On the Challenges of Deploying Privacy-Preserving Synthetic Data in the Enterprise

Lauren Arthur [1] **Jason Costello** [1] **Jonathan Hardy** [1] **Will O'Brien** [1] **James Rea** [1] **Gareth Rees** [1] **Georgi Ganev** [1 2]

Abstract

Generative AI technologies are gaining unprecedented popularity, causing a mix of excitement and apprehension through their remarkable capabilities. In this paper, we study the challenges associated with deploying synthetic data, a subfield of Generative AI. Our focus centers on enterprise deployment, with an emphasis on privacy concerns caused by the vast amount of personal and highly sensitive data. We identify 40+ challenges and systematize them into five main groups – i) generation, ii) infrastructure & architecture, iii) governance, iv) compliance & regulation, and v) adoption. Additionally, we discuss a strategic and systematic approach that enterprises can employ to effectively address the challenges and achieve their goals by establishing trust in the implemented solutions.

1. Introduction

Recently, Generative AI has made significant advancements, with applications and capabilities spanning text, code, image, video, speech, and structured data (Sequoia Capital, 2022; Gartner, 2023). The acceptance of Generative AI products has reached unprecedented levels. For instance, OpenAI's ChatGPT attracted an estimated 100M active users in January alone (Reuters, 2023). This wide-spread adoption, however, has not spread to large organizations as they feel unprepared (KPMG, 2023) while facing a plethora of concerns, including exposing themselves to ethical, security, privacy, robustness, copyright, compliance and legal risks (Carlini et al., 2021; Weidinger et al., 2021; CNN, 2023; Entrepreneur, 2023; TechCrunch, 2023b). On the contrary, many companies have banned internal use of products like ChatGPT and GitHub Copilot (TechCrunch, 2023a).

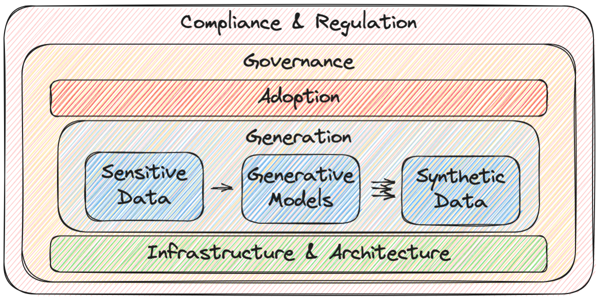

Figure 1: Main challenges of synthetic data deployment.

Motivation. As it is still unclear how/if said concerns can be resolved, we focus on synthetic data, a more established subfield of Generative AI. On the one hand, synthetic data is produced by similar generative models, e.g., GANs (Goodfellow et al., 2014), Transformers (Sohl-Dickstein et al., 2015), and Diffusion Models (Vaswani et al., 2017). On the other, it is typically trained on smaller-scale tabular datasets (vs. text/image), owned by a single entity (vs. public). This ownership provides enhanced control over the entire data generation process, contributing to increased trustworthiness. Moreover, synthetic data has attracted the attention of reputable organizations (Royal Society, 2023; UN, 2023; OECD, 2023) and regulators (ICO, 2022; FCA, 2023). However, these studies lack in-depth insights into the practical challenges and considerations that companies encounter when deploying synthetic data solutions.

In the digital age, enterprises have accumulated and safeguarded immense volumes of their users' sensitive and personal data, making it one of most highly valued asset in their possession (NYT, 2018; 2019). Unfortunately, many organizations impede data access, maintain data silos, and discourage data sharing, which undermines efforts to capitalize on the potential business and social benefits (Gartner, 2021). Privacy-preserving synthetic data presents a promising solution that holds immense potential in addressing these challenges in an ethical and trustworthy way.

Main Contributions. Our findings can be summarized:

1. We identify and explore 40+ challenges of deploying privacy-preserving synthetic data in large enterprises spanning beyond machine learning, unlike previous

[1] Hazy, London, UK [2] University College London, London, UK. Correspondence to: Georgi Ganev <georgi@hazy.com>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

studies (Assefa et al., 2020; Jordon et al., 2022).

2. We categorize the challenges into five groups: generation, infrastructure & architecture, governance, compliance & regulation, and adoption (displayed in Fig. 1).
3. We discuss the necessity of a structured approach that can help organizations frame the challenges and scale synthetic data deployment while establishing trust.

In Sec. 2, we discuss the main challenges and provide discussion in Sec. 3, while in App. A, we define some key concepts including synthetic data and Differential Privacy (DP).

2. Main Challenges

In this section, we discuss the challenges associated with deploying synthetic data in an enterprise context. We provide an overview of common considerations/obstacles and highlight those that are specifically relevant to enterprises.

2.1. Generation

We examine the challenges of training, generation, evaluation from a technical perspective. We discuss data/models operations and governance in Sec. 2.2 and 2.3.

Data Preparation. Assuming there is data access, the following challenges need to be taken into consideration:

- *Data quality*: ensuring high-quality and diverse train data includes collecting rich up-to-date data from various sources, validating/cleaning it, maintaining consistent pipelines, integrating legacy formats/systems.
- *Data preprocessing*: essential preprocessing steps include standardizing and normalizing formats, addressing missing values and outliers, handling custom entities and business rules, labeling and annotating data, and conducting feature engineering across databases.

Non-technology companies were lacking adequate investment in both areas just a few years ago (McKinsey, 2019).

Generating Process. Even though synthetic data could be presented as a panacea solution, in fact, there is no one-model-fits-all use cases (Tao et al., 2022) as different generative approaches/models are better suited for different use tasks and settings (Ganev et al., 2023). Important challenges in selecting the best candidate model(s) include:

- *Data domain*: each data domain (e.g., single table, time-series/sequential, multi-table) presents its own unique challenges. While a wide range of solutions are available for the former (Zhang et al., 2017; Xie et al., 2018; Jordon et al., 2018; McKenna et al., 2021; 2022), there are only a few privacy-preserving options for the latter two (Lin et al., 2020; Xu et al., 2023).
- *Use case*: the use case with its associated task (e.g., capturing statistics, preserving query answers, classification, etc.) and complexity as well as the chosen evaluation criteria could limit the choice.
- *Domain expertise*: different domains might require (enterprise) specific knowledge to be encoded into the model to achieve higher utility/scalability.
- *Robustness*: edge cases, outliers require extra attention.
- *Models comparison*: due to the uncertainty/variability of how a model would perform a priori, often it is necessary to train and compare different models.

Privacy. While applying DP provides formal mathematical privacy protections, it comes with its own challenges:

- *Threat model*: it is necessary to assess whether DP addresses the relevant threats – it successfully defends vs membership/reconstruction inference attacks (Hilprecht et al., 2019). However, DP does not protect company-specific values such as categorical ids/names.
- *Privacy unit/budget*: consistent definition of the privacy unit is required across the entire end-to-end pipeline. Setting the privacy budget is context-specific and challenging (Hsu et al., 2014). Also, various DP mechanisms allocate their budget differently.
- *Utility effect*: DP usually reduces utility (unfortunately, disproportionally affecting outliers and minorities (Stadler et al., 2022; Ganev et al., 2022)).
- *Implementation*: Coding DP mechanisms is complex; practitioners should rely on proven/public solutions.

Evaluation. Evaluation is also pivotal yet challenging area:

- *Quality*: trust requires measurability of the desired properties of synthetic data – utility, fidelity, diversity, authenticity, fairness, etc. Not only are they hard to measure, but defining them is also challenging (Alaa et al., 2022; van Breugel & van der Schaar, 2023).
- *Privacy and auditability*: auditing the model ensures the privacy of synthetic data by avoiding leakage of undesirable properties. (Houssiau et al., 2022a;b).

Jordon et al. (2022); Cummings et al. (2023) provide further discussion on the technical challenges of DP synthetic data.

2.2. Infrastructure & Architecture

Building on an enterprise platform entails requirements in system design and architecture, such as security, scalability, and distributed state management (Fowler, 2012).

Network Topologies. Deploying software in enterprise networks is challenging for various reasons:

- *Complexity*: unique organizational needs and structure drive networked system topology (Bano et al., 2016).
- *Legacy*: achieving agility/change (Leffingwell, 2007) in legacy systems is challenging for new technologies.
- *Cloud multiplicity*: "cloud transformation" (Jamshidi et al., 2013; Avram, 2014) brought various options to

consider: on-premises, hybrid, public, and multi-cloud.

Distributed Data. Data distribution across multiple at-rest locations poses significant barriers:

- *Segmentation*: organizational structure and data-centric regulations (PCI SSC, 2013) result in data spanning multiple security tiers and availability zones.
- *Discovery*: visibility of existent datasets for synthetic data use can become non-trivial at enterprise scale.
- *Security*: at sensitive levels, data may be limited to a small subset of enterprise services and operations staff for network ingress, typically in "break glass" scenarios (Brucker & Petritsch, 2009).

Data Access for Training. Model training necessitates data access, posing a significant security challenge to address. In addition to adhering to best industry best practices including *principle of least privilege* (Chen & Crampton, 2007) and *defense in depth*, data access poses further challenges:

- *Training at source*: training executed at or nearby the data within the same security tier.
- *Temporary privilege escalation*: short-lived delegation of control to ephemeral compute tasks.
- *Avoid data leakage*: ensuring all computation is done in-memory without spilling to disk.
- *Encryption*: all intermediate data encrypted at-rest.

Data/Model Security. There is an asymmetry in security provisions between data at-rest environments and compute environments used for synthetic data model training and generation. This presents an architectural challenge, and solutions may vary across different security contexts.

Data/Model Management. Synthetic data integration into enterprise data systems causes several challenges:

- *Versioning*: data versioning, synchronization, refreshment and metadata management (Nargesian et al., 2019) are all important for up-to-date synthetic data.
- *Compatibility*: model portability across environments and synthetic data compatibility with various locations/formats both need to be considered.
- *Compute resources*: high resource requirements for model training and data storage necessitate load distribution across multiple machines and efficient management of compute-intensive resource allocations.

2.3. Governance

As enterprises deploy synthetic data, careful governance of the lifeycle processes and efficiency including analysis, policies, communication, and scaling become crucial.

AI Governance. AI governance is still in early stages, which poses challenges for enterprises using synthetic data models to develop governance frameworks:

- *Cross-team collaboration*: creating a unified governance framework is challenging when involving multiple stakeholders, including i) model builders (data scientists/engineers), ii) data owners/controllers, domain experts, and iii) governing stakeholders (legal/privacy experts) (Butcher & Beridze, 2019; PwC, 2023).
- *Limited monitoring*: the lack of industry standards for monitoring synthetic data models makes tool building and/or selection difficult.

Scalability. Enterprise segmentation and diverse legislation across countries present challenges due to limited knowledge exchange and differing regulatory requirements:

- *Access control*: complex team structures hinder i) tracking ownership of data used by synthetic data models, ii) implementing precise access control, and iii) preventing data exploitation (e.g., proper disposal of data and the copy problem (Trask et al., 2020)).
- *Security*: models should be tested for potential vulnerabilities such as data poisoning, query injection attacks.
- *Policies*: data governance frameworks, privacy policies, and compliance procedures are resource-intensive to create and disseminate.
- *Environment controls*: clear internal guidance on the operational scope of synthetic data models is crucial for enterprises with multiple environments (e.g., staging vs production) of varying data sensitivity levels.

Ethics. Aligning on the ethics of synthetic data at various organizational levels can be challenging and time-consuming, considering privacy, explainability, fairness, agency, and sourcing considerations at scale (UKSA, 2022).

- *Explainability and interpretability*: ensuring synthetic data explainability is challenging due to ongoing internal governance, knowledge, and research gaps, yet it plays a crucial role in instilling trust (Ohm, 2009).
- *Transparency*: obtaining explicit user consent and publicizing synthetic data policies builds trust but opens up risk of scrutiny and reputational damage.

2.4. Compliance & Regulation

Large corporations are facing growing difficulties in adhering to regulations governing data protection and AI laws.

Landscape. On the one hand, over 120 countries have adopted data protection and AI legislation (EDPS, 2022) (e.g., GDPR, HIPAA, AI Act) while regulators are issuing companies more/higher-value fines (CMS, 2023). On the other, enterprises are caught between laws balancing innovation, competition, ethical considerations, and regulatory compliance. Gal & Lynskey (2023) warn that synthetic data could exacerbate these challenges.

Data Protection. Understanding and adhering to regulation is time/resource intensive. Main considerations include:

- *Privacy engineering*: integrating privacy protections across the design and development cycles include privacy by design, data minimization, informed consent, accountability, transparency, purpose limitation.
- *Anonymization*: A29WP (2014); ICO UK (2021) assert that *singling out*, *linkability*, and *inferences* risks need to be reduced for sufficient anonymization under the lenses of *motivated intruder* test. Yet there is not enough clarity or concrete guidelines (FCA, 2023).

Tech-Legal Gap. Mapping technology to legal concepts, and vice versa, is a complex process requiring careful consideration and expertise. Bellovin et al. (2019); López & Elbi (2022); Ganev (2023) argue that applying DP to synthetic data techniques can address privacy concerns, confirmed by empirically evaluations (Giomi et al., 2022).

2.5. Adoption

In recent years, the number of AI capabilities used by organizations has more than doubled (McKinsey, 2022). In this section, we highlight the key challenges when driving synthetic data adoption at scale.

Strategic Alignment & Outcomes. Unspecified metrics and ROI measures, along with a lack of strategic alignment, can hinder achieving business outcomes:

- *Vision & strategy*: without a clear synthetic data vision, aligned with the broader business strategy, behaviors and target outcomes are unlikely to materialize. Alignment across the business on objectives will drive more measurable value (Accenture, 2021).
- *Business metrics & ROI*: misalignment on business metrics and ROI measures can jeopardize decision-making, investment, and performance tracking.

Operational Effectiveness. Operational challenges include workflow adaptation, team training, and cost management:

- *Existing workflows*: New technologies will create fundamental changes in workflows, roles, and culture (HBR, 2019). Optimizing producer/consumer workflows is crucial for efficient synthetic data generation, utilization, and feedback loops.
- *Capability ownership*: a lack of coordination managed by a central team can slow adoption due to conflicting priorities and communication issues.
- *Skills gap & training*: Organizations are struggling to find employees with the combination of skills and knowledge to unleash the full potential of AI (MIT Professional, 2018). Thus, adequate upskilling of synthetic data users is essential to address knowledge disparities and aid adoption.
- *Cost management*: implementing synthetic data at scale incurs costs (e.g., monetary investments, resource allocation, infrastructure requirements) that must be carefully managed before benefits are realized.

Organizational Change. User skepticism, stakeholder buy-in, and effective communication impact change:

- *User scepticism*: growing scrutiny and concerns over Generative AI could hinder synthetic data adoption.
- *Change management*: successful transformation programs shift from a sole focus on technology to account for the human experience (EY, 2021). As synthetic data becomes integrated across the enterprise, failure to guide organizational change and adapt the culture can impede adoption.
- *Leadership & influencer buy-in*: overlooking leadership buy-in and cultural influencers can reduce confidence and trust in a change program (HBR, 2023).

3. Discussion

Given the wide-ranging applications and benefits of synthetic data, there are technical, architectural, governance, and adoption challenges related to deploying it within an enterprise. However, these challenges are not obstructions, nor do they need to be addressed simultaneously.

State of Play. Unlike nascent Generative AI technologies, synthetic data is delivering on its initial promises in commercial environments. The focus now turns to how it can be adopted at scale as a core technology in an enterprise data strategy. Early adopters have invested in proofs of concept (POCs) that have demonstrated tangible business value, such as increased efficiency, accelerated innovation, and reduced compliance risk (Benedetto et al., 2018; Nature, 2023) but deployment and adoption should be approached holistically and prioritised based on strategic objectives.

Structured Approach. Organizations can adopt a structured approach to anchor their synthetic data programs, which includes assessing external dependencies, guiding activities related to deployment, governance, and adoption. By involving diverse stakeholders and the appropriate expertise, focusing initially on low-risk use cases with quick time to value, and creating awareness about synthetic data, organizations can gain buy-in and establish trust.

Finally, in App. B, we outline a simplified three-staged process which prioritizes focus areas to address.

4. Conclusion

We have identified and categorized numerous challenges associated with large-scale deployment of synthetic data. We believe our work will be valuable for practitioners and professionals interested in adopting synthetic data solutions.

References

A29WP. Opinion on anonymisation techniques. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf, 2014.

Accenture. Aligning data strategy and digital transformation. https://www.accenture.com/hk-en/insights/strategy/aligning-data-strategy-digital-transformation, 2021.

Alaa, A., Van Breugel, B., Saveliev, E. S., and van der Schaar, M. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *ICML*, 2022.

Assefa, S. A., Dervovic, D., Mahfouz, M., Tillman, R. E., Reddy, P., and Veloso, M. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *ACM ICAIF*, 2020.

Avram, M. G. Advantages and challenges of adopting cloud computing from an enterprise perspective. *Procedia Technology*, 2014.

Bano, M., Zowghi, D., and Sarkissian, N. Empirical study of communication structures and barriers in geographically distributed teams. *IET software*, 2016.

Bellovin, S. M., Dutta, P. K., and Reitinger, N. Privacy and synthetic datasets. *STLR*, 2019.

Benedetto, G., Stanley, J. C., Totty, E., et al. The creation and use of the SIPP synthetic Beta v7. 0. *US Census Bureau*, 2018.

Brucker, A. D. and Petritsch, H. Extending access control models with break-glass. In *ACM SACMAT*, 2009.

Butcher, J. and Beridze, I. What is the State of Artificial Intelligence Governance Globally? *The RUSI Journal*, 2019.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., Oprea, A., and Raffel, C. Extracting training data from large language models. In *USENIX Security*, 2021.

Chen, L. and Crampton, J. Inter-domain role mapping and least privilege. In *ACM SACMAT*, 2007.

CMS. GDPR Enforcement Tracker. https://www.enforcementtracker.com/, 2023.

CNN. Don't tell anything to a chatbot you want to keep private. https://edition.cnn.com/2023/04/06/tech/chatgpt-ai-privacy-concerns/index.html, 2023.

Cummings, R., Desfontaines, D., Evans, D., Geambasu, R., Jagielski, M., Huang, Y., Kairouz, P., Kamath, G., Oh, S., Ohrimenko, O., Papernot, N., Rogers, R., Shen, M., Song, S., Su, W., Terzis, A., Thakurta, A., Vassilvitskii, S., Wang, Y.-X., Xiong, L., Yekhanin, S., Yu, D., Zhan, H., and Zhang, W. Challenges towards the Next Frontier in Privacy. *arXiv:2304.06929*, 2023.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 2014.

EDPS. Data Protection concerns all of us. https://edps.europa.eu/press-publications/press-news/blog/data-protection-concerns-all-us_en, 2022.

Entrepreneur. History Has Shown What Happens to Companies that Shy Away from New Tech, So Why Are So Many Afraid of Generative AI? https://www.entrepreneur.com/leadership/why-are-so-many-companies-afraid-of-generative-ai/446198, 2023.

EP and Council. Article 4 GDPR Definitions. https://gdpr-info.eu/art-4-gdpr/, 2016.

EY. Why change management is crucial during technologic transformations. https://www.ey.com/en_be/consulting/why-change-management-is-crucial-during-technologic-transformations, 2021.

FCA. Synthetic data call for input feedback statement. https://www.fca.org.uk/publication/feedback/fs23-1.pdf, 2023.

Fowler, M. *Patterns of Enterprise Application Architecture*. Pearson Education, 2012.

Gal, M. and Lynskey, O. Synthetic Data: Legal Implications of the Data-Generation Revolution. *109 Iowa Law Review*, 2023.

Ganev, G. When synthetic data met regulation. In *ICML Workshop on Generative AI and Law*, 2023.

Ganev, G., Oprisanu, B., and De Cristofaro, E. Robin Hood and Matthew Effects: Differential privacy has disparate impact on synthetic data. In *ICML*, 2022.

Ganev, G., Xu, K., and De Cristofaro, E. Understanding how Differentially Private Generative Models Spend their Privacy Budget. *arXiv:2305.10994*, 2023.

Gartner. Data Sharing Is a Business Necessity to Accelerate Digital Business. https://www.gartner.com/smarterwithgartner/data-sharing-is-a-business-necessity-to-accelerate-digital-business, 2021.

Gartner. Beyond ChatGPT: The Future of Generative AI for Enterprises. https://www.gartner.com/en/articles/beyond-chatgpt-the-future-of-generative-ai-for-enterprises, 2023.

Giomi, M., Boenisch, F., Wehmeyer, C., and Tasnádi, B. A unified framework for quantifying privacy risk in synthetic data. In *PETs*, 2022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *NIPS*, 2014.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. Logan: membership inference attacks against generative models. In *PoPETs*, 2019.

HBR. Building the AI-Powered Organization. https://hbr.org/2019/07/building-the-ai-powered-organization, 2019.

HBR. Getting Employee Buy-In for Organizational Change. https://hbr.org/2023/02/getting-employee-buy-in-for-organizational-change, 2023.

Hilprecht, B., Härterich, M., and Bernau, D. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. In *PoPETs*, 2019.

Houssiau, F., Cohen, S. N., Szpruch, L., Daniel, O., Lawrence, M. G., Mitra, R., Wilde, H., and Mole, C. A Framework for Auditable Synthetic Data Generation. *arXiv:2211.11540*, 2022a.

Houssiau, F., Jordon, J., Cohen, S. N., Daniel, O., Elliott, A., Geddes, J., Mole, C., Rangel-Smith, C., and Szpruch, L. TAPAS: a toolbox for adversarial privacy auditing of synthetic data. In *NeurIPS Workshop on SyntheticData4ML*, 2022b.

Hsu, J., Gaboardi, M., Haeberlen, A., Khanna, S., Narayan, A., Pierce, B. C., and Roth, A. Differential privacy: an economic method for choosing epsilon. In *IEEE CSF*, 2014.

ICO. Chapter 5: privacy-enhancing technologies (PETs). https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf, 2022.

ICO UK. Chapter 2: how do we ensure anonymisation is effective? https://ico.org.uk/media/about-the-ico/documents/4018606/chapter-2-anonymisation-draft.pdf, 2021.

Jamshidi, P., Ahmad, A., and Pahl, C. Cloud migration research: a systematic review. *IEEE TCC*, 2013.

Jordon, J., Yoon, J., and Van Der Schaar, M. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *ICLR*, 2018.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. Synthetic Data–what, why and how? *arXiv:2205.03257*, 2022.

KPMG. KPMG Generative AI Survey. https://info.kpmg.us/news-perspectives/technology-innovation/kpmg-generative-ai-2023.html, 2023.

Leffingwell, D. *Scaling software agility: best practices for large enterprises*. Pearson Education, 2007.

Lin, Z., Jain, A., Wang, C., Fanti, G., and Sekar, V. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *ACM IMC*, 2020.

López, C. A. F. and Elbi, A. On the legal nature of synthetic data. In *NeurIPS SyntheticData4ML*, 2022.

McKenna, R., Miklau, G., and Sheldon, D. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *JPC*, 2021.

McKenna, R., Mullins, B., Sheldon, D., and Miklau, G. Aim: An adaptive and iterative mechanism for differentially private synthetic data. *PVLDB*, 2022.

McKinsey. Driving impact at scale from automation and AI. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Driving%20impact%20at%20scale%20from%20automation%20and%20AI/Driving-impact-at-scale-from-automation-and-AI.ashx, 2019.

McKinsey. The state of AI in 2022—and a half decade in review. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review#review, 2022.

MIT Professional. Four key barriers to the widespread adoption of AI. https://professional.mit.edu/news/articles/four-key-barriers-widespread-adoption-ai, 2018.

Nargesian, F., Zhu, E., Miller, R. J., Pu, K. Q., and Arocena, P. C. Data Lake Management: Challenges and Opportunities. *PVLDB*, 2019.

Nature. Synthetic data could be better than real data. https://www.nature.com/articles/d41586-023-01445-8, 2023.

NYT. As Facebook Raised a Privacy Wall, It Carved an Opening for Tech Giants. https://www.nytimes.com/2018/12/18/technology/facebook-privacy.html, 2018.

NYT. Twelve Million Phones, One Dataset, Zero Privacy. https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html, 2019.

OECD. Emerging privacy-enhancing technologies. https://www.oecd-ilibrary.org/content/paper/bf121be4-en, 2023.

Ohm, P. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 2009.

PCI SSC. Information Supplement: PCI DSS Cloud Computing Guidelines. https://listings.pcisecuritystandards.org/pdfs/PCI_DSS_v2_Cloud_Guidelines.pdf, 2013.

PwC. Managing the risk of generative AI. https://explore.pwc.com/generativeai?_pfses=D8nsC9bP5NQMW25zxpYx69tC, 2023.

Reuters. ChatGPT sets record for fastest-growing user base. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/, 2023.

Royal Society. From privacy to partnership: the role of PETs in data governance and collaborative analysis. https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/From-Privacy-to-Partnership.pdf, 2023.

Sequoia Capital. Generative AI: A Creative New World. https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Stadler, T., Oprisanu, B., and Troncoso, C. Synthetic data – anonymization groundhog day. In *Usenix Security*, 2022.

Tao, Y., McKenna, R., Hay, M., Machanavajjhala, A., and Miklau, G. Benchmarking differentially private synthetic data generation algorithms. *PPAI*, 2022.

TechCrunch. Apple reportedly limits internal use of AI-powered tools like ChatGPT and GitHub Copilot. https://techcrunch.com/2023/05/19/apple-reportedly-limits-internal-use-of-ai-powered-tools-like-chatgpt-and-github-copilot/, 2023a.

TechCrunch. The current legal cases against generative AI are just the beginning. https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning/, 2023b.

Trask, A., Bluemke, E., Garfinkel, B., Cuervas-Mons, C. G., and Dafoe, A. Beyond privacy trade-offs with structured transparency. *arXiv:2012.08347*, 2020.

UKSA. Ethical considerations relating to the creation and use of synthetic data. https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/, 2022.

UN. The United Nations Guide on privacy-enhancing technologies for official statistics. https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf, 2023.

van Breugel, B. and van der Schaar, M. Beyond Privacy: Navigating the Opportunities and Challenges of Synthetic Data. *arXiv:2304.03722*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *NeurIPS*, 2017.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., and Gabriel, I. Ethical and social risks of harm from language models. *arXiv:2112.04359*, 2021.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. Differentially private generative adversarial network. *arXiv:1802.06739*, 2018.

Xu, K., Ganev, G., Joubert, E., Davison, R., Van Acker, O., and Robinson, L. Synthetic data generation of many-to-many datasets via random graph generation. In *ICLR*, 2023.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. Privbayes: Private data release via bayesian networks. *ACM TODS*, 2017.

A. Preliminaries

We introduce the foundational concepts that we use throughout the paper.

Enterprise. For the purposes of this paper, large enterprises (typically with over $1B revenue and 100,000 customers) are established organizations with diverse business units and extensive legacy infrastructure. They possess vast and diverse data, which is often distributed and inconsistent. While they face burdensome regulations and the risk of reputational damage, their cautious approach to innovation is in contrast to the agility of startups and SMEs.

Personal and Proprietary Data. EP and Council (2016) define personal data as "any information relating to an identified or identifiable natural person" while the latter as someone who can directly or indirectly be identified, by reference to an identifier such as name, id number, location, etc. On the other hand, proprietary data refers to privileged/confidential information owned by an organization that provides a competitive advantage (e.g., trade secrets, commercial/financial information) and is not public.

(Privacy-Preserving) Synthetic Data. Synthetic data is data generated by a purpose-built mathematical model or algorithm (i.e., machine learning generative model) trained on real data, with the aim of solving a (set of) task(s) (Jordon et al., 2022). Unfortunately, unless models are trained with explicit privacy-preserving mechanisms, they could memorize and leak the privacy of input records (Hayes et al., 2019; Carlini et al., 2019; Stadler et al., 2022). The state-of-the-art method to protect against such risks is to train models while satisfying Differential Privacy (DP) (Dwork & Roth, 2014), whereby noisy/random mechanisms provably minimize the exposure of all records. The level of exposure is quantified by an input parameter ϵ, also known as the privacy budget.

We focus on privacy-preserving synthetic data, generated by DP models, even though synthetic data use cases could go beyond privacy (e.g., data augmentation, fairness, simulation, etc. (Jordon et al., 2022; van Breugel & van der Schaar, 2023)). However, privacy remains a central focus in all applications, given the need to handle sensitive data, which poses potential risks of financial loss, reputational damage, and loss of customers trust if mishandled.

B. Deployment Phases

As discussed in Sec. 2, there are several challenges when deploying high quality enterprise synthetic data. In this section, we outline a simplified three-stage approach and highlight the core areas to consider.

B.1. Initial Phase

Organizations embarking on their synthetic data journey can gain valuable insights from case studies and expert conversations with synthetic data professionals, including technology vendors, regulatory bodies, and system integrators.

- *Proofs of concept*: select initial use cases that will deliver strong proof points and quick time to value to prove the technology. Ensure the results and business outcomes are communicated with the relevant stakeholders to build awareness and advocacy, including how the synthetic data performs compared to previously used techniques (e.g., anonymization/masking).
- *Education*: as synthetic data is considered an emerging technology, education and foundational knowledge sharing is necessary for technical and non-technical audiences. An additional aim should be to increase visibility across the enterprise to spark further use cases.
- *Governance and metrics*: agree a core set of metrics to compare and showcase the results of synthetic data. Begin to consider governance frameworks and performance monitoring required for transitioning into the scaling phase.

B.2. Scaling Phase

Many synthetic data early adopters are now in the Scaling Phase where the focus is on expansion. This stage can be challenging for enterprises to navigate: delivery capacity has to grow to meet demand at the same time as the organisational change program and governance framework evolves to enable further adoption.

- *Strategic alignment*: as with any large-scale AI deployment, aligning the long-term business value of synthetic data with the broader strategy is crucial for setting direction and driving the right behaviours. The earlier key stakeholders are involved in this alignment exercise, the greater the chances of success in the later stages.
- *Regulation & data tracking*: as adoption grows, organizations need to keep abreast of regulatory changes and internal governance should be updated accordingly – covering both synthetic data models and synthetic data usage. This should be incorporated as part of the governance framework developed in the initial phase.
- *Scaling architecture*: to counter the complexity of data ownership and network topologies as noted in Sec. 2.2, data flows should be mapped as much as possible before scaling synthetic data.
- *Monitoring & auditing*: monitoring the models with logging, dashboards, and audit trails provide greater visibility across the enterprise and may be necessary for compliance.

B.3. Future Phase

The focus areas in the Future Phase will be largely driven by the progress of synthetic data technologies, the evolving regulatory landscape and rate of adoption within enterprises. Synthetic data will replace the use of real data across business areas as the production and consumption process becomes more streamlined.

- *Seamless & integrated use*: as synthetic data is adopted across the enterprise, it should be embedded in the culture as a core component of the data strategy. There may be scope for more seamless ways to attain and consume synthetic data, for example, on-demand data and models via a marketplace.
- *Synthetic data advisory*: organizations will have established strong governance frameworks to reach this stage. They may wish to utilize knowledge and experience to advise external bodies and contribute to industry best practices.