Draft Study Material

DATA ANNOTATOR

(Qualification Pack: Ref. Id. SSC/Q8120)
Sector: Information Technology-Information Technology Enable Services (IT-ITeS)

(Grade XI)

विद्यया ऽमृतमश्नुते
एन सी ई आर टी
NCERT

PSS CENTRAL INSTITUTE OF VOCATIONAL EDUCATION
(a constituent unit of NCERT, under Ministry of Education, Government of India)
Shyamla Hills, Bhopal- 462 002, M.P., India
https://www.psscive.ac.in

Preface

Vocational Education is a dynamic and evolving field, and ensuring that every student has access to quality learning materials is of paramount importance. The journey of the PSS Central Institute of Vocational Education (PSSCIVE) toward producing comprehensive and inclusive study material is rigorous and time-consuming, requiring thorough research, expert consultation, and publication by the National Council of Educational Research and Training (NCERT). However, the absence of finalized study material should not impede the educational progress of our students. In response to this necessity, we present the draft study material, a provisional yet comprehensive guide, designed to bridge the gap between teaching and learning, until the official version of the study material is made available by the NCERT. The draft study material provides a structured and accessible set of materials for teachers and students to utilize in the interim period. The content is aligned with the prescribed curriculum to ensure that students remain on track with their learning objectives.

The contents of the modules are curated to provide continuity in education and maintain the momentum of teaching-learning in vocational education. It encompasses essential concepts and skills aligned with the curriculum and educational standards. We extend our gratitude to the academicians, vocational educators, subject matter experts, industry experts, academic consultants, and all other people who contributed their expertise and insights to the creation of the draft study material.

Teachers are encouraged to use the draft modules of the study material as a guide and supplement their teaching with additional resources and activities that cater to their students' unique learning styles and needs. Collaboration and feedback are vital; therefore, we welcome suggestions for improvement, especially by the teachers, in improving upon the content of the study material.

This material is copyrighted and should not be printed without the permission of the NCERT-PSSCIVE.

Deepak Paliwal
(Joint Director)
PSSCIVE, Bhopal

Date: 29 September, 2024

STUDY MATERIAL DEVELOPMENT COMMITTEE

Members

Monika Sharma, Assistant Professor in IT-ITeS (Contractual), Department of Engineering and Technology, PSSCIVE, NCERT, Bhopal

Member Coordinator

Deepak D. Shudhalwar, Professor (CSE), Head, Department of Engineering and Technology, PSSCIVE, NCERT, Bhopal, Madhya Pradesh

TABLE OF CONTENTS

Module 1 Basics of Data Annotation

Module Overview

In this unit, you will get knowledge about data annotation and its role in Machine Learning. You will first start by understanding the concept of Machine Learning and, Artificial Intelligence, as well as the working of Machine Learning. You will also explore the different types of Machine Learning and also look into the important processes of training and testing in Machine Learning. You'll also learn about the various applications of Machine Learning and discover some essential software tools used in this field.

Moving forward, you will understand the concept of Data Annotation for Business, highlighting its importance, its applications, and its use in Artificial Intelligence. You will understand the role of Data Annotation in AI applications for businesses, exploring its uses and the advantages it offers. As you progress, you will explore the real-world Use Cases and Applications of Data Annotation, demonstrating its importance in different scenarios. Finally, you will more understand about Data Annotation, including its types, methods, and the various data types it involves, with a special focus on Computer Vision tasks and techniques. By the end of this unit, you will be ready about the fundamentals of Data Annotation.

Learning Outcomes

After completing this module, you will be able to:

- Understand the fundamental concepts and scope of machine learning and its significance in data-driven industries.
- Explore the role of data annotation in improving business processes and decision-making through accurate data labeling.
- Analyze real-world applications of data annotation across various industries to enhance machine learning models.
- Identify different types of data annotation, methods, and data formats to optimize machine learning outcomes.

Module Structure

Session 1. Introduction to Machine Learning

Session 2. Data Annotation for Business

Session 3. Use Cases and Applications of Data Annotation

Session 4. Data Annotation – Types, Methods and Data Types

Session 1. Introduction to Machine Learning

Once upon a time, there was a young girl named Amy who loved to draw. She would spend hours sketching and colouring her artwork. But sometimes, she felt stuck when deciding what to draw next. One day, Amy's teacher told her about a special computer program that used machine learning. This program could analyse her previous drawings and suggest new ideas based on her style. Excited, Amy tried it out and was amazed at the creative suggestions it gave her. With the help of machine learning, Amy's drawings became even more creative and special. She found out that technology could help her be even better at drawing and come up with lots of cool ideas for her art. Figure 1.1 illustrates the power of Machine learning.

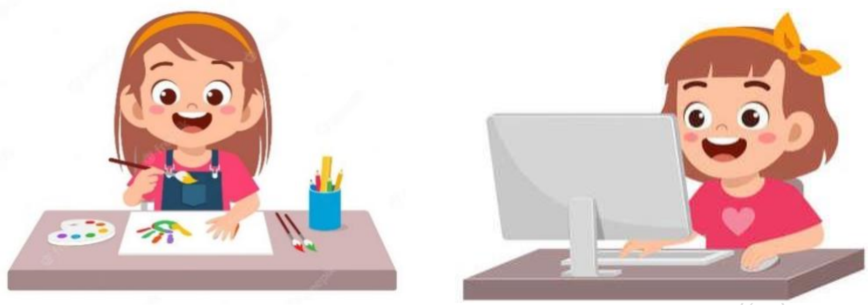


Figure 1.1: Power of Machine Learning

In this chapter, you will understand the concept of Machine Learning, the workings of machine learning, features, and types of machine learning. And in the end, you will also get familiar with the different machine-learning software tools.

1.1. Machine Learning- An Introduction

Definition: Machine Learning is a part of artificial intelligence. It is all about making computer programs that can learn from data and things that happened before. It is like training a computer to get better at something by showing it lots of examples, as shown in figure 1.2.

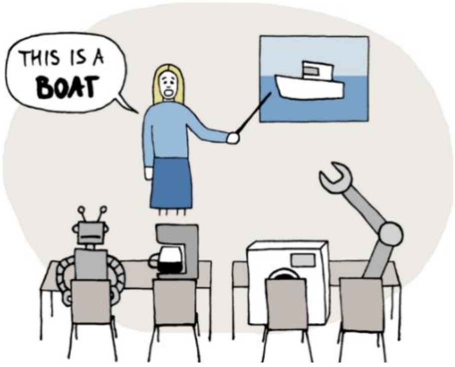


Figure 1.2: Example of training a machine/computer by showing example

For example, you have a pet dog ‘Max’. Max was a puppy, you taught him how to do tricks like "sit" and "roll over." You showed him these tricks over and over until he learned them. Max got better at performing the tricks the more he practiced. As shown in figure 1.3.

Figure 1.3: Instructing a dog (Source: https://www.vecteezy.com/)

Imagine teaching computers to learn independently, as Max taught tricks. As shown in figure 1.4, the human gives instructions to the machine, and it responds to him.

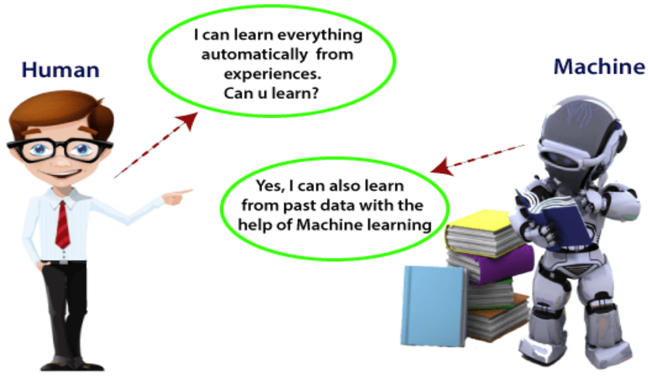

Human
I can learn everything automatically from experiences. Can u learn?
Machine
Yes, I can also learn from past data with the help of Machine learning

Figure 1.4: Human and machine communication (Source: Javatpoint)

1.1.1. Artificial Intelligence (AI)- Artificial Intelligence is a branch of computer science focused on developing computer systems that can imitate human intelligence. It combines the words "Artificial" and "Intelligence," meaning "human-like thinking power." In simpler terms, AI, or Artificial Intelligence, computers learn to think and make decisions like humans. It is like teaching a computer to think and learn on its own.

Figure 1.6 illustrates the example of a voice assistant, where you say something to your phone voice assistant and it responds accordingly. Examples are Google Voice Assistant, Siri, and Alexa.

Figure 1.6. Example of voice assistant

Difference between AI and Machine Learning

S.No.	Artificial Intelligence	Machine Learning
1.	Artificial intelligence is a type of technology that lets machines act like humans.	Machine learning is a type of artificial intelligence (AI) that lets a machine learn directly from previous data without being programmed explicitly.
2.	AI aims to create intelligent computer systems as smart as humans so they can solve difficult problems.	The objective of machine learning (ML) is to enable machines to gain knowledge from data, enhancing their ability to provide accurate outputs.
3.	In the field of AI, we develop intelligent systems capable of performing tasks similar to humans.	In machine learning (ML), we teach machines using data to perform a specific task and produce accurate results.
4.	Machine learning and deep learning are the primary subcategories within the field of AI.	Deep learning is a primary subset of machine learning.
5.	AI is mostly used for tasks like Siri, customer service chatbots, enjoying online games, robots, and so on.	The primary applications where machine learning is used are Google search algorithms, online recommendations, etc.

1.1.2. Deep Learning- Deep learning is a type of computer technology that helps computers to learn and understand things, like humans learn. In deep learning, computers use something called "neural networks," which are like virtual brains made of many tiny parts (neurons). These virtual brains learn and make decisions based on the information they receive. Figure 1.7 illustrates the graphical representation of neurons.

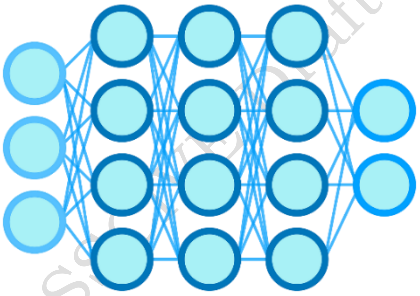


Figure 1.7: Deep learning neurons

Deep learning helps computers to identify patterns in things such as pictures, sounds, or even text. It is used in facial recognition on the phone, helping self-driving cars to see the road, or making suggestions on social media.

Difference between Deep Learning and Machine Learning

S.No.	Deep Learning	Machine Learning

1.	Deep learning algorithms depend on large amounts of data, so we need to input large amounts of data to achieve good performance.	Machine learning depends on large amounts of data, it can also work on smaller amounts of data.
2.	Deep learning models typically require more time for training, but they can execute tests relatively quickly.	Machine learning algorithms require less time for model training compared to deep learning, but they may take a longer time for model testing.
3.	Deep learning models require a large amount of data to operate effectively, a high-end system with GPUs is required.	Machine learning models require less data, and can effectively work on low-end machines.
4.	Deep learning models can process both structured and unstructured data since they rely on the layers of artificial neural networks.	Structured data is mostly needed for machine learning models.
5.	Deep learning models are appropriate for resolving challenging issues.	Machine learning models can be used to solve straightforward or a little bit challenging issue.

1.2. Working of Machine Learning

In Machine Learning, a system learns from old data. It creates models to make predictions and guesses the result upon receiving new data. It can make better predictions with lots of data because it makes a more intelligent model. As shown in figure 1.8, the working model of machine learning is given below.

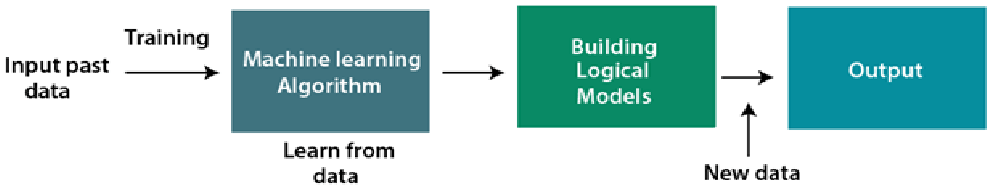


Figure 1.8: Working of Machine Learning (Source: JavaTpoint)

Here is a simplified explanation of the working of Machine Learning:

Step 1: Input past data: In this phase, we collect many examples, like various fruit pictures, for the computer to learn from. These pictures are the computer's training materials. It learns by repeatedly studying these pictures to identify patterns and differences between the fruits. As figure 1.9 illustrates, giving input to the computer.

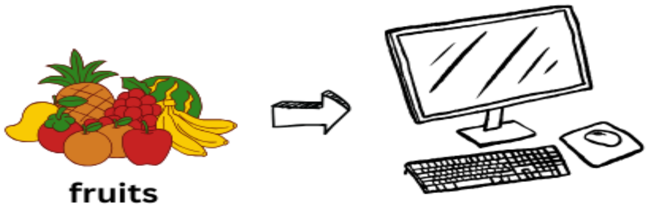


Figure 1.9: Giving input to the computer

Step 2: Machine Learning Algorithm: It is a set of instructions that helps the computer learn from the input data and make decisions. As figure 1.10 illustrates, applying an algorithm for giving instructions to the computer.

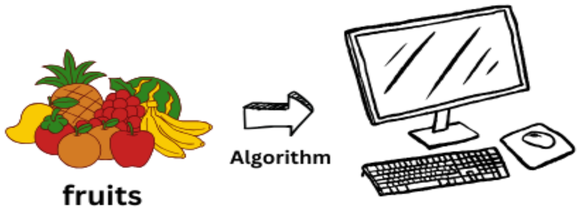


Figure 1.10: Applying the algorithm to the computer

Step 3: Learning from Data: The computer uses the algorithm to analyse the input data and find patterns or connections. For example, your friend looks at the fruit pictures and notices that apples are round and red, bananas are long and yellow, and oranges are orange and bumpy. Figure 1.11 shows a computer analysing the pattern of the given input data.

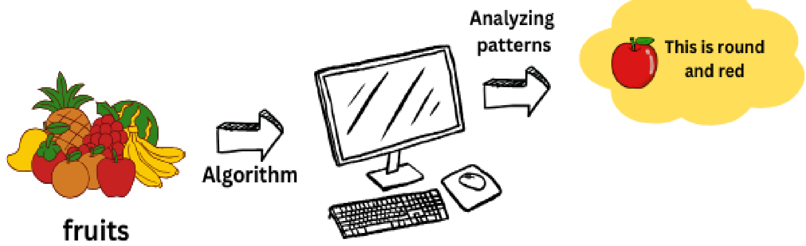


Figure 1.11: Analysing patterns

Step 4: Building Logical Models: The computer creates a set of rules based on what it learned, so it can recognize patterns in new data. Your friend writes down rules like "if it is round and red, it is probably an apple" or "if it is long and yellow, it is likely a banana". Figure 1.12 shows computer learning by analysing input data patterns.

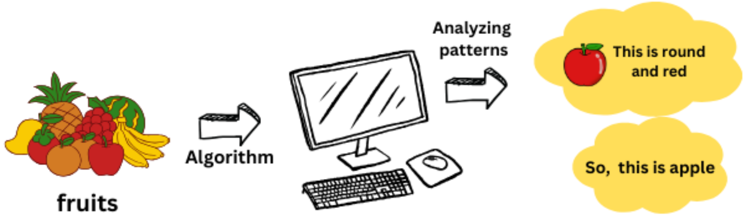


Figure 1.12: Computer learning

Step 5: Taking New Data: Now, the computer analyses new data and makes predictions or decisions. If you show the computer a new picture of a round and red fruit, it will predict that it is likely an apple. Figure 1.13 illustrates, giving new data to the computer.

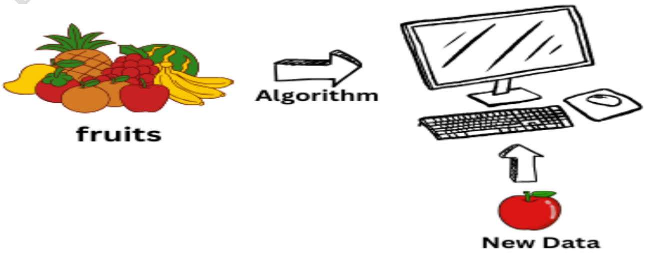


Figure 1.13: Giving new data to the computer

Step 6: Getting Output: The computer gives you an answer based on what it learned. The computer tells you, "This is an apple because it is round and red". As figure 1.14 shows, the computer learned and gave output.

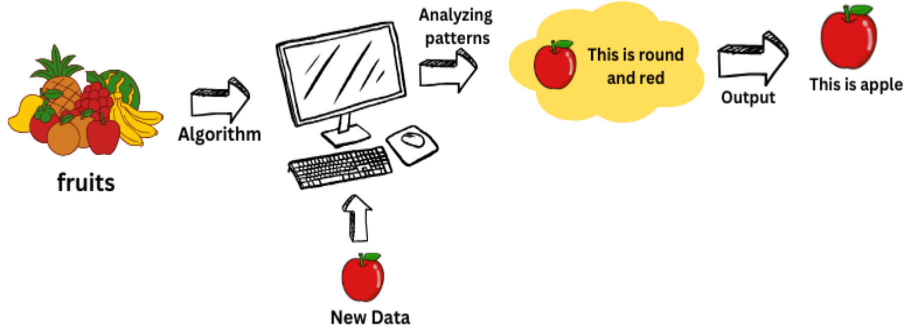


Figure 1.14: Computer learned and giving output

1.2.1. **Phases of Machine Learning**

Machine Learning lets computers learn by themselves. There are different phases of machine learning. These steps help the computer to find answers to the problems. Phases of Machine Learning are:

- Problem Understanding
- Data Collection
- Data Annotation
- Data Wrangling
- Model Development, Training, and Evaluation
- Model Deployment and Maintenance

a) **Problem Understanding**

In this phase, we identify a specific problem we want the computer to solve. For example, we want to identify different types of animals in pictures. Figure 1.15 illustrates below.

Figure 1.15: Problem understanding

b) **Data Collection**

In this phase, we collect lots of information or data related to the problem. It includes collecting different information that computers can use to learn. To identify the animal problem, we need lots of pictures of different animals. As figure 1.16 illustrates.

Figure 1.16: Different types of animals

c) Data Annotation

Data annotation provides labels or hints to the computer. At this step, we label or tag the data, which means we tell the computer what's in each picture. Like "This is a dog," or "This is a cat." As shown in figure 1.17.

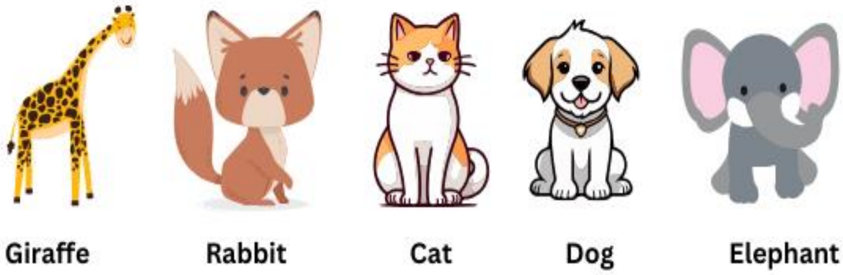


Figure 1.17: Labeling the animals

d) Model Development, Training, and Evaluation

In this phase, we create a computer program (model) and teach it using the labelled data. It is similar to teaching a pet new trick. This process is like teaching the computer to recognize animals by showing the pictures and telling the output. As shown in figure 1.18.

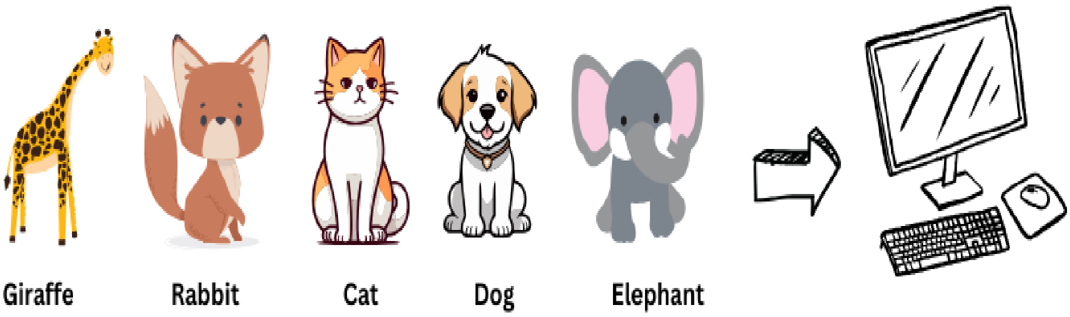


Figure 1.18: Labelled data given to the computer

e) Model Deployment and Maintenance

Once the computer has learned well, we can use it to identify animals in new pictures. We also need to make sure it keeps working correctly by updating it if needed. As shown in figure 1.19.

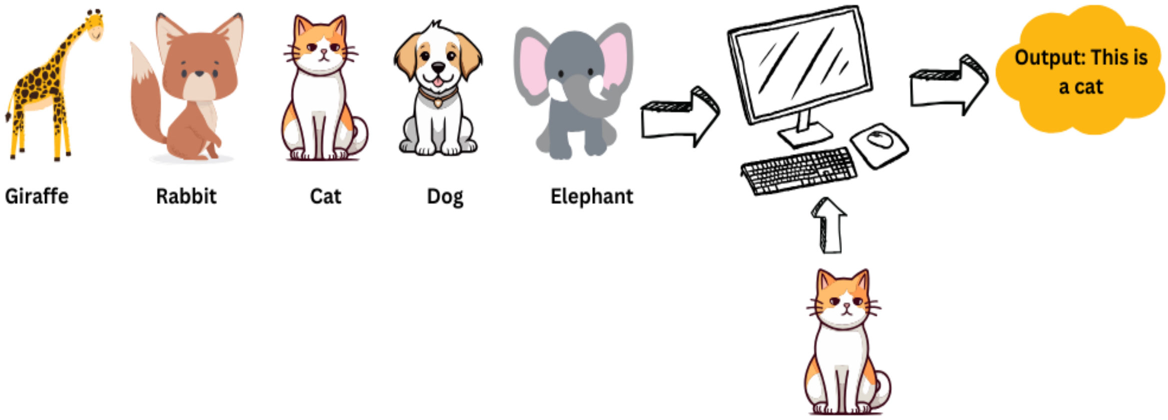


Figure 1.19: Computer learned

1.2.2. **Features of Machine Learning**

- Machine Learning systems learn from data.
- Data is analysed using machine learning to find various patterns.
- It can automatically improve by analysing historical data.
- Machine Learning systems learn from data, which means they improve their performance over time.
- Machine Learning finds applications in various fields, including healthcare, finance, marketing, robotics, and more, enhancing efficiency and decision-making.

1.3. Types of Machine Learning

Machine learning applications are grouped into four primary categories: Supervised Learning, Unsupervised Learning, Reinforcement learning, and Semi-Supervised Learning.

I. **Supervised Learning**

Supervised learning is a type of machine learning where the computer learns from known input data. It is like being a teacher for a computer. You provide it with examples and correct answers, so it learns to make predictions on its own. Imagine helping a friend learn to identify fruits: you show pictures of apples, oranges, and bananas, and your friend learns their names from those examples. As shown in figure 1.20.

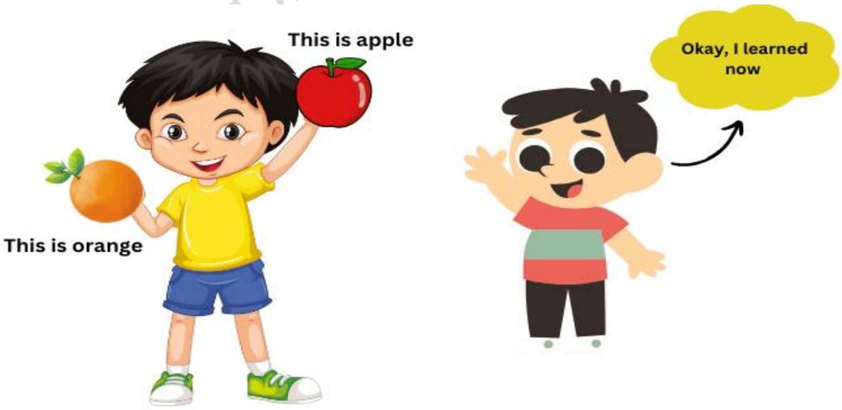


Figure 1.20: A boy telling his friend about fruits

Table 1.1: Consider the following information about people attending a seminar as an example. The data includes the age and gender of the persons, as well as an "Adult" or "Teenager" designation for each.

Gender	**Age**	**Label**
Male	13	Teenager
Male	21	Adult
Female	15	Teenager
Male	16	Teenager
Male	25	Adult
Female	23	Adult
Male	22	Adult

Types of Supervised Learning Algorithms

a) Random Forest

b) Support Vector Machine (SVM)

c) Logistic Regression

d) Naive Bayes Classifier

e) Linear Regression

f) Decision Trees

g) K-Nearest Neighbour (KNN)

a) **Random Forest**

Random Forest is a computer algorithm used in machine learning. It is a group of decision-making trees working together to make more accurate predictions. Each tree makes its guess, and then they vote to decide the final answer. Random Forest is like asking multiple friends for their opinions on a movie, and then it combines their ratings to decide if it is good or bad. As shown in Figure 1.21.

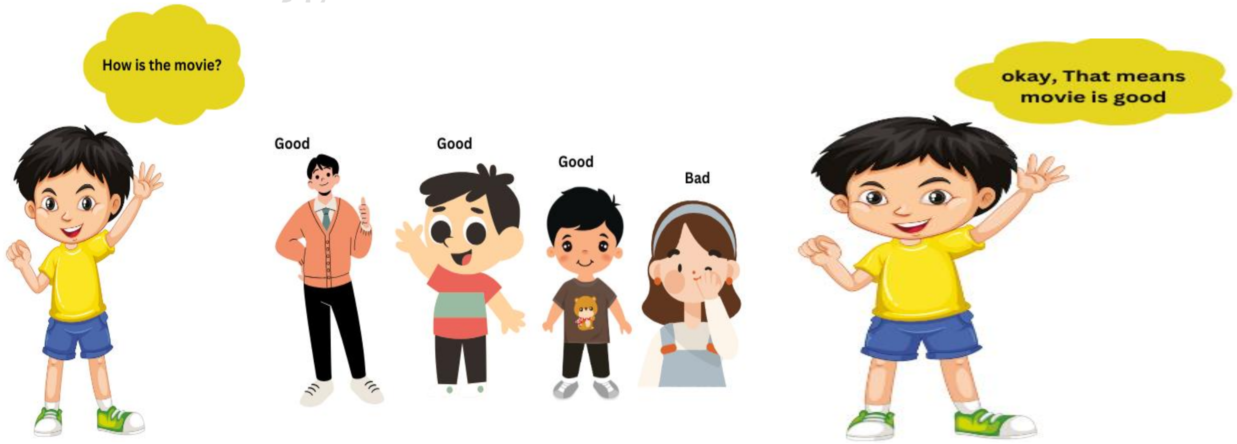


Figure 1.21: Example of Random Forest

b) **Support Vector Machine (SVM)**

Support Vector Machine (SVM) is a method of drawing a line to separate different types of things in a picture. This line is the decision boundary. For example, imagine you have pictures of apples and bananas. SVM helps find the best line that separates apples from bananas. It is like drawing a line so that apples are on one side and bananas are on the other. As illustrated in Figure 1.22.

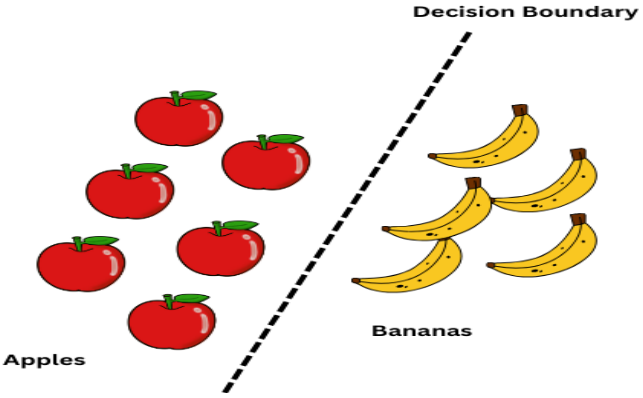


Figure 1.22: Example of Support Vector Machine

c) **Logistic Regression**

Logistic Regression is a machine learning method used for making predictions. The outcome is binary, which means it has only two possible values "yes" or "no." For example, suppose you want to predict whether a student will pass or fail an exam based on the number of hours they studied.

d) **Naive Bayes Classifier**

Naive Bayes Classifier is a simple and efficient machine learning model used for sorting things into categories. For example, you have a bunch of fruits and you want to know if they are apples, bananas, or oranges. Naive Bayes helps by just looking at their colour, shape, and size. If a fruit is red and round, it is likely an apple.

e) **Linear Regression**

Linear Regression is like drawing a straight line through points on a graph. For example, it can predict a house's price based on its size. As shown in Figure 1.23.

Figure 1.23: Example of Linear Regression

f) **Decision Trees**

Decision Trees are a way to make decisions or predictions by breaking down a complex problem into a sequence of simpler decisions. For example, as shown in Figure 1.24, If it is sunny outside and you are not getting late then wear shorts, otherwise wear pants.

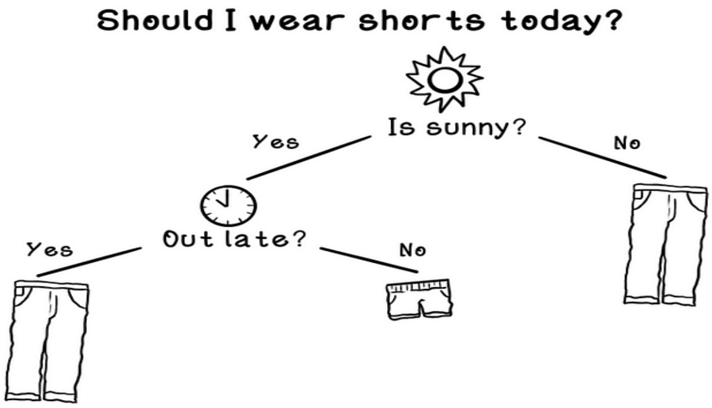


Figure 1.24: Decision Trees Example

g) **K-Nearest Neighbour (KNN)**

K-Nearest Neighbours (K-NN) is a type of machine learning algorithm that helps us make decisions based on our nearest neighbours decisions. In other words, it finds the most similar data points to make predictions. For example, ask your nearby friends for movie recommendations. If they liked a movie, you might enjoy it too.

Advantages of supervised learning

1. Supervised learning helps the model make predictions based on past experiences.
2. In supervised learning, we can easily identify which groups belong to which categories.
3. This type of learning is handy for solving real-life problems like fraud detection, stopping unwanted messages, etc.

Disadvantages of supervised learning

1. Supervised learning models might not be great at very hard tasks.
2. Supervised learning does not work well if the test data is not similar to the training data.
3. It takes a long time to train the supervised learning computer.

II. **Unsupervised Learning**

Unsupervised Learning is a type of machine learning where the algorithm tries to find patterns in data without any predefined labels. It aims to discover hidden relationships or groupings within the data on its own. For example, you have to sort a bag of colourful candies without labels, based on their colours and shapes. As illustrated in Figure 1.25.

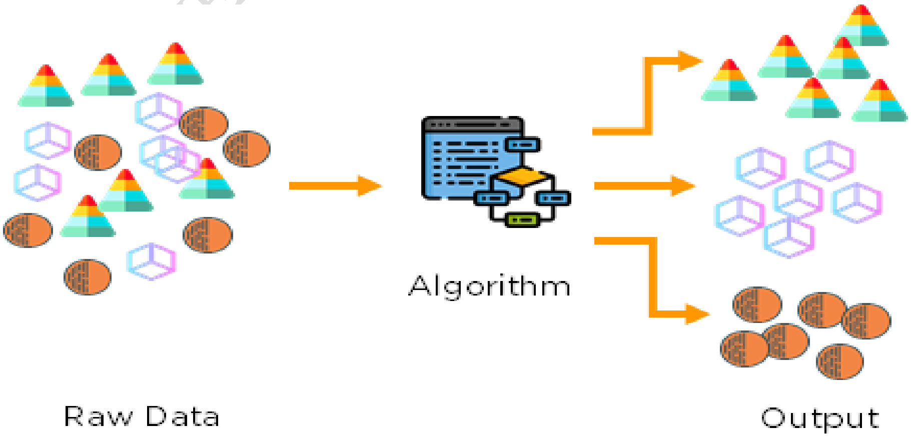


Figure 1.25: Unsupervised Learning

Example: Consider the following table 1.2 information about people attending a seminar as an example. The data includes the age and gender of the persons.

Table 1.2: Information about people attending a seminar.

Gender	Age
Male	13
Male	21
Female	15
Male	16
Male	25
Female	23
Male	22

Types of Unsupervised Learning: Unsupervised Learning will be categorized into two types:

1. **Clustering**
2. **Association**

1. Clustering

Clustering is a method of organizing data into groups or clusters based on similarities between data points. It is like putting similar things together. For example, the collection of mixed fruits is separated into different groups by analysing their size, shape, or colour. As shown in Figure 1.26.

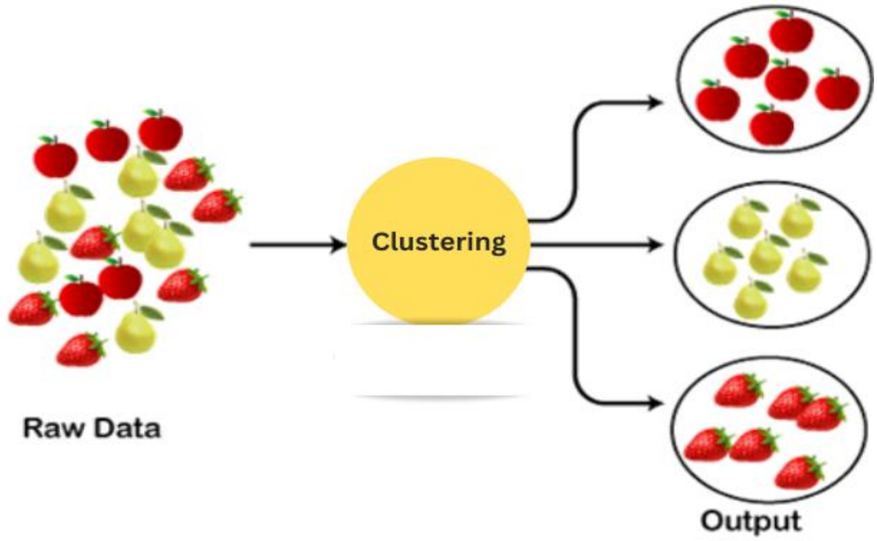


Figure 1.26: Clustering Example

Types of Clustering: Types of Clustering algorithms in Unsupervised Learning are:

a) Hierarchical Clustering

b) K-Means Clustering

c) Principal Component Analysis

d) Singular Value Decomposition

e) Independent Component Analysis

a) **Hierarchical Clustering**

Hierarchical Clustering is a way to organize data into a tree-like structure, where similar items are grouped, and these groups can be further divided into subgroups. The groups of animals and birds are divided into several groups, as shown in Figure 1.27.

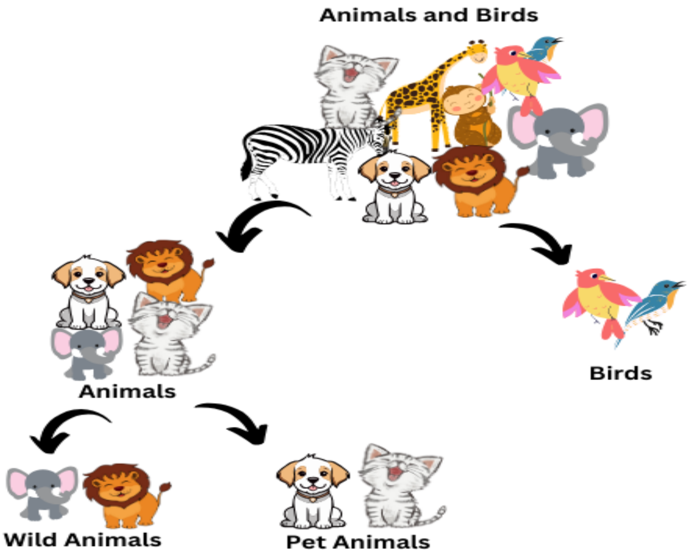


Figure 1.27: Example of Hierarchical Clustering

b) **K-means Clustering**

K-means Clustering is a method in clustering that aims to divide a dataset into groups, or "clusters," where data points in the same group are more similar to each other than to those in other groups. For example, sorting marbles by their colours. As shown in Figure 1.28.

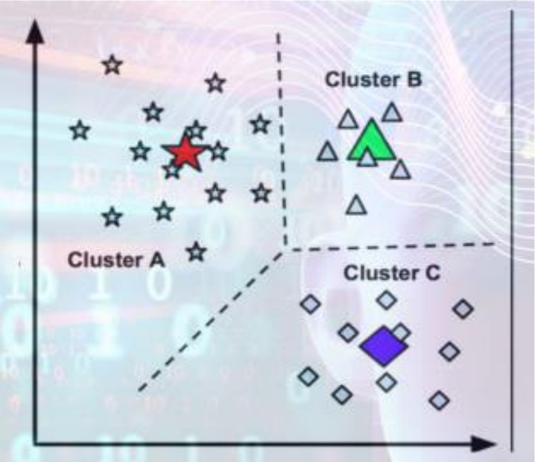


Figure 1.28: Example of K-means Clustering (source: https://dev.to/piyushbagani15)

c) **Principal Component Analysis**

Principal Component Analysis (PCA) is a method that simplifies complex data. It helps you identify the important points in your data. For example, you have a cake with many ingredients, but you want to know which ones give it the better taste. PCA helps you identify those key ingredients in your data.

d) **Singular Value Decomposition (SVD)**

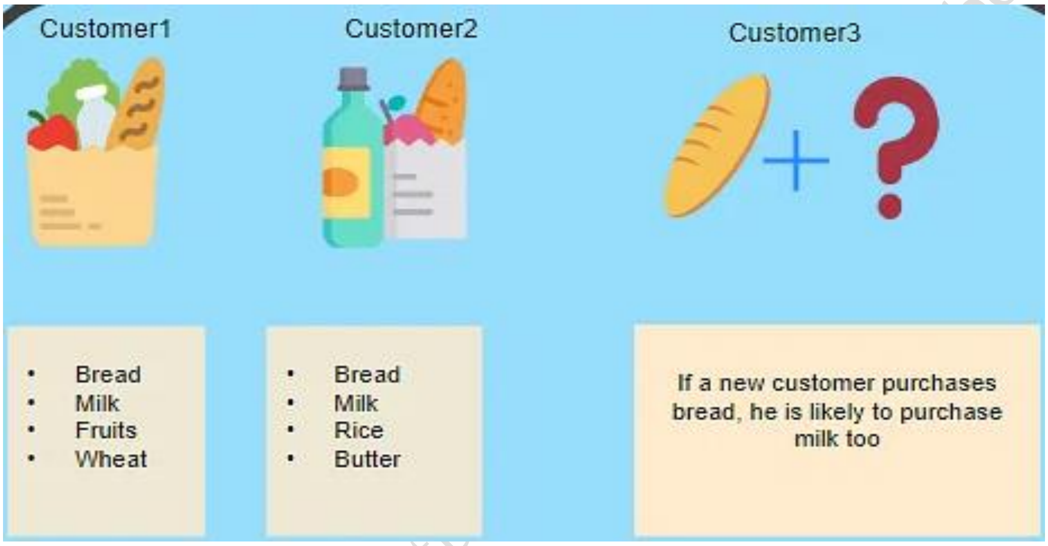
Singular Value Decomposition (SVD) is like taking a difficult mathematical problem and breaking it down into simpler pieces that are easier to work with. It is a way to understand and analyse data by separating it into its fundamental components.

e) **Independent Component Analysis**

Independent Component Analysis (ICA) is a method that helps us find hidden patterns or sources within a mixture of data. It separates complex data into its original, independent components. For example, ICA helps you to separate mixed candies and tells you the count of each colour in the bowl.

2. Association

Association in unsupervised learning refers to finding patterns, connections, or relationships among data items without any specific guidance or labels. For example, as customers shop, we can notice patterns. If one customer buys bread and milk, and another customer also buys bread, they might buy milk too. This technique is useful for recommending online shopping. As shown in Figure 1.29.



- Bread
- Milk
- Fruits
- Wheat

- Bread
- Milk
- Rice
- Butter

If a new customer purchases bread, he is likely to purchase milk too

Figure 1.29: Example of Association

Advantages of Unsupervised Learning

a) Compared to supervised Learning, unsupervised Learning is used for more complex problems since it lacks labelled input data.

b) Unsupervised Learning is preferred because unlabelled data is simpler to obtain than labelled data.

Disadvantages of Unsupervised Learning

a) Due to the lack of a comparable output, unsupervised Learning is more challenging than supervised Learning.

b) As the input data is not labelled and the algorithms do not know the precise output in advance, the outcome of the unsupervised learning method may be less accurate.

III. Semi-Supervised Learning

Semi-supervised learning is a type of machine learning where the model is trained on a dataset that contains both labelled and unlabelled data. In this, only a portion of the data is labelled, while the majority remains unlabelled. Semi-supervised learning is like having a few labelled pictures and using them to guess the names of other fruits in the unlabelled pictures. As illustrated in Figure 1.30.

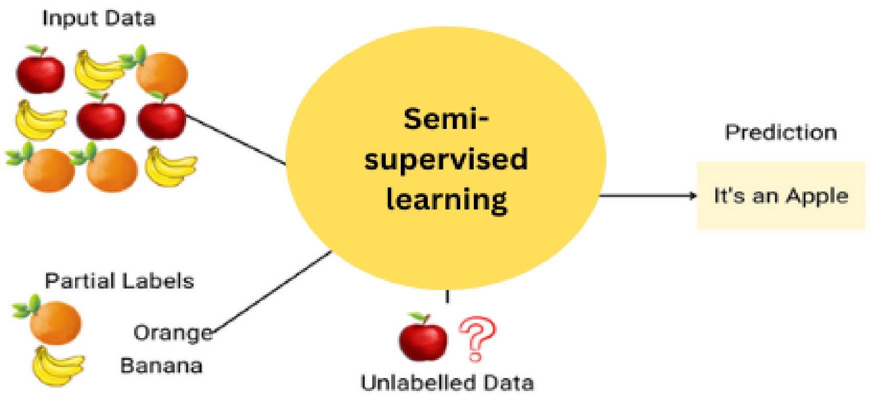


Figure 1.30. Semi-Supervised Learning

Working process of Semi-Supervised Learning

a) **Starting with Labelled Data:** You begin with a small amount of labelled data, where you know the right answers. For example, you have labelled images of cats and dogs. As illustrated in Figure 1.31.

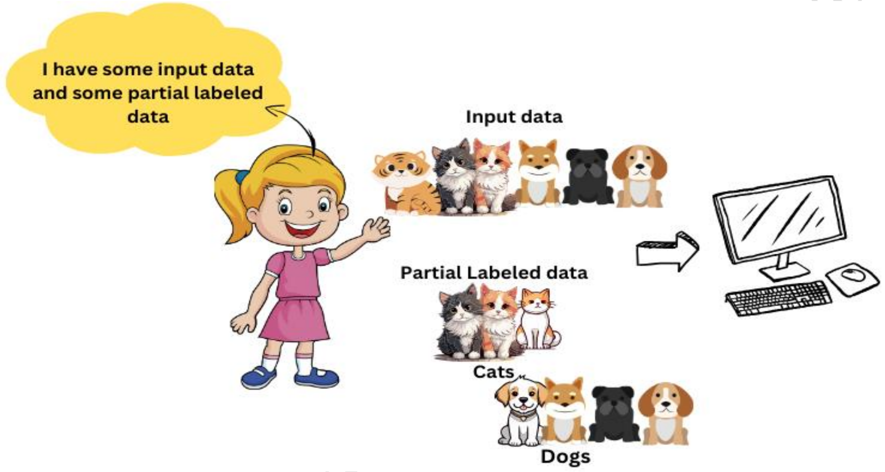


Figure 1.31: Input data of animals

b) **Learning from Labelled Data:** You use this labelled data to learn patterns, features, and distinctions between categories. For example, analysing the shape of ears, eyes, colour, etc. as illustrated in Figure 1.32.

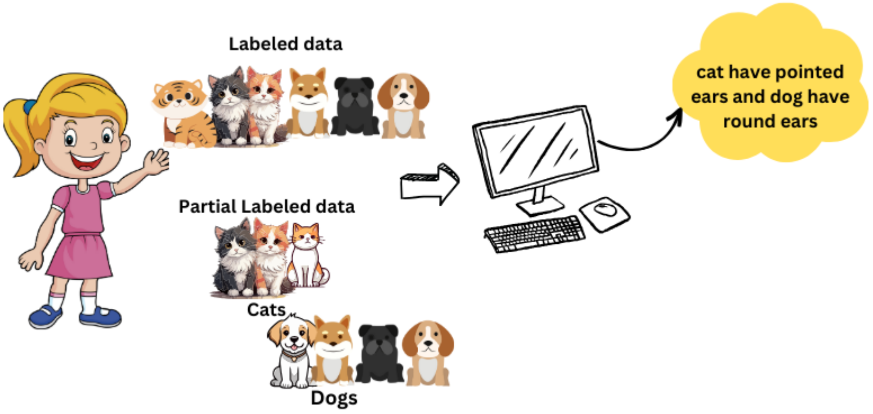


Figure 1.32: Learning from labelled data

c) **Applying to Unlabelled Data:** Now, you take what you've learned from the labelled data and apply it to a larger set of unlabelled data. You use your knowledge to make educated guesses about the categories of the unlabelled data. The figure illustrates 1.33.

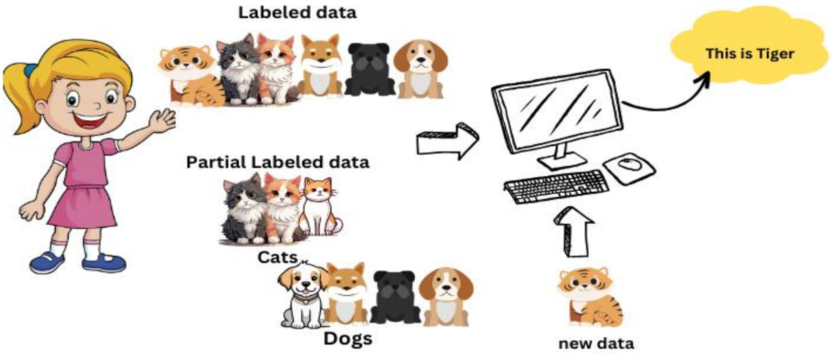


Figure 1.33: Applying to unlabelled data

d) **Improving the Model:** As you get more labelled data, you continue to refine and improve your model by adjusting the guesses you made for the unlabelled data.

Applications of Semi-Supervised Learning: Semi-supervised learning models are rapidly being used in real-world settings across a range of industries. Some of the most significant applications include the following:

- Text Classification
- Image Recognition
- Anomaly Detection
- Medical Diagnosis
- Recommendation Systems
- Speech Recognition
- Language Translation

IV. Self-supervised Learning

Self-supervised learning is a way for computers to learn from data without needing humans to label everything. In this, a computer can learn to recognize cats and dogs in pictures by looking at unlabelled images and figuring out the differences between them.

Applications of Self-supervised Learning: The main uses of self-supervised learning are as follows:

- Face detection
- Weather Forecasting
- 3D Rotation
- Stock Price Predictions
- Spam detection

V. Reinforcement Learning

Reinforcement Learning is a way for computers to learn to make decisions by trying different actions and learning from the outcomes. For example, you are training a dog. You teach tricks, and if it performs well, you give treats, as shown in Figure 1.34.

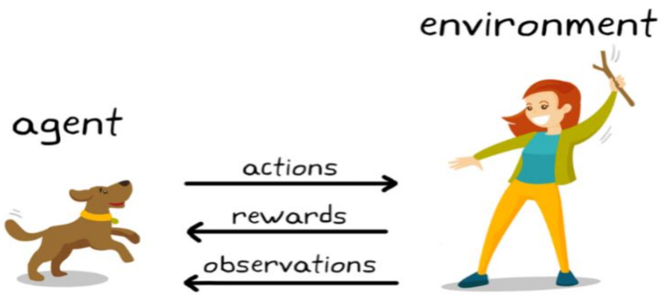


Figure 1.34: Reinforcement Learning

Applications of Reinforcement Learning: The following are a few examples of real-world uses for reinforcement learning:

- Virtual Assistants
- Robotics
- Autonomous Vehicles
- Game Playing
- Automation of the production industry
- Trading and Finance

1.4. Training and Testing in Machine Learning

In machine learning, "training" refers to the phase where a computer program or model learns from data. In training, the model learns from a dataset with examples and correct answers. It uses these examples to understand patterns, relationships, and features in the data.

During “testing”, the trained model is evaluated on new, unseen data to examine its performance and accuracy.

Training dataset: A training dataset is an important part of machine learning, It has a set of examples that a machine learning model uses to learn patterns and make predictions. It consists of input data paired with the correct output or label.

Test dataset: A test dataset is another important part of machine learning used to evaluate the performance of a trained model. It is a separate set of examples that the model has not seen during training. The primary purpose of a test dataset is to examine the performance of the model and generalize its learned patterns to new, unseen data.

1.4.1. Working on training and testing in Machine learning

Machine learning algorithms provide computers with the ability to deal with problems and make predictions based on previous observations or experiences. The model is tested using test data, it has been properly trained by using the training data.

The training and testing process consist of three steps:

1. **Input:** First, train the model by providing it with input data for training.
2. **Define:** In this step, pair the input data with the correct answers. The model learns patterns and features from this data.
3. **Test:** Finally, check the model's performance using a new dataset it has never seen before.

1.4.2. Difference between Training and Testing data

S.No.	Training Data	Testing Data
1.	Teaching the model to learn.	Check the model performance.
2.	Contains labelled data for learning.	Not having labelled data (unlabelled).
3.	Helps to set up patterns for the model.	Tests if the model's patterns work.
4.	Accuracy is not measured on training data.	Accuracy on test data is measured.

1.5. Applications of Machine Learning

Machine Learning has become a part of our daily lives. We use AI to help us in many ways without even realizing it. Let's explore some natural ways ML is used in our lives to understand its importance better.

- Recommendations to Find Products on E-commerce Sites.
- Suggestions to Identify routes and traffic. for example, Google Maps.
- Social Media Marketing.
- Detecting Email Spam.
- Search Engine Help.
- Medical Help from Machines.
- Predicting Cab Prices.
- Speech Recognition: Google Voice Assistant.
- Face Recognition in Photos.

1.6. Machine Learning Software Tools

There are many computer programs for Machine Learning that you can find in the market. Here are some of the most well-known ones among them.

i. **Scikit-learn**

Scikit-learn is a tool used for developing machine learning using Python. It is like a library for the Python programming language. The graphical representation is shown in Figure 1.35.

scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysi
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Figure 1.35: Scikit-learn

Features:

a) It can help with mining and analysing data.

b) It offers models and methods for various tasks like classification, regression, clustering, reducing dimensions, picking models, and getting data ready.

c) It also helps to test and train your models.

d) It is easy to use.

e) It works with most operating systems.

Benefits:

a) The user guide is easy to understand.

b) You can change the settings of a particular method while you are using it.

c) Scikit-Learn is an open-source tool.

d) It is cost free.

e) It uses Python libraries.

ii. **TensorFlow**

TensorFlow offers a JavaScript library that's useful for machine learning. It provides tools (called APIs) to create and train models. The graphical representation is shown in Figure 1.36.

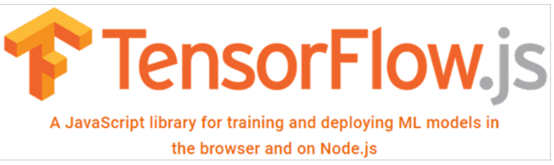

TensorFlow.js
A JavaScript library for training and deploying ML models in the browser and on Node.js

Figure 1.36: TensorFlow

Features:

a) It can help in creating and training models.

b) You can use TensorFlow.js to use models you've already made.

c) It is suitable for neural networks.

d) Provides a library for dataflow programming.

e) TensorFlow is a machine learning framework

Benefits:

a) You can use it by adding script tags or installing it through NPM.

b) It can even help estimate human poses.

c) It is an end-to-end open source platform.

d) It offers different workflows to train and develop models.

iii. **Weka**

These machine-learning tools are helpful for data mining. Graphical representation is shown in Figure 1.37.

WEKA
The University of Waikato
Machine Learning Group at the University of Waikato
Project
Software
Book
Courses
Publications
People
Related

Weka 3: Data Mining Software in Java

Figure 1.37: WEKA

Features:

a) It can prepare data.

b) It can classify data.

c) It can predict outcomes (regression).

d) It can group similar data points (clustering).

e) It can create visual representations.

f) It can find associations in data.

Benefits:

a) It offers online learning courses.

b) Its algorithms are easy to understand.

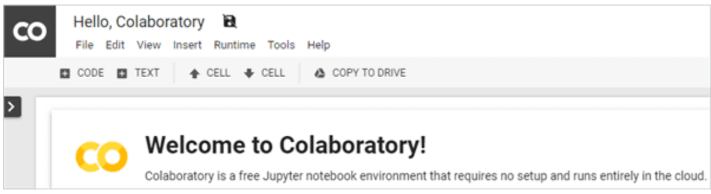
c) It is beneficial for students, too.

d) It can work on different platforms.

e) It is easy to use.

iv. **Colab**

Google Colab is an online platform that works with Python. It helps create machine learning programs using PyTorch, Keras, TensorFlow, and OpenCV libraries. The graphical representation is shown in Figure 1.38.


CO
Hello, Colaboratory
File Edit View Insert Runtime Tools Help
CODE TEXT CELL CELL COPY TO DRIVE
CO

Welcome to Colaboratory!

Colaboratory is a free Jupyter notebook environment that requires no setup and runs entirely in the cloud.

Figure 1.38: Colab

Features:

a) It supports learning about machine learning.

b) It helps in conducting machine learning research.

c) It supports libraries of PyTorch, Keras, TensorFlow, and OpenCV.

d) It is open-source.

e) It is easy to use.

Benefits:

a) It is accessible through your Google Drive.

b) Colab is a free cloud service.

c) It enhances the performance of different ML tools.

d) It supports Pytorch, TensorFlow and Keras.io.

e) It is cost free.

v. **Keras.io**

Keras is a tool for creating neural networks. It is excellent for fast research and is written in Python. Graphical representation is shown in Figure 1.39.

Figure 1.39: Keras

Features:

a) It makes quick prototyping easy.

b) It supports convolution networks.

c) It helps in recurrent networks.

d) It works by combining two networks.

e) It gives results on both the CPU and GPU.

Benefits:

a) Easy for users to work with.

b) It can be customized.

c) It can be extended as needed.

d) Keras uses a Python library to help its users.

e) It is user-friendly.

Comparison Chart

Table 1.3: A list of machine learning tools comparison is enlisted below:

Tools	**Platform**	**Cost**	**Written in Language**	**Algorithms or Features**
Scikit Learn	Linux, Mac OS, Windows	Free.	Python, Cython, C, C++	Classification, Regression Clustering Dimensionality reduction.
TensorFlow	Linux, Mac OS, Windows	Free	Python, C++, CUDA	Offers a programming library for data flow tasks.
Weka	Linux, Mac OS, Windows	Free	Java	Classification, Regression Clustering Association rules mining
Colab	Cloud Service	Free	-	Supports libraries of PyTorch, Keras, TensorFlow, and OpenCV
Keras.io	Cross-platform	Free	Python	API for neural networks

SUMMARY

- In Machine Learning (ML), which is a part of Artificial Intelligence (AI), machines learn using algorithms and their own experience to give predictions for a specific problem.
- The process of Machine Learning (ML) begins by collecting training data and giving it to the ML algorithm. The accuracy of the predictions made by the algorithm depends on this data. After that, the algorithm is tested with different data to check if its predictions are correct. If the predicted results match the expected ones, the algorithm is good at predicting. Now, new data can be input into the algorithm for accurate predictions.
- In Supervised Learning, the training data comes with labels. Usually, each piece of training data is linked to a specific group or category it belongs to.
- There are a few classification algorithms, like Naïve Bayes, Decision Trees, and Support Vector Machine.
- In Unsupervised learning, the training data doesn't have labels. Instead of using labels, patterns are used to sort the input data.

- In Clustering, we look for patterns or groups in a collection of data that don't have categories. Clustering algorithms analyse the data and group them based on their patterns or similarities.
- There are different types of clustering methods, such as Hierarchical Clustering, K-Means Clustering, K-Nearest Neighbour, Clustering, Principal Component Analysis, Singular Value Decomposition, and independent Component Analysis.
- Semi-supervised learning uses both labelled and unlabelled data for training. In this type of learning, some data has labels while some don't. The labels from labelled data help the algorithm understand the data, and then it learns from the unlabelled data to give them appropriate labels. This way, the algorithm learns and predicts outcomes from new, unknown data.
- In reinforcement learning, we find data with better rewards through trial and error. This type of learning has three main parts: the agent, the actions it does, and the environment it interacts with.

Check Your Progress

A. **Multiple-Choice Questions**

1. Identify the learning method in which labelled training data is used. (a) Supervised learning (b) Unsupervised learning (c) Reinforcement learning (d) Semi-supervised learning
2. Machine learning is a subset of: (a) Deep Learning (b) Artificial Intelligence (c) Reinforcement Learning (d) None of the above
3. What are the usual types of problems in machine learning? (a) Association (b) Classification (c) Clustering (d) All of the above
4. Identify the applications of ML. (a) Face recognition (b) Social Media Marketing (c) Email Spam identification (d) All of the above
5. The term machine learning was used in which year? (a) 1958 (b) 1959 (c) 1960 (d) 1961
6. Which is the machine learning algorithm that can be used with unlabelled data. (a) Clustering algorithms (b) Classification algorithms (c) All of the above (d) None of the above
7. Which of the following algorithms are used in Machine learning? (a) Support Vector Machines (b) K-Nearest Neighbours (c) Naive Bayes (d) All of the above
8. The unsupervised learning problems can be grouped as _______ (a) Association (b) Clustering (c) Both A and B (d) None of the above
9. The term machine learning was coined by __________ (a) Guido van Rossum (b) James Gosling (c) Arthur Samuel (d) None of the above
10. Which machine learning models learn to make choices by getting rewards and feedback for what they do? (a) Reinforcement learning (c) Supervised learning (c) Unsupervised learning (d) All of the above

B. **Fill in the blanks.**

1. Supervised learning uses ________ data for training.

2. Unsupervised learning involves finding ________ in data.
3. Machine learning is a subset of ________.
4. In reinforcement learning, models learn by receiving ________ for their actions.
5. Machine learning allows computers to learn from ________.
6. The three main types of machine learning are supervised, unsupervised, and ________.
7. Semi-supervised learning combines elements of both ________ and ________ learning.
8. Clustering is a common task in ________ learning, where similar data points are grouped.
9. In reinforcement learning, agents receive ________ as feedback for their actions.
10. In semi-supervised learning, a small portion of the data is ________, while the rest is ________.

C. **True and False**

1. Machine learning is a branch of artificial intelligence focusing on developing systems that can learn from data.
2. Labelled training data is used to train the machine learning model in supervised learning.
3. Clustering is a common task in supervised learning where data points are grouped based on similarities.
4. Unsupervised learning is used when the training data is labelled with the correct answers.
5. Semi-supervised learning uses both labelled and unlabelled data to train the model.
6. In reinforcement learning, agents learn by interacting with the environment and receiving rewards.
7. Decision trees are a standard algorithm used in unsupervised learning.
8. K-means clustering is an example of unsupervised learning.
9. Random Forest is an example of a deep-learning algorithm.
10. Support Vector Machines (SVM) is a classification algorithm used in supervised learning.

D. **Short Answer Questions**

1. Explain Supervised Learning with examples.
2. What is Clustering?
3. Suggest some real-life applications of Machine Learning.
4. Explain the Random Forest algorithm.
5. List any three machine learning software tools.
6. What is reinforcement Learning? Give some real-life examples.
7. Explain semi-supervised learning. Suggest some real-life applications.
8. What are the advantages and disadvantages of unsupervised learning?
9. What are the types of supervised learning algorithms?
10. Explain the phases of Machine Learning.

Session 2. Data Annotation For Business

In a quiet village, curious Arjun helped his family run a grocery store. They wanted to understand their customers better. One day, Arjun learned about 'Data Annotation,' like labels on his school notebooks. Arjun took pictures of veggies in their store, labelled them, and used a computer to understand which one's customers liked. This helped his family stock what people wanted and reduce food waste. Arjun was happy with his idea as shown in Figure 2.1.

Figure 2.1: Happy Arjun at the grocery shop (Source: pixta.jp and istockphoto)

This chapter will cover the use of Data Annotation in business. In which you will be able to understand the applications of data annotation, uses of data annotation, and importance of data annotation in business application

2.1. Data Annotation

Data Annotation adds labels or tags to data and helps computers to understand it. This helps to train machine learning models for better understanding and accurate predictions. For example, image recognition, labels objects like buildings, persons, or trees. As illustrated in Figure 2.2.

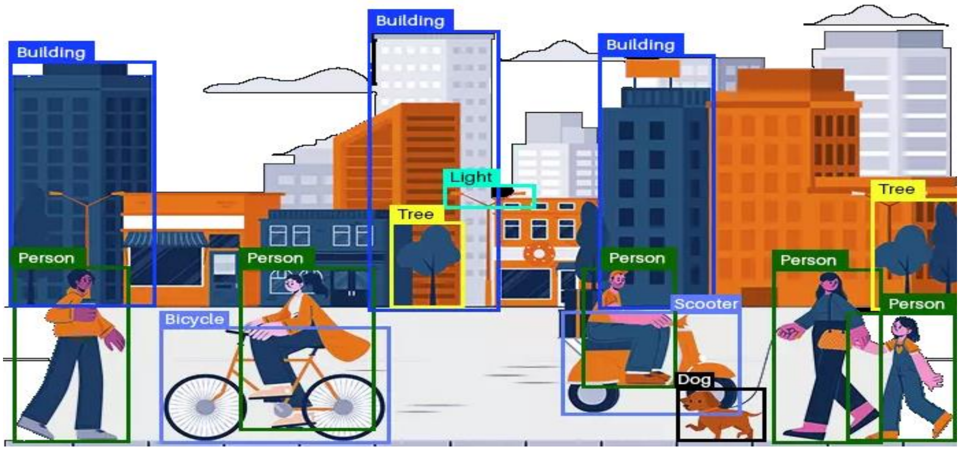


Figure 2.2: data annotation (source: https://www.anolytics.ai/)

2.1.1. Applications of data annotation

Some examples of Data Annotation are used in various business applications:

- E-Commerce
- Healthcare

- Automotive Industry
- Agriculture
- Real Estate, etc.

2.2. Data Annotation in AI

Data annotation is important for AI because it changes raw data into a form that AI algorithms can understand and learn from. It is useful in AI.

2.2.1. Uses of Data Annotation in AI

Data annotation is very useful in AI. Here are some uses of data annotation in AI:

- **Training Machine Learning Models:** AI algorithms in supervised learning, require labelled data to learn patterns and make correct predictions. Data annotation provides the labelled examples needed to train these AI models effectively.
- **Improving Accuracy:** Well-annotated data allows AI models to learn from various examples, to perform higher accuracy in tasks like image recognition, speech understanding, and text analysis.
- **Enabling Unsupervised Learning:** Data annotation can involve grouping similar data in unsupervised learning. This labelled data helps algorithms to find hidden patterns, clusters, or relationships.

2.2.2. Natural Language Processing (NLP)

Natural Language Processing (NLP) is a field of artificial intelligence (AI) that focuses on the interaction between computers and natural language. It allows computers to read, understand, and generate human language. It is like teaching computers to understand and work with human language, just like humans do.

2.2.3. Computer Vision

Computer vision is like giving computers the ability to see and understand the world through visual data, just like humans do with their eyes. It is a field of artificial intelligence (AI) that focuses on teaching computers to interpret and make sense of images and videos. Computer vision includes tasks like Image Classification, Object Detection, Semantic Segmentation, and Instance Segmentation.

2.2.4. Computer vision Tasks

The computer vision includes various tasks as follows:

- Image Classification
- Object Detection
- Semantic Segmentation
- Instance Segmentation

2.2.5. Importance of Data Annotation in Business Applications

Data annotation is very important for businesses because it allows them to train machine learning models that can handle tasks and help them make better decisions.

Data Annotation is important in business applications for several key reasons:

- Data Annotation ensures the quality of AI processes.
- Data Annotation allows AI to understand customer preferences and behaviours.
- Data Annotation reduces the risk of fraud activities.

- Data Annotation enables AI to automate tasks.
- Data Annotation ensures the scalability of AI applications.

2.2.6. Benefits of Data Annotation

Data annotation is helpful by using supervised learning methods, like training a machine learning model to make accurate predictions. Some benefits of data annotation:

- Data annotation can improve the accuracy of machine learning models.
- Data annotation speeds up the machine learning model development process.
- Good quality data annotation saves time and money by minimizing rework.
- Data annotation helps machine learning models manage big datasets.

Summary

- Data annotation is a method to mark content that machines can understand through computer vision or training with machine learning (ML).
- Data annotation uses natural language processing (NLP) and comes in different formats like text, pictures, and videos.
- Data annotation involves recognizing information in different types like text, video, or images.
- Labelled datasets are needed for supervised machine learning to help computers understand input data.
- Properly labelled data is essential for training computer vision-based machine learning models too. Various methods can be used for data annotation to create datasets for specific needs.

Check Your Progress

A. **Multiple Choice Questions**

1. What is data annotation? (a) Organizing data in spreadsheets (b) Labeling data with meaningful information (c) Deleting unnecessary data (d) Encrypting data for security
2. Why is data annotation important for business applications? (a) It saves electricity (b) It boosts employee morale (c) It improves the accuracy of AI models (d) It makes computers faster
3. How does accurate data annotation affect AI system performance? (a) It doesn't affect performance (b) It can make AI models worse (c) It improves the AI model's accuracy and reliability (d) It only affects the AI model's speed
4. Which of the following is not a task included in the field of computer vision? (a) Image Classification (b) Object Detection (c) Semantic Segmentation (d) Speech Recognition
5. What is the primary goal of computer vision in artificial intelligence? (a) Teaching computers to understand human language (b) Enhancing data security in visual data (c) Teaching computers to see and interpret images and videos (d) Training machine learning models for natural language processing

6. In which field of artificial intelligence does Natural Language Processing (NLP) focus? (a) Image recognition (b) Speech understanding (c) Interpreting visual data (d) Interaction between computers and natural language
7. What is the primary function of well-annotated data in AI? (a) Teaching AI models how to label data (b) Enabling unsupervised learning (c) Improving accuracy in image recognition (d) Enhancing data security
8. What is the primary purpose of data annotation in the context of business? (a) Enhancing data security (b) Adding labels or tags to data (c) Improving data visualization (d) Automating data analysis
9. How does data annotation benefit machine learning models? (a) It speeds up data processing (b) It helps computers understand data (c) It replaces the need for machine learning (d) It prevents data from being used for AI
10. Does data annotation impact business decision-making? (a) No, it is just for technical purposes (b) Yes, it enhances decision-making with more accurate insights (c) It only impacts decisions made by robots (d) It makes decision-making slower

B. **Fill in the blanks**

1. _____________is the process of adding labels or information to data to make it understandable for machines.
2. Data Annotation is the process of adding ____________ to raw data to make it understandable for machines.
3. Annotated data is beneficial by using ___________ methods.
4. ________________ Technology helps computers understand what people say.
5. Annotated data ensures the ________ and ________ of AI-driven processes.
6. Supervised learning may require __________ data to learn patterns and make accurate predictions.
7. Data Annotation is the process of adding meaningful labels, tags, or __________ to raw data.
8. Computer vision-based machine learning model can't be taught without properly ________ data.
9. ___________ enable AI to automate tasks traditionally done by humans.
10. Annotated data enables AI to understand customers __________ and _________.

C. **State whether True or False**

1. Data annotation involves Labeling data to make it understandable for computers.
2. Data annotation is not essential for business applications.
3. Accurate data annotation can lead to better AI system performance.
4. Data annotation only affects AI-related applications, not other aspects of business.
5. Data annotation is unnecessary for training AI models, as they can learn from raw data.
6. Data annotation plays a role in improving customer experiences in various industries.
7. Data annotation is limited to the technology sector and doesn't impact other industries.
8. Data annotation is a one-time process and doesn't have ongoing benefits.

9. Data annotation can help optimize business operations and automate repetitive tasks.
10. Data annotation's impact on business decision-making is minimal.

D. **Short answer questions**

1. What is data annotation?
2. What are the applications of data annotation?
3. Explain the uses of data annotation in business.
4. How does unsupervised learning help in data annotation?
5. Explain the importance of data annotation in business applications.
6. Explain computer vision.
7. How does Automation help in AI?
8. What are the benefits of data annotation?
9. How does Annotated data enable AI to understand customer preferences and behaviours?
10. How Natural Language Processing (NLP) is helpful in data annotation?

Session 3. Use Cases And Applications of Data Annotation

Rani was an artist in a city, who learned about Data Annotation. It helped her add extra information to her artwork using technology. People loved her art, and soon her art gallery displayed it. Other artists in the city also began using Data Annotation to make their art more interesting.

Figure 3.1: Rani using data annotation applications

In this chapter, you will understand the use cases of data annotation and the various applications of data annotation.

3.1. Use cases of Data Annotation

The various use cases of data annotation in various fields are as follows:

Search Engine Efficiency

Search engines use a lot of data to improve search results. Such as a person's search history, age, and location. Annotations help the search engine show the best results for each person. As illustrated in Figure 3.2.

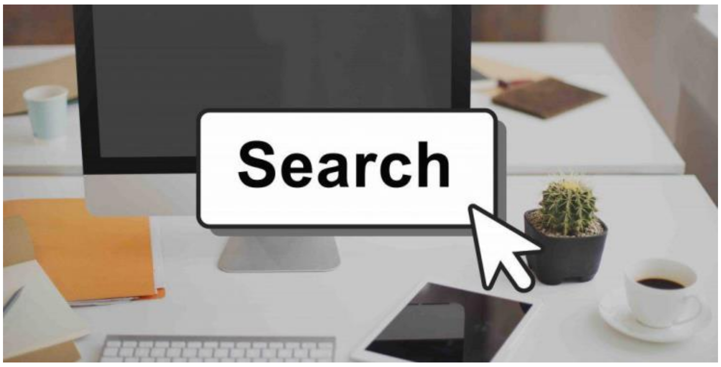


Figure 3.2: Search Engine (source: https://tarjama.com/)

Development of Facial recognition software

Machines can learn and identify different facial features. They put dots on faces to show things like the face's length, eye shape, nose, and more. These dots are saved in a computer memory. This helps if the same faces are seen again. Face unlock features on phones and laptops are using this technology. As illustrated in Figure 3.3 below.

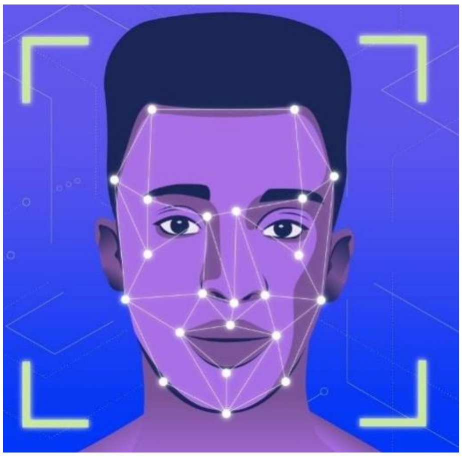


Figure 3.3: Facial recognition (Source: https://www.rootstrap.com/)

Production of data for Automated vehicles

Automated vehicles use pictures to learn. These vehicles must understand road signs, stay in their lanes, and go safely around other vehicles. This helps them detect lanes on the road, see things around them, and understand objects in front of them. It is like putting boxes, shapes, and labels on the pictures to help the cars understand better. As illustrated in Figure 3.4.

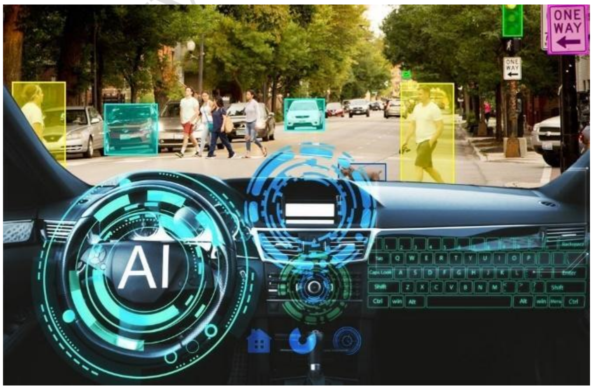


Figure 3.4: Automated vehicles (Source: https://www.einfochips.com/)

Medical breakthroughs

Data annotation is also useful in the medical field of pathology and neurology. It helps to find patterns for faster and more accurate diagnoses. As illustrated in Figure 3.5.

Figure 3.5: Medical (Source: https://www.medicalnewstoday.com/)

3.2. Data Annotation Applications

Unmanned Vehicles

Supervised Machine Learning is essential for self-driving vehicles like cars and trucks driving alone. Also, robots that deliver packages use it. These machines need lots of specially labelled data to work well. This helps them stay in their lanes, see people walking, and notice traffic lights.

Manufacturing

In the year 2035, experts think that AI (a smart computer technology) could make 16 different types of jobs, like in factories, much better. AI is changing how factories work. It is helping them use machines that put things together by themselves. It also helps find mistakes and keep workers safe. As illustrated in Figure 3.6.

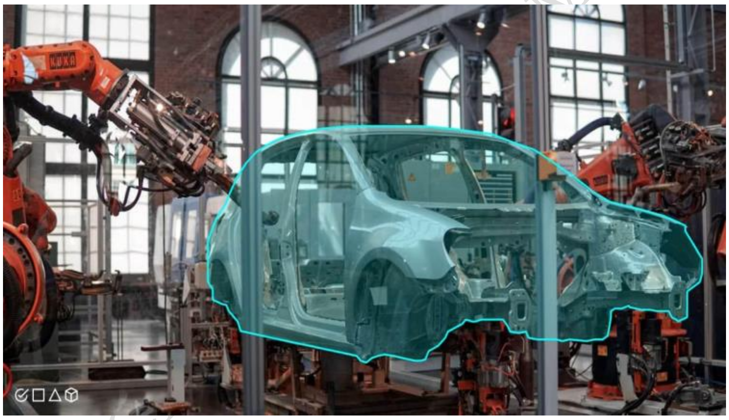


Figure 3.6: (source: https://labelyourdata.com/)

Healthcare

AI is useful in the healthcare field. Lots of labelled data can bring significant changes to this industry. It can help with important things like studying genes, making new medicines, and finding problems in X-ray and MRI images. As illustrated in Figure 3.7.

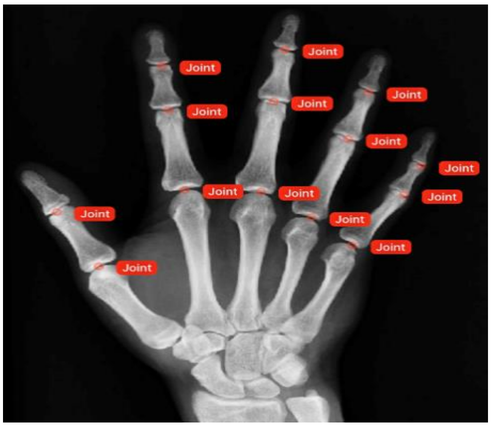


Figure 3.7: Healthcare (source: https://www.v7labs.com/)

Insurance and Banking

In insurance, data annotation helps computers learn to find fraud in claims, decide on risks for customers, and understand customer feelings from their messages. In banking, data annotation teaches computers to predict if someone can pay back a loan, catch bad transactions, and make helpful chatbots.

Agriculture

In agriculture, Data Annotation is helpful. It helps farmers understand their crops and fields better. Data Annotation also helps in identifying different types of plants and pests. This information helps them make better decisions about caring for their crops and getting more food from their fields. As illustrated in Figure 3.8.

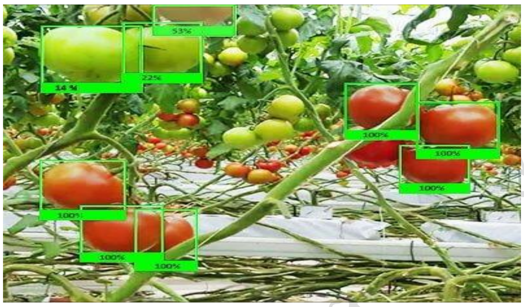


Figure 3.8: Agriculture (source: https://www.anolytics.ai/)

Retail

Data Annotation helps a lot in the field of retail. It helps to put products in the correct categories so you can find them online quickly. It also allows stores to suggest things you might like to buy. Data Annotation helps stores understand what people feel about products by looking at their reviews. As illustrated in Figure 3.9.

Figure 3.9: Retail (source: https://mindy-support.com/)

Animal management

Using drones to care for your animals and plants on a farm by using "image annotation" technology. This means you can add special notes or labels to pictures taken by the drones. For example, if you have animals, these notes can tell you important details about them. As illustrated in Figure 3.10.

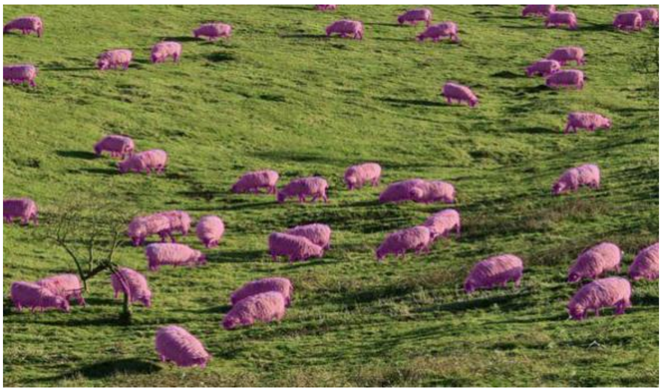


Figure 3.10: Animal management (source: https://keymakr.com/)

Geosensing

Data annotation helps us to understand the types of soil available in different places. It can also detect water bodies, buildings, roads, and other land. As illustrated in Figure 3.11.

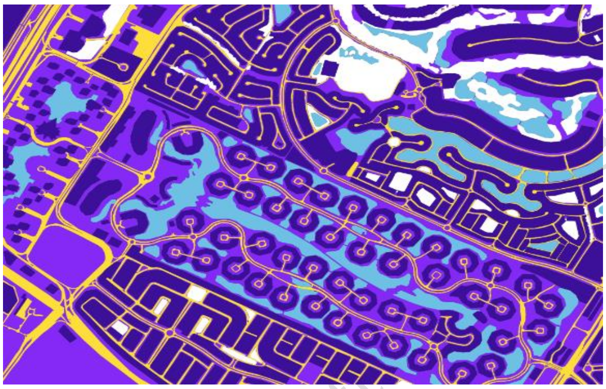


Figure 3.11: Geosensing (source: https://humansintheloop.org/)

Crop health surveillance

Computer Vision helps us to identify plant diseases caused by bugs or fungi. Carefully adding the correct labels of bugs to pictures, helps the computers to learn and understand the look of these harmful things. As illustrated in Figure 3.12.

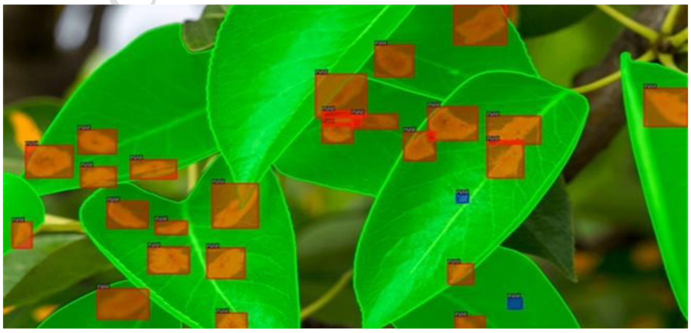


Figure 3.12: Crop health surveillance (source: https://www.linkedin.com/)

Surveillance and Protection

Computers use labelled data to identify and sort things like objects, people, and other details in pictures and videos. For this purpose, image annotation is used. This is important because computers cannot understand pictures. As illustrated in Figure 3.13.

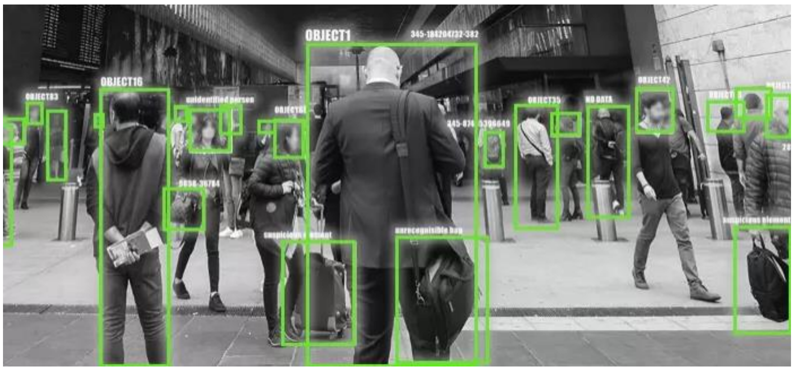


Figure 3.13: Surveillance and Protection (source: https://www.cogitotech.com/)

Image annotation can help in Surveillance (cameras) to do many things. For example:

- **Pedestrian Tracking:** It can follow people's movements and keep an eye on them.
- **Facial Recognition:** It can recognize people's faces, which helps catch thieves or people who shouldn't be there.
- **Traffic Motion:** It can track vehicles to make sure traffic flows smoothly.
- **Thermal Vision and Night Vision:** It can even "see" in the dark using special technology.
- **Crowd Detection:** It can be noticed as many people are in one place.

Robotics

Businesses use robots because they are cheap, work fast, and can do many things without needing people's help. These robots are like machines can do things like people do. People teach these robots using a lot of pictures with unique labels. These labels act as instructions for the robots to understand. Robots can be used in agriculture, teaching, shops, etc. As illustrated in Figure 3.14.

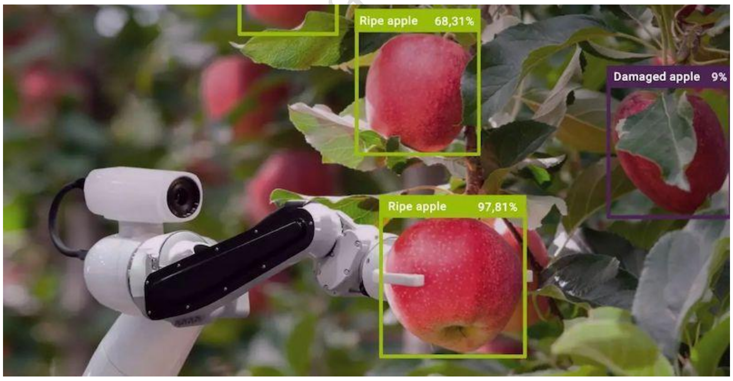


Figure 3.14: Robotics (source: https://roboticsandautomationnews.com/)

Analytics in sports

In sports, technology uses 'data Labeling' and 'picture annotation' to add notes and labels to pictures. This helps experts understand the game better. Labelled pictures are useful for sports, computers use these labelled pictures to track player movements accurately. As illustrated in Figure 3.14.

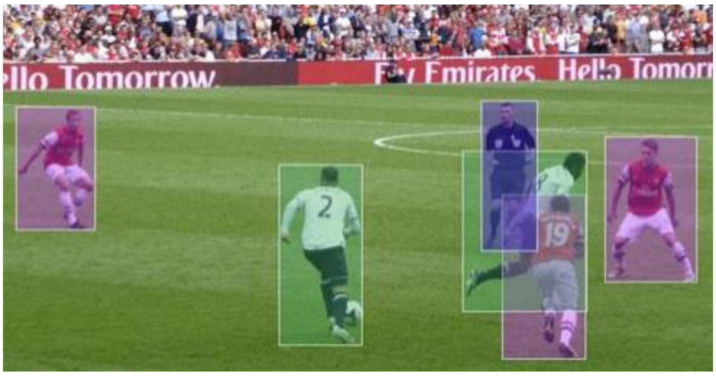


Figure 3.14: Analytics in sports (source: https://www.superannotate.com/)

Fashion Industry

Choosing the right outfit is easier with smart technology. It uses 'data categorization' and 'image annotation' to understand stylish clothes. You can even get fashion advice and find out the trendy. The computer can tell you where to find it just by looking at a picture of something you want. As illustrated in Figure 3.15.

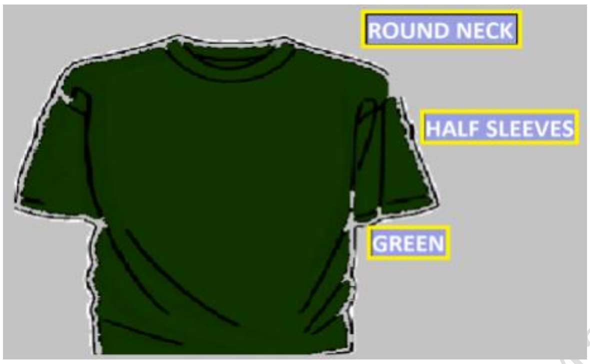


Figure 3.15: Fashion Industry (source: https://www.precisebposolution.com/)

Shop automation

Image annotation improves online shopping. It adds extra features to stores for a better customer experience, which is vital for retail companies. As illustrated in Figure 3.16.

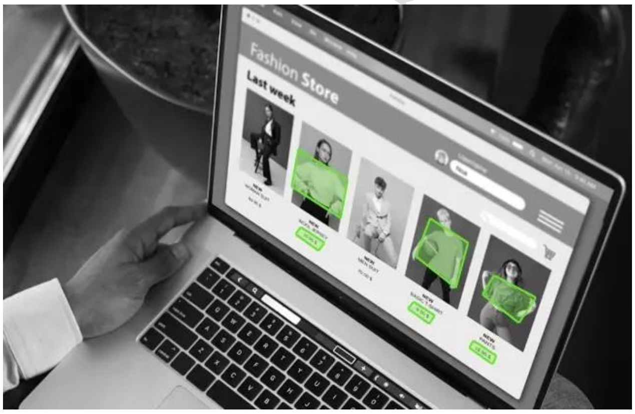


Figure 3.16: Shop Automation (source: https://www.cogitotech.com/)

Check Your Progress

A. **Multiple Choice Questions (MCQ)**

1. Data Annotation Applications involve which of the following tasks? (a) Baking cookies (b) Sequencing, Classification, and Mapping (c) Building rockets (d) Playing sports
2. What area does "Unmanned Vehicles" fall under as a Data Annotation Application? (a) Cooking (b) Entertainment (c) Automotive (d) Gardening
3. Data Annotation can be used to enhance Surveillance and Protection in which field? (a) Space exploration (b) Fashion industry (c) Security and safety (d) Food industry
4. Which industry is NOT listed as an application of Data Annotation? (a) Agriculture (b) Education (c) Healthcare (d) Retail
5. What does "Geosensing" refer to in Data Annotation Applications? (a) Cooking different cuisines (b) Sensing emotions (c) Sensing geographical information (d) Sensing cosmic events
6. Which of the following is NOT a real-life application of Data Annotation? (a) Animal management (b) Virtual reality gaming (c) Surveillance and Protection (d) Geosensing
7. What does "Healthcare" involve as a Data Annotation Application? (a) Creating fashion designs (b) Developing social media platforms (c) Improving medical treatments (d) Designing new buildings
8. How can Data Annotation impact "Retail"? (a) By producing new movies (b) By improving online shopping experiences (c) By designing video games (d) By organizing music concerts
9. In which field can Data Annotation be applied for "Crop health surveillance"? (a) Marine biology (b) Movie production (c) Farming and agriculture (d) Astronomy
10. Which industry can benefit from Data Annotation regarding "Insurance and Banking"? (a) Music Industry (b) Travel and tourism (c) Finance and insurance (d) Circus entertainment

B. **Fill in the blanks.**

1. Data Annotation Applications involve tasks like Sequencing, Classification, __________, Mapping, and more.
2. The goal of Data Annotation in Healthcare is to contribute to medical breakthroughs and improve ____________.
3. One of the Data Annotation Applications is the development of Facial _______________ software.
4. The retail industry can make use of Data Annotation to optimize shopping experiences and improve _______________.
5. Data Annotation plays a role in Crop Health Surveillance to monitor and improve the health of ______________.
6. Geosensing involves collecting and annotating data related to ______________ locations and attributes.
7. Data Annotation plays a role in improving the efficiency of __________ engines by refining search results.

8. Data Annotation is essential for creating detailed ________ maps and geographic information.
9. Classification tasks involve labeling data into distinct ______________ or categories.
10. In the Healthcare sector, Data Annotation assists in diagnosing and treating __________ conditions.

C. **State whether true or false**

1. Data Annotation is labeling data to make it understandable for machines.
2. Data Annotation can enhance the accuracy of facial recognition software.
3. Data Annotation has no significance in the Healthcare industry.
4. Data Annotation is not useful for improving the efficiency of search engines.
5. Data Annotation is crucial for assessing risks in the Insurance and Banking sectors.
6. Data Annotation has no role in improving product quality in the Manufacturing sector.
7. Retailers can benefit from Data Annotation to better understand customer preferences.
8. Data Annotation is only used for text data and has no application in image or video data.
9. Data Annotation is not necessary for developing autonomous vehicles.
10. Data Annotation is only used for creating maps and geosensing applications.

D. **Short answer questions**

1. What are the different use cases of Data Annotation?
2. How machine learning uses Data Annotation Application?
3. How does data annotation help in the healthcare sector?
4. How does Data Annotation contribute to the development of facial recognition software?
5. What industries benefit from Data Annotation in terms of insurance and banking?
6. How does Data Annotation contribute to improving search engine efficiency?
7. In what ways can Data Annotation be utilized in the agriculture sector?
8. What role does Data Annotation play in the field of retail?
9. How does Data Annotation contribute to geosensing and mapping applications?
10. What are the various categories of Data Annotation applications?

Session 4. Data Annotation – Types, Methods & Data Types

Ria lived in a forest village and loved watching birds. One day, she drew birds and labelled their colours in her notebook. Her friend Raju told her about data annotation, and that her drawings could teach computers. They marked important things in pictures, so computers could learn. Raju showed her how to do it on a computer. Ria was amazed that her simple drawings could teach computers about birds. She learned that data annotation was like giving instructions to computers using pictures. Now, whenever Ria saw birds, she knew she was helping computers learn too.

Figure 4.1: Anjali drawing a bird (Source: https://storyweaver.org.in/)

In this chapter, you will first understand the concept of computer vision and its tasks. After that, you will get knowledge about different types of data annotation and their methods. Apart from this, you will also get knowledge about manual data annotation, data annotation using crowdsourcing, data annotation based on usage, data-driven data annotation, and data annotation using an artificial intelligence API.

4.1 Computer Vision

Computer vision gives computers the ability to see and understand the world through images and videos, just like humans use their eyes to see and understand. It is a field of artificial intelligence (AI) that focuses on teaching computers to analyse and make sense of visual information.

4.2 Computer Vision Tasks

The main tasks of computer vision are as follows:

- Image Classification
- Object Detection
- Object Tracking
- Image Segmentation
- Pose Estimation

4.2.1 Image classification

Image Classification method sorting images into different categories or labels based on their content. It is a fundamental task in computer vision. In this, an algorithm or machine learning model assigns a label or category to an image based on the features and patterns it detects within the image. For example, you have a collection of animal photos, and you want to automatically classify them into "Cats" and "Dogs." As illustrated in Figure 4.2.

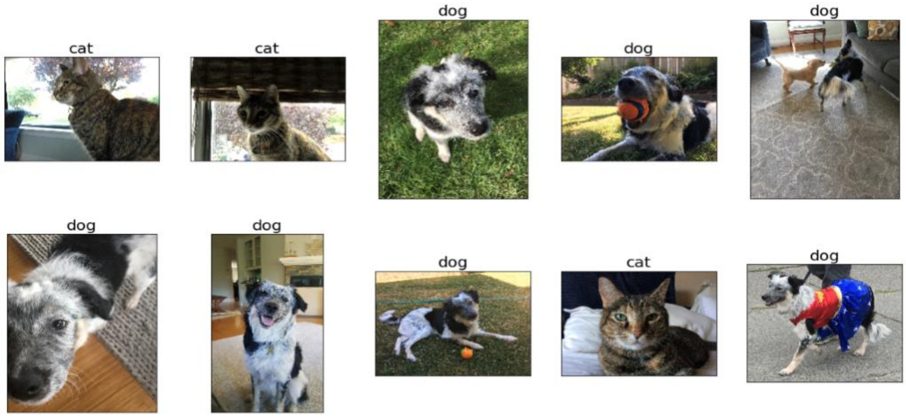


Figure 4.2: Image Classification

4.2.1.1 Types of Image Classification

There are two main types of Image Classification:

a) **Binary Class Classification and**

b) **Multi-Class Classification.**

a) **Binary Class Classification (Two Tags Only):** In binary classification, an image is categorized into one of two classes or labels. It is like a yes-or-no decision. For example, classifying images as either "cat" or "not a cat," "spam" or "not spam," or "fraudulent" or "legitimate."

b) **Multi-Class Classification (Multiple Tags):** In multi-class classification, an image is assigned to multiple classes or labels. It is used for classification with more than two distinct categories. For example, classifying images of fruits like "apple," "banana," "orange," or "grape."

4.2.2 Object Detection

Object Detection teaches computers to identify and locate objects within images or videos. Object Detection draws bounding boxes around each object to show their precise locations in the image, and also tells the names of the objects present in an image. For example, you have a photo that contains multiple objects, such as cars, dogs, and bicycles, as shown in Figure 4.3.

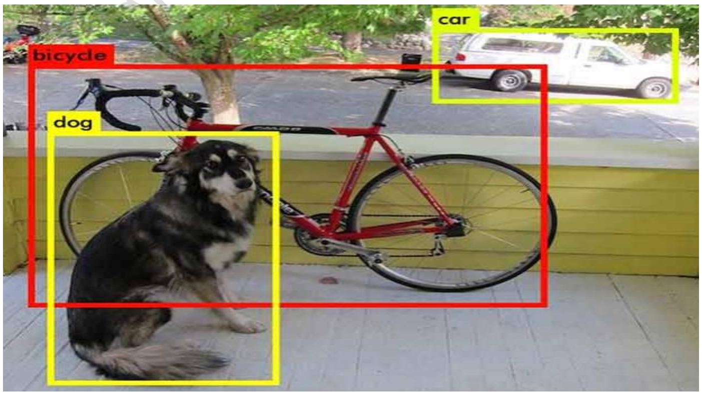


Figure 4.3: Object detection (source: https://towardsdatascience.com/)

4.2.2.1 Object Detection Applications

The various object detection applications are as follows:

- Autonomous Vehicles
- Surveillance
- Image-Based Search Engines

4.2.3 Object Tracking

Object Tracking is used to monitor the movement of objects in videos or image sequences. It involves identifying a specific object in the initial frame of a video and then keeping track of that object as it moves through subsequent frames. For example, in surveillance systems, object tracking can be used to follow a person or a vehicle as they move across different camera views. As shown in Figure 4.4.

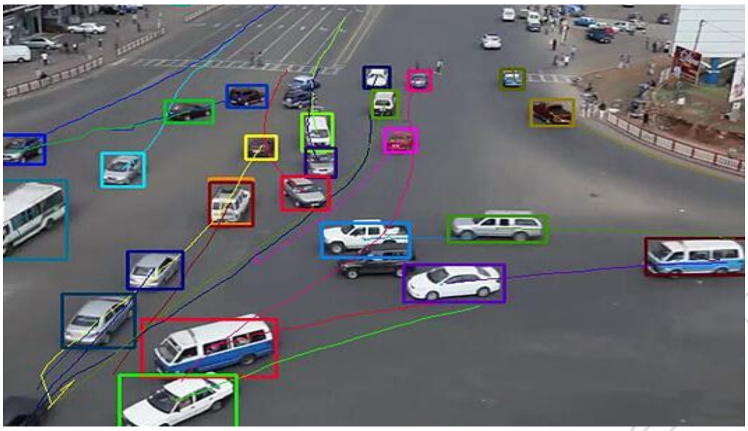


Figure 4.4: Object Tracking

4.2.3.1 Object Tracking Applications

Object tracking has applications as follows:

- Autonomous vehicles,
- robotics,
- CCTV cameras
- Various other fields where monitoring objects over time are crucial.

4.2.4 Image segmentation

The image Segmentation method divides an image into multiple meaningful regions or segments, where each segment represents a separate object or part of the image. Image segmentation would not only identify the cars, streets, persons, and buildings but also correctly outline and label each of them as separate entities. As illustrated in Figure 4.5.

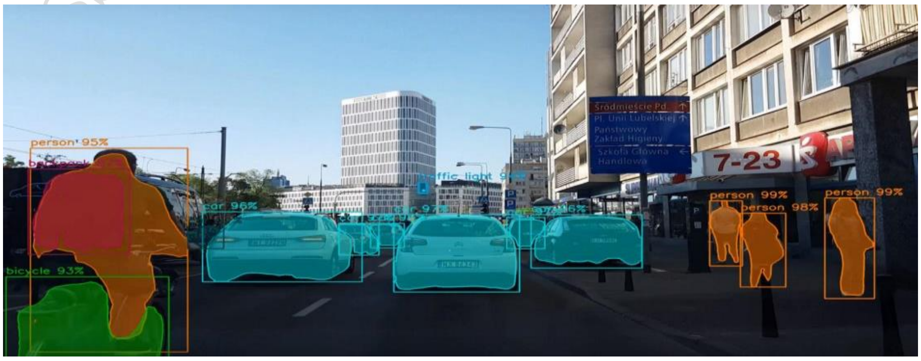


Figure 4.5: (source: https://imageannotation.home.blog/)

4.2.5 Pose estimation

Pose estimation is a task in computer vision that allows machines to identify human figures and understand their bodies positioned in videos and pictures. It assists machines in identifying specific body parts, like determining the position of a person's knee in an image. As illustrated in Figure 4.6.

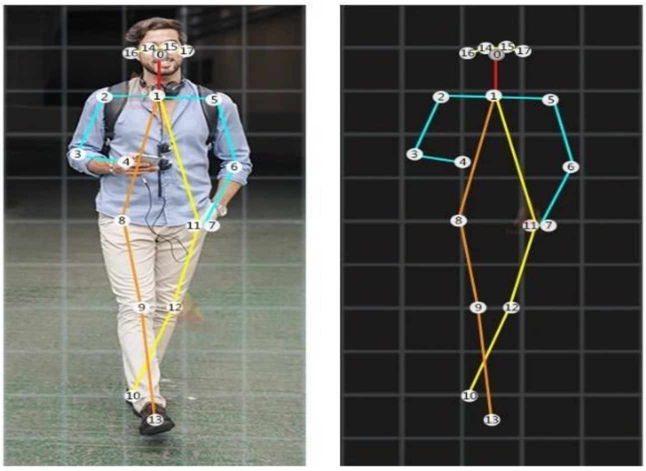


Figure 4.6: Example of Pose Estimation

4.3 Data Annotation Types

Data annotation involves different types of methods to provide additional information to data. These types help computers understand and interpret data more effectively. Some common data annotation types are Image Annotation, Text Annotation, Audio Annotation, and Video Annotation.

4.3.1 Image Annotation

Image annotation means Labeling or adding notes to objects in pictures to help computers understand and identify different objects, shapes, or areas within the images. This allows computers to perform various tasks, such as identifying objects, remembering things, and even colouring different parts of images. As illustrated in Figure 4.7.

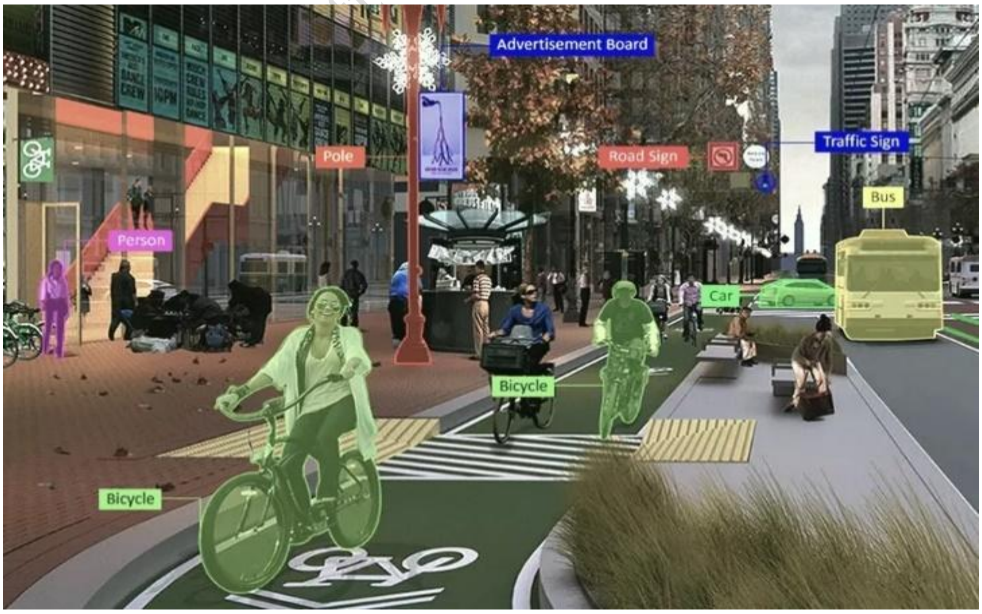
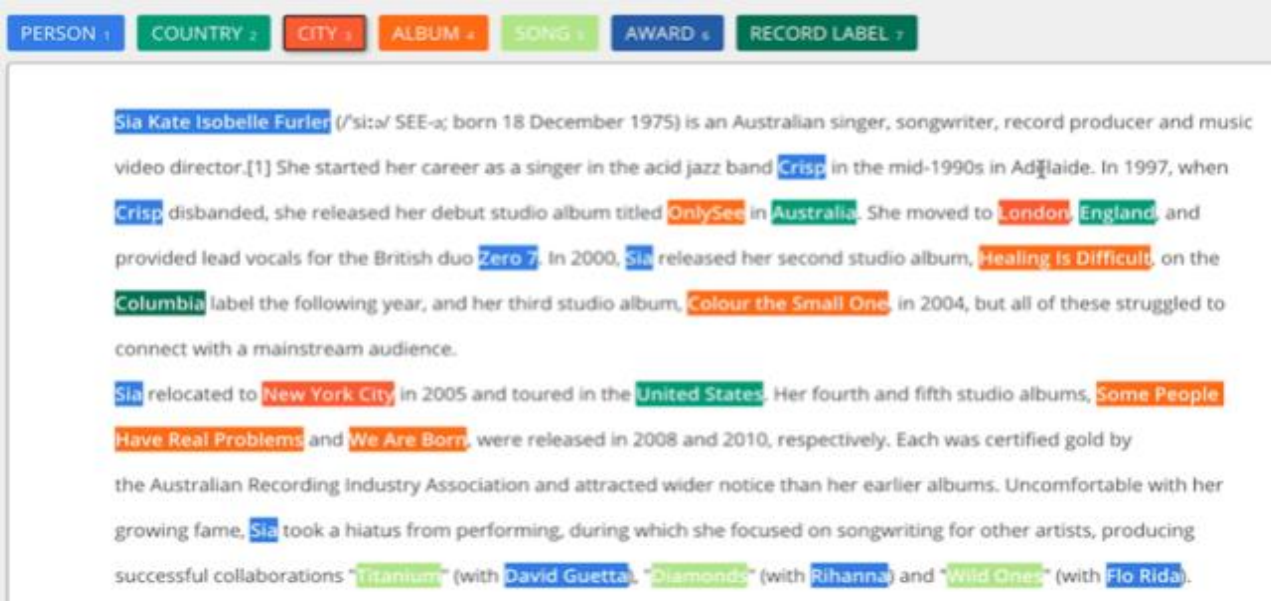


Figure 4.7: Image Annotation (source: https://www.cogitotech.com/)

4.3.2 Text Annotation

Text annotation is a process where human annotators add labels to text documents to help computers understand and analyse the content. Adding labels to text helps machines identify the important words in a sentence. In this, we use tags to highlight important things like keywords, phrases, or sentences in the text. Sometimes, we also tag feelings like "sad" or "happy" to help machines understand the emotions or intentions behind the words in certain situations. Figure 4.8 illustrates the example of text annotation.



Sia Kate Isobelle Furler (/'siːə/ SEE-ə; born 18 December 1975) is an Australian singer, songwriter, record producer and music video director.[1] She started her career as a singer in the acid jazz band Crisp in the mid-1990s in Adelaide. In 1997, when Crisp disbanded, she released her debut studio album titled OnlySee in Australia. She moved to London, England, and provided lead vocals for the British duo Zero 7. In 2000, Sia released her second studio album, Healing Is Difficult, on the Columbia label the following year, and her third studio album, Colour the Small One, in 2004, but all of these struggled to connect with a mainstream audience.

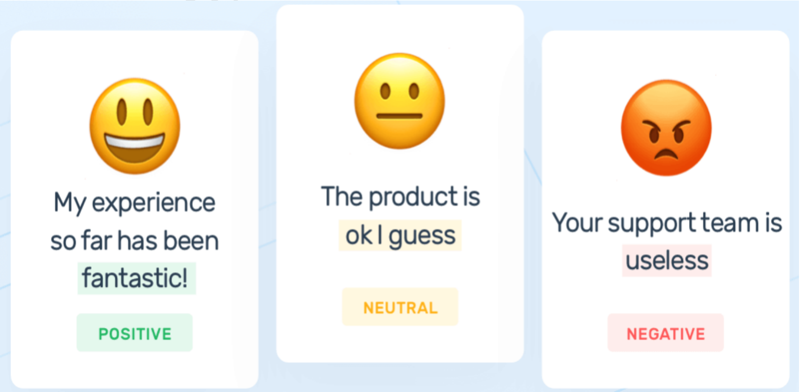
Sia relocated to New York City in 2005 and toured in the United States. Her fourth and fifth studio albums, Some People Have Real Problems and We Are Born, were released in 2008 and 2010, respectively. Each was certified gold by the Australian Recording Industry Association and attracted wider notice than her earlier albums. Uncomfortable with her growing fame, Sia took a hiatus from performing, during which she focused on songwriting for other artists, producing successful collaborations "Titanium" (with David Guetta), "Diamonds" (with Rihanna) and "Wild Ones" (with Flo Rida).

Figure 4.8: Example of Text Annotation

4.3.2.1 Types of Text Annotation

There are three types of text annotation, such as sentiment analysis, named entity recognition, and part-of-speech.

a) **Sentiment analysis-** Sentiment analysis means figuring out if a text is positive (happy), negative (sad), or neutral (neither). For example, when you read a review, it helps you know whether it is a good or bad review. Figure 4.9 shows an example of sentiment analysis.



My experience so far has been fantastic!

POSITIVE

The product is ok I guess

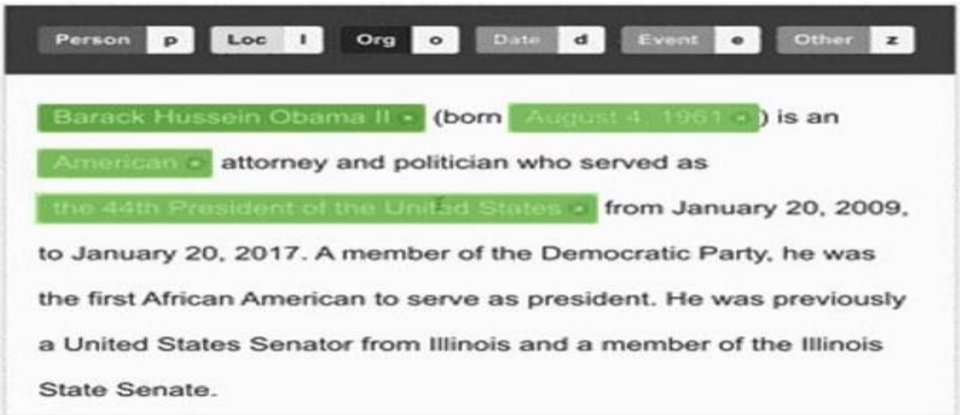
NEUTRAL

Your support team is useless

NEGATIVE

Figure 4.9: Example of Sentiment Analysis (source: https://monkeylearn.com/)

b) **Named Entity Recognition-** Named Entity Recognition (NER) is like a detective for text. It is a computer program that scans through sentences and finds important things like a person's name, places, dates, and more. It helps computers to understand the text and make it useful for tasks like organizing information or answering questions about a text. Figure 4.10 illustrates the example of named entity recognition.



Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to serve as president. He was previously a United States Senator from Illinois and a member of the Illinois State Senate.

Figure 4.10: Example of Named Entity Recognition (source: https://www.cogitotech.com/)

c) **Parts of Speech-** Parts of Speech (POS) are like word roles in a sentence. They tell us if a word is a noun (like 'dog'), a verb (like 'run'), an adjective (like 'happy'), and more. POS helps us understand the working of words together to make sentences. Figure 4.11 illustrates the example of parts of speech (POS).

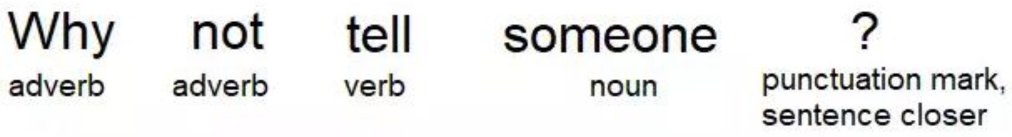


Figure 4.11: Example of Parts Of Speech (source: https://www.altexsoft.com/)

4.3.3 Audio Annotation

Audio annotation is used for adding labels or notes to audio files to help computers understand and process the sound. This process attaches descriptions to sounds so that computers can identify and understand them. Audio annotation is essential for developing and improving voice-based software. With tools, it improves audio quality and removes background noise, tasks like transcribing customer support calls, converting speech to text, and adding details like gender and sentiment become easier. This is important for the E-commerce sector and various other fields. Figure 4.12 illustrates the example of audio annotation.

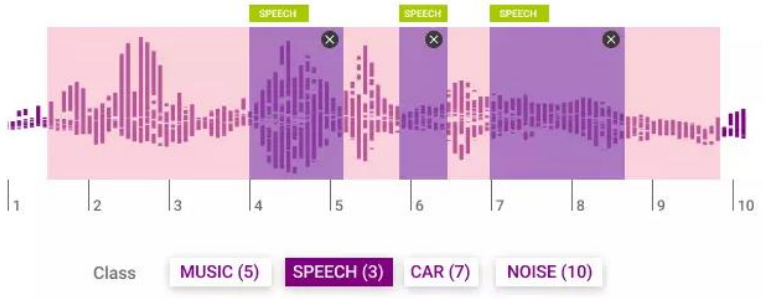


Figure 4.12: Example of Audio Annotation

4.3.3.1 Types of Audio Annotation

There are different types of audio annotation:

a) **Speech-To-Text Transcription:** Speech-To-Text Transcription is like turning spoken words into written text. It is a process where a computer listens to what's said in an audio recording, like a conversation or a speech, and then converts it into written words. This is useful for tasks like creating subtitles for videos, making audio searchable, etc. As illustrated in Figure 4.13.

Figure 4.13: Speech to Text Transcription

b) **Audio labeling:** Audio or speech labeling is like giving names to sounds or spoken words. It is a process where you listen to audio recordings and describe what you hear. For example, you might label parts of the audio as "dog barking," "car engine," or "person talking." This helps computers understand and categorize different sounds in the recordings. It is useful for tasks like voice recognition or analyzing audio data. As illustrated in Figure 4.14.

Figure 4.14: Example of Audio labeling (dog barking) (Source: https://www.vectorstock.com/)

Audio tagging: Audio tagging is like adding labels or tags to audio clips to describe what's in them. It is a way to organize and categorize audio files by attaching keywords or labels. For example, you might tag an audio clip as "jazz music," "birdsong," or "traffic noise." Audio tagging is used in music libraries. As illustrated in Figure 4.15.

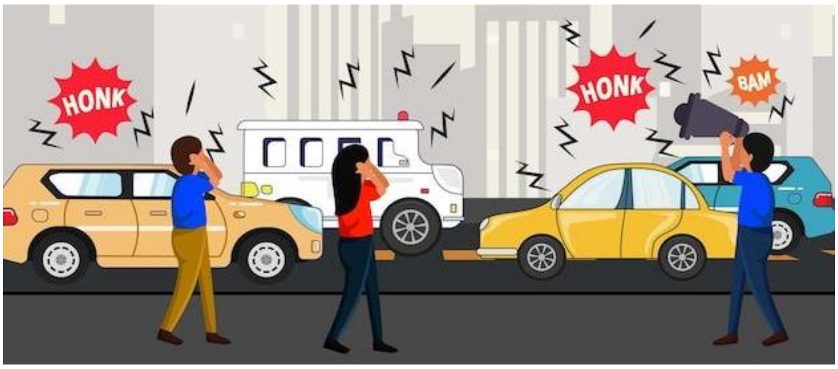


Figure 4.15: Example of Audio tagging (Traffic noise) (Source: https://www.freepik.com/)

c) **Speaker Identification:** Marking who is speaking in an audio recording, which is helpful for transcribing conversations or identifying different voices.

d) **Emotion Labeling:** Adding labels to indicate the emotional tone of the speech, such as happy, sad, angry, etc. As illustrated in Figure 4.16.

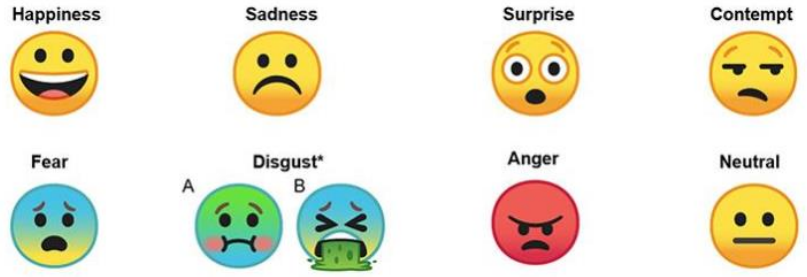


Figure 4.16: Example of Emotion Labeling (Source: https://www.asia-research.net/)

4.3.4 Video Annotation

Video annotation is the process of labeling and adding information to videos. Such as images, and videos must highlight specific details to help machines understand and analyze the content. Video annotation involves marking objects, drawing bounding boxes around them, and assigning labels to them. This allows machines to recognize and distinguish different objects or actions in the video. As illustrated in Figure 4.17.

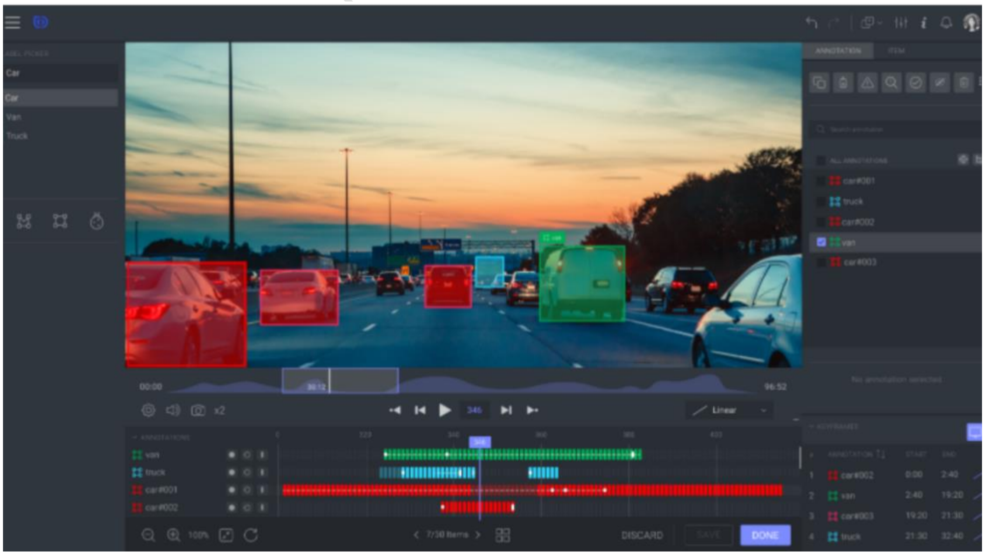


Figure 4.17: Example of Video Annotation (source: https://dataloop.ai/)

4.4 Annotation Techniques

Bounding boxes

Bounding boxes are like rectangles drawn around objects. They are like invisible rectangles we draw around objects in pictures or images. Imagine you have a picture of a cat. To show where the cat is, you draw a rectangle around it. This rectangle is the bounding box. It is like putting a frame around the cat. Figure 4.18 shows the visual representation of bounding boxes. You can see here that each object is covered inside a rectangular box. As illustrated in Figure 4.18.

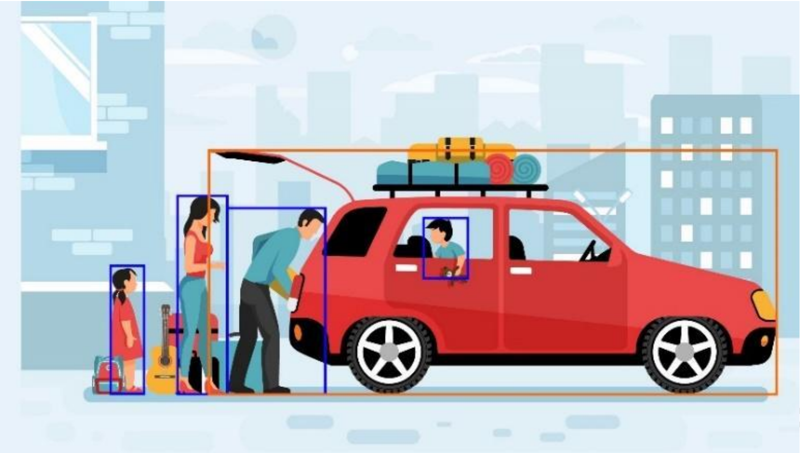


Figure 4.18: Bounding boxes (Source: https://www.anolytics.ai/)

Two types of Bounding boxes:

- **2D Bounding boxes**: In 2D Bounding boxes, a rectangular or square outline is drawn around the object we want to focus on. This is done to make it clear that there is an object in the picture. Figure 4.19 illustrates the 2D bounding boxes.

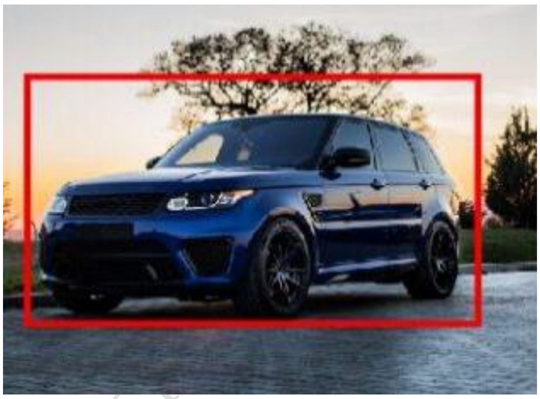


Figure 4.19: Example of 2D bounding boxes

- **3D Bounding boxes:** 3D bounding box annotation is a method used to mark and outline objects in a three-dimensional (3D) space. Instead of just drawing a rectangle around an object as in 2D annotation, 3D bounding box annotation creates a cuboid or box that fully encloses the object in three dimensions (length, width, and height) within a 3D space. As illustrated in Figure 4.20.

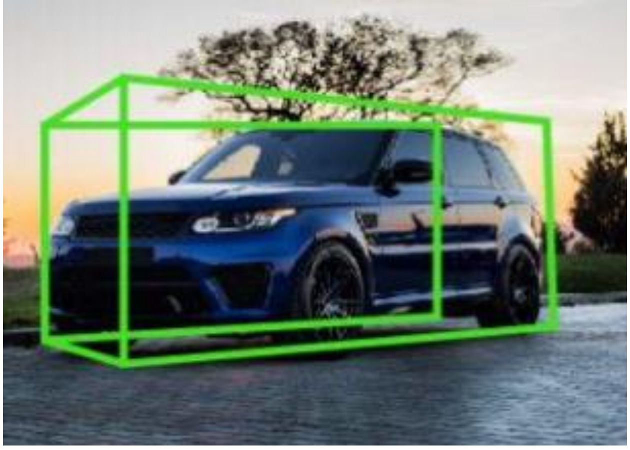


Figure 4.20: Example of 3D bounding boxes

Lines and splines

In adding extra information to data, we use lines and curved lines to find and understand lanes. These lanes are important for self-driving cars to know where to go. The example is shown in Figure 4.21.

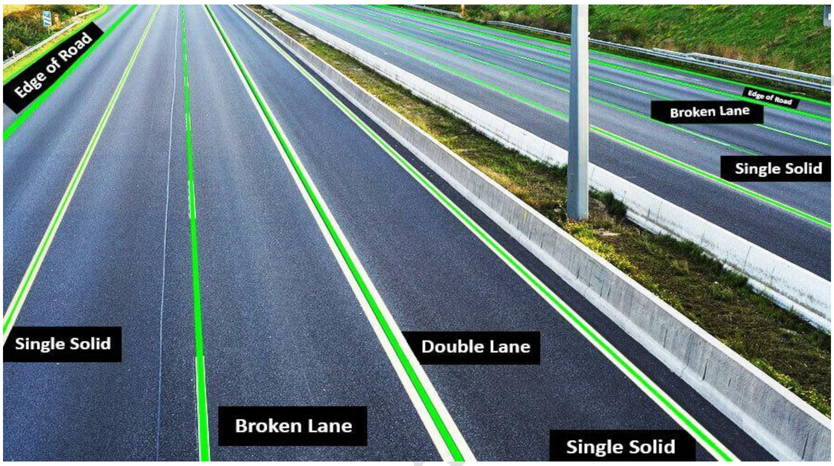


Figure 4.21: Example of Lines and splines (Source: https://tarjama.com/)

Semantic segmentation

Semantic segmentation is a way of labeling different parts of an image with colors or marks. Each object in an image gets its color so the computer can identify objects easily in an image. This helps computers understand the details and boundaries of objects in an image. For example, in a street picture, semantic segmentation can help the computer separate the road, sidewalks, buildings, and cars into different areas. The visual representation of Semantic segmentation is shown in Figure 4.22.

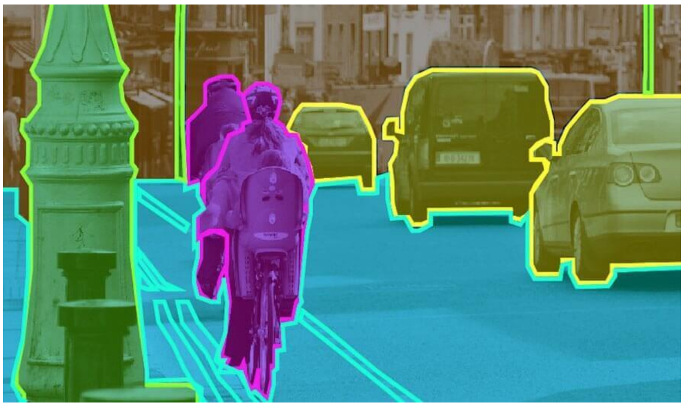


Figure 4.22: Semantic Segmentation (Source: https://www.anolytics.ai/)

Instance segmentation

Instance segmentation is a technique used in computer vision to do three important things in an image:

- Detection: It finds and identifies all the objects in the image.
- Segmentation: It outlines and separates each object accurately, often with pixel-level precision.
- Classification: It assigns a category or label to each object, distinguishing one object from another.

Consider a picture of cats and dogs below, in Figure 4.23. Semantic segmentation can show that there are dogs and cats in a picture, but it can't tell you the number count of each. Whereas, for Instance segmentation, you not only find the dogs and cats but also count the number of dogs and cats. So, you can count them easily in the picture.

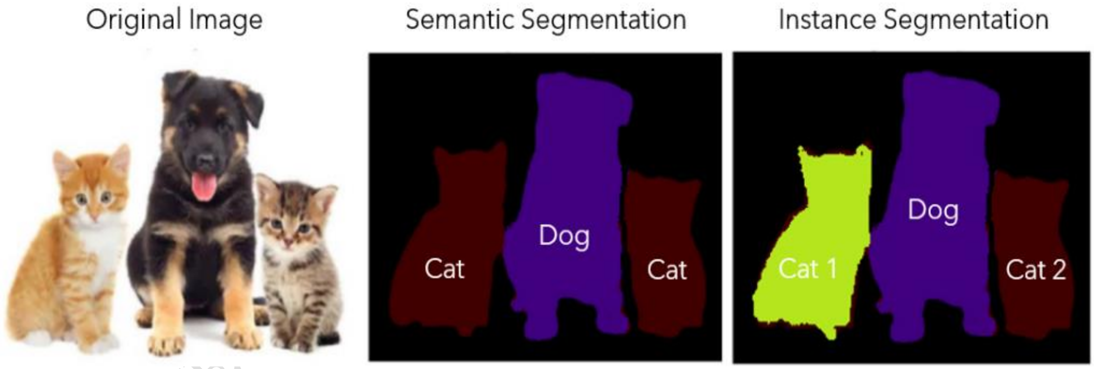


Figure 4.23: Example of instance segmentation

3D cuboids

3D cuboids are like boxes with three dimensions: length, width, and height. They are used to show the shape and size of objects in a 3D space. Imagine a shoebox but with height, width, and depth. These cuboids are useful in computer vision to show the boundaries of objects in a three-dimensional world.

For example, suppose you are teaching a computer to identify different objects in a room. In that case, you might use 3D cuboids to draw boxes around things like chairs, tables, and books to help the computer understand their size and position, as shown in Figure 4.24. These boxes are called 3D cuboids. With 3D cuboids, computers can understand things in a 3D world. This is great for making driverless smart cars. They can figure out how close things are to them by using these boxes and driving safely.

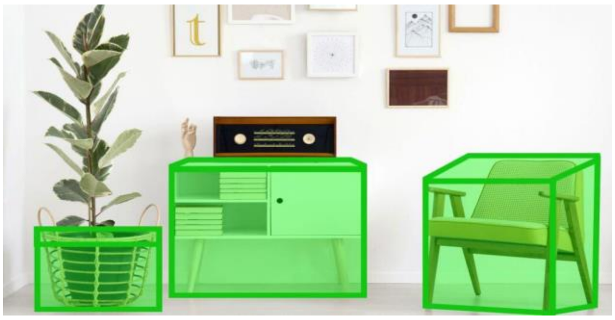


Figure 4.24: Example of 3D Cuboids (Source: https://www.anolytics.ai/)

Polygonal segmentation

Polygonal segmentation is a method used to outline the shapes of objects in an image using lines that connect points. It is like tracing the edges of things with a pencil but on a computer. Instead of just using simple boxes, you draw more complex shapes. These shapes are made of straight lines that connect points, forming a polygon. This helps computers to understand objects' exact shape and outline in a picture. For example, if you want to show the outline of a car in an image, you would use polygonal segmentation to draw a shape that fits around the car's body, tyres, and headlight as shown in Figure 4.25.

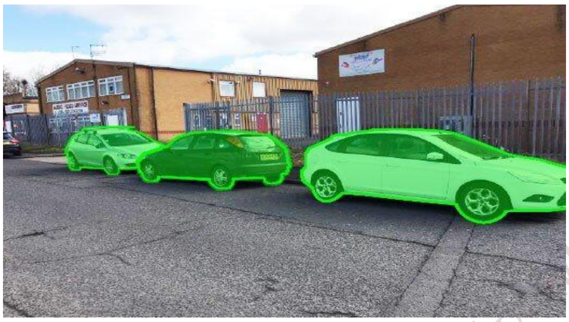


Figure 4.25: Example of Polygonal segmentation (Source: https://www.anolytics.ai/)

Landmark and key-point

Landmark and key-point annotation is a technique for marking important points or spots in an image. These points could be an object's corners, edges, or specific parts. It is like putting dots on a picture to show particular places. This helps computers to understand where important things are located and positioned. For example, if you are teaching a computer to recognize faces, you might mark the eyes, nose, and mouth as landmarks to help the computer know where these features are in the picture. The visual representation is shown in Figure 4.26.

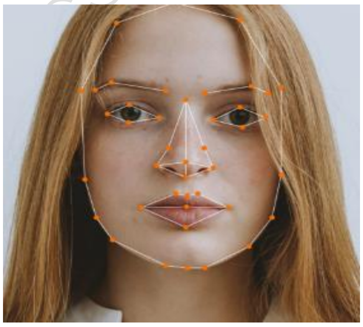
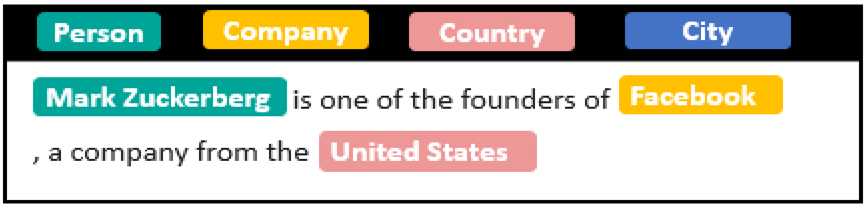


Figure 4.26: Example of Landmark and key-point (Source: https://smartone.ai/)

Entity annotation

Entity annotation labels specific words, phrases, or parts of a text to identify their meaning. It is like highlighting important information in a text. This helps computers understand the context and meaning of different parts of a document. For example, in a news article, you might label the person's name, places, dates, and other important details. Entity annotation is a way to tag important words in sentences so that computers can understand them better. As illustrated in Figure 4.27.



Mark Zuckerberg is one of the founders of Facebook , a company from the United States

Figure 4.27: Example of Entity Annotation (source: https://tarjama.com/)

3D LiDAR

3D LiDAR (Light Detection and Ranging) annotation is a critical process in the field of computer vision and autonomous systems. 3D LiDAR Annotation is marking objects in a 3D map created by lasers, like putting labels on cars, persons, and buildings in a self-driving car's view to help it drive safely. Figure 4.28 illustrates the example of 3D LiDAR.

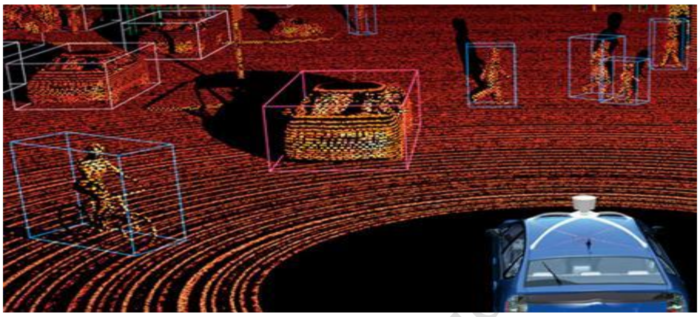


Figure 4.28: Example of 3D LiDAR

4.5 Data Annotation Methods

4.5.1 Manual Data Annotation

In the initial phases of a project, manual data annotation can be carried out if the datasets are small or the aim is to create a prototype swiftly. In such cases, developers review the data and label the samples based on specific annotation criteria.

Advantages: This approach has various benefits:

a) It demands minimal oversight of data annotation activities.

b) Engineers better grasp the data than many other experts, enhancing its quality.

c) Engineers might discover unique insights about the data that can enhance an algorithm.

d) This method allows for greater control and accuracy.

e) It has the ability to capture complex information.

Disadvantages: This approach has various disadvantages:

a) A small team is responsible for annotation challenges, leading to lower quality and slower progress.

b) This approach may not be feasible for handling a large dataset.

c) It is time-consuming process

d) It involves adding metadata to datasets by human annotators

e) It can lead to poor quality due to human errors.

4.5.2 Data Annotation using Crowdsourcing

Using crowd assistance for data annotation offers a cost-effective and scalable solution. Platforms such as Amazon Mechanical Turk and Crowd flower are well-known examples of crowdsourcing services for this purpose.

Benefits:

a) It is Affordable.

b) It is capable of handling large amounts of data.

c) It is fast.

d) It has the ability to obtain large labelled data at a low cost.

e) It allows for fast turnaround times and scalability.

Drawbacks:

a) Requires knowledge of various crowdsourcing methods;

b) Requires implementation of quality control measures.

c) It suffers from inaccuracies and inconsistent Annotations.

d) It has biased annotations.

e) It has less data privacy and security.

4.5.3 Data Annotation based on the usage

Data annotation can be skipped if data already has labels. For example, a bank's loan data comes with approvals and denials. This data can train a machine learning model. Raw annotations might need cleaning but can be used for more refined annotation processes, like crowd-sourcing or manual annotation. This approach helps keep costs lower compared to complete annotation.

Advantages:

a) Economical or low-priced

b) Suitable for handling large datasets

c) Requires minimal administration if the data-producing business process is well-organized.

d) It improves the accuracy of output.

e) It improves the experience for users.

Disadvantages:

a) Possibility of noise in user-generated data or content

b) Additional post-processing is often required.

c) It has less data privacy.

d) It has less data security.

e) It suffers with poor quality data.

4.5.4 Data-Driven Data Annotation

In AI projects, simple rules can address part of the data. If this subset is representative and high-quality, it can be used to train a machine-learning model that works well for the entire dataset. Consider extracting job responsibilities from job descriptions. While most job descriptions lack structure with natural language and flexible HTML, some include headers like 'Responsibilities:' and bullet-pointed duties. Such job descriptions' HTML codes reveal a framework such as:

```
<h3>Data Annotation</h3>
<ul>
<li>TEXT</li>
...</ul>
```

Advantages: No cost involved; it needs very little management; it can handle a large amount of data.

Disadvantages: Extraction patterns might not always be very accurate, and the patterns might not cover all types of data, limiting the performance of the trained model can work for different situations.

4.5.5 Data Annotation Using an Artificial Intelligence API

Using existing APIs for your app's tasks to process data samples efficiently. Review and adjust the results for accuracy before integrating them into your application, similar to usage-based data annotation. The main difference is cost. For instance, in building a sentiment analysis system, using a sentiment analysis API can assign scores to texts. While AI APIs may not fully solve your task, they aid in data filtering. For example, with just a face detection API, you can detect faces in images before using another API for emotion tagging.

Advantages: Cost-effective, simple to handle, and suitable for large collections of books.

Disadvantages: It involves expenses, and the AI API might occasionally create errors that introduce label inaccuracies.

Summary

- Different ways to do data annotation are: bounding boxes, creating lines and curves, categorizing areas in images, outlining 3D shapes, marking polygons, pinpointing landmarks, and Labeling entities.
- Methods for data annotation include: doing it manually, getting help from crowdsourcing, deciding based on how the data will be used, following a data-driven approach, and using an Artificial Intelligence API for annotation.
- Examples of data types in data annotation are text annotation, audio annotation, image annotation, and video annotation.

Check Your Progress

A. **Multiple Choice Questions**

1. What type of Data Annotation involves adding labels and descriptions to images? (a) Text Annotation (b) Audio Annotation (c) Image Annotation (d) Video Annotation
2. Which Data Annotation Method marks specific points on an image? (a) Bounding boxes (b) Lines and splines (c) Landmark and key-point (d) Semantic segmentation
3. Which Data Annotation Type is most suitable for transcribing spoken words into text? (a) Image Annotation (b) Text Annotation (c) Audio Annotation (d) Video Annotation
4. Which Data Annotation Method is used to outline the boundaries of objects in images? (a) Bounding boxes (b) Lines and splines (c) 3D cuboids (d) Entity annotation
5. What kind of Data Annotation involves marking specific areas within an image? (a) Entity annotation (b) Polygonal segmentation (c) Bounding boxes (d) Video Annotation
6. Which Data Annotation Type labels specific actions or events in a video? (a) Image Annotation (b) Audio Annotation (c) Text Annotation (d) Video Annotation
7. What does Audio Annotation involve? (a) Adding descriptions to images (b) Labeling objects in videos (c) Transcribing spoken words into text (d) Marking key points on images
8. Which Data Annotation Method labels objects in a 3D space? (a) Bounding boxes (b) 3D cuboids (c) Lines and splines (d) Landmark and key-point
9. Which Data Annotation Type is suitable for marking specific words or phrases in a document? (a) Text Annotation (b) Image Annotation (c) Audio Annotation (d) Video Annotation
10. What is the primary purpose of Video Annotation? (a) Adding text descriptions to videos (b) Labeling specific actions or events in a video (c) Transcribing audio in videos (d) Highlighting key points in videos

B. **Fill in the blanks.**

1. Image Annotation involves adding labels to ___________.
2. Text Annotation is used to annotate and label ___________ data.
3. Audio Annotation helps in Labeling and transcribing ___________ content.
4. Video Annotation involves annotating and identifying objects or actions in ___________.
5. Bounding boxes are commonly used in image annotation to outline ___________.
6. 3D cuboids are used to annotate objects in ___________ space.
7. Key points are commonly used in image annotation to identify important ___________.
8. Lines and splines are used in annotation to outline ___________ shapes.
9. Video annotation helps in tracking and identifying ___________ within videos.
10. Audio annotation involves transcribing spoken ___________ into text.

C. **State whether true or false**

1. Image Annotation involves Labeling and annotating textual data.
2. Text Annotation is a method used for Labeling and transcribing audio content.

3. Audio Annotation helps in identifying and Labeling objects in images.
4. Video Annotation is used to annotate and identify objects or actions in videos.
5. Bounding boxes are commonly used in image annotation to outline objects.
6. Semantic segmentation focuses on Labeling individual segments within an image.
7. 3D cuboids are used to annotate objects in two-dimensional space.
8. Polygonal segmentation involves annotating objects with only straight-line edges.
9. Landmark and key-point annotation identifies specific points in an object.
10. Entity annotation is used to label and tag specific entities in text.

D. **Short answer questions**

1. What is the purpose of Video Annotation?
2. Describe the concept of Bounding Boxes in image annotation.
3. How does Semantic Segmentation work?
4. What is the role of Landmark and Key-point annotation?
5. Explain Entity Annotation.
6. Explain the significance of Text Annotation.
7. Explain Image annotation and its working.
8. Explain audio annotation and its types.
9. How are Bounding boxes useful in annotation?
10. Explain 3D cuboids.

Session 5. Tools of Data Annotation

Meet Priya, a fashion-loving girl with a dream. She wanted to be a fashion designer but had to go through many fashion images to learn the latest trends. Then, Priya discovered data annotation tools with her aunt's help. These tools were like magic pens for computers. Priya drew boxes around dresses, shoes, and accessories in pictures. The computer learned what each item was and showed her similar designs. Priya loved it! She created her virtual fashion gallery and used the computer to spot trends and design outfits faster. Years later, Priya became a successful fashion designer, as shown in figure 5.1.

Figure 5.1. Priya working as a Fashion designer (Source: istockphoto)

In this chapter, you will gain knowledge about different data annotation tools. In the end, you will understand open-source data annotator tools.

5.1. Tools of Data Annotation

The tools of data annotation help you to annotate machine learning training data in a production setting. These tools can be software that works online and offline. These tools are available for business use, either for renting or purchasing. They can annotate various data types like images, videos, text, audio, spreadsheets, and sensor data. Here are some of the data annotation tools:

Colabler

The "Colabler" data annotation tool is software designed to help people annotate various data types. It provides a platform where multiple users can collaborate in adding labels, marks, or other annotations to different kinds of data, such as images, videos, texts, and more. This tool enables efficient teamwork and sharing of tasks, allowing individuals to contribute to the annotation process collectively. It is particularly useful for projects requiring multiple annotators' input, ensuring consistency and accuracy in the annotated data. The picture representation of the Colabler tool is illustrated in Figure 5.2.

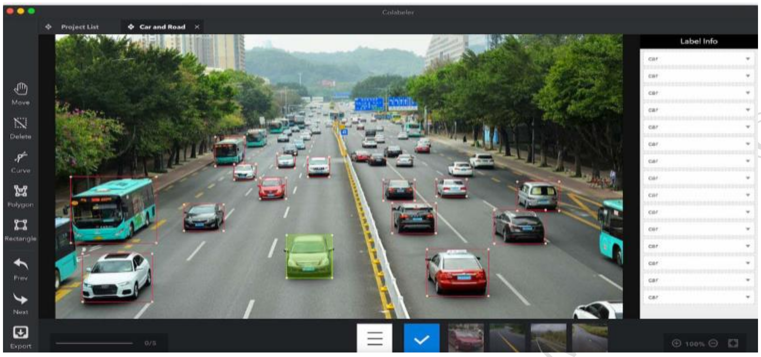
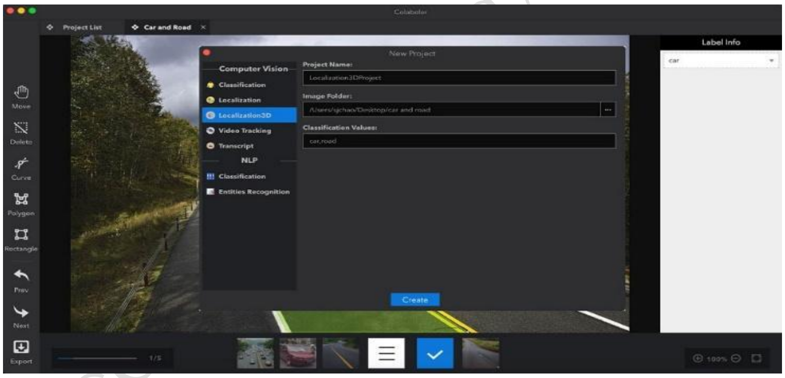


Figure 5.2. Colabler Tool (Source: http://www.colabeler.com/)



Project Name: Localization3DProject

Image Folder: /Users/qchao/Desktop/car and road

Classification Values: car,road

Figure 5.3. Image classification in Colabler Tool (Source: http://www.colabeler.com/)

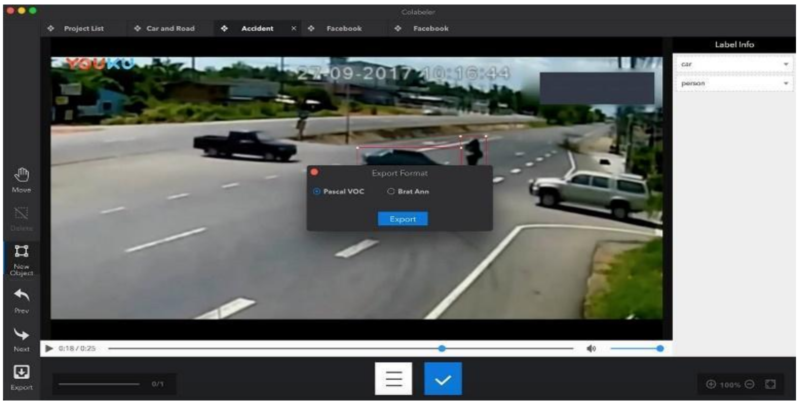
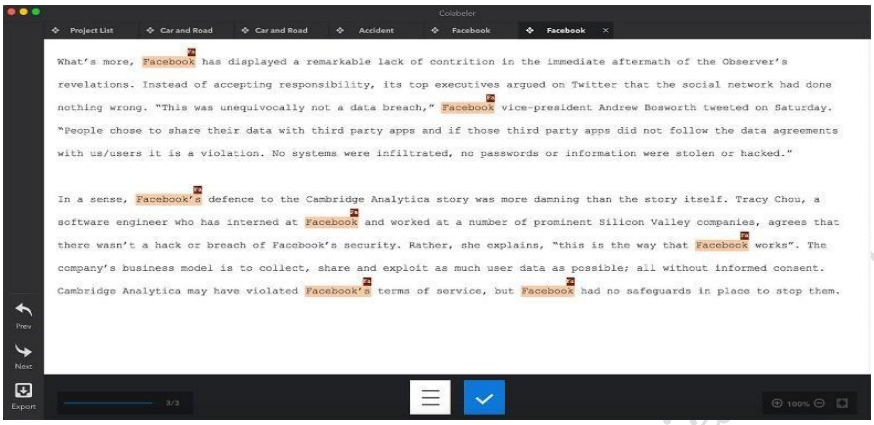


Figure 5.4. Video Labeling using CoLabeler (Source: http://www.colabeler.com/)



What's more, Facebook has displayed a remarkable lack of contrition in the immediate aftermath of the Observer's revelations. Instead of accepting responsibility, its top executives argued on Twitter that the social network had done nothing wrong. "This was unequivocally not a data breach," Facebook vice-president Andrew Bosworth tweeted on Saturday. "People chose to share their data with third party apps and if those third party apps did not follow the data agreements with us/users it is a violation. No systems were infiltrated, no passwords or information were stolen or hacked."

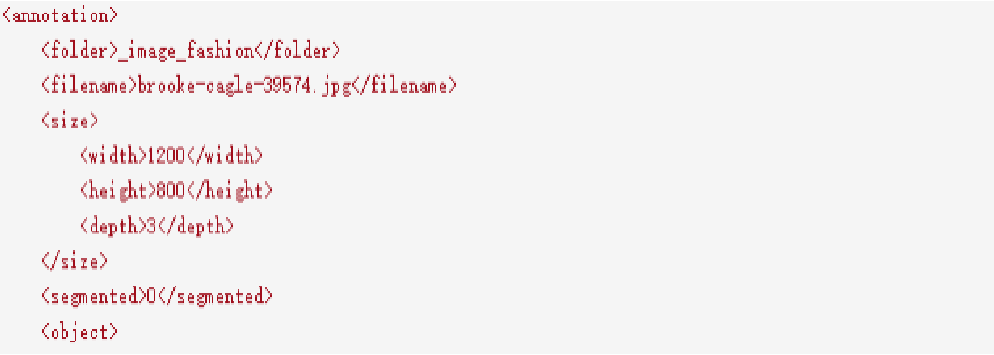
In a sense, Facebook's defence to the Cambridge Analytica story was more damning than the story itself. Tracy Chou, a software engineer who has interned at Facebook and worked at a number of prominent Silicon Valley companies, agrees that there wasn't a hack or breach of Facebook's security. Rather, she explains, "this is the way that Facebook works". The company's business model is to collect, share and exploit as much user data as possible; all without informed consent. Cambridge Analytica may have violated Facebook's terms of service, but Facebook had no safeguards in place to stop them.

Figure 5.5. Text labeling using CoLabeler (Source: http://www.colabeler.com/)

As shown in Figure 5.2, Figure 5.3, 5.4, and Figure 5.5, this Colabeler tool is designed for AI data Labeling and works with images, videos, and texts. With this tool, you can perform various types of data annotations, including:

- Image classification
- Marking objects with bounding boxes
- Outlining shapes with polygons
- Drawing curves
- Locating objects in 3D
- Tracing actions in videos
- Categorizing text
- Labeling entities in text

Furthermore, the tool supports custom task plugins, allowing users to design their Labeling tools. It includes features for exporting files in PascalVoc XML format (the same format used by ImageNet) and CoreNLP file format. This tool is accessible on Windows, Mac, CentOS, and Ubuntu operating systems. It is compatible with the annotation format provided by PascalVoc XML. An example is provided below, as shown in Figure 5.6.



```
<annotation>
    <folder>_image_fashion</folder>
    <filename>brooke-cagle-39574.jpg</filename>
    <size>
        <width>1200</width>
        <height>800</height>
        <depth>3</depth>
    </size>
    <segmented>0</segmented>
    <object>
```

Figure 5.6. Annotation method using PascalVoc XML

Labelbox

This platform has strong AI-based data Labeling tools. It helps users manage data by automatically Labeling it and training models. Users can work together with the team on it. This tool supports importing and exporting annotations in various formats, ensures high-quality labels, and works well with popular cloud platforms.

Labelbox offers the flexibility to change the tools to fit your unique requirements. It offers different services and solutions, including:

- Data extraction from documents
- Security surveillance
- Healthcare applications like ultrasonography and digital pathology
- Property assessment for insurance purposes
- Agriculture tasks like crop weed detection and livestock monitoring.
- Transportation needs, particularly for safe-driving cars.

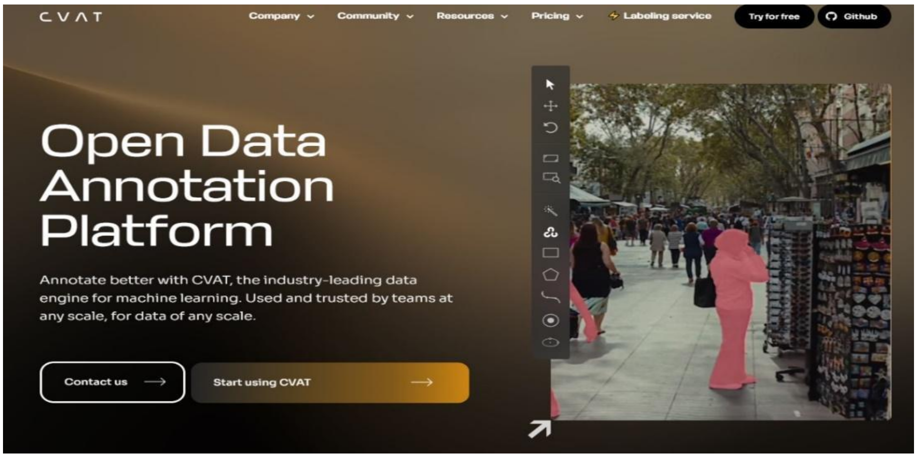
LabelBox Toolkit: LabelBox offers support for various data annotations, including bounding boxes, lines, and points. The toolkit provided by LabelBox includes:

- Instance segmentation tool (pen & super pixels): This tool helps segment instances using pens and super pixels, dividing them into distinct pixel ranges to analyse various components.
- Super pixel tool: This tool aids in the detailed analysis of instances by breaking them down into different pixel ranges.
- Draw over objects: With this tool, you can create outlines around objects.
- Brush: This tool has a unique radius, different from a standard paintbrush.
- Eraser: This tool is used to remove existing selections.

Open Source Tools: CVAT, VGG Image Annotator

CVAT (Computer Vision Annotation Tool)

CVAT, which stands for Computer Vision Annotation Tool, is a free tool for Labeling images and videos to help with machine learning and computer vision projects. It has a user-friendly interface and supports different types of Labeling like object detection, segmentation, image classification, and key point annotation. A visual representation of CVAT is shown in Figure 5.7.



Open Data Annotation Platform

Annotate better with CVAT, the industry-leading data engine for machine learning. Used and trusted by teams at any scale, for data of any scale.

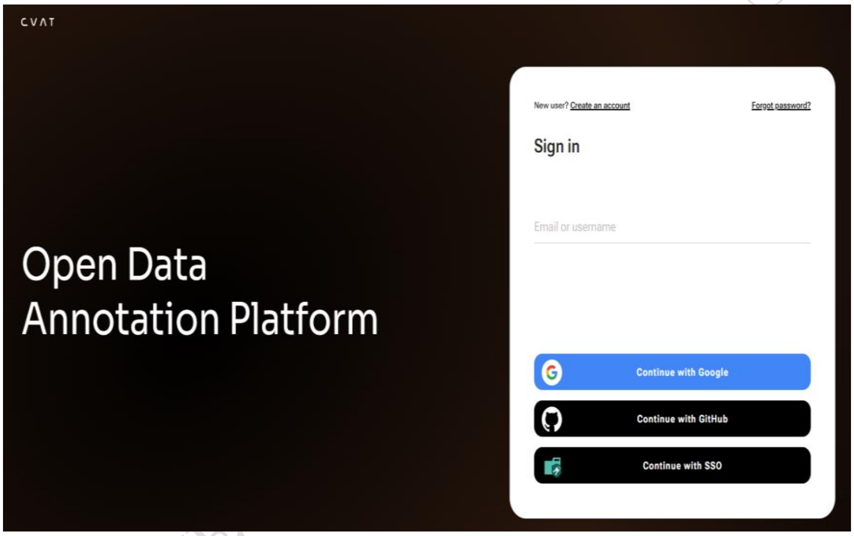
Figure 5.7: Graphical representation of CVAT Tool homepage (source: https://www.cvat.ai/)

CVAT Registration

To use CVAT, you need to create an account or log in if you already have an account. There are two steps:

1. User Registration
2. Account Access

To make an account or sign in, visit the App CVAT login page, as shown in Figure 5.8.



Open Data Annotation Platform

Figure 5.8: CVAT Login page

1. **User Registration**

If you want to sign up as a regular user (not an admin), follow these steps:

i. Click on "Create an account." As shown in figure 5.9.

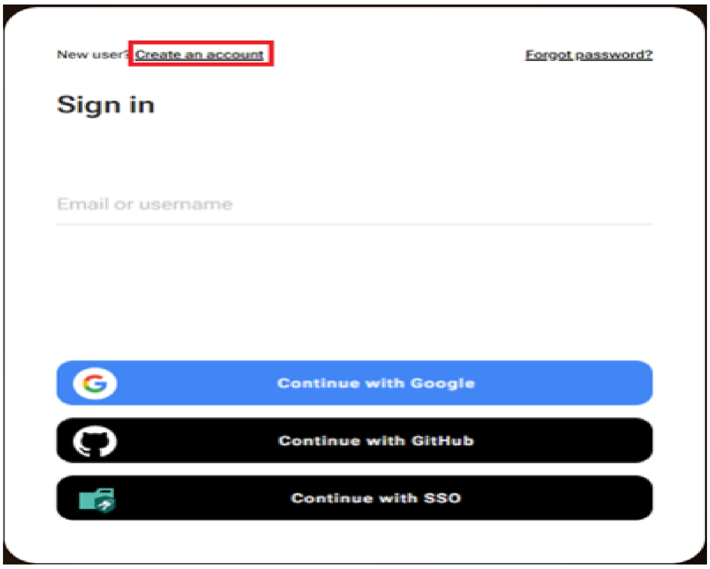


Figure 5.9: User Sign in page

ii. Complete all the empty spaces with the required information, agree to the terms of use, and click the "Create an account" button. Your username is created automatically using your email. You can change it if necessary. As shown in Figure 5.10.

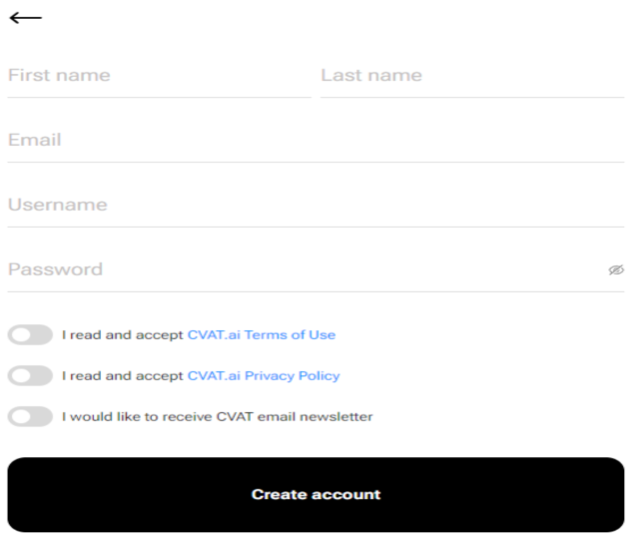

←
First name
Last name
Email
Username
Password
I read and accept CVAT.ai Terms of Use
I read and accept CVAT.ai Privacy Policy
I would like to receive CVAT email newsletter
Create account

Figure 5.10. User registration page

iii. If you want to register using Google or GitHub, click the button with the service's name and follow the instructions on the screen.

2. **Account Access**

If you want to use your account, here are the following steps:

i. Open the login page.

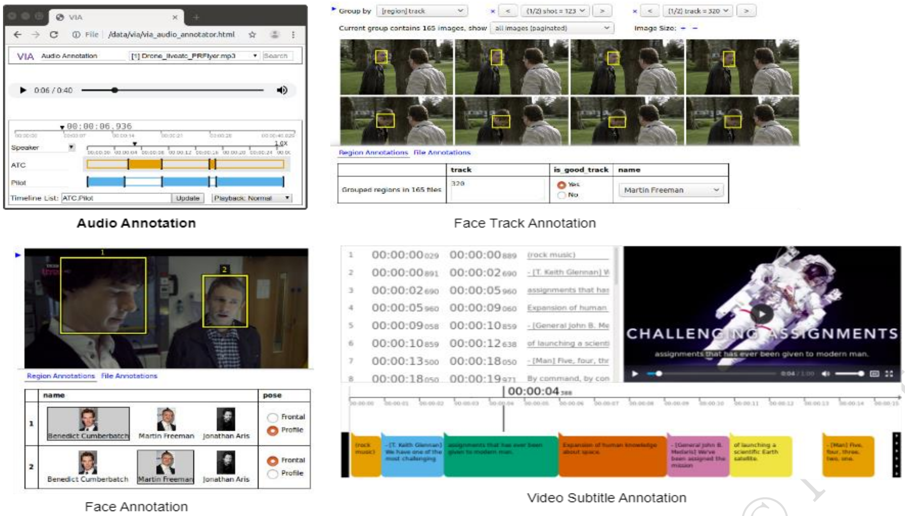
ii. Put in your username or email. The password box will show up.

iii. Type in your password and press Next.

iv. If you prefer to use Google or GitHub, click the button for the service.

VGG Image Annotator

The VGG Image Annotator is designed to simplify the manual annotation of images, audio, and videos. VIA is a free and open-source web application developed using HTML, JavaScript, and Cascading Style Sheets without relying on external libraries. It allows direct use within web browsers, reducing the need for downloads or complex setups. The entire VIA tool is accessible through a single HTML page, less than 400 kilobytes, and it can even work offline on most modern web browsers. Visual representations of VGG applications are shown in Figure 5.11.


Audio Annotation

Face Track Annotation

Face Annotation

Video Subtitle Annotation

Figure 5.11: VGG Annotation platform

(Source: VGG Annotation web page - https://www.robots.ox.ac.uk/~vgg/software/via/)

5.2. **Documentation Annotation tools**

Tools like Adobe and MS Word facilitate document annotation.

i. Annotation using Word: You can also use Microsoft Word for document annotation. Here is a step-by-step guide on how to use Microsoft Word for document annotation.

Step 1: Open Your Document

- Open Microsoft Word on your computer.
- Click "File" at the top left corner and choose "Open" to open the document you want to work on.

Step 2: Select Text

- Read your document and find the part to which you want to add a note.
- Click and drag your mouse to select the text you want to annotate.

Step 3: Add a Comment

- Look for the "Review" tab at the top of the screen.
- Click on "New Comment." A little speech bubble will appear next to the text you selected.

Step 4: Write Your Note

- Click inside the speech bubble to write your annotation or note.
- Type in your thoughts, questions, or explanations about the selected text.

Step 5: Save Your Work

- Don't forget to save your document to keep your annotations safe.

- Click on "File" again and choose "Save" or "Save As."

Step 6: Read and Reply to Comments

- If someone else added comments, you can see them by clicking on the speech bubbles.
- You can also reply to their comments by typing inside the speech bubbles.

Step 7: Resolve Comments

- After you read and understand a comment, you can "resolve" it.
- This means you addressed the comment, and it would not show up as a new comment anymore.

Step 8: Share Your Document

- You can share your document with others if you are working on a group project.
- Go to "File," choose "Share," and follow the steps to share your annotated document.

Step 9: Close Your Document

- Upon completing your work on the document, you can close it.
- Make sure you save any changes before closing.

ii. Annotate using Adobe: Annotation adds notes, highlights, and other markings to a document to provide additional information or emphasize specific parts. Adobe offers powerful tools that make annotation easy, whether working with PDFs, images, or other documents.

Adobe Acrobat, for example, is a widely used tool for annotating PDFs. Here is the process to do it:

Step 1: Open Your PDF

- Open Adobe Acrobat on your computer.
- Find and open the PDF document you want to annotate.

Step 2: Use Annotations

- Look for the "Comment" feature at the top of the screen and click on it.
- Read the PDF to understand what you need to annotate.

Step 3: Add Annotations

- Choose the annotation tools you want to use.
- To highlight text, select the highlighter tool and drag it over the text.
- To underline or strikethrough text, choose those tools and select the text.
- Pick the appropriate tools to add text boxes or sticky notes and place them where you want.

Step 4: Save Your Work

- Click on "File" and choose "Save" to keep your annotations in the PDF.

Step 5: Share Your Annotations

- You can use the "share with others" icon to share your annotated PDF with others.
- Follow the steps to send the file to your friends or classmates.

Step 6: Keep Your PDF Safe

- Remember not to lose your work. Keep the PDF in a safe place on your computer.

5.3. Data Management Standards - General Principles and Standards

Data Management Standards are a set of rules and guidelines that organizations follow to ensure that their data is handled, stored, and used effectively and consistently. These standards help maintain data quality, security, and integrity throughout its lifecycle. Here are some general principles and standards that are commonly included in data management standards:

Information included in a protocol:

- Basic guidelines for keeping records (such as sample annotations).

Standards for outlining experimental designs:

- Allow for reusing previously gathered information (both yours and others).
- Avoid unnecessary repetition by providing proven methods for experimental design, data gathering, and analysis.
- Make it easy to combine data from various studies.
- Enhance the possibility of collaboration and information sharing.

Data Ownership and Responsibility: Clearly define who owns the data and who is responsible for its accuracy, security, and proper usage. This helps avoid confusion and ensures accountability.

Data Quality: Set standards for data accuracy, completeness, and reliability. Regularly monitor and address data quality issues to maintain reliable and trustworthy information.

Data Security: Establish protocols for protecting sensitive and confidential data.

Data Documentation: Require complete records of data sources, definitions, and transformations. This helps us understand the data better.

Data Governance: Set up a clear way to make decisions about data. This means making roles for people who take care of data, creating steps to follow, and making rules for managing data.

Data Privacy: Follow the rules about privacy, and make sure you collect, store, and use personal or sensitive data in the right way as per the laws.

SUMMARY

- Data annotation tools used in the field of data management and AI.
- It covers popular tools like Colabler, Labelbox, CVAT, and VGG Image Annotator.
- Documentation Annotation focuses on annotation tools specifically designed for document-based data, including Annotation using Word and annotation using Adobe.
- Data Security measures to protect data from unauthorized access and breaches.
- Data Privacy addresses the need to protect sensitive and personal information in data.

Check Your Progress

A. **Multiple choice questions**

1. Which of the following is a popular open-source tool used for data annotation? (a) Colabler (b) Labelbox (c) CVAT (d) All of the above
2. What is the primary purpose of using the VGG Image Annotator? (a) Text annotation (b) Image classification (c) Audio annotation (d) Video annotation
3. Which tool supports easy and fast prototyping, as well as neural networks and recurrent networks? (a) Labelbox (b) Colabler (c) CVAT (d) Keras
4. Which tool supports image classification, bounding boxes, polygons, curves, and 3D localization? (a) Colabler (b) CVAT (c) Labelbox (d) VGG Image Annotator
5. Which tool is known for supporting various data formats like images, videos, and 3D data? (a) Colabler (b) CVAT (c) VGG Image Annotator (d) Labelbox
6. Which data annotation tool supports various annotation methods, including bounding boxes, lines, polygons, and curves? (a) Colabler (b) Labelbox (c) CVAT (d) VGG Image Annotator
7. Which data annotation tool is known for its AI-enabled data labeling solutions and API support? (a) Colabler (b) Labelbox (c) CVAT (d) VGG Image Annotator
8. Which tool is designed to annotate images, texts, and videos? (a) Colabler (b) Labelbox (c) TensorFlow (d) CVAT
9. Which tool supports creating accounts or logging in using Google or GitHub authentication? (a) Colabler (b) Labelbox (c) CVAT (d) VGG Image Annotator
10. Which tool enables custom task plugins and supports exporting in PascalVoc XML format? (a) Colabler (b) Labelbox (c) CVAT (d) VGG Image Annotator

B. **Fill in the blanks.**

1. Labelbox is a popular platform used to create high-quality ___________ datasets.
2. CVAT is an open-source tool designed for ___________ and ___________ of data.
3. VGG Image Annotator is a free tool developed by the Visual Geometry Group for annotating and ___________ images.
4. Data annotation involves adding ___________ to raw data to make it understandable for machines.
5. Image bounding boxes are rectangular regions drawn around objects of interest to indicate there ___________ in an image.
6. Polygon annotation involves creating and outlining irregular ___________ shapes around objects in images.
7. Key points annotation involves marking specific ___________ on objects, often used for human pose estimation.
8. Text entity recognition is the process of identifying and Labeling specific ___________ in text data, such as names or dates.
9. Semantic segmentation is an annotation technique where each pixel in an image is assigned a ___________ label.

10. CVAT is an open-source tool used for annotating and Labeling data, particularly in the context of ____________ tasks.

C. **State whether true or false**

1. Labelbox is an open-source tool for data annotation.
2. TensorFlow is primarily used for image annotation.
3. Colabler is a proprietary tool for data annotation.
4. CVAT stands for Computer Vision Annotation Tool.
5. VGG Image Annotator is an open-source tool for annotating images.
6. Labelbox is mainly used for text annotation tasks.
7. Microsoft develops TensorFlow.
8. Colabler supports custom task plugins for creating unique labeling tools.
9. CVAT is only available for Windows operating systems.
10. VGG Image Annotator has limited support for different annotation formats.

D. **Short answer questions**

1. What is the purpose of data annotation in machine learning?
2. What types of data can be annotated using CoLabeler?
3. What are the uses of labelbox?
4. Briefly describe the open-source tool CVAT and its primary use.
5. How to create a user registration in the CVAT tool.
6. What is the purpose of the VGG Image Annotator?
7. Write down the steps of annotation using Word.
8. Write down the steps of annotation using Adobe.
9. What are the various types of annotation performed by using colabeler
10. What kinds of services does the labelbox tool offer?

Module 2 Data Curation and Labeling

Module Overview

In this unit, you will explore Data Curation, which is about choosing and organizing data neatly. You will discover its significance, data curation working, its key characteristics, and the tools used for this task. Then, you will move on to Data Labeling, where you can add labels to data to help machines understand it better. You will explore its importance, the step-by-step working process, and the tools used for it.

After that, you will get knowledge about Data Annotation in AI, which is like giving extra information to businesses. We will explore its role, the services in business it offers, and data errors in annotation. Lastly, this unit introduces you to some handy open-source tools like CVAT and LabelMe, which make data annotation easy. By the end of this chapter, you will be ready to prepare data for AI projects.

Learning Outcomes

After completing this module, you will be able to:

- Understand the importance of data curation and its role in maintaining data quality for AI systems.
- Learn the process of data labeling and its impact on improving the accuracy of machine learning models.
- Explore the significance of data annotation in AI systems and how it enables machine learning models to perform effectively.
- Evaluate various open-source tools available for data annotation and their practical applications in AI projects.

Module Structure

Session 1. Data Curation

In a quiet village, there lived a girl named Reeba who loved books. She dreamt of creating a library with the best collection of books for her fellow villagers. However, she faced a challenge – finding and organizing the books. One day she heard about something called "Data Curation." With data curation, Reeba learned how to collect, organize, and catalog books efficiently. She created a digital catalog, making it easy for everyone to find their favorite books. Reeba's library became a hub of knowledge, all thanks to the magic of data curation, which showed how valuable it is in managing information for the benefit of everyone as shown in Figure 1.1

Figure 1.1: Reeba's digital library

In this chapter, you will learn about the concept of data curation, uses of data curation, types of data used for data curation, and tools of data curation.

1.1 Data Curation

Data curation for AI means cleaning, selecting, and arranging data to make it acceptable for use in AI and machine learning. Data curation aims to arrange correct, useful, and high-quality data for learning. Data curation means removing duplicate data, fixing mistakes, filling in missing parts, and making sure that data is consistent. Data curation makes sure AI systems get high-quality data, so they can make the right predictions and provide useful results.

1.1.1 Data Curation in Machine Learning

In machine learning, data curation means finding, arranging, annotating, improving, and maintaining data. It is really important because it helps to make good datasets that we use to train, test, and make sure machine learning models work well.

To get data ready for machine learning, you need to start planning even before you get the data. Once you have the right data, you can start getting it ready for training. This means making sure the data is set up in the best way for the computer to understand it. This step has four parts-

1. **Formatting-** Data often comes in different ways. Formatting means putting it all in one way, like putting different types of customer information together in one list.

2. **Labeling-** Labels are like tags that help the computer to understand the data. For example, if we are teaching a computer to drive a car, we need to label the data with things like "car," "people," "traffic signs," and "sidewalks."
3. **Data Cleaning-** Sometimes information is missing in the data. We fix these mistakes and fill in the missing parts.
4. **Feature extraction-** Start examining the data, and choose the important features. This makes the computer learn faster and use less memory.

1.1.2 Stages of Data Curation

The process of Data Curation has three main steps that happen one after the other-

- **Preserving-** Collecting data from different places and then organizing and managing it properly is called Preserving.
- **Sharing-** Ensuring that data can be easily found and used by verified people in the future.
- **Discovering-** Reusing data in new ways and combining it to create something new is part of the discovering step.

1.1.3 Modes of Data Curation

Data curation can be done in two ways: manual and AI-based.

- **Manual Curation:** Manual data curation means hiring a person to collect and organize data. It can be expensive, take a lot of time, and have errors. It can also be difficult to maintain and challenging for the person doing it.
- **AI-based Curation:** AI-based curation means using a special computer program to perform the curator's job. It is not faster and more efficient. With AI tools, handling large and complicated data is easy, and the curator does not need to stress about it.

1.1.4 Importance of Data Curation

Data curation is important because it helps keep all the important information organized. Without it, people might struggle to find the data they need, and they might not feel confident about its accuracy. Companies and employees can face problems without data curation, such as:

- Data remains unused even though it is needed.
- Poor data quality
- Missing, old, or unorganized information about the data.
- Unorganized data.

1.1.5 Uses of data curation

Data curation is used in various ways to enhance the quality and usefulness of data for different purposes-

- **Machine Learning-** Data curation helps in preparing clean and well-structured datasets for training machine learning models. It ensures that the data is accurate, relevant, and labeled properly, leading to better model performance and predictions.
- **Research and Analysis-** In research, data curation ensures that the collected data is reliable and can be used for drawing accurate results.
- **Data Preservation-** Curating data is important for preserving historical, scientific, and cultural information for future generations.

- **Data Sharing-** Curated data is shared and used by others, promoting collaboration and knowledge exchange.
- **Educational Purposes-** Curated data can be used in educational settings to teach students about real-world scenarios, research methods, and data analysis techniques.

1.2 Types of Data Used for Data Curation

Several types of data are involved in the data curation process-

- **Structured Data**- This type of data is organized into tables, rows, and columns, making it easy to store and analyse. Examples include databases and spreadsheets.
- **Unstructured Data**- Unstructured data does not have a fixed format and can include things like text, images, audio, and video. It requires special methods to organize and make sense of.
- **Metadata**- Metadata is data about data. It provides details about data creation and its representation.
- **Time-Series Data**- This type of data is collected at specific time intervals, like stock market prices, weather data, and sensor readings.
- **Geospatial Data**- Geospatial data includes information related to locations on Earth's surface, like GPS coordinates, addresses, and maps.
- **Categorical Data**- Categorical data consists of distinct categories or groups, like types of fruits, colors, or genders.
- **Numerical Data**- Numerical data consists of numbers and can be further categorized into discrete (whole numbers) and continuous (decimal numbers) data.
- **Text Data**- Text data includes written words, sentences, and paragraphs. It is often found in documents, social media posts, and articles.
- **Image Data**- Image data includes visual information captured in pictures or graphics.
- **Audio Data**- Audio data consists of sounds and recordings, like music, speech, or other audio clips.

1.3 Working Process of Data Curation

The data curation working process includes a set of actions, starting with collecting data, and then preparing, cleaning, and improving it. If data is well-organized, machine learning models work better and can be applied to data they have not seen before.

Here's a simple working process of data curation. As illustrated in Figure 1.2-

- **Data collection-** Data is collected from different places like databases, websites, social media, and even devices like IoT. The data can be of different types, such as text, images, or organized data i.e. files and databases.
- **Data cleaning-** This involves fixing any missing information, removing duplicates, dealing with unusual values, and making sure the data is all in a consistent format.
- **Data annotation-** In this, we need to add labels to the data.
- **Data transformation-** After the data is cleaned and labeled, we might need to change its format so that machine learning can understand it better.

- **Data integration-** We collect data from various places and must ensure that it fits together in a meaningful way. This might involve arranging data sequentially or merging datasets that share common information.
- **Data maintenance-** With time, data might have to be changed or added to with new data. Keeping the dataset up-to-date makes sure it stays helpful for machine learning tasks that continue.

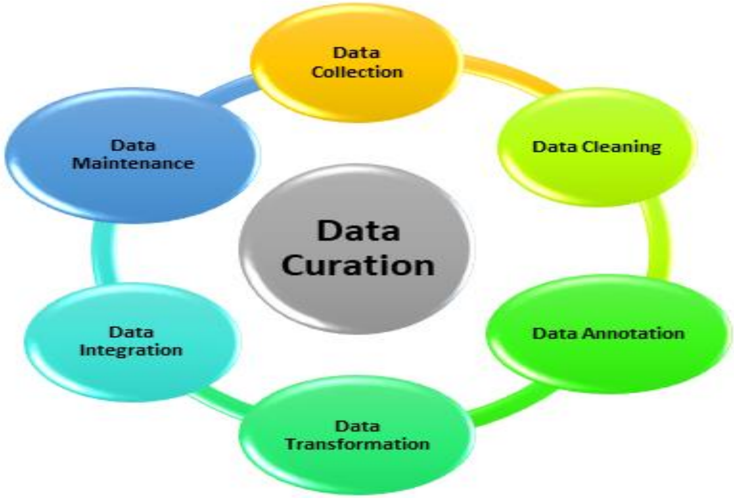


Figure 1.2: The working flow of Data Curation

1.4 Characteristics of Data Curation

- **Identify patterns-** Data curation helps to understand patterns in information more easily.
- **Data Management-** Data curation makes it easier to manage data. It helps in taking care of data to keep it safe and valuable.
- **Supports data governance-** Data governance authority makes rules for how to handle data.

1.5 Activities of Data Curation

- **Adding Context-** Adding context means adding extra information, like sources and credits, to the dataset. This helps us understand the source of data and its use.
- **Validating and Adding Information-** We validate and add metadata, and organize the details about a dataset in a way that computers can understand. This helps us to find the data more easily.
- **Citing the data**- Data users should give credit to the data creators if they use the data.
- **Protecting Privacy**- Private or personal information is removed or changed so that no one can know who it belongs to. This keeps a person's information safe and confidential.
- **Data Verification-** A data creator examines the dataset. This is done to make sure that the data is correct and accurate.
- **De-identification:** De-identification means taking away or covering up personal information to keep it private.

1.6 Different types of data available for the data curation process

Data curation means collecting and arranging data so that a business can easily use and understand it. This process includes handling both the actual data and extra information about it, known as metadata.

- Text
- Image
- Audio
- Video
- Scientific measurements
- Education data
- Sales data
- Health care data

1.7 Keywords to identify various sources of data sets

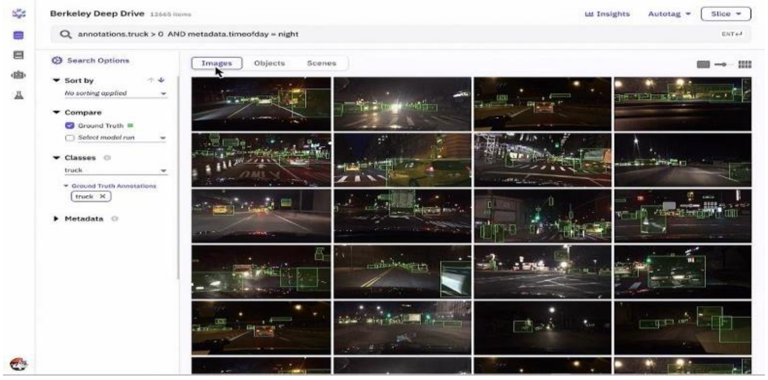
- Feature Selection and curation for Finance data
- Signature image cleaning
- Medical image cleaning
- Chemical information cleaning
- Heart disease prediction using EEG curation
- Vaccines Impact based on data curation
- Data Wrangling
- Data cleaning and curation before prediction
- Geospatial analysis of data with data curation
- Weather data curation
- Customer purchase behavior curation

1.8 Sample datasets available

- Fisher's Iris Flower Dataset,
- Labeled Faces in the Wild Dataset
- Google Data set - https://datasetsearch.research.google.com/
- Paper with code data set - https://paperswithcode.com/datasets
- Visual data - https://visualdata.io/discovery
- Data Hub – data set - https://datahub.io/
- U.S. Govt. open data set - https://data.gov/

1.9 Tools of Data Curation

1. **Scale Nucleus-** In 2020, Scale introduced a tool called Scale Nucleus. It works with image data. It also has useful features like finding specific details about the data, finding mistakes with easy-to-understand measures, and connecting with their API (but it cannot be used on your servers yet). The picture representation of the Scale Nucleus is shown below in Figure 1.3.



annotations.truck > 0 AND metadata.timeofday = night

Sort by: No sorting applied

☑ Ground Truth

□ Select model run

Classes: truck

Ground Truth Annotations: truck ✕

Figure 1.3: Screen capture from Scale Nucleus

Advantages-

a) It connects directly with Scale's labeling team.

b) It can see predictions and labels to find where the model is not working well.

c) It can search for similar images in the dataset.

d) It is easy to use.

e) It has a user-friendly interface.

Disadvantages-

a) It does not work as well with big datasets that do not have labels.

b) You can only use it within the scale system.

c) You have to choose data by hand using the user interface.

d) It is time-consuming.

e) It lacks data security.

2. **Labelbox-** Labelbox's solution concentrates on the part of the machine learning process where you keep improving the training data. The platform is set up to do three main things- adding labels to data, figuring out the performance of the model, and deciding the data to use next. It also makes it easy for teams to work together on projects, which is useful for teams that are far apart. The picture representation of Labelbox is shown below in Figure 1.4.

Advantages-

a) It works directly with Labelbox's labeling tool.

b) It can use embeddings to group similar images.

c) It can be used on your computer

d) It is easy to use.

e) It is a user-friendly software

Disadvantages-

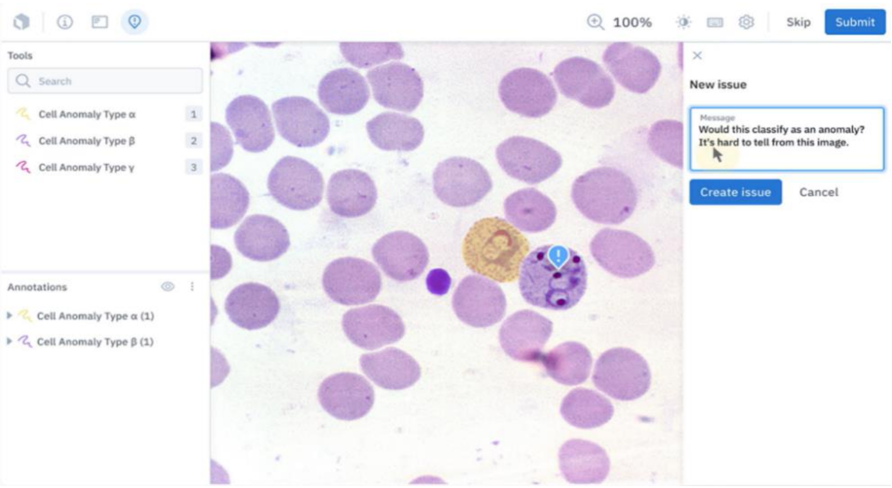
a) Need to choose data manually using the user interface.

b) It is time-consuming.

c) It lacks data security.

d) It suffers with slow speed.

e) It is costly.



Would this classify as an anomaly? It's hard to tell from this image.

Figure 1.4- Screen capture from Labelbox

3. **Lightly-** Lightly is a software that helps organize and improve data for computer vision. Unlike other tools, it can handle a large number of images. It uses a special kind of learning to group similar data in a dataset. With Lightly's methods, it can make a balanced and good set of data for training models. This helps models learn better and avoids problems like overfitting and biases that can make models fail. The picture representation of Lightly is shown below in Figure 1.5.

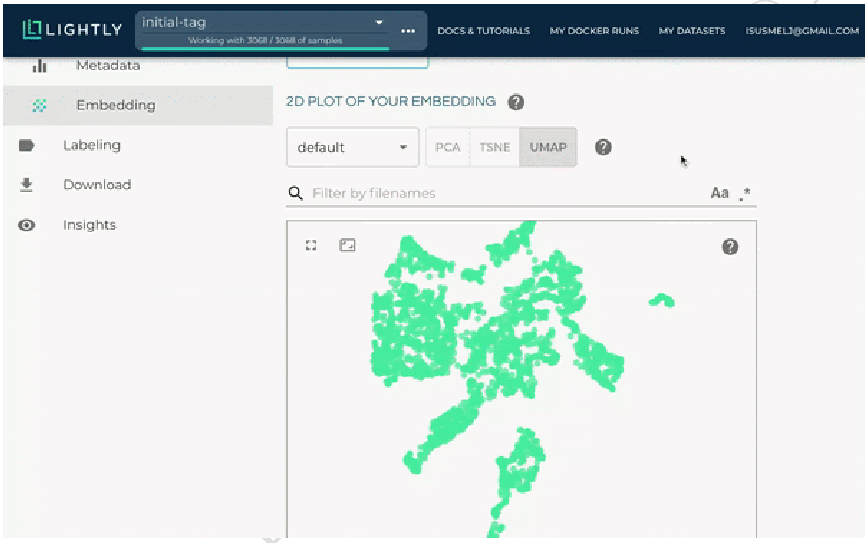


Figure 1.5. Screen capture from Lightly

Advantages-

a) It can pick data using smart learning techniques.

b) It works with videos and can handle lots of frames in a dataset

c) It can be used on your computer

d) It works without needing labels

e) It is easy to use.

f) It is user-friendly software.

Disadvantages-

a) It concentrates more on data than fixing model issues.

b) It is time-consuming.

c) It lacks data security.

d) It suffers from low dataset quality

e) Its speed is slow.

1.10 Real-world applications of Data curation

Here are four real-world examples of data curation in machine learning-

Autonomous vehicle- Autonomous vehicles use machine learning with data like pictures and videos from cameras. People mark these pictures to show things like pedestrians, other cars, and traffic signs. They remove bad pictures and change them to a suitable format for training. This makes autonomous vehicles safe and dependable. As illustrated in Figure 1.1.

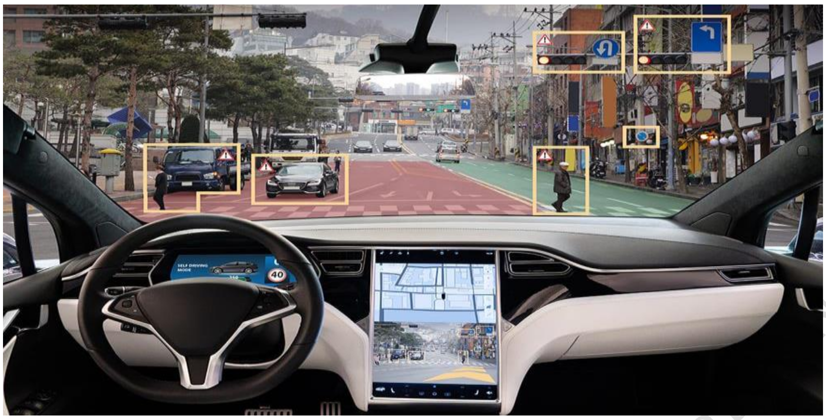


Figure 1.6: Autonomous vehicle (source: https://bernardmarr.com/)

Scientific Research- In science, data curation is like taking care of important information. Scientists use it to keep, organize, and share their research data. This helps them find and use the data again, which can lead to discoveries and progress in their field of study. As illustrated in Figure 1.7.

Figure 1.7: Scientific research (source: https://slate.com/)

ImageNet- ImageNet is a big collection of pictures used to teach computers to recognize things. It has more than 14 million pictures that people carefully labeled and grouped into categories. These categories are like folders that hold similar things. ImageNet helps computers get good at understanding pictures, and it shows well-organized and labeled data can help make machine learning better. As illustrated in Figure 1.8.

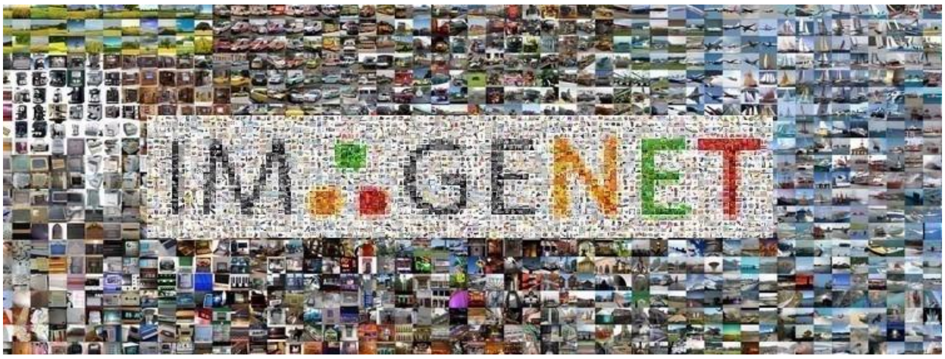


Figure 1.8: ImageNet (source: https://cv.gluon.ai/)

Healthcare- In healthcare, making sure data is clean and useful is very important. This includes data from things like medical records, wearable devices, and medical pictures. As illustrated in Figure 1.9.

Figure 1.9: Healthcare (source: https://www.gehealthcare.com/)

Financial Services- In the financial world, data curation plays a role in managing money-related activities such as investments and loans. It ensures that all financial information is stored securely and can be reviewed, if necessary. This helps prevent fraud and adds transparency to the financial industry. As illustrated in Figure 1.10.

Figure 1.10: Financial services (source: https://www.pwc.com/)

Twitter sentiment analysis- Sentiment analysis is a standard task in natural language processing. For example, a collection of tweets are collected, and people want to see if they're happy, sad, or just okay. First, they clean up the tweets (like deleting duplicate data or dealing with missing stuff), then they say if each tweet is positive, negative, or in-between. After that, they change the words into numbers so computers can understand. As illustrated in Figure 1.11.

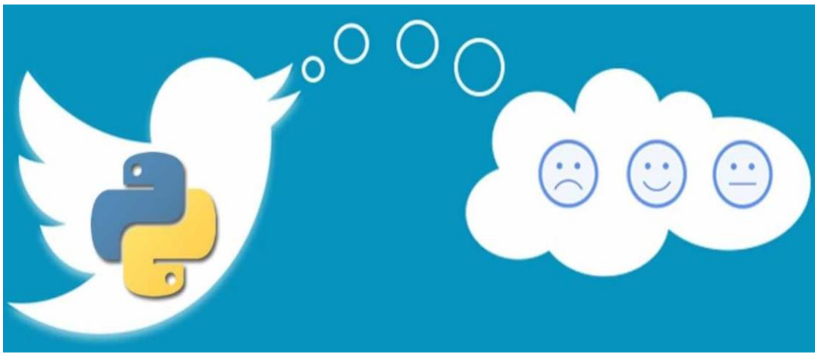


Figure 1.11: Twitter sentiment analysis (source: https://iliyaz.hashnode.dev/)

Government- In government, data curation helps keep important records like census data, legal documents, and historical papers safe and organized. This makes sure that this information can be easily used and understood by people in the future.

1.11 Challenges of data curation

Data curation is important for businesses, but it also has some challenges as follows:

a) **Data Accuracy-** Getting accurate data is one of the biggest challenges in the data cycle. If the main data source is not right, everything else that depends on it will also go wrong. This can lead to bad decisions that can hurt a business and its owners.

b) **Dealing with huge data-** Dealing with huge data can be challenging in data curation. Companies are making lots of data, so it is a big task to handle all of it.

c) **Data diversity-** Another challenge in data curation is dealing with different types of data. Data can be of various kinds, like organized data, unorganized data, and partly organized data. This means companies need to be ready to manage and curate all these different types of data properly based on what they need.

d) **Data Quality-** Data quality is an important challenge in data curation. It means that the data needs to be correct, not confusing, and updated. This is important because if the data is not accurate, it might lead to wrong results that are used for tasks like machine learning.

e) **Data security-** Data security is a big problem in data curation. It means that organizations have to make sure that their important data is not seen or stolen by hackers. To prevent this, organizations need to use strong security methods to make sure their data is safe all the time.

1.12 Benefits of data curation

a) It improves data quality.

b) It improves model performance.

c) It helps us understand data more effectively.

d) Data curation helps in finding and removing duplicate data.

e) It improves efficiency.

f) It reduces storage costs.

g) It increases organizational productivity.

h) It safeguards important data resources.

i) Data curation enhances customer connections and boosts customer loyalty.

1.13 Drawbacks of data curation

There are several drawbacks or challenges associated with data curation-

a) **Time-Consuming-** Data curation can be time-consuming, especially dealing with large and complex datasets. It requires thorough cleaning, organization, and validation.

b) **Data Privacy-** Curation involves handling sensitive data, and raising concerns about privacy and security. Proper measures must be taken to protect personal or sensitive information.

c) **Data Loss-** During the curation process, there's a risk of losing or altering valuable data, especially if not handled carefully.

d) **Data Volume-** As data grows exponentially, managing and curating large volumes of data becomes increasingly difficult.

e) **Changing Data-** Data is not static; it changes over time. Keeping curated data up-to-date and relevant requires ongoing effort.

Summary

- Data curation is the process of managing and organizing data for various purposes.
- Data curation can be done manually or using AI-based tools.
- Data curation involves stages like preserving data, sharing it, and discovering its value.
- Data curation is essential for ensuring data quality and reliability.
- Data curation is used in machine learning, research, data preservation, sharing, and education.
- Data curation involves stages such as data collection, cleaning, annotation, transformation, integration, and maintenance.
- Data curation involves managing data with attention to accuracy, relevance, and completeness.
- Tools like Scale Nucleus, Labelbox, and Lightly are used for data curation.

Check Your Progress

A. Multiple Choice Questions (MCQ)

1. What is data curation? (a) Creating new data (b) Organizing and managing data (c) Deleting data (d) Storing data in a random order
2. Why is data curation important? (a) To generate new data (b) To organize data for analysis (c) To hide data from others (d) To make data inaccessible
3. What does data curation involve? (a) Mixing up data from different sources (b) Storing data in one place without any organization (c) Organizing, managing, and preserving data (d) Deleting all data to free up storage space
4. What is the goal of data curation? (a) To create as much data as possible (b) To make data confusing and disorganized (c) To make data valuable and usable (d) To delete all unnecessary data

5. What is the main purpose of data curation for researchers and scientists? (a) To make data difficult to access (b) To keep data private (c) To ensure data is reliable and shareable (d) To delete data after a project is complete
6. What is the purpose of preserving data in data curation? (a) To forget about the data (b) To keep data safe and accessible for the future (c) To modify data for personal use (d) To hide data from others
7. Data curation is essential in which of the following industries? (a) Food and beverage (b) Entertainment (c) Healthcare and medicine (d) Sports
8. Which of the following is a drawback of poor data curation? (a) Enhanced data accessibility (b) Time-consuming and data loss (c) Faster data processing (d) Increase collaboration among researchers
9. What is a popular open-source tool used for data curation? (a) Cooking app (b) Social networking platform (c) Fitness tracker (d) Labelbox
10. Which of the following scenarios requires data curation? (a) Organizing a music playlist (b) Sorting a drawer of random items (c) Keeping a diary of personal thoughts (d) Storing patient medical records

B. Fill in the blanks

1. Data curation is a process of collecting, organizing, managing, and __________data.
2. When data is properly _____________, it becomes more reliable.
3. Data __________ involves assigning labels to data.
4. In data curation, data is __________ cleaned and organized to make it organized and useful.
5. __________ is a drawback of data curation.
6. Data curation assists in finding and removing __________ data, making storage and processing more efficient.
7. Organizations need to use strong security methods to make sure their data is ________ all the time.
8. Data curation prevents _________ and adds transparency to the financial industry.
9. ___________ is a standard task in natural language processing.
10. ___________ is a big collection of pictures used to teach computers to recognize things.

C. State whether true or false

1. Data curation involves only the collection of data.
2. Data curation ensures consistent data quality over time.
3. Data annotation is not a part of the data curation process.
4. Effective data curation makes data less accessible for different purposes.
5. Data curation is essential only in the field of mathematics.
6. Data curation helps preserve data integrity and prevent loss.
7. Data curation involves labeling data to make it understandable and retrievable.
8. Data curation is not important in scientific research.
9. Data curation does not involve handling missing or incomplete data.

10. Data curation is primarily focused on collecting as much data as possible without any selection process.

D. Short Question Answer

1. What is data curation?
2. What are the activities of data curation?
3. Explain any two tools of data curation.
4. What are the uses of data curation?
5. What are the challenges of data curation?
6. Explain any two real-life applications of data curation.
7. What are the characteristics of data curation?
8. Explain the working process of data curation.
9. Explain the data curation importance in machine learning.
10. Explain any three types of data used in data curation.

Session 2. Data Labeling

In a quiet village, curious Ajay loved to observe the birds that visited her garden. One day, he heard about data labeling. It was like giving names to things so computers could understand. Ajay realized this could help his village too. He started labeling the fruits and vegetables at his local market. This helped the shopkeepers keep track of their stock. Soon, the village started labeling more things, like crops and animals. This made life easier and improved farming. Ajay's curiosity brought positive change to his village through data labelling as shown in Figure 2.1.

Figure 2.1: Shopkeepers using data labeling (Source: https://www.istockphoto.com/)

In this chapter, you will learn about the concepts of data labeling and data labeling tools.

2.1 Data Labeling

Data labeling is like giving names or labels to different kinds of information, like pictures, videos, text, or sounds. These labels tell us what category the data belongs to. Data labeling step in machine learning where we mark things in raw data like images, videos, sounds, or text. We put labels on them to help the machine-learning model make correct guesses and estimates. This helps a machine learning model learn how to recognize those categories if it sees similar data without labels.

2.1.1 Labels and Features in Machine Learning

1. **Labels-** Labels, also called tags, are like name tags for pieces of data. They help us know what that data is all about. Labels are like the final answers that the model predicts. For example, in Figure 2.2., we might have labels like "cat" or "dog." In audio, labels could be the words spoken. These labels help the ML model learn from the dataset. We train a model using supervised techniques, we give it a dataset with labels. With this labeled training data, the model can make accurate predictions, if it is given new data to test.

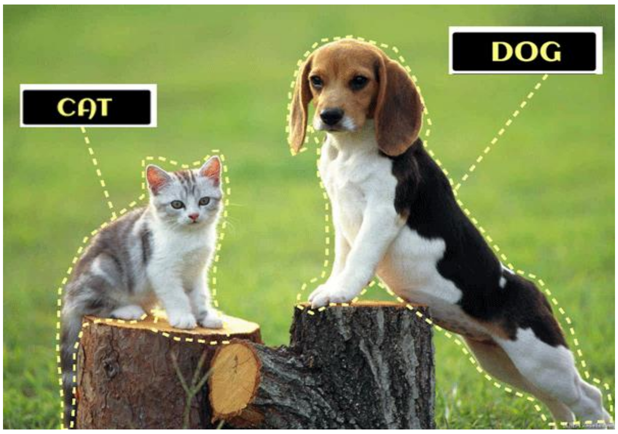


Figure 2.2- Example of Labels (Source- Javatpoint)

2. **Features-** Features are the pieces of information that ML systems use to understand and make predictions. Each piece of information is like a building block for the ML model. Each column in a dataset is a feature. We can create new features by using the existing information, called feature engineering. By considering the same example shown in Figure 2.2. The features are height, weight, ears, eyes, nose, shape, size, etc.

2.1.2 Importance of data labeling

Labeling data is important because the machine needs to learn from examples. This process is called supervised learning. Labeled data has both the input and the correct output labeled. This helps the machine learn to categorize things correctly. Machines learn from labeled data, they can recognize patterns in new, unorganized data.

For example, shown in Figure 2.3, we have images of various animals such as cats, and dogs. Our goal is to create a machine-learning algorithm that can identify these animals. By providing the model with labeled data, it can easily identify and categorize the images into their respective groups – cats and dogs.

However, if the images are not labeled, the machine learning model must independently identify each image's different characteristics, like color, body shape, facial features, and other details.

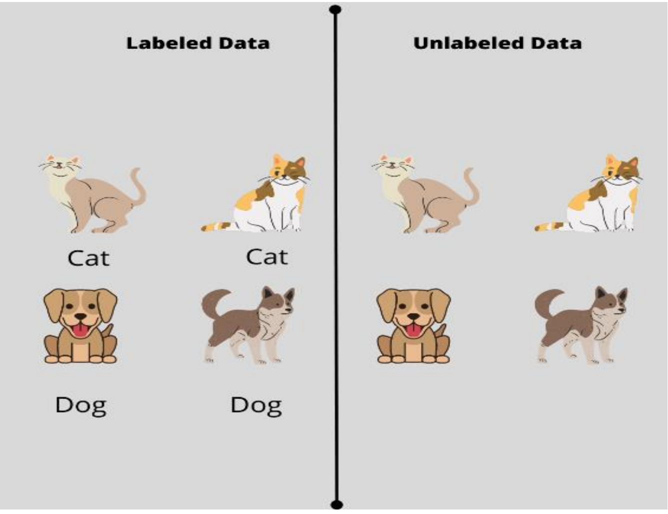


Figure 2.3 - Example of Labeled and Unlabeled Data

2.1.3 Labeled data

Labeled data is any data that has been given a characteristic, category, or set of attributes. Examples of labeled data include a picture of a cat, a person's height, and a product's pricing.

2.1.4 Unlabeled data

Unlabeled data means information that does not have any descriptions, tags, or labels attached to it. It is like having a collection of pictures without knowing the details of any picture or having a list of words without knowing their meanings.

2.1.5 Labeled data vs. Unlabeled data

To teach machine learning methods, computers need data that has been categorized, but there are differences between labeled and unlabeled data-

Labeled Data	**Unlabeled Data**
Data with labels or categories.	Data without labels or categories.
Example: Photos of cats and dogs labeled as "cat" or "dog".	Example: Photos with no labels.
Used to train computers to recognize and classify objects.	Used to explore data patterns and clustering.
Used for supervised learning.	Used for unsupervised learning.
Requires humans to label or categorize the data.	Does not require humans to add labels.

2.2 Working process of Data Labeling

Data labeling is a process that helps computers understand and learn from data. The working process of data labeling is shown in the following-

Step 1- Data Collection- The first step in any machine learning project is getting the correct raw data, like pictures, sounds, videos, and words.

Step 2- Preparing for Labeling- Before labeling, you need to decide what you want the computer to learn from the data. For example, if you are working with images of animals, you want the computer to recognize different animals.

Step 3- Data annotation- Data annotation is like tagging data with labels. Experts add these labels to help the model learn and make predictions, especially in pictures.

Step 4- Quality assurance- The data needs to be good, trustworthy, exact, and consistent. The quality of the data depends on carefully adding labels to each piece of data. Experts also regularly review and correct any errors in a process. This process is called Quality Assurance (QA) and it helps to keep the data accurate and reliable.

Step 5- Model training and testing- Training the model with labeled data helps it make predictions. Testing it with unlabeled data checks its accuracy. For example, if it is not right 96% out of 1000 times, it is not good.

2.3 Types of Data Labeling

There are numerous ways to label data, and the method used depends on the data and how the labeled dataset will be used. Some of the most popular ways to label data are-

a) **Binary data labeling-** This process includes giving a label of "true" or "false," or '1' or '0,' to each data point. For example, in a dataset with animal images, a label of "true" could mean the image has a cat, and "false" if it does not.

b) **Multiclass data labeling-** This usually means giving labels to more than two categories in the dataset. For example, in a dataset with animal images, each image might be labeled with the names of different animal species like "cat," "dog," "bird," and so on.

c) **Multi-label data labeling-** This involves giving more than one label to each data. For example, in a dataset with animal images, each image might be labeled with the names of the animals in the image and the environment they're in (like "cat in a grassy field"). This is often used for text-to-image projects.

d) **Semantic data labeling-** This includes adding detailed explanations to the data. For example, in a dataset with animal images, each image might have detailed descriptions of the animal's looks, behavior, and surroundings.

e) **Structured data labeling-** It means labeling data in an organized way, like using a table or database. For example, a dataset with customer feedback might have labels for different topics in each feedback, along with the customer's name and contact details.

f) **Unstructured data labeling-** This means labeling data that is not structured in an organized way. For tasks like NLP and speech recognition, a dataset of audio recordings might have labels with written versions of the spoken words in each recording.

2.4 Human-in-the-loop

Human-in-the-loop (HITL) means people and computers working together to train and test a smart program, improving it with human guidance. HITL is used if the computer cannot figure out a problem on its own.

2.4.1 Human-in-the-loop in Machine Learning

Good data is important for teaching machines to make accurate decisions, and here Human-in-the-Loop machine learning helps. Human-in-the-loop means combining human and machine intelligence in a continuous cycle. The working process of Human-in-the-Loop is as follows:

Training: Humans provide the machine with data and teach it how to make decisions. They help the machine learn the right criteria.

Testing: The machine tries to make decisions on its own, but sometimes it makes mistakes. Humans check their work and see where it goes wrong.

Validation: The machine is tested again to make sure it is making the right decisions.

2.5 Synthetic data

Synthetic data is "made-up" data created by computers instead of being collected from the real world. It is used to teach machines or test computer programs. Imagine making a fake recipe with pretend ingredients to practice cooking. Synthetic data helps computers learn or practice without using real information.

2.5.1 Synthetic data in data labeling

Synthetic data plays a helpful role in data labeling by providing more examples for training and testing as shown below:

Enhancing Training: You are teaching a computer program to recognize things, like cats in pictures, and you need many examples. If you do not have enough real pictures of cats, you can create fake ones. These artificial examples help the computer learn better.

Testing and Validation: After training, make sure the computer can still recognize things correctly. Synthetic data helps in this stage too. You can use it to test if the computer can spot cats, even if it has never seen those exact pictures before.

Diverse Scenarios: Sometimes, you want to teach the computer to recognize things in different situations, like cats in rainy weather or at night. Real-world data might not cover all these scenarios, so synthetic data can fill the gaps.

2.5.2 Data Labeling Tools

Open-source tools- These tools are available for anyone to use for free, although there are some restrictions for commercial purposes. They are really useful for learning about machine learning and AI, working on personal projects, and trying out early business applications of AI.

1. **Labelbox-** Labelbox is a well-known platform for data labeling that enables teams to handle, annotate, and collaborate on data. It comes with an easy-to-use interface for labeling various types of data like images, text, and videos. Labelbox supports different ways of annotation, like drawing bounding boxes, and polygons, and making classifications. It provides various image annotation options, including object detection, semantic segmentation, and image classification. For text, Labelbox allows you to highlight named entities, sentiment analysis, and text classification. Additionally, the platform also supports annotating videos, which involves tracking objects as they move and identifying actions taking place in the videos. The screen of the labelbox is shown in Figure 2.4(a) and Figure 2.4(b) below.

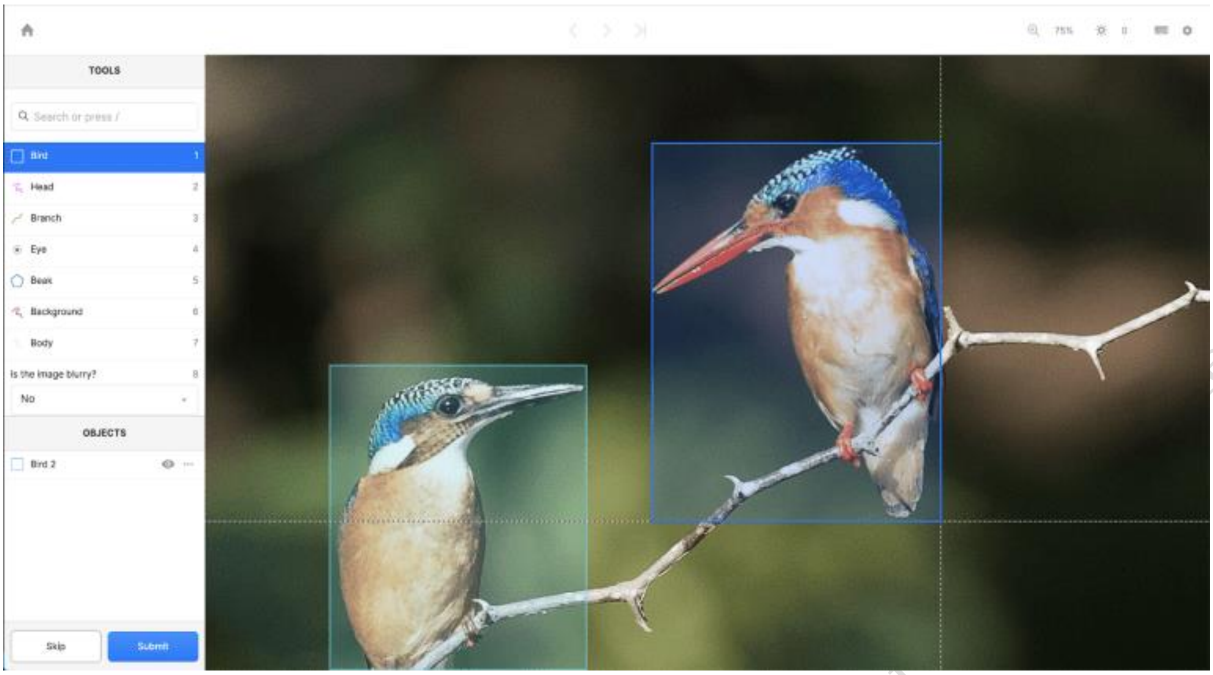


Figure 2.4(a)- Screen capture from Labelbox (Source- Labelbox and neptune.ai)

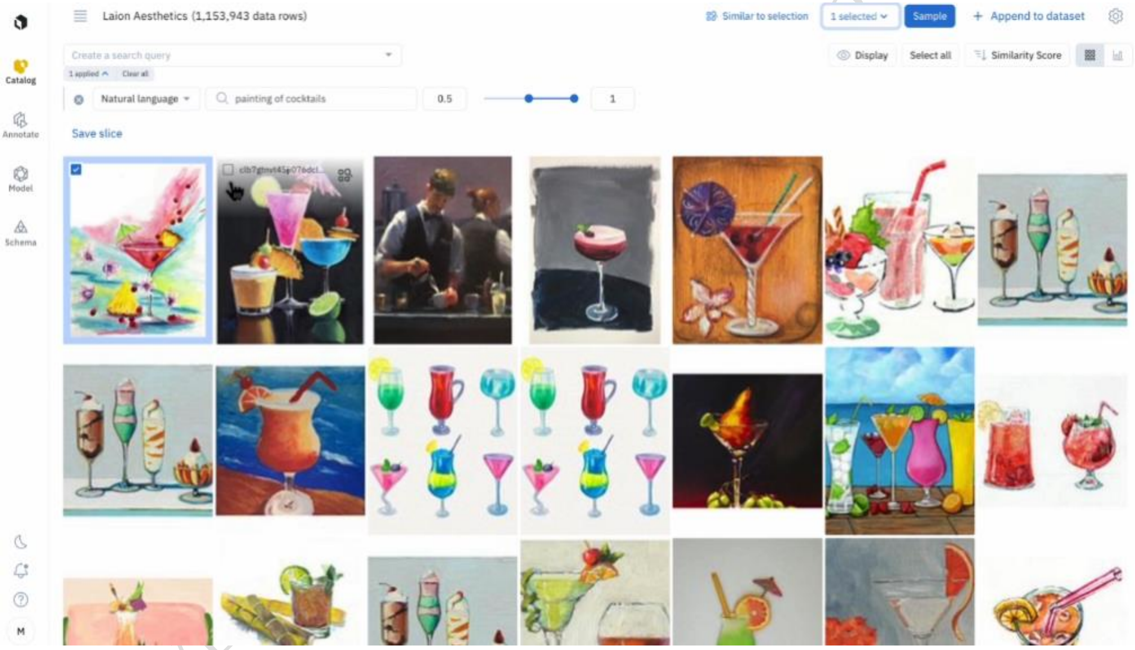


Figure 2.4(b)- Screen capture from Labelbox (Source- Labelbox and neptune.ai)

Benefits of Labelbox

a) Everyone in the machine learning team can work together from one place.

b) Easy to perform tasks with complete communication.

c) Repeat your process with active learning to perform highly accurate labeling and create improved datasets.

d) It is easy to use.

e) It is user-friendly software.

2. **SuperAnnotate-** SuperAnnotate is a data labeling platform that uses AI technology to assist in labeling images and videos. It offers tools to draw bounding boxes, polygons, and segment instances in images and videos. It also allows multiple people to

collaborate and provides various tools for annotating images, such as identifying objects, image classification, and semantic segmentation. It also offers tools for annotating videos to identify actions in the video. The screen of SuperAnnotate is shown in figure 2.5(a) and 2.5(b)

Benefits of SuperAnnote

a) This platform can make smart predictions and learn better as you work, which makes datasets more accurate.

b) It uses tricks from past learning to improve the efficiency of its tasks.

c) You can label things yourself or let the system help, and it checks to make sure the labels are good.

d) It is easy to use.

e) It is user-friendly software.

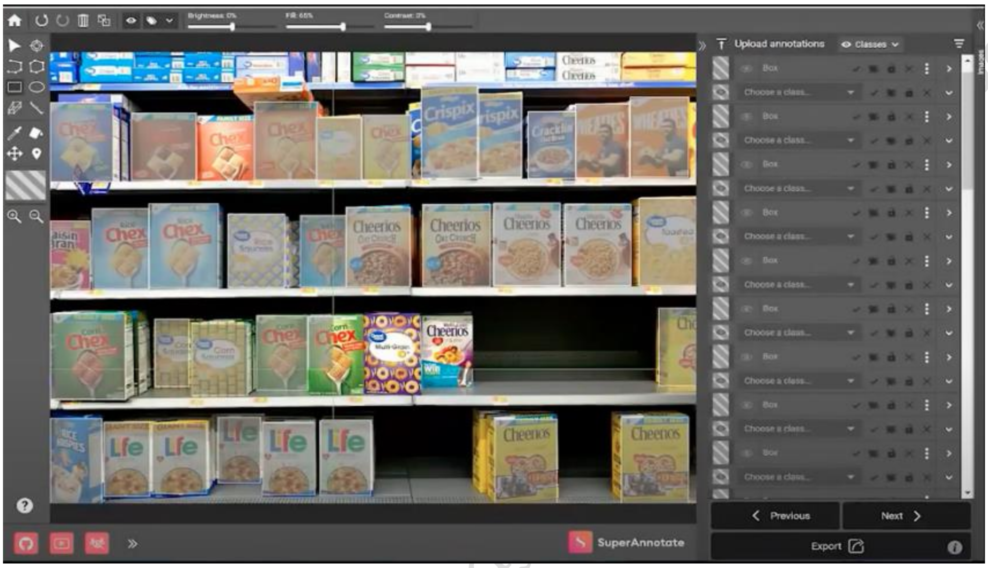


Figure 2.5(a) Screen capture of SuperAnnotate (Source- SuperAnnotate and neptune.ai)

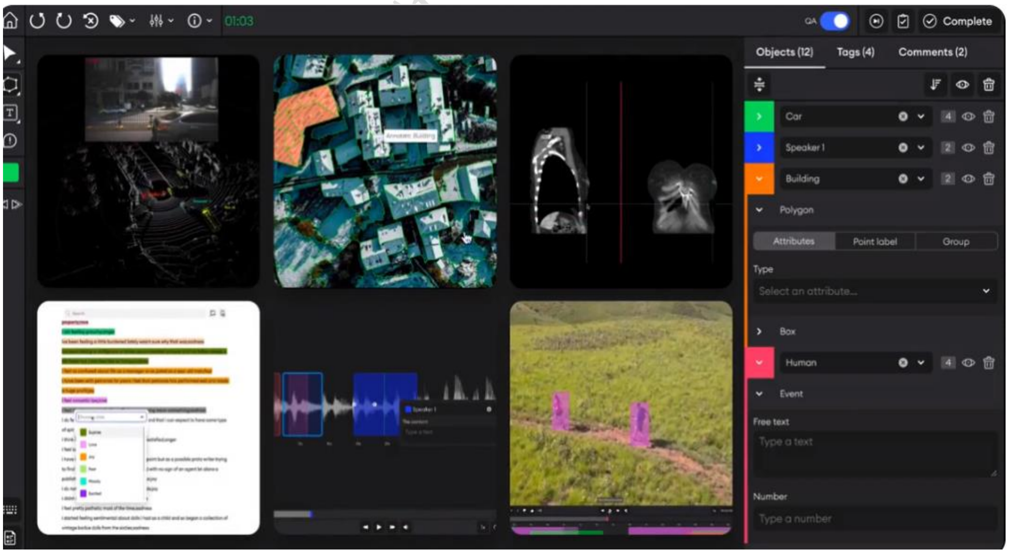


Figure 2.5(b) Screen capture of SuperAnnotate (Source- SuperAnnotate and neptune.ai)

3. **CVAT-** CVAT, which stands for Computer Vision Annotation Tool, is a strong tool i.e. free to use. It is great for labeling images and videos in computer work. It helps in image classification, image segmentation, object detection, and 3D data annotation. It

might be a bit difficult to understand the working of this tool. It is accessible only through the Google Chrome browser. The screen of SuperAnnotate is shown in Figure 2.6.

Benefits of CVAT

a) This tool is free to use.

b) It works well for adding labels to images and videos.

c) It can add labels automatically.

d) It is easy to use.

e) It is user-friendly software.

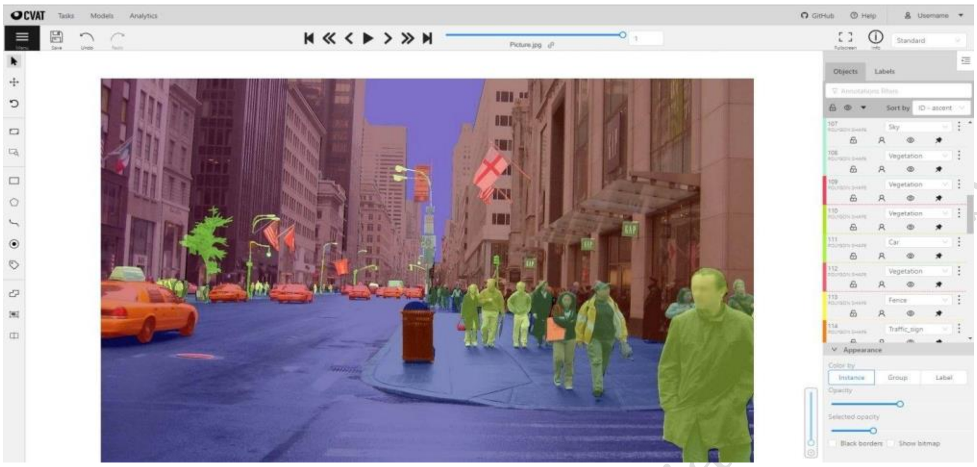


Figure 2.6- Screen of CVAT (Source- CVAT and https://viso.ai/)

2.6 Use-cases of Data Labeling

1. **Computer vision**

In computer vision, you start by marking important parts of images, like drawing boxes or highlighting pixels. This creates a dataset for training. Computers learn to understand images like humans do, by analyzing digital pictures to gather important information. To do this, computers require-

a) **Image annotation** means putting labels on images to describe objects in them.

b) **Video annotation** is about adding labels to frames of a video, which are like pictures taken from the video.

2. **Natural Language Processing**

In natural language processing, you begin by highlighting important text or labeling it. This builds the dataset for training the computer. This helps the computer to understand if the text is positive or negative, or if it mentions a person or place. You can also extract text from pictures, PDFs, or other files by drawing boxes around it.

3. **Audio processing**

Audio processing is used to change different types of sounds, into a way that computers can understand. We label spoken or heard words, organizing them into groups to train computers.

2.7 Data Labeling Approaches

Data labeling is an important part of making powerful machine-learning models. Even though it might seem easy, it is quite important to do it correctly. So, companies that want to use data labeling methods, need to think about many things to find the best way to label data. Here are a few approaches to label data:

a) **Manual Labeling-** In manual data labeling, humans are involved. Data annotation experts are responsible for putting tags or labels on different parts of datasets. Manual labeling means humans add labels to data. This can take a lot of time and mistakes might happen if the person labeling is not careful or familiar with the task.

b) **Automatic Labeling-** Data labeling can be automatic, especially for large datasets with well-known objects. Specially trained machine learning models can add labels automatically, but they work best with accurate initial datasets. However, even with high-quality data, it can be difficult to cover all unique situations or provide the best labels.

c) **In-house data labeling-** In-house data labeling means that the people who work for the company label the data themselves. This way, the labels are accurate and easy to track. But, it takes a lot of time and is best for companies with many resources.

d) **Outsourcing-** Outsourcing is useful if a company hires people from outside to label their data. These people might be freelancers or workers from other companies. This is good for big, short-term projects. But, it can take time to organize and manage this kind of work.

e) **Crowdsourcing-** Crowdsourcing is a quick and cheap way to get data labeled. It uses the internet to give small tasks to many people. These people are usually freelancers on a crowdsourcing website. They help label data, like pictures of plants and animals. This is easier because it does not need special skills. A well-known example of crowdsourcing is Recaptcha.

f) **Synthetic Labeling-** In this way, new project data is made using the already their data. This makes the data better and the work faster. But it needs a lot of computer power and resources, which makes it cost more.

g) **Programmatic labeling-** Programmatic labeling uses a computer program that helps to label the data. This makes things faster and needs less human work because the computer follows a script. But even though it is automated, humans still need to check for any technical issues to make sure it is right.

2.8 Benefits and Challenges of Data Labeling

As a big part of machine learning, data labeling has benefits and challenges too. It helps make predictions right, but it can be expensive. Here are some benefits and challenges of data labeling-

Benefits

a) Precise Predictions- We label data accurately, and help models to get trained with better data. This helps models give the right answers.

b) Better Data for Models- Data labeling makes data more useful for models. For example, we can change categories to yes or no for models to understand better.

c) Data Consistency- In-house tools can provide reliable data consistently over the long term.

d) Feedback system- A feedback loop for annotations guarantees ongoing performance monitoring and enhancement.

e) Auto Data Labeling- Automatic data labeling involves the use of specialized programmatic algorithms in machine learning to handle specific tasks or objects.

Challenges

Data labeling can be tricky, and there are common challenges that people face-

a) It is time-consuming and costly.

b) In this human error occurs.

c) It lacks data security.

d) It suffers from low dataset quality.

e) It has poor workforce management.

SUMMARY

- Data labeling involves assigning labels or tags to data points for machine learning and analysis.
- Labels are target values we want to predict, while features are the data attributes used for prediction.
- Data labeling is crucial for supervised machine learning, as it provides labeled examples for training models.
- Labeled data has known labels or categories, while Unlabeled data means information that does not have any descriptions, tags, or labels attached to it.
- The data labeling process includes data collection, preparation, annotation, quality assurance, and model training/testing.
- Various data labeling types include binary, multiclass, multi-label, semantic, structured, and unstructured data labeling.
- Human-in-the-loop refers to the involvement of humans in the data labeling process, ensuring accuracy and quality.
- Tools like Labelbox, SuperAnnotate, and CVAT assist in data labeling, streamlining the annotation process for better efficiency.

Check Your Progress

A. Multiple Choice Questions

1. What is data labeling? (a) Sorting data alphabetically (b) Adding tags or labels to data (c) Deleting data from a dataset (d) Changing data formats
2. Why is data labeling important for machine learning? (a) It makes data colourful (b) It helps computers see and understand data (c) It creates new data from scratch (d) It adds sound to the data
3. Which type of learning uses labeled data for training? (a) Supervised learning (b) Unsupervised learning (c) Semi-supervised learning (d) Reinforcement learning
4. Which of the following is NOT a common approach for data labeling? (a) Outsourcing (b) Crowdsourcing (c) Handwriting (d) In-house labeling

5. What is the main challenge of data labeling? (a) It is too easy and quick (b) It is expensive and time-consuming (c) It requires no human involvement (d) It never requires any corrections
6. What is one benefit of accurate data labeling? (a) It increases the size of the data (b) It decreases the model's accuracy (c) It helps the model make better predictions (d) It adds complexity to the model
7. Which type of data can be labeled for machine learning? (a) Only images (b) Only audio files (c) Any data, like images, text, audio, etc. (d) Only video files
8. What is the purpose of Quality Assurance (QA) checks in data labeling? (a) To make the data look more colourful (b) To ensure data is never labelled (c) To ensure the accuracy and quality of labeled data (d) To increase the size of the dataset
9. Which type of learning uses both labeled and unlabeled data? (a) Supervised learning (b) Unsupervised learning (c) Semi-supervised learning (d) Reinforcement learning
10. What is the purpose of adding labels to data? (a) To make the data look better (b) To make the data smaller in size (c) To provide context and meaning to the data (d) To hide the data from others

B. Fill in the blanks

1. Data labeling is the process of adding ____________ to raw data.
2. Labeled data is essential for training ____________.
3. In image annotation, ________________ is used to mark objects.
4. Text annotation involves labeling and categorizing ______________.
5. Data labeling helps machines ____________ patterns and features.
6. One type of data labeling is __________________ where each item is labeled.
7. Data labeling helps create ______________ sets for machine learning.
8. Tools like CVAT and Labelbox help in ______________ data labeling.
9. Data labeling is important for helping computers understand and __________________ patterns.
10. Data labeling can be done automatically using ______________ techniques.

C. State whether True or false

1. Data labeling helps computers understand patterns in the data.
2. Data annotation is only applicable to images.
3. Data labeling can be done manually by human annotators.
4. Automatic data labeling does not require any input from humans.
5. Data labeling is not important for training accuracy machine learning models.
6. Data labeling is essential for supervised machine learning tasks.
7. Labeling data involves adding annotations or labels to the data.
8. Data labeling can improve the performance of machine learning models.
9. Data labeling is only used in the field of computer vision.
10. Data labeling is a one-time process and does not require continuous updates.

D. Short Answer Question

1. What do you understand by data labeling?

2. Explain the use cases of data labeling.
3. What are the types of data labeling?
4. What is manual data labeling?
5. Explain any two data labeling tools.
6. What do you understand about labels and features in machine learning?
7. What is labeled data?
8. What are the benefits of data labeling?
9. Explain automatic data labeling.
10. What are the challenges in data labeling?

Session 3. Data Annotation in AI

In a city, Maya dreamed of starting an online clothing store. To make her store appealing, she needed to organize thousands of clothing images. Maya discovered "Data Annotation," a way to label pictures for computers. With data annotation, she labeled colors, styles, and clothing types. Customers loved it, and Maya's online store became a successful store. As shown in figure 3.1.

Figure 3.1: Maya's clothing business

In this chapter, you will understand the use of data annotation in business, data annotation services, market demand for annotation, and open-source tools for annotating the data.

3.1 Data Annotation Role in AI

Data labeling ensures AI and ML project scalability. Humans classify data, photos, and videos, enabling machines to make predictions. ML algorithms depend on data labeling for crucial properties. The growing demand for data labeling has businesses turning to service providers. Updated with new trends helps to identify effective annotation and labeling practices for your business.

3.1.1 Data Annotation in Business

Data annotation is the process of categorizing and tagging data to train machine learning models. This includes manually labeling various types of data like images, text, and audio with relevant details such as object labels, bounding boxes, and transcriptions. The goal of data annotation is to build a labeled dataset that helps machine learning models learn and understand the data accurately. Figure 3.2 demonstrates the way the picture annotation assists in determining if a fruit possesses a disease or not.

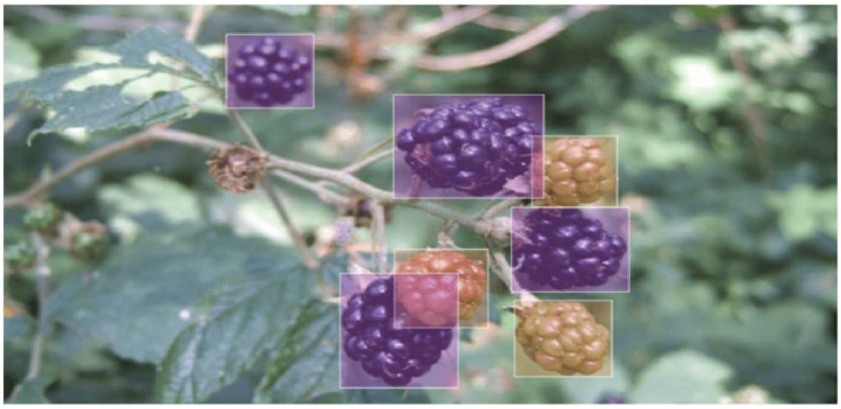


Figure 3.2: Image annotation identifying fruit affected by disease or not

3.1.2 Data Annotation Services in Business

Many tools and services can help companies annotate data. Because of developments in computer vision and AI, businesses offer better products and services. AI used to be for big tech companies but now is not used in fields like farming and medicine.

Here are the top three reasons to use data annotation services:

- Data preparation for AI takes time, but using third-party help makes it faster. It frees you to focus on other tasks and also reduces AI bias.
- Data annotation makes AI user-friendly and more effective. It helps users get quick answers and solves problems, especially in applications like chatbots and search engines.
- AI models are successful if they give good results. Proper data annotation reduces errors, and AI models can provide accurate and effective outcomes.

3.1.3 Specialized data annotation for certain sectors

Data labeling is used in the business field in several different ways. Some businesses only use one way to annotate their data, while others use a mix of different methods. Some of the most important types of data annotation for different areas are shown below:

- **Medical:** Medical data annotation is used to add clinical observations like medical records, pictures, electronic health records, and so on. This type of data annotation is useful for making systems that use computer vision to automate medical data processing and diagnose diseases.
- **Retail:** Retail data annotation is the process of annotating product pictures, customer records, and feedback. This kind of annotation helps build and train strong AI and machine learning models.
- **Finance:** Finance data annotation is the process of adding notes and comments to financial information. Annotation helps AI/ML systems, like those that seek out fraud and compliance problems, accomplish their tasks better.

- **Autonomous Vehicles:** Automotive data annotation is a specialized method that can be used to annotate data received by cameras and lidar sensors on autonomous vehicles. This annotation type makes it easier to develop models that can find objects and other data points in the surroundings for use in autonomous vehicles.
- **Industry:** Industrial data annotation can be used to annotate many different kinds of data, such as manufacturing pictures, maintenance records, safety records, data from quality control systems, and so on. This way of annotating data helps to make tools that can spot issues throughout the manufacturing process while keeping staff safe.

3.1.4 Market demand for data annotation

The three main factors that are fueling the market demand for data annotation tools include:

- The market demand for data annotation tools is being driven by efficient automated labeling tools and the increasing use of cloud-based computing resources for annotating large datasets.
- The increasing need for businesses to precisely label huge amounts of training data for AI projects is boosting the demand for data annotation tools.
- The growing investments in autonomous driving technology and the resulting need for well-labeled data to enhance machine learning models for driverless vehicles are contributing to the increased demand for data annotation tools.

3.1.5 Demand for Data Annotation in 2023

The demand for data labeling is expected to rise by 2030, mainly because of the expansion of machine learning tools and algorithms in both business applications and research. Additionally, data annotation is becoming increasingly essential for national security and surveillance reasons.

The newest trends in data labeling are helping to connect two important ways that machine learning works: supervised and unsupervised learning. This means that artificial intelligence is getting closer to being as smart as humans.

Below are seven trends that could affect the data marking market in 2023.

i. **Image, text, and video data annotation-** The data annotation market grows due to images, especially in industries like cars, energy, entertainment, healthcare, and more. Text data has become more important, especially in e-commerce, research, and social media.

ii. **Automated Data annotation-** AI is becoming more important in data annotation as it can automatically extract complex details from datasets.

iii. **Data labeling tools-** The data annotation tools market is booming globally. This growth is driven by ongoing tech advancements that require big datasets with annotations for learning and insights.

iv. **Healthcare sector-** Data annotation will boost AI in healthcare. AI-powered imaging spots issues and generates patient reports, helping healthcare professionals.

v. **Technology Market Growth-** Modern technologies such as artificial intelligence, machine learning, IoT, and robots create tons of data.

vi. **Quality Control Procedures-** Quality assurance (QA) is important. Careful data labeling is important for quality, especially in unusual cases. Skilled professionals can find and fix issues in big datasets.

The data-driven sector is taken over by predictive annotation-

Industries and Technologies- Several industries and technologies require more labeled data. These include:

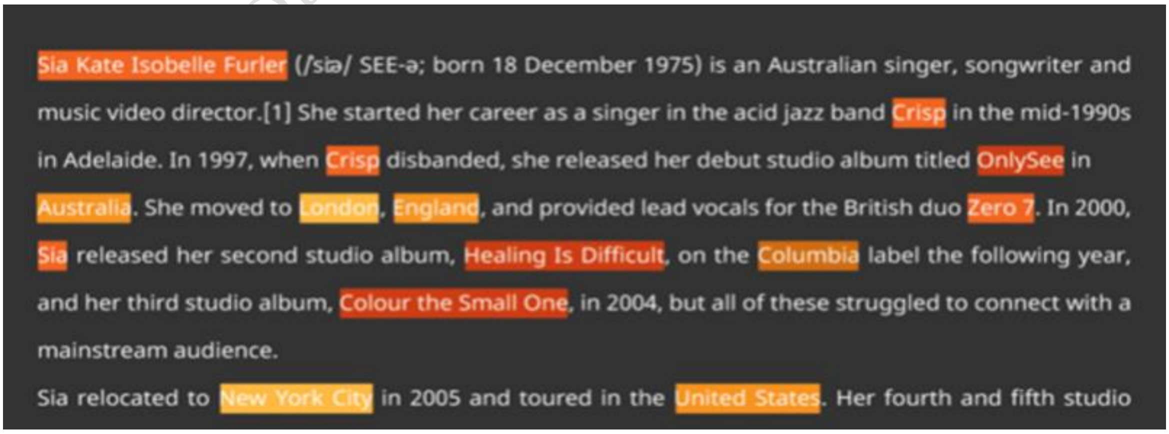
- **AI in healthcare-** Data annotation will transform healthcare using better computer vision and medical imaging. Labeling tools help AI in medicine, reduce human work, and advance medical research.
- **Active Learning-** Training machine learning models in a new way: create and label datasets together. It gives accurate results with fewer labels. This is called active learning, using both labeled and unlabeled data smartly.
- **Cloud-Based Services-** The need for cloud services is growing, especially in digital marketing and online shopping. They use labeled data to understand customer behavior. Cloud-based AI platforms will make labeling easier for tasks like recognizing faces, landmarks, and objects.
- **Social Media-** Data annotation helps measure emotions on social media like Twitter and Facebook. It manages data growth and short messages. Smart labeling is needed to analyze and monitor social media visuals securely.

3.2 Data Errors in Annotation

There are three categories of data errors:

i. **Mislabeled Data**

Annotating and labeling data properly is a must for a model to work. We try to find specific things in the text, but it can be tricky because it is not always clear where certain entities belong. In the same way, there are many ways to get labeling wrong such as annotating an image, video, or audio file. For example, as shown in Figure 3.3 consider the text annotation above: people/band names are marked in orange, countries in lighter orange, cities in yellow, and album titles in red. Without proper guidance, an annotator might struggle to recognize that "Crisp" is a band name, and "Healing is Difficult" is an album title, especially if they lack knowledge about music or the artist.



Sia Kate Isobelle Furler (/ˈsiːə/ SEE-ə; born 18 December 1975) is an Australian singer, songwriter and music video director.[1] She started her career as a singer in the acid jazz band Crisp in the mid-1990s in Adelaide. In 1997, when Crisp disbanded, she released her debut studio album titled OnlySee in Australia. She moved to London, England, and provided lead vocals for the British duo Zero 7. In 2000, Sia released her second studio album, Healing Is Difficult, on the Columbia label the following year, and her third studio album, Colour the Small One, in 2004, but all of these struggled to connect with a mainstream audience.

Sia relocated to New York City in 2005 and toured in the United States. Her fourth and fifth studio

Figure 3.3: Example of Mislabeled data (Source: Telus International)

This highlights the importance of being careful about assigning labels to training data. It is important to guarantee that the labels are obtained during data collection and annotation—whether done manually or automated.

To avoid errors, we can give clear instructions to annotators, whether human or using platforms. Quality checks can also help prevent labeling mistakes that could hurt the model's accuracy and business outcomes.

ii. **Inaccurate Labels**

Your algorithm will find it difficult to accurately identify objects if a label is not correct. Inaccurate labels can have serious effects on object detection. Common examples include:

a) Bounding boxes/polygons that are too large or not well-fitted.

b) Labels that do not encompass the whole object.

c) Labels that overlap with other objects.

For example, as shown in figure 3.4. if you are creating a computer vision model to spot tigers in the wild, your labels should accurately cover the complete visible area of the tiger—neither more nor less.

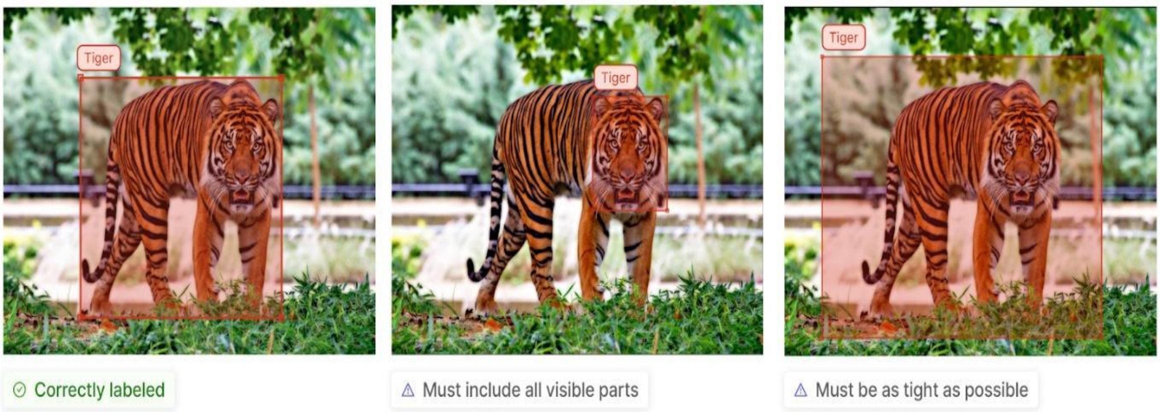


Figure 3.4. Example of inaccurate labels (source: https://encord.com/)

iii. **Inaccuracy in the Naming Procedure**

Training data bias is a common concern. If labeling needs specific knowledge or if annotators are too similar, bias can sneak in. For instance, consider labeling breakfast dishes like 'Hagelslag' and 'vegemite' as illustrated in Figure 3.5(a) and 3.5(b) respectively. American annotators might not recognize them as morning foods, causing bias. It is better to involve annotators from different regions for accurate data reflecting each culture's cuisine.

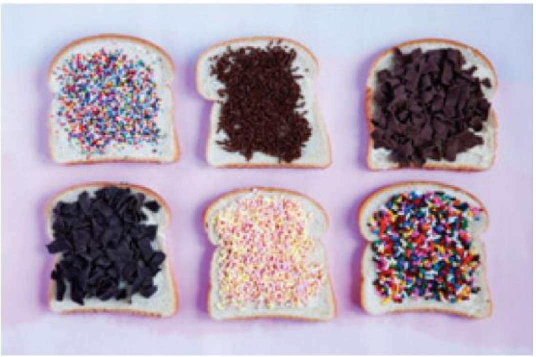


Figure 3.4 (a) - Hagelslag

Figure 3.4 (b) – Vegmite

iv. **Missing Label**

Certain objects might have been left without labels altogether. Fixing such errors, along with others, can be a time-intensive process that sometimes demands starting the data-labeling from scratch. Annotators must be careful and patient to prevent falling into this common trap. This can be exemplified by referring to the image shown in Figure 3.5.

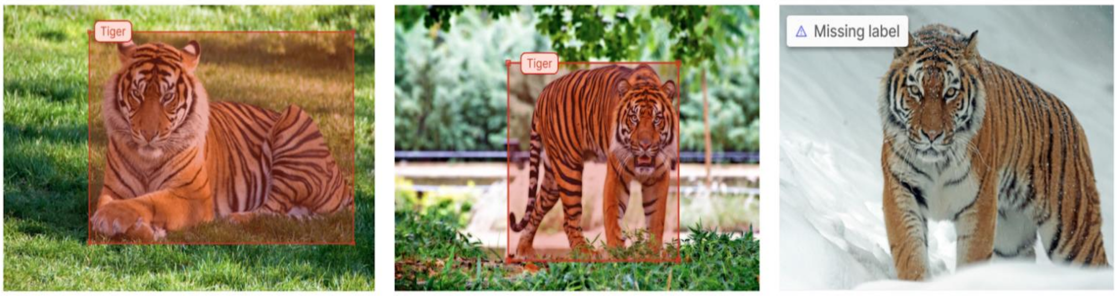


Figure 3.5. example of missing label (source: https://encord.com/)

Identifying a mistake in the dataset can become easier by having a review step where annotators double-check their work for quality before submitting it. These steps are to be followed, to prevent errors completely. Annotators should closely review the dataset before starting a task, and practicing patience is important. Being patient helps them catch any errors that might otherwise waste their time.

3.3 Object Annotation Errors

Here are some common errors that can occur during object annotation:

a) **Misclassification:** Assigning the wrong category to an object, like labeling a car as something else.

b) **Incorrect Attribute:** Not accurately describing an object's state, such as marking a moving car as parked.

c) **Missing Annotations:** Failing to annotate an object that should have been labeled.

d) **Redundancy:** Annotating an item multiple times, it only needs one label.

e) **Dimension Mismatch:** Annotations did not match the object's actual size, leading to incorrect sizing.

An illustrative example of object detection and classification can be seen in Figure 3.6.

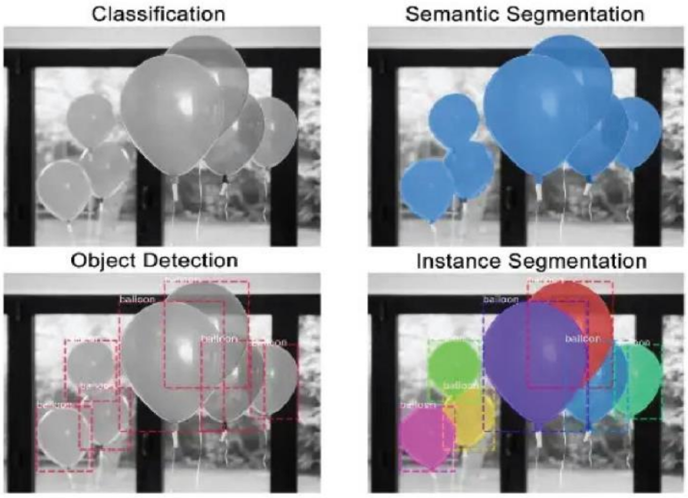


Figure 3.6. Object Detection and classification using Annotation

3.4 Data Annotation challenges

i. **High-Quality Training Datasets**

The success of AI/ML projects depends on good labeled data. Models learn from this data, so it must be precise. Errors in labeling or bounding boxes are a big problem for analytics companies, with serious consequences. AI/ML models work best and are trained with high-quality data.

ii. **Lack of data security**

Data annotation companies must follow strict rules to protect sensitive client info. Annotators can't use risky devices or share data randomly, especially in projects like military technology where security is critical for missions.

iii. **Low dataset quality**

Achieving high-quality data can be tough, but It is crucial for companies. Labelers need to be consistent and follow the rules. There are two types of data quality: subjective and objective.

In subjective quality, labels decide on their own because there is no single right answer, depending on language, location, or culture. Objective quality has one right answer, but sometimes labelers might not know enough or the rules are not clear.

iv. **Smart tools & assisted annotations**

The best way to annotate data in the future is by combining both manual and automatic methods, called a hybrid annotation model. Using AI-assisted tools is a smart solution. They automate tasks, improve processes, check data quality, and make work easier.

3.5 Data Annotation tools and their business purposes

Annotation tools markup different content types like text, photos, and databases. They are visually appealing and user-friendly, making them great for efficient processes. You can annotate important content like research and business reports on a whiteboard or PowerPoint slide. Quality annotation software is used in fields like gene ontology.

i. **nTask-** It is the best annotation tool and project management application. You can link and share files related to your tasks, making it great for collaboration.

ii. **Filestage-** This cloud service helps businesses collaborate, review, and approve content in one place. It lets companies share files, track versions, and create custom workflows for faster feedback and approval from multiple reviewers.

iii. **Annotate-** This easy-to-use data annotation tool makes the process fun. It is useful from start to finish, helping you present your data to your target audience.

iv. **PDF Annotator-** It is a tool for annotating PDFs on Windows. You can add notes, images, and more to any PDF file, making it easy to include notes on top of pages.

v. **DrawBoard Projects-** This software is excellent for marking up PDFs. You can use text and polygon tools to add notes and share your documents with clients, colleagues, and friends.

vi. **Doccano-** Machine learning pros can use Doccano, a free and open-source annotation tool. It labels data for tasks like sentiment analysis and text summarization. You can create datasets quickly, even collaboratively, and use them on mobile devices. It supports multiple languages and has a RESTful API.

vii. **Markup Hero-** Markup Hero is a user-friendly web tool for snapshots and notes. It has many markup options like callouts, highlights, and more. Your notes can be edited anytime. It helps you stay organized and share your annotations easily with colleagues through links or in documents.

3.5.1 Open Source and Freeware Alternatives

Companies decide to use free software or open-source tools for data annotation, instead of paying for expensive services. This approach finds a middle ground between doing everything independently and relying heavily on costly commercial solutions.

3.5.1.1 Widely Used Open Source/ Freeware Annotation Tools

Make sure that the tool's methods for creating and managing data structures, such as classes and attributes, meet your specific use case requirements. While many tools can work with various use cases, others specialize in specific types of labeling. For example, if you plan to use computer vision for tasks like classification, object annotation, or semantic segmentation, your chosen annotation tool should be able to annotate images effectively for all these purposes.

CoLabeler- CoLabeler is a free tool available for download, installation, usage, and sharing, similar to open-source software. It allows for creating bounding boxes, 2-D point annotations, and text annotations. A screen capture of CoLabeler is shown in Figure 3.7.

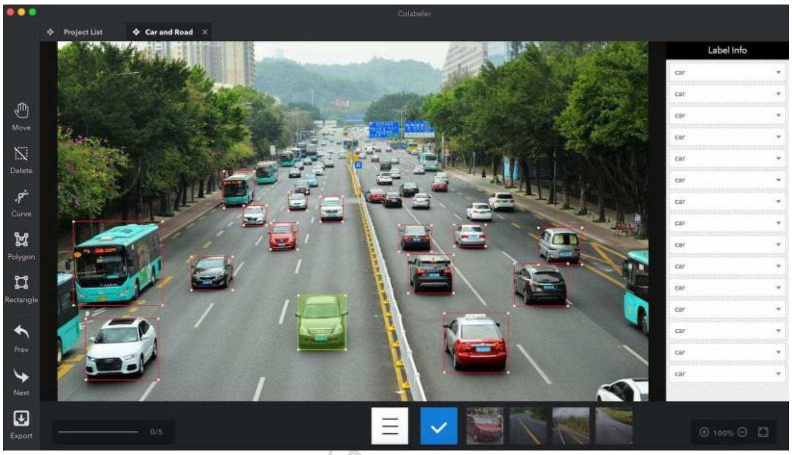


Figure 3.7: Screen capture of CoLabeler (Source: http://www.colabeler.com/)

Main Features:

1. It can perform Multi-type Labeling Tasks (Image classification, bounding box, polygon, curve, 3D localization, Video trace, text classification, and text entity labeling).
2. It is easy to use.
3. It has a user-friendly interface
4. It supports a custom task plugin, so you can create your label tool.
5. It can Support multiple platforms Windows/Mac/CentOS/Ubuntu.

Labelbox- Labelbox is a tool for training data. It was made in 2018 and quickly became one of the most popular tools for labeling data very fast. It lets you annotate with polygons, bounding boxes, lines, and more complicated tools for labeling. A screen capture of Labelbox is shown in Figure 3.3.

Main Features:

1. Works smoothly with data labeling services.

2. Effective statistics for label effectiveness.
3. It is easy to use.
4. It has a user-friendly interface
5. It is fast in speed.

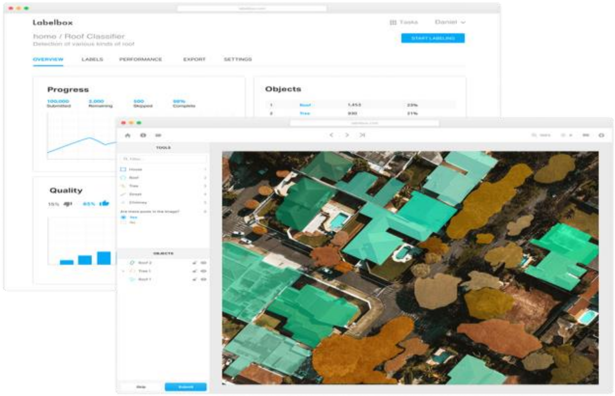


Figure 3.8: Screen capture of Labelbox (Source: https://www.v7labs.com/)

Plainsight- Plainsight collects top-quality datasets, develops and maintains incredibly accurate models, and deploys them on-site for tasks like animal detection, counting, and monitoring. These solutions not only automate manual tasks but also deliver the highest levels of accuracy, resulting in significant time and cost savings for users. A screen capture of Plainsight is shown in Figure 3.9.

Vision AI, Simply Done.

Accelerate enterprise value generation with innovative computer vision strategies, deep learning expertise, visual data science toolsets, and fully managed services.

Figure 3.9: A screen capture of Plainsight

Main Features:

1. It is easy to connect.
2. It can collect accurate training datasets to provide high-quality training for models.
3. It can easily train models.
4. It is easy to use.
5. It has a user-friendly interface.

Superannotate- Superannotate is an all-in-one platform for marking images and videos, making computer vision tasks easier. It helps create training datasets for tasks like object detection and video tracking. You can use various tools like boxes, polygons, and brushes for precise annotations. A screen capture of Superannotate is shown in Figure 3.10.

Main Features:

1. Superpixels for precise semantic segmentation.
2. Robust quality control systems with advanced features.
3. Compatibility with various formats through image conversion.
4. It is easy to use.
5. It has a user-friendly interface.

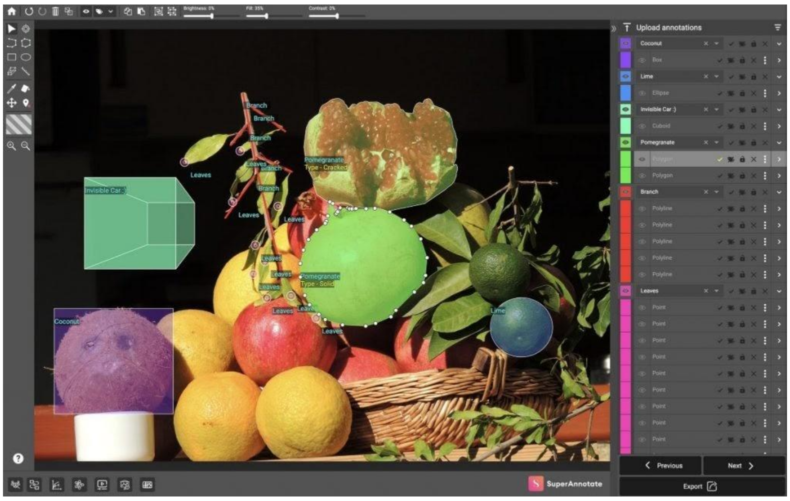


Figure 3.10: Screen capture of Superannotate (Source: https://www.v7labs.com/)

LabeIimg- LabelImg is a Python-based graphical tool designed for annotating images by labeling objects with bounding boxes. It allows you to export your annotations as XML files following the PASCAL VOC format. In its standard version, LabelImg supports a single annotation type, which is bounding boxes or rectangles. However, you can extend its capabilities by adding other shapes using a code available on GitHub. A screen capture of Labelimg is shown in Figure 3.11.

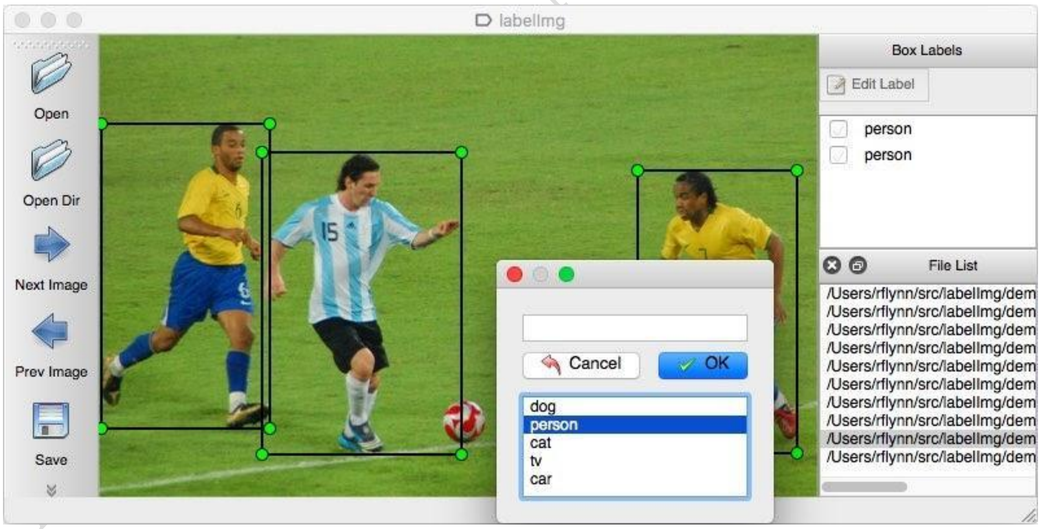

labelImg
Open
Open Dir
Next Image
Prev Image
Save
Box Labels
Edit Label
person
person
File List
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
/Users/rflynn/src/labelImg/dem
Cancel
OK
dog
person
cat
tv
car

Figure 3.11: Screen capture of Labelimg (Source: https://www.v7labs.com/)

Main Features:

1. Annotations are stored as XML files in the PASCAL VOC format.
2. It requires local installation.
3. It is primarily designed for image annotation.
4. It is easy to use.

5. It has a user-friendly interface.

LabelMe- LabelMe is an online tool made by MIT Computer Science and Artificial Intelligence Lab. LabelMe lets you use six types of annotations, like drawing polygons, rectangles, circles, lines, points, and lines in sequences. But, the drawback is – you can only save and export files in JSON format. A screen capture of Labelimg is shown in Figure 3.12.

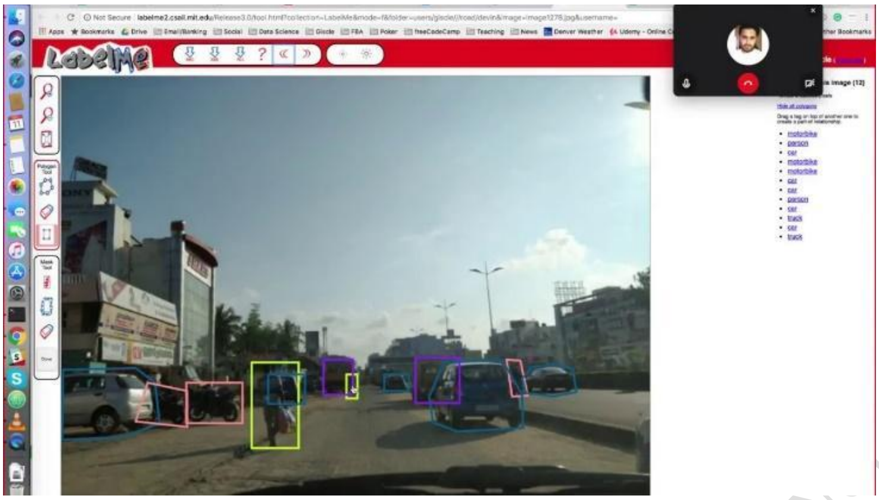


Figure 3.12: Screen capture of LabelMe (Source: https://www.v7labs.com/)

Main Features:

1. It can edit control points easily.
2. It can remove segments and polygons.
3. It supports six different types of annotations.
4. Handy file list for organization.
5. It is easy to use.

Dataloop- Dataloop is a cloud-based annotation platform that comes with built-in tools and automation to create high-quality datasets. Dataloop provides tools for fundamental computer vision jobs such as detection, classification, key points, and segmentation. And, it works with both image and video data. A screen capture of Dataloop is shown in Figure 3.13.

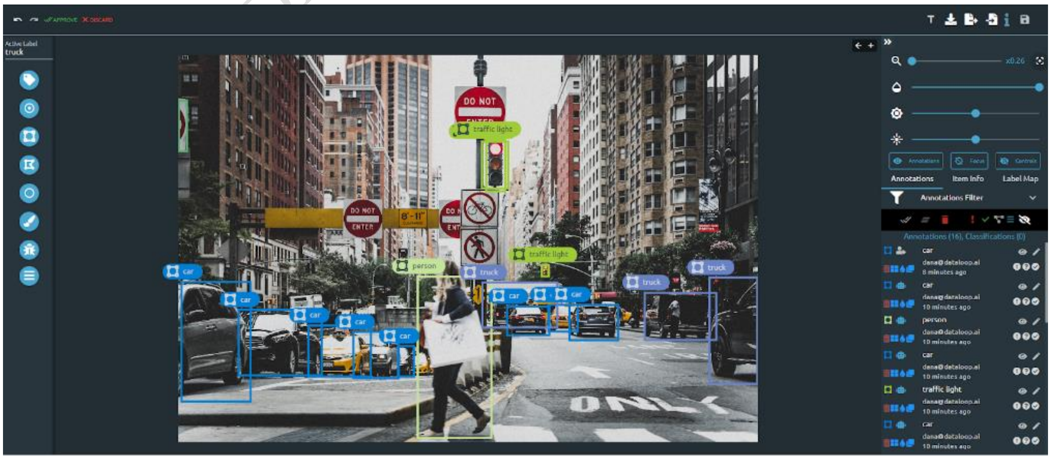


Figure 3.13: A screen capture of Dataloop (Source: https://www.v7labs.com/)

Main Features:

1. Labeling with assistance from AI models.
2. It supports various data types.
3. It is easy to use.
4. Advanced team workflows with a simplified data indexing and search system.
5. Compatibility with video data.

VoTT- VoTT is a flexible tool for importing and exporting data from your device or cloud storage. It works on Windows, Linux, or OSX and as a web app on browsers. It supports polygon and rectangle shapes, has project metrics and shortcuts, and exports data in different formats, including CSV and Microsoft Cognitive Toolkit (CNTK). A screen capture of VoTT is shown in Figure 3.14.

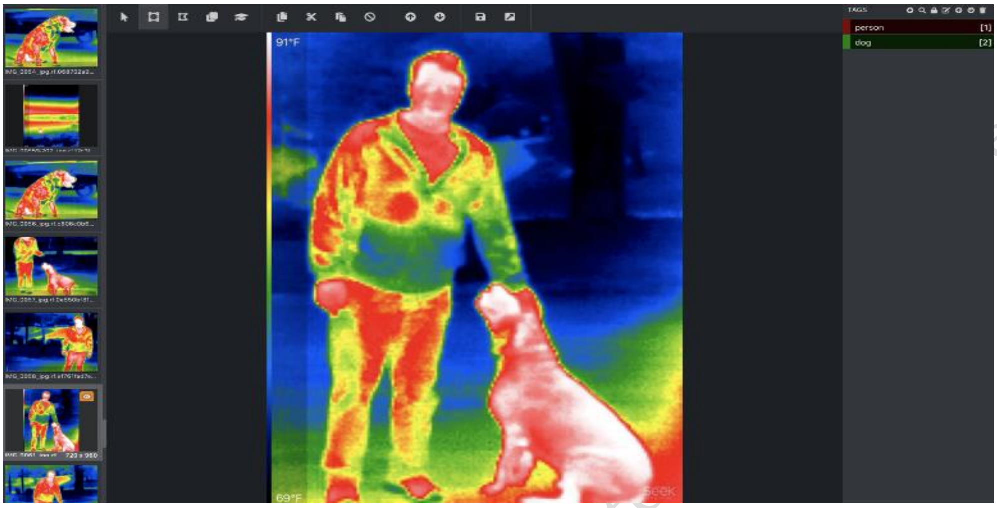


Figure 3.14: Screen capture of VoTT (Source: https://blog.roboflow.com/)

ImgLab: ImgLab is a free web tool for annotating images. It is easy to use, does not need installation, and does not use much computer power. You can resize images for different devices, use AI Face ID, and apply filters and adjustments for different styles. A screen capture of ImgLab is shown in Figure 3.15.

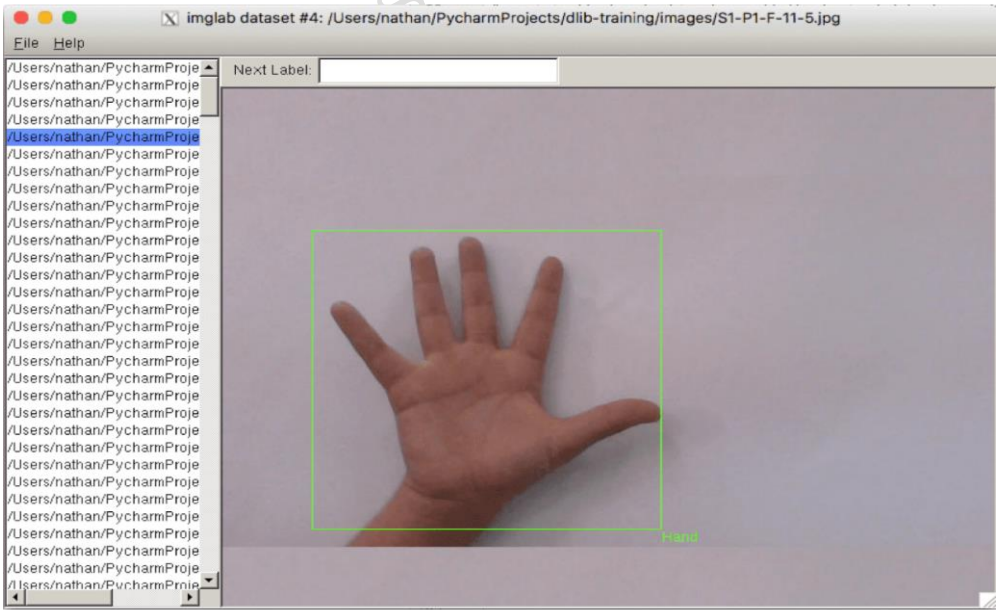

imglab dataset #4: /Users/nathan/PycharmProjects/dlib-training/images/S1-P1-F-11-5.jpg
File
Help
Next Label:
Hand

Figure 3.15: Screen capture of ImgLab (Source: ImgLab and https://www.v7labs.com/)

Main Features:

1. ImgLab is an open-source image annotation tool.
2. It has basic IDE features.
3. It supports multiple label types and file formats.
4. It is cost-free software.
5. It is easy to use.

SUMMARY

- Data annotation plays an important role in AI, providing labeled data for training machine learning models.
- In business, data annotation is important for tasks like image and text analysis, leading to improved decision-making.
- Businesses often rely on data annotation services to efficiently label their data.
- Various sectors, including medical, retail, finance, autonomous vehicles, and industry, require specialized data annotation.
- In 2023, there is a growing demand for image, text, and video data annotation, along with automated annotation and data labeling tools. The healthcare sector and technology market are experiencing significant growth.
- Data annotation errors include mislabeled data, inaccurate labels, naming procedure inaccuracies, and missing labels.
- Object annotation errors encompass misclassification, incorrect attributes, missing annotations, redundancy, and dimension mismatches.
- Data Annotation Tools like nTask, Filestage, PDF Annotator, DrawBoard Projects, Markup Hero, and Doccano serve various business purposes in data annotation.
- Open-source and freeware alternatives like CoLabeler, Labelbox, Plainsight, Superannotate, Labelimg, LabelMe, Dataloop, VoTT, and ImgLab offer cost-effective solutions for data annotation needs.

Check Your Progress

A. **Multiple Choice Questions**

1. What is the primary role of data labeling in AI and ML projects? (a) Data protection (b) Data organization (c) Scalability (d) Data encryption
2. Which of the following is NOT a type of data annotation? (a) Image labelling (b) Text annotation (c) Audio segmentation (d) Video object detection
3. Why do businesses use data annotation services? (a) To increase data security (b) To decrease the demand for AI models (c) To reduce AI bias (d) To speed up data preparation for AI projects
4. In which sector is medical data annotation commonly used? (a) Automotive (b) Retail (c) Finance (d) Healthcare
5. What is the main driver of market demand for data annotation tools? (a) Decreased use of AI in various industries (b) Reduced need for large datasets (c) The demand for

labeling tools for autonomous driving technology (d) Efficient automated labeling tools and cloud-based computing resources

6. What is the expected trend in data labeling by 2030? (a) Decreased demand due to reduced AI applications (b) Increased demand for national security reasons (c) A shift towards unsupervised learning (d) Growth in machine learning tools and algorithms
7. What is the significance of active learning in data annotation? (a) It involves labeling data independently (b) It requires manual labeling of the entire dataset (c) It helps achieve accurate results with fewer labels (d) It is not related to data annotation.
8. What is one way to prevent misclassification in data annotation? (a) Use annotators from different regions (b) Avoid annotators with specific knowledge (c) Keep labeling instructions vague (d) Assign labels randomly
9. Which data annotation tool supports labeling with assistance from AI models? (a) LabelImg (b) LabelMe (c) Dataloop (d) VGG Image Annotator
10. What is the advantage of using cloud-based services for data annotation? (a) Lower data quality (b) Reduced data security (c) Improved understanding of customer behaviour (d) Limited support for object annotation

B. **Fill in the blanks**

1. Data labeling helps machines understand data by categorizing and tagging it with relevant details such as labels and ______________.
2. Data annotation makes AI more ______________ and efficient by providing labeled data for training.
3. In the field of ______________, data annotation is used to automate tasks like diagnosing diseases using computer vision.
4. One of the main reasons businesses turn to data annotation services is to speed up data ______________ for AI projects.
5. Inaccurate labels can lead to difficulties in object ______________ for computer vision models.
6. Data annotation tools like Labelbox and Superannotate help create training datasets for tasks such as ______________ detection.
7. The demand for data labeling is expected to grow due to the increasing use of machine learning in various industries, including ______________.
8. Quality control procedures are important in data annotation to ensure the accuracy of ______________.
9. Smart tools and assisted annotations are becoming more popular to automate tasks and improve data annotation ______________.
10. CoLabeler and LabelMe are examples of ______________ annotation tools that are freely available for use.

C. True or False

1. CoLabeler is a paid annotation tool.
2. The demand for data annotation tools is increasing
3. Data labeling is not necessary for training machine learning models in AI projects.
4. Data annotation helps machine learning models understand and learn from data.

5. Data annotation aims to build datasets that enhance machine learning model understanding.
6. Smart tools and assisted annotations do not impact data labeling efficiency.
7. Data annotation is not important for quality control in AI/ML projects.
8. Data annotation tools like Labelbox and Superannotate are used in computer vision.
9. Data annotation is not relevant to the success of AI/ML projects.
10. Data annotation tools are used to advance AI in healthcare.

D. Short Question Answers

1. What is the role of data annotation in AI?
2. What are some specialized sectors where data annotation is used?
3. Explain market demand for data annotation tools.
4. Why is data security crucial in data annotation services?
5. What is the significance of quality control in data annotation?
6. What are the various open-source tools for data annotation?
7. What are the data annotation challenges?
8. What are the data errors in data annotation?
9. What are the Object annotation errors?
10. Explain Data Annotation Services in Business.

Session 4. Data Annotation Open Source Tools

Once in a small town, there was a curious teenager named Aiden who loved animals, plants, antiques, and fancy cars. One day, Aiden discovered some special computer tools that could teach computers about these topics, and they were free for everyone to use. Aiden eagerly used these tools to show pictures of animals, plants, and objects to the computer. Aiden's efforts even helped a bird protection group in identifying bird species and caring for them. The town soon joined in, using these tools to educate computers about various subjects. Aiden's story showcased that computers could learn and grow through these free tools, making the world a smarter place. As illustrated in Figure 4.1.

Fig. 4.1: Aiden working on the computer (Source: https://www.freepik.com/)

In this chapter, you will learn about open-source tools used for data annotation. And hands-on practice with these tools.

4.1 Data Annotation Open-source tools

4.1.1 CVAT

CVAT, which stands for Computer Vision Annotation Tool, is a well-known tool for labeling images and videos. It was created by Intel and can be used online with some restrictions or installed on your computer. CVAT is a widely used free tool for adding labels to your pictures and videos.

CVAT is used to tag data for computer vision tasks such as:

- ➢ Image Classification
- ➢ Image Segmentation
- ➢ Object Detection
- ➢ Object Tracking
- ➢ Pose Estimation

i. **Begin with CVAT**

To access the CVAT, you're required to make an account or sign in to an account you already have. It has two steps:

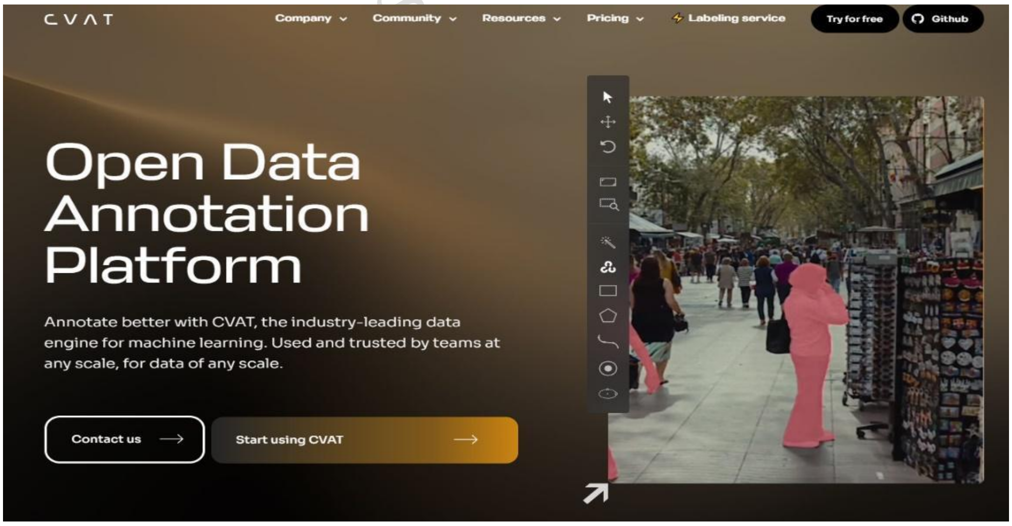
1. Registration
2. Login

After the login, you have to do the following tasks:

3. Start Labeling
4. Start Annotating

1. **Registration**

- ➢ Visit the CVAT login page, shown in Figure 4.1, to make an account or sign in. In the figure 4.2 below, you can see the CVAT login page.

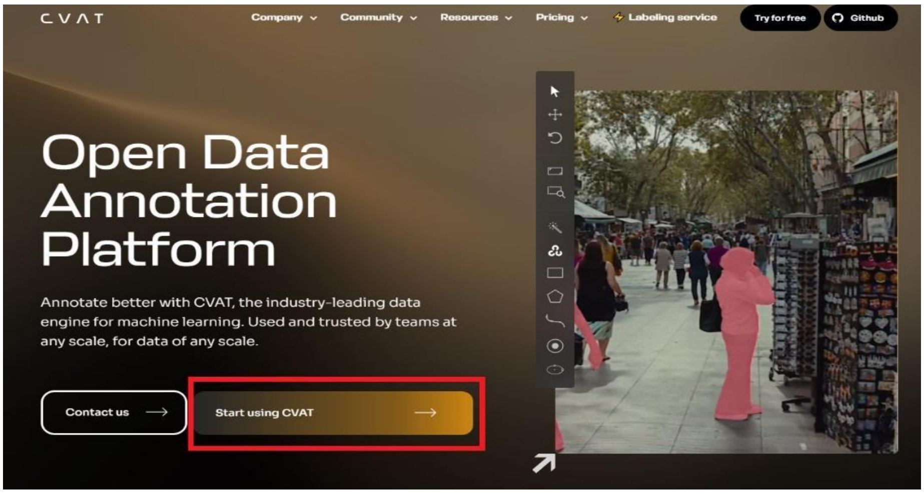


Open Data Annotation Platform

Annotate better with CVAT, the industry-leading data engine for machine learning. Used and trusted by teams at any scale, for data of any scale.

Figure 4.2: CVAT Homepage

- ➢ For registration or login, first, you have to click on "Start using CVAT". As shown in figure 4.3.

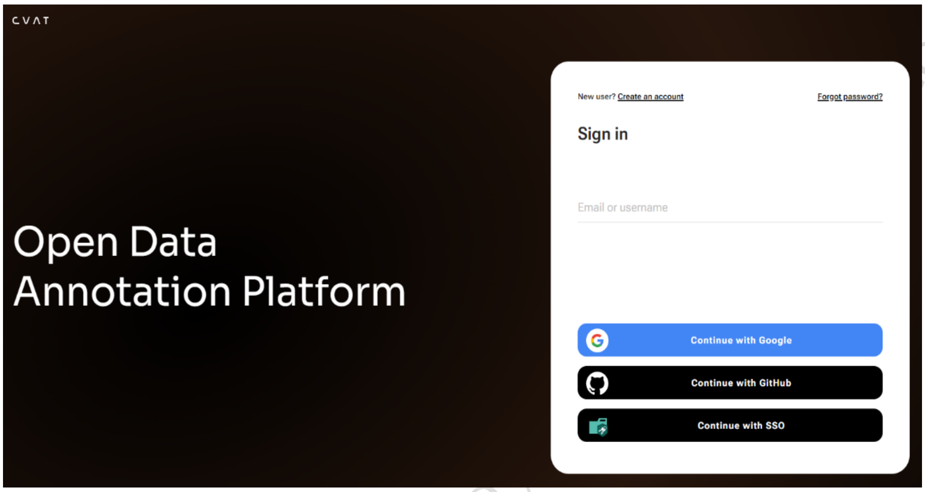


Open Data Annotation Platform

Annotate better with CVAT, the industry-leading data engine for machine learning. Used and trusted by teams at any scale, for data of any scale.

Figure 4.3: Click on start using CVAT

➢ After clicking, this will open another webpage. As illustrated in figure 4.4.



Open Data Annotation Platform

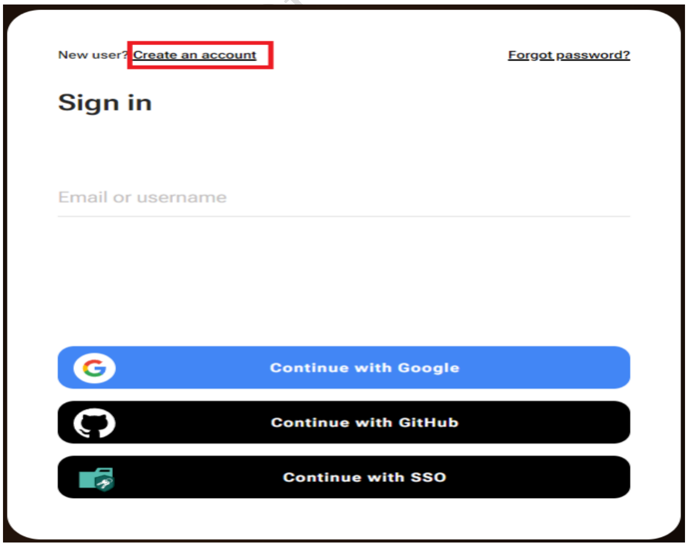
Sign in

Email or username

Figure 4.4: CVAT Sign-in Page

➢ If you are a new user and want to make a regular account (not as admin), just follow these steps:

i. Click on "Create an account." As illustrated in Figure 4.5.

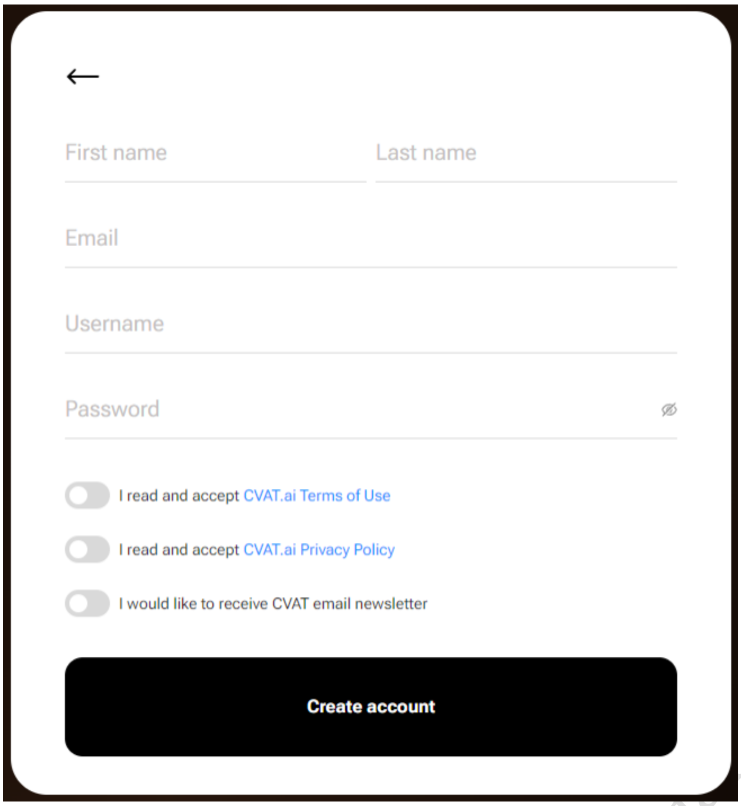


Sign in

Email or username

Figure 4.5: Create an account

ii. The following form will appear to fill. You have to fill in all the required details. As illustrated in Figure 4.6.



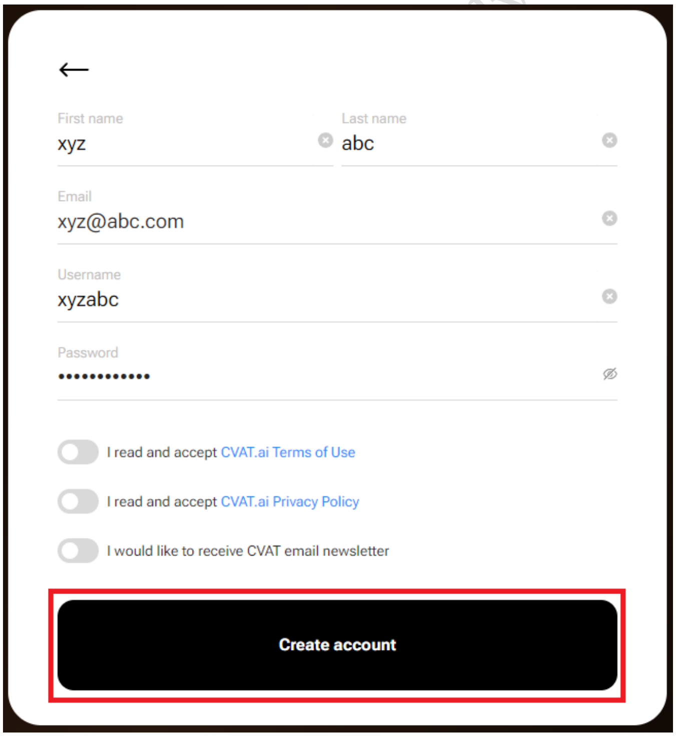
○ I read and accept CVAT.ai Terms of Use

○ I read and accept CVAT.ai Privacy Policy

○ I would like to receive CVAT email newsletter

Figure 4.6: Create account form

iii. Fill in all the empty spaces with the required information they ask for. click on their terms of use and privacy policy, and then tap on the "Create an account" button. As illustrated in Figure 4.7.



First name: xyz

Last name: abc

Email: xyz@abc.com

Username: xyzabc

Password: ••••••••••••

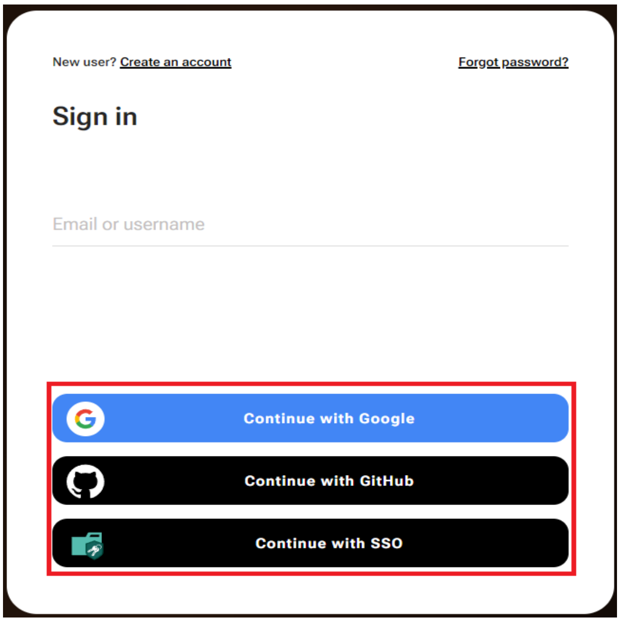
○ I read and accept CVAT.ai Terms of Use

○ I read and accept CVAT.ai Privacy Policy

○ I would like to receive CVAT email newsletter

Figure 4.7: Click on “Create account”

- If you prefer signing up with Google or GitHub, just click the button of the service name you want to use, and then follow the instructions on the screen. As shown in Figure 4.8.



New user? Create an account

Forgot password?

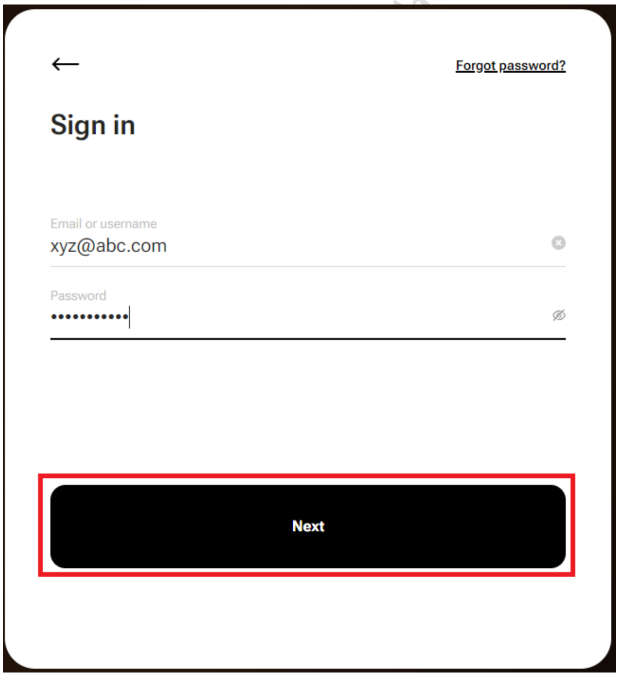
Sign in

Email or username

Figure 4.8: Create an account using other services

1. **Login/Sign in**

- To access your account, follow these steps:
- Go to the login page.
- Enter your username or email. This will make the password box appear. As shown in figure 4.4.
- Type in your password and click "Next."
- You are successfully logged in.



Forgot password?

Sign in

Email or username

xyz@abc.com

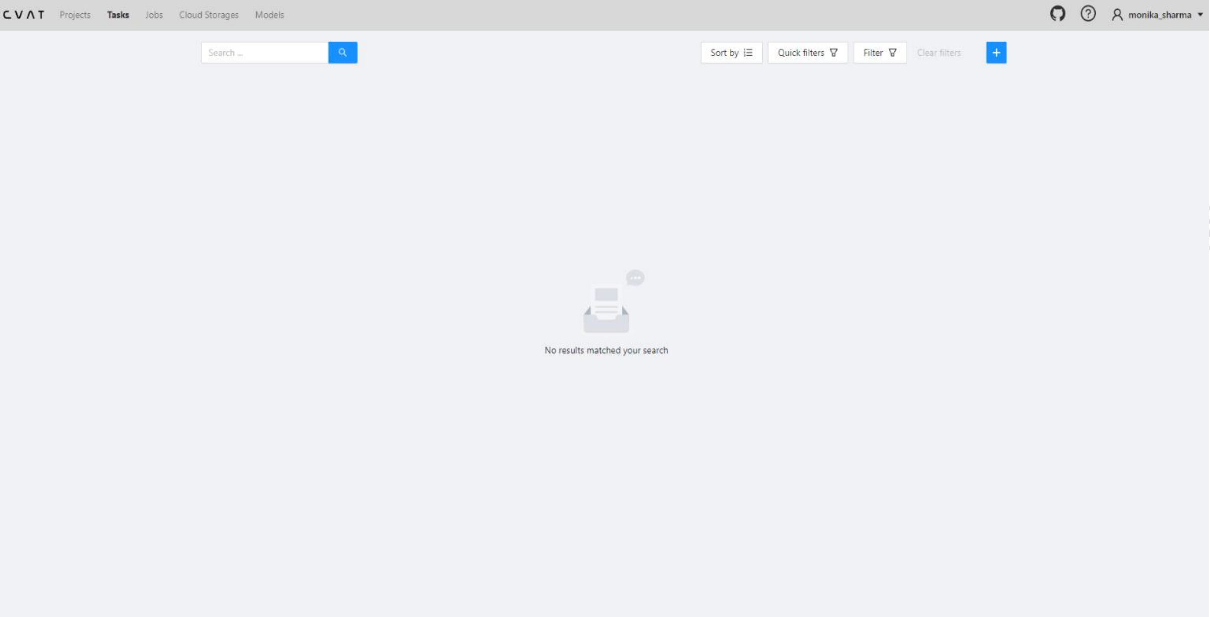
Password

•••••••••••

Figure 4.9: Sign in by registered email id and password

2. **Start Labeling**

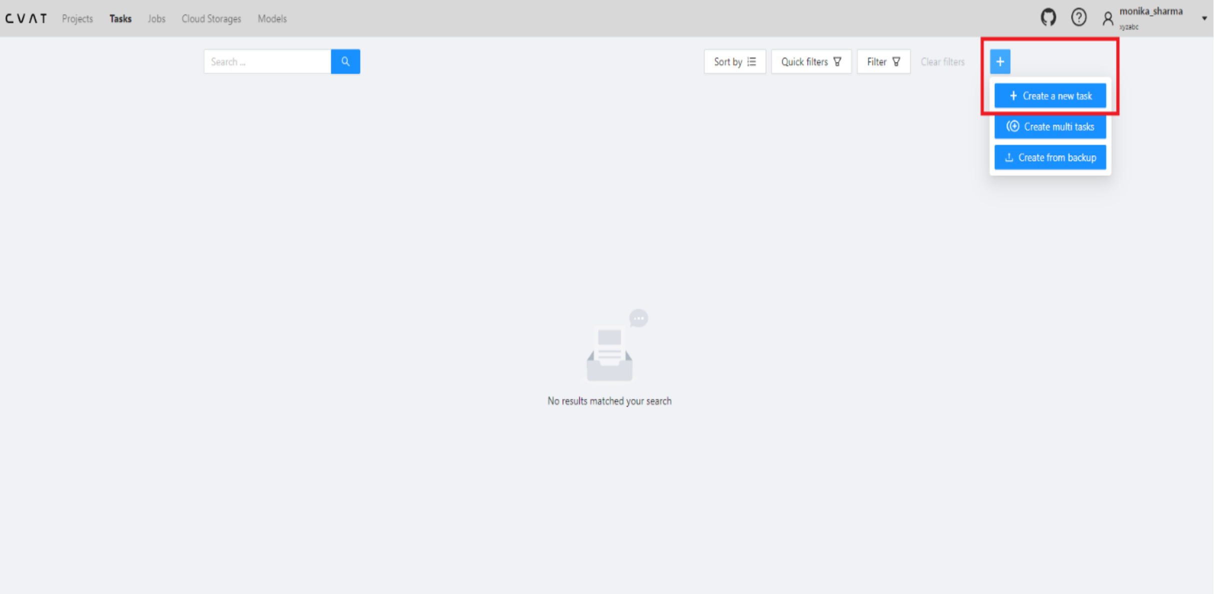
➢ Once you have completed those steps, login, and you will arrive at this page. As shown in figure 4.10.



No results matched your search

Figure 4.10: CVAT dashboard

➢ Now, you can create a new task, by clicking on the "+" and then tap on "create a new task". As illustrated in Figure 4.11.



No results matched your search

Figure 4.11: creating a new task

➢ After clicking, a new form will appear on your screen. As illustrated in Figure 4.12.

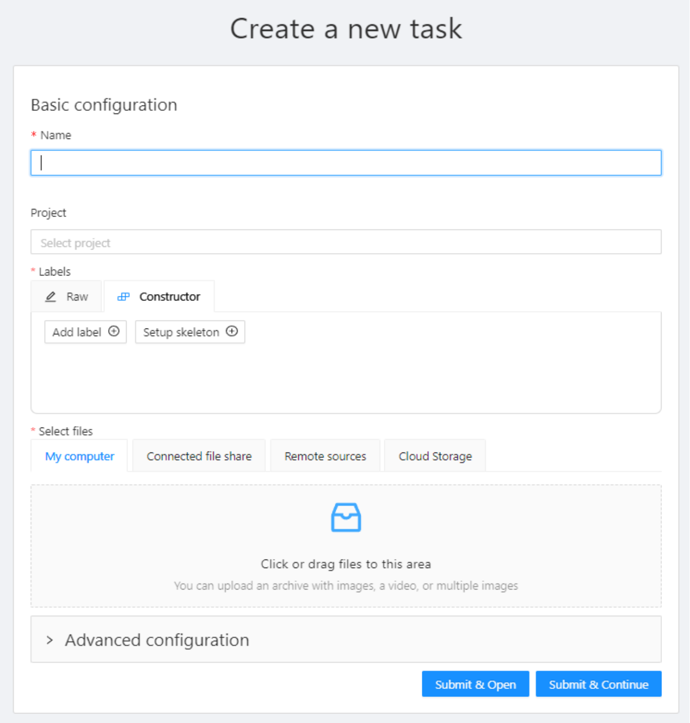
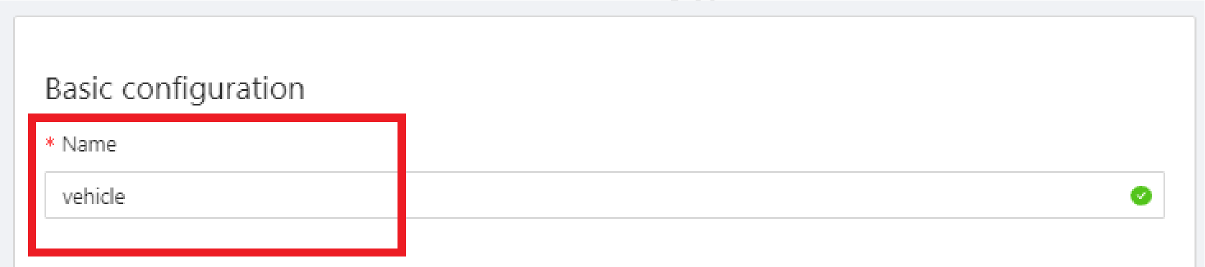


Figure 4.12: CVAT's labeling task

➢ Provide the following information:

1) Fill in the "Name" field with the name of your new task. As shown in figure 4.13.


Basic configuration

* Name

vehicle

Figure 4.13: CVAT's labeling task

2) Choose a project from the "Projects" drop-down menu to associate the task with a specific project (Optional). If not, you can leave this field empty. As shown in figure 4.14.

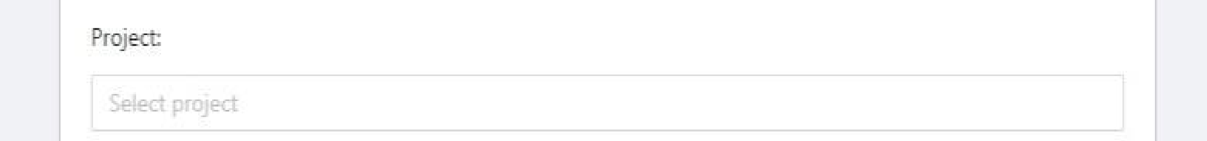


Figure 4.14: Choose a project

3) Go to the Constructor tab and click on "Add label." This will open the label constructor menu.

- Enter the label's name in the "Label name" field. As shown in figure 4.15.

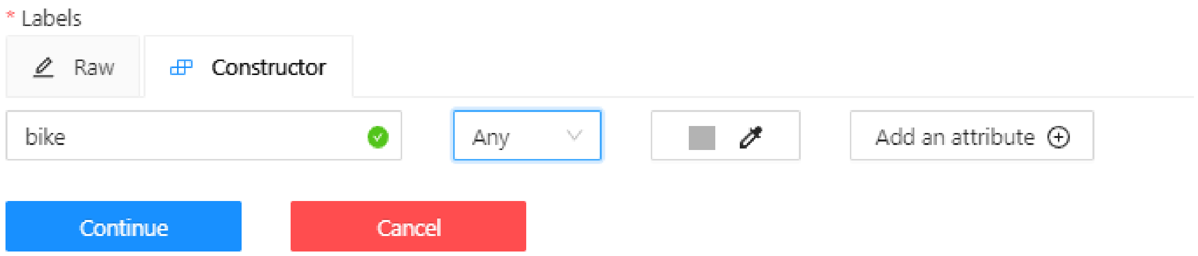


Figure 4.15: Add labels

- (Optional) If you want to use the label with a specific shape tool, choose the shape from the "Label shape" drop-down. As shown in figure 4.16.

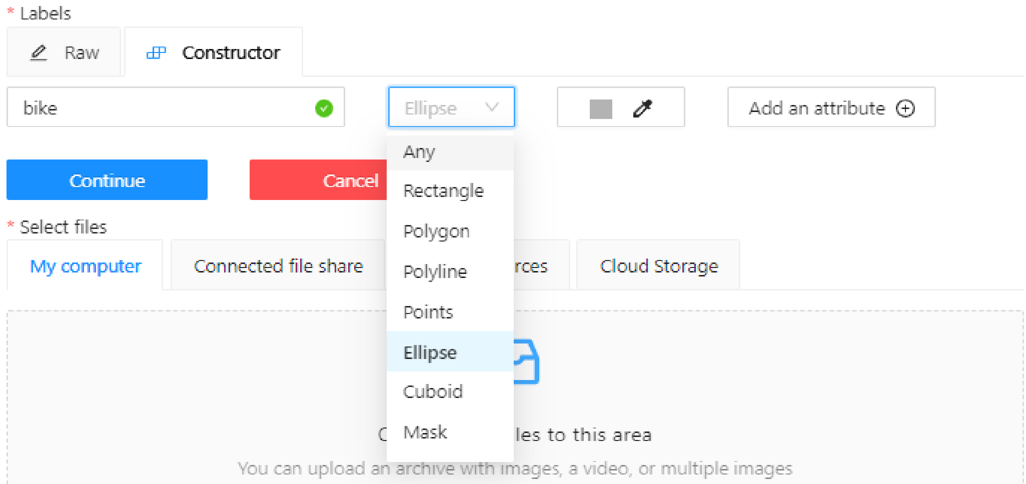


Figure 4.16: Pick a shape

- (Optional) Pick a color for the label. As shown in Figure 4.17.

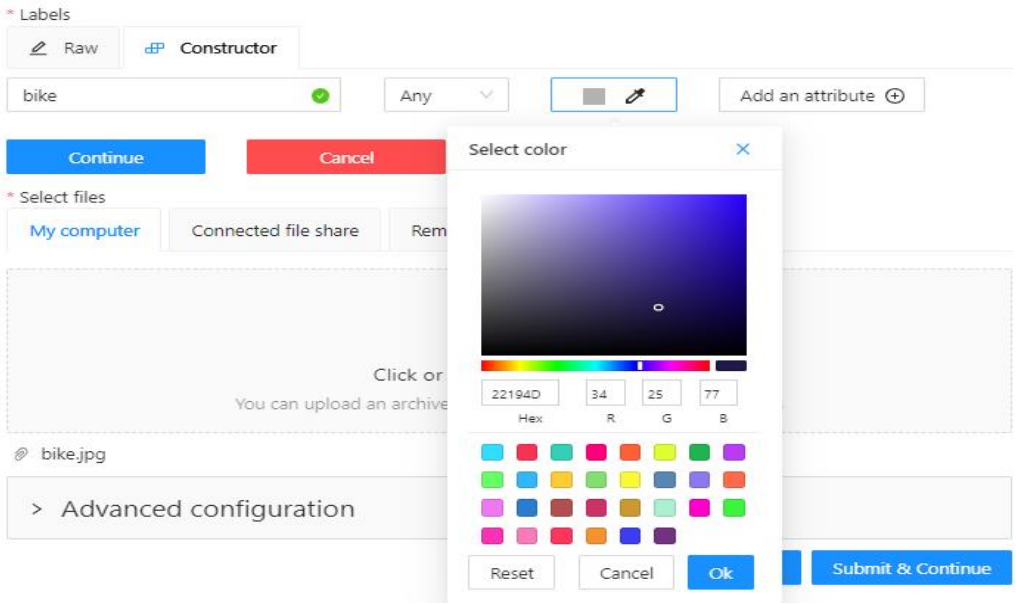
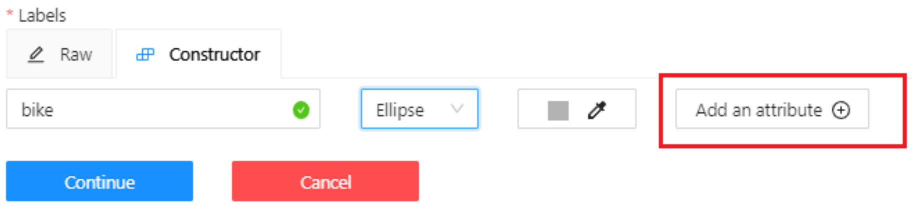


Figure 4.17: Adding color

4) If needed, add an attribute and configure its settings. As illustrated in figure 4.18.



* Labels

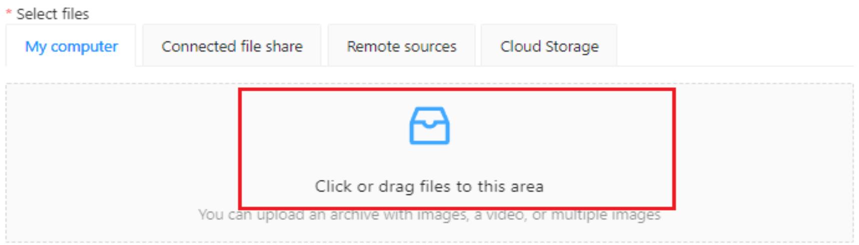
Raw | Constructor

bike | Ellipse | Add an attribute ⊕

Continue | Cancel

Figure 4.18: Adding attribute

5) Click "Select files" to choose the files you want to annotate. As shown in figure 4.14.



* Select files

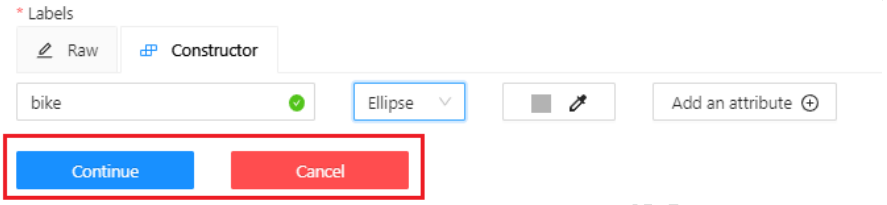
My computer | Connected file share | Remote sources | Cloud Storage

Click or drag files to this area

You can upload an archive with images, a video, or multiple images

Figure 4.19: Add files

6) Click "Continue" to save the label and create another one, or "Cancel" to exit the current label and go back to the labels list. As shown in figure 4.20.



* Labels

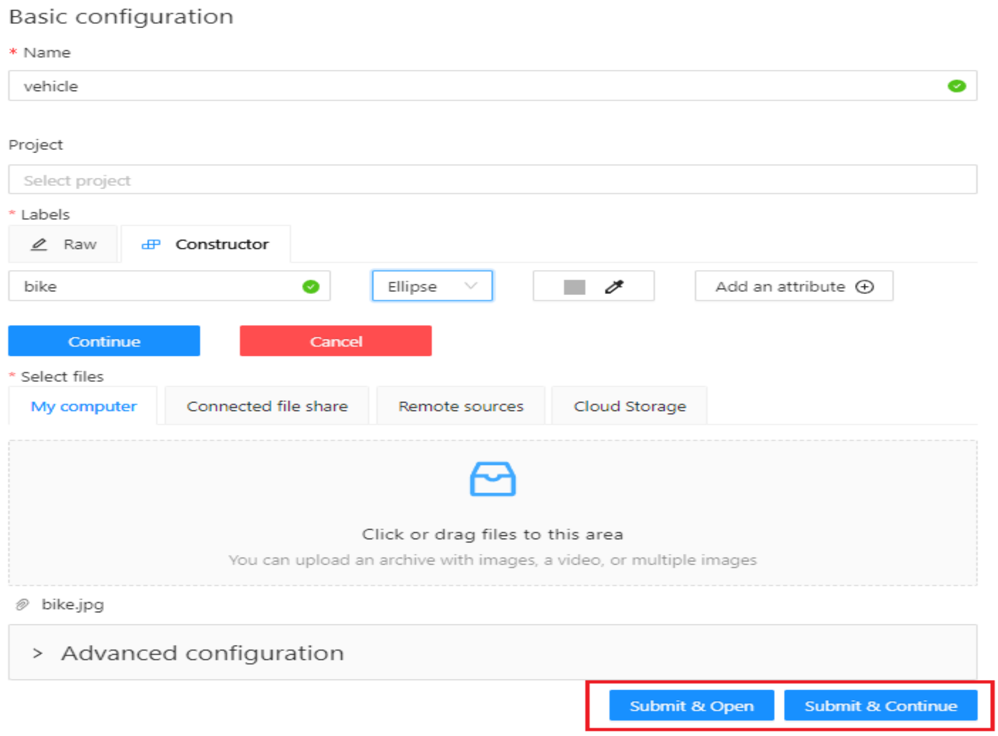
Raw | Constructor

bike | Ellipse | Add an attribute ⊕

Continue | Cancel

Figure 4.20: Click continue or cancel

7) Click "Submit and open" to save the setup and open the new task, or "Submit and continue" to save the setup and begin a new task. As shown in figure 4.21.



Basic configuration

* Name

vehicle

Project

Select project

* Labels

Raw | Constructor

bike | Ellipse | Add an attribute ⊕

Continue | Cancel

* Select files

My computer | Connected file share | Remote sources | Cloud Storage

Click or drag files to this area

You can upload an archive with images, a video, or multiple images

bike.jpg

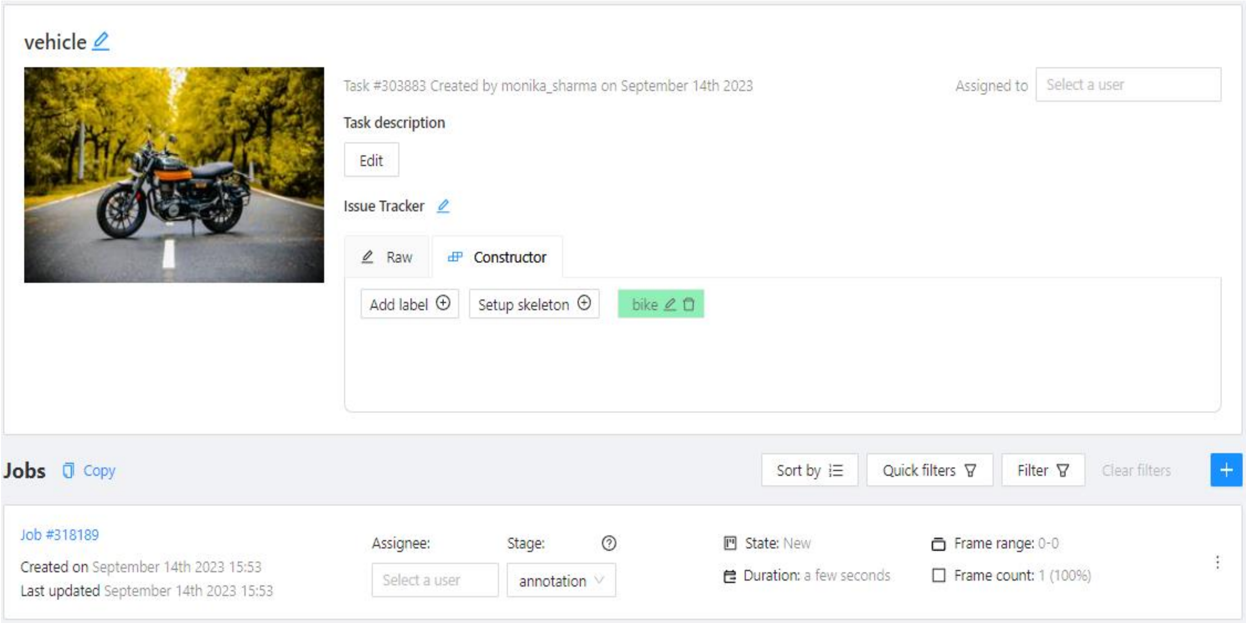
> Advanced configuration

Submit & Open | Submit & Continue

Figure 4.21: Click "Submit and open" or "Submit and continue"

8) After clicking on “Submit & open”, your file is uploaded to the server. And the project is created.

9) After that, it will continue you to the next page. That shows your created project details. As illustrated in figure 4.22.



vehicle

Task #303883 Created by monika_sharma on September 14th 2023

Assigned to: Select a user

Job #318189

Created on September 14th 2023 15:53

Last updated September 14th 2023 15:53

Assignee: Select a user

Stage: annotation

State: New

Duration: a few seconds

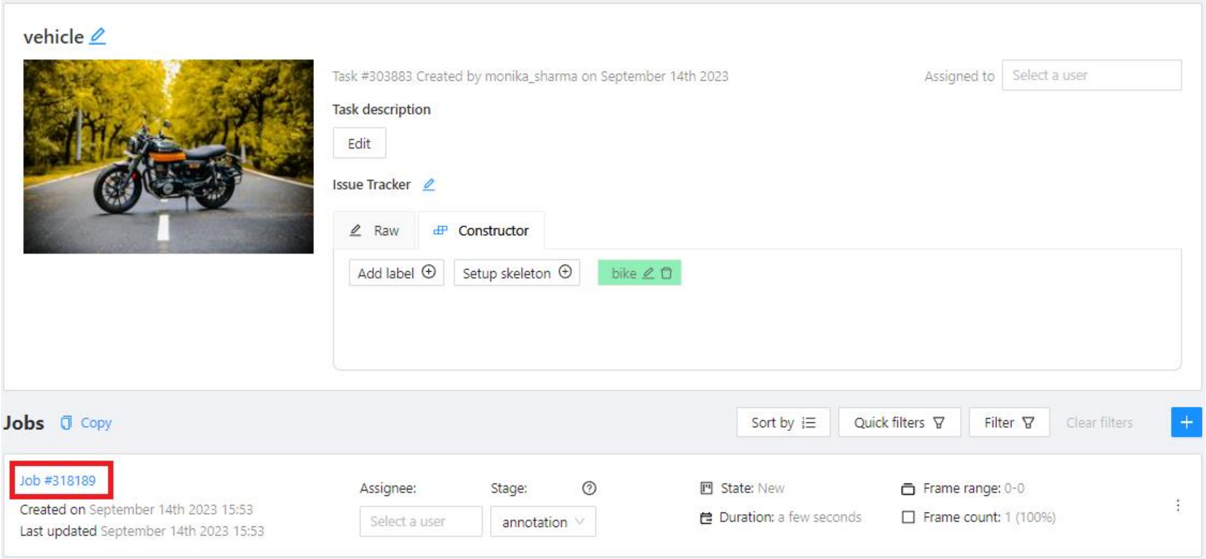
Frame range: 0-0

Frame count: 1 (100%)

Figure 4.22: Created project details

3. **Start Annotating**

a) Now, you click on the job below. As shown in figure 4.23.



vehicle

Task #303883 Created by monika_sharma on September 14th 2023

Assigned to: Select a user

Job #318189

Created on September 14th 2023 15:53

Last updated September 14th 2023 15:53

Assignee: Select a user

Stage: annotation

State: New

Duration: a few seconds

Frame range: 0-0

Frame count: 1 (100%)

Figure 4.23: click on the job

b) After clicking on the job, it will redirect you to the new page. As illustrated in figure 4.24.

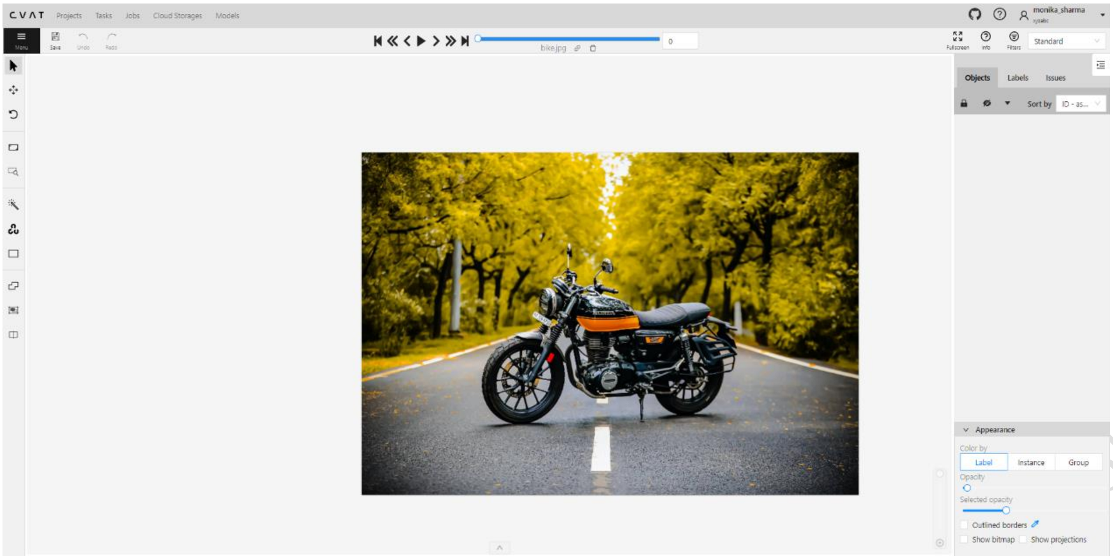


Figure 4.24: New page of CVAT

c) The tools in the Controls sidebar will only work with the chosen shape types. As shown in figure 4.25.

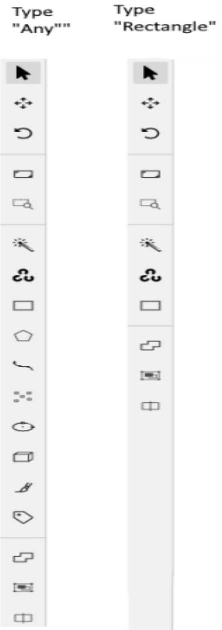
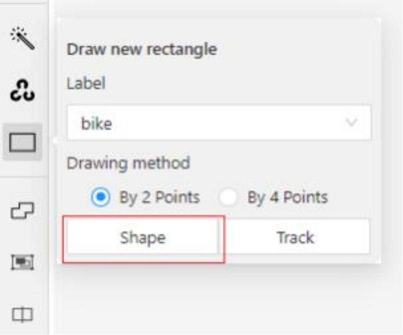


Figure 4.25: Controls sidebar

d) In the control's sidebar, choose a rectangle or any shape and click on "Shape". As shown in figure 4.26.



Draw new rectangle

Label

bike

Drawing method

● By 2 Points ○ By 4 Points

Shape Track

Figure 4.26: Click "Shape"

e) Now, draw a rectangle around the object in an image. As shown in figure 4.27.

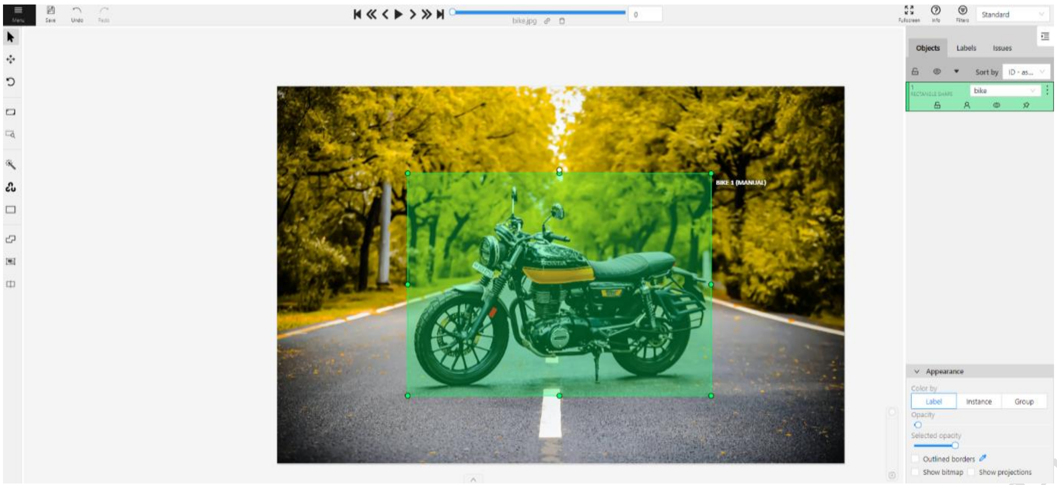


Figure 4.27: Draw a rectangle around the object.

f) After this, you will see that the image is successfully annotated. As shown in figure 4.28.

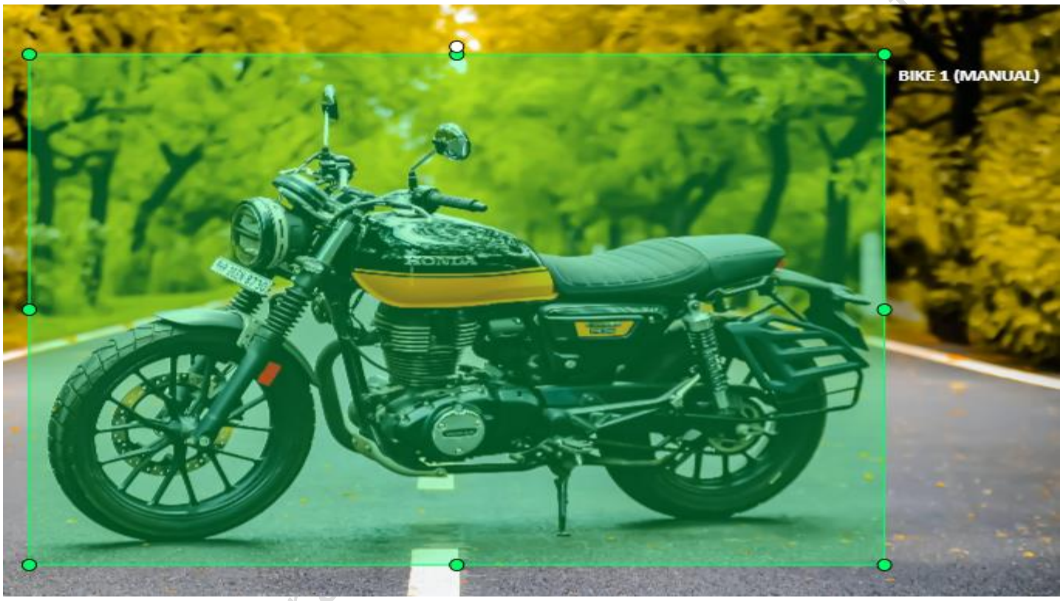


Figure 4.28: Annotated image

4.1.1 LabelMe

LabelMe is a free tool for adding labels to pictures. It helps to create annotations for object detection, classification, and segmentation of computer vision datasets. You can mark and label things in the pictures using shapes like polygons, rectangles, circles, lines, and points.

LabelMe Installation

LabelMe provides different ways to install it, including methods specific to different computer systems and a Python method that works on Windows, Linux, and macOS. For our example, we use the Python installation method.

For installing LabelMe by using Python, we need to run the following command:

Once the installation is complete, you can launch LabelMe using the following command:

LabelMe will open in a new window. As shown in figure 4.24.

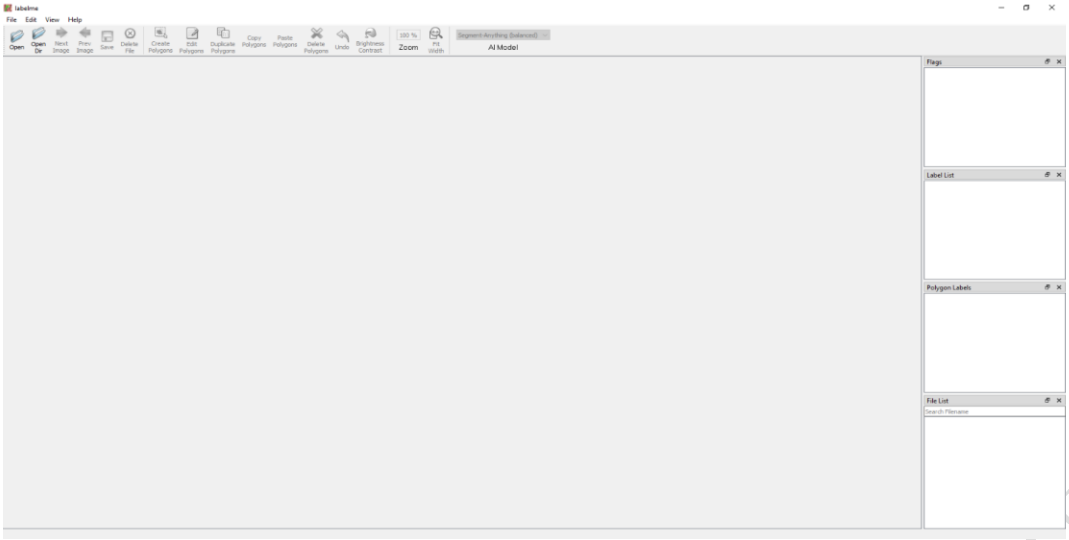


Figure 4.29: LabelMe interface

Upload Images to LabelMe

➢ To label images in LabelMe, click on "Open dir" and pick a folder containing the images you want to work on. You can also click "Open" to choose individual images. As illustrated in figure 4.30:

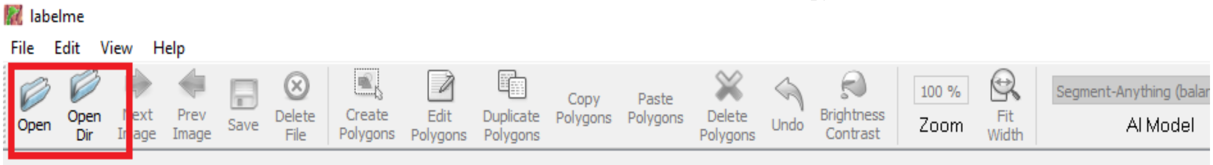


Figure 4.30: Click "Open dir" or "Open"

➢ The pictures from the folder you selected in LabelMe will appear on the screen. We can now start adding labels to these images. As illustrated in figure 4.31.

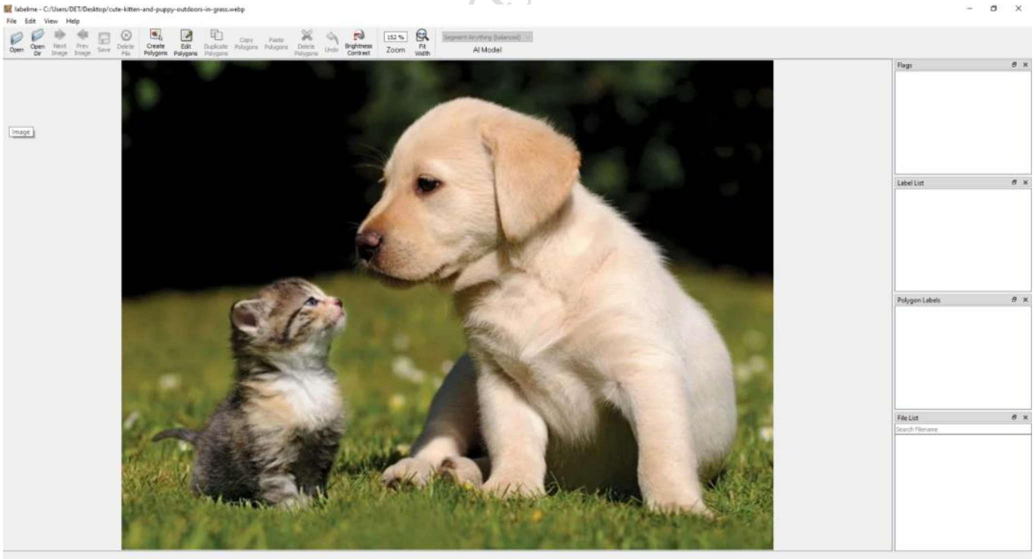


Figure 4.31: Add pictures

➢ To create an annotation, select the "Create Polygons" option from the application sidebar. As illustrated in figure 4.32.

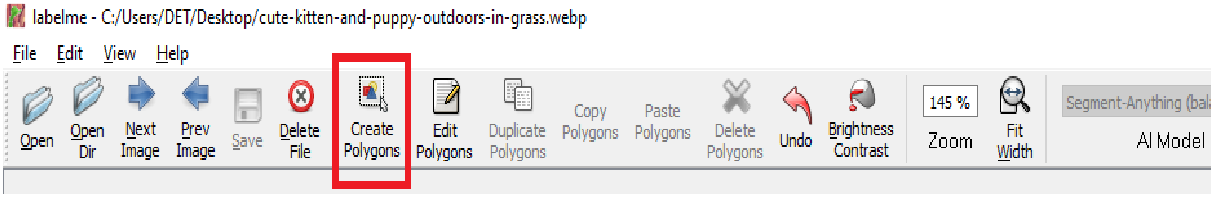


Figure 4.32: Create Polygons

- ➢ Next, click on "Edit" in the application command tray to select the type of polygon you want to make. As illustrated in figure 4.33.

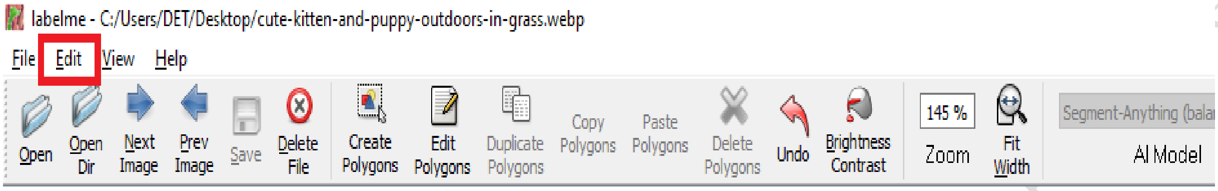
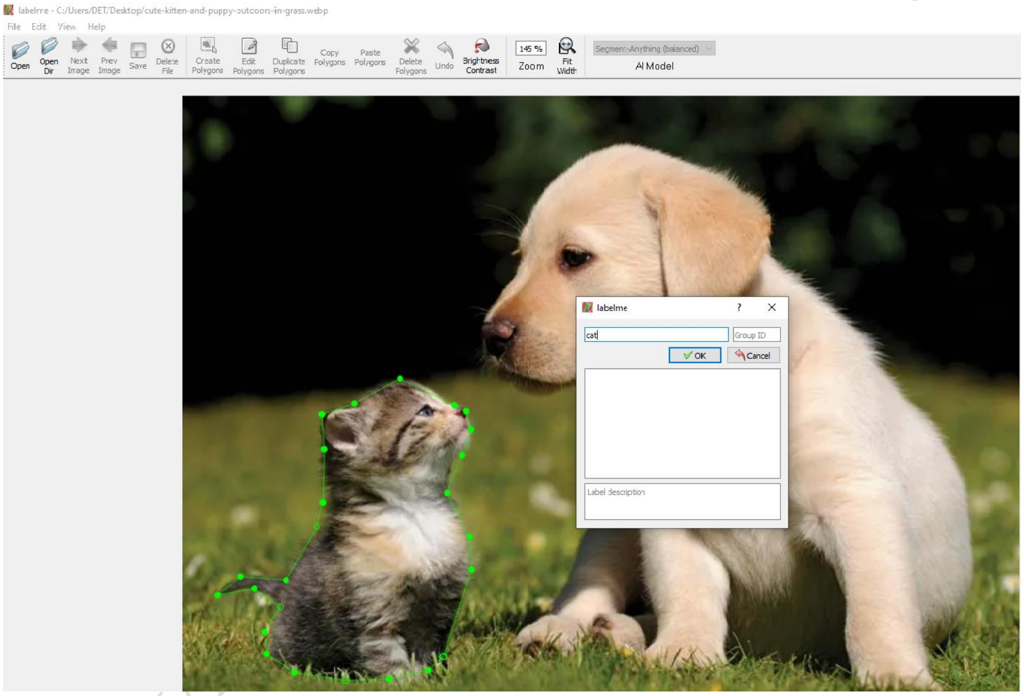


Figure 4.33: Click Edit

- ➢ Next, you can draw polygons to outline the image. And then add a label to the selected polygon. As illustrated in figure 4.34.



cat

Figure 4.34: Draw polygons

- ➢ Here, you can also choose the "Create Rectangle" option to use the rectangle tool for drawing a bounding box. After selecting this tool, click anywhere on the image to begin drawing the bounding box. As illustrated in figure 4.35.

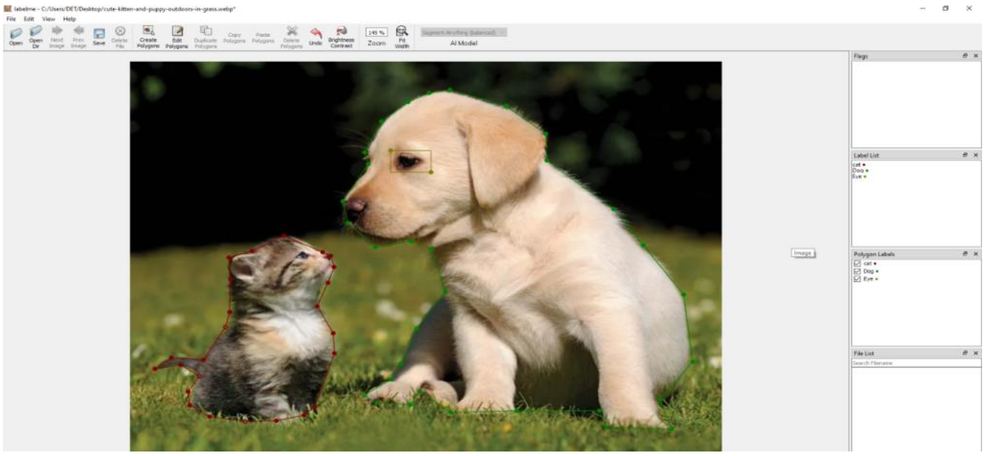


Figure 4.35: Create Rectangle

- ➢ To create a bounding box, click on a point, then drag your cursor to draw the box. Once you are done, click your mouse or trackpad to confirm the box. As illustrated in figure 4.36.

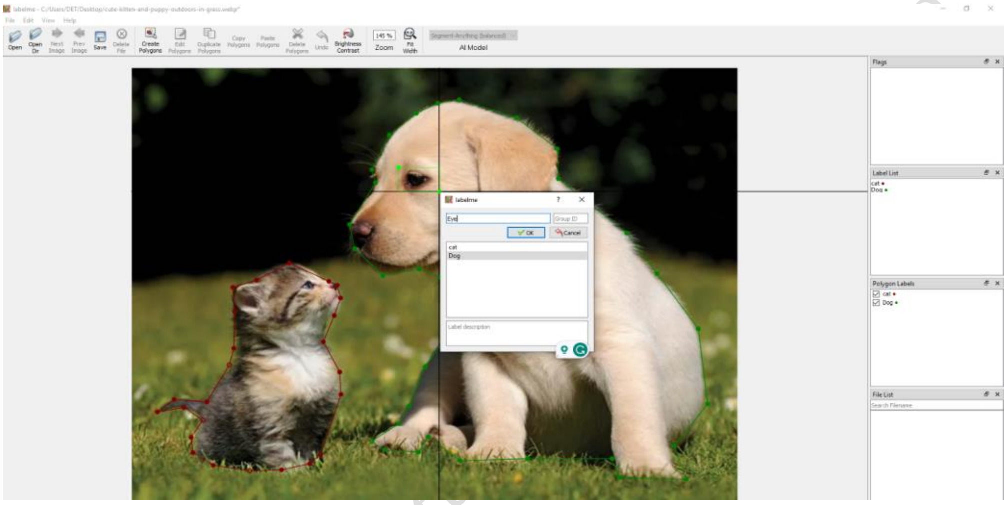


Figure 4.36: Create a bounding box

- ➢ Now, you can save your file. This is the process to annotate or label the images.

Check Your Progress

A. Multiple Choice Questions

1. What does CVAT stand for in the context of data annotation tools? (a) Computer Vision Analysis Tool (b) Computer Vision Annotation Tool (c) Computer Vision Automation Tool (d) Computer Vision Assessment Tool
2. What types of data annotation tasks can be performed using CVAT? (a) Image Classification and Image Segmentation (b) Object Tracking and Pose Estimation (c) c) Image Classification and Object Detection (d) Object Tracking and Image Segmentation
3. Which method allows you to sign up with Google or GitHub in CVAT? (a) Registration (b) Labeling (c) Annotation (d) Installation

4. Which operating systems is LabelMe compatible with when using the Python installation method? (a) Windows only (b) Linux only (c) macOS only (d) Windows, Linux, and macOS
5. What is one of the key features of CVAT as mentioned in the chapter? (a) It requires a subscription fee for use (b) It was created by Google (c) It is not suitable for labeling videos (d) It is a widely used free tool
6. Which of the following tasks is NOT mentioned as a capability of LabelMe in the chapter? (a) Object Detection (b) Classification (c) Pose Estimation (d) Optical Character Recognition (OCR)
7. What is the first step you need to take to access CVAT, as described in the chapter? (a) Start labeling (b) Create an account (c) Select the type of polygon (d) Download the software
8. How can you draw a bounding box when using the "Create Rectangle" option in LabelMe? (a) By clicking on a point and dragging the cursor (b) By drawing a circle (c) By using the "Create Polygons" tool (d) By selecting from predefined shapes
9. What is the primary purpose of LabelMe in the context of computer vision? (a) To create digital art (b) To label and annotate images for computer vision tasks (c) To generate text labels for images (d) To convert images into 3D models
10. Which of the following is true about CVAT's availability, as per the chapter? (a) It can only be used online without any restrictions (b) It is exclusively available for Linux users (c) It can be used both online with some restrictions and installed on a computer (d) It is a paid tool with no free version

B. **Fill in the blanks**

1. CVAT stands for _______________ in the context of data annotation tools.
2. You can use CVAT online with some restrictions or install it on your _______________.
3. LabelMe is a free tool for adding labels to pictures and helps create annotations for object detection, classification, and _______________.
4. LabelMe provides different ways to install it, including methods specific to different _______________ systems.
5. To label images in LabelMe, click on "Open dir" and pick a folder containing the images you want to work on, or click "Open" to choose individual _______________.
6. In LabelMe, you can draw bounding boxes by clicking on a point, then dragging your cursor to draw the _______________.
7. The primary purpose of LabelMe in the context of computer vision is to label and annotate images for various computer vision _______________.
8. To access CVAT, you are required to go through two steps: registration and _______________.
9. One of the key features of CVAT is that it is a widely used _______________ tool for adding labels to images and videos.
10. Using the Python installation method, LabelMe can be installed on Windows, Linux, and _______________ operating systems.

C. **True or False**

1. CVAT is exclusively available for offline use and cannot be used online.

2. LabelMe provides a variety of installation methods, including options specific to different operating systems.
3. LabelMe allows you to create annotations for object detection, classification, and segmentation tasks.
4. LabelMe is a paid tool, and there is no free version available.
5. CVAT and LabelMe are both tools used for audio data annotation.
6. To access CVAT, you must first create an account or sign in to an existing one.
7. CVAT can be used to label data for computer vision tasks such as object tracking and pose estimation.
8. The Python installation method for LabelMe is only compatible with Windows operating systems.
9. In LabelMe, you can draw polygons to outline images and add labels to them.
10. Both CVAT and LabelMe are open-source tools that are freely available for users.

D. **Short Question Answer**

1. What does CVAT stand for, and what is its primary function in data annotation?
2. How can you access CVAT, and what are the key steps involved in the process?
3. What types of data annotation tasks can you perform using CVAT?
4. Describe the installation methods available for LabelMe and which one is used in the provided example.
5. What are some of the shapes you can use in LabelMe to mark and label objects in images?
6. How can you start annotating images in LabelMe once you've launched the application?
7. What is the primary purpose of LabelMe in the context of computer vision?
8. How can you draw bounding boxes in LabelMe, and what is their significance in object annotation?
9. What are the primary methods for LabelMe installation?
10. Are both CVAT and LabelMe open-source tools, and how does their accessibility differ?

Module 3 — Manage and Plan Work Requirements

Module Overview

In this module, you will explore Annotation Project Management. In this, you will start with an introduction and then learn about the process of beginning a new project. You will discover the importance of defining your annotation project and managing timelines effectively. You will also look into the significance of gathering detailed work requirements and prioritizing different work areas, along with understanding the essential Annotation Guidelines.

Next, you will understand the concept of Work Accuracy in Annotation, where you will understand the use of annotations that serve their intended purpose, control timelines, and coordinate project schedules. You will also explore different workforce options, understand their limitations, and learn about selecting the right annotation tool for the job. By the end of this unit, you will be well-prepared to manage and plan work requirements for successful annotation projects.

Learning Outcomes

After completing this module, you will be able to:

- Learn how to plan, manage, and execute annotation projects efficiently to meet business objectives.
- Understand the importance of accuracy in annotation tasks and strategies to enhance precision in labeled data.

Module Structure

Session 1. Annotation Project Management

Session 2. Work Accuracy in Annotation

Session 1. Annotation Project Management

Once upon a time in a little village, three best friends named Daisy, Sam, and Anuj found a puzzling, ancient book with strange drawings. They decided to work together and annotate the book, writing down what they thought each drawing meant. As they completed the annotations, they realized the book held the secret to finding a hidden garden filled with colourful flowers. Excited, they followed the directions from the book, solving riddles and puzzles along the way. In the end, their teamwork paid off, and they discovered the magical garden, bringing happiness to their village. This adventure taught them that when friends work together and communicate well, they can achieve wonderful things. As illustrated in Figure 1.1.

Figure 1.1: Example of teamwork (Source: https://blog.icons8.com/)

In this chapter, you will learn about annotation projects, annotation guidelines, and types of required annotations, and assess the quality of annotation.

1.1 Introduction

Annotated data means that it has been labeled by humans with specific tags or labels. Nowadays, machine learning depends a lot on this labeled data, and every organization needs to have clear guidelines for how they handle annotation projects.

1.1.1 Beginning a new project

- Before you start a new annotation project, make sure to identify the important people or groups involved, known as key stakeholders.
- Set ambitious end goals for your project by gathering input from all the important stakeholders. By establishing and communicating these goals, everyone involved in the project will have a clear understanding of its purpose, and it will ensure that all efforts are directed toward achieving the same objectives.
- Effective communication is crucial throughout the entire duration of the project, ensuring that the project's end goals remain in focus, whether it's a small cross-team collaboration or a larger single-team effort.

1.1.2 Defining the Annotation Project

In defining the parameters of the annotation project, it is important to involve all stakeholders in a collaborative effort. This ensures that the project's design aligns with the desired high-level end goals.

- Identify the data that requires annotation and specify the types of annotations that are necessary for the project.
- Determine the most effective method for collecting annotations. This might involve using subject matter experts, partial automation, or micro-tasking, depending on the project's requirements.
- For planning the annotation project, make sure to consider the budget, the resources available, and the project timeline for planning and managing your annotation project.

- Identify the purpose of the annotations, such as whether they will be used to automate tasks (like training a machine learning model) or if they will be directly given to the end users.

1.1.3 Managing Timelines

Involving all stakeholders in creating project timelines is crucial. Timelines convey expectations, limitations, and interconnections, which can significantly affect both the budget and the project's overall success.

- Stakeholders should work together to develop project timelines to ensure that expectations are well-handled and different viewpoints are taken into consideration.
- Stakeholders should make sure that the timelines are precise and comprehensive. This should include sufficient information about the projects, their interdependencies, and the important milestones.
- For setting up timelines, it is necessary to allocate time for developing guidelines and providing training to the workforce.

1.2 Roles, responsibilities, and limits of the responsibilities in a working environment

At the beginning of an annotation project, it is essential to identify key stakeholders. Typically, there are three main ones:

- The project manager's role involves determining the practical use of the project and understanding its impact on the business. They decide which features to work on first based on what the users want, and they might be experts in the subject of the project.
- The manager of the annotation project has important responsibilities. They oversee the project every day, select and lead the team, make sure the data is consistent and high-quality, and they might also know a lot about the subject of the project.
- The engineering manager's main job is to put the solution that needs annotations into action. They also guide technical plans.

1.3 Importance of gathering detailed work requirements and prioritizing work areas

Requirements gathering is the process of figuring out exactly what your project needs from the beginning to the end. This process starts the project, but you will keep track of your project's requirements throughout the entire project.

There are some advantages to gathering requirements, such as:

- Makes customers and clients happier
- Speeds up the delivery of the product
- Decreases the need for reworking
- Builds trust
- Improves communication

1.3.1 Use cases, desired functionality, and intended audience

All stakeholders can see the project from the viewpoint of the people who will use it, they can better concentrate on achieving a valuable outcome and plan the data annotation process more strategically. To achieve this, the product manager should provide all stakeholders with an overview of the product to help them understand the user’s needs.

To get a clear picture of the target user’s expectations, it can be beneficial for the stakeholders to have a direct conversation with the users.

1.3.2 Success Measurement

To determine if the project has been successful, all parties need to be on the same page. This means they should agree on project milestones, such as the requirements for a basic working product.

1.3.3 Select Techniques for Consistent and Clear Communication

The important people involved in an annotation project should decide on a way to communicate clearly.

- Effective methodologies for annotation projects include the Dynamic Systems Development Method, Kanban, and Scrum.
- Stakeholders should, at a minimum, collaborate to create a shared project roadmap, maintain an issue backlog, and hold regular meetings for the annotation project.

1.3.4 Setting the Project’s Goals

To make sure the project goes well, everyone involved should work together to decide the project's main characteristics.

- Determine which data needs annotations and which types of annotations are required.
- Choose the most suitable method for acquiring these annotations, which may involve experts, automation, or task segmentation.
- Factor in the budget, available resources, and deadlines.
- Consider the intended use of the annotations, whether for automating tasks (e.g., training machine learning models) or providing them directly to users.

1.3.5 Data Identification and Types of Required Annotations

The stakeholders should identify the characteristics of the data to be annotated, the method for selecting samples, and the types of annotations that should be included. Studying the data is an important step in defining the project. It assists project members in making informed choices about the types of annotations needed by providing insights into the data's characteristics, limitations, patterns, and unique situations.

1.4 Annotation Guidelines

To keep communication clear and consistent with the workforce, use annotation guidelines as the primary reference. These guidelines should be readily available within the annotation tool, providing all necessary information for the project. For starting a new project, consider reviewing past projects for guidance and inspiration.

- For creating annotation guidelines, it is important to consider complexity and length. The guidelines should match the complexity of the project and the capabilities of the workforce responsible for annotations.
- Annotation guidelines should explain tool usage and annotation procedures.
- Alongside the guidelines, provide examples to demonstrate label usage.
- In the annotation guidelines, explain the project's main goal to help motivate the workforce.
- Ensure that the guidelines match other project documents to reduce confusing instructions for the workforce.

1.4.1 Creating, Testing, and Maintaining Annotation Guidelines

- Assigning one person to be responsible for the guidelines helps them to keep consistent and clear.
- Annotation project managers should use a short testing period where a small number of items are annotated following the guidelines. Afterward, reviewers should check if any guideline improvements are needed.
- It is a good idea to have the guidelines visible within the annotation tool, making it convenient for the workforce to access them while working.
- Annotation project managers should inform the workforce about guideline updates through written communication. This ensures that all annotators receive the same information at the same time, promoting consistent annotations. Any changes to the guidelines that may affect previously annotated data should be thoroughly considered and explicitly approved by all stakeholders. These changes can impact the project's cost and timeline.

1.4.2 Staffing and Training the Workforce

The training of annotators for an annotation project depends on factors like the project's nature, timeframe, preferences, and available resources of the managing group. However, training is a crucial step in the annotation process. It ensures that annotators fully grasp the project and create annotations that are accurate and consistent.

i. **Use the Guidelines for Training**

It is important to use the guidelines for training the workforce to maintain consistency, regardless of whether annotators join the project. If there are any additions to the guidelines, provide extra training through written materials or recorded videos. This way, everyone gets the same information.

ii. **Real-time questions from the workforce should be encouraged**

During the training phase, it is important to invite questions from the workforce. This helps clear up any confusion or misunderstandings about the guidelines as soon as possible. If answering questions in real time is not feasible, project managers should have follow-up meetings with the annotators to address any issues.

iii. **Assess quality during the training period**

It is important to evaluate quality during the training phase. Quality assurance methods can be used to provide feedback to the workforce and identify the annotator's performance.

iv. **Provide written feedback during the training period**

Annotation project managers should give written feedback during training so that the workforce can refer to it. If needed, project managers should also revise the project's guidelines based on tester feedback and testing results.

1.5 Assess Quality of Annotation

During training, it is important to analyse the quality of annotations created by the workforce. This helps provide feedback to the workforce and assess the performance of individual annotators and the workforce as a whole. Methods for assessing quality include:

- "Gold" tasks

- Annotation redundancy with targeted quality assurance (QA)
- Random QA

i. **"Gold" tasks**

"Gold" tasks are similar to answer keys that project managers have prepared in advance. They serve as useful training tools, particularly in projects where annotations must strictly adhere to predefined guidelines. "Gold" tasks help annotation project managers provide rapid feedback to the workforce because they contain pre-determined correct annotations. However, creating these tasks for training purposes takes time as project managers need to explore the data thoroughly to ensure they have a representative set of "gold" tasks.

ii. **Annotation Redundancy with Targeted Quality Assurance**

Annotation redundancy involves multiple annotators independently annotating the same work items. Annotation project managers can then focus on quality assurance by reviewing items with differing annotations, as these differences may indicate confusion or ambiguity in the guidelines and potential issues with accuracy or consistency.

While it's important to prioritize work items with differing annotations for review, annotation project managers should also review a percentage of items with identical annotations to ensure their validity and adherence to project guidelines. Moreover, the annotation tool needs to support annotation redundancy.

iii. **Random Quality Assurance**

Random QA involves randomly selecting work items from each annotator's annotations for quality checks. This approach is suitable for the annotation tool that does not support redundancy and there is a need to review a large amount of data quickly. It may not promptly pinpoint errors or areas of confusion.

1.6 Managing Process of Annotation

There are some few important steps to manage and control the process of Annotation are given below:

i. **Review the Workforce and the Annotation Tool**

It's crucial to perform one last evaluation of both the annotation tool and workforce choices. Check they are satisfied with policy constraints and meet product requirements. This includes verifying that the vendor Statement of Work (if applicable) matches the project plans, confirming that annotators possess the necessary expertise, and guaranteeing that the communication process with annotators is well-defined.

ii. **Establish attainable goals for quality, timeliness, and productivity**

Using data from the pilot and training phases, it is important to set appropriate targets for quality, timeliness (speed tasks are completed), and productivity (number of tasks done) in the annotation project. This makes it easier to establish realistic goals because these periods closely resemble the real workforce and project conditions.

For defining productivity and timeliness goals, consider the complexity of the data, guidelines, project, and how the annotation tool operates. Also, account for potential downtime due to technical problems or delays in communication with project managers. These targets may need periodic adjustments as the annotation project progresses.

1.7 Quality assurance procedure

Once the training period is completed, stakeholders should decide how to evaluate the quality of annotations during the actual project and the frequency of such assessments. Similar to the training phase, it's important to continuously check the accuracy and consistency of the annotations throughout the project. Here are a few methods for assessing annotation quality:

- "Gold-level" tasks
- Redundant annotations with focused quality assurance
- Implementing random Quality Assurance.

i. **"Gold-level" tasks**

During annotating, make sure to include "gold" tasks along with the regular work items in the annotators' queues. This helps periodically evaluate their performance. It is recommended to have at least 5% of "gold" tasks among the work items. This percentage may need to be higher if annotators are new or if many items do not require annotation, which helps ensure annotators are focused and accurate.

ii. **Redundant annotations with focused quality assurance**

Another method to assess quality during annotation is through redundancy and targeted QA. Prioritize reviewing work items that had disagreements (no consensus) among annotators, as these are likely to have errors or highlight guideline ambiguities. Consider prioritizing annotators whose work items differ the most from others, as they may have more errors in their annotations and could benefit from additional training.

iii. **Implementing random Quality Assurance**

Random Quality Assurance (QA) can also be used to check the quality during the annotation process. This means randomly selecting some work items from each annotator for assessment. The number of items in this random QA sample should be feasible for annotation project managers to review individually and determine if the annotations are valid and consistent.

1.8 Address Annotator Collaboration Directly

The preference for annotator collaboration can vary based on the workforce, project, and quality assurance method. For example, by using annotation redundancy for quality control, annotators should refrain from working together as it can influence the results negatively. Similarly, collaboration should be avoided if it could introduce bias, especially on tasks requiring subjective judgment. Additionally, if annotators have similar expertise and experience levels, working together may lead to inconsistent or inaccurate annotations. In such cases, it's better to seek clarification from annotation project managers instead of collaborating.

1.9 Workforce with Written Feedback

Providing written feedback to the workforce is important for a successful annotation project. It helps annotators to understand the guidelines better, leading to more accurate and consistent annotations. To help annotators improve their work, feedback should include examples of errors along with clear written explanations detailing the mistakes. This explanation should reference specific rules or examples from the guidelines and clarify their application.

Giving feedback in written form ensures uniform guidance for all annotators at once. This supports consistency in annotations and allows project managers to monitor. Providing feedback quickly helps to prevent future issues and keeps the project on track. Keeping a record of feedback allows managers to see if it is improving the annotations or if they need to make changes to the guidelines, project, or workforce.

1.10 Monitor Figures for Productivity, Timeliness, and Quality

Annotation project managers should evaluate the productivity, timeliness, and quality of the workforce. Productivity is about the number of tasks completed, timeliness is how quickly they're done, and quality measures how well they're done. These statistics help with project management and deadlines.

The annotation tool should automatically calculate and keep track of these statistics for each annotator and the entire team. If it cannot do this, it should allow easy exporting of data to another platform for calculation. For example, information about tasks identified as errors by project managers should be available for quality assessments. To measure timeliness and productivity, data from time-stamped audit trails for each task should also be included.

1.11 Technical Support for Annotation Tool

The stakeholders should determine who will handle technical issues with the annotation tool during the project, especially if the workforce will use it outside of regular business hours. They should plan for potential scenarios like the server or tool going offline unexpectedly and establish procedures for addressing such issues promptly. This ensures that annotators can continue their work smoothly and prevents project delays.

1.12 Completion of Annotation Project

- After the project is finished, it is crucial to check all the annotations. This makes sure that the annotations are accurate and consistent before they are used for further purposes.
- If you need more annotations for your project even after completing the initial project, it is essential to think about improving the project's planning, training, and workforce to enhance the efficiency of the next round of annotation collection.
- It is essential to have processes for identifying data drift and anomalies that might need extra annotations.

Summary

- Annotated data is labeled by humans with specific tags or labels.
- Machine learning relies on labeled data, necessitating clear annotation project guidelines.
- Set ambitious project goals and communicate them effectively.
- Plan strategies for consistent communication among stakeholders.
- Identify data requiring annotation and specify annotation types.
- Choose effective methods for collecting annotations.
- Understand the user's perspective to plan data annotation effectively.

Check Your Progress

A. **Multiple Choice Questions**

1. What does "annotated data" mean in the context of machine learning? (a) Data labeled with tags or labels by humans (b) Data generated by AI algorithms (c) Data collected by automated systems (d) Data without any labels
2. What is the purpose of "gold tasks" in assessing annotation quality? (a) Tasks that are annotated in gold color (b) A set of predefined correct annotations (c) Tasks related to gold mining (d) Tasks that require special attention
3. How can project managers evaluate the productivity, timeliness, and quality of the annotation workforce? (a) By ignoring these metrics (b) By manually calculating statistics (c) By using an annotation tool with automated tracking (d) By asking annotators for self-assessment
4. When defining the parameters of an annotation project, why is it important to involve all stakeholders? (a) To increase project complexity (b) To ensure project design aligns with high-level goals (c) To avoid project delays (d) To assign blame in case of project failure
5. In the context of an annotation project, what does "QA" stand for? (a) Quality Analysis (b) Quick Assessment (c) Quantitative Analysis (d) Quality Assurance
6. What does "annotated data" refer to in the context of machine learning? (a) Data that is encrypted for security (b) Data that has been labeled with specific tags or labels by humans (c) Data that is processed by artificial intelligence algorithms (d) Data that is stored in the cloud
7. Which method can be used to assess the quality of annotations during an annotation project? (a) Randomly selecting work items from each annotator's annotations (b) Assigning gold tasks to annotators (c) Redundant annotations with targeted quality assurance (d) All of the above
8. What are the three main aspects that annotation project managers should evaluate in an annotation project? (a) Speed, budget, and resources (b) Quality, timeliness, and productivity (c) Complexity, guidelines, and project timeline (d) Redundancy, automation, and accuracy
9. Which method can be used to assess annotation quality when the annotation tool does not support redundancy? (a) Random Quality Assurance (b) Gold tasks (c) Reviewing identical annotations (d) Collaboration
10. Which of the following is NOT one of the aspects that annotation project managers should evaluate? (a) Complexity of the data (b) The color of the annotation guidelines (c) Project Timeline (d) Guidelines' adherence

B. **Fill in the blanks**

1. The annotation tool should _____________ calculate and keep track of these statistics for each annotator and the entire team.

2. Annotation project managers should evaluate the productivity, timeliness, and _______ of the workforce.
3. Providing written feedback to the workforce is important for a ___________ annotation project.
4. The stakeholders should determine who will ________ technical issues.
5. Random Quality Assurance can also be used to check the _________ during the annotation process.
6. It is recommended to have at least 5% of __________ tasks among the work items.
7. "Gold" tasks are similar to _________ keys that project managers have prepared in advance.
8. Annotation redundancy involves ________ annotators independently annotating the same work items.
9. Giving feedback in __________ form ensures uniform guidance for all annotators.
10. Random QA involves __________ selecting work items from each annotator's annotations for quality checks.

C. **True or False**

1. Annotated data is essential for machine learning and is labeled by humans with specific tags or labels.
2. Defining the parameters of an annotation project should not involve all stakeholders.
3. Annotation guidelines should not match other project documents to avoid confusion.
4. Random Quality Assurance (QA) is a suitable method for assessing annotation quality when redundancy is not supported.
5. Collaboration among annotators is always beneficial for annotation projects.
6. Assigning one person to be responsible for annotation guidelines is unnecessary.
7. Checking annotations for accuracy is essential after project completion.
8. Effective communication is not crucial in annotation projects involving a small team.
9. Written feedback helps annotators improve their work.
10. Effective communication is crucial in annotation projects.

D. **Short Question Answer**

1. Explain Roles, responsibilities, and limits of the responsibilities in a working environment.
2. Explain the Importance of gathering detailed work requirements and prioritizing work areas.
3. Define the Annotation Project goals.
4. How to Begin a New project?
5. What are the techniques for Consistent and Clear Communication?
6. Explain setting a Project's Goals.
7. Explain Annotation Guidelines.
8. What are the methods for assessing the quality of Annotation?
9. Explain the important steps to manage and control the process of Annotation.
10. Explain Random Quality Assurance.

Session 2. Work Accuracy in Annotation

Avanti, a student, joined an annotation project about historical documents. She initially struggled to understand the old texts but sought help and practised regularly. Her annotations improved over time, and she discovered an important historical fact. Her dedication showed that accuracy in annotations can uncover hidden treasures from the past. As illustrated in figure 2.1.

Figure 2.1.: Avanti working

In this chapter, you will learn about work accuracy in annotation, selecting the workforce type, selecting an annotation tool, anomalies, and data drift.

2.1 Work Accuracy in Annotation

The current set of annotations should be assessed by stakeholders to determine if they are enough to create a basic usable product. It is assumed that more annotations will be added later, especially for machine learning models through active learning.

Budget and resource availability can significantly influence how a project is defined, as they may restrict the choice of workforce and overall project scope. The choice of the workforce can affect the quantity and nature of annotations produced, especially the data's complexity and characteristics. The project's schedule can also influence how we define it. If we have a tight schedule, we may not be able to do as many different types of annotations.

2.1.1 Intended use of the annotations

The way annotations are utilized should be based on the type of data and the project's ultimate objectives. Stakeholders should consider the potential benefits of automation, particularly if annotations are sent directly to end users. If the project's goal is to predict the future with new data, take that into account. If it is about identifying human-recognizable patterns in the data, think about using a rules-based model.

2.1.2 Controlling Timelines

Creating project timelines involves all stakeholders communicating expectations, limitations, and interdependencies, which can have a significant impact on the project's budget and outcome.

- Collaborate with stakeholders to create project timelines that consider different viewpoints and maintain clear expectations.

- Ensure that timelines are thorough and include essential details about projects, dependencies, and key milestones.
- Include time for policy development and worker training timelines.

2.1.3 Coordinate Project Schedules

Project timelines should reflect the project's significance, main objectives, and practical constraints. Collaborate with all involved parties to create an initial timeline with clear descriptions of key milestones. For reviewing project timelines, each stakeholder should provide feedback based on their unique viewpoints:

- The annotation project manager should ensure that the deadlines are realistic, considering their knowledge of the dataset, the project's complexity, and the available workforce.
- Any issues or uncertainties related to the data or annotation process should be openly discussed with all involved parties and, if necessary, documented as potential risks.
- The engineering manager should make sure that the project timelines coordinate with the development needs of the product solution.
- The product manager needs to make sure that the project schedule coordinates with product requirements, user expectations, and deadlines.

2.1.4 Ensure Timelines are Clear and Detailed

After the initial project discussions, stakeholders should make a detailed schedule. This schedule should have specific points or goals for different parts of the project, such as:

- Getting the data ready and preparing the annotations.
- Making the solution.
- Developing the product.

Dependencies between milestones (like needing a specific amount of annotated data before starting the solution) should be made clear. Each milestone should have a description and examples. And they should have set timeframes, like in sprints.

Milestone times adapt to project complexity and data volume. Large annotation tasks or extensive machine learning model testing requires multiple sprints. The annotation and engineering managers should consult their past experiences or experienced managers for timeline estimations and communicate clearly with the product manager. Milestones can be divided into Program Increments for better time allocation for dealing with multiple timelines.

2.1.5 Include Time for Creating Guidelines and Training the Workforce

Allocate time for guidelines and training to avoid confusion and delays. Communicate complexity to stakeholders for understanding.

2.2 Selecting the Workforce Type

Choosing the workforce for the annotation project depends on data privacy, project complexity, and budget limitations. The annotation project manager should be involved in selecting the workforce, and the decision should not be made by product or engineering managers alone.

- The workforce options are important because each type has its advantages and drawbacks.
- The limitations of choosing the workforce are important.

2.2.1 Understand the Workforce Options

To make sure we choose the right workforce for an annotation project, we need to look at different workforce options and consider their pros and cons. There are mainly three types of workforces:

- **The Crowd**

Crowdsourcing is cost-effective and quick to start, ideal for simple tasks. Micro-tasking is used for quality control, but supervision is challenging due to remote, platform-based training.

- **Vendors**

Vendors offer versatility, expertise, and confidentiality. They excel in complex projects with close supervision and data privacy needs.

- **Full-Time Employees**

Full-time employees are vital for complex, confidential, and specialized annotation projects, especially if data is sensitive.

2.2.2 Understand the limitations of Workforce Selection

a) **Data Privacy Restrictions**

For choosing the workforce and annotation platform for a project, it is important to consider the type of data involved. Data privacy, security, and intellectual property issues will determine which workforce type and annotation platforms can be used. For seeking guidance from official sources like an organization's Internal Review Board or Legal/Compliance departments.

b) **Project Complexity**

For deciding on the workforce for a project, it's important to consider the complexity of the project. Consider factors like the quantity of work that needs to be done, the complexity of annotations, whether subject matter expertise is needed, and the amount of supervision and quality assurance that will be required. Also, consider the time needed to manage the project effectively.

c) **Annotation Volume**

Annotation volume refers to the number of annotations you need for the project. If you need a lot of annotations, it is a good idea to choose a workforce that can handle a big workload, like the Crowd or vendors who can scale up easily. If privacy restrictions make it impossible to use workforce types like the Crowd or vendors, you might need to change the project timeline or find a way to achieve the minimum viable product with fewer annotations.

d) **Annotation Complexity**

Annotation complexity involves the level of precision and detail the annotations need to be. For example, it includes deciding which specific parts of text or data should be annotated to show its importance or relevance.

e) **Required Subject Matter Expertise**

Subject matter expertise means having knowledge that goes beyond what can be found in written documents or manuals. It involves having a deep understanding of a specific field or topic. In some cases, the workforce needs to already have specific knowledge about a subject because teaching them that knowledge would take too much time and money.

f) **Project Lead Time**

The time available before the project begins should be considered to choosing a workforce, as it can affect whether it is feasible to hire a vendor.

2.3 Selecting an Annotation Tool

For deciding an annotation tool, stakeholders must carefully consider and balance several factors:

- Engineering and annotation project managers should aim to use established infrastructure. It's better to choose an existing and well-supported annotation platform or tool rather than creating a new one or using a makeshift process.
- The main reasons for choosing a tool should be to protect data privacy and provide access to the workforce.
- Stakeholders should also assess the technical backend needs.
- It's also important to consider the tool, which will support project management requirements.
- Stakeholders should also define the preferred user interface and user experience features for both the annotation tool and the project interface.

2.3.1 Use Standard Infrastructure

Implementing and maintaining new annotation tools can result in technical challenges and governance problems in the long run. Whenever feasible, opt for an existing annotation platform or tool that is already supported by the organization.

2.3.2 Ensure Data Privacy and Access by the Workforce

Intellectual property, data privacy, and security concerns affect the choice of workforce and annotation tools for a project. Stakeholders should select a tool that fits into the project workflow and is accessible to the chosen workforce.

2.3.3 Evaluate Technical Back-End Requirements

The chosen tool should, at the very least, allow for the smooth uploading, downloading, storage, and processing of data. It should also be compatible with standardized and user-friendly data formats such as JSON and XML.

2.3.4 Facilitate Project Management Needs

The tool should have built-in project management features that allow annotation project managers to keep track of the project's progress, communicate with the workforce, spot any problems with the annotation process, and ensure quality control.

In particular, the tool should be able to meet the following requirements for annotation project managers:

- Share written guidelines and updates with the workforce.
- Manage work queues, including setting priorities and orders.
- Access audit logs for work items.
- Carry out quality assurance, including using annotation redundancy and "gold" tasks.
- Give feedback to the workforce.
- Monitor quality, productivity, and timeliness metrics.

2.3.5 Establish Desired UI/UX Features

The chosen tool should offer a customizable interface that matches the needs of the annotation project. This helps the workforce perform annotations efficiently and accurately. It's important to make sure that the tool's user interface features don't influence the workforce in a way that could affect the accuracy or consistency of the annotations.

The workforce should be able to easily:

- Read the project guidelines and updates;
- View the relevant data in the work item;
- Add, edit, or delete annotations;
- Move through each work item or work queue, including going back to previously completed work items if necessary;
- Flag work items with potentially incorrect or unsuitable data if needed.

2.3.6 Identify anomalies and data drift

Regardless of the future purpose of the annotations, ongoing funding and resources are needed for quality control and monitoring during continuous annotation. It's crucial to have processes in place to identify the following two things:

Data drift: Data drift refers to the gradual changes that occur over time in the distribution of annotation labels or other aspects of the data. This can cause errors in rule-based systems or machine-learning models. To keep models and solutions accurate, continuous annotation is necessary because data is never static.

Anomalies: Anomalies are sudden and often temporary changes in data caused by external events, unlike the gradual changes of data drift. Detecting anomalies is important, and it might be necessary to shift from automated to human-based processes or increase annotations.

SUMMARY

- Assess current annotations and plan for future additions, especially for machine learning models.
- Budget and resource availability impact workforce selection and annotation project scope.
- Align annotation use with data type and project objectives, exploring automation options.
- Collaboratively create comprehensive project timelines that include dependencies and milestones.
- Coordinate project schedules based on significance, objectives, and constraints, with stakeholder input.
- Develop detailed schedules with specific goals, milestones, and timeframes, considering project complexity.
- Allocate time for creating guidelines and training the workforce to prevent delays.
- Choose the workforce type based on data privacy, project complexity, and budget limitations.
- Understand workforce options like the Crowd, vendors, and full-time employees, considering project needs.

Check Your Progress

A. **Multiple Choice Questions**

1. What should stakeholders assess to determine if annotations are sufficient for creating a basic usable product? (a) Budget constraints (b) Complexity of the dataset (c) The current set of annotations (d) Project schedule
2. Which factor can significantly influence how a project is defined, including the choice of workforce and project scope? (a) Project Timeline (b) Data privacy (c) Stakeholder preferences (d) Annotation tool selection
3. What is a key consideration when creating project timelines? (a) Stakeholder hierarchy (b) Including as many milestones as possible (c) Collaboration and clear expectations (d) Excluding policy development and worker training time
4. In the context of coordinating project schedules, who should ensure that deadlines are realistic? (a) The engineering manager (b) The annotation project manager (c) The product manager (d) The project coordinator
5. Which workforce type is ideal for simple tasks that require quick scalability? (a) Vendors (b) Full-time employees (c) The Crowd (d) Remote freelancers
6. What is a critical consideration when selecting a workforce for an annotation project? (a) The complexity of annotations (b) The availability of open-source annotation tools (c) The size of the project budget (d) The location of the workforce
7. What is the main purpose of collaborating with stakeholders when creating project timelines? (a) To assign tasks to each stakeholder (b) To maintain clear expectations and consider different viewpoints (c) To ensure all stakeholders are satisfied with the timeline (d) To speed up the project by minimizing discussions
8. How can budget and resource availability impact the definition of an annotation project? (a) They do not influence the project's scope (b) They can expand the range of workforce options (c) They only affect the project's timeline (d) They may restrict the choice of workforce and overall project scope
9. What is the purpose of defining milestones with set timeframes in a project schedule? (a) To create flexibility in project timelines (b) To create ambiguity and encourage creativity (c) To enable stakeholders to set their timelines (d) To provide specific goals and deadlines for different project phases
10. Why is it important to allocate time for creating guidelines and training the workforce in annotation projects? (a) To reduce complexity (b) To avoid involving stakeholders (c) To prevent confusion and delays (d) To speed up the annotation process

B. **Fill in the blanks**

1. Data drift refers to the ________ changes that occur over time in the distribution of annotation labels.
2. Anomalies are sudden and often _______ changes in data caused by external events.
3. Annotation complexity involves the level of _______ and the detail of the annotations.
4. Implementing and maintaining new annotation tools can result in _______ challenges and governance problems.

5. Choosing the workforce for the _________ project depends on data privacy, project complexity, and budget limitations.
6. Creating project timelines involves all stakeholders to _________ expectations, limitations, and interdependencies.
7. Project timelines should reflect the project's significance, main objectives, and ________ constraints.
8. The product manager needs to make sure that the _______ schedule coordinates with product requirements, user expectations, and deadlines.
9. Milestones can be divided into _______ Increments for better time allocation when dealing with multiple timelines.
10. Vendors offer versatility, ___________, and confidentiality.

C. **True or False**

1. Continuous annotation is necessary to address data drift and anomalies in the data.
2. The size of the project budget is the primary consideration when choosing a workforce for annotation projects.
3. When selecting a workforce, annotation volume and annotation complexity should not be considered.
4. Automation is beneficial for all annotation projects, regardless of their objectives or data type.
5. Detailed project schedules should include clear timeframes for milestones and dependencies.
6. Data privacy and security concerns do not influence the choice of annotation tools.
7. The complexity of annotations and the need for subject matter expertise are important factors to consider when choosing a workforce.
8. The choice of workforce for annotation projects is solely the responsibility of product or engineering managers.
9. Project timelines should involve collaboration with stakeholders to ensure clear expectations and consider different viewpoints.
10. Annotations may need to be updated or expanded, especially in machine learning projects through active learning.

D. **Short Answer Question**

1. Explain Work Accuracy in Annotation.
2. Explain the use of the annotations.
3. How to coordinate Project Schedules?
4. How to select the Workforce Type?
5. Explain Workforce Options.
6. Explain limitations on Workforce Selection.
7. What are the several factors for selecting an Annotation Tool?
8. What do you understand about anomalies?
9. How to establish Desired UI/UX Features?
10. Explain data drift.