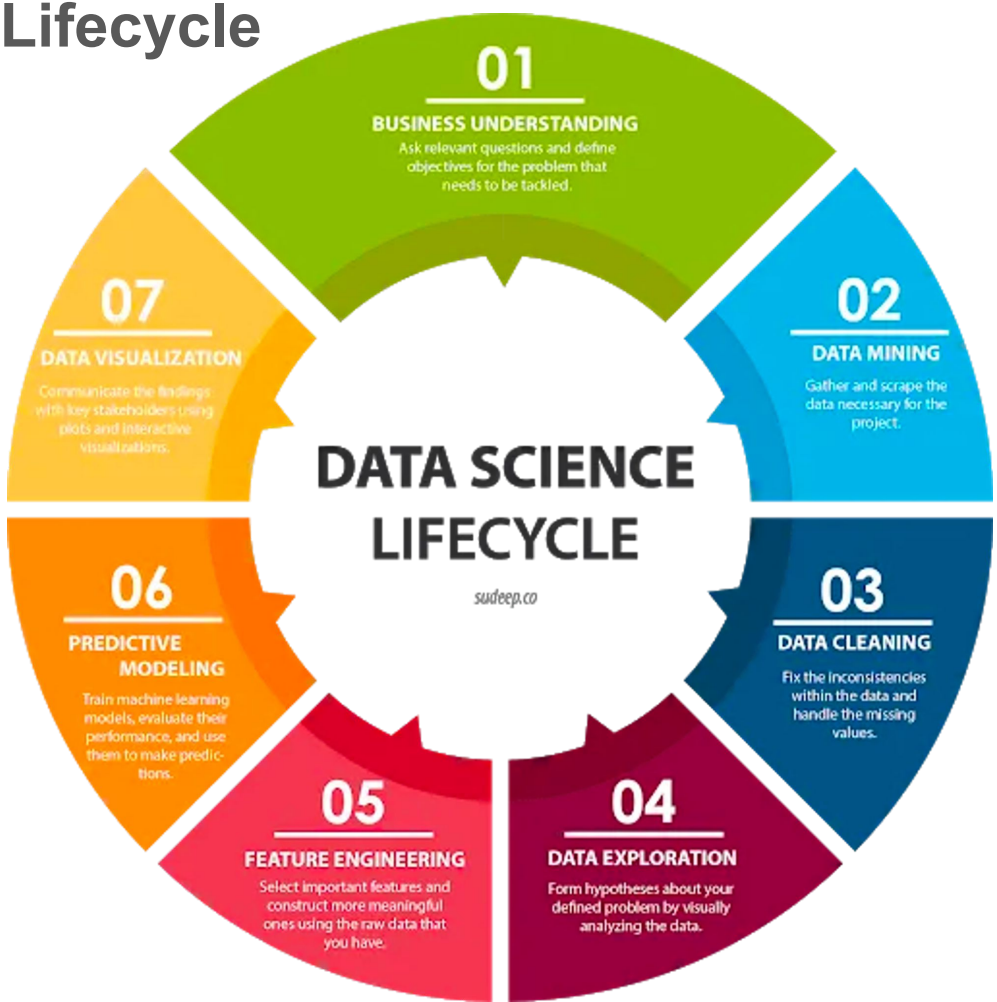
Introduction to
Data Science

Fırat Öncü
Sr. MLOps Engineer

Contact:
+90 537 619 36 49
f.firatoncu@gmail.com

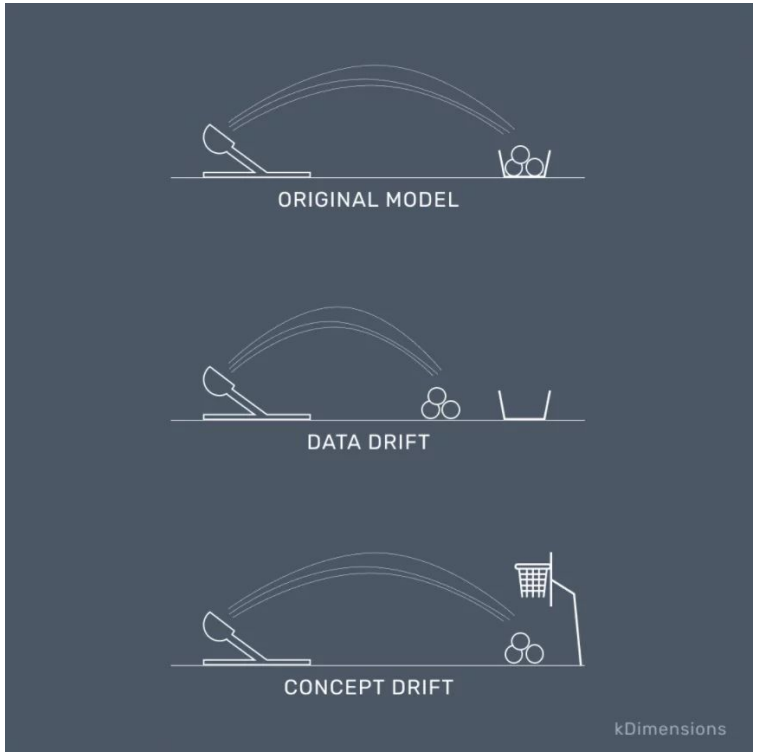
Data Science Nedir?

Data Science Lifecycle



Data Drifts

Data Drift



Concept Drift

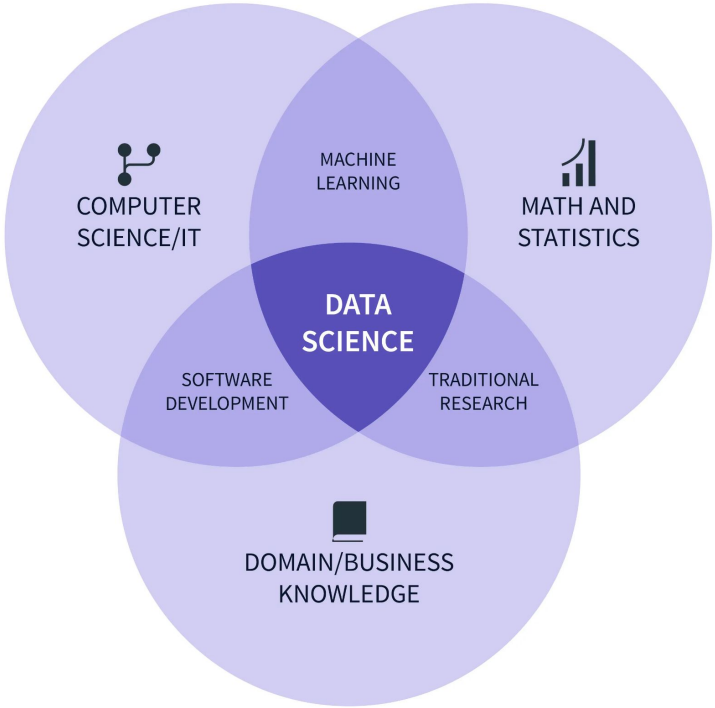
Data Scientists’ Skill Set

Programlama Dilleri

- **Python**
  - Temel Fonksiyonlar
  - Kütüphaneler
  - Kodlama Standartları
  - Test ve Validasyon Akışları
- **SQL**
  - MySQL, MSSQL, BQL, HQL..
  - Join Yapıları
- **Unix**
  - Ubuntu, Debian, CentOS

Analitik Bilimler

- Matematik
- Olasılık
- İstatistik



Altyapı

- Veritabanları (Hadoop, MongoDB, ELK)
- Cloud (GCP, S3, Azure)
- **Yardımcı Araçlar**
  - MlFlow
  - Airflow

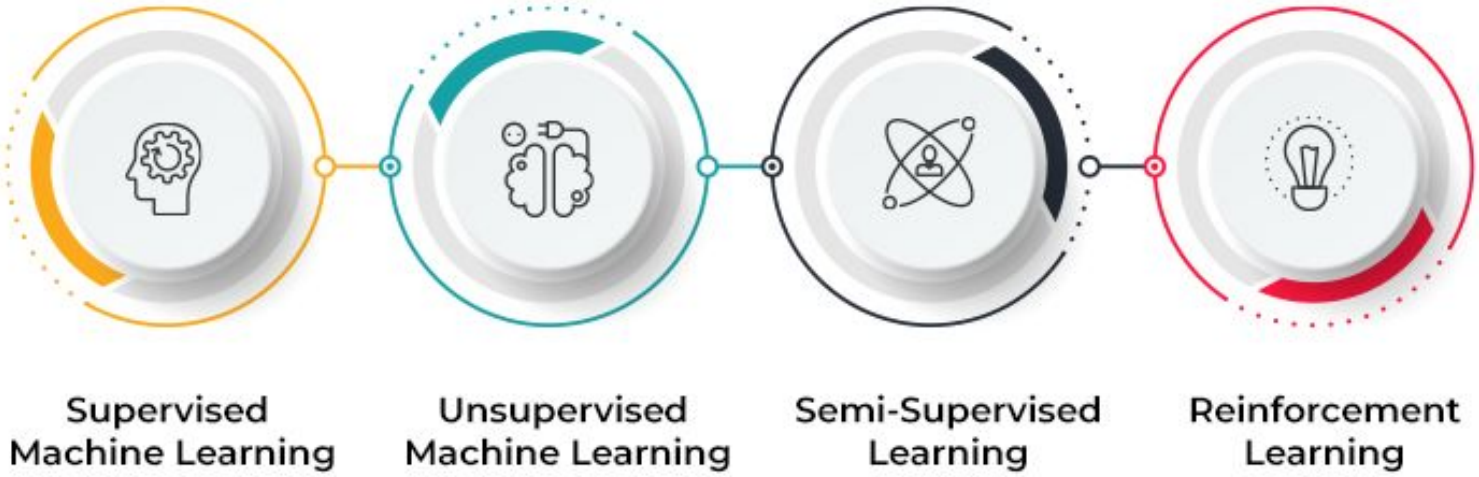
Raporlama

- Veri Görselleştirme
- Çıktıların Yorumlanması

Diğer Beceriler

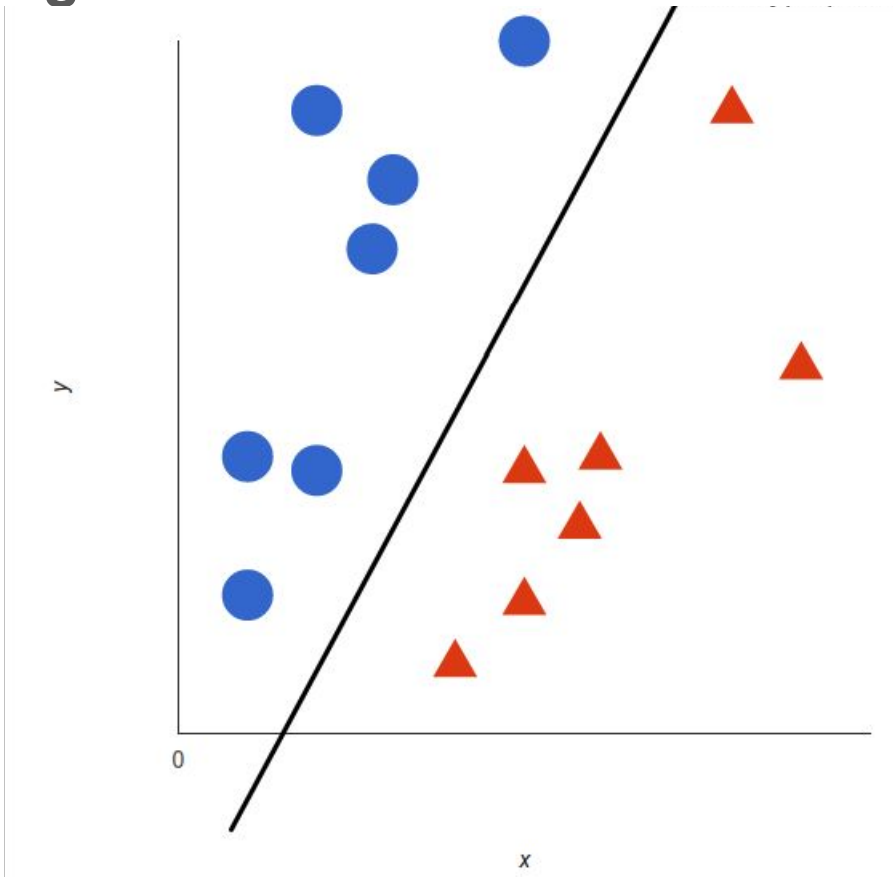
- Analitik Düşünce
- İletişim
- Hikaye Anlatıcılığı
- Sürekli Öğrenme
- Takım Uyumu

Machine Learning Nedir?

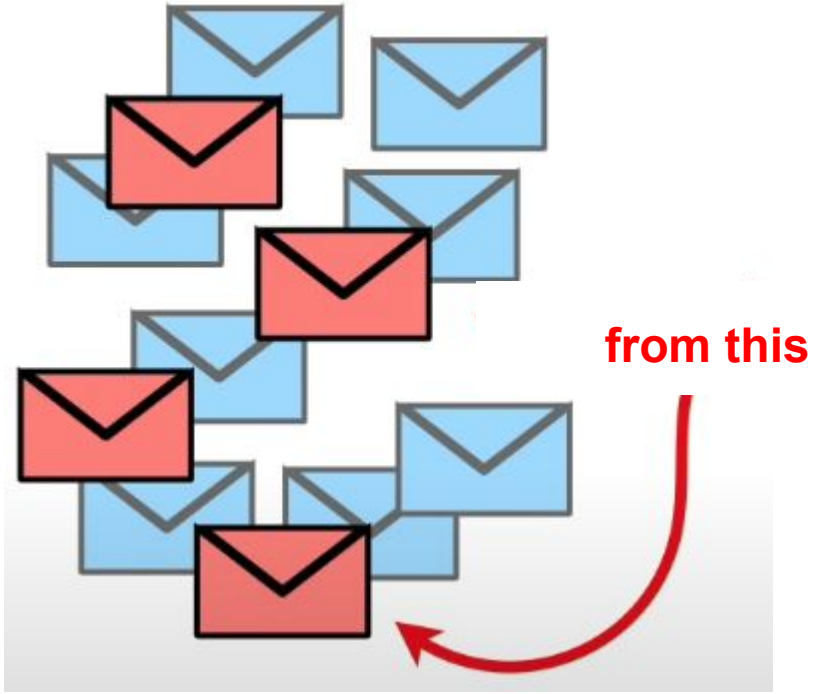


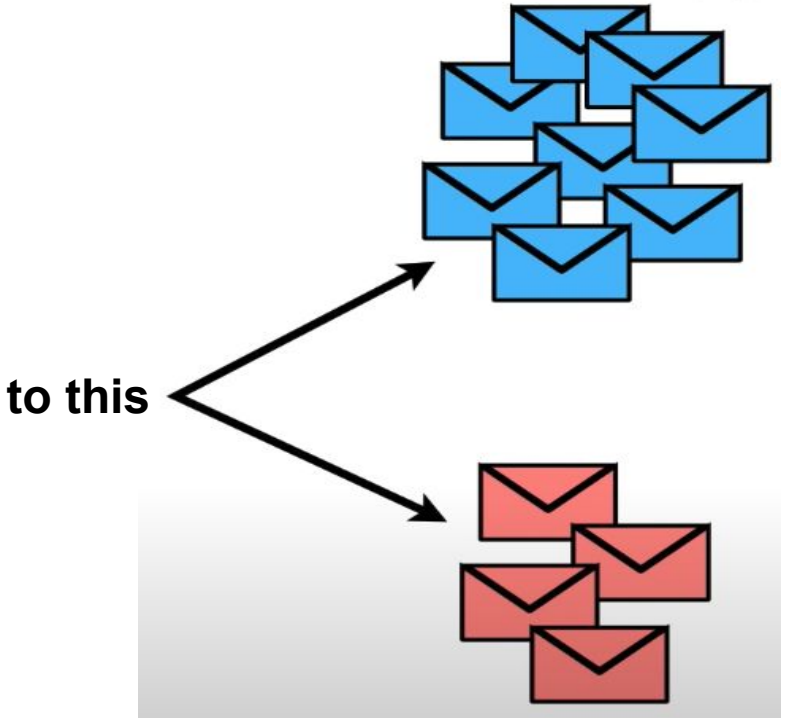
Supervised Learning

- Classification

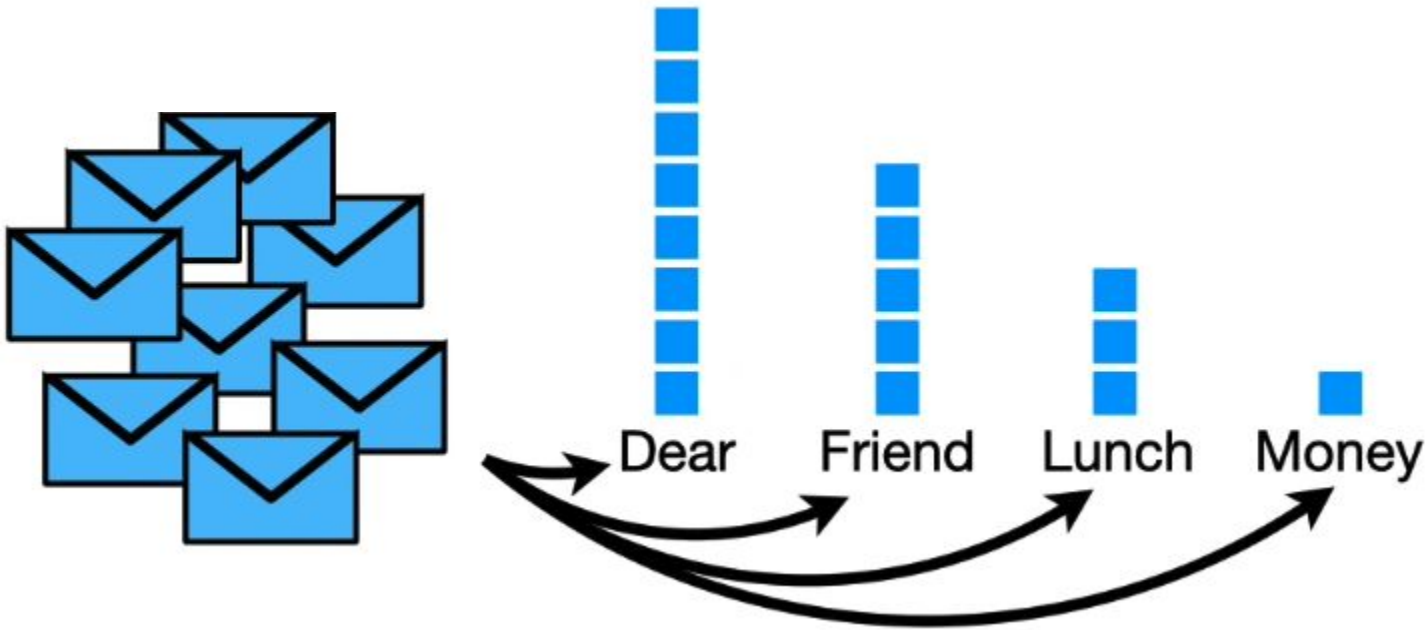


Naive Bayes

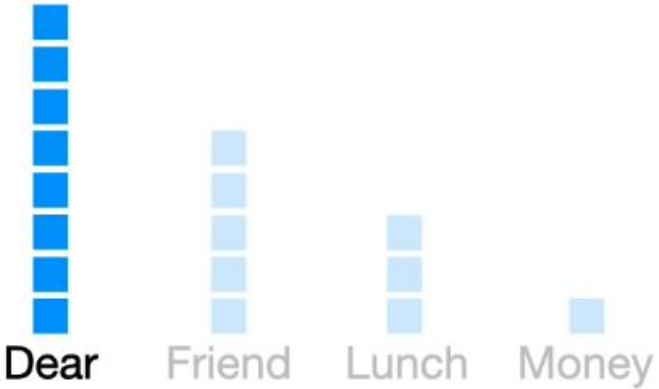




Naive Bayes

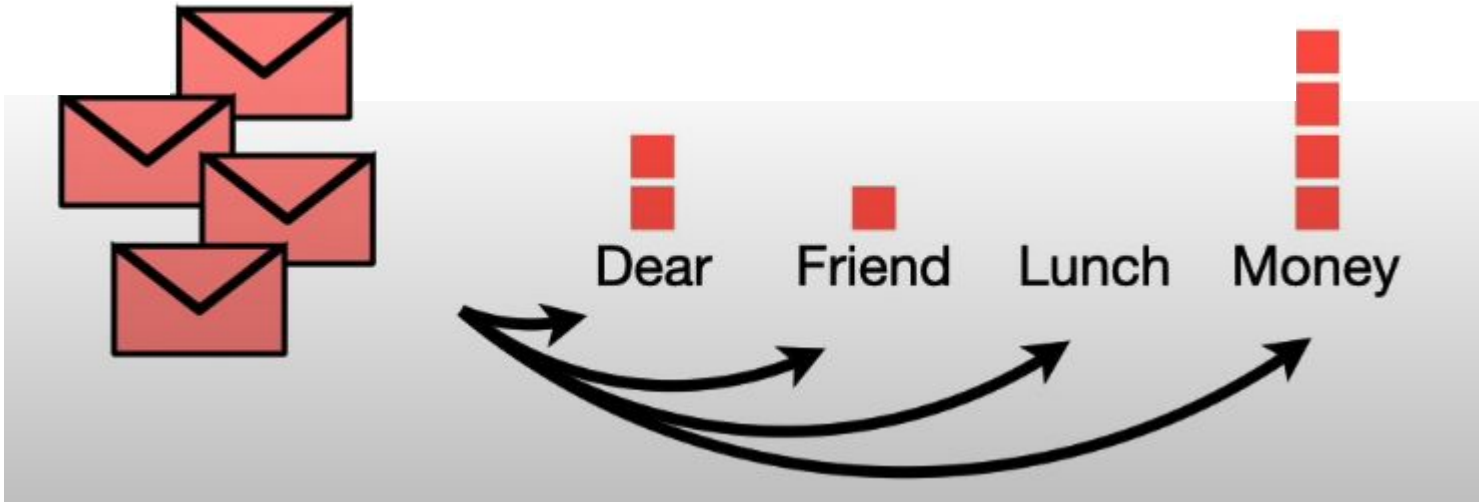


Naive Bayes

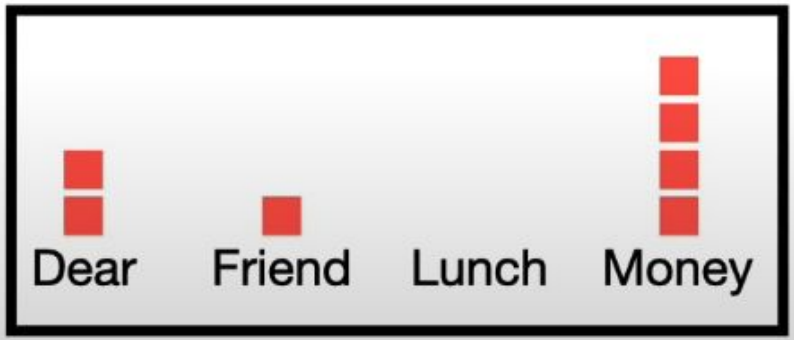


$$p(\textbf{Dear} \mid \textbf{Normal}) = \frac{8}{17} = 0.47$$

Naive Bayes



Naive Bayes



$$p(\text{Dear} \mid \text{Spam}) = \frac{2}{7} = 0.29$$

Naive Bayes

p(**Dear** | **N**) = 0.47
p(**Friend** | **N**) = 0.29
p(**Lunch** | **N**) = 0.18
p(**Money** | **N**) = 0.06

p(**N**) = 0.67

p(**S**) = 0.33

p(**Dear** | **S**) = 0.29
p(**Friend** | **S**) = 0.14
p(**Lunch** | **S**) = 0.00
p(**Money** | **S**) = 0.57

Naive Bayes

$$p(\mathbf{N}) \times p(\mathbf{Dear} \mid \mathbf{N}) \times p(\mathbf{Friend} \mid \mathbf{N})$$

$$0.67 \times 0.47 \times 0.29 = 0.09$$

$p(\mathbf{N}) = 0.67$

$p(\mathbf{Dear} \mid \mathbf{N}) = 0.47$

$p(\mathbf{Friend} \mid \mathbf{N}) = 0.29$

$p(\mathbf{Lunch} \mid \mathbf{N}) = 0.18$

$p(\mathbf{Money} \mid \mathbf{N}) = 0.06$

Naive Bayes

$$p(S) \times p(\text{Dear} \mid S) \times p(\text{Friend} \mid S)$$

$$0.33 \times 0.29 \times 0.14 = 0.01$$

$p(S) = 0.33$

$p(\text{Dear} \mid S) = 0.29$

$p(\text{Friend} \mid S) = 0.14$

$p(\text{Lunch} \mid S) = 0.00$

$p(\text{Money} \mid S) = 0.57$

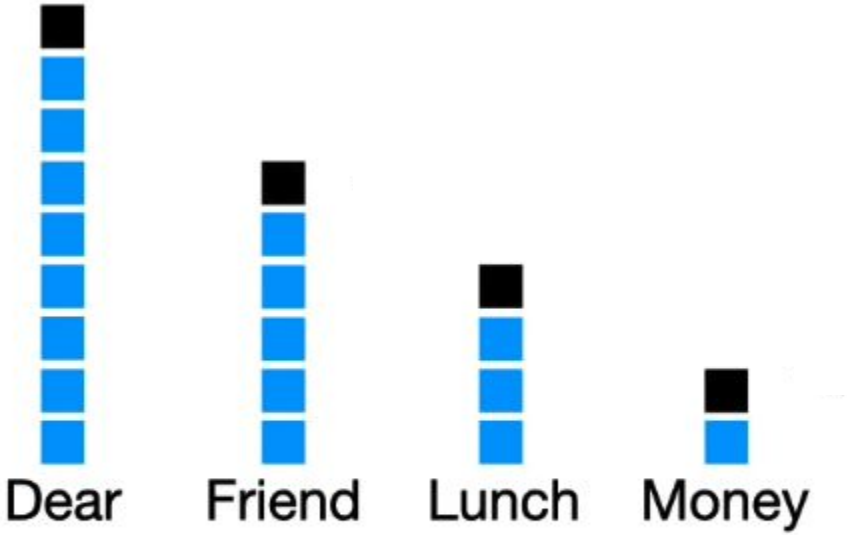
Lunch Money Money Money Money

$$p(\mathbf{N}) \times p(\mathbf{Lunch} \mid \mathbf{N}) \times p(\mathbf{Money} \mid \mathbf{N})^4 = 0.000002$$

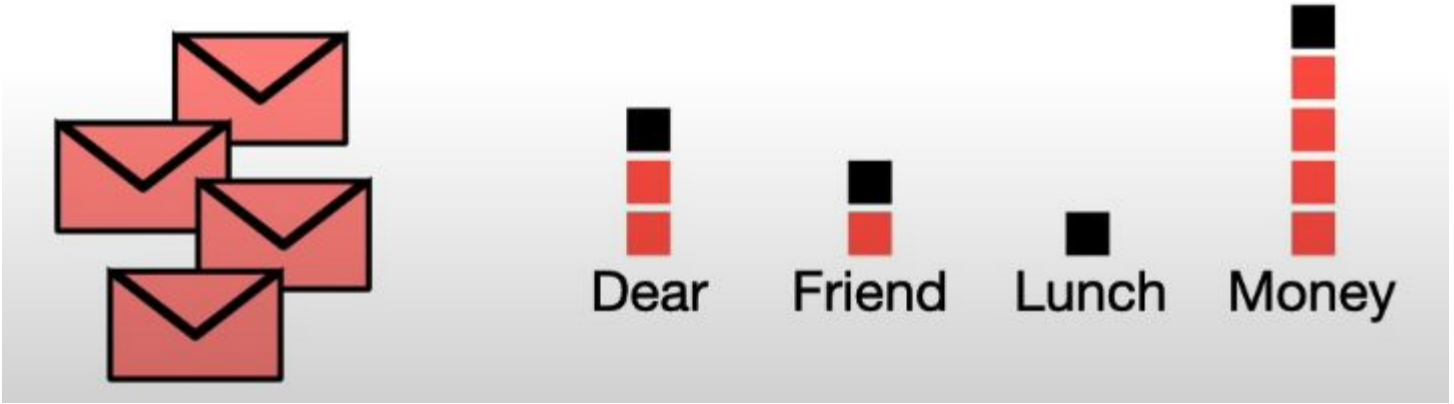
$$p(\mathbf{S}) \times p(\mathbf{Lunch} \mid \mathbf{S}) \times p(\mathbf{Money} \mid \mathbf{S})^4 = 0$$

Naive Bayes

p(N) = 0.67



Naive Bayes



Naive Bayes

Lunch Money Money Money Money

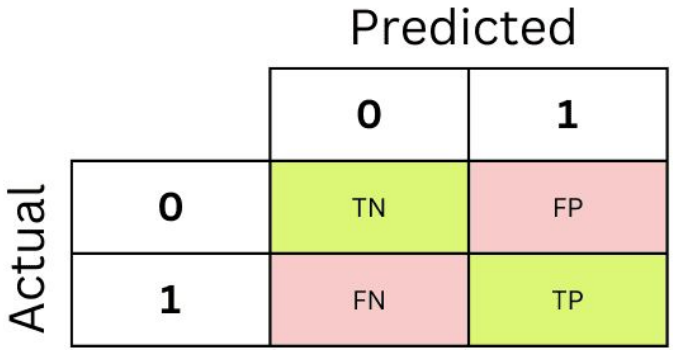
$$p(\mathbf{N}) \times p(\mathbf{Lunch} \mid \mathbf{N}) \times p(\mathbf{Money} \mid \mathbf{N})^4 = 0.00001$$

$$p(\mathbf{S}) \times p(\mathbf{Lunch} \mid \mathbf{S}) \times p(\mathbf{Money} \mid \mathbf{S})^4 = 0.00122$$

Naive Bayes

Confusion Matrix

Normal - N
Spam - P



Precision: Spam olarak işaretlenen verilerin ne kadarı gerçekten spam

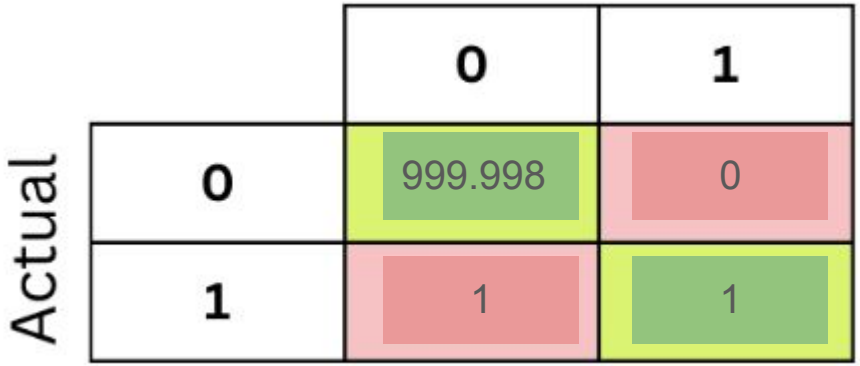
Recall: Gerçekte spam olan verilerin ne kadarı tespit edildi

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

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

Naive Bayes

Sağlık Sektörü - Maximum Precision



Gerçek Veri:
Akciğer Kanseri Sayısı: 2
Sağlıklı İnsan Sayısı: 999.999

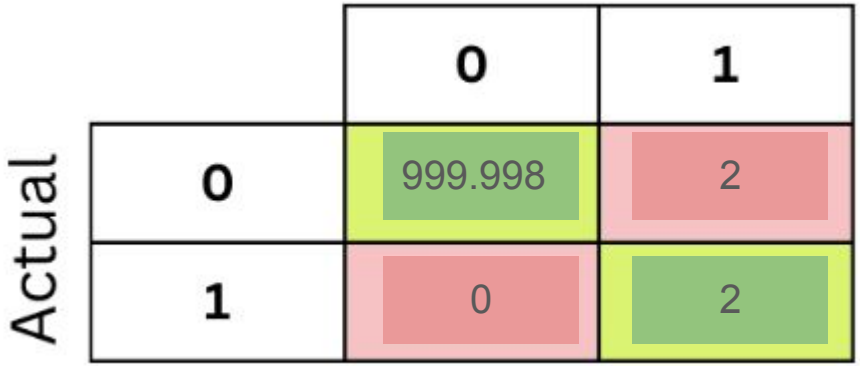
Precision = 1 / 1 + 0 = 1

Recall = 1 / 1 + 1 = 0.5

Tahmin:
Akciğer Kanseri Sayısı: 1
Sağlıklı İnsan Sayısı: 1.000.000

Naive Bayes

Sağlık Sektörü - Maximum Recall



Gerçek Veri:
Akciğer Kanseri Sayısı: 2
Sağlıklı İnsan Sayısı: 999.999

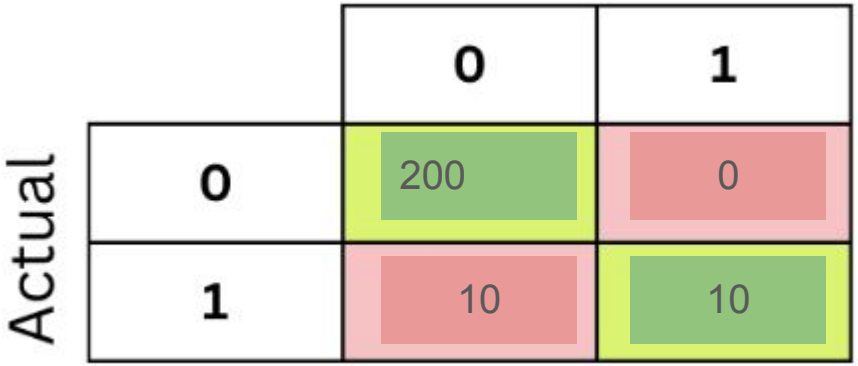
Precision = 2 / 2 + 2 = 0.5

Recall = 2 / 2 + 0 = 1

Tahmin:
Akciğer Kanseri Sayısı: 4
Sağlıklı İnsan Sayısı: 1.000.000

Naive Bayes

Spam Mail: Maximum Precision



Gerçek Veri:
Spam Mail Sayısı: 20
Normal Mail Sayısı: 200

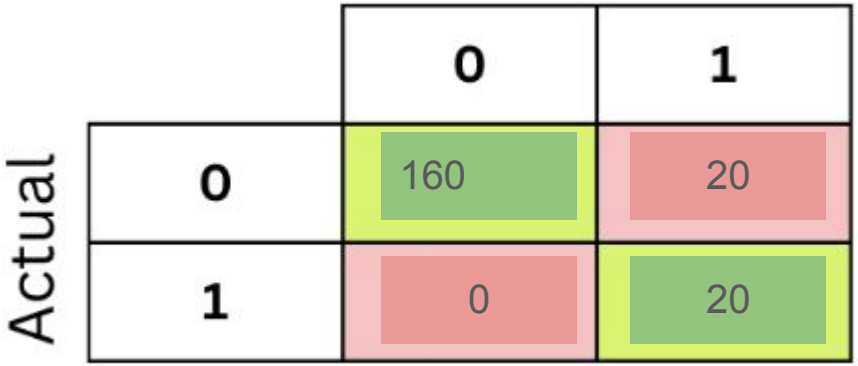
Precision = 10 / 10 + 0 = 1

Recall = 10 / 10 + 10 = 0.5

Tahmin:
Spam Mail Sayısı: 10
Normal Mail Sayısı: 210

Naive Bayes

Spam Mail: Maximum Recall



Gerçek Veri:
Spam Mail Sayısı: 20
Normal Mail Sayısı: 200

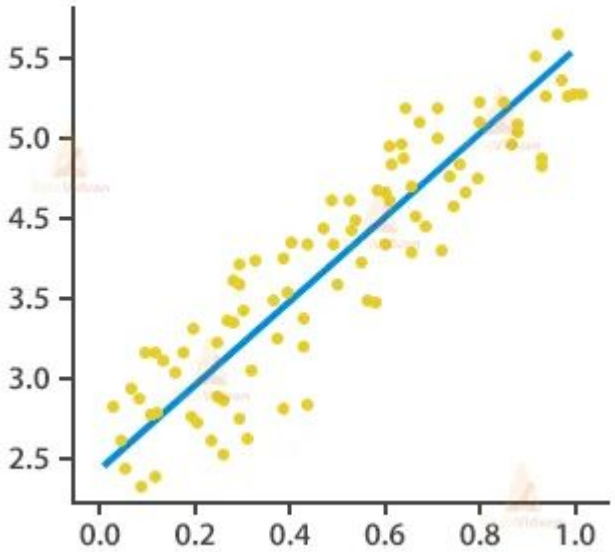
Precision = 20 / 20 + 20 = 0.5

Recall = 20 / 20 + 0 = 1

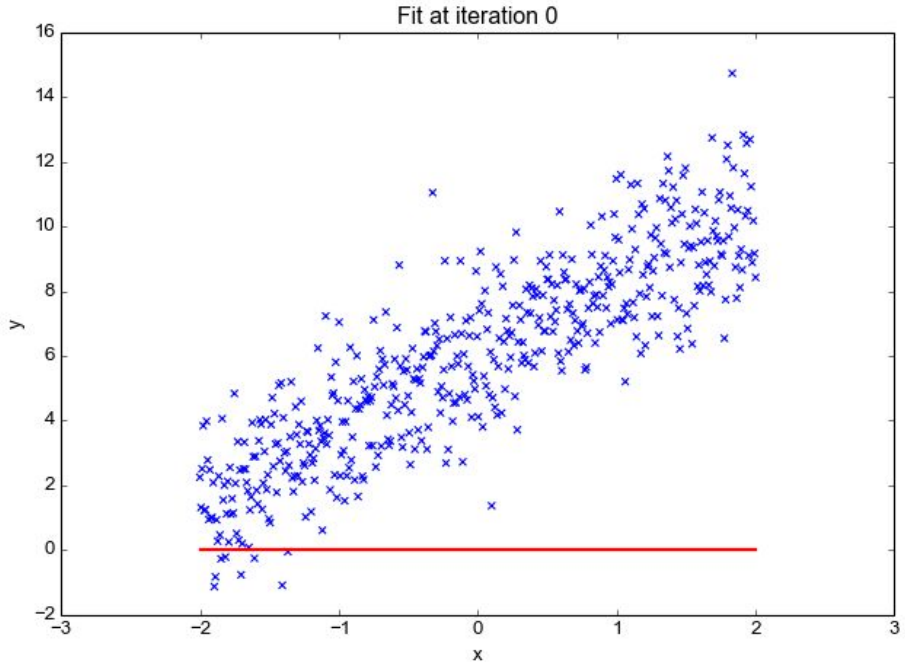
Tahmin:
Spam Mail Sayısı: 40
Normal Mail Sayısı: 160

Supervised Learning

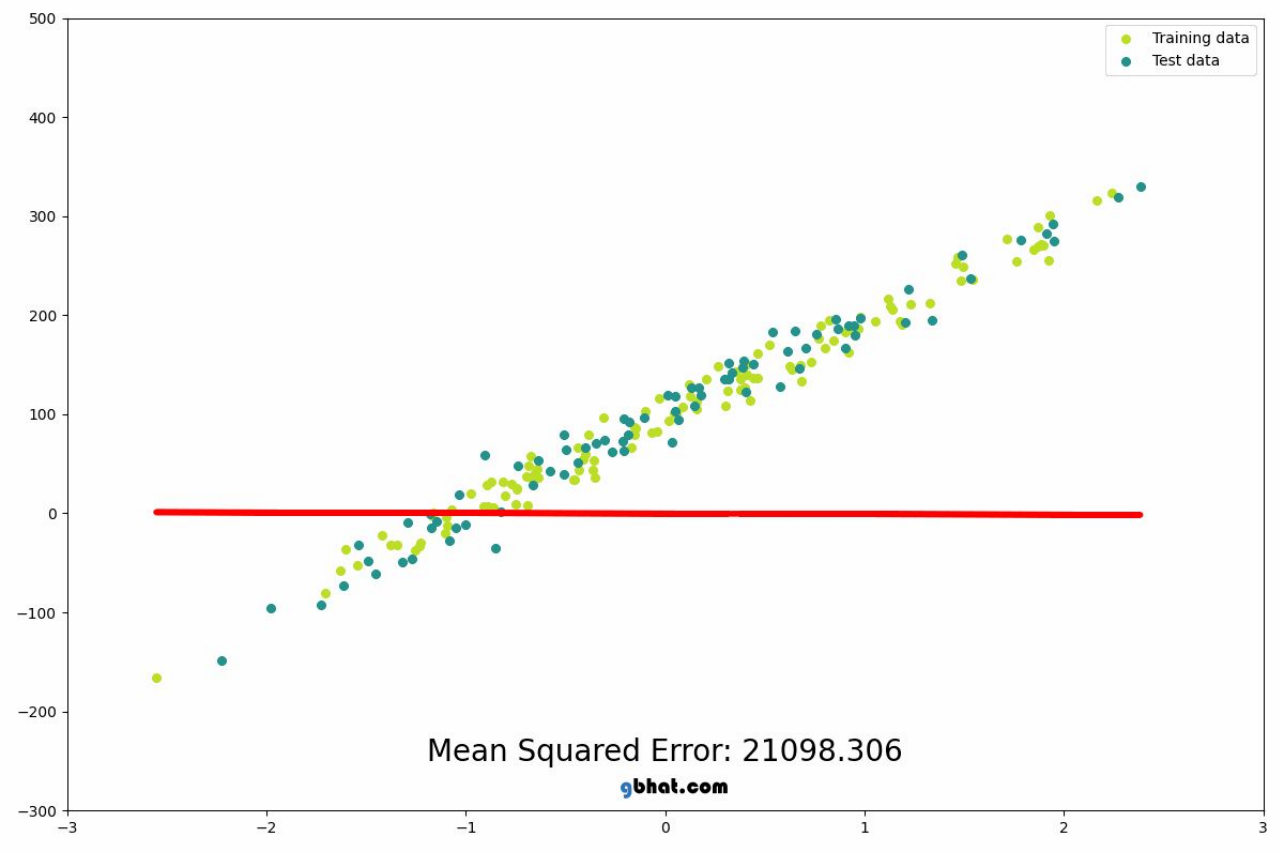
- Regression



Linear Regression

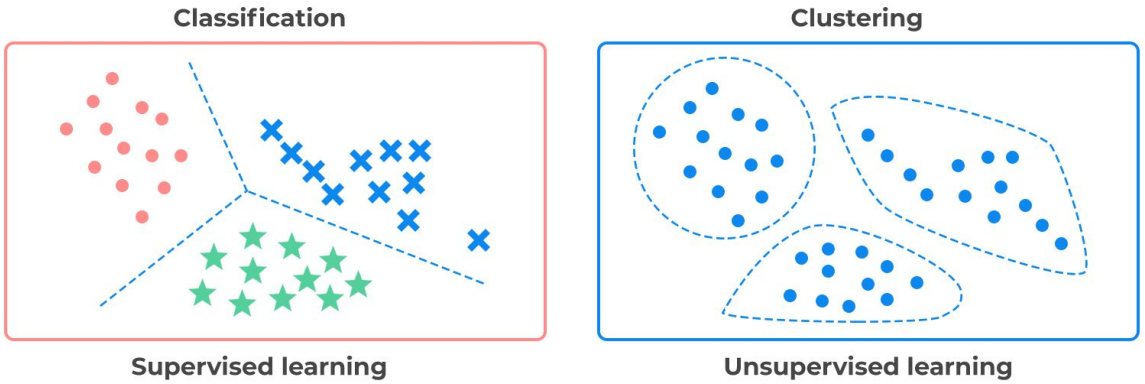


Linear Regression



Unsupervised Learning

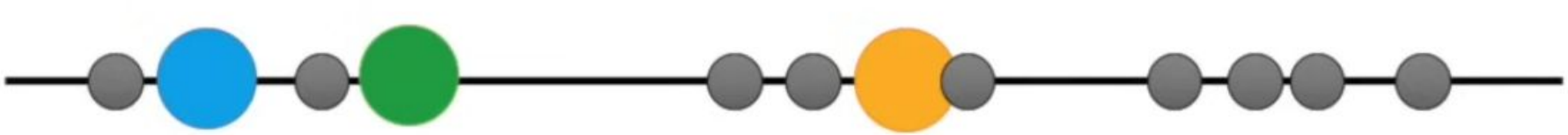
- Clustering



K-Means Clustering

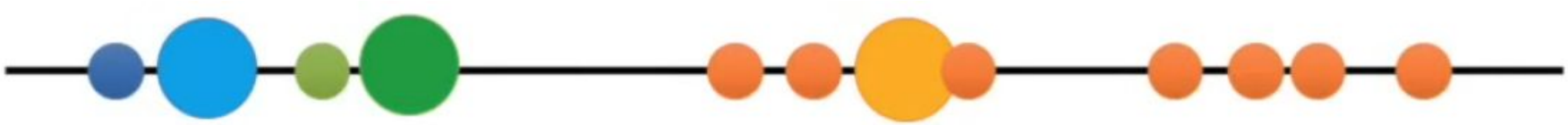
K-Means Clustering

1st attempt



K-Means Clustering

1st attempt

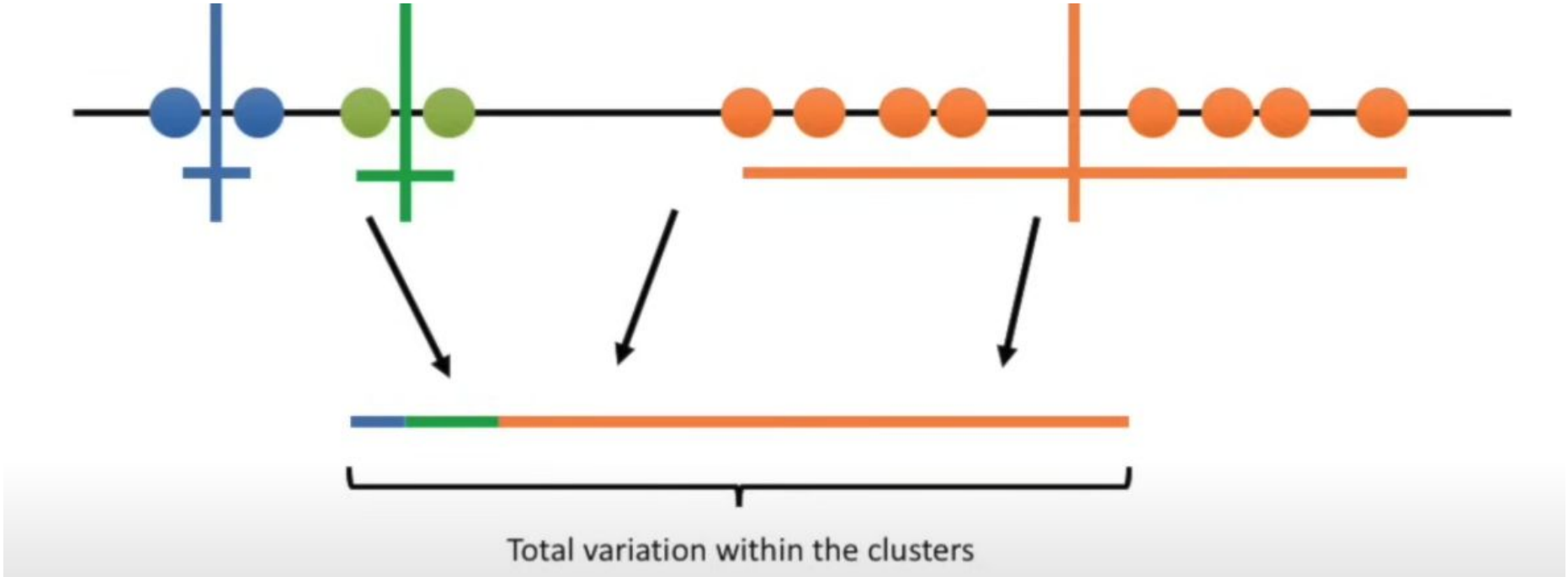


K-Means Clustering

1st attempt

K-Means Clustering

1st attempt

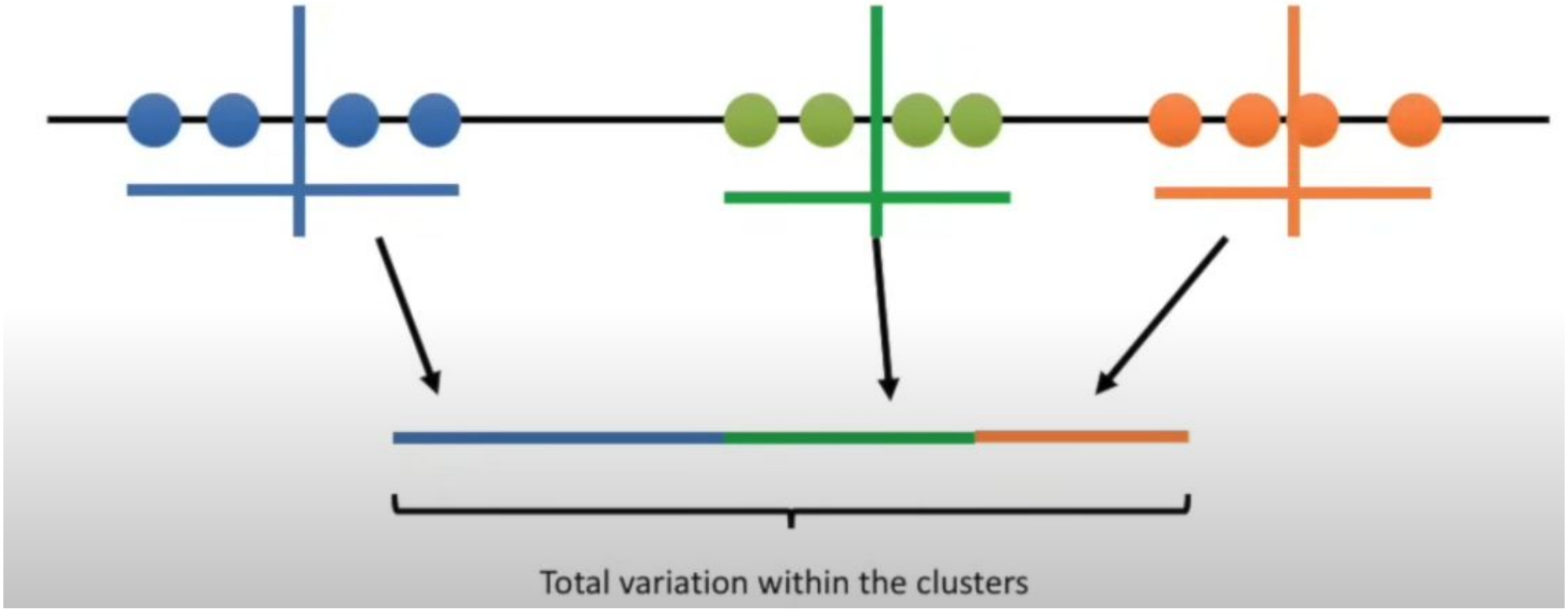


K-Means Clustering

2nd attempt

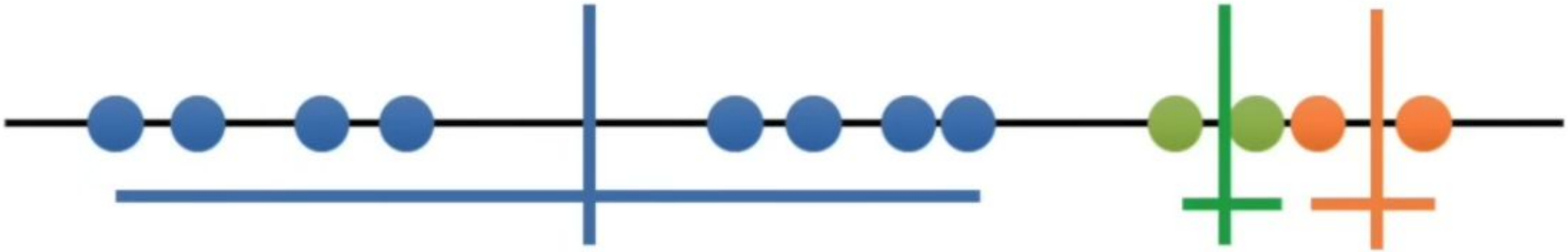
K-Means Clustering

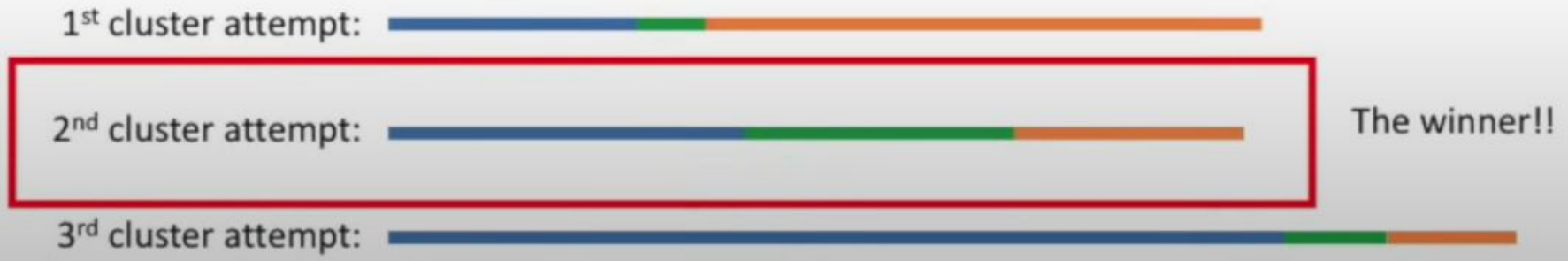
2nd attempt



K-Means Clustering

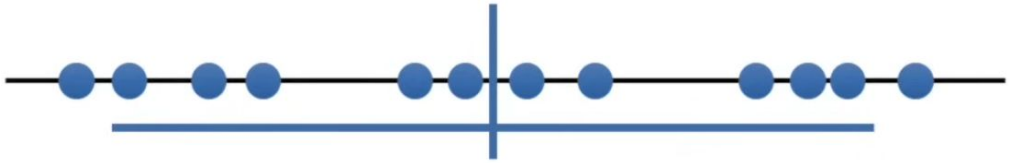
3rd attempt



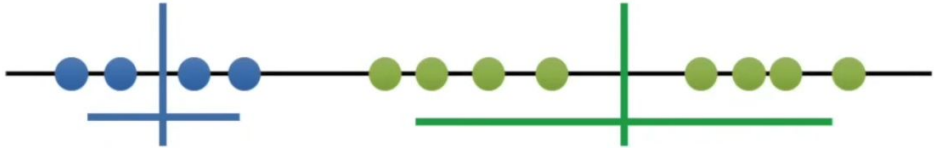


K-Means Clustering

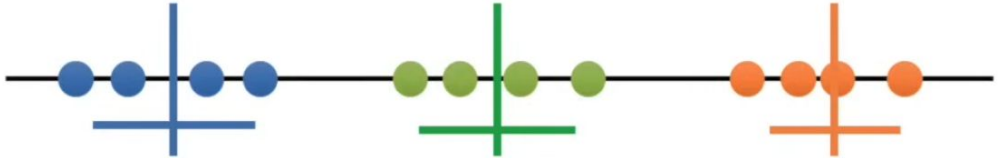
Deciding on K



K = 1



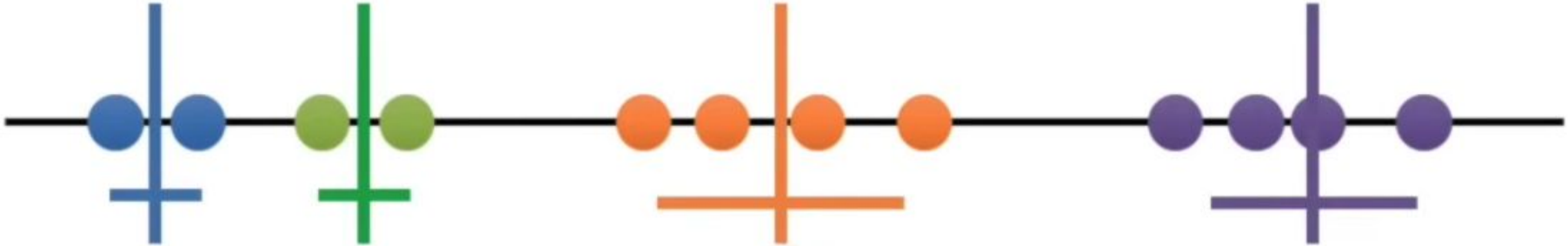
K = 2



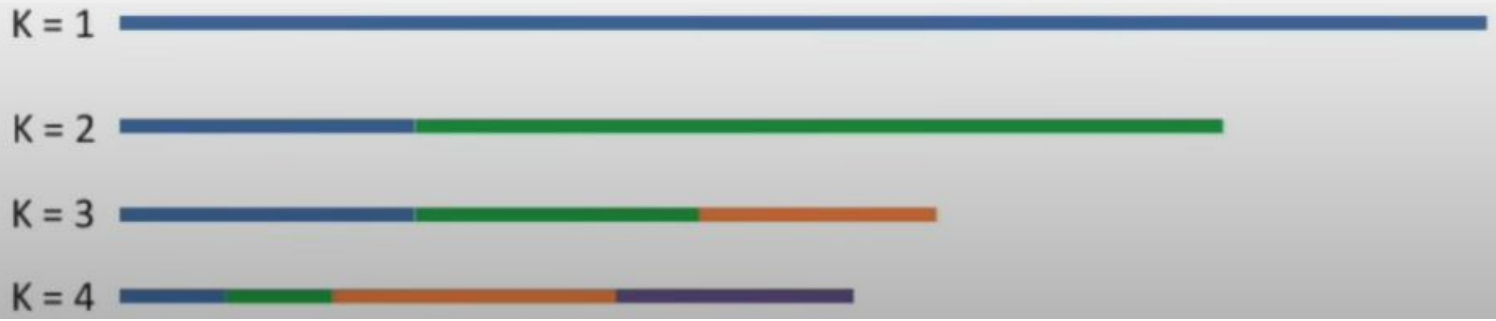
K = 3

K-Means Clustering

Deciding on K

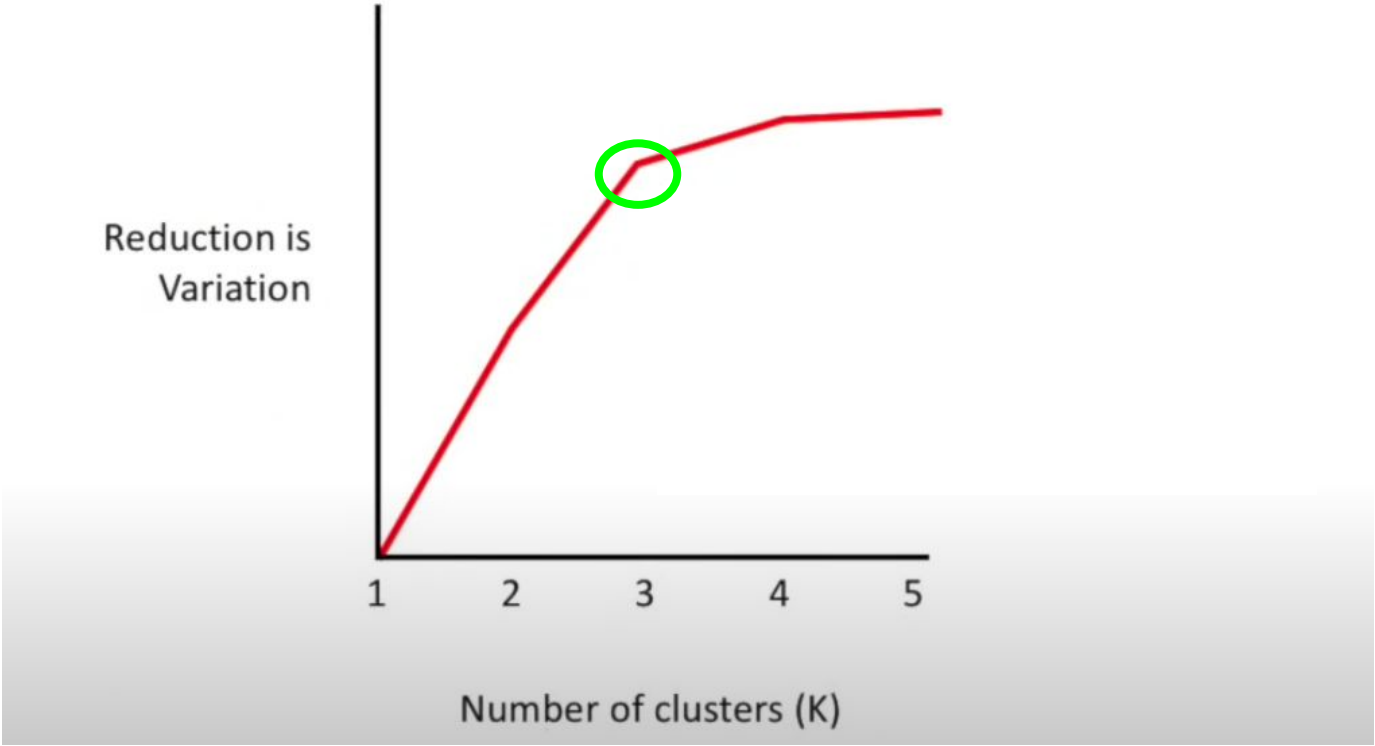


K = 4



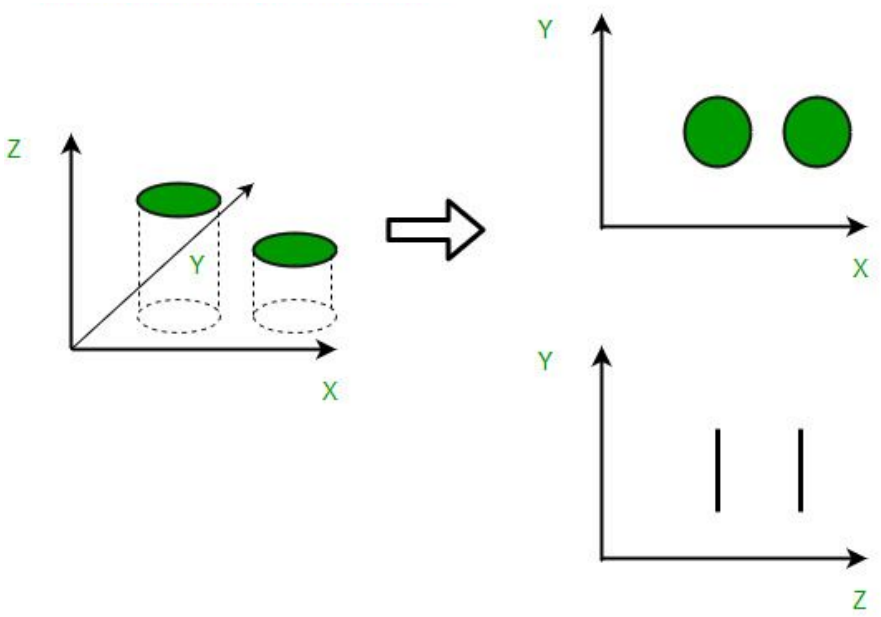
K-Means Clustering

Deciding on K

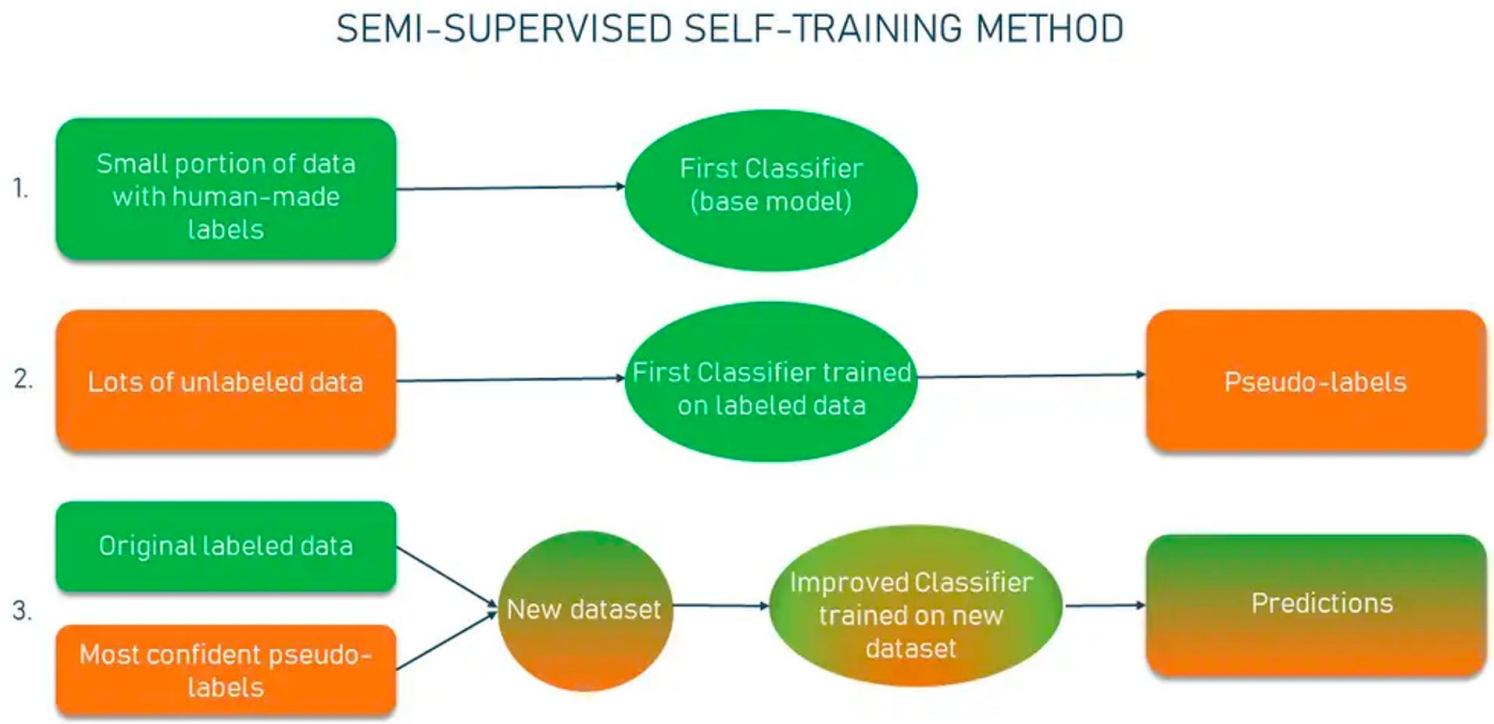


Unsupervised Learning

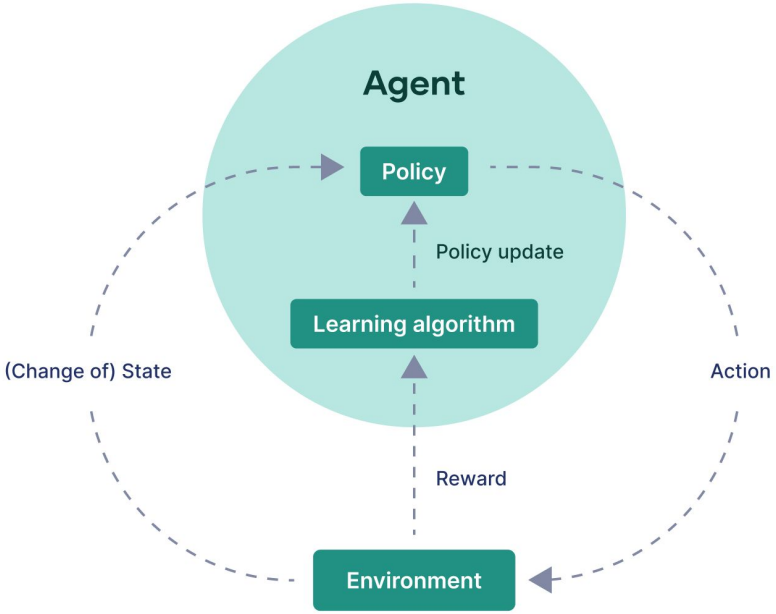
- Dimension Reduction



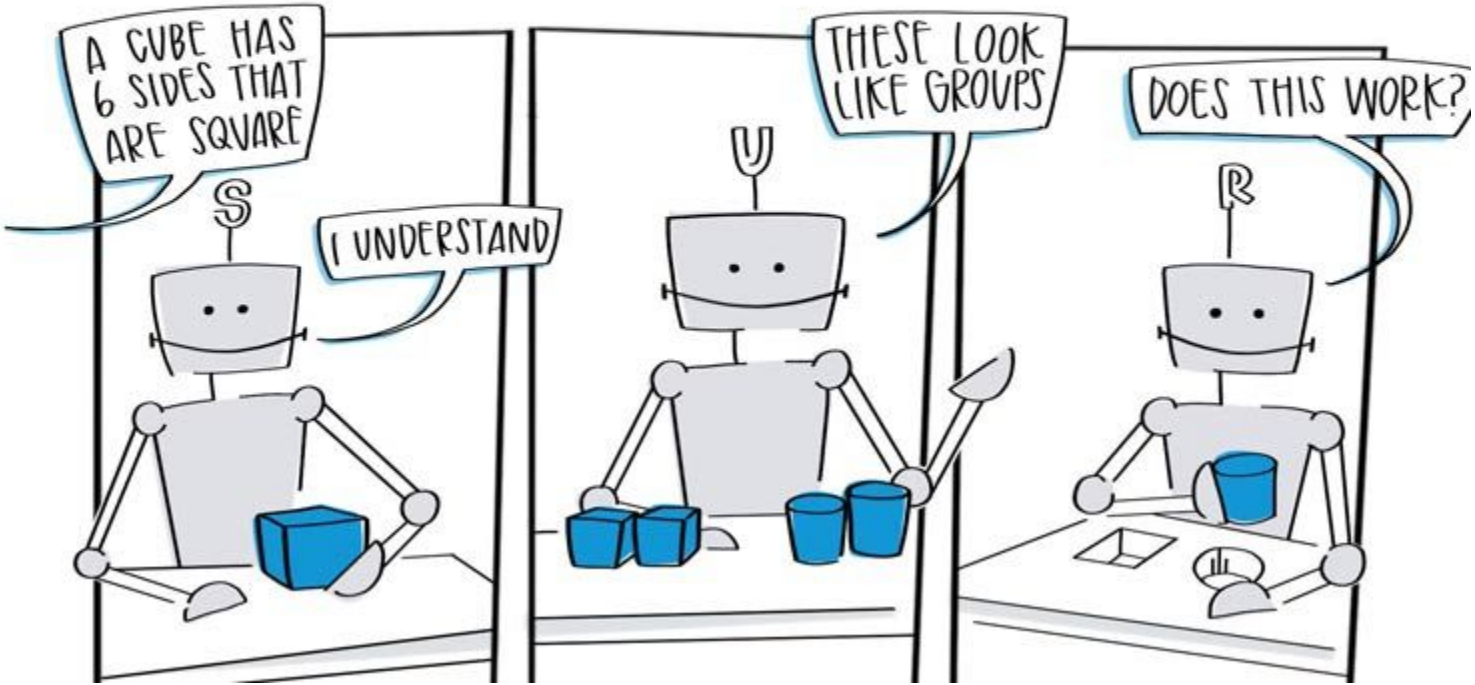
Semi-Supervised Learning



Reinforcement Learning



MACHINE LEARNING



Three classes of learning problems

Supervised Learning

Data: (x, y)
x is an input data, y is a label (e.g. photo with label "cat")

Goal: Learn to map input to output i.e. $x \rightarrow y$

An example: to classify

This is a cat

Unsupervised Learning

Data: x
x is data, there's no labels!

Goal: Learn an underlying structure of the data.

An example: Comparison

The two things are alike

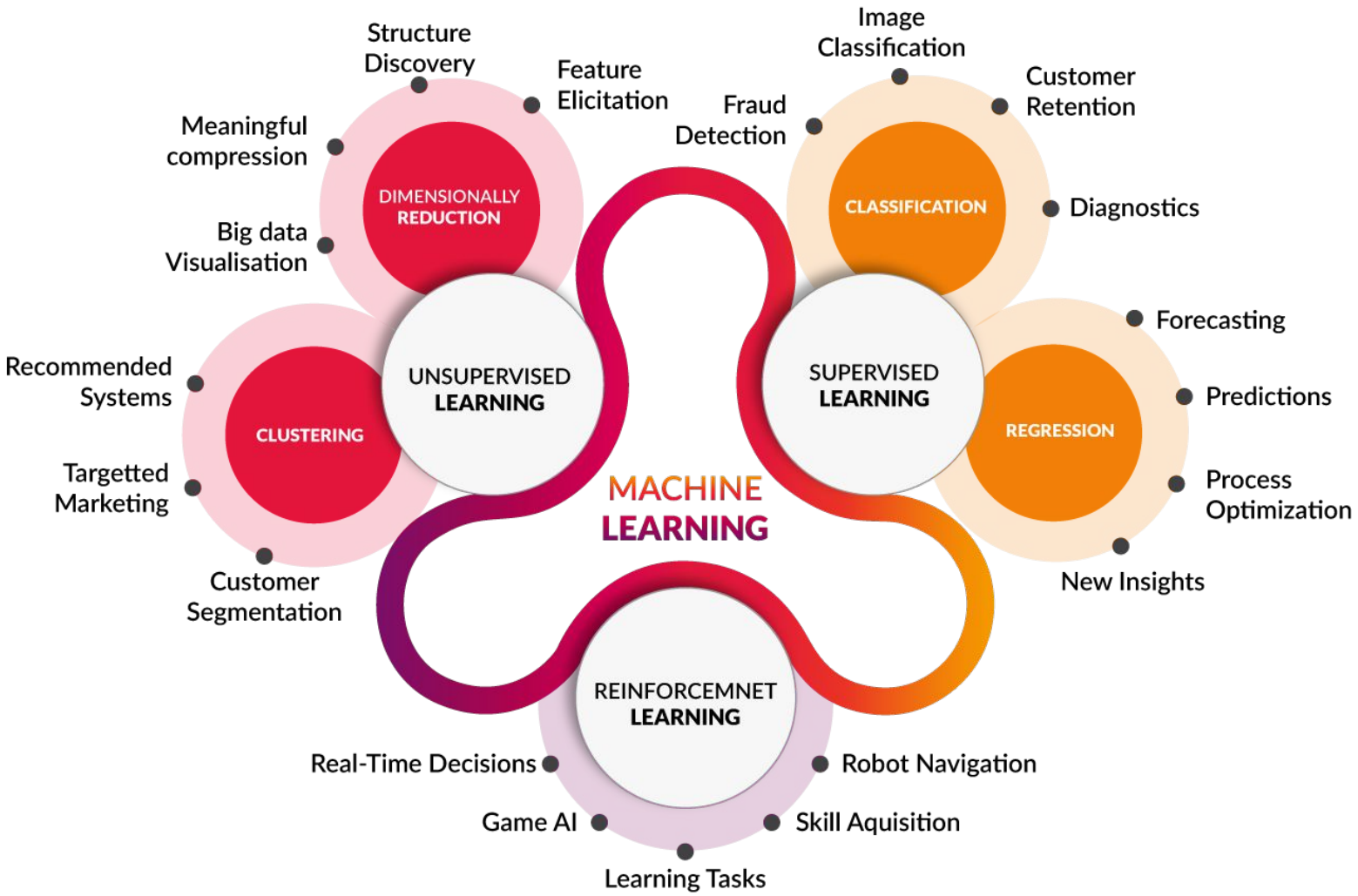
Reinforcement Learning

Data: No data, Only state-action pairs (s, a).

Goal: Maximize future reward over many time steps.

An example: reward = joy

Interaction with the cat gives joy


MACHINE LEARNING
UNSUPERVISED LEARNING
SUPERVISED LEARNING
REINFORCEMNET LEARNING
DIMENSIONALLY REDUCTION
CLUSTERING
CLASSIFICATION
REGRESSION
Structure Discovery
Feature Elicitation
Meaningful compression
Big data Visualisation
Image Classification
Customer Retention
Fraud Detection
Diagnostics
Recommended Systems
Targetted Marketing
Customer Segmentation
Forecasting
Predictions
Process Optimization
New Insights
Real-Time Decisions
Robot Navigation
Game AI
Skill Aquisition
Learning Tasks

Deep Learning