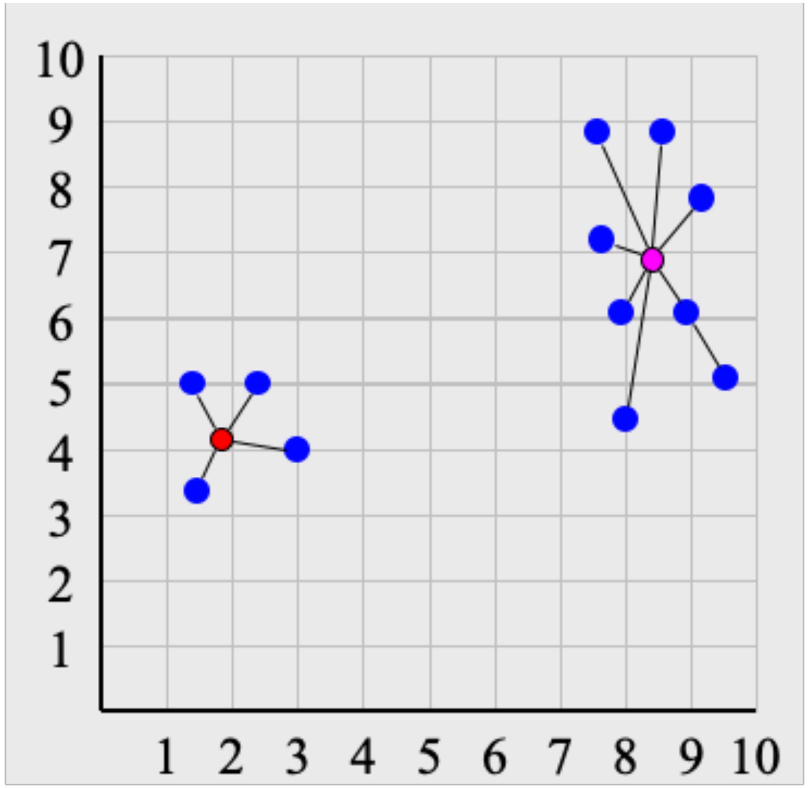
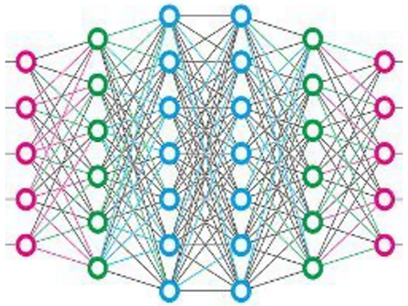
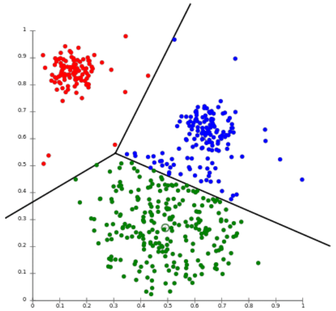
CS 445
Introduction to Machine Learning

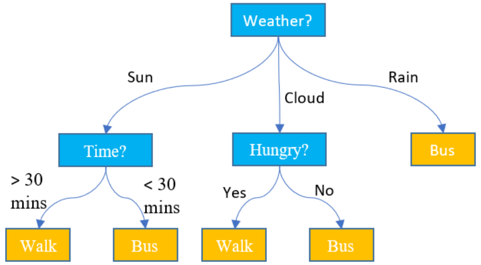
Unsupervised Learning
Clustering

Instructor: Dr. Kevin Molloy









Weather?

Sun

Cloud

Rain

Time?

Hungry?

Bus

> 30 mins

< 30 mins

Yes

No

Walk

Bus

Walk

Bus

Learning Objectives

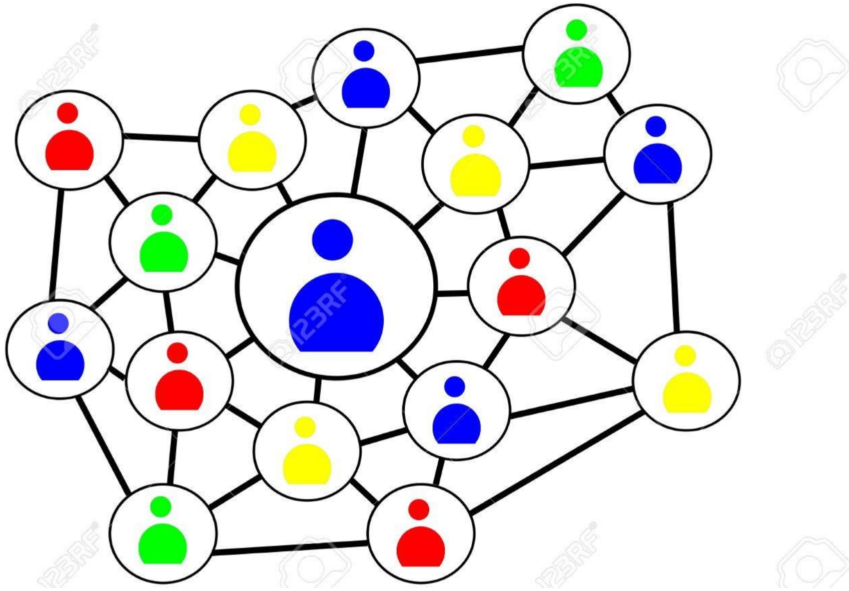
- Define unsupervised learning
- Define k-means clustering and sketch an outline of Lloyd's algorithm

Is the entire world labeled?

Problem with labels and training data are supervised learning problem

Here are some problems that we would like to utilize machine learning:

- Organize a set of newspaper articles by topic
- Grouping customers together by their interest
- Recommending movies to watch based on your interests
- Identifying outliers or anomalies in your data
- Perform a social network analysis to identify groups of individuals



Cocktail Party

The Cocktail party effect refers to our brain's ability to focus our attention on a particular conversation, despite the large amount of noise in a room.

Original recordings: Mic 1 🔊

Mic 2 🔊

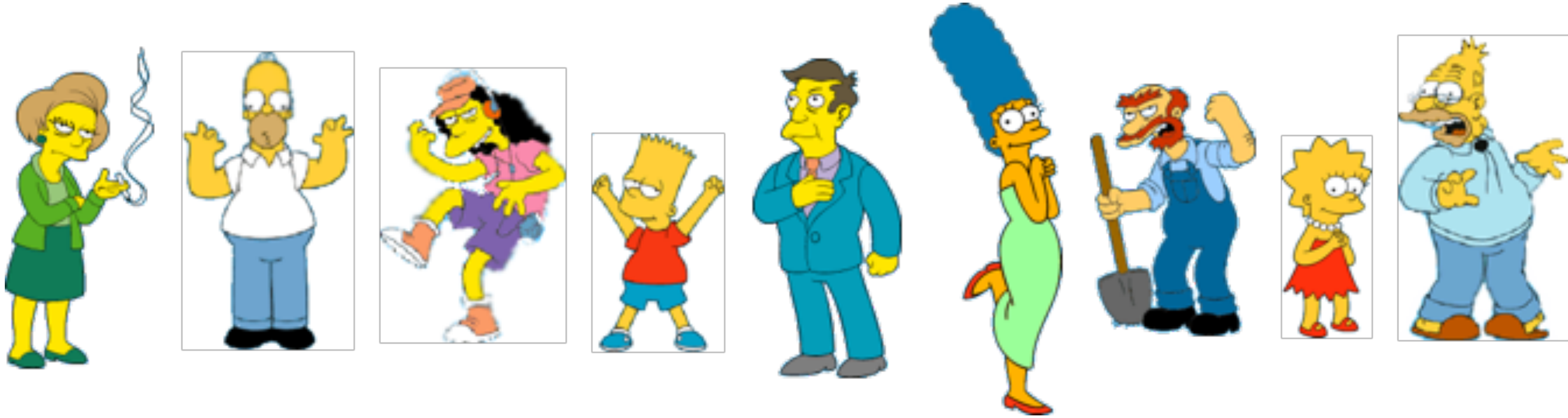
Separated: Source 1 🔊

Source 2 🔊

Audio courtesy of Te-Won Lee, ICASSP 1998 paper.

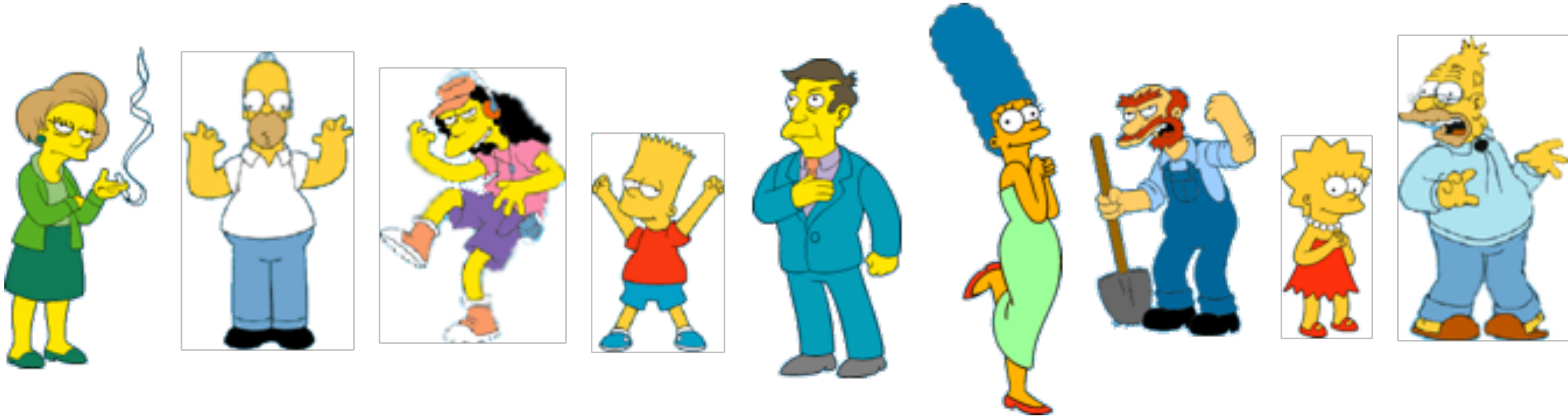
Clustering

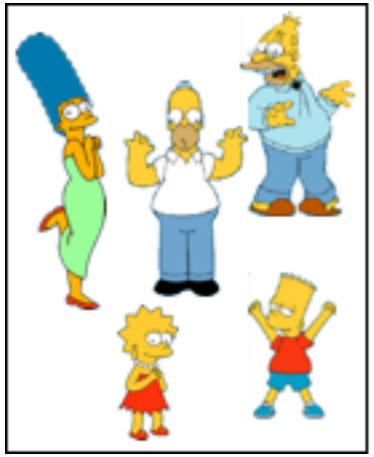
Grouping "like" objects together.



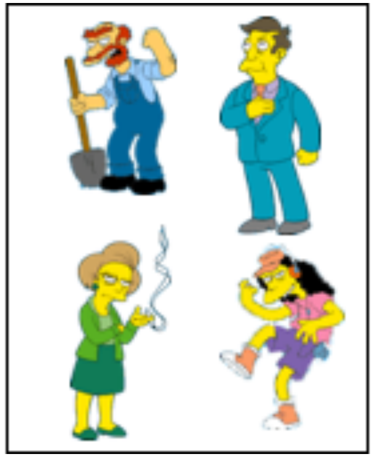
Clustering

Grouping "like" objects together.

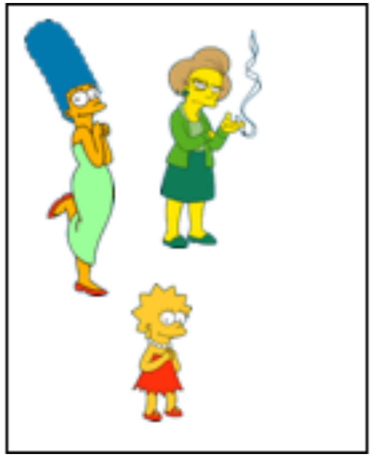




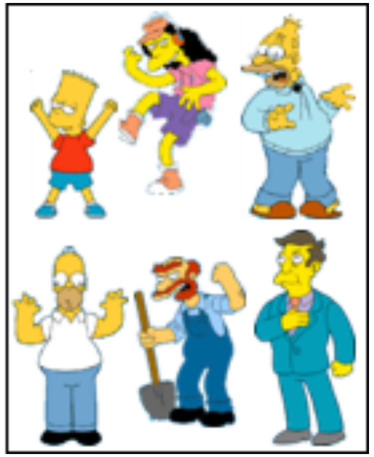
Simpon's Family



School Employees



Females



Males

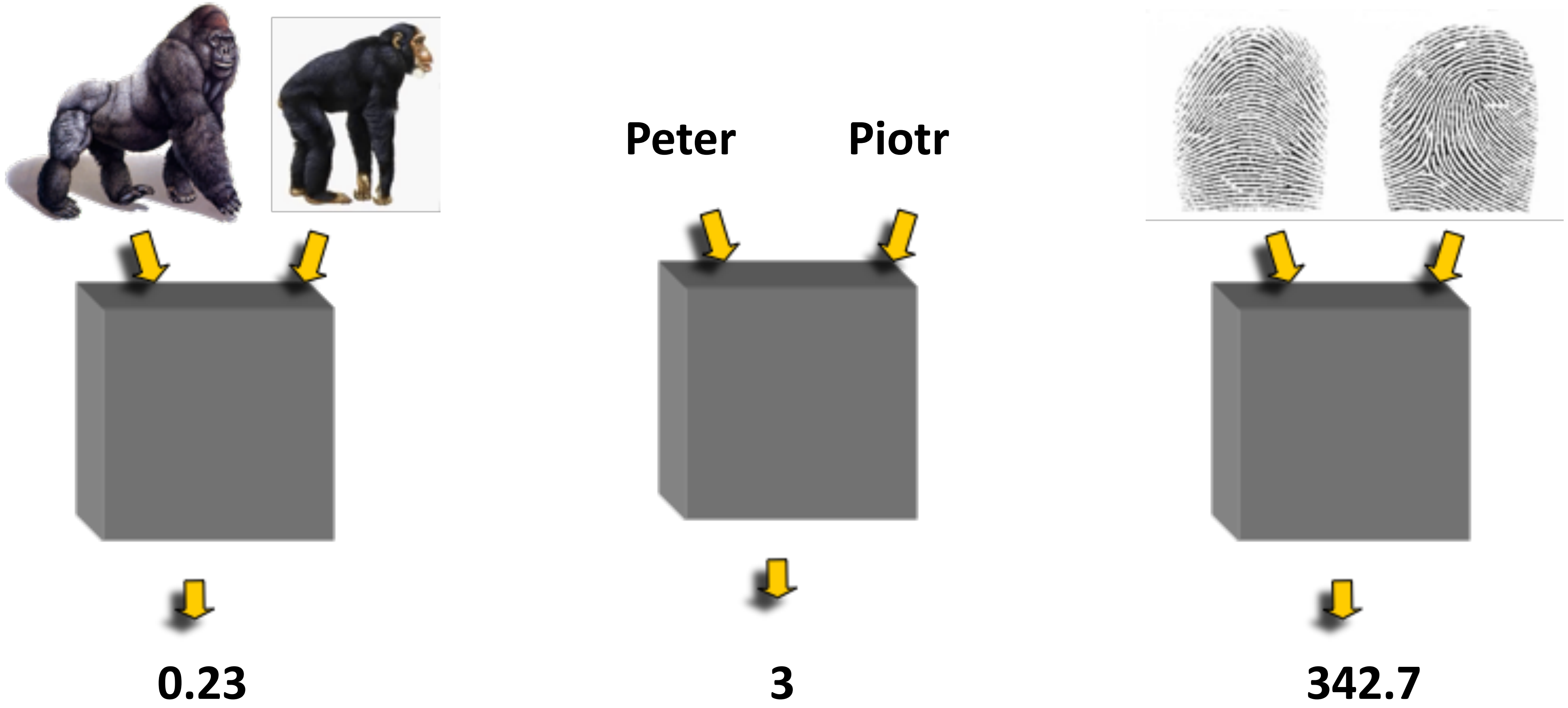
Clustering is **subjective**

Defining Similarity – Let's Start with Distance

Definition: Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) between O1 and O2 is a real number denoted by $D(O_1, O_2)$.

Defining Similarity – Let's Start with Distance

Definition: Let O_1 and O_2 be two objects from the universe of possible objects. The distance (dissimilarity) between O1 and O2 is a real number denoted by $D(O_1, O_2)$.



Peter **Piotr**

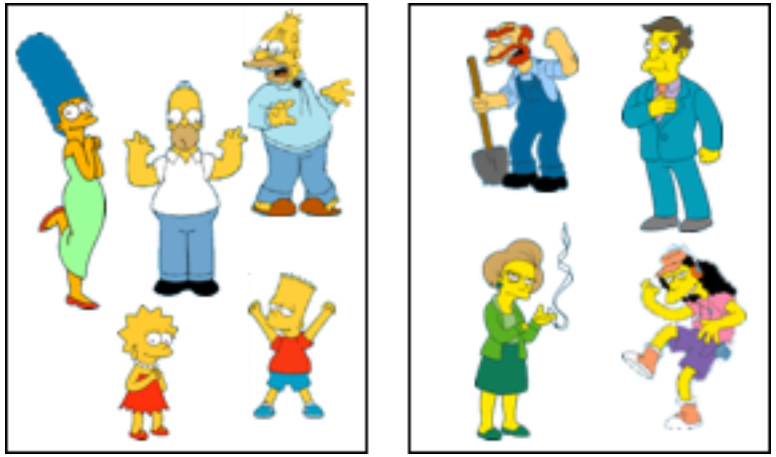
0.23

3

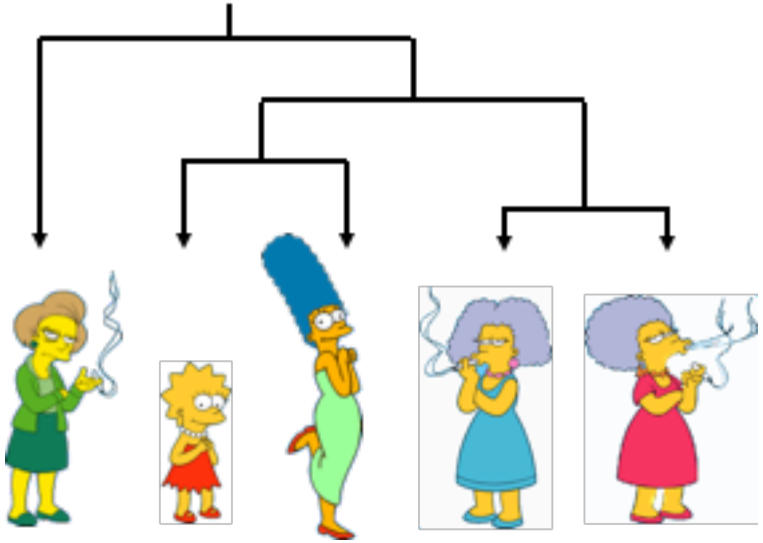
342.7

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

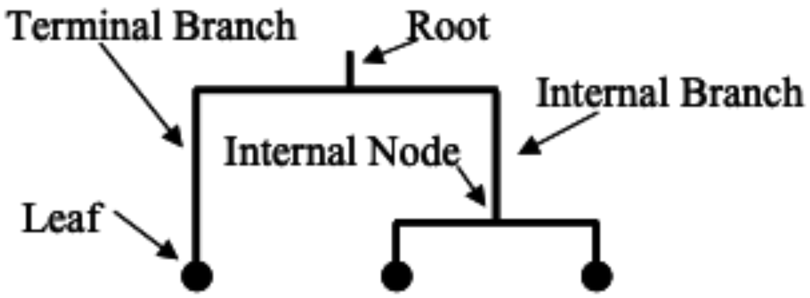


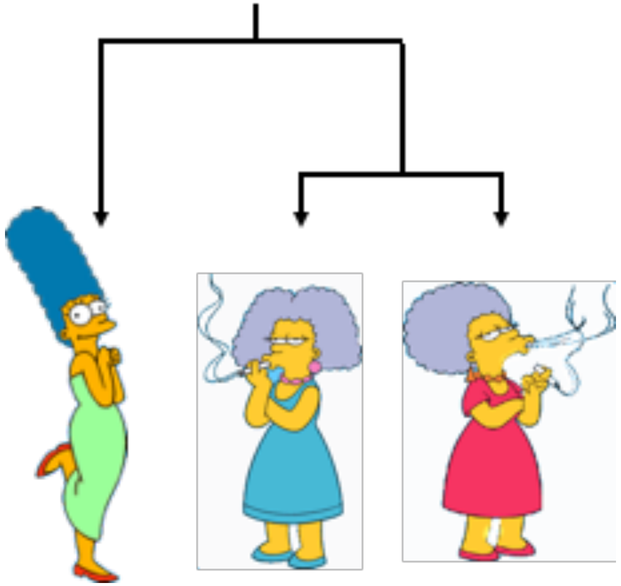
Desirable Properties of a Clustering Algorithm

- Scalability (in terms of both space and time)
- Ability to deal with different types of data
- Able to deal with noise and outliers
- Insensitive to order of input records

Visualizing Similarity

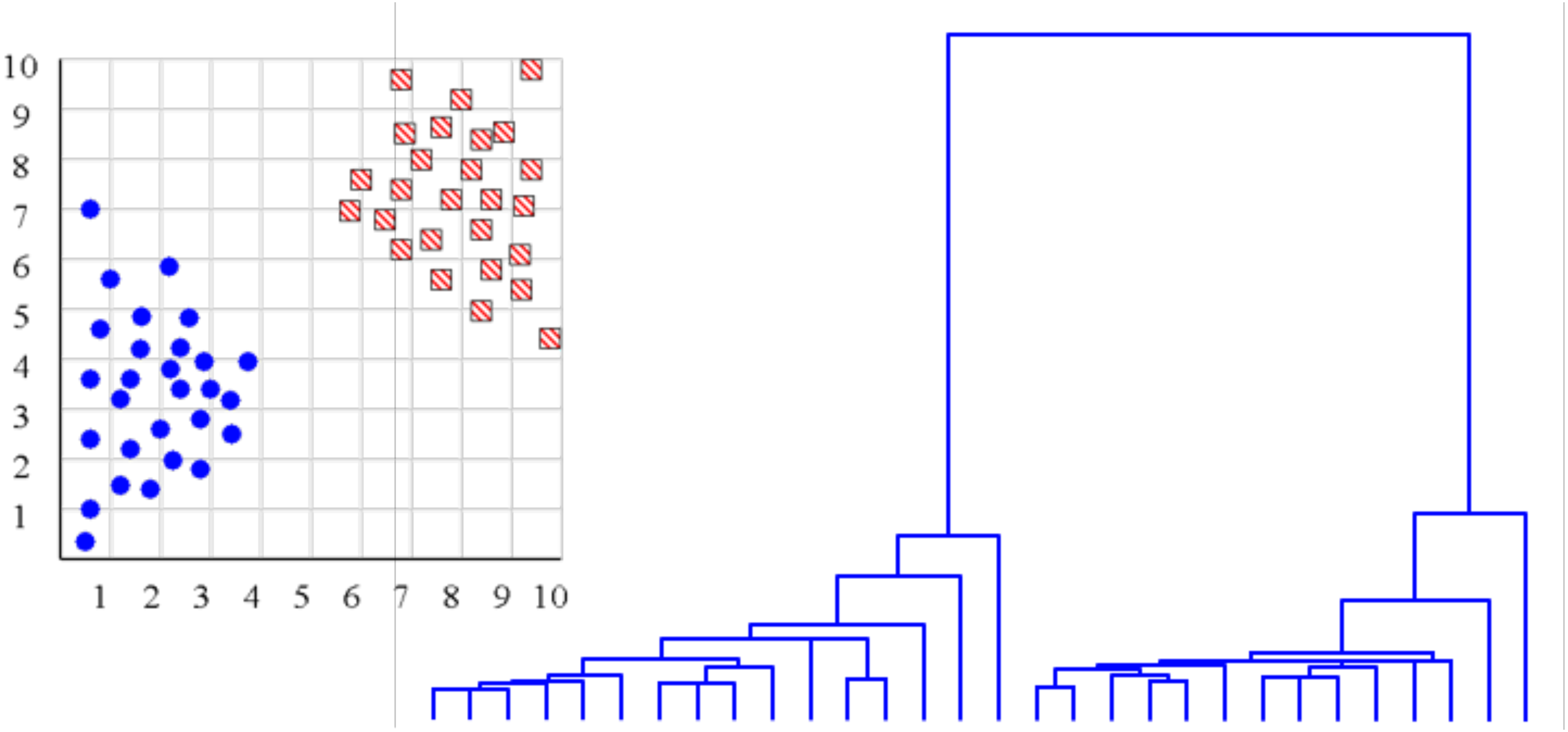
In order to better appreciate and evaluate the examples given, we introduce the dendrogram.





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

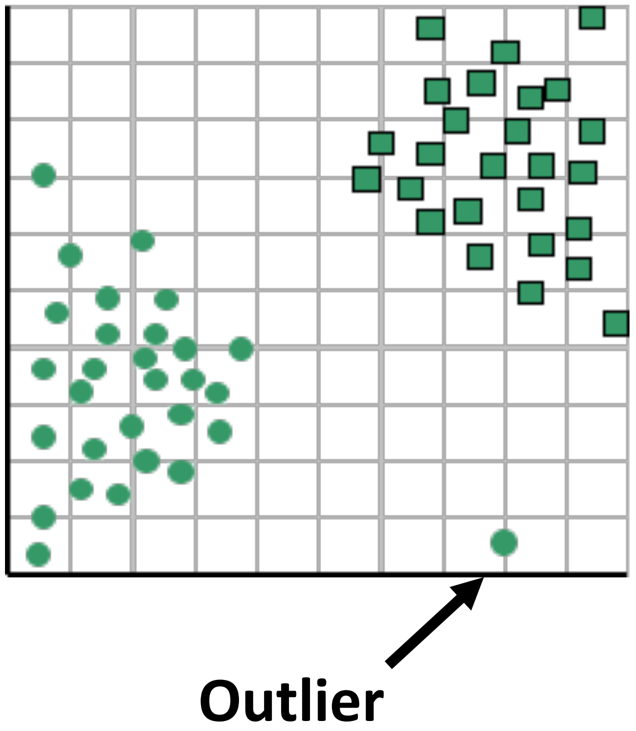
Identifying Clusters with Dendrograms

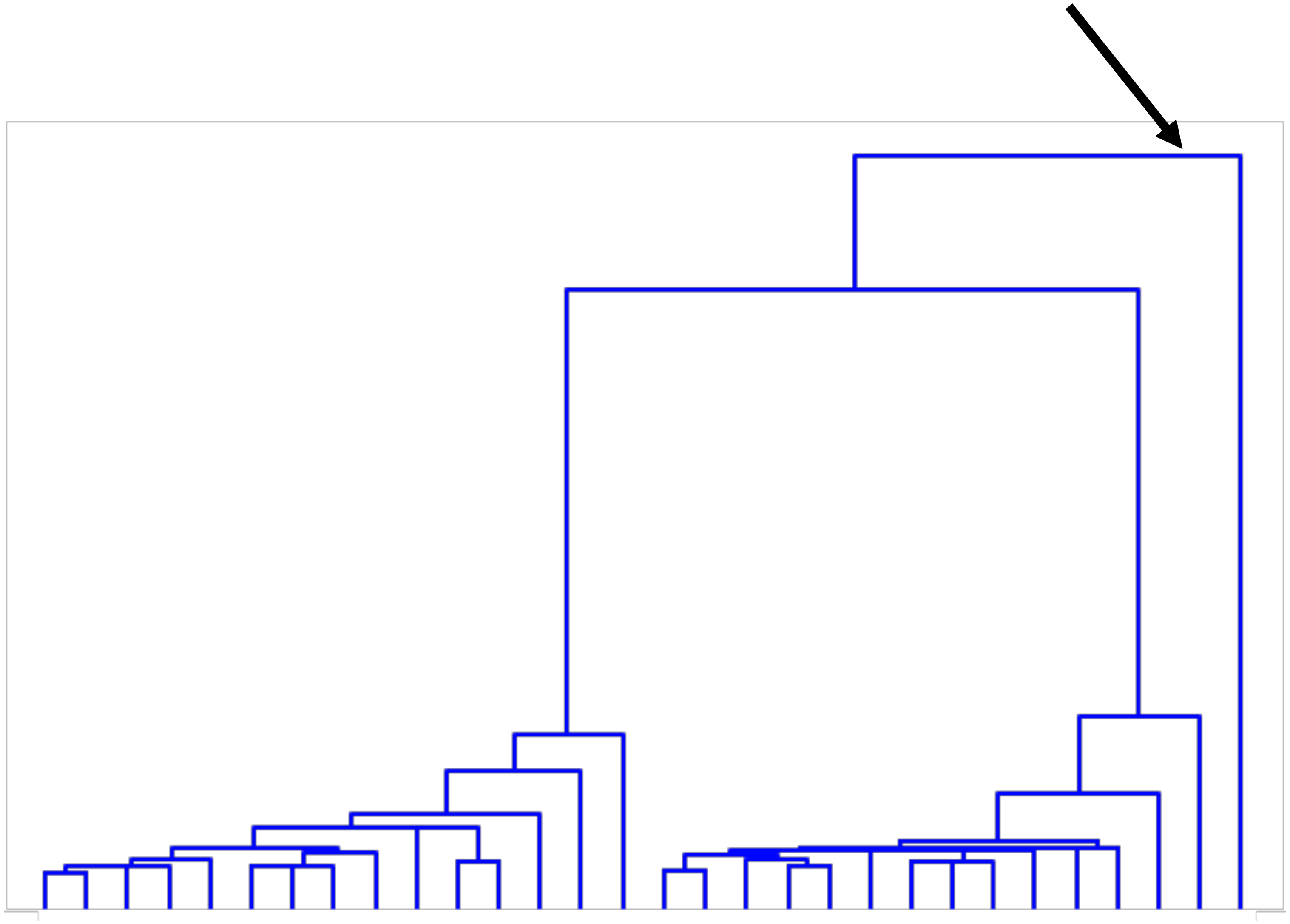


How many clusters? Dendrogram suggests two clusters (if life were only this easy).

Identifying Clusters with Dendrograms

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





Hierarchical Clustering

Which tree provides the "best" clustering? How many trees can we build with *n* data points.

Hierarchical Clustering

Which tree provides the "best" clustering? How many trees can we build with *n* data points.   There are (2n -3)! / [2$^{(n-2)}$(n-2)!]

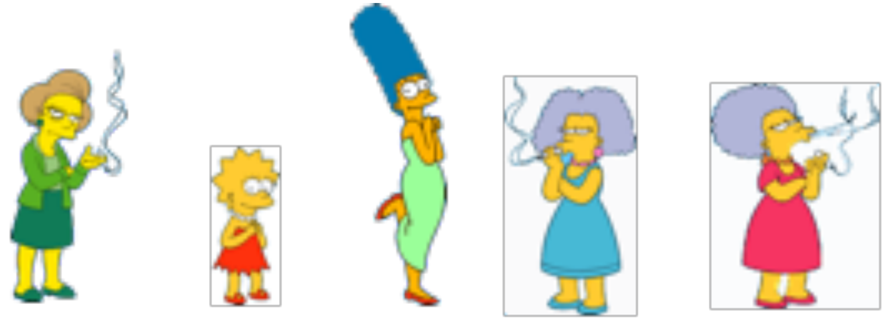
# of leaves	# of possible Dendrograms
2	1
3	3
4	15
5	105
…	…
10	34,459,425

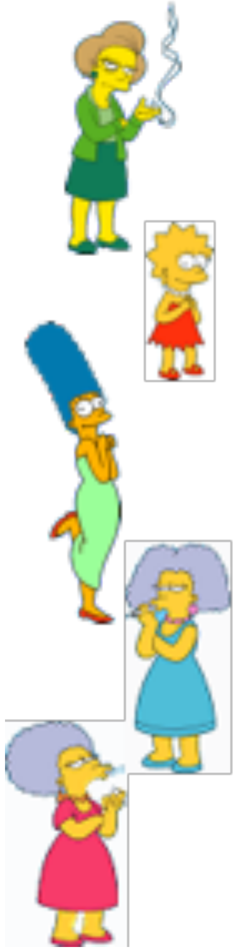
Top-Down (divisive): Starting with all the data in a single cluster, consider every possible way to divide the cluster (into 2 pieces). Chose the best decision and recursively do the same on both sides.

Bottom-up (aggomerative): 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.

Distance Matrix

Construct a distance matrix which contains the distances between every pair of objects.



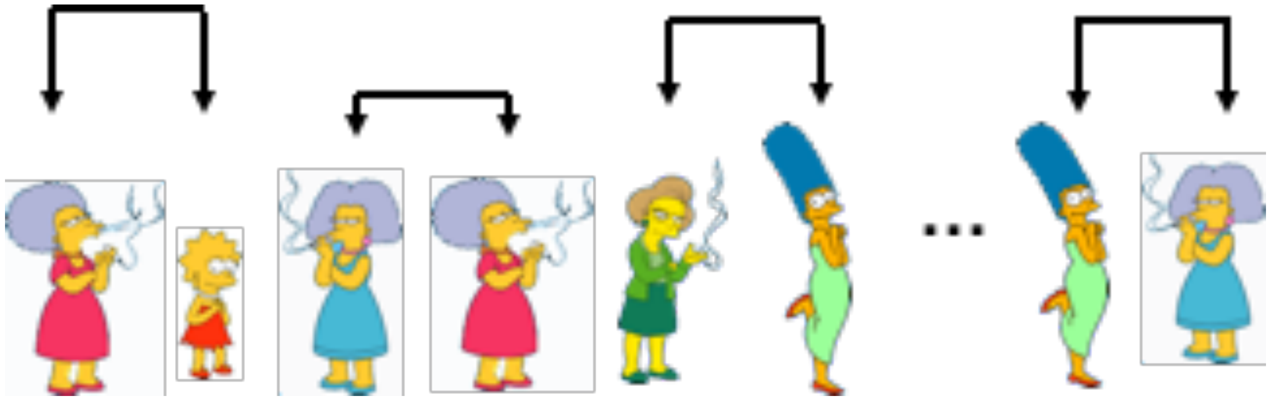


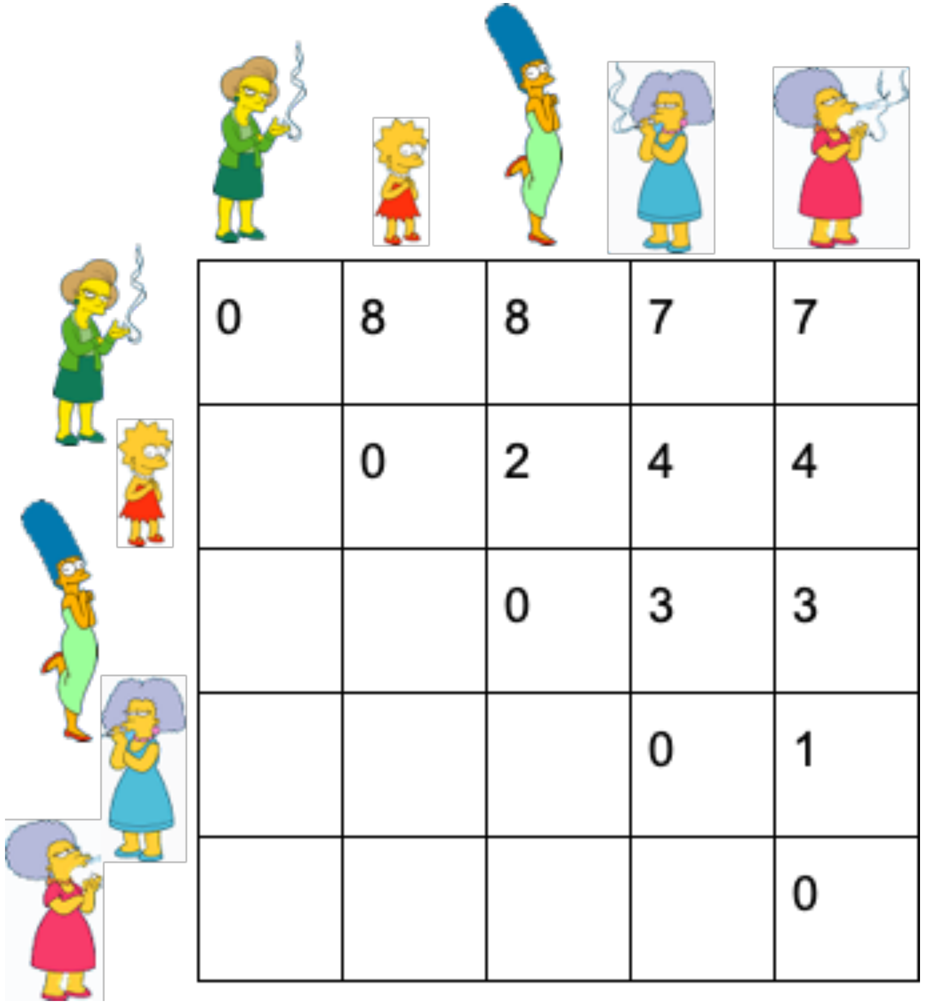
0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

Bottom-Up Example

Start with each item in its own cluster, find the best pair (most similar, smallest distance) to merge into a new cluster. Repeat.

Consider all possible merges (how many in the first round):



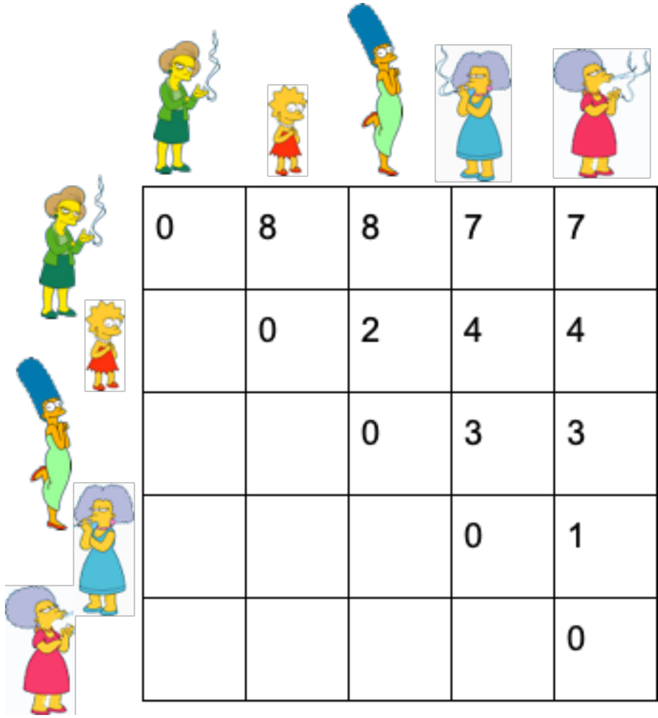


0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0

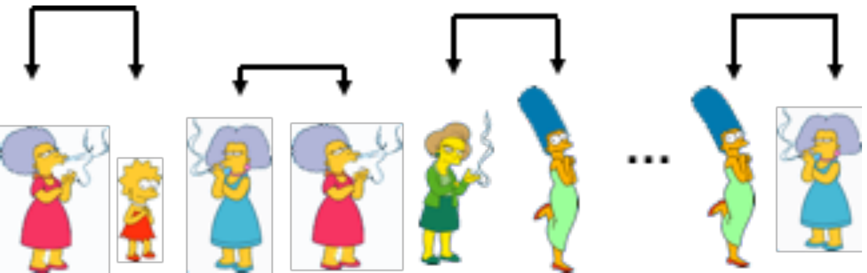
Bottom-Up Example

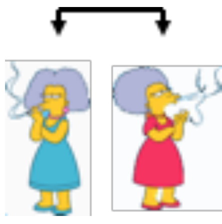
Start with each item in its own cluster, find the best pair (most similar, smallest distance) to merge into a new cluster. Repeat.

Consider all possible merges (how many in the first round):



0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0





Points to Cluster Distances

Distance matrix provides distances between all points. What is the distance between an object and a cluster?

Single linkage (nearest neighbor): Distance between clusters is the distance of the two closest objects in the different clusters.

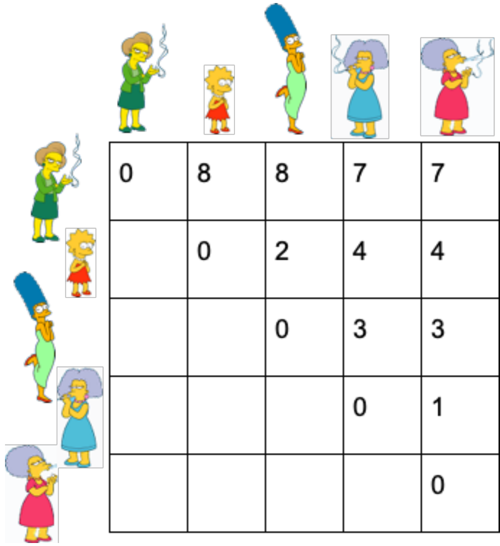
Complete linkage (furthest neighbor): Distance between clusters is the greatest distance between any two objects in the different clusters.

Group average linkage: average distance between all pairs of objects in the different clusters..

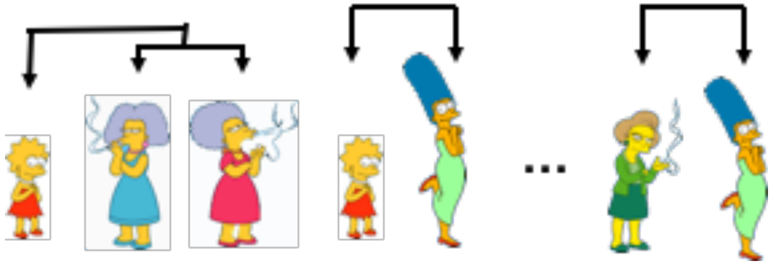
Wards linkage: Minimize the variance of the merged clusters.

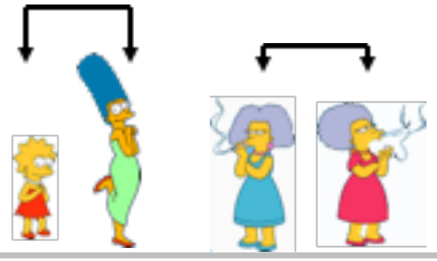
Bottom-Up Example

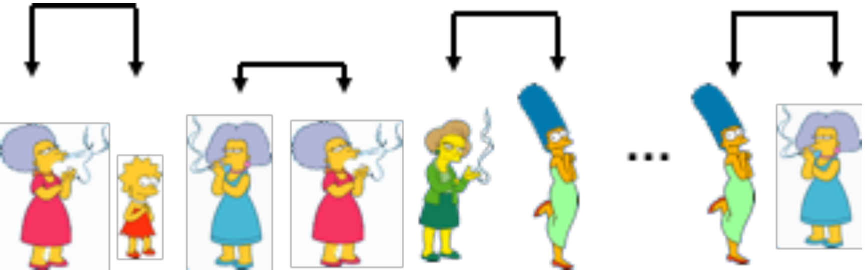
Start with each item in its own cluster, find the best pair (most similar, smallest distance) to merge into a new cluster. Repeat.

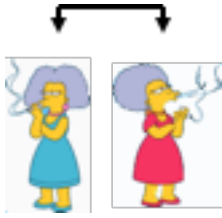


0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0



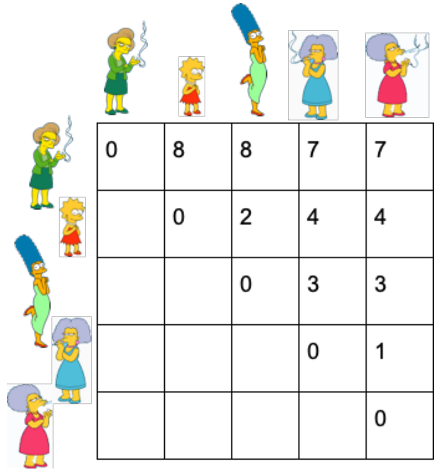




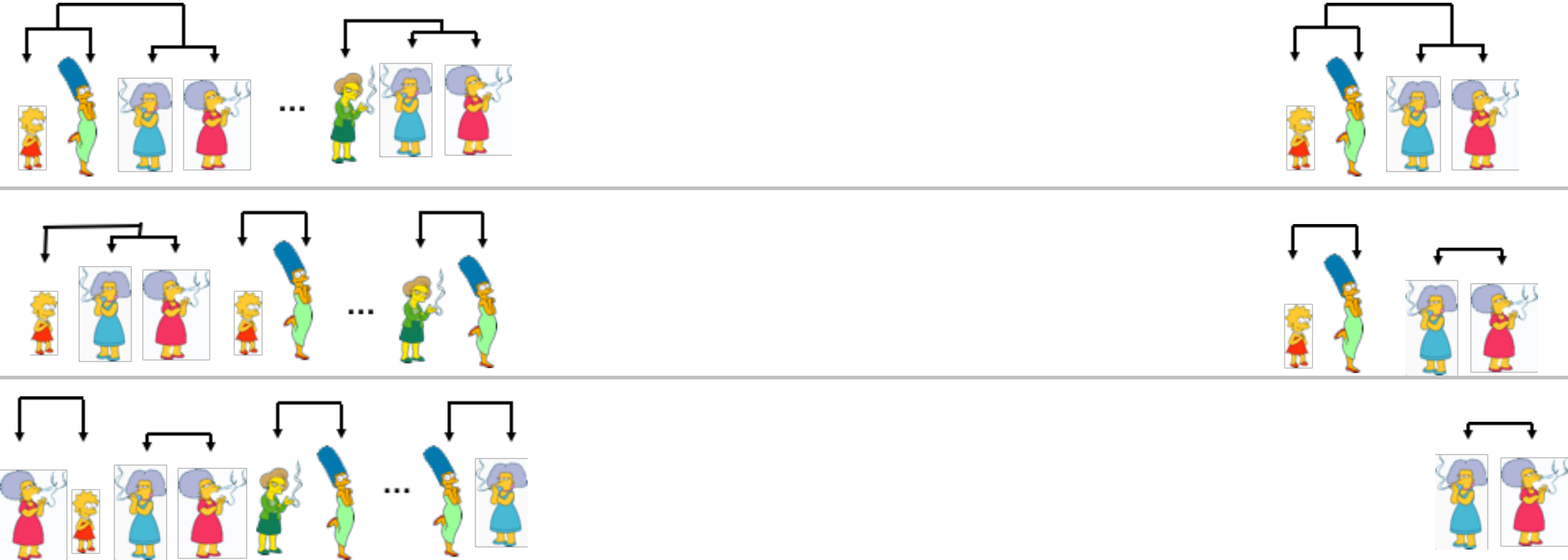


Bottom-Up Example

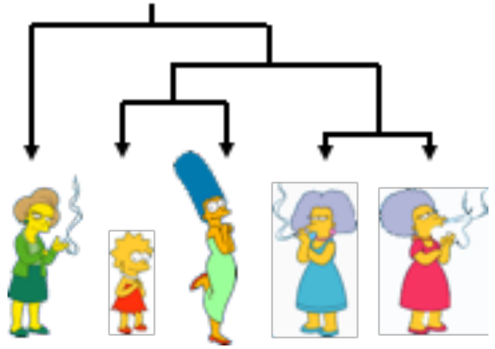
Start with each item in its own cluster, find the best pair (most similar, smallest distance) to merge into a new cluster. Repeat.



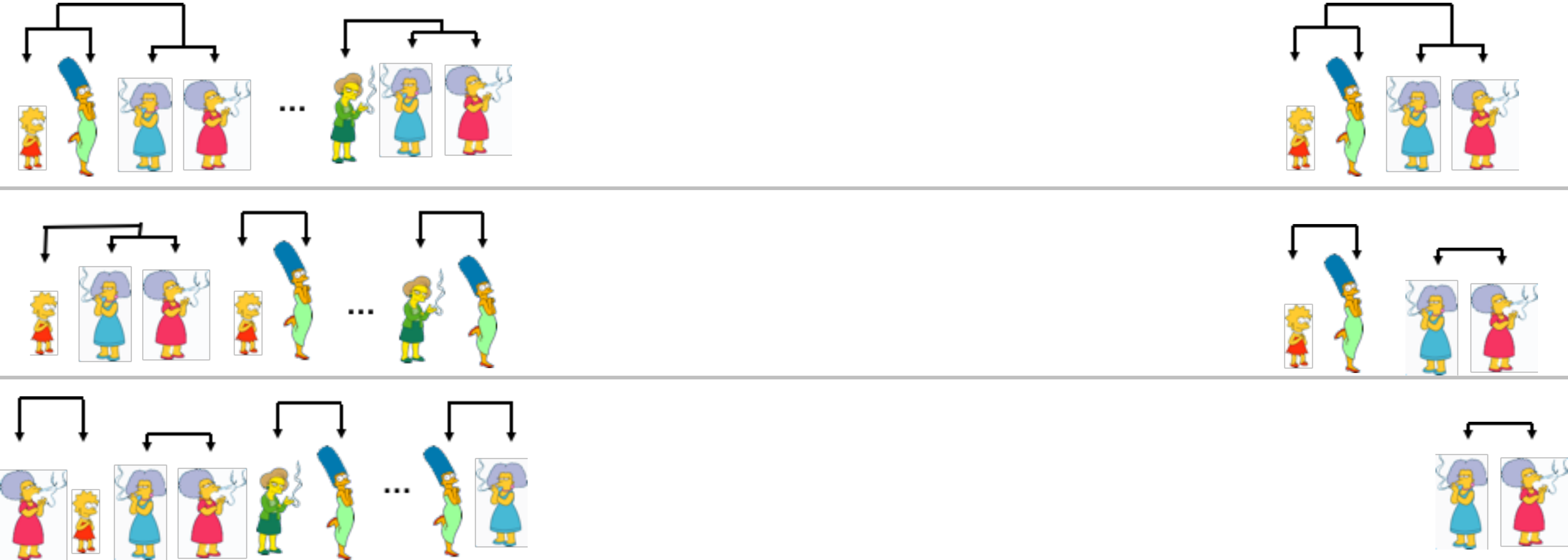
0	8	8	7	7
	0	2	4	4
		0	3	3
			0	1
				0



Bottom-Up Example



Start with each item in its own cluster, find the best pair (most similar, smallest distance) to merge into a new cluster. Repeat.



Hierarchical Clustering Summary

Advantages:

- Number of clusters is not a hyperparameter
- Hierarchical nature maps nicely into human intuition of domains

Distadvantages:

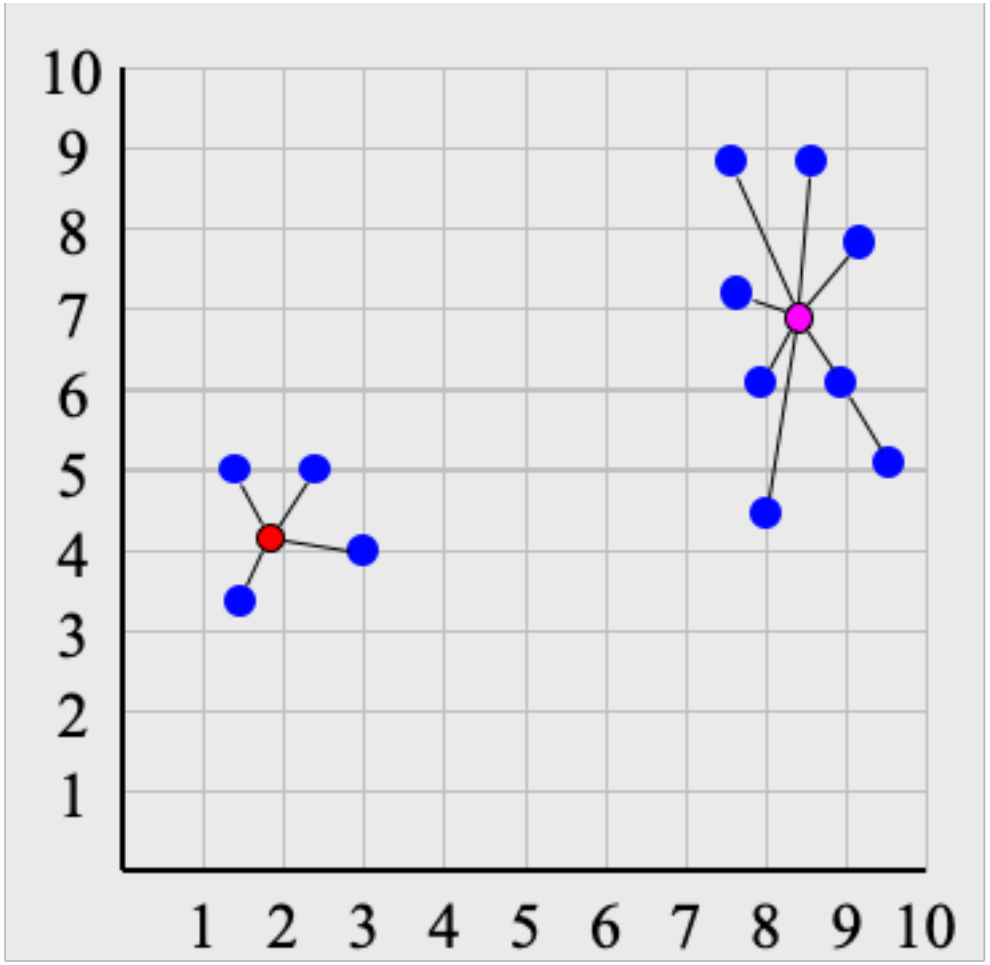
- Does not scale well (expensive to compute)
- Heuristic search finds local optima.

Partitional Clustering

- Each instance is placed in exactly one of K non-overlapping clusters.
- Hyperparameter K is defined by the user.

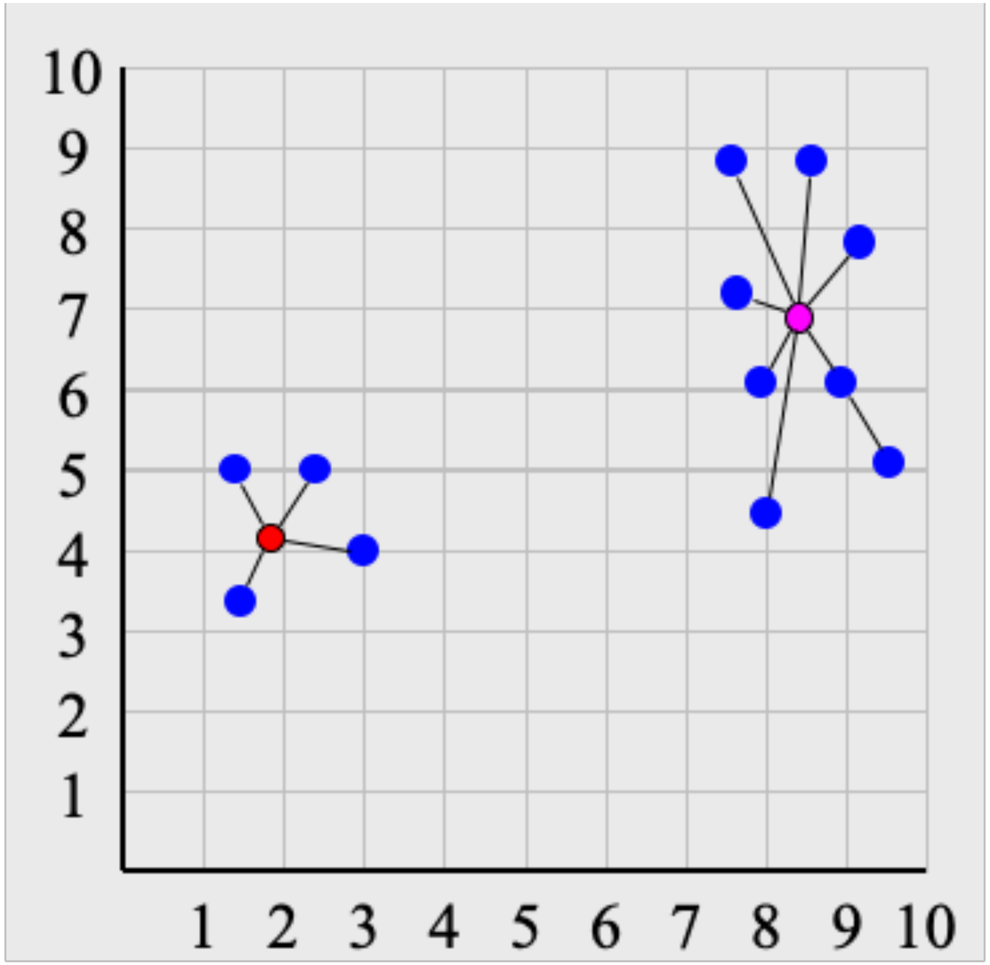
A centroid is the "mean" location of each cluster (shown in red). It is NOT one of the data points (it is computed by the clustering algorithm).

The residual sum of squares (RSS) is the sum of the squared distances from the data points to the respected centroid. Partitional clustering attempts to minimize the RSS.



K-Means Clustering

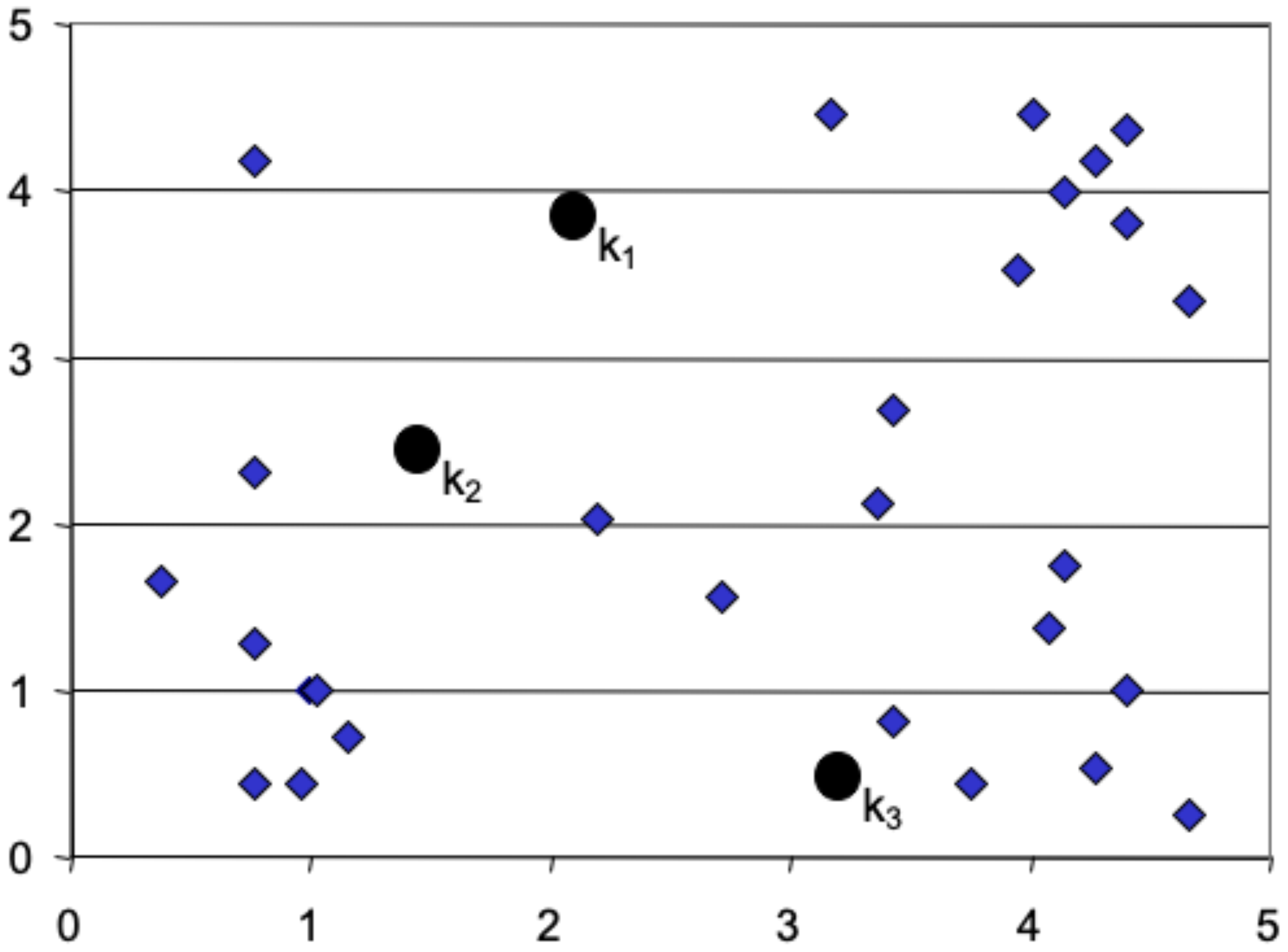
- Partitional clustering
- Each cluster has a centroid
- Number of clusters defined in advanced (it is a hyperparameter)
- Optimal K-Means clustering is NP-complete, thus, we will use an approximation method known as Llyod's algorithm



K-Means Clustering Step 1

Algorithm: k-means (Lloyds) distance metric: Euclidean k = 3

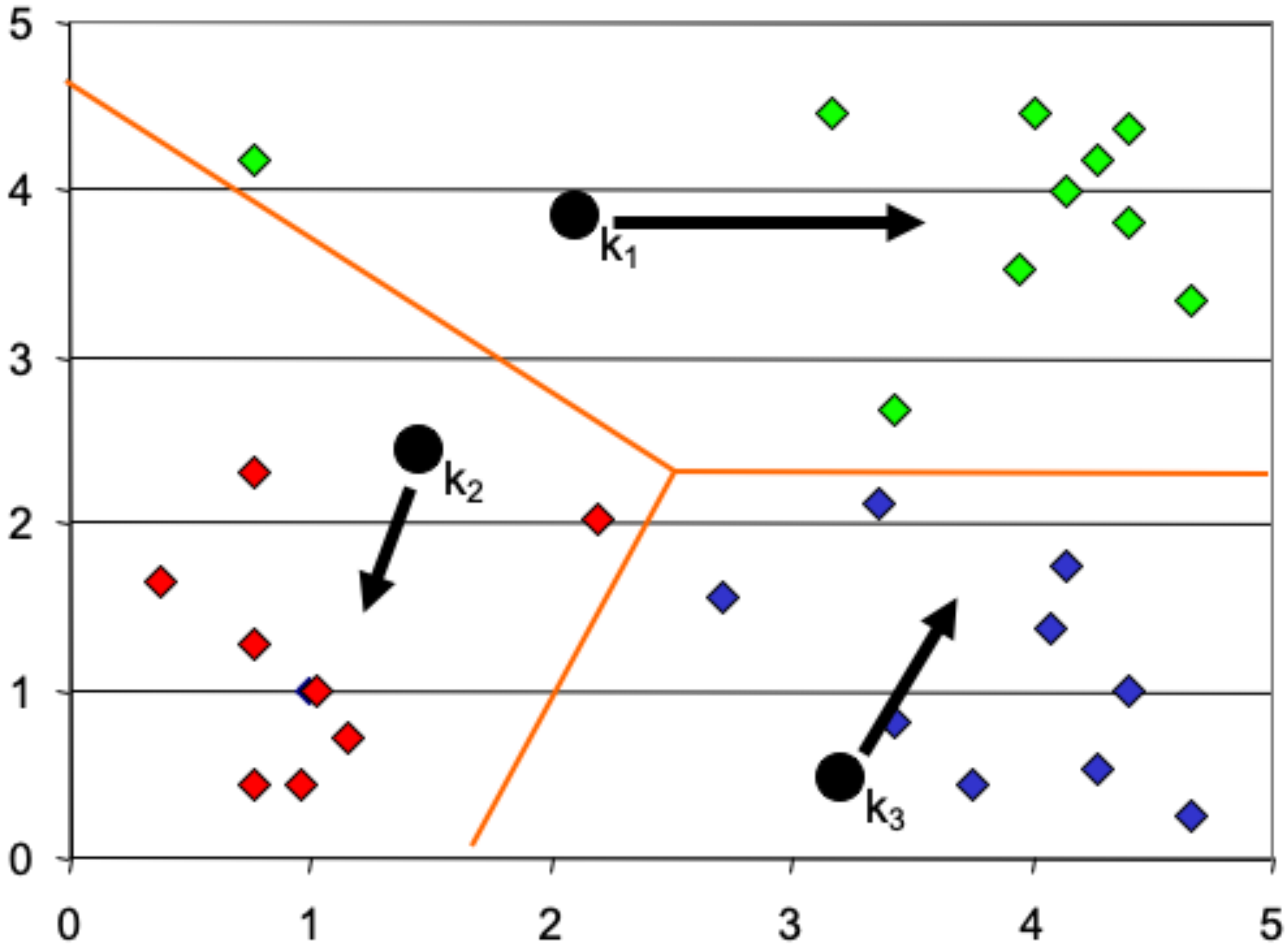
Randomly place the centroid (k_1, k_2, k_3)



K-Means Clustering Step 2

Algorithm: k-means (Lloyds) distance metric: Euclidean k = 3

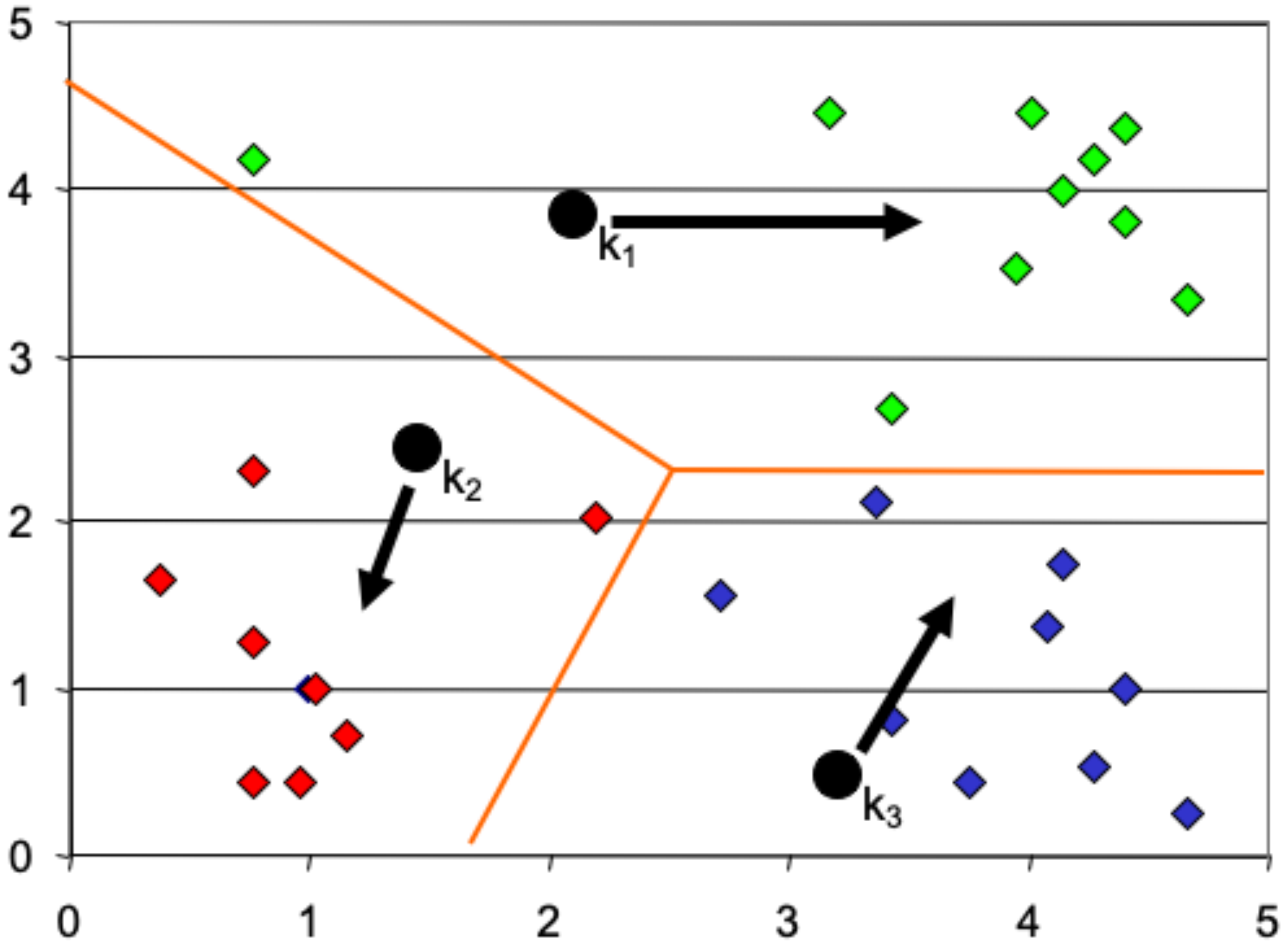
1. Randomly place the centroid (k_1, k_2, k_3)
2. Color the points according to the centroid that they are closest. If no points are colored differently, stop.



K-Means Clustering Step 3

Algorithm: k-means (Lloyds) distance metric: Euclidean k = 3

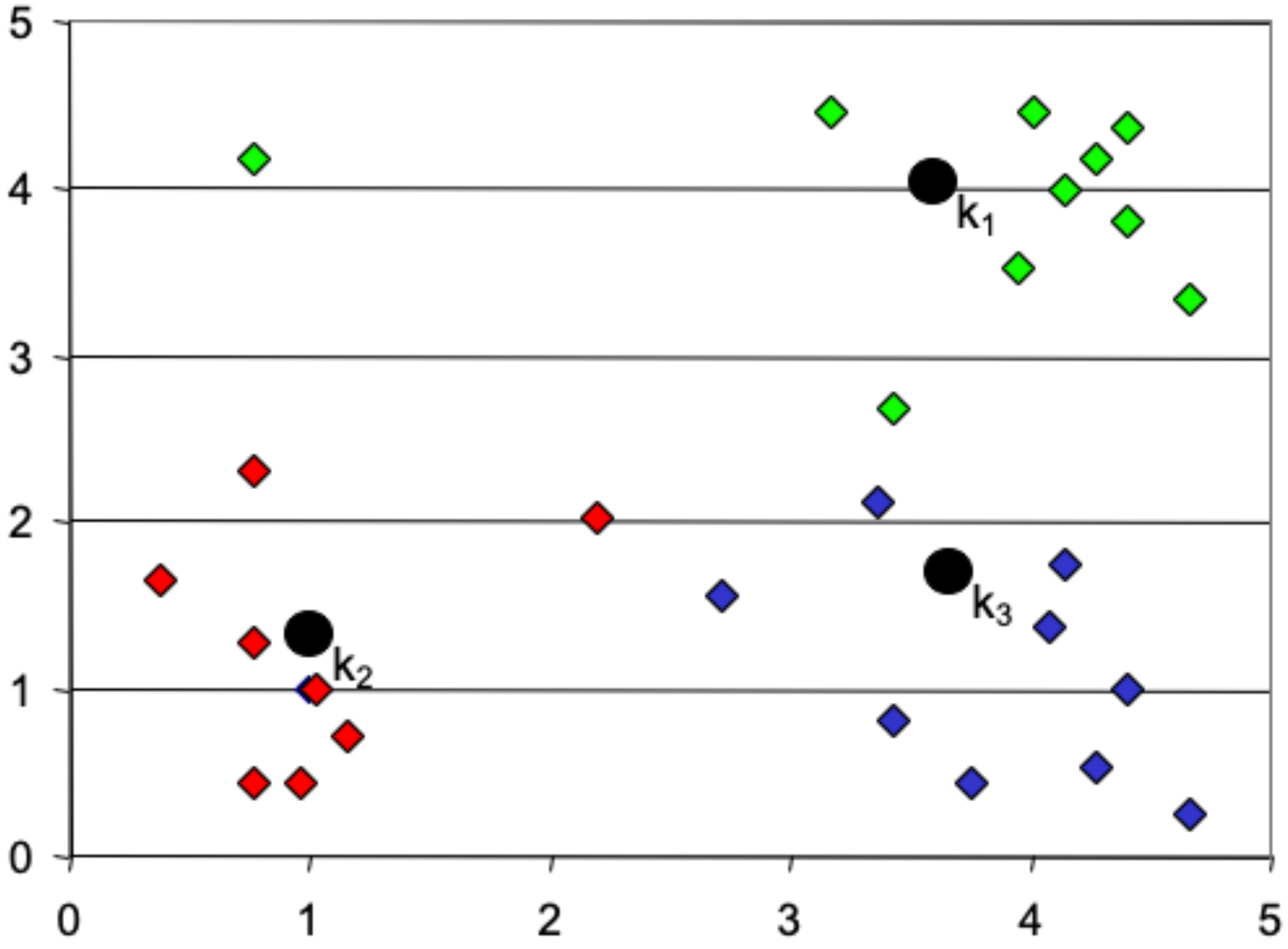
1. Randomly place the centroid (k_1, k_2, k_3)
2. Color the points according to the centroid that they are closest. If no points are colored differently, stop.
3. Recompute the centroids



K-Means Clustering Step 4

Algorithm: k-means (Lloyds)   distance metric: Euclidean   k = 3

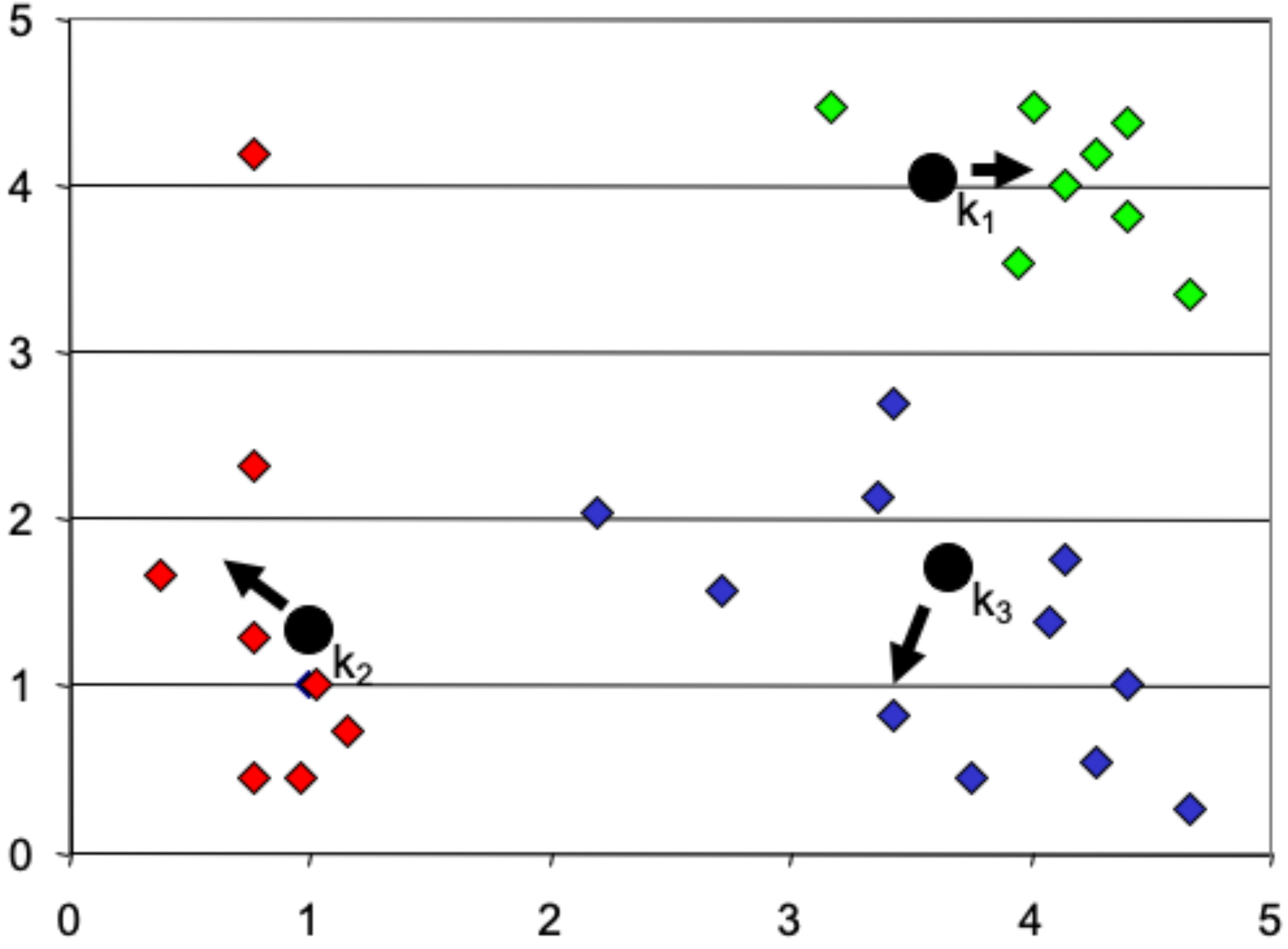
1. Randomly place the centroid (k_1, k_2, k_3)
2. Color the points according to the centroid that they are closest. If no points are colored differently, stop.
3. Recompute the centroids
4. Jump to step 2



K-Means Clustering Step 5

Algorithm: k-means (Lloyds)   distance metric: Euclidean   k = 3

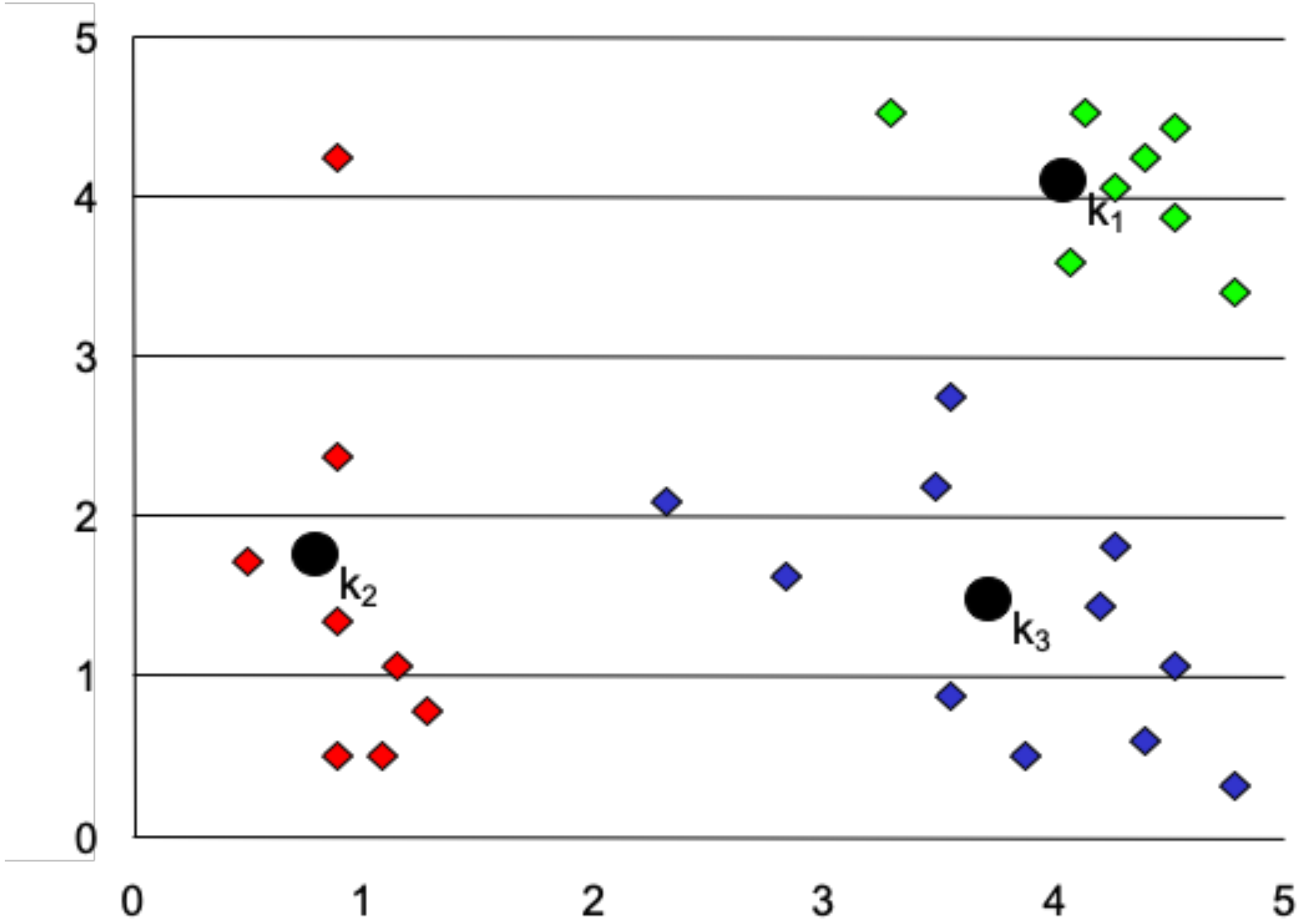
1. Randomly place the centroid (k_1, k_2, k_3)
2. Color the points according to the centroid that they are closest. If no points are colored differently, stop.
3. Recompute the centroids
4. Jump to step 2



K-Means Clustering Step 6

Algorithm: k-means (Lloyds)   distance metric: Euclidean   k = 3

1. Randomly place the centroid (k_1, k_2, k_3)
2. Color the points according to the centroid that they are closest. If no points are colored differently, stop.
3. Recompute the centroids
4. Jump to step 2



K-Means Clustering

Advantages:

- Relatively efficient (especially if you can store the distance matrix in memory)

Disadvantages:

- Results dependent on the initial centroids
- How do we define the "mean" of categorical data?
- Need to specify k in advance.
- Unable to handle noisy data and outliers