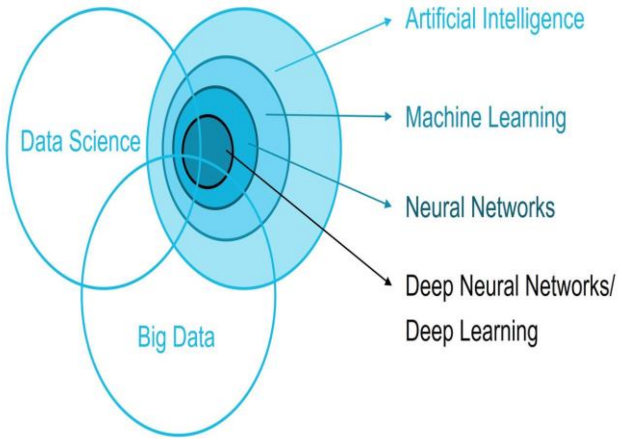
Introduction

Module 1

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.
- Data science is also commonly referred to as knowledge discovery, machine learning, predictive analytics, and data mining; each term has a slightly different meaning depending on the context.

Why Data Science

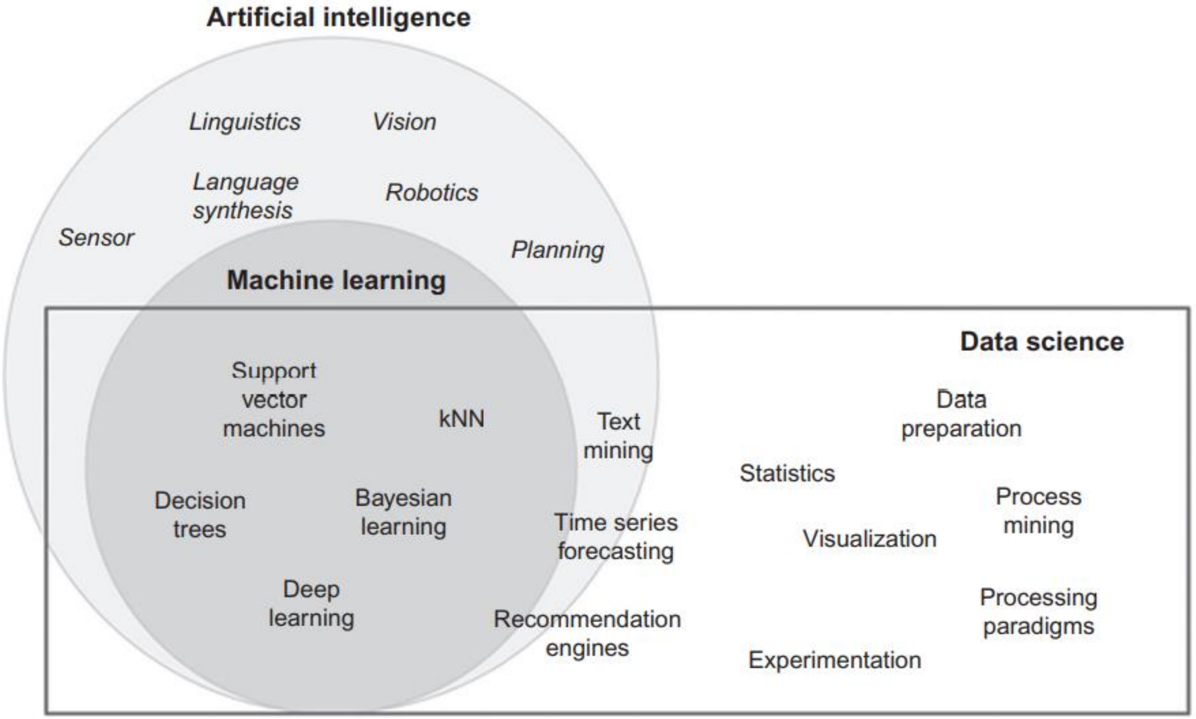
- The technology revolution has brought about the need to process, store, analyze, and comprehend large volumes of diverse data in meaningful ways.
- The scale of data volume and variety places new demands on organizations to quickly uncover hidden relationships and patterns.
- This is where data science techniques have proven to be extremely useful.
- Data science is a compilation of techniques that extract value from data.
- To get meaningful results from any data, a major effort of preparing, cleaning, scrubbing, or standardizing the data is required.

Why Data Science

- The technology revolution has brought about the need to process, store, analyze, and comprehend large volumes of diverse data in meaningful ways.
- The scale of data volume and variety places new demands on organizations to quickly uncover hidden relationships and patterns.
- This is where data science techniques have proven to be extremely useful.
- Data science is a compilation of techniques that extract value from data.
- To get meaningful results from any data, a major effort of preparing, cleaning, scrubbing, or standardizing the data is required.

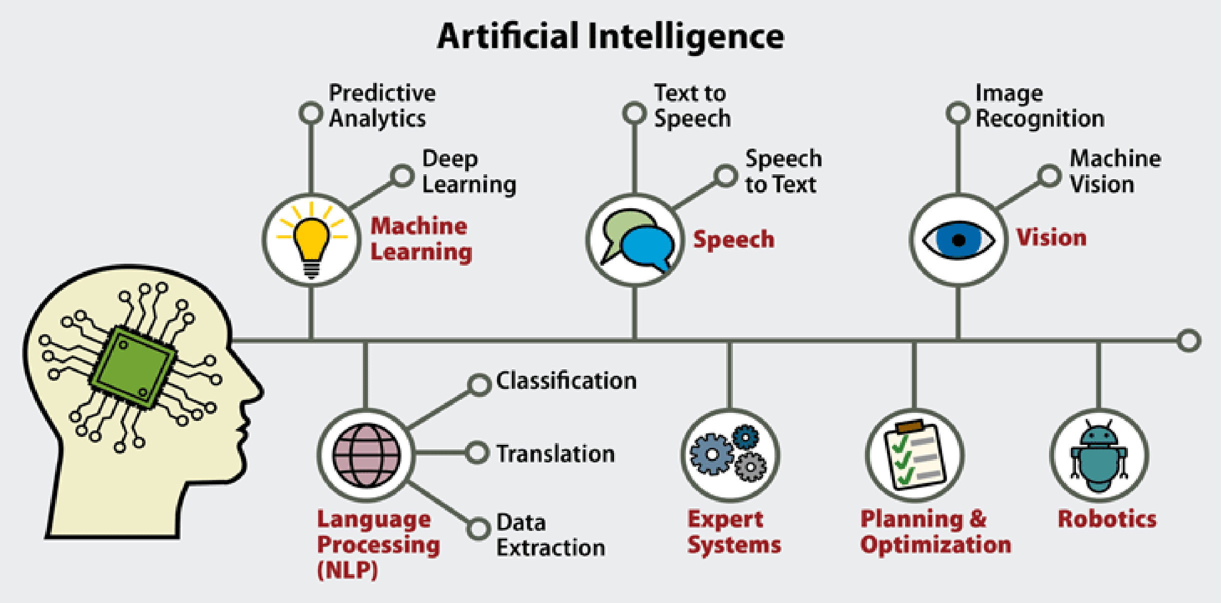
AI, MACHINE LEARNING, AND DATA SCIENCE

- Artificial intelligence, Machine learning, and data science are all related to each other.
- They are often used interchangeably and conflated with each other in popular media and business communication.
- However, all of these three fields are distinct depending on the context.



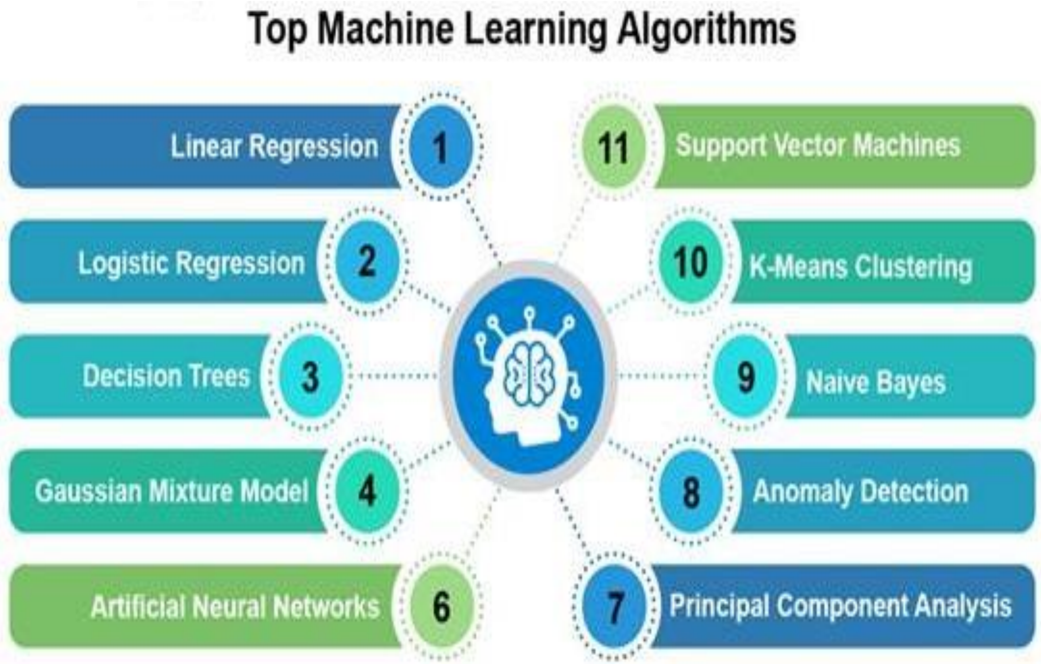
AI- Artificial Intelligence

- Artificial intelligence is about giving machines the capability of mimicking human behavior, particularly intellectual functions.
- Examples : facial recognition, automated driving etc.
- The techniques that fall under artificial intelligence are : linguistics, natural language processing, decision science, bias, vision, robotics, planning, etc.

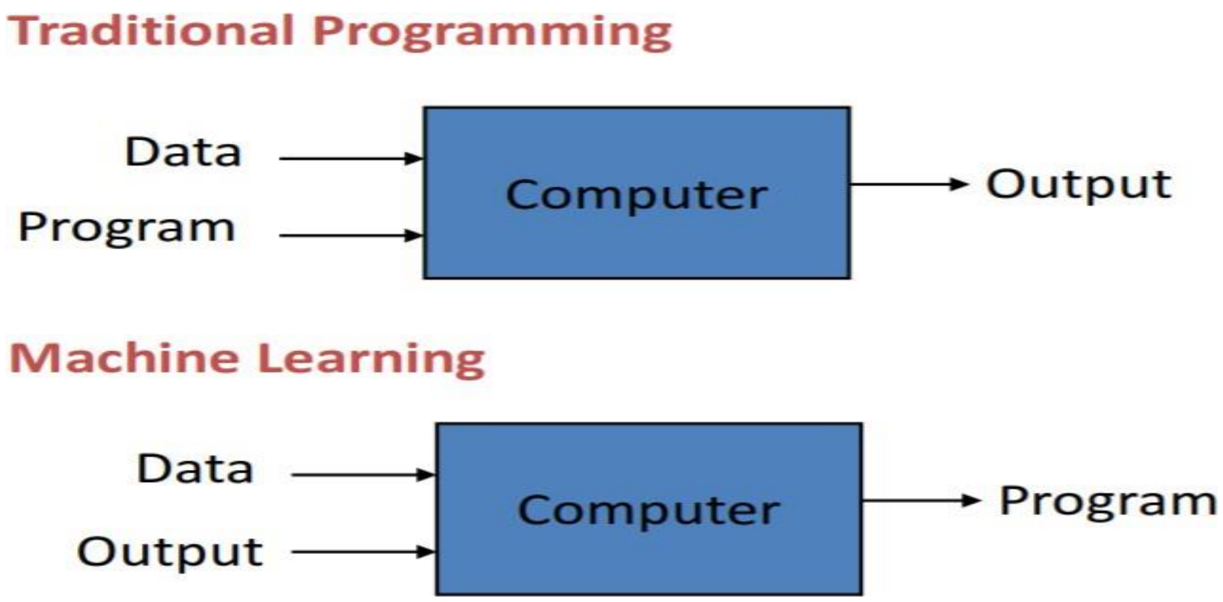


MACHINE LEARNING

- Machine learning can either be considered a sub-field or one of the tools of artificial intelligence.
- It provides machines with the capability of learning from experience.
- Experience for machines done using data. The data that is used to teach machines is called training data.



MACHINE LEARNING…



- A program, transforms input signals into output signals using predetermined rules and relationships.
- Machine learning take both the known input and output (training data) to derive a model for the program to convert input to output.

DATA SCIENCE

- 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.
- 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- Features

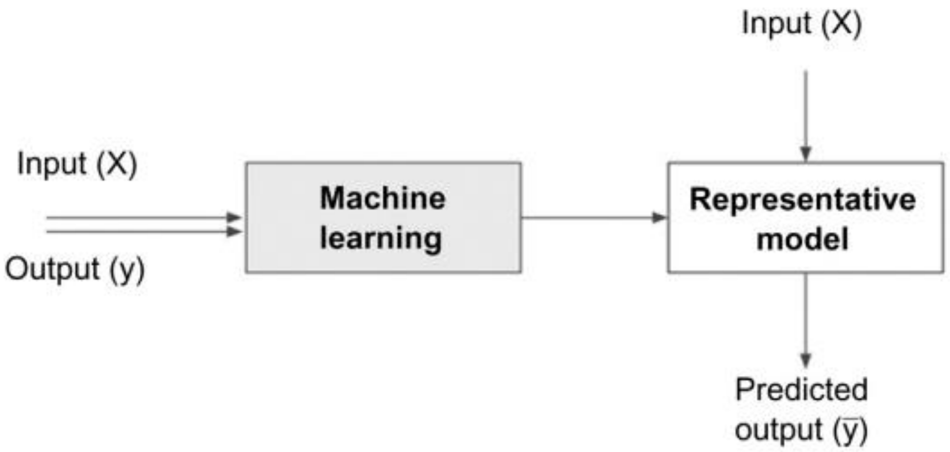
(i) Extracting meaningful Patterns

- Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns or relationships within a dataset to make important decisions.
- One of the key aspects of data science is the process of generalization of patterns from a dataset and it should be valid for both dataset and new unseen data.
- The ultimate objective of data science is to find potentially useful conclusions.

DATA SCIENCE- Features

(ii) 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.
- Modeling is a process in which a representative abstraction is built from the observed dataset.
- For example, based on credit score, income level, and requested loan amount, a model can be developed to determine the interest rate of a loan.
- Once the representative model is created, it can be used to predict the value of the interest rate, based on all the input variables.



DATA SCIENCE- Features

(ii) Building Representative Models….

- Data science is the process of building a representative model that fits the observational data.

- This model serves two purposes:
  - (i) It predicts the output (interest rate) based on the new and unseen set of input variables (credit score, income level, and loan amount).
  - (ii) The model can be used to understand the relationship between the output variable and all the input variables.

DATA SCIENCE- Features

(iii) Combination of Statistics, Machine Learning, and Computing

- To extract knowledge, 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.
- One of the key ingredients of successful data science is substantial prior knowledge about the data and the business processes.

DATA SCIENCE- Features

(iv) Learning Algorithms

- The application of sophisticated learning algorithms for extracting useful patterns from data differentiates data science from traditional data analysis techniques.
- 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.

DATA SCIENCE- Features

(v) Associated Fields

- The associated fields that data science heavily relies on
  - Descriptive statistics: Computing mean, standard deviation, correlation, and other descriptive statistics
  - Exploratory visualization: The process of expressing data in visual coordinates
  - Dimensional slicing: Online analytical processing (OLAP) applications, mainly provide information on the data through dimensional slicing, filtering, and pivoting
  - Hypothesis testing: 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.
  - 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.

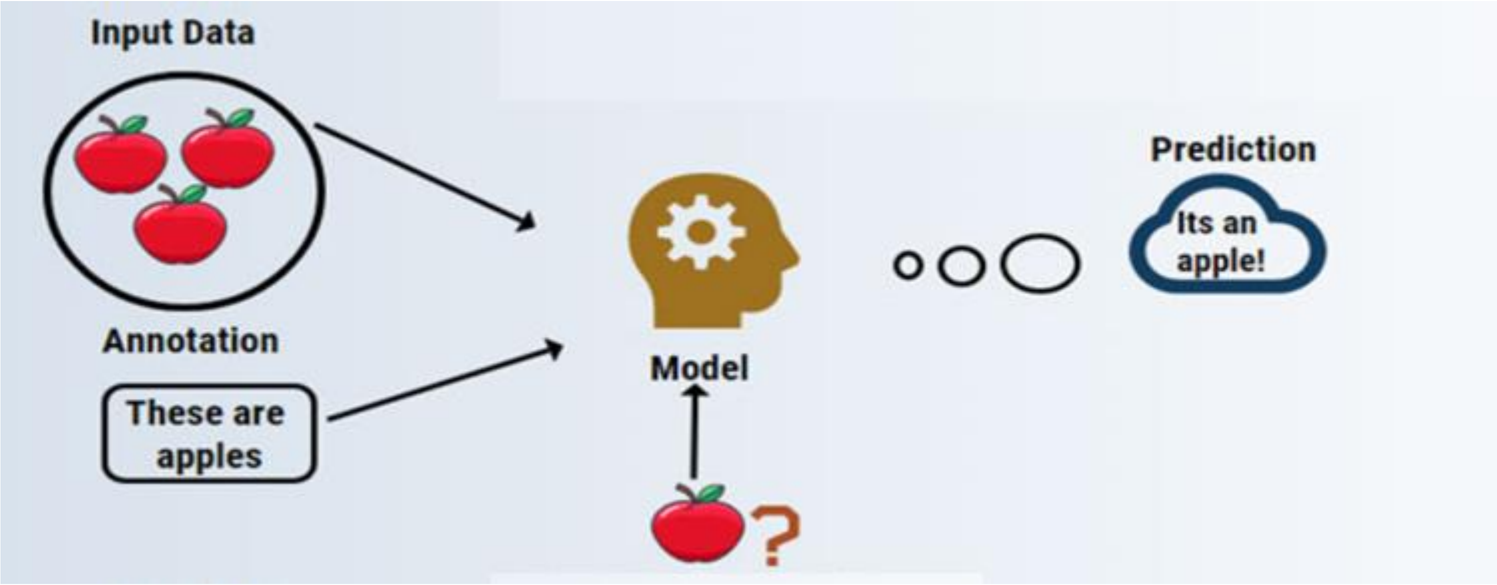
1. DATA SCIENCE CLASSIFICATION

- Data science problems can be broadly categorized into
  - Supervised Learning
  - Unsupervised Learning

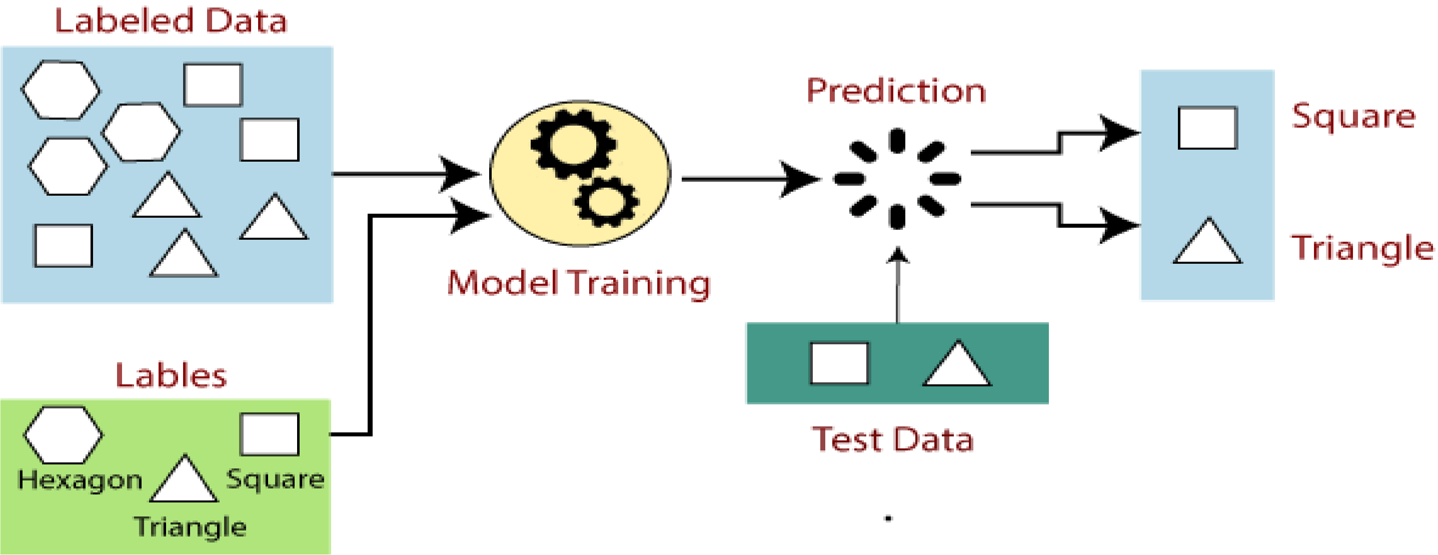
- Supervised Learning tries to infer a function or relationship using labeled training data and uses this function to map new unlabeled data.
- Supervised techniques predict the value of the output variables based on a set of input variables.
- To do this, a model is developed from a training dataset where the values of input and output are previously known.
- The output variable that is being predicted is known as a class label or target variable.

1.1 DATA SCIENCE CLASSIFICATION- Supervised Learning

Ex1:

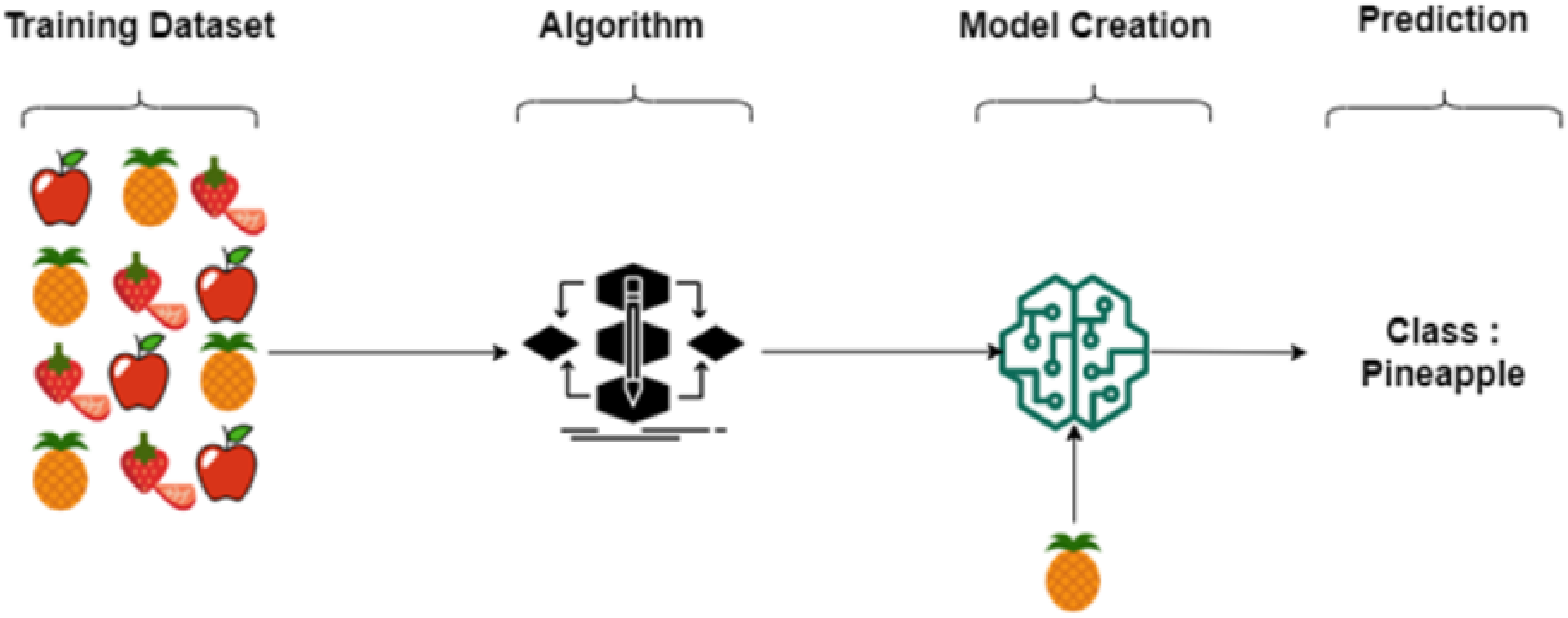


Ex2:

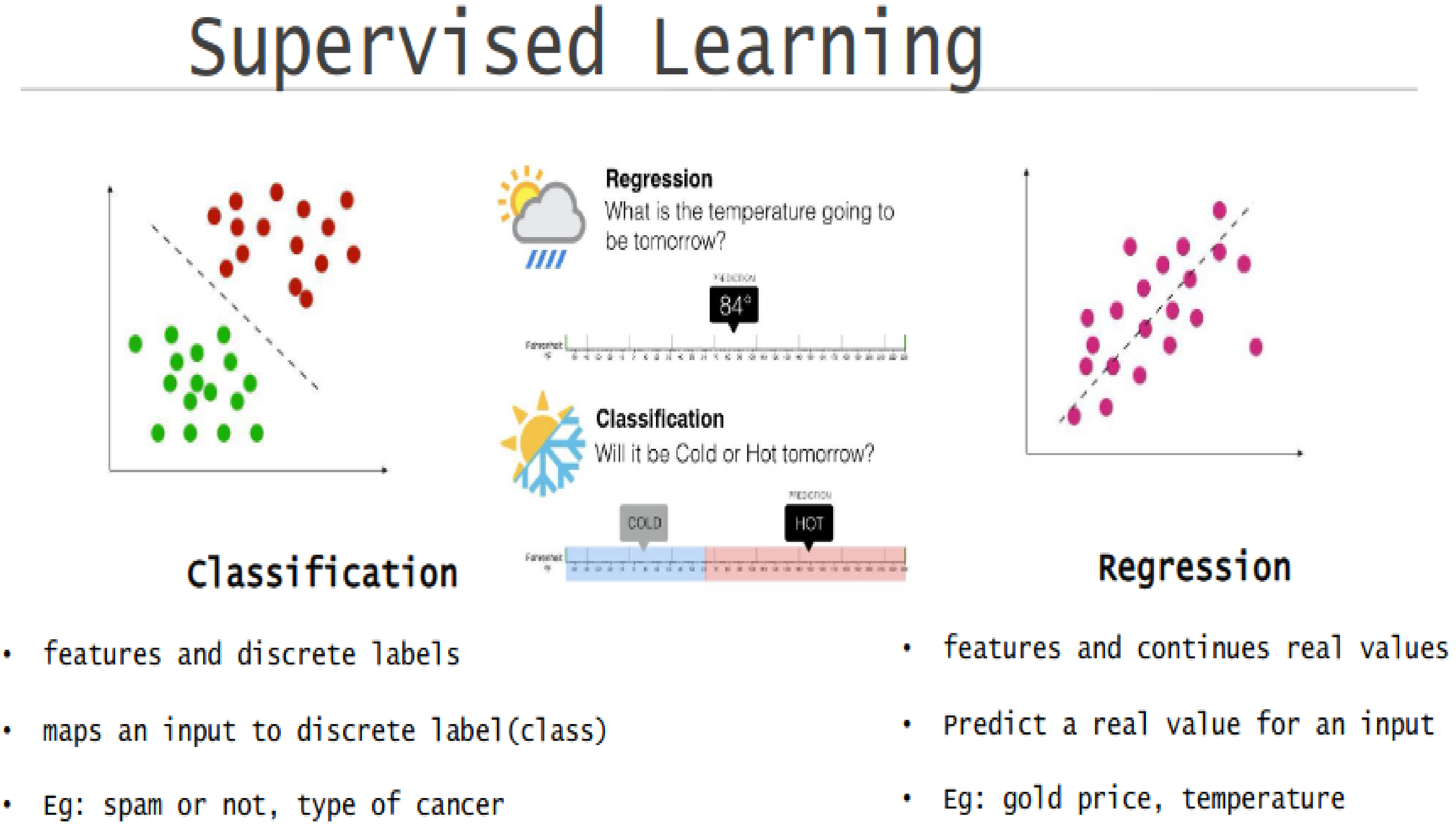


DATA SCIENCE CLASSIFICATION- Supervised Learning

Supervised Learning



DATA SCIENCE CLASSIFICATION- Supervised Learning



Supervised Learning

Regression

What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?

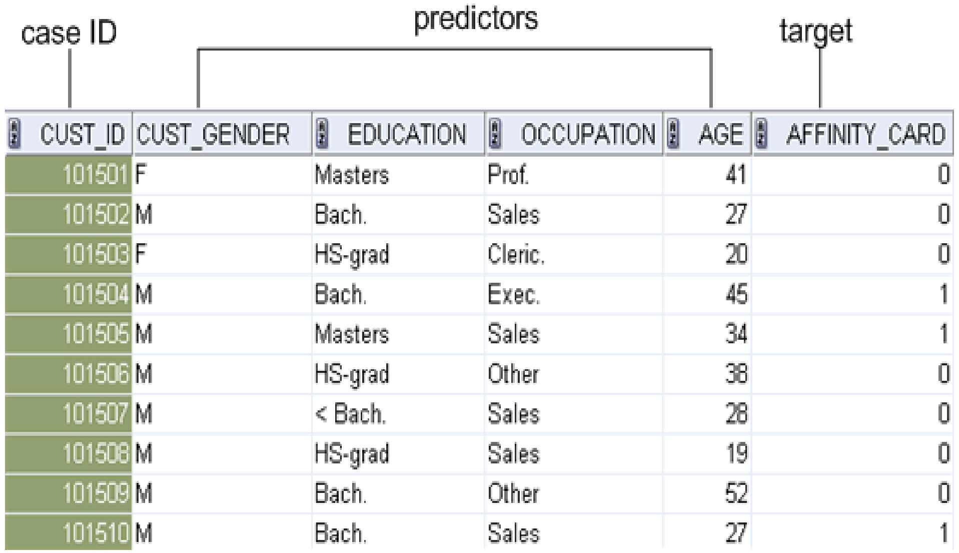
Classification

- features and discrete labels
- maps an input to discrete label(class)
- Eg: spam or not, type of cancer

Regression

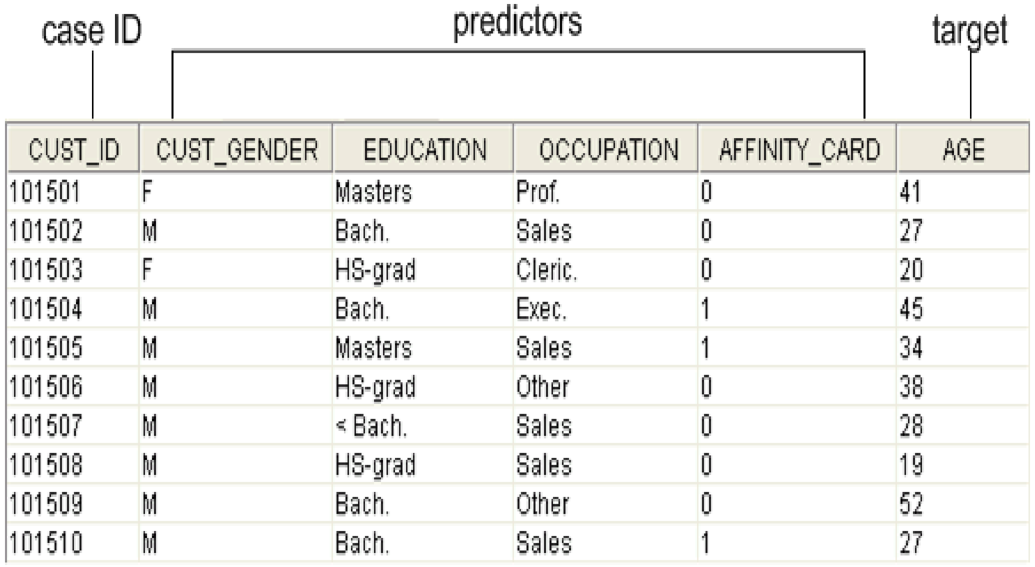
- features and continues real values
- Predict a real value for an input
- Eg: gold price, temperature

DATA SCIENCE CLASSIFICATION- Supervised Learning



CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AGE	AFFINITY_CARD
101501	F	Masters	Prof.	41	0
101502	M	Bach.	Sales	27	0
101503	F	HS-grad	Cleric.	20	0
101504	M	Bach.	Exec.	45	1
101505	M	Masters	Sales	34	1
101506	M	HS-grad	Other	38	0
101507	M	< Bach.	Sales	28	0
101508	M	HS-grad	Sales	19	0
101509	M	Bach.	Other	52	0
101510	M	Bach.	Sales	27	1

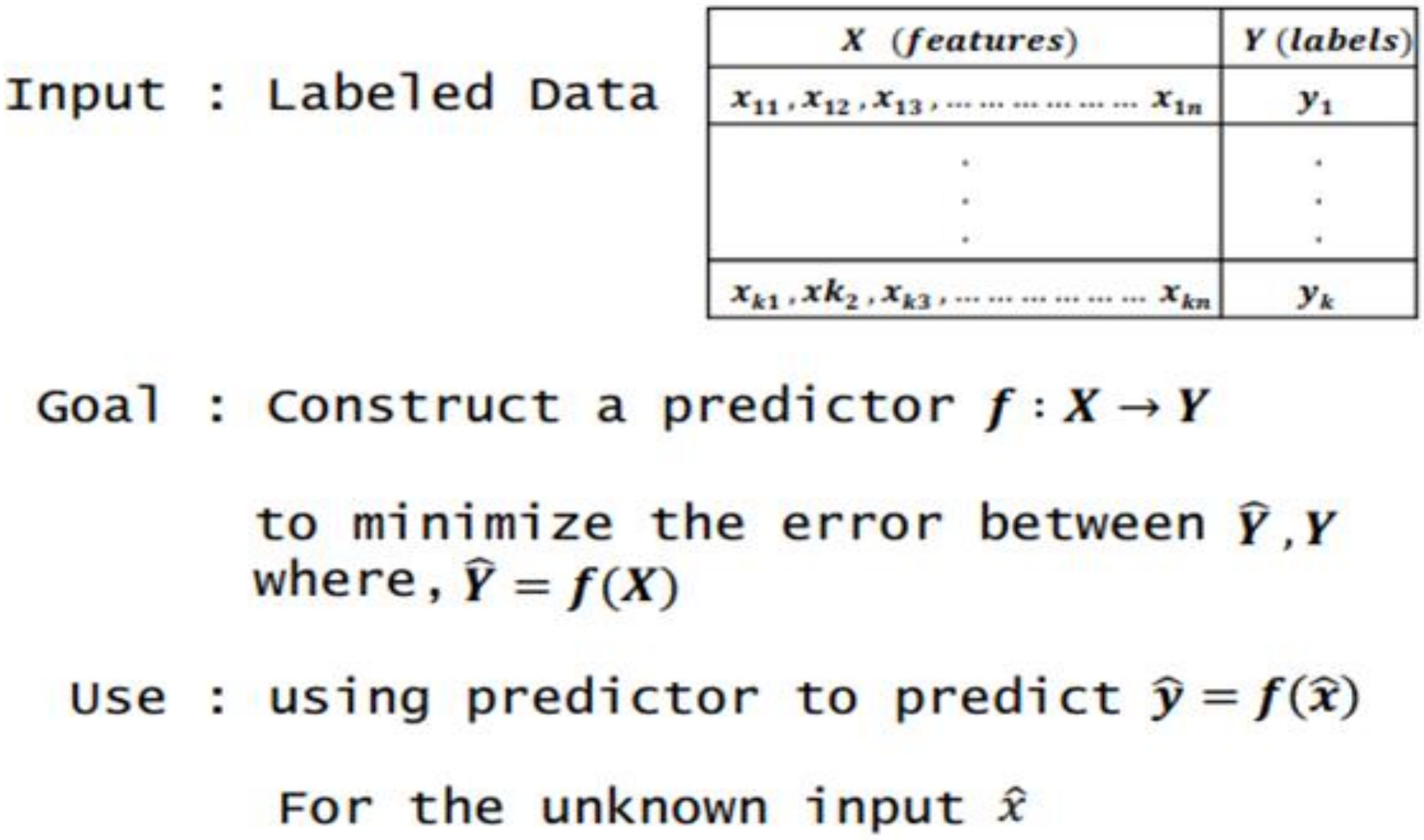
Classification



CUST_ID	CUST_GENDER	EDUCATION	OCCUPATION	AFFINITY_CARD	AGE
101501	F	Masters	Prof.	0	41
101502	M	Bach.	Sales	0	27
101503	F	HS-grad	Cleric.	0	20
101504	M	Bach.	Exec.	1	45
101505	M	Masters	Sales	1	34
101506	M	HS-grad	Other	0	38
101507	M	< Bach.	Sales	0	28
101508	M	HS-grad	Sales	0	19
101509	M	Bach.	Other	0	52
101510	M	Bach.	Sales	1	27

Regression

DATA SCIENCE CLASSIFICATION- Supervised Learning



Input : Labeled Data

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

Goal : Construct a predictor $f : X \rightarrow Y$

to minimize the error between $\widehat{Y}, Y$
where, $\widehat{Y} = f(X)$

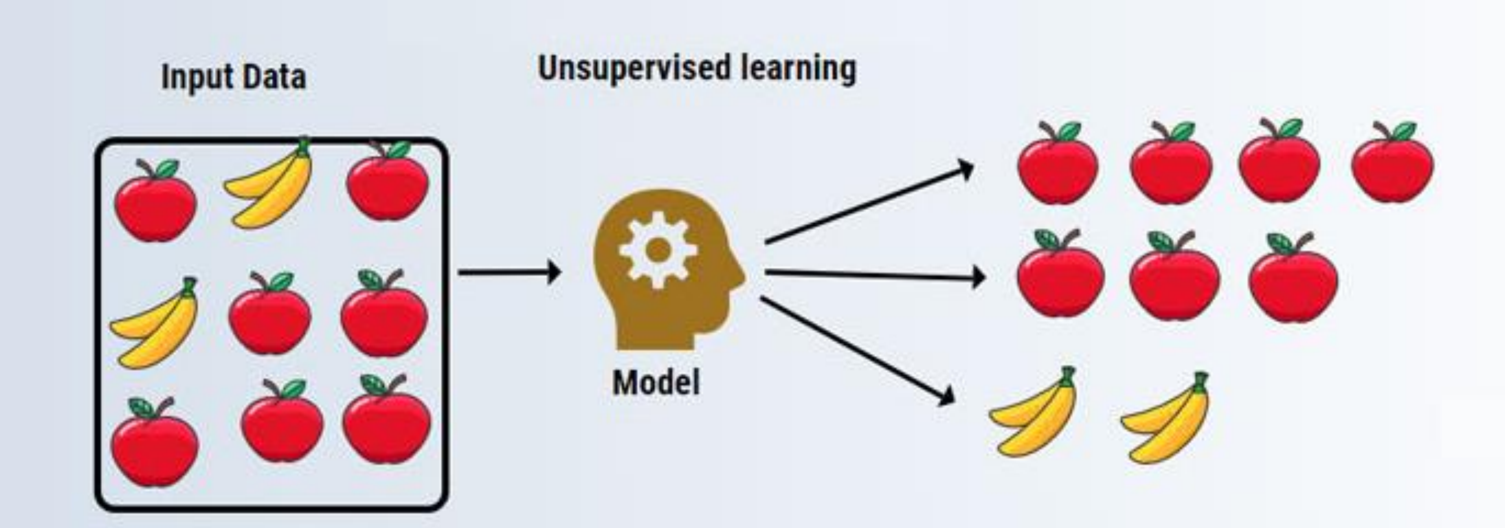
Use : using predictor to predict $\widehat{y} = f(\widehat{x})$

For the unknown input $\hat{x}$

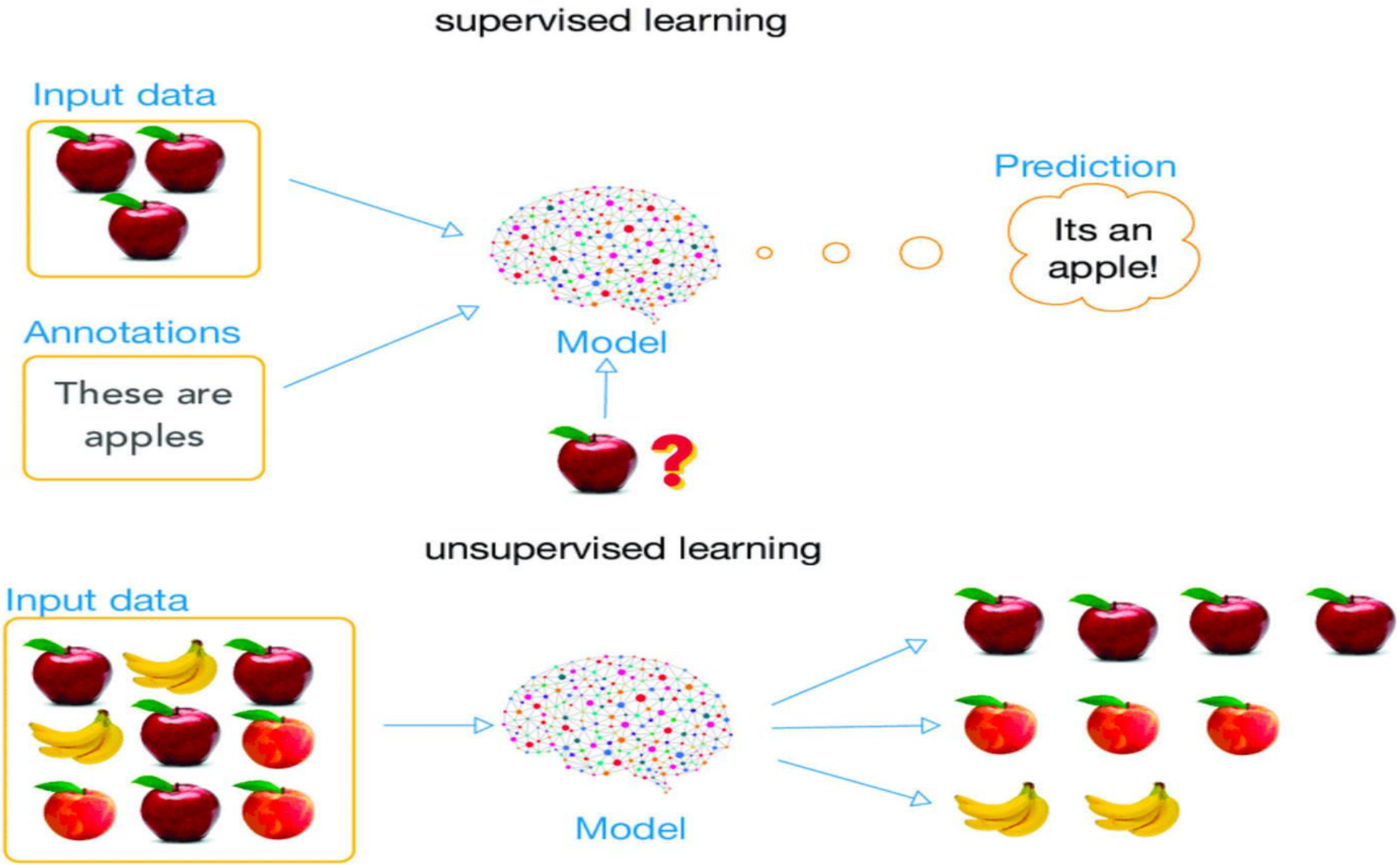
1.2 DATA SCIENCE CLASSIFICATION- Unsupervised Learning

- 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 technique is to find patterns in data based on the relationship between data points themselves.

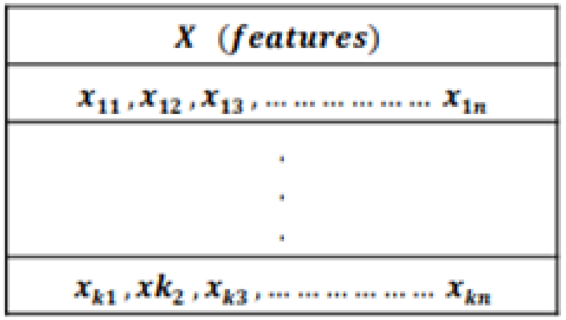


DATA SCIENCE CLASSIFICATION



DATA SCIENCE CLASSIFICATION

Input : Unlabeled Data



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

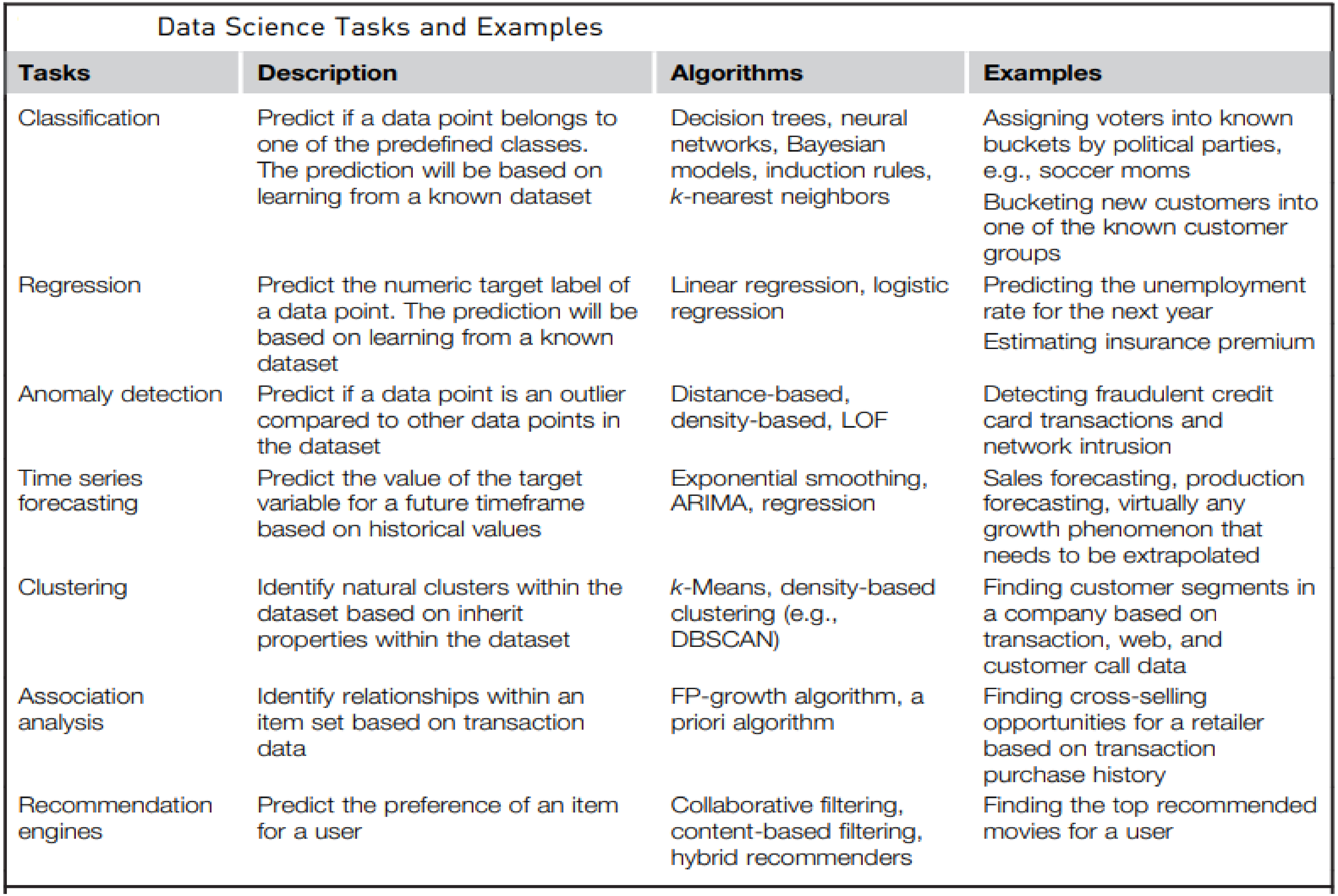
Goal : Construct a analyzer to find the hidden relationship between inputs

$$x_1, x_2, x_3, \dots, x_k$$

Use : Group or associate inputs according to their similarity

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.
- Data science algorithms can be implemented by custom-developed computer programs in almost any computer language.
  - For example, a classification task can be solved using many different learning algorithms such as decision trees, artificial neural networks, k-NN etc.
- 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.
- R, RapidMiner, Python, SAS Enterprise Miner, etc., can be used to implement these algorithms.



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

Data Science process

- The standard data science process involves
  - (1) understanding the problem
  - (2) preparing the data samples
  - (3) developing the model
  - (4) applying the model on a dataset to see how the model may work in the real world
  - (5) deploying and maintaining the models.

Data Science process- Framework

- One of the most popular data science process frameworks is CRoss Industry Standard Process for Data Mining (CRISP-DM) and was developed by a consortium of data mining companies.

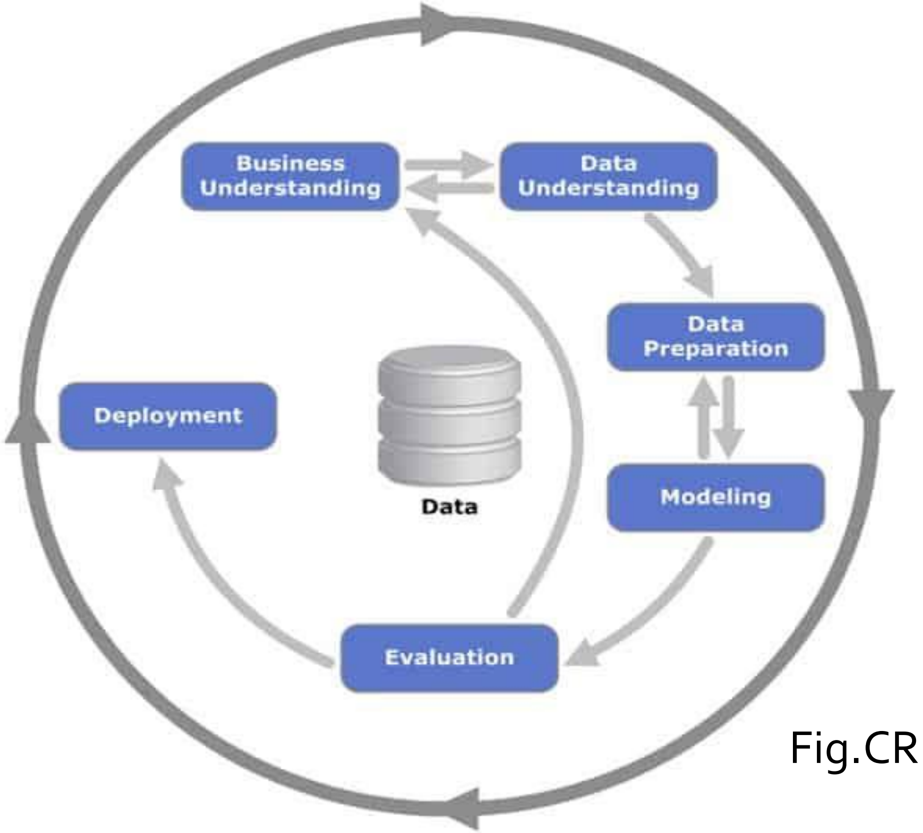
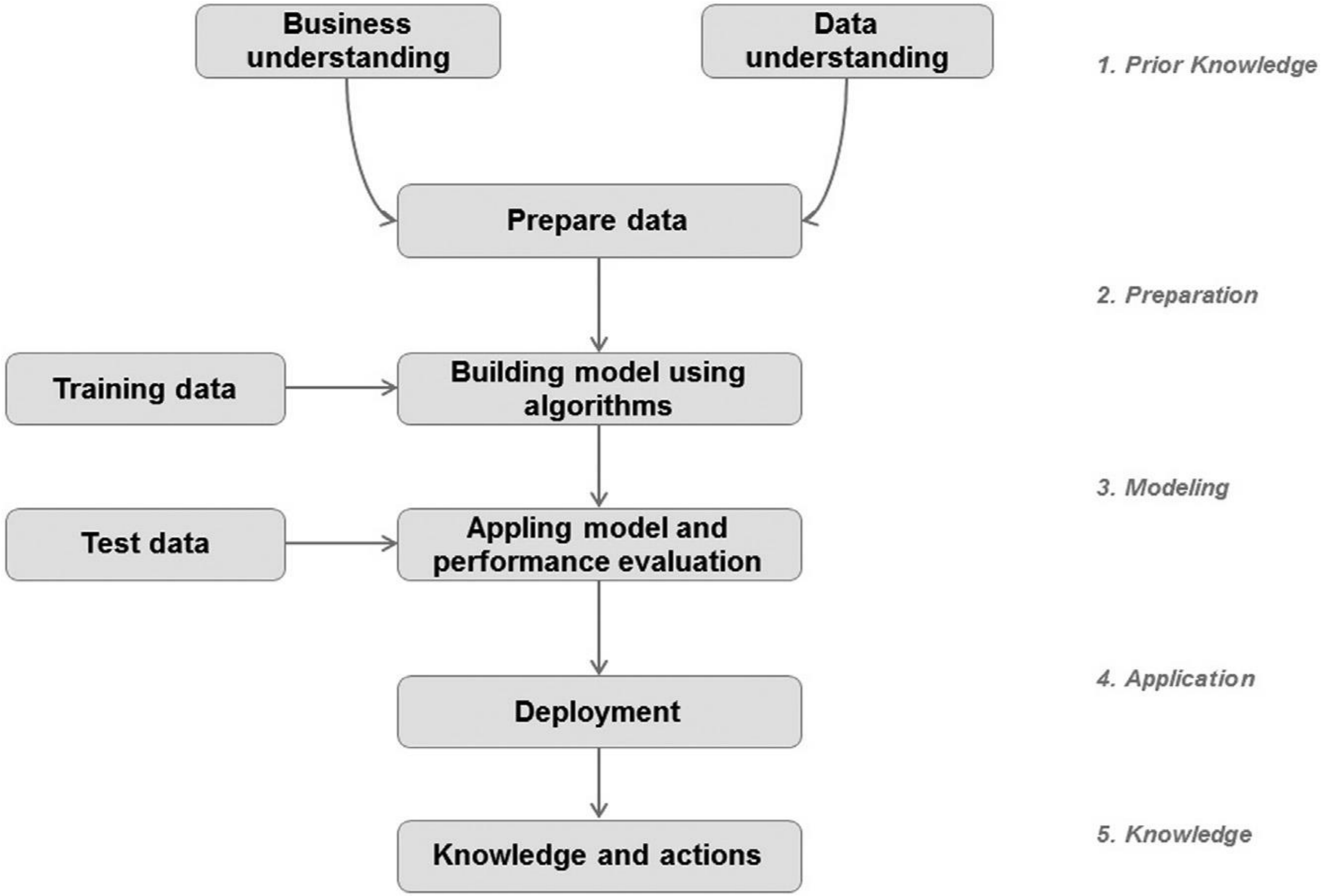


Fig.CRISP data mining framework.

Data Science process- Framework

- The CRISP-DM process is the most widely adopted framework for developing data science solutions.
- Other data science frameworks are SEMMA, an acronym for Sample, Explore, Modify, Model, and Assess developed by the SAS Institute; DMAIC, is an acronym for Define, Measure, Analyze, Improve, and Control, used in Six Sigma practice.
- The Selection, Preprocessing, Transformation, Data Mining, Interpretation, and Evaluation- framework is ofen used in the knowledge discovery process.

2. The Data Science Process



2.1 PRIOR KNOWLEDGE

- 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 solved, how it fits in the business context, and what data is needed in order to solve the problem.

2.1.1 Objective

- The data science process starts with a need for analysis, a question, or a business objective. This is possibly the most important step in the data science process.
- 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.
- As an iterative process, it is common to go back to previous data science process steps, revise the assumptions, approach, and tactics.

2.1.2 Subject Area

- The false or spurious signals are a major problem in the data science process. It is up to the practitioner to sift through the exposed patterns and accept the ones that are valid and relevant to the answer of the objective. Hence, it is essential to know the subject matter, the context, and the business process generating the data.

2.1.3 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. Terminologies include:

2.1.3 Data….

- A **dataset** (example set) is a collection of data with a defined structure.
- A **data point** (record, object or example) is a single instance in the dataset.
- An **attribute** (feature, input, dimension, variable, or predictor) is a single property of the dataset.
- A **label** (class label, output, prediction, target, or response) is the special attribute to be predicted based on all the input attributes.
- **Identifiers** are special attributes that are used for locating to individual records.

Eg: Sample Dataset

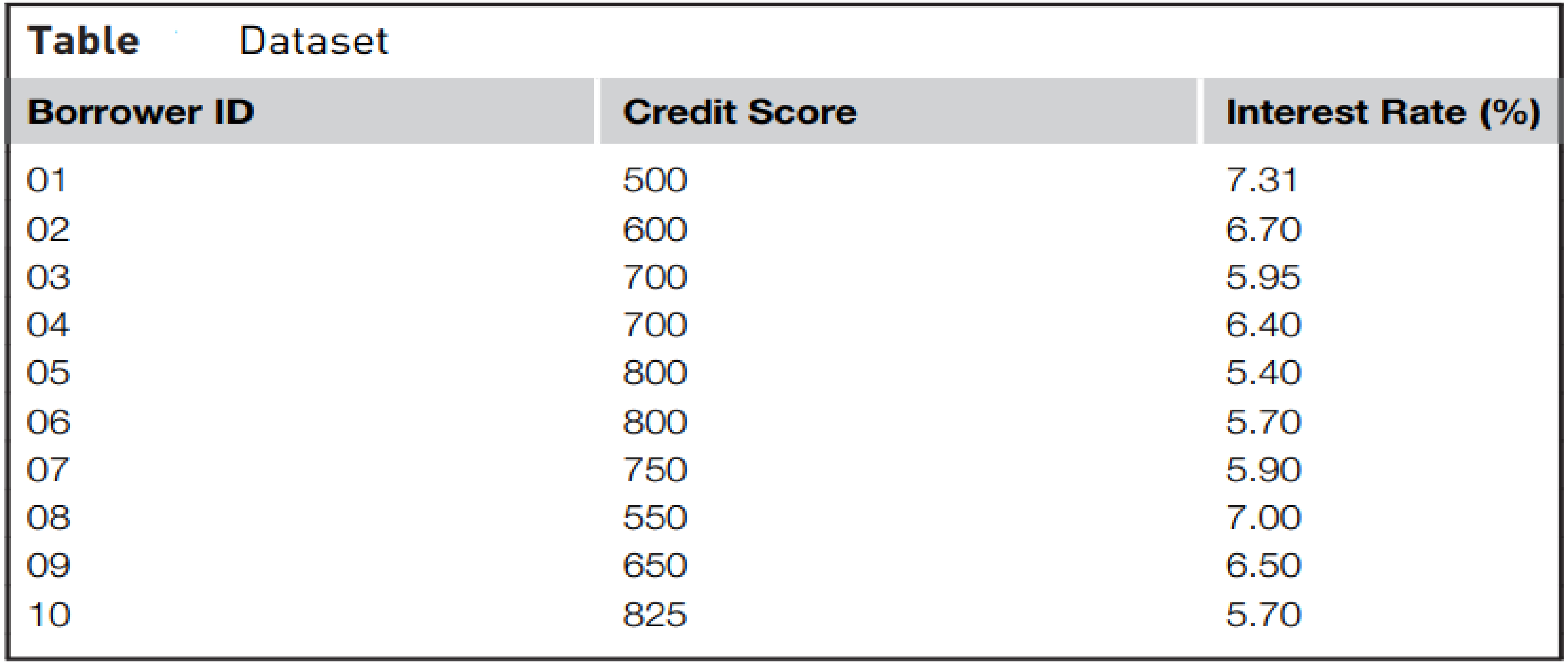


Table Dataset

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

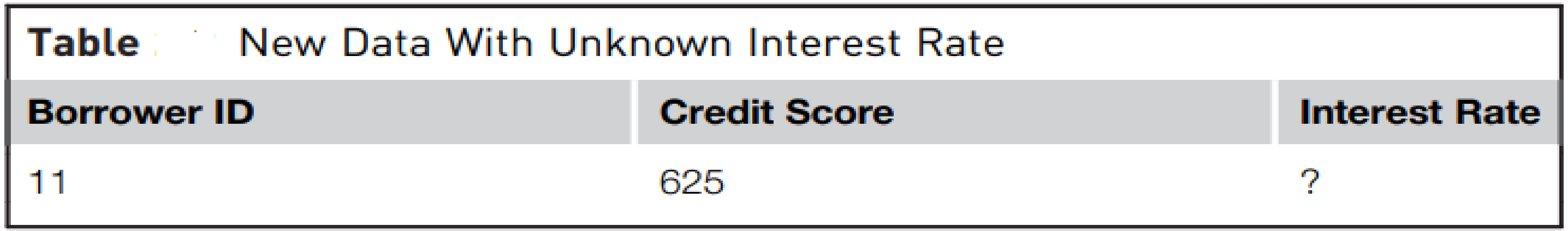


Table New Data With Unknown Interest Rate

Borrower ID	Credit Score	Interest Rate
11	625	?

2.1.4 Causation Versus Correlation

- The correlation between the input and output attributes doesn’t guarantee causation. Hence, it is important to frame the data science question correctly using the existing domain and data knowledge. In this data science example, the interest rate of the new borrower with an unknown interest rate will be predicted based on the pattern learned from known data.

2.2 DATA PREPERATION

- Preparing the dataset to suit a data science task is the most time-consuming part of the process.
- Most of the data science algorithms would require data to be structured in a tabular format with records in the rows and attributes in the columns.

2.2.1 Data Exploration

- Data preparation starts with an in-depth exploration of the data and gaining a better understanding of the dataset. It 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.

2.2.1 Data Exploration…

- They can expose the structure of the data, the distribution of the values, the presence of extreme values, and highlight the inter-relationships within the dataset.
- Descriptive statistics like mean, median, mode, standard deviation, and range for each attribute provide an easily readable summary of the key characteristics of the distribution of data. Also the visual plot of data points provides an instant grasp of all the data points condensed into one chart.

2.2.2 Data Quality

- Data quality is an important concern. Errors in data will affect the representativeness of the model. Organizations use data alerts, cleansing, and transformation techniques to improve and manage the quality of the data and store them in companywide repositories called data warehouses.

2.2.2 Data Quality....

- The data cleansing practices include elimination of duplicate records, isolating outlier records that exceed the bounds, standardization of attribute values, substitution of missing values, etc.

2.2.3 Missing Values

- There are several methods to deal with this problem, but each method has pros and cons. The **first step** of managing missing values is to understand the reason behind why the values are missing. Knowing the source of a missing value will often guide which mitigation methodology to use. The missing value can be substituted with a range of artificial data (mean, minimum, or maximum value, depending on the characteristics of the attribute) so that the issue can be managed. This method is useful if the missing values occur randomly and the frequency of occurrence is quite rare.

2.2.3 Missing Values..

- Alternatively, all the data records with missing values or records with poor data quality can be ignored. This method reduces the size of the dataset.

2.2.4 Data Types and Conversion

- The attributes in a dataset can be of different types, such as continuous numeric, integer numeric, or categorical. Different data science algorithms impose different restrictions on the attribute data types.
- In case of linear regression models, the input attributes have to be numeric. 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-5 400, good-5 600, excellent -5700, etc.

2.2.4 Data Types and Conversion...

- Similarly, numeric values can be converted to categorical data types by a technique called **binning**, where a range of values are specified for each category.
- For example, a score between 400 and 500 can be encoded as “low” and so on.

2.2.5 Transformation

- Many data science algorithm compares the values of different attributes and calculates distance between the data points. Normalization prevents one attribute dominating the distance results because of large values. One solution is to convert the range of attributes to a more uniform scale from 0 to 1 by normalization. This way, a consistent comparison can be made between the two different attributes with different units.

2.2.6 Outliers

- Outliers are anomalies in a given dataset. Outliers may occur because of correct data capture or erroneous data capture. The presence of outliers needs to be understood and will require special treatments.

2.2.7 Feature Selection

- In a dataset, not all the attributes are equally important in predicting the target. The presence of some attributes might be counterproductive. Some of the attributes may be highly correlated with each other. As the number of dimensions in the data increase, data becomes sparse in high-dimensional space. This condition degrades the reliability of the models, especially in the case of clustering and classification.
- Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection

2.2.8 Data Sampling

- Sampling is a process of selecting a subset of records as a representation of the original dataset for use in data analysis. Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling.
- To build a model, it is necessary to segment the datasets into training and test samples. The training dataset is sampled from the original dataset using simple sampling or class label specific sampling. Stratified sampling is a process of sampling where each class is equally represented in the sample.

2.3 MODELING

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

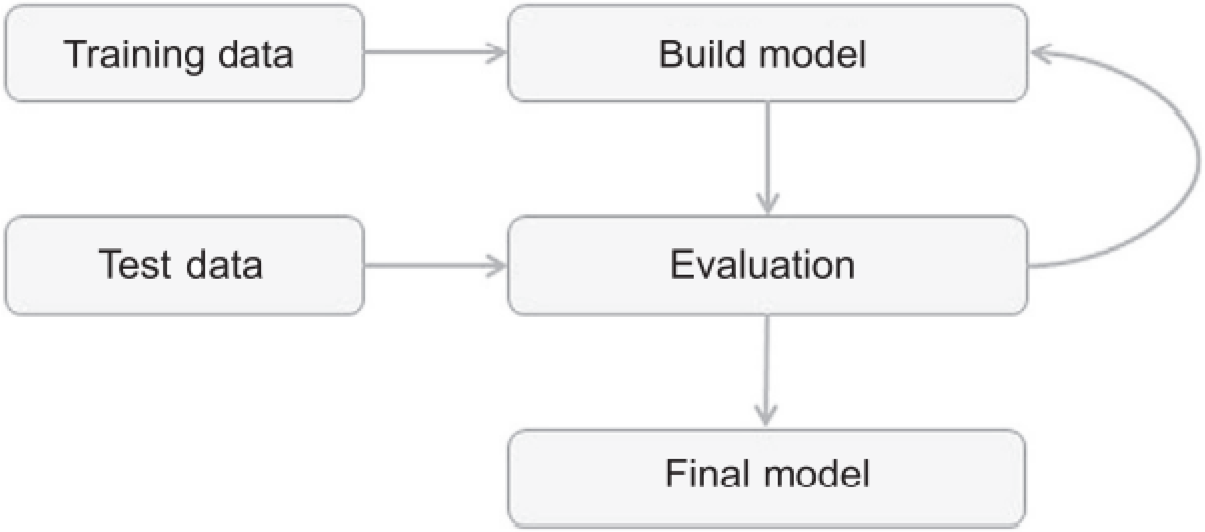


Fig. Steps in Modelling Process

2.3.1 Training & Testing Dataset

- The modeling step creates a representative model from the data. The dataset used to create the model, with known attributes and target, is called the training dataset. The validity of the created model will also need to be checked with another known dataset called the test dataset or validation dataset.
- To facilitate this process, the overall known dataset can be split into a training dataset and a test dataset. A standard rule is two-thirds of the data are to be used as training and one-third as a test dataset.

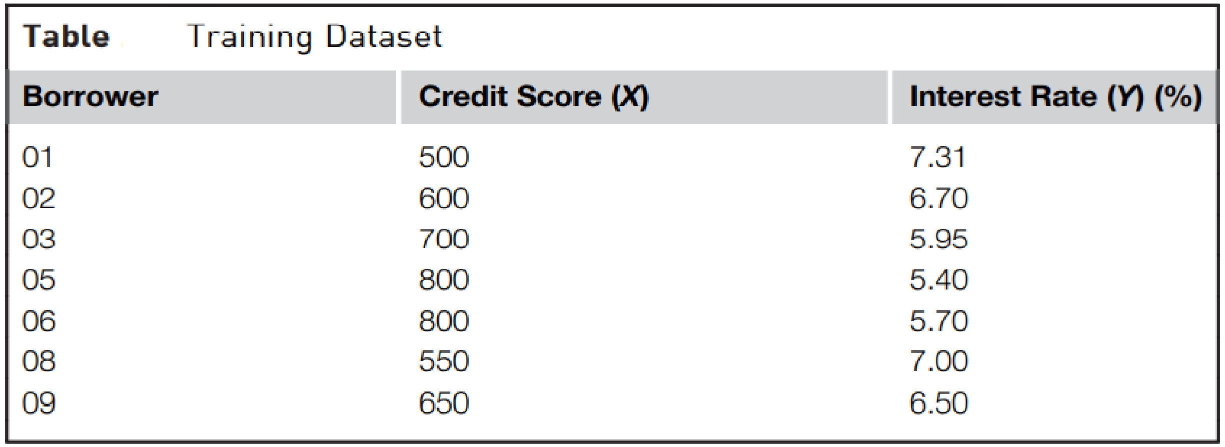


Table Training Dataset

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

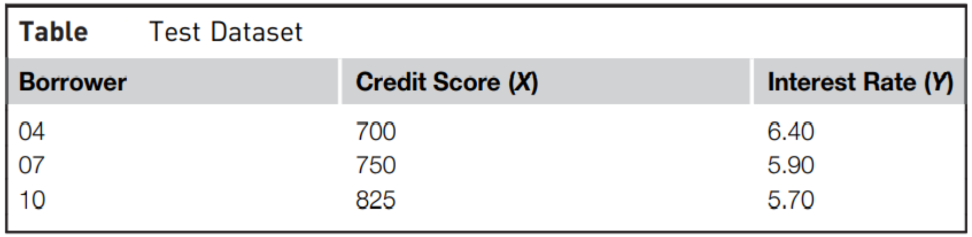


Table Test Dataset

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

2.3.2 Learning Algorithms

- 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.

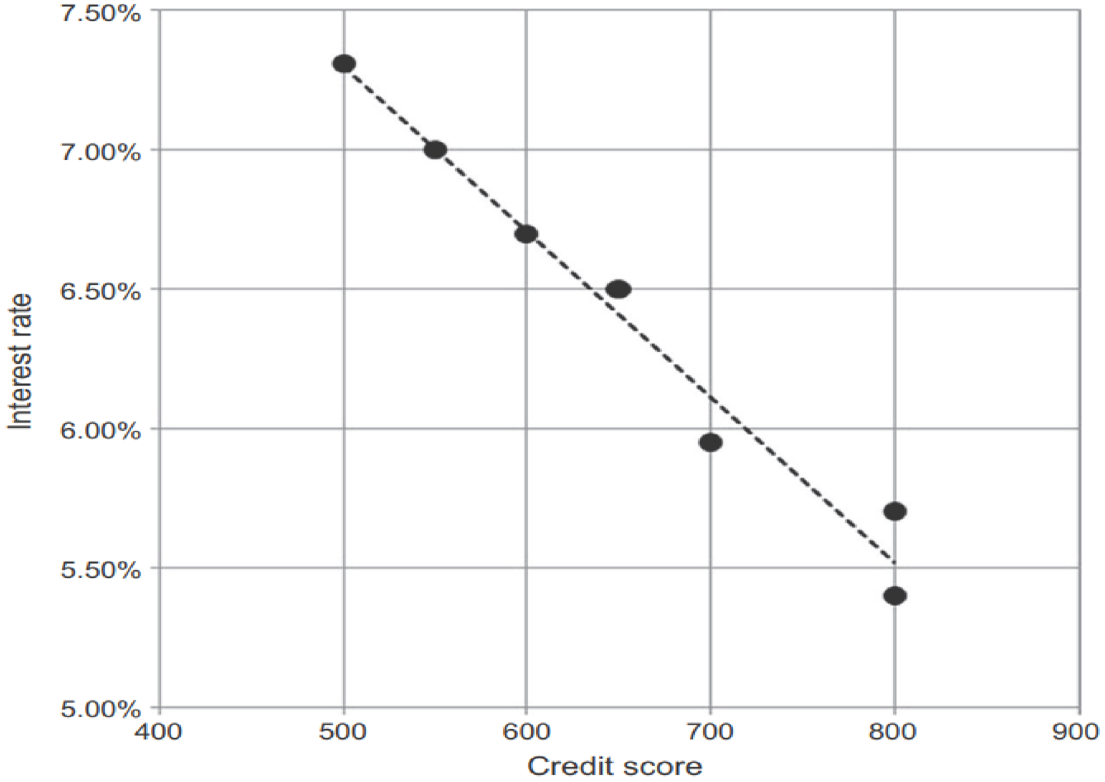
Eg: Linear Regression Technique

- In linear regression, its objective is to fit a straight line through the data points in a scatterplot. The line is built in such a way that the sum of the squared distance from the data points to the line is minimal. The line can be expressed as:

$$y = a * x + b$$

  - where y is the output or dependent variable, x is the input or independent variable, b is the y-intercept, and a is the coefficient of x. The values of a and b can be found in such a way to minimize the sum of the squared residuals of the line.

2.3.2 Learning Algorithms.....



2.3.3 Evaluation of the Model

- Once the model is created, its evaluation is done using Test set data.

2.3.3 Evaluation of the Model..

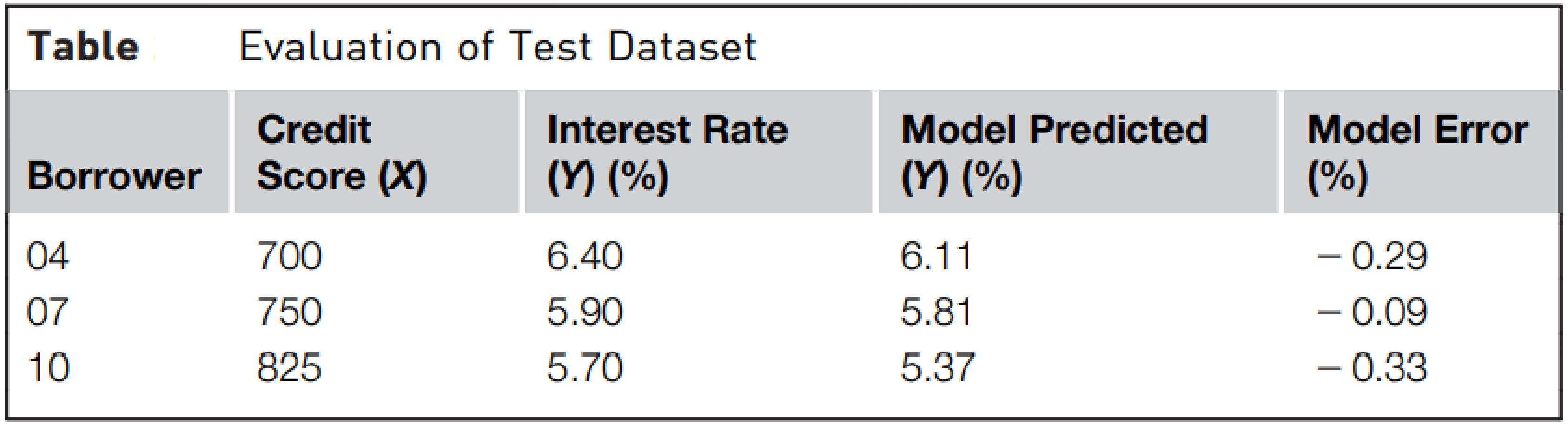


Table Evaluation of Test Dataset

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

- The above table provides the three testing records where the value of the interest rate is known; these records were not used to build the model.
- The actual value of the interest rate can be compared against the predicted value using the model, and thus, the prediction error can be calculated. As long as the error is acceptable, this model is ready for deployment.

2.3.4 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.
- Even though the ensemble model has multiple base models within the model, it acts and performs as a single model. Most of the practical data science applications utilize ensemble modeling techniques

2.3. Modeling….

- At the end of the modeling stage of the data science process, one has
- (1) analyzed the business question;
- (2) sourced the data relevant to answer the question;
- (3) selected a data science technique to answer the question;
- (4) picked a data science algorithm and prepared the data to suit the algorithm;
- (5) split the data into training and test datasets;
- (6) built a generalized model from the training dataset;
- (7) validated the model against the test dataset.

The derived model can now be used to predict target for the problem considered.

2.4. 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
  - Assimilation

2.4.1 Model Readiness

- The production readiness part of the deployment determines the critical qualities required for the deployment objective. Consider the case of determining whether a consumer qualifies for a loan….

2.4.1 Model Readiness….

- The consumer credit approval process is a real-time endeavor. Either through a consumer-facing website or through a specialized application for frontline agents, the credit decisions and terms need to be provided in real-time as soon as prospective customers provide the relevant information. It is optimal to provide a quick decision while also proving accurate. The decision-making model has to collect data from the customer, integrate third-party data like credit history, and make a decision on the loan approval and terms in a matter of seconds. The critical quality of this model deployment is real-time prediction.

2.4.2 Technical Integration

- Currently, data science automation tools or coding use R or Python to develop models. Data science tools save time as they do not require the writing of custom codes to execute the algorithm, allows the analyst to focus on the data, business logic, and exploring pattern from the data.

- The models created by data science tools can be ported to production applications by utilizing the Predictive Model Markup Language (PMML). PMML provides a portable and consistent format of model description which can be read by most data science tools. This allows the flexibility for practitioners to develop the model with one tool and deploy it in another tool or application.

2.4.3 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. There are trade-offs to be made between production responsiveness and modeling build time. The quality of prediction, accessibility of input data, and the response time of the prediction remain the critical quality factors in business application.

2.4.4 Model Refresh

- It is quite normal that the conditions in which the model is built change after the model is sent to deployment. 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. Creating a maintenance schedule is a key part of a deployment plan that will sustain a relevant model.

2.4.5 Assimilation

- 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. The challenge for the data science practitioner is to articulate the findings, establish relevance to the original business question, quantify the risks in the model, and quantify the business impact.
- This challenge can be addressed by focusing on the end result, the impact of knowing the discovered information, and the follow-up actions, instead of the technical process of extracting the information through data science

2.5 KNOWLEDGE

- Data science provides various options in terms of algorithms and parameters within the algorithms. Using these options, we can extract the right information from data, and is a bit of an art. This can be developed with practice.
- Not all discovered patterns lead to incremental knowledge. Again, it is up to the practitioner to invalidate the irrelevant patterns and identify the meaningful information.