Synthetic data puts privacy at the heart of AI projects

MARJO JOHNE
SPECIAL TO THE GLOBE AND MAIL
PUBLISHED FEBRUARY 21, 2023
UPDATED YESTERDAY

Infinity.AI is using synthetic data to create AI-powered fitness coaching from a range of virtual coaches.
INFINITY.AI

Lina Colucci predicts a future where artificial intelligence-aided sensors in factories and warehouses spot hazards instantly, and where fitness buffs can bypass their neighbourhood gym in favour of an avatar workout coach who will count reps in real-time, correct lapses in form and be available at all times.

That future is coming fast, thanks to synthetic data – tabular or visual information generated from computer algorithms to feed machine learning and predictive analytics models without setting off concerns about privacy and confidentiality.

“Synthetic data provides a way for engineers and developers to work on innovations that would normally require real-world data, which is becoming increasingly hard to get,” says Ms. Colucci, co-founder of Infinity.AI Inc. The San Francisco startup creates synthetic computer

vision data for developers looking to quickly build AI models for applications ranging from warehouse safety and robotic-powered inventory to virtual fitness coaching.

“When you talk to machine learning engineers today,” she explains, “they’ll tell you the biggest blocker of progress in their work is lack of access to data. Synthetic data removes that block.”

As privacy laws have tightened in Canada – where a new federal privacy bill, with provisions for AI systems, was introduced last summer – and in other parts of the world, use of synthetic data is growing. By 2024, according to U.S. research firm Gartner Inc., synthetic data will account for 60 per cent of all information used to develop AI and analytics projects.

Synthetic data is currently used across a wide range of applications in various sectors – from sensor training for self-driving vehicles, to market behaviour simulation in financial services. Synthetic data and AI have a closed-loop, mutually beneficial relationship: Synthetic data is created with AI, and AI models are built with synthetic data.

“You start with the real data set – for example, data from a clinical trial – and you train the AI model to learn the patterns in that data,” says Dr. Khaled El Emam, Canada Research Chair in Medical AI at the University of Ottawa and senior scientist, focused on privacy-enhancing technologies, at the Children’s Hospital of Eastern Ontario (CHEO) Research Institute. “Then you can generate new data from the AI model.”

Synthetic data bears the same mathematical and statistical properties as the source data, and preserves correlations among data variables, so that trends in the source data set are also reflected in the generated data set. But it contains no information that could compromise privacy. Unlike data that’s been merely “de-identified” (scrubbed of identifying details), synthetic data sets are entirely separate, and they can’t be linked back to the source.

“The generated data doesn’t pertain to specific individuals but contains patterns,” Dr. El Emam says. “And when you analyze the synthetic data, you learn the same things you would have learned from the real data.”

Synthetic data can also be simulated so that, instead of merely cloning a data set, the algorithm augments the real data with synthetic values.

“For example, to create a warehouse system that automatically detects spills, you’d have to feed the machine learning hundreds, even thousands, of images that would teach it to

recognize what a spill looks like," Ms. Colucci says. "You can either go out and take photos of different kinds of spills – of different sizes, shapes, colours, textures and under different lighting – or you can generate synthetic images of spills based on just a few real-world images."

The benefits from synthetic data are promising. In medical research, it's improving research quality and results by, for example, simulating patients from under-represented racial and socioeconomic groups to reduce bias in a study. This ability to simulate data could also solve a persistent challenge in researching treatments for pediatric and rare diseases: small patient groups that have, historically, made it difficult to prove whether or not a new drug works.

In Alberta, a not-for-profit organization called Health Cities has built synthetic data for a project aimed at preventing opioid addiction.

"Alberta has about 400,000 data points spanning seven years that include pharmacy data, ER visits, diagnostic data and administrative data," says Health Cities CEO Reg Joseph. "With this we can start looking at prescribing and usage habits and all kinds of metrics to find patterns that can help inform practices to prevent addiction."

A similar predictive data project is being advanced at the U.S. Veterans Health Administration, which oversees about 1,300 medical sites that look after roughly nine million former soldiers. About two years ago, the administration – known commonly as the VA – launched a predictive data analysis project in hopes of identifying veterans at high risk of suicide.

An important early step in the project was the creation of synthetic data – based on real-life information such as medical records, hospital registration and calls to suicide hotlines – which can be shared with external companies the VA has hired to build predictive models.

"Without this synthetic data, the VA's ability to work with industry would be a bit handcuffed," says Josh Rubel, chief commercial officer at Israel-based MDClone Ltd., which generated the synthetic data for the VA. "Now they've got 25 companies working with them on this project."

The use of synthetic data doesn't always have to be project-based, Mr. Rubel says. Other MDClone customers – such as The Ottawa Hospital, Jewish General Hospital in Montreal and Nova Scotia Health Authority – use synthetic data regularly to uncover patterns that can help with health care quality improvement and resource management. These patterns could relate

to emergency department admissions following patient discharge from hospital, or to patients on a particular medication who develop a certain condition.

“Synthetic data allows (our customers) to make these kinds of queries multiple times a day, on demand,” Mr. Rubel says. “The use of synthetic data in health care is proliferating and I think it will open up a lot of opportunities, including the potential for organizations that have common problems to share data with each other and compare with their peers.”

It isn’t just data that’s up for sharing. At the Massachusetts Institute of Technology in Cambridge, Mass., a team of scientists created an open-source platform in 2020 to give other organizations access to software for generating synthetic data.

The Synthetic Data Vault – which was recently folded into a spinoff company called DataCebo – had no shortage of users, says co-founder Kalyan Veeramachaneni, a principal research scientist at MIT’s Laboratory for Information and Decision Systems.

“We’ve had more than a million downloads,” he says. “The Korean customs office used it to see if they could identify which people should be pulled aside for inspection, and a school in the Netherlands used it for assessing whether they were being meritocratic in granting admissions. They also created an algorithm to project future admissions.”

The potential applications for synthetic data are as many as can be imagined, Mr. Veeramachaneni says. But where he sees viable, high-value uses are in areas where projected outcomes or incidents on which a dataset would be based are rare.

“For example, human resources attrition, or certain types of machine failures,” he says. “It would be useful in insurance because 95 per cent of us don’t make claims, so an insurance company that wants to predict claims or losses to set better pricing would need synthetic data.”

For the time being, synthetic data won’t be replacing real-world data. Today, researchers who use synthetic data to arrive at a conclusion will typically then validate their results against real-world data. The two work hand in hand, and will continue to do so for the foreseeable future.

“We don’t know enough – we’re still trying to figure out how robust our synthetic data is, and we see in publications that there are biases inherent in AI,” Mr. Joseph at Health Cities says. “But the promise is there for sure.”