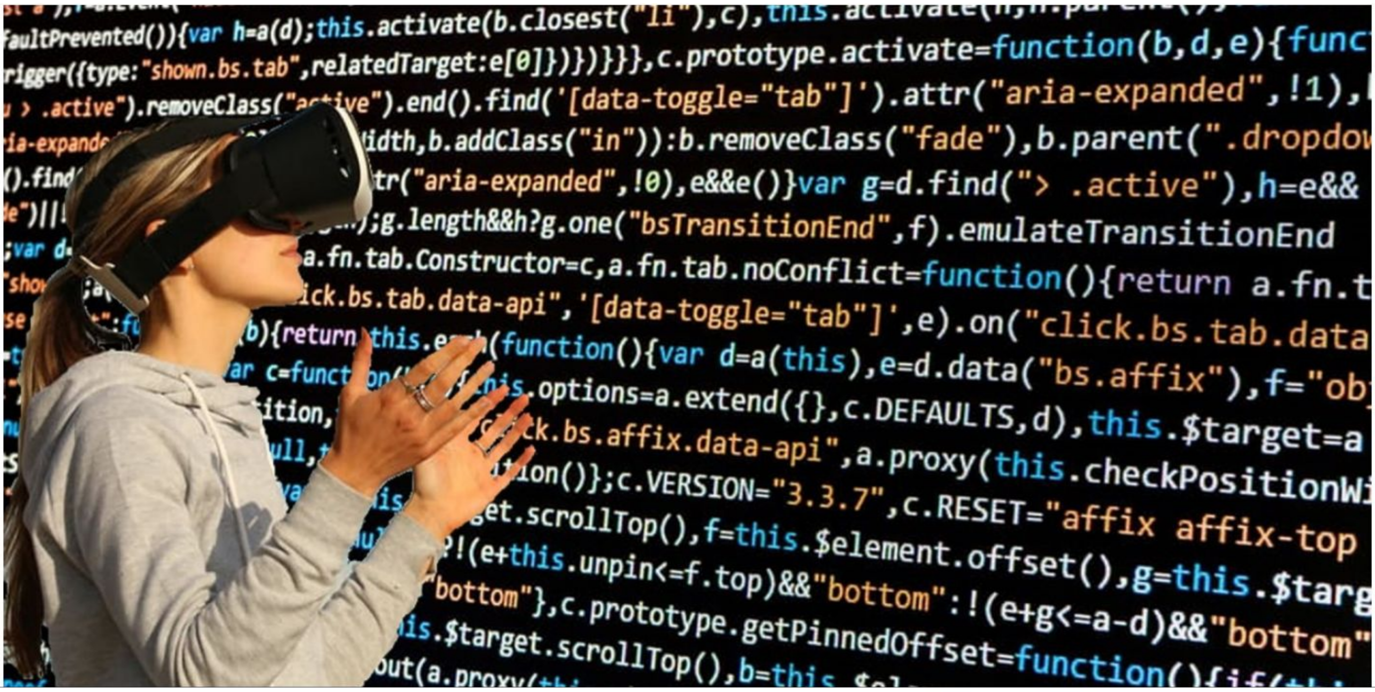
What Is Synthetic Data?

Synthetic data generated from computer simulations or algorithms provides an inexpensive alternative to real-world data that's increasingly used to create accurate AI models.

June 8, 2021 by GERARD ANDREWS



Share

Reading Time: 7 mins

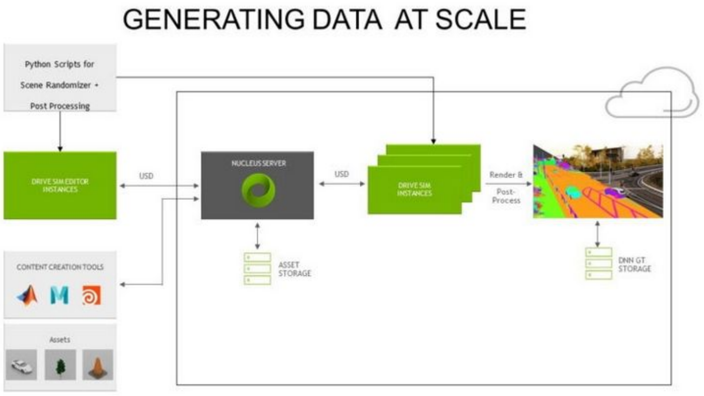
Data is the new oil in today's age of AI, but only a lucky few are sitting on a gusher. So, many are making their own fuel, one that's both inexpensive and effective. It's called synthetic data.

What Is Synthetic Data?

Synthetic data is annotated information that computer simulations or algorithms generate as an alternative to real-world data.

Put another way, synthetic data is created in digital worlds rather than collected from or measured in the real world.

It may be artificial, but synthetic data reflects real-world data, mathematically or statistically. Research demonstrates it can be as good or even better for training an AI model than data based on actual objects, events or people.



Users can generate synthetic data for autonomous vehicles using Python inside NVIDIA Omniverse.

That's why developers of deep neural networks increasingly use synthetic data to train their models. Indeed, a 2019 survey of the field calls use of synthetic data "one of the most promising general techniques on the rise in modern deep learning, especially computer vision" that relies on unstructured data like images and video.

The 156-page report by Sergey I. Nikolenko of the Steklov Institute of Mathematics in St. Petersburg, Russia, cites 719 papers on synthetic data. Nikolenko concludes "synthetic data is essential for further development of deep learning … [and] many more potential use cases still remain" to be discovered.

The rise of synthetic data comes as AI pioneer Andrew Ng is calling for a broad shift to a more data-centric approach to machine learning. He's rallying support for a benchmark or competition on data quality which many claim represents 80 percent of the work in AI.

ALL NVIDIA NEWS

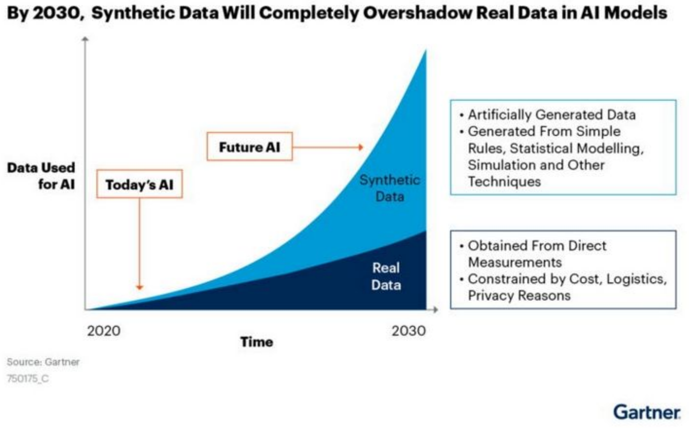
A Breakthrough Preview: JIDU Auto Debuts Intelligent Robo-01 Concept Vehicle, Powered by NVIDIA DRIVE Orin

The Data Center's Traffic Cop: AI Clears Digital Gridlock

3D Environment Artist Jacinta Vu Sets the Scene 'In the NVIDIA Studio'

Powered Up: 5G and VR Accelerate Vehicle Battery Design

From Code to Clinic, Smart Hospital Tech Boosts Efficiency, Sustainability in Medicine



Source: Gartner
750175_C

Synthetic data will become the main form of data used in AI. Source: Gartner, "Maverick Research: Forget About Your Real Data – Synthetic Data Is the Future of AI," Leinar Ramos, Jitendra Subramanyam, 24 June 2021.

In a June 2021 report on synthetic data, Gartner predicted by 2030 most of the data used in AI will be artificially generated by rules, statistical models, simulations or other techniques.

"The fact is you won't be able to build high-quality, high-value AI models without synthetic data," the report said.

Augmented and Anonymized Versus Synthetic Data

Most developers are already familiar with data augmentation, a technique that involves adding new data to an existing real-world dataset. For example, they might rotate or brighten an existing image to create a new one.

Given concerns and government policies about privacy, removing personal information from a dataset is an increasingly common practice. This is called data anonymization, and it's especially popular for text, a kind of structured data used in industries like finance and healthcare.

Augmented and anonymized data are not typically considered synthetic data. However, it's possible to create synthetic data using these techniques. For example, developers could blend two images of real-world cars to create a new synthetic image with two cars.

Why Is Synthetic Data So Important?

Developers need large, carefully labeled datasets to train neural networks. More diverse training data generally makes for more accurate AI models.

The problem is gathering and labeling datasets that may contain a few thousand to tens of millions of elements is time consuming and often prohibitively expensive.

Enter synthetic data. A single image that could cost $6 from a labeling service can be artificially generated for six cents, estimates Paul Walborsky, who co-founded one of the first dedicated synthetic data services, AI.Reverie.

Cost savings are just the start. "Synthetic data is key in dealing with privacy issues and reducing bias by ensuring you have the data diversity to represent the real world," Walborsky added.

Because synthetic datasets are automatically labeled and can deliberately include rare but crucial corner cases, it's sometimes better than real-world data.

What's the History of Synthetic Data?

Synthetic data has been around in one form or another for decades. It's in computer games like flight simulators and scientific simulations of everything from atoms to galaxies.

Donald B. Rubin, a Harvard statistics professor, was helping branches of the U.S. government sort out issues such as an undercount especially of poor people in a census when he hit upon an idea. He described it in a 1993 paper often cited as the birth of synthetic data.

"I used the term synthetic data in that paper referring to multiple simulated datasets," Rubin explained.

"Each one looks like it could have been created by the same process that created the actual dataset, but none of the datasets reveal any real data — this has a tremendous advantage when studying personal, confidential datasets," he added.

In the wake of the Big Bang of AI, the ImageNet competition of 2012 when a neural network recognized objects faster than a human could, researchers started hunting in earnest for synthetic data.

Within a couple years, "researchers were using rendered images in experiments, and it was paying off well enough that people started investing in products and tools to generate data with their 3D engines and content pipelines," said Gavriel State, a senior director of simulation technology and AI at NVIDIA.