HumanSignal

humansignal.com

Data Labeling Overview for Machine Learning and Data Science

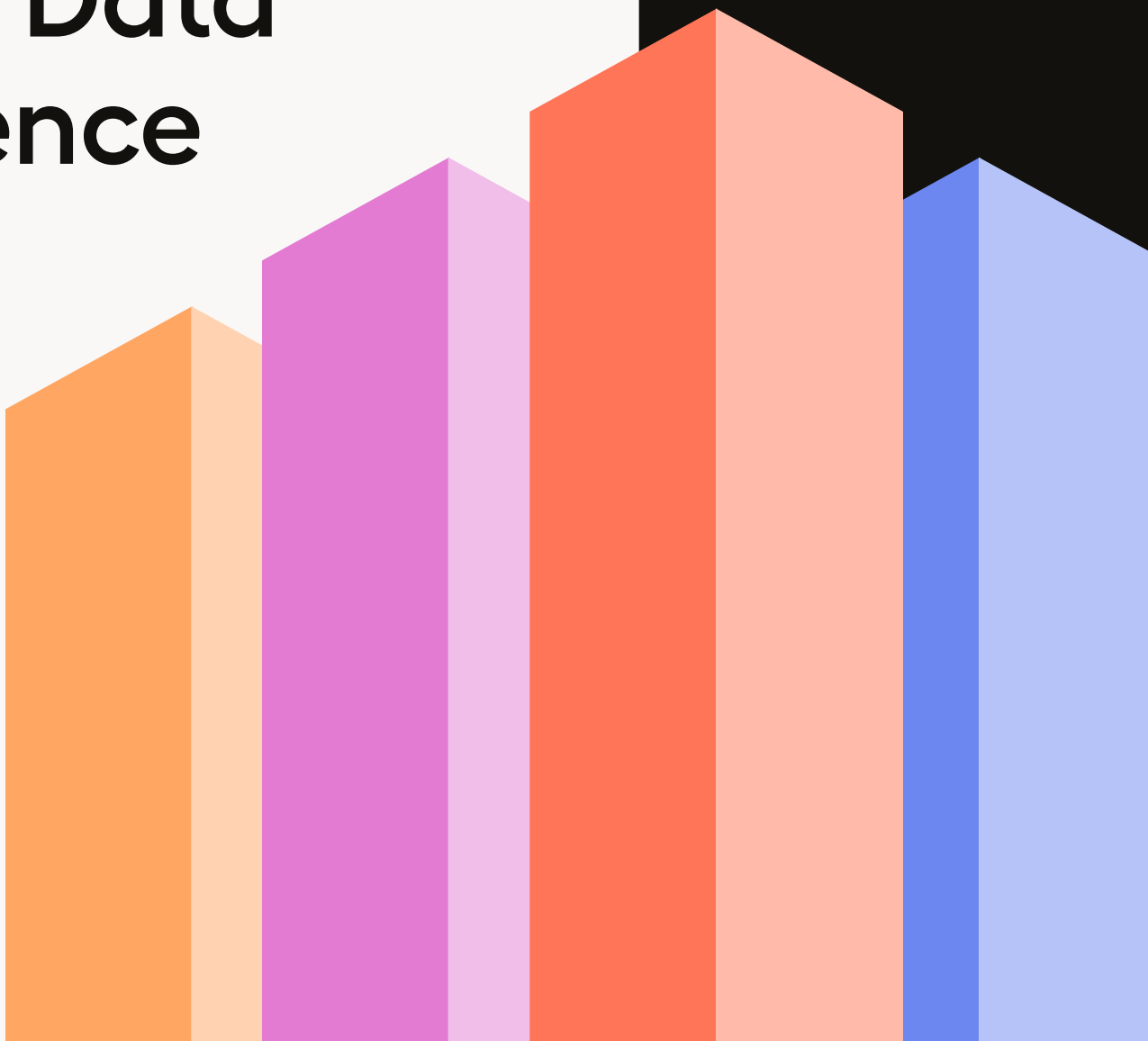


Table of Contents

Introduction

The paradox is that data is the most overvalued and undervalued aspect of machine learning today. While having data is a good foundation for AI, raw data on its own doesn't have enough built-in meaning to teach or train a machine learning model. However, when that data is given proper context, organized effectively, and trained, companies can use it to perform meaningful ML tasks.

With this knowledge, there has recently been a paradigm shift in the AI domain from a "model/algorithm" focus to a "data-centric" focus. This shift aims to optimize the data quality processes involving data collection, preparation, and processing to more effectively program ML/AI models.

A study by McKinsey shows that teams are now investing as much as 80% of their time in data preparation and processing for the best AI/ML outcomes. This primary focus on the quality of the data and its related processes is termed a data-centric AI approach.

Data-centric AI is a rapidly growing, data-first approach to building AI systems using high-quality data from the start and continually enhancing the dataset to improve the machine learning model's performance. One fundamental aspect of the data-centric AI approach is data labeling, where informative labels are added to raw data samples to provide proper context to raw data for training machine learning models.

The Importance of Data Labeling

Data labeling adds a critical layer of metadata that draws the connection between raw data and the exact prediction your model is learning to make. For example, to build an ML algorithm that can identify defects on a production line, you'd need approximately thousands of images of different properly labeled product defects, with which data scientists can train the ML model to identify defects.

Data labeling is an indispensable part of the ML process and optimizing data labeling quality, accuracy, and speed is of utmost importance. An internal data labeling team that oversees a combination of manual labeling by subject matter experts, programmatic labeling with a human-in-the-loop model, and even outsourcing labeling for common knowledge tasks is one data labeling strategy that is becoming increasingly popular.

This data labeling whitepaper outlines four core pillars of data labeling: Data, People, Process and Technology, and highlights how to operationalize these pillars to build an efficient and scalable data labeling process for successful ML outcomes.

Data

Optimizing the data collection and storage process through a solid data pipeline.

Technology

Technology provides the tools that the other three pillars use to implement their processes.

People

Data labeling requires human expertise. By automating as much as possible, we can focus human time on the most strategic and impactful tasks.

Process

Organizations need to have a well-designed, efficient, and scalable process that they can actively monitor to ensure high-quality and accurate results.

01

Data

Optimizing the data collection and storage process through a solid data pipeline is the first step in creating an effective data labeling process for ML. You'll need a pipeline that collects the correct data (and its related metadata) and stores it in the proper format to enable data labeling and future discoverability.

Machine learning projects have data originating from a variety of sources. Datasets may include structured and unstructured data, public and proprietary data, and different data types, such as images, audio, documents, and videos. So, ensuring that you have the correct data in a suitable format is essential before data labeling can commence.

Here are some standard guidelines for assessing and preparing data for labeling and ML.

Understand The Problem

Before collecting data, you must first define the problem you're trying to solve. It would be best to answer questions like:

- What is the goal of the ML project?
- How much data would you need?
- Where is that data going to come from?
- Are you going to be making use of supervised or unsupervised learning?

Asking questions like this early on helps ensure that the data labeling and ML process is more successful. For example, you might want to leverage supervised learning if you are solving a classification or regression problem, like determining if a patient has a disease or predicting stock market price.

On the other hand, unsupervised learning is a better fit to solve clustering or associative problems such as customer segmentation or building recommendation systems. While supervised learning uses labeled input and output data, an unsupervised learning algorithm does not. Hence, the problem you want to solve will determine the data you will collect and the labeling used.

Collect the Right Type of Data

The quality of your data labels for your ML model is only as accurate as the data you feed into it. You'll need to include input data that is useful for predicting your target outcome. For example, to build an ML model for detecting credit card fraud, you'll need information on the transaction amount, the transaction location, and the cardholder's primary location. You'll also need examples of both the positive and negative outcomes (legitimate and fraudulent transactions).

Additionally, you'll need a significant number of examples for the model to feel confident in its predictions. With more data, you can get a higher degree of accuracy. Also, collect and use data that is an unbiased representation of the population or distribution you're modeling. For example, suppose manual credit systems give certain groups of people poor credit ratings, and you plan to label that data to train the ML models. In this case, the models will replicate and may amplify the original system's biases. Therefore, at every stage of data collection, it is important to question where the data is coming from, whose bias affected earlier decisions, and what changes need to be made in the data accordingly to use it for ML purposes.

Data collection is an iterative process because even when the model has been deployed, users will add new datasets in production, and the loop continues.

Store Data Using the Correct Methods

To carry out effective labeling and increase the accuracy of the ML models, you need to ensure that the data you collect is stored in the right format. Storing the data and associated labels together in a big data store like a data warehouse or lake simplifies the management and understanding of your dataset and any edge cases contained within it.

When examining the storage requirements for AI workloads, consider scalability and accessibility. Your storage system must be able to scale to meet the demands of the ML models as data increases. Depending on your use case and data privacy concerns, you might decide to store data on-prem or utilize cloud storage. Cloud object storage is becoming increasingly popular, as it offers the scalability uniquely suited to support massive quantities of data.

Data must also be continuously accessible from the data stores. Consider how your data will be accessed by the team members (Data Scientists, Data Engineers, and Data Labelers) who require it. The storage platform you choose also needs to support robust APIs to accommodate varieties of data for the different people who require it.

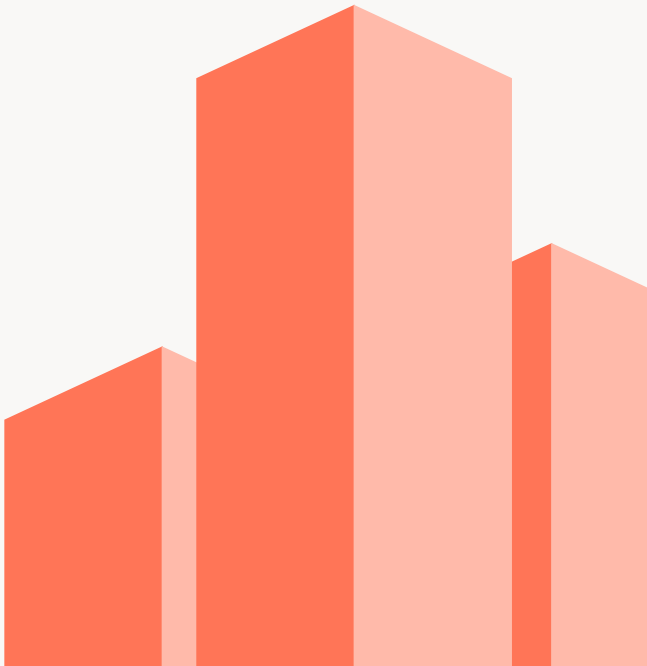
Ensure Data Consistency

Bad data can significantly impact your labeling and the performance of ML models. Clean the data by identifying and correcting or deleting errors, noise, and missing values and by making the data consistent. For example, if the values in a data field have different units (e.g., inches, meters, feet), convert them to a standard unit. Additionally, identify all the different representations of the same value and convert them to a single representation. For example, convert “Dir of Engineering,” “Director of Engineering,” “Dir Engineering,” and other variations as appropriate.

Missing values are another typical issue with data. Even for the same data field, missing values can be represented in various ways: NULL, None, NaN, and so forth. The more missing values, the less useful the data. Investigating why the data is missing data is also essential.

Leverage Data Pipelines

Use data pipelines to seamlessly do the data collection, formatting, aggregation, and transformation. To ensure data quality, you can also add automatic unit and end-to-end tests to the pipeline process. These tests help ensure that there are no missing or erroneous values or duplicates.



02

People

Data labeling requires human expertise. Of course, we want to automate as much as possible in the process so we can focus human time on the most strategic and impactful tasks. For example, using humans to transcribe and decode speech to text will be laborious; machines can do that faster. But humans provide context, experience, and reasoning to augment the automation process. Human minds are not rigid structures like machines. They offer better precision and accuracy in data annotation projects, especially when analyzing sentiment and emotion.

Consider the example of sarcasm in social media content moderation. A Facebook post might read, “Gosh, you’re so smart!” However, that could be sarcastic in a way that a robot would miss.

Your data labeling team will most likely be an offshoot of your data science team, so having the right people and group structure in the data science team will lead to better strategies and more effective processes for data labeling. While there isn’t an outright data team management standard, a few proven methods exist to get the best result for your ML teams.

Define Team Roles and Responsibilities

Data science is a team sport. Structure your team in a way that each member of the team is assigned tasks that fit their strengths. For example:

Data Engineers

Responsible for creating pipelines that convert data into usable formats for other data consumers

Data Annotators

Help enrich data by labeling it so machines can recognize it

Data Scientists

Responsible for extracting value from data using statistical techniques and machine learning

MLOps Engineers

Deploy machine learning models and ensure they are functional in production

That said, for smaller teams, these roles would most likely overlap. However, it's imperative people understand how they fit into a process. Each data team member should understand the process, their role, and what they need to achieve.

For the labeling team to be successful, you will need to develop a repeatable process and leverage tools and technology to centrally manage workflows, assign tasks, and assess results for quality. Data annotation managers are responsible for managing and monitoring the data labeling team and for assessing quality to ensure labeled data is accurate.

Leverage Domain Experts and Internal Data Labeling

Domain knowledge is valuable in labeling data for effective ML. Subject matter experts can often detect minor discrepancies and edge cases that others might not see. Collaborating with domain experts to create and label datasets is essential to build ML models that deliver value. For example, in the agricultural industry, ML teams could collaborate with farmers and others knowledgeable about the farming sector's practical issues to annotate data that adequately represent the problem being solved.

However, not all data labeling tasks require domain knowledge of 20 years of experience. You can find ways to share domain expert knowledge with the rest of the team by providing labeling instructions via documented text and visuals.

Improve Communication

Effective communication is an important component of the success of any data science team and is relevant to every aspect of a data science or ML project. Promote open communication avenues where data labelers, scientists, and other stakeholders can easily collaborate, share ideas, and suggest improvements. Short daily standup meetings are a great example of this, where team members can share status updates and highlight potential blockers with the team. Based on our experience, we also recommend a tightly closed feedback loop for communicating with your labeling team. This helps to make fast changes, such as modifying your labeling workflow or iterating data features.

Based on research by Hivemind, working with an internal group of full-time data annotators and interacting with them via open lines of communication will often yield better results than third-party data labeling services. However, for large data volumes, a common strategy is to leverage technology that helps identify the most important subset of data to label, or to programmatically label the bulk of the dataset, and then apply internal data labeling resources to complete or review the highest value tasks. Whichever strategy you choose, ensure every member of the data labeling team is educated on the project rules and requirements. Doing this makes it easier to ensure data labeling quality.

Address Human Bias

Humans are naturally biased, and human annotators can unconsciously bring their biases to the table when labeling datasets. The impact of bias can be harmful to algorithm performance and the ML product. For example, in 2015, Amazon realized that the algorithm used for hiring employees was found to be biased against women. This occurred because the algorithm's dataset was based on the information within the resumes submitted over the past 10 years, which belonged mostly to men.

Additionally, cultural and social perspectives can contribute to bias when labeling data. To build an AI system that makes unbiased data-driven decisions, the internal labeling team can discuss annotation differences and clean the training data to be free from conscious and unconscious assumptions on race, gender, or other ideological concepts.

Working with an internal group of full-time data annotators and interacting with them via open lines of communication will often yield better results than third-party data labeling services.

03

Process

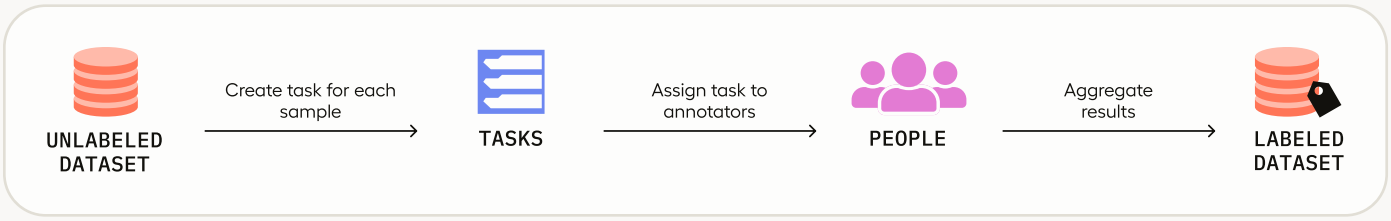
For a machine learning project to be successful, organizations need to have a well-designed, efficient, and scalable process that they can actively monitor to ensure high-quality and accurate results. Data labeling sits at the core of the ML process. Think of it as an assembly line that takes source data as raw inputs and creates meaningful metadata in a format that machine learning algorithms can understand and use to make predictions as outputs.

Data labeling is an iterative process. The data labeling process isn't complete after you ship the first ML version. There is a lot of trial and error involved. Based on feedback from the model, you are constantly optimizing the models and redefining your goals for the next deployment iteration. Therefore, there has to be a process for constant feedback, monitoring, optimization, and testing.

While every data labeling project is unique, typically, projects will align to one of three common stages.

Data Labeling for Initial Model Training

The data labeling process for training your initial ML model encompasses generating the unlabeled dataset and breaking it into labeling tasks assigned to annotators who follow instructions to label each sample properly. After assessing quality and fixing discrepancies, the process completes with your labeled data set, which is fed to your ML models.



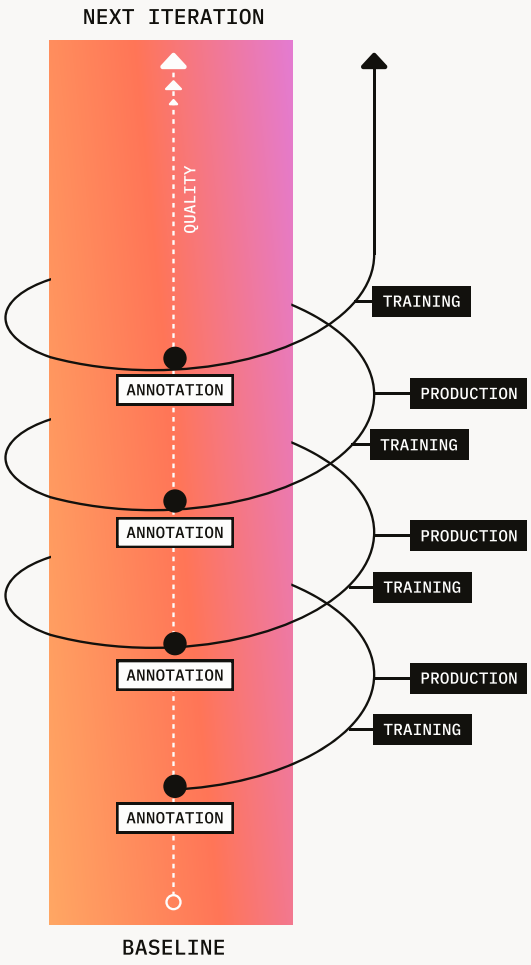
Data Labeling for Model Fine-tuning

Once a model is in production, you may need to label additional data to correct a model whose accuracy has degraded or because there are now new data that, once labeled, will enhance the model's predictive accuracy. For example, a sentiment classifier for an eCommerce site pre-trained on data from eBay might not perform very well with your user data. This scenario is one of the examples of data drift—situations when the original dataset differs from the actual data being sent to the model. Data drift is a signal to continue the labeling process to get the model back on track.

Here, the data labeling process involves running the pre-trained model over the initial dataset and logging the predictions. After this, we explore how the pre-trained model performs with our dataset. For example, is there a specific bias common to the algorithm, and in what areas is it failing? Taking into account this analysis, we can start relabeling our new dataset. We recommend you follow the workflow below:

1. Start labeling a fraction of your data that will be aligned with the dataset distribution you want.
2. Once you sense the data, you can start distributing labeling randomly.

Once this is done, prepare your dataset for training, test the model again, and keep iterating the process until you get more accurate results.



Human in the Loop and Active Learning

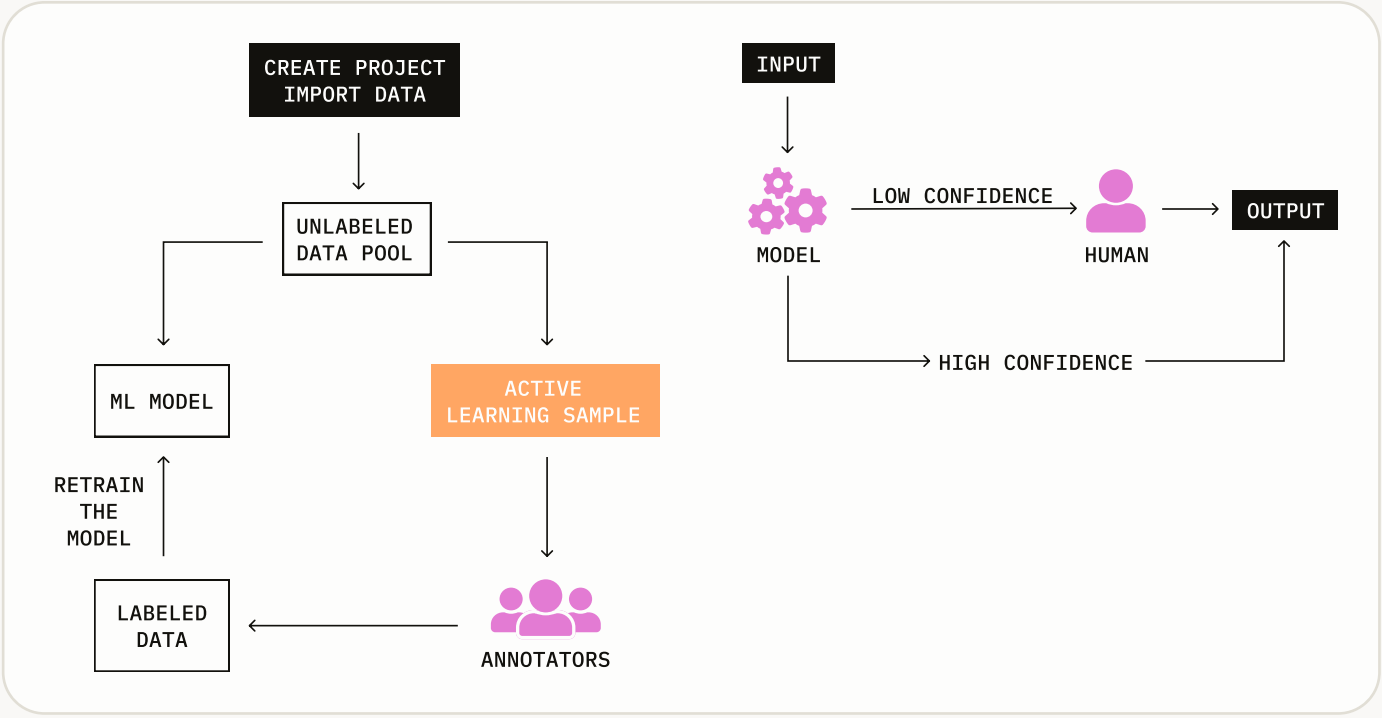
Because of the tremendous quantity of data needed to train ML models, manual data labeling can be time consuming. One way to speed up the data labeling process is to combine manual data labeling with artificial intelligence to enrich and label a dataset faster, a technique known as “Human in the loop” (HITL).

The HITL process begins by having humans label a dataset sample for training a seed model. The seed model is then used to train machines to identify and label data points in a transfer-learning model. This transfer-learning model automatically labels the rest of the dataset used in the ML model.

Once the model is in production, it often uses a prediction (certainty) score to automatically send high-confidence predictions to users and low-confidence predictions to humans to review, for example, if a machine is unconfident about a certain decision, like if a particular image is a cat. Humans can score those decisions, effectively telling the model, “yes, this is a cat” or “nope, it’s a dog.” This approach of humans handling low-confidence predictions and feeding them back to the model to learn from is referred to as active learning.

In active learning, the model gives answers when it’s sure and confident about what it knows, but it’s free to ask questions to the human whenever it is in doubt or wants to learn more. HITL encompasses active learning approaches as well as the creation of datasets through human labeling. This mechanism is a constant process that occurs all through the ML life cycle, from initial training to model fine-tuning. It allows humans and machines to interact continuously, making development faster and more cost efficient.

The key to a well-oiled data labeling process is tracking each process time, taking constant feedback from team members, and constantly looking for improvement opportunities. Here are some best practices to improve your data labeling process.



Create a Tagging Taxonomy That's Unique to Your Use Case

Tagging taxonomies help create a shared framework and understanding of an organization's domain, products, and services. See an example of Google's taxonomy site.

Creating tagging taxonomies lays the foundation for effective data labeling; they allow annotators to label and classify datasets quickly. Suppose we have a large dataset of images taken from cameras on the street. Each image may feature a variety of objects useful in creating algorithms for driverless cars. In order to extract the information from the images, we must first define a hierarchical representation of that information. After doing this, we then begin the labeling process by taking these raw, unlabeled images and labeling them with the high-level classes (e.g., 'vehicles,' 'street signs,' 'traffic signals,' and 'pedestrians').

> **In order to extract the information from the images, we must first define a hierarchical representation of that information.**

Design Good Instructions

Good instructions are an essential factor in getting good human labeling results. Instructions give the human label information about how to apply labels to your data and help maintain consistency. The instructions should provide sample labeled data as well as other explicit instructions.

Avoid making the instructions too long. It is ideal if a labeler can review and understand them within 20 minutes. The instructions should describe the task's concept as well as specifics on how to categorize the data. For example, for a bounding box labeling task, describe how you want labelers to draw the bounding box. Should it be a tight box or a loose box? If there are several instances of the object, should they draw one big bounding box or multiple smaller boxes?

> **The instructions should describe the task's concept as well as specifics on how to categorize the data.**

Employ Data-specific Labeling Techniques

Consider the type of data you will be labeling and choose a data annotation technique.
There are some standard techniques for labeling different data types, including:

Image Classification

Image annotation is used to mark objects and key points on an image. Image annotation helps train ML algorithms to identify objects for medical imaging, sports analytics, and even fashion. Examples here include bounding boxes, Polygonal Segmentation, Semantic Segmentation, key points, and Landmark.

Text Annotation

Text data annotation is used for natural language processing (NLP) so that AI programs can understand what humans are saying and act on their commands. Examples include Sentiment analysis, Intent, Semantic, Named entity recognition (NER), and linguistics.

Voice Annotation

Voice/audio annotation is used to produce training data for chatbots, virtual assistants, and voice recognition systems. Types of voice annotation include Speech Labeling, music classification, and Speech-to-Text Transcription.

Video Annotation

Video annotation is used to power machine learning algorithms used in self-driving cars and for motion capture technology used for making animations and video games. Video annotation combines image, text, and voice annotation techniques because video data has all three components.

The goal is to ensure that the labels perfectly match the data type the model will utilize. Choosing the wrong technique can completely alter the model's training.

Establish a Quality Assurance Process

Conduct a QA check by regularly auditing data labeling tasks, including examining the data labeling process from start to finish and bringing in subject matter experts to check the accuracy of the labels. Compare the annotations in your dataset with an ideal set of annotations to verify that your model accurately reflects the real-life conditions in which you plan to use it.

Ensure everyone in your team is adding labels the same way, consistently. If not, there will be a lot of confusion in the dataset. For example, labeling helicopters as “helicopter” and “chopper” in the same dataset can confuse the model. Let multiple annotators label the same samples and use agreement matrices to identify and resolve labeling issues.

Also, regularly test the quality and performance of the overall data labeling process by randomly selecting samples to review and tracking specific metrics, like time taken to label and label accuracy over time, to evaluate strategies and see what's working. Take note of common mistakes and adjust your labeling guidelines based on the results.

Use a Proper Project Management Framework

A solid project management framework ensures that data labeling teams can execute ML projects on time. Agile methodologies like Scrum and Kanban are usually best suited for the iterative and experimental nature of data science and labeling projects. They allow teams to focus on their highest-priority tasks while enabling tasks to be reprioritized as needed. In the Agile approach, data scientists, data engineers, data labelers, and other stakeholders work together to build ML models quickly, test them in production, and refine them in rapid iterations.

Investigate which project management methodology best fits your data labeling project and data science team structure. You can even consider combining several methods or creating one that best fits your specific use case.

04

Technology

Technology provides the tools that the other three pillars use to implement their processes. As more organizations move from a model-centric to a data-centric AI approach, there is now an expansive and rapidly growing set of technology and tools available to support this movement. From synthetic data generation tools that help create AI-generated datasets to data pipeline tools that automate the movement and aggregation of data to data labeling tools that allow for fast and efficient data tagging and enrichment, the number of tools and frameworks accessible to data teams is expanding.

Your data labeling technology should seamlessly fit into and connect to the other technology products in your ML pipeline. For better labeling quality and ultimately more accurate ML models, it is also advisable to have a single platform for data labeling. This allows for standardized data processes and easy collaboration between data annotators and the relevant stakeholders.

Here are some important capabilities to consider when choosing a data labeling software.

Data Source Integration

The first thing you should look for in a data labeling tool is if it supports data imports for the present and future data sources you use. Luckily, most data labeling platforms, like our own Label Studio, support data imports from popular data sources. You can upload the dataset directly to the platform or provide the link to the dataset on a public URL.

Multi-data Type Support

Your data labeling tool should be able to support all your different data types (images, text, video, audio) in the different formats you use. For example, .bmp, .gif, .jpg, and .png for images and .txt, and .json for text. Investigate the different data labeling platforms and find one that supports your data type and format. Having one data labeling technology for all data types drastically simplifies your data labeling process because data labelers only need to learn one tool, and you have a centralized platform to manage and monitor the labeling process.

Computer Vision

- Image classification
- Object Detection
- Semantic Segmentation

Audio & Speech

- Classification
- Speaker Diarization
- Emotion Recognition
- Audio Transcription

Text, NLP and Documents

- Classification
- Named Entity Recognition
- Question Answering
- Sentiment Analysis

Time Series

- Classification
- Segmentation
- Event Recognition

Video

- Classification
- Object Tracking

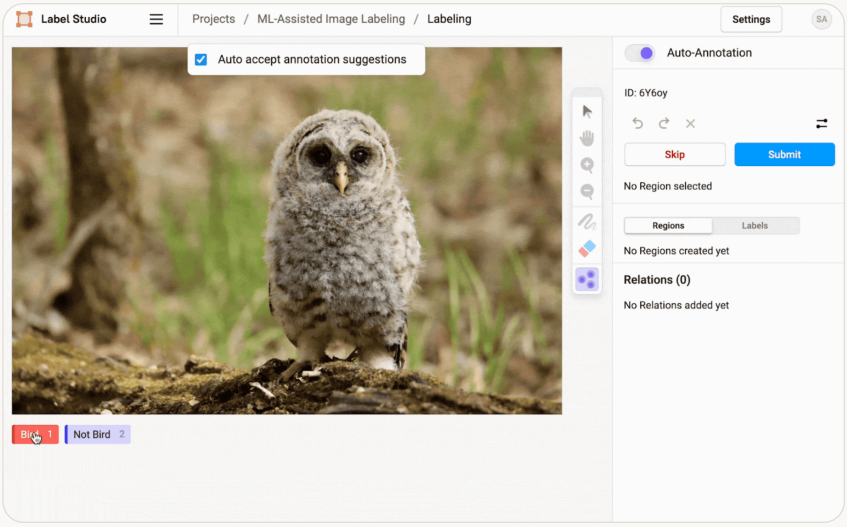
Multi-Domain

- Dialogue Processing
- Optical Character Recognition
- Time Series with Reference

Auto Labeling

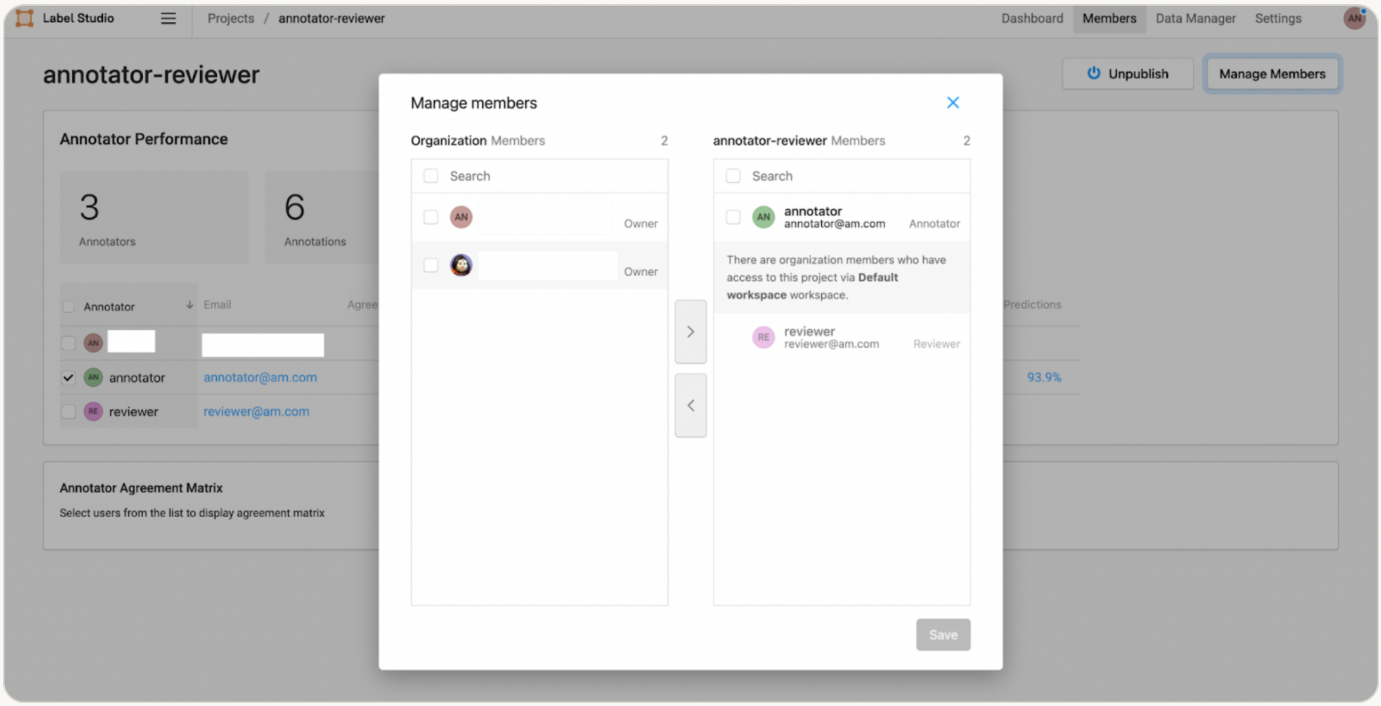
Having automated data labeling in your labeling tool can speed up the data labeling process and greatly reduce the workload on your team, especially when dealing with large datasets.

For example, for an image classification task, the tool can pre-select an image class for your data annotators to verify. For audio transcriptions, the model displays a transcription that data annotators can modify.



Team and Project Management

You will likely have multiple data labeling projects over time. You may also manage internal and external annotation teams. Your data labeling software should simplify the administration and management of multiple projects, onboarding internal and external team members, and assigning annotators to projects.



Security

Annotators will be given access to your systems and data, and there is usually a non-trivial amount of turnover on a data labeling team. Thus, security is paramount. Your data labeling software should be able to integrate with your Single Sign-on (SSO) solution, audit user activity, and support role-based access controls (RBAC).

Workflow Management

Your data labeling software should support workflow customization to orchestrate the movement of data labeling tasks. Having your data labeling software manage and enforce your workflow makes sure critical steps, especially quality management, are completed and even raises alerts if certain conditions aren't met.

Quality Management

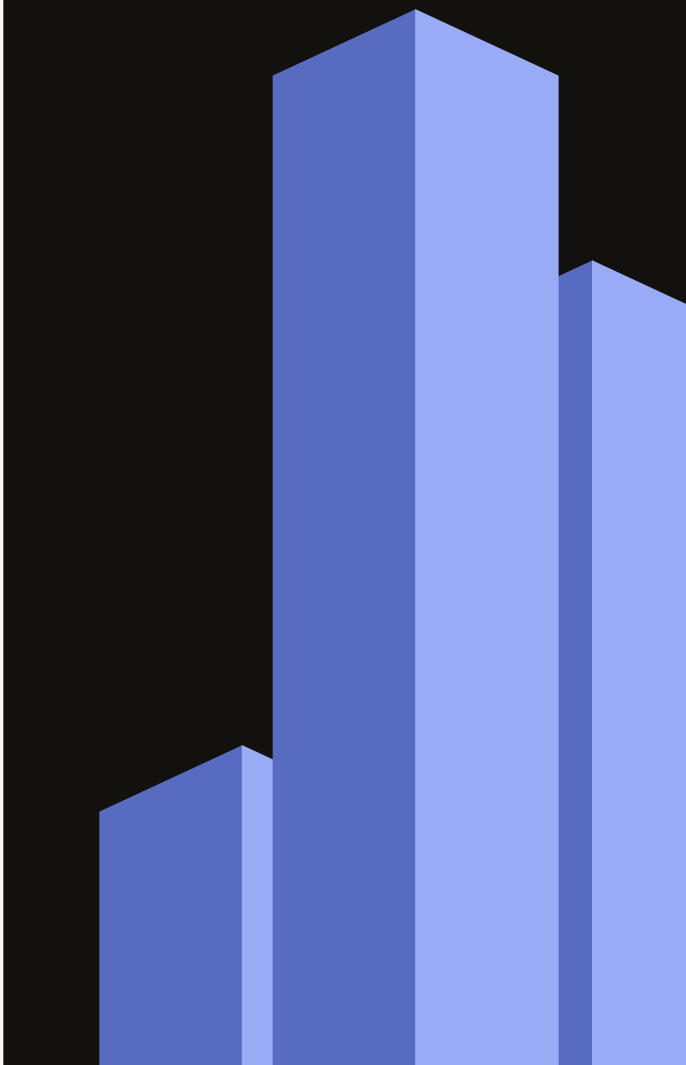
Quality data labels lead to more accurate ML models. Therefore, your data labeling software should provide mechanisms (metrics, reports, alerts, etc.) that make quality management easy and comprehensive.

Reporting and Analytics

Once your labeling projects are running at scale, having reporting and in-product analytics in your data labeling tool helps provide critical feedback on the health of your labeling project. Reports and analytics show how labeling tasks are progressing through the assembly line, identify bottlenecks, highlight areas of optimization, and keep quality management front and center for all stakeholders.

Data labeling software is integral to running a well-organized and efficient data labeling process. As with any software, you can build your own, use an open-source product, or purchase a commercial offering. While each option has pros and cons, we generally advise against building your own software. In the long run, it is hard to scale, tends to be more expensive, and lacks advanced capabilities and security features.

Your data labeling software should be able to integrate with your Single Sign-on (SSO) solution, audit user activity, and support role-based access controls (RBAC).



Conclusion

Improve Your ML Success Rate With Data-centric AI

According to a report by LXT, 70% of mid to large US firms spend $1 million or more of their budget on AI. But despite the millions invested, the number of failed AI projects has increased. A BCG study found that 70% of companies report little to no impact from investments in AI. According to Alation's State of Data Culture Report, 87% of employees cite data quality issues as the number one reason their organizations failed to successfully implement AI and machine learning. A data-centric AI approach helps to fix this by focusing on optimizing the data quality rather than model quality alone.

Companies need to invest resources into high-quality data to ensure the successful delivery of any machine learning project. A scalable and efficient data labeling process is a core of the data-centric AI process and helps organizations reap the benefits of their ML initiatives.

70%

of mid to large US firms spend $1 million or more of their budget on AI

70%

of companies report little to no impact from investments in AI

87%

of employees cite data quality issues as the number one reason their organizations failed to successfully implement AI and machine learning

HumanSignal is a software company that's pioneering the next generation of data management for machine learning, with a focus on data labeling and data preparation. Our web-based platform powers the work of data scientists and machine learning engineers, helping them unlock the value of organizational data, one customer at a time. HumanSignal HQ is based in San Francisco, CA.

Get A Free Trial

humansignal.com