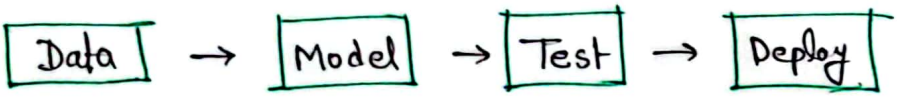
CampusX Machine Learning

Batch VS Online ML:

BATCH Learning (offline): A technique, when you train your model with whole dataset.

To train model with whole dataset, if the dataset is really big, it is difficult to run that model on the server. It will be costly and time consuming. We need to train this model in a personal pc When the training will be done, then it can be deployed in the server.



Disadvantage:

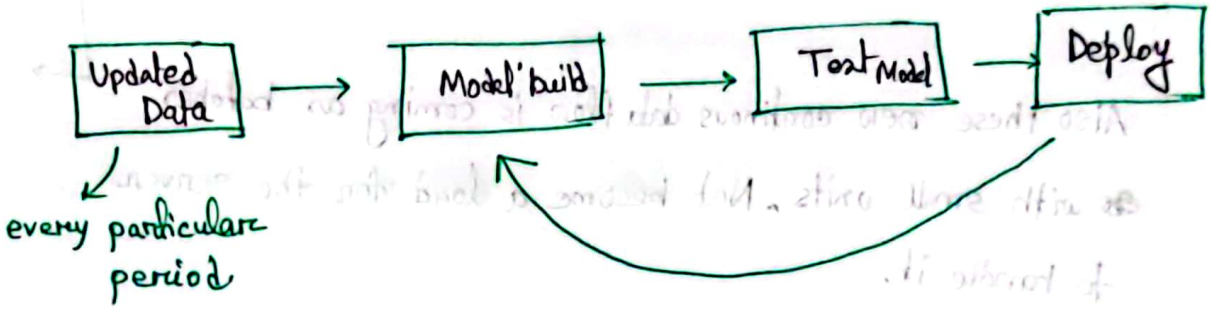
- Model becomes static. It can't be learned from new data Because after training in a offline pc, it is not getting trained in the server.

For example, Netflix recommendation system.

If you use Batch technique, you can't make a good recommendation system for Netflix because every week in the netflix database new movies are getting added. You have to train your model with

the new data sets continuosly in order to make your recommendation system's accuracy good.

So, the flowchart should look something like this →



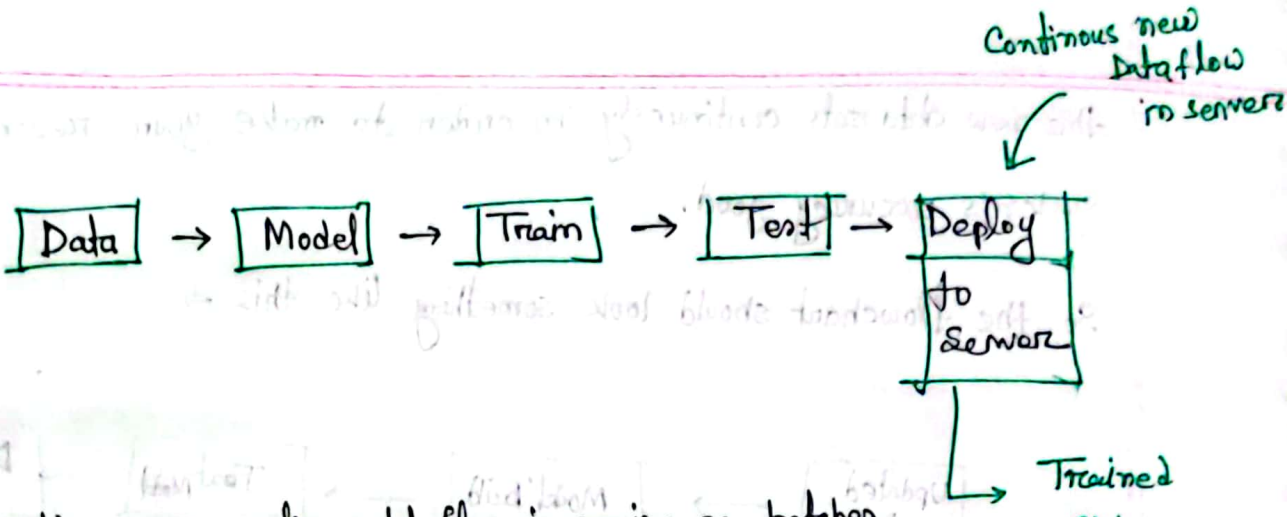
other disadvantages:

- limited Hardware
- Availability
  - Data is not available before a particular time event

This issuses can be solved by online Learning technique.

Online Machine Learning:

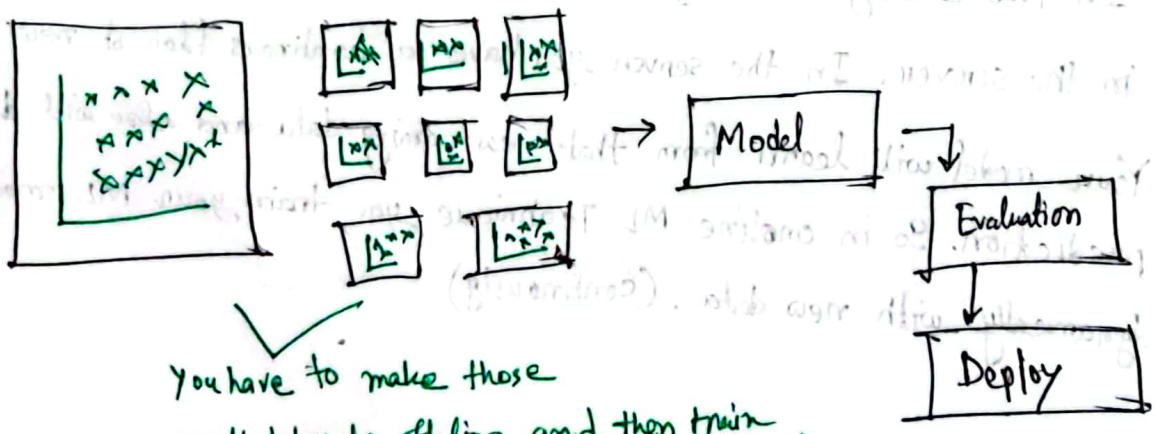
In this case, you train your model with a new dataset, test it and deploy in the server. In the server you have a continous flow of new data Your model will learn from that new coming data and also will do prediction. So in online ML Technique you train your ML model dynamically with new data. (continously)



Also these new continous data flow is coming as batches with small units. Not become a load for the server to handle it.

Learning Rate in Online ML: How fast or slow your model should behave while learning from new datasets coming as small Batches that should be set which is a difficult task to do.

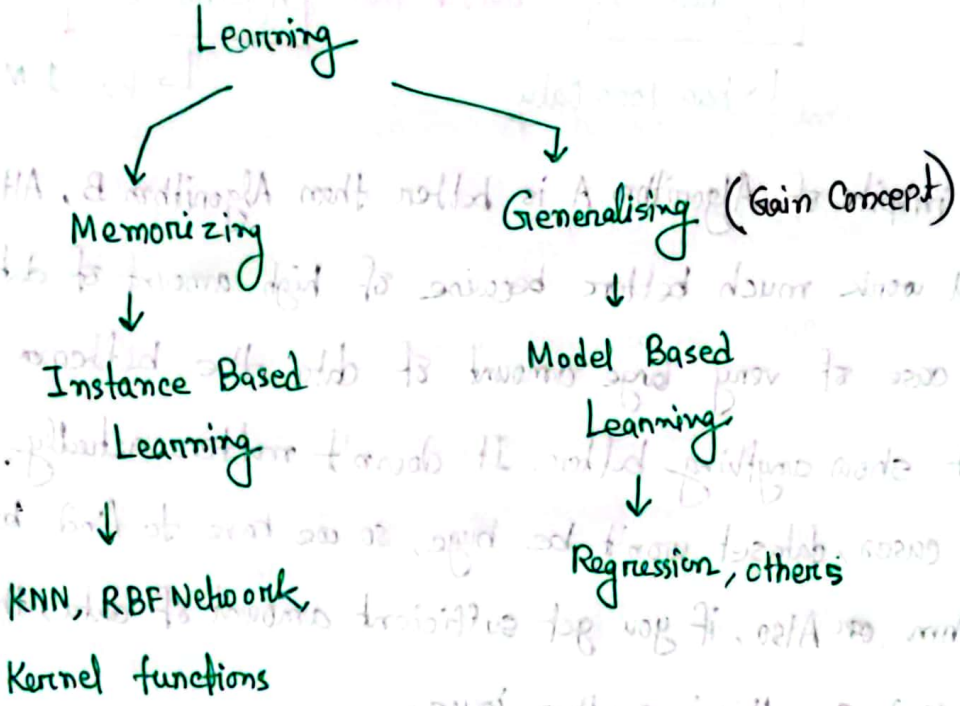
Out of core learning: If the data is so big that you can't load the dataset in your offline batch technique, then what you can do is, you can use the online ML technique by providing small batches of data periodically to train the whole large dataset.



Disadvantage:

1) Tricky to use
2) Risky → (You have to always provide security for your incoming data)
   Like anomaly detection

Instance Based Learning Vs Model Based Learning:



Problems and Challanges of Machine Learning:

1. **Data Collection:** If you don't have a ready made data, you have face challanges to find and use data. You have to get the data through some API or scrapping the web

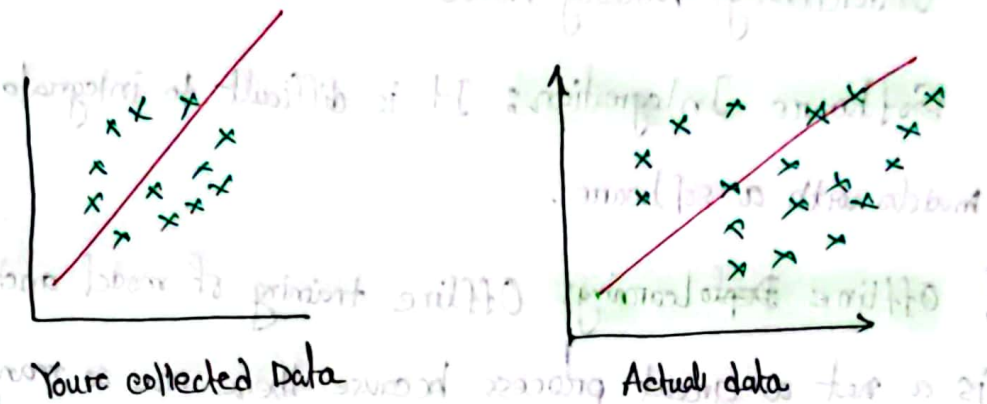
2. **Insufficient/Labelled Data:**

Algorithm A better than Algorithm B

↳ has 1000 data ↳ has 1 Million data

Here inspite of Algorithm A is better than Algorithm B, Although Algo B will work much better because of high amount of data.
So in case of very large amount of data the bettering algorithm doesn't show anything better. It doesn't matter actually. But most of the cases, dataset won't be huge, so we have to find better suitable algorithm. Also, if you get sufficient amount of data, they may not be labelled. So, this is another issue.

[3] **Non Representative Data:** Suppose you are gathering data, you gathered the data but it was not a proper representative of the population



That can be a problem when the data is not collected properly, can't become a representative of the whole population.

[4] **Poore quality Data:** Most of the real dataset would be of poore quality. They can be mixed in columns, they can be none, nan, in different format, in different types, messy etc. We have clean and transform ours data according to ours ml model need.

[5] Garbage in → ML Model → Garbage Out

[5] **Irrelelvant Features:** Sometimes, there are more than enough features, features that we don't need, features which are in multiple colums but need to be together in 1 column, same columns from like me model needs only one. So these are the problems. So we have to do FE here.

[6] **Overfitting:** Already Noted

[7] **Underfitting:** Already Noted

[8] **Software Integration:** It is difficult to integrate ML models with a software.

[9] **Offline ~~Deplo~~ Learning:** Offline training of model and deployment is a not a great process because there are so many issues related to it (already discussed)

[10] **Cost involved:** To get Online ML ~~to~~ environments, deployment, using servers and so on we have to buy cloud infrastructure which is very costly.