Competence Center on Data Anonymization

Matthias Templ

Institut für Datenanalyse und Prozessdesign
School of Engineering
Zürcher Hochschule für Angewandte Wissenschaften

Statistikertreffen, ZHAW, 15.06.2021

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

Personal history de-anonymization game

... born in [...]

Personal history de-anonymization game

... born in [...]

... Master, PhD and Habilitation in [...]

Personal history de-anonymization game

... born in [...]

... Master, PhD and Habilitation in [...]

... stopover in [...]

Personal history de-anonymization game

... born in [...]

... Master, PhD and Habilitation in [...]

... stopover in [...]

... finally arrived in [...]

Personal history de-anonymization game

... born in Linz, Austria

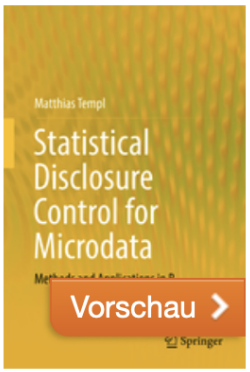
... PhD and Habilitation in Vienna, Austria

... stopover in Olomouc, Czech Republic

... finally arrived in Winterthur, Switzerland

Background in Anonymization

- Projects in/with Statistics Austria, OECD, IHSN, Paris21, Weltbank, EU, Helsana, Swisscom, ESTHER/MEIRU, SBB, Stadt and Canton of Zurich, EnerSuisse, Workshops, ...
- Publications in SDC, e.g. in the Journal of Statistical Software (sdcMicro, simPop)
- Springer-Book Statistical Disclosure Control



Statistical Disclosure Control for Microdata

Methods and Applications in R

Autoren: **Templ**, Matthias

- Lecture *Advanced Survey Statistics: Disclosure Control* at the free Univ. of Berlin, Bamberg and Trier (2019, 2020, 2021).

Overview of the presentation

1. Different concepts for different problems
2. Quantifying disclosure risks
3. *(no time for) Anonymization)*
4. Data utility
5. Competence Center at IDP and ZHAW on Data Anonymization

Overview of the presentation

1. Different concepts for different problems
2. Quantifying disclosure risks
3. *(no time for) Anonymization)*
4. Data utility
5. Competence Center at IDP and ZHAW on Data Anonymization

> *Any anonymization should be very data and use case specific with the paradigm of high data utility and if possible move towards open access data.*

Typical problems (traditional approaches)

At the university

- Students/Researchers need data for their research that including personal information

In Business

- Companies store and distribute (internally or externally) data that includes customer information

In Health and Demograhics Survilleance Systems (HDSS)

- Data with detailed spatial and demographical information need to be shared to other organisations and within an organization.

and much more ...

Is pseudo-anonymisation sufficient?

Pseudo-anonymisation is often confused with anonymisation.

- The Massachusetts Group Insurance Commission (GIC) had an *excellent* idea in the mid-1990s: to publish *anonymised* data on government employees that included every single hospital visit.
- William Weld - the then Governor of Massachusetts - assured the public that the GIC had absolutely protected patients' privacy by deleting identifiers

Is pseudo-anonymisation sufficient?

- A student searched in GIC data for the Governor's hospital records. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 people with seven postcodes.
- For $20 she bought the complete electoral rolls from the city, a database containing, among other things, the name, address, postcode, date of birth and gender of each voter.
- By merging this data with the GIC data, Sweeney found Governor Weld with ease, although the record was pseudo-anonymized.
- The student sent the governor's health files (which contained diagnoses and prescriptions) to his office.

Is pseudo-anonymisation sufficient?

- A student searched in GIC data for the Governor's hospital records. She knew that Governor Weld resided in Cambridge, Massachusetts, a city of 54,000 people with seven postcodes.
- For $20 she bought the complete electoral rolls from the city, a database containing, among other things, the name, address, postcode, date of birth and gender of each voter.
- By merging this data with the GIC data, Sweeney found Governor Weld with ease, although the record was pseudo-anonymized.
- The student sent the governor's health files (which contained diagnoses and prescriptions) to his office.

As a result, the legislators have known about for over 25 years:

pseudonymization is different than anonymization

pseudonymized data $\neq$ anonymized data.

Different concepts for different problems

- sharing individual data?
- privacy-by-design black box methods
- providing aggregated information, predictions, or individual information?
- different needs for different type of user

Different concepts for anonymization (1/2)

- **Traditional anonymization** to anonymize data
  - for scientific or public-use files.
  - data are modified under the paradigm of high data utility.

Different concepts for anonymization (1/2)

- **Traditional anonymization** to anonymize data
  - for scientific or public-use files.
  - data are modified under the paradigm of high data utility.
- **Remote execution**, **secure lab**, and **remote access**
  - for access to pseudo-anonymized data on a secure server
  - involves high costs, especially remote execution

Different concepts for anonymization (1/2)

- **Traditional anonymization** to anonymize data
  - for scientific or public-use files.
  - data are modified under the paradigm of high data utility.
- **Remote execution**, **secure lab**, and **remote access**
  - for access to pseudo-anonymized data on a secure server
  - involves high costs, especially remote execution
- **Query servers** with perturbed aggregated output
  - e.g. for aggregated information in public dashboards
  - optimization methods versus differential privacy methods
  - sophisticated methods when subtotals + totals and hierarchies are present in multidimensional tabular data.

Different concepts for anonymization (2/2)

- **Privacy preserving computation**. Privacy by design / black box methods to receive predictions on sensitive variables of *test data* without access to *training data*
  - **Differential privacy** to noise output/predictions
  - **Federated learning** (PATE, ...) for calculations on distributed data sets on clients side on their (training) data

Different concepts for anonymization (2/2)

- **Privacy preserving computation**. Privacy by design / black box methods to receive predictions on sensitive variables of *test data* without access to *training data*
  - **Differential privacy** to noise output/predictions
  - **Federated learning** (PATE, . . .) for calculations on distributed data sets on clients side on their (training) data
- **Synthetic data**
  - machine learning methods to simulate synthetic data from original data
  - may serve as training data in machine learning
  - may serve as *twin* data for the public or sharing within an organization
  - for augmented data or in form of population data

Anonymization and de-facto anonymity

- (ISO/TS 25237:2008) Anonymization: *Process that removes the association between the identifying data set and the data subject.*
- (Traditional) Anonymisation involves the **use of complex methods** of statistical disclosure control.
- *Absolute anonymity* is not possible and is not required by laws on privacy (keyword: **de-facto anonymity**)

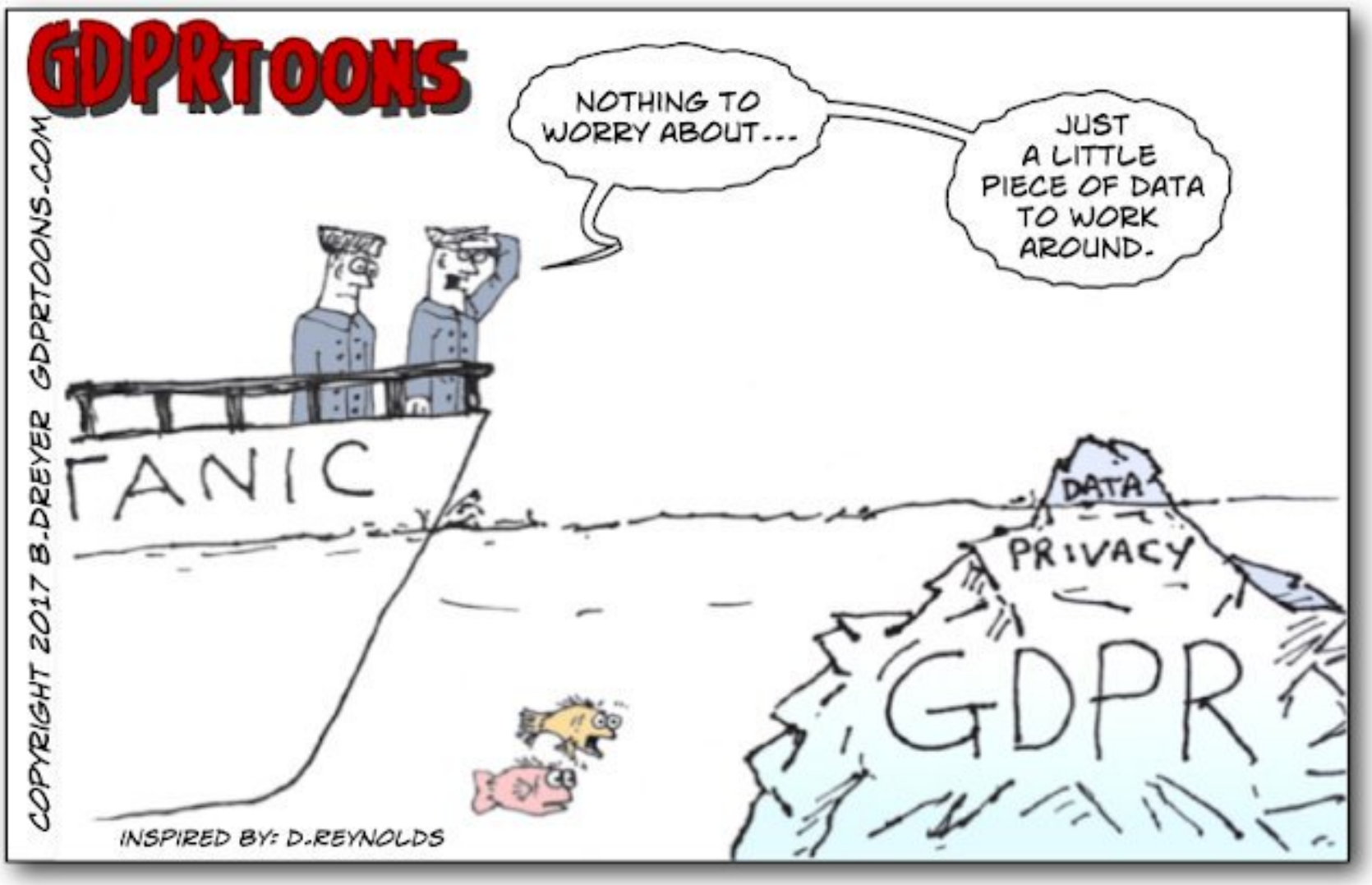
Anonymization and de-facto anonymity

- (ISO/TS 25237:2008) Anonymization: *Process that removes the association between the identifying data set and the data subject.*
- (Traditional) Anonymisation involves the **use of complex methods** of statistical disclosure control.
- *Absolute anonymity* is not possible and is not required by laws on privacy (keyword: **de-facto anonymity**)

de-facto anonymity

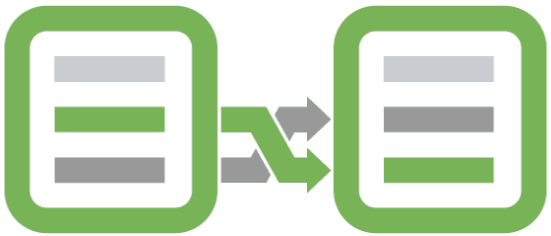
If the **effort is higher** data is to be re-identified **as the benefit** we speak of **de-facto anonymity**.

Everything simple?



Types of re-identification

1. **Identity disclosure**. Link of the record with **external data** so that person is identified.
   - Example 1 on next slide
   - Example 2: Persons including information about mental disorder. **Record Linkage** of the quasi-identifiers (e.g. age, gender, occupation, municipality) with external data containing names of persons.
   - If the link for a person is successful, the data attacker now knows the names of persons having mental disorder.



MATCH

Example: Matching of key variables (1)

Records provided (QID's and keys residence $\times$ occupation $\times$ gender)

name	residence	occ.	sex	# $\{1, .., n\} \in$ key	Income	...
x	Stadel	Prof	M	1	176456	...
x	Winterthur	architect	M	18	143111	...
x	...	...	...	...	...	...

External data set (the intruders external knowledge)

name	place of residence	occ.	gender	# $\{1, .., n\} \in$ key	...
Max Muster	Stadel	Prof	M	1	...
Jo Johann	Winterthur	architect	M	18	...
Nils Nilson	Winterthur	architect	M	18	...
...	...	...	...	...	...

Example: Matching of key variables (1)

Records matched (QID's and keys residence × occupation × gender)

name	residence	occ.	sex	# $\{1, .., n\} \in$ key	Income	...
Max Muster	Stadel	Prof	M	1	176456	...
x	Winterthur	architect	M	18	143111	...
x	...	...	...	...	...	...

External data set (the intruders external knowledge)

name	place of residence	occ.	gender	# $\{1, .., n\} \in$ key	...
Max Muster	Stadel	Prof	M	1	...
Jo Johann	Winterthur	architect	M	18	...
Nils Nilson	Winterthur	architect	M	18	...
...	...	...	...	...	...

Example: Matching of key variables (1)

Records matched (QID's and keys residence × occupation × gender)

name	residence	occ.	sex	# $\{1, .., n\} \in$ key	Income	...
Max Muster	Stadel	Prof	M	1	176456	...
x	Winterthur	architect	M	18	143111	...
x	...	...	...	...	...	...

External data set (the intruders external knowledge)

name	place of residence	occ.	gender	# $\{1, .., n\} \in$ key	...
Max Muster	Stadel	Prof	M	1	...
Jo Johann	Winterthur	architect	M	18	...
Nils Nilson	Winterthur	architect	M	18	...
...	...	...	...	...	...

→ Max Muster is clearly matchable.

Types of re-identification

2. **Attributes Disclosure**.

- Example: A health study publishes statistics in which **all** people with nationality of Austria and age between 45 and 50 has dementia:

	key variables			sensitive variable
	Nat.	age	region	dementia
1	Aut	45–50	Winterthur	yes
2	Aut	45–50	Winterthur	yes
3	Aut	45–50	Winterthur	yes
4	Aut	45–50	Winterthur	yes

Types of re-identification

2. **Attributes Disclosure**.

- Example: A health study publishes statistics in which **all** people with nationality of Austria and age between 45 and 50 has dementia:

	key variables			sensitive variable
	Nat.	age	region	dementia
1	Aut	45–50	Winterthur	yes
2	Aut	45–50	Winterthur	yes
3	Aut	45–50	Winterthur	yes
4	Aut	45–50	Winterthur	yes

$\Longrightarrow$

Types of re-identification

2. **Attributes Disclosure**.

- Example: A health study publishes statistics in which **all** people with nationality of Austria and age between 45 and 50 has dementia:

	key variables			sensitive variable
	Nat.	age	region	dementia
1	Aut	45–50	Winterthur	yes
2	Aut	45–50	Winterthur	yes
3	Aut	45–50	Winterthur	yes
4	Aut	45–50	Winterthur	yes

$\Longrightarrow$ I cannot remember what to say

Types of re-identification

2. **Attributes Disclosure**.

- Example: A health study publishes statistics in which **all** people with nationality of Austria and age between 45 and 50 has dementia:

	key variables			sensitive variable
	Nat.	age	region	dementia
1	Aut	45–50	Winterthur	yes
2	Aut	45–50	Winterthur	yes
3	Aut	45–50	Winterthur	yes
4	Aut	45–50	Winterthur	yes

$\Longrightarrow$ we learn: Every **individual** of Austrian nationality in age group [45-50] living in Winterthur has dementia.

Types of re-identification

2. **Attributes Disclosure**.

- Example: A health study publishes statistics in which **all** people with nationality of Austria and age between 45 and 50 has dementia:

	key variables			sensitive variable
	Nat.	age	region	dementia
1	Aut	45–50	Winterthur	yes
2	Aut	45–50	Winterthur	yes
3	Aut	45–50	Winterthur	yes
4	Aut	45–50	Winterthur	yes

$\Longrightarrow$ we learn: Every **individual** of Austrian nationality in age group [45-50] living in Winterthur has dementia.

3. **Inferential Disclosure**. model-based estimation of the value of a sensitive variable: when the quality of prediction is too high.

Trad. anonymisation in practice: rough procedure

1) **RISK** Measurement of risk 

- Sample or population? Micro data or tabular data?
- Which data sources with overlapping populations exist on the *market*?
- Determination of a so-called **disclosure scenario**.
- Individual risk (of each individual person) and global risk

Trad. anonymisation in practice: rough procedure

1) **RISK** Measurement of risk 

- Sample or population? Micro data or tabular data?
- Which data sources with overlapping populations exist on the *market*?
- Determination of a so-called **disclosure scenario**.
- Individual risk (of each individual person) and global risk

2) 

Anonymisation 

- Traditional methods or synthetic data generation?
- Categorical variables and/or continuous variables?
- Clusters and hierarchical structures present in data?

Trad. anonymisation in practice: rough procedure

1) **RISK** Measurement of risk 

- Sample or population? Micro data or tabular data?
- Which data sources with overlapping populations exist on the *market*?
- Determination of a so-called **disclosure scenario**.
- Individual risk (of each individual person) and global risk

2) 

Anonymisation 

- Traditional methods or synthetic data generation?
- Categorical variables and/or continuous variables?
- Clusters and hierarchical structures present in data?

3) 

Measurement of the utility 

- Global procedures or data-specific comparisons?
- What is the analysis of interest of the users?

Re-identification scenarios

1. Nosy neighbour scenario

- The data recipient has detailed personal information about a specific (or some) person(s).
- Example: Celebrities in NYC Taxi, tip

2. The archive (matching) scenario

- Match via key variables with other data sources ("Archives' ') which contain clear names or ID's (*Record Linkage* problem)
- Re-identify people through successful matches

...

There are more scenarios, but they are less common

Disclosure Risk

The most important and complicated part of SDC is not to apply anonymisation methods, but the measurement of the re-identification risk of individuals.

- for register/population data, risk determination is easier.
- non-trivial for survey samples and/or for data with missing values
- non-trivial for data with missing values

Disclosure Risk for populations

Concept of the **Uniqueness**:

- By combining several variables (the QID's), an individual can uniquely be identified in the data record.
- A key is unique if its frequency is 1 (only one person has the combination of characteristics defined by the key. Example: the key Postcode **8404**, citizienship **Austria**, **male**, **age 45**)

Disclosure Risk for populations

Concept of the **Uniqueness**:

- By combining several variables (the QID's), an individual can uniquely be identified in the data record.
- A key is unique if its frequency is 1 (only one person has the combination of characteristics defined by the key. Example: the key Postcode **8404**, citizienship **Austria**, **male**, **age 45**)

Concept of *k*-**anonymity**:

- Each combination of key variables contains at least *k* observations
- Often we want to ensure 3-anonymity

Disclosure Risk for populations

Concept of the **Uniqueness**:

- By combining several variables (the QID's), an individual can uniquely be identified in the data record.
- A key is unique if its frequency is 1 (only one person has the combination of characteristics defined by the key. Example: the key Postcode **8404**, citizienship **Austria**, **male**, **age 45**)

Concept of *k*-**anonymity**:

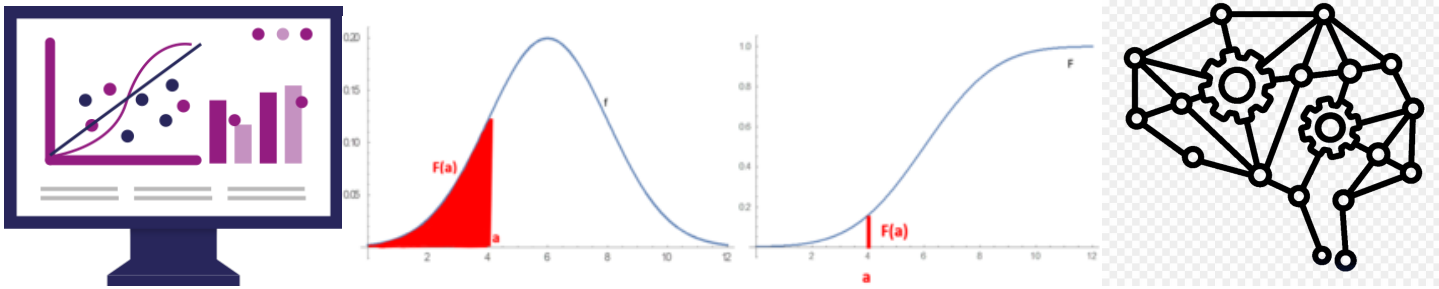
- Each combination of key variables contains at least *k* observations
- Often we want to ensure 3-anonymity

For survey data, risk assessment is much more complicated.

Methods for anonymisation of data

Different groups of methods:

- Methods that generalize or suppress values. Examples are recoding or local suppression
- Methods which perturb data. Examples are Adding Noise, Post-Randomization Method (PRAM), Microaggregation and Shuffling.
- Methods for generating synthetic data



Utility

- After data has been anonymised, it is important to assess the **information loss** and the **data quality**.
- Comparing results from original and anonymised data (tables, regression models, distributions, . . .)
- Comparison of indicators
- Propensity score matching methods
- Etc.

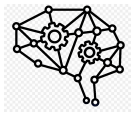
If the loss of data is high, anonymisation should be considered.

Trade-Off and iterative approach (Anonymisation $\leftrightarrow$ Utility)

Software

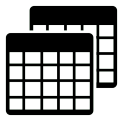
sdcMicro (Templ et al., Journal of Statistical Software, 2016)

- state-of-the-art software
- can handle more complex data
- with click-App (for the browser)
- is programmed very efficiently (C++ code, parallel computing)



simPop (Templ et al., Journal of Statistical Software, 2017)

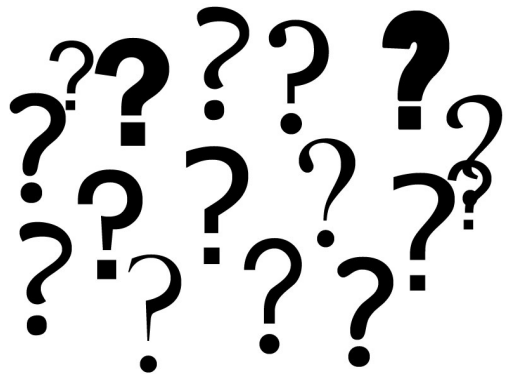
- for the creation of synthetic data sets
- unlike other software, can also handle more complex data structures



sdcTable and **cellKey** (Author: B. Meindl)

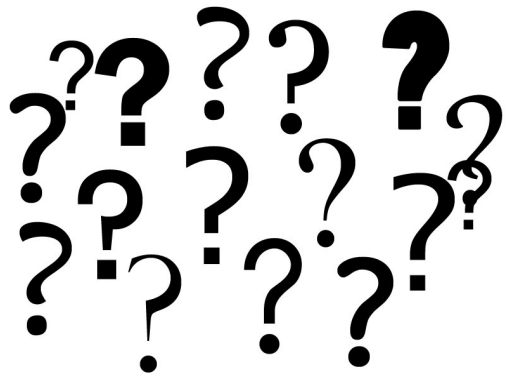
- For the confidentiality of tables (aggregated information)

Difficulties in practice



- Unfortunately there is no general solution and no standardized procedure
- Anonymisation varies from case to case. Strongly data- and case-dependent
- Years of experience necessary

Difficulties in practice



- Unfortunately there is no general solution and no standardized procedure
- Anonymisation varies from case to case. Strongly data- and case-dependent
- Years of experience necessary

This is the reason why we need a competence center on data anonymization :)

Projects at IDP on data anonymization (2017-2021)

- Anonymisierung von Daten der **Helsana** Gesundheitswissenschaften (LINK)
  - Anoymisierung der Datenbank über Versicherte inklusive Apotheken- und Krankenhaus-rechnungen und -aufenthalten

Projects at IDP on data anonymization (2017-2021)

- Anonymisierung von Daten der **Helsana** Gesundheitswissenschaften (LINK)
  - Anoymisierung der Datenbank über Versicherte inklusive Apotheken- und Krankenhaus-rechnungen und -aufenthalten
- Evaluierung der Daten-Anonymisierung bei Swisscom (LINK)
  - Anonymisierung des Data lake und der Anonymisierungsplatform + Mobilitätsdaten in Applikationen/Dashboards

Projects at IDP on data anonymization (2017-2021)

- Anonymisierung von Daten der **Helsana** Gesundheitswissenschaften (LINK)
  - Anoymisierung der Datenbank über Versicherte inklusive Apotheken- und Krankenhaus-rechnungen und -aufenthalten
- Evaluierung der Daten-Anonymisierung bei Swisscom (LINK)
  - Anonymisierung des Data lake und der Anonymisierungsplatform + Mobilitätsdaten in Applikationen/Dashboards
- Datenanonymisierung für die SBB (LINK)
  - Kurse und Mobilitätsdaten/Trajektoriendatenanonymisierung

Projects at IDP on data anonymization (2017-2021)

- Anonymisierung von Daten der **Helsana** Gesundheitswissenschaften (LINK)
  - Anoymisierung der Datenbank über Versicherte inklusive Apotheken- und Krankenhaus-rechnungen und -aufenthalten
- Evaluierung der Daten-Anonymisierung bei Swisscom (LINK)
  - Anonymisierung des Data lake und der Anonymisierungsplatform + Mobilitätsdaten in Applikationen/Dashboards
- Datenanonymisierung für die SBB (LINK)
  - Kurse und Mobilitätsdaten/Trajektoriendatenanonymisierung
- Smart-Meter Datenanonymisierung für EnerSuisse (LINK)
  - Suppressions bei Abfragen von Datenbanken

Projects at IDP on data anonymization (2017-2021)

- Anonymisierung von Daten der **Helsana** Gesundheitswissenschaften (LINK)
  - Anoymisierung der Datenbank über Versicherte inklusive Apotheken- und Krankenhaus-rechnungen und -aufenthalten
- Evaluierung der Daten-Anonymisierung bei Swisscom (LINK)
  - Anonymisierung des Data lake und der Anonymisierungsplatform + Mobilitätsdaten in Applikationen/Dashboards
- Datenanonymisierung für die SBB (LINK)
  - Kurse und Mobilitätsdaten/Trajektoriendatenanonymisierung
- Smart-Meter Datenanonymisierung für EnerSuisse (LINK)
  - Suppressions bei Abfragen von Datenbanken
- Bessere De-Identifizierung durch statistische Anonymisierung für Gesundheitsdaten der malawischen Bevölkerung (LINK)
  - Anonymisierung von longitudinalen Gesundheitsdaten

Projects at IDP on data anonymization (2017-2021)

- Anonymisierung von Daten der **Helsana** Gesundheitswissenschaften (LINK)
  - Anoymisierung der Datenbank über Versicherte inklusive Apotheken- und Krankenhaus-rechnungen und -aufenthalten
- Evaluierung der Daten-Anonymisierung bei Swisscom (LINK)
  - Anonymisierung des Data lake und der Anonymisierungsplatform + Mobilitätsdaten in Applikationen/Dashboards
- Datenanonymisierung für die SBB (LINK)
  - Kurse und Mobilitätsdaten/Trajektoriendatenanonymisierung
- Smart-Meter Datenanonymisierung für EnerSuisse (LINK)
  - Suppressions bei Abfragen von Datenbanken
- Bessere De-Identifizierung durch statistische Anonymisierung für Gesundheitsdaten der malawischen Bevölkerung (LINK)
  - Anonymisierung von longitudinalen Gesundheitsdaten
- Fellowship DIZH *Anonymisation and estimation of the re-identification risk of personal data*

Competence center on data anonymization

Activities:

1. Will go online as a *Lab* in a few weeks
2. Will list projects and references
3. Will promote data anonymization
4. Workshops, courses, and presentations

Competence center on data anonymization

Activities:

1. Will go online as a *Lab* in a few weeks
2. Will list projects and references
3. Will promote data anonymization
4. Workshops, courses, and presentations
5. Collaboration is highly welcome

Linking the lab with other people and institutes, and to topics not covered by IDP

Competence center on data anonymization

Activities:

1. Will go online as a *Lab* in a few weeks
2. Will list projects and references
3. Will promote data anonymization
4. Workshops, courses, and presentations
5. Collaboration is highly welcome

   Linking the lab with other people and institutes, and to topics not covered by IDP

6. Consultancy
7. Research
8. Software development

Competence center on data anonymization

Activities:

1. Will go online as a *Lab* in a few weeks
2. Will list projects and references
3. Will promote data anonymization
4. Workshops, courses, and presentations
5. Collaboration is highly welcome

   Linking the lab with other people and institutes, and to topics not covered by IDP

6. Consultancy
7. Research
8. Software development

Thank you for your attention and follow-up discussions