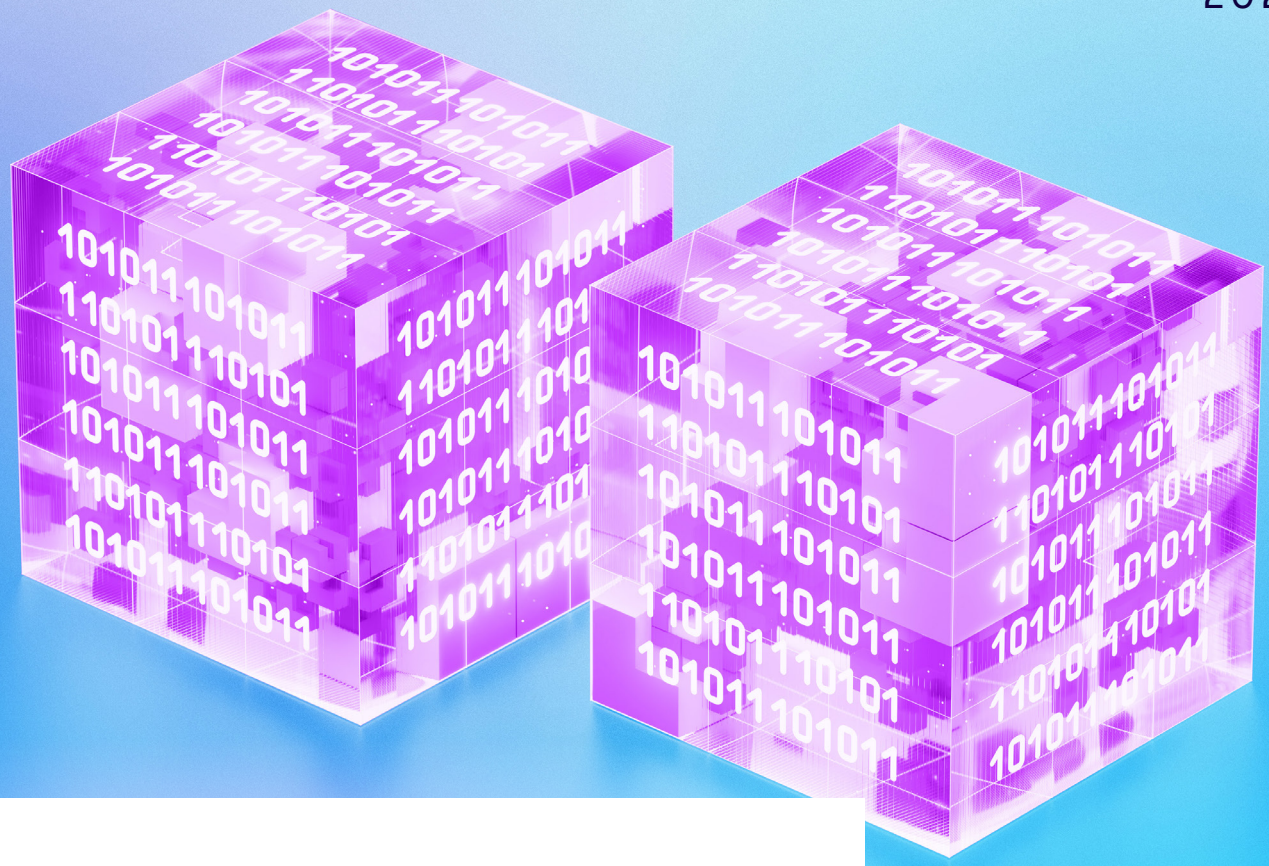
VIEWPOINT

ARTHUR D LITTLE
2024



SYNTHETIC DATA: FACILITATING INNOVATIVE SOLUTIONS

Synthetic data set to alter future of data science

Synthetic data (artificially generated data that mimics real-world data) enables AI development by addressing data scarcity and privacy concerns, enhancing model performance, enabling testing and validation, mitigating bias and security vulnerabilities, and facilitating prototyping. As we explore in this Viewpoint, synthetic data helps create robust, secure, unbiased AI systems by overcoming data challenges and privacy limitations.

AUTHORS

Sri Rajagopal

Corrine Bai

Mark Rowland

WHAT IS SYNTHETIC DATA?

In a world where data overload and storage are significant issues for business, one might wonder why we might want more data. The truth is, despite the current volume of data at our exposure, there are still gaps that make data innovation challenging. Synthetic data closely replicates the characteristics of real-world data without containing exact data points. Data scientists use algorithms and simulations to produce data that maintains the same statistical properties as the authentic data it emulates.

The use of synthetic data dates back to the 1970s. Many of the first systems and algorithms needed data to operate. Today, limited computational power, difficulties gathering large amounts of data, and privacy disclosure issues are orienting efforts toward generating synthetic data, turning this domain into a key strategic advantage.

Synthetic data can be a substitute for real-world data (or in addition to it) to power data-driven decision-making. With more businesses using machine learning (ML) and AI in to gain strategic advantage, synthetic data is gaining value due to its flexibility in training data sets. It lets data scientists create large customizable data sets and scenarios in controlled environments, enabling thorough testing and validation of models.

HOW DOES IT DIFFER?

Initially, synthetic data mainly took the form of unstructured data like synthetic images and videos. Early applications focused on generating visual and multimedia content with references to the original data points in a qualitative way. Today, synthetic data generation is capable of referring to structured data formats as input samples, thereby retaining the individual data points and their interrelationships. Structured synthetic data, with the complex interconnectedness between the data points it represents, offers a huge opportunity for business.

DESPITE THE CURRENT VOLUME OF DATA AT OUR EXPOSURE, THERE ARE STILL GAPS THAT MAKE DATA INNOVATION CHALLENGING

Unlike a mock dataset created at random, synthetic data retains the statistical information, inherent relationships, and nuances found in the real data. Synthetic data can simulate various scenarios for testing, training, and validating models, ensuring that the data used in such processes is structurally and functionally similar to data one might encounter in real-world applications. Synthetic datasets accurately reflect real-life scenarios and distributions, making them more valuable in a business setting than randomly generated mock data. Businesses can use structured, synthetic data for a wide variety of applications without compromising accuracy or privacy.

As data analytics and AI models become increasingly crucial in business decision-making, synthetic data can play a crucial role — it enables accurate, sophisticated analytics while complying with privacy regulations, saving money, and reducing risks. For example, organizations can accelerate their innovation cycles using synthetic data. Rapid prototyping, testing, and development become more efficient, helping businesses bring products and services to market faster and more efficiently. This Viewpoint details the benefits and risks of synthetic data and looks at relevant use cases in high-growth industries, where synthetic data can provide immediate value.

SYNTHETIC DATASETS ACCURATELY REFLECT REAL-LIFE SCENARIOS AND DISTRIBUTIONS

There are several approaches to generating synthetic data. Generative machine models learn how a dataset is generated from a probabilistic model and then create synthetic data by sampling from the learned distribution. Generative AI (GenAI) techniques (e.g., generative adversarial networks [GANs] and variational autoencoders [VAEs]) are options for synthetic data generation (see Figure 1). These models learn from existing data to generate new samples that closely resemble the original data distribution. By capturing the underlying patterns and structures of the real data, GenAI models can produce synthetic data that is statistically similar to the original but does not contain any sensitive or personally identifiable information. According to Gartner, by 2024, 60% of data used to train AI models will be synthetically generated. At that rate, synthetic data will overshadow AI models by 2030 (see Figure 2).

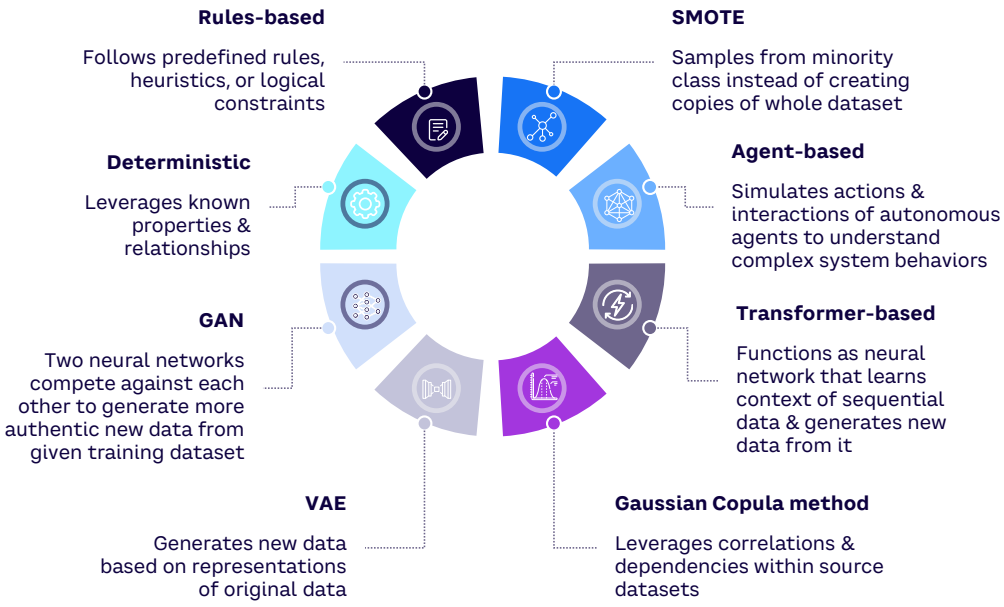
BENEFITS OF SYNTHETIC DATA

Thanks to its ability to mimic the quality of real-life data and capture the interconnectedness of individual data points, businesses across industries are turning to synthetic data. There are several advantages to this approach, as described below.

Addresses privacy concerns

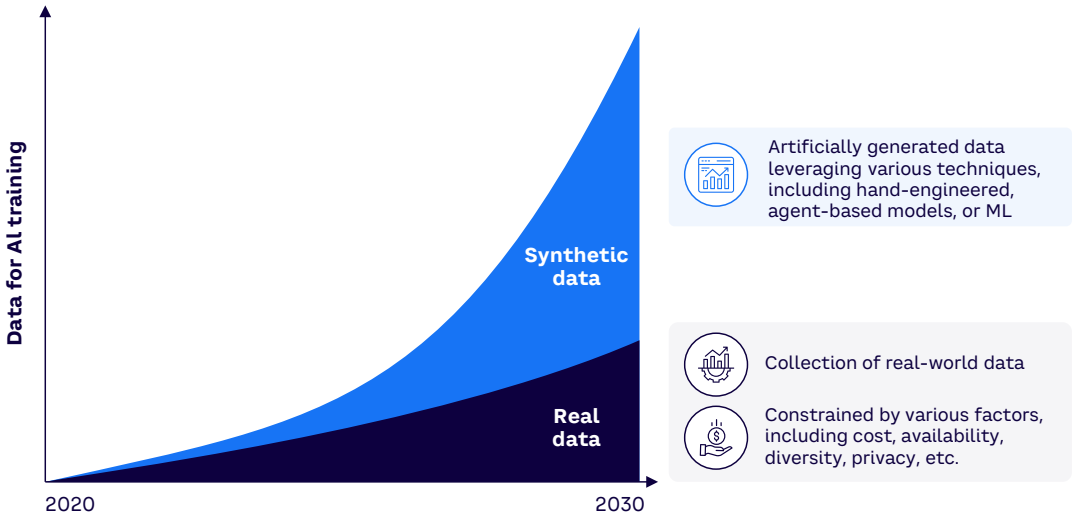
Synthetic data greatly reduces the risk of exposing sensitive or Personally Identifiable Information (PII). However, it isn't a complete solution to privacy issues. Since synthetic data builds on the statistical and inter-relational qualities of real data, it is up to the end user to decide how much information the synthetic dataset reveals about the real data from which it was generated. Businesses must determine how much information about the original data the synthetic data will reveal based on use cases. For example, for internal projects where synthetic data comes from in a secure environment, a less stringent privacy standard might be acceptable.

Figure 1. Synthetic data-generation methods



SMOTE = Synthetic minority oversampling technique
Source: Arthur D. Little

Figure 2. Synthetic data expected to overshadow real data in AI models by 2030



Source: Arthur D. Little, Gartner

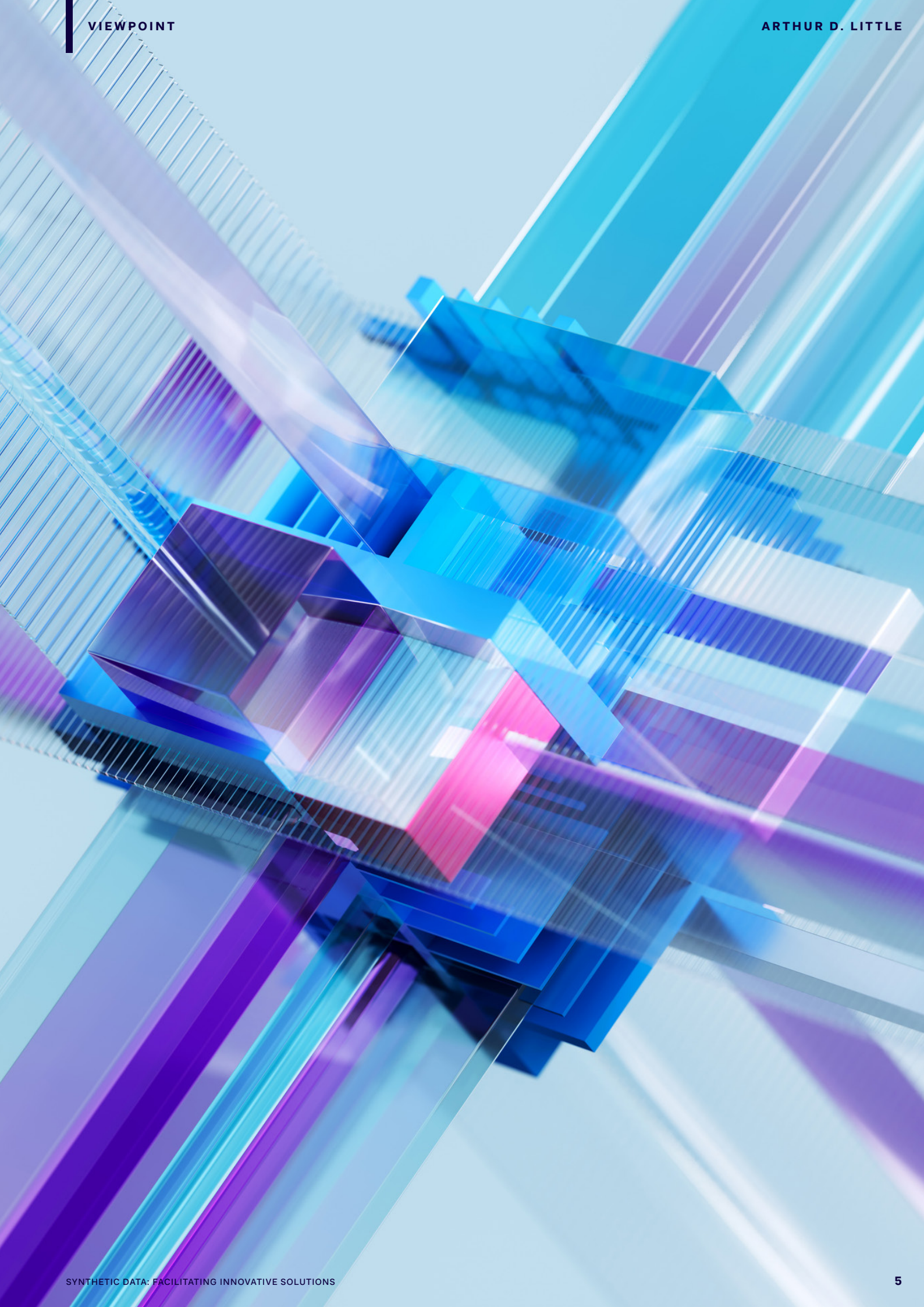
SYNTHETIC DATA CONCEALS INFORMATION CONTAINED IN REAL DATA

Enhances safety & security

Synthetic data is an excellent way to protect data from leaks. Should an unauthorized party access sensitive data with the intent of leaking or misusing it, synthetic data conceals the individual information contained in the real data, minimizing the impact of a potential data breach. This is extremely useful when organizations need to share data with third parties: synthetic data lets vendors or partners work with realistic data without exposing any sensitive/PII data collected by the business. Synthetic data also enhances data availability — having eliminated worry about leaking original data, organizations are more comfortable sharing their data for research and other purposes.

Offers high level of flexibility

Synthetic data helps companies easily create customized datasets. For example, a company could alter the quality of the real data to remove bias or test what-if situations using augmented synthetic data. When generating synthetic data, a user can produce samples that are statistically accurate but devoid of the historical bias that exist in the original data. Such datasets can be highly valuable in training ML models, mitigating the risk of amplifying bias. Similarly, a user can generate data that reflects one or more hypothetical situations, changing the distribution of the dataset while maintaining the same causal structure. Synthetic data's flexibility lets companies explore specific scenarios or needs that real-world data might underrepresent. However, one must be careful to maintain the same quality as the original data in modeling the causal structure and carefully monitor the ML training process when using modified synthetic data.



Provides cost savings

Developers need large, carefully labeled datasets to train AI models. AI models are essentially neural networks, so the larger and more diverse the training data, the more accurate the model. Of course, gathering and labeling the tens of millions of elements required in a training dataset requires a significant up-front investment in manpower and capital. Synthetic data solves this problem. According to an estimate by Paul Walborsky, cofounder of one of the first dedicated synthetic data services, a single image that would have cost US $6 from a labeling service can be artificially generated for six cents. It is easy to imagine the cost-saving impact synthetic data could make, especially when businesses require data at a high volume to power complex processes like AI model training.

Provides competitive advantage

Synthetic data lets users power AI models with fresh, authentic data. Today's AI models consume vast amounts of publicly available data from the Internet. As a result, the current AI landscape relies heavily on the same set of information, compromising the effectiveness of the resulting AI models due to the outdated nature and inaccuracy. Synthetic data lets companies create new datasets that fill in gaps (often called "blind spots") that existing data might not cover and expand on proprietary data that isn't readily available. Synthetic data helps AI models perform more accurate and reliable analysis, deriving data-based insights to drive innovation that could power long-term growth.

Synthetic data drawbacks

By no means is synthetic data a cure-all solution. Like all AI, it is susceptible to the "garbage in, garbage out" dilemma. The models that generate synthetic data are only as good as the data they are trained on. If the training data contains biases and limitations, the synthetic data produced will inherit these flaws. Additionally, there are several other potential issues with synthetic data that users should consider:

1. **Ethical concerns.** Utilizing synthetic data in sensitive domains, such as medical diagnostics, can be problematic. Inaccuracies in the data may lead to significant risks and adverse outcomes.
2. **Model degradation.** If synthetic data is not periodically updated to mirror changes in the real world, the effectiveness of AI models may diminish over time.
3. **Introduction of bias.** Care must be taken in the generation of synthetic data to avoid the introduction of bias, which can result in flawed AI models.
4. **Validation challenges.** It is challenging to validate the accuracy of synthetic data. Consequently, it is uncertain whether AI models trained on such data will perform effectively in real-world scenarios.

USE CASES BY INDUSTRY

Below are several ways businesses can leverage engaging synthetic data across industries to drive innovation and maximize efficiency.

Financial services

In banking, synthetic data has become an essential way to extract the full potential from data and further train ML models to satisfy customers and business goals without compromising privacy. The applications of this technology within advanced analytics and ML development are extensive, from improving fraud detection and market simulations to data exchange and increased collaboration across teams. Synthetic data is compliant with all General Data Protection Regulation (GDPR) requirements and allows companies, including banks, to generate the large datasets they need without running up against legal and ethical issues or losing informational or statistical properties.

Over the next few years, as banking continues its digital transformation and fintech integration, synthetic data will become even more valuable for customer acquisition, helping banks with advanced data analytics, mortgage and credit decision assessments, and many other key strategic elements.

Telecom

Data has always been extremely important to the telecom industry. To create systems and offer services that meet the rapid changes in the industry, companies rely on data to derive insights and stay on top of trends in customer behavior. However, the massive data volume these businesses deal with can be unwieldy (they collect vast amounts of data from millions of users, including call records, location data, and Internet usage).

Additionally, 80%-85% of customer data is locked away due to lack of consent. Many telecoms struggle with the high cost of protecting the sensitive nature of the vast amount of data they collect while remaining unable to unlock actionable insights when customer consent is absent.

Synthetic data steps into power analytics on a scale that original data could not achieve. It enables telecom companies to gain deep insights into customer behavior, preferences, and usage patterns without compromising individual privacy. This allows telecoms to offer more targeted and personalized services, enhancing customer experiences with synthetic versions of real data. Additionally, they can assist other businesses in optimizing campaigns and recommendations, as private information is abstracted and masked with newly generated data that maintains statistical integrity.

Healthcare

The sensitive nature of patient data makes it difficult for healthcare companies to effectively acquire, manage, and use it to perform business analytics. Thanks to stringent patient-privacy laws across the world (e.g., HIPAA, the UK Data Protection Act, the German Federal Data Protection Act), healthcare data tends to be so obscure and fragmented that organizations cannot effectively use it to understand the patient journey. At times, this has suppressed innovation in the pharmaceutical industry. In 2022, for instance, the Public Health Agency of Canada found that slow or restricted health data sharing had a negative impact on pandemic response. There is also significant bias in healthcare data, including a huge gap in women's health data. Gender and race bias can hinder an organization's understanding of various disease states and slow research in areas with unmet needs.

Since the predominant focus of synthetic data methodologies is on the accurate representation of entire populations — rather than the replication of individual entities — there is no direct link between individual data points in a synthetic dataset and the individual data points in the real sample. According to research by ADL Cutter contributor Khaled El Emam et al. in the Journal of Medical Internet Research, synthetic data generated from clinical data offers four to five times more protection against identify disclosure than the real dataset. When done correctly, synthetic data can be invaluable in healthcare data sharing.

Synthetic data's ability to preserve the interrelationships between certain data points makes it extremely useful to the pharmaceutical and healthcare industries, as diseases are more accurately represented by a disease development span than a specific point in time. Synthetic data has the ability to represent electronic health records and biometric measurements with tabular and time-series generation, creating a complete patient journey that can improve care quality and inform the development of new therapeutics.

Manufacturing

Managing defects is an ongoing challenge for manufacturers, and there is often a lack of sufficient data to train models that could identify and anticipate them. When defects do arise, they cause more harm than just revenue loss on a single product — they often cause supply chain disruptions, lead to a loss of competitive advantage, and/or waste human and monetary resources. When a statistically significant amount of real-world data isn't available, synthetic data can help grow the dataset to more effectively train a model to discover defects.

In cases where manufacturing defect data isn't readily available, such as a new assembly that hasn't yet generated enough real-world data, synthetic data can jumpstart AI training, contributing as much as 90% of the process. The resulting AI model doesn't just preemptively flag defects, it can help a business understand how it would perform under a variety of operating conditions.

> **AI CAN HELP A BUSINESS UNDERSTAND HOW IT WOULD PERFORM UNDER A VARIETY OF OPERATING CONDITIONS**

Energy

The energy sector depends heavily on insights gleaned from consumer behavior. However, the unpredictability of human behavior and challenges in acquiring and interpreting real-world consumer data have led to an underutilization of data analytics in decision-making in this sector. Synthetic data can help by generating realistic, anonymized datasets that replicate real-life behavior patterns. This is instrumental in creating consumption profiles for populations and can aid in predictive maintenance. By bridging the gaps in areas where data collection is limited and refining existing datasets, synthetic data enhances the accuracy of predictive models. Moreover, it safeguards privacy by excluding sensitive information, such as demographic details, geographic data, and income levels.

CONCLUSION

EMBRACING SYNTHETIC DATA FOR ENHANCED DECISION-MAKING

THE ROLE OF INFORMATION IN SHAPING BUSINESS STRATEGIES AND FOSTERING INNOVATION CANNOT BE OVERSTATED

In today's data-driven world, the role of information in shaping business strategies and fostering innovation cannot be overstated. As businesses increasingly rely on data for insights and growth, they also face significant challenges, including the availability of large, industry-specific datasets and the need to comply with stringent privacy laws. Here are some key ways in which synthetic data transforms business:

1. **Data augmentation.** Synthetic data can be tailored to specific needs without risking sensitive information exposure, enhancing data availability for decision-making.
2. **AI training.** With increasing adoption of AI across various sectors, the demand for high-quality data to train AI models is more critical than ever. Synthetic data provides a viable solution by offering a diverse, yet privacy-compliant dataset that can help in developing more accurate and robust AI systems.
3. **Scenario simulation.** Synthetic data enables testing of various strategies as well as anticipation of challenges, which is crucial for risk management.
4. **Cost-effectiveness.** Synthetic data reduces expenses related to data collection and processing.

By integrating synthetic data into their operations, companies can enhance decision-making processes and maintain a competitive edge in the market.

Arthur D. Little has been at the forefront of innovation since 1886. We are an acknowledged thought leader in linking strategy, innovation and transformation in technology-intensive and converging industries. We navigate our clients through changing business ecosystems to uncover new growth opportunities. We enable our clients to build innovation capabilities and transform their organizations.

Our consultants have strong practical industry experience combined with excellent knowledge of key trends and dynamics. ADL is present in the most important business centers around the world. We are proud to serve most of the Fortune 1000 companies, in addition to other leading firms and public sector organizations.

For further information, please visit www.adlittle.com.