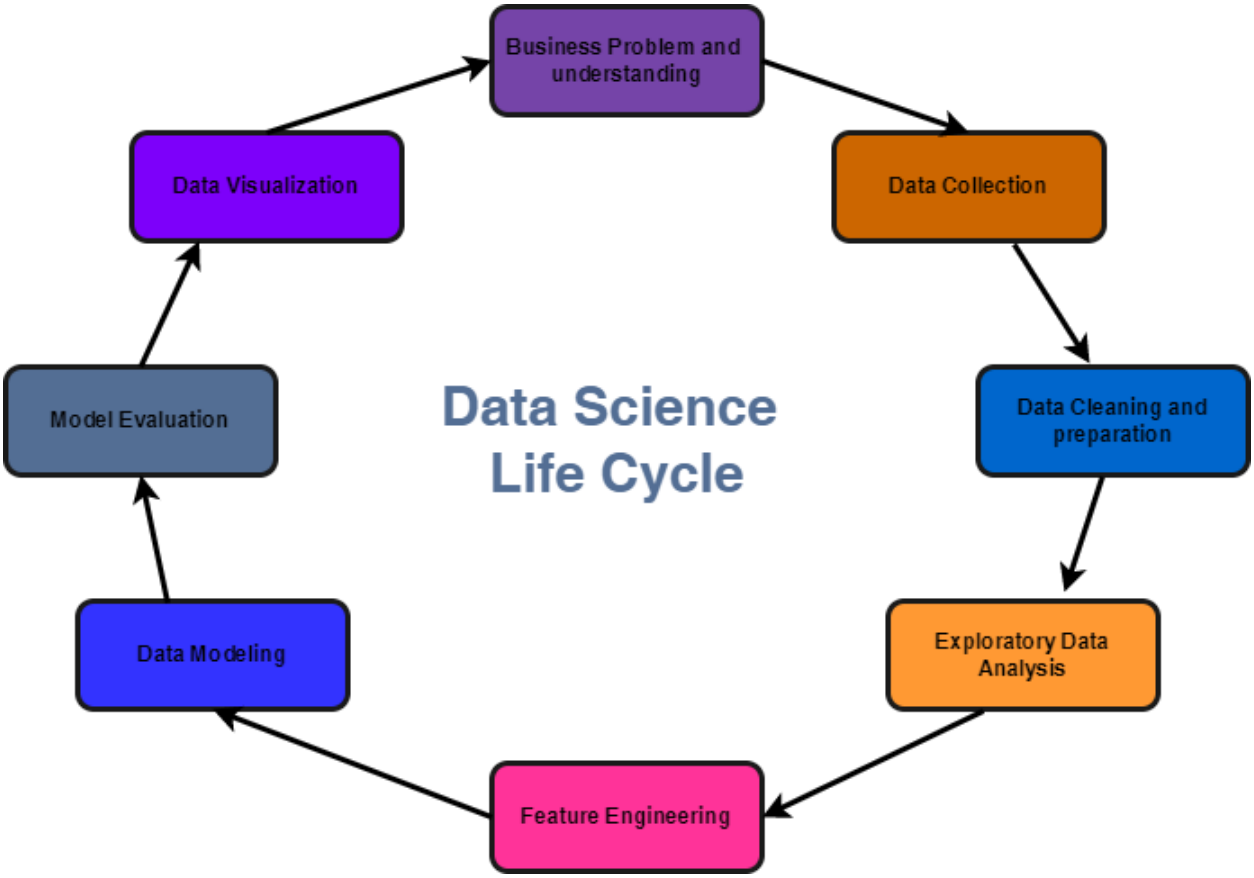
Data Science

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains.The goal of "Stastical Computing with R" is to learn the most important tools in R that will allow in data science.



Steps of Data Science Life Cycle

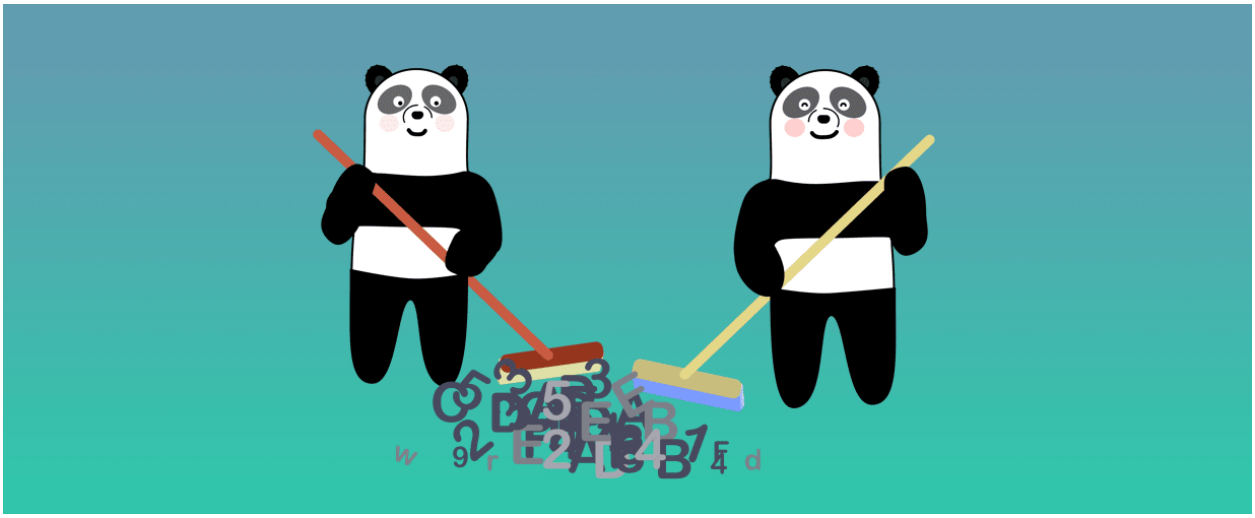
1. Business Problem and Understanding

In order to build the business model, it is necessary to understand the business problem i.e the problem faced by customer,requirement and what to predict from the model.In such case, it is necessary to consult with the domain expert and understand the underlying problem faced by the system.Here problem and goal of project is define.

2. Data Collection

After understanding the business problem, collect data from different sources such as web scraping, third party API,etc. The collected data must be according to the business requirement. Let, If the problem is of medical domain then the feature collected must be age, gender, sex,etc.

3. Data Cleaning and Preparation



Data collections are raw data; it may be structured or unstructured. The data may be null, may contain outliers, or in any format.So transform, remove null values, remove outliers.Format the data as the business requirement.

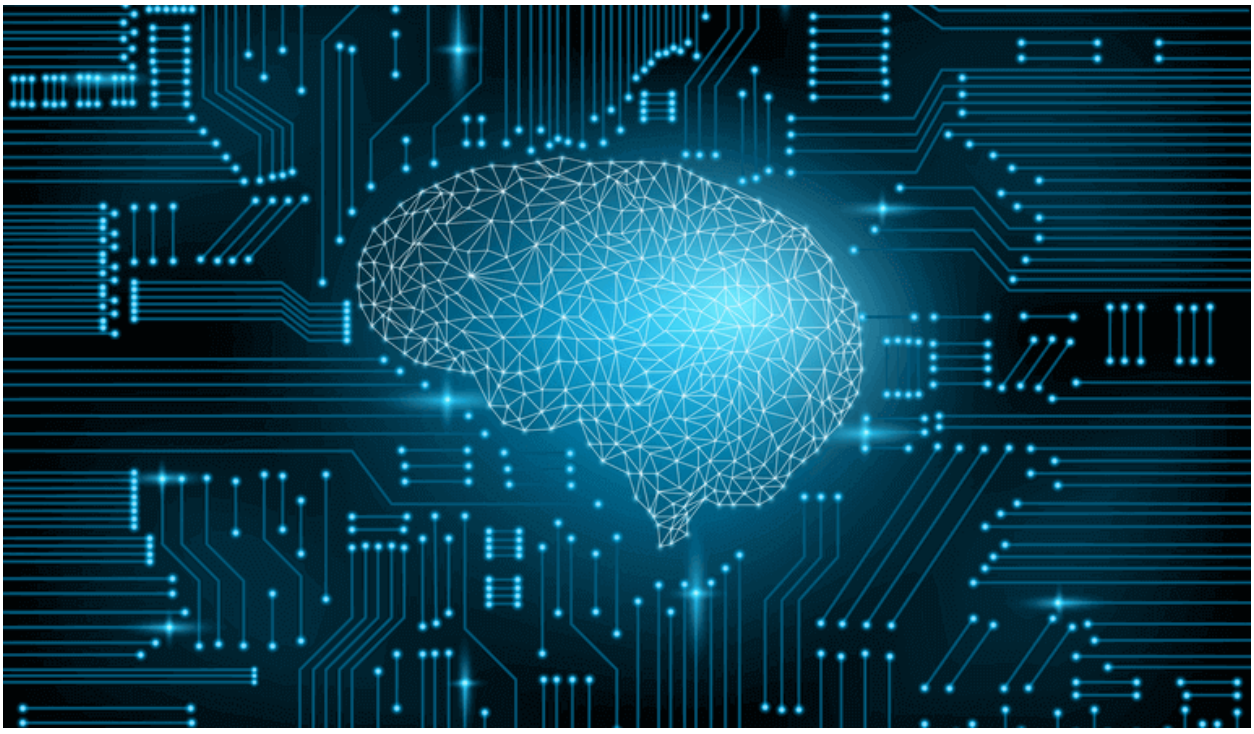
4. Exploratory Data Analysis(EDA)

After data cleaning, explore the feature in the dataset with the help of visualization, statistics and other technique.It is used to analyze the data to summarize the feature by plotting or using statistical methods.

5. Feature Engineering

In this phase the important features are selected as the business requirement by removing the noise from the dataset.

6. Data Modeling

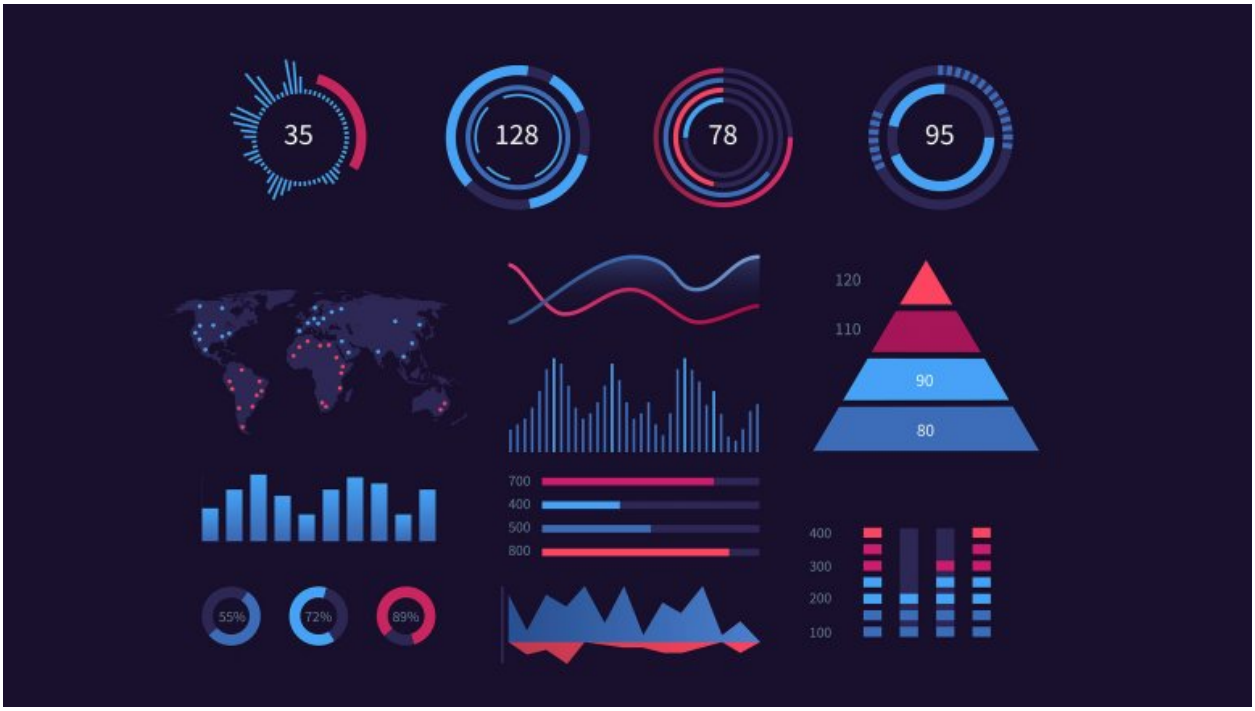


After an important feature is selected then the data are trained to the machine to find pattern.Here different supervised and unsupervised learning methods are applied.Supervised method consist of regression and classification problem and unsupervised method consist like clustering problem.

7. Model Evaluation

After model training, model evaluation helps to find the best model that represents our data and how well the chosen model will work in future.In some cases the model evaluated from training data may not be accepted as there may be the overfit model. So, different methods are applied to remove overfit models such as cross-Validation.

8. Data Visualization



After the best fitted model is selected the result is visualized using different tools like Power BI in the form of graphs to communicate with the client or the end user. It also provides clear and actionable insights to take business decisions.