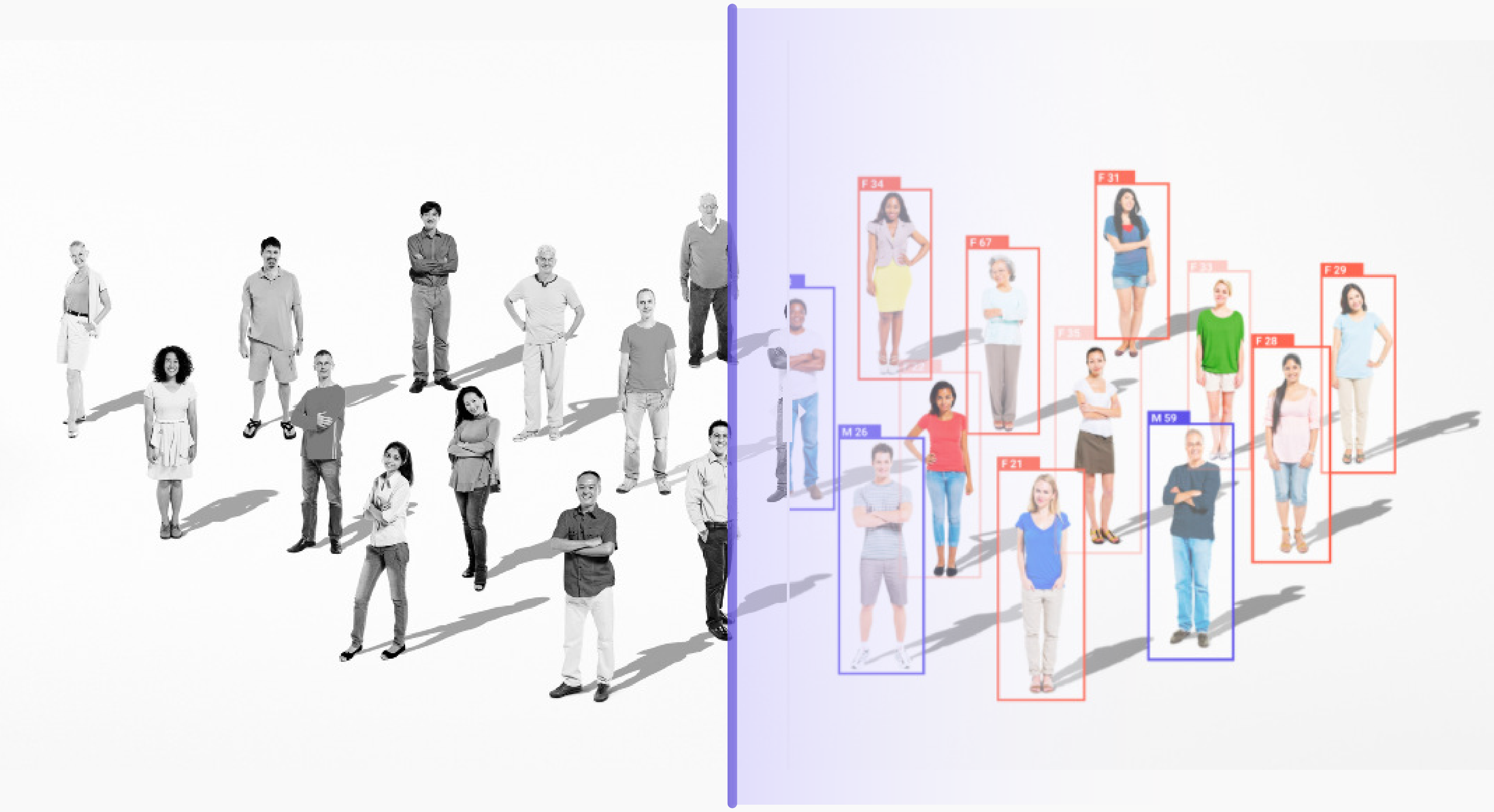
DeepLobe



Your Definitive Guide to Data Labeling

for Machine Learning and Deep Learning Projects

Table of Contents

Data has become a foundational requirement for any project in recent years. There is no shortage of data sources that generate large volumes of structured and unstructured data. Of late, unstructured data is proving to be an equal contributor to drawing actionable insights by leveraging machine learning and deep learning technologies when this data is used for AI modeling projects. However, to do so, enterprises need tools and human resources to label that data to train, validate, and build quality models with high accuracy.

We understand how important data labeling is for any machine learning and deep learning project. Our eBook acts as a comprehensive handbook to data labeling requirements as we dive deeper into the essential elements of this vital but time-consuming task along with the best practices to label the data, and what to look for when choosing the right data labeling platform.

Let's start with the basics!

What is data labeling?

The AI life cycle begins with collecting the data and organizing it. Data labeling - also synonymous with data annotation/tagging/classification in many scenarios - is the process of identifying raw data and giving meaningful and informative labels to this data to provide some context for the machine learning models to learn from it. Curating the data manually by adding keywords to the unstructured data enables the machines to automatically recognize the concepts that these keywords describe.

COLLECT

Data Sourcing

Data Ingestion

ANALYZE

Data Profiling

Data Labelling

ORGANIZE

AI Model Training & Building

Deploy & Iterate

> A recent study conducted by a group of authors from Stanford University, Salesforce AI Research, University of California, and Amsterdam University Medical Centers reveals that **“key to most AI tasks is the availability of a sufficiently large, labeled data set with which to train AI models.”** They also highlighted that, “However, it is challenging to obtain large-scale high-quality annotations for AI models.“ [1]

Data labelling process model

Data Preparation

- Define the problem statement concerning objects, features, scenarios or the amount of data to be labelled
- Data collection and selection for the desired features and scenarios

1

Setting up Labelling Tasks

- Create simple labelling instructions and guidelines for annotators
- Choose the right annotation tool that can easily integrate with the data pipeline and is easy to use
- Configure the tool to visualize raw data
- Hire talent and train them

2

Labelling Process

- Make annotations
- Perform manual and automated checks. If mistakes are detected then fix them by remaking the annotations
- Check for quality

3

Data Transformation

- Model ingestion
- Training the model
- Testing the model. Dependinig on the result obtained, if more data is needed to train, the process starts back from data collection

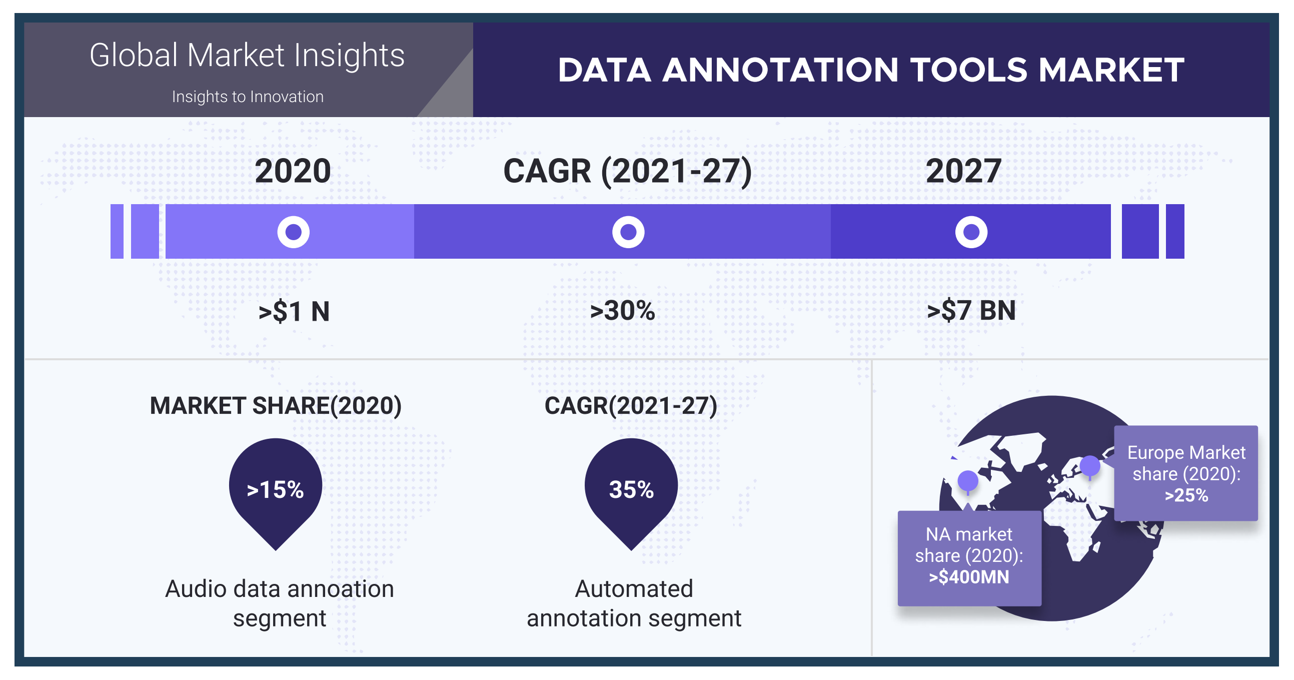
4

Labeling the data - Need of the hour

Imagine a self-driving car and the level of accuracy that the AI model should consist of as there won't be any room for second-guessing. The model's accuracy improves drastically if it is trained on the data that has been annotated with parameters like colors, shapes, signs, angles, and sizes. Getting that kind of labeled data is the tricky part.

Data labeling is slowly transforming as an industry of its own. AI and machine learning models require large volumes of labeled data to understand it as humans do. Maintaining high-quality training datasets should be of utmost importance for the models to work more accurately in real-world problems.

The Global Market Insights highlighted that the global data annotation tools market size exceeded USD 1 billion in 2020 and is expected to reach USD 7 billion by 2027. [2]



Data labeling enables the machines to understand the precise conditions as it helps increase the accuracy of data used to train machine learning algorithms. The most efficient characteristic of labeled data in AI is its ability to update the dataset. As and when new unstructured data comes in, the AI automatically labels it and uses it as a dataset. This improves the accuracy of the AI model as it gets better and better over a period of time, without the need to change a single line of code.

Prerequisites to build a high-quality dataset

Data volume

More data for better training

Accuracy

Accurately labeled data

Relevance

Data relevant to the full scope of the use case

Benefits of sophisticated and AI-powered data labeling

Better quality of training data

Improves accuracy of the output

Enhanced end-user experiences

Types of data labeling

Data labeling is a task that the machine learning algorithm has to perform from the given data. For instance, if a machine learning model is built to detect and inspect defects, then a dataset consisting of images of rust or cracks is fed to the model. The corresponding annotations would be polygons of localization of the rust and cracks, and tags for naming them.

Below are some common AI domains with their data annotation types.

1. Computer vision

Image Classification

Categorizing images into classes by adding tags to them. Unique tags represent the number of unique classes that the model can classify. It can be further divided to

- Binary class classification, where only two tags are assigned
- Multiclass classification, where multiple tags can be assigned

Image Segmentation

Separating objects in the images at pixel level from their backgrounds and other objects in the same image.

Object Detection

Detecting and locating objects in images with labeled bounding boxes.

Pose Estimation

Detecting key points (human joints like elbows, wrists, etc) in the human body and correlating them for obtaining the pose.

2. Natural language processing (NLP)

NLP is the analysis of human languages and initially requires manually identifying the important sections of the text by drawing bounding boxes and tagging them with specific labels. Such NLP models are used for sentiment analysis, entity name recognition, and optical character recognition among others.

Entity Annotation

Annotating specific features like nouns, keyphrases, parts of speech, and other structured texts from documents.

Text Classification

Classifying documents into predefined categories with one or multiple labels based on sentiment (for sentiment analysis) or the basis of the topic of the text (for topic categorization).

Phonetic Annotation

Labeling the commas, semicolons, and full stops present in the text. This annotation is specifically necessary for chatbots.

3. Audio processing

Audio annotation is labeling all kinds of sounds like speaker identification, wildlife noises (such as barks, etc), surrounding sounds (such as breaking of glass, etc) by converting them into a structured format for further processing with the help of NLP algorithms. It, first, requires manually transcribing the audio into written text and then adding tags and categorizing the audio to be used as the training dataset.

Improving the efficiency and accuracy of data labeling

The simplest approach of labeling to increase efficiency and accuracy is by labeling all the data at hand and creating the ground truth for the machine learning model. However, there are a few best practices like

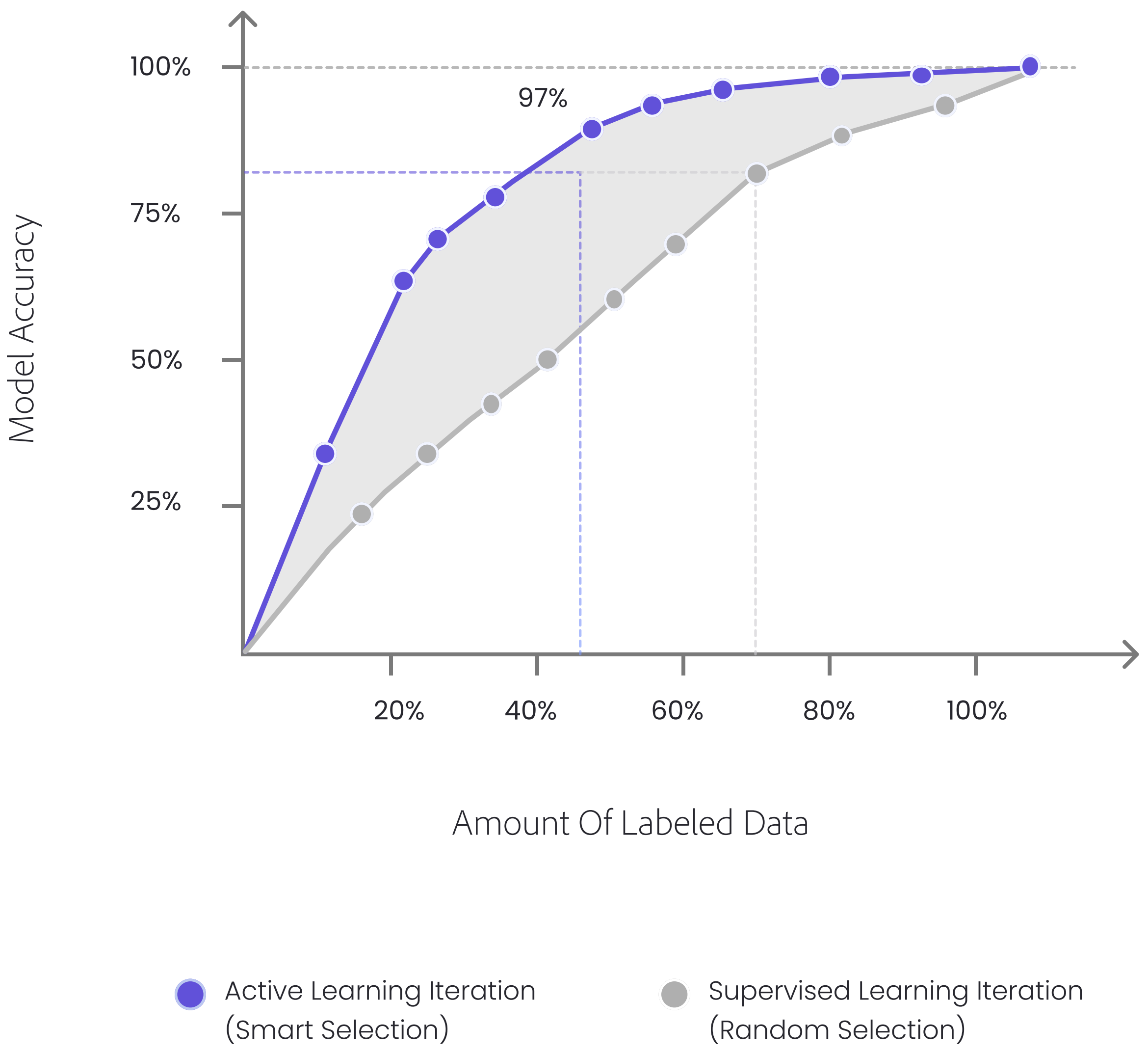
- Opting for intuitive and streamlined task interfaces to minimize the cognitive load context switching for data labelers
- A periodical and continuous audit of labels to keep a check on their accuracy and update them as and when required
- Leveraging machine learning to identify the most useful data from the unlabeled data. This process is known as **active learning**.

How does active learning help?

Active learning is the science of applying machine learning algorithms to simplify and automate the process of data labeling in order to reduce the cost of data labeling. In this semi-supervised approach, the data annotators select an initial sample from the unlabeled data with the aim of providing more reliable labeling using as few labeled instances as possible.

Based on the results at each step, the annotators incrementally and selectively label more data to the system until every data point is labeled. It has been observed that active learning works best on large volumes of unstructured and unlabeled data such as tweets, news articles, images, etc. This approach is beneficial for a variety of machine learning use cases like biomedical imaging, fraud detection, industrial equipment recognition, and many more.

Model "Learning Curve" [3]



How does active learning work?

Step 1

Label raw data

Step 2

Training model on labeled data

Step 3

Identifying the low-confidence data samples

Step 4

Label each data sample and add it to the training data

Step 5

Retraining the model with the new training data

Step 6

Monitor the labeling process and accuracy

Repeat from Step 2 until the model achieves the desired accuracy.

Essential components of data labeling

A high-performing data labeling pipeline will require a strategic combination of workforce with technical knowledge, tools, and operations that can consistently deliver high accuracy across entire datasets. Below are a few essential considerations for organizations thinking to implement this concept to accelerate high-quality data processing.

Annotation tools

Security

Pricing

Data quality

Workforce

Annotation tools

Annotation tools are a prerequisite to any data labeling process. There are specialized tools for various annotation types, which makes it obvious that choosing a specific tool will depend on your problem statement and that supports your out-of-the-box requirements. As the labeling process grows, new integrations with the annotation tools will develop in the pipeline. Therefore, choosing the right annotation tool that uses modern web technologies for easy integrations will make the job easy. You only need to decide whether you want to build it or buy it from a third party.

While choosing your annotation tool, consider the following

1. Filtering tools based on your use case
2. Decide - to build or to buy

Consider the following when deciding whether to build or buy a data labeling solution

Factors	**Build** Plan, design, build, test, integerate, maintain and scale	**Buy** Integerate scalable solution instantly
Data and types	Less data on a specific data type	Large volumes of specific data types
Scope of modificatins	One time solution	Oppurtunity to scale and expand use cases
Type of use case	Specific to the organization	Generic
Cost incurrence	Oppurtunity, building and maintaining costs	Predictable costs
Project timeline and resources	Time consuming huge financial budgets	Less time consuming, quick deployments
Workforce with domain expertise	AI, machine learning, data science, data collection and annotation	No or insuffieciently exerienced employees in these domains
Project management expertise	Yes	No
Continuity and reliability	Dependancy on internal resources	Continuous access to dedicated teams

3 Organization size and growth [4]

Start Early stage companies establishing process	**Scale** Growth stage companies accelerating development	**Sustain** At-scale companies consolidating control for long haul
Open Source Labelling Tool · DIY approach to workforce · Low nominal cost · Quick starting · Common task types	Self Built and Managed · Maximum control and security · Stack Integeration · Change management agility	Self Built and Managed · Maximum control and security · Stack Integeration · Change management agility
Crowd Sourcing Platform · DIY approach to tooling · Prototype and refined task details · On demand turn around	Commercial Software · Out of box/lightly customized · Balance of cost and control · Competetive features · API Integeration · Resource-lite deployment	Commercial Software · Fully custmized · Bespoke tooling · Resource-lite deployment

4 Offers easy integrations with modern technologies

5 Has an easy interface and provides a better user experience

Data quality

The quality of the data labels is the most critical part of data labeling. It refers to the accuracy across the overall dataset. Optimizing the labeling pipeline for the highest quality with available resources is a continuous process as the quality depends on various factors -

such as the functioning of annotation tools, ambiguity in their guidelines, the expertise of the workforce, quality assurance workflows, and the type/nature of the data itself.

Ways to measure the quality of data labeling

Test questions

Where the quality is measured based on correct and incorrect tasks.

Sample review

Where a small set of sampled annotations are reviewed by expert annotators for accuracy.

Consensus

Where the same task is assigned to several labelers and the correct answer is the one that comes from the majority of labelers.

Workforce model

Large volumes of data need the right amount of data labeling workforce that caters to its needs. To scale the data labeling functions more effectively,

- Decide upon the workforce strength depending on the data volume
- Allocate the workforce depending on the frequency of labeling
- Hire annotators only after understanding and analyzing the above two
- Measure the labelers productivity in terms of speed and accuracy
- Enable the feedback and review mechanisms with the labeling teams

The organization planning to choose the right personnel should consider the following labeling workforce approaches during its decision-making process:

In-house

Where the employees within the organization take part in the labeling process.

Outsource

Where the organization can hire a group of labelers, who are also known as cloud workers.

Third-party companies

Where companies specialized in offering data labeling services are hired.

Crowdsourcing

Where the organization can hire mass groups who perform the labeling tasks on crowdsourcing platforms on the internet.

Labelling workforce approaches

Approach	Description	Advantages	Disadvantages
Inhouse	Labeling task assignment to internal data science teams	· Higher accuracy · Progress tracking · Predictable results	· Time consuming · Expensive
Outsource	Task assignment to remotely working cloud annotators	· Annotator's skill evaluation	· The need to organize workflow · Comparatively expensive than crowdsourcing and third-party companies

Approach	Description	Advantages	Disadvantages
Third-party companies	Hiring an external company for the specific labeling task	- Quality assurance - Cost-effective - Time-saving	- Risk of leak of the labeled data
Crowd sourcing	Hiring freelancers from crowdsourcing platforms	- Cost savings - Faster results	- Low quality - Lack of confidentiality

Pricing

Another essential element for data labeling is pricing. The model a data labeling service uses to determine pricing can affect the overall cost and quality of the data. Pricing is a complex process as a slight variation in the data type, annotation types, the number of classes, speed, and/or the volume of the data can influence the pricing.

Critical factors that influence the labeling price

Duration of the project

Is it long-term or a one-time?

Quality, cost, and turn-around time

Evaluate and rank these in order of importance

Pricing model

Pay per hour or pay per annotation

Internal costs

Recurred if any part of the labeling process is carried out internally

Security

A data labeling service should comply with the necessary regulatory and other requirements based on the level of security the data needs. There should be a facility where the work can be done securely, with the right training, policies, and processes.

How can you protect your data?

- Conduct background checks of the annotators
- Making it mandatory for labelers to sign an NDA or any other similar document that outlines the data security requirements
- Training the annotators on the security protocols related to data
- Measuring the workforce for compliance
- Prohibiting the annotators from using devices like mobile, etc in the workplace
- Disabling the download feature on the devices that label the data
- Video monitoring the physical security of the workplace

A checklist to determine when to invest in a data labeling platform

- ✓ Do you have the required workforce to take on larger data labeling projects?
- ✓ Are your data scientists, engineers, labelers, and product managers wrangling between multiple systems?
- ✓ Do you have full visibility in recruiting the data labeling workforce, managing them, and evaluating their performances?
- ✓ Do you have secure and reliable processes that comply with regulations and ensure data security?

Factors to consider for a data labeling platform

Now, let's understand what to look for when deciding on which data labeling platform to use as it's a fact that high-quality labeled datasets are critical to developing high-performance AI models. There are numerous platforms to consider, with some combination of annotation tools, task design frameworks, and the right workforce to annotate the data. Whether to choose a platform that performs simple and straightforward labeling tasks, or the one that can handle complex and subjective tasks depends on the specific use case.

The following factors and critical questions should be considered while choosing the right data labeling platform:

Technology

- Can the platform build and distribute high-volume data labeling tasks?
- Can the platform support the machine learning needs of the organization?
- Is the platform powered with AI in order to prefill labels using prebuilt ML models, custom trained models, or third-party models?
- Does the platform support processes and workflows like multi-stage annotation workflows, complex classifications, etc.

The expertise of the platform vendor

- The level of expertise the vendor has in working on similar use cases
- Can the provider deliver enterprise-level data labeling?
- Is the provider technically sound enough to process and employ best practices to scale and improve the data quality?
- Are the use case management and support included in the provider's pricing?

Quality control

- Is the platform equipped with necessary quality control and assurance tools?
- Does the provider deliver an optimized and targeted quality control strategy based on the use case and budget?
- Examine the QA methodologies that the provider is using to evaluate the labeling accuracy.
- Can the provider adapt and maintain quality in case of scaling?

Tools

- Determine if the tools can annotate images, videos, texts, and/or audio.
- Are the tools configurable for a range of use cases as and when the use cases evolve?
- Is the platform provider constantly investing in R&D to improve the tooling capabilities, accuracy, and efficiency?
- Is the platform equipped with task management tools to assign labeling tasks across teams depending on the skill sets and specialization?

Platform security

- Is the platform meeting certain industry standards like CMMC, DFARS, etc?
- Are the data encryption and platform access controls well defined?
- Is the provider flexible enough to accommodate the IT and data teams' security requirements additionally?
- Are the annotators ready to sign an NDA?

Conclusion

The application of artificial intelligence (AI) and machine learning (ML) has become an essential element for many organizations that tend to innovate products and services, improve productivity, and disrupt their respective industries. To bring these AI solutions to the real world, a large amount of high-quality labeled data is required to feed and train the ML models. Machine learning models totally depend on labeled data, as such data increases the efficiency of every AI project, thus making a strong AI system for every business and industry. And if you have such high-quality labeled data that you want to leverage in order to capture the hidden opportunities with data-driven decisions and are unsure on the "how to proceed" and "what's next?" scenarios, DeepLobe is the solution.

References

1. Biological data annotation via a human-augmenting AI-based labeling system
   https://www.nature.com/articles/s41746-021-00520-6#author-information
2. Global Market Insights
   https://www.gminsights.com/industry-analysis/data-annotation-tools-market
3. KDNuggets
   https://www.kdnuggets.com/2018/10/introduction-active-learning.html
4. Cloud Factory
   https://blog.cloudfactory.com/steps-for-choosing-data-labeling-tool

DeepLobe

About DeepLobe

DeepLobe is a leading enterprise API platform and a best-in-class machine learning development tool enabling rapid and custom model building and iteration of unstructured images, videos, and text. With an unmatched developer experience, DeepLobe comes with a repository of pre-trained, out-of-the-box AI models that detect explicit content, embed images and predict various attributes, detect objects, etc.

A flag-ship product from SoulPage IT Solutions, DeepLobe leverages computer vision and deep learning technologies for making them easily accessible through API and derives innovative embedded data-rich insights from any text, image, or video.

www.deeplobe.ai