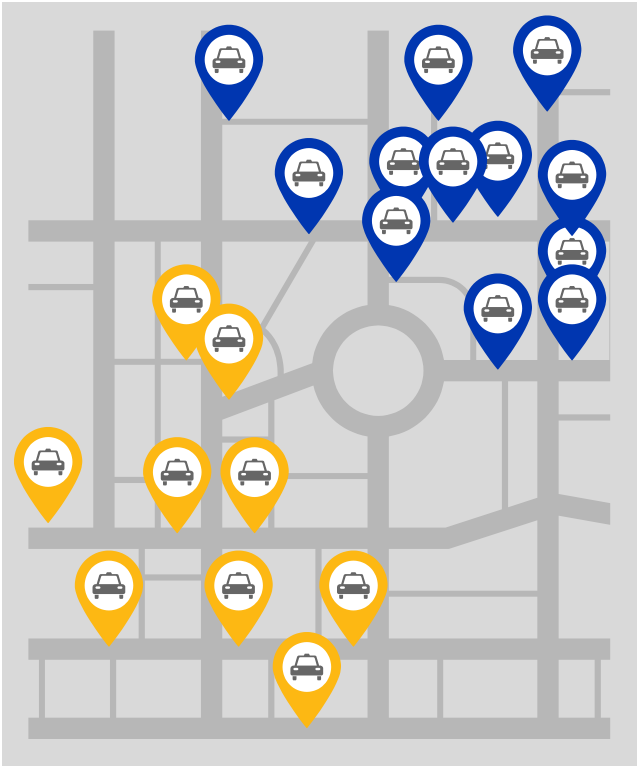
Taxi Trajectory Analysis

EDAA - G04

Diogo Rodrigues
Eduardo Correia
João Sousa

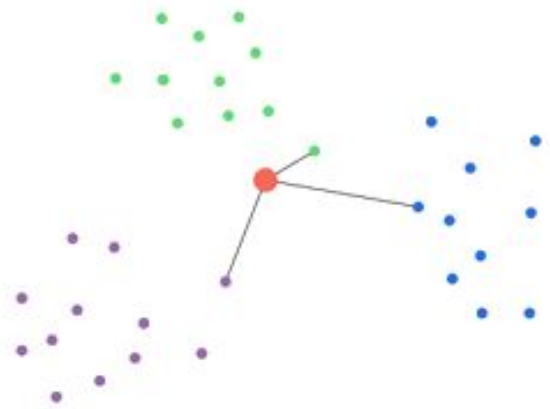
Problem recap

Clustering of the **processed data points** in the first part of the project.

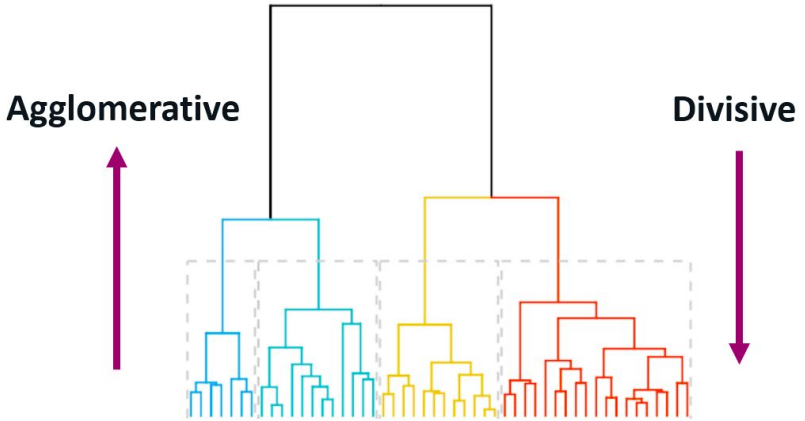


Clustering

Two approaches:



k-means



Hierarchical Clustering

k-means

Pseudocode

Input:

```
    D = {t₁, t₂, …, t□} // Set of elements
    K                   // Numbers of desired clusters
```

Output:

```
    C                   // Set of clusters
```

Procedure:

```
    Assign initial values for m₁, m₂, …, m□
    Repeat
        Assign each item tᵢ to the clusters which has the closest mean;
        Calculate new mean for each cluster;
    Until convergence criteria is met
```

k-means

Complexity Analysis

Time complexity:

- $O(N \times K \times I)$

Space complexity:

- $O(N \times (D + K))$

***N* -** Number of points

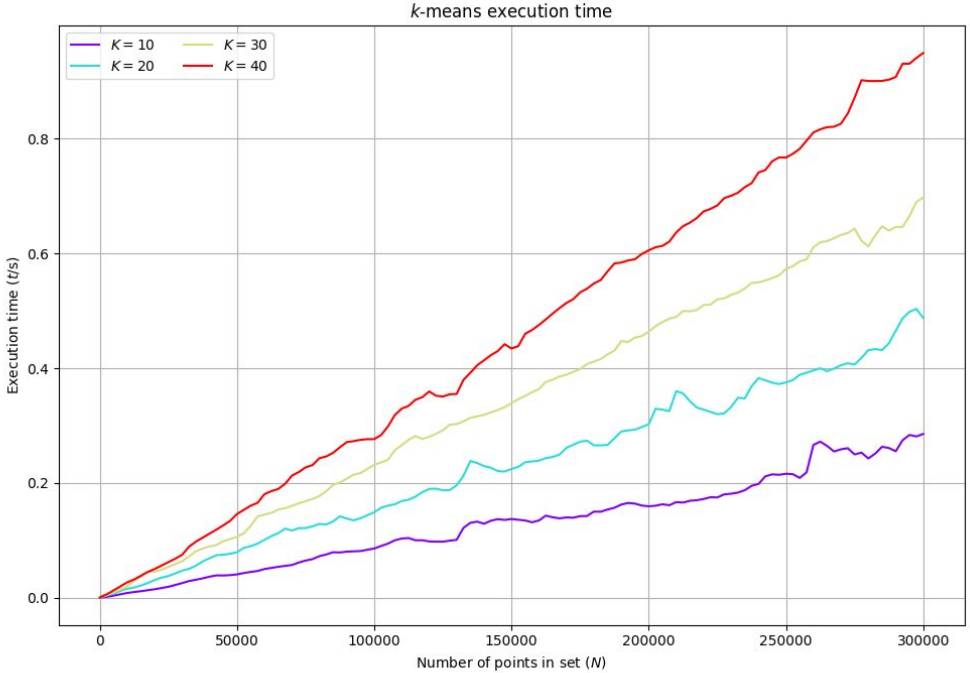
***D* -** Number of dimensions

***K* -** Number of centroids

***I* -** Number of iterations

k-means

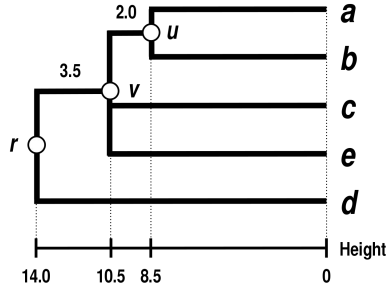
Empirical Analysis



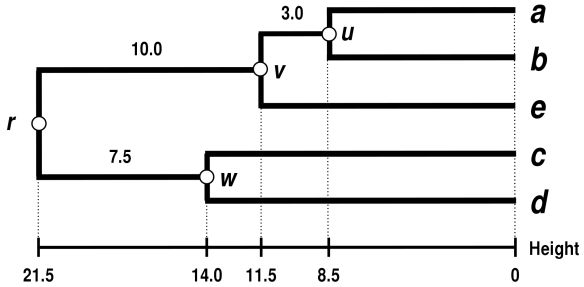
Averaged 5 samples, 4-points exponential moving average

Hierarchical Clustering

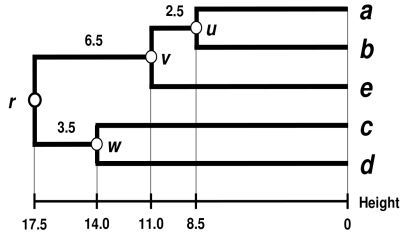
- There are many methods to perform hierarchical clustering, each yielding different results for the same distance matrix
- They essential difference is on how they link clusters together



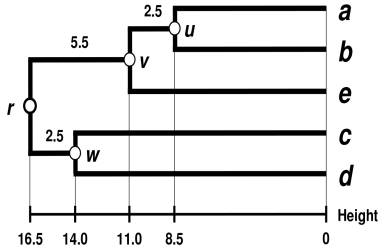
Single-linkage



Complete-linkage



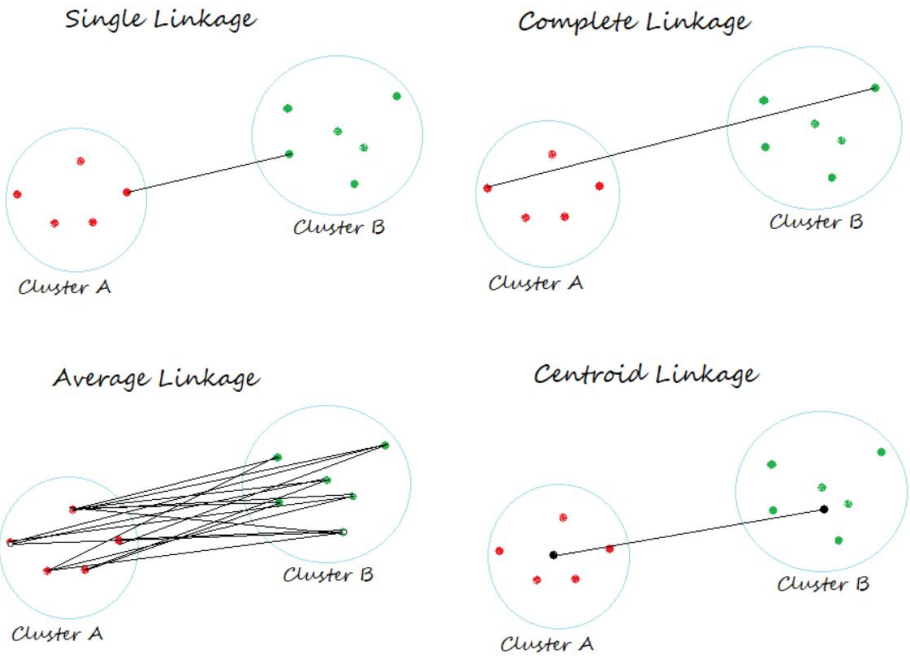
WPGMA



UPGMA

Hierarchical Clustering

Linkage Criteria



Hierarchical Clustering

UPGMA

- Stands for **U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic **M**ean
- Simple agglomerative (bottom-up) hierarchical clustering method
- Constructs a rooted tree (dendrogram) that reflects the structure present in a pairwise similarity matrix (or a dissimilarity matrix).
- At each step, the nearest two clusters are combined into a higher-level cluster.

UPGMA

Pseudocode

Input:

```
    D    // Distance matrix
```

Output:

```
    T    // Hierarchical Tree
```

Procedure:

```
    For i = Size of D
        Find closest pair in distance matrix D;
        Add new node in tree T;
        Update distance matrix D;
    End For
```

UPGMA

Complexity Analysis

Time complexity:

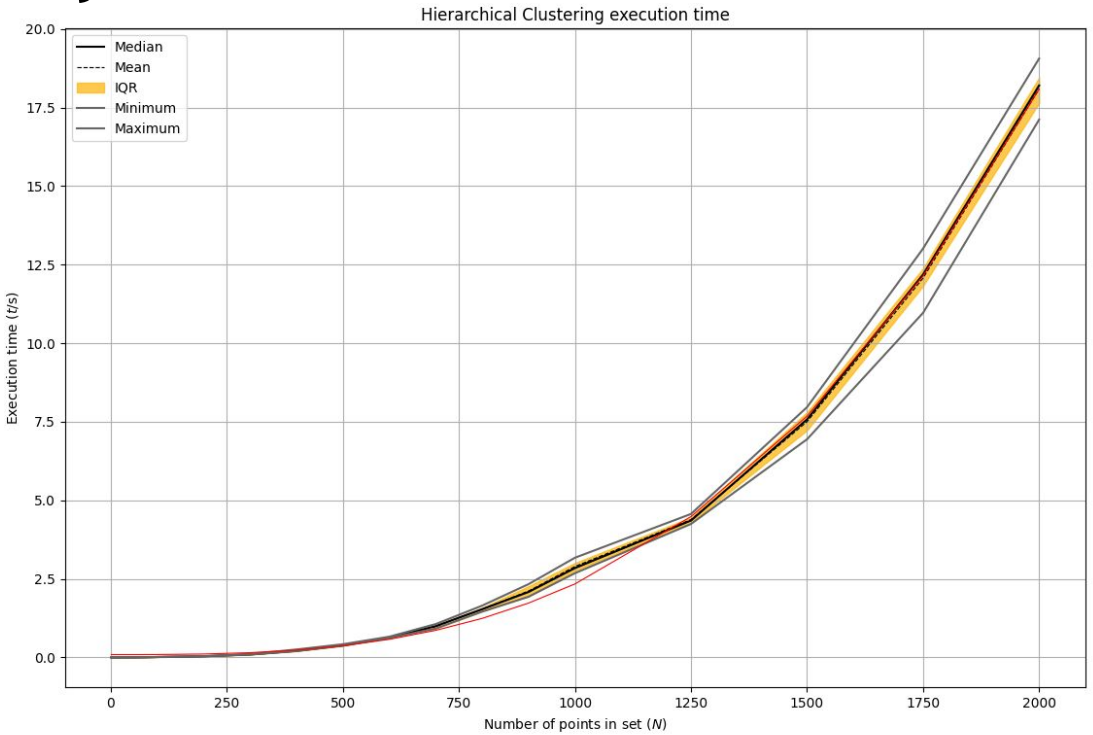
- $O(n^3)$ - Trivial implementation
- $O(n^2 \log n)$ - Using a heap for each cluster to keep its distances from other cluster
- $O(n^2)$ - Fionn Murtagh implementation

Space complexity:

- $O(n^2)$ - $n \times n$ Distance matrix

UPGMA

Empirical Analysis



Q&A