AI Generated Synthetic Data – A Good Way to Protect Privacy?

Dr. Marcel Neunhoeffer

IAB Brown Bag Talk – Young Researcher Network

About Me

- PhD in Political Science at the University of Mannheim (GANs for Social Scientists)
- 2021-2023: Researcher/Postdoc at Boston University
- Since August 2023: Postdoc at KEM
- Part of the AnigeD Project (Forschungsverbund “Anonymität bei integrierten und georeferenzierten Daten”)

We are in the decade of Generative AI – Why should we, as Social Scientists, care?

Generative AI is the fastest-growing technology ever – and it has come a long way since 2014

2014

2023

Generative AI is transforming social science research

NEW RESEARCH QUESTIONS

- Fake news/disinformation
- Biased outputs
- Ethical concerns

ADVANCED METHODS

- Synthetic data
- Multiple Imputation

PRODUCTIVITY ENHANCEMENTS

- Coding support
- Text production

We are in the decade of Generative AI – Why should we, as Social Scientists, care?

Generative AI is the fastest-growing technology ever – and it has come a long way since 2014

2014

2023

Generative AI is transforming social science research

NEW RESEARCH QUESTIONS

- Fake news/disinformation
- Biased outputs
- Ethical concerns

ADVANCED METHODS

- Synthetic data
- Multiple Imputation

PRODUCTIVITY ENHANCEMENTS

- Coding support
- Text production

What is Synthetic Data?

When Rubin (1993) introduced the idea of fully synthetic data , there was considerable appeal to releasing data that represented "no actual individual's" responses, and skepticism regarding its feasibility. Subsequent research has adequately demonstrated the feasibility. However, the basic question "How much protection does the synthetic data methodology provide?" remained largely unanswered.

Source: Abowd & Vilhuber 2008

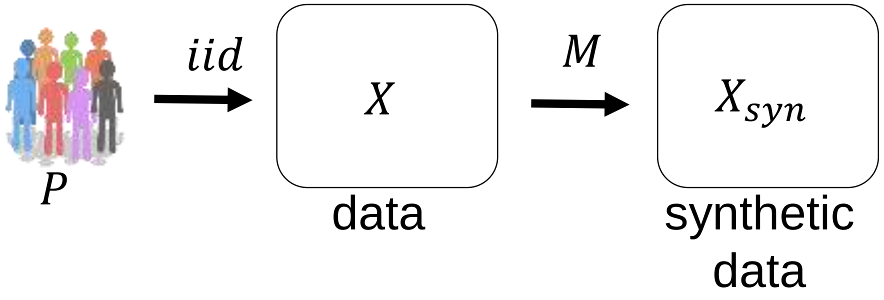
The idea of synthetic data sets is similar. A statistical process is used to extract information from an actual data set collected from a set of respondents and is reexpressed as a collection of artificial or synthetic data sets for public consumption. This allows wide dissemination of the informational content of the actual data set and, at the same time, limits the exposure to potential inadvertent or malicious disclosure of sensitive information about the respondents.

Source: Raghunatan 2021

A promising alternative to address the trade-off between broad data access and disclosure protection is the release of synthetic data. With this approach, a model is fitted to the original data and draws from this model are used to replace the original values. Depending on the desired level of protection, only some records (partial synthesis) or the entire dataset (full synthesis) are replaced by synthetic values.

Source: Drechsler & Haensch 2023

What is Synthetic Data?



A brief Introduction to Generative Adversarial Nets (GANs)

The central promise

Generative AI learns distributions from data and then generates new samples from the underlying distribution (mainly images and text).

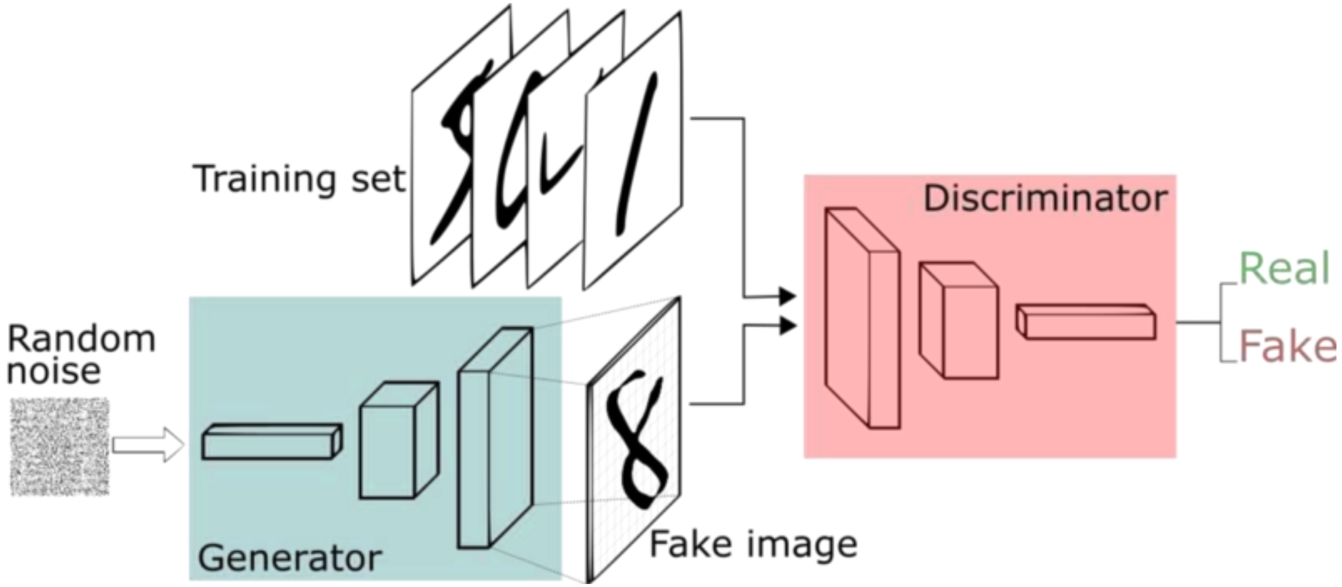
Different Generative AI models

based on, e.g.,

– Generative Pre-Trained Transformers (GPT) (Radford et al., 2018),
– Diffusion models (Sohl-Dickstein et al., 2015), or
– **Generative Adversarial Nets (GAN) (Goodfellow et al., 2014)**

are used today.

A brief introduction to GANs:



Silva, 2018

Synthetic Data Can Leak Information about the Real Data!

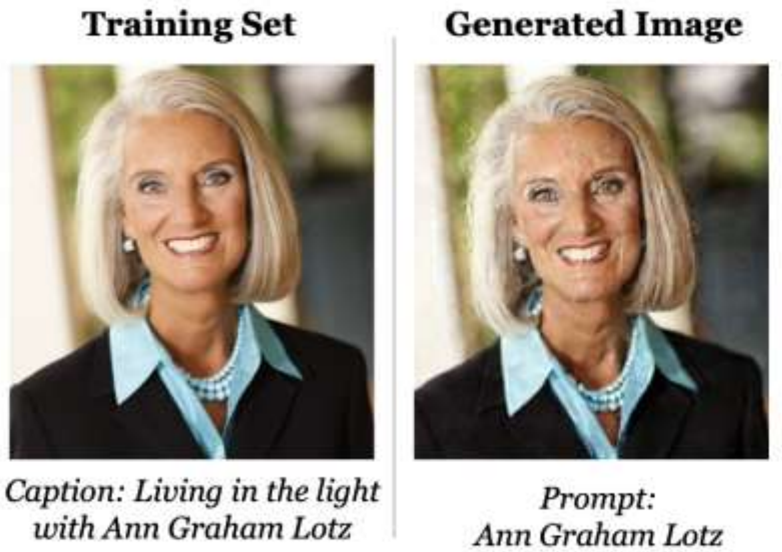
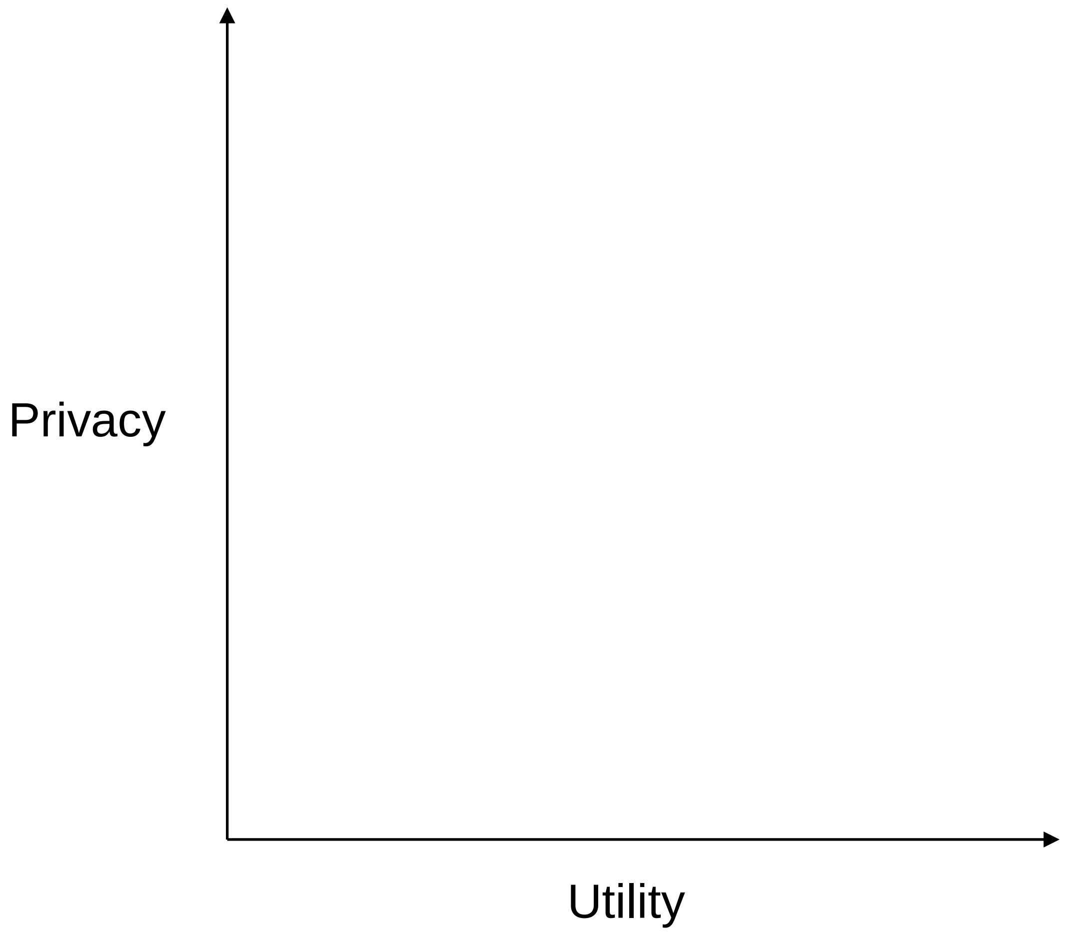


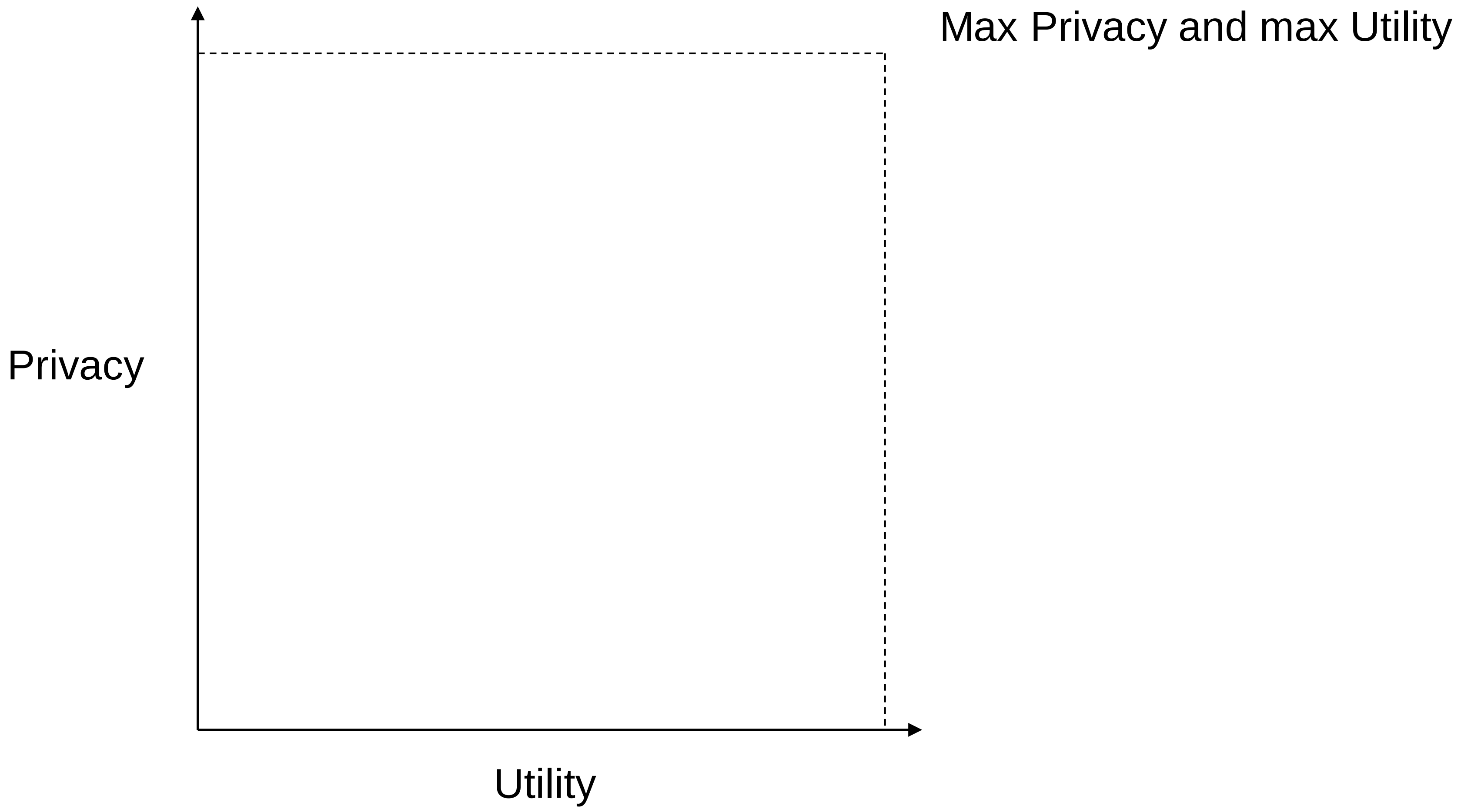
Figure 1: Diffusion models memorize individual training examples and generate them at test time. **Left:** an image from Stable Diffusion's training set (licensed CC BY-SA 3.0, see [49]). **Right:** a Stable Diffusion generation when prompted with "Ann Graham Lotz". The reconstruction is nearly identical (ℓ_2 distance = 0.031).

Source: [Carlini et al. 2023]()

The Privacy-Utility Tradeoff



The Privacy-Utility Tradeoff



A Primer on Differential Privacy

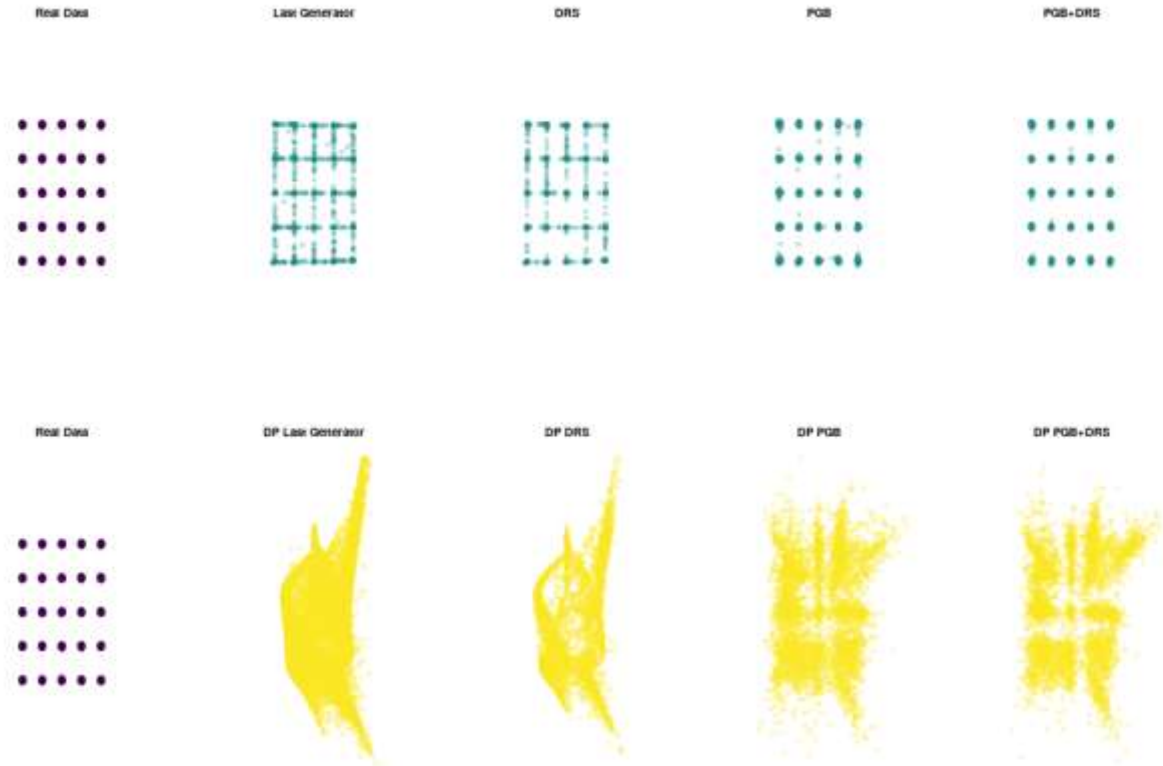
Definition 1 (Differential Privacy (DP) (Dwork et al., 2006)). *A randomized algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{R}$ with output domain $\mathcal{R}$ (e.g. all generative models) is (ε, δ)-differentially private (DP) if for all adjacent datasets $X, X' \in \mathcal{X}^n$ and for all $S \subseteq \mathcal{R}$: $P(\mathcal{A}(X) \in S) \leq e^{\varepsilon} P(\mathcal{A}(X') \in S) + \delta$.*

Training GANs with Formal Privacy Guarantees (**Neunhoeffer**, Wu & Dwork 2021)

Starting points:

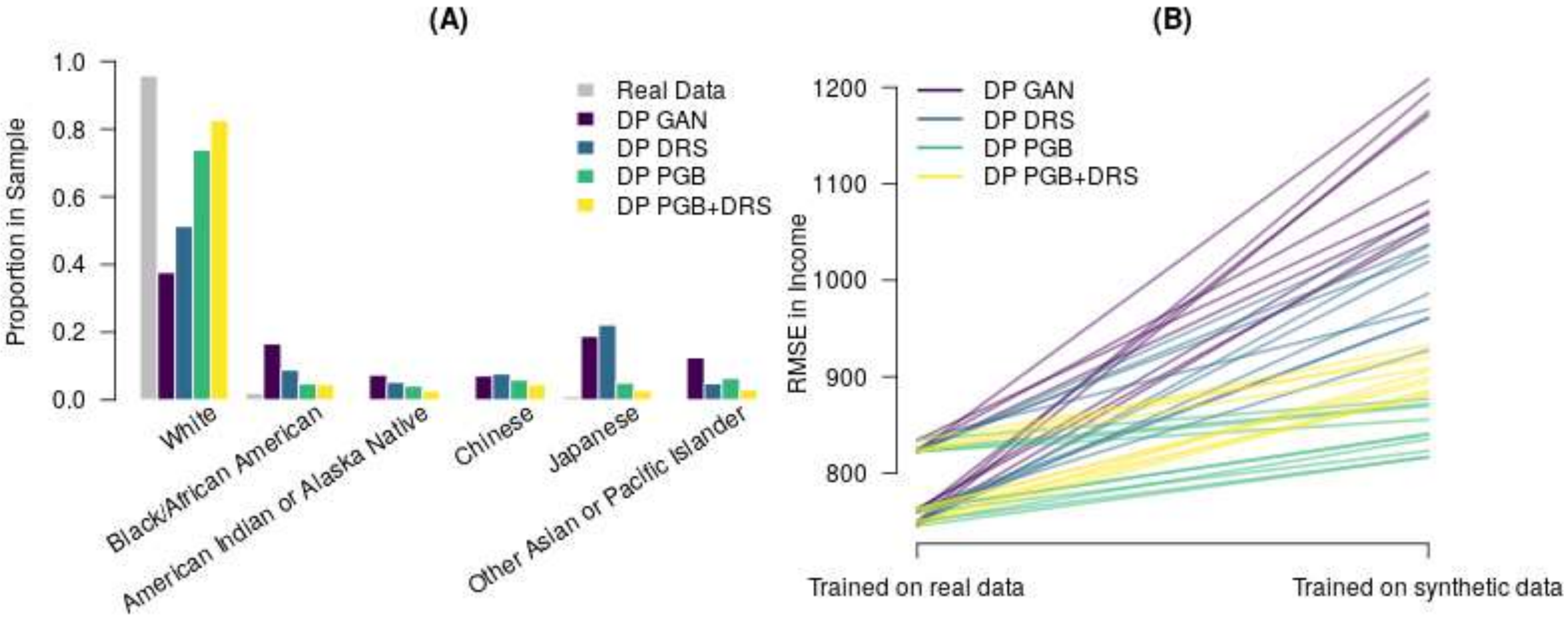
- Synthetic data without formal privacy guarantees is not sufficient
- Training GANs with differential privacy makes convergence harder
- But a mixture of generators can contain information for good synthetic data

Outcomes:



- Introduction of Private PGB to generate high-quality samples from a sequence of GAN generators
- Our experiments show this novel method improves upon a standard private GAN approach across a collection of quality measures

Different Ways to Measure Utility



Different Ways to Measure Utility

Table 1: Predicting Titanic Survivors with Machine Learning Models trained on synthetic data and tested on real out-of-sample data. Median scores of 20 repetitions with independently generated synthetic data. With differential privacy ϵ is 2 and δ is $\frac{1}{2N} \approx 5.6 \times 10^{-4}$.

	GAN	DRS	PGB	PGB + DRS
Logit Accuracy	0.626	0.746	0.701	**0.765**
Logit ROC AUC	0.591	0.760	0.726	**0.792**
Logit PR AUC	0.483	0.686	0.655	**0.748**
RF Accuracy	0.594	0.724	0.719	**0.742**
RF ROC AUC	0.531	0.744	0.741	**0.771**
RF PR AUC	0.425	0.701	0.706	**0.743**
XGBoost Accuracy	0.547	0.724	0.683	**0.740**
XGBoost ROC AUC	0.503	0.732	0.681	**0.772**
XGBoost PR AUC	0.400	0.689	0.611	**0.732**
	DP GAN	**DP DRS**	**DP PGB**	**DP PGB +DRS**
Logit Accuracy	0.566	0.577	0.640	**0.649**
Logit ROC AUC	0.477	0.568	0.621	**0.624**
Logit PR AUC	0.407	0.482	0.532	**0.547**
RF Accuracy	0.487	0.459	0.481	**0.628**
RF ROC AUC ROC AUC	0.512	0.553	0.558	**0.652**
RF PR AUC PR AUC	0.407	0.442	0.425	**0.535**
XGBoost Accuracy	0.577	0.589	0.609	**0.641**
XGBoost ROC AUC	0.530	0.586	**0.619**	0.596
XGBoost PR AUC	0.398	0.479	0.488	**0.526**