Heartex

Data Labeling Overview for Machine Learning and Data Science

Introduction

Raw data on its own doesn't have enough built-in meaning to teach or train a machine learning model. Data labeling or annotation is the process that gives meaning to your raw data, adding a critical layer of metadata that draws the connection between raw data and the prediction your model is learning to make. The quality and accuracy of data labeling is directly correlated with the accuracy of predictions delivered by your models once in production.

At scale, data labeling can feel like a daunting, labor-intensive process. It is! And, you may be inclined to cut corners to speed-up the process. Don't! The metadata generated by labeling is extremely valuable intellectual property completely unique to your business. Generate good labels and your competitors won't know what hit them.

In this data labeling overview, we will outline the core aspects of data labeling, including data, process, people, and technology. After reading this overview, you should understand the key components of data labeling and how to organize these components together to build a successful, efficient, and repeatable data labeling system for your organization.

Data

Machine learning projects benefit from data originating from a variety of sources. Datasets may include structured and unstructured data and data with different data types, such as images, audio, text and documents, videos, and more. For example, a dataset that consists of audio recordings of customer support interactions has media files (audio). Transcribing the audio recordings results in text that can be added to your dataset. The transcribed text could be processed to extract the named entity, for example the problem text. The original audio recordings could be segmented based on emotion. All of the data and metadata could then be used to train a machine learning model to execute the same tasks automatically for every new audio recording, with prediction results used for analytics.

Storing the data and associated labels together in a single data store simplifies the management and understanding of your dataset and any edge cases contained within it. The dataset can be continuously updated with new items, new or updated labels, or repurposed for other models.

People

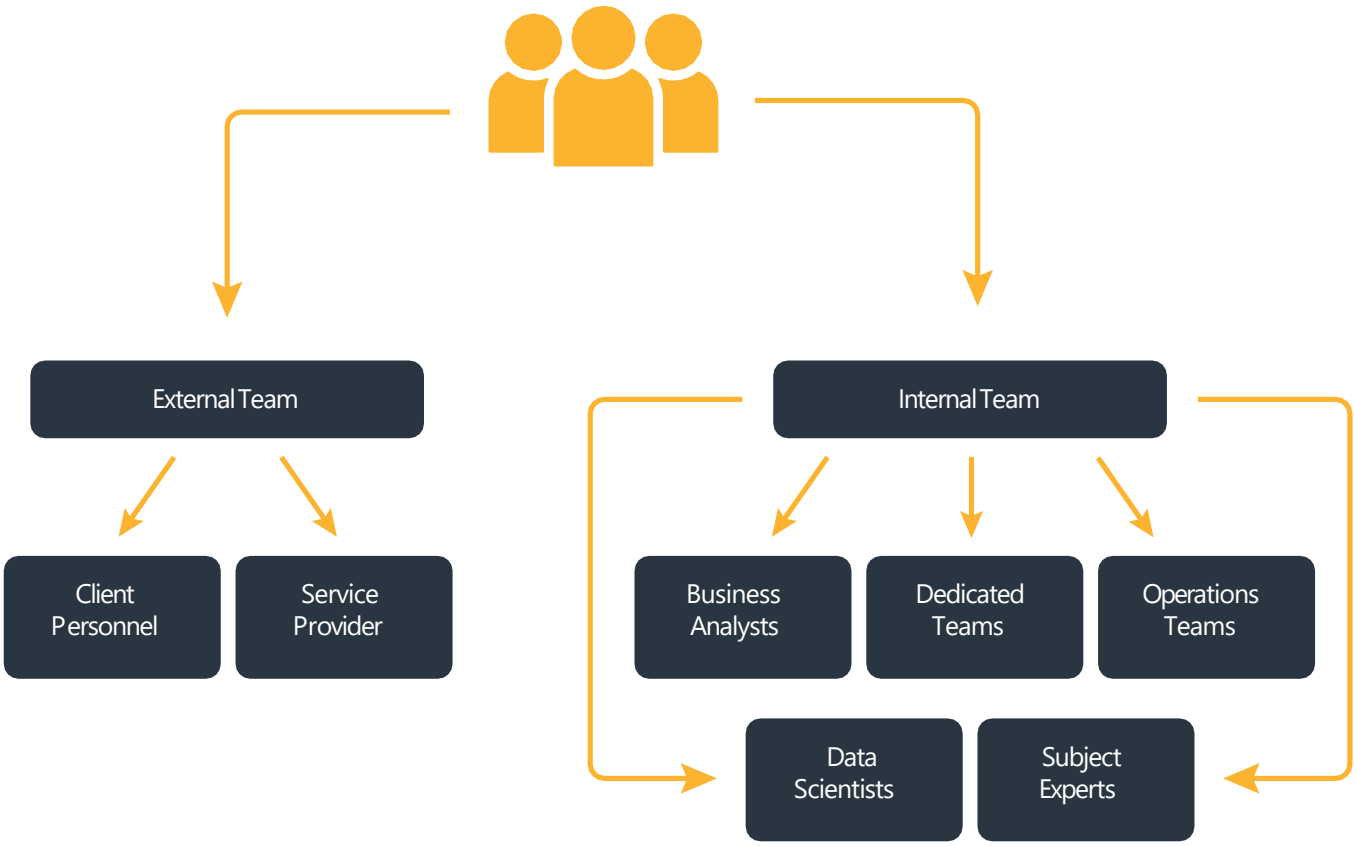
People, ignoring source data for the moment, are arguably the most critical aspect to any data labeling project. Process and technology are important too, but are there to support and optimize your data labeling team. Only people can understand the meaning of the task, evaluate the source data, give their decision, and generate the metadata that gives meaning to raw data. People must have the necessary domain knowledge, and in some cases domain expertise, to successfully and accurately label data.

To assemble your data labeling team, you can either hire new Annotators or utilize existing members of your organization. Alternatively, there are third-party services that you can leverage to outsource your labeling tasks. The volume of labeling required, necessary domain knowledge, and other internal factors will influence your decision to build an internal team or work with a service provider. It is also possible to have an internal team and a third party service working together, especially when the internal team needs temporary support to get through an urgent or large set of tasks.

! NOTE: A common strategy is to leverage a service provider when a project requires a significant amount of data labeling and to then use internal resources once the initial, high-volume, labeling tasks are complete.

Regardless of the makeup of your data labeling team, you must ensure the people have the domain knowledge to successfully complete their tasks.

For the labeling team to be successful, you will need to develop a repeatable process and leverate tools and technology to centrally manage workflows, assign tasks, and assess results for quality. With that in mind, Data Scientists or managers will be responsible for managing and monitoring the data labeling team and for assessing quality to ensure labeled data is accurate.



Process

Data labeling is, in and of itself, a process. Think of it as an assembly line that takes source data in as raw inputs and creates meaningful metadata, in a format that machine learning algorithms can understand and use to make predictions, as outputs. For a machine learning or data science project to be successful, you need to have a well-designed, efficient, and scalable process that can be actively monitored to ensure high-quality and accurate results.

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

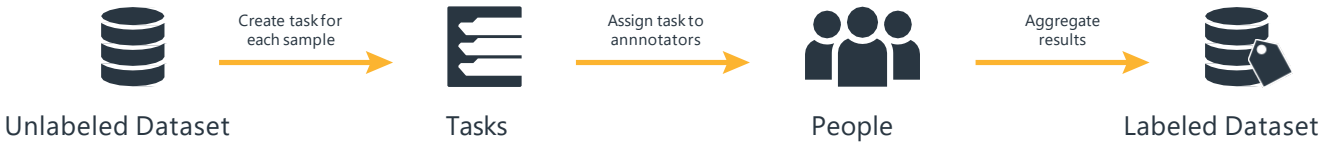
- Initial model training
- Model fine-tuning
- Human in the loop

Most aspects of the data labeling process are common across all three categories. However, there are important differences in each category that need to be understood and factored into your process. We'll discuss those shortly, but first, let's define the common attributes and components.

- Unlabeled dataset: This data is your raw input and consists of source data, that once labeled, will be used to train your machine learning model.
- Instructions: Clearly document instructions for your data labeling team. Most importantly, describe the data they will be reviewing and the decision(s) they are tasked to make. Additionally, provide instructions, including distinct steps, that annotators must follow and documentation for how to use any relevant tools.
- Labeling tasks: Individual samples from your unlabeled dataset are the labeling tasks.
- Annotators: Labeling tasks are assigned to people, or annotators, on your data labeling team.
- Labeled Dataset: The aggregated results of your labeling tasks make up your labeled dataset.

Labeling for Initial Model Training

The data labeling process for your initial model encompasses generating the unlabeled dataset, breaking it into labeling tasks that are assigned to Annotators who follow instructions to properly label each sample. After assessing quality and fixing any discrepancies, the process completes with your labeled dataset.



Labeling for Model Fine-tuning

There will be occasions when you need to label additional data to improve the accuracy of an existing model. A common way to initiate a new machine learning model, especially if you don't have a lot of your own data, is to start with an open source model that was trained with a large dataset. While this gives you a great starting point, you will still need to label samples from your own dataset and retrain the model to ensure your unique data is represented during training.

Once a model is in production, you may need to label additional data to correct a model that's accuracy has degraded or because there is now new data that, once labeled, will enhance the model's predictive accuracy. You may also want to label samples that your production model is least sure about when giving a prediction, this is known as active learning.

Regardless of the reason, you will need to design your process to include steps to label additional data points to keep improving model accuracy.

Human in the Loop

The third category, human in the loop, is used when the predictions of a production model need to be reviewed by an Annotator (the human in the loop) to ensure only highly accurate results are delivered to users. When the Annotator reviews the production predictions, they also add the newly labeled data to the labeled dataset to support future model training.

If the volume of predictions is too high to manually review each one, then you can use a prediction (certainty) score to automatically send high-confidence predictions to users and low-confidence predictions to your data labeling team.

Label Accuracy and Quality Management

The quality and accuracy of data labeling rests in the hands of the individuals on your data labeling team. As we all know, people are not perfect. We make mistakes, have biases, and may lack domain knowledge to always make the right decision when labeling. While understandable, incorrect labels have a devastating impact on the accuracy of your production models.

As Data Scientists, we are ultimately responsible for the accuracy and quality of the labeled dataset. Therefore, we must include quality checks and balances into the labeling process. This will take additional time and resources, however it is much more efficient to identify label quality issues before model training begins.

A very successful technique to improve labeling quality is called Collaborative Data Labeling. Collaborative Data Labeling is when multiple Annotators review and label the same unlabeled samples. While this may seem inefficient at first, it has proven to be an effective way to improve labeling quality and accuracy. It can also improve overall process throughput because you find that data labeled with 100% agreement by 3 Annotators, for example, proves to be consistently accurate and thus requires fewer downstream quality management tasks.

Whether you implement Collaborative Data Labeling or not, you will need to build quality management and monitoring into your data labeling process. Again, it is much better to find labeling quality issues before the labeled dataset moves on to the next step, typically model training, in your data science lifecycle.

Technology

There is an expansive, and rapidly growing, set of technology and tools available to support and manage the end-to-end lifecycle of data science and machine learning projects. The full range of technology is far too great for us to even summarize in this overview article. Instead, we provide a brief summary of the capabilities you should look for specific to data labeling software.

Data labeling software supports both Annotators and the actual act of labeling data as well as the management, orchestration, and quality assessment of the overall process. Important capabilities of data labeling software include:

Data Type Support

Most organizations label data with a variety of different data types. Leveraging a single solution that supports all of your data types drastically simplifies your data labeling process. Annotators only need to learn one tool and Data Scientists and managers have a centralized platform to manage, monitor, and optimize the labeling process.

Team & Project Management

It is likely that you will have multiple data labeling projects over time. You will also likely be adding, removing, and reassigning Annotators to different projects. As discussed earlier, you may be managing internal teams, external resources, or a hybrid of internal and external resources. Your data labeling software should simplify the administration and management of multiple projects, onboard internal and external team members, and assign 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 integrate with your Single Sign-on (SSO) solution, audit user activity, and support role-based access controls (RBAC).

Workflow Management

Labeling software, customized to your process, orchestrates the movement of tasks through the workflow. Having the 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

If you remember just one thing from this overview (hopefully, you remember more), remember the correlation between label quality and production model quality. Labeling software should also remind you of the importance of quality by providing mechanisms (metrics, reports, alerts, etc) that make quality management both easy and comprehensive.

Reporting and Analytics

Once your labeling projects are running at scale, reporting and in-product analytics provide important 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.

> As you can see, data labeling software is integral to running a well organized and efficient process. As with just about any software, you have the option of building your own, using an open source product, or purchasing a commercial offering. While each of these options have their pros and cons, we generally advise against building your own software. In the long run, the system is usually not built for scale, tends to be more expensive, and generally lacks the sophistication in capabilities and security.

Conclusion

Creating a well-functioning, scalable, and efficient data labeling process that yields accurate labels and ultimately, accurate production models is hard. Hopefully, this brief overview has provided you with enough details to start planning your new labeling process or improving your existing ones. The best place to start is documenting and mapping out your ideal, quality-focused process first. Then evaluate the different ways to staff your data labeling team and estimate the number of people needed to meet your timeline. Finally, you can adopt labeling software that best facilitates your process and supports your team. And, like most things in life, things will change over time. Try to instill flexibility in your process and people and adopt software that can evolve with you.

Heartex offers both open source and commercial editions of Label Studio, our data labeling software. Contact us if you'd like to learn more about Heartex Label Studio.