**Synthetic Data Generation: Definition, Types, Techniques, and Tools**

Synthetic data can be defined as artificially annotated information. It is generated by computer algorithms or simulations. Synthetic data generation is usually done when the real data is either not available or has to be kept private because of personally identifiable information (PII) or compliance risks. It is widely used in the health, manufacturing, agriculture, and eCommerce sectors.

In this article, we will learn more about synthetic data, synthetic data generation, its types, techniques, and tools. It will provide you the knowledge required to help in producing synthesized data for solving data-related issues.

Table of Contents

[1. What is synthetic data?](https://www.turing.com/kb/synthetic-data-generation-techniques#what-is-synthetic-data?)

[2. Why is synthetic data required?](https://www.turing.com/kb/synthetic-data-generation-techniques#why-is-synthetic-data-required?)

[3. Synthetic data generation](https://www.turing.com/kb/synthetic-data-generation-techniques#synthetic-data-generation)

[3.1. Real data vs synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#real-data-vs-synthetic-data)

[3.2. Advantages of synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#advantages-of-synthetic-data)

[3.3. Characteristics of synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#characteristics-of-synthetic-data)

[3.4. Uses of synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#uses-of-synthetic-data)

[4. Types of synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#types-of-synthetic-data)

[5. Varieties of synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#varieties-of-synthetic-data)

[6. Synthetic data generation methods](https://www.turing.com/kb/synthetic-data-generation-techniques#synthetic-data-generation-methods)

[6.1. Based on the statistical distribution](https://www.turing.com/kb/synthetic-data-generation-techniques#based-on-the-statistical-distribution)

[6.2. Based on an agent to model](https://www.turing.com/kb/synthetic-data-generation-techniques#based-on-an-agent-to-model)

[6.3. Using deep learning](https://www.turing.com/kb/synthetic-data-generation-techniques#using-deep-learning)

[7. Synthetic data generation tools](https://www.turing.com/kb/synthetic-data-generation-techniques#synthetic-data-generation-tools)

[8. Generating synthetic data using Python-based libraries](https://www.turing.com/kb/synthetic-data-generation-techniques#generating-synthetic-data-using-python-based-libraries)

[9. Challenges and limitations while using synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#challenges-and-limitations-while-using-synthetic-data)

[10. Real-world applications using Synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#real-world-applications-using-synthetic-data)

[11. Future of synthetic data](https://www.turing.com/kb/synthetic-data-generation-techniques#future-of-synthetic-data)

[12. Wrapping up](https://www.turing.com/kb/synthetic-data-generation-techniques#wrapping-up)

What is synthetic data?

Synthetic data is information that is not generated by real-world occurrences but is artificially generated. It is created using algorithms and is used to test the dataset of operational data. This is mainly used to validate mathematical models and train the synthetic data for [deep learning models](https://www.turing.com/kb/making-deep-learning-models-generalize-better).

The advantage of synthetic data usage is that it reduces constraints when you use regulated or sensitive data. And creates the data requirements as per specific requirements which can’t be attained with authentic data. Synthetic datasets are usually generated for quality assurance and software testing.

The disadvantage of synthetic data includes inconsistencies that take place while you try and replicate the complexity found within the original data and its inability for replacing authentic data straightforwardly because you will still need accurate data for producing useful results.

Why is synthetic data required?

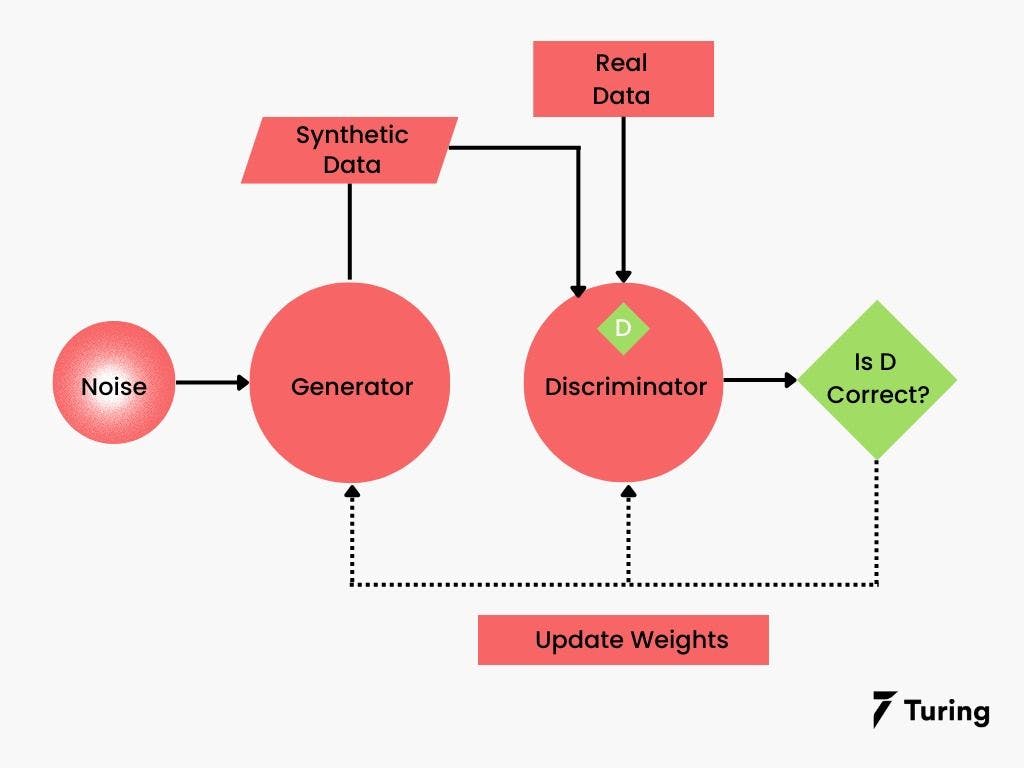
For three main reasons, synthetic data can be an asset to businesses for privacy concerns, faster turnaround for product testing, and training machine learning algorithms. Most data privacy laws restrict businesses in the way they handle sensitive data.

Any leakage and sharing of personally identifiable customer information can lead to expensive lawsuits that also affect the brand image. Hence, minimizing privacy concerns is the top reason why companies invest in synthetic data generation methods.

For entirely new products, data usually is unavailable. Moreover, human-annotated data is a costly and time-consuming process. This can be avoided if companies invest in synthetic data, which can instead be quickly generated and help in developing reliable machine learning models.

Synthetic data generation

A process in which new data is created by either manually using tools like Excel or automatically using computer simulations or algorithms as a substitute for real-world data is called synthetic data generation.



This fake data can be generated from an actual data set or a completely new dataset can be generated if the real data is unavailable. The newly generated data is nearly identical to the original data. Synthetic data can be generated in any size, at any time, and in any location.

Although it is artificial, synthetic data mathematically or statistically replicates real-world data. It is similar to the real data that is collected from actual objects, events, or people for training an AI model.

Real data vs synthetic data

Real data is gathered or measured in the actual world. Such data is created every instant when an individual uses a smartphone, a laptop, or a computer, wears a smartwatch, visits a website, or makes a purchase online. These data can also be generated through surveys (online and offline).

Synthetic data, on the contrary, is generated in digital environments. These data are fabricated in a way that successfully imitates the actual data in terms of basic properties, except for the part that was not acquired from any real-world occurrences.

With various techniques to generate synthetic data, the training data required for machine learning models are available easily, making the option of synthetic data highly promising as an alternative to real data. However, it cannot be stated as a fact whether synthetic data can be an answer to all real-world problems. This does not affect the significant advantages that synthetic data has to offer.

Advantages of synthetic data

Synthetic data has the following benefits:

Customizable: It is possible to create synthetic data to meet the specific needs of a business.

Cost-effective: Synthetic data is an affordable option compared to real data. For instance, real vehicle crash data for an automotive manufacturer will be more expensive to obtain than to create synthetic data.

Quicker to produce: Since synthetic data is not captured from real-world events, it is possible to generate as well as construct a dataset much faster with suitable tools and hardware. This means that a huge volume of artificial data can be made available in a shorter period of time.

Maintains data privacy: Synthetic data only resembles real data, but ideally, it does not contain any traceable information about the actual data. This feature makes the synthetic data anonymous and good enough for sharing purposes. This can be a boon to healthcare and pharmaceutical companies.

Characteristics of synthetic data

[Data scientists](https://www.turing.com/jobs/remote-data-scientist) aren't concerned about whether the data they use is real or synthetic. The quality of the data, with the underlying trends or patterns, and existing biases, matters more to them.

Here are some notable characteristics of synthetic data:

Improved data quality: Real-world data, other than being difficult and expensive to acquire, is also likely to be vulnerable to human errors, inaccuracies, and biases, all of which directly impact the quality of a machine learning model. However, companies can place higher confidence in the quality, diversity, and balance of the data while generating synthetic data.

Scalability of data: With the increasing demand for training data, data scientists have no other option but to opt for synthetic data. It can be adapted in size to fit the training needs of the machine learning models.

Simple and effective: Creating fake data is quite simple when using algorithms. But it is important to ensure that the generated synthetic data does not reveal any links to the real data, that it is error-free, and does not have additional biases.

Data scientists enjoy complete control over how synthetic data is organized, presented, and labeled. That indicates that companies can access a ready-to-use source of high-quality, trustworthy data with a few clicks.

Uses of synthetic data

Synthetic data finds applicability in a variety of situations. Sufficient, good-quality data remains a prerequisite when it comes to machine learning. At times, access to real data might be restricted due to privacy concerns, while at times it might appear that the data isn't enough to train the machine learning model.

Sometimes, synthetic data is generated to serve as complementary data, which helps in improving the machine learning model. Many industries can reap substantial benefits from synthetic data:

Banking and financial services

Healthcare and pharmaceuticals

Automotive and manufacturing

Robotics

Internet advertising and digital marketing

Intelligence and security firms

Types of synthetic data

While opting for the most appropriate method of creating synthetic data, it is essential to know the type of synthetic data required to solve a business problem. Fully synthetic and partially synthetic data are the two categories of synthetic data.

Fully synthetic data does not have any connection to real data. This indicates that all the required variables are available, yet the data is not identifiable.

Partially synthetic data retains all the information from the original data except the sensitive information. It is extracted from the actual data, which is why sometimes the true values are likely to remain in the curated synthetic data set.

Varieties of synthetic data

Here are some varieties of synthetic data:

Text data: Synthetic data can be artificially generated text in [natural language processing](https://www.turing.com/kb/natural-language-processing-understanding-analyzing-generating-text-with-python) (NLP) applications.

Tabular data: Tabular synthetic data refers to artificially generated data like real-life data logs or tables useful for classification or regression tasks.

Media: Synthetic data can also be synthetic video, image, or sound to be used in [computer vision](https://www.turing.com/kb/all-you-need-to-know-about-computer-vision) applications.

Synthetic data generation methods

For building a synthetic data set, the following techniques are used:

Based on the statistical distribution

In this approach, you have to draw numbers from the distribution by observing the real statistical distributions, similar factual data should be reproduced. In some situations where real data is not available, you can make use of this factual data.

If a data scientist has a proper understanding of the statistical distribution in real data, he can create a dataset that will have a random sample of distribution. And this can be achieved by the normal distribution, chi-square distribution, exponential distribution, and more. The trained model’s accuracy is heavily dependent on the data scientist’s expertise in this method.

Based on an agent to model

With this method, you can create a model which will explain observed behavior, and it will generate random data with the same model. This is fitting actual data to the known distribution of data. Businesses can use this method for synthetic data generation.

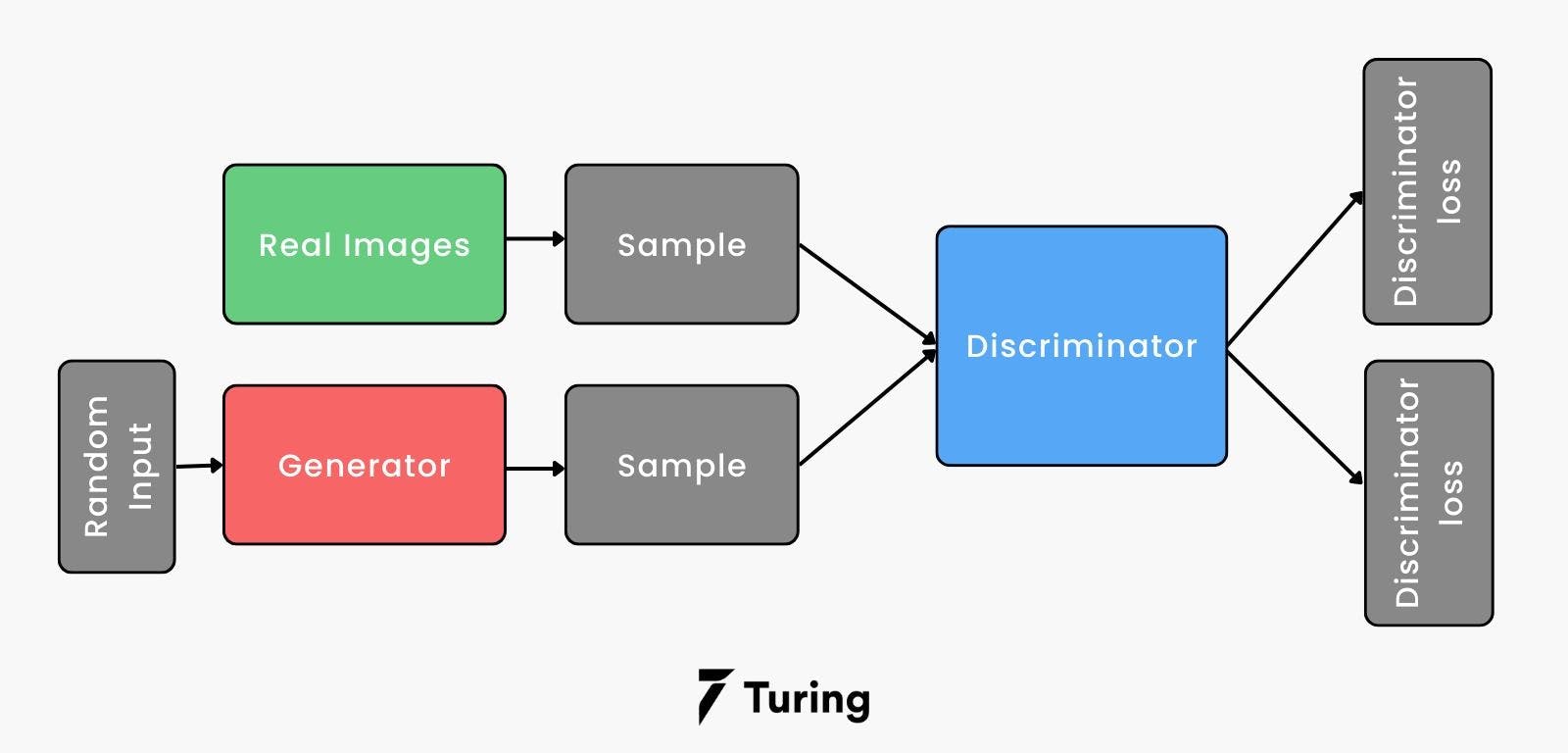
Apart from this, other machine learning methods can be used to fit the distributions. But, when the data scientist wants to predict the future, the decision tree will overfit because of the simplicity and going up to full depth.

Also, in certain cases, you can see that a part of the real data is available. In such a situation, businesses can use a hybrid approach to build a dataset based on statistical distributions and generate synthetic data using agent modeling based on real data.

Using deep learning

The use of deep learning models which will employ a Variational autoencoder or Generative Adversarial Network model uses methods for generating synthetic data.

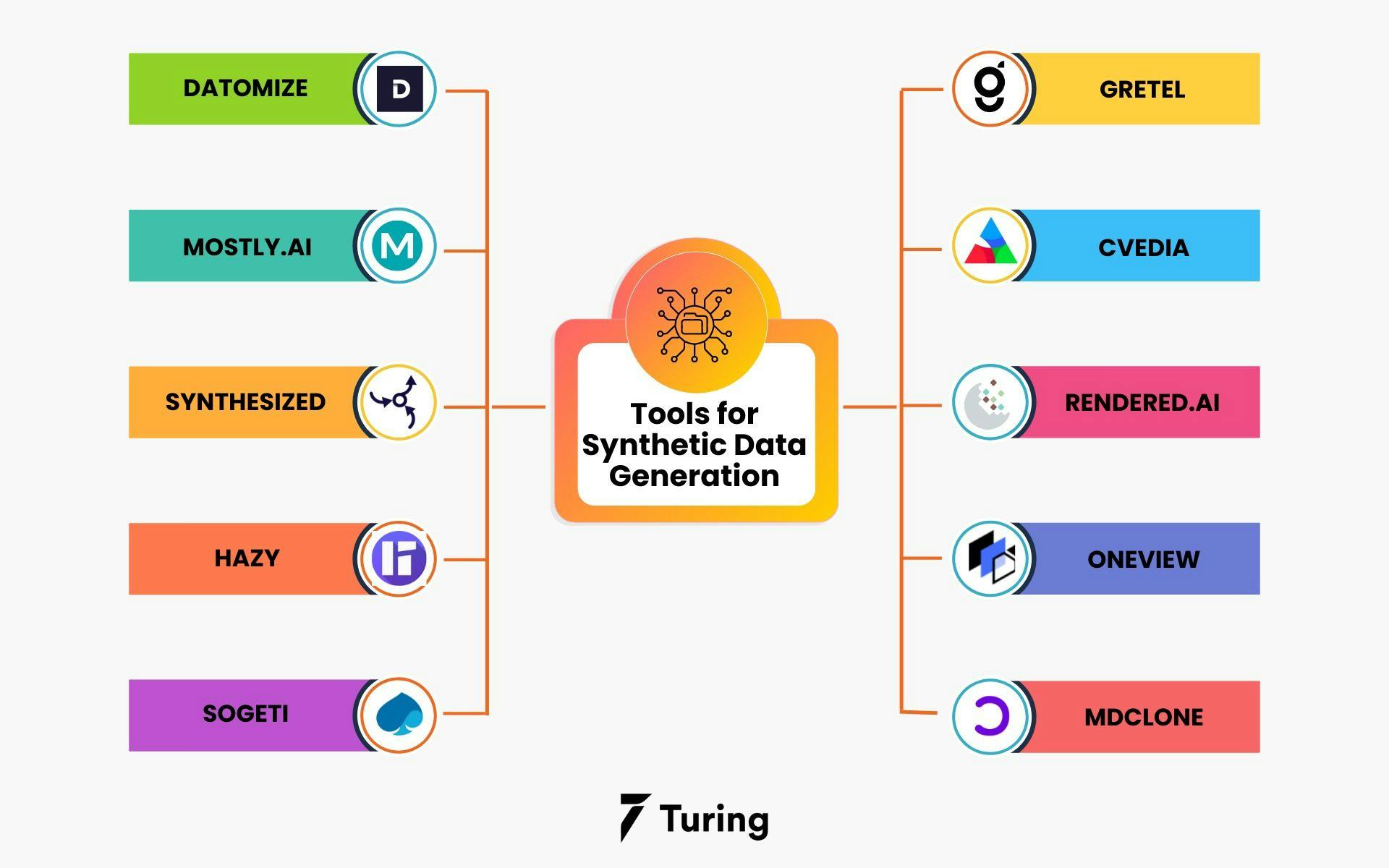
VAEs are unsupervised machine learning model types that contain encoders to compress and compact the actual data while the decoders analyze this data for generating a representation of the actual data. The vital reason for using VAE is to ensure that both input and output data remain extremely similar.



[GAN models and adversarial networks](https://www.turing.com/kb/generative-models-vs-discriminative-models-for-deep-learning) are two competing neural networks. GAN is the generator network that is responsible for creating synthetic data. An adversarial network is the discriminator network, which functions by determining a fake dataset and the generator is notified about this discrimination. The generator will then modify the next batch of data. In this way, the discriminator will improve the detection of fake assets.

There is another method for generating additional data known as Data Augmentation. But, it is not synthetic data. This method is a process where new data is added to an existing dataset. This is known as data anonymization, and a set of such data is not synthetic data.

Synthetic data generation tools



Synthetic data generation is now a widely used term along with machine learning models. As it is AI, using a tool for generating synthetic data plays a vital role. Here are some tools which are used for the same:

Datomize: Datomize has an Artificial Intelligence or Machine Learning model which is majorly used by world-class banks all over the globe. With Datomize, you can easily connect your enterprise data services and process high-intensity data structures and dependencies with different tables. This algorithm will help you in extracting behavioral features from the raw data and you can create identical data twins with the original data.

MOSTLY.AI: MOSTLY.AI is a synthetic data tool that enables AI and high-priority privacy while extracting structures and patterns from the original data for preparing completely different datasets.

Synthesized: Synthesized is an all-in-one AI dataOps solution which will help you with data augmentation, collaboration, data provisioning, and secured sharing. This tool generates different versions of the original data, and also tests them with multiple test data. This helps in identifying the missing values and finding sensitive information.

Hazy: Hazy is a synthetic data generation tool that aims to train raw banking data for fintech industries. It will let the developers ramp up their analytics workflows by avoiding any fraudulence while collecting real customer data. You can generate complex data during financial service generations and store it in silos within the company. But, sharing real financial data for research purposes is severely limited and restricted by the government.

Sogeti: Sogeti is a cognitive-based solution that helps you with data synthesis and processing. It uses Artificial Data Amplifier technology which reads and reasons with any data type, whether it's structured or unstructured. ADA uses deep learning methods to mimic recognition capabilities and sets it apart.

Gretel: Gretel is the tool that is specifically built to create synthetic data. It is a self-proclaimed tool that generates statistically equivalent datasets without giving out any sensitive customer data from the source. While training the model for data synthesis, it compares the real-time information by using a sequence-to-sequence model for enabling the prediction while generating new data.

CVEDIA: Packed with different machine language algorithms, CVEDIA provides synthetic computer vision solutions for improved object recognition and AI rendering. It is used for a variety of tools, and IoT services for developing AI applications and sensors.

Rendered.AI: Rendered.AI generates physics-based synthetic datasets for satellites, robotics, healthcare, and autonomous vehicles. It is a no-code configuration tool and API for engineers to make quick changes and analytics on datasets. They can perform data generation on the browser and it will enable easy operation on ML workflows without much computing power.

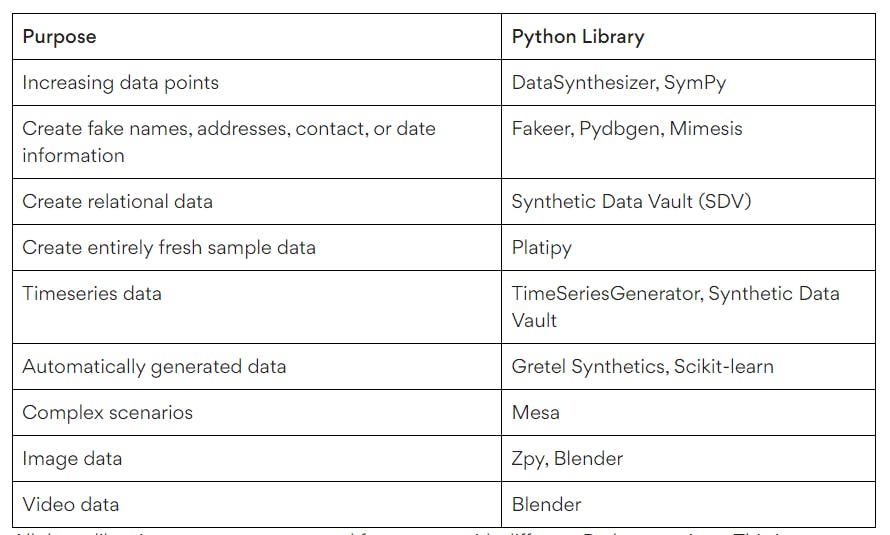
Oneview: Oneview is a data science tool that uses satellite images and remote sensing technologies for defense intelligence. Using mobiles, satellites, drones, and cameras, this algorithm will help object detection even where there are blurred images or lower resolutions. It will provide accurate and detailed annotations on the virtually created imagery which will closely resemble the real-world environment.

MDClone: MDClone is a dedicated tool that is majorly used in healthcare businesses for generating an abundance of patient data which will allow the industry to harness the information for personalized care. But, for accessing clinical data, researchers should depend on mediators and the process was slow and limited. MDClone offers a systematic approach for democratizing healthcare data for research, synthesis, and analytics without disturbing sensitive data.

Generating synthetic data using Python-based libraries

A few Python-based libraries can be used to generate synthetic data for specific business requirements. It is important to select an appropriate Python tool for the kind of data required to be generated.

The following table highlights available Python libraries for specific tasks.



All these libraries are open-source and free to use with different Python versions. This is not an exhaustive list as newer tools get added frequently.

Challenges and limitations while using synthetic data

Although synthetic data offers several advantages to businesses with data science initiatives, it nevertheless has certain limitations as well:

Reliability of the data: It is a well-known fact that any machine learning/deep learning model is only as good as its data source. In this context, the quality of synthetic data is significantly associated with the quality of the input data and the model used to generate the data. It is important to ensure that there are no biases in source data else those may be very well reflected in the synthetic data. Additionally, the quality of the data should be validated and verified before using it for any predictions.

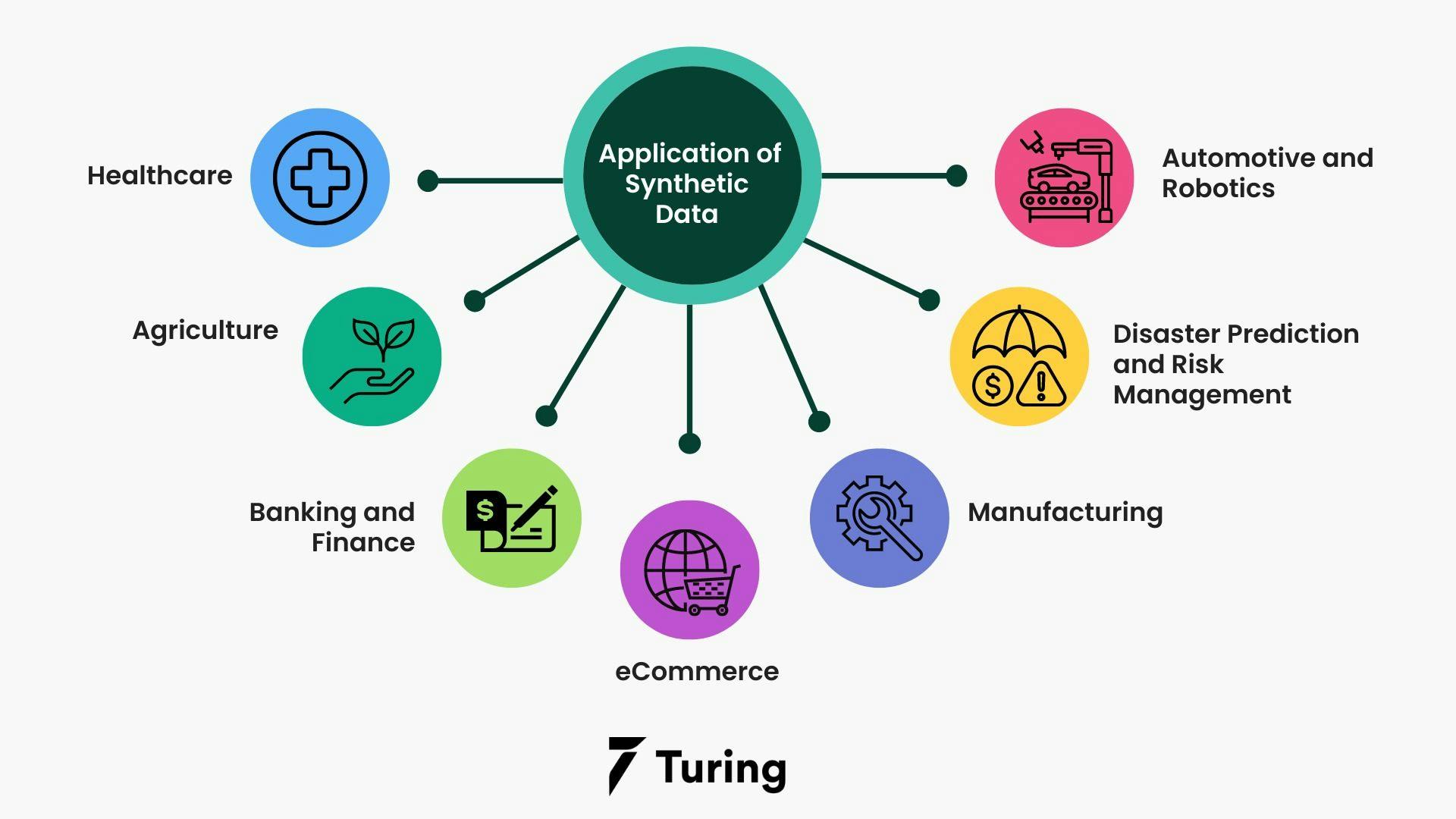
Replicating outliers: Synthetic data can only resemble real-world data, it cannot be an exact duplicate. As a result, synthetic data may not cover some outliers that exist in genuine data. Outliers in the data might be more important than normal data.

Requires expertise, time, and effort: While synthetic data might be easier and inexpensive to produce when compared with real data, it does require a certain level of expertise, time, and effort.

User acceptance: Synthetic data is a new notion, and people who have not seen its advantages may not be ready to trust the predictions based on it. This means that awareness about the value of synthetic data to drive more user acceptance needs to be created first.

Quality check and output control: The goal of creating synthetic data is to mimic real-world data. The manual check of the data becomes critical. For complex datasets generated automatically using algorithms, it is imperative to ensure the correctness of the data before implementing it in machine learning/deep learning models.

Real-world applications using Synthetic data



Here are some real-world examples where synthetic data is being actively used.

Healthcare: Healthcare organizations use synthetic data to create models and a variety of dataset testing for conditions that don’t have actual data. In the field of medical imaging, synthetic data is being used to train AI models while always ensuring patient privacy. Additionally, they are employing synthetic data to forecast and predict trends of diseases.

Agriculture: Synthetic data is helpful in computer vision applications that assist in predicting crop yield, crop disease detection, seed/fruit/flower identification, plant growth models, and more.

Banking and finance: Banks and financial institutions can better identify and prevent online fraud as data scientists can design and develop new effective fraud detection methods using synthetic data.

eCommerce: Companies derive the benefits of efficient warehousing and inventory management as well as an improved customer online purchase experiences through advanced machine learning models trained on synthetic data.

Manufacturing: Companies are benefitting from synthetic data for predictive maintenance and quality control.

Disaster prediction and risk management: Government organizations are using synthetic data for predicting natural calamities for disaster prevention and lowering the risks.

Automotive & Robotics: Companies make use of synthetic data to simulate and train self-driving cars/autonomous vehicles, drones, or robots.

Future of synthetic data

We have seen different techniques and advantages of synthetic data in this article. Now, we will want to understand ‘Will synthetic data replace the real-world data?’ or ‘Is synthetic data the future?’.

Yes, synthetic data is highly scalable and smarter than real-world data. But creating accurate synthetic data will require more effort than creating it using an AI tool. When you want to generate correct and accurate synthetic data, you need to have a thorough knowledge of AI and should have specialized skills in handling risky frameworks.

Also in the dataset, there should not be any trained models which will skew it and make it far from reality. This will adjust the datasets by creating a true representation of the real-world data and considering the present biases. You can generate synthetic data using this method and can fulfill your goals.

It is well-known that synthetic data aims at facilitating data scientists on accomplishing new and innovative things which will be tougher to achieve with real-world data, so you can surely assume that synthetic data is the future.

Finally

You will come across many situations where synthetic data can address the data shortage or the lack of relevant data within a business or an organization. We also saw what techniques can help to generate synthetic data and who can benefit from it. Furthermore, we discussed some challenges involved in working with synthetic data, along with a few real-life examples of industries where synthetic data is being used.

Real data will always be preferred for business decision-making. But when such real raw data is unavailable for analysis, synthetic data is the next best solution. However, it needs to be considered that to generate synthetic data; we do require data scientists with a strong understanding of data modeling. Additionally, a clear understanding of the real data and its environment is crucial too. This is necessary to ensure that the data being generated is as close to the actual data as possible.