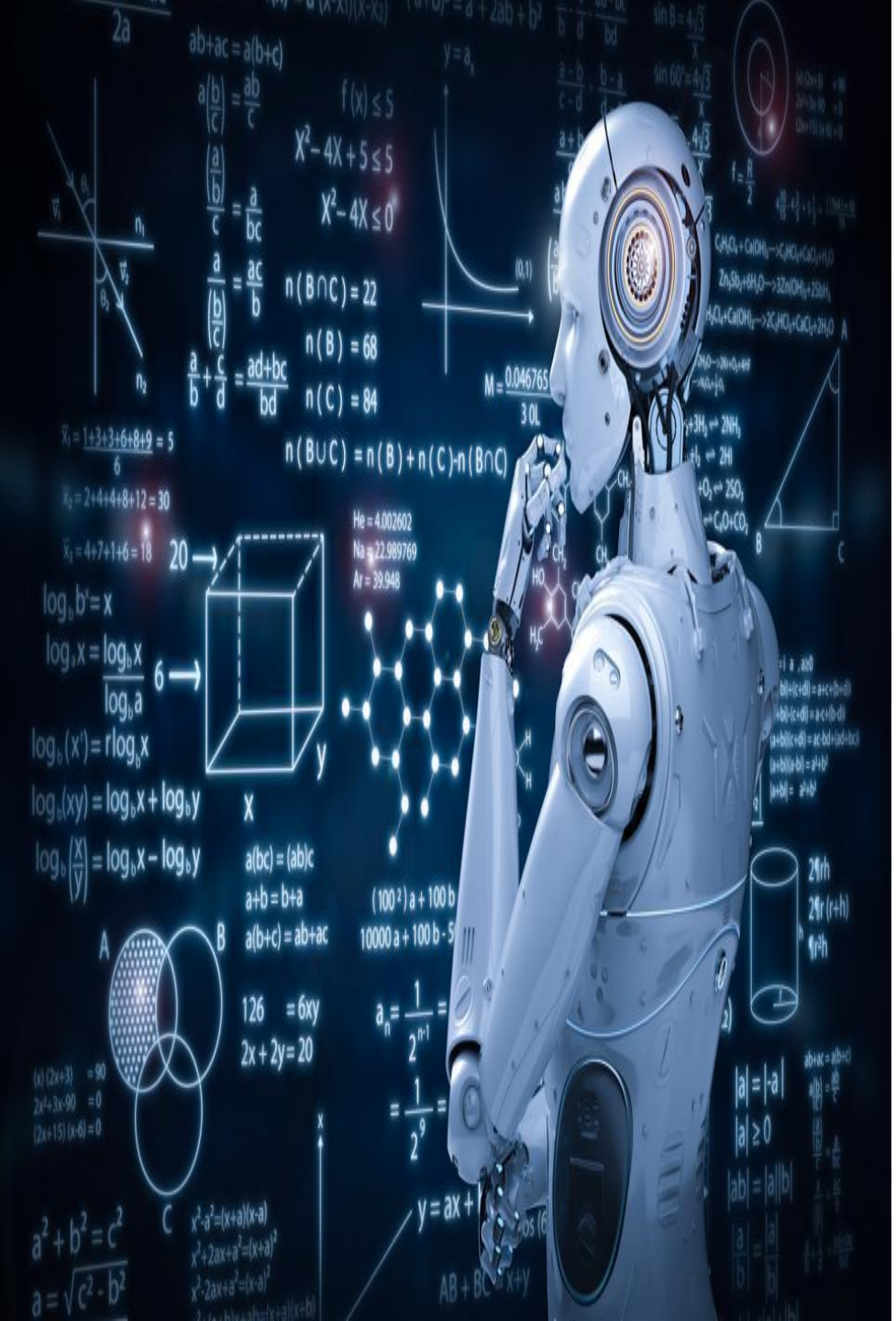


Practical Machine Learning

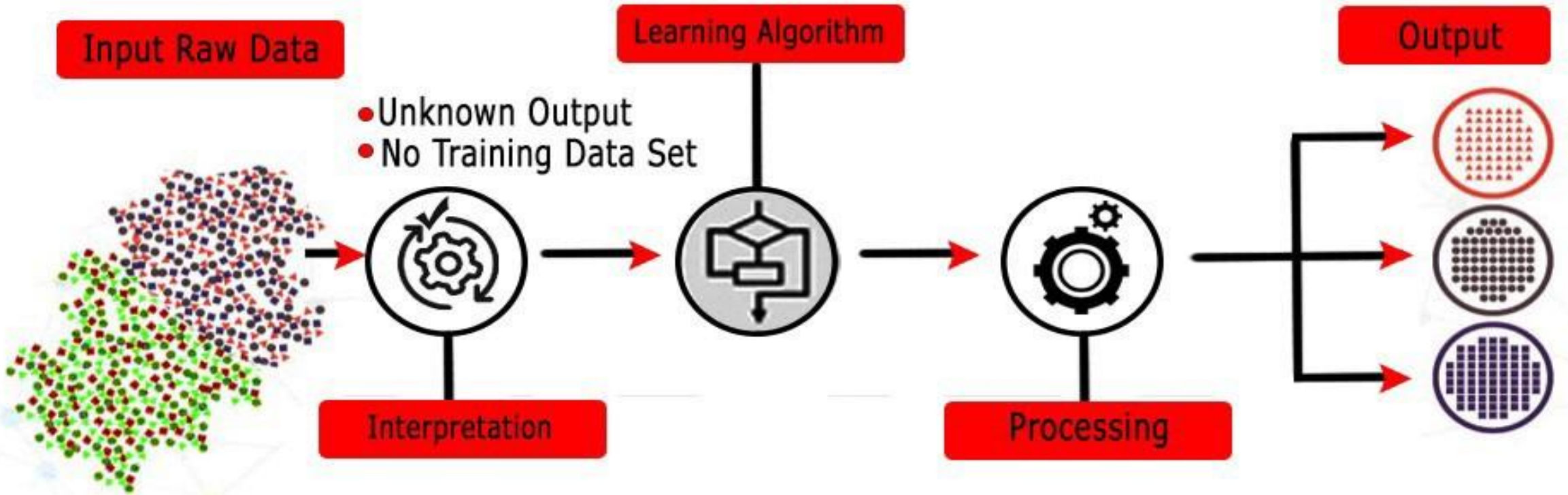
Day 14: Mar23 DBDA

Kiran Waghmare

Agenda

- Clustering
- K-Means
- Hierarchical
- DB-SCAN

Unsupervised Learning



Clustering

Clustering:

- Unsupervised learning
- Requires data, but no labels
- Detect patterns e.g. in
  - Group emails or search results
  - Customer shopping patterns
  - Regions of images
- Useful when don't know what you're looking for
- But: can get gibberish

Machine Learning: Clustering

By color

By shape

By size

etc...

Cluster by Type

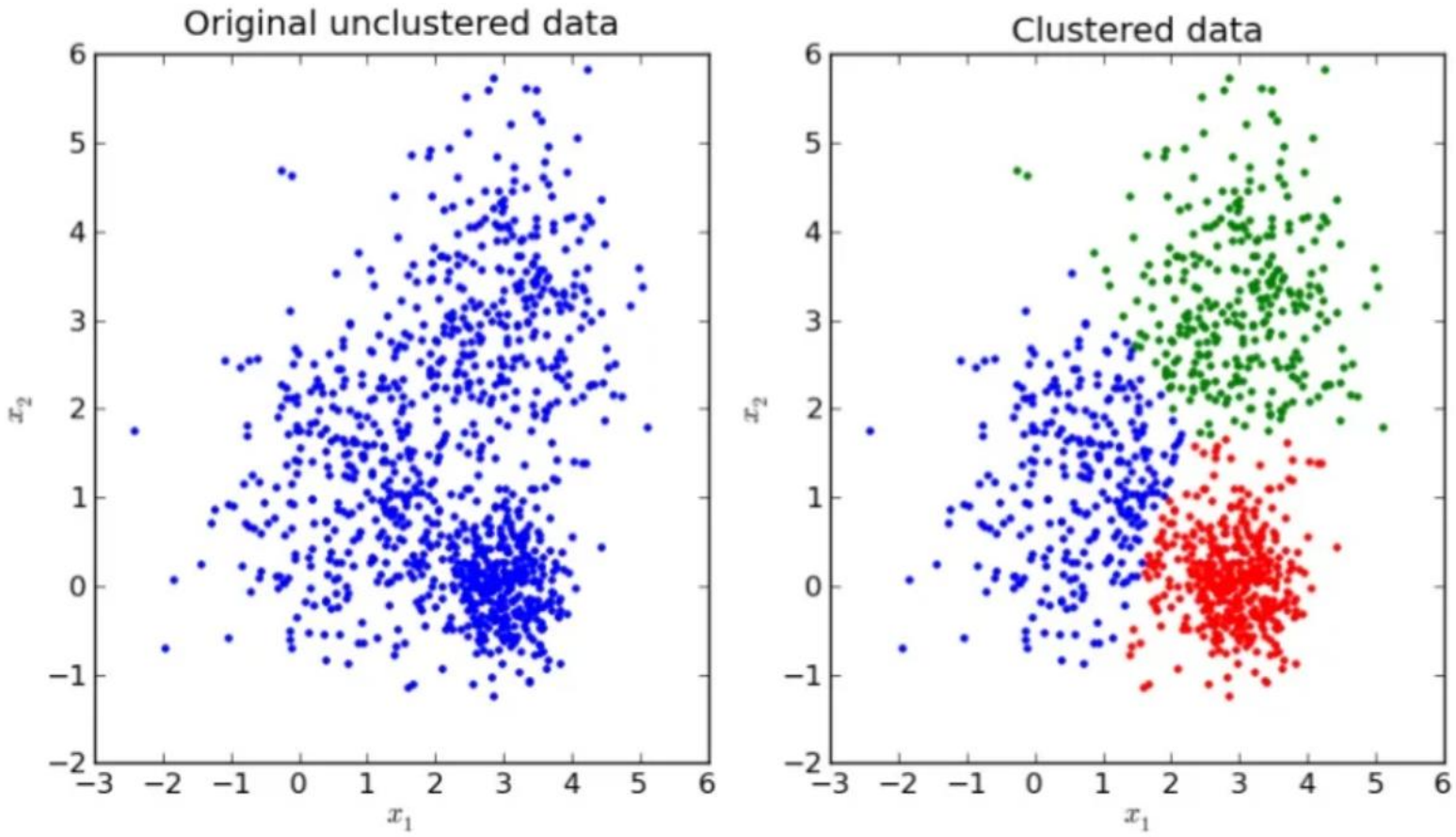
Clustering:

- Clustering is the task of gathering samples into groups of similar samples according to some predefined similarity or dissimilarity measure

sample

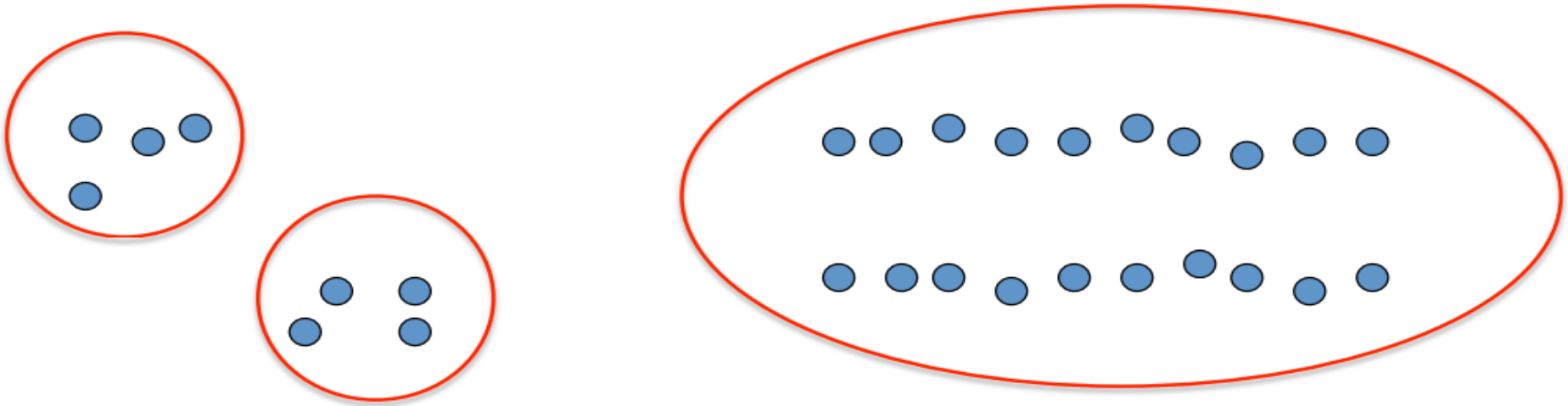
Cluster/group

- In this case clustering is carried out using the Euclidean distance as a measure.



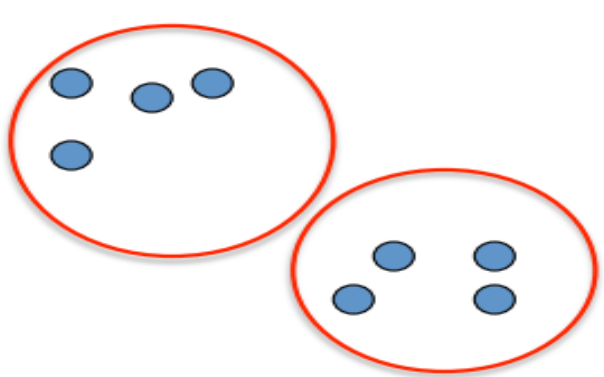
Clustering

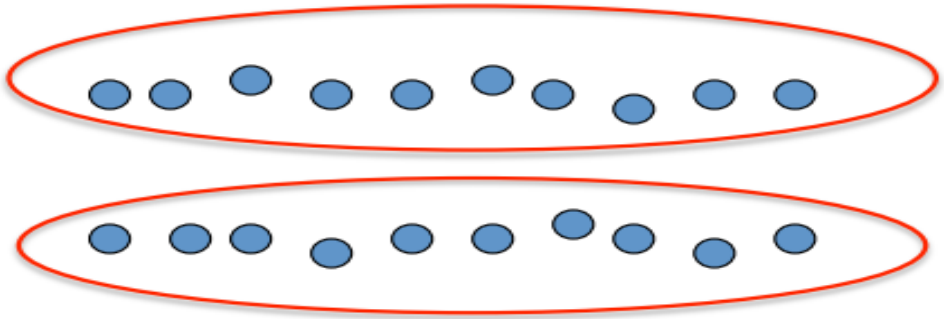
- Basic idea: group together similar instances
- Example: 2D point patterns



Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns





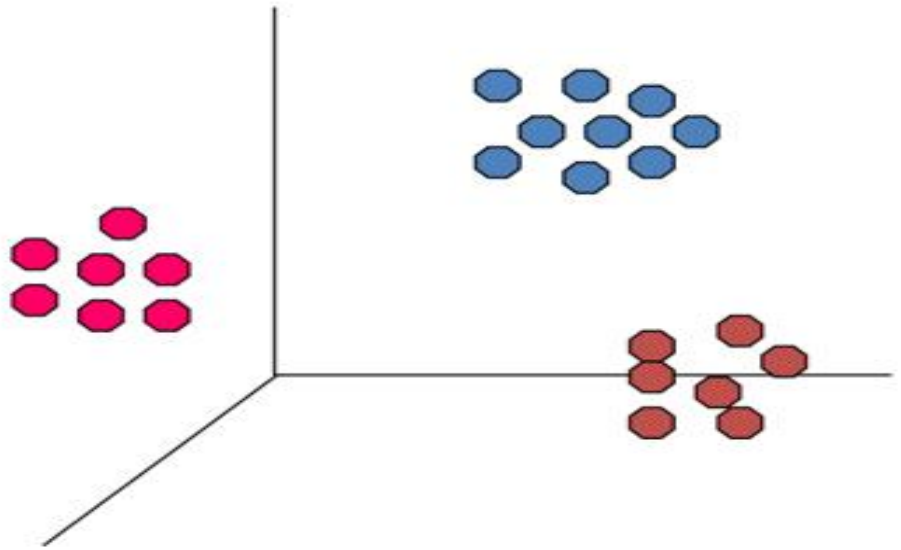
- What could "similar" mean?
  - One option: small Euclidean distance (squared)

$$\text{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$

  - Clustering results are crucially dependent on the measure of similarity (or distance) between "points" to be clustered

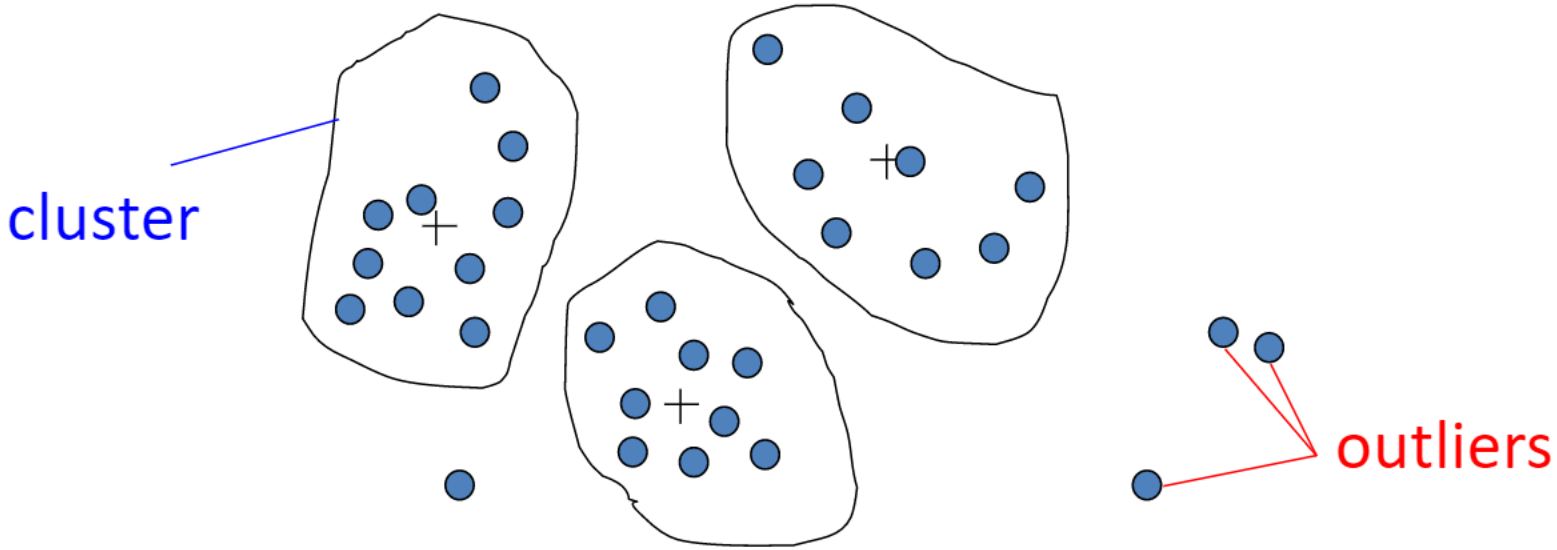
What is clustering?

- A **grouping** of data objects such that the objects **within a group are similar** (or related) to one another **and different from (or unrelated to) the objects in other groups**



Outliers

- **Outliers** are **objects that do not belong to any cluster** or form clusters of very small cardinality



- In some applications we are interested in discovering outliers, not clusters (outlier analysis)

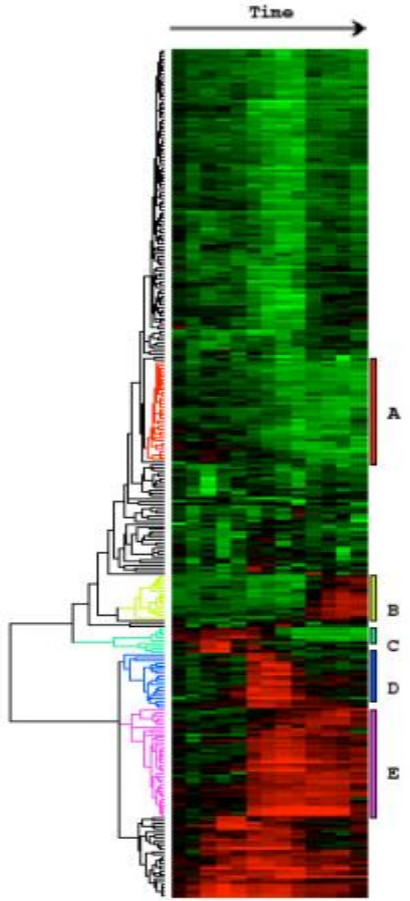
Clustering examples

Image segmentation

Goal: Break up the image into meaningful or perceptually similar regions

Clustering examples

Clustering gene expression data



Eisen et al, PNAS 1998

Feature Extraction → **Graph Contruction** → **Graph Embedding** → **Clustering**

- i) L2-norm
- ii) PCA
  - w/ optional ICA BSS
- iii) UFL
  - RICA
  - SFT

- i) Similarity graph
  - scaling factor
  - similarities
  - kNN criterion

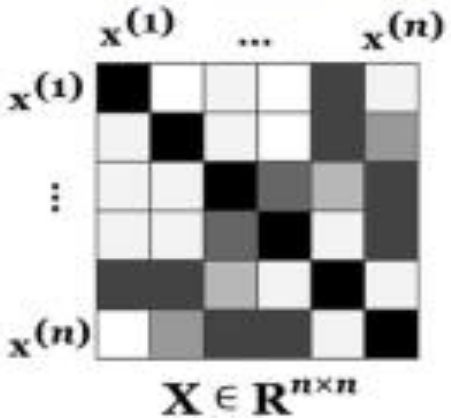
- i) Eigendecomp.
- ii) GNMF
- iii) ICA
  - used as BSS

- i) K-means
  - check clusters
  - NMI, ACC

SFT features

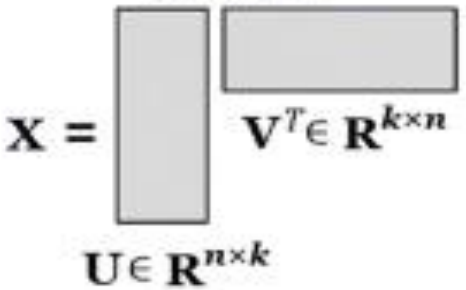
→

Similarity matrix



→

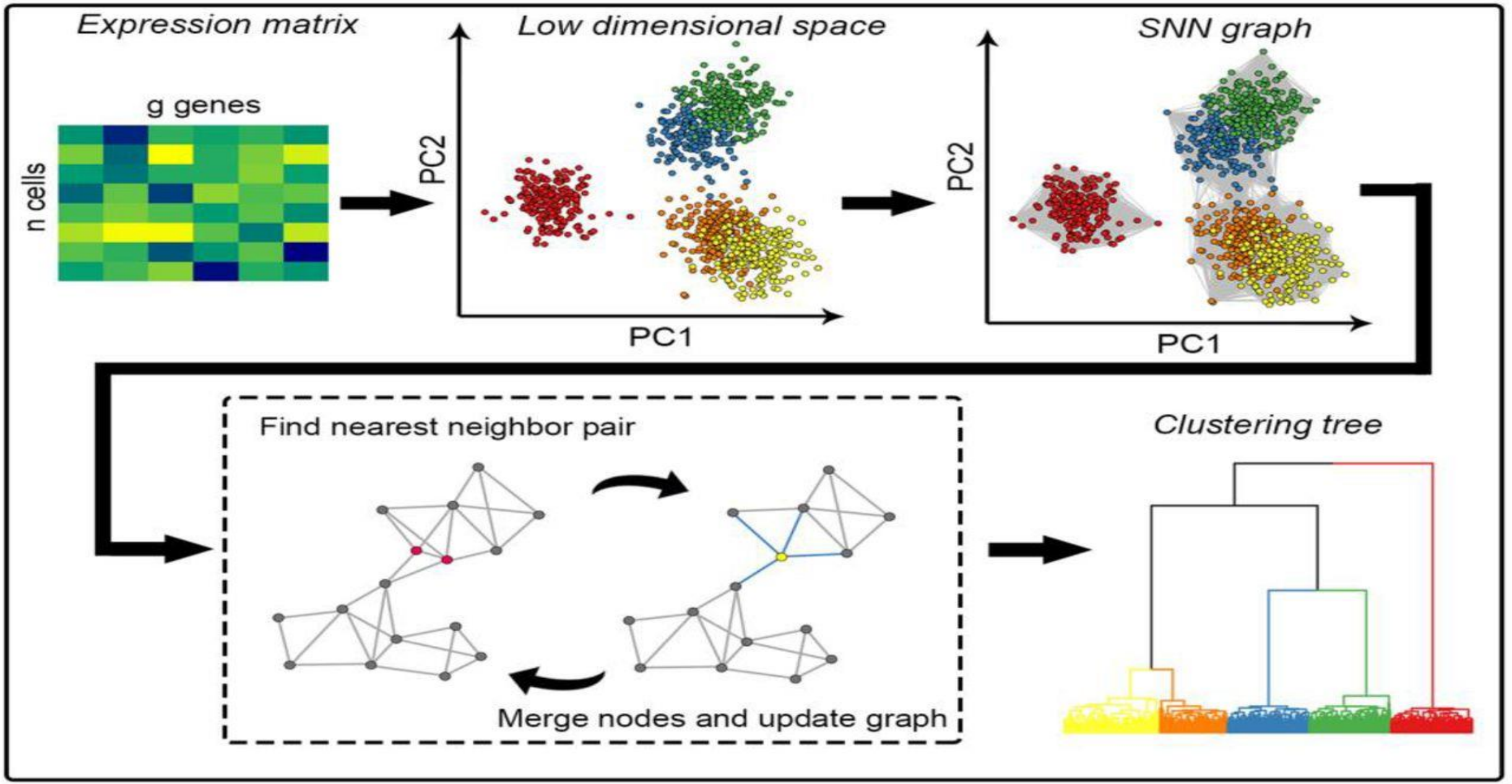
GNMF



→

Clusters

Workflow of HGC



Applications 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
- Bioinformatics
  - Cluster similar proteins together (similarity wrt chemical structure and/or functionality etc)
- Many more...

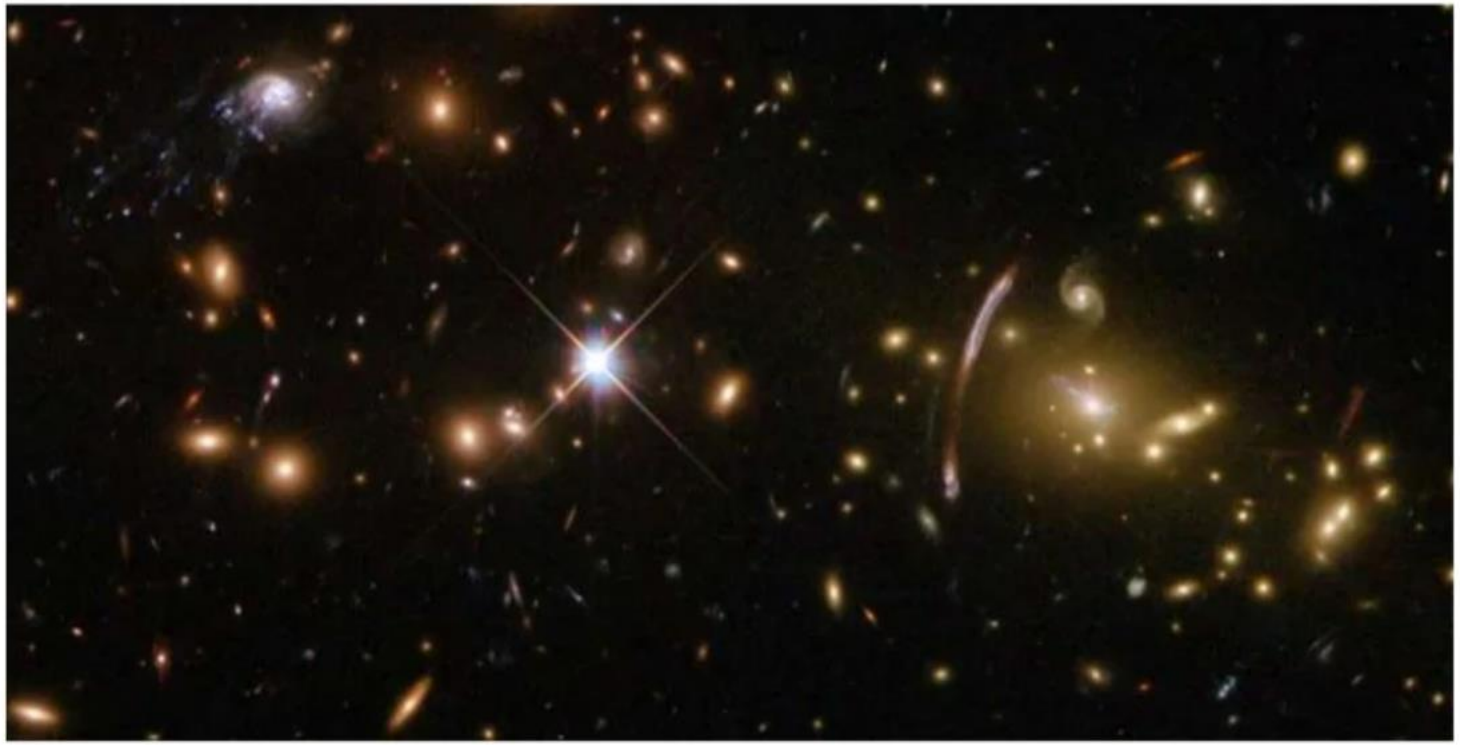
Clustering:

- What is clustering good for
  - **Market segmentation** - group customers into different market segments
  - **Social network analysis** - Facebook "smartlists"
  - **Organizing computer clusters** and data centers for network layout and location
  - **Astronomical data analysis** - Understanding galaxy formation

Clustering:

- What is clustering good for
  - **Market segmentation** - group customers into different market segments
  - **Social network analysis** - Facebook "smartlists"
  - **Organizing computer clusters** and data centers for network layout and location
  - **Astronomical data analysis** - Understanding galaxy formation

Galaxy Clustering:

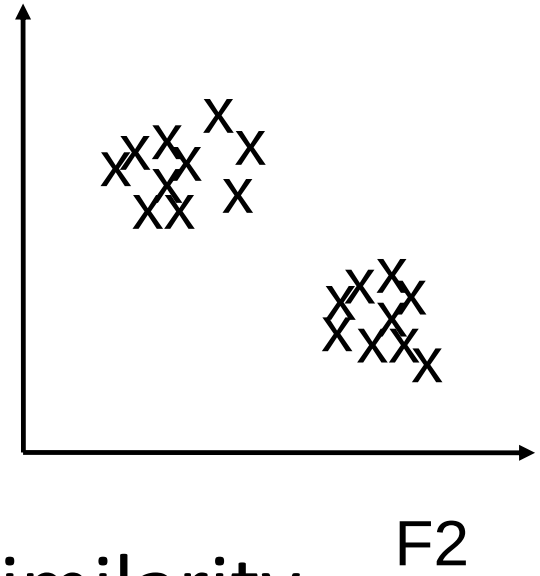


Credit: HST

- Multi-wavelength data obtained for galaxy clusters
  - **Aim:** determine robust criteria for the inclusion of a galaxy into a cluster galaxy
  - **Note:** physical parameters of the galaxy cluster can be heavily influenced by wrong candidate

Goal of Clustering

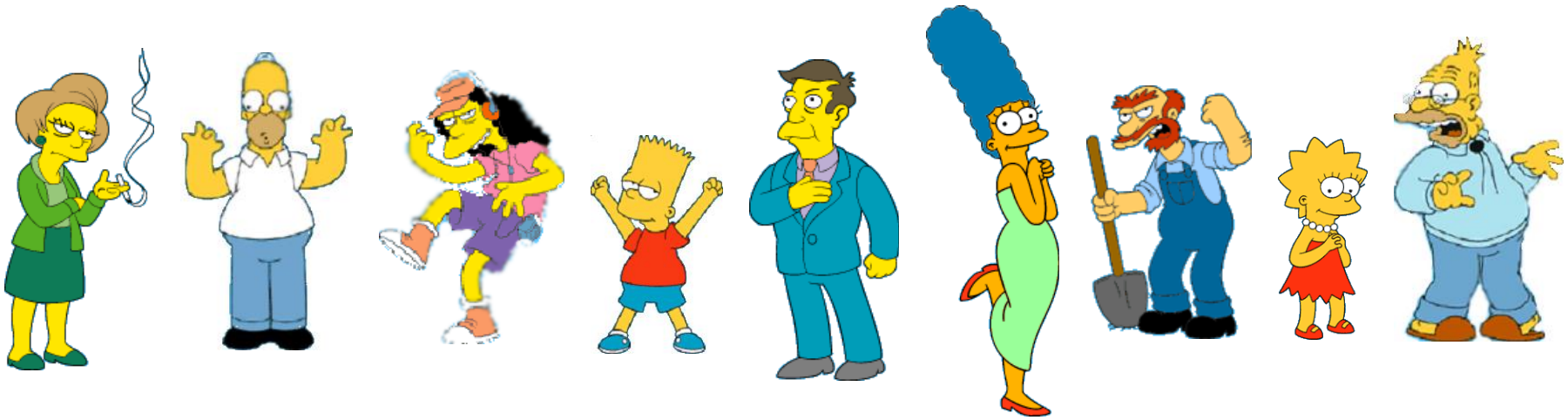
- Given a set of data points, each described by a set of attributes, find clusters such that:
  - Inter-cluster similarity is maximized
  - Intra-cluster similarity is minimized



- Requires the definition of a similarity measure

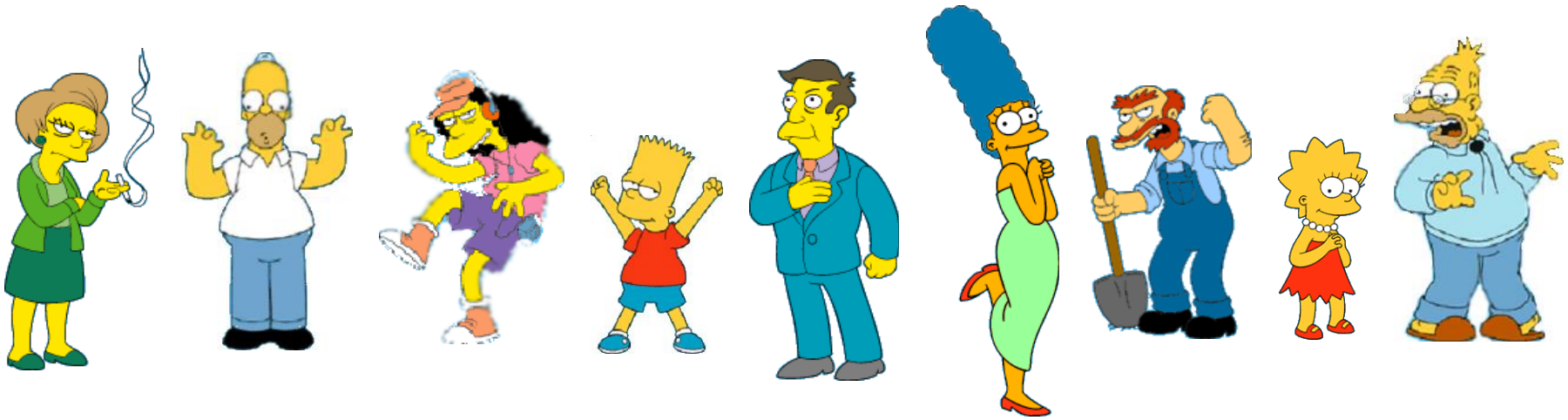
What is a natural grouping of these objects?

Slide from Eamonn Keogh

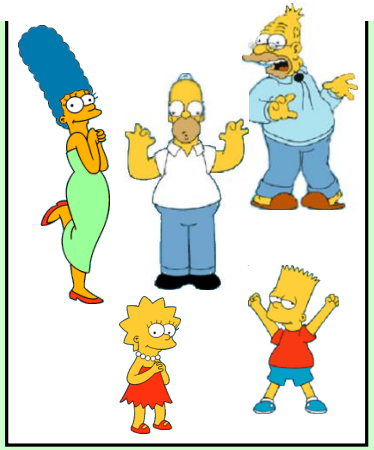


What is a natural grouping of these objects?

Slide from Eamonn Keogh

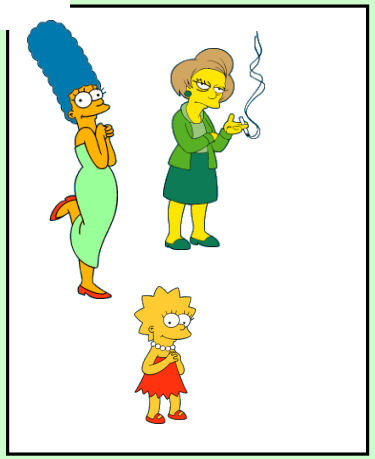


Clustering is subjective



Simpson's Family

School Employees



Females

Males

What is Similarity?

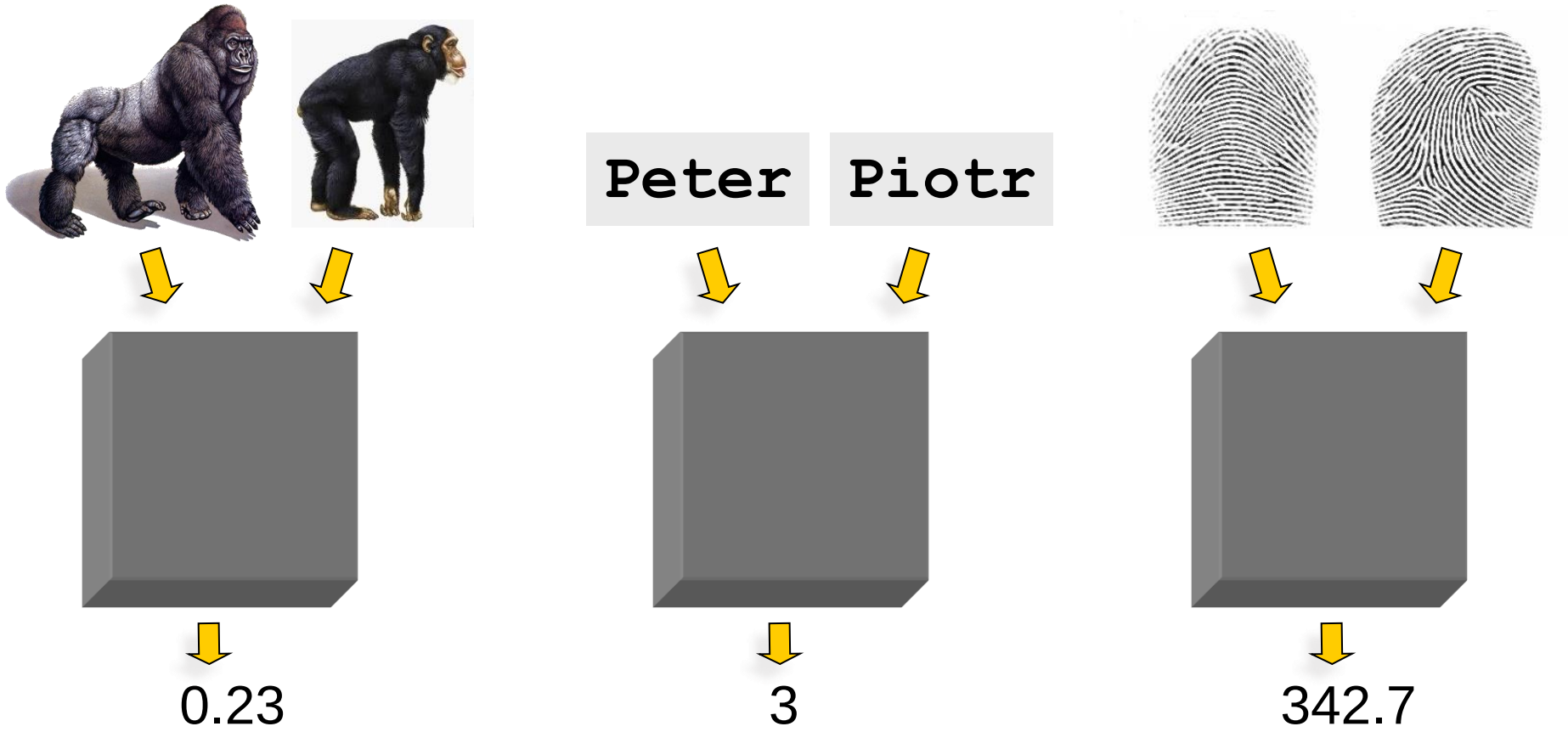
Slide based on one by Eamonn Keogh

Similarity is hard to define, but…
"*We know it when we see it*"

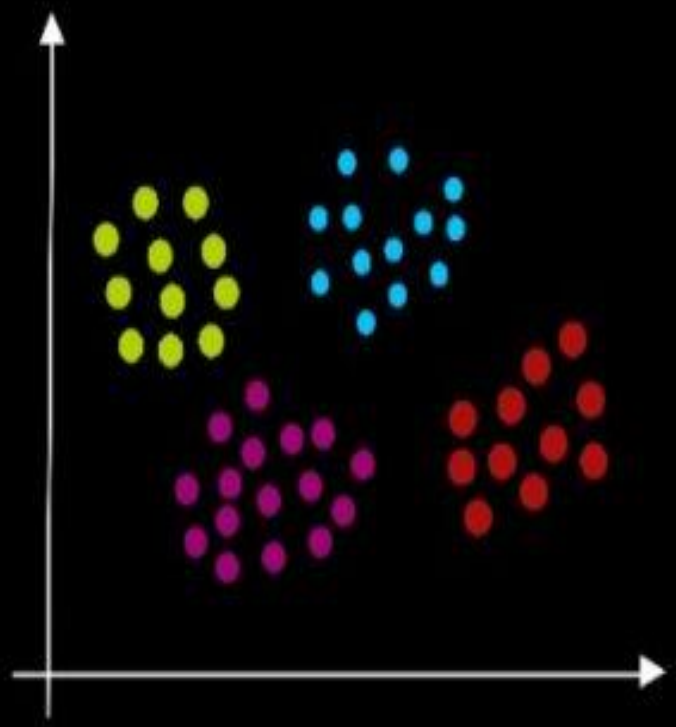
Defining Distance Measures

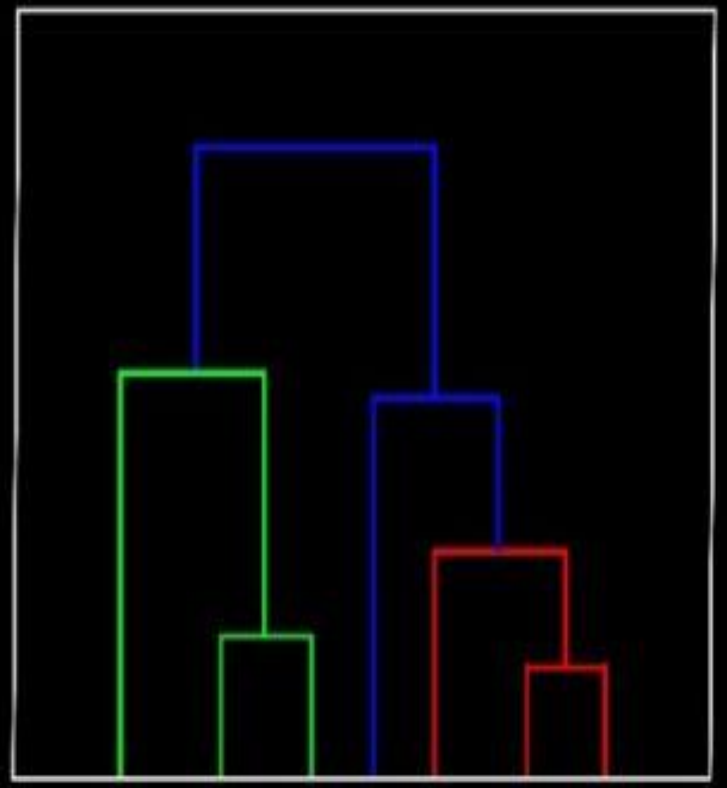
Slide from Eamonn Keogh

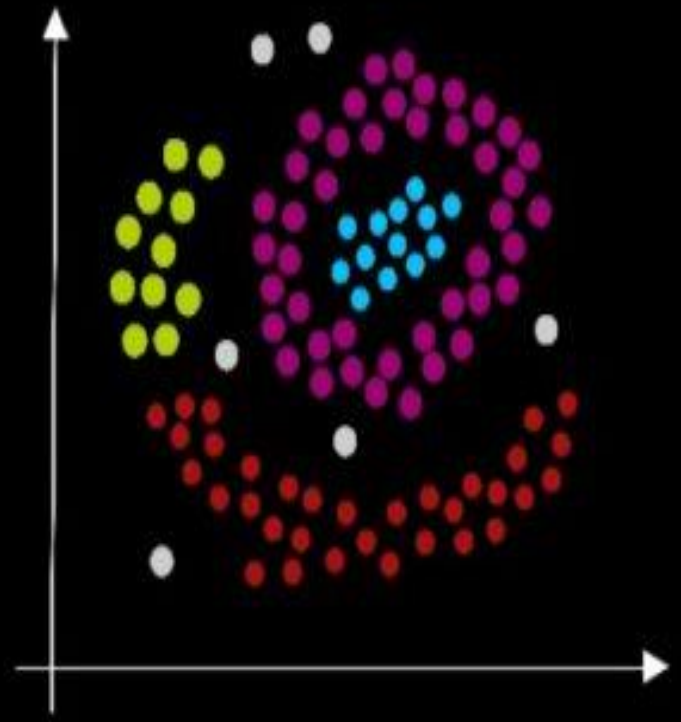
Definition: Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) between O_1 and O_2 is a real number denoted by $D(O_1,O_2)$



CLUSTERING IN MACHINE LEARNING





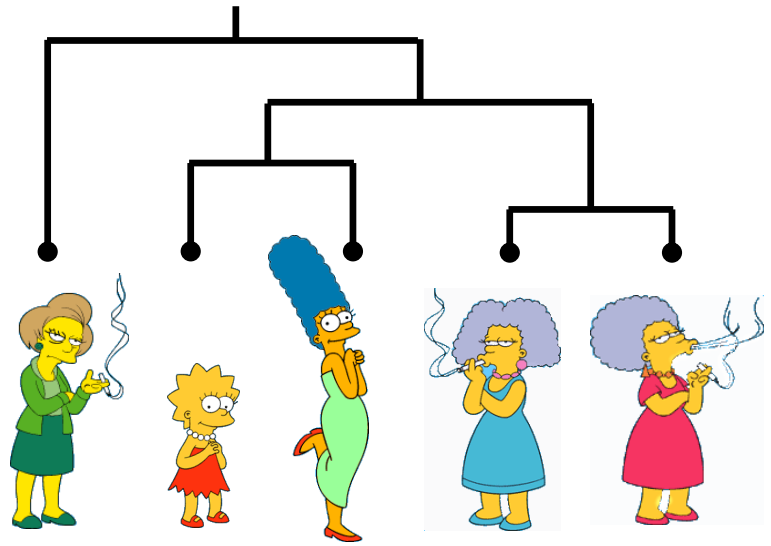


Slide based on one by Eamonn Keogh

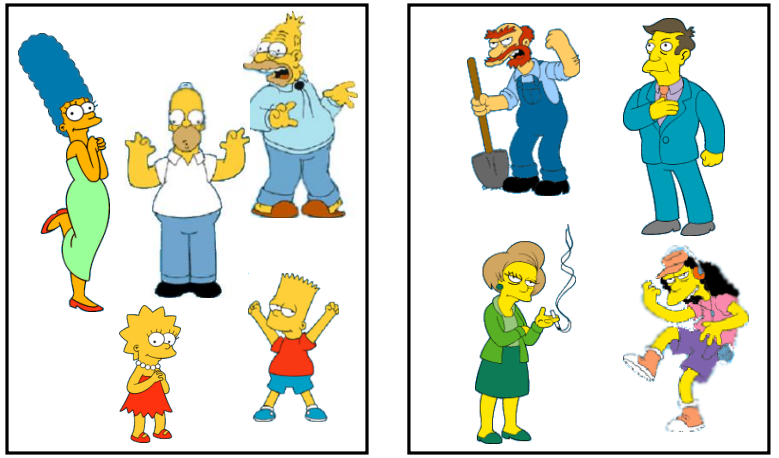
Two Types of Clustering

- **Partitional algorithms:** Construct various partitions and then evaluate them by some criterion
- **Hierarchical algorithms:** Create a hierarchical decomposition of the set of objects using some criterion

Hierarchical

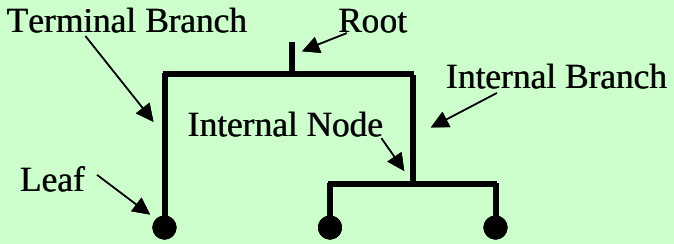


Partitional

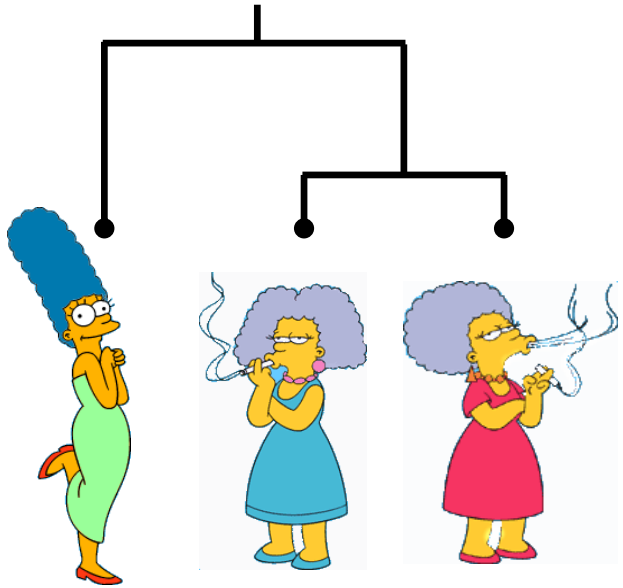


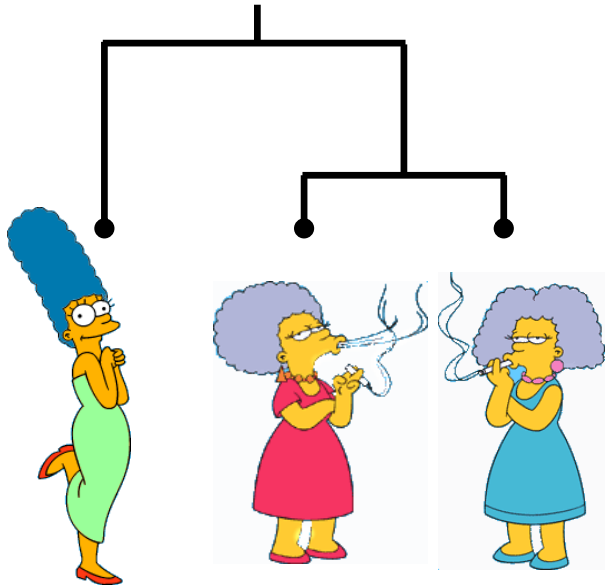
Slide based on one by Eamonn Keogh

Dendogram: A Useful Tool for Summarizing Similarity Measurements



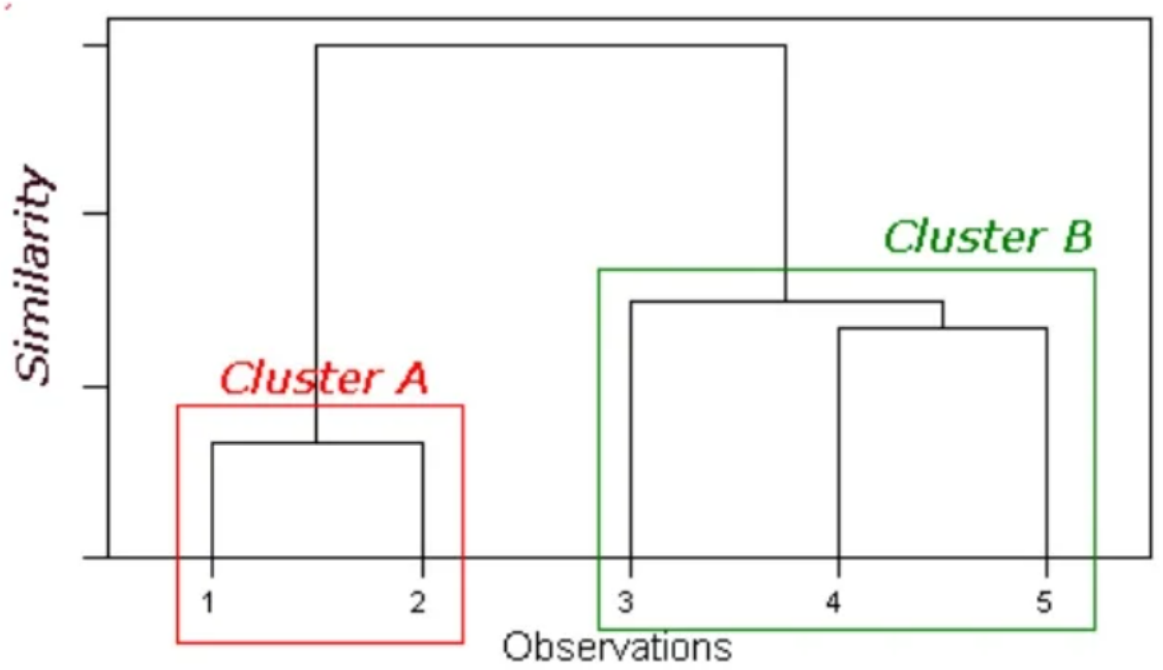
The similarity between two objects in a dendrogram is represented as the height of the lowest internal node they share.





Clustering Algorithms :

- Hierarchy methods
  - statistical method used to build a cluster by arranging elements at various levels

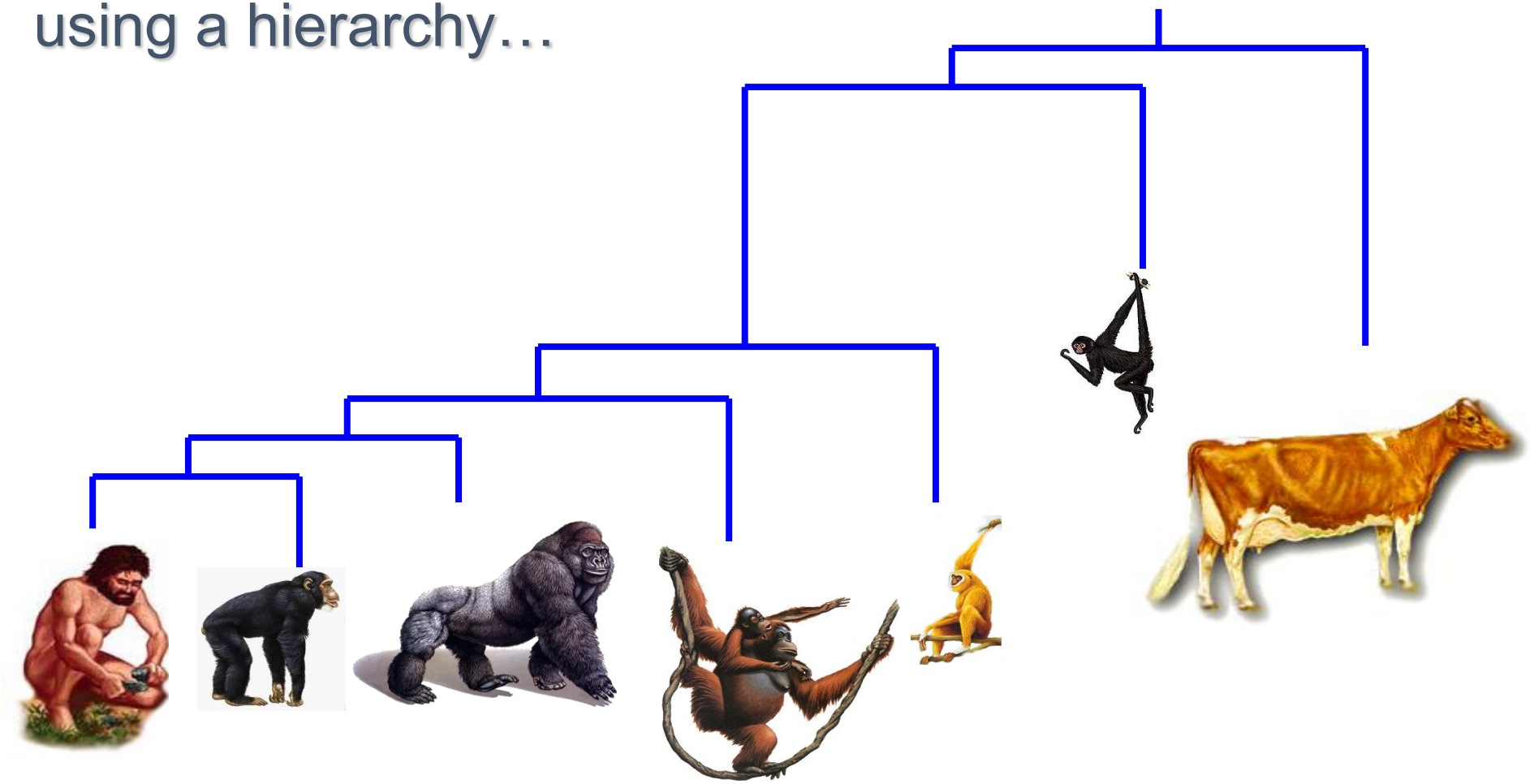


Dendogram:

- Each level will then represent a possible cluster.
- The height of the dendrogram shows the level of similarity that any two clusters are joined
- The closer to the bottom they are the more similar the clusters are
- **Finding of groups from a dendrogram is not simple and is very often subjective**

Slide based on one by Eamonn Keogh

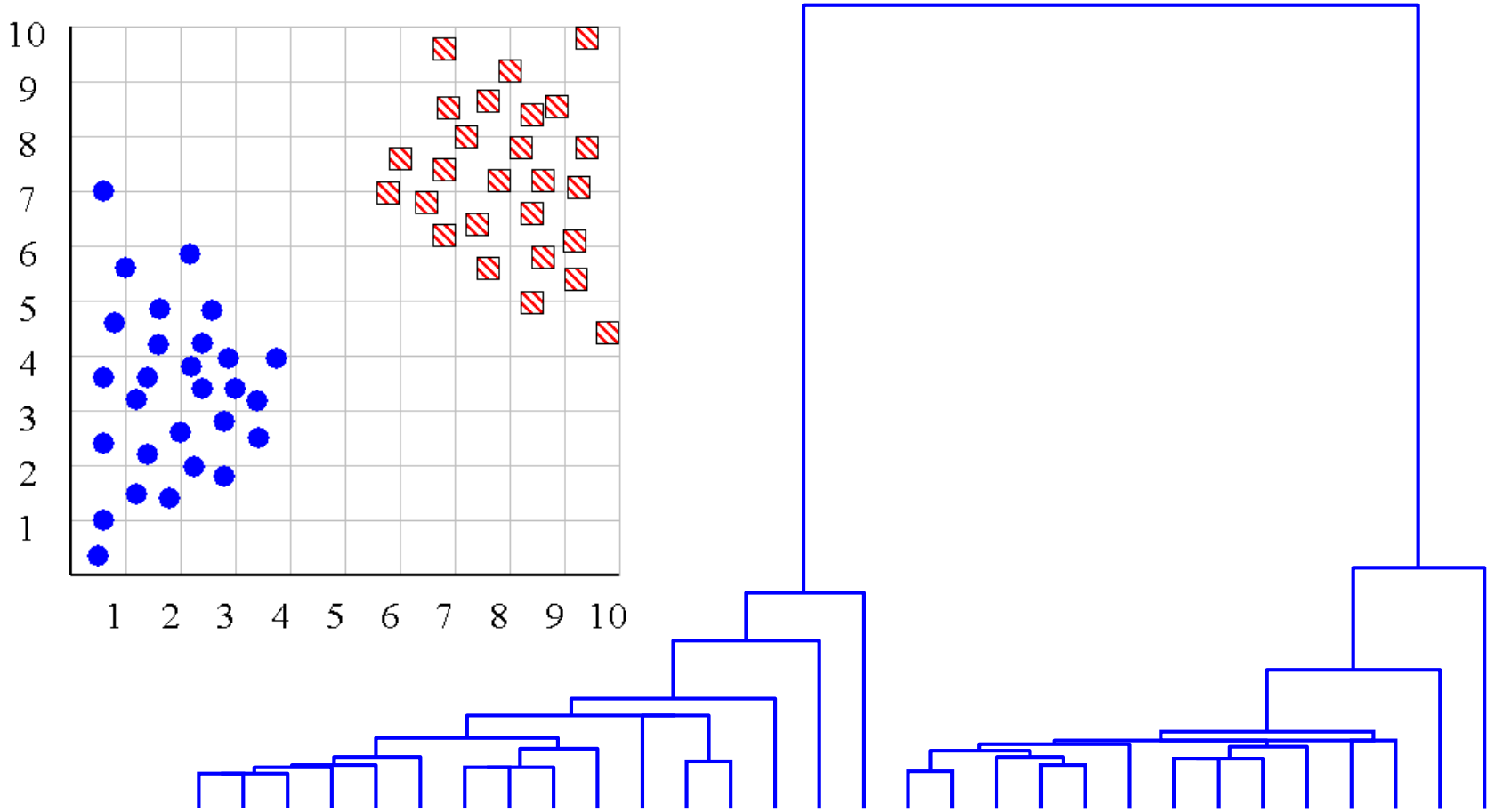
There is only one dataset that can be perfectly clustered using a hierarchy…



(Bovine:0.69395, (Spider Monkey 0.390, (Gibbon:0.36079,(Orang:0.33636,(Gorilla:0.17147,(Chimp:0.19268, Human:0.11927):0.08386):0.06124):0.15057):0.54939);

Slide based on one by Eamonn Keogh

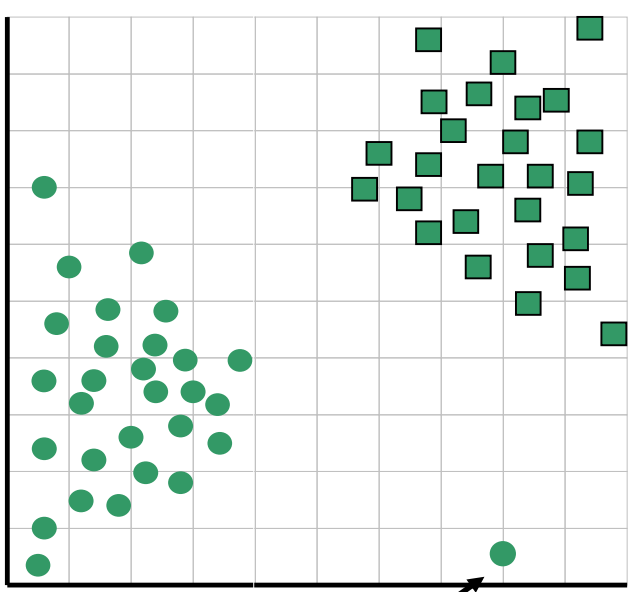
We can look at the dendrogram to determine the “correct” number of clusters.



Slide based on one by Eamonn Keogh

One potential use of a dendrogram: detecting outliers

The single isolated branch is suggestive of a data point that is very different to all others



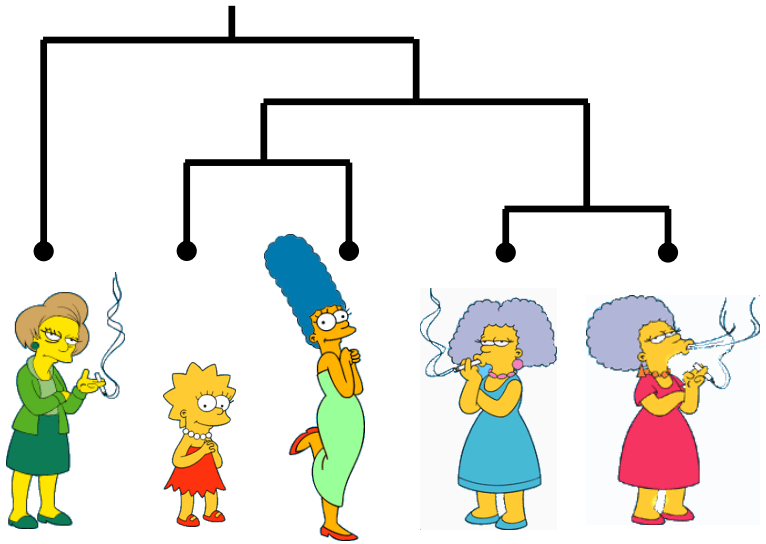
Outlier

Hierarchical Clustering

Slide based on one by Eamonn Keogh

The number of dendrograms with n leafs $= (2n-3)!/[(2^{(n-2)})(n-2)!]$

Number of Leafs	Number of Possible Dendrograms
2	1
3	3
4	15
5	105
...	…
10	34,459,425



Since we cannot test all possible trees we will have to heuristic search of all possible trees. We could do this..

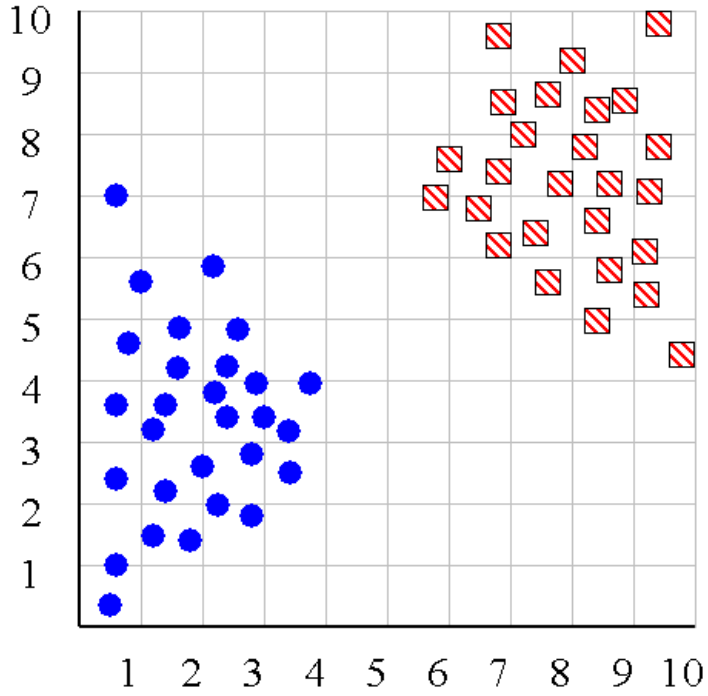
Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

Top-Down (divisive): Starting with all the data in a single cluster, consider every possible way to divide the cluster into two. Choose the best division and recursively operate on both sides.

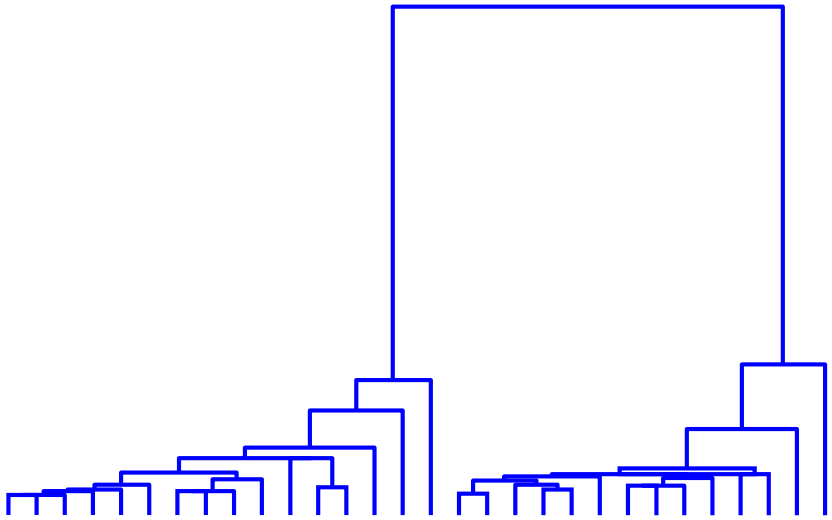
Slide based on one by Eamonn Keogh

We know how to measure the distance between two objects, but defining the distance between an object and a cluster, or defining the distance between two clusters is non obvious.

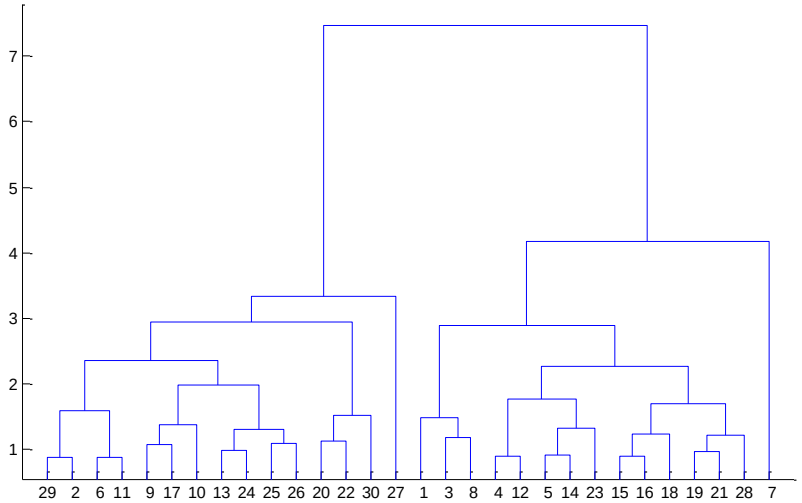
- **Single linkage (nearest neighbor):** In this method the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters.
- **Complete linkage (furthest neighbor):** In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors").
- **Group average linkage:** In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters.



Slide based on one by Eamonn Keogh



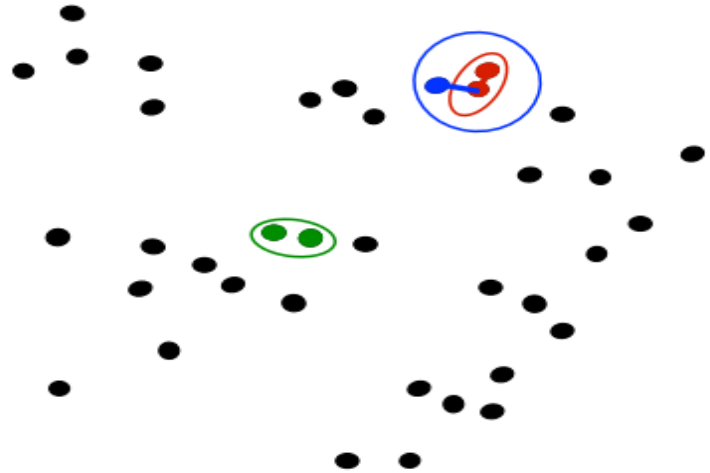
Single linkage



Average linkage

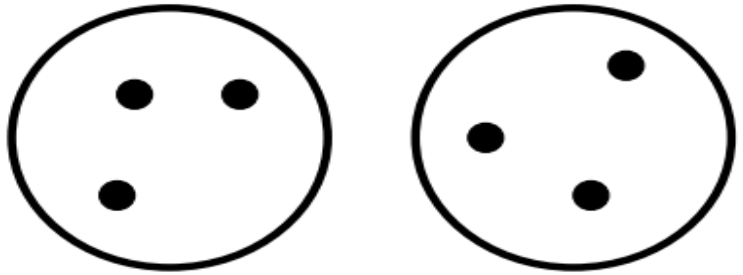
Agglomerative Clustering

- Agglomerative clustering:
  - First merge very similar instances
  - Incrementally build larger clusters out of smaller clusters
- Algorithm:
  - Maintain a set of clusters
  - Initially, each instance in its own cluster
  - Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster
    - Stop when there's only one cluster left
- Produces not one clustering, but a family of clusterings represented by a dendrogram



Agglomerative Clustering

- How should we define “closest” for clusters with multiple elements?

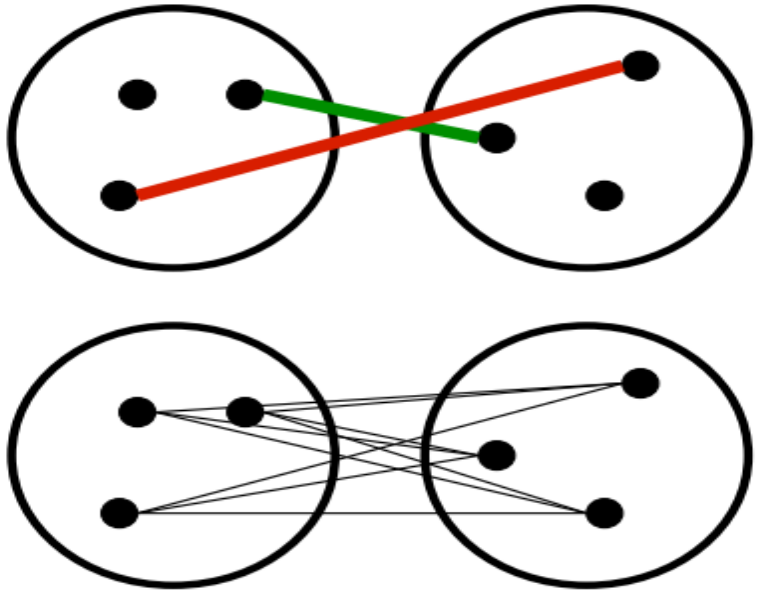


Agglomerative Clustering

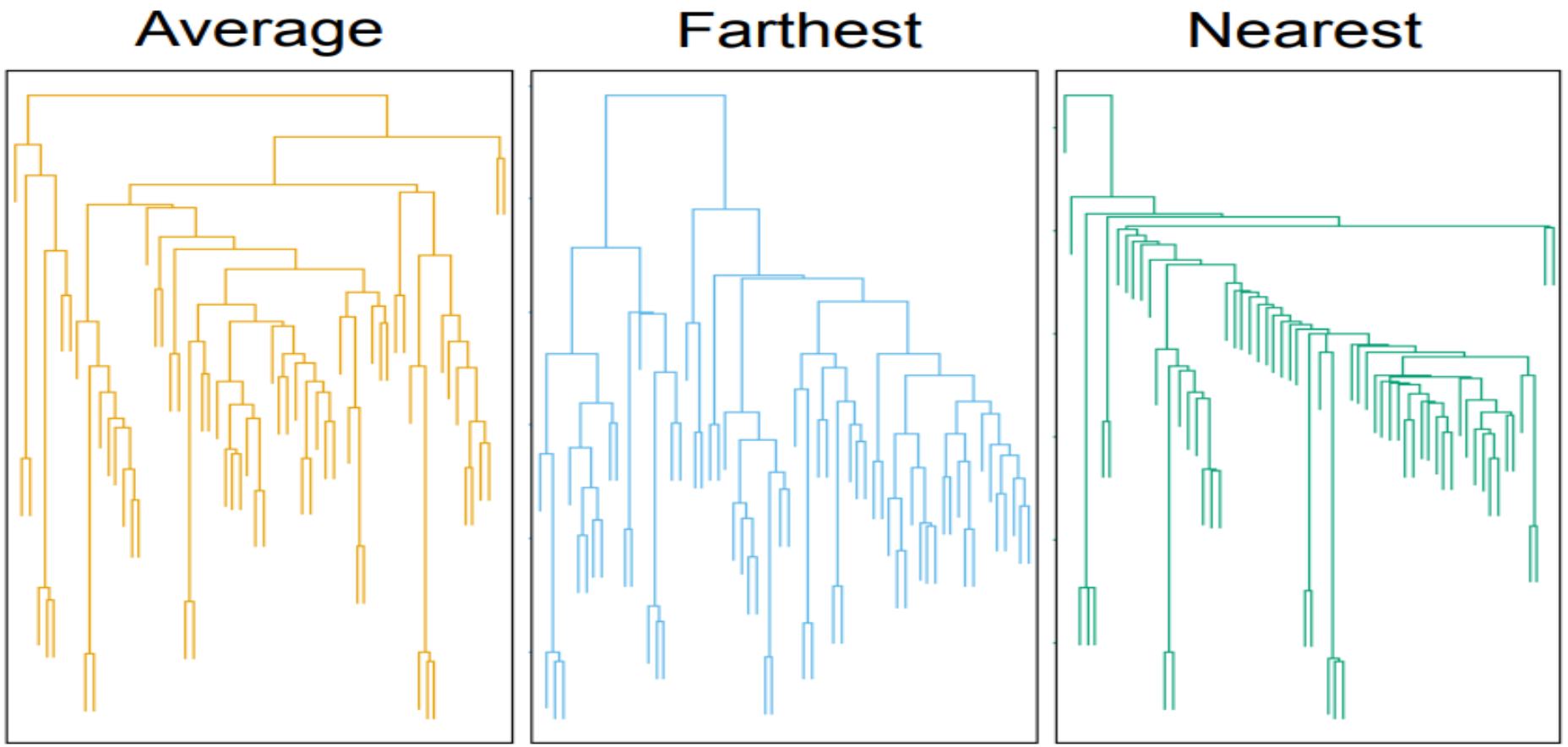
- How should we define "closest" for clusters with multiple elements?

- Many options:
  - Closest pair (single-link clustering)
  - Farthest pair (complete-link clustering)
  - Average of all pairs

- Different choices create different clustering behaviors

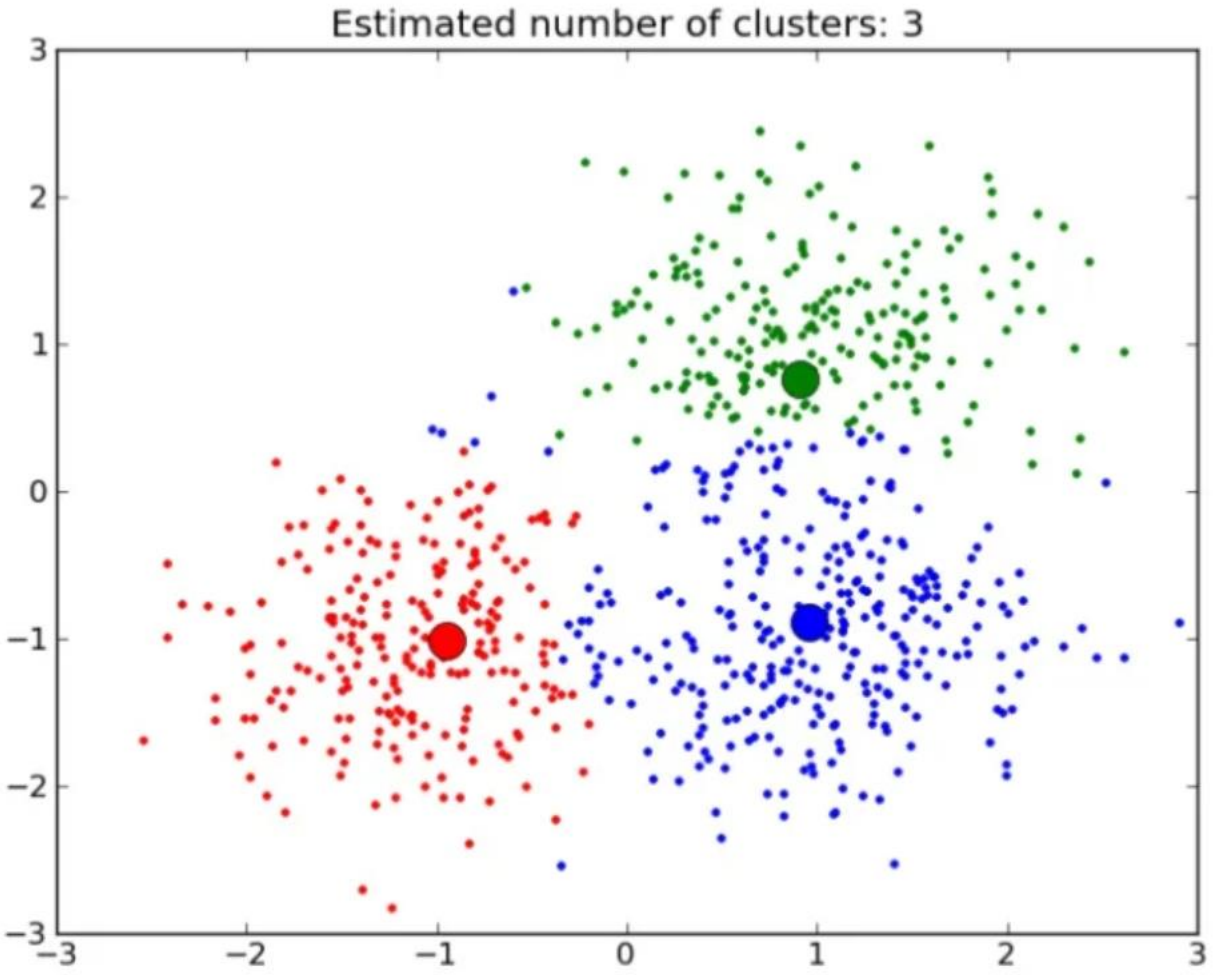


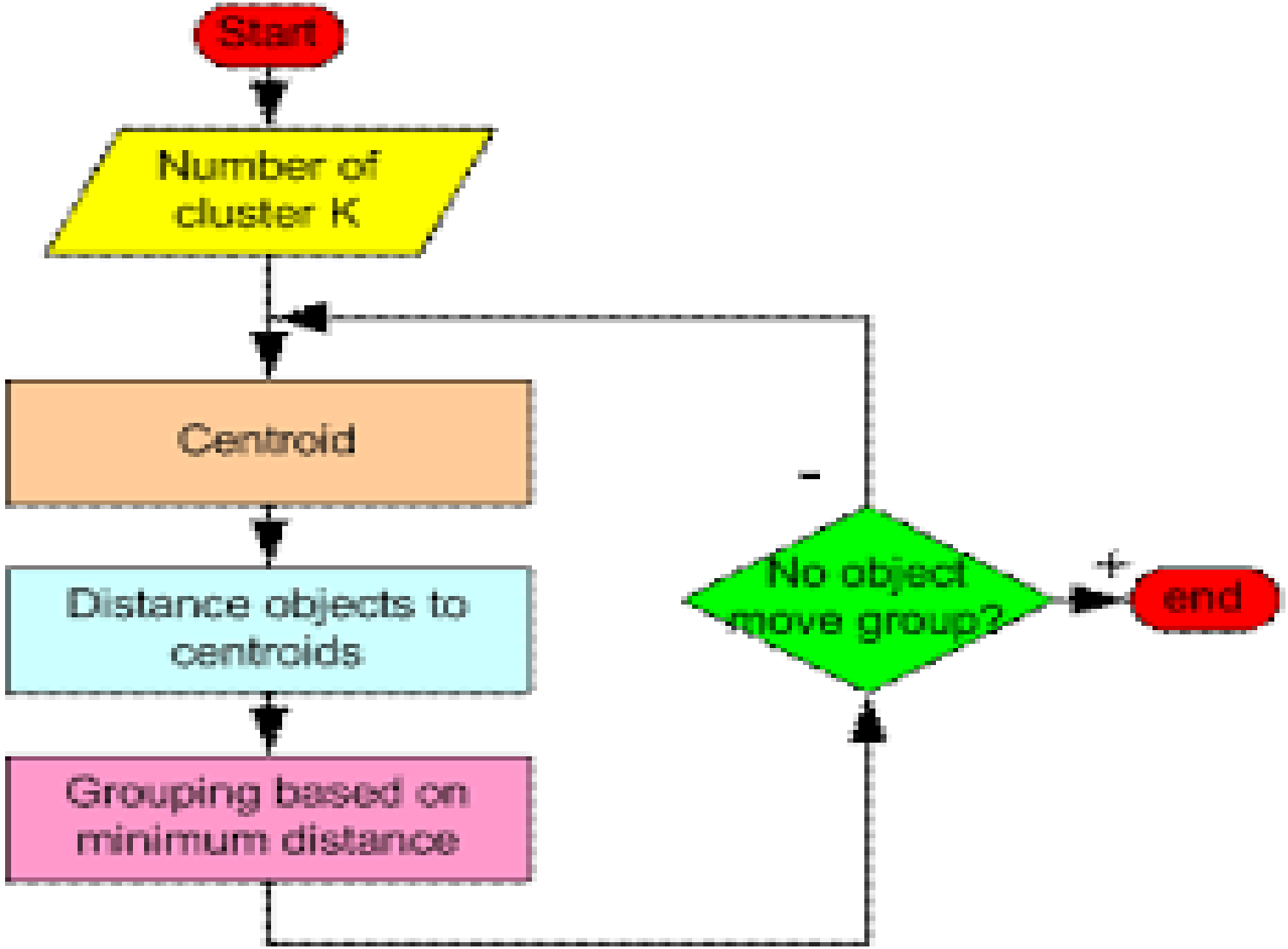
Clustering Behavior



Mouse tumor data from [Hastie *et al.*]

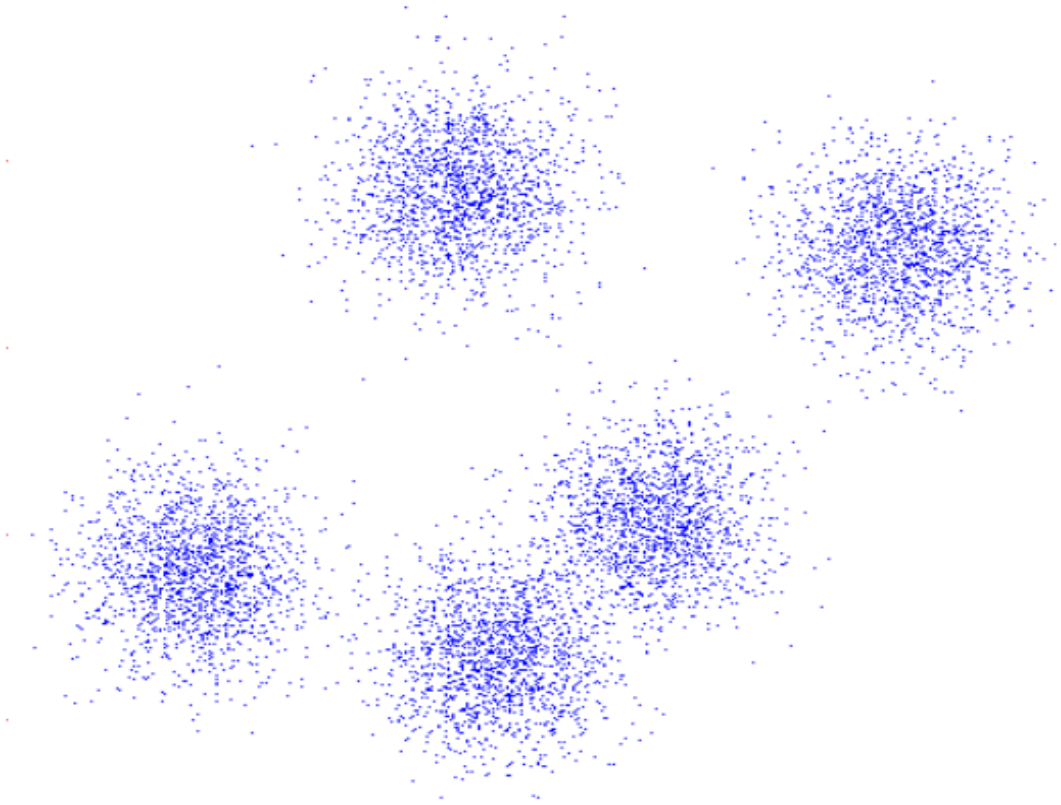
K-means:



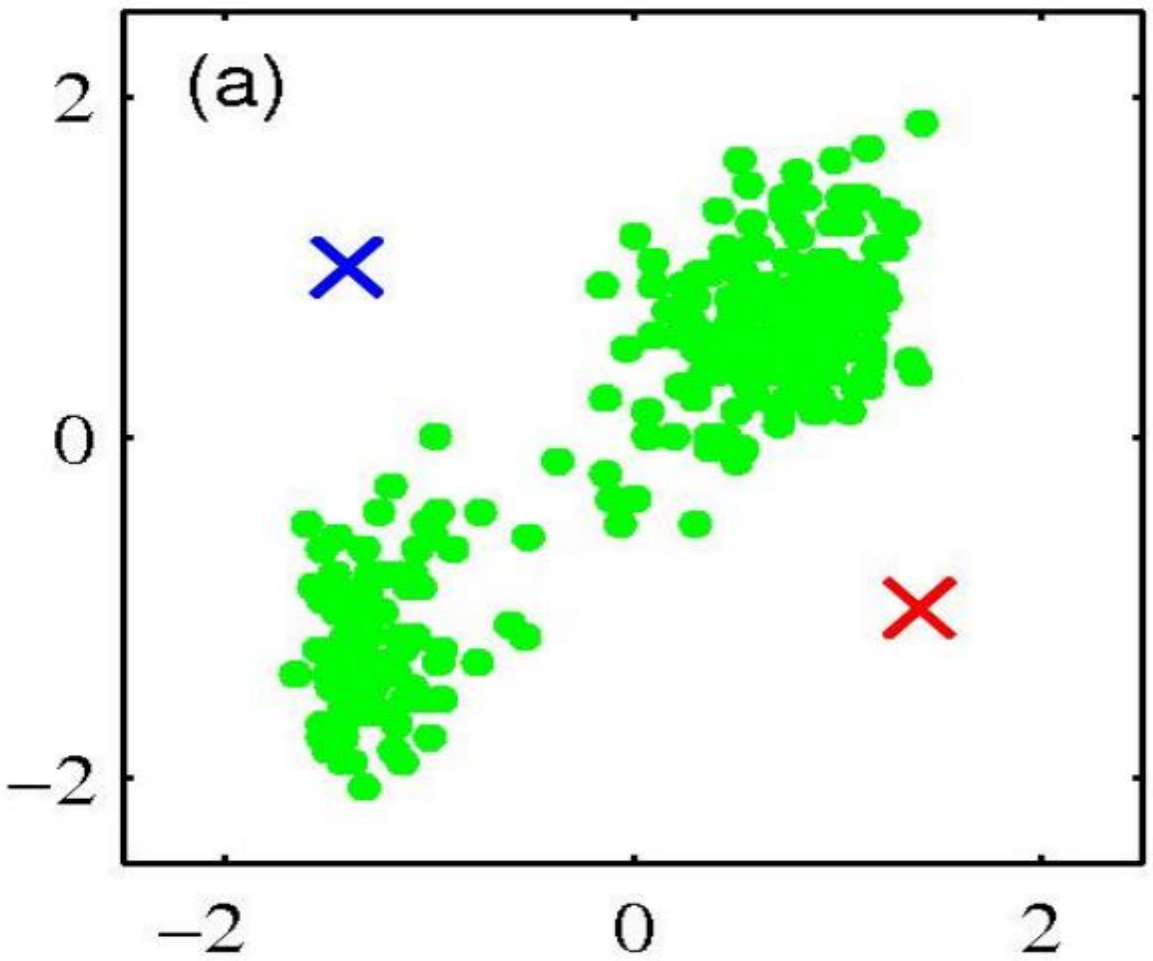

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

K-Means

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



K-means clustering: Example



- Pick *K* random points as cluster centers (means)

Shown here for *K*=2

Slide based on one by Eamonn Keogh

Partition Algorithm 1: k-means

1. Decide on a value for *k*.

2. Initialize the *k* cluster centers (randomly, if necessary).

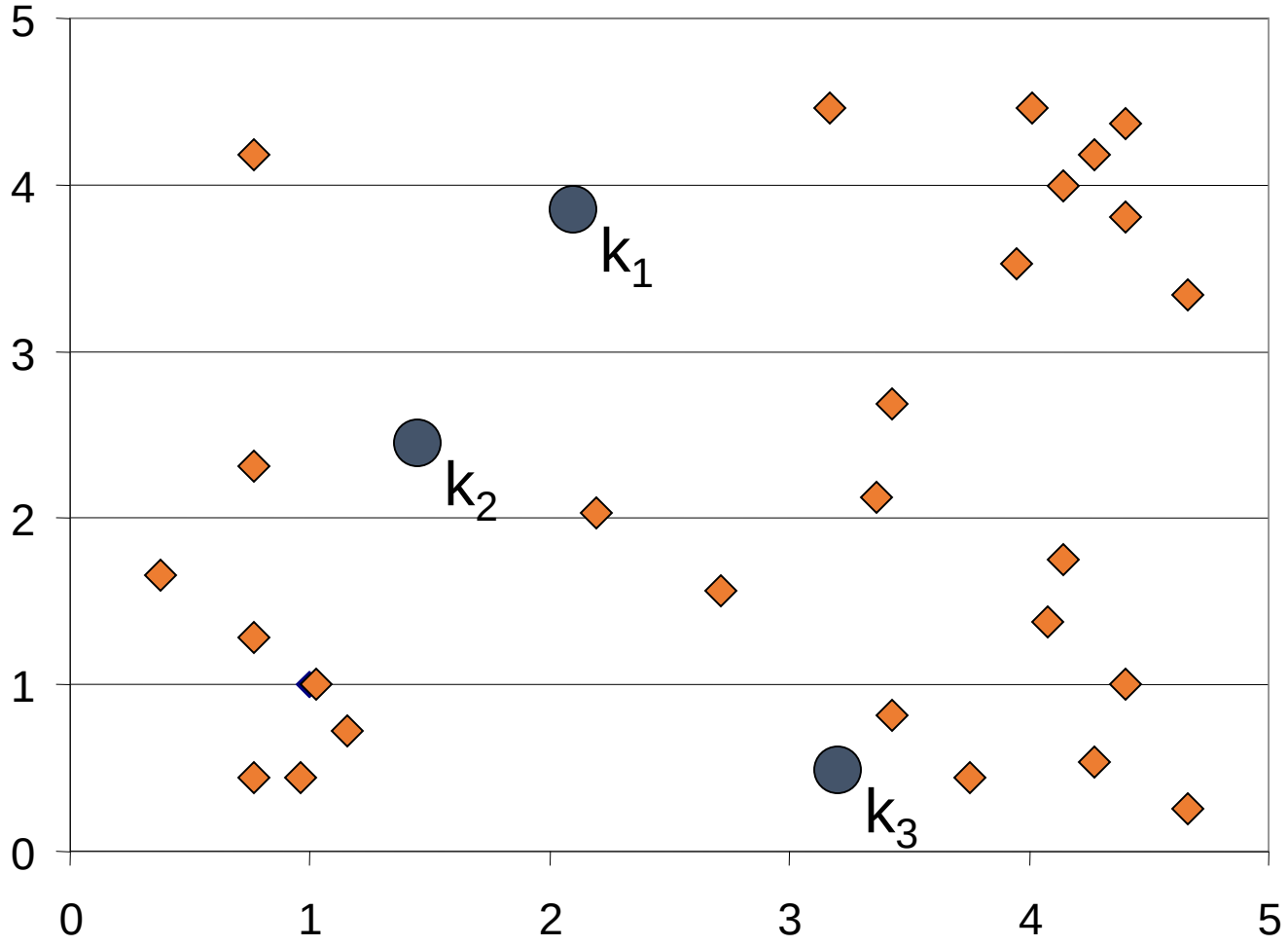
3. Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.

4. Re-estimate the *k* cluster centers, by assuming the memberships found above are correct.

5. If none of the *N* objects changed membership in the last iteration, exit. Otherwise goto 3.

K-means Clustering: Step 1

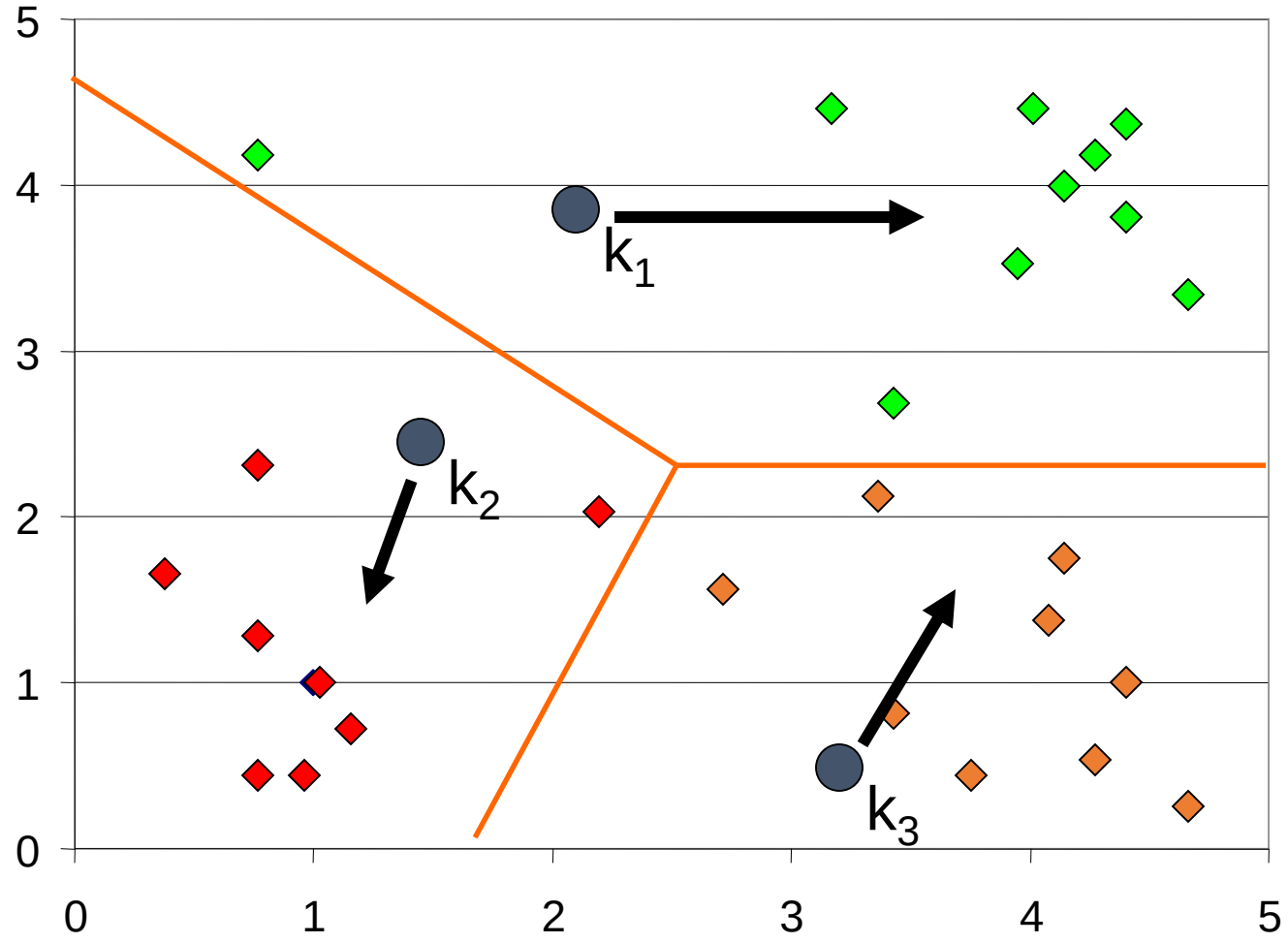
Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh

K-means Clustering: Step 2

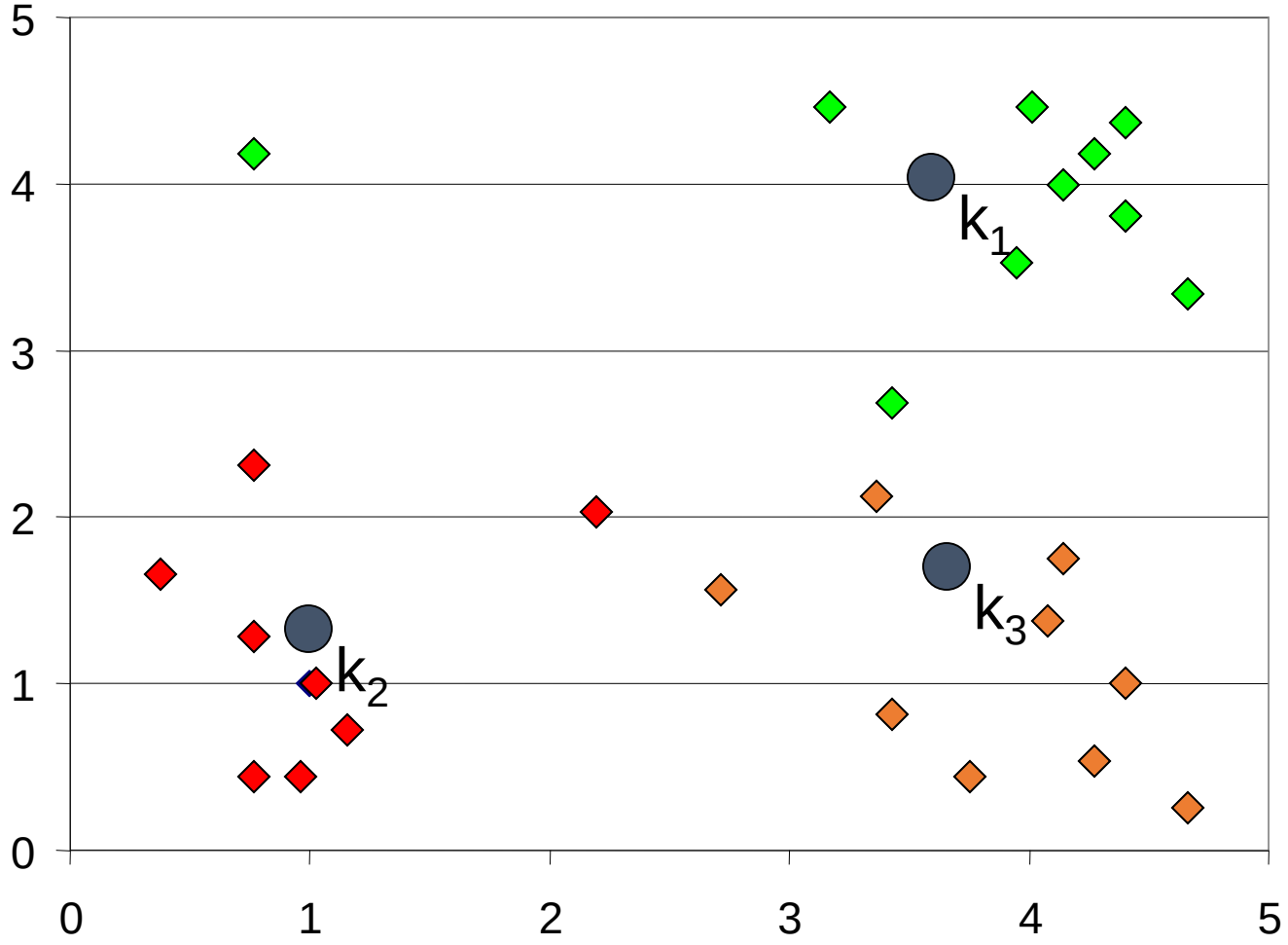
Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh

K-means Clustering: Step 3

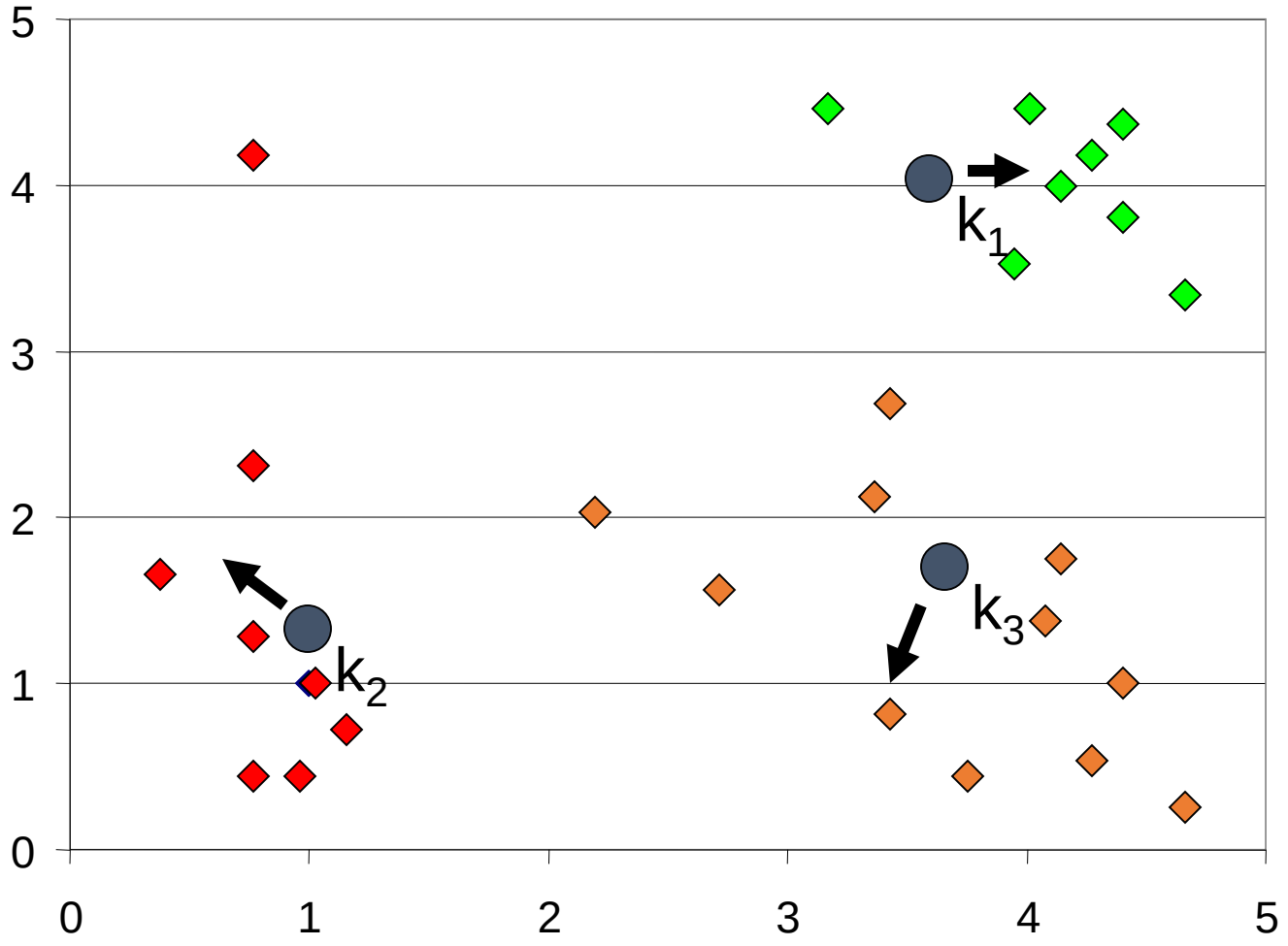
Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh

K-means Clustering: Step 4

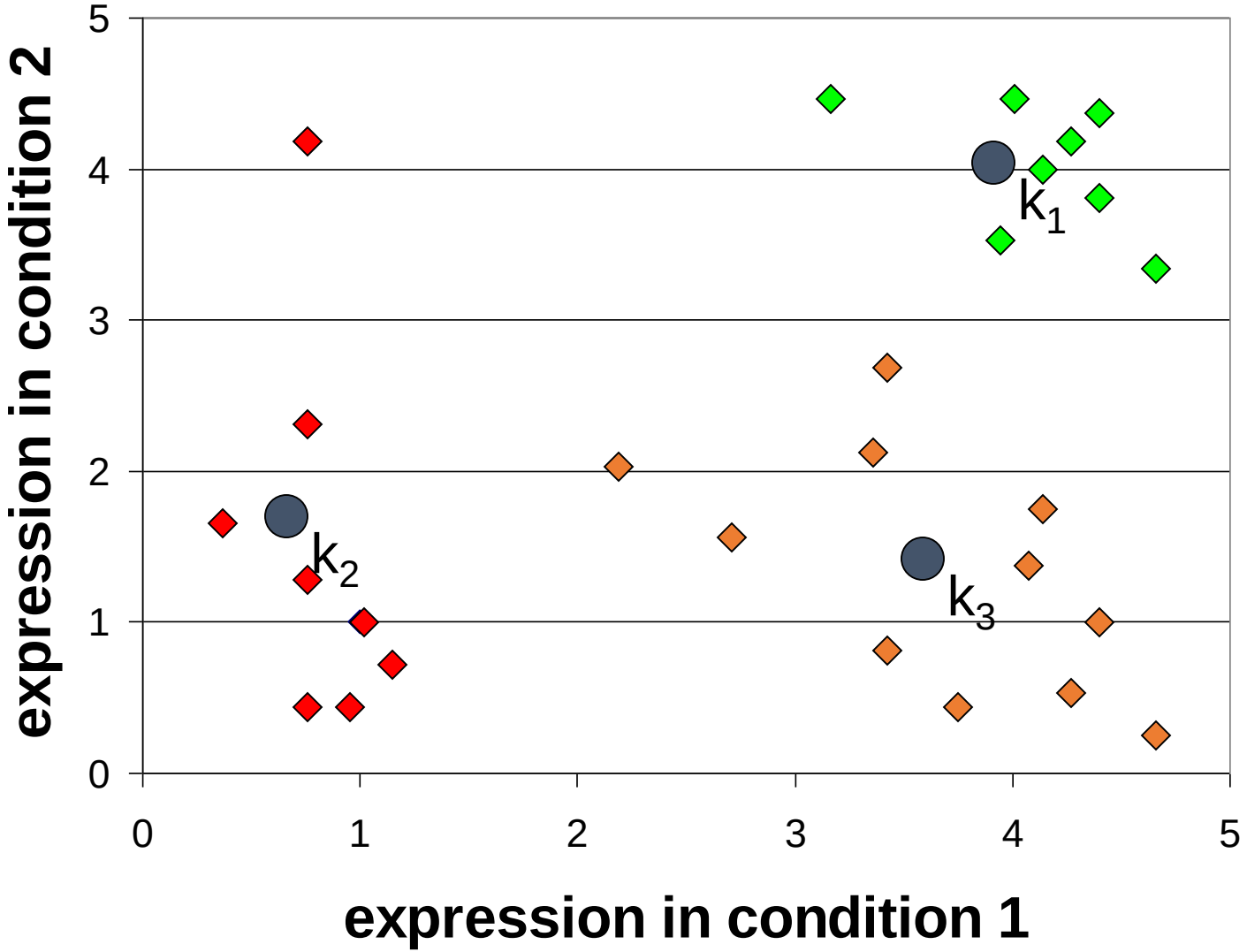
Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh

K-means Clustering: Step 5

Algorithm: k-means, Distance Metric: Euclidean Distance



Slide based on one by Eamonn Keogh