Data Science

Unit –I

B.Tech. VI Semester AIM
2024-25

By
Dr. Loshma Gunisetti,
Professor and Head , AIML Dept.

Text Book:

- Data Science Concepts and Practice, Vijay Kotu, Bala Deshpande, 2nd Edition, Morgan Kaufmann Publishers.

Reference Books:

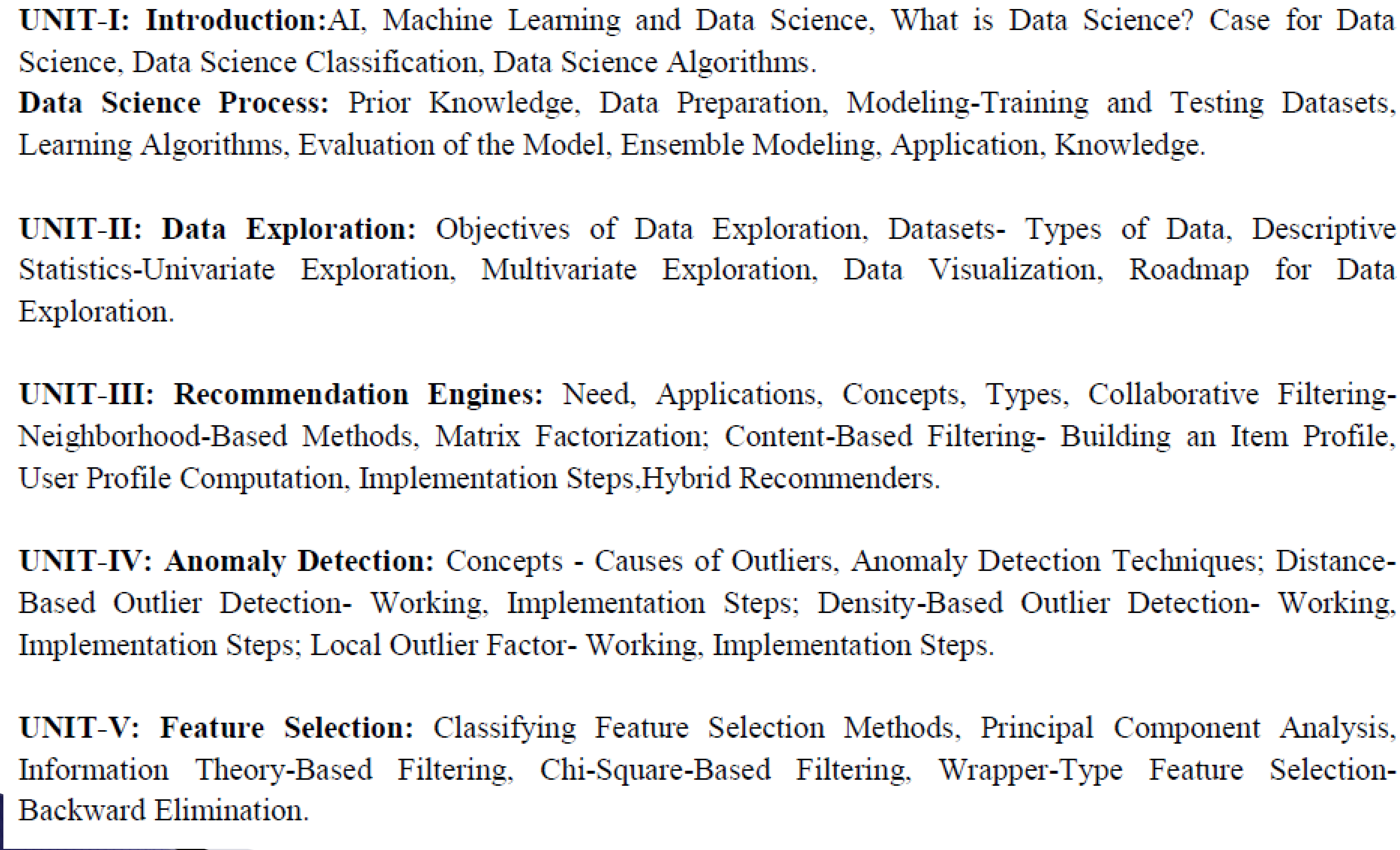
1. An Introduction to Data Science, Jeffrey S. Saltz,Jeffrey M. Stanton, Sage Publications.
2. The Art of Data Science, Roger D Peng, Elizabeth Matsui, Lean Publishing.
3. Data Science for Business, Foster Provost, Tom Fawcett,O'Reilly Media.

Course Outcomes

After successful completion of this course, the student will be able to:

CO	**Course Outcomes**	**Knowledge Level**
1	**Discuss the fundamental concepts of Data Science.**	**K2**
2	Illustrate Exploratory Data Analysis.	K2
3	Explain the Concepts of Recommendation Engines.	K2
4	Explain various Anomaly Detection Techniques.	K2
5	Discuss Feature Selection techniques.	K2

Syllabus



UNIT-I: Introduction:AI, Machine Learning and Data Science, What is Data Science? Case for Data Science, Data Science Classification, Data Science Algorithms.

Data Science Process: Prior Knowledge, Data Preparation, Modeling-Training and Testing Datasets, Learning Algorithms, Evaluation of the Model, Ensemble Modeling, Application, Knowledge.

UNIT-II: Data Exploration: Objectives of Data Exploration, Datasets- Types of Data, Descriptive Statistics-Univariate Exploration, Multivariate Exploration, Data Visualization, Roadmap for Data Exploration.

UNIT-III: Recommendation Engines: Need, Applications, Concepts, Types, Collaborative Filtering-Neighborhood-Based Methods, Matrix Factorization; Content-Based Filtering- Building an Item Profile, User Profile Computation, Implementation Steps,Hybrid Recommenders.

UNIT-IV: Anomaly Detection: Concepts - Causes of Outliers, Anomaly Detection Techniques; Distance-Based Outlier Detection- Working, Implementation Steps; Density-Based Outlier Detection- Working, Implementation Steps; Local Outlier Factor- Working, Implementation Steps.

UNIT-V: Feature Selection: Classifying Feature Selection Methods, Principal Component Analysis, Information Theory-Based Filtering, Chi-Square-Based Filtering, Wrapper-Type Feature Selection-Backward Elimination.

Unit I Contents

- **Introduction:**
  - AI, Machine Learning and Data Science
  - What is Data Science?
  - Case for Data Science
  - Data Science Classification
  - Data Science Algorithms.
- **Data Science Process:**
  - Prior Knowledge
  - Data Preparation
  - Modeling
    - Training and Testing Datasets
    - Learning Algorithms
    - Evaluation of the Model
    - Ensemble Modeling
  - Application
  - Knowledge

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.

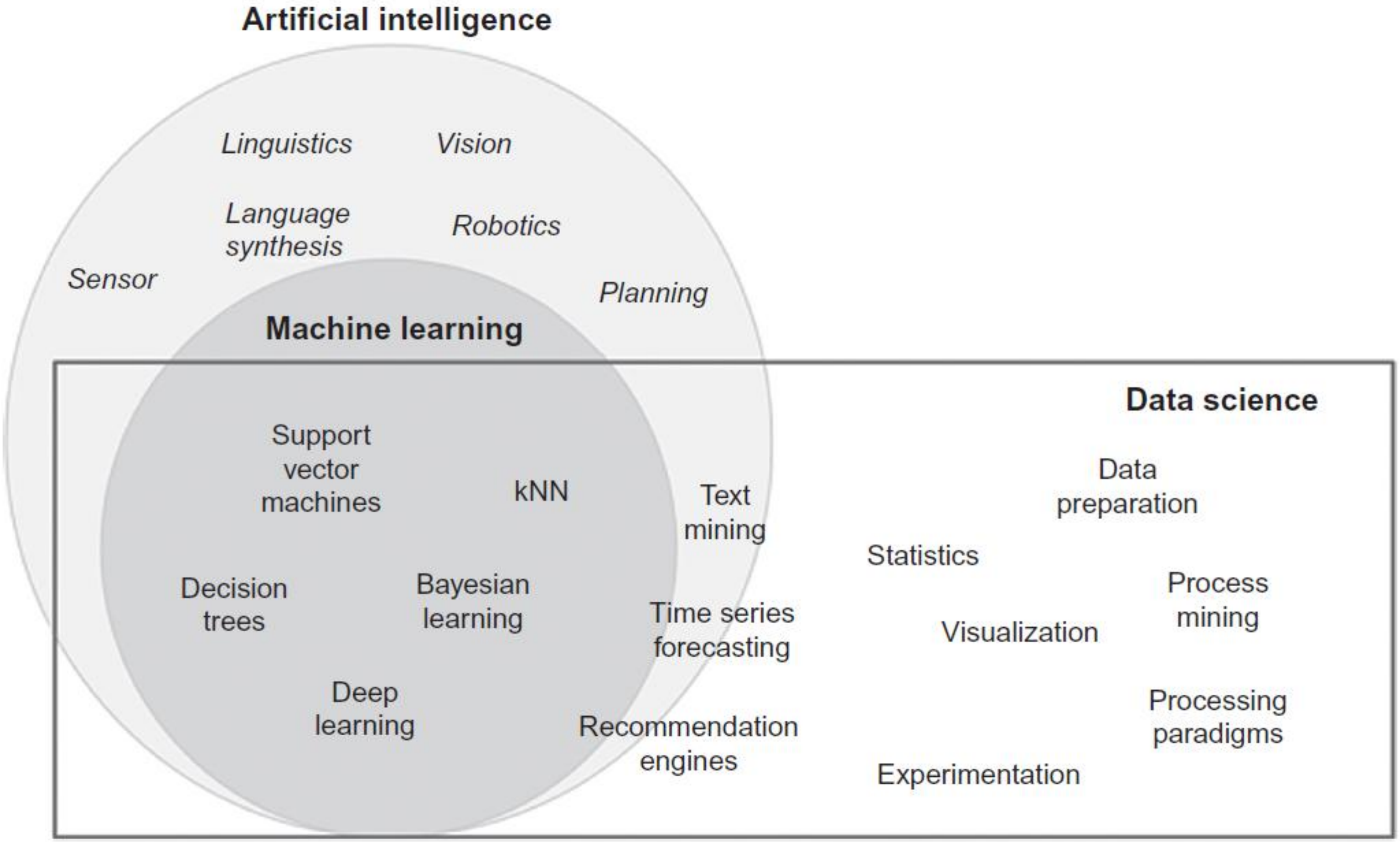
Introduction

- 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.
- As the ability to collect, store, and process data has increased, in line with Moore’s Law - which implies that computing hardware capabilities double every two years, data science has found increasing applications in many diverse fields.

Introduction

- The process involved in data science, however, has not changed since those early days and is not likely to change much in the foreseeable future.
- 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. But what may change is the automation available to do this.

Introduction : AI, Machine Learning and Data Science



Artificial intelligence, machine learning, and data science.

Introduction : AI, Machine Learning and Data Science

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

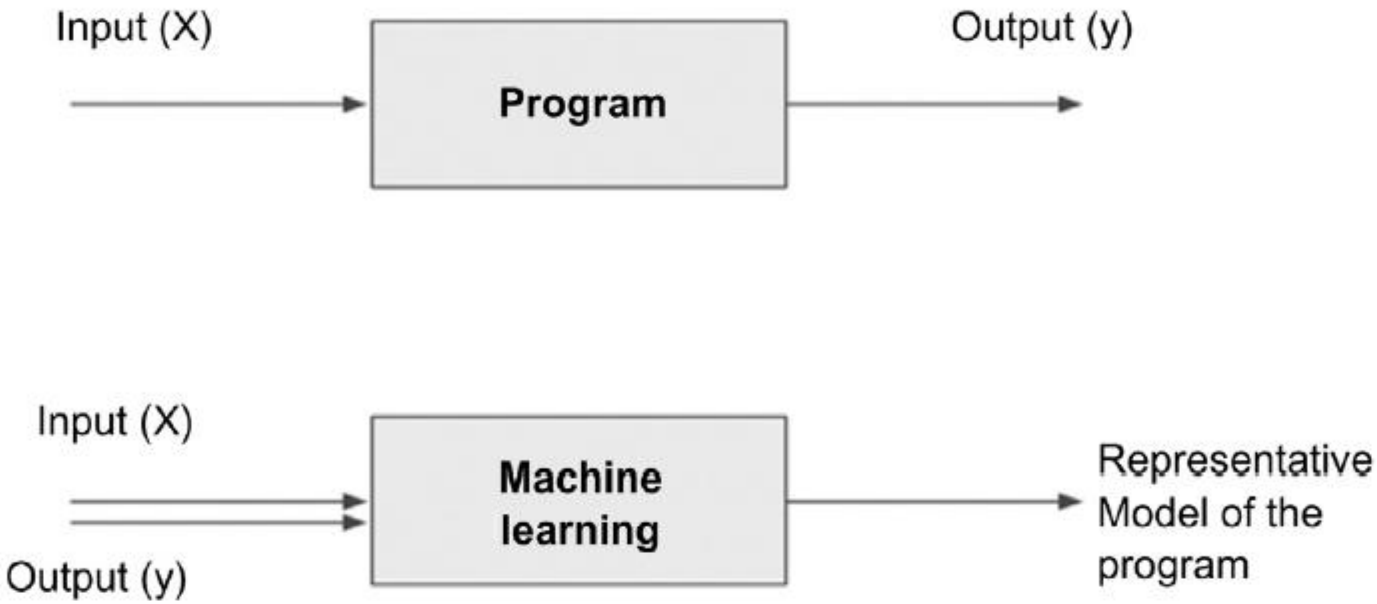
Introduction : AI, Machine Learning and Data Science

- 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.

Introduction : AI, Machine Learning and Data Science

- 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.

Introduction : AI, Machine Learning and Data Science



Traditional program and machine learning.

Introduction : AI, Machine Learning and Data Science

- 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.
- For example, many organizations like social media platforms, review sites, or forums are required to moderate posts and remove abusive content.
- How can machines be taught to automate the removal of abusive content?

Introduction : AI, Machine Learning and Data Science

- The machines need to be shown examples of both abusive and non-abusive posts with a clear indication of which one is abusive.
- The learners will generalize a pattern based on certain words or sequences of words in order to conclude whether the overall post is abusive or not.
- The model can take the form of a set of “if-then” rules.
- Once the data science rules or model is developed, machines can start categorizing the disposition of any new posts.

Introduction : AI, Machine Learning and 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.
- In the context of how data science is used today, it relies heavily on machine learning and is sometimes called data mining

Introduction : AI, Machine Learning and Data Science

- 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).

What is Data Science?

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.

What is Data Science?

Extracting Meaningful Patterns

- The generalization should be valid, not just for the dataset used to observe the pattern, but also for new unseen data.
- Data science is also a process with defined steps, each with a set of tasks.
- The term novel indicates that data science is usually involved in finding previously unknown patterns in data.
- The ultimate objective of data science is to find potentially useful conclusions that can be acted upon by the users of the analysis.

What is Data Science?

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.
- 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.

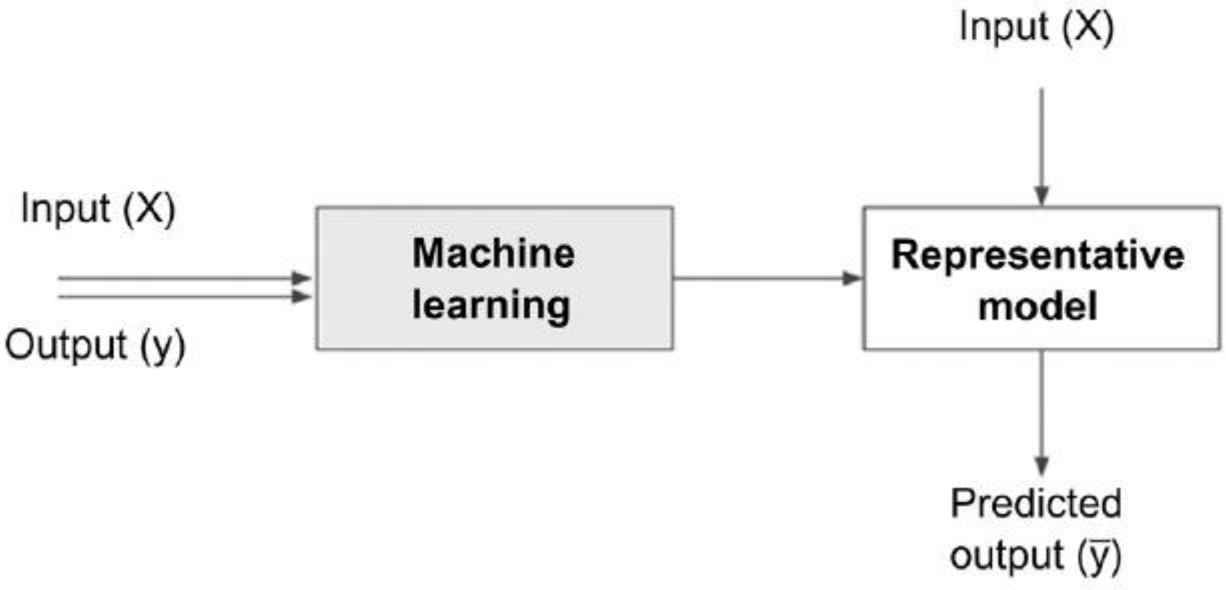
What is Data Science?

Building Representative Models

- For this task, previously known observational data including credit score, income level, loan amount, and interest rate are needed.

What is Data Science?

Building Representative Models



Process of generating a Data Science model

What is Data Science?

Building Representative Models

- 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.

What is Data Science?

Building Representative Models

- For example, does income level really matter in determining the interest rate of a loan?
- Does income level matter more than credit score?
- What happens when income levels double or if credit score drops by 10 points?
- A Model can be used for both predictive and explanatory applications.

What is Data Science?

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.

What is Data Science?

Combination of Statistics, Machine Learning, and Computing

- 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.
- 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.

What is Data Science?

Combination of Statistics, Machine Learning, and Computing

- 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.

What is Data Science?

Learning Algorithms

- We can 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

What is Data Science?

Learning Algorithms

- Some algorithms are based on the foundations of Bayesian probabilistic theories and regression analysis, originating from hundreds of years ago.
- 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.

What is Data Science?

Learning Algorithms

- 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.

What is Data Science?

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. They are used in the exploration stage of the data science process.

What is Data Science?

Associated Fields

- **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.

What is Data Science?

Associated Fields

- **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).

What is Data Science?

Associated Fields

- **Dimensional slicing:** 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).

What is Data Science?

Associated Fields

- **Dimensional slicing:** 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).

What is Data Science?

Associated Fields

- **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.

What is Data Science?

Associated Fields

- **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.

What is Data Science?

Associated Fields

- **Business intelligence:** Business intelligence (BI) 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. BI 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

- 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.
- 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

- Data science is one such paradigm that can handle large volumes with multiple attributes and deploy complex algorithms to search for patterns from data.
- Key motivation for using data science techniques are:
    - Volume
    - Dimensions
    - Complex Questions

Case for Data Science

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.
- As data become more granular, the need to use large volume data to extract information increases

Case for Data Science

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.

Case for Data Science

Dimensions

- 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 data space.

Case for Data Science

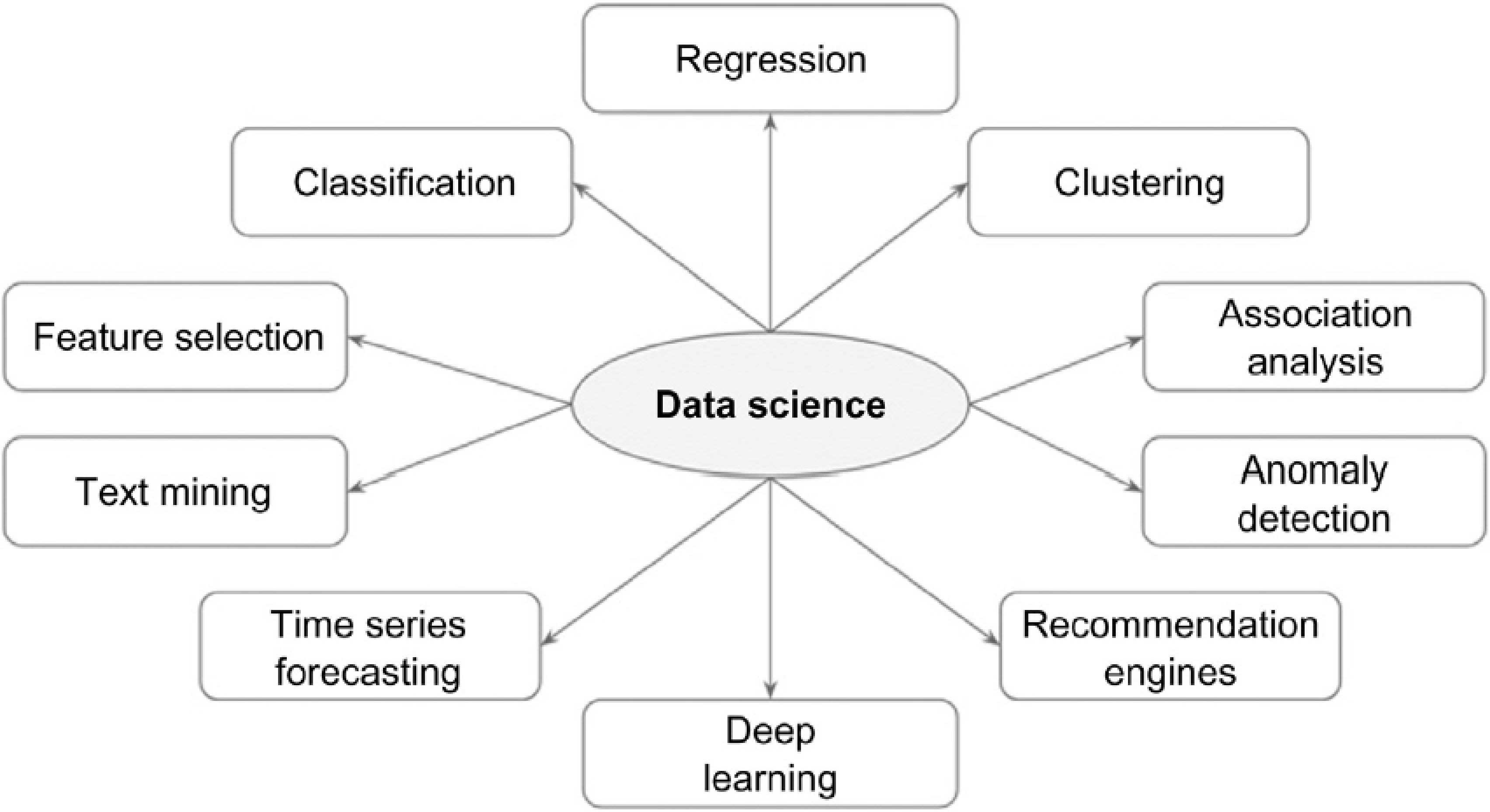
Complex Questions

- 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.

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



Data science tasks

Data Science Classification

- **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).
- **Deep learning** is a more sophisticated artificial neural network that is increasingly used for classification and regression problems.

Data Science Classification

- **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.
- 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.

Data Science Classification

- **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.
- 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

- **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 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.

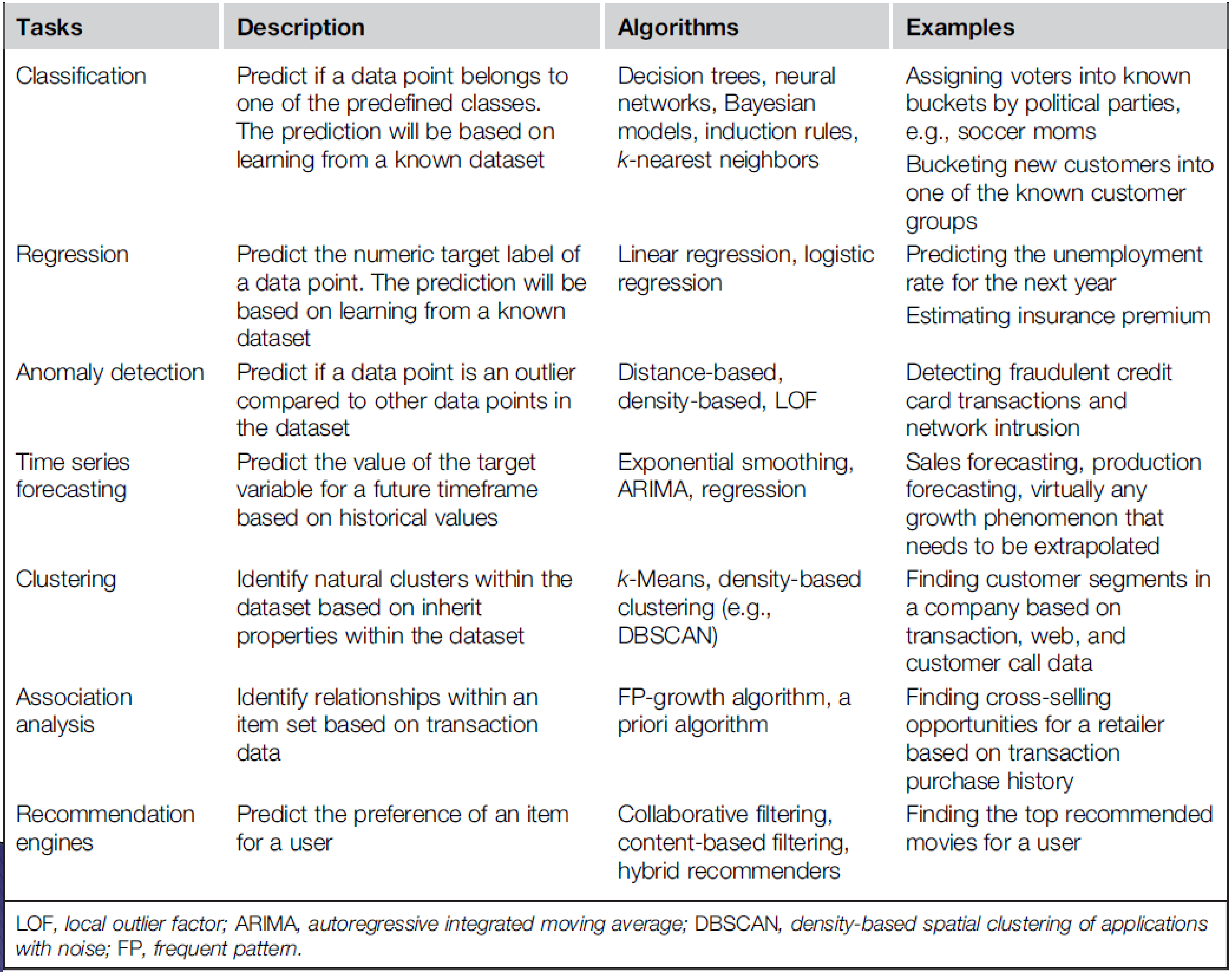
Data Science Algorithms

- 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

- 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



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.*

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.
- The standard data science process involves

(1) understanding the problem,

(2) preparing the data samples,

(3) developing the model,

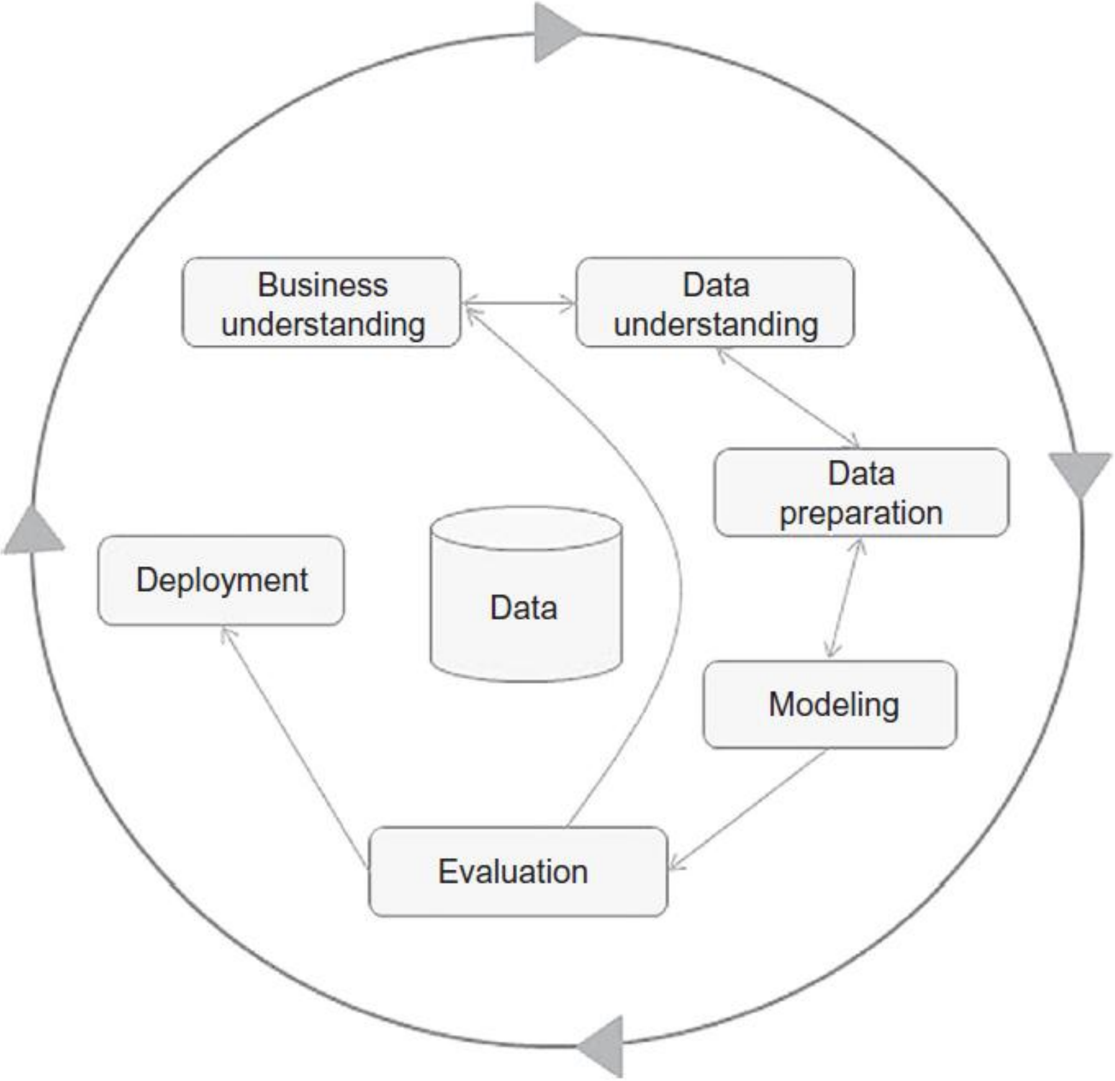
(4) applying the model on a dataset to see how the model may work in the real world, and

(5) deploying and maintaining the models.

Data Science Process

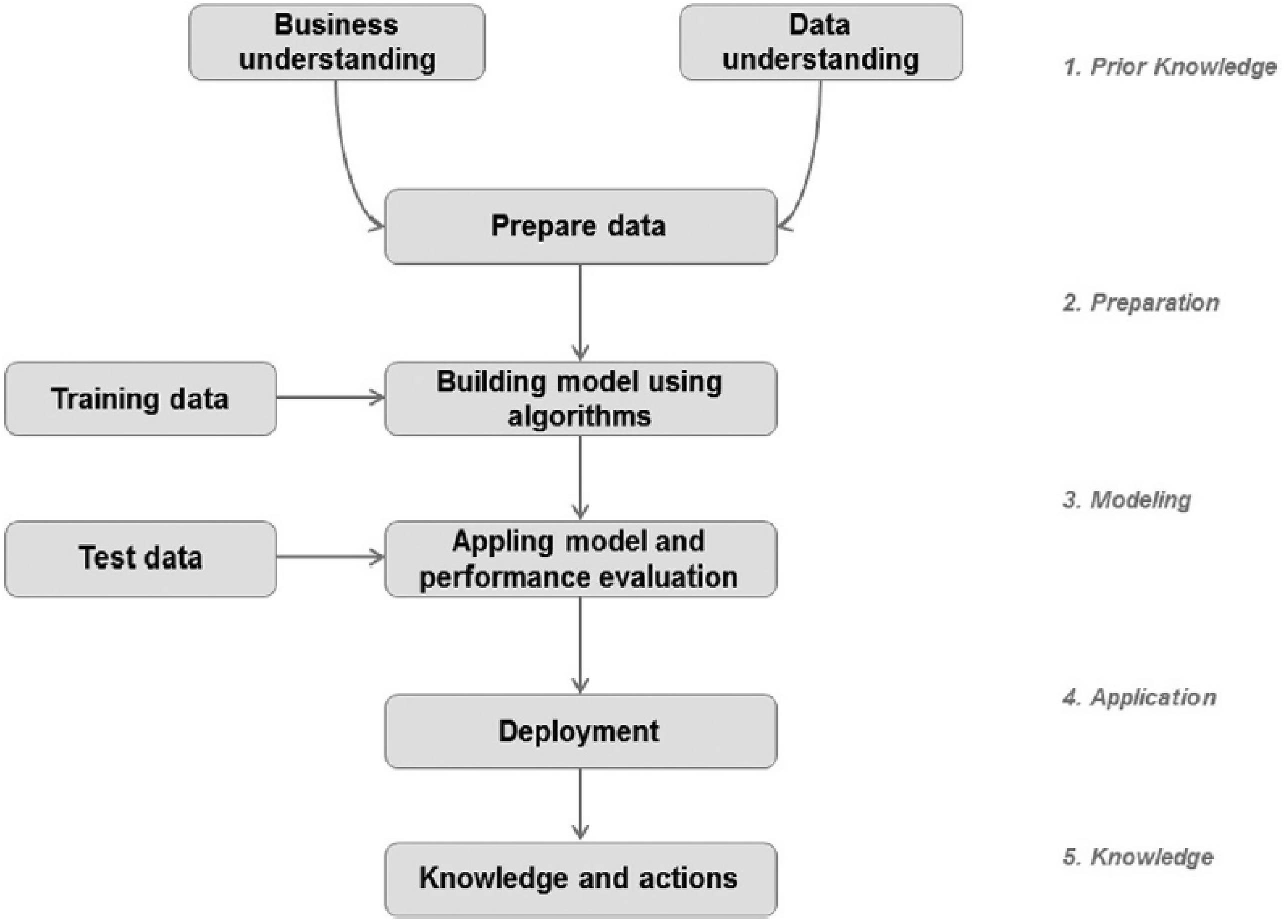
- One of the most popular data science process frameworks is **Cr**oss **I**ndustry **S**tandard **P**rocess for Data Mining (CRISP-DM), which is an acronym for Cross Industry Standard Process for Data Mining

Data Science Process



CRISP data mining framework

Data Science Process



Data Science Process

1. Prior Knowledge

- Prior knowledge refers to information that is already known about a subject
- **Objective**
- The data science process starts with a need for analysis, a question, or a business objective.
- For example, For consumer loan business, the business objective of this hypothetical case is: If the interest rate of past borrowers with a range of credit scores is known, can the interest rate for a new borrower be predicted?

Data Science Process

1. **Prior Knowledge**

- **Subject Area**
- The process of data science uncovers hidden patterns in the dataset by exposing relationships between attributes.
- 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 question. Hence, it is essential to know the subject matter, the context, and the business process generating the data

Data Science Process

1. **Prior Knowledge**

- **Subject Area**
- If the objective is to predict the lending interest rate, then it is important to know how the lending business works, why the prediction matters, what happens after the rate is predicted, what data points can be collected from borrowers, what data points cannot be collected because of the external regulations and the internal policies, what other external factors can affect the interest rate, how to verify the validity of the outcome, and so forth.

Data Science Process

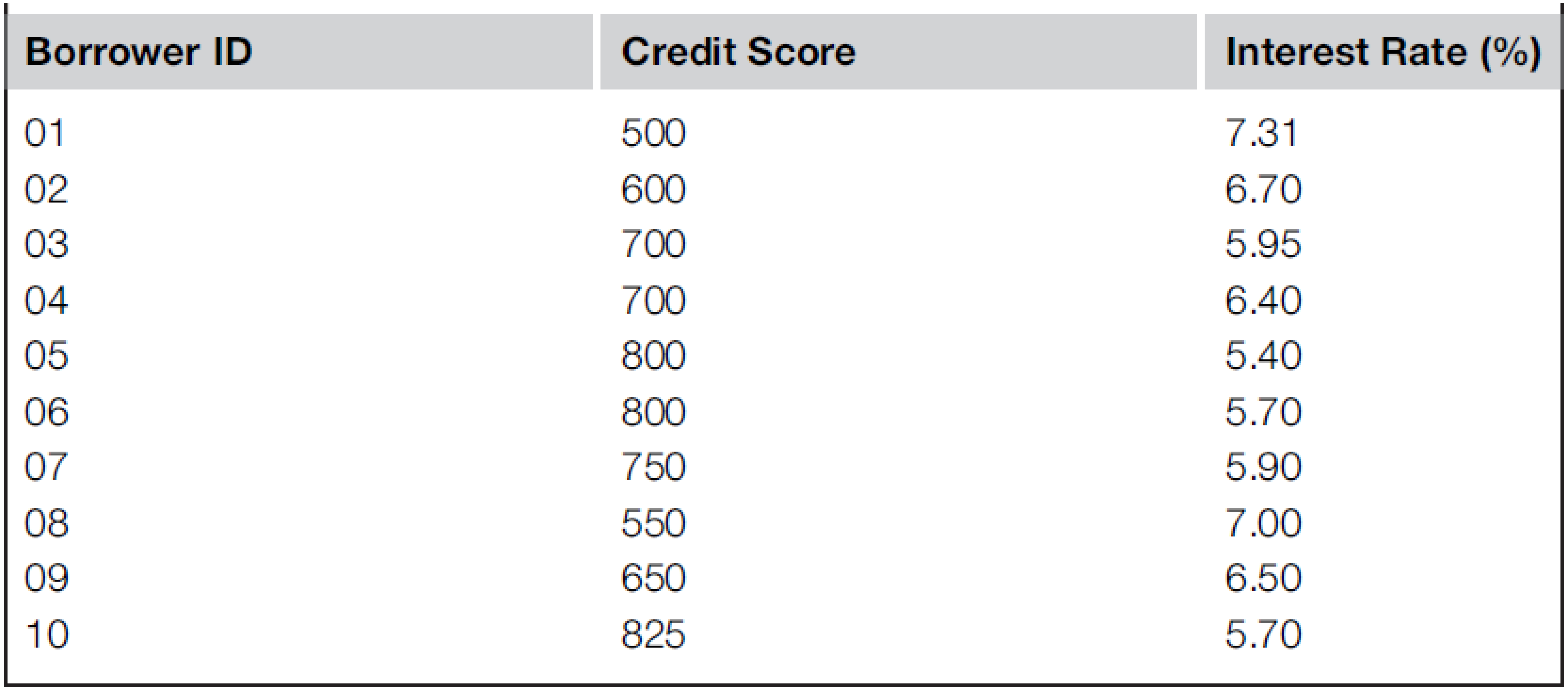
1. Prior Knowledge

- **Data**
- Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process.
- It is critical to recognize that an inferred model is only as good as the data used to create it
- 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.

Data Science Process

1. Prior Knowledge

- **Data**



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

Dataset

Data Science Process

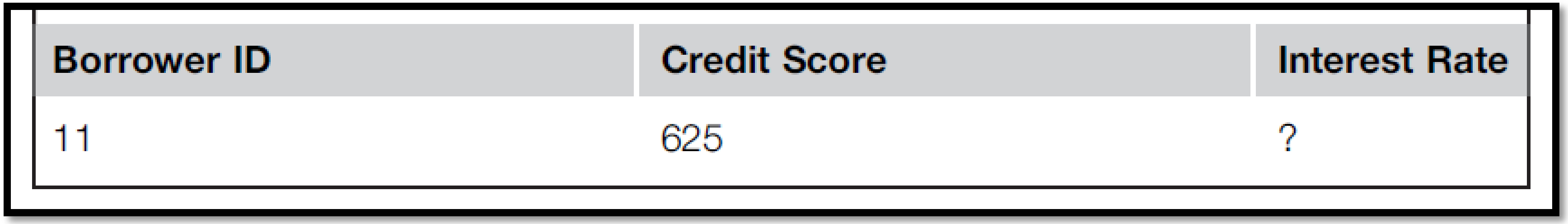
1. Prior Knowledge

- **Data**
- 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

Data Science Process

1. Prior Knowledge

- **Data**
- Identifiers are special attributes that are used for locating or providing context to individual records



Borrower ID	Credit Score	Interest Rate
11	625	?

New Data With Unknown Interest Rate

Data Science Process

1. Prior Knowledge

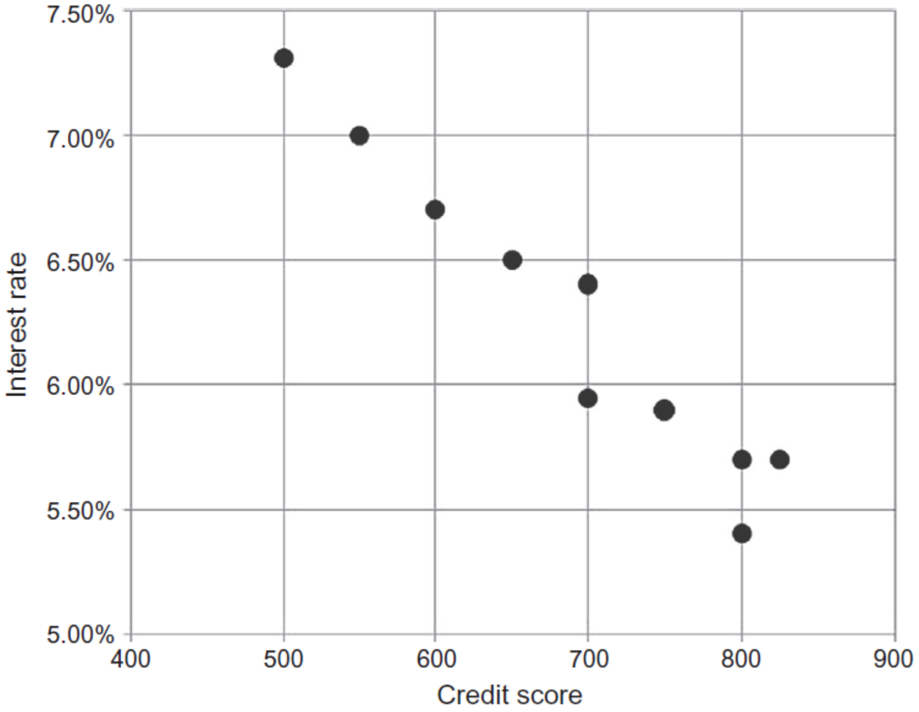
- **Causation Versus Correlation**
- Based on the data, can the credit score of the borrower be predicted based on interest rate? The answer is yes-but it does not make business sense.
- From the existing domain expertise, it is known that credit score influences the loan interest rate. Predicting credit score based on interest rate inverses the direction of the causal relationship.
- 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.

Data Science Process

2. Data Preparation

- **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.

Scatterplot for interest rate dataset



Data Science Process

2. Data Preparation

- **Data quality**
- Data quality is an ongoing concern wherever data is collected, processed, and stored
- 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
- It is critical to check the data using data exploration techniques in addition to using prior knowledge of the data and business before building models.

Data Science Process

2. Data Preparation

- **Missing Values**
- 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 so that the issue can be managed with marginal impact on the later steps in the data science process.
- 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). This method is useful if the missing values occur randomly and the frequency of occurrence is quite rare.

Data Science Process

2. Data Preparation

- **Missing Values**
- Alternatively, to build the representative model, all the data records with missing values or records with poor data quality can be ignored. This method reduces the size of the dataset.
- Some data science algorithms are good at handling records with missing values, while others expect the data preparation step to handle it before the model is inferred.
- For example, k-nearest neighbor (k-NN) algorithm for classification tasks are often robust with missing values. Neural network models for classification tasks do not perform well with missing attributes, and thus, the data preparation step is essential for developing neural network models.

Data Science Process

2. Data Preparation

- **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. 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=400, good=600, excellent=700, etc.

Data Science Process

2. Data Preparation

- **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
- **Transformation**
- In some data science algorithms like k-NN, the input attributes are expected to be numeric and normalized, because the 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.

Data Science Process

2. Data Preparation

- **Transformation**
- For example, consider income (expressed in USD, in thousands) and credit score (in hundreds). The distance calculation will always be dominated by slight variations in income. One solution is to convert the range of income and credit score 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.

Data Science Process

2. Data Preparation

- **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).
- Regardless, the presence of outliers needs to be understood and will require special treatments. The purpose of creating a representative model is to generalize a pattern or a relationship within a dataset and the presence of outliers skews the representativeness of the inferred model. Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

Data Science Process

2. Data Preparation

- **Feature Selection**
- Not all the attributes are equally important or useful in predicting the target.
- The presence of some attributes might be counterproductive.
- 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 due to the curse of dimensionality
- In general, the presence of more detailed information is desired in data science because discovering nuggets of a pattern in the data is one of the attractions of using data science techniques.

Data Science Process

2. Data Preparation

- **Feature Selection**
- 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.
- It leads to a more simplified model and helps to synthesize a more effective explanation of the model.
- **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, such as a similar mean.

Data Science Process

2. Data Preparation

- **Data Sampling**

- Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling.
- In most cases, to gain insights, extract the information, and to build representative predictive models it is sufficient to work with samples.
- Theoretically, the error introduced by sampling impacts the relevancy of the model, but their benefits far outweigh the risks.
- In the build process for data science applications, it is necessary to segment the datasets into training and test samples.

Data Science Process

2. Data Preparation

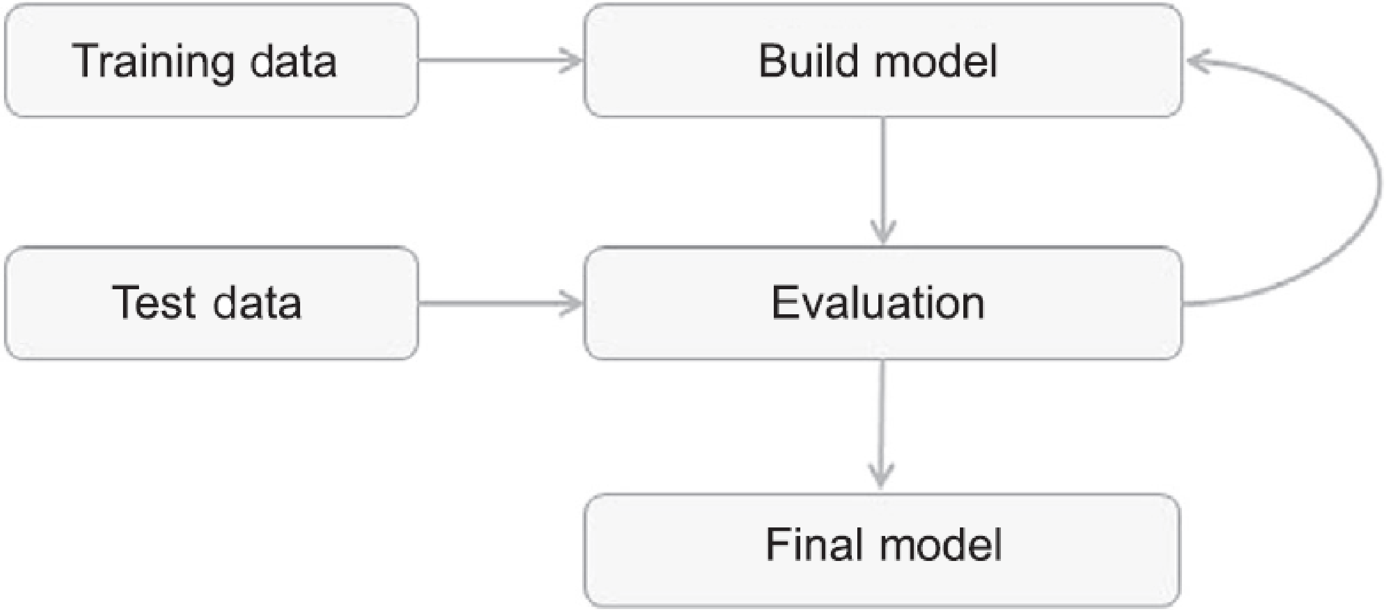
- **Data Sampling**

- Stratified sampling is a process of sampling where each class is equally represented in the sample; this allows the model to focus on the difference between the patterns of each class that is, normal and outlier records.
- In classification applications sampling is used create multiple base models, each developed using a different set of sampled training datasets. These base models are used to build one meta model, called the ensemble model, where the error rate is improved when compared to that of the base models

Data Science Process

3. Modeling

- A model is the abstract representation of the data and the relationships in a given dataset.
- A simple rule of thumb like "*mortgage interest rate reduces with increase in credit score*" is a model



Modeling steps of Predictive data science

Data Science Process

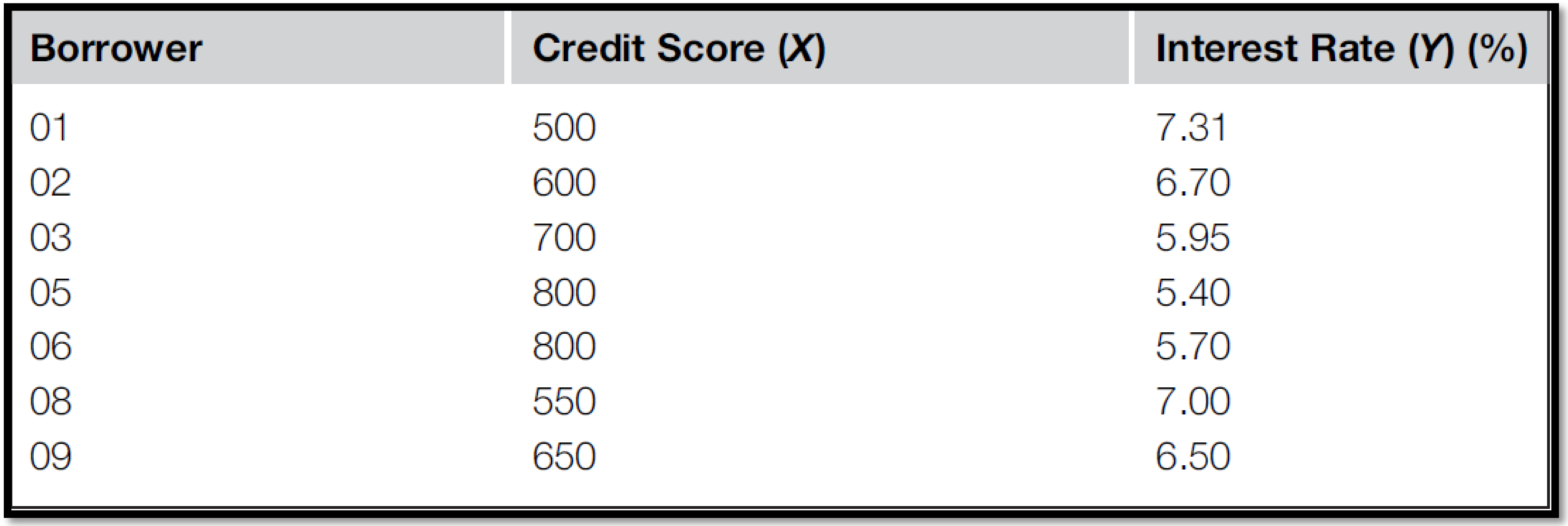
3. Modeling

- Classification and regression tasks are **predictive** techniques because they predict an outcome variable based on one or more input variables.
- Predictive algorithms require a prior known dataset to learn the model
- Association analysis and clustering are descriptive data science techniques where there is no target variable to predict; hence, there is no test dataset.
- Both predictive and descriptive models have an evaluation step.

Data Science Process

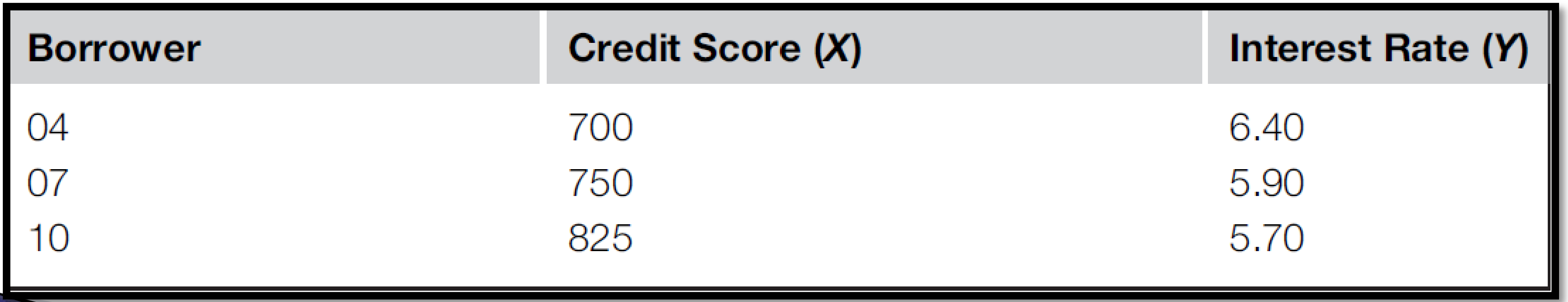
3. Modeling

- **Training and Testing Datasets**



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

Training Dataset



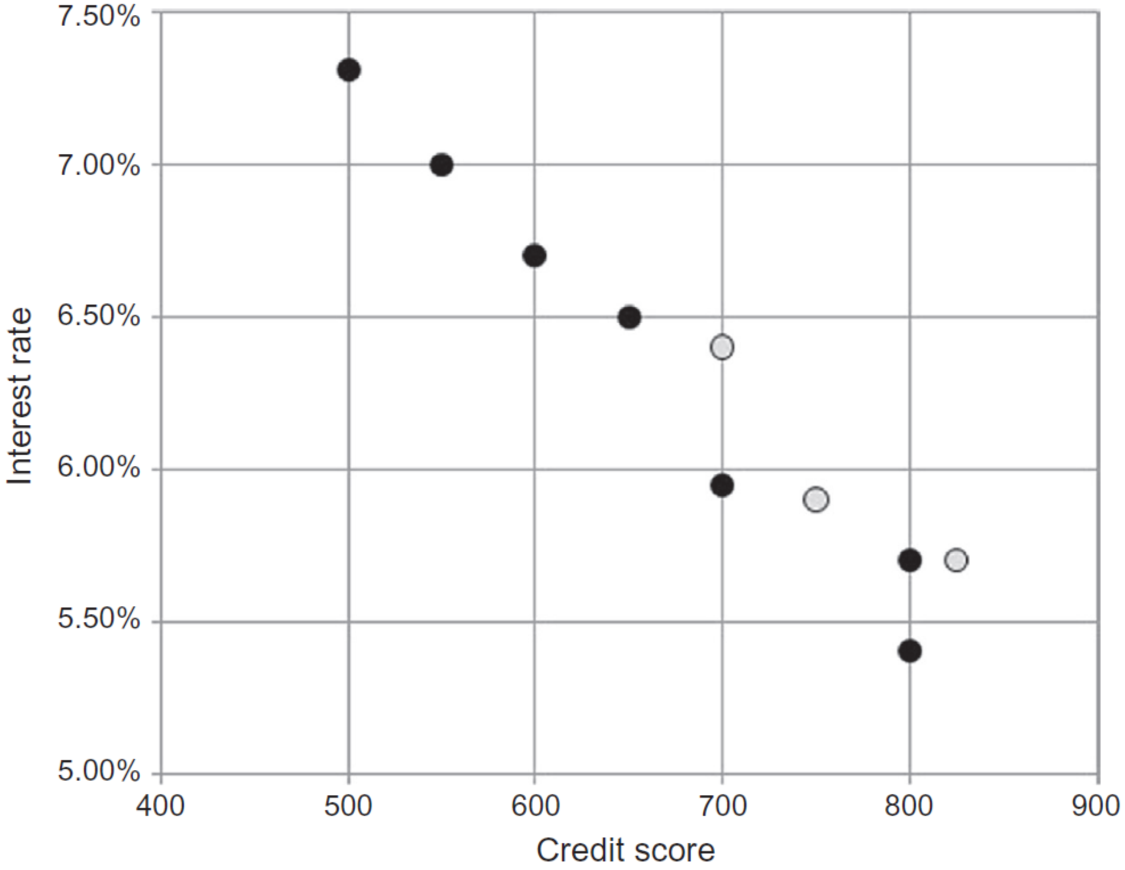
Borrower	Credit Score (*X*)	Interest Rate (*Y*)
04	700	6.40
07	750	5.90
10	825	5.70

Test Dataset

Data Science Process

3. Modeling

- **Training and Testing Datasets**



Scatterplot of training and test data

Data Science Process

3. Modeling

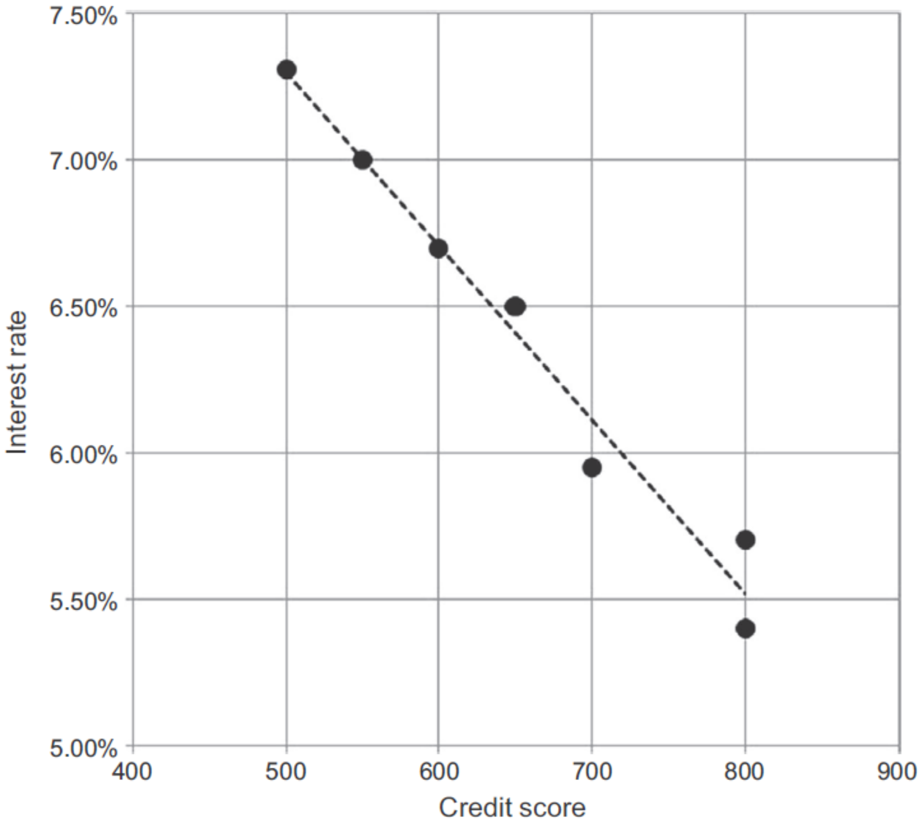
- **Learning Algorithms**
- The business question and the availability of data will dictate what data science task (association, classification, regression, etc.) can to be used
- The objective of simple linear regression can be visualized as fitting a straight line through the data points in a scatterplot
- The line can be expressed as:

  y= a * x + b
- The values of a and b can be found in such a way so as to minimize the sum of the squared residuals of the line

Data Science Process

3. Modeling

- **Learning Algorithms**



Regression model

Data Science Process

3. Modeling

- **Learning Algorithms**
- For eg.

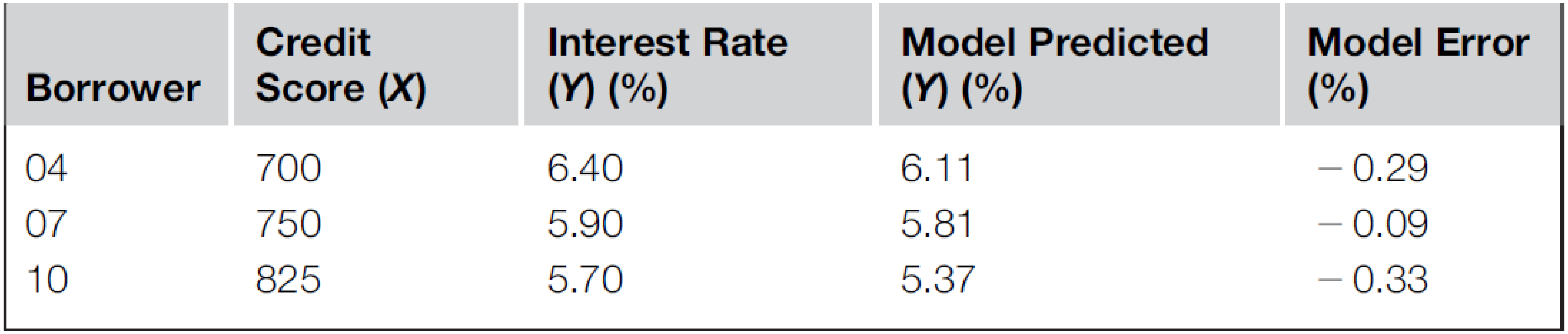
$$y = 0.1 + \frac{6}{100,000}x$$

$$\text{Interest rate} = 10 - \frac{6 \times \text{credit score}}{1000}$$

Data Science Process

3. Modeling

- **Evaluation of the Model**



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

Evaluation of Test Dataset

- A model should not memorize and output the same values that are 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

Data Science Process

3. Modeling

- **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. The approach seeks the wisdom of crowds in making a 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

Data Science Process

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; and

(7) validated the model against the test dataset.

This model can now be used to predict the interest rate of new borrowers by integrating it in the actual loan approval process.

Data Science Process

4. Application

- Deployment is the stage at which the model becomes production ready or live. In business applications, the results of the data science process have to be assimilated into the business process—usually in software applications.
- The model deployment stage has to deal with: assessing model readiness, technical integration, response time, model maintenance, and assimilation
- **Production Readiness**
- The production readiness part of the deployment determines the critical qualities required for the deployment objective.
- Consider two business use cases: determining whether a consumer qualifies for a loan and determining the groupings of customers for an enterprise by marketing function.

Data Science Process

4. Application

- **Production 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.

Data Science Process

4. Application

- **Production Readiness**
- Segmenting customers based on their relationship with the company is a thoughtful process where signals from various customer interactions are collected.
- Based on the patterns, similar customers are put in cohorts and campaign strategies are devised to best engage the customer.
- For this application, batch processed, time lagged data would suffice.
- The critical quality in this application is the ability to find unique patterns amongst customers, not the response time of the model.
- The business application informs the choices that need to be made in the data preparation and modeling steps.

Data Science Process

4. Application

- **Technical Integration**
- Currently, it is quite common to use data science automation tools or coding using 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. This allows the analyst to focus on the data, business logic, and exploring patterns from the data.
- The models created by data science tools can be ported to production applications by utilizing the Predictive Model Markup Language (PMML) or by invoking data science tools in the production application
- PMML provides a portable and consistent format of model description which can be read by most data science tools.

Data Science Process

4. Application

- **Technical Integration**
- This allows the flexibility for practitioners to develop the model with one tool (e.g., RapidMiner) 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. These models are represented by simple equations and the “if-then” rule, hence, they can be ported easily to most programming languages.

Data Science Process

4. Application

- **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.

Data Science Process

4. Application

- **Model Refresh**
- The key criterion for the ongoing relevance of the model is the representativeness of the dataset it is processing.
- 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.

Data Science Process

5. Knowledge

- The data science process starts with prior knowledge and ends with posterior knowledge, which is the incremental insight gained.
- As with any quantitative technique, the data science process can bring up spurious irrelevant patterns from the dataset.
- Not all discovered patterns lead to incremental knowledge.
- It is up to the practitioner to invalidate the irrelevant patterns and identify the meaningful information.