Infosys®
Navigate your next

VIEWPOINT

SCALING AI: DATA OVER MODELS

Machine learning models have generated much hype. But without clean, labeled data, their outcomes are flawed. Humans have traditionally been used to do the labeling, but bias can creep in, and costs often escalate. Instead, a combination of intelligent learners and a programmatic data creation approach is required.

Infosys® | Knowledge Institute

Artificial intelligence (AI) and its hearty accomplice machine learning (ML) are used in many novel efforts: self-driving cars, cancer detection, and cashier-less shops. However, these systems are increasingly complicated, and the models used are data hungry. They require high-quality data that is free from the unconscious bias of the programmers and AI experts sifting through it.

To overcome these issues, firms from Google to Amazon and beyond are employing labeling companies to lay the groundwork. The industry is awash in new outfits, such as San Francisco-based Scale AI and Sama (formerly known as Samsource), a startup with teams of human data labelers. All will label gargantuan datasets for a fee.

However, the fees are usually high. And there is no guarantee that the output will be comprehensive, unbiased, and free from noise. Further, the work takes too long for Agile companies that release new products every week. Lytx, a San Diego company that sells systems to gauge tiredness in truck drivers, says that it takes 10,000 hours of 20-second labeled video clips to train its AI system. The amount of video increases to 5 million hours when deployed at scale.[1] That time eats away at project success and reduces efficiency.

AI consultancy Cognilytica reckons that firms will spend $4.1 billion on data labeling by 2024.[2] At Infosys, we believe that 25% to 60% of ML project costs go toward manual labeling and validation.

The human touch

Even the most advanced technology can require the human touch. That is why firms outsource this sort of work to startups. Some ML tasks need highly skilled experts to understand the problem and to label the data correctly, such as with legal contracts and health care use cases. In other scenarios, data labeling is deskilled and gamified, capturing the wisdom of crowds. The reCAPTCHA project, for instance, is used to both verify identity and to create datasets for digitizing books and building language models. This kind of gamification can power the learning of AI to solve problems at an unimaginable scale while also reducing data bias. In two telling scenarios, the Foldit project gamified data labeling to help create antiviral drugs for both HIV and the coronavirus that led to this current pandemic.[3,4] It did this by asking users to play a simple game of connecting dots, which in turn represented the tertiary structure of proteins.

At the corporate level, firms can tap into their partner ecosystems to create rich datasets that have built-in privacy, a boon for those in the financial services and health care industries. In a system known as cooperative computing, each partner develops its own models and data, the output of which is shared with other firms in the consortium. For example, Google recently made COVID-19 datasets freely available to scientists during the pandemic. However, such data must be consolidated and encoded properly so that different users can use it efficiently and effectively.

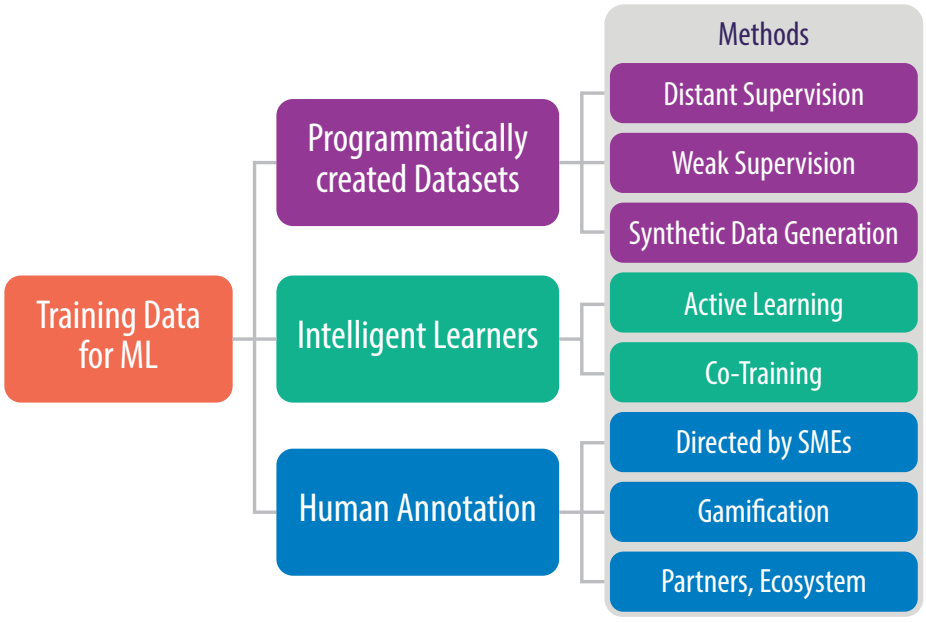
But in all these cases, human error and biased judgment enter the picture. Human data labeling can have errors as high as 18% across projects we've encountered.

What's needed is a faster and more effective way of preparing data for hungry ML algorithms, with the right governance in place to ensure models use data that is fair and unbiased.

> Gamification can solve problems at scale and reduce data bias

Two further methods to label data include "intelligent learners" and the programmatic creation of data. (Figure 1)

Figure 1. The three categories of creating machine learning data



Source: Infosys

Intelligent learners: Active learning and co-training

Intelligent learners generally increase the efficiency of creating labeled datasets. In "active learning," a classifier examines unlabeled data and picks parts of this data for further human labeling. This makes the process active rather than passive. It also increases data quality, as the classifier has control over data selection and picks only areas that haven't been optimized for ML. In one legal use case (labeling contractual clauses), we found that active learning increased data accuracy from 66% to 80%[5] even when using fewer data points. Labeling time and cost were also significantly lower; avoiding tagging by subject matter experts reduced costs by 18%.

However, co-training raises the game to a new level. In this paradigm, orthogonal views of the data are used to specify which features are needed for training. This optimizes the labeling process. For instance, one view of a case law document might be visual, without any labels given. The other view might be textual, with human labeling used. Bringing these two views together, the intelligent learner can use the first visual view to create new labels along with appropriate confidence intervals. This cycle is repeated; each cut of data boosts model accuracy.

Humans are still in the loop in both of these cases. To scale AI effectively, a large organization needs to bring down costs and increase the speed of data labeling. In this instance, firms create datasets programmatically.

Distant supervision and weak supervision

Programmatic data creation is currently the best way to do AI at scale, with either distant or weak supervision. For both, a labeling function is programmed to create labels from input datasets. By combining noisy signals, distant and weak supervision can resolve conflicting labels without access to any sort of "ground truth." The process can be run on wide and diverging datasets.

In distant supervision, noise-free training data is produced using distant knowledge bases. For instance, in the case of an initial public offering filing, a report could be generated that gives hard evidence about financial performance. An ML algorithm would have to read a standard report, extract circled entities, and find the standard name of the attribute.

Distant supervision would look in many data sources and databases and then map the financial metrics to the corresponding sentences in which they appear. Training data would be created using several transformation functions. However, there might still be noise in the label, determined by the type and number of knowledge bases the training data refers to. The real challenge, though, is that finding useful distant knowledge bases is difficult. ML engineers need domain experts to help them uncover the appropriate information. Using this method properly, however, we've found that 25,000 records of financial data can be created in one week, with 98% ML model accuracy.

But what happens when data needs to be sourced from unreliable avenues? In this case, it is wise to use weak supervision. One such use case is scoring customer sentiment on social media. With weak supervision, training data is created by crawling through social feeds and using specific hashtags in a classification procedure. In this case, labels are created from labeling functions using both weak and strong signal data.[6]

Synthetic data generation

Distant and weak supervision use databases and labeling functions to get to the right answer. But sometimes that data just isn't available.

One option is to make up the data, as Amazon has done with its new Amazon Go Stores.[7] The company uses graphics software to create virtual shoppers, which in turn train computer vision algorithms to work out what real-world shoppers are choosing. Other examples of synthetic data generation abound. Nvidia, the chipmaker, released a paper in 2018 that described a method for creating synthetic training data for self-driving cars.[8] The authors concluded that the algorithms worked better than those trained on real-world data alone. An otherworldly example comes from the Perseverance mission to Mars, where the entire Martian landscape was synthetically captured.

An important element is that the synthetic data has the same representative characteristics as the real-world data from which it is derived. Further, this data must have exposure to converse use cases and outliers, reducing uncertainty while ensuring that data is fair, safe, reliable, and inclusive.

Mathematical generators

There are three types of synthetic data creation listed in the literature. The first, mathematical generators, need knowledge of the underlying statistical distribution, or objective function. For example, call center customers often follow a Poisson mathematical distribution and are defined by teletraffic theory. Recently, one of the most promising statistical generators is the knockoff generator, which scans large datasets to uncover the features used in a given ML decision. It does

Figure 2. Recommendations for introducing data labeling at corporate scale

Actionable Recommendations	Use when…
Programmatic data creation is currently the best way to do AI at scale, with either **distant** or **weak supervision**.	You want to build and deploy AI fast without human labeling.
Some ML tasks need SMEs to label data correctly (legal contracts, healthcare use cases). **Active learning** increases labeling accuracy, without need for SMEs.	You want to achieve better model performance with less data.
For data shared between corporations, known as **cooperative computing**, data must be consolidated and encoded so that different users can use it efficiently and effectively.	Robust datasets are needed for innovative new corporate ML models at scale and speed
Synthetically generate when no data is available or outliers/edge cases are rare in real-world data.	Synthetic data and knock-offs will make your ML models safe, reliable, fair and inclusive.

Source: Infosys

this by uncovering features that aren't important. Such data can be used to find variables that indicate a biomarker for detecting diseases — such as triple negative breast cancer — and are used in AI governance to ensure black box systems are explainable.

Agent-based modeling

Another technique is known as agent-based modeling. The technically difficult approach boils down to creating a model that explains emergent behavior and then produces random data with the same model. To do this, the synthetic data generator uses an agent that acts based on a policy function. Data generation itself takes advantage of probability mass functions to determine what actions are needed to get from one state to another. Even if data is sparse, this sort of routine can create a vast amount of data with minimum noise.

Turing learning

Finally, Turing learning is worthy of a mention. The idea is to build a machine that combines multiple labeling functions — sometimes noisy — into one de-noised output. Of course, the output could have conflicting labels, which is why a Turing learner (basically a generative adversarial network) automatically combines the label outputs. Those are then fed into a discriminative model to come to the right answer, in which case they pass the Turing test with human-level accuracy.

Better balance sheets

Labeling data properly is all-important to AI projects and takes up a disproportionate amount of time. About one-quarter of the average ML task is spent labeling, compared to just 3% devoted to developing algorithms.[9] Large corporations are likely to struggle with those trade-offs as they seek to scale AI into every part of their business. However, active learning, distant supervision, and synthetic data generation can significantly reduce costs, increase speed, and improve the data quality required for powerful AI models to work effectively.

> Active learning, distant supervision and synthetic data generation reduces costs, increases velocity and improves data quality

This level of automation reduces the dependence on human labor in developing countries, the sort typically employed by startups such as Sama and Scale AI. It can also increase a firm's corporate governance credentials by ensuring that unconscious bias doesn't creep into decisions made by machines. With 18% of data labeled by humans flagged as incorrect or worse, the use of more sophisticated data labeling techniques can lead to greater customer confidence, reduced time to market of innovative products, and potentially better balance sheets.

References

1. If data is the new oil, these companies are the new Baker Hughes, Jeremy Kahn, Feb 4, 2020, Fortune
2. For AI, data are harder to come by than you think, June 13, 2020, The Economist
3. Foldit game leads to AIDS research breakthrough, Elizabeth Armstrong Moore, Sep 19, 2011, Cnet
4. Scientists Use Online Game to Research COVID-19 Treatment, Emma Yasinski, March 9, 2020, The Scientist
5. A system and method for active machine learning using multi-criteria decision making based on human and machine metrics, Rajeshwari Ganesan & Bhavana Bhasker & Niraj Kunnumma, October 2020, IPO Journal
6. Weak Supervision: A New Programming Paradigm for Machine Learning, Alex Ratner & Paroma Varma & Braden Hancock & Chris Re, March 10, 2019, The Stanford AI Lab Blog
7. See Ref 2
8. Meta-Sim: Learning to Generate Synthetic Datasets, October 27, 2019, Nvidia
9. See Ref 2

Authors

Rajeshwari Ganesan
AVP – Solution Consulting, Infosys
rajeshwari_ganesan@Infosys.com

Sivan Veera
Senior Principal – Enterprise Applications, Infosys
kumarasivan.v01@infosys.com

Harry Keir Hughes
Senior Consultant – Infosys Knowledge Institute
harrykeir.hughes@infosys.com

About Infosys Knowledge Institute

The Infosys Knowledge Institute helps industry leaders develop a deeper understanding of business and technology trends through compelling thought leadership. Our researchers and subject matter experts provide a fact base that aids decision making on critical business and technology issues.
To view our research, visit Infosys Knowledge Institute at infosys.com/IKI

For more information, contact askus@infosys.com

Infosys®
Navigate your next