Out-of-the-Box Article

BIR

Business Information Review
2024, Vol. 0(0) 1–5
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02663821241231101
journals.sagepub.com/home/bir
Sage

Synthetic data

Preeti Patel
London Metropolitan University, UK

Abstract

The rise of data-driven businesses poses a number of significant challenges for contemporary organisations. These include legal and ethical considerations arising from the use of personal data, the growing challenges of information security, and the difficulty managing the volume of data generated in business transactions of different kinds. The exponential growth of data continues unabated with global data volumes reaching 181 zettabytes by 2025, and with 90% of the world's data generated in the last two years alone. This massive growth can be mainly attributed to data gathered by Internet of Things (IoT) and related sensory devices, in addition to data generated through the human use of digital tools and applications. Given this abundance of real-world data, in what context could synthetic data be necessary? This paper highlights the growing organisational use of synthetic data and explores where and how it can be optimally used. It examines the ethical aspects of synthetic data usage, the need to garner public perception and acceptance, and the key aspects of traceability, accountability and risk mitigation.

Keywords

Synthetic data, privacy-enhancing technology, data sharing, ethics of synthetic data

Synthetic data

Information and data by its very nature is becoming ubiquitous and pervasive (Tredinnick, 2022), and this drives the ways in which businesses are adapting to the datascape. The information profession has long been at the forefront of the uses of data in the workplace, and in understanding the ways in which data can sometimes be misused (Tredinnick, 2023). While acknowledging the strategic value of data - sometimes referred to as 'the new gold' (van Laarhoven-Smits, 2023) or 'the new oil' (Jonk et al., 2023) and also known as 'the new water' - is of prime concern to businesses, so too is the mitigation of risks that accompany the use of data in commercial contexts. An organisation's Chief Data Officer (CDO) will prioritise in a data strategy the key aspects of data collection, analytics, and mobility for effective decision-making as well as the requirements of governance, compliance, and security. Given the heavy reliance on real-world data, in what ways can synthetic data be useful to businesses?

The uses of synthetic data

Synthetic data describes artificially generated data that mimics the structure, properties, and characteristics of data generated from the real-world. Instead of being collected from actual observations or events, synthetic data is created through mathematical models, algorithms, or other simulation methods. The purpose of synthetic data is to replicate the statistical properties, structure, and patterns found in real data while protecting the privacy of individuals and sensitive information. This type of data is often used in various fields, including machine learning, data analysis, and software testing, where access to real data may be restricted due to privacy concerns or data scarcity. Synthetic data allows researchers, developers, and data scientists to simulate scenarios, test algorithms, and train models without compromising privacy or relying solely on limited real-world datasets.

From the 1970s onwards, digitization paved the way for software synthesisers, but the idea of fully synthetic data was not proposed until 1993 by Rubin. The early use of synthetic data was in the realm of computer vision where artificial drawings and 3D imagery were required to train the vision systems. In parallel there has been a growing need over the decades for privacy protection. The natural evolution to the current day requirements and uses for synthetic data are myriad. Indeed, with the accessibility of big data, it

Corresponding author:
Preeti Patel, Computer Science and Applied Computing, London Metropolitan University, 166-220 Holloway Road, London N7 8DB, UK.
Email: p.patel@londonmet.ac.uk

is possible for AI systems to generate fake faces for non-existent individuals, craft new artworks, or create virtual but realistic landscapes – none of which exist in the real world.

Despite sometimes being labelled with derogatory terms such as 'fake', 'artificial', 'dummy' or 'simulated', the value of synthetic data as an alternative to sensitive real-world data arises from the advantages it confers in some contexts. It can be used to conceal individuals' identities and offering privacy against re-identification. It can be used to generate at scale data at a fraction of the cost of standard data acquisition. It can facilitate mechanisms to improve fairness, bias and robustness of the AI models that generate data. With these potential benefits in mind, synthetic data is rapidly emerging as an essential tool for organizations, primarily aimed at safeguarding the privacy of data subjects (including customers and employees). This leads to a more secure, convenient, and expedited means of sharing data that might otherwise be considered sensitive, such as protected characteristics or other identifiable personal information.

In this context, synthetic data is an important privacy-enhancing technology (PET), offering privacy-preserving alternatives to real world sensitive information. In general, privacy-enhancing technologies (PETs) allow for the collection, processing, analysis and sharing of information while protecting the privacy of sensitive data. In data-driven contexts PETs offer a promising way for organisations and governments to maintain data privacy whilst unlocking the economic benefits of the untapped potential of data. Lundy-Bryan (2021) states that by 2030 PET-enabled data marketplaces, in which governments, corporates, individuals, and machines trade data freely and securely will be second only to the Cloud market. PETs have the potential to democratise data access, create borderless computing infrastructure, support digital free trading, and enable civic technology.

Data supply chains have been vastly extended and aid in-house insight and predictive capability, making a digital ecosystem that allows multiple parties to collaborate on the same data, whilst keeping within regulatory requirements. Lundy-Bryan (2021) predicts that by 2030 collaborative computing will be the largest new market, encompassing both internal and external data collaboration; and has coined the term partnership-enhancing technology to distil the concepts of collaborative computing which will give rise to entirely new data collaboration applications uncovering a new avenue of data-based business models.

There are several reasons why organisations might consider turning to synthetic data. In larger organisations, data silos and legacy systems are often the root cause of data unavailability; where there is simply not enough data, or it is difficult to combine there lies a potential to synthesise it. Furthermore, in today's highly regulatory landscape, organisations require to be legally compliant when it comes to data protection, and this may limit the use of the original data. Another challenge is that data flows, data sharing and data access within an organisation may be inhibited due to security concerns, for example some information may be too sensitive to be migrated to a cloud infrastructure. In addition, for many organisations the financial costs associated with acquiring data from third parties or preparing and accessing it in-house may be prohibitive. Syntheticus advocate the democratisation of data through synthetic data, (Syntheticus 2023) and foresee the levelling of the playing field with smaller businesses being able to compete with larger businesses, such as Facebook, Google and Amazon who have in the past excelled at collecting and leveraging huge amounts of data.

Synthetic data can also encompass text, image, video, audio, and its applications are broad and span many businesses. For example, Amazon's Sagemaker Ground Truth (Barth, 2022) can be used to generate synthetic images and associated labels and annotations at massive scale for use in computer vision training applications. Amazon also used synthetically developed text-to-speech data to train Alexa to speak with an Irish accent (Tinchez and Czarnowska, 2023) allowing for a multi-speaker, multi-accent platform. In another example, Google's self-driving car project, Waymo, uses synthetic data to aid analysis and fine-tuning of sensors used in autonomous vehicles. In the absence of data portraying real-world scenarios, which Waymo calls long-tail cases, such as pedestrians dressed in costumes or a motorcycle travelling at speed without its rider, synthetic data generation plays a vital role (Ambrosio, 2022). Additionally, in a bid to circumnavigate complaints of IP abuse, privacy and data access, companies such as OpenAI and Microsoft are using synthetic data to train their ever-hungry generative AI models. Government departments such as the Ministry of Defence (Walker, 2020), whose datasets are particularly sensitive are becoming reliant on synthetic data as a way to cooperate with the data science community.

The Financial Conduct Authority (FCA) recognises the increasing level of synthetic data exploration within financial services. The FCA (FCA 2022) contends that data sharing is controlled within the sector and is only conducted in accordance with strict data privacy laws, which third-party providers such as FinTechs, RegTechs and also start-ups may find challenging to access, as a result of the complex due diligence requirements. However, financial data, such as consumer transaction records, account payments, or trading data, is sensitive personal data that is subject to data protection obligations, as well as often being commercially sensitive. Interesting examples from the finance world, where bank transactions are highly confidential and where there are no real public datasets, include using synthetic data to investigate anti-money laundering and atypical fraud behaviours. Many financial institutions, such as, American Express and J. P. Morgan, have declared their ability to generate statistically accurate synthetic data from financial transactions to perform fraud detection (ACM, 2021).

Generating synthetic data

Let us now turn to how synthetic data is generated. There are many techniques to obscure private and sensitive information, including redaction, masking/replacing, coarsening/de-precision, mimicking, and simulation. Data can be synthesised either manually with tools such as Excel or automatically with computer algorithms. So synthetic data is generated in a purely digital environment, whilst real-world data is created every time a person uses a laptop, smartphone, computer, smartwatch, or Web site.

Machine learning methods can be used to generate synthetic data, but it is the deep learning models that use Generative Adversarial Networks (GAN) that are able to create statistically identical data through a continual algorithmic cycle of generating unreal data and learning how to discriminate the unrealness. Whilst GAN technology has been notoriously used to create deepfake videos of, for example, Presidents giving speeches that they never actually did, it can also be used to fight financial fraud. Of concern is the view that by 2030, synthetic data use will outweigh real data in AI models, (Linden, 2022).

Nevertheless, the generation of synthetic data raises concerns of its own. The Office for National Statistics (ONS) places importance on the need to consider the ethics of generating and sharing synthetic data, in particular the notion that the original data's outliers may not be replicated which could lead to inaccuracies and distortion (UKSA 2022). The ONS Synthetic Data Spectrum (Figure 1) is a classifying scale used to determine how closely synthetic datasets resemble the original data (fidelity) and the perceived disclosure risk.

The majority of business organisations will use low fidelity structural synthetic data due to its low risk, as the data reflects the original data much less closely. In comparison high fidelity synthetic data mirrors closely the complex relationships between different variables, and as a result poses a greater risk of disclosure. In some situations, it may be possible to purposefully introduce noise to datasets to minimise the potential of disclosure. Typically, it is the low fidelity structural synthetic data that is provided by commercial synthetic data providers, As an aside, these providers market their products as being 'risk-free', for example, the London-based company Hazy states that "with no real data there is no risk", similarly the tagline for a Swedish company Syndata AB is "solving real data problems with no real risk", and lastly a start-up called Synthesized markets their product to give "10x the performance, 0 risks".

When considering the ethics of synthetic data, healthcare and pharma companies need to be particularly mindful. Generating synthetic data at scale seems an attractive proposition with the known benefit of preserving patient confidentiality whilst having vast volumes of data for fast-tracked clinic trailing. The Swiss pharma Roche have successfully generated large scale synthetic medical data to anonymise data and protect patient privacy, whilst complying with GDPR and HIPAA requirements Statice (2022).

The use of synthetic data in healthcare research, education, health IT and public health is becoming more established as the particular need to link (and subsequently share) data originating from multiple sources and in multiple formats becomes important (Gonzales et al., 2023). The processes for gaining data access, particularly within the healthcare sector, where the need for data use agreements, submission and approval of full protocol, completion of data request forms, ethics review approvals and the costs associated with non-public domain data make it virtually impossible to conduct clinical research and improvements in patient care. Hence the use of synthetic data to create an electronic health record (HER) dataset with patient identifiable information and other sensitive information replaced with artificial data in order to avoid reidentification, or conversely a synthetic dataset could be made of HER records where all the original data is synthesised to produce a completely fictitious record (Giuffre and Shung, 2023).

The digital twin concept of integrating real world data with virtual representations of physical assets is important for space exploration, manufacturing, construction, automotive and utilities. However, the use of digital twins in healthcare is

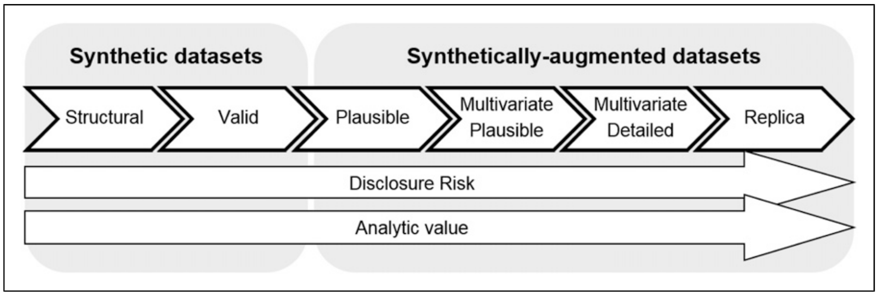

Figure 1. ONS methodology working paper series number 16 – Synthetic Data Pilot. Source: Office for National Statistics.

currently underdeveloped, where synthetic data could be used to create realistic models to simulate different scenarios such as patient volume fluctuations, equipment availability and staff training levels. Synthetic data could also be used to create digital twins of patients to optimise treatment plans and improve patient outcomes. Synthetic data can improve healthcare, but measures to protect patient well-being and maintain ethical standards are key to promote responsible use.

As the use of synthetic data becomes widespread, the need to consider public perceptions and understanding of its uses, benefits, limitations, and potential risks will need to be transparently communicated to stakeholders. Generators of synthetic data should be mindful of ensuring the quality and validity of synthetic data, and be able to describe how it was generated, how it differs from the original and how it could be reproduced.

Data sharing and data exchange is an aspiration for many. The ideal of an open government in which intensive use of information and communication technologies facilitate transparency, public collaboration, and civic participation (Tai, 2021), may be one which can be realised in part through synthetic data and other privacy-enhancing technologies. With the recent Parliamentary decision to establish a data bridge with the US, from October 2023, “UK businesses and organisations will be able to make use of this data bridge to safely and securely transfer personal data to certified organisations in the US” (GOV, 2023).

Synthetic data and the information professions

The benefits of synthetic data to organisations are in a sense limitless. While real-world data remains the preferred choice in many contexts, synthetic data holds the potential to bridge data access gaps, not only for commerce but also for research and evidence-based policymaking. The work of regulatory agencies will need to focus on promoting transparency and accountability, whilst proposing strategies for risk mitigation. A synthetic dataset chain of custody to maintain data integrity, traceability and accountability will become crucial to building the necessary trust. Businesses will need to ensure accountability by design and have in place governance processes that ensure a degree of human oversight during the design and implementation of synthetic datasets. The auditing at regular intervals, provision of clear documentation and continuous risk evaluation will be necessary for businesses who plan to make heavy use of synthetic datasets, either developed in-house or purchased from either open source or proprietary synthetic data providers. Of this latter source, businesses need to be vigilant with regards to the quality of synthetic datasets, especially as currently there are no data standards or benchmarks for accuracy and levels of privacy provided. The reality is that as global regulations governing data subject privacy get tighter and as people are more privacy conscious and therefore less likely to give consent, businesses face a shortage of customer data and therefore, every company who utilises AI technology, will at the very least experiment with synthetic data.

It is in this general oversight that the information profession has an important role to play in both the generation and use of synthetic data. Information professionals often have good oversight of the uses and sources of information within corporate contexts, and generally a professional awareness of the legal, regulatory and ethical concerns relating to information and data. In addition, information professionals often mediate between technical disciplines and business management. While information professionals may not be directly involved in creative synthetic data applications their professional expertise gives them an ideal perspective from which to understand the uses, and issues in adopting synthetic data solutions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Preeti Patel https://orcid.org/0009-0003-1806-6198

References

ACM (2021), Companies Beef Up AI Models with Synthetic Data, The Wall Street Journal, available at: https://cacm.acm.org/news/254385-companies-beef-up-ai-models-with-synthetic-data/fulltext (accessed 10 October 2023)

Ambrosio J (2022) How Waymo is using ML to Build a Scalable, Autonomous Driver, Google Waymo, available at: https://exchange.scale.com/public/blogs/how-ml-waymo-building-scalable-autonomous-driver-dmitri-dolgov (accessed 10 October 2023)

Barth A (2022) Amazon Sagemaker Ground Truth Now Supports Synthetic Data Generation, AWS News Blog, available at: https://aws.amazon.com/blogs/aws/new-amazon-sagemaker-ground-truth-now-supports-synthetic-data-generation/ (accessed 10 October 2023)

FCA (2022), Synthetic Data to Support Financial Services Innovation, Financial Conduct Authority, available at: https://www.fca.org.uk/publication/call-for-input/synthetic-data-to-support-financial-services-innovation.pdf (accessed 10 October 2023)

Giuffre M, Shung DL (2023) Harnessing the power of synthetic data in healthcare: innovation, application and privacy.

Nature Portfolio Digital Medicine 6: 186. DOI: 10.1038/s41746-023-00927-3

Gonzales A, Guruswamy G, Smith SR (2023) Synthetic data in health care: a narrative review. *PLOS Digital Health* 2(1). DOI: 10.1371/journal.pdig.0000082.

GOV UK (2023), UK-US Data Bridge Supporting Documents, Department for Science, Innovation and Technology, available at: https://www.gov.uk/government/publications/uk-us-data-bridge-supporting-documents/ (accessed 10 October 2023)

Jonk G, Thota B, MacGillvray S (2023) Is Data Really the New Oil?, Kearney, available at: https://www.kearney.com/service/analytics/article/is-data-really-the-new-oil (accessed 4th January 2024)

Linden A (2022) Is Synthetic Data the Future of AI?, Gartner Q&A, available at: https://www.gartner.com/en/newsroom/press-releases/2022-06-22-is-synthetic-data-the-future-of-ai (accessed 10 January 2024)

Lundy-Bryan L (2021) Privacy Enhancing Technologies Part 2 - The Coming Age of Collaborative Computing, Lunar Ventures – Insights Series, downloaded from: https://lunar.vc (accessed 10 October 2023)

Statice (2022), Roche Shares Synthetic Clinical Data for Machine Learning - A Case Study, Statice (now Anonos.com), downloaded from https://www.statice.ai/case-study/roche-synthetic-clinical-data#get-case-study (accessed 10 October 2023)

Syntheticus (2023) Democratising Data Access with Synthetic Data Whitepaper, Syntheticus, (downloaded from https://syntheticus.ai/democratizing-data-access-with-synthetic-data-whitepaper (accessed 10 October 2023)

Tai K-T (2021) Open Government research over a decade: a systematic review. *Government Information Quarterly* 38(2).

Tinchez G, Czarnowska M (2023) How Alexa Learned to Speak with an Irish Accent, Amazon Science, available at: https://www.amazon.science/blog/how-alexa-learned-to-speak-with-an-irish-accent (accessed 10 October 2023)

Tredinnick L (2022) Ubiquitous information, *Business Information Review* 39(1): 9–11.

Tredinnick L (2023) Dangerous Data: analytics and information behaviour in the commercial world. *Business Information Review* 40(1): 10–20.

UKSA (2022). Ethical Considerations Relating to the Creation and Use of Synthetic Data, UK Statistics Authority, available at: https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-relating-to-the-creation-and-use-of-synthetic-data/pages/5/ (accessed 10 October 2023)

van Laarhoven-Smits E (2023) If Data Is the New Gold, How Do You Use it to Create Value, EY: Switzerland. Available at: https://www.ey.com/en_ch/consulting/how-forward-thinking-organizations-are-becoming-data-driven

Walker K (2020) Synthetic data: Unlocking the power of data and skills for machine learning, Data in Government, available at: https://dataingovernment.blog.gov.uk/2020/08/20/synthetic-data-unlocking-the-power-of-data-and-skills-for-machine-learning/ (accessed 10 October 2023)

Author biography

Professor Preeti Patel is currently the Head of Computer Science and Applied Computing at London Metropolitan University. Prior to entering the Higher Education sector, she worked within the IT industry for organisations including British Telecom, The Wellcome Foundation and General Electric Information Services. She has a keen interest in international education and has been previously involved with the NCC global programmes for a number of years and continues to work with international partners. Her current research interests include FAIR and FATE data challenges, synthetic data, big data fusion, the data science curriculum, AI-generative models for Data Science and learning-related issues for database environments and enhancement of student engagement. Professor Patel is a Principal Fellow of the Higher Education Academy (Advance HE), a member of the British Computer Society, the Chartered Institute for IT and an associate Member of the Chartered Management Institute.