Universität Zürich[UZH]

# Data anonymization: A workflow

## How and why anonymize data

Jiří Novák

November 20, 2024

# About me – Jiří Novák

- From Prague, Czech Republic,
- PhD in Statistics
- Research topics
  - Data privacy  and data anonymization
  - Statistical disclosure control

- Responsible for data confidentiality of Czech Census 2021
- Currently working at FHNW on a SNF project: *Harnessing event and longitudinal data in industry and health sector through privacy preserving technologies*
- PhD student of Carolin Strobl with topic*: Data anonymization of longitudinal psychological data*

https://www.linkedin.com/in/jiri-novak-8b748718/

# Data anonymization

My work is *Data Anonymization* in the context of the field of **Statistical Disclosure Control** (SDC)

**Statistical Disclosure Control** seeks to protect statistical data in such a way that they can be released without giving away confidential information that could be linked to specific individuals or entities.

Importance of Data Anonymization

- 1. **Principle**
  Statistical records of individual persons, businesses, or events used to produce Statistics are strictly confidential.
- 2. **Legal**
  Laws mandate protection of business and personal data, regulating publication of private information.
- 3. **Quality**
  Trust in confidentiality ensures respondents provide accurate information.
- 4. **Ethical**
  Disclosing information that can be linked to specific individuals or entities is unethical.

# Relevance to Open Science

**Open Science, Open Access, Open Data** are important trends in the scientific community.

Research data that results from publicly funded research should be FAIR:

**Findable,**
**Accessible,**
**Interoperable,**
**Reusable**

— therefore replicable, transparent, trustworthy, verifiable and accountable

— Principle: **As open as possible, as closed as necessary**

Commission Recommendation (EU) 2018/790 on access to and preservation of scientific information

# Key Concepts

## Disclosure

- Disclosure occurs when new information is revealed via released data.
- (1) **Identity disclosure**: Revealing the identity of an individual.
- (2) **Attribute disclosure**: Revealing sensitive attributes of an individual.
- (3) **Inferential disclosure**: Making inferences about an individual based on the released data.
- (4) **Membership disclosure**: Determining whether a specific individual is included in a dataset

## Re-identification risk

- Risk that an intruder can link a record in the released data to a specific individual in the population.

## Data utility

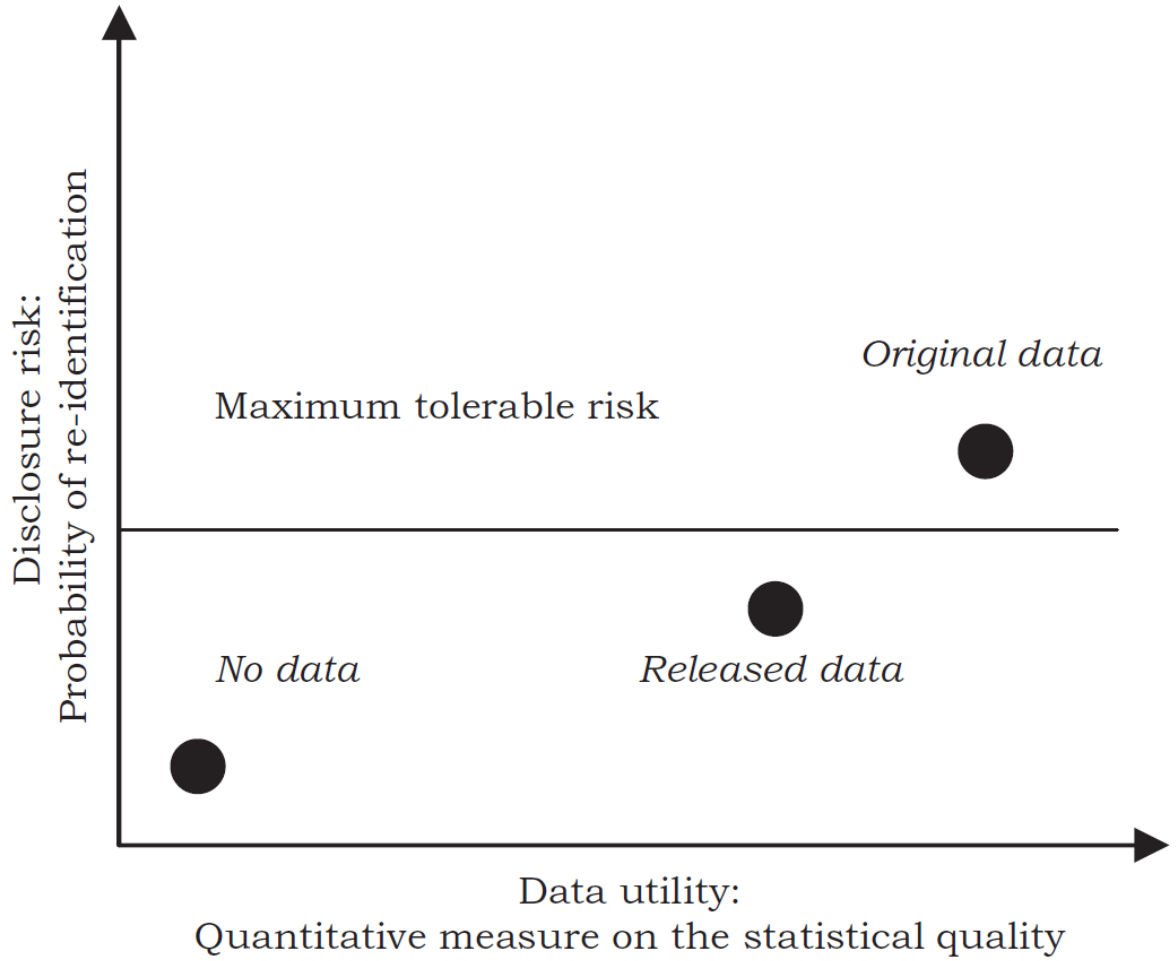
- Usefulness of the data for the intended purpose.

# Risk vs. Utility

The goal is to find a balance between risk and utility, so there is a **risk-utility trade-off**.

— **Risk**: the probability of a disclosure event occurring.


— **Utility**: the usefulness of the data for the intended purpose.




R-U confidentiality map (Duncan et al.,2001)

# Disclosure control methods

## 1. Masking original data

### a) Non-perturbative masking
Generalize data to hide identities without changing actual values.

### b) Perturbative masking
Add noise or alter data values.

## 2. Generating synthetic data

### a) Parametric methods
Statistical models based on the data‘s assumed distribution

### b) Non-parametric methods
Techniques that do not assume an underlying distribution

### c) Generative Adversarial Networks (GANs)
Machine learning models that generate synthetic data by training two neural networks in tandem.

# Mostly AI

https://mostly.ai/

MOSTLY AI's synthesizer uses a proprietary AI-driven approach that draws on deep learning principles similar to those found in generative adversarial networks (GANs).

You can use point-and-click environment, or python code.

# Libraries used

**a) os and sys**

- Set the working directory to the folder

OS Module

**a) Matplotlib**

- For visualization



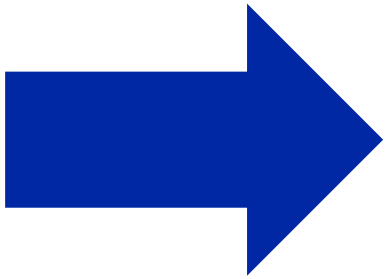
**b) Pandas**

- For data manipulation


**c) MostlyAI**

- Synthetize data with Mostly AI

# Challenges and error



**Open word document Problems.docx**

# Thank you for the attention

Swiss Data Anonymization Competence Center
https://swissanon.ch

**Contact**
Jiří Novák
PhD student
jiri.novak@uzh.ch