Statistiker:innen Treffen 2026:

Workshop on data anonymization: Theory

Reto Bürgin, bure@zhaw.ch
Institut for Data Science, ZHAW

2026-09-08

About me

- Education: Statistician (Master of Science ETH, PhD University of Geneva)
- Current positions:
  - *Since 2021*: Teaching and R&D at ZHAW, Institut for Data Science
    Teaching of data anonymization methods in connection with survey data
  - *Since 2015*: Statistician/ Data Analyst at LEP AG, R&D group
    Collection and analysis of nursing documentation data
- Selected former positions:
  - *2015-2020*: Federal Office of Public Health (FOPH), premiums and financial risks KVG
    Work on problem that data must be anonymous at the time of collection; development of an iterative process
  - *2017-2020*: Bern University of Applied Science
    Support of projects involving sensitive health data

Outline

1. The DSEmbedded 2.0 project

2. Workshop on data anonymization

Outline

1. The DSEmbedded 2.0 project

2. Workshop on data anonymization

The DSEmbedded 2.0 project

- Funded by swissuniversities and ZHAW
- Project lead: Simon van Rekum (ZHAW Finance & Services)
- Team: Tibor Pimentel (SML), Hedi Strahm (F&S), Tatiana Feketeová (F&S), Reto Bürgin
- Follow-up project of Integrating discipline-specific Know How into Data Stewardship
- DSEmbedded 2.0 has a thematic focus on **Data anonymization** and **Data archiving**
- Duration: July 2025 to the end of 2026

The DSEmbedded 2.0 project

Activities:

- **Consulting** services available to ZHAW researchers on data anonymization and data archiving (so far one request for qualitative, one request for quantitative data)
- **Guidelines** (e.g. see Zenodo: Illustration einer Anonymisierung von quantitativen Umfragedaten in R
- **Workshops** Workshop in January, StatistikerInnen Treffen 2026
- **Networking** (now being part of Swiss Data Anonymization Competence Center and SRDSN node "Personal and Sensitive Data"
- **Applied research and development of tools**

Outline

1. The DSEmbedded 2.0 project

2. Workshop on data anonymization

What is Data Anonymization?

Data anonymization is the process of transforming

> *personal data, i.e., a set of data records that can be linked with a identifiable person (or, more generally, the protected entity)*

into "anonymous" data where

- Any personal reference is removed or only possible with disproportionate effort
- Re-identification is practically impossible under reasonable assumptions, even when combined with other data
- The process is irreversible according to legal standards (e.g., General Data Protection Regulation of the European Union, GDPR)
- The balance between privacy protection and data utility is maintained

The relevance of data anonymization today

- Recommendations for science practices such as the UNESCO Recommendation on Open Science (UNESCO, 2021), but also research funds such as the SNSF, **propagate or even require the publication of research data**, which often entails data anonymization
- In industry, data anonymization is often used to create synthetic data.
- At the same time, the increase of digitally available data collections, powerful IT infrastructures and software and vast digital networks **make re-identification increasingly easier**
- As a result, **data protection regulations** such as General Data Protection Regulation of the European Union (European Parliament and Council of the European Union, 2016, GDPR) **have become more precise and rigorous** and imply higher requirements on data anonymization

Here

- **Issues** related to data anonymization
- Main **types and scenarios of re-identification**
- **Variable types** with respect to anonymity and how to identify them in real datasets
- Basic **metrics for risk quantification**
- Overview of **data anonymization methods**
- Fundamentals of **measuring information loss**

Data Security vs. Data Protection

- **Data Security**: Protection of digital data, e.g., in a database, against unwanted actions by unauthorized users or attackers
- **Data Protection**: Protection of personal rights during data processing. *Anonymization is part of data protection*

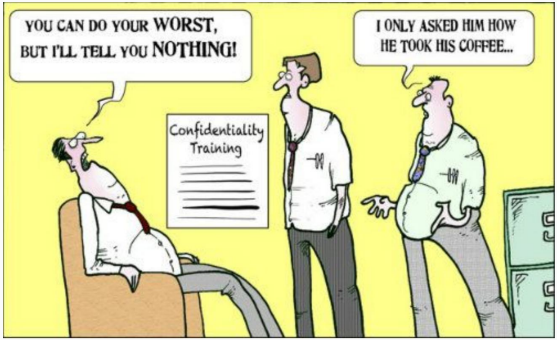
We are interested in:

1. Quantifying the risk of attackers re-identifying individuals, given that the attacker has unauthorized access to the data
2. Methods to reduce this risk $\Rightarrow$ "coarsening" the data so that re-identification for each person is minimal while preserving the information content of the data

Anonymization vs. Pseudonymization vs. Encryption

- **Encryption**: *The process of disguising a message in such a way as to hide its substance is called encryption. An encrypted message is ciphertext.* (Schneier, 1996, P. 1)
- **Pseudonymization**: *The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, ...* Art. 4(5) GDPR
  $\Rightarrow$ **Remove directly identifying personal reference using encryption**
- **Anonymization**: *process of transformation ...*, see Slide 8
  $\Rightarrow$ **Remove any personal reference**

Training on Confidentiality



Unfortunately, this does not reflect daily practice!

Misconceptions (1)

"Pseudonymization of direct-identifiers = anonymous data" ⇒ wrong

Data before pseudonymization:

	Name	Municipality	Gender	Age	Occupation	Fare Dodger
1	Max Muster	Bex	m	40	Teacher	yes
2	Mia Berg	Egg	f	37	Doctor	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Data after pseudonymization of the **direct** identifier "Name" by code:

	Name	Municipality	Gender	Age	Occupation	Fare Dodger
1	!dsf(...	Bex	m	40	Teacher	yes
2	i9r.p...	Egg	f	37	Doctor	no
⋮	⋮	⋮	⋮	⋮	⋮	⋮

⇒ Individuals may still be re-identified via **indirect** identifiers (municipality, etc.) combined with an external data source.

Is Pseudonymization of Direct-Identifiers Sufficient?

Pseudonymization is often confused with anonymization.

- The Massachusetts Group Insurance Commission (GIC) published pseudonymized individual data on hospital visits of state employees in the mid-1990s
- William Weld, the then Governor of Massachusetts, assured the public that the GIC had absolutely protected patient privacy by deleting identifiers

Is Pseudonymization of Direct-Identifiers Sufficient?

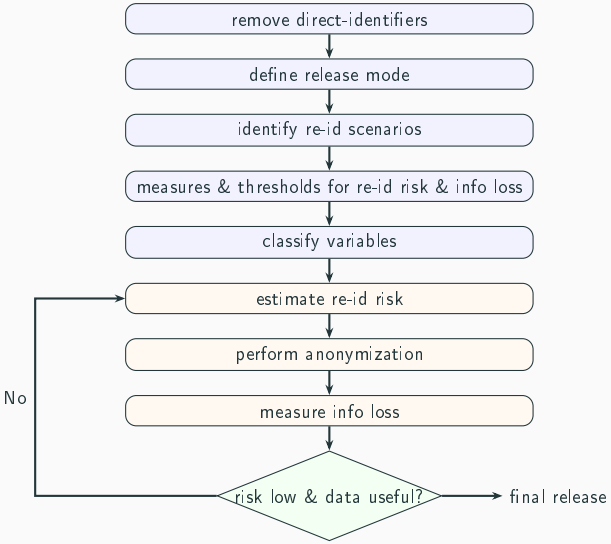
- A student searched the GIC data for the governor's hospital records. She knew that Governor Weld resided in Cambridge, Massachusetts, a city with 54'000 inhabitants and seven ZIP codes and had a hospital stay
- For $20, she purchased the complete voter lists from the city of Cambridge – a database that included, among other things, the name, address, ZIP code, date of birth, and gender of every voter
- By simply merging these data with the GIC data, Sweeney easily found Governor Weld, even though the dataset was pseudonymized
- The student sent the governor's health records (which contained diagnoses and prescriptions) to his office in a dramatic fashion

See also Sweeney (2015)

Misconceptions (2)

- *"With a usage agreement and confidentiality agreement, anonymization is not necessary"*
  $\Rightarrow$ helpful, but wrong
- *"Within a company, data does not need to be anonymized"*
  $\Rightarrow$ wrong
- *"Data quality suffers from anonymization"*
  $\Rightarrow$ wrong, good data anonymization can drastically reduce the re-identification risk for individuals without drastically altering the data.

Anonymization in Practice: Rough Procedure



Data Release Modes

In principle, **anonymization** occurs during **data sharing**. Common modes are:

- **Internal Sharing**: Generally less strict anonymization, as IT security and access rights can be controlled. The intrinsic motivation to re-identify individuals is usually low
- **External Sharing**: For example, for scientific studies. Stricter anonymization, usually accompanied by a usage agreement and confidentiality agreement
- **Publication ("Public Use")**: Strong anonymization, re-identification risk should ideally be zero!

Variable Types

Variable types:

- **Direct-identifiers**: Variables that precisely identify statistical unit, e.g. full name, insurance number, address, etc. They sould be globally unique.
- **Quasi-identifiers**: A.k.a indirect-identifiers or key variables. Variables that, when considered together, can be used to identify individual units, e.g. ZIP code, age, gender
- **Sensitive variables**: E.g. health status
- **Design Variables** e.g., stratification variables, survey weights
- **Other** (remainder) i.e. non-confidential variables, design variables etc.

Notes:

- Variables may be assigned to multiple types (no disjunction)
- There is no "true" classification as it is context-specific and depends on assumptions (e.g. what information intruders have)
- Identifying these quasi-identifiers is fundamentally hypothetical (how can I know what information a potential attacker possesses?) and therefore difficult to determine

Variable Types

Direct-identifiers:

- Direct-identifiers should be deleted or pseudonymized
- If an ID is to be retained, it is usually sufficient to replace it with randomly assigned serial numbers.

```
ID <- c("Anna", "Christoph", "Nima", "Reto", "Thoralf")
(ID.Pseudo <- sample(seq_along(ID)))
## [1] 4 5 2 3 1
```

- A more professional method is “hashing”: It transforms a string into a fixed-length, non-readable one. This process cannot be reversed, reproducibility is possible when using the same (securely stored) seed or salt. Disadvantage: Very memory-intensive!

```
# devtools::install_github("paulhendricks/anonymizer")
library("anonymizer")
name_salt <- salt(
  "Reto Bürgin",
  .n_char = 5,
  .seed = 123)
name_salt
## [1] "osncjReto Bürginosncj"
hash(name_salt, .seed = 123)
## [1] "13f13d902615556ff24add55fe312722214bdf5a88e319322fb0cec305ebf6b4"
```

Variable Types

Your turn: Practice your variable classification skills by classifying the following variables:

1. ZIP
2. Awarness of climate change
3. insurance number
4. language
5. person wears corrective eyeglasses

Variable types

My suggestion for classification:

1. ZIP $\Rightarrow$ quasi-identifier
2. Awarness of climate change $\Rightarrow$ other
3. insurance number $\Rightarrow$ direct-identifier
4. language $\Rightarrow$ quasi-identifier
5. person wears corrective eyeglasses $\Rightarrow$ sensitive variable

Variable types

Discussion:

- Disagreements between classifiers seem to be common
- Recommendation: Classification by at least 2 persons plus a consensus round
- My current research suggests that generative AI may be used as a “third opinion”.

Re-identification scenarios

- **Identity Disclosure**: The attacker has external data with direct-identifiers and a set of quasi-identifiers (gender, age, place of residence, etc.) that are also present in the data to be protected. Re-identification occurs by merging the two data sources.
- **Attribute Disclosure**: By example, the SBB publishes the following statistics:

	Indirect RID Variables				Sensitive Variables	
	Gender	Age Group	Residence	Nationality	Fare Dodger	Count
1	m	20-29	Winterthur	AT	never	0
					once	0
					multiple	10

  Do you know someone living in Winterthur who is Austrian? $\Rightarrow$ Must be a fare dodger.
- **Inferential Disclosure**: Values of sensitive variables can be predicted very accurately using models/algorithms (model built on data, prediction for external data containing the predictors and direct-identifiers)

The re-identification risk is measured differently depending on the scenario.

Re-identification scenarios

Example Identity Disclosure:

- External data of attacker:

	Name	Residence	Occupation	Gender
1	Max Muster	Stadel	Cashier	M
2	Jana Meiser	Zurich	Cook	F

- Your data to be protected:

	Residence	Occupation	Gender	Income	...
1	Stadel	Cashier	M	75'000	
2	Zurich	Cook	F	90'000	

- Your data linked with identifiers:

	Residence	Occupation	Gender	Income	...	Name
1	Stadel	Cashier	M	75'000	...	Max Muster
2	Zurich	Cook	F	90'000	...	Jana Meister

This is possible since (here) there is only 1 male cashier in Stadel ("unique case")

Estimating the Re-Identification Risk: Uniqueness

Uniqueness: A combination of quasi-identifiers (so-called cells, or keys) occurs only once

- The corresponding person is unique with respect to the quasi-identifiers
- As a result, the person can be uniquely identified via a join with external data that contain the quasi-identifiers and direct-identifiers ⇒ Very high re-identification risk

Example: `eusilcP` data; synthetically generated data from real Austrian EU-SILC (European Union Statistics on Income and Living Conditions) data.

```
library("simPop"); data("eusilcP");
head(eusilcP, 2)
##       hid        region hsize eqsize eqIncome pid      id age gender ecoStat citizenship py010n
## 39993   1 Upper Austria     2    1.5    11128   1 0000101  25   male       1       Other  16693
## 39994   1 Upper Austria     2    1.5    11128   2 0000102  24 female       4          AT      0
##       py050n py090n py100n py110n py120n py130n py140n hy040n hy050n hy070n hy080n hy090n hy110n
## 39993      0      0      0      0      0      0      0      0      0      0      0      0      0
## 39994      0      0      0      0      0      0      0      0      0      0      0      0      0
##       hy130n hy145n  main
## 39993      0      0  TRUE
## 39994      0      0 FALSE
```

Estimating the Re-Identification Risk: Uniqueness

```
colnames(eusilcP)
##  [1] "hid"          "region"       "hsize"        "eqsize"       "eqIncome"     "pid"
##  [7] "id"           "age"          "gender"       "ecoStat"      "citizenship" "py010n"
## [13] "py050n"       "py090n"       "py100n"       "py110n"       "py120n"       "py130n"
## [19] "py140n"       "hy040n"       "hy050n"       "hy070n"       "hy080n"       "hy090n"
## [25] "hy110n"       "hy130n"       "hy145n"       "main"
```

We assume the following variables to be quasi-identifiers: `region`, `age`, `gender`, `citizenship`. Then we compute the frequency of each combination of those quasi-identifiers:

```
tab <- as.data.frame(table(
  region = eusilcP$region, age = eusilcP$age,
  gender = eusilcP$gender, citizenship = eusilcP$citizenship,
  useNA = "always"))

head(tab[tab$Freq == 1,], 2)
##         region age gender citizenship Freq
## 304  Carinthia  29   male          AT    1
## 809 Vorarlberg  79   male          AT    1
```

E.g., there is only one male Austrian person with age 29 in region Kärnten.

Estimating the Re-Identification Risk

Further criteria for risk assessment:

- k-**anonymity**: Data satisfy k-anonymity if each cell occurs for at least k respondents
  $\Rightarrow$ reduces the risk of **identity disclosure** (increasing with higher k).
  No unique cases $\Rightarrow$ $k = 2$-*anonymity is satisfied*.
- l-**diversity**: Data satisfy (distinct) l-diversity if, for each cell, there are at least l different values in the sensitive variables
  $\Rightarrow$ reduces the risk of **attribute disclosure** (increasing with higher l).

k-anonymity and l-diversity

	quasi-identifiers		f_k	sensitive var.	l_k
	Gender	Age group		Fare dodger	
1	m	20–29	3	several times	2
2	m	20–29	3	never	2
3	m	20–29	3	never	2
4	f	10–19	3	once	1
5	f	10–19	3	once	1
6	f	10–19	3	once	1

- The two cells m/20–29 and f/10–19 each occur 3 times $\Rightarrow$ the data satisfy (at most) $k = 3$-anonymity.
- Within the cell m/20–29 there are 2 different values with respect to fare dodging, while for f/10–19 there is only one $\Rightarrow$ the data satisfy (only) $l = 1$-diversity.

Estimating the Re-Identification Risk: Uniqueness

Remark: Depending on the software, groups with missing values (NAs) may result:

```
tab[is.na(tab$citizenship) & tab$Freq == 1, ]
##           region age gender citizenship Freq
## 9061  Burgenland   5   male        <NA>    1
## 9111  Burgenland  10   male        <NA>    1
## 10091 Burgenland   8 female        <NA>    1
## 10141 Burgenland  13 female        <NA>    1
```

Such "cells" are usually not counted as "unique cases", because the attacker is likely to not know the missing values in your data.

Methods for data anonymization

Methods for data anonymization include:

1. **Non-perturbative techniques**: Suppress values or reduce detail without altering true data points.
2. **Perturbative techniques**: Insert random deviations from the true value.
3. **Synthetic data generation**: Construct artificial datasets that preserve key statistical properties and relationships.

Software environments such as the R **sdcMicro** package provide algorithms that ensure anonymized datasets satisfy target re-identification risk thresholds (e.g., $k = 5$-anonymity) by selectively modifying a minimal set of values.

Methods for data anonymization

Specific Methods

- **Categorical variables**
  - **Global Recoding**: Combines multiple values or categories into broader, higher-frequency groups. Can also be applied to continuous variables by binning.
  - **Local Suppression**: Replaces specific sensitive values with missing values (`NA`). Typically applied algorithmically after recoding to eliminate residual risk.
  - **Post-Randomization (PRAM)**: Probabilistically swaps categorical values based on a predefined transition probability matrix.
- **Continuous variables**
  - **Microaggregation**: Clusters observations into possibly homogeneous groups, then replaces original values with group summary statistics (e.g., mean). Preserving multivariate covariance is the primary challenge.
  - **Noise Addition**: Adds or multiplies random values to original measurements. Noise can be correlated to preserve covariance structures.
  - **Shuffling**: Swaps continuous values between records while retaining marginal distributions and correlation structures.

Measuring information loss

- **Main Approaches:**
  - **Element-wise comparisons**: Direct measurement of distances or frequencies between the original and anonymized datasets.
    E.g., counting missing/suppressed values, comparing contingency tables.
  - **Quality indicators**: Comparison of key summary statistics (e.g., target or sensitive variables) computed on original vs. anonymized data.
    E.g., differences in point estimates, regression coefficients, correlation matrices.
- While element-wise comparisons are widespread, they provide limited insight into actual analytical utility.

Notes

- Legal aspects (e.g., data protection law) are not covered here
- Challenge: Trade-off between re-identification risk and information value
  Only reduce information that increase the re-identification risk!
- Data anonymization is highly data-specific. Different datasets have varying levels of information value, data sensitivity differs, etc.
- Additional (data-specific) problems:
  - **Randomly sampled data**: A record that is unique in the sample may not be unique in the underlying population ($\Rightarrow$ use of sampling weights).
  - **Household data**: Re-identification via combination of residents, e.g., if there is only one household where a doctor lives with a bricklayer
  - **Longitudinal data**: Re-identification via trajectories, e.g., if there is only one person who moved from Bauma to Meiringen between 2020 and 2021

Recommended literature

- Templ, M. (2017). Statistical disclosure control for microdata. Springer.
- European Data Protection Board EDPB (2025) *Guidelines 01/2025 on Pseudonymisation*. Retrieved from https://www.edpb.europa.eu/... (Accessed: 2025-10-31)
- Hundepool, A. et al. (2025). *Handbook on statistical disclosure control*. In (2. Aufl.). European Commission, Eurostat: Publications Office of the European Union. Retrieved from https://sdctools.github.io/... (Accessed: 2025-10-31)
- Kleiner, B. & Heers, M. (2024). *Quantitative data anonymisation: practical guidance for anonymising sensitive social science data* (Software-Handbuch Nr. 23). Retrieved from https://forscenter.ch/... (Accessed: 2026-01-22)

End of the workshop

Questions?

Literaturangaben i

European Parliament and Council of the European Union. (2016, 4). *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)*. (The regulation was adopted on 27 April 2016 and published in the Official Journal on 4 May 2016. It entered into force on 25 May 2018.)

Schneier, B. (1996). *Applied cryptography: Protocols, algorithms, and source code in c* (2nd ed.). New York: John Wiley & Sons.

Sweeney, L. (2015, September 28). Only you, your doctor, and many others may know. *Technology Science*, *2015092903*.

UNESCO. (2021). *Recommendation on Open Science*. Paris: United Nations Educational, Scientific and Cultural Organization. (Adopted by the General Conference at its 41st session)