SYNTHO

SYNTHO GUIDE

Unlock privacy sensitive data with synthetic data

TABLE OF **CONTENTS**

ABOUT **SYNTHO**

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

As leading provider of synthetic data software, Syntho's mission is to empower businesses 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, our ultimate goal is to foster an open data economy where information can be freely shared and utilized without compromising privacy.

Accordingly, Syntho has been honored with prestigious accolades such as the **Philips Innovation Award** and **Unesco's Challenge** at VivaTech. Additionally, Syntho has been recognized as a "to watch" generative AI startup by **NVIDIA**.

COOPERATION WITH **LEADING ORGANIZATIONS**

Recognitions

NVIDIA

SAS

UNESCO

Client References

Cedars Sinai

Erasmus MC
University Medical Center Rotterdam

Centraal Bureau voor de Statistiek

PHILIPS
Innovation Award

ERASMUS UNIVERSITEIT ROTTERDAM

INTRODUCTION

Currently, our world is undergoing a digital revolution, which has been fueled by the emergence of data-driven solution such as: ***software, business intelligence and, artificial intelligence***.

These innovative technologies have the potential to reshape industries, drive growth, and unlock unprecedented opportunities. However, it is crucial to acknowledge that the effectiveness of these solutions relies heavily on the quality and accessibility of data.

The Importance of Data Privacy

Data privacy has become a paramount concern in the modern era, as individuals and organizations seek to protect sensitive information and maintain control over personal data. Stricter data privacy regulations have been introduced to safeguard user privacy and ensure responsible data handling practices. While these regulations provide essential protections, they have inadvertently led to a significant portion of data being locked away, limiting its potential for utilization.

Currently, approximately **50%** of data remains inaccessible due to strict data privacy regulations.
This staggering statistic highlights the challenges businesses face in harnessing the full power of data-driven solutions.

As a result, a staggering **$4T** worth of missed data opportunities has been estimated globally.
This untapped potential represents a lost chance for businesses to gain valuable insights, improve decision-making, and drive innovation.

WHY DOES CLASSIC 'ANONYMIZATION' **NOT WORK ANYMORE?**

In order to address privacy concerns in datasets or databases, conventional 'anonymization' techniques have often been employed. These techniques share a common goal of manipulating original data to impede the identification of individuals. The process typically involves the following steps:

1. Initial removal of direct personal identifiers, such as names.
2. Aggregation of indirect information, such as age.
3. Continued manipulation of the remaining data.

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

- **Privacy risk** - even with the application of classic anonymization techniques, a privacy risk persists. While these techniques make it more challenging to identify individuals, it is not impossible to do so.
- **Destroying data** - the more extensive the anonymization efforts, the greater the privacy protection. Unfortunately, this also leads to data destruction. For analytics purposes, this is undesirable since it can result in poor insights due to the loss of valuable information.
- **Time-consuming** - implementing traditional 'anonymization' techniques is a time-consuming endeavor. The effectiveness of these techniques varies depending on the dataset and data type, necessitating significant effort to ensure proper application.

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

1 2 3

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

WHY SHOULD ORGANIZATIONS **START USING SYNTHETIC DATA?**

50%
Of data for AI will be unlocked by privacy enhancing techniques

Unlock data and valuable insights

AI-generated synthetic data unlocks valuable insights by enabling organizations to utilize sensitive data that would otherwise be inaccessible. With privacy-enhancing techniques like synthetic data generation, up to 50% of locked data can be unlocked, empowering organizations to adopt a data-driven strategy and gain a competitive edge. This recognition is driving wider adoption and increased innovation in AI and machine learning powered by AI-generated synthetic data.

30%
More profits for companies that earn and maintain digital trust with customers

Gain digital trust

In today's digital world, trust is crucial. AI-generated synthetic data builds trust by avoiding the use of real individuals' sensitive information, demonstrating commitment to privacy and security. Companies that earn digital trust can see a 30% increase in profits. Adopting AI-generated synthetic data enables innovation and competitive advantages, showcasing a prioritization of trust compared to organizations that don't. As data reliance grows, responsible data policies and AI-generated synthetic data adoption will continue to expand.

70%
Increase in industry collaborations expected with use of privacy tools

Drive industry collaborations

Organizations collaborate in the data-driven world for innovation and competitiveness, but privacy concerns and data silos limit sharing sensitive data. AI-generated synthetic data mimics real-world data securely, facilitating collaboration and reducing risks. Embracing it results in a 70% increase in industry collaborations, unlocking opportunities for faster innovation. As collaboration gains recognition, wider adoption of privacy-enhancing techniques like AI-generated synthetic data is expected.

1,000,000hrs
Millions of hours saved by organizations that embrace synthetic data

Realize speed and agility

Organizations need agility to stay competitive in today's business landscape. Privacy regulations create slack and dependencies when handling personal data. AI-generated synthetic data reduces reliance on real-world data, saving time and resources. This agility accelerates tech development and deployment, providing a competitive advantage. As organizations recognize the importance of minimizing dependencies and embracing an agile approach, we can expect wider adoption and increased innovation in AI-generated synthetic data.

THE DUTCH DPA ABOUT
USING PERSONAL DATA AS TEST DATA

AUTORITEIT PERSOONSGEGEVENS

Vragen van organisaties over testen

Mag ik testen met persoonsgegevens bij de ontwikkeling van een systeem of applicatie?

Dat is niet aan te raden. Testen is een complex proces, waarvoor zorgvuldigheid en meerdere gescheiden omgevingen nodig zijn. Het testen met persoonsgegevens brengt namelijk risico's met zich mee.

Aparte grondslag

De mensen van wie u persoonsgegevens verwerkt, verwachten niet dat u hun gegevens ook voor testdoeleinden gaat gebruiken. Dat betekent onder meer dat u voor het testen een aparte grondslag moet hebben.

Niet noodzakelijk

Verder is het vaak niet noodzakelijk om te testen met persoonsgegevens, omdat er meestal alternatieven mogelijk zijn. Dat is een van de redenen dat testen met persoonsgegevens moeilijk in overeenstemming te brengen is met de AVG.

Eind van het [illegible]

[illegible]persoonsgegevens in het nieuwe systeem inlezen. En ook die verwerking moet zeer zorgvuldig gebeuren.

"Testing with personal data is *difficult to reconcile with the GDPR*"

What is allowed?

Welke gegevens kan ik wel gebruiken om testen uit te voeren?

U kunt bijvoorbeeld onderzoeken of er synthetische gegevens of testdata ('dummy data') beschikbaar zijn. Stel daarbij altijd vast dat de dataset die u wilt gebruiken niet alsnog persoonsgegevens bevat.

De Rijksdienst voor Identiteitsgegevens biedt bijvoorbeeld een reeks test-burgerservicenummers aan.

[illegible] een nieuwe applicatie [illegible] inzetten.

"You can explore the availability of synthetic data or mock data"

Read more from the AP official website.

PRODUCT
OVERVIEW

AI-GENERATED **SYNTHETIC DATA**

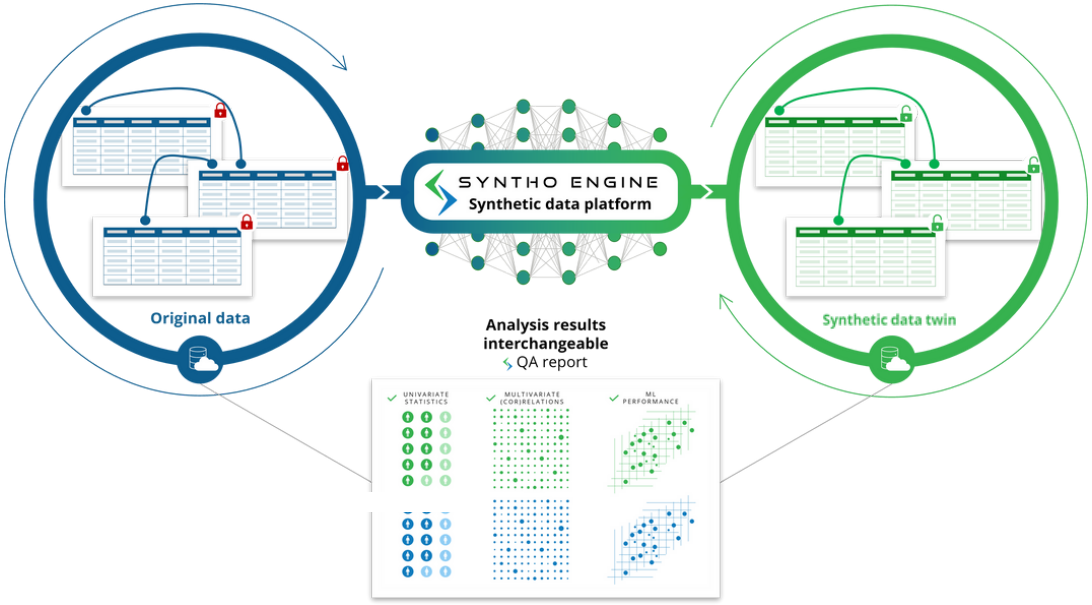
Organizations dealing with highly sensitive data often encounter challenges in utilizing and sharing this information with stakeholders. Due to the sensitive nature of the data, traditional usage and sharing methods are not viable options. Consequently, these organizations miss out on data-driven innovation 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. This unlocks significant benefits for these organizations, including reduced risk, increased data availability, and faster access to data.

Therefore, 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*

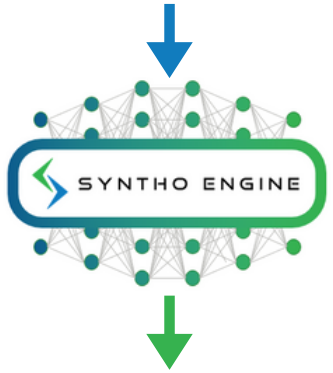
Through our advanced **Syntho Engine** software, we harness the capabilities of artificial intelligence to create synthetic data twin that closely mimic the original (including sensitive) data.



HOW DOES
SYNTHETIC DATA GENERATION WORK?

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

The **Syntho Engine** revolutionizes data generation by creating completely new and artificially generated datapoints. This approach eliminates any privacy risks as synthetic data is entirely new and artificially generated data.

The key difference lies in our utilization of AI to model the synthetic data with a diligent precision, ensuring that the statistical patterns, relationships, and characteristics of the original data are preserved. This preservation allows the synthetic data to be effectively utilized for analytics purposes.

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 easy, fast and is scalable**

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

01

Easy to use

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

02

Fastest

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

03

Massive data

Support for massive data with minimal computing resources

04

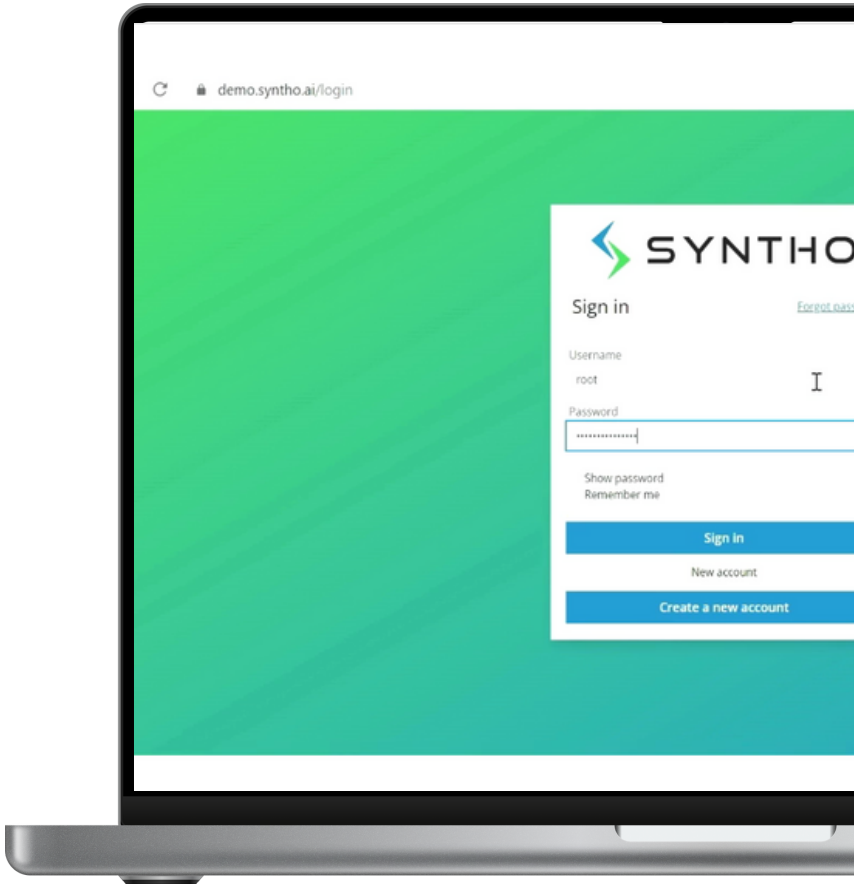
Time-series

Enhanced support for time series / data longitudinal data

05

Extended features

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



THE SYNTHO ENGINE PLATFORM: **EXTENDED FEATURES**

PII scanner and mockers

The PII Column scanner detects PII in a user's database using shallow and deep scans, and suggests mockers for each PII entity as replacements.

PII scanner in open text

Syntho's PII scanner identifies and handles PII in open text data by removing or replacing it with placeholders or mock values.

Advanced mocker

Mock data substitutes real or sensitive information, simplifying data generation by creating it from scratch or following predefined rules.

New connectors

Syntho offers connectors for easy configuration of synthetic data generation and supports over 20 databases and filesystem connectors for an integrated approach.

Subsetting

Database subsetting creates smaller or larger subsets of a database while maintaining referential integrity. It helps businesses expand data or reduce computation costs by working with smaller subsets.

Workspace sharing

Workspace sharing enables collaborative and scalable use of synthetic data in organizations. Teams can work together or separately within the same workspace, with role-based access levels and permissions.

Questions?

Discover the power of synthetic data generation. Schedule a demo with us to explore how our platform can benefit your data-driven solutions.

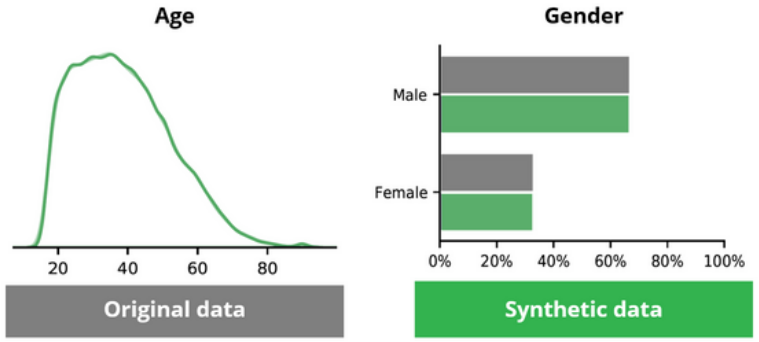
Learn more

Book a demo

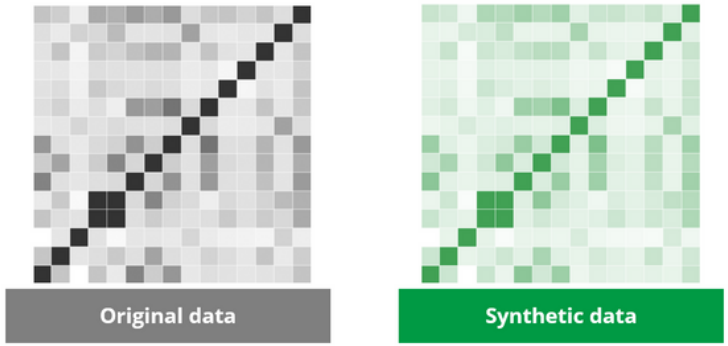
QUALITY EVALUATION OF **THE SYNTHETIC DATA**

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).

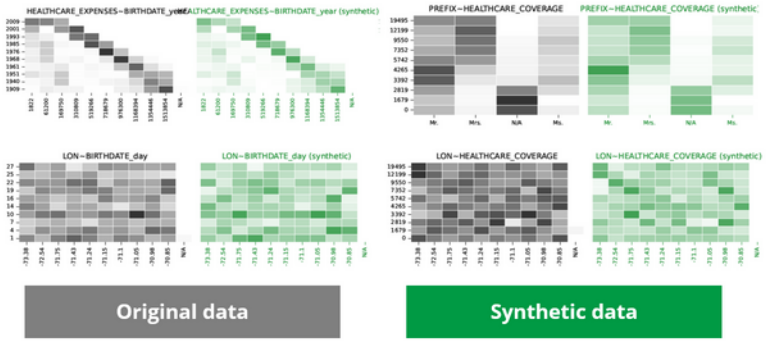
- The distributions, the frequency of variables in the dataset, are **similar**.



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



- Of course, our quality assurance report contains many more.



THE SAS DATA EXPERTS APPROVED
OUR AI GENERATED SYNTHETIC DATA

Read more about the assessement

We take great pride in our partnership with **SAS**, as their esteemed data experts have conducted a thorough evaluation and endorsement of our synthetic data.

During the assessment, we used 4 powerful machine learning models: *neural network, logistic regression, gradient boosting and the random forest.* These models were utilized to predict churn in a telecom use case, with the area under the curve serving as a performance indicator for machine learning accuracy.

We trained them using three distinct datasets:

1. **Original data**
2. **Anonymized data**
3. **Synthetic data generated** by Syntho.

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**.

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.

Challenge

Utilizing personal data or original production data as test data is prohibited, and traditional alternative methods are outdated, time-consuming, and require manual effort. These methods lack the ability to reflect production data accurately and often introduce "legacy-by-design" challenges.

Our solution

Accelerate the delivery and release of cutting-edge software solutions with high-quality AI-generated synthetic test data.

- Production-like data
- Privacy by design
- Easy, fast, and agile

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.

Challenge

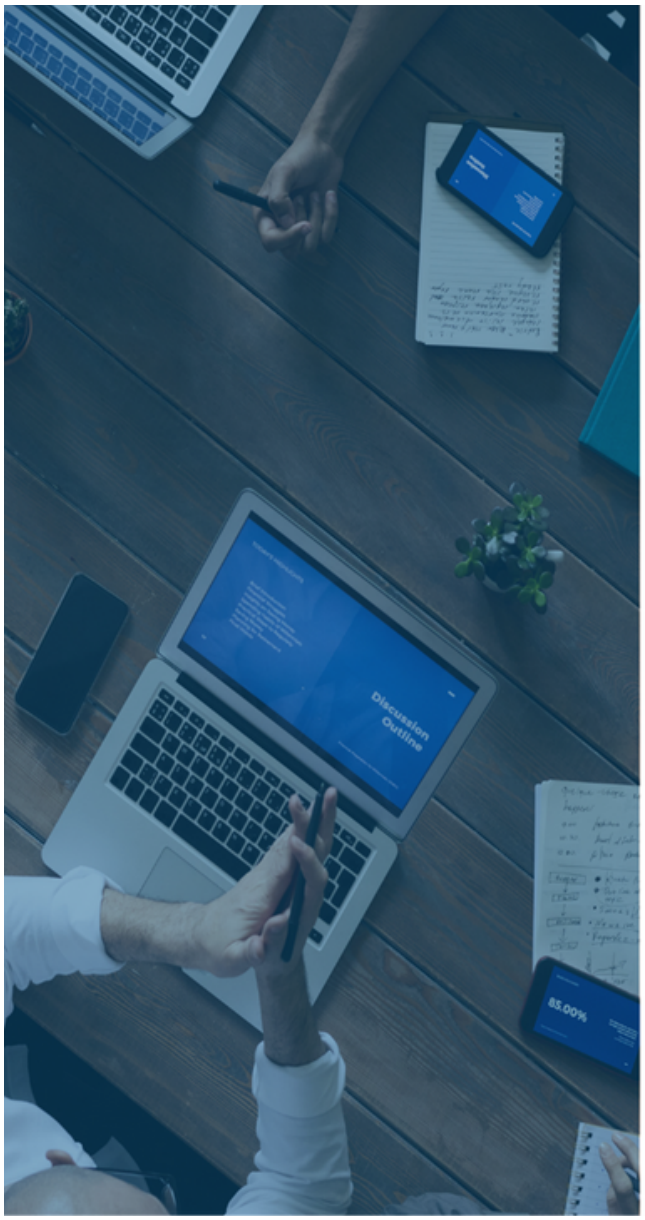
Traditional anonymization methods are not effective as they still pose privacy risks, destroy data, and require significant time and effort for access and anonymization.

Our solution

Harnessing the capabilities of AI, Syntho creates a data twin of the original data.

- Unlock (sensitive) data
- As-good-as-real data
- Easy, fast, and scalable

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. Accessing or utilizing real customer data for products.

Challenge

Product demos face the challenge of demonstrating realistic scenarios and data without compromising customer privacy or security. Using real customer data for demos may involve sensitive information, such as personally identifiable data, transaction details, or proprietary datasets.

Our solution

We offer a synthetic data solution that addresses the challenge of utilizing real customer data for product demos.

- Errorless, high-quality demo data
- Tailor your product demo
- Easy, fast, and agile

MORE INFORMATION

Wim Kees Janssen
CEO & Founder

Synthetic Data - Real People!

Though, we are experts in synthetic data, our team is real, so if you have any questions, do not hesitate to contact **Wim Kees Janssen** via **email (kees@syntho.ai)** or visit our website ***www.syntho.ai*.**

kees@syntho.ai

syntho.ai/meet

syntho.ai/careers

SYNTHO