A Rising Opportunity from Synthetic Data: Generative Data Science

Guang Cheng*

February 3, 2026

Abstract

Statistics has historically been grounded in the analysis and modeling of real world data produced by human activities, scientific experiments or natural processes. The rapid rise of AI is fundamentally altering this landscape: *synthetic data* - generated by algorithms rather than observed from reality - is becoming a new data source in science, industry, and policy. This shift challenges long-standing assumptions about what data are, how they should be evaluated, and what it means to draw statistical insights from them.

This position paper examines emerging phenomena, articulates key conceptual challenges, and outlines open research questions in synthetic data. It calls for the recognition of a new research frontier - Generative Data Science - devoted to the statistical principles underlying the generation, evaluation, and governance of synthetic data. I argue that synthetic data necessitates a fundamental rethinking of core statistical notions, including fidelity, utility, privacy, and trust, and that these dimensions are intrinsically coupled rather than independently optimizable. Taken together, these observations and considerations motivate a generative paradigm for statistical modeling and inference as an essential pillar of modern statistical science [12, 55].

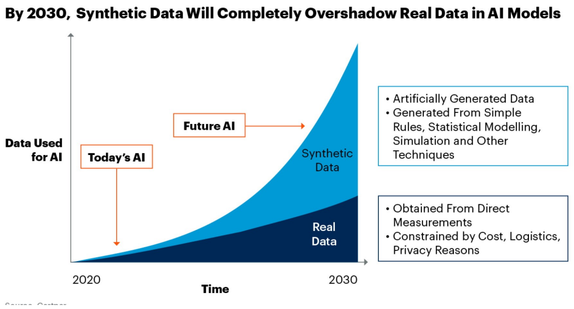
Key Words: Synthetic Data, Generative Data Science, Large Language Model, Privacy Preserving, Watermark

*Professor, Department of Statistics and Data Science, UCLA, CA, 90095. Email: guangcheng@ucla.edu

1 Introduction

What we now refer to as synthetic data has appeared for decades under various names and in multiple forms - from simulated data in the physical sciences, to imputed values in statistical analysis, to images generated by GANs. Recent attention to synthetic data is driven largely by the proliferation of AI-generated content, the growing demand for large and diverse datasets, and the increasing difficulty of sharing sensitive information. These pressures arise across a broad range of domains, including computer vision, natural language processing, network science, and tabular data analysis. Looking back to 2022, one already observes widespread adoption of synthetic data in both research and industry; looking ahead to 2030, its influence is poised to expand even further as a cornerstone of modern data science; see Figure 1.

Figure 1: Synthetic data include images, text, tables, graphs etc.



This position paper begins by highlighting two contemporary forces that make synthetic data increasingly necessary. The first is concerned with the (empirical) scaling law of large language models (LLMs) [19], which characterize LLM performance as

$$\text{LM loss } L(N, D) = \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + C,$$

where N denotes model size and D denotes the size of the pre-training dataset. Beyond increasing model capacity, the availability of sufficiently large training corpora constitutes a fundamental bottleneck in further scaling LLM performance under this law. Moreover, recent work [38] demonstrates that data quality also plays a crucial role: by removing "easy" training samples, the scaling behavior can be improved from a power-law regime to an exponential regime. Unfortunately, high-quality human-generated real data is rapidly running out by 2026–2032 as projected in [44]; see Figure 2. In this context, the synthesis of high-quality data emerges as a natural - and potentially indispensable - solution.

A second force stems from regulatory constraints, including the General Data Protection Regulation [13] and the California Consumer Privacy Act [5], which impose strict requirements on data collection, sharing, and processing. In sectors such as banking and digital marketing, synthetic tabular data has therefore emerged as a modern solution - preserving statistical properties while preventing the disclosure of sensitive individual information.

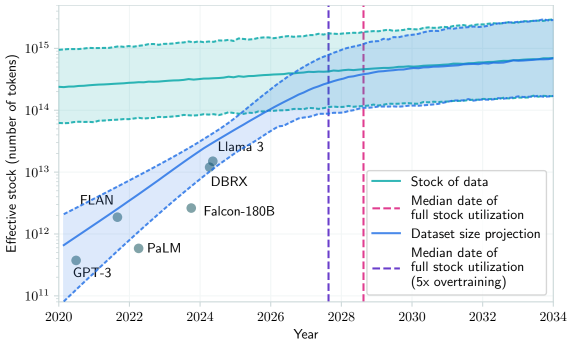
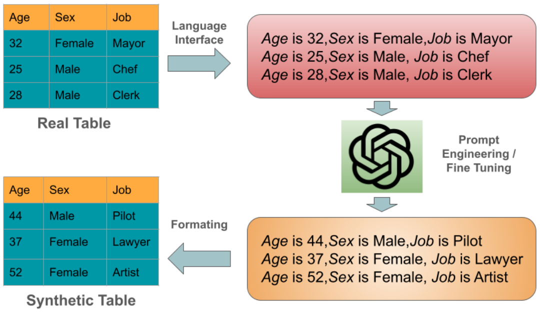


Figure 2: Projections of the effective stock of human-generated public text and dataset sizes used to train LLMs.

Technically, data synthesizers fall into three broad categories; for brevity, I focus on tabular data synthesizers: (i) classical (non-)parametric statistical models, such as PrivBayes [57, 27]; (ii) deep generative models, including GAN-based approaches such as CT-GAN [52] and diffusion-based methods such as AutoDiff [40, 41]; and (iii) language-model-based approaches, either as auto-regressive generators [25, 37], or as semantic encoders within hybrid generative pipelines, e.g., CTSyn [29]. Figure 3 illustrates the third paradigm. Somewhat surprisingly, pre-trained LLMs can generate tabular data with high statistical similarity - even in mixed-type settings with complex constraints - and can augment small datasets with diverse and realistic samples by leveraging web-scale knowledge and in-context learning capabilities.

From a classical information-theoretic perspective, one might expect synthetic data to offer little benefit, since it is ultimately generated from the raw data. This raises a natural question: *can artificial data truly create "something out of nothing"?* However, this intuition is difficult to reconcile with empirical observations, as numerous synthetic data solutions have already been successfully deployed in practice - for enhancing adversarial robustness [51], imputing missing values [34], preserving privacy [52], and improving both the utility and even the fairness of downstream tasks [54, 56, 3]. Nevertheless, several fundamental questions remain unresolved: *Can we reliably train models on synthetic data? To what extent might sensitive information in the original data be leaked through synthetic data? How should authorship and ownership of synthetic data be defined?* This position paper surveys recent progress (largely biased towards the author lab's works) in this still empirical landscape and argues that these reviews collectively point toward a promising new research frontier in statistics - what one may refer to as *Generative Data Science*.

Figure 3: LLMs can be adapted for synthetic tabular data generation.



Age	Sex	Job
32	Female	Mayor
25	Male	Chef
28	Male	Clerk

Real Table

Age is 32,*Sex* is Female,*Job* is Mayor
Age is 25,*Sex* is Male, *Job* is Chef
Age is 28,*Sex* is Male, *Job* is Clerk

Age is 44,*Sex* is Male,*Job* is Pilot
Age is 37,*Sex* is Female, *Job* is Lawyer
Age is 52,*Sex* is Female, *Job* is Artist

Age	Sex	Job
44	Male	Pilot
37	Female	Lawyer
52	Female	Artist

Synthetic Table

2 Large Language Models

Large language models (LLMs) have emerged as a major source of synthetic data. In this section, we use LLMs as a representative synthesizer to make the challenges of generative data science concrete in a broader context.

At a high level, an LLM defines a probabilistic model over sequences of discrete tokens and generates text autoregressively by repeatedly sampling the next token conditioned on previously generated tokens. Given a tokenized sequence $x_1, \ldots, x_T$, the model factorizes the joint distribution as

$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}),$$

where the conditional distributions are parameterized by a deep neural network trained on large corpora of human-generated text in a self-supervised manner. From a statistical standpoint, this formulation places LLMs within the classical framework of probabilistic generative modeling, albeit at an unprecedented scale and complexity.

The magic power of LLMs is closely tied to their ability to leverage massive pretraining datasets and to generalize through contextual learning. During generation, an LLM does not simply reproduce memorized training examples; rather, it synthesizes text by recombining learned linguistic, semantic, and structural patterns. This capability makes LLMs a particularly compelling engine for synthetic data generation. Recent empirical studies demonstrate that LLM-generated text can augment training corpora, enhance robustness, and improve downstream tasks such as reasoning and instruction following.

At the same time, the use of LLMs as text synthesizers raises fundamental challenges for statistical evaluation. Unlike classical simulated data, synthetic text lacks a natural notion of ground truth. Consequently, fidelity is often assessed using embedding-based similarity measures, such as BERTScore, which emphasize semantic alignment rather than exact distributional agreement. While these metrics have proven useful in practice, it is often unclear which aspects of the data-generating process they faithfully capture.

More broadly, evaluating LLM-generated text exposes a tension between surface-level similarity and functional utility. Text that appears fluent and semantically coherent may still exhibit hallucinations, biases, or subtle distributional shifts that affect downstream tasks. Conversely, text that deviates lexically from real data may nevertheless preserve task-relevant structure. These observations suggest that evaluation frameworks must move beyond perceptual similarity and toward criteria grounded in statistical, computational, and decision-theoretic considerations.

Finally, LLMs introduce unique challenges for privacy and governance in synthetic data generation. High-capacity autoregressive models are known to exhibit memorization under certain conditions, leading to potential leakage of sensitive training data. This phenomenon complicates the interpretation of privacy guarantees and motivates the development of auditing tools - such as membership inference attacks and watermarking - specifically tailored to discrete, sequential data. Understanding how privacy risks scale with model size, training data composition, and generation strategies remains an important open problem for generative data science.

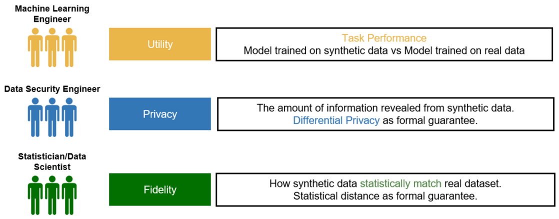
3 Generative Data Science

As illustrated by the case of LLMs, generative data science confronts a range of conceptual and technical challenges. Consequently, it remains premature to delineate the full scope of this emerging field. Accordingly, this position paper focuses on one foundational aspect: the *evaluation* of synthetic data.

Such evaluation is typically organized around three key dimensions - fidelity, utility, and privacy - as summarized in Figure 4. Fidelity assesses how well synthetic data statistically match real data; utility measures their effectiveness when used to train downstream models; and privacy quantifies the extent to which sensitive information from the original data may be inferred from the synthetic data.

3.1 Statistical Fidelity

Figure 4: Evaluation of synthetic data from three perspectives.


Machine Learning Engineer
Utility
Task Performance
Model trained on synthetic data vs Model trained on real data
Data Security Engineer
Privacy
The amount of information revealed from synthetic data.
Differential Privacy as formal guarantee.
Statistician/Data Scientist
Fidelity
How synthetic data statistically match real dataset.
Statistical distance as formal guarantee.

In computer vision, fidelity is commonly understood as the degree to which generated images resemble real images in terms of perceptual quality, realism, and accuracy. A widely used metric is the Fréchet Inception Distance (FID), which measures discrepancies between feature distributions extracted from a pre-trained Inception-v3 network. An analogous metric in natural language processing is the BERTScore, which relies on contextual embeddings derived from BERT [8]. Although these metrics - and their numerous variants - are broadly adopted in practice, their theoretical underpinnings remain limited. Existing theoretical analyses are largely synthesizer-specific; for example, [27] provides guarantees tailored to PrivBayes only. A general, systematic framework for fidelity evaluation is still lacking.

However, over the past two decades, the statistics literature has developed a rich set of tools for estimating distances between high-dimensional probability distributions, which are directly relevant to fidelity assessment in synthetic data. One recent example is the fidelity metric for synthetic tabular data proposed in [43], which is grounded in a statistical discrimination framework. This framework is generic in the sense that it can be applied to arbitrary embedding spaces while avoiding the curse of dimensionality. To fully realize its practical potential, however, additional engineering effort and adaptation across different data modalities remain necessary.

3.2 Machine Learning Utility

As discussed in the introduction, another primary use of synthetic data is to train downstream machine learning models. Both theoretical and empirical studies, e.g., [54, 30], provide evidence supporting its practical value. However, the examples below underscore that there is no single, universal notion of "utility." Rather, utility spans a spectrum of task- and metric-dependent evaluations.

For instance, [7] demonstrated that the perceived utility of synthetic data can depend strongly on the choice of evaluation metric. In their study, tabular data generated by Bayesian networks (BNs) and neural networks (NNs) yielded markedly different outcomes across metrics: BN-based synthetic data outperformed on F1 score and precision, whereas NN-based synthetic data achieved superior AUROC and recall. This example highlights that conclusions about utility are inherently tied to the downstream task and the criteria used for assessment.

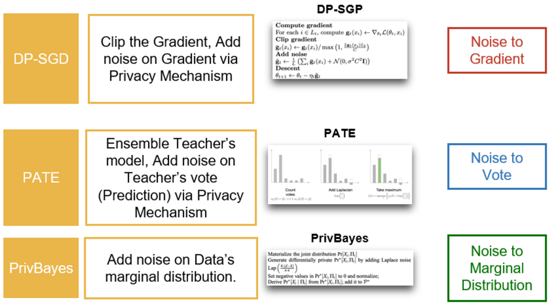
Beyond metric dependence, recent work has revealed more subtle and consequential challenges. A recent *Nature* paper [36] showed that generative models trained recursively on synthetic data are susceptible to *model collapse*, exhibiting progressively degraded performance over successive training iterations. Subsequent studies [11, 45, 10] further demonstrate that even partial inclusion of synthetic data in the training distribution can induce distributional drift, which in turn contributes to collapse. Notably, the proliferation of AI-generated content makes some degree of synthetic data contamination in large web-scraped training corpora increasingly unavoidable.

Encouragingly, recent works suggest that model collapse can be mitigated through careful management of training data. One strategy is to enlarge synthetic datasets according to a superlinear growth schedule [53]. Another is to incorporate fresh real data into recursive training loops [16]. More generally, these methods develop weighting schemes that optimally balance newly collected real samples against previously generated synthetic samples. Notably, the "golden ratio" derived theoretically in [16] has been empirically validated in a follow-up study [23].

3.3 Privacy-Preserving Capability

When synthetic data are shared, a fundamental question is how much sensitive information from the original training data can, in principle, be inferred by an adversary. *Differential privacy (DP)* provides a widely adopted formalism for reasoning about this risk by treating data release as an information-releasing mechanism and requiring that the presence or absence of any single individual in the training data has only a limited influence on the distribution of released outputs. Motivated by this principle, a growing line of work has proposed DP data synthesis algorithms that aim to preserve privacy by injecting calibrated noise into the *training or generation process*, including DP-GANs [58, 50] and DP diffusion models [9]. Figure 5 illustrates three representative mechanisms for incorporating noise into data synthesizers.

Figure 5: Three ways to preserve differential privacy in data synthesizers


DP-SGD
Clip the Gradient, Add noise on Gradient via Privacy Mechanism
DP-SGP
Compute gradient
For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$
Clip gradient
$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i)/\max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$
Add noise
$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$
Descent
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$
Noise to Gradient
PATE
Ensemble Teacher's model, Add noise on Teacher's vote (Prediction) via Privacy Mechanism
Count votes
Add Laplacian
Take maximum
Noise to Vote
PrivBayes
Add noise on Data's marginal distribution.
PrivBayes
Materialize the joint distribution $\Pr[X_i, \Pi_i]$
Generate differentially private $\Pr^*[X_i, \Pi_i]$ by adding Laplace noise
Set negative values in $\Pr^*[X_i, \Pi_i]$ to 0 and normalize;
Derive $\Pr^*[X_i \mid \Pi_i]$ from $\Pr^*[X_i, \Pi_i]$; add it to $\mathcal{P}^*$
Noise to Marginal Distribution

A critical caveat is that DP guarantees are typically established at the level of the parameters or training processes of data synthesizers, whereas how these guarantees translate to the released synthetic data themselves remains unclear - or, at least, insufficiently understood in a rigorous manner, even when invoking standard post-processing arguments; also see Section 7 of [22]. Bridging this gap requires a deeper understanding of the probability measure on synthetic datasets induced by the generator.

That said, privacy leakage from synthetic data can often be meaningfully analyzed in a task-specific manner. For example, when synthetic data are used to estimate a population-level statistic such as median income, it is natural to define and assess DP with respect to that particular task. This tension between task-agnostic privacy guarantees for synthesizers and task-specific privacy analyses for synthetic data highlights a fundamental conceptual challenge and motivates further theoretical development.

Complementing differential privacy, a widely used empirical approach for assessing privacy risks in synthetic data is through *membership inference attacks* (MIAs). MIAs formalize privacy leakage as an adversary's ability to infer whether a candidate record was included in the training set of a generative model. As such, MIAs provide an interpretable and computationally efficient lens for diagnosing privacy leakage in high-capacity generative models.

Formally, let $T = \{x_i\}_{i=1}^n$ be a training dataset sampled from a population distribution $\mathbb{P}$, and let a generative model G trained on T produce a synthetic dataset $S \sim G(T)$. Given a target data point $x^\star$ from $\mathbb{P}$, an adversary seeks to decide whether $x^\star \in T$ by computing a score $f(x^\star)$ from information allowed by a *threat model* and comparing it to a threshold γ:

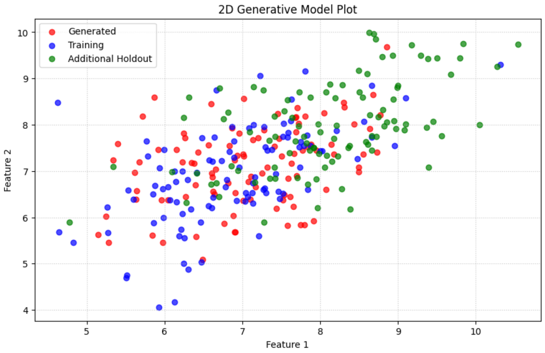
$$A(x^\star) = \mathbb{I}\{f(x^\star) > \gamma\}.$$

Threat models[1] range from black-box settings (access to S and possibly a reference set $R \sim \mathbb{P}$), to shadow-box settings (the above plus knowledge of the model implementation), to white-box settings (the above plus access to the model parameters of G). Attack performance is typically evaluated using standard binary classification metrics (e.g., AUC–ROC), with stronger attacks interpreted as evidence of greater privacy risk.

Across these settings, MIAs exploit the intuition that if a generative model has excessive capacity, then the synthetic dataset S will be *overly similar* to the training data T relative to an independent reference dataset R. Figure 6 illustrates this intuition, and also that without applying an MIA, it is difficult to identify which synthetic samples leak privacy by visual inspection alone.

Concretely, two broad classes of MIAs can be distinguished: distance-based approaches, e.g., [47], and (local) likelihood-based approaches, e.g., [48], which construct the score function f in fundamentally different ways. Compared with state-of-the-art methods [39, 21, 31], these approaches operate under realistic threat models for data release while incurring substantially lower computational cost—often several orders of magnitude less. While MIAs do not replace formal privacy guarantees, they provide a practical, task-agnostic diagnostic that is conceptually simple and capable of revealing material privacy risks. As such, MIAs constitute an essential tool for empirically characterizing when and where synthetic data may leak information about the training data.

Figure 6: Privacy auditing via membership inference attacks. An adversary uses synthetic outputs S and an additional holdout set R to infer a candidate data point $x^\star$ is from the training dataset T or not.



In the literature, there are two related notions to privacy leakage: one is overfitting and another is memorization. They are often used interchangeably. Conceptually, however, overfitting is a population-level phenomenon, referring to a generator that approximates the training distribution too closely relative to the underlying population distribution, whereas memorization is a sample-level phenomenon, whereby individual training records - often outliers - are explicitly reproduced in synthetic outputs. While these phenomena are distinct, they are closely related in practice. Distance-based MIAs [47] are particularly effective for detecting memorization, whereas (local) likelihood-based MIAs [48] are more naturally

[1] Note that terminology for threat models is used inconsistently across the literature.

aligned with diagnosing overfitting, although the resulting privacy risk assessments are often highly correlated.

3.4 Trade-offs among Fidelity, Utility and Privacy

A key insight emerging from recent studies is that privacy, fidelity, and utility are fundamentally intertwined. Intuitively, synthetic data with high fidelity tend to be more useful for downstream model training, but less effective at preserving the privacy of the underlying real data; see the benchmark results in [46]. While this intuition is broadly correct, a growing body of empirical evidence reveals that the relationships among these three dimensions are far more nuanced. Such observations underscore the need for a more systematic and rigorous theoretical framework to characterize these trade-offs.

Due to space constraints, we illustrate this subtle interplay with a single representative example concerning the trade-off between privacy and utility. Consider three different classes of DP data synthesizers, each with distinct strengths and limitations; see Figure 5. As reported in [30], when privacy constraints are relaxed - corresponding to increasing the privacy budget ϵ in ϵ-DP - models trained with DP-SGD (Differentially Private Stochastic Gradient Descent) [2] tend to exhibit a persistent degradation in utility that does not recover even under weaker privacy requirements. In contrast, PATE (Private Aggregation of Teacher Ensembles) [35] achieves competitive utility when privacy constraints are loose, but its performance deteriorates sharply as stricter privacy guarantees are imposed. By comparison, PrivBayes [57], a Bayesian network–based DP synthesizer, demonstrates relatively strong performance across both stringent and relaxed privacy regimes.

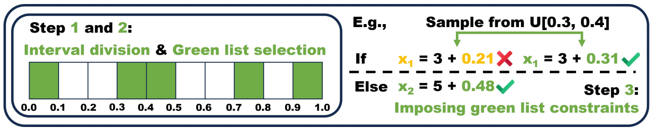
3.5 Statistical Watermarking

The final topic addressed in this position paper concerns *provenance and trust* in synthetic data, which extends beyond the traditional dimensions of fidelity, privacy, and utility. Distinguishing AI-generated content from human-generated content - across modalities including images, text, audio, and tabular data - remains an open challenge, with far-reaching implications for misinformation [6], plagiarism [26], and copyright infringement [32].

While post hoc detection methods based on classification have been widely studied, they are inherently fragile to adversarial paraphrasing, stylistic variation, and the rapid evolution of generative models [28]. In contrast, watermarking has emerged as a promising alternative by embedding statistical signals that are imperceptible to humans yet reliably detectable by machines, thereby enabling scalable and robust attribution of generated content [1]. Although substantial progress has been made in watermarking for unstructured modalities such as text [24] and images [49], comparable techniques for structured tabular data remain largely unexplored. Developing robust watermarking methods for synthetic tabular data is therefore an open problem with unique challenges, particularly for data scientists whose work predominantly relies on tabular data.

Let me briefly review a series of recent works [17, 15, 18] on watermarking for synthetic tabular data. Consider a table whose entries are continuous-valued. Inspired by [24], the unit interval $[0, 1]$ is first partitioned into $2m$ equal sub-intervals, which are grouped into m disjoint pairs. A pseudo-random key is then used to select one sub-interval from each pair,

Figure 7: Three-step watermarking scheme for generative tabular data.


Step 1 and 2:
Interval division & Green list selection
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
E.g.,
Sample from U[0.3, 0.4]
If x_1 = 3 + 0.21 ✗ x_1 = 3 + 0.31 ✓
Else x_2 = 5 + 0.48 ✓
Step 3:
Imposing green list constraints

forming a designated “green list.” During generation, if a value falls outside the green list, its fractional part is replaced by a sample from the nearest green sub-interval; see Figure 7 for an illustration.

To detect the embedded watermark, one computes, for each column i, the proportion of generated values that fall within the corresponding green sub-intervals, denoted by T_i. Under the null hypothesis of no watermark, the statistic satisfies the asymptotic distribution

$$2\sqrt{n}\left(T_i - \tfrac{1}{2}\right) \;\Rightarrow\; \mathcal{N}(0,1),$$

which enables watermark verification via standard hypothesis testing. Despite its simplicity, this procedure has proven both effective and revealing, exposing several fundamental phenomena governing watermarking behavior.

In particular, these studies uncover a *robustness–fidelity trade-off* that has received limited attention in prior work. Informally, it can be shown that to effectively erase the watermark via additive Gaussian noise, an attacker must inject noise with standard deviation at least on the order of $\Omega(1/m)$, matching the scale of the watermark perturbation itself. Consequently, increasing m improves fidelity by reducing distortion to the original distribution, but simultaneously reduces robustness, since a smaller noise variance suffices for watermark removal. This result provides a quantitative characterization of the cost of removal. Further analysis[2] reveals that balancing fidelity, robustness, and detectability may depend sensitively on the choice of metric. In particular, when fidelity is measured via f-divergence, the optimal trade-off between fidelity and detectability is characterized by [17].

The trade-offs among robustness, fidelity, and verifiability can thus be characterized in a relatively transparent manner for tabular data. By contrast, the discrete and autoregressive structure of large language models introduces substantially greater complexity for comprehensive analysis across all three dimensions - particularly given the need to preserve fluency and semantic coherence after watermarking. One promising direction in this setting is topic-based watermarking [33].

4 Path Forward

This position paper only scratches the surface of the rapidly evolving landscape of synthetic data and generative data science. Much of the existing theoretical work remains confined to stylized or toy settings, with only tentative connections to real-world systems. Looking ahead, we believe that carefully designed data synthesizers have the potential not merely to replicate existing data distributions, but to *actively enhance* the capabilities of large

[2] A softened version of the proposed watermarking can further improve the robustness–fidelity trade-off.

language models - particularly in multi-step reasoning [14], causal understanding [42] and physical intelligence [4] - by generating relevant pre-training or fine-tuning data.

At the same time, the current state of generative data science exposes several fundamental gaps where rigorous theoretical understanding is urgently needed. Among these, a central open question concerns how to justify and quantify the value of synthetic data from a *computational* perspective. One promising viewpoint is to regard computation as a means of "charging" real data: by injecting computational effort, the same underlying information can be reorganized into forms that are more structured and more readily exploitable by learning algorithms under computational constraints. This perspective complements, rather than replaces, the classical information-theoretic framework.

From a practical standpoint, the development of *tabular-native foundation models* - for table reasoning, synthesis, and prediction - emerges as a particularly important open direction, and one that is well suited to the expertise of data scientists. While recent LLM-based approaches alleviate data scarcity through large-scale pretraining, their inductive biases are primarily linguistic rather than tabular-structural. As a result, it remains unclear whether language-driven generation can faithfully capture heterogeneous schemas, numerical dependencies, and domain-specific constraints inherent to structured data. Tabular-native foundation models, pre-trained across diverse tables and domains, offer a principled path toward learning a "world model" of structured data. CTSyn [29] and TabPFN [20] represent an early step towards this direction.

Addressing these challenges will require close integration of statistical theory, algorithmic innovation, and empirical validation, and constitutes a fertile and timely direction for future research in generative data science.

Acknowledgment: This position paper grew out of a series of generative AI courses taught by the author at UCLA, spanning levels from senior undergrad to advanced doctoral student. The author gratefully acknowledges financial support from the National Science Foundation (NSF–CNS 2247795), the Office of Naval Research (ONR N00014-22-1-2680), and a gift from Cisco.

References

[1] Scott Aaronson and Hendrik Kirchner. Watermarking GPT outputs. 2023. Technical report / blog post.

[2] 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*, CCS '16, page 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

[3] Meshi Bashari, Yonghoon Lee, Roy Maor Lotan, Edgar Dobriban, and Yaniv Romano. Statistical inference leveraging synthetic data with distribution-free guarantees. *arXiv preprint arXiv:2509.20345*, 2025.

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim

Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026.

[5] California Attorney General. California consumer privacy act (ccpa), 2024. Accessed: 2024-09-22.

[6] Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368, 2024.

[7] Guang Cheng, Chi-Hua Wang, Vamsi Potluru, Tucker Balch, and C. Cheng. Downstream task-oriented generative model selections on synthetic data training for fraud detection models. In *ACM International Conference on AI in Finance – Workshop*, 2022.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[9] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023.

[10] Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[11] Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024.

[12] David Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017.

[13] GDPR.eu. Gdpr compliance, 2024. Accessed: 2024-09-22.

[14] Anna Goldie, Azalia Mirhoseini, Hao Zhou, Irene Cai, and Christopher D. Manning. Synthetic data generation & multi-step rl for reasoning & tool use. *arXiv preprint*, arXiv:2504.04736, 2025.

[15] Bochao Gu, Hengzhi He, and Guang Cheng. Watermarking generative categorical data. *arXiv preprint arXiv:2411.10898*, 2024.

[16] Hengzhi He, Shirong Xu, and Guang Cheng. Golden ratio weighting prevents model collapse. *arXiv preprint arXiv:2502.18049*, 2025.

[17] Hengzhi He, Shirong Xu, Alexander Nemecek, Jiping Li, Erman Ayday, and Guang Cheng. Optimal watermark generation under type i and type ii errors. *arXiv preprint*, arXiv:2512.05333, 2025. stat.ME.

[18] Hengzhi He, Peiyu Yu, Junpeng Ren, Ying Nian Wu, and Guang Cheng. Watermarking generative tabular data. *arXiv preprint arXiv:2405.14018*, 2024.

[19] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[20] Niklas Hollmann, Simon Müller, Lukas Purucker, et al. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–326, 2025.

[21] Florimond Houssiau, James Jordon, Samuel N Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. Tapas: a toolbox for adversarial privacy auditing of synthetic data. *arXiv preprint arXiv:2211.06550*, 2022.

[22] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how? *CoRR*, abs/2205.03257, 2022.

[23] Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls. *arXiv preprint arXiv:2510.01631*, 2025.

[24] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR, 23–29 Jul 2023.

[25] Tung Sum Thomas Kwok, ChiHua Wang, and Guang Cheng. Greater: Generate realistic tabular data after data enhancement and reduction. In *ICDE Workshop: Data Engineering Meets Large Language Models*, 2025.

[26] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647. ACM, 2023.

[27] [First] Li, Chi-Hua Wang, and Guang Cheng. Statistical theory of differentially private marginal-based data synthesis algorithms. In *International Conference on Learning Representations (ICLR)*, 2023.

[28] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 2023.

[29] [First] Lin, [First] Xu, [First] Yang, and Guang Cheng. Ctsyn: A foundational model for cross tabular data generation. In *ICLR*, 2025.

[30] Yucong Liu, Chihua Wang, and Guang Cheng. On the utility recovery incapability of neural net-based differential private tabular training data synthesizer under privacy deregulation. *arXiv preprint arXiv:2211.15809*, 2022.

[31] Matthieu Meeus, Florent Guepin, Ana-Maria Cretu, and Yves-Alexandre de Montjoye. *Achilles' Heels: Vulnerable Record Identification in Synthetic Data Publishing*, page 380–399. Springer Nature Switzerland, 2024.

[32] Felix B. Mueller, Rebekka Görge, Anna K. Bernzen, Janna C. Pirk, and Maximilian Poretschkin. LLMs and Memorization: On quality and specificity of copyright compliance. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 984–996. ACM, 2024.

[33] Alexander Nemecek, Yuzhou Jiang, and Erman Ayday. Topic-based watermarks for large language models. *arXiv preprint*, arXiv:2404.02138, 2024.

[34] Yidong Ouyang, Liyan Xie, Chongxuan Li, and Guang Cheng. Missdiff: Training diffusion models on tabular data with missing values. In *ICML workshop on Structured Probabilistic Inference and Generative Modeling*, 2023.

[35] Nicolas Papernot, Martín Abadi, Úlfar Erlingsson, Ian J. Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *ArXiv*, abs/1610.05755, 2016.

[36] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

[37] Aivin Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers. *arXiv preprint arXiv:2302.02041*, 2023.

[38] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[39] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, Boston, MA, August 2022. USENIX Association.

[40] Namjoon Suh, [First] Lin, Din-Yin Hsieh, [First] Honarkhah, and Guang Cheng. Autodiff: Combining auto-encoder and diffusion model for tabular data synthesizing. In *NeurIPS Workshop on SyntheticData4ML*, 2023.

[41] Namjoon Suh, Yuning Yang, Din-Yin Hsieh, Qitong Luan, Shirong Xu, Shixiang Zhu, and Guang Cheng. Timeautodiff: Combining autoencoder and diffusion model for time series tabular data synthesizing. *arXiv preprint arXiv:2406.16028*, 2024.

[42] Omar Swelam, Lennart Purucker, Jake Robertson, Hanne Raum, Joschka Boedecker, and Frank Hutter. Does tabpfn understand causal structures? *arXiv preprint*, arXiv:2511.07236, 2025.

[43] Lan Tao, Shirong Xu, Chi-Hua Wang, Namjoon Suh, and Guang Cheng. Discriminative estimation of total variation distance: A fidelity auditor for generative data. *arXiv preprint arXiv:2405.15337*, 2024.

[44] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: will we run out of data? limits of llm scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[45] Lecheng Wang, Xianjie Shi, Ge Li, Jia Li, Xuanming Zhang, Yihong Dong, Wenpin Jiao, and Hong Mei. Theoretical proof that generated text in the corpus leads to the collapse of auto-regressive language models, 2025.

[46] Joshua Ward, Xiaofeng Lin, , Chi-Hua Wang, and Guang Cheng. Synth-mia: A testbed for auditing privacy leakage in tabular data synthesis. *arXiv preprint arXiv:2509.18014*, 2025.

[47] Joshua Ward, ChiHua Wang, and Guang Cheng. Data plagiarism index: Characterizing the privacy risk of data-copying in tabular generative models. In *KDD Workshop on GenAI Evaluation*, 2024.

[48] Joshua Ward, Chi-Hua Wang, and Guang Cheng. Privacy auditing synthetic data release through local likelihood attacks. *arXiv preprint arXiv:2508.21146*, 2025.

[49] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[50] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint*, arXiv:1802.06739, 2018.

[51] [First] Xing, [First] Song, and Guang Cheng. Why do artificially generated data help adversarial robustness? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[52] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[53] Shirong Xu, Hengzhi He, and Guang Cheng. A probabilistic perspective on model collapse. *arXiv preprint arXiv:2505.13947*, 2025.

[54] Shirong Xu, Will Wei Sun, and Guang Cheng. Utility theory of synthetic data generation. *arXiv preprint arXiv:2305.10015*, 2023.

[55] Bin Yu and Rebecca L. Barter. *Veridical Data Science: The Practice of Responsible Data Analysis and Decision Making.* The MIT Press, Cambridge, MA, 2024. Adaptive Computation and Machine Learning series.

[56] Xianli Zeng, Edgar Dobriban, and Guang Cheng. Bayes-optimal classifiers under group fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[57] Jun Zhang, Graham Cormode, Cecilia Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes : private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 2014.

[58] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model. *arXiv preprint*, arXiv:1801.01594, 2018.