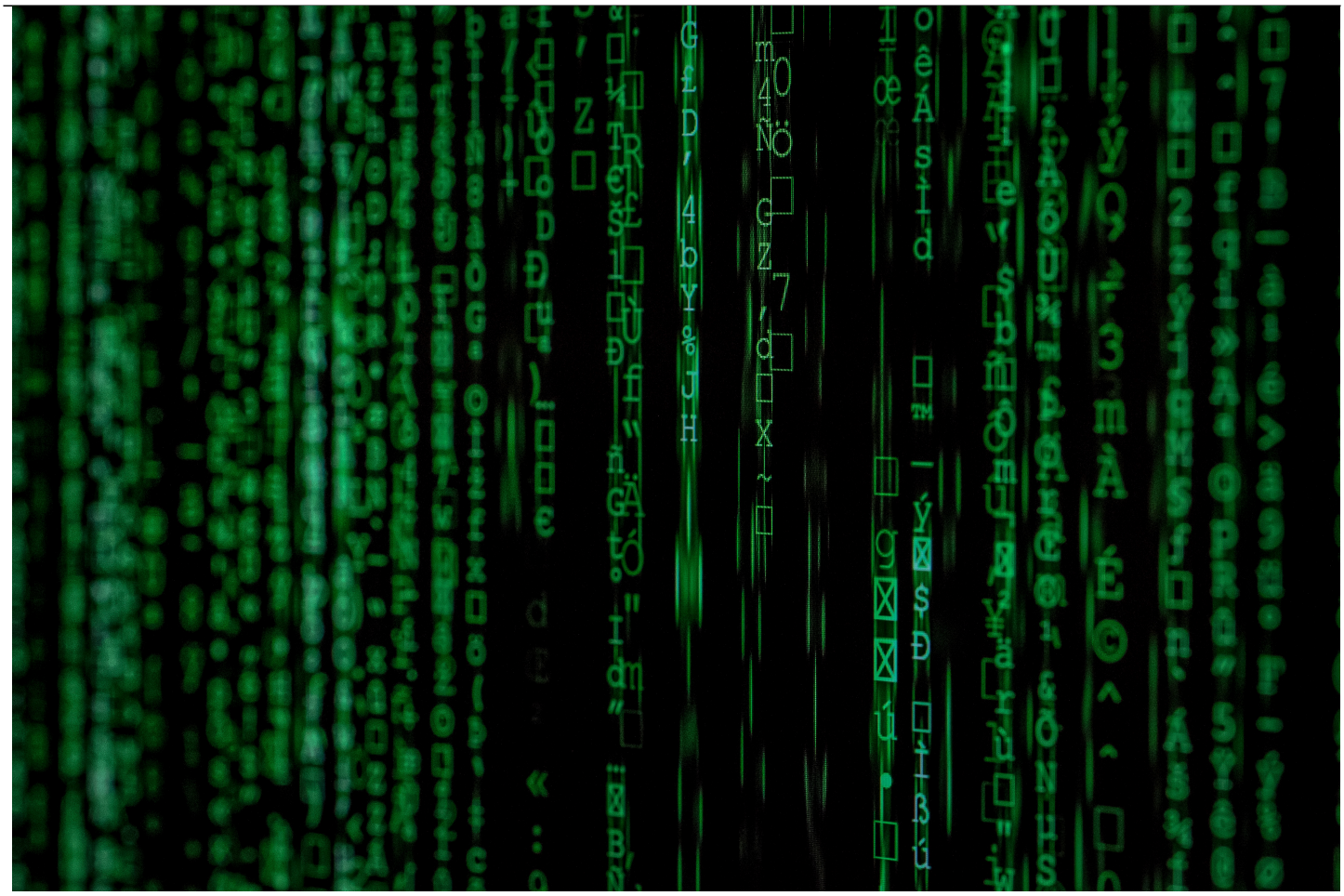
??????????

??????????

???Sergio Bruccoleri



Data fuels artificial intelligence (AI). Any AI-based product, ranging from Alexa to Netflix, needs reliable data to teach itself how to be more effective. But what happens when AI lacks enough real data to train itself? This is where synthetic data comes into play.

What Is Synthetic Data, and How Is It Used?

Synthetic data consists of data generated with the assistance of AI. Synthetic data is based on a set of real data. After being fed real data, a computer simulation or algorithm generates synthetic data to train an AI model. Research demonstrates it can be as good or even better for training an AI model than data based on actual objects, events or people. Here is a more technical definition courtesy of this post:

> Synthetic data generation describes a method of producing artificial datapoints from a real dataset. The new data is supposed to mimic the original data such that the two datasets

cannot be distinguished from one another, not even by human domain experts or computer algorithms. Having more data with similar properties to the original can be useful in a variety of ways. For example, machine learning models often improve in performance, the more training data is fed to them. Using synthetic data, more and complementary data can be created that eventually might improve a model.

The underlying principles can be a lot to digest. But the important thing to remember is that synthetic data is created in digital worlds rather than collected in the real world.

Why would someone need synthetic data to train an AI? Well, there are plenty of reasons. One of them is cost: companies spend billions of dollars each year (collectively) on data acquisition and solutions to manage, process, and analyze real data. Synthetic data can reduce the cost of data acquisition by reducing the amount of real data needed to train AI models.

In addition, synthetic data can help solve a practical problem: training AI models when real data is hard to come by. Examples include:

- Shell is reportedly using synthetic data to build models to detect problems that rarely occur; for example, Shell created synthetic data to help models to identify deteriorating oil lines.
- A technology company might need synthetic data to learn how to identify and block spam text messages. In this example, getting access to personal text messages to train an algorithm with real data might not be possible because of privacy laws.
- A retailer that wants to program an algorithm to identify attempted fraud might require synthetic data if the retailer lacks access to a large set of fraudulent transactions. With synthetic fraud data, new fraud detection methods can be tested and evaluated for their effectiveness.

The blog post "Top 20 Synthetic Data Use Cases & Applications in 2022" contains many more use cases.

At Centific, we used synthetic data for example to help a search engine understand concepts related to relatively new domains such as 3D printing in order to provide more relevant results. Collecting some of these queries in the real world at the moment was not as effective due to the low market share that a topic such as 3D printing has. To prepare for when public interest (and subsequent search queries) will increase, the client worked with us to use synthetic data to prepare the engine in giving relevant results.

Another example: collecting voice samples in small spaces, such as in a car. A business might need to do so in order to test how a voice application works in this particular setting. In the past, we might have simply put two or more people in a car to train that use case. But during the COVID-19 pandemic, scenarios like these are becoming much harder to cover. Synthetic data has helped us overcome this limitation. It mimics what real data would look like.

How Does Synthetic Data Get Created?

A business can create synthetic data in a number of ways. As this post points out, neural networks or Bayesian networks are used in order to generate new data. For example, variational autoencoders

learn patterns in data by utilizing encoding and decoding techniques or autoregressive models that are used to generate synthetic images.

AI research and deployment company OpenAI created a model, GPT-3, which produces predictive text. Per OpenAI, GPT-3 generates coherent paragraphs of text, achieves state-of-the-art performance on many language modeling benchmarks, and performs rudimentary reading comprehension, machine translation, question answering, and summarization—all without task-specific training.

OpenAI notes that this capability can help create applications such as:

- AI writing assistants.
- More capable dialogue agents.
- Unsupervised translation between languages.
- Better speech recognition systems.

Synthetic data can be used for malicious purposes such as manufacturing misleading news articles, automate the production of spam/phishing content, or impersonating others. Here's the rub, though: to fight criminals, organizations need to learn how they think and act, and few criminals are going to share how they manipulate data. Synthetic data, based on a limited data set, can be an effective solution to the lack of real data available.

How Should a Business Get Started Using Synthetic Data?

Understand first that this is a young and evolving field. Synthetic data is imperfect, and it's subject to the same caveats of using real data to train an AI model. For example, as noted, biased source data will create a biased outcome. And if you are using synthetic data in a dynamic, ever-changing field where new insights are constantly being generated – our search example with electric vehicles comes to mind — you need a way to update the algorithm to account for new data. We suggest keeping a diverse team of people in the loop to manage the sourcing of the initial data and then the ongoing oversight of the algorithm. Only human annotators or subject matter experts can ensure the data is correct. People can potentially post-edit synthetic data before it reaches the target AI model.

At Centific, we apply a human-in-the-loop approach. Our diverse team of subject matter experts uses our OneForma platform to teach AI models to make accurate decisions. Human and AI collaboration requires tight integration of human operations, machine learning, and user experience design. The human safety net serves as an extra feedback loop for model training.

Contact Centific

Our OneForma platform and global team together are the key to train your AI models and make your next generation products faster, better, smarter, and inclusive. Contact us to learn more.

Photo by Markus Spiske on Unsplash

-
-
-
-
-