datanami.com

Training Data: Why Scale Is Critical for Your AI Future

10-12 minutes

via Shutterstock

Data is the fuel that drives AI. But there's a big difference in the quality of fuel you can put into your AI engine. If your enterprise can create the biggest stockpile of the highest quality training data, it will likely win the AI race, but getting there is no easy task.

For all the advanced skills that data scientists possess, there's no escaping the fact they often spend up to 80% of their time cleaning and prepping data. Without good, clean data to feed into machine learning algorithms, the data scientist can't be sure that the model will predict anything worthwhile.

And so data scientists spend much of their time doing what amounts to data janitorial work. While data scientists will always want to personally inspect some of the data that's being used to

train machine learning models, there clearly are better uses of data scientists' time.

This situation has spawned a cottage industry of data labeling outfits. You provide the raw data and a description of what you're after, and the data labeling company will distribute the work to its human workforce, who will apply the labels that will tell algorithms what to look for, leaving data scientists to concentrate on improving the model.

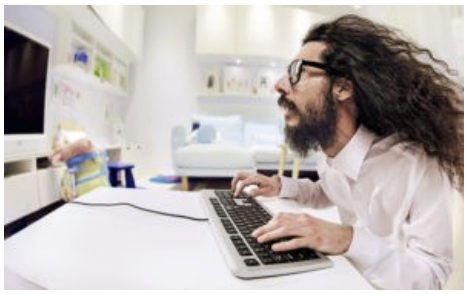
One of the data labeling outfits that's helping companies to stockpile high quality AI fuel for American enterprises is Alegion. Nathaniel Gates founded the Austin, Texas company seven years ago as a crowdsourcing company for general purpose business automation tasks. But about three years ago, Gates noticed a change in the types of tasks customers were requesting.

"We didn't really know quite what we were doing at that point," Gates tells *Datanami*. "They needed 97% to 99% accurate data, and they needed it very large scale. You had to knock us on the head a few times for us to realize, oh holy cow, this is not going away. And we ended up rebuilding the platform, our whole technology stack, around very specifically the construction and development of very high quality training data."

Garbage In, Garbage Out

The potential for bad data to negatively impact AI projects is not a theoretical threat. In fact, bad data threatens the very existence of AI, at least as we have come to define it. That's why data scientists spend so much of their time making absolutely certain the data they're feeding into their machine learning algorithms is as good as

it can be.



Are you treating your data scientists like data janitors? (conrado/Shutterstock)

Alegion finds itself working closely with data science teams to obtain training data. “We’re very much on the nose of the pain for these data science teams, because it’s still a garbage in, garbage out world,” Gates says.

“There’s a lot of frustration when they train their models with deficient data and then they question why they have a deficient machine learning model out the backend,” he continues. “We end up having a lot of frustrated data scientists who approach us and say ‘We need 10 times the scale of what we’re trying to do now with double the accuracy,’ and that’s where were able to solve their pain.”

Deep learning, which is the driving force behind modern AI today, needs a lot of data, and Alegion is positioned to serve it up. The bulk of the company’s use cases fall into one of three buckets: labeling data for computer vision algorithms (images or video); labeling data for natural language processing (audio or transcribed text); or performing entity data resolution for other types of enterprise data.

The company has dozens of clients across retail, manufacturing, healthcare, and financial services. Companies can use Alegion’s

training data platform with their own workforce of human curators, or they can contract with Alegion to provide the curators via third-party services like Amazon's Mechanical Turk or a government-backed Malaysian entity called eRezeki.

Alegion provides data for just about any AI project, except for self-driving cars, Gates notes. "It could be tracking objects as they move through frames of video," he says. "It could be listening to call audio between a firetruck and dispatch and tracking speakers and categorizing the speakers. There's lots of different categorization and annotation that we do to customers data in a very defined taxonomy so it can be used for training."

Scaling Up For AI

Creating curated training data is one aspect of the Alegion platform. But the platform helps automate additional steps, including model validation and scoring, and evaluating edge cases.

"Step one is we need to train this algorithm, so they'll send us thousand and thousands of images or video and we'll do the initial labeling and training," he says. "But steps two through 100 is the constant iteration where we're looking at the output of their machine learning's best guess, and we're scoring that."

The accuracy of training data is critical for successful machine learning (Image source: Google)

Getting people to label is just 25% of what you need, Gates says. To deliver a truly useful model will often involve identifying edge cases where the machine learning models don't perform well enough, for whatever reason.

"What you must have is accuracy, and accuracy at scale is very difficult to achieve," he says. "You have to be sure you have the processes and methods behind the scenes to make sure that you are getting to accuracy."

For example, if you're building a chatbot, it's fairly easy to find the 10 most popular questions that consumers have for a given retail operation, and train a chatbot to answer those questions. You can very quickly get a chatbot with 60% to 65% accuracy, Gates says.

"But the other 35% is a hodgepodge of several thousand other reasons of why they might call in, and you have to train and train and train to find enough use cases to identify commonalities and suss out all of those edge cases," he says. "But that's what separates a model that's 65% accurate from a model that's at 90% efficacy…The devil very much is in the downstream details."

One way the Alegion platform boosts accuracy is by using consensus grade work, where it asks three human workers the same question, and measures their responses. If they all agree, then it might be passed it to the workers with 95% certainty. But if they disagree, Alegion might solicit the response of a higher rated worker, or somebody with a specialized credentials, such as handling sensitive HIPAA data.

Alegion tracks every human worker in its system, and measures the accuracy of the data labeling decisions they make. "Unless you go through and build all of that," he says, "you will not have an

assurance of accuracy, in whatever method you're using to do your labeling to begin with."

AI Helping AI

In addition to providing labeled data to train AI models, Alegion itself uses AI models to enhance the quality of the training data. One way it does this is by using machine learning to pre-label certain data elements, such as the quantity or quality of berries in photographs taken by drones.

Machine learning will be essential for providing large volumes of accurate training data for AI (Bakhtiar Zein/Shutterstock)

"Initially we might require three human judgements to get to the level of quality the customer needs," he says. "But after a few hundred thousand, we might be able to start pre-labeling, such that now we only need two workers or one worker…The customer benefits from that cost savings, and the moat gets bigger and bigger."

In Gates view, we must lean on AI to generate the large amounts of high quality training data that downstream AI models will need. "The only economically viable solution is that a large portion of that heavy lifting is machine-derived and not human-derived," he says.

At Alegion, AI currently augments human-derived data only 20% of the time, mostly in computer vision use cases, where neural

network-powered AI is proving itself capable of labeling data with human-like precision. The goal is to get to 50% by the end of the year, he says. General enterprise data, however, is much more variable, and still requires lots of human eyeballs to ensure high accuracy.

“The more that we can move to AI, the more cost savings the customer can recognize,” Gates says. “But what that really means is, with the same amount of budget they can train more data. If the customer has a $100,000 per month training budget, and they’re able to train 5x more because of the interaction with AI in the training effort now, that has a huge amount of value to the customer, because they’re going to suss out more edge cases and get to even higher confidence in their model.”

Pursuit of Model Perfection

Modern AI is still in its infancy, and most enterprises are content to establish the efficacy of machine learning with some relatively easy use cases. But as enterprises rack up the AI wins, they will see what an enormous opportunity lies before them to continuously use data to improve business processes across the board.

The enterprises with the best training data will win (Joe Techapanupreeda/Shutterstock)

“I don’t think the broader enterprise market gets it yet. Right now they’re just trying to chase basic ROI in their ML efforts,” Gates

says. “The more mature customers understand that it doesn’t stop there. The training continues. It’s a never-ending pursuit of perfection in these models.”

The winners in individual vertical industries will be the enterprises who have the models with the highest confidence, Gates says. “That’s where you’re Google Maps beats the Apple Maps because it’s been training for longer. It has more data points. That’s why Google Assistant works better than Siri. More training, more edge cases, more variability that has been sussed out at this point.

“I’m quite convinced,” Gates continues, “that in the future, these models are going to be mandated to continuously and always self-improve and identify new edge cases and new deficiencies within their confidence models, and get those trained and get more holes plugged to have the most dominant ML model within those respective industries.”

Related Items:

‘Lifelong’ Neural Net Aims to Slash Training Time

Training Your AI With As Little Manually Labeled Data As Possible

Three Ways Biased Data Can Ruin Your ML Models