SYNTHO

SYNTHETIC DATA IN HEALTHCARE

REPORT

Explore the role of synthetic data in the healthcare sector

INTRODUCTION

Currently, our world is undergoing a digital revolution, which is driven by data-driven solution such as **software, business intelligence** and **artificial intelligence.**

In reality, those solutions are only as good as the data that can be utilized. Because of strict data privacy regulations, **50%** of all data is locked, resulting in **4 trillion dollars** of untapped data opportunities.

In the last few years, many healthcare organizations have suffered from data breaches, which are the most costly in this industry since hospitals and care centers handle sensitive patient data.

Because of these complex and constantly changing regulatory requirements in healthcare, AI became increasingly appealing to these organizations.

100%

more compliance costs for companies that lack privacy protection

$67.4bn

is the projected value of the AI in healthcare market by 2027

93%

of healthcare organizations have experienced a data breach

CLASSIC 'ANONYMIZATION'

To overcome this on datasets or databases, one typically applies classic 'anonymization' techniques that all have in common that they manipulate 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 continues to manipulate the data.

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

- **Privacy risk** - you will always have a privacy risk. It makes that only harder, but not impossible to identify individuals.

- **Destroying data** - the more you anonymize, the better you protect your privacy, but at the same time you destroy your data more. 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

1 2 3

Classic anonymization

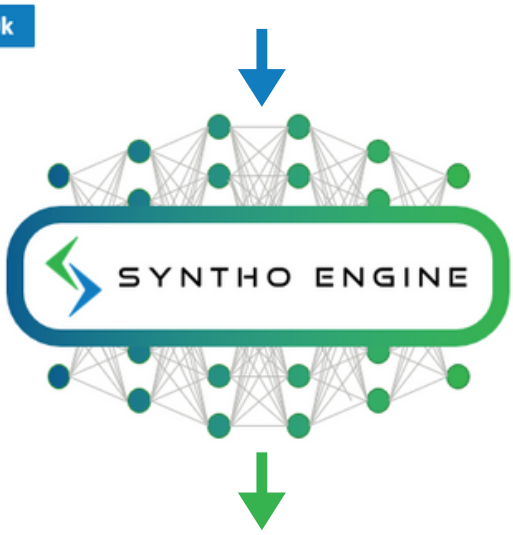
Name	Age	Gender	Item	Price	Data
xxx	25-30	Female	Cloth	€100 - €200	March
xxx	70-75	Male	IT	€600 - €700	March
xxx	40-45	Male	Drink	<€5	March
...	...	...	...	...	...
xxx	40-45	Male	Cloth	€20 - €30	March

N=100k

AI GENERATED SYNTHETIC 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



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

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 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 is able to generate completely new artificially generated datapoints. Hence, there are no privacy risks, because synthetic data is completely new and artificially generated data and individuals simply do not exist anymore.

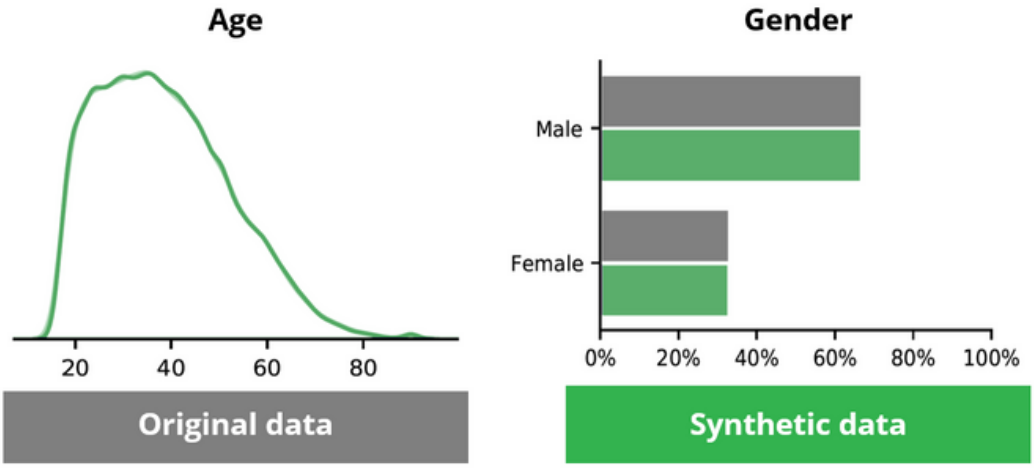
A key difference, we apply AI to model the synthetic data in such a way that we preserve those statistical patterns, relations and characteristics to such an extent that it can even be used for analytics.

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

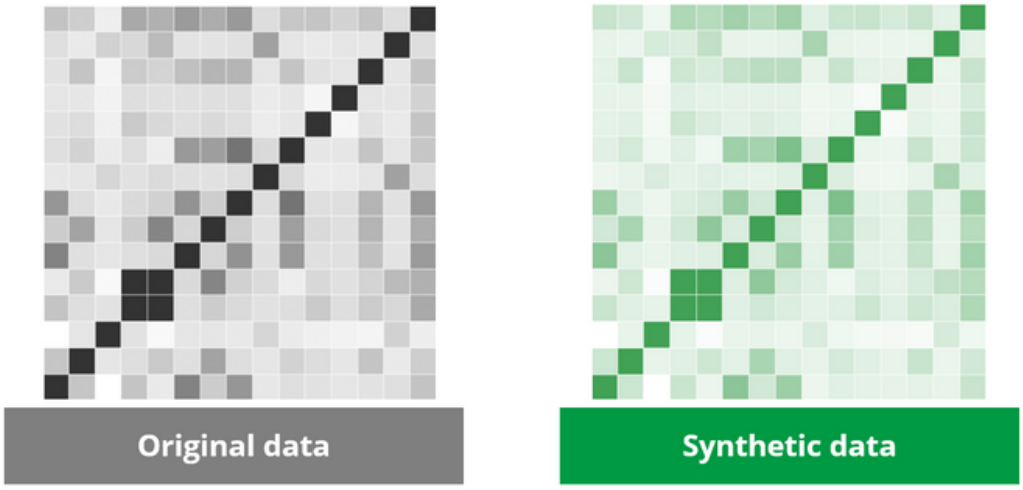
- **as good as real** and statistically identical to the original data,
- **works easy, fast and is scalable**,
- and there is **no privacy risk.**

Our *data quality report* proves this with our comparison of the original data in grey with the synthetic data in green.

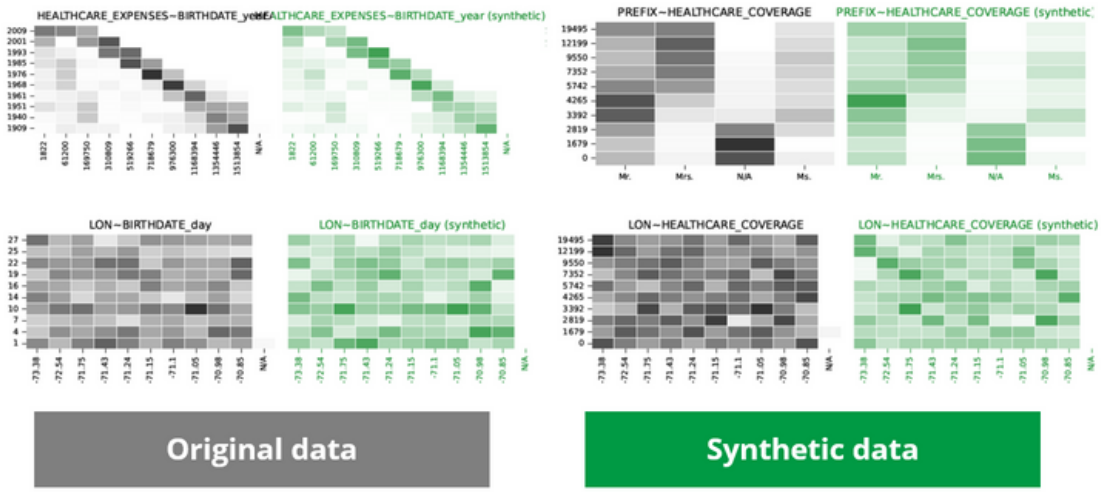
- The distributions, the frequency of variables in the dataset, are similar.



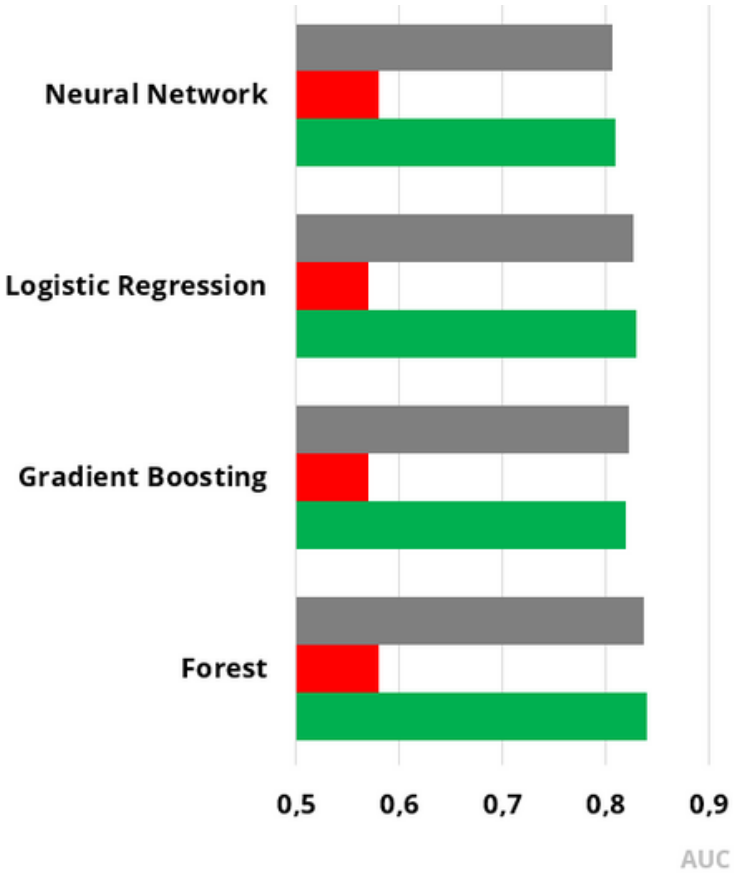
- The correlations, the relationship between variables, are also similar.



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



SAS COLLABORATION



Original data
Anonymized data
Synthetic data

Nevertheless, we are most proud of our collaboration with SAS since they are data experts, who 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 using 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.

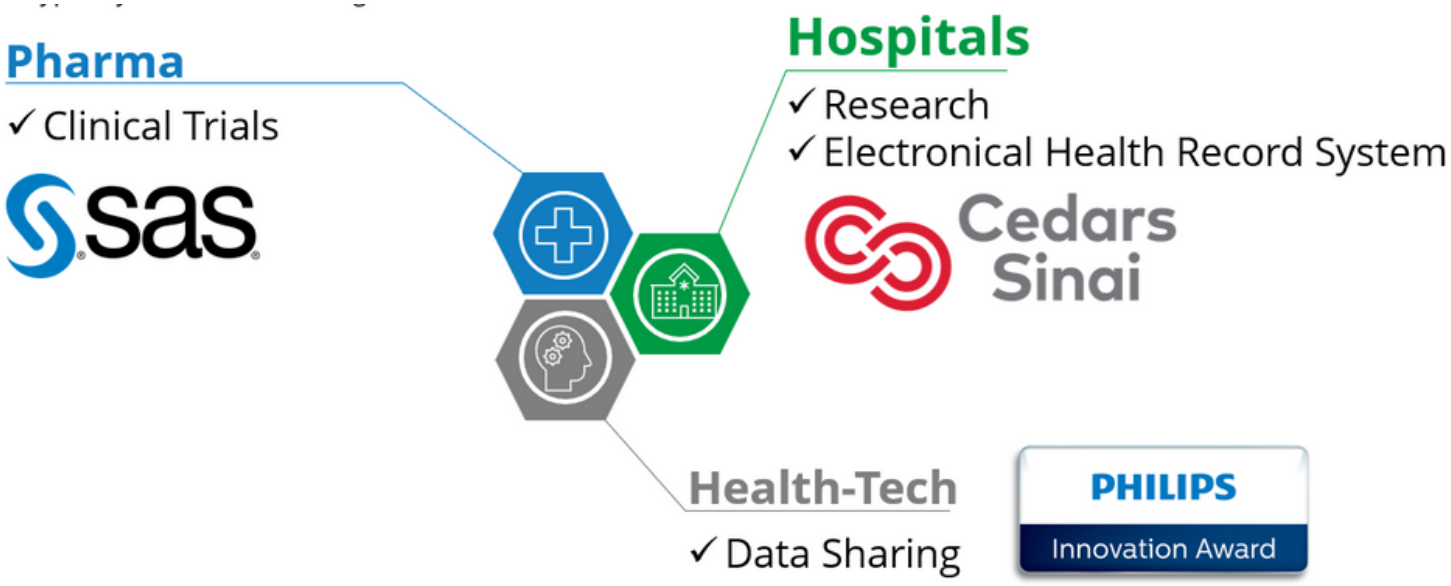
These are the results and conclusions from this assessment:

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

ORGANIZATIONS WE WORK WITH

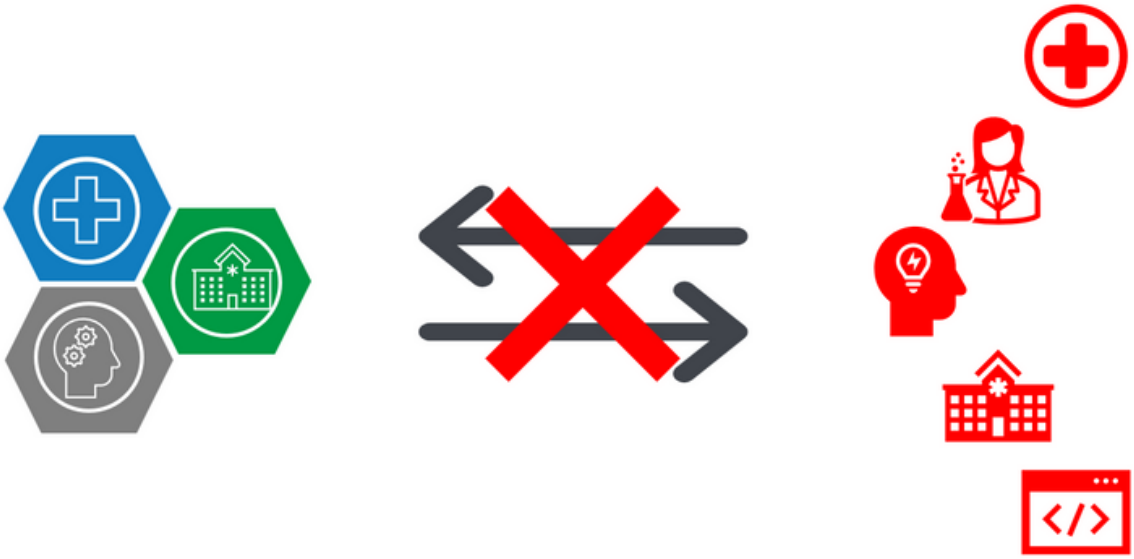
We work with top tier organizations:

- in **pharma**, typically with clinical trials,
- with **hospitals**, typically relating to research or on data from an electronic health record system,
- and with **health-tech** organizations where the focus is on data sharing.



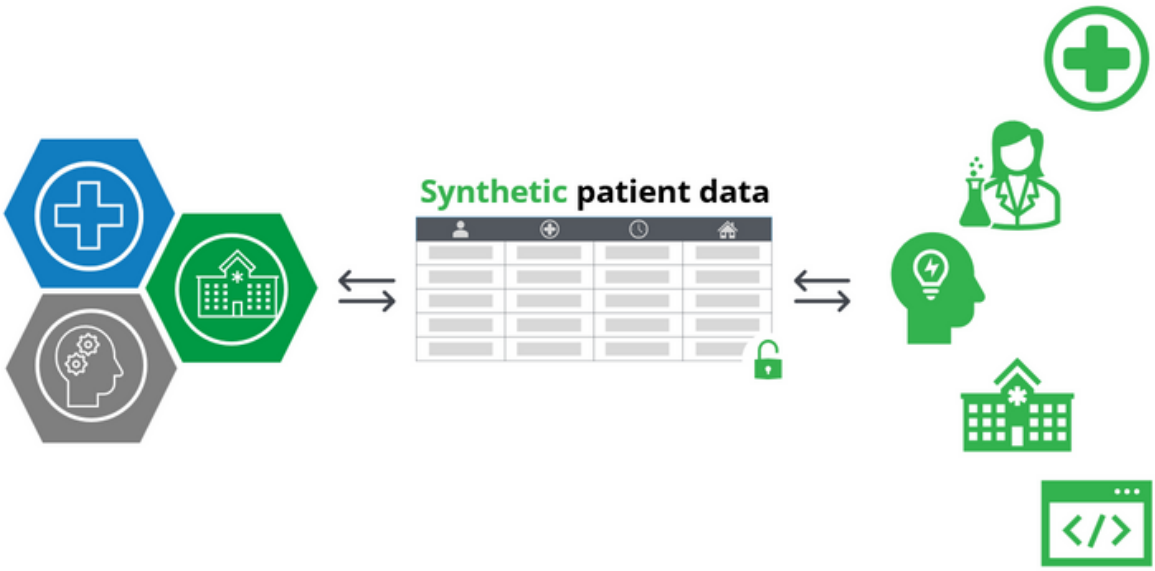
Use, share and sell data is challenging

Highly sensitive data is typically collected by those hospitals, pharmaceuticals and health-tech organizations and cannot simply be used and shared with stakeholders. Consequently, those organizations cannot realize data-driven innovation and they miss data opportunities



Use freely, share and sell synthetic data

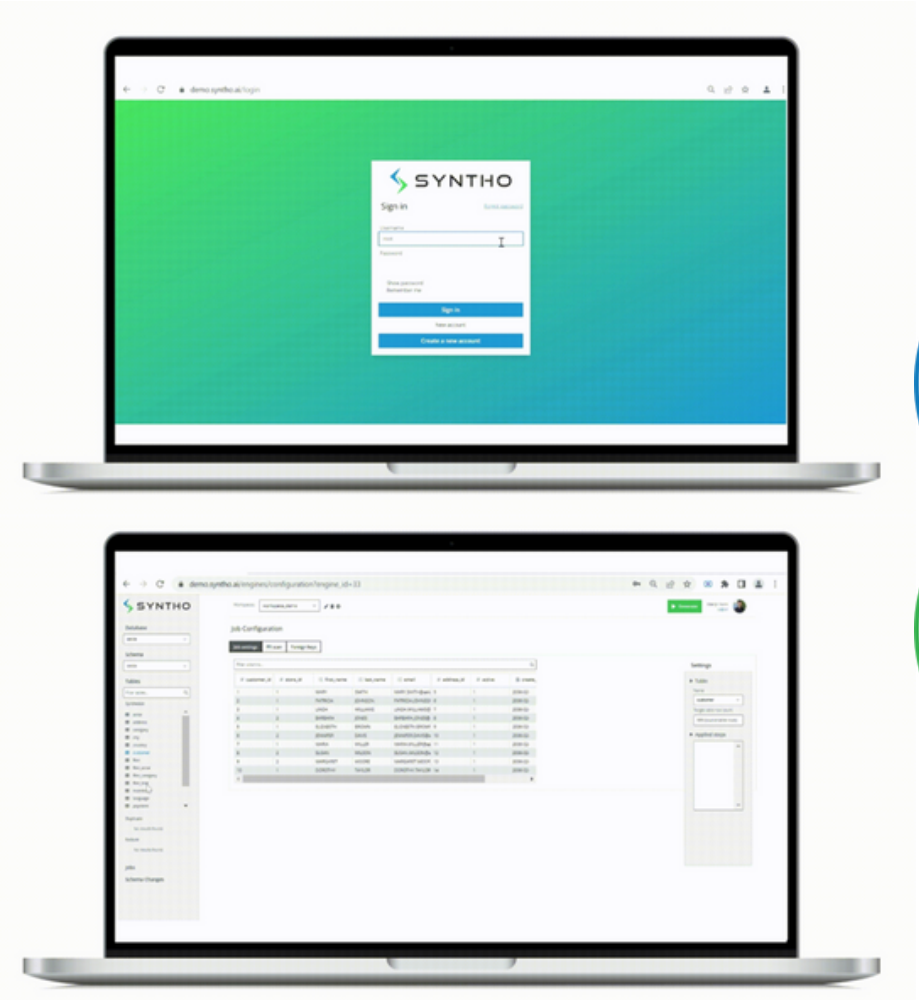
The solution is to share the data in a synthetic form to unlock this data, benefiting for those organizations from less risk, more data and faster data access. After our visit, those organizations can test, develop and innovate based on synthetic data.

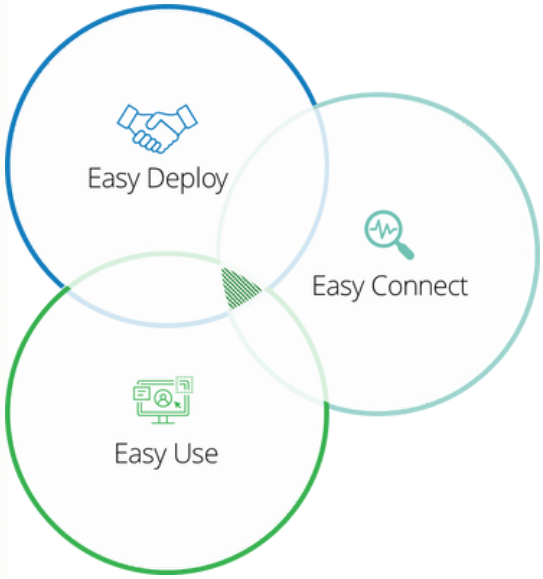


SOLUTION

Hence, we build a self-service synthetic data generation platform and optimized it on 3 axes:

- **Easy deployment** so that we can deploy in the safe environment of the customer.
- **Easy connect** so that we can connect with the source and target data for an end-to-end integrated approach.
- And **easy use** so that anyone can generate and benefit from the value of synthetic data





That's how we are able to unlock that **50%** of data to realize the **4 trillion dollars** of data opportunities.

MORE INFORMATION

Syntho is a data technology organization with a strong expertise in privacy enhancing technologies (PET), headquartered in Amsterdam, Netherlands. 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*)** or visit our website ***www.syntho.ai.***

SYNTHO