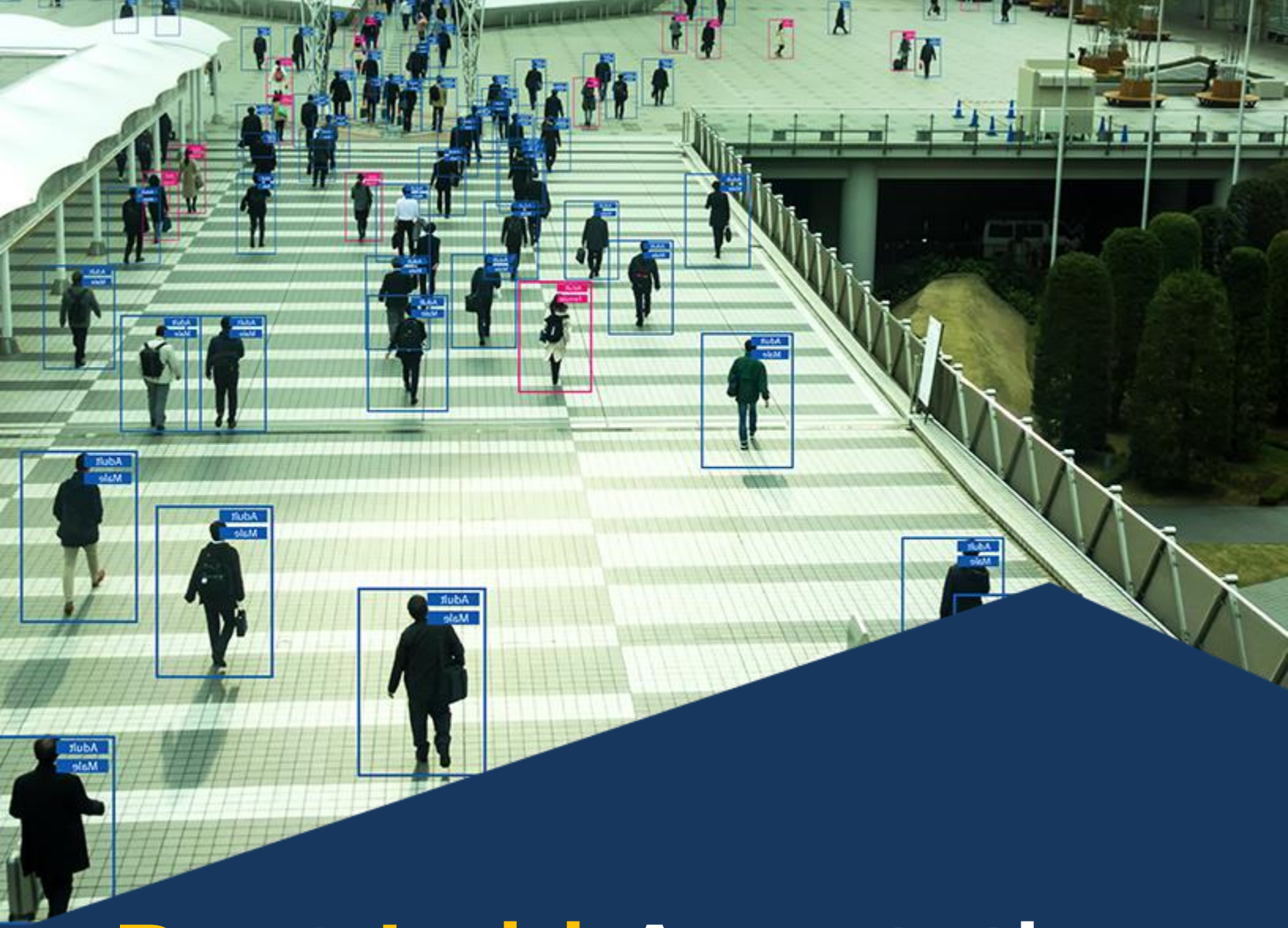

Best-Laid Annotation Plans for the Best-Paid Machine Learning Model

Practices, Procurement, and Business Perspectives

COGITO

Preface

It’s profoundly awestricken to see the conversational AI acquiesced in Amazon's Alexa responding wisely to human voice commands. Technology is magical — it never ceases to amaze us with its magnificence. *Artificial intelligence (AI) is the new kid on the block — this marvel of modern technology has got mankind stunned with a whole bit of breakthroughs in the form of internet-of-things (IoT), algorithm-integrated machine learning models, and automated industry-specific systems.* There are no two ways about it — the future of human civilization lies in artificial intelligence and AI-based machine learning models.

Automation, AI, and machine learning have turned out to be the new-fangled interests of the scientific society of humankind. Frankly speaking, the whole course of chores in the future world will be under complete command & control of AI-powered machines, and we have this testimony before us in the shape of self-driving cars from Google & Tesla Motors and automated assembly of robotic systems at manufacturing facilities. Artificial intelligence lurks in the background of everything we do online or even offline today — whether it's leafing through your Facebook newsfeed, searching Google for a product or service, or booking a train or flight ticket online.

Data — as we all see it — is at the helm in the world of virtual technology. It holds lord-like dominance around every inlet and outlet of the enormous global tech infrastructure. Though the influx of data can be attributed to several fingers of technology — social media, cloud computing, and this latest evolution of artificial intelligence (AI) & machine learning (ML) are most known to have disrupted the dynamics of data distribution & usage. While every technology has its sense and significance for collecting, using, and applying data, AI & ML models are a whole different ball of wax in the territory of technology — the most sought-after data-driven thing of recent times.

It, in the light of what we have discussed above, leads to the fact that developing AI-integrated machine learning models, in reasonable likelihood, is a great deal of data & data science that goes into sourcing the suitable datasets for AI development — and then — preparing the correct training data for machine learning models for a successful computer vision program. The data fed into the platforms and machines work in integration with AI algorithms and computer vision setup to power up process automation. These intrinsically interlinked mechanisms enable the machine learning system to make accurate predictions or do things on its own with minimal or entirely no human intervention.

This ebook is very much inclined to clarify the concepts of data annotation and data labeling, i.e., the essential prerequisites in the course of AI integration into machine learning models. The end purpose of the book is to provide you with an insight into the significance of data and data-inclined processes, e.g., data annotation and data labeling — that propel the path to the development of quality training data for AI and machine learning models.

Introduction

Enclosing datasets with points and lines in layman's terms is called data annotation, while data labeling is said to be a form of meta-information provided to the set of annotated graphics. It is, however, crucial that when teaching a computer, it is given a sufficient amount of data to learn from — the data that has been accurately annotated with the bounding boxes and labeled with the right metatext.

For a model to decide and take action, it must be trained to recognize and grasp a specific set of data and respond in line to automate certain tasks. Building an AI or a machine learning model that acts like a human requires plenty of training data. Training data needs to be correctly categorized and annotated for a specific use case, and this refers to data annotation. Experts can help you build and improve AI implementations through quality data annotation and labeling expertise. The result, thus, is an enhanced customer experience solution, such as product recommendations, relevant search engine results, computer vision, speech recognition, chatbots, and more.

Data labeling, on the contrary, is identifying raw data (images, text files, videos, etc.) and adding one or more informative labels to provide the machine learning model with the proper context. Labels show the characteristics of visual and audible media, whether a photo contains a car or bike, which words were voiced in an audio recording, or if an x-ray test report includes a malignancy. Data labeling is done for a variety of use cases, including computer vision, natural language processing, and speech recognition. It is up to your idea how you want to harness the potential of data labeling to automate your industrial practices.

Both annotation and data labeling processes are reliant on human expertise. Annotators need to bind images and assign boxes to them, whereas data labels need to manually label the media with the right metatext — in order for machine learning algorithms to understand the data they are fed. There are different ways to annotate and label data: video annotations, text annotations, or semantic annotations are the primary types of annotations. The whole thing about annotation and data labeling concludes that the attributes that matter to a machine learning algorithm can't be calculated without the correct data, i.e., the data which is annotated and labeled as it should ideally be.

AI is all around us. However, it's yet in the developing stage — where computers still, at some or other point, need manual intervention to perform specific tasks. The entire scheme of data annotation and labeling calls for a human approach. The experience & technical expertise of human-in-the-loop is needed for the sake of precision and accuracy, though automated annotation techniques are there for an easy go with training data development.

(1) Image Annotation — Application and Use Cases

At the time of developing the AI-enabled machine learning model, the first and foremost requirement is the relevant training datasets. Only the suitable datasets can power up the AI algorithms to build the necessary cognizance in a machine.

In the visual perception-based AI-based machine learning model, you need images containing the objects that are around us in the real world. To make the object of interest identifiable to machine learning models, the visual data collected in the form of images or videos need to be annotated and labeled with the correct annotation and labeling techniques while ensuring the quality and accuracy of the annotated data. Image annotation is a technique used to produce such annotated images at scale. The demand for image annotation in machine learning and AI- is substantial in terms of model success.

(1.1) Image Annotation

Image annotation is the process of marking the object of interest identifiable to machine learning models. To make things recognizable in the images, they are annotated with enclosed or bounding boxes and metadata for the explanation of the objects. When a large amount of similar data is fed into a model, it becomes trained enough to recognize the objects. The data and algorithm fed into a machine learning model empower it with human-like intelligence, which sticks up there to help the machine detect and predict real-life things & events.

(1.2) Applications of Image Annotation

Annotated images are primarily used to make the machine learn how to detect different types of objects. But there are other applications of image annotation as per their features and functions, ML algorithms compatibility, and its use for automation in various industries.

Here are some essential industrial applications of image annotation:

I. Detecting the Objects of Interest

Detecting objects in images is an important application of image annotation. A picture consists of many objects or things. A machine wouldn't need to observe each object in order to recognize the picture. By using computer vision technology, the image annotation technique annotates and makes these objects detectable so they can be identified.

II. Recognizing Types of Objects

After detecting the object, it is also essential to recognize what kind of object it is. a human, animal, or non-living object like a vehicle, street pole, or other man-made objects visible in the natural environment. Image annotation is key to recognizing and marking the objects of interest in a particular picture or video.

Object detection and recognition go simultaneously, and while annotating the objects in different cases, the notes or metadata are added to describe the attributes and nature of the object so that machines can easily identify such things and store the information for the future.

III. Classifying Objects of Interest

Objects in a picture do not always belong to the same category. For instance, a dog can be seen by a man, so classified or categorized objects in the images can distinguish them. Classifying objects in images is also an important application of the use of annotations.

IV. Object localization

is done with the use of image annotation practice along with image classification. In this image annotation approach, multiple techniques are used to annotate objects and classify them into different categories for detecting and categorizing things using visual perception-based artificial intelligence.

V. Object Segmentation of Single Class

It is important to classify objects according to their attributes as well as their category, location, and class to make it clear what the thing is. Semantic segmentation image annotation classifies the pixels in an image based on the class to which each belongs.

Images are annotated primarily for the purpose of making AI models or machine learning algorithms more accurate in their learning about objects in the images. An image annotation method is used to produce precise results in various scenarios in deep learning-based AI models for semantic segmentation.

VI. Recognizing Humans' Faces

Today, smartphones and security cameras can recognize faces using artificial intelligence (AI). What was the breakthrough in AI that made that possible? With image annotation, the human faces can be recognized through computer vision with the ability to identify individuals from a database and distinguish them from the crowd by way of the perspective of the security surveillance system.

The dimensions and various points of the human face, such as the chin, ears, eyes, nose, and mouth, are annotated in image annotation algorithms for facial recognition.

(2) Data Annotation vs. Data Labeling

While both data labeling and annotation are often used interchangeably, their process, use, and how it applies to the machine learning model are quite different. Data annotation is basically the technique of marking the data so that the machine can understand and memorize the input data using machine learning algorithms. Data labeling, also called data tagging, means to attach some meaning to different types of data in order to train a machine learning model. Labeling identifies a single entity from a set of data.

(2.1) What is Data Annotation?

Simply providing a computer with massive amounts of data and expecting it to learn to speak is not enough. The data has to be collected and presented in such a way that a computer can easily recognize patterns and inferences from the data. This is usually done by adding relevant metadata to a set of data. Any metadata tag used to mark up elements of the dataset is called an annotation over the input.

In machine learning, data must be annotated, or to simply say, labeled, so that the system could easily recognize it. But, for the algorithms to learn effectively and efficiently, the annotation of data must be accurate and relevant to the job the computer is tasked with. Simply put, data annotation is the technique of labeling the data so that the machine can comprehend and apply the input data in the due course of automation or prediction.

(2.2) What is Data Labeling?

Data comes in many different forms, such as text, images, audio, and video. To enable machines to recognize the data through AI algorithms, the data needs to be labeled. Data labeling, as the name suggests, is the process of identifying raw data to attach meaning to different types of data in order to train a machine learning model. When the data is labeled, it is used for training advanced algorithms to recognize patterns in the future.

(2.3) How Data Annotation Differ from Data Labeling

Annotation is a process used in supervised machine learning for training data sets to help machines understand and recognize the input data and act accordingly. Labeling is used to identify key features present in the data while minimizing human involvement. Labeling is a cornerstone of supervised machine learning, and various industries still rely heavily on manually annotating and labeling their data. Because poor labeling may lead to compromised AI, tagging or annotating must be done accurately so that they can be used for AI applications.

Basically, labeling is adding metadata to the data to make it more informative and meaningful to machines so that they can understand and learn from it. Labels may indicate whether or not an image contains a person, in which language an audio file is recorded, or what kind of action is performed in a video.

(2.3) Purpose behind Annotation and Data Labeling

It is still common for various industries to provide annotated and labeled data by hand. Labeling is essential to supervised machine learning. As far as NLP is concerned, labels are used to identify dataset features. Data annotations, on the other hand, are used for visual-based perception models.

As a general rule, labeling is more difficult than annotations. Annotations can be used for computer vision, but labeling is for training advanced algorithms to identify patterns in the future. In order to develop an NLP-based AI model, these two processes must be done with absolute accuracy.

(3) Types Of Data Annotation

(3.1) Image/Video Annotation

Understanding the types of image annotations used in creating the machine learning training datasets for deep learning-based AI models should have helped you understand the applications of image annotations in artificial intelligence and machine learning. Get a sneak peek into the application of different kinds of image annotation in various industries and the business arena with use cases of AI-based models.

i. Bounding Box Annotation to Easily Detect the Objects

Bounding box annotation is one of the most popular techniques used to detect the objects in an image. The objects of interest are annotated either in a rectangular or square shape to make the object recognizable to machines through computer vision. All types of AI models like self-driving cars, robots, autonomous flying objects, and AI security cameras rely on data created by bounding box annotation.

ii. Semantic Segmentation to Localize Objects in Single Class

To recognize, classify, and segment the objects in a single class, semantic image segmentation is used to annotate the objects for machines' more precise and accurate detection. It is the process of dividing the images into multiple segments of an object having different semantic

definitions. Autonomous vehicles and drones need such training data to improve the performance of their AI model.

iii. 3D Point Cloud Annotation to Detect the Minor Objects

The image annotation applications not only include object detection or recognition but can also measure or estimate the types and dimensions of the object. 3D point cloud annotation is the technique that helps to make such objects detectable by machines through computer vision. Self-driving cars are the use cases where these training data sets are created through 3D point cloud annotation. This image annotation helps to detect the object with additional attributes, including lane, sideways, and path detection.

iv. Landmark Annotation to Detect Human Faces & Gestures

Landmark annotation is another kind of image annotation that is used to detect human faces. The machine learning models like AI cameras, which are used for security surveillance in gadgets and smartphones are one of the best examples that work on the back of Landmark annotation. The AI-based security cameras and surveillance systems work to identify human faces and recognize gestures and various human postures. Landmarking is also used in sports analytics to analyze the human poses performed while playing outdoor games. Cogito provides the landmark point annotation with the next level of accuracy for the precise detection of human faces or their poses.

v. 3D Cuboid Annotation to Detect the Object with Dimension

Detecting the dimensions of the object is also crucial for AI models to get a more accurate measurement of various objects. The 2D images are annotated by capturing all the dimensions visible in the image to build a ground truth dataset for the 3D perception of the objects of interest. Again autonomous vehicles, AI robots, and visual perception models are used to detect indoor objects like carton boxes with the dimension that need such annotated images, created through 3D cuboid annotation.

vi. Polygon Annotation to Detect Asymmetrical Shaped Objects

Similarly, polygon annotation is used to annotate objects that are in irregular shapes. Coarse or asymmetrical objects can be made recognizable through the polygon image annotation technique. Mainly road markings or other objects are annotated for self-driving cars. And autonomous flying objects like drones viewing the objects from an aerial view can detect or recognize such things when trained with training data sets created through polygon annotation for precise object detection.

vii. Polyline/Splines/Line Annotation for Lane or Path Detection

Lines, Polylines, and Splines are similar image annotations used to create the training data sets allowing computer vision systems to consider the divisions between essential regions of an image. The boundaries, annotating lines, or splines are helpful to detect lanes for self-driving cars. Road surface markings that indicate the instructions for driving on the road need to be also made understandable to autonomous vehicles. Polyline annotation divides one region from another region.

(3.2) Audio Annotation

A subset of data labeling – audio annotation – is a critical technique for building well-performing natural language processing (NLP) models that offer many benefits to organizations, such as analyzing text, speeding up customer responses, recognizing human emotions, etc.

Audio annotation involves classifying components of audio that come from people, animals, the environment, instruments, and so on. In the course of audio annotation, data scientists specify the labels or "tags" by using software and pass the audio-specific information to the NLP model being trained.

i. Sound Labeling

Here, the annotators segregate and label the sound files provided with a recording. Typically, they mark the sound files with specific keywords.

ii. Event Tracking

In testing or training audio data, there is little control over the sound events that overlap numerous times. Event tracking is done to check the performance and efficiency

of sound signals and tracking systems in different scenarios like those we encounter in our daily routine, where the source of sound can hardly be heard in the distance

iii. Speech to Text

Developing NLP (Natural Language Processing) technology is in fact an integral part of it. Transcription of recorded speech involves both transcribing the audio into text and labeling the sounds and words of the person delivering the speech. The correct punctuation should be used, and the sentence should be carefully crafted.

iv. Audio Classification

Audio classification involves listening to and analyzing audio recordings. These machines can distinguish voice commands from sound commands based on the data they gather. Having audio annotations such as these, text-to-speech applications, and virtual assistants can be enhanced.

(3.3) Text Annotation

Among the most exciting areas of AI development is natural language processing (NLP). The use of NLP solutions like chatbots, automatic speech recognition, and sentiment analysis enhances productivity and efficiency across a wide range of businesses.

Recent innovations in NLP have shown potential to assist the speech impaired with communication abilities through AI-enabled speech recognition devices. However, none of these great technologies would be possible without text annotation and the companies that provide these annotation services.

Entity annotation is one of the essential processes in the generation of chatbot training datasets and other NLP training data. It is the act of locating, extracting, and tagging entities in text. Types of entity annotation include:

> **Named entity recognition (NER):** This is to annotate the entities with proper names.
> **Keyphrase tagging:** This is for locating and labeling the keywords and/or keyphrases in the text data.
> **Part-of-speech (POS) tagging:** This is for discernment and annotation of all the functional elements of speech (adjectives, nouns, adverbs, verbs, etc.).

i. Identifying Grammatical Characters

Entity annotation teaches NLP models how to identify parts of speech, named entities, and keyphrases within a text. In this task, annotators read the text thoroughly, locate the target entities, highlight them on the annotation platform and choose from a predetermined list of labels. To help NLP models learn about named entities further, entity annotation is often paired with entity linking.

ii. Entity Linking

Whereas entity annotation is the location and annotation of certain entities within a text, entity linking is the process of connecting those entities to larger repositories of data about them. Types of entity linking include:

iii. End-to-end entity linking

The collaborative process of first analyzing and annotating entities within a text (named entity recognition) and engaging in entity disambiguation.
Entity disambiguation: Process of linking named entities to knowledge databases about them.
Search functions and user experience can both be improved by entity linking. It is the responsibility of annotations to link labeled entities in the text to URLs that provide additional information about the entity.

iv. Text Classification

Often referred to as text classification or document categorization, text classification involves annotators reading long passages of text or small sentences. Annotators must analyze the content, discern the subject, intent, and sentiment within it and classify it based on a predetermined list of categories. Whereas entity annotation is the labeling of individual words or phrases, text classification is the process of annotating an entire body or line of text with a single label. Related text annotation types include:

v. Document Classification

This includes the classification of documents used to help with the sorting and recall of text-based content.

v. Product Categorization

Crucial for eCommerce sites, product categorization is the sorting of products or services into intuitive classes and categories to help improve search relevance and user experience. Sometimes annotators are shown product descriptions, product images, or both. The annotators would then choose from a list of departments or categories that the client has provided.

(3.4) Sentiment Annotation

This involves the classification of text based on the emotion, opinion, or sentiment within the text. Because text classification is a broad category, various annotation types like product categorization or sentiment annotation are technically just specialized forms of text classification.

Emotional intelligence is one of the most difficult fields of machine learning. Sometimes it is difficult even for humans to guess the genuine emotion behind a text message or email. It is exponentially more difficult for a machine to determine connotations hidden in texts that use sarcasm, wit, or other casual forms of communication. To help machine learning models understand the sentiment within the text, the models are trained with sentiment-annotated text data.

More broadly referred to as sentiment analysis or opinion mining, sentiment annotation is the labeling of emotion, opinion, or sentiment inherent within a body of text. Annotators are given texts to analyze and must choose which label best represents the emotion or opinion within the text. A simple example would be the analysis of customer reviews. Annotators would read the reviews and label them as positive, neutral, or negative.

When built correctly with accurate training data, a strong sentiment analysis model can accurately detect the sentiment in user reviews, social media posts, and more. The sentiment analysis model would then allow businesses to track public opinion about their products, allowing the companies to develop future strategies or alter current plans accordingly.

(3.5) Linguistic Annotation

Also referred to as corpus annotation, linguistic annotation simply describes the process of tagging language data in text or audio recordings. With linguistic annotation, annotators are tasked with identifying and flagging grammatical, semantic, or phonetic elements in the text or audio data. Types of linguistic annotation include:

Discourse annotation: Linking of anaphors and cataphors to their antecedent or postcedent subjects. Ex: James broke the chair. He felt awful about it.
Part-of-speech (POS) tagging: Annotation of the different function words within a text.
Phonetic annotation: Labeling of intonation, stress, and natural pauses in speech.
Semantic annotation: Annotation of word definitions.
Linguistic annotation: To create AI training datasets for a variety of NLP solutions such as chatbots, virtual assistants, search engines, machine translation, and more.

These were all 5 types of data annotation commonly used in machine learning today.

(4) Types of Data Labeling

Data is the fuel for Machine Learning-based AI developments that provide the set of patterns that the machines can recognize and understand, thereby storing such information for future references. A considerable amount of such datasets is required to ensure and widen the horizon of Machine Learning, analyze the various scenarios, and predict with accurate results.

But the problem here lies in producing many quality datasets. This is one of the biggest challenges for Machine Learning engineers. Data labeling, especially for visual perception AI models like self-driving cars, drones, or robotics, needs annotated images to understand the surroundings better and take actions accordingly.

Data labeling is making objects identifiable to machines through Computer Vision. Image annotation is the technique used to manually annotate images using tools and software. But this process takes time and lots of effort to produce the Machine Learning training datasets as per various needs.

(4.1) Manual vs. Auto Labeling

Technically, there are two approaches – Automated data labeling and Manual data labeling used to annotate the images. Both methods have their advantages and disadvantages. So, to select one, we need to be wise in distinguishing which one is the best approach to produce high-quality training datasets for Machine Learning. Apart from the aforementioned two, the data labeling techniques, an advanced AI-assisted data annotation method is also used in the industry for a better and more efficient way of producing the machine learning training data sets at the best pricing.

(4.2) Manual Data Labeling

The manual data labeling process seems very simple, but it is very time-consuming, and it requires more skills and effort to manually annotate the objects in the images. In this process, annotators are presented with a series of raw, unlabeled data like images or videos and are tasked with labeling it according to a set of rules or specific data labeling techniques.

Let's take an example. While annotating images, various types of image annotations like bounding boxes, polygon, point cloud annotation, and semantic segmentation are used to make the objects in the pictures recognizable to machines through computer vision technology.

Bounding box or polygon annotations are two of the easiest and cheapest annotation techniques, taking less time and effort, while semantic segmentation takes more time and fine-grade annotations. Objects are selected from a given list in the manual data annotation process.

Compared to the automated labeling process, if you would assume that it takes a user 10 seconds to draw a bounding box around an object and select the object class from a given list. If there are datasets of around

100,000 images with 5 objects per image, it will take about 1,500 person-hours to label, and it would cost approximately $10K to label this data.

(4.3) Quality Check

Apart from the primary annotation process, a quality check is another stage where the manual data labeling process takes enormous time. During this, it would take a trained user approximately one second to check off each bounding box annotation, resulting in an increase of 10% in the cost of labeling.

While few workflows may choose to adopt consensus-based workflow, the time and the money spent are proportional to the number of users that work on overlapping tasks for consensus. To simplify, if you have deployed three users to label the same image three times, you must pay for all three annotations, thereby increasing the annotation cost.

In manual data labeling, these two steps are crucial – first, data labeling, and second, checking and verifying to ensure the quality of annotations. In automated data labeling, the data labeling and verification process take less time than manual annotation.

But now, thanks to AI and Machine Learning, automated data labeling techniques have come a long way. Though not all automated systems are developed equally, in many cases, the use of AI-based labeling needs more human participation to be corrected promptly by the AI. Hence, there needs to be a conscious effort to ensure how the overall workstream is affected by an AI-enabled system.

(4.4) Automatic Data Labeling

To reduce time and effort in data labeling, we need to understand the workflow process and find out the key areas wherein we can make the system automated.

A. Auto-Label AI labels raw, unlabeled data;
B. A human user verifies the label;
C. If data is labeled correctly, it is added to the pool of labeled training data;
D. If data is incorrectly labeled, this data is considered valuable for re-training the Auto-label AI and a human labeler who will proceed to correct the errors on a trial basis;
E. Once the data is labeled at a satisfactory level, it is then used to re-train the Auto-Label AI and subsequently added to the pool of labeled training data;
F. Finally, ML teams use the compiled labeled training data to train the various models.

Let's assume that a well-trained AI model is available and is making accurate predictions in most of the cases, and only the edge cases need human assistance to check and correct the errors. The Data Labeling Process contains most of the manual process which is usually done by humans.

Even for AI models like autonomous vehicles that need a considerable amount of training datasets to train the machine learning algorithms, most data points only have the marginal value to improve the model's performance. Though edge-cases in the long-tail distribution of data (like car accident scenes) are rare, they help to improve model performance.

Hence, AI companies are in a quest to develop the most accurate and highest performing Machine Learning model. However, the sizes of datasets are not matters of concern. Instead, the game of finding and gathering the most edge cases is vital to training the AI model.

Hence, it is more important for a company to set aside the unimportant cases of data that AI models are already well-trained on and devote more time and effort sensibly in labeling the high-value data. Thus, the AI-assisted data labeling process provides a better capacity to produce accurate datasets making AI possible in various emerging fields.

(4.5) AI-Assisted Data Annotation for Speed & Accuracy

Though the automated data labeling process provides multiple times faster speed labeling and the advancement of technology brings more efficiency and quality, Human-in-the-loop machine learning is vital to ensure quality and accuracy while labeling the data for Machine Learning. Using the AI-assisted labeling process makes the labeling task easier and provides them with ideas on which object in the image needs to be annotated. Once the data is annotated, annotators can manually check or fill the area left behind or need proper annotation.

Cogito is one of the top data annotation companies, providing AI-assisted data labeling services to deliver the most advanced annotation platform to produce high-quality training datasets for Machine Learning and deep learning-based AI models.

The use of the right AI-enabled tool and data labeling software is vital to producing a considerable volume of training data for different types of models. Data labeling with Machine Learning is multiple times faster. It gives an edge to annotators to utilize their skills in ensuring quality and accuracy, making every data understandable to machines for appropriate predictions.

(5) What is Training Data in ML Model?

Training data sets are the key ingredients for machine learning to build a well-organized functional model that can work with the highest relevancy and give accurate results. And it is tough to find such training data that can fit into your demand for machine learning. This discourse clarifies how to create training data for machine learning with simple processes and other alternatives to gather such data at an affordable cost.

(5.1) Building Training Dataset

To build a functional model, you must keep in mind the flow of operations involved in building a high-quality dataset. To solve a particular problem through an AI-integrated machine learning model, the data should be accurate and authenticated. Just take an example if you want to determine a person's height, then other features like gender, age, weight, or the size of clothes are among the other factors considered seriously.

In this example, a person's clothes, weight, and height are essential, while color and fabric material will not add value to training data. Such irrelevant data have very little weightage while predicting the person's height, and as per the machine learning golden rules, the larger the data, the results would be a better help to create robust models.

(6) How Much Training Data Is Sufficient

Training data is the crucial input to machine learning (ML) and having the right quality and quantity of data sets is vital to getting accurate results. The larger the training data available for the ML algorithm, the more it will help the model perceive the diverse types of objects, making it easier to recognize them when used in real-life predictions.

But the question here is, how will you decide how much training is enough for your machine learning. As insufficient data will affect your model prediction accuracy while more than enough data can give the best results, you can manage the big data or massive quantity of datasets. It also requires deep learning or a more complex way to feed such data into algorithms.

Many factors decide how much training data is required for machine learning, like your model complexity, machine learning algorithms, and data training or validation process. And in some cases, how much data is needed to demonstrate that one model is better than another.

(6.1) Complexities of the Problem and Learning Algorithms

One of the significant aspects when selecting training data for machine learning is the complexity of the problem. It means that the unknown underlying function that relates to your variable inputs to the output varies as per the ML model type.

Similarly, the complexity of the machine learning model algorithm is another important factor considered while choosing the correct quantity of data sets. The algorithm is used to inductively learn the unknown underlying mapping function from specific examples to make the best use of training data and integrate the same into the machine learning model.

(6.2) Using the Statistical Heuristic Rule

In statistical terms, there are many components considered as a factor of the number of classes, a characteristic of the number of input features, and an aspect of the number of model parameters. And there are statistical heuristic methods available that allow you to calculate a suitable sample size.

In the factor of the number of classes, x independent examples need to be selected for each type, where x

could be tens, hundreds, or thousands depending upon the parameter range. While input features must be X% more examples than input features and in model parameters, there must be independent examples for each parameter in the model.

(6.3) Model Skill vs. Data Size Evaluation

While choosing the training data set for machine learning, you can design a study that can evaluate the model skill required against the size of the training dataset. To perform this study, plot the result of your model prediction as a line plot with training dataset size on the x-axis and model skill on the y-axis, which will give you an idea of how much the quantity of data affects the dexterity of the model while solving a specific problem with machine learning.

You can use a learning curve in which you will be able to project the amount of data required to develop a skillful model or perhaps how small data you need before touching an inflection point of diminishing returns. So, you can perform the study with available data and single-performing algorithms like random forest and suggest you develop robust models in the context of a well-rounded understanding of the problems.

(6.4) More Data Required for Nonlinear Algorithms

Nonlinear algorithms are usually known as one the most powerful machine learning algorithms. As they are capable of learning the complex nonlinear relationships between inputs and output features. If you are using nonlinear algorithms, you need an adequate amount of data sets and need to hire a machine learning engineer that can work with such applied mathematics.

Such wisely-developed algorithms are flexible and even nonparametric at times, i.e., they can help find out the number of parameters required to model a problem while also helping out to figure out the values of those parameters. The predictions with such models vary based on the particular data used to train them, resulting in lots of data required for such model training.

(6.5) Don't Wait for More Data. Get Started with What You Have

You need to get sufficient training data for your ML. However, waiting to acquire such data for long days is not a sensible decision. Don't let the problem of the training dataset size stop you from getting started on your model prediction problem-solving.

Get started with the data you can, use what you have, and check how effective models are on your problem. Acquire something, then take action to understand better what you have for further analysis, and then increase the data you have with augmentation or collect more data from your domain to make your model training more accurate.

(6.6) Quality vs. Quantity of Training Data

The quality and quantity of training data are some of the most essential factors machine learning engineers or data scientists take into serious consideration while developing a model. However, in the future, it will become more apparent how much training data suffices for machine learning model development. Still, it is clear now that "the more, the better," hence if you can gain as much data and use the information as you want. It would be better for you, but waiting for bid data acquisition for a long time can delay your projects.

(7) What are Datasets?

Data is an essential component of any AI model and the sole reason for the spike in the popularity of machine learning that we witness today. Because of the availability of data, scalable ML algorithms became viable as actual products that can bring value to a business rather than being a by-product of its fundamental processes.

(7.1) It's a Collection of Data Pieces

A dataset in machine learning is simply a collection of data pieces that can be treated by a computer as a unit for analytic and prediction purposes. This means that the data collected should be made uniform and understandable for a machine that doesn't see data the same way as humans do. For this, after collecting the data, it's essential to preprocess it by cleaning and completing it, as well as annotate the data by adding meaningful tags readable by a computer.

A vast amount of datasets are required for developing training data for the machine learning models. The training data is annotated with the appropriate bounding boxes & pointers, which are further labeled with the relevant metatext so that the machine can learn the patterns and memorize the same for predictions and automation. Artificial Intelligence and Machine Learning companies have enormous demands for such annotated data to train their ML algorithm.

It might look like an easy and straight task to collect a dataset for your AI project while you focus on building the machine learning model and pour most of your energy and time into that. Nevertheless, in practice, dealing with data frequently takes the most time because of its sheer scale. Therefore, it is important to grow familiarity with what a dataset is, how it is collected, and what features a dataset should have.

(8) Training, Testing, and Validation Datasets in Machine Learning

Datasets are typically used for more than just training. Usually, a single training dataset is divided into several parts so that it can be checked how well the model was trained. An independent testing dataset is commonly used for this purpose. Additionally, a validation dataset, while not strictly necessary, helps you avoid making biased predictions when training the same data.

The training data, test data, and validation data hold distinct values in the course of machine learning model development. While all three data types are split up from one large dataset, each one has its own distinct use in ML modeling. Let’s start with an elaborative definition of each term:

(8.1) Training Data

The training data is used to build the machine learning algorithm. This properly annotated and labeled data serves as input data for developing an AI-integrated machine learning model. The training data is fed into a machine learning model to result in the automation of the process involved. Training data, in general, makes the true foundation of an AI-enabled machine learning model as this is what directs the machine to make appropriate decisions in the course of automation.

(8.2) Validation Data

Validation data is the data that hasn’t been evaluated before. This unseen, unused, or unevaluated data comes into use to commence the first test of the AI integration to check whether the machine learning model is able to make the right prediction or not. Not every data scientist uses the validation data. However, it can provide some helpful insight into the optimized hyperparameters — which validate data’s accessibility and usability for the machine learning model.

(8.3) Test Data

Test data is for the purpose of revalidation. Data scientists do a double check of the machine learning model whether it is functional enough to make the right predictions or not. While the training and validation data are labeled with the appropriate metatext, the testing data usually rests unlabeled as it is seeped into the machine just to monitor the performance metrics of a machine learning model. This is for the final real-time check just to confirm if the ML algorithm has been trained effectively or not.

As the training data is to train a machine learning model, the test data is to confirm if the model works properly. While each of these three datasets, i.e., training validation, and test data has its significance for building AI-powered ML models, the purpose behind their application into the machine learning model is different.

(9) Principal Sources of Datasets

The sources for collecting a dataset vary depending on the need and purpose of machine learning projects. It strongly depends on the project, budget, and size of a business for which the computer vision program is to be developed. The best approach is to collect the data from the web or IoT sensors that directly hold a connection with the business objectives. You may, however, find that this gives you the most control over your data collection, but it is a more complex process and requires more time, money, and resources.

Other ways, like auto-generated datasets, require significant computational powers and are not suitable for any project. There are large comprehensive repositories of public datasets that can be downloaded freely and used for the training of your machine learning algorithm.

The obvious advantage of free datasets is that they're, well, free. On the other hand, you'll most likely need to tune any of such downloadable datasets to fit your project since they were built for others initially and won't fit precisely into your custom-built ML model. Still, this is an option of choice for many startups, as well as small and medium-sized businesses, since it requires fewer resources to collect a proper dataset.

(9.1) Pre-labeled Datasets & Their Reliability

Pre-labeled datasets can be accessed and collected from different sources. The source either might be a private dataset developing firm working in the AI and machine learning space or it might be a governmental firm that receives or processes a truckload of data in the course of its industry or business operations.

Here are some considerable benefits of pre-labeled data:

- Pre-labeled datasets are promptly available to use;
- They save you huge on costs;
- They save you a huge time.

Having access to pre-labeled datasets, at one point, might seem lucrative at first as they bring you a time-and-cost saving proposition. However, they might not be as reliable or workable as on-demand custom-developed datasets. The pre-labeled datasets might not be up-to-the-mark or the standards set for the quality dataset needed for a machine learning project. As far as images are concerned, the resolution might be too poor to even rightly locate and mark the objects for labeling and annotation. This goes the same with the pre-labeled text, audio, and video data.

Despite a bunch of benefits of pre-labeled datasets, the pre-labeled datasets carry a bundle of drawbacks as follows:

- The pre-labeled datasets miss the quality mark as required for the machine learning model;
- The pre-labeled datasets disappoint when it comes to quality;
- The abundance of pre-labeled datasets might be full of non-relevant data;
- The pre-labeled data may need more time to filter out the quality data that matters to the machine learning project.

(9.2) Selecting the Appropriate Data Annotation & Labeling Tool

When selecting an annotation tool, there are several factors that stakeholders should consider and balance. Engineering and annotation project managers should endeavor to use standard infrastructure. It is wise to prefer the selection of an existing and supported annotation platform or tool over the implementation of a new tool or ad hoc workflow.

The primary drivers for tool selection should be to ensure data privacy and access by the workforce. The stakeholders should also evaluate technical back-end requirements. It is also essential to consider how the tool will facilitate project management needs. Stakeholders should also establish desired UI/UX features for the annotation tool and project interface.

i. Use standard infrastructure
Implementation and maintenance of new annotation tools lead to technical debt over time and potential governance issues. Whenever possible, choose an existing annotation platform or tool that is already supported by the organization.

ii. Ensure data privacy and access by the workforce
As outlined in the previous section, intellectual property, data privacy, and security considerations impact the workforce type and which annotation tools can be used for an annotation project. Stakeholders should agree on a tool that can be used in the overall project workflow and that the chosen workforce can access.

iii. Evaluate technical back-end requirements
At a minimum, the selected tool should enable efficient upload, download, storage, and data processing. The tool should also support standardized and easy-to-use data formats like JSON and XML.

iv. Facilitate project management needs

Project management functionalities should be built through wisely picked tools that can help annotation project managers monitor the status of the project, communicate with the workforce, identify issues with the annotation process, and control quality. Specifically, the selected tool should accommodate annotation project managers' need to:

- convey written guidelines and updates to the workforce;
- control work queues, including priority/order;
- access audit logs for work items;
- perform quality assurance, including leveraging annotation redundancy and "gold" tasks;
- provide feedback to the workforce; and
- track quality, productivity, and timeliness metrics.

(9.3) Establishing Desired UI/UX Features

The selected tool should have an interface that can be customized to the annotation project in order to give the workforce the ability to execute the annotations in an efficient and accurate manner. Stakeholders should ensure that the UI/UX features of the tool do not bias the workforce, affecting the validity or reliability of the annotations. The workforce should have the ability to efficiently:

- access the written guidelines and updates for the project;
- read/view the relevant data in the work item;
- add/edit/delete annotations;
- navigate through each work item/work queue, including revisiting previously submitted work items, if that is desirable; and
- flag work items with potentially corrupt or unsuitable data, if that is desirable.

(9.4) Common Problems with the Data Annotation Tools

Before moving on to a large number of capabilities that one might look for in annotation tools, it is worthwhile to consider a set of general problems that many tools present that can reduce their utility. A tool might in theory have a capability you need, but it might be so

difficult to use or so buggy as to make it practically impossible to leverage that capability.

Ask your data annotation partner to avoid a set of such annotation tools that feature any one of the following problems in the course of the data annotation process:

1) Inadequate importing, exporting, or conversion
2) Lack of documentation or support
3) Difficult to learn
4) Poor user interface
5) Difficult with installation
6) Unstable, slow, or buggy
7) Unavailability of technical support in the hour of need

While an annotation tool may look sound with its easy-to-play user interface, it may disappoint when it comes to stability. You wouldn't like your outsourced data annotation partner working on any set of tools that are buggy, slow, and may pose risk to your data security. Having the right annotation tools by your side is a step to conquer the challenges that might strike in the course of the annotation process.

(9.5) Betting on the Right One for Your Annotation Project

The annotation tools which come with the right documentation & user manual and easy installation process are worth picking for your annotation project. Moreover, also make sure that you are provided with trouble-free access to the technical support in the hour of need — just in case the annotation tool stops responding to the controls and commands.

The nature of the workforce and annotation tools to use for an annotation project is influenced by intellectual property considerations, data privacy, and security. Stakeholders should choose a tool that can be integrated into the overall project workflow and one which is accessible to the workforce.

(10) Getting Your Training Data Ready

A quality dataset should correspond to specific standards. For smooth and fast training, make sure your dataset is relevant and well-balanced. Try to use live data and consult with experienced professionals about the volume of the data and the source to collect it.

Raw data is an excellent place to start from, but you obviously cannot just shove it into a machine learning algorithm and hope it offers you valuable insights into your customers' behaviors. There are quite a few steps you need to take before your dataset becomes usable.

(10.1) Collect

The first thing to do when you're looking for a dataset is to decide the sources you'll be using to collect the data. Usually, there are three types of sources you can choose from: the freely available open-source datasets, the Internet, and the generators of artificial data. Each of these sources has its pros and cons and should be used in specific cases. We'll talk about this step in more detail in the next section of this article (click here to read it now).

(10.2) Preprocess

There's a principle in data science that every experienced professional adheres to. Start by answering this question:

has the dataset you're using been used before? If not, assume this dataset is flawed. If yes, there's still a high probability you'll need to re-appropriate the set to fit your specific goals. After covering the sources, we'll discuss the features that make up a proper dataset (you can click here to skip to that section now).

(10.3) Annotate

After ensuring your data is clean and relevant, you also need to make sure it's understandable for a computer to process. Machines do not understand the data the same way as humans do (they aren't able to assign the same meaning to the images or words as we). This step is where

many businesses often decide to outsource since keeping a trained annotation professional is not always viable. We have a great article on building an in-house labeling team vs. outsourcing this task to help you understand which way is the best for you.

(11) Data Collection Mechanisms

Data is a crucial source for any invention. Collecting the data from reliable sources timely is vital to preserving its essence and usability. Before you set yourself for anything with your machine learning project, make sure that you have the right and accurate data in the kitty that applies to your machine learning project. Instead of relying on your in-house team and resources, partnering with a dedicated data collection firm is the wise approach. A company delicately working on data collection can bring you filtered and high-quality data for your next machine learning initiatives.

Here are the data collection mechanisms that data collection companies take leverage of to provide you with well-optimized data for your machine learning model:

(11.1) Natural Language Utterance Data Collection

Different people can use different words, phrases, and different sentences to say anything having the same meaning. Experts can easily find the different ways people express the same request and systematize it with data in order to make it teachable to various innovative technologies like AI and Machine learning to develop a suitable functional model.

(11.2) Data Collection with Data Annotation & Tagging

Data collection is not enough. We also need to add meaning to this data. We annotate videos, audios, images, and text to make the information valuable for building, training, and testing with machine learning algorithms. There are companies that can help you with data

collection and analysis, annotation, and tagging projects in various languages for any location with expertise in semantic annotation, object recognition, data processing services, and data enrichment services.

(11.3) Scalable Data Collection at Faster Speed

We are scalable to handle a large amount of data with heaps of requests every month for such data operations to make an enriched database. We are scalable to deal with big companies powered by our hardworking distributed team members with improved technology. Our effective screening process and custom software help to deliver a high-quality, scalable data service.

(12) Picking the Right Annotation Partner

Engaging your team and expensive resources in collecting the datasets is a faulty play when it comes to getting your data ready for the machine learning project. It is advisable to focus on picking the right data collection partner instead of going in-house for data collection. Every data collection mechanism has its value in the course of data collection and developing training data for AI-enabled machine manufacturing.

Cogito, Anolytics, and sort of other industry players are doing well in the market when it comes to annotating and labeling datasets for developing training data for AI-powered machine learning models. Choosing the right data collection and annotation partner will determine the success of your machine learning project.

Keep in mind the following attributes when picking a data collection & annotation expert/company to partner with for your data annotation project:

i. Industry exposure for The industry-specific data collection & annotation

The company you decide to ally with for the data collection and annotation must have market exposure in the industry domain of your relevance. Also, take count of the data collection approaches and data annotation tools & techniques that are utilized in the course. This check on the approaches, tools, and techniques will help you determine the goals, quality of the work delivery, and possible deadline for your machine learning project.

ii. Market repo & credibility as an annotation Expert

This is key to the success of your annotation project. You are paying your money, you have your choices. However, make sure that the partner you are picking for data collection and annotation has that credibility in the market. The market repo refers to the experience and expertise of a data annotation company and being sure about the reputation of your prospective annotation partner brings you assurance for the timely accomplishment of your machine learning project while not missing anything on the quality front.

iii. The one that ensures data security

Confidentiality of your data is something you would not appreciate being at risk. It's a concern that must be pondered over while on the lookout for a data annotation partner for collaboration — this is even more worrying when you are outsourcing to or through a third party. To ensure the security of your data when outsourcing annotation and labeling assignments to a third-party vendor, make sure that you are aware of their security protocols for threat-sensitive data.

iv. Assurance of quality output

How your AI-enabled machine learning model performs is completely dependent on the training data it has been fed — the training data that is the matter of the quality

annotation work. Before you step ahead for picking the annotation partner, have a know-how of the types of the quality control mechanisms and approaches they have in place. Also look for their sourcing practices, i.e., where they get the experts on board to handle expert-specific annotation tasks and what is their experience and expertise level in annotation & data labeling.

(13) Pricing

You pay your money, you make your choices. However, it will be helpful to decide on the money part before you confirm your partnership with a data annotation company. Deciding on the budget is the key prerequisite before proceeding with an annotation and data labeling company. It is also wise to have a vendor by your side to help you determine the budget for your machine learning model development while also helping you with an honest evaluation of the annotation assignment.

(14) Request a Proof-of-Concept

Having narrowed down all your choices around choices for a data annotation collaboration, go for a pilot project before you headfirst. The pilot project refers to a small annotation task that you can assign to the annotation company that you can request for a proof-of-concept. This is an intelligent approach to get an insight into the annotation work that your prospective data annotation partner is capable of delivering. With the proof-of-concept, you will be able to wisely decide on the delivery deadlines and quality of annotations and labels by your to-be annotation partner.

The more you engage with your annotation partner through conversation, the more are the chances of success of your machine learning model. As you grow familiar with your future data annotation and labeling partner, you prepare yourself to make an informed decision about picking the right partner to undertake your data annotation and labeling tasks.

(15) Planning Training Data Set

To build a functional model, you must keep in mind the flow of operations involved in building a high-quality dataset. The datasets need to be authenticated by experts when it comes to solving a particular problem through an AI-powered machine learning model. The first step toward creating machine learning data sets is selecting the right data sets with the correct number of features for datasets.

The data should be reliable and should have the least number of missing values because over 25 to 30% of missing values are not considered during the training of machines. However, in certain conditions, the relationship among different machine features is too intricate to comprehend and autonomously integrate with the appropriate machine learning training data. In such cases, you must manually impute and handle the missing values to get better results.

For example, if an organization has borrowed a loan from the bank, and a feature containing a value like the GDP of the country is available with only 30% missing values. If this feature has a very high weightage to predict whether the institution can repay the loan or not, then this feature needs to be considered with high priority. While if this feature doesn't have any importance in developing the AI model, then there is no need to include such data.

I. Collecting the Right Datasets

The first and foremost thing to start from is collecting the right and relevant datasets for the machine learning model. The machine needs to learn from the datasets that are in keeping with artificial intelligence.

II. Processing of Training Data

Once you have selected the correct data, the processing of data also involves a selection of the correct data from the complete data set and building a training set, and the entire process is given here step-by-step to process the training data service.

III. Organize and Format

As the data can be distributed into different files or folders, so you need to find the relation between the datasets and preprocess to form a dataset of required dimensions. And data also needs to be transformed into a universal language before organizing into a particular format.

IV. Cleaning of Data

This is one of the essential stages in data processing, where you must remove unwanted characters from the data and clean up the missing values. The missing values can be removed or replaced, and if there are such missing values, they can be replaced or removed to get more accurate results.

V. Extraction of Features

This stage involves the process of analysis and optimization of the number of features. Here you must find out which aspects are essential for prediction and select them for faster computations and low memory consumption. For example, if you are dealing with image classification, then remove all the irrelevant images.

VI. Conversion of Training Data

After data selection and processing, you must convert the data into a meaningful dataset. The data conversion process also involves several steps that are discussed below.

VI. **Conversion of Training Data**
After data selection and processing, you must convert the data into a meaningful dataset. The data conversion process also involves several steps that are discussed below.

VII. **Scaling**
It is essential when a dataset is placed while considering the linear data set like bank data. If the transaction amount feature is necessary, then data must be scaled to build a robust model. The correlation matrix, the Pearson method, is used to find the relationship between such things to perform this action. If definite values do not scale data, this might lead to a misunderstanding of the data.

VIII. Disintegration and Composition
This step involves breaking up a particular feature to build better training data for the model you understand more comprehensively. The splitting of the time-series element is one of the best examples of data disintegration in which you can extract the days, months, years, hours, minutes, seconds, etc., from a specific sample. Separating and processing such features may result in better accuracy to get a better output.

IX. **Composition**
What remains now is to combine unique features into a single part to get more accurate or meaningful data. The composition of data sets combined with distinct features can be said to be an authentic or high-quality dataset that can be used for machine learning. The more the data, the more precise and accurate the predictions would also be.

(16) A Check on the Quality of Training Data

Training datasets are the key ingredients for machine learning to build a well-organized functional model that can work with the highest relevancy and give accurate results. And it is tough to find such training data that can fit into your demand for machine learning. To make sure that the training data that you have created for your machine learning project has quality up to the mark, check it on the following standards:

A poor quality training data for your machine learning model is not good from any angle. And until you feed the right data your AI model will not give you accurate results. If you train the computer vision system with incomplete data sets it can give disastrous results in certain AI-enabled autonomous vehicles or healthcare.

(17) All in All, the Quality of AI Training Data is the Key

Generating the high-quality training data for AI-powered machine learning requires skilled annotators who can carefully annotate and label the information like text, images, or videos as your algorithm compatibility to make the perception model successful. Moreover, consistency in providing a high-quality image is more important, and only well-resourced organizations can provide such consistent data annotation service. Actually, there are a few quality control methods discussed below you can use to ensure the quality of data for your machine learning or AI projects.

Standard Quality-Assurance Methods:

I. **Benchmarks or Gold Sets Method**
This process helps to measure the accuracy by comparing the annotations to a “gold set” or vetted example. And it helps to measure how much a set of annotations from a group or individual meets the benchmark set for such a task.

II. **Overlap or Consensus Method**
This process helps measure the consistency and agreement among the group. And it is done by dividing the sum of agreeing on data annotations by the total number of annotations. This is one of the most common methods of quality control for AI or ML projects with annotation objective rating scales.

III. Auditing Method
The auditing method of checking the quality of training data measures the accuracy by having reviewed the labels by experts either by checking on the spot or by reviewing all. This method is crucial for projects where auditors review and retread the content until and unless it reaches the highest level of accuracy.

IV. **Detailed For In-Depth Quality Assessment**
To monitor the quality of data annotations, these baseline quality measurements are one of the solid methods. But if AI projects are different from each other, then organizations need to establish quality assessments in a customized way for a specific initiative. And only highly-experienced leaders can organize the in-depth quality control analysis by considering the process discussed below.

V. **Multi-layered Quality Evaluation Metrics**
This method is multiple quality measurement metrics that help to leverage the methods of quality measurement which have been already discussed. It can ensure maintaining the accuracy level at best while not delaying the project.

VI. **Weekly Data Deep Monitoring Process**
Under this method, a project management team is implemented to examine the data on a weekly basis and also set the expanded productivity and quality score. For example, if you need 92% accurate data, you can set the goal of 95% and try to ensure the annotation process exceeds your goals.

VII. **Management Testing and Auditing**
To build the quality-assurance skill set of your project manager you can ask them to carry out annotation work and quality audits to make them get first-hand experience of the annotation process. This method helps the management team have a 360-degree view of the projects and a full understanding of the entire annotation process.

VIII. **Get High-Quality Training Data for Unbiased Decisions**
The approach to ensure the machine learning training data quality also ensures the accurate algorithms, and it can also rationalize the potential bias in different types of AI projects. Bias can be distinct as uneven voice or facial recognition performance for different types of genders, speech patterns, or ethnicities. Therefore, organizations need to actively build diversity into the data teams by defining goals, metrics, roadmaps, and algorithms used to develop such models while also avoiding biases at the project level.

Hiring a data talent team is easier said than done but if the composition of your team is not representing the population your algorithm training will be affected. Yes, there is no doubt that the unavailability of high-quality training data is one of the prime reasons for AI & ML project failure. However, there are numerous quality assurance processes vital for AI development. Hence, quality training data is not only good for algorithm training but also helps to make the model work in the real world.

(18) Getting a Handle on Your Data Annotation Project

(18.1) Identifying Your Key Stakeholders

Identifying key stakeholders is an essential first step in the annotation process. Typically, there are three in case of an annotation project — the Project Manager, Annotation Project Manager, and Engineering Manager.

I. **The Project Manager is key to the success of your annotation project as he is the one who:**

- Leads from the front.
- Comprehend the objectives and end goals.
- Determines the project needs potential.
- Decides on the practical application for the project.
- prioritizes features based on user needs. may have subject matter expertise.

II. **Annotation Project Manager is the second lead after the project manager and who holds accountability for:**

- The day-to-day oversight of the annotation project.
- Managing the workforce.
- Ensuring data quality and consistency.
- The subject matter expertise.

III. **Engineering Manager who is primarily responsible for:**

- Getting the annotations to work in line with the implementation plan.
- Determining the tools, techniques, and other resources that go into the annotation process.
- Providing key input as to technical strategy and feasibility.

(18.2) Setting Up an Uninterrupted Communication Channel

Irrespective of whether the annotation project is conducted by a single team or a larger cross-functional group, communication among all stakeholders is very important. Communication should be clear and regular between all project stakeholders throughout the whole process. By doing so, the end result will remain in focus.

To get the most out of an annotation project, key stakeholders need to agree on a method for regular and clear communication, since the way in which they communicate determines:

- Methods, dynamic systems, and annotation development approach.
- Implementation plan of the annotation project that works effectively.
- shared development roadmap.
- Create a backlog of issues, and
- Schedule further meetings as and when required.

Managing the roadmap and backlog, and determining meeting agendas should be placed in the hands of the annotation project manager since he or she will be the most familiar with any issues that arise in the course of the annotation process.

(18.3) Establish High-level End Goals

Engage your project stakeholders in the discussion about any existing alternatives as well as the expected business impact of the project in addition to establishing the desired functionality. Discussing current solutions to the problem being addressed will help the stakeholders understand what can be done better, how those improvements can be achieved, and the value of those improvements.

An understanding of the end goals of your annotation project will give stakeholders a better view of the overall significance of the project — which will further help them determine the resource availability and overall timelines for the timely accomplishment of your annotation project.

(18.4) Consider Budget, Resource Availability, and Timeline

Depending on the nature and complexity of the data set, the type of workforce may affect the kinds and numbers of annotations that can be produced. Budget and resources will necessarily influence the way the project is defined because it is the budget that can fold and unfold the scope of resources and workforce to be employed for carrying out the annotation tasks. Timeline considerations may also significantly influence the way a project is defined. This is because a tight schedule may limit the manner in which annotations are completed.

(18.5) Consider How Annotations Will Be Used

Stakeholders should carefully consider the possibilities of automating annotations when passed directly onto end-users. When project goals include making future predictions on new data. The type and size of the machine learning model are the key determinants when it comes to deciding on the usability and implementation of the annotated and labeled data for the AI-enabled machine learning model. Stakeholders should consider a rules-based model when the goal of the project is to capture human-identifiable patterns in the data.

(18.6) In-house vs. Outsourced Fully Managed vs. Crowdsourced Data Labeling

Work with a team of experts. It requires far more time and money to find a skilled workforce than it does to find a workforce without the annotation and labeling expertise. Stakeholders should explore different workforce options to weigh both their benefits and limitations, in order to ensure that an appropriate workforce is selected for an annotation project. This workforce can be composed of three types:

(18.6.1) In-House

The in-house annotation handling approach comes in handy when it comes to handling sensitive data protected under law. you should consider hiring full-time employees. However, it adds up to the cost of hiring and managing an in-house team of full-time employees alongside the infrastructure composed of multiple workstations. The workforce new to the AI, ML, and annotation industry may also require training and supervision. The only thing that goes in favor of in-house annotation is the assurance of data security. Otherwise, in-house annotation is the most cost-intensive proposition among all three.

(18.6.2) Outsourced Fully Managed

It is a low-cost option and the most feasible when it comes to getting annotation and labeling tasks done within means and budget. An outsourced data annotation partner, with its low-cost team of annotation experts, can fully manage the whole hog of your annotation and labeling tasks. An outsourced annotation partner may have access to certain internally developed tools and nonpublic data as long as they are bound by confidentiality/nondisclosure agreements.

Communication and supervision of the annotation things going on is, however, a challenge when working with the outsourced partner — but this gets compensated with the quality and on-time work delivery at a cost suited to your budget.

(18.6.3) Crowdsourced Annotation

Though getting annotation and labeling things done through crowdsourcing is relatively cost-effective among all three workforce options, it lacks expertise when it comes to handling the core annotation or industry-specific labeling tasks. With the annotation handled by crowdsourced annotators, you have limited, or in some instances, entirely no scope to communicate and supervise with the individuals engaged.

There is limited opportunity for supervisors and managers to supervise and communicate with individual annotators as they receive training through written instructions available on the crowdsourcing platform. For quality assurance, the Crowd's work may be subjected to gold tasks, double-blind validation, or annotator random sampling, depending on the crowdsourcing platform used.

(18.6.4) Finding the Right Workforce

It is important to consider the pricing structure of an annotation project when engaging vendors. The type of incentive offered to annotators could affect their behavior. An author may increase annotation efficiency with transactional pricing, while a flat rate with quality metrics may increase annotation quality.

The time needed to develop and revise annotation guidelines and to select and train workforce members should also be included. If annotation guidelines are poorly designed, or the workforce is inadequately prepared, there may be wasted effort and lingering back-and-forth with the annotators that will lead to more time spent and unnecessary uncertainty during the annotation process.

(18.6.5) What makes an outsourced fully-managed workforce feasible to go with

An outsourced fully-managed workforce is worth considering when you have high concerns for the quality of the annotation work. The feasibility of an outsourced pool of specialists who work in sync is realized when the AI-based machine learning model is too big to be handled by the in-house team or by the crowdsourced experts. Allying with an outsourced annotation is relatively a profitable proposition for your annotation project when there are demands for a dedicated team with industry-specific annotation expertise.

Bottom Line

The scope and size of data collection, acquisition, annotation, and labeling have amped up with the growing opportunities for automation and AI integration into industrial practices. Besides unfolding the stratum of a rock-solid future of AI and machine learning, the expansion has leaped out the demand for quality training data, i.e., the data that is adequately annotated and labeled.

Data is predicted to be crucial to the success of companies in the future. Handling all the data will become increasingly difficult as the data volume grows. Cogito, an accomplished data annotation and labeling company, herein provides some sense about the significance of Big Data, collection and usage practices through this public and industrial discourse while casting some light on current data annotation and labeling practices.

Hopefully, rifling through the e-book might have given you a complete sense of what a successful machine learning model looks like and how you can set your plans and priorities for data annotation & labeling processes without being on the fritz and failure.

About Us

Cogito is a hybrid data labeling platform following model-assisted labeling (MAL) approach to cater industry's leading businesses. The MAL model leverages a human workforce to label a relevant portion of the training dataset which enables training of the AI application. Playing an important part as human-in-the-loop, our solutions encompass business verticals ranging from Retail, Manufacturing, Building, and Construction, to Medical, Food processing, E-commerce, and more.

www.cogitotech.com

COGITO