EXPLORE AI

Clustering Pt 1

K-means and Hierarchical clustering

14 March 2023

Objectives

By the end of this session you should be able to...

- Understand the concept of clustering
- Apply a calculation given a distance metric formula and data points
- Describe the algorithms of K-means clustering and hierarchical clustering, up to determining k clusters
- Distinguish between the different linkage methods

1 / **What is clustering?**

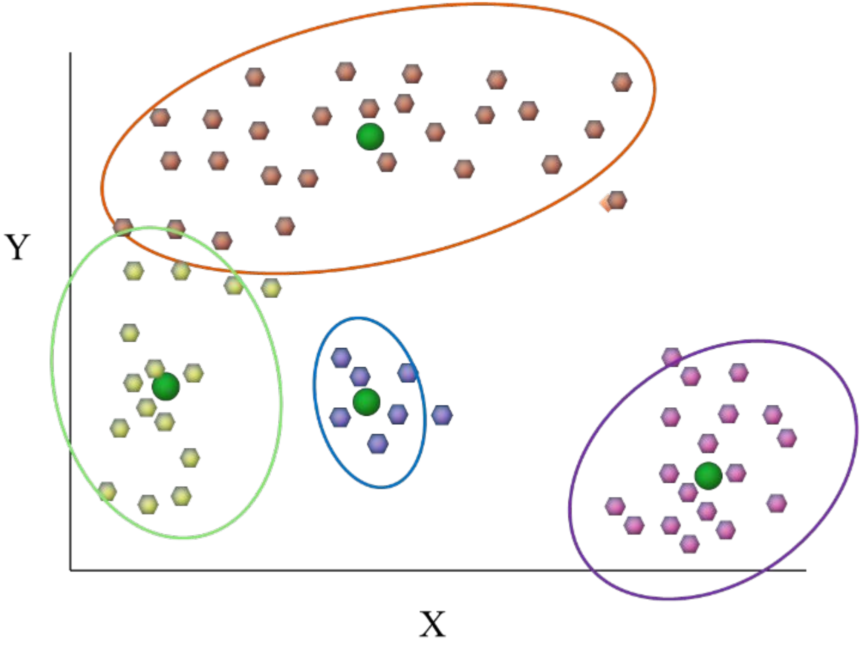
2 / **Similarity**

3 / **K-means clustering**

4 / **Hierarchical clustering**

What is Clustering?

- A cluster is a **group of similar things or people positioned or occurring closely together**
- The key idea of clustering is **similarity** and difference
- Clustering with regards to data means grouping data points based on similarities and differences in the data
  - Similar data points will belong to the same cluster



Please do not copy without permission. © ExploreAI 2023.

1 / **What is clustering?**

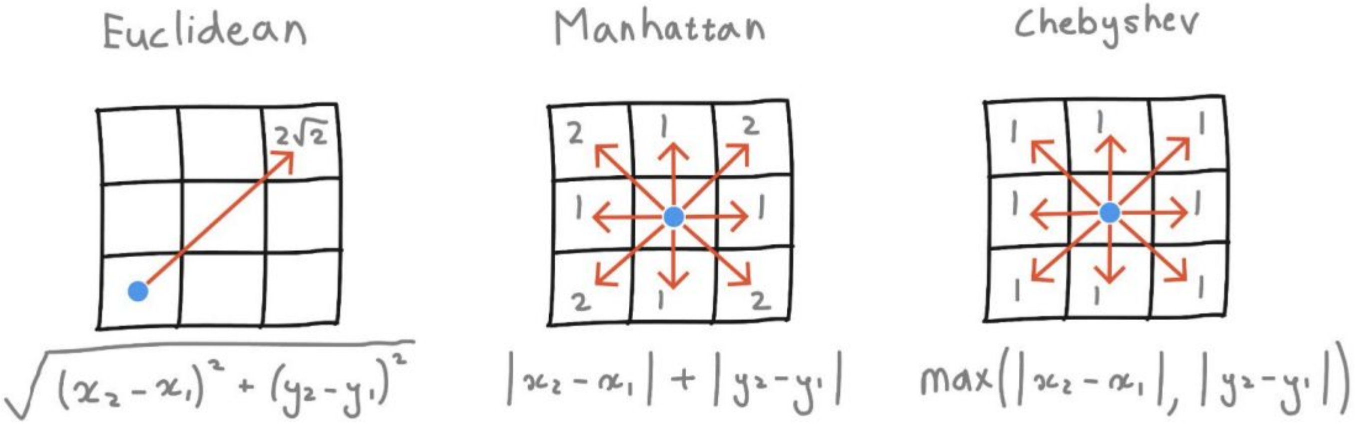
2 / **Similarity**

3 / **K-means clustering**

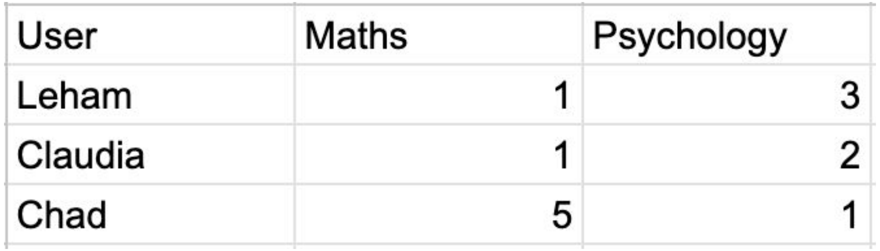
4 / **Hierarchical clustering**

Similarity

- The easiest way to determine how close (how similar) data points are, we use distance formulas
  - In this way, we're asking "what is the distance between these two points?"
- A small distance indicates they are similar, a larger distance indicates they are different



Similarity - Euclidean distance



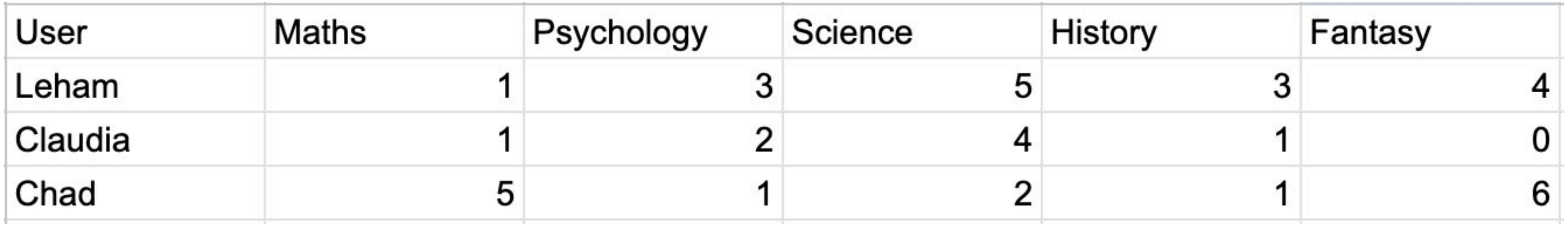
User	Maths	Psychology
Leham	1	3
Claudia	1	2
Chad	5	1

Leham - Claudia : SQRT(((1-1)^2) + ((3-2)^2)) = 1

Leham - Chad : SQRT(((1-5)^2) + ((3-1)^2)) = 4.47

Claudia - Chad : SQRT(((1-5)^2) + ((2-1)^2)) = 4.12

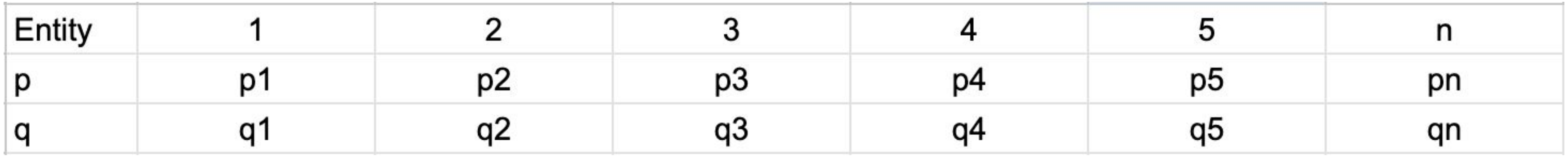
Similarity



User	Maths	Psychology	Science	History	Fantasy
Leham	1	3	5	3	4
Claudia	1	2	4	1	0
Chad	5	1	2	1	6

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}.$$

Similarity



Entity	1	2	3	4	5	n
p	p1	p2	p3	p4	p5	pn
q	q1	q2	q3	q4	q5	qn

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2}.$$

1 / **What is clustering?**

2 / **Similarity**

3 / **K-means clustering**

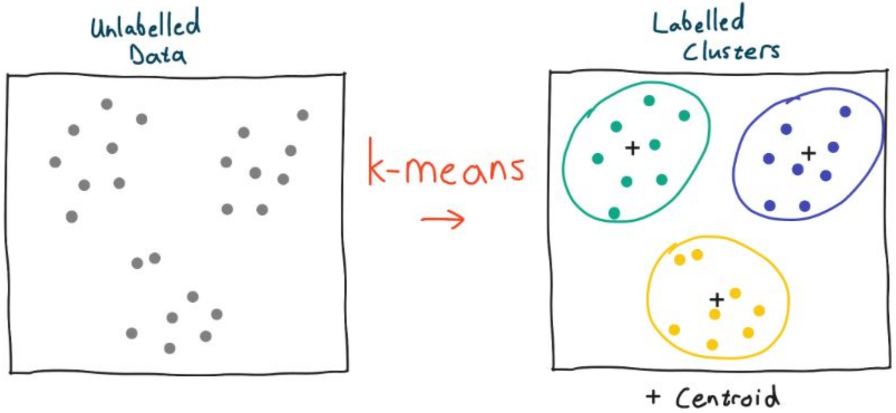
4 / **Hierarchical clustering**

K-means Clustering

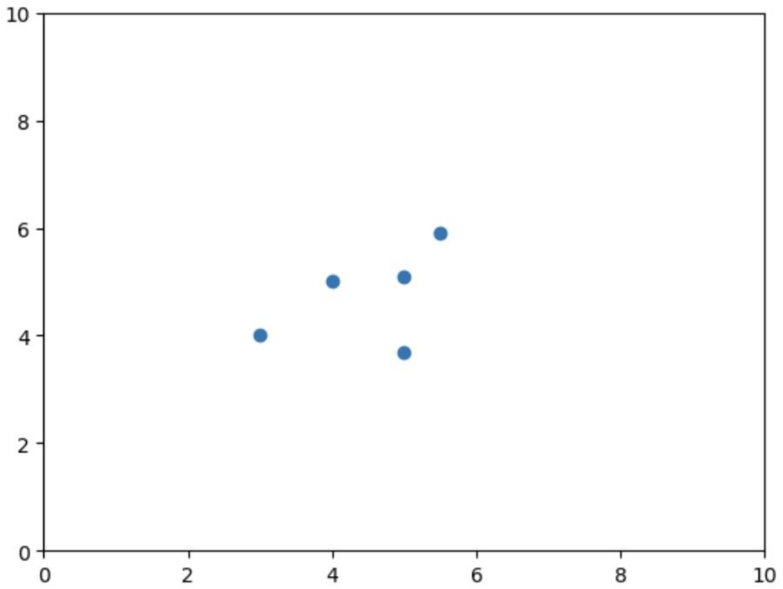
K-means is a **hard clustering** method - clusters, or groups that result, are **distinct** and **non-overlapping**

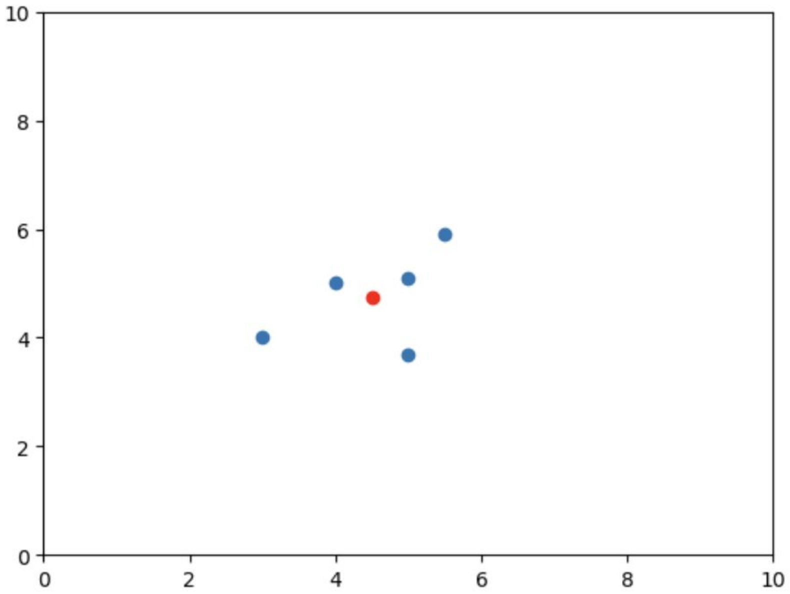
K-means makes use of **centroids** to help assign data points to particular clusters

It is a **non-deterministic** clustering method, meaning that running the algorithm multiple times may not always lead to the same results

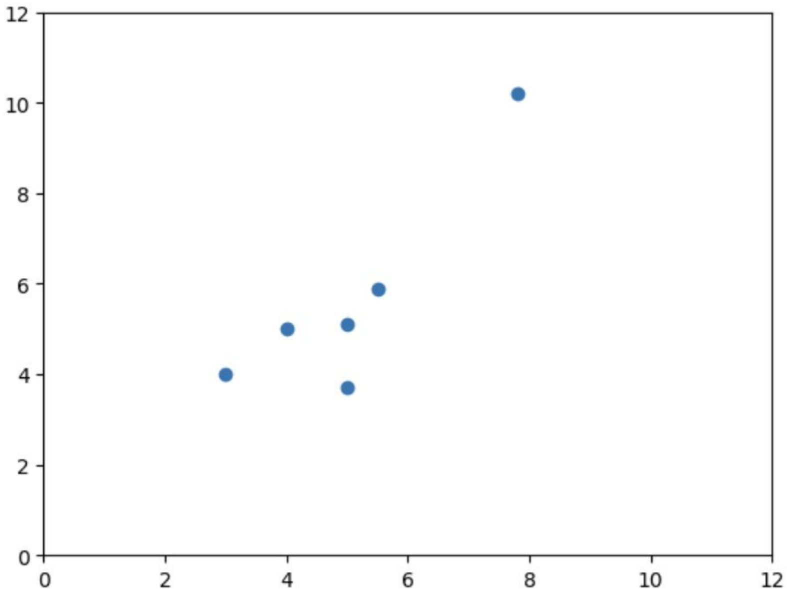


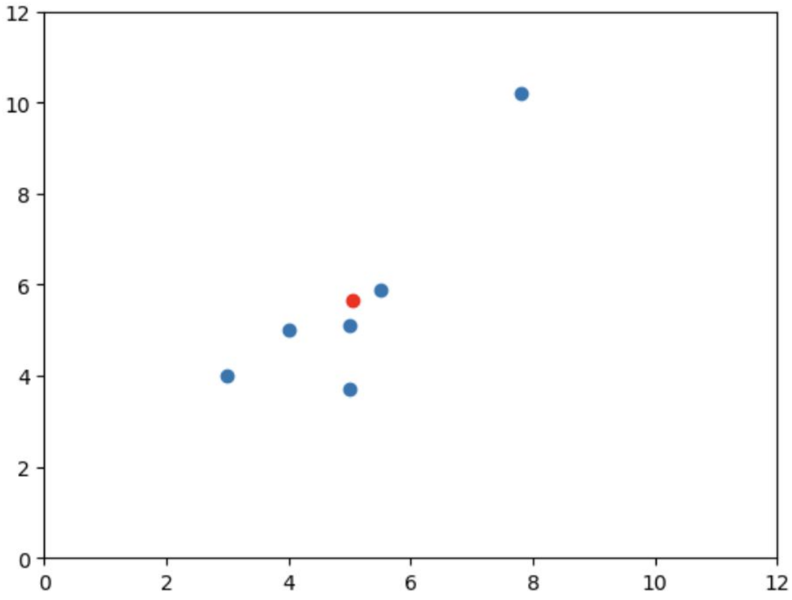
Wait. What's a centroid?





Wait. What’s a centroid?

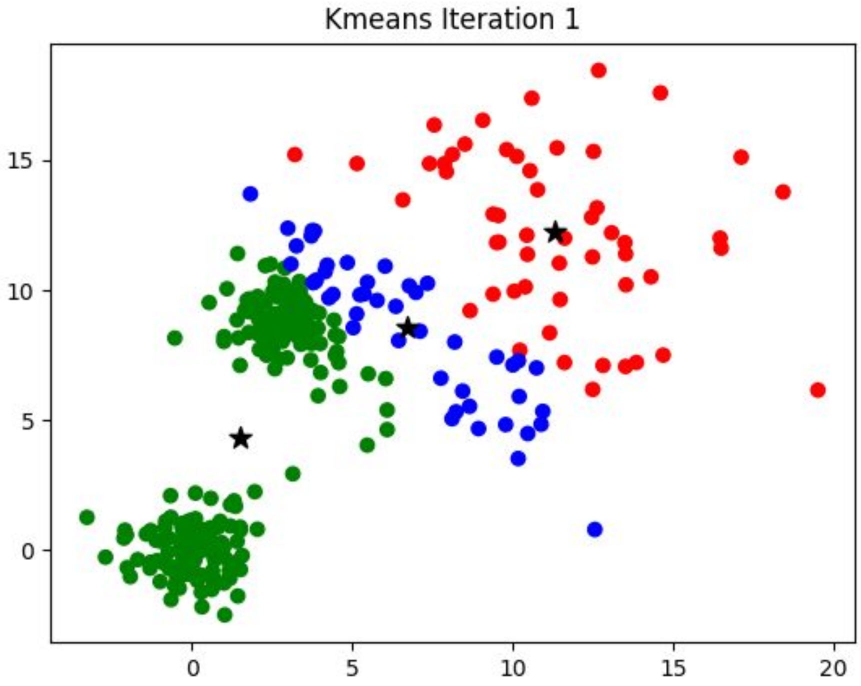




K-means Clustering

1. Pick the number of clusters K
2. Randomly assign each observation a cluster number from 1 to K
3. Compute each cluster centroid.
4. Assign each observation to the cluster whose centroid is closest (using a chosen distance metric)
5. Update cluster centroids and data point membership
6. Repeat until we reach a stopping condition

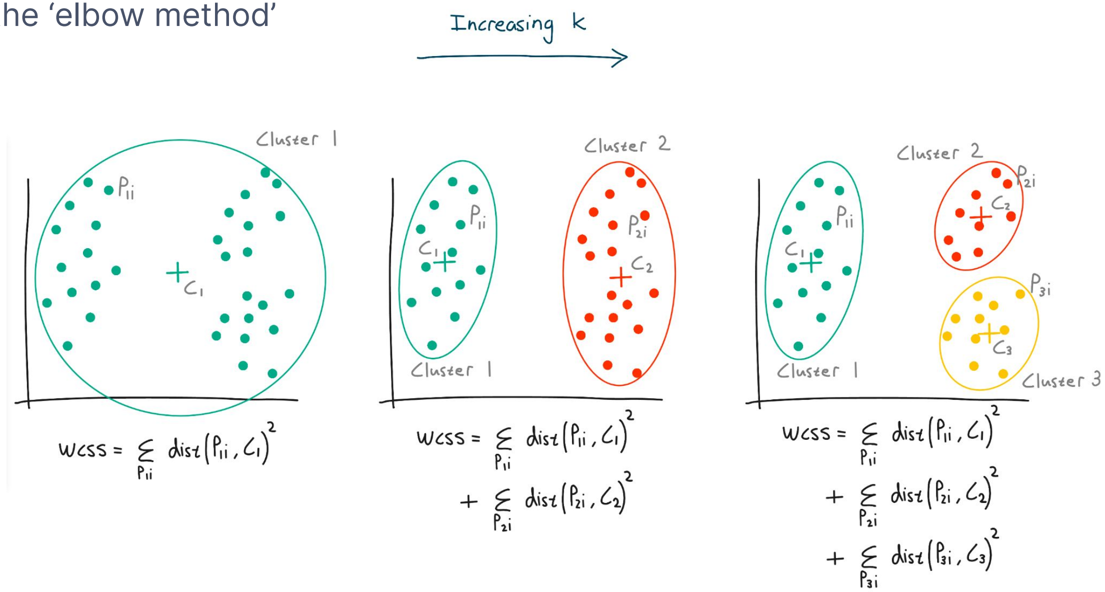
https://www.naftaliharris.com/blog/visualizing-k-means-clustering/



K-means Clustering

How do we decide on the number of clusters, K?

- Using the **within-cluster sum of squares** (WCSS) - a measure of similarity of the data points in the cluster
- Summed squares of distances between each point and the centroid
- We find an optimal K using the ‘elbow method’

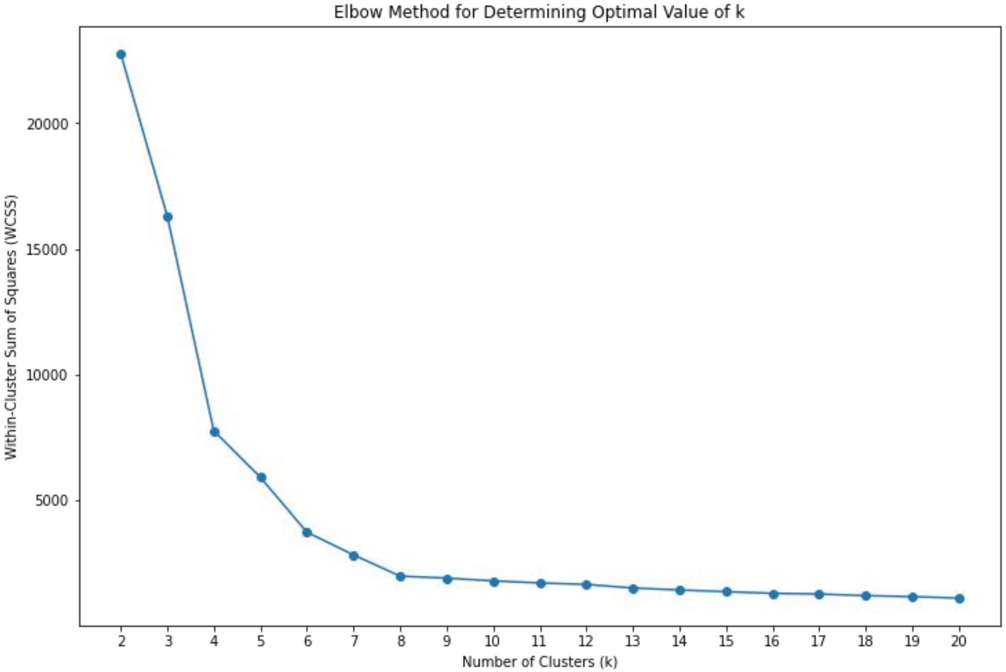


Increasing k
Cluster 1
P_{1i}
C_1
$WCSS = \sum_{P_{1i}} dist(P_{1i}, C_1)^2$
Cluster 1
Cluster 2
P_{1i}
P_{2i}
C_1
C_2
$WCSS = \sum_{P_{1i}} dist(P_{1i}, C_1)^2 + \sum_{P_{2i}} dist(P_{2i}, C_2)^2$
Cluster 1
Cluster 2
Cluster 3
P_{1i}
P_{2i}
P_{3i}
C_1
C_2
C_3
$WCSS = \sum_{P_{1i}} dist(P_{1i}, C_1)^2 + \sum_{P_{2i}} dist(P_{2i}, C_2)^2 + \sum_{P_{3i}} dist(P_{3i}, C_3)^2$

K-means Clustering

How do we decide on the number of clusters, K?

- Using the **within-cluster sum of squares** (WCSS) - a measure of similarity of the data points in the cluster
- We find an optimal K using the ‘elbow method’



- *This can also be done with other measures - the **between-cluster sum of squares** (BCSS) or the **CH index** - both detailed in the notebook*

K-means Clustering

Advantages

- Computationally efficient
- Suitable for large datasets
- Easy to implement and interpret
- Easily adaptable to new datasets

Disadvantages

- Lacks consistency (non-deterministic)
- Sensitive to scaling
- Dependent on initial values (initialise k)

1 / **What is clustering?**

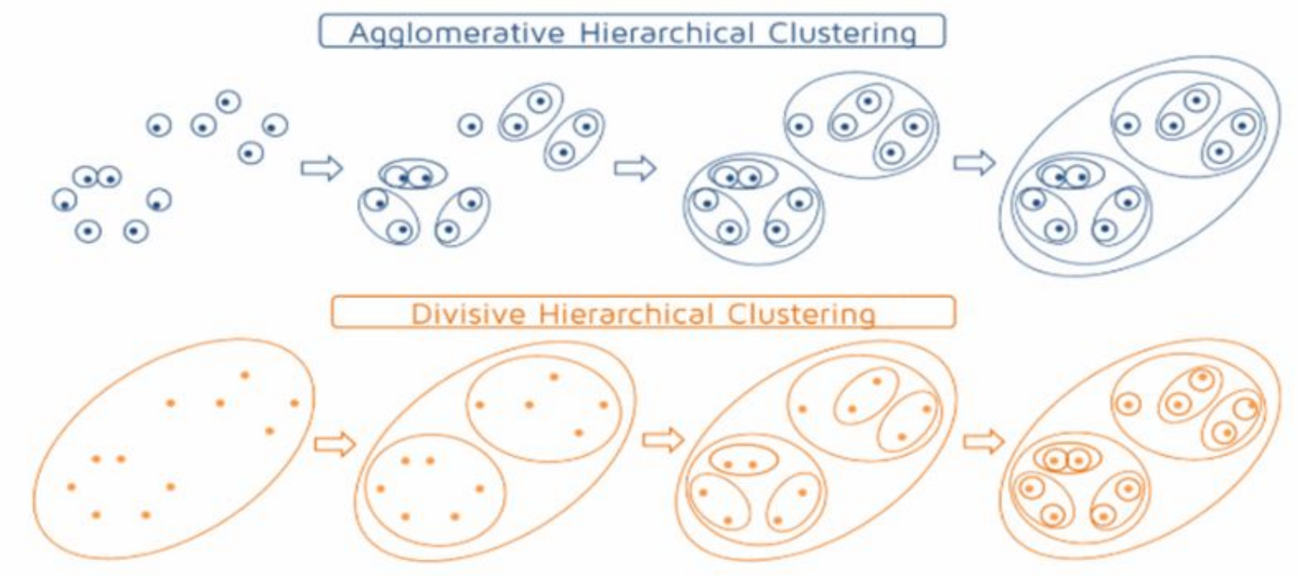
2 / **Similarity**

3 / **K-means clustering**

4 / **Hierarchical clustering**

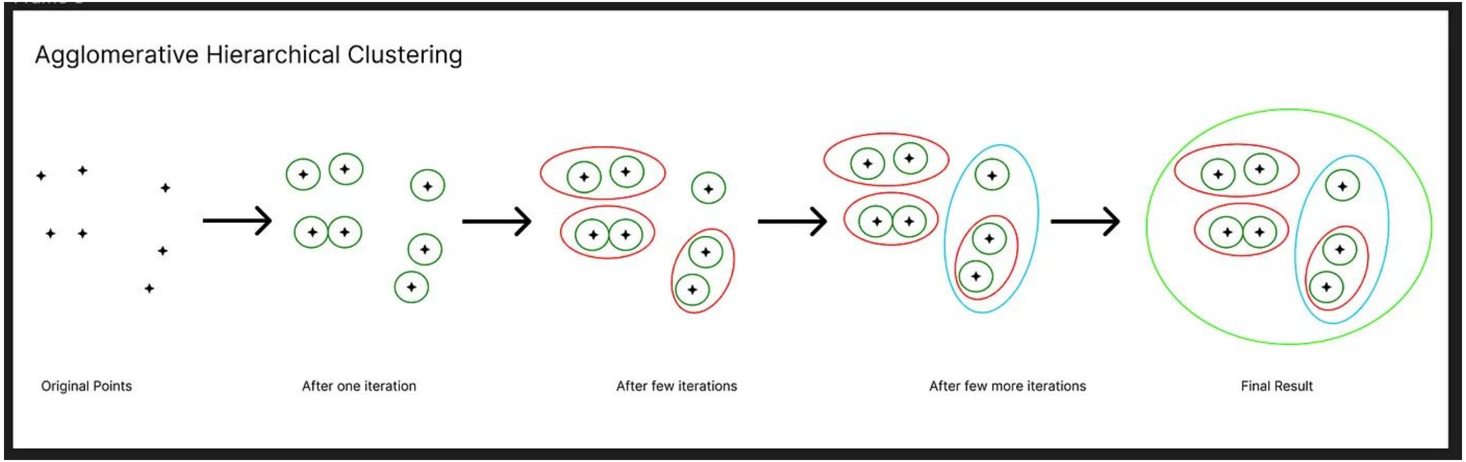
Hierarchical Clustering

Data is arranged in a tree-like structure such that parent clusters contain smaller child clusters, which also have their own child clusters, etc. → a **hierarchy**!


Agglomerative Hierarchical Clustering
Divisive Hierarchical Clustering

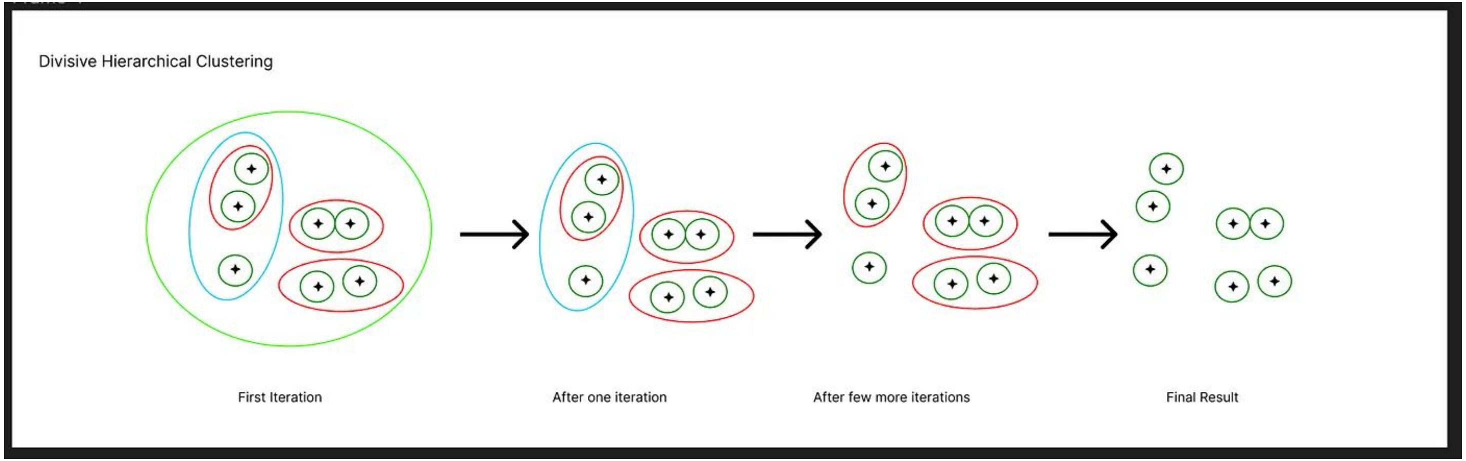
Agglomerative Hierarchical Clustering

1. Each data point begins as its own “cluster”
2. Two "most similar" clusters (determined by a distance measure) join to form a new cluster
3. Repeated iteratively until all data points belong to the same cluster
4. Determine a suitable number of clusters

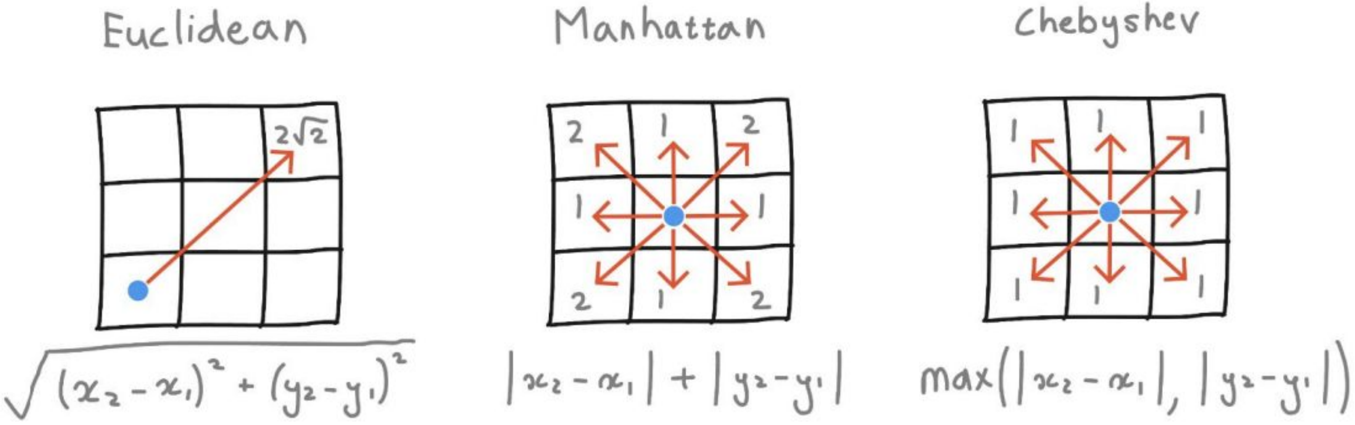


Divisive Hierarchical Clustering

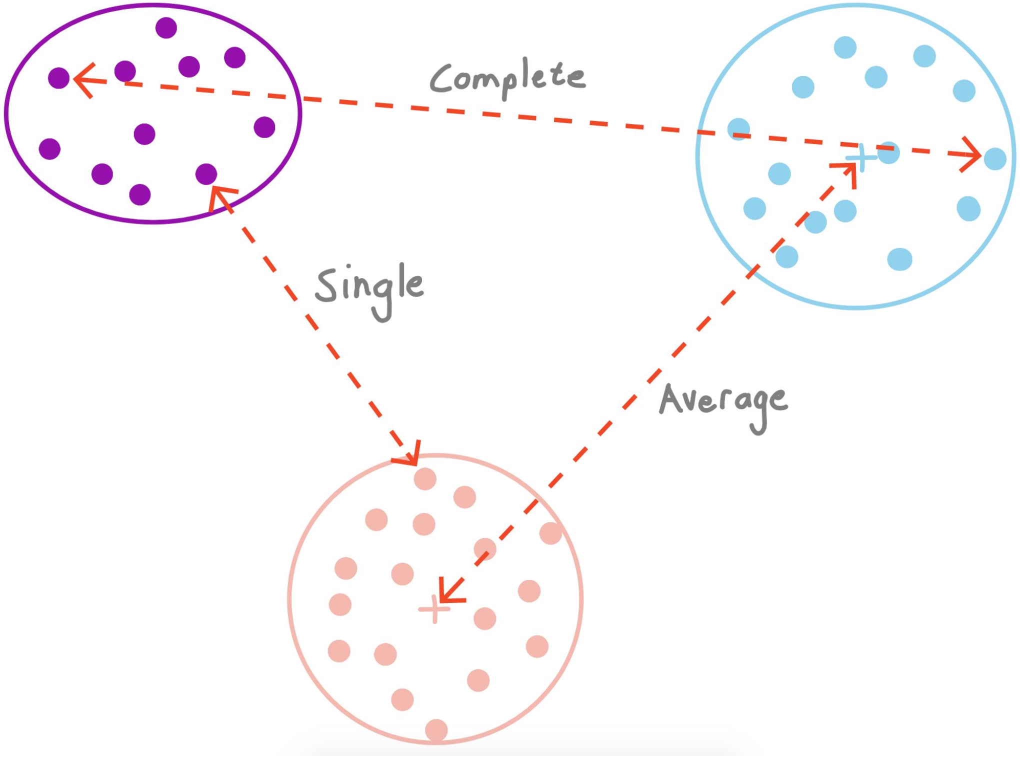
1. All data points start in one cluster altogether
2. The initial single cluster is broken down into child clusters
3. Repeated iteratively until all data points belong to their own cluster
4. Determine a suitable number of clusters



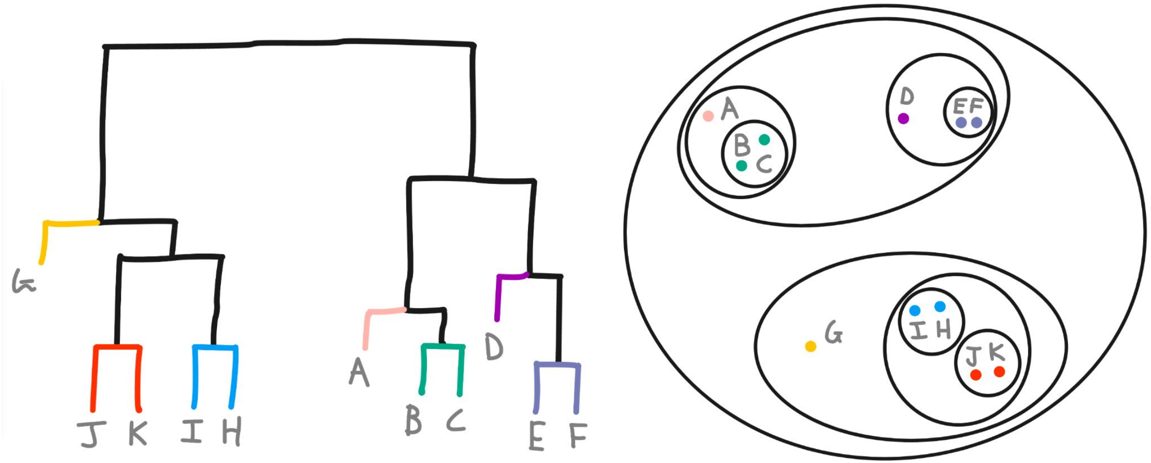
Similarity


Euclidean
$2\sqrt{2}$
$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$
Manhattan
2 1 2
1 1
2 1 2
$|x_2-x_1|+|y_2-y_1|$
Chebyshev
1 1 1
1 1
1 1 1
$\max(|x_2-x_1|,|y_2-y_1|)$

Linkage methods

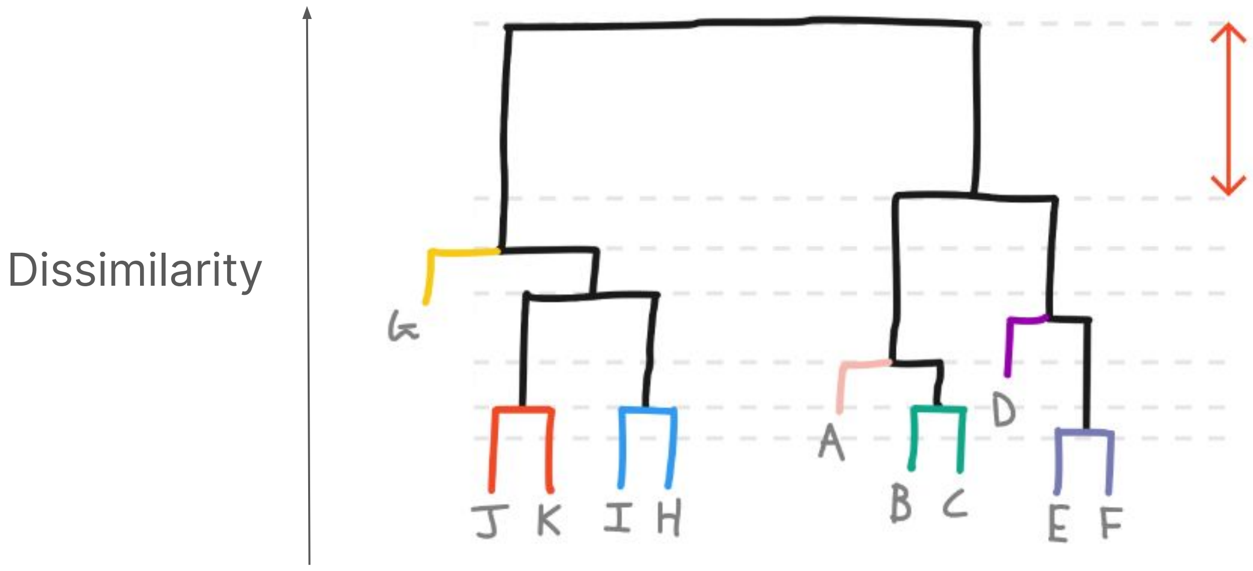


Dendrograms & determining n clusters

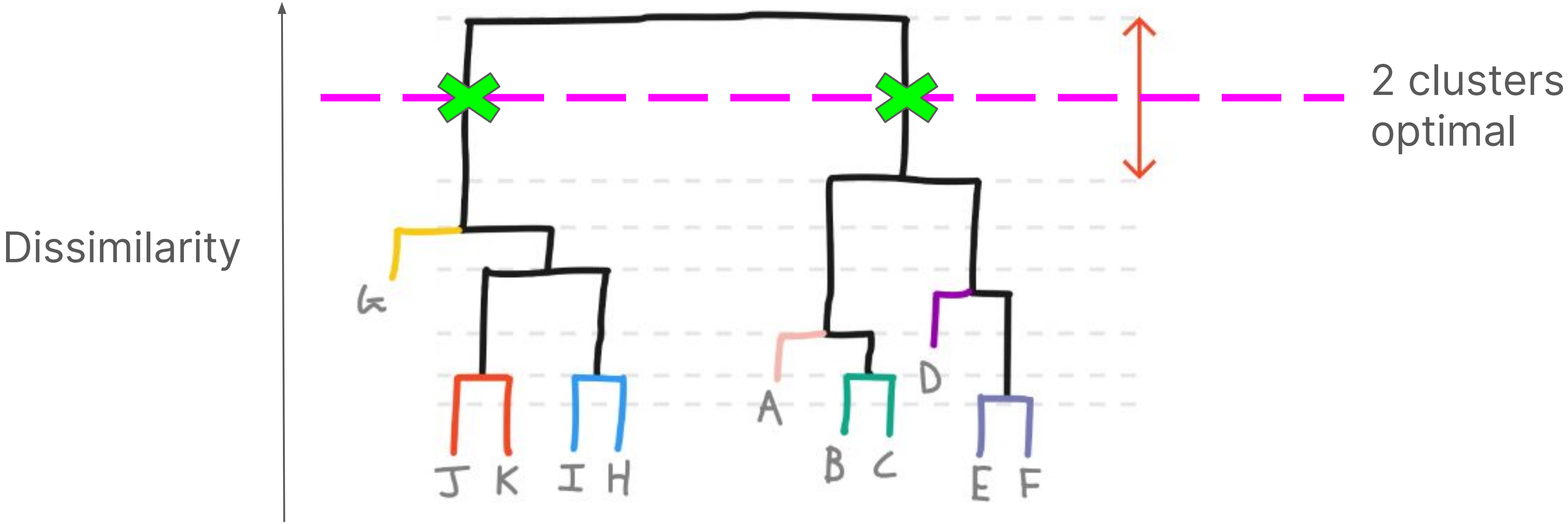


A, B, C, etc = Clusters

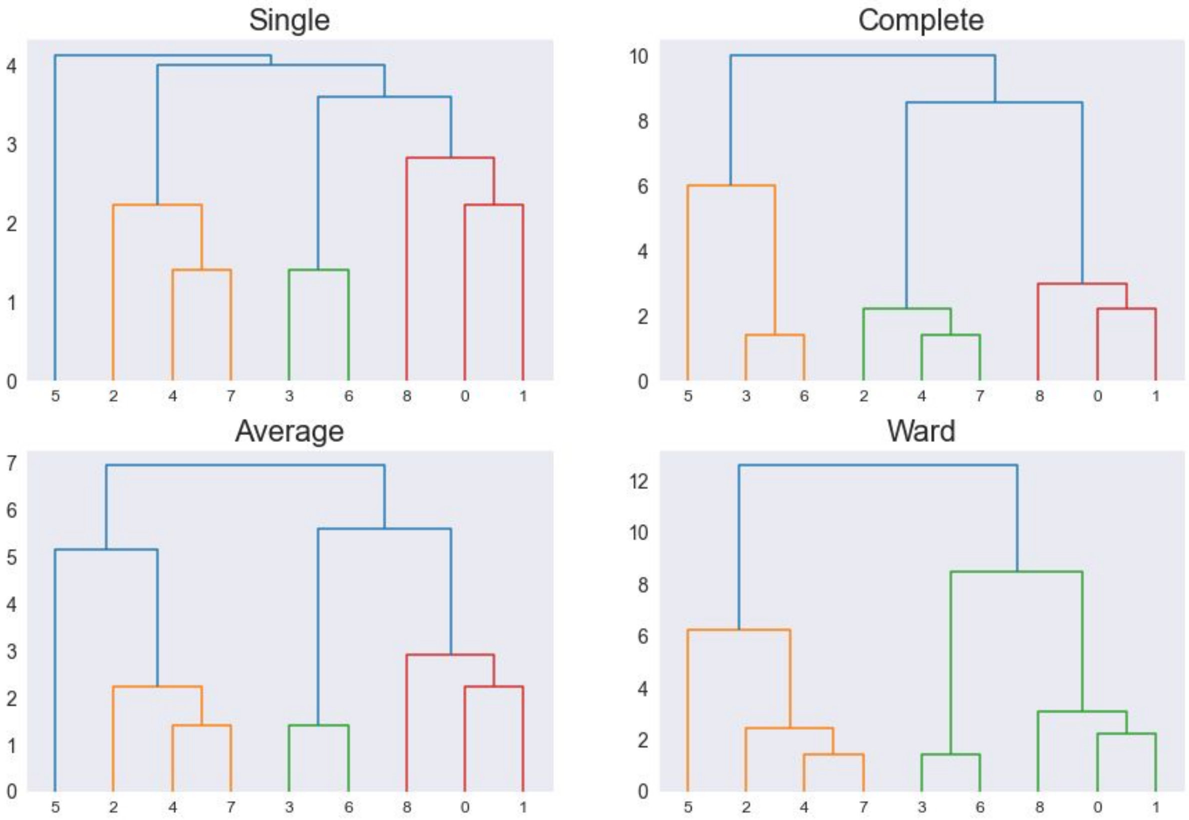
Dendrograms & determining n clusters



Dendrograms & determining n clusters



Dendrograms & determining n clusters



Hierarchical Clustering

Advantages

- Simple to implement
- No prior information about the number of clusters in the data is required

Disadvantages

- Algorithm takes much longer to run
- Runtime worsens as the size of the dataset increases.
- Difficult to determine ideal number of clusters from dendrogram

We’ve got clusters…now what?

Once we have clusters, we can use these clusters to analyse our data further. From what we’ve identified as the optimal number of clusters, are there any interesting trends or characteristics of the data points that fall into particular clusters?

→ Are there any interesting similarities in certain features in a particular cluster?

→ Are any clusters much larger than others? Why?

→ Anywhere you can identify that more/less clustering would be needed?

→ Would your data benefit from dimensionality reduction before retrying clustering?

→ Can you think of any ways dimensionality reduction and clustering could be useful in regression or classification problems?

Objectives

By the end of this session you should be able to...

- Understand the concept of clustering
- Apply a calculation given a distance metric formula and data points
- Distinguish between K-means and hierarchical clustering
- Describe the algorithms of K-means clustering and hierarchical clustering, up to determining k clusters
- Distinguish between the different linkage methods