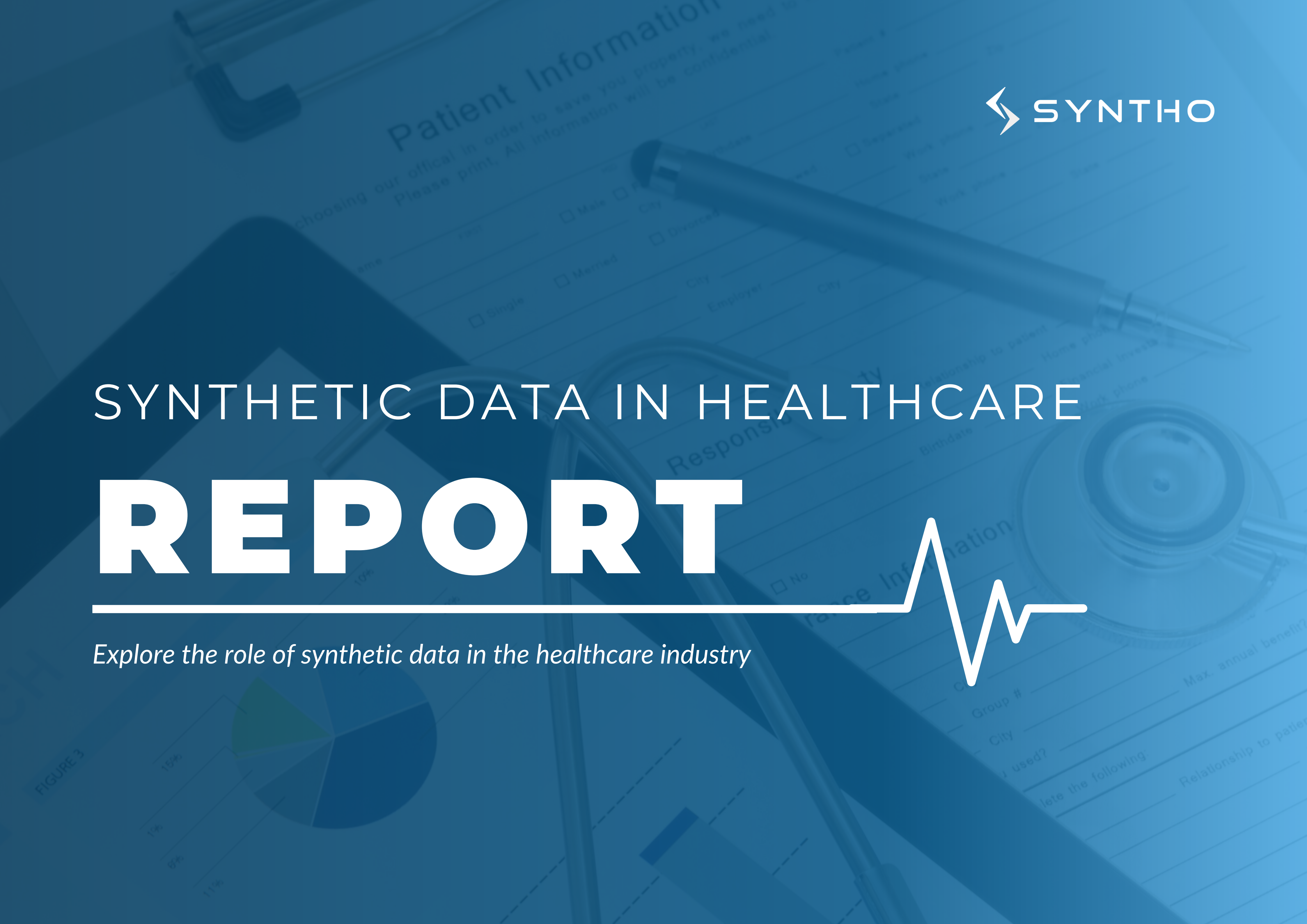


SYNTHO

SYNTHETIC DATA IN HEALTHCARE

REPORT

Explore the role of synthetic data in the healthcare industry

INTRODUCTION

Healthcare organizations rely heavily on data to make evidence-based medical decisions, tailor treatments to individual patients, and drive medical research, all of which contribute to better patient outcomes, operational efficiency, and advancements in medical knowledge and technology.

Synthetic data offers a valuable solution for healthcare organizations, ensuring data privacy while still enabling the generation of realistic and non-sensitive datasets. This empowers researchers, clinicians, and data scientists to innovate, validate algorithms, and conduct analysis without compromising patient privacy, especially crucial in the phase of rising data breaches that have plagued the healthcare sector in recent years.

Given the complex and constantly changing **regulatory landscape** in healthcare, artificial intelligence has become increasingly appealing to these organizations as a means to overcome these challenges effectively.

$67.4bn

expected AI Healthcare market value by 2027

60%

consumers lack sufficient access to patient data

95%

identify theft cases specifically target health records

60%

healthcare IT will use AI for automation and decision-making by 2024

HEALTHCARE DATA ACCESS AND CHALLENGES

Healthcare organizations' data usage is important as it enables evidence-based medical decisions, personalized treatments, and medical research, ultimately leading to enhanced patient outcomes, improved operational efficiency, and advancements in medical knowledge and technologies. Innovative data solutions can significantly benefit healthcare organizations by providing privacy-preserving alternatives and empowering researchers, clinicians, and data scientists to innovate, validate algorithms, and conduct analysis without compromising patient privacy.

HOSPITALS

- Improve patient care
- Leverage data for predictive analytics capabilities
- Protect Personal Health Information (PHI) from the Electric Health Record System (EHR, MHR)
- Lack of data and an unbalanced dataset

PHARMA & LIFE SCIENCES

- Share data and collaborate efficiently with health systems, payers, and related institutions to solve bigger problems
- Overcome data silos
- Perform studies and clinical trials to understand the drug product's impact (efficacy) on a new disease
- Complete a full analysis quickly

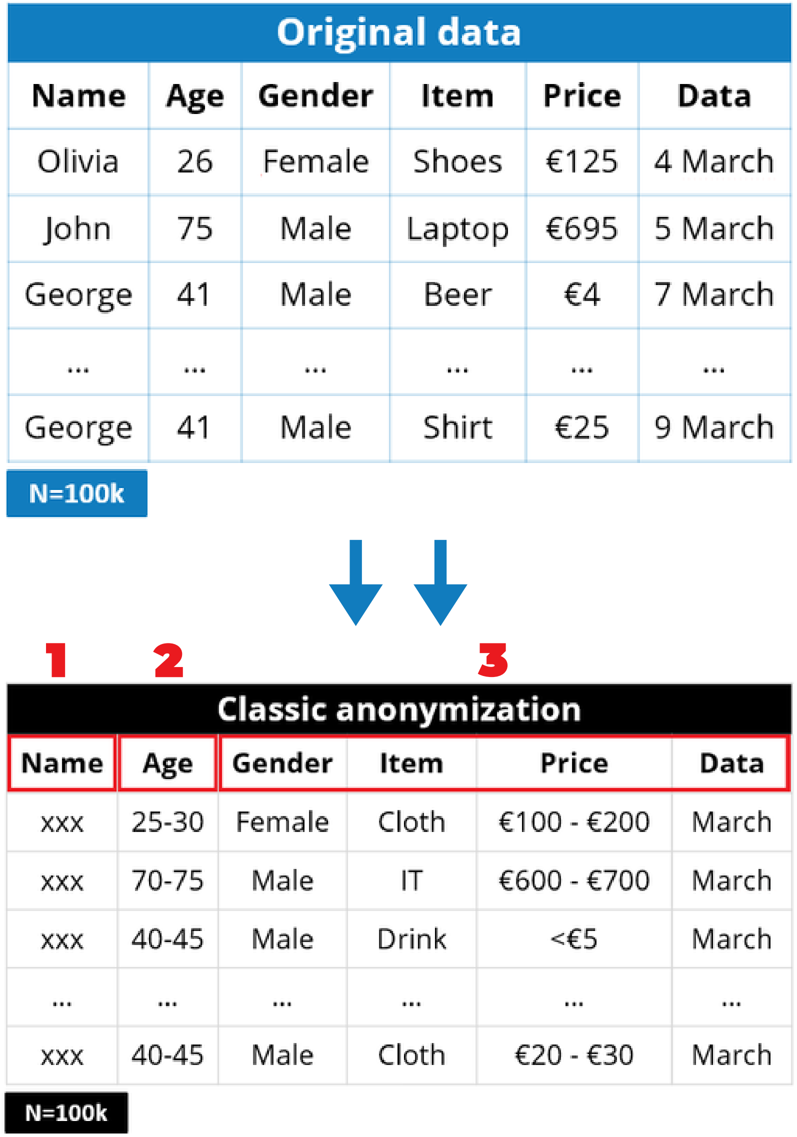
ACADEMIC & CLINICAL RESEARCH

- Accelerate the pace of data-driven research by providing the ability to access data faster and easier
- Solution for generating and sharing data in support of precision healthcare
- Check project feasibility before submitting for original data access

CLASSIC 'ANONYMIZATION'

In order to address privacy concerns in datasets or databases, one typically applies classic 'anonymization' techniques. These ones have one thing in common, they manipulate original data to hinder tracing back individuals.

1. It starts with **deleting** the direct personal identifiers, such as names.
2. Then the indirect information will be **aggregated**, like age.
3. And one will continue to **manipulate** the data.



Original data

Name	Age	Gender	Item	Price	Data
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
...	...	...	...	...	...
George	41	Male	Shirt	€25	9 March

N=100k

Classic anonymization

Name	Age	Gender	Item	Price	Data
xxx	25-30	Female	Cloth	€100 - €200	March
xxx	70-75	Male	IT	€600 - €700	March
xxx	40-45	Male	Drink	<€5	March
...	...	...	...	...	...
xxx	40-45	Male	Cloth	€20 - €30	March

N=100k

Classic 'anonymization' is not a solution, because of:

- **Privacy risk** - you will always have a privacy risk. Applying those classic anonymization techniques makes it harder, but not impossible to identify individuals.
- **Destroying data** - the more you anonymize, the better you protect your privacy, but the more you destroy your data. This is not what you want for analytics, because destroyed data will result in bad insights.
- **Time-consuming** - it is a solution that takes a lot of time because those techniques work differently per dataset and per datatype.

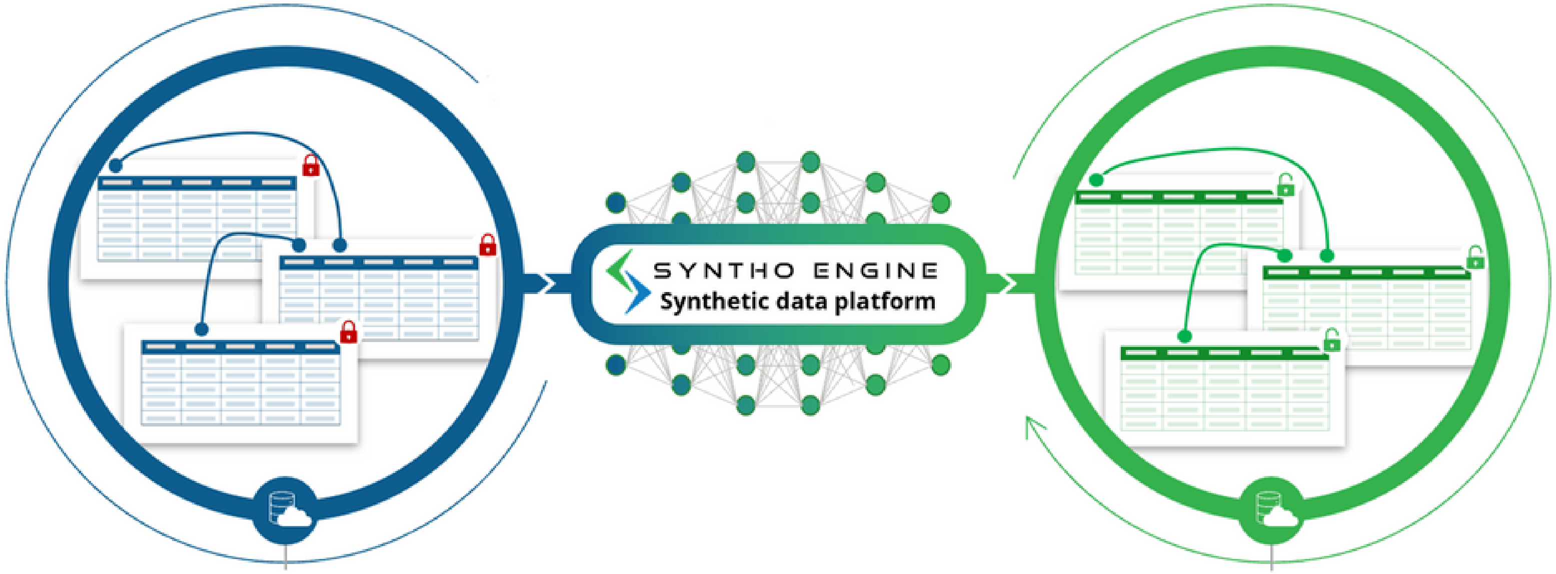
WHAT IS SYNTHETIC DATA?

Organizations dealing with highly **sensitive data** often encounter challenges in utilizing and sharing this information with stakeholders. Due to the sensitive nature of the data, traditional usage and sharing methods are not viable options. Consequently, these organizations miss out on data-driven innovation opportunities and the ability to harness the full potential of their data.

As **Syntho aims to solve the global** privacy dilemma, we are actively shaping the future of data privacy through utilizing and sharing *AI-generated synthetic data, which is artificially generated information that mimics real-world data patterns and characteristics without containing any actual, sensitive information.* This unlocks significant benefits for these organizations, including reduced risk, increased data availability, and faster access to data.

Privacy by design is a key **driver for business success** because it:

- *Gains digital trust*
- *Boosts data and insights*
- *Facilitates industry collaborations*
- *Realizes speed and agility*



TYPES OF SYNTHETIC DATA

Three types of synthetic data do exist within the synthetic data umbrella.

Dummy data / mock data

Dummy data is randomly generated data (e.g. by a mock data generator). Consequently, characteristics, relationships, and statistical patterns that are in the original data are not preserved, captured, and reproduced in the generated dummy data.

Rule-based generated synthetic data

Rule-based generated synthetic data is synthetic data generated by a pre-defined set of rules. Examples of those pre-defined rules could be that you would like to have synthetic data with a certain minimum value, maximum value, or average value.

Synthetic data generated by artificial intelligence (AI)

As you expect from the name, synthetic data generated by artificial intelligence (AI) is synthetic data generated by an artificial intelligence (AI) algorithm. The AI model is trained on the original data to learn all characteristics, relationships, and statistical patterns. Thereafter, this AI algorithm is able to generate completely new data points and models those new data points in such a way that it reproduces the characteristics, relationships, and statistical patterns from the original dataset. This is what we call a synthetic data twin.

WHY DO HEALTH ORGANIZATIONS CONSIDER SYNTHETIC DATA?

- **Privacy-sensitive data.** Health data is the most privacy-sensitive data with even stricter (privacy) regulations.
- **Urge to innovate with data.** Data is a key resource for health innovation, as the health vertical is understaffed, and over-pressured with the potential to save lives.
- **Data quality.** Anonymization techniques destroy data quality, while data accuracy is crucial in health (e.g. for academic research and clinical trials).
- **Data exchange.** The potential of data as a result of collaborative data exchange between health organizations, health systems, drug developers, and researchers is enormous
- **Reduce costs.** Healthcare organizations are under extreme pressure to reduce costs. This could be realized via analytics, for which data is needed.

WHY TO CHOOSE SYNTHO?

Syntho's platform places healthcare organizations at the forefront of data transformation. It seamlessly handles time series and event data, crucial elements often found in healthcare data, ensuring comprehensive support for the unique intricacies of the industry. With extensive experience and adaptability, Syntho accommodates a wide array of healthcare data types, including EHRs, MHRs, surveys, clinical trials, claims, patient registries, and more.

What sets Syntho apart is its forward-looking approach, with a product roadmap finely tuned to the needs and goals of leading health organizations in the **United States, Japan,** and **Europe**.

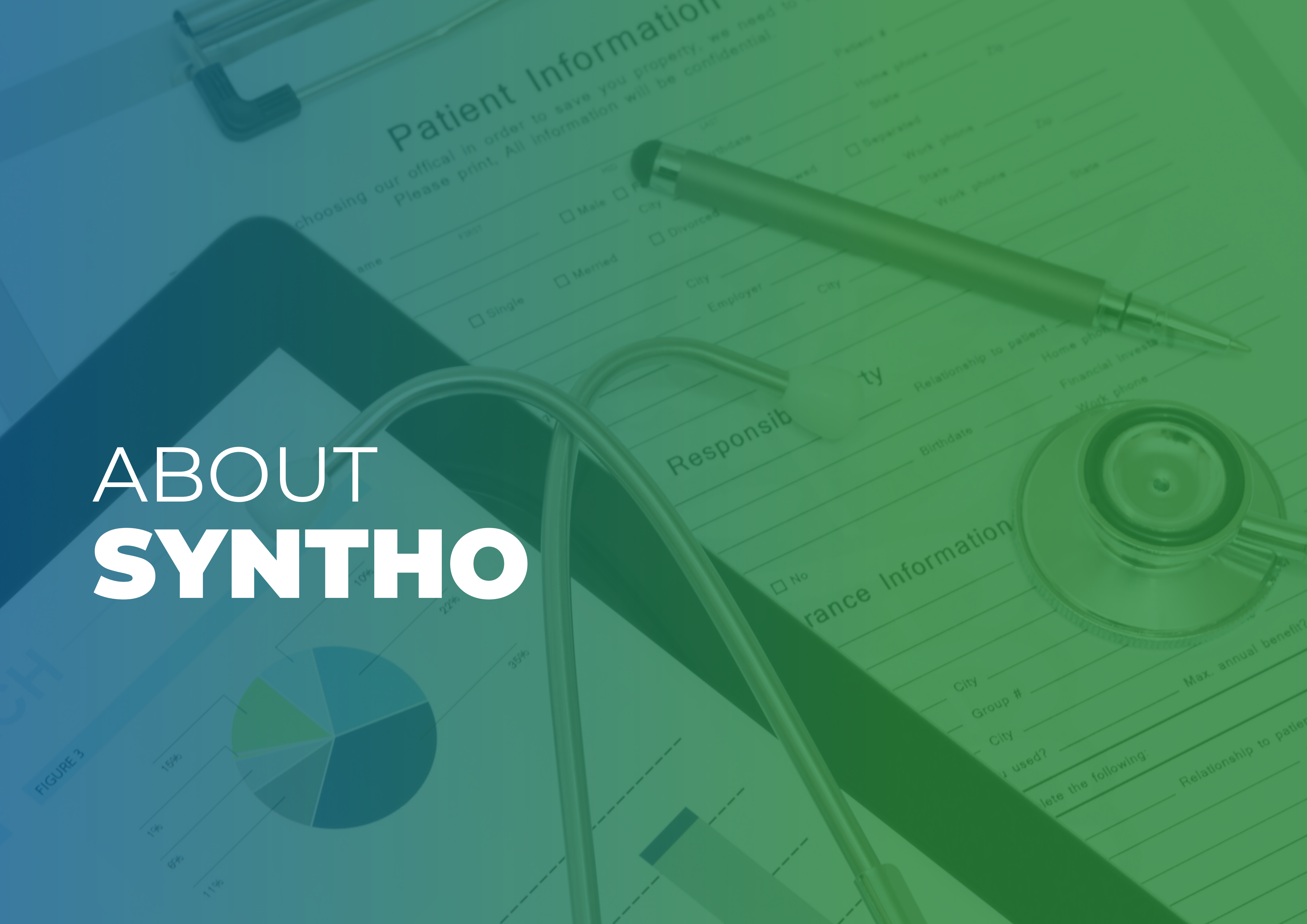
When you choose Syntho, you're not just accessing cutting-edge technology; you're embracing a partnership that drives innovation and excellence in healthcare data management.

Questions?

Do you want to discover more about of synthetic data? Request our Syntho Guide or contact us

Request Guide

Contact us



ABOUT SYNTHO

ABOUT SYNTHO

Founded in 2020, **Syntho** is an Amsterdam-based startup that is revolutionizing the healthcare industry by using AI-generated synthetic data.

As a **leading provider of synthetic data** software, Syntho's mission is to empower healthcare organizations worldwide by enabling them to generate and leverage high-quality synthetic data on a large scale. Our innovative solutions are accelerating the data revolution, unlocking privacy-sensitive data, and significantly reducing the time required to access relevant and sensitive information. With these advancements, healthcare data can be freely shared and utilized without compromising privacy.

Accordingly, Syntho has been honored with prestigious **Awards** like:

Winner of
Philips Innovation Award

Winner of global **SAS Hackathon** for Healthcare & Life Sciences

Shortlisted by **NVIDIA** as Generative AI startup to watch

Winner of
gender bias challenge of **Unesco**

ORGANIZATIONS WE WORK WITH

Recognitions

NVIDIA

sas

PHILIPS

Innovation Award

UNESCO

Client References

Cedars Sinai

Erasmus MC
University Medical Center Rotterdam

Centraal Bureau voor de Statistiek

LUMC
Leiden University Medical Center

Erasmus Universiteit Rotterdam

lifelines

QUALITY EVALUATION OF THE SYNTHETIC DATA

1. Syntho's Quality Assurance (QA) Report

We provide a comprehensive quality assurance report for every synthetic data run, that demonstrates the accuracy of the synthetic data compared to the original data.

2. External assessment by the data experts of SAS

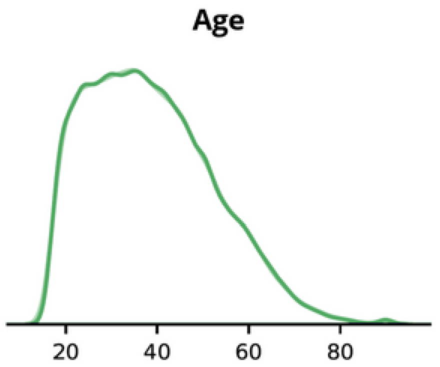
The data experts from SAS approved our AI-generated synthetic data for the use of model development as part of a real case study for a telecom customer.

Our synthetic data is approved by the data experts of SAS

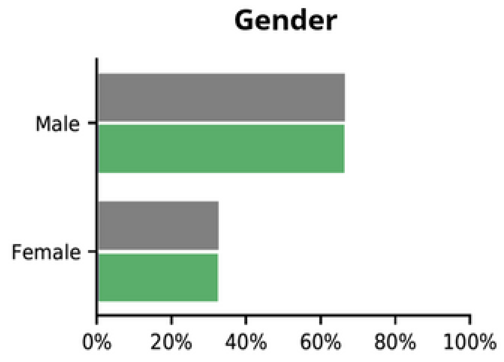
QUALITY EVALUATION OF THE SYNTHETIC DATA

We provide tangible evidence of data quality through our data quality report, which presents a clear visual comparison between the original data (depicted in grey) and the synthetic data (highlighted in green).

- The **distributions**, the frequency of variables in the dataset, are similar.

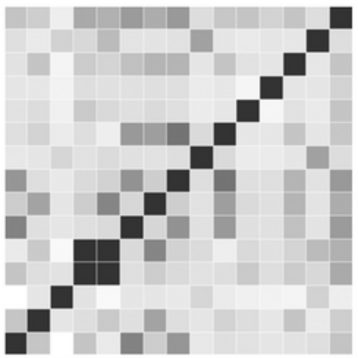


Original data

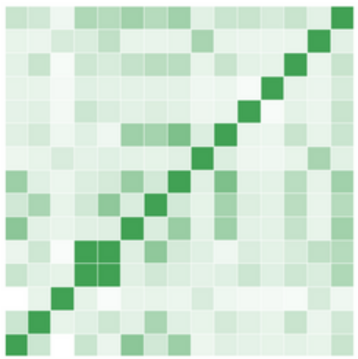


Synthetic data

- The **correlations**, the relationship between variables, are also similar.



Original data

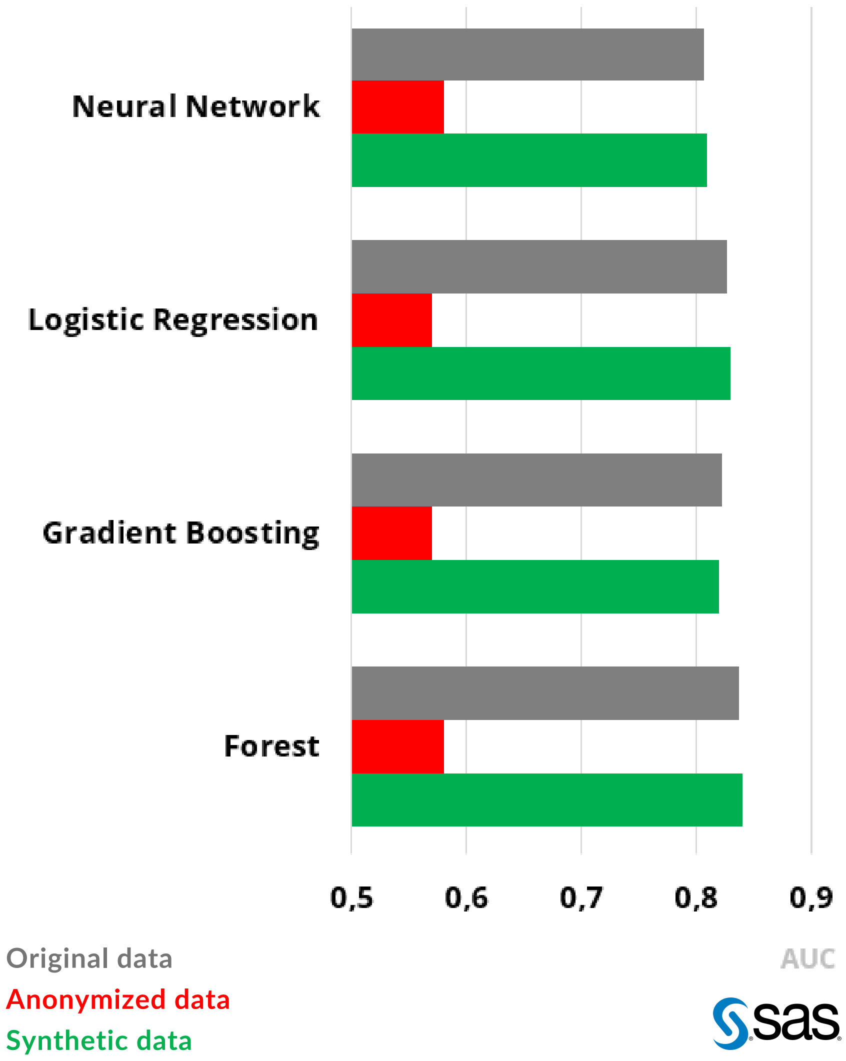


Synthetic data

To get more insights between the original and synthetic data, contact us to request our **Quality Report**:

Request the report

THE SAS DATA EXPERTS APPROVED OUR AI GENERATED SYNTHETIC DATA



In partnership with **SAS**, we have conducted a synthetic data evaluation.

We did the assessment of synthetic data to do the churn prediction for the Telecom company. During the assessment, we used 4 machine learning models: *neural network*, *logistic regression*, *gradient boosting*, and *the random forest*. These models were utilized, with the area under the curve serving as a performance indicator for machine learning accuracy.

These are the results and conclusions from this assessment:

- **Synthetic data** demonstrates **comparable** performance in comparison to the **original data,** indicating its effectiveness as a reliable alternative.
- **Anonymized data** shows **the lowest** performance when compared to **the synthetic data,** underscoring the limitations of anonymization techniques.
- These findings support the efficacy of our solution, which is **easy, fast and scalable**.

Read more about the assessement

PRODUCT
OVERVIEW

THE SYNTHO ENGINE PLATFORM

Syntho provides a self-service synthetic data generation platform to unlock your data and to take away legitimate privacy concerns.

The key pillars of our platform:

- **Privacy-by-design:** Our solution is built with *privacy-by-design* principles, guaranteeing that data remains secure throughout the deployment process.
  - *Choose from deployment options such as on-premise, private cloud, Syntho cloud, or any other environment.*
- **Maximized Data Accuracy:** Our platform excels at replicating synthetic data to closely mimic real data, achieving an "as-good-as-real" level of accuracy. It ensures that all patterns and relationships present in the original data are captured faithfully.
- **Easy use:** Our user-friendly platform enables anyone to generate and benefit from synthetic data. Enjoy an intuitive and straightforward experience through our self-service platform.

THE SYNTHO ENGINE PLATFORM: KEY DIFFERENTIATORS

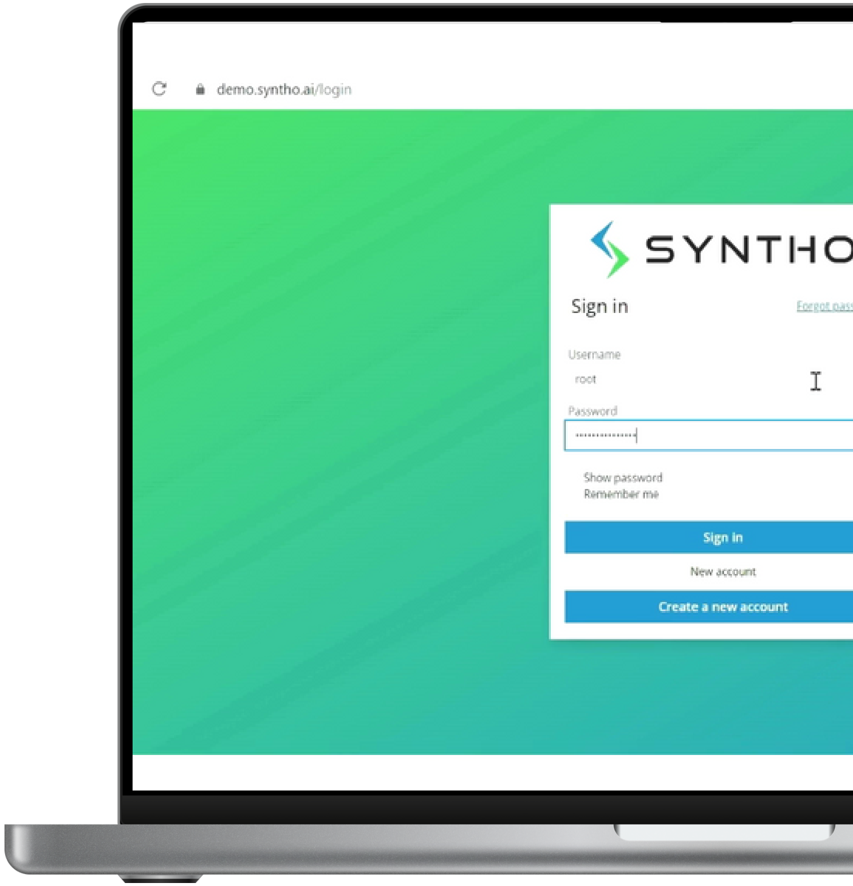
Easy to use
AI-Generated Synthetic data in just 3 steps with our user-friendly self-service platform

Fastest
Fastest synthetic data generation (benchmarked against open source and commercial)

Massive data
Support for massive data with minimal computing resources

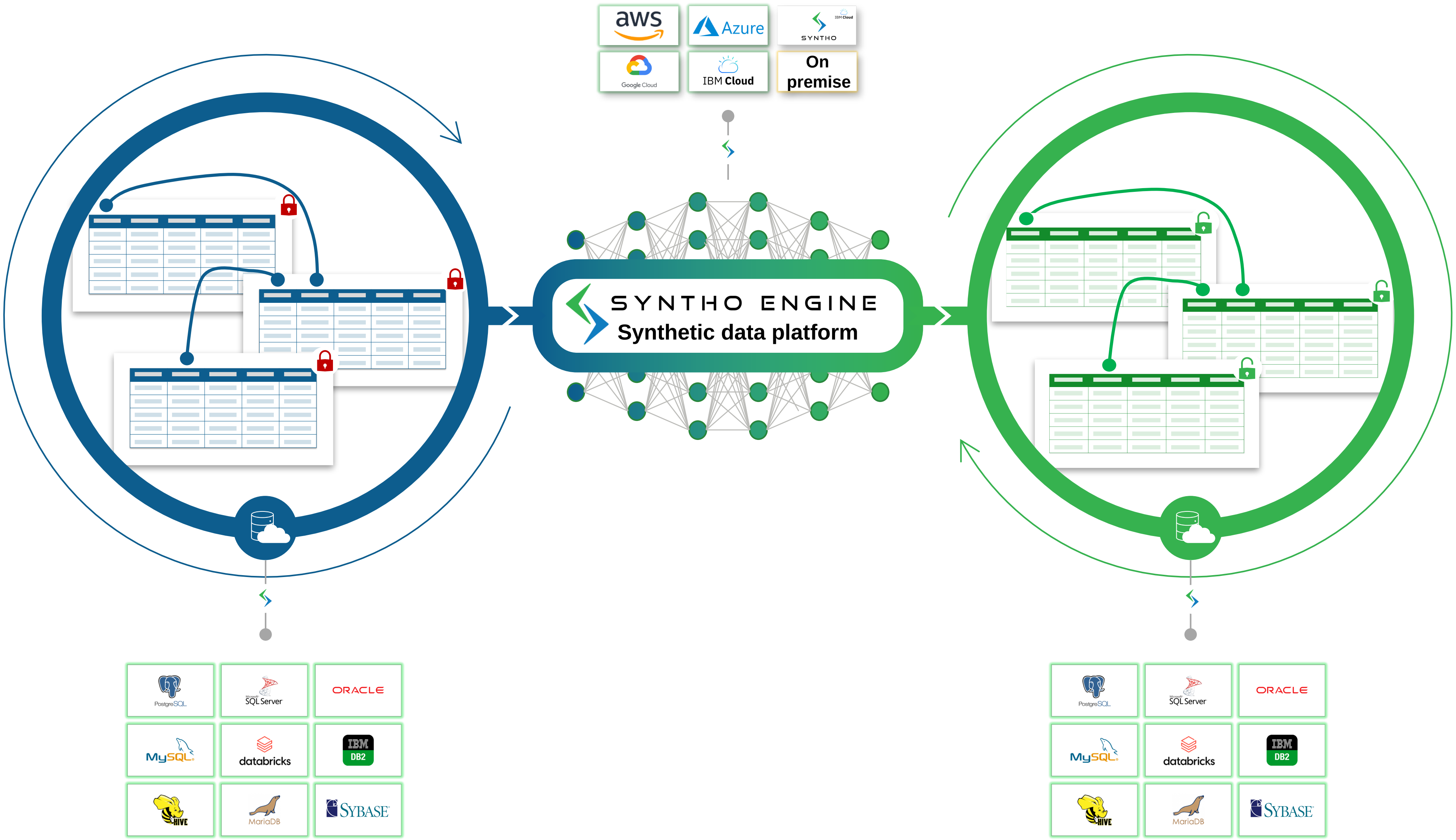
Time-series
Enhanced support for time series / data longitudinal data

Extended features
Such as PII scanner and mockers; PII scanner in open texts; advanced mockers; new connectors; subsetting; workspace sharing


demo.syntho.ai/login

Username: root

AN END-TO-END INTEGRATED PLATFORM



AI GENERATED SYNTHETIC DATA

Syntho is on a mission to solve the global privacy dilemma and enable the open data economy, where data can be used and shared freely and privacy is guaranteed. Hence, we build the future of data privacy with AI-generated synthetic data.

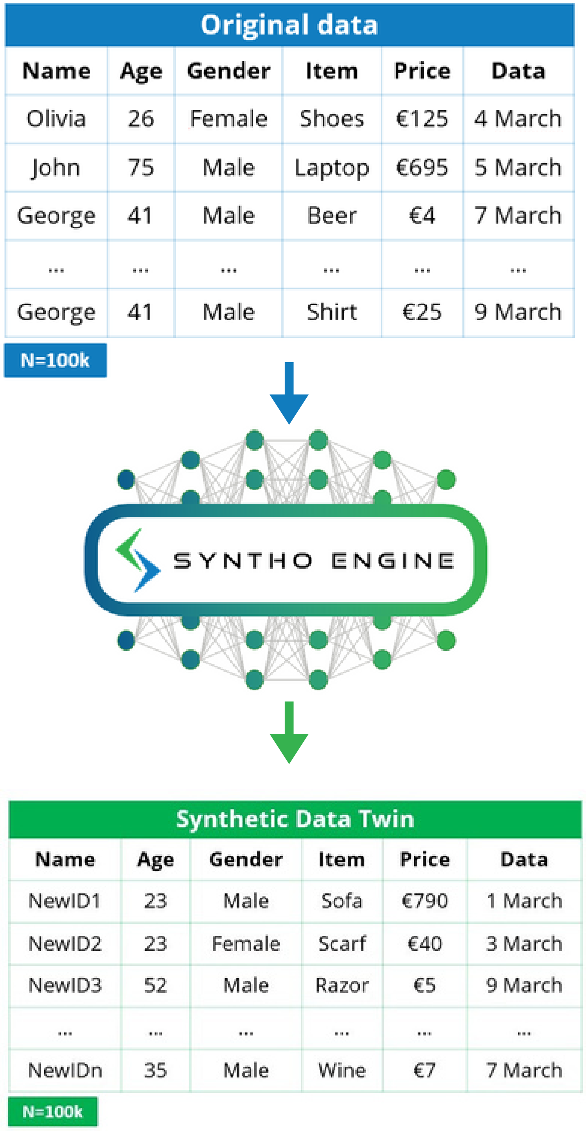
Our **Syntho Engine** learns by utilizing the power of AI all statistical patterns, relations, and characteristics that are in the original data.

The Syntho Engine generates completely new and artificially generated data points. Hence, there are no privacy risks, because synthetic data is completely new and artificially generated data, and individuals simply do not exist anymore.

The key difference, we apply AI to model the synthetic data in such a way that we preserve those statistical patterns, relations, and characteristics to such an extent that it can even be used for analytics.

As a result, this **synthetic data twin** is:

- **as good as real** and statistically identical to the original data,
- there is **no privacy risk**
- and works **easily, fast, and scalable**.



Original data

Name	Age	Gender	Item	Price	Data
Olivia	26	Female	Shoes	€125	4 March
John	75	Male	Laptop	€695	5 March
George	41	Male	Beer	€4	7 March
...	...	...	...	...	...
George	41	Male	Shirt	€25	9 March

N=100k

Synthetic Data Twin

Name	Age	Gender	Item	Price	Data
NewID1	23	Male	Sofa	€790	1 March
NewID2	23	Female	Scarf	€40	3 March
NewID3	52	Male	Razor	€5	9 March
...	...	...	...	...	...
NewIDn	35	Male	Wine	€7	7 March

N=100k

USE CASES

USE CASE 1:
AI-GENERATED SYNTHETIC DATA AS TEST DATA

Testing and development with representative test data is essential to delivering state-of-the-art software solutions. Using personal data or original production data as test data is not allowed and alternative methods are outdated and introduce “legacy-by-design”, because they do not reflect production data, are time-consuming, and require manual work.

Challenge

- Data is locked due to privacy reasons
- Data misses the business rules
- Limited data sharing with external/offshore test teams
- Time and resource consumption to build test data sets (60+ days)
- It is hard to recreate complex databases with referential integrity

Synthetic data helps with it

- Accurate production-like data
- Privacy by design by deploying on clients' secure infrastructure
- Fast data generation (2+ hours)
- Keep test data up-to-date
- Replicate the BD business rules and keep referential integrity
- No domain and programming knowledge is required
- Data subletting allows the creation of a smaller or larger representative subset of a database

USE CASE 2:
SYNTHETIC DATA FOR DATA ANALYTICS

Data-driven solutions are only as good as the data that they can utilize. This is challenging to get access to private data due to strict data/privacy regulations. Classic anonymization is not a solution as it is still a privacy risk, it destroys the data, getting access, and anonymizing it is time-consuming.

Challenge

- Privacy and legal regulations lock data
- Low data quality because of data anonymization
- Transactional, behavioral, and location data can not be used as is
- AI/ML model quality is low and not ready to be used in production
- Unbalanced dataset

Our solution

- Accurate and full data
- As good as real, quality assurance mimics the original data
- Unsampling of underrepresented data
- Safe data-sharing across stakeholders
- Transactional, behavioral, and location data improve AI/ML model performance

USE CASE 3:
SYNTHETIC DATA FOR PRODUCT DEMOS

Product demos play a crucial role in showcasing the capabilities and value of a product to potential customers. As well as new employees' education, when they need to have a safe environment to learn the systems.

Challenge

- Cross-border data sharing is prohibited by regulators
- Lack of high-quality test data for vendor selection
- Product demos cannot include real data
- Employee education should performed on real data, without privacy violation

Synthetic data helps with it

- Synthetic data is not personal data, as it does not contain one-to-one relationships
- Provide an opportunity to share as good as real data with vendors and perform a high-quality selection process
- Showcase the product/platform capabilities based on real flows and user scenarios

MORE INFORMATION

Syntho Syntho is the Amsterdam based startup that is revolutionizing the tech industry with AI-generated synthetic data. It was founded in 2020 with the goal of solving the privacy dilemma and enable the open data economy, where data can be used and shared freely and privacy guaranteed. Syntho enables organisations to boost innovation in a privacy-preserving way by providing AI software for synthetic data. Syntho is the winner of the 2020 Philips Innovation Award.

Wim Kees Janssen
CEO & Founder

If you have any questions regarding synthetic data, do not hesitate to contact us via **email (*kees@syntho.ai*),** schedule a demo or a deep-dive.

www.syntho.ai

SYNTHO