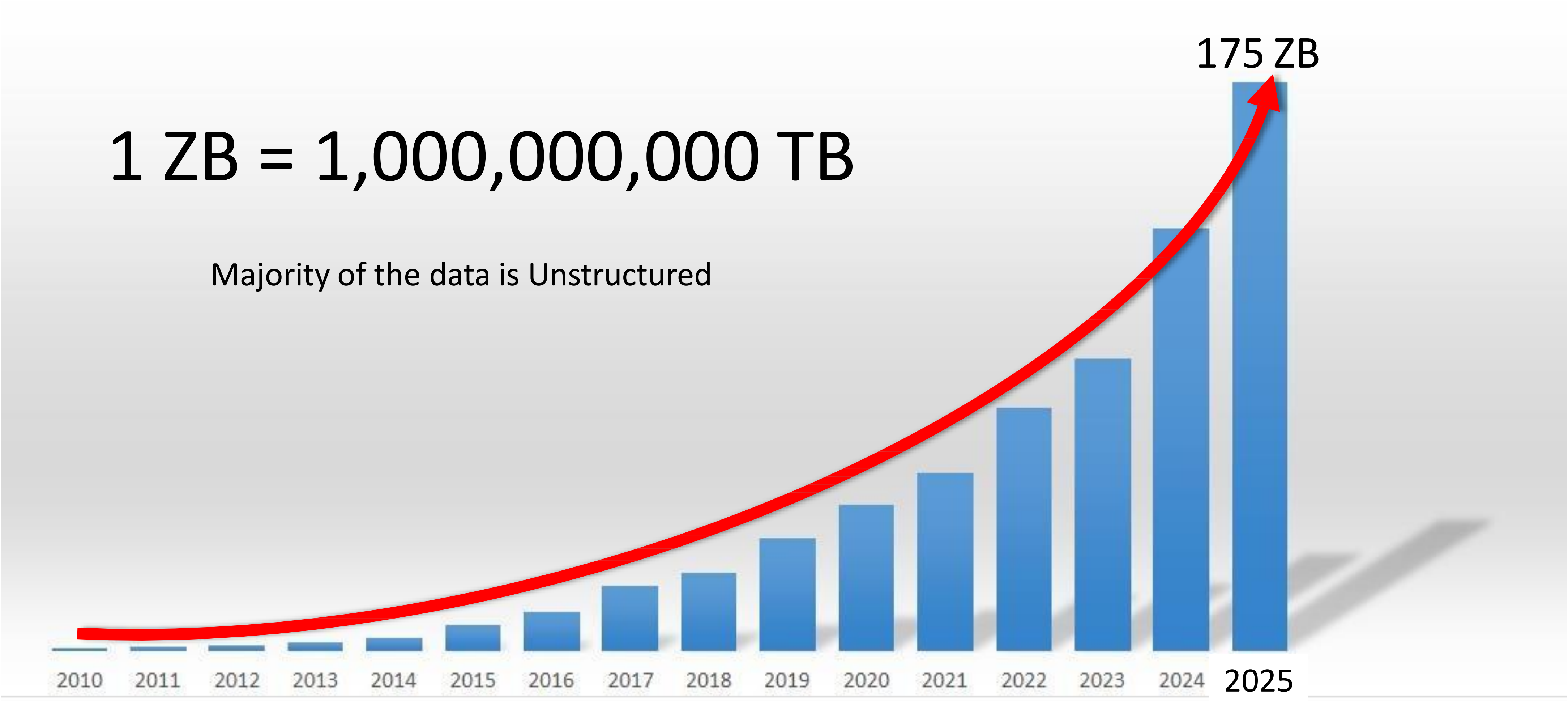
Machine Learning Introduction

Data Growth – IDC-Seagate November, 2018



1 ZB = 1,000,000,000 TB

Majority of the data is Unstructured

Application of Data Science and Machine Learning

Automobile

Preventive Maintenance

Banking

Fraud/Default Prevention

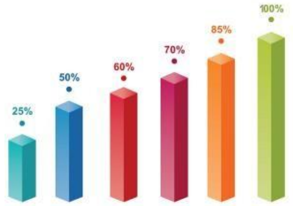
Healthcare

Predict Disease

Media

Content Personalisation

Telecom



Increase sale

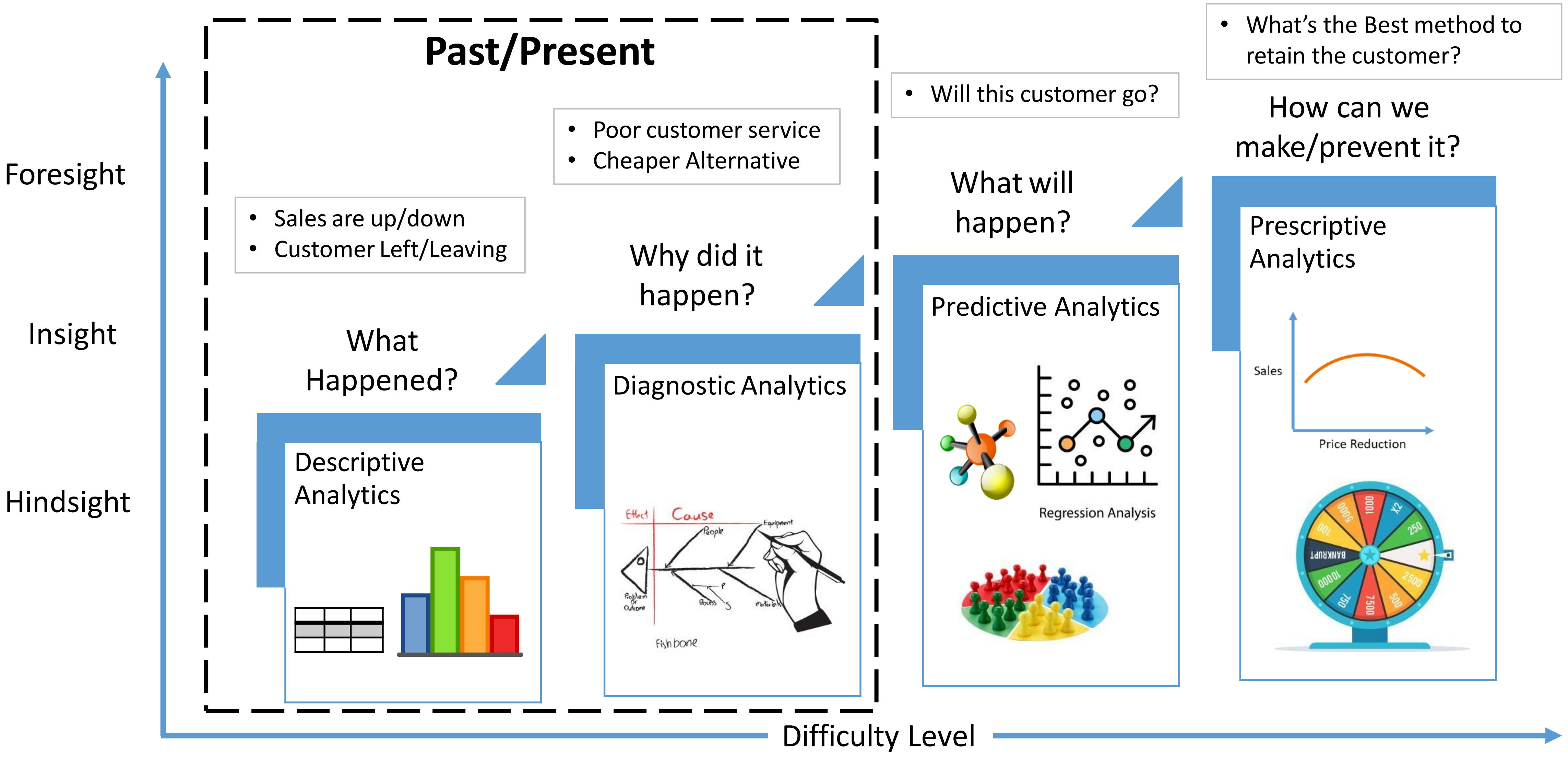
Benefits of Data Science and Machine Learning

- ✓ Faster decisions

- ✓ Develop insights that are beyond human capabilities

- ✓ Act at the right time and take advantage of opportunities, converting them into closed deals.

Types of Analytics



Past/Present

Foresight

Insight

Hindsight

- Sales are up/down
- Customer Left/Leaving

What Happened?

Descriptive Analytics

- Poor customer service
- Cheaper Alternative

Why did it happen?

Diagnostic Analytics

- Will this customer go?

What will happen?

Predictive Analytics

- What’s the Best method to retain the customer?

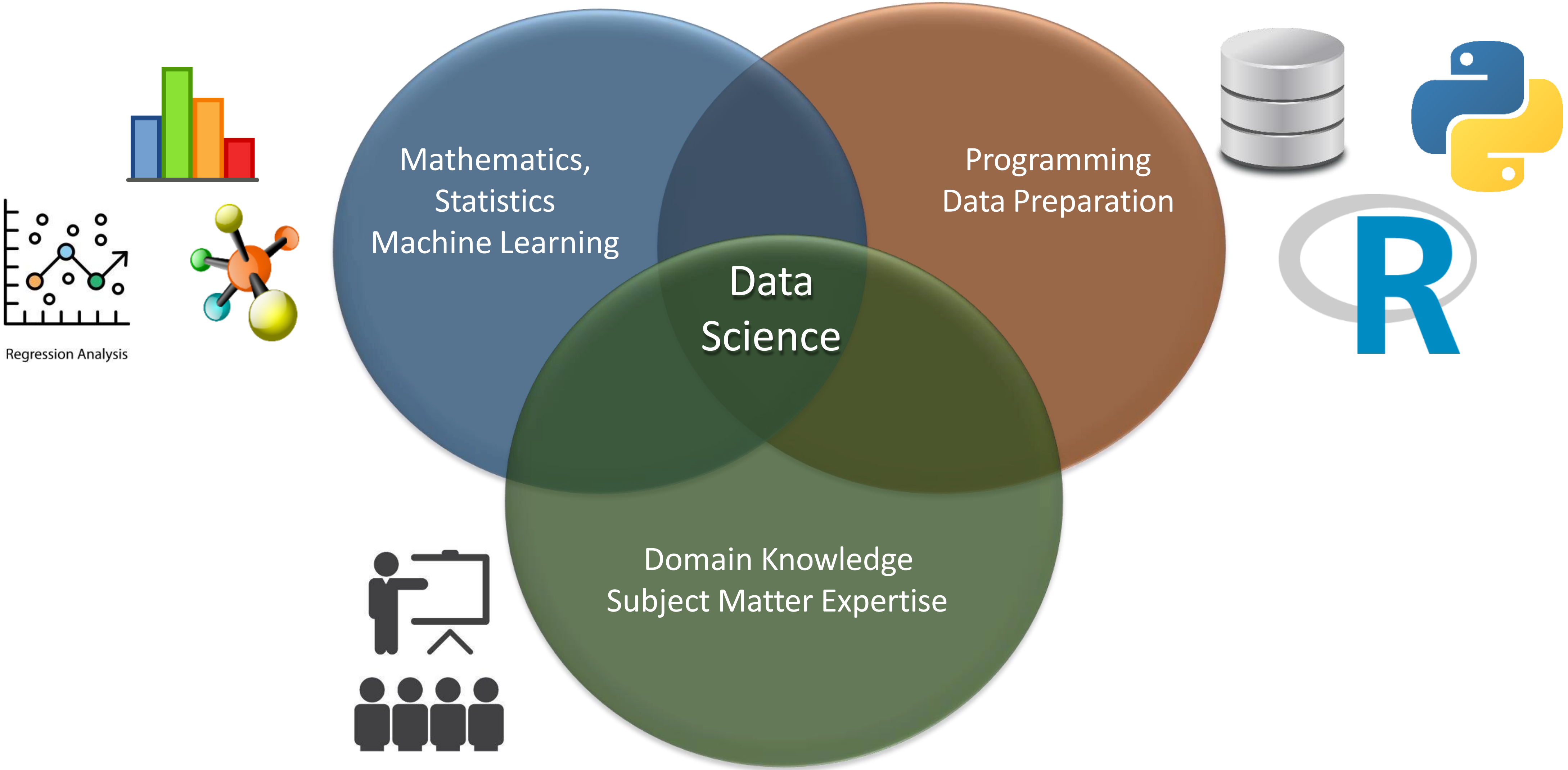
How can we make/prevent it?

Prescriptive Analytics

Difficulty Level

Courtesy – Gartner Report and analysis

What is Data Science?

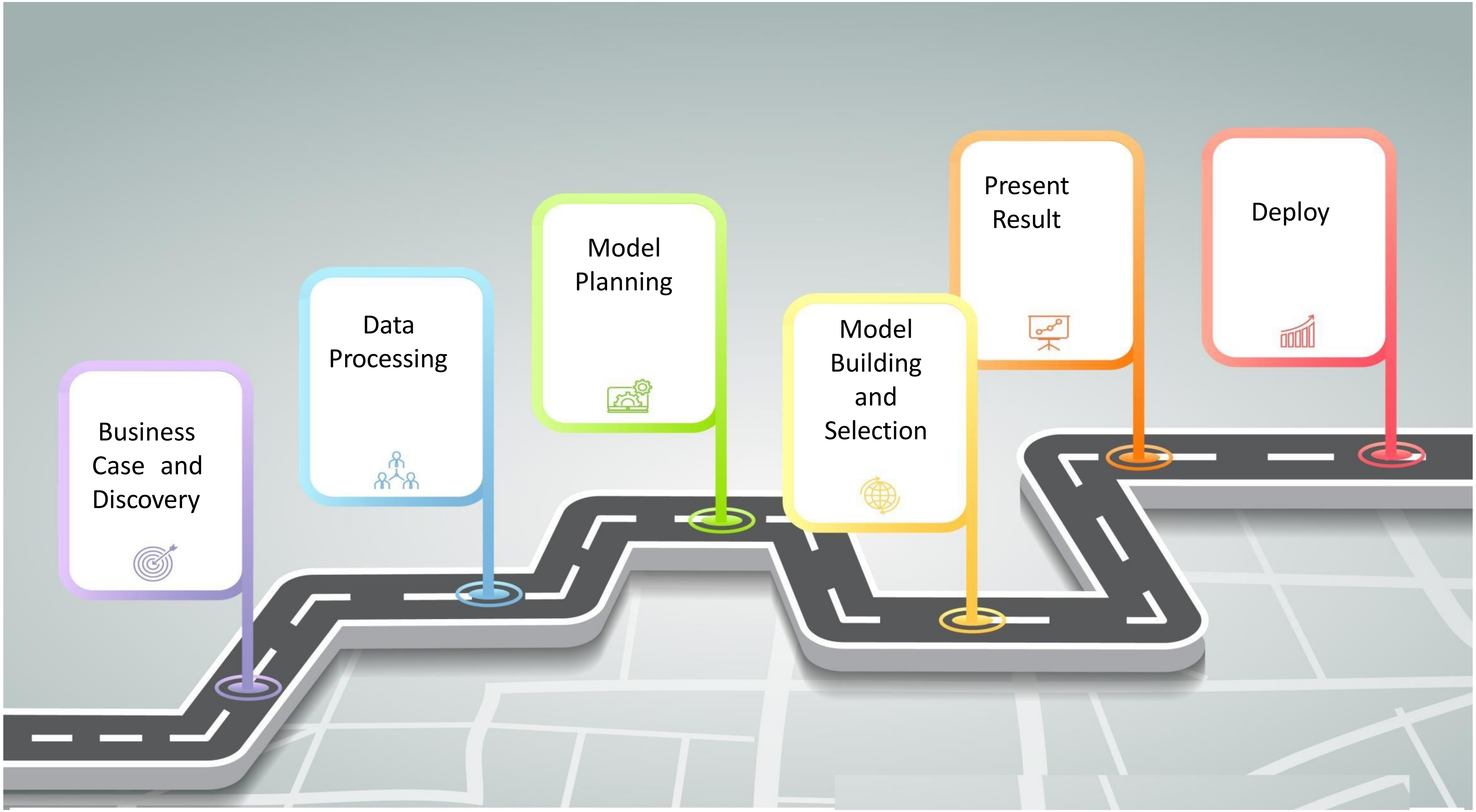


Mathematics,
Statistics
Machine Learning

Programming
Data Preparation

Domain Knowledge
Subject Matter Expertise

Data
Science


Business Case and Discovery
Data Processing
Model Planning
Model Building and Selection
Present Result
Deploy

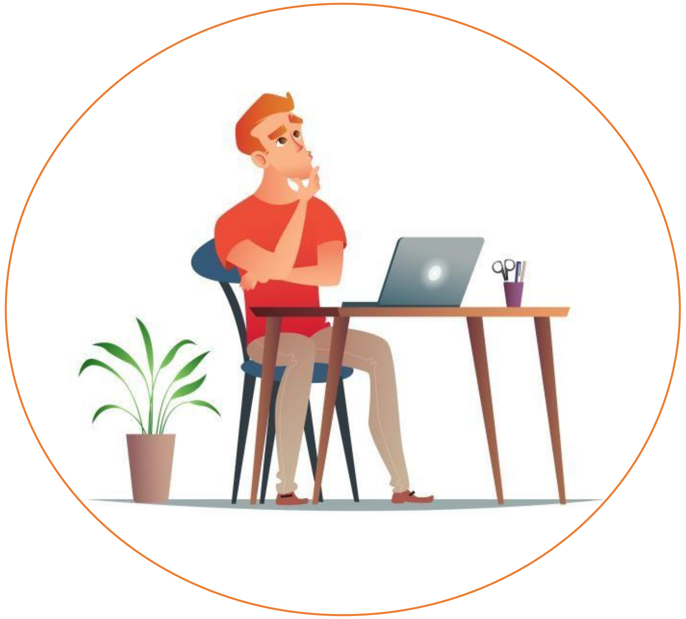
Business Case and Discovery

Stakeholders Discussions

What’s the End Goal?

How much time and budget we have

Past attempts

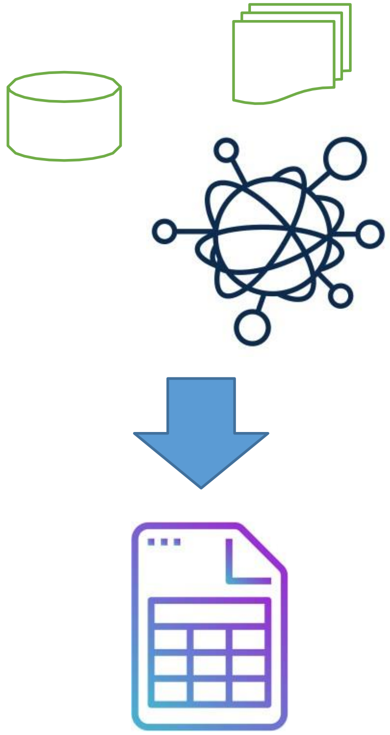


What kind of data is available

Data Processing

Data Mapping → Data Cleaning → Data Transformation → Sample the Data

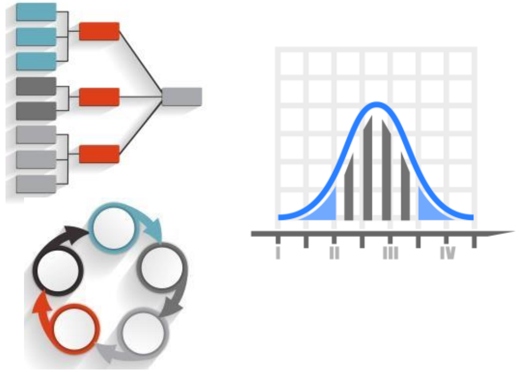
Data Mapping



Data Cleaning

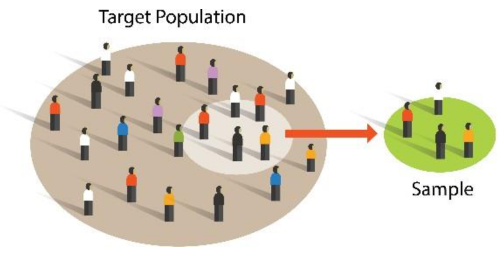
- Data Quality
- Missing Data
- Noisy Data
- Outlier Treatment

Data Transformation



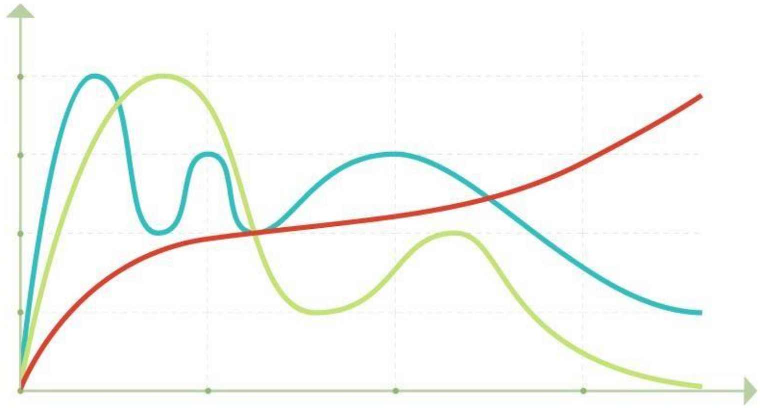
- Format conversion
- Data Normalization
- Statistical imputation
- Feature Engineering

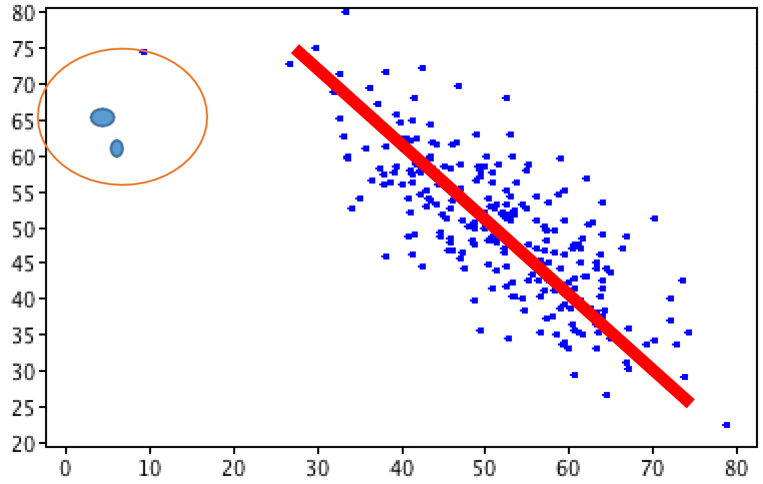
Sample the Data

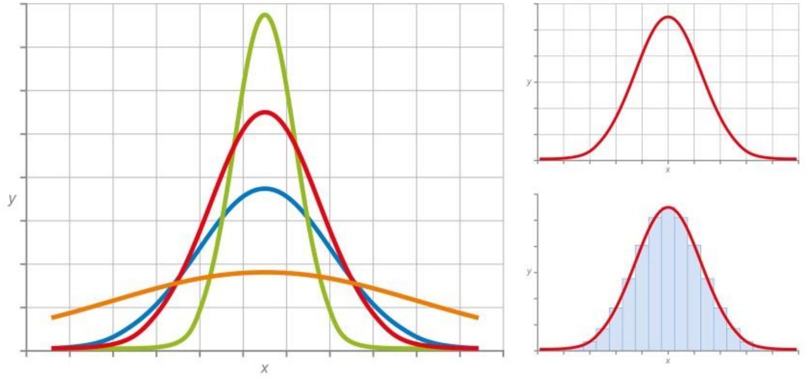


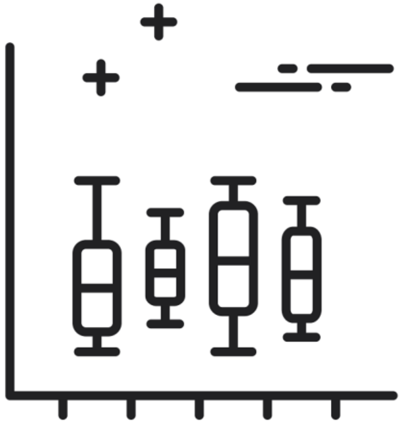
- Data Sampling
- Data Split
- Data Binning

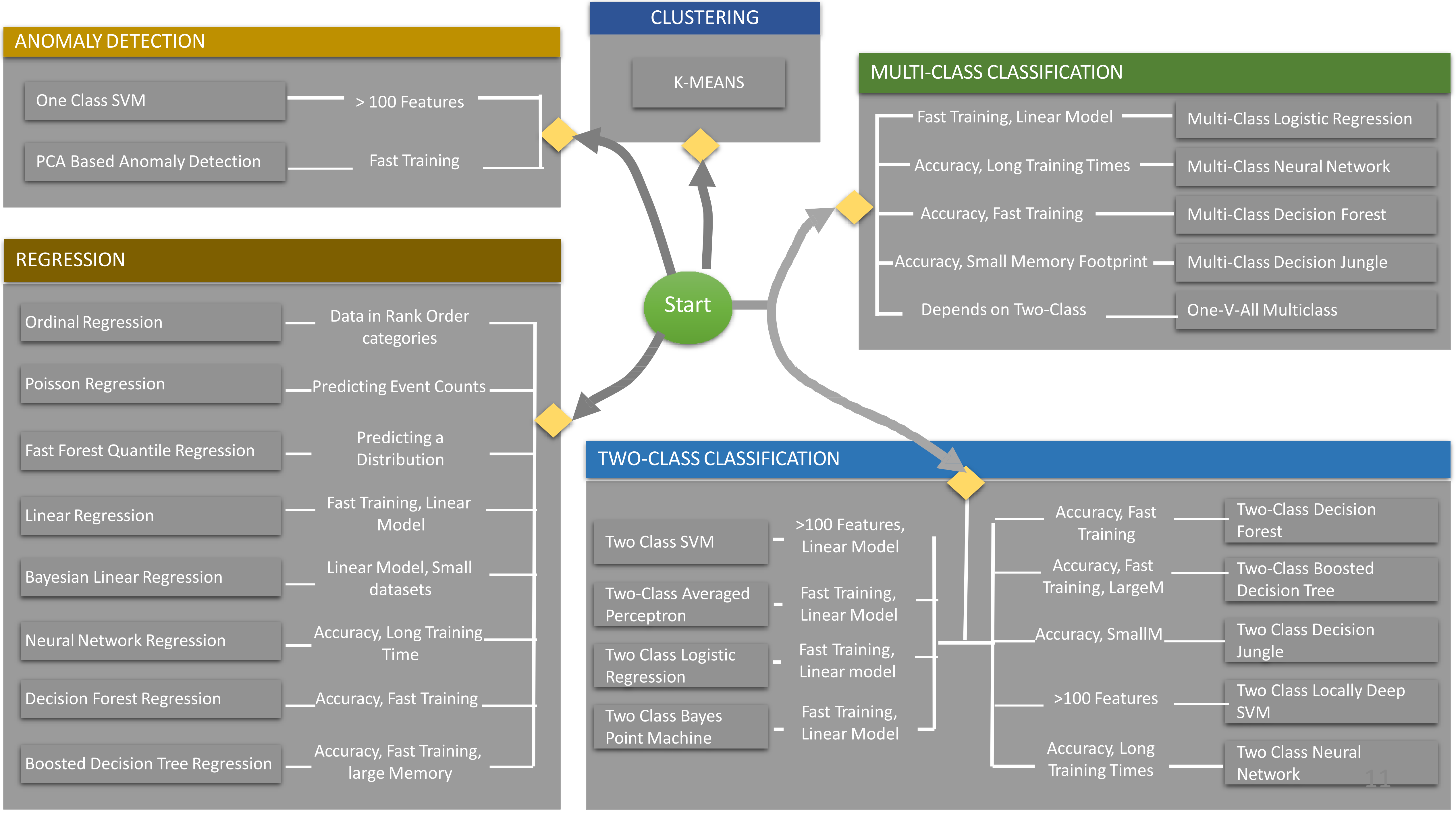
Exploratory Data Analysis











Start

ANOMALY DETECTION

One Class SVM	> 100 Features
PCA Based Anomaly Detection	Fast Training

CLUSTERING

K-MEANS

MULTI-CLASS CLASSIFICATION

Fast Training, Linear Model	Multi-Class Logistic Regression
Accuracy, Long Training Times	Multi-Class Neural Network
Accuracy, Fast Training	Multi-Class Decision Forest
Accuracy, Small Memory Footprint	Multi-Class Decision Jungle
Depends on Two-Class	One-V-All Multiclass

REGRESSION

Ordinal Regression	Data in Rank Order categories
Poisson Regression	Predicting Event Counts
Fast Forest Quantile Regression	Predicting a Distribution
Linear Regression	Fast Training, Linear Model
Bayesian Linear Regression	Linear Model, Small datasets
Neural Network Regression	Accuracy, Long Training Time
Decision Forest Regression	Accuracy, Fast Training
Boosted Decision Tree Regression	Accuracy, Fast Training, large Memory

TWO-CLASS CLASSIFICATION

Two Class SVM	>100 Features, Linear Model
Two-Class Averaged Perceptron	Fast Training, Linear Model
Two Class Logistic Regression	Fast Training, Linear model
Two Class Bayes Point Machine	Fast Training, Linear Model

Accuracy, Fast Training	Two-Class Decision Forest
Accuracy, Fast Training, LargeM	Two-Class Boosted Decision Tree
Accuracy, SmallM	Two Class Decision Jungle
>100 Features	Two Class Locally Deep SVM
Accuracy, Long Training Times	Two Class Neural Network

What to consider while choosing an algorithm?

Predicting Categories

Predicting Continuous Value

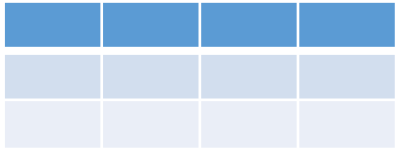
Finding Unusual Data Points

Discovering Structure

Model Building and Selection

Train Model

Cross Validation



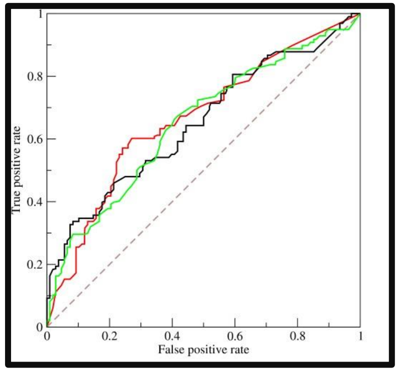
Parameter Tuning

Parameter 1 →

↓ Parameter 2

	1	2	3
A	A, 1	A, 2	A, 3
B	B, 1	B, 2	B, 3
C	C, 1	C, 2	C, 3
D	D, 1	D, 2	D, 3

Select Model



Present the results

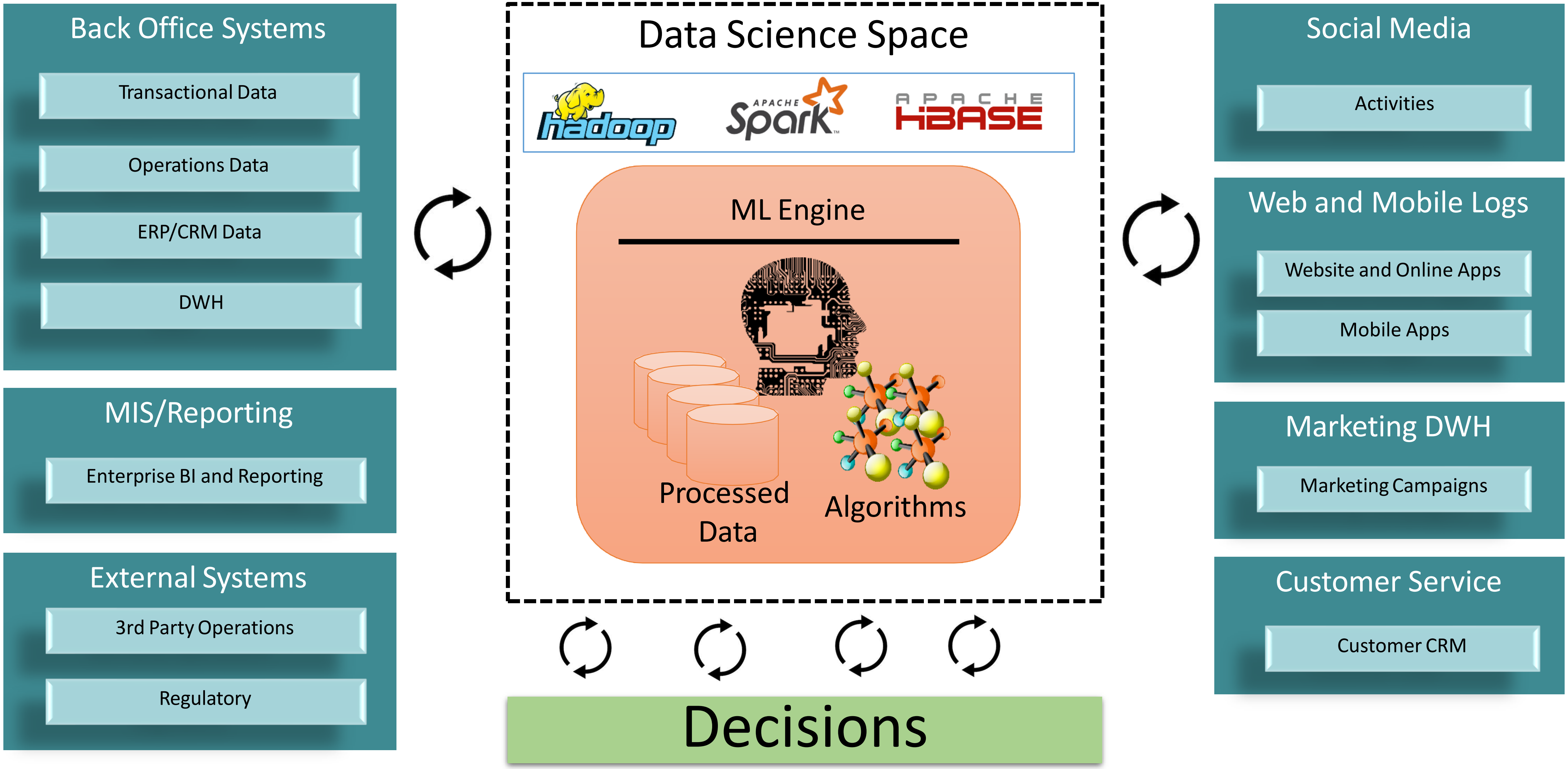
- Explain the process of model planning and selection

- Explain the findings; correlations, causes, variable selections

- Communicate the results

- Explain the process of operationalization

Deployment



Back Office Systems

Transactional Data

Operations Data

ERP/CRM Data

DWH

MIS/Reporting

Enterprise BI and Reporting

External Systems

3rd Party Operations

Regulatory

Data Science Space

ML Engine

Processed Data

Algorithms

Decisions

Social Media

Activities

Web and Mobile Logs

Website and Online Apps

Mobile Apps

Marketing DWH

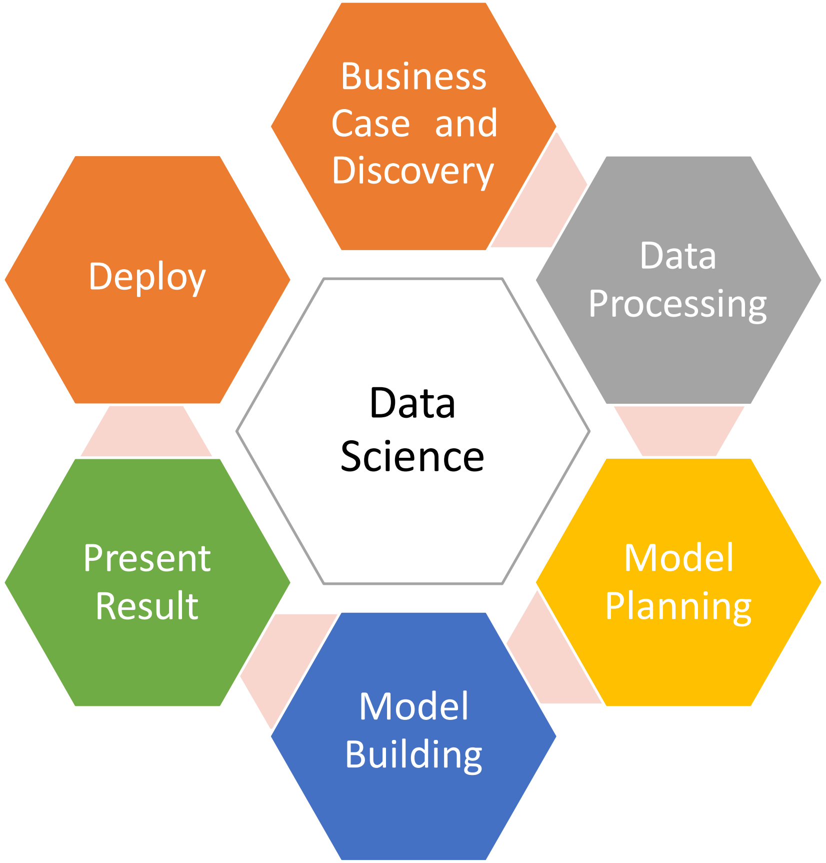
Marketing Campaigns

Customer Service

Customer CRM

Skills Required to be a Data Scientist

- Soft Skills
  - Domain knowledge
  - Communication
  - Analytical skills
  - Curiosity
  - Common Sense



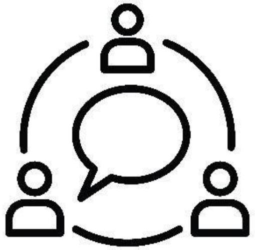
- Technical Skills
  - Mathematics
  - Statistics
  - File handling or database
  - Machine Learning
  - Python or similar
  - Tableu or similar visualization

Soft Skills

Understanding of the data elements based on domain expertise

Domain knowledge

Discovery phase as well as presenting findings to the stakeholders



Communication

Analyse various relationships among data features.

Analytical Skills

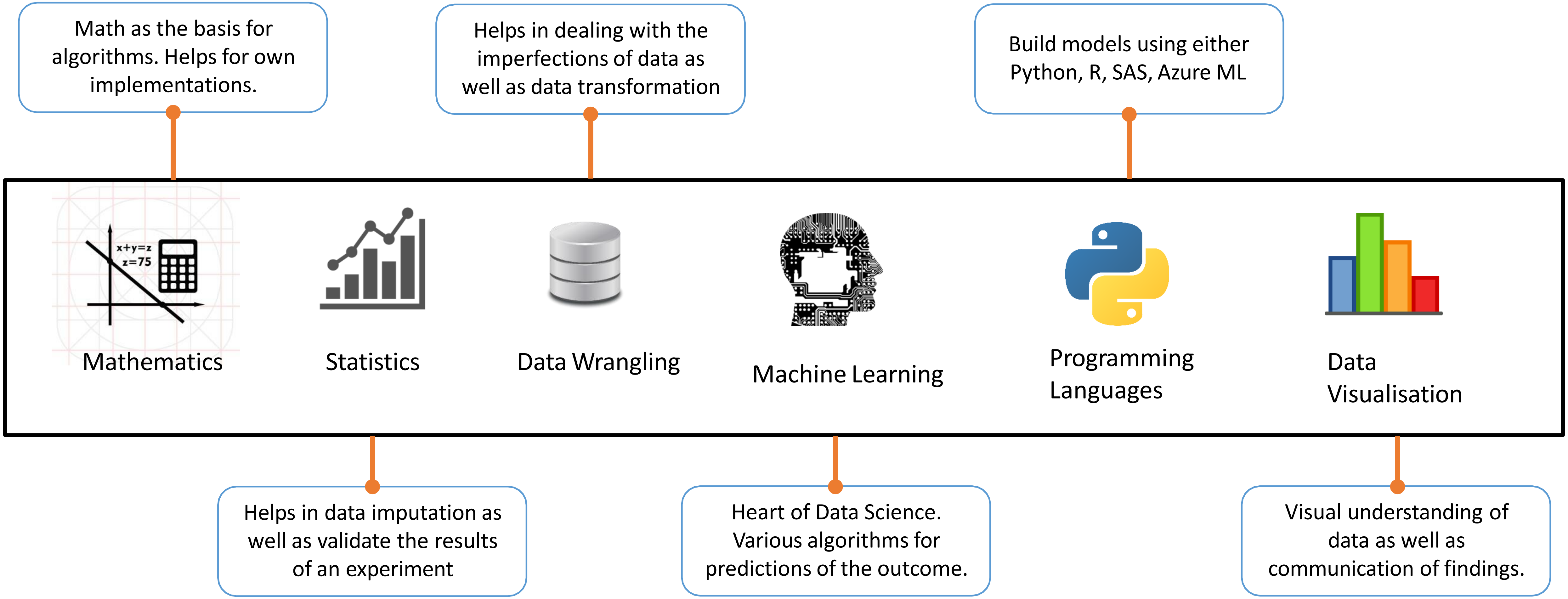
Asking the right questions to gain deeper understanding.

Curiosity

Is it making sense on normal beliefs?

Common Sense

Technical Skills



Math as the basis for algorithms. Helps for own implementations.

Helps in dealing with the imperfections of data as well as data transformation

Build models using either Python, R, SAS, Azure ML

Mathematics

Statistics

Data Wrangling

Machine Learning

Programming Languages

Data Visualisation

Helps in data imputation as well as validate the results of an experiment

Heart of Data Science. Various algorithms for predictions of the outcome.

Visual understanding of data as well as communication of findings.