Machine Learning

Data Science Process

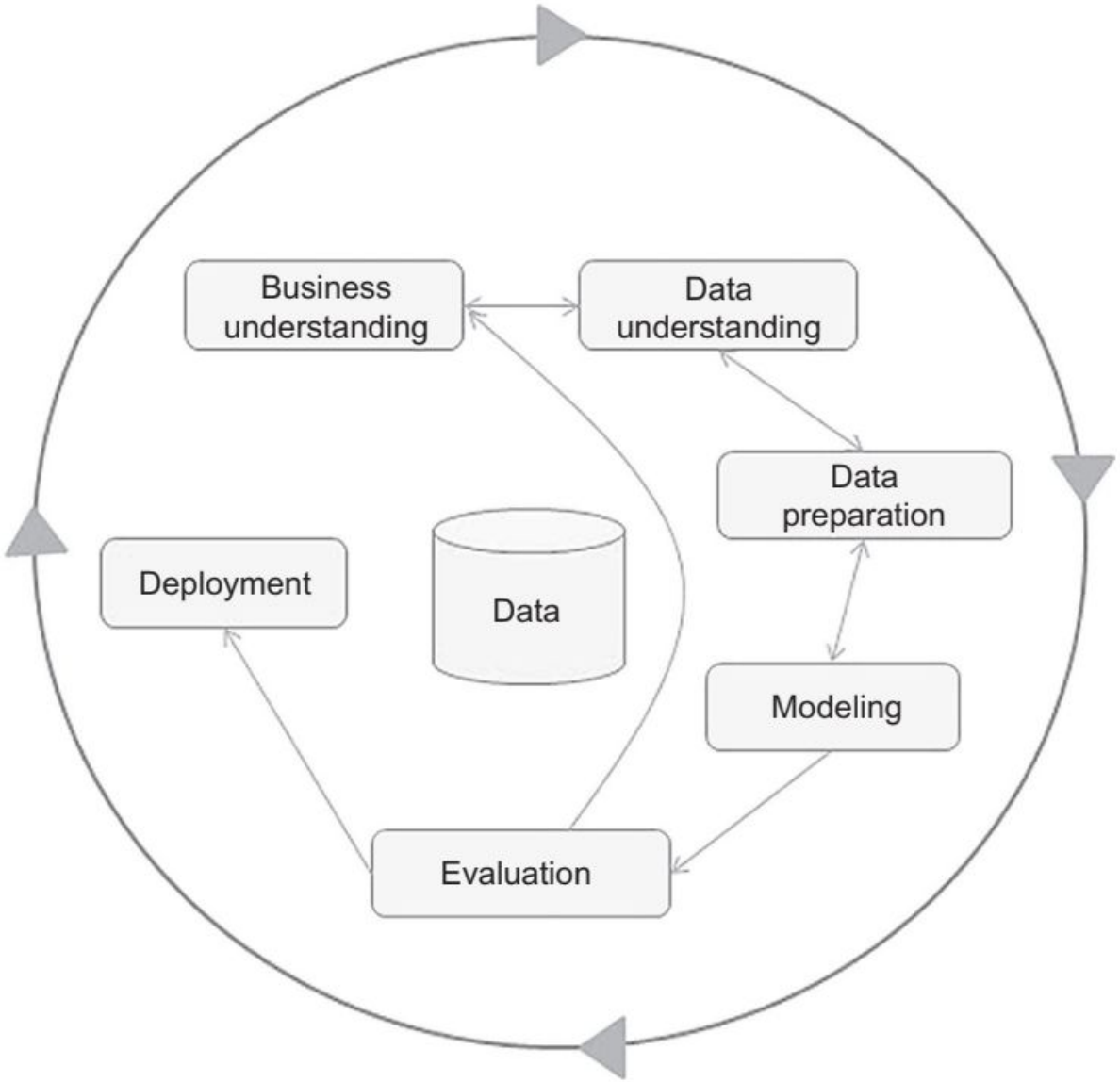
Data Science Process

- 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).
- This framework was developed by a consortium of companies involved in data mining.
- The CRISP-DM process is the most widely adopted framework for developing data science solutions.

CRISP Data Mining Framework



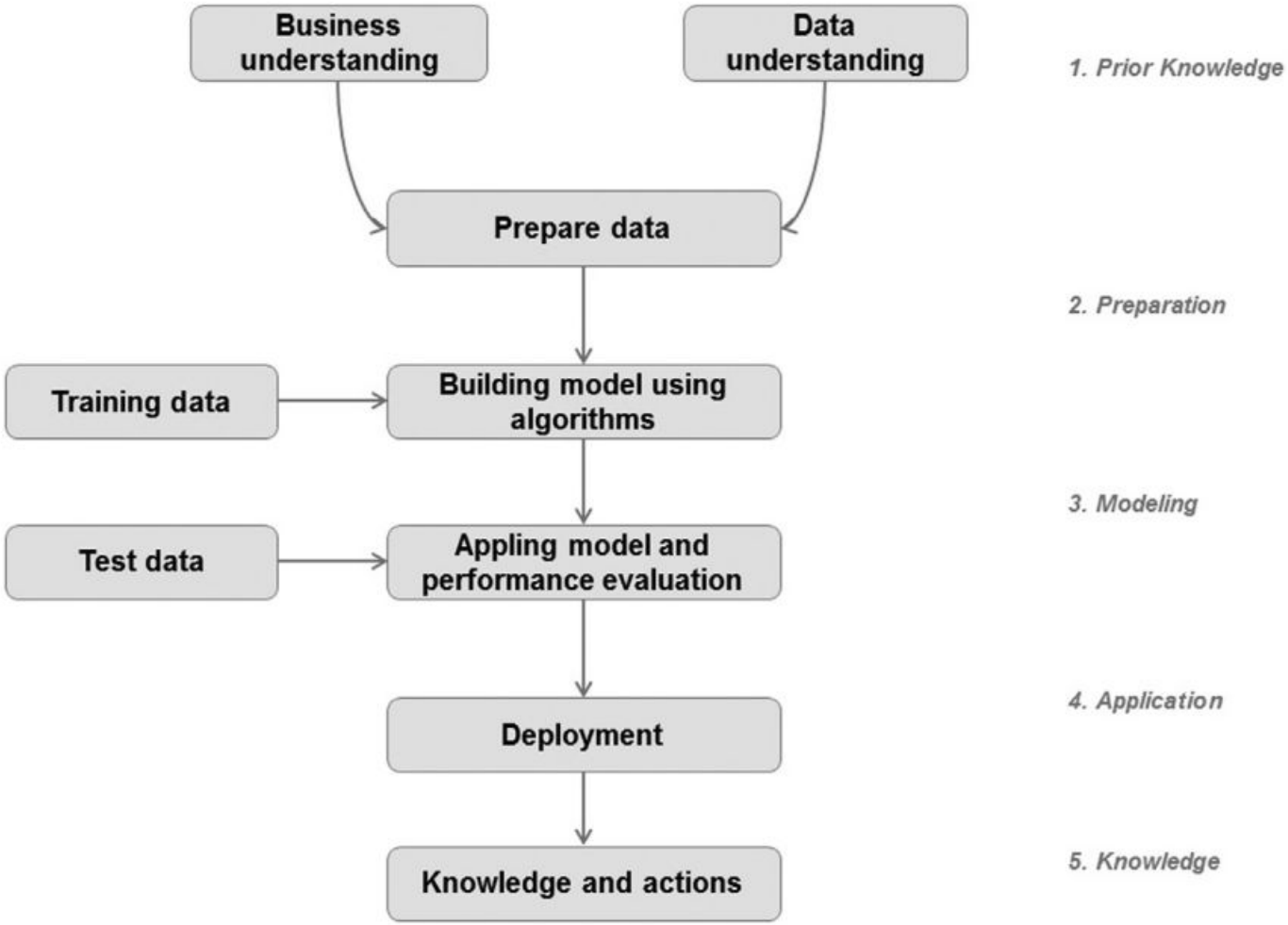
Data Science Process Framework

- 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 used in the knowledge discovery in databases process.

Data Science Process Framework

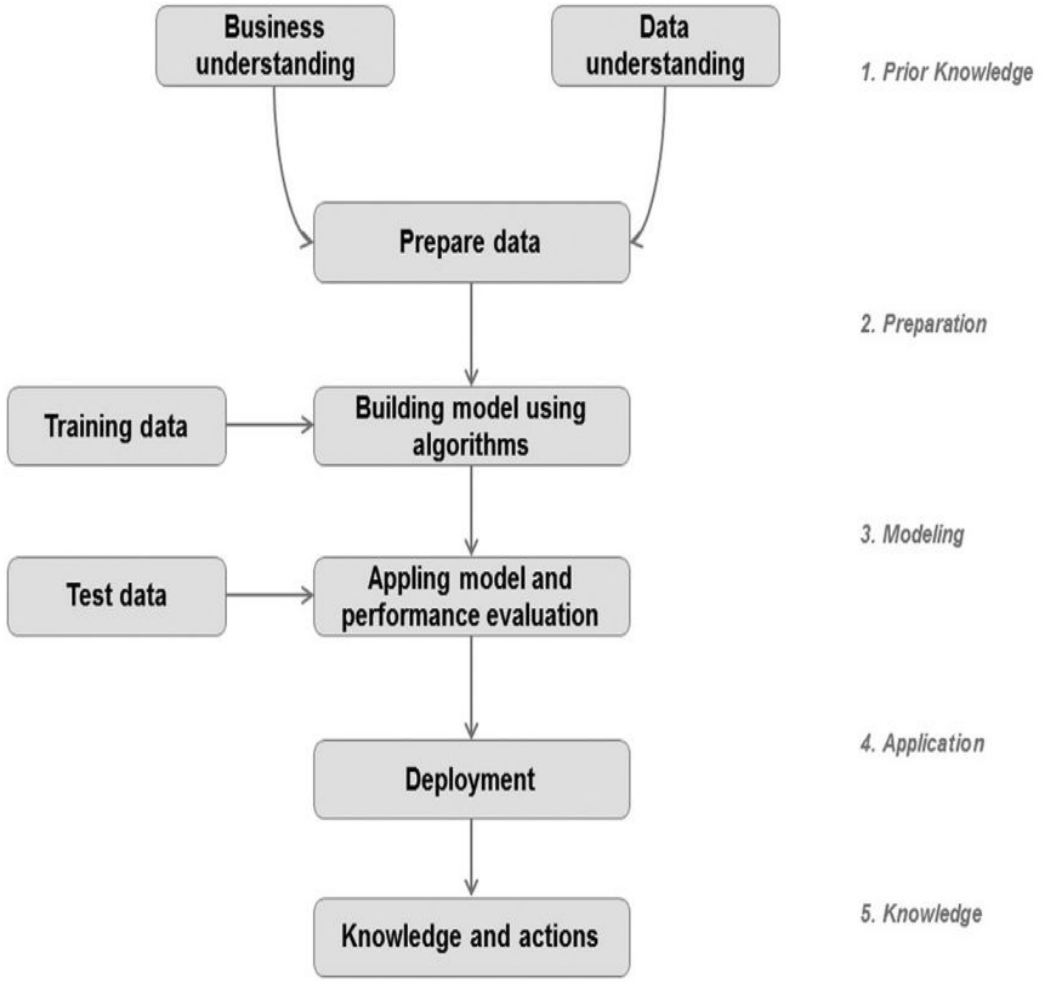
- The steps within the data science process
  - are not linear
  - have to undergo many loops, go back and forth between steps
  - at times go back to the first step to redefine the data science problem statement.

Data Science Process



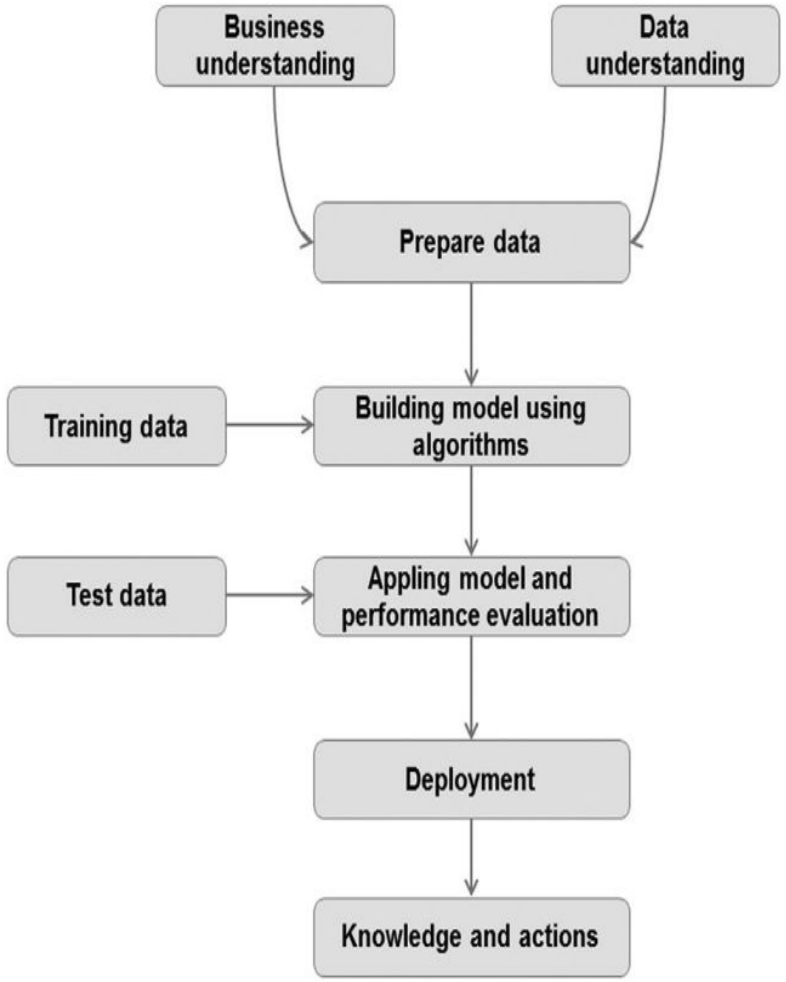
Data Science Process

- The fundamental objective is to address the analysis question.
- The problem at hand could be
  - a segmentation of customers,
  - a prediction of climate patterns, or
  - a simple data exploration.



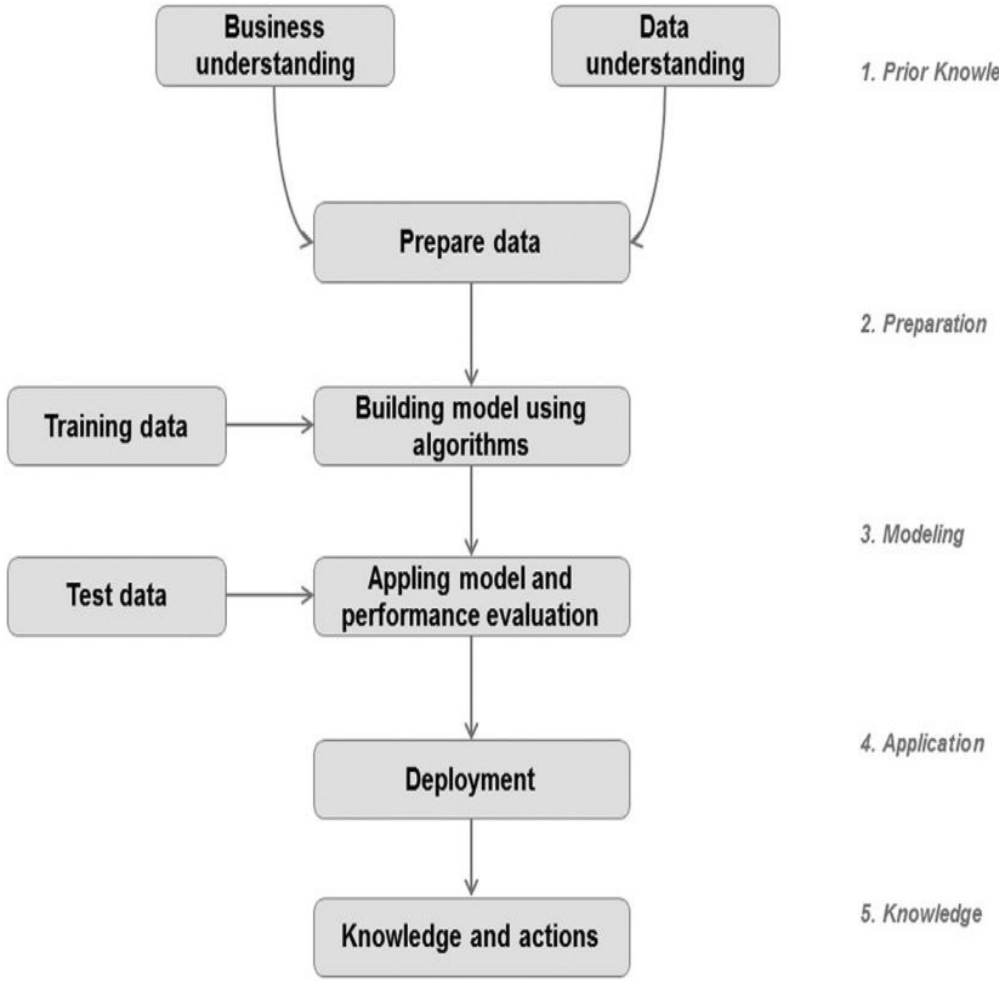
Data Science Process

- The learning algorithm used to solve the business question could be
  - a decision tree,
  - an artificial neural network, or
  - a scatterplot.



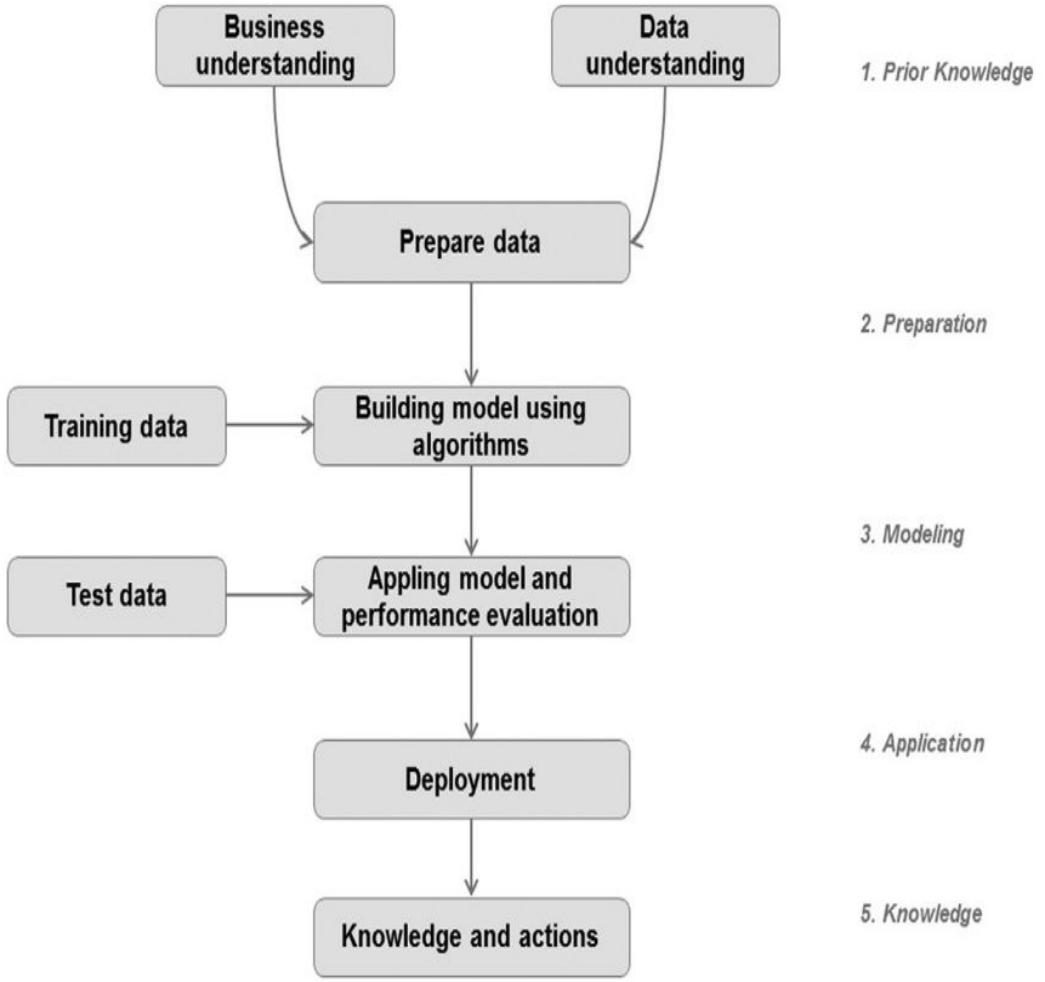
Data Science Process

- Perhaps the most visible and discussed part of data science is the third step: modeling.
- It is the process of building representative models that can be inferred from the sample dataset,
- which can be used for
  - either predicting (predictive modeling)
  - or describing the underlying pattern in the data (descriptive or explanatory modeling).



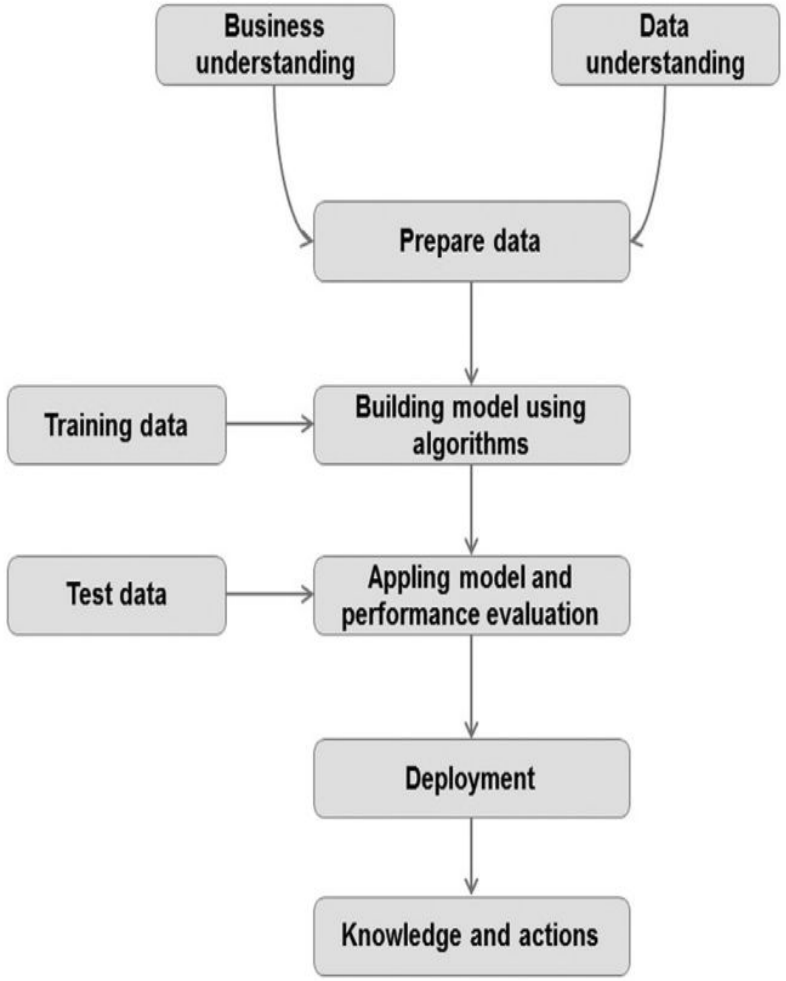
Data Science Process

- The most time-consuming part of the overall data science process is not the model building part, but the preparation of data, followed by data and business understanding.
- There are many data science tools, both open source and commercial, available on the market that can automate the model building.



Data Science Process

- Crucial to the success of the data science process:
  - asking the right business question,
  - gaining in-depth business understanding,
  - sourcing and preparing the data for the data science task,
  - mitigating implementation considerations,
  - integrating the model into the business process, and,
  - most useful of all, gaining knowledge from the dataset



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.
- Involves:
  - Objective
  - Subject area
  - Data
  - Causation vs Correlation

Prior Knowledge

- Objective
  - The data science process starts with a need for analysis, a question, or a business objective.

Prior Knowledge

- Subject Area
  - it is essential to know the subject matter, the context, and the business process generating the data.

Prior Knowledge

- Data
  - Understanding how the data is collected, stored, transformed, reported, and used is essential to the data science process.
- Some of the terminology used in the data science process are discussed:
  - 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.
  - 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.
  - Identifiers are special attributes that are used for locating or providing context to individual records.

Prior Knowledge

Table 2.1 Dataset

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

Prior Knowledge

Table 2.2 New Data With Unknown Interest Rate

11	625	?

Prior Knowledge

- 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.

Data Preparation

- 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.
- If the data is in any other format, the data would need to be transformed by applying pivot, type conversion, join, or transpose functions, etc., to condition the data into the required structure.

Data Preparation

- Data preparation involves:
  - Data Exploration
  - Data Quality
  - Missing Values
  - Data Types and Conversation
  - Transformation
  - Outliers
  - Feature Selection
  - Data Sampling

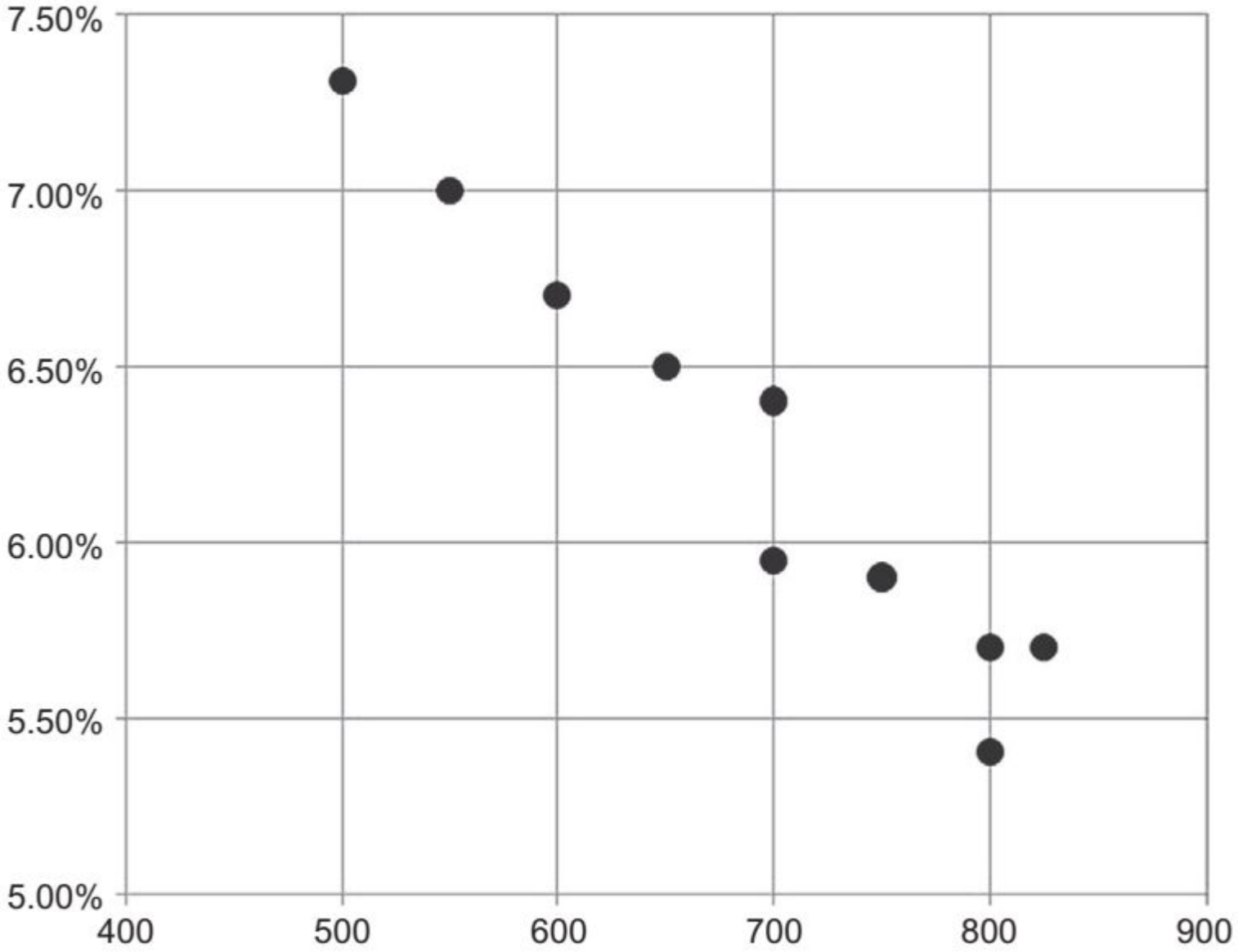
Data Preparation

- Data exploration
  - also known as exploratory data analysis,
  - provides a set of simple tools to achieve basic understanding of the data.

Data Preparation

- Data exploration
  - approaches involve computing descriptive statistics and visualization of data.
  - 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.
  - On the other hand, a visual plot of data points provides an instant grasp of all the data points condensed into one chart.

Data Preparation



Data Preparation

- Data quality
  - Data quality is an ongoing concern wherever data is collected, processed, and stored.

Data Preparation

- Missing Values
  - One of the most common data quality issues is that some records have missing attribute values.
  - There are several different mitigation methods to deal with this problem, but each method has pros and cons.

Data Preparation

- Data Types and Conversion
  - The attributes in a dataset can be of different types, such as
    - continuous numeric,
    - integer numeric, or
    - categorical.

Data Preparation

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

Data Preparation

- Outliers
  - Outliers are anomalies in a given dataset.
  - Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

Data Preparation

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

Data Preparation

- 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.

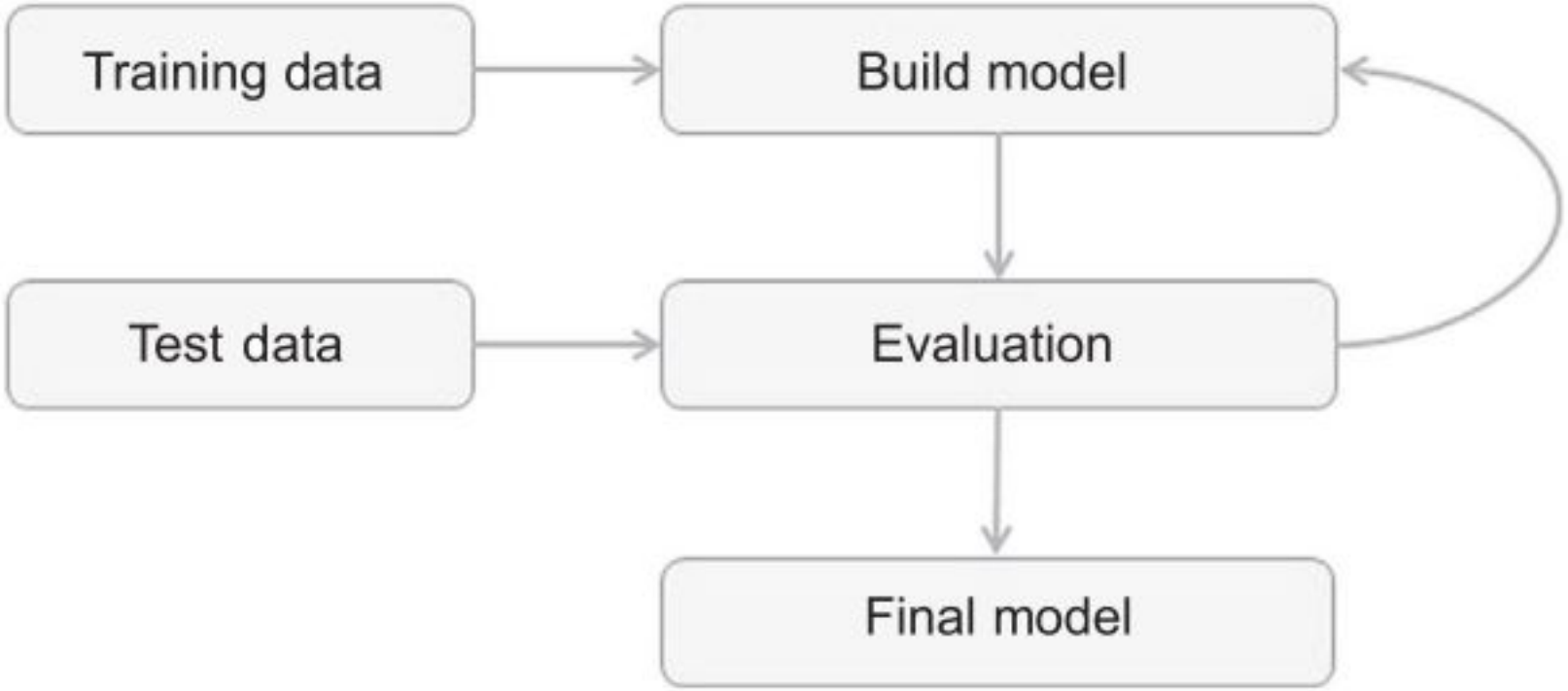
Modeling

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

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.

Modeling Steps



Modeling

- Association analysis and clustering are descriptive data science techniques where there is no target variable to predict; hence, there is no test dataset.
- However, both predictive and descriptive models have an evaluation step.

Training and Testing Datasets

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

Training and Testing Datasets

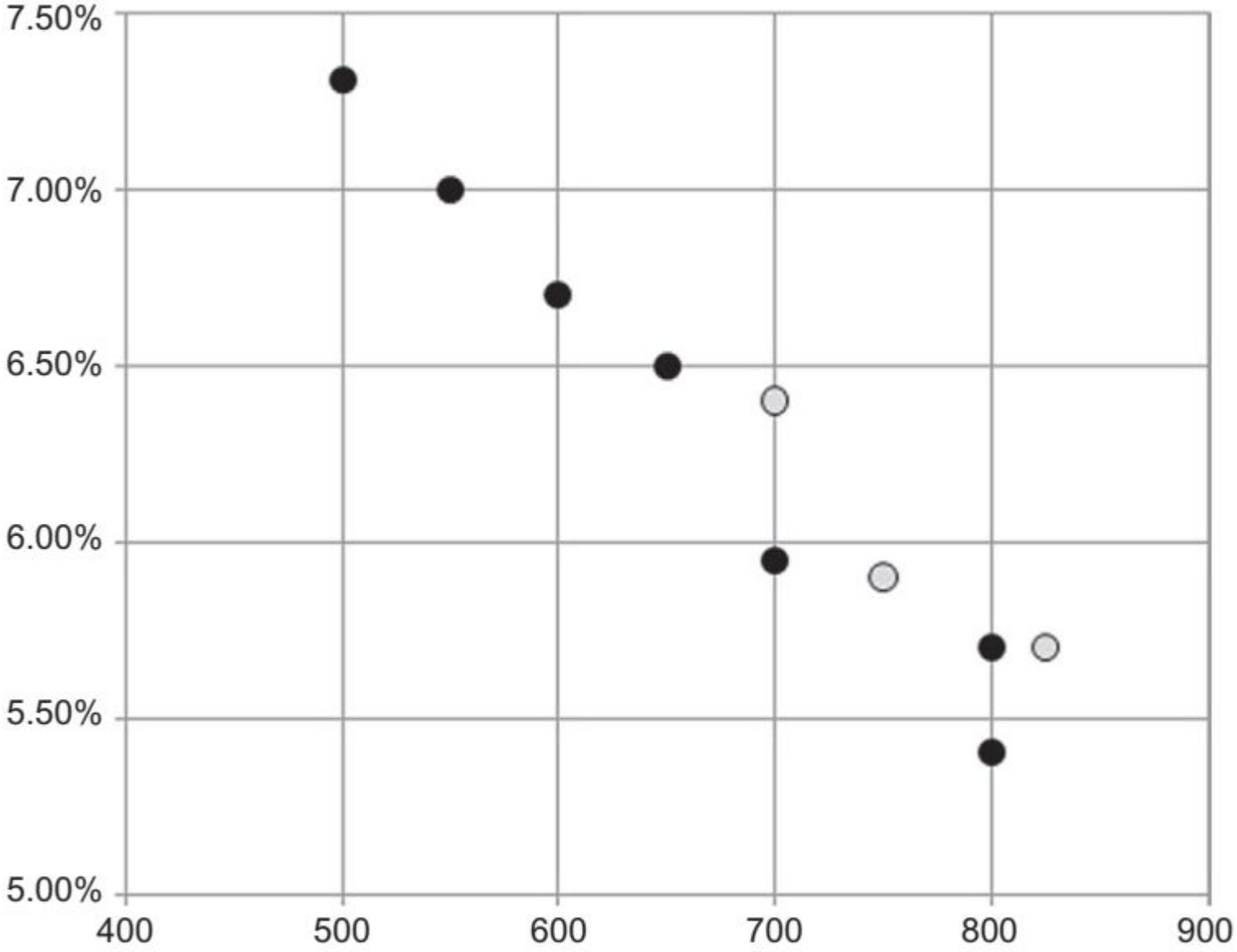
Table 2.3 Training Dataset

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

04	700	6.40
07	750	5.90
10	825	5.70

Training and Testing Datasets



Learning Algorithms

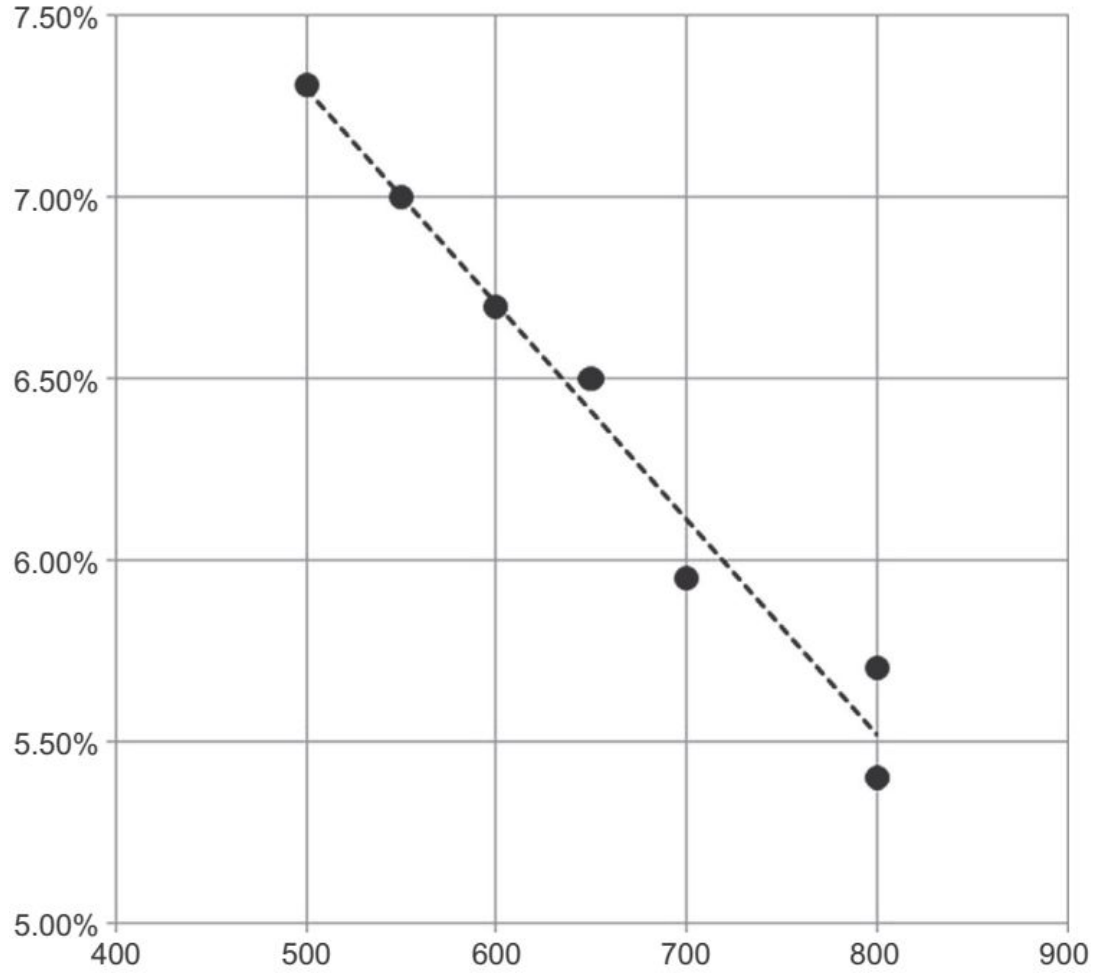
- The business question and the availability of data will dictate what data science task (association, classification, regression, etc.,) can to be used.
- The practitioner determines the appropriate data science algorithm within the chosen category.
- For example, within a classification task many algorithms can be chosen from: decision trees, rule induction, neural networks, Bayesian models, k-NN, etc.
- It is not uncommon to use multiple data science tasks and algorithms to solve a business question.

Learning Algorithms

- The shown prediction is a regression problem. A simple linear regression technique will be used to model and generalize the relationship.
- The training set of seven records is used to create the model and the test set of three records is used to evaluate the validity of the model.

Learning Algorithms

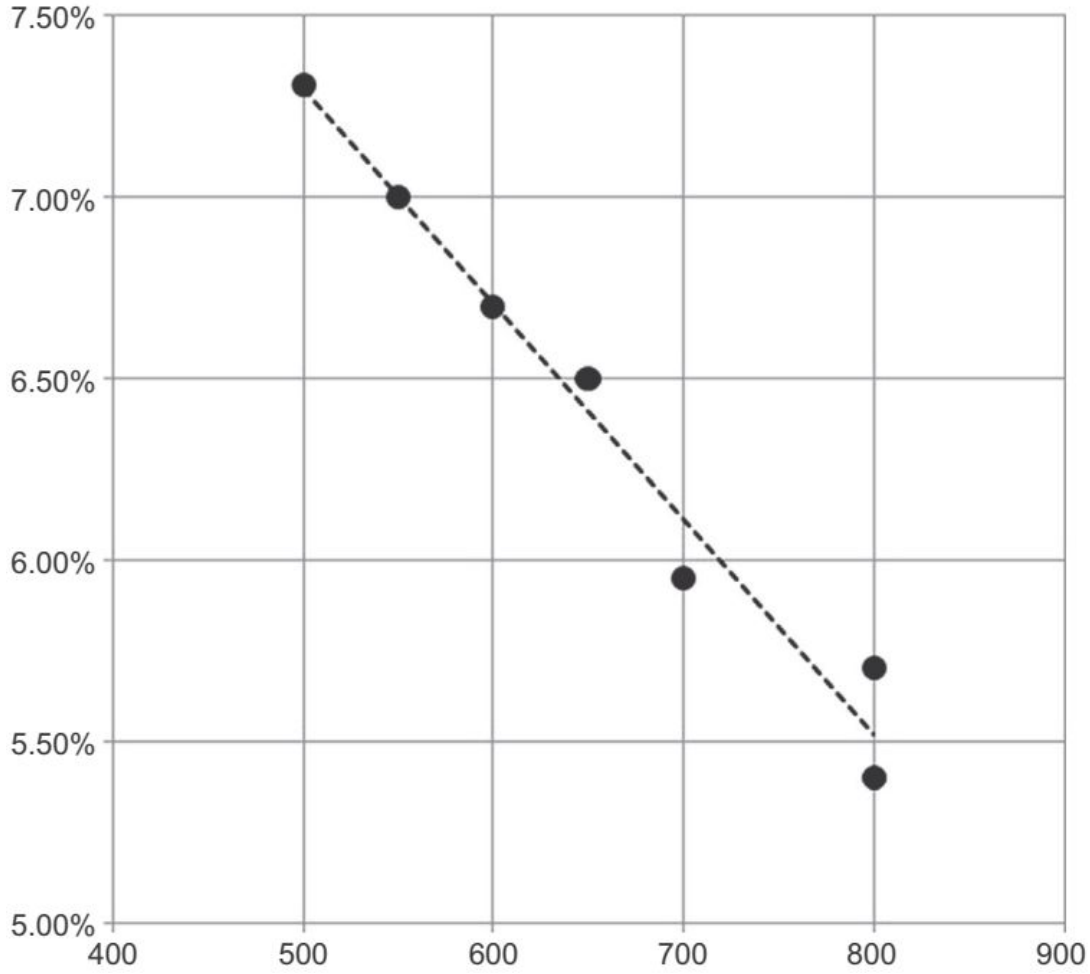
- The objective of simple linear regression can be visualized as fitting a straight line through the data points in a scatterplot.



Learning Algorithms

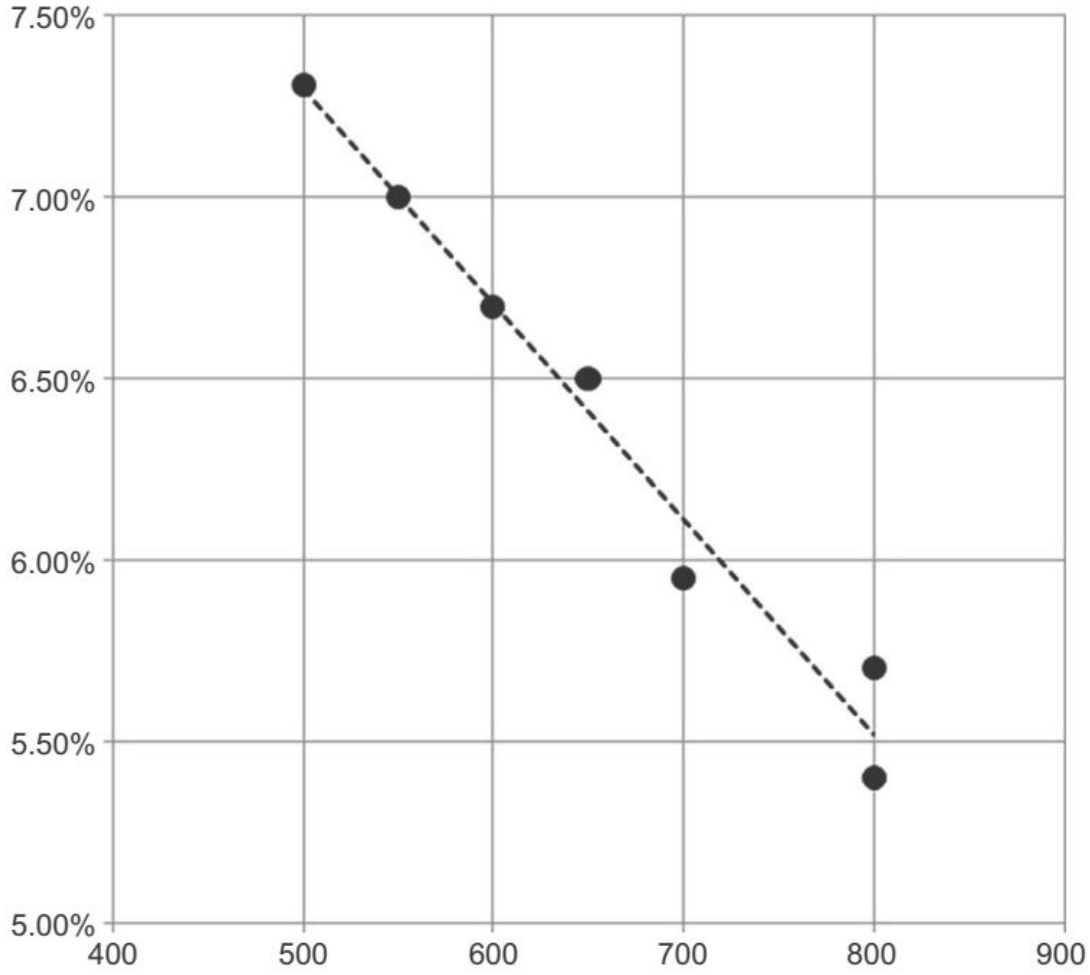
- The line has to be built in such a way that the sum of the squared distance from the data points to the line is minimal. The line can be expressed as:
- Y = a * x + b

- 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 so as to minimize the sum of the squared residuals of the line.



Learning Algorithms

- The line shown serves as a model to predict the outcome of new unlabeled datasets.



Evaluation of the Model

- The model generated in the form of an equation is generalized and synthesized from seven training records.
- The score in the equation can be substituted to see if the model estimates the rate for each of the seven training records.
- The estimation may not be exactly the same as the values in the training records.
- 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.

Evaluation of the Model

Table 2.5 Evaluation of Test Dataset

			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 the Model

- Provides the three testing records where the value of the rate is known; these records were not used to build the model.
- The actual value of the 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.
- The error rate can be used to compare this model with other models developed using different algorithms like neural networks or Bayesian models, etc.

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.

Processes at end of 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.