Data Science Workshop

Practical . Python

By: Ali Hamdi

Introduction to Data Science

01 Life cycle

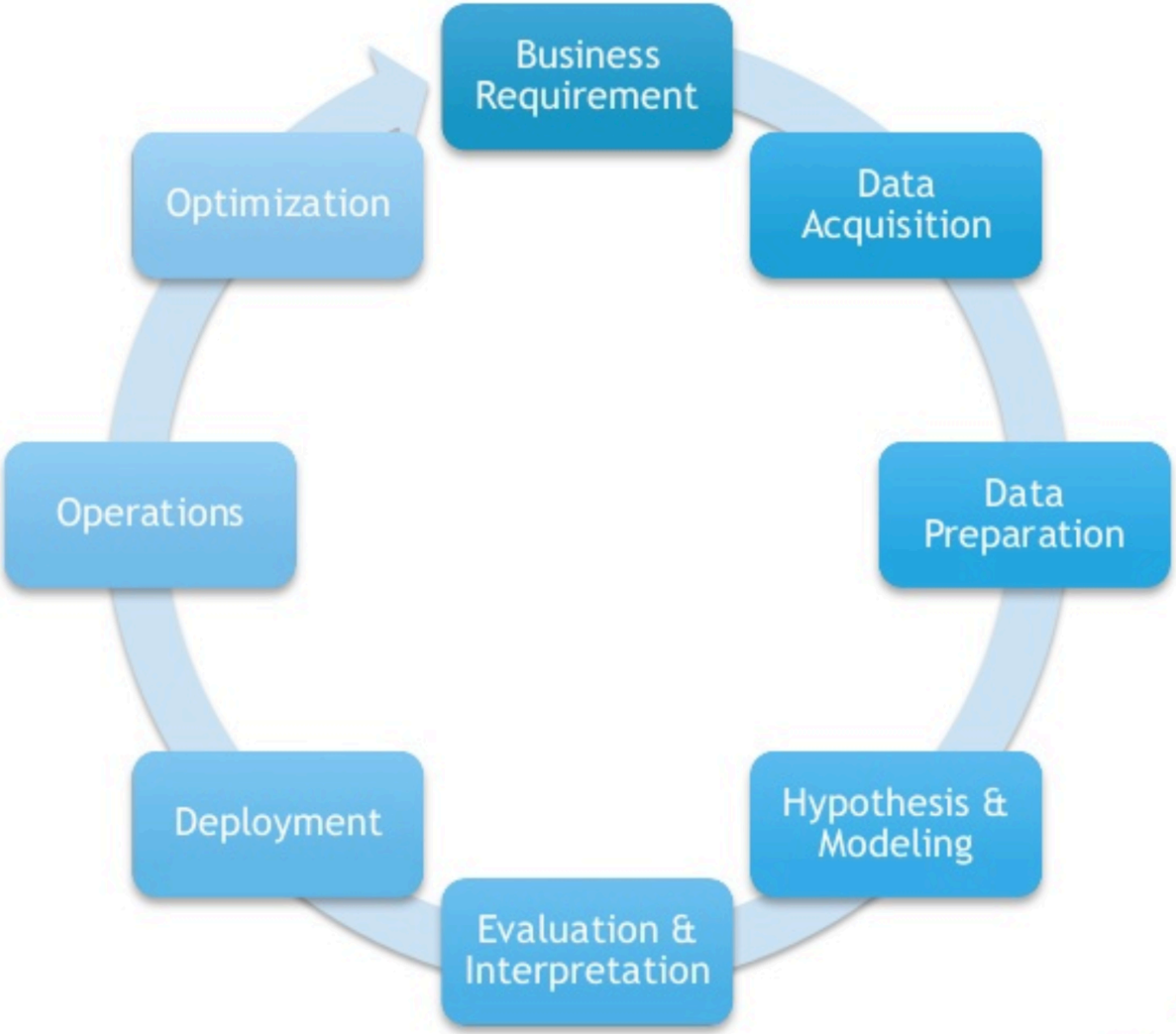
02 Data science team

03 Applications

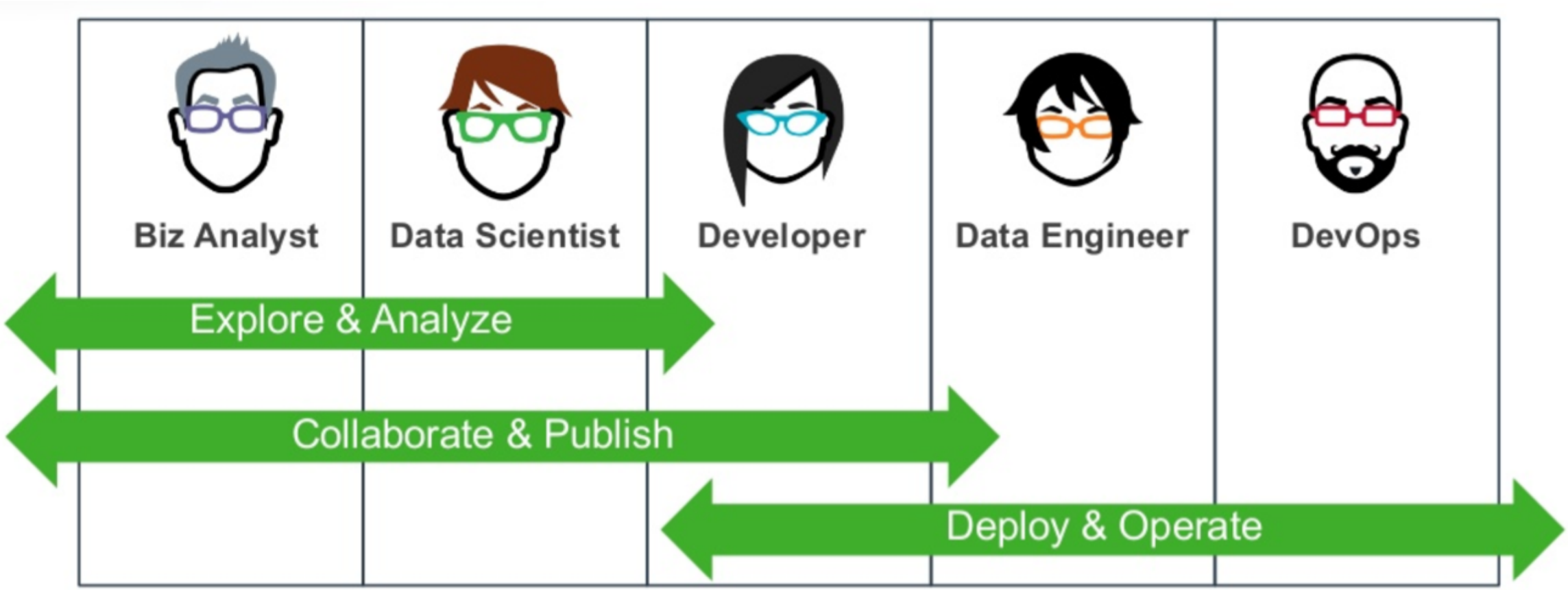
04 Research directions

05 Jobs

Data Science Project Life Cycle



Data science team



Data Science Application

- Internet Search
- Digital Advertisements
- Recommender Systems
- Image Recognition
- Speech Recognition
- Gaming
- Airline Route Planning
- Fraud and Risk Detection
- Self Driving Cars
- Robots

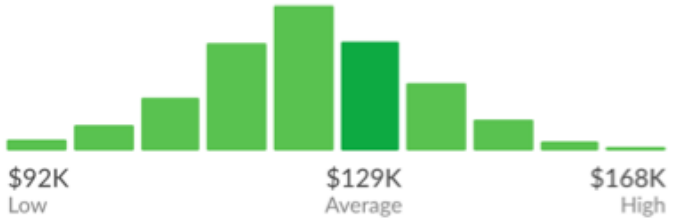
Research directions

- Knowledge Discovery & Data Mining.
- Machine Learning & Deep Learning.
- Visual Computing & Multimedia Analytics.
- Spatial & Context-Aware Data Management.
- Recommender Systems & Preference Analytics.
- Natural Language Processing & Text Mining.

Data Scientist Salary

Average Base Pay

$128,549 /yr



Data Scientist	Facebook	65 salaries	$135,117/yr ($100K – $180K)
Data Scientist	Microsoft	44 salaries	$123,556/yr ($94K – $152K)
Data Scientist	IBM	40 salaries	$109,177/yr ($81K – $144K)
Data Scientist	Booz Allen Hamilton	38 salaries	$84,450/yr ($58K – $140K)
Data Scientist	Capital One	28 salaries	$106,750/yr ($79K – $131K)
Data Scientist	Nielsen	23 salaries	$73,725/yr ($63K – $85K)
Data Scientist	Airbnb	21 salaries	$126,287/yr ($93K – $170K)
Data Scientist	Twitter	19 salaries	$135,360/yr ($120K – $155K)
Data Scientist	Uber	18 salaries	$123,686/yr ($103K – $152K)
Data Scientist	KPMG	16 salaries	$96,922/yr ($80K – $122K)
Data Scientist	Civis Analytics	16 salaries	$76,284/yr ($61K – $95K)
Data Scientist	LinkedIn	16 salaries	$132,059/yr ($111K – $261K)
Data Scientist	Rang Technologies	15 salaries	$104,480/yr ($92K – $133K)
Data Scientist	Apple	14 salaries	$144,833/yr ($112K – $180K)

glassdoor.com

Python Programming

01

Environment Setup

02

Python Basics

Data Exploratory and Analysis

Manipulation

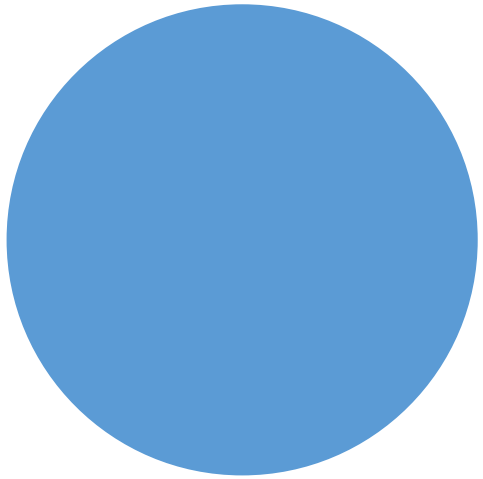
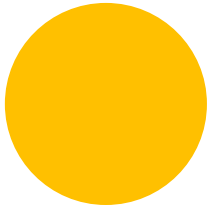
- Numpy
- Pandas

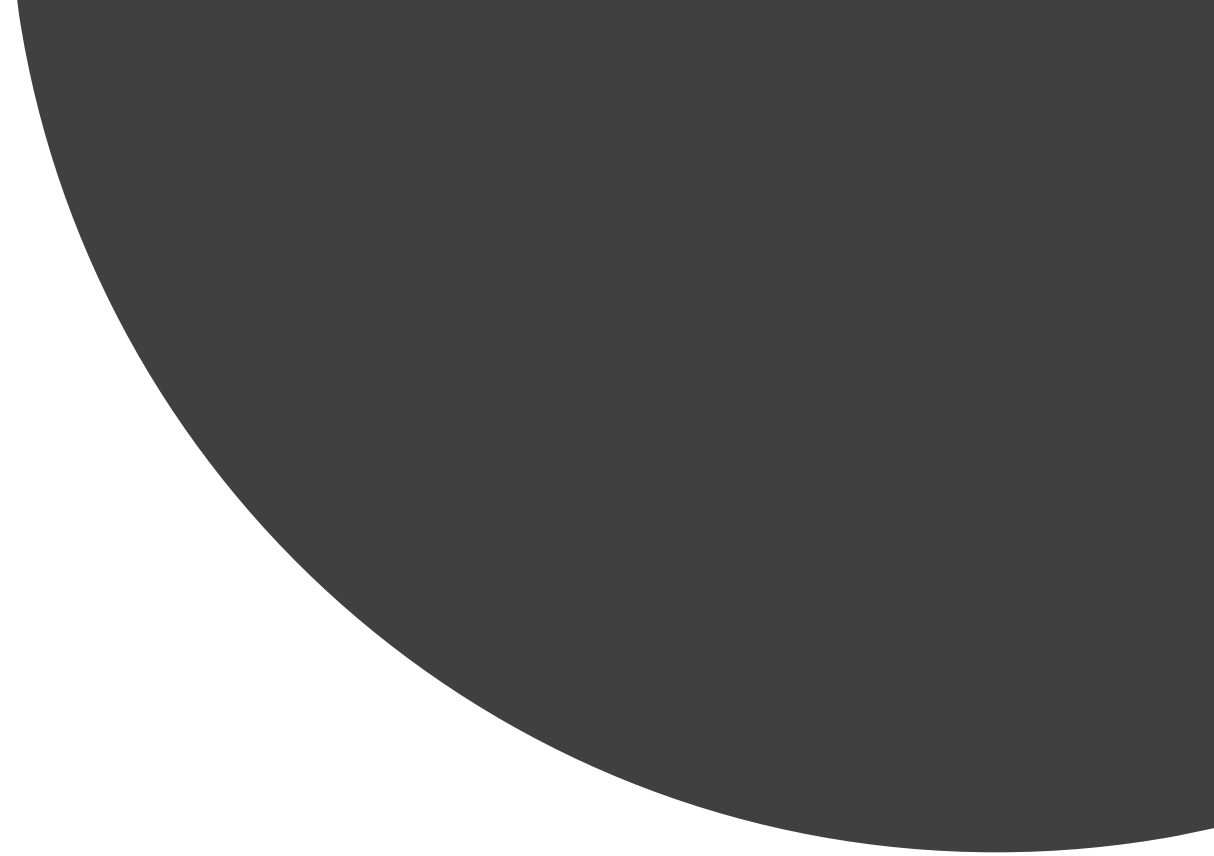
Visualization

- Matplotlip

Retrieval

- BeautifulSoup



Break

Machine Learning

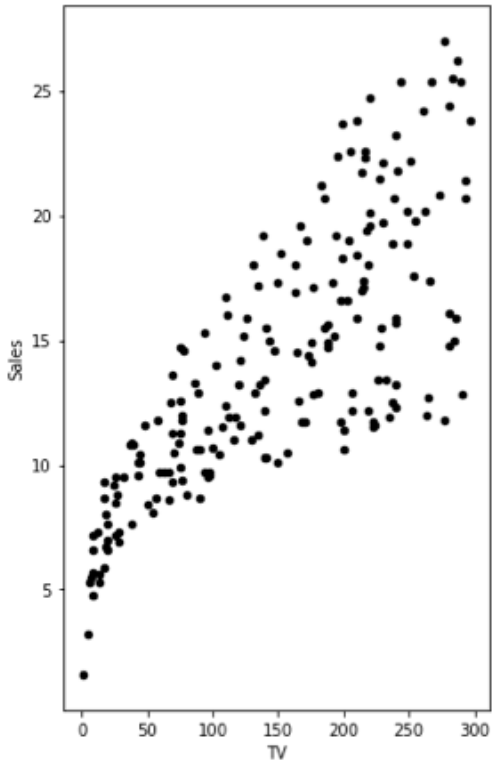
- Regression
- Classification
- Clustering
- Association Rules

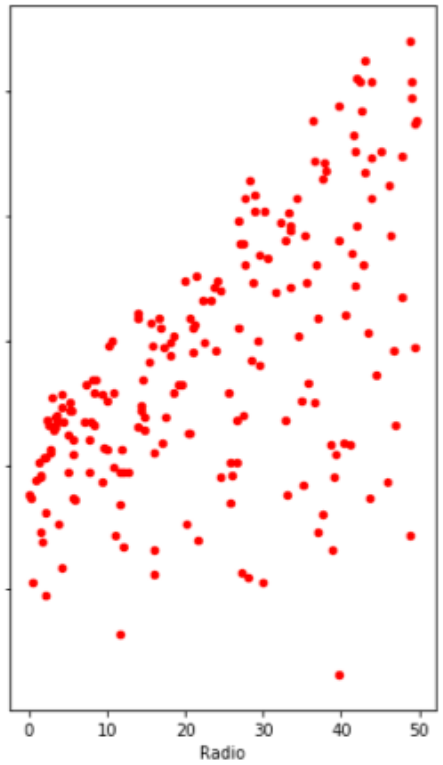
Supervised & Unsupervised Learning

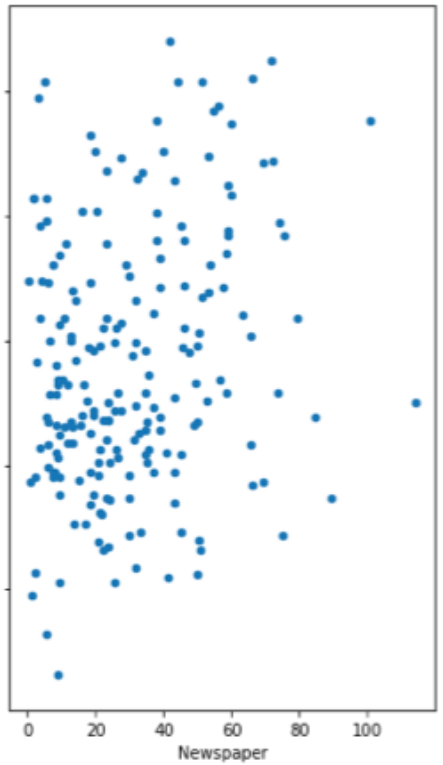
	Continuous	Categorical
Supervised	Regression	Classification
Unsupervised	Dimensionality Reduction	Clustering

Regression

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2





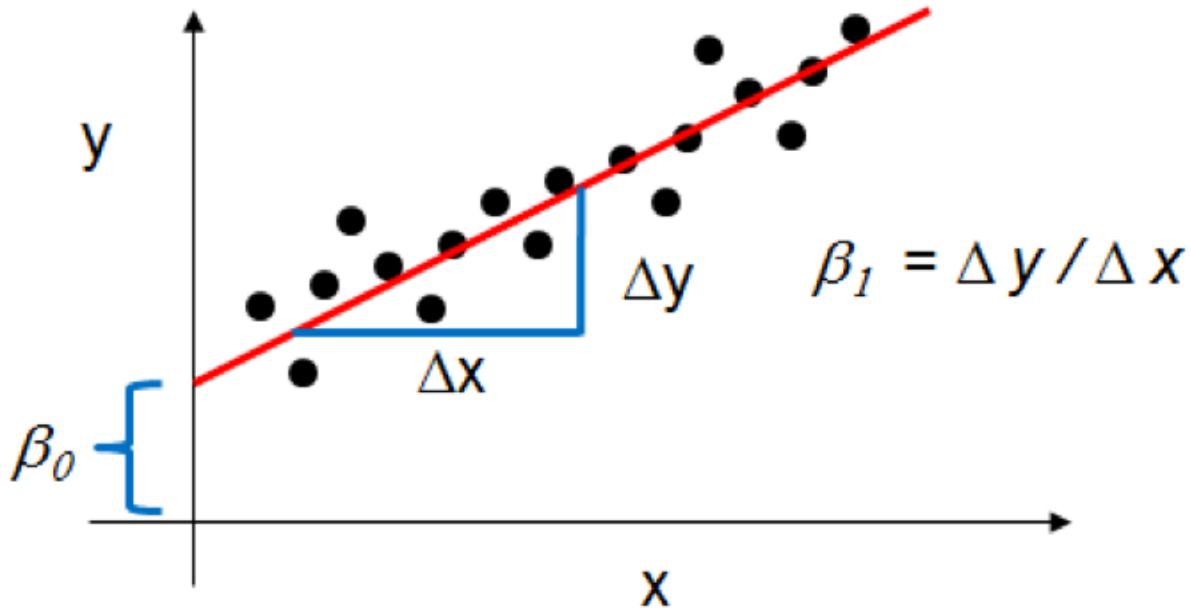


Simple Linear Regression

- To predict a **quantitative response** using a **single feature**.

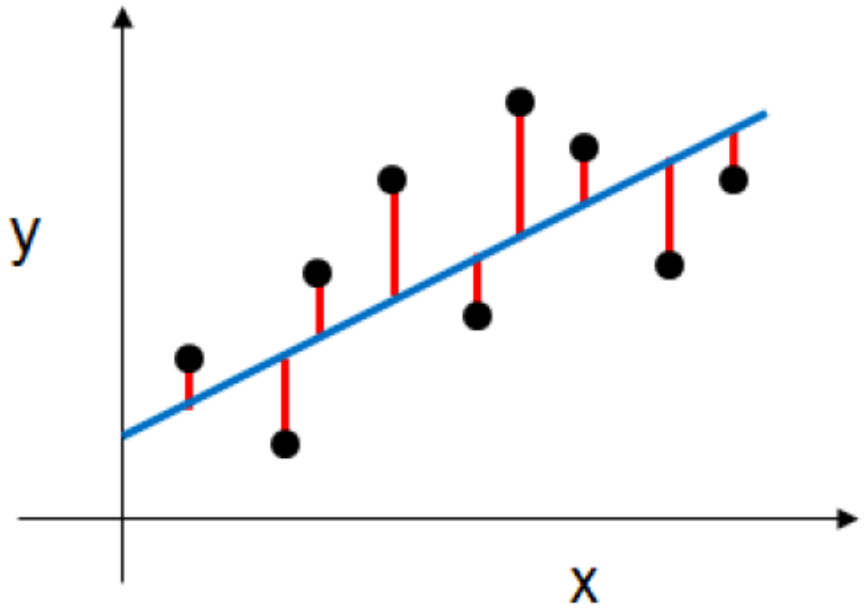
$$y=\beta 0+\beta 1x$$

- y is the response (Dependent Variable)
- x is the feature (Independent variable)
- β0 is the **intercept** (the value of y when x=0)
- β1 is the **slope** (the change in y divided by change in x)



Estimating Model Coefficients

- Coefficients are estimated using the **least squares criterion**
  - Minimizes the **sum of squared residuals** or "sum of squared errors"



Model Prediction

$$SS_{residuals} = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2$$

Observed Result

Multiple Linear Regression

- Using multiple features:

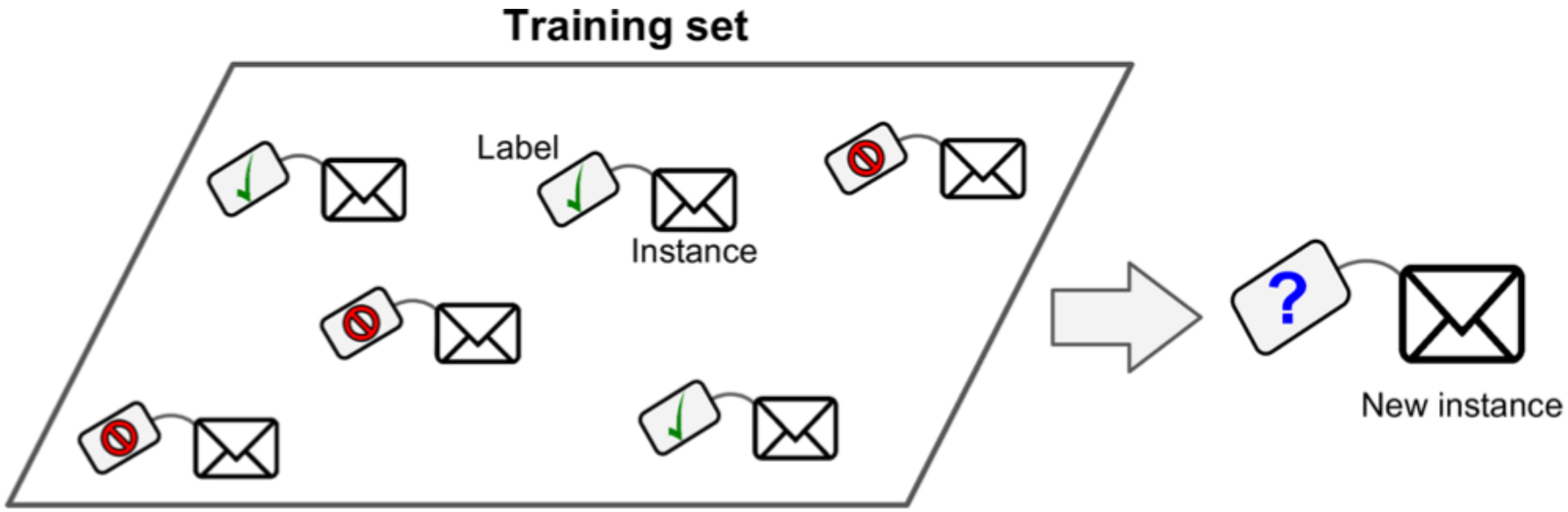
$$y = \beta 0 + \beta 1x1 + \ldots + \beta nxn$$

- x represents different feature and β is the feature coefficient :

$$y = \beta 0 + \beta 1 \times TV + \beta 2 \times Radio + \beta 3 \times Newspaper$$

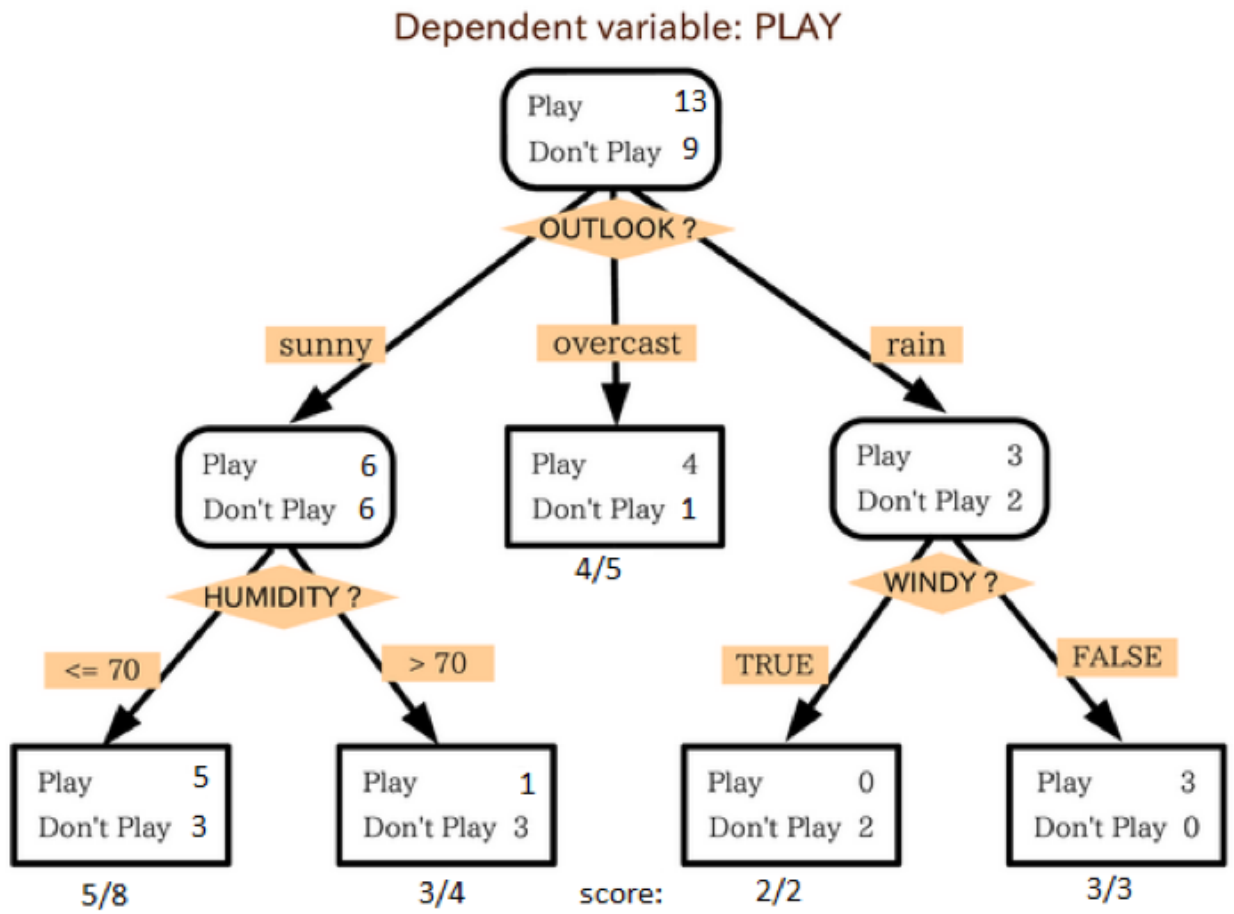
Machine Learning / Statistical classification

To identify a new observation belongs to which category, based on training data containing observations whose category membership is known.

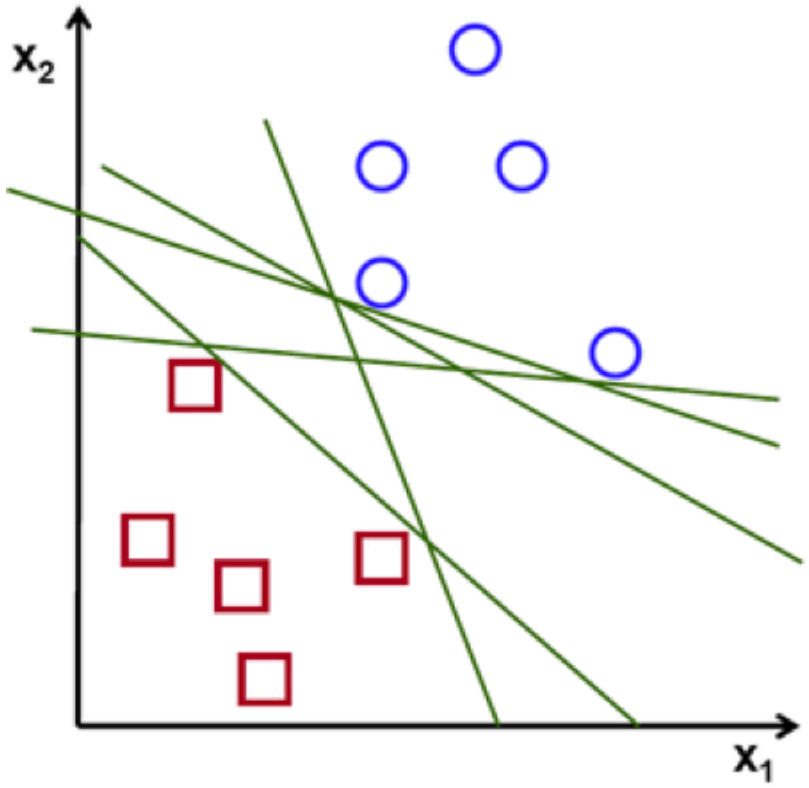


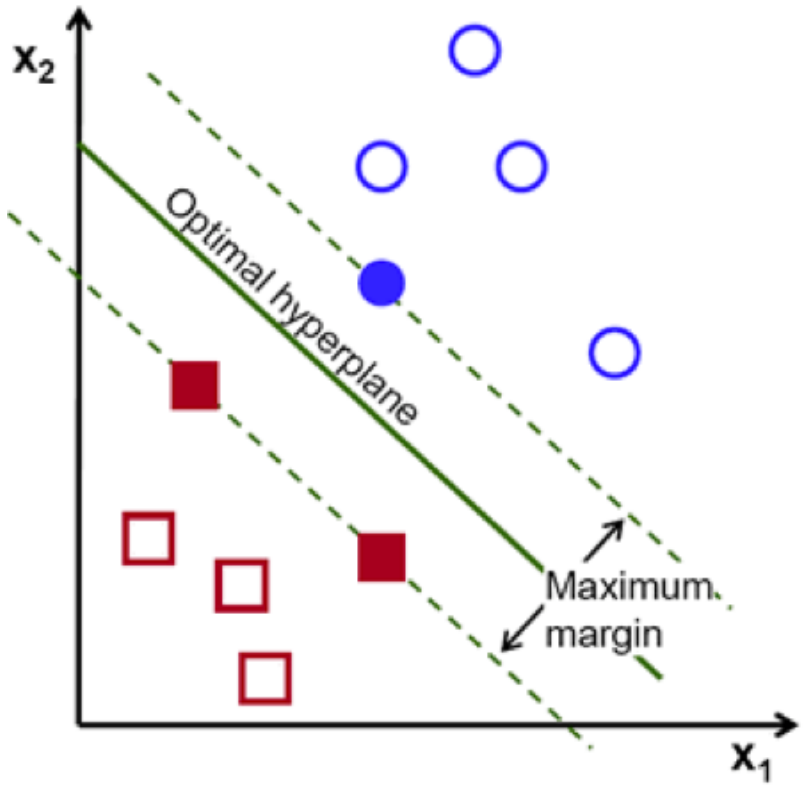
Decision Tree

Split the data into two or more homogeneous sets, based on most significant attributes to make as distinct groups as possible.



Support Vector Machine (SVM)





Naive Bayes

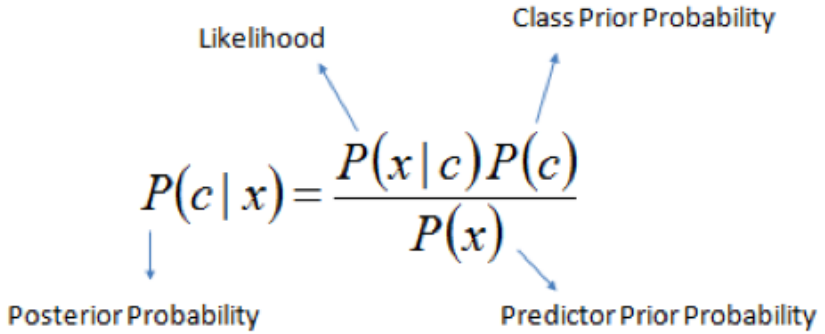
Will players will pay if weather is sunny?
P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)
P (Sunny |Yes) = 3/9 = 0.33,
P(Sunny) = 5/14 = 0.36,
P(Yes)= 9/14 = 0.64
P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

14 data instances

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

$P(c|x)$ is the posterior probability of *class* (c) given *predictor* (x).
$P(c)$ is the prior probability of *class*.
$P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
$P(x)$ is the prior probability of *predictor*.

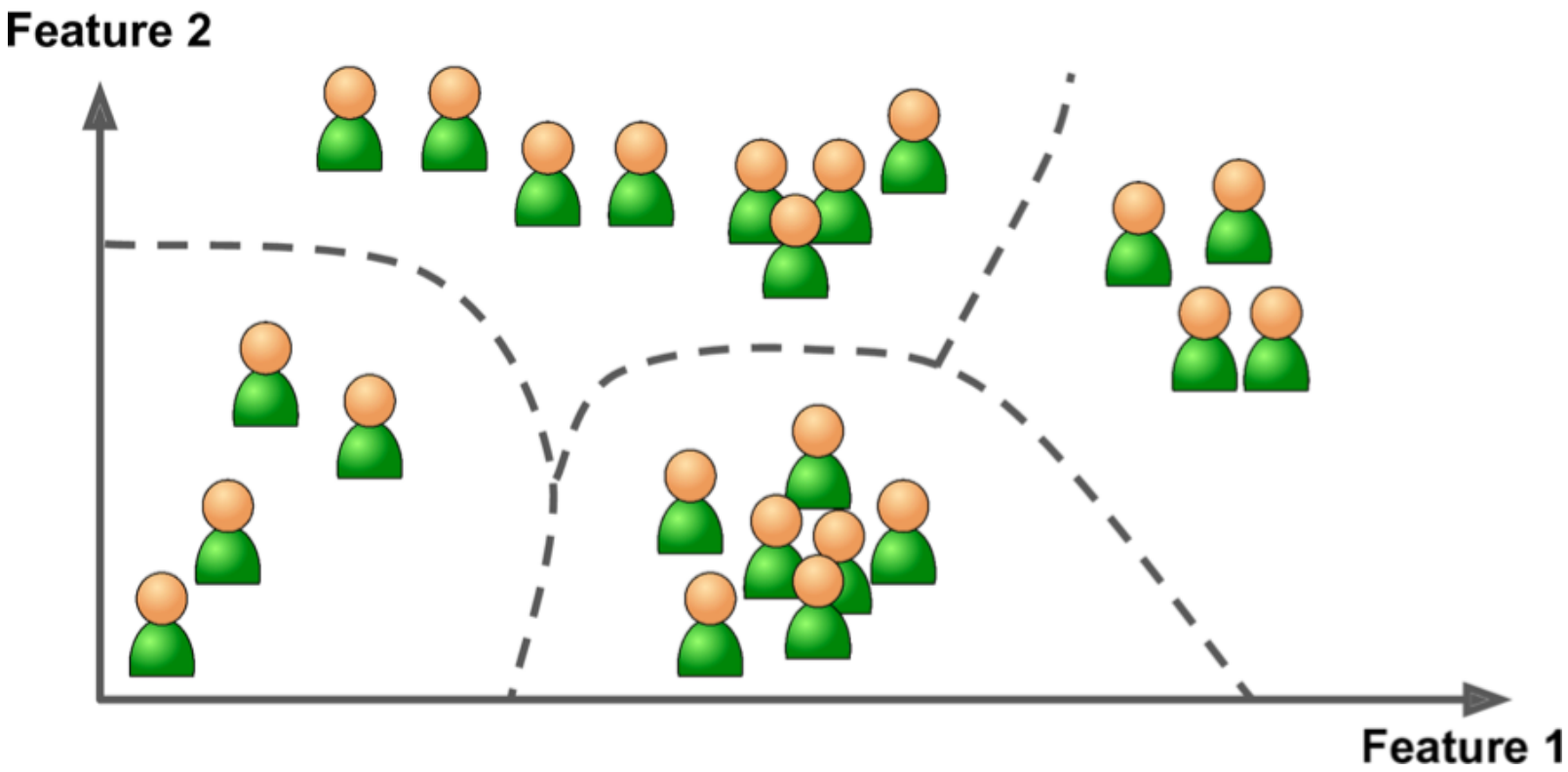
Clustering

Blog's visitors dataset:

Clustering is to group similar visitors without labeling.

For example, the visitors can be clustered based on their gender and the visited sections.

This may help you target your posts for each group.



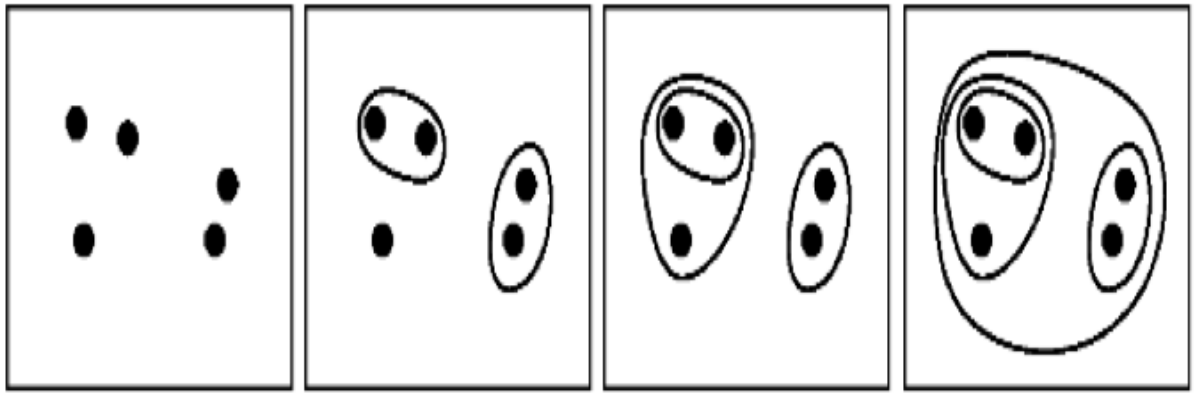
Hierarchical Clustering

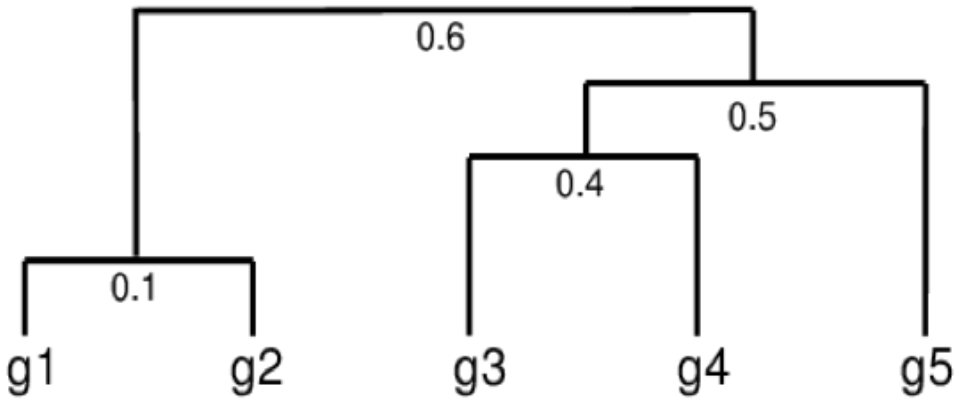
- **Agglomerative** or "bottom up" approach.

Each observation starts as a cluster, and pairs of clusters are grouped in one cluster, moves up the hierarchy.

- **Divisive** or "top down" approach.

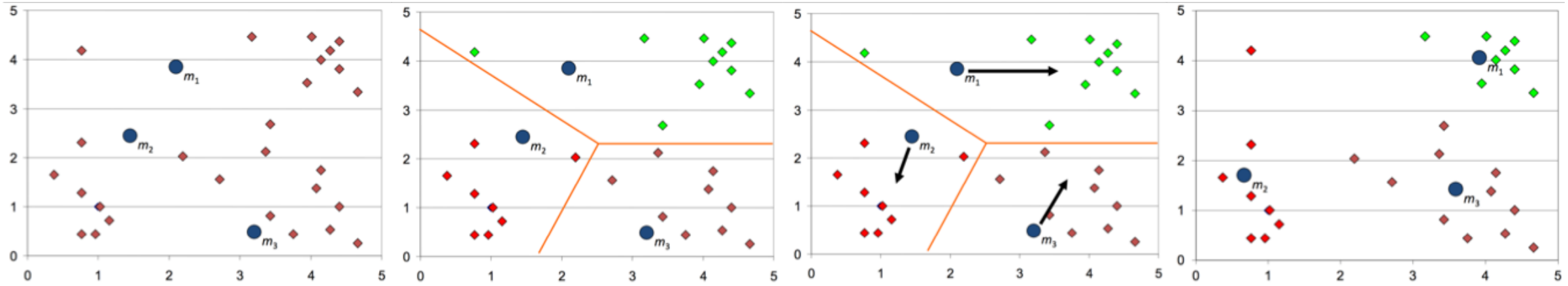
All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.





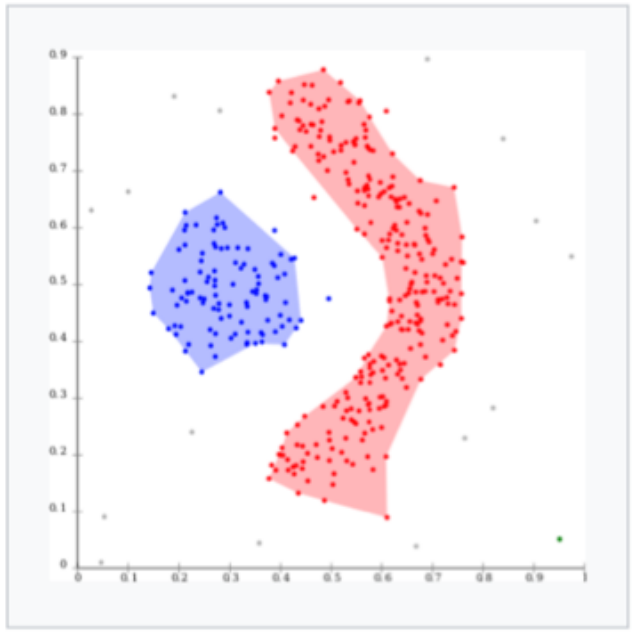
Partitional / Centeriod-based Clustering

- K-means:
  - Define K number of clusters.
  - For each object x_i
    - Calculate the distances between x_i and the K centroids
    - (Re)assign x_i to the cluster whose centroid is the closest to xi
  - Update the cluster centroids based on current assignment

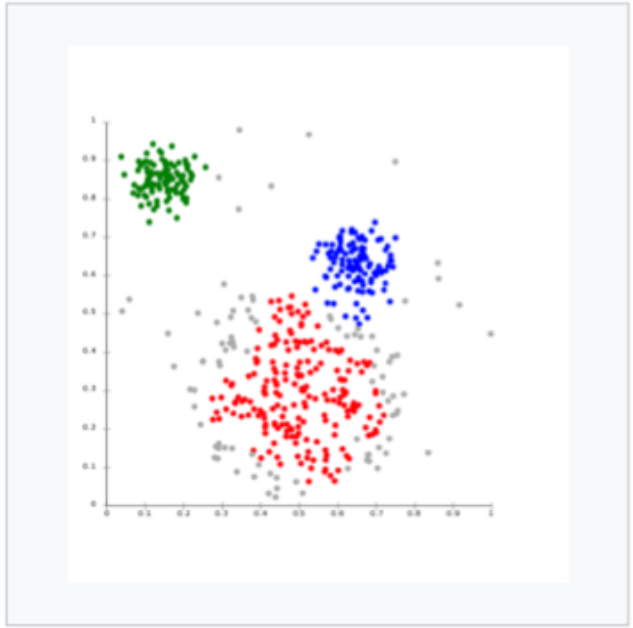


Density-based clustering

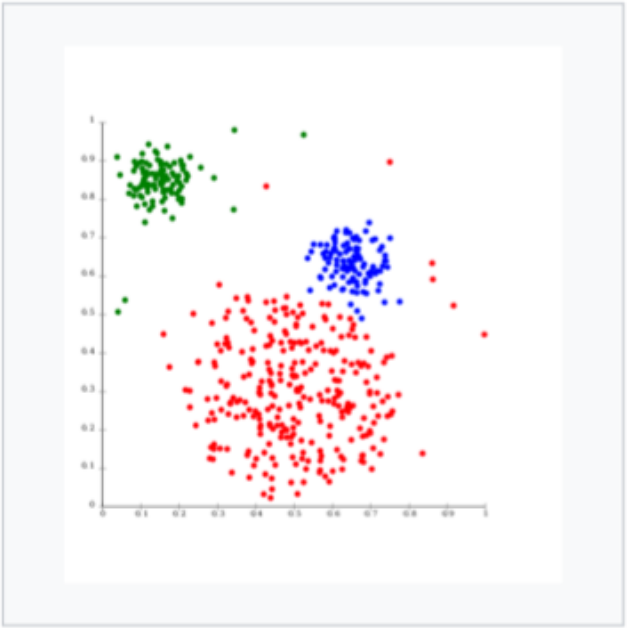
- The objective is to cluster the objects in areas of higher density.
- Objects that are not close to any cluster considered as noise and border points between clusters.



Density-based clustering with DBSCAN.



DBSCAN assumes clusters of similar density, and may have problems separating nearby clusters



OPTICS is a DBSCAN variant that handles different densities much better

Association Rules / Apriori Algorithm

Association rules analysis is a technique to uncover how items are associated to each other. There are three common ways to measure association.

Transaction 1	🍎 🥫 🍚 🍗
Transaction 2	🍎 🥫 🍚
Transaction 3	🍎 🥫
Transaction 4	🍎 🍐
Transaction 5	🍼 🥫 🍚 🍗
Transaction 6	🍼 🥫 🍚
Transaction 7	🍼 🥫
Transaction 8	🍼 🍐

$$\text{Support}\{🍎\} = \frac{4}{8}$$

how popular an itemset is?

$$\text{Confidence}\{🍎 \rightarrow 🥫\} = \frac{\text{Support}\{🍎, 🥫\}}{\text{Support}\{🍎\}}$$

how likely item Y is purchased when item X?

$$\text{Lift}\{🍎 \rightarrow 🥫\} = \frac{\text{Support}\{🍎, 🥫\}}{\text{Support}\{🍎\} \times \text{Support}\{🥫\}}$$

how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is.

Thank You !

Video tutorials: https://goo.gl/JqevFk

Source code: https://goo.gl/FAYr51