SYNTHO

SYNTHO GUIDE

Unlock privacy sensitive data with synthetic data

Table of contents

About Syntho

Syntho is a data technology organization with a strong expertise in AI-generated synthetic data, headquartered in Amsterdam, Netherlands.

It was founded in 2020 with the goal of solving the global privacy dilemma and enable the open data economy, where data can be used and shared freely and privacy guaranteed.

As winner of the **2020 Philips Innovation Award**, Syntho enables organisations to boost innovation in a privacy-preserving way by providing AI software for synthetic data generation.

Introduction

Currently, our world is undergoing a digital revolution, which is accelerated by data-driven solution such as:

- ***software***
- ***business intelligence***
- ***artificial intelligence***

In reality, those solutions are only as good as the data that can be utilized

50%

Of data is locked due to strict data privacy regulations

$4T

Worth of 4 Trillion dollars of missed data opportunities, due to strict data privacy regulations

Data privacy is real, but offer opportunities!

70%

Increase in industry collaborations expected with use of privacy tools

30%

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

60%

of all training data for AI will be synthetically generated by 2024

Gartner

Why classic 'anonymization' does not work anymore?

To overcome this on 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 One starts with deleting the direct personal identifiers, such as names.

2 Then the indirect information will be aggregated, like age.

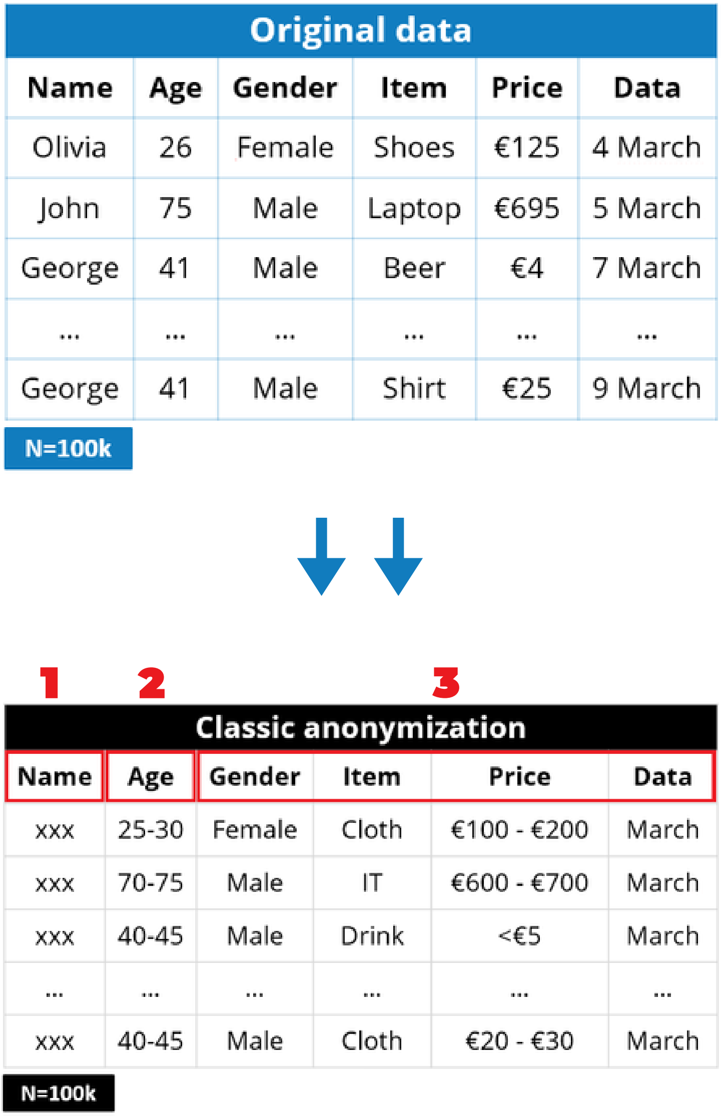
3 And one will continue to manipulate the rest of the data.

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

- **Privacy risk** - you will always have a privacy risk. Applying those classic anonymization techniques makes it only 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 different per dataset and per datatype.



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

AI-generated synthetic data

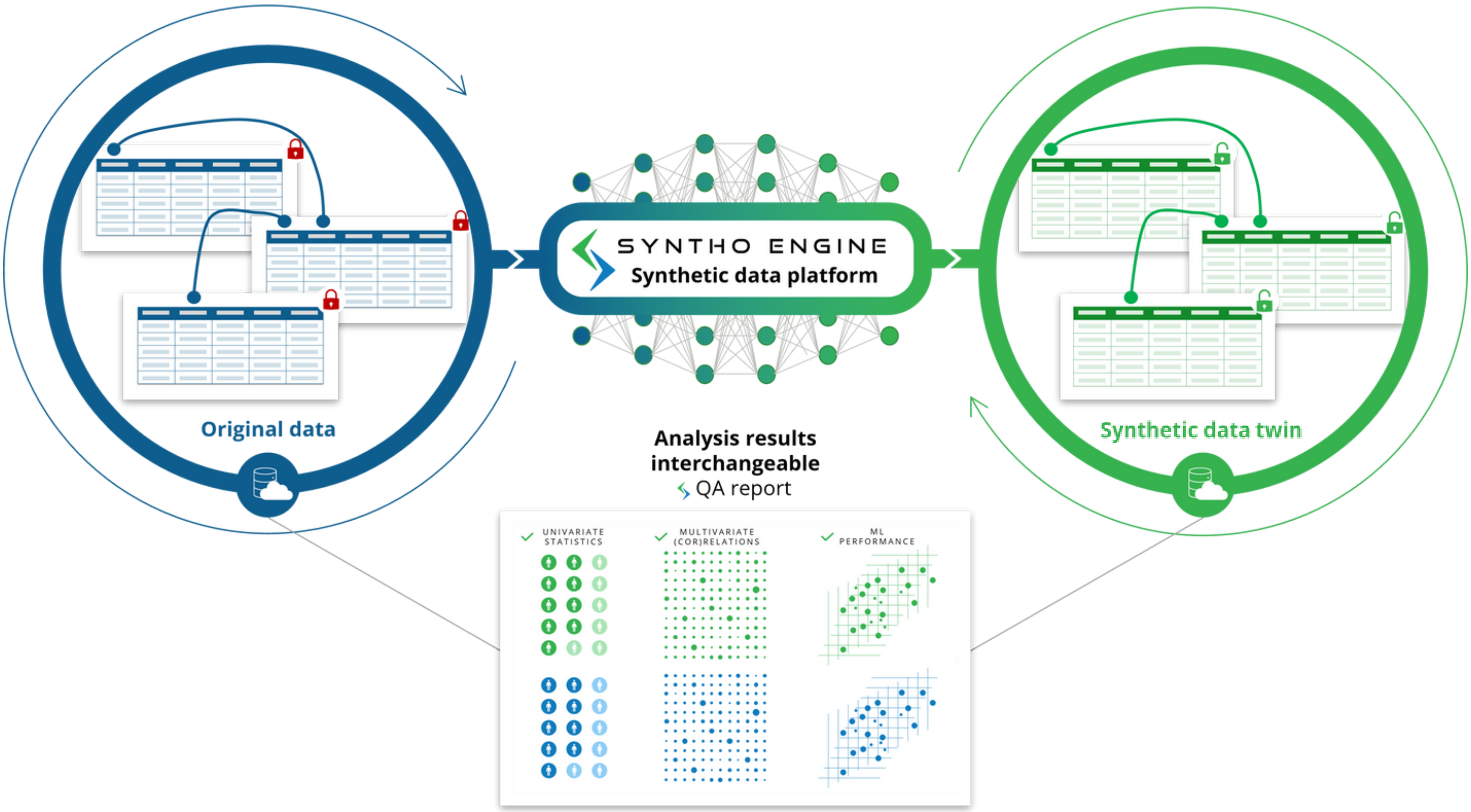
As Syntho's main goal is to solve the global privacy dilemma, we build the future of data privacy with AI generated synthetic data.

Synthetic data is artificially generated data that mimics real-world data.

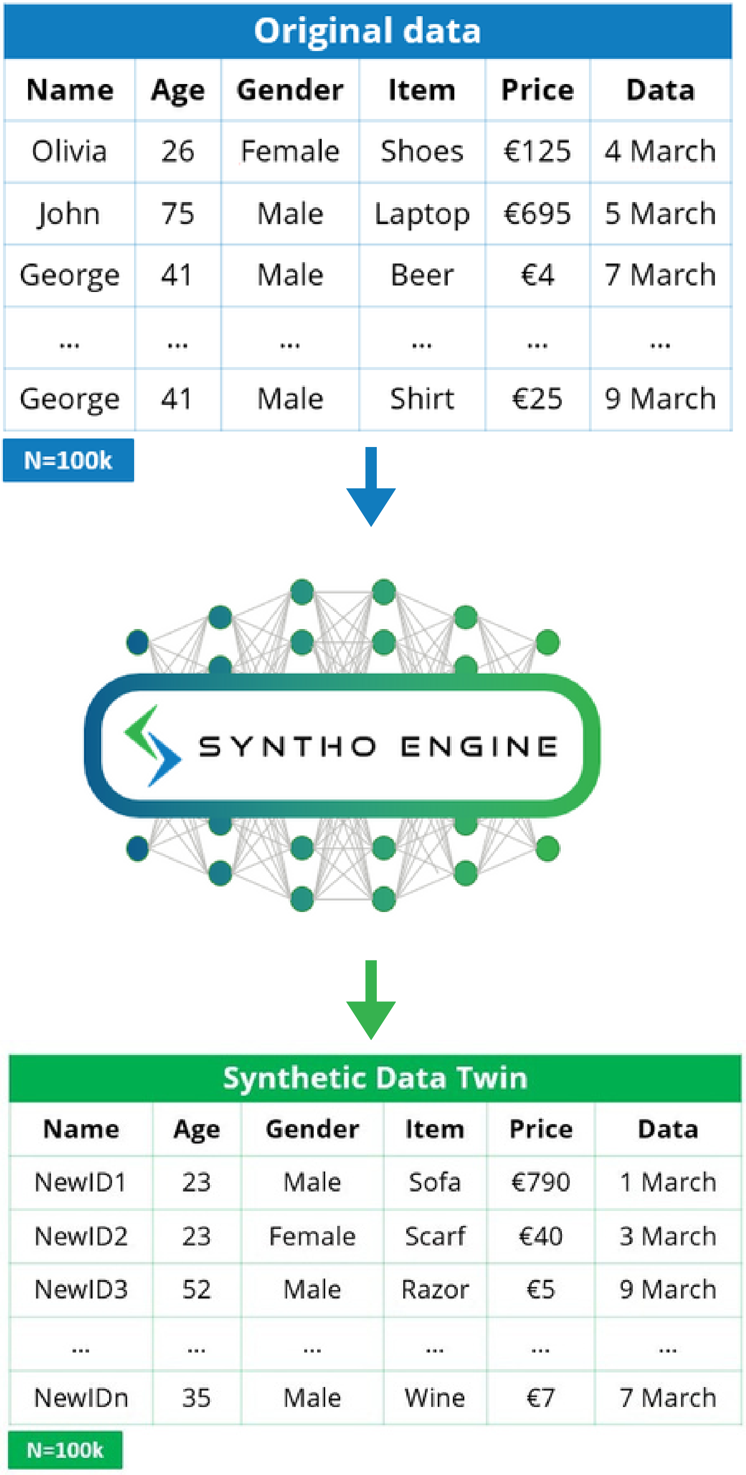
Why?
Privacy by design is a key driver for business success, because it:

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

Our **Syntho Engine** software mimics (sensitive) data by utilizing the power of AI to generate a synthetic data twin of the original 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** generates completely new and artificially generated datapoints. Hence, there are no privacy risks, because synthetic data is a 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.

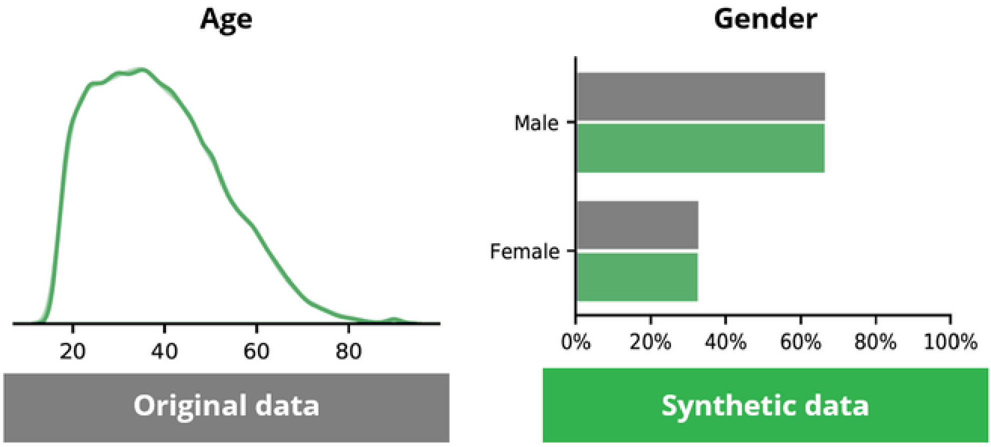
As a result, this *synthetic data twin* is:

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

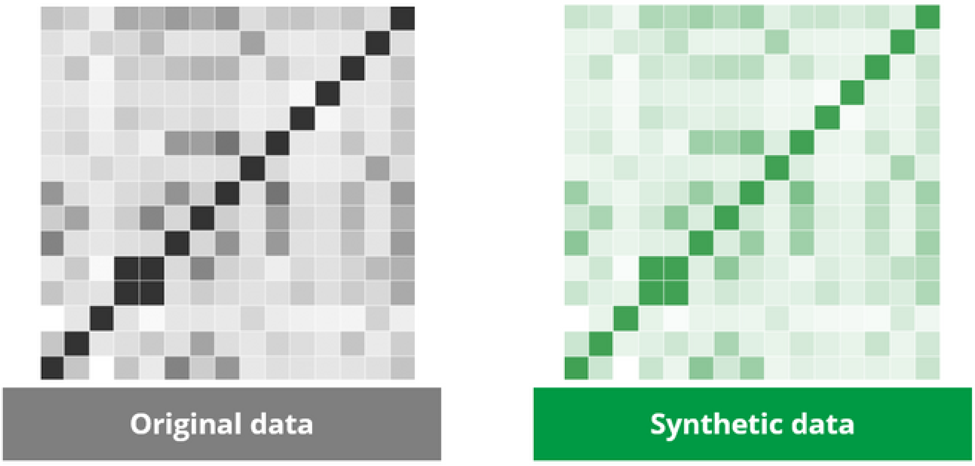
Our data quality report

We prove this with our data quality report, where we compare the original data in grey with the synthetic data in green.

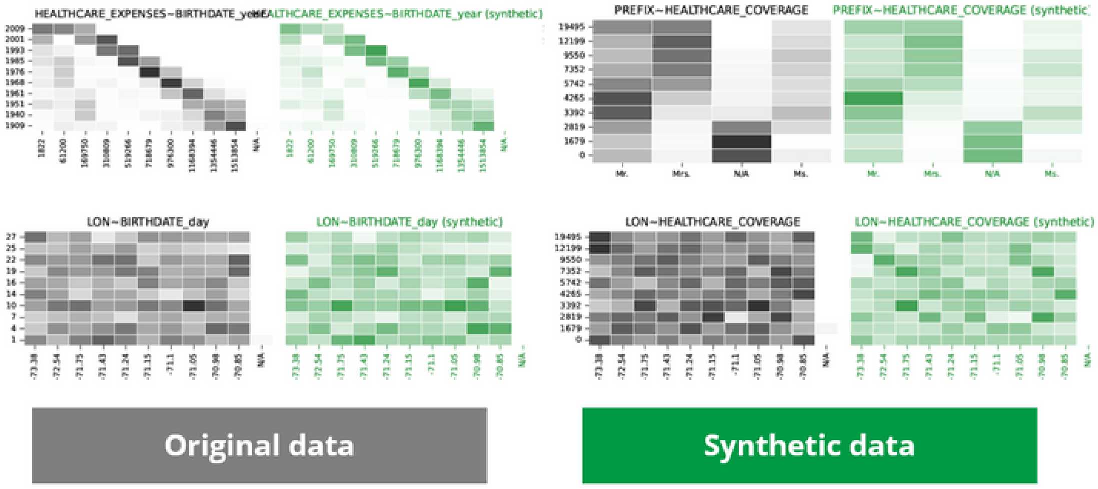
- The distributions, the frequency of variables in the dataset, are similar.



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



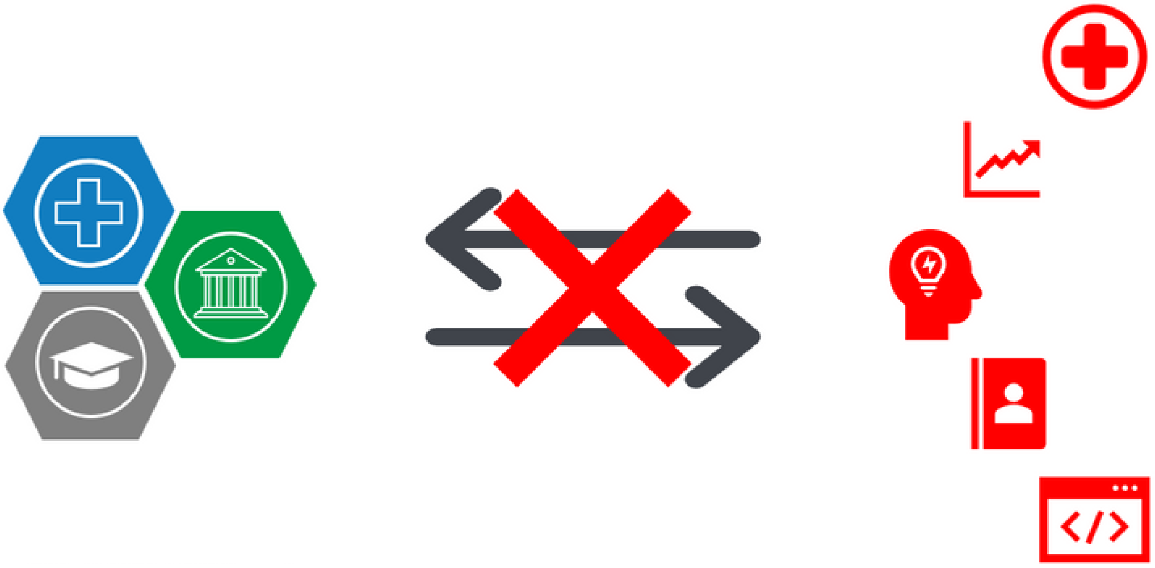
- Of course, our quality assurance report contains many more.



How does it work in practice?

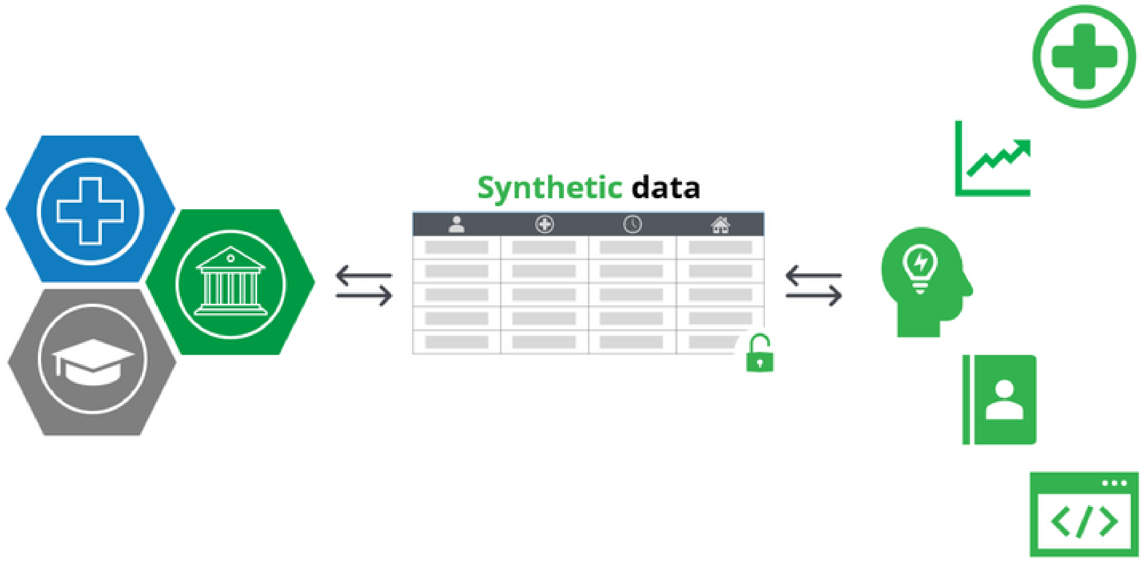
Using, sharing and selling data is challenging

Highly sensitive data is typically collected by organizations that work with the most privacy sensitive information. This data cannot be simply used and shared with stakeholders. Consequently, those organizations cannot realize data-driven innovation and they miss data opportunities

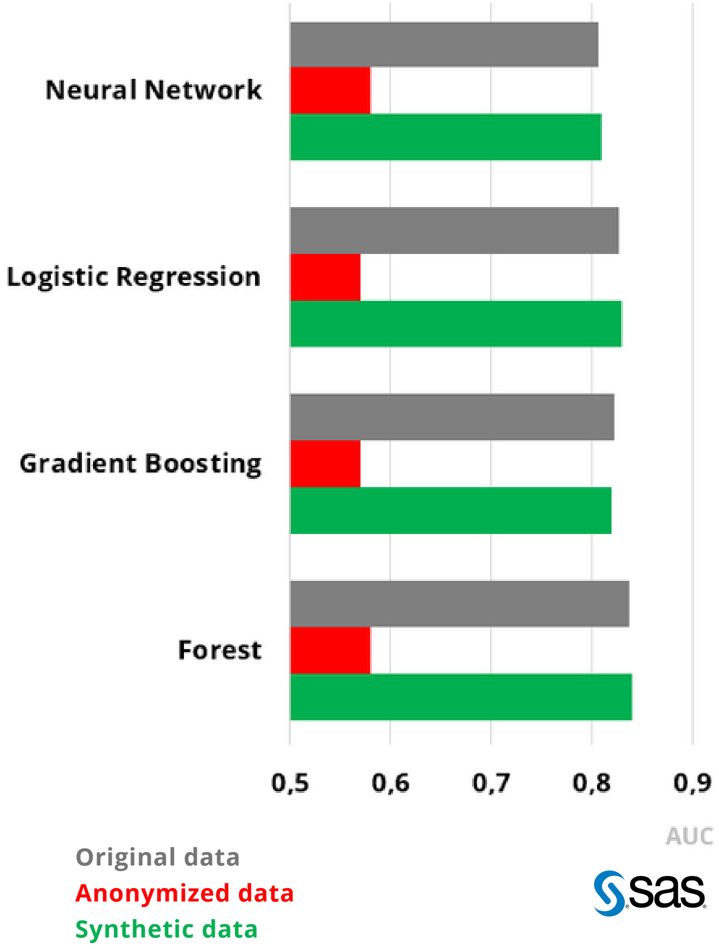


Freely using, sharing and selling synthetic data

Our solution: share the data in synthetic form to unlock this data. Benefits for those organizations: Less risk, More data and Faster data access. After our visit, those organizations can test, develop and innovate based on synthetic data.



The SAS data experts approved our AI generated synthetic data



We are very proud of our collaboration with SAS, because their data experts assessed and approved our synthetic data.

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

1. We trained them on the **original data**
2. We trained them on **anonymized data**
3. And we trained them on **synthetic data** from Syntho.

These are the results and conclusions from this assessment:

- **Synthetic data** show **similar** performance in comparison to the **original data**
- **Anonymized data** shows **worst** performance in comparison to the **synthetic data**
- In a solution that works **easy, fast and scalable**.

Client cases

1

Smart test data

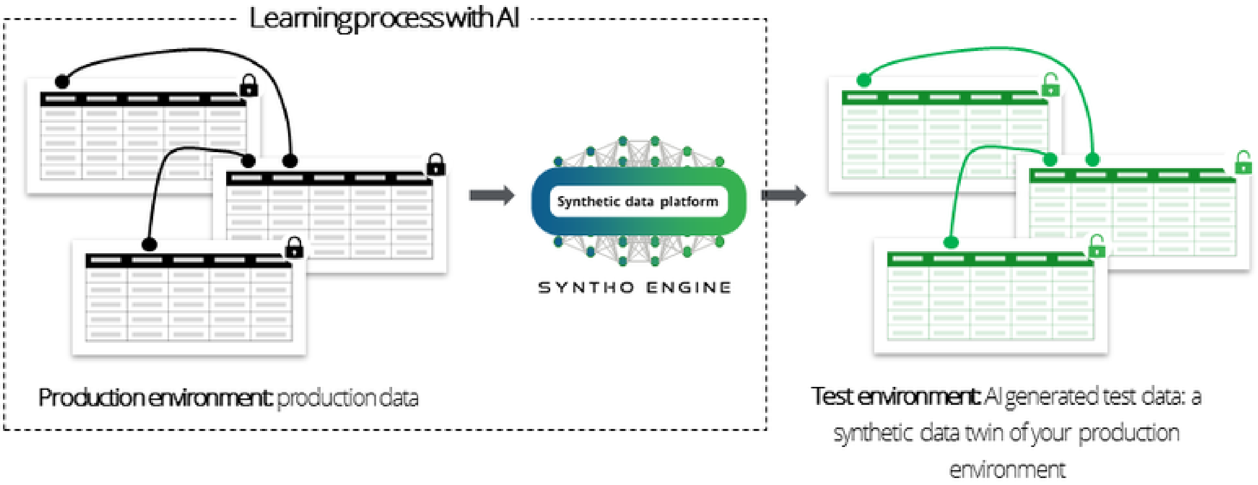
Testing and developing with high quality test data is essential to deliver state-of-the-art software solutions. Using original production data seems obvious, but is not allowed due to (privacy) regulations. This introduces challenges for many organizations in getting the test data right.

Issue

Classic Test Data Management (TDM) tools* fail, because they introduce "legacy-by-design“. The test data from those classic TDM tools”:

- Does not reflect production data
- Works slow and time consuming
- Requires manual work

** Examples: anonymized data, scrambled data, dummy data etc.*


Learning process with AI
Synthetic data platform
SYNTHO ENGINE
Production environment: production data
Test environment: AI generated test data: a synthetic data twin of your production environment

Solution

Mimic your production data with AI to generate a synthetic data twin of your production data:

- Production-like test data
- Privacy-by-design
- Easy, fast and scalable
- One-click end-to-end refresh of your entire test environment within an hour by the power of AI
- Intelligent data augmentation and simulation

Impact

Deliver state-of-the-art software solutions with AI generated test data:

- Smart test data
- Spot bugs faster and earlier in the testing cycle
- Release faster and shorten the time-to-market
- Utilize test and development resources smarter
- Improve overall test, development and delivery quality
- Realize speed and agility

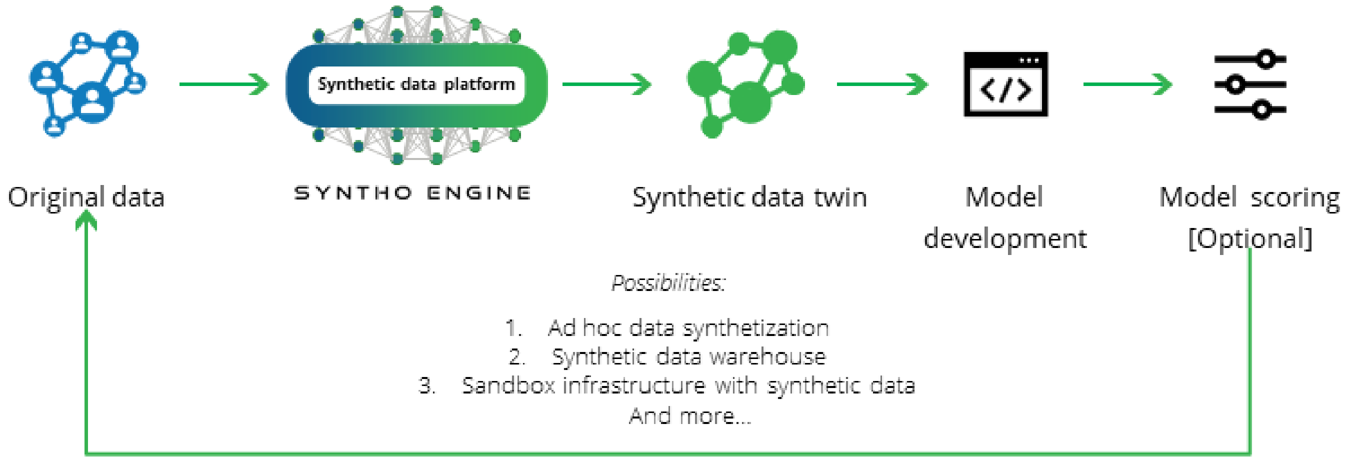
2 Synthetic data for analytics

Having a strong data foundation with easy and fast access to usable, high quality data is essential to develop models (e.g. dashboards [BI] and advanced analytics solutions [AI & ML]). However, many organizations are affected by a sub-optimal data foundation, where data cannot simply be used and shared.

Issue

A sub-optimal data foundation, where data cannot simply be used and shared:

- Data is locked and cannot be touched, while data access is critical
- Getting access to data takes ages
- Classic anonymization does not work
- Bureaucracy around data access requests that introduce slack



Possibilities:

1. Ad hoc data synthetization
2. Synthetic data warehouse
3. Sandbox infrastructure with synthetic data

And more...

Mimic (sensitive) data with AI to generate synthetic data twins:

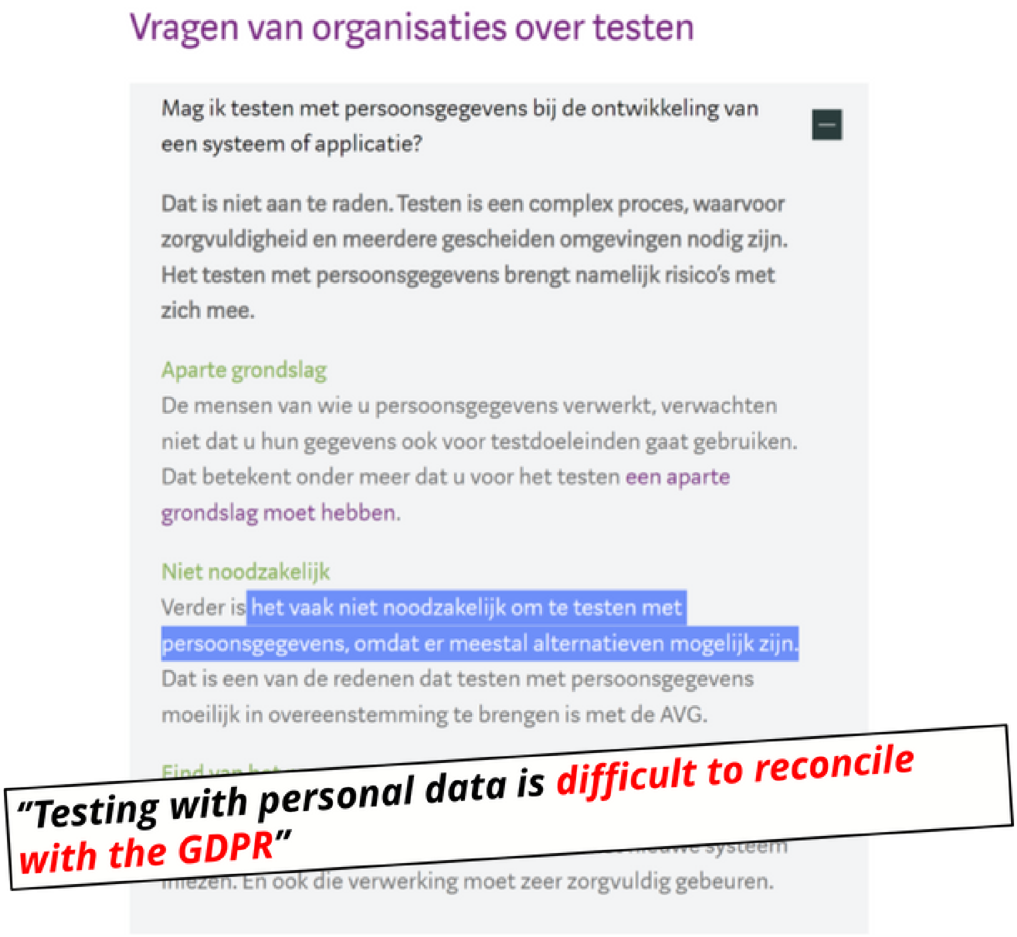
- As-good-as-real data that is statistically identical in comparison to the original data
- Bypass internal processes, risk assessments, data access requests and similar time-consuming overhead
- Unlock your full data potential
- Easy, fast and scalable

Build your strong data foundation with easy and fast access to usable, high-quality data:

- Be smarter than (and even beat) the competition
- Leverage new and more innovation opportunities
- Unlock data, and thereby valuable insights
- Mitigate overhead

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.

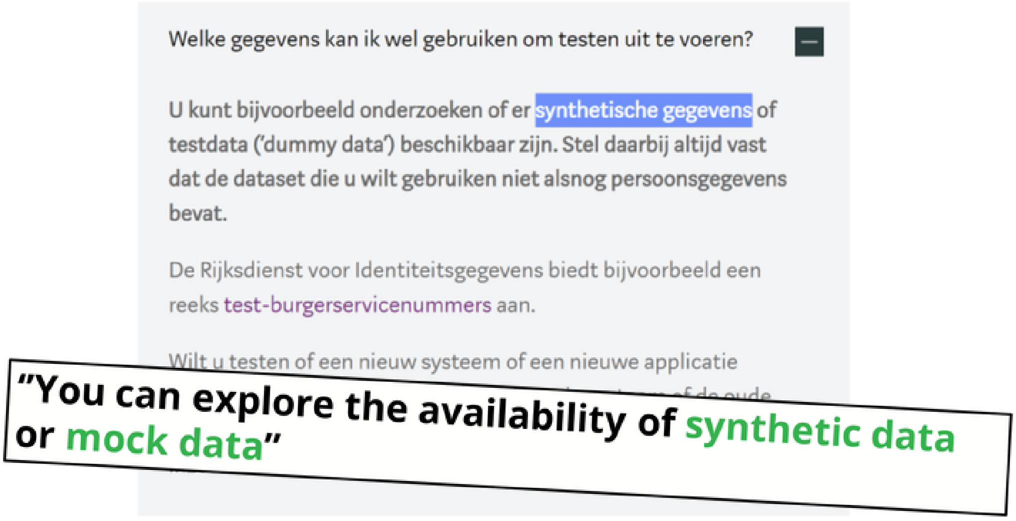
Eind van het

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.

Wilt u testen of een nieuw systeem of een nieuwe applicatie

of de oude

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

Read more from the AP official website.

Our platform

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

The key benefits of our platform:

→ Easy deployment

We typically deploy in the safe environment of the customer so that (sensitive) data never leaves the safe and trusted environment of the customer.

- Deployment options: *on-premise, private cloud, Syntho cloud or any other environment of your choice*

→ Easy connect

We support various integrated connectors so that you can connect with the source-environment (where the original data is stored) and the target-environment (where you want to write your synthetic data to) for an end-to-end integrated approach.

- *20+ tool integrators & database connectors*

→ Easy use

Our platform is optimized for easy use so that anyone can generate and benefit from the value of synthetic data via our easy to use self-service platform.

→ Maximized Data Accuracy

We maximize the data accuracy for every synthetic data generation job and demonstrate this via our data quality report. In addition, the SAS data experts assessed and approved our synthetic data from an external point of view.

→ Automatic data detection

Syntho automatically detects the data types, schemas and formats to maximize data accuracy. For multi-table database, we support automatic table relationship inference and synthesis to preserve referential integrity.

- *Maximized data accuracy for every generated synthetic dataset or database.*

PII detection and data augmentation

Our platform offers a PII scan that scans your entire database on PII elements. We support various intelligent data augmentation features and mockers to generate (PII) data from scratch

- *Data augmentation, mock data and automatic PII detection & generation.*

Minimal Computational Requirements

We optimized our platform to minimize computational requirements (e.g. no GPU required), without compromising on the data accuracy. In addition, we support auto scaling, so that one can synthesize huge databases.

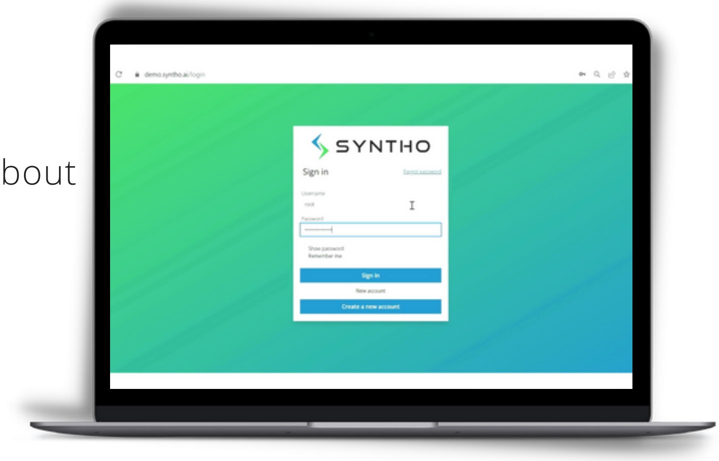
Complex data support

Next to all regular types of tabular data, the Syntho Engine supports complex data types and complex data structures.

- *All tabular data*
- *Time series*
- *Multi-table databases*
- *Open text*

As result, that's how we are able to **unlock** that **50%** of data to **realize** the **$4T** of data opportunities.

Book a demo today to learn more about how you can benefit from using our synthetic data generation platform.



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