# How to start your synthetic data journey?

Your quickstart guide to synthetic data

# Educate your team and all citizen data scientists about data anonymization risks and synthesize wherever possible!

1

## Review current practices of data anonymization

What was sufficient to protect data a few years ago is no longer good enough. Legacy anonymization techniques, like data masking, randomization, pseudonymization, or obfuscation do not protect your data sufficiently.[1] In addition, they are often implemented in a way that doesn't meet regulatory standards. Map your data anonymization practices and make sure you don't have any privacy blindspots![2] Synthetic data doesn't classify as personal data; therefore it is exempt from privacy regulations altogether.

What's more, it's impossible to link synthetic data points to original subjects, since synthetic subjects are AI generated with no 1:1 relationship to the original, hence protecting your customers. Educate your team and all citizen data scientists about data anonymization risks and synthesize wherever possible!

2

## Find the right synthetic data vendor

Although open source synthetic data generators are available, they might come with serious limitations regarding accuracy and privacy. Their performance can be volatile and is highly dependent on the community behind it. Closed source offers more sophisticated capabilities and commercial services you can count on. Choose a vendor with in-depth experience in your industry, capable of augmenting as well as synthesizing data. Demand automated privacy and accuracy quality assurance. Use third party research from trusted sources, such as Gartner and Forrester, to identify viable and robust players.

1 Semantic Web Enabled Record Linkage Attacks on Anonymized Data, Jacob Miracle and Michelle Cheatham http://ceur-ws.org/Vol-1750/paper-03.pdf

2 The Digital Banking Blindspot: Emerging Privacy Enhancing Technologies, Mobey Forum https:// mobeyforum.org/mobey-forum-report-reveals-privacy-blind-spot-in-digital-banking-industry/

3

## Start with tabular data

Gartner recommends starting your synthetic data exploration by synthesizing tabular data.[3] Identify valuable tabular assets, such as transaction data, CRM databases and other valuable, but privacysensitive assets, and create a synthetic data road map. Prioritize easy to implement use cases, such as cross-border data sharing or synthetic test data.

4

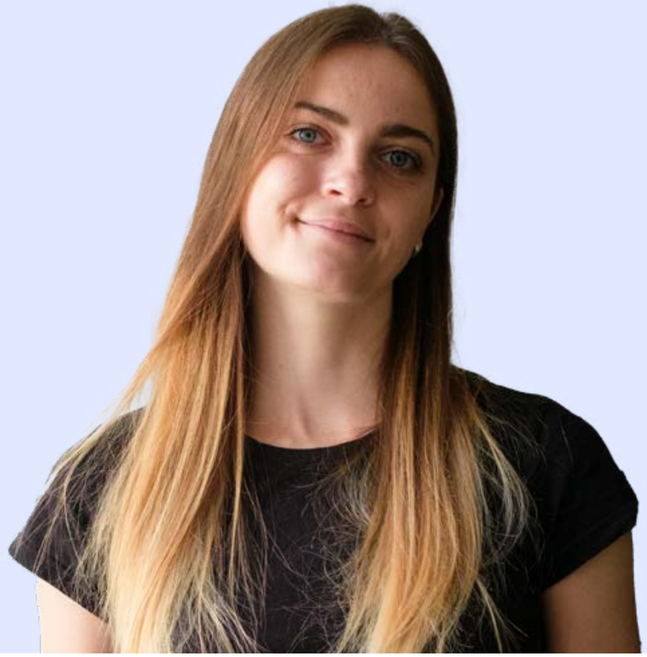
## Set up a synthetic data excellence center

Managing data access requests takes up the majority of time and resources. By setting up a Synthetic Data Excellence Centre, you can provide a quick, painless, compliant, and fully audited process to request synthetic versions of data. The Synthetic Data Excellence Center should also serve as the space where project opportunities are identified by establishing the value chain between the business goal, the data challenge and the synthetic data solution.

5

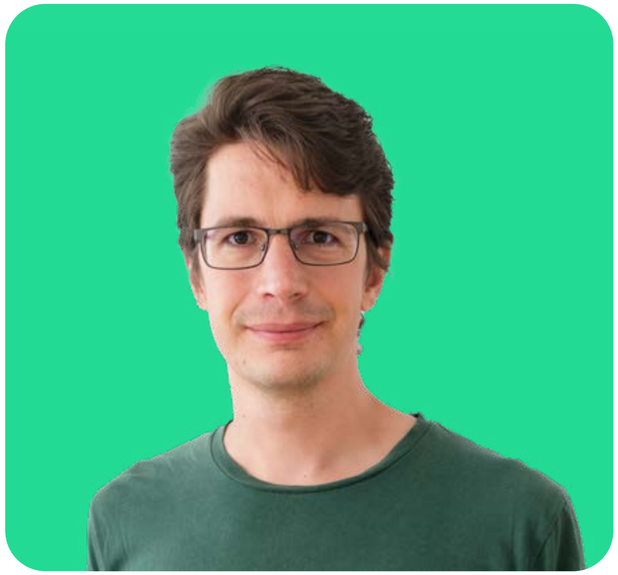
## Create synthetic data lakes

Set up synthetic data lakes to mirror your most valuable and insightful data assets. Colleagues across your organization can use it as a self-service data center to access private-by-design, productionquality data. By making synthetic data flow freely throughout your organization, true data-centricity is born: data-driven decision making and data literacy increases and works in a self-reinforcing fashion.



3 Semantic Web Enabled Record Linkage Attacks on Anonymized Data, Jacob Miracle and Michelle Cheatham http://ceur-ws.org/Vol-1750/paper-03.pdf

# The most important questions and answers about synthetic data for the initiated.



## Notes for your data science team

**What data types can you synthesize?**

MOSTLY AI's platform can synthesize numerical, categorical, datetime, short text (ex. transaction text), and geographic data. All data must be provided in a tabular format.

**Is time series data supported?**

Yes. Time series data is modeled in a two-table setup. The first table, called the subject table, contains unique identifiers. The second table, called the linked table, contains events belonging to a unique identifier. For example, user accounts and their transactions

**How is privacy guaranteed?**

Privacy is built into the generation process in multiple ways. The model uses a random generative process to avoid direct duplicates in the synthetic data. Outlier handling protects column-wise privacy by ensuring that unique values don't occur in the synthetic data.

**How do you guarantee that outliers don't persist in the synthetic data, potentially leaking sensitive information?**

Outliers are handled in two different ways, depending on the data type. For numerical data, any values between the 10th and 90th percentile are clipped away. For categorical data, values appearing more than n times are replaced based on a sliding scale.

**How are the privacy and accuracy metrics defined?**

After each synthetic generation is complete, a custom QA report is generated. We have open-sourced our metrics in a Python library.[4]

**Can synthetic data preserve referential integrity?**

Yes. Generation is done by first generating identifiers, then their associated attributes. This way, keys referenced in the subsequent tables are guaranteed to exist.

4 Virtual Data Lab, https://github.com/mostly-ai/virtualdatalab

**What is the quality of MOSTLY AI's synthetic data for AI/ML model training?**

Always highly accurate. Small fluctuations in the accuracy depend on how much data the model was trained with, how long the model is trained for, and how complex the model becomes. Overall, synthetic data can capture 80%-99% underlying patterns of the original data. We have done extensive research[5] [6] covering the use of synthetic data in ML training. The results are consistently on par or better than training with real data.

5 The World's Most Accurate Synthetic Data Platform? Let's check the Numbers! https://mostly.ai/2020/09/25/the-worlds-most-accurate-synthetic-data-platform/

6 Boost your Machine Learning Accuracy with Synthetic Data https://mostly.ai/2020/08/07/boost-machine-learning-accuracy-with-synthetic-data/

### Can synthetic data reproduce business rules?

MOSTLY AI's synthetic data generator can reproduce business rules implicitly. Due to the random nature of the generative process, small violations may occur. The ability to incorporate hard business rules is an upcoming feature.

### Is there a UI?

MOSTLY AI's synthetic data platform has an intuitive interface, with drag and drop functionality, interactive runtime graphs, and the ability to queue multiple runs. Once a model is created, it is possible to generate more sets of synthetic data without having to wait through training time again.

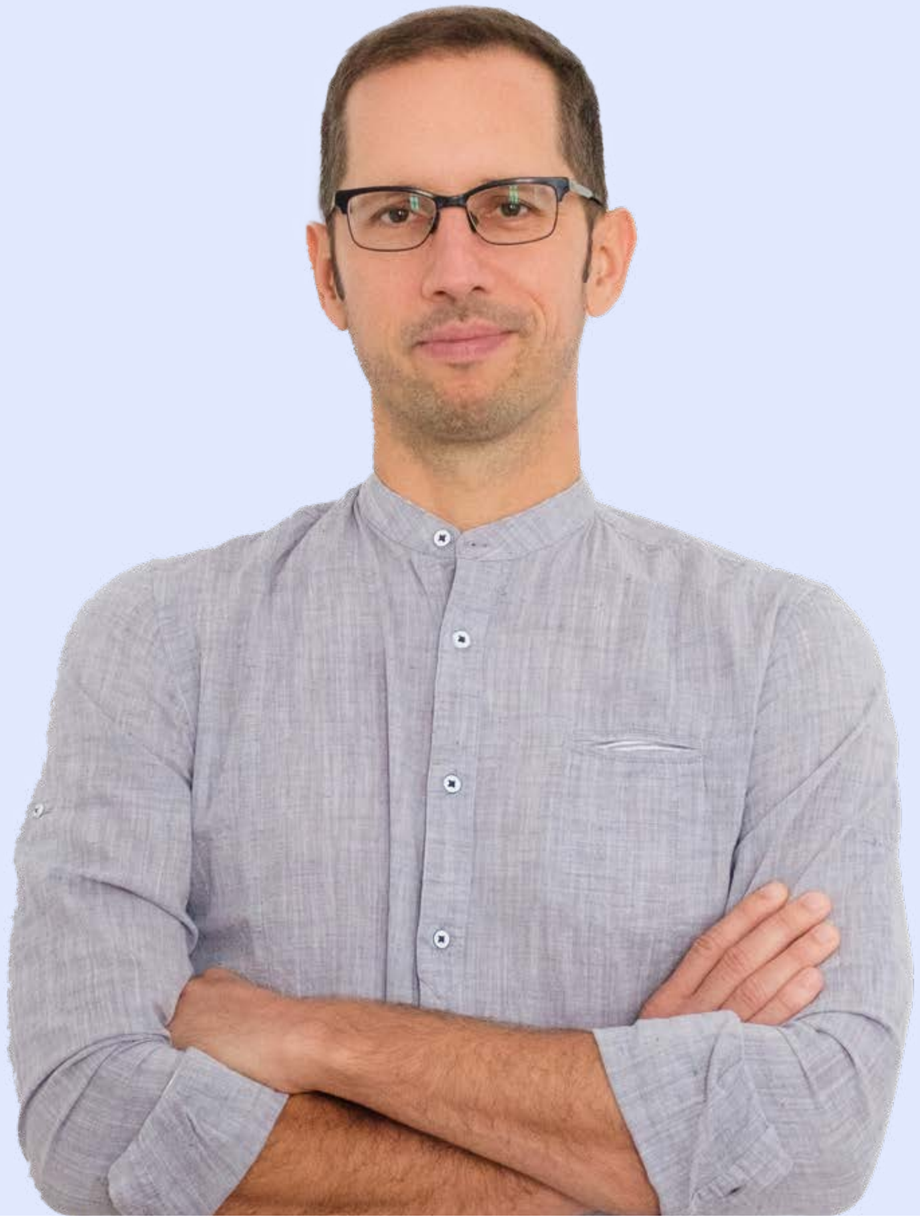
### Is there a non-UI feature that allows automated data pipelines in production?

MOSTLY AI has an API feature and the ability to read from network drives. These two features allow easy integration for automated data pipelines in production

- AP
- Read data from network drives

### Is the application a cloud solution or on-premise installation?

MOSTLY AI's synthetic data platform is available for both an on-cloud and an on-premise installation.



# About MOSTLY AI

**MOSTLY AI is the leading synthetic data company globally. Its platform enables enterprises across industries to unlock, share, fix and simulate data.**

Thanks to the advances in artificial intelligence MOSTLY AI's synthetic data looks and feels just like real data, is able to retain the valuable, granular-level information, yet guarantee that no individual is ever getting exposed. This enables businesses to drive innovation and digital transformation, overcome data silos, improve machine learning models as well as application testing capabilities. MOSTLY AI was founded in 2017 and is headquartered in Vienna, Austria. Its global operation includes customers in a variety of verticals, including banking, insurance and telecommunications.

**Contact:** hello@mostly.ai

**Vienna office (HQ)**
MOSTLY AI Solutions MP GmbH
Hegelgasse 21/3 · 1010 Vienna · Austria

**New York office**
MOSTLY AI Inc.
500 7th Ave 8th floor · New York · NY 10018 · United States

▷ Talk to one of our experts

MOSTLY·AI