**Introduction To Data Science STUDY NOTE**

In this session we'll talk about

1. **Data To Insight**
2. **What is Machine Learning?**
3. **How Machine Learning Works?**
4. **What is Big Data ?**

**What is Data Science?**

Data Science is nothing but and technique by which we can actually achieve Data Driven Decision. So, whenever we take any decision instead of our knowledge and if we take the decision based on the data that is actually Data Driven Decision and to actually achieve this thing the Science is Data Science.

* **Data Driven Decision(DDD) Making**:

All the organization collect data that is internal data(internal to the organization) or external data,after collecting the data they store it in the big data Hadoop system or maybe in a SQL database or maybe no SQL database or maybe it is a streaming data and they can store it in a streaming database if is a time series database they can store it in a time series database so there is a different kind of option. But high level tries to understand in any organization to collect all those data and store the data depending on how the data type is and then they process the data in this time you need a requirement that you have to process in real time. But sometimes you have a requirement that you process it batch to overnight.



After the collecting, storing and processing you have to analyze. Now,these are the two cases, we use any machine learning algorithms, maybe some time use Forecasting, some time use Predictive Modeling or Optimization, Simulation or Artificial Intelligence or whatever ultimately we come off with a system.

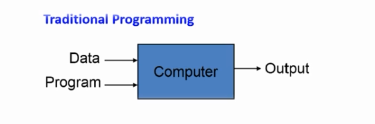
The system may be Recommendations System would provide an Action Items from your prescription that you can simply do this Contextual Intelligence that since the price of the raw materials is high you should not buy it right now. So everything and anything is associated with Action Item and from the action, you can achieve two thing-Cost Reduction & Profit Maximization that is actually your objective.

**What is Machine Learning?**

Now you create a rule and algorithm and you pass a new data. Suppose you have written if else statement and you are trying to create a system which can actually detect a spam. So, in a mail, you are looking for each of those mail and you are trying to figure out whether this is spam or not. If it is spam then you put into a spam folder.

* **Traditional Programming:**

Now, if you would try in a Traditional Programming way, then you would create different rules so you would write if else statement, while statement etc. and all those things. A new data would come which is a new mail and based on your program it would give you an output. Whether this is spam or not but this base is actually a Rule Based approach. Where your rule accepts or rejects based on the Rule Based but it would not change the situation.



Suppose you have created a rule that each mail contains more than ten images then it is a spam but unfortunately, you receive a good mail from your friend and which is contain more than 10 images but it is not actually a spam. Here your programme marked this as a spam but once you say no that is not a spam then it would not learn again. But if you see in our Gmail what’s happened you try to train your model, you say no this is not a spam but next time whenever you received from your friend more than 10 images it would not mark it spam why because Gmail spam system is based on ML Based is not rule based.

* **Machine Programming**

What is different between Rule Based and ML Based?

ML Based actually adopt the new situation and it improves with new each data and on the other hand Rule Based accept or reject based on rule doesn’t change the situation.

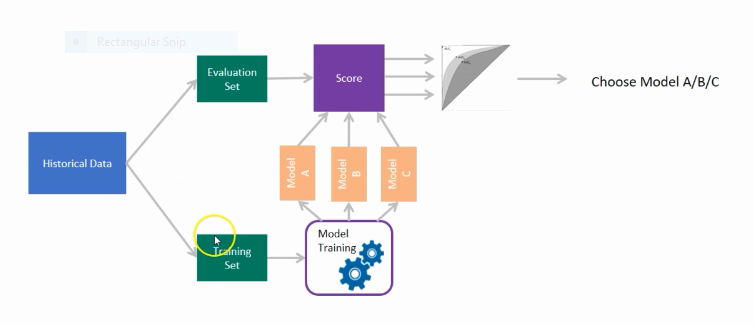


we have a historical data, historical data is nothing just historical email data that is last or previous 6 month to 1 year ago message. whatever the mail came that is historical data and it has an output also. So whenever you marked that it is not a spam or said it is not a spam, then those are placed in your outbox.whether it can be spam or not.

Now, It will be trained in a Machine Learning Algorithm, now put some algorithm, SVM algorithm or other, a lots of algorithms are there.So, now just pass on some data, pass on the output and my machine will learn itself and it will create an algorithm. So, Algorithm is nothing just a rule to be followed in calculations or other problem-solving operations, for a computer. So it would now say that if there is normally more than 10 images then it would be spam but if the images more than ten and comes from your friend then it is not a spam. Since you have passed the historical data and output to your machine and your system can learn using the machine learning algorithm. So, next time when it comes from your friend it owns marked it a spam.

**What is Machine Learning? How it works?**

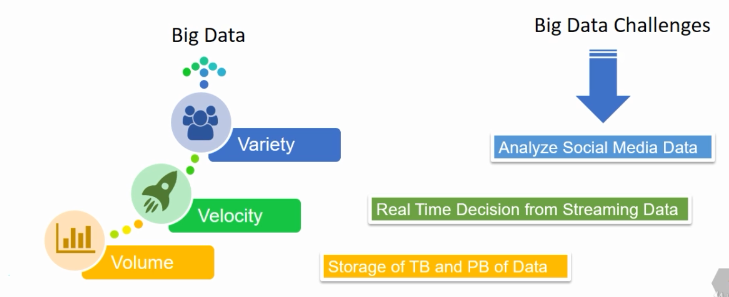
Here we don’t create any algorithms rather we pass an algorithm and the system learn itself and now your last 6 month all the emails are in historical data which is divided into 2 trains (Training Set) and Test(Evaluation Set) then your five month data is training set and one month data is in evaluation set and then we pass different different algorithms like Linear Regression, Logistic Regression, and SVM and then we evaluate the model and then best on output we choose which is the highest accuracy. That’s why machine learning work.



Think tomorrow you have a new data and it also becomes in the history of the data and the same way it comes in the training data set so your model change the data set every day and every second and then the best on that it would predict. The model is tested on an unseen data (Evaluation Set) and the model score is calculated for each of the algorithms. Thereafter best algorithm is chosen.

**What is Big Data and the Connection between Big Data and Data Science**-

Big data is more familiar to 4Vs but here we discussed 3Vs



**Volume**:  
If your data volume is so large you cannot store it in a simple sequel database. That’s would be a big data challenge then you have to store the Terabytes and PetaBytes data. The data is unknown value, such as Twitter data feeds, click streams on a webpage or a mobile app or sensor enabled equipment.

**Velocity**:  
The data is coming with so much Velocity so may be from a GPS or from a twitter or from a facebook every time in a nanosecond there are so many tweets and so many facebook post and you have to store all of them, that against a big data challenge that’s why you have to use the Hadoop system. Velocity deals with the speed at which data flows in from the sources of business processes, application logs, networks, social media sites, sensors, mobile devices etc.

**Variety**:  
Similarly, Variety refers to the many types of data which is different in nature. Earlier all those data are stored in very structured format but today datas are unstructured, not structured at all. Because all those videos, images are unstructured data. This refers to the inconsistency which can be shown by the data at times.Now,since the Big Data are very different in nature to actually understand to get the insight out of this data is also very difficult that’s why we need to have these Decision Science and we need Machine Learning Algorithm to learn itself because it not easy that manually you would actually extract the insight from the data that’s why we need all these Machine Learning Algorithm to learn itself.

This brings an end to this post, I encourage you to re read the post to understand it completely if you haven’t and THANK YOU.