# Synthetic Tabular Data: Methods, Attacks and Defenses

Graham Cormode
Meta and University of Warwick
Coventry, UK
gcormode@meta.com

Shripad Gade
Meta
Menlo Park, USA
shripadgade@meta.com

Samuel Maddock
Meta and University of Warwick
Coventry, UK
smaddock@meta.com

Enayat Ullah
Meta
Menlo Park, USA
enayat@meta.com

## ABSTRACT

Synthetic data is often positioned as a solution to replace sensitive fixed-size data sets with a source of unlimited matching data, freed from privacy concerns. There has been much progress in synthetic data generation over the last decade, leveraging corresponding advances in machine learning and data analytics. In this tutorial, we survey the key developments and the main concepts in tabular synthetic data generation, including paradigms based on probabilistic graphical models and on deep learning. We provide background and motivation, before giving a technical deep-dive into the methodologies. We also address the limitations of synthetic data, by studying attacks that seek to retrieve information about the original sensitive data. Finally, we present extensions and open problems in this area.

## KEYWORDS

synthetic data, differential privacy, marginal distributions, membership inference

**PVLDB Reference Format:**
Graham Cormode, Shripad Gade, Samuel Maddock, and Enayat Ullah. Synthetic Tabular Data: Methods, Attacks and Defenses. PVLDB, 18(12): 5448 - 5450, 2025.
doi:10.14778/3750601.3750692

## 1 MOTIVATION AND OVERVIEW

A common scenario in many data-focused applications is when there is a valuable dataset but its contents are very sensitive. For instance, this could be a dataset of customers with their personal details and purchases, or of hospital patients with information on their health conditions. The dataset would be very useful to share with data scientists or ML engineers, but due to privacy concerns it is not appropriate to make the data available in its original form.

Instead we would like to create a new dataset that shares the characteristics of the original data, but is entirely fabricated. This is referred to as "Synthetic Data Generation". Being completely made up, intuitively we would believe that the synthetic data is freed of privacy concerns, and can be shared more easily than the original source. However, things are not so simple: if the synthetic data is very similar to the original data, it may leak sensitive information about its source. Meanwhile, if the synthetic data does not resemble the original data, then it is not a very useful substitute. Research in synthetic data generation is concerned with walking this tightrope: balancing fidelity and privacy, whilst also taking into account expressivity (the richness of the models), and efficiency (computational cost).

Synthetic data can take many forms, depending on the domain. We might want to generate synthetic text, synthetic images and videos, or synthetic three-dimensional objects. However, in this tutorial we focus on the core case of synthetic tabular data: data which is most naturally represented within a structured table. This captures many problems in data management, where we can consider the tables as relations from a database; and in machine learning, where the rows are examples and the columns are features.

In this tutorial, we will give an overview of the state-of-the-art in synthetic tabular data. We will describe the objectives and desiderata for synthetic data, and how they are achieved. We will use detailed examples to show how techniques have developed from simplistic modeling to leveraging complex cutting-edge machine learning models, and the tradeoffs along this path. A number of different lenses can be used to view the task of generating synthetic data: a statistical lens, which seeks to find a parsimonious model of the original data from which new examples can be sampled; or a machine learning perspective, which seeks to train a model that can generate examples that are sufficiently realistic to fool a classifier; we can also adopt the framing of generative AI, where the objective is to create data based on many real-world examples of tables and the context of a specific target. We will also consider the limitations of synthetic data generation, and how adversaries can try to use the output synthetic data to learn private information about the data that the model was trained on. We will address defenses against such attacks based on formal privacy guarantees, and discuss how the relative success of attacks can be used as a metric for the empirical level of privacy that the synthetic data obtains. We conclude with a consideration of extensions to other forms of data and other scenarios, and open problems for the community to work on.


Proceedings of the VLDB Endowment, Vol. 18, No. 12 ISSN 2150-8097.
doi:10.14778/3750601.3750692

## 2 OUTLINE OF THE TUTORIAL

**Introduction and Motivation**

- The background to synthetic data generation and motivating factors: legal and policy reasons to want synthetic data. What problems can synthetic data solve? When is synthetic data not appropriate?
- The desiderata for synthetic data: fidelity, privacy, efficiency, expressivity, and how can each of these be specified or measured. The lack of a universal notion of "utility", and alternatives: similarity measures (such as workload loss) or evaluations on representative downstream tasks
- Formalizing the privacy requirements for synthetic data. When privacy and utility are in competition, and when they can be complementary. A brief introduction to differential privacy and the basic methods to achieve it by random noise addition to statistics [8].

**Marginal-based methods**

- Initial work on synthetic tabular data came from the statistics community, where the key concepts include marginals and probabilistic graphical models (PGMs) [16].
- We begin with some initial approaches to building synthetic data. For instance, a heuristic approach is to create a synthetic example by interpolating between one or more real examples, such as via SMOTE [4].
- We consider two modeling approaches from opposite extremes. The first is to treat each attribute as independent from each other, and to learn the correlation of each one with a target attribute: this yields the Naive Bayes approach [30]. The second is to attempt to fit a model to the full distribution of all attributes, via a multiplicative weights approach, MWEM [13].
- An important middle way is achieved by using probabilistic graphic models to describe the data. A first approach was PrivBayes, which makes use of Bayesian networks [33]. This approach relies on treating data generation as a form of inference. In building the model, the effort is split between structure learning (choosing the graph) and parameter estimation (instantiating marginal distributions, a.k.a. 'marginals').
- Successive approaches have expanded on this graphical modeling approach, by varying the class of models considered, and the approach taken to learn the model structure. These include private optimization via PrivMRF [5], and model building via PrivSyn [34] and the MST-based approach [23].
- The state-of-the-art methods are also based on marginal generation. These make use of the "select-measure-generate" approach: given a target workload of queries to answer, select a next marginal that will give the biggest increase in accuracy, measure that marginal (with DP noise), and use the current set of published marginals to generate a set of synthetic data. The methods here are AIM [24] and RAP++ [29].
- Several approaches have sought to extend marginal-based tabular data generation in different directions. The PrivLava algorithms seeks to support multi-table generation via latent variables [6]; JAM-PGM addresses the case when some of the training data is considered public [10]; and Private-GSD makes use of genetic algorithms to better fit the training data [18].

**Deep learning-based methods**

- Generative Adversarial Networks (GANs) produce synthetic data by training a generator to try to fool a discriminator [12]. However, applying GANs for tabular data is challenging, due to the mix of categorical and numerical features.
- CTGAN was developed specifically to generate tabular data via Gaussian Mixture Models and sampling [31]. A version of this approach with differential privacy, DP-CTGAN [9] can be obtained by adopting the generic DP-SGD approach to training [1]. Alternatives based on variational autoencoders (VAEs) were also proposed by Xu *et al.* [31]
- An alternate method is the GEM, which trains a generator network on the noisy marginals that are measured, following the "select-measure-generate" paradigm [20].
- TabDDPM [17] adapts diffusion models to tabular synthetic data. Subsequently Zhang *et al.* proposed TabSyn [32] building directly on TabDDPM.
- LLM-based approaches include GReaT [3], which fine-tunes a pretrained LLM on textually encoded tabular data, and SynLM [27] which trains a transformer-based language model for tables from scratch via DP-SGD.

**Attacks and defenses**

- We consider synthetic data as an object of attack, and describe the potential for information leakage. An empirical privacy measurement is modeled as a game between an attacker ("the adversary") and the data curator. Results vary depending on what access the attacker has to the model and to auxiliary information.
- Membership Inference attacks are used to measure vulnerability, based on the ratio of true positives to false positives.
- Density-based attacks test if training examples are overfitted by the model, and estimate the likelihood of a target output [14, 15]. Important examples include DOMIAS [28] and MAMA-MIA [11].
- The *shadow modeling* approach trains a classifier on public data to recognize when a point is drawn from the private input. This approach has found privacy vulnerabilities in several implementations of synthetic tabular data generation [2].

**Advanced Topics and Open Problems**

- We also consider generating other types of data, such as synthetic text and synthetic graphs.
- Recent work has considered the case when the reference data is held by many distributed individuals. This leads to efforts on both distributed [25, 26] and federated [21] synthetic data generation.
- We describe cases where some attributes of the data are public, which allows higher utility [10, 19, 22].

## 3 LOGISTICS

**Timing and plan.** The tutorial spans three hours to cover all the above listed topics in detail. The outline timing plan is:

- Introduction, motivation and overview- 25 minutes
- Marginal-based algorithms - 45 minutes
- Deep learning based methods - 40 minutes
- Attacks and defenses - 30 minutes
- Advanced Topics and Open Problems - 30 minutes
- Additional Q&A - 10 minutes

The tutorial slides are available from the tutorial website, at *https://sites.google.com/view/synthetic-tab-data-tutorial*, along with links to other relevant material including an accompanying survey [7].

**Intended Audience and Background Knowledge.** The tutorial is intended to be accessible to all participants at VLDB, and so makes minimal assumptions on prior knowledge. The main methodological approaches covered are statistical (probabilities, distributions, probabilistic graphical modes and inference), machine learning (generative adversarial networks, diffusion models, gradient descent), and privacy (differential privacy, membership inference attacks). While some awareness of each of these topics would be useful, we do not expect any knowledge of any of them, and instead introduce just the definitions needed. We avoid giving formal mathematical analysis, and rather focus on building intuition around the different methods and their pros and cons. We refer to the relevant research literature for more in-depth study of the algorithms and their properties. To this end, the tutorial does not include any formal proofs, but instead provides an overview and intuition of the key concepts. As a result the tutorial is intended to be suitable for starting researchers or for those with expertise in other areas seeking to understand the key concepts in synthetic data generation.

## 4 ABOUT THE PRESENTERS

**Graham Cormode** is a research scientist at Meta in the UK. Since 2013, he is a professor at the University of Warwick, UK. He has worked extensively on privacy-preserving computations and efficient data analytics.

**Shripad Gade** is a research scientist at Meta in the USA. He completed his PhD at the University of Illinois, Urbana-Champaign. He has previously worked on federated learning and multi-party computation for private optimization. His current focus is on synthetic data generation and its applications within Meta.

**Samuel Maddock** is a final-year PhD student at the University of Warwick and a research intern at Meta. His research focus is on private federated learning, with emphasis on private pre- and post-training, and federated synthetic data generation.

**Enayat Ullah** is a research scientist at Meta in the USA. He received his PhD from Johns Hopkins University. He works at the intersection of privacy, optimization and machine learning. His current work addresses differential privacy, membership inference attacks, and empirical measures of privacy.

## ACKNOWLEDGMENTS

This work is supported in part by EPSRC grant EP/V044621/1, the UKRI Prosperity Partnership Scheme (FAIR) under the EPSRC Grant EP/V056883/1, and The Alan Turing Institute.

## REFERENCES

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *ACM CCS*, 2016.
[2] M. S. M. S. Annamalai, G. Ganev, and E. De Cristofaro. "what do you want from theory alone?" experimenting with tight auditing of differentially private synthetic data generation. In *USENIX Security*, pages 4855–4871, 2024.
[3] V. Borisov, K. Seßler, T. Leemann, M. Pawelczyk, and G. Kasneci. Language models are realistic tabular data generators. *arXiv:2210.06280*, 2022.
[4] K. W. Bowyer, N. V. Chawla, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011.
[5] K. Cai, X. Lei, J. Wei, and X. Xiao. Data synthesis via differentially private markov random field. *Proc. VLDB Endow.*, 14(11):2190–2202, 2021.
[6] K. Cai, X. Xiao, and G. Cormode. Privlava: Synthesizing relational data with foreign keys under differential privacy. *ACM SIGMOD*, 1(2):142:1–142:25, 2023.
[7] G. Cormode, S. Maddock, E. Ullah, and S. Gade. Synthetic tabular data: Methods, attacks and defenses. *CoRR*, abs/2506.06108, 2025.
[8] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.
[9] M. L. Fang, D. S. Dhami, and K. Kersting. DP-CTGAN: Differentially private medical data generation using ctgans. In *Intl. Conf. on AI in medicine*, 2022.
[10] M. Fuentes, B. C. Mullins, R. McKenna, G. Miklau, and D. Sheldon. Adaptively incorporating public information for private synthetic data. In *AISTATS*, 2024.
[11] S. Golob. *Privacy Vulnerabilities in Marginals-based Synthetic Data.* University of Washington, 2024.
[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *NeurIPS*, 2014.
[13] M. Hardt and G. N. Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *IEEE FOCS*, pages 61–70, 2010.
[14] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro. Logan: Membership inference attacks against generative models. *arXiv:1705.07663*, 2017.
[15] B. Hilprecht, M. Härterich, and D. Bernau. Monte carlo and reconstruction membership inference attacks against generative models. *PETS*, 2019.
[16] D. Koller and N. Friedman. *Probabilistic Graphical Models - Principles and Techniques.* MIT Press, 2009.
[17] A. Kotelnikov, D. Baranchuk, I. Rubachev, and A. Babenko. Tabddpm: Modelling tabular data with diffusion models. In *ICML*, 2023.
[18] T. Liu, J. Tang, G. Vietri, and S. Wu. Generating private synthetic data with genetic algorithms. In *ICML*, 2023.
[19] T. Liu, G. Vietri, T. Steinke, J. Ullman, and S. Wu. Leveraging public data for practical private query release. In *ICML*. PMLR, 2021.
[20] T. Liu, G. Vietri, and S. Wu. Iterative methods for private synthetic data: Unifying framework and new methods. In *NeurIPS*, 2021.
[21] S. Maddock, G. Cormode, and C. Maple. FLAIM: Aim-based synthetic data generation in the federated setting. In *ACM SIGKDD*, 2024.
[22] S. Maddock, S. Gade, G. Cormode, and W. Bullock. Leveraging vertical public-private split for improved synthetic data generation. *arXiv*, 2504.10987, 2025.
[23] R. McKenna, G. Miklau, and D. Sheldon. Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *J. Priv. Confidentiality*, 11(3), 2021.
[24] R. McKenna, B. Mullins, D. Sheldon, and G. Miklau. AIM: an adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.*, 15(11):2599–2612, 2022.
[25] S. Pentyala, M. Pereira, and M. De Cock. Caps: Collaborative and private synthetic data generation from distributed sources. *arXiv:2402.08614*, 2024.
[26] M. Pereira, S. Pentyala, A. Nascimento, R. T. d. Sousa Jr, and M. De Cock. Secure multiparty computation for synthetic data generation from distributed data. *arXiv:2210.07332*, 2022.
[27] A. Sablayrolles, Y. Wang, and B. Karrer. Privately generating tabular data using language models. *arXiv:2306.04803*, 2023.
[28] B. Van Breugel, H. Sun, Z. Qian, and M. van der Schaar. Membership inference attacks against synthetic data through overfitting detection. *arXiv:2302.12580*, 2023.
[29] G. Vietri, C. Archambeau, S. Aydöre, W. Brown, M. Kearns, A. Roth, A. A. Siva, S. Tang, and Z. S. Wu. Private synthetic data for multitask learning and marginal queries. In *NeurIPS*, 2022.
[30] G. I. Webb. Naïve bayes. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning and Data Mining*. Springer, 2017.
[31] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni. Modeling tabular data using conditional GAN. In *NeurIPS*, pages 7333–7343, 2019.
[32] H. Zhang, J. Zhang, B. Srinivasan, Z. Shen, X. Qin, C. Faloutsos, H. Rangwala, and G. Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. *arXiv:2310.09656*, 2023.
[33] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: Private data release via bayesian networks. *ACM TODS*, 42(4):25:1–25:41, 2017.
[34] Z. Zhang, T. Wang, N. Li, J. Honorio, M. Backes, S. He, J. Chen, and Y. Zhang. Privsyn: Differentially private data synthesis. In *USENIX Security*, 2021.