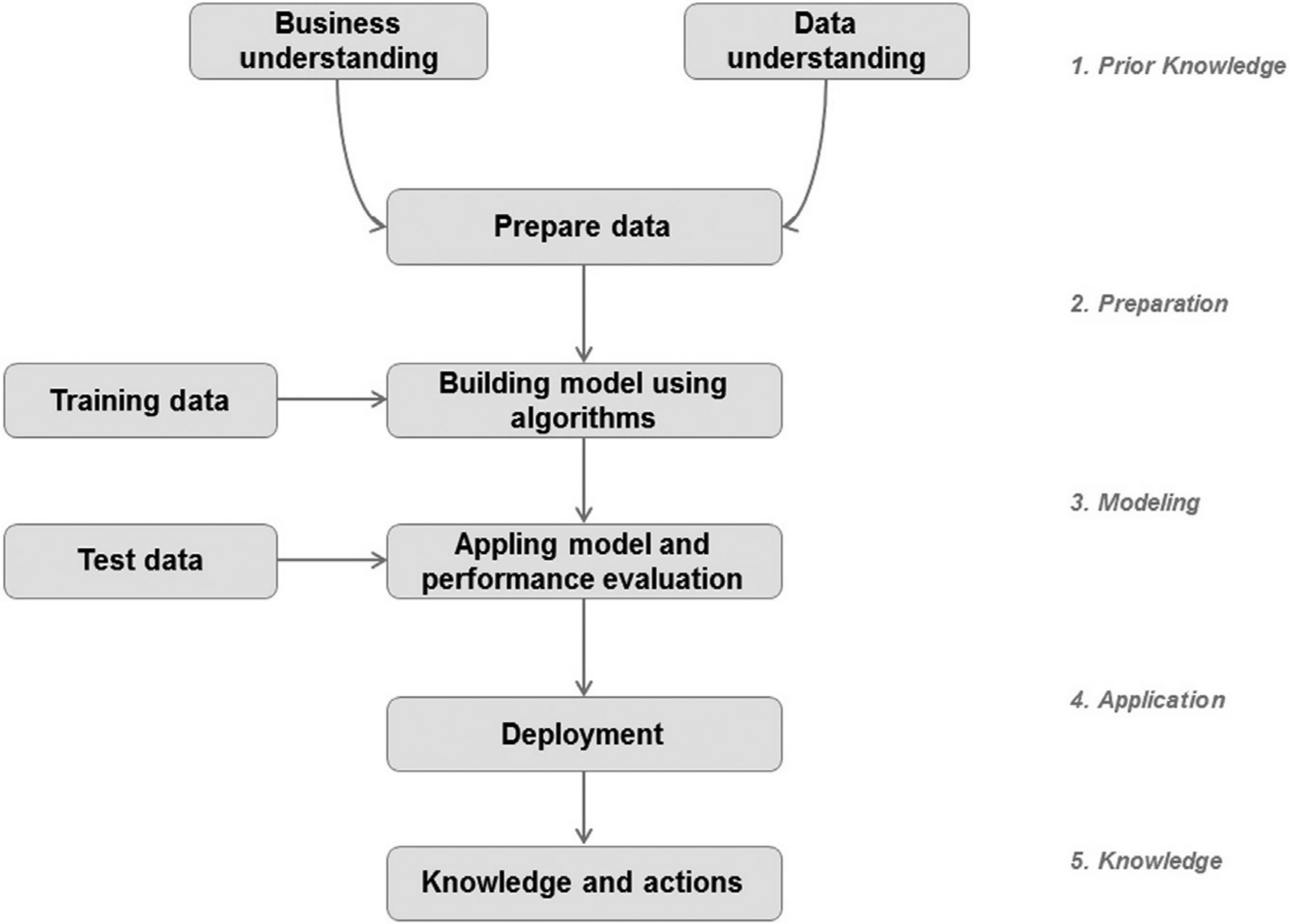
DATA SCIENCE PROCESS



1.PRIOR KNOWLEDGE

- Prior knowledge refers to information that is already known about a subject.
- This helps to define
  - what problem is being solved
  - how it fits in the business context
  - What data is needed in order to solve the problem

1.1 Objective

- The data science process starts with a need for analysis, a question, or a business objective
- Consumer loan business- Example
- 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?

1.2 Subject Area

- The process of data science uncovers hidden patterns in the dataset by exposing relationships between attributes.
- The false or spurious signals are a major problem in the data science process.
- It is up to the practitioner to analyse the exposed patterns and accept the ones that are valid and relevant to the situation/problem

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.
- This part of the step surveys all the data available to answer the business question and narrows down the new data that need to be sourced.
- There are quite a range of factors to consider: quality of the data, quantity of data, availability of data, gaps in data, does lack of data compel the practitioner to change the business question.

Data Contd...

- The objective of this step is to come up with a dataset to answer the business question through the data science process.
- It is critical to recognize that an inferred model is only as good as the data used to create it.

Terminologies

- A data set(example set) is a collection of data with a defined structure.This structure is also sometimes referred to as a “data frame”.
- A data point (record, object or example) is a single instance in the dataset.Each row in the Table is a data point.Each instance contains the same structure as the dataset.
- An attribute (feature, input, dimension, variable, or predictor)is a singleproperty of the dataset.Each column in the Table is an attribute.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

1.4 Causation Versus Correlation

- Correlation means there is a relationship or pattern between the values of two variables.
- Causation means that one event causes another event to occur. Causation can only be determined from an appropriately designed experiment

Correlation Contd..

- There are three ways to describe the correlation between variables.
- Positive correlation: As xxx increases, yyy increases.
- Negative correlation: As xxx increases, yyy decreases.
- No correlation: As xxx increases, yyy stays about the same or has no clear pattern.
-

Causation

- Ice Cream Sales and Shark Attacks
- There was a study that was done that found a strong correlation between the ice cream sales and number of shark attacks for a number of beaches that were sampled.

  Conclusion: Increasing ice cream sales causes more shark attacks ??

  Better explanation: The confounding variable is temperature. Warmer temperatures cause ice cream sales to go up. Warmer temperatures also bring more people to the beaches, increasing the chances of shark attacks. This is known as common response, where two variables (ice cream sales and shark attacks) are both responding to changes in some third variable (temperature).

2 .DATA PREPARATION

- Preparing the dataset to suit a data science task is the most time-consuming part of the process.
- It is extremely rare that datasets are available in the form required by the data science algorithms.
- 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 type conversion, join, or transpose functions, etc., to condition the data into the required structure.

2.1 Data Exploration

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

2.2 Data Quality

- What if an attribute value has a recorded value of 900 (beyond the theoretical limit) ?
- If there was a data entry error? Errors in data will impact the representativeness of the model.
- Organizations use cleansing, and transformation techniques to improve and manage the quality of the data and store them in companywide repositories called *data warehouses.*
- Data sourced from well-maintained data warehouses have higher quality, as there are proper controls in place to ensure a level of data accuracy for new and existing data.
- The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bounds, standardization of attribute values, substitution of missing values, etc.
- In a given dataset , how does one know if the data are accurate?
- Data quality is an ongoing concern wherever data is collected, processed, and stored.

2.3 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.
- The first step of managing missing values is to understand the reason behind why the values are missing.
- Tracking the the data source can lead to the identification of systemic issues during data capture or errors in data transformation.
- 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.

2.4 Data Types and Conversion

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

2.6 Outliers

- Outliers are anomalies in a given dataset
- Outliers may occur because of correct data capture (few people with income in tens of millions) or erroneous data capture (human height as 1.73 cm instead of 1.73 m).
- Detecting outliers may be the primary purpose of some data science applications, like fraud or intrusion detection.

2.7 Feature Selection

- Many data science problems involve a dataset with hundreds to thousands of attributes.
- Not all the attributes are equally important or useful in predicting the target
- 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
- Reducing the number of attributes, without significant loss in the performance of the model, is called feature selection. It leads to a more simplifiedmodel and helps to synthesize a more effective explanation of the model.

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 or modeling.
- The sample data serve as a representative of the original dataset.
- Sampling reduces the amount of data that need to be processed and speeds up the build process of the modeling.
- Stratified sampling is a process of sampling where each class is equally represented in the sample

3. MODELING

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

3.1Training and Testing Datasets

- The modeling step creates a representative model inferred 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 of thumb is two-thirds of the data are to be used as training and one-third as a test dataset.

3.2 Learning Algorithms

- The business question and the availability of data will dictate what data science task (association, classification, regression, etc.,) can to be used.
- The practitioner determines the appropriate data science algorithm within the chosen category.
- For example, within a classification task many algorithms can be chosen from: decision trees, neural networks, Bayesian models, k-NN, etc.

3.3 Evaluation of the Model

- The model is tested with known records ; where these records were not used to build the model.
- The actual value of the oputput 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