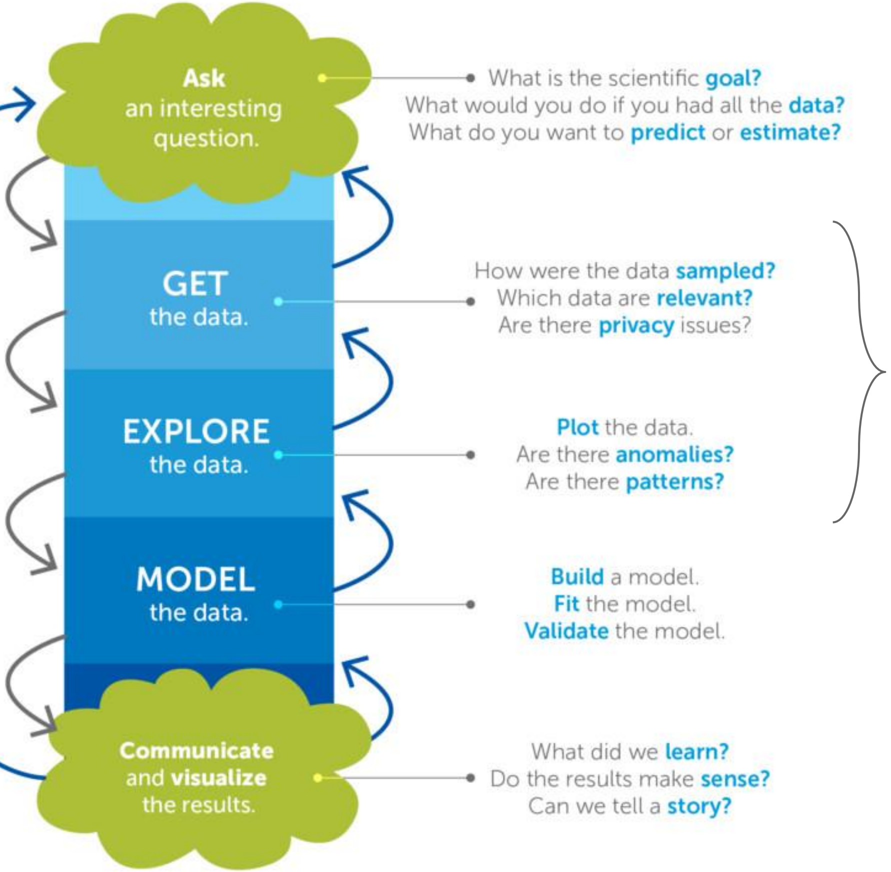
Workshop
Part 4 - Data Mining

by Kanda Tiwatthanont

Data Mining Process



Ask an interesting question.

- What is the scientific **goal?** What would you do if you had all the **data?** What do you want to **predict** or **estimate?**

GET the data.

- How were the data **sampled?** Which data are **relevant?** Are there **privacy** issues?

EXPLORE the data.

- **Plot** the data. Are there **anomalies?** Are there **patterns?**

MODEL the data.

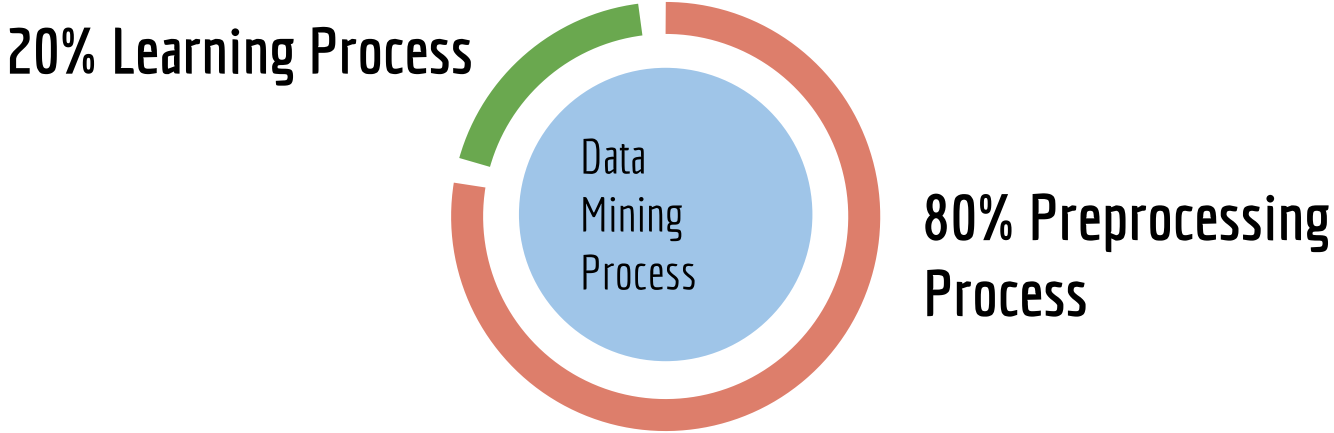
- **Build** a model. **Fit** the model. **Validate** the model.

Communicate and visualize the results.

- What did we **learn?** Do the results make **sense?** Can we tell a **story?**

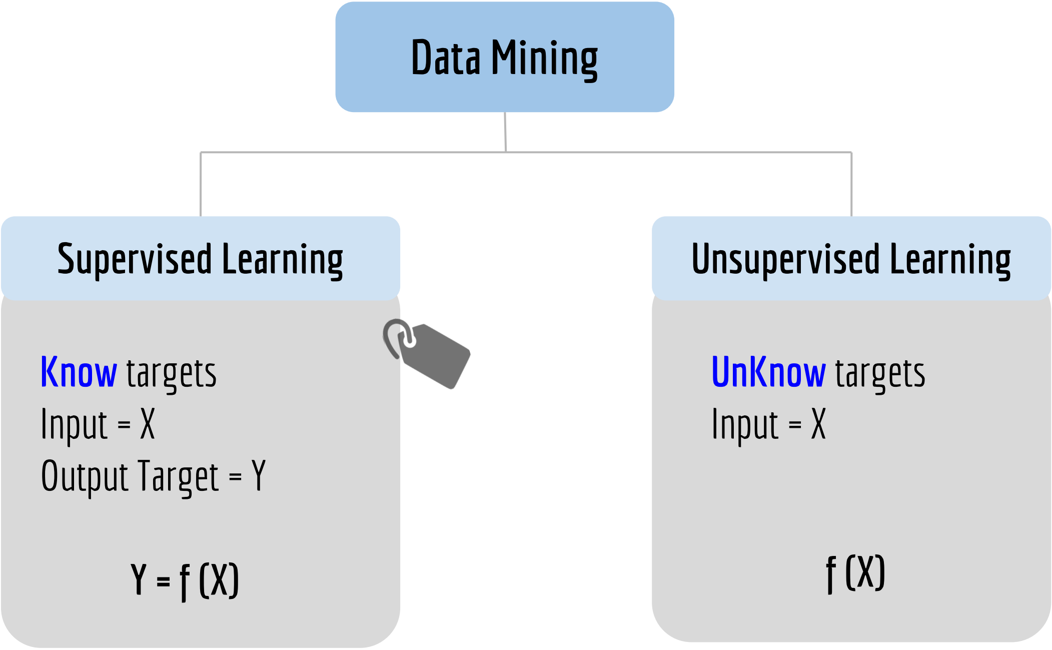
Preprocessing Process

Data Mining Process



Ref : http://www.dbjournal.ro/archive/14/14_3.pdf

Data Mining



Data Mining

Supervised Learning

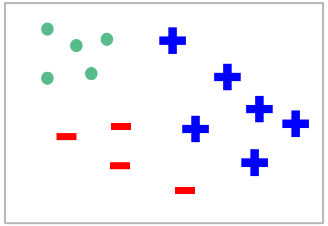
Know targets
Input = X
Output Target = Y

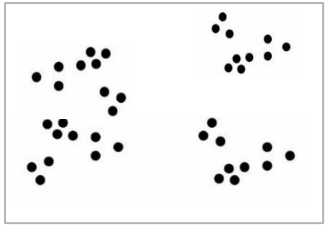
Y = f (X)

Unsupervised Learning

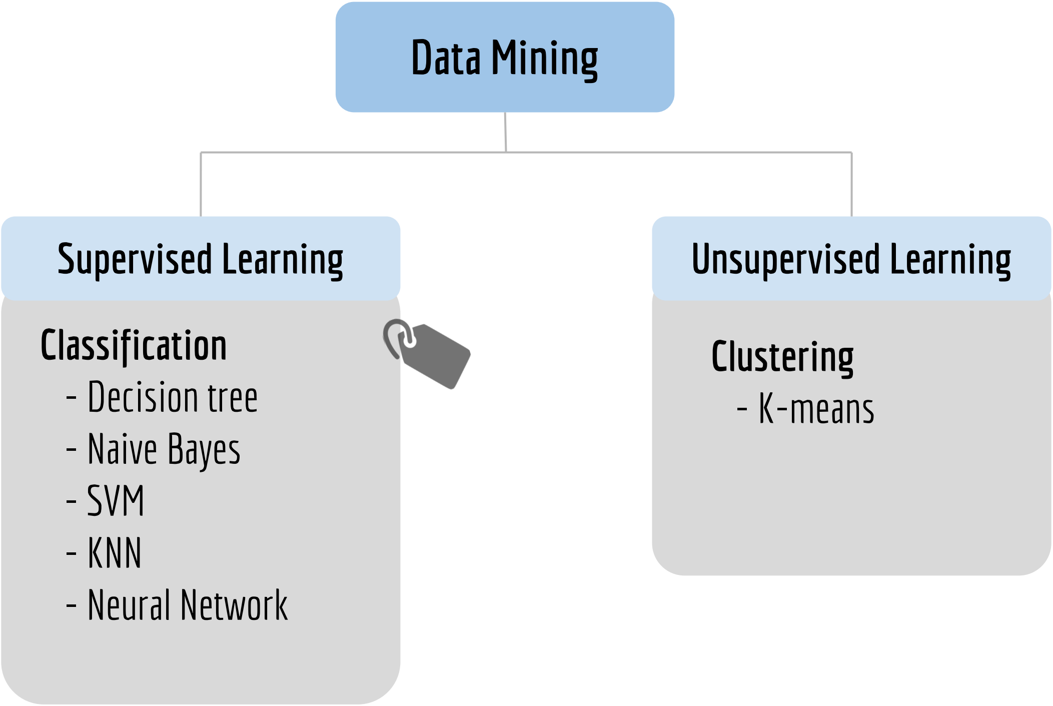
UnKnow targets
Input = X

f (X)





Data Mining



Data Mining

Supervised Learning

Classification

- Decision tree
- Naive Bayes
- SVM
- KNN
- Neural Network

Unsupervised Learning

Clustering

- K-means

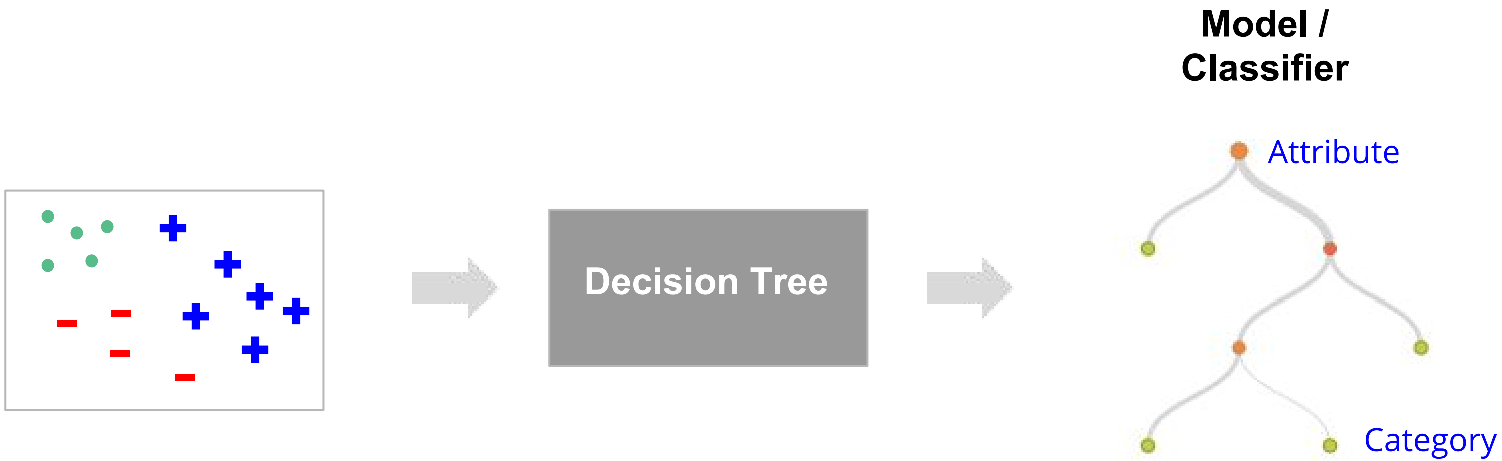
Overview

Part 4 : Data Mining (DM)

- Tasks
    - Classification with **DT** or **SVM**
    - Clustering with **k-mean**
- Model Evaluation
- Hands-on
    - Scikit-learn -- Machine Learning Tool for Data Scientist
    - Try predicting data

Classification - Decision Tree

Decision Tree - find the best attribute for decision node



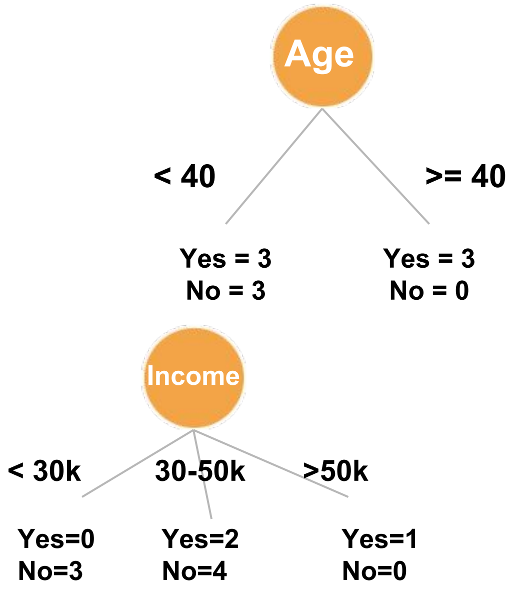
Decision Tree

Model / Classifier

Attribute

Category

Classification - Decision Tree



Decision Tree

#1: Select the best attribute for root node

#2 Create branches for all possible values

#3: Split instances into subsets

Loop : Repeat recursively (#1,#2,#3) for each branch

Until: All instances of the subset have the same class, or have a single value

Classification - Decision Tree

Decision Tree

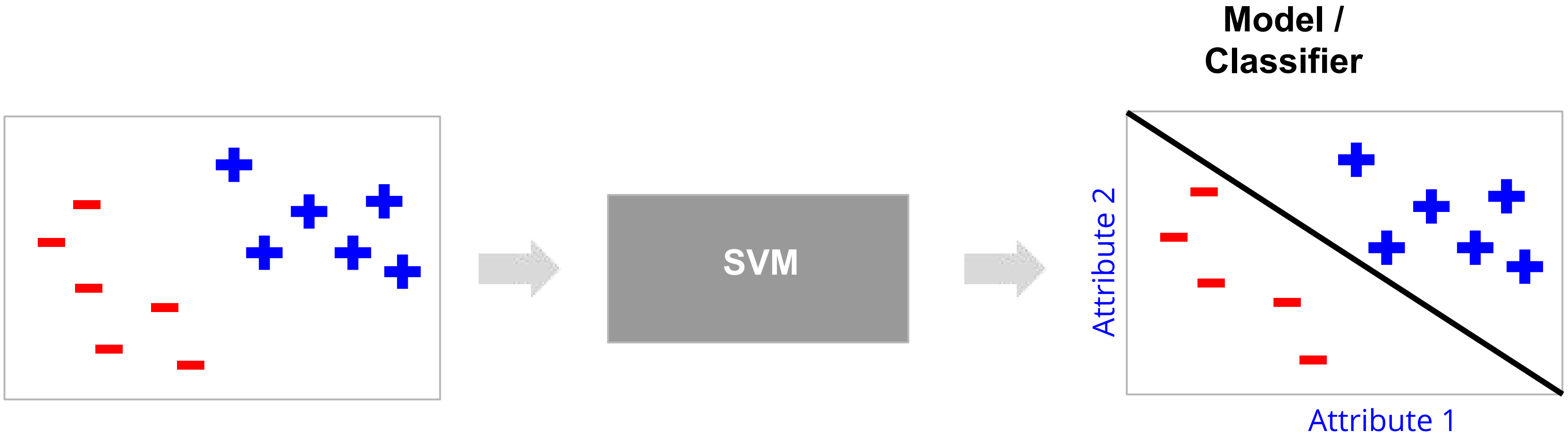
The best attribute

The way to select the best attribute

- CART Algorithm → **Gini**
- ID3 Algorithms → **Information Gain**
- C4.5 Algorithms → **Entropy**

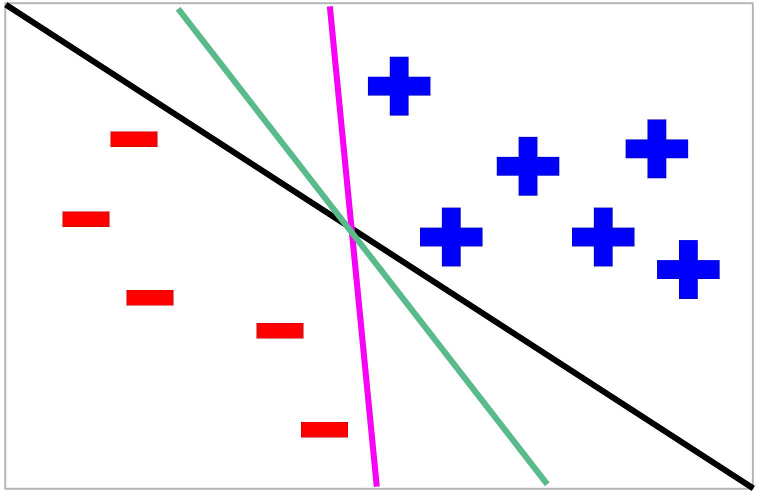
Classification - Support Vector Machine (SVM)

SVM - find the optimal hyperplane



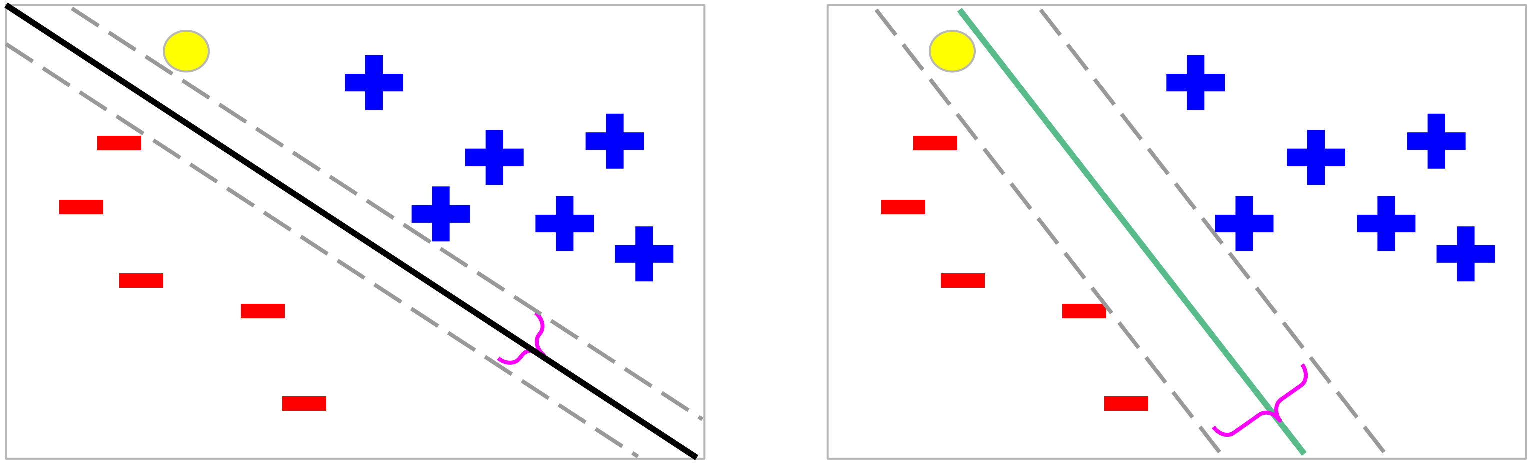
Classification -SVM

SVM - What is the optimal hyperplane ?



Classification - SVM

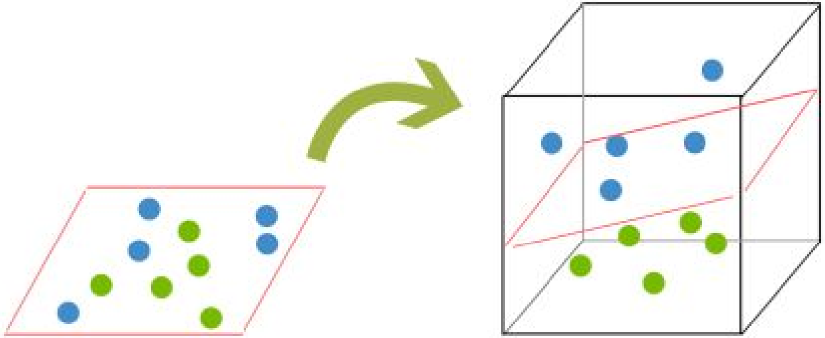
SVM - What is the optimal hyperplane ?



find optimal hyperplane that maximum margin

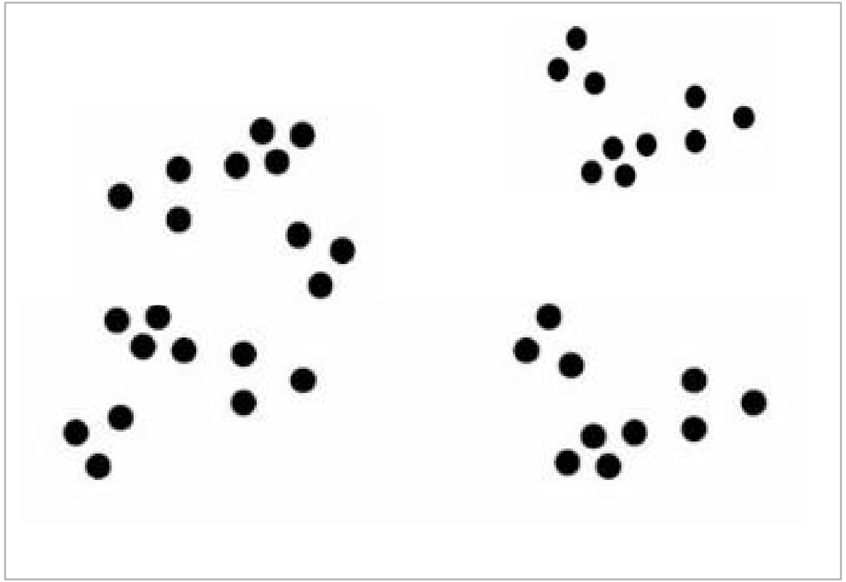
Classification - SVM

SVM - Effective on high dimensional data



Clustering - k-means

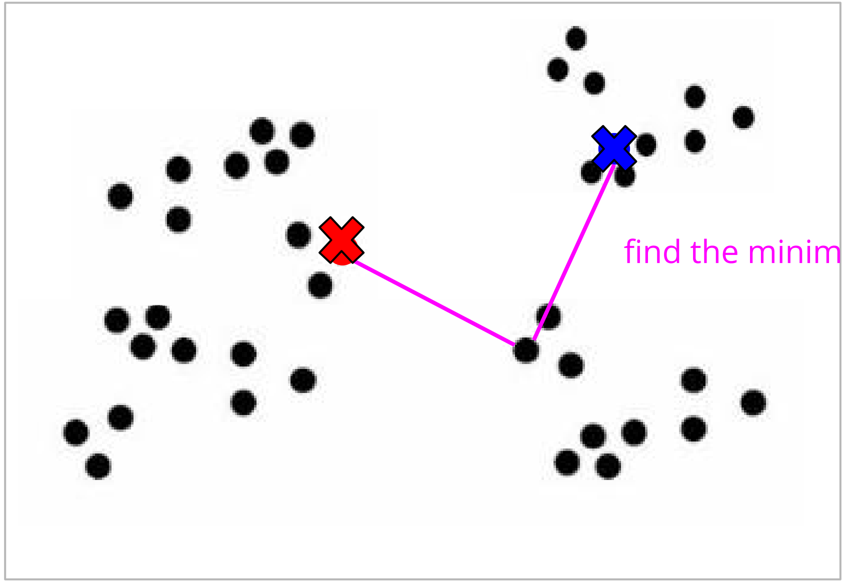
k-means - define k centroids, find the closest cluster to each data point



Clustering - k-means

k-means - define **2** centroids, find the closest cluster to each data point

Round 1

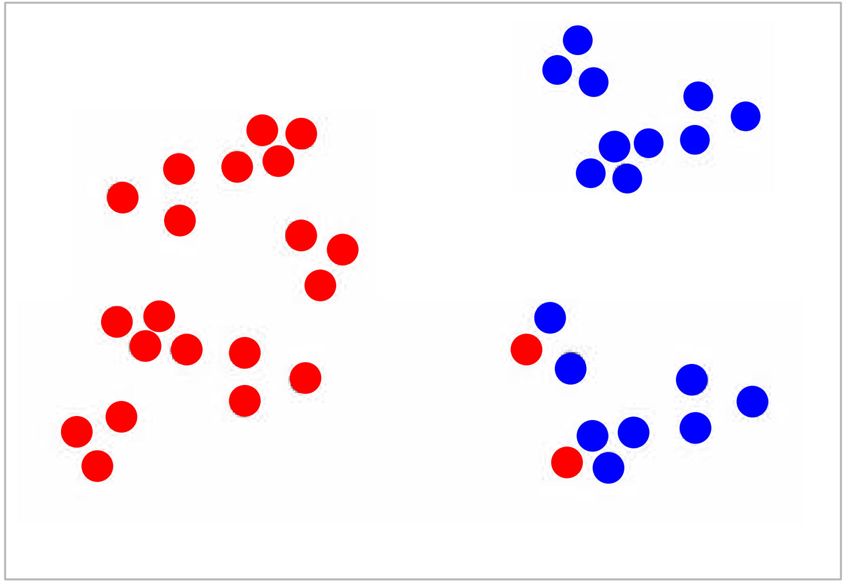


find the minimum distance

Clustering - k-means

k-means - define **2** centers, find the closest cluster to each data point

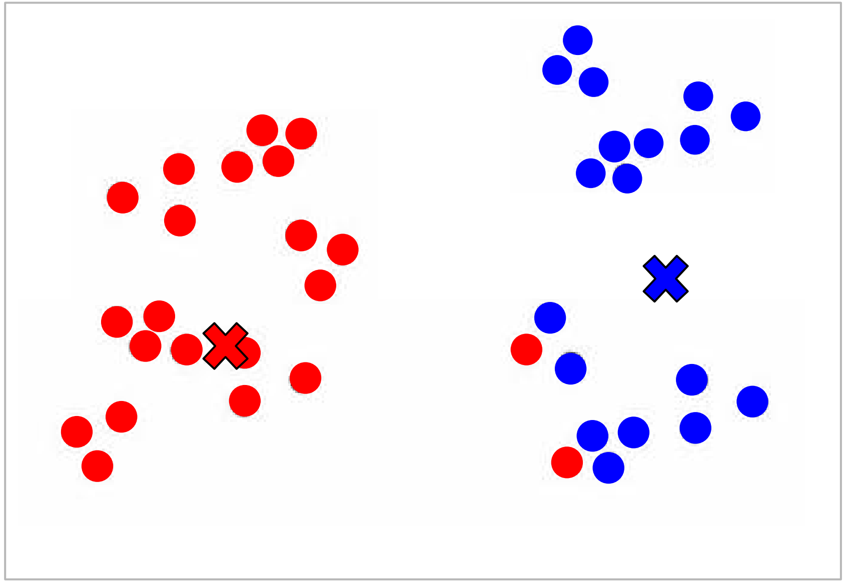
Round 1



Clustering - k-means

k-means - define 2 centers, find the closest cluster to each data point

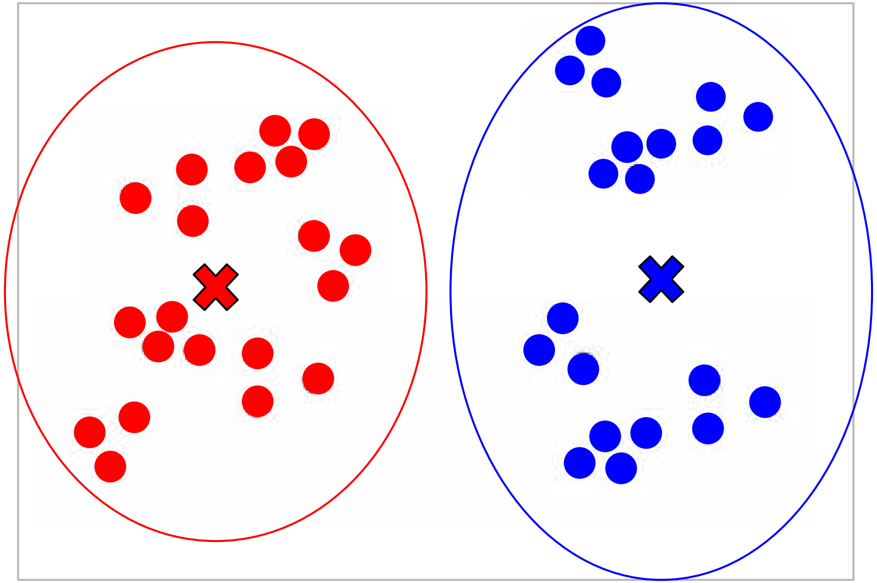
Round 2



Clustering - k-means

k-means - define **2** centers, find the closest cluster to each data point

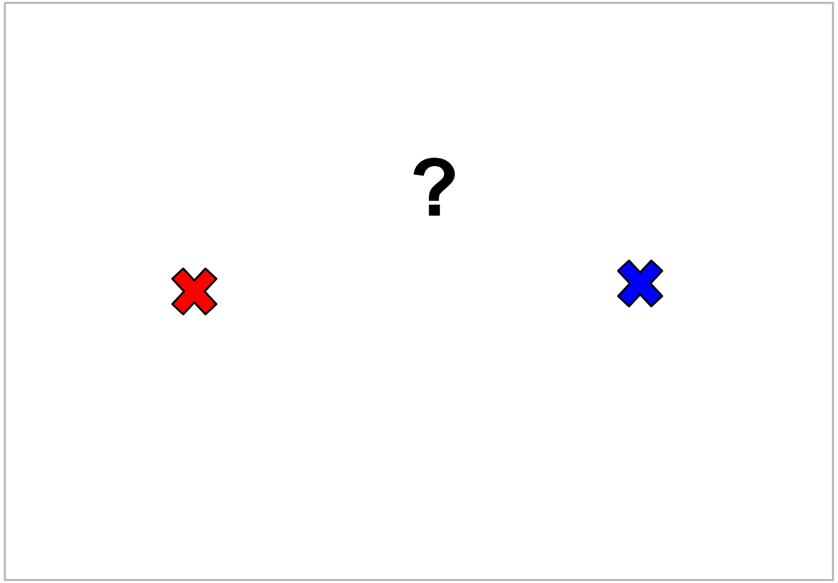
Round n



Clustering - k-means

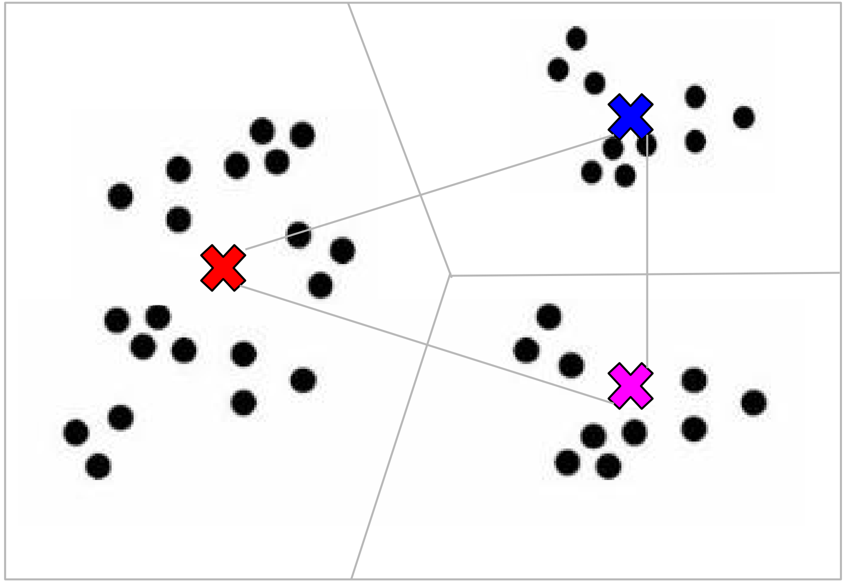
k-means - define 2 centers, find the closest cluster to each data point

Unseen data

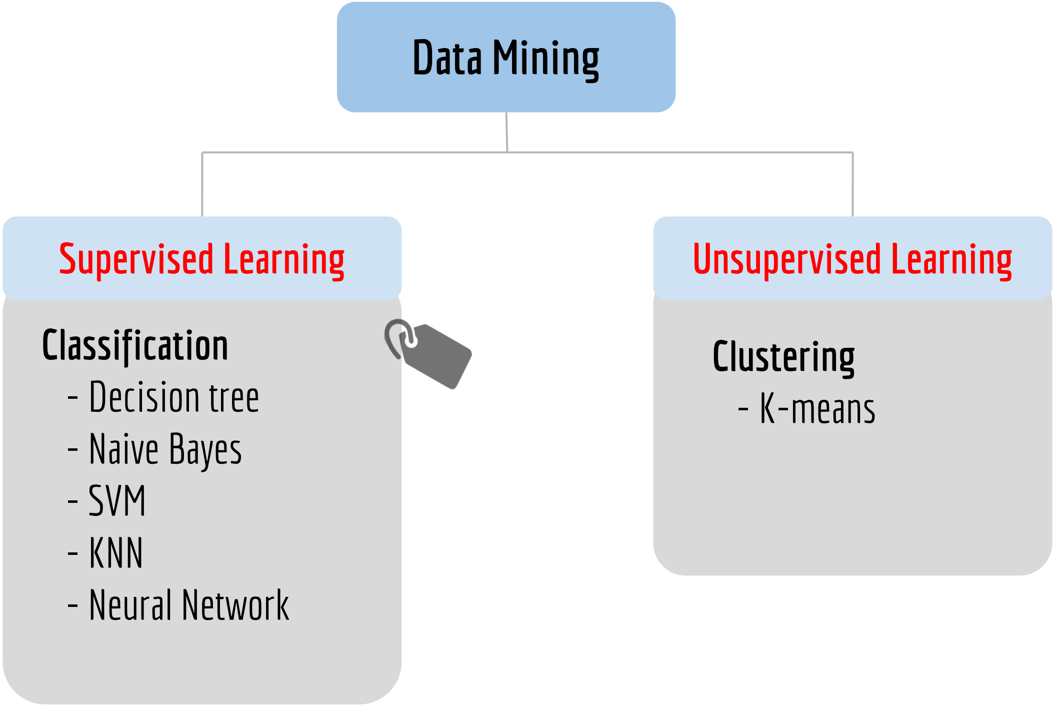


Clustering - k-means

k-means - define 3 centers, find the closest cluster to each data point



Data Mining



Data Mining

Supervised Learning

Classification
- Decision tree
- Naive Bayes
- SVM
- KNN
- Neural Network

Unsupervised Learning

Clustering
- K-means

Supervised or Unsupervised Learning ?

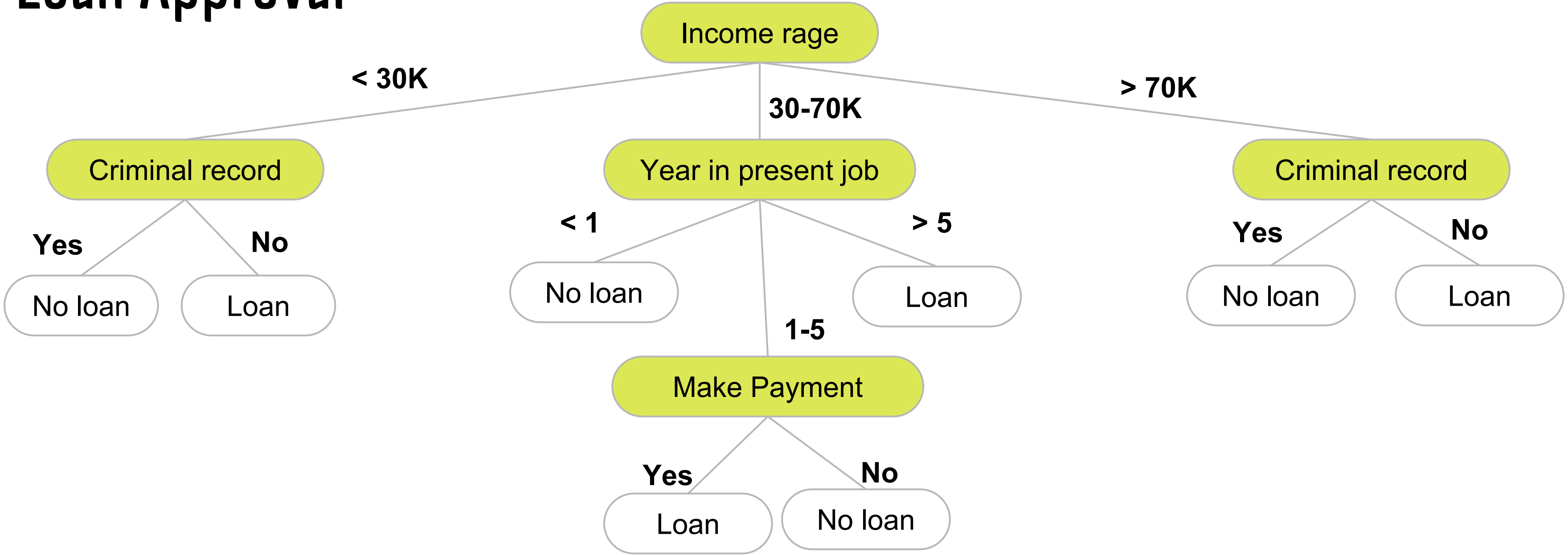
Supervised or Unsupervised Learning ?

Bank loan dataset

Customer ID	Sex	Income	Year in present job	Make Payment	Criminal record	Decision
1	M	72,000	15	Yes	No	Loan
2	F	35,000	3	Yes	Yes	No Loan
3	M	28,000	2	No	No	Loan
...						

Application of *Decision Tree*

Loan Approval



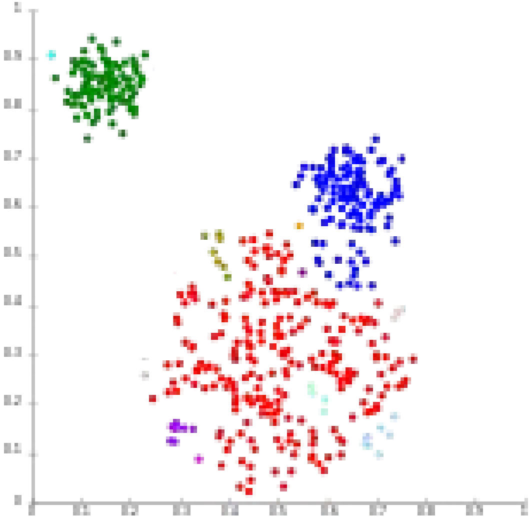
Ref : https://www.youtube.com/watch?v=0sJMMbjjjZM

Supervised or Unsupervised Learning ?

Insurance: Identifying groups of motor insurance policy holders with a high average claim cost.

Application of *Clustering*

Insurance



🟥

Retire
Officer

🟦

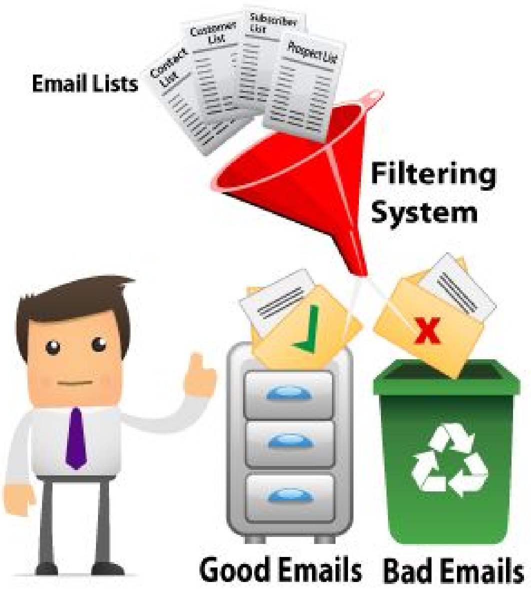
The
Yo-Pro

🟩

The New
driver

Ref : https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/

Supervised or Unsupervised Learning ?



Application of *Naive Bayes* or *SVM*

Spam Filtering

Email Lists
Contact List
Customer List
Subscriber List
Prospect List
Filtering System
Good Emails
Bad Emails

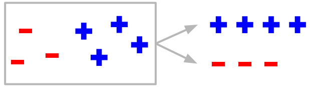
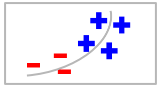
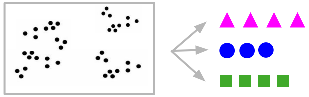
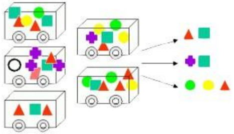
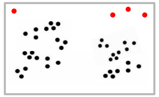
Ref : https://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/

Why Spam Filtering does *not* use Decision Tree ?

Supervised or Unsupervised Learning ?

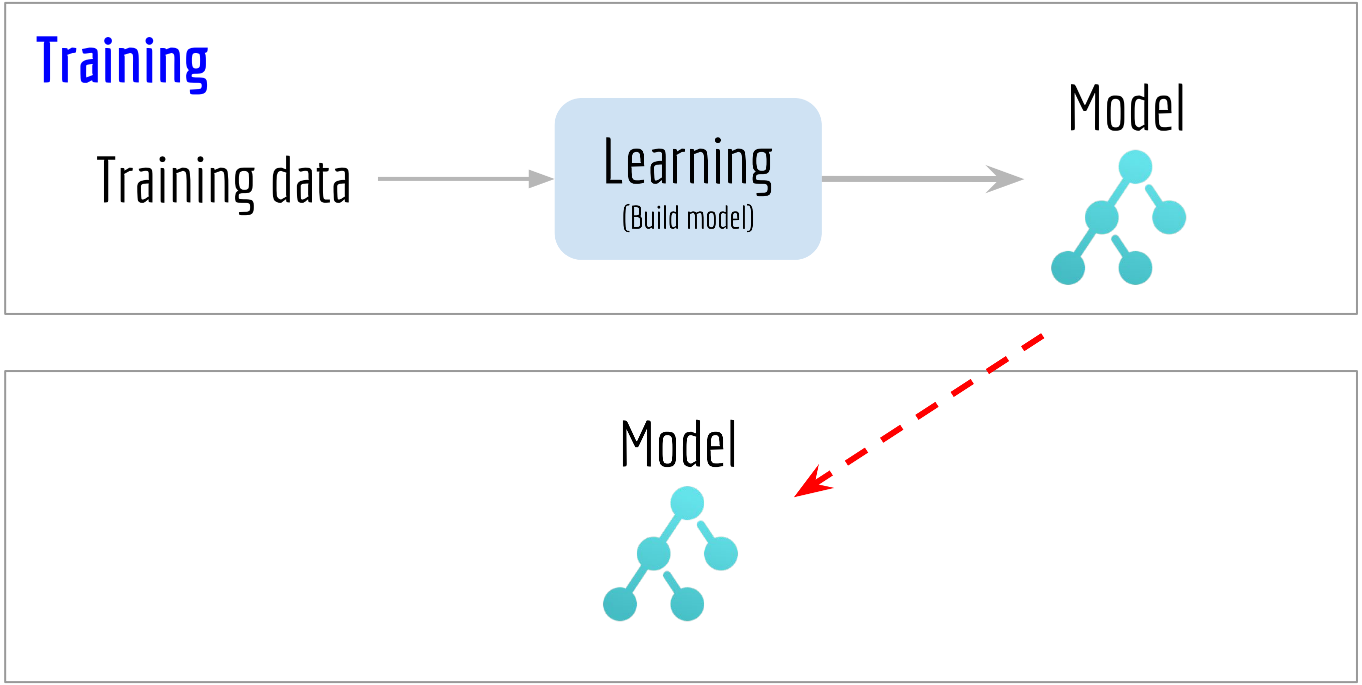
data
techniques
information
Text
Mining
language
human
Document

Data Mining Tasks

Techniques	Algorithms
Classification 	Decision Tree Naive bayes SVM
Regression 	Linear Regression
Clustering 	K-means
Association 	Apriori FP-Growth
Anomaly Detection 	One class SVM

Are those all about Data Mining ?

Model Evaluation



Training

Training data

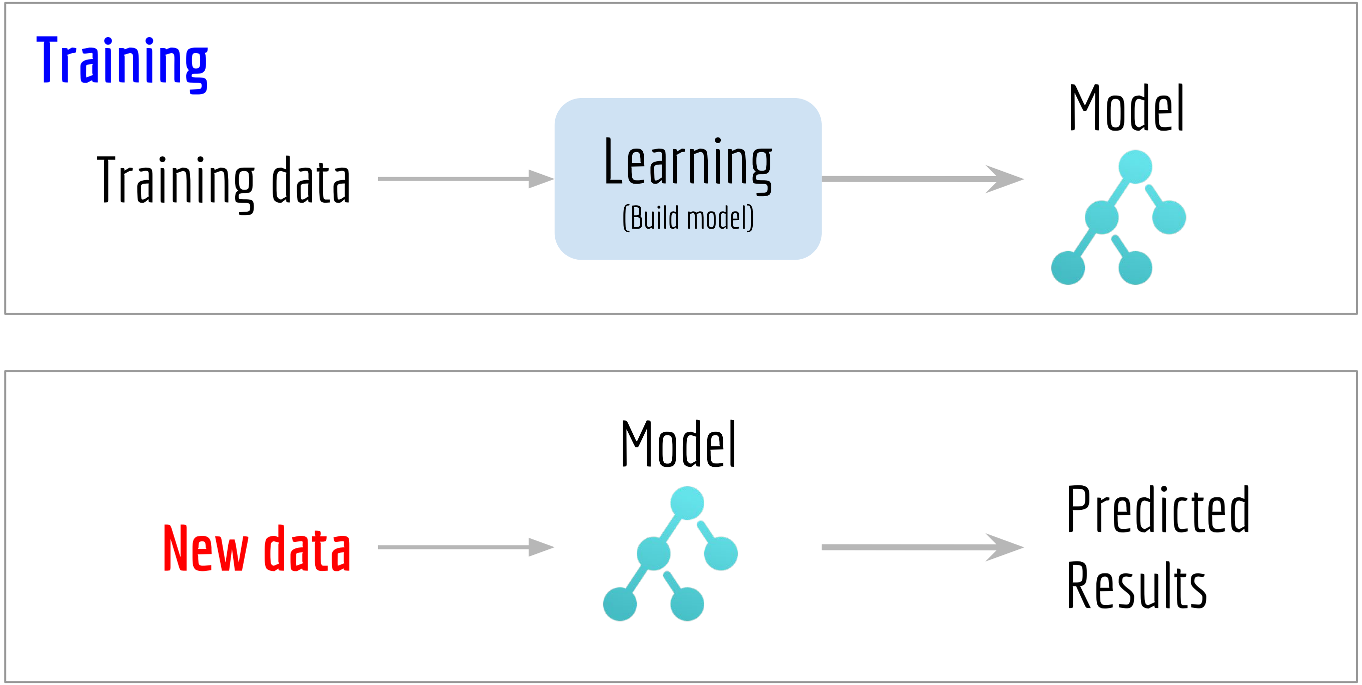
Learning

(Build model)

Model

Model

Model Evaluation



Training

Training data

Learning

(Build model)

Model

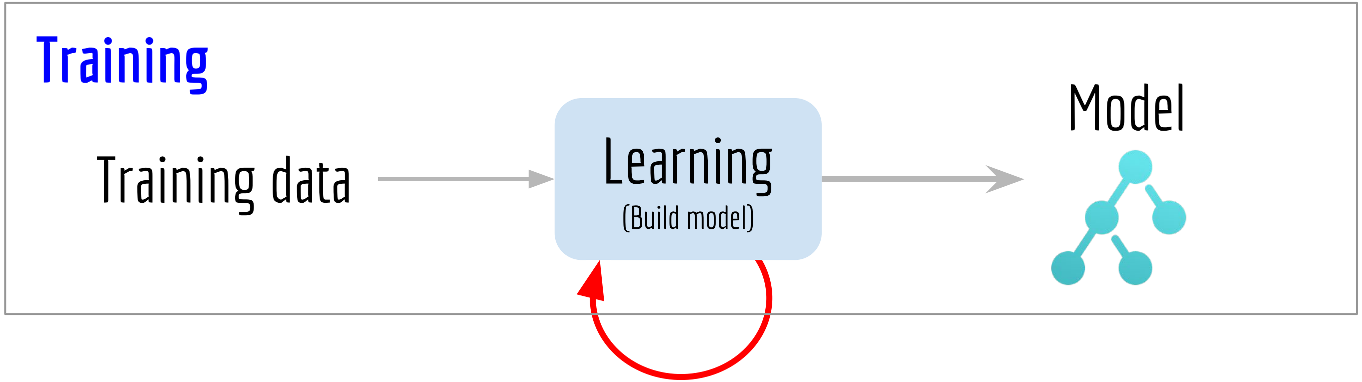
New data

Model

Predicted Results

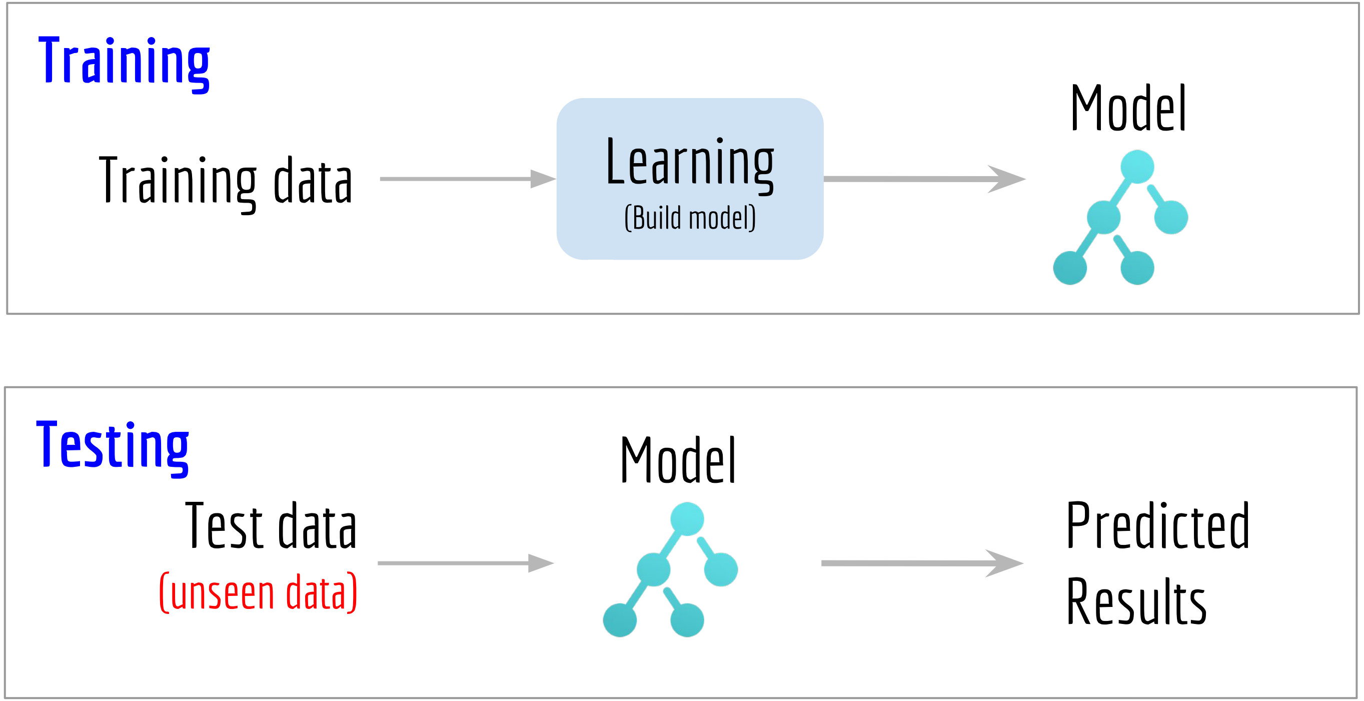
Good Enough Model ?

Model Evaluation



k-fold Cross Validation

Model Evaluation



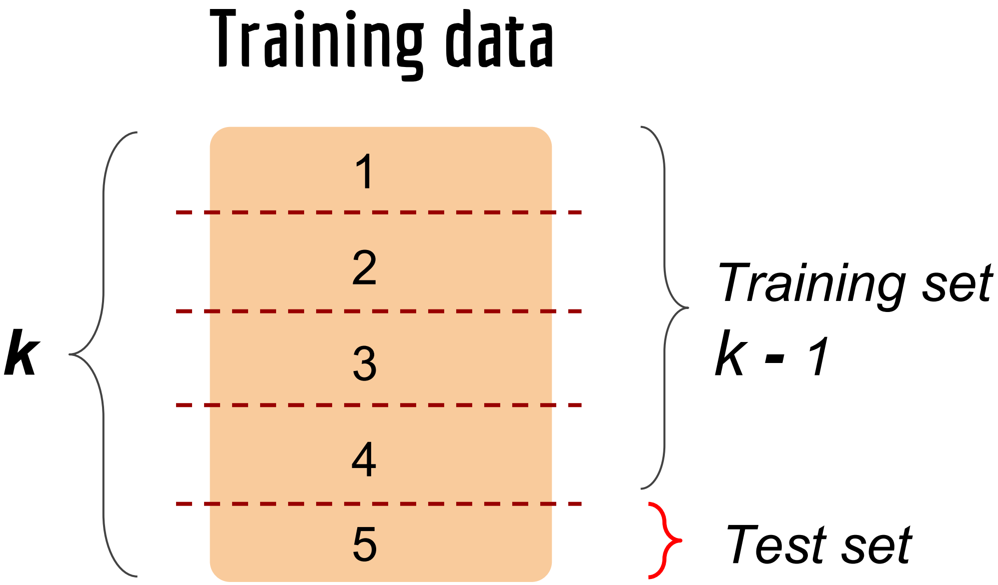
Training

Training data → Learning (Build model) → Model

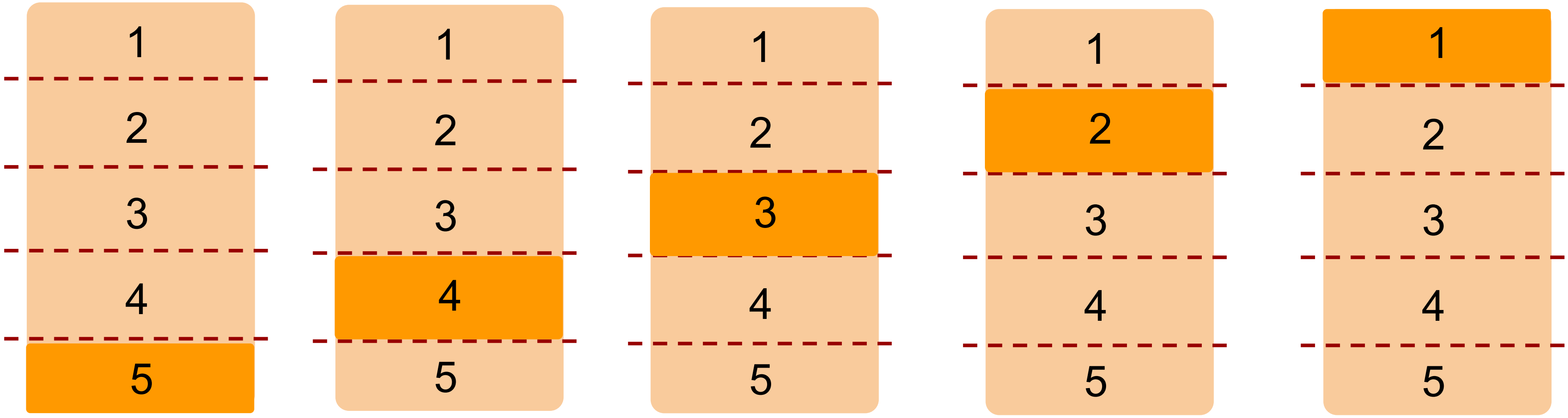
Testing

Test data (unseen data) → Model → Predicted Results

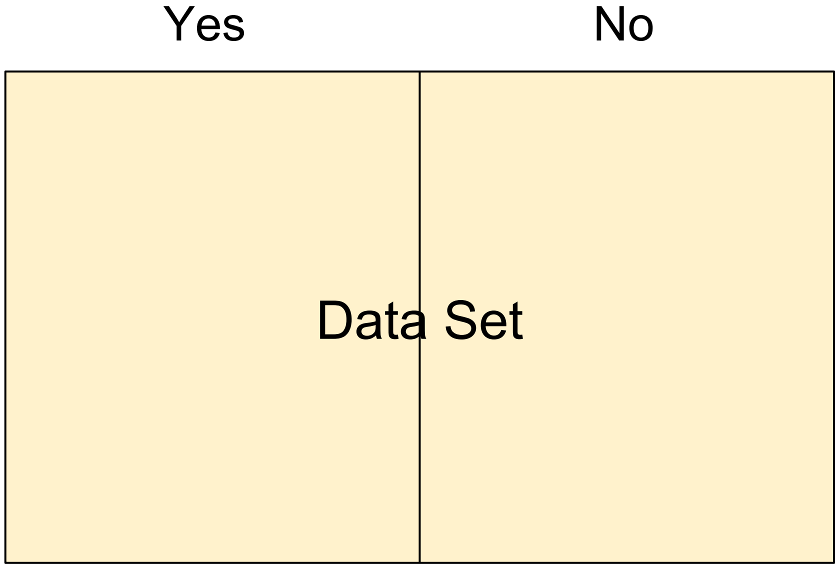
k-fold Cross Validation



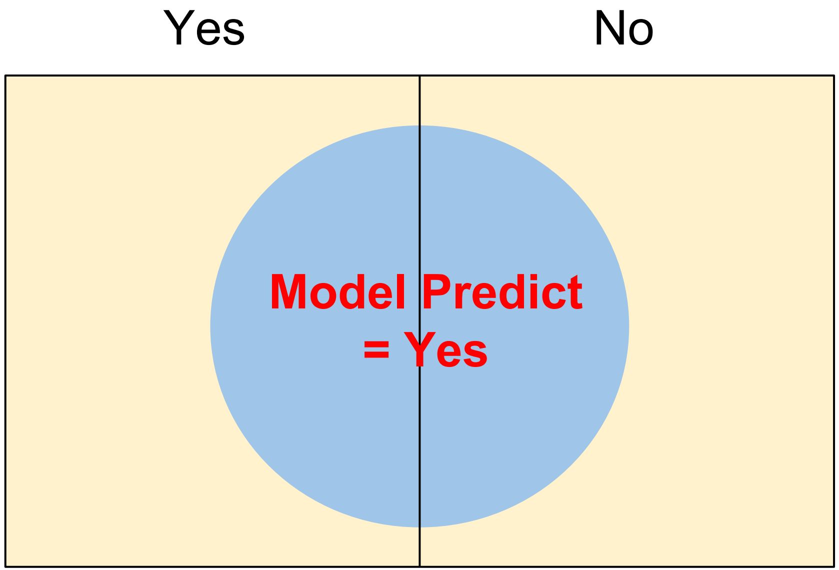
k-fold Cross Validation



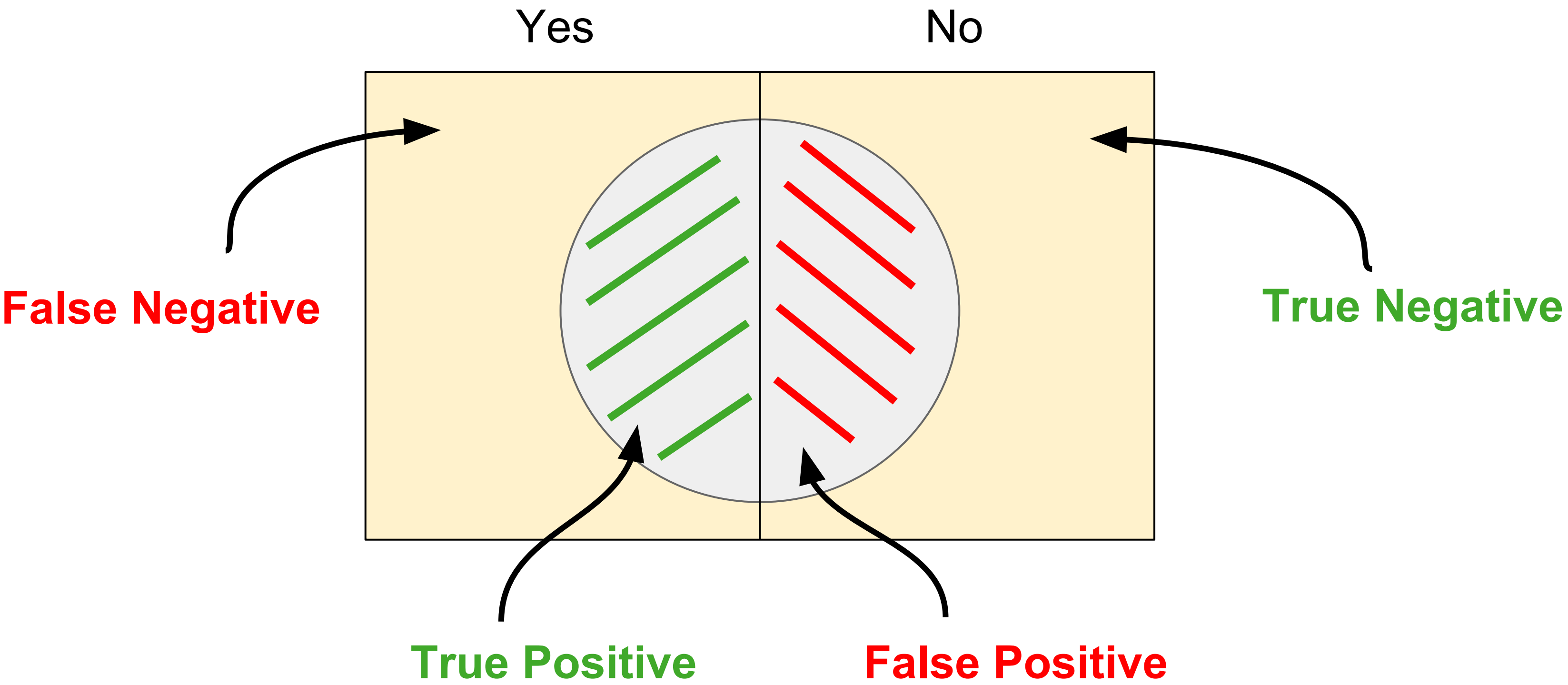
Evaluation Method



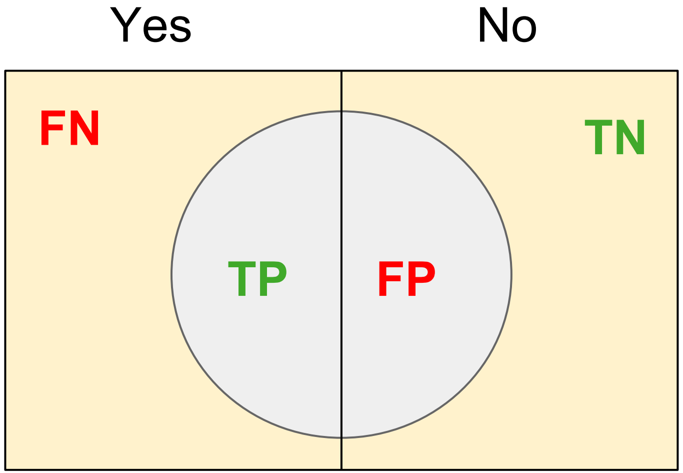
Evaluation Method



Evaluation Method



Evaluation Method



Accuracy

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$\frac{TP}{TP + FP}$$

Recall

$$\frac{TP}{TP + FN}$$

F1-measure

Precision x Recall

จบ