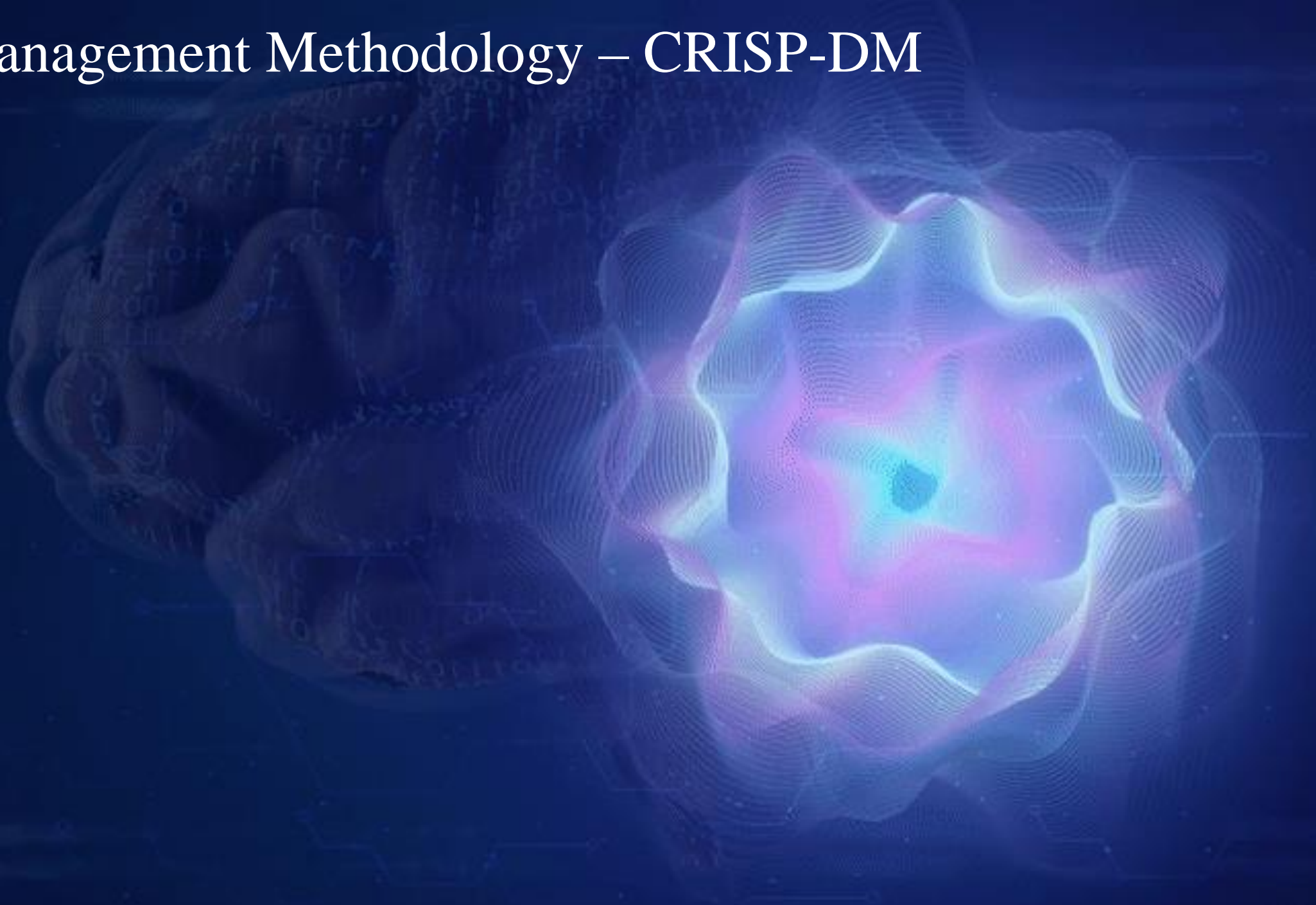


# Project Management Methodology – CRISP-DM



**1**

**Various PM methodologies**

**2**

**Frameworks for building DM solutions**

**3**

**KDD, CRISP-DM, SEMMA**

**4**

**Our Unique Methodology – Almost Always Works!**

**5**

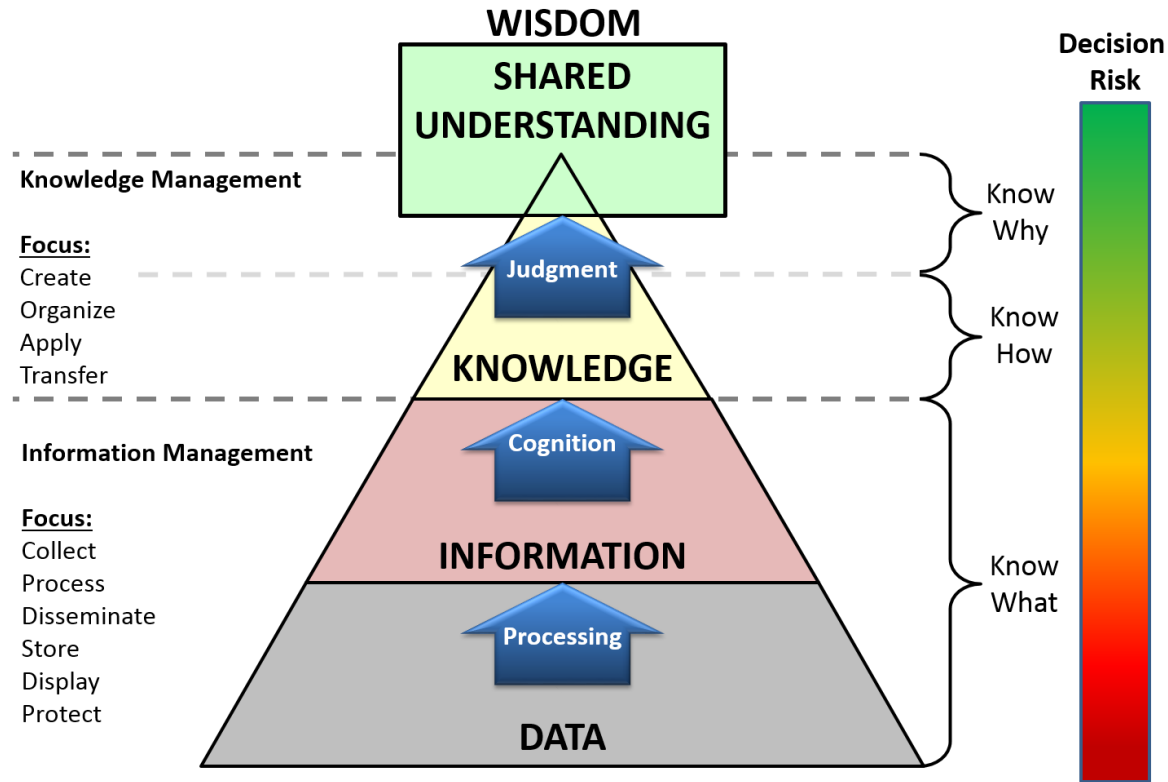
**Deep Dive into Each Stage**

*Knowledge Discovery Databases (KDD) process model*

*Cross Industrial Standard Process for Data Mining (CRISP – DM)*

*Sample, Explore, Modify, Model and Assess (SEMMA)*

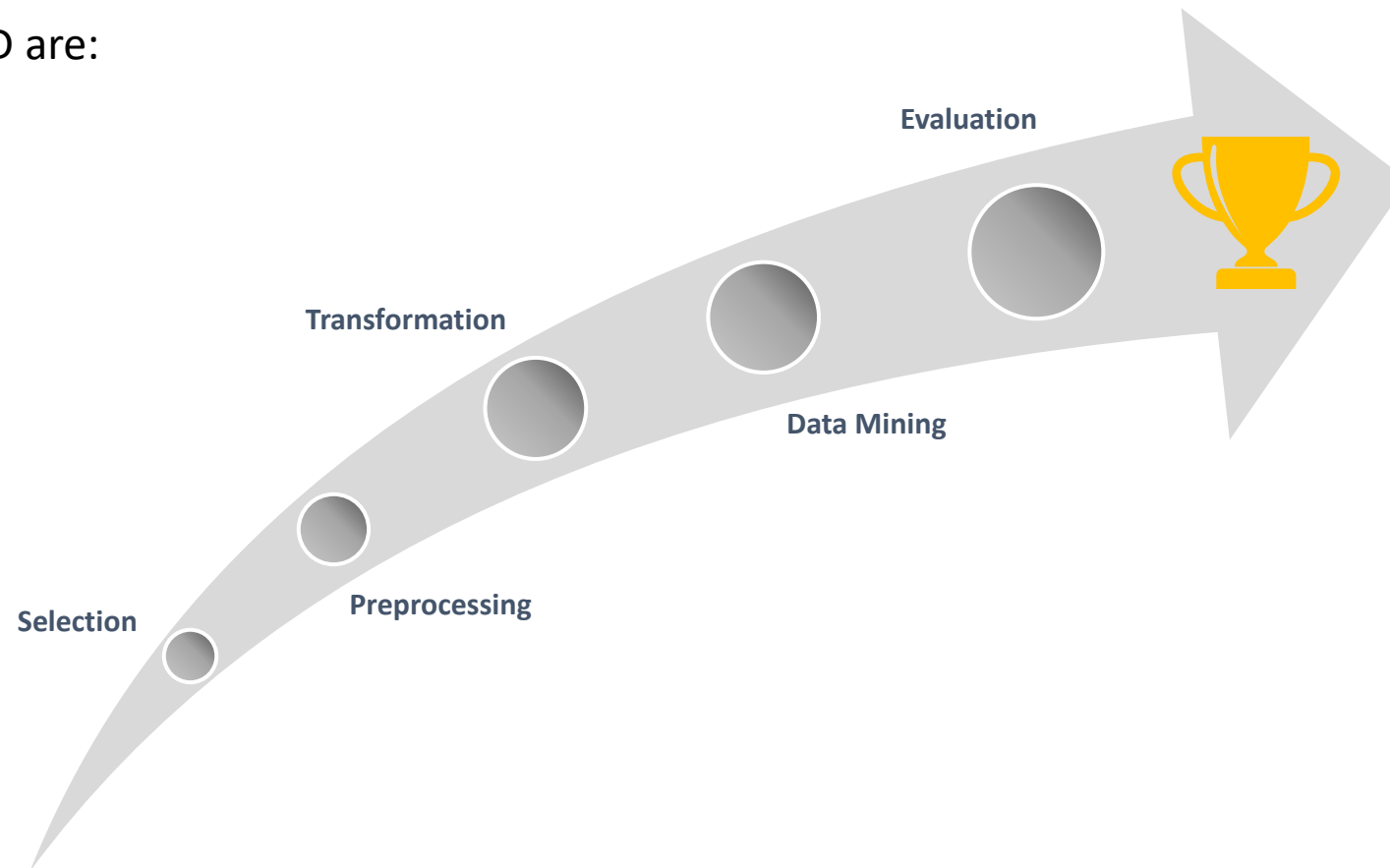
## Knowledge Management Cognitive Pyramid



Source: Wikipedia

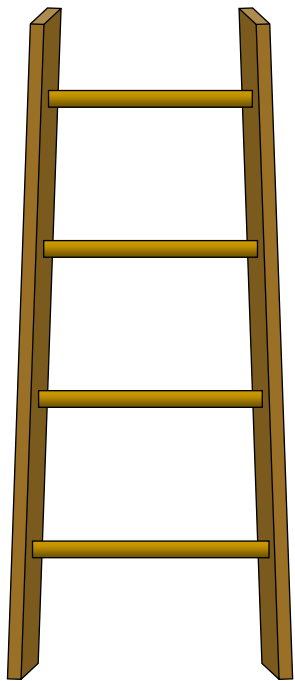
Knowledge Discovery Databases center around the overall process of knowledge discovery from data that covers the entire life cycle of data that includes how the data are stored, how it is accessed, how algorithms can be scaled to enormous datasets efficiently, how results can be interpreted and visualized

The five stages of KDD are:



SEMMA are the sequential steps to build machine learning models incorporated in 'SAS Enterprise Miner', a product by SAS Institute Inc., one of the largest producers of commercial statistical and business intelligence software

The five sequential steps of SEMMA are:



**Assess**

**Model**

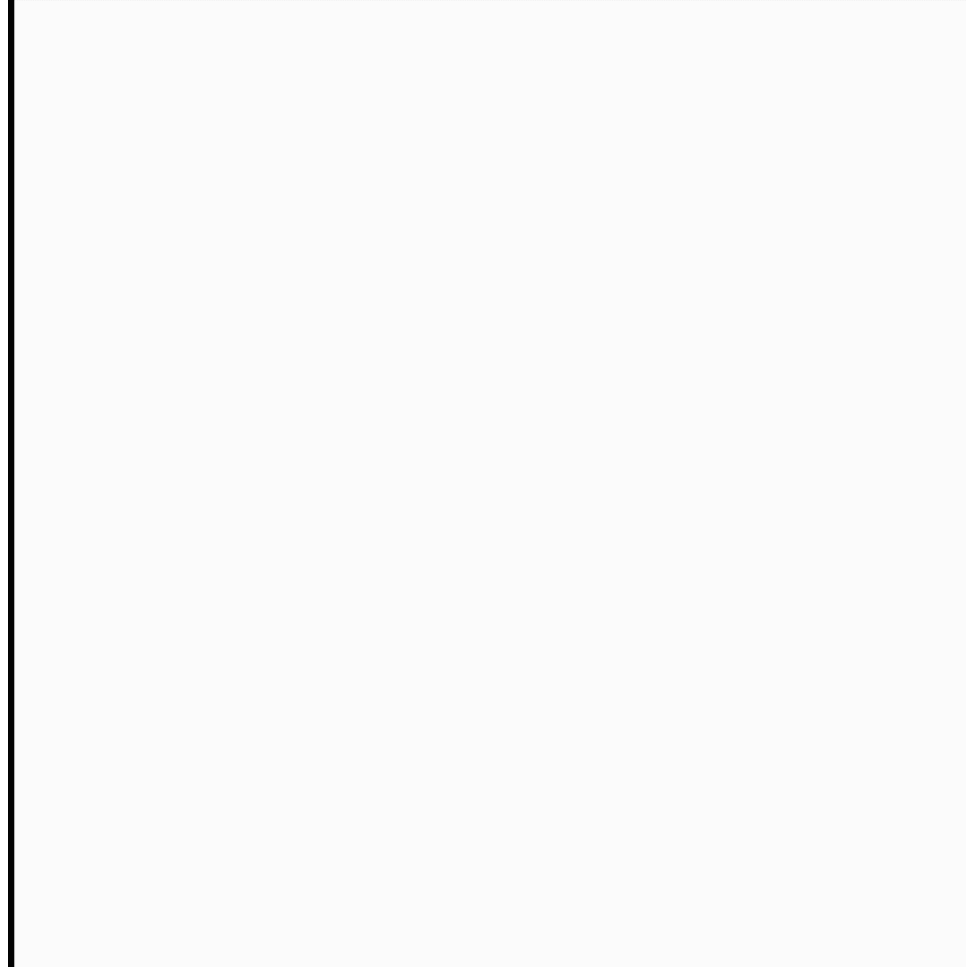
**Modify**

**Explore**

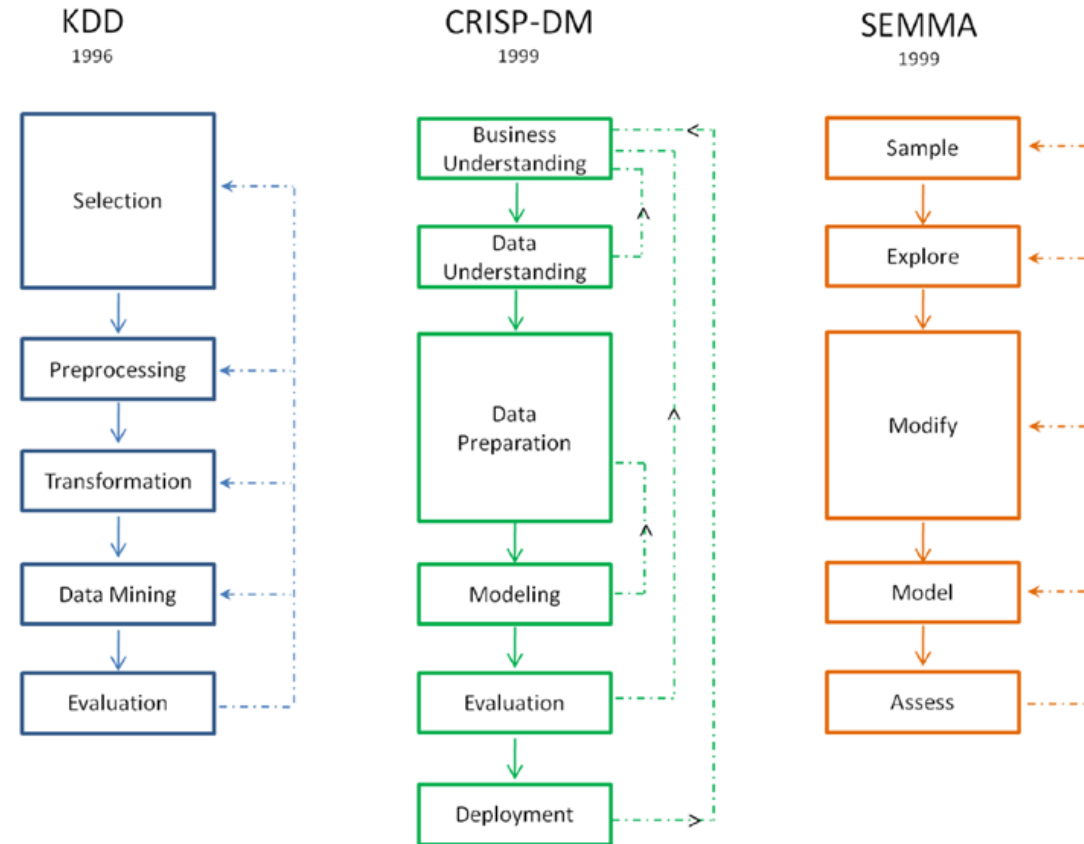
**Sample**

CRISP-DM was established by the [European Strategic Program on Research in Information Technology](#) initiative with an aim to create an unbiased methodology that is not domain dependent.

6 Step iterative process



## Summary of data mining frameworks



## The Four Key Steps of Business Understanding Phase of CRISP-DM

- . Define Business Problem
- . Assess and Analyze Scenarios
- . Define Data Mining Problem
- . Project Plan

# 1. Business Understanding – Define Business Problem



Business Problem A: Significant proportion of customers who take loan are unable to repay

Business Objective: Minimize Loan Defaulters

Business Constraints: Maximize Profits



Business Problem B: Significant proportion of customers are complaining that they did not do the credit card transaction

Business Objective: Minimize Fraud

Business Constraints: Maximize Convenience



Business Problem C: Yield of crop is not improving year on year

Business Objective: Maximize Yield

Business Constraints: Minimize Cost



Business Problem D: Present Recommendation System is not effective

Business Objective: Maximize Cross-selling & Up-selling

Business Constraints: Minimize Coupon Fatigue



Business Problem E: Google Adwords Strategy is not effective

Business Objective: Maximize Click Through Rate

Business Constraints: Minimize Cost Per Click

# 1. Business Understanding – Assess & Analyze Scenarios

As-Is state analysis from the perspective of :

- Data

- Human Resources & their available time

- Risks

What is required:

- Hardware & Software

- Human Resources

Record Assumptions & Constraints of each requirement

Verify these assumptions & constraints considering data availability

Perform Risk Management for:

- Timelines

- Human Resources

- Data

- Hardware & Software

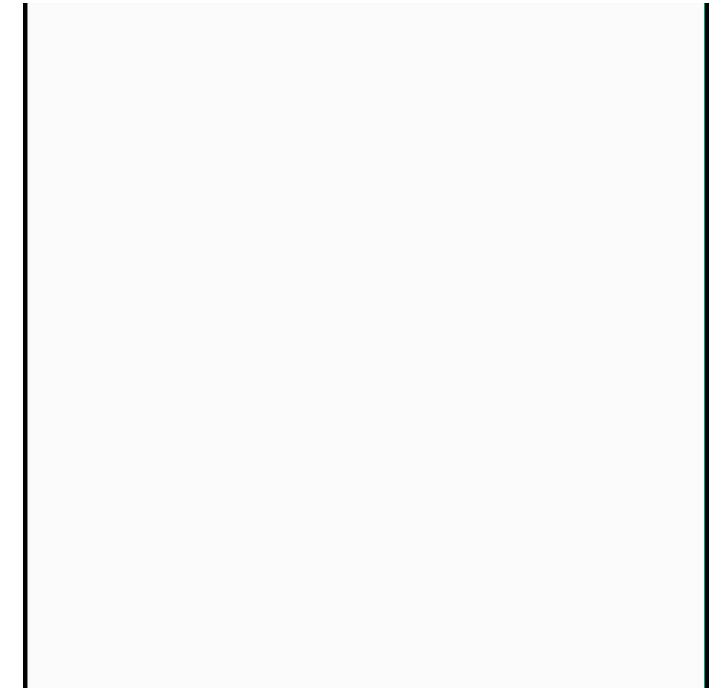
- Financial Aspects

Define success criteria

Document ROI

# 1. Business Understanding – Define Data Mining Problem

- Pre-analysis phase
- Input to this will be Success criteria and business problem along with risks, assumptions & constraints
- Technical discussions with Data Scientists, Data Analysts, Data Engineers, Architects, etc.
- Understand on what ML, Data Mining techniques and algorithms are suitable for the given business problem to be solved
- High level design for end to end solution architecture along with integration into existing customer infrastructure
- Success criteria from Data Science perspective, e.g. no overfitting with accuracy of  $> 75\%$ . Depends on industry - Social sciences or Medical sciences



## Project Plan Components:

- High Level Timelines
- Allocated Human Resources
- Allocated Hardware and Software
- Risks and Risk Response Plans
- High Level Deliverables along with Success Criteria for each of 6 phases of CRISP-DM
- Highlight One-time activities and Iterative activities pictorially

## Phase Gate Check Points:

- Project Charter
- Definition of Business Objectives
- Success Criteria for Business as well as Data Mining
- Cost Allocation and Resource Planning (Hardware as well as Software)
- ML and DM techniques and algorithms to be applied including workflow of Data from exploration to deployment
- Project plan for all 6 phases of CRISP-DM with timelines and risks identified at each phase

### The Four Key Steps of Data Understanding Phase of CRISP-DM

- Data Collection
- Data Description
- Exploratory Data Analysis
- Data Quality Analysis

## 2. Data Understanding – Data Collection

Data is something which can be measured

Data is a plural form of Datum (Latin word for “Given”. Datum is Singular)

Which in turn	Which in turn	Which in turn	Which in turn	Which in turn
Data is measure	-> Used for Analysis	-> Used for Modeling	-> Used for Prediction	-> Used for Optimization
for Management				

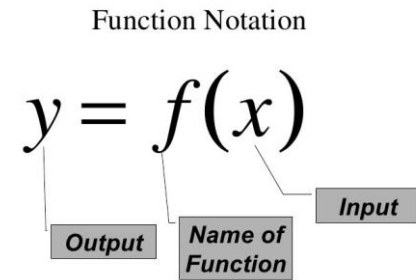
a. Primary Data Sources (Data Collected at source) - E.g., Surveys, Experiments, Interviews, Focus groups, etc.

- i. Costly
- ii. Time Consuming / Low Quality
- iii. Get Exact Data

b. Secondary Data Sources (Already available - Internal / External) - E.g., Sales Records, Industry Reports, ERP, CRM, Open Data Sources, etc.

- i. Easy / Quick Access to Data
- ii. Less in Cost
- iii. Irrelevant Data (Sometimes)

## 2. Data Understanding – Data Collection



### Data Types:

- a. Continuous
  - i. Interval
  - ii. Ratio
- b. Discrete
  - i. Categorical
    - Binary
    - Multiple
      - 1. Nominal
      - 2. Ordinal
  - ii. Count

#### *Different Names of Y:*

Response, Dependent, Regressand, Explained variable, Criterion, Measures variable, Experimental variable, Label, Outcome...

#### *Different Names of X:*

Explanatory, Predictor, Independent, Covariates, Regressors, Factors, Carriers, Controlled variable, Manipulated variable, Exposure variable, Input....

#### *Different Names for Rows:*

Records, Observations, Cases, Entries, Entities...

#### *Different Names for Columns:*

Features, Columns, Variables.....

## 2. Data Understanding – Data Collection

Quantitative vs Qualitative

Balanced vs Imbalanced

Structured vs Unstructured as well as semi-structured (Raw format)

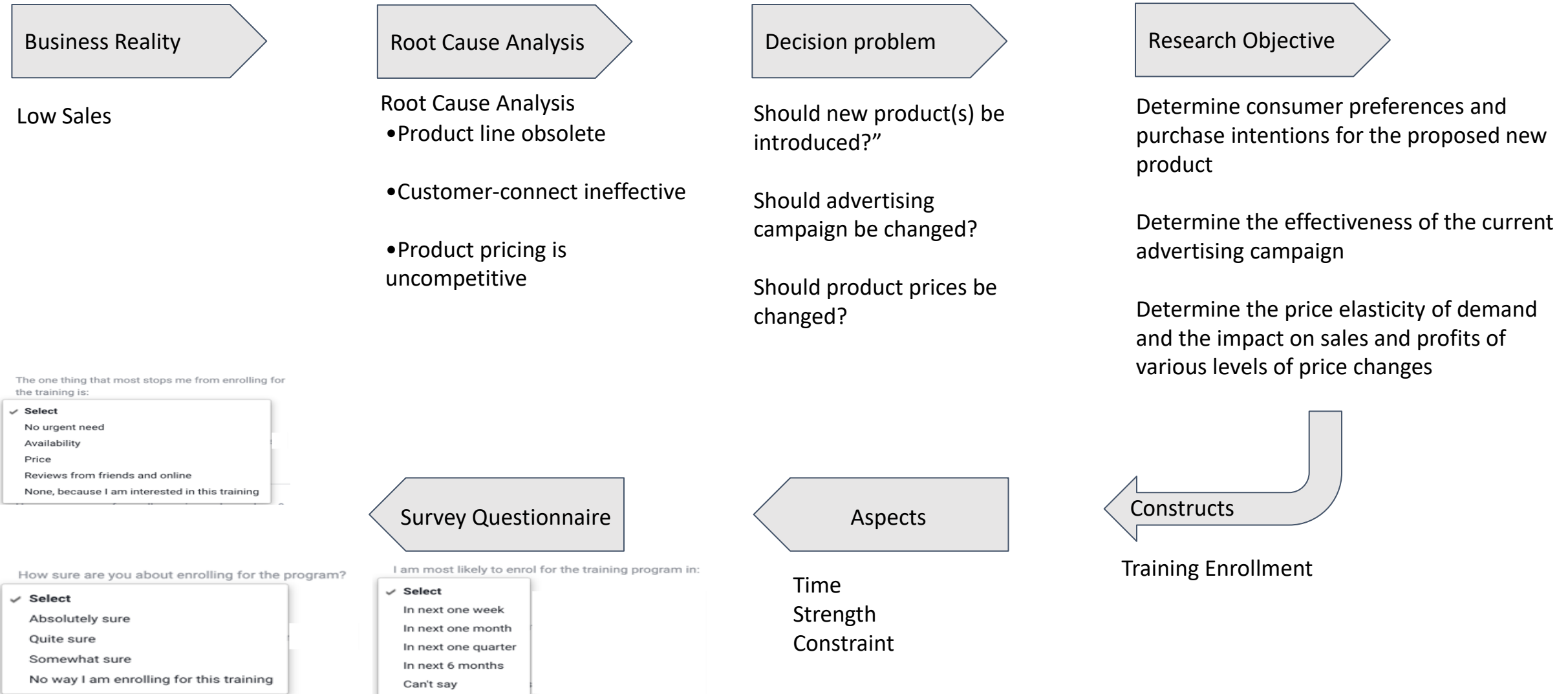
Batch Data vs Streaming Data

Big Data vs Not Big Data

Time Series vs Cross sectional vs Panel / Longitudinal Data

# 2. Data Understanding – Data Collection

## Data Collection using Survey



## 2. Data Understanding – Data Collection

### Data Collection using Design of Experiments (DoE)

Coupon marketing

10% discount vs 20% discount

Expiry Date

2 days expiry vs 10 days expiry

Distance

5 kms vs 10 kms

Combination

10% discount & 2 kms radius vs 12% discount & 3 kms radius

DATA, DATA, DATA!!!!

- Effects of Fertilizers A, B, C on the Yield of the crop should be compared
- Scientist experimenting this also considers soil types - Clay, Silt, Sandy Soil
- Compare the fertilizer effects based on soil type
- For each Soil Type, the scientist chooses 5 representative equal sized plots
- Scientist assigns each fertilizer to each of the equal sized plots RANDOMLY
- Based on this DoE, yield is observed, and data collected

# 2. Data Understanding – Data Collection

## Data Collection using Design of Experiments (DoE)

What factors impact?

If food, then veg vs non-veg, cuisine variant, .....day of the week, time of the day.....

Do a trial / experiment

Figure 1: Effect of Mobile Promotions via Temporal Targeting

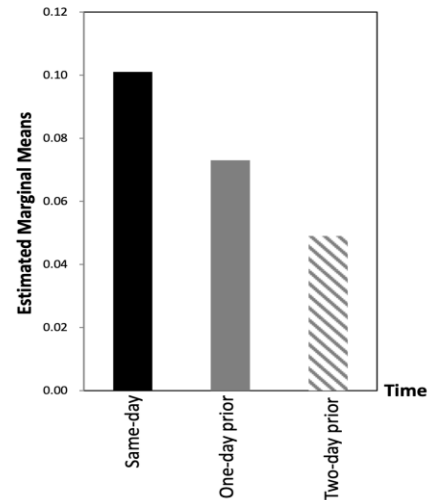


Figure 2: Effect of Mobile Promotions via Geographical Targeting

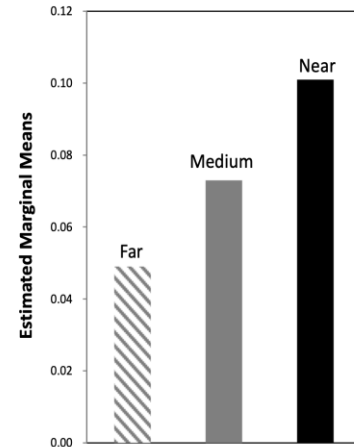
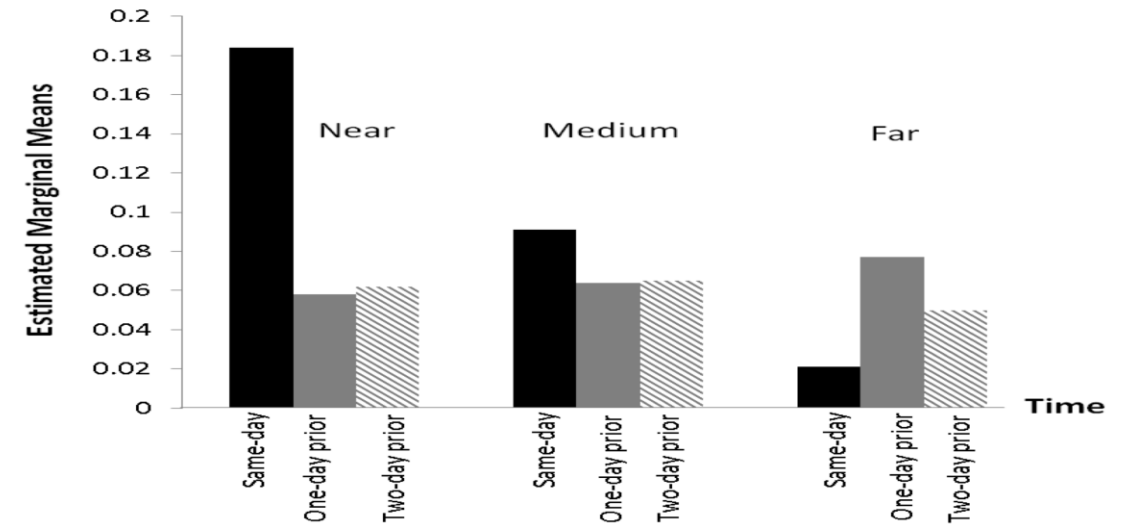


Figure 3: Effect of Mobile Promotions via Combining Temporal and Geographical Targeting



## 2. Data Understanding – Data Description

Data Description is all about performing Initial Analysis

### Data Sources

RDBMS

SQL

NoSQL

Big Data

Record of Origin (ROO)

Record of Reference (ROR)

### Data Volume

Size

Number of records

Total databases

Tables

### Data Attributes & their description

Variables

Data Types

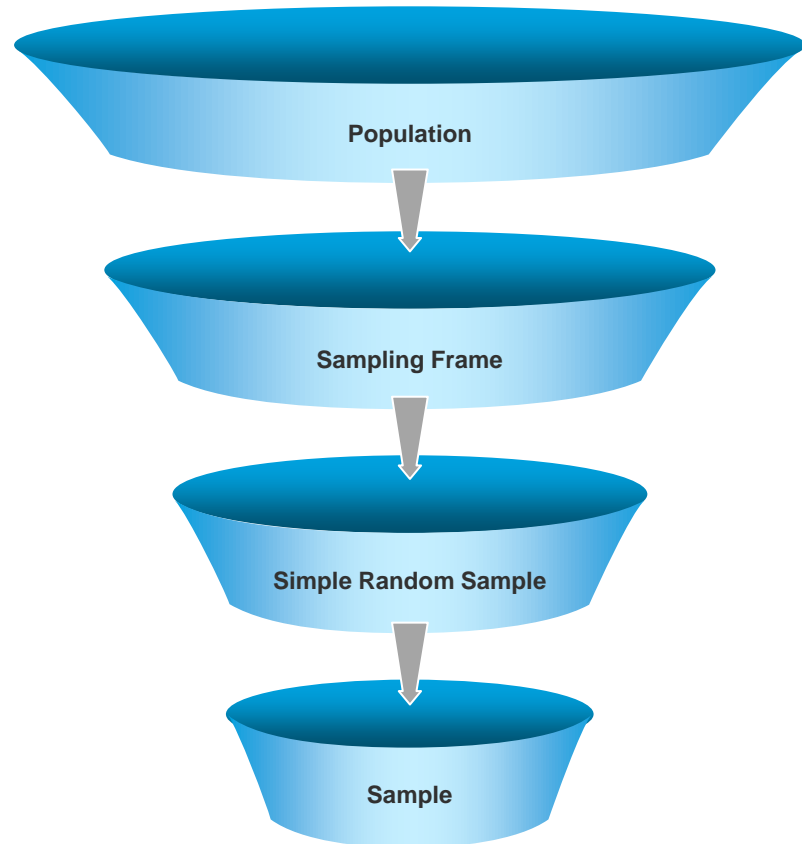
Internal Data Sources	External Data Sources
Transaction Data	Survey Data e.g <a href="http://www.mospi.gov.in/">http://www.mospi.gov.in/</a>
Customer Preference Data	Biometric Data e.g. <a href="https://www.dhs.gov/immigration-statistics">https://www.dhs.gov/immigration-statistics</a>
Experimental Data	Third Party Data e.g. <a href="http://aws.amazon.com/datasets">http://aws.amazon.com/datasets</a>
Customer Relationship Data	Government & Quasi Government Agencies e.g. <a href="http://data.gov/">http://data.gov/</a>
	Social Sites Data e.g. <a href="https://en.wikipedia.org/">https://en.wikipedia.org/</a>

# 2. Data Understanding – Exploratory Data Analysis

## Inferential Statistics / Analysis

Sampling - Balanced vs Imbalanced Data sets

Hypothesis Testing - Parametric vs Non-parametric



## For Imbalanced Datasets

- Random Resampling - Under & Over Sampling
- Stratified Sampling
- Re-substitution
- K fold Cross Validation
- Leave-one-out (N-fold cross-validation)
- SMOTE - Synthetic Minority Oversampling Technique
- MSMOTE - Modified SMOTE
- Cluster based sampling
- Ensemble Techniques - Bagging & Boosting
  - Ada Boost
  - Gradient Tree Boosting
  - XG Boost

# 2. Data Understanding – Exploratory Data Analysis

## Descriptive Statistics / Analysis

First Moment Business Decision / Measures of Central Tendency

Mean, Median, Mode

Second Moment Business Decision / Measures of Dispersion

Variance, Standard Deviation, Range

Third Moment Business Decision

Skewness

Fourth Moment Business Decision

Kurtosis

## Graphical Representation

### Univariate

Box Plot / Box & Whisker Plot

- i. Primary purpose – Identify outliers
- ii. Secondary purpose – Identify shape of distribution

Histogram

- i. Primary purpose – Identify shape of distribution
- ii. Secondary purpose – Identify outliers

Q-Q Plot (Quantile - Quantile) - Data are normal or not

## 2. Data Understanding – Exploratory Data Analysis

### Bivariate

#### Scatter Plot

##### i. Primary purposes

1. Direction – Positive, Negative, no correlation
2. Strength – Strong, moderate, weak – Subjective;  
Objective – Correlation Coefficient;  $r$ : -1 to +1;  
 $|r| > 0.85 \Rightarrow$  Strong;  $|r| < 0.4 \Rightarrow$  Weak
3. Linear or Non-linear / Curvilinear

##### ii. Secondary purposes

1. Clusters
2. Outliers

## 2. Data Understanding – Data Quality Analysis

Focus of this step is only in identifying the potential errors, shortcomings and issues with data

- Identify Outliers
- Identify Missing Data
- Identify Different levels of granularity
- Validation and Reliability
- Inconsistent Data
- Wrong information due to data errors (manual / automated) - AAA or Gage R & R
- Wrong metadata information

## 2. Data Understanding – Data Quality Analysis

Four errors to be avoided during Data Collection

- Random Errors - Measurement device (thermometer) faulty or Person measuring does mistakes. Leads to False Positives e.g. Cancer
- Systematic Errors - Social desirability bias of Trump on Twitter. Wearable devices data is of wealthy customers
- Errors of choosing what to measure - Rather than choosing a person from top university for a job, may be we need to look at their social network which guided them through series of events, which resulted in them joining the top school. High SAT score is not just based on high IQ, it depends on access to good tutors and purchasing good study material. Someone might like a subject and hence got a high GPA, but can we guarantee such a success in other fields
- Errors of exclusion - Not capturing women data pertaining to cardiovascular diseases. Election in US, not having data of colored women candidates. Chief Diversity Officer in big firms is a solution!

### The Three Key Steps of Data Preparation Phase of CRISP-DM

- Data Integration
- Data Wrangling
- Attribute Generation & Selection

# 3. Data Preparation – Data Integration & Wrangling

Data Integration is invoked when there are multiple datasets to be integrated or merged

Appending - Multiple datasets with same attributes / columns

Merging - Multiple datasets having different attributes using a common attribute

Data Wrangling or Data Munging

Clean, Wrangle, Curate, and Prepare the data

- Outlier Analysis / Treatment - 3 R technique - Rectify, Remove, Retain
- Special Case of Outlier Analysis: What if there are 100000 outliers out of 100 Million records?
- Handling Missing Data - Imputation - Mean, Median, Mode, Regression, Hot Deck, KNN, etc.
- Data Transformation - Log, Exp, Sqrt, Reciprocal, Box-Cox, Johnson - Done when data are non-normal, Data suffers from heteroscedasticity or Collinearity problem, etc.
- Data Normalization / Standardization - Used to make data Unitless and Scale Free
- Discretization / Binning / Grouping
- Dummy Variable Creation - One Hot Encoding
- Heterogeneous Data
- Handling Data Inconsistencies
- Fixing incorrect metadata and annotations
- Handling Ambiguous Attribute Values
- Curating and Formatting data into required formats - CSV, JSON, relational

# 3. Data Preparation – Attribute Generation & Selection

Attribute Generation is also called as Feature Extraction or Feature Engineering. Using your given variables, try to apply domain knowledge to come up with more meaningful derived variables

Attribute Selection is shortlisting a subset of features or attributes based on

- Attribute importance
- Quality
- Relevancy
- Assumptions
- Constraints

Feature Extraction

Derived Features

Normalized Features

Feature Selection

Hypothesis Testing

Information Gain - Decision Tree

Variable Importance Plot -

Random Forest

Lasso / Ridge Regression

## Derived Features

### Raw Input

- **Time** of **current** transactions
- **Place** of **current** transactions
- **Time** of **previous** transactions
- **Place** of **previous** transactions

### Derived Feature - 1

**Distance** (Prev - Current)

**TimeLag** (Prev - Current)

### Derived Feature - 2

**Velocity** (Prev - Current)

$$\text{Velocity (Prev} \rightarrow \text{Current)} = \frac{\text{Distance (Prev} \rightarrow \text{Current)}}{\text{TimeLag (Prev} \rightarrow \text{Current)}}$$

# 3. Data Preparation – Attribute Generation & Selection

## Normalized Features

### Raw Feature

Total Card Balance

Total Card Payment

Total Debt

### Normalized Feature

Total Card Balance / Total Credit Limit

Total Card Payment / Total Card Balance

Total Debt / Annual Income

$\log(\text{Total Debt}) / \log(\text{Annual Income})$

### The Four Key Steps of Modeling Phase of CRISP-DM

- Selecting Model Techniques
- Model Building
- Model Evaluation and Tuning
- Model Assessment

# 4. Modelling: Selecting Model Techniques & Model Building

- Supervised learning

- Predict an output  $y$  when given an input  $x$

- Predict a categorical class: classification

- Predict a numerical value: prediction

- Predict user PREFERENCE from a large pool of options: Recommendation

- Predict RELEVANCE of an entity to a “query”: Retrieval

- Unsupervised learning

- Reinforcement learning (learning from “rewards”)

- Semi-supervised learning (combines supervised + unsupervised)

- Active learning, Transfer learning, Structured prediction

# 4. Modelling: Selecting Model Techniques & Model Building

Data Mining (Cross sectional / Panel)

a. Supervised Learning / Machine Learning / Predictive Modelling (Y known)

i. Regression Analysis (Interpret the parameters)

1. Y= Continuous -> Linear Regression
2. Y = Discrete (2 categories) -> Logistic Regression
3. Y = Discrete (> 2 categories) -> Multinomial / Ordinal Regression
4. Y = Count -> Poisson / Negative Binomial Regression
5. Excessive Zero – ZIP (Zero Inflated Poisson) / ZINB (Zero Inflated Negative Binomial) / Hurdle

ii. KNN - K Nearest Neighbor

iii. Naïve Bayes

iv. Black Box Techniques (No interpretation exists)

1. Neural Network
2. Support Vector Machine

v. Ensemble Techniques

1. Stacking
2. Bagging (Random Forest)
3. Boosting (Decision Tree, Gradient Boosting, XGB, Adaboost)

# 4. Modelling: Selecting Model Techniques & Model Building

## a. Data Mining Unsupervised Learning (Y unknown)

### i. Clustering / Segmentation – Reduce the row

1. K-Means / non-hierarchical – Upfront determine the # of clusters – Scree plot / Elbow curve
2. Hierarchical / Agglomerative – Dendrogram
3. DBSCAN - Density-Based Spatial Clustering of Applications with Noise
4. OPTICS - Ordering Points to Identify Cluster Structure
5. CLARA - Clustering Large Applications
6. K-medians / K-Medoids / K-modes

### ii. Dimension Reduction - Reduce the columns

1. PCA (Principal Component Analysis), Factor Analysis
2. SVD (Singular Value Decomposition)

### iii. Association Rules / Market Basket Analysis / Affinity Analysis

1. Support
2. Confidence
3. Lift Ratio  $> 1 \Rightarrow$  Antecedent & Consequent have strong association

## iv. Recommender Systems

### v. Network Analytics

1. Degree
2. Closeness
3. Betweenness
4. Eigenvector
5. Page Rank

### vi. Text Mining & NLP (Natural Language Processing)

1. BoW - Bag of Words
2. TDM / DTM
3. TF / TFIDF

# 4. Modelling: Selecting Model Techniques & Model Building

## a. Forecasting / Time Series

### i. Model Based Approaches

#### 1. Trend

- a. Linear
- b. Exponential
- c. Quadratic

#### 2. Seasonality

- a. Additive
- b. Multiplicative

### ii. Data Based Approaches

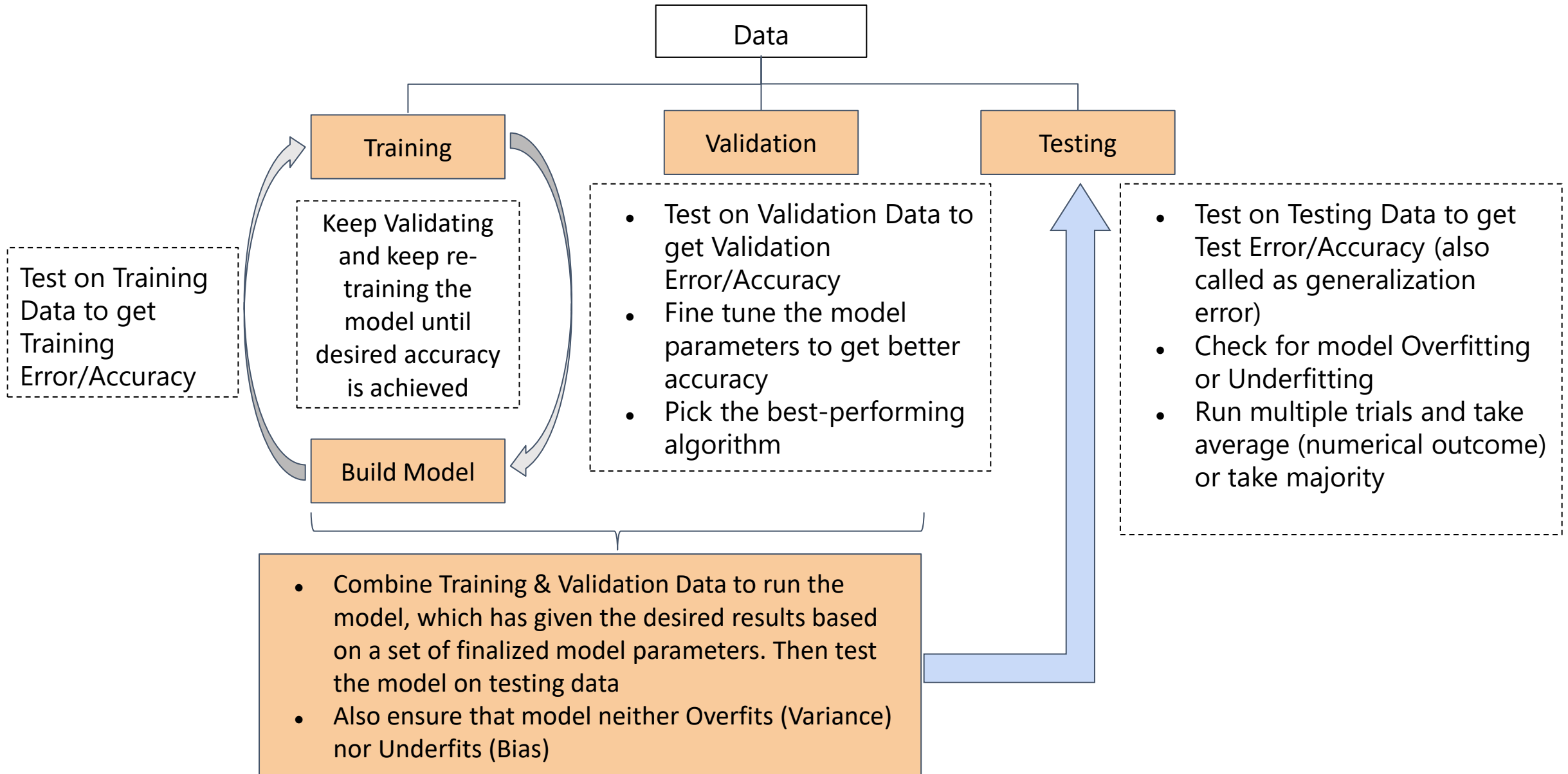
#### 1. AR

#### 2. MA

#### 3. ES

- a. SES
- b. Holts / Double Exponential Smoothing
- c. HoltWinters / Winters

# 4. Modelling: Selecting Model Techniques & Model Building

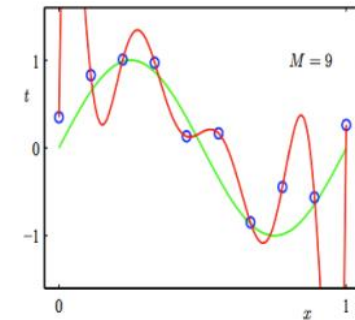
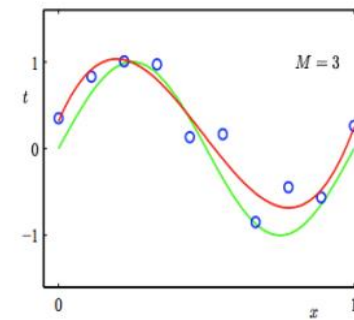
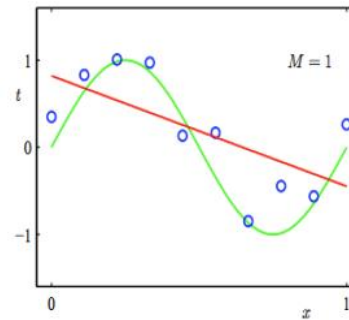


# 4. Modelling: Model Evaluation & Tuning

Overfitting (Variance) vs Underfitting (Bias)

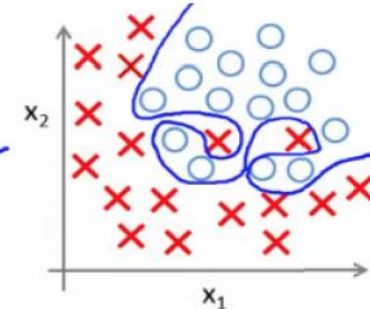
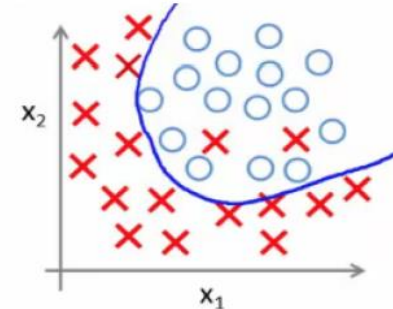
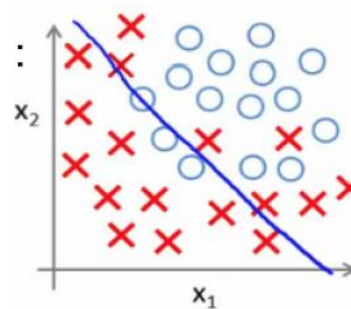
Regression

Classification



predictor too inflexible:  
cannot capture pattern

predictor too flexible:  
fits noise in the data

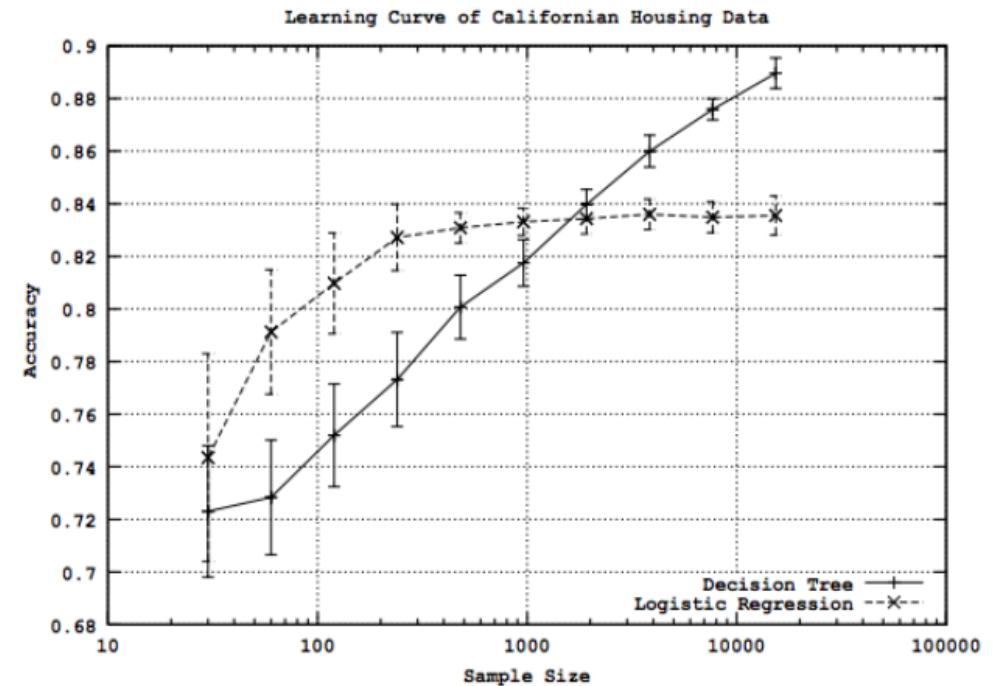


# 4. Modelling: Model Evaluation & Tuning

## Choosing Training, Validation & Testing Datasets

For Balanced Datasets:

- Split randomly to avoid bias
- Take 60% into Training, 20% into Validation & 20% into Testing
- Take 80% into Training, 20% into Testing etc.
- Large test set => estimate future error as accurately as possible (vs)
  - Large training set => better estimates
- How large should a training set be?



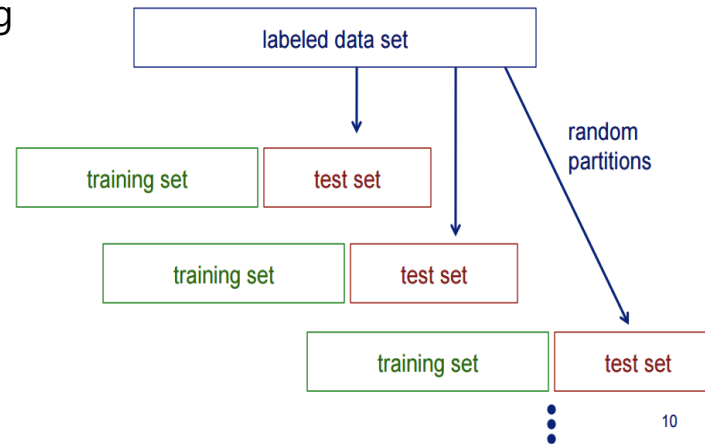
# 4. Modelling: Model Evaluation & Tuning

## Choosing Training, Validation & Testing Datasets

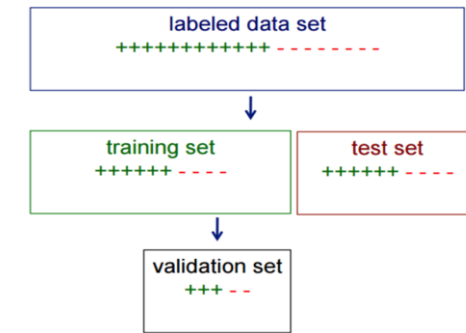
### For Imbalanced Datasets

- Random Resampling - Under & Over Sampling
- Stratified Sampling
- Resubstitution
- K fold Cross Validation
- Leave-one-out (N-fold cross-validation)
- SMOTE - Synthetic Minority Oversampling Technique
- MSMOTE - Modified SMOTE
- Cluster based sampling
- Ensemble Techniques - Bagging & Boosting
  - Ada Boost
  - Gradient Tree Boosting
  - XG Boost

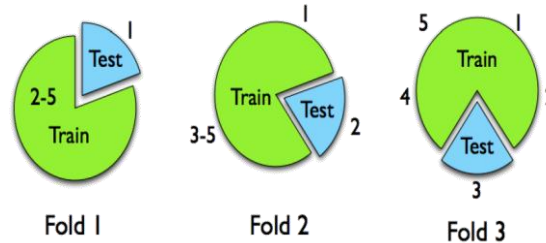
We can artificially increase training set size using random resampling



When randomly selecting training or validation sets, we may want to ensure that class proportions are maintained in each selected set



K-fold cross validation



iteration	train on	test on	correct
1	s <sub>2</sub> s <sub>3</sub> s <sub>4</sub> s <sub>5</sub>	s <sub>1</sub>	11 / 20
2	s <sub>1</sub> s <sub>3</sub> s <sub>4</sub> s <sub>5</sub>	s <sub>2</sub>	17 / 20
3	s <sub>1</sub> s <sub>2</sub> s <sub>4</sub> s <sub>5</sub>	s <sub>3</sub>	16 / 20
4	s <sub>1</sub> s <sub>2</sub> s <sub>3</sub> s <sub>5</sub>	s <sub>4</sub>	13 / 20
5	s <sub>1</sub> s <sub>2</sub> s <sub>3</sub> s <sub>4</sub>	s <sub>5</sub>	16 / 20

# 4. Modelling: Model Evaluation & Tuning

Errors & Accuracy Measures: Y is Continuous

- Mean error  $ME = \frac{1}{T} \sum_{t=1}^n e_t$

- Mean absolute deviation  $MAD = \frac{1}{n} \sum_{t=1}^n |e_t|$

- Mean squared error  $MSE = \frac{1}{n} \sum_{t=1}^n e_t^2$

- Root mean squared error  $RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$

- Mean percentage error  $MPE = \frac{1}{n} \sum_{t=1}^n \frac{e_t}{Y_t}$

- Mean absolute percentage error  $MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{e_t}{Y_t} \right|$

Actual data	Prediction Model 1	Error from model 1	Prediction Model 2	Error from model 2
100	101	1	110	10
200	199	-1	190	-10
300	301	1	310	10
400	399	-1	390	-10

Accuracy % + Error % = 100%

Accuracy + Error = 1

# 4. Modelling: Model Evaluation & Tuning

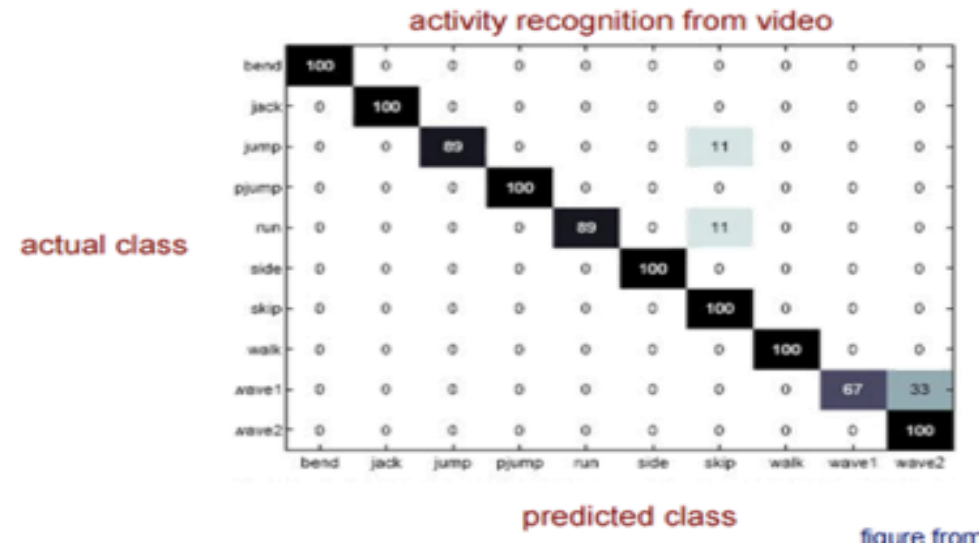
Errors & Accuracy Measures: Y is Discrete

Discrete in 2 categories

		actual class	
		positive	negative
predicted class	positive	true positives (TP)	false positives (FP)
	negative	false negatives (FN)	true negatives (TN)

$$\text{accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Discrete in multiple categories



What if there is different cost for misclassification? For e.g. Earth Quake

- E.g. Earthquake prediction
- False positive: Cost of preventive measures
- False negative: Cost of recovery
- Detection Cost (Event detection)
- $\text{Cost} = C_{FP} * FP + C_{FN} * FN$

## 4. Modelling: Model Assessment

Key 5 points to consider for Model Assessment are as follows:

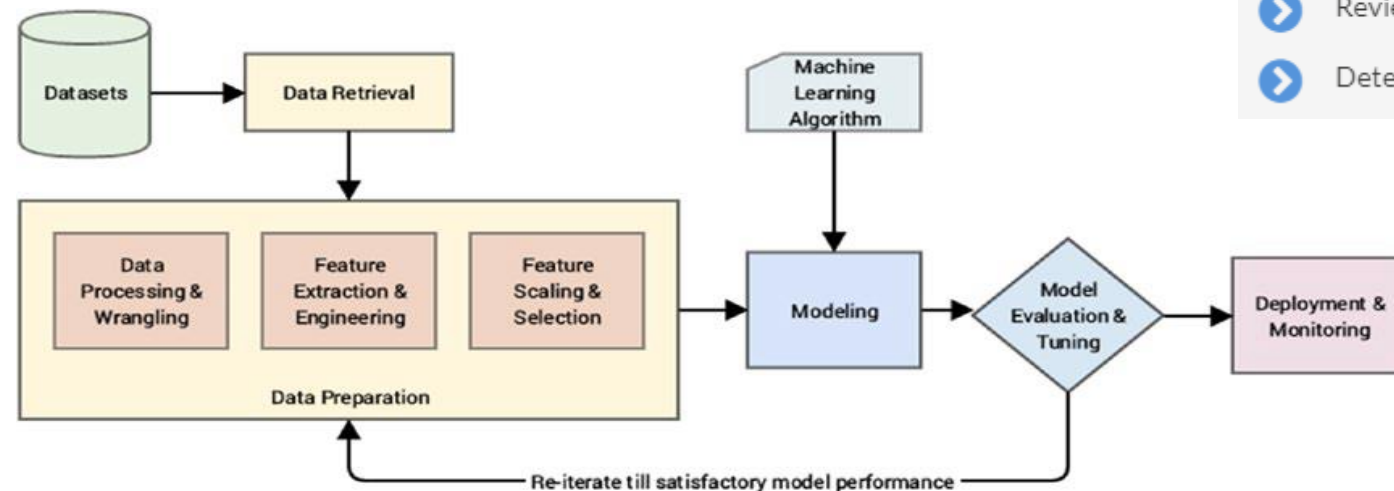
- Model performance and success criteria agreed upon earlier are in synchronization
- Model Results should be repeatable and reproducible
- Model is in line with the Non-functional requirements such as Scalable, Maintainable, Robust and Easy to Deploy
- Model evaluation gives satisfactory results
- Model is meeting business requirements

# 5. Evaluation

Key points to consider for 5th Phase of CRISP-DM includes:

- Ranking final models based on the quality of results and their relevancy based on alignment with business objectives
- Any assumptions or constraints that were invalidated by the models
- Cost of deployment of the entire Machine Learning pipeline from data extraction and processing to modeling and predictions
- Any pain points in the whole process? What should be recommended? What should be avoided?
- Data sufficiency report based on results
- Final suggestions, feedback, and recommendations from solutions team and SMEs

*A standard Machine Learning pipeline*



- Evaluate results
- Review process
- Determine next steps

## 6. Deployment

Key points to consider for 6th & Final Phase of CRISP-DM includes:

Transition from Development to Production is seamless

Proper plan for deployment based on

- Human Resources
- Servers
- Hardware
- Software, etc.

Model is saved and then deployed on proper servers and systems

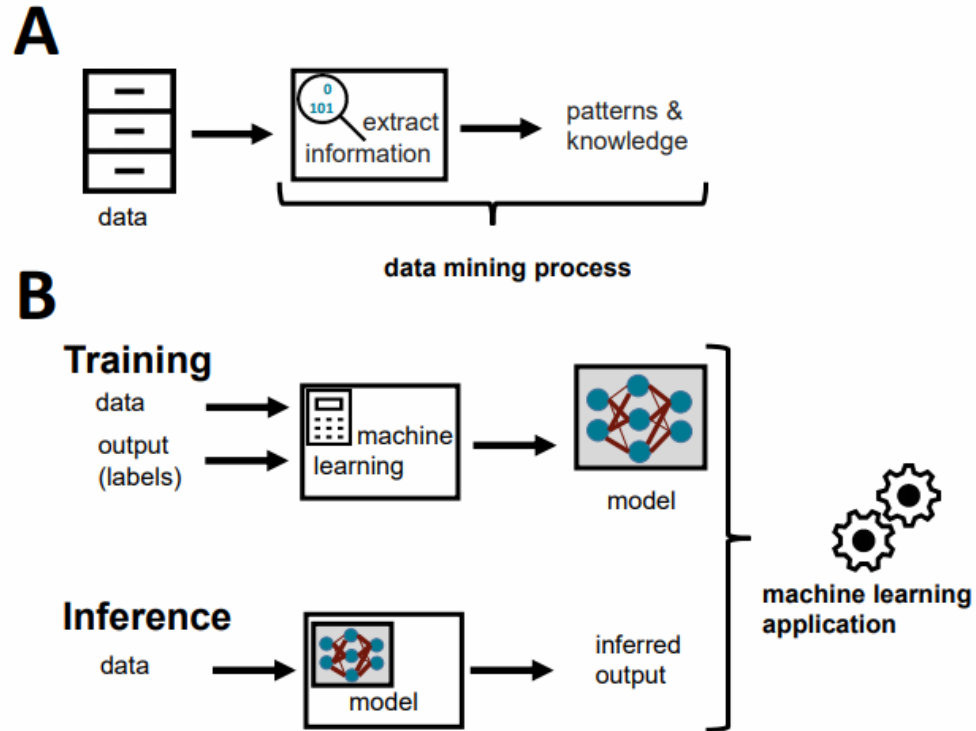
Deployment Plan & Maintenance Plans have to be prepared

Regular maintenance & monitoring on the model

- Check for results and their validity
- Retire, Replace & Update

- Plan deployment
- Plan monitoring and maintenance
- Produce final report
- Review project

# CRISP-ML(Q)



- CRISP-TDM
- CRISP-DM0
- CRISP-EM
- CRISP-MED-DM

CRISP-ML(Q)	CRISP-DM
Business & Data Understanding	Business Understanding
	Data Understanding
Data Preparation	Data Preparation
Modeling	Modeling
Evaluation	Evaluation
Deployment	Deployment
Monitoring & Maintenance	-

## Business and Data Understanding

- Define the Scope of the ML Application
- Success Criteria
  - Business Success Criteria
  - ML Success Criteria
  - Economic Success Criteria
- Feasibility
  - Applicability of ML technology
  - Legal constraints
  - Requirements on the application
- Data Collection
  - Data version control
- Data Quality Verification
  - Data description
  - Data requirements
  - Data verification
- Review of Output Documents

## Data Preparation

- Select Data
  - Feature selection
  - Data selection
  - Unbalanced Classes
- Clean Data
  - Noise reduction
  - Data imputation
- Construct Data
  - Feature engineering
  - Data augmentation
- Standardize Data
  - File format
  - Normalization

## Modeling

- Literature research on similar problems
- Define quality measures of the model
- Model Selection
- Incorporate domain knowledge
- Model training
- Using unlabeled data and pre-trained models
- Model Compression
- Ensemble methods
- Assure reproducibility
  - Method reproducibility
  - Result reproducibility
  - Experimental Documentation

## Evaluation

- Validate performance
  - Determine robustness
  - Increase explainability for ML practitioner & end user
  - Compare results with defined success criteria

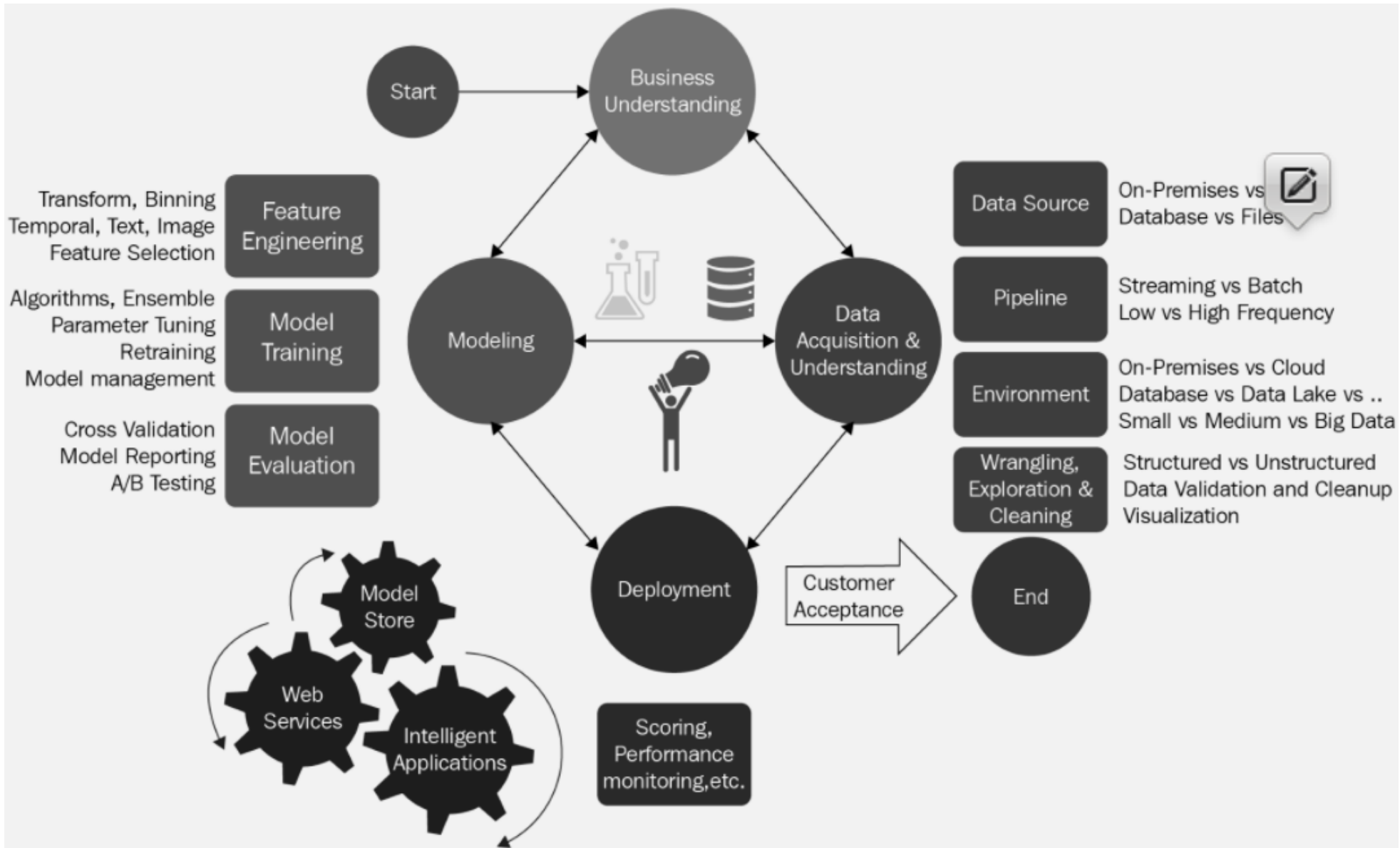
## Deployment

- Define inference hardware
- Model evaluation under production condition
- Assure user acceptance and usability
- Minimize the risks of unforeseen errors
- Deployment strategy

## Monitoring and Maintenance

### Causes of Violation:

- Non-stationary data distribution
- Degradation of hardware
- System updates
- Monitor
- Update



## Five broad stages

- Business understanding
- Data acquisition and understanding
- Modeling
- Deployment
- Customer acceptance

## Business understanding

### Two Tasks

#### Defining objectives

Model targets, Relevant questions, Roles and milestones, Success metrics

#### Identifying data sources

Data that has an impact on the question, directly or indirectly

Data that directly measures the model target and the important features

## Deliverable

- Charter document
- Data sources
- Data dictionaries

## Data acquisition and understanding

- Ingest data
- Explore data
- Data pipeline

## Deliverable

- Data quality report
  - Solution architecture
  - Checkpoint decision
- 

## Modeling

- Feature engineering
- Model training

## Deliverable

- Feature sets
  - Model report
  - Checkpoint
- 

## Deployment

- A dashboard that shows the health and key metrics of the prediction system
- A modeling report that shows the deployment details
- A solution architecture document capturing the various components of the solution

## Customer acceptance

- **System validation:** Confirming that the deployed model and data pipeline meet the stakeholders needs
- **Project hand-off:** Hand the system off to the group that is going to run the system in production

- You have learnt about the 4 Stages of Analytics
- You have learnt about the Data Analytics Project Management Steps (CRISM – DM)