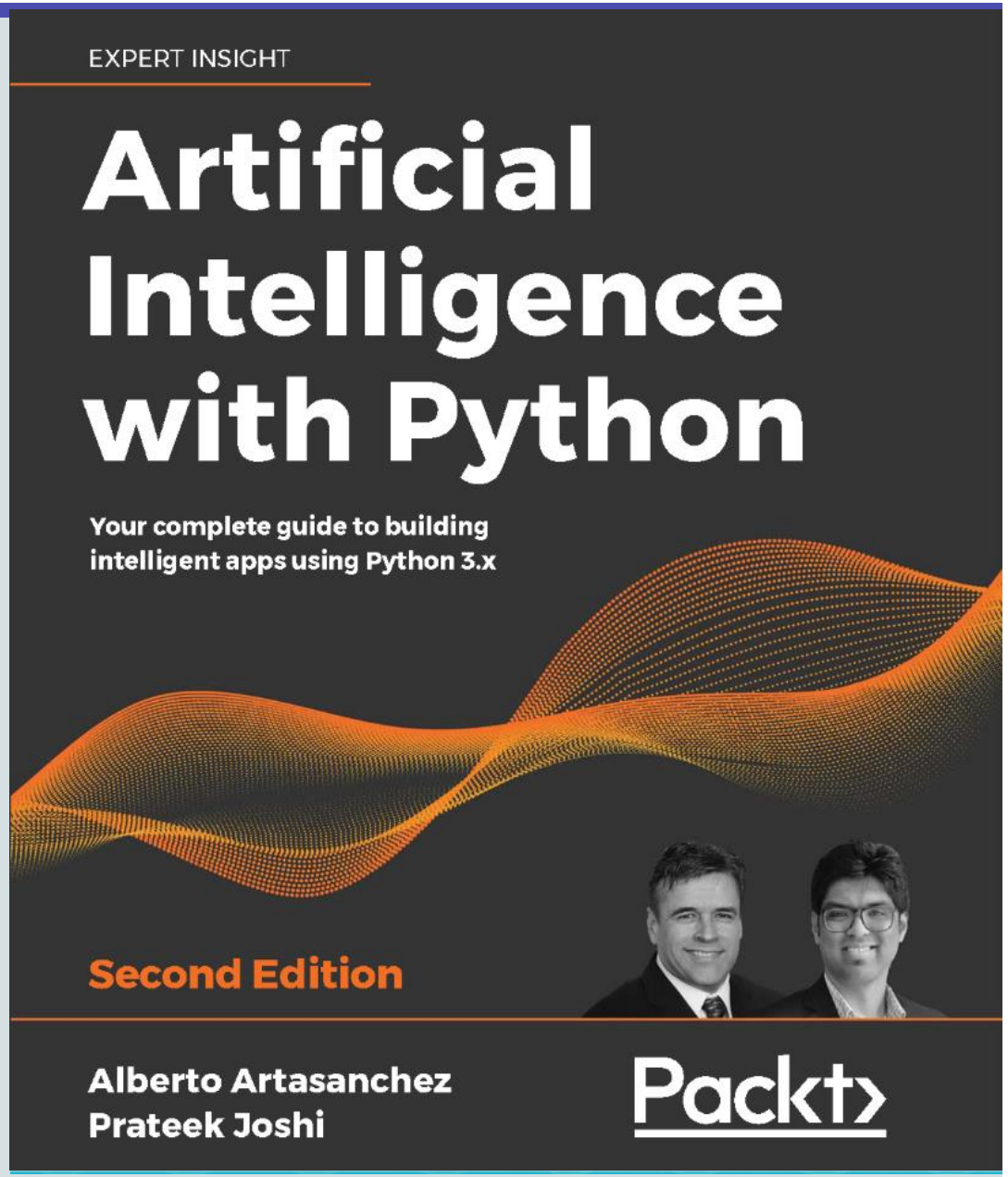
University of South-Eastern Norway
School of Business

Lecture 4: Detecting Patterns with Unsupervised Learning

Sinuo Wu

Course: AI for Business Applications (AI3000)



University of South-Eastern Norway
School of Business

Do it before we start:

Download Data From Canvas – AI3000 - Files – Day4 Practice

Sinuo Wu

Course: AI for Business Applications (AI3000)

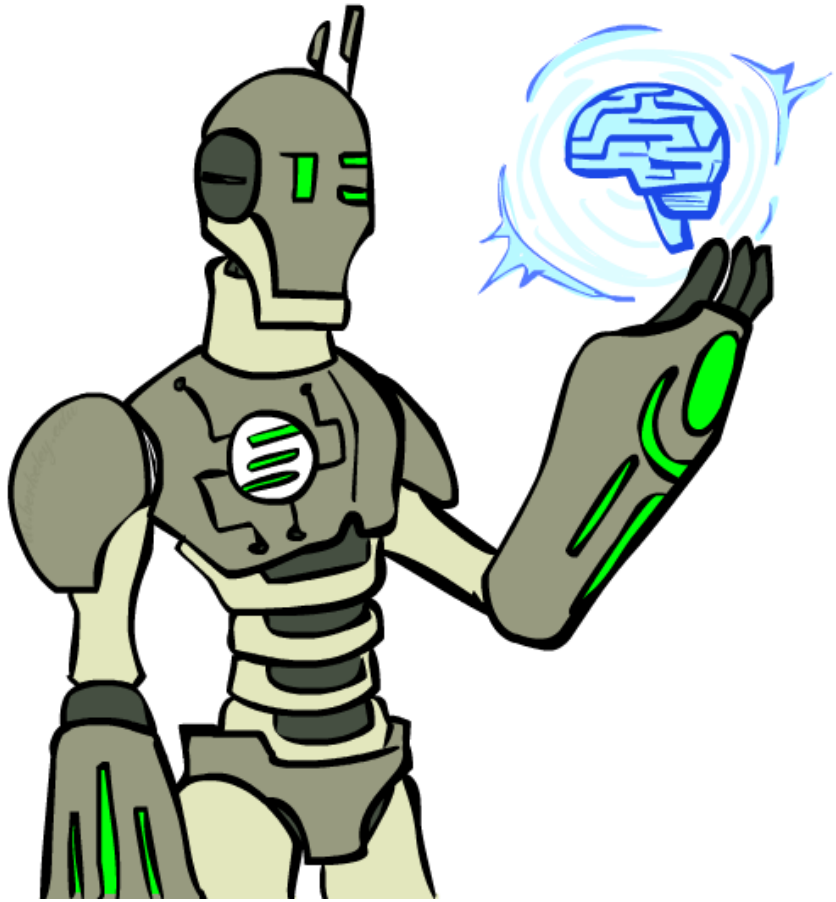
University of South-Eastern Norway
School of Business

Quiz

- Enter room number or Scan the QR code

Today

- Pattern recognition
- Unsupervised Learning
- Clustering data
  - K-mean algorithm
  - Gaussian mixture model
  - Affinity Propagation model



Pattern Recognition

Pattern recognition

- Pattern recognition is a process of description, grouping, and classification of patterns

- A pattern is an entity that could be associated with a name

Pattern Recognition

- In the context of artificial intelligence and machine learning, pattern recognition refers to **algorithms and techniques used to automatically identify patterns in data.**
- Because of big data and ML technologies emergence, a lot of data became available that was previously either deduced or speculated.
- In other words, now that we “knew more”, we moved from the goal of getting information itself to analyzing and understanding the data that was already coming to us

Pattern Recognition

- Of all the tools used in Big Data, pattern recognition is in the center.
  - It comprises the core of big data analytics-it gets the juice out of the data and uncovers the meaning of hidden behind it.

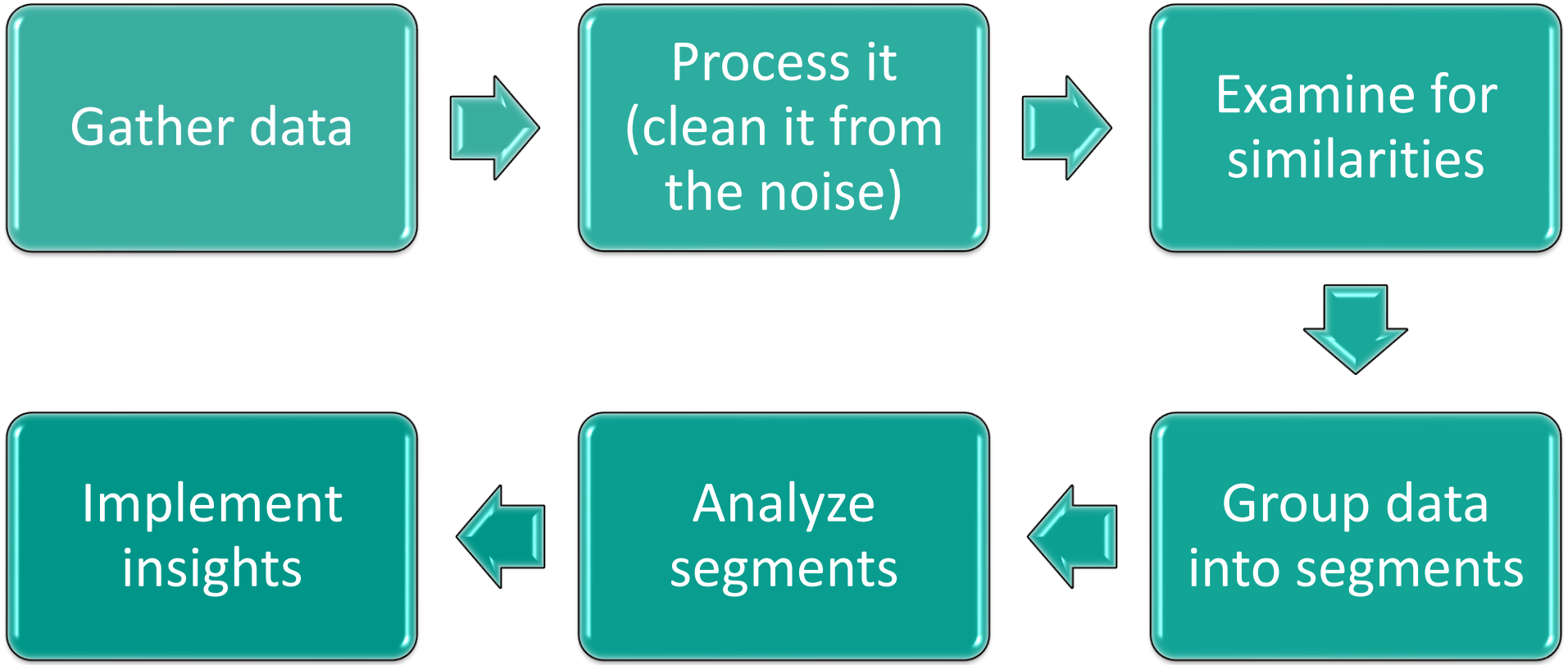
- Patter recognition gives a strategic advantage for the company which makes it capable of continuous improvement and evolution in the ever-changing market.

Example

Pattern Identification

- Identifying patterns in data
  - These patterns tell the data stories through ebbs and flows, spikes and flat lines
- Data itself can be anything:
  - Text
  - images
  - Sound
  - Sentiment
  - Any information on the sequential nature can be processed by pattern recognition algorithms, making the sequence comprehensible and enabling its practical use.

Pattern Recognition Process



Applications

- **Stock Market Forecasting:** Pattern recognition is used for comparative analysis of the stock exchanges and predictions of the possible outcomes.

- **Audience Research:** Pattern recognition refers to analyzing available user data and segmenting it by selected features. Google Analytics provides these features.

Applications

- Machine learning makes it possible to discover patterns in **supply chain data** by relying on algorithms that quickly pinpoint the most influential factors to a supply networks' success, while constantly learning in the process.

- Discovering new patterns in supply chain data has the potential to revolutionize any business. Machine learning algorithms are finding these new patterns in supply chain data daily, without needing manual intervention or the definition of taxonomy to guide the analysis.

Unsupervised Learning

Unsupervised Learning

We, as human beings, can perform lots of learning & pattern recognition tasks easily; **Consider listening to a conversation in a crowded room:**

- How do we separate a single source (speaker) from all the others?
- How do we process the words?

Consider further recognizing a face in a crowded scene:

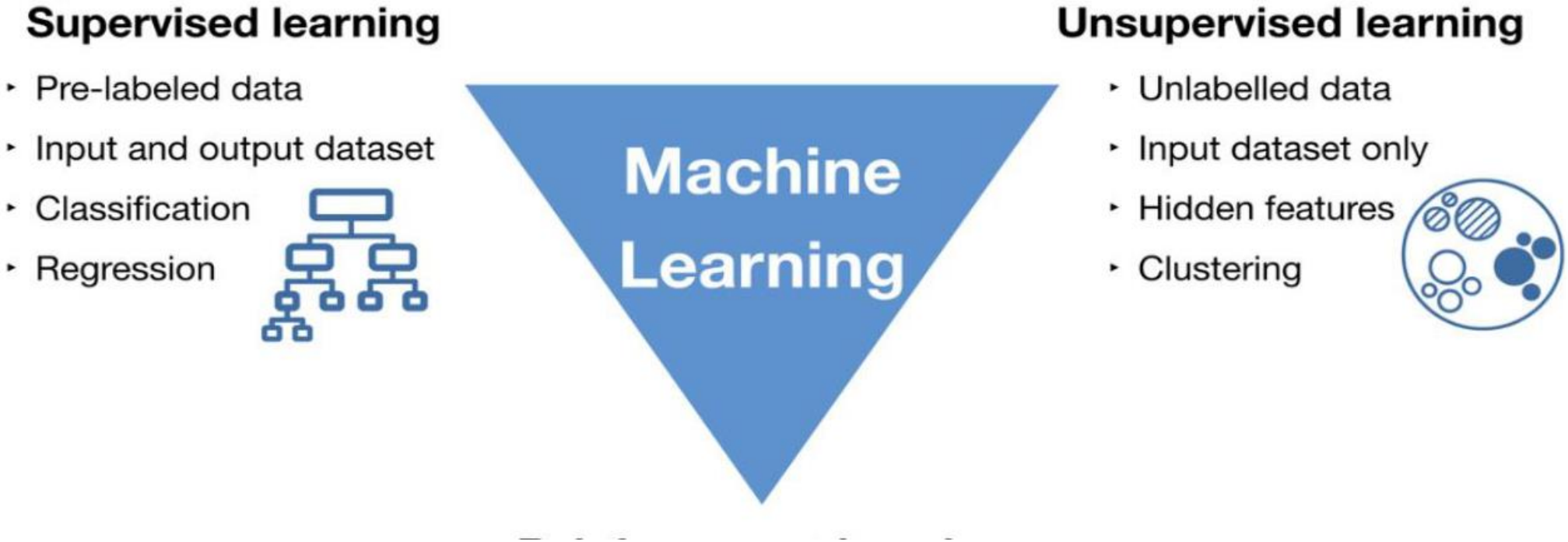
- How do we identify faces?
- How do we identify a particular face?

Unsupervised Learning

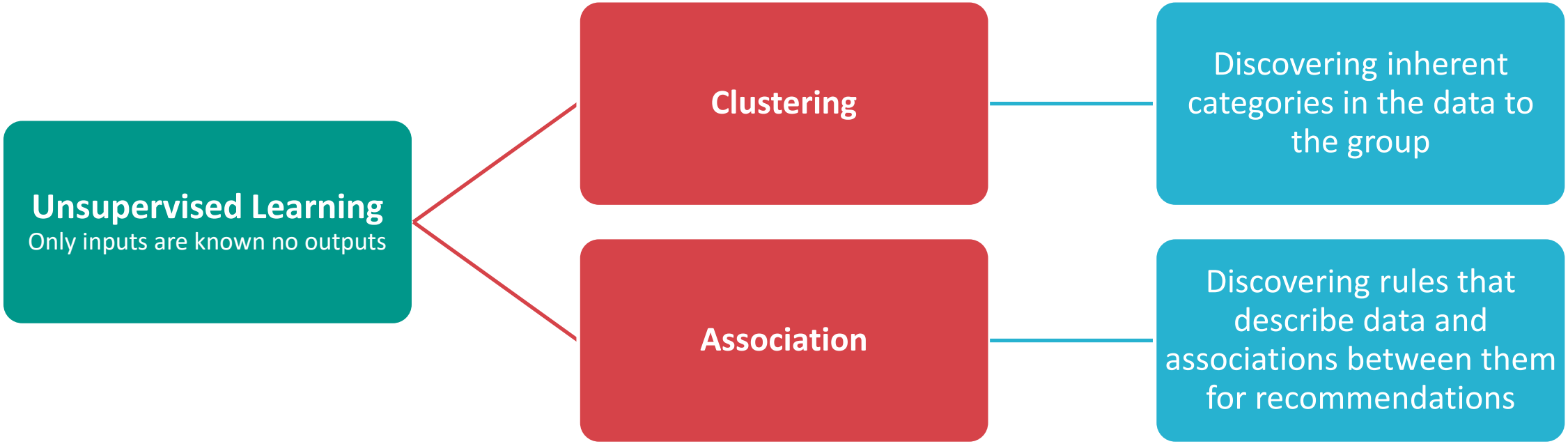
- A technique with the idea to explore hidden gems / patterns.
- To find some intrinsic structure in data.
- **Available data have no target attribute.**
- Machine Learning Algorithm takes training examples as the set of attributes/features alone.

Unsupervised Learning

Figure 1



Unsupervised Learning



Unsupervised Learning

- Applications:
  - Market segmentation
  - Stock markets
  - Natural language processing
  - Computer vision

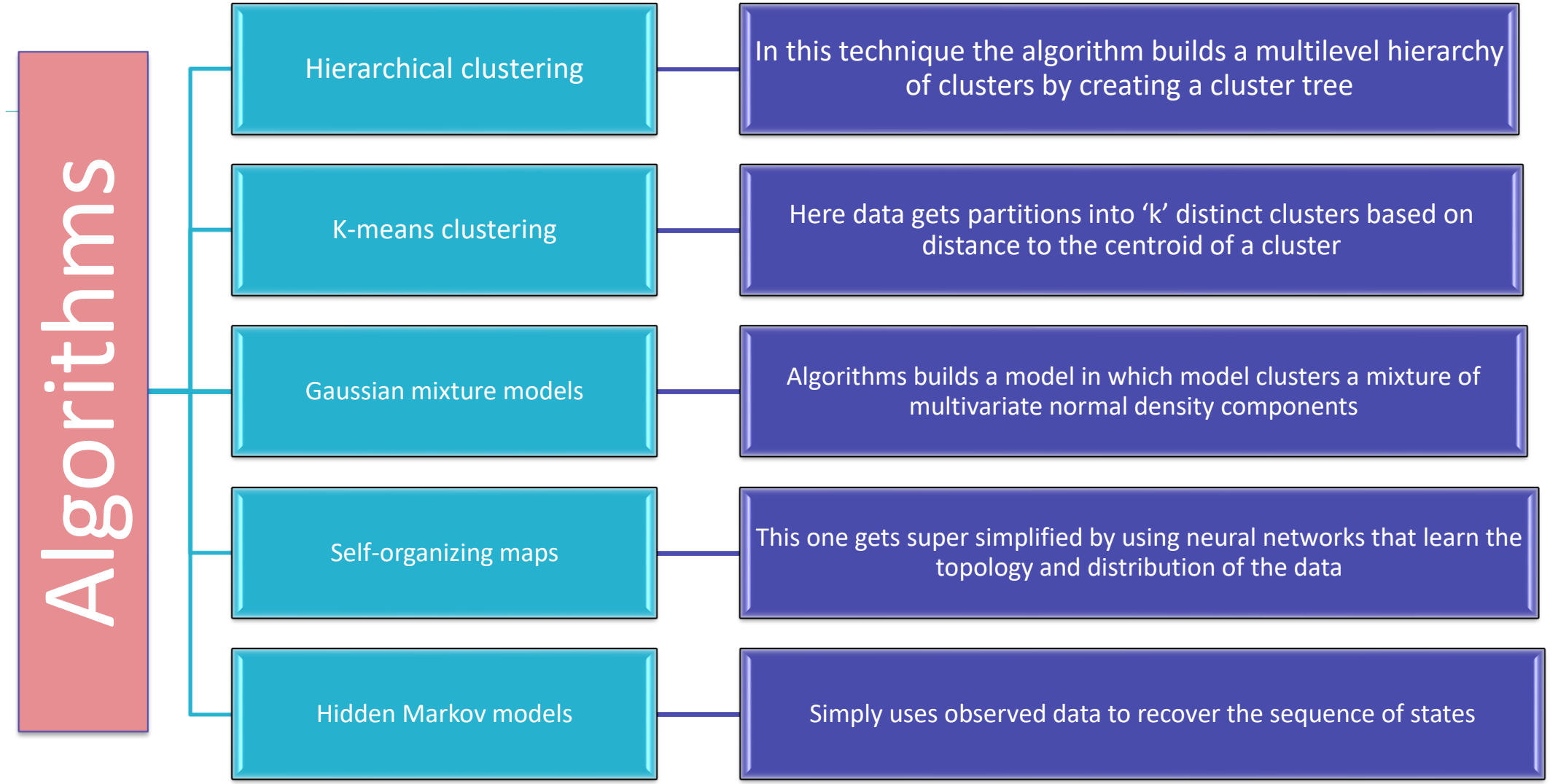
- Common scenarios
  - Data Exploration
  - Outlier Detection
  - Pattern Recognition

Strengths

- It can detect what human eyes can not understand
- The potential of hidden patterns can be very powerful for the business or even detect extremely amazing facts, fraud detection etc.
- Output can determine the un-explored territories and new ventures for businesses. Exploratory analytics can be applied to understand the financial, business and operational drivers behind what happened.

Weakness

- Unsupervised learning is harder as compared to supervised learning.
- It can be a costly affair, as we might need external expert look at the results for some time.
- Usefulness of the results; difficult to confirm since no answer labels are available.


Algorithms
Hierarchical clustering
In this technique the algorithm builds a multilevel hierarchy of clusters by creating a cluster tree
K-means clustering
Here data gets partitions into ‘k’ distinct clusters based on distance to the centroid of a cluster
Gaussian mixture models
Algorithms builds a model in which model clusters a mixture of multivariate normal density components
Self-organizing maps
This one gets super simplified by using neural networks that learn the topology and distribution of the data
Hidden Markov models
Simply uses observed data to recover the sequence of states

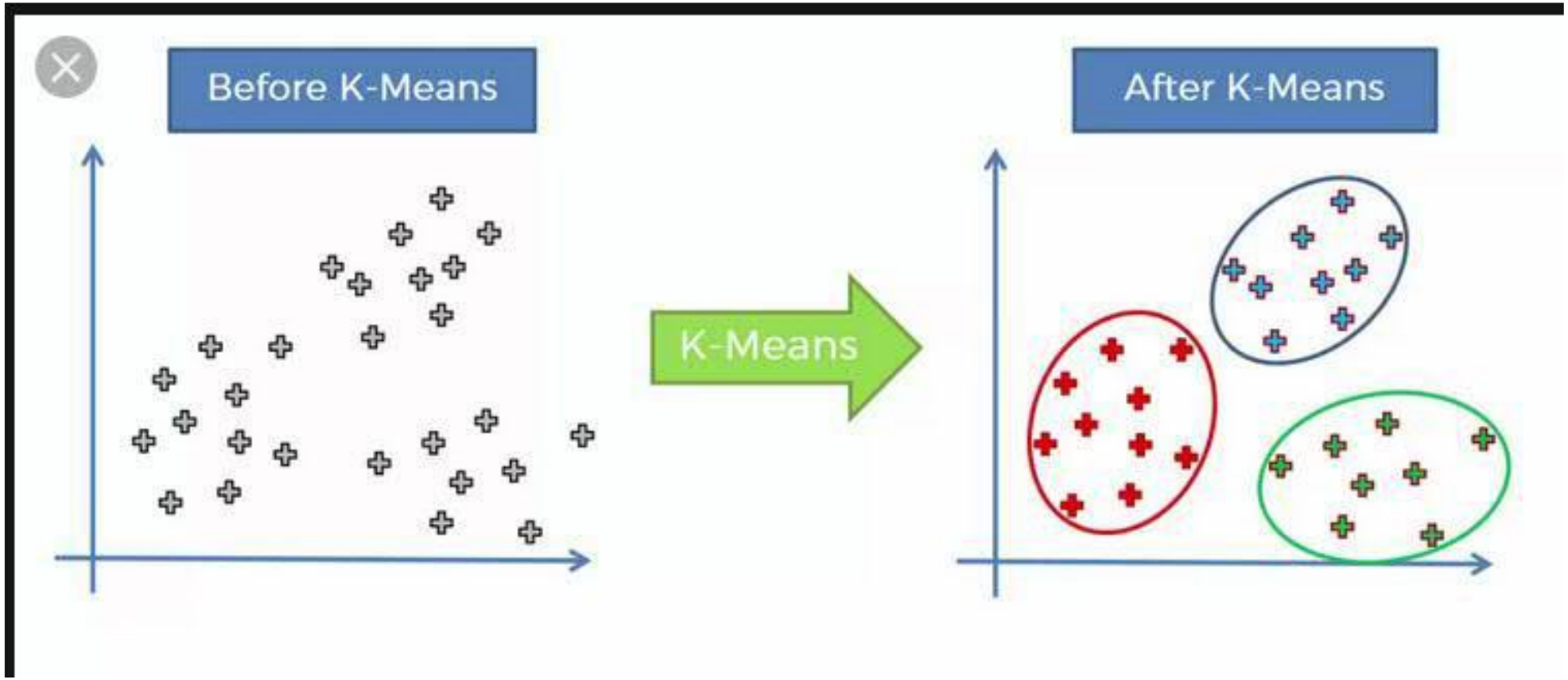

Have a Break!

K-Means

K-means

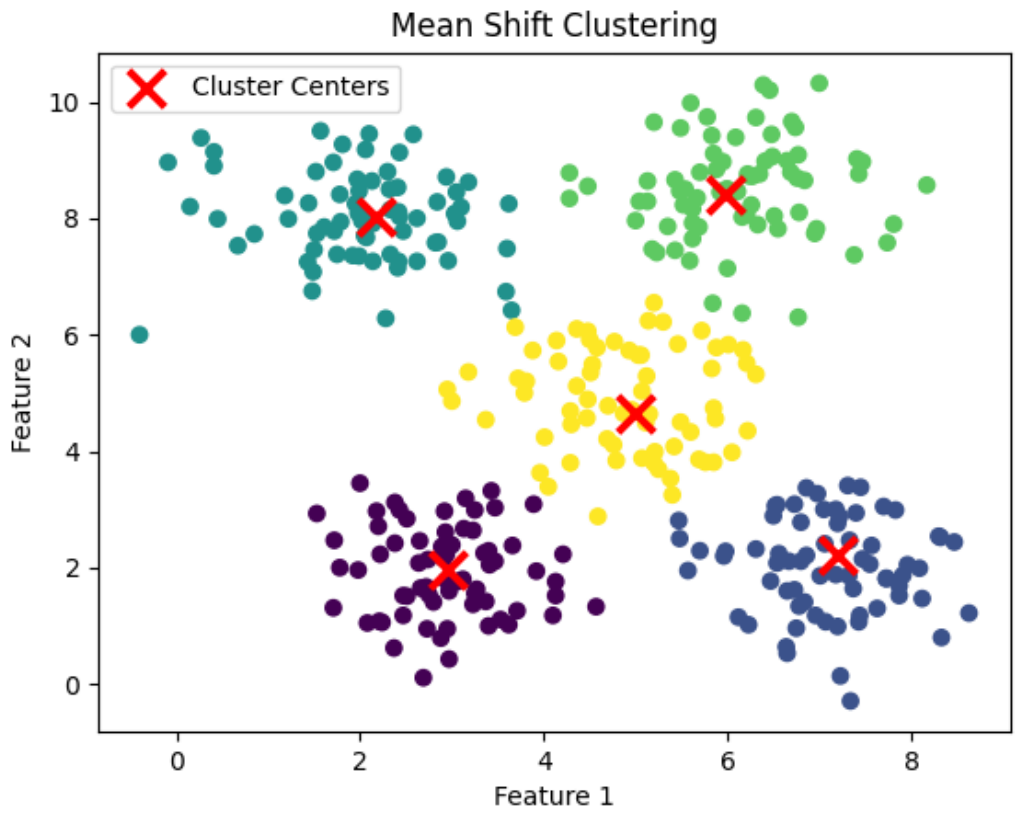
- The K-Means algorithm is a well-known algorithm for clustering data.
- In order to use it, the number of clusters is assumed beforehand.
- The data is segmented into *K* subgroups using various data attributes.
- The number of clusters is fixed, and the data is classified based on that number.
- The main idea here is that we need to update the locations of the centroids with each iteration.
  - A centroid is the location representing the center of the cluster. We continue iterating until we have placed the centroids at their optimal locations.

K-mean



The number of clusters (Mean Shift algorithm)

```
import matplotlib.pyplot as plt
from sklearn.cluster import MeanShift, estimate_bandwidth
from itertools import cycle

# Load data from input file
X = np.loadtxt('data_clustering.txt', delimiter=',')

# Estimate the bandwidth of X
bandwidth_X = estimate_bandwidth(X, quantile=0.1, n_samples=len(X))

# Cluster data with MeanShift
meanshift_model = MeanShift(bandwidth=bandwidth_X, bin_seeding=True)
meanshift_model.fit(X)

# Extract the centers of clusters
cluster_centers = meanshift_model.cluster_centers_
print('\nCenters of clusters:\n', cluster_centers)

# Estimate the number of clusters
labels = meanshift_model.labels_
num_clusters = len(np.unique(labels))
print("\nNumber of clusters in input data =", num_clusters)

# Plot the points and cluster centers
plt.figure()
markers = 'o*xvs'
for i, marker in zip(range(num_clusters), markers):
    # Plot points that belong to the current cluster
    plt.scatter(X[labels==i, 0], X[labels==i, 1], marker=marker, color='r')

    # Plot the cluster center
    cluster_center = cluster_centers[i]
    plt.plot(cluster_center[0], cluster_center[1], marker='o',
            markerfacecolor='black', markeredgecolor='black',
            markersize=15)

plt.title('Clusters')
plt.show()
```



Evaluation (Silhouette scores)

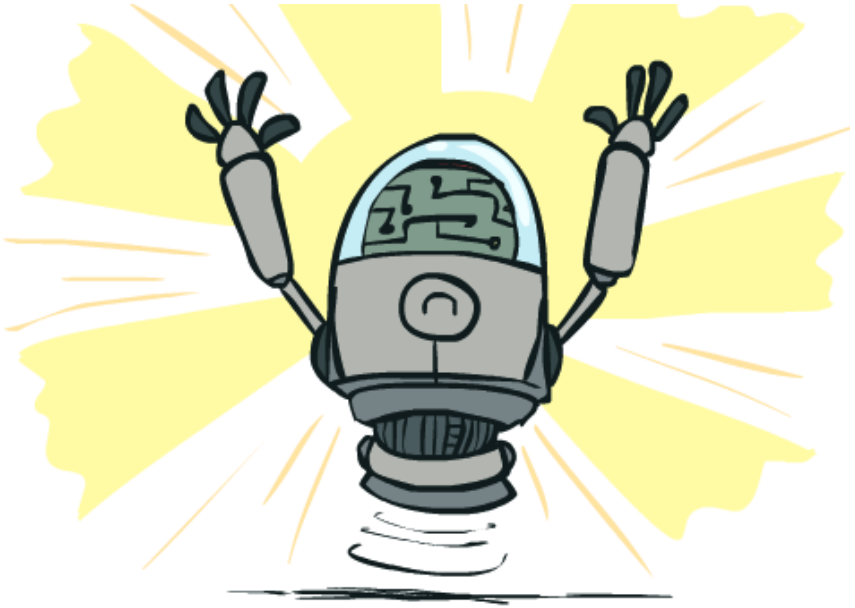
- Silhouette refers to a method used to check the consistency of clusters in our data.
- The silhouette score is a metric that measures how similar a data point is to its own cluster.

$$Silhouette\ score = (p - q) / max(p, q)$$

- The value of the silhouette score range lies between -1 to 1. A score closer to 1 indicates that the data point is very similar to other data points in the cluster, whereas a score closer to -1 indicates that the data point is not similar to the data points in its cluster.

Practice

TRY THE COMPETITION



Inspiration

https://www.youtube.com/watch?v=zPJtDohab-g

Example

Why Customer Segments Matters ?

The resulting five segments proved attitudinally differentiated and demographically distinct.

	Young Achievers	Concerned Moms	Financially Mature	Ho Hum	Solo Content
Demographics	Younger Skews male	Young, Middle Age Mostly female	Mature Skews male	Middle Age Mostly female	Mature Male and Female
Attitudes	Early adopters, technical Driven, Risk taker Price sensitive	Use social media, but not otherwise technical Don't know where to begin Price sensitive	Recognize value of insurance Confident about financial matters Least price sensitive	Late adopters Risk averse Not primary decision makers and not thinking about LI	Use social media Mistrustful of financial inst. Least interest in LI

CRM EXECUTIVE SUMMIT 10TH ANNIVERSARY
MERKLE

Case 1

Mall Customer data is a dataset that has hypothetical customer data. It puts you in the shoes of the owner of a supermarket. You have the customer data, and you need to divide the customers into various groups.

The data includes the following features:

1. Customer ID
2. Customer Gender
3. Customer Age
4. Annual Income of the customer (in Thousand Dollars)
5. Spending score of the customer (based on customer behavior and spending nature)


Have a Break!

Report

One representative from each group report for:

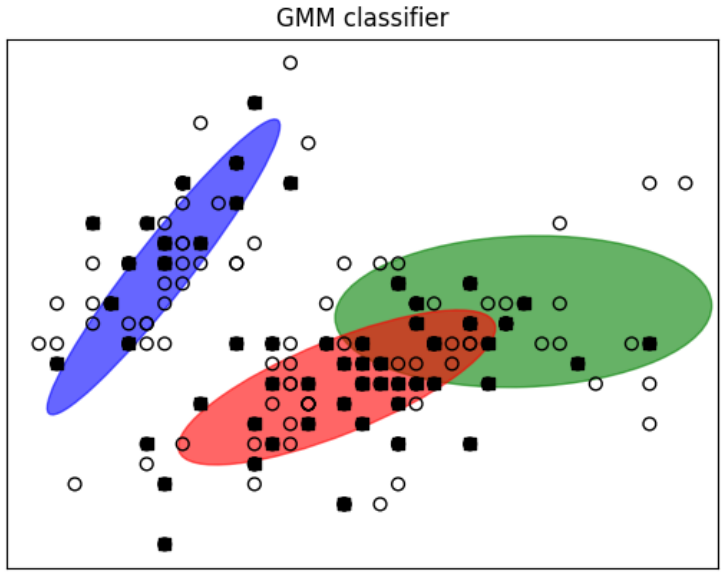
1. What is your results
2. Explain your results

Gaussian Mixture Models

Gaussian Mixture Models

Mixture Model: A type of probability density model where it is assumed that the data is governed by several **component distributions**.

Eg. Shopping habits of all the people in South America



Coding

```
# Load the iris dataset

# Extract the number of classes

# Build and Train GMM

# Draw boundaries

# Plot the data

# Compute predictions for training and testing data (evaluate)
```

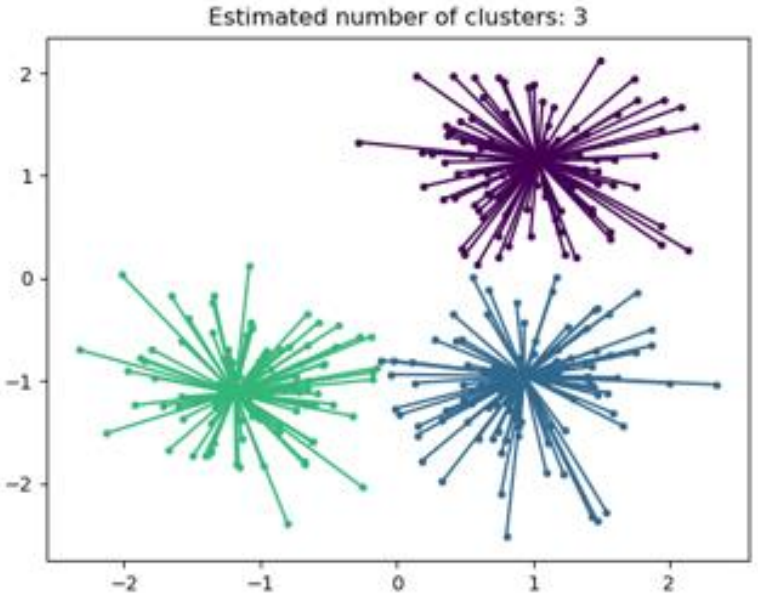
Difference

- https://www.youtube.com/watch?v=C7jhwN6H9LU

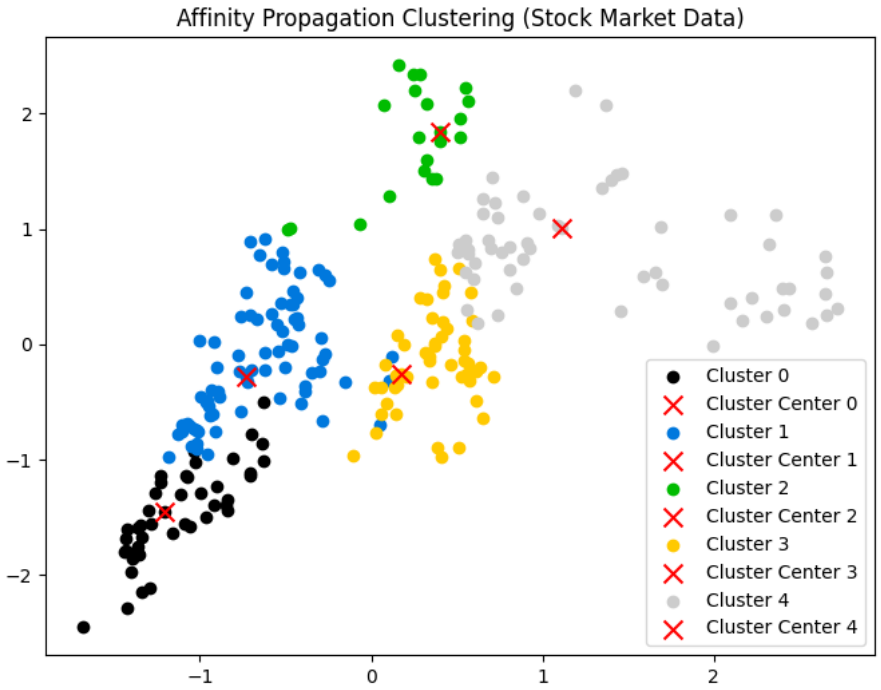
Affinity Propagation model

AP Model

- A clustering algorithm that doesn't require us to specify the number of clusters beforehand.
- It finds out representatives of clusters, called exemplars, using a technique called message passing.
- It simultaneously considers all training data points as potential exemplars. It then passes messages between the data points until it finds a set of exemplars.



Coding



Deprecation of the **"Open"** and **"Close"** attributes in the latest version of the **yfinance** library. Use DataFrame by calling the history symbol.

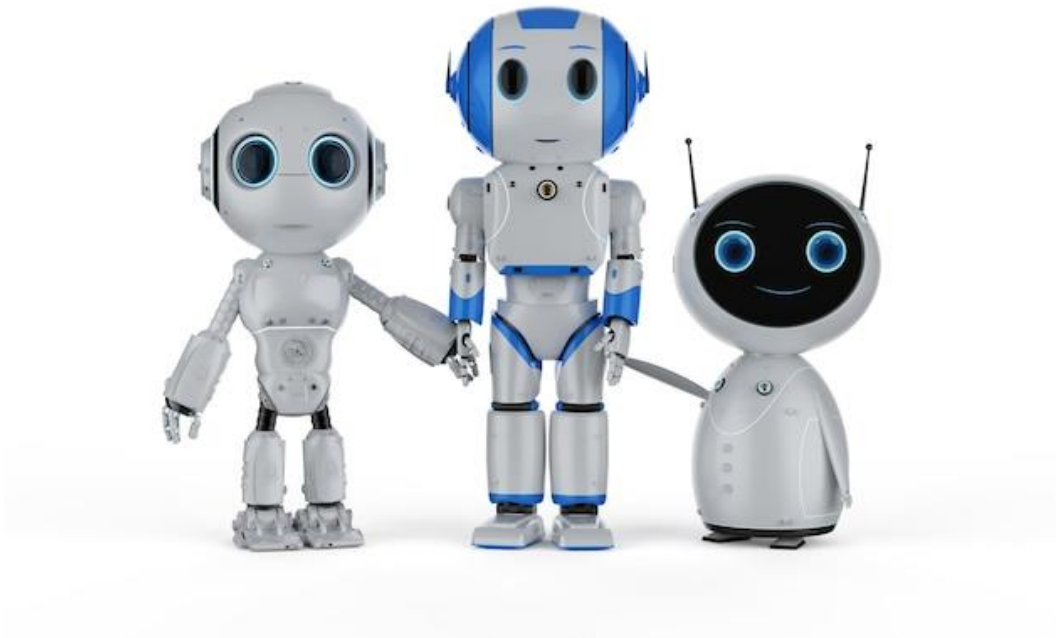
```
import datetime
import json
import numpy as np
import matplotlib.pyplot as plt
from sklearn import covariance, cluster
import yfinance as yf


# Input file containing company symbols
input_file = 'company_symbol_mapping.json'

# Load the company symbol map
with open(input_file, 'r') as f:
    company_symbols_map = json.loads(f.read())

symbols, names = np.array(list(company_symbols_map.items())).T

# Load the historical stock quotes
start_date = datetime.datetime(2019, 1, 1)
end_date = datetime.datetime(2019, 1, 31)
quotes = [yf.Ticker(symbol).history(start=start_date, end=end_date)
                for symbol in symbols]

# Extract opening and closing quotes
opening_quotes = np.array([quote.Open for quote in quotes]).astype(np.float)
closing_quotes = np.array([quote.Close for quote in quotes]).astype(np.float)

# Compute differences between opening and closing quotes
quotes_diff = closing_quotes - opening_quotes

# Normalize the data
X = quotes_diff.copy().T
X /= X.std(axis=0)

# Create a graph model
edge_model = covariance.GraphLassoCV()

# Train the model
with np.errstate(invalid='ignore'):
    edge_model.fit(X)

# Build clustering model using Affinity Propagation model
_, labels = cluster.affinity_propagation(edge_model.covariance_)
num_labels = labels.max()

# Print the results of clustering
print('\nClustering of stocks based on difference in opening and closing quotes:\n')
for i in range(num_labels + 1):
    print("Cluster", i+1, "==>", ', '.join(names[labels == i]))
```

Case 2

Try stock_symbols = ["AAPL", "MSFT", "GOOGL", "AMZN", "TSLA", "CGC", "BABA"]


Have a Break!

Group Up!



Work on your assignment!

Guidance

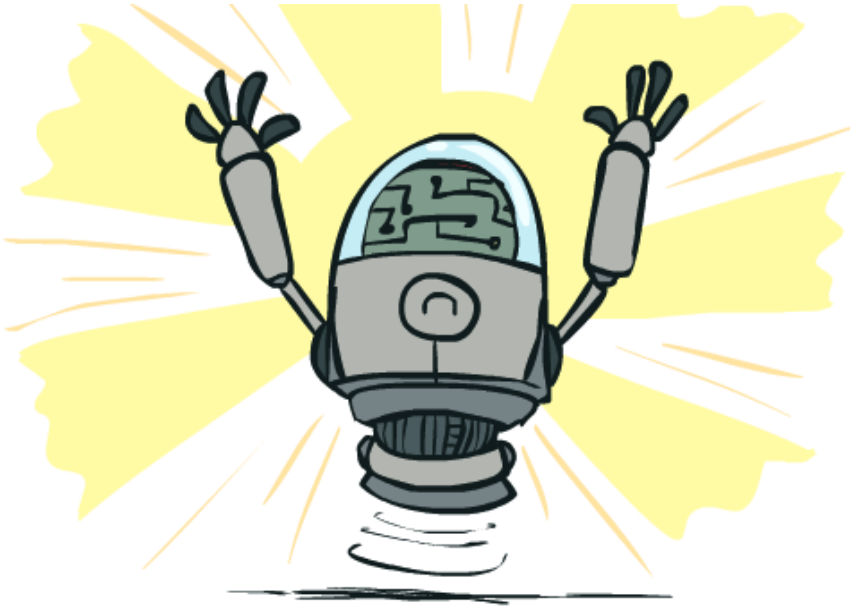
https://towardsdatascience.com/customer-segmentation-using-k-means-clustering-d33964f238c3

https://lisalondon.medium.com/applying-k-means-clustering-model-to-customer-segmentation-4254386c7563

https://medium.datadriveninvestor.com/using-an-unsupervised-machine-learning-algorithm-to-detect-different-stock-market-regimes-5c6354a1826a

Room Change: B207

ASSIGNMENT PROCESS



THANKS