Data Science

Introduction

Dr. K R Remesh Babu

Professor & Head

Department of Information Technology

Government Engineering College Idukki

1. Introduction

Introduction

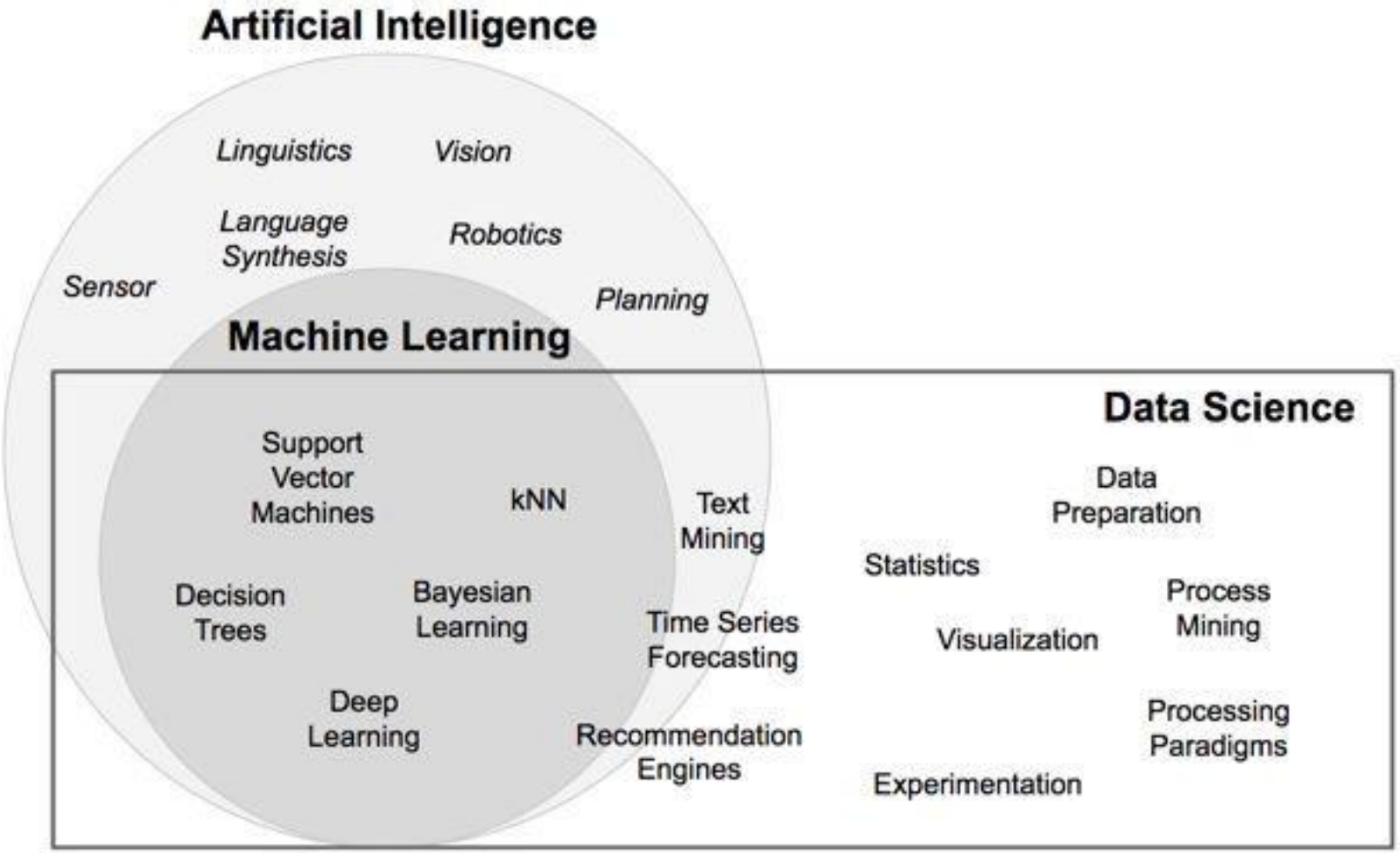
- Why Data Science?
  - We have run out of adjectives and superlatives to describe the growth trends of data. The technology revolution has brought about the need to process, store, analyze, and comprehend large volumes of diverse data in meaningful ways.
- Generally, Data science is a collection of techniques used to extract value from data. It has become an essential tool for any organization that collects, stores, and processes data as part of its operations.

However, the value of the stored data is zero unless it is acted upon.

AI, Machine Learning, and Data Science

- Artificial intelligence is about giving machines the capability of mimicking human behavior, particularly cognitive functions.
- Machine learning can either be considered a sub-field or one of the tools of artificial intelligence, is providing machines with the capability of learning from experience. Experience for machines comes in the form of data.
- Data science is the business application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics. It is an interdisciplinary field that extracts value from data.

What is Data Science



What is Data Science?

- **Data science starts with data**, which can range from a simple array of a few numeric observations to a complex matrix of millions of observations with thousands of variables.
- Data science **utilizes certain specialized computational methods** in order to discover meaningful and useful structures within a dataset.
- The discipline of data science **coexists** and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence (BI).

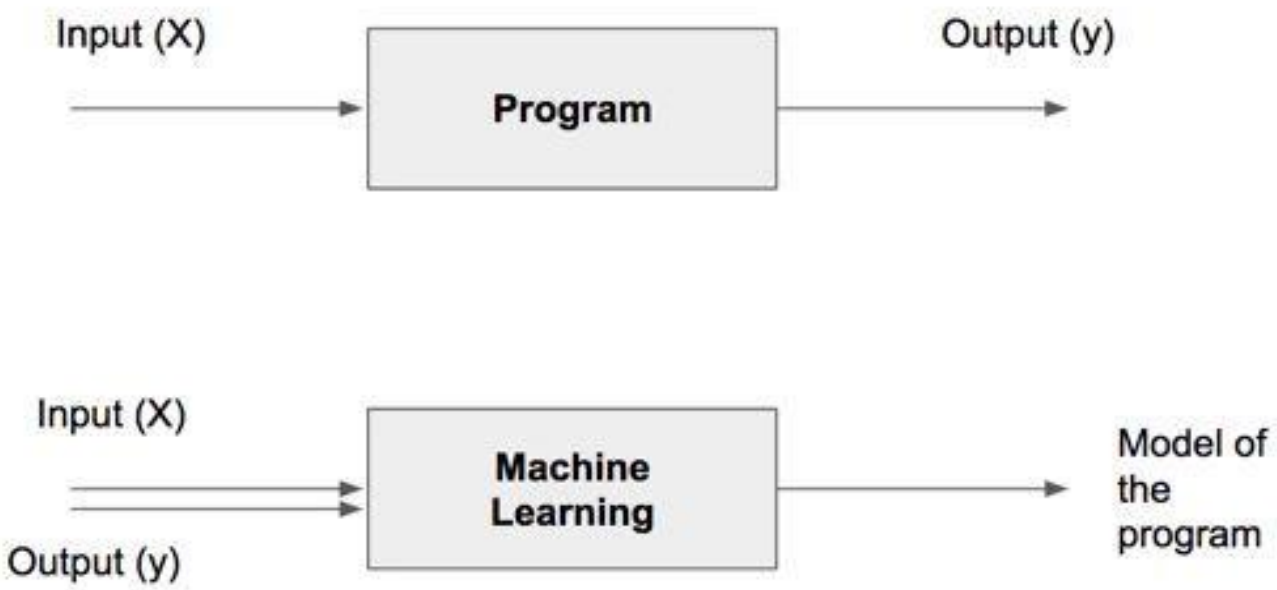
Data Science

- Extracting Meaningful Patterns
- Building Representative Models
- Combination of Statistics, Machine Learning, and Computing
- Learning Algorithms
- Associated Fields

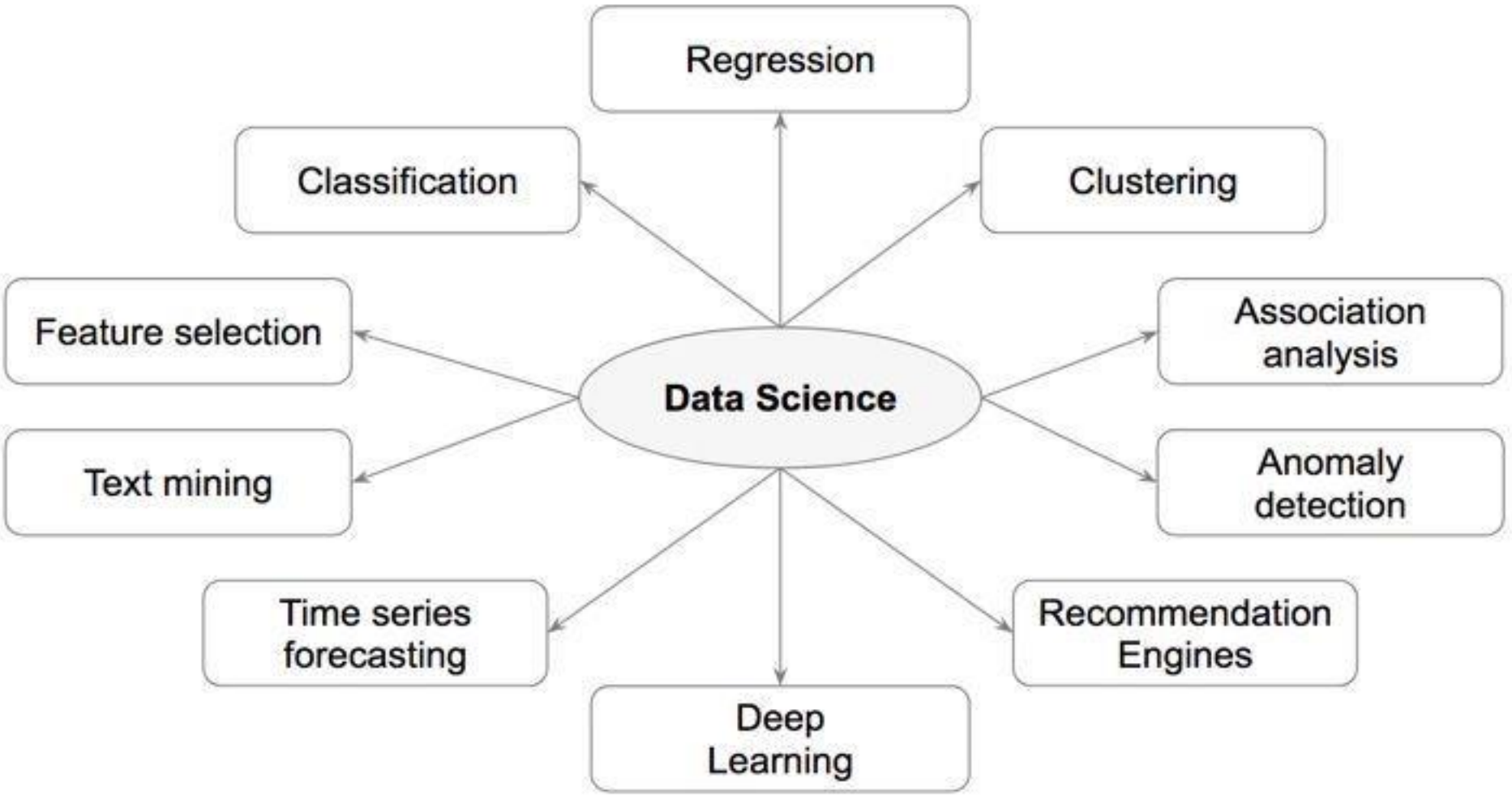
Data Science

- The ultimate objective of data science is to find potentially useful conclusions that can be acted upon by the users of the analysis.
- Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or relationships within a dataset in order to make important decisions.
- One of the key aspects of data science is the process of generalization of patterns from a dataset. The generalization should be valid, not just for the dataset used to observe the pattern, but also for new unseen data. Data science is also a process with defined steps, each with a set of tasks.
- The term novel indicates that data science is usually involved in finding previously unknown patterns in data.
- The ultimate objective of data science is to find potentially useful conclusions that can be acted upon by the users of the analysis

Models


Input (X)
Program
Output (y)
Input (X)
Output (y)
Machine Learning
Model of the program

Types of Data Science



Tasks	Description	Algorithms	Examples
Classification	Predict if a data point belongs to one of predefined classes. The prediction will be based on learning from known data set.	Decision Trees, Neural networks, Bayesian models, Induction rules, K nearest neighbors	Assigning voters into known buckets by political parties eg: soccer moms. Bucketing new customers into one of known customer groups.
Regression	Predict the numeric target label of a data point. The prediction will be based on learning from 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, LOF	Fraud transaction detection in credit cards. Network intrusion detection.
Time series	Predict if the value of the target variable for future time frame based on history values.	Exponential smoothing, 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 - DBSCAN	Finding customer segments in a company based on transaction, web and customer call data.
Association analysis	Identify relationships within an itemset based on transaction data.	FP Growth, Apriori	Find cross selling opportunities for a retailor based on transaction purchase history.