# Methodology for measuring the usefulness and privacy of synthetic data

Maryline Laurent (TSP)

Companies need to store large data sets on their customers in order to enrich their knowledge base. To avoid the regulatory constraints of the GDPR, they must, for certain purposes (e.g., using data to create artificial intelligence models), anonymize their data while meeting the dual requirement that the anonymized data be useful (viable) to the company and that it not reveal any information relating to privacy.

A new approach is emerging, that of synthetic data generators (e.g., MostlyAI, Syndiffix), particularly those based on GANs (generative adversarial networks). These generators are trained using real customer data and are then able to generate as much "synthetic data" as desired, with the advantage that this data resembles real data but does not actually correspond to any customer, which suggests lower risks to personal privacy.

In this context, synthetic data is of interest to companies, but it must still be useful, i.e., it must be meaningful enough to demonstrate its effectiveness in relation to the company's needs.

The internship will consist of taking the research a step further than the work carried out by TSP students. In particular, it will involve:

- familiarizing yourself with metrics for evaluating the usefulness of synthetic databases compared to real databases (in particular SDMetric [1])
- identifying synthetic datasets that have multiple uses on Kaggle [2]
- measuring the different utility metrics—overall and for each purpose—for at least one dataset, and analyzing them
- providing the results of the developments obtained on GitHub

## Required expertise

- Background in AI or a strong interest in statistics
- Development in Python

## Expected deliverables

- The software code

- A report detailing all considerations, difficulties, and achievements

## References:

[1] https://docs.sdv.dev/sdmetrics/reports/quality-report/whats-included#column-pair-trends [2] https://www.kaggle.com