THE VIABILITY BOUNDARY OF DIFFERENTIAL PRIVACY

Arinbjörn Kolbeinsson
K01
Reykjavík, Iceland
arinbjorn@k01.is

Benedikt Kolbeinsson
K01
Reykjavík, Iceland
benedikt@k01.is

ABSTRACT

Differentially private synthetic data can enable data sharing without compromising individual privacy, but DP-SGD adds noise that can destroy utility when training data is scarce. How much data is enough is poorly understood. We characterise a sharp *viability boundary*, a training set size below which DP models produce random-chance output and above which they approach non-private baselines. Across six tabular datasets spanning healthcare, census and ecology domains, we find that the ratio N/d (training samples per encoded dimension) consistently predicts this transition, with viability emerging between $N/d \approx 50$ and 300. The boundary is insensitive to model size. The data cost of strong privacy is sublinear, with $\epsilon = 1$ requiring only $\sim 2.5\times$ more data than $\epsilon = 10$, well below formal DP-ERM predictions. A controlled dimension-reduction experiment confirms that N/d, not N alone, drives viability. These results give practitioners an actionable heuristic: check N/d before investing in DP synthetic data generation, and prefer feature engineering over data collection when the ratio is too low.

1 INTRODUCTION

When sensitive data cannot be shared directly, synthetic data generated under differential privacy offers a path forward. Recent high-profile deployments span government, research and healthcare: the US Census Bureau applied differential privacy to the 2020 Decennial Census (Abowd et al., 2022), NIST ran a public competition on differentially private synthetic data generation (McKenna et al., 2021) and healthcare is a growing adoption area (Giuffrè & Shung, 2023). The standard training mechanism, DP-SGD (Abadi et al., 2016), clips per-sample gradients to norm C and adds Gaussian noise with standard deviation σC, producing a per-step noise vector whose norm scales as $\sigma C\sqrt{p}$ where p is the number of model parameters.

Below some dataset size, this noise overwhelms the gradient signal and the generative model cannot learn. Above that size, the noise averages out over training steps and the synthetic data becomes viable for downstream tasks. We call the transition between these regimes the *viability boundary*: the minimum training set size N_{viable} at which differentially private synthetic data retains useful predictive signal.

Prior work has established that DP-SGD requires substantially more data for competitive accuracy (Tramèr & Boneh, 2021) and benchmarks compare DP generators at fixed dataset sizes (Tao et al., 2022), but neither identifies the dataset-size threshold at which utility collapses. A recent review of 74 healthcare DP studies finds that strict privacy often degrades accuracy, particularly for smaller or more complex datasets (Mohammadi et al., 2026), without quantifying how small is too small. The practical question is direct: given N records and d encoded features, will DP-SGD produce useful synthetic data? No prior work maps this boundary.

Why this matters. The domains that need differentially private synthetic data most are those with limited data. A hospital with 30 000 electronic health records and 400 clinical features sits at $N/d \approx 75$. A rare-disease registry may have only a few thousand patients. A census bureau must release high-dimensional microdata under formal privacy guarantees. In each case the core question is identical: is this dataset large enough for DP to work? An incorrect answer means either wasted

compute (training a model that produces noise) or unnecessary caution (withholding data that could safely be shared). Healthcare is the most acute case. Single-institution datasets are inherently small, credentialled access requirements limit data pooling (Johnson et al., 2016) and privacy regulations such as HIPAA and GDPR prevent the "just collect more data" solution available to technology companies (Price & Cohen, 2019).

What we find. We characterise the viability boundary empirically across six datasets spanning healthcare and non-healthcare domains, three independent seeds and multiple privacy budgets. Three findings emerge. First, a clear viability boundary exists across all six datasets tested (Section 4.1). The transition from random-chance output to viable synthetic data occurs over a narrow range of training set sizes, and the ratio N/d (training samples per encoded dimension) predicts the transition zone: below $N/d \approx 50$, DP synthetic data carries negligible utility; above $N/d \approx 300$, it is consistently viable. Second, the boundary appears insensitive to model size (Section 4.2). Three MLP VAE architectures spanning a $14\times$ range in parameter count (11k to 157k) share the same boundary bracket on the Adult dataset, suggesting the threshold is driven by data sufficiency rather than model capacity (Tramèr & Boneh, 2021). Third, the data cost of strong privacy is sublinear (Section 4.3). Reducing the privacy budget from $\epsilon = 10$ to $\epsilon = 1$ requires only $\sim 2.5\times$ more data, well below both the $\sim 20\times$ from per-step noise scaling (Bottou et al., 2018) and the $\sim 4.6\times$ from formal DP-ERM bounds (Bassily et al., 2014) (Section 2.2). For $\epsilon \geq 2$, the additional data cost relative to $\epsilon = 10$ is negligible.

Contributions. Our contributions are threefold: (1) we empirically map the viability boundary across six datasets with three-seed validation, identifying N/d as a consistent predictor of the transition zone; (2) we show the boundary is insensitive to model size via a three-architecture comparison on Adult; and (3) we show the data cost of $\epsilon = 1$ is $\sim 2.5\times$, below the $\sim 4.6\times$ from DP-ERM theory, validated on Adult, ACS Income and NHANES.

2 Background

2.1 Differential Privacy and DP-SGD

A randomised mechanism $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy (Dwork et al., 2006; Dwork & Roth, 2014) if, for all neighbouring datasets D, D' differing in a single record and all measurable sets S:

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S] + \delta \tag{1}$$

where ϵ controls the privacy loss bound and δ the probability of exceeding it.

DP-SGD (Abadi et al., 2016) adapts stochastic gradient descent for differential privacy through two modifications: each per-sample gradient is clipped to a maximum ℓ_2 norm C, and Gaussian noise with standard deviation σC is added to the aggregated batch gradient. The privacy budget ϵ is consumed over training; at fixed noise multiplier σ, more training steps yield a higher (weaker) ϵ. Tight accounting of this composition uses the privacy random variable (PRV) framework or Rényi DP (Mironov, 2017).

2.2 The Signal-to-Noise Argument

The viability boundary arises from a tension between gradient signal and DP noise. Over $T = N \cdot E/B$ training steps (where E is epochs and B is batch size), the accumulated gradient signal grows in proportion to $\sqrt{T}$. The DP noise vector added per step has norm $\sim \sigma C \sqrt{p}$, where p is the number of model parameters; this quantity is fixed regardless of N. For the MLP VAE architectures used here, p is dominated by the first encoder layer ($d \times 256$ weights), so $p \propto d$ and the two quantities are interchangeable up to a constant.

The resulting signal-to-noise ratio scales as:

$$\text{SNR} \propto \frac{\sqrt{N}}{\sigma\sqrt{d}} \tag{2}$$

For learning to succeed, N must be large enough that accumulated signal overcomes noise. Treating each step independently yields a per-step prediction: $N_{\text{viable}} \propto \sigma^2 d$, consistent with standard SGD

convergence rates under gradient noise (Bottou et al., 2018). However, formal DP-ERM bounds are tighter: excess risk scales as $\sqrt{d}/(n\epsilon)$ (Bassily et al., 2014), which, since $\epsilon \sim 1/\sigma$, gives $N_{\text{viable}} \propto \sigma\sqrt{d}$ once noise averaging over steps is accounted for.

Empirically, the scaling with σ is gentler still, due to optimiser adaptation and the fact that batch gradient variance shrinks with N. Quantifying the gap between these theoretical predictions and observed behaviour is the central empirical question of this paper.

2.3 Synthetic Data Utility Evaluation

We evaluate synthetic data quality using TSTR (Train on Synthetic, Test on Real) (Esteban et al., 2017), a standard utility metric for tabular synthetic data (Xu et al., 2019). A downstream classifier (XGBoost) is trained on synthetic data and evaluated on a held-out real test set, yielding an AUC score that measures whether the synthetic data preserves predictive signal. An oracle upper bound is obtained by training XGBoost directly on real training data, bypassing synthetic generation entirely. The baseline (BL) is distinct: it trains a VAE *without* DP noise, generates synthetic data, and evaluates via TSTR. The gap between oracle and baseline measures information loss from the synthetic pipeline itself; the gap between baseline and DP measures the additional cost of privacy. To compare across datasets with different baseline performance levels, we define a normalised DP signal:

$$\text{signal} = \frac{\text{DP AUC} - 0.5}{\text{peak BL AUC} - 0.5} \tag{3}$$

where the denominator is the peak baseline AUC for each dataset (a constant). A value of 1 indicates the DP model recovers the full discriminative signal of the best baseline; 0 indicates random-chance performance (AUC $= 0.5$). We define a DP model as *viable* when the normalised signal reaches 0.5, meaning it recovers at least half the peak baseline signal. Using a per-dataset constant as the denominator avoids instability when the baseline is itself near random at low N. This is an absolute utility measure rather than a same-N comparison; the raw DP and baseline curves at each N are shown separately in Figure 1.

2.4 VAEs for Tabular Synthetic Data

Variational autoencoders (Kingma & Welling, 2014) learn a generative model by jointly training an encoder $q_\phi(z \mid x)$ that maps data to a latent representation z and a decoder $p_\theta(x \mid z)$ that reconstructs the input. Training maximises a β-weighted evidence lower bound:

$$\mathcal{L}(\theta, \phi; x) = -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \beta \cdot \text{KL}(q_\phi(z|x) \| p(z)) \tag{4}$$

where the first term is the reconstruction loss and the second regularises the latent space towards a standard Gaussian prior. For tabular data, the decoder uses mixed output heads: mean squared error for continuous features, cross-entropy for categorical features and binary cross-entropy for binary features.

VAEs and other deep generative models are widely used for synthetic healthcare data (Jordon et al., 2022; Xu et al., 2019; Yoon et al., 2020); however, without differential privacy, synthetic data generated from such models has been shown to leak membership information (Stadler et al., 2022). We integrate DP-SGD via Opacus (Yousefpour et al., 2021), which wraps the standard PyTorch training loop with per-sample gradient computation, clipping and noise addition.

3 Experimental Setup

3.1 Datasets

We select six datasets spanning healthcare and non-healthcare domains with encoded dimensionalities from $d = 98$ to $d = 414$ (Table 1). The three non-healthcare datasets are standard ML benchmarks: **Adult** (UCI), predicting income above \$50k from census features (152 encoded dimensions after one-hot encoding); **ACS Income** (Ding et al., 2021), the same prediction task on American Community Survey data for California (98 dimensions, 156k records); and **Covertype** (UCI), a 7-class forest cover type classification (205 dimensions).

Table 1: Datasets used in viability boundary experiments. N_{train} is the number of training records; d is the encoded feature dimensionality after one-hot encoding; Oracle is the AUC of XGBoost trained directly on real data (mean $\pm$ std over 3 seeds).

Dataset	Domain	N_{train}	d	Target	Metric	Oracle
Adult	Demographics	34,188	152	Income >\$50k	ROC AUC	$.927 \pm .002$
ACS Income	Survey	156,532	98	Income >\$50k	ROC AUC	$.890 \pm .000$
Covertype	Ecology	464,809	205	Cover type (7-class)	Wt. AUC	$.960 \pm .000$
Diabetes-130	Clinical	71,236	213	Ext. LOS $\geq$7d	ROC AUC	$.844 \pm .003$
MIMIC-III	Clinical	30,830	414	In-hosp. mortality	ROC AUC	$.889 \pm .005$
NHANES	Public health	47,529	146	Diabetes	ROC AUC	$.834 \pm .002$

The three healthcare datasets represent the data-constrained settings that motivate this work. **Diabetes-130** (Strack et al., 2014) contains 71k diabetic hospital encounters; we bin the three ICD-9 diagnosis columns (each with $\sim$700 codes) into 18 clinical chapters, reducing d from 2,475 to 213. This dimensionality reduction is critical for DP viability (Section 7). The target is extended length of stay ($\geq$7 days). **MIMIC-III** (Johnson et al., 2016) contains 30k ICU stays with demographics, first-24-hour vitals and labs, and ICD diagnoses (414 encoded dimensions); the target is in-hospital mortality (11.4% prevalence). Its high dimensionality makes it the hardest dataset for DP. **NHANES** (Centers for Disease Control and Prevention, 2024) covers 47,529 participants across multiple survey cycles (146 dimensions) with diabetes as the prediction target; like Diabetes-130 it reaches DP viability at moderate privacy, though it falls short at $\epsilon = 1$.

3.2 Model Architecture

All experiments use the same MLP VAE: an encoder mapping d-dimensional input through hidden layers of 256 and 128 units to a 32-dimensional latent space, and a symmetric decoder. The decoder uses mixed output heads (MSE for continuous, cross-entropy for categorical and BCE for binary features). Parameter counts range from $\sim$160k (ACS, $d = 98$) to $\sim$360k (MIMIC-III, $d = 414$), varying only due to input dimensionality. Training uses 30 epochs, batch size 512, learning rate 10^{-3} and KL weight $\beta = 0.1$ for both baseline and DP models. For the model size experiment (Section 4.2), we additionally test two smaller architectures on Adult: a medium model ([64, 32] $\rightarrow$ 8-dim latent, $\sim$25k parameters) and a small model ([32, 16] $\rightarrow$ 4-dim latent, $\sim$11k parameters).

3.3 DP-SGD Configuration

The primary cross-dataset comparison uses $\epsilon \approx 10$ with $\delta = 10^{-5}$, representing moderate privacy. Gradient clipping norm is fixed at $C = 5.0$ throughout. The noise multiplier σ is calibrated for each (N, epochs) combination using the PRV accountant to hit the target ϵ; because the number of training steps $T = N \cdot E/B$ changes with N, σ must be recalibrated at each dataset size. For the epsilon-boundary experiments (Section 4.3), we sweep $\epsilon \in \{1, 2, 3, 5, 10, 20\}$ on Adult, $\epsilon \in \{1, 2, 5, 10\}$ on ACS Income and $\epsilon \in \{1, 2, 5, 10\}$ on NHANES.

3.4 Evaluation Protocol

For each configuration, the trained VAE generates N synthetic samples. An XGBoost classifier (default hyperparameters) is then trained on the synthetic data and evaluated on the held-out real test set. Each configuration is run with 3 independent seeds (42, 123, 456), each training the VAE from scratch and generating synthetic data; a single XGBoost classifier (default hyperparameters) is trained per synthetic dataset. We report mean $\pm$ standard deviation over the 3 seeds. Error bars in all figures reflect this cross-seed variability. A model is classified as *viable* when its normalised DP signal (Section 2) reaches 0.5, meaning the DP model recovers at least half the predictive signal of the best non-private synthetic data. The baseline at each N uses the identical architecture and training procedure without DP noise. We define N_{viable} as the smallest evaluated N at which the mean normalised signal across seeds exceeds 0.5. Where the boundary falls between two evaluated grid points, we report the bracketing range (e.g. [20k, 25k]). All experiments ran on cloud GPUs with parallel execution across N values and seeds.

4 RESULTS

We present results in three parts: establishing the viability boundary across datasets, testing whether model capacity shifts it, and characterising how the boundary scales with privacy budget ϵ.

4.1 THE VIABILITY BOUNDARY IS CONSISTENT

Figure 1 plots TSTR AUC against training set size N for all six datasets, comparing baseline (no DP) and DP models at $\epsilon \approx 10$. Two patterns are immediately visible.

First, every dataset exhibits two distinct thresholds: a lower N at which the baseline begins to learn and a higher N at which the DP model follows. The gap between these thresholds is the data cost of privacy. The transitions are clear; models jump from near-random AUC ($\sim$0.50) to near-baseline performance over a narrow range of N, rather than improving gradually.

Second, at sufficient N the DP model closes the gap to within noise of the baseline on several datasets. On ACS Income at $N = 100$k and Diabetes-130 at $N = 71$k the DP model slightly outperforms the baseline (by 0.010 and 0.037 AUC respectively), likely because DP noise acts as mild regularisation, though the differences are small enough to be within experimental variance.

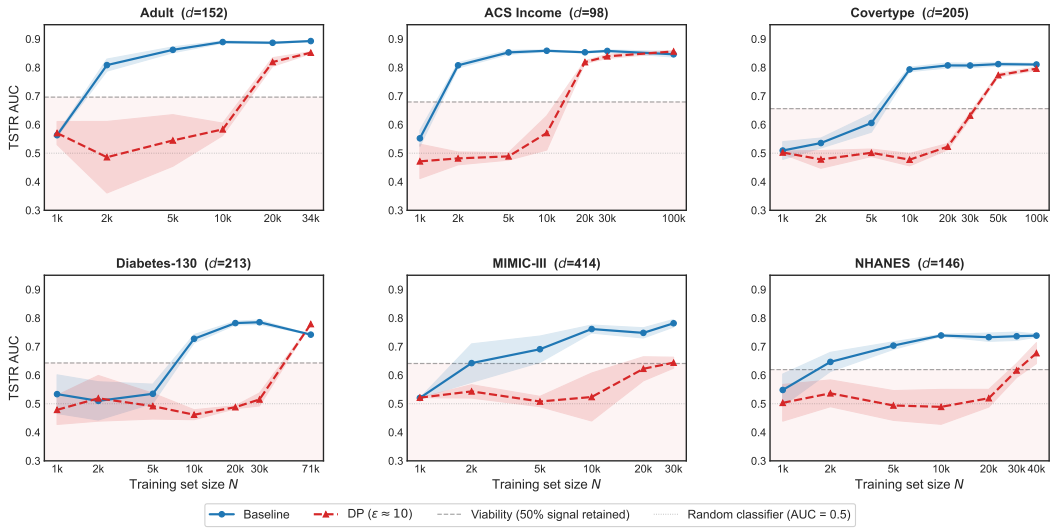

Figure 1: Viability boundary across six datasets. Baseline (solid) and DP at $\epsilon \approx 10$ (dashed) TSTR AUC vs training set size N. Error bars: ± 1 std over 3 seeds. All datasets show a clear transition from random-chance to near-baseline performance. DP requires more data but approaches baseline at high N.

Table 2: Cross-dataset viability boundary summary at $\epsilon \approx 10$. N_{viable}: smallest N where normalised DP signal ≥ 0.5 (DP recovers $\geq$50% of peak baseline signal). Data cost: ratio of DP to baseline boundary ($N_{\text{viable}}/N_{\text{BL}}$).

Dataset	Domain	d	BL Boundary	DP Boundary	N/d	Data Cost
Adult	Demographics	152	$\sim$2k	$\sim$20k	$\sim$132	$\sim$10$\times$
ACS Income	Survey	98	$\sim$10k	$\sim$30k	$\sim$306	$\sim$3$\times$
Covertype	Ecology	205	$\sim$10k	$\sim$50–100k	$\sim$244–488	$\sim$5–10$\times$
Diabetes-130	Clinical	213	$\sim$10k	$\sim$71k	$\sim$333	$\sim$7$\times$
MIMIC-III	Clinical	414	$\sim$5k	$\sim$30k†	$\sim$72	$\sim$6$\times$
NHANES	Public health	146	$\sim$5k	$\sim$32k	$\sim$219	$\sim$6$\times$

†Just reaches viability (normalised signal 51%) at dataset maximum.

Table 2 summarises the viability boundaries. All six datasets reach or approach viability at $\epsilon \approx 10$, though the two highest-dimensional health datasets are borderline: MIMIC-III ($d = 414$) just reaches viability at its full $N = 30$k (signal 51%), while NHANES ($d = 146$) crosses viability between $N = 30$k and 40k. These cases illustrate how close real healthcare datasets sit to the boundary, even though they were not designed with DP in mind.

N/d as a predictor. When the normalised DP signal is plotted against the samples-per-dimension ratio N/d (Figure 2), the six datasets collapse onto a common S-shaped trajectory despite spanning different domains, dimensionalities ($d = 98$ to 414) and classification tasks. Below $N/d \approx 50$, DP synthetic data is uniformly useless. Between $N/d \approx 50$ and 300, models transition from collapsed to viable. Above $N/d \approx 300$, DP is consistently viable.

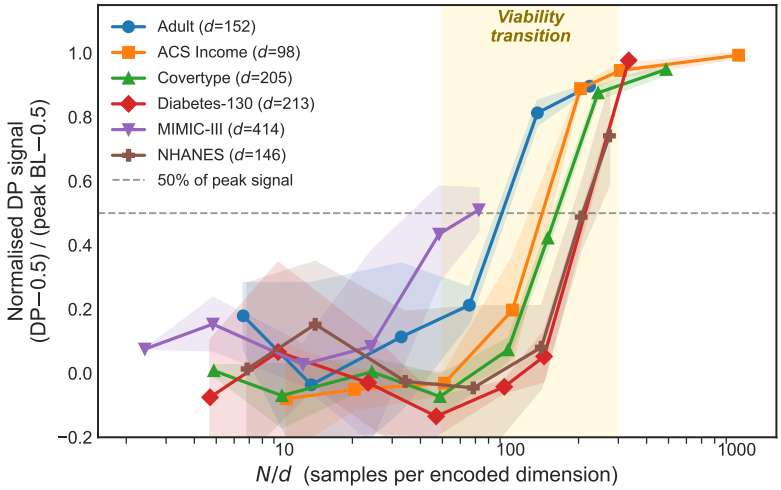

Figure 2: Normalised DP signal, defined as $(\text{DP AUC} - 0.5)/(\text{peak baseline AUC} - 0.5)$, plotted against the samples-per-dimension ratio N/d for all six datasets. A value of 1 means the DP model recovers the full discriminative signal of the best baseline; 0 means random-chance performance. Despite spanning different domains, dimensionalities ($d = 98$ to 414) and classification tasks, the datasets follow a common S-shaped trajectory. The shaded band marks the viability transition zone ($N/d \approx 50$–300), below which DP synthetic data carries negligible or highly degraded utility.

Controlled validation via dimension reduction. The cross-dataset collapse in Figure 2 could reflect confounds beyond N/d. To isolate the effect of d, we repeat the Adult experiment with a reduced feature set: dropping four high-cardinality categoricals (native_country, occupation, education, relationship) reduces d from 152 to 73. At $N = 5$k, both feature sets produce random-chance DP output ($N/d = 33$ and 68). At $N = 15$k, the reduced model ($N/d = 205$, TSTR AUC 0.817) exceeds the full model ($N/d = 99$, AUC 0.805), though the difference is within noise. At $N = 30$k both converge (0.850 vs 0.843). The boundary shifts to lower N with reduced d as predicted, confirming that N/d drives viability rather than N or d alone.

4.2 The Boundary Is Insensitive to Model Size

To test whether model capacity drives the viability boundary, we train three MLP VAE architectures on Adult at $\epsilon \approx 10$: the standard large model (157k parameters), a medium variant (25k) and a small variant (11k), spanning a $14\times$ range. Table 3 shows the results.

All three architectures transition in the same N bracket: collapsed at $N = 10$k, viable at $N = 15$k. The $14\times$ difference in parameter count does not shift the boundary location. Near the boundary, the smaller models achieve slightly higher AUC at $N = 15$k (medium: 0.860 vs large: 0.805), likely because fewer parameters reduces the dimensionality of the noise subspace relative to the gradient signal, and the smaller hypothesis class is less prone to overfitting the DP noise. However, the boundary *location* is the same.

This suggests the viability threshold is driven by data sufficiency rather than model capacity, at least within the MLP VAE family. This is consistent with Sander et al. (2023), who show that the DP-SGD privacy budget depends on total injected noise rather than architectural details. We note the limitation: this experiment covers a single dataset (Adult) and a single architecture family. De et al. (2022) show that scaling model size together with pre-training improves DP accuracy in discriminative image models; whether pre-trained generative models exhibit different boundary behaviour is an open question.

Table 3: DP TSTR across model sizes on Adult. All three models transition in the same N bracket despite 14× size difference. ✓ = viable (normalised signal ≥ 0.5, i.e. AUC ≥ 0.695); × = collapsed.

Architecture	Params	N=5k	N=10k	N=15k	N=20k
Large [256,128]→32	157k	0.379 ×	0.554 ×	0.805 ✓	0.834 ✓
Medium [64,32]→8	25k	0.418 ×	0.683 ×	0.860 ✓	0.844 ✓
Small [32,16]→4	11k	0.477 ×	0.527 ×	0.749 ✓	0.834 ✓
Baseline (large)		0.851	0.885	0.886	0.890

4.3 The Data Cost of Privacy Is Sublinear

Section 2.2 presents two theoretical predictions for how N_{viable} scales with noise. The per-step SNR argument predicts $N_{\text{viable}} \propto \sigma^2$; the formal DP-ERM bound predicts the gentler $N_{\text{viable}} \propto \sigma$. On Adult, the noise multiplier increases from $\sigma \approx 0.8$ at $\epsilon = 10$ to $\sigma \approx 3.7$ at $\epsilon = 1$ (both calibrated at reference $N = 20$k), a ratio of $4.6\times$. Because σ is recalibrated at each N to hit the target ϵ (Section 3), the actual σ at the boundary point differs slightly; we use a fixed reference N for comparability across ϵ. The per-step prediction expects $\sim 20\times$ more data; the DP-ERM prediction expects $\sim 4.6\times$. We test these predictions by sweeping ϵ on Adult, ACS Income and NHANES.

Table 4: Epsilon-boundary on Adult. 3-seed validation for $\epsilon \leq 3$; seed 42 for $\epsilon \geq 5$. σ at reference $N = 20$k. DP-ERM prediction ($N_{\text{viable}} \propto \sigma$) anchored at $\epsilon = 5$.

ϵ	σ	N_{viable}	Cost vs $\epsilon \geq 5$	DP-ERM
≥5	≤1.1	∼12k	1×	12k
3	1.5	∼18k	∼1.5×	∼16k
2	2.1	[20k, 25k]	∼1.9×	∼23k
1	3.7	[30k, 40k+]	∼2.5×	∼40k

Adult ($d = 152$). Table 4 shows the epsilon-boundary results. For $\epsilon \geq 5$ the boundaries are indistinguishable at ∼12k; moderate privacy is effectively free in terms of data. Below $\epsilon = 5$ the boundary shifts rightward monotonically (∼18k at $\epsilon = 3$, ∼22k at $\epsilon = 2$, ∼30k at $\epsilon = 1$), but the increase is gentle compared to the DP-ERM prediction, which would require ∼40k records at $\epsilon = 1$. Even $\epsilon = 1$ is achievable with ∼30k samples, only $\sim 2.5\times$ the $\epsilon \geq 5$ requirement.

Three-seed validation confirms these results are robust: at $\epsilon = 2$ the cross-seed standard deviation is 0.028–0.036, while $\epsilon = 1$ is noisier (std 0.127 at $N = 30$k), consistent with the higher variance expected under stronger noise.

Table 5: Epsilon-boundary comparison across datasets. For $\epsilon \geq 2$, the data cost is negligible. Only $\epsilon = 1$ requires more data.

ϵ	Adult (d=152)	ACS Income (d=98)
≥5	∼12k	[10k, 30k]
2	[20k, 25k]	≤30k (= ϵ=10)
1	[30k, 40k+]	[30k, 50k]

Cross-dataset validation. Table 5 compares the epsilon boundaries on Adult and ACS Income. Both datasets show the same pattern: for $\epsilon \geq 2$ the boundaries cluster near the $\epsilon = 10$ value, and only $\epsilon = 1$ requires meaningfully more data. ACS Income shows an even gentler scaling, with $\epsilon \geq 2$ all viable at the same N as $\epsilon = 10$; its lower dimensionality ($d = 98$ vs 152) may contribute. On NHANES, $\epsilon = 1$ ($\sigma \approx 3.05$) never reaches viability even with all available data, while $\epsilon \geq 2$ remains viable (Figure 3).

We do not claim a specific scaling exponent; the data are too sparse to distinguish between, say, $\sigma^{0.5}$ and $\sigma^{0.8}$. The qualitative conclusion is clear: the data cost of privacy is sublinear even relative to the formal DP-ERM prediction of $N_{\text{viable}} \propto \sigma$. For $\epsilon \geq 2$, the cost is negligible. Only $\epsilon = 1$ carries a measurable cost, and it is modest ($\sim 2.5\times$ on Adult).

5 WHY PRIVACY COSTS LESS DATA THAN EXPECTED

The per-step SNR argument (Section 2.2) predicts $N_{\text{viable}} \propto \sigma^2$ ($\sim$20$\times$); formal DP-ERM bounds predict $N_{\text{viable}} \propto \sigma$ ($\sim$4.6$\times$). We observe $\sim$2.5$\times$. Three complementary mechanisms account for the progression from per-step to formal to observed scaling. This is a qualitative account, not a formal proof; each mechanism is individually well-understood, but their joint interaction under DP-SGD has not been analysed.

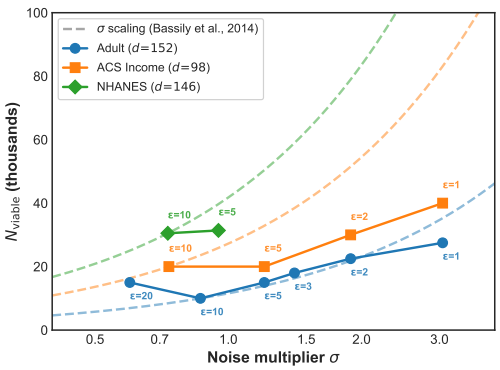

Figure 3: Observed N_{viable} vs noise multiplier σ (log scale) for Adult, ACS Income and NHANES, compared against DP-ERM σ scaling (dashed). DP-ERM bounds predict $\sim$4.6$\times$ more data from $\epsilon = 10$ to $\epsilon = 1$; in practice the cost is $\sim$2.5$\times$, sublinear even relative to this tighter bound. NHANES at $\epsilon = 1$ never reaches viability. Per-epsilon curves: Appendix A.

Noise averaging over gradient steps. The per-step argument treats each gradient step independently, but the model accumulates learning over $T = N \cdot E/B$ steps. Since DP noise is independent across steps, it cancels as $1/\sqrt{T}$. More data means more steps at fixed epochs, so the effective noise after T steps scales as $\sigma/\sqrt{T} \propto \sigma/\sqrt{N}$. This changes the viability condition from $N_{\text{viable}} \propto \sigma^2 d$ to $N_{\text{viable}} \propto \sigma\sqrt{d}$ (recall $p \propto d$), recovering the scaling predicted by formal DP-ERM bounds (Bassily et al., 2014). Under this scaling, the σ ratio of 4.6$\times$ predicts an N ratio of $\sim$4.6$\times$. Our observed $\sim$2.5$\times$ is gentler still, suggesting additional factors.

Shrinking relative noise. What determines learning is not the absolute DP noise but noise relative to the stochastic gradient variance. At moderate σ (e.g. $\epsilon \geq 5$), the DP noise is already comparable to or smaller than the inherent stochastic gradient noise; further reducing σ has negligible effect. This explains the $\epsilon \geq 5$ clustering observed in Section 4.3: the bottleneck shifts from DP noise to finite-sample estimation error, and privacy becomes free in the sense that it is no longer the binding constraint. This is consistent with recent theoretical work showing that privacy cost vanishes in the overparameterised regime with sufficient data (Bombari & Mondelli, 2025).

Clipping and convergence dynamics. DP-SGD clips per-sample gradients to norm C. Early in training, gradients are large and heavily clipped, discarding signal. As the model converges, gradient norms shrink below C and clipping becomes non-binding. With more data the model reaches this low-gradient regime sooner, creating a positive feedback loop: more data leads to faster convergence, less clipping and better SNR. The combined effect of noise averaging, shrinking relative noise and convergence dynamics produces the observed sublinear scaling.

Limitations. We cannot distinguish which factor dominates. The $\sqrt{T}$ averaging alone gets within $\sim$2$\times$ of the observed scaling; the remaining gap could be explained by any combination of the other factors. The account assumes Adam-style optimisation with a fixed learning rate; different optimisers or schedules could change the dynamics. A formal treatment would require characterising the convergence rate of DP-SGD for VAE objectives, which remains an open problem.

6 RELATED WORK

DP and data requirements. DP learning needs "much more data" or better features to match non-private accuracy (Tramèr & Boneh, 2021), but this finding concerns discriminative models at fixed N rather than generative models across dataset sizes. Practical guides for DP-SGD tuning note that sufficient data is needed for good privacy-utility tradeoffs (Ponomareva et al., 2023), without quantifying the N–ϵ relationship.

DP synthetic data. Several benchmarks compare DP generators on tabular data. Ganev et al. (2024) vary ϵ and N jointly, finding non-monotonic behaviour and identifying minimum dataset sizes per generator (DP-WGAN $\geq$4k, PATE-GAN $\geq$8k), but do not systematically map the boundary as a function of ϵ. Other benchmarks (Tao et al., 2022; Bowen & Liu, 2020) compare generators across settings but do not map the collapse boundary as a function of N. On the evaluation side, TSTR (Esteban et al., 2017) remains the standard utility metric for synthetic data, which we adopt throughout. Among generators, AIM (McKenna et al., 2022), a marginal-based DP method, consistently outperforms other marginal-based approaches; we present a preliminary comparison in Appendix B. Without DP, synthetic data has been shown to leak membership information (Stadler et al., 2022), motivating the practical importance of understanding the data cost of adding formal guarantees.

Scaling laws and theory. Recent work derives scaling laws for DP training. McKenna et al. (2025) use a noise-batch ratio framework for language models, showing that data and privacy budgets interact synergistically; though their setting (LLM loss curves) differs from synthetic data viability thresholds, it provides theoretical grounding for sublinear scaling. The total-noise perspective of Sander et al. (2023), who derive scaling laws that reduce hyperparameter search cost by $100\times$, helps explain why per-step σ^2 analysis overestimates the data cost we observe. In the overparameterised regime, Bombari & Mondelli (2025) prove that privacy cost vanishes when the model is sufficiently overparameterised relative to the training set, consistent with our finding that $\epsilon \geq 5$ boundaries converge. On the theoretical side, DP-ERM excess risk scales as $\sqrt{d}/(n\epsilon)$ (Bassily et al., 2014), implying $N_{\text{viable}} \propto \sigma$ (a $\sim$4.6$\times$ data cost from $\epsilon = 10$ to $\epsilon = 1$). Our observed $\sim$2.5$\times$ is modestly below this bound, a more expected theory-to-practice gap than the $\sim$20$\times$ from per-step analysis. Complementary lower bounds confirm the dimension dependence: counting queries under DP require sample size $\propto \sqrt{d}$ (Bun et al., 2014), and minimax optimal rates for private estimation exhibit the same structure (Cai et al., 2021).

Healthcare DP. A recent review of 74 healthcare DP studies finds that strict privacy often degrades accuracy, particularly for smaller or more complex datasets (Mohammadi et al., 2026); our work quantifies what "smaller" means. DP also has disparate impact on underrepresented subgroups (Bagdasaryan et al., 2019); since subgroup size is effectively a small-N problem, our viability boundary framework could be applied per-subgroup.

Positioning. No prior work has empirically mapped $N_{\text{viable}}(\epsilon)$ for DP synthetic data, shown the scaling is sublinear relative to σ^2 theory, or demonstrated consistent viability boundaries across diverse datasets. The practical finding that strong privacy costs $\sim$2.5$\times$ the data, below even the $\sim$4.6$\times$ from formal DP-ERM theory, is new.

7 Discussion

Our N/d heuristic makes the recommendation of Tramèr & Boneh (2021) concrete. If $N/d < 100$, DP-SGD is unlikely to produce viable synthetic data. Consider dimension reduction, PATE-based mechanisms or federated approaches instead. In the transition zone ($N/d \approx$ 100–300), DP may work at moderate privacy ($\epsilon \geq 5$) but strong privacy ($\epsilon \leq 2$) may require more data. Above $N/d \approx 300$, DP-SGD is likely viable even at strong privacy levels. Strong privacy is cheaper than commonly assumed: $\epsilon = 1$ costs only $\sim$2.5$\times$ more data than $\epsilon = 10$, and for $\epsilon \geq 5$ the cost is effectively zero. Reducing d is at least as valuable as increasing N: on Diabetes-130, binning ICD-9 codes reduced d from 2,475 to 213, making DP viable at $N =$ 71k, and the controlled d-reduction experiment on Adult (Section 4.1) confirms this effect directly. This aligns with the theoretical prediction: Bun et al. (2014) and Cai et al. (2021) show that DP sample complexity grows with $\sqrt{d}$, so reducing dimensionality directly lowers the data requirement. Recent theory strengthens this connection: He et al. (2025) show that DP synthetic data generation can achieve error rates depending on the intrinsic dimension of the data rather than the ambient dimension, and Zhou et al. (2021) show that projecting gradients onto the dominant subspace allows DP-SGD to bypass the ambient-dimension dependence. Our empirical findings with standard DP-SGD tooling on real datasets are consistent with these predictions. The encoded dimension d serves as a practical

upper bound on the effective complexity, and the width of the transition band ($N/d \approx 50$–300) may partly reflect variation in how tightly ambient dimension tracks intrinsic dimension across datasets.

The healthcare case is acute. A recent review of 74 studies finds that strict privacy often degrades accuracy, particularly for smaller or more complex datasets (Mohammadi et al., 2026), but leaves the threshold unquantified. Our results answer their open question: the binding constraint is not N alone but N/d. MIMIC-III has 30k records, enough for many ML tasks, but its 414 encoded features place it at $N/d = 72$, just above the collapse zone. The subgroup dimension compounds this. Bagdasaryan et al. (2019) showed that DP has disparate impact on underrepresented groups. Since a subgroup of size N_s has effective ratio N_s/d, minorities within an already-borderline dataset may fall below the viability boundary even when the full dataset is above it. Per-subgroup viability analysis is a natural extension of our framework.

Above $N/d \approx 300$, DP matches or slightly exceeds the baseline on several datasets, and $\epsilon \geq 5$ costs no additional data, consistent with theoretical predictions that privacy cost vanishes in the overparameterised regime (Bombari & Mondelli, 2025). This shifts the practical question from "can we afford privacy" to "what ϵ do we need." The viability boundary is mechanism-dependent. We ran AIM (McKenna et al., 2022) on Adult with 20 seeds across the transition zone and find mean viability at $N \approx 100$ ($N/d < 1$), roughly $100\times$ lower than DP-SGD on the same dataset, though with high run-to-run variance (Appendix B).

Limitations. All experiments use a single model family (MLP VAE). De et al. (2022) show that pre-trained models combined with scale can substantially improve DP accuracy in discriminative settings; pre-trained generative models may shift the boundary, but this remains untested. The cross-dataset comparison uses $\epsilon = 10$; the epsilon sweep covers three datasets. The model size and dimension reduction experiments each cover one dataset. We use TSTR AUC as the sole utility metric, and the 50% normalised signal threshold is somewhat arbitrary. However, the sharp S-shaped transition (Figure 2) means the exact threshold matters less than the transition zone itself. A more fundamental concern is downstream validity: Montoya Perez et al. (2024) show that DP synthetic data at low ϵ can inflate Type I error rates, suggesting that TSTR viability may be necessary but not sufficient for valid statistical inference. The AIM comparison (Appendix B) covers one dataset; we did not test PATE or DP-FTRL. Training used 30 epochs throughout, which may not be optimal for all datasets.

8 Conclusion

The prevailing assumption that differential privacy requires prohibitive amounts of data has slowed adoption, particularly in healthcare where data is scarce and privacy is non-negotiable. Our results replace this assumption with a quantitative answer: check N/d. Below $\sim$100, DP-SGD synthetic data is not viable; above $\sim$300, it works even at strong privacy. The data cost of moving from $\epsilon = 10$ to $\epsilon = 1$ is $\sim 2.5\times$, not the $\sim 20\times$ that naive theory suggests. For many real datasets, the binding constraint is not privacy but dimensionality, and feature engineering can shift the boundary more effectively than data collection.

These findings open several directions. Extending the viability boundary to other generative architectures (DP-CTGAN, diffusion models) would test whether N/d generalises beyond VAEs. Per-subgroup viability analysis could reveal whether demographic minorities face different boundaries, connecting privacy guarantees to fairness. On the theory side, tightening the link between DP-ERM bounds and observed sublinear scaling remains open. For healthcare specifically, multi-site aggregation strategies that increase effective N without centralising data offer a practical path for datasets that currently fall below the boundary.

LLM Usage Disclosure

We used an LLM tool for A) editing prose for clarity and concision and B) generating implementation scaffolding (such as boilerplate code, configuration templates, plotting/figure scripts and routine processing utilities). All scientific contributions (problem formulation, hypotheses, experimental design, dataset selection, model architecture, DP-SGD integration, preprocessing choices and interpretation of results) were developed by the authors. All LLM-assisted text and code were reviewed, edited and validated by the authors.

ACKNOWLEDGEMENTS

We thank Bogdan Kulynych for suggesting the AIM comparison experiments. This work was supported by the Technology Development Fund (Tækniþróunarsjóður) under grants 2423970-601 and 2525674-601.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, 2016.

John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. The 2020 Census disclosure avoidance system TopDown algorithm. *Harvard Data Science Review*, (Special Issue 2), 2022. doi: 10.1162/99608f92.529e3cb9.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 464–473. IEEE, 2014. doi: 10.1109/FOCS.2014.56.

Simone Bombari and Marco Mondelli. Privacy for free in the overparameterized regime. *Proceedings of the National Academy of Sciences*, 122(15):e2423072122, 2025. doi: 10.1073/pnas.2423072122.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173.

Claire McKay Bowen and Fang Liu. Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2):280–307, 2020. doi: 10.1214/19-STS742.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 1–10, 2014. doi: 10.1145/2591796.2591877.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *Annals of Statistics*, 49(5):2825–2850, 2021. doi: 10.1214/21-AOS2058.

Centers for Disease Control and Prevention. National Health and Nutrition Examination Survey (NHANES). `https://www.cdc.gov/nchs/nhanes/`, 2024.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162. PMLR, 2022.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*, volume 9. 2014. doi: 10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography Conference*, pp. 265–284, 2006.

Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017.

Georgi Ganev, Kai Xu, and Emiliano De Cristofaro. Graphical vs. deep generative models: Measuring the impact of differentially private mechanisms and budgets on utility. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2024. doi: 10.1145/3658644.3690215.

Mauro Giuffrè and Dennis L. Shung. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. *npj Digital Medicine*, 6:186, 2023. doi: 10.1038/s41746-023-00927-3.

Yiyun He, Thomas Strohmer, Roman Vershynin, and Yizhe Zhu. Differentially private low-dimensional synthetic data from high-dimensional datasets. *Information and Inference: A Journal of the IMA*, 14(1):iaae034, 2025. doi: 10.1093/imaiai/iaae034.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality*, 11(3), 2021. doi: 10.29012/jpc.778.

Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. AIM: An adaptive and iterative mechanism for differentially private synthetic data. *Proceedings of the VLDB Endowment*, 15(11): 2599–2612, 2022.

Ryan McKenna, Yangsibo Huang, Amer Sinha, Borja Balle, Zachary Charles, Christopher A. Choquette-Choo, Badih Ghazi, Georgios Kaissis, Ravi Kumar, Ruibo Liu, Da Yu, and Chiyuan Zhang. Scaling laws for differentially private language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Ilya Mironov. Rényi differential privacy. *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017.

Marziyeh Mohammadi, Mohsen Vejdanihemmat, Mahshad Lotfinia, Mirabela Rusu, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. Differential privacy for medical deep learning: methods, tradeoffs, and deployment implications. *npj Digital Medicine*, 9:93, 2026. doi: 10.1038/s41746-025-02280-z.

Ileana Montoya Perez, Parisa Movahedi, Valtteri Nieminen, Antti Airola, and Tapio Pahikkala. Does differentially private synthetic data lead to synthetic discoveries? *Methods of Information in Medicine*, 63(01/02):035–051, 2024. doi: 10.1055/a-2385-1355.

Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H. Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. How to DP-fy ML: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77:1113–1201, 2023. doi: 10.1613/jair.1.14649.

W. Nicholson Price and I. Glenn Cohen. Privacy in the age of medical big data. *Nature Medicine*, 25(1):37–43, 2019. doi: 10.1038/s41591-018-0272-7.

Tom Sander, Pierre Stock, and Alexandre Sablayrolles. TAN without a burn: Scaling laws of DP-SGD. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 29937–29949. PMLR, 2023.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium*, pp. 1451–1468, 2022.

Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014, 2014.

Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Benchmarking differentially private synthetic data generation algorithms. In *AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2022.

Florian Tramèr and Dan Boneh. Differentially private learning needs better features (or much more data). In *International Conference on Learning Representations*, 2021.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32, 2019.

Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, 2020.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.

Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private SGD with gradient subspace identification. In *International Conference on Learning Representations*, 2021.

APPENDIX

A PER-EPSILON TSTR CURVES

Figure 4 shows the full per-epsilon TSTR curves underlying the summary in Section 4.3 and Figure 3.

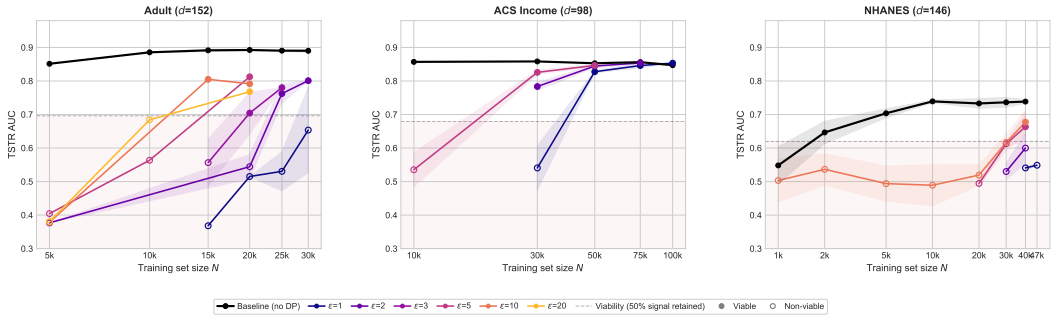

Figure 4: TSTR AUC vs training set size N at multiple privacy budgets ($\epsilon = 1, 2, 3, 5, 10, 20$) for Adult, ACS Income, and NHANES. Filled markers indicate viable configurations (above the 50% signal threshold); open markers indicate non-viable. The epsilon curves cluster tightly: for $\epsilon \geq 2$, the viability boundary is nearly identical to $\epsilon = 10$. Only $\epsilon = 1$ requires meaningfully more data. See Figure 3 for the summary.

B AIM: A MARGINAL-BASED BASELINE

We focus on DP-SGD throughout the main paper because it is the general-purpose DP training mechanism: it handles continuous features natively, scales to high-dimensional data and underpins most deep generative models. However, for low-dimensional tabular data, marginal-based methods like AIM (McKenna et al., 2022) avoid the gradient-noise bottleneck by operating on discretised contingency tables rather than model parameters.

To test whether the viability boundary generalises across mechanism classes, we ran AIM on Adult at $\epsilon = 10$, $\delta = 10^{-5}$ with 17–20 seeds per N value. Continuous features were quantile-binned into 20 bins (fitted on the full training set for stability), giving an effective dimensionality of $d_{\text{eff}} = 151$. Table 6 reports the distribution of TSTR AUC across seeds.

Table 6: AIM TSTR AUC on Adult ($\epsilon = 10$, $d_{\text{eff}} = 151$) across 20 seeds. DP-SGD on the same dataset requires $N \approx$ 10k–15k for viability; AIM reaches mean viability (≥ 0.7) at $N = 100$.

N	N/d	Mean	Std	Min	Median	Max
30	0.2	0.520	0.116	0.368	0.516	0.786
50	0.3	0.563	0.116	0.329	0.542	0.773
75	0.5	0.640	0.126	0.399	0.680	0.796
100	0.7	0.698	0.085	0.507	0.715	0.844
125	0.8	0.720	0.079	0.560	0.732	0.840
150	1.0	0.742	0.049	0.610	0.751	0.808
200	1.3	0.740	0.059	0.599	0.766	0.802
300	2.0	0.722	0.080	0.500	0.740	0.828
500	3.3	0.728	0.065	0.540	0.750	0.795

AIM's mean viability boundary is at $N \approx 100$, roughly $100\times$ lower than DP-SGD on the same dataset. At $N/d < 1$, AIM already produces useful synthetic data, whereas DP-SGD requires $N/d \approx$ 100–300. This confirms that the viability thresholds in the main paper are specific to DP-SGD, not to differential privacy itself.

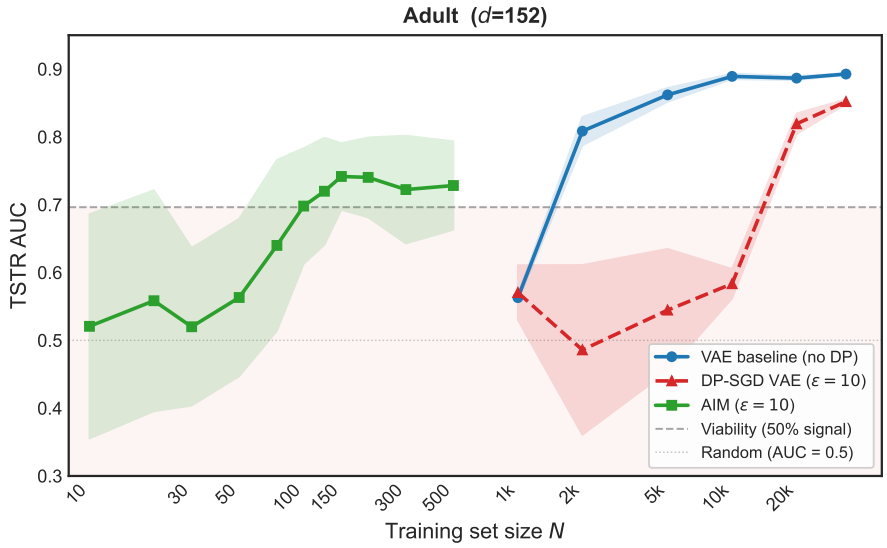

Figure 5: AIM vs DP-SGD viability boundary on Adult ($\epsilon = 10$). AIM (green) crosses the viability threshold at $N \approx 100$; DP-SGD (red, dashed) requires $N \approx 10$k–15k. Shaded regions show ± 1 std across seeds.

Run-to-run variance is high at low N and drops at $N \approx 150$: standard deviation falls from 0.12 (at $N \leq 75$) to 0.05 (at $N = 150$). AIM selects marginal queries adaptively; at low N, the quality of the output depends on which marginals happen to be selected. This variance is qualitatively different from DP-SGD, where noise is applied uniformly to all gradient coordinates.

Even at $N = 500$ ($N/d = 3.3$), the worst of 17 seeds scores 0.54, below viability. AIM is viable *on average* from $N \approx 100$ but not *reliably*: no N tested produces viable output on every seed. For practitioners requiring consistent output quality, marginal-based methods may need multiple runs with quality checks.

That AIM with DP outperforms a non-DP VAE at comparable N (Figure 5) reflects the generator architecture, not the privacy mechanism. Marginal-based methods estimate low-order statistics directly from data, which requires far fewer samples than training a neural network. The DP-SGD viability boundary reported in the main paper therefore conflates two costs: the data needed to train the generator and the additional data needed to absorb DP noise. AIM eliminates the first.