DANS Open Day
Open data, open science

Synthetic Data for Open Science

Dr. Chang Sun, Assistant Professor
Department of Advanced Computing Sciences
Maastricht University

DANS Open Day
11 June 2026

DANS

The open-science paradox

The data we could learn the most from is often the data we cannot easily share.

Sensitive datasets such as health records, social surveys, and registries are rich in information and contain deep insights, but difficult to access, reuse, and combine

A repository's dilemma

Sensitive data can be preserved safely, but not shared easily.

Access requests, agreements, and waiting times slow down reuse and limit new research.

What if we could share a faithful synthetic version instead?

What is synthetic data?

What is synthetic data?

Data created by a model that learned from the real data.

Synthetic data is **structurally and statistically similar to the real data.**

- at the **individual** sample level (e.g., synthetic data should not include prostate cancer in a female patient)
- at the **population** level (e.g., marginal and joint distributions of features)
- at the **machine learning/statistical analysis utility** level (i.e. the analysis results on synthetic data are close to the results on real data).
- offers strong **privacy guarantees** to prevent adversaries from extracting any sensitive information.

Why we need synthetic data?

Data replication

When data is **unavailable** or **costly to access**, synthetic data can be used to simulate the characteristics of the real data, to facilitate **open science** and **education**.

Data privacy

When real data contains sensitive information, and sharing it can raise **privacy concerns**, synthetic data can be used with preserving statistical characteristics of real data without exposing sensitive information

Data augmentation

When work with **small** or **imbalanced** datasets, synthetic data methods can augment or supplement existing data to **increase data size and diversity**.

Data creation

When the research needs data that is **rare or difficult to collect** in the real world, synthetic data can generate **rare situations, under-represented groups**, allowing model training in a broader range of scenarios/populations.

How do we generate synthetic data?

Different generative architectures

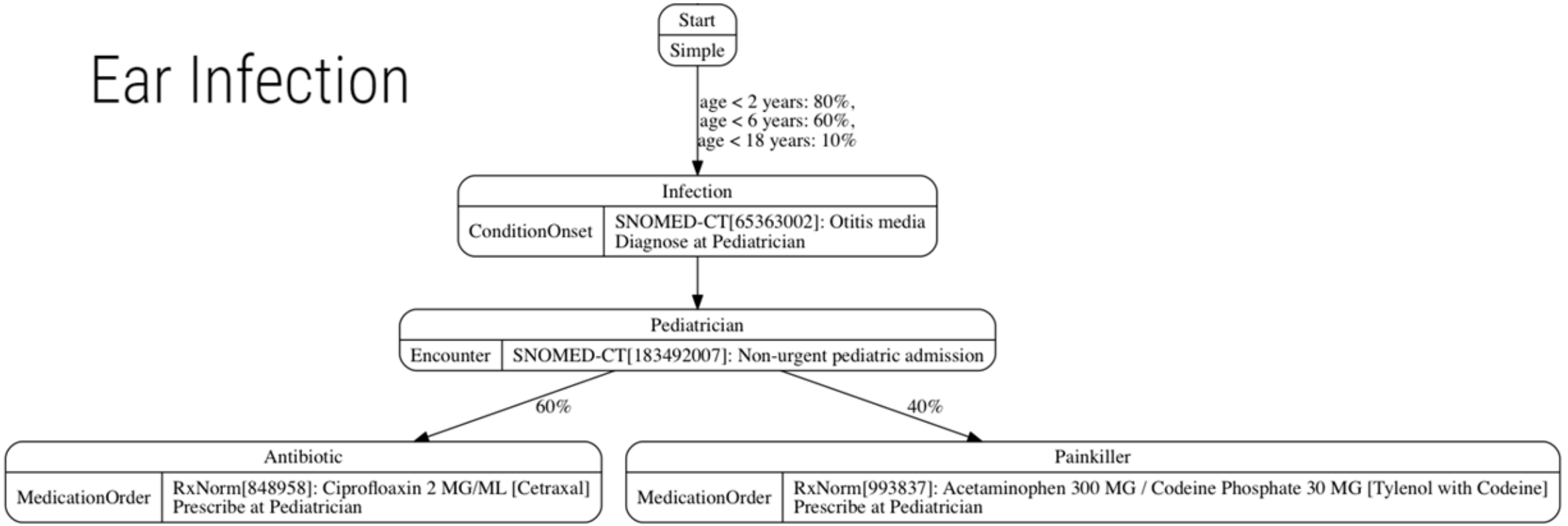
1. Domain knowledge based
2. Distribution based
3. Neural Networks (Deep learning)

Supervised learning

Unsupervised learning

Different generative architectures

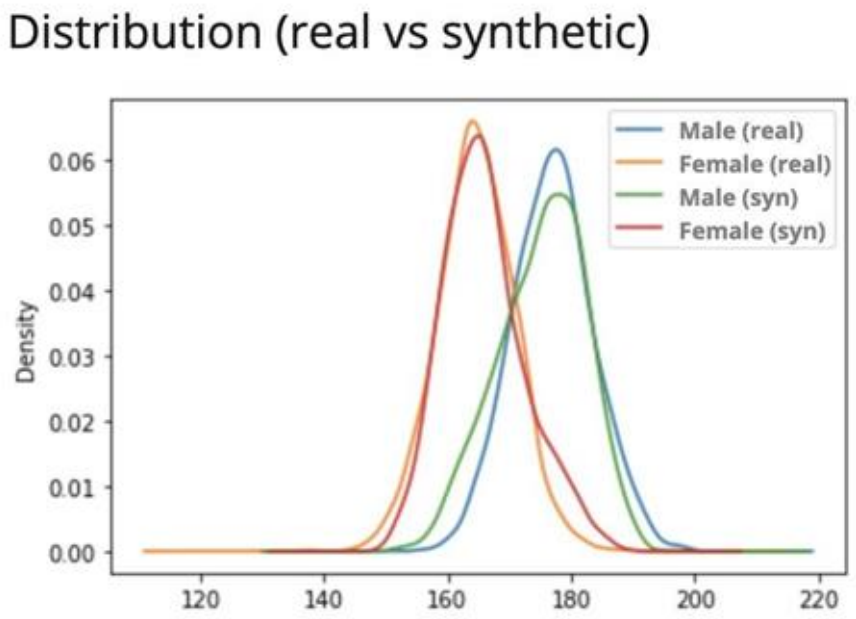
1. **Domain knowledge based**



[1] Walonoski J, Kramer M, Nichols J, Quina A, Moesel C, Hall D, Duffett C, Dube K, Gallagher T, McLachlan S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. J Am Med Inform Assoc. 2018;25(3):230–8.

Different generative architectures

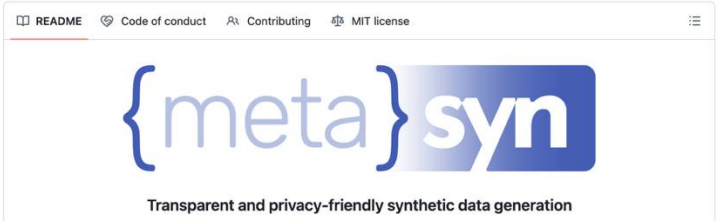
1. Domain knowledge based
2. **Distribution based**



g)

Independent Marginals: sampling from the empirical marginal distributions of each variable.

+ Computationally efficient
- Unable to capture statistical dependencies across variables



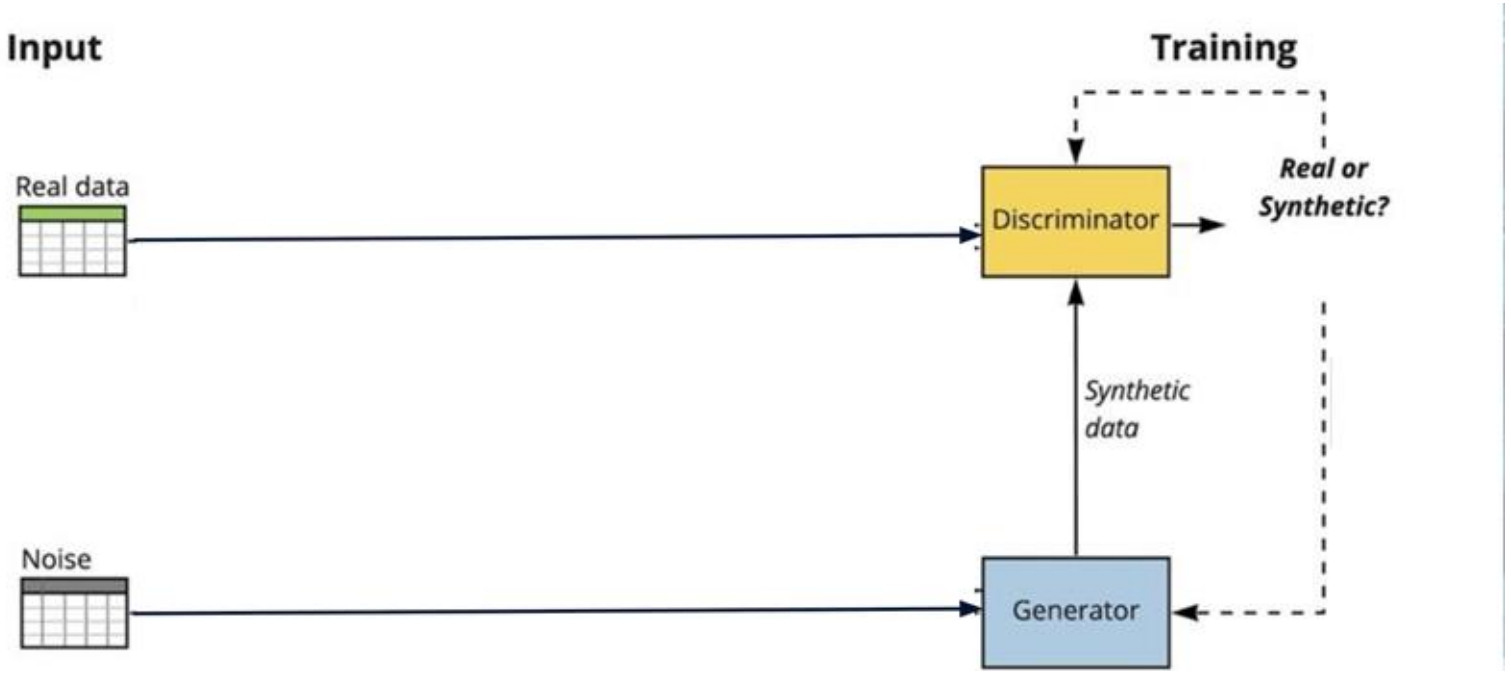
Related work of synthetic data generation

1. Domain knowledge based
2. Distribution based
3. **Neural Networks** (Generative adversarial network, Variational autoencoder, Diffusion model)

Generative Adversarial Network (GAN): trains and leverages two opposing neural network models (a generator and a discriminator) in a competitive manner.

Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

DP-CGANS: Differentially Private Conditional Generative Adversarial NetworkS

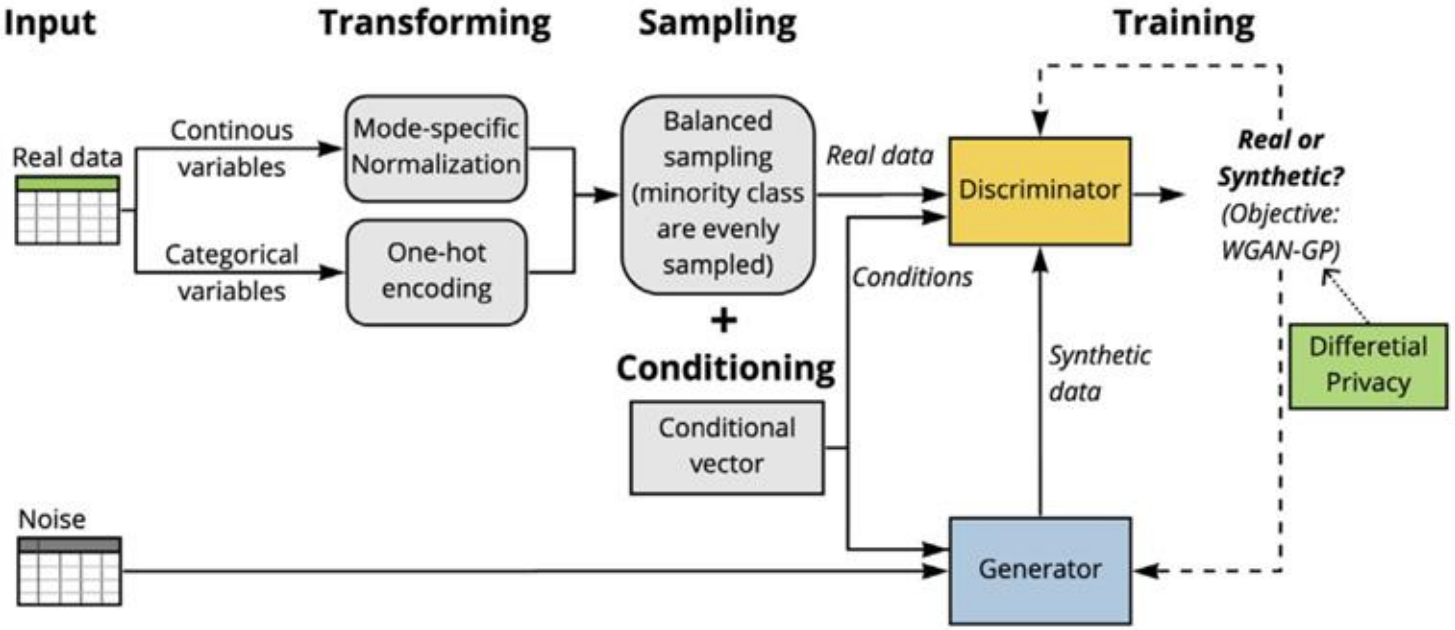


Sun, Chang, Johan van Soest, and Michel Dumontier. "Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy." Journal of Biomedical Informatics 143 (2023): 104404.

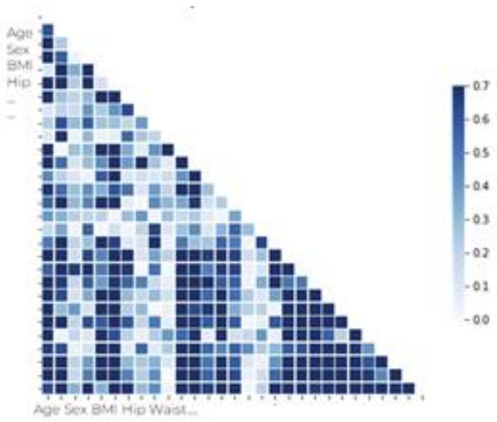
DP-CGANS: Differentially Private Conditional Generative Adversarial NetworkS

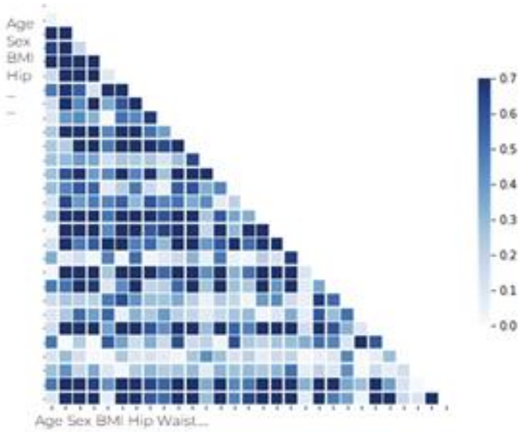
To tackles challenges:

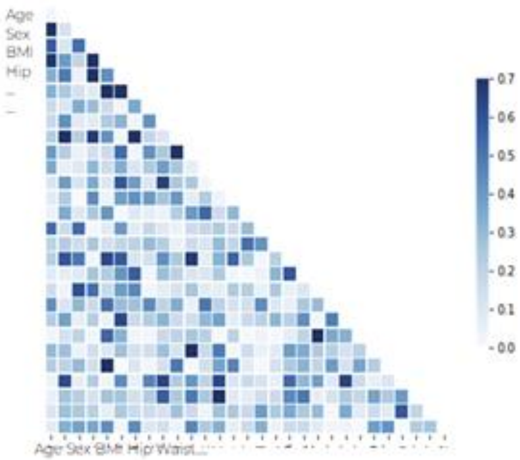
- Mixed data types
- Imbalanced variables
- Capturing the correlations and dependencies between variables
- Privacy guarantee

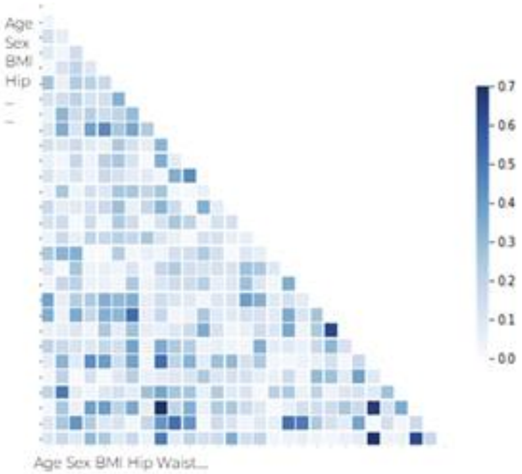


Sun, Chang, Johan van Soest, and Michel Dumontier. "Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy." Journal of Biomedical Informatics 143 (2023): 104404.









Employment	Occupation Class	Movement time	Mobility_lim	Depression medication	History of depression
Unemployed	Low occ class	11.887071463024785	No	yes	No
Employed	Not working	1.2580274245171872	Yes	no	Yes
Other	Not working	8.774445241442756	Yes	yes	No
Other	Low occ class	27.71435519708517	Yes	no	Yes
Employed	Low occ class	7.022962046269205	Yes	no	No
Unemployed	Low occ class	8.394498024617405	Yes	no	No
Employed	Intermediate occ class	22.118454250030194	Yes	yes	Yes
Other	Other	11.723942069202064	No	no	Yes
Other	Other	1.3802864356861302	Yes	no	No
Other	Not working	27.71435519708517	Yes	yes	Yes
Employed	Low occ class	6.636072967291062	No	yes	Yes

Employment	Occupation Class	Movement time	Mobility limitation	Depression medication	History of depression
Employed	High occ class	11.683110035839919	No	no	No
Unemployed	Not working	5.198166849876685	No	no	No
Employed	Self-employed	2.0694206532500825	Yes	no	No
Unemployed	Not working	3.8583851070704736	No	no	No
Unemployed	Not working	14.26477181650862	No	no	No
Unemployed	Not working	2.3634242654562385	Yes	yes	Yes
Employed	Intermediate occ class	15.346579554425292	No	no	No
Unemployed	Not working	4.760861567433586	No	no	No
Employed	Intermediate occ class	11.540673369706266	No	no	No
Unemployed	Not working	10.392846949564621	No	no	No
Employed	High occ class	4.254728929923106	No	no	No
Unemployed	Not working	2.5083105632833282	Yes	no	No

Training

Maastricht University

Assessing synthetic data quality

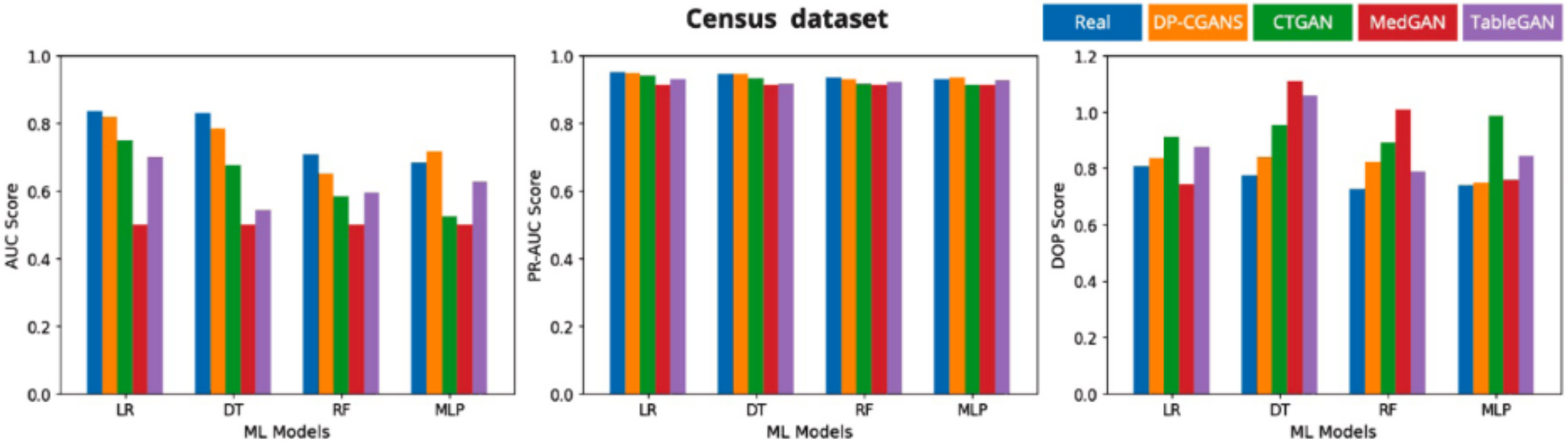
Synthetic data that are **structurally** and **statistically** similar to the real data, with a **privacy** guarantee**.**

- **Statistical Similarity**
  - Single variable (distribution)
  - Variable pairs (correlation)

- **Utility (machine learning performance)**
  - Logistic regression
  - Random Forest …

- **Privacy Cost**
  - Identity disclosure (e.g., Hamming distance …)
  - Attribute disclosure (e.g, KNN …)

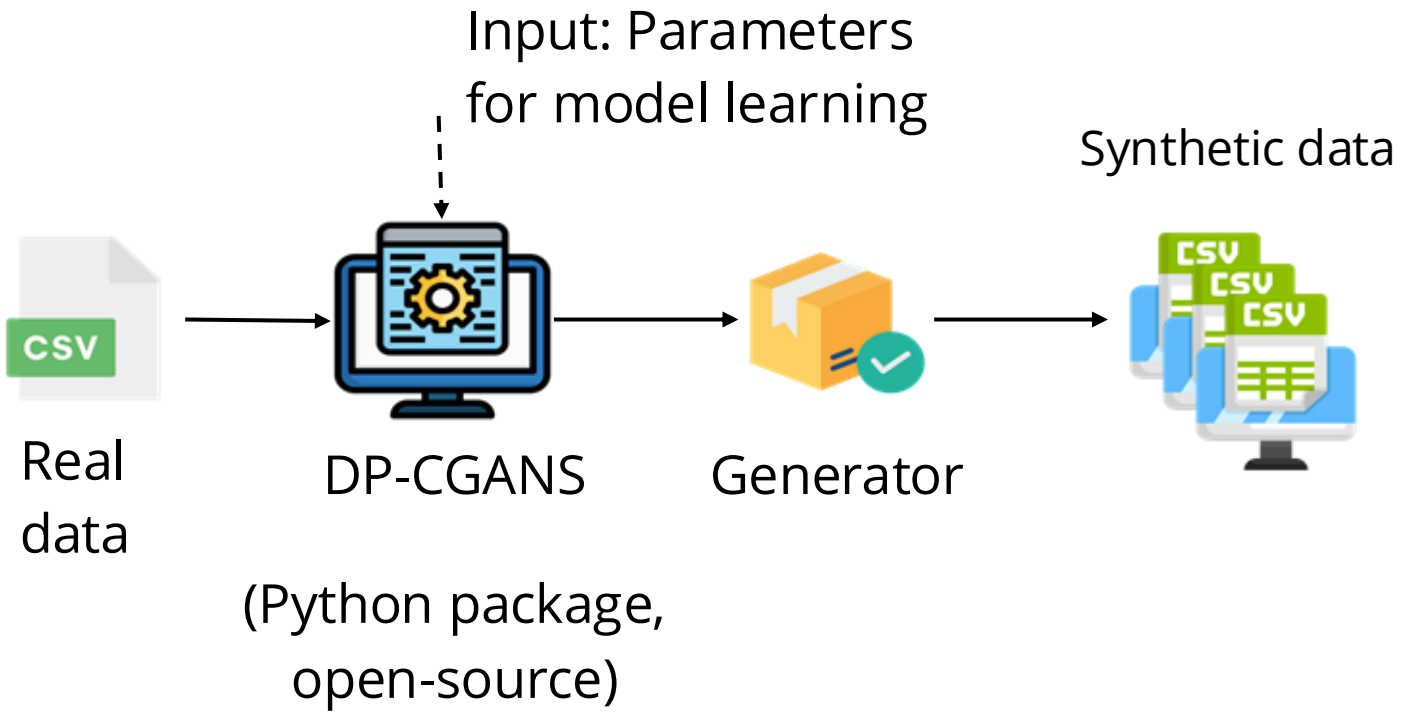
One result example (ML performance)



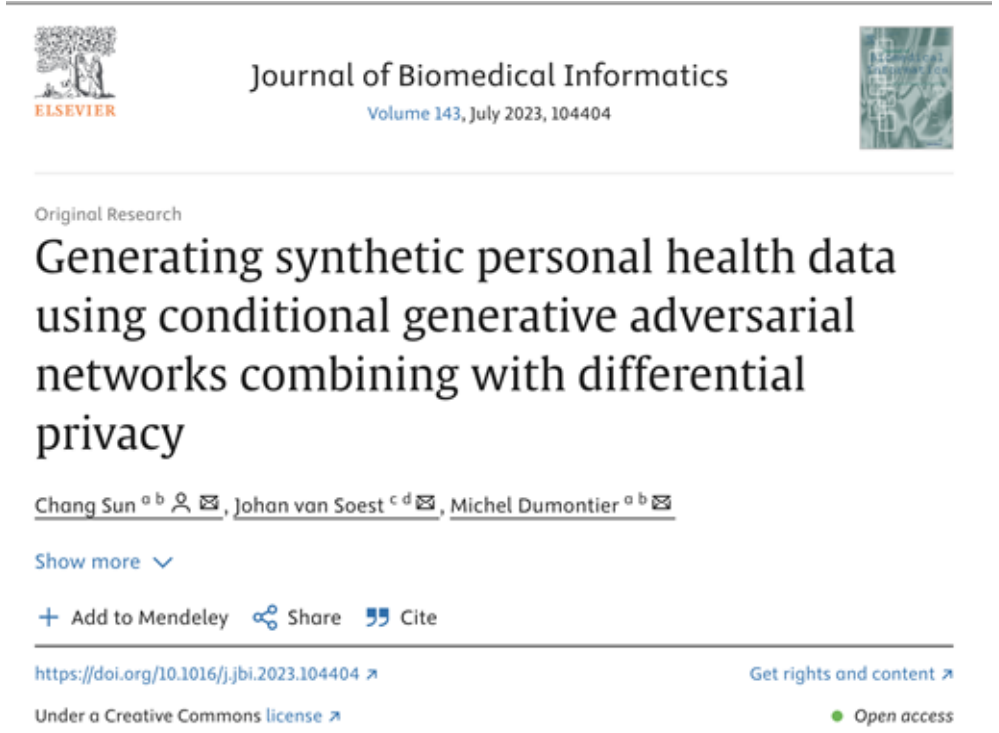
Sun, Chang, Johan van Soest, and Michel Dumontier. "Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy." Journal of Biomedical Informatics 143 (2023): 104404.

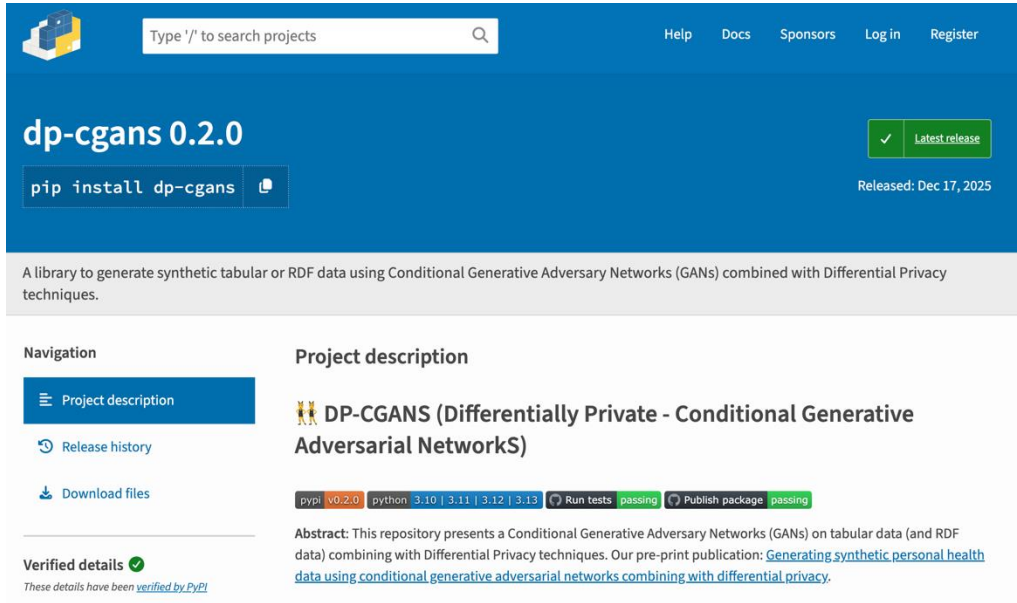
Maastricht University

Sounds complicated but Easy to Use



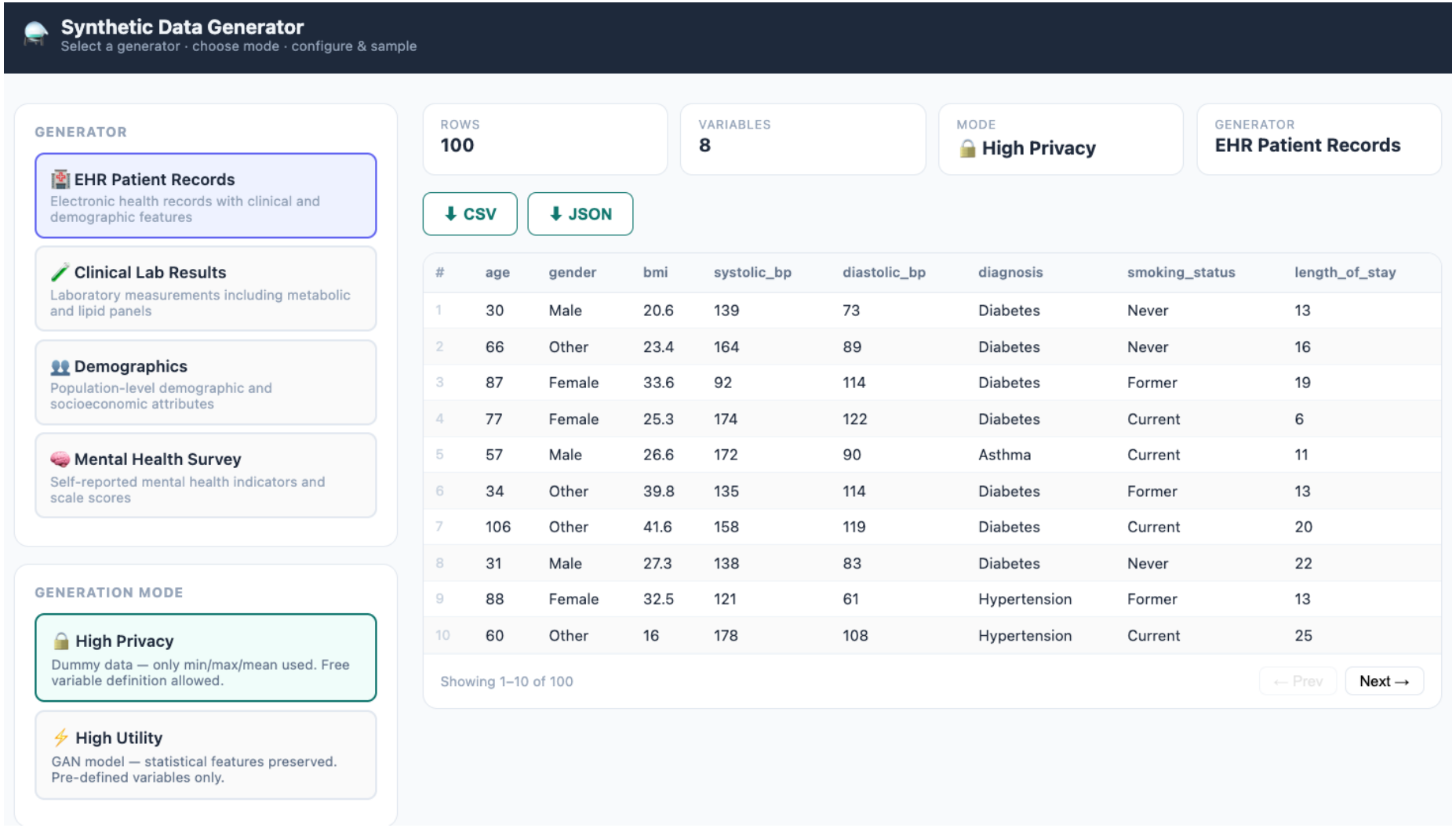
(Python package,
open-source)





Maastricht University

User interface



How do we choose synthetic data (generator)?

Privacy and Utility Trade-off

Meta-Syn

DP-CGANS model

More privacy

More utility

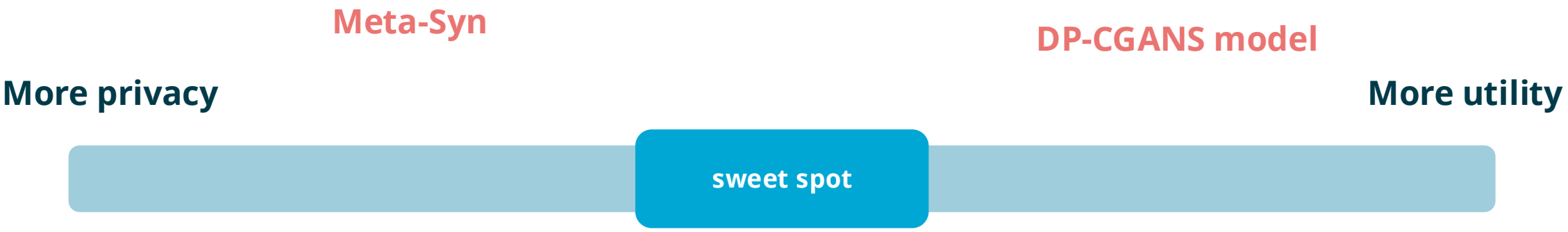
Push too far and the data collapses toward the average. Safe, but it loses statistical characteristics

Push too far the other way and it gets too close to real individuals.

Faithful enough to be useful.

Private enough to be safe.

TDCC-SSH: Synthetic data leveraging the potential of sensitive data in SSH research



Open Discussion:

What this means for a data repository

Open the closed

Publish a synthetic version **alongside** a restricted dataset.

Anyone can explore, prototype, and decide if it fits — before requesting the real data.

A synthetic front door

Make the synthetic layer the **open entry point** and keep the real data behind controlled access.

- Treat each synthetic dataset as a citable object — its own metadata, provenance, and DOI.
- Findable and reusable by default, while sensitive data stays protected.

Safe data for teaching and testing

- Run courses and tutorials on realistic synthetic data data.
- Let others test and reproduce a pipeline without ever touching real records.

What should be considered

- Some privacy risk always remains → we measure and report it with synthetic data.
- Rare and extreme cases could be lost → we are working on rare cases generation
- Always validate against the question you actually care about → Choose right evaluation

It is a powerful tool that should be used responsibly.

Where DANS could lead

- Offer a **synthetic twin** next to restricted datasets.
- Set **metadata standards** for synthetic data objects.
- Share **simple guidance** for evaluating quality and privacy.

Key messages

- Synthetic data shares the real data **patterns**, not the **individual records**.
- It turns closed data into an **open starting point**.
- Faithful enough to be useful, private enough to be safe.
- Repositories are the natural place to make the synthetic data useful and impactful.

DANS Open Day
Open data, open science

Synthetic Data for Open Science

Dr. Chang Sun, Assistant Professor
Department of Advanced Computing Sciences
Maastricht University

Thank you!

DANS