DATA LABELING

# Data labeling in 2020: The Ultimate Guide for executives and labelers

JUNE 15, 2020 · 9 MINUTE READ

Since 2010s, companies have been heavily investing in machine learning. Supervised learning is the most common form of machine learning today. Supervised learning algorithms need to be fed with labeled instances. This increases the importance of data labeling solutions.

Therefore, data labeling tools (open source vs proprietary), service providers and alternatives to data labeling are important aspects of a company's data labeling strategy:



## What is data labeling?

Supervised machine learning algorithms learn from labeled data, data that has been tagged with labels. Programmers do not explicitly program machine learning algorithms on how to make decisions, they program the models that learn from labeled data.

Data labeling, also called data annotation/tagging/classification, is the process of preparing tagged (i.e. labeled) data sets for machine learning. Machine learning models learn to recognize repetitive patterns in labeled data. After a sufficient amount of labeled data is processed, machine learning models can identify the same patterns in data that have not been labeled.

# What are its applications?

Machine learning models have 3 processes during their life cycle:

- **Initial model training:** Enables the model to infer outputs (e.g. there is a cat in the image) from inputs (e.g. image file)
- **Inference:** Model makes inferences with accompanying confidence levels. Not all models provide confidence levels and confidence levels provided by models do not always correspond to the actual probability of success of the inference.
- **Continuous learning:** The output of machine learning models is inferences. In cases where inference confidence is low, it is likely that the model output is wrong. If the model prediction is to be used in a business decision, businesses would prefer to have a human review model output and correct it as necessary. The corrected output can be fed back into the machine learning model to constantly improve model performance. This process of improving model performance with humans is an example of a human-in-the-loop system.

Data labeling is required both in initial model training and in continuous improvement.

# Why is it important now?

Businesses are adopting AI technology to automate decision-making and benefit from new business opportunities, but it not as easy as it seems. In fact, a McKinsey article listed data annotation as the most challenging limitations to AI adoption in the industry.

Data labeling enables machines to gain an accurate understanding of real-world conditions and opens up opportunities for a wide variety of businesses and industries. Grand View Research estimates data labeling services industry revenues to be USD 316 million in 2018. They expect the industry to grow to USD 1.6 billion by 2025 and register a 26.6% CAGR over the course of the forecast period.

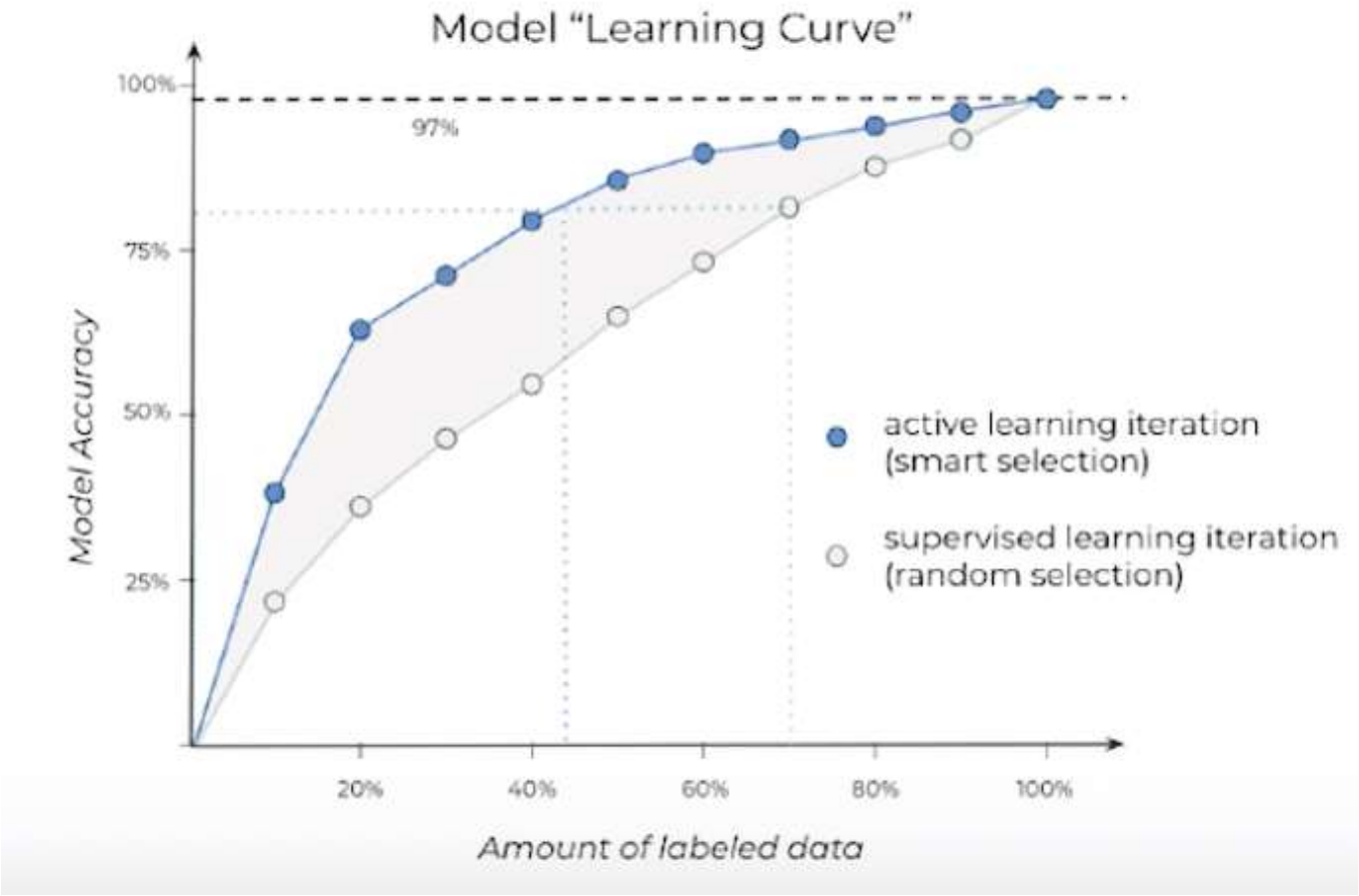
According to McKinsey, AI has the potential to deliver additional global economic activity of around $13 trillion by 2030. Data labeling is critical for achieving that potential. For example, having better-labeled data than competitors provides superiority in the machine learning industry.

# How can we increase data labeling efficiency?

The simplest labeling approach, labels all data at hand, creating ground truth for the machine learning algorithm. However, it is also possible to focus on parts of unlabeled data that will create the most learning when they are labeled. Active learning aims to achieve that reducing labeling time and cost.

## Active Learning

Active learning is a semi-supervised approach of data annotation with the aim of more reliable labeling using as few labeled instances as possible. In active learning, annotator selects an initial sample from unlabeled data. Depending on results in each step, the annotator incrementally selects and labels more data to the system during the process until every data point is labeled.




Source: FigureEight

Active learning approaches include:

**Membership Query Synthesis**: The active learner (ML model that uses active learning) generates a synthetic instance and requests a label for it. Label may not be possible to be produced by a human worker in all cases. For example, in the case of an ML model that is trying to solve a scientific problem (e.g. the medium that makes a certain yeast grow in the optimal rate) the instance can be a scientific experiment and label can be the result of the experiment.

**Stream-Based Selective Sampling**: This method assumes that getting an unlabeled instance is free. Then the algorithm selects an instance one-by-one and decides to label or ignore it based on its informativeness and region of uncertainty.

**Pool-Based sampling**: This method assumes that there is a large but not infinite pool of unlabeled instances. Then, it ranks all instances according to informativeness measurement and selects the best queries to label.

# What are alternatives of data labeling?

Data labeling is required for supervised learning. Semi-supervised or unsupervised learning approaches do not require data labeling. However, unsupervised or semi-supervised learning approaches are currently not the best performing approaches for most machine learning applications.

## Generative adversarial networks – GANs (Semi-supervised learning example)

GANs are a semi-supervised learning method. They separate the model into two sub-models: a generator network and a discriminator network. The generator network creates new examples of data. Discriminator network classifies those examples as real or fake. GANs method enables supervised learning with fewer human interaction.

## Reinforcement learning (Unsupervised learning example)

In cases where programmers can build a formula for evaluating different inputs, manual labeling is not necessary.

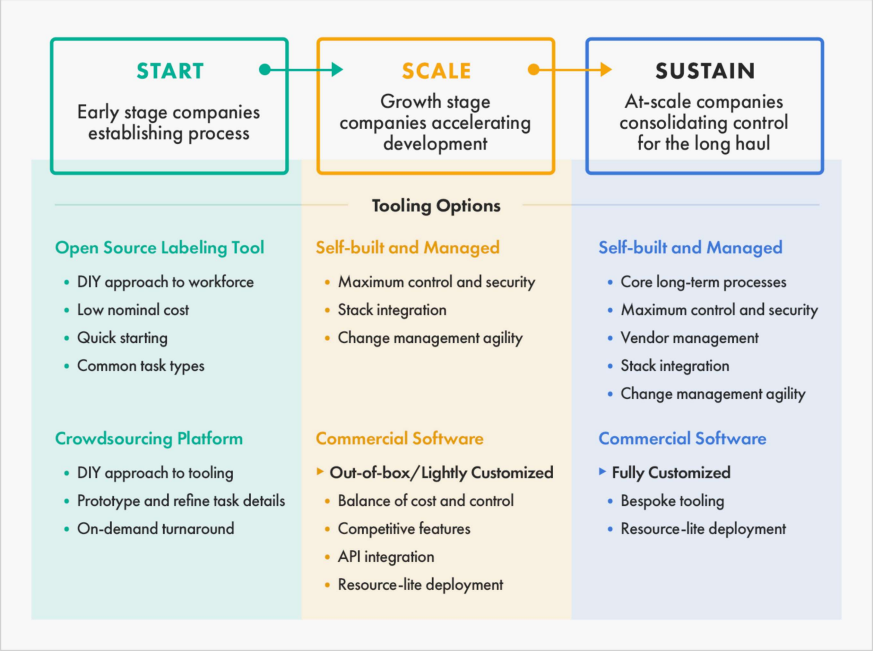
Let's assume that we need to build a chess playing system. We need a way to evaluate different board positions so the system can decide whether its next move will improve its probability of winning. We can not have humans estimate a players' probability of winning a chess game from

the location of chess pieces on the board. This is because there are too many combinations of chess pieces on a chessboard. However, developers can formulate a function for assessing a players' probability of winning a chess game from the location of chess pieces on the board. Using such a function, a chess playing system can play millions of games against itself and learn how to play chess.

## What are the things to pay attention to while choosing a data labeling software?

Your choices about data labeling tools will be an essential factor in the success of the machine learning model:

- You need to determine the type of data labeling the organization needs. There are labeling tools for text, image, video and audio. Data security, storage, quality and supported file types differ between tools specialized for different file types.
- Labeling accuracy is an important aspect of data labeling. High-quality data generates better model performance. When it is low-quality, the machine learning model struggles to learn. The main reason for low-quality data labeling is the capability of people and processes; however, technology can also play a role. For example, some data labeling tools pre-process unstructured data with their machine learning models and partially label the data with high-confidence output from their models. These make it easier for personnel to label data and increase labeling accuracy.
- Necessary features: The more complex requirements you have, the more effort will be spent to customize a labeling tool for your companies' needs. Simplifying your requirements will make it easier to adapt a data labeling tool with minimal customization.



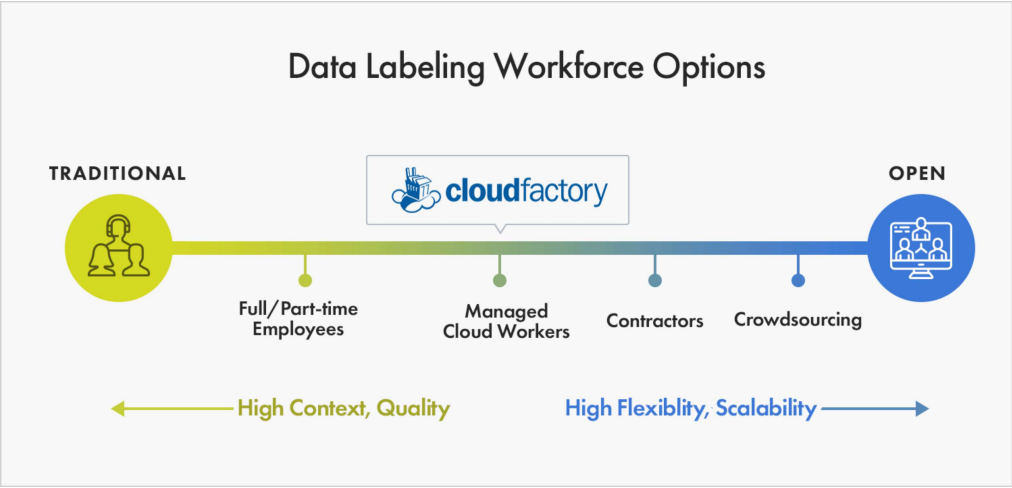
Source: Cloudfactory

# What are different types of data labeling software?

Data labeling tool may be divided into different types of categories.

- Some tools are designed to perform in a single segment such as text, image, audio or video while few of them can perform in multiple domains simultaneously.
- Annotation tools can also differ from each other according to their approaching method. Some tools such as MedTagger and imagetagger, are suitable for collaborative data labeling. Organizations that prefer to use crowdsourcing can pick such a tool for labeling.
- There are also different types of labeling features that differ in each tool. Bounding boxes, polygonial annotation and semantic segmentation are the most common features in the labeling market.
- Finally, open-source labeling tools are also available in the annotation market. Businesses may prefer this type of labeling tool depending on their organization's needs.

# Who should run your data labeling program?




Source: Cloudfactory

There are four different types of data labeling workforce options, each with different pros and cons.

## Full/Part-time Employees (In-house)

The enterprise assigns tasks to an in-house labeling team.

**Pros:** This is a great solution to provide quality output and track progress.

**Cons:**

- Since recruiting is slow, this is slow to scale.
- If you are operating in a developed country, this is the most expensive approach. Most outsourcing providers would staff their teams with employees in developing companies. Therefore outsourcing is likely to be a cheaper alternative.

## Outsourced employees (Managed cloud workers )

The organization combines vetted, trained, and actively managed, remotely located data labelers with an in-house team so that you can assure the quality of labeled data.

**Pros:** High quality labels if your organization can manage the labeling process successfully.

**Cons:** This approach is considerably higher priced than crowdsourcing and contractors.

Tasks can change as development teams train and tune their models, so labeling teams must be able to adapt and make changes in the workflow quickly. Therefore it is important to build a feedback loop between AI project teams and data labelers.

## Data labeling companies (Contractors)

You can hire data labeling companies that will charge you based on their output volume and manage all aspects of data labeling for you.

**Pros:**

- This is one of the most cost-effective solutions
- It saves your team from managing the data labeling effort.

**Cons:** It is important to ensure that your contractor will not later sell the labeled data to a competitor. NDAs can be used to enforce this.

However, if you are solving a rare problem and building a machine learning solution is the focus of your business, you should probably not rely on contractors for data labeling. In most cases, this should not be a concern because very few companies have rare problems. For example, if you need to have a model for identifying license plate numbers, this is not a rare problem. Many

companies have this problem and they have years of data on it. No single company would have an advantage in building a solution for identifying license plate numbers, this is an area where contracts can label the data or build machine learning models.

However, if you have access to data on how search engine visitors click links, you should probably not share it. Sorting search results is a rare problem that is the focus of only a few companies. Only Google, Bing and Baidu have sizable data on this problem so these companies would be better off building their own solutions.

### Crowdsourcing

In this approach, the organization uses a crowdsourcing platform to send data tasks to a large number of data labelers at once.

**Pros:**

- Lowest costs
- Fast results

**Cons:**

- May result in lower quality
- No confidentiality

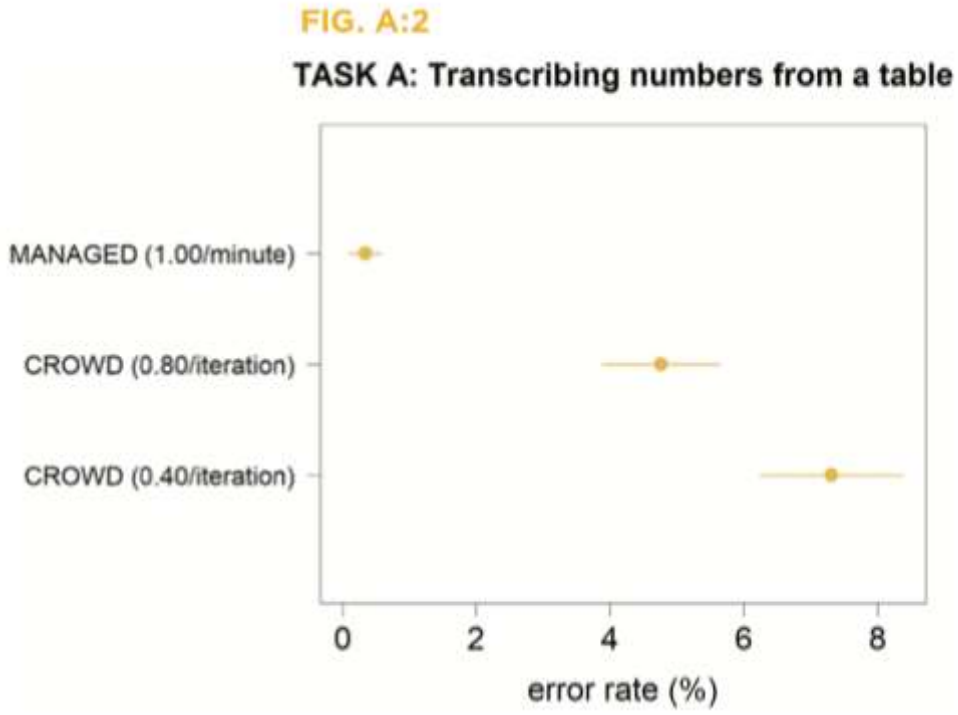
## What is the quality trade-off in outsourced vs crowdsourced data labeling?

If cost is a bigger constraint than data quality, companies can choose crowdsourcing. However, crowdsourced solutions are less accurate than outsourced workers that are managed by your team. Cloudfactory ran a study comparing the quality of work of managed vs crowdsourced workers. In this study, three tasks are examined:

1. Basic transcription,
2. Assess sentiment from text,
3. Categorize an event from unstructured text.

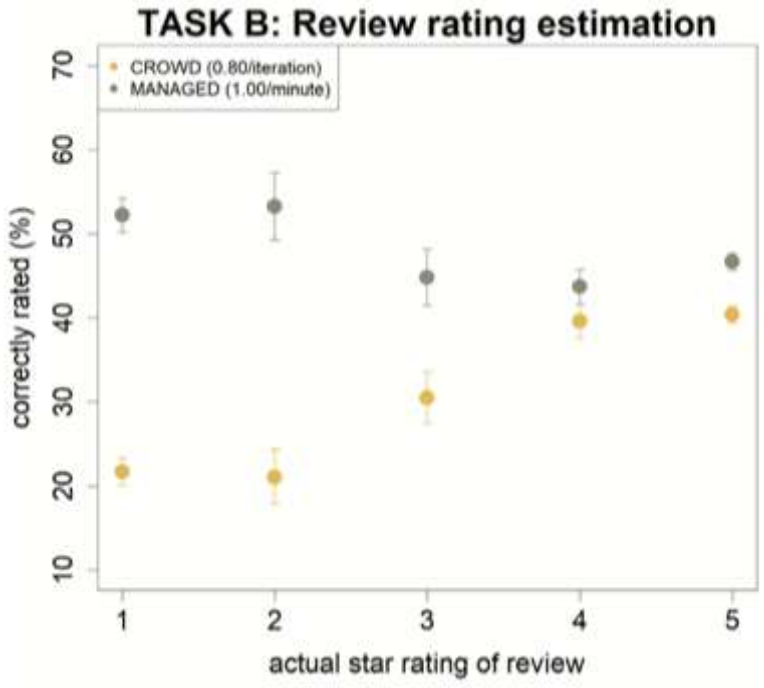
In all tasks, the accuracy levels of managed workers have been higher crowdsourced workers.

- In basic transcription task, managed workers' error rate is ~1% which is significantly better than crowdsourced workers with 4-8% error rate:



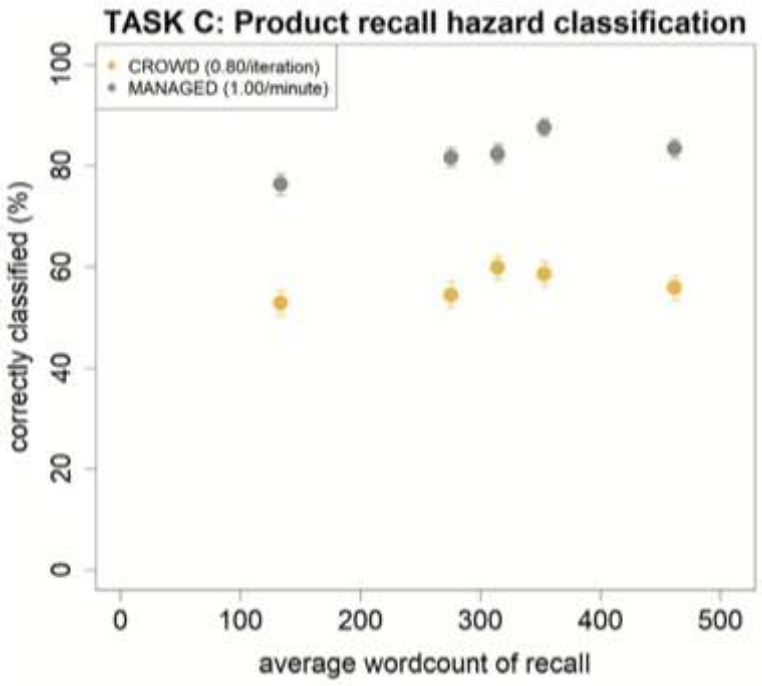

Source: Cloudfactory

- In sentiment analysis task, the average accuracy of managed workers and crowdsourced workers are 50% and 40%, respectively.




Source: Cloudfactory

- In categorizing an event from unstructured text task, managed workers labeled with 25% higher accuracy than crowdsourced workers, accuracy rate of 80% and 60%, respectively.

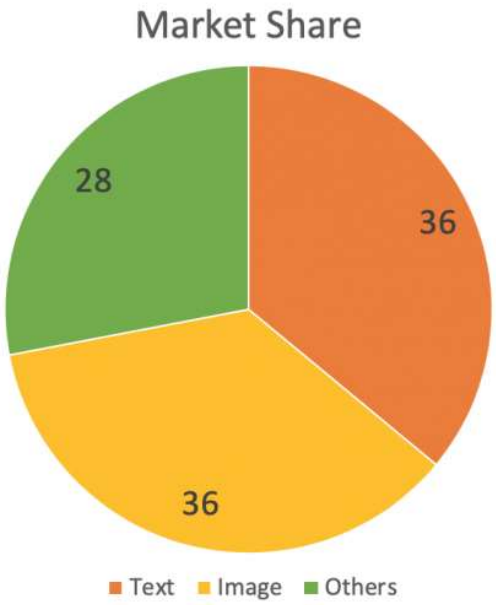



Source: Cloudfactory

# What are the different types of data labeling companies?

The data labeling services market is segmented into three categories.

- Text Labeling: The text segment has the largest market share. Use cases include sentiment tagging where humans tag text with the sentiment (e.g. anger, happiness, etc.) expressed in the text.
- Image Labeling: These tools make images readable for machines. There are different types of image labeling techniques such as bounding boxes, polygonal segmentation, line annotation, landmark annotation, 3D cuboids, semantic segmentation, etc.
- Others: Others segment involves audio and video labeling.




If you want to learn more about companies in the data annotation market, please check our comprehensive, sortable list of annotation vendors.

If you still have questions, contact us:

Let us find the right vendor for your business

**Sources:**

List of different types of data labeling software: Github