to tell in polynomial time
- Finding an **optimal** anonymization is not easy
  - NP-hard: reduction from k-dimensional perfect matching*

➔ Heuristic solutions

A. Meyerson and R. Williams. On the complexity of optimal k-anonymity. In PODS’04.

Solving k-anonymity: Algorithms

- **Datafly**
- **Incognito**
- **SaNGreeA**
- Mondrian
- Flash

Datafly

- Properties:
  - Global (full-domain) generalization algorithm
  - Heuristics: for generalization selects the attribute with the greatest number of distinct values (iteratively until k-anonymity is satisfied)
  - Not necessarily minimal generalization

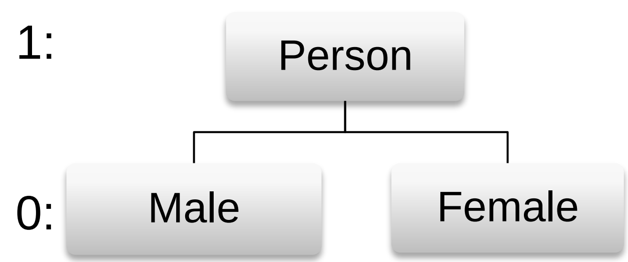
Datafly: example (k=2)

Birthdate	Sex	Zipcode	Disease
21.1.'76	Male	53715	Flu
13.4.'86	Female	53715	Hepatitis
28.2.'76	Male	53703	Bronchitis
21.1.'76	Male	53703	Broken Arm
13.4.'86	Female	53706	Sprained Ankle
28.2.'76	Female	53706	Hang Nail

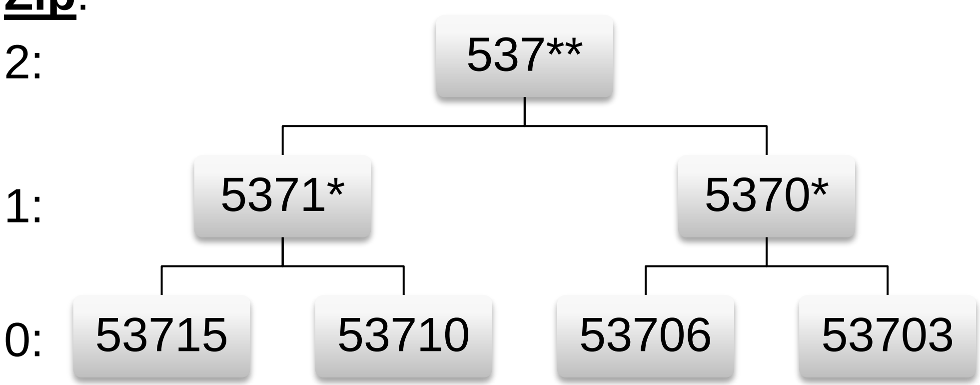
While not 2-anonymous:
generalise the attribute with the greatest number of distinct values

Start → *Birthdate* (or *Zipcode*)

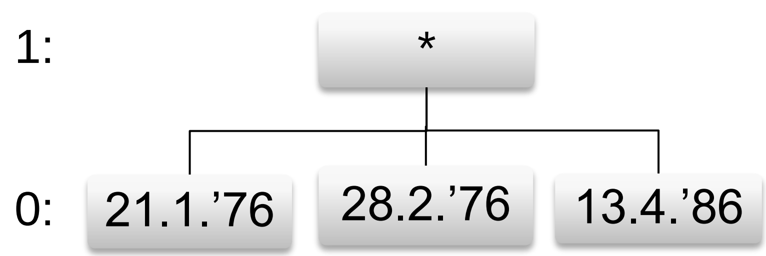
Sex:



Zip:



Birthdate:



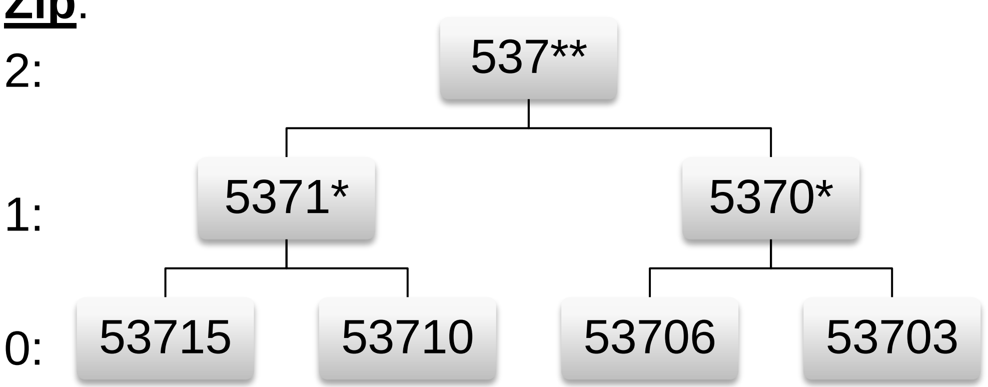
Datafly: example (k=2)

Birthdate	Sex	Zipcode	Disease
*	Male	53715	Flu
*	Female	53715	Hepatitis
*	Male	53703	Bronchitis
*	Male	53703	Broken Arm
*	Female	53706	Sprained Ankle
*	Female	53706	Hang Nail

While not 2-anonymous:
generalise the attribute with the greatest number of distinct values

2-anonymous? NO!

Zip:



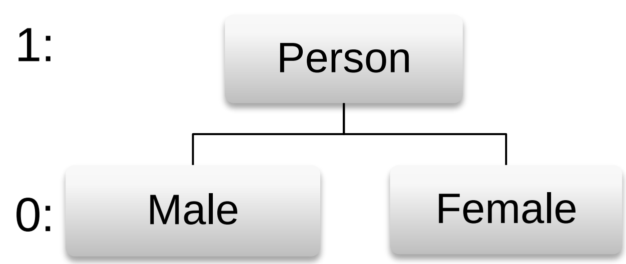
Datafly: example (k=2)

Birthdate	Sex	Zipcode	Disease
*	Male	5371*	Flu
*	Female	5371*	Hepatitis
*	Male	5370*	Bronchitis
*	Male	5370*	Broken Arm
*	Female	5370*	Sprained Ankle
*	Female	5370*	Hang Nail

While not 2-anonymous:
generalise the attribute with the greatest number of distinct values

2-anonymous? NO!

Sex:



Datafly: example (k=2)

Birthdate	Sex	Zipcode	Disease
*	*	5371*	Flu
*	*	5371*	Hepatitis
*	*	5370*	Bronchitis
*	*	5370*	Broken Arm
*	*	5370*	Sprained Ankle
*	*	5370*	Hang Nail

<u>While not 2-anonymous</u>:
generalise the attribute with the greatest number of distinct values

2-anonymous? YES ☺

2-minimal generalization?

Datafly: example (k=2)

Birthdate	Sex	Zipcode	Disease
*	*	5371*	Flu
*	*	5371*	Hepatitis
*	*	5370*	Bronchitis
*	*	5370*	Broken Arm
*	*	5370*	Sprained Ankle
*	*	5370*	Hang Nail

Consider:

Birthdate	Sex	Zipcode	Disease
*	*	53715	Flu
*	*	53715	Hepatitis
*	*	53703	Bronchitis
*	*	53703	Broken Arm
*	*	53706	Sprained Ankle
*	*	53706	Hang Nail

Datafly: example (k=2)

Birthdate	Sex	Zipcode	Disease
*	*	5371*	Flu
*	*	5371*	Hepatitis
*	*	5370*	Bronchitis
*	*	5370*	Broken Arm
*	*	5370*	Sprained Ankle
*	*	5370*	Hang Nail

2-anonymous? YES ☺

2-minimal generalization? NO!

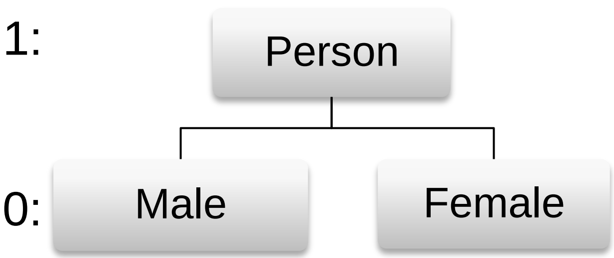
Incognito

- Properties:
  - Generates the set of all possible *k*-anonymous full-domain generalizations of the dataset
  - Iterative bottom-up breadth-first search
  - *k*-minimal generalization
  - Maximizing the number of equivalence classes

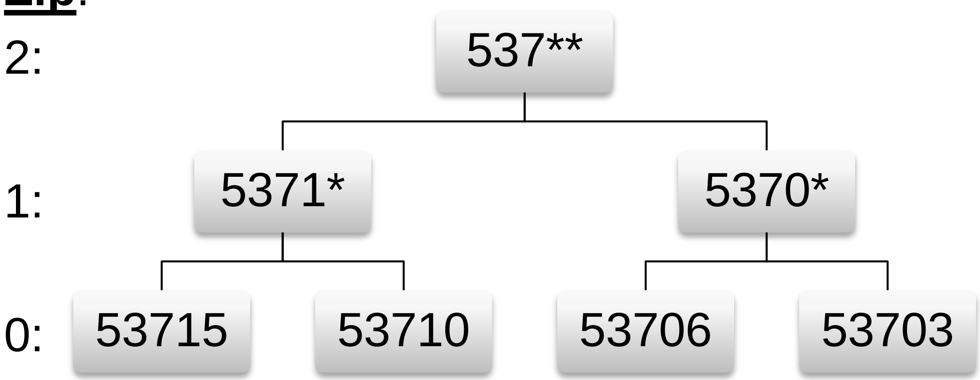
Incognito

Birthdate	Sex	Zipcode	Disease
21.1.'76	Male	53715	Flu
13.4.'86	Female	53715	Hepatitis
28.2.'76	Male	53703	Bronchitis
21.1.'76	Male	53703	Broken Arm
13.4.'86	Female	53706	Sprained Ankle
28.2.'76	Female	53706	Hang Nail

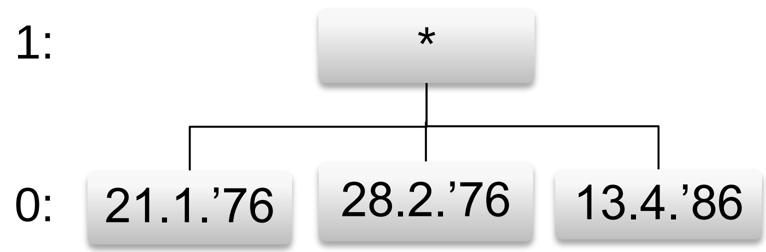
Sex:



Zip:



Birthdate:



Incognito

Birthdate	Sex	Zipcode	Disease
21.1.’76	Male	53715	Flu
13.4.’86	Female	53715	Hepatitis
28.2.’76	Male	53703	Bronchitis
21.1.’76	Male	53703	Broken Arm
13.4.’86	Female	53706	Sprained Ankle
28.2.‘76	Female	53706	Hang Nail

Birth.1

↑

Birth.0

Frequency set:

21.1.’76 : 2
13.4.’86 : 2
28.2.’76 : 2

✓ 2-anonymous with respect to „Birth.0“

Incognito

Birthdate	Sex	Zipcode	Disease
21.1.'76	Male	53715	Flu
13.4.'86	Female	53715	Hepatitis
28.2.'76	Male	53703	Bronchitis
21.1.'76	Male	53703	Broken Arm
13.4.'86	Female	53706	Sprained Ankle
28.2.'76	Female	53706	Hang Nail

Sex1

↑

Sex0

Frequency set:

Male : 3

Female : 3

✓ 2-anonymous with respect to „Sex0“

Incognito

Birthdate	Sex	Zipcode	Disease
21.1.'76	Male	53715	Flu
13.4.'86	Female	53715	Hepatitis
28.2.'76	Male	53703	Bronchitis
21.1.'76	Male	53703	Broken Arm
13.4.'86	Female	53706	Sprained Ankle
28.2.'76	Female	53706	Hang Nail

Zip2

↑

Zip1

↑

Zip0

Frequency set:

53715 : 2
53703 : 2
53706 : 2

✓ 2-anonymous with respect to „Zip0“

Incognito

Birthdate	Sex	Zipcode	Disease
21.1.’76	Male	53715	Flu
13.4.’86	Female	53715	Hepatitis
28.2.’76	Male	53703	Bronchitis
21.1.’76	Male	53703	Broken Arm
13.4.’86	Female	53706	Sprained Ankle
28.2.‘76	Female	53706	Hang Nail

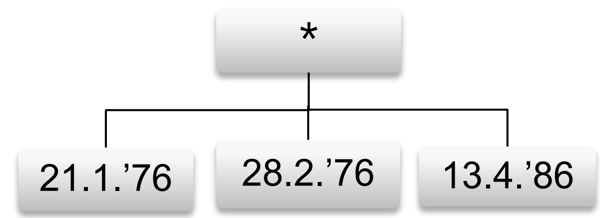
$$
\begin{tikzcd}
 & \langle \text{Birth.1,Sex1} \rangle & \\
\langle \text{Birth.1,Sex0} \rangle \arrow[ur] & & \langle \text{Birth.0,Sex1} \rangle \arrow[ul] \\
 & \langle \text{Birth.0,Sex0} \rangle \arrow[ul] \arrow[ur] &
\end{tikzcd}
$$

Frequency set:

<21.1.‘76, Male> : 2
<13.4.‘86, Female> : 2
<28.2.‘76, Male> : 1
<28.2.‘76, Female> : 1

Incognito

Birthdate	Sex	Zipcode	Disease
21.1.’76	Male	53715	Flu
13.4.’86	Female	53715	Hepatitis
28.2.’76	Male	53703	Bronchitis
21.1.’76	Male	53703	Broken Arm
13.4.’86	Female	53706	Sprained Ankle
28.2.‘76	Female	53706	Hang Nail



<Birth.1,Sex1>

<Birth.1,Sex0> ↗ **<Birth.0,Sex1>** ↖

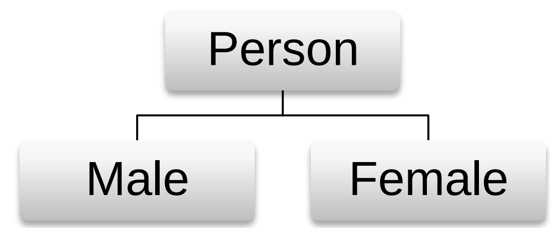
Frequency set:

<*, Male> : 3
<*, Female> : 3

✓ 2-anonymous with respect to <Birth.1,Sex0>

Incognito

Birthdate	Sex	Zipcode	Disease
21.1.'76	Male	53715	Flu
13.4.'86	Female	53715	Hepatitis
28.2.'76	Male	53703	Bronchitis
21.1.'76	Male	53703	Broken Arm
13.4.'86	Female	53706	Sprained Ankle
28.2.'76	Female	53706	Hang Nail



<Birth.1,Sex1>

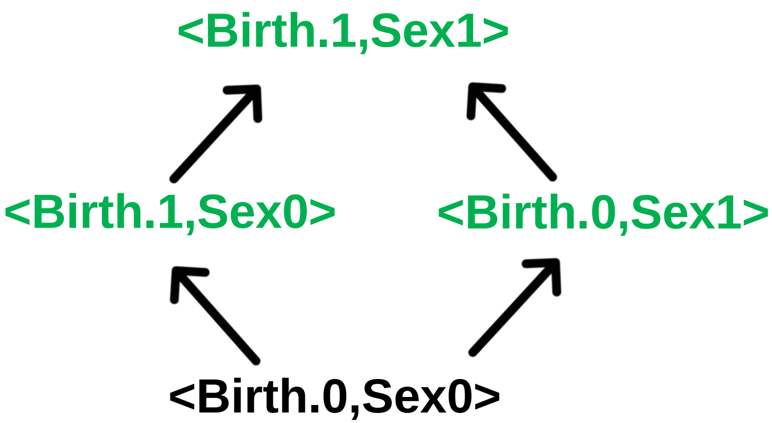
<Birth.1,Sex0> <Birth.0,Sex1>

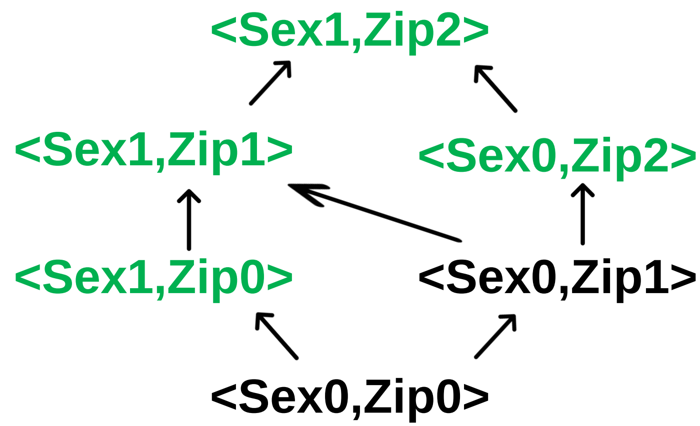
Frequency set:

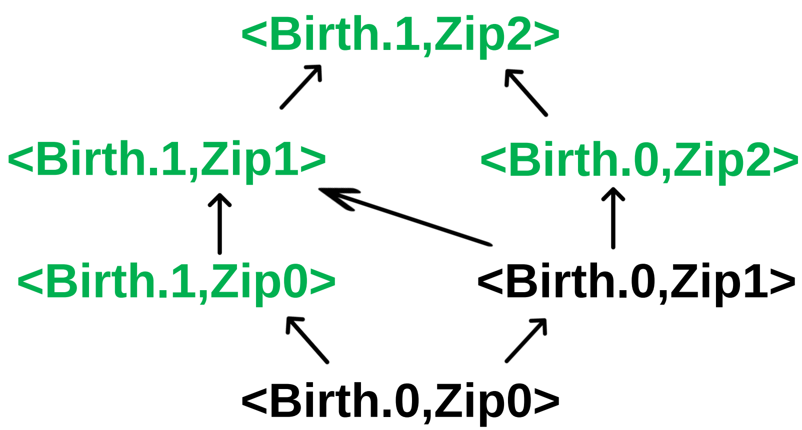
<21.1.'76, Person> : 2
<13.4.'86, Person> : 2
<28.2.'76, Person> : 2

✓ 2-anonymous with respect to <Birth.0,Sex1>

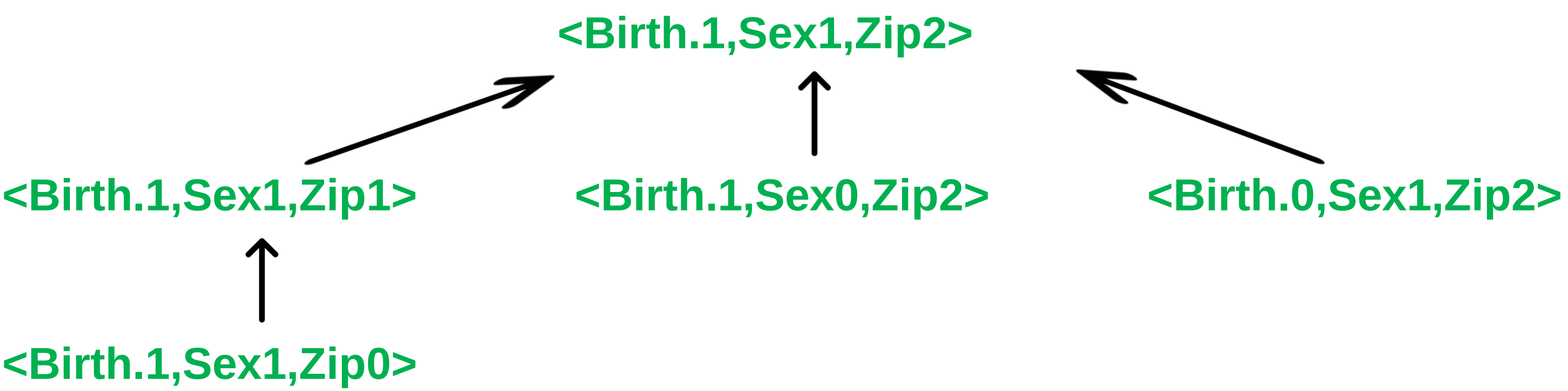
Incognito


<Birth.1,Sex1>
<Birth.1,Sex0>
<Birth.0,Sex1>
<Birth.0,Sex0>


<Sex1,Zip2>
<Sex1,Zip1>
<Sex0,Zip2>
<Sex1,Zip0>
<Sex0,Zip1>
<Sex0,Zip0>


<Birth.1,Zip2>
<Birth.1,Zip1>
<Birth.0,Zip2>
<Birth.1,Zip0>
<Birth.0,Zip1>
<Birth.0,Zip0>

Incognito



Birthdate	Sex	Zipcode	Disease
21.1.’76	Male	53715	Flu
13.4.’86	Female	53715	Hepatitis
28.2.’76	Male	53703	Bronchitis
21.1.’76	Male	53703	Broken Arm
13.4.’86	Female	53706	Sprained Ankle
28.2.‘76	Female	53706	Hang Nail

Incognito

<Birth.1,Sex1,Zip2>

<Birth.1,Sex1,Zip1> <Birth.1,Sex0,Zip2> <Birth.0,Sex1,Zip2>

<Birth.1,Sex1,Zip0>

3 equivalence classes

Birthdate	Sex	Zipcode	Disease
*	*	53715	Flu
*	*	53715	Hepatitis
*	*	53703	Bronchitis
*	*	53703	Broken Arm
*	*	53706	Sprained Ankle
*	*	53706	Hang Nail

Incognito

<Birth.1,Sex1,Zip2>

<Birth.1,Sex1,Zip1>

<Birth.1,Sex0,Zip2>

<Birth.0,Sex1,Zip2>

2 equivalence classes

<Birth.1,Sex1,Zip0>

3 equivalence classes

Birthdate	Sex	Zipcode	Disease
*	Male	537**	Flu
*	Female	537**	Hepatitis
*	Male	537**	Bronchitis
*	Male	537**	Broken Arm
*	Female	537**	Sprained Ankle
*	Female	537**	Hang Nail

Incognito

<Birth.1,Sex1,Zip2>

<Birth.1,Sex1,Zip1>

<Birth.1,Sex0,Zip2>

2 equivalence classes

<Birth.0,Sex1,Zip2>

3 equivalence classes

<Birth.1,Sex1,Zip0>

3 equivalence classes

Out: dataset with the greatest number of equivalence classes

Birthdate	Sex	Zipcode	Disease
21.1.’76	*	537**	Flu
13.4.’86	*	537**	Hepatitis
28.2.’76	*	537**	Bronchitis
21.1.’76	*	537**	Broken Arm
13.4.’86	*	537**	Sprained Ankle
28.2.‘76	*	537**	Hang Nail

SaNGreeA

- Properties:
  - Greedy clustering algorithm
  - User-specified generalization hierarchies for each categorical attribute
  - Numerical attributes are generalized on the fly – no fixed categories needed
- GIL function – measures the amount of generalization

N = set of numerical attributes

$$GIL(cl) = |cl| \cdot \left(\sum_{j=1}^{s} \frac{size(gen(cl)[N_j])}{size(min_{x \in N}(X[N_j]), max_{x \in N}(X[N_j]))} \right)$$

→ „how large is the generalised range compared to the total range of the attribute“

$$+ \sum_{j=1}^{t} \frac{height(A(gen(cl)[C_j]))}{height(H_{C_j})}$$

→ „how many steps up the hierarchy we need to take out of total # of hierarchy levels“

C = set of categorical attributes

SaNGreeA: example (k=2)

	Age	Sex	Zipcode	Disease
t1	43	Male	53715	Flu
t2	35	Female	53715	Hepatitis
t3	32	Male	53703	Bronchitis
t4	43	Male	53703	Broken Arm
t5	28	Female	53706	Sprained Ankle
t6	33	Female	53706	Hang Nail

c1

- Initiate the cluster *c1* with the record *t1*
- Add another record *t* to *c1*:
  - Calculate GIL for each available *t* and *c1*
    - E.g. if we would add *t2* to *c1*, *Age* would need to be generalised to the range [35-43] and *Sex* to *
    - Hence, GIL(c1,t2) = size of range [35-43] / size of total Age range[28-43]
      + #steps taken in Sex gen.hierarcy / #tot. steps in Sex gen.hierarchy
      + #steps in Zipcode gen.hierarchy / #tot. steps in Zipcode gen.hierarchy

    GIL(c1,t2) = 8/15 + 1/1 + 0/2 = 1,53
  - Choose a record with min GIL

SaNGreeA: example (k=2)

	Age	Sex	Zipcode	Disease
t1	43	Male	53715	Flu
t2	35	Female	53715	Hepatitis
t3	32	Male	53703	Bronchitis
t4	43	Male	53703	Broken Arm
t5	28	Female	53706	Sprained Ankle
t6	33	Female	53706	Hang Nail

c1

GIL(c1,t2) = 8/15 + 1/1 + 0/2 = 1,53

GIL(c1,t3) = 11/15 + 0/1 + 2/2 = 1,73

GIL(c1,t4) = 0/15 + 0/1 + 2/2 = 1 → **Min GIL**

GIL(c1,t5) = 15/15 + 1/1 + 2/2 = 3

GIL(c1,t6) = 10/15 + 1/1 + 2/2 = 2,67

SaNGreeA: example (k=2)

	Age	Sex	Zipcode	Disease	
t1	43	Male	537**	Flu	c1
t2	35	Female	53715	Hepatitis	c2
t3	32	Male	53703	Bronchitis	
t4	43	Male	537**	Broken Arm	c1
t5	28	Female	53706	Sprained Ankle	
t6	33	Female	53706	Hang Nail	

- Initiate the next cluster *c2* with the record *t2*
- Add another record *t* to *c2*:
  - Calculate GIL for each available *t* and *c1*
    - E.g. if we would add *t3* to *c2*, *Age* would need to be generalised to the range [32-35], *Sex* to * and *Zipcode* to 537**
    - Hence, GIL(c2,t3) = size of range [32-35] / size of total Age range[28-43]
      + #steps taken in Sex gen.hierarcy / #tot. steps in Sex gen.hierarchy
      + #steps in Zipcode gen.hierarcy / #tot. steps in Zipcode gen.hierarchy

      GIL(c2,t3) = 3/15 + 1/1 + 2/2 = 2,2
  - Choose a record with min GIL

SaNGreeA: example (k=2)

	Age	Sex	Zipcode	Disease	
t1	43	Male	537**	Flu	c1
t2	35	Female	53715	Hepatitis	c2
t3	32	Male	53703	Bronchitis	
t4	43	Male	537**	Broken Arm	c1
t5	28	Female	53706	Sprained Ankle	
t6	33	Female	53706	Hang Nail	c2

GIL(c2,t3) = 3/15 + 1/1 + 2/2 = 2,2

GIL(c2,t5) = 7/15 + 0/1 + 2/2 = 1,47

GIL(c2,t6) = 2/15 + 0/1 + 2/2 = 1,13 → **Min GIL**

SaNGreeA: example (k=2)

	Age	Sex	Zipcode	Disease	
t1	43	Male	537**	Flu	c1
t2	[33,35]	Female	537**	Hepatitis	c2
t3	32	Male	53703	Bronchitis	c3
t4	43	Male	537**	Broken Arm	c1
t5	28	Female	53706	Sprained Ankle	c3
t6	[33,35]	Female	537**	Hang Nail	c2

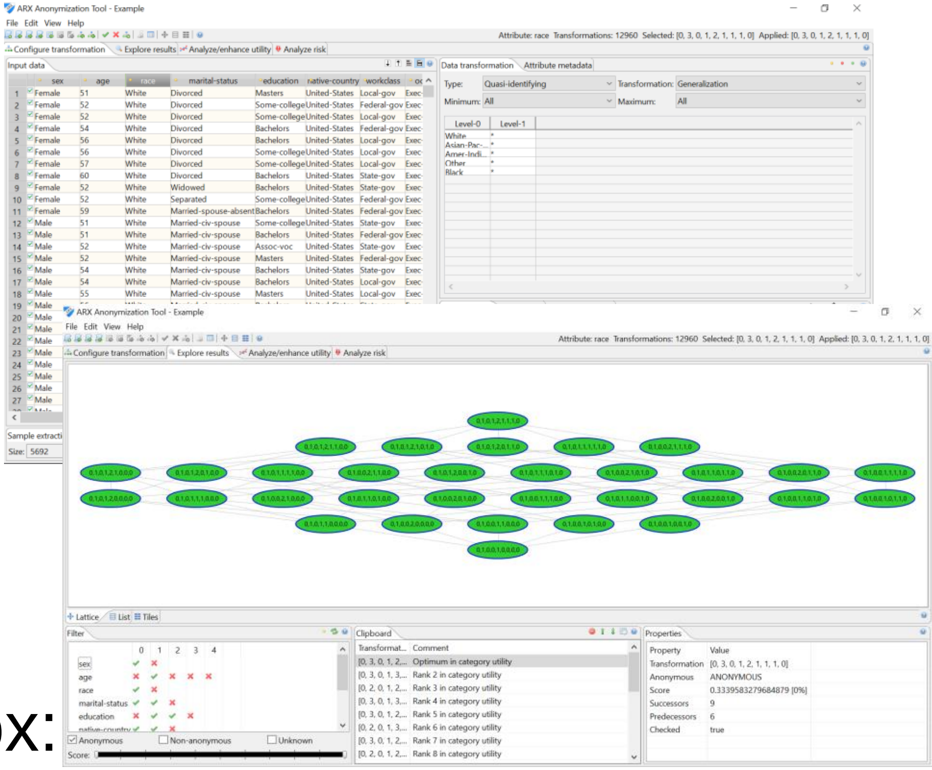
SaNGreeA: example (k=2)

	Age	Sex	Zipcode	Disease	
t1	43	Male	537**	Flu	c1
t2	[33,35]	Female	537**	Hepatitis	c2
t3	[28,33]	*	5370*	Bronchitis	c3
t4	43	Male	537**	Broken Arm	c1
t5	[28,33]	*	5370*	Sprained Ankle	c3
t6	[33,35]	Female	537**	Hang Nail	c2

SaNGreeA: example (k=2)

Age	Sex	Zipcode	Disease
43	Male	537**	Flu
43	Male	537**	Broken Arm
[33,35]	Female	537**	Hepatitis
[33,35]	Female	537**	Hang Nail
[28,33]	*	5370*	Bronchitis
[28,33]	*	5370*	Sprained Ankle

Solving k-anonymity: Tools



- ARX:
  - Flash algorithm
  - https://arx.deidentifier.org/
- Amnesia:
  - https://amnesia.openaire.eu/
- UTD Anonymization Toolbox:
  - Datafly, Incognito, Mondrian
  - http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php
- Microaggregation tool:
  - https://github.com/CrisesUrv/microaggregation-based_anonymization_tool

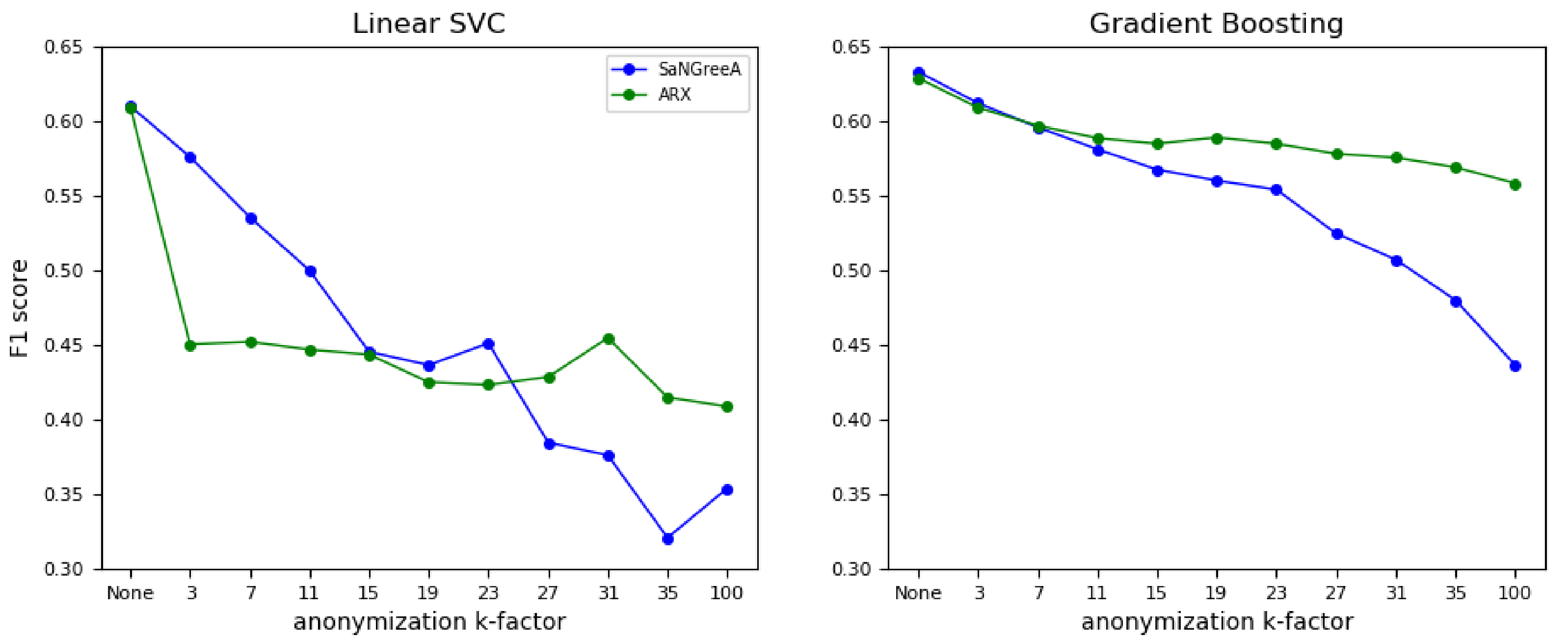
k-Anonymity: Effects on utility

- Two main approaches to evaluate the effect of k-anonymisation on the data utility
  - Measured directly on the data (“information loss metric”)
    - Precision (steps in the hierarchy), Discernibility Metric (how many records can be distinguished), non-uniform entropy, ...
  - Measured by the effect on utility for a certain task/model
    - E.g. Train a machine learning model, and evaluate difference in effectiveness measures

Emam et al., Globally Optimal k-Anonymity for De-Identification of Health Data. 2009

Effects on Utility

- Increasing the level of anonymity, the information loss also increases



- Local (SaNGreeA) vs Global (Flash/ARX) transformation

* Experiments on Adult dataset (target: education-num)

Attacks Against K-Anonymity

- Complementary Release Attack
  - Different releases can be linked together to compromise *k*-anonymity
  - Solution:
    - Consider all of the released tables before release the new one, and try to avoid linking
  - Other data holders may release some data that can be used in this kind of attack.
    - Hard to be prevented completely

Attacks Against K-Anonymity

- k-Anonymity does not provide **privacy** if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeinity Attack

Bob	
Zipcode	**Age**
47678	27

Background Knowledge Attack

Alan	
Zipcode	**Age**
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Breast Cancer
476**	3*	Heart Disease
476**	3*	Breast Cancer
476**	3*	Breast Cancer

Outline

- Privacy: definitions and motivation
- Pseudonimisation
  - Record-Linkage Attack
- Anonymisation
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
- Data watermarking and fingerprinting

L-diversity principles

- Each equivalence class has at least *l* well-represented sensitive values

Bob	
Zipcode	**Age**
47678	27

?

Alan	
Zipcode	**Age**
47673	36

?

A 3-anonymous patient table

Zipcode	Age	Disease
476**	20-40	Heart Disease
476**	20-40	Heart Disease
476**	20-40	Breast Cancer
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Breast Cancer
476**	20-40	Heart Disease
476**	20-40	Heart Disease
476**	20-40	Breast Cancer

Machanavajjhala, Gehrke & Kifer. l-Diversity: Privacy Beyond k-Anonymity. 2006

L-diversity principles

- L-diversity principle:
  - A q-block (equivalence class) is l-diverse if contains at least *l* ‘well represented” values for the sensitive attribute S
  - A table is *l*-diverse if every q-block is *l*-diverse
  - Different variations: distinct, entropy, recursive ***l***-diversity

Machanavajjhala, Gehrke & Kifer. l-Diversity: Privacy Beyond k-Anonymity. 2006

l-Diversity: variations

- Distinct *l*-diversity
  - Each equivalence class has at least *l* well-represented sensitive values
  - Limitation:
    - Doesn't prevent a ***probabilistic inference attack***
    - Example
      - 10 tuples in one equivalent class
      - The "Disease" variable contains one "Flu", one "Heart Disease", and eight "Cancer"
      - This satisfies ***3***-diversity, but an attacker can still affirm that the target person's disease is "***Cancer***" with the accuracy of 80%.

#	Zipcode	Age	Disease
1	476**	2*	Cancer
2	476**	2*	Flu
3	476**	2*	Cancer
4	476**	2*	Cancer
5	476**	2*	Cancer
6	476**	2*	Cancer
7	476**	2*	Cancer
8	476**	2*	Heart Disease
9	476**	2*	Cancer
10	476**	2*	Cancer

l-Diversity: variations

- Entropy *l*-diversity
  - Each equivalence class not only must have enough different sensitive values, but also the different sensitive values must be distributed evenly enough.
  - It means the entropy of the distribution of sensitive values in each equivalence class is at least $\log_2(l)$

$$H(X) = \mathrm{E}(I(X)) = \sum_{i=1}^{n} p(x_i) I(x_i) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

  - Sometimes too restrictive – when some values are very common, entropy of the entire table may be very low

l-Diversity: variations

- Recursive *(c,l)*-diversity
  - Less conservative notion
  - “The most frequent value does not appear too “frequently
  - s_1, …. s_m possible values of attribute in a q-block
  - $n_{(q,sm)}$ = count of that value
    - sorted descending & referred to as r_1 .. r_m
  - A q-block is (c,2) diverse if, for a specified c:
    - $r_1 < c(r_2 + \ldots + r_m)$
  - Recursively (if more than two sensitive values)
    - $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$

Limitations of l-Diversity

- l-diversity may be **difficult** or **unnecessary**

- Example: a single sensitive attribute
  - Two values: HIV positive (1%) and HIV negative (99%)
  - Very different degrees of sensitivity

  - *l*-diversity may be unnecessary
    - 2-diversity is unnecessary for an equivalence class that contains only negative records

  - l-diversity is difficult to achieve
    - Suppose there are 10000 records in total
    - To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes

Limitations of l-Diversity

- l-diversity is insufficient to prevent attribute disclosure

Bob	
Zip	***Age***
47678	27

Zipcode	Age	Salary	Disease
476**	2*	3K	Gastric Ulcer
476**	2*	4K	Gastritis
476**	2*	5K	Stomach Cancer
4790*	≥40	6K	Gastritis
4790*	≥40	11K	Flu
4790*	≥40	8K	Bronchitis
476**	3*	7K	Bronchitis
476**	3*	9K	Pneumonia
476**	3*	10K	Stomach Cancer

Similarity Attack

- Conclusions:
  - Bob's salary is in [3k,5k], which is relatively low
  - Bob has some stomach-related disease
- l-diversity does not consider semantic meanings of sensitive values

Outline

- Privacy: definitions and motivation
- Pseudonimisation
  - ➢ Record-Linkage Attack
- Anonymisation
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
- Data watermarking and fingerprinting

t-closeness

- ***k*-anonymity** prevents identity disclosure but not attribute disclosure
- To solve that problem ***l*-diversity** requires that each eq. class has at least *l* values for each sensitive attribute
- ***t*-closeness** requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the attribute in the overall table

t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. Li et al., 2007

t-closeness

Bob	
Zip	***Age***
47678	27

?

Similarity Attack

Zipcode	Age	Salary	Disease
476**	<40	3K	Gastric Ulcer
476**	<40	9K	Pneumonia
476**	<40	5K	Stomach Cancer
4790*	≥40	6K	Gastritis
4790*	≥40	11K	Flu
4790*	≥40	8K	Bronchitis
476**	<40	7K	Bronchitis
476**	<40	4K	Gastritis
476**	<40	10K	Stomach Cancer

- Privacy = information gain of an observer
- Distribution of the sensitive attribute in each equivalence class should be *similar* to distribution of the sensitive attribute in the whole table

t-closeness

- Privacy is measured by the information gain of an observer
- Information Gain = (Posterior Belief – Prior Belief)
- Q = the distribution of the sensitive attribute in the whole table
- P = the distribution of the sensitive attribute in equivalence class

t-closeness Principle

- An equivalence class is said to have t-closeness
  – If the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold *t*

- A table is said to have *t*-closeness
  – If all equivalence classes have t-closeness.

Distance between two distributions

- Given two distributions
  - $P = (p_1, p_2, ..., p_m)$
  - $Q = (q_1, q_2, ..., q_m)$,
- Variational distance: $$\mathrm{D}[\mathbf{P},\mathbf{Q}] = \sum_{i=1}^{m} \frac{1}{2}|p_i - q_i|.$$
- Earth Movers Distance:

$$\mathrm{D}[\mathbf{P},\mathbf{Q}] = \frac{1}{2}\sum_{i=1}^{m}|p_i - q_i| = \sum_{p_i \geq q_i}(p_i - q_i) = -\sum_{p_i < q_i}(p_i - q_i)$$

  - (Or something else..)

Similarity Attack Example

	ZIP Code	Age	Salary	Disease
1	4767*	≤ 40	3K	gastric ulcer
3	4767*	≤ 40	5K	stomach cancer
8	4767*	≤ 40	9K	pneumonia
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
2	4760*	≤ 40	4K	gastritis
7	4760*	≤ 40	7K	bronchitis
9	4760*	≤ 40	10K	stomach cancer

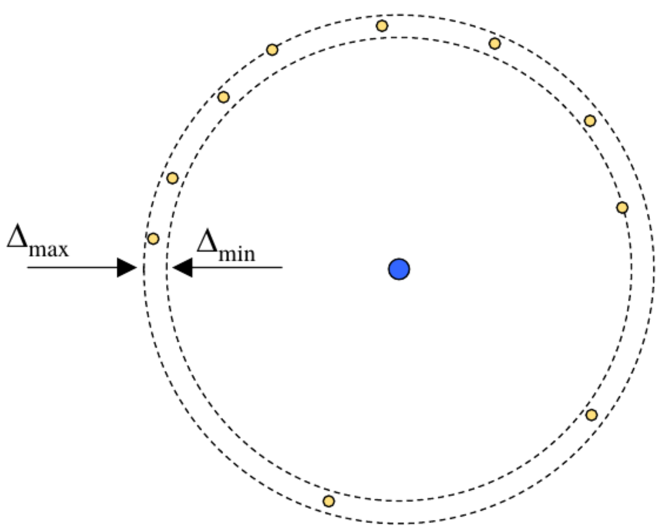
- 0.167-closeness for *Salary* and 0.278-closeness for *Disease*

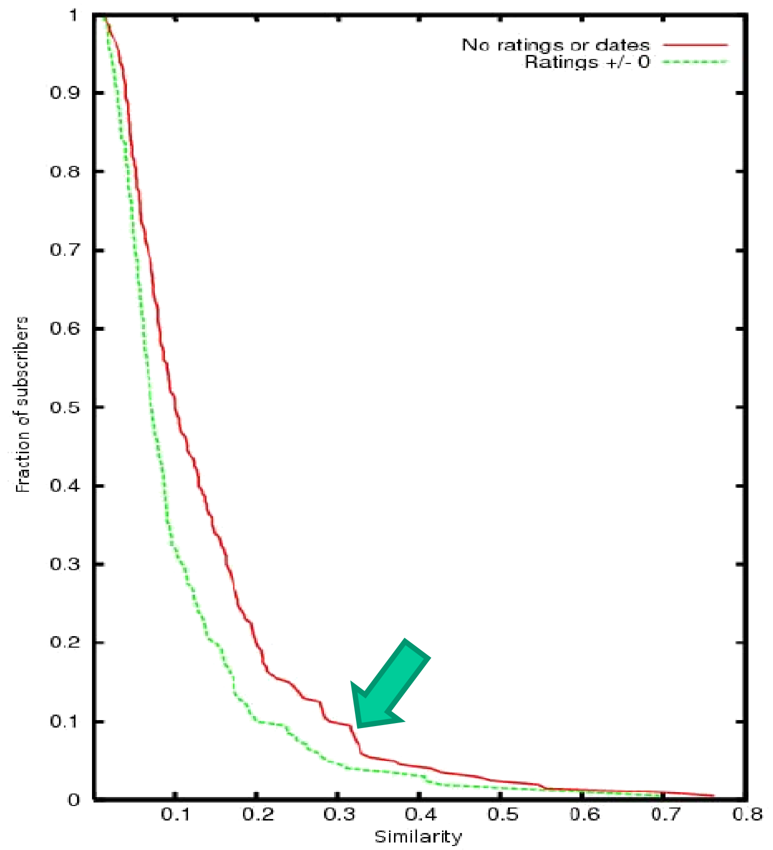
l-t: Conclusion

- *l*-diversity and *t*-closeness add additional guarantees for the privacy of the individuals
- They however further limit the data utility
- Search for *k/l/t*-minimal distortion more complex
- Adds two more parameters to set – which values??

Anonymisation: other limitations

- In very high-dimensional spaces data matrices often get very sparse
  - ➔ Only a few items are actually similar to each other

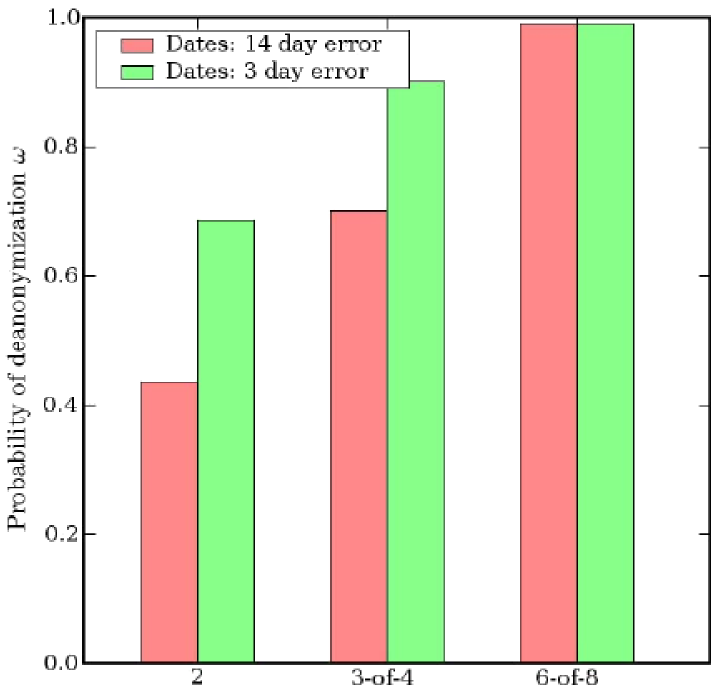


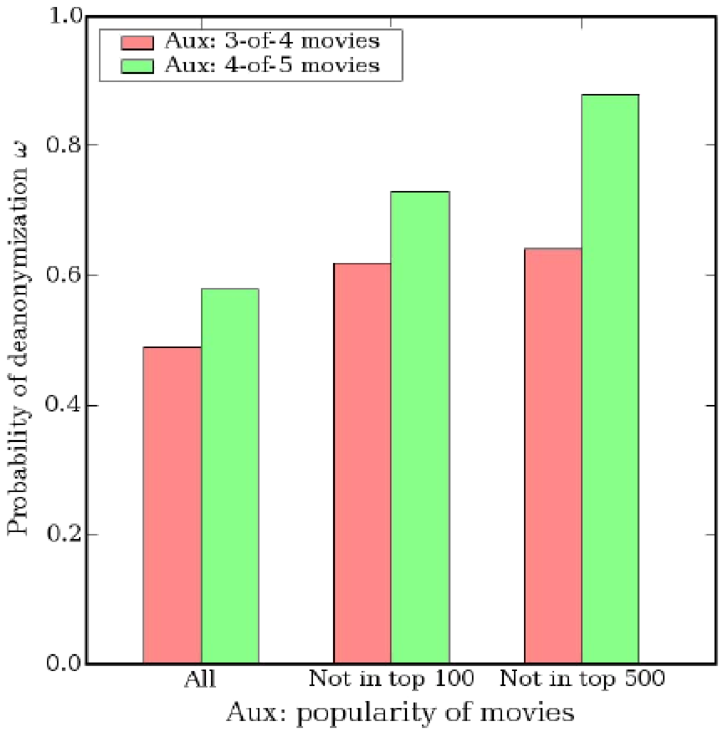


Robust De-anonymization of Large Sparse Datasets. Arvind Narayanan and Vitaly Shmatikov. 2008

Anonymisation: other limitations

- In very high-dimensional spaces data matrices often get very sparse
  - Makes re-identification easier





Robust De-anonymization of Large Sparse Datasets. Arvind Narayanan and Vitaly Shmatikov. 2008

Anonymisation: conclusion

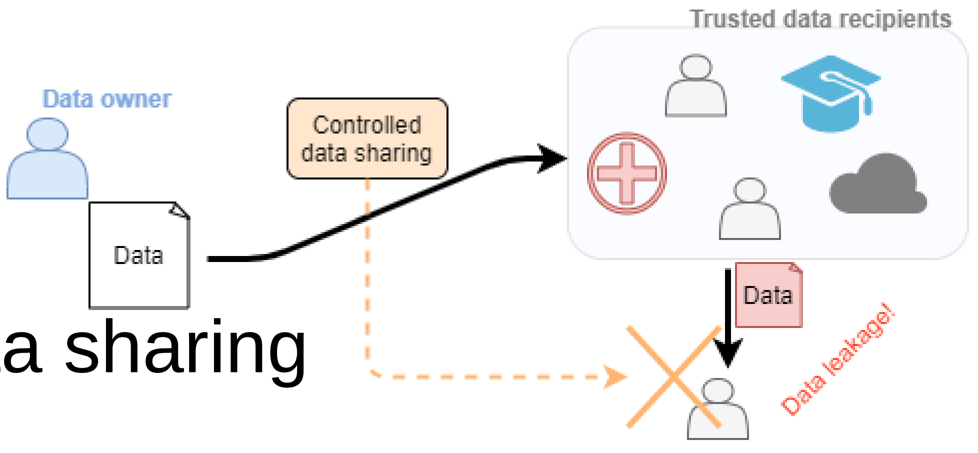
- Approaches like k/l/t-* prevent certain types of attacks
  - Identification, background, similarity, ...
  - Has effects on the data utility
  - It is difficult to assess what other data is available
  - It is not clear what a required level for ***k*** is
- Still
  - There aren’t many alternatives around
    - Differential privacy the one likely most often mentioned
  - Still frequently used approach when you need to publish data to the “public”
  - Makes it more GDPR compliant

Outline

- Privacy: definitions and motivation
- Pseudonimisation
  - Record-Linkage Attack
- Anonymisation
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
- Data watermarking and fingerprinting

Digital property protection: motivation

- Why protecting the data?
  - Data owner used a lot of resources to collect/create the data (money, human experts, time…)
  - Sensitive data (e.g. medical data) needs to be shared with researchers
    - Privacy implications: only the trusted parties get the data and should not share it further



- The goal: controlled data sharing
  - Share full data
  - Trace the unauthorised data re-distribution

Data fingerprinting and watermarking

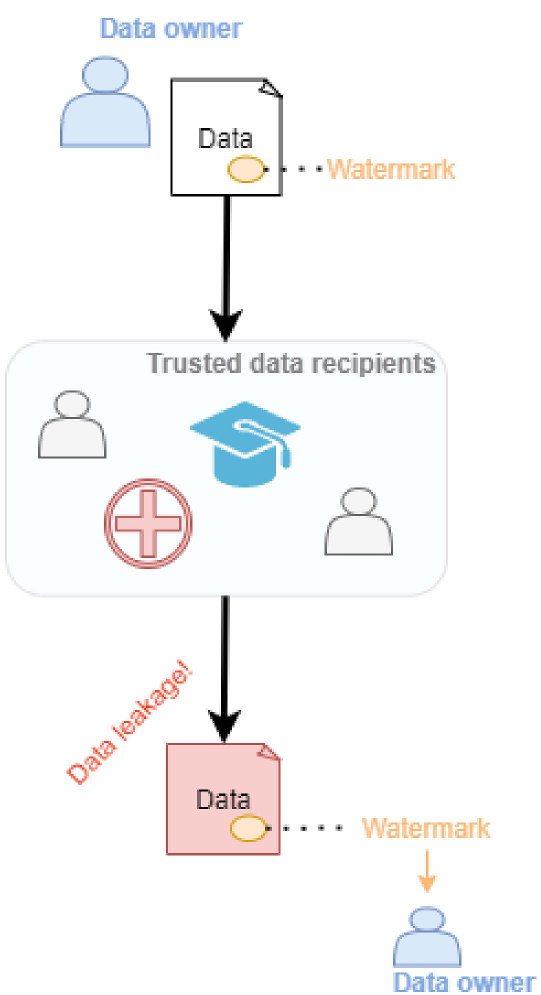
- Embedding owner‘s signature into the data
  - Applying tailored modifications to the data which only the owner is able to extract

Age	Blood Pressure	Diabetes
32	64	1
31	66	0
50	72	1
48	70	0

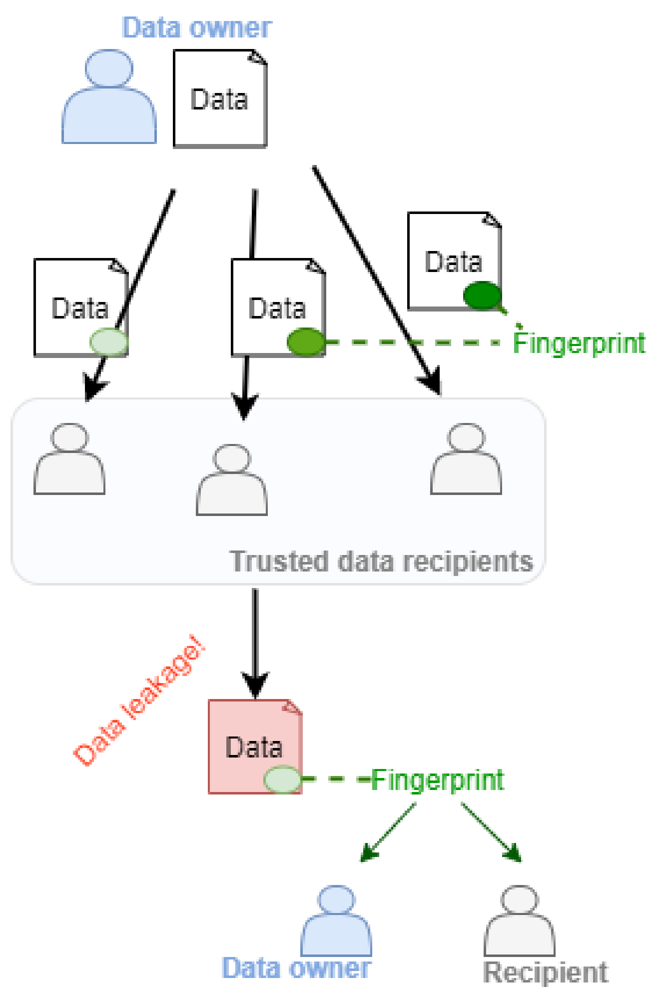
Age	Blood Pressure	Diabetes
33	64	1
31	**68**	0
50	72	1
47	70	0

Watermarking vs. fingerprinting

Watermark: identifies the owner



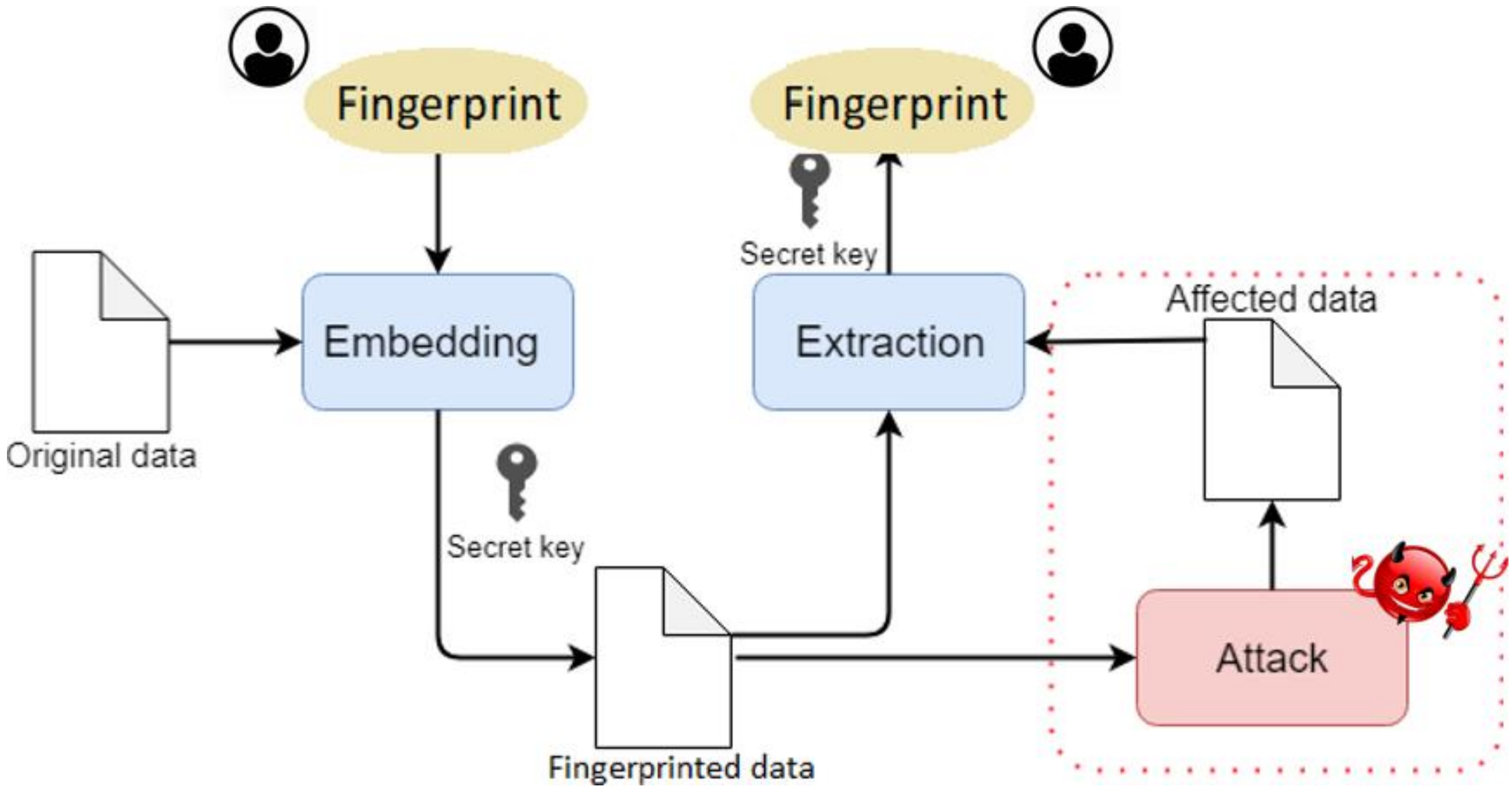
Fingerprint: owner & recipient



Fingerprinting – (a bad) example

https://twitter.com/pnikosis/status/1592823543498436611

The workflow

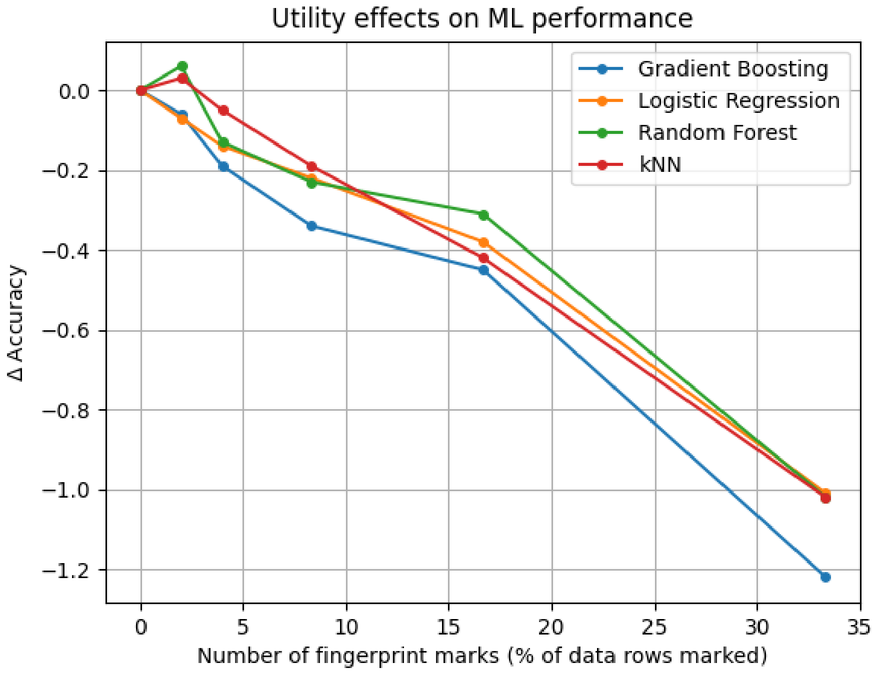


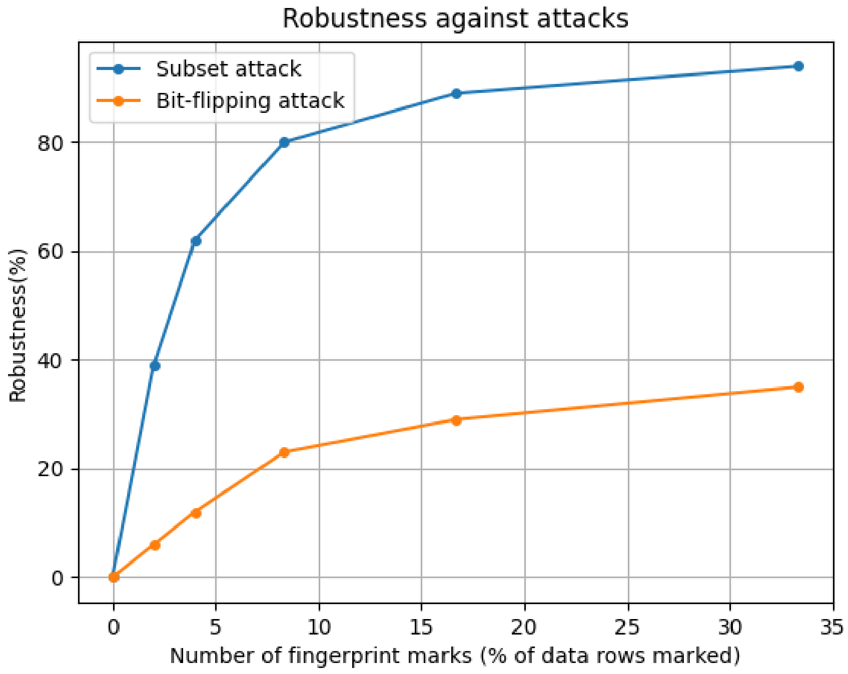
The schemes: fingerprinting

- Owner's secret key used for
  - Fingerprint creation
  - Embedding pattern
- Create distinct fingerprint for each data recipient
  - Fingerprint = bitstring (output of a hash function seeded by the secret key)
- Embed the fingerprint bits following the embedding pattern:
  - Pseudorandom number generator seeded by the secret key outputs the locations in the dataset to be modified with fingerprint bits
    - bit ($location_i$) = bit ($location_i$) x $fingerprint_i$ (if fingerprint=1 ➔ change)
- Fingerprint extraction: reverse insertion (possible only by knowing the secret key!)

Robustness vs utility

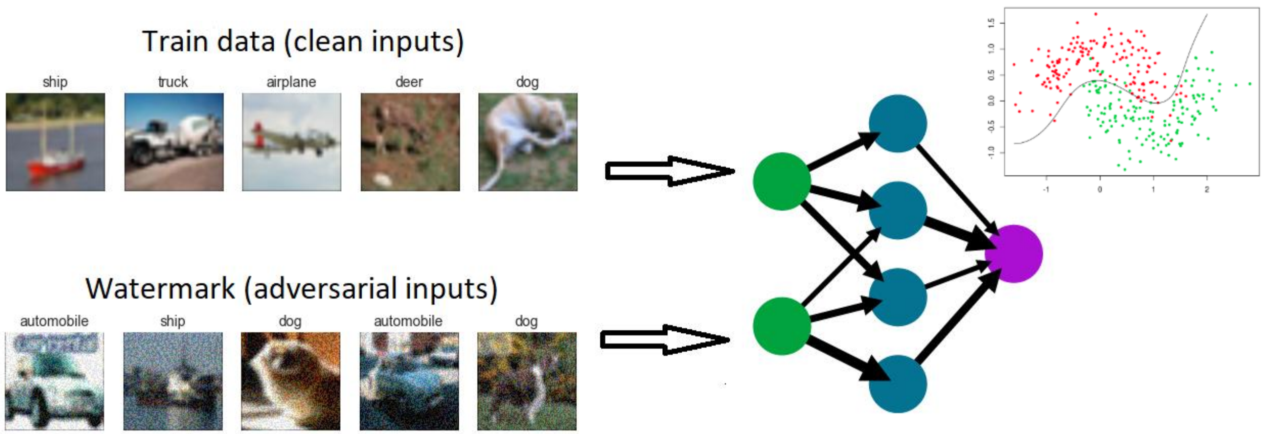
- Robustness against attacks -> maximise modifications
- Preserve data utility! -> minimise modifications
  - Trade-off!





Watermarking ML/DL models

- Protecting the ownership of ML/DL models
- The same idea: Embedd the owner‘s signature into the model
  - E.g. modify decision boundary of DNN by learning specifically tailored input data (adversarial input)



- More about this later in adversarial ML lecture! ☺

WM & FP: Conclusions

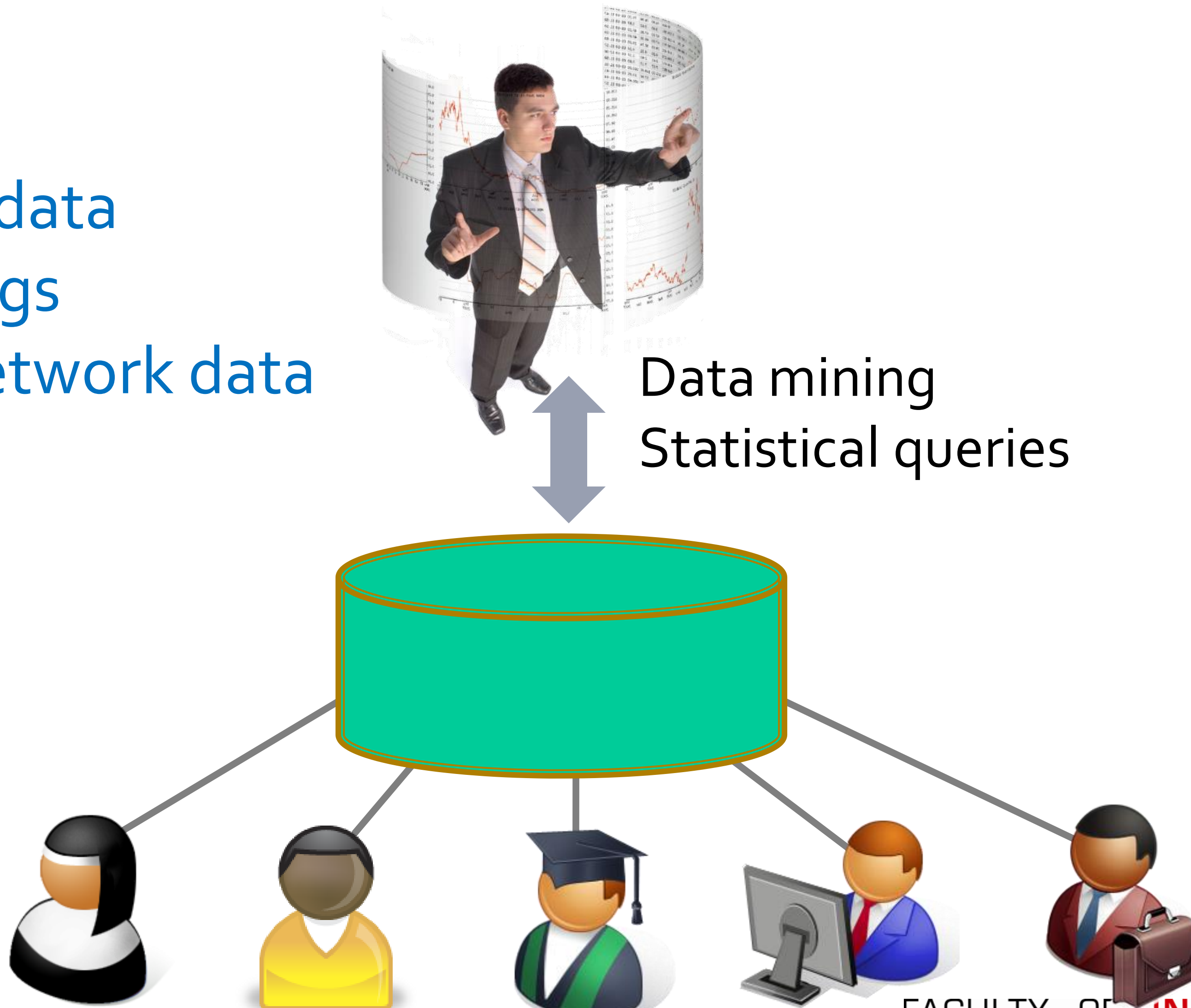
- Watermarking and fingerprinting allow sharing the data with a possibility of:
  - Ownership verification
  - Identification of unauthorised usage of data (only fingerprint)
- Requires modifying the data
- Robustness of a fingerprint vs. data utlity:
  - Stronger fingerprints decrease the utility more

Outline

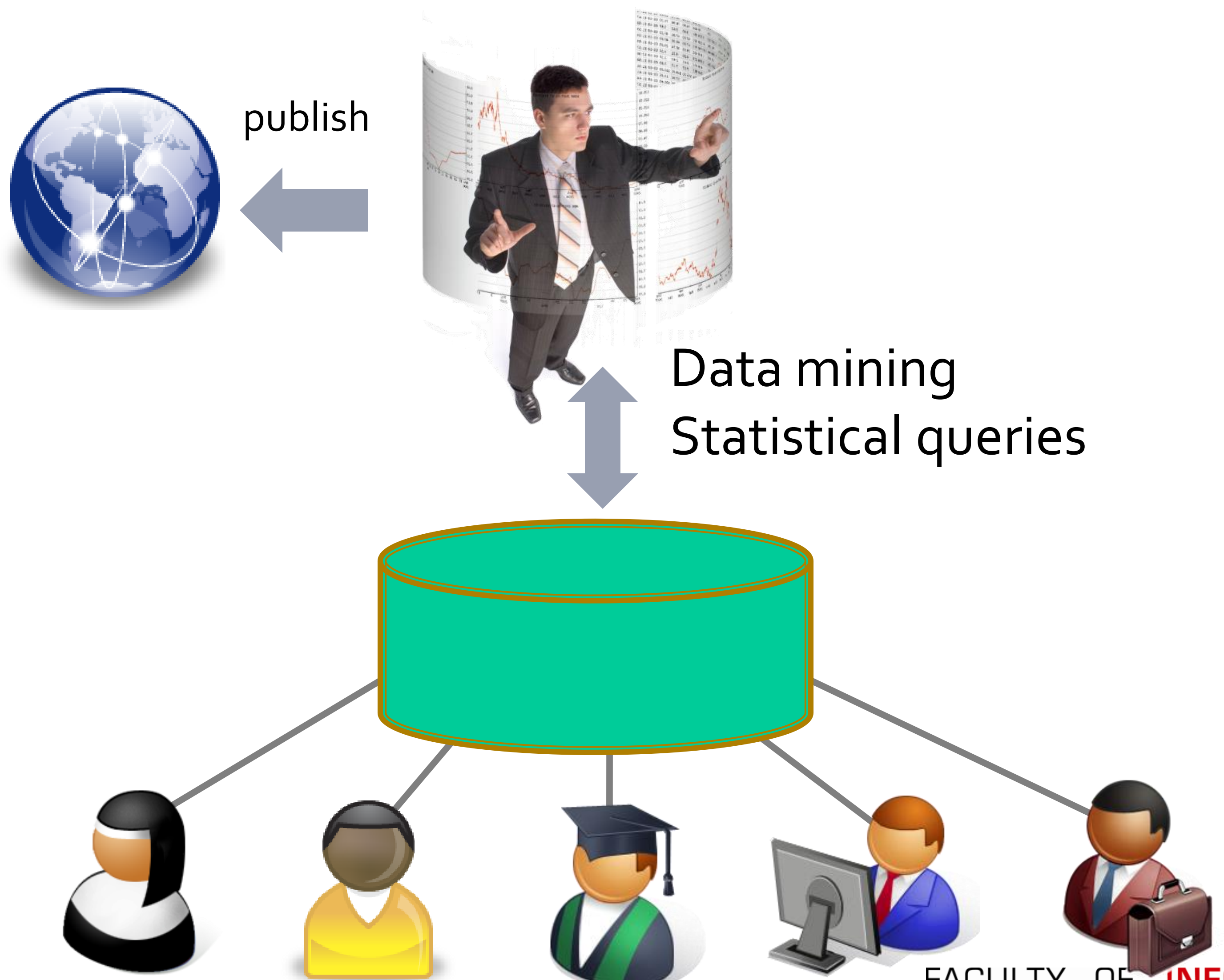
- Privacy: definitions and motivation
- Pseudonimisation
  - Record-Linkage Attack
- Anonymisation: setting & threat models
  - *k*-anonymity
  - *l*-diversity
  - *t*-closeness
- Data watermarking and fingerprinting

General Setting

Medical data
Query logs
Social network data
...



General Setting



General Threats to Privacy

- Identity disclosure
- Attribute disclosure
- Membership disclosure

*(assuming **data** being published/analysed; other threats applicable e.g. if models are published / distributed, e.g. model inversion)*

General Threats to Privacy

- Identity disclosure (or re-identification)
  - Means that an individual can be linked to a specific data entry

ID	Birthdate	Sex	Salary
?	02.03.1995	Male	3 950€
?	03.04.2006	Male	2 870€
?	01.02.1994	Female	3 720€

Tanja

- Attribute disclosure
- Membership disclosure

General Threats to Privacy

- Identity disclosure

- Attribute disclosure
  - May be achieved even without linking to a specific item in a dataset
  - Discloses sensitive attributes from the dataset with which individuals are not willing to be linked with, e.g. the salary of a person
  - Possible when **knowing values of some attributes of a record**

ID	Birthdate	Sex	Education	Salary
Tom	01.02.1984	F	Tertiary	?
Tanja	02.03.1995	M	Secondary	?

→

Salary
4 720€
3 950€

- Membership disclosure

General Threats to Privacy

- Identity disclosure

- Attribute disclosure

- Membership disclosure
  - Inference allows an attacker to determine whether or not data about an individual is contained in a dataset
  - Does not directly disclose any information from the dataset itself
    ➔ but may allow an attacker to infer meta-information
  - Deals with implicit sensitive attributes: attributes of an individual that are not contained in the dataset, but are globally true for all/most records in the dataset

General Threats to Privacy

- Identity disclosure (or re-identification)
  - Means that an individual can be linked to a specific data entry

Tanja ➔

ID	Birthdate	Sex	Salary
?	02.03.1995	Male	3 950€
?	03.04.2006	Male	2 870€
?	01.02.1994	Female	3 720€

  - From the identification it also follows that an attacker can learn all sensitive information contained in the data entry about the individual

    ➔ **automatically** leads to **attribute** and **membership** disclosure

- Attribute disclosure

- Membership disclosure

General Threats to Privacy

- Identity disclosure
- Attribute disclosure
- Membership disclosure

➔ **Which methods discussed last week counter which disclosure type(s)?**

Questions ?