UNIVERSITY of GREENWICH

Extracting Truth From Fiction:

Synthetic Data for Anti Money Laundering and FinCrime: The Future and Significance of "True Data" in FinCrime model training.

Prof. George Samakovitis
Professor of FinTech
School of Computing & Mathematical Sciences
University of Greenwich, UK

Background to the problem: what is it?

What is Synthetic Data?

- Artificially generated data
  - Typically generated through algorithms with the intention of standing-in for real data.
- Aiming to emulate behaviour of original without revealing attributes of
  - Data
  - Models used to generate

Why is it useful?

- Simulate scenarios not otherwise attainable
- Simulate multiple / parallel scenarios at scale
- Much richer / extensive datasets

What are the possible alternatives?

- Anonymization/pseudo-anonymization
  - Masking
  - Noise addition
- Aggregation
  - Joining datasets (summary data from sources)
  - High risk of deanonymisation
- Federated learning (FL)
  - ‘Data-agnostic’ collaborative training
  - *‘Model inference attacks’* as core risk
- Homomorphic encryption
  - Data leakage / side channel attacks risk

Privacy

Expand Access

Testing & QA

Scenario Modelling

ML Model Augmentation

Background to the problem: **why** do we need it?

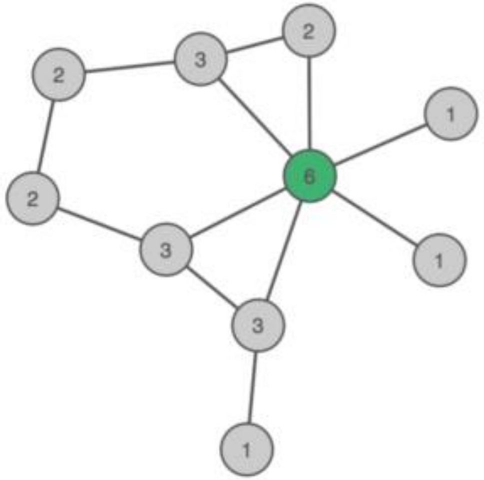


Fig 1: Transaction visibility limitation

The limited line of sight of any individual data custodian allows it to perform KD solely on intelligence gathered from within its scope…

…financial crime activity is performed *across* a transactions network, involving multistage / multiparty / multi-asset transfers.

…***Confidential Data Pooling*** *approaches have proved ineffective on account of data sharing regulation*

The *Collective Intelligence conundrum*

'the inability to share transaction data for knowledge discovery across networks, on account of PII protection and privacy constraints'

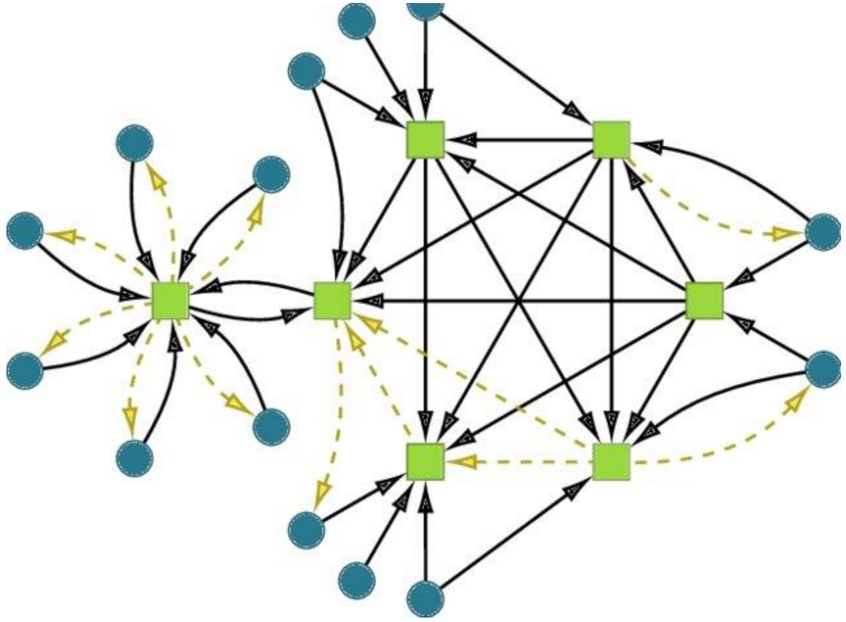


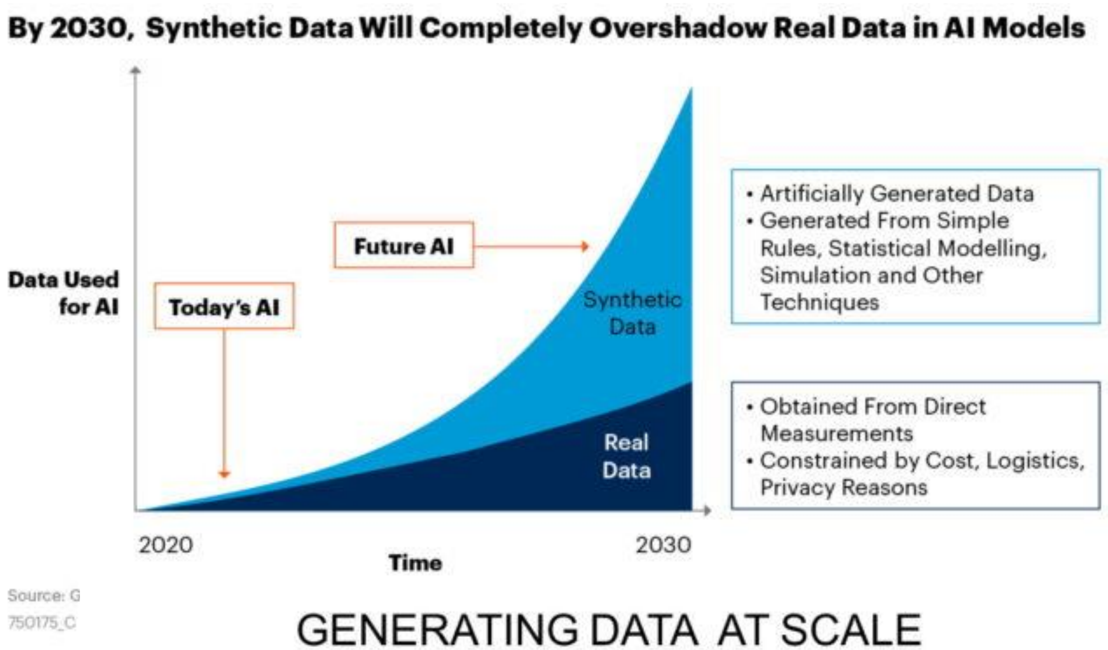
Fig 2: Transaction flow in ecosystem

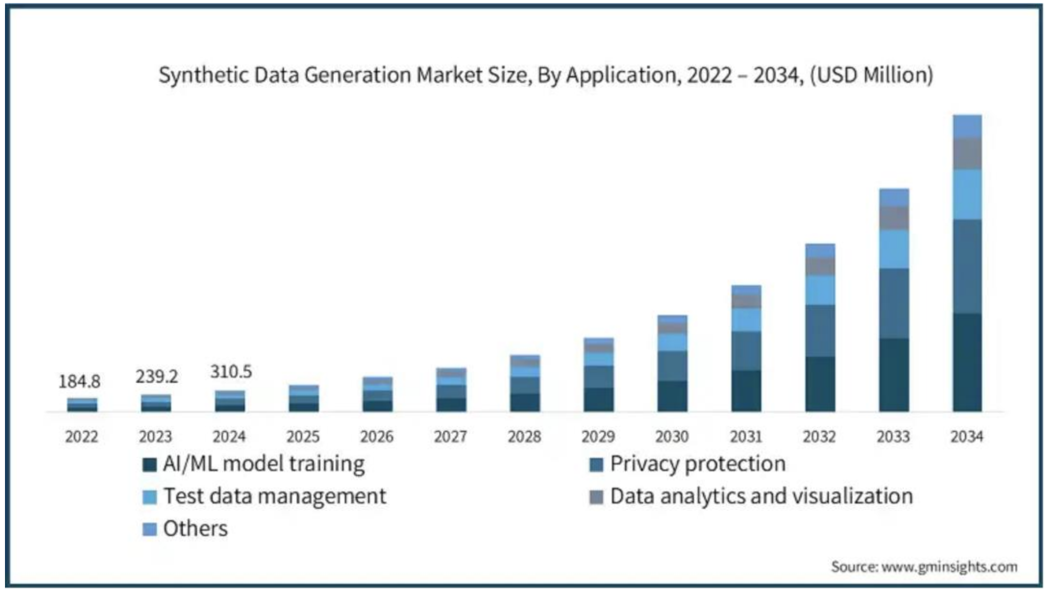
UNIVERSITY of GREENWICH

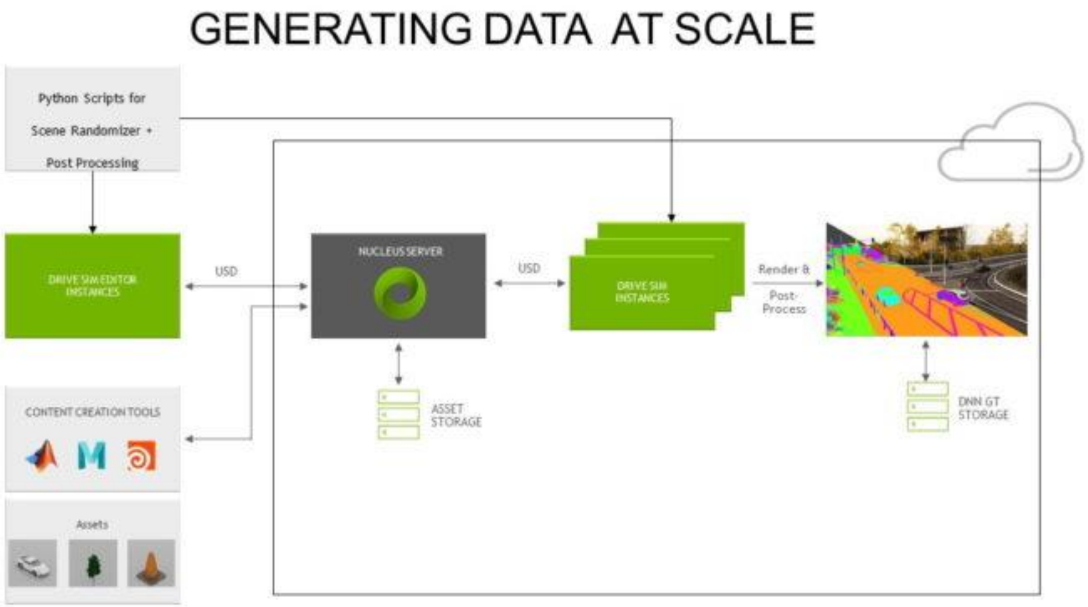
Problems?

Scaling synthetic data use

- Early indications project extensive use;
  - GenAI - enhanced fidelity
- Challenges
  - *Curse of recursion (Shumailov et al, 2023)*
    related solutions? (Gerstgrasser et al, 2024)
  - 'Data Laundering'
  - GenAI governance models for *data provenance*



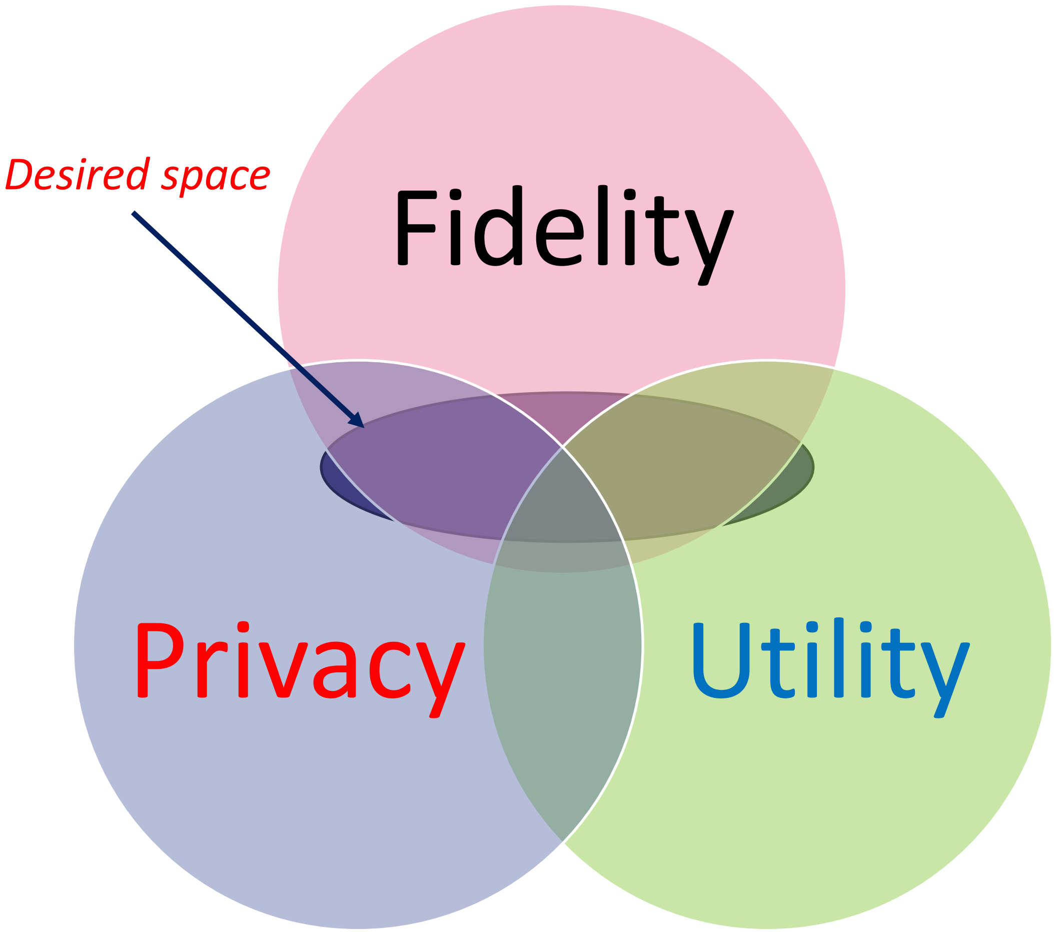




Source: https://www.gminsights.com/industry-analysis/synthetic-data-generation-market

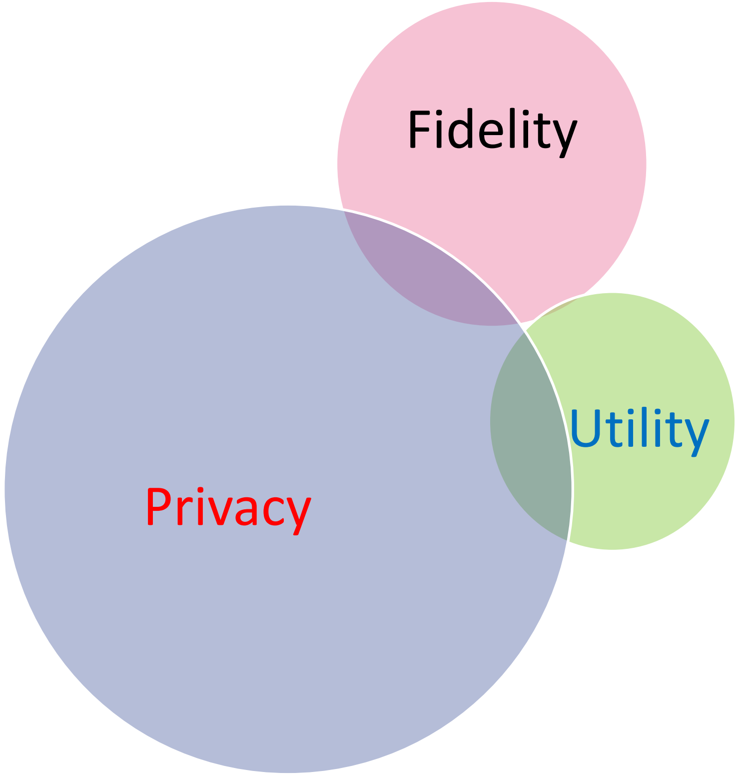
Source: NVIDIA: https://blogs.nvidia.com/blog/what-is-synthetic-data/

Synthetic Data Tradeoffs: budgets on privacy / fidelity



Synthetic Data Tradeoffs: budgets on privacy / fidelity

High privacy budget

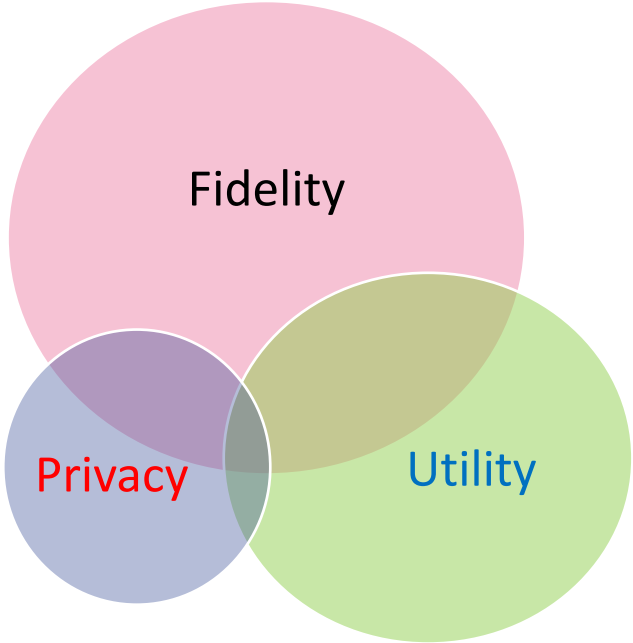


Budgeting as a risk-driven strategy

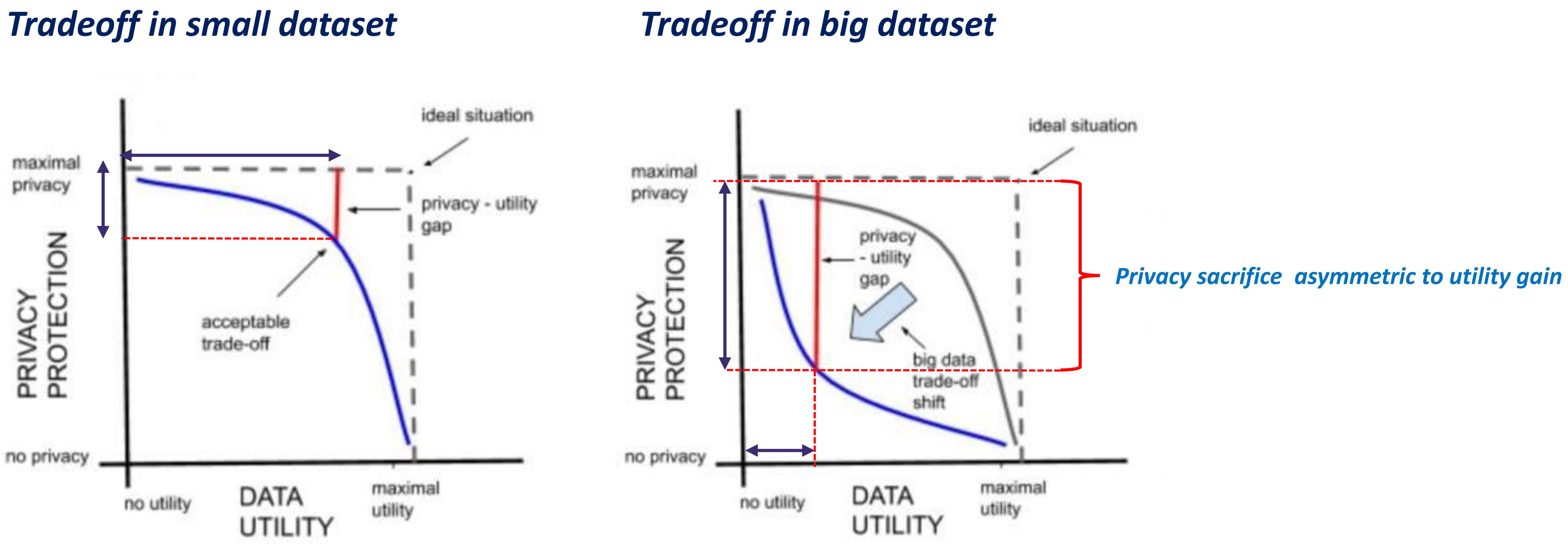
- Data quality correlated to *Fidelity*
  - Statistical behaviour accuracy
  - Often associated with ‘veracity’ of datapoints
- *Fidelity* often inversely related to *Privacy*
  - ‘Noisy’ datasets harder to infer
  - Unknown which statistical behaviours are important (hard to optimise by selection)
- *Privacy* inversely related to *Utility*
  - ‘Noisy’ datasets hard to carry rare events (e.g. fraud / ML)
  - Fully synthetic datasets (ABM) hard to assess for utility

Risk of mistraining Machine Learning models

Low privacy budget



Privacy – Utility tradeoffs



Adapted from: https://mostly.ai/blog/only-a-little-bit-re-identifiable

UNIVERSITY of GREENWICH

Approaches to Synthetic Data

Agent-based Modelling (ABM)

Pros

- Generates complete picture
- Granular control
- Diminished data privacy concerns (?)

Cons

- Computationally expensive
- Model attack concerns
- Requires complex behaviours to be pre-defined (extensive domain knowledge)

Machine Learning (GAN; DP-GANs; P-PGM; VAE…)

Pros

- Can capture complex interactions
- Effective for high dimensionality
- Model temporal relationships

Cons

- Can be computationally expensive
- Privacy risks (overfitting)

Statistical Methods

Pros

- Established
- Interpretable

Cons

- Limited utility
- Privacy risks

Open Source

SMART DATA FOUNDRY

k2view

MOSTLY·AI

gretel™

SDV The Synthetic Data Vault

Altreva™ real-world forecasts from virtual markets

hazy

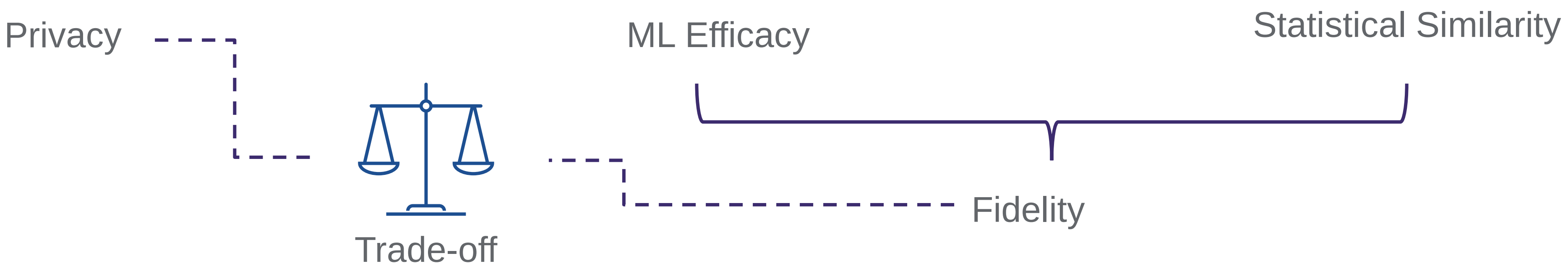
Comparative Analysis of Synthetic Data Generation Techniques

Technique	Realistic	Computational Cost	Privacy Protection	Suitability for Fraud Detection
Rule-Based	Low	Low	High	Limited
Statistical Sampling	Medium	Low	Medium	Moderate
GANs	High	High	Low	Strong
VAEs	High	High	Medium	Strong
Agent-Based	High	High	Medium	Strong
Differential Privacy	Medium	Medium	High	Moderate

Tradeoffs in “strong” candidate models are mainly:

- Adaptiveness to new cases (*cf. fraud et al*)
- Computational cost
- Expert manual intervention

Assessment Criteria: What's important depends on the use case…



Privacy

- Correct Attribution Probability
- Identical or matched rows
- Inference attacks (data / model)
- Protection of rare categories or outliers
- Comparison of training to holdout
- Differential privacy
- Overfitting

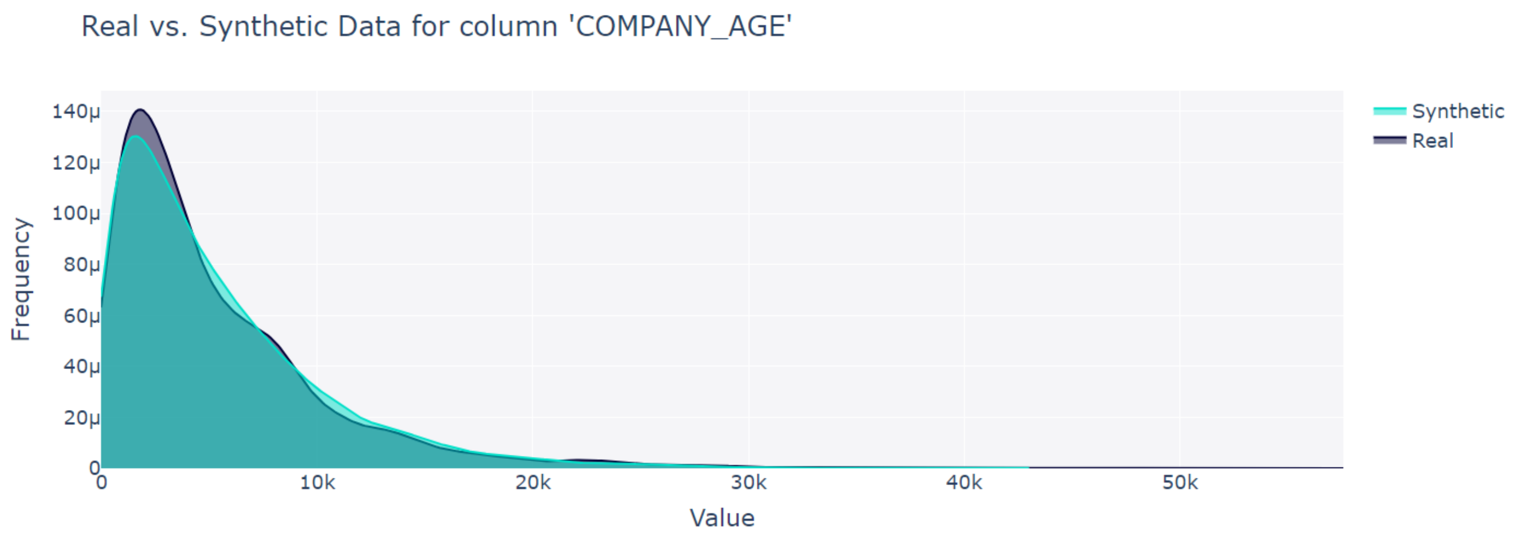
ML Efficacy

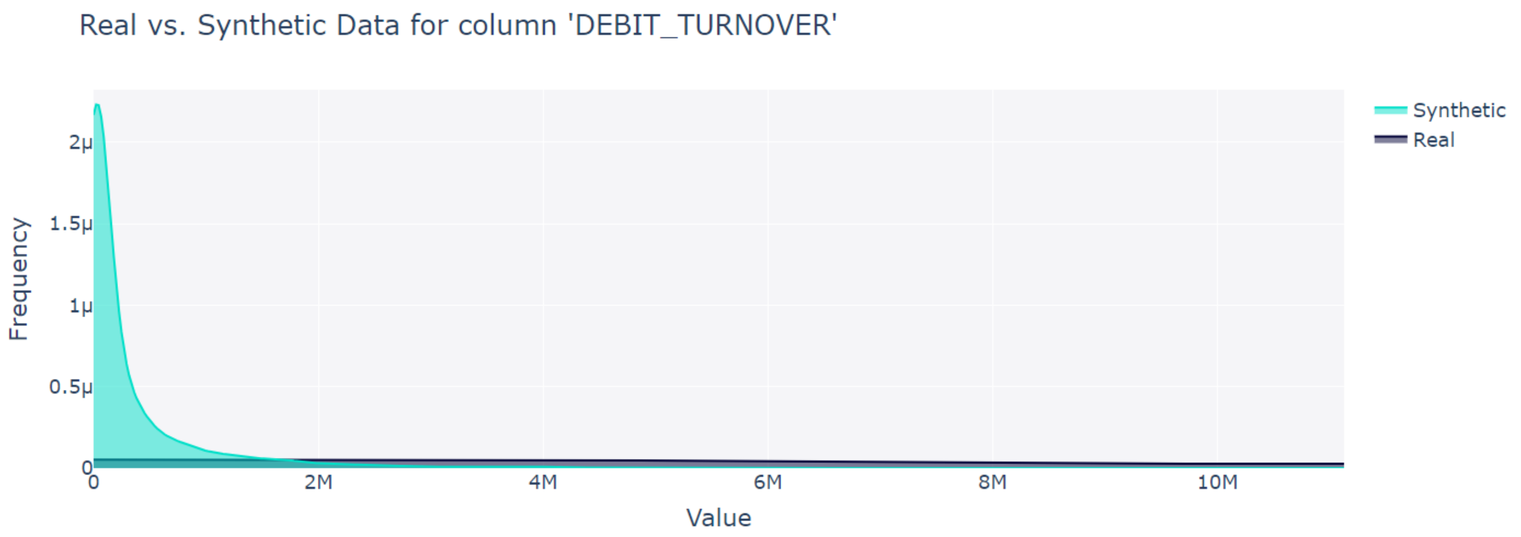
- Cross-evaluate models built using synthetic data on real/holdout data and vice versa
- Compare existing models to synthetic data models and quantify the difference
- Test discriminators

Statistical Similarity

- Basic statistics (mean, median, SD)
- Correlation similarity
- Categorical and range coverage
- KS-Complement (Kolmogorov-Smirnov statistic)
- TV-Complement (based on Total Variation Difference)
- Aggregations or derived metric similarity

Comparison based on statistical similarity





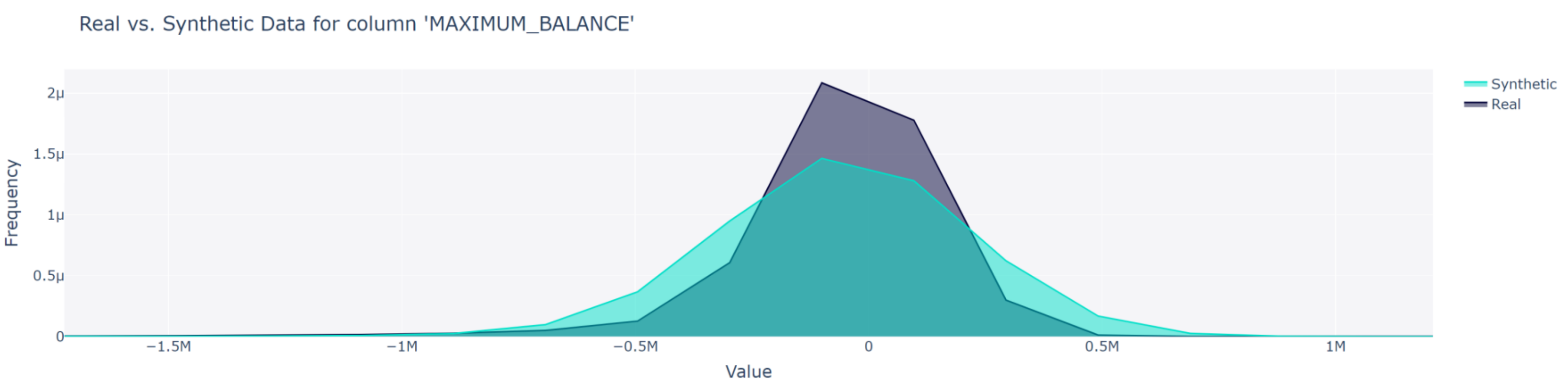
Gaussian Copula - SDV

Statistical Model

- *Models dependencies between variables*
- *Assumes linear relationships in handling dependencies*
- *Computationally efficient*

Comparison based on statistical similarity

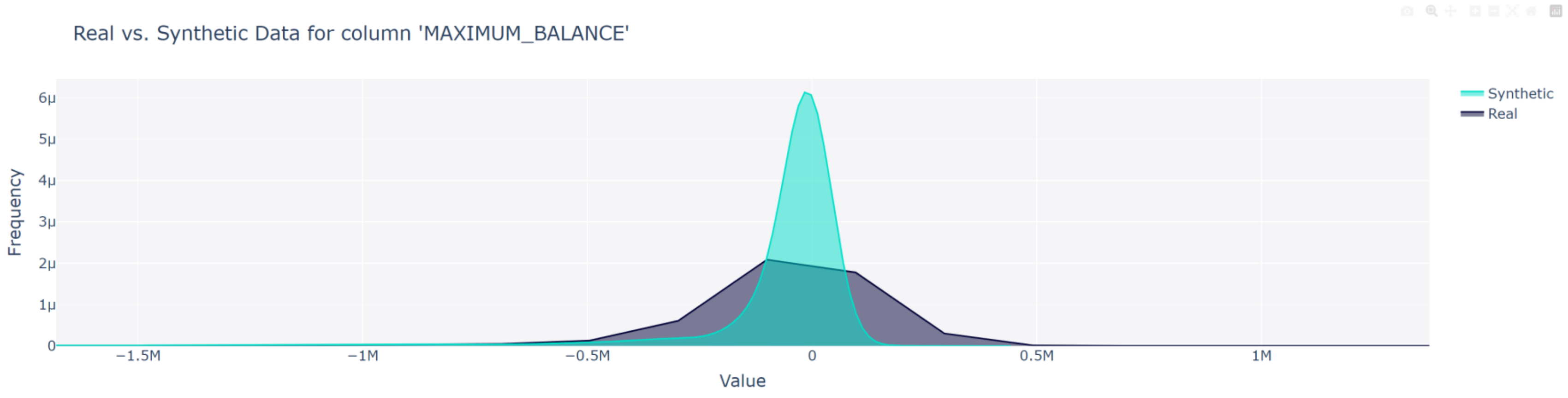
- Univariate distribution – based on subset of 10k entities



Forest Flow Diffusion

XG-Boost w Diffusion models:

- *Iterative noise injection*
- *Addresses mixed data types*
- *Flow-based models improve complex data distribution modeling*



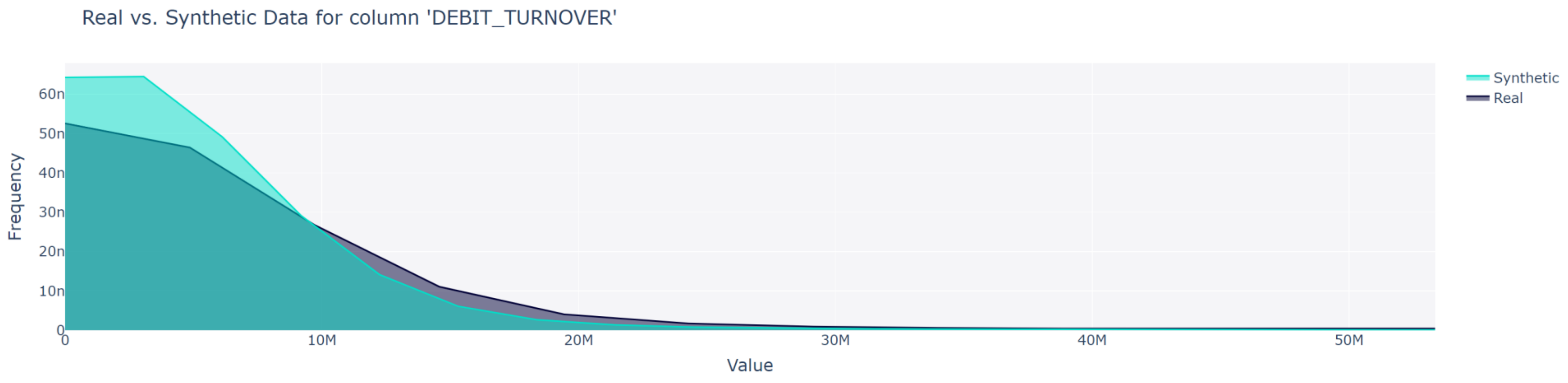
CT-GAN (1000 epochs)

Conditional Tabular GAN:

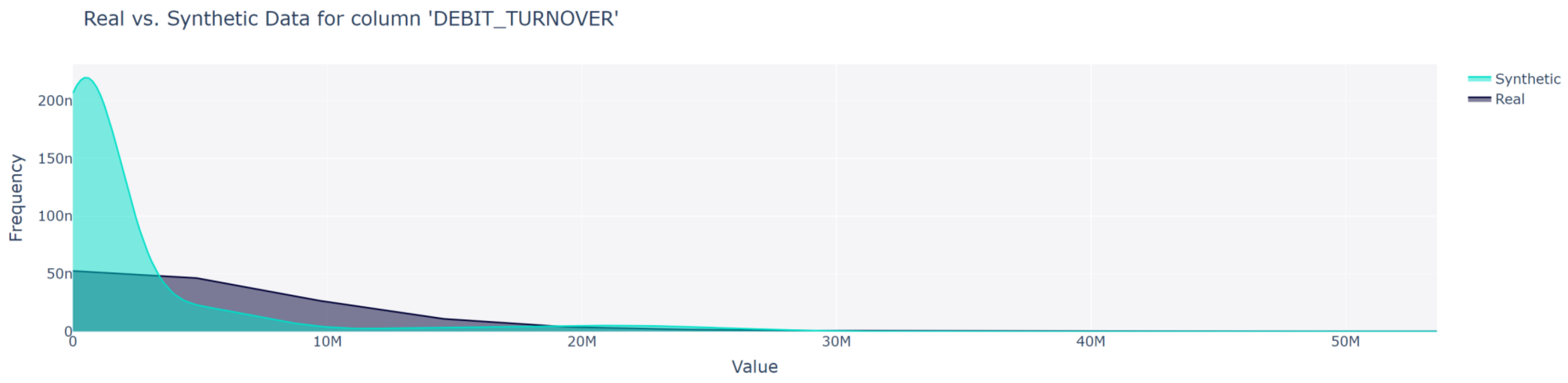
- *tabular data synthesis*
- *Addresses mixed data types / imbalanced distributions*
- *Enhanced privacy preservation*

Comparison based on statistical similarity

- Univariate distribution – based on subset of 10k entities



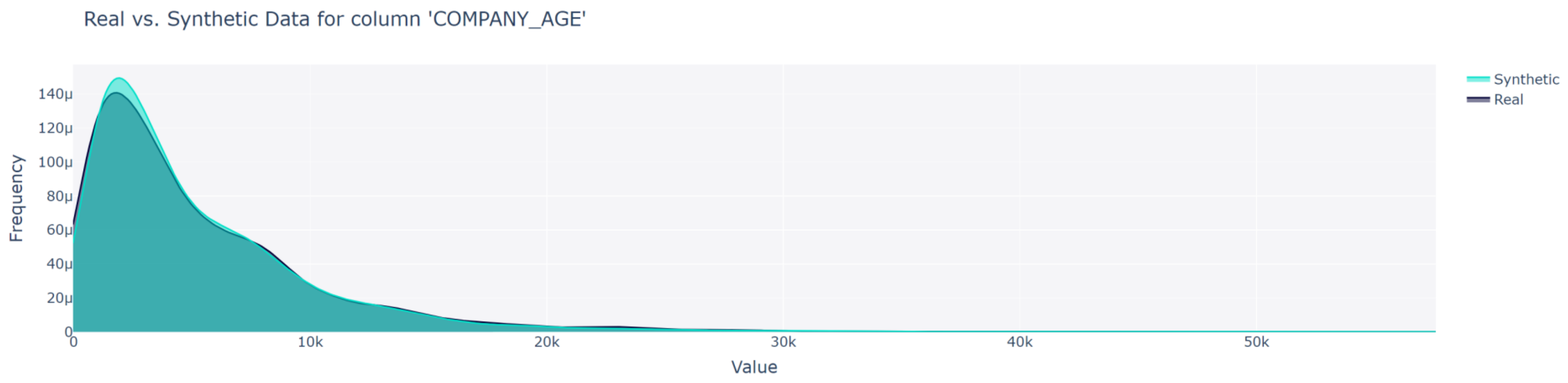
Forest Flow Diffusion



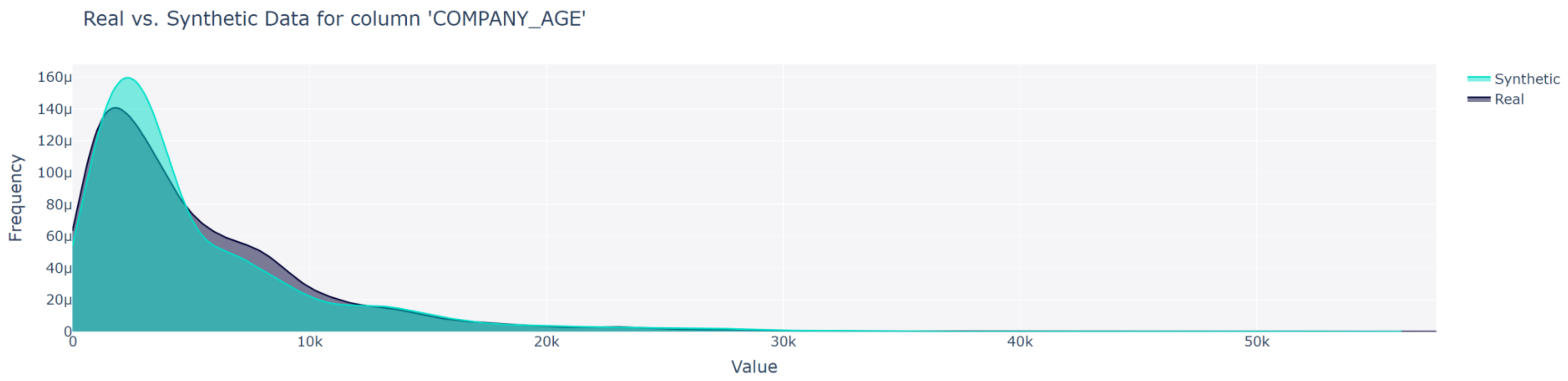
CT-GAN (1000 epochs)

Comparison based on statistical similarity

- Univariate distribution – based on subset of 10k entities



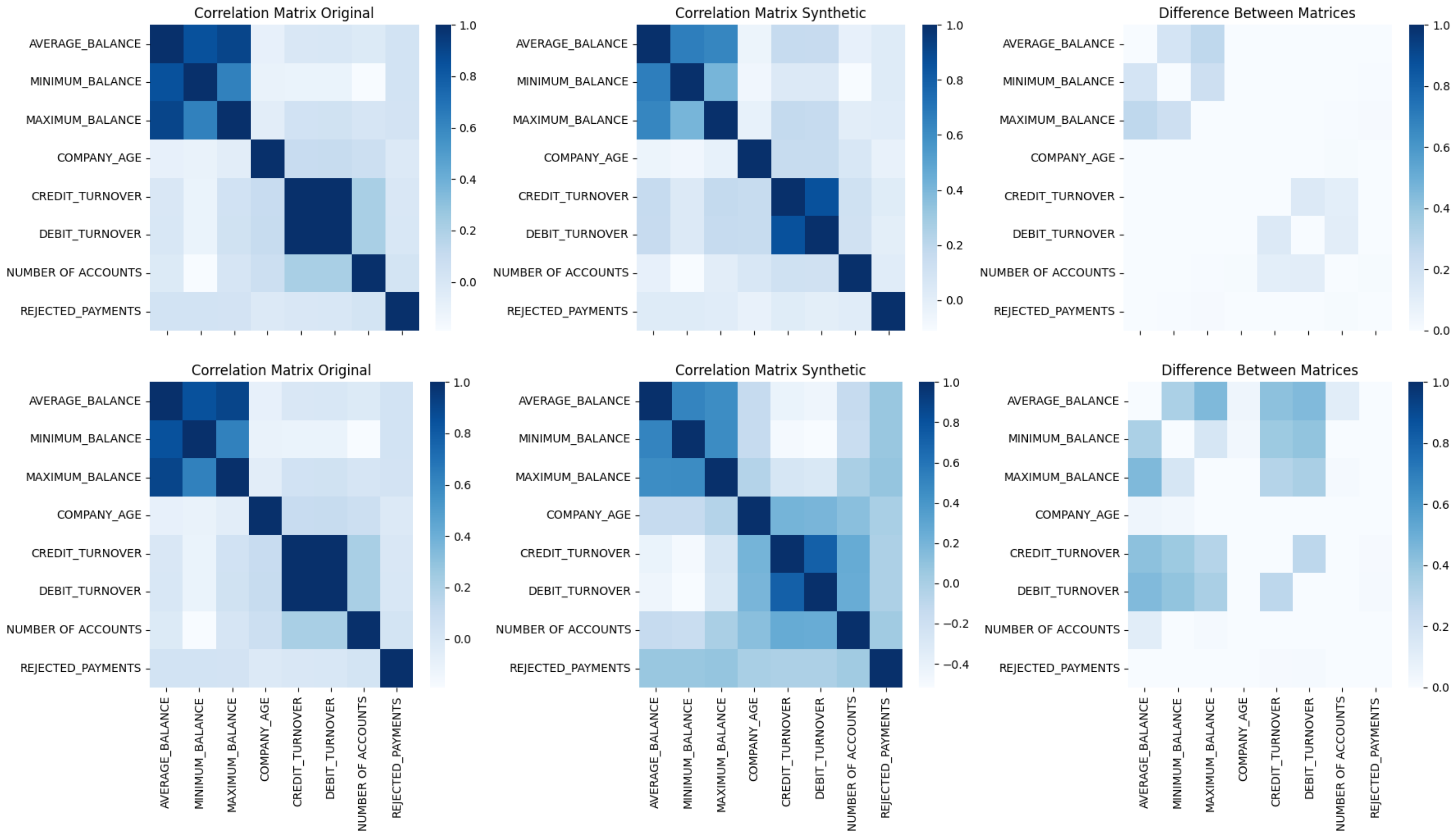
Forest Flow Diffusion



CT-GAN (1000 epochs)

Comparison based on statistical similarity

- Correlation – based on subset of 10k entities

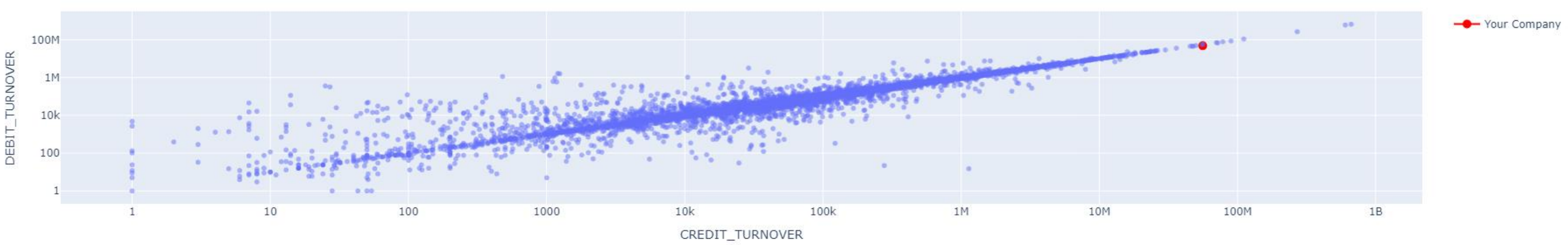


Forest Flow Diffusion

CT-GAN (1000 epochs)

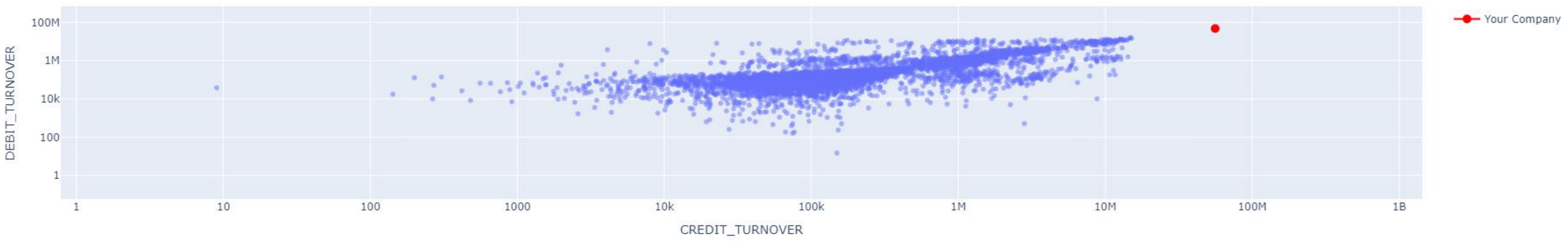
Comparison based on statistical similarity

- Bivariate distribution with hypothetical company – based on subset of 10k entities



Original sample

Forest Flow Diffusion



CT-GAN (1000 epochs)

Agent based modelling

Agent-based Modelling (ABM)

Pros
- Generates complete picture
- Granular control
- Little/no privacy concerns (?)

Cons
- Computationally expensive
- Model attack concerns
- Requires complex behaviours to be pre-defined (extensive domain knowledge)

Machine Learning (GAN; DP-GANs; P-PGM; VAE…)

Pros
- Can capture complex interactions
- Effective for high dimensionality
- Model temporal relationships

Cons
- Can be computationally expensive
- Privacy risks (overfitting)

Statistical Methods

Pros
- Established
- Interpretable

Cons
- Limited utility
- Privacy risks

Agent-Based Simulation

Models financial transactions using autonomous agents representing banks, customers, and fraudsters, simulating interactions based on predefined behavioural rules.

- Can model evolving fraud patterns
- Useful for stress testing financial crime models
- Computationally expensive
- Requires expert domain knowledge to design accurate agent behaviours / frequent manual intervention

UNIVERSITY of GREENWICH

FCA FINANCIAL CONDUCT AUTHORITY

Search

About us | Firms | Markets | Consumers | News | Publications

Home > Firms > Authorised Push Payment synthetic data

Authorised Push Payment synthetic data

Print Page

Share page

APP Fraud Dataset Success Criteria

ıdated: 12/03/2025

ı Payment (APP) fraud synthetic data, which covers
əs, transactions, telecom data, and fraud to improve

5 million, highlighting the urgent need for advanced technological
e FCA and Payment Systems Regulator (PSR) hosted an APP Fraud
: limited access to data for innovation. This led to the creation of a
raud detection innovations while safeguarding consumer privacy.

ated through agent-based simulations, a modelling approach that
ıns of approximately 20,000 individuals over two years. They cover
a wide range of data, including:

- individual and business identities
- bank account details
- phone call and SMS metadata
- fraud instances (both reported and successful)

The data is structured across 4 synthetic banks and 2 synthetic telecom operators. The banking data is formatted to reflect what would typically be accessible by an individual with high-level access, allowing visibility into fields like unredacted personal information of account holders, detailed transaction narratives, and destination account details for payments. Similarly, the telecom data is unredacted, providing access to call and text histories for each data subject. Both the banking and telecom data encompass information on individuals and businesses, and it is possible to find instances where fraudsters have used business accounts to receive fraudulent payments.

Digital Sandbox
Authorised Push Payment synthetic data
Digital Sandbox pilots

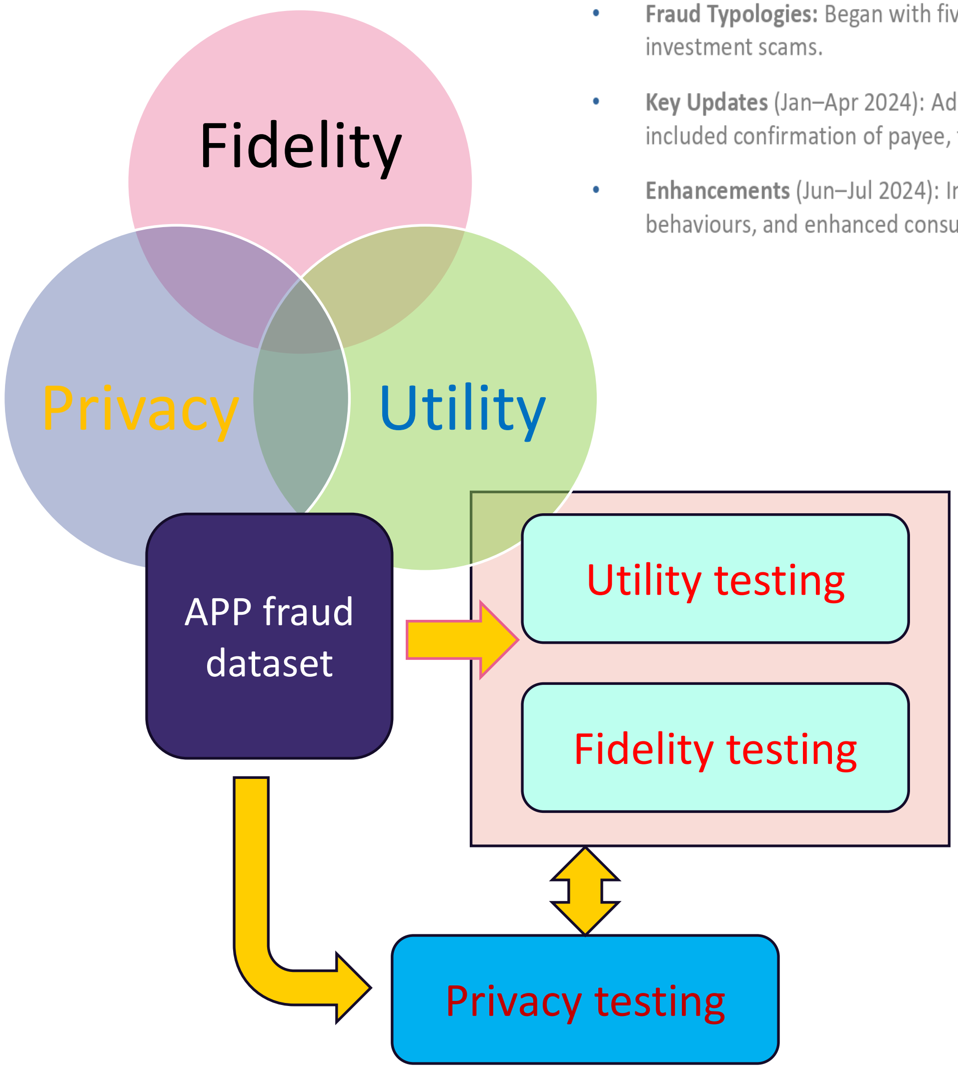
Related content
Read the APP fraud dataset evaluation report
FCA and COLC Authorised Push Payment synthetic data launch
Report on using synthetic data in financial services

UNIVERSITY of GREENWICH

APP fraud dataset project

Dataset Highlights & Enhancements

- **Initial Dataset (Sep 2023):** Featured 5.2GB of data across 37 datasets, including:
  - 15 million transactions
  - 58 million data points
  - 61,000 fraud attempts spanning two years, covering 20,000 synthetic individuals' bank and transaction data.
- **Fraud Typologies:** Began with five types—bank, police, and family impersonation, advance fee, and purchase scams—later expanding to include romance and investment scams.
- **Key Updates** (Jan–Apr 2024): Added foreign exchange transactions, transaction currency, and improvements to ethnicity, identity, and documentation data. New features included confirmation of payee, transaction categories, and enhanced accuracy for social finance and family data.
- **Enhancements** (Jun–Jul 2024): Introduced scam refunds, dynamic susceptibility, and scammer demographics. Adjusted typology frequencies, refined scammer behaviours, and enhanced consumer profiles with income and family data.



A three-stage project

1. *ML-efficacy testing (utility).*
2. *Evaluation of privacy-preserving attributes for the APP Fraud Synthetic Dataset.*
3. *Evaluation of statistical fidelity of the APP Fraud Synthetic Dataset.*

Defining a 'dial' to adjust depending on privacy/fidelity trade-off

FinTech use of synthetic data: *Implications*

Caveats & opportunities from extending Use Cases

For industry

- **Increased number of available datasets**
  - Enhancing model training depth
  - Collaboration mechanics *(synthetic data sharing?)*
- **Market maturity yet to be achieved**
  - Level of confidence to synthetic data?
  - Regulatory / legislative complexities
    - **Explainability** (…of model decision)
    - **Interpretability** (...of model mechanics)
- **Governance trade-offs**
  - Impacts on data supply chain for AML/CTF
  - Data Governance / Model Governance more prominent
  - ***Better Data Sharing <-> Sharing Better Data***
- **Confidential Data Pooling (CDP) Revamped?**
  - In-house synthetic data generative capability
  - Models for Industry-owned **CDP Sandbox**
  - Synthetic Data Sharing – model retraining

For regulators

- **A renewed role of the regulator as:**
  - Process custodian (vetted generative models)
  - Issuer for *'generative data governance'* rulesets
  - Sandbox utility provider (see UK FCA model)
- **FinTech Innovation support**
  - Data Privacy Laws / EU GDPR amendments
  - Widening accessibility & cost to training data

Summary Notes

- Collective Intelligence as goal
- Significant value of collaboration (reliable deliverables not otherwise feasible)
- Trilemma and privacy budgets as critical
- GenAI limitations
  - running out of usable real data without collaboration
  - Fincrime is primarily rare-event based
  - Over-generalisation may lead to unrealistic data
  - Privacy concerns (model inversion / model poisoning attacks et al)
- Potentially new models for data sharing / training?
- Potentially calls for change in Privacy Laws and related regulations

UNIVERSITY of GREENWICH

References

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., ... & Koyejo, S. (2024). Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*.

Fan, L., Chen, K., Krishnan, D., Katabi, D., Isola, P., & Tian, Y. (2024). Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7382-7392).

Ge, T., Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Prof. Georgios Samakovitis, MEng, MSc, MBA, SFHEA

Professor of FinTech
School of Computing & Mathematical Sciences
University of Greenwich