2019-11

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Practical data anonymization

Felix Schönbrodt
Ludwig-Maximilians-Universität
München

OSC
LMU Open Science Center

www.nicebread.de
www.researchtransparency.org
@nicebread303

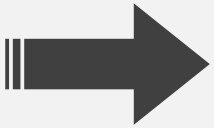
Within sphere of GDPR!

Not relevant for GDPR!

personal data

pseudonymization

anonymous



Scientific Use Files

- Can I share non-anonymized / high risk data?
  - Yes, in a restricted way.
  - Limit access to necessary groups (e.g., researchers): Scientific Use Files
- Neutral third party does access control
  - All users must sign a confidentiality agreement
- If nothing else is possible: At least make metadata open

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

PsychData

Levels of data access

leibniz-psychology.org

PsychArchives

- **0**: Everybody can download data and reuse it for any purpose. Anonymous tracking of download counts.
- **0+**: Like (0), but users have to register and log in; tracking who downloads what.
- **1**: Scientific Use File with standard contract: Only members of research institutions, only scientific use. Contract that no reidentification is attempted; data have to be deleted after usage. Users can download data after signing contract, without interaction with data provider.
- **2**: Custom Scientific Use File: Like (1), but extended with custom contract (can contain anything); data provider have to unlock data set for each request.
- **3**: Secure computing: Users have to (physically) come to the ZPID in Trier and do their computations in a secure room.

https://leibniz-psychology.org/angebote/archivieren/

Higher risk of reidentification

- Data sets with few persons (few rows)
- High-dimensional (many columns)
- Variables with unique values (e.g., only one person has age 39)
- Unique combinations of variables (e.g., the only male with age 39)
- Rare values (e.g., rare diseases)
- Dyadic data: Reidentify yourself, get your partner for free.
- A knowledgeable attacker (jealous romantic partner, co-worker, deine Mudda)
- The most plausible attacker is often near (jealous partner, lab assistant)
- Risk = Damage x Likelihood

Fitness tracking app Strava gives away location of secret US army bases

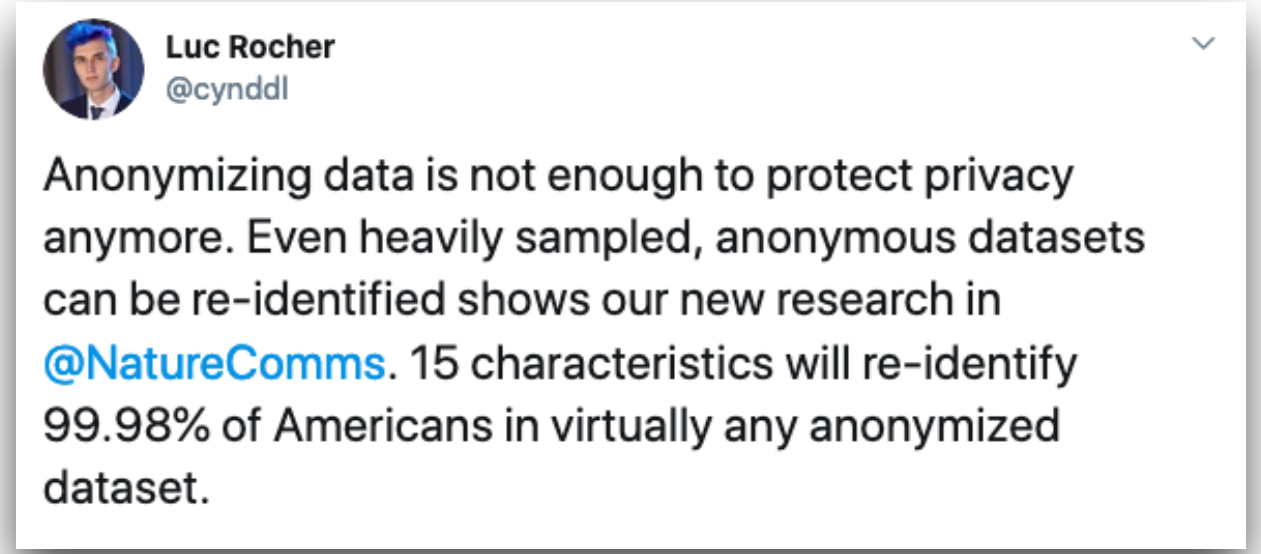
Data about exercise routes shared online by soldiers can be used to pinpoint overseas facilities

- **Latest: Strava suggests military users 'opt out' of heatmap as row deepens**

▲ A military base in Helmand Province, Afghanistan with route taken by joggers highlighted by Strava. Photograph: Strava Heatmap

https://www.theguardian.com/world/2018/jan/28/fitness-tracking-app-gives-away-location-of-secret-us-army-bases?utm_content=buffer74888&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Slide adapted from Ruben Arslan (https://osf.io/n2dsq/)



„For example, date of birth, location (PUMA code), marital status, and gender uniquely identify 78.7% of the 3 million people in this population“

https://twitter.com/cynddl/status/1153711987878223873?s=20
https://aircloak.com/on-the-failure-of-anonymization/

Rocher, L., Hendrickx, J. M., & Montjoye, Y.-A. de. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*, 1–9. doi:10.1038/s41467-019-10933-3

k-anonymity

Samarati, Pierangela; Sweeney, Latanya (1998). "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression" (PDF). Harvard Data Privacy Lab. Retrieved April 12, 2017

k-anonymity

- „A release of data is said to have the **k-anonymity property** if the information for each person contained in the release cannot be distinguished from at least *k*-1 individuals whose information also appear in the release.“
- See Wikipedia: https://en.wikipedia.org/wiki/K-anonymity

k-anonymity

Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Name	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
Anna	21	Weiblich	76189	Grippe
Louis	35	Männlich	77021	Krebs
Holger	39	Männlich	63092	Haarausfall
Frederic	23	Männlich	63331	Muskelzerrung
Anika	25	Weiblich	76121	Grippe
Peter	31	Männlich	77462	Vergiftung
Tobias	38	Männlich	77109	Demenz
Charlotte	19	Weiblich	83133	Karies
Sarah	27	Weiblich	89777	Akne

- Do binning and suppression (i.e., remove values) ➞

k-anonymity

Suppressed

Binned

Binned

Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Name	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
*	20 < Alter < 25	Weiblich	76*	Grippe
*	30 < Alter < 40	Männlich	77*	Krebs
*	20 < Alter < 40	Männlich	63*	Haarausfall
*	20 < Alter < 40	Männlich	63*	Muskelzerrung
*	20 < Alter < 25	Weiblich	76*	Grippe
*	30 < Alter < 40	Männlich	77*	Vergiftung
*	30 < Alter < 40	Männlich	77*	Demenz
*	18 < Alter < 28	Weiblich	8*	Karies
*	18 < Alter < 28	Weiblich	8*	Akne

k-anonymity: Equivalence classes

equivalence class = same values of identifying variables

each equivalence class has at least 2 members —> $k = 2$

	Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Äquivalenzklasse	**Name**	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
A	*	20 < Alter < 25	Weiblich	76*	Grippe
	*	20 < Alter < 25	Weiblich	76*	Grippe

	Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Äquivalenzklasse	**Name**	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
B	*	30 < Alter < 40	Männlich	77*	Krebs
	*	30 < Alter < 40	Männlich	77*	Vergiftung
	*	30 < Alter < 40	Männlich	77*	Demenz

	Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Äquivalenzklasse	**Name**	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
C	*	20 < Alter < 40	Männlich	63*	Haarausfall
	*	20 < Alter < 40	Männlich	63*	Muskelzerrung

	Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Äquivalenzklasse	**Name**	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
D	*	18 < Alter < 28	Weiblich	8*	Karies
	*	18 < Alter < 28	Weiblich	8*	Akne

k-anonymity: Still not anonymous?

	Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Äquivalenzklasse	**Name**	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
A	*	20 < Alter < 25	Weiblich	76*	Grippe
	*	20 < Alter < 25	Weiblich	76*	Grippe

	Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Äquivalenzklasse	**Name**	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
B	*	30 < Alter < 40	Männlich	77*	Krebs
	*	30 < Alter < 40	Männlich	77*	Vergiftung
	*	30 < Alter < 40	Männlich	77*	Demenz

	Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Äquivalenzklasse	**Name**	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
C	*	20 < Alter < 40	Männlich	63*	Haarausfall
	*	20 < Alter < 40	Männlich	63*	Muskelzerrung

	Identifikator	Quasi-Identifikatoren			Sensibles Attribut
Äquivalenzklasse	**Name**	**Alter**	**Geschlecht**	**PLZ**	**Krankheit**
D	*	18 < Alter < 28	Weiblich	8*	Karies
	*	18 < Alter < 28	Weiblich	8*	Akne

Homogeneity Attack:
All members of a class have the same sensitive attribute.

Background knowledge attack:
The neighbor of Bob (who has age 35) knows his postal code (63*; well, he's his neighbor) and knows that Bob is in this table. He also knows that Bob has no hair loss
—> reidentified

Choice of *k*?

- „Even though a minimum *k* value of 3 is often suggested, a common recommendation in practice is to ensure that there are **at least five similar observations (k = 5).** It is uncommon for data custodians to use values of *k* above 5, and quite rare that values of *k* greater than 15 are used in practice." (El Emam & Dankar, 2008)

- „Based on the recommendations made in the IOM report and the available precedents for public release of health data, EMA believes that it is advisable to **set the threshold to a conservative level of 0.09.** […] Further information about the methodology to calculate the risk of re-identification is available in the literature, such as for instance that the probability of re-identification of a record in a data set **is 1 divided by the frequency of trial participants with same category/value of a set of the quasi identifiers (group size).**" (External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use)

  - A threshold of 0.09 implies ***k* =1 / 0.09 = 11**

- See also http://www.appliedclinicaltrialsonline.com/de-identifying-clinical-trials-data

Choice of *k*?

- „Da für einen Antrag aus dem BMG, bei dem auch das Regionalkennzeichen auf Kreisebene ausgewertet wurde, bereits vom BMG und der BfDI eine Absenkung der MFZ auf 5 ohne Nachweis der Erfordernis als akzeptabel bewertet wurde, **könnte es politisch akzeptabel sein, eine MFZ von 5 einheitlich vorzugeben.**“
  MFZ = Mindestfallzahl; Informationssystem Versorgungsdaten (Datentransparenz) Evaluationsbericht 07/2013-02/2016 Teil 2: Bewertungsmatrix der Handlungsoptionen

- At another place, they discuss choices of ***k*=10** and ***k*=30** and prefer the latter.

1.4.4 Empfehlung

Hier kann von Seitens der Datenaufbereitungsstelle bzw. des DIMDI keine klare Empfehlung ausgesprochen werden. Für eine gesetzliche Fixierung der MFZ auf 10 oder gar 5 spricht, dass damit

- die Untersuchung von insbesondere seltenen Ereignissen oder Merkmalskombinationen erleichtert wird,
- die Datenaufbereitungsstelle von der Pflicht entbunden wird, den potentiellen Erkenntnisgewinn und das Re-Identifizierungsrisiko der Versicherten im Falle eines Antrags auf Senkung der MFZ gegeneinander abzuwägen und
- die Antragsbearbeitung damit zumindest etwas beschleunigt wird.

Gegen eine gesetzliche Fixierung der MFZ auf 10 oder gar 5 spricht, dass

- damit das Re- Identifizierungsrisiko der Versicherten steigt und
- sich die Frage stellt, ob auf Basis derart kleiner Fallzahlen noch aussagekräftige Schlussfolgerungen gezogen werden können, sich also der angestrebte Erkenntnisgewinn tatsächlich realisieren lässt.

p. 34

Practical anonymization

Practical Anonymization

1. Remove obviously identifying variables (names, email addresses, addresses, phone numbers, social security numbers, birth date, IP address, …)
   ➞ 18 HIPAA Identifiers
2. Remove variables that are not necessary to reproduce your results
   - E.g., meta-data such as login time, server log entries, browser and OS version, free text entries
3. Identify the identifying variables („key variables“, i.e., those variables which an attacker could know)
4. Detect unique variable values in your data set of all key variables
   (univariate, i.e. only looking at one variable each time)

Practical Anonymization

5. Make them *k*-anonymous (choose *k*).
   Most common techniques:
   - Binning (age ➞ age groups, rare categories ➞ ,,other'')
   - Bottom / Top-Coding (= Winsorizing): cap high and low outliers at some value (e.g., all incomes > 200.000 € are set to 200.000)
   - Microaggregation: Construct small clusters, aggregate only some rows of a variable by replacing them with the mean.
   - Fuzzing (e.g., add 0.1 standard deviations of noise)
   - Suppression: Remove unique entries
   - Recoding: E.g., recode specific dates to ,,days since study start"

- The **research question** (What do I need the data for?) determines how you anonymize: How you define the boundaries of the bins, which variable you start with, whether you prefer binning or microaggregation, …

Practical Anonymization

6. Create *k*-anonymity for unique *combinations* (multivariate, i.e. looking at all possible combinations of variables)
7. ! Anonymization techniques change the truthfulness of a data set (you lose information!) → find a good balance between anonymity and reusability

- Loss of information is proportional to the choice of *k*

1. Remove obviously identifying variables (names, email addresses, addresses, phone numbers, social security numbers, birth date, IP address, …)
2. Remove variables that are not necessary to reproduce your results
3. Identify the identifying variables („key variables“, i.e., those variables which an attacker could know)
4. Detect unique variable values in your data set of all key variables (univariate, i.e. only looking at one variable each time)
5. Make them *k*-anonymous (choose *k*).
   Most common techniques:
   - Binning (age ➞ age groups, rare categories ➞ „other“)
   - Bottom / Top-Coding (= Winsorizing): cap high and low outliers at some value (e.g., all incomes > 200.000 € are set to 200.000)
   - Microaggregation: Construct small clusters, aggregate only some rows of a variable by replacing them with the mean.
   - Fuzzing (e.g., add 0.1 standard deviations of noise)
   - Suppression: Remove unique entries
6. Create *k*-anonymity for unique *combinations* (multivariate, i.e. looking at all possible combinations of variables)

Let's practice in *R*!

Excursus: Read and write encrypted files in R

- We want a solution where you do not have to type in the password every time, but also do not store the plain password in your R script.
- keyringr package allows to access the operating system's password manager (macOS: Keychain, Windows Data Protection API, Linux Gnome Keyring)
- Store credentials on an encrypted USB drive, the R script points to the key file location, OS asks for password.
  - https://blog.revolutionanalytics.com/2015/12/securely-storing-your-secrets-in-r-code.html

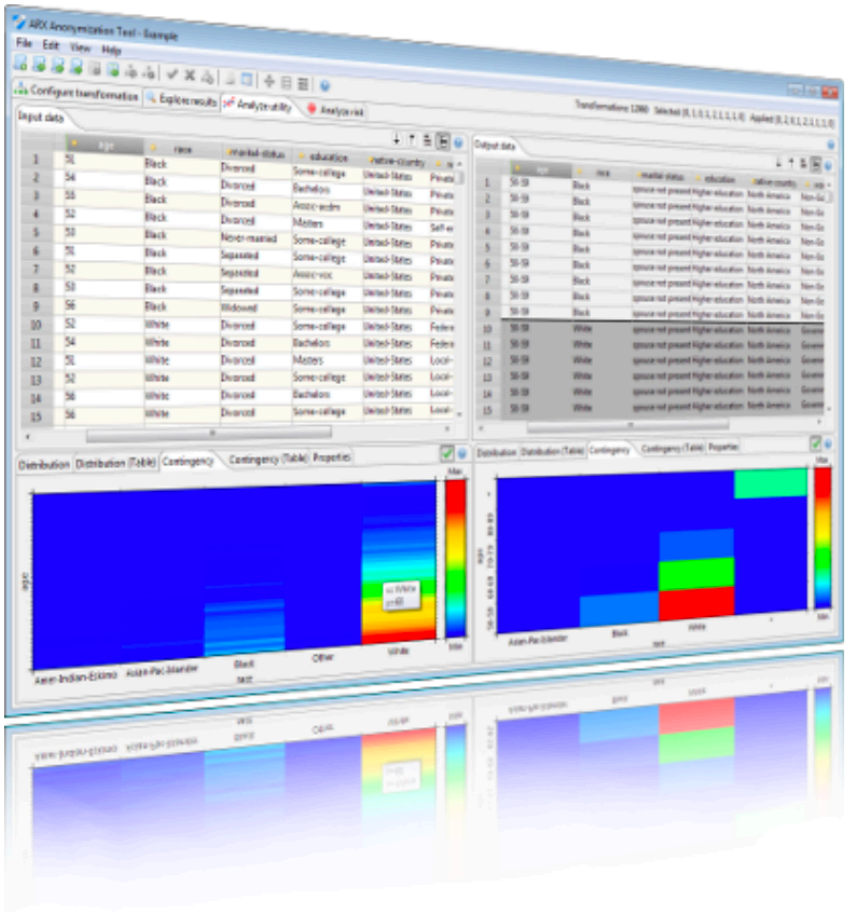
ARX

Data Anonymization Tool

ARX is a comprehensive open source software for anonymizing sensitive personal data. It supports a wide variety of (1) privacy and risk models, (2) methods for transforming data and (3) methods for analyzing the usefulness of output data.

The software has been used in a variety of contexts, including commercial big data analytics platforms, research projects, clinical trial data sharing and for training purposes.

ARX is able to handle large datasets on commodity hardware and it features an intuitive cross-platform graphical user interface. You can find further information here, or directly proceed to our downloads section.



https://arx.deidentifier.org/

Outlook:
What holds the future?

What does the future hold?

openhumans.org

midata.coop

33n

Give participant more insight, control over their data?
Move beyond traditional notions of privacy?

Slide adapted from Ruben Arslan (https://osf.io/n2dsq/)

Personal Health Train

https://www.dtls.nl/fair-data/personal-health-train/

Synthetic data

- Create a synthetic data set that recreates key statistical properties of the original data set, but no single row is identical to the original.
- „… will only be valid if the model used to construct the synthetic data is the true mechanism that has generated the observed data, which is very difficult, if at all possible, to achieve“ (Nowok, Raab, & Dibben, 2016).
- Use case: Other researchers can explore the synthetic dataset without risk of reidentification, final analysis is run (maybe by data holders) on original data set.
- What could be recreated?
  - Means, min, max, SDs of each column
  - Correlation matrix
  - Granularity/specific values of variables
  - preserve the missing value structure
- e.g., *R* package *synthpop* (e.g., http://gradientdescending.com/generating-synthetic-data-sets-with-synthpop-in-r/)
- Generative Adversarial Nets (GANs): A generator neural net creates synthetic data; an independent discriminator net tries to distinguish the new data from the original data set. Both learn form each other, until the discriminator cannot distinguish generated from real data any more.

And even more …

- Secure multi-party computation
- Differential privacy
- In-database computation: Analysts never see the raw data; send their analytical queries to the data base
- …