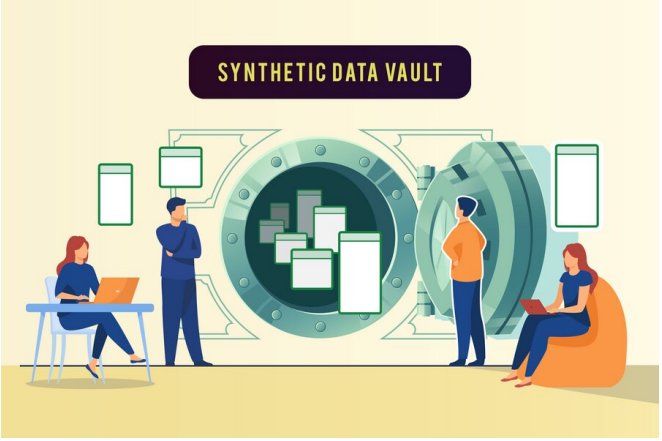
The real promise of synthetic data

19 October 2020



After years of work, MIT's Kalyan Veeramachaneni and his collaborators recently unveiled a set of open-source data generation tools — a one-stop shop where users can get as much data as they need for their projects, in formats from tables to time series. They call it the Synthetic Data Vault. Credit: Arash Akhgari

Each year, the world generates more data than the previous year. In 2020 alone, an estimated 59 zettabytes of data will be "created, captured, copied, and consumed," according to the International Data Corporation—enough to fill about a trillion 64-gigabyte hard drives.

But just because data are proliferating doesn't mean everyone can actually use them. Companies and institutions, rightfully concerned with their users' privacy, often restrict access to datasets—sometimes within their own teams. And now that the COVID-19 pandemic has shut down labs and offices, preventing people from visiting centralized data stores, sharing information safely is even more difficult.

Without access to data, it's hard to make tools that actually work. Enter synthetic data: artificial information developers and engineers can use as a stand-in for real data.

Synthetic data is a bit like diet soda. To be effective, it has to resemble the "real thing" in certain ways. Diet soda should look, taste, and fizz like regular soda. Similarly, a synthetic dataset must have the same mathematical and statistical properties as the real-world dataset it's standing in for. "It looks like it, and has formatting like it," says Kalyan Veeramachaneni, principal investigator of the Data to AI (DAI) Lab and a principal research scientist in MIT's Laboratory for Information and Decision Systems. If it's run through a model, or used to build or test an application, it performs like that real-world data would.

But—just as diet soda should have fewer calories than the regular variety—a synthetic dataset must also differ from a real one in crucial aspects. If it's based on a real dataset, for example, it shouldn't contain or even hint at any of the information from that dataset.

Threading this needle is tricky. After years of work, Veeramachaneni and his collaborators recently unveiled a set of open-source data generation tools—a one-stop shop where users can get as much data as they need for their projects, in formats from tables to time series. They call it the Synthetic Data Vault.

Maximizing access while maintaining privacy

Veeramachaneni and his team first tried to create synthetic data in 2013. They had been tasked with analyzing a large amount of information from the online learning program edX, and wanted to bring in some MIT students to help. The data were sensitive, and couldn't be shared with these new hires, so the team decided to create artificial data that the students could work with instead—figuring that "once they wrote the processing software, we could use it on the real data," Veeramachaneni says.

This is a common scenario. Imagine you're a software developer contracted by a hospital. You've been asked to build a dashboard that lets patients

access their test results, prescriptions, and other health information. But you aren't allowed to see any real patient data, because it's private.

Most developers in this situation will make "a very simplistic version" of the data they need, and do their best, says Carles Sala, a researcher in the DAI lab. But when the dashboard goes live, there's a good chance that "everything crashes," he says, "because there are some edge cases they weren't taking into account."

High-quality synthetic data—as complex as what it's meant to replace—would help to solve this problem. Companies and institutions could share it freely, allowing teams to work more collaboratively and efficiently. Developers could even carry it around on their laptops, knowing they weren't putting any sensitive information at risk.

Perfecting the formula—and handling constraints

Back in 2013, Veeramachaneni's team gave themselves two weeks to create a data pool they could use for that edX project. The timeline "seemed really reasonable," Veeramachaneni says. "But we failed completely." They soon realized that if they built a series of synthetic data generators, they could make the process quicker for everyone else.

In 2016, the team completed an algorithm that accurately captures correlations between the different fields in a real dataset—think a patient's age, blood pressure, and heart rate—and creates a synthetic dataset that preserves those relationships, without any identifying information. When data scientists were asked to solve problems using this synthetic data, their solutions were as effective as those made with real data 70 percent of the time. The team presented this research at the 2016 IEEE International Conference on Data Science and Advanced Analytics.

For the next go-around, the team reached deep into the machine learning toolbox. In 2019, Ph.D. student Lei Xu presented his new algorithm, CTGAN, at the 33rd Conference on Neural Information Processing Systems in Vancouver.

CTGAN (for "conditional tabular generative adversarial networks) uses GANs to build and perfect synthetic data tables. GANs are pairs of neural networks that "play against each other," Xu says. The first network, called a generator, creates something—in this case, a row of synthetic data—and the second, called the discriminator, tries to tell if it's real or not.

"Eventually, the generator can generate perfect [data], and the discriminator cannot tell the difference," says Xu. GANs are more often used in artificial image generation, but they work well for synthetic data, too: CTGAN outperformed classic synthetic data creation techniques in 85 percent of the cases tested in Xu's study.

Statistical similarity is crucial. But depending on what they represent, datasets also come with their own vital context and constraints, which must be preserved in synthetic data. DAI lab researcher Sala gives the example of a hotel ledger: a guest always checks out after he or she checks in. The dates in a synthetic hotel reservation dataset must follow this rule, too: "They need to be in the right order," he says.

Large datasets may contain a number of different relationships like this, each strictly defined. "Models cannot learn the constraints, because those are very context-dependent," says Veeramachaneni. So the team recently finalized an interface that allows people to tell a synthetic data generator where those bounds are. "The data is generated within those constraints," Veeramachaneni says.

Such precise data could aid companies and organizations in many different sectors. One example is banking, where increased digitization, along with new data privacy rules, have "triggered a growing interest in ways to generate synthetic data," says Wim Blommaert, a team leader at ING financial services. Current solutions, like data-masking, often destroy valuable information that banks could otherwise use to make decisions, he said. A tool like SDV has the potential to sidestep the sensitive aspects of data while preserving these important constraints and relationships.

One vault to rule them all

The Synthetic Data Vault combines everything the group has built so far into "a whole ecosystem," says Veeramachaneni. The idea is that stakeholders—from students to professional software developers—can come to the vault and get what they need, whether that's a large table, a small amount of time-series data, or a mix of many different data types.

The vault is open-source and expandable. "There are a whole lot of different areas where we are realizing synthetic data can be used as well," says Sala. For example, if a particular group is underrepresented in a sample dataset, synthetic data can be used to fill in those gaps—a sensitive endeavor that requires a lot of finesse. Or companies might also want to use synthetic data to plan for scenarios they haven't yet experienced, like a huge bump in user traffic.

As use cases continue to come up, more tools will be developed and added to the vault, Veeramachaneni says. It may occupy the team for another seven years at least, but they are ready: "We're just touching the tip of the iceberg."

More information: Modeling Tabular Data using Conditional GAN. arXiv:1907.00503 [cs.LG] arxiv.org/abs/1907.00503

Provided by Massachusetts Institute of Technology