DATA SCIENCE

Prepared By,

SHELLY SHIJU GEORGE

Assistant Professor

DATA SCIENCE - INTRODUCTION

- 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.
- Data science techniques rely on finding useful patterns, connections, and relationships within data.
- Being a buzzword, there is a wide variety of definitions and criteria for what constitutes data science.
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining.

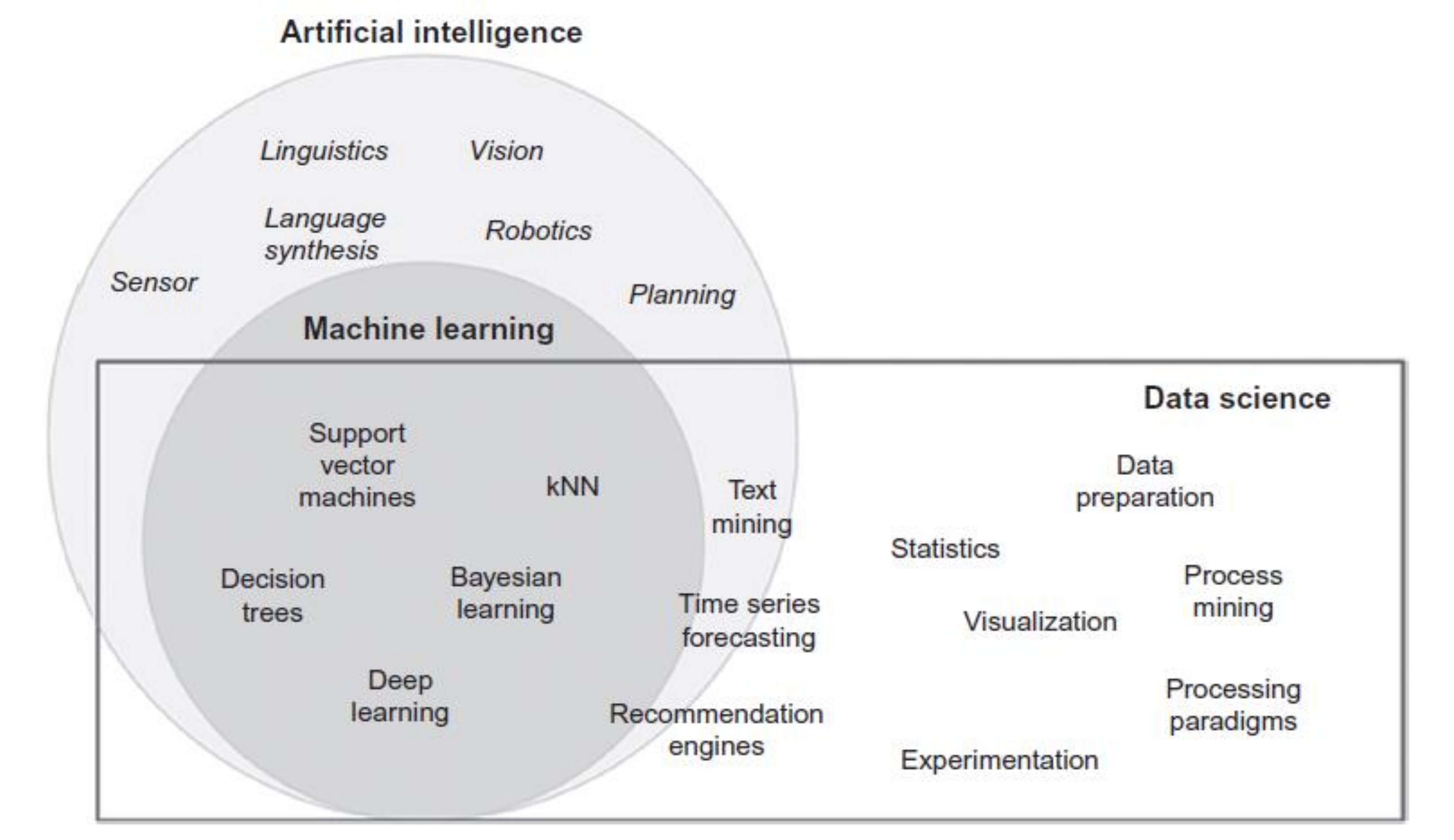
DATA SCIENCE – INTRODUCTION (CONTINUED)

- The use of the term science in data science indicates that the methods are evidence based, and are built on empirical knowledge, more specifically historical observations.
- To get meaningful results from any data, a major effort preparing, cleaning, scrubbing, or standardizing the data is still required, before the learning algorithms can begin to crunch them.

AI, MACHINE LEARNING, AND DATA SCIENCE

- Artificial intelligence, Machine learning, and data science are all related to each other.
- However, all of these three fields are distinct depending on the context.
- The following diagram shows the relationship between artificial intelligence, machine learning, and data science.

AI, MACHINE LEARNING, AND DATA SCIENCE (CONTINUED)



AI, MACHINE LEARNING, AND DATA SCIENCE (CONTINUED)

- Artificial intelligence is about giving machines the capability of mimicking human behavior, particularly cognitive functions.
- Examples would be: facial recognition, automated driving, sorting mail based on postal code.
- In some cases, machines have far exceeded human capabilities (sorting thousands of postal mails in seconds) and in other cases we have barely scratched the surface (search “artificial stupidity”).
- There are quite a range of techniques that fall under artificial intelligence: linguistics, natural language processing, decision science, bias, vision, robotics, planning, etc.
- Learning is an important part of human capability. In fact, many other living organisms can learn.

AI, MACHINE LEARNING, AND DATA SCIENCE (CONTINUED)

- 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 that is used to teach machines is called training data.
- Machine learning turns the traditional programing model upside down.
- A program, a set of instructions to a computer, transforms input signals into output signals using predetermined rules and relationships.

AI, MACHINE LEARNING, AND DATA SCIENCE (CONTINUED)

- Machine learning algorithms, also called “learners”, take both the known input and output (training data) to figure out a model for the program which converts input to output.
- Once the data science rules or model is developed, machines can start categorizing the disposition of any new posts.

AI, MACHINE LEARNING, AND DATA SCIENCE (CONTINUED)

- 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.
- In the context of how data science is used today, it relies heavily on machine learning and is sometimes called data mining.
- Examples of data science user cases are: recommendation engines that can recommend movies for a particular user, a fraud alert model that detects fraudulent credit card transactions, find customers who will most likely churn next month, or predict revenue for the next quarter.

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).

EXTRACTING MEANINGFUL PATTERNS

- 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.
- Data science involves inference and iteration of many different hypotheses.
- 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.

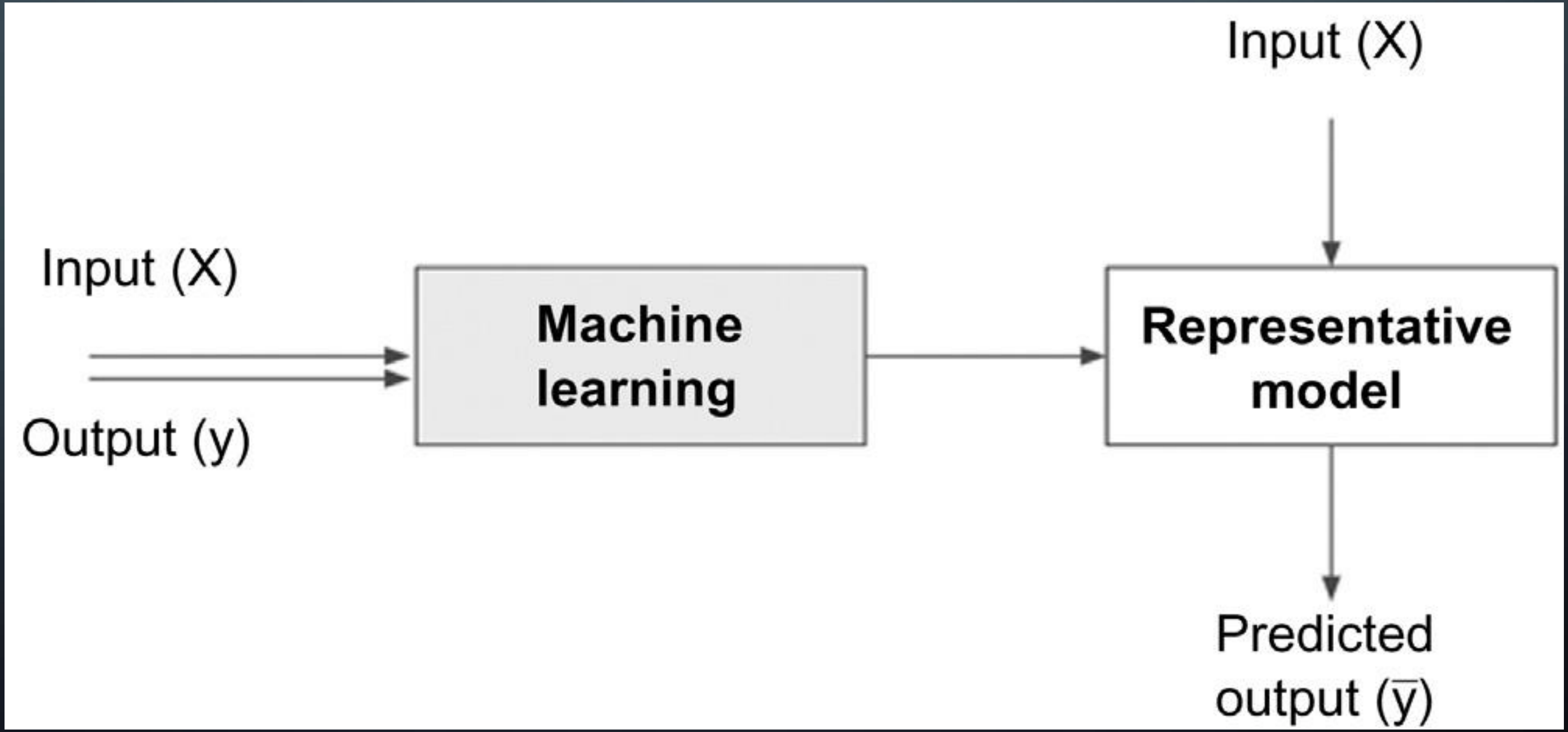
EXTRACTING MEANINGFUL PATTERNS (CONTINUED)

- 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.

BUILDING REPRESENTATIVE MODELS

- In statistics, a model is the representation of a relationship between variables in a dataset.
- It describes how one or more variables in the data are related to other variables.
- Modeling is a process in which a representative abstraction is built from the observed dataset.

BUILDING REPRESENTATIVE MODELS (CONTINUED)



BUILDING REPRESENTATIVE MODELS (CONTINUED)

- Data science is the process of building a representative model that fits the observational data.
- This model serves two purposes: on the one hand, it predicts the output (interest rate) based on the new and unseen set of input variables (credit score, income level, and loan amount), and on the other hand, the model can be used to understand the relationship between the output variable and all the input variables.
- A Model can be used for both predictive and explanatory applications.

COMBINATION OF STATISTICS, MACHINE LEARNING, AND COMPUTING

- In the pursuit of extracting useful and relevant information from large datasets, data science borrows computational techniques from the disciplines of statistics, machine learning, experimentation, and database theories.
- The algorithms used in data science originate from these disciplines but have since evolved to adopt more diverse techniques such as parallel computing, evolutionary computing, linguistics, and behavioral studies.
- One of the key ingredients of successful data science is substantial prior knowledge about the data and the business processes that generate the data, known as **subject matter expertise**.

COMBINATION OF STATISTICS, MACHINE LEARNING, AND COMPUTING (CONTINUED)

- Like many quantitative frameworks, data science is an iterative process in which the practitioner gains more information about the patterns and relationships from data in each cycle.
- Data science also typically operates on large datasets that need to be stored, processed, and computed.
- This is where database techniques along with parallel and distributed computing techniques play an important role in data science.

LEARNING ALGORITHMS

- We can also define data science as a process of discovering previously unknown patterns in data using **automatic iterative methods**.
- The application of sophisticated learning algorithms for extracting useful patterns from data differentiates data science from traditional data analysis techniques.
- Many of these algorithms were developed in the past few decades and are a part of machine learning and artificial intelligence.
- Some algorithms are based on the foundations of Bayesian probabilistic theories and regression analysis, originating from hundreds of years ago.

LEARNING ALGORITHMS (CONTINUED)

- These iterative algorithms automate the process of searching for an optimal solution for a given data problem.
- Based on the problem, data science is classified into tasks such as classification, association analysis, clustering, and regression.
- Each data science task uses specific learning algorithms like decision trees, neural networks, k-nearest neighbors (k-NN), and k-means clustering, among others.
- With increased research on data science, such algorithms are increasing, but a few classic algorithms remain foundational to many data science applications.

ASSOCIATED FIELDS

- While data science covers a wide set of techniques, applications, and disciplines, there a few associated fields that data science heavily relies on.

The techniques used in the steps of a data science process and in conjunction with the term “data science” are:

- Descriptive statistics
- Exploratory visualization
- Dimensional slicing
- Hypothesis testing
- Data engineering
- Business intelligence

ASSOCIATED FIELDS (CONTINUED)

1. Descriptive statistics

- Computing mean, standard deviation, correlation, and other descriptive statistics, quantify the aggregate structure of a dataset.
- This is essential information for understanding any dataset in order to understand the structure of the data and the relationships within the dataset.
- They are used in the exploration stage of the data science process.

ASSOCIATED FIELDS (CONTINUED)

2. Exploratory visualization

- The process of expressing data in visual coordinates enables users to find patterns and relationships in the data and to comprehend large datasets.
- Similar to descriptive statistics, they are integral in the pre- and post-processing steps in data science.

ASSOCIATED FIELDS (CONTINUED)

3. Dimensional slicing

- Online analytical processing (OLAP) applications, which are prevalent in organizations, mainly provide information on the data through dimensional slicing, filtering, and pivoting.
- OLAP analysis is enabled by a unique database schema design where the data are organized as dimensions (e.g., products, regions, dates) and quantitative facts or measures (e.g., revenue, quantity).
- With a well defined database structure, it is easy to slice the yearly revenue by products or combination of region and products.
- These techniques are extremely useful and may unveil patterns in data (e.g., candy sales decline after Halloween in the United States).

ASSOCIATED FIELDS (CONTINUED)

4. Hypothesis testing

- In confirmatory data analysis, experimental data are collected to evaluate whether a hypothesis has enough evidence to be supported or not.
- There are many types of statistical testing and they have a wide variety of business applications (e.g., A/B testing in marketing).
- In general, data science is a process where many hypotheses are generated and tested based on observational data.
- Since the data science algorithms are iterative, solutions can be refined in each step.

ASSOCIATED FIELDS (CONTINUED)

5. Data engineering

- Data engineering is the process of sourcing, organizing, assembling, storing, and distributing data for effective analysis and usage.
- Database engineering, distributed storage, and computing frameworks (e.g., Apache Hadoop, Spark, Kafka), parallel computing, extraction transformation and loading processing, and data warehousing constitute data engineering techniques.
- Data engineering helps source and prepare for data science learning algorithms.

ASSOCIATED FIELDS (CONTINUED)

6. Business intelligence

- Business intelligence helps organizations consume data effectively.
- It helps query the ad hoc data without the need to write the technical query command or use dashboards or visualizations to communicate the facts and trends.
- Business intelligence specializes in the secure delivery of information to right roles and the distribution of information at scale.
- Historical trends are usually reported, but in combination with data science, both the past and the predicted future data can be combined.
- BI can hold and distribute the results of data science.

CASE FOR DATA SCIENCE

- In the past few decades, a massive accumulation of data has been seen with the advancement of information technology, connected networks, and the businesses it enables.
- This trend is also coupled with a steep decline in data storage and data processing costs.
- The applications built on these advancements like online businesses, social networking, and mobile technologies unleash a large amount of complex, heterogeneous data that are waiting to be analyzed.

CASE FOR DATA SCIENCE (CONTINUED)

- Traditional analysis techniques like dimensional slicing, hypothesis testing, and descriptive statistics can only go so far in information discovery.
- A paradigm is needed to manage the massive volume of data, explore the inter-relationships of thousands of variables, and deploy machine learning algorithms to deduce optimal insights from datasets.

CASE FOR DATA SCIENCE (CONTINUED)

- A set of frameworks, tools, and techniques are needed to intelligently assist humans to process all these data and extract valuable information.
- Data science is one such paradigm that can handle large volumes with multiple attributes and deploy complex algorithms to search for patterns from data.
- Each key motivation for using data science techniques are explored here.

1. VOLUME

- The sheer volume of data captured by organizations is exponentially increasing.
- The rapid decline in storage costs and advancements in capturing every transaction and event, combined with the business need to extract as much leverage as possible using data, creates a strong motivation to store more data than ever.

1. VOLUME (CONTINUED)

- As data become more granular, the need to use large volume data to extract information increases.
- A rapid increase in the volume of data exposes the limitations of current analysis methodologies.
- In a few implementations, the time to create generalization models is critical and data volume plays a major part in determining the time frame of development and deployment.

2. DIMENSIONS

- The three characteristics of the Big Data phenomenon are **high volume, high velocity, and high variety.**
- The variety of data relates to the multiple types of values (numerical, categorical), formats of data (audio files, video files), and the application of the data (location coordinates, graph data).
- Every single record or data point contains multiple attributes or variables to provide context for the record.

2. DIMENSIONS (CONTINUED)

- For example, every user record of an ecommerce site can contain attributes such as products viewed, products purchased, user demographics, frequency of purchase, clickstream, etc.
- Determining the most effective offer for an ecommerce user can involve computing information across these attributes.
- Each attribute can be thought of as a dimension in the dataspace.

2. DIMENSIONS (CONTINUED)

- The user record has multiple attributes and can be visualized in multidimensional space.
- The addition of each dimension increases the complexity of analysis techniques.
- A simple linear regression model that has one input dimension is relatively easy to build compared to multiple linear regression models with multiple dimensions.

2. DIMENSIONS (CONTINUED)

- As the dimensional space of data increase, a scalable framework that can work well with multiple data types and multiple attributes is needed.
- In the case of text mining, a document or article becomes a data point with each unique word as a dimension.
- Text mining yields a dataset where the number of attributes can range from a few hundred to hundreds of thousands of attributes.

3. COMPLEX QUESTIONS

- As more complex data are available for analysis, the complexity of information that needs to get extracted from data is increasing as well.
- If the natural clusters in a dataset, with hundreds of dimensions, need to be found, then traditional analysis like hypothesis testing techniques cannot be used in a scalable fashion.
- The machine-learning algorithms need to be leveraged in order to automate searching in the vast search space.

3. COMPLEX QUESTIONS (CONTINUED)

- Traditional statistical analysis approaches the data analysis problem by assuming a stochastic model, in order to predict a response variable based on a set of input variables.
- A linear regression is a classic example of this technique where the parameters of the model are estimated from the data.
- These hypothesis-driven techniques were highly successful in modeling simple relationships between response and input variables.
- However, there is a significant need to extract nuggets of information from large, complex datasets, where the use of traditional statistical data analysis techniques is limited.

3. COMPLEX QUESTIONS (CONTINUED)

- Machine learning approaches the problem of modeling by trying to find an algorithmic model that can better predict the output from input variables.
- The algorithms are usually recursive and, in each cycle, estimate the output and “learn” from the predictive errors of the previous steps.
- This route of modeling greatly assists in exploratory analysis since the approach here is not validating a hypothesis but generating a multitude of hypotheses for a given problem.
- In the context of the data problems faced today, both techniques need to be deployed.
- John Tuckey, in his article “We need both exploratory and confirmatory,” stresses the importance of both exploratory and confirmatory analysis techniques.

DATA SCIENCE CLASSIFICATION

- Data science problems can be broadly categorized into supervised or unsupervised learning models.
- **Supervised or directed data science** tries to infer a function or relationship based on labeled training data and uses this function to map new unlabeled data.
- Supervised techniques predict the value of the output variables based on a set of input variables.
- To do this, a model is developed from a training dataset where the values of input and output are previously known.

DATA SCIENCE CLASSIFICATION (CONTINUED)

- The model generalizes the relationship between the input and output variables and uses it to predict for a dataset where only input variables are known.
- The output variable that is being predicted is also called a class label or target variable.
- Supervised data science needs a sufficient number of labeled records to learn the model from the data.

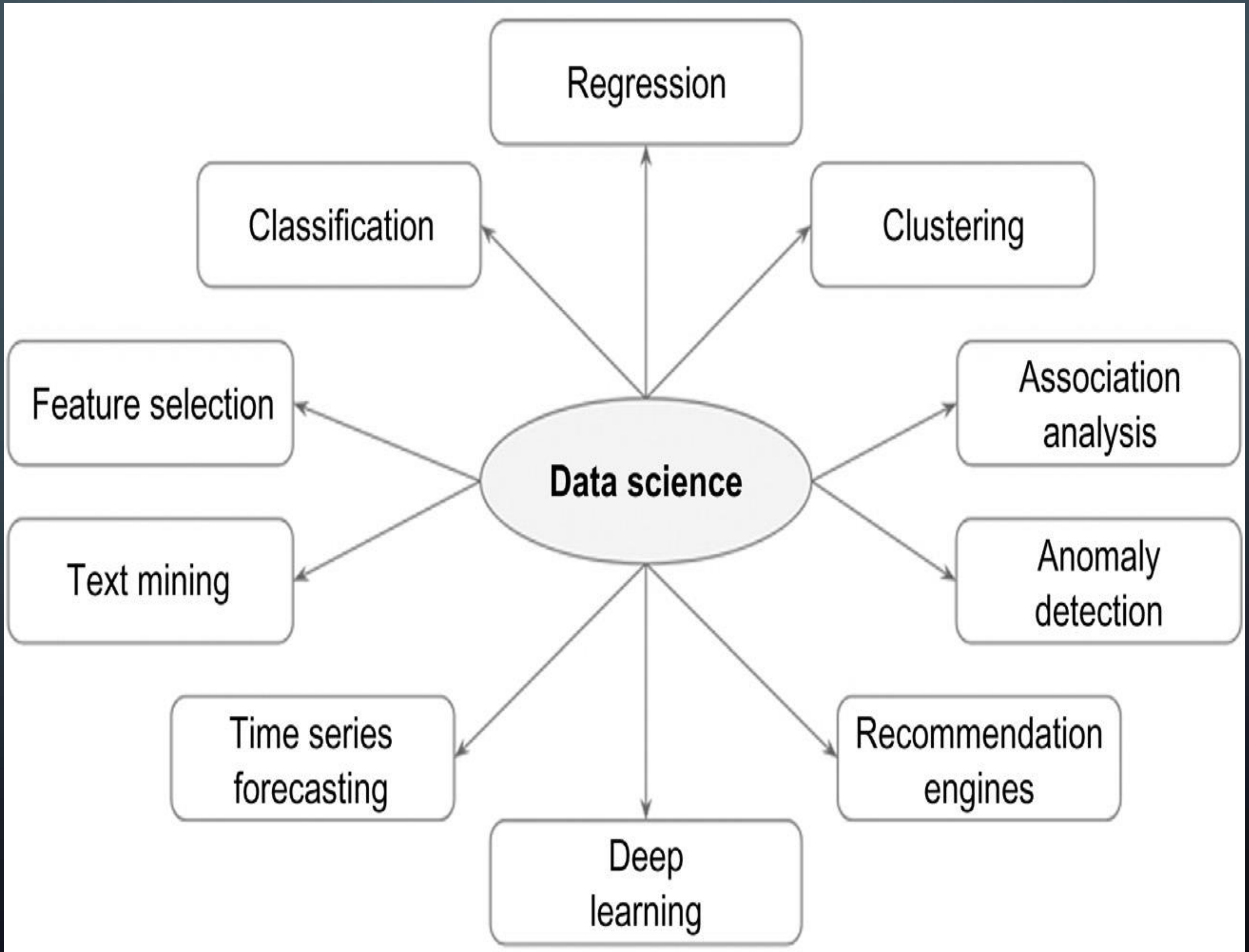
DATA SCIENCE CLASSIFICATION (CONTINUED)

- **Unsupervised or undirected data science** uncovers hidden patterns in unlabeled data.
- In unsupervised data science, there are no output variables to predict.
- The objective of this class of data science techniques, is to find patterns in data based on the relationship between data points themselves.
- An application can employ both supervised and unsupervised learners.

DATA SCIENCE CLASSIFICATION (CONTINUED)

- Data science problems can also be classified into tasks such as: classification, regression, association analysis, clustering, anomaly detection, recommendation engines, feature selection, time series forecasting, deep learning, and text mining.

DATA SCIENCE TASKS



DATA SCIENCE CLASSIFICATION (CONTINUED)

- **Classification and regression techniques** predict a target variable based on input variables.
- The prediction is based on a generalized model built from a previously known dataset.
- In regression tasks, the output variable is numeric (e.g., the mortgage interest rate on a loan).
- Classification tasks predict output variables, which are categorical or polynomial (e.g., the yes or no decision to approve a loan).

DATA SCIENCE CLASSIFICATION (CONTINUED)

- **Deep learning** is a more sophisticated artificial neural network that is increasingly used for classification and regression problems.
- **Clustering** is the process of identifying the natural groupings in a dataset.
- For example, clustering is helpful in finding natural clusters in customer datasets, which can be used for market segmentation.
- Since this is unsupervised data science, it is up to the end user to investigate why these clusters are formed in the data and generalize the uniqueness of each cluster.

DATA SCIENCE CLASSIFICATION (CONTINUED)

- In retail analytics, it is common to identify pairs of items that are purchased together, so that specific items can be bundled or placed next to each other.
- This task is called **market basket analysis or association analysis**, which is commonly used in cross selling.
- **Recommendation engines** are the systems that recommend items to the users based on individual user preference.

DATA SCIENCE CLASSIFICATION (CONTINUED)

- **Anomaly or outlier detection** identifies the data points that are significantly different from other data points in a dataset.
- Credit card transaction fraud detection is one of the most prolific applications of anomaly detection.
- **Time series forecasting** is the process of predicting the future value of a variable (e.g., temperature) based on past historical values that may exhibit a trend and seasonality.

DATA SCIENCE CLASSIFICATION (CONTINUED)

- **Text mining** is a data science application where the input data is text, which can be in the form of documents, messages, emails, or web pages.
- To aid the data science on text data, the text files are first converted into document vectors where each unique word is an attribute.
- Once the text file is converted to document vectors, standard data science tasks such as classification, clustering, etc., can be applied.
- **Feature selection** is a process in which attributes in a dataset are reduced to a few attributes that really matter.

DATA SCIENCE CLASSIFICATION (CONTINUED)

- A complete data science application can contain elements of both supervised and unsupervised techniques.
- Unsupervised techniques provide an increased understanding of the dataset and hence, are sometimes called **descriptive data science**.
- As an example of how both unsupervised and supervised data science can be combined in an application, consider the following scenario.
- In marketing analytics, clustering can be used to find the natural clusters in customer records.
- Each customer is assigned a cluster label at the end of the clustering process.
- A labeled customer dataset can now be used to develop a model that assigns a cluster label for any new customer record with a supervised classification technique.

DATA SCIENCE ALGORITHMS

- An algorithm is a logical step-by-step procedure for solving a problem.
- In data science, it is the blueprint for how a particular data problem is solved.
- Many of the learning algorithms are recursive, where a set of steps are repeated many times until a limiting condition is met.
- Some algorithms also contain a random variable as an input and are aptly called **randomized algorithms**.

DATA SCIENCE ALGORITHMS (CONTINUED)

- A classification task can be solved using many different learning algorithms such as decision trees, artificial neural networks, k-NN, and even some regression algorithms.
- The choice of which algorithm to use depends on the type of dataset, objective, structure of the data, presence of outliers, available computational power, number of records, number of attributes, and so on.
- It is up to the data science practitioner to decide which algorithm (s) to use by evaluating the performance of multiple algorithms.

DATA SCIENCE ALGORITHMS (CONTINUED)

- There have been hundreds of algorithms developed in the last few decades to solve data science problems.
- Data science algorithms can be implemented by custom-developed computer programs in almost any computer language.
- This obviously is a time consuming task.
- In order to focus the appropriate amount of time on data and algorithms, data science tools or statistical programing tools, like R, RapidMiner, Python, SAS Enterprise Miner, etc., which can implement these algorithms with ease, can be leveraged.

DATA SCIENCE ALGORITHMS (CONTINUED)

- These data science tools offer a library of algorithms as functions, which can be interfaced through programming code or configurated through graphical user interfaces.
- The following Table provides a summary of data science tasks with commonly used algorithmic techniques and example cases.

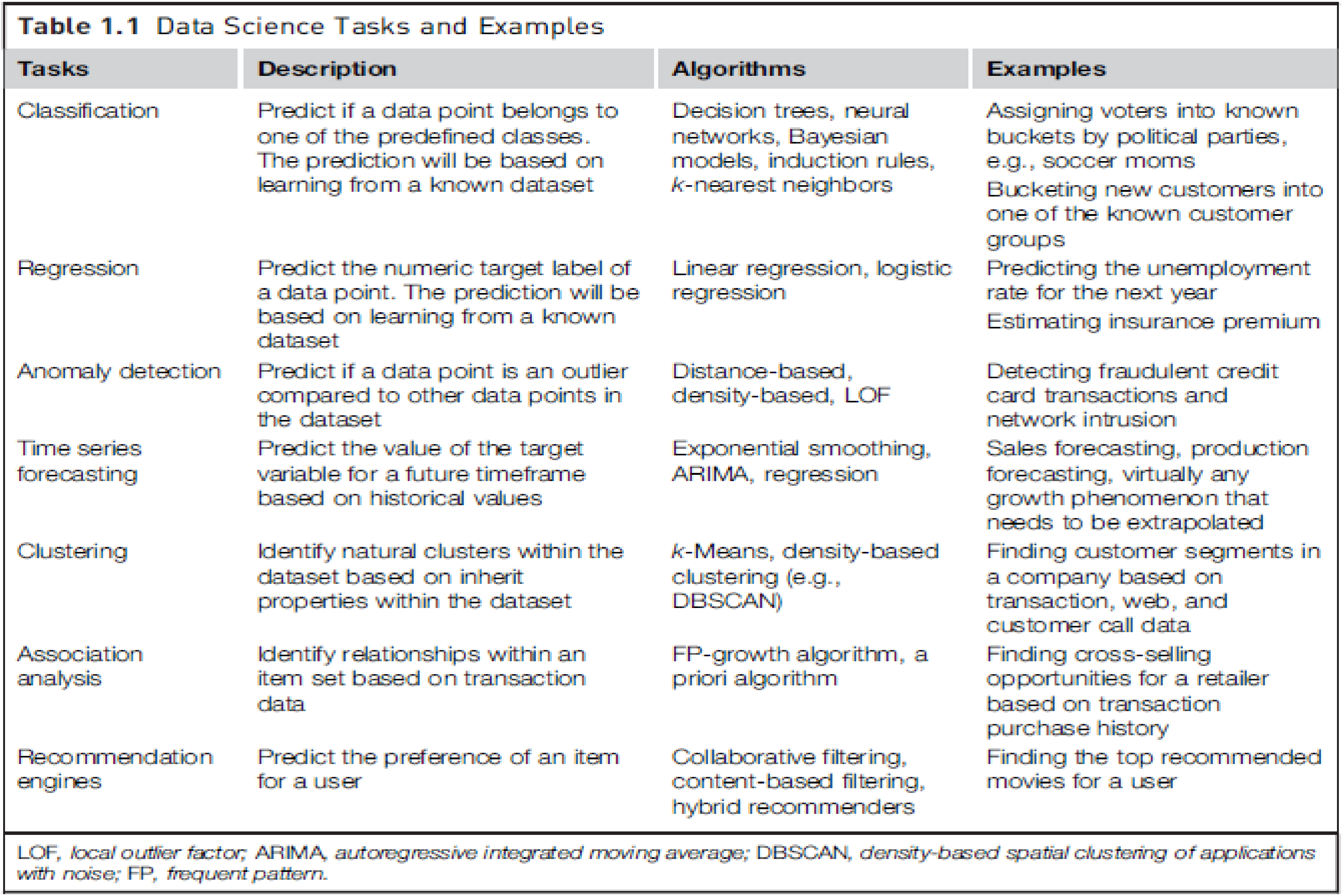


Table 1.1 Data Science Tasks and Examples

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 dataset	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 dataset	Linear regression, logistic regression	Predicting the unemployment rate for the next year Estimating insurance premium
Anomaly detection	Predict if a data point is an outlier compared to other data points in the dataset	Distance-based, density-based, LOF	Detecting fraudulent credit card transactions and network intrusion
Time series forecasting	Predict the value of the target variable for a future timeframe based on historical values	Exponential smoothing, ARIMA, regression	Sales forecasting, production forecasting, virtually any growth phenomenon that needs to be extrapolated
Clustering	Identify natural clusters within the dataset based on inherit properties within the dataset	*k*-Means, density-based clustering (e.g., 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	FP-growth algorithm, a priori algorithm	Finding cross-selling opportunities for a retailer based on transaction purchase history
Recommendation engines	Predict the preference of an item for a user	Collaborative filtering, content-based filtering, hybrid recommenders	Finding the top recommended movies for a user

LOF, *local outlier factor;* ARIMA, *autoregressive integrated moving average;* DBSCAN, *density-based spatial clustering of applications with noise;* FP, *frequent pattern.*

CORE ALGORITHMS

- **Classification** is the most widely used data science task in business.
- The objective of a classification model is to predict a target variable that is binary (e.g., a loan decision) or categorical (e.g., a customer type) when a set of input variables are given.
- The model does this by learning the generalized relationship between the predicted target variable with all other input attributes from a known dataset.
- Each algorithm differs by how the relationship is extracted from the known training dataset.

DECISION TREES

- **Decision trees** approach the classification problem by partitioning the data into purer subsets based on the values of the input attributes.
- The attributes that help achieve the cleanest levels of such separation are considered significant in their influence on the target variable and end up at the root and closer-to-root levels of the tree.
- The output model is a tree framework than can be used for the prediction of new unlabeled data.

RULE INDUCTION

- **Rule induction** is a data science process of deducing “if-then” rules from a dataset or from the decision trees.
- These symbolic decision rules explain an inherent relationship between the input attributes and the target labels in the dataset that can be easily understood by anyone.

NAÏVE BAYESIAN

- **Naïve Bayesian** algorithms provide a probabilistic way of building a model.
- This approach calculates the probability for each value of the class variable for given values of input variables.
- With the help of conditional probabilities, for a given unseen record, the model calculates the outcome of all values of target classes and comes up with a predicted winner.

K-NN ALGORITHM

- Why go through the trouble of extracting complex relationships from the data when the entire training dataset can be memorized and the relationship can appear to have been generalized?
- This is exactly what the **k-NN algorithm** does, and it is, therefore, called a **"lazy" learner** where the entire training dataset is memorized as the model.

ARTIFICIAL NEURAL NETWORKS

- Neurons are the nerve cells that connect with each other to form a biological neural network in our brain.
- The working of these interconnected nerve cells inspired the approach of some complex data problems by the creation of **artificial neural networks**.
- The neural networks section provides a conceptual background of how a simple neural network works and how to implement one for any general prediction problem.
- Later on this is extended to deep neural networks which have revolutionized the field of artificial intelligence.

SUPPORT VECTOR MACHINES (SVMS)

- **Support vector machines (SVMs)** were developed to address optical character recognition problems: how can an algorithm be trained to detect boundaries between different patterns, and thus, identify characters?
- SVMs can, therefore, identify if a given data sample belongs within a boundary (in a particular class) or outside it (not in the class).

ENSEMBLE LEARNERS

- **Ensemble learners** are “meta” models where the model is a combination of several different individual models.
- If certain conditions are met, ensemble learners can gain from the wisdom of crowds and greatly reduce the generalization error in data science.

REGRESSION METHODS

- The simple mathematical equation y=ax+b is a linear regression model.
- Regression Methods, describes a class of data science techniques in which the target variable (e.g., interest rate or a target class) is functionally related to input variables.

LINEAR REGRESSION

- The simplest of all function fitting models is based on a linear equation, as previously mentioned.
- Polynomial regression uses higher-order equations.
- No matter what type of equation is used, the goal is to represent the variable to be predicted in terms of other variables or attributes.
- Further, the predicted variable and the independent variables all have to be numeric for this to work.

LOGISTIC REGRESSION

- Addresses the issue of predicting a target variable that may be binary or binomial (such as 1 or 0, yes or no) using predictors or attributes, which may be numeric.

ASSOCIATION ANALYSIS

- Supervised data science or directed data science predict the value of the target variables.
- Two important unsupervised data science tasks: Association Analysis and Clustering.
- Ever heard of the beer and diaper association in supermarkets?
- Apparently, a supermarket discovered that customers who buy diapers also tend to buy beer.
- While this may have been an urban legend, the observation has become a poster child for association analysis.

ASSOCIATION ANALYSIS (CONTINUED)

- Associating an item in a transaction with another item in the transaction to determine the most frequently occurring patterns is termed **association analysis**.
- This technique is about, for example, finding relationships between products in a supermarket based on purchase data, or finding related web pages in a website based on clickstream data.
- It is widely used in retail, ecommerce, and media to creatively bundle products.

CLUSTERING

- **Clustering** is the data science task of identifying natural groups in the data.
- As an unsupervised task, there is no target class variable to predict.
- After the clustering is performed, each record in the dataset is associated with one or more cluster.
- Widely used in marketing segmentations and text mining, clustering can be performed by a range of algorithms.

CLUSTERING (CONTINUED)

- The **k-means clustering** technique identifies a cluster based on a central prototype record.
- **DBSCAN clustering** partitions the data based on variation in the density of records in a dataset.
- **Self-organizing maps** create a two dimensional grid where all the records related with each other are placed next to each other.