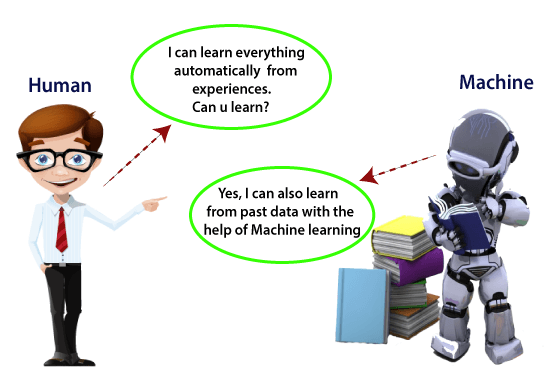
# MACHINE LEARNING

Machine learning is about extracting knowledge from data.

*Enables machines to learn from past data or experiences without being programmed.*



## Datasets for Machine Learning

The field of ML depends on datasets for preparing models and making precise predictions.

Datasets assume a vital part in the progress of AIML projects .

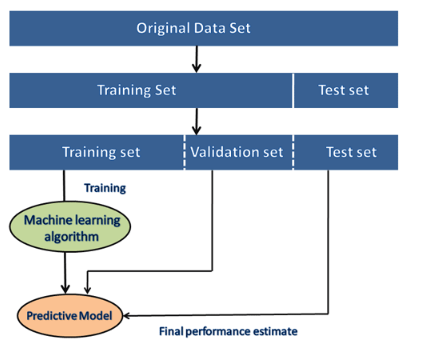
The most supported file type for a tabular dataset is **"Comma Separated File,"** or **CSV.** But to store a "tree-like data," we can use the JSON file more efficiently.

### Types of data in datasets

* **Numerical data:**Such as house price, temperature, etc.
* **Categorical data:**Such as Yes/No, True/False, Blue/green, etc.
* **Ordinal data:**These data are similar to categorical data but can be measured on the basis of comparison.

### Data Pre-processing:

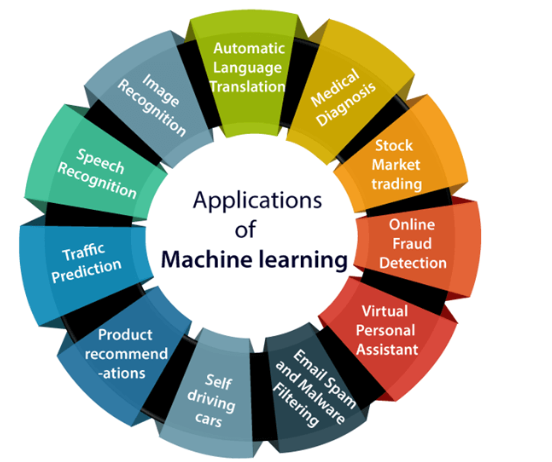
Data pre-processing is a fundamental stage in preparing datasets for machine learning. It includes changing raw data into a configuration reasonable for model training.



In building ML applications, datasets are divided into two parts:

* **Training dataset:**
* **Test Dataset**

### Applications of ML



### Free Popular sources for Machine Learning datasets

1. Kaggle Datasets

2. UCI Machine Learning Repository

3.Datasets via AWS

4. Google's Dataset Search Engine

5. Microsoft Datasets

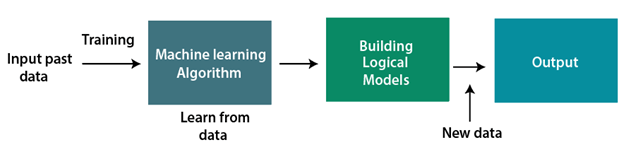
6. Awesome Public Dataset Collection

7. Government Datasets

8. Computer Vision Datasets

9. Scikit-learn dataset

## Machine Learning

A machine learning system builds prediction models, learns from previous data, and predicts the output of new data whenever it receives it.   


### TYPES:

1.Supervised Learning

2.UnSupervised Learning

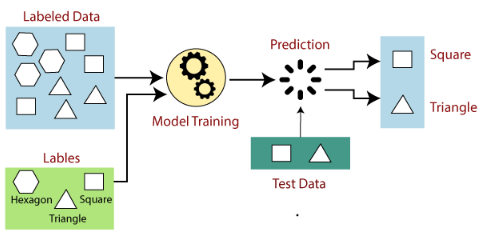
3.Reinforecement Learning

#### SUPERVISED LEARNING

In supervised learning, **sample labeled data are provided to the machine** learning system for training, and the system then predicts the output based on the training data.

Here input & output are given to the model to train. Input data is already tagged with the correct output.

Eg. Image recognition in google image



Supervised again classified into

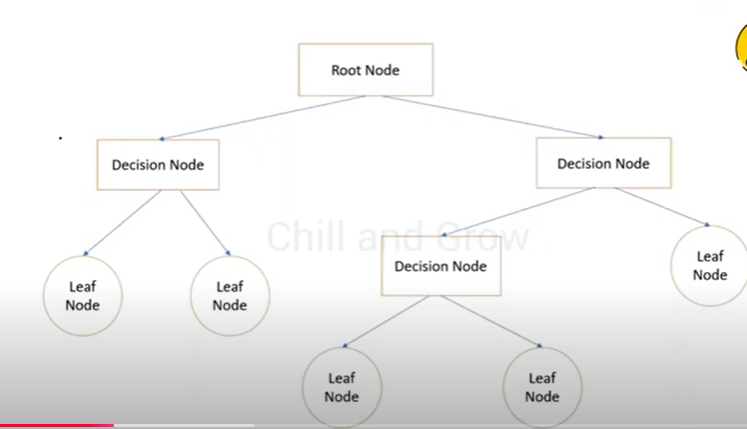
#### 1.Classification

Classification algorithms are used when the output variable is **categorical**, which means there are two classes such as **dog-cat**, Male-Female

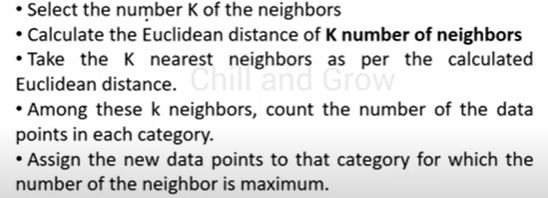
**1.Decision Tree**

In Decision data set will contain input & output , based on that , decision tree will prepare & predict the output

Attribute selection method



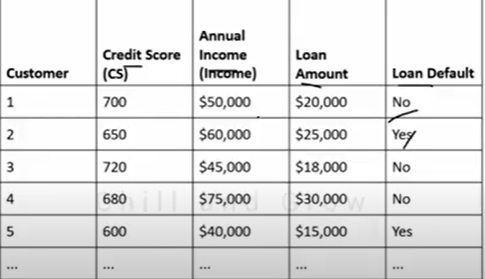
**2 KNN (K nearest neighbhour ) Algorithm**:

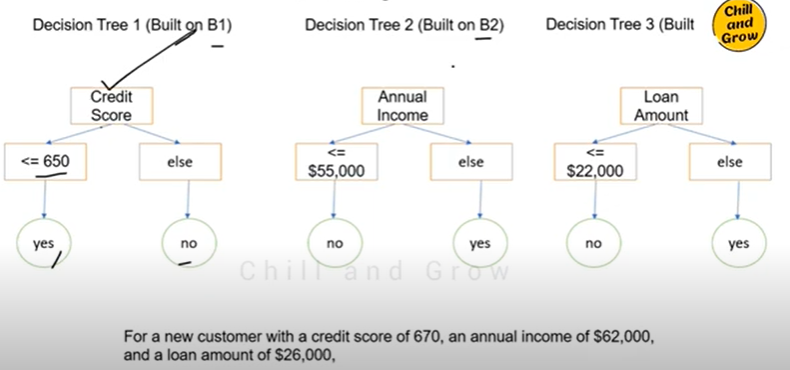


We train one or more categories of data. By giving new set will determine which category it belongs to. Shortest distance, majority of the category

**3 Random Forest ALgorithm**:

Creates multiple decision tree during training and combines their prediction to make more accurate

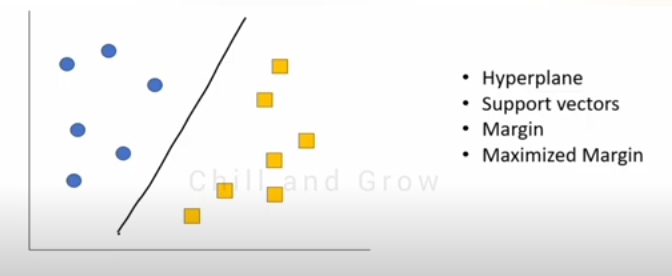




This use bagging technique – group the random data and combines the highest score of each group

**4 Support Vector Machine Algorithm:**

To Categorize the dataset, there are linear SVM & non-linear SVM

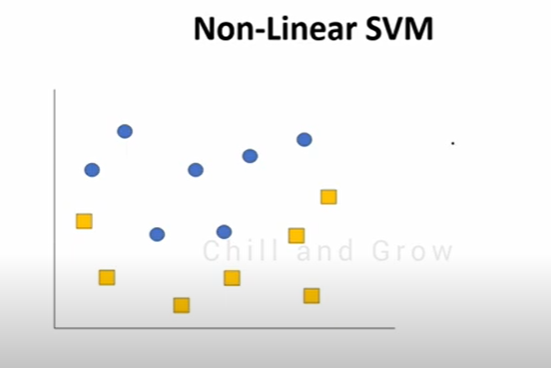
**Linear SVM**

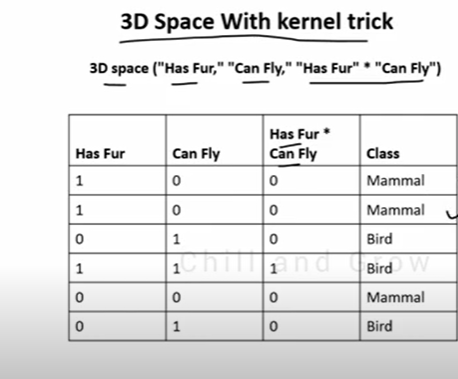
Draw a line between them that line is hyperplane.The very first data after the line is support vectors.

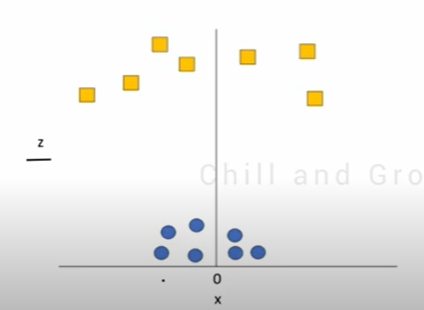
Training the features of data . And giving one data & asking it to predict

**Non-liner SVM**

Here we connect separate two category.so we need to connect 2d to 3d dataset and follow the same in linear SVM







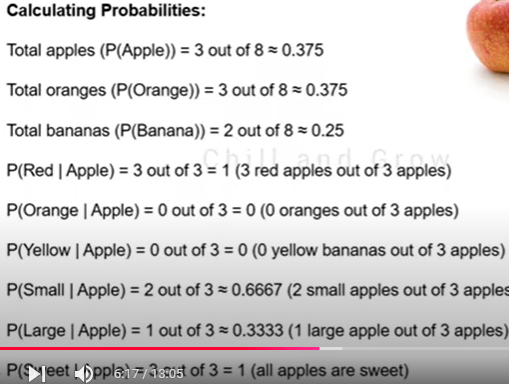
**5 Naïve Bayes' Classifier**

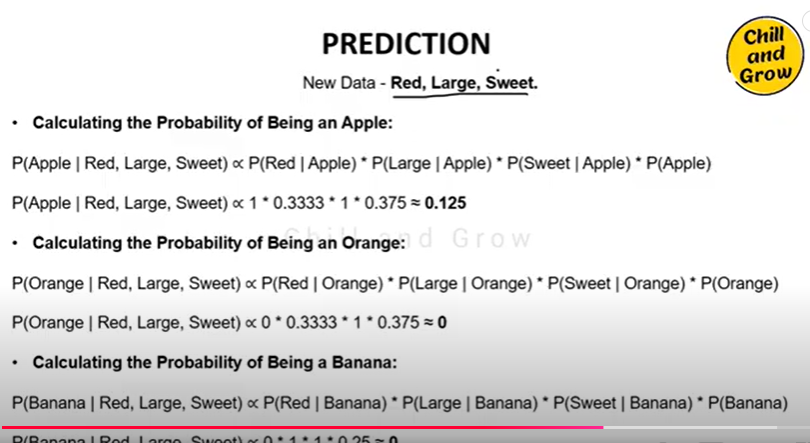
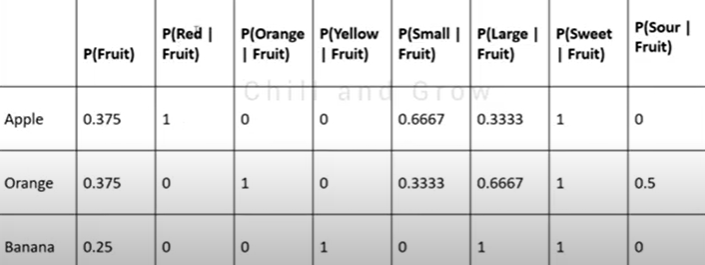
It is based on bayes theorem in probability

Naïve Bayes Classifier Algorithm

From







#### 2.Regression

Regression algorithms are used if there is a **relationship between the input variable and the output variable.**

Eg: weather forecast

In regresion, our dataset will have input & output.

Using those will draw a graph & plot, here can calculate the shortest distance between them.shorter the distance the more ability to match.

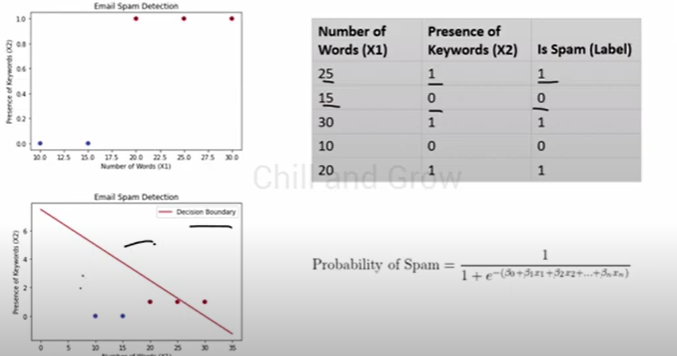
2.1 **Linear Regression**:

Relationship between input & output variable . Simply plot . Draw a line using y=mx+b equation where m,b can be find using **Least square method**.The line drawn should be nearest to the points plot.If new data comes, it will use the equation to predict the output using the existing data

2.2 **Logistic Regression:**

Relationship between one or more input variable using specific keywords.

Eg Sms Scam or not



Output will return nearest of 0 or 1 using sigmoid function. 0- not spam, 1-spam

#### UNSUPERVISED LEARNING

The training is provided to the machine with the set of **data that has not been labeled**, classified, or categorized, and the algorithm needs to act on that data without any supervision.

We dont give output data to train the model

The machine tries to find useful insights from the huge amount of data.

Supervised again classified into

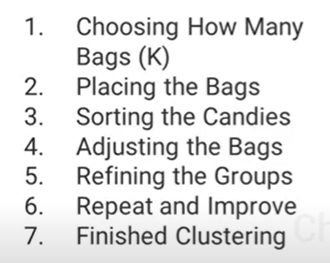
##### 1.Clustering

Clustering means grouping of data. Group the unlabelled dataset.

***"A way of grouping the data points into different clusters, consisting of similar data points.*** *The objects with the possible similarities remain in a group that has less or no similarities with another group.****"***

**1 K-means clustering**:

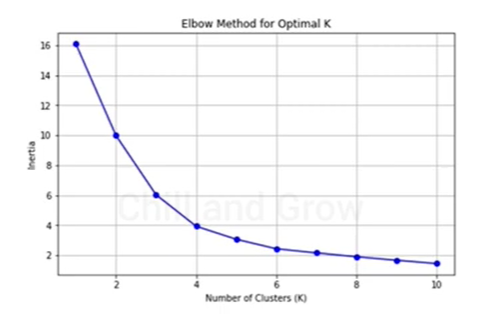
<https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning>



Here K is group (number of clusters). K-means **finds the patterns in the data** which is helpful **to group the data.**

Elbow method to choose k

In the below graph x- k choosen, y distance between data and cluster center



Bend area point will be the correct way of choosing K.

**2 Hierarchical Clustering**

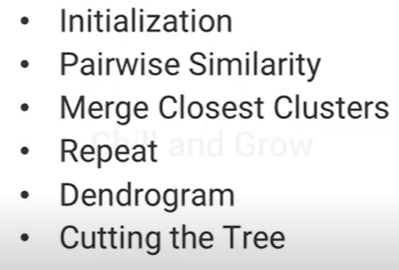
<https://www.javatpoint.com/hierarchical-clustering-in-machine-learning>

Used to group unlabeled data into a group.

Bottom –top approach is the common approach to cluster data. This is known as agglomerative - It will make every data as a single cluster, find similarities by closet distance between clusters between them and merge the similar clusters, this repeat

Finding closest distance between clusters can be done by technique called linkage-Single (closest distance between clusters), complete (farthest),

Tree-like structure after merging is called dendrogram.



##### 2.Association

Learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

#### REINFORCEMENT LEARNING

Reinforcement learning is a **feedback-based learning** method.

The agent learns automatically with this feedback and improves its performance.

To work with machine line, prerequisites are:

Fundamental knowledge of probability and linear algebra.

The ability to code in any computer language, especially in Python language.

Knowledge of Calculus, especially derivatives of single variable and multivariate functions.

**Mean** - The average value

**Median** - The mid-point value

**Mode** - The most common value

STANDARD DEVIATION

Standard deviation is a number that describes how spread out the values are.

A low standard deviation means that most of the numbers are close to the mean (average) value.

A high standard deviation means that the values are spread out over a wider range.

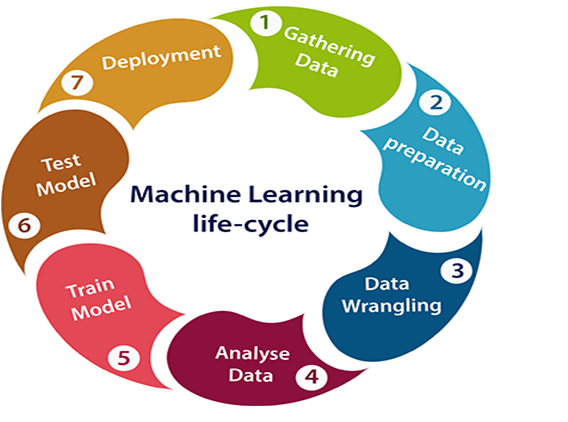
PERCENTILE:

It is a percentage of a particular value in a data set or below

Percentiles are used in statistics to give you a number that describes the value that a given percent of the values are lower than.

meaning that 75% of the people are 43 or younger.

**Lifecycle of ML**



**Bias**:

Bias – constanly predict which is not true value

Bias happens because of training very less data .We need to train a machine with huge set of data to get less biased model.

**Variance:**

Predict false data when giving new set of data to predict. This is because machine doesnot understand the data properly

**UnderFitting**:

Predict wrong data from both train and test data

**OverFitting**:

It predict only on trained data and in new data

# NUMPY

# PANDAS

# PREPROCESSING TECHNIQUES

# NLP (Natural Language Processing)

Nlp – Teaching computers to **understand human natural language** that can be tamil,english,chinese so on.

It helps to understand & analyse the text and respond back

Eg. Language translator, auto complete words ,Chatbox

The best example is **chatBox** – we can ask any question to it & it will respond back to us as a reply. Behind the process,the text we are asking is preprocessed first.

## Text Preprocessing

**Tokenization**

Breaking down text into smaller units, such as words or sentences.

**Stemming & Lemmatization**

Reduce words to their base or root forms.

Stemming will break words to base word -> running – run

Lemmatization will give a simple synonym of the word -> went – go

**Stemming Types in NLTK lib**

* [Porter Stemmer](https://www.geeksforgeeks.org/introduction-to-stemming/)
* [Lovins Stemmer](https://www.geeksforgeeks.org/introduction-to-stemming/)
* [Dawson Stemmer](https://www.geeksforgeeks.org/introduction-to-stemming/)
* [Krovetz Stemmer](https://www.geeksforgeeks.org/introduction-to-stemming/)
* [Xerox Stemmer](https://www.geeksforgeeks.org/introduction-to-stemming/)

**Parts-Of-Speech tagging**

Labeling each word in a sentence with its corresponding part of speech, such as noun, verb, adjective

<https://www.geeksforgeeks.org/nlp-part-of-speech-tagged-word-corpus/>

**Stop-watch removal**

Removing common, unimportant words like a,is,the,an so on

**The Text- preprocessing also includes**

Convert the text to lowercase, remove punctuation, numbers, special characters, and HTML tags.

Handling Emojis and Emoticons - Converting emojis to their textual representation.

Spell Checking

**Freely available pre-trained models**

TensorFlow hub

FastText from facebook

GPT3

**Python Open-source Libraries to work with NLP**

NLTk

Spacy

Gensim

Scikit

TensorFlow

pyTorch

Huggin Face

Regular expression in python is necessary for pattern matching - <https://regex101.com/> this website would be helpful

## Text Representation

Conversion of input text into numerical form.

Here also we can use neural network concepts

**Bag Of Words:**

Taking or counting repeating words in a text.

Used in text classification or sentiment analysis

**Term Frequency (TF) – Inverse Document Frequency (IDF)**

Finds importance of a word in a document

**Word Embedding**

Converting text to numerical representations, which are close to the text,most possibility to related text

Word Embedding or Word Vector is a numeric vector input that represents a word in a lower-dimensional space. It allows words with similar meanings to have a similar representation.

## Semantic Analysis:

The process of classifying whether a block of text is positive, negative, or neutral

For this analysis we can train a model using machine learning,neural networks

## Named Entity Recognition

It is a technique to identify important things mentioned in a text where it is name of a person,organization or normal verbs.

**Rule based NER**

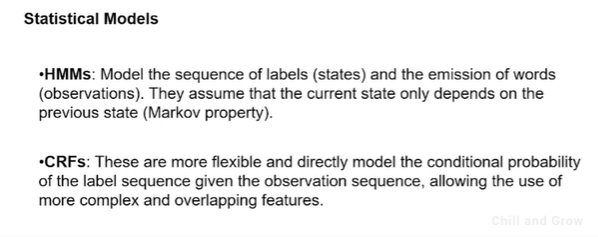
We define rules to identity

Orthografic features – if captilazise word in between the sentences can be name or location

Lexical – new follwing york is a location

Syntactic feature –Parts of speech tag

Dictionary based feature -



# VECTOR EMBEDDING

# RNN (Recurrent Neural Networks)

Text Represented can be improved by using RNN .Rnn used in auto complete,language translation.

Recurrent Neural Network (RNN) is a type of Neural Network where the **output from the previous step is fed as input to the current step.**

The main and most important feature of RNN is its **Hidden state**, which remembers some information about a sequence.

It will add some weightage and predict data

Loss function is calculated for metrix

Gradient – to improve network

Input data – encoder

Output data – decoder in context of deep learning.

# LSTM (Long-Short Term Memory)

LSTM networks are an extension of recurrent neural networks ([RNNs](https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/)) mainly introduced to handle situations where RNNs fail.

LSTMs can retain information for a long time while forgetting irrelevant information.

It has gates to remember & forget memories – Forget gate & input gate

# Attention Mechanism

A computer method for prioritizing **specific information** in a given context is called the attention mechanism of [deep learning](https://www.geeksforgeeks.org/deep-learning-tutorial/). During translation or question-answering activities, attention is used in [natural language processing](https://www.geeksforgeeks.org/natural-language-processing-nlp-tutorial/) to align pertinent portions of the source phrase.

# Transformers

Transformer Architecture is a model that uses self-attention that transforms **one whole sentence into a single sentence.**

This is a big shift from how older models work step by step, and it helps overcome the challenges seen in models like RNNs and LSTMs.

# BERT

<https://jalammar.github.io/illustrated-bert/>

BERT is a powerful NLP model that **revolutionized how language models understand and process text**.

It uses a **bidirectional transformer architecture**, allowing it to understand the **full context of a word** in a sentence.

BERT has been pre-trained on vast amounts of data and can be fine-tuned for specific tasks such as question answering, text classification, and more.

Its ability to generate **contextual word embeddings** has set new benchmarks for many NLP tasks and made it a cornerstone of modern language understanding applications.

# LLM (Large-Language Models)

It is a neural network with trillions of parameters.

Other than statistical model, it reads reinforecement learning with human feedback.

ChatGpt uses Gpt llm

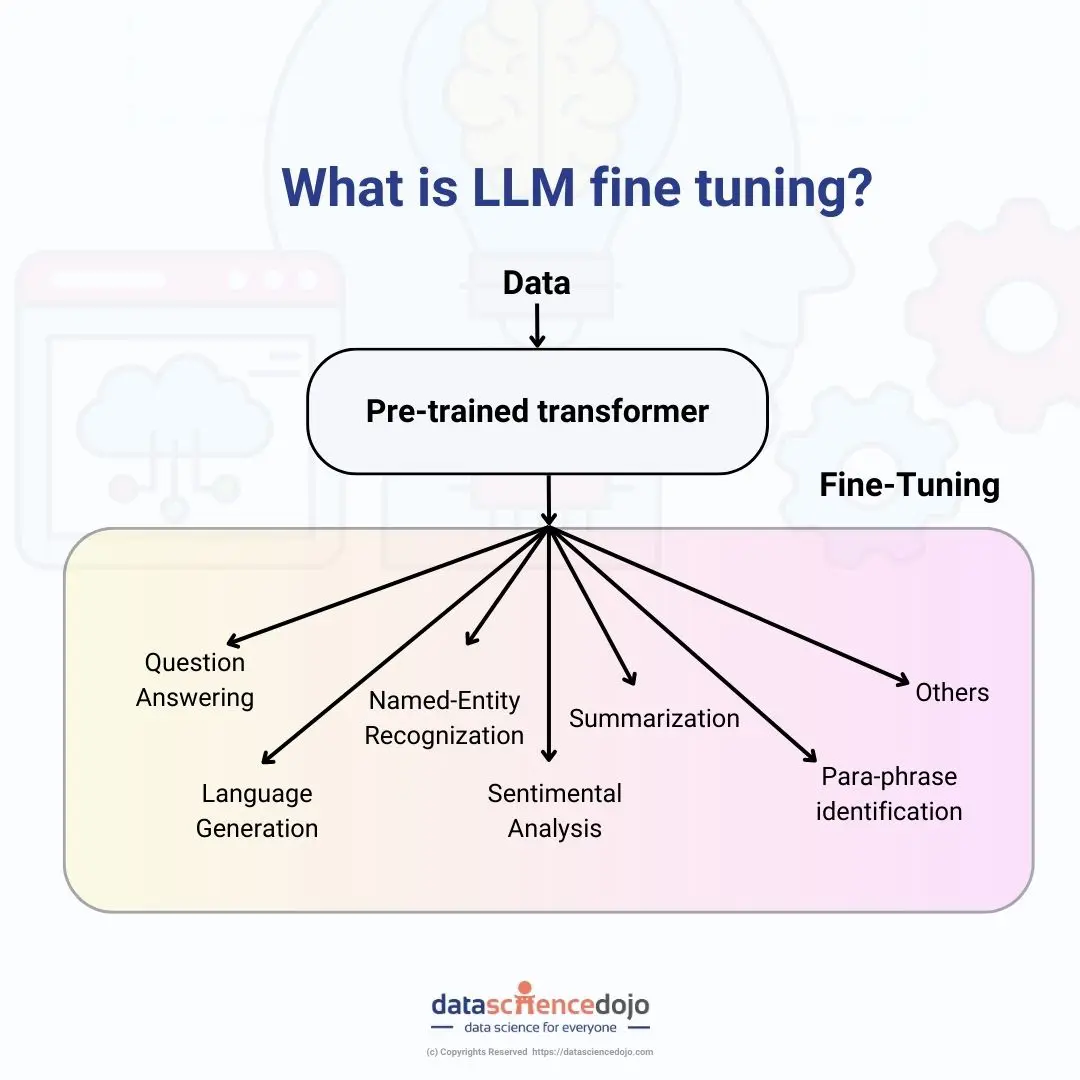
Google uses Palm2

Meta uses Llama

LLM works on data being trained on.

# Fine-Tuning

Taking existing LLM and training it with specific data to improve specific features.



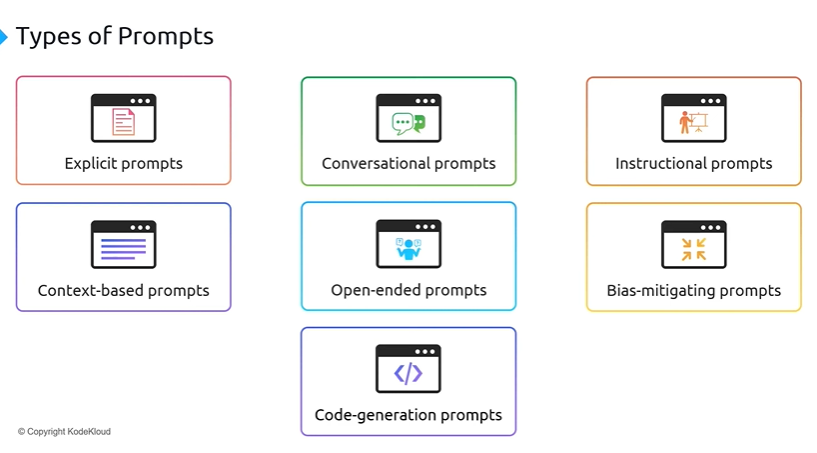
# Prompting Techniques

<https://www.geeksforgeeks.org/what-is-prompt-engineering-the-ai-revolution/>

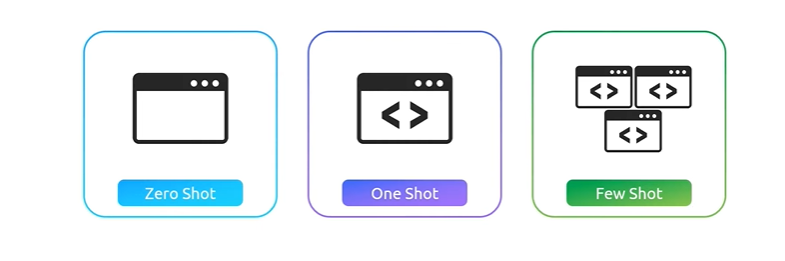
**Prompt engineering** is the process of creating effective prompts that enable AI models to generate responses based on given inputs

**Prompts** are short pieces of text that are used to provide context and guidance to machine learning models.

**Types Of Prompts**



**Prompting Techniques**



**1.Zero Shot Prompting**

Direct questioning to the model without any detailed derivations. It simply retrieves data from trained dataset

Eg: Write a poem about peace

Translate this sentence from english to hindi

**2.One Shot Prompting**

Providing a single example for the question or giving hint to the model.

Giving example, it will learn small amount of data from the input

Eg: Give a step-by-step process to prepare pizza

Give a symptoms and treatment for malaria

**3. Few Shot Prompting**

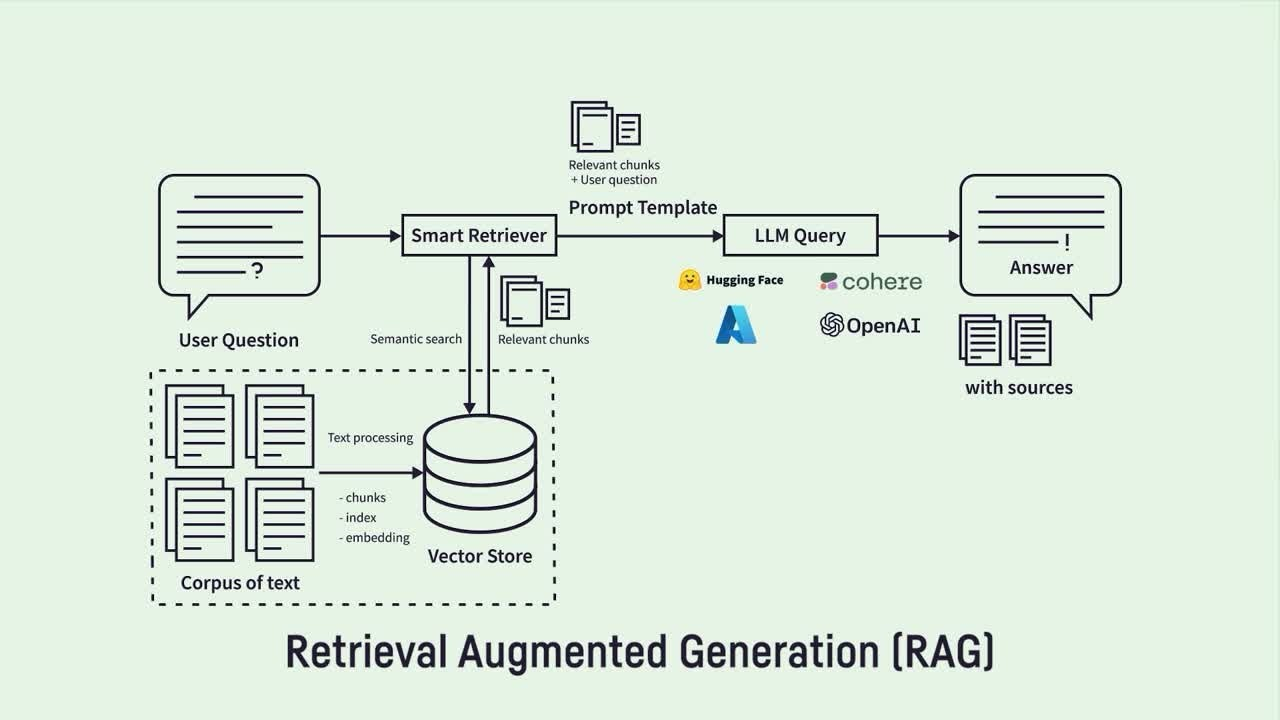
Providing a model with several examples to the model

# Vector DB – RAG (Retrieval-Augmented Generation)

A vector database is a collection of data stored as mathematical representations.

Vector databases use **Semantic or similarity Search** that allow applications to connect **relevant items** together. Vectors that are clustered together are similar and likely relevant to each other. This can help users search for relevant information.

**RAG (Retrieval Augmented Generation)**



We use RAG when our model has no source or is out of date.   
When user asks question, it will create query to retrieve vector from the vector DB, which already have all the information in the form of vector,using similar search and give it to the llm as a response

* **Vector Databases**: There are several **open-source vector DBs** like **FAISS**, **Milvus**, and **Annoy** that anyone can use to store and retrieve vector embeddings.
* **LLM Access**: LLMs cannot automatically query or retrieve data from a vector DB unless they are integrated into an RAG **system** that manages the querying process. You need to write the logic to connect the LLM to the vector DB via a retrieval pipeline.
* **Custom Setup**: Developers must set up the integration, allowing the LLM to issue queries, retrieve relevant data from the vector DB, and use that data to generate a response.

# Langchain and Lama index

**LangChain** and **LlamaIndex** are both powerful frameworks designed to help developers build applications that combine **large language models (LLMs)** with external data sources, like databases, APIs, or documents. They provide abstractions and tools that make it easier to integrate LLMs with various retrieval and reasoning systems.

NLP

**Data Preprocessing**

(tokenization,Stemming,Lemmatization)

**Text Representation** (Represent Vector space)

(BOW, IF-IDF, word embedding like wordToVec,GloVe)

**Model Selection**

(1-Traditional –ML algorithm

2- deep learning – RNN, LSTM

3- Pre-Trained models –LLM, BERT

To get the expected results)

**Pipeline of NLP**

