Deloitte.

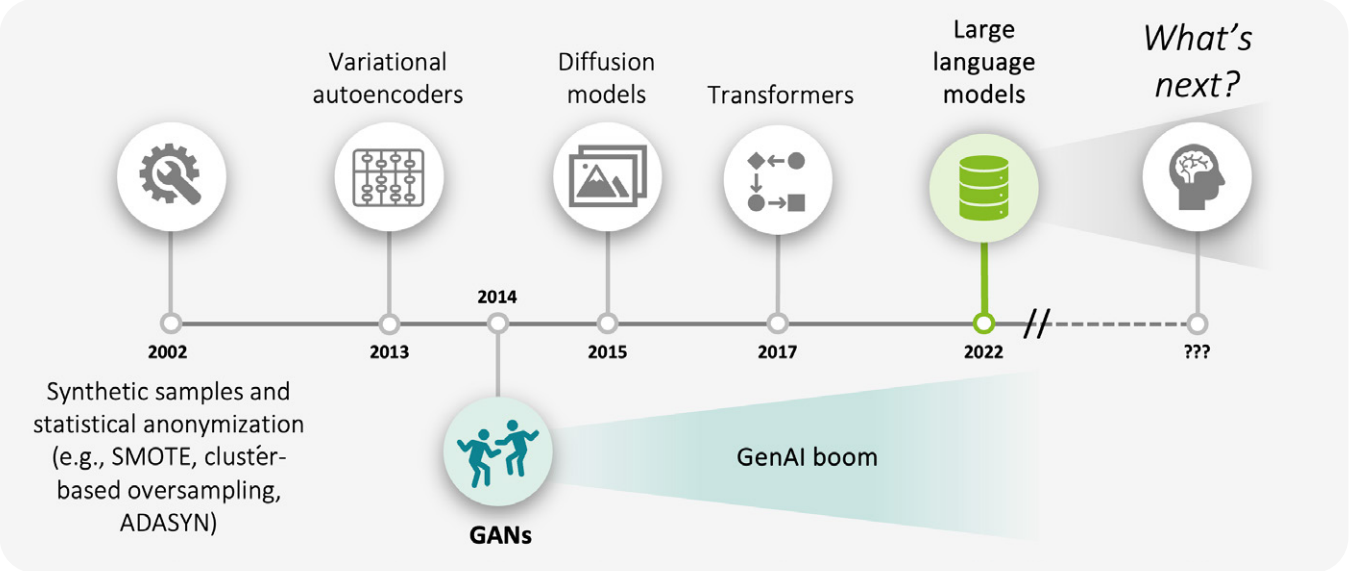
Transformational Impacts of Generative AI on Synthetic Data Generation

Authors: **Adita Karkera, Mike Greene, Ashley Hall, Landon Henderson, Kelly Lewis**

A report by Gartner predicts that by 2030, most data in AI models will be synthetic data.[1] Although synthetic data has been around since the latter half of the 20th century, Generative AI (GenAI) is changing the game for federal organizations and businesses to use synthetic data more effectively. It enables faster and more accurate synthetic data than ever before, prompting a reevaluation of how organizations can better deploy AI solutions while also protecting citizen privacy and upholding public trust.

What is synthetic data?

Synthetic data, at its core, is artificially generated data that resembles real data, consisting of the same relationships and trends that are present in the real data. This data could be text data, numerical data, images, videos, or even sound. With all the latest GenAI buzz, it can be easy to miss that synthetic data is by no means a new concept. It's been around in less mature constructs since the 1960s. It was used to tackle problems like generating data in computer games or simulating macro- and micro-level phenomena like galaxies and atoms in scientific modeling.[2] In a 1993 paper considered to be the official birth of synthetic data, Rubin popularized its use for preserving confidentiality, in this case to offset missing survey responses in the US Census,[3] but much of the more recent techniques boomed in the early 2000s.



Prior to this boom, synthetic data techniques took the form of rules-based approaches, in which generated data follows human-defined rules or constraints regarding relationships among data elements. While those rules-based processes got synthetic data started, they have some drawbacks. To scale a rules-based solution, the user would have to manually consider thousands of variables and their interactions to simulate missing data that is needed to predict, for example, incidences of heart attacks or emergency room admissions to optimize hospital staffing. The human labor involved to define those complex variable relationships, such as those among age, prior hospitalizations, and other comorbidities, is an insurmountable dilemma. Data simulated in this manner may disregard outliers or naturally occurring irregularities. And to top it all off, since humans are defining the rules in the first place, you're likely already beginning with human biases or mistakes encoded directly into the data.

In 2002, Synthetic Minority Oversampling Technique (SMOTE) kick-started an era of more sophisticated synthetic data generation algorithms. With SMOTE, new synthetic data points (oversampled data) are interpolated from existing underrepresented data (minority data) to create a more balanced and comprehensive dataset. It is also primarily used for statistical anonymization to protect sensitive data like protected health information (PHI), as discussed in the example above. While this interpolation-based approach is more scalable than prior rules-based approaches, it is limited in its ability to generate realistic data points.

Unleashing the power of GenAI

GenAI changes the game of synthetic data with robust, realistic data generation created by a machine learning (ML) model taught on a real dataset.[4] The GenAI model learns from the real data on its own without humans telling it how to create the synthetic data. GenAI emulates the meaning and relationships within the data, at scale and at lowered cost, and is of significant value for sensitive, sparse, or limited data. When implemented appropriately, synthetic data using GenAI offers substantial value and potential to improve your organization's processes and achieve its missions.

Let's consider three use cases to highlight the power of GenAI and synthetic data, while reserving how to get started to do something similar in the "Getting started" section on page 5.

1. Protecting PII, PHI, and sensitive data

Synthetic data and GenAI can be used to protect personally identifiable information (PII), PHI, and sensitive data. Let's say there are medical claims to analyze, but we can't obtain a sufficient amount because the data is sensitive and can't be shared between departments. Sensitive data often takes time to obtain approvals for data sharing and requires anonymization. Sometimes users are unwilling to share or use it at all, leaving no options to use the data effectively for advanced AI techniques. To circumvent this problem, we can apply synthetic data generation using GenAI, resulting in benefits that include cost and time savings, minimized risk of sensitive data leaking, and increased accuracy versus older methods of synthetic data generation.

2. Filling in data gaps

It is very common, especially in the federal space, to work with datasets that have missing fields. When citizens apply for benefits at various agencies or fill out their annual tax returns, for example, they may not fill out all necessary fields. This lack of data makes it difficult for government programs to perform operational duties including data analysis and advanced AI techniques. The more data you have, especially the more *complete* data you have, the more accurately an AI model can learn from the information it's given and predict mission outcomes. Sufficiently large and comprehensive datasets produced by GenAI can offset gaps in data collection to better execute your program's missions and serve the public.

3. "Debiasing" a dataset

We can use synthetically generated data to "debias" a dataset in which certain subpopulations are over- or underrepresented in the data. Perhaps an organization's HR department decided it wanted to save time reading résumés by screening them with a model that learns from several years of prior successful hires to the company. But what if the prior hires trended toward more men than women or toward more civilians than veterans? To only use the company's previously hired employees' résumés then introduces AI-enabled bias to select certain candidate profiles.[5] With synthetic data, you can limit AI-enabled model bias by increasing the underrepresented population in the data, and with GenAI, you can have a more efficient and accurate option than previous algorithms such as SMOTE.

What to look out for

Without careful preventive measures, a GenAI model can inadvertently release private or confidential information. Since the proliferation of GenAI, bad actors can generate their own synthetic data for illegal ends, like the $35 million scam in 2020 that used a synthetically generated CEO voice to authorize such a considerable monetary transfer.[6]

Another kind of attack seeks to capture an individual's private information. A membership inference attack (MIA) reverse engineers the architecture of a model to infer whether specific sensitive information was used within the data used to teach the model. Models are especially vulnerable to MIAs if they are overfit, meaning the model is overconfident when predicting real-world examples. This tends to happen when the model learns from the data it's trained on so thoroughly, that it can't generalize well to new data.

Proper quality assurance must be implemented at every stage, from model development to testing to deployment, to mitigate data privacy and security risks. Such attention can also result in generating increased stakeholder trust. You can safeguard against malicious data attacks, unintentional data breaches, and bad actors using synthetic data with preventive measures such as those listed below.

1. **Fraud detection models** can pinpoint anomalous activity and bad actors taking advantage of synthetic data. For example, fraud detection can help flag images generated by AI for someone looking to get ahead in job applications or to forge medical insurance. Under the hood of one such fraud detection model is a statistical phenomenon known as Benford's law. It ascribes a higher probability of lower leading digits like 1s and 2s than higher leading digits in a list of numbers. If the pixel values in an image, for example, deviate significantly from this expected pattern, the image is likely to have been generated by AI.[7]

2. **Membership inference tests** can measure the amount of privacy leakage in machine learning (ML) models by performing an MIA on the model to test its defensibility. If the results of this test are unfavorable, an ML model's defense can be bolstered by various techniques, such as differential privacy. This introduces "noise," or small intentional random variations in the training data, to conceal the true data and improve an AI model's robustness against attacks.

3. **Red teaming** is a standard cybersecurity technique that simulates system infiltration in a safe setting. The "red team" acts as the malicious attackers, executing the best tactics real attackers could use to uncover security gaps. Based on the red team's discoveries, your team can then address vulnerabilities and improve your organization's readiness against data security threats.

4. **AI model optimization** prevents overfitting so that the model performs well on new predicted cases. Techniques to optimize your model's performance will depend on the type of model and its susceptibilities. A trained team of data professionals can enhance a model's accuracy to more confidently protect the model's data.

Synthetic data is no substitute when real data is already comprehensive and readily accessible, or when the real data's usage is constrained by ethical considerations. If the data is being used to justify important decisions or policies, then more precise real-world data may be a better strategy and garner more trust. This is distinctly apparent in a life-or-death situation, such as using data to inform disease diagnoses or train AI-piloted fighter jets. Human life should always be at the center of all AI development. While synthetic data augments scalability, affordability, privacy, and efficiency, it should never be implemented at the expense of welfare and safety. With these challenges and ethical considerations in mind, determining whether synthetic data is right for your organization is the first step toward optimizing your data's impact.

Getting started

With the advent of GenAI, there are more opportunities for synthetic data to take center stage across many different industries and organizations. With time, applications can continue to grow as we challenge our current ways of using data. How can you navigate the newness and harness the potential of synthetic data?

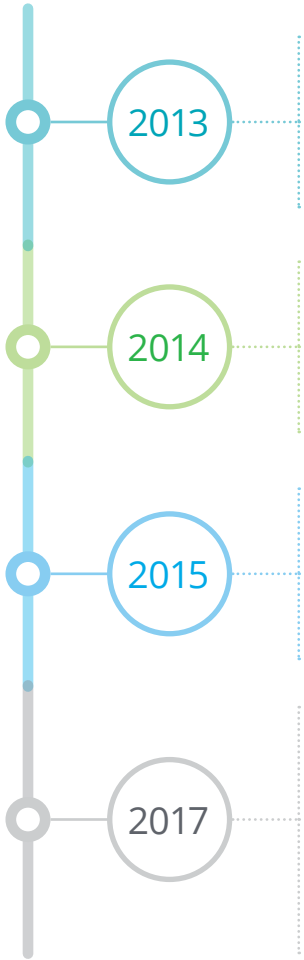
Start small. Identify existing use cases and areas of opportunity within your organization.

Organize a team. Gather a team of data professionals who can explore and formulate more concrete synthetic data via GenAI proofs of concept catered to your goals and needs.

Engage stakeholders. Get internal and external leaders on board. Determine the appropriate actionable proofs of concept that enhance your organization's data privacy, efficiency, or accessibility.

Consolidate data. Create a centralized, one-stop shop as a starting resource for people in your organization to use synthetic data, facilitating data sharing and exploration of useful applications.

Explore models. Explore your AI modeling options for synthetic data generation. All these GenAI models have been developed and refined over the past decade with applications in text, video, or image generation as well as enhancing or modifying these types of data. The right one for you depends on your use case,[8] but you can start with any of the below, noting the year these techniques were founded.



Variational autoencoders (VAEs): A successor to classical autoencoders, VAEs are a type of neural network that first encodes data to compress it, then decodes data to reconstruct it.[9] In this way, VAEs can generate new examples from a portion of the original information, identify anomalies present in the data, and fill in any incomplete data.

Generative adversarial networks (GANs): A GAN plays a game with itself using two neural networks. The first, the "generator," looks at randomly sampled values from a normal distribution and generates a synthetic data element. The second, the "discriminator," attempts to discern the real data from the fake data. Over many tries, the generator figures out how to create a realistic output that tricks the discriminator.[10]

Diffusion models: Diffusion models work in a similar way to VAEs where the first step is to change the data and then to revert it. However, for diffusion, more noise or variation is added to the data and then the model attempts to remove the noise to uncover the original data.[11] They are a current staple for image generation due to the high-quality level of detail they produce.

Transformers: Transformers put a new spin on the encoder-decoder logic of VAEs. The differentiator is an attention mechanism that gives the model the ability to focus on different parts of the input data at the same time.[12] This is revolutionary for understanding and contextualizing relationships between data. Transformers are at the heart of many text-related tasks such as sentiment analysis to understand attitudes toward a topic or entity recognition to identify people, places, and things and large language models (LLMs), which are built upon Transformer architecture.

Innovate and iterate. Employ and train the people who can make it happen. Encourage a culture of innovation and revisit your organization's missions and strategies[13] to arrive at the most valuable final products.

Synthetic data is not new, but its possibilities are expanding rapidly with GenAI. Whether your goal is to preserve the privacy and security of your data or to improve its reliability by mitigating bias, embracing the future of synthetic data with the transformational impacts of GenAI has the potential to better serve your organization's data and the people who rely on it.

Authors

Adita Karkera, GPS Chief Data Officer
Chief Data Officer, Deloitte Government & Public Services

Adita Karkera serves as the Chief Data Officer for Deloitte Consulting LLP's Government & Public Services practice and is a fellow at the Deloitte AI Institute for Government. With more than 23 years of industry experience, Adita is an award-winning executive, dedicated to improving public service. She serves on numerous industry boards and data management industry forums, including the CDO Magazine Editorial Board. She is a pioneer in articulating the importance of data literacy especially in accelerating advanced analytics and trustworthy AI adoption in government.

LinkedIn: www.linkedin.com/in/aditakarkera/ X: https://twitter.com/akark_datagirl?lang=en

Mike Greene
Technology Fellow, Artificial Intelligence & Data Engineering,
Deloitte Government & Public Services

Mike Greene is a data scientist passionate about helping our clients use data and advanced analytical solutions to crack the biggest problems. For more than 15 years, he has designed, built, and implemented statistical, behavioral, and machine learning solutions for public- and private-sector organizations to improve outcomes. Mike holds an AB in mathematics from the University of Chicago, and an AM in statistics from Harvard University. He has been a member of the American Statistical Association since 2003 and proudly knows more computer languages than spoken languages.

LinkedIn: https://www.linkedin.com/in/mike-greene-a026221/

Ashley Hall
Manager
Government & Public Services
Deloitte Consulting LLP

Landon Henderson
Senior Consultant
Government & Public Services
Deloitte Consulting LLP

Kelly Lewis
Consultant
Government & Public Services
Deloitte Consulting LLP

Endnotes

1. Linden, A. (2022, June 22). "*Is Synthetic Data the Future of AI?*". Gartner Press-releases.
2. Andrews, G. (2021, June 8). "*What Is Synthetic Data?*". NVIDIA Blogs.
3. Rubin, D. (1993). "*Discussion: Statistical disclosure limitation*". Journal of Official Statistics.
4. Buren et al., (2024). "*Maximizing the public good: How Generative AI can enhance government programs and services*". Deloitte.com.
5. Parikh, N. (2021, October 14). "*Understanding Bias In AI-Enabled Hiring*". Forbes Human Resources Council.
6. Brewster, T. (2021, October 14). "*Fraudsters Cloned Company Director's Voice In $35 Million Heist, Police Find*". Forbes.com.
7. Bonettini et al., (2021, January). "*On the use of Benford's law to detect GAN-generated images*". 2020 25th International Conference on Pattern Recognition (ICPR).
8. Buren et al., (2024). "*Maximizing the public good: How Generative AI can enhance government programs and services*". Deloitte.com.
9. Kingma, D., Welling, M. (2013, 20 December). "*Auto-Encoding Variational Bayes*". https://doi.org/10.48550/arXiv.1312.6114.
10. Goodfellow et al., (2014, 10 June). "*Generative Adversarial Networks*". https://doi.org/10.48550/arXiv.1406.2661.
11. Dickstein et al., (2015, 12 March). "*Deep Unsupervised Learning using Nonequilibrium Thermodynamics*". https://doi.org/10.48550/arXiv.1503.03585.
12. Vaswani et al., (2017, 12 June). "*Attention Is All You Need*". https://doi.org/10.48550/arXiv.1706.03762.
13. Karkera et al., (2023, 17 July). "*Introduction to CDO 2.0*". Deloitte Insights.

Acknowledgments

The authors thank Harlan Simpson for her assistance in drafting and editing. The authors also thank the numerous reviewers who provided invaluable feedback on earlier drafts of this paper.

Deloitte.

About Deloitte

Designed by CoRe Creative Services. RITM1881182