



# Unsupervised Learning: hierarchical clustering

Mirco Schönfeld  
[mirco.schoenfeld@uni-bayreuth.de](mailto:mirco.schoenfeld@uni-bayreuth.de)



# Strategies of Clustering

## Hierarchical Agglomerative Clustering

Each point is in its own cluster

Clusters are combined based on their “closeness”

Combination stops when undesirable clusters occur

## Point assignment

Initial clusters are estimated

Points are considered in some order

Points are assigned to clusters into which they best fit

# Examples: Hierarchical Clustering



```
WHILE more than one cluster left
DO
    pick the best two clusters to merge
    combine those two clusters into one cluster
END
```

# Examples: Hierarchical Clustering



```
WHILE more than one cluster left
DO
    pick the best two clusters to merge
    combine those two clusters into one cluster
END
```

How will clusters be represented?

How will we choose which clusters to merge?

This is the agglomerative approach (bottom up).  
A divisive approach exists as well which starts  
with one cluster that is recursively split



# Hierarchical Clustering: Represent Clusters

We need to combine nearest/closest clusters.

Key question: how to represent the „location“ of each cluster to tell which pair of clusters is closest?

In Euclidean spaces: each cluster has an average of its points – the **centroid**

In Non-Euclidean spaces:

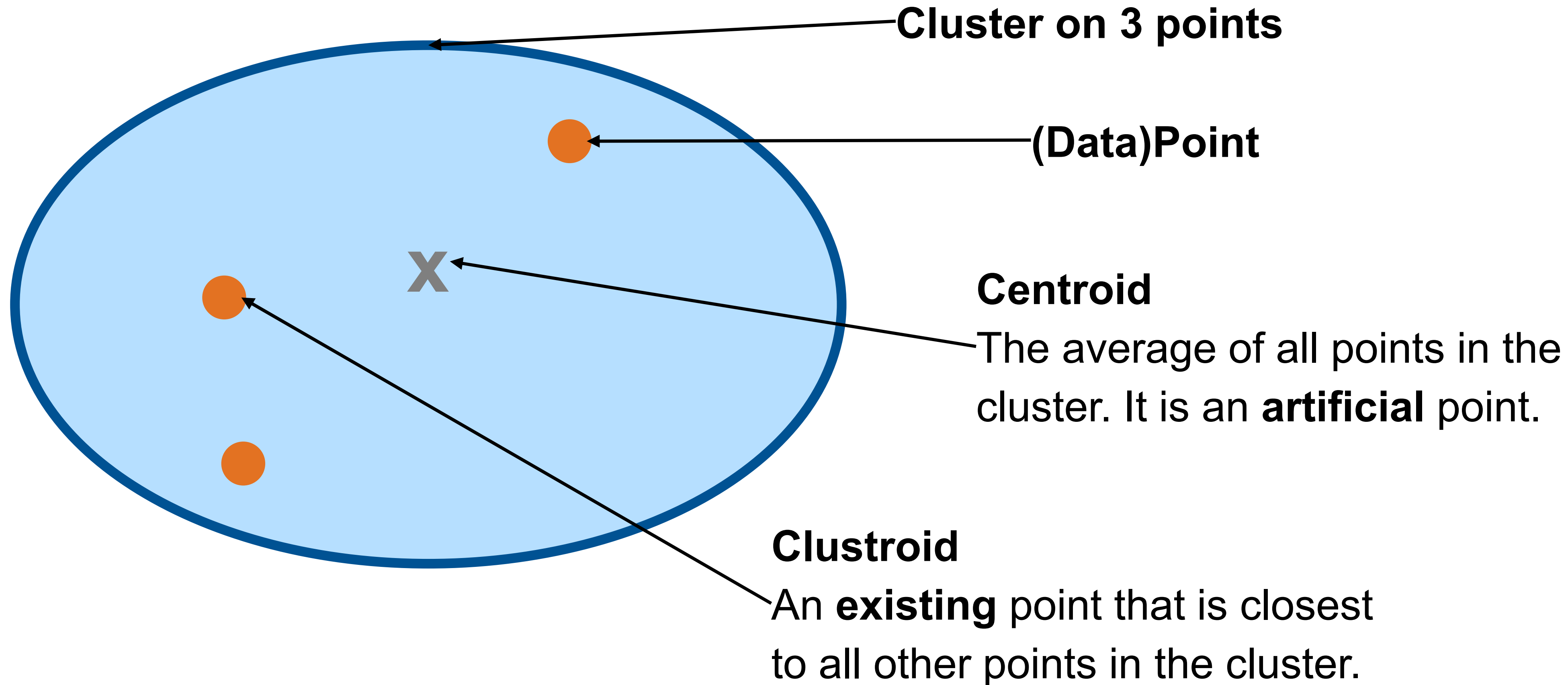
Only „locations“ are the points themselves

We do not have an average of points

Choose a **clustroid** which is a point closest to other points



# Centroids and Clustroids



Determining the clustroid, i.e. the point being closest to all other points:

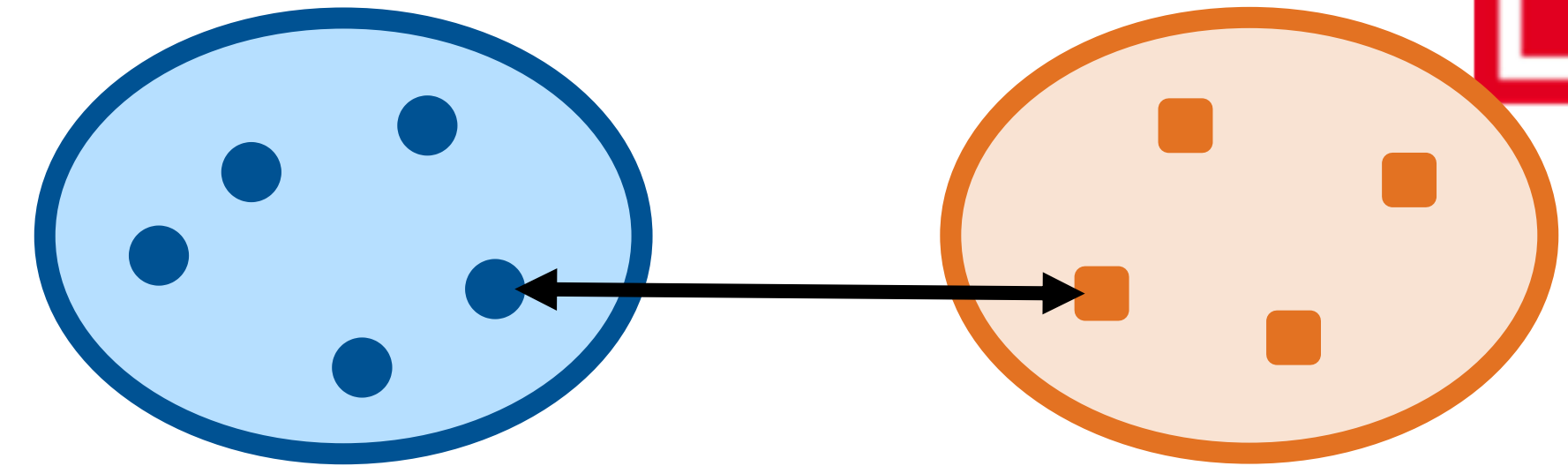
- Point with smallest maximum distance to other points
- Point with smallest average distance to other points
- More complicated notions



# Hierarchical Clustering: Compare Clusters

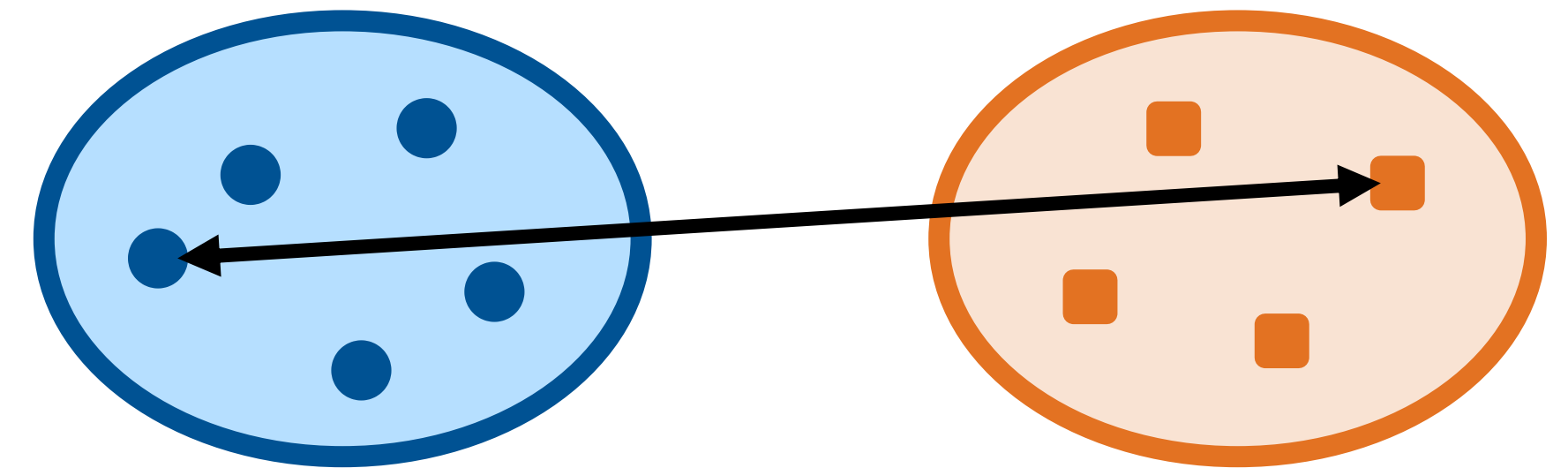
## Single-linkage:

Minimum distance (roughly maximum similarity)



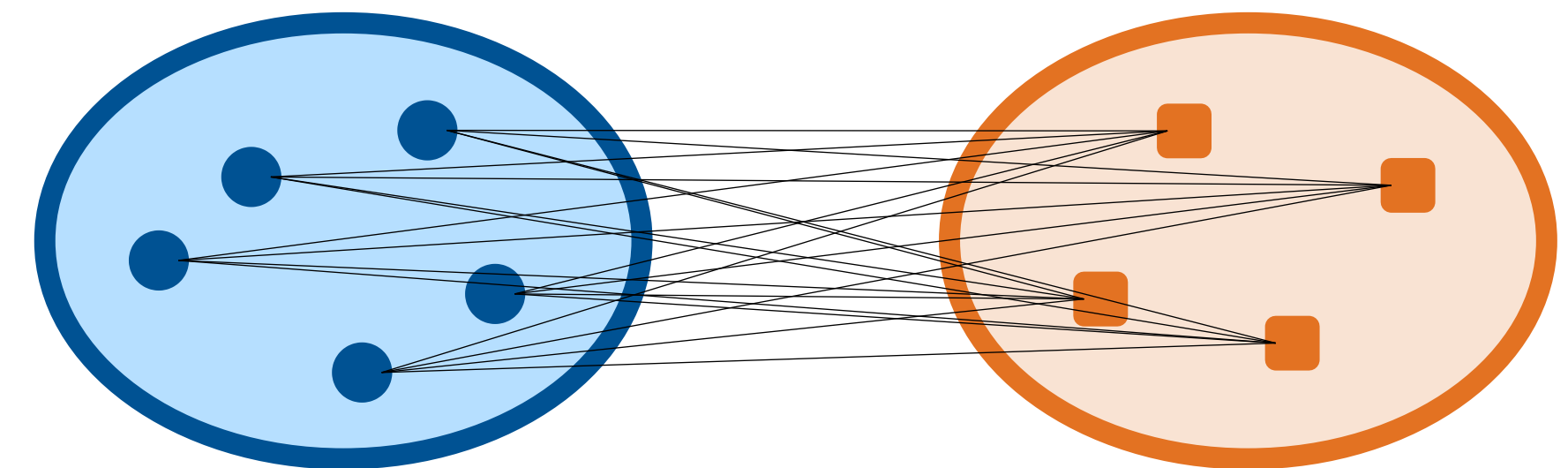
## Complete-linkage:

Maximum distance (roughly minimum similarity)



## Average-linkage:

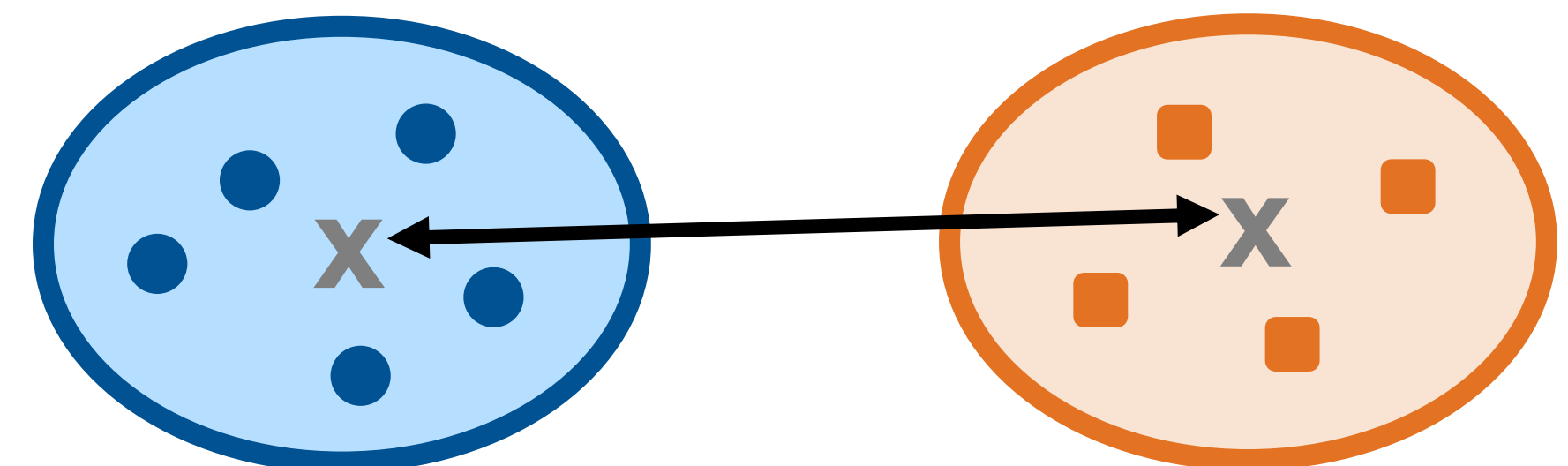
Average distance



## Centroid-linkage:

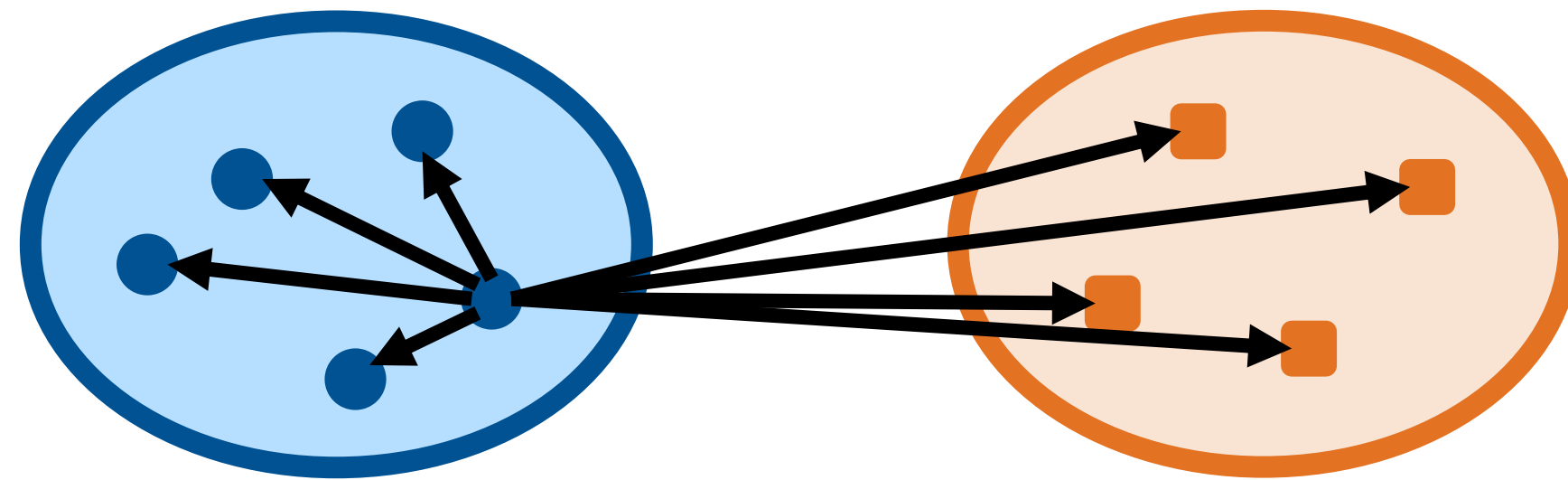
Distance between cluster centroids.

Only for Euclidean spaces.



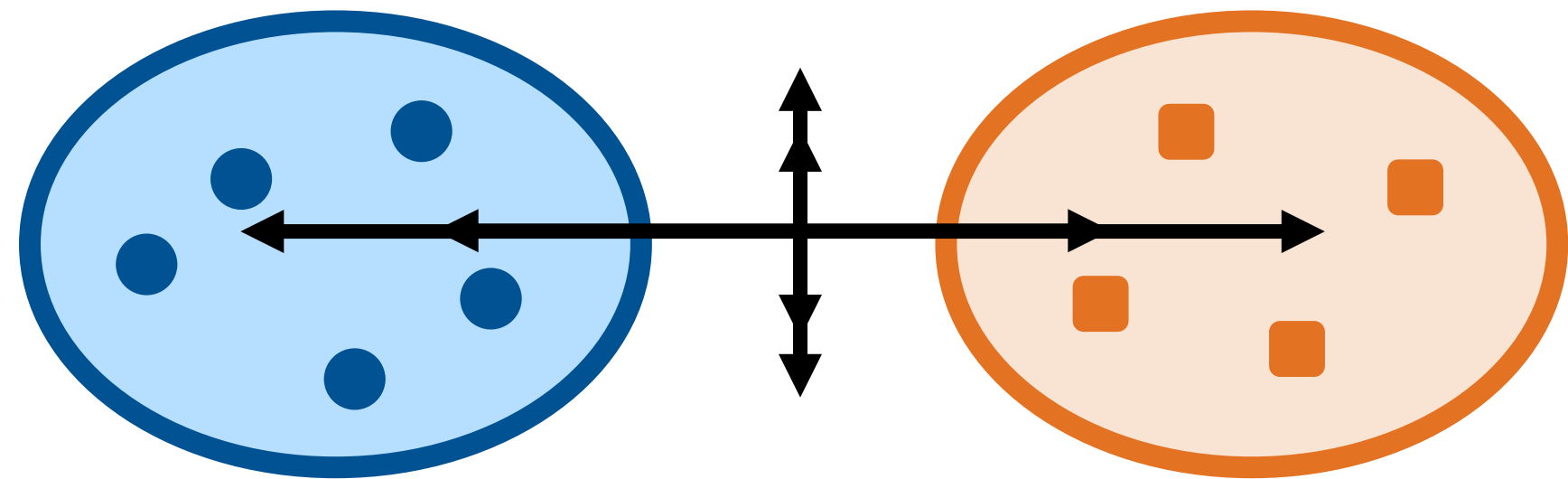


# Hierarchical Clustering: Compare Clusters



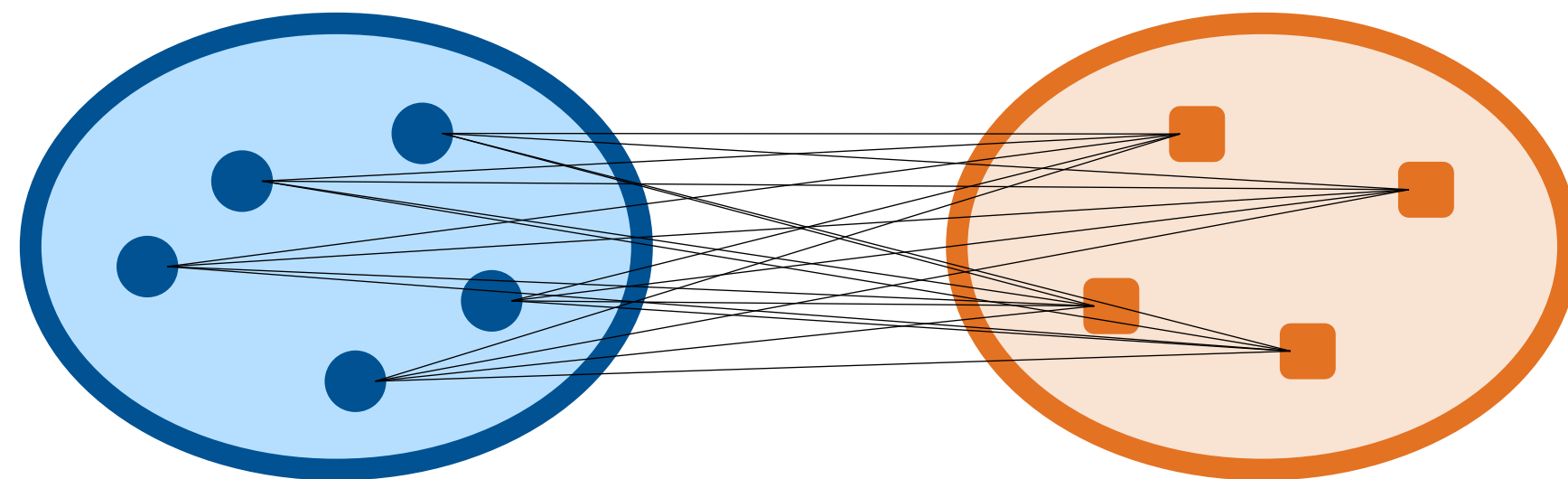
## Min-Max-linkage:

Best maximum distance (best minimum similarity)



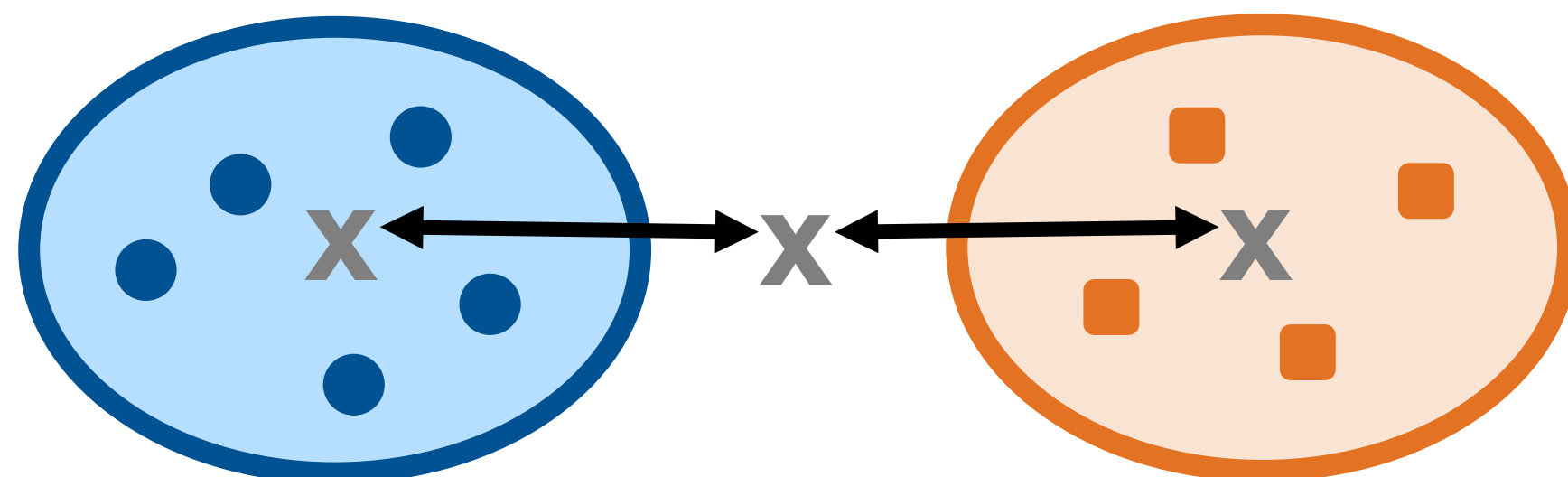
## Ward-linkage:

Minimum increase of squared error



## McQuitty (WPGMA):

Average distance to the previous two clusters.  
Recursive definition

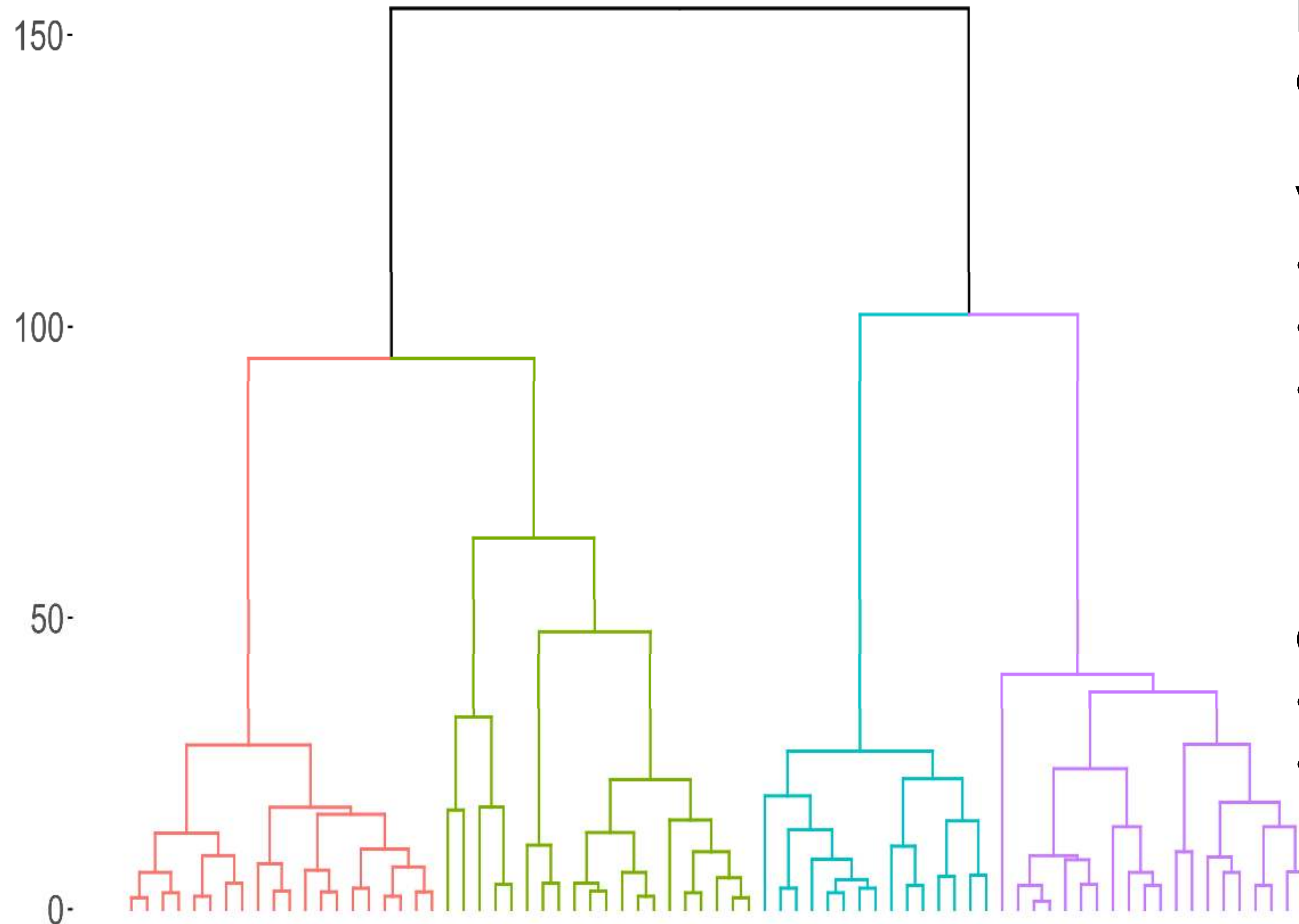


## Median-linkage:

Distance between cluster midpoints.  
Recursive definition



# From Dendrograms to Clusters



Hierarchical clustering outputs a **dendrogram**, but not „clusters“

Various strategies to select a clustering:

- Choose visually interesting branches
- Cut tree horizontally
- Other scientific approaches using cluster distances, densities, sizes, clustered objects, ....

Questions:

- Are clusters allowed to overlap?
- How to handle outliers?



# Hierarchical Clustering: why and why not?

## Pro:

- Very general. Supports any distance metric
- Number of clusters doesn't need to be known beforehand

## Contra:

- Unbalanced cluster sizes
- Outliers
- Slow for large datasets