

Synthetische data

Een oplossing voor datatoegangsproblemen



Universiteit Utrecht



Statistics
Netherlands



ODISSEI

Thom Volker^{1,2} & Peter Gerbrands¹

¹Utrecht University; ²Statistics Netherlands

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

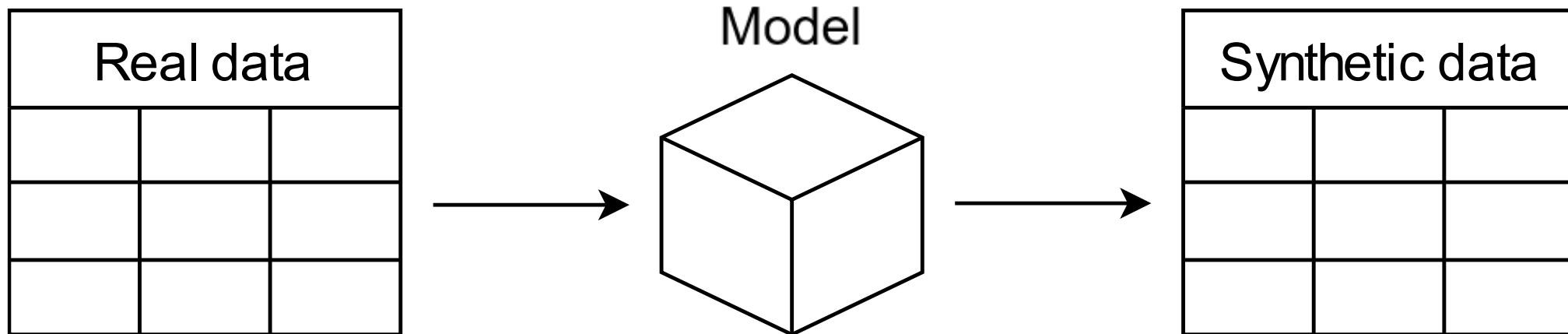
simulated data

digital twin

public use file

Synthetic (test) data generation

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

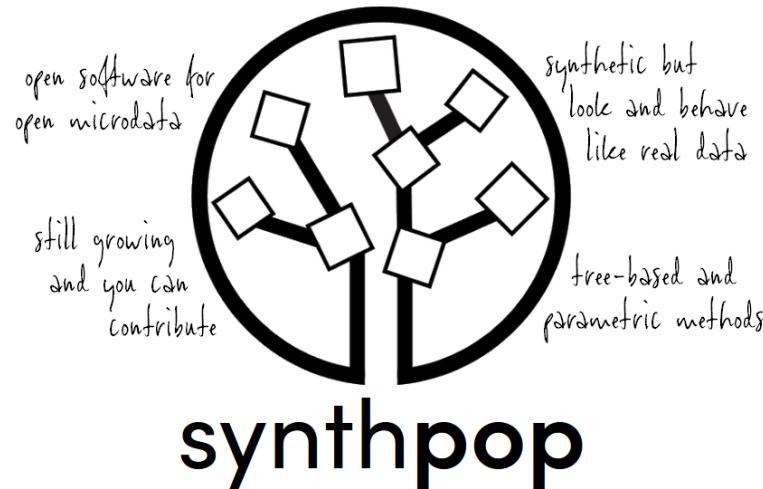
Model-based test data

- Train models on real world data that comply with observed statistical distributions
- Generative Adversarial Network (GAN) or Variational Autoencoder (VAE)
- Good for ML training and privacy preserving analysis
- Difficult for multiple tables with cross-table dependencies

Rule-based test data

- Create data by using logic and predefined rules.
- High control and customization to business/domain rules.
- Random sampling from predefined distributions, then shuffling and augmentation.
- Good for functional- and regression testing.
- Limited usability for fuzz testing

Synthetic data software



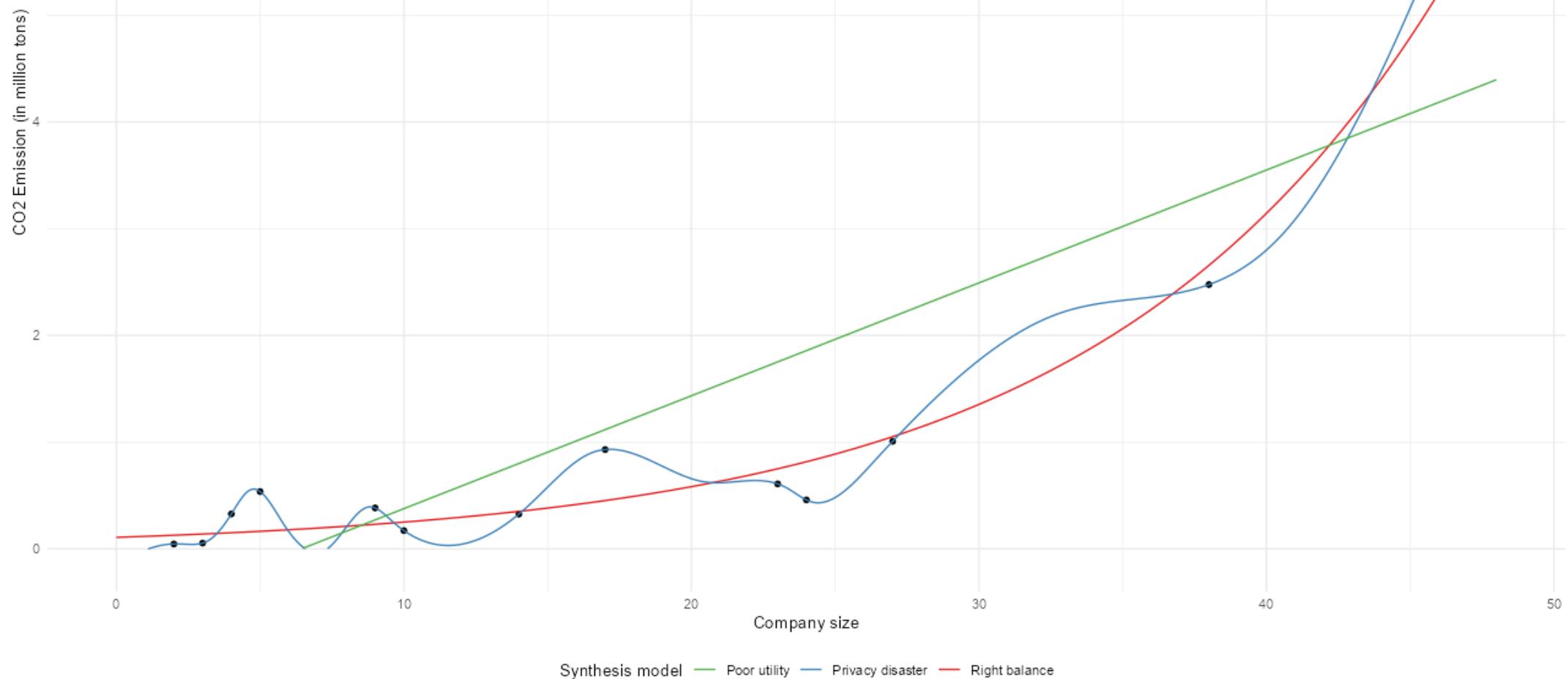
Gretel Synthetics



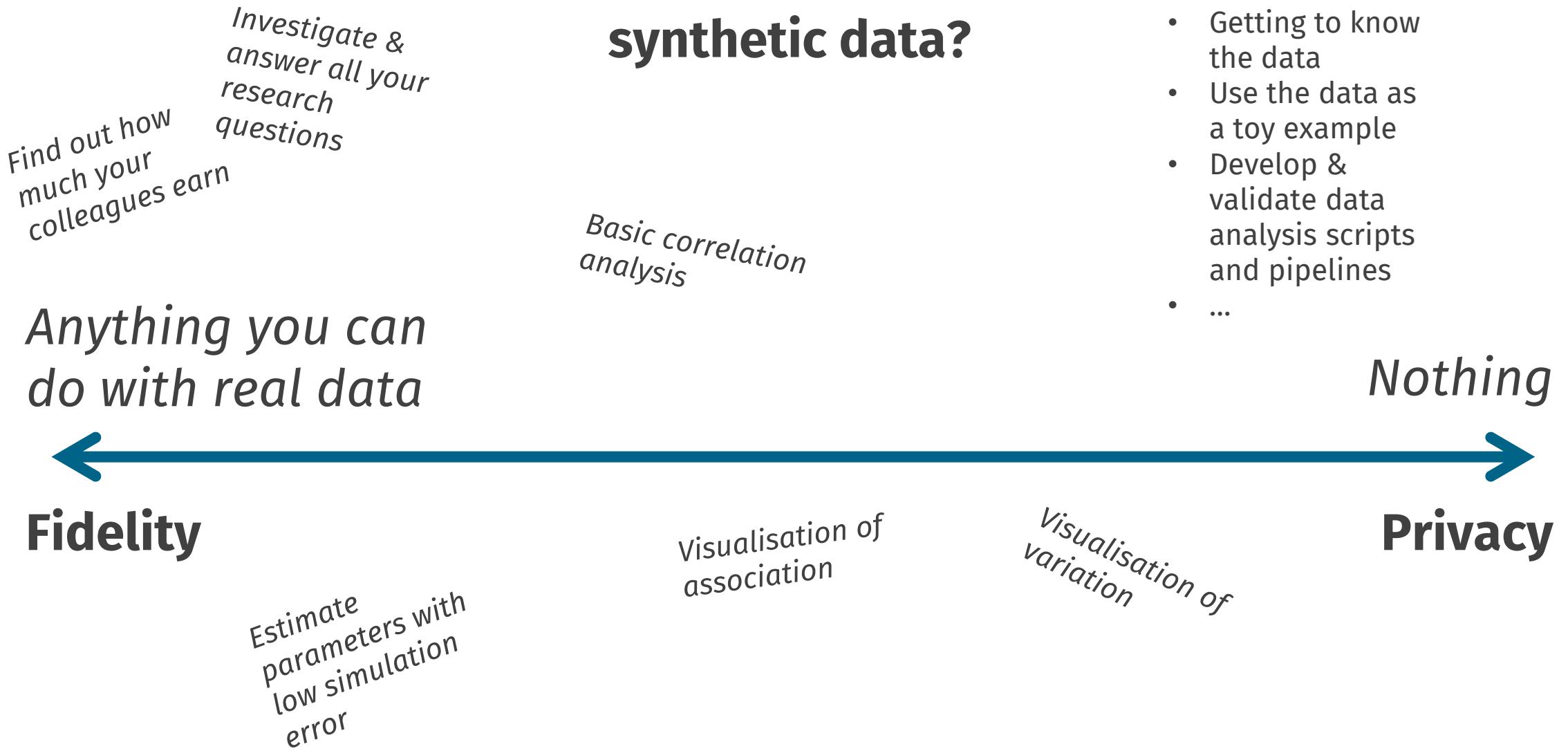
Privacy – fidelity tradeoff

- 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**”

Privacy – fidelity tradeoff



What can we do with the synthetic data?



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

You record emission data, and certain high emitters (“outliers”) are in the data. The synthetic data model may generate much higher emission data.

Synth. data may "leak" information

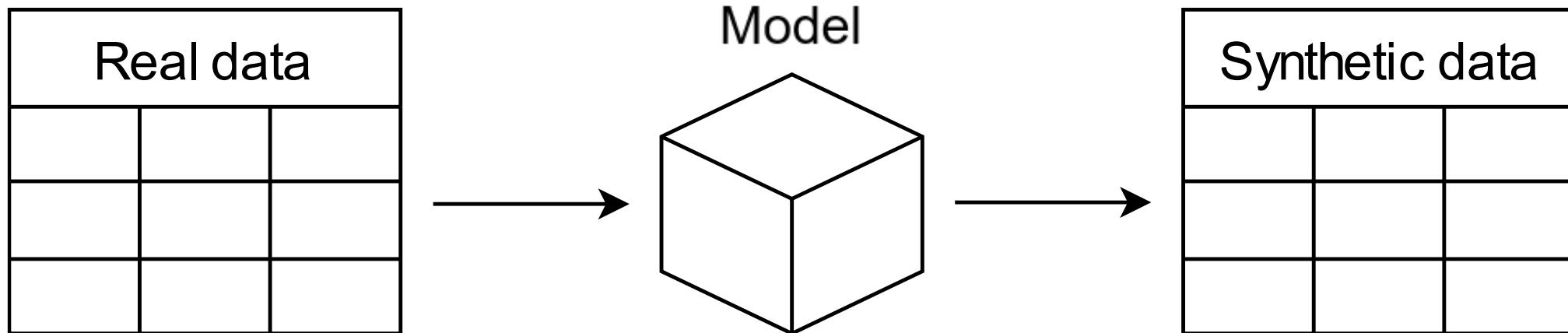
Attribute disclosure

- We might reveal information about someone that was not public

Example

Data shows that all companies in a certain region emit more than 1 million tons of CO₂. If I know that a certain company is in a certain region, I now know that it produced more than 1 million ton CO₂ emission.

This needs *disclosure control*



Synthetic data for YOUTH

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



YOUth Cohort Study

Home About YOUth Request YOUth data YOUth data collection YOUth publications YOUth open science News

About YOUth

Request YOUth data

- › The data request system
- › Approved data requests
- › Poster Template

YOUth data collection

YOUth publications

YOUth open science

News

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 \(NL\)](#)



YOUth Cohort Study

[About YOUth](#) [**Request YOUth data**](#) [YOUth data collection](#) [YOUth publications](#) [YOUth open science](#) [News](#)

About YOUth

Request YOUth data

- Synthetic test data
- The data request system
- Approved data requests
- Poster Template

YOUth data collection

YOUth publications

YOUth open science

News

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.

The {meta}syn solution

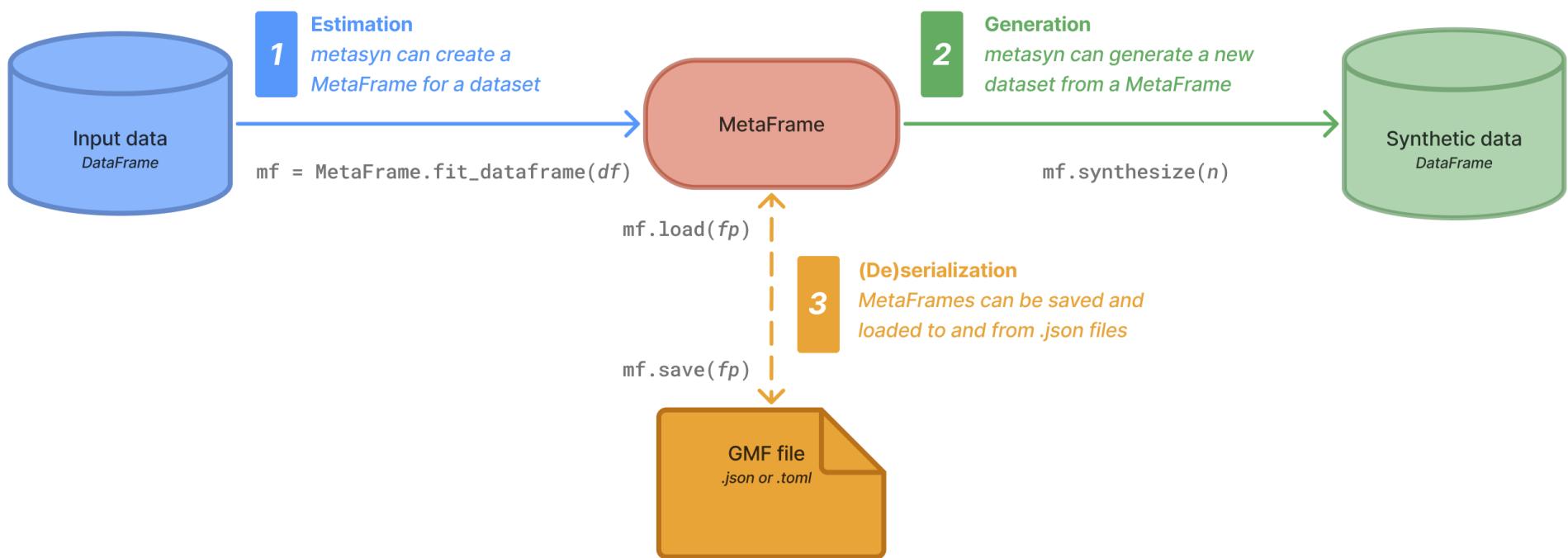
Step 1: make the model simple and transparent

- 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



<https://github.com/sodascience/metasyn>

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 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



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

Python Package using

TDCC-SSH: Synthetic data

- Enabling the potential of synthetic data in the social and health sciences
- Funded by NWO
- <https://tdcc-synthetic-data.nl/>

<https://tdcc-synthetic-data.nl/>