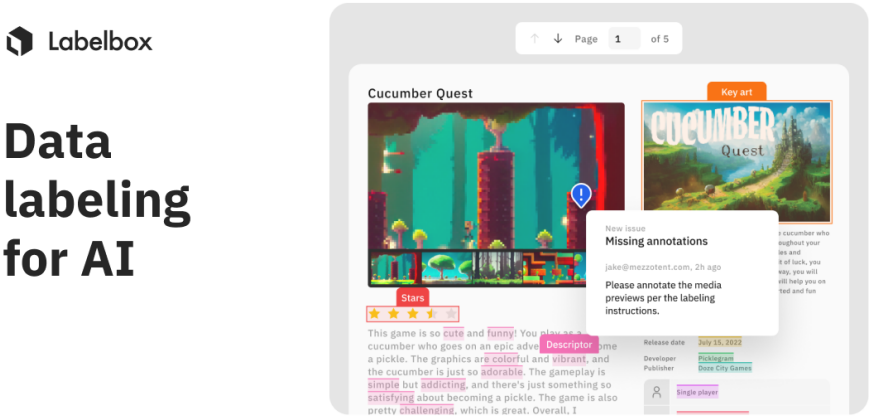
Data labeling for AI



Having an efficient data labeling process is an important foundation for any successful AI product. Your model is only as good as the data it's trained with, and part of the training process includes getting your data labeled quickly and accurately.

However, many companies typically approach this process by gathering and labeling as much data as they possibly can to train their model. In reality, AI teams need to focus on the quality of their data alongside quantity.

Having larger, low-quality datasets prolong the data labeling process and makes getting to production AI harder. Wading through a vast amount of unstructured data to get accurately labeled data requires a tremendous amount of patience, organization, and time. Ensuring that you have high quality data will save you time and money from decreased labeling costs.

What is data labeling?

Data labeling is the task of annotating data such as images, text, videos or audio with the purpose of helping to teach a machine learning model to make similar annotations. Labels can include bounding boxes and segmentation masks for image and text data, for example.

The data labeling process typically involves human-powered work in order to manually curate datasets, and in some cases, computer-assisted help. The types of labels are predetermined by a machine learning engineer and are chosen to give a machine learning model specific information about what is shown in the data in order to teach the model from these examples. Labels can be as simple as deciding whether a photo contains a human all the way down to labeling different parts of a human face such as the eyes, nose, lips, etc.

The process of data labeling also helps machine learning engineers hone in on important factors that determine the overall precision and accuracy of their model. Example considerations include possible naming and categorization issues, how to represent occluded objects, how to deal with parts of the image that are unrecognizable, etc.

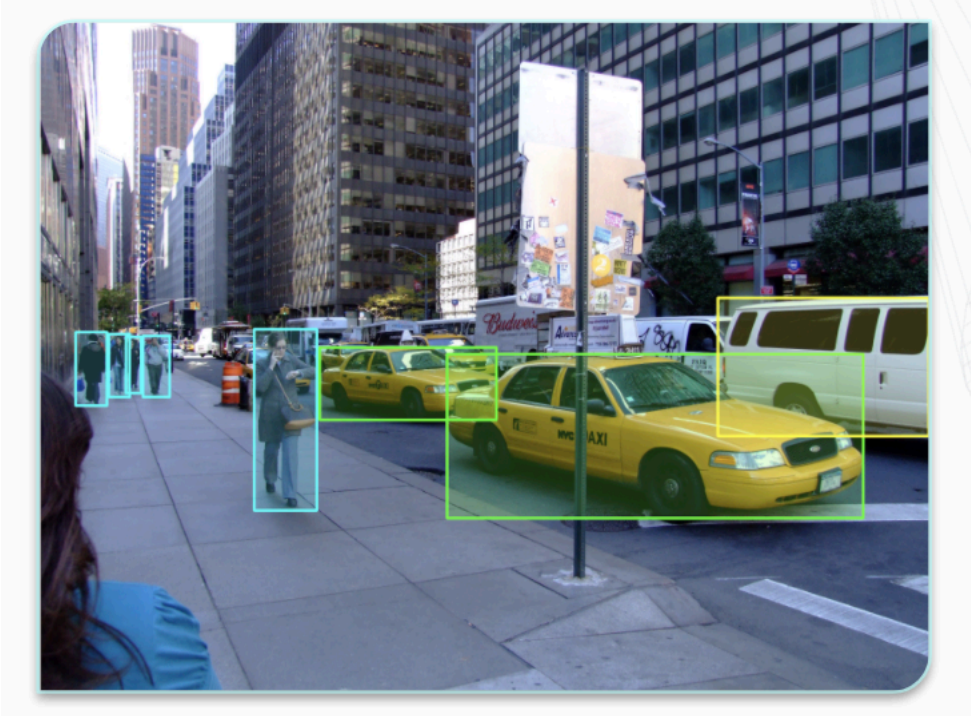
How does data labeling work and why is it important?

Data labeling is a central part of the data pre-processing workflow for machine learning. Data labeling structures data to make it meaningful.

This labeled data is then used to train a machine learning model to find “meaning” in new, relevantly similar data. Throughout this process, machine learning practitioners strive for both quality and quantity. Accurately labeled data coupled with a larger quantity creates more useful deep learning models, as the resulting machine learning model bases their decisions on all the labeled data.

To illustrate from the example below, a human labeler applies a series of labels on an image asset by applying bounding boxes to the relevant objects, otherwise known as image labeling or image annotation.

In the example below, pedestrians are marked in blue and taxis are marked in yellow, while trucks are marked in yellow. Accurately identifying the cars from the pedestrians will yield a more successful model, which is defined as a model that can make accurate predictions when presented with new data (which in this case, are images of objects in a street view).



This process is then repeated, and depending on the business use case and project, the quantity of labels on each image can vary. Some projects will require only one label to represent the content of an entire image (e.g. image classification). Other

projects could require multiple objects to be tagged within a single image, each with a different label (e.g., bounding boxes).

What are the different types of data labeling?

There are many fields of AI, each working with a different type of data and requiring different data labeling types. The most common fields are computer vision for image and video, natural language processing (NLP) for text, and audio processing for speech recognition.

Data labeling for computer vision with image and video

A computer vision model is built to interpret visual data from images and videos to identify, classify, and extract further information about objects that appear in the data. The data labeling process for this type of model includes labeling images, much like in the example above. The computer vision model would then be trained with the labeled data to categorize images, recognize the position of objects, or identify objects of importance in an image. A real-world use case for this type of model includes helping retailers manage inventory by identifying different products on a shelf and the quantity of their stock.

Data labeling for NLP

NLP is a branch of AI that gives models the ability to understand natural language as it is spoken or written. This form of data labeling requires labelers to identify important sections of text or tag text with specific labels to train the model. The model would then develop the ability to understand and interpret the text, even when it's worded slightly differently.

A common real-world use case for this model is a chatbot built for customer support. Using this model, a chatbot would be able to understand the question, “When is my package being delivered?” even when phrased differently by different customers, such as “When will my package be delivered?” or “What is the delivery date of my package?” and answer accordingly.

Audio processing for speech recognition

Audio processing converts sounds into structured data so it can be used for model training and improvement. This data labeling process actually goes hand-in-hand with NLP, as it typically requires the audio to first be transcribed into text before it is labeled.

A common real-world use case for this is any type of virtual assistant commands

← All guides

What is data labeling?

How does data labeling work and why is it important?

How does Labelbox support data labeling?

Data labeling projects begin by identifying and instructing human labelers (otherwise known as annotators) to perform labeling tasks. Annotators must be thoroughly trained on the specifications and guidelines of each annotation project, as every use case, team, and organization will have different requirements.

In the specific case of images and videos, once the annotators are trained on how to annotate or label the data, they will begin labeling hundreds or thousands of images or videos, often using home-grown or open-source labeling tools. Advanced AI teams, however, will have a data-centric AI platform that facilitates an efficient labeling process.

A data-centric AI platform is software that is designed to have all the necessary tools for labeling any data modality. This type of software also promotes an iterative approach to data labeling. Instead of using one large dataset to train your model, Labelbox equips AI teams with the tools they need to label data in smaller batches. This approach means AI teams give more supervision and feedback at the beginning of the project and create a more agile process. This type of approach prioritizes two-way collaboration between the labelers and AI teams to ensure that the data labeling process is efficient and accurate.

According to a recent study from Stanford University researchers, this agile, data-centric approach results in anywhere between a 10% to 50% reduction in the amount of training data needed depending on the task at hand. This in turn translates into significant time and cost savings during the data labeling process.

In addition to enabling this iterative approach to data labeling, Labelbox also include additional features that specifically help optimize your data labeling projects.

High-performance data labeling tools

When looking for the right AI platform for your team, it’s important to ensure that the software supports enough labels and annotations per asset without sacrificing loading times. This way, you’ll be able to use the AI platform for both simple and complex use cases, which may be a requirement in the future for your team.

Customization based on ontology requirements

The ability to configure an AI platform to your exact data structure (ontology) requirements enables you to ensure consistency and scalability in the data labeling process as your use cases expand. Labelbox provides a convenient way to copy your ontology across multiple projects so that you can make cascading changes or use an existing ontology as a starting point rather than starting from scratch.

Labelbox allows you to configure the label editor to your ontology requirements. Bring additional attachments such as text, videos, images, overlays, or even custom widgets to aid data labelers to create perfect labels.

An emphasis on performance for a wide array of devices

A data-centric AI platform includes an intuitive user interface, which helps lower the cognitive load on labelers and enables fast data labeling. Even on lower spec PCs and laptops, high performance is critical for professional annotators who are working in an editor all day.

A simple, intuitive UI reduces friction in the data labeling process.

Seamlessly connect your data via Python SDK or API for easy data labeling

Stream data into an AI platform and push labeled data into training environments like TensorFlow and PyTorch. Labelbox was built to be developer friendly and API-first, so you can use it as infrastructure to scale up and connect your ML models to accelerate data labeling productivity and orchestrate active learning.

Simplified data import without writing and maintaining your own scripts.

Benchmarks & consensus for data labeling

Quality is measured by both the consistency and the accuracy of labeled data. The industry standard methods for calculating data quality are benchmarks (aka gold standard), consensus, and review.

Figuring out what combination of these quality assurance procedures is right for your machine learning project is an essential part of an AI data scientist's job.

Quality assurance is an automated process that operates continuously throughout your training data development and improvement processes. With Labelbox consensus and benchmark features, you can automate consistency and accuracy tests. These tests allow you to customize the percentage of your data to test and the number of labelers that will annotate the test data.

Benchmarks in action, highlighting the example labeled asset with a gold star.

Collaboration and performance monitoring

Having an organized system to invite and supervise all your labelers during the data labeling process is important for both scalability and security. A data-centric AI platform should include granular options to invite users and review the work of each one.

With Labelbox, setting up a project and inviting new members is extremely easy, and there are many options for monitoring their performance, including statistics on seconds needed to label an image. You can implement several quality control mechanisms, including activating automatic consensus between different labelers or setting gold standard benchmarks.

Seamless collaboration between data science teams, domain experts, and dedicated internal & external labeling teams.

Final thoughts on data labeling

The traditional method of training your model with one large training dataset is no longer effective. Machine learning has moved past this approach to be more agile: carefully curating datasets to accelerate the data labeling process and train the model, examining its performance, and modifying the next dataset accordingly.

A data-centric AI platform like Labelbox promotes this iterative process and enables AI teams with tools to accelerate their data labeling process — empowering ML teams to create AI ready datasets. As such, investing in the right platform is key for deploying successful AI products. Try Labelbox for free.

Continue reading

Programmatically launch human data jobs for RLHF and evaluation

Learn how to harness the SDK to manage human data labeling jobs for RLHF and model evaluation. With just a few steps, you can set up the SDK, import various types of data,

Evaluating leading text-to-speech models

Discover how to employ a more comprehensive approach to evaluating leading text-to-speech models using both human preference ratings and automated evaluation techniques.

Metrics-based RAG Development with Labelbox

Learn how to optimize your Retrieval-Augmented Generation (RAG) applications by focusing on key metrics like context recall and precision.

Try Labelbox today

Get started for free or see how Labelbox can fit your specific needs by requesting a demo

Start for free

We enable breakthroughs