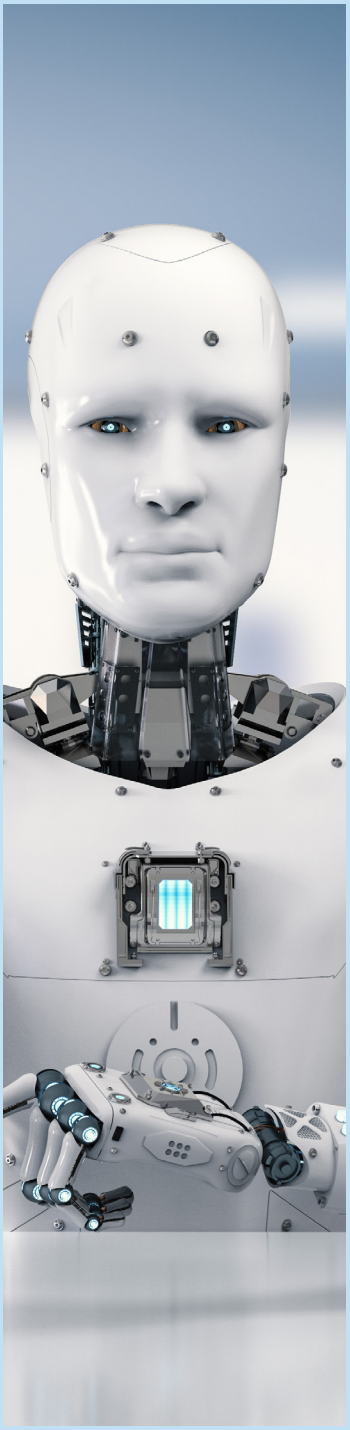
VIEW POINT

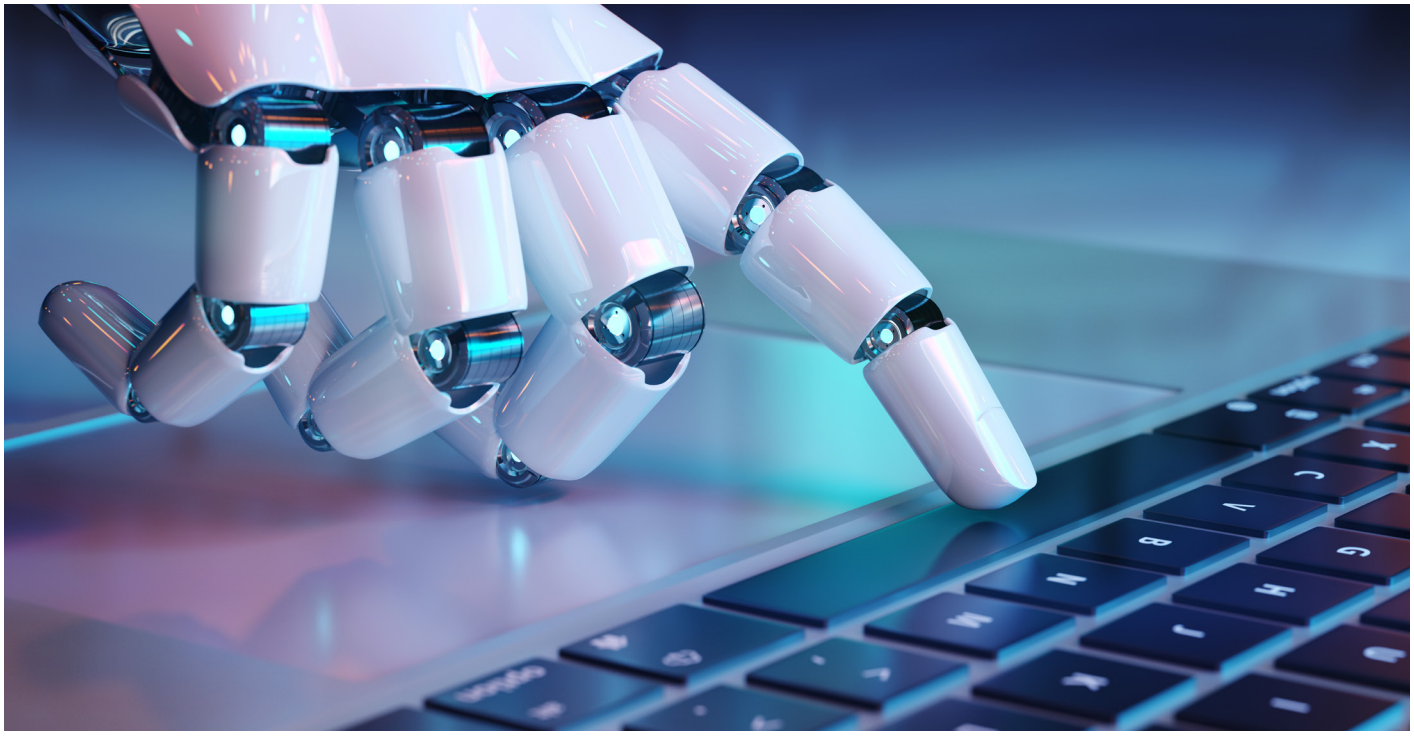


# DATA ANNOTATION: THE LIFE BLOOD OF AI/ML APPLICATIONS


## Abstract

Can you imagine going anywhere today without your GPS turned on? Following that arrow around on red, yellow, and blue lines which depict roads is the only way that we can keep ourselves from getting lost. Unless you are trekking in the wild, gone are the days of those large, unwieldy maps and stopping on the road to ask for directions. That is just one example of the magic of AI and ML. Now imagine if those lines were not roads but rivers and the place you wanted to go to, was labelled wrongly. That is data annotation gone wrong. Hence, if the text and images aren't labelled correctly, your AI/ML application cannot function correctly. That is how important Data Annotation is to AI/ML.


Infosys®
Navigate your next



## Relationship between AI/ML and data annotation

AI and ML have completely changed our lives both in terms of convenience and efficiency. Whether it is Alexa who listens to your every command and complies, or a self-driving car or even just predictive replies on your emails, everything today is dependent on AI/ML.

McKinsey has estimated that AI will be able to power additional economic activity to the tune of $13 trillion by 2030. 70% of customer engagement and usage is expected to flow through machine learning (ML) applications, mobile messaging, and AI-powered chatbots.

And, in order to function correctly, AI and ML depend on well annotated data. It is the backbone of any AI/ML model. The only way an image-based AI application can identify an image as a "car" is if several similar images have already been labelled as cars. Without annotated data, there is no machine learning application.

## Basics of data annotation

Data annotation or data labelling is needed to create training datasets which enable machine learning models to understand what it is that they are looking for in the highly cluttered, real-world situation.

Computers cannot process information like humans. They need to be trained on what it is that is being seen and how it is to be interpreted before being sentient* about it. Data annotation helps in the interpretation by labelling or annotating the objects or text so that the computer can recognise them in the real world, where the data is not already annotated. Only once a machine learning model passes through training, is it allowed to go live. Hence, it's evident that data quality determines the success of a machine learning model.

Based on the rate at which data is currently created, it is estimated that there will be a daily output of over 400 exabytes of data by 2025. Estimates from GM Insights also predict an annual growth rate of 30% for the data annotation tools market for the next six years.

Given the clear correlation between correctly annotated data and the success of a machine learning model, it should not come as a surprise that about 80% of the project development time for any AI/ML project is spent in preparing the data. As illustrated above, the smallest of errors can prove disastrous for the application.

## Data annotation types

Different types of data annotation techniques can be used, with the choice determined by the purpose of the machine learning model. Some of these types are below.

**Text annotation:** Trains machines to understand text in a better way through keyword, intent, emotion or concept identification.

**Text categorisation:** Assigns categories to sentences so that users can find what they are looking for easily.

**Image annotation:** labels images so that machines can recognize the annotated image as a distinct object using bounding boxes, classification, and segmentation.

**Video annotation:** Like in image annotation, uses bounding boxes to annotate each frame of the video or video annotation tools which capture movement.

**Equity annotation:** Helps the machine understand unstructured sentences by categorising and tagging related words.

Intent extraction: Helps machines, especially applications like chat bots, understand the intent behind a certain sentence or question. For e.g., 'How do I buy tickets?' v/s 'I would like to cancel my tickets'.

Once labelled, these data sets need to be refreshed periodically to stay relevant. According to McKinsey's Global Institute, about 75 per cent of the models require their data to be refreshed monthly, and the rest, weekly.

## Supervised or unsupervised machine learning – is the human touch needed?

In supervised machine learning, human beings annotate the data. In unsupervised learning, machines need to be able to connect the dots and learn to annotate by themselves. However, as the project becomes more complex, it becomes preferable to involve human effort in the annotation process.

Then, there is the "human-in-the-loop" learning where labelling happens within the user experience, by many people instead of one. For example, in Google Docs, when a user types a document and clicks on the squiggly line underneath a word to change the spelling, the word gets tagged against its correct spelling.

This kind of learning where human judgement is combined with machine performance helps ML models perform well.

## Build or buy? The eternal quandary

For supervised learning which requires human intervention, there exist data annotation tools which help specialists annotate training data. These are either on-premises or cloud-based solutions. Companies can either employ external vendors to do complex data annotation or create their own tools in house that are either custom built or use freeware or open-source tools available on the internet.

Until a few years ago, there weren't too many data annotation tools in the market. The early AI/ML application developers ended up building their own tools. However, that's not the case today. Sometime in 2018, a lot of data annotation tools became available offering a range of services and features for data labelling. That led to the debate within the organisations on whether to build or to buy.

Organisations who want to build their own data annotation tools have the ability to control the entire process, ensure security of data, maintain quality, make changes quickly, set the workflow as per priorities and even be able to include the AI tooling as part of their intellectual priority. However, there are the issues of needing to build the technical know-how from scratch and also the upfront costs of setting up the infrastructure and personnel required.

On the other hand, buying a tool or engaging a vendor to do the data annotation means that organisations can avoid the upfront development and personnel costs. It affords them the time and resources to focus on the core project and help accelerate the project timeline. Tooling vendors having worked with other customers can incorporate industry best practices into the tools and ensure performance metrics. They may also have multiple tools in their arsenal to suit the organisation's requirements. However, flexibility of the service provider, security, compliance, and legal issues remain valid concerns.

## As AI continues to grow, so will data annotation

Data annotation is an essential cog in the wheel of AI and machine learning. AI and machine learning are here to stay, so is data annotation. This industry will continue to grow while adding more nuances and complexities that AI and machine learning applications will be bringing into our daily lives. While the human loop is essential in data annotation, supporting them with data annotation tools will ensure the standards of quality required to enable effective data annotation support for AI/ML applications.

*For organisations on the digital transformation journey, agility is key in responding to a rapidly changing technology and business landscape. Now more than ever, it is crucial to deliver and exceed organisational expectations with a robust digital mindset backed by innovation. Enabling businesses to sense, learn, respond, and evolve like living organisms will be imperative for business excellence. A comprehensive yet modular suite of services is doing precisely that. Equipping organisations with intuitive decision-making automatically at scale, actionable insights based on real-time solutions, anytime/anywhere experience, and in-depth data visibility across functions leading to hyper-productivity, Live Enterprise is building connected organisations that are innovating collaboratively for the future.

For more information, contact infosysbpm@infosys.com

Infosys®
Navigate your next