Innovation Insight for Synthetic Data

Published 7 February 2022 - ID G00757632 - 14 min read

By Analyst(s): Arun Chandrasekaran, Alexander Linden, Anthony Mullen, Farhan Choudhary

Initiatives: Digital Future; Artificial Intelligence

Synthetic data can be an effective supplement or alternative to real data, providing access to annotated data to build accurate, extensible AI models. Enterprise architecture and technology innovation leaders should evaluate its benefits, risks, use cases and tech ecosystem to reap business value.

Additional Perspectives

- Summary Translation: Innovation Insight for Synthetic Data (24 February 2022)

Overview

Key Findings

- Poor data quality, lack of adequate data, siloed data and bias in training data are among the top data-related challenges with AI initiatives.
- Artificial data generated using AI techniques (synthetic data) can improve AI model accuracy, time to value, aid with regulatory compliance and lower the cost of data acquisition.
- Synthetic data can be both structured and unstructured data, with tabular data and images being the most commonly used forms today.
- There is a growing ecosystem of startups in this space, with levers of competition being the impact on model performance, types of data they can generate, privacy filters they offer and the industry use case focus.

Recommendations

To best sense and respond to disruptions using technology innovation, enterprise architecture and technology innovation leaders should:

- Develop guidelines, in conjunction with analytics, security and legal teams, on appropriate usage of synthetic data.
- Educate the internal stakeholders through training programs on the benefits and limitations of synthetic data and institute guardrails to mitigate challenges (such as user skepticism, inadequate data validation and improper feature engineering).
- Conduct a POC to verify vendor claims and validate use case fit. Choose vendors that can generate realistic synthetic datasets for your use cases, provide tools to measure the effectiveness of synthetic datasets, and provide privacy filters to comply with regulations and internal compliance mandates.
- Measure and communicate the business value, success and failure stories of synthetic data initiatives, as this creates realistic expectations on the art of what's possible and provides opportunities for continuous exploration.

Strategic Planning Assumptions

By 2025, the use of synthetic data and transfer learning will reduce the volume of real data needed for machine learning by 70%.

By 2025, synthetic data will reduce personal customer data collection, avoiding 70% of privacy violation sanctions.

Introduction

> **"In God we trust, all others bring data."**
>
> — *W Edwards Deming*

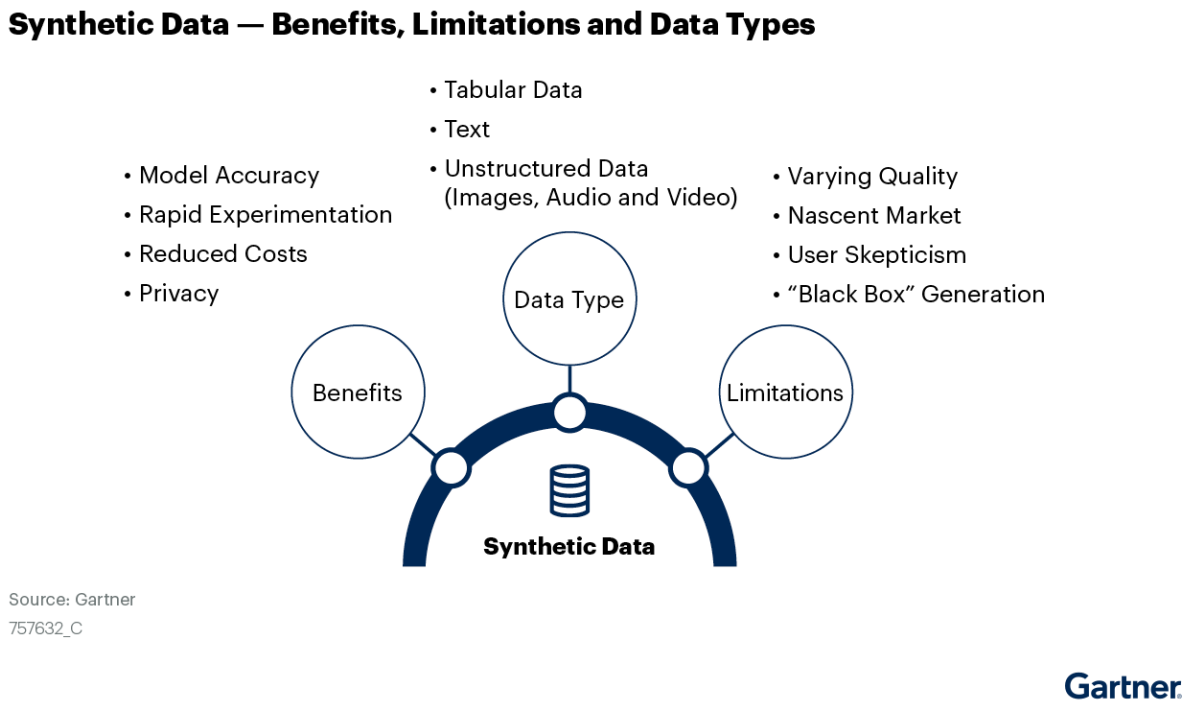
Data is the lifeline for digital businesses. However, getting access to real-world data that is of high quality (that is, clean, well-labeled and audited for bias) is often challenging for enterprises. Synthetic data can address this challenge to enable faster, accurate and responsible AI initiatives.

Synthetic data is a class of data that is artificially generated. It can be used to train AI models for scenarios such as:

- Use cases for which data must be guaranteed to be anonymous or for which privacy must be preserved (such as with uses of medical data).
- Augmentation of real data, especially where the cost of data collection is high.
- There is a need to balance class distribution within existing training data (such as with population data).
- Emerging AI use cases for which limited real data is available.

The key benefits, limitations and types of synthetic data are summarized in Figure 1.

Figure 1: Synthetic Data — Benefits, Limitations and Data Types



While it may be tempting to dismiss synthetic data as "fake data," therein lies its power. Synthetic data is critical, since it can be generated to meet specific needs or conditions that are not available in real data, which makes it potentially context-aware, domain-specific and privacy-enhanced. We believe that synthetic data is important for the future of AI because it solves one of the most pervasive and critical challenges that AI systems face today — the lack of domain-specific, well-labeled, high-volume data at a reasonable cost.

Description

Synthetic data is a class of data that is artificially generated (that is, not obtained from direct observations of the real world). Data can be generated using different methods, such as statistically rigorous sampling from real data, semantic approaches, generative adversarial networks, or by creating simulation scenarios where models and processes interact to create completely new datasets of events.

While there is a lot of hype around the promise of synthetic data recently, it isn't a completely novel idea. Synthetic data has been around for decades and has been used in applications such as computer games and scientific simulations. However, the recent advancements in computing power coupled with the advent of generative adversarial networks (GANs) have been a game changer in our ability to generate faster, cheaper and more accurate synthetic data.

Synthetic data can be generated using several techniques, such as:

- **Statistical Distribution:** Simple ways to generate synthetic data include observing the statistical distribution of real data and replicating the same through discrete or joint distributions to produce similar data. This can be an effective technique for simple and small-scale use cases involving tabular data.
- **Variational autoencoders (VAEs):** VAEs are a form of neural net. VAEs consist of an encoder and decoder. The encoder takes objects and compresses them into more condensed representations while retaining the main features. These representations can then be mapped onto a two-dimensional space where similar objects are clustered. New objects are generated by decoding a point in the dimensional space, say, between two objects. For example, a VAE could create an image that looks like a mix of a drone and a hovercraft. VAEs are relatively easy to implement and train. However, if the original data is highly heterogeneous, VAEs can suffer from reconstruction errors.
- **Generative adversarial networks (GANs):** GANs are a form of neural net, where two neural networks are trained in an adversarial fashion. A generator generates the data, while the discriminator evaluates whether it's fake or real. When both networks are trained together, the discriminator needs to learn from patterns in real data, where the generator learns to outsmart the discriminator by producing more realistic samples from its random input. The advantage of GANs is their ability to learn the characteristics of "real" data quickly and iterating faster toward accurate representation. GANs are specifically powerful with unstructured data, such as images, although they may be more complex to train.

- **Others:** There are other techniques, such as gaming engines, and physics-based approaches for generating synthetic datasets.

Synthetic data can include both structured and unstructured data. The most common forms of synthetic data include:

- **Tabular data:** This form of synthetic data is structured and stored in a table in rows and columns. This mimics structured data stored in data warehouses and can also include forms such as time series data.
- **Text:** Text-based synthetic data has been challenging historically due to massive amounts of pretraining needed and the contextual aspects of natural language. However, recent advancements in large language transformer models, such as BERT and GPT-3, have made text-based synthetic data possible. These models are massively pretrained on large datasets, including public repositories such as Wikipedia and Project Gutenberg books library, which enables better accuracy and cross-domain applications.
- **Unstructured data:** This form of synthetic data includes images, audio and videos. Use cases often include generating images or sounds of people, where the individuals aren't real for privacy preservation. This is one of the fast-growing use cases for next-generation AI applications, such as self-driving cars, geolocation services, healthcare patient images and cashierless stores.

Synthetic data within enterprises can be used in a variety of ways — either fully synthetic, where there is no real data or in a hybrid manner (more common); where the real data is augmented or replaced with synthetic data; or where synthetic data is used to enhance class distribution and provide adequate quantity of well-labeled data.

Benefits and Uses

While synthetic data is not quite the panacea it is painted to be in the popular media, there are real, tangible benefits to its use in the data pipeline and in product development, such as:

- **Increased accuracy of ML models.** Real-world data is happenstance and does not contain all permutations of conditions or events possible in the real world. Synthetic data can counter this by generating data at the edges or for conditions not yet seen. More permutations of event data and the fact that synthetic data can be labeled automatically make for more accurate models. Today's state of the art in computer vision and natural language use cases is to use synthetic data in conjunction with real data to train models using transfer learning.
- **A faster data pipeline.** By switching to synthetic data, you can speed up (or avoid) internal processes, lengthy contractual efforts, legal blockers or hosting challenges.
- **The ability to safely experiment.** Synthetic data "unlocks" signals in private and sensitive data that otherwise could not be examined. It also creates new opportunities to use cloud services or multiparty analytics in an effective way.
- **A reduction of costs.** Synthetic data reduces the cost of time and money to create, buy, collect and label data for improved quality.

There is great applicability across all industry verticals (see Table 1).

Table 1: Synthetic Data Use Cases

(Enlarged table in Appendix)

Use Case	Analysis
Pool and share data between territories, companies and business units.	Although many organizations wish to collaborate, data security and privacy often inhibit them. Synthetic data allows for one or more synthetic datasets to be shared in a business ecosystem to generate new value in a safe way. In many government organizations, collaborating internally can be very difficult due to policy and time, so the principle can also apply to internal data sharing.
Open up the use of third-party analytics consultants.	By allowing third parties access to synthetic datasets, organizations can expand their pool and talent of analytics workers.
Usage for vendor evaluation and procurement.	Synthetic data can be more easily provided to third parties to test their platforms.
Test data for software engineering.	This allows for test data to better cover future real-world situations and account for seasonal or other statistical gaps in datasets.
Enable cloud adoption and migration.	Security and compliance risks often stymy the adoption of cloud services. By moving the synthetic counterparts of some datasets to the cloud, organizations can take advantage of the wide array of services previously not permissible.
Create opportunities for data monetization.	Although it is possible that a marketplace for synthetic domain data will appear in the next few years, to date this is a rare use case/possibility.
Help combat data losses due to data retention policies.	Data from customers is not stored in perpetuity. Often, data protection laws will require the removal of data after a certain time period. However, this will affect your ability to include historical temporal, behavioral and seasonal data in model development. By creating a synthetic snapshot, this data need not be "lost."

Source: Gartner (January 2022)

Some specific uses we see of synthetic data in particular Industries include:

- **Finance:** Synthetic generation of fraud events, multiagent simulations to explore market behaviors (such as pension investments and loans) and for better lending decisions.

- **Automotive:** Testing robotic and automotive systems is slow and costly. Synthetic data allows the automated vehicle (or its digital twin counterpart) to explore a dynamic environment safely and quickly.
- **Healthcare:** Share data with external parties but still maintain patient confidentiality (for example, in clinical trials).
- **Retail:** Autonomous check-out systems, cashierless stores, analysis of customer demographics and improving inventory management are some examples of synthetic data usage in retail.
- **Defense:** The power of synthetic data in defense is its ability to provide training data for even the rarest occurrences, which enhances the risk assessment and threat mitigation capabilities.

Risks

While synthetic data techniques can score quite highly for cost-effectiveness and privacy, they do have risks and limitations, such as:

- The quality of synthetic data often depends on the quality of the model that created it and the input "seed" dataset.
- If seed datasets change, it is necessary to regenerate synthetic data using the new characteristics for it to enable meaningful model accuracy.
- Using synthetic data requires additional verification steps, such as the comparison of model results with human-annotated, real-world data, to ensure the fidelity of results.
- Beyond the technology challenges, user skepticism might be a hard challenge for synthetic data to overcome, as users may perceive it to be "inferior" or "fake" data.
- As synthetic data gains broader adoption, more questions will be posed on the openness of the data generation techniques and the efficacy of complete privacy guarantee, particularly in sensitive use cases, such as clinical trials or demographic surveys.
- If fringe or edge cases are not part of the seed dataset, they will not be synthetized. This means the handling of such borderline cases must be carefully accommodated.

Adoption Rate

Synthetic data is an emerging technology with a low degree of enterprise adoption. While tabular data and image data use cases may have less than 5% of target market adoption, the usage of synthetic text, video and audio is nascent, at less than 1%. The increased VC funding, emergence of more than 50 different startups in this space and its substantial business benefits point to a robust future for this technology. Moreover, data science and data fabric vendors will be embedding synthetic data generation capabilities either through organic or inorganic means as part of a platform play.

Although synthetic data adoption is in its infancy stage in enterprise environments, there are several emerging use cases for synthetic data that are impacting our daily lives. Examples include:

- Amazon was able to accelerate the Alexa support for new languages (such as Hindi, Spanish and Brazilian Portuguese) by using synthetic text data.
- Nationwide uses synthetic data to share data safely and faster with third-party partners to accelerate co-innovation.

Alternatives

Real data is the obvious alternative to synthetic data. There are use cases where the complexity and inability to prove the veracity of synthetic data may lead to simply using real data. Regulators are just now waking up to the benefits and risks of using synthetic data and may nudge organizations toward not using synthetic data for some scenarios. In privacy preservation use cases, techniques such as data masking and differential privacy may be deemed as alternatives to synthetic data.

Synthetic data is still nascent, with most organizations being either unaware of it or merely experimenting with it. While it holds tremendous potential, whether it can deliver substantial business benefits for a variety of enterprise use cases is yet to be seen and proven.

Recommendations

Enterprise architecture and technology innovation leaders must:

- Develop guidelines, in conjunction with analytics, security and legal teams, on appropriate usage of synthetic data.

- Educate the internal stakeholders through training programs on the benefits and limitations of synthetic data and institute guardrails to mitigate challenges (such as user skepticism, inadequate data validation and improper feature engineering).
- Conduct a POC to verify vendor claims and validate use case fit. Choose vendors that can generate realistic synthetic datasets for your use cases, provide tools to measure the effectiveness of synthetic datasets, and provide privacy filters to comply with regulations and internal compliance mandates.
- Measure and communicate the business value, success and failure stories of synthetic data initiatives, as this creates realistic expectations on the art of what's possible and provides opportunities for continuous exploration.

Representative Providers

Table 2: Sample Synthetic Data Vendors

(Enlarged table in Appendix)

Vendor Name	HQ/Location	Year of Founding	Data Type Specialization
Aindo	Trieste, Italy	2018	Tabular & time series data
Bitext	Las Rozas, Madrid, Spain	2008	Multilingual text data
CVEDIA	London, England	2016	Image data
Datagen	Tel Aviv, Yafo, Israel	2018	Image data
Diveplane	Raleigh, North Carolina, U.S.	2017	Tabular and time series data
Edgecase.AI	Hingham, Massachusetts, U.S.	2017	Image data
Facteus	Beaverton, Oregon, U.S.	2010	Tabular data
Gretel.AI	San Diego, California, U.S.	2019	Tabular, Text and time series data
Hazy	London, England	2017	Tabular data
IBM	Armonk, NY	1911	Tabular and time series data
LeapYear	San Francisco, California, U.S.	2015	Tabular data
MOSTLY.AI	Vienna, Austria	2017	Tabular and text data
Neuromation	San Francisco, California, U.S.	2017	Image data
Parallel Domain	Palo Alto, California, U.S.	2017	Image data
Rendered.AI	Bellevue, Washington, U.S.	2019	Image data
Scale AI	San Francisco, California, U.S.	2016	Image, video, unstructured (text) and tabular data
Statice	Berlin, Germany	2018	Tabular data
Superb AI	San Mateo, California, U.S.	2018	Tabular and image data
Syndata	Stockholm, Sweden	2020	Tabular and text data
Synth	London, England, U.K.	2020	Tabular data
Synthesis AI	San Francisco, California, U.S.	2019	Image & video data
Synthesized	London, England	2018	Tabular data
Syntho	Amsterdam, Holland	2020	Tabular data
Superb AI	San Mateo, California, U.S.	2018	Tabular and image data
Tonic.ai	San Francisco, California, U.S.	2018	Tabular data
YData	Seattle, Washington, U.S.	2019	Tabular & time series data

Source: Gartner (January 2022)

Evidence

Detailed briefings were conducted with several vendors represented in this research.

Synthetic Data, J.P. Morgan.

The Real Deal About Synthetic Data, MIT Sloan Management Review.

How Synthetic Data Could Save AI, VentureBeat.

Differentially Private Synthetic Data, NIST.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Innovation Insight for Generative AI

Predicts 2022: Artificial Intelligence Core Technologies

Top Trends in Data and Analytics for 2021: From Big to Small and Wide Data

Preserving Privacy While Using Personal Data for AI Training

Maverick* Research: Forget About Your Real Data — Synthetic Data Is the Future of AI

Top Strategic Technology Trends for 2022: Generative AI

Cool Vendors in AI Governance and Responsible AI

Table 1: Synthetic Data Use Cases

Use Case	*Analysis*
Pool and share data between territories, companies and business units.	Although many organizations wish to collaborate, data security and privacy often inhibit them. Synthetic data allows for one or more synthetic datasets to be shared in a business ecosystem to generate new value in a safe way. In many government organizations, collaborating internally can be very difficult due to policy and time, so the principle can also apply to internal data sharing.
Open up the use of third-party analytics consultants.	By allowing third parties access to synthetic datasets, organizations can expand their pool and talent of analytics workers.
Usage for vendor evaluation and procurement.	Synthetic data can be more easily provided to third parties to test their platforms.
Test data for software engineering.	This allows for test data to better cover future real-world situations and account for seasonal or other statistical gaps in datasets.
Enable cloud adoption and migration.	Security and compliance risks often stymy the adoption of cloud services. By moving the synthetic counterparts of some datasets to the cloud, organizations can take advantage of the wide array of services previously not permissible.
Create opportunities for data monetization.	Although it is possible that a marketplace for synthetic domain data will appear in the next few years, to date this is a rare use case/possibility.

Use Case	*Analysis*
Help combat data losses due to data retention policies.	Data from customers is not stored in perpetuity. Often, data protection laws will require the removal of data after a certain time period. However, this will affect your ability to include historical temporal, behavioral and seasonal data in model development. By creating a synthetic snapshot, this data need not be “lost.”

Source: Gartner (January 2022)

Table 2: Sample Synthetic Data Vendors

Vendor Name	*HQ/Location*	*Year of Founding*	*Data Type Specialization*
Aindo	Trieste, Italy	2018	Tabular & time series data
Bitext	Las Rozas, Madrid, Spain	2008	Multilingual text data
CVEDIA	London, England	2016	Image data
Datagen	Tel Aviv, Yafo, Israel	2018	Image data
Diveplane	Raleigh, North Carolina, U.S.	2017	Tabular and time series data
Edgecase.AI	Hingham, Massachusetts, U.S.	2017	Image data
Facteus	Beaverton, Oregon, U.S.	2010	Tabular data
Gretel.AI	San Diego, California, U.S.	2019	Tabular, Text and time series data
Hazy	London, England	2017	Tabular data
IBM	Armonk, NY	1911	Tabular and time series data
LeapYear	San Francisco, California, U.S.	2015	Tabular data
MOSTLY.AI	Vienna, Austria	2017	Tabular and text data
Neuromation	San Francisco, California, U.S.	2017	Image data
Parallel Domain	Palo Alto, California, U.S.	2017	Image data
Rendered.AI	Bellevue, Washington, U.S.	2019	Image data

Vendor Name	*HQ/Location*	*Year of Founding*	*Data Type Specialization*
Scale AI	San Francisco, California, U.S.	2016	Image, video, unstructured (text) and tabular data
Statice	Berlin, Germany	2018	Tabular data
Superb AI	San Mateo, California, U.S.	2018	Tabular and image data
Syndata	Stockholm, Sweden	2020	Tabular and text data
Synth	London, England, U.K.	2020	Tabular data
Synthesis AI	San Francisco, California, U.S.	2019	Image & video data
Synthesized	London, England	2018	Tabular data
Syntho	Amsterdam, Holland	2020	Tabular data
Superb AI	San Mateo, California, U.S.	2018	Tabular and image data
Tonic.ai	San Francisco, California, U.S.	2018	Tabular data
YData	Seattle, Washington, U.S.	2019	Tabular & time series data

Source: Gartner (January 2022)