Synthetic data for YOUth

Pilot project update

TDCC-SSH
Social Sciences & Humanities

NWO

Erik-Jan van Kesteren
Assistant professor
Utrecht University / ODISSEI

Programme today

- Context and goals of the pilot
- Privacy-friendly synthetic data: primer & metasyn
- Current status of synthetic data for YOUth

- Targeted discussion: what's next to use the synthetic data?

tdcc-synthetic-data.nl

Work packages

Package	Title	Description
WP1	Synthetic data generation tools	Creating open source software (metasyn and DP-CGANS) to produce privacy-friendly and realistic synthetic data.
WP2	Integrating tools in data repositories	Creating a plug-in to allow synthetic data generation in DANS Data Station SSH in the ingest pipeline and/or the user interface.
WP3	Synthetic data use-cases in SSH	Creating pilot implementations of synthetic data at various partner institutions.
WP4	Legal and privacy constraints	Creating a whitepaper or publication on how to overcome implementation issues with synthetic versions of sensitive data.
WP5	Outreach and project management	Organising events, creating teaching materials, and ensuring our project runs smoothly.

The full proposal of this project is openly available: doi:10.5281/zenodo.15697035

Synthetic data pilot with YOUth

- Goal: improve accessibility of YOUth data for prospective researchers

- Stretch goal: code-to-data
  Don’t give access to real data to researchers
  Researchers send in their code
  YOUth runs it and researchers receive only results

Request YOUth data

YOUth encourages and facilitates extensive and appropriate use of its data by bona fide research organisations and bona fide researchers. Please note that requests for biological material are temporarily unavailable.

Please read the following documents before submitting a data request:

- YOUth Data Access Protocol (PDF)
- Interactive YOUth prospectus or prospectus as Excel document (an overview of all available data)
- Preview of the online data request form
- Instructions preregistration YOUth Registry

When you are ready to submit a data request, you can access the data request module. For additional questions, you can contact us at youthonderzoek@uu.nl.

Researchers who wish to use YOUth data in a project that has not yet been funded and intend to mention the use of YOUth data in a grant proposal first need to contact us at youthonderzoek@uu.nl. The proposed project will be evaluated by the YOUth Management Team. If approved, the YOUth Management Team will sign a Statement of Intent to share YOUth data.

Below you can find the YOUth template for the Statement of Intent in English and in Dutch:

- Statement of Intent (English)
- Intentieverklaring (NL)

Request YOUth data

YOUth encourages and facilitates extensive and appropriate use of its data by bona fide research organisations and bona fide researchers. Please note that requests for biological material are temporarily unavailable.

Just want to try out our data?

Go to our synthetic data request. Usually, a request for the synthetic data is granted within 24 hours.

Please read the following documents before submitting a data request:

- YOUth Data Access Protocol (PDF)
- Interactive YOUth prospectus or prospectus as Excel document (an overview of all available data)
- Preview of the online data request form
- Instructions preregistration YOUth Registry

When you are ready to submit a data request, you can access the data request module. For additional questions, you can contact us at youthonderzoek@uu.nl.

Researchers who wish to use YOUth data in a project that has not yet been funded and intend to mention the use of YOUth data in a grant proposal first need to contact us at youthonderzoek@uu.nl. The proposed project will be evaluated by the YOUth Management Team. If approved, the YOUth Management Team will sign a Statement of Intent to share YOUth data.

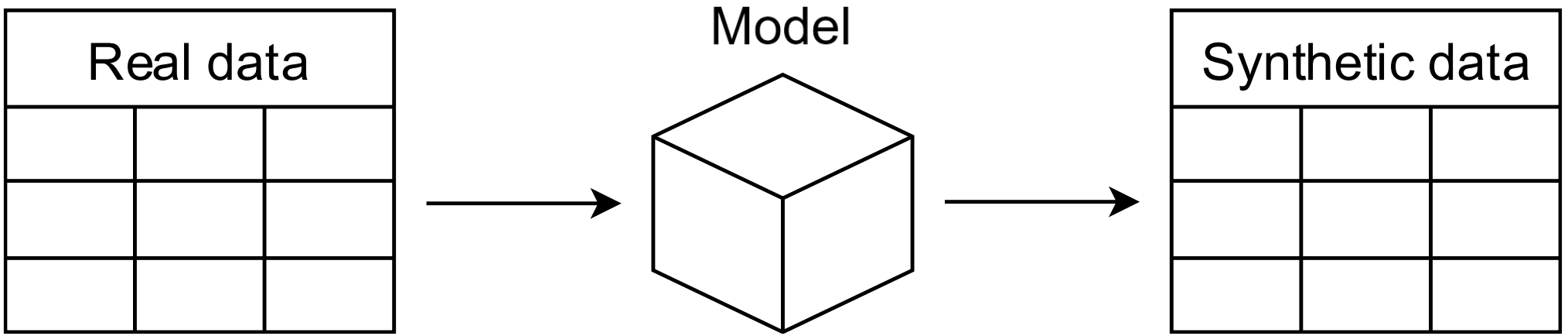
Primer on synthetic data

Our view of synthetic data

Synthetic data is generated from a model
As opposed to real, natural, collected data

fake data
generated data
simulated data
digital twin
public use file

Generating synthetic data

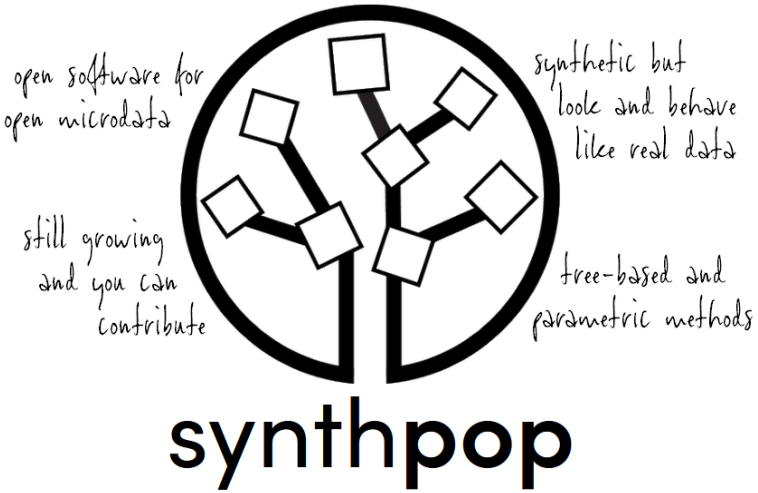


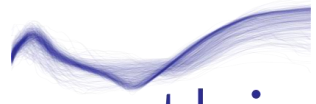
Generating synthetic data

- To create synthetic data, you need a **generative model**
- You can **learn** (fit, estimate) its parameters using real data
- Then you can **generate** (predict, infer, synthesize) new data from this fitted model

- Examples:
  - A normal distribution with parameters μ, σ
  - A histogram with bins and proportions
  - A generative adversarial network with a million parameters

Synthetic data software





Gretel Synthetics

https://utrechtuniversity.github.io/privacy-engineering-tools/synthetic-data/

Privacy – fidelity tradeoff

- Every parameter in the data-generating model contains **information** about the observations in the real data
- The more complex your model (e.g., many parameters), the closer the synthetic data will be to the true data: "**fidelity**"

Crucially, per statistical fact:

- When the model has as many parameters as data points, we can exactly recreate the real data
- At that point, there is no more privacy / disclosure control

What can we do with the synthetic data?

Investigate & answer all your research questions

Find out how much your colleagues earn

Basic correlation analysis

- Getting to know the data
- Use the data as a toy example
- Develop & validate data analysis scripts and pipelines
- ...

Anything you can do with real data

Nothing

Fidelity

Privacy

Estimate parameters with low simulation error

Visualisation of association

Visualisation of variation

What can go wrong?

Synth. data may "leak" information

Identity disclosure

- We might reproduce someone's data exactly in the synthetic data
- We might indirectly reveal that someone is in the dataset (related to *differential privacy*)

Example

A hospital records income for billing, now a billionaire enters the hospital. The model generates much higher incomes.

Synth. data may "leak" information

Attribute disclosure

- We might reveal information about someone that was not public

Example

Hospital data shows that every woman in the hospital between 56-60 has cancer. If I know the 57-year-old neighbour was in the hospital, I now know that she has cancer.

Synth. data may "leak" information

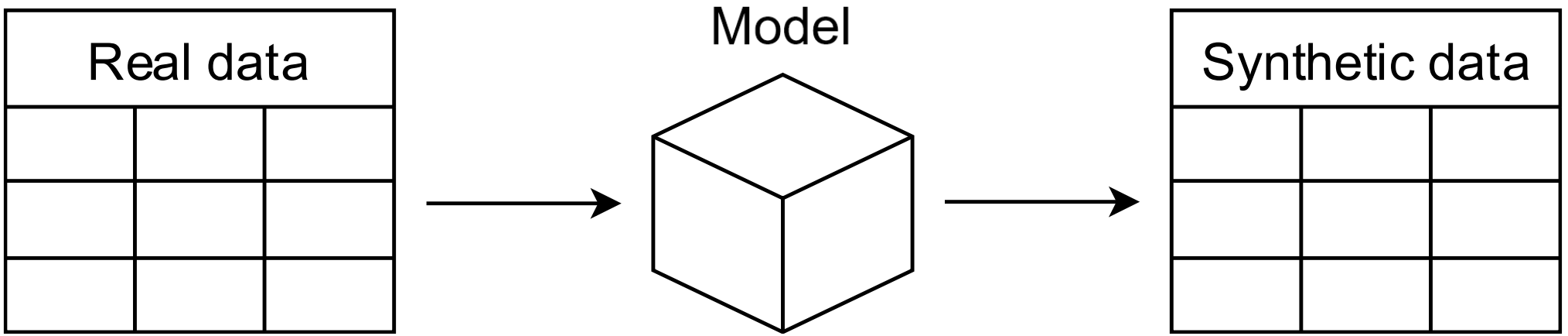
Inferential disclosure

- Attributes in the data may be used to improve knowledge about other sensitive attributes not in the data

Example

If I know your age, home value, industry, number of hours worked, then I may know your net income very well.

This needs *disclosure control*



The {meta}syn solution

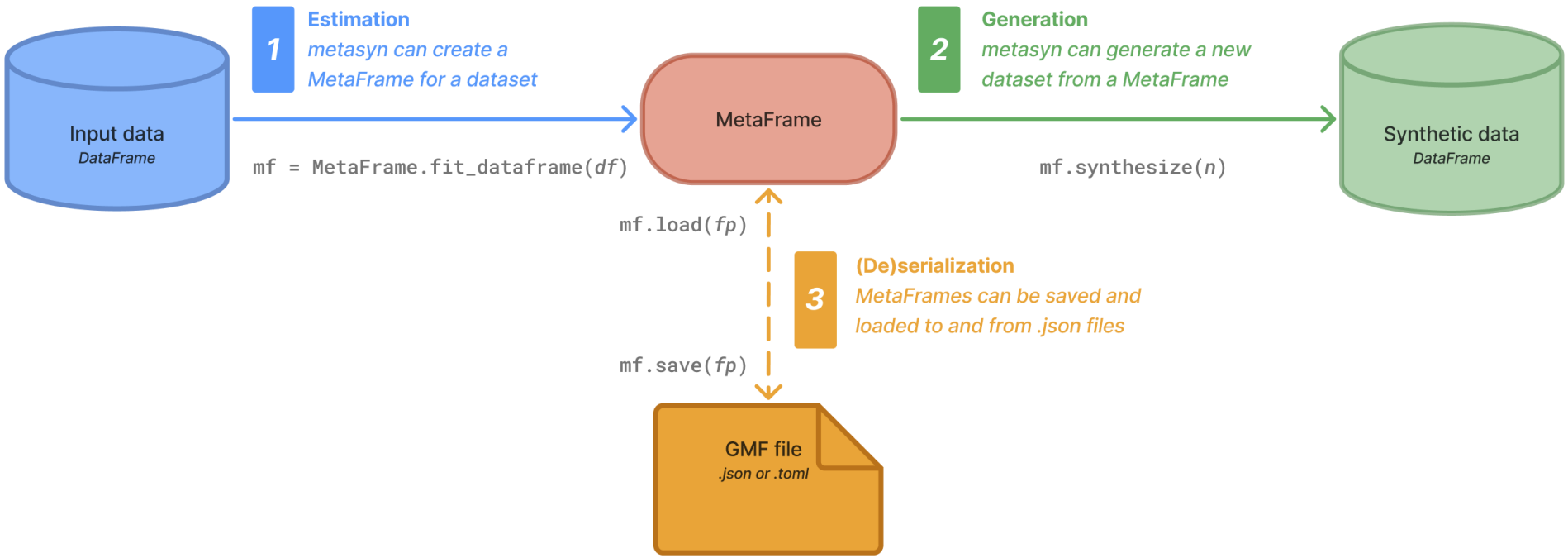
Step 1: make the model simple and transparent

- No multivariate information (!) just a simple model for each variable
- You (yes you!) can inspect the model through metadata

Step 2: allow for automatic disclosure control rules

- Today: Eurostat / CBS disclosure control rules
- Cool fancy plugin system: every org. their own rules!

{meta}syn python package



Synthetic data for YOUth

github.com/sodascience/synthetic_youth_pilot

synthetic_youth_pilot Public

Edit Pins | Unwatch 2 | Fork 1 | Star 0

main | 1 Branch | 0 Tags | Go to file t | Add file | Code

vankesteren Start presentation YOUth 26ead9c · last week 7 Commits

docs	Start presentation YOUth	last week
output	initial commit	6 months ago
raw_data	initial commit	6 months ago
src	change file structure for more organization	4 months ago
.gitignore	initial commit	6 months ago
.python-version	initial commit	6 months ago
LICENSE	Create LICENSE	5 months ago
README.md	update filenames	6 months ago
pyproject.toml	update code for metasyn v2	last week
synthesize.py	update code for metasyn v2	last week
test_analysis.py	change file structure for more organization	4 months ago
uv.lock	update code for metasyn v2	last week

README | MIT license

YOUth pilot privacy-friendly synthetic data

Python uv

This repository implements a pilot for creating privacy-friendly questionnaire datasets from the YOUth cohort. It is built on metasyn with the disclosure control plugin.

About

Synthetic data pilot for YOUth study questionnaires, using metasyn

synthetic-data questionnaire-survey youth-data

Readme
MIT license
Activity
Custom properties
0 stars
2 watching
1 fork
Report repository

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

Python 100.0%

Suggested workflows

Based on your tech stack

Discussion

Discussion questions

- (When) would you be OK with openly available synthetic data as shown here?
- Can the metadata be made openly available? (e.g., multiple choice answer labels)
- How would a code-to-data pipeline look?
- Which steps do we need to take to move to synthetic YOUth data availability?

https://tdcc-synthetic-data.nl