newData.ai
data-driven success starts here

Why Fake Data Matters

by Taymour | February 24, 2022

Data is the lifeblood of business today. It's what we use to make informed decisions about where to allocate our resources, how to improve our products and services, and who our target market is. Of course, relevant data must exist and be of satisfactory quality to add value, and it mustn't get into the wrong hands for security and privacy reasons. Data must also be sufficient for building machine learning models. When data doesn't meet these criteria, synthetic data can be helpful. Synthetic data generation can be faster, more adaptable, and more scalable than real-world data. It may also be easier and less expensive to acquire. Synthetic data can be incredibly beneficial for businesses in several ways, which we'll explore in this blog post.

Data Acquisition: Cost and Speed

The major cost of synthetic data is the upfront development phase. After that, generating data becomes increasingly more cost-effective than collecting real information. Fake data is often seen as being easier and less expensive to acquire compared to real data for a number of reasons.

- First, in many cases it can be generated automatically, without the need for manual labor.
- Second, it is often not subject to the same legal restrictions as real data. For example, it can be much easier to generate synthetic cornea images than it is to collect real ones, due to all the regulations around collecting biometric data.
- Third, synthetic data can be generated in controlled environments, which makes it much easier to ensure that the data is of high quality.

- Finally, synthetic data can be generated at scale more easily than real data, making it more cost-effective in many cases.

Use Case 1: Anonymizing Data for Privacy and Security Purposes

One of the key challenges in healthcare and financial services is protecting patient and customer privacy while still being able to use data for research and analytics. A common approach to this problem is anonymization, or de-identifying data by removing personal information like names and addresses. However, this approach can often lead to data that is too abstract to be useful. Synthetic data offers a potential solution to this problem, as it's generated by algorithms that mimic real data's statistical properties, but without any actual personal information. It can therefore be used for research and analytics without compromising privacy. In addition, synthetic data can be generated specifically for a particular application, making it more useful than anonymized data. As privacy concerns continue to grow, synthetic data may play an increasingly important role in health care and financial services.

Both anonymization and de-identification involve removing personal identifiers from data, but the two approaches have some important differences. Anonymization is the process of irreversibly transforming data so it can no longer be linked back to an individual, which means once data has been anonymized, it can never be used to identify an individual, even if the anonymization process is reversed. De-identification, on the other hand, is a process of removing personal identifiers from data while still retaining the ability to link the data back to an individual. Therefore, de-identified data can be used for research or statistical purposes, as long as the individuals involved can never be re-identified. While both anonymization and de-identification serve the same basic purpose, choosing the right approach based on a given situation's specific needs is important.

The Health Insurance Portability and Accountability Act (HIPAA) is a US federal law that establishes standards for handling protected health information (PHI). PHI is any information related to an individual's health, including medical records, insurance information, and other personal health data. HIPAA requires that covered entities take steps to protect PHI's confidentiality, and it imposes strict penalties for unauthorized disclosures. De-identified data is not subject to these restrictions, so it can be freely shared without concern for violating HIPAA. However, it's important to note that de-

identified data is still considered PHI if it can be used to identify an individual. For this reason, organizations should take care to ensure that de-identified data is properly anonymized before sharing it.

Another consideration is that the world has become a global village, which makes data- sharing easier than ever before. However, due to PHI's sensitive nature, it cannot be sent overseas without compromising patient confidentiality. Synthetic data with anonymized information, on the other hand, can be analyzed overseas without violating HIPPA. That way, patients' privacy is protected, but important research can still be conducted.

Use Case 2: Developing and Testing Software

Synthetic data is often used to develop and test software because it can be generated to account for all scenarios, including those that may be rare (or even impossible) in real life. Algorithms that mimic real-world conditions can generate this data, or developers can create it manually. In either case, synthetic data can be an invaluable tool for testing software to ensure it works correctly in all situations.

In some cases, no real-world data exists that would be relevant to testing the software, so synthetic data is the only option. In other cases, the amount of data required to thoroughly test the software is prohibitively expensive or time-consuming to collect. In either case, using synthetic data can help ensure that the software is of high quality and will work correctly when it is finally released, which allows software developers to test their products in a controlled environment, without having to rely on potentially unavailable or incomplete real-world data. In some cases, synthetic data can be more effective than real-world data, as it can be specifically designed to cover all the potential scenarios the software might encounter. This ability makes it an essential tool for developing and testing software.

Use Case 3: Simulating Real World Events

In recent years, the focus on using synthetic data to create realistic simulations has increased, due to the fact that synthetic data can more accurately represent real-world conditions than traditional methods. For example, when creating a car-accident simulation, synthetic data can create a more realistic representation of the physics involved. In addition, synthetic data can create simulations that

are not possible with real-world data. By manipulating the properties of synthetic data, for instance, it's possible to simulate a black hole. As synthetic data use becomes more widespread, increasingly more simulations will likely be created that would not be possible without it.

For example, synthetic data and autonomous vehicles are a natural fit for each other, owing to the difficulties and significance of "edge cases" in the world of AVs. Collecting real-world driving data for every conceivable scenario that an automated vehicle may encounter on the road just isn't feasible. Given how unpredictable and ill-defined the world is, it would take hundreds of years of real-world driving to gather all the information necessary to create a genuinely safe autonomous vehicle. To remedy this situation, AV companies created sophisticated simulation engines that generate the required data volume to train their AI systems thoroughly. This technology allows us to generate thousands—or even millions—of different driving scenario permutations such as changing other cars' positions on the road, adding or removing pedestrians, increasing or decreasing vehicle speeds, adjusting the weather conditions, and so forth.

Use Case 4: Balancing Machine Learning Models

In machine learning, data is the foundation on which models are built and trained. Without enough high-quality data, producing accurate predictions can be difficult, especially for imbalanced datasets that contain a disproportionate amount of examples from one class (e.g., positive or negative sentiment). In these cases, data augmentation—the process of artificially generating new examples—can be used to balance the dataset and improve the model's performance.

Synthetic data can be used to balance machine learning models in a number of ways. For example, if a dataset is skewed toward a particular class (e.g., it has more data points for males than females), synthetic data can be generated to even out the class distribution. This technique ensures that the model is trained on a more balanced dataset, and thus is less likely to overfit to the majority class.

Additionally, synthetic data can be used to augment existing datasets, which is particularly useful when limited real-world data is available. However, it's important to train a high-quality model. By generating additional synthetic data points, the model can be trained on a larger, more representative dataset.

Finally, synthetic data can be used to create entirely new data, which is useful when real-world data isn't available or is too difficult to collect. For example, synthetic medical records can be generated to train predictive models without violating patient privacy.

Clearly, synthetic data plays an important role in machine learning and can be used in a variety of ways to improve machine learning models' performance.

Data that is difficult or impossible to collect can also be used with the principle of a dangerous collection. For example, if your AI algorithm needs to find a needle in a haystack, synthetic data can generate rare events so that an AI model can accurately learn from it. This is especially useful in cases where real data may be too expensive or difficult to collect. For example, if you are training an AI model to detect rare disease symptoms, it would be very expensive and time-consuming to collect a large enough dataset of real data. In this case, you can use synthetic data to generate a dataset that is large enough for your AI model to learn from. Another example where synthetic data can be used is when you are training an AI model to recognize objects in images. If you have a dataset of images that do not contain the object you want your AI model to recognize, you can use synthetic data to generate images that do contain the object.Consider this: Some of the most beneficial uses of AI are focused on 'rare' events. By its very nature, rare data is hard to collect. Going back to the automotive example, car crashes don't happen very often, so you rarely have a chance to collect this data. But with synthetic data, you can simulate different crash scenarios and choose how many crashes you want to simulate .

Synthetic data is becoming an increasingly important tool for businesses and researchers. It can be used to anonymize data, develop and test software, balance machine learning models, and simulate real world events. Among synthetic data's chief benefits are its low cost and fast generation time. For these reasons alone, businesses would be well-advised to consider using synthetic data as a supplement to their real data in order to get the most value from their data sets.