Evaluating Privacy & Utility in Synthetic Tabular Data

Presented & Proposed by:
Brandon Ismalej

California State University, Northridge

May 22, 2025

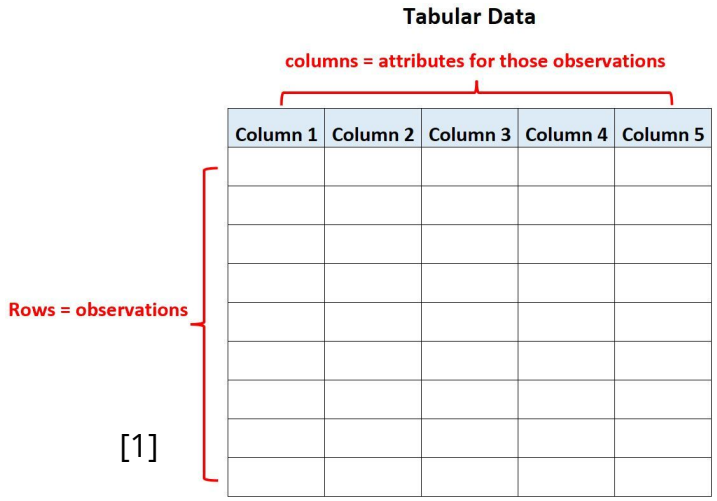
Why synthetic data? Why this project?

- Synthetic data has become increasingly and widely used for privacy-preserving data sharing and ML-based tasks

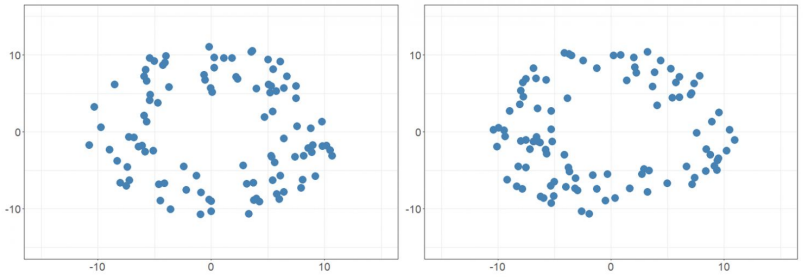
- Tabular data appears in healthcare, finance, and other sensitive domains
- … but synthetic ≠private by default, privacy can still be leaked

- Need to **measure both privacy and utility** to evaluate real-world safety

  Privacy?? Utility??



[1]



Original data Synthetic data

The synthetic data retains the structure of the original data but is not the same

[2]

Research Question & Goals

To what extent does synthetic tabular data leak private information from training records, and how well does it support downstream ML tasks?

- Quantify privacy risk using membership inference attacks (MIAs)
- Measure utility by training models on synthetic data, testing on real holdout
- Analyze privacy-utility tradeoff
- Evaluate across multiple models and datasets

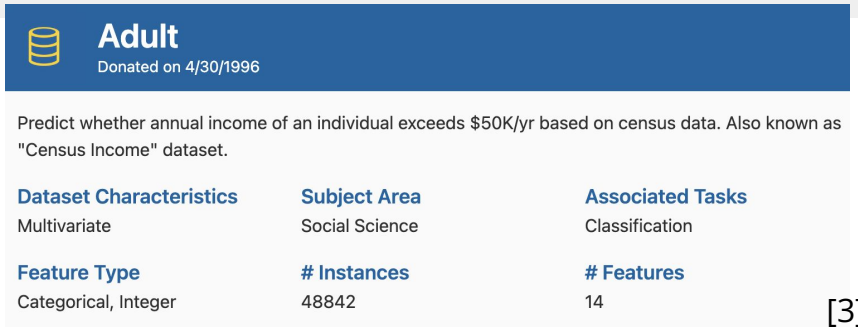
Datasets and Synthetic Data Generators

Datasets:

- UCI Adult
- Bank Marketing

Synthetic Generators:

- CTGAN (deep generative model)
- TVAE (variational autoencoder)
- GaussianCopula (statistical baseline)

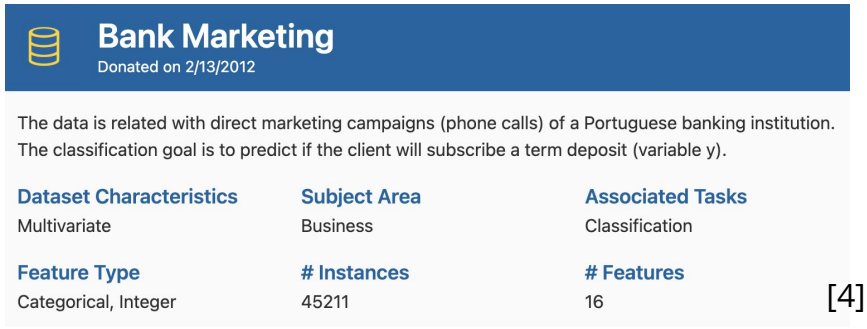


Adult
Donated on 4/30/1996

Predict whether annual income of an individual exceeds $50K/yr based on census data. Also known as "Census Income" dataset.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Social Science	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer	48842	14

[3]

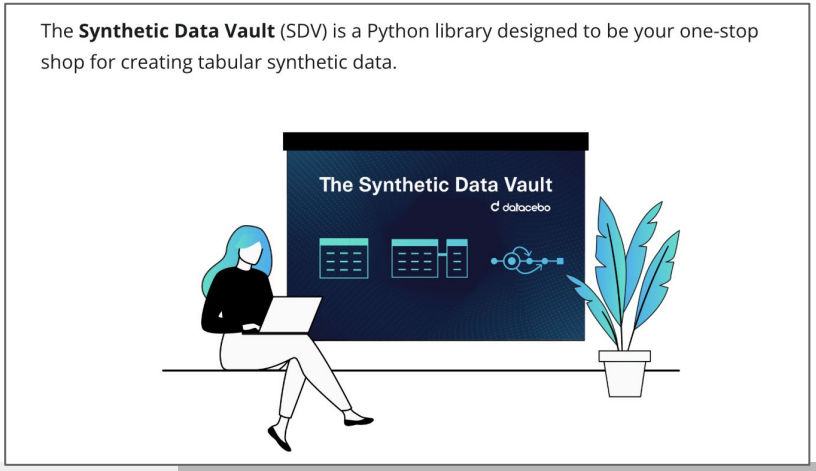


Bank Marketing
Donated on 2/13/2012

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y).

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Business	Classification
Feature Type	**# Instances**	**# Features**
Categorical, Integer	45211	16

[4]



The **Synthetic Data Vault** (SDV) is a Python library designed to be your one-stop shop for creating tabular synthetic data.

[5]

How Is Privacy Evaluated?

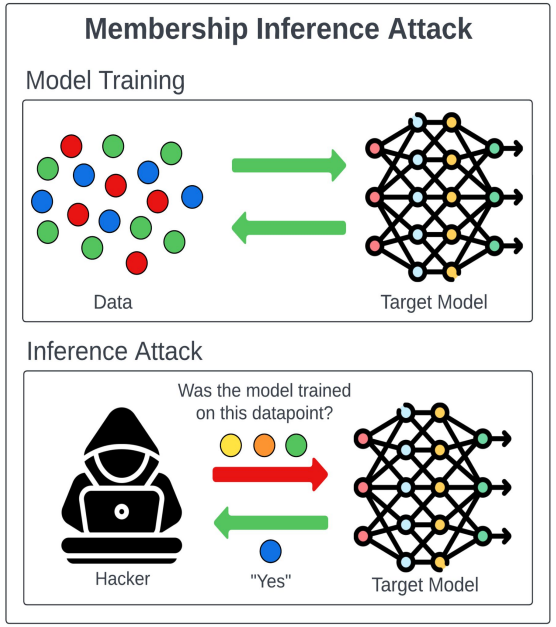
Membership Inference Attacks (MIAs)

- *Can an attacker guess if a real person's data was used in training?*
- Attempts to infer whether a specific indiv.'s data was part of model's training set [5]

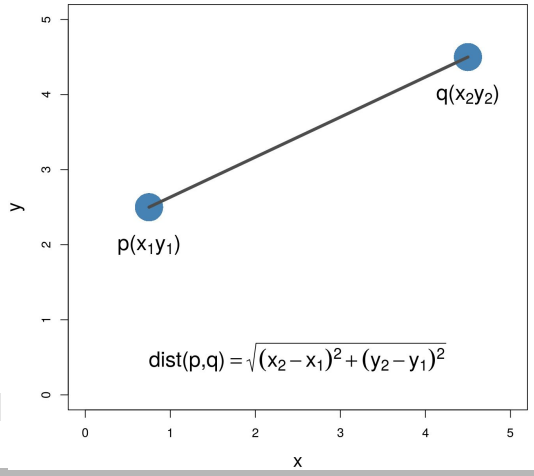
Three Evaluations:

- **Distance-based** (unsupervised): If real training records are "closer" to synthetic data
- **Model-based:** logistic regression trained to predict membership
- **Worst-case:** Focus on only 10% of most exposed records

Metric: AUC (0.5 = safe, >0.5 = leakage)



[6]

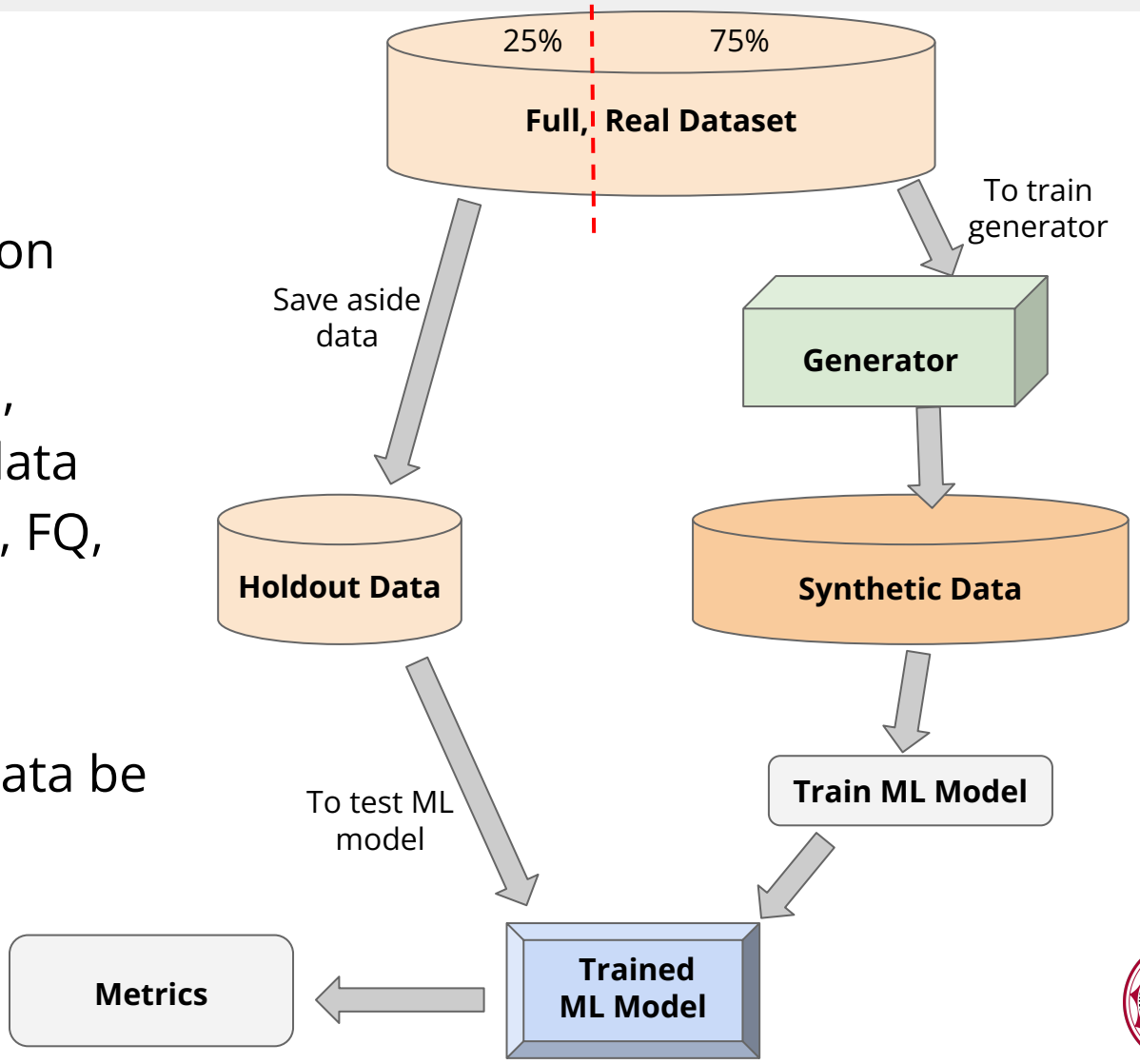


[7]

How Is Utility Evaluated?

- Train ML models on synthetic data
- Test them on real, unseen holdout data
- Metrics: Accuracy, FQ, ROC AUC

Goal: Can synthetic data be used for useful ML?



Privacy-Utility Tradeoff

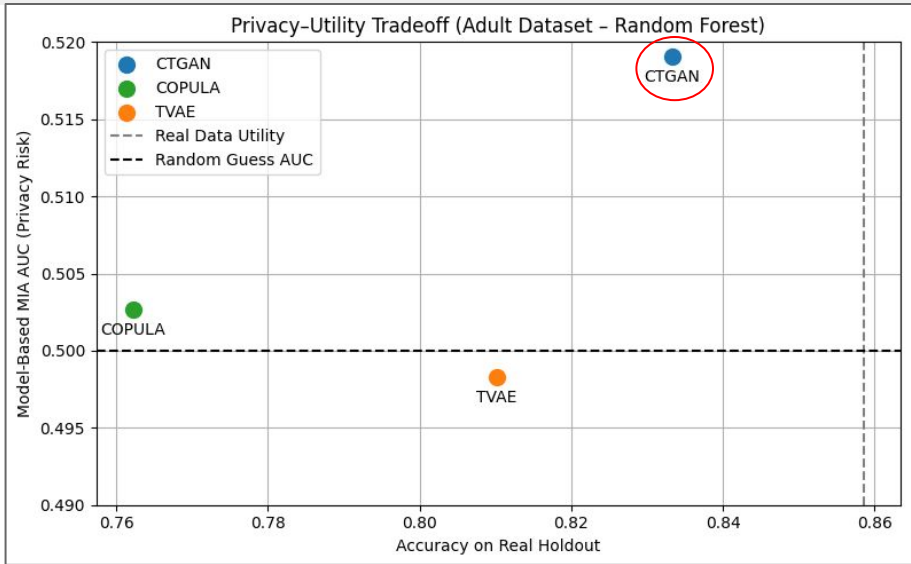
- Not just “how private?” or “how useful?” … but both
- Tradeoff plots compare:
  - X: Accuracy on real holdout
  - Y: Privacy Risk (MIA AUC)

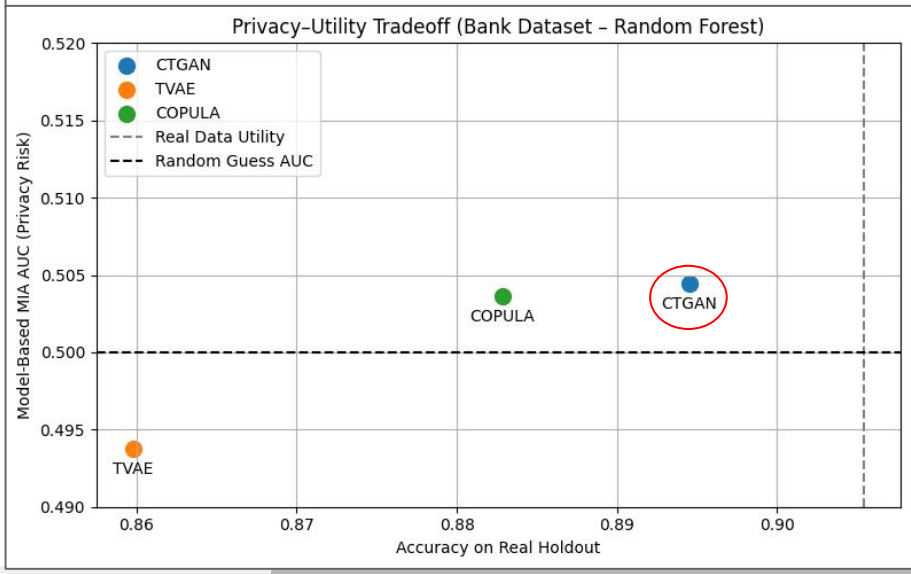
Tradeoff Concept:

- Stronger utility often comes with higher privacy leakage

Theoretical “Sweet Spot”

- High utility, no privacy risk





Summary

- This work implements a complete empirical evaluation of tabular synthetic data across three dimensions:
  - **Privacy risk:** via membership attacks (distance-based, model-based, worst-case)
  - **Predictive utility:** using classification models trained on synthetic data, test on real holdout sets
  - **Privacy-utility tradeoff:** visualized to assess balance between risk and usefulness
- The evaluation is applied to:
  - **2 real-world datasets:** UCI Adult, Bank Marketing
  - **Three generation methods:** CTGAN, TVAE, GaussianCopula
  - **Two classifiers for utility assessment:** Logistic regression, Random forest

References

[1]Z. Bobbitt, “What is Tabular Data? (Definition & Example),” Statology, Mar. 25, 2022. https://www.statology.org/tabular-data/ (accessed May 14, 2025).

[2]K. Walker, “Synthetic data: Unlocking the power of data and skills for machine learning – Data in government,” Blog.gov.uk, Aug. 20, 2020. https://dataingovernment.blog.gov.uk/2020/08/20/synthetic-data-unlocking-the-power-of-data-and-skills-for-machine-learning/ (accessed May 15, 2025).

[3]“UCI Machine Learning Repository,” Uci.edu, 2019. https://archive.ics.uci.edu/dataset/222/bank+marketing (accessed May 15, 2025).

[4]“UCI Machine Learning Repository,” Uci.edu, 2019. https://archive.ics.uci.edu/dataset/2/adult (accessed May 15, 2025).

[5]“Welcome to the SDV! | Synthetic Data Vault,” Sdv.dev, May 07, 2025. https://docs.sdv.dev/sdv (accessed May 15, 2025).

[6]Mindgard, “AI Under Attack: Six Key Adversarial Attacks and Their Consequences,” Mindgard.ai, Jan. 22, 2025. https://mindgard.ai/blog/ai-under-attack-six-key-adversarial-attacks-and-their-consequences (accessed May 14, 2025).

[7]M. Wayland, “5 Nearest neighbours | An Introduction to Machine Learning,” Github.io, Mar. 12, 2019. https://bioinformatics-training.github.io/intro-machine-learning-2019/nearest-neighbours.html (accessed May 15, 2025).