ANONOS®

Synthetic Data Generation

Open-source or commercial vendors?

A guide to building vs. buying

Table of Content

Check out pages 19 and 20 for build vs. buy comparison table and decision tree!

Introduction

Synthetic data is expected to completely replace real data in AI models by 2030, according to the famous estimate by Gartner. The market for synthetic data continues to grow. According to Cognilytica, its size will reach $1.15B by 2027, up from $110M in 2021.

The use of synthetic data is growing across many industries, and you may wonder how to get started.

Today, we can divide the synthetic data market into two groups: commercial and open-source solutions. While commercial vendors offer off-the-shelf functionalities, open-source solutions provide toolboxes for custom programming and modifications.

Finding a synthetic data solution to meet your needs isn't easy. The cost of making the wrong decision may be high.

Based on our eleven years of experience in the field, we have compiled a few things to consider when deciding whether to build your own solution with open-source tools or to buy commercial software.

In this guide, we compare key aspects of open-source and commercial synthetic data solutions and analyze them based on a few important elements of a healthy data project.

Here's what you'll learn

- Key differences between open-source and commercial solutions
- An overview of open-source synthetic data tools
- Building vs. buying: What you need to know beforehand

Disclaimer

Anonos is a data privacy technology vendor founded in 2012 to to expand and expedite enterprise data use. The guide draws on our team's eleven years of experience supporting enterprises with data protection to provide the best answer possible.

01

An overview of open-source and commercial synthetic data solutions,

In the current landscape of structured synthetic data generation software, two categories stand out: commercial and open source.

- **Commercial vendors'** software offers platforms and frameworks that plug into your data pipeline and provide synthetic dataset generation and evaluation functionality out-of-the-box.
- **Open-source tools** offer code for synthetic data generation that you can modify, enhance and use to build your own solution.

There are several technological approaches (GANs, VAEs, etc. to generate synthetic data that might influence your choice. We won't cover this aspect in this guide, but you can read more about synthetic data generation methods in this article.

In terms of functionality and services, **most commercial vendors usually offer some form of privacy guarantee** meaning that mechanisms in the synthetic data are meant to prevent the re-identification of an individual from the original data. While open-source solutions generally provide only some additional functionality for assessing privacy and utility, you are still free to **build the functionality you require.**

Commercial vendors offer SaaS, professional services, support, and licensing based on **monthly or annual fees.** Some vendors offer free trials or free plans.

It is mostly **free or low-cost** to use open-source solutions, so they are an attractive option for projects with a smaller budget. Although you get only a limited number of additional services, you can get started with many of the tools using their communities and tutorials.

A list of open-source synthetic data tools

Copulas
USA
Python library for modeling multivariate distributions and sampling from them using copula functions

CTGAN
USA
SDV's collection of deep learning based synthetic data generators for single table data

DataGene
UK
Tool to train, test, and validate datasets, detect and compare dataset similarity between real and synthetic datasets

DoppelGANger
USA
Synthetic data generation framework based on generative adversarial networks (GANs)

DP_WGAN-UCLANESL
This solution trains a Wasserstein generative adversarial network (w-GAN) that is trained on the real private dataset

DPSyn
Algorithm for synthesizing microdata while satisfying differential privacy

Faker
UK
Python package that generates fake data (Note: this tool does not generate synthetic data but offers dummy data)

Generative Adversarial Nets
Repository that shows how to create synthetic time-series data using generative adversarial networks (GANs)

Gretel.ai
USA
Commercial synthetic data vendor that offers open-source functionality

Mimesis
Russia
Fake data generator for Python

mirrorGen
USA
Python tool that generates synthetic data based on user-specified causal relations among features in the data

Open SDP
Community
Open online community for sharing educational analytic tools and resources

> The synthetic data market is expected to reach $1.15B by 2027.
>
> **Cognylitica, 2022**

A list of open-source synthetic data tools

Plait.py

Program for generating fake data from composable yaml templates

Pydbgen
USA

Python package that generates a random database table based on the user's choice of data types

Smart noise synthesizer
USA

Differentially private open-source synthesizer for tabular data

Synner
UAE

Tool to generate real-looking synthetic data by visually specifying the properties of the dataset

Synth
UK

Data-as-code tool that provides a simple CLI workflow for generating consistent data in a scalable way

Synthea
USA

Synthetic patient generator that models the medical history of synthetic patients

Synthetic data vault
USA

Tools for generating synthetic data for tabular, relational, and time series data

TGAN

Generative adversarial training for generating synthetic tabular data

Tofu
UK

Python library for generating synthetic UK Biobank data

Twinify
Finland

Software package for privacy-preserving generation of a synthetic twin to a given sensitive data set

Ydata

Synthetic structured data generator by YData, a commercial vendor

Did we forget to mention a cool open-source synthetic data generation tool?

Let us know

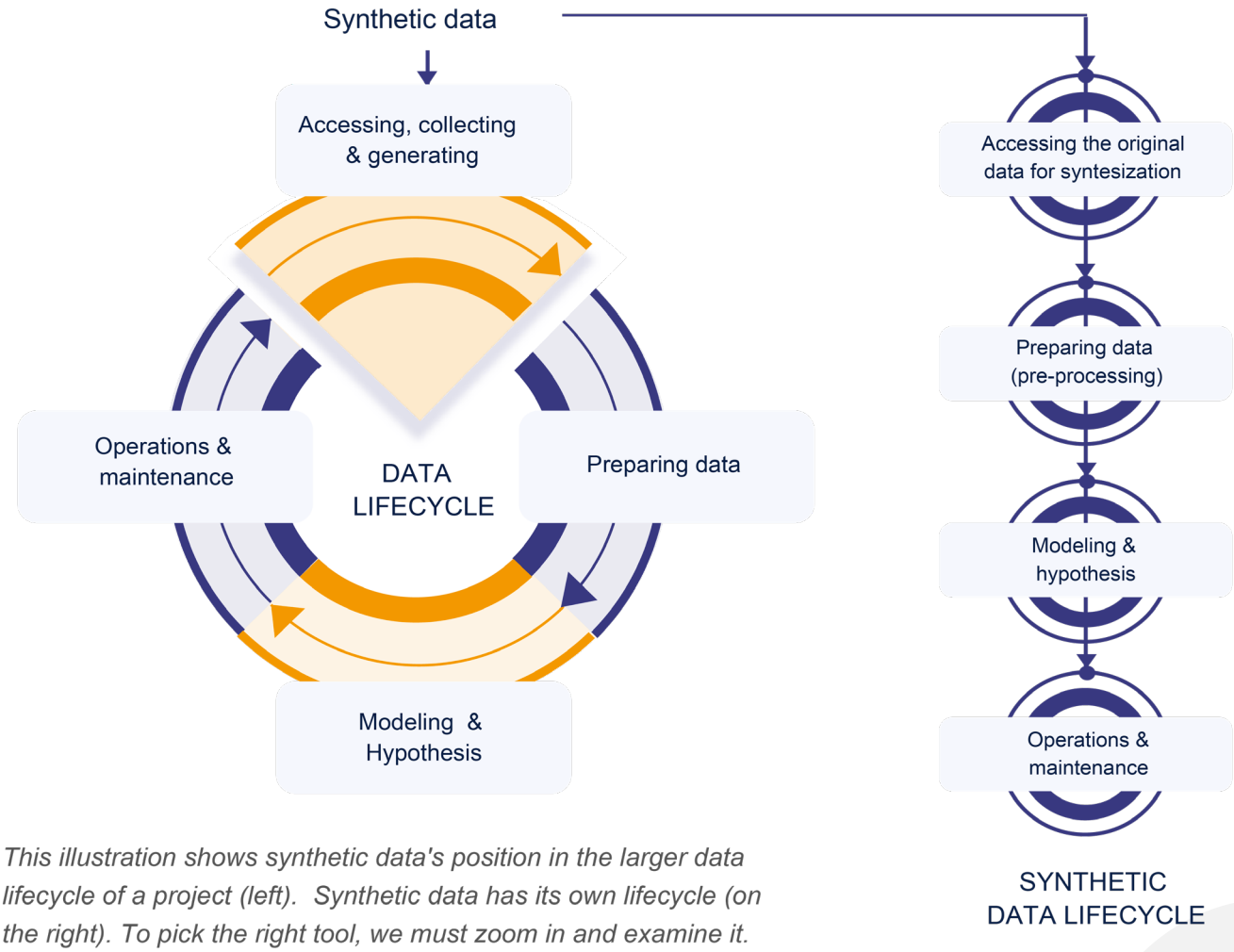
02

Synthetic data project assessment: Criteria for building vs. buying

What aspects should you consider when deciding on building vs. buying?

Whatever the choice, your synthetic data tool will integrate into the larger data lifecycle of the project or use case you are developing. **To pick the right tool, we must zoom in and examine the specific steps of that synthetic data project. The constraints on these steps will help you determine which tool is most appropriate.**

The process of generating and using synthetic data involves different considerations, from getting your first project running to complex privacy assessments. In the next paragraphs, we will focus on the most important aspects.



This illustration shows synthetic data's position in the larger data lifecycle of a project (left). Synthetic data has its own lifecycle (on the right). To pick the right tool, we must zoom in and examine it.

Data access needs

Whether you acquire data externally, gather it internally, or plan to synthesize completely new datasets, access has a big impact on everything that follows.

Quickly accessing and sharing data with stakeholders is often the difference between a successful project and a failure.

How you obtain the data for your synthetic data project will determine what tools you need. For example, is this a one-off test project that you already have the data file for? Is it likely that you will frequently retrieve and save data in your database environment?

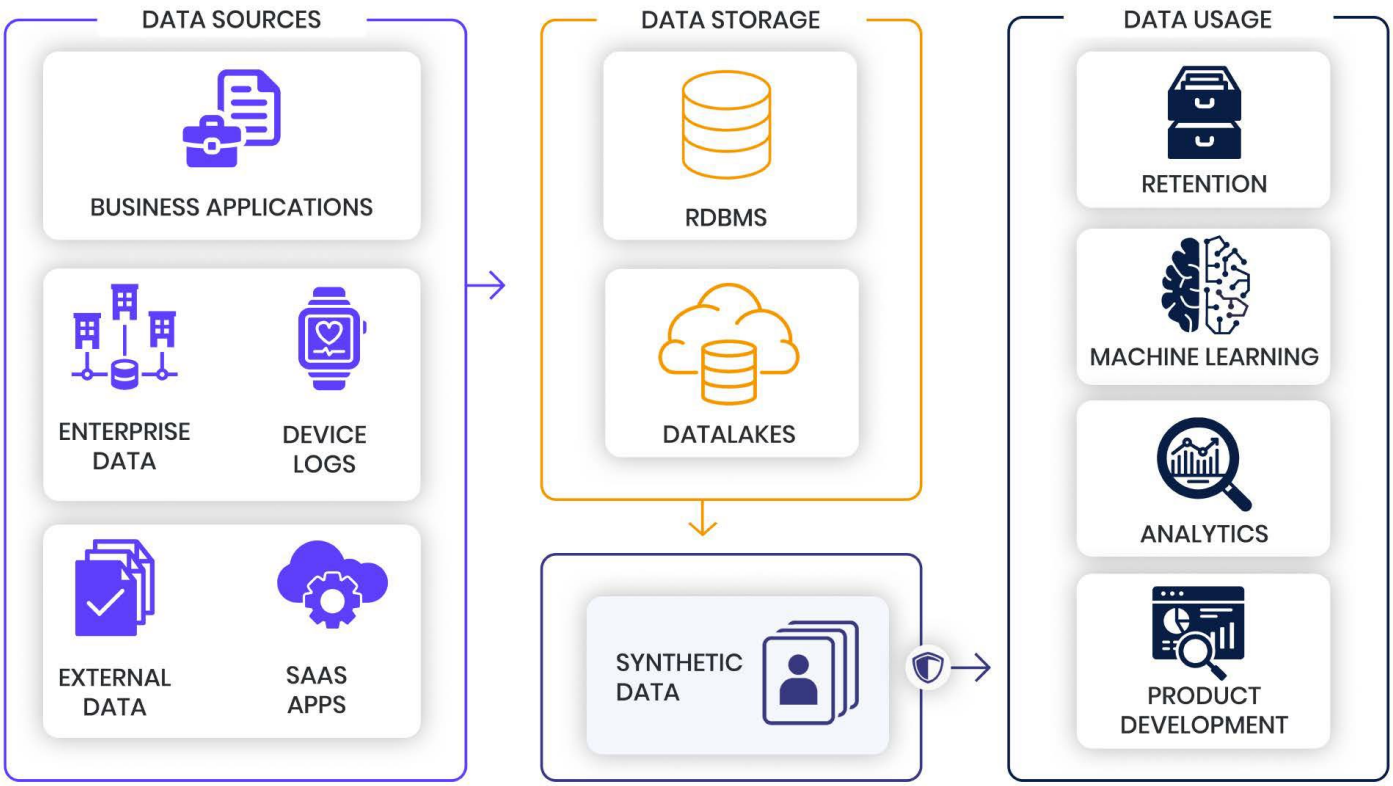
Consider how you will handle potential data access issues now and in the future.

Choose a commercial vendor

Commercial vendors offer plug-and-play solutions with data storage connectors and security functionalities for managing data access levels and roles. **If you work in a big company where typical data projects involve cross-departmental work and/or third parties,** it may be expensive and time-consuming to build open source-based data access capabilities.

Choose an open-source solution

If you are developing or testing **one specific use case that you do not intend to scale**, open source offers you the ability to develop a 100% custom data access functionality that will precisely address your needs without the need to purchase yearly or monthly commercial vendor licenses.



Synthetic data in a typical data pipeline.

Preparing original data

Data preparation is among the most time-consuming and important phases of data projects. To train machine learning applications with synthetic data, for example, data scientists have to prepare the original data for synthesization.

While commercial vendors offer automated pre-processing features, some open-source solutions might require you to prepare your original data for synthesization manually. **The list of pre-processing tasks can quickly grow the more complex your dataset is.**

As part of the data preparation phase, you will also need to specify the rules and requirements that will apply to your synthetic dataset. For example, for driving license datasets, you might want to include only people over the age of 18. The more complex your original dataset is, and the more complicated your rules and dependencies are, the more time it will take to manually set up your custom rules.

Commercial vendors, who work directly with various types of customers and use cases, tend to offer out-of-the-box functionality for a wider range of use cases and data types. Having the necessary support and expertise in a wide range of issues is a big advantage of commercial solutions.

Data preparation is among the most time-consuming and important phases of data projects.

Choose a commercial vendor

For **complex datasets** that require a lot of pre-processing and custom rules. Depending on the vendor, you'll get features to automate pre-processing.

Choose an open-source solution

If you are working with **small and simple datasets**, apply straightforward rules and do not plan to scale your project.

Synthetic data utility and quality assessment

At this stage, data teams analyze and extract valid business perceptions from data. Developing hypotheses and modeling based on synthetic data requires evaluating its quality and utility first.

Data utility refers to the analytical completeness and validity of the data. **This means that synthetic data with a high utility should provide, from an analytical point of view, the closest value to the real-world data it's made from.**

To evaluate your synthetic data's utility, you will need to compare an original and a synthetic dataset on various properties. For example, you can compare their pairwise dependencies, responses to generic evaluations like marginal distributions, use-case-specific evaluations like SQL query counts, or machine learning predictions.

To get good utility from complex datasets, it is often necessary to iterate the process, improving it little by little, adding more (or better tricks to extract utility from data while keeping an eye on privacy.

The utility is also closely tied to your use case. For some use cases, you should check pairwise dependencies, while this metric is irrelevant for others.

You need to determine what utility evaluation metrics are appropriate for your case.

Some commercial vendors provide a catalog of "utility" evaluations you can use out-of-the box regardless of your use case to assess the quality of the synthetic data.

Alternatively, if you only need to maintain a few specific statistical aspects of the original data in your synthetic set, the open-source solution provides those and there is no need to go beyond.

Choose a commercial vendor

Your data complexity will determine how long it takes to validate the results. You'll need to validate your evaluation metrics on several datasets to be confident that they are representative. If that's the case, using commercial vendor tools might be more convenient.

Their utility evaluations can run on datasets of varying complexity, and the output can be adjusted and controlled.

Additionally, some commercial vendor solutions offer ways to assess how the synthetic data performs on your machine learning models.

Choose an open-source solution

Open-source tools are a good solution for simple datasets or if you have the time and expertise to cover a wide range of situations encountered in different datasets. Open-source can be a good option in situations where you require **simple utility verifications and when the complexity of your original dataset is limited.**

From an analytical point of view, synthetic data should provide the closest value to real-world data.

Synthetic data privacy assessment

Another crucial aspect of data access is privacy. To share synthetic data derived from data containing personal information, you need to ensure it can withstand re-identification attacks.

When companies use synthetic data as an anonymization method, a balance must be met between utility and the level of privacy protection.

The biggest question around synthetic data is how to assess the privacy risks. Privacy evaluations are of the utmost importance and are difficult to build internally.

If you want to comply with the GDPR definition of anonymized personal data, the current WP29 recommendation is to test your anonymized dataset against the three known privacy attacks against re-identification:

- Possibility to **single out** an individual in the dataset, meaning some of the records that identify an individual can be isolated from the data.
- Possibility to **link at least two records** that concern the same data subject within a dataset or between two separate datasets. A linkage attack is an attempt at such an event.
- Possibility to **infer information in the dataset**, meaning the values of a set of attributes can be deduced from the data. An attribute inference attack is an attempt at such an event.

Privacy is an empirical field, and without experts, it is hard to calculate the risks, run privacy attacks and get the approval of the DPO (Data Protection Officer. If you are building your own synthetic data solution with open-source tools and need a strong privacy guarantee, we recommend involving data privacy experts to develop and verify the privacy evaluations you need.

Keep in mind that building a privacy evaluation is also time-consuming. Depending on the complexity of the use case, it can take between 3 to 6 months to research, develop, test, and approve synthetic data privacy with/for a DPO.

Anonymeter®

Want to evaluate the privacy of your synthetic data?

Try **Anonymeter**, an open-source tool to comprehensively evaluate the three key indicators of factual anonymization for synthetic data.

CNIL.
COMMISSION NATIONALE INFORMATIQUE & LIBERTÉS

"The results produced by the tool Anonymeter should be used by the data controller to decide whether the residual risks of re-identification are acceptable or not, and whether the dataset could be considered anonymous. Anonymeter is a valuable tool, relevant in the context of personal data protection."

CNIL Technology Experts Department

Choose a commercial vendor

Vendor solutions are recommended if you are **using synthetic data as a privacy mechanism and need strong technical guarantees that it is legally anonymized**. The commercial vendor is the better option when your synthetic datasets are generated based on sensitive original data and need to be shared across multiple departments or external stakeholders. This is because privacy evaluators are already built-in and tested.

Choose an open-source solution

Some open-source solutions offer privacy evaluators. However, those metrics might not be robust enough to provide comprehensive and legally meaningful evaluations of the privacy risks which compliance professionals can understand. **Open source would be a good fit if you don't need to involve DPOs or demonstrate compliance in your project.**

Ease of use

Now, think of who needs access to synthetic data generation processes in your team or company. Sometimes, it is not just data scientists but DPOs, managers, or even CEOs.

For technical users, open-source tools are relatively easy to use. Some open-source solutions have Discord or Slack channels where users can ask questions and solve issues collectively. However, most open-source tools only are a developer toolkit or a library, which isn't suited for non-technical users.

Commercial vendors typically offer ready-to-use platforms with GUIs (graphical user interfaces) and expert support. Non-technical users will be able to autonomously use these platforms to generate or access synthetic datasets.

Additionally, commercial platforms support custom data types and some are tailored to a particular industry, such as healthcare.

Choose a commercial vendor

When you have **a specialized case** like healthcare data. When you are not ready to configure everything yourself or/and other team members need to have access and be able to understand and use the tool.

Choose an open-source solution

When you work with tech-savvy stakeholders and/or **want to have 100% control over the functionality and independence from third-party software.** Go for open-source solutions that have communities around them to get support when needed.

Setup, operations, and maintenance

This step ensures your data project is sustainable. Monitoring the performance ensures that there are no errors or mistakes left for it to work accurately in the future.

The cost of setting up and running your synthetic data project varies greatly.

Creating a unit devoted to synthetic data can cost hundreds of thousands of dollars/euros. On the other hand, if your project is small, free open-source software like SDV may suffice.

Choose a commercial vendor

When you have the budget and plans to scale your data project. Commercial solutions are usually flexible enough to **scale with your project.**

Choose an open-source solution

When you have a **short-term data project** and/or do not plan to grow it.

03

Average cost estimates for building vs. buying

Let's consider the following example to estimate the potential budget needs of a synthetic data project. Suppose you need to run a small synthetic data project without needing extensive (or any) privacy evaluations and DPO approvals. You plan to only use the synthetic data within your data science team and no external stakeholders will be involved.

Let's take a look at the approximate costs of such a project.

Disclaimer

To illustrate the potential costs of a synthetic data project, we have included rough estimates and ballpark figures. These numbers should be used as a guide rather than as a strict blueprint. The actual costs will vary depending on your team size, your infrastructure, needs, and data.

Time and cost to basic functionality

Time and cost to basic functionality allows calculating an initial time and monetary investment for a rather simple synthetic data project.

Building

- For instance, if you are building a synthetic data project for one specific use case, an optimistic estimate to **generate a synthetic dataset of limited complexity** is approximately **30 days** (in our experience) at a loaded cost of **€800 per day** (based on average data scientist salary and technical infrastructure costs).
- For basic functionality, this means an investment of **€24.000 in total.**

Buying

- An out-of-the-box vendor solution for the same type of use case will cost **a minimum of €3.000 per month** (low-average cost based on the price of our own solution and other vendors) for a production license **with basic functionality** (assuming no fixed cost other than the license is paid).
- **A yearly license can cost anywhere between €36.000 (for smaller scale projects, SMEs) and €100.000 and higher (for larger scale projects and larger companies).**

Time and cost to build additional cases

Once you've built your first use case or dataset, the effort to accomplish the same for additional datasets is not negligible. Your budget plan should include costs for building additional use cases.

Building

- Additional cases might not take as much time as the first one. We estimate that it may take about **10 days** to build an additional case at the same loaded rate of **€800 per day**.
- All in all, we're looking at approximately **€8.000 of investment**.

Buying

The commercial vendor also takes time and costs to set up the software. However, once everything is in place, your project can scale up faster. Vendor solutions are often agnostic to the dataset and type of data, while opensource tools might be.

- In an average of **5 days** to set everything up for additional use cases, we'll be looking at **€4.000**.

Maintenance

The next item on the budget is maintenance which includes updates, tweaks, and repairs.

Building

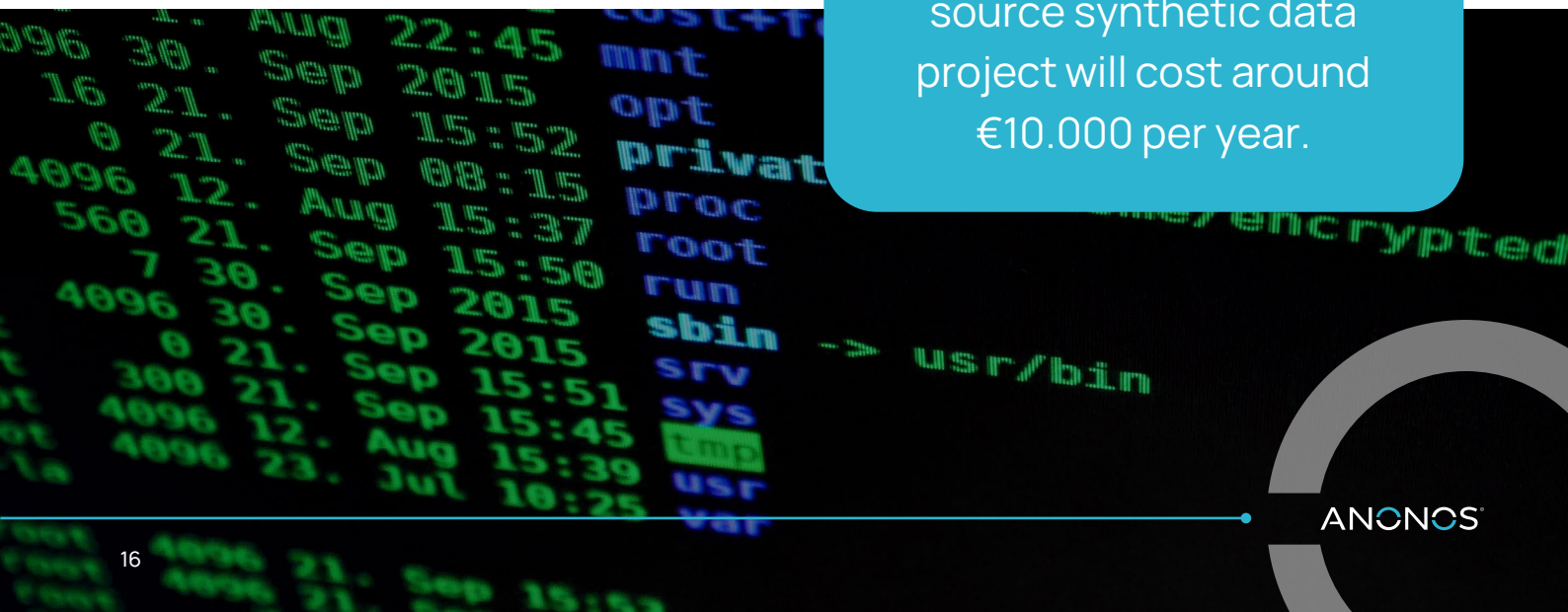
Taking care of maintenance on your own means DIYing or setting up technical support contracts, or assigning a team member to do it.

With the same **€800 loaded cost** and a rough estimate of 1 day per month spent on maintenance, we have **€9.600** for annual maintenance.

Buying

The commercial vendor takes care of the software maintenance as long as you pay for the license.

One day per month of maintenance for an open-source synthetic data project will cost around €10.000 per year.



ANONOS

Support

With technical support, reviewing documentation, logging tickets, investigating, and fixing bugs can be done in-house by your own team, but keep in mind that it will require regular budget allocations. The bigger your project gets, the more support it will need.

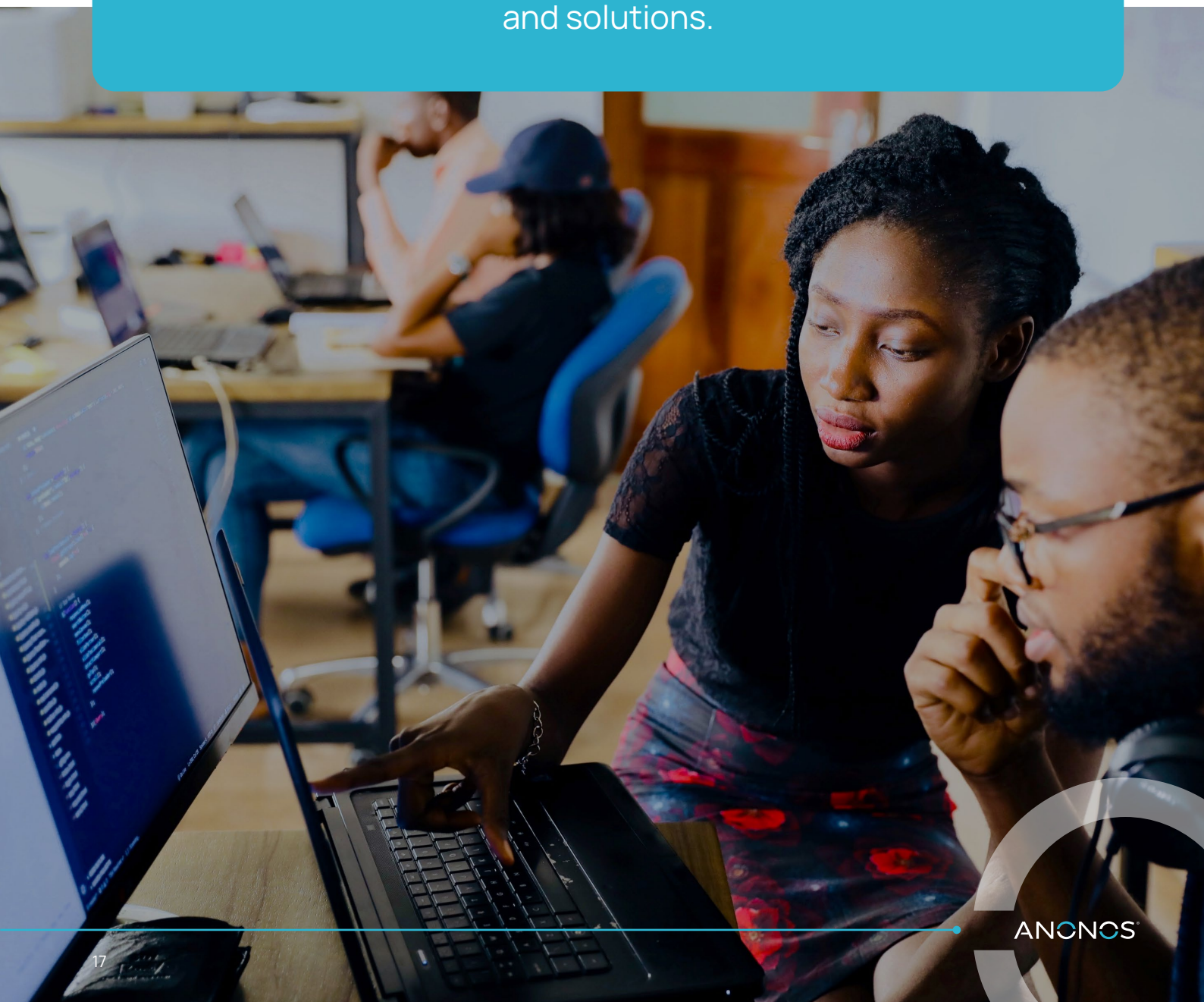
Building

- Open-source tools often have **communities where users share their issues and offer ideas and solutions.**

Buying

- In the case of a commercial solution, you can usually **rely on the vendor to ensure the product is running smoothly** and advise you on any technical issues.

Open-source synthetic data tools often have online communities where users can share their problems and solutions.



Summary

Open-source synthetic data tools will be an excellent choice if you:

- Have **a one-shot**, specific data project(s) that you are **not planning to scale**.
- Want to get into the ins and outs of technology and **experiment with synthetic data generation.**
- Would benefit from having **access to a strong community** of other users.
- Have the necessary **technical knowledge** and can allocate **resources** to development, maintenance, and optimization.
- Have relatively **simple datasets** that would make evaluating the quality of the output easier.
- **Don't need** to go through thorough **privacy/compliance checks**.
- **Develop an in-house synthetic data solution** and want to use open source code elements.

Commercial synthetic data solutions will be a good fit if you:

- Need to involve **external stakeholders and share access** to your synthetic data project with them.
- Work with **highly sensitive data**, need to ensure the privacy of synthetic data and run it by a DPO.
- Work with **complex datasets and need utility evaluations** compatible with datasets of varying nature and complexity.
- **Do not have resources for maintenance** and support of your synthetic data tool.
- Need expert support and **graphical user interfaces**.
- Need a tool suitable for **non-technical users**.
- Want to **scale your synthetic data projects** and use cases.
- Need to **get started on synthetic data quickly**

An alternative: Taking a hybrid approach

Sometimes you **don't have to choose between open-source and commercial vendors** because you can take advantage of both with a hybrid approach.

For instance, you're happy with what you've built in-house, but your project scales. It may be necessary to perform extensive privacy evaluations in order to share your synthetic data. In this case, you can perform privacy assessments using scientifically proven commercial vendor functionality.

You can also use vendor expertise and services when you need expert help building your own tool.

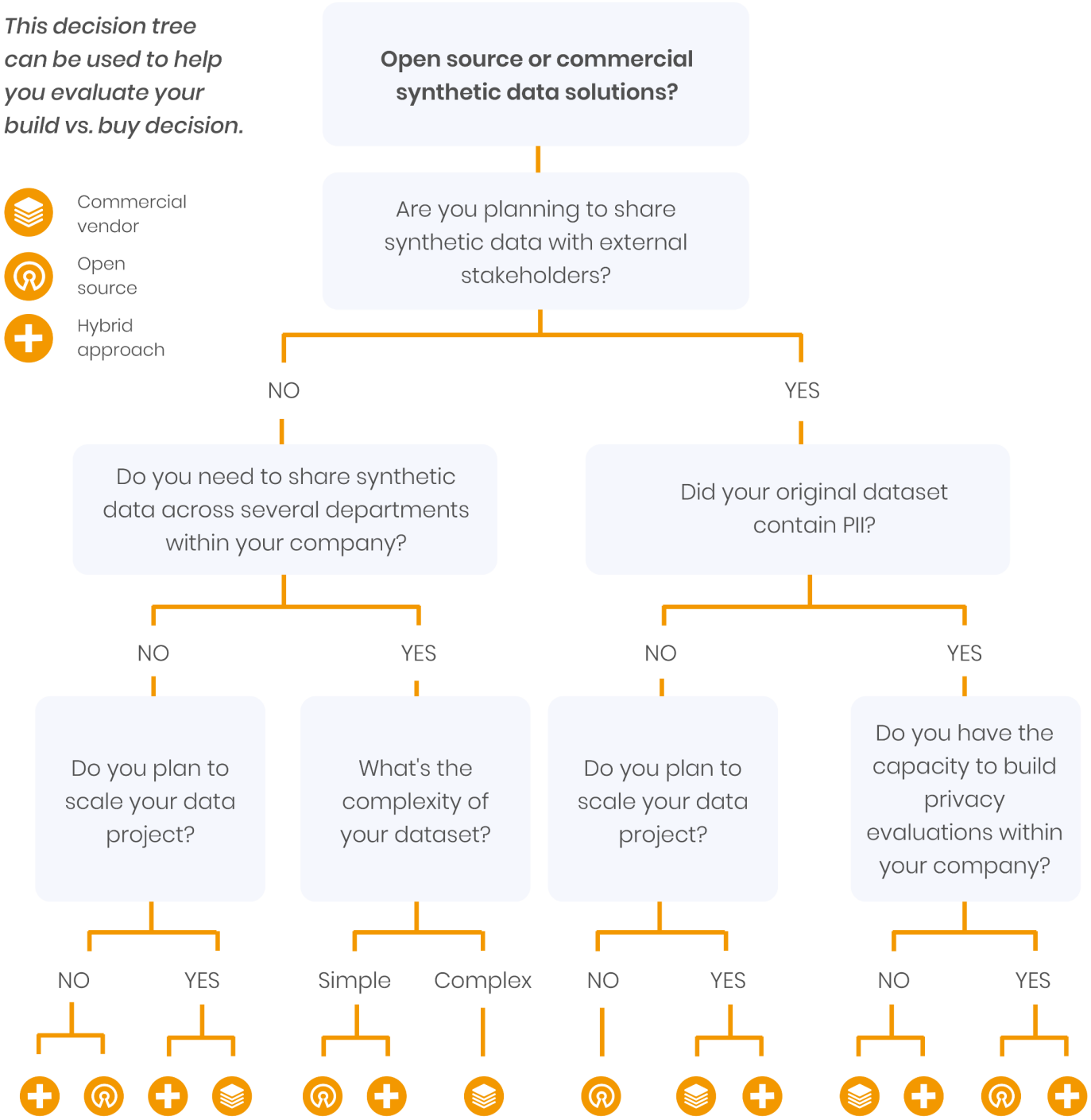
Conclusions

Commercial vendors and open-source tools both have amazing features. The specifics and goals of your data project will determine which category suits you best. When making your decision, consider your project's use case, its complexity, your stakeholders' needs, maintenance budget, team capacity, and, of course, the security of your sensitive data.

	Built with the open- source tool	Buy from the commercial vendor
Data access	Build custom data access & privacy assessment functionality yourself. Good for **testing specific use case** with no intention to scale; **sharing the output only within your team**.	Plug-and-play functionality for managing data access levels. Built-in privacy evaluators. Good for **large companies**; cross-departmental work and/or **third party involvement**.
Preparing original data	Some solutions might require to prepare data for synthesization manually. Good for working with **small & simple datasets**; applying **straightforward rules**; no plans to scale.	Automated pre-processing, functionality for a wider range of use cases & data types. Good for **complex datasets** that require a lot of preprocessing & **custom rules**.
Synthetic data utility and quality assessment	Simple utility verifications. Good for original datasets of **limited complexity**	Utility assessment for datasets of varying complexity. The output can be adjusted and controlled. Good for **complex datasets**; when utility validation needs to be done on **several datasets**; adjusting and **controlling the output**.
Synthetic data privacy assessment	Some privacy evaluations may be present. Use when you **don't need to involve DPOs** (Data Protection Officers or **demonstrate compliance** in your project.	Privacy evaluators are already built-in and tested. Good for using synthetic data **as a privacy mechanism**; strong **technical guarantees** that the synthetic data is **legally anonymized**; ensuring privacy for synthetic datasets that are generated **based on sensitive original data**.
Time & cost to basic functionality	**30 days** at a loaded cost of **€800 per day**. **€24.000 total**.	A minimum of **€3.000 per month** for a production license. **€36.000 - €100.000 & higher yearly.**
Time & cost to build additional cases	**10 days** to build an additional case at a loaded rate of **€800 per day**. **€8.000 total.**	An average of **5 days** to set everything up for additional use cases. **€4.000 total.**
Maintenance	**One day per month at €800** loaded rate. **€9.600 per year.**	Software maintenance **included in the license fees**. No additional cost.
Support	**Do it yourself**. Review docs, log tickets, investigate bugs.	**Basic support included** in license fees. Advanced support **may require additional costs**.
Ease of use	Relatively **easy to use for technical users**; community & tutorials to help you get started; DIY for specialized use cases. Need to configure everything yourself. Good for **technical users**; simple datasets and straightforward use cases.	Well-documented, **multiple user interfaces** (Web, HTTP & Python API and CLI), expert **support for custom data types** & for extending the functionality. Good for **specialized cases** like healthcare data; **nontechnical users**; getting **started quickly**.

Build vs. buy decision tree

This decision tree can be used to help you evaluate your build vs. buy decision.



ANONOS

About Anonos

Anonos is an innovator in enterprise data privacy, security and enablement.

Our globally patented, award-winning Data Embassy software digitizes and technologically enforces privacy and security policies, so organizations can safely optimize their sensitive data assets across departments or around the globe. The solution uses a combination of Privacy-Enhancing Technologies to transform source data into lawful variations that can travel everywhere, thanks to the embedded technical controls that ensure compliance.

From testing through production, we provide enterprises with full-spectrum data protection for all their use cases regardless of environment for faster speed to insight and value.

Anonos. Data without the drama.

In 2022, Anonos acquired Berlin-based Statice GmbH, a synthetic data software provider.

CONTACT US

Follow us

ANONOS®