Clustering

Ankit Gupta

Room E105

Email: ankit.gupta@vsb.cz, gupta.ankit894@gmail.com

Clustering

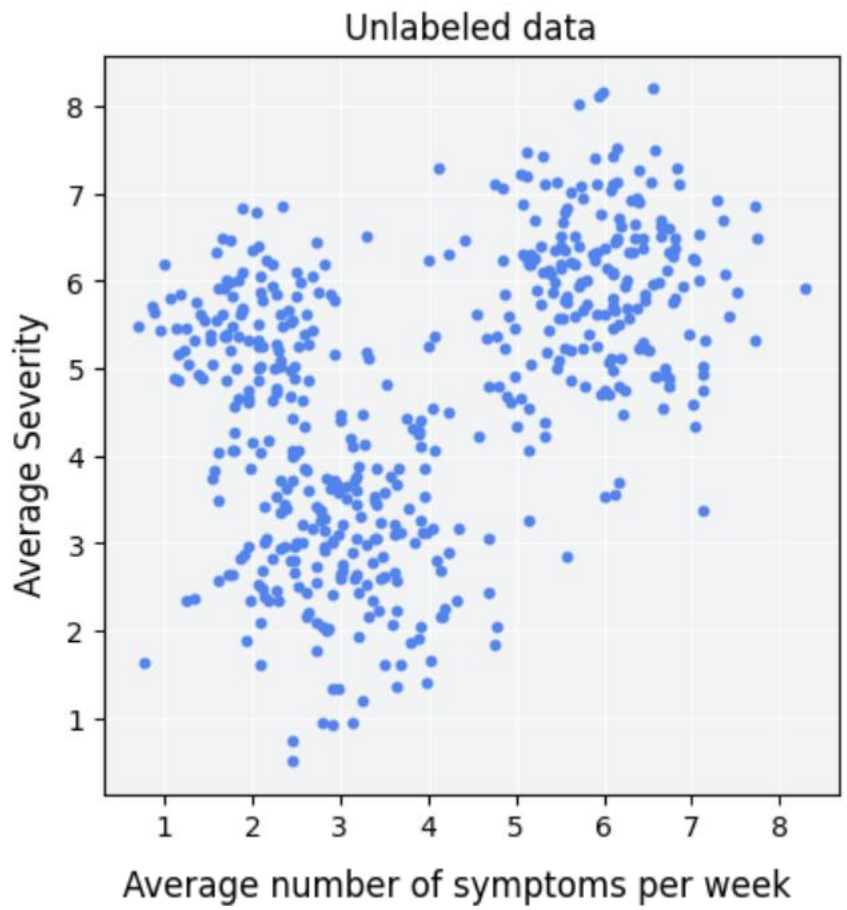
Clustering- unsupervised technique to group unlabeled samples based on similarities with the following conditions:

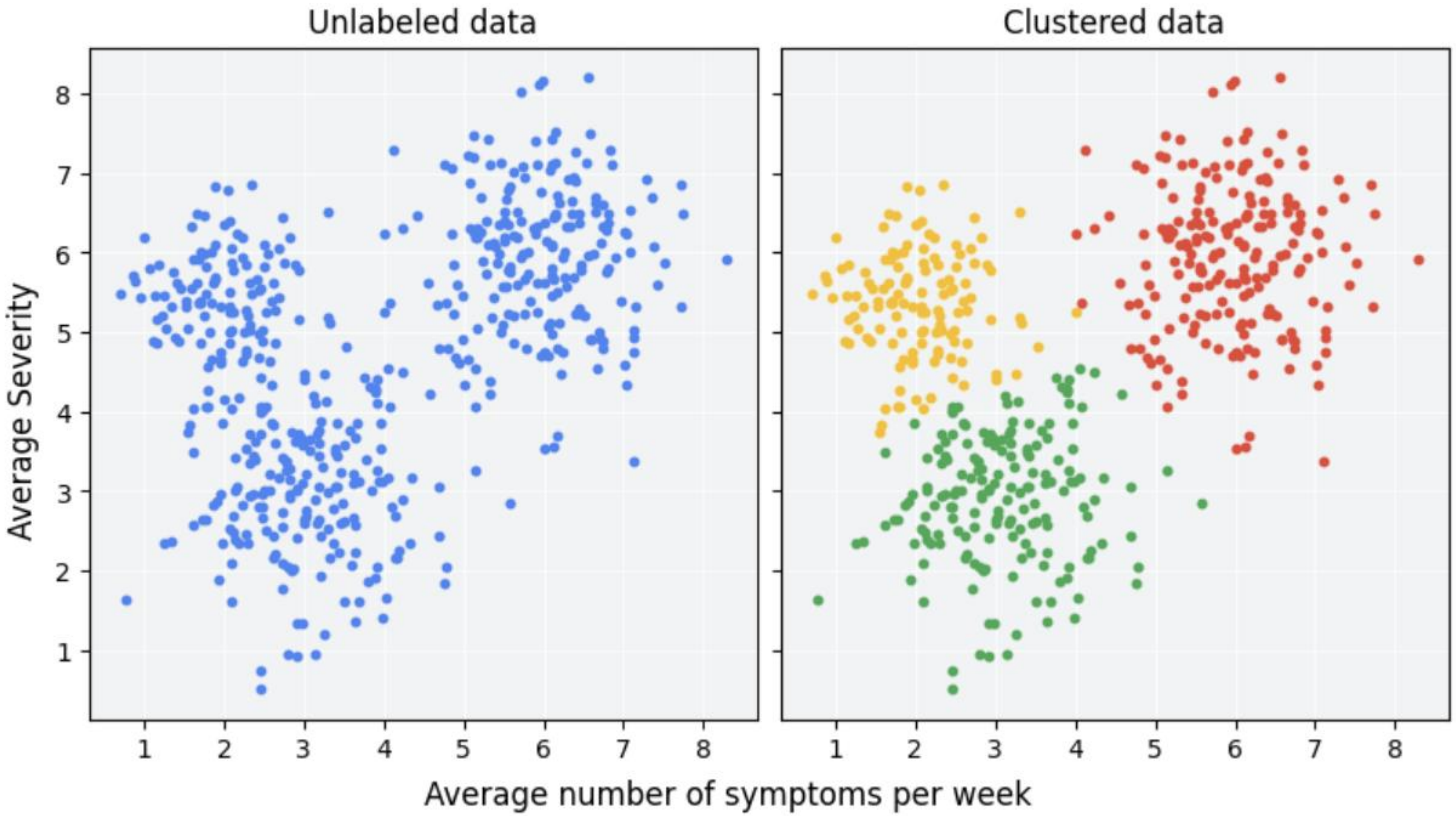
1. **High intra-cluster similarity.**
2. **Low inter-cluster similarity.**

For example:

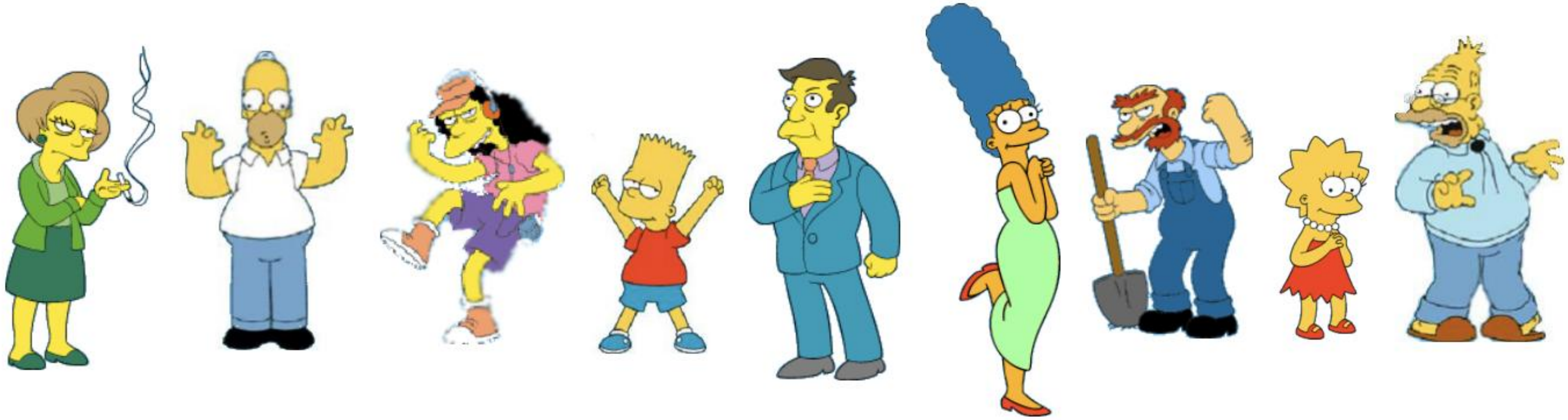
1. Patient study for evaluating a new treatment protocol.

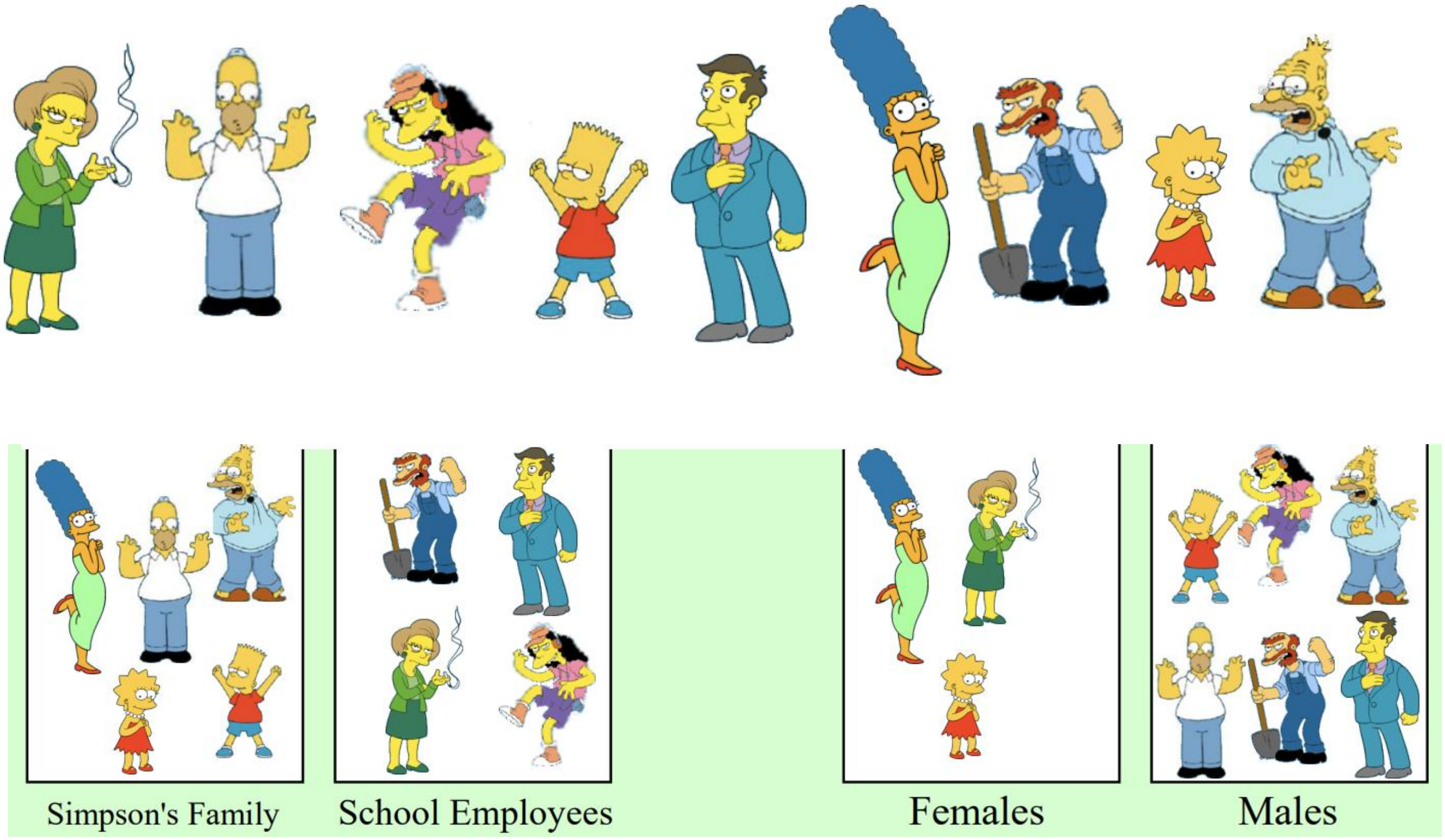
Question: How many times a patients experience symptoms and severity of symptoms?





How to reach a conclusion that this data can be clustered into n clusters?

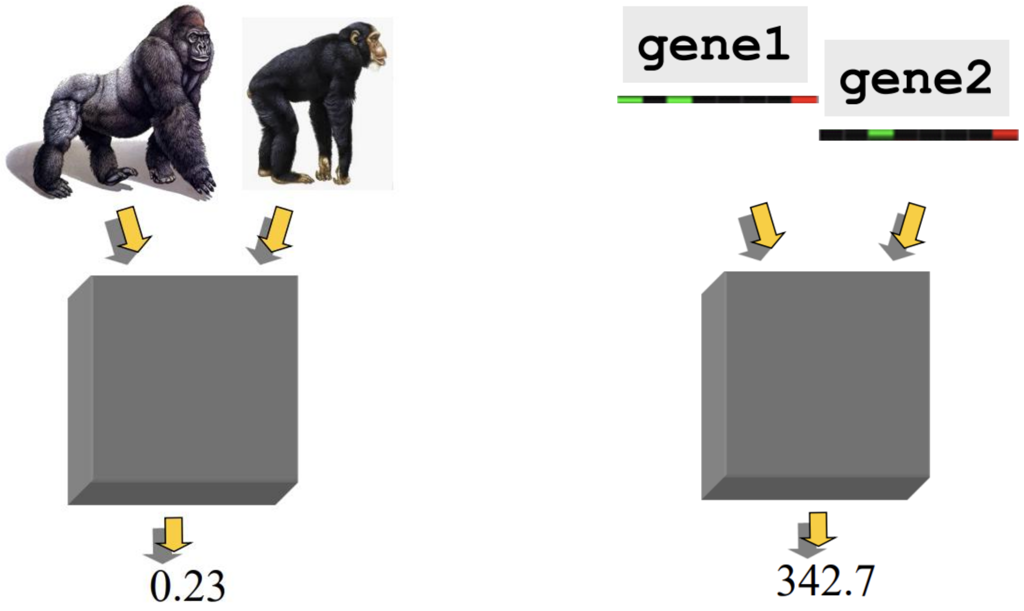



Simpson's Family
School Employees
Females
Males

What do you mean by Similarity?

- The quality or state of being similar; likeliness; resemblance; similarity of features.
- **Pragmatic Approach**

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



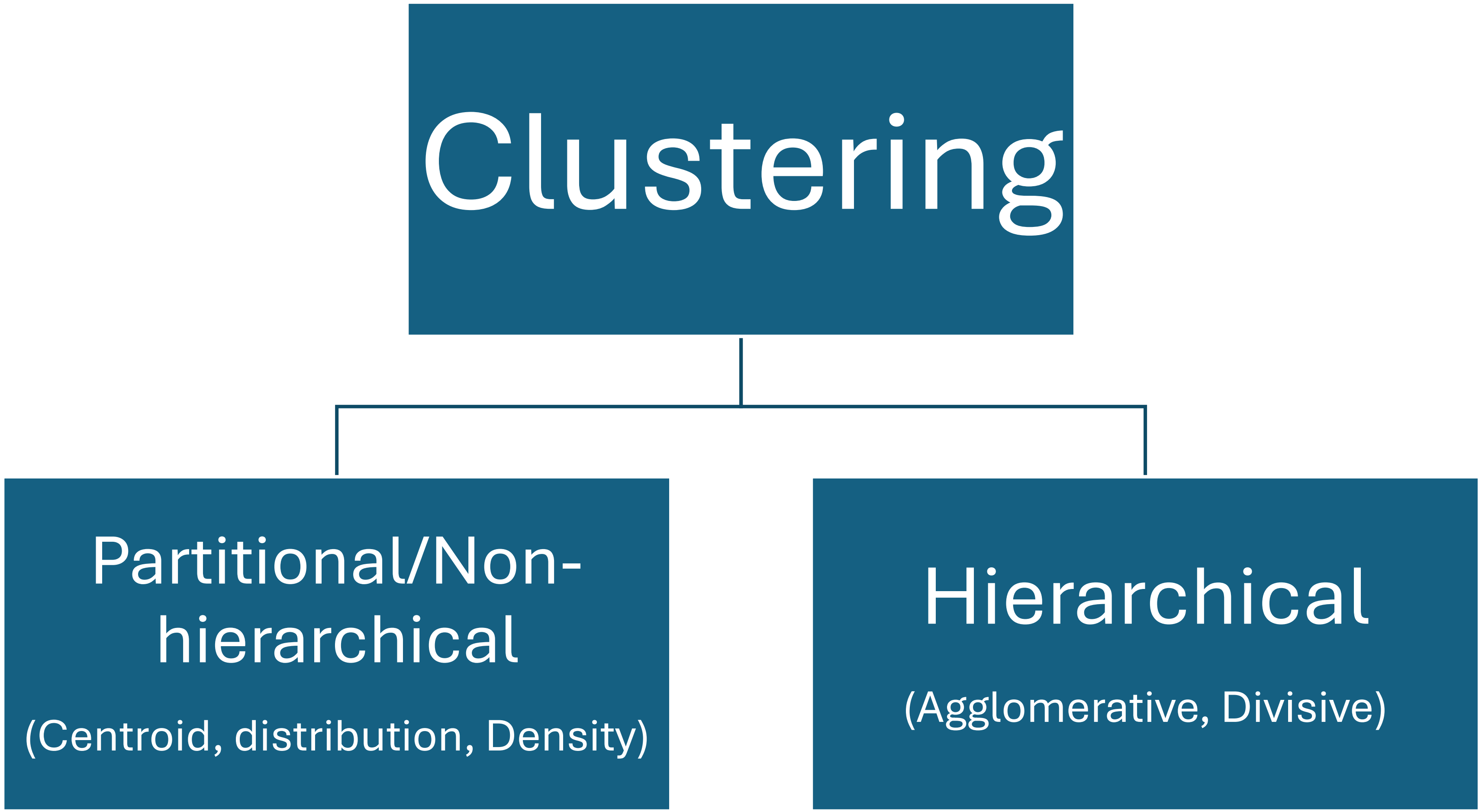
A few examples:

- Euclidian distance

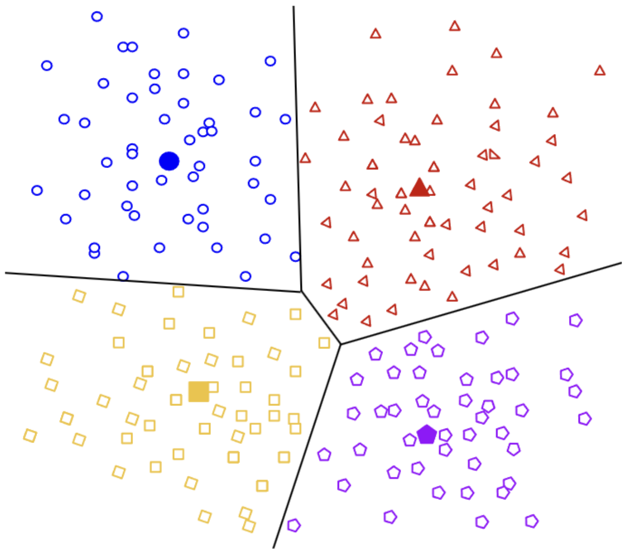
$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Correlation coefficient

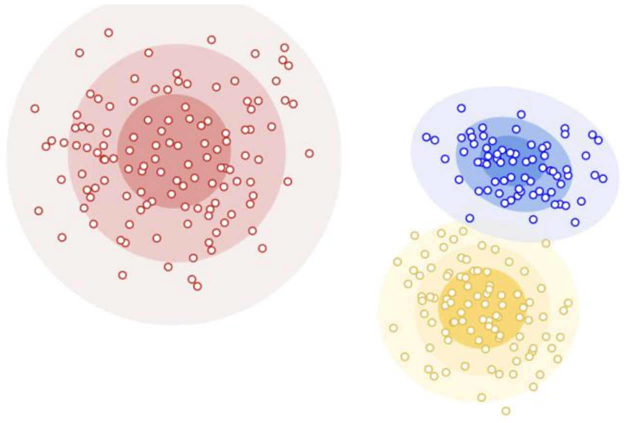
$$s(x,y) = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \sigma_y}$$


Clustering
Partitional/Non-hierarchical
(Centroid, distribution, Density)
Hierarchical
(Agglomerative, Divisive)

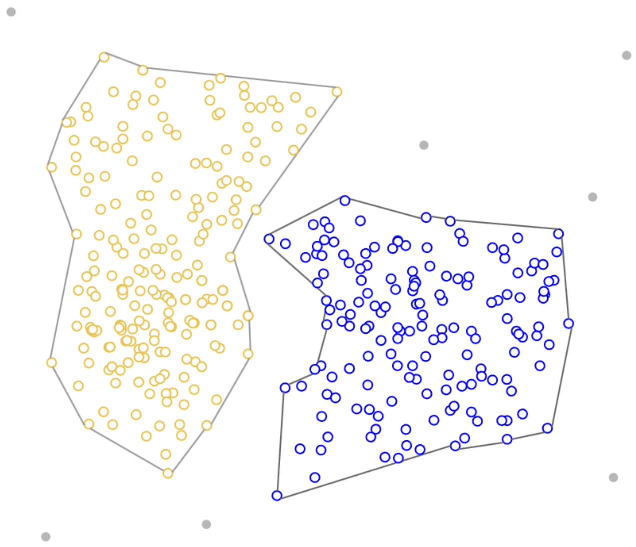
Partitional/Hierarchical Clustering



Centroid Based clustering



Distribution based clustering



Density based clustering

Algorithm *k-means*

1. Decide on a value for K, the number of clusters.

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. Repeat 3 and 4 until none of the N objects changed membership in the last iteration.

K-means Clustering

- Data={1,2,3,5,6}
- Let's assume there are three clusters, **C{i}**, **C{j}**, **C{k}** with initial centroid as **1,3**, and **5**.
- Let's say we want **{2}** to assign a **cluster** from three clusters. To do that let's calculate the **Euclidean distance between 2 and centroids of clusters**.
- E{2,1}=$\sqrt{(2-1)^2} = 1$, E{2,3}=$\sqrt{(2-3)^2 = 1}$, and E{2,5}=$\sqrt{(2-5)^2}$ =9. Let's keep 2 in the cluster with centroid 1.
- New clusters are c{i}={1,2}, C{j}={3}, C{k}={5}.

K-means Clustering

- Re-estimate cluster by calculating centroid:
  - The centroid of **C{i}=(1+2)/2=1.5**
  - The centroid of **C{j}=3** (No change since there is only one element in the set).
  - The centroid of **C{k}=5** (No change since there is only one element in the set).

- Let's try to assign {6} a cluster.
- E{6,1.5}=$\sqrt{(6-1.5)^2} = 4.5$, E{6,3}=$\sqrt{(6-3)^2} = 3$, and E{6,5}=$\sqrt{(6-5)^2}$ =1. 6 will be assigned to C{k}.

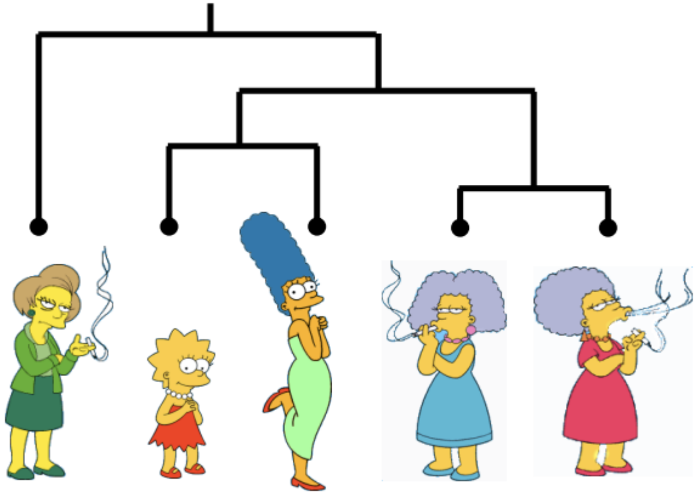
Update the centroids: C{i}=1.5, C{j}=3, C{k}=5.5. Keep on repeating these steps, until elements keep on shifting from one cluster to another.

Final clusters C{i}={1,2}, C{j}=3, c{k}={5,6}.

Agglomerative Hierarchical

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

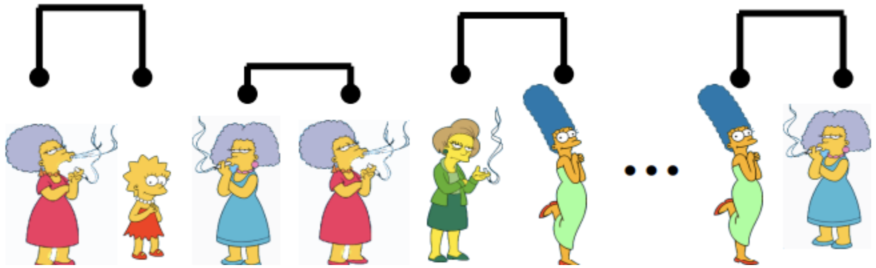


Bottom-Up (agglomerative): Starting with each item in its own cluster, find the best pair to merge into a new cluster. Repeat until all clusters are fused together.

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.

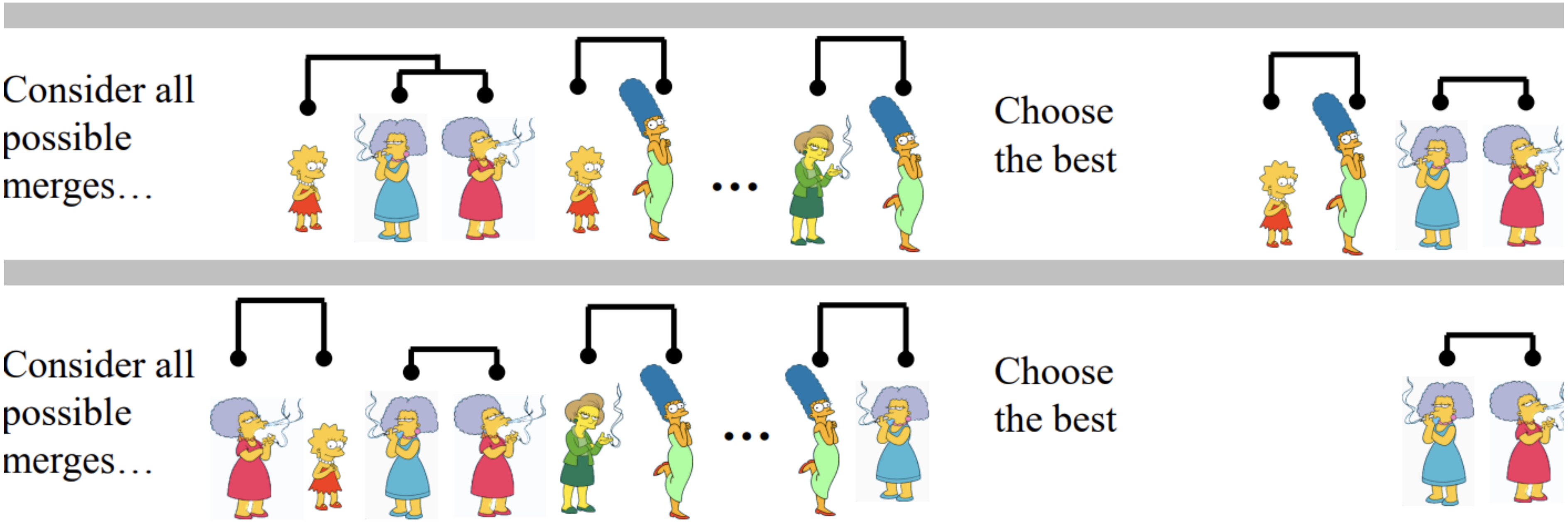
Consider all possible merges…



Choose the best

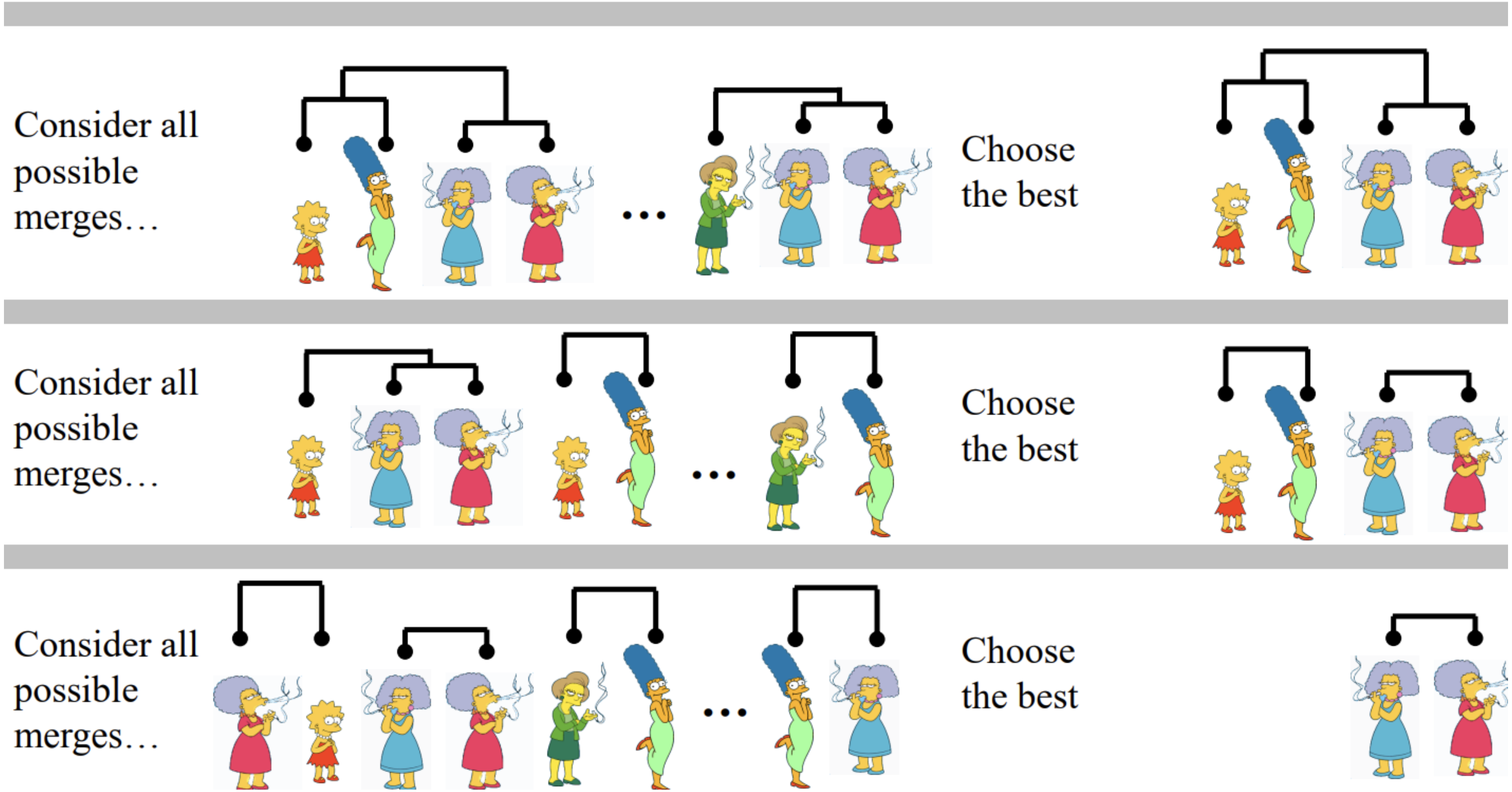
Bottom-Up (agglomerative):

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.

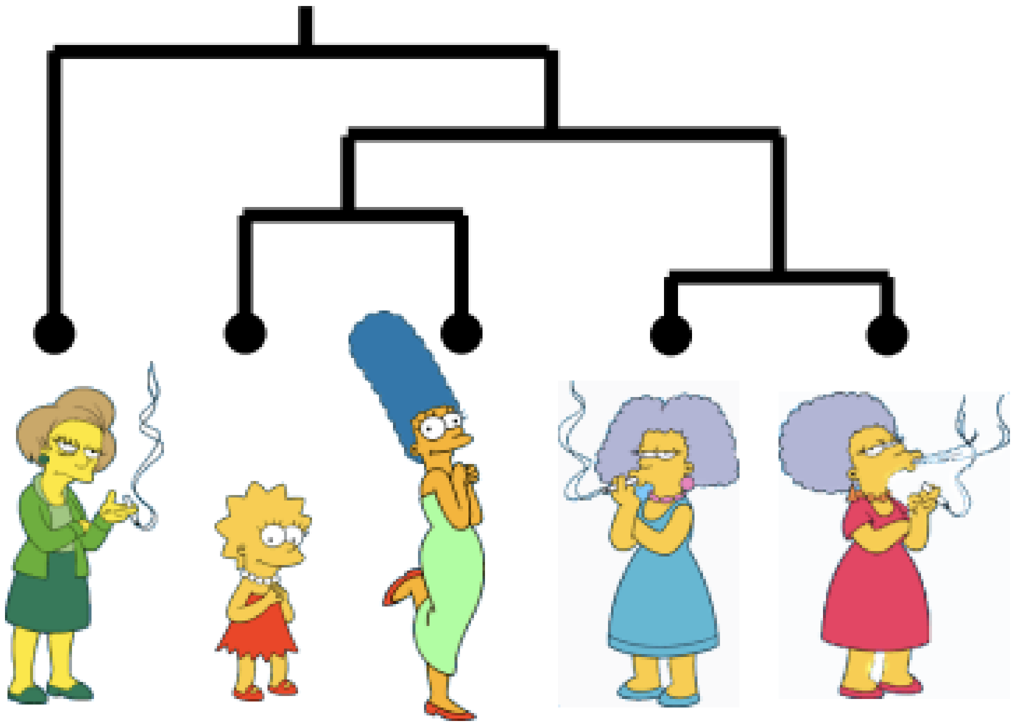


Bottom-Up (agglomerative):

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.



Final Result



How to do it with clusters?

Problem

Distance matrix

	1	2	3	4	5
1	0	2	6	10	9
2	2	0	3	9	8
3	6	3	0	7	5
4	10	9	7	0	4
5	9	8	5	4	0

Rule: Cluster points based on minimum distance

D{(i,j),k}=min (D{i,j}, D{i,k})

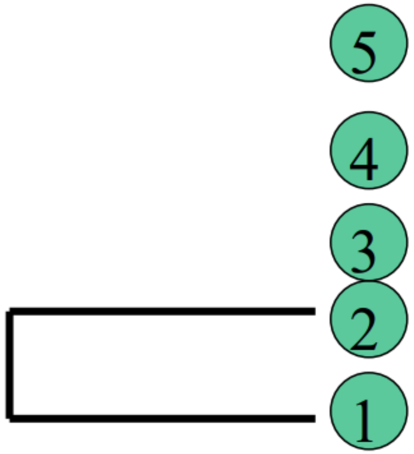
Example:D{(1,2),3} = min(D{1,3}, D{2,3})
= min(6,3)
=3

Agglomerative Clustering

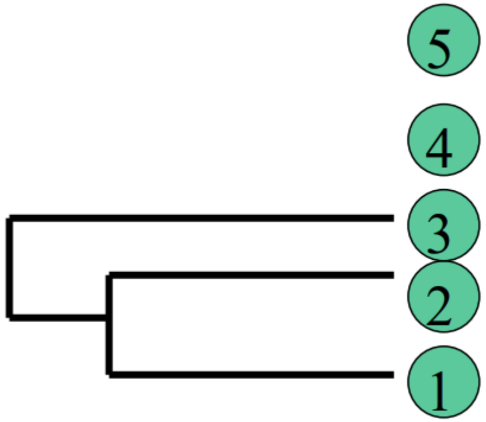
	1,2	3	4	5
1,2	0	3	9	8
3	3	0	7	5
4	9	7	0	4
5	8	5	4	0

	1,2,3	4	5
1,2,3	0	7	5
4	7	0	4
5	5	4	0

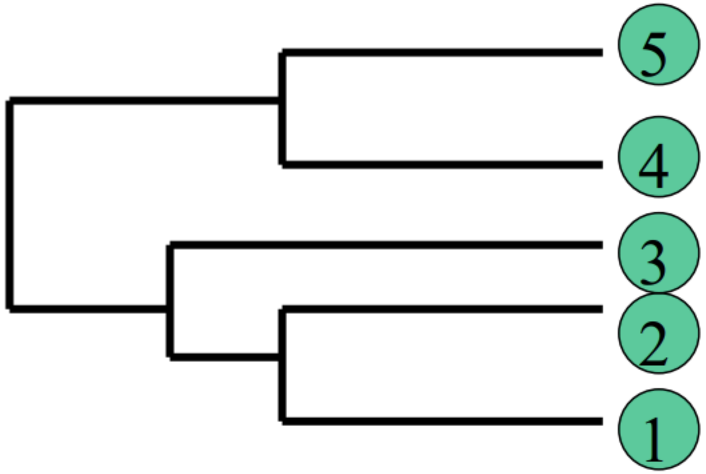
	1,2,3	4,5
1,2,3	0	5
4,5	5	0



1.



2.



3.

Divisive Clustering (Top Down)

- It is just **opposite of agglomerative clustering** (Split the points based on dissimilarity).
- Rules:
  - Start with **all data points in one cluster**.
  - **Divide the cluster into two smaller clusters** by finding dissimilar points. Repeat the process: For each of the new clusters, repeat the splitting process:
    - C**hoose the cluster with the most dissimilar points** (opposite to what we do in agglomerative clustering).
    - **Split it again** into two smaller clusters.

Divisive Clustering in Action (Same Example)

- Cluster splitting:

	1	2	3	4	5
1	**0**	**2**	**6**	**10**	**9**
2	**2**	**0**	**3**	**9**	**8**
3	**6**	**3**	**0**	**7**	**5**
4	**10**	**9**	**7**	**0**	**4**
5	**9**	**8**	**5**	**4**	**0**

Let's assume that cluster C{i} has all points:
C{i}={1,2,3,4,5} and another set C{j}={}

Calculate dissimilar points using average distance:
Dist{1}=(Dist{1,2}+Dist{1,3}+Dist{1,4}+Dist{1,5})/4
Dist{1} = (2+6+10+9)/4 = 6.75
Dist{2} = (2+3+9+8)/4 = 5.5
Dist{3} = (6+3+7+5)/4 = 5.25
Dist{4} = (10+4+7+4)/4 = 7.5
Dist{5} = (9+8+5+4)/4 = 6.5

Most dissimilar point is {4}, since it has the maximum distance.

C{i}={1,2,3,5}, C{j}={4}.

Divisive Clustering in Action (Same Example)

- Cluster splitting:

	1	2	3	4	5
1	**0**	**2**	**6**	**10**	**9**
2	**2**	**0**	**3**	**9**	**8**
3	**6**	**3**	**0**	**7**	**5**
4	**10**	**9**	**7**	**0**	**4**
5	**9**	**8**	**5**	**4**	**0**

Calculating dissimilar points calculates the
mean distance between intra-set points (C{i})– distance of current point with the points in another set (C{j}).

Calculate dissimilar points using average distance:
Dist{1}=(Dist{1,2}+Dist{1,3}+Dist{1,5})/3-Dist{1,4}/1

Dist{1} = (2+10+9)/3 - 10 = -3
Dist{2} = (2+3+8)/3 - 9 = -4.7
Dist{3} = (6+3+5)/3 - 7 = -2.4
Dist{5} = (9+8+5)/3 - 4 = 3.3

Most dissimilar point is {4}, since it has the maximum distance.

C{i}={1,2,3}, C{j}={4,5}.

Divisive Clustering in Action (Same Example)

- Cluster splitting:

	1	2	3	4	5
1	0	2	6	10	9
2	2	0	3	9	8
3	6	3	0	7	5
4	10	9	7	0	4
5	9	8	5	4	0

If all distances are negative, split clusters based on the distance between pairs.

Calculate dissimilar points using average distance:
Dist{1}=-5.5, Dist{2}=-4, Dist{3}=-1.5.

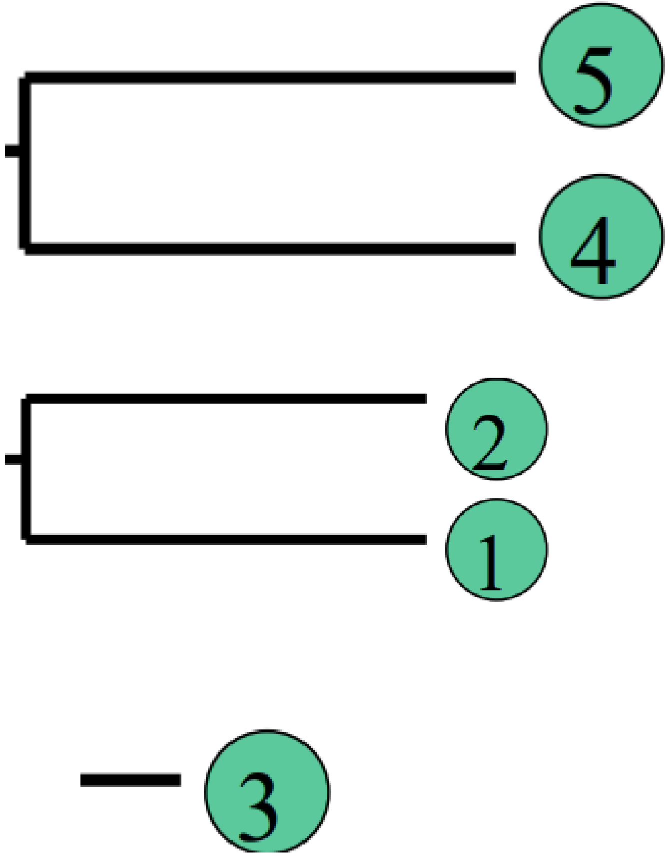
Check the distance between the points of cluster, and split

C{i}={1,2,3}, C{j}={4,5}.

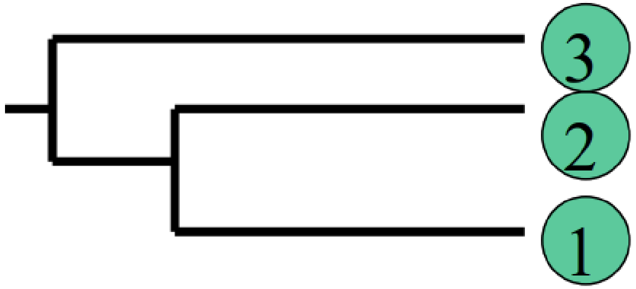
Diameter{4,5}=Dist{4,5}
= 4
Diameter({1,2,3})=max{Dist{1,2}, Dist{2,3}, Dist{1,3}}
Diameter({1,2,3})={2,3,6}
=6 (Dist{1,3})

Split {1} and {3} into two clusters
C{i}={1,2}, C{j}={4,5}, C{k}={3}.

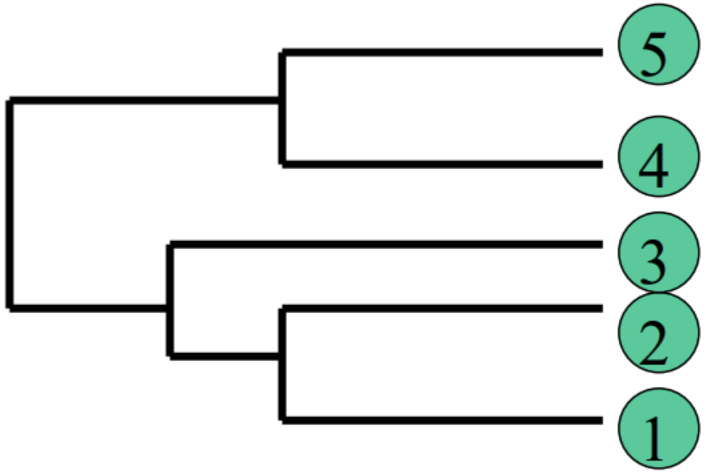
Final Result



1.



2.



3.

Let’s try some code!!!

GitHub