Generating synthetic data in finance: opportunities, challenges and pitfalls

Samuel Assefa[1], **Danial Dervovic**[1], **Mahmoud Mahfouz**[1]
Tucker Balch [1], **Prashant Reddy** [1], **and Manuela Veloso**[1]
[1] JPMorgan Chase & Co.,
samuel.a.assefa@jpmorgan.com

Abstract

Financial services generate a huge volume of data that is extremely complex and varied. These datasets are often stored in silos within organisations for various reasons, including but not limited to, regulatory requirements and business needs. As a result, data sharing within different lines of business as well as outside of the organisation (*e.g.* to the research community) is severely limited. It is therefore critical to investigate methods for synthesising financial datasets that follow the same properties of the real data while respecting the need for privacy of the parties involved in a particular dataset.

This introductory paper aims to highlight the growing need for effective synthetic data generation in the financial domain. We highlight three main areas of focus for the academic community: 1) Generating realistic synthetic datasets. 2) Measuring the similarities between real and generated datasets 3) Ensuring the generative process satisfies any privacy constraints.

Although these challenges are also present in other domains, the extra regulatory and privacy requirements add another dimension of complexity and offer a unique opportunity to study the topic in financial services. Finally, we aim to develop a shared vocabulary and context for generating synthetic financial data using two types of financial datasets as examples.

1 Introduction

Financial data contains some of the most sensitive and personally identifiable attributes of customers. Using and sharing such data, especially for research purposes outside of the organisations that generate it, is severely restricted. One approach to address this limitation is the generation of synthetic data. The primary directive in generating synthetic financial data is therefore protecting the privacy of customers and entities involved in generating a particular synthetic data set. This directive has been enshrined into law in various jurisdictions, notably the GDPR laws in the European Union [1], and in the United States FERPA [2] and HIPAA [3], relating to educational and medical data privacy, respectively. The issue of data privacy is of great importance in public opinion, as evidenced by reactions to the Facebook/Cambridge Analytica scandal [4] and the numerous data breaches [5] that have occurred over the last few decades, alongside the financial markets' response to these breaches [6].

We define synthetic data as data obtained from a generative process that learns the properties of the real data. Such processes are strictly different from the most commonly used data obfuscation techniques (*e.g.* anonymisation or removing certain sensitive attributes) as our intention is to synthesise new samples that are related to but can not be mapped back to the real data. Some of the requirements for such generative processes include:

33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Workshop on AI in Financial Services: Data, Fairness, Explainability, Trustworthiness, and Privacy. Vancouver, Canada.

- Capability for generating many different data types, such as numeric, binary and categorical.
- The process should be able to generate an arbitrary number of data points, with as many features as desired to good fidelity.
- The privacy characteristics of the dataset can be precisely tuned against how realistic the data is.

We shall concentrate primarily on considerations of *privacy* over *security*, that is, we are in a setting where we wish to share a dataset without compromising information about any given entity within the dataset. An attacker is assumed to have full access to the synthetic dataset, as well as to the generating algorithm, but not the original dataset.

1.1 Motivation

Some the basic use cases and motivations for synthetic data generation in finance are highlighted below.

Internal data use restrictions. Regulatory requirements may prevent data sharing between different lines of business within a company. Alternatively, teams may wish to begin working with data before the relevant approvals have been made.

Lack of historical data. There is a limited amount of historical data to study certain events (*e.g.* flash crashes in the market, recessions, new regimes of behaviour) that make studying the underlying mechanisms very challenging. It is useful in various such settings to have counterfactual data for testing strategies and inferences.

Tackling class imbalance. For use cases such as fraud detection, the datasets are usually highly imbalanced, and traditional machine learning and anomaly detection techniques often fail. Realistic synthetic data, along with appropriate data imputation techniques offer a promising approach to tackle this challenge.

Training advanced Machine Learning models. Large scale advanced machine learning (*e.g.* deep learning) is often carried out using cloud services, requiring compute resources and vast quantities of training data. Institutions may not be able to upload training data to these services for a number of reasons. Synthetic data can be used to train models, which can then be brought back on premise to be used on real data. Moreover, training on synthesised data offers some protection from "membership inference attacks", wherein model parameters can be used to extract training data.

Data sharing. By sharing data between institutions and within the research community, better solutions can be found for technical problems faced by financial institutions. Sharing of synthetic data allows financial institutions to do this in a way that satisfies their data sharing restrictions.

2 Examples of financial data

For the purposes of this paper, we consider two broad classes of financial data, *retail banking* and *market microstructure* data. Retail financial data arises from operations facing the general public. This includes, but is not limited, to transaction data, loan applications, customer service logs, etc. Market microstructure data refers to the data maintained and distributed by exchanges detailing historical limit order book orders and snapshots for a given financial asset. Generating synthetic data in each domain presents its own challenges, which we go through in turn.

2.1 Retail Banking Data

A famous example of a retail-type dataset is the UCI Adult dataset [7], where the binary feature of whether a particular person has an income of over \$50k/year is predicted by various census information, such as age, profession, marital status and gender. This data set is tabular, mimicking how data is typically stored in a relational database. Most examples of synthesised data sets are also tabular.

There have been a number of technical approaches suggested for protecting privacy in tabular data. The two main approaches are *k-anonymity* [8] and *differential privacy* [9]. A dataset is said to have the k-anonymity property if the information for each entity contained in the dataset cannot be distinguished from $k-1$ entities also appearing in the release. Differential privacy is a technical condition satisfied by a randomised algorithm ensuring that the distribution of the output of the algorithm is bounded when applied to two "adjacent" datasets, with the definition of adjacent being dependent on the application. In the most typical case, tabular datasets are adjacent when they differ by a single row. Typically, a differentially private algorithm is parameterized by the pair (ϵ, δ), where smaller values of ϵ denote greater privacy and δ denotes a failure probability. Differential privacy enjoys the property of *composability*, whereby differentially private algorithms can be chained with their output remaining differentially private. Moreover, differentially private algorithms are provably safe against attackers with side information [9]. The k-anonymity paradigm suffers from the *curse of dimensionality* [10], in that one has to destroy a rapidly increasing fraction of a particular dataset as the number of columns in the table grows to ensure k-anonymity. As such, differential privacy is the most popular technical solution for releasing data privately. Data synthesis methods using differential privacy for tabular data are well covered in the review by Bowen and Liu [11].

One must also be cognizant of data *deanonymization*: whence an anonymized dataset is combined with existing data to personally identify individuals in the dataset. This has been demonstrated by Daries et al. [12] where data from massive open online courses (MOOCs) was deanonymized and by Archie et al. [13], who deanonymized the Netflix prize dataset using Amazon review data. Indeed, the US Census is recognised as having been vulnerable to such attacks in previous iterations [14].

2.2 Market Microstructure Data

Markets data typically come in the form of time series describing, for example, a stock price over time. The granularity of the data is dependent on the frequency of the trading activity by the market participants. In the past decade, the rise of algorithmic trading, and specifically, high frequency trading has resulted in a significant increase in the amount of data available for research. However, access to such data sets is very limited and therefore an effort to synthesize such data sets using real data is needed.

Of particular interest to the research community is limit order book data. A limit order book is used by exchanges to match buyers and sellers of a particular security [15]. It is an electronic record of the outstanding orders in the market and represents a snapshot in time describing the supply and demand of the security. It is based on the continuous double auction mechanism whereby participants can submit both buy and sell orders and expect their trades to match instantaneously if a corresponding trade on the opposing side is present. Exchanges offer various order types. The two main types are market and limit orders. A market order is an instruction to buy/sell a specific amount of an asset without specifying the price. In contrast, a limit order specifies the price that should not be exceeded in the case of a buy order or gone below in case of a sell order.

The main technical challenge in synthesising such order book data is that of representing aggregate decisions of many independent actors with differing risk tolerance, rationality and motives. In addition, generating realistic datasets requires defining an appropriate distance between datasets, then accurately measuring how close the two datasets are. The empirical properties of limit order book data have been studied extensively in the literature and are often referred to as *stylized facts* of the real limit order book data. It is important to make sure that the empirical properties of the synthetic data follows, as close as possible, those of real order book data. For example, it is empirically shown that lower spreads (the difference between the best bid and ask prices) are observed during period of high trading volumes and that trading volumes are typically highest at the beginning and end of the trading day. These are two examples, of many, which would need to be taken into consideration when synthesizing synthetic data in the market microstructure domain [16].

3 Techniques for synthetic data generation

We shall focus mainly on synthetic data generation with privacy guarantees. The following sections highlight a selected list of techniques used for financial data generation.

3.1 Tabular Data

A number of techniques have been proposed for tabular data generation. For a comprehensive survey of these methods see the survey by Surendra and Mohan [17].

In the Data Mining literature Eno and Thompson [18] define an XML-based synthetic data definition language (SDDL) from which synthetic data may be generated. The algorithm generating the SDDL effectively "inverts" a decision-tree classifier. This method carries no guarantees on privacy. In a similar fashion, there are methods based on other classical machine learning classifiers, such as support vector machines [19] and Random Forests [20]. The drawback with these methods is that the better the classifier, the higher the risk of leaking data, with no tunable privacy parameter.

Abowd and Vilhuber [21] provide a Bayesian inspired differentially private synthetic data release of multidimensional tabular data. Zhang et al. [22] present a synthetic data generator based on Bayesian networks that is also differentially private. This method is effective, with tunable privacy parameters but suffers from the drawback of growing substantially with each new feature added. The method devised by Li et al. [23] generates data according to the histogram of each feature, linking the features via a copula. This method enjoys privacy guarantees but scales poorly with a growing number of features. The Gibbs sampling based method by Park and Ghosh [24] gives strong privacy guarantees and scales well. This technique is however limited to categorical variables.

These techniques for generating synthetic tabular data with privacy protection all suffer to a varying degree from the following limitations. Most differential privacy frameworks represent a row of a given table as a bit string with length equal to the *domain size*, which grows exponentially in the number of columns of the table. This representation quickly becomes impractical to use. The second limitation of this representation is that most high-dimensional datasets are very sparse, resulting in the noise being added to generate the privacy completely washing out the real data, rendering the released dataset unsuitable as an approximation to the true dataset. A more thorough discussion of these limitations can be found in Zhang [25].

Agent-based modelling (ABM) has been used in the context of synthesizing payments data, for instance in modelling a bank's payment processing system [26] and investigating the macroscopic impact of a disruptive event on the flow of interbank payments [27]. Synthetic data for a retail shoe store has been created using ABM by Lopez-Rojas and Axelsson [28]. This generated data intrinsically respects privacy constraints if calibration is manually carried out. If an automated quantitative method is used there is a risk of data leakage. Calibration methods in the ABM literature currently do not explicitly preserve privacy, a survey on ABM calibration methods can be found in [29].

3.2 Synthetic financial time series

There has been a lot of work in generating synthetic financial time series data and in releasing differentially private data streams, but little in the way of synthetic financial time series data with privacy guarantees. The classical approach is to propose a simple statistical model such as autoregressive or GARCH (Generalized Autoregressive Conditional Heteroscedasticity) models for financial time series, and then fit real-world data to this model using maximum-likelihood. These methods carry the advantage of being easy to fit and interpret, but rely on strong assumptions and are unable to reproduce many of the statistical features of many financial time series [30]. Examples of more modern approaches using neural architectures are QuantGAN [31] and the work by Fu et al. [32] for modelling log returns of stocks and associated classical time series models. These methods provide no privacy guarantees and have yet to be conclusively shown not to be explicitly memorising data. Agent-based models are often used to replicate the dynamics of financial markets and then used to derive financial time series approximating those seen live. This topic lies outside the scope of this paper, see [33], [34] , [15, Section 5C] and references therein for more on agent based models in finance.

3.3 Stream data with privacy guarantees

Publishing stream data with privacy guarantees was first addressed by Dwork [35] and Dwork et al. [36]. In this approach, stream data is viewed as a bit string with a continuous counter of the number of ones observed being reported. This model is called *event-level* privacy as individual events in the stream are protected. One can also protect *user-level* privacy, where the presence of a particular user in a dataset is masked from an adversary. This is not necessarily the most natural model of privacy for time series data. A privacy model called w-event privacy, protecting events occurring within any window of w timestamps, was proposed by Kellaris et al. [37] with associated data release algorithms. The model used here is one in which an organisation wishes to publish continuous counts of a finite number of different events occurring. There is also the notion of d-privacy, where d is a metric defined on datasets that generalises adjacency of databases for differential privacy, defined by Chatzikokolakis et al. [38]. This model allows for preserving the privacy of continuous quantities. These two models were combined by Fioretto and Van Hentenryck [39] into the notion of (w, α)-indistinguishability, protecting blocks of continuously-valued data, which is more relevant to financial data. The authors present their OPTSTREAM algorithm for releasing time series data under this model of privacy. Limitations of these methods have not yet been settled upon by the research community, but representation of this data is likely to lead to same scaling issues as in tabular data as the number of streams being simultaneously released increases. There is a yet no established approach in synthesising financial time series data, in particular market microstructure data.

3.4 Unstructured Data

Neural network based methods such as those presented by Shokri and Shmatikov [40], Acs et al. [41] and [42] extend differentially private synthetic data generation to the domain of unstructured data such as images and audio. Although highly promising, these methods are currently in their infancy and suffer from the usual problems neural networks face [43, 44]. Moreover, these methods protect data on an individual level, that is, the presence of a singular data point (x, y) within the training set is protected. This notion of privacy is insufficient in the case of, say, a image generator for faces, with the training set containing the same face at different angles. In this example one would need to protect against an attacker ascertaining the presence of this entire group of images in the training set.

Controlling the the trade off between noise and privacy [9] in an optimal way for synthetic data remains an open problem.

4 Evaluating generative models

As argued by Donoho [45], a large part of the success of machine learning as a field comes from the *common task framework*, namely having a set of benchmark datasets and a real-valued metric that was used to compare model quality. With this in mind, there has been a large body of work attempting to evaluate generative models using a single real-valued metric, as well as a standardisation of datasets being used.

In particular, we briefly focus on the success in the computer vision community with regard to generated image data. While of a very different nature to data that is of financial origin, image data shares many salient properties. Both types of data can often be qualitatively evaluated well by humans to some degree, and come from high-dimensional multi-model distributions. Indeed, these properties are what lead classical methods such as density estimation to be poorly performing at generating convincing synthetic data. In the paper by Theis et al. [46], the authors argue against using classical methods such as log-likelihood and Parzen window estimates as an evaluation metric for generative models of high-dimensional image data. They provide explicit examples of where these metrics fail, for instance when using nearest neighbour evaluation in the dataset of generated images, whereby perceptually small changes can lead to large changes in Euclidean distance and vice versa. For datasets with a large number of images, small changes can make the nearest neighbour qualitatively different. There is theoretical justification for this being the rule rather than the exception, see for instance [47]. Metrics such as the Inception Score [48], Fréchet Inception Distance [49] and precision/recall [50] have been devised to combat these problems, but suffer from their own disadvantages. This type of analysis and creation of metrics is what is currently missing in the literature for generating financial data, time series in particular.

Evaluating generative models for tabular data is still a nascent discipline with, to our knowledge, no widely established benchmarks, datasets or metrics. At the time of writing, the most common method is to adapt classical statistical methods such as the Kolmogorov-Smirnov test [51, 52], although more recently some effort has gone towards developing specialised metrics [53]. However, none of these methods provide a single real-valued metric for evaluating such models.

A number of different techniques are used in evaluating generative models for time series-type data. For example, Wiese et al. [31] and Li et al. [54] estimate statistical distances between the test data and generated data. Indeed, in light of the discussion by Theis et al. [46] one can not rely on these measures as a metric for realism. Alternatively, the evaluation of certain quantities such as Value-at-Risk (VaR) [55] is "backtested", whereby historical data for a time period is used to train a generative model, at which point the true VaR observed in the subsequent historical period is compared with that obtained by the generative model [32]. This method suffers from a dearth of data since there is only one history. There are alternatively, certain "stylized facts" that are widely observed in market microstructure data, that are often evaluated qualitatively [56, 57]. The stylized facts thus far have yet to be converted into a single real-valued metric in a convincing way.

5 Towards a compact representation of real data

Given the preceding discussion, it is worthwhile to consider a framework for how synthetic data can best be represented and transferred between different parties. Moreover, we can examine how existing techniques fit into this framework. The following properties of such a representation are desirable:

Privacy preserving. It is important to define the groups, entities and events whose privacy is to be protected, with tunable parameters for the trade-off between privacy and realism.

Human readable. It is desirable that the format in which synthetic data, or a generative model describing the synthetic data be readily interpretable without undue difficulty. In finance this is especially relevant as regulators, internal colleagues and other parties must have confidence in the representation and its privacy properties. An overly technical representation is less likely to be relevant and foster trust.

Compact. The representation of synthetic data should be such that it is significantly smaller than the real data it is derived from, and should be reconstructible on the receiver's end, ideally with open source software. The synthetic data should also require little technical know-how to generate, so that staff with direct access can synthesise the data on-premise.

With regard to the first point in the framework above, (ϵ, δ)-differential privacy provides a good model for tabular data, as seen in Section 3.1. For stream data and more structured data there is still not a consensus in the literature as to the best way for privacy to be defined. Indeed, even for tabular data differential privacy is not the last word, with future directions suggested by Kifer and Machanavajjhala [58].

On the second point, Eno and Thompson [18] define a synthetic data language for tabular data, which is is a promising direction although lacking any privacy guarantees. The *synthetic data vault* [59] is a more contemporary approach, with a synthetic database specified via JSON. Some aspects of the correlations between variables are captured, although a solution more tailored to financial data driven from more exotic distributions is required.

The most common ways of sharing synthetic data are either providing the generated dataset itself, or sharing learned parameters for the generative process. These two approaches have contrasting readability and compactness properties that need to be balanced on a case by case basis.

6 Conclusions

In this paper we have highlighted the unique challenges faced in generating financial data, and we have summarised the state of related literature at the time of writing. We describe the types of financial data that need to be protected by institutions, namely tabular data in retail banking and time series of market microstructure data. The state of the art in privacy-preserving generation and release of tabular data is more advanced than the equivalent task in stream data, perhaps due to the increased complexity of the latter.

Despite the various pitfalls described in our paper, we strongly believe in the importance and utility of researching methods to synthesize financial data as it will aid in building models to tackle issues of fairness and trustworthiness in financial market operations.

Disclaimer

References

[1] Chris Jay Hoofnagle, Bart van der Sloot, and Frederik Zuiderveen Borgesius. The European Union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.

[2] Joan Sieber. Family Educational Rights and Privacy Act (FERPA). *Journal of Empirical Research on Human Research Ethics*, 2(1):101–101, 2007.

[3] I. Glenn Cohen and Michelle M. Mello. HIPAA and Protecting Health Information in the 21st Century. *JAMA*, 320(3):231–232, 2018.

[4] Brian Tarran. What can we learn from the Facebook—Cambridge Analytica scandal? *Significance*, 15(3): 4–5, 2018.

[5] Long Cheng, Fang Liu, and Danfeng (Daphne) Yao. Enterprise data breach: causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(5): e1211, 2017.

[6] Niaz Kammoun, Ahmed Bounfour, Altay Özaygen, and Rokhaya Dieye. Financial market reaction to cyberattacks. *Cogent Economics & Finance*, 7(1):1645584, 2019.

[7] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[8] Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, 1998.

[9] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[10] Charu C. Aggarwal. On K-anonymity and the Curse of Dimensionality. In *Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 901–909. VLDB Endowment, 2005.

[11] Claire McKay Bowen and Fang Liu. Comparative study of differentially private data synthesis methods. *arXiv:1602.01063 [stat.ME]*, 2019.

[12] Jon P. Daries, Justin Reich, Jim Waldo, Elise M. Young, Jonathan Whittinghill, Andrew Dean Ho, Daniel Thomas Seaton, and Isaac Chuang. Privacy, Anonymity, and Big Data in the Social Sciences. *Commun. ACM*, 57(9):56–63, September 2014.

[13] Maryam Archie, Sophie Gershon, Abigail Katcoff, and Aaron Zeng. Who's watching? de-anonymization of netflix reviews using amazon reviews. Technical report, MIT, 2018.

[14] Logan Kugler. Protecting the 2020 census. *Commun. ACM*, 62(7):17–19, June 2019.

[15] M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. *Quantitative Finance*, 13(11):1709–1742, 2013.

[16] Frédéric Abergel, Marouane Anane, Anirban Chakraborti, Aymen Jedidi, and Ioane Muni Toke. *Limit order books*. Cambridge University Press, 2016.

[17] H. Surendra and H. S. Mohan. A review of synthetic data generation methods for privacy preserving data publishing. *International Journal of Scientific & Technology Research*, 4(8):95–101, 2015.

[18] J. Eno and C. W. Thompson. Generating synthetic data to match data mining patterns. *IEEE Internet Computing*, 12(3):78–82, May 2008.

[19] Jörg Drechsler. Using support vector machines for generating synthetic datasets. In *Proceedings of the 2010 International Conference on Privacy in Statistical Databases*, PSD'10, pages 148–161, Berlin, Heidelberg, 2010. Springer-Verlag.

[20] Gregory Caiola and Jerome P. Reiter. Random forests for generating partially synthetic, categorical data. *Trans. Data Privacy*, 3(1):27–42, April 2010.

[21] John M. Abowd and Lars Vilhuber. How protective are synthetic data? In Josep Domingo-Ferrer and Yücel Saygın, editors, *Privacy in Statistical Databases*, pages 239–246, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[22] Jun Zhang, Graham Cormode, Cecilia M. 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*, SIGMOD '14, pages 1423–1434, New York, NY, USA, 2014. ACM.

[23] Haoran Li, Li Xiong, Lifan Zhang, and Xiaoqian Jiang. DPSynthesizer: Differentially Private Data Synthesizer for Privacy Preserving Data Sharing. *Proc. VLDB Endow.*, 7(13):1677–1680, August 2014.

[24] Yubin Park and Joydeep Ghosh. PeGS: Perturbed Gibbs Samplers That Generate Privacy-Compliant Synthetic Data. *Trans. Data Privacy*, 7(3):253–282, December 2014.

[25] Jun Zhang. *Algorithms for Synthetic Data Release under Differential Privacy*. PhD thesis, Nanyang Technological University, 2016.

[26] M. Galbiati and K. Soramäki. An agent-based model of payment systems. *Journal of Economic Dynamics and Control*, 35(6):859 – 875, 2011.

[27] Luca Arciero, Claudia Biancotti, Leandro D'Aurizio, and Claudio Impenna. Exploring agent-based methods for the analysis of payment systems: A crisis model for starlogo tng. *Journal of Artificial Societies and Social Simulation*, 12(1):2.

[28] Edgar Alonso Lopez-Rojas and Stefan Axelsson. Using the retsim fraud simulation tool to set thresholds for triage of retail fraud. In Sonja Buchegger and Mads Dam, editors, *Secure IT Systems*, pages 156–171, Cham, 2015. Springer International Publishing.

[29] D. Platt. A comparison of economic agent-based model calibration methods. *arXiv:1902.05938 [q-fin.CP]*, 2019.

[30] Ruey Tsay. *Analysis of Financial Time Series*. Wiley & Sons, 3rd edition, 2010.

[31] Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant GANs: Deep Generation of Financial Time Series. *arXiv:1907.06673 [q-fin.MF]*, 2019.

[32] Rao Fu, Jie Chen, Shutian Zeng, Yiping Zhuang, and Agus Sudjianto. Time series simulation by conditional generative adversarial net. *arXiv:1904.11419 [stat.ML]*, 2019.

[33] David Byrd, Maria Hybinette, and Tucker Hybinette Balch. ABIDES: Towards High-Fidelity Market Simulation for AI Research. *arXiv:1904.12066 [cs.MA]*, 2019.

[34] Blake LeBaron. Agent-based Computational Finance. In Leigh Tesfatsion and Kenneth L. Judd, editors, *Handbook of Computational Economics*, volume 2 of *Handbook of Computational Economics*, chapter 24, pages 1187–1233. Elsevier, 2006.

[35] Cynthia Dwork. Differential Privacy in New Settings. In *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 174–183, Philadelphia, PA, USA, 2010. Society for Industrial and Applied Mathematics.

[36] Cynthia Dwork, Moni Naor, Toniann Pitassi, and Guy N. Rothblum. Differential Privacy Under Continual Observation. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 715–724, New York, NY, USA, 2010. ACM.

[37] Georgios Kellaris, Stavros Papadopoulos, Xiaokui Xiao, and Dimitris Papadias. Differentially Private Event Sequences over Infinite Streams. *Proc. VLDB Endow.*, 7(12):1155–1166, August 2014.

[38] Konstantinos Chatzikokolakis, Miguel E. Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. Broadening the scope of differential privacy using metrics. In Emiliano De Cristofaro and Matthew Wright, editors, *Privacy Enhancing Technologies*, pages 82–102, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[39] Ferdinando Fioretto and Pascal Van Hentenryck. Optstream: Releasing Time Series Privately. *J. Artif. Int. Res.*, 65(1):423–456, May 2019.

[40] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 909–910, Sep. 2015.

[41] G. Acs, L. Melis, C. Castelluccia, and E. De Cristofaro. Differentially Private Mixture of Generative Neural Networks. *IEEE Transactions on Knowledge and Data Engineering*, 31(6):1109–1121, June 2019.

[42] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In Michele Berlingerio, Francesco Bonchi, Thomas Gärtner, Neil Hurley, and Georgiana Ifrim, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 510–526, Cham, 2019. Springer International Publishing.

[43] A. Rahimi and B. Recht. Reflections on random kitchen sinks. 2017. URL `http://www.argmin.net/2017/12/05/kitchen-sinks/`.

[44] L. H. Gilpin, D. Bau, Yuan B. Z., A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. *The 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018).*, 2018.

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

[46] L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016. URL `http://arxiv.org/abs/1511.01844`. arXiv:1511.01844.

[47] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1lWUoA9FQ`.

[48] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 2234–2242, USA, 2016. Curran Associates Inc.

[49] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6629–6640, USA, 2017. Curran Associates Inc.

[50] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 5234–5243, USA, 2018. Curran Associates Inc.

[51] Alejandro Mottini, Alix Lheritier, and Rodrigo Acuna-Agost. *Presented at the 2018 ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models.*, 2019. URL `https://arxiv.org/abs/1807.06657`.

[52] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305, Boston, Massachusetts, 18–19 Aug 2017. PMLR.

[53] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *arXiv:1907.00503 [cs.LG]*, 2019. URL `https://arxiv.org/abs/1907.00503`.

[54] Junyi Li, Xintong Wang, Yaoyang Lin, Arunesh Sinha, and Michael P. Wellman. Generating realistic stock market order streams, 2019. URL `https://openreview.net/forum?id=rke41hC5Km`.

[55] Philippe Jorion. *Value at Risk: The New Benchmark for Managing Financial Risk*. McGraw-Hill, 3rd edition, 2006.

[56] Frédéric Abergel, Marouane Anane, Anirban Chakraborti, Aymen Jedidi, and Ioane Muni Toke. *Limit Order Books*. Cambridge University Press, 2016.

[57] Jean-Philippe Bouchaud, Julius Bonart, Jonathan Donier, and Martin Gould. *Trades, quotes and prices: financial markets under the microscope*. Cambridge University Press, Cambridge, 2018.

[58] Daniel Kifer and Ashwin Machanavajjhala. A Rigorous and Customizable Framework for Privacy. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS '12, pages 77–88, New York, NY, USA, 2012. ACM.

[59] N. Patki, R. Wedge, and K. Veeramachaneni. The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 399–410, Oct 2016.