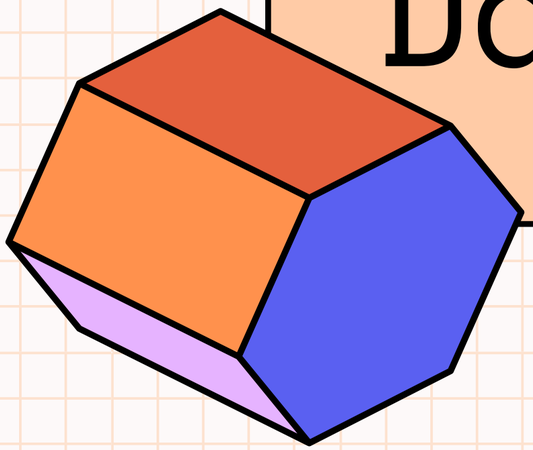
Labelling of Dataset

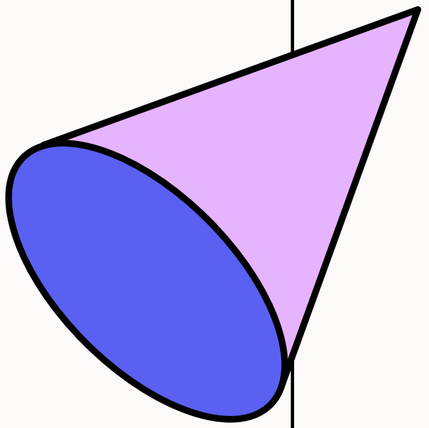
Team Members:

Saumil Shah-115
Virag Shah -116
Kathan Shah-112

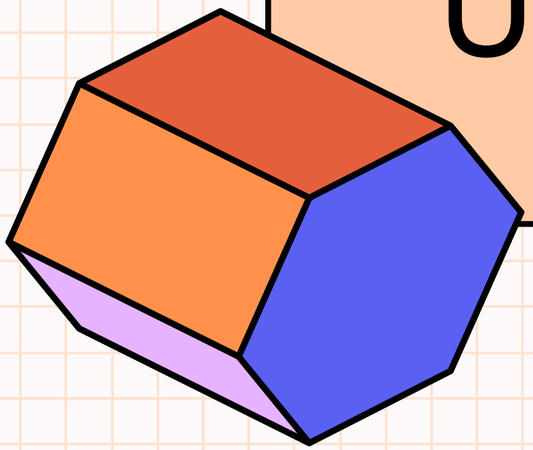
Introduction to Dataset Labelling



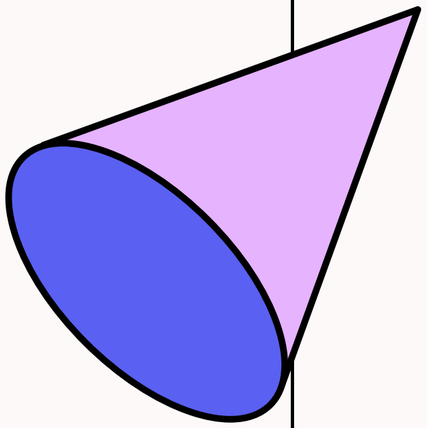
Dataset labelling is defined as, dataset labelling is the process in machine learning in which raw data such as images, text files, videos, etc, can be identified and to provide the context it allows to add one or more labels that are meaningful and informative so that the model of machine learning can learn something from it, it also allows to label a dataset in machine learning and in supervised learning the dataset labelling is the important part of data pre-processing so for classification it can label the input and output of to provide learning basis for future data processing.



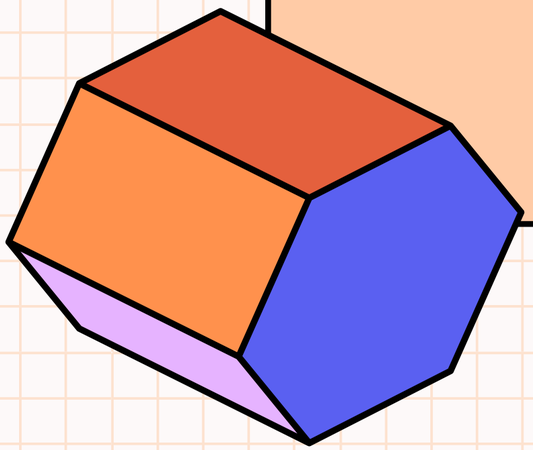
Labeled Data vs. Unlabelled Data



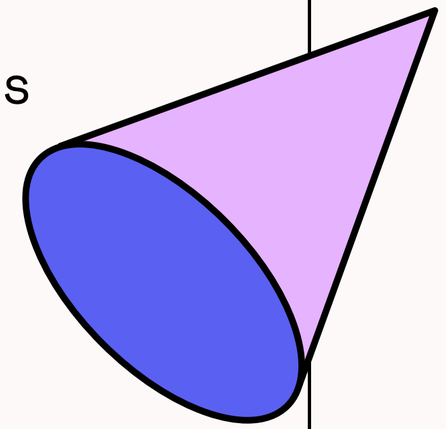
- In data labeling, data is labeled, but in machine learning, both labeled and unlabelled data are used. So, what is the difference between them?
- Labeled data is data that has some predefined tags such as name, type, or number. For example, an image has an apple or banana. At the same time, unlabelled data contains no tags or no specified name.
- Labeled data is used in Supervised Learning techniques, whereas Unlabelled data is used in Unsupervised Learning.
- Labeled data is difficult to get, whereas Unalabled data is easy to acquire.



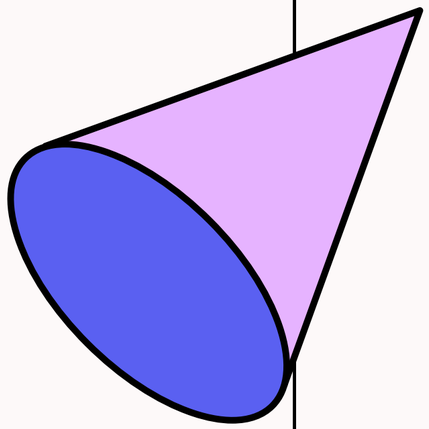
What is DataSet Labelling?



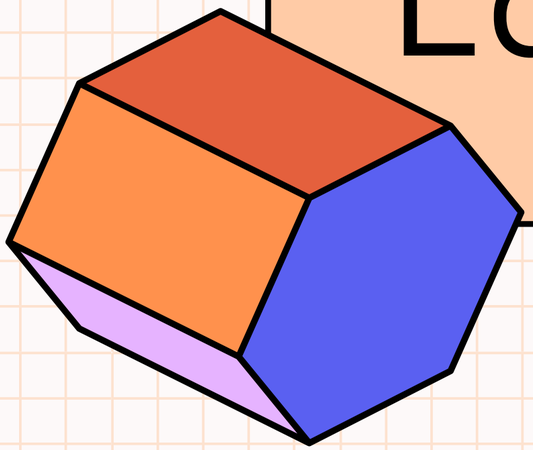
- The dataset labelling is the machine learning process to identify the raw data that also allows labelling the informative data, as well as meaningful data to provide context to it, and machine learning can use that data to learn from it.
- The labelling of data is the critical process because it can add context to data before using that in the training model, so that the data labelling helps us to select a correct approach when we want to improve the scalability factor and the quality factor,



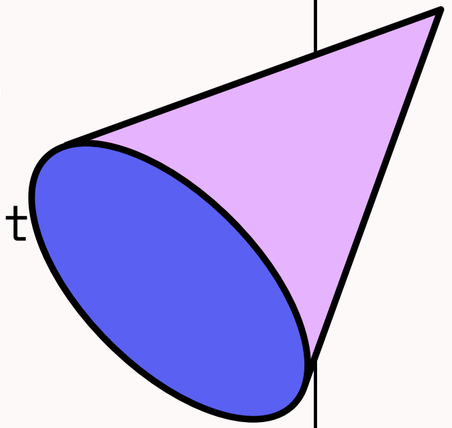
- **example**, if we have any photo then labelling works to indicate whether the photo has animal or car and that word may come out in recording of the audio this also happen if we have an x-ray report in which it about x-ray report of having a tumor, so the dataset labelling is very important when we have a variety of use cases having the computerized vision, processing of the natural language, and recognition of the speech.



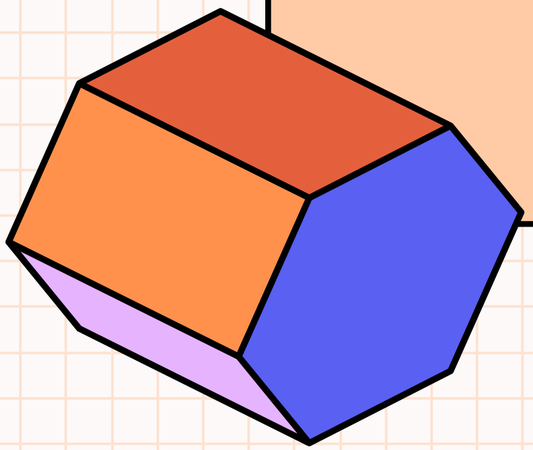
How does Data Labelling works?



- Today, most practical machine learning models utilize supervised learning, which applies an algorithm to map one input to one output.
- For supervised learning to work, you need a labeled set of data that the model can learn from to make correct decisions. Data labeling typically starts by asking humans to make judgments about a given piece of unlabeled data.
- For example, labelers may be asked to tag all the images in a dataset where “does the photo contain a bird” is true. The tagging can be as rough as a simple yes/no or as granular as identifying the specific pixels in the image associated with the bird.

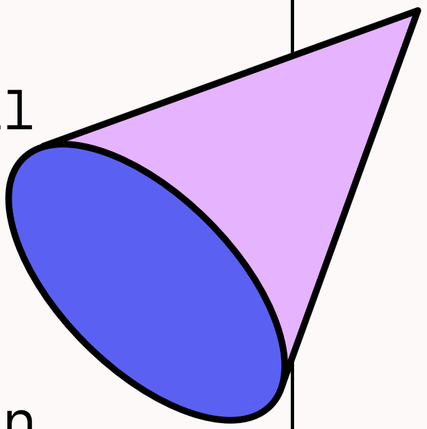


Types of Data Labelling



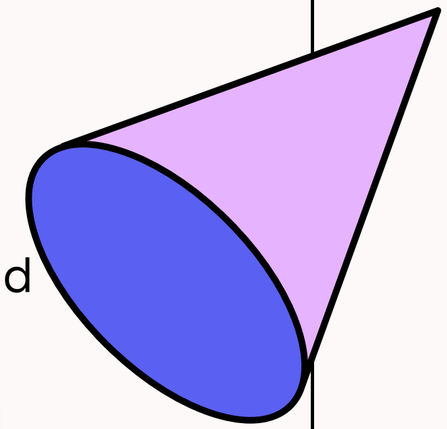
1. Natural Language Processing

Natural language processing is a part of artificial intelligence in which machines can understand natural language, we can say that it is like an intermediate between humans and machines which allows the machines to understand and operate human language invaluable way, the working of it depends on the application which is being developed, it uses hidden models to convert the words into the text and to understand the language and context it divides each part of the sentence into part of speech.



2. Audio Processing

This is also a type of data labelling in which audio processing can convert all kinds of sounds into machine learning format, it creates different types of noises and sounds of breaking glass, etc, in audio processing first the audio is converted into written text and then taking deeper information the audio can be categorized into a dataset and it allows to add different tags according to the audio, as per the characteristics of the dataset segmentation divide the objects into different parts.



Thank You!