Anonymization of longitudinal demographic data

Jiří Novák

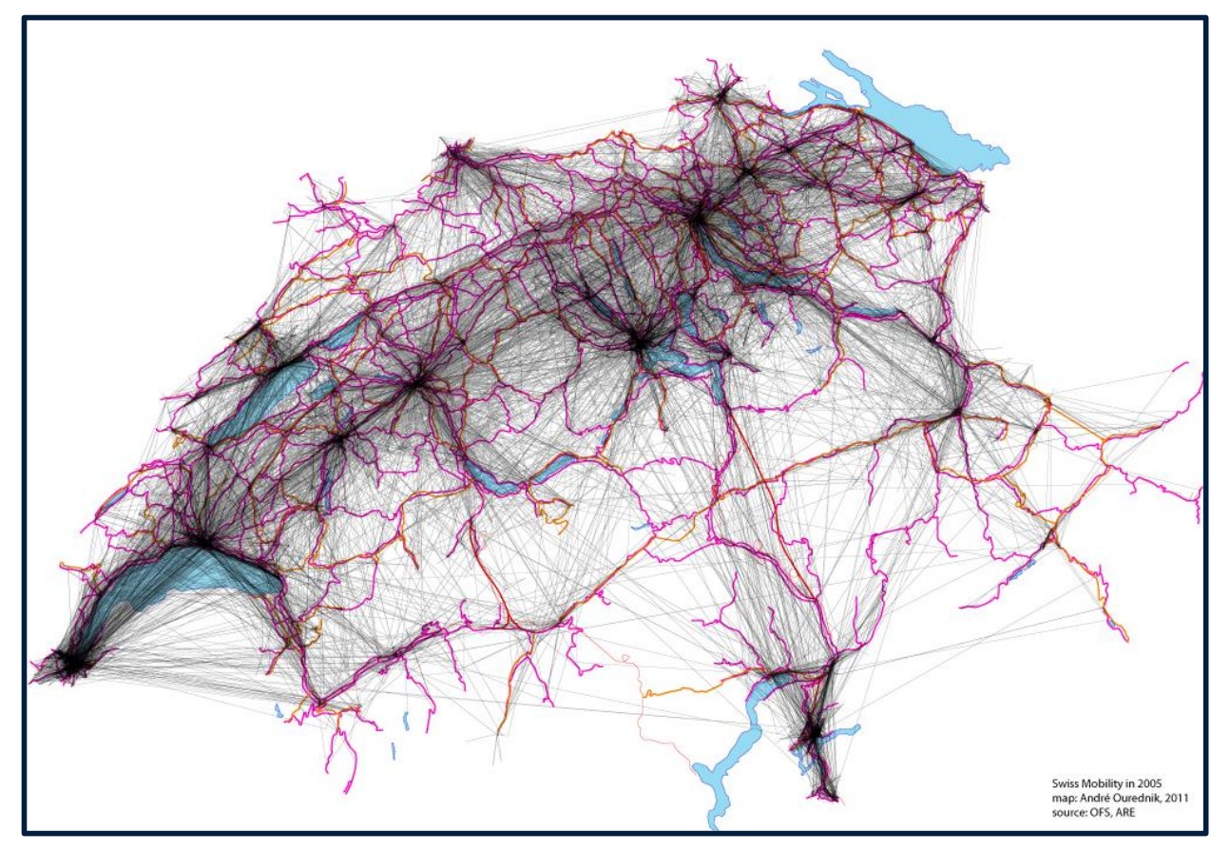
- *University of Zürich*
- *University of Applied Sciences Northwestern Switzerland*
- *Swiss Data Anonymization Competence Center*

BACKGROUND: Many longitudinal datasets contain **demographic variables that require proper protection against disclosure.** These datasets are also invaluable sources of information for researchers in fields such as demographics, medicine, psychology, transportation, social science, economics, and many more.

To enable dissemination, we can use these methods:

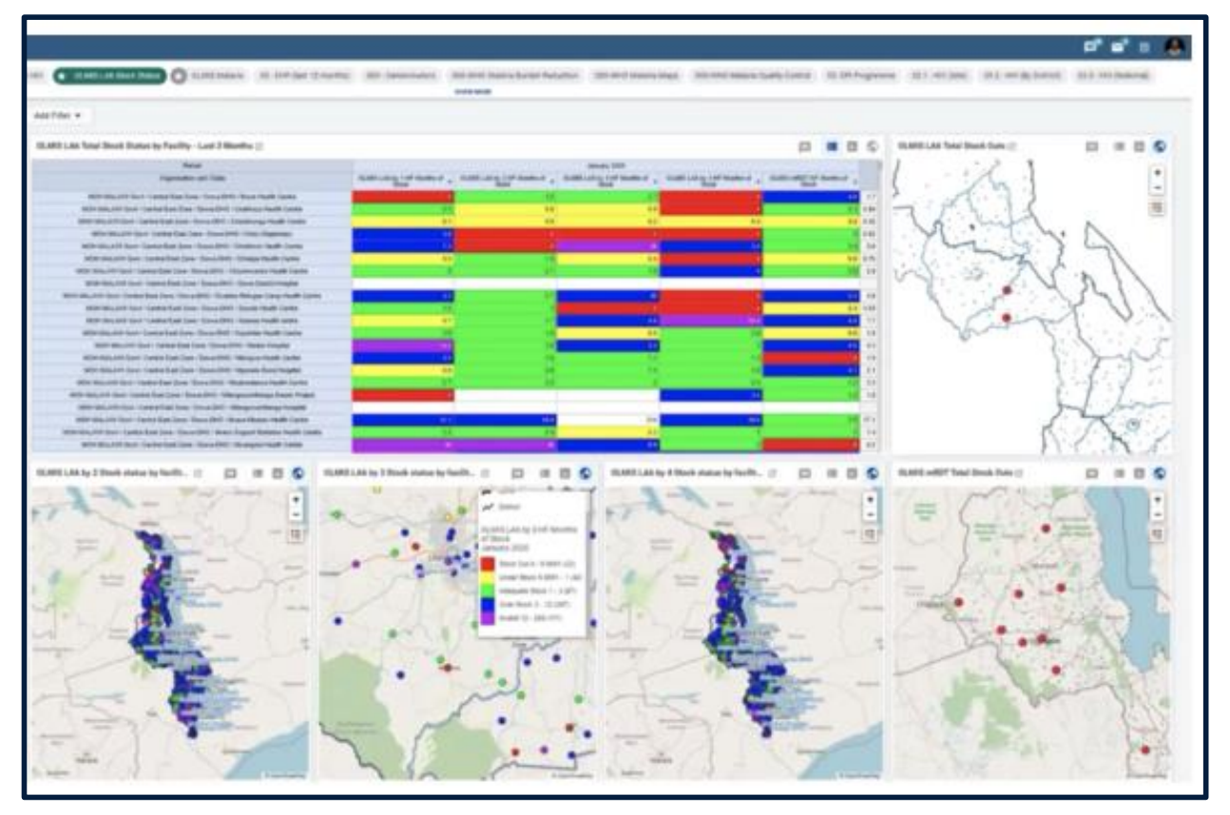
- **Statistical Disclosure Control (SDC)**
  - protects the data, prevents re-identification
- **Synthetic data generation**
  - mimics the original data
  - creates artificial data that can be safely disseminated

Mobility Tracking Example



Mobility in Switzerland: Microcensus on transport behavior 2005

Public Health Example



Open health management system in Malawi

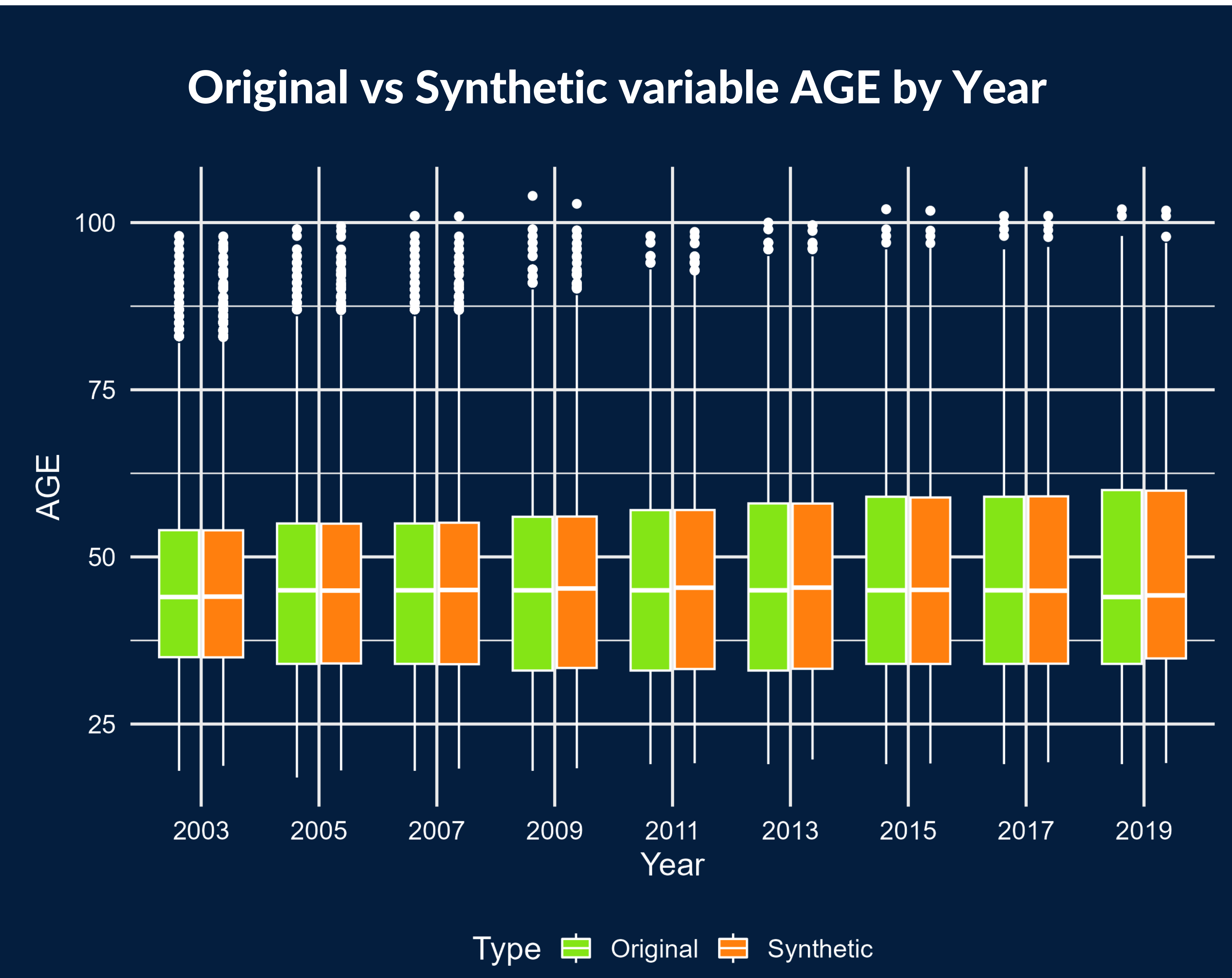
Open Science, Open Access, Open Data

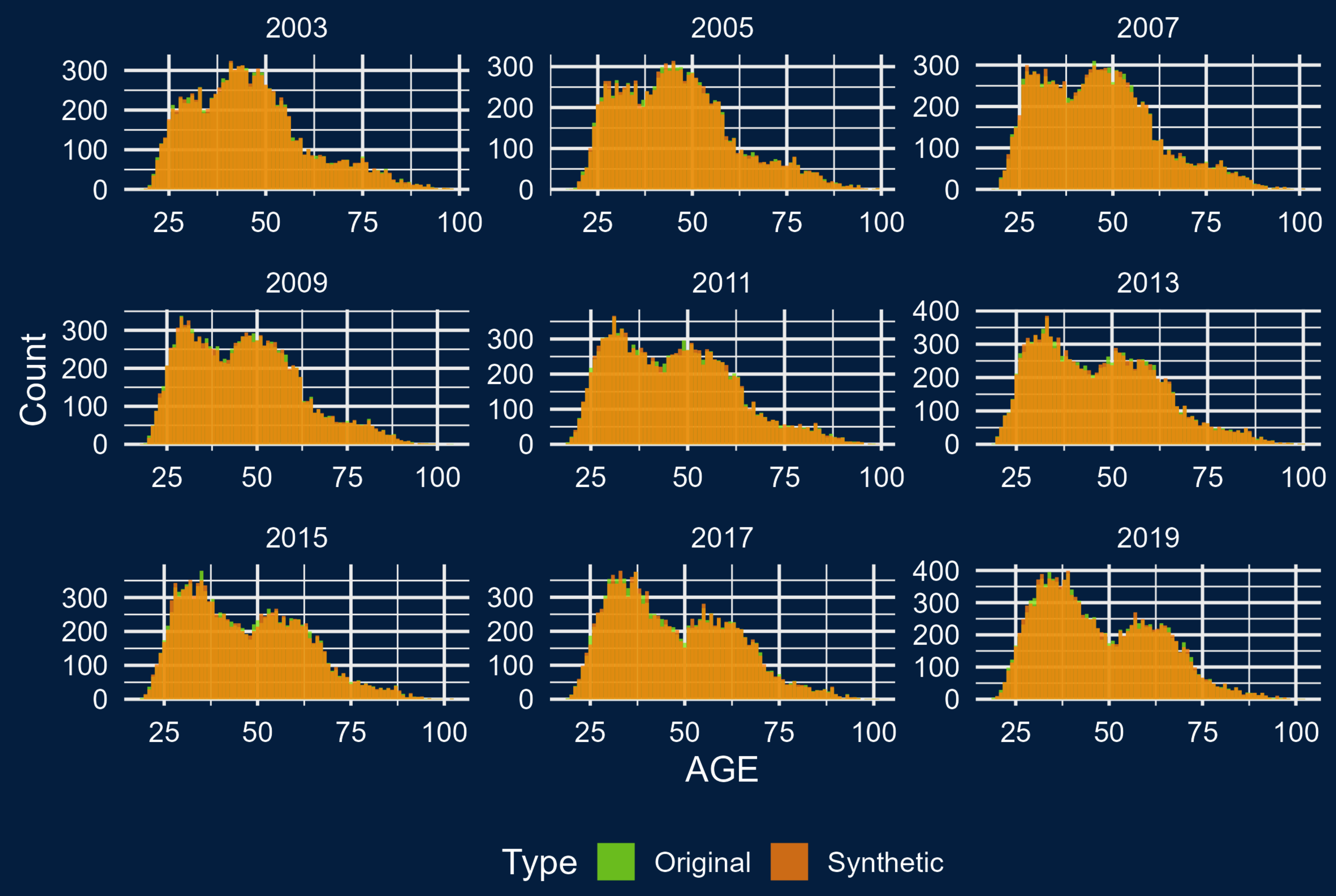
Research data that results from publicly funded research should be

- **findable**, **accessible**, **interoperable**, **reusable** ('FAIR principles')
- therefore **replicable**, **transparent**, **trustworthy**
- as open as possible, as closed as necessary

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

Illustrative example — PSID Data





Take a picture to download

Feel free to reach out and discuss!

METHODOLOGY

A key concern with the disclosure of personal data is whether an attacker can gain any new information about an individual.

- **SDC** is traditional approach to protecting outputs
  - **Non-perturbation methods** (reduce provided information)
    - Local suppression (delete high-risk records)
    - Global recoding (create broader categories)
  - **Perturbation methods** (modify data)
    - Noise masking
    - Record swapping
    - Microaggregaation

Traditional SDC methods alone are insufficient to protect longitudinal data. It is necessary to also use a more modern approach.

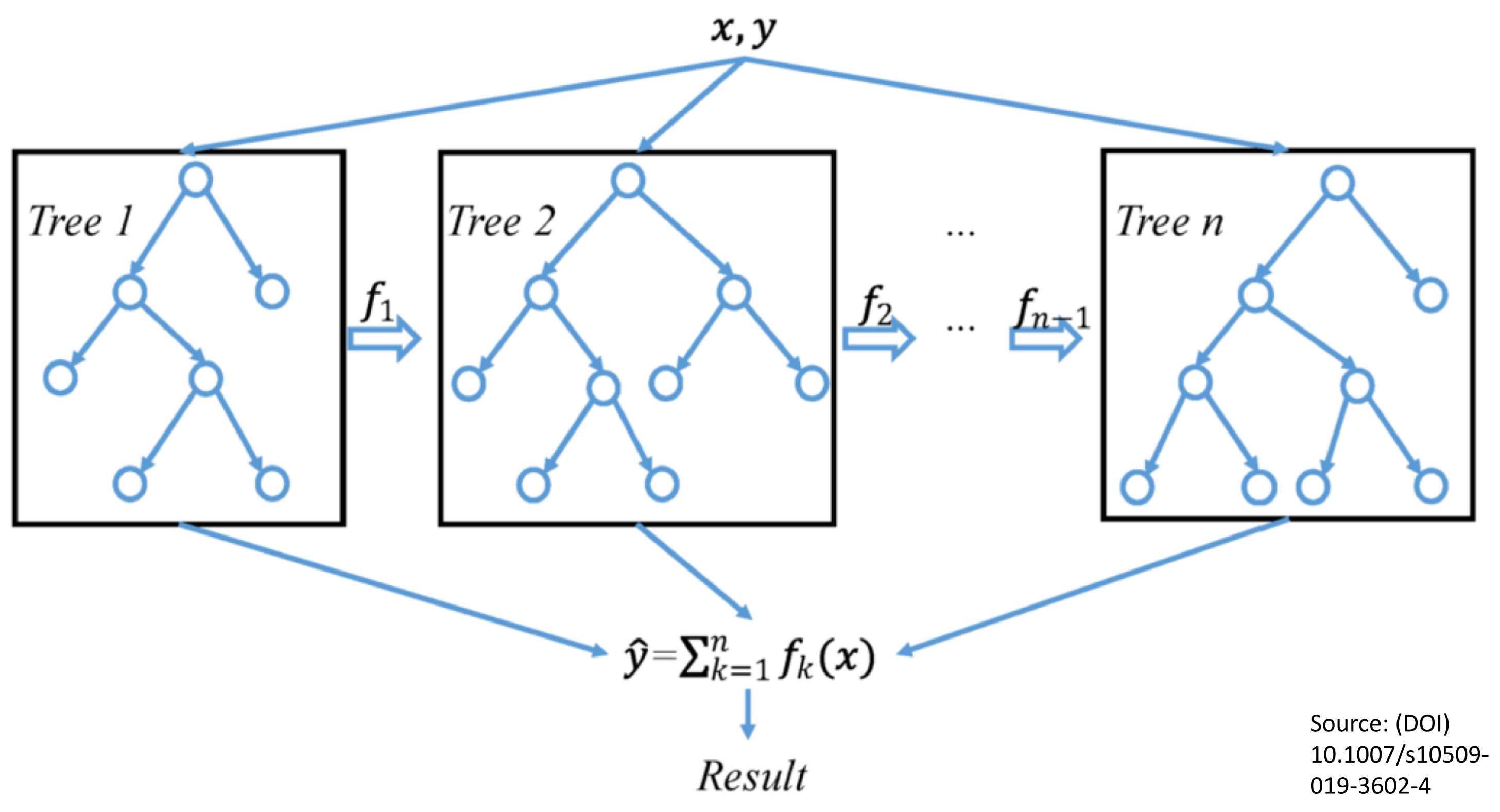
- **Synthetic data** — *replace some or all of the observed values by sampling from appropriate probability distributions so that the essential statistical features of the original data are preserved.*

Challenges of anonymizing longitudinal data

- Data Granularity
- Temporal Uniqueness
- Dynamic Features
- Consistency in Anonymization
- Loss of Data Utility
- Re-identification Risk
- Updating Anonymized Data

In illustrative example, the synthesizer utilized the XGBoost algorithm, which was adapted for longitudinal data.

- **XGBoost** is a distributed, optimized gradient boosting system using an iterative decision tree algorithm, with each tree learning from the residuals of previous trees.



Source: (DOI) 10.1007/s10509-019-3602-4

Acknowledgments

This work was funded by the Swiss National Science Foundation with grant number 211751: "*Harnessing event and longitudinal data in industry and health sector through privacy preserving technologies*".