

ARX – Anonymizing Data in Theory and Practice

Prof. Dr. Fabian Prasser

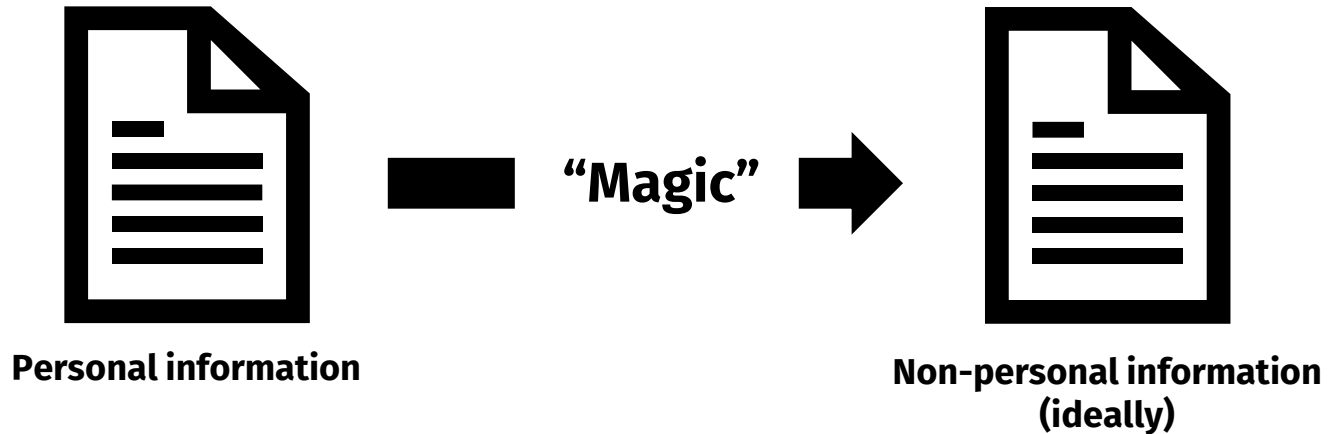


Outline

- 1. Recap (cf. workshop “Reducing risk: An introduction to data anonymization” by Kristi Thompson) and background on risk-based anonymization**
- 2. Adversarial modelling**
- 3. Introduction to the ARX Data Anonymization Tool**
- 4. Practical exercises**
 - Basic project setup
 - Anonymisation, risk and utility assessment
 - Using different transformation methods
 - Advanced anonymization techniques
- 5. Recommended readings**
- 6. Q & A**

1. Background on risk-based anonymization

Basic concept of data anonymization



Anonymization of tabular data

Age	Sex	ZIP	Weight	Diagnosis
55	Male	81539	71	C25.0 Malignant neoplasm of head of pancreas
76	Male	81675	80	C25.0 Malignant neoplasm of head of pancreas
66	Male	81929	85	C25.0 Malignant neoplasm of head of pancreas
81	Male	80802	79	C25.1 Malignant neoplasm of body of pancreas
74	Male	81249	88	C25.2 Malignant neoplasm of tail of pancreas
71	Female	80335	69	C18.2 Malignant neoplasm of ascending colon
64	Female	80339	71	C18.4 Malignant neoplasm of transverse colon
69	Male	80637	75	C18.7 Malignant neoplasm of sigmoid colon
55	Female	80638	77	C18.7 Malignant neoplasm of sigmoid colon
61	Male	81667	67	C18.7 Malignant neoplasm of sigmoid colon

Age	Sex	ZIP	Weight	Diagnosis
72,0	Male	81***	[80, 90[C25.- Malignant neoplasm of pancreas
72,0	Male	81***	[80, 90[C25.- Malignant neoplasm of pancreas
72,0	Male	81***	[80, 90[C25.- Malignant neoplasm of pancreas
62,7	---	80***	[70, 80[C18.- Malignant neoplasm of colon
62,7	---	80***	[70, 80[C18.- Malignant neoplasm of colon
62,7	---	80***	[70, 80[C18.- Malignant neoplasm of colon

Sampling

Aggregation

Deletion

Masking

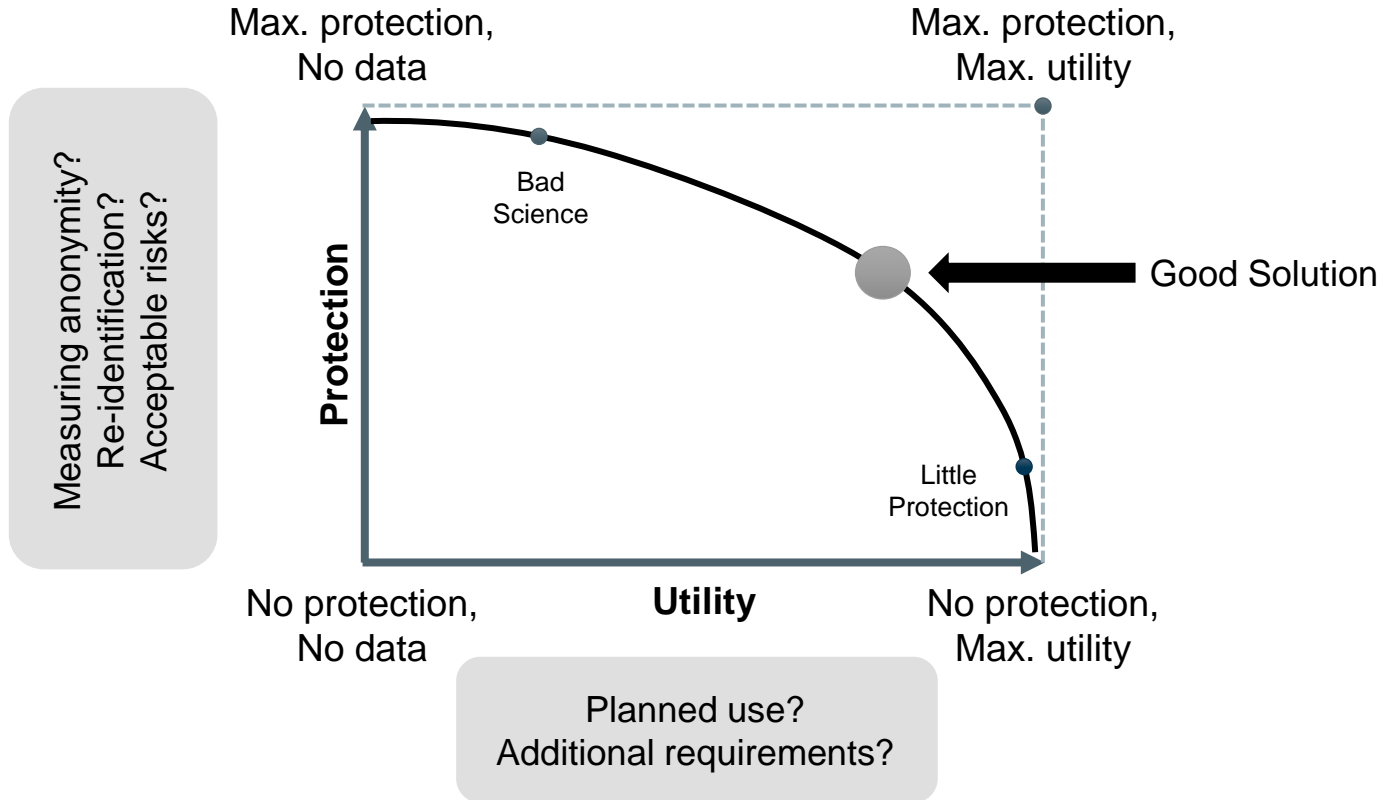
Categorization

Generalization



→ k-Anonymity (with k=3) and (ϵ, δ) -Differential Privacy (with $\epsilon \approx 0.92$ and $\delta \approx 0.22$)

Risk-utility trade-off

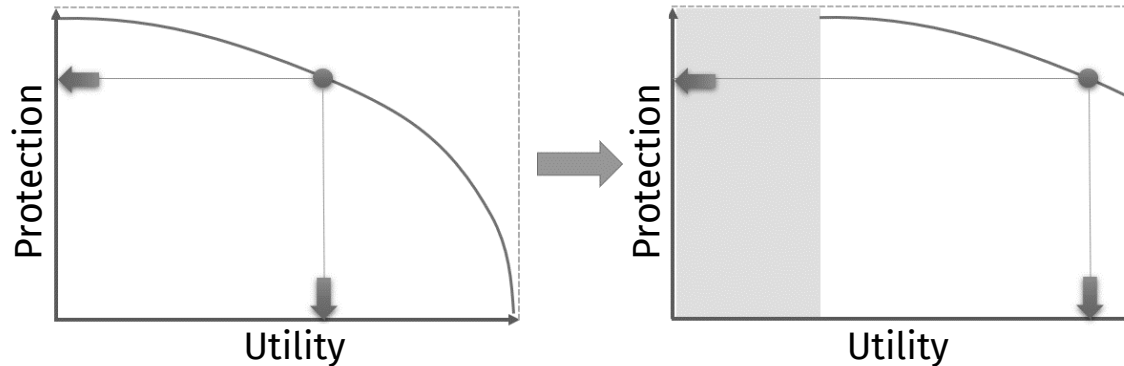


Risk-based anonymization

- Anonymization as part of a multi-layered approach to secure data sharing
- Consider the whole process of sharing data and using shared data (cf. Five Safes)



- Enables risk-based approach when captured by adversarial model



→ Cf. Differential Privacy

2. Adversarial modelling

Finding quasi-identifiers (1)

- Catalog-based approach: Comparison with variables classified as "risky" in laws and guidance documents (cf. "HIPAA List").
- Qualitative risk analysis according to Malin et al.

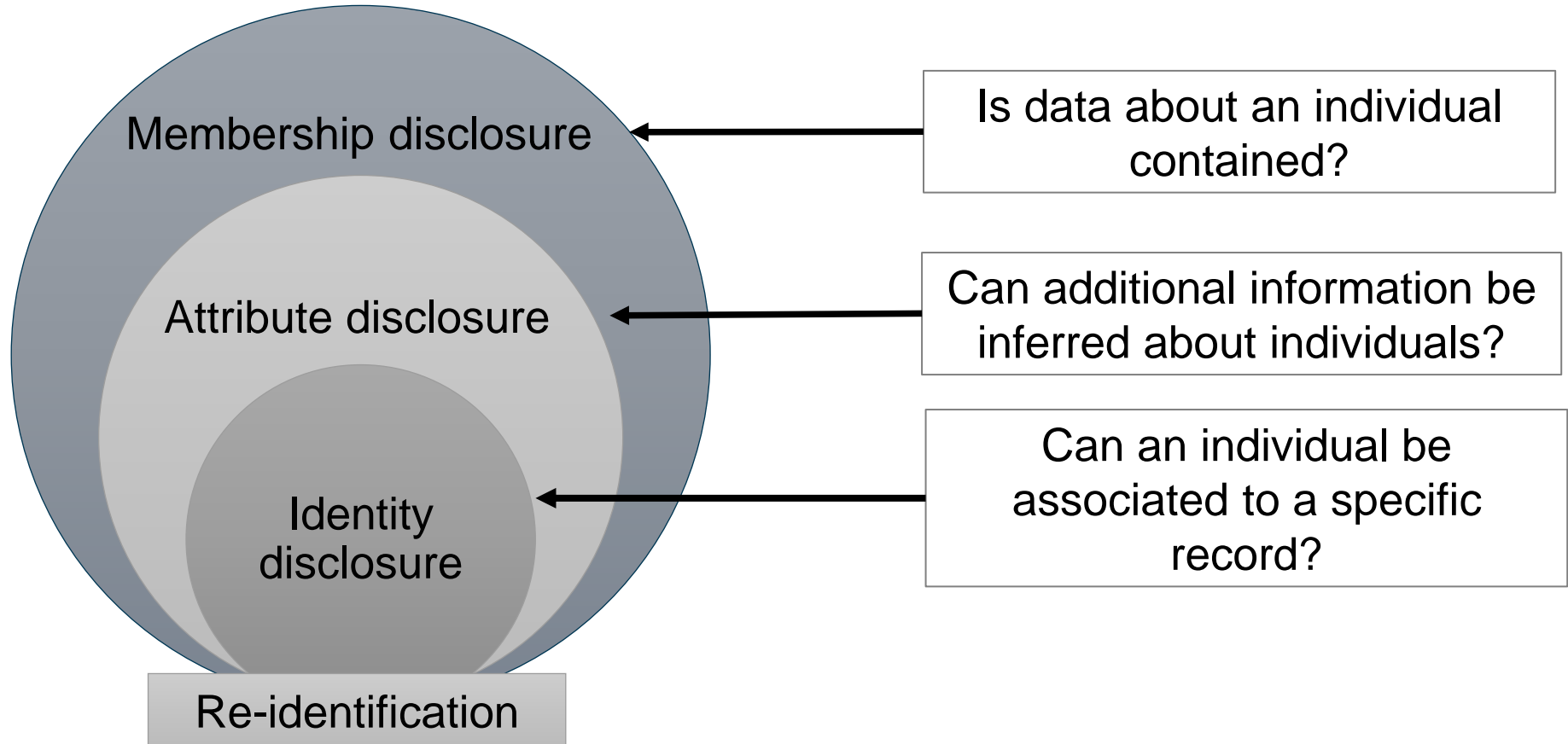
Replicability refers to the probability that specific traits or characteristics will manifest again in individuals who are impacted.	Low: blood pressure measurements may fluctuate. High: medical history tends to remain constant.
Availability: which accessible external assets might hold features that can be reproduced.	Low: lab findings are mostly known only in the medical sector. High: demographic data can be found in official records.
Distinctness: degree to which attributes allow for the differentiation of individuals.	Low: eye colour. High: rare genetic mutation.

- Followed, for example, by a threshold approach
- Quantitative methods: uniqueness, separation

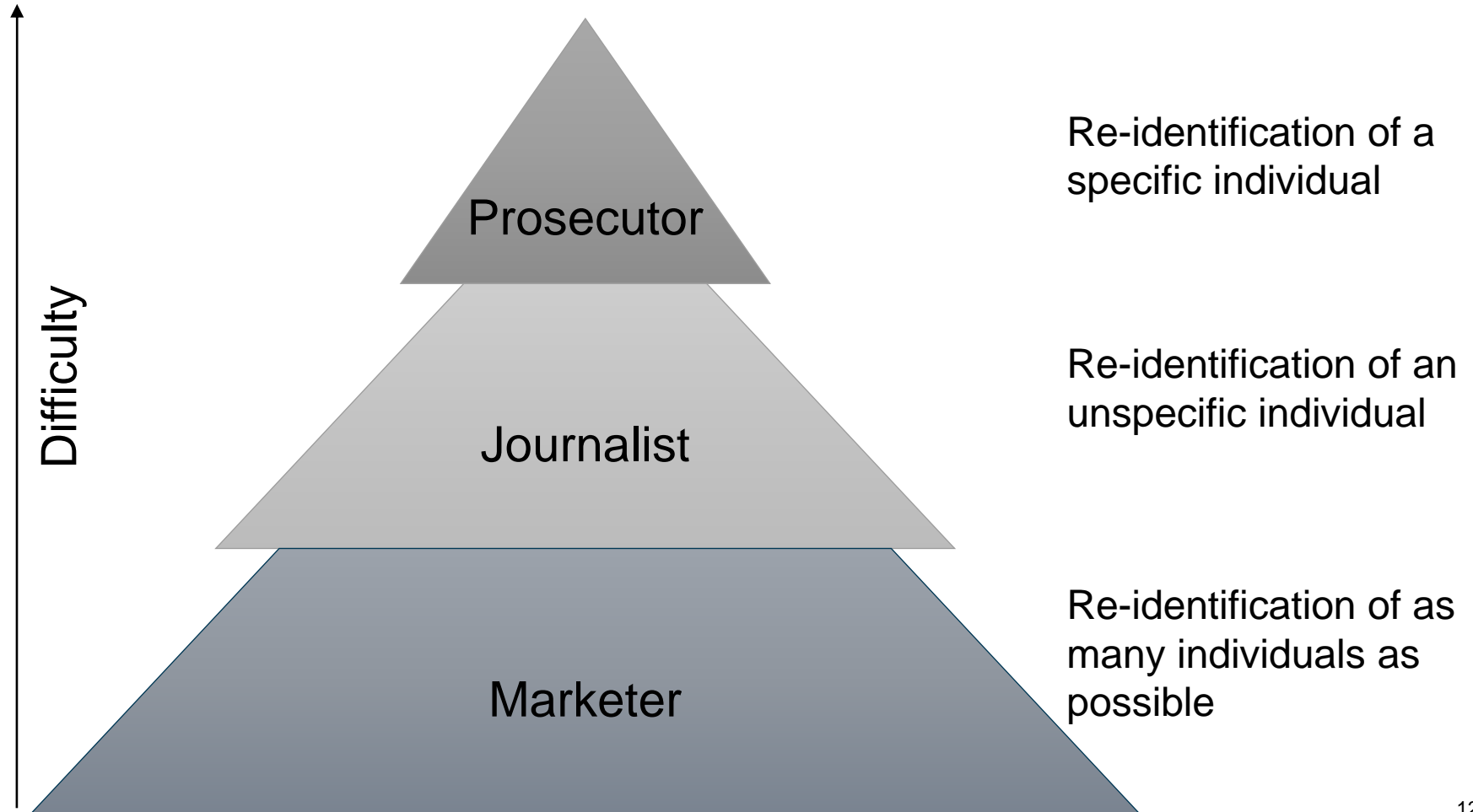
Finding quasi-identifiers (2)

Variable	Replicable	Available	Distinguish.	Key
Age at diagnosis	3	3	3	Yes (9)
Gender	3	3	2	Yes (8)
Month first diagnosis	1	3	2	Yes (6)
Year first diagnosis	1	3	2	Yes (6)
Uncomplicated phase	1	2	1	No (4)
Complicated phase	1	2	2	No (5)
Critical phase	1	2	2	No (5)
Recovery phase	1	2	1	No (4)
Vasopressors in complicated phase	1	1	2	No (4)
Vasopressors in critical phase	1	1	2	No (4)
Invasive ventilation in critical phase	1	1	2	No (4)
Superinfection in uncomplicated phase	1	1	2	No (4)
Superinfection in complicated phase	1	1	2	No (4)
Superinfection in critical phase	1	1	2	No (4)
Symptoms in recovery phase	1	1	2	No (4)
Last known patient status	1	2	2	No (5)

Reasoning about goals of adversaries (1)



Reasoning about goals of adversaries



3. Introduction to the ARX Data Anonymization Tool

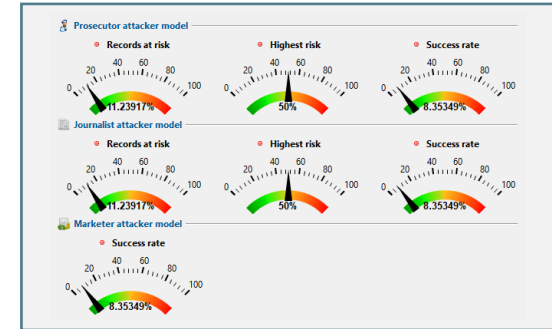
Key Facts (1)

- **Comprehensive tool for anonymizing tabular data:** Developed with a specific focus on health data
- **Continuous development:** Released in 2014 and maintained ever since
- **Open source:** Permissive Apache 2.0 license
- **Privacy models:** Supports e.g. k-Anonymity, δ -Presence, k-Map, Pop. Uniqueness, l-Diversity, t-Closeness, (ϵ, δ) -Differential privacy
- **Transformation methods:** Supports e.g. global and local schemes, generalization, random sampling, record/attribute/cell suppression, aggregation, categorization
- **Methods for utility assessment:** generic models, specialized methods, e.g. for privacy-preserving machine learning

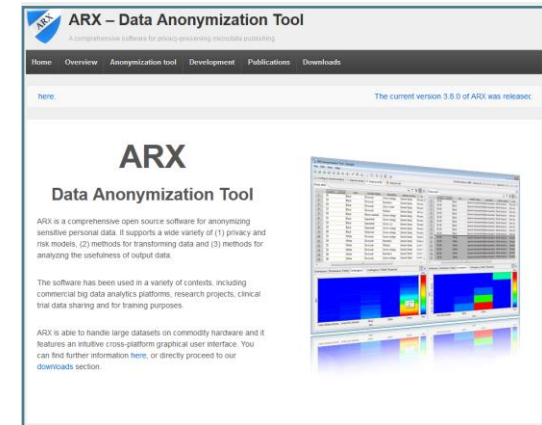


Key Facts (2)

- **Scalable:** Significantly outperforms related tools, used to anonymise datasets with billions of records. Includes algorithms for handling (relatively) high-dimensional data. Typically provides output data with more utility, requiring less resources.
- **Graphical tool:** Used in education and training by commercial and public institutions in several countries.
- **Comprehensive API:** Available as a well-designed software library for the Java ecosystem. Real-world applications often use the API to implement complex anonymization pipelines.



Software-supported determination of the re-identification risk of a medical research dataset



<https://arx.deidentifier.org/>

Key Facts (3)

- **Wide range of applications:** Creation of open datasets and used to build anonymization pipelines in several domains, e.g. by telecom providers, health insurances.
- **Industry friendly:** Integrated into several commercial products, core algorithms adopted by SAP HANA.
- **Widely used:** More than 50.000 downloads.
- **Active open source community:** 550 stars, 28 contributors on GitHub (10/2023)
- **Based in science:** Structures and algorithms implemented into ARX have been described in more than 20 journal and conference papers

Graphical frontend

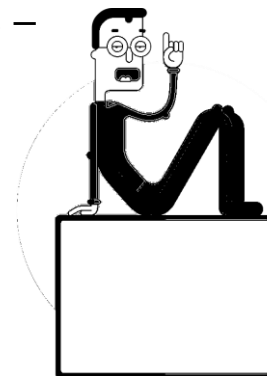
The screenshot displays the ARX Anonymization Tool interface, which is divided into several functional areas:

- Configuration:** The top-left panel shows the 'Input data' table with columns for sex, race, marital-status, and education. It also includes a 'Transformations' list and a 'Privacy criteria' section.
- Exploration:** The top-right panel shows a grid of data points with various colors (green, yellow, red) indicating different risk levels. A 'Clipboard' and 'Properties' window are also visible.
- Risk analysis:** The bottom-right panel features a 'Distribution of risk' chart showing the percentage of records with different risk levels. Below the chart is a table of 'Re-identification risks' with columns for Measure, Value [%], and Population.
- Quality analysis:** The bottom-left panel shows 'Summary statistics' and 'Contingency' tables, along with a heatmap visualization of the data.

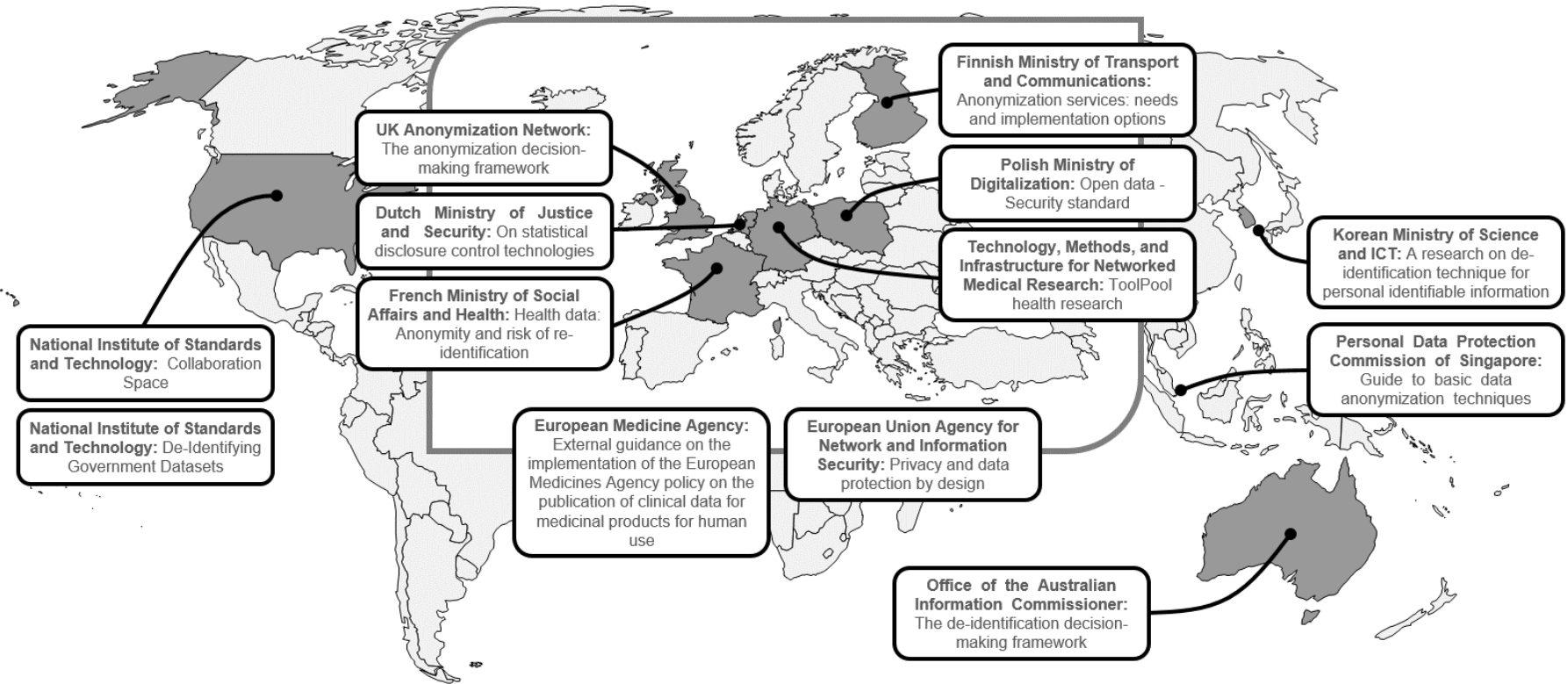
A central diagram consists of four colored quadrants (red, green, blue, yellow) arranged in a circle, with arrows indicating a clockwise cycle between them. The quadrants are labeled: Configuration (red), Exploration (green), Risk analysis (yellow), and Quality analysis (blue).

Examples of guidelines mentioning ARX (1)

- European Medicines Agency. EMA/90915/2016 – external guidance on the implementation of the European medicines agency policy on the publication of clinical data for medicinal products for human use; 2018.
- European Union Agency for Network and Information Security. Privacy and data protection by design; 2015.
- UKAN. The anonymisation decision-making framework; 2016.
- Office of the Australian Information Commissioner. The de-identification decision-making framework; 2017.
- French Ministry of Solidarity and Health. Health data: anonymity and risk of re-identification; 2015.
- Finnish Ministry of Transport and Communications. Anonymization services – requirements and implementation options; 2017.
- Personal Data Protection Commission of Singapore. Guide to basic data anonymisation techniques; 2018.
- Polish Ministry of Digitalization. Open data - Security standard; 2018.
- Dutch Ministry of Justice and Security. On statistical disclosure control technologies; 2018.
- Korean Ministry of Science and ICT. A research on de-identification technique for personal identifiable information; 2016.



Examples of guidelines mentioning ARX (2)



World Map provided by simplemaps.com

4. Practical exercises

4.1 Basic project setup

- Importing data
- Specifying metadata
- Creating generalization hierarchies
- Specifying privacy guarantees and transformation options

4.2 Anonymisation, risk and utility assessment

- Executing the anonymization process
- Performing a risk analysis
- Performing a utility analysis
- Searching for alternative solutions
- Adjusting parameters

4.3 Using different transformation methods


- Changing the transformation configuration
- Studying the impact of different transformation options


4.4 Advanced anonymization techniques

- Considering population properties
- Using privacy models protecting against attribute or membership inference
- Differential privacy

5. Recommended readings


Scientific articles

SCIENTIFIC DATA 


 Check for updates

OPEN
ARTICLE

Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19

Carolin E. M. Jakob¹, Florian Kohlmayer², Thierry Meurers^{3,4}, Jörg Janne Vehreschild^{1,5,6} & Fabian Prasser^{3,4} 

The Lean European Open Survey on SARS-CoV-2 Infected Patients (LEOSS) is a European registry for studying the epidemiology and clinical course of COVID-19. To support evidence-generation at the rapid pace required in a pandemic, LEOSS follows an Open Science approach, making data available to the public in real-time. To protect patient privacy, quantitative anonymization procedures are used to protect the continuously published data stream consisting of 16 variables on the course and therapy of COVID-19 from singling out, inference and linkage attacks. We investigated the bias introduced by this process and found that it has very little impact on the quality of output data. Current laws do not specify requirements for the application of formal anonymization methods, there is a lack of guidelines with clear recommendations and few real-world applications of quantitative anonymization procedures have been described in the literature. We therefore believe that our work can help others with developing urgently needed anonymization pipelines for their projects.

SCIENTIFIC DATA 

OPEN

Data Descriptor: Open University Learning Analytics dataset

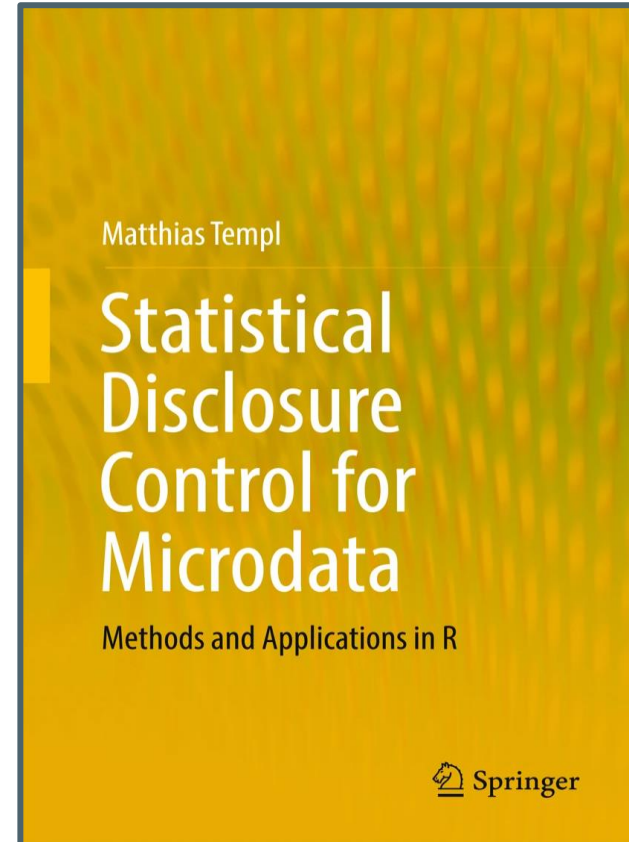
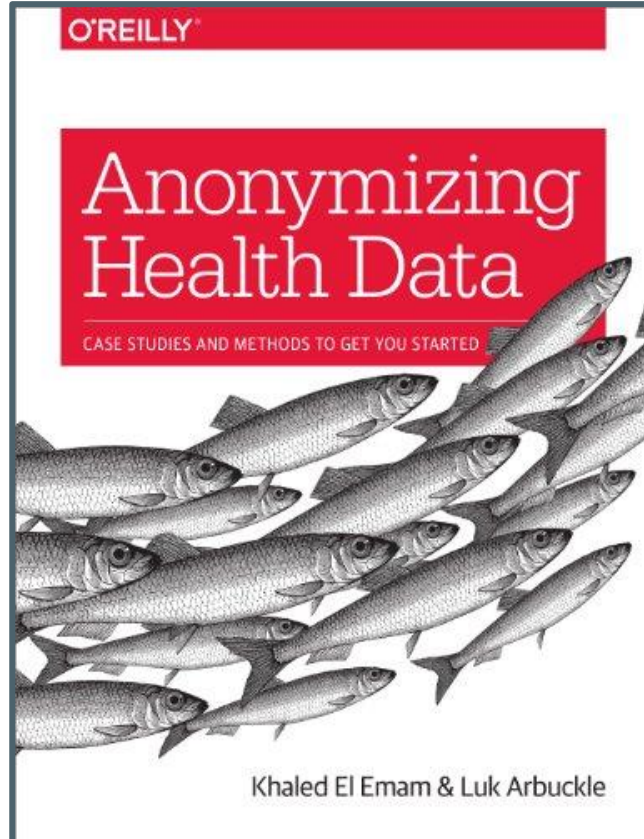
Jakub Kuzilek^{1,2}, Martin Hlosta¹ & Zdenek Zdrahal^{1,2}

Received: 20 July 2017
Accepted: 13 October 2017
Published: 28 November 2017

Learning Analytics focuses on the collection and analysis of learners' data to improve their learning experience by providing informed guidance and to optimise learning materials. To support the research in this area we have developed a dataset, containing data from courses presented at the Open University (OU). What makes the dataset unique is the fact that it contains demographic data together with aggregated clickstream data of students' interactions in the Virtual Learning Environment (VLE). This enables the analysis of student behaviour, represented by their actions. The dataset contains the information about 22 courses, 32,593 students, their assessment results, and logs of their interactions with the VLE represented by daily summaries of student clicks (10,655,280 entries). The dataset is freely available at https://analyse.kmi.open.ac.uk/open_dataset under a CC-BY 4.0 license.

Design Type(s)	time series design • data integration objective • observation design
Measurement Type(s)	learning behavior
Technology Type(s)	digital curation
Factor Type(s)	temporal_interval
Sample Characteristic(s)	Homo sapiens

Books



Tools

sdcmicro GUI

Reset SDC problems
[View SDC problem](#)

View/Analyze existing sdcProblem

Show summary

Expose variables

Add linked variables

Create new IDs

Anonymize categorical variables

Recoding

k-Anonymity

PSM (simple)

PSM (expert)

Suppress values with high risk

Anonymize numerical variables

Top-bottom coding

Microaggregation

Adding noise

Risk swapping

Suppress values with high risk

This method allows to suppress (set to NA) values in the selected key variables for records that have an individual risk higher than the specified threshold.

Select key variable for suppression: **roof**

Threshold for individual risk: **0.00**

roof

Frequency

Individual Risk

Suppress 4580 values with high risk in 'roof'

Variable selection

Variable name	Type	Suppressions
roof	cat. key variable	0
income	num. key variable	0
savings	num. key variable	0

Additional parameters

Parameter	Value
number of records	4584
alpha	1
random seed	0

k-anonymity

k-anonymity	Modified data	Original data
2-anonymity	0 (0.00%)	0 (0.00%)
3-anonymity	0 (0.00%)	0 (0.00%)
4-anonymity	0 (0.00%)	0 (0.00%)

Risk in numerical key variables

sdcmicro

6. Questions and answers

Thank you for your attention!