Synthetic Data: Description, Benefits and Implementation

The quality and volume of data are critical to the success of AI algorithms. Real-world data collection is expensive and time-consuming. Furthermore, due to privacy regulations, real-world data cannot be used for research or training in most situations, such as healthcare and the financial sector. Another disadvantage is the data's lack of availability and sensitivity. To power deep learning and artificial intelligence algorithms, we need massive data sets. **Synthetic data**, a new area of artificial intelligence, relieves you of the burdens of manual data acquisition, annotation, and cleaning. Synthetic data generation solves the problem of acquiring data that would otherwise be impossible to obtain. Synthetic data generation will produce the same results as real-world data in a fraction of the time and with no loss of privacy.

Visual simulations and recreations of real-world environments are the focus of synthetic data generation. It is photorealistic, scalable, and powerful data that was created for training using cutting-edge computer graphics and data generation algorithms. It is highly variable, unbiased, and annotated with absolute accuracy and ground truth, removing the bottlenecks associated with manual **data collection** and annotation.

Why is synthetic data required?

Businesses can benefit from synthetic data for three reasons: privacy concerns, faster product testing turnaround, and training machine learning algorithms.
Most data privacy laws limit how businesses handle sensitive data. Any leakage or sharing of personally identifiable customer information can result in costly lawsuits that harm the brand's reputation. As a result, one of the primary reasons why companies invest in synthetic data and synthetic data generation techniques is to reduce privacy concerns.
Any previous data remains unavailable for completely new products. Furthermore, human-annotated data is an expensive and time-consuming process. This can be avoided if businesses invest in synthetic data, which can be generated quickly and used to develop reliable machine learning models.

What is the creation of synthetic data?

Synthetic data generation is the process of creating new data as a replacement for real-world data, either manually using tools like Excel or automatically using computer simulations or algorithms. If the real data is unavailable, the fake data can be generated from an existing data set or created entirely from scratch. 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. Despite being artificial, synthetic data mathematically or statistically reflects real-world data. It is similar to real data, which is collected from actual objects, events, or people in order to train an AI model.

Real data vs. synthetic data

Real data is measured or collected in the real world. Such information is generated every time a person uses a smartphone, laptop, or computer, wears a smartwatch, accesses a website, or conducts an online transaction. Furthermore, surveys can be used to generate real data (online and offline).
In digital contexts, synthetic data is produced. With the exception of the portion that was not derived from any real-world occurrences, synthetic data is created in a way that successfully

mimics the actual data in terms of fundamental qualities. The idea of using synthetic data as a substitute for actual data is very promising because it may be used to provide the training data that machine learning models require. But it's not certain that artificial intelligence can solve every issue that arises in the real world. The substantial benefits that synthetic data has to provide are unaffected by this.

Where can you use synthetic data?

Synthetic data has a wide range of applications. When it comes to machine learning, adequate, high-quality data is still required. Access to real data may be restricted due to privacy concerns at times, while there may not be enough data to train the machine learning model satisfactorily at others. Synthetic data is sometimes generated to supplement existing data and aid in the improvement of the machine learning model.
Many sectors can benefit greatly from synthetic data:
1. Banking and financial services
2. Healthcare and pharmaceuticals
3. Internet advertising and digital marketing
4. Intelligence and security firms
5. Robotics
6. Automotive and manufacturing

Benefits of synthetic data

Synthetic data promises to provide the following benefits:

Customizable:

To meet the specific needs of a business, synthetic data can be created.

Cost-effective:

In comparison to genuine data, synthetic data is a more affordable solution. Imagine a producer of automobiles that needs access to crash data for vehicle simulations. In this situation, acquiring real data will cost more than producing fake data.

Quicker to produce:

It is possible to produce and assemble a dataset considerably more quickly with the right software and hardware because synthetic data is not gathered from actual events. This translates to the ability to quickly make a large amount of fabricated data available.

Maintains data privacy:

The ideal synthetic data does not contain any information that may be used to identify the genuine data; it simply closely mimics the real data. This characteristic makes the synthetic data anonymous and suitable for dissemination. Pharmaceutical and healthcare businesses may benefit from this.

Some real-world applications of synthetic data

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

Healthcare:

In situations where actual data is lacking, healthcare institutions are modeling and developing a variety of tests using synthetic data. Artificial intelligence (AI) models are being trained in the area of medical imaging while always maintaining patient privacy. In order to forecast and predict disease patterns, they are also using synthetic data.

Agriculture:

In computer vision applications that help with crop production forecasting, crop disease diagnosis, seed/fruit/flower recognition, plant growth models, and more, synthetic data is useful.

Banking and Finance:

As data scientists create and develop more successful fraud detection algorithms employing synthetic data, banks and financial institutions will be better able to detect and prevent online fraud.

Ecommerce:

Through advanced machine learning models trained on synthetic data, businesses gain the benefits of efficient warehousing and inventory management, as well as an improved customer online purchase experiences.

Manufacturing:

Companies are benefiting from synthetic data for predictive maintenance and quality control.

Disaster prediction and risk management:

Government agencies are using synthetic data to predict natural disasters in order to prevent disasters and lower risks.

Automotive & Robotics:

Synthetic data is used by businesses to simulate and train self-driving cars, autonomous vehicles, drones, and robots.

Synthetic Data Generation by TagX

TagX focuses on accelerating the AI development process by generating data synthetically to fulfill every data requirement uniquely. TagX has the ability to provide synthetically generated data that are pixel-perfect, automatically annotated or labeled, and ready to be used as ground truth as well as train data for instant segmentation.

Final Thoughts

In some cases, synthetic data may be used to address a company's or organization's lack of relevant data or data scarcity. We also investigated the methods for creating artificial data and the potential users. Along with a few examples from actual fields where synthetic data is used, we discussed some of the challenges associated with working with it.
When making business decisions, the use of actual data is always preferable. When such true raw data is unavailable for analysis, realistic data is the next best option. However, it should be noted that in order to generate synthetic data, data scientists with a solid understanding of data modeling are required. A thorough understanding of the actual data and its surroundings is also required. This is necessary to ensure that, if available, the generated data is as accurate as possible.