Opportunities and Challenges in linking and reusing education data in the Netherlands

Chang Sun, Birgit Wouters, Carlos Utrilla Guerrero, Michel Dumontier
Maastricht University

Maastricht University
Institute of Data Science

ODISSEI

ODISSEI Project Task 3.1

Task objective:

To explore the use of the novel privacy-preserving distributed analytics technologies for social science research

Research questions:

- What is the impact of the presence of special educational needs (SEN) students on the social and emotional development of students without SEN?
- What is the impact of the presence of SEN students on teacher work stress?

ODISSEI Project Task 3.1

Task objective:

To explore the use of the novel privacy-preserving distributed analytics technologies for social science research

Research questions:

- What is the impact of the presence of special educational needs (SEN) students on the social and emotional development of students without SEN?
- What is the impact of the presence of SEN students on teacher work stress?

Informed consent
Underaged children

ODISSEI

ODISSEI Project Task 3.1

Task objective:

To explore the use of the novel privacy-preserving distributed analytics technologies for social science research

Research questions:

- What is the impact of the presence of special educational needs (SEN) students on the social and emotional development of students without SEN?
- What is the impact of the presence of SEN students on teacher work stress?

However, we face the legal challenges in linking and reusing education data in the Netherlands.

the opportunity of using **synthetic data** in light of the **GDPR**

ODISSEI

Our objective for Period 3

to develop a synthetic data generator framework from technical and ethical-legal perspectives that will enable the examination of the trade-off between data privacy and the potential utilization of synthetic representations

Our actions for Period 3

Centraal Bureau voor de Statistiek

SURF

We will study

1. the **quality of synthetically generated data** to real world data as a function of privacy cost
2. the **quality of preservation of multi-attribute relation**s in the face of increased individual variation
3. the **utility of synthetic data** in certain kinds of social science research
4. both EU and Dutch **law, regulation and policies** pertaining to the generation and use of synthetic data from personal data.

Maastricht University Institute of Data Science

NRO NATIONAAL REGIEORGAAN ONDERWIJSONDERZOEK

Inspectie van het Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap

ODISSEI

Synthetic Data: 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) [1];

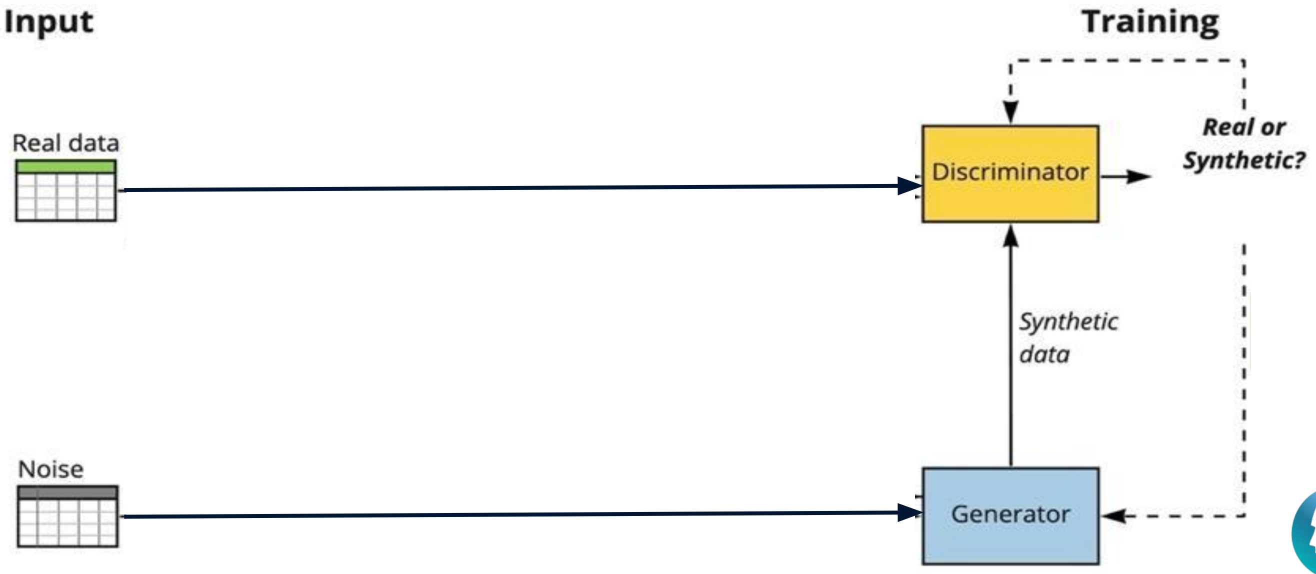
- at the **population** level (e.g., marginal and joint distributions of features).

- 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.

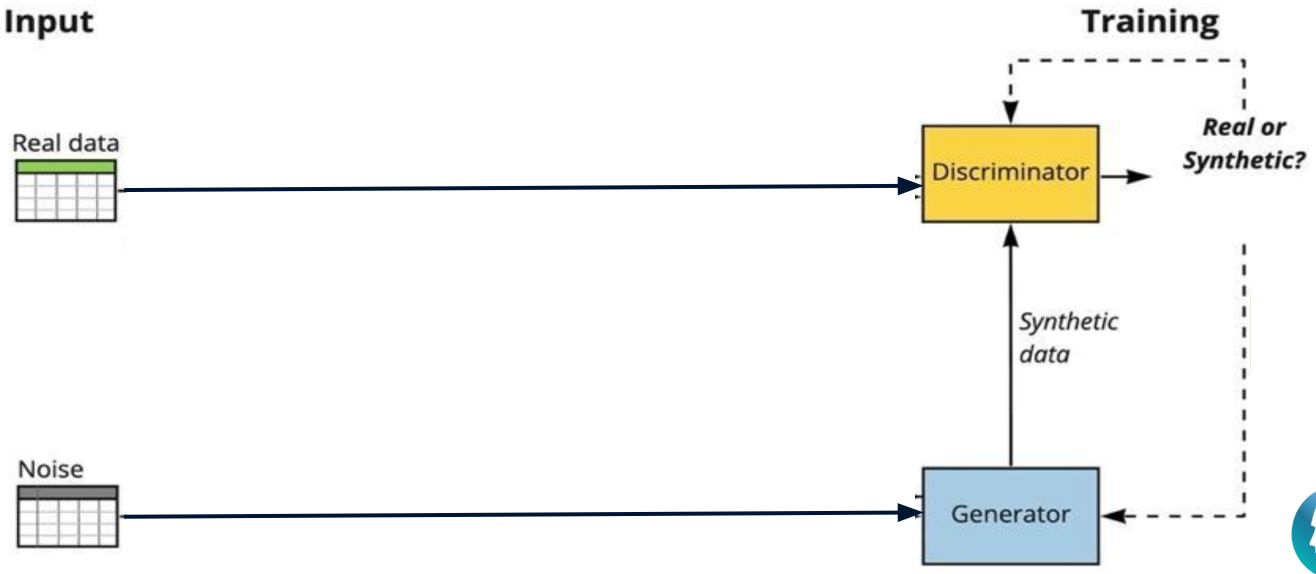
Our approach: Differentially Private Conditional Generative Adversarial NetworkS (DP-CGANS)

Trains and leverages two opposing neural network models (a generator and a discriminator) in a competitive manner.



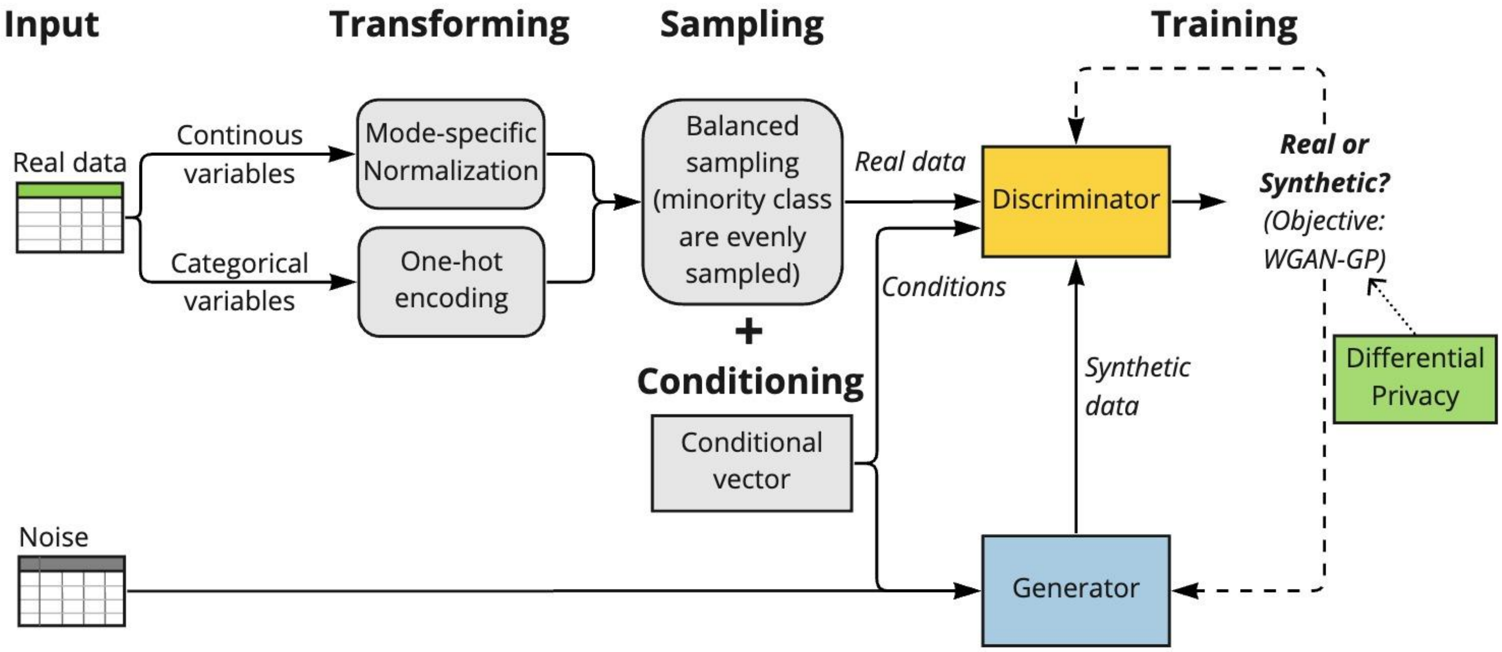
Our approach: Differentially Private Conditional Generative Adversarial NetworkS (DP-CGANS)

Trains and leverages two opposing neural network models (a generator and a discriminator) in a competitive manner.



Our approach: Differentially Private Conditional Generative Adversarial NetworkS (DP-CGANS)

Trains and leverages two opposing neural network models (a generator and a discriminator) in a competitive manner.



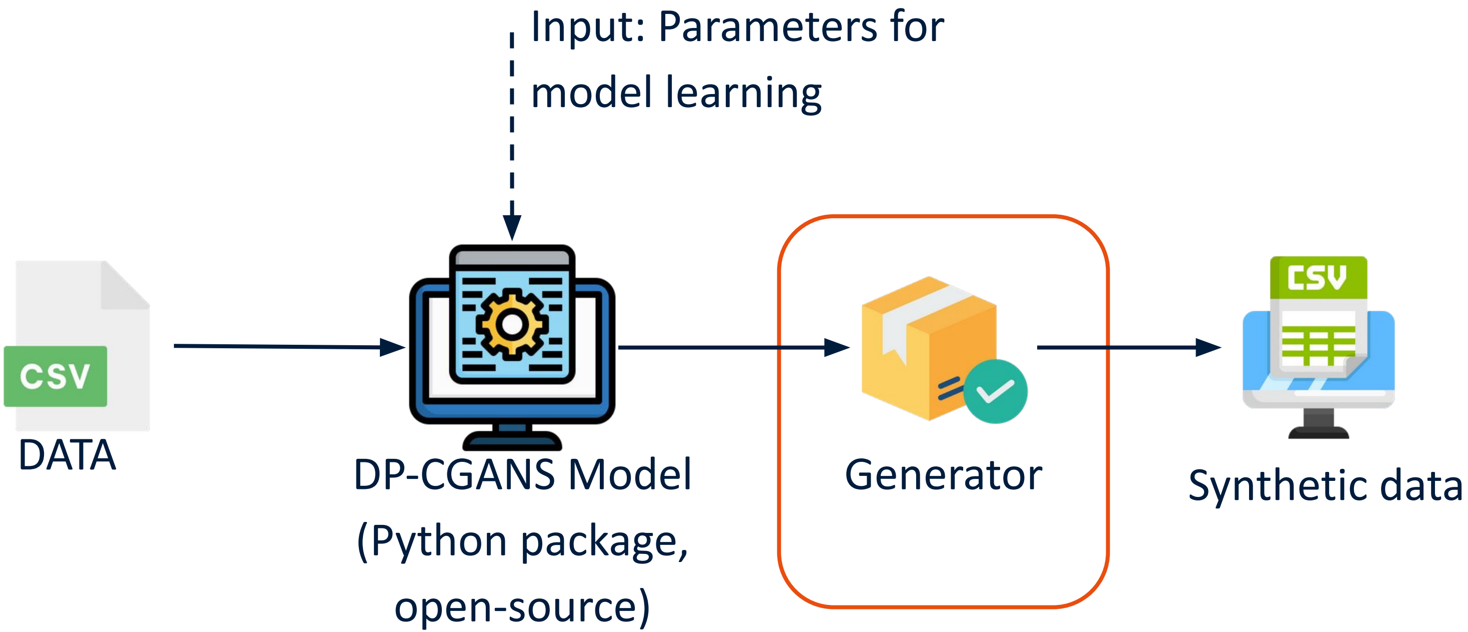
DP-CGANS structure - Differential Privacy (DP)

Uses a solid mathematical formulation to measure the privacy and provide theoretical privacy guarantees by typically adding noise when training the models.

DP protects the participation of individual data point in the datasets.

- replacing or removing one data point with another one will not make an observable change in the results

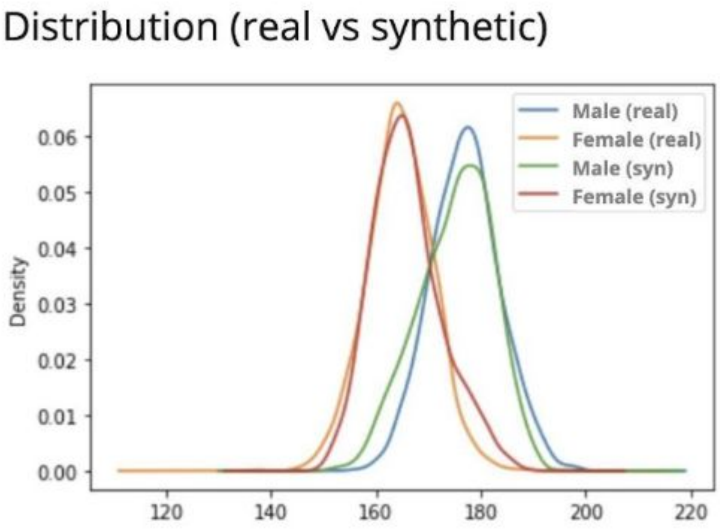
Sounds complicated but Easy to Use

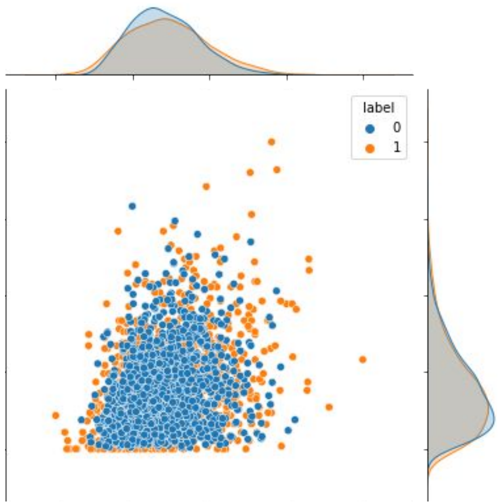


Maastricht University
Institute of Data Science

ODISSEI

How synthetic data looks like





Privacy vs Utility

Privacy budget **ε=100**

Privacy < Utility

Sex	Age	Height	Weight	Waist	Hip
male	78	184.65	132.80	97.178	119.92
female	68	202.47	58.252	113.85	109.60
male	57	169.98	81.766	116.70	106.25
male	69	196.88	82.293	71.857	122.92
male	80	164.19	79.838	79.849	114.13
female	74	192.93	84.968	62.205	113.87

Privacy budget **ε=0.01**

Privacy > Utility

Sex	Age	Height	Weight	Waist	Hip
male	30	135.98	52.19	62.77	72.56
male	30	135.98	52.19	62.77	72.56
male	30	135.98	52.19	62.77	72.56
male	50	135.98	52.19	62.77	72.56
male	30	135.99	52.19	62.77	72.56
female	50	135.98	52.19	62.77	72.56

Privacy <------------ | ------------> Utility

How we balance this trade-off?

Privacy vs Utility

The goal of legal WP is to create a legal framework allowing for the use of synthetic data as an alternative to real-world data that meets the standards set by EU law (in particular the GDPR), regulation and policy.

Privacy <------------ | ------------> Utility

Key question: when is the synthetic dataset sufficiently different from the original dataset to be classified as truly anonymous

Recent progress

Inspectie van het Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap

Centraal Bureau voor de Statistiek

SURF

- Using student enrolment data, measures of youth support and youth protection data, medication data from students in primary and secondary schools (CBS Microdata)

- Using ODISSEI Secure Supercomputer (OSSC) to train the generator collaborating Inspectorate of Education, SURF, and CBS)

- Next: evaluate the quality of synthetic data with different level of privacy preservation by comparing the analysis performance on real and synthetic data *(e.g., on the analysis of the effect of SEN students on the cognitive (and socio-emotional) outcomes of non-SEN students*

Discussion

Opportunities: call for more use cases for synthetic data

- to replace real data in some research studies?
- for training or education purpose?

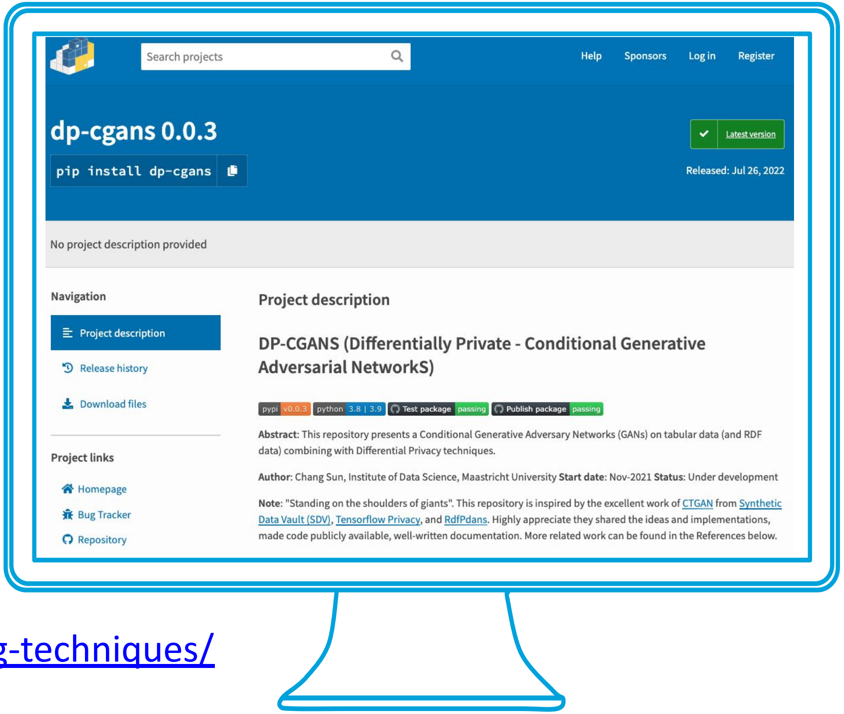
Questions:

- When is the synthetic dataset sufficiently different from the original dataset to be classified as truly anonymous?
- What if the synthetic dataset (strongly) resembles an individual contained in the real dataset?
- What is the accuracy of a model trained with synthetic data to infer/predict individual attributes?

THANKS!

Further discussion and questions?
chang.sun@maastrichtuniversity.nl

Task webpage:
https://odissei-data.nl/en/privacy-preserving-techniques/



Pypi: https://pypi.org/project/dp-cgans/
Github: https://github.com/sunchang0124/dp_cgans

Maastricht University
Institute of Data Science

ODISSEI

Thanks to partners:

- **Maastricht University,** who brings expertise in data science, data management, distributed data analysis
- **Netherlands Initiative for Education Research**, who maintains close connections to the schools who will make available the education-related data,
- **Inspectorate of Education**, who seeks to perform the analysis, and
- **Statistics Netherlands**, who has relevant microdata and has established a research-grade secure computing environment to undertake privacy-preserving research.
- **SURF (ODISSEI Secure Supercomputer, OSSC)**, who provides high performance computation environment and facilitates our generative model to run on a GPU

NRO NATIONAAL REGIEORGAAN ONDERWIJSONDERZOEK

Inspectie van het Onderwijs
Ministerie van Onderwijs, Cultuur en Wetenschap

Centraal Bureau voor de Statistiek

SURF

Maastricht University
Institute of Data Science

ODISSEI