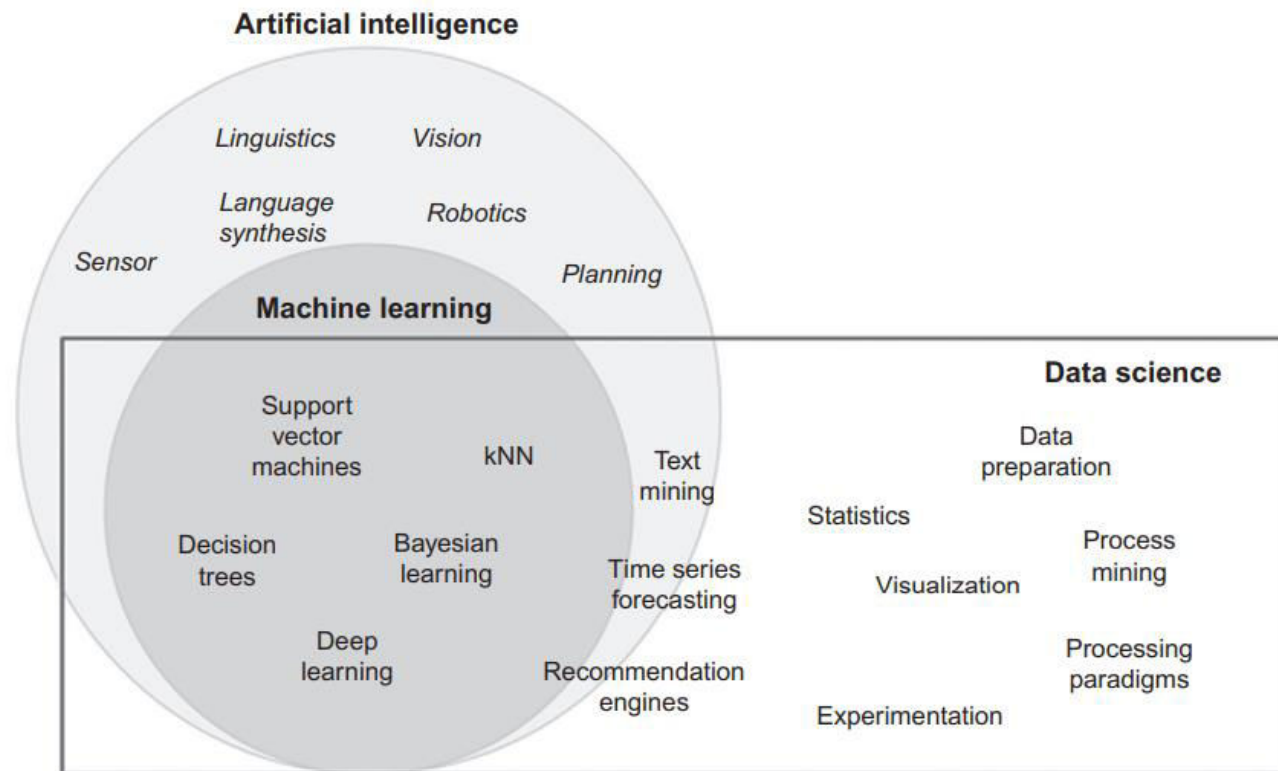# MODULE 1

Syllabus

Introduction to data science, Data science classification, Data science process - Prior knowledge, Data preparation, Modelling, Application, Data exploration - Data sets, Descriptive statistics for univariate and multivariate data

Data visualisation – Histogram, Quartile plot, Distribution chart, Scatter plot, Bubble chart, Density chart
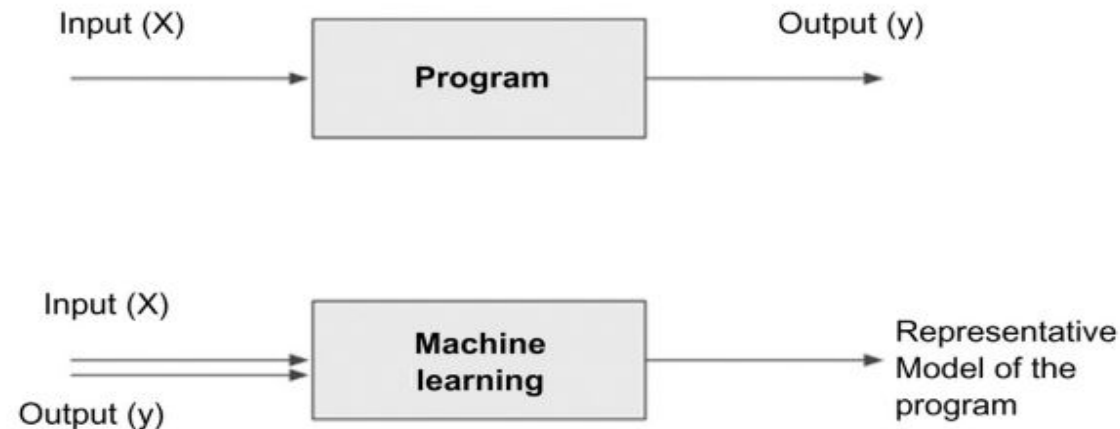
INTRODUCTION

Data Science

- Data science is a collection of techniques used to extract value from data.

- Data science techniques rely on finding useful patterns, connections, and relationships within data.

- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining.

- 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.

- Artificial intelligence, Machine learning, and data science are all related to each other.

Artificial intelligence

Linguistics *Vision*

Language synthesis *Robotics*

Sensor *Planning*

Machine learning

Data science

Support vector machines kNN Text mining

Statistics

Data preparation

Decision trees Bayesian learning Time series forecasting

Visualization

Process mining

Deep learning Recommendation engines

Experimentation

Processing paradigms

FIGURE 1.1

Artificial intelligence, machine learning, and data science.

- Artificial intelligence is about giving machines the capability of mimicking human behaviour. Examples would be: facial recognition, automated driving etc.

- 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 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.

Input (X) → **Program** → Output (y)

Input (X)
Output (y) → **Machine learning** → Representative Model of the program

FIGURE 1.2
Traditional program and machine learning.
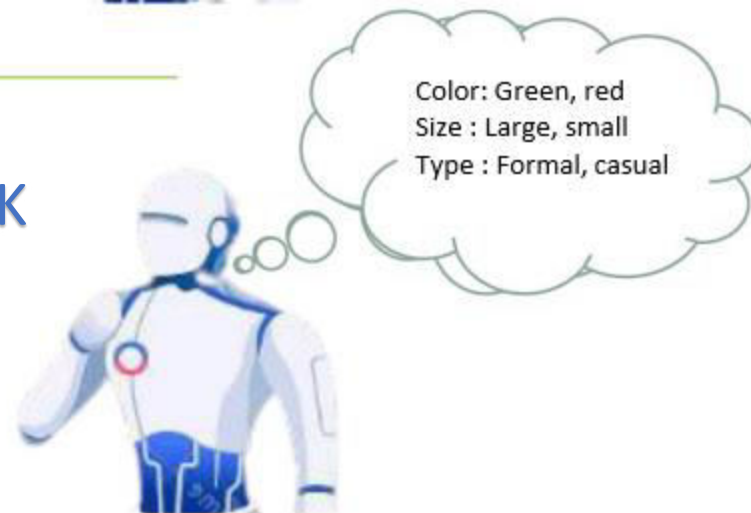
Data science vs. AI vs. ML

Data Science

- based on strict analytical evidence
- deals with structured & unstructured data
- includes various data operations

Artificial Intelligence

- imparts human intellect to machines
- uses logic and decision trees
- includes machine learning

Machine Learning

- subset of AI
- uses statistical models
- machines improve with experience

What is Data Science

- Data science is the business application of machine learning, artificial intelligence, and other quantitative fields like statistics, visualization, and mathematics.

- 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 or predict revenue for the next quarter.

- 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.

Key features and motivations:-

1. *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.

The ultimate objective of data science is to find potentially useful conclusions that can be acted upon by the users of the analysis.

2.Building Representative Models

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.
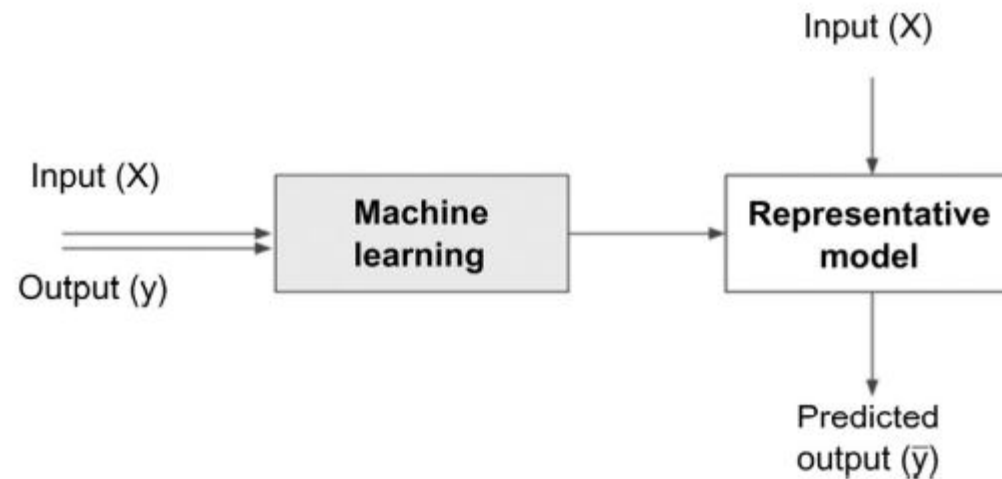
3. Combination of Statistics, Machine Learning, and Computing

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.

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.

4. Learning Algorithms

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.

Each data science task uses specific learning algorithms like decision trees, neural networks, k-nearest neighbors (k-NN), and k-means clustering, etc

5. Associated Fields

- 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.

- 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.
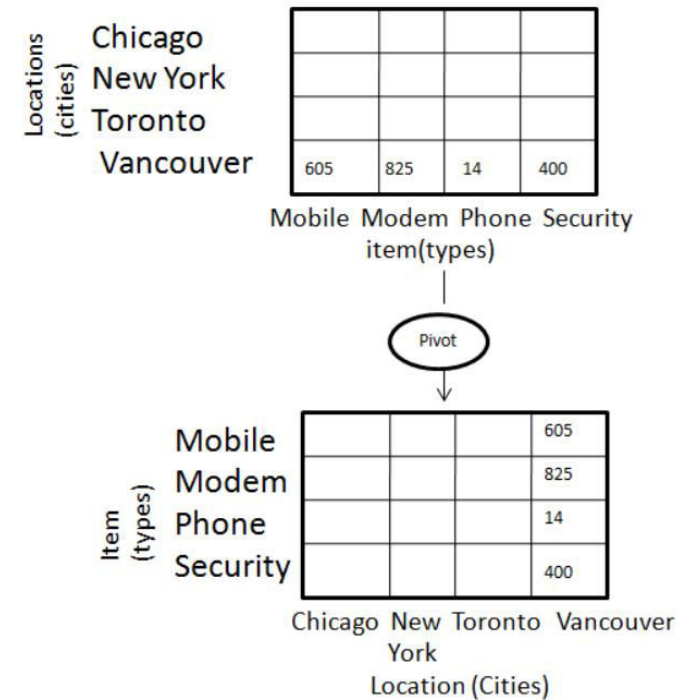
- Dimensional slicing: Online analytical processing (OLAP) applications, mainly provide information on the data through dimensional slicing, filtering, and pivoting

slicing
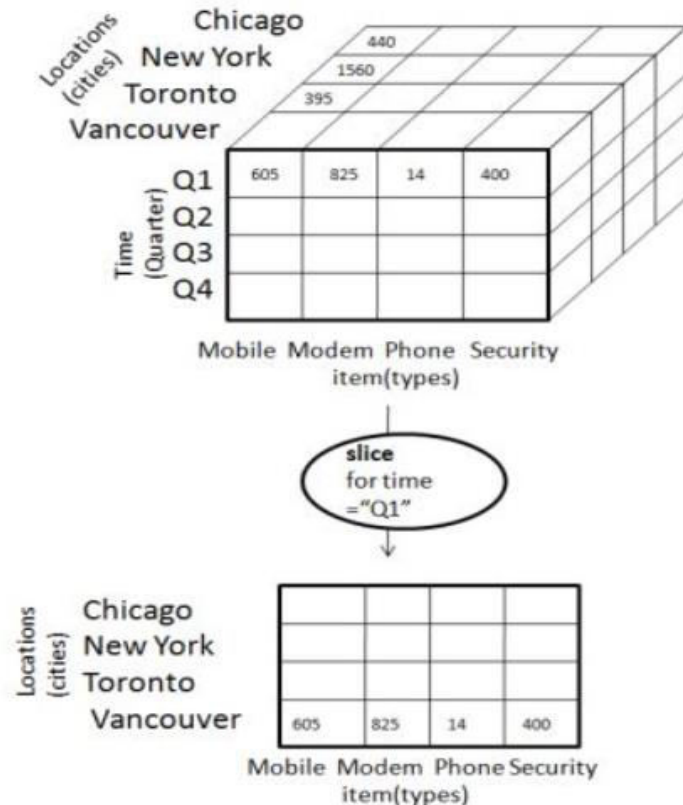
pivoting

- Hypothesis testing: It is a statement or assumption that is either true or false. In general, data science is a process where many hypotheses are generated and tested based on observational data.

- Data engineering: Data engineering is the process of sourcing, organizing, assembling, storing, and distributing data for effective analysis and usage.

- Business intelligence: Business intelligence helps organizations consume data effectively. Business intelligence (BI) refers to the procedural and technical infrastructure that collects, stores, and analyzes the data produced by a company's activities.
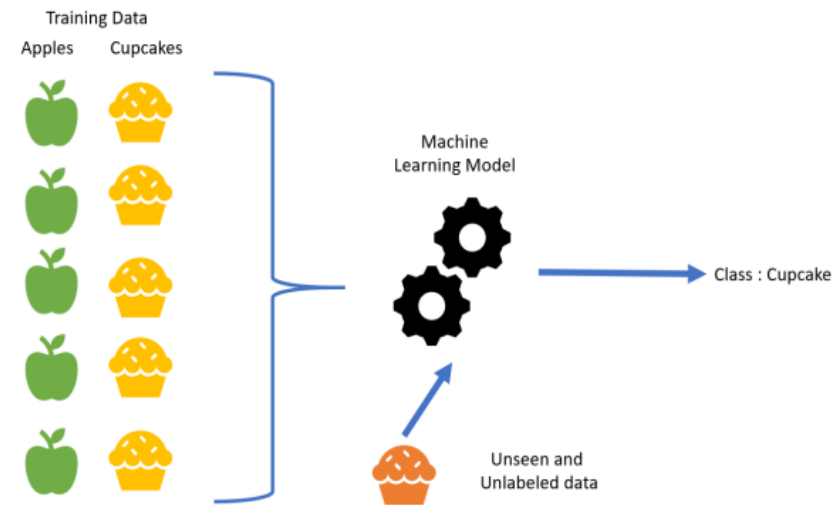
DATA SCIENCE CLASSIFICATION

- Data science problems can be broadly categorized into

    supervised or unsupervised learning models.

Supervised 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.

- 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

Input : Labeled Data

X (features)	Y (labels)
$x_{11}, x_{12}, x_{13}, \dots\dots\dots\dots\dots x_{1n}$	y_1
.	.
.	.
.	.
$x_{k1}, xk_2, x_{k3}, \dots\dots\dots\dots\dots x_{kn}$	y_k

Training Data

Apples   Cupcakes

Machine
Learning Model

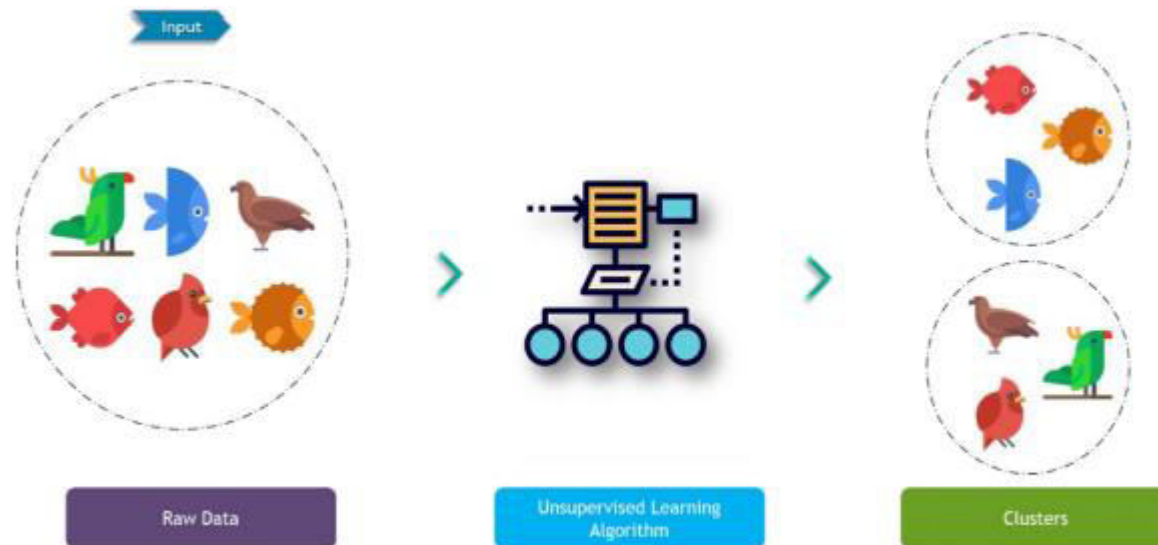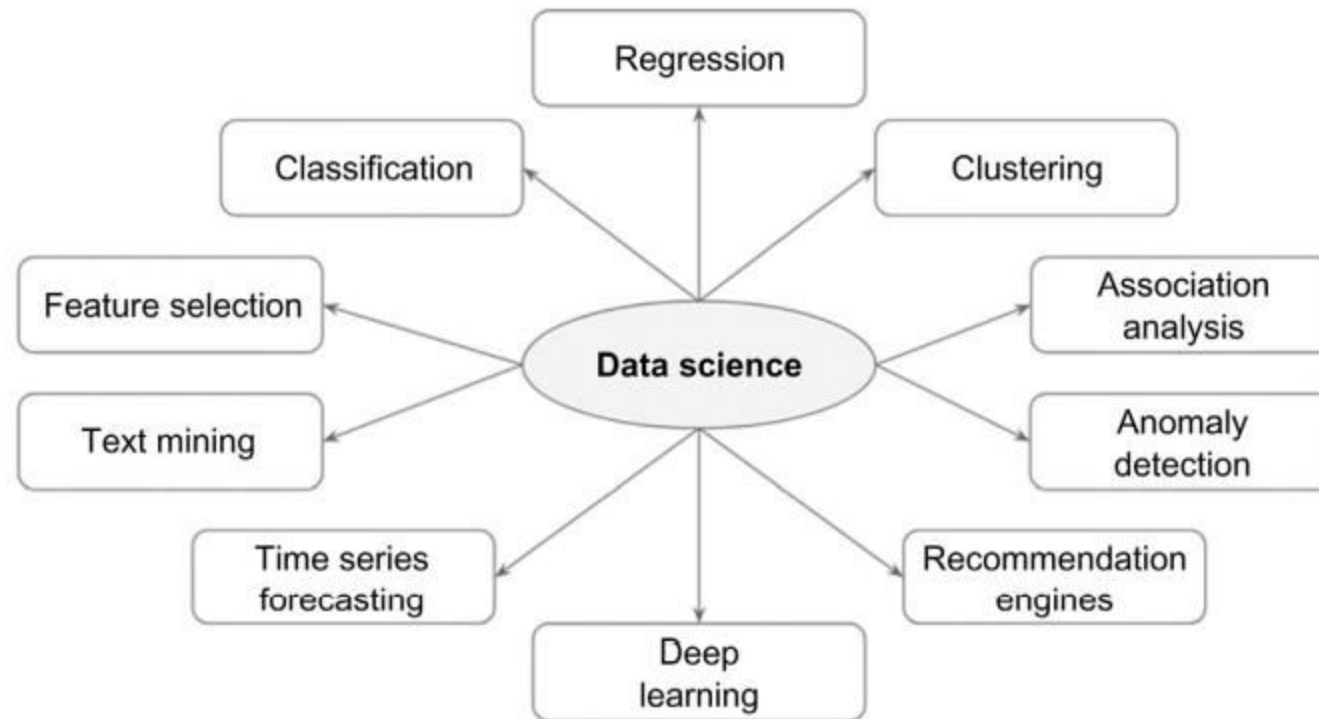Class : Cupcake

Unseen and
Unlabeled data

Unsupervised Models

- 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

Input : Unlabeled Data

X (*features*)
$x_{11}, x_{12}, x_{13}, \ldots \ldots \ldots \ldots \ldots x_{1n}$
.
.
.
$x_{k1}, xk_2, x_{k3}, \ldots \ldots \ldots \ldots \ldots x_{kn}$

Input

Raw Data

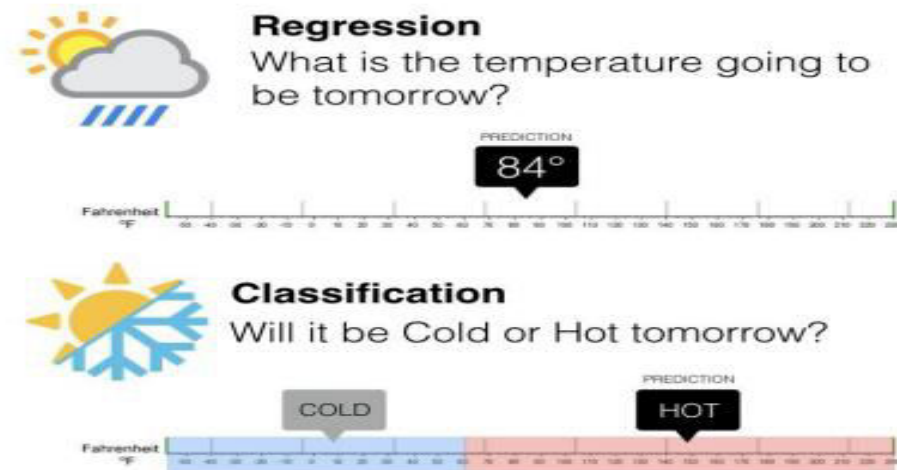Unsupervised Learning Algorithm

Clusters

FIGURE 1.4

Data science tasks.

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 <u>numeric</u>. Classification tasks predict output variables, which are <u>categorical</u>.

Regression
What is the temperature going to be tomorrow?

PREDICTION
84°

Fahrenheit °F

Classification
Will it be Cold or Hot tomorrow?

PREDICTION
COLD   HOT

Fahrenheit °F

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.

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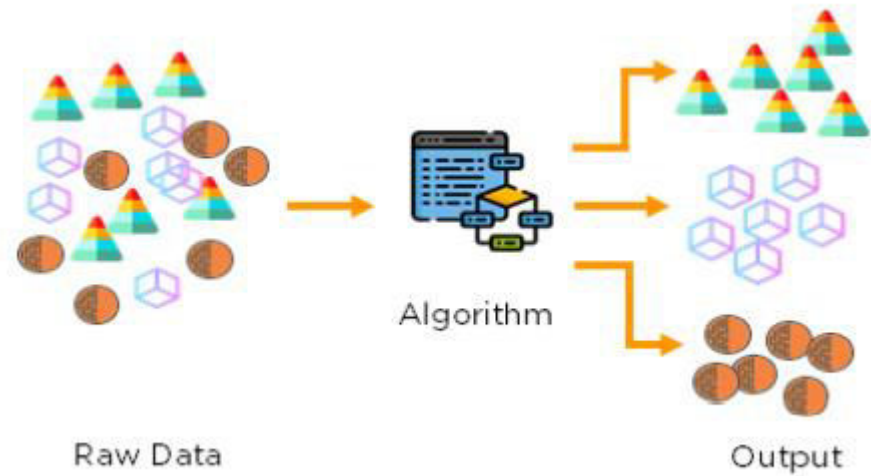
In retail analytics, it is common to identify pairs of items that are purchased together, so that specific items can be bundled or placed next to each other. This task is called *market basket analysis* or *association analysis.*
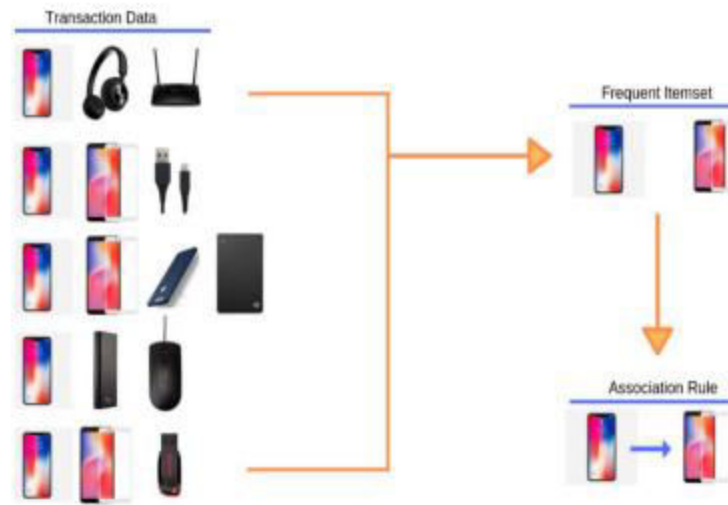
Recommendation engines are the systems that recommend items to the users based on individual user preference.

Anomaly or outlier detection identifies the data points that are significantly different from other data points in a dataset.

OUTLIERS

Time series forecasting is the process of predicting the future value of a variable (e.g., temperature) based on past historical values.

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.

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, 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*.

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

DATA SCIENCE PROCESS

The methodical discovery of useful relationships and patterns in data is enabled by a set of iterative activities collectively known as the *data science process.*
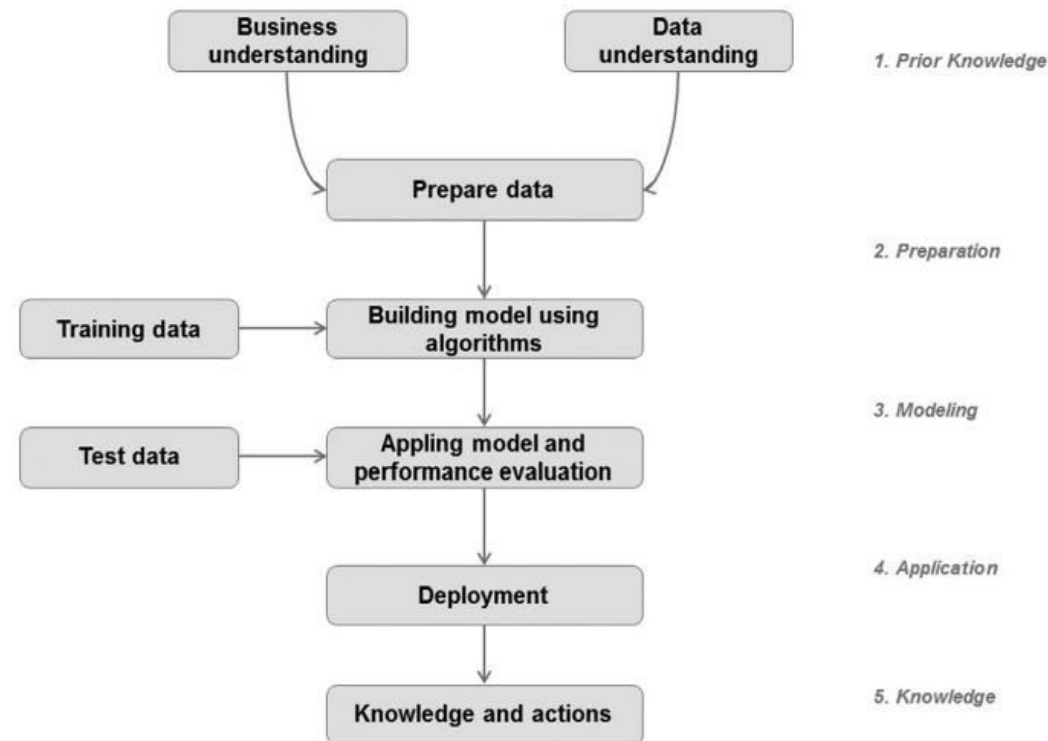
FIGURE 2.2
Data science process.

Prior knowledge refers to information that is already known about a subject.

The prior knowledge step in the data science process helps to define what problem is being solve and what data is needed in order to solve the problem.

Objective:Without a well-defined statement of the problem, it is impossible to come up with the right dataset and pick the right data science algorithm.

Subject Area:The process of data science uncovers hidden patterns in the dataset by exposing relationships between attributes. But the problem is that it uncovers a lot of patterns. The false signals are a major problem in the data science process. . Hence, it is essential to know the subject matter, the context, and the business process generating the data.

Data:Similar to the prior knowledge in the subject area, prior knowledge in the data can also be gathered. Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process.

- A dataset is a collection of data with a defined structure.

This structure is also sometimes referred to as a "data frame".

● A data point (record, object) is a single instance in the dataset. Each instance contains the same structure as the dataset.

● An attribute (feature, input, dimension, variable, or predictor) is a single property of the dataset. Attributes can be numeric, categorical, date-time, text, or Boolean data types.

- A label (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes. In Table interest rate is the output variable.

● Identifiers are special attributes that are used for locating or providing context to individual records. For example, common attributes like names, account numbers, and employee ID numbers are identifier attributes. Identifiers are often used as lookup keys to join multiple datasets

Table 2.1 Dataset

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

Table 2.2 New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

Causation Versus Correlation: Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables.

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.
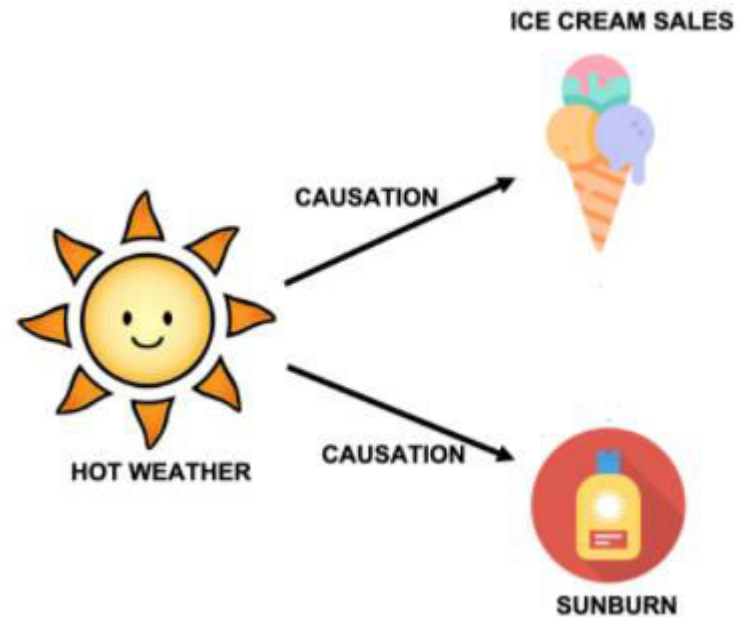
DATA PREPARATION

Preparing the dataset to suit a data science task is the most time-consuming part of the process. It is extremely rare that datasets are available in the form required by the data science algorithms.

1.Data Exploration:Data exploration, also known as exploratory data analysis, provides a set of simple tools to achieve basic understanding of the data. Data exploration approaches involve computing descriptive statistics and visualization of data.
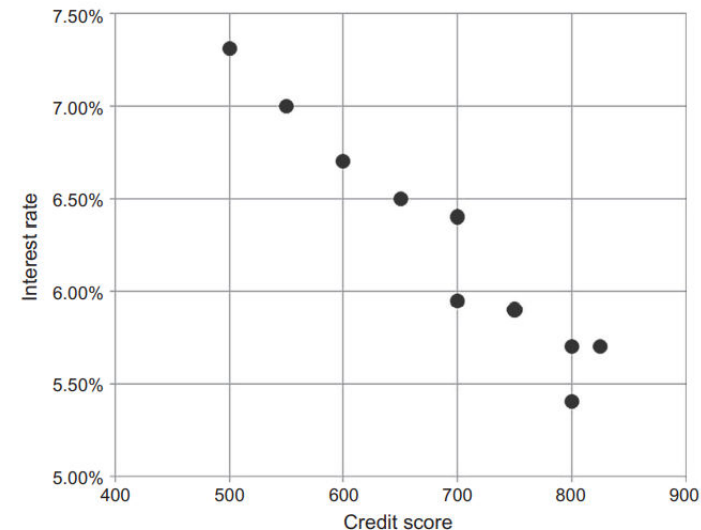
FIGURE 2.3

Scatterplot for interest rate dataset.

2.Data Quality:Data quality is an ongoing concern wherever data is collected, processed, and stored.

Organizations use data alerts, cleansing, and transformation techniques to improve and manage the quality of the data and store them in companywide repositories called <u>data warehouses.</u>

Data sourced from well-maintained data warehouses have higher quality, as there are proper controls in place to ensure a level of data accuracy for new and existing data.

The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bound, substitution of missing values, etc.
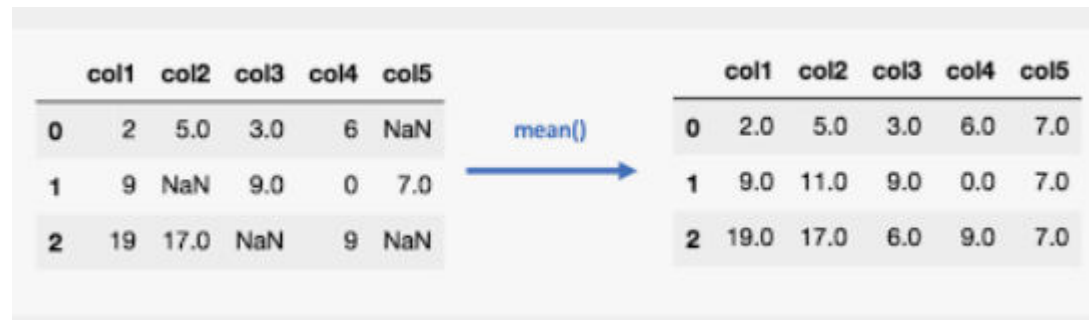
3. Missing Values

One of the most common data quality issues is that some records have missing attribute values. For example, a credit score may be missing in one of the records.

The first step of managing missing values is to understand the reason behind why the values are missing.

The missing value can be substituted with a range of artificial data so that the issue can be managed.

Missing credit score values can be replaced with a credit score derived from the dataset (mean, minimum, or maximum value, depending on the characteristics of the attribute).

	col1	col2	col3	col4	col5			col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN	mean()	0	2.0	5.0	3.0	6.0	7.0
1	9	NaN	9.0	0	7.0		1	9.0	11.0	9.0	0.0	7.0
2	19	17.0	NaN	9	NaN		2	19.0	17.0	6.0	9.0	7.0

4. Data Types and Conversion

The attributes in a dataset can be of different types, such as continuous numeric (interest rate), integer numeric (credit score), or categorical.

For example, the credit score can be expressed as categorical values (poor, good, excellent) or numeric score. Different data science algorithms impose different restrictions on the attribute data types.

If the available data are categorical, they must be converted to continuous numeric attribute. A specific numeric score can be encoded for each category value, such as poor=400, good=600, excellent=700, etc.

Numeric values can be converted to categorical data types by a technique called binning

5.Transformation

In some data science algorithms , the input attributes are expected to be numeric and normalized.

Normalization prevents one attribute dominating the results because of large values.

For example, consider income (expressed in USD, in thousands) and credit score (in hundreds). One solution is to convert the range of income and credit score to a more uniform scale from 0 to 1 by normalization.

6. Outliers

Outliers are anomalies in a given dataset. Outliers may occur because of correct data capture (few people with income in tens of millions) or erroneous data capture (human height as 1.73 cm instead of 1.73 m). Presence of outliers needs to be require special treatments.

Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

7.Feature Selection

Many data science problems involve a dataset with hundreds to thousands of attributes.

Not all the attributes are equally important or useful in predicting the target.

Some of the attributes may be highly correlated with each other, like annual income and taxes paid.

A large number of attributes in the dataset significantly increases the complexity of a model and may degrade the performance of the model.

Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection.

8.Data Sampling

Sampling is a process of selecting a subset of records as a representation of the original dataset for use in data analysis or modeling. The sample data serve as a representative of the original dataset with similar properties. Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling.

In the build process for data science applications, it is necessary to segment the datasets into training and test samples. The training dataset is sampled from the original dataset using simple sampling.
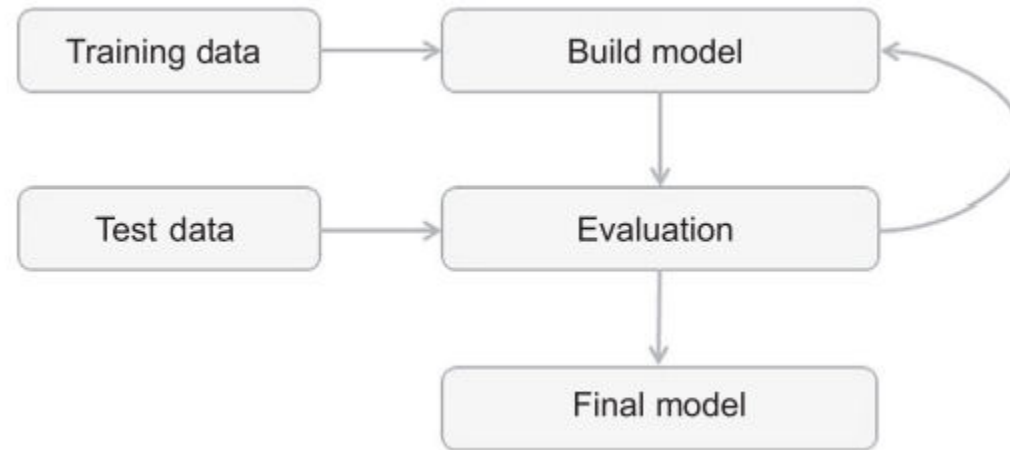
Stratified sampling is a process of sampling where each class is equally represented in the sample.

Sampling

SAMPLING
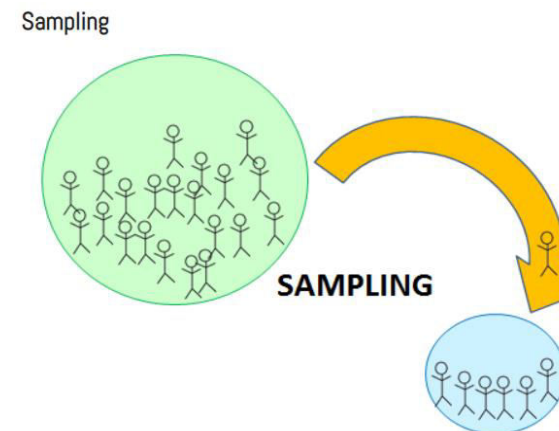
Modelling

A model is the abstract representation of the data and the relationships in a given dataset.

1. Training and Testing Datasets

The dataset used to create the model, with known attributes and target, is called the <u>training dataset</u>.

The validity of the created model will also need to be checked with another known dataset called the <u>test dataset or validation dataset</u>.

To facilitate this process, the overall known dataset can be split into a training dataset and a test dataset.

A standard rule of thumb is two-thirds of the data are to be used as training and one-third as a test dataset

Table 2.3 Training Dataset

Borrower	Credit Score (X)	Interest Rate (Y) (%)
01	500	7.31
02	600	6.70
03	700	5.95
05	800	5.40
06	800	5.70
08	550	7.00
09	650	6.50

Table 2.4 Test Dataset

Borrower	Credit Score (X)	Interest Rate (Y)
04	700	6.40
07	750	5.90
10	825	5.70

2.Learning Algorithms

The business question and the availability of data will dictate what data science task (association, classification, regression, etc.,) can to be used.

The practitioner determines the appropriate data science algorithm within the chosen category.

For example, within a classification task many algorithms can be chosen from: decision trees, rule induction, neural networks, Bayesian models, k-NN, etc.

2. Evaluation of the Model

Model evaluation is used to test the performance of the model.

Training set is used for the purpose of model building and test set is used to test the performance of the model.

The estimation may not be exactly the same as the values in the training records. The phenomenon of a model memorizing the training data is called overfitting.

An overfitted model just memorizes the training records and will underperform on real unlabeled new data.

The model should generalize or learn the relationship between credit score and interest rate. To evaluate this relationship, the validation or test dataset, which was not previously used in building the model, is used for evaluation.

Table 2.5 Evaluation of Test Dataset

Borrower	Credit Score (X)	Interest Rate (Y) (%)	Model Predicted (Y) (%)	Model Error (%)
04	700	6.40	6.11	− 0.29
07	750	5.90	5.81	− 0.09
10	825	5.70	5.37	− 0.33

3.Ensemble Modeling

Ensemble modeling is a process where multiple diverse base models are used to predict an outcome.

The motivation for using ensemble models is to reduce the generalization error of the prediction.

As long as the base models are diverse and independent, the prediction error decreases when the ensemble approach is used.
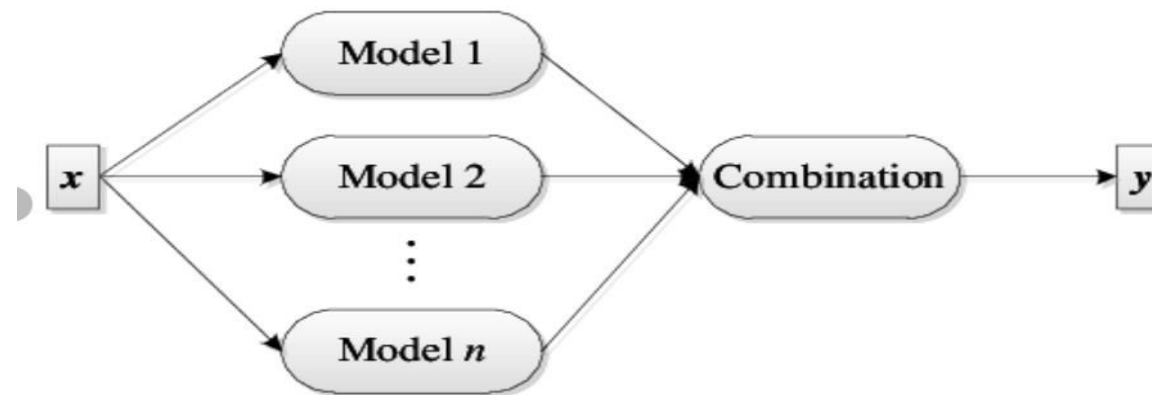
APPLICATION

Deployment is the stage at which the model becomes production ready or live.

The model deployment stage has to deal with: assessing model readiness, technical integration, response time, model maintenance, and assimilation.

1. Production Readiness

The production readiness part of the deployment determines the critical qualities required for the deployment objective.

Consider two business use cases:

i)determining whether a consumer qualifies for a loan- critical quality of this model deployment is real-time prediction

ii)determining the groupings of customers for an enterprise by marketing function- critical quality in this application is the ability to find unique patterns amongst customers, not the response time of the model.

2. Technical Integration

Currently, it is quite common to use data science automation tools or coding using R or Python to develop models.

It is flexible to develop the model with one tool and deploy it in another tool or application.

Some models such as simple regression, decision trees, and induction rules for predictive analytics can be incorporated directly into business applications and business intelligence systems easily.

3.Response Time

Data science algorithms, like k-NN, are easy to build, but quite slow at predicting the unlabeled records.

Algorithms such as the decision tree take time to build but are fast at prediction.

The quality of prediction, accessibility of input data, and the response time of the prediction remain the critical quality factors in business application

4.Model Refresh

It is quite normal that the conditions in which the model is built change after the model is sent to deployment.

For example, the relationship between the credit score and interest rate change frequently based on the prevailing macroeconomic conditions. Hence, the model will have to be refreshed frequently.

The validity of the model can be routinely tested by using the new known test dataset and calculating the prediction error rate.

If the error rate exceeds a particular threshold, then the model has to be refreshed and redeployed.

5.Assimilation

In the descriptive data science applications, deploying a model to live systems may not be the end objective.

The objective may be to assimilate the knowledge gained from the data science analysis to the organization.

For example, the objective may be finding logical clusters in the customer database so that separate marketing approaches can be developed for each customer cluster. Then the next step may be a classification task for new customers to bucket them in one of known clusters.

The association analysis provides a solution for the market basket problem, where the task is to find which two products are purchased together most often.

KNOWLEDGE

The data science process provides a framework to extract nontrivial information from data.

With the advent of massive storage, increased data collection, and advanced computing paradigms, the available datasets to be utilized are only increasing.

To extract knowledge from these massive data assets, advanced approaches need to be employed, like data science algorithms, in addition to standard business intelligence reporting or statistical analysis.

Though many of these algorithms can provide valuable knowledge, it is up to the practitioner to skillfully transform a business problem to a data problem and apply the right algorithm.

Data science, like any other technology, provides various options in terms of algorithms. Using these options to extract the right information from data .