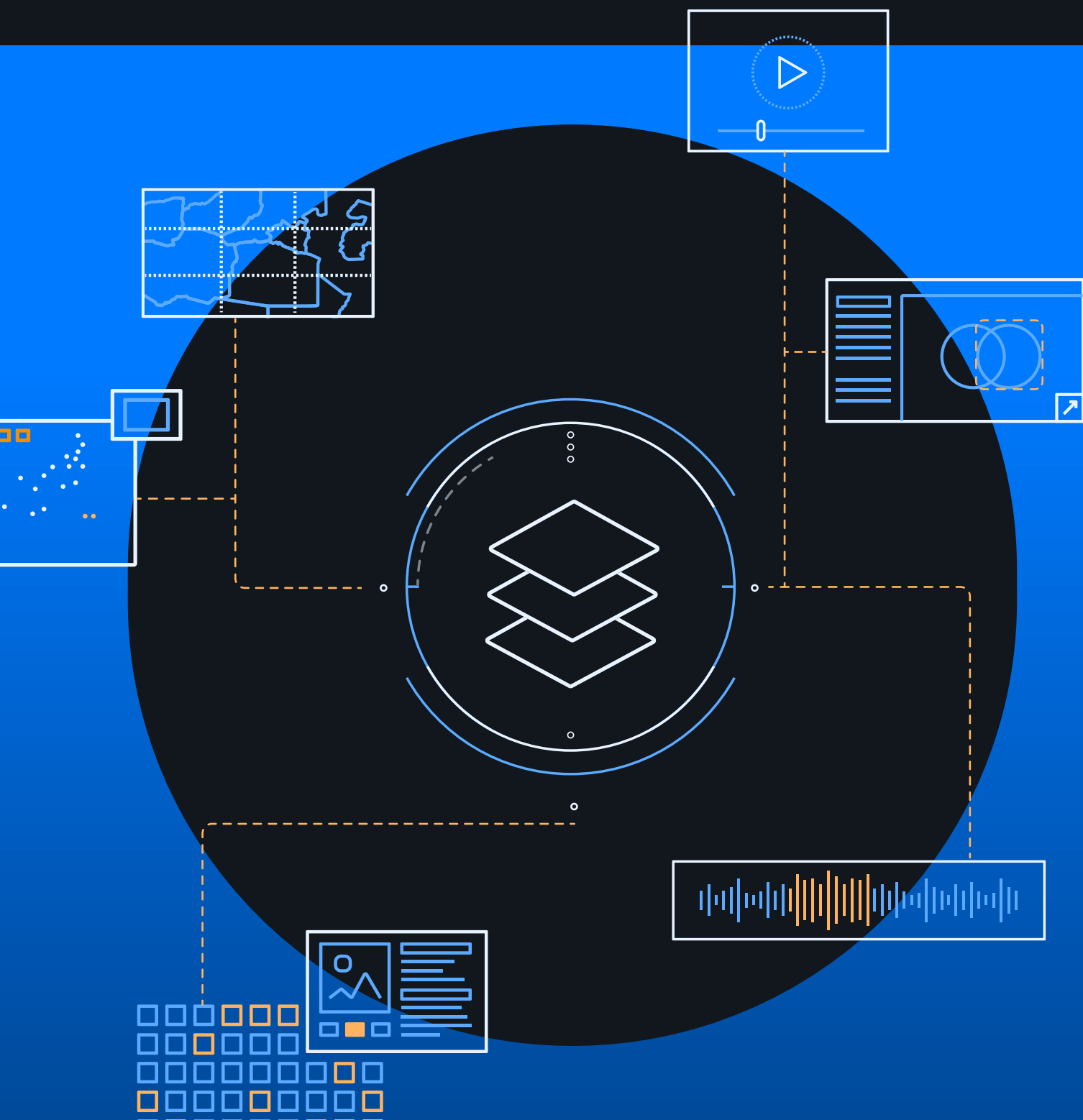


The complete guide to training data platforms



00

Content

Overview	1
Common challenges with AI data and model training	2
What is a training data platform	3
Who benefits from a training data platform?	5
When to start using a training data platform	6
Business value and ROI	7
How to choose a training data platform	12

01

Overview

There is a tremendous amount of data being generated around the world today. The task of collecting relevant information, organizing it, and piecing it together to make accurate predictions seems like a monumental task, but this is the everyday job of data science and machine learning teams.

In recent years, teams have capitalized on technological advancements in artificial intelligence and are making it possible for companies to deploy deep learning technology across nearly every industry, particularly with computer vision and natural language processing where cameras or the human eye were traditionally involved in making decisions. The trend towards broad business adoption has been accelerated by the emergence of enabling technologies such as open-source machine learning models and greater, less cost-prohibitive access to processing power.

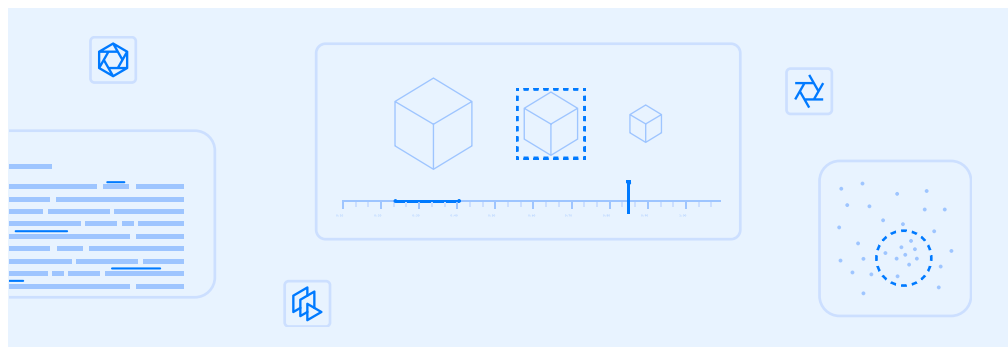
What stands in the way of AI realizing its full potential are the costs and uncertainties associated with creating training data. Teams often use an amalgam of home grown and open source tools to manage their training data or rely on external data labeling vendors. These solutions present myriad challenges to AI teams.

- Patchwork systems can be difficult to use, maintain, and scale over time
- Engineers end up spending more time preparing, augmenting, processing, and developing tooling for training data than actually training models
- External labeling services can be opaque with how they save, use, and process data, which presents considerable risks to enterprise AI teams working with proprietary data
- It can be hard to estimate the time and costs required to produce the training data for a specific project
- Ensuring that the training data produced is high quality can require quality management workflows that are difficult to implement and slow down the process

With more teams embracing a data-centric approach to AI, it's more important than ever to produce high-quality training data. To do this, AI teams need a powerful, reliable, integrated solution to optimize their training data processes: a training data platform.

A training data platform unites data, people, and processes in one place to empower teams with faster iteration and more accurate models.

An AI team armed with a training data platform can train their models more quickly, scale their labeling operations easily, and realize true AI-driven competitive advantages.



Common challenges with AI data and model training

With traditional ML training techniques, one of the primary challenges was to establish a faster iteration cycle for the model. With the increasingly popular (and successful) data-centric approach to ML, producing a steady stream of high-quality training data has become a priority. These processes — iterating on the model and producing quality training data — are not mutually exclusive, and improving one will likely improve the other.

Creating and iterating on training data

Before being fed into an ML framework, data has to be aggregated, transformed, cleaned, augmented, and — in most cases — labeled. This process consumes roughly 80% of the resources in an average ML project, far exceeding other categories like algorithm development, model training, and deployment. Data management, in other words, is the engine powering modern AI and ML.

Producing training data is no longer just about using a basic interface to draw boxes around objects. Teams that prioritize the quality of data — especially those moving towards a data-centric approach to AI — need to establish a robust labeling operation that addresses the following challenges.

Ineffective labeling tools and services

ML teams often use open source or home-grown tools for labeling their assets. These tools can present a myriad of challenges to its users, however. They might:

- Have a large learning curve or a less intuitive interface that slows down labelers
- Be limited in terms of what can be labeled and how, forcing teams to settle for a less precise ontology that will ultimately result in a less performant model
- Require labelers to waste time waiting for assets, depending on how assets are distributed
- Make it difficult or even impossible for remote labelers to do their job by requiring them to work odd hours or be onsite

Some ML teams turn to external labelers who use AI in combination with a labeling team to produce training data, but this often presents a security risk. Teams may have no control or visibility over what happens to their data when it's with the third party labeling service. Their labeled data — usually considered valuable IP — can be copied, sold, or used to train another algorithm.

Delays from quality management and labeling iteration

The integration of quality management workflows is essential to generating high-quality training data. ML teams often require the use of consensus scores, benchmarks, and review queues to identify unsatisfactorily labeled assets, provide feedback, and iterate on them. When added to a patchwork system of open source and in-house tooling, however, these additional workflows can delay an already slow process.

Poor ontology management

Labeling ontologies are usually unique to each ML project, and are often either created from scratch or copied and modified from an existing ontology for a similar project. Creating it may be simple at first, depending on the use case, but the task will likely grow in complexity as labeling teams get to work and come across edge cases that don't neatly fit into the categories and/or classes identified in the ontology. When ontologies are edited during an ongoing labeling project however, labels created before the change may need to be edited as well. If these are missed, the quality of training data will be reduced.

Miscommunication

Turning raw data into accurate and consistent training data is a team effort. Engineers, domain experts, labelers, product leaders, and managers must work together while playing different roles. During the labeling process, labelers may have questions that need to be answered by domain experts or project managers. Often, issues found by labelers will require managers to edit the ontology and/or labeling instructions.

This makes collaboration an integral part of the labeling process, and lack of a standardized, easy, and timely way to talk to each other can result in unresolved issues that lower the quality of training data.

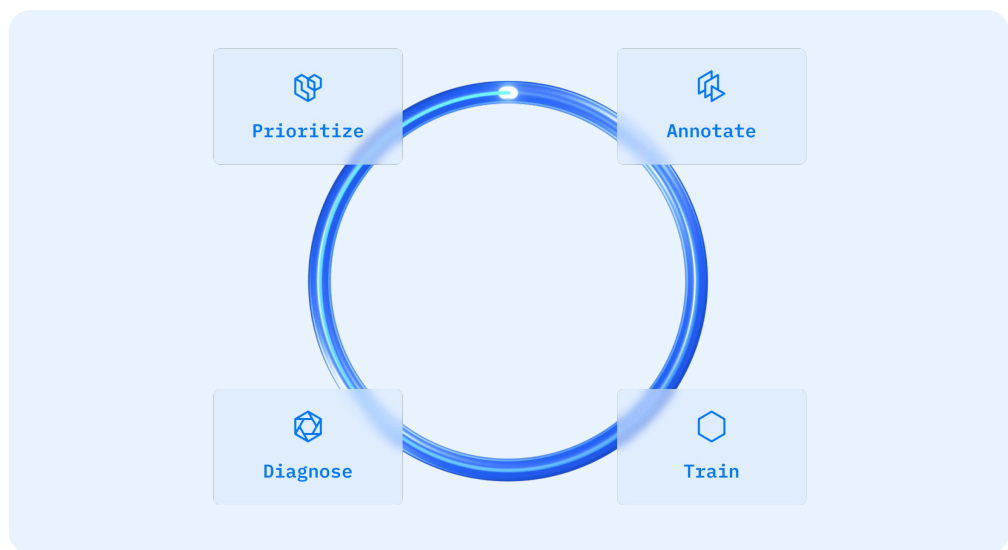
Creating an efficient model iterating process

Even if an ML team has devised a fast, seamless pipeline to create high-quality training data, they still have to contend with the model iteration cycle, which often takes weeks — a sluggish pace compared to the minutes-to-hours speed of the standard cycle for software development.

While those dedicated to advancing ML technology work on the problem of iteration speed, mature ML teams have found that gaining a better understanding of model performance and curating subsequent training datasets accordingly can significantly increase model performance from one cycle to the next. Typically, models require increasing amounts of training data for smaller performance gains as the number of iterations increases.

With this active learning method based on model performance, however, teams are realizing steady performance gains with relatively small amounts of training data.

To use this method, however, ML teams need a way to gain a deeper understanding of their model performance than typical aggregate metrics provide, and a way to explore their unstructured data, organize it, and curate their datasets. Tooling for these requirements is rare and difficult to build, leaving ML teams without a dedicated training data platform with slow and less effective iterations.



Taking supervised learning into production AI is a continuous process similar to DevOps, with workflows that emphasize consistent testing, deployment, and iteration.

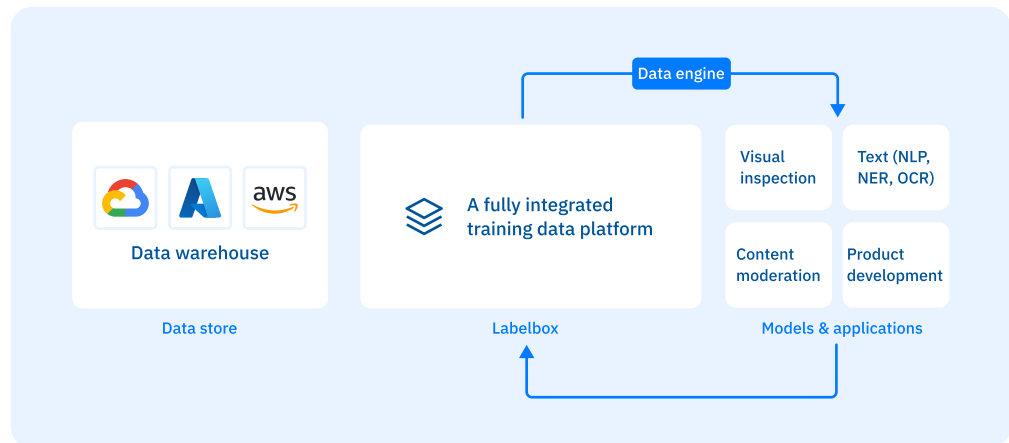
03 What is a training data platform?

A training data platform (TDP) empowers teams with software solutions to label data efficiently and accurately for all their AI projects by bringing together all the people, processes, and data in one place. The primary benefits of a typical TDP include giving organizations the ability to create and track their labeled data and to optimize the quality of their training data before feeding it into their machine-learning models.

The TDP functions not merely as a tool, but helps to define a new way of working that enables companies to become more effective with their training data efforts. This is typically accomplished by supplying users with a convenient interface to handle common workflows such as project creation, ontology set up, the labeling operation itself.

Best-in-class TDPs also provide advanced features that combine collaboration and easy setup for experimental workflows with labeling automation and labeler performance metrics to help AI teams set up the most efficient labeling operation for high quality training data.

A leading TDP will also offer model error analysis tools and solutions to help visualize, filter, and organize large unstructured datasets for advanced active learning methods that are proven to speed up model iteration and increase performance.



Benefits of a TDP

Most training data platforms in the market will offer the following benefits and features:

1. **Easy uploading via raw data or link uploads**, which can scale to support millions of pieces of data
2. **Support for common data types**, including image, video, and text data
3. **Fast and configurable annotation tools** to support all types of labeling, including polygon, rectangle, line and point segmentation, as well as assisted annotation
4. **Flexible editing tools to meet common ontology requirements**, such as custom attributes, hierarchical relationships, and infinite nesting
5. **Real-time labeling queuing** to enable the management of hundreds or thousands of labelers all working in parallel
6. **Role and permission-based access** to conveniently give users secure access to your organization and specific projects
7. **Complete visibility, security, and control over all your data** throughout the labeling process. Labeled data is usually considered IP, and should be secured at all times
8. **Seamless integration for quality management workflows** such as consensus, benchmarking, and review queues
9. **API support** for fully automated human in the loop ML pipelines

Benefits of a best-in-class TDP

An exceptional TDP will also offer additional features that enable ML teams to build a powerful data engine for all their AI requirements. This includes a fully optimized labeling operation that can be partially automated, used for any data type, and configured to meet the specific requirements of any AI project. It also includes support for advanced training data and model training techniques such as active learning, with tools for in-depth model error analysis and unstructured data curation.



A way to **visualize, explore, and organize large unstructured datasets** of multiple data types. This feature can be used to curate smaller datasets for labeling for AI teams using active learning methods. Selecting the data most likely to improve model performance will help the model reach required accuracy levels faster and with less training data.



Automated labeling workflows such as **model-assisted labeling**, which save significant time and costs. Automating part of the labeling process helps generate training data faster, speeding up model iteration.



Support for a **wide range of data types**, including less common formats such as tiled images, video, DICOM images, audio data, and PDF documents.



Built-in collaboration tools to enable labelers and other stakeholders ask and answer questions, raise issues, and receive feedback within the labeling interface and directly on the asset instead of toggling between applications.



An **SDK and/or API** that allows teams to configure the labeling tool as required for their specific use case and use any feature programmatically. An SDK or API can also seamlessly connect the TDP to the team's data warehouse and model training systems so that data can flow throughout the engine with little manual intervention.



Model diagnostics tools to empower AI teams with model performance metrics that go deeper than a typical aggregate analysis. This will help the team pinpoint areas of low confidence and find edge cases. Information can then be used to select the next dataset for labeling. This active learning method is proven to produce and maintain high accuracy levels in algorithms.



Labeling operations services as an optional addition to best-in-class software, helping ML teams build their ontologies, find external labeling teams, create labeling instructions, monitor labeling performance and quality, and more

04

Who benefits from a training data platform?

Organizations investing in AI will need to equip themselves with the tools and processes to produce high-quality training data in order to build performant models and applications. Below are the specific types of team members that can derive outsized benefits when adopting a training data platform.

AI product managers will be empowered to quickly deploy and maintain applications with timely insights into labeling productivity, label quality, and improved integration between the models and data labeling pipelines. They'll also be able to implement advanced workflows to automate labeling, leverage active learning techniques, and build a data engine that optimizes model training.

Data scientists will be able to ensure that labeled data (the source code for ML model) is high quality and reflects required business objectives.

Business leaders will be able to track areas of AI investment to key business priorities to better correlate ML projects to business value, as well as develop and maintain high-performing production AI fast to gain and keep competitive edge.

Labeling operations leaders will be able to measure labeling efficiency and performance to optimize workflows.

Labeling teams can access the fastest and most intuitive tools for labeling training data.

05

When to start using a training data platform

Many enterprise ML teams find that they outgrow in-house and open-source labeling solutions. Even small teams with simple and/or short term labeling projects can benefit from a TDP — and may be able to realize many operational efficiencies for free! If you and your ML team are experiencing any of the challenges listed below, you may want to consider moving to a TDP.

Your ML and data science team is spending too much valuable time building and maintaining training data infrastructure. As is often the case with key business infrastructure, there are hidden costs to building. Investing in a best-in-class solution might

² [The AI-Powered Enterprise](#), CapGemini Research Institute, 2020

seem more expensive in the short term, but a training data platform will likely end up saving your team hours of time and expertise (not to mention dollars) in the long run, especially as your ML efforts scale.

Your training data workflows are starting to become unreliable and pose risks to your IP.

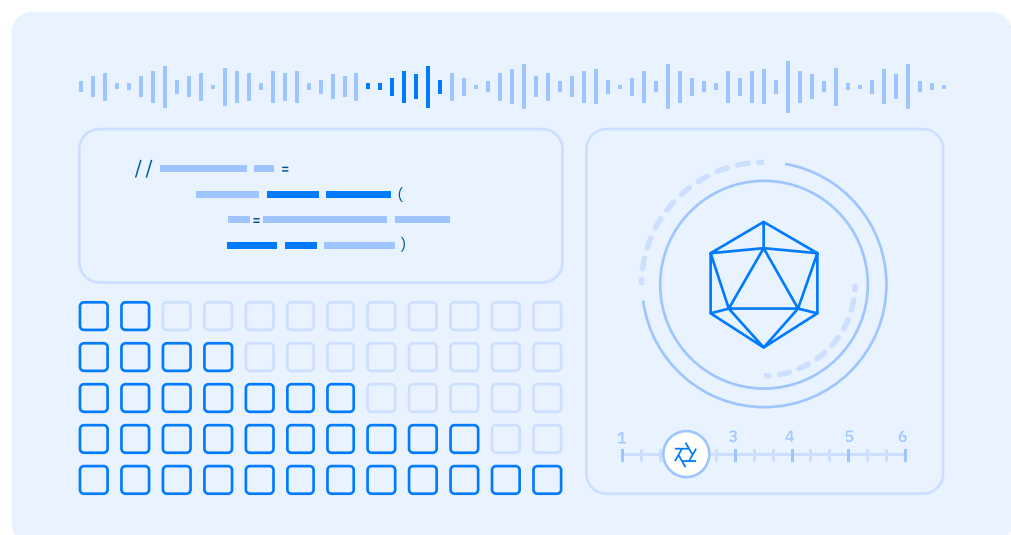
A TDP can support both cloud-hosted environments and on-premises solutions to equip data labelers with secure access to datasets. They can also provide extra security measures such as encryption at rest with AES-256 and Auth0 for authentication.

Scaling your labeling pipeline is becoming a bottleneck. TDPs are built to handle large amounts of data and enable large teams to work in parallel as they label, review, and manage training data. These platforms make it easy to grow your ML team and label more data.

You need to diversify your training data labeling services. With a TDP, you'll have full visibility and granular insight into the performance of labeling teams, freeing you from dependence on any one vendor for your labeling services. You can add and remove labelers easily from any project and ensure that only the most qualified teams work on your use case.

You have multiple ML teams in your organizations as a shared service. A TDP can function as a single source of truth for defining, storing, and accessing training data across an entire organization.

You need greater visibility into preventing bias in your AI product. A TDP will make it easy for your team to add quality management measures such as benchmarking, consensus, and review workflows without creating bottlenecks in your labeling pipeline. If you choose to invest in a best-in-class TDP, you may also be able to access model diagnostics and data curation features that help you gain a better understanding of model performance and compile training datasets accordingly.



06

Business Value and ROI

As data labeling needs scale, ML teams will require more data management, quality and performance monitoring, easier ways to experiment with and implement advanced techniques to improve the speed and efficiency of their labeling operations, and quality assurance processes. Even teams with smaller scale projects and datasets, however, can realize significant benefits from a TDP.

Compared to those using in-house labeling tools, a TDP helps teams save months in custom R&D, given that it can be difficult to accurately define the scope and construct a solution for needs across engineering and product groups. Building dedicated software to handle one of the most time intensive parts of their workflow means spending less time on infrastructure planning, resource allocation, and preparing for the unknown. In the absence of a TDP, companies find that their internal tools are generally not built for usability, scalability, or cross-team support.

In addition to reducing costs, organizations armed with a TDP across industries and of every size and maturity level have developed higher-quality training data and deployed performant, high-value applications. A [session at a recent virtual conference](#) covered the benefits that ML teams realized within just one year of using a TDP.

“We've seen a 40% increase in the speed of annotations, increased the quality of our annotations, and there was a massive reduction in the back and forth we had to do.”

Hong Noh, Senior Product Manager at Criteo

“Having a tool to collect metrics on agreement and errors is helpful. We can communicate what we want labeled correctly, and have a review process.”

Matthew McAuley, Senior Data Scientist at Allstate

07

How to choose a training data platform

Choosing the right TDP for your organization is no easy task. More options are available now than ever before, and you'll need to ensure that the platform will meet all your immediate and future requirements for high-quality training data.

Be sure to consider the following as you shop:

- What data types will you need to label? Will you need to label different types of data in the future?
- Who will be labeling your data, and how will they access it? Will your labeling team grow in the future?
- How will your data need to be labeled? Your TDP will need to support all current and future labeling methods and ontologies.
- Will your team need to use advanced labeling techniques to increase speed and efficiency, such as automated labeling?
- Can your TDP provide full visibility and control over your data throughout the labeling process? Don't forget that your labeled data is a valuable asset; many enterprises consider it as IP.
- How will your TDP help you increase iteration speed as your model trains?

For a more comprehensive checklist of what to look for in a TDP, check out our [definitive guide to buying a training data platform](#).

Labelbox is an enterprise-grade, best-in-class training data platform built by and for those who create AI. Every day, organizations of all sizes, including Fortune 500 companies and leading startup teams, use Labelbox to annotate, manage, and iterate on their training data — as well as gain a deeper understanding of model performance during and after training, and explore and organize their unstructured data. Watch [this brief video demo to learn how it works](#).

Labelbox

Learn more about our offerings, sign up for a demo, or start using our free version today at www.labelbox.com.