Syntheticus®

SIX

Unlocking New Possibilities in AI-Driven Projects with Synthetic Data:

The Story of SIX

Content

What Is Synthetic Data? 2

How Is Synthetic Data Generated? 3

How Synthetic Data Changes AI and LLM Projects 4

Case Study: Deep Dive into the SIX Use Case 5

Meet SIX 5

The Challenge 6

How Synthetic Data Benefits SIX 7

Deep Dive into the Proof of Concept 8

The Results 10

Conclusion 12

When working on Artificial Intelligence (AI) projects, having access to a large amount of data is essential for making informed decisions. But oftentimes, the data available is limited, biased, or outdated, making it difficult to extract useful insights.
That's where synthetic data comes in.

Synthetic data is a type of artificial data created by computer algorithms to replicate the patterns and characteristics of real-world data. Think of it as a "digital sibling" of your data but with more flexibility and diversity. By using synthetic data, you enhance or even replace real-world data to achieve more accurate and comprehensive results. And this isn't just a theoretical concept.

Synthetic data has already proven valuable in various use cases, such as analytics, test data management, and AI. In this case study, we will explore how SIX, a leading financial services provider in Switzerland, leveraged synthetic data to enhance their projects, gain data-driven insights, and create valuable business outcomes.

What Is Synthetic Data?

At its core, synthetic data is a computer-generated representation of real-world data. But unlike real-world data, which is collected from various sources such as sensors, surveys, or databases, synthetic data is created from scratch based on predefined rules or statistical models. This gives it distinct advantages over traditional data sources, such as flexibility, scalability, and privacy.

Synthetic data aims to provide additional data points similar to real-world data, allowing for more comprehensive analysis and modeling. It supplements existing data sets or creates new ones, enabling a more detailed look into specific trends and patterns.

Gartner

Gartner named "Synthetic Data" and "Differential Privacy" as one of its Top Strategic Technology Trends and estimates that 60% of large enterprises will be leveraging one or more of these techniques by 2025.

Forbes

Forbes named 'Synthetic Data' as one of the Top 10 transformative CV Trends in 2024, further highlighting its growing importance.

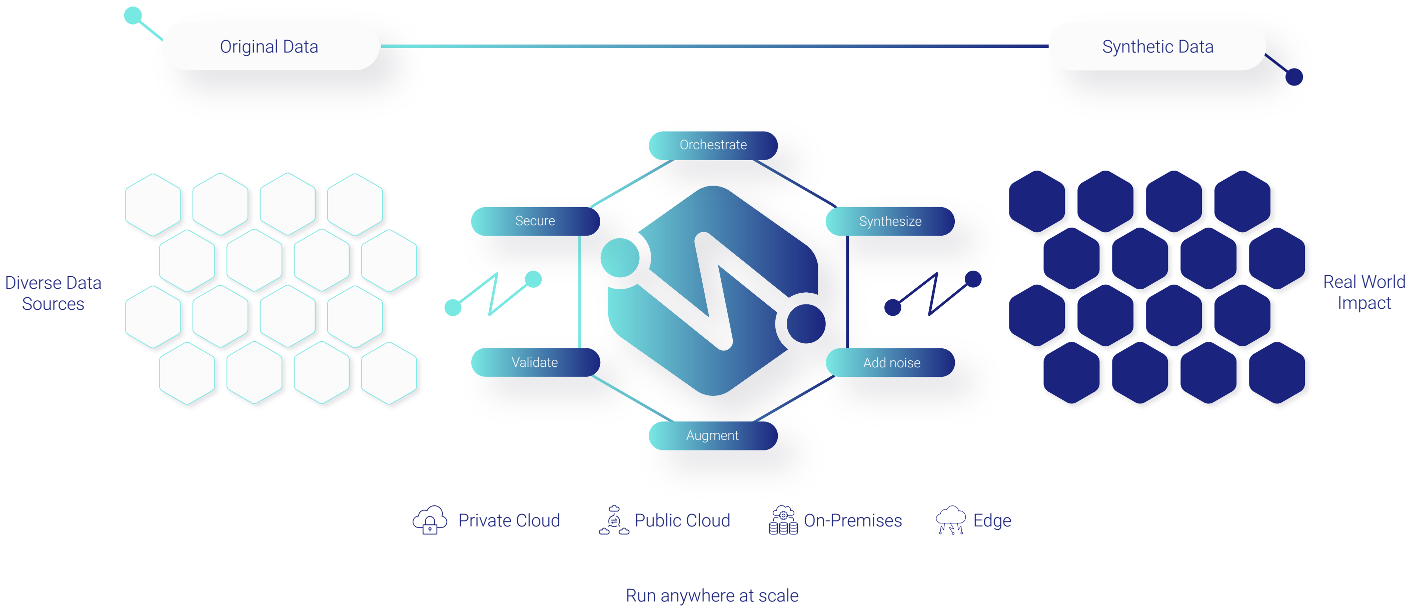
Since synthetic data isn't considered "Personal Identifiable Information (PII)," according to privacy regulations such as GDPR, all the data can be safely used and collaborated on without worrying about privacy breaches or compliance.

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.

Since synthetic data isn't considered "Personal Identifiable Information (PII)," according to privacy regulations such as GDPR, all the data can be safely used and collaborated on without worrying about privacy breaches or compliance.

How is Synthetic Data Generated?

Synthetic data is created artificially using Generative AI techniques, such as Generative Adversarial Networks (GANs), Large Language Models,and Variational Autoencoders (VAEs) in combination with Differential Privacy. With these techniques, Syntheticus® creates synthetic data that realistically mimics the original data while preserving privacy.



Once the original data is collected, the platform orchestrates the proprietary machine learning algorithms to synthesize, secure, validate, augment, and enrich it. This results in synthetic data that looks and behaves like the original one without revealing PII or risking compliance issues.

How Synthetic Data Changes AI and LLM Projects

AI projects often face significant challenges due to data scarcity and access limitations. With the EU AI Act and its new regulations for AI system operations, data availability is an even bigger obstacle. Conventional privacy protection techniques like pseudonymization or legacy anonymization, such as data masking, are not GDPR compliant, leading companies to turn towards synthetic data.

Synthetic data overcomes these challenges by providing high-quality, compliant data used to train and test AI models with minimal bias or risk of non-compliance. Let's dive deeper into a few ways synthetic data changes AI-driven projects:

Improved AI Model Performance

Organizations often struggle with the pressing need for highly accurate predictive models in their machine learning projects. Synthetic data delivers high-quality datasets that enhance testing and refinement, ultimately reducing bias and improving model accuracy, leading to more reliable outcomes.

Unrestricted Access to Critical Data

Informed decision-making is at the core of all AI and machine learning projects. However, data silos and ethical considerations obstruct access to vital insights, resulting in missed opportunities. Synthetic data provides a cost-effective and secure solution, granting organizations unrestricted access to essential data.

Legal Compliance

In a world of increasing regulations, GDPR takes center stage. Stricter requirements for data fairness and increased scrutiny driven by regulations such as GDPR and the EU AI Act demand a responsible approach to AI and machine learning. Synthetic data offers a robust solution by providing GDPR-compliant datasets, mitigating risk, and ensuring legal complianc.

Reduced Storage Costs

Traditional data storage requires large amounts of physical space, as well as resources for maintenance and security. Synthetic data offers a solution by reducing the need for storing and managing large datasets, freeing up valuable resources that can be allocated to other critical areas of business operations.

Enhanced Data Fairness

Data fairness refers to the unbiased representation of all segments of a population. However, traditional methods often fail to achieve data fairness due to bias in underlying datasets. Synthetic data plays an important role in mitigating bias and ensuring model fairness by providing diversified datasets essential for model training and testing, while adhering to ethical ML standards.

Simplified Collaboration for Innovation

Synthetic AI data eliminates the complexities associated with sharing sensitive, personal, or classified data by providing an alternative that holds the same statistical value but does not violate privacy or security concerns.It allows teams to work together on AI and machine learning projects without the barriers traditionally associated with data privacy.

Transparency and Trust with Explainable AI

Synthetic AI data eliminates the complexities associated with sharing sensitive, personal, or classified data by providing an alternative that holds the same statistical value but does not violate privacy or security concerns.It allows teams to work together on AI and machine learning projects without the barriers traditionally associated with data privacy.

Case Study: Deep Dive into a SIX Use Case

Meet SIX

SIX is a leading global financial company providing services related to transaction security, financial information processing, payment transactions, and building a digital infrastructure. SIX has stood for innovation and stability in the global financial markets since 2008 and operates the infrastructure for the financial centers in Switzerland and Spain, thus ensuring the flow of information and money between financial market players.

SIX offers exchange services, financial information, and banking services, aiming to increase efficiency, quality, and innovative capacity along the entire value chain. The company is owned by around 120 national and international financial institutions. SIX's close relationship with them guarantees the financial infrastructure and processes stability, proximity to clients' evolving business needs, and competitive prices.

The Challenge

To unlock the full potential of their data, drive data-driven insights for their AI projects, and create business value, SIX employees need constant access to a wide range of data from every corner of the company.

However, data scientists at SIX struggle to have constant access to the most up-to-date original data due to complex and evolving regulations that limit access to specific datasets, thus hindering their ability to unlock its full potential. This results in an inability to deliver data-driven insights on time and lost opportunities for the company.

The barriers that lead to the challenges mentioned before are, amongst others:

Legal & Compliance

Data is subject to complex and evolving legislation, making it difficult to unlock and use for AI-driven projects.

Privacy regulations

Regulations often impose strict limits on the ability to share data and the extent to which it can be used, further complicating the process of leveraging data for AI initiatives.

IT silos

Data from different business units is often stored in disconnected silos, making it difficult to access and use for AI analysis.

Adding to these challenges, analyses are often conducted locally rather than on the cloud, which restricts the speed and scalability of data-driven activities at SIX. This limitation in computational performance results in inefficient data analytics.

How Synthetic Data Benefits SIX

To overcome these challenges, **SIX joined forces with Syntheticus®**, a powerful synthetic data platform that leverages advanced Privacy-Enhancing Technologies, such as Generative AI and Differential Privacy, to generate realistic and accurate synthetic data.

Gartner

According to Gartner, 80% of traditional financial services companies may struggle to compete effectively by 2030. To avoid this fate, banks must develop a strategy to stay ahead of the curve, and synthetic financial data is the key to achieving this goal.

Here are some of the key benefits that synthetic data can offer to your organization:

Improved Data Quality and Diversity

Synthetic data enriches AI models by offering diverse scenarios, enhancing prediction accuracy.

Enhanced Scalability

Synthetic data supports unlimited AI model growth, overcoming limitations of traditional

Improved Access to Critical Data

Synthetic data breaks data silos, providing secure access to vital insights, fostering revenue opportunities.

Enhanced Collaboration and Knowledge Sharing

Privacy-preserving synthetic data encourages inter-organizational collaboration and knowledge sharing.

Improved Risk Management

Synthetic data allows risk simulation, minimizing losses, and reducing resources needed for model development.

Better Innovation and Experimentation

Synthetic data enables safe experimentation with AI-driven ideas, products, and services.

Improved Regulatory Compliance

Synthetic data ensures regulatory compliance while training AI models and safeguarding sensitive data.

Deep Dive into the Proof of Concept

To better understand the use cases for synthetic data and its potential impact on SIX, Syntheticus® conducted a Proof of Concept (POC) as a part of project collaboration with Constructor Learning and SIX.

The POC is divided into two parts: the first part highlights the original data and the generated differentially private synthetic data. In the second part, the original and synthetic data are compared and evaluated according to specific metrics.

Part One: Data

The underlying financial transaction data consists of customers' personal and financial information. The data scientist's task is to thoroughly understand the customer's loan default risk to better predict future loans.

The original data consists of 100k rows and 11 columns. However, this data is sensitive and contains sensitive information such as the customer's name, address, and bank account number. As a result, the original data cannot be shared outside or within the organization. This poses a challenge for financial institutions looking to collaborate with external partners and data scientists who can't gain full access to the real-world data they need to perform their analysis.

To overcome this problem, Syntheticus® was used to generate synthetic data with almost identical statistical properties as the original data but without containing any PII. We truncated the original dataset to 5k rows for the sake of simplicity. The generated differentially private synthetic data consists of 5k rows and 11 columns.

Input data

100k rows, 11 columns
Contains sensitive customer information

Output data

5k rows, 11 columns
Matching statistical properties, without PII

Part Two: Evaluation

The evaluation is based on two different methods - **utility and privacy.**

Utility evaluation was based on comparing the models built on the original and synthetic data. We used standard classification algorithms to compare the performance of both datasets. The results show four different metrics as well as the final score:

Basic statistics	0.9974
Correlation column correlations	0.9218
Mean Correlation between synthetic and real columns	0.8903
1 - MAPE Estimator results	0.9467
Similarity Score	**0.9390**

The similarity score aggregates different evaluation metrics to cover all aspects of data. It is calculated by taking the mean of these different metrics. The goal is to get a single value representing a synthetic dataset's proximity to a real dataset.

Using the **privacy evaluation**, the results show two metrics:

Duplicate rows between sets (real/synthetic)	0/0
Nearest neighbor mean	2.7591
Nearest neighbor std	0.5285

Two metrics were used to assess the level of privacy. The first is a simple analysis of whether any rows in the synthetic dataset are identical to corresponding rows in the real dataset. Generally, this is not desired, and possible regularisation or checks might be required to prevent it. The second metric is the mean and standard deviation distance between each synthetic record and the most similar real record. The desired outcome is a high mean with a low standard deviation, showing that all synthetic records have a sufficiently large distance from their closest real record while still forming a valid dataset.

The Results

The results of the POC have shown that the synthetic data generated by Syntheticus® is of high quality in terms of privacy and utility: it preserves the original statistical structure while protecting sensitive data. This opens up new opportunities for SIX, allowing their analytics team to leverage synthetic data for advanced analytics and collaborate in a secure and privacy-preserving manner.

Using synthetic data generated by Syntheticus®, SIX's data analytics team collaborates and computes at scale while being 100% compliant with applicable regulations.
Additionally, synthetic data enables cloud infrastructure usage and ensures data security while allowing the execution of complex analysis and predictive models. This results in speeded-up development cycles, a higher degree of scalability, increased overall data literacy, and increased efficiency throughout the organization.

At SIX, synthetic data is used to:

1. Populate data warehouses or sandboxes
2. Run all sorts of analytics like Machine or Deep Learning
3. Serve as test data for product development
4. Share internally or externally for secure collaboration

Using Syntheticus® to seamlessly generate high-quality datasets, SIX was able to run predictive models, testing, and other analytical tasks while ensuring consistent results and reducing overall time to data.

Overall, the promising results of the SIX POC have shown that other AI-driven companies should leverage Syntheticus® to create secure and accurate synthetic data that will allow them to:

Securely collaborate and compute on synthetic data at scale to derive data-driven insights.

Stay compliant and de-risk potential privacy fines and reputation damages.

Reduce wait times and administrative burdens, leading to a positive impact on data-driven attitude.

Harness the power of robust cloud infrastructure for efficient AI model development and deployment.

Democratize data within their organization, leading to an increase in overall data literacy, business value, and a heightened competitive edge in the AI landscape.

Conclusion

The financial services industry is undergoing a significant transformation driven by advances in technology and the growing importance of data. The future of banking will be shaped by the ability to use data-driven decision-making and successful digital transformation. One of the biggest challenges in this context is handling bank-specific and personal data and its processing by artificial intelligence.

Synthetic financial data is likely to become even more important for institutions in the years to come. The increased availability of synthetic data will allow them to leverage AI and machine learning to make more informed decisions while complying with data privacy and regulatory requirements. As new technologies like blockchain and distributed ledgers become widespread, the potential applications for synthetic data will only grow.

Synthetic financial data will become an indispensable tool for financial institutions in areas such as risk management, compliance, fraud prevention, and customer segmentation. As synthetic financial data solutions continue to develop, the possibilities are endless. By planning thoughtfully, taking ethical considerations into account, and understanding its potential uses, financial institutions can take advantage of this opportunity and create a data-driven future in banking.

Syntheticus®

Ready to explore the power of synthetic data for your AI projects?

Sign up for a free demo

and learn how Syntheticus® advances your AI-driven projects while protecting customer privacy.

syntheticus.ai