Business Anaytics

M. Tech QROR – 2nd yr (2024)

Learning: Supervised vs. Unsupervised

(Predictive Analytics)

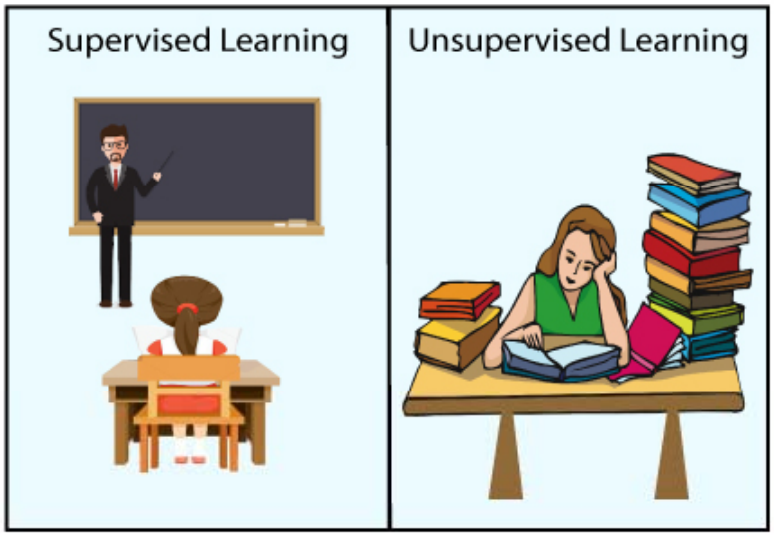
Dr. Prasun Das

SQC & OR Unit

Indian Statistical Institute

e-mail: prasun@isical.ac.in

Learning with Data



Supervised and Unsupervised learning are the basic two techniques of learning.

- **Supervised learning:** models are trained using labeled data.
- **Unsupervised learning:** patterns are inferred from the unlabeled data.

❑ Reinforcement Learning: depends on behaviour (state, reward & action/decision) of environment (dynamic) [ex. *autonomous driving system – cars, flights, robots*]

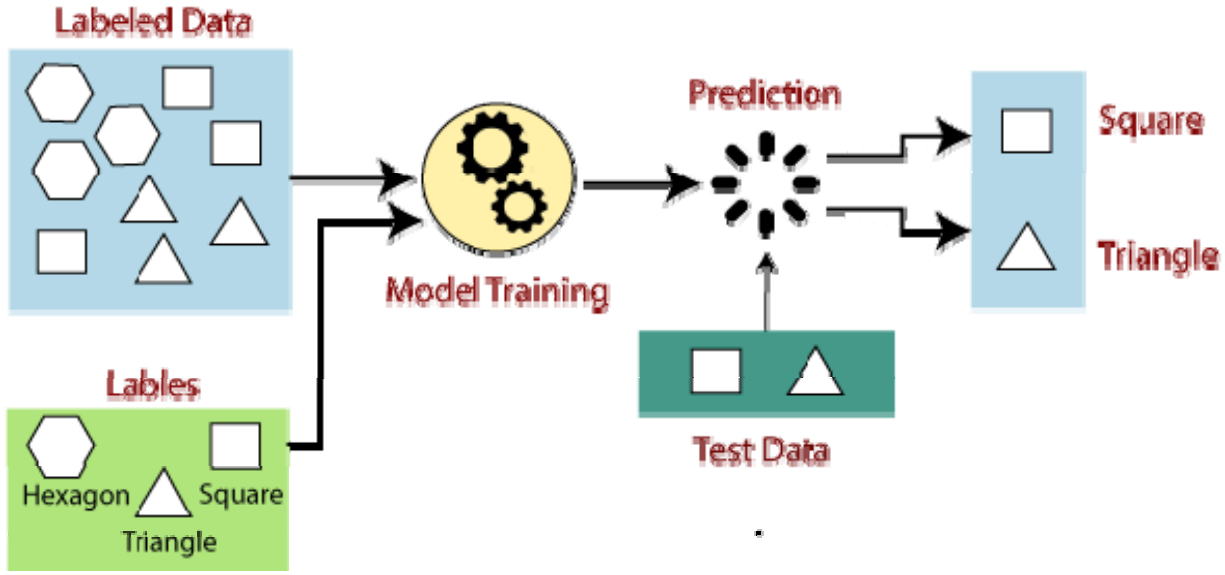
Supervised Learning

Supervised learning is the type of learning in which machines (computing frames) are trained using well "labelled" training data, and on basis of that data, machines learn and predict the output. The labelled data means some input data is already tagged with the correct output.

The training data, provided to the machines, work as the Supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

Supervised Learning: How it works?



Step-0: provide Training Data (labelled): shapes (as per the number of sides) like square, rectangle, triangle, and hexagon.

Step-1: train the model for each shape (data).

Step-2: test the trained model using the test dataset

Step-3: identify the shape (output Prediction)

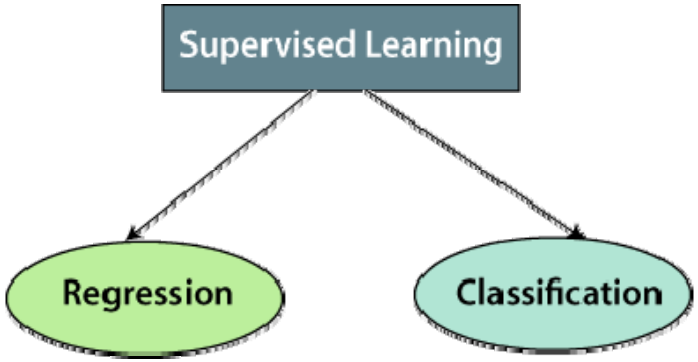
Supervised Learning: Broad Steps

- Determine the type of training dataset (input vs. correct output)
- Collect/Gather the labelled training data
- Split the dataset into **training dataset, test dataset,** and/or **validation dataset**.
- Determine the input features of the training dataset, which should have enough knowledge so that the model can accurately predict the output
- Determine the suitable training algorithm for the model (SVM/DT/RF/LR/ANN etc.)
- Execute the algorithm on the training dataset. Sometimes we need validation sets as the control parameters, which are the subset of training datasets
- Evaluate the accuracy of the model by providing the test set. If the model predicts the correct output, which means our model is accurate towards generalization.

Supervised Learning: Types

1. Regression

Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market trends, etc.



- Linear Regression
- Regression Trees
- Non-Linear Regression
- Bayesian Linear Regression
- Polynomial Regression

2. Classification

Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc.:

- Random Forest
- Decision Trees
- Logistic Regression
- Support Vector Machine

Supervised Learning: Advs & Disadvs

Advantages:

1. The model can predict the output on the basis of prior experiences.
2. We can have an exact idea about the classes of objects.
3. It helps us to solve various real-world problems such as **fraud detection, spam filtering**, etc.

Disadvantages:

1. Not suitable for handling the complex tasks.
2. Cannot predict the correct output if the test data is different from the training dataset.
3. Training required lots of computation times.
4. We need enough knowledge about the classes of object.

Unsupervised Learning

In Unsupervised learning, models are trained using unlabeled dataset and are allowed to act on that data without any supervision. It can be compared to learning which takes place in the human brain while learning new things.

The goal of unsupervised learning is to find the underlying structure (patterns) of the dataset, group that data according to similarities, and represent that dataset in a compressed format.

Example:

Suppose, an input dataset contains images of different types of cats and dogs.

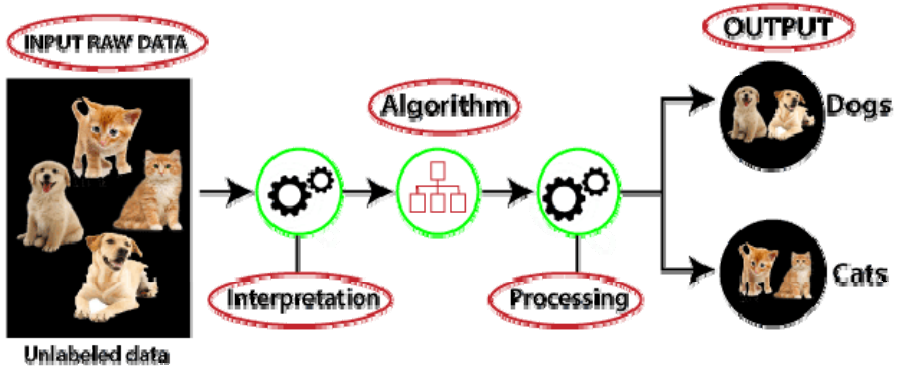
The task of an unsupervised learning algorithm is to identify the features of those images on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images.

Unsupervised Learning – Why to use?

Importance:

- It is helpful for finding useful insights from the data.
- It is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- It works on unlabeled and uncategorized data which make unsupervised learning more important as because, in real-world, we do not always have input data with the corresponding output.

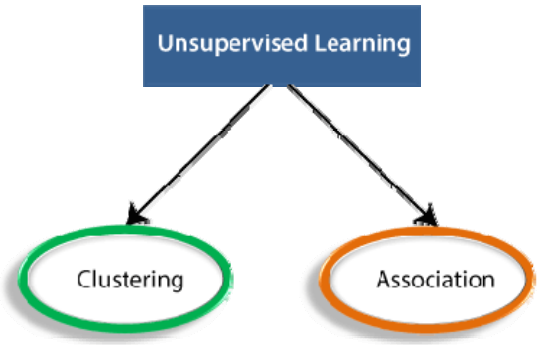
Here, we have taken an unlabeled input data, which is fed to the model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms (such as k-means clustering, Decision tree etc.) to divide the data objects into groups according to the similarities and difference between the objects.



Unsupervised Learning – Types

1. Clustering

Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.



- ✓ K-means clustering
- ✓ KNN (k-nearest neighbors)
- ✓ Hierarchal clustering
- ✓ Anomaly detection
- ✓ Principle Component Analysis
- ✓ Independent Component Analysis
- ✓ Factor Analysis (EFA, CFA)
- ✓ Apriori algorithm
- ✓ Singular value decomposition
- ✓ Structural Equation Modeling

2. Association

An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

Unsupervised Learning: Advs & Disadvs

Advantages:

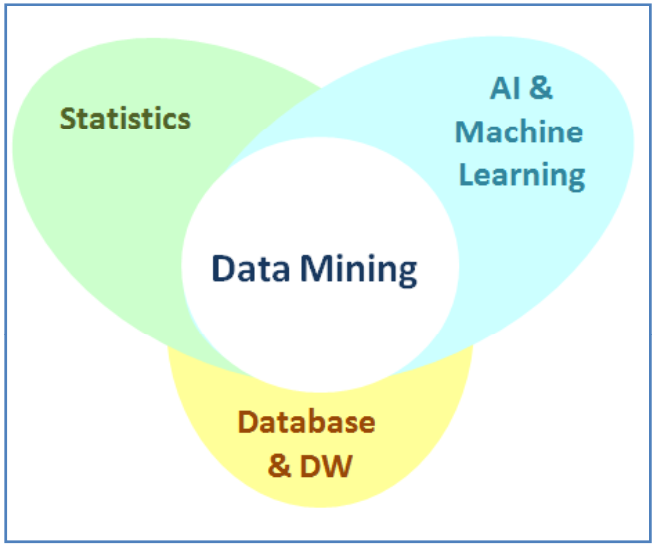
1. It is used for more complex tasks as compared to supervised learning because, in unsupervised learning, we don't have labeled input data.
2. It is preferable as it is easy to get unlabeled data in comparison to labeled data (ground truth).

Disadvantages:

1. It is intrinsically more difficult than supervised learning as it does not have corresponding output.
2. The result of an unsupervised learning algorithm might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

Data Mining

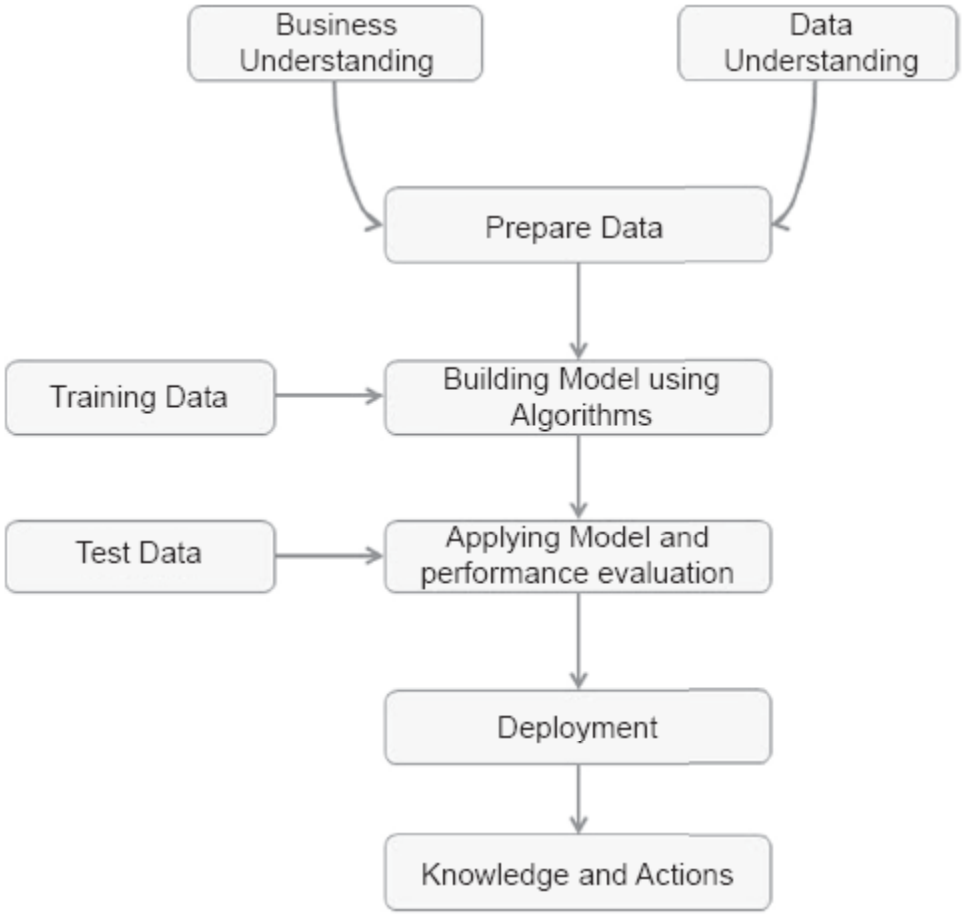
1. Extraction of knowledge (finding the hidden signals / rules / patterns) from Data (often BIG in nature, i.e., large data set) to explain the past.
2. A multi-disciplinary field which combines statistics, machine learning, artificial intelligence and database theories & management technology.
3. Typically operates on large data sets that need to be stored, processed, and computed. Hence, database techniques along with parallel and distributed computing techniques play an important role.



Data Mining: Tasks with Examples

Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of the predefined classes. The prediction will be based on learning from a known data set.	Decision trees, neural networks, Bayesian models, induction rules, k-nearest neighbors	Assigning voters into known buckets by political parties, e.g., soccer moms Bucketing new customers into one of the known customer groups
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from a known data set.	Linear regression, logistic regression	Predicting unemployment rate for next year Estimating insurance premium
Anomaly detection	Predict if a data point is an outlier compared to other data points in the data set.	Distance based, density based, local outlier factor (LOF)	Fraud transaction detection in credit cards Network intrusion detection
Time series	Predict the value of the target variable for a future time frame based on historical values.	Exponential smoothing, autoregressive integrated moving average (ARIMA), regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the data set based on inherit properties within the data set.	k-means, density-based clustering (e.g., density-based spatial clustering of applications with noise [DBSCAN])	Finding customer segments in a company based on transaction, web, and customer call data
Association analysis	Identify relationships within an item set based on transaction data.	Frequent Pattern Growth (FP-Growth) algorithm, Apriori algorithm	Find cross-selling opportunities for a retailer based on transaction purchase history

Data Mining Process: Generic Steps



1. Prior Knowledge

2. Preparation

3. Modeling

4. Application

5. Knowledge

Data Mining Process: Business Problem

The data mining process starts with an analysis need, a question or a business objective. This is possibly the most important step in the data mining process (*Shearer*, 2000).

Process: consumer loan business

Business Objective: *If we know the **interest rate** of past borrowers with a range of **credit scores**, can we predict interest rate for a new borrower?*

Correlation vs. Causation: *Can we predict the credit score of the borrower based on interest rate?*
The answer is yes—but it doesn't make business sense.

Note: The correlation between the input and output attributes doesn't guarantee causation. Hence, it is very important to frame the data mining question correctly using the existing domain and data knowledge.

In this ex., we need to predict the interest rate of the new borrower with unknown interest rate, based on available credit score (Business Objective).

Data Mining Process: Data Preparation

Consists of:

a) Data Quality – Cleaning, Transformation, Reduction (Preparatory Analytics)
b) Data Visualization – Visual Analytics
c) Data Exploration – Exploratory Data Analysis (Exploratory & Descriptive Analytics)
d) Data Sampling – Training, Testing, Validation
e) Data Modelling – Statistical / ML / Others

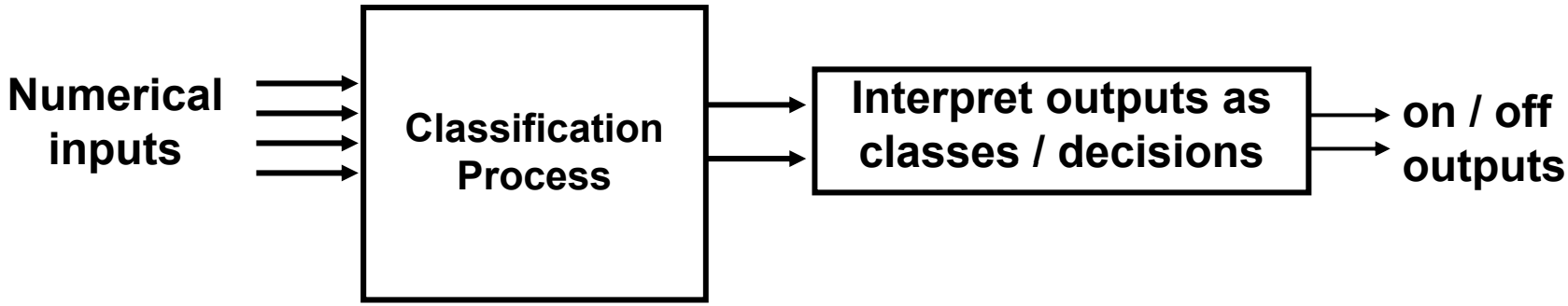
Data Mining Process: Data Modeling

A **Model** is the abstract representation of the data and its relationships in a given data set. A simple statement like “mortgage interest rate reduces with increase in credit score” is a model; although there is not enough quantitative information to use in a business scenario, it provides directional information to abstract a relationship between credit score and interest rate.

Borrower ID	Credit Score	Interest Rate
01	500	7.31%
02	600	6.70%
03	700	5.95%
04	700	6.40%
05	800	5.40%
06	800	5.70%
07	750	5.90%
08	550	7.00%
09	650	6.50%
10	825	5.70%

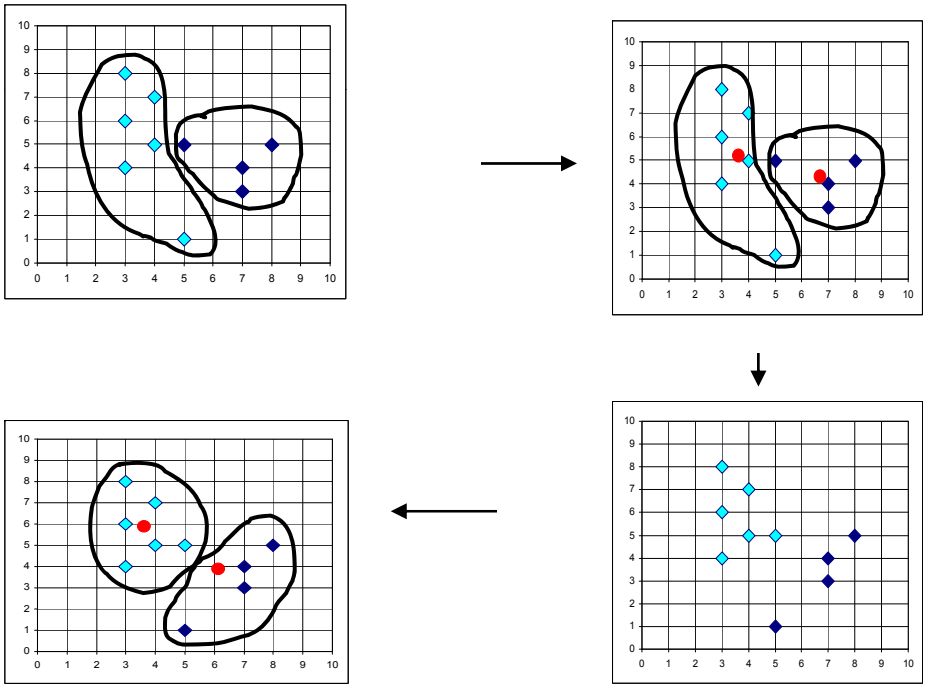
Data Modeling: Types

1. **Classification:** to determine the process of identifying similar groups of objects and developing rules for assigning new objects into one of the existing groups.



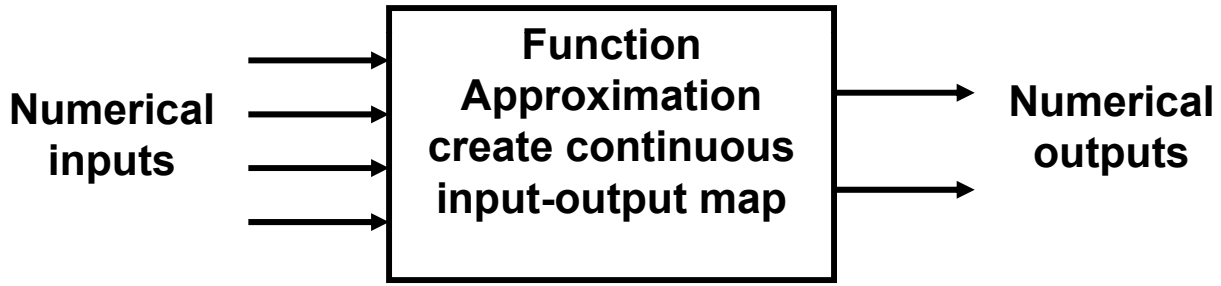
Data Modeling: Types

2. **Clustering:** Grouping of objects on the basis of similarities or distances (dissimilarities) in such a way that objects belonging to the same cluster resemble each other, whereas objects in different clusters are dissimilar.



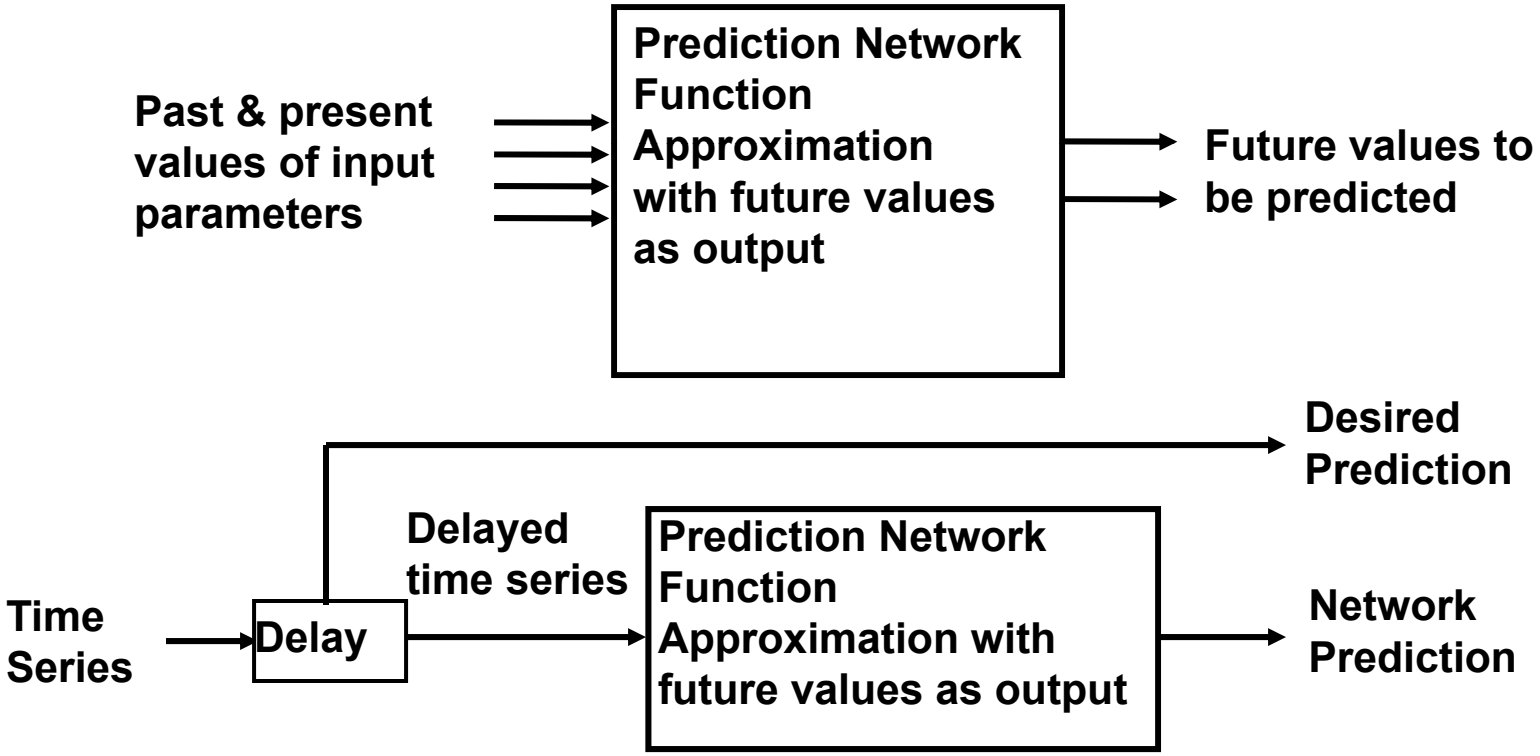
Data Modeling: Types

3. **Prediction:** to determine future value(s) of the response variable that are associated with a specific combination of predictor variable values.



Data Modeling: Types

4. **Forecasting:** to determine future value(s) of the response variable at future time point(s) that are associated with a specific combination of past time dependent input variable values.



Q. What we learnt ??