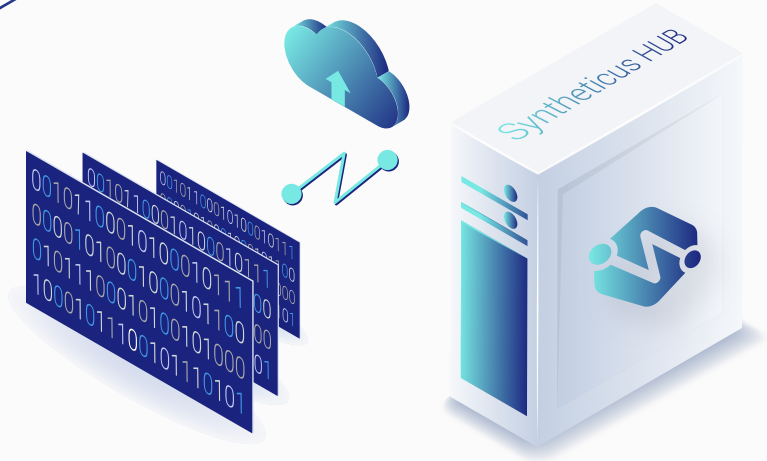
Syntheticus

Whitepaper:

Democratizing Data Access with Synthetic Data



Content

Introduction

Privacy is a growing concern in today's data-driven world. As AI algorithms become more sophisticated and able to process larger and more complex datasets, the potential to infer sensitive information from data is increasing. This has led organizations to look for ways to protect the privacy of their customers and employees.

Apart from privacy issues, companies face the challenge of **data availability**. Collecting, labeling, training, and maintaining large datasets is time-consuming, expensive, and often not feasible. This is particularly true for datasets with rare events or hard-to-label data points, such as medical images or sensor readings.

One of the solutions to modern data challenges is using **AI-generated synthetic data**. Synthetic data is artificially generated, using new and advanced machine learning algorithms, instead of collecting and curating sensitive and risky real-world data.

This whitepaper will guide you through synthetic data's benefits and use cases, how it is generated, and how it helps organizations in various industries democratize access to data and enhance data privacy.

What is Synthetic Data?

In short, synthetic data is artificial data generated to mimic original data while protecting the privacy of real-world data sources. It looks and behaves like the original data—but without any personally identifiable information (PII), secure records, or other sensitive data points.

Since synthetic data does not have one-to-one correlations with real data, it is used in place of real-world data for training machine learning (ML) models, software testing, and analytics. It is also used for filling in gaps when there is not enough real data to work with.

Due to its privacy-preserving properties, synthetic data is particularly beneficial in healthcare, insurance, and financial services—where ethical use of customer data is mandatory.

Benefits of Using Synthetic Data

The most important benefit of synthetic data is not exposing sensitive data of companies and individuals in any way. Highly privacy-concerned organizations, such as healthcare companies and financial institutions, use synthetic data to train their models without violating compliance regulations.

Let's imagine a pharmaceutical company with access to patient data. To use that data for model training, they would need to anonymize the records, and even then, there is a risk of reidentification. By using synthetic data, the company trains its models without touching the real patient data, thus avoiding any potential data breaches.

Synthetic data is also cost-effective. Generating synthetic data requires far less time and money than collecting, labeling, and curating real data. Additionally, the generated data is not impacted by outliers or missing values that may occur in real datasets. It's scalable, reliable, and can be constantly refreshed to reflect the latest trends and dynamics.

Another benefit of synthetic data is eliminating the bureaucratic burden of accessing sensitive data. Even for internal use, companies usually have to go through complex approval processes to justify the need for sensitive data access. With synthetic data, such restrictions don't apply, and companies are free to access data as often as needed without going through any red tape.

In addition, synthetic data is used to fill in the gaps of incomplete datasets. In industries such as healthcare and finance, many datasets do not have enough entries to draw reliable conclusions. With synthetic data, companies create larger datasets with more varied information and insights.

Finally, synthetic data is essential for ExplainableAI and provides insights into how models behave. By running the models on synthetic data, organizations can better understand how the models make decisions and interpret the results more accurately.

Additionally, synthetic data helps organizations comply with ethical AI principles and avoid risks of bias. Syntheticus Hub platform takes part in the Ethical AI vetter database of AI startups providing ethical services, proving its commitment to ethical AI principles.

Synthetic Data Applications

Advanced Analytics

Corporate data is growing in number, size, and complexity. Companies are constantly looking for ways to analyze their data more efficiently while maintaining the privacy and data security. Cloud solution providers (CSPs) offer the most effective data analytics tools, such as Google Analytics, to extract value from data within organizations.

However, organizations must comply with data protection and privacy regulations that limit access to these tools. Synthetic data is ideal for such scenarios, as it provides real-world insights without exposing sensitive information.

Machine Learning/AI

Synthetic data is also used to train and deploy AI/ML algorithms. Instead of data scientists struggling with limited or low-quality datasets when working with machine learning, they can leverage synthetic data to generate large, high-quality datasets to improve the accuracy and performance of their models.

Whether it's used for predictive modeling, forecasting, or financial risk management, synthetic data significantly improves the performance and results of advanced analytics systems. Additionally, it reduces the cost and time associated with data management, storage, and analysis.

Software Development and Testing

Software development and testing require large-scale datasets to ensure the accuracy and reliability of their applications. As companies develop more complex applications, their demand for data increases. Synthetic data helps developers and testers build more realistic test cases and environments, which allows them to evaluate applications better.

Software companies use synthetic data to generate vast amounts of meaningful data to test their applications' performance and scalability. This helps them reduce time-to-market, improve customer experience, and reduce development costs.

Cybersecurity

Organizations need to protect their data from malicious actors, and synthetic data is an effective way to do so. It gives organizations the same insight as original data without exposing sensitive information to cyber threats.

Synthetic data is also used to generate realistic datasets for cyber threat simulation. It helps security teams detect, prevent, and respond to threats and malicious attacks by providing a realistic dataset for training machine learning models.

Advanced Analytics

Synthetic data increases efficiency and accuracy in companies' advanced analytics solutions.
Using Syntheticus' data generation capabilities, organizations create realistic datasets to feed their analytics processes and extract valuable insights without worrying about privacy and security violations.

Machine Learning/AI

The Syntheticus Hub platform allows organizations struggling with limited or low-quality datasets to generate high-quality synthetic data that improves their advanced analytics systems' accuracy, performance, and results. By leveraging synthetic data for their ML/AI algorithms, companies reduce the cost and time associated with data management, storage, and analysis.

Software Development and Testing

Syntheticus' data generation capabilities make it an ideal tool for developers looking to build smarter, more accurate software and applications. With access to high-quality synthetic datasets, you reduce development time, test and refine your code and applications, and improve the accuracy of your software solutions.

Cybersecurity

Syntheticus helps security teams create realistic data to fortify internal security and protect their sensitive digital assets from malicious attacks. Synthetic data retains all the statistical value of original datasets while eliminating identifiable characteristics that make it easy to reverse engineer and misuse, making it the ideal tool for cybersecurity teams.

Many industries already leverage the potential of synthetic data. As we mentioned, **financial institutions** use it to simulate data for fraud detection, financial crime, AML, credit risk, operational risk, and asset management. **Healthcare and pharmaceutical companies** use it to work on clinical trials where real patient data is unavailable, or there is insufficient data to reach meaningful conclusions. **Insurance companies** leverage synthetic data to simulate real-world datasets and improve their predictive capabilities, allowing them to make more accurate risk assessments and price insurance policies more effectively.

Synthetic Data Generation Methods

When talking about generating synthetic data, it's important to note there are several different approaches and techniques that can be used. The most commonly used methods involve **Generative adversarial network (GAN) models, VAE (Variational Autoencoders)**, and **Gaussian Copula (Statistics based)** in combination with **differential privacy**.
Each method has its benefits and drawbacks depending on the use case and data type.

In this paper, we will take a closer look at how GANs and differential privacy (DP) are used to generate synthetic data that is both realistic and privacy-preserving.

Generative Adversarial Networks (GANs)

GANs are a type of machine learning algorithm consisting of two separate neural networks, a generator, and a discriminator. The basic idea of a GAN is that the generator looks for statistical distributions or patterns in a chosen dataset and produces synthetic data with similar distributions or patterns. At the same time, the discriminator looks at both synthetic and real data to identify any differences between them. If it can spot the difference between the originals and the new sample, it sends a signal to the generator to make changes to the generated synthetic data until they become indistinguishable from real data.

"Synthetic data holds the key to unlocking the power of AI while protecting the privacy of sensitive information. With the combination of Generative Adversarial Networks (GANs) and differential privacy, we generate synthetic data that accurately reflects the real data while ensuring that private information remains protected."

Dr. Valerio Mazzone, Co-founder & CTO

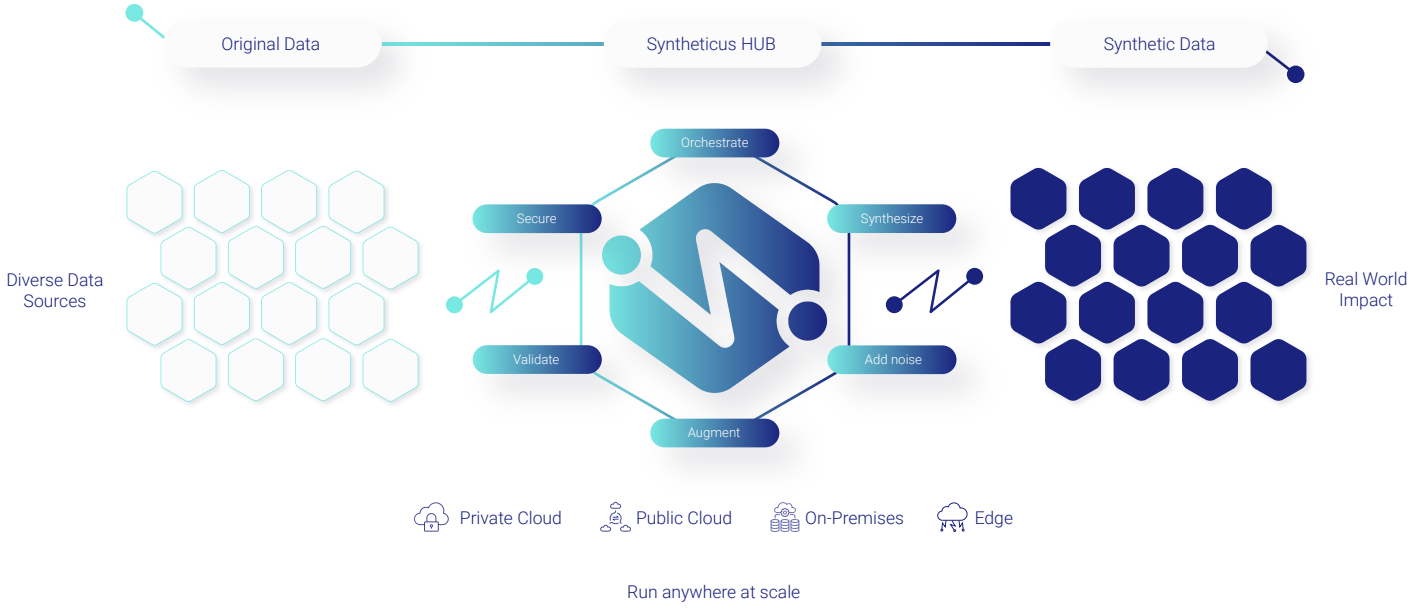
Differential privacy

Differential privacy is a concept in computer science that provides a rigorous definition of privacy for statistical data analysis. It involves adding carefully calibrated noise to the data during the analysis process, which makes it difficult for an attacker to determine whether a particular individual's data is included in the dataset. This ensures that synthetic data generated using differential privacy cannot be reverse-engineered to reveal personal information.

The key idea behind differential privacy is that the amount of noise added to the data should be carefully calibrated to ensure that the privacy of individual data points is protected while still allowing for accurate and meaningful data analysis. This is achieved by using the concept of epsilon-differential privacy, which defines a privacy budget epsilon that determines the maximum amount of noise that can be added to the data.

How it Works: Synthetic Data Generation in Practice

The good news is, you don't need to understand all of the mathematics and computer science behind synthetic data generation for it to be useful in practice. The Syntheticus Hub platform makes synthetic data generation easy and accessible to anyone.



1 The process starts by collecting the **original data** from diverse data sources, which is used as input. The platform is data-agnostic, meaning it can ingest all data types, such as structured (table, relational tables, time series) and unstructured (images, audio, JSON, etc.). The data can be collected either through event-based drag & drop or through connections to all the major databases such as Oracle, MongoDB, etc.

2 After the input is collected, the Syntheticus Hub **orchestrates** a series of proprietary and external algorithms to **synthesize** data that accurately mimics the original data.

3 Differential privacy **adds noise** to the Generative AI algorithms while they're running to ensure generated synthetic data is protected against reidentification attacks.

4 **Data augmentation** and enrichment techniques add additional variability to the synthetic data, tackling data scarcity challenges and making it even more realistic and harder to reverse-engineer.

5 The platform **validates** the generated data to ensure it looks and behaves like the original data and meets the quality requirements.

6 Syntheticus Hub **secures** the data, algorithm, and model orchestration by applying Confidential Computing techniques leveraging cutting-edge solutions from our global strategic partner Cysec.

7 Finally, the synthetic data is ready for **real-world impact**. It can be shared with other applications, systems, and users while staying compliant with privacy regulations.

Evaluating the Quality of Synthetic Data

One of the main challenges in using synthetic data is determining its quality.
Many factors affect the quality of synthetic data, including the dataset size, the number of variables included, and how well it mimics real data.

To evaluate the quality, synthetic data is measured against three key dimensions: **fidelity**, **utility**, and **privacy**.

Syntheticus Hub deploys various utility and privacy metrics to evaluate the quality of synthetic data. We can also easily customize our platform by adding custom external metrics into the flow. Our platform then orchestrates all the metrics accordingly to ensure accuracy. With our comprehensive suite of metrics, we are able to evaluate the quality of synthetic data in multiple dimensions with improved accuracy and speed.

There's no global standard yet to determine the appropriate quality regarding privacy, utility, or fidelity. The quality needs to be assessed on an individual use case basis.

However, the IEEE Standards Association has set up an IC Expert Group **to set a standard for structured privacy-preserving synthetic data.**

Syntheticus takes part in this expert group, with our CEO, Aldo Lamberti, elected as Vice Chair.

IEEE SA STANDARDS ASSOCIATION

Privacy and Security Considerations of Using Synthetic Data

Synthetic data solves many problems: it can help organizations protect their data, generate large datasets, and reduce bias. But is it 100% safe and secure in every situation?

No data is 100% safe and secure, but synthetic data can provide a level of privacy and security that is hard to achieve with real data. The key is ensuring that the generated synthetic data does not contain any sensitive information or private characteristics of individuals and has been properly anonymized and de-identified.

So what are the underlying privacy risks that follow synthetic data?

To determine realistic parameters and generate synthetic data, you will need to process at least some amount of real data. This process itself carries some risk, as there is always a chance of data leakage while transferring and processing information. Where that real data comes from and how it flows should be carefully monitored. If it consists of any sensitive information about individuals, the data analytics team must process it in compliance with data protection laws.

In other words, the GDPR may still apply to the researchers' activities when producing synthetic datasets.

In some cases, further alteration of the synthetic data may be necessary to ensure it is truly anonymized and de-identified. For example, if the real data contains an unusual individual with a rare medical condition or a specific combination of traits, and synthetic data shows a similar individual to remain statistically valid, there is a risk that the anonymity of the dataset may be breached.

Therefore, organizations should pay special attention to privacy when using synthetic data. Recognizing the potential risks and ensuring all necessary steps have been taken to protect the data is just as important as evaluating its quality and accuracy.

Finally, organizations should consider their own security infrastructure when generating or utilizing synthetic data. It is important to ensure that the data is stored in a secure environment with proper access control measures and that all related systems have been properly configured to protect the data from malicious attacks.

The Road Ahead

Synthetic data governance is quickly becoming a critical part of data-driven organizations.

Research firm Gartner projects that synthetic data will completely overshadow real data in AI models by 2030. It named "Synthetic Data" and "Differential Privacy" as one of its Top Strategic Technology Trends that will, according to their widely referenced study, replace 60% of the data used for the development of AI and analytics projects by 2024.

Forbes named 'Synthetic Data' as one of The 5 Biggest Data Science Trends in the previous year, further highlighting its growing importance.

According to the European Data Protection Supervisor (EDPS), "Synthetic data is a technical solution to a legal problem," enhancing technology privacy, mitigating bias, and democratizing access to data. In their regularly published TechSonar report on emerging technologies, EDPS mentions synthetic data as one of the most promising technologies worth monitoring.

As the use of synthetic data continues to grow, its impact will only become more pronounced in the coming years. With its potential to democratize access to data while minimizing risk, synthetic data will transform the economics of data. It will undercut the strength of proprietary datasets and create new opportunities for businesses to become more data-driven.

While tech giants like Google, Facebook, and Amazon have achieved market dominance by collecting and leveraging vast amounts of real data, synthetic data — with its ability to provide deep insights while limiting privacy risks — will eventually level the competitive playing field. By **democratizing access to data at scale**, synthetic data will open up new possibilities for smaller businesses to compete with their much larger counterparts.

Additionally, synthetic data will threaten the livelihoods of companies specializing in the massive data labeling industry, which has become the lifeblood of machine learning in recent years. It will also open up an entirely new market, providing businesses with a much-needed alternative to costly, time-consuming data preparation.

The implications of synthetic data are far-reaching, and its potential to revolutionize the way we use data is undeniable. As businesses around the world continue to invest in synthetic data, its impact will become increasingly visible — and undeniable.

References and Recommended Literature

Dankar, F., Ibrahim, M., 2021, Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation, https://www.mdpi.com/2076-3417/11/5/2158

Emam, K., 2020, Accelerating AI with Synthetic Data, https://www.oreilly.com/library/view/accelerating-ai-with/9781492045991/

Marr, B., 2021, The 5 Biggest Data Science Trends In 2022, https://www.forbes.com/sites/bernardmarr/2021/10/04/the-5-biggest-data-science-trends-in-2022/?sh=5ce7b25740d3

Nikolenko, I., 2021, Synthetic Data for Deep Learning

Syntheticus, 2022, Everything You Need to Know About Synthetic Data, https://syntheticus.ai/guide-everything-you-need-to-know-about-synthetic-data

Syntheticus, 2022, The benefits of using synthetic data in cybersecurity, https://syntheticus.ai/blog/the-benefits-of-using-synthetic-data-in-cybersecurity

TechSonar Report, 2021, https://edps.europa.eu/system/files/2021-12/techsonar_2021-2022_report_en.pd

Syntheticus

Ready to explore the potential of synthetic data?

Sign up for a free demo

and learn how Syntheticus advances your data-driven projects and helps your business stay secure, compliant, and competitive.