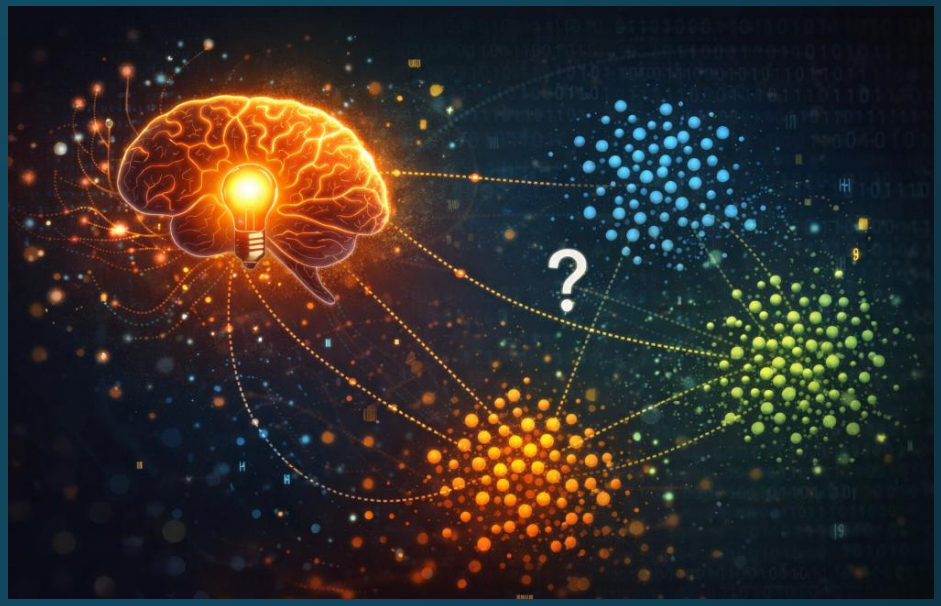
Data Analysis and Knowledge Discovery

Lecture 5-6

Faculty of Mathematics and Computer Science
Babeș-Bolyai University

Sergiu Limboi, PhD Teaching Assistant

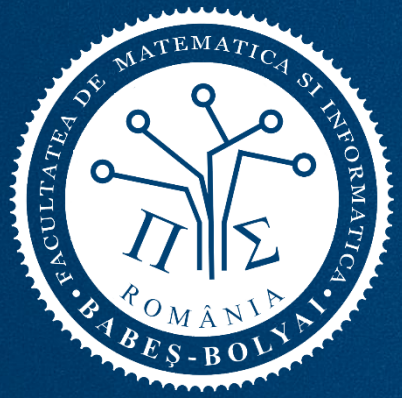
Motto: “When labels are missing, patterns become your truth.”



Discovering Structure in Chaos: Unsupervised Learning & Clustering

AGENDA

- Warm-Up
- Real world problem
- Machine learning
- Unsupervised learning
- Clustering
- What is a cluster?
- Similarity measures
- Clustering techniques
- Hierarchical clustering
- Partitioning clustering
- Clustering evaluation
- Applications of clustering
- Industry case studies
- Key Takeaways



Warm-Up

Faculty of Mathematics and Computer Science

Warm-Up

Go to www.menti.com and enter the code **4699 0459**

or use the QR code

Real world problem

Real world problem

- Netflix does NOT know your “true taste label”
- Banks do NOT know “customer types”
- TikTok does NOT know “user categories”

- But they still group users.

Machine learning

Machine learning

- Learning can be explained as the activity of acquiring knowledge or understanding it, as well as the ability to assimilate concepts through study, training, or experience.

- Machine Learning refers to changes in systems that perform tasks associated with various topics and concepts from Artificial Intelligence, such as diagnosis, prediction, planning, control, or detection.

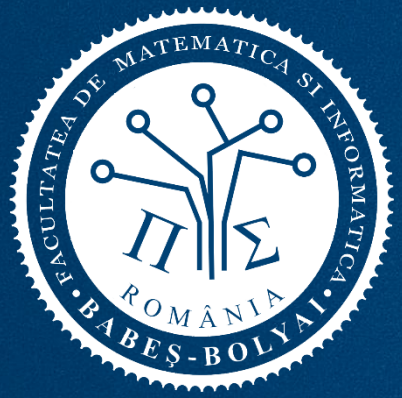
- The goal is the design and development of algorithms and methods that allow a computational system to *"learn"*.

Machine learning

- **Why should computer systems learn?**
  - Some tasks can be defined **only through examples**, therefore, we need to provide **input–output pairs**, even when there is no explicit relationship between input data and results.
  - There may exist **hidden or unknown information** that reflects correlations or relationships in the data.
  - The **large volume of information** can be difficult for the human mind to encode or process manually.
- We assume there is an unknown function f that maps inputs to outputs.
- The goal of the learning algorithm is to discover or approximate this function based on data

Machine learning

- Types of learning:
  - **Supervised Learning**
    - We sometimes know (at least approximately) the values of the function f for m cases in a training set.
    - After the training phase, the system can infer what the function would be for a new set of data.
  - **Unsupervised Learning**
    - We only have a set of vectors (a dataset) for which the function is unknown.
    - It involves dividing (partitioning) the dataset into subsets or groups.
    - In this case, the value of the function corresponds to the name of the subgroup (class) to which the input vector belongs.
  - **Reinforcement Learning**
    - The system interacts with the environment and can receive rewards or penalties.
  - **Semi-Supervised Learning**
    - We have a **small amount of labelled data** and a **large amount of unlabelled data**.
    - The model learns from both:
      - labelled data → provides guidance
      - unlabelled data → helps discover structure
    - This approach improves performance when labelling data is expensive or limited.



Unsupervised learning

Unsupervised learning

- **Basic Idea**
  - The system receives information (sequences such as $x_1, x_2, \dots$) but it is not provided with predefined outputs obtained beforehand, and it does not receive rewards from the environment.
- The unsupervised style is more natural to the brain than the supervised one.
- Humans and animals learn by analysing their environment and identifying objects and events around them.

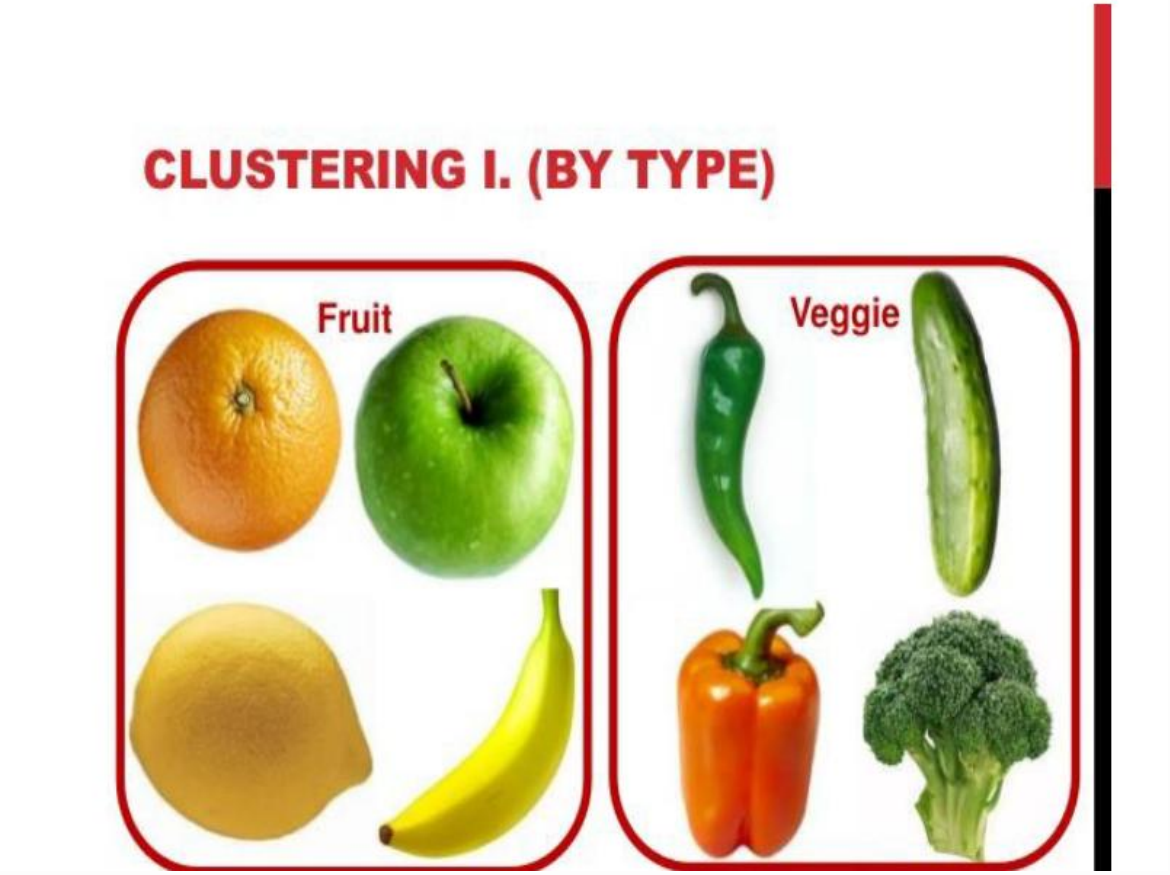
Unsupervised learning

- Can we group fruits based on their characteristics?

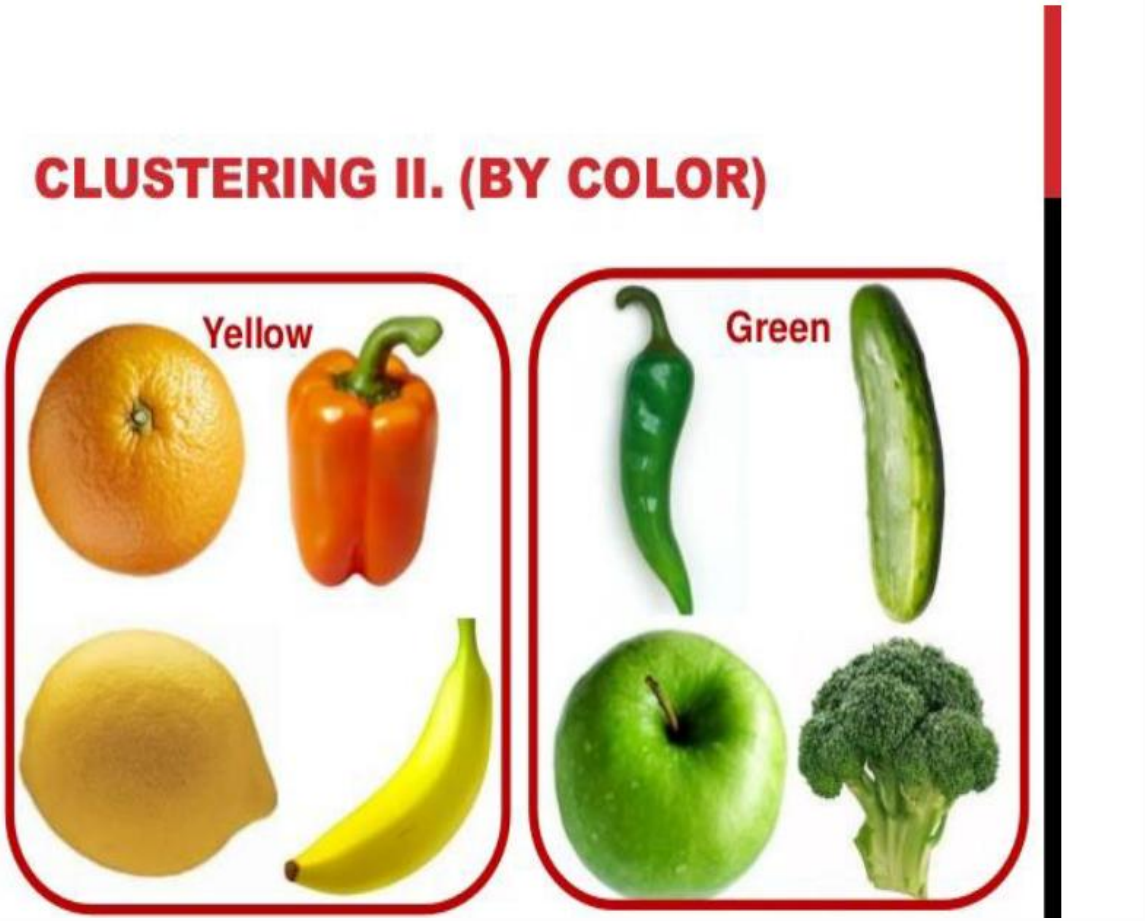
Unsupervised learning- examples

- What happens if we have both fruits and vegetables?

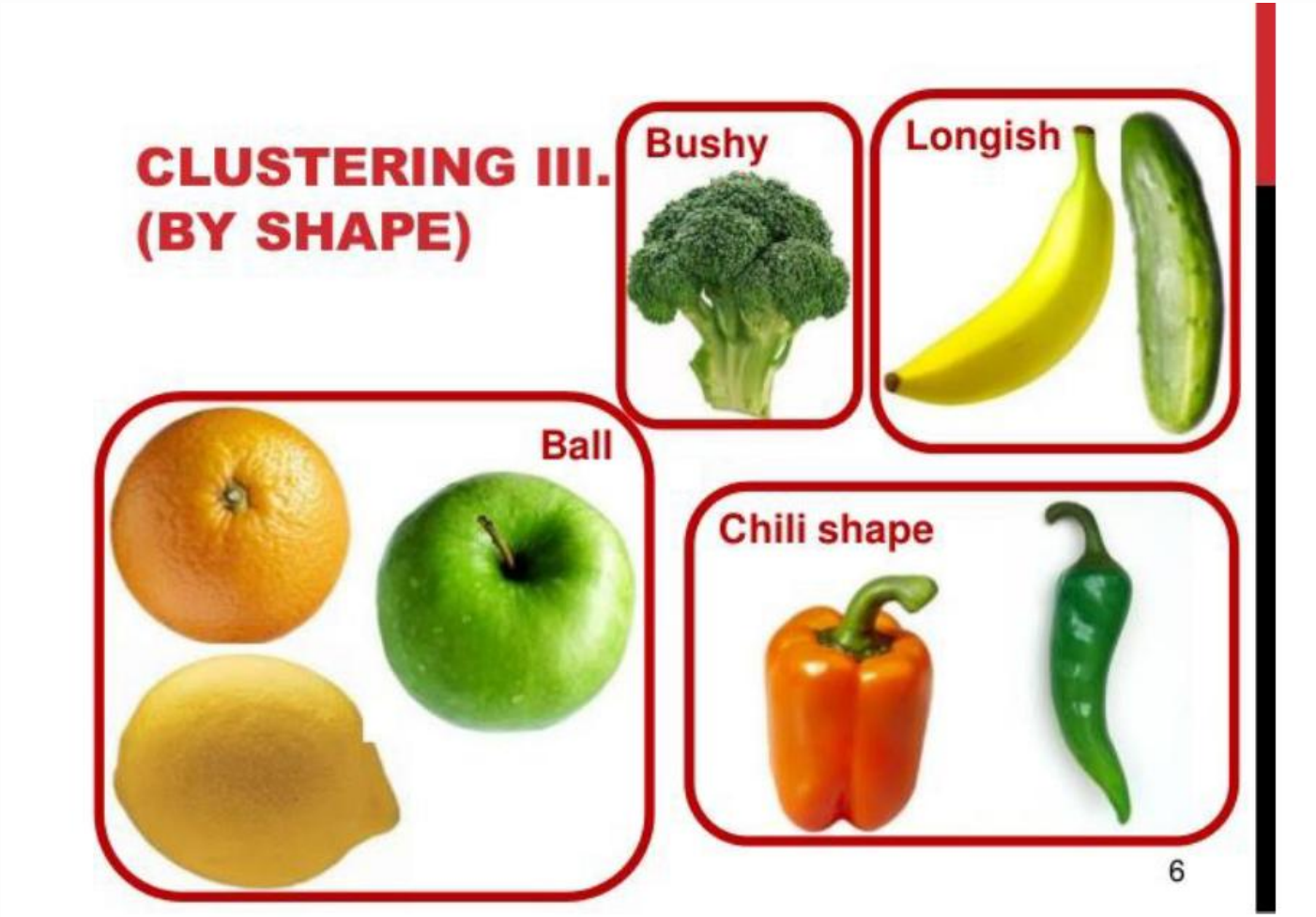
Unsupervised learning- examples



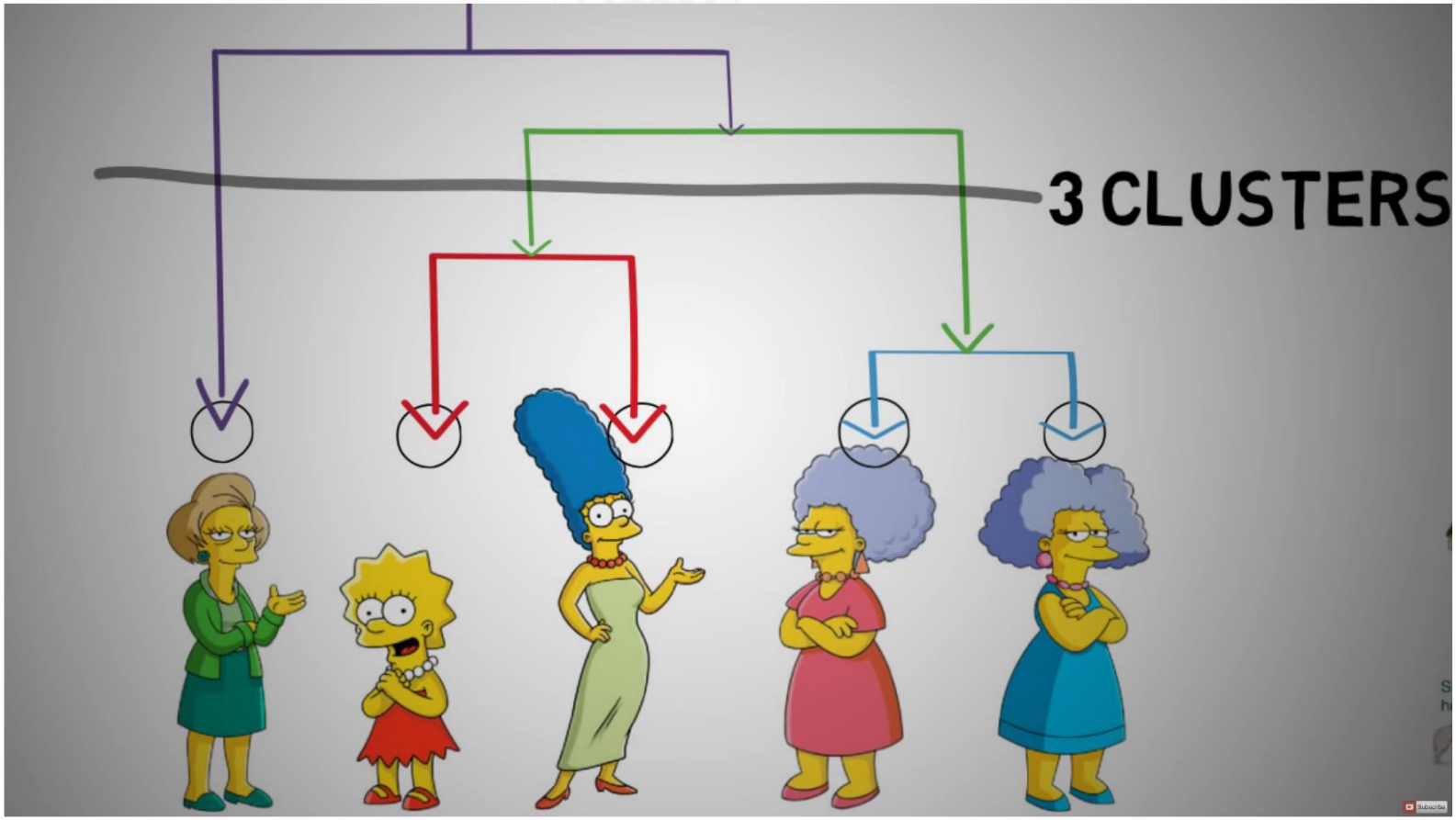
Unsupervised learning- examples



Unsupervised learning- examples



Unsupervised learning- examples



Unsupervised learning

- **"What do we want the system to learn if we do not provide external information?"**
  - Discovering groups (classes) within the initial dataset
  - Extracting features (attributes) that describe the input data in a more compact way
  - Identifying natural similarities (patterns) within the data

- Utility in Classifying Unlabelled Objects
  - Starting from a set of instances, the goal is to find mechanisms to group objects based on similarities, such that objects within the same group have maximum similarity.

Unsupervised learning- approaches

- Data clustering
- Self-organizing maps (SOMs)
- Association rules
- Expectation-Maximization (EM) algorithms
- Independent Component Analysis (ICA)
- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)
- Method of moments

- **Clustering methods** → grouping data
- **Dimensionality reduction** → PCA, ICA, SVD
- **Probabilistic methods** → EM
- **Rule discovery** → association rules


FACULTATEA DE MATEMATICA SI INFORMATICA
ROMÂNIA
BABEŞ-BOLYAI

Clustering

Clustering (Data Grouping)

- Clustering is the process of grouping objects (instances, observations) based on their characteristics.

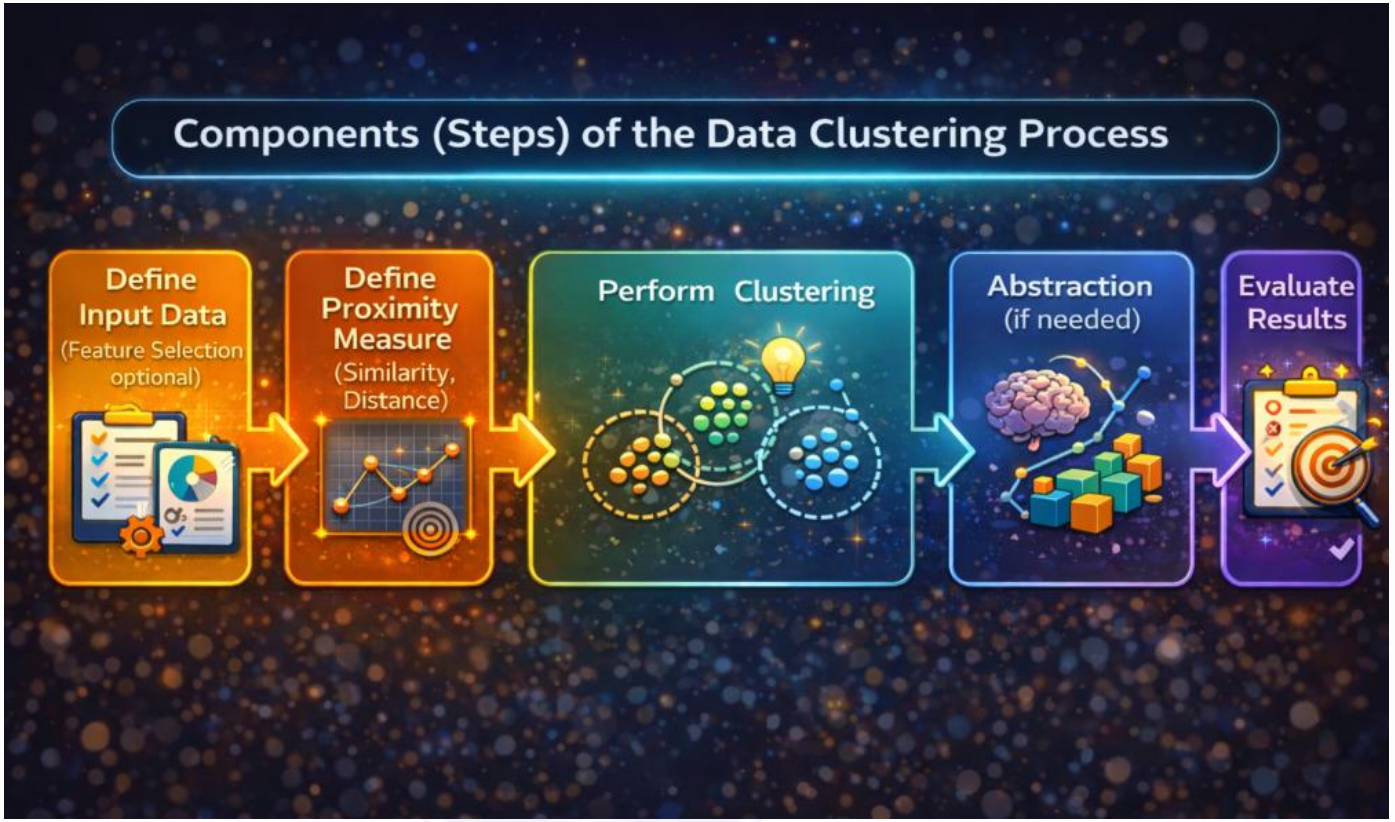
- The goal of this activity is that objects within the same group are similar to each other and different from objects in other groups.

- Homogeneity within a group (called a cluster) and maximum separation between groups characterize this process.

Clustering

- In supervised classification, a collection of labelled data is given, and the problem is to assign labels to a new set of data.
- In clustering, the problem is to group a collection of unlabeled data into meaningful groups.
- To some extent, labels are associated with the groups, but these labels are derived from the initial data itself (they are not provided from an external source); these are called data-driven labels.

Clustering


Components (Steps) of the Data Clustering Process
Define Input Data
(Feature Selection optional)
Define Proximity Measure
(Similarity, Distance)
Perform Clustering
Abstraction
(if needed)
Evaluate Results

Clustering

- Defining the data refers to identifying:
  - the number of classes (clusters)
  - the number and types of attributes in the dataset
  - During this stage, feature selection can also be applied
- A proximity or similarity measure is, in fact, a distance function used to determine how similar or different two objects are.
- Abstraction is the process of extracting a simple and compact representation of the dataset.
- The goal is simplicity (easy to understand), achieved through a concise description of each group of objects, using terms specific to clustering (e.g., medoids, centroids).
- Evaluation of Results
  - External evaluation can be used, which involves comparing the obtained structure with a prior (ground truth) structure
  - Or internal evaluation, which assesses how well the structure fits the intrinsic properties of the original dataset

What is a cluster?

What is a cluster?

So tell me how many clusters do you see?

How many clusters?

Six Clusters

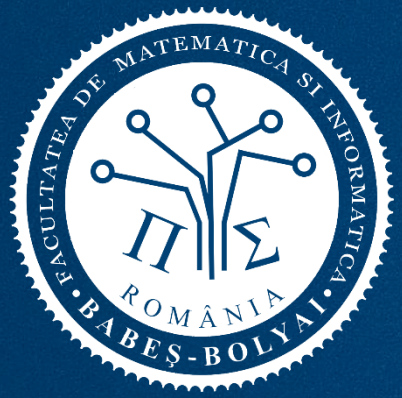
Two Clusters

Four Clusters

What is a cluster?

- Cormack (1971) and Gordon (1999) established that a cluster can be defined based on two internal properties:
  - **homogeneity**
  - **external isolation (separation)**

- Vipin Kumar presents several definitions of a cluster:
  - **Separation-based definition**:
    - A cluster is a set of points where each point is **closer to another point within the set** than to any point outside the set.
  - **Centre-Based Definition of a Cluster**
    - A cluster is a set of objects in which each object is **closer to the centre of its own cluster** than to the centre of another cluster.
  - **Similarity-Based Definition of a Cluster**
    - A cluster is a set of objects that are **like each other**, but **different from objects in other groups**.



Similarity measures

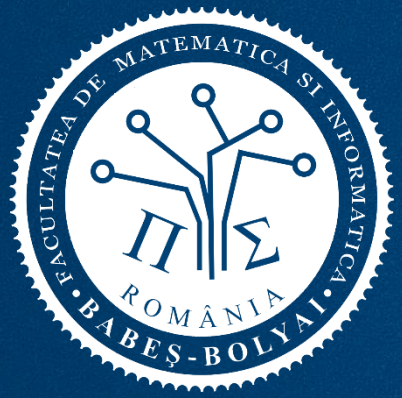
Similarity measure

- A similarity function "measures" how good a particular group is.
- In general, the term used for such measures is **proximity** or **neighbourhood**.
- Two instances are considered "close" when the **dissimilarity (distance) is small** or the **similarity is high**.
- Minimum distance → Maximum similarity
- Euclidean distance: $d(i,j)=\sqrt{\sum_{k=1}^{n}|x_{ik}-x_{jk}|^2}$
- Manhattan distance: $d(i,j)=\sum_{k=1}^{n}|x_{ik}-x_{jk}|.$

Similarity measure

- Cosine similarity $\cos(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|}$
  - Two vectors are:
    - **Very similar** → angle ≈ 0° → cosine ≈ 1
    - **Unrelated** → angle ≈ 90° → cosine ≈ 0
    - **Opposite** → angle ≈ 180° → cosine ≈ -1

- Jaccard similarity: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$
  - Where A, and B are sets



Clustering techniques

Clustering techniques

- Hierarchical Methods
  - Agglomerative algorithms
  - Divisive algorithms
- Partitioning Methods
  - k-medoids
  - k-means
- Density-based algorithms : DBSCAN, OPTICS, Mean Shift
- Algorithms for Large-Scale Data
- Probabilistic algorithms: Gaussian mixture models (GMM)

Clustering techniques

- Anil K. Jain considers that clustering methods can be organized based on several criteria:
  - Agglomerative vs. Divisive→ refers to how groups are formed
  - Hard (crisp) vs. Fuzzy (soft) → refers to how instances are assigned to groups

- Assignment Types
  - Hard assignment
    → an object belongs to only one group throughout the process
  - Fuzzy (soft) assignment
    → each instance has degrees of membership to multiple groups

Hierarchical clustering

Hierarchical clustering

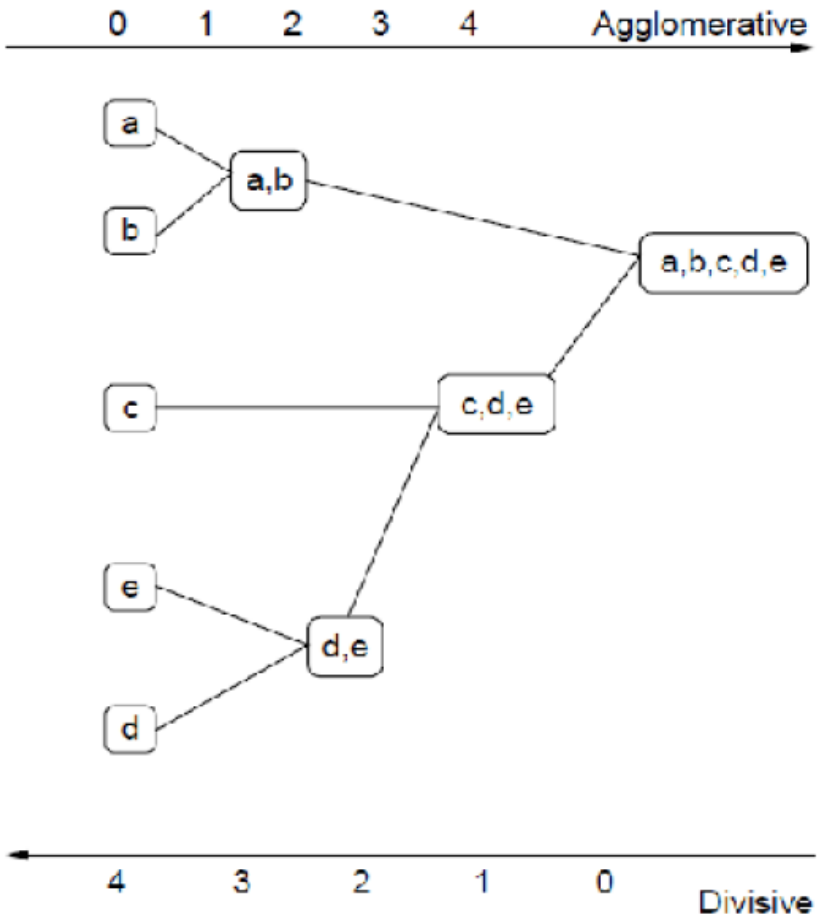
- In hierarchical classification, data is not partitioned into a fixed number of groups in a single step.

- Hierarchical techniques are divided into:
  - Agglomerative methods
  - Divisive methods

- Agglomerative methods
  → perform a series of successive merges, combining *n* instances into groups

- Divisive methods
  → perform successive splits, separating *n* instances into smaller, more precise groups

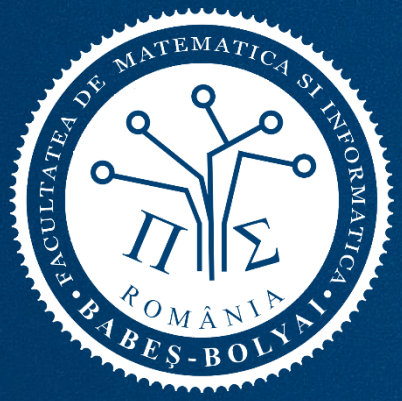
Hierarchical clustering



Hierarchical clustering

- **Agglomerative Approach**
  - Starts with individual objects
  - Gradually merges objects that have maximum similarity (minimum distance)
  - Continues until:
    - all objects are merged into a single group, or
    - the desired number of clusters is reached

- **Divisive Approach**
  - Starts with a single group containing all objects
  - Gradually splits the group into two, assigning objects to new clusters such that similarity within each group is maximized
  - Continues splitting until:
    - each cluster contains a single object, or
    - the desired number of clusters is reached



Partitioning clustering

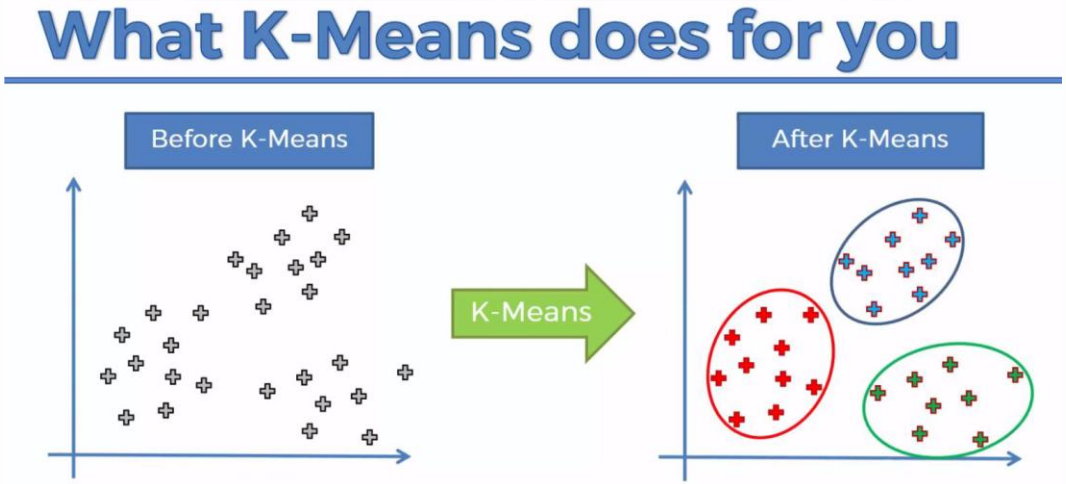
Partitioning clustering

- A partitioning clustering algorithm produces a partition of the dataset, compared to hierarchical clustering which produces a structure.
- The goal of partitioning methods is to construct or find a partition of k clusters from a dataset with *n* instances.
- These techniques include several heuristic methods, such as:
  - k-means
  - k-medoids
- k-means algorithm
  → assumes that each cluster is represented by its centre (centroid)
- k-medoids algorithm
  → assumes that each cluster is represented by one of its actual objects (medoid)

K-Means

- The name comes from representing k clusters based on the mean of the objects within each cluster.

- Such an object, defined as the average of the instances in a cluster, is called a centroid.



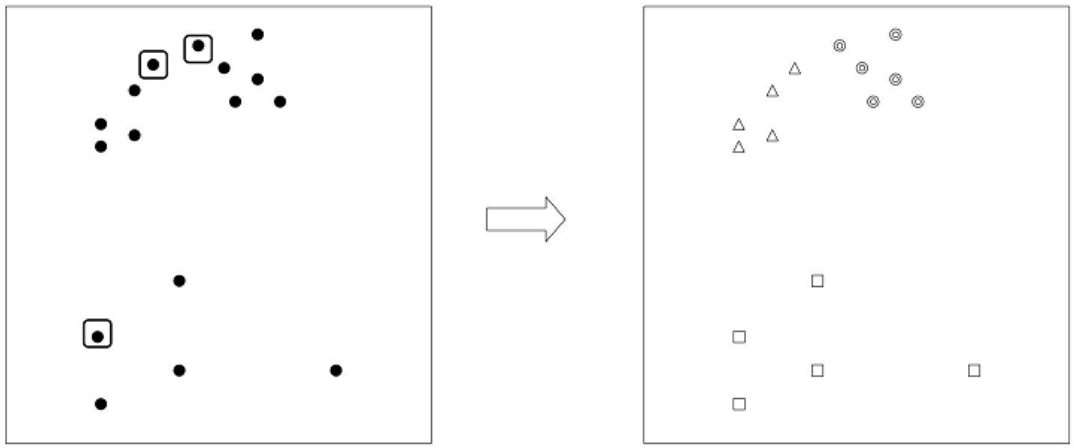
K-Means

- Algorithm :
  - Select k instances as the initial centroids
  - Assign each object to the nearest centroid
  - Recalculate the centroid of each cluster
  - Repeat from step 2 until no more changes occur

- Choosing Initial Centroids
  - This is a key step in the k-means procedure.
  - It is very easy to randomly select *k* objects as centroids, but the results are usually not satisfactory.
  - Another approach is to run the algorithm multiple times, each time with different randomly chosen centroids.

K-Means

- **Natural Choice of Centroids**
  - The initial centroids are often chosen from **dense regions**, so that objects are **well separated**, and two centroids are not selected from the **same group**.



Choosing the Optimal Number of Centroids (K)

- ELBOW METHOD (Most intuitive)
  - We measure how **compact clusters are** using:
    - **WCSS (Within-Cluster Sum of Squares)**
      → distance of points to their centroid
  - **Formula intuition**
    - Small distance → good cluster
    - Large distance → bad cluster

$$\text{WCSS} = \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

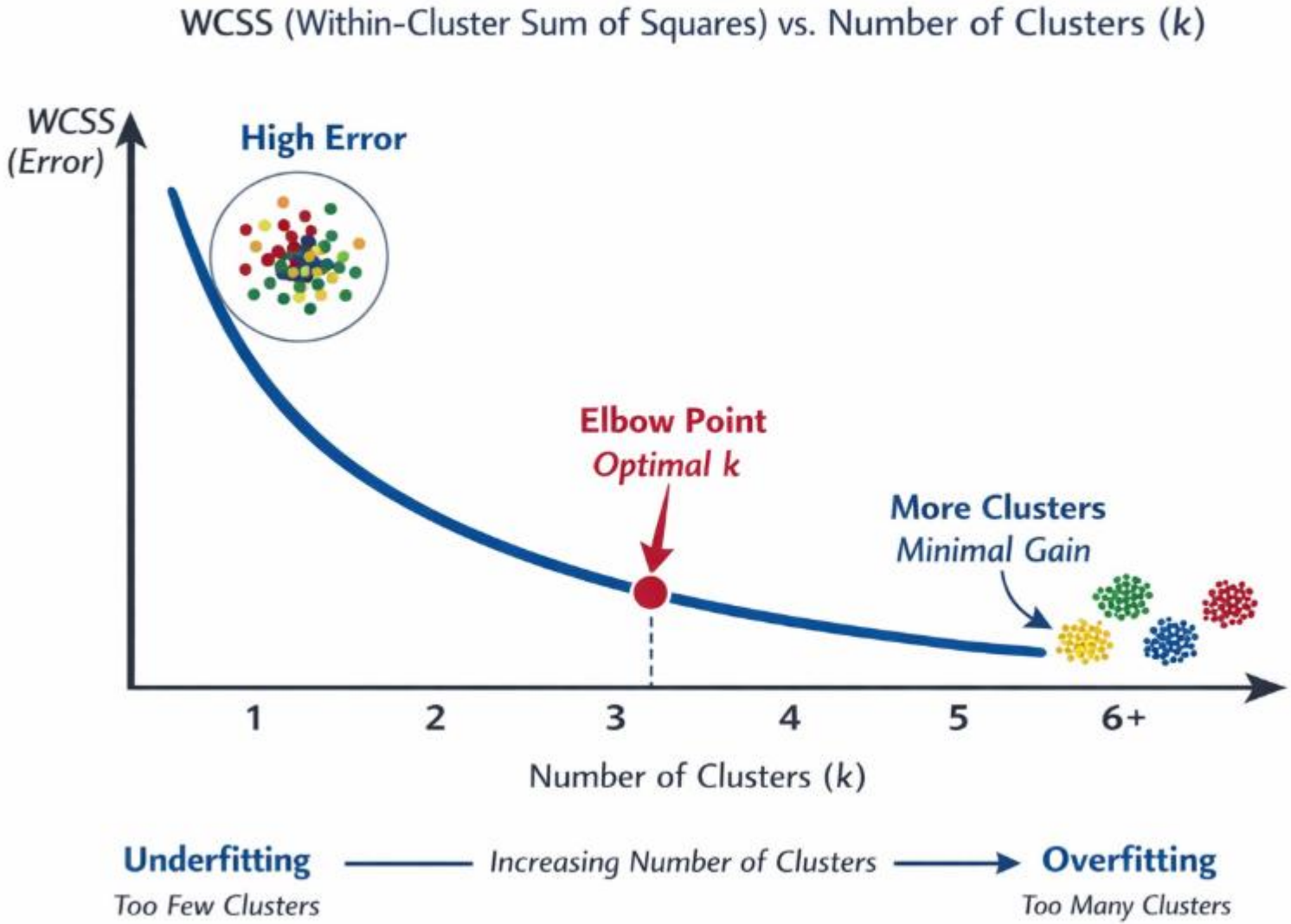
- k= The total number of clusters.
- S_i=The set of data points in cluster i
- X_j=a data point within cluster j
- μ_i =centroid of cluster i

Choosing the Optimal Number of Centroids (K)

ELBOW METHOD (Most intuitive)

- **Process (step-by-step)**
  - Choose a range of K values
    e.g., K = 1 to 10
  - For each K:
    - run K-Means
    - compute WCSS
  - Plot:
    X-axis = K
    Y-axis = WCSS
- K = 1 → very high error
- K increases → error decreases
- At some point → improvement slows
  - This point = **ELBOW**

The Elbow Method



Choosing the Optimal Number of Centroids (K)

- SILHOUETTE METHOD (More robust)
  - Measures: **how well each point fits in its cluster**

$$s(i) = \frac{(b(i) - a(i))}{Max\{a(i), b(i)\}}$$

  - $a(i)$ is the average distance between object i and the other objects in its own cluster
  - $b(i)$ is the minimum average distance between object i and the objects in the other clusters (clusters that i does not belong to)

  - Values between -1 and 1
  - +1 → perfect clustering
  - 0 → overlapping clusters
  - -1 → wrong assignment

Choosing the Optimal Number of Centroids (K)

- SILHOUETTE METHOD (More robust)
  - Process
  - Try multiple K values
  - Compute average silhouette score
  - Choose K with highest score

- Use Elbow:
  - quick estimation
  - large datasets

- Use Silhouette:
  - need better accuracy
  - smaller datasets

K-Means -Example

- Given the points:
  $P_1(2,3),\ P_2(3,1),\ P_3(4,2),\ P_4(11,5),\ P_5(12,4),\ P_6(12,6),$
  $P_7(7,5),\ P_8(8,4),\ P_9(8,6)$

- Apply **K-Means** starting with initial centroids:
  $K_1 = P_2$ and $K_2 = P_8$

- Use **Euclidean distance**

K-Means - Example

- Iteration 1: assign each point to the nearest centroid
  - **Distances to** $K_1 = (3, 1)$ **and** $K_2 = (8, 4)$

- $d(\mathrm{P1}, \mathrm{K1}) = \sqrt{(2-3)^2 + (3-1)^2} = \sqrt{1+4} = \sqrt{5} \approx 2.23$
- $d(\mathrm{P1}, \mathrm{K2}) = \sqrt{(2-8)^2 + (3-4)^2} = \sqrt{36+1} = \sqrt{37} \approx 6.08$

$\rightarrow P_1 \in C_1$

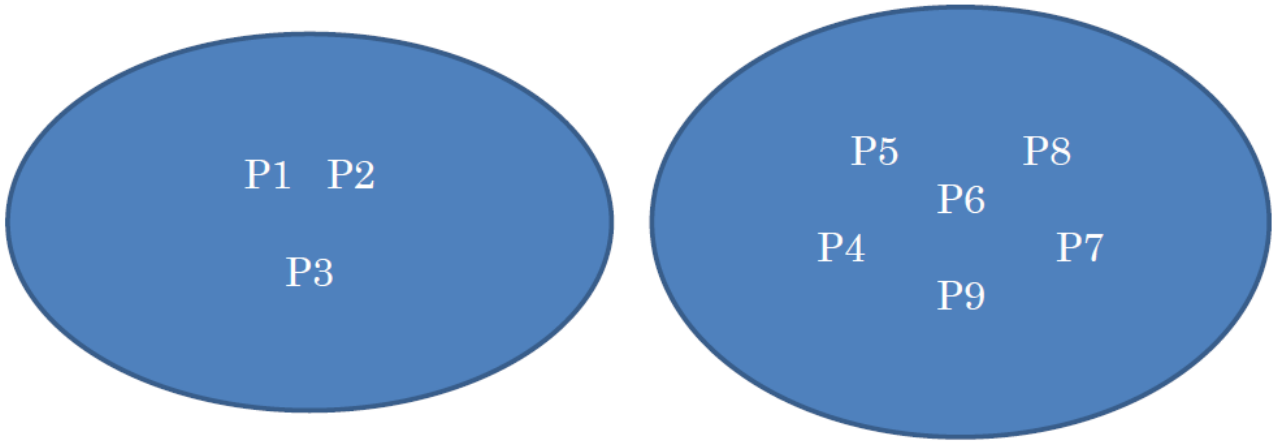
- $d(P_3, K_1) = \sqrt{(4-3)^2 + (2-1)^2} = \sqrt{2} \approx 1.41$
- $d(P_3, K_2) = \sqrt{(4-8)^2 + (2-4)^2} = \sqrt{20} \approx 4.47$

$\rightarrow P_3 \in C_1$

.....

K-Means - Example

- $P_4 \in C_2$
- $P_5 \in C_2$
- $P_6 \in C_2$
- $P_7 \in C_2$
- $P_7 \in C_2$
- $P_8 \in C_2$



K-Means - Example

- Recompute centroids
  - New centroid for C_1
  - Points: $P_1(2, 3),\ P_2(3, 1),\ P_3(4, 2)$

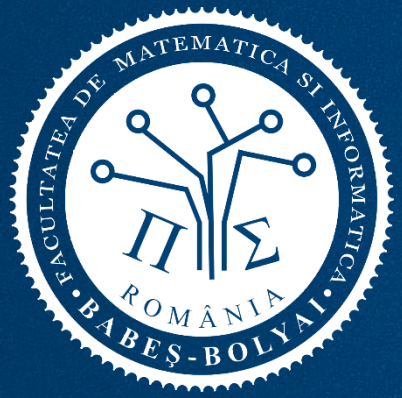
$$K_1' = \left(\frac{2+3+4}{3}, \frac{3+1+2}{3}\right) = \left(\frac{9}{3}, \frac{6}{3}\right) = (3, 2)$$

  - New centroid for C_2
  - Points: $P_4(11, 5),\ P_5(12, 4),\ P_6(12, 6),\ P_7(7, 5),\ P_8(8, 4),\ P_9(8, 6)$
    - $K_2' = \left(\frac{58}{6}, \frac{30}{6}\right) = (9.67, 5)$

- Iteration 2: reassign points
- Now use:
  - $K_1 = (3, 2), K_2 = (9.67, 5)$
  - If we recompute the assignments, all points remain in the same clusters:
  - $C_1 = \{P_1, P_2, P_3\}$
  - $C_2 = \{P_4, P_5, P_6, P_7, P_8, P_9\}$
- So the algorithm has converged.

K-Medoids

1. Select **k representative objects**, called **medoids**
2. Replace one of the selected medoids with a **non-selected object**
   - Compute the distance between each non-selected object and the **closest candidate medoid**
   - Sum all these distances
   - This total represents the **cost of the current configuration**
3. Select the configuration with the **lowest cost**
   - If a better configuration is found, repeat step 2
4. If no better configuration exists:
   - Assign each non-selected object to the **nearest medoid**
   - Stop the algorithm



Clustering evaluation

Clustering evaluation

- **Why Evaluate the Clustering Process?**
  - To **compare different clustering algorithms**
  - To **compare two groups (clusters)**
  - To **compare sets of clusters**
- **Aspects of Clustering Validation**
  - Determining the **optimal number of clusters**
  - Finding **relationships between the discovered structures** in the dataset and **external information** related to the input data
  - Evaluating how well the clustering results **fit the dataset itself**, without using external information (i.e., using only the data)
- We evaluate clustering to understand if the groups are meaningful, stable, and useful — even without ground truth.

Clustering evaluation

- **Clustering Evaluation Approaches**
  - **External evaluation**
    → based on prior information about the data (e.g., known labels)
  - **Internal evaluation**
    → based only on intrinsic information (the dataset itself), without using external sources

Internal evaluation

- Dunn index
  - The Dunn Index determines the **ratio between the smallest inter-cluster distance** and the **largest intra-cluster distance** within a partition.

$$D = \frac{\min_{1 \le i < j \le n} d(i,j)}{\max_{1 \le k \le n} d'(k)},$$

  - $d(i, j)$represents the **inter-cluster distance** (distance between cluster $i$and cluster j)
  - d′(k) represents the **intra-cluster distance** (distance within cluster k)
  - The distance between two clusters can be defined using any metric, for example, the **distance between the centroids of the clusters**
  - The intra-cluster distance $d'(k)$can be defined as the **maximum distance between any pair of elements within cluster** k

- **Higher Dunn Index = better clustering**
  - large separation between clusters
  - small spread within clusters

Internal evaluation

- Silhouette score
- Inertia (WCSS-Within-Cluster Sum of Squares)
- Davies-Bouldin index:

$$BD = \frac{1}{c}\sum_{i=1}^{c} Max_{i \neq j} \left\{ \frac{d(X_i) + d(X_j)}{d(c_i, c_j)} \right\}$$

  - c represents the number of clusters
  - $d(X_i)$and $d(X_j)$ represent the distances between all objects in clusters X_i and X_j and their respective centroids
  - $d(c_i,c_j)$ is the distance between the centroids of clusters $c_i$and c_j
  - lower values of this index indicate better clustering solutions

Internal evaluation

- Calinski-Harabasz Index (Variance Ratio Criterion)
  - It measures the ratio between:
    - Between-cluster dispersion (how far clusters are from each other)
    - Within-cluster dispersion (how compact clusters are)

$$CH = \frac{\text{trace}(B_k)/(k-1)}{\text{trace}(W_k)/(n-k)}$$

  - B_k →Between-cluster dispersion
    - Measures how far cluster centroids are from the global mean
    - Higher = clusters are well separated
  - W_k →Within-cluster dispersion
    - Measures how far points are from their own centroid
    - Lower = clusters are compact
  - k = number of clusters
  - n = number of data points
  - trace (sum of diagonal elements → total variance)
  - A higher score indicates better clustering.

External evaluation

- Accuracy
- Precision
- Recall
- Specificity
- Purity
- …

Summary of evaluation measures

- Inertia: Lower is better but always evaluate in relative terms.
  - Compare inertia between different K values
- Silhouette Score: Closer to +1 is better.
- Davies-Bouldin Index: Lower is better.
- Dunn index: higher is letter.
- Calinski-Harabasz Index: Higher is better.

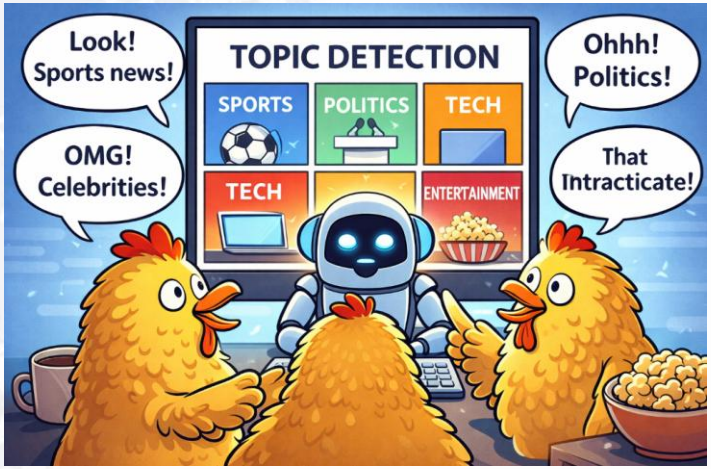
Applications of clustering

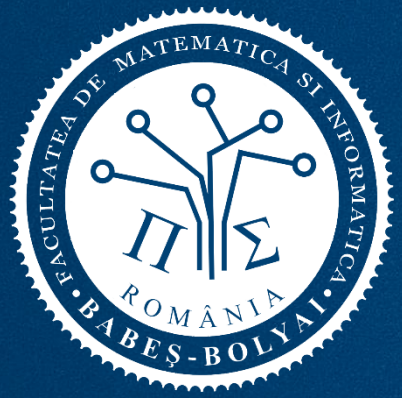
Applications of clustering

- Marketing and Commerce
  - **Segmentation** = the process of organizing customers into groups based on:
    - preferences for certain products
    - characteristics
    - expectations

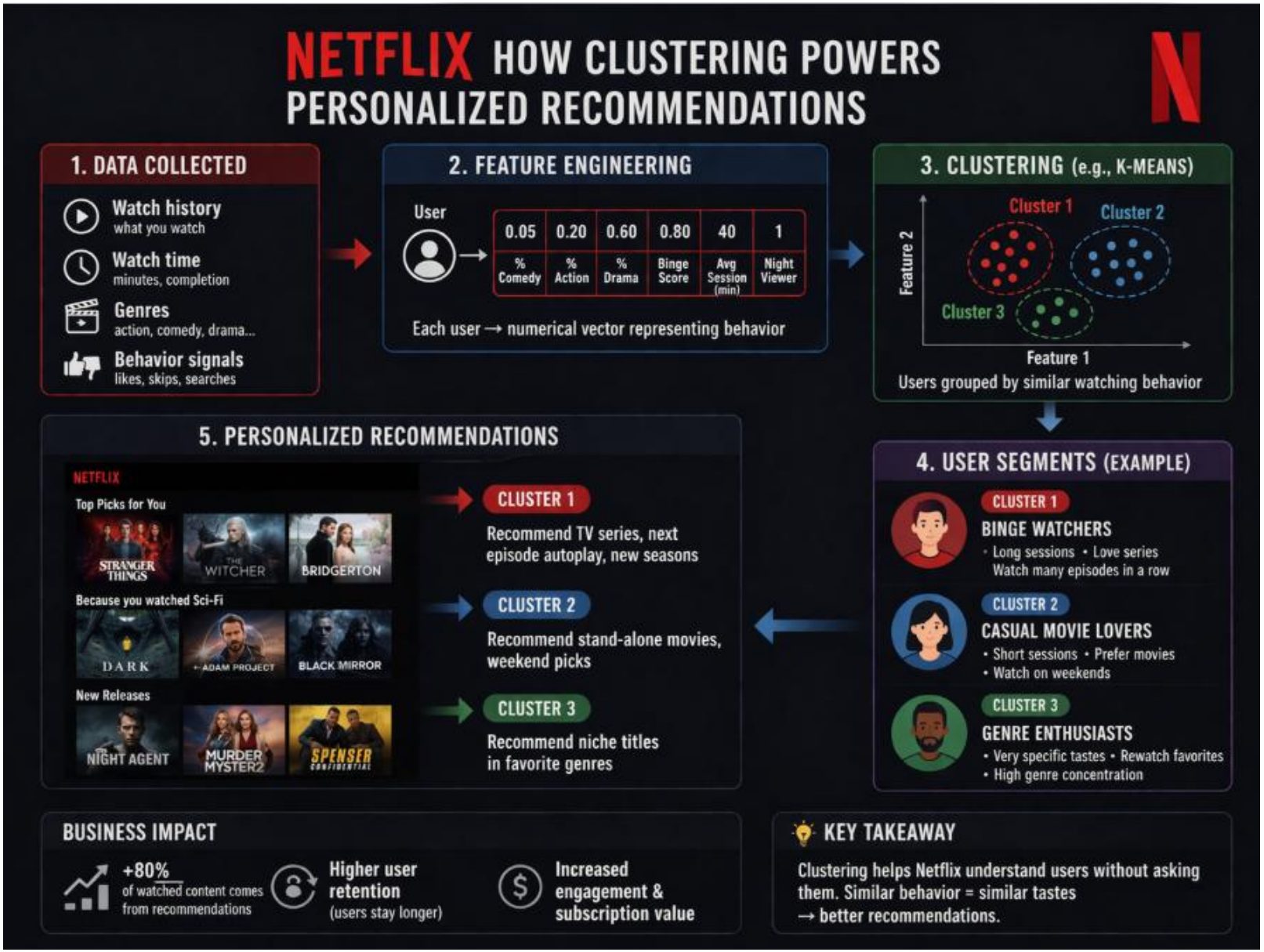
Applications of clustering

- Bioarchaeology
- Recommendation systems
- Economy
- Healthcare
  - Group patients by symptoms
  - Analyse gene expression
- Social media
  - news grouping
  - Topic detection



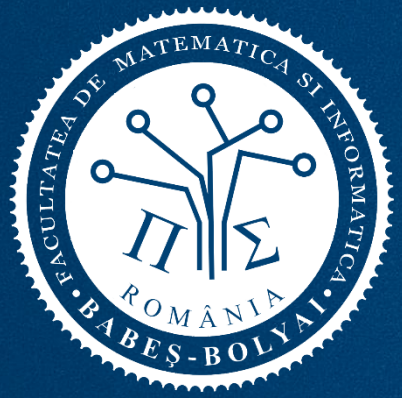


Industry case studies


NETFLIX HOW CLUSTERING POWERS PERSONALIZED RECOMMENDATIONS
1. DATA COLLECTED
Watch history
what you watch
Watch time
minutes, completion
Genres
action, comedy, drama...
Behavior signals
likes, skips, searches
2. FEATURE ENGINEERING
User
0.05 % Comedy
0.20 % Action
0.60 % Drama
0.80 Binge Score
40 Avg Session (min)
1 Night Viewer
Each user → numerical vector representing behavior
3. CLUSTERING (e.g., K-MEANS)
Cluster 1
Cluster 2
Cluster 3
Feature 2
Feature 1
Users grouped by similar watching behavior
4. USER SEGMENTS (EXAMPLE)
CLUSTER 1
BINGE WATCHERS
• Long sessions • Love series
Watch many episodes in a row
CLUSTER 2
CASUAL MOVIE LOVERS
• Short sessions • Prefer movies
• Watch on weekends
CLUSTER 3
GENRE ENTHUSIASTS
• Very specific tastes • Rewatch favorites
• High genre concentration
5. PERSONALIZED RECOMMENDATIONS
NETFLIX
Top Picks for You
STRANGER THINGS
THE WITCHER
BRIDGERTON
Because you watched Sci-Fi
DARK
THE ADAM PROJECT
BLACK MIRROR
New Releases
THE NIGHT AGENT
MURDER MYSTERY 2
SPENSER CONFIDENTIAL
CLUSTER 1
Recommend TV series, next episode autoplay, new seasons
CLUSTER 2
Recommend stand-alone movies, weekend picks
CLUSTER 3
Recommend niche titles in favorite genres
BUSINESS IMPACT
+80% of watched content comes from recommendations
Higher user retention (users stay longer)
Increased engagement & subscription value
KEY TAKEAWAY
Clustering helps Netflix understand users without asking them. Similar behavior = similar tastes
→ better recommendations.


NETFLIX
How Netflix Recommends Shows
Clustering Users by Viewing Habits
Binge Watchers
Loves TV Series
Casual Viewers
Prefers Movies
Horror Fans
Watches Scary Films
Popular TV Shows
Top Movies
Horror Movies
STRANGER THINGS
Recommented For You!
Suggested For You!
Just For You!
AI Algorithm
Analyzing Your Viewing History


Spotify
HOW SPOTIFY USES CLUSTERING TO PERSONALIZE MUSIC EXPERIENCES
Turning billions of listening behaviors into perfectly tailored recommendations
THE CHALLENGE
Deliver the right music to the right listener from 100+ million tracks to 600+ million users.
THE DATA SPOTIFY COLLECTS
Listening history (plays, skips, repeats)
Listening time & context (time of day, device)
Liked & disliked songs
Artist & genre interactions
Playlist creation & follows
Search & discovery behavior
FEATURE ENGINEERING
Raw behavior is transformed into meaningful features:
Genre affinity (e.g., % pop, % rock)
Energy level (acoustic, danceability)
Listening recency & frequency
Session duration & skipping rate
THE CLUSTERING PROCESS
Spotify uses advanced machine learning and clustering (e.g., K-Means, Hierarchical, Gaussian Mixture Models) to group users with similar listening behaviors.
1. USER VECTORIZATION
Convert users into numerical behavior vectors
2. CLUSTERING
Group users based on similar patterns
3. VALIDATION
Evaluate clusters with silhouette score & business relevance
EXAMPLE LISTENER CLUSTERS
POP ENTHUSIASTS
High energy
Love catchy pop, follow charts, many skips
STRATEGY: Deliver new releases, top hits playlists
INDIE EXPLORERS
Curious & diverse
Deep catalog listeners, niche artists, low skips
STRATEGY: Discover Weekly, indie playlists, concert suggestions
WORKOUT BOOSTERS
High tempo
Listen to hype music during active hours, often on mobile
STRATEGY: Workout mixes, high-energy personalized playlists
CHILL SEEKERS
Calm & relaxing
Prefer acoustic, lo-fi, sleep sounds, nighttime listeners
STRATEGY: Sleep, Lo-Fi, Focus mode, calm mixes
Good evening
Music
Podcasts & Shows
Discover Weekly
Chill Mix
Pop Rising
Release Radar
Made for you
Your Daily Mix
Discover Weekly
The Indie Explorer
Arlo Waves
Home
Search
Your Library
BUSINESS IMPACT
+35% higher user engagement
+30% increase in listening time
+25% improvement in user retention
Personalization at scale → stronger loyalty, more listening, higher subscription value.
KEY CHALLENGES
Dynamic tastes
Users' preferences change constantly
Cold start problem
New users have little to no listening history
Diversity vs. relevance
Balancing familiarity with new discoveries
KEY TAKEAWAY
Spotify uses clustering to understand listener behavior without labels, powering hyper-personalized experiences that make every user feel like "Spotify just gets me."



Key Takeaways

Key Takeaways

- Clustering represents an important branch of unsupervised learning, with significant applications in many domains
- Clustering greatly simplifies manual work, which can sometimes be difficult and time-consuming
- unsupervised = discovery
- clustering = grouping

Thank you for your attention — questions, thoughts, or challenges?

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
BABEȘ-BOLYAI UNIVERSITY

1 Mihail Kogălniceanu Street,
Cluj-Napoca, Cluj, România

www.cs.ubbcluj.ro