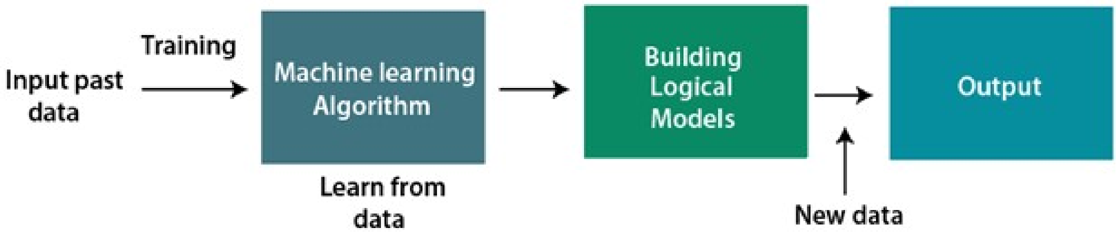
1. What does one mean by the term "machine learning"?

Ans: Machine Learning is a technology which enables computers to learn automatically from past data. **Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information.**

A Machine Learning system **learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it**. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. The below block diagram explains the working of Machine Learning algorithm:



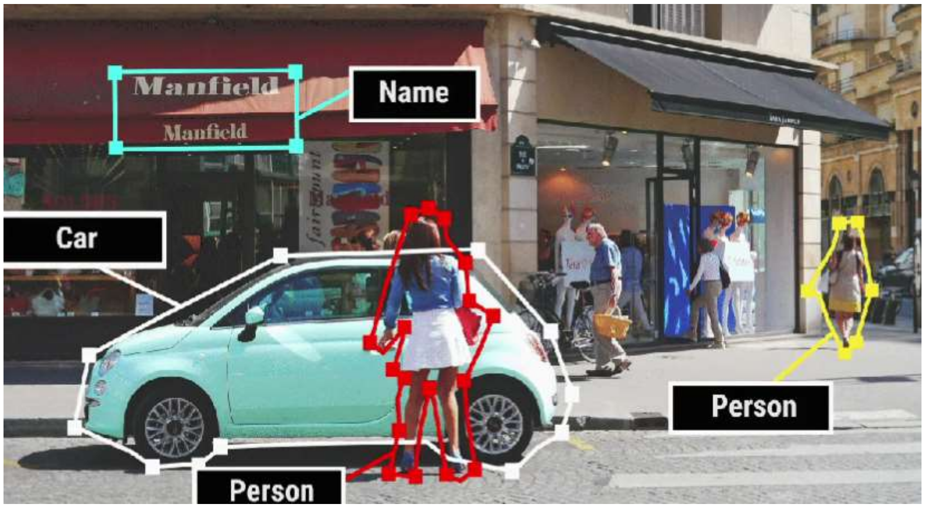
2. Can you think of 4 distinct types of issues where it shines?

Ans:

1. Supervised Learning
2. Unsupervised Learning
3. Semi- Supervised Learning
4. Reinforcement Learning

3. What is a labeled training set, and how does it work?

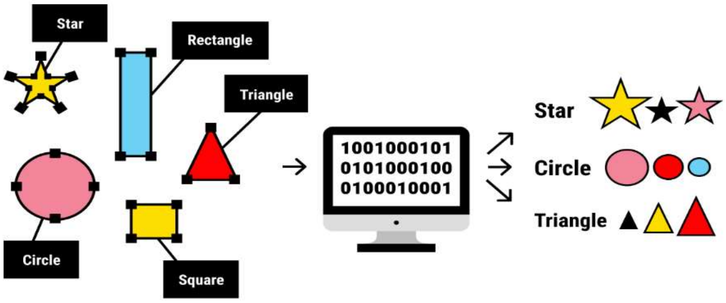
Ans: As the name suggests, labeled data (aka annotated data) is when you put **meaningful labels**, add tags, or assign classes to the raw data that you've collected. An image recognition system and have already collected several thousand photographs. Labels would be telling the AI that the photos contain a ‘person’, a ‘tree’, a ‘car’, and so on as shown below.



Labeled data makes the training process much more efficient and simpler. The idea behind labeling data is to teach the ML Model to recognize patterns according to the task or target. This way, after the training process, the input of new unlabeled data will lead to predictable labels.

To put it simply, this means that you add labels to data and set a target, and the ML Model **learns by example**. The process of assigning the target labels is what we know as annotation. After the training period ends, your machine will be able to identify the presence of a 'person', a 'car', or a 'tree' in the new photos. Not only that but the ML Model trained on labeled data can be used for complex forecasting (e.g., predicting the prices on the stock market or suggesting additional products for the customer.

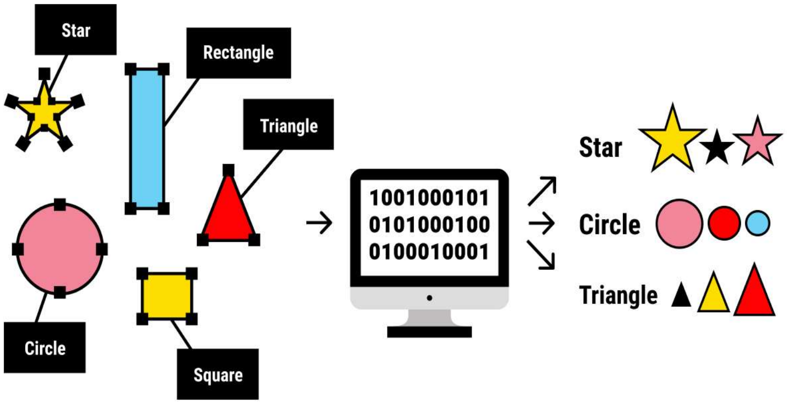
Below is an example of labeled data.



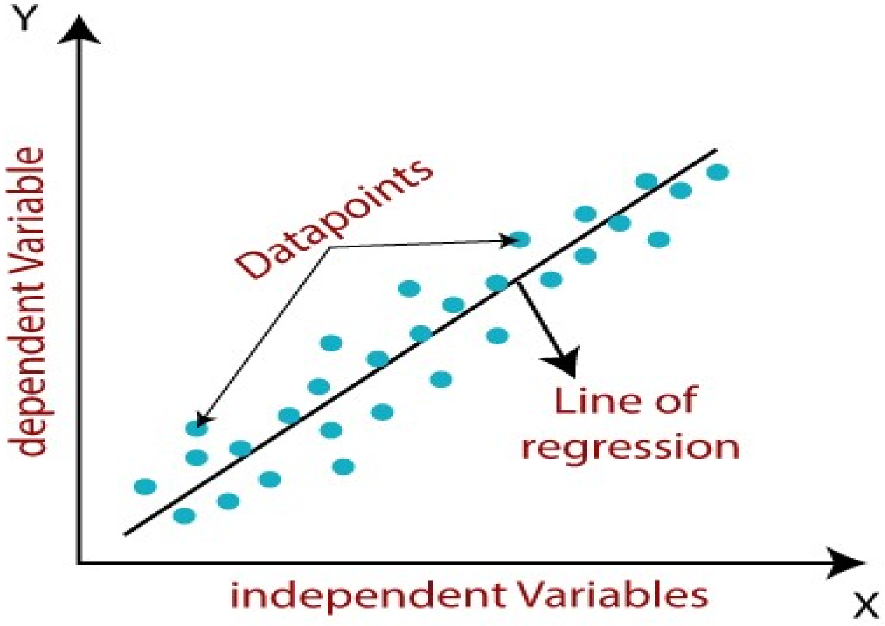
4. **What are the two most important tasks that are supervised?**

Ans:

1. **Classification**



2. **Regression**



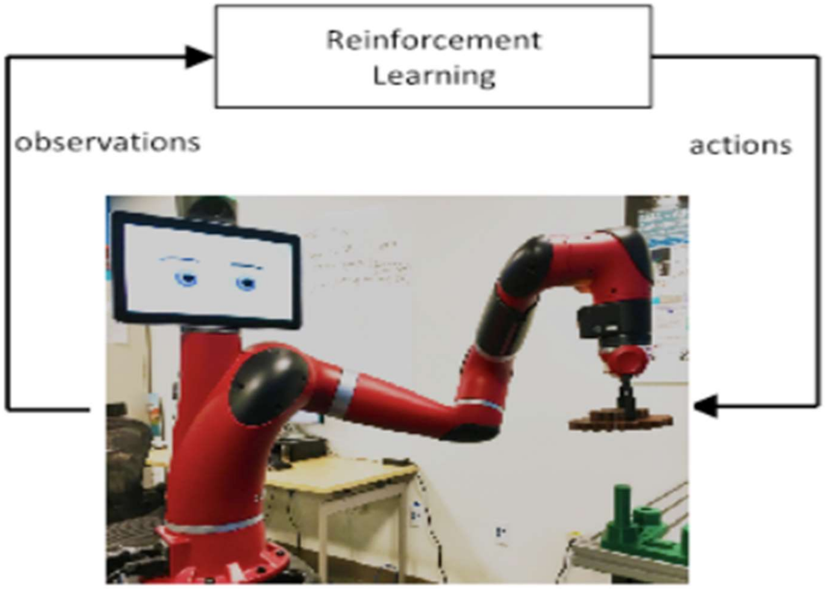
5. **Can you think of four examples of unsupervised tasks?**

Ans:

1. **Customer segmentation**, or understanding different customer groups around which to build marketing or other business strategies.
2. **Genetics**, for example clustering DNA patterns to analyze evolutionary biology.
3. **Recommender systems**, which involve grouping together users with similar viewing patterns in order to recommend similar content.
4. **Anomaly detection**, including fraud detection or detecting defective mechanical parts (i.e., predictive maintenance).

6. State the machine learning model that would be best to make a robot walk through various unfamiliar terrains?

Ans: The best Machine Learning algorithm to allow a robot to walk in unknown terrain is Reinforced Learning, where the robot can learn from response of the terrain to optimize itself.



7. Which algorithm will you use to divide your customers into different groups?

Ans: We will use following Clustering Algorithms:

1. K-Means
2. DBSCAN

3. Hierarchical Clustering
4. Mini-Batch K-means
5. Gaussian Mixture Model

Out of all the above K-Means gives best results for customer segmentation.

8. Will you consider the problem of spam detection to be a supervised or unsupervised learning problem?

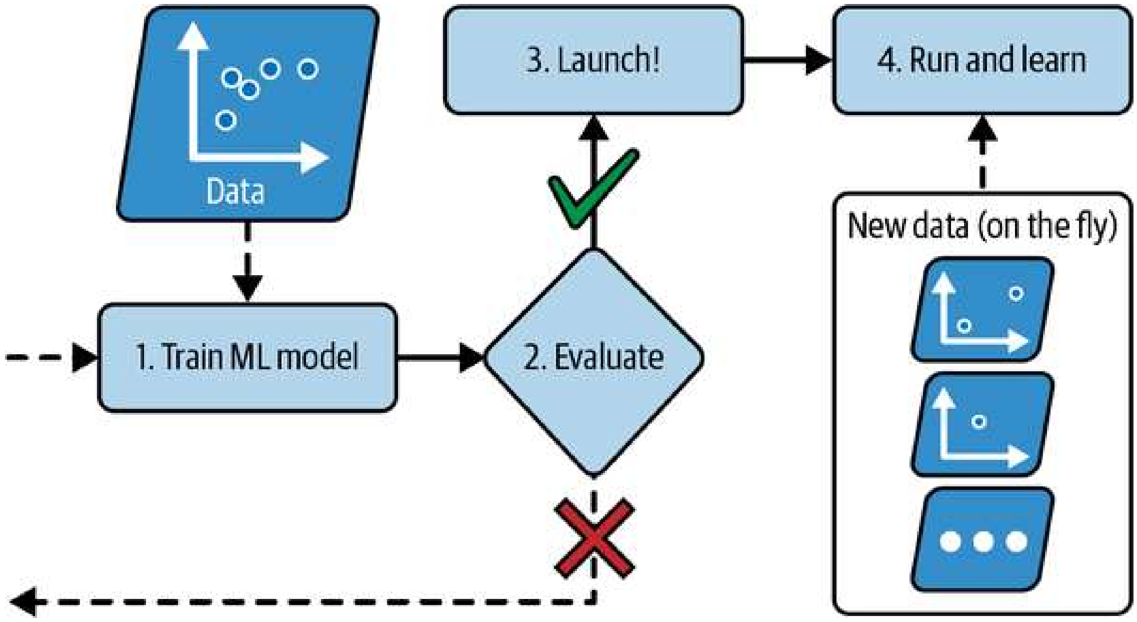
Ans: The Spam Detection will come under supervised learning since in spam detection & filtering, classification algorithms are used. These algorithms classify an email as spam or not spam. The spam emails are sent to the spam folder.

9. What is the concept of an online learning system?

Ans: In online learning, we train the system incrementally by feeding it data instances sequentially, either individually or in small groups called mini-batches. Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives.

Online learning is useful for systems that need to adapt to change extremely rapidly (e.g., to detect new patterns in the stock market). It is also a good option if you have limited computing resources; for example, if the model is trained on a mobile device.

Below figure shows the complete working of online learning Model.



10.What is out-of-core learning, and how does it differ from core learning?

Ans: Out-of-core learning system is a system that can handle data that cannot fit into your computer memory. It uses online learning system to feed data in small bits.

11.What kind of learning algorithm makes predictions using a similarity measure?

Ans: Learning algorithm that relies on a similarity measure to make predictions is instance-based algorithm

12.What's the difference between a model parameter and a hyperparameter in a learning algorithm?

Ans: Model parameter determines how a model will predict given a new instance; model usually has more than one parameter (i.e. slope of a linear model). Hyperparameter is a parameter for the learning algorithm, not of a model.

13.What are the criteria that model-based learning algorithms look for? What is the most popular method they use to achieve success? What method do they use to make predictions?

Ans: Model based learning algorithm search for the optimal value of parameters in a model that will give the best results for the new instances. We often use a cost function or similar to determine what the parameter value has to be in order to minimize the function. The model makes prediction by using the value of the new instance and the parameters in its function.

14.Can you name four of the most important Machine Learning challenges?

Ans:

1. **Poor Quality of Data**
   Data plays a significant role in the machine learning process. One of the Significant issues that machine learning professionals face is the absence of good quality data. Unclean and noisy data can make the whole process extremely exhausting. We don't want our algorithm to make inaccurate or faulty predictions. Hence the quality of data is essential to enhance the output. Therefore, we need to ensure that the process of data pre-processing which includes removing outliers, filtering missing values, and removing unwanted features, is done with the utmost level of perfection.
2. **Underfitting of Training Data**
   This process occurs when data is unable to establish an accurate relationship between input and output variables. It simply means trying to fit in undersized jeans. It signifies the data is too simple to establish a precise relationship. To overcome this issue:

1. Maximize the training time
2. Enhance the complexity of the model
3. Add more features to the data
4. Reduce regular parameters
5. Increasing the training time of model

3. **Overfitting of Training Data**
   Overfitting refers to a machine learning model trained with a massive amount of data that negatively affect its performance. It is like trying to fit in Oversized jeans. Unfortunately, this is one of the significant issues faced by machine learning professionals. This means that the algorithm is trained with noisy and

biased data, which will affect its overall performance. Let's understand this with the help of an example. Let's consider a model trained to differentiate between a cat, a rabbit, a dog, and a tiger. The training data contains 1000 cats, 1000 dogs, 1000 tigers, and 4000 Rabbits. Then there is a considerable probability that it will identify the cat as a rabbit. In this example, we had a vast amount of data, but it was biased; hence the prediction was negatively affected.

We can tackle this issue by:

1. Analyzing the data with the utmost level of perfection
2. Use data augmentation technique
3. Remove outliers in the training set
4. Select a model with lesser features

4. Lack of Training Data

The most important task you need to do in the machine learning process is to train the data to achieve an accurate output. Less amount training data will produce inaccurate or too biased predictions. Let us understand this with the help of an example. Consider a machine learning algorithm similar to training a child. One day you decided to explain to a child how to distinguish between an apple and a watermelon. You will take an apple and a watermelon and show him the difference between both based on their color, shape, and taste. In this way, soon, he will attain perfection in differentiating between the two. But on the other hand, a machine-learning algorithm needs a lot of data to distinguish. For complex problems, it may even require millions of data to be trained. Therefore, we need to ensure that Machine learning algorithms are trained with sufficient amounts of data.

15.What happens if the model performs well on the training data but fails to generalize the results to new situations? Can you think of three different options?

Ans: If the model performs poorly to new instances, then it has overfit on the training data.

To solve this, we can do any of the following three:

1. Get more data.
2. Implement a simpler model.
3. Eliminate outliers or noise from the existing data set.

16.What exactly is a test set, and why would you need one?

Ans: Test set is a set that you test your model (fit using training data) to see how it performs. Test set is necessary so that you can determine how good (or bad) your model performs.

17.What is a validation set's purpose?

Ans: Validation set is a set used to compare between different training models.

18.What precisely is the train-dev kit, when will you need it, how do you put it to use?

Ans:

The goal of **dev-set** is to rank the models in term of their accuracy and helps us decide which model to proceed further with. Using Dev set we rank all our models in terms of their accuracy and pick the best performing model. i.e. dev set ranks models similar to a search engine like google rank pages and then pick the top model and hence act as a filter to remove bad models.

While having dev set split, first training algorithm makes the choice for optimal parameters and then those parameters are used on dev data to help us find best model architecture as compared to both choices made together by learning algorithm itself.

Dev set helps us in reducing the complexity of diagnosis if things won't go fine i.e. we'll be able to assign error to either choice of parameters or picking up model architecture very concretely.

Lack of dev set and using only the training set doesn't give you clue about which choice went wrong and luck rather than skill will be helpful to debug your learning algorithm there and to make decision further to improve the model accuracy.

19.What could go wrong if you use the test set to tune hyperparameters?

Ans:

If you tune hyperparameters using the test sets, then it may not perform well on the out-of-sample data because the model is tuned just for that specific set.