



DATA MINING (Week 1)

DSBA CURRICULUM DESIGN

FOUNDATIONS

Data Science Using
Python

Statistical Methods
for Decision Making

CORE COURSES

Advanced
Statistics

Data Mining
(Week-1/6)

Predictive Modelling

Machine Learning

Time Series
Forecasting

Data Visualization

DOMAIN APPLICATIONS

Financial Risk
Analytics

Web & Social Media
Analytics

Marketing Retail
Analytics

LEARNING OBJECTIVE OF THIS MODULE

- Clustering
- CART & Model Performance Measures
- Random Forest
- Neural Network

LEARNING OBJECTIVES OF THIS SESSION -

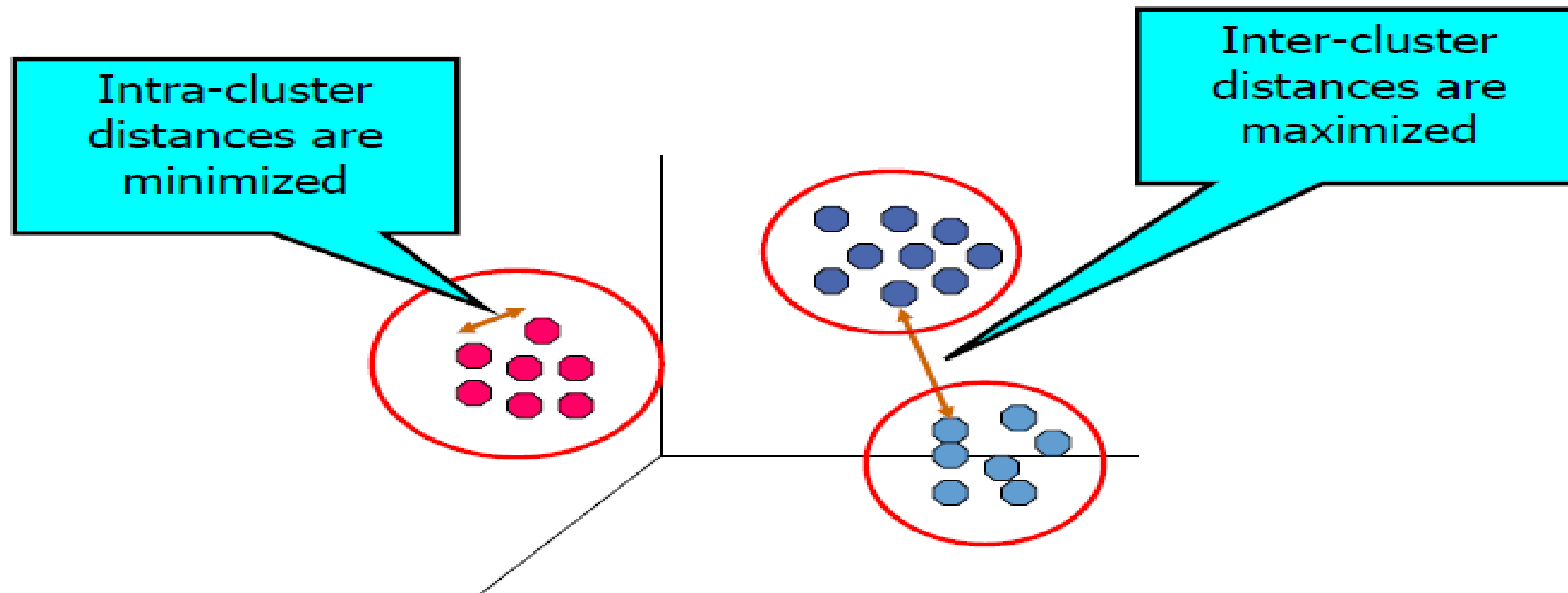
- Hierarchical Clustering
- K- Mean Clustering

TRY ANSWERING THE FOLLOWING

- What is the formula to calculate Euclidean Distance?
- Is Clustering a part of Supervised learning?
- Which clustering technique is preferred for dealing with large data- Hierarchical or K-Mean Clustering?



BROAD OVERVIEW



This is classic case of Unsupervised learning

Few Application of Clustering

Image processing

- cluster images based on their visual content

Web

- Cluster groups of users based on their access patterns on webpages
- Cluster webpages based on their content

Market Segmentation

- customers are segmented based on demographic and transaction history information, and a marketing strategy is tailored for each segment

Market structure analysis

- identifying groups of similar products according to competitive measures of similarity

Finance

- cluster analysis can be used for creating *balanced portfolios*

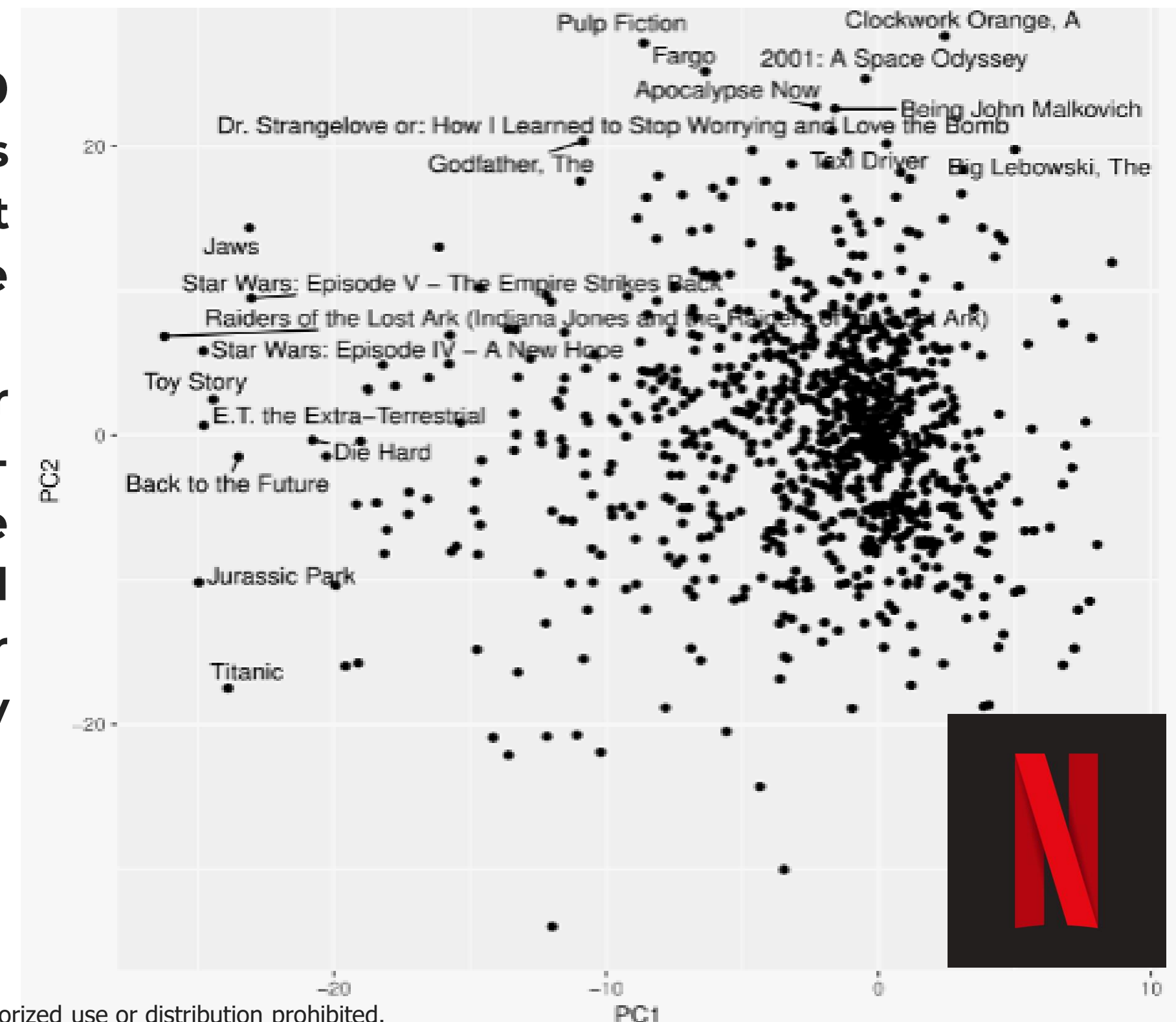
Industry Application - Netflix' Movie Recommendation System

In 2006 Netflix rolled out a contest with a prize money of 1 million \$ to develop a movie recommendation system which could deliver 10% reduction in terms of root mean square error compared to their own in-house developed algorithm called Cinematch.

They had a huge database of over 100 million ratings of 17,770 movies from 4,80,189 customers. Now anything that requires sharper customer targeting can never be achieved without forming clusters. But basis what features or attributes do we form cluster here?

Genre is the answer. It is quite evident that every movie lover has a liking for certain type of movies e.g. Action-Adventure-Science Fiction or Crime-Thrillers etc. Around 10 clusters were formed using 19 different genres. More meaningful recommendations could be made knowing the cluster membership of subscribers, and finally the desired accuracy was achieved after 3 years of rigorous competition.

Reference: <https://www.netflixprize.com/>
<https://www.thrillist.com/entertainment/nation/the-netflix-prize>



Industry Application - Clustering the stocks by monthly returns

It is a common practice to apply machine learning algorithms in order to predict stock price movement i.e. Upwards or Downwards i.e. essentially a classification problem. But is it wise to club all the listed stocks together and try to fit them to a model? Well, not really!

Let's say we take a dataset where we have stock returns from January to November, and basis this we want to predict if a given stock's price will move up or go down in December. It is observed that the prediction accuracy on such datasets without cluster analysis is around 50-55% whereas the same significantly improves to ~70% when models were developed specific to each cluster. These clusters can be obtained using K-means clustering, typically with the value of k between 3 to 5.

Cluster analysis helps group the stocks of similar nature together and hence improves prediction accuracy.

Similar Reference: <https://econwpa.ub.uni-muenchen.de/econ-wp/fin/papers/0509/0509022.pdf>



50.22	↑	4.04	↑	0.07%	0 (N/A)	3000	110
58.76	↑	2.58	↑	0.02%	0 (N/A)	2300	90
58.04	↑	1.86	↑	1.63%	0 (N/A)	2445	90
56.28	↑	0.10	↑	0.01%	0 (N/A)	-2130	101
57.60	↑	1.42	↑	0.00%	0 (N/A)	3027	80
58.67	↑	2.49	↑	0.01%	0 (N/A)	3520	80
16.05	↑	2.05	↑	0.04%	0 (N/A)	3430	90
05.76	↑	0.76	↑	0.01%	0 (N/A)	1400	100
02.31	↑	2.31	↑	0.01%	0 (N/A)	37%	90
67.97	↑	1.97	↑	0.00%	0 (N/A)	1401	10
48.24	↓	-1.76	↓	0.01%	0 (N/A)	120	80
66.59	↑	0.59	↑	0.01%	0 (N/A)	5004	-8
14.28	↑	2.28	↑	-0.01%	0 (N/A)	4000	110
69.78	↑	0.78	↑	0.00%	0 (N/A)	5073	90
34.55	↑	4.55	↑	0.00%	0 (N/A)	9	-10
20.70	↑	8.70	↑	0.44%	0 (N/A)	29	-20
32.40	↑	4.40	↑	0.12%	0 (N/A)	127	80
98.43	↓	-0.57	↓	0.78%	0 (N/A)	86	-10
5.94	↑	5.94	↑	1.06%	0 (N/A)	81	-10
			↑	3.32%	0 (N/A)	300	30
			↓	-0.06%	0 (N/A)	912	20
			↑	0.05%	0 (N/A)	1200	-10
			↑	0.32%	0 (N/A)	1200	110

CASE STUDY - Strategy to group Engineering Colleges

You are an independent trainer who would like to pitch your Data Science training program to a set of Engineering colleges. You have data of 26 colleges after survey using questionnaires. Each college has been given a score for 5 performance criteria-Teaching, Fees, Placement, Internship & Infrastructure. Ratings are in the standardized scale of 1 to 5 where 5 has a higher weightage than 1. Segment the colleges into groups and come up with your pitch recommendations for each segment.

CASE STUDY - Clustering of Banks

We have a transaction details of 515 banks which include number of DD taken, Withdrawals, Deposits, Area of the branch and Average Walk-Ins. Profile the banks into segments and come up with recommendations for each segment.



ANY QUESTIONS



HAPPY LEARNING