



[Supercharging large language model inference.](#)

# How to Organize Data Labeling for Machine Learning: Approaches and Tools

*The main challenge for a data science team is to decide who will be responsible for labeling, estimate how much time it will take, and what tools are better to use.*

By [AltexSoft](#).

[comments](#)

If there was a data science hall of fame, it would have a section dedicated to labeling. The labelers' monument could be Atlas holding that large rock symbolizing their arduous, detail-laden responsibilities. ImageNet — an image database — would deserve its own stele. For nine years, its contributors manually annotated more than 14 million images. Just thinking about it makes you tired.

While labeling is not launching a rocket into space, it's still seriously business. Labeling is an indispensable stage of data preprocessing in [supervised learning](#). Historical data with predefined target attributes (values) is used for this model training style. An algorithm can only find target attributes if a human mapped them.

Labelers must be extremely attentive because each mistake or inaccuracy negatively affects a dataset's quality and the overall performance of a predictive model.

How to get a high-quality labeled dataset without getting grey hair? The main challenge is to decide who will be responsible for labeling, estimate how much time it will take, and what tools are better to use.

We briefly described labeling in the article about the [general structure of a machine learning project](#). Here we will talk more about labeling approaches, techniques, and tools.

Search KDnuggets...



Develop skills  
for today's  
data-driven world

Earn your Northwestern master's  
degree online.

Northwestern

DATA SCIENCE  
School of Professional Studies

APPLY NOW

[Learn from leading data science experts.](#)

## Latest Posts

Free Courses That Are Actually Free:  
Computer Science Edition

Has Europe Gone Too Far? The Delicate  
Dance of Regulation and Innovation

Get an Additional 30% off Courses –  
Offer Ends in 3 Days!

5 LLM Tools I Can't Live Without

How To Improve the Performance of a  
RAG Model

7 Steps to Mastering Coding for Data  
Science

## Top Posts

## Labeling approaches

The choice of an approach depends on the complexity of a problem and training data, the size of a data science team, and the financial and time resources a company can allocate to implement a project.

7 Free Online Python REPLs

7 Steps to Mastering Coding for Data Science

How Natural Language Processing of Unstructured Data is Improving Healthcare Outcomes

Partial Functions in Python: A Guide for Developers

5 LLM Tools I Can't Live Without

10 GitHub Repositories for Deep Learning Enthusiasts

Ollama Tutorial: Running LLMs Locally Made Super Simple

Free Courses That Are Actually Free: Programming Edition

Fundamentals of Effective Prompt Engineering

How to Write Basic SQL Queries in BigQuery



Get the FREE ebook 'The Great Big Natural Language Processing Primer' and 'The Complete Collection of Data Science Cheat Sheets' along with the leading newsletter on Data Science, Machine Learning, AI & Analytics straight to your inbox.

Your Email

**SIGN UP**

By subscribing you accept KDnuggets Privacy Policy

# PROS AND CONS OF LABELING APPROACHES

Approach	Description	Pros	Cons
Internal labeling	Assignment of tasks to an in-house team	<ul style="list-style-type: none"> <li>✓ Predictable results</li> <li>✓ High accuracy of labeled data</li> </ul>	<ul style="list-style-type: none"> <li>✗ It takes much time</li> </ul>



Machine Learning  
MLOps  
NLP



JOIN NEWSLETTER

Outsourcing	Recruitment of temporary employees on freelance platforms, posting vacancies on social media and job search sites	<ul style="list-style-type: none"> <li>✓ The ability to evaluate applicants' skills</li> </ul>	<ul style="list-style-type: none"> <li>✗ The need to organize workflow</li> </ul>
Crowdsourcing	Cooperation with freelancers from crowdsourcing platforms	<ul style="list-style-type: none"> <li>✓ Cost savings</li> <li>✓ Fast results</li> </ul>	<ul style="list-style-type: none"> <li>✗ Quality of work can suffer</li> </ul>
Specialized outsourcing companies	Hiring an external team for a specific project	<ul style="list-style-type: none"> <li>✓ Assured quality</li> </ul>	<ul style="list-style-type: none"> <li>✗ Higher price compared to crowdsourcing</li> </ul>
Synthetic labeling	Generating data with the same attributes of real data	<ul style="list-style-type: none"> <li>✓ Fewer constraints for using sensitive and regulated data</li> <li>✓ Training data without mismatches and gaps</li> <li>✓ Cost- and time-effectiveness</li> </ul>	<ul style="list-style-type: none"> <li>✗ High computational power required</li> </ul>
Data programming	Using scripts that programmatically label data to avoid manual work	<ul style="list-style-type: none"> <li>✓ Automation</li> <li>✓ Fast results</li> </ul>	<ul style="list-style-type: none"> <li>✗ Lower quality dataset</li> </ul>



## In-house labeling

That old saying *if you want it done right, do it yourself* expresses one of the key reasons to choose an internal approach to labeling. That's why when you need to ensure the highest possible labeling accuracy and have an ability to track the process, assign this task to your team. While in-house labeling is much slower than approaches described below, it's the way to go if your company has enough human, time, and financial resources.

Let's assume your team needs to conduct a sentiment analysis. Sentiment analysis of a company's reviews on social media and tech site discussion sections allows businesses to evaluate their reputation and expertise compared with competitors. It also gives the opportunity to research industry trends to define development strategy.

You will need to collect and label at least 90,000 reviews to build a model that performs adequately. Assuming that labeling a single comment may take a worker 30 seconds, he or she will need to spend 750 hours or almost 94 work shifts averaging 8 hours each to

complete the task. And that's another way of saying three months. Considering that the median hourly rate for a data scientist in the US is \$36.27, labeling will cost you \$27,202.5.

You can streamline data labeling by **automating** it with semi-supervised learning. This training style entails using both labeled and unlabeled data. A part of a dataset (e.g. 2000 reviews) can be labeled to train a classification model. Then this multiclass model is trained on the rest of the unlabeled data to find target values — positive, negative, and neutral sentiments.

The implementation of projects for various industries, for instance, finance, space, healthcare, or energy, generally require expert assessment of data. Teams consult with domain experts regarding principles of labeling. In some cases, experts label datasets by themselves.

Altexsoft has built the DolGrind app aimed at diagnosing and monitoring bruxism for Dutch startup Sleep.ai. Bruxism is excessive teeth grinding or a jaw clenching while awake or asleep. The app is based on a noise classification algorithm, which was trained with a dataset consisting of more than 6,000 audio samples. To define recordings related to teeth grinding sounds, a client listened to samples and mapped them with attributes. The recognition of these specific sounds is necessary for attribute extraction.

### **Advantages**

**Predictable good results and control over the process.** If you rely on your people, you're not buying a pig in a poke. Data scientists or other internal experts are interested in doing an excellent job because they are the ones who'll be working with a labeled dataset. You can also check how your team is doing to make sure it follows a project's timeline.

### **Disadvantages**

**It's a slow process.** The better the quality of the labeling, the more time it takes. Your data science team will need additional time to label data right, and time is usually a limited resource.

### **Crowdsourcing**

Why spend additional time recruiting people if you can get right down to business with a crowdsourcing platform?

Amazon Mechanical Turk (MTurk) is one of the leading platforms that offer an on-demand workforce. Clients register there as requesters, create and manage their projects with one or more HITs (Human Intelligence Tasks) on the Mechanical Turk Requester website. The website provides users with an easy-to-use interface for creating labeling tasks. MTurk representatives claim that with its wide community of workers, labeling thousands of images can take few hours instead of days or weeks.

Another global online marketplace, Clickworker, has more than 1 million contractors ready

to be assigned to image or video labeling and sentiment analysis tasks. The first stages of workflow are similar to the ones on MTurk. Task processing and allocation phases differ. Registered employers place their orders with predefined specifications and demands, the platform team drafts a solution and posts a required set of work on the order platform for freelancers, and the magic begins.

### **Advantages**

**Fast results.** Crowdsourcing is a reasonable option for projects with tight deadlines and large, basic datasets that require using powerful labeling tools. Tasks like categorization of images of cars for computer vision projects, for instance, won't be time-consuming and can be accomplished by a staff with ordinary — not arcane — knowledge. Speed can also be achieved with decomposition of projects into microtasks, so freelancers can do them simultaneously. That's how Clickworker organizes workflow. MTurk clients should break down projects into steps themselves.

**Affordability.** Assigning labeling tasks on these platforms won't cost you a fortune. Amazon Mechanical Turk, for instance, allows for setting up a reward for each task, which gives employers freedom of choice. For example, with a \$0.05 reward for each HIT and one submission for each item, you can get 2,000 images labeled for \$100. Considering a 20 percent fee for HITs consisting of up to nine assignments, the final sum would be \$120 for a small dataset.

### **Disadvantages**

Inviting others to label your data may save time and money, but crowdsourcing has its pitfalls, the risk of getting a low-quality dataset being the main one.

**Inconsistent quality of labeled data.** People whose daily income depends on the number of completed tasks may fail to follow task recommendations trying to get as much work done as possible. Sometimes mistakes in annotations can happen due to a language barrier or a work division.

Crowdsourcing platforms use quality management measures to cope with this problem and guarantee their workers will provide the best possible services. Online marketplaces do so through skill verification with tests and training, monitoring of reputation scores, providing statistics, peer reviews, audits, as well as discussing outcome requirements beforehand. Clients can also request multiple workers to complete a specific task and approve it before releasing a payment.

As an employer, you must make sure everything is good from your side. Platform representatives advise providing clear and simple task instructions, using short questions and bullet points, and giving examples of well and poorly done tasks. If your labeling task entails drawing bounding boxes, you can illustrate each of the rules you set.

*Clear illustration of image labeling dos and don'ts*

You must specify format requirements and let freelancers know if you want them to use specific labeling tools or methods. Asking workers to pass a qualification test is another strategy to increase annotation accuracy.

### **Outsourcing to individuals**

One of the ways to speed up labeling is to hunt for freelancers on numerous recruitment, freelance, and social networking websites.

Freelancers with different academic backgrounds are registered on the [UpWork](#) platform. You can advertise a position or look for professionals using such filters as skill, location, hourly rate, job success, total revenue, level of English, and others.

When it comes to posting job ads on social media, LinkedIn, with its 500 million users, is the first site that comes to mind. Job ads can be posted on a company's page or advertised in the relevant groups. Shares, likes, or comments will ensure that more interested users see your vacancy.

Posts on Facebook, Instagram, and Twitter accounts may also help find a pool of specialists faster.

### **Advantages**

**You know who you hire.** You can check applicants' skills with tests to make sure they will do the job right. Given that outsourcing entails hiring a small or midsize team, you'll have an opportunity to control their work.

### **Disadvantages**

**You have to build a workflow.** You need to create a task template and ensure it's intuitive. If you have image data, for instance, you can use [Supervising-UI](#), which provides a web

interface for labeling tasks. This service allows the creation of tasks when multiple labels are required. Developers recommend using Supervising-UI within a local network to ensure the security of data.

If you don't want to create your own task interface, provide outsource specialists with a labeling tool you prefer. We'll tell more about that in the tool section.

You are also responsible for writing detailed and clear instructions to make it easy for outsourced workers to understand them and make annotations correctly. Besides that, you'll need extra time to submit and check completed tasks.

### **Outsourcing to companies**

Instead of hiring temporary employees or relying on a crowd, you can contact outsourcing companies specializing in training data preparation. These organizations position themselves as alternative to crowdsourcing platforms. Companies emphasize that their professional staff will deliver high-quality training data. That way a client's teams can concentrate on more advanced tasks. So, partnership with outsourcing companies feels like having an external team for a period of time.

Outsourcing companies, such as CloudFactory, Mighty AI, LOA, and DataPure, mostly label datasets for training computer vision models. CrowdFlower and CapeStart also conduct sentiment analysis. The former allows for analyzing not only text but also image and video files. In addition, CrowdFlower clients have an option to request a more complex method of sentiment analysis. Users can ask leading questions to find out why people reacted to a product or service in a certain way.

Companies offer various service packages or plans, but most of them don't give pricing information without a request. A plan price usually depends on a number of services or working hours, task complexity, or a dataset's size.

*CloudFactory allows for calculating service price according to the number of working hours*

### **Advantages**

**High-quality results.** Companies claim their clients will get labeled data without inaccuracies.

### **Disadvantages**



**It's more expensive than crowdsourcing.** Although most companies don't specify the cost of works, the example of CloudFactory's pricing helps us understand that their services come at a slightly higher price than using crowdsourcing platforms. For instance, labeling

90,000 reviews (if the price for each task is \$0.05) on a crowdsourcing platform will cost you \$4500. To hire a professional team of 7 to 17 people not including a team lead, may cost \$5,165-5200.

Find out whether a company staff does specific labeling tasks. If your project requires having domain experts on board, make sure the company recruits people who will define labeling principles and fix mistakes on the go.

---

## Our Top 3 Course Recommendations

-  1. [Google Cybersecurity Certificate](#) - Get on the fast track to a career in cybersecurity.
  2. [Google Data Analytics Professional Certificate](#) - Up your data analytics game
  -  3. [Google IT Support Professional Certificate](#) - Support your organization in IT
- 

## More On This Topic

- [How I Did Automatic Image Labeling Using Grounding DINO](#)
- [Organize, Search, and Back Up Files with Python's Pathlib](#)
- [Machine Learning's Sweet Spot: Pure Approaches in NLP and Document Analysis](#)
- [Automated Machine Learning with Python: A Comparison of Different...](#)
- [Data Analytics: The Four Approaches to Analyzing Data and How To...](#)
- [Multi-label NLP: An Analysis of Class Imbalance and Loss Function...](#)

Get the FREE ebook 'The Great Big Natural Language Processing Primer' and 'The Complete Collection of Data Science Cheat Sheets' along with the leading newsletter on Data Science, Machine Learning, AI & Analytics straight to your inbox.

Your Email

**SIGN UP**

By subscribing you accept KDnuggets Privacy Policy



Information from your device can be used to personalize your ad experience.

[Do not sell or share my personal information.](#)