Governing Synthetic Data in the Financial Sector

August 27, 2025

Taylor Spears
University of Edinburgh Business School
taylor.spears@ed.ac.uk

Kristian Bondo Hansen
Copenhagen Business School
kbh.msc@cbs.dk

Ruowen Xu
Warwick Business School
ruowen.xu@wbs.ac.uk

Yuval Millo
Warwick Business School
yuval.millo@wbs.ac.uk

Abstract

Synthetic datasets, artificially generated to mimic real-world data while maintaining anonymization, have emerged as a promising technology in the financial sector, attracting support from regulators and market participants as a solution to data privacy and scarcity challenges limiting machine learning deployment. This paper argues that synthetic data's effects on financial markets depend critically on how these technologies are embedded within existing machine learning infrastructural "stacks" rather than on their intrinsic properties. We identify three key tensions that will determine whether adoption proves beneficial or harmful: (1) data circulability versus opacity, particularly the "double opacity" problem arising from stacked machine learning systems, (2) model-induced scattering versus model-induced herding in market participant behaviour, and (3) flattening versus deepening of data platform power. These tensions directly correspond to core regulatory priorities around model risk management, systemic risk, and competition policy. Using financial audit as a case study, we demonstrate how these tensions interact in practice and propose governance frameworks, including a synthetic data labelling regime to preserve contextual information when datasets cross organizational boundaries.

1 Introduction

Synthetic datasets, artificially generated to mimic real-world data while maintaining anonymization (Jordon et al., 2022; Nikolenko, 2021), have emerged as a promising technology in the financial sector. While artificial data has long been used in financial markets for a variety of applications, in recent years advances in generative Artificial intelligence (AI) have allowed for the creation of artificial datasets that are indistinguishable from data generated by real-world processes. Because modern synthetic data promise to match the broad distributional properties of real-world data without reproducing them, synthetic data has attracted interest from a diverse range of financial market stakeholders who see it as a potential solution to a broad range of

technical and regulatory problems, including data privacy and sharing issues, problems related to data bias, and model overfitting (Assefa et al., 2020; Blach et al., 2024; Potluru et al., 2024). Support for increased use of synthetic data has even come from regulators and policymakers, such as the Financial Conduct Authority (FCA) and the European Commission (FCA, 2024; Di Girolamo, Hledik, and Pagano, 2024). This is somewhat surprising given that, in recent years, regulators have come to approach innovations in financial modelling with caution given the central role that novel financial modelling practices played in the 2008 financial crisis (MacKenzie, 2011) and earlier episodes, such as the 1987 stock market crash (MacKenzie, 2004). Indeed, even as regulators have embraced synthetic data, they are showing a deep uneasiness about the potential implications of AI on financial markets and institutions (Bank of England 2025).

In response to this enthusiasm for synthetic data—bordering on hype in some cases (Ravn, 2025)—an emerging academic literature in Critical Data Studies (CDS) and related fields has taken an increasingly sceptical eye to the use of synthetic data generation technologies (e.g., Steinhoff, 2024; Jacobsen, 2023; Offenhuber, 2024; Whitney and Norman, 2024). Yet there has been relatively little work exploring the potential effects of synthetic data on financial markets, despite significant interest in these techniques among market participants and regulators.

This paper serves as a call for more attention to the infrastructural dimension of synthetic data (Bowker and Star, 1999; Bernards and Campbell-Verduyn 2019; Westermeier, Campbell-Verduyn, and Brandl, 2025). Rather than focusing on the properties of synthetic data *itself*, this paper advocates for closer attention to the way that synthetic data generation techniques are being embedded into the broader "stack" of technologies, standards, and infrastructures that constitute machine learning (ML) systems (Straube 2016; Hansen 2024). As work in infrastructure studies has emphasised, embedding new technologies into existing financial market infrastructures can generate unexpected effects by reconfiguring relationships between different layers of the stack and among infrastructure owners and users (Jensen and Morita, 2017; Paraná, 2025). This infrastructural perspective, we argue, is critical for anticipating synthetic data's potential impact on markets. Whether synthetic data generation technologies prove to be beneficial or harmful, we argue, depends on the infrastructures and practices into which they are embedded. To this end, this paper identifies three broad sets of concerns around synthetic data which, depending on how synthetic data generators are integrated into market infrastructures, will create either beneficial or harmful effects. These are a tension between (i) synthetic data's capacity to facilitate data sharing versus the tendency of synthetic data generators to create new forms of model opacity, (ii) synthetic data's capacity to diversify data-driven decision-making by generating data corresponding to alternative futures and pasts versus its potential to induce new forms of model-induced isomorphism among financial market participants, and finally (iii) its potential political-economic effects on the market concentration of incumbent data platforms and owners.

2 Framing synthetic data

Synthetic data is typically defined in terms of its *use*, rather than the techniques used to generate it, which are diverse. Jordan, et al. (2022), writing in a widely cited report sponsored by the

Royal Society, define it as "data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s)" (Jordan, et al. 2022: 5). The concept of using artificially-generated data for statistical purposes is much older than modern data science and ML. Some of the earliest references to "synthetic data" come from papers published in the applied economics literature in the 1960s and 1970s, in which researchers associated with national statistical agencies constructed "synthetic microdata" sets by merging datasets collected from different samples and matching demographic information in order to study questions requiring data from multiple datasets (cf., Ruggles and Ruggles, 1968; Sisson 1979). In the financial markets, artificial data produced via Monte Carlo simulation has long been used for a variety of problems, including derivatives pricing and the estimation of Value-at-Risk (VaR) for risk management purposes (Jackel, 2002).

However, recent years have seen the emergence of synthetic data as a distinct research and practice area in the fields of data science and ML. There are several reasons for this. First, recent years have seen the emergence of new generative AI techniques—such as generative adversarial networks (GANs) and large language models (LLMs)—capable of generating synthetic data that is virtually indistinguishable from "real-world" data (Goodfellow, et al., 2020; Nadas et al. 2025). These ML techniques produce data that exhibit a level of realism that cannot be easily matched using more traditional statistical techniques, such as Monte Carlo simulation applied to stochastic differential equations, the workhorses of much of quantitative finance. GANs, for instance, are the ML technology that underpins the production of many "deep fakes": images of humans that are nearly indistinguishable from photographs of real people. Partly as a consequence of these modelling developments, modern synthetic data generating techniques are extremely diverse (Lu et al., 2023), ranging from 3D modelling techniques in the field of computer vision, to traditional simulation-based methods, such as ABMs agent-based models (ABMs) (Axtell and Farmer, 2022), to advanced generative AI techniques like variational auto-encoders (VAEs), as well as GANs and LLMs (Kingma and Welling, 2019). While simulation-based methods generate data from scratch by mimicking real-world processes, ML-driven techniques instead use "real" data to train a neural network to produce synthetic copies of that data that match its broad distributional properties, but which do not match the original data on a case-by-case basis.

Second, apart from these modelling developments, interest in synthetic data has grown because it provides a technological 'fix' to several barriers that constrain the adoption of ML. This is particularly true in domains where there are significant concerns around data privacy, or in situations where there are inadequate quantities of "real-world" data to train ML models (Hansen and Spears 2025). In the case of privacy-sensitive applications of ML, synthetic data is sought because it can, in principle, closely match the distributional properties of real-world data – e.g. the share of borrowers of a particular background who default on a loan – without revealing any personally-identifiable information about the specific individuals contained in the original "real-world" dataset. In activities such as credit risk modelling, synthetic data is extremely attractive to ML practitioners because the General Data Protection Regulation (GDPR) and other regulations make it increasingly cumbersome for financial services firms to share credit

default data across, and in some cases within, firms. Likewise, where privacy is not an issue but data is scarce, synthetic data generators are attractive because they allow ML practitioners to generate large quantities of data that are statistically indistinguishable from real-world data to train their models. For example, one of the most prominent uses of synthetic data for this purpose is associated with an ML-driven system for automated options hedging developed jointly by quants at J.P. Morgan and ETH-Zurich known as *Deep Hedging* (Buehler et al. 2019). Because real-world options data is too sparse to train a deep learning system, the Deep Hedging system is trained using data produced by using what quants call a *market generator* – a separate ML system that can produce large quantities of realistic synthetic market data using a smaller amount of real-world price data (c.f. Buehler et al. 2023; Kondratyev and Schwartz 2019).

In the context of finance, certain global systemically important banks (G-SIBs), including J.P. Morgan and Goldman Sachs, have embraced synthetic data and actively promote its use in different contexts (Assefa, 2020; Bansel and Stefani, 2024; Coletta et al., 2021). A report by the International Monetary Fund notes the acceleration in the use of synthetic data in finance and highlights cost-effectiveness, privacy-protection, and debiasing as features that make synthetic data attractive to the financial industry (Shabsigh and Boukherouaa, 2023). While financial regulators have expressed serious concerns about the risks of machine learning (ML) and AI on financial markets, they have instead shown considerable enthusiasm for the use of synthetic data. This is due to the way that synthetic data generation technology could help to meet regulatory objectives, including the promotion of competition in the financial markets and the enforcement of data protection laws. For example, the UK Financial Conduct Authority (FCA) has produced several reports on possible use cases of synthetic data in the financial industry; in these reports, it emphasizes that the shareability of synthetic data could contribute to "democratizing data access" by lowering barriers to entry for challenger firms seeking to bring new financial products to market (FCA, 2023). The European Union has adopted a similar attitude with the development of their Data Hub on their Digital Finance Platform. The Data Hub is meant to facilitate data exchange between financial firms and supervisory authorities through the provision of synthetic supervisory data for the purpose of testing and training AI and ML models (Girolamo et al., 2024). However, some regulators and institutions have also pointed to potential risks associated with the use of synthetic data in finance. For example, The US Commodity Futures Trade Commission (CFTC) has raised concerns about data quality in financial risk management if "data gaps" are filled with synthetic data that could lead to inaccurate information. The CFTC's main concern is with the risk of so-called "AI hallucinations" (Romero et al., 2024). Similarly, the UK Financial Conduct Authority (FCA, 2024) has highlighted the importance of careful judgment in the creation of synthetic data. They emphasize that the judgement in selecting appropriate technologies and models is crucial, depending on specific use cases and evaluation metric. For instance, the acceptable levels of privacy, utility, and fidelity[1] of synthetic data are likely to be influenced by its intended application, which involves trade-off and tension.

Although the use of synthetic data in the financial industry has yet to receive attention in

[1]Fidelity pertains to the measure of how closely the synthetic dataset resembles the real dataset, specifically in terms of statistical similarity between the synthetic data and the original real data.

economic sociology and in SSF, a budding literature on synthetic data is emerging in CDS. This emerging body of CDS research covers a range of themes including privacy (Munkholm and Weihn, 2025), governance and regulation (Beduschi, 2024; Gal and Lynskey, 2023), surveillance (Ravn, 2024a; Ridgway and Malevé, 2024), capital and labour (Steinhoff, 2024), synthetic media such as deepfakes or other deceitful media (de Vries, 2020; Ferrari and McKelvey, 2023; Fitzgerald, 2024; Martin and Newell, 2024), ethico-politics (Helm, Lipp, and Pujadas, 2024; Jacobsen, 2024; Ravn, 2024b), the representational logics of data types (Offenhuber, 2024; Susser and Seeman, 2024), data pollution (Wiehn, 2024), and the construction of synthetic data promises in science and industry (Ravn, 2025). Particularly relevant to this paper are discussions of synthetic data as a technology that offsets risk associated with the use of ML for process- and decision-automation. Jacobsen (2023) conceptualizes synthetic data as a "technology of risk" that "de-risks" ML models by attributing risk solely to the real-world data on which such models are trained: "By shifting the domain of risk to the 'real' dataset," Jacobsen (2023: 8) argues, "synthetic data promise to be the means by which algorithms can be rendered free from their own manufactured uncertainties." Understood as a derisking technology, synthetic data generation becomes a technology that enables the development and use of ML and AI models.

We build on this recent work by shifting analytical attention from the properties of synthetic data itself to the ways that synthetic data generation techniques are being embedded within existing *stacks* in finance, particularly those that underpin the operation of automated machine learning systems. In employing the term 'stack', we build on a growing body of work in sociology, political economy, and science and technology studies (STS) that employs stack-based theorizing to understand developments at the intersection of computing and finance (Straube, 2016; Caliskan, 2020; Hansen, 2025; MacKenzie, Caliskan, Roomerskirchen, 2023). This work has in common an attentiveness to the way that modern computing and financial products are typically created by linking or "stacking" together multiple technologies and infrastructures. The term "stack" has its origins in computer networking; it originally referred to the "stacked" nature of protocols that underpin modern Internet communications, particularly in the Open Systems Interconnection (OSI) model. Stack diagrams, like that of the OSI, capture the logical dependency of protocols "higher" in the stack on those "lower" in the stack. At the bottom of the stack is the physical transmission of bits of information across the material infrastructure of the network: cables, switches, and servers. On top of this are layered basic Internet protocols - e.g. TCP/IP, the DNS system, and application-level protocols, like HTTP for information displayed on the web. At the top of the protocol "stack" are user applications that depend on these lower "layers" to function. The stack diagram is thus not a description of layering in Euclidean space, but one of logical dependency: changes "lower" in the stack propagate upward, which can lead to failure or unexpected effects on layers higher in the stack. Stack-based theorizing in the social sciences has its origin in the writing of Bratton (2016), who uses the term to describe an emergent political order that has emerged consisting of interlinked computing infrastructure, protocols, and applications, which intersects and challenges more traditional forms of state-centred power.

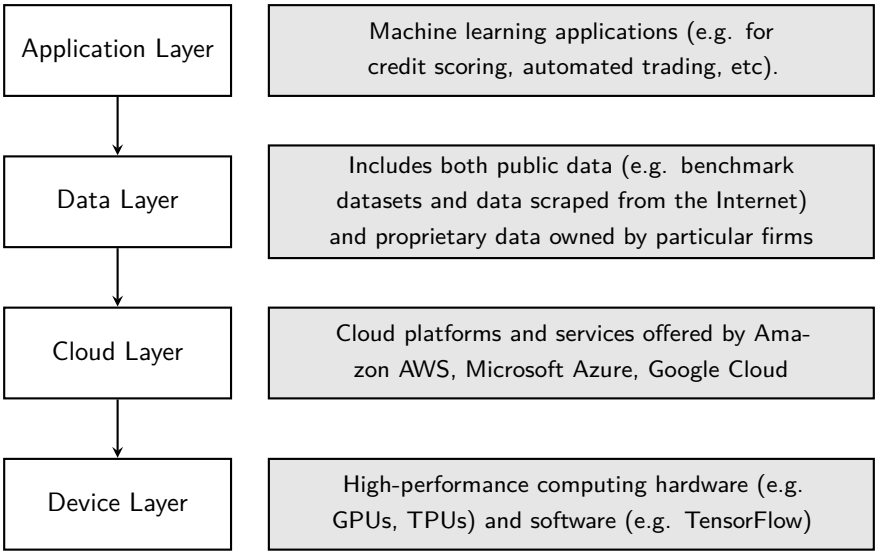

Figure 1: Illustration of the Machine Learning Stack

In the case of machine learning systems, it has become increasingly common for practitioners to refer to a similar layered stack that organizes the hardware, data and software that organizes modern automated ML systems. We provide a simple four-layer depiction of an ML stack in Figure 1. At the bottom is what we call the device layer, which includes the physical computing hardware underpinning modern ML systems, such as graphics and tensor processing units (GPUs and TPUs), networking infrastructure, and the software used for training neural networks (e.g. TensorFlow). Above this is the cloud layer, consisting of cloud platforms and services provided by Amazon, Microsoft, and Google, from which ML practitioners can access computing infrastructure at the device layer through standardized interfaces. Stacked above this layer is a data layer, which captures the data platforms and systems that produce and store data that are used to train financial ML systems. This layer encompasses both proprietary and public datasets. And finally on top is the application layer, which corresponds to the ML systems that are developed and trained by practitioners for specific purposes (e.g. trading, risk management, fraud detection, etc.). With the emergence of generative AI techniques, particularly large language models (LLMs), some have argued that a new, fifth layer is emerging that sits on top of the cloud and data layers: that of foundation model providers like OpenAI, Anthropic, and Mistral (Gambacorta and Shretti, 2025). Both proprietary foundation models, such as GPT-4, and open-source synthetic data generators making direct use of the data layer are likely to be increasingly used to generate synthetic data in financial markets (Nadas, et al., 2025).

In this paper, we argue that synthetic data generation techniques are likely to significantly alter the data layer of the stack, which is likely to propagate uncertain effects 'up the stack' into the application layer as the use of synthetic data generation technologies shapes financial market participant behaviour and reconfigures power relations between different layers of the financial ML stack. In making this argument, we follow a line of thinking in infrastructure studies which has broadly emphasised the way that changes to the design or operation of market infrastructures can induce significant changes in market behaviour (Paraná, 2025; Arnoldi, 2016;

Spears, 2019; Pinzur, 2016). In what follows, we outline three tensions related to the use of synthetic data generation as a technological enabler of ML/AI in finance. What the discussions of these tensions suggest is that there are governance-issues looming as the use of synthetic data in finance picks up.

3 Synthetic data generation technology: three tensions

3.1 First tension: Data circulability versus opacity

First, synthetic data generation techniques have a remarkable capacity to decouple data from the conditions and local context of their production, as data collected for one purpose, will be regenerated to serve different predictive purpose, enabling applications beyond the original intent. This, indeed, is key to its promise to enable enhanced data sharing among firms that are proving attractive to financial market participants and regulators alike. Yet, as a long tradition of work in Science and Technology Studies (STS) has emphasized, the production of data is deeply shaped by social, historical, and discipline-specific practices (Bowker and Star 1999; Bowker et al., 2019). Far from being an abstract point about the ineluctably socially constructed nature of data, this context can be critically important for consumers of synthetic data—namely, ML system developers and engineers—using it to train ML algorithms in contexts distant from that in which the original data was created. By decoupling data from the context of its production, consumers of synthetic data may overlook its limitations, biases, including its inherent temporal structures (Preda, 2008). Financial data related to past transactions differs significantly from those concerning ongoing transactions. The temporal structures embedded within synthetic data generation influence how users respond to this data, as the timing of data availability and release may critically shape their contextual reactions in decision-making processes.

This tension is particularly acute in cases where synthetic data is used to address a paucity of real-world data that otherwise prevents the adoption of ML systems. In the options markets, for instance, historical data on options transactions is relatively limited compared to the vast number of quoted options prices offered by dealers and exchanges daily. Historically this has limited the development of automated systems for options market making (Buehler et al., 2019). Likewise, in the equities markets, historical market data are limited except for high-frequency price data and thus not available at the scope required to properly train ML models for systematic investment strategies operating over longer timescales (Arnott et al., 2019). In these and other "data-limited regimes", domains in which data are scarce and/or expensive to obtain (Hoffmann et al. 2019), synthetic data generators are being used to generate large quantities of historical data from relatively small real-world historical datasets (cf., Heaton and Witte, 2019; Buehler et al., 2020; Limmer and Horvath, 2023). However, these synthetic data generators run the risk of amplifying noise in the original, rather limited, training datasets.

In addition to the original context and conditions underpinning the production of the "real-world" data used to produce it, synthetic data are also shaped by design choices made by the developer of the synthetic data generator itself. Synthetic data generation techniques are numerous; they span ML techniques, non-ML techniques such as agent-based models, and even

more traditional statistical techniques such as Gaussian copulas. Even in the case of "data-driven" ML models, there are several key design decisions that must be made by the model builder and cannot be "learned" from data itself. Among other things, these include how many layers the neural network will have, what type of activation functions to use between layers, which regularization technique to use (Mullainathan and Speiss, 2017). All these decisions involve a mix of experimentation and subjective judgment on the part of the model developer. These choices, which can shape the synthetic data produced by the generator, ultimately must be made through a combination of domain-specific, theoretical, and experimental knowledge, along with knowledge of how the synthetic data will be used in practice. Synthetic data are thus not only a byproduct of the conditions underpinning the production of the "real-world" data upon which they are derived, but they are also fundamentally "model-laden" in a way that is difficult to detect than data produced via a classical statistical model using Monte Carlo Simulation (Bokulich, 2020; Offenhuber, 2024). Although no data are "raw" and all are somewhat "cooked" (Bowker, 2005; Gitelman, 2013) or even "model-filtered" (Edwards, 1999), synthetic datasets are ultra-processed in the sense that they are derivatives transposed from already "cooked" data.

For this reason, when synthetic data is used to train other automated systems, it can potentially induce hidden forms of model overfitting (i.e., the situation in which a model learns patterns that are specific to the training data that do not reflect the patterns they are likely to encounter when put into production (Arnott et al., 2019)). Drawing on classic work on the opacity of ML systems (e.g., Burrell 2016), we refer to this as the "double opacity" of synthetic data, which arses from the way that synthetic data generators effectively "stack" multiple forms of black boxed ML on top of each other.

3.2 Second tension: Model-induced herding versus scattering

A second set of tensions relate to how the embedding of synthetic data generator technology into the data layer of the ML stack may shape the behaviour of ML systems and that of financial market participants using these systems to make decisions. On the one hand, we know from classic work in the SSF that when financial models are embedded into market participants' cognitive or decision-making processes, they can exert isomorphic pressures on their behaviour, which can have a destabilizing effect on markets (e.g., MacKenzie and Millo, 2003; Beunza and Stark, 2012; Svetlova, 2012). Strategy isomorphism is not a new phenomenon in the hedge fund sector, nor is it specific to quant funds. Studies have shown that inter-personal and inter-institutional social ties and communication practices between competing funds can be conducive to strategy conformism and increase herding risk in the sector (Kellard et al., 2016; Millo, Spence, and Valentine, 2023). However, herding can also be intentional as a reflexive calculated decision to imitate others, which is not the same as succumbing to irrationality (Beunza and Garud, 2007; cf. Lange, 2016). Along similar lines, Beunza and Stark (2012) have shown how derivatives traders use models to get social cues on what their competitors think. By taking competitors' social cues into consideration, traders perform "reflexive modelling" whereby they adjust and fine-tune their trading strategies to create dissonance vis-à-vis the competition. While reflexive modelling can improve trading through the creation of dissonance, errors can on the other

hand accumulate if many funds use the same models, which creates resonance (Beunza and Stark, 2012). The possibility of doing reflexive modelling is, however, limited in contemporary automated trading where the human trader has become more of an appendage to an automated system than an executor of a trading strategy. In areas like high-frequency trading, resonance in the form of cognitive interdependence is largely replaced by infrastructural and algorithmic interdependence (Borch, 2016).

Given the potentially widespread application of synthetic data generators in the near future, this is a realistic possibility that needs to be considered. In general, regulators have expressed concerns about the potential systemic risks associated with deep learning, although not specifically deep learning-based synthetic data generation techniques, such as GANs or LLMs. Not so long before his term at the helm of the US Securities and Exchange Commission (SEC) came to an end in 2024, Gary Gensler shared concerns about the systemic risk threat he believes that AI poses to financial market stability, an issue he had previously raised with specific emphasis on the role of deep learning (Gensler and Bailey, 2020). In a short YouTube video, in the series 'Office Hours with Gary Gensler', Gensler argued that a rather small number of leading generative AI companies and cloud providers are dominating the market. These companies' dominant position has, he stated, implications for the economy writ large, but also for the financial industry more specifically. The core issue is that when financial firms build downstream AI applications, as they do at a large scale, they rely on only a few base models or "data aggregators". Data aggregators are foundational models or AI model development infrastructures on and through which AI applications are built. Gensler's concern is that the widespread proliferation of AI applications, built on a few base models, could create network interconnectedness, monocultures of model design, limited model explainability, and uniformity of data (Gensler and Bailey, 2020: 4-5).[2] This cocktail of problems would, Gensler believes, increase systemic risks, including herding risks, among financial market participants. Because model risk management tends to work at the micro-prudential (individual firm-level), these types of interconnection and interoperability risks tend to evade those guardrails (Gensler and Bailey, 2020: 4-5).

While Gensler articulates a very real concern, because synthetic data can potentially allow for the generation of more diverse training sets than would otherwise be available to market participants, their widespread use may instead diversify market participant behaviour, a phenomenon that we call "model-induced scattering", in contrast to herding. This is particularly the case in which synthetic data is used to augment historical data to prevent what is known as "backtest overfitting", a well-known problem where ML algorithms over-index on past events, thereby ignoring possible events that could have happened but did not and which might happen in the future.

Backtest overfitting is arguably the biggest headache of developers of ML solutions for trading and investment management (Arnott et al., 2019). A backtest consists of two phases: an in-sample training phase and an out-of-sample test and validation phase. The training and test

[2] A similar sounding concern was raised in the CFTC's Technology Advisory Committee's report on responsible AI in financial markets. Here they point to what the term "critical infrastructure dependence risk" that arises when financial firms become "more reliant on a small number of AI vendors and platforms" (Romero et al., 2024: 64-65).

datasets come from the same dataset, but the former tends to be larger than the latter. Moreover, the training set is labelled in a way that it is known what the model is taught to predict, whereas the test set only contains labels for evaluation purposes. A model is overfitting if it performs impeccably well on training data but poorly on test data (Hansen, 2020; Mullainathan and Spiess, 2017). If a model performs poorly both in- and out-of-sample means that the model is too simple to learn underlying patterns in the data, which means that the model is underfit. Models that overfit in backtests are likely to perform poorly if put into production. While data quantity is the primary culprit when it comes to backtest overfitting risk, data quality is another potential course of overfitting. Sometimes the root cause of the overfitting problem is exactly a combination of limited historical data and poor data quality. As hedge fund quant and researcher Marcos Lopez de Prado (2018) points out, there is this dual issue with historical market data: rarely is there enough of it to properly train ML models and, secondly, history might not be the best proxy for what the future will look like.

Synthetic data generation promises to alleviate this data quant-qual problem that is causing of financial ML models to overfit in the backtest. Framing the issue, Lopez de Prado (2018: 170) notes that training on historical data can cause the trading strategy to become "so attached to the past that it becomes unfit for the future" (Lopez de Prado, 2018: 170). This data quality issue can be resolved, Lopez de Prado (2018: 170) continues if the parameters for the trading rules are derived "directly from the stochastic process that generates the data, rather than engaging in historical simulations". The problem of data quantity, on the other hand, simply concerns the size or lack thereof of the training sets used for model development. What synthetic data generation promises in terms of addressing the data quantity issue associated with backtest overfitting is to produce multiple virtual training sets. As Lopez de Prado (2020: 9) frames it, "while it is easy to overfit a model to one test set, it is hard to overfit a model to thousands of test sets for each security" Collapsing the problem of data quantity and that of data quality as it pertains to backtest overfitting risk, he summarises:

> We can use historical series to estimate the underlying data-generating process, and sample synthetic data sets that match the statistical properties observed in history. Monte Carlo methods are particularly powerful at producing synthetic data sets that match the statistical properties of a historical series. [...] The main advantage of this approach is that those conclusions are not connected to a particular (observed) realization of the data-generating process but to an entire distribution of random realizations. (Lopez de Prado, 2020: 9)

The judgment here lies in balancing these two tensions—understanding when to prioritize the data herding and the potential for stronger collective performance versus the benefits of diverse, scattered approaches that can avoid overfitting and better account for variability and uncertainty.

3.3 Third tension: Flattening versus deepening of data platform power

A final set of uncertainties relates to the potential political-economic effects of synthetic data generation technologies on financial markets. On the one hand, with the growth of platform-based business models in finance, competitive advantage in markets is increasingly secured through control and ownership of the data layer of the financial ML stack, which can be used to train ML models (Birch, Cochrane, and Ward, 2021). It is for this reason that the Economist famously quipped in 2017 that "the most valuable resource is no longer oil, but data" (Economist 2017).[3] In the financial markets, this trend is manifest in the growing power of data-driven fintech platforms, the entry of so-called "TechFins" into credit issuance, and the growing centrality of data infrastructure provision as a core component of incumbent financial institutions' business (cf., Cornelli et al., 2023; Hansen and Borch, 2022; Petry, 2021). Because synthetic data generation technologies can decouple the information content of data from the data platforms that own them, they may potentially have a levelling effect on financial institutions insofar as ownership of data assets may become a weaker source of a platform's competitive advantage. Indeed, it is in part for this reason that regulators such as the Financial Conduct Authority (FCA) have supported the use of synthetic data technology, due to its promise to make datasets available that could promote competition in markets by allowing new fintechs to challenge the market power of incumbent financial institutions, such as banks (FCA, 2021; 2022). However, it is also possible that synthetic data generation technology may allow for the further monetization of the data layer by platform owners, thereby increasing—rather than challenging— their power.

LLMs represent one synthetic data generation technology in which this tension between the flattening and deepening of platform power is likely to play out. Unlike traditional machine learning techniques, or even generative AI techniques like GANs, training new LLMs is an extremely capital-intensive process, one that only a relatively small number of firms possess the resources to carry out. In March 2023, Bloomberg LLP announced BloombergGPT, a specialized LLM trained on Bloomberg's significant proprietary data assets (Wu et al. 2023). To the extent that LLMs will become a critical technology for synthetic data production, tools like BloombergGPT will allow firms such as Bloomberg to reinforce their platform dominance. At the same time, foundation model providers like Anthropic are developing specialized LLM pipelines for financial services applications, which allow financial institutions to use a customized LLM platform on their own proprietary data without training such a model from scratch. By taking advantage of general purpose LLMs' capacity to generate realistic synthetic data from a small number of example cases (what is known as "few shot" learning), these product offerings by major foundation model providers may instead erode financial data owners' own platform advantages (Meng et al. 2023; Ren et al. 2025).

[3] We recognise that the adage "data is the new oil" is older than the *Economist* article. It was allegedly declared by the British mathematician Clive Humby in 2006 (Authur, 2013).

4 Examining the interplay of the three tensions: the case of financial audit

Our discussion so far has focussed on the use of synthetic data in a financial markets context. However, the growth of machine learning in financial services has largely taken place in areas peripheral to the trading floor, such as compliance, risk management, and customer management (Spears and Hansen 2025; Bank of England and FCA 2022; Bank of England 2025). Therefore, to examine how these tensions may play out in the future, it is helpful to consider the application of synthetic data in financial audit, another non-trading domain where there has been significant investment in machine learning. In this domain, the three tensions that we discussed above are likely to be particularly acute. In contrast to ML systems in financial markets, where data sources are typically evaluated pragmatically by their capacity to improve a model's predictive power, accounting data is evaluated according to the norms of generally accepted accounting practices, which emphasises the importance of accounting information's *representational faithfulness* (IFRS Foundation, 2018). Auditing, as an established practice, thus relies on a clear and verifiable trail of real data sources, enabling auditors to sample, trace, verify, and draw conclusions about an entity's financial position. This creates inherent tensions between the uses of synthetic data and the norms of accounting. One recent episode that exemplifies those tensions is the case of Frank, a student loan fintech that J.P. Morgan Chase acquired in 2021. Not long after the acquisition, it emerged that Charlie Javice, Frank's CEO prior to its acquisition, had inflated its reported customer metrics during the due-diligence process with J.P. Morgan by using synthetic data generation techniques to produce fraudulent customer data (Staiger 2025).

Yet, in the domain of audit, there is growing interest in the use of ML to improve the efficiency of financial audits and reduce auditor workload, which is substantial (Brown-Liburd et al., 2015). According to an April 2024 survey by KPMG, sixty-three percent of corporate board members surveyed by the firm across 1,800 companies believed that auditors should prioritize the use of AI for identifying risks and anomalies (KPMG 2025: 14). In this domain, the use of synthetic data is likely to be prioritized given the constraints that auditors face in using and aggregating company-specific data. Consider the problem of building a ML classifier to detect anomalous financial accounting data, which might be indicative of fraud or reporting error (Aftabi, et al., 2023; Wang et al., 2024). Here one would encounter several problems that synthetic data generators promise to address. The first is the issue of privacy and data protection. Granular financial data often involves personally identifiable information; because data protection regulations such as the EU General Data Protection Regulation (GDPR) constrain the circulation of such information, a model developer would likely encounter difficulty assembling enough "real" granular financial accounting data to build a robust model. However, even if a large enough dataset of "real" financial data *could* be assembled, an auditor would encounter a second problem: anomalous data are, by definition, rare. If one were to build an ML classifier on a dataset consisting of a comparatively large number of transactions cases and a small number of anomalous cases, the ML classifier would likely exhibit a high false positive rate (i.e., incorrectly classifying anomalous cases as "normal"). ML practitioners refer to this form of overfitting as a "class imbalance" problem (Fernandez et al., 2018; Singh, Ranjan, and Tiwari, 2022).

To address these two issues, a model developer could use a synthetic data generator, trained on a relatively small number of "real" transactions, to build a balanced synthetic dataset of anomalous and typical data. This synthetic dataset could then be circulated to another firm, where it is used to train an ML classifier to detect anomalous data, which could then be embedded into an auditor's workflow. In principle, the use of synthetic data here *could* significantly improve an auditor's capacity to detect potential misstatements of fact. However, the potential impact of synthetic data usage in this case will depend on how the three tensions we identify above play out in practice. First, consider the "double opacity" of the synthetic data used to train the classifier model. Because these synthetic data are already labelled as "anomalous" or "typical", they already embody subjective judgments regarding the materiality of the misstatement. Yet a crucial professional responsibility of auditors is to determine whether a given misstatement is "material" in the sense that it "could reasonably be expected to influence the economic decisions of users taken on the basis of the financial statements" (FRC 2016; Carpenter, et al. 1994). In this context, the double opacity of synthetic data may have the unintended effect of undermining auditors' autonomy in exercising professional scepticism and oversight regarding traceability. Second, we might consider the potential effects of this technology on the behaviour of financial statement users, namely investors. Consistent with Gensler's concerns about the impact of data aggregators on financial markets, one can imagine how widespread adoption of a particular anomaly detection model by auditors could shape investor behaviour in systematic ways. For instance, if the original data used to train the synthetic data generator upon which the fraud detection model depends does not generalize well to the financial data encountered by auditors using the model, it could potentially fail to detect certain types of material omissions in financial statements. On the other hand, this problem could be ameliorated if synthetic data generators are employed by model developers who are "close" to model end-users (e.g., auditors) and who have a high-level of awareness of the intended use of the synthetic data.

Finally, one could consider the potential political-economic effects of the widespread use of synthetic data on the market for auditing services. At present, this market is highly concentrated, with the Big Four accounting firms EY, Deloitte, KPMG, and PwC, earning 98% of audit fees from the FTSE 350 in 2022 (Financial Reporting Council, 2023). The Big Four accounting firms are also seeking to develop AI agents (PwC, 2025) that will be trained on extensive proprietary data to enhance user experience. The development and adoption of synthetic data could have a similar role on the market for audit services to the one expected by financial regulators on the financial markets. The development of publicly available synthetic datasets to detect financial misstatements could help to lower barriers to entry for challenger firms in the market for audit services. On the other hand, the Big Four may also be able to use synthetic data generation technology to find new ways to monetize their existing data assets.

5 Conclusion

This paper has examined the use of synthetic data generators in the financial sector and discussed regulatory and governance-issues that arise from the increasing use of such data for the purpose of developing and training ML and AI models.

Our analysis makes two key contributions to the emerging literature on synthetic data in finance. First, we argue for moving beyond assessments of synthetic data that focus on its intrinsic properties. Instead, we focus on how synthetic data generation technologies will likely be integrated into existing financial practices and infrastructures, and their resulting effects on financial markets and institutions. To this end, we employ the metaphor of the financial ML "stack", the layered technology infrastructure underlying ML systems, to explore how synthetic data may shape financial market participant behaviour and the power of existing financial data platforms. Second, our article demonstrates that the adoption of synthetic data generation technologies by financial market participants presents a "double-edged sword" for financial regulators. Whether these technologies are likely to, on net, amplify or ameliorate model-related risks, depends on the specific ways that these techniques are layered into existing practices and infrastructures underpinning ML systems. In particular, we identify three important tensions between the promises and potential risks of generating and using synthetic data in the sector. First, we noted a tension between synthetic data's capacity to increase the circulability of proprietary datasets and the attendant risk of increased model opacity that such sharing would induce. Second, we discussed the tension between the synthetic data's capacity to diversify market participant behaviour by facilitating the generation of diverse training sets for ML algorithms versus its potential capacity to induce new forms of isomorphic behaviour among participants. Finally, we examined the tension between synthetic data's ability to promote competition in the financial sector by lowering entry barriers in data-intensive markets and its potential to reinforce the market power of data platform owners by creating new opportunities for data monetization.

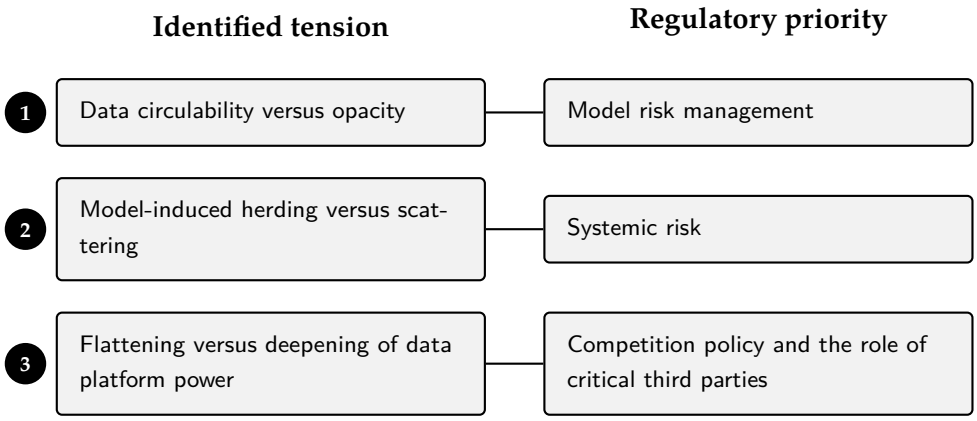

Figure 2: Correspondence between synthetic data generation tensions and key regulatory priorities

It is important to note that the three tensions we identify in this paper broadly correspond to three existing priority areas for financial regulators. The first tension, between model circulability and opacity, broadly aligns with regulators' concerns around model risk management. Our second tension, between model-induced scattering and isomorphism, aligns closely with regulators' macroprudential concerns around systemic risk. Finally, our third tension – concerning the flattening and deepening of data platforms' power – speaks to emerging concerns around competition policy in financial services and the role of critical third parties. We propose that

regulators as well as firms involved in the generation and use of synthetic data in the financial sector reflect on these tensions to better navigate the necessary trade-offs of wielding synthetic data.

Take as an example the issue of double opacity. Current model risk regulations (SR 11-7 in the US, SS1/23 in the UK) establish comprehensive governance, validation, and testing requirements for financial models, applying these standards to both internally-developed models and externally-sourced models (e.g. from software vendors). These regulations also contain validation requirements for assessing data quality. SR 11-7, the model risk regulation for US financial institutions, requires that data be assessed for accuracy and relevance, while SS1/23 provides more extensive requirements to ensure that data be suitable for intended use, representative, and free from potential bias. However, a weakness of these regulations in the case of synthetic data is that they place the validation burden on the data user without ensuring that synthetic data providers disclose relevant contextual information that data users might need to satisfy these requirements. Crucially, synthetic data is itself a model output, and the increased circulation of synthetic data—even to address legitimate regulatory and compliance issues like data protection—is likely to induce new model linkages and dependencies across institutions that may go unnoticed by current model risk regulations. Ensuring that metadata capturing the context and limitations of synthetic data is included when such data is shared could help to ameliorate these risks by allowing synthetic data users to better understand its limitations.

For this reason, we suggest that regulators and policymakers consider developing a synthetic data labelling regime to ensure that the context underpinning the production of synthetic datasets is maintained at the point they are transmitted across organizational boundaries and between market participants. Meanwhile, we propose to recognise that each tension represents a trade-off that requires careful judgement. For instance, there may be a tension between data privacy and data utility, where maximising one may compromise the other. This trade-off necessitates that stakeholders weigh the benefits and risks associated with each option to make informed decisions.

References

Aftabi, S. Z., Ahmadi, A., & Farzi, S. (2023). Fraud detection in financial statements using data mining and GAN models. *Expert Systems with Applications*, *227*, 120144.

Arnoldi, J. (2016). Computer algorithms, market manipulation and the institutionalization of high frequency trading. *Theory, Culture & Society*, *33*(1), 29–52. Retrieved from `https://doi.org/10.1177/0263276414566642`

Arnott, R., Harvey, C. R., & Markowitz, H. (2019). A backtesting protocol in the era of machine learning. *The Journal of Financial Data Science*, *1*(1), 64–74.

Arthur, C. (2013). *Tech giants may be huge, but nothing matches big data.* The Guardian. Retrieved from `https://www.theguardian.com/technology/2013/aug/23/tech-giants-data` (23 August 2013)

Assefa, S., Dervovic, D., Mahfouz, M., Blach, T., Reddy, P., & Veloso, M. (2020). *Generating synthetic data in finance: Opportunities, challenges and pitfalls.* SSRN Scholarly Paper. doi: 10.2139/ssrn.3634235

Austin, A. A., Carpenter, T. D., Christ, M. H., & Nielson, C. S. (2021). The data analytics journey: Interactions among auditors, managers, regulation, and technology. *Contemporary Accounting Research*, *38*(3), 1888–1924.

Axtell, R. L., & Farmer, J. D. (2022). *Agent-based modeling in economics and finance: Past, present, and future* (Working Paper No. 2022-10). INET Oxford.

Balch, T., Potluru, V. K., Paramanand, D., & Veloso, M. (2024). *Six levels of privacy: A framework for financial synthetic data.* arXiv. doi: 10.48550/arXiv.2403.14724

Bank of England and Financial Conduct Authority. (2022). *Machine learning in UK financial services* (Tech. Rep.). Bank of England. Retrieved from `https://www.bankofengland.co.uk/-/media/boe/files/report/2019/machine-learning-in-uk-financial-services.pdf?la=en&hash=F8CA6EE7A5A9E0CB182F5D568E033F0EB2D21246`

Bansel, J., & Stefani, S. (2024). *Synthetic data generator for financial contracts.* Goldman Sachs Developer Blog. Retrieved from `https://developer.gs.com/blog/posts/synthetic-data-generator` (Accessed on 21 March 2024)

Beduschi, A. (2024). Synthetic data protection: Towards a paradigm change in data regulation? *Big Data & Society*, *11*(1), 20539517241231277. doi: 10.1177/20539517241231277

Bernards, N., & Campbell-Verduyn, M. (2019). Understanding technological change in global finance through infrastructures. *Review of International Political Economy*, *26*(5), 773–789. Retrieved from `https://doi.org/10.1080/09692290.2019.1625420`

Birch, K., Cochrane, D. T., & Ward, C. (2021). Data as asset? the measurement, governance, and valuation of digital personal data by big tech. *Big Data & Society*, *8*(1), 20539517211017308.

Bokulich, A. (2020). Towards a taxonomy of the model-ladenness of data. *Philosophy of Science*, *87*(5), 793–806. doi: 10.1086/710516

Bowker, G. C. (2005). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Brown-Liburd, H., Issa, H., & Lombardi, D. (2015). Behavioral implications of big data's impact on audit judgment and decision making and future research directions. *Accounting Horizons*, *29*(2), 451–468.

Buehler, H., Gonon, L., Teichmann, J., & Wood, B. (2019). Deep hedging. *Quantitative Finance*, *19*(8), 1271–1291.

Buehler, H., Horvath, B., Lyons, T., Arribas, I. P., & Wood, B. (2020). *A data-driven market simulator for small data environments.* arXiv. doi: 10.48550/arXiv.2006.14498

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512.

Caliskan, K. (2020). Platform works as stack economization: Cryptocurrency markets and exchanges in perspective. *Sociologica*, *14*(3), Article 3. Retrieved from `https://doi.org/10.6092/issn.1971-8853/11746`

Coletta, A., Prata, M., Conti, M., & Balch, T. (2021). Towards realistic market simulations:

A generative adversarial network approach. In *Proceedings of the 2nd ACM international conference on AI in finance* (*ICAIF'21*). doi: 10.1145/3490354.3494411

Cornelli, G., Frost, J., Gambacorta, L., Rau, P. R., Wardrop, R., & Ziegler, T. (2023). Fintech and big tech credit: Drivers of the growth of digital lending. *Journal of Banking & Finance*, *148*, 106742.

De Vries, K. (2020). "you never fake alone. creative AI in action.". *Information, Communication & Society*, *23*(14), 2110–2127.

Di Girolamo, F., Hledik, J., & Pagano, A. (2024). *Synthetic data in the data hub of the digital finance platform* (Tech. Rep.). EU Science Hub. doi: 10.2760/83055

Edwards, P. N. (1999). Global climate science, uncertainty and politics: Data-laden models, model-filtered data. *Science as Culture*, *8*(4), 437–472. doi: 10.1080/09505439909526558

FCA Innovation Digital Sandbox. (2025). *Authorised push payment (APP) fraud dataset evaluation* (Tech. Rep.). Financial Conduct Authority. Retrieved from `https://www.fca.org.uk/publication/external-research/app-fraud-dataset-evaluation-report.pdf` (Accessed on 18 August 2025)

Ferrari, F., & McKelvey, F. (2023). Hyperproduction: A social theory of deep generative models. *Distinktion: Journal of Social Theory*, *24*(2), 338–360. doi: 10.1080/1600910X.2022.2137546

Financial Conduct Authority. (2021, April). *Supporting innovation in financial services: the digital sandbox pilot* (Tech. Rep.). UK Financial Conduct Authority.

Financial Conduct Authority. (2022, March). *Synthetic data to support financial services innovation* (Tech. Rep.). UK Financial Conduct Authority.

Financial Conduct Authority. (2023, February). *Synthetic data call for input: Feedback statement* (Tech. Rep.). UK Financial Conduct Authority.

Financial Conduct Authority. (2024, March). *Using synthetic data in financial services* (Tech. Rep.). UK Financial Conduct Authority.

Financial Reporting Council. (2023, December). *Audit market and competition developments: A snapshot* (Tech. Rep.). Financial Reporting Council.

Fitzgerald, A. (2024). Why synthetic data can never be ethical: A lesson from media ethics. *Surveillance & Society*, *22*(4), 477–482. doi: 10.24908/ss.v22i4.18324

Gal, M. S., & Lynskey, O. (2023). Synthetic data: Legal implications of the data-generation revolution. *Iowa Law Review*, *109*, 1087–.

Gambacorta, L., & Shreeti, V. (2025). *The AI supply chain* (Tech. Rep.). Bank for International Settlements. Retrieved from `https://www.bis.org/publ/bppdf/bispap154.htm`

Gensler, G., & Bailey, L. (2020). *Deep learning and financial stability.* SSRN Scholarly Paper. doi: 10.2139/ssrn.3723132

Gitelman, L. (Ed.). (2013). *"Raw Data" Is an Oxymoron*. Cambridge, MA: MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Hansen, K. B. (2020). The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data & Society*, *7*(1), 2053951720926558. doi: 10.1177/2053951720926558

Hansen, K. B. (2025). The stack inversion: On algo-centrism and the complex architecture of automated financial securities trading systems. *Science, Technology, & Human Values*, *50*(5),

932–961. doi: 10.1177/01622439241269983

Hansen, K. B., & Borch, C. (2022). Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance. *Big Data & Society*, *9*(1), 20539517211070701.

Hansen, K. B., & Spears, T. (2025). *Making data problems doable: The case of synthetic data in financial markets.* (Working Paper prepared for Special Issue in *The Information Society*)

Helm, P., Lipp, B., & Pujadas, R. (2024). Generating reality and silencing debate: Synthetic data as discursive device. *Big Data & Society*, *11*(2), 20539517241249447. doi: 10.1177/20539517241249447

Hoffmann, J., Bar-Sinai, Y., Lee, L. M., Andrejevic, J., Mishra, S., Rubinstein, S. M., & Rycroft, C. H. (2019). Machine learning in a data-limited regime: Augmenting experiments with synthetic data uncovers order in crumpled sheets. *Science Advances*, *5*(4), eaau6792. doi: 10.1126/sciadv.aau6792

Jäckel, P. (2002). *Monte carlo methods in finance*. John Wiley & Sons.

Jacobsen, B. N. (2023). Machine learning and the politics of synthetic data. *Big Data & Society*, *10*(1), 205395172211453. doi: 10.1177/20539517221145372

Jacobsen, B. N. (2024). The logic of the synthetic supplement in algorithmic societies. *Theory, Culture & Society*, *41*(4), 41–56. doi: 10.1177/02632764231225768

Jensen, C. B., & Morita, A. (2016). Introduction: Infrastructures as ontological experiments. *Ethnos*, *82*(4), 615–626. doi: 10.1080/00141844.2015.1107607

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., … Weller, A. (2022). *Synthetic data – what, why and how?* arXiv. Retrieved from `http://arxiv.org/abs/2205.03257`

Kellard, N., Millo, Y., Simon, J., & Engel, O. (2016). *Close communications: Hedge funds, brokers and the emergence of herding.* British Journal of Management. doi: 10.1111/1467-8551.12158

Kingma, D. P., & Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, *12*(4), 307–392.

Korenhof, P., Giesbers, E., & Sanderse, J. (2023). Contextualizing realism: An analysis of acts of seeing and recording in digital twin datafication. *Big Data & Society*, *10*(1), 20539517231155061. doi: 10.1177/20539517231155061

Kornberger, M., et al. (Eds.). (2019). *Thinking infrastructures* (Vol. 62). Emerald Publishing Limited.

KPMG. (2025). *AI in financial reporting and audit: Navigating the new era* (Tech. Rep.). Retrieved from `https://assets.kpmg.com/content/dam/kpmg/xx/pdf/2024/04/ai-in-financial-reporting-and-audit-web.pdf`

Limmer, Y., & Horvath, B. (2023). Robust hedging GANs. *Applied Mathematical Finance*. (arXiv) doi: 10.1080/1350486X.2024.2440661

Lopez de Prado, M. (2018). *Advances in financial machine learning*. Hoboken, NJ: Wiley.

Lopez de Prado, M. (2020). *Machine learning for asset managers*. Cambridge, MA: Cambridge University Press.

Lu, Y., Shen, M., Wang, H., & Wei, W. (2023). *Machine learning for synthetic data generation: A review.* arXiv. Retrieved from `http://arxiv.org/abs/2302.04062`

MacKenzie, D. (2004). The big, bad wolf and the rational market: portfolio insurance, the 1987 crash and the performativity of economics. *Economy and Society*, *33*(3), 303–334.

MacKenzie, D. (2011). The credit crisis as a problem in the sociology of knowledge. *American Journal of Sociology*, *116*(6), 1778–1841.

MacKenzie, D., Caliskan, K., & Rommerskirchen, C. (2023). The longest second: Header bidding and the material politics of online advertising. *Economy and Society*, *52*(3), 554–578. doi: 10.1080/03085147.2023.2238463

MacKenzie, D., & Millo, Y. (2004). Constructing a market, performing theory: The historical sociology of a financial derivatives exchange. *American Journal of Sociology*, *109*(1), 107–145.

Martin, A., & Newell, B. (2024). Synthetic data, synthetic media, and surveillance. *Surveillance & Society*, *22*(4), 448–452. doi: 10.24908/ss.v22i4.18334

McCosker, A. (2024). Making sense of deepfakes: Socializing AI and building data literacy on github and youtube. *New Media & Society*, *26*(5), 2786–2803. doi: 10.1177/14614448221093943

Meng, Y., Michalski, M., Huang, J., Zhang, Y., Abdelzaher, T., & Han, J. (2023). *Tuning language models as training data generators for augmentation-enhanced few-shot learning*. arXiv:2211.03044. Retrieved from `https://doi.org/10.48550/arXiv.2211.03044`

Millo, Y., Spence, C., & Valentine, J. (2023). The field of investment advice: The social forces that govern equity analysts. *The Accounting Review*, *98*(7), 457–477. doi: 10.2308/TAR-2021-0140

Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. doi: 10.1257/jep.31.2.87

Munkholm, J. L., & Weihn, T. (2025). Synthetic data: Serving privacy. In S. O. Søe, T. Wiehn, R. F. Jørgensen, & B. Valtýsson (Eds.), *Beyond privacy: People, practices, politics* (pp. 137–154). Policy Press.

Nadas, M., Diosan, L., & Tomescu, A. (2025). Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, *13*, 134615–134633. Retrieved from `https://doi.org/10.1109/ACCESS.2025.3589503`

Nikolenko, S. I. (2021). *Synthetic data for deep learning*. Cham: Springer Nature.

Offenhuber, D. (2024). Shapes and frictions of synthetic data. *Big Data & Society*, *11*(2), 20539517241249390. doi: 10.1177/20539517241249390

Paraná, E. (2025). AI as financial infrastructure? In C. Westermeier, M. Campbell-Verduyn, & B. Brandl (Eds.), *Cambridge global handbook on financial infrastructures* (pp. 386–400). Cambridge University Press.

Petry, J. (2021). From national marketplaces to global providers of financial infrastructures: Exchanges, infrastructures and structural power in global finance. *New Political Economy*, *26*(4), 574–597.

Pinzur, D. (2016). Making the grade: Infrastructural semiotics and derivative market outcomes on the chicago board of trade and new orleans cotton exchange, 1856–1909. *Economy and Society*, *45*(3–4), 431–453. Retrieved from `https://doi.org/10.1080/03085147.2016.1225360`

Potluru, V. K., Borrajo, D., Coletta, A., Dalmasso, N., El-Laham, Y., Fons, E., … Balch, T. (2024).

Synthetic data applications in finance. SSRN Scholarly Paper. doi: 10.2139/ssrn.3634235

Power, M. (2022). Theorizing the economy of traces: From audit society to surveillance capitalism. *Organization Theory*, *3*(3), 26317877211052296. doi: 10.1177/26317877211052296

Preda, A. (2008). Technology, agency, and financial price data. In T. Pinch & R. Swedberg (Eds.), *Living in a material world: Economic sociology meets science and technology studies.* Cambridge, MA: MIT Press.

PwC. (2025). *AI agents: Transforming the tax experience.* Retrieved from `https://www.pwc.com/us/en/services/tax/library/tax-ai-agents.html`

Ravn, L. (2024a). *The overlooked politics of synthetic data performance metrics.* Internet Policy Review. Retrieved from `https://policyreview.info/articles/news/politics-of-synthetic-data-performance-metrics/1761` (Retrieved May 16, 2024)

Ravn, L. (2024b). Synthetic training data and the reconfiguration of surveillant assemblages. *Surveillance & Society*, *22*(4), 460–465. doi: 10.24908/ss.v22i4.18319

Ravn, L. (2025). The fabrication of synthetic data promises: Tracing emerging arenas of expectations and boundary work. *Big Data & Society*, *12*(1), 20539517241307915. doi: 10.1177/20539517241307915

Ren, J., Du, Z., Wen, Z., Jia, Q., Dai, S., Wu, C., & Dong, Z. (2025). *Few-shot LLM synthetic data with distribution matching.* arXiv:2502.08661. Retrieved from `https://doi.org/10.48550/arXiv.2502.08661`

Ridgway, R., & Malevé, N. (2024). Synthetic data and reverse image search: Constructing new surveillant indexicalities. *Surveillance & Society*, *22*(4), 466–471. doi: 10.24908/ss.v22i4.18332

Romero, C. G., Lee, S. W., Lee, N. T., Smith, T., & Biaglioli, A. (2024). *Responsible AI in financial markets: Opportunities, risks & recommendations* (Tech. Rep.). Technology Advisory Committee under the CFTC. (2 May 2024)

Rossi, L., Harrison, K., & Shkloversuski, I. (2024). The problems of LLM-generated data in social science research. *Sociologica*, *18*(2), 145–168. doi: 10.6092/issn.1971-8853/19576

Ruggles, R., & Ruggles, N. D. (1975). The role of microdata in the national economic and social accounts. *Review of Income and Wealth*, *21*(2), 203–216.

Samiolo, R., Spence, C., & Toh, D. (2024). Auditor judgment in the fourth industrial revolution. *Contemporary Accounting Research*, *41*(1), 498–528.

Sisson, C. A. (1979). The synthetic micro data file: A new tool for economists. *Journal of Agricultural Economics Research*, *31*(3), 1–10.

Spears, T. (2019). Discounting collateral: Quants, derivatives and the reconstruction of the 'risk-free rate' after the financial crisis. *Economy and Society*, *48*(3), 342–370. Retrieved from `https://doi.org/10.1080/03085147.2018.1525153`

Spears, T., & Hansen, K. B. (2025). The use and promises of machine learning in financial markets: From mundane practices to complex automated systems. In C. Borch & J. P. Pardo-Guerra (Eds.), *Oxford handbook on the sociology of machine learning* (pp. 421–439). Oxford University Press. doi: 10.1093/oxfordhb/9780197653609.013.6

Staiger, A. (2025). *Learning from jp morgan's $175 million due diligence error.* Association of Certified Fraud Examiners. Retrieved from `https://www.acfe.com/acfe-insights-blog/blog`

-detail?s=jpmorgan-175-million-dollar-due-diligence-error-charlie-javice

Steinhoff, J. (2024). Toward a political economy of synthetic data: A data-intensive capitalism that is not a surveillance capitalism? *New Media & Society*, *26*(6), 3290–3306. doi: 10.1177/14614448221099217

Steinhoff, J., & Hind, S. (2025). Simulation and the reality gap: Moments in a prehistory of synthetic data. *Big Data & Society*, *12*(1), 20539517241309884. doi: 10.1177/20539517241309884

Straube, T. (2016). Stacked spaces: Mapping digital infrastructures. *Big Data & Society*, *3*(2), 2053951716642456. Retrieved from https://doi.org/10.1177/2053951716642456

Susser, D., & Seeman, J. (2024). Critical provocations for synthetic data. *Surveillance & Society*, *22*(4), 453–459. doi: 10.24908/ss.v22i4.18335

Svetlova, E. (2012). On the performative power of financial models. *Economy and Society*, *41*(3), 418–434.

Talwar, D., Guruswamy, S., Ravipati, N., & Eirinaki, M. (2020). Evaluating validity of synthetic data in perception tasks for autonomous vehicles. In *2020 IEEE international conference on artificial intelligence testing (AITest)*. doi: 10.1109/AITEST49225.2020.00018

The Economist. (2017). *The world's most valuable resource is no longer oil, but data.* The Economist. Retrieved 2025-03-25, from https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data (6 May 2017)

Wang, R., Liu, J., Zhao, W., Li, S., & Zhang, D. (2025). Auditbench: A benchmark for large language models in financial statement auditing. In *Ai for research and scalable, efficient systems.* doi: 10.1007/978-981-96-8912-5_3

Westermeier, C., Campbell-Verduyn, M., & Brandl, B. (Eds.). (2025). *Cambridge global handbook on financial infrastructures*. Cambridge University Press.

Whitney, C. D., & Norman, J. (2024). Real risks of fake data: Synthetic data, diversity-washing and consent circumvention. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency* (pp. 1733–1744). doi: 10.1145/3630106.3659002

Wiehn, T. (2024). Synthetic data: From data scarcity to data pollution. *Surveillance & Society*, *22*(4), 472–476.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., … Mann, G. (2023). *Bloomberggpt: A large language model for finance.* arXiv:2303.17564. Retrieved from https://doi.org/10.48550/arXiv.2303.17564