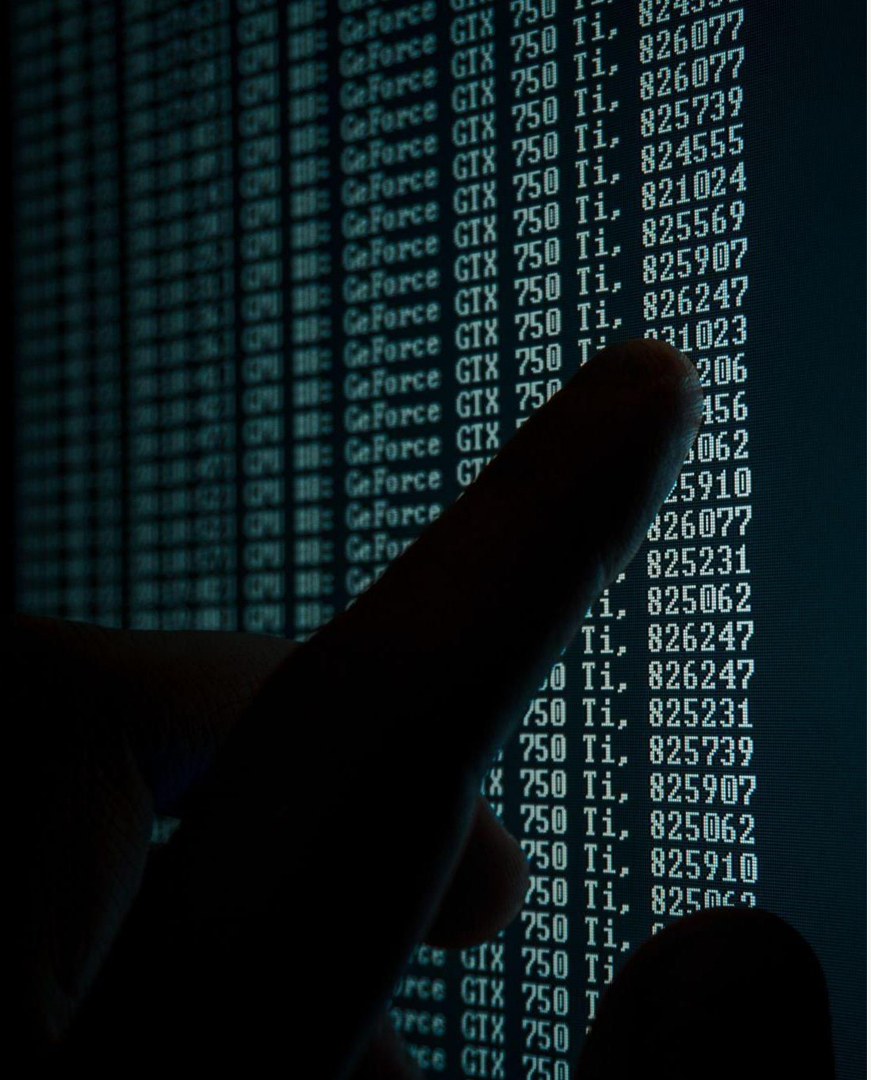


Privacy-Preserving Synthetic Data Generation

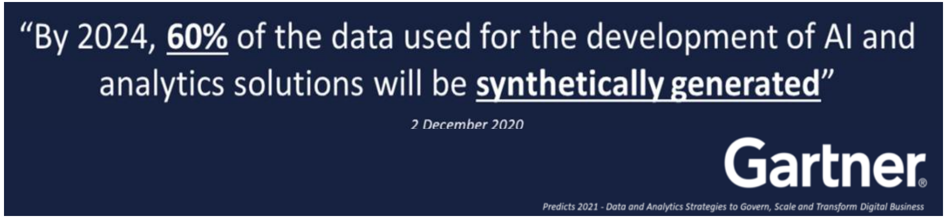
Marta Batlle and Finn Janson

25 July 2022

Table of contents

1. Background
2. What is synthetic data generation?
3. How can synthetic data be generated?
4. Quality and privacy evaluation of synthetic data
5. Why synthetic data at Roche?
6. The challenges of synthetic data
7. Q&A

The rapid adoption of Synthetic Data Generation



"By 2024, **60%** of the data used for the development of AI and analytics solutions will be **synthetically generated**"

2 December 2020

Gartner

Predicts 2021 - Data and Analytics Strategies to Govern, Scale and Transform Digital Business

NEURIPS RESEARCH TOPICS: NUMBER of ACCEPTED PAPERS on PRIVACY in AI, 2015–21

Source: NeurIPS, 2021; AI Index, 2021 | Chart: 2022 AI Index Report

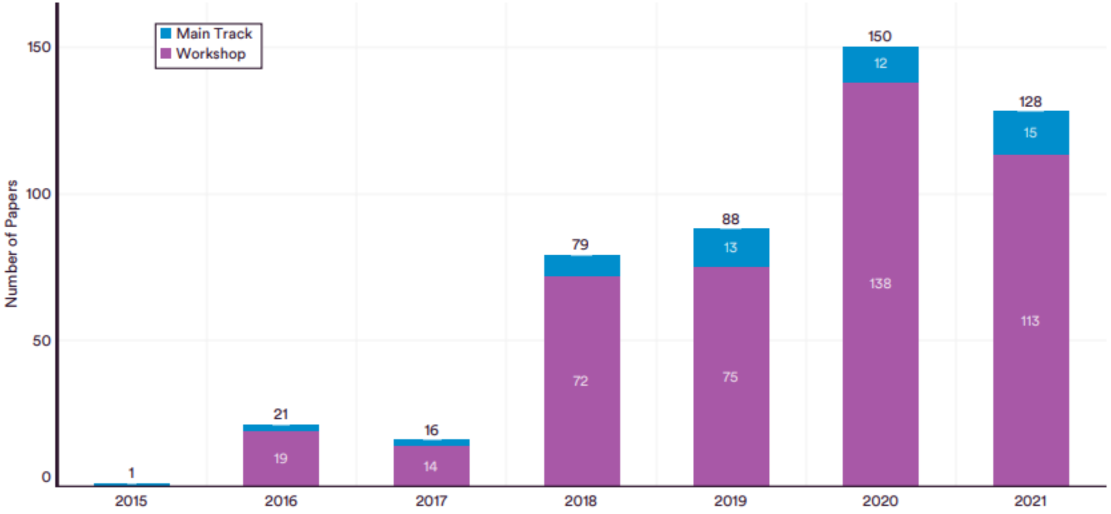
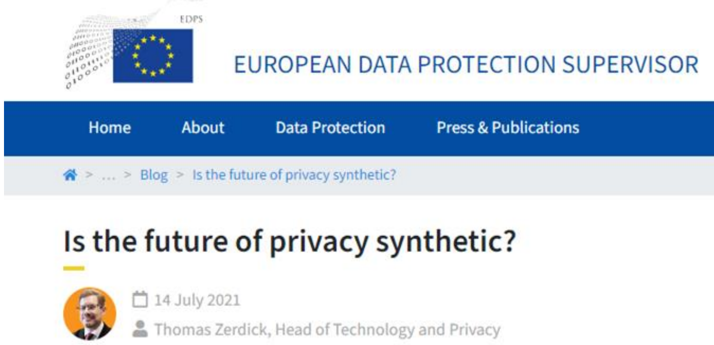


Figure 3.3.6



EDPS

EUROPEAN DATA PROTECTION SUPERVISOR

Is the future of privacy synthetic?

14 July 2021

Thomas Zerdick, Head of Technology and Privacy

The core problems of sharing healthcare data

Due to privacy concerns, it is often **difficult or impossible to access real healthcare data**

Many existing anonymisation techniques highly compromise the **quality of the data** that is accessible

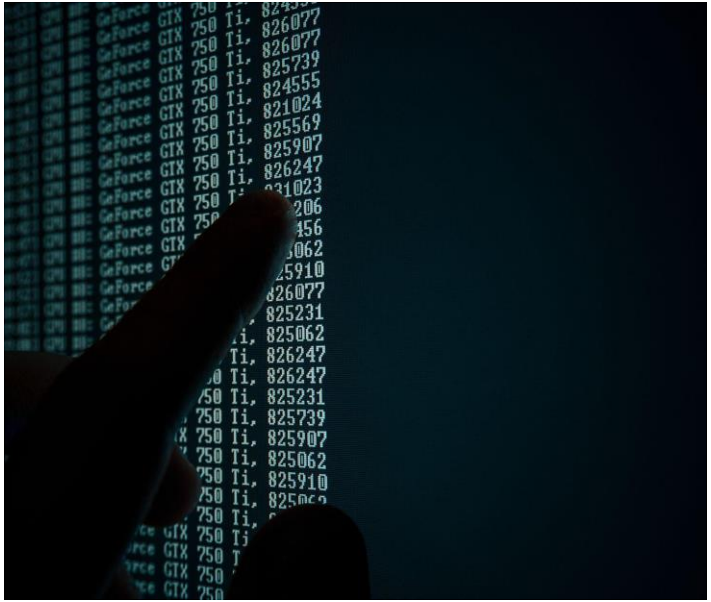
Legal uncertainties establish a gray area for existing anonymisation techniques

Limitations of existing solutions - a GDPR perspective

	Key coded (Raw Data)	GDPR Pseudonymisation	Risk-based de identification	Homomorphic encryption	Federated learning
Definition	Direct identifiers are replaced with a unique code	Direct (name) and indirect identifiers (address) are replaced	Different level of de-identification applied based on risk of the dataset	Allows data to remain encrypted while it's being processed	Train an algorithm across multiple decentralized servers holding local data
Challenges	Re-identification is easy with access to key	To use only for the specified purpose (pseudoanonymised)	Challenging to assess the right level of compromise	Does not resolve data compliance on it's own	Infrastructure, negotiation, data standardisation
Privacy risk	🟠	🟠	🟢	🟢	🟢
Data utility	🟢	🟢	🟡	🟢	🟢
Ease of use	🟢	🟢	🟡	🔴	🔴
Cost/access	🟢	🟢	🟡	🔴	🟡

What is Synthetic Data generation?

What is synthetic data generation?



A modelling technique that allows us to generate a dataset that retains the **same statistical properties** as the original dataset, **without compromising privacy**

How can it be generated?

What are Generative Adversarial Networks (GANs)?

A model to generate synthetic data

"Art forger"

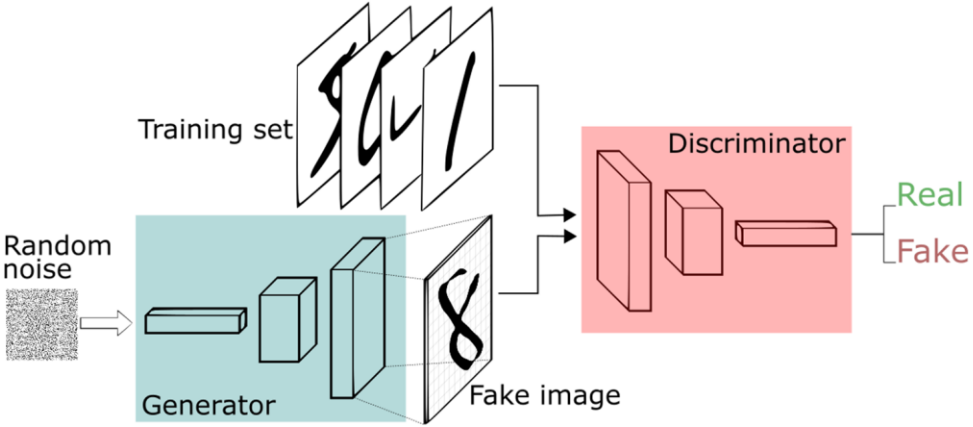
The generator attempts to create fake data

"Art inspector"

The discriminator tries to distinguish between real and fake data

Two competing neural networks - **generator and discriminator**

Architecture of Generative Adversarial Networks (GANs)



The **generator** attempts to fool the **discriminator**

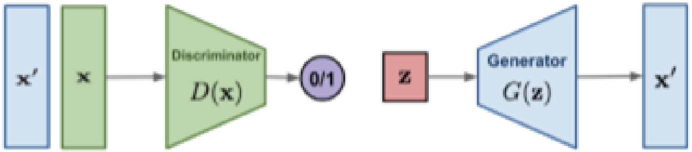
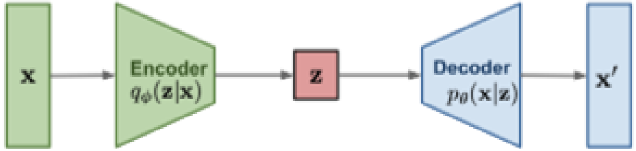
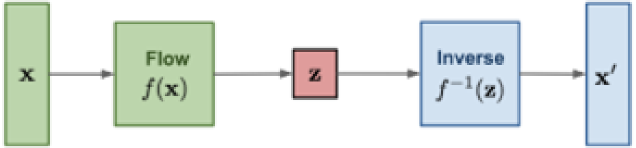
Discriminator & Generator may reach Nash equilibrium

Discriminator may be discarded for security reasons

Generator is used to create synthetic data

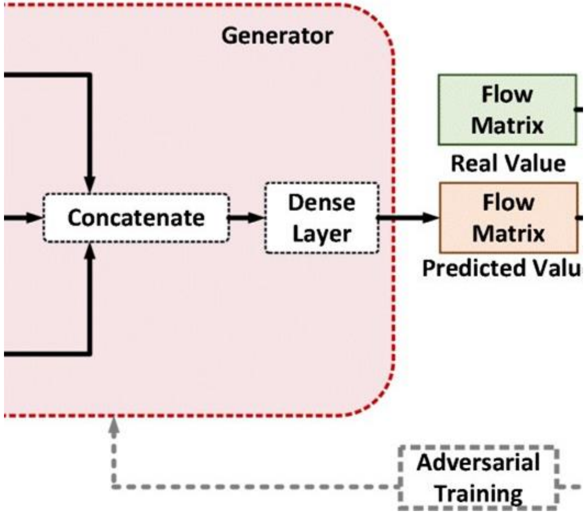
Alternatives to GANs

Other models for generating synthetic data

Model	Advantages	Architecture
GAN: minimax the classification error loss	higher quality data generated than other methods (typically)	 $\mathbf{x}'$ $\mathbf{x}$ Discriminator $D(\mathbf{x})$ 0/1 $\mathbf{z}$ Generator $G(\mathbf{z})$ $\mathbf{x}'$
VAE (variational autoencoder): maximize evidence lower bound	invertible, stable training	 $\mathbf{x}$ Encoder $q_\phi(\mathbf{z}\|\mathbf{x})$ $\mathbf{z}$ Decoder $p_\theta(\mathbf{x}\|\mathbf{z})$ $\mathbf{x}'$
Flow-based generative model: minimize negative log-likelihood	invertible, stable training	 $\mathbf{x}$ Flow $f(\mathbf{x})$ $\mathbf{z}$ Inverse $f^{-1}(\mathbf{z})$ $\mathbf{x}'$

Hybrid models

- Flow-GANs, a generative adversarial network for which we can perform exact likelihood evaluation, thus supporting both adversarial and maximum likelihood training, can help with mode collapse while having a effective regularization



Improving Generative Models

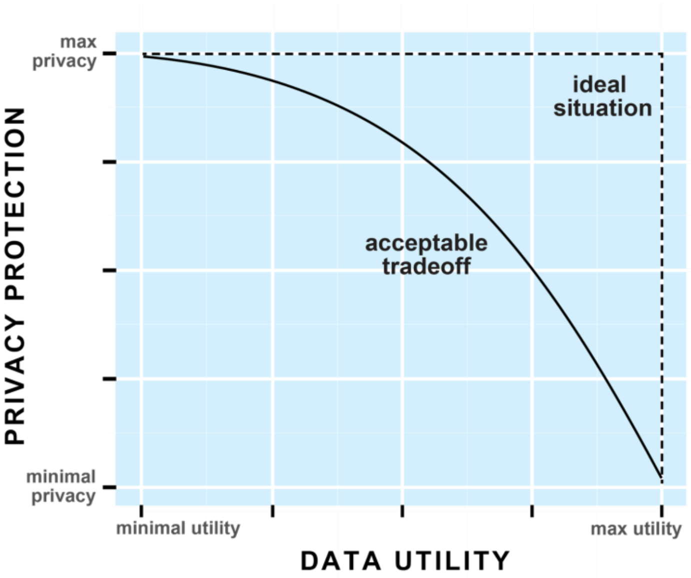
- As demonstrated with Flow-Gans, there are **numerous ways to improve the synthetic data generation process**
- These can be achieved by:
    - **Changing objection functions**
    - **Trying different neural architecture paradigms**
    - **Leveraging large-scale foundational models** (e.g Inception model trained on Imagenet)

Quality and privacy evaluation of synthetic data

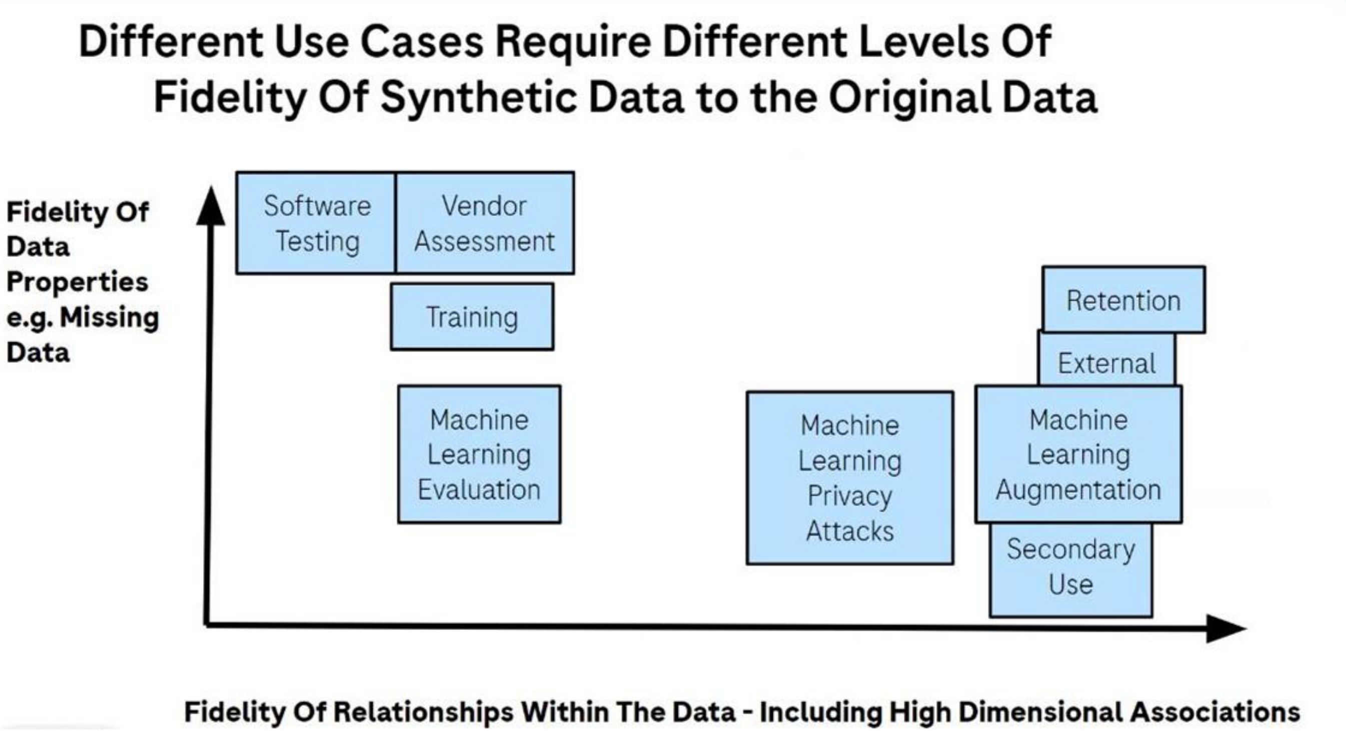
A trade-off between quality and privacy

The higher the privacy/noise, the lower the quality of the synthetic data

Although the statistical properties of real and synthetic datasets are similar, there is **a tradeoff between the privacy and quality** of a dataset based on the level of noise added



A trade-off between quality and privacy



Different Use Cases Require Different Levels Of Fidelity Of Synthetic Data to the Original Data

Fidelity Of Data Properties e.g. Missing Data
Software Testing
Vendor Assessment
Training
Machine Learning Evaluation
Retention
External
Machine Learning Privacy Attacks
Machine Learning Augmentation
Secondary Use
Fidelity Of Relationships Within The Data - Including High Dimensional Associations

Quality evaluation

How do we assess the quality of the synthetic data?

Feature distributions

Feature correlations

Distance functions

ML performance

Privacy evaluation

How do we assess the privacy of the synthetic data?

Article 29 Working Party (EU, 2014)

Within Europe, from a legal standpoint, there are 3 main criteria to define whether a dataset is anonymised or not:

Singling out

The ability to isolate some or all of the records that identify an individual

Linkability

The ability to link a record in one database to a record in the real database

Inference

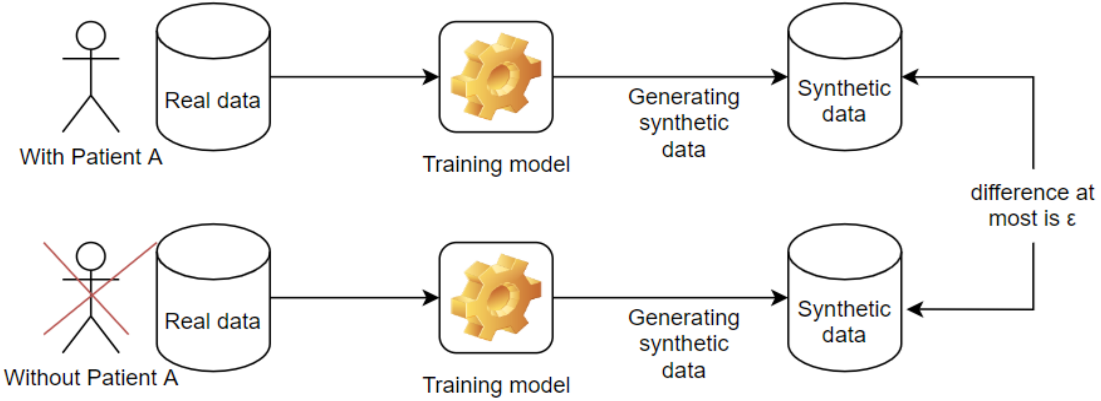
The possibility of deducing the value of an attribute from the values of another

What is differential privacy (DP)?

A mathematical framework used to ensure privacy

- A framework for measuring the privacy guarantees provided by an algorithm
- Epsilon ε value (or "privacy budget") can be seen as the amount of noise applied to the data.

  Higher ε value=higher privacy (and lower quality)
- Epsilon ε value in DP measures the privacy loss when removing or adding one entry



Why synthetic data at Roche?

Synthetic data generation at Roche

What do we have, what do we need

The problem

- Current regulations challenge the process of data sharing
- Other data anonymization techniques reduce the quality of results

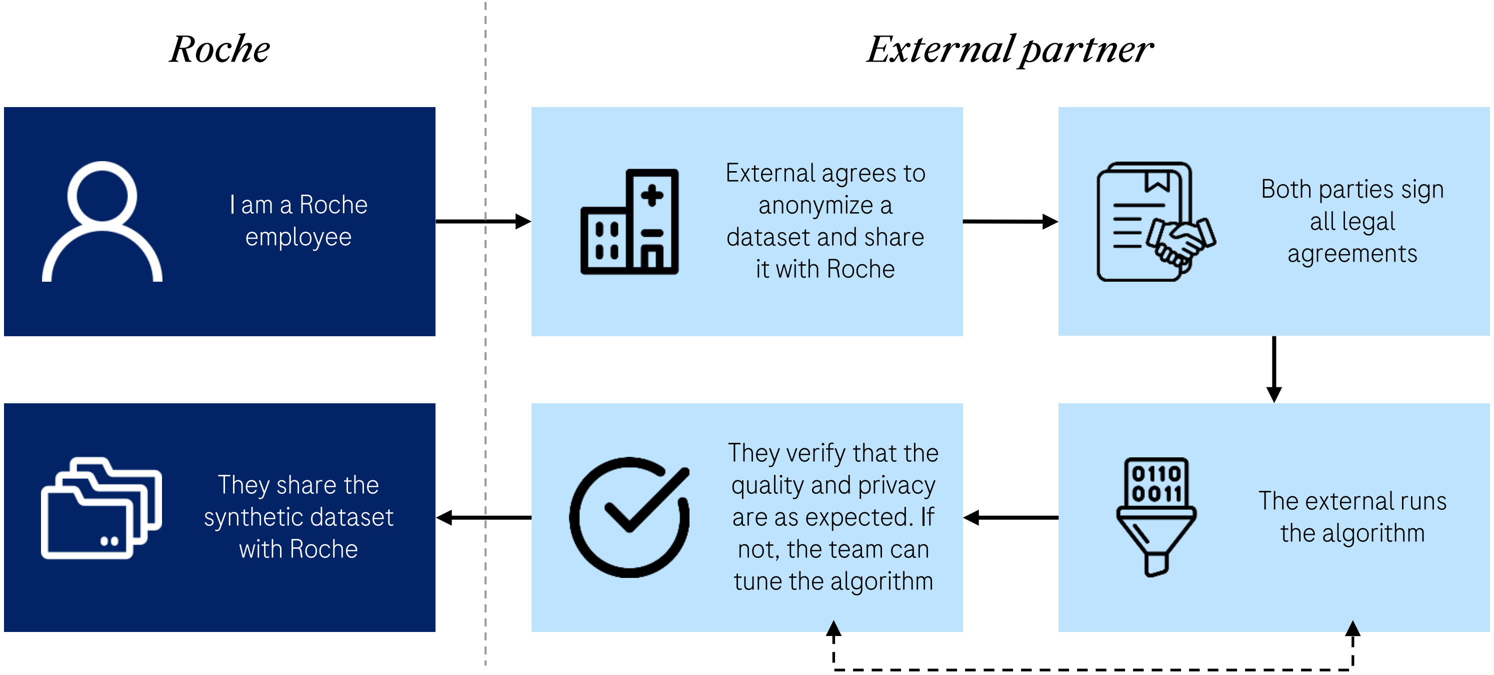
We have

- A team with great expertise on synthetic data algorithms

We need

- To improve the quality of synthetic datasets
- A quality and privacy evaluation that we trust

How would the process work?



Roche

External partner

I am a Roche employee

External agrees to anonymize a dataset and share it with Roche

Both parties sign all legal agreements

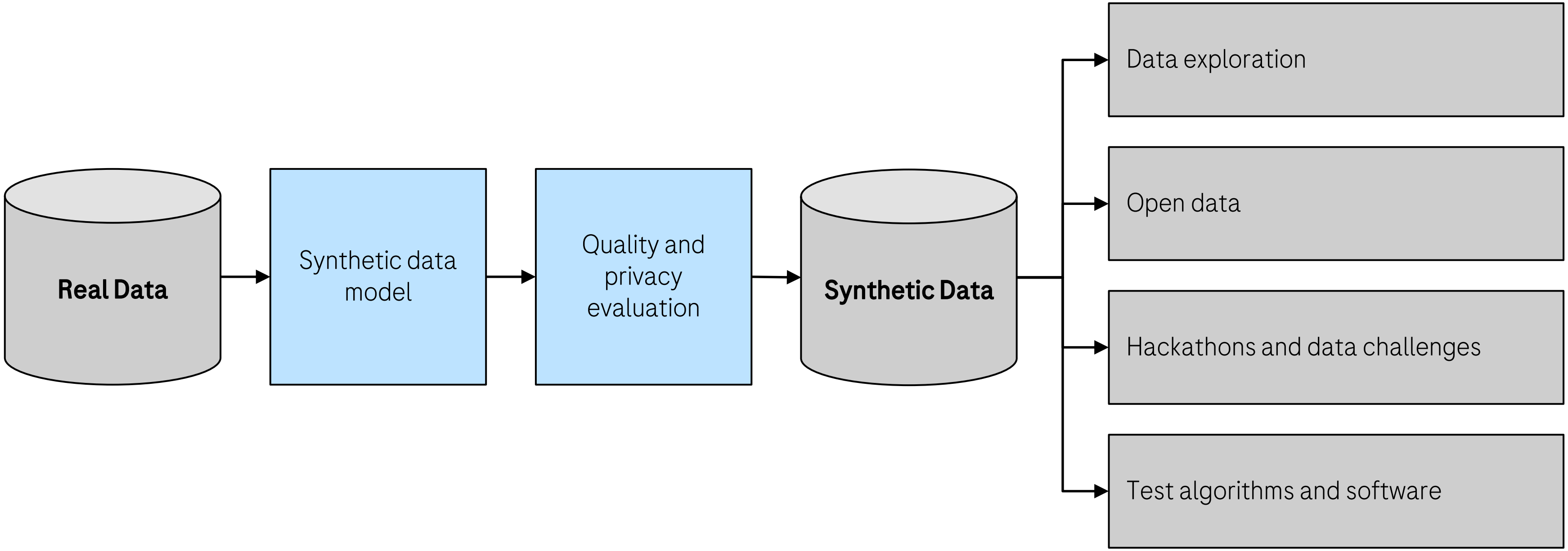
The external runs the algorithm

They verify that the quality and privacy are as expected. If not, the team can tune the algorithm

They share the synthetic dataset with Roche

The solution

To answer Roche's needs



Real Data

Synthetic data model

Quality and privacy evaluation

Synthetic Data

Data exploration

Open data

Hackathons and data challenges

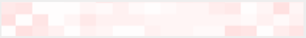
Test algorithms and software

The challenges of synthetic data

Bottlenecks in the creation & adoption of Synthetic Data

- Generating high-quality synthetic data requires big high-quality datasets
- Generating longitudinal datasets is significantly harder than cross-sectional datasets
- Trustability may slow adoption of this technology
- Finding the suitable trade-off for quality-privacy can be difficult and subjective

Non-longitudinal dataset example



Longitudinal dataset example

Q&A

Please contact us if you are interested in academic collaborations!