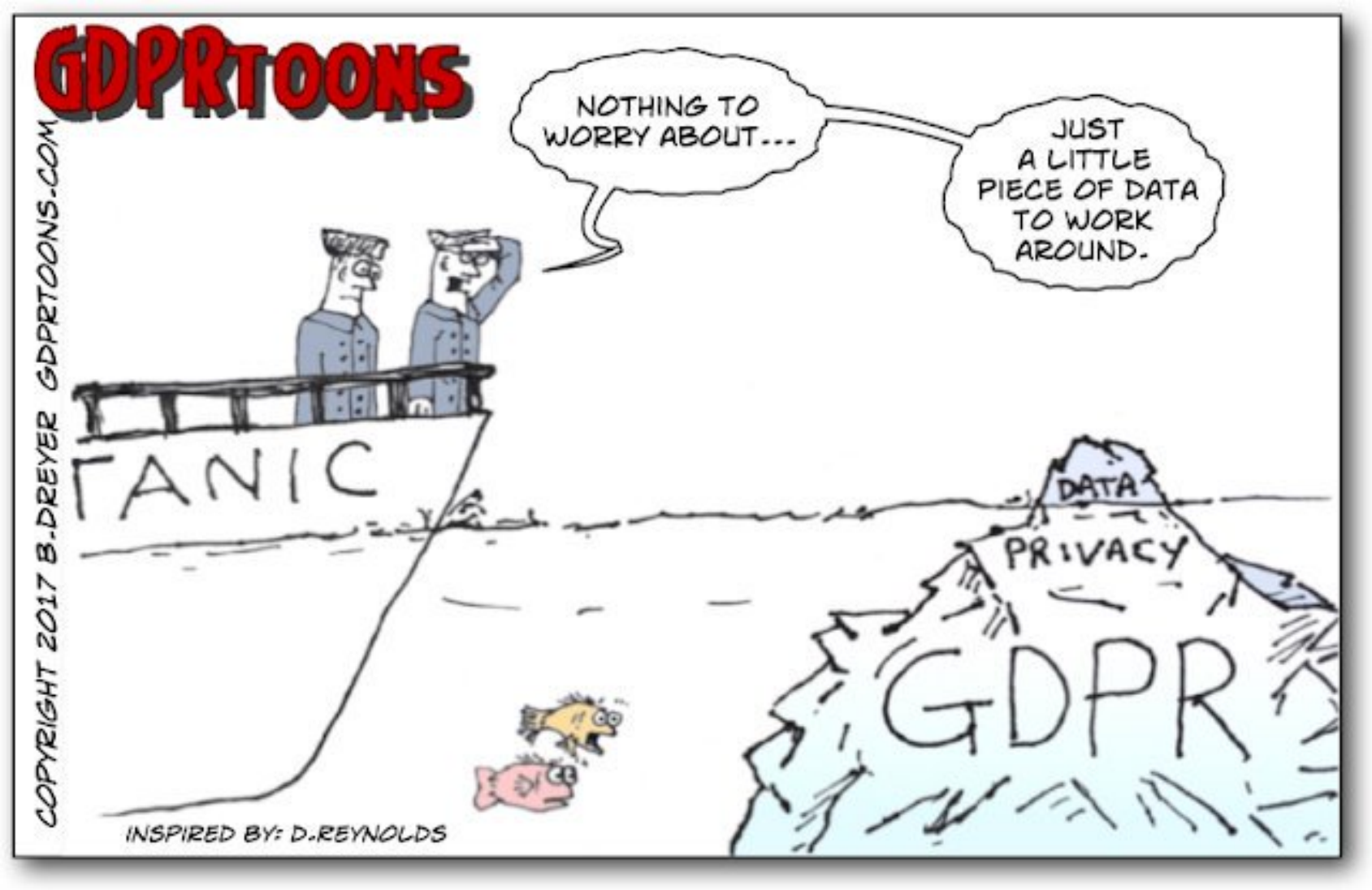
Stakeholder-Day

on Data Anonymization and Data Synthetization

SwissAnon Team

25.03.2024 (Olten)

The reason why we are here



Agenda

Part I

1. Round of introductions
2. SNF Bridge Discovery Project
   - Research
   - Software + brief work-in-progress
3. Swiss Data Anonymization Competence Center
4. A few unresolved in the Swiss data protection law and the GDPR

Part II

5. Discussion, Collaboration, Ideas, Exchange
   - What are unsolved issues in practice with respect to anonymization? Gaps in Switzerland.
   - What do you expect from the competence center to deliver?
   - What are the next steps? Potential joint activities and opportunities for collaboration.

Part III

6. Apero

Introductionary round

Important for networking between **all of you**.

Our team

https://swissanon.ch/team/

Introductionary round

Important for networking between **all of you**.

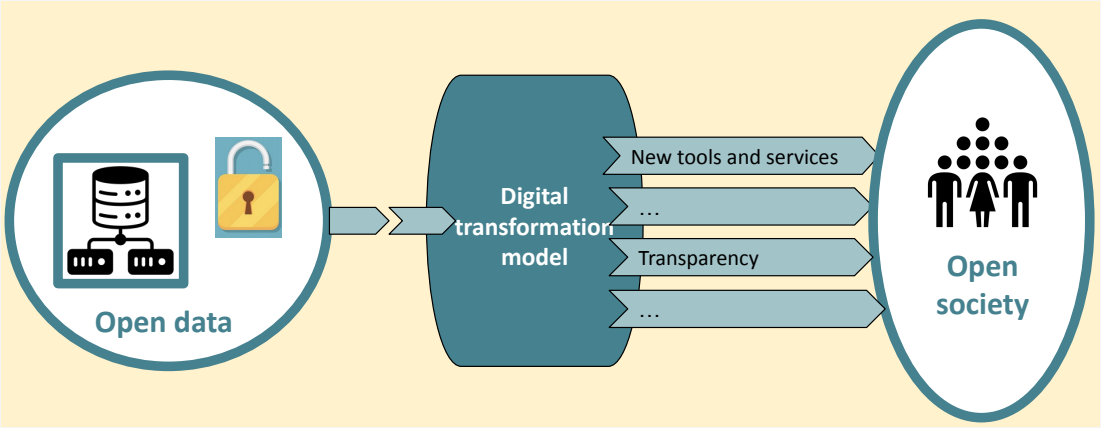
Our team

https://swissanon.ch/team/

Your turn

- Please introduce yourself
- Introduce yourself and tell us about your specific interests in data anonymization/synthetization

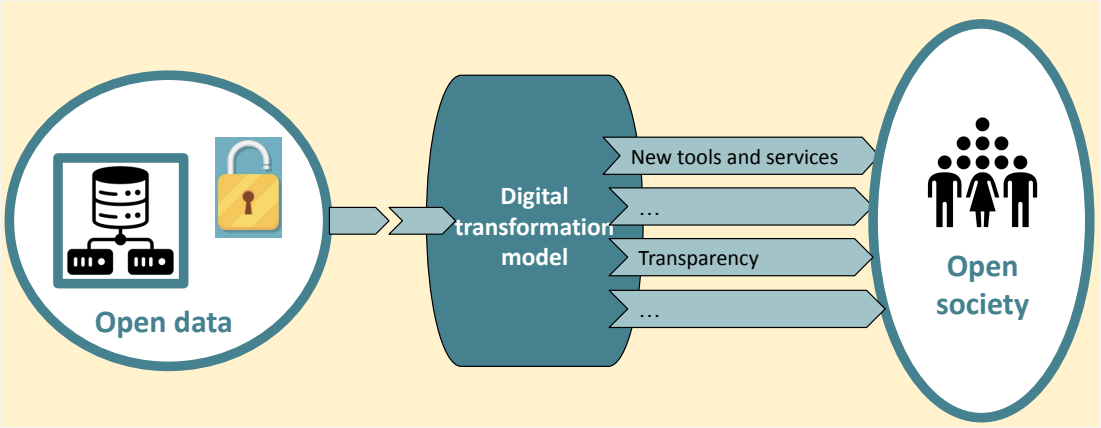
Challenge. Open society versus privacy of personal data



1. But what about **personal data** whose use is restricted by data protection laws?

open **anonymized** data

Challenge. Open society versus privacy of personal data

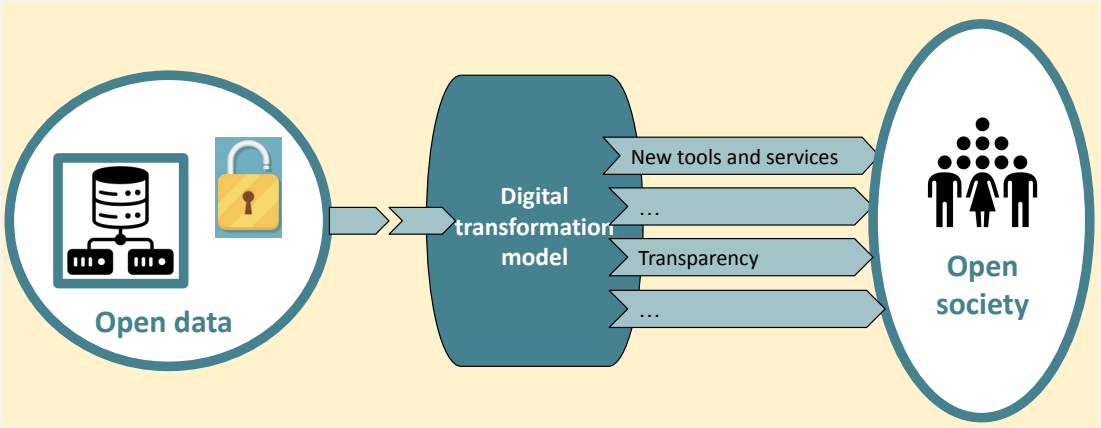


1. But what about **personal data** whose use is restricted by data protection laws?

open **anonymized** data

2. Longitudinal information shows processes over time

Challenge. Open society versus privacy of personal data



1. But what about **personal data** whose use is restricted by data protection laws?

open **anonymized** data

2. Longitudinal information shows processes over time

Need: Anonymisation methods for longitudinal data

Research objectives. Example Mobility Tracking

Example: Tracking people's movement pattern throught mobile phones including customer information (age, gender, nat., . . .)

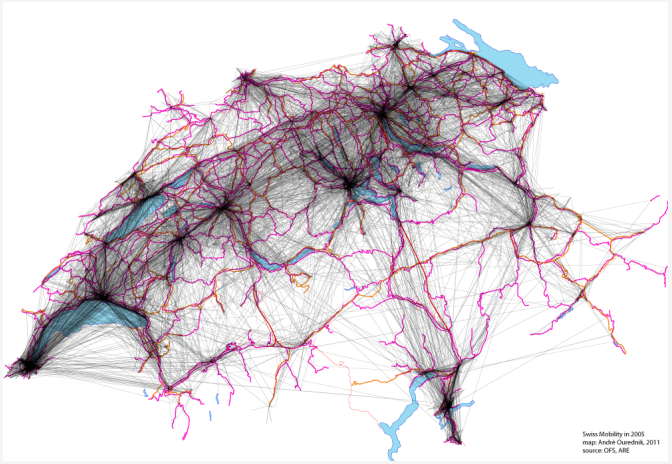


Figure 1: Mobility in Switzerland: Microcensus on transport behaviour 2005. Source: BFS

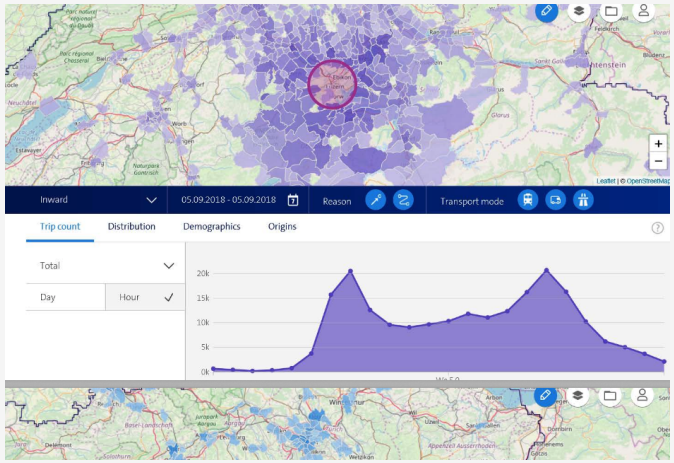


Figure 2: Swisscom's mobility insight platform with movement statistics

Needs: Methods and free and open-source Software

Access to data: Better quality, cost-efficient, and very useful for transport and traffic planners.

Research objectives. Example Public Health

Researchers and institutions: easy and cost-effective access to health and surveillance data. Analysis of the health system.

ident	case	typind	intdate	birth_date	sex	maritalstatus	schoolever	schooltype	schoolstd
<chr>	<dbl+l>	<dbl+lb>	<date>	<date>	<dbl+l>	<dbl+lbl>	<dbl+lbl>	<dbl+lbl>	<dbl>
2397…	0 [Con…	1 [Ran…	2009-05-25	1944-06-15	2 [Mal…	2 [Married]	1 [Ever]	1 [Prima…	8
2617…	1 [Cas…	NA	2002-07-28	1955-06-15	2 [Mal…	NA	NA	NA	NA
1040…	0 [Con…	3 [Sic…	2008-11-10	1978-06-15	1 [Fem…	3 [Divorced…	1 [Ever]	1 [Prima…	7
1040…	1 [Cas…	NA	2008-11-12	1978-06-15	1 [Fem…	3 [Divorced…	1 [Ever]	1 [Prima…	7
1040…	1 [Cas…	NA	2009-10-30	1978-06-15	1 [Fem…	3 [Divorced…	1 [Ever]	1 [Prima…	5
5382…	1 [Cas…	NA	1992-04-07	1924-06-15	2 [Mal…	NA	NA	NA	NA
1262…	1 [Cas…	NA	1991-02-25	1935-06-15	1 [Fem…	NA	NA	NA	NA
4792…	1 [Cas…	NA	1999-01-18	1961-06-15	1 [Fem…	5 [Previous…	1 [Ever]	1 [Prima…	8
1205…	1 [Cas…	NA	1986-01-21	1920-06-15	2 [Mal…	NA	NA	NA	NA
3122…	0 [Con…	3 [Sic…	2008-09-19	1945-06-15	1 [Fem…	3 [Divorced…	0 [Never]	NA	NA

… with 100 more variables: occupation <dbl+lbl>, occgphead <dbl+lbl>, prevtbtreat <dbl+lbl>, prevtbtreatyr <dbl>, tbtype <dbl+lbl>, tbcpu <dbl+lbl>, mic_resH <dbl+lbl>, cultposever <dbl+lbl>, outcome_cpu <dbl+lbl>, starteps <date>, fineps <date>, cough <dbl+lbl>, swgl <dbl+lbl>, practype <dbl+lbl>, tbtype1 <dbl+lbl>, tbtype2 <dbl+lbl>, tbtype3 <dbl+lbl>, mic_resD <dbl+lbl>, mic_res2m <dbl+lbl>, mic_resC <dbl+lbl>, cult_res <dbl+lbl>, cult_res2m <dbl+lbl>, cult_resC <dbl+lbl>, drugres_s <dbl+lbl>, drugres_i <dbl+lbl>, drugres_r <dbl+lbl>, drugres_e <dbl+lbl>, drugres_p <dbl+lbl>, drugres_c <dbl+lbl>, …

Figure 3: MEIRU Population Health and Surveillance data.

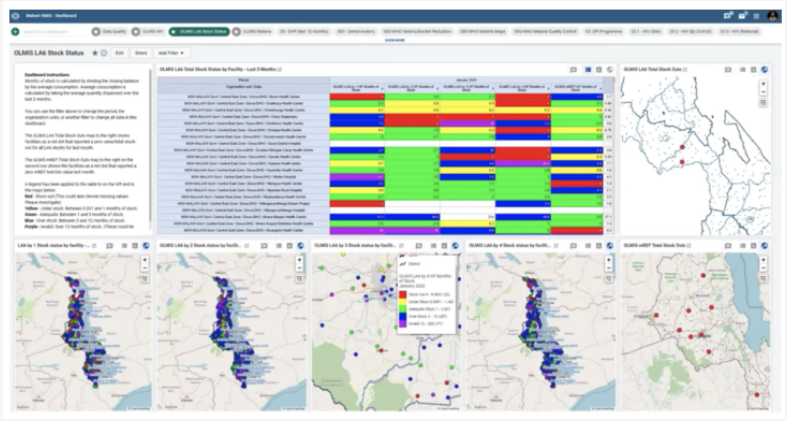


Figure 4: Open health management system / Malawi.

Research objectives. Example Public Health

Researchers and institutions: easy and cost-effective access to health and surveillance data. Analysis of the health system.

ident	case	typind	intdate	birth_date	sex	maritalstatus	schoolever	schooltype	schoolstd
<chr>	<dbl+l>	<dbl+lb>	<date>	<date>	<dbl+l>	<dbl+lbl>	<dbl+lbl>	<dbl+lbl>	<dbl>
2397…	0 [Con…	1 [Ran…	2009-05-25	1944-06-15	2 [Mal…	2 [Married]	1 [Ever]	1 [Prima…	8
2617…	1 [Cas…	NA	2002-07-28	1955-06-15	2 [Mal…	NA	NA	NA	NA
1040…	0 [Con…	3 [Sic…	2008-11-10	1978-06-15	1 [Fem…	3 [Divorced…	1 [Ever]	1 [Prima…	7
1040…	1 [Cas…	NA	2008-11-12	1978-06-15	1 [Fem…	3 [Divorced…	1 [Ever]	1 [Prima…	7
1040…	1 [Cas…	NA	2009-10-30	1978-06-15	1 [Fem…	3 [Divorced…	1 [Ever]	1 [Prima…	5
5382…	1 [Cas…	NA	1992-04-07	1924-06-15	2 [Mal…	NA	NA	NA	NA
1262…	1 [Cas…	NA	1991-02-25	1935-06-15	1 [Fem…	NA	NA	NA	NA
4792…	1 [Cas…	NA	1999-01-18	1961-06-15	1 [Fem…	5 [Previous…	1 [Ever]	1 [Prima…	8
1205…	1 [Cas…	NA	1986-01-21	1920-06-15	2 [Mal…	NA	NA	NA	NA
3122…	0 [Con…	3 [Sic…	2008-09-19	1945-06-15	1 [Fem…	3 [Divorced…	0 [Never]	NA	NA

… with 100 more variables: occupation <dbl+lbl>, occgphead <dbl+lbl>, prevtbtreat <dbl+lbl>, prevtbtreatyr <dbl>, tbtype <dbl+lbl>, tbcpu <dbl+lbl>, mic_resH <dbl+lbl>, cultposever <dbl+lbl>, outcome_cpu <dbl+lbl>, starteps <date>, fineps <date>, cough <dbl+lbl>, swgl <dbl+lbl>, practype <dbl+lbl>, tbtype1 <dbl+lbl>, tbtype2 <dbl+lbl>, tbtype3 <dbl+lbl>, mic_resD <dbl+lbl>, mic_res2m <dbl+lbl>, mic_resC <dbl+lbl>, cult_res <dbl+lbl>, cult_res2m <dbl+lbl>, cult_resC <dbl+lbl>, drugres_s <dbl+lbl>, drugres_i <dbl+lbl>, drugres_r <dbl+lbl>, drugres_e <dbl+lbl>, drugres_p <dbl+lbl>, drugres_c <dbl+lbl>, …

Figure 3: MEIRU Population Health and Surveillance data.

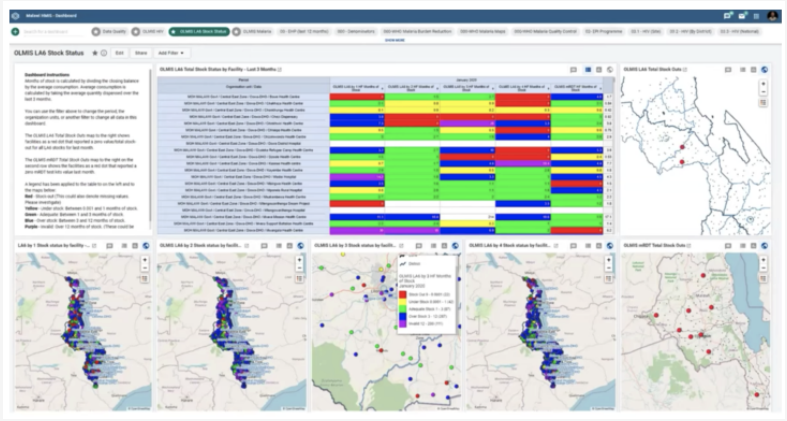


Figure 4: Open health management system / Malawi.

Needs: Software and methods anonym./synthetic data.

- Different number of events for persons + event order is crucial
- Information changes over time, e.g. careers ↗ disclosure risk

Summary research gap to be solved

1. Trajectory/movement data:
   - Algorithms are computational inefficient, not proven regarding data utility after anonynmization, spares out socio-economic information.
2. Event history data:
   - Surveillance data: Insufficient approach (Heldal 2011). Problem definition: Templ, Kanjala, and Siems (2022)
   - hospital@home event data: no high-quality open data available.
   - (anonymization of network data including time-related information)
3. Disclosure risk and data utility of complex data
4. Many companies business model: synthetic data using deep learning methods
   - adapted methods are not disclosed
   - Synthetic (twin) data is not even a silver bullet (Stadler, Oprisanu, and Troncoso 2020) $\Rightarrow$ **Innovation needed**

Challenges when anonymizing longitudinal data (1/2)

- **Temporal Uniqueness**: The patterns of behavior or events over time can make an individual recognizable. (e.g. sequence of medical treatments).
- **Re-identification Risks**: More data points about an individual $\Rightarrow$ higher re-identification risk
- **Consistency in Anonymization**: Ensuring consistent anonymization across time points is challenging, e.g. event order.
- **Data Granularity:** The finer the time granularity provided (e.g., hourly data vs. daily data), the easier to identify persons and the harder to anonymize.

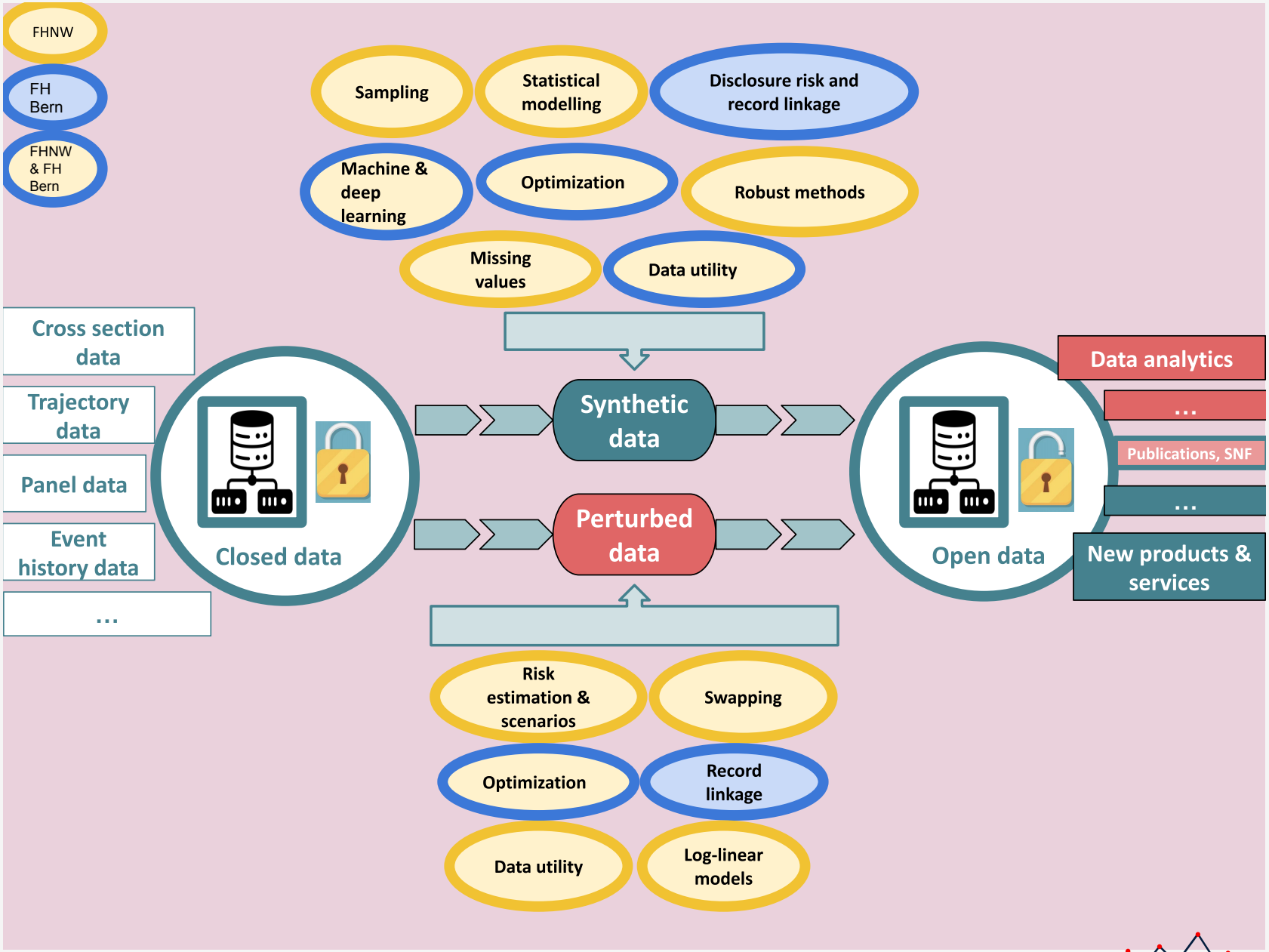
Challenges when anonymizing longitudinal data (2/2)

- **Loss of Data Utility:** Higher for longitudinal data thus more anonymization.
- **Updating Anonymized Data:** As new data points are added over time, ensuring that the updated dataset remains anonymous and consistent.
- **Dynamic Features:** In some datasets, features or variables might be added, removed, or modified over time.
- ...

And challanges are even bigger to synthetize longitudinal data, because of considering

- variablities over time
- preserving temporal dependencies
- data drift (primaritly covariate shift)
- consistency and validity (age increase as simplified example)

The road to open data



Software from us

about consulting references research publications repositories team contact

repositories

Members of our team have co-developed several extensive R packages for statistical confidentiality, synthetic data generation and record linkage

cran/RecordLinkage

! This is a read-only mirror of the CRAN R package repository. RecordLinkage — Record Linkage Functions for Linking and Ded

C 6 13

sdcTools/sdcMicro

sdcMicro

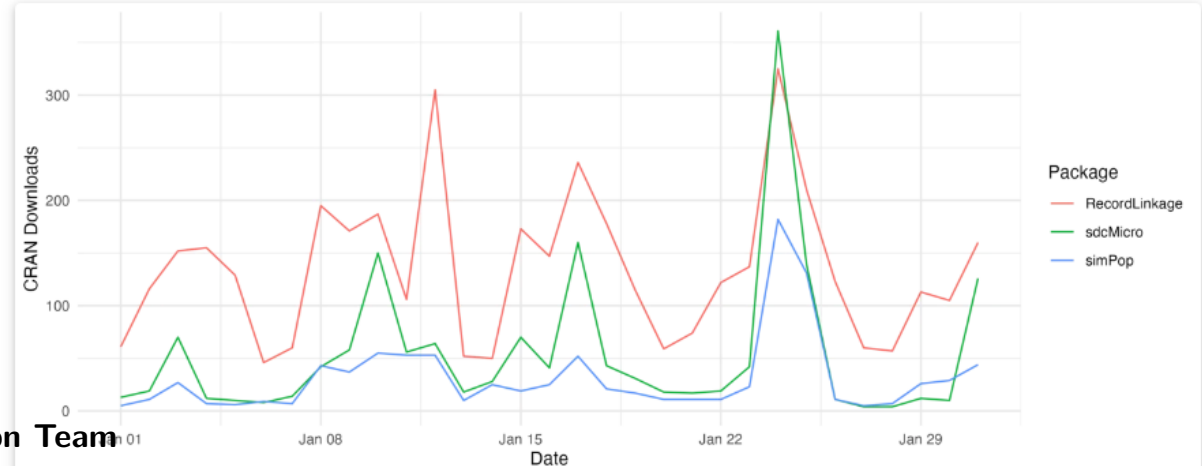
R 73 23

statistikat/simPop

Simulation of Synthetic Populations for Survey Data Considering Auxiliary Information

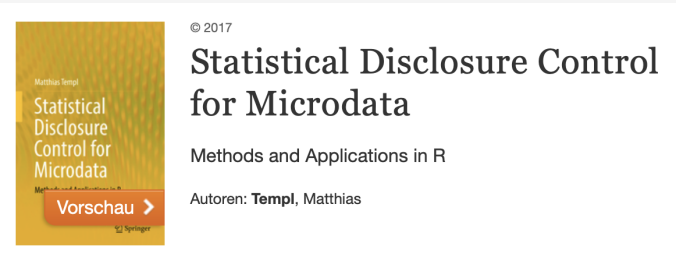
R 28 7

CRAN Downloads through RStudio Server for the Month of January 2024



Software 1: sdcMicro - Statistical Disclosure Control

- Developed since 2007, S4-class implementation
- Core-Publication Journal of Statistical Software 2015 [sdcMicro]
- "sdcMicro" Springer-Book Statistical Disclosure Control [Link]



- Core tool from Eurostat's working group sdcTools
- Several funds from World Bank, Eurostat, etc.
- Suggested tool by World Bank and IHSN: https://www.ihsn.org/software/disclosure-control-toolbox

Software 1: sdcMicro - Statistical Disclosure Control

Data anonymization for cross-sectional data:

Method | Software	μ-**Argus**	**BioMed Tools**	**sdcMicro**	**sdcMicroGUI**	**IHSN**
recoding	✔	✔	✔	✔	✔
k-anonymity	✔	✔	✔	✔	✔
l-diversity		✔	✔	✔	✔
t-closeness, etc.		✔			
suda2			✔		✔
individual risk (IR)	✔		✔	✔	✔
IR on households	✔		✔	✔	✔
GR with log-lin mod.			✔		
recoding	✔	✔	✔	✔	(✔)
local suppression	(✔)	(✔), Arx: ✔	✔	✔	(✔)
pram	✔		✔	✔	✔
target record swapping			✔		
adding correlated noise			✔	✔	✔
microaggregation	✔	Arx: ✔	✔	✔	✔
shuffling			✔	✔	
utility measures	(✔)	Arx: (✔)	✔	✔	
reproducibility			✔	✔	✔
GUI	(✔)	✔		✔	
CLI		Arx, Amnesia: ✔	(✔)		✔
reporting	✔		✔	✔	
platform independent		Arx: ✔	✔	✔	✔
free and open-source		Arx, Amnesia: ✔	✔	✔	✔

Software 2: simPop - Synthetic data

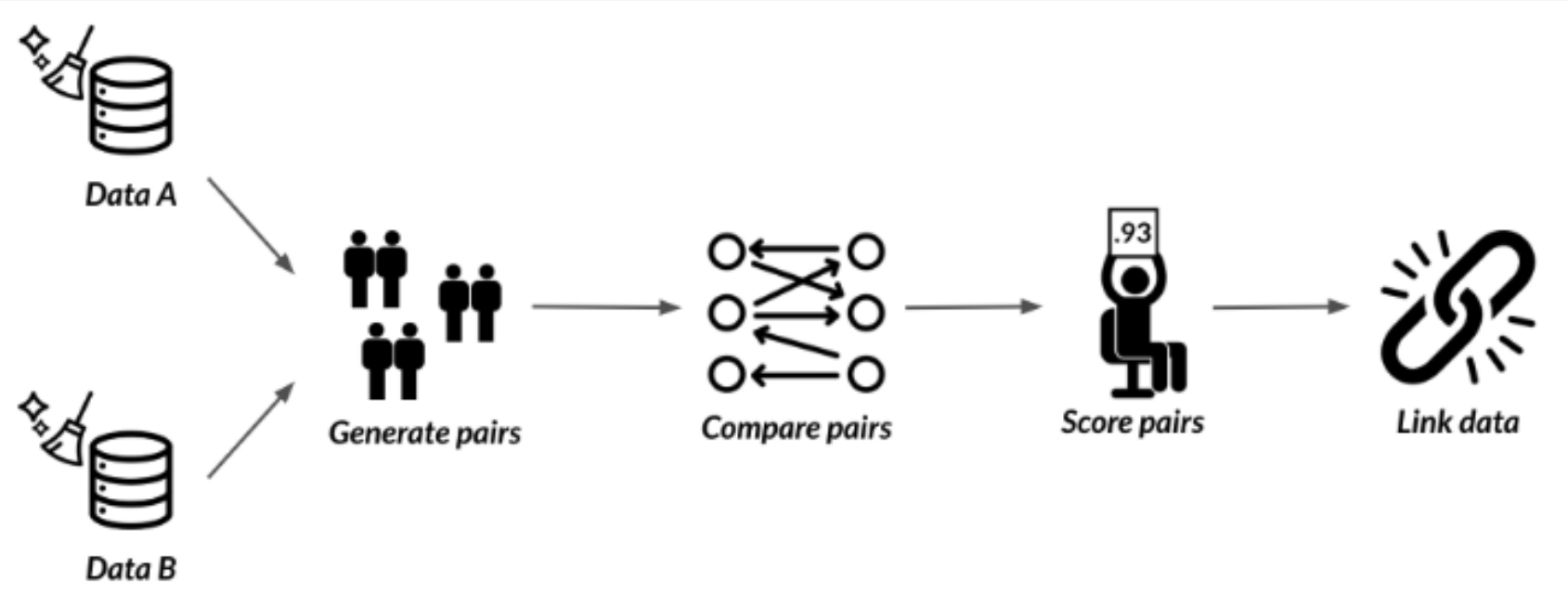
- Synthetic generation of "twin'' data sets using machine learning methods
- Can consider hierarchical and cluster structures of data
- Can deal with samples from populations
- Can construct realistic structures, e.g. sensible household compositions of persons.

[CRAN]

[JSS Paper]

Software 3: RecordLinkage

- Functions for linking and deduplicating data sets.



More info: Sariyar and Borg (2010)

[RJ Paper]

[CRAN]

Translation and innovation

- Translation after methodological research.

> We research, we develop generic FOSS software, we produce open data and we consult $\Rightarrow$ Leadership

Translation and innovation

- Translation after methodological research.

> We research, we develop generic FOSS software, we produce open data and we consult $\Rightarrow$ Leadership

- Innovation with strong economic impact foreseeable
  - **Swiss competence centre on data anonymisation**
  - Consultancy and applications for companies and institutions
- Innovation with strong societal impact
  - without open data no digital transformation
  - **open (anonymized) data $\Rightarrow$ open society** (new products, democracy/transparency, quality), ...

Translation and innovation

- Translation after methodological research.

> We research, we develop generic FOSS software, we produce open data and we consult ⇒ Leadership

- Innovation with strong economic impact foreseeable
  - **Swiss competence centre on data anonymisation**
  - Consultancy and applications for companies and institutions
- Innovation with strong societal impact
  - without open data no digital transformation
  - **open (anonymized) data ⇒ open society** (new products, democracy/transparency, quality), ...

> Enable digital transformation by the development of **anonymization/synthetization** methods for **longitudinal person-related** data.

Legal reasons for anonymization

Both under European and Swiss law, there are two legal grounds for anonymization:

- **Protecting personal data**: with anonymous data no consent or legal basis for processing is necessary (Rec. 26: "The protection principles do not apply to data that is anonymized to the extent that the data subject is no longer identifiable.")
- **Principle of data minimization**: Personal data must be de-identified as much as possible without jeopardizing the research purpose.

A few unresolved in the GDPR

GDPR Article 4: "... an identifiable person is one who can be identified directly or indirectly by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity."

- **Circular definition**: indication that there is a mixture of concepts, e.g., Criterion of identity versus principle of individuation
- **Philosophical insights might help**: epistemology (how to identify?) versus ontology (what is real-world identity in fact?)
- **In practice**: When are we safe with our privacy protection measures? Unresolved issue

A few unresolved in the GDPR

- **Criterion of identity**: standard for determining whether two things are the same
- **Principle of individuation**: features that distinguish one thing from another
- **In practice**: Relational identity for identification in addition to property-based identification (ontology of identity is still an issue)

> **Example**: the name of a patient is not just an attribute of the patient, but a relation between the patient and those authorities that have certified and validated that name assignment. Hence, relations such as "identified by" or "certified by" are part of the relational identity of the patient.

More info: Sariyar and Holm (2022) [Paper on identity]

Relative Anonymity

- **Patrick Beyer versus Germany case (European Court of Justice)**: "IP address is personal data when held by an ISP, but does not constitute personal data if held by a party that does not have the “means likely reasonably to be used to identify the individual"
- **Even without IP addresses we might have a problem (Recital 30 of the GPDR)**: "Natural persons may be associated with online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags. This may leave traces which, in particular when combined with unique identifiers and other information received by the servers, may be used to create profiles of the natural persons and identify them."

Anonymization vs Pseudonymization

- **Definition of pseudonymization of WP29 is too broad for a differentiation**: "Processing personal data in such a way that it can no longer be attributed to a specific “data subject” without the use of additional information, which must be kept separately and be subject to technical and organisational measures to ensure non-attribution."
- **Anonymous data**: "If any possible actor is unable to directly or indirectly identify data subjects with means reasonably likely to be used now or in the future."
- **Strictly pseudonymous data**: "if in presence of technical and organizational measures of the pseudonymization (not general), the intended recipients are unable to directly identify data subjects."

In Switzerland

- Data is deemed to relate to an identified person if it is linked to a specific person. GDPR requires only a singling-out.
- De-facto anonymity is the standard, not absolute anonymity
- Art. 32 & 33 of HRA shows that consent is not required to anonymize non-genetic health-related data.
- De-identification results in anonymized or pseudonymized data.
- Cave: Discrimination is still possible with anonymized data
- Data Protection Impact Assessments and Mandatory Security Breach Reporting are not yet fully standardized
- Finnish Social Science Archive (FSD) in contrast: 42 characteristics to be removed, but singling-out, linkability, and attribute inference may still be possible.

Swiss Data Anonymization Center

- Supports you in secure data anonymization according to the requirements of the GDPR and Swiss national laws on data privacy.
- It leads research on key topics in data anonymization and synthesis of data, and it serves researchers, businesses, and practitioners with tailored solutions

Consulting + Research + Software

https://swissanon.ch

References I

Heldal, J. 2011. "ANONYMISED INTEGRATED EVENT HISTORY DATASETS FOR RESEARCHERS." In *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Tarragona, Spain*.

Sariyar, Murat, and Andreas Borg. 2010. "The RecordLinkage Package: Detecting Errors in Data." *The R Journal* 2 (2): 61–67. https://doi.org/10.32614/RJ-2010-017.

Stadler, Theresa, Bristena Oprisanu, and Carmela Troncoso. 2020. "Synthetic Data – Anonymisation Groundhog Day." arXiv. https://doi.org/10.48550/ARXIV.2011.07018.

Templ, M. 2017. *Statistical Disclosure Control for Microdata: Methods and Applications in R*. Cham, Switzerland: Springer International Publishing.

Templ, M., C. Kanjala, and I. Siems. 2022. "Privacy of Study Participants in Open-Access Health and Demographic Surveillance System Data: Requirements Analysis for Data Anonymization." *JMIR Public Health Surveill* 8 (9): e34472. https://doi.org/10.2196/34472.

Templ, M., A. Kowarik, and B. Meindl. 2015. "Statistical Disclosure Control for Micro-Data Using the R Package SdcMicro." *Journal of Statistical Software, Articles* 67 (4): 1–36. https://doi.org/10.18637/jss.v067.i04.

Templ, M., B. Meindl, A. Kowarik, and O. Dupriez. 2017. "Simulation of Synthetic Complex Data: The R Package SimPop." *Journal of Statistical Software* 79 (10): 1–38. https://doi.org/10.18637/jss.v079.i10.