DIVIDI ET IMPERA: ENHANCING SYNTHETIC DATA FIDELITY THRU DATA PARTITIONING

Yordanos Nebiyou Yifru

Salvatore Distefano

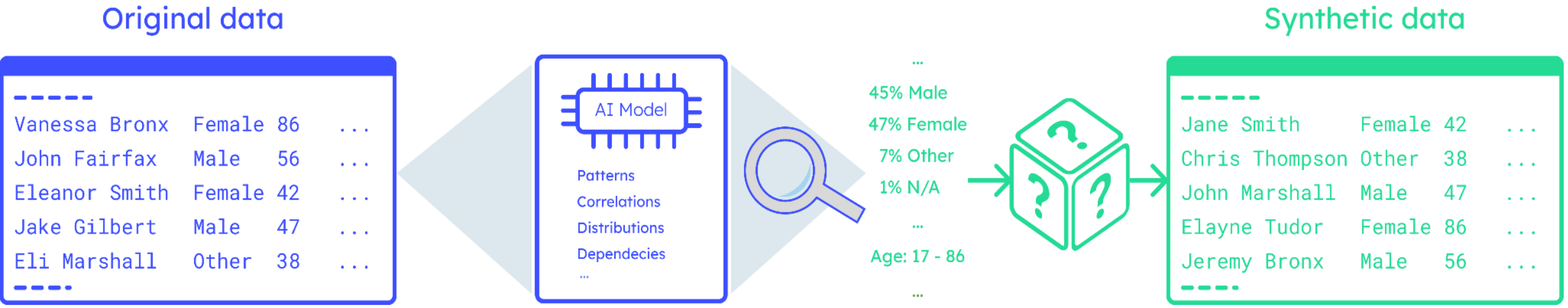
University of Messina

yordanos.yifru@studenti.unime.it
sdistefano@unime.it

ItaData 2025 - The 4th Italian Conference on Big Data and Data Science - 9-11 September 2025, Turin, Italy

SYNTHETIC DATA

Synthetic data *are generated by an on purpose-built mathematical model or algorithm to support one or more data science tasks.*



THE NEED FOR SYNTHETIC DATA

Data Quality:
Low-quality or noisy datasets can mislead models and lead to inaccurate predictions.

Data Scarcity:
Many applications—especially involving rare events, new domains, or proprietary systems—have insufficient or unavailable real data.

Data Privacy:
Legal, ethical, and technical constraints (e.g., GDPR, HIPAA) often prohibit open sharing or use of real data.

Data Fairness:
Historical or sampling biases in datasets can cause unfair or discriminatory results.

METHODS OF SYNTHETIC DATA GENERATION

Generative AI Models
(GANs, VAEs, Transformers)

Statistical Modeling
(Copulas, Bayesian Networks)

Hybrid Models
(Combination of Generative and Statistical Methods)

Rule-Based Systems

LIMITATIONS OF CURRENT METHODS

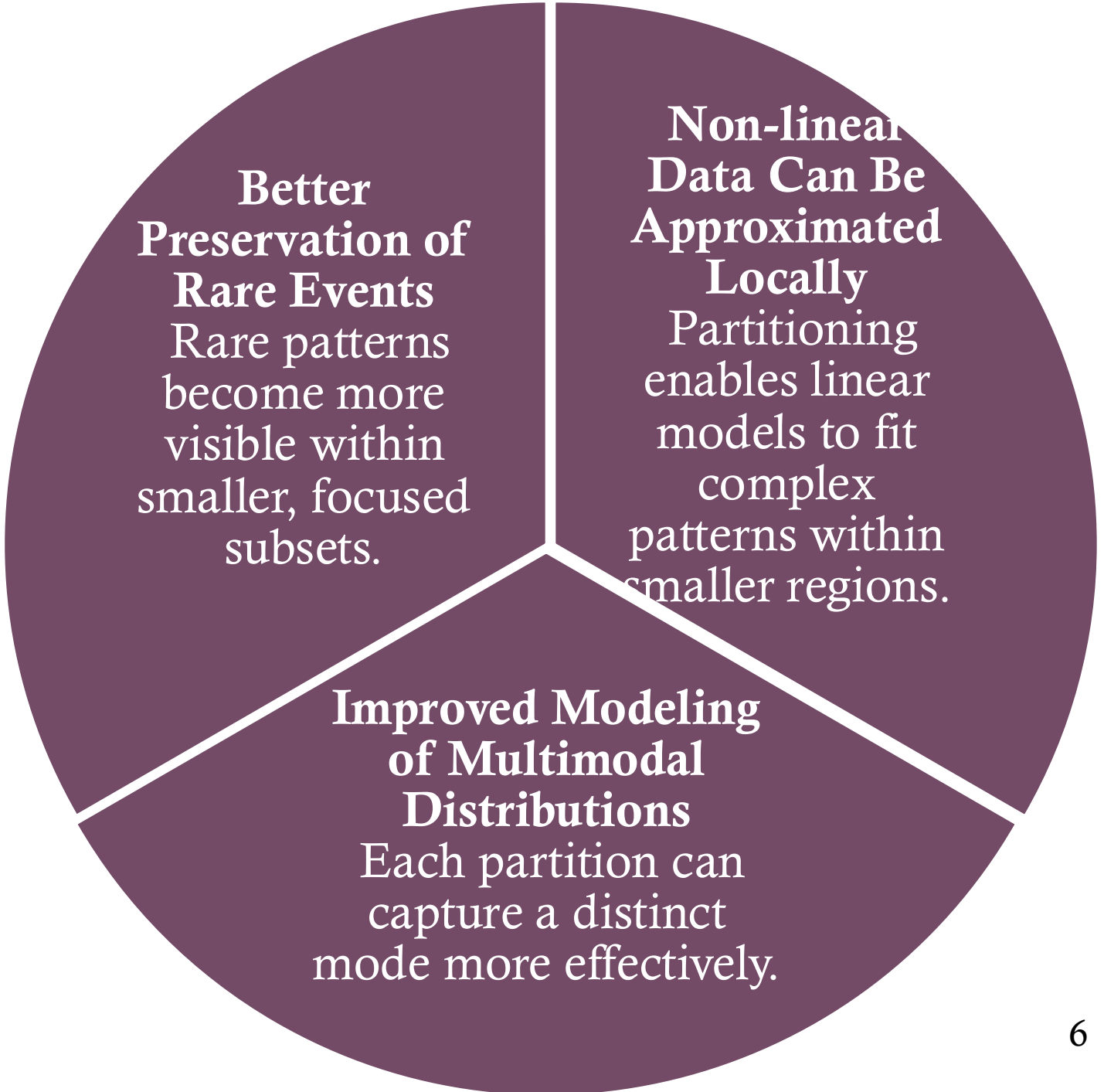
Assumption of Linear Dependencies
e.g., *Gaussian Copula*

Assumption of Unimodal Distributions
e.g., *TVAE*

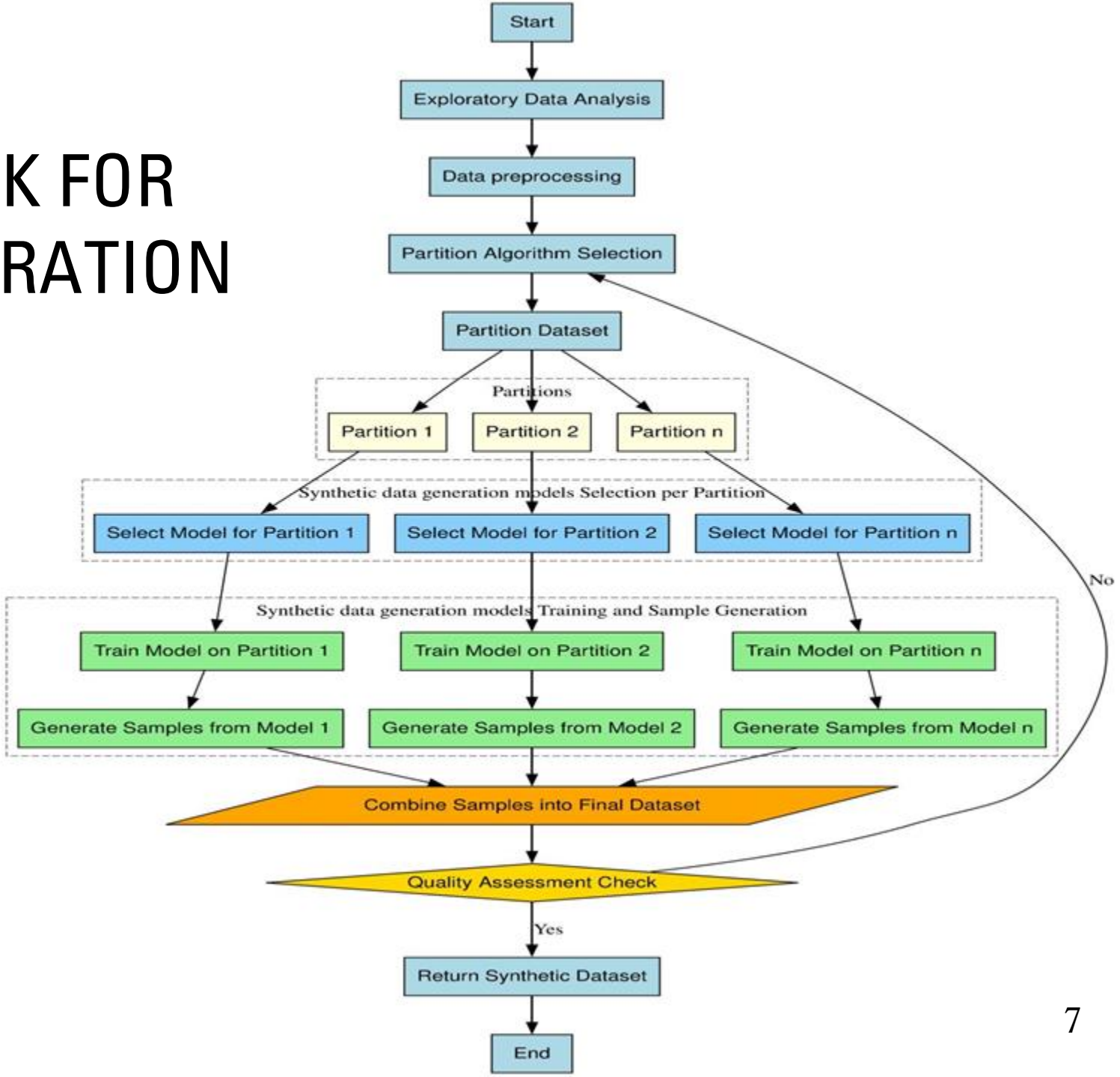
Poor Handling of Rare Events
Difficulty in accurately modeling low-frequency patterns or outliers

DATA PARTITIONING IS THE KEY!

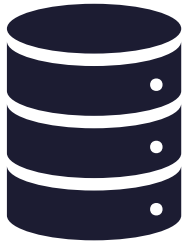
- Divide et impera
- Euclidean and non Euclidean data
- Piecewise linearization/grouping/categorization
- Fine grained
- Large dataset



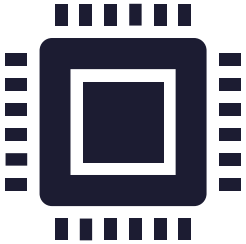
PROPOSED FRAMEWORK FOR SYNTHETIC DATA GENERATION



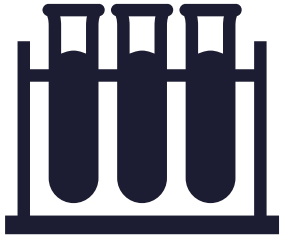
CASE STUDY: TABULAR DATA



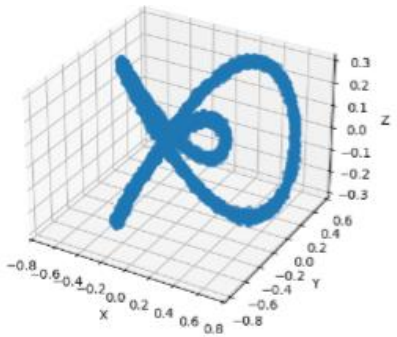
Partitioning Strategy:
Hierarchical Binary Splitting Algorithm decompose the data into subsets.



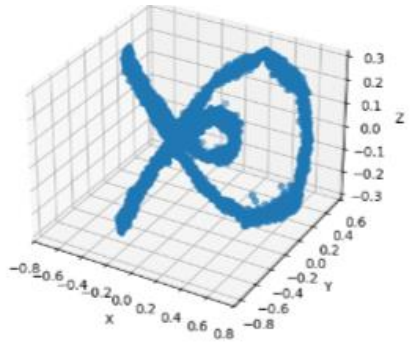
Generation Model:
On each partition the *Gaussian Copula* is adopted to generate synthetic data.



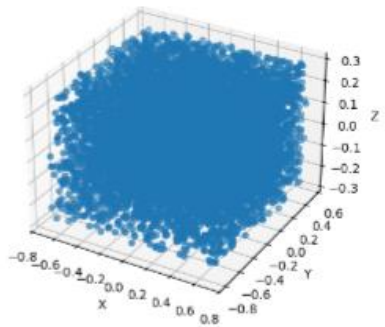
Evaluation:
Tests on both *real-world tabular data* and a *geometric dataset*.



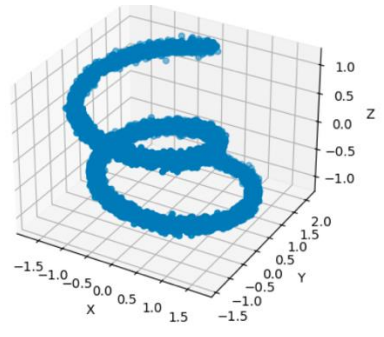
(a) Original



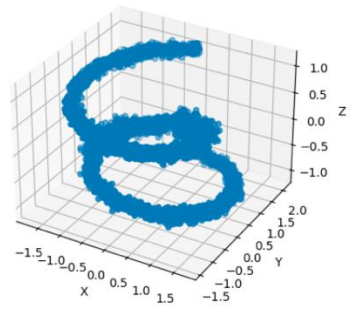
(b) PGC



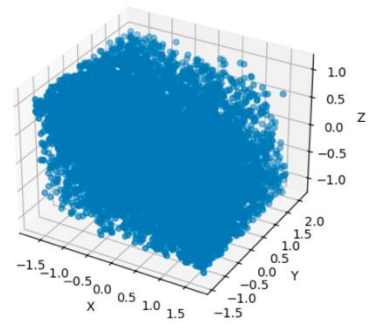
(c) GGC



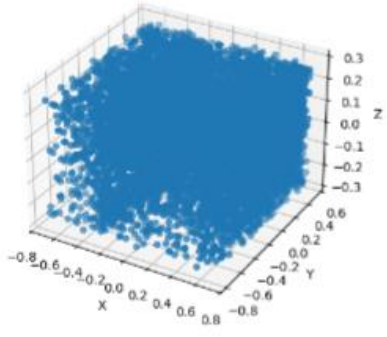
(a) Original



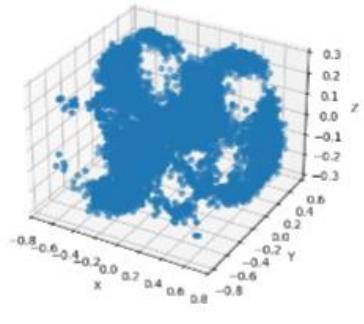
(b) PGC



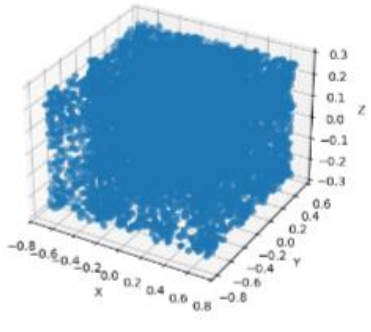
(c) GGC



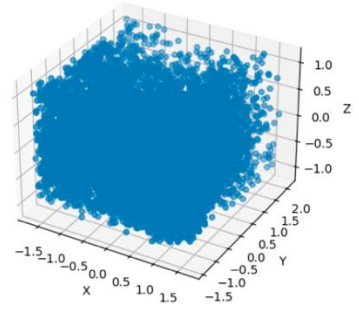
(d) CTGAN



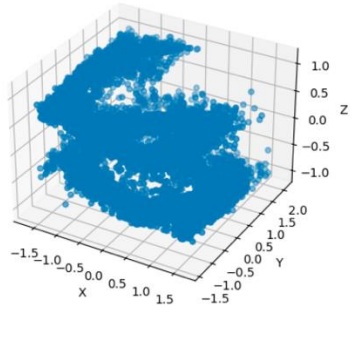
(e) TVAE



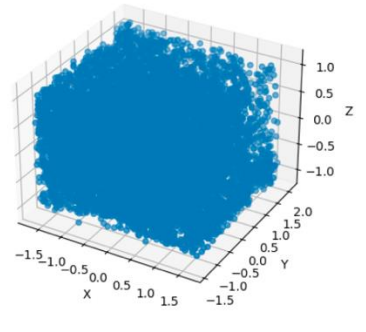
(f) CopulaGAN



(d) CTGAN



(e) TVAE



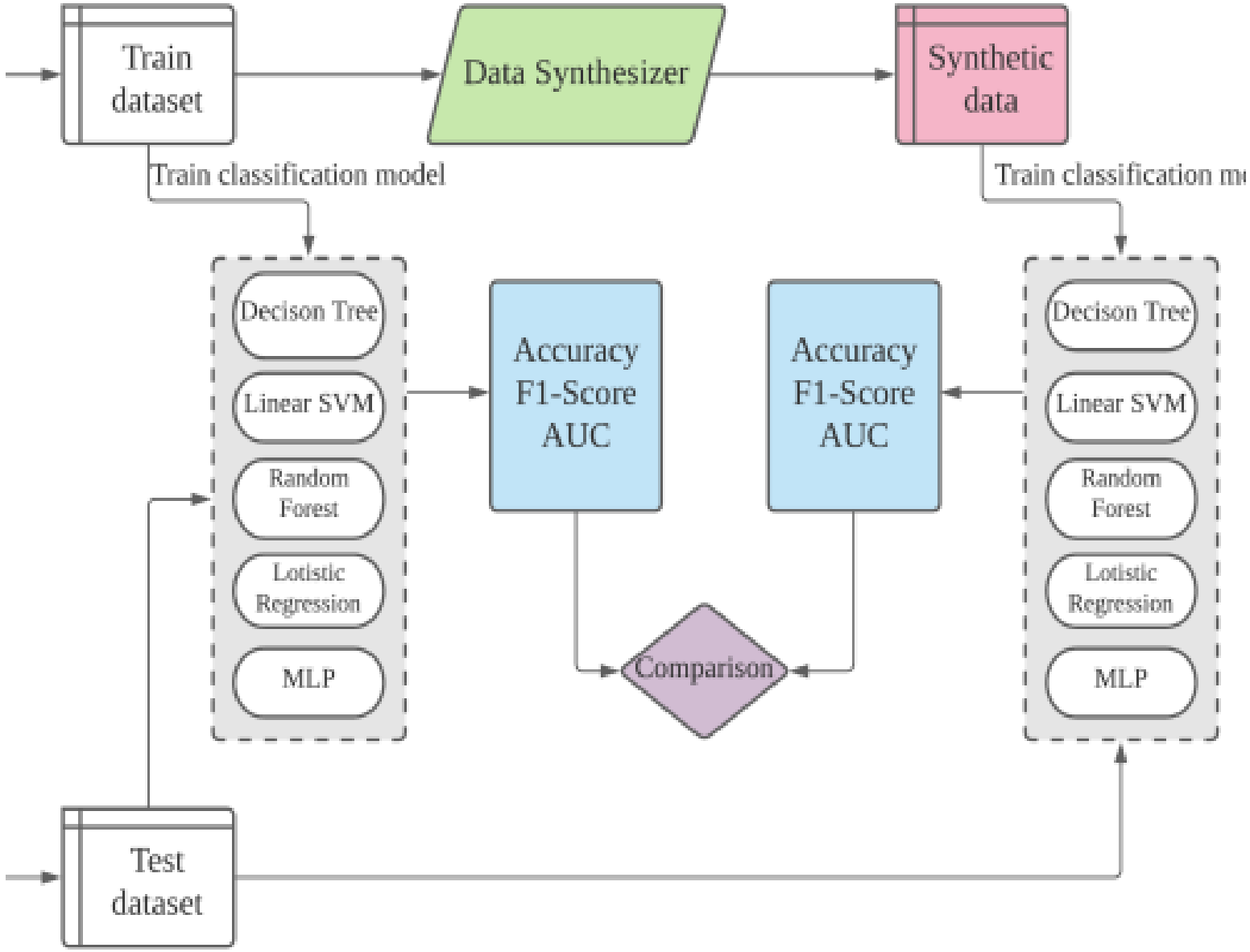
(f) CopulaGAN

GEOMETRIC DATASET EXPERIMENTS

REAL DS EXPERIMENTS

- **4 Different Real Datasets**
- **Evaluation Metrics**

1. **Statistical Similarity**
   Jensen-Shannon Divergence (JSD), Wasserstein Distance (WD), Correlation Difference
2. **Privacy**
   Distance to Closest Record (DCR), Nearest Neighbour Distance Ratio (NNDR)
3. **Machine Learning Utility**
   Performance of models trained on synthetic data compared to real data



Model	Average WD	Average JSD	Correlation Distance
Gaussian Copula with partitioning	**0.0088**	**0.0249**	**1.0137**
CTGAN	0.0631	0.1024	3.8806
Gaussian Copula without partitioning	0.0918	0.0923	2.4465
CopulaGAN	0.0870	0.1057	4.1020
TVAE	0.0462	0.0875	1.9061

RESULTS: STATISTICAL FIDELITY

- The **Gaussian Copula with Partitioning** achieved the **best performance across all statistical metrics**.
- This indicates that **partitioning preserves both marginal distributions and correlation structure** more effectively than non-partitioned models.

Model	DCR	NNDR
Gaussian Copula with partitioning	1.3620	0.6924
CTGAN	1.5795	0.7529
Gaussian Copula without partitioning	**1.6256**	0.7254
CopulaGAN	1.4146	0.7592
TVAE	1.4883	**0.7936**

RESULTS: PRIVACY

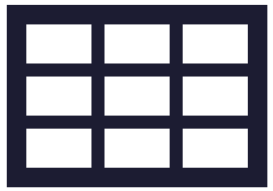
- The **Gaussian Copula with Partitioning** generally exhibits **lower privacy** than the version without partitioning.
- **Partitioning reduces the size of training subsets**, increasing the risk that **synthetic records resemble real ones** too closely.

RESULTS: ML UTILITY

- The **Gaussian Copula with Partitioning** outperformed the standard Gaussian Copula (without partitioning).
- It achieved an average **F1-score gain of approximately +43 percentage points (pp)**.
- This represents a **substantial improvement** in the quality of generated data.

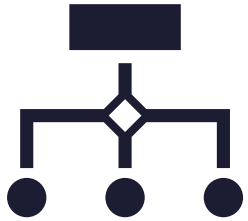
Model	Accuracy (%)	F1 Score	AUC
Gaussian Copula with partitioning	**2.96**	**0.0320**	**0.0180**
CTGAN	6.39	0.1118	0.0800
Gaussian Copula without partitioning	31.19	0.4600	0.1690
CopulaGAN	7.00	0.1600	0.0350
TVAE	11.80	0.1500	0.0360

CONCLUSION



Partitioning improves the quality of synthetic data, but …

- Granularity: fine-grained, more homogeneous subsets **enhances modeling accuracy**
- **Reduce privacy** due to increased similarity with real records.



Outlook for experiments on non tabular (multimedia, non structured data)

- Assessing different partitioning algorithms (SAFE metrics, …)
- GAN, Diffusion model, TVAE, …
- New partitioning techniques (e.g. based on semantic dimensions

THAT'S ALL FOLKS!

Many thanks!

Questions?