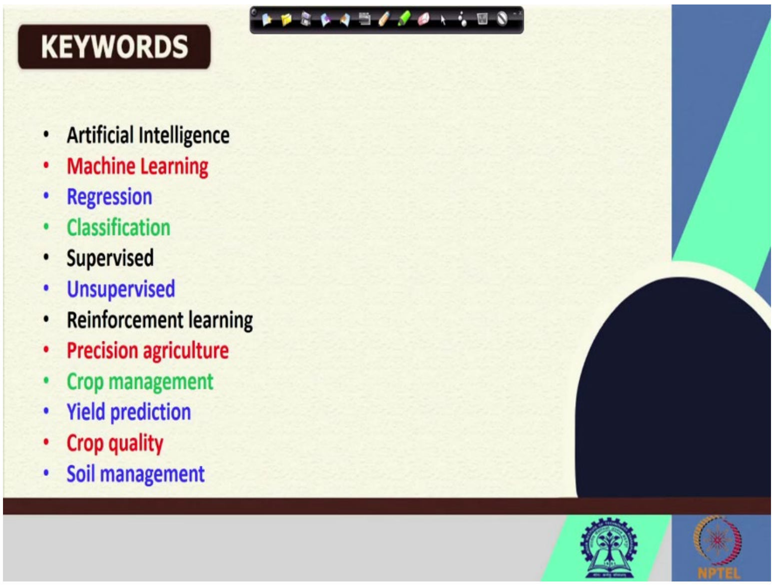
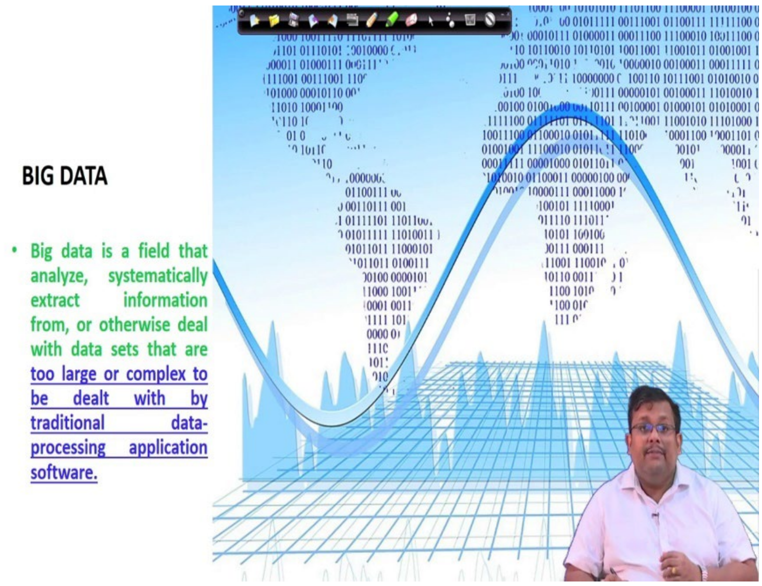
Machine Learning for Soil and Crop Management
Professor Sumsubhra Chakraborty
Department of Agriculture and Food Engineering
Indian Institute of Technology, Kharagpur
Lecture 02
General Overview of ML and DL Applications in Agriculture (Continued)

Welcome friends to this second lecture of week 1 of NPTEL online certification course of Machine Learning for Soil and Crop Management. In this week, we are discussing about general overview of Machine Learning and deep learning applications in agriculture.

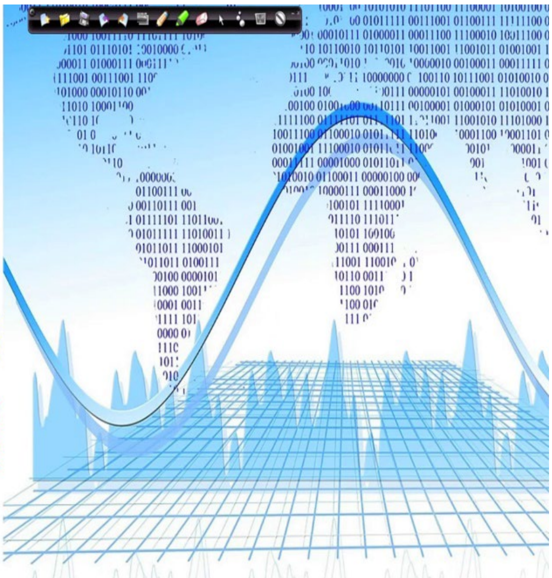
(Refer Slide Time: 00:40)





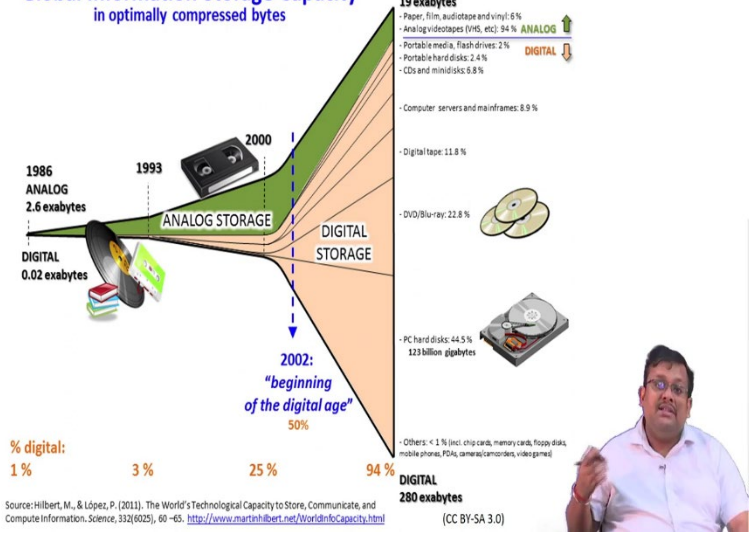
BIG DATA

- Big Data: large volume of data, which is produced by digital technologies
- Needs large storage capabilities in addition to editing, analyzing, and interpreting
- Interpretation of big data: considerable potential to add value for society, environment, and decision-makers



Global Information Storage Capacity

in optimally compressed bytes

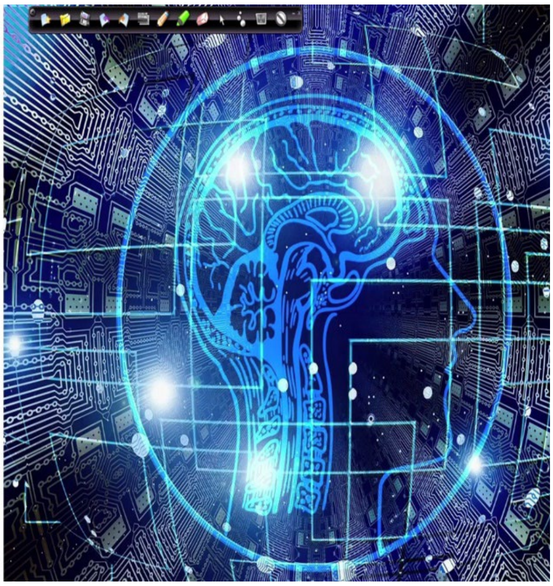


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

CHALLENGES OF BIG DATA

- Big data encompass challenges on account of their so-called "5-V" requirements
  - ❖ Volume
  - ❖ Variety
  - ❖ Velocity
  - ❖ Veracity
  - ❖ Value
- The conventional data processing techniques are incapable of meeting: emergence of ML !!

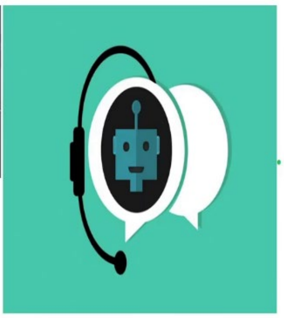
ARTIFICIAL INTELLIGENCE



AI

- The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

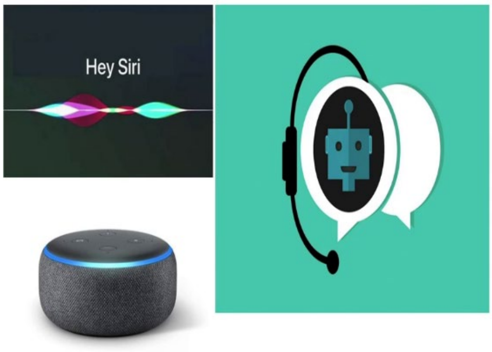
- Google (Web search)
- Recommendation system (YouTube, Amazon, Netflix)
- Human speech understanding (Siri or Alexa)
- Self-driving car (Tesla)



AI

- AI systems work by ingesting large amounts of labeled training data, analyzing the data for correlations and patterns, and using these patterns to make predictions about future states
- In this way, a chatbot that is fed examples of text chats can learn to produce lifelike exchanges with people, or an image recognition tool can learn to identify and describe objects in images by reviewing millions of examples

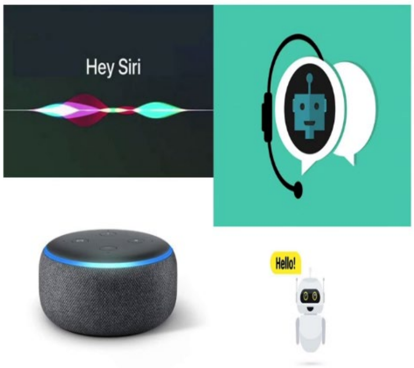
AI



Three cognitive skills:

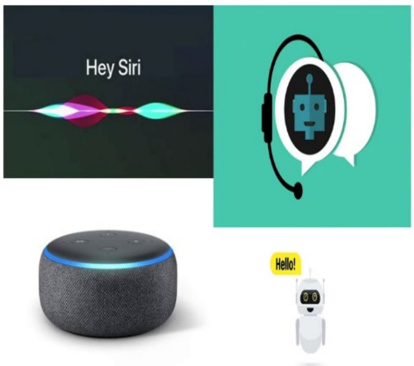
- Learning
- Reasoning
- Self-correction

AI: Advantages



1. Beneficial for detail-oriented jobs
2. Reduced time for data-heavy jobs
3. Offers consistent results
4. AI-powered virtual agents: efficient and always available

AI: Disadvantages



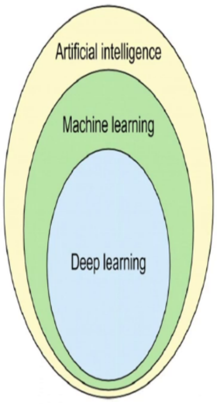
1. Costly
2. Requires deep technical knowledge
3. Limited experts to build AI tools
4. Lacking generalization from one task to another

MACHINE LEARNING

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy

MACHINE LEARNING

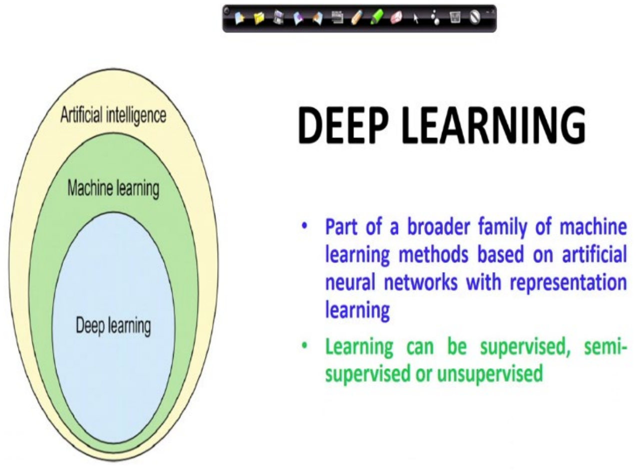
- The term machine learning was coined in 1959 by Arthur Samuel, an American IBMer and pioneer in the field of computer gaming and artificial intelligence.
- Objectives:
    - To classify data based on models which have been developed
    - To make predictions for future outcomes based on these models

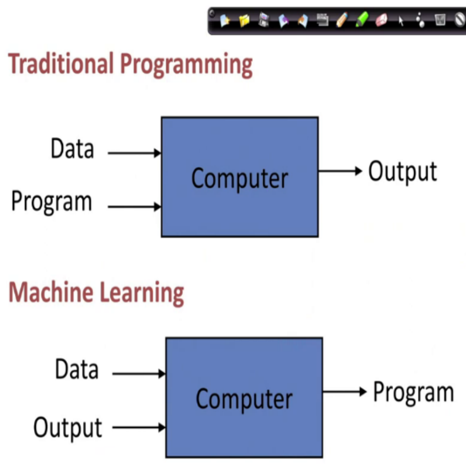


ML vs. AI

- ML learns and predicts based on passive observations, whereas AI implies an agent interacting with the environment to learn and take actions that maximize its chance of successfully achieving its goals

Judea Pearl in The Book of Why



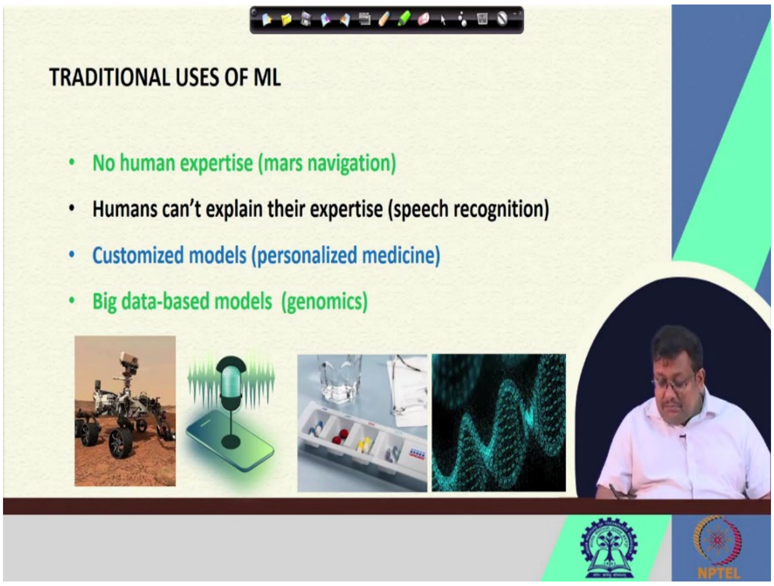


And in our first lectures, we have already discussed some important concepts like big data, we have also come, discuss the features of the big data, what are the changes, global changes in information storage capacities, temporal changes in global information storage capacities, we have also discussed the challenges of the big data.

We have discussed what is Artificial Intelligence also the application of Artificial Intelligence and then we have discussed about the cognitive skills of Artificial Intelligence, there are advantages and disadvantages we have discussed and also we have discussed then Machine Learning and who coined the term Machine Learning their objectives, the difference and relationship between Machine Learning, deep learning and Artificial Intelligence.

And also we have discussed about deep learning and the difference between traditional programming and Machine Learning.

(Refer Slide Time: 01:45)

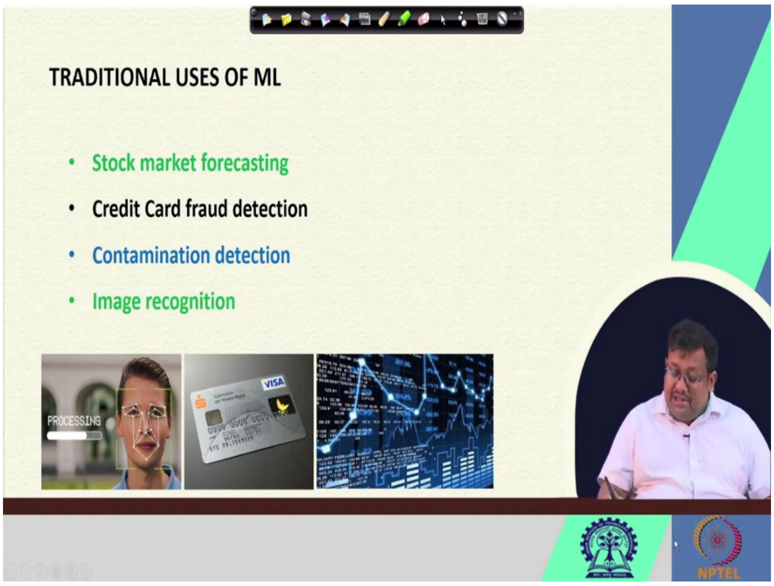


So, today we are going to discuss, today we are going to start discussing the traditional uses of Machine Learning. So, as I have already told you that, Machine Learning is a subset of Artificial Intelligence, this Machine Learning is more or less statistical learning and the traditional uses of Machine Learning we generally we can see when there is no human expertise for example, when there is a Mars Rover, which navigate in the Mars soil.

So, that uses the Machine Learning to identify the path. Then, also, we apply Machine Learning when human cannot explain their expertise for example, speech recognition is an important aspect where we apply the Machine Learning. When, when you want to do some customized, customization of personalized, personalized medicine, we rely on Machine Learning.

And also in case of genomics, I have already told you that it depends on big data models. So, handling that big data models is done by the Machine Learning approaches or rather deep learning approaches. So, these are the some of the traditional, traditional uses of Machine Learning.

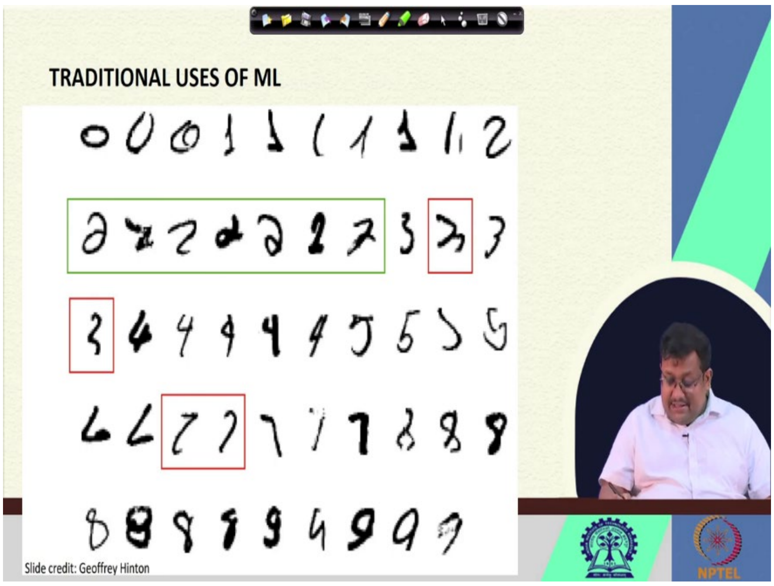
(Refer Slide Time: 03:23)



Apart from them, you can also see the Machine Learning application in stock market forecasting, in credit card fraud detection, in contamination detection, in different media and also image recognition. So, the application of Machine Learning is extensive and the application of, Machine Learning is continuously evolving in different different, sectors.

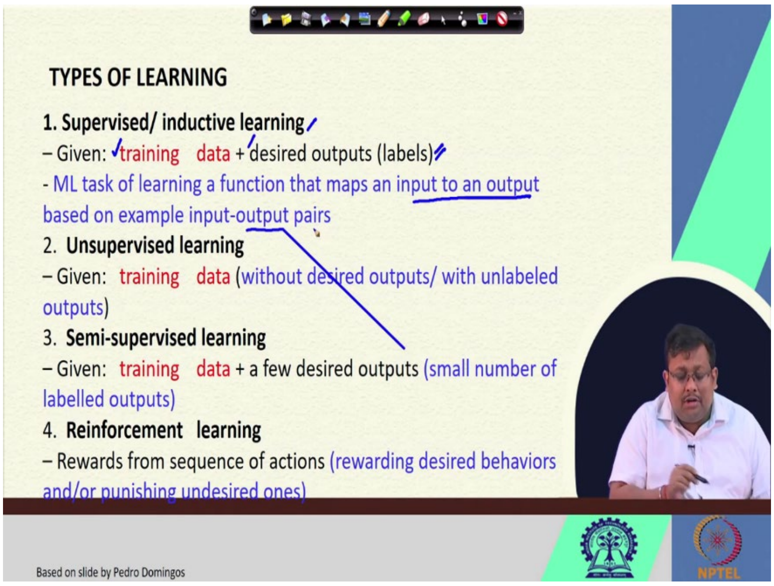
Some sectors where we traditionally depend on human force, nowadays, Machine Learning is taking their place helping to reduce the burden of different different activities. So, the, although the traditional uses of Machine Learning is huge, new new sectors are being utilized, in the, in the, in the, in the world of Machine Learning.

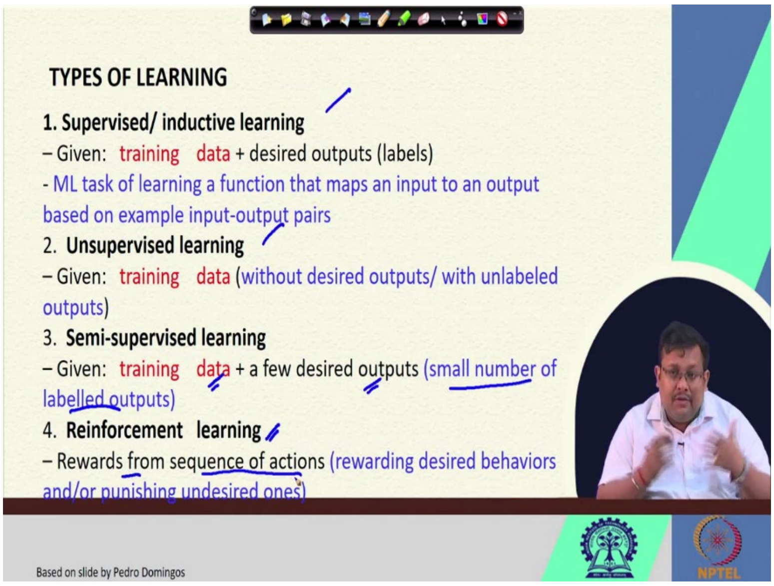
(Refer Slide Time: 04:29)



Some example, sometime identifying the handwritten digits, this is a classic example, where we have seen the identification of the handwritten digits. So, sometime human cannot do it perfectly. So, in that case, Machine Learning also helps. So, Machine Learning can helps in identifying these handwritten digits. So, for and reducing the human induced errors.

(Refer Slide Time: 05:04)





So, let us now see what are the tradition, what are the types of learning; Machine Learning or statistical learning. The first one is the first one is supervised or inductive learning. So, in the supervised learning, what we, what are the features? So, in the supervised learning, we give both training data that means input data and also the desired outputs are given with their proper labels.

So, both training data and desired outputs are given with labels. So, Machine Learning task of learning is a function that map's an input to an output based on example, input output pairs. So, this, in this input output pairs when both of them are labelled then or specifically the outputs are labelled then we call it supervised learning. Just opposite when we only incorporate the training data without the desired outputs or with unlabelled outputs then it is called the unsupervised learning.

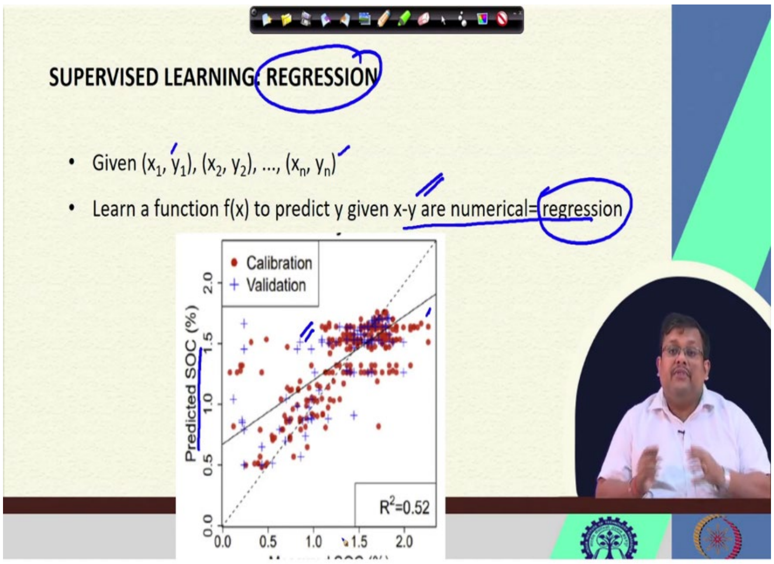
So, in case of unsupervised learning, the outputs are not labelled. The third category is semi-supervised learning. Now, in case of semi-supervised learning, we have training data, also, we have a few desired outputs or small number of labelled outputs. So, we can see that semi supervised learning is an intermediate between supervised learning and unsupervised learning.

Again, in case of supervised learning, the outputs are labelled in case of unsupervised learning the outputs are non-labelled or unlabelled and semi-supervised learning, the a few of the outputs are labelled. A fourth category learning is also there that is called reinforcement learning. Now, what is reinforcement learning?

Reinforcement learning is a learning which rewards from the sequence of actions which gets, which counts on the rewards from the sequence of actions to determine its future course or path of actions. I will give you some examples. So, that is called reinforced learning, rewarding generally in case of reinforcement learning, reward is desired behaviours and, or punishing undesired ones.

So, generally what happens in case of reinforced learning it gives rewards to desired behaviours which will help to achieve the goal of the problem and it gives, it punishes the undesired step which will not help to get or achieve the desired objective that is called reinforcement learning. We will, will see an example.

(Refer Slide Time: 08:58)



So, let us see some examples of supervised learning. Now, if we divide the supervised learning supervised learning of 2 types, one is regression and other is classification. So, in case of regression, let us see regression. In case of regression, where we have given x1 y1 where x1 is the input and y1 is output x2 y2 up to xn yn and our objective is to learn a function fx to predict y given x y our numerical.

So, when you go for the regression, when both x and y are numerical, and continuous. So, then we only can go with the regression. When y is categorical then, we call it classification problem; we will discuss later. But in case of regression both x and y are numerical and continuous.

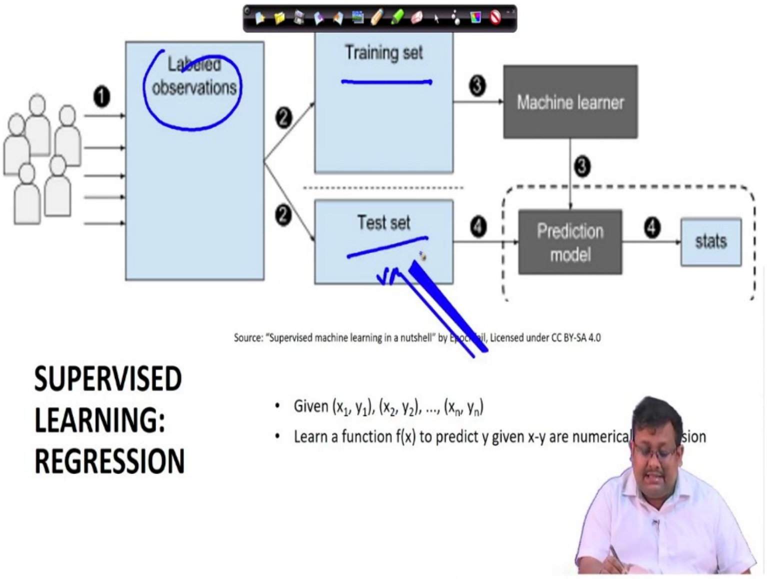
So, you can see that this is a class regression problem where the there is a relationship between predicted soil organic carbon and measured soil organic carbon, we wanted to discuss the relationship between the predicted soil organic carbon and measured soil organic carbon and here you can see both calibration and training data set as well as validation and testing data sets.

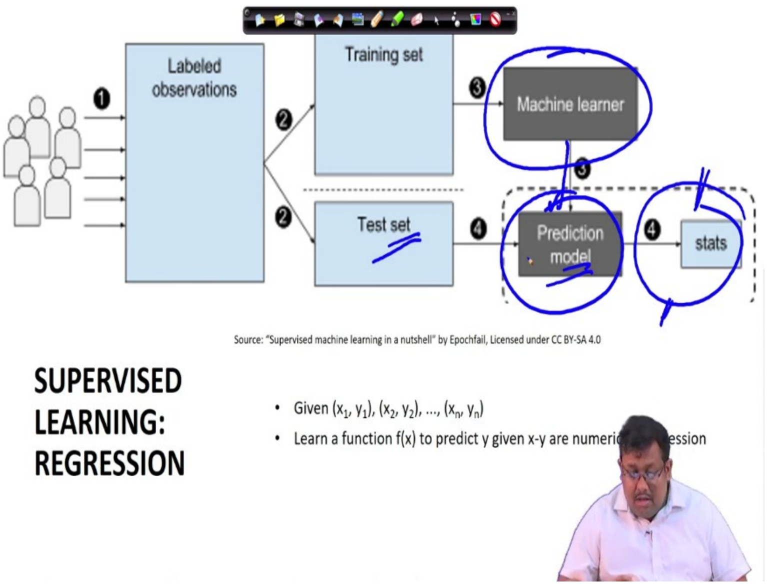
So, what is calibration data set? Calibration data set is the data set based on which we did we first fit a model and what is testing data set or validation data set? Testing data set is a subset of the data or some time it could be totally new or independent. These independent data set which we require to validate the accuracy of the training model is called the validation sample or testing data sets.

Sometime our training model could be highly accurate however, our validation gives very low performance. So, that is why we need to validate each and every calibration model. We cannot base our conclusion based on only validation data, calibration data set, we have to validate the model.

So, here one example is given here you can see x and y here x is the actual soil organic carbon whereas, the y is the predicted soil organic carbon and we wanted to see the relationship be using a regression method and this is an example of regression. I hope now it is clear to you.

(Refer Slide Time: 12:09)



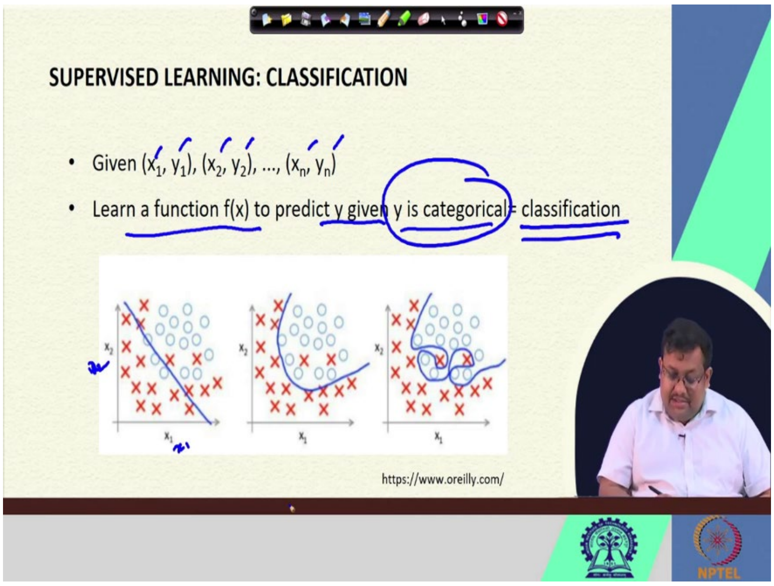


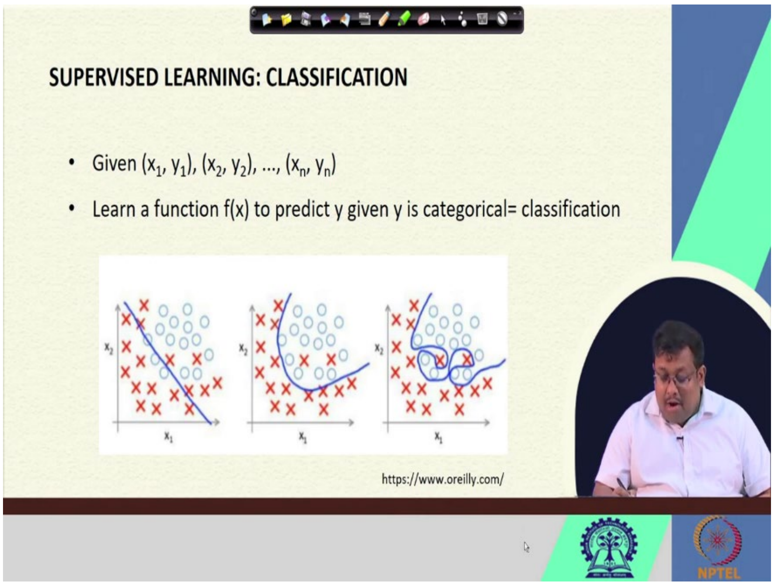
Now, so, if we talk about the overview of supervised learning, we can see that here we are incorporating the level observations and once we are incorporating the labelled observation, we can subdivide them into a training set or calibration set or test set or validation set.

And using the training set, we can create the machine learner and the prediction model and this machine learner can use this training set to create the prediction model and the validation set or test set is used to validate this prediction model and finally, generate the statistics. So, that the user can determine whether this prediction model is accurate or inaccurate.

So, this is the brief overview of supervised regression or supervised learning through regression. Another category of supervised learning is there that is classification.

(Refer Slide Time: 13:39)

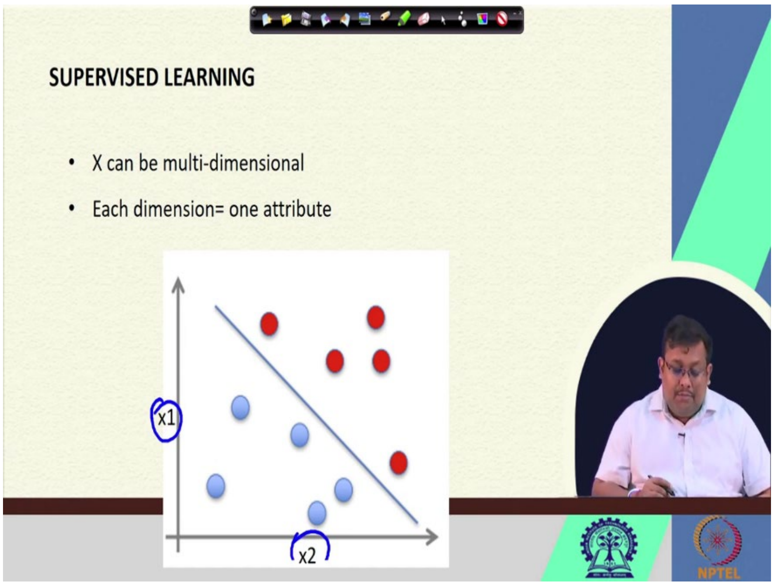




Now, what is classification? In case of classification as I have told you that just like in case of it just like in case of prediction, we have given x1 y1, x2 y2 up to xn yn. But in case of classifications remember that y is always categorical. So, when we want to learn a function fx to predict y given y is categorical then we call this problem as a classification problem. For example, here you can see there are two inputs x1 and x2 and we wanted and these are the observations.

These are all the observations which are mixed together and we want to develop a function which can separate the two different categories and maintaining maximum homogeneity. So that is called classification problem. And remember in this case, a classification problem our target is always categorical in nature. So, we can draw the boundary in different fashion. You can see Linear boundaries can be drawn and more intricate complex boundary also can be done in case of classification problem.

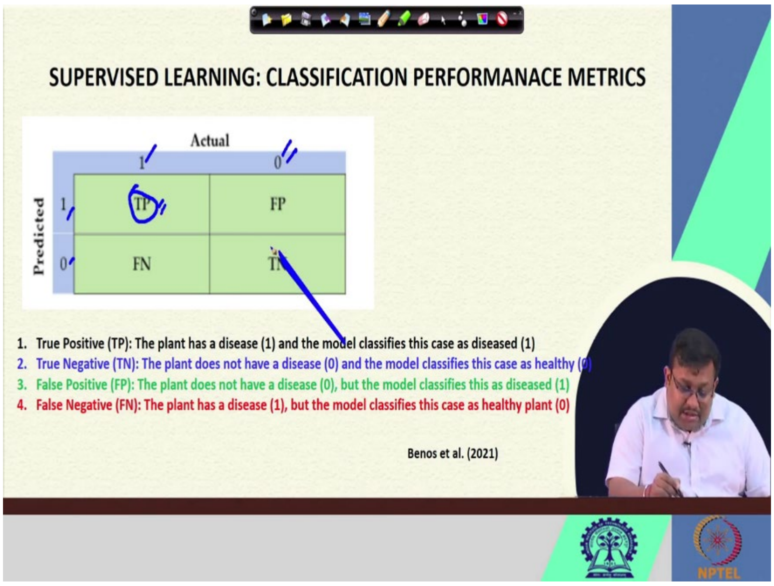
(Refer Slide Time: 15:09)

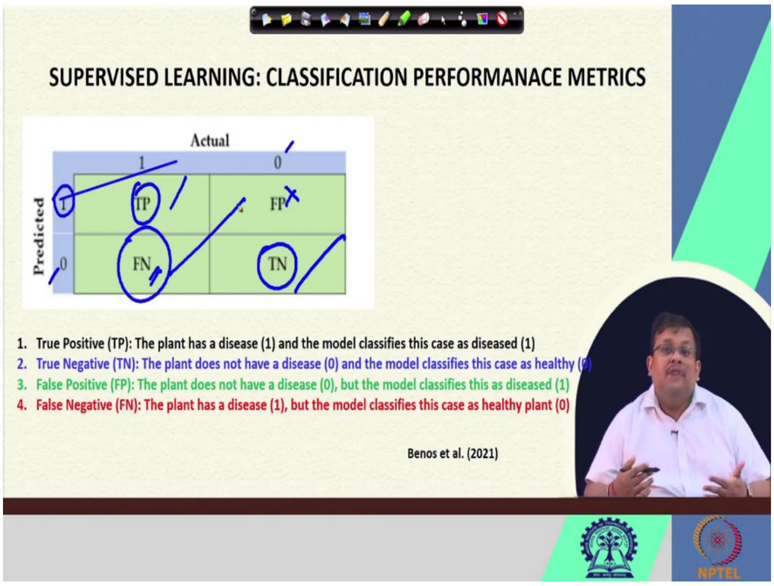


Now, as I have shown in my previous slide, any classification problem are supervised learning in that sense, in case of classification problem there could be, the problem could be multi-dimensional. So, we can use more than one dimension of the data x1 x2 you can see here and each dimension is equal to 1 attribute.

So, if we use multi dimension to classify the data that is also possible whereas, each dimension represents an attribute or input in the data. So, we have covered what is supervised learning, what is supervised prediction and supervised classification.

(Refer Slide Time: 16:04)





Now, the performance metrics of any supervised classification method can be described in terms of confusion matrix and you can see here this is a confusion matrix which gives the score or which gives the number of the observation which are actually rightly classified or wrongly classified.

So, you can see here suppose, we are discussing some plant which we have both healthy plants as well as diseased plants and using some classification scheme we wanted to see the accuracy of our classification scheme to perfectly separate the samples in the healthy plants as well as the disease plants.

So, this is a perfect this is a confusion matrix and these are the actual these are the actual these 1 and 0 are discussing the actual observations and these 1, 0 in the column vertical showing the predicted observations. Now, in case of actual observation 1 generally denote the healthy, 1 generally denote the disease plant whereas 0 indicate the healthy plant, which does not have any disease.

So, if we can have 4 different types of outcomes. True positive will be here when the plant has a disease and the model classifies these as this case has the disease, so it is perfect the plant actually has the disease and our classification model perfectly classify it as a disease plant, so it is called true positive. And what is second possibility, possibility is true negative.

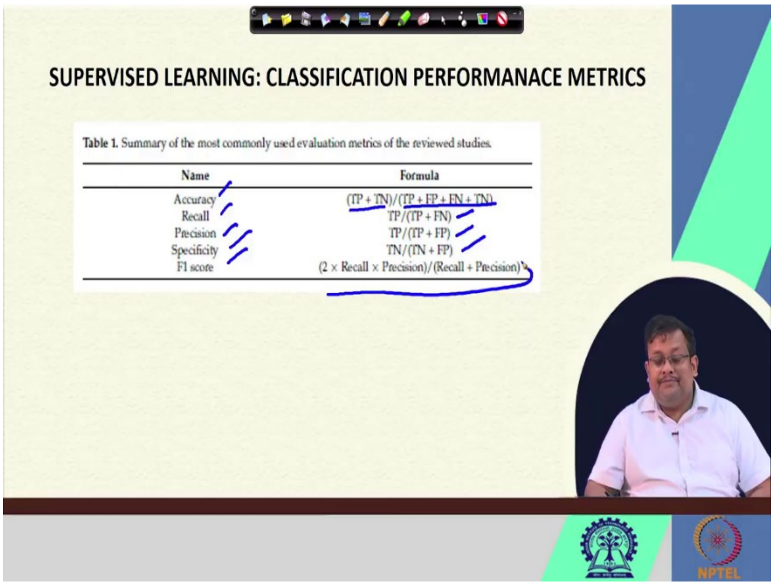
True negative means the plant does not have the disease and our classification model classifies and healthy. This is called true negative. So, we can see true positive and true negative. So, you can see these diagonals are always true because it true negative comes under these 0 and 0. So, that means it is okay, that means it is correct. The plant does not, plant is healthy, and the model also classifies it as healthy.

But there are two other options also. For example, false positive. False positive means when the plant does not have a disease, but the model classifies is as a healthy plant. I am sorry, model classifies as the disease and false negative is the plant has a disease but the model classifies this case as healthy plant that is called false negative.

So, these are some of the examples. So here false positive means the plant does not have a disease, here, which is actual, but the prediction gives the label of one. So that is called the false positive and finally, false negative means the plant has a disease that is actually 1, but the model classifies this as a healthy plant. So, this is called false negative. So, you can see.

These diagonal elements like true positive and true negative gives the correctly classified samples, whereas, these diagonals are off diagonal points are given, are generally using the wrongly classified samples. So, this is how we calculate the classification, performance, we call it performance metrics. When we will discuss the classification in subsequent in our in our coming weeks, we will discuss this confusion matrix in more details.

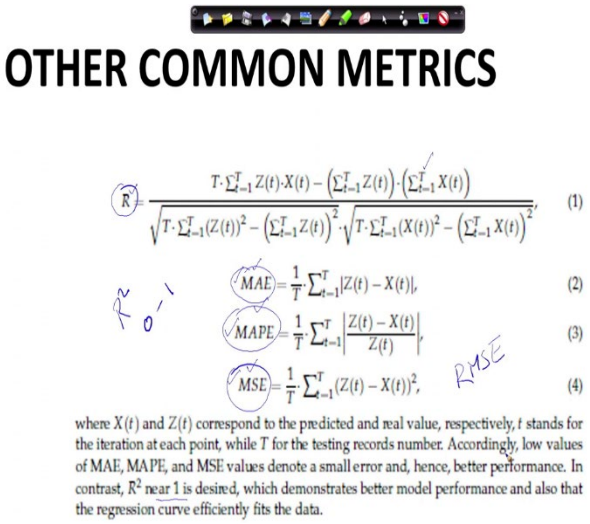
(Refer Slide Time: 20:55)



Now, based on the confusion matrix, there are most commonly used evaluation matrix are there, some of them are mentioned here. For example, if you want to see the accuracy, for accuracy you have to overall accuracy you have to sum the true positive and true negative and then you have to sum the divided by the whole number of samples and then recall has this formula TP by TP plus FN. Then precision has TP by TP plus FP.

Specificity has TN by TN plus FP and finally, F1 score is using this formula. So, based on the confusion matrix, it is possible to derive different types of other evaluation matrices also.

(Refer Slide Time: 22:02)



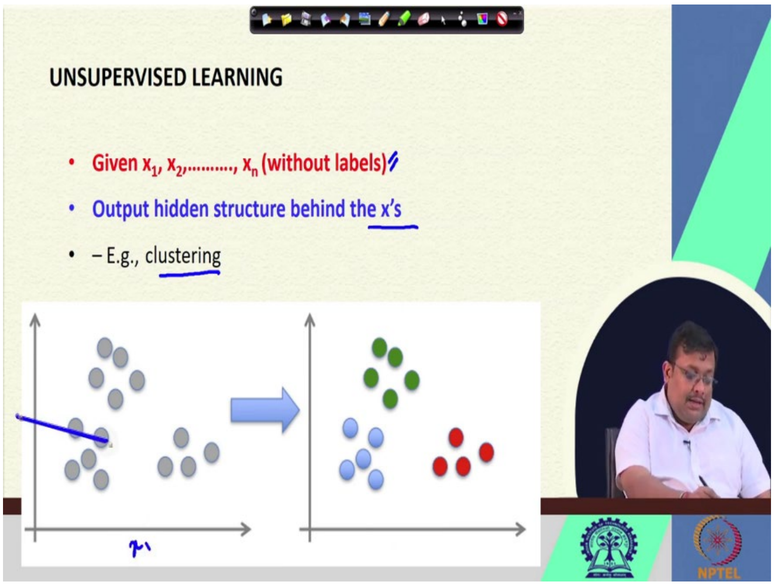
Among other common matrices, specially which we use for regression purpose, these are basically the correlation coefficient you can see here, this R stands for the correlation coefficient. Apart from that correlation coefficient we also denote R MSE which is root mean squared error and then the mean absolute error or MAE and then mean absolute percent error or MAPE and mean squared error that is MSE.

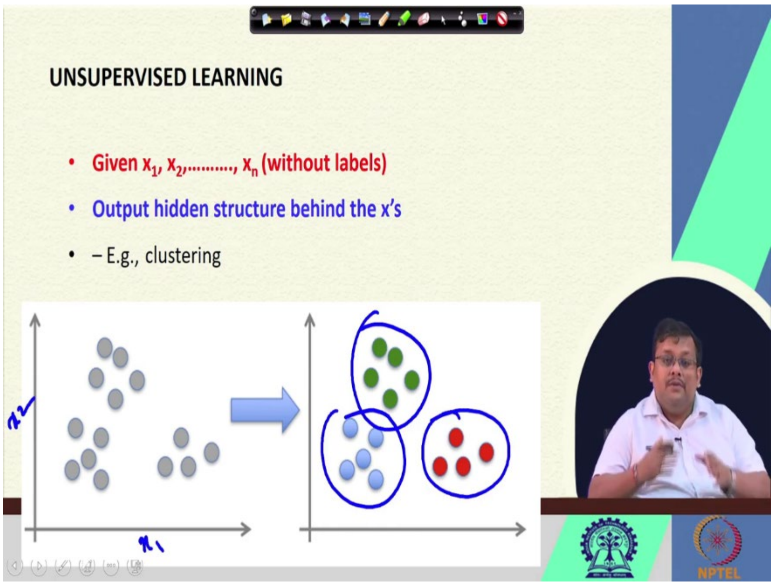
There is also root mean square error RMSE which is also an very common metric for determining the model accuracy. Remember in this formula, Xt is basically stands for the predicted value whereas Zt stands for the real value and t stands for the iteration at each point were the capital T, here this capital T for the testing record numbers, records number and then accordingly low values of all 3 matrices MAE, MAPE, MSE and also RMAC values denote a small error and hence better performance.

We always want our model to show lower values of these error estimates MAE, MAPE, MSE as well as our RMAC. We will, we will discuss these in details in our coming lectures. Also, note one thing that this is a correlation coefficient. However, there is another coefficient and this correlation coefficient generally varies from minus 1 to plus 1 whereas, if you take a square of R that is our square which is called coefficient of determination or regression coefficient that also used as a very important metric for model accuracy.

So, these R square values generally varies from one 0 to 1 and when the R squared value is near to 1 that is more desert and which demonstrates the better model performance and also that the regression curve efficiently fits the data. So, these are the common matrix matrices for regression problem, whereas, we have already discussed the confusion matrix for classification accuracy determination.

(Refer Slide Time: 24:42)



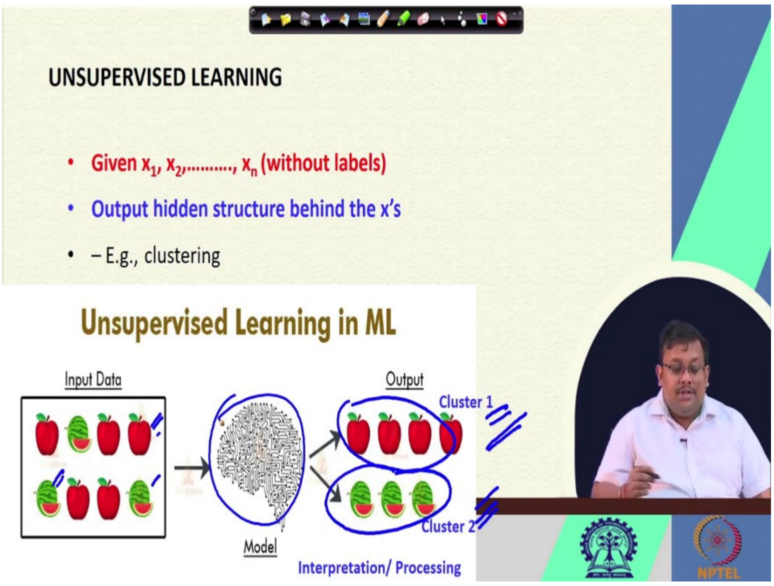


Now, let us see the examples of data set, let us discuss about the unsupervised learning. Now, in case of unsupervised learning x1, x2 up to xn we have but without the labels. We do not have any labels for the observations or outputs. So, output hidden structure behind the x. So, so, we need some kind of Machine Learning applications to extract the output from there and label them.

So, here we are not labelling them, but the Machine Learning algorithm is labelling those outcomes. So, example is clustering. So, here you can see that if we consider this is x1 and this is x2 and sorry, this is x1 and this this is x2. So, using these two attributes, if we can classify the samples, they are not labelled, but if we can make use, these three clusters, this is the example of unsupervised learning.

So, here we are not telling beforehand that which sample belong to which class, but we let the Machine Learning algorithm decide which sample belong to which class based on certain features of that observations.

(Refer Slide Time: 26:20)

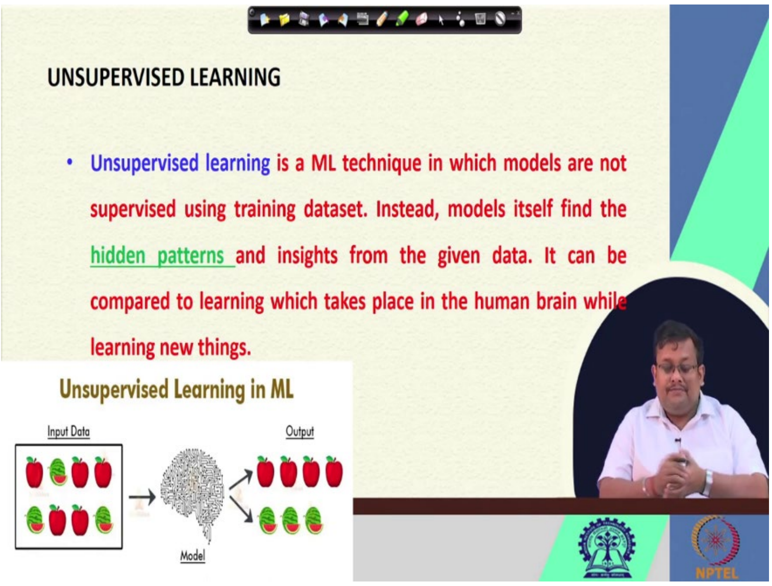


So, one example I can show you here for unsupervised learning, unsupervised learning you can see here there are some inputs data suppose, there are some input data where both apples and watermelons are there and in this unsupervised learning or clustering model, it helps to identify and clustered these apples and these watermelons in two separate cluster. So, this is an example of unsupervised learning.

Remember, we are not labelling that this is an apple, this is an apple and this is a watermelon, when we are incorporating the data, we are incorporating the whole data into the model and model decide based on certain feature and then classify them into two cluster. Cluster 1 will be a apple cluster and cluster 2 will be the watermelon cluster.

And this clustering is based on certain features. Suppose, the shape and certain colour. So, these will be identified by the model itself, the user is not incorporating these features. So, user will not tell the model that these are the labels, but model itself interpret based on certain features of the observations. So, this is called unsupervised learning.

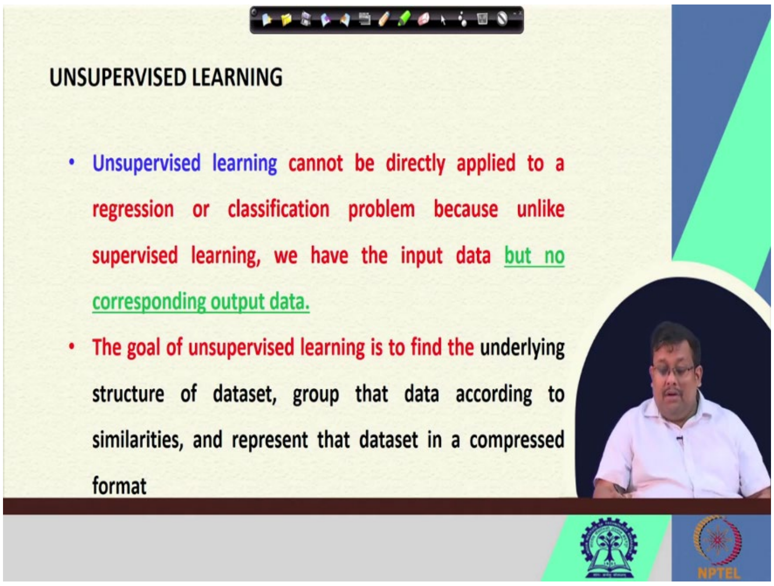
(Refer Slide Time: 27:56)



So, unsupervised learning if we define unsupervised learning, unsupervised learning is a Machine Learning technique in which models are not supervised using training data set instead, models itself find the hidden pattern, just like I told you, it itself find that the hidden pattern and insight from the given data and it can be compared to learning which takes place in the human brain, while learning new things.

When our brain is learning new things just by exploring sometimes, they are not clearly labelled, we are perceptive men and we are making perception and also, we are learning by your experience and we can easily differentiate them by seeing some of the features. So, this is an example of unsupervised learning.

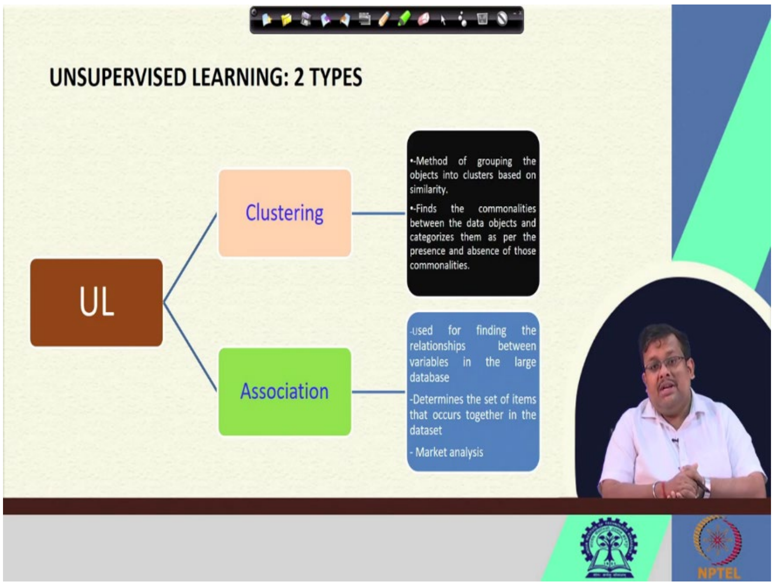
(Refer Slide Time: 28:48)



Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data we do not have any corresponding output data, because it is not labelled. So, the goal of unsupervised learning is to find the underlying structure of the dataset or group that data according to the similarities and we, these unsupervised learning can group the data according to the similarities and represent that data set in a compressed format.

So, just like we have seen in case of apple and watermelon example, these, based on some similarities, they are grouping the data, grouping the apples in a single cluster and also grouping the watermelon in another cluster. So, based on some similarities, like their colour, their shapes and, based on these features, they are clustering them in a single clusters.

(Refer Slide Time: 30:01)



So, this is example of unsupervised clustering. Unsupervised learning are of two types. One is clustering and another is association. And clustering basically shows us the method of grouping the objects into cluster based on the similarity and finds the commonalities between the data objects and the categorize them as per the presence or absence of those commonalities.

So, in case of apple and watermelon example, we have seen that type of classification or clustering based on some commonalities. Another example or another category of unsupervised learning is association. Association are used to find the the relationship between the variables in the large database and determines the set of items that occurs together in the data set.

And generally, these association rules are useful in the market analysis. So, guys, let us wrap up for lecture here we have learned something in this lecture, we have learned the, the supervised learning and unsupervised, learning different categories of supervised learning, classification regression, some classification matrices, regression matrices, and we have seen unsupervised learning or clustering, and we have seen their types also.

So, in the next lecture, we will start from here and we will discuss the reinforcement learning and, how reinforced learning is important in the Machine Learning domain and also the difference between Machine Learning and difference between the supervised learning as well as the reinforced learning and other aspects of Machine Learning and their application in the field of agriculture. Thank you. Let us meet in our lecture three.