

Qinyi liu

City university of Macau

An abstract network diagram on a black background. It features numerous nodes, represented by small white circles with black outlines, connected by thin, curved lines in shades of orange, teal, and yellow. The lines create a complex web of connections across the frame, with some lines being straight and others curved, suggesting a dynamic or evolving network structure.

CAN SYNTHETIC DATA
MAKE RESEARCH DATA
MORE OPEN?
OPPORTUNITIES AND
PRIVACY RISKS IN THE
AGE OF AI



WHAT I HAVE BEEN WORKING ON

- Qinyi Liu, Assistant Professor at City University of Macau
- Ph.D. in Centre for the Science of Learning and Technology (SLATE), University of Bergen, Norway
- My PhD is about using privacy-enhancing techniques like synthetic data generation to address privacy issues in datasets

Scaling while privacy preserving: A comprehensive synthetic tabular data generation and evaluation in learning analytics

73

2024

Q Liu, M Khalil, J Jovanovic, R Shakya

Proceedings of the 14th Learning Analytics and Knowledge Conference, 620-631

Can synthetic data be fair and private? A comparative study of synthetic data generation and fairness algorithms

45

2025

Q Liu, O Deho, F Vadiiee, M Khalil, S Joksimovic, G Siemens

Proceedings of the 15th International Learning Analytics and Knowledge ...

Ensuring privacy through synthetic data generation in education

32

2025

Q Liu, R Shakya, J Jovanovic, M Khalil, J de la Hoz-Ruiz
British journal of educational technology 56 (3), 1053-1073

See more here! Website: qinyiliu.site

HUGE ATTENTION HAS BEEN GIVEN TO SYNTHETIC DATA

- *"Right now we have exhausted practically the entirety of human knowledge for training AI... The only way to augment real-world information is synthetic data."* — **Elon Musk**
- *"Today, human-generated knowledge is 99%. In 10 years, it will probably be 1%. The vast majority of knowledge will be generated by AI—it's going to be synthetic generated intelligence."* — **Jensen Huang**



You collected the perfect dataset.

- ✓ Rich, longitudinal patient histories.
- ✓ High statistical power.
- ✓ Ready for global collaboration.



But you cannot share it.

Record ID: **UIO-992-8A**

Patient Age: **42 Years Old**

Location: **Oslo, Norway**

Diagnosis: **Type 2 Diabetes**

Outcome: **Full Remission**

GDPR PRIVACY BLOCK

The Reality

Direct records contain sensitive personal identifiers.

Strict ethical constraints prevent raw data release.

Progress halts completely.

THE CURRENT IMPASSE

Two broken options.



The Data Silo

Locked away securely.
Zero scientific reproducibility.



Heavy Anonymization

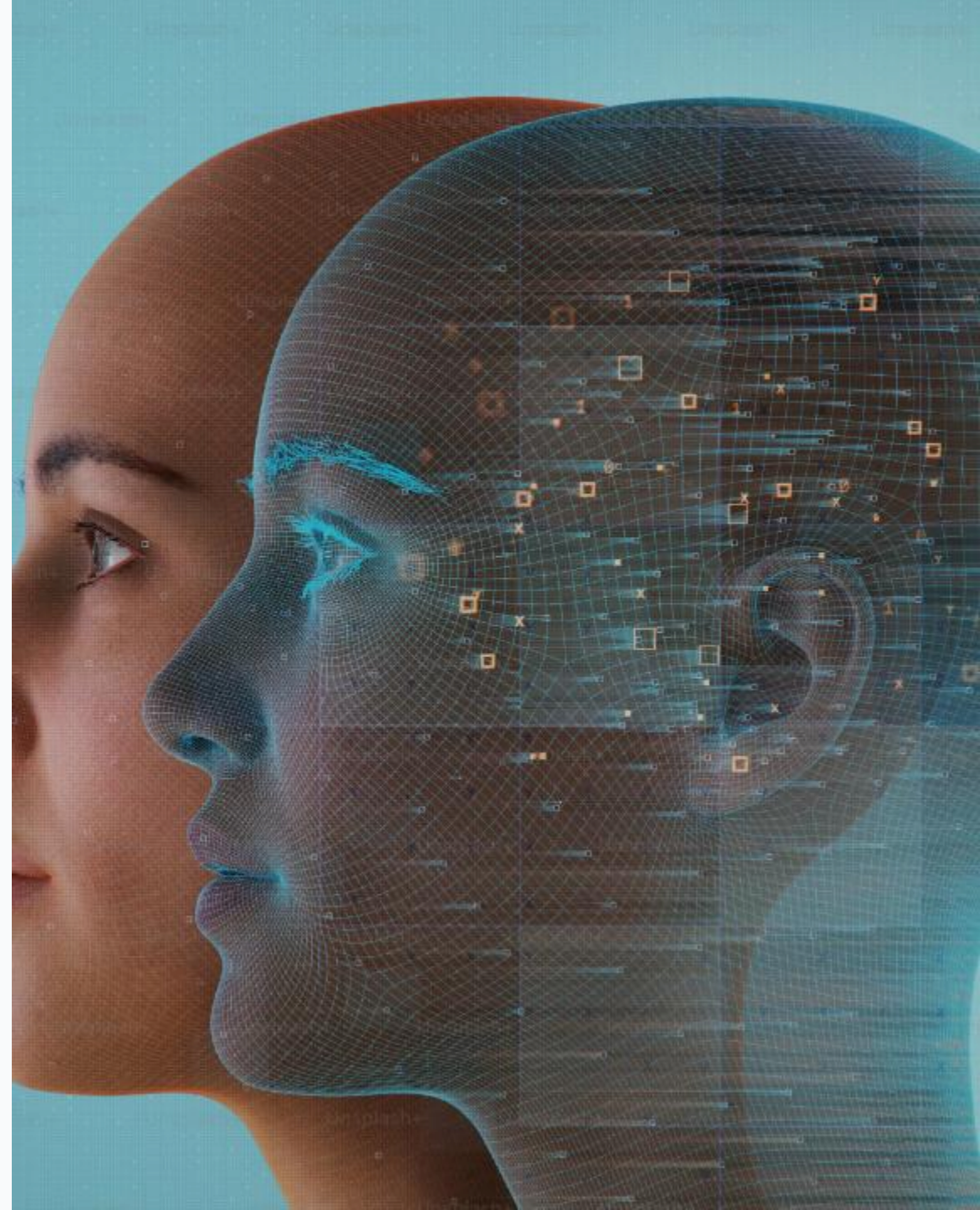
Information deeply aggregated.
Destroyed analytical utility.

A NEW PARADIGM

What if we share the pattern, not the actual data?

Enter AI-generated synthetic data.

Statistical fidelity maintained. Privacy preserved.

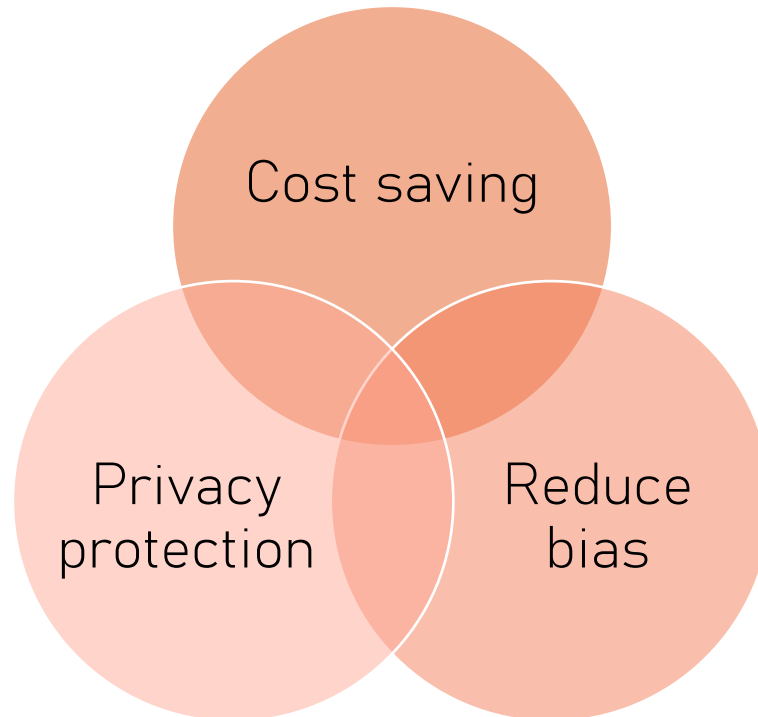


TODAY'S PLAN

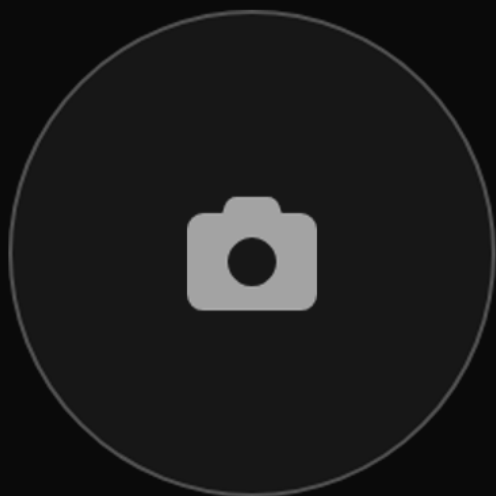
- What is Synthetic data, and what opportunities can it bring?
- How to generate synthetic data
- How to evaluate synthetic data
- One example of using synthetic data for open science
- Limitation of synthetic data

WHAT IS SYNTHETIC DATA?

- “Generally, synthetic data is defined as the artificially annotated information generated by computer algorithms or simulations” [1].
- Synthetic data can



The Photograph



Real data.
Captures a specific person.

The Painting



Synthetic data.
Statistically realistic, but fictional.

How this enables Open Science.



Secure Vault

Original raw data remains isolated.



AI Generation

Model learns the statistical distributions.



Open Access

Publish synthetic data openly for code-testing.

THE MODALITIES

A flavor for every discipline.



Tabular

CLINICAL RECORDS, SURVEYS

Spreadsheets that preserve complex variable correlations without real humans.



Time-Series

ECGS, CLIMATE SENSORS

Dynamic data capturing accurate temporal dependencies and natural rhythms.



Text

INTERVIEWS, MEDICAL NOTES

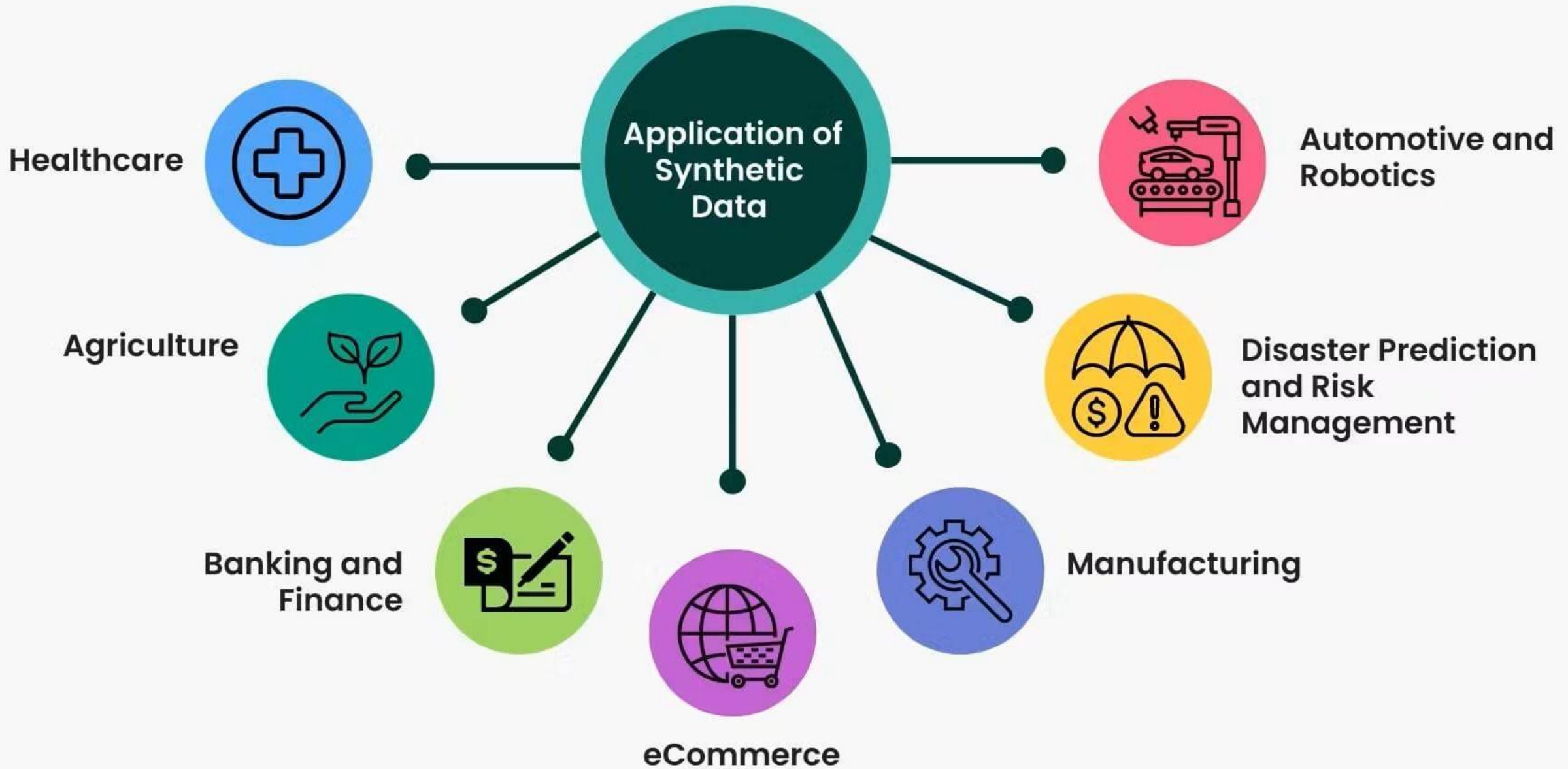
Transcripts reflecting specific jargon and emotional tone for NLP training.



Image & Video

AUTONOMOUS DRIVING, SCANS

Beyond privacy: Simulating dangerous or ultra-rare "edge cases" safely.



HOW TO GENERATE SYNTHETIC DATA

- Statistical & Copula Models
- Deep Learning: GANs, VAEs, Diffusion
- LLMs & Agentic Generation

STATISTICAL & PROBABILISTIC MODELS

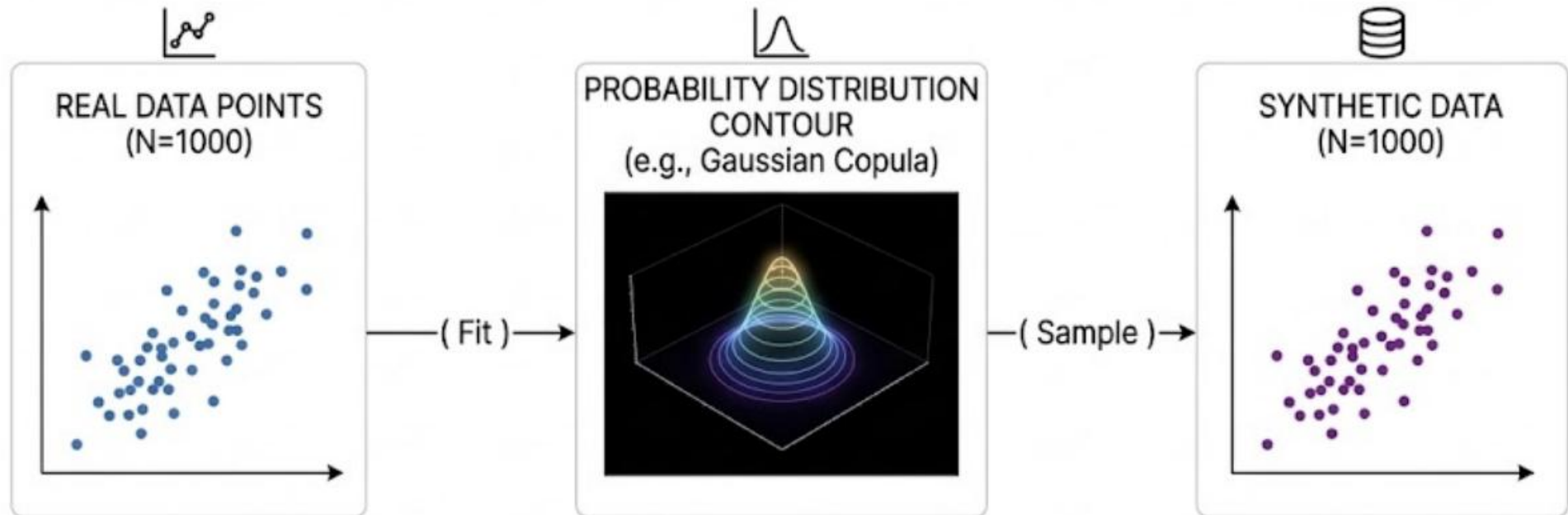
Representative algorithms: Gaussian Copulas, Bayesian Networks, Conditional Independent Models.

Working principle: Fits the marginal distribution and covariance matrix of the original data, then randomly samples from them.

Applicable scenarios: tabular data.

Advantages and disadvantages: Extremely high interpretability, extremely fast computation, and easy to prove with differential privacy; however, it cannot capture extremely complex nonlinear relationships or unstructured data.

STAT METHODS



DEEP

GENERATIVE

MODELS

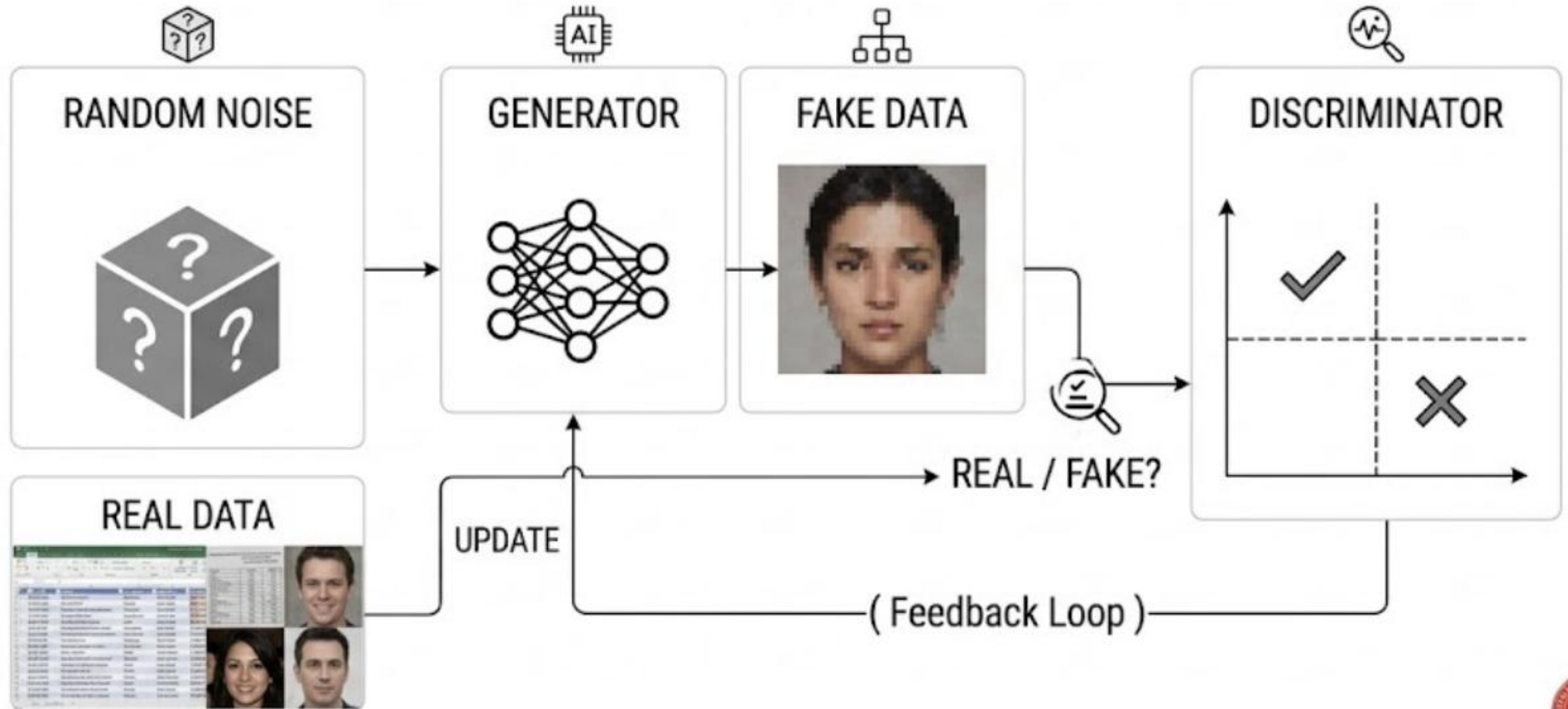
Representative Algorithms:

GANs (Generative Adversarial Networks): such as CTGAN (for tables), TimeGAN (for time series), StyleGAN (for images).

VAEs (Variational Autoencoders): such as TVAE.

Diffusion Models: such as TabDDPM (table diffusion model), Stable Diffusion/FLUX (for images).

SYNTHETIC DATA GENERATION: DEEP LEARNING (GANs)



DEEP

GENERATIVE

MODELS

Working Principle: Implicitly fits extremely complex real-world data distributions through learning the high-dimensional latent space of a neural network.

Applicable Scenarios: Complex tables, multidimensional time series, medical images, and visual data for autonomous driving.

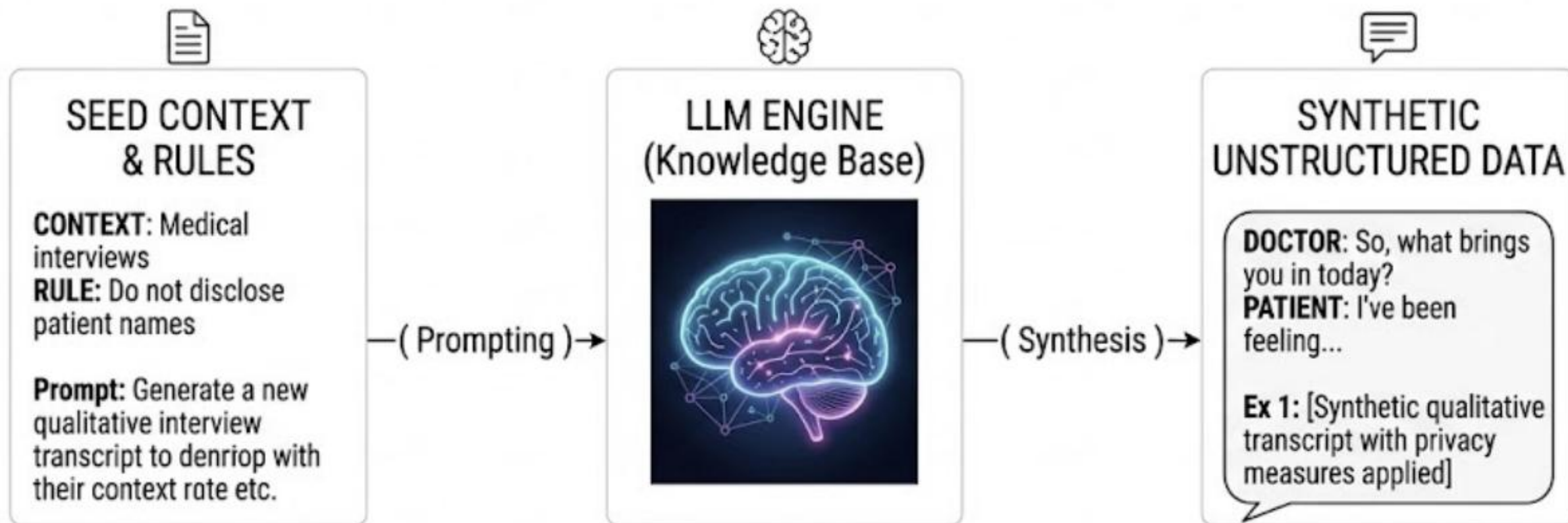
Advantages and Disadvantages: Extremely strong fitting ability, capable of capturing deep and complex relationships; however, it requires significant computational resources and may be susceptible to **mode collapse or overfitting**.

LLM-BASED SYNTHETIC GENERATION

Representative Methods: Prompt Engineering (Few-shot example generation), Agentic Workflows, Self-Instruct.

Working Principle: Leveraging the powerful world knowledge and contextual understanding capabilities of LLM, data is directly "written" based on the Prompt or Agent interaction logic.

SYNTHETIC DATA GENERATION: LLM-BASED APPROACH



FOUR MAIN PARADIGMS OF LLM SYNTHETIC DATA GENERATION

- In context seed generation
- Agentic Persona Simulation
- Structured Schema Generation
- Model Distillation & Synthetic Reasoning Chains

IF YOU KNOW SOME CODING

Python



```
from sdv.single_table import CTGANSynthesizer
from sdv.metadata import SingleTableMetadata

# 1. Extract metadata structure from the raw sensitive dataset
metadata = SingleTableMetadata.load_from_dict(raw_data_dict)

# 2. Initialize the deep generative model (e.g., CTGAN)
model = CTGANSynthesizer(metadata)

# 3. Fit the model on real data (executed inside your secure environment)
model.fit(real_data)

# 4. Generate 10,000 synthetic rows (ready for open sharing)
synthetic_data = model.sample(num_rows=10000)
```

Just 5- 8 rows, bang!

IF YOU DON'T HAVE CODING EXPERIENCE



YData

- Home
- Data Catalog
- Connectors
- Labs
- Synthetic Data
- Pipelines
- Account
- Support

**Hello Fabiana Clemente, welcome to Quickstart!**

 Fabiana Clemente  Not Shared  Add Contributors

Quickstart

Your Recent Activity

DATA CATALOG
Access, profile and manage datasets

**START WITH**
Upload and connect to your datasets through an easily configurable interface.
[+ Add Dataset](#)

**ONBOARD**
Download an example dataset to quickstart with Fabric.
[Download Credit Card dataset](#)
[Download Loan Approval dataset](#)

**EXPLORE**
Learn more about your data with quality warnings and visualizations.

1 Datasets

SYNTHETIC DATA
Data privacy & enrichment

**CREATE**
Train a synthetic data generator from your data catalog datasets.
[+ Create Synthetic Data](#)

**GENERATE**
Sample new data from your synthetic data generator.
[No Synthetic Data, create first](#)

LABS
Familiar code environments

**CODE**
Create data workflows using your favorite packages and YData SDK.
[+ Create Lab](#)

GET HELP

Documentation
[YData Fabric Docs](#)
[YData Academy](#)

SDK
[Token](#)
[SDK Docs](#)

Videos
[Data Catalog](#)
[Labs](#)

IF YOU DON'T HAVE CODING EXPERIENCE

YData

[Home](#)

[Data Catalog](#)

[Connectors](#)

[Labs](#)

[Synthetic Data](#)

[Pipelines](#)

[Account](#)

[Support](#)

Quickstart ▾

1. Select a Dataset (Census)

Search Dataset By Name, Connector or Type

 **Census**
Upload CSV
Tabular
2023-11-28 15:01

2. Select columns to generate and confirm metadata

Column Name	Variable Type	Data Type	Format
☒ Age	INT	Numerical	
☒ Workclass	STRING	Categorical	
☒ fnlwgt	INT	Numerical	
☒ Education	STRING	Categorical	
☒ education.num	INT	Categorical	
☒ Marital status	STRING	Categorical	
☒ Occupation	STRING	Categorical	

Quickstart ▾

☒	Marital status	STRING ▾	Categorical ▾
☒	Occupation	STRING ▾	Categorical ▾
☒	Relationship	STRING ▾	Categorical ▾
☒	Race	STRING ▾	Categorical ▾
☒	Gender	STRING ▾	Categorical ▾

1 - 10 of 15

|< < 1 2 > >|

3. Anonymize Columns ▾

4. Configuration ^

Display Name*

Census - Synth

Target

[None] ▾

Privacy Level*

High Fidelity ▾

☐ Conditional Sampling ⓘ

None ▾

Data Generation

Number of Rows



Sample History



You have not generated data yet.
Generate new data by choosing the number of records.



Home

Data Catalog

Connectors

Labs

Synthetic Data

Pipelines

Account

Support

Quickstart

Synthetic Data > Census - Synth

Use in Labs

Retrain Model



Overview

Generation

Metadata

Api Docs

Status

Ready

Data Source

Census

Dataset Type

Tabular

Created

2023-11-28 17:04

Metadata Summary

Selected Columns

14 out of 15

Target

None

Column's Data Types

Categorical 57%

Numerical 36%

String 7%

Privacy Level

HIGH_FIDELITY

Conditional Sampling

None

Status Details

Training process is being prepared

Synthetic data is training

Ready

Synthetic data is ready to generate.

Go to Generation

Check out the Quality Report

Fidelity Score



Utility Score



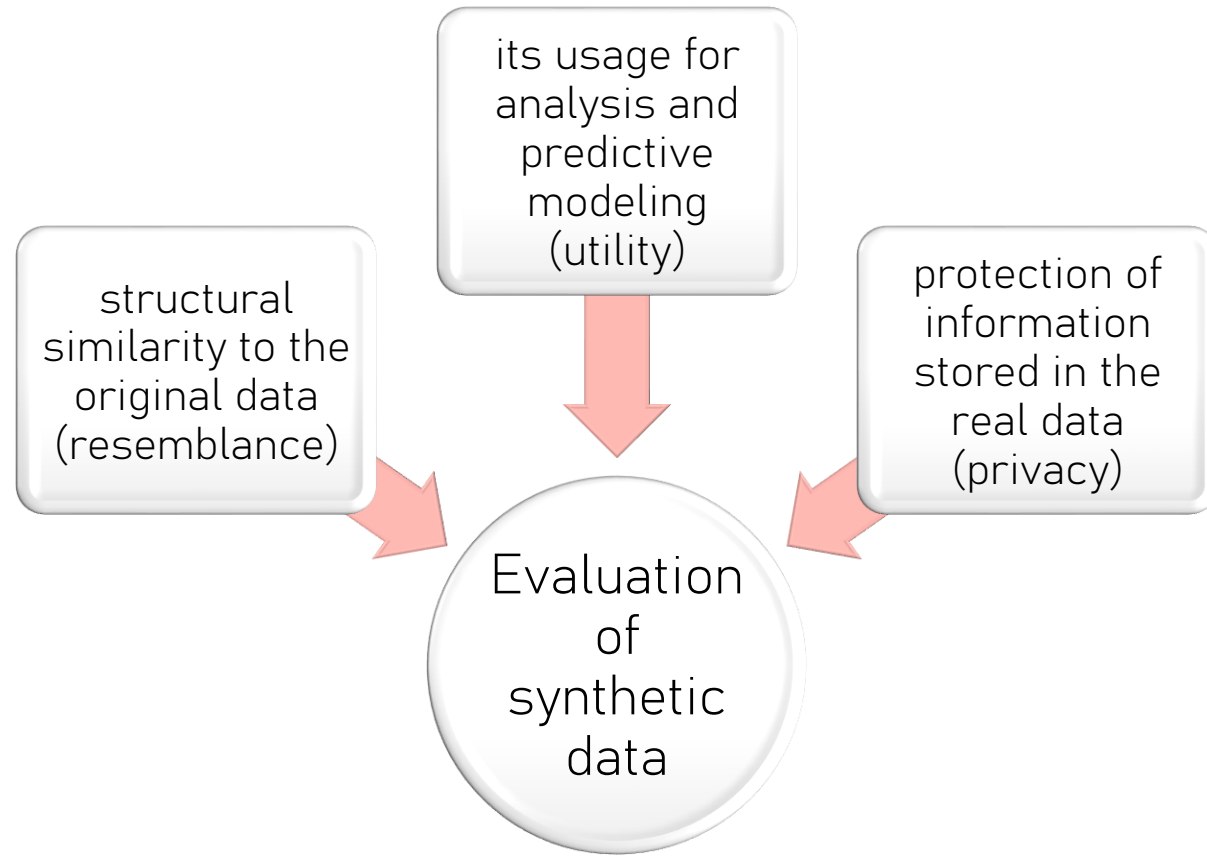
Privacy Score

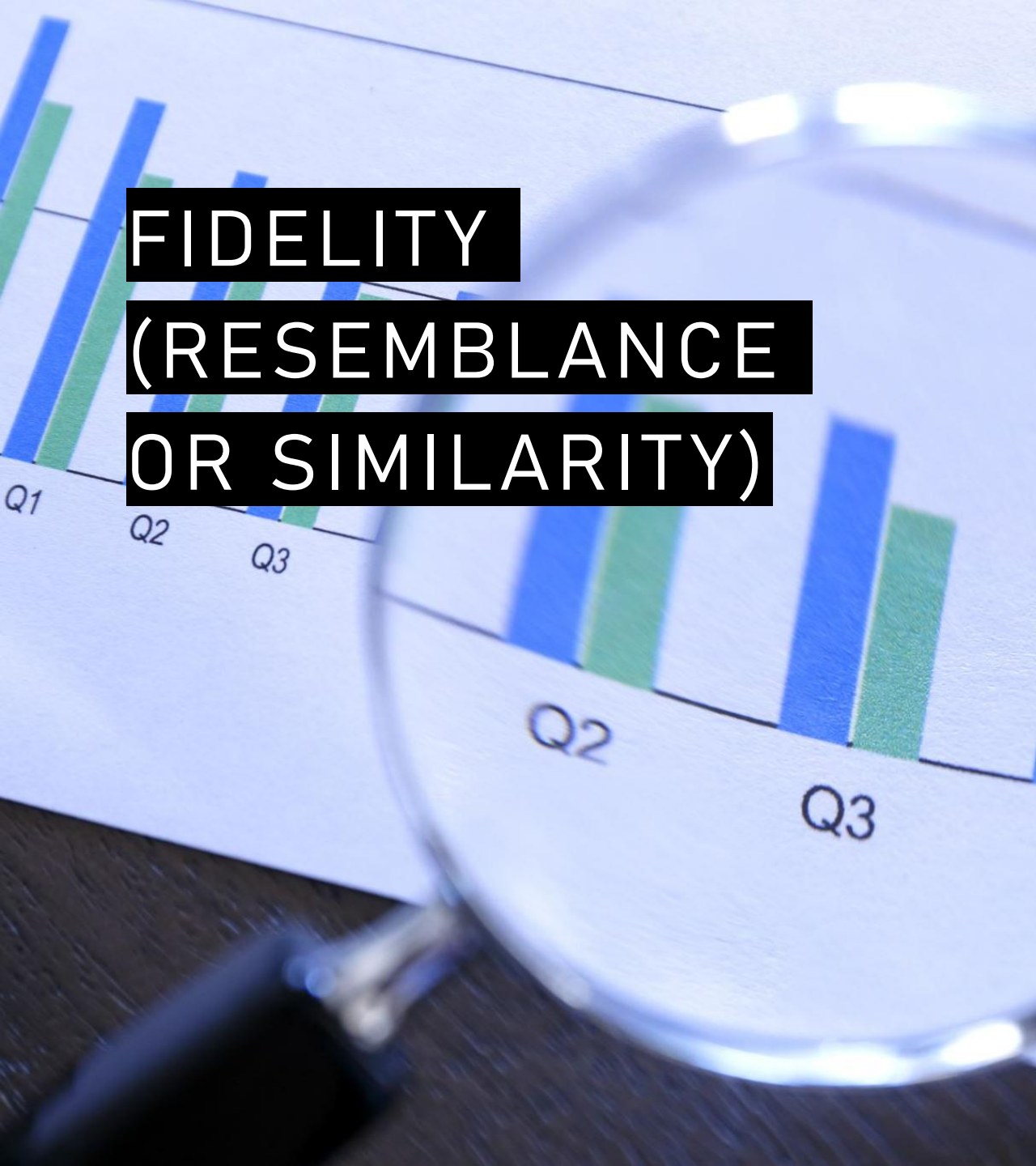


A report showing the synthetic data quality metrics of the dataset based on the defined metadata.

Download Quality Report

SYNTHETIC DATA EVALUATION

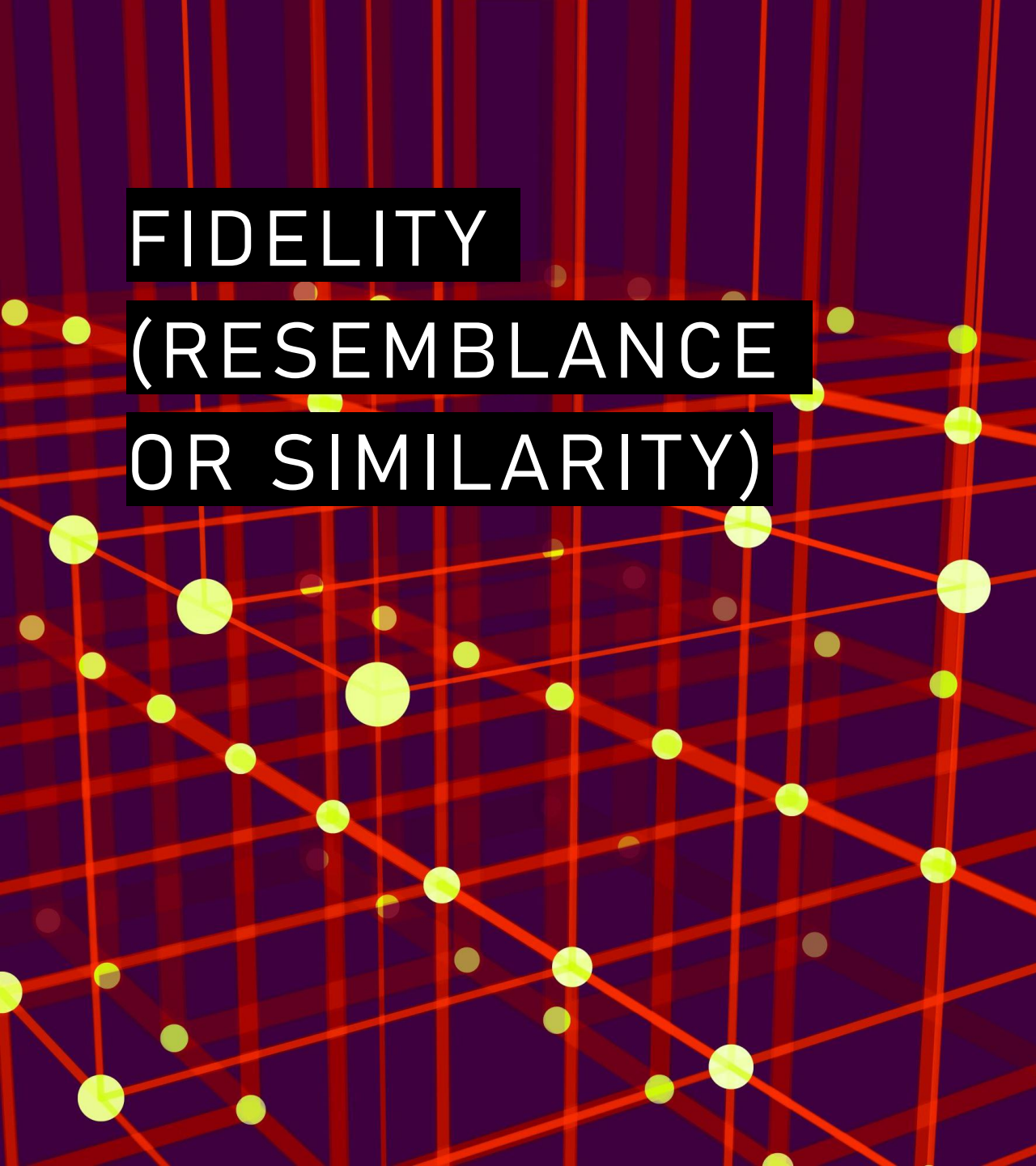




FIDELITY

(RESEMBLANCE OR SIMILARITY)

- The core question: "Does the synthetic data resemble real data in appearance and distribution?"
- Tabular Data: Examine the probability distribution of univariate data (e.g., KS-Test/Kolmogorov-Smirnov test) and the correlation heatmap of multivariate data.
- Image Data: Examine the FID (Fréchet Inception Distance). This is the gold standard for image synthesis; the lower the FID score, the closer the feature distribution of the synthetic image is to that of a real photograph.

The background of the slide features a dark purple field with a network of thin, bright red lines. Scattered throughout this network are numerous small, glowing yellow circles of varying sizes, some of which are connected by the red lines, creating a complex, web-like pattern.

FIDELITY (RESEMBLANCE OR SIMILARITY)

- Text Data: Examine the perplexity and the degree of overlap of the embeddings (semantic vectors) in high-dimensional space.
- Time-Series Data: Examine the autocorrelation function and dynamic time warping (DTW).

UTILITY

- The core question is: "If I use this fake data to run analysis/train the model, will the conclusions drawn still hold true in the real world?"
- Evaluation Metrics:
 - AUC / R^2 Gap: Is performance loss minimal?
 - Inference Alignment: Are p -values & coefficients consistent?

UTILITY

- [Benchmark: TRTR]
- Real Training Data $\longrightarrow$ Train ML/Stats Model $\longrightarrow$ Test on Real Data $\longrightarrow$ Score A (e.g., AUC = 0.88)
- [Utility Test: TSTR]
- Synthetic Data $\longrightarrow$ Train ML/Stats Model $\longrightarrow$ Test on Real Data $\longrightarrow$ Score B (e.g., AUC = 0.86)

PRIVACY

Distance-based privacy:
"Did the model
'memorize' a specific
real subject, leading to a
privacy leak?"

Attack-based privacy: Is
the synthetic data model
or released synthetic
data robust? Let's do
some attacks and try
them out

SPECIFIC METRICS

- DCR (Distance to Closest Record): Measures how far each synthetic data point is from its nearest original real data point. A distance of 0 indicates that the model is "copying" the original data, posing a serious vulnerability.
- MIA (Membership Inference Attack): Simulates an attacker attempting to infer whether a specific patient is present in the original dataset using synthetic data.

THE EVALUATION TRIANGLE

There can be more: such as fairness!

Can Synthetic Data be Fair and Private? A Comparative Study of Synthetic Data Generation and Fairness Algorithms

Liu et al., 2025

[1. Fidelity]
(Statistical Similarity)



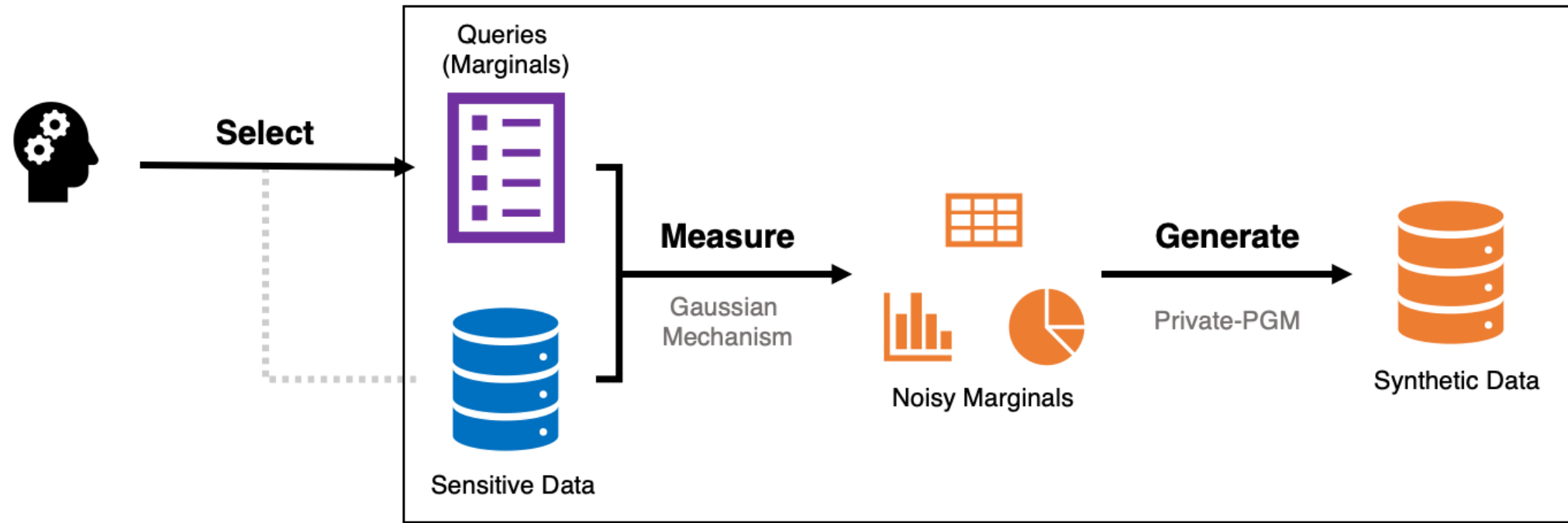
[2. Utility]
(TSTR / Downstream Task)

[3. Privacy]
(DCR / Memorization)

SPEAKING OF PRIVACY IN SYNTHETIC DATA

What if AI-generated data accidentally leaks information about a patient?

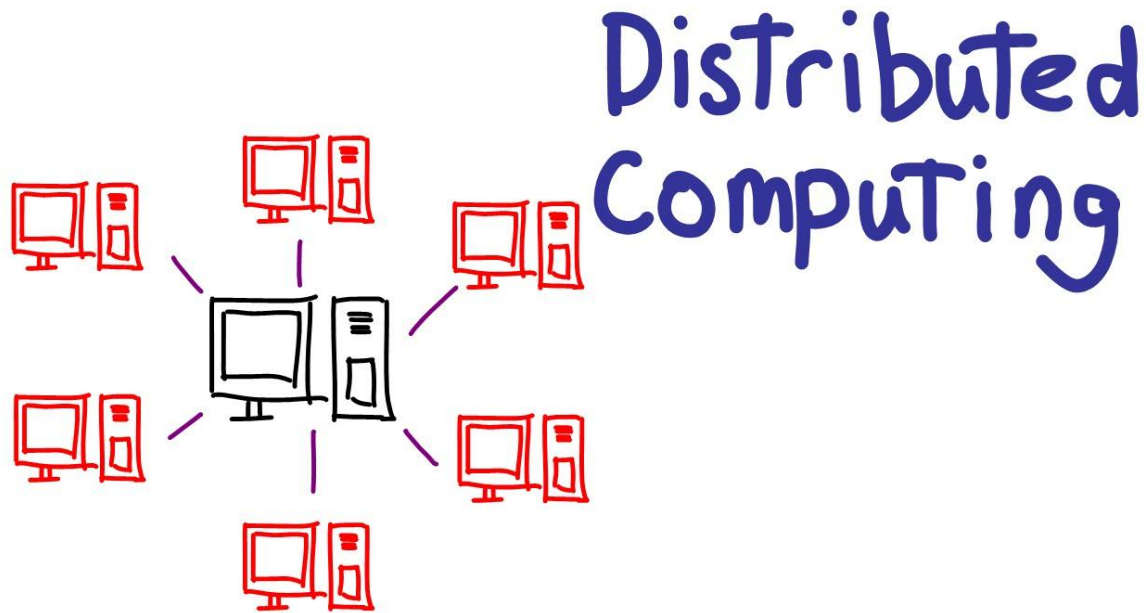
What if the data is scattered across 5 different hospitals/institutions, and none of them can send the data to the others?



DIFFERENTIAL PRIVATE SYNTHETIC DATA

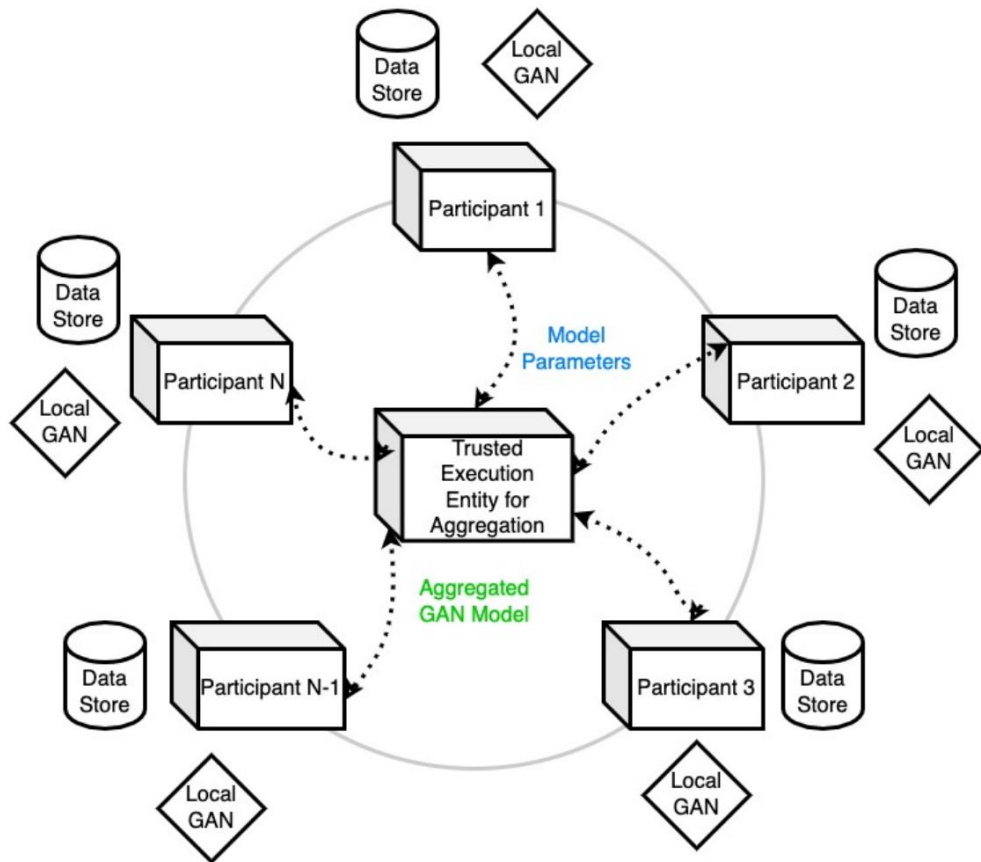
- [Pic sources](#)
- By latest AI attacks, with only synthetic data alone, it's not safe

FEDERATED LEARNING + SYNTHETIC DATA



- The core challenge: Data is scattered across three hospitals in Oslo, USA, and London. Due to GDPR restrictions, none of the three hospitals can aggregate the raw data onto a central server.
- In short: "Bring the model to the data, not the data to the model."

FEDERATED LEARNING + SYNTHETIC DATA



- The three hospitals do not share a single line of raw data. We train the AI generation models locally separately and upload only the model's 'weight parameters' to a central server for aggregation.
- Finally, we use this 'joint society's AI' to generate a global synthetic data set (Global Synthetic Dataset) that combines the strengths of each company. "

ASPIRE PROJECT IN UIB

- Accelerating Privacy and Data Protection Measures Using Synthetic Data Generation (ASPIRE)
- ASPIRE aims to develop novel synthetic data generation methods that robustly ensure privacy and data protection using two innovative techniques: Differential Privacy and Federated Learning.
- For education data and health data both

Project Period:

January 2025 – December 2028.

Funded By:

Equinor ASA.

Project Leader:

Mohammad Khalil.

Project Members:

- Prospect PhD Candidate (currently in the process of hiring)
- Qinyi Liu
- Farhad Vadiee.

Project Partners:

To be announced.

ILLUSTRATIVE EXAMPLE: A PRIVACY-CONSCIOUS SHARING WORKFLOW

[1. THE DILEMMA]

- 5,000 sensitive patient records (EHR)
- Locked under strict GDPR & TSD security rules
- Direct cross-border sharing was strictly impossible

[3. THE OPEN SCIENCE IMPACT]

- Public synthetic dataset released on Zenodo
- Global collaborators built & debugged algorithms without ever accessing raw patient rows

[2. THE SYNTHETIC PIPELINE]

- Trained CTGAN generator inside secure vault
- Outputted 10,000 privacy-safe synthetic rows
- Applied strict Differential Privacy (DP) checks

[4. THE VALIDATION RESULT]

- High Utility: TSTR Model AUC 0.85 (vs Real 0.87)
- Strong Privacy: Zero leaks (DCR safety score)
- Limitation: Rare gene interactions were smoothed

LIMITATIONS OF SYNTHETIC DATA



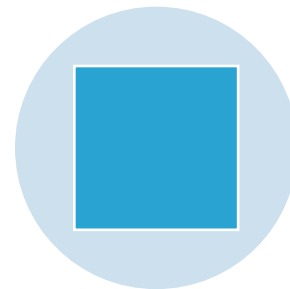
Not Automatically
Anonymous –
overfitting problem



Garbage In, Garbage
Out & Amplified Biases



Information Loss & The
"Ground Truth" Limit



Hallucination – LLM

TAKE AWAY: PRACTICAL RULE FOR SYNTHETIC DATA IN OPEN SCIENCE

1. Before you generate	2. Before you share	Before you reuse
Define the target modality & complexity before picking a model	Run the evaluation test (fidelity, utility, privacy)	Use as a "sandbox": develop code & test hypotheses safely
Don't over-engineer: start with simple models	See if more audits are needed	Check the performance reports (if there are)
Document the generation parameters transparently for reproducibility	Report and limitations: rare samples loss	

USEFUL LINKS

- Common synthetic data generation packages (if you have coding experience):
 - <https://github.com/statice/awesome-synthetic-data>
 - <https://github.com/ql909/awesome-synthetic-data-education>
 - <https://github.com/Data-Centric-AI-Community/fg-data-synthetic>
 - <https://github.com/vanderschaarlab/synthcity>
 - <https://github.com/sdv-dev/sdv>
- Common synthetic data generation dashboard (if you don't know how to coding)
 - https://docs.synthetic.ydata.ai/latest/ydata_fabric_app/
- Mostly.ai

QUESTIONS?

- Contact me at qyliu@cityu.edu.mo
- Slides will be upload to my website (<https://qinyiliu.site/>), feel free to download or send request