EuroVis 2016
Machine Learning Methods
in Visualization for Big Data
6 June 2016, Groningen, the Netherlands

Hierarchical Clustering

Jaakko Peltonen[1,2]

[1]Aalto University, Department of Computer Science
[2]University of Tampere, School of Information Sciences

A single partitioning clustering - not enough?

A single partitioning clustering result e.g. from k-means is just one of many clusterings for the data.

- How many clusters are needed? Maybe more than one level of detail is needed!

- One could run several clusterings (e.g. k-means) at different numbers of clusters - but results have no clear relationship to each other.

Hierarchical clustering builds a tree of clusterings with changing number of clusters (detail level). Each level arises from the next by a simple **merge/split**.

Agglomerative hierarchical clustering

The most common hierarchical clustering is **agglomerative hierarchical clustering**.

- In a bottom-up strategy, at first each data item is its own cluster.
- Iteratively, two ”closest” clusters are merged, until only one cluster remains.

- The reverse is called divisive clustering.

Agglomerative hierarchical clustering

Three types of agglomerative clustering:

- **Single-linkage:** distance of two clusters is the minimum between their members

$$d(c_1, c_2) = min_{i \in c_1} min_{j \in c_2} d(x_i, x_j)$$

- **Average-linkage (aka UPGMA):** distance of two clusters is the average between their members

$$d(c_1, c_2) = E_{i \in c_1}[E_{j \in c_2}[d(x_i, x_j)]]$$

- **Complete-linkage:** distance of two clusters is the maximum between their members

$$d(c_1, c_2) = max_{i \in c_1} max_{j \in c_2} d(x_i, x_j)$$

Agglomerative hierarchical clustering

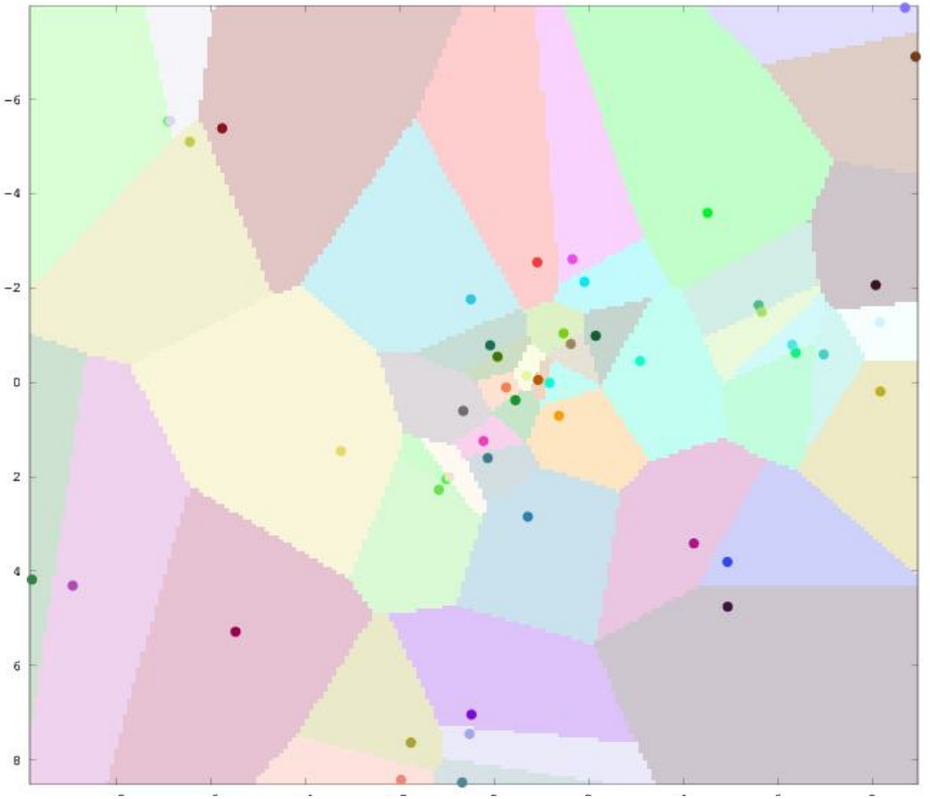
Pseudocode: compute distances between all clusters, merge the two closest ones, repeat.

In each iteration, the number of active clusters decreases by 1.

A new data item can be assigned to a cluster by the same rules as the merging rules (minimum/average/maximum distance).

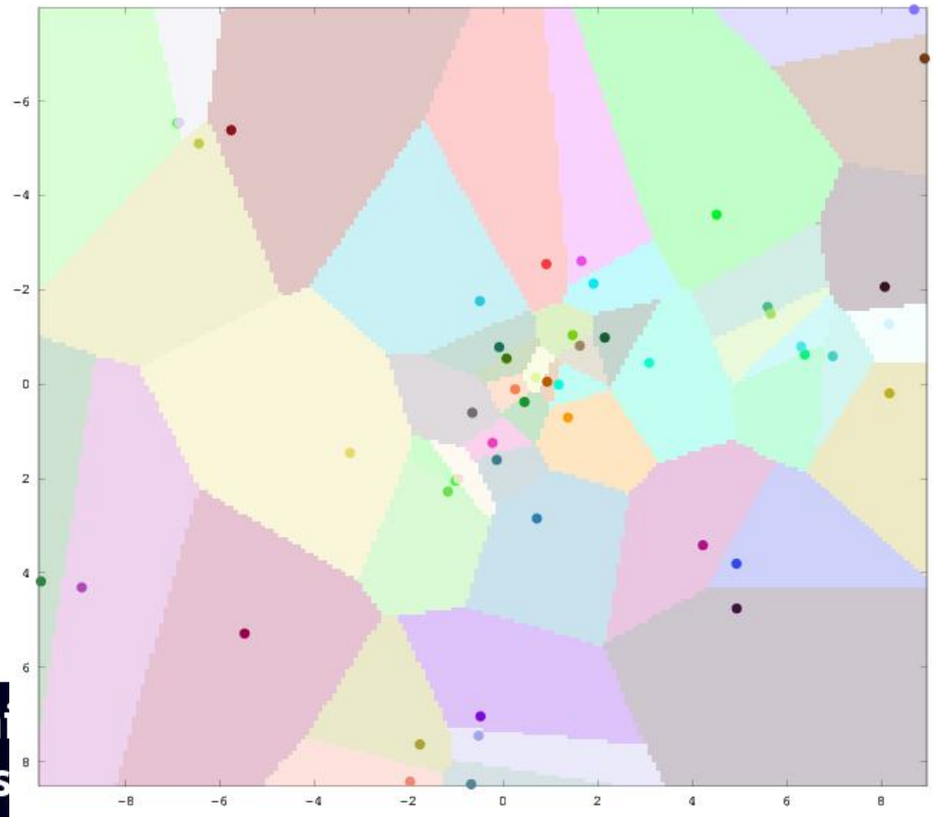
Example

50 clusters

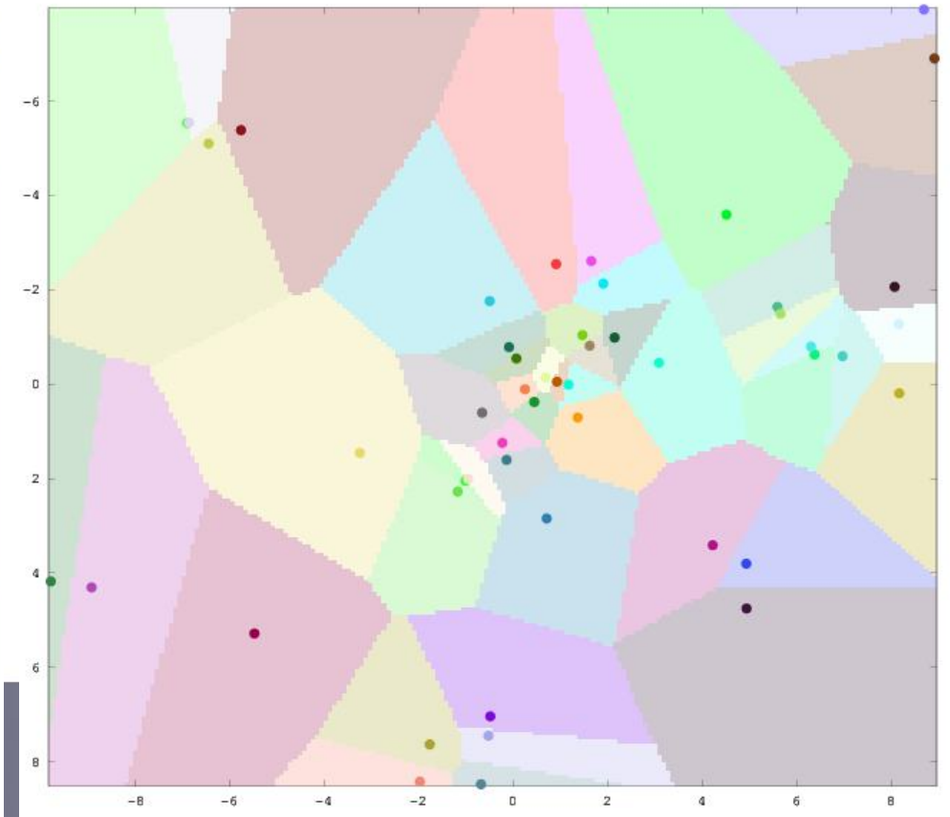


single-linkage

average linkage



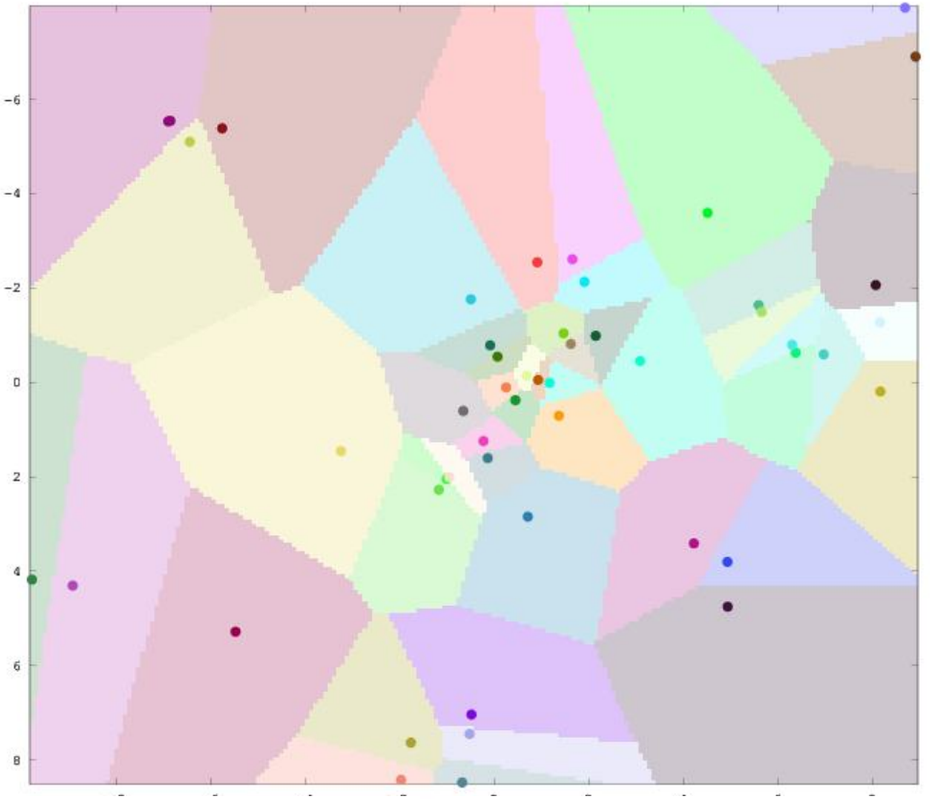
complete linkage



Example

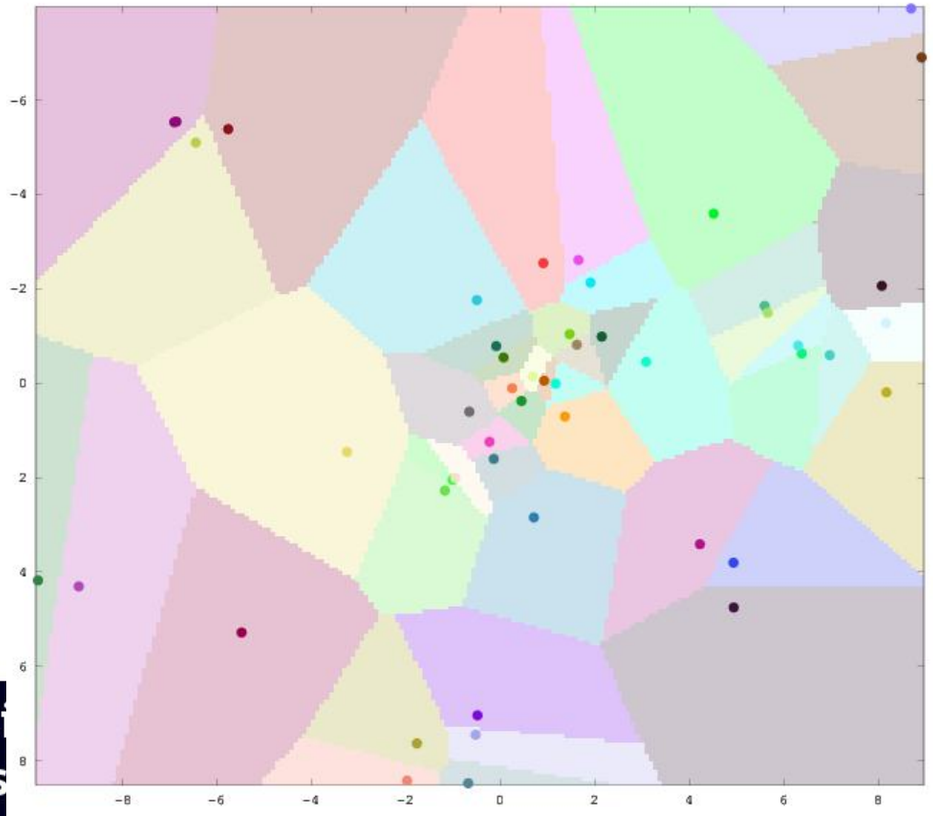
49 clusters

(merge in top left corner)

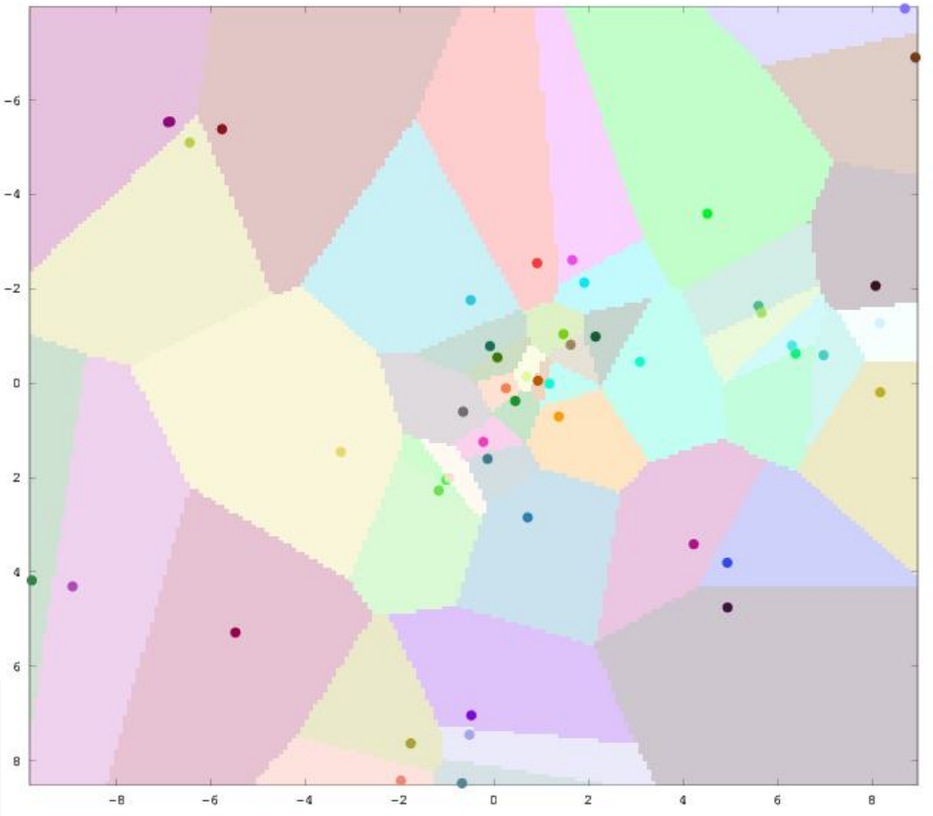


single-linkage

average linkage



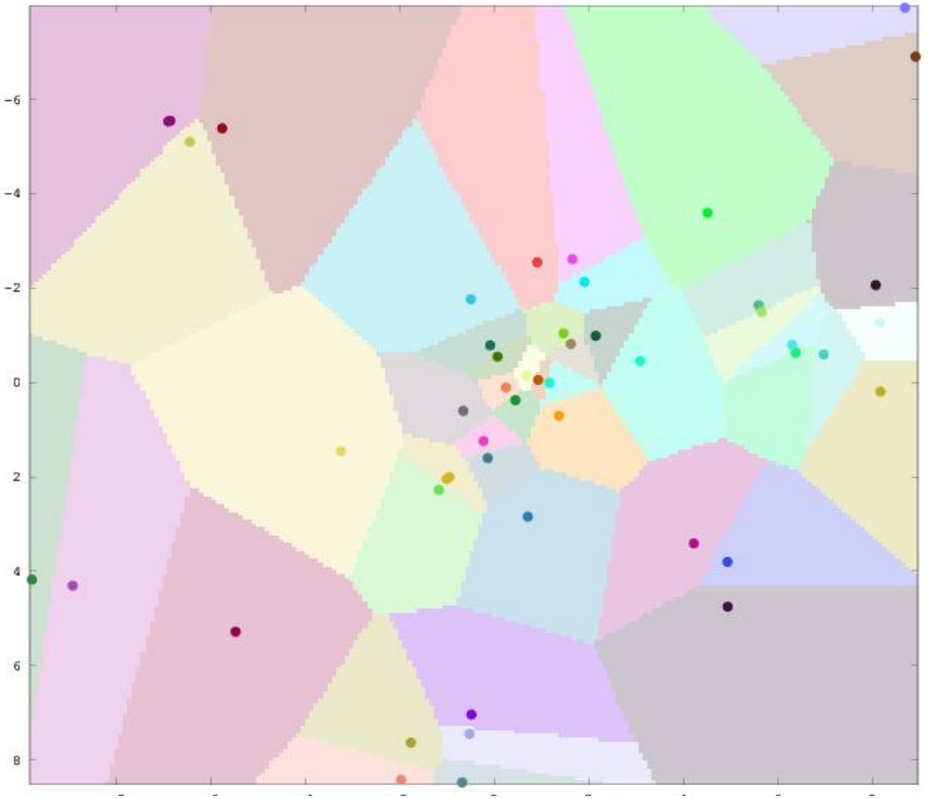
complete linkage



Example

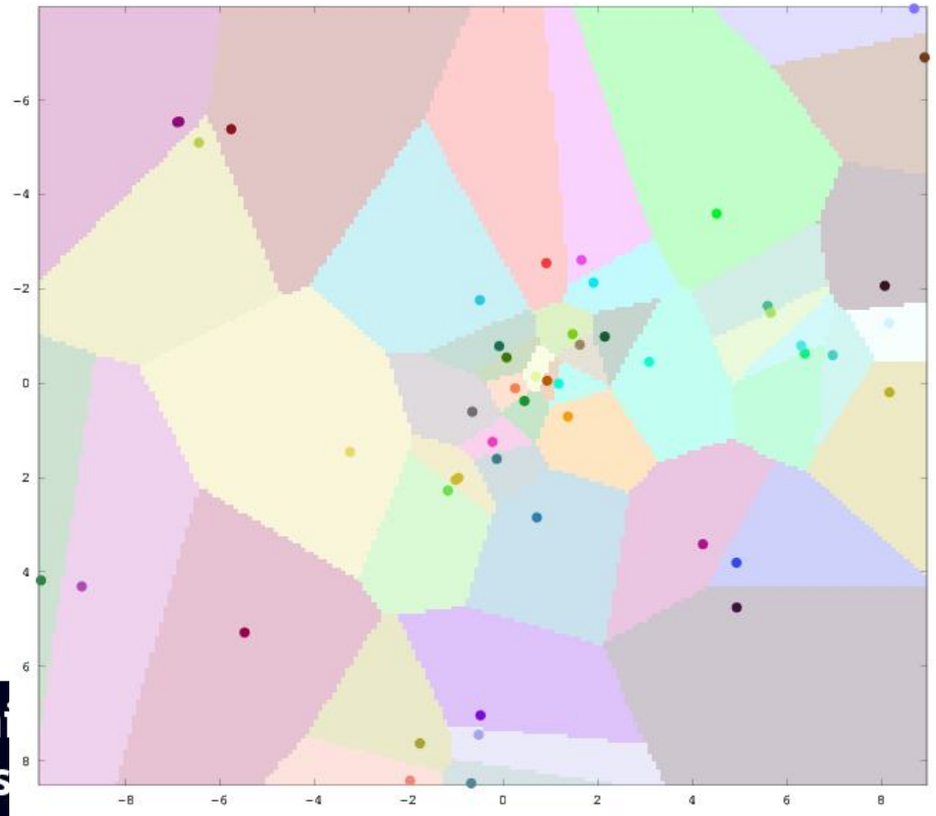
48 clusters

(merge below middle)

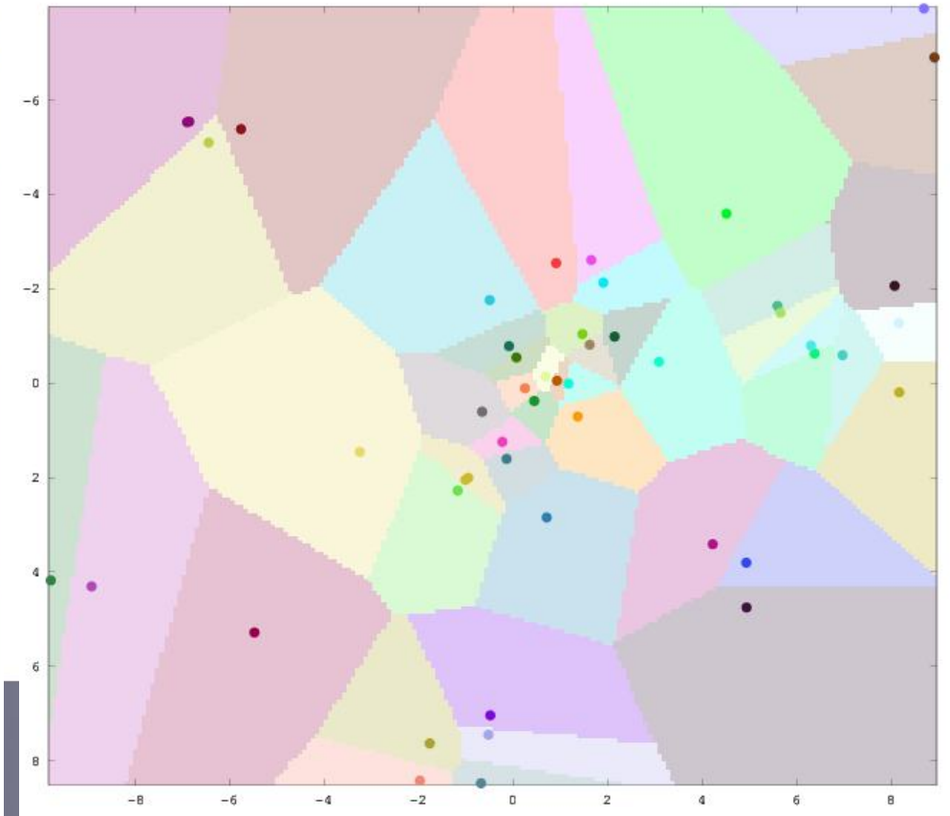


single-linkage

average linkage



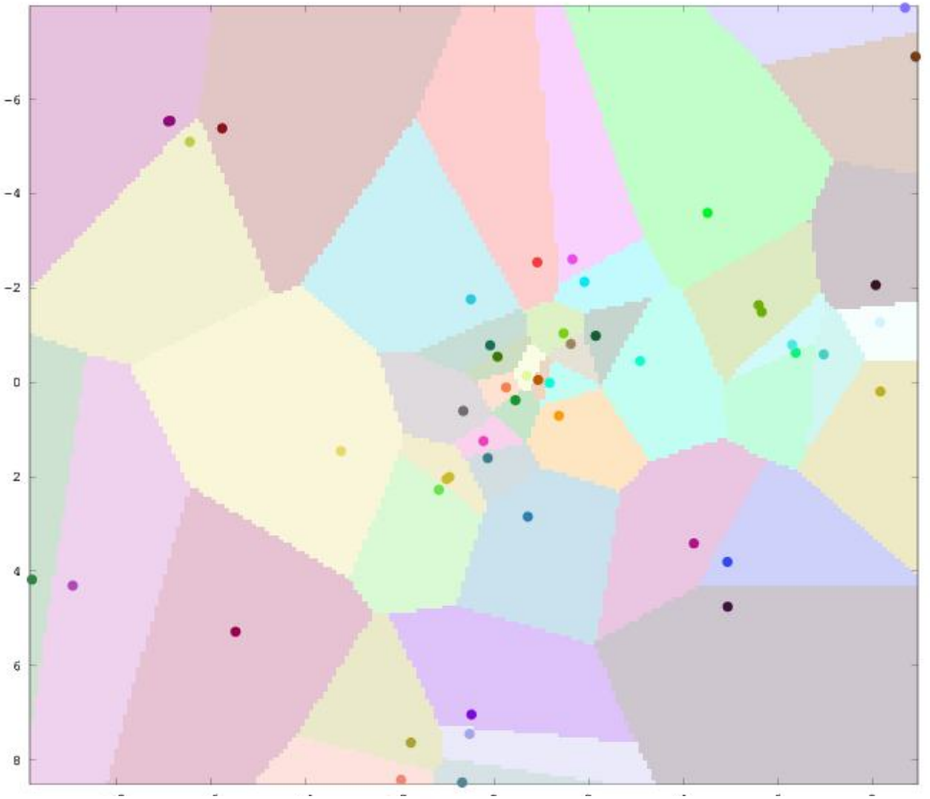
complete linkage



Example

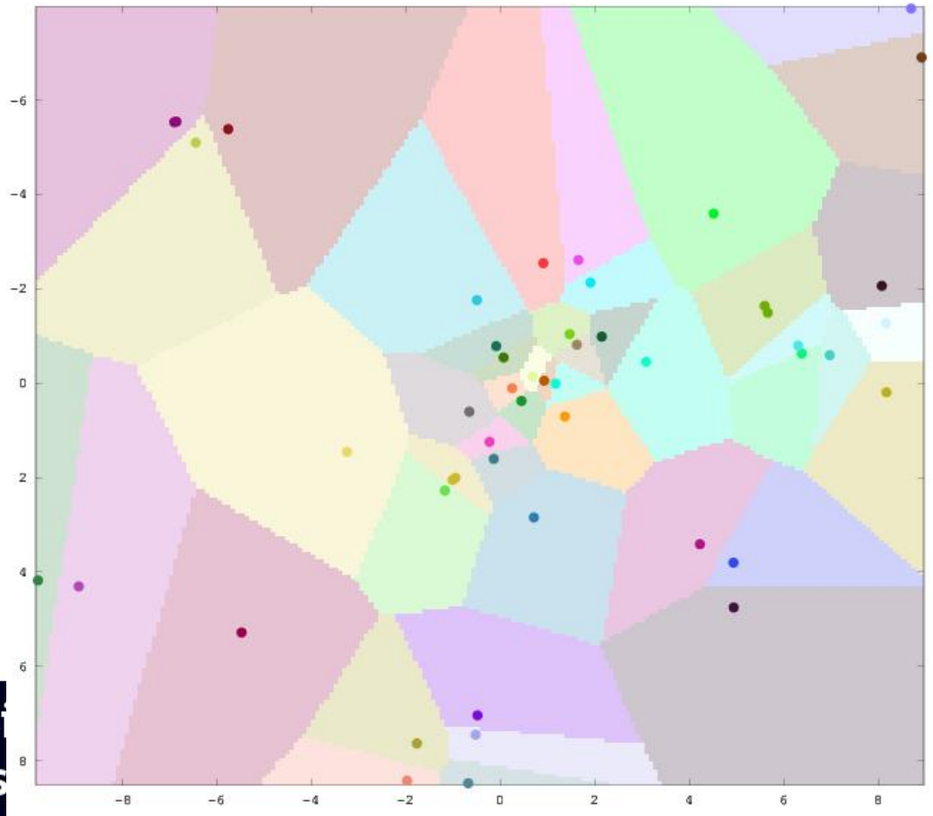
47 clusters

(merge at right)

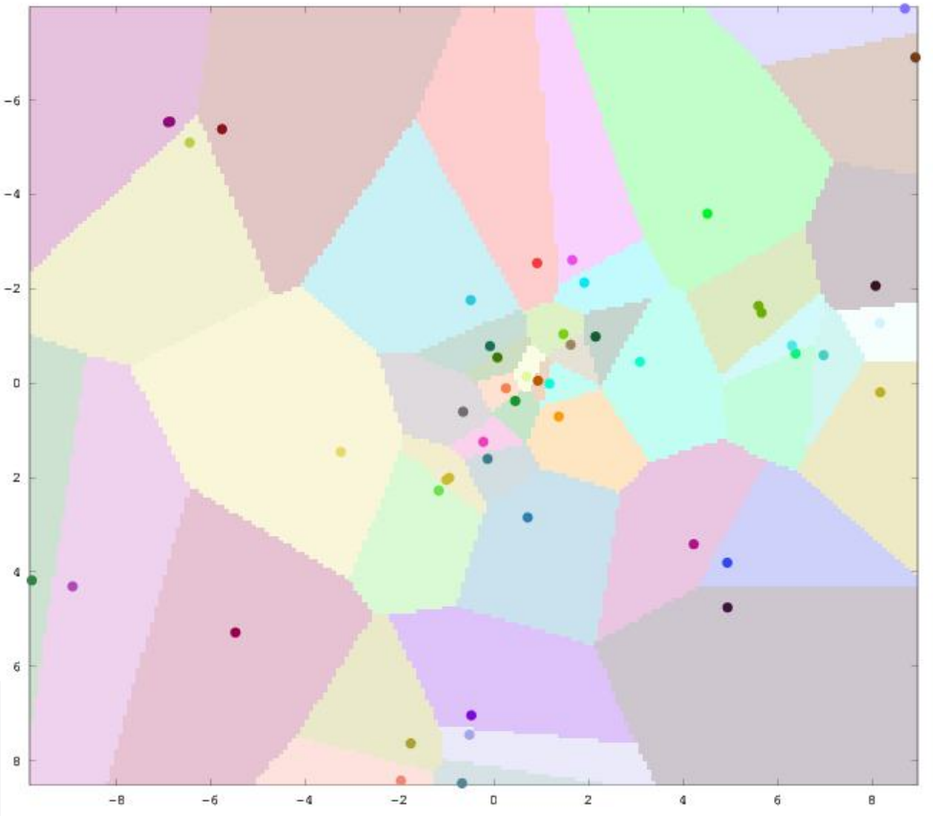


single-linkage

average linkage

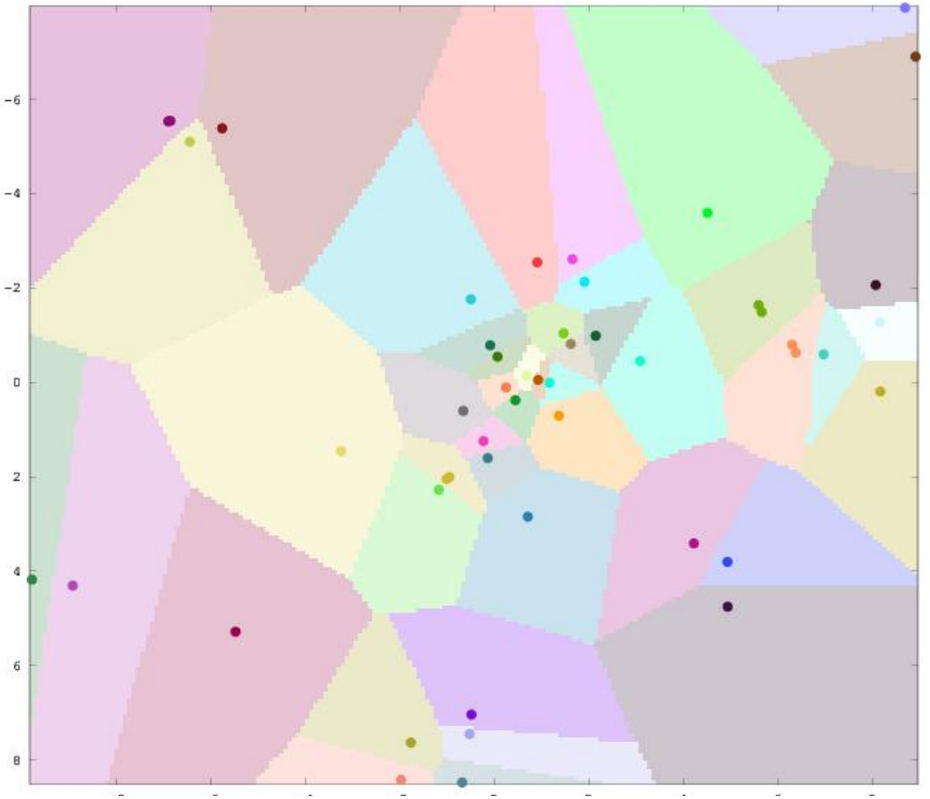


complete linkage



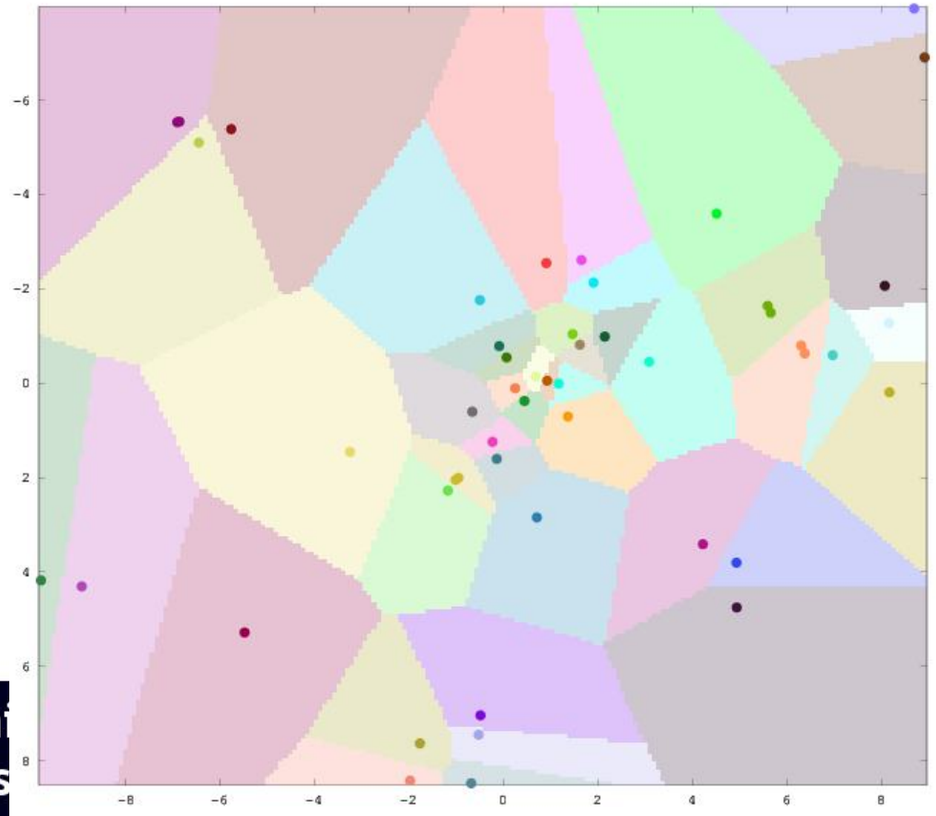
Example

46 clusters

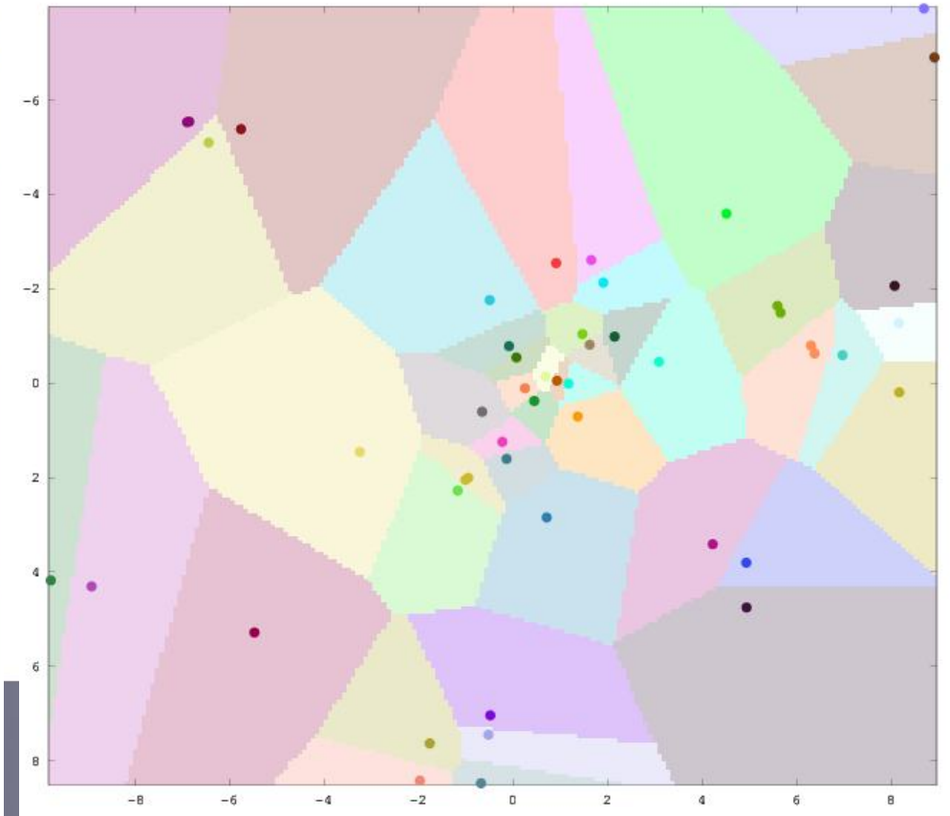


single-linkage

average linkage

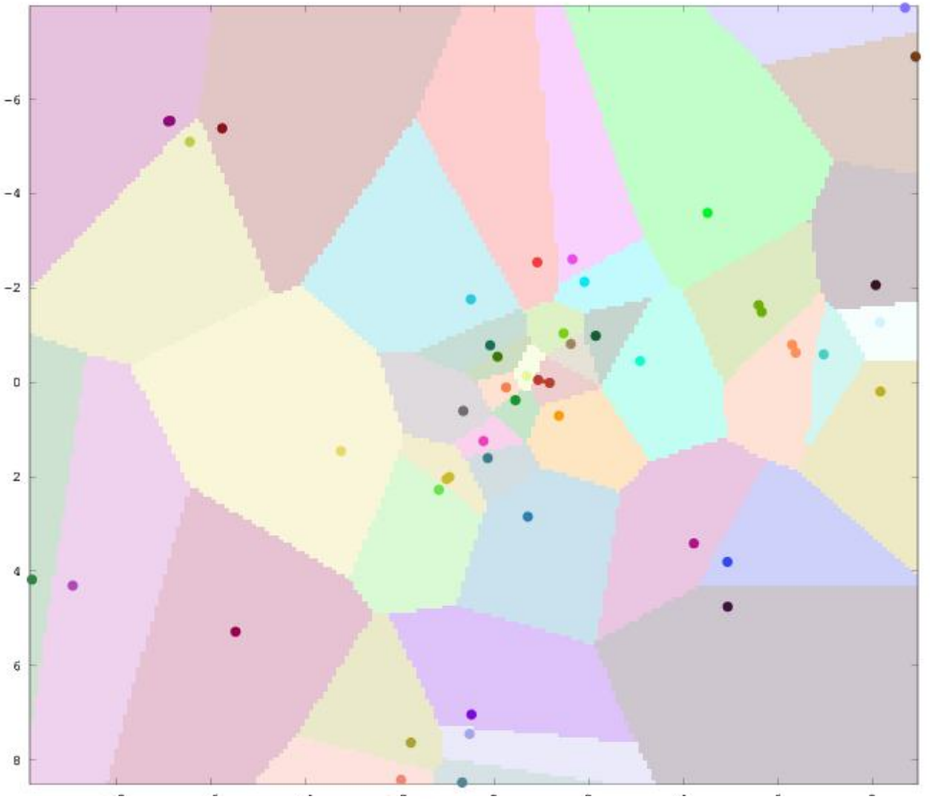


complete linkage



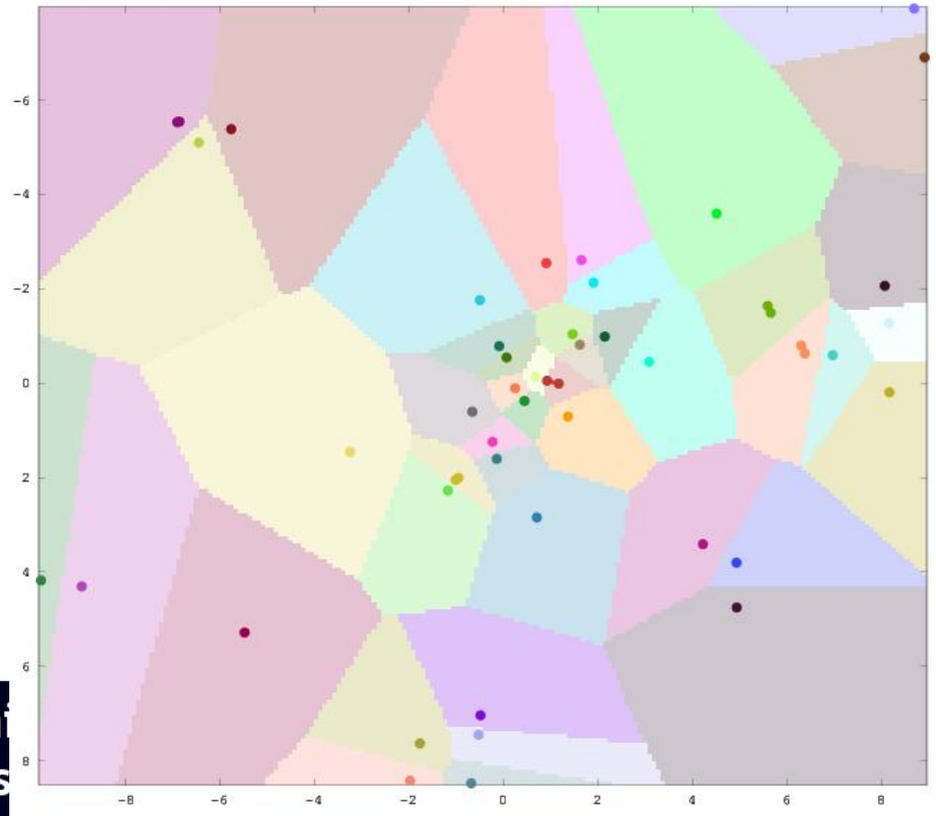
Example

45 clusters

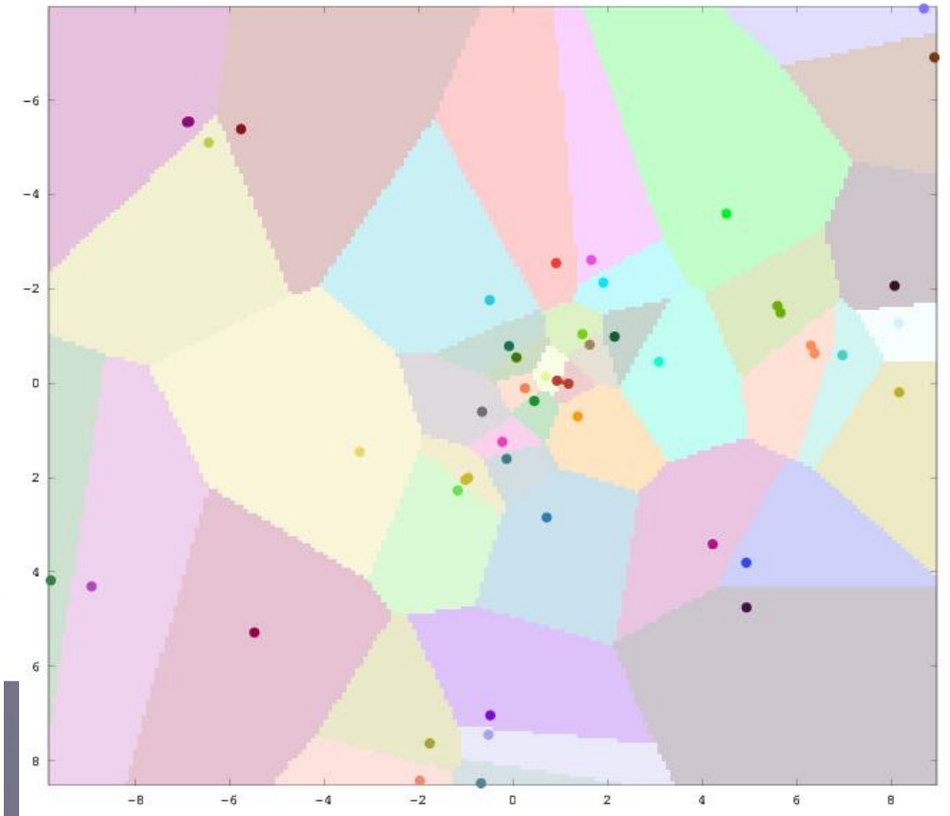


single-linkage

average linkage

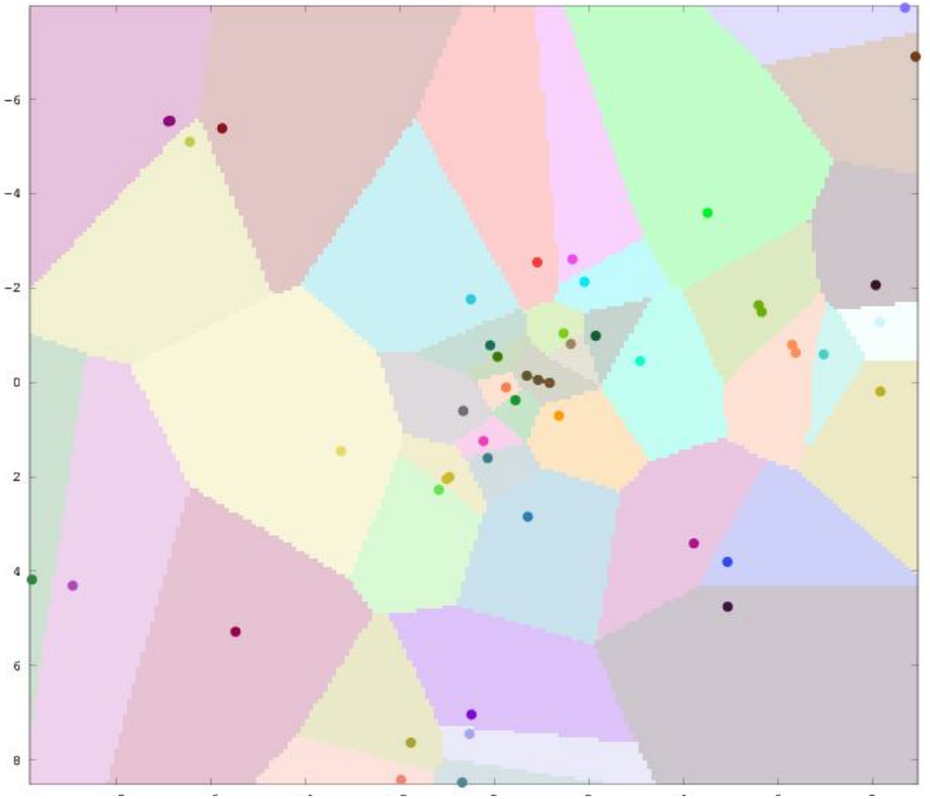


complete linkage



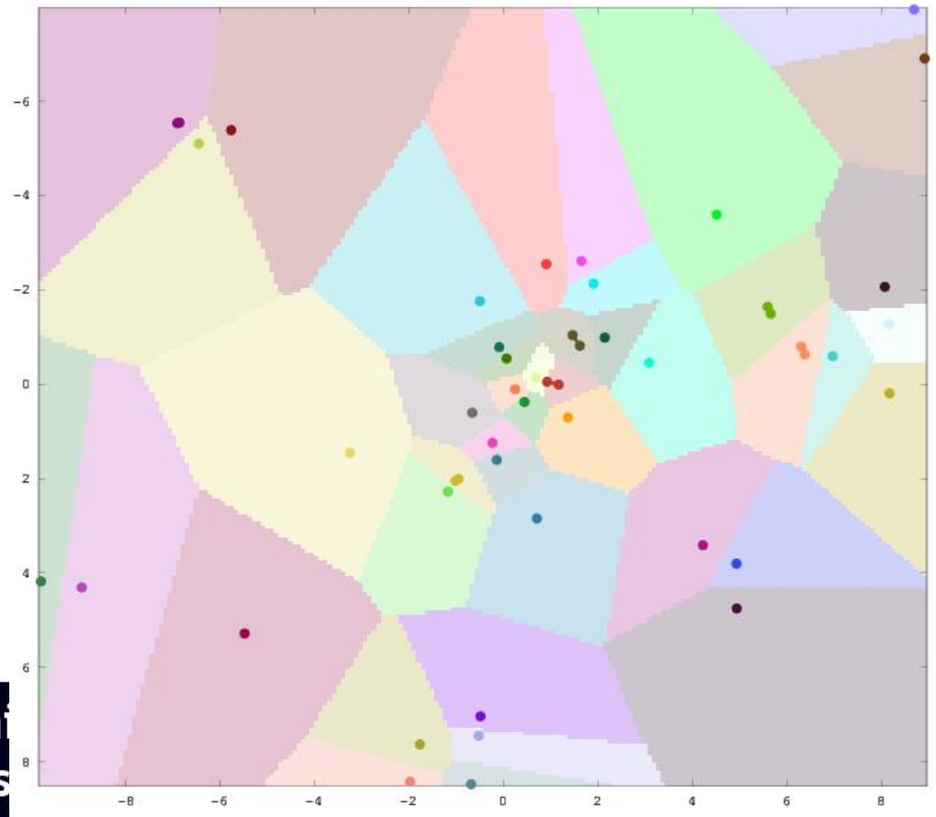
Example

44 clusters

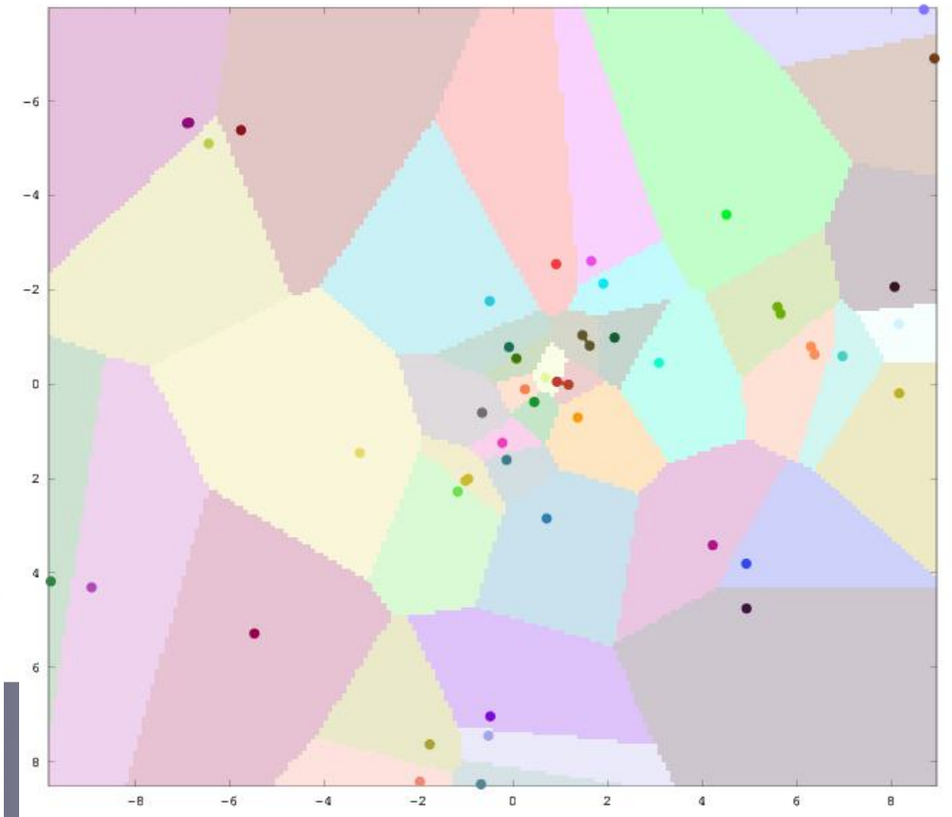


single-linkage

average linkage

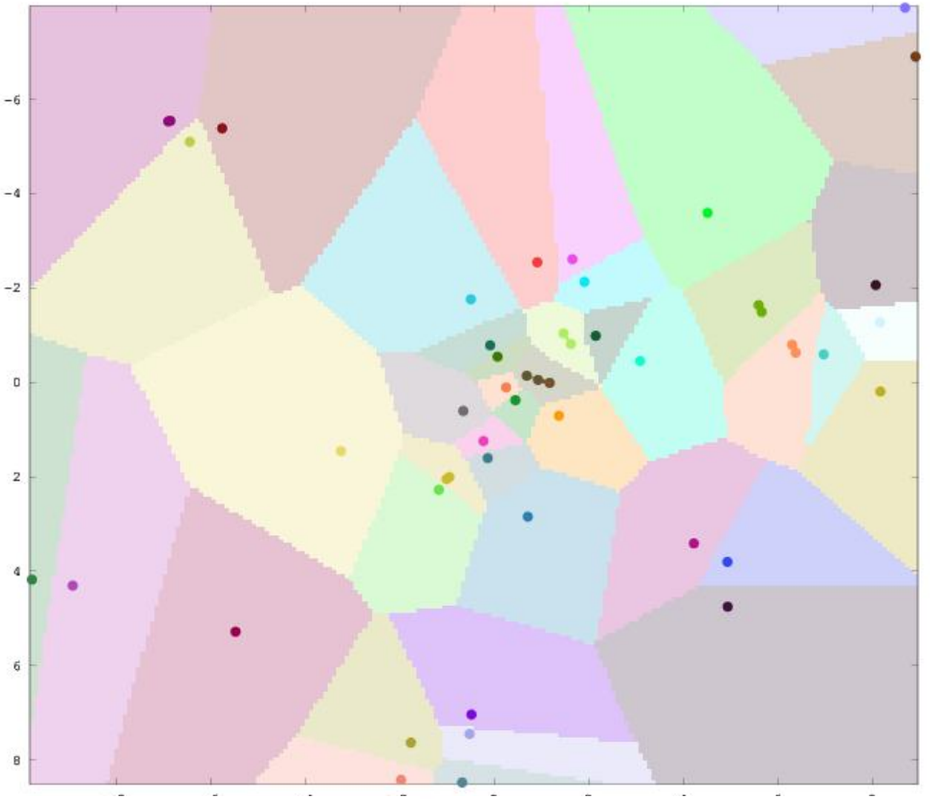


complete linkage



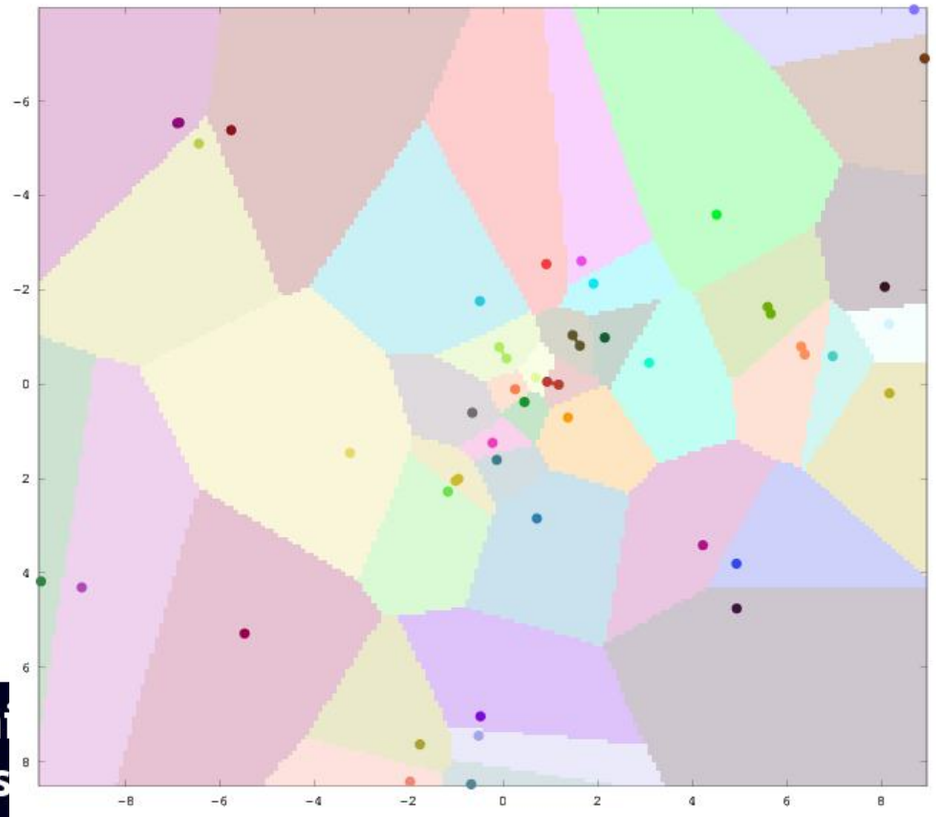
Example

43 clusters

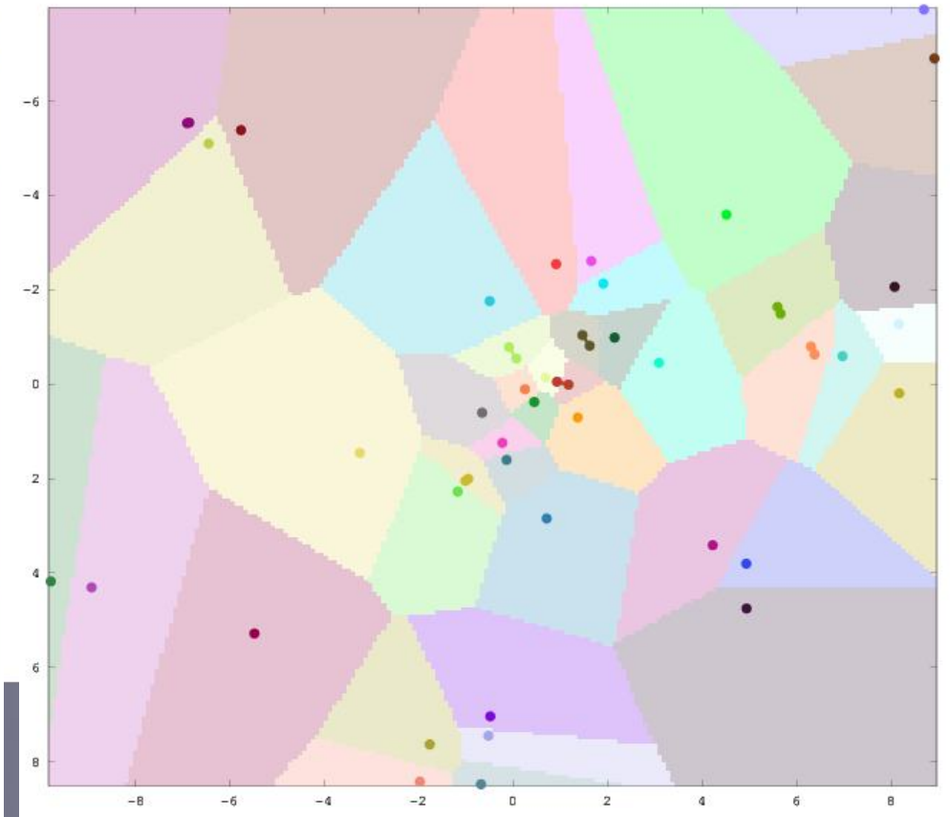


single-linkage

average linkage

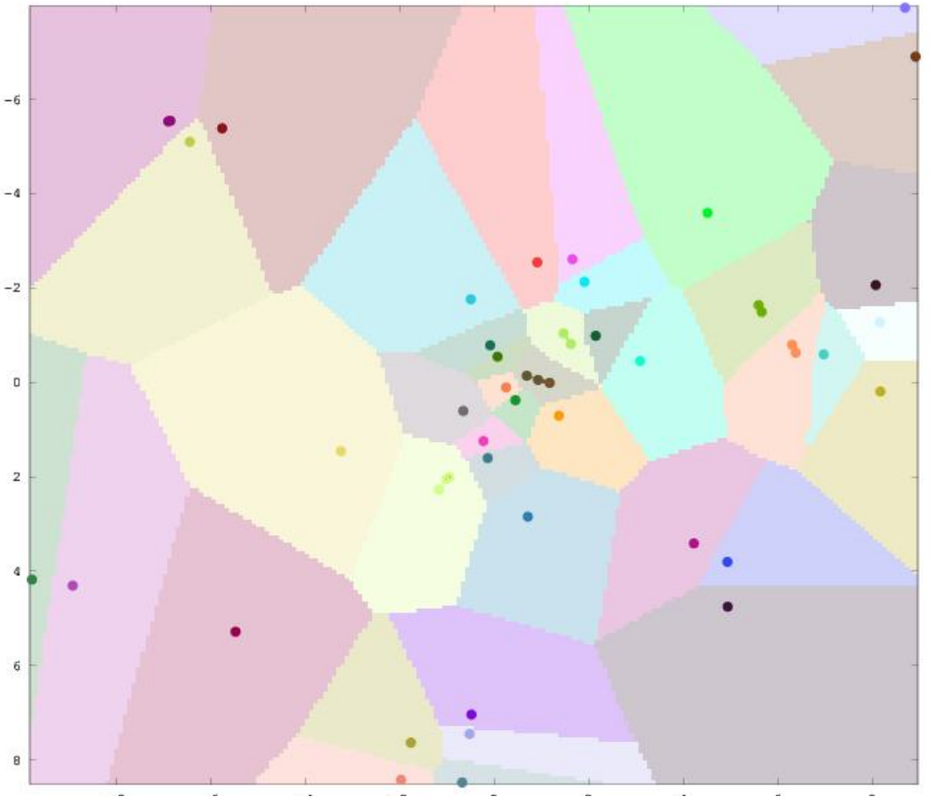


complete linkage



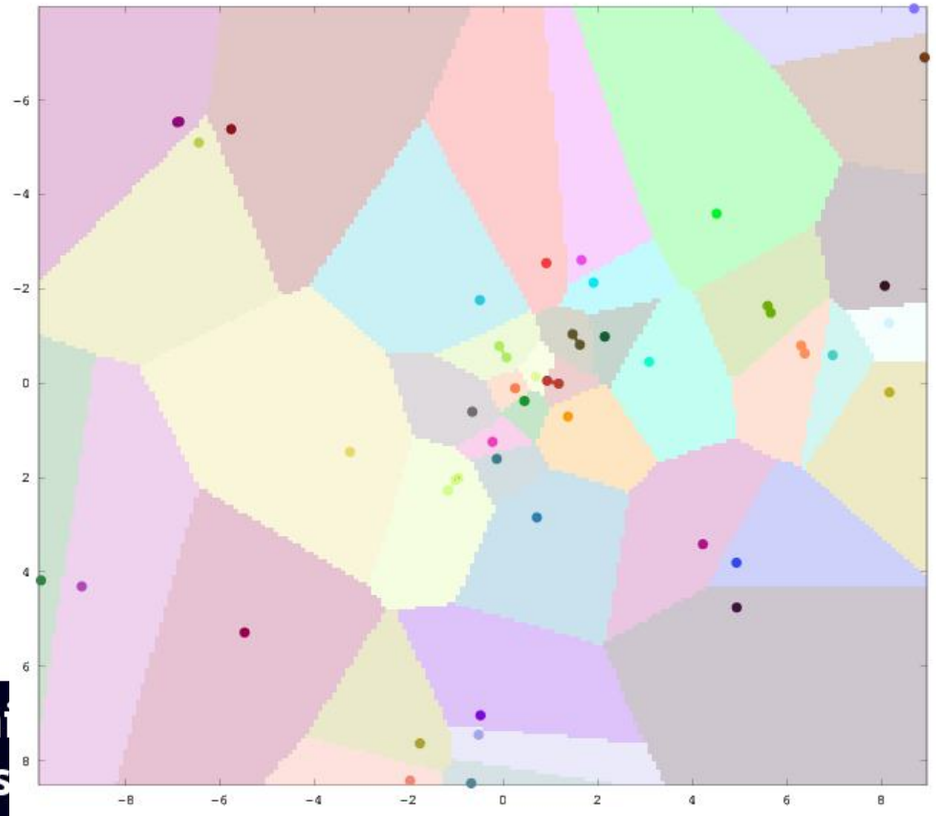
Example

42 clusters

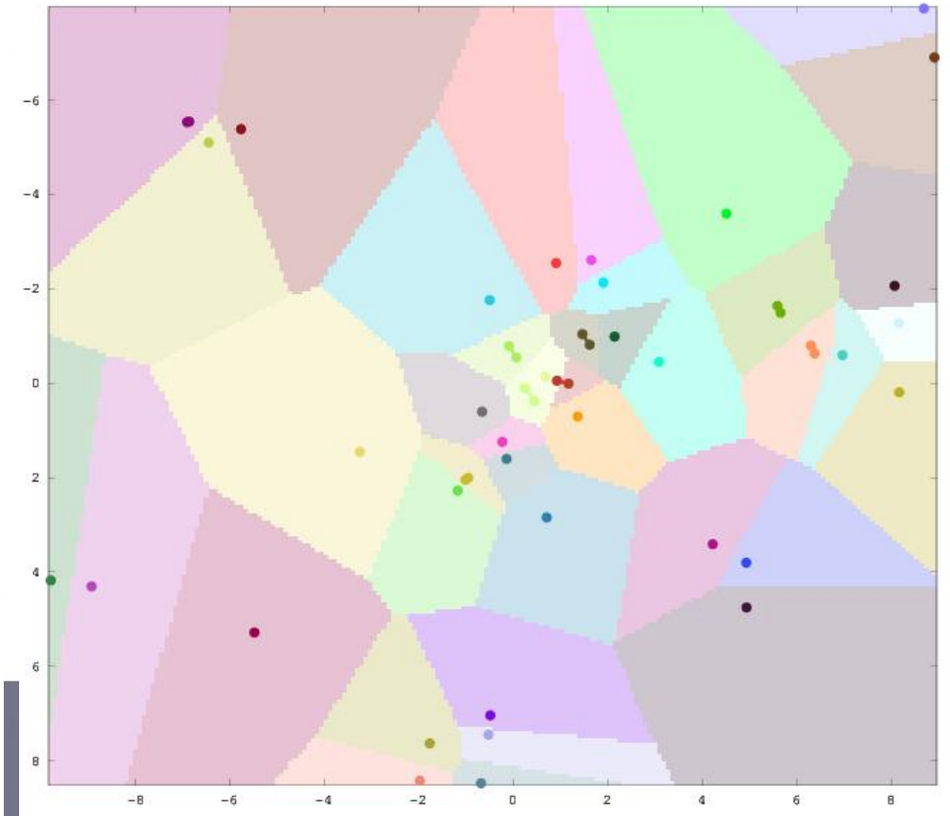


single-linkage

average linkage

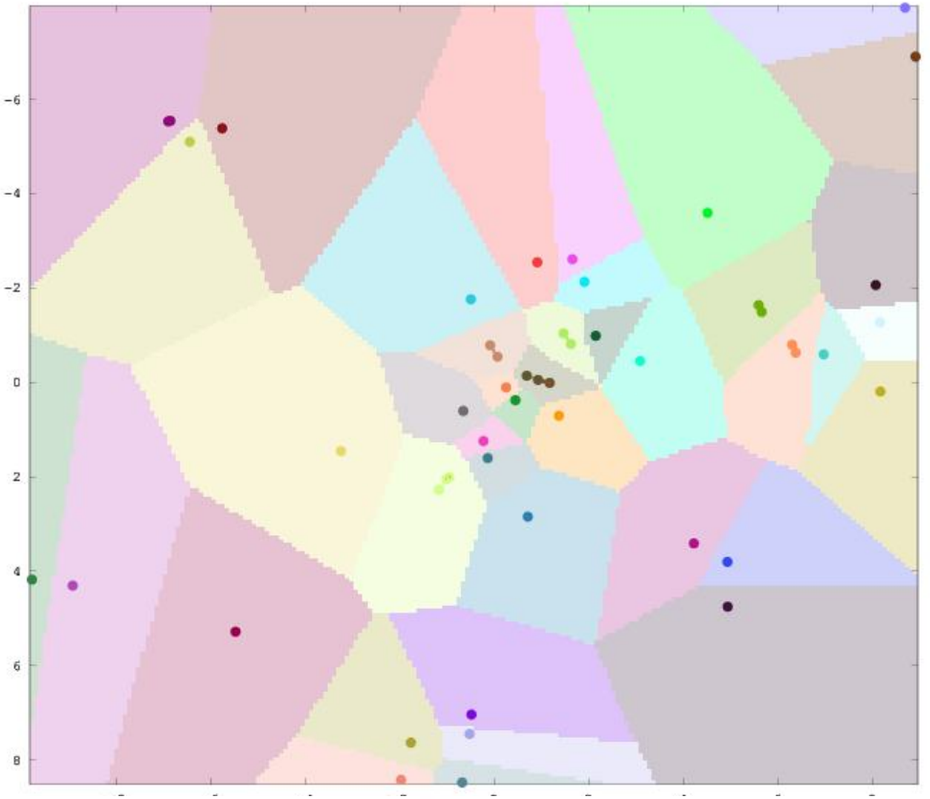


complete linkage



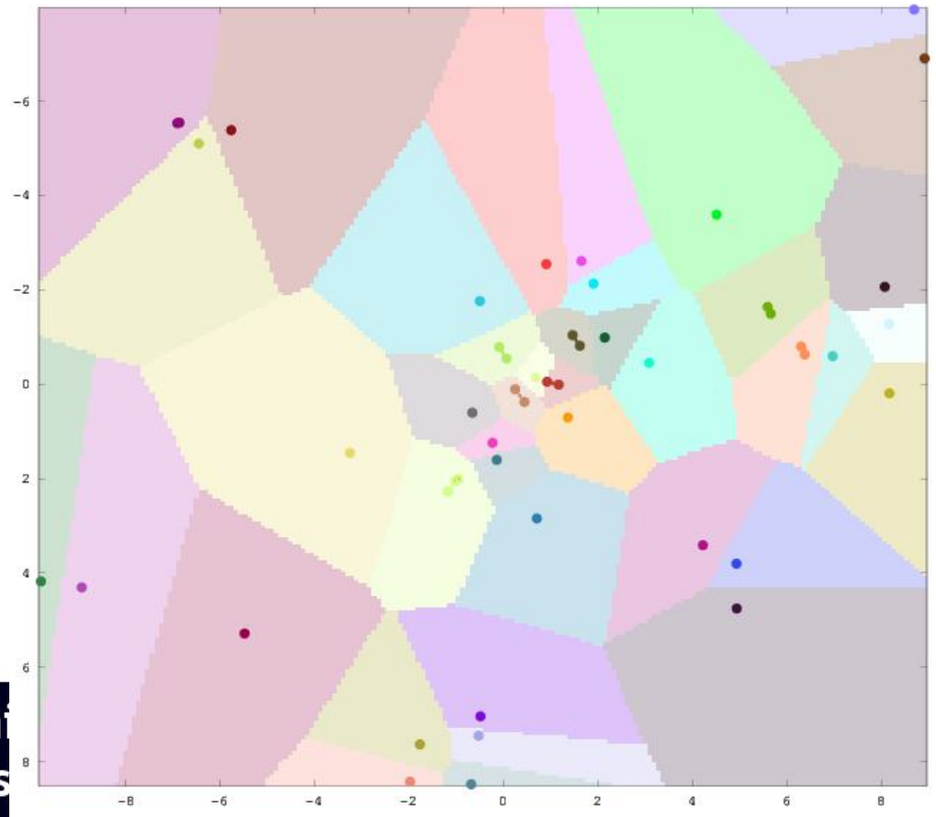
Example

41 clusters

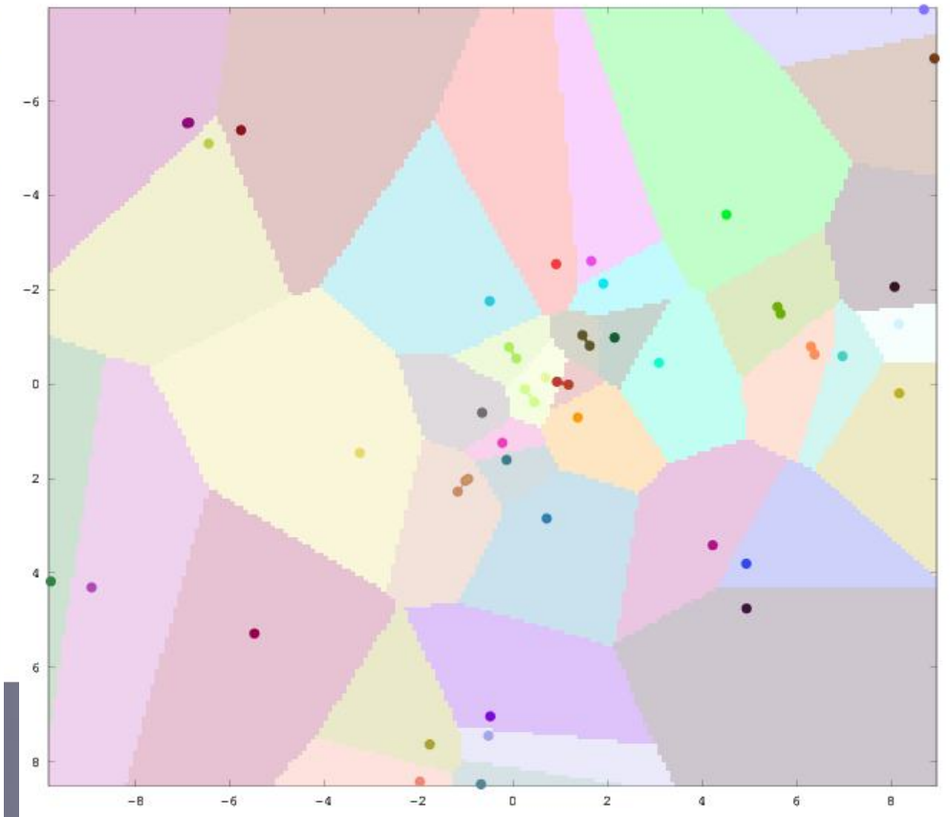


single-linkage

average linkage

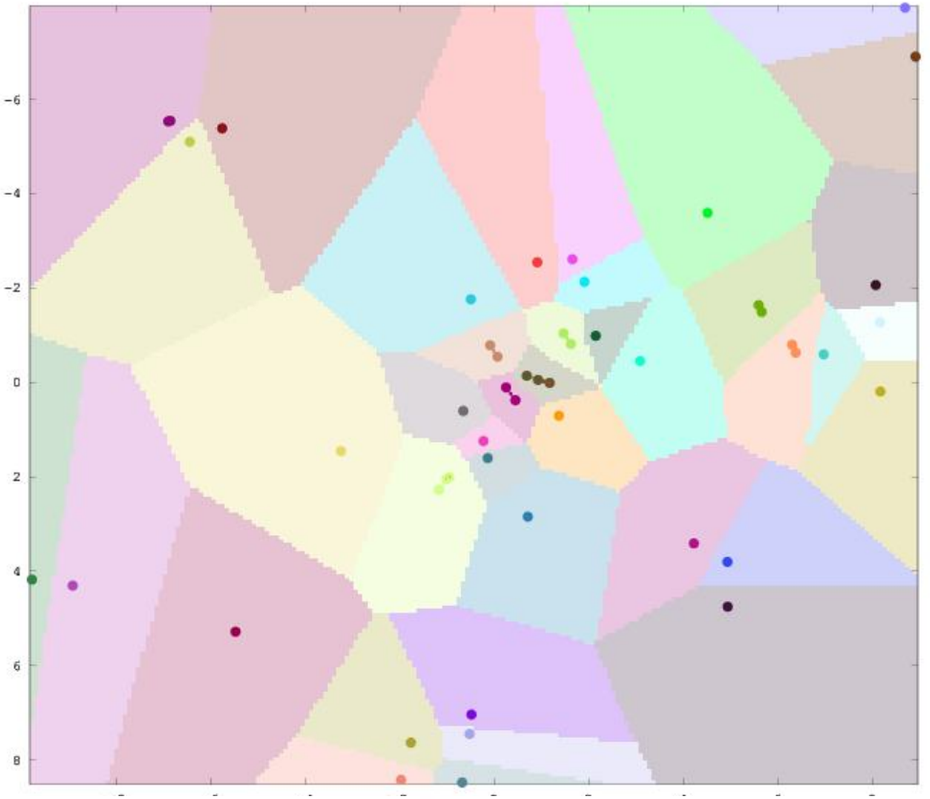


complete linkage



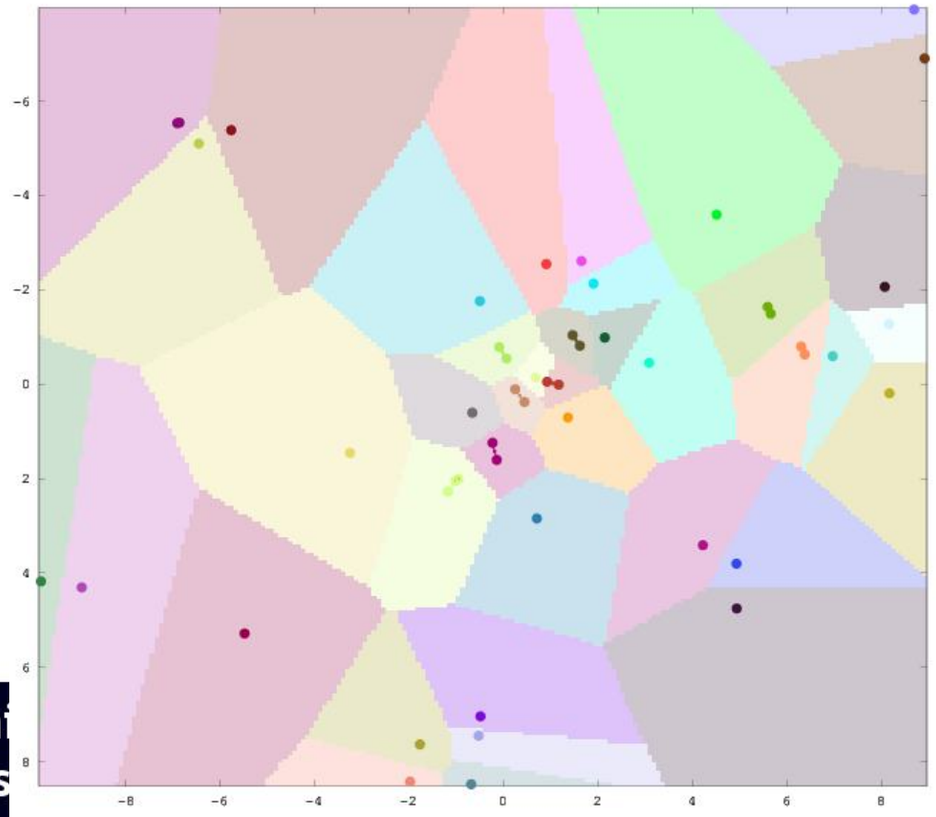
Example

40 clusters

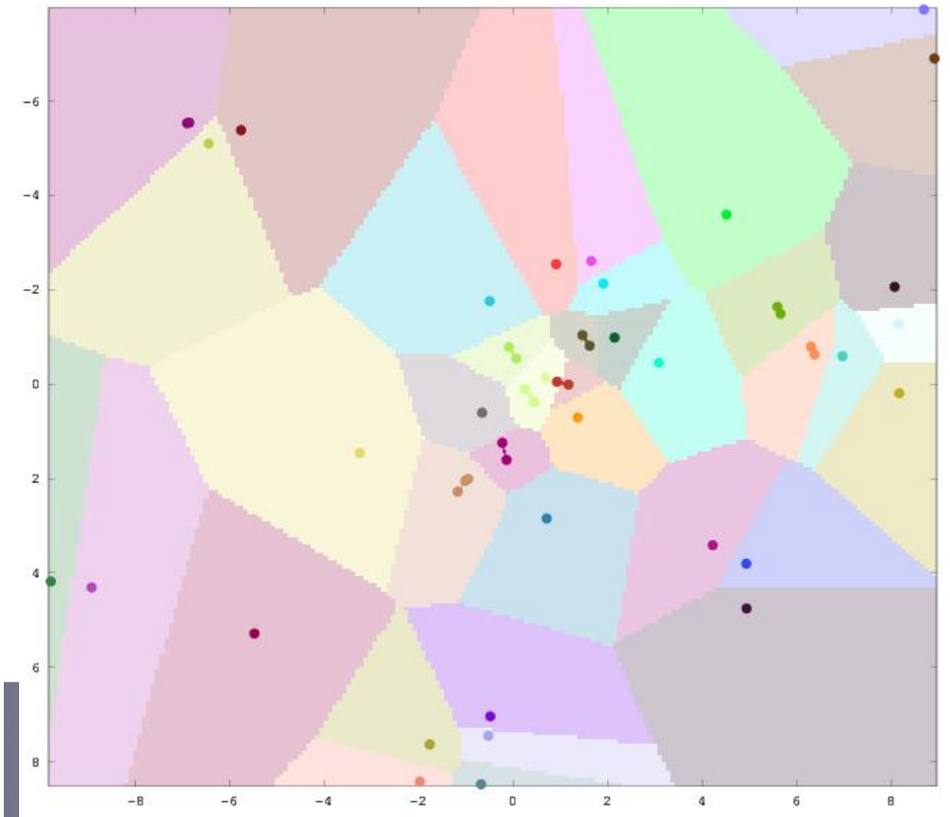


single-linkage

average linkage

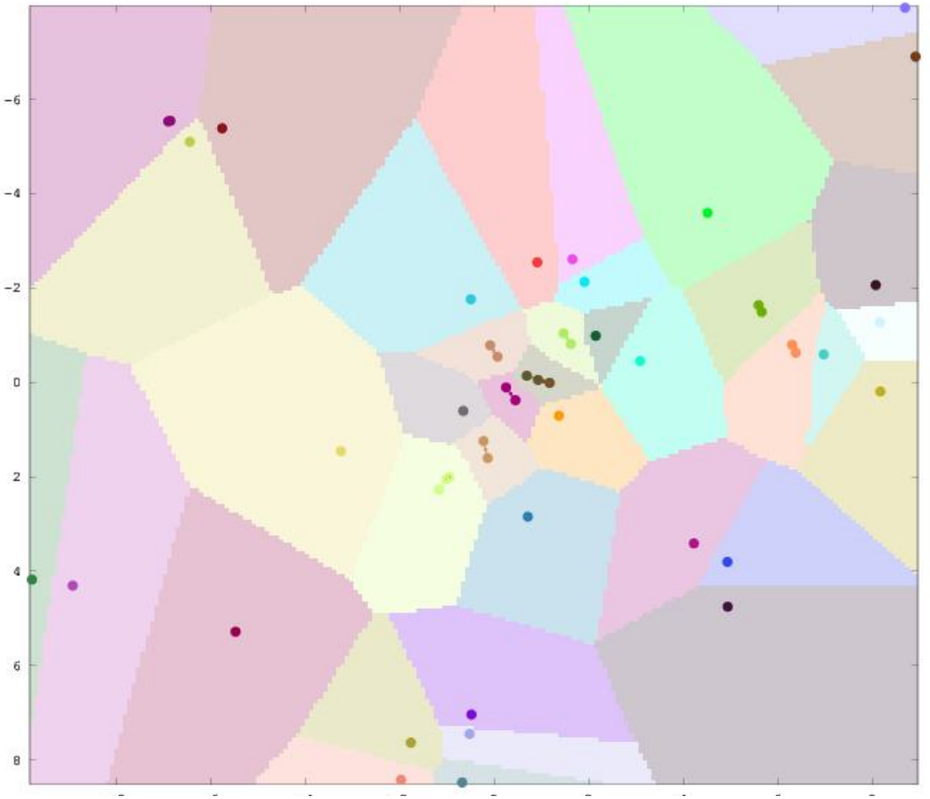


complete linkage



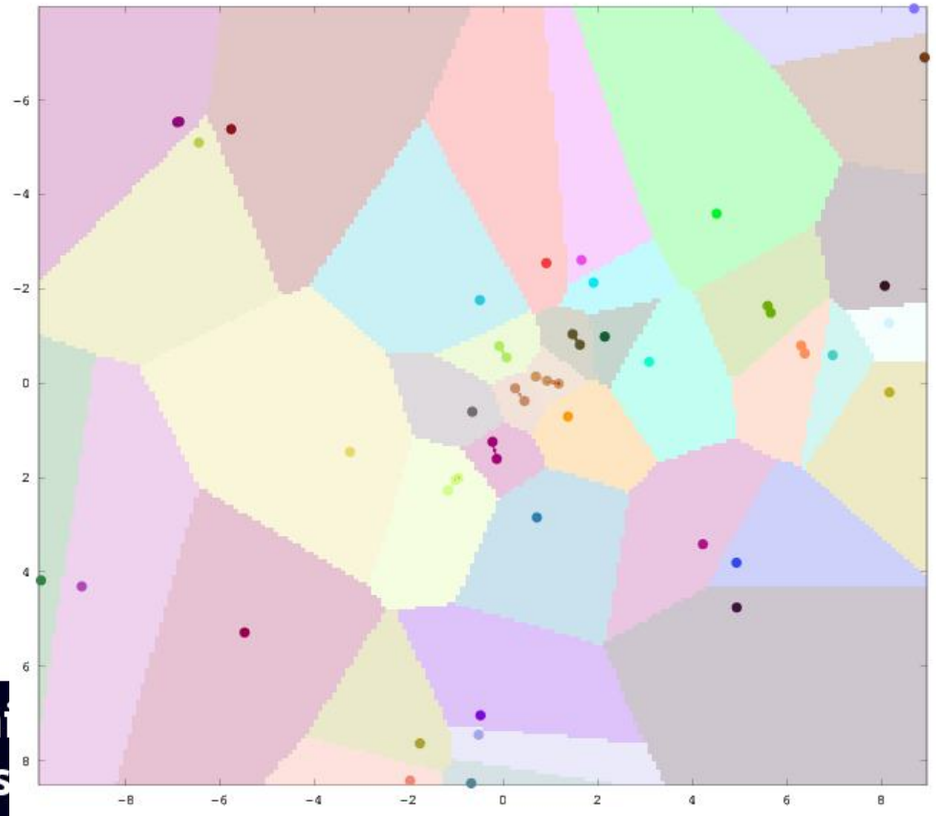
Example

39 clusters

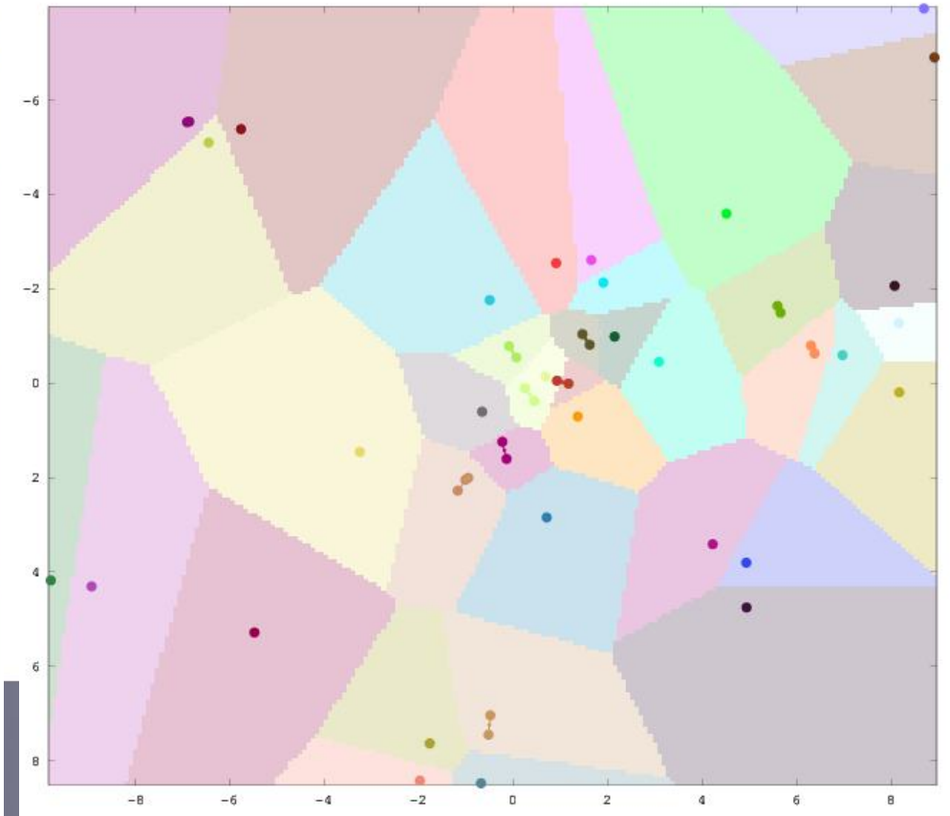


single-linkage

average linkage

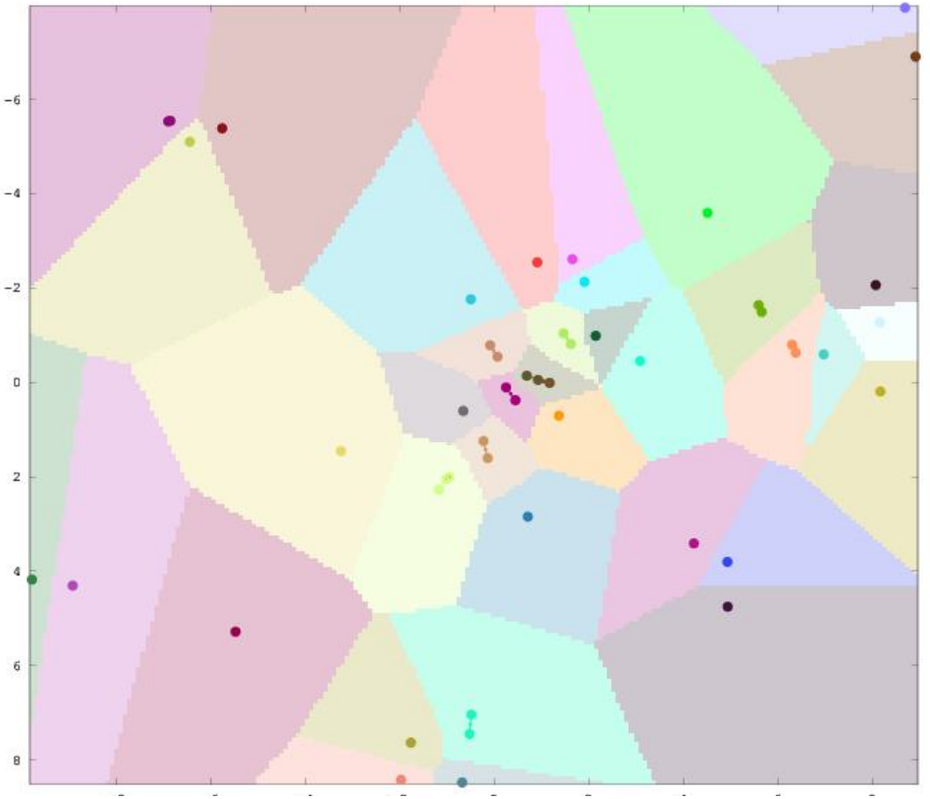


complete linkage



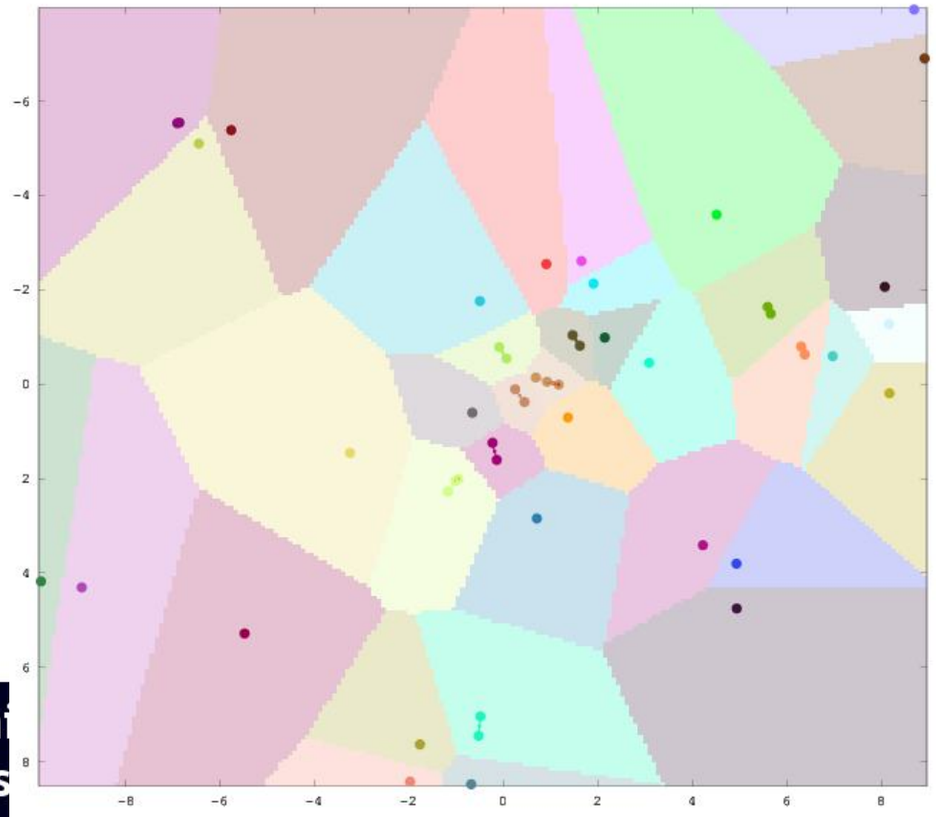
Example

38 clusters

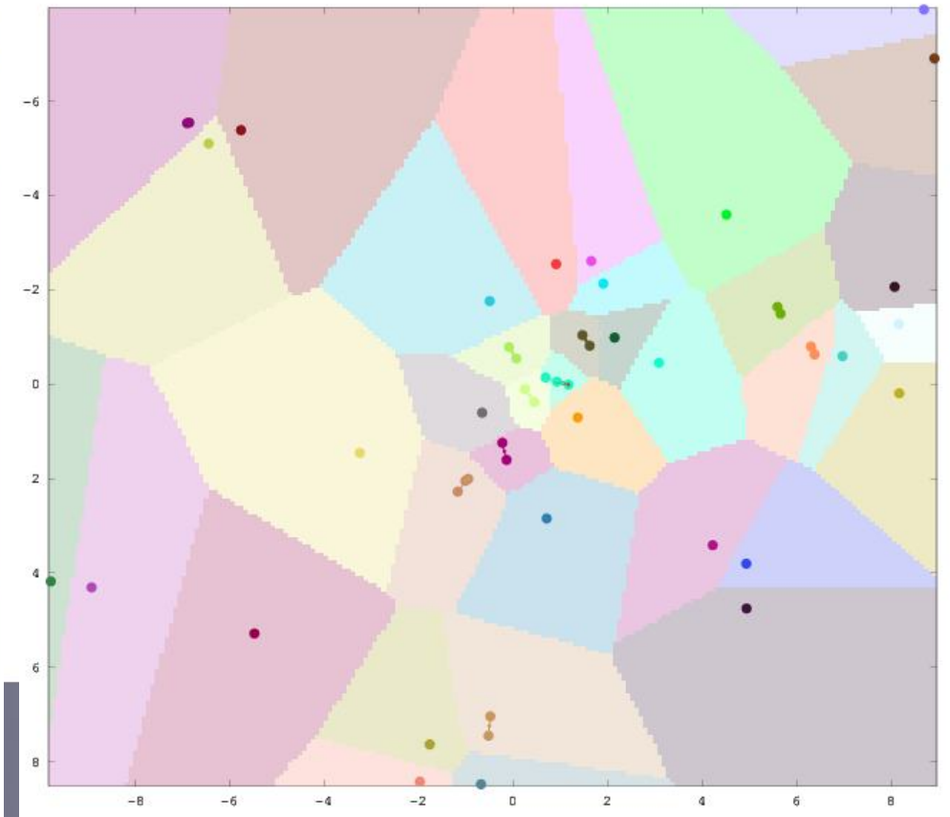


single-linkage

average linkage

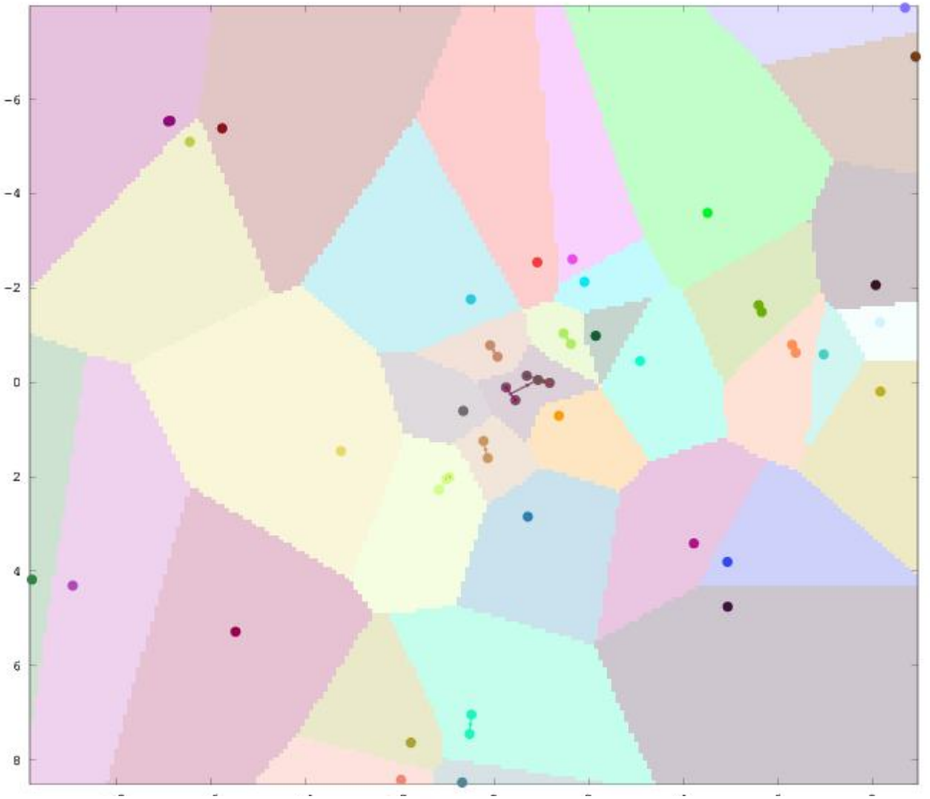


complete linkage



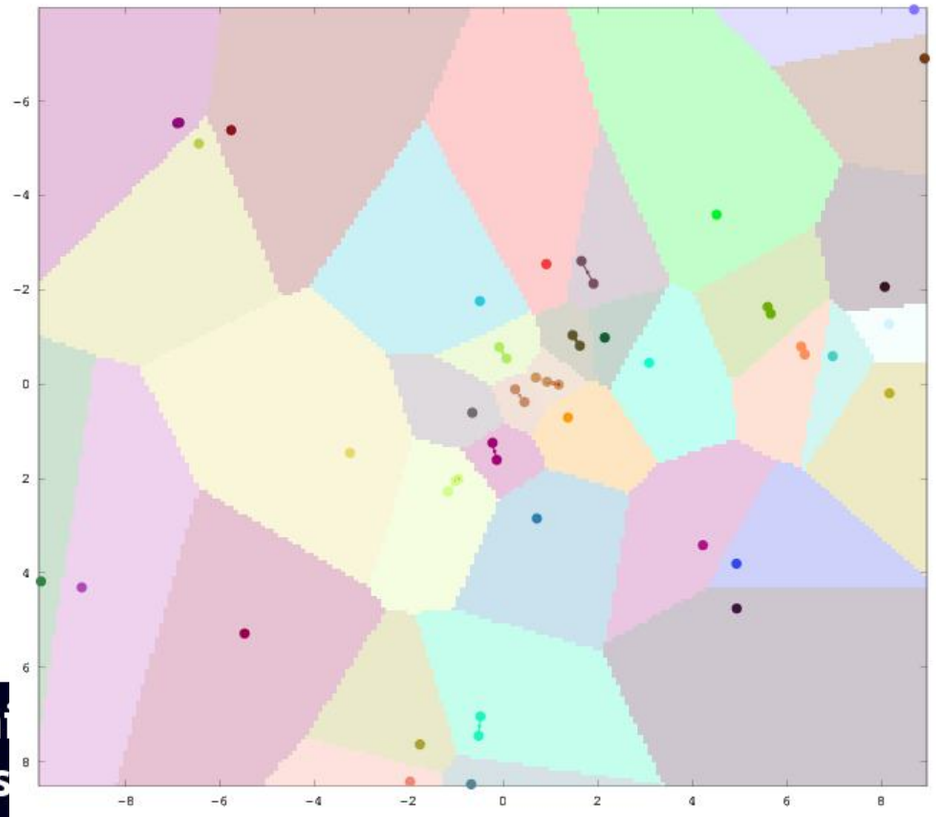
Example

37 clusters

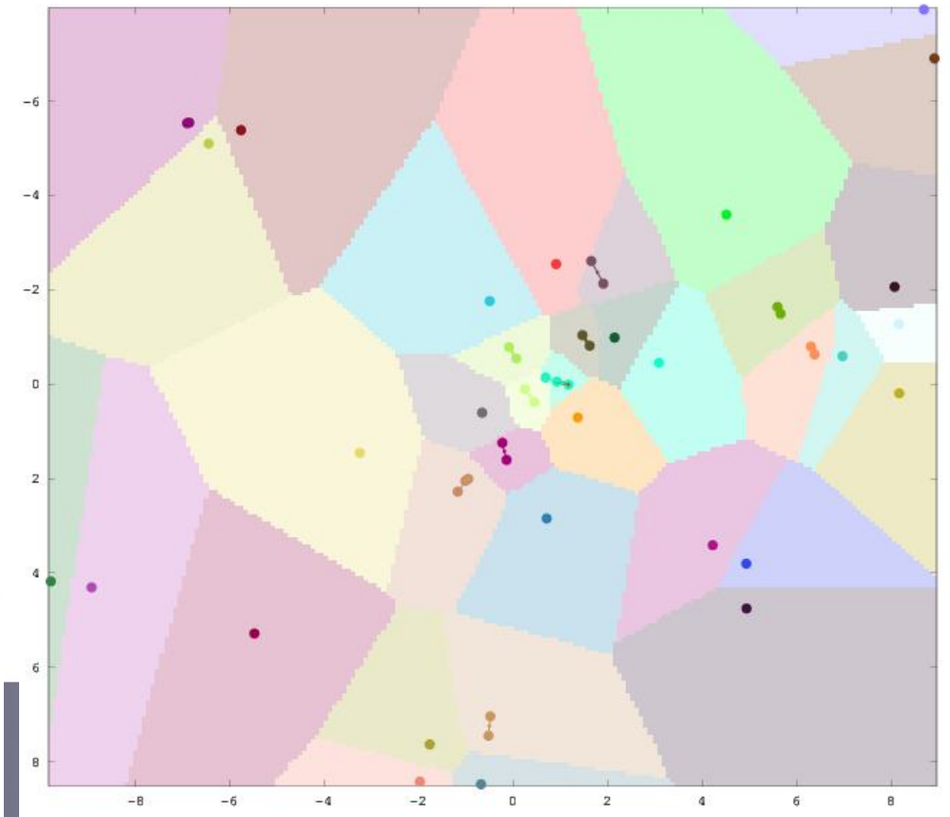


single-linkage

average linkage

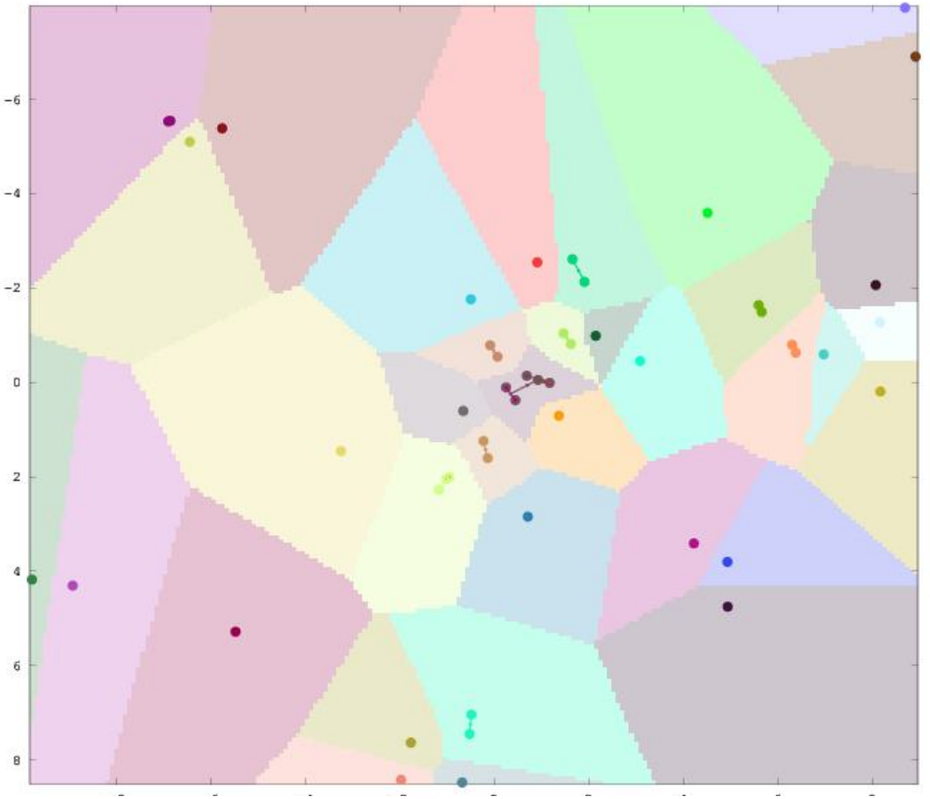


complete linkage



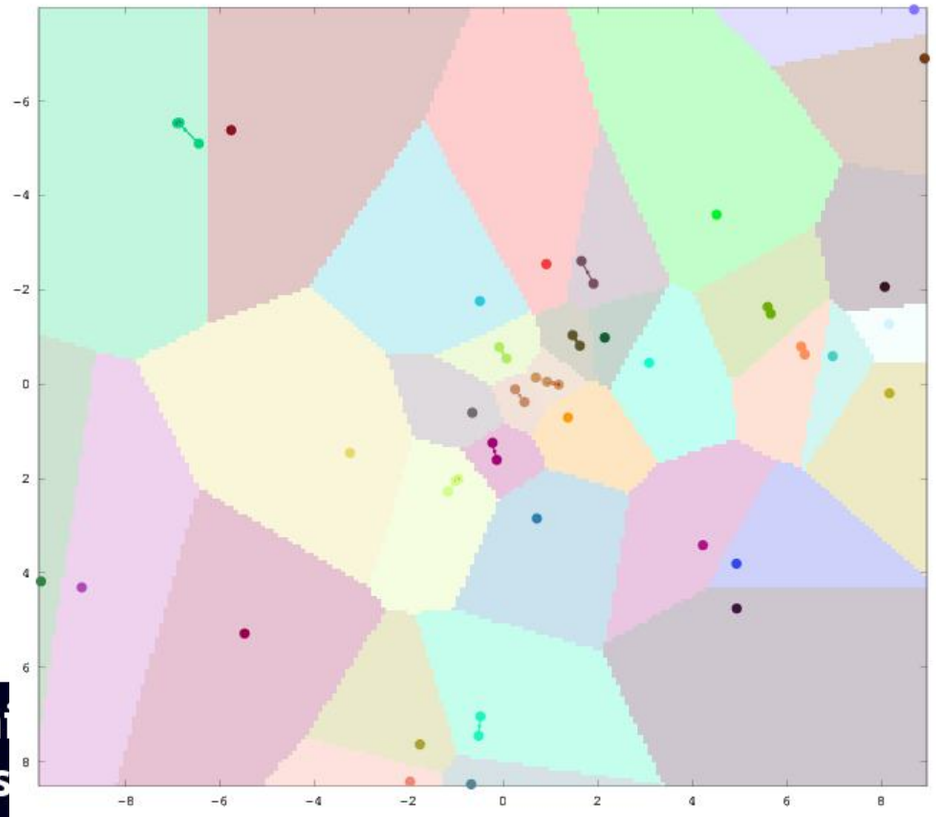
Example

36 clusters

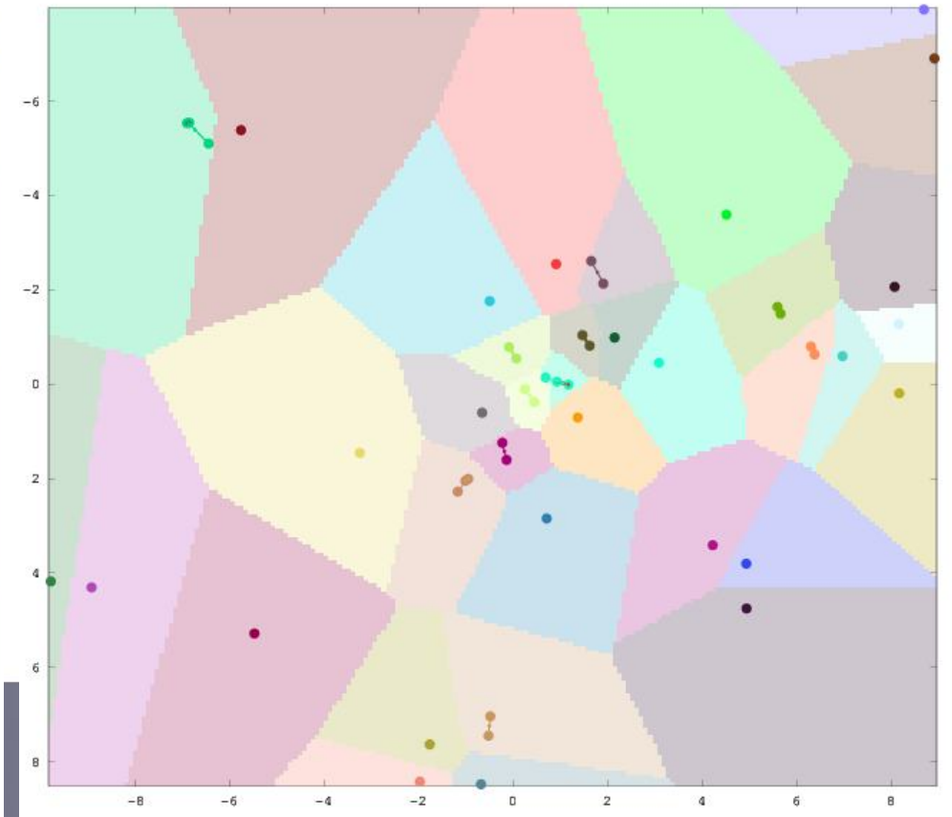


single-linkage

average linkage

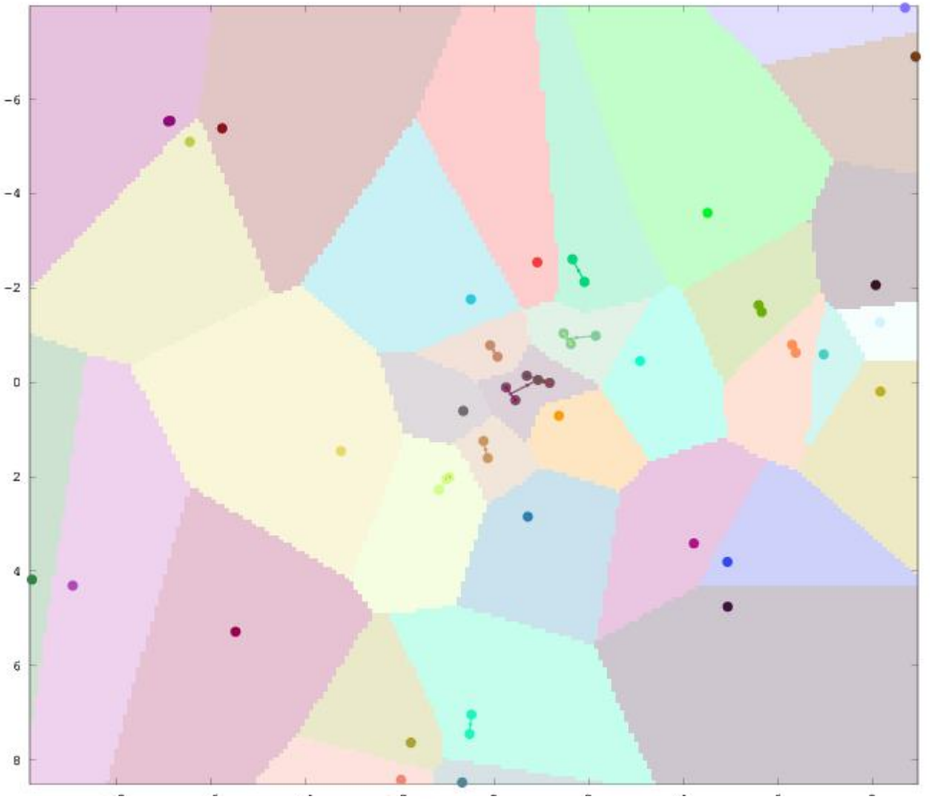


complete linkage



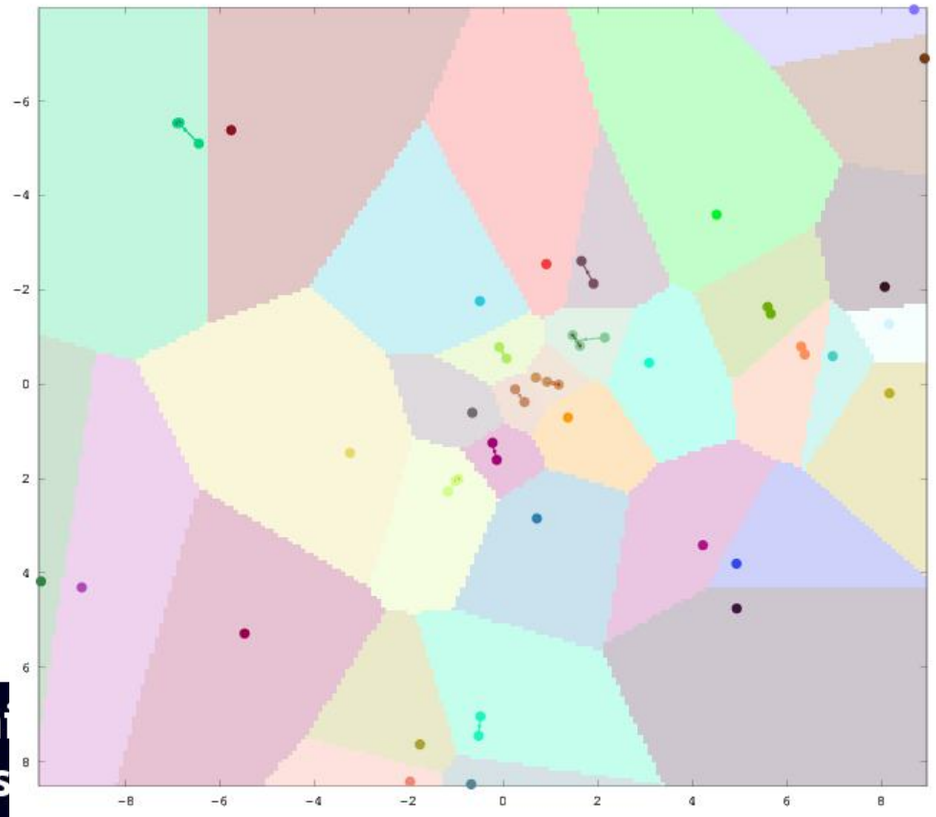
Example

35 clusters

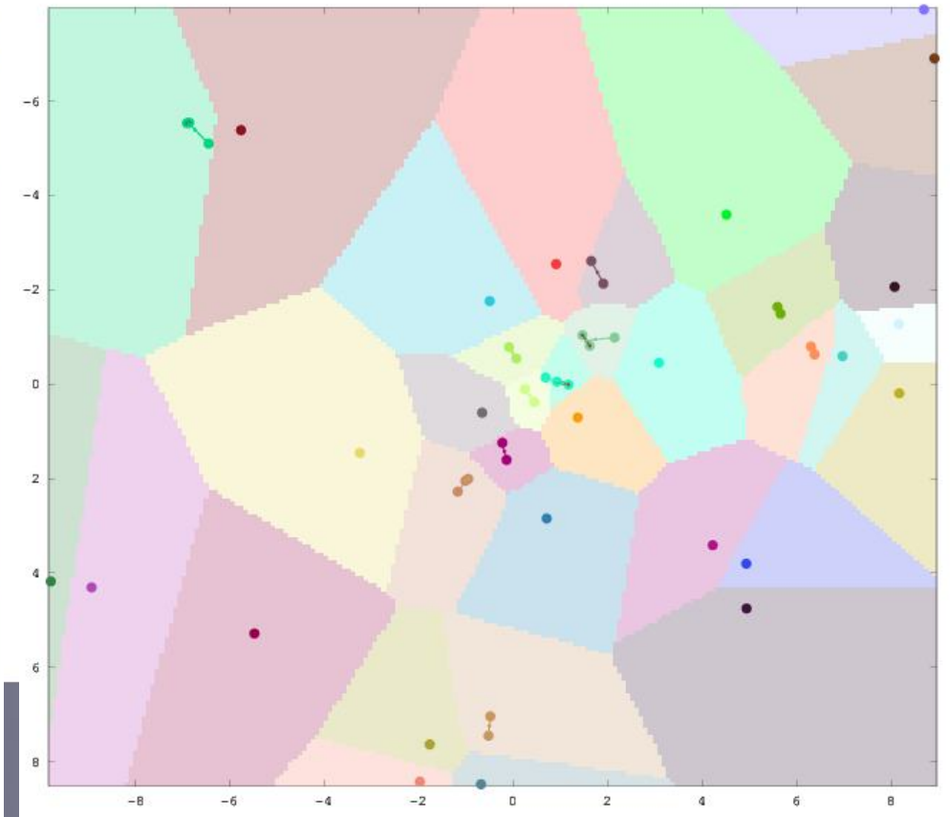


single-linkage

average linkage

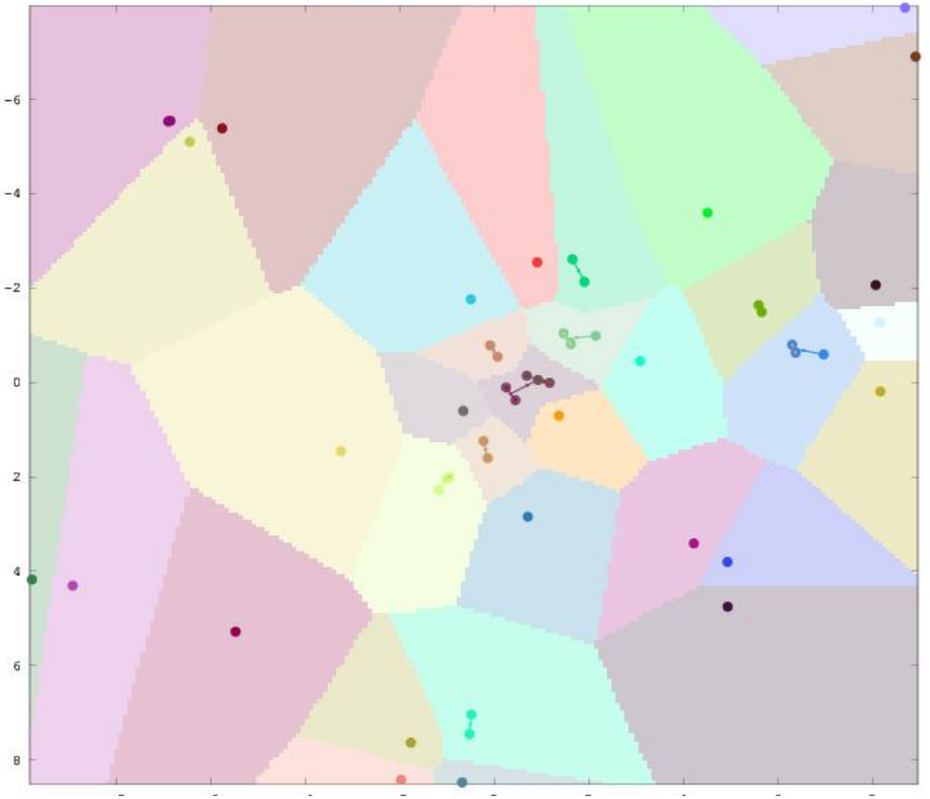


complete linkage



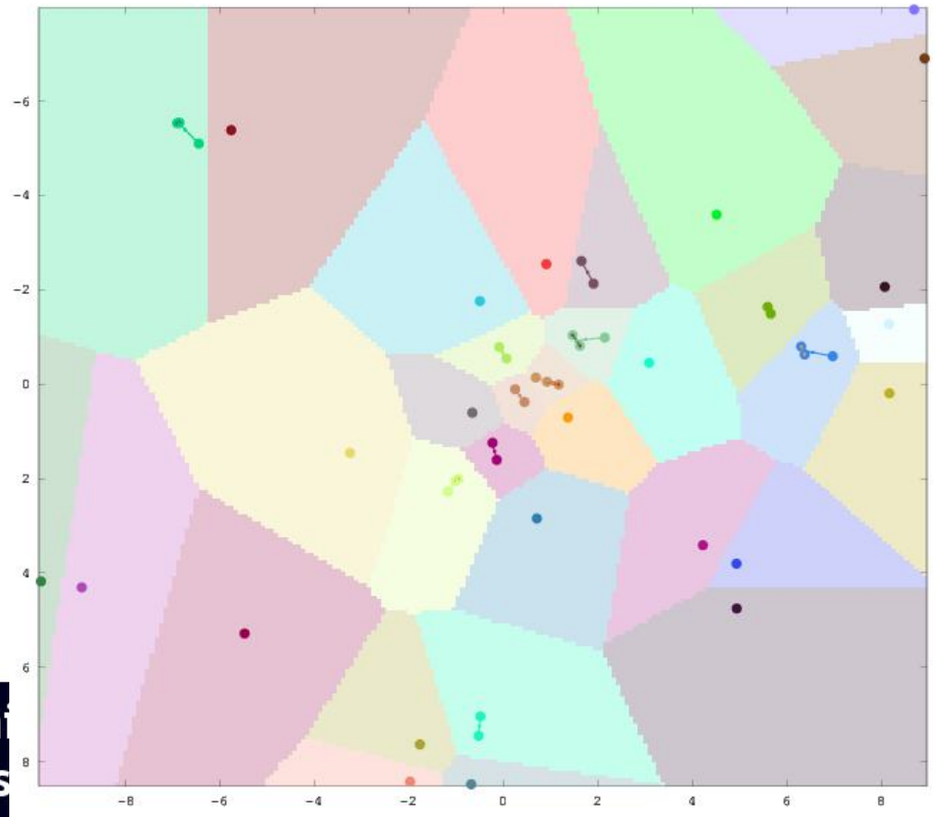
Example

34 clusters

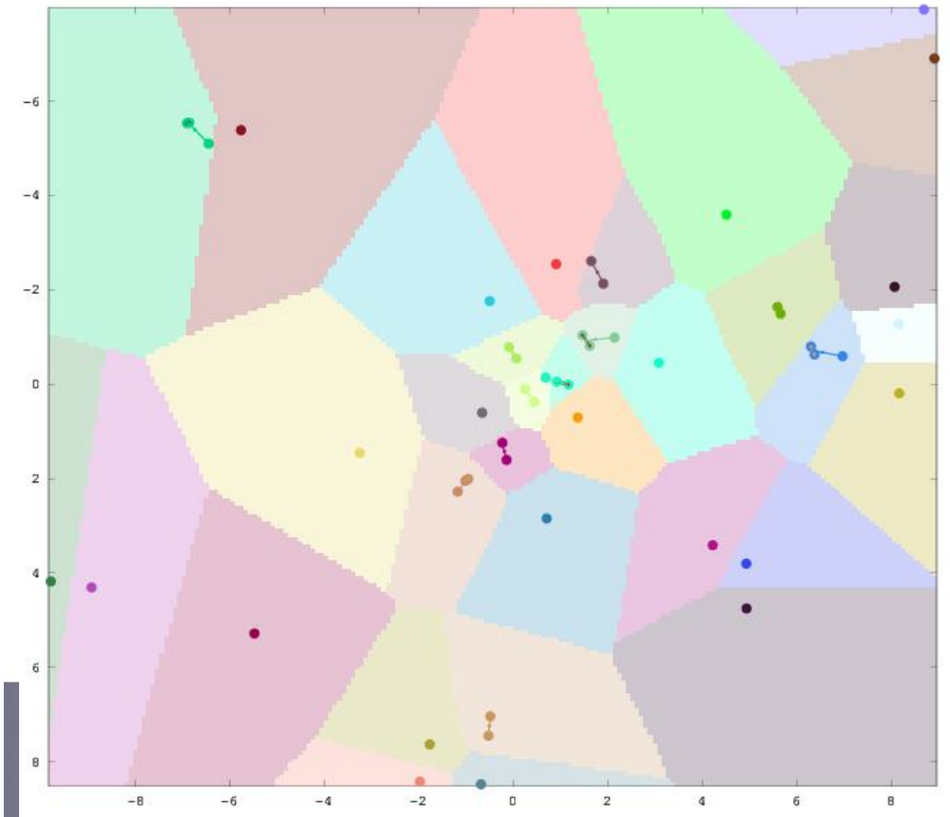


single-linkage

average linkage

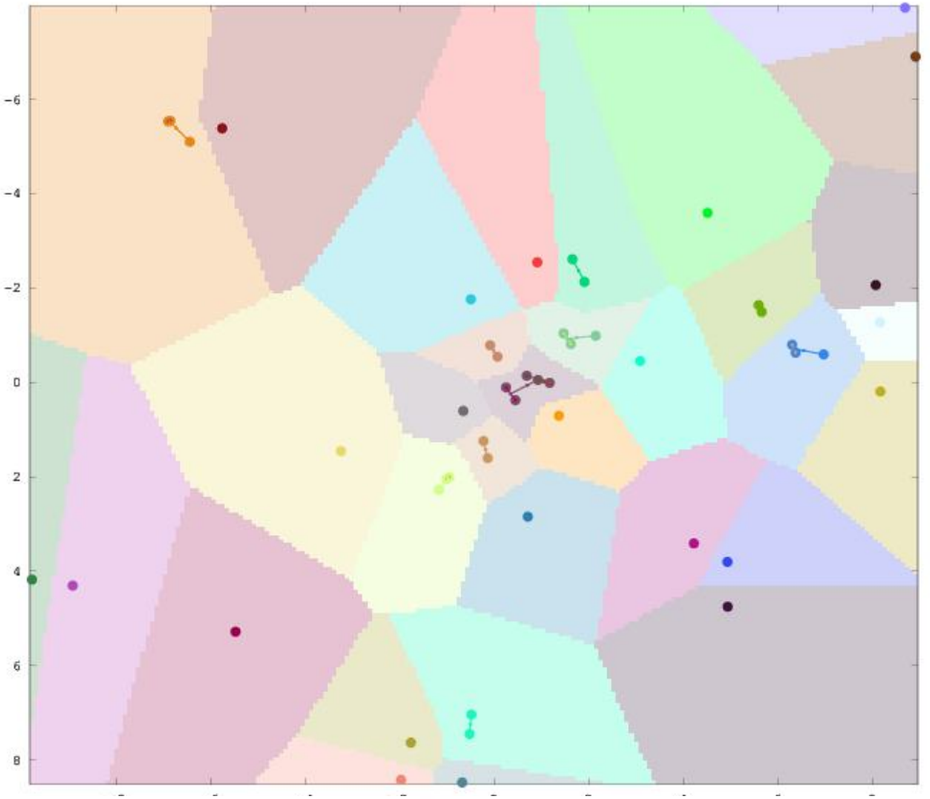


complete linkage



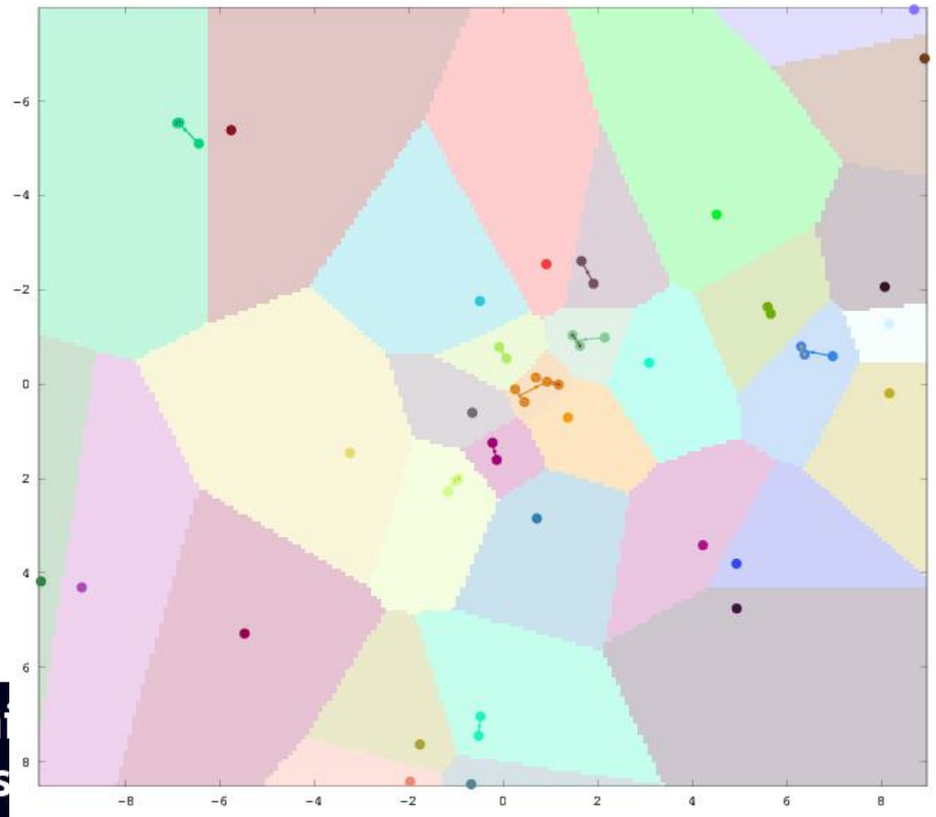
Example

33 clusters

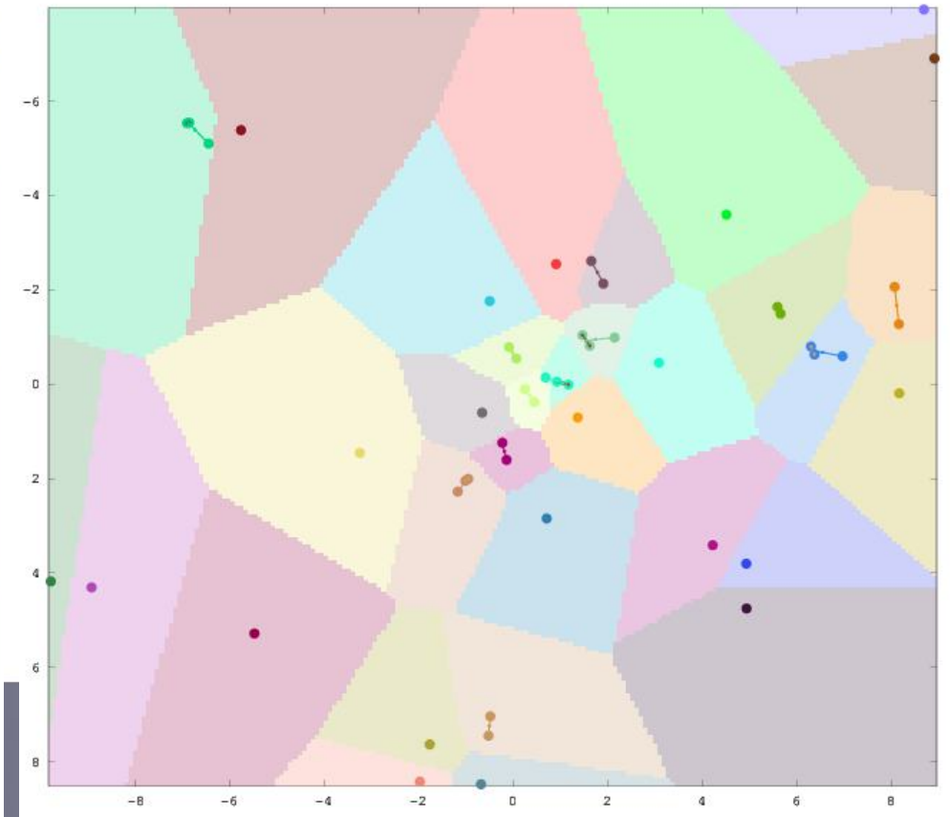


single-linkage

average linkage



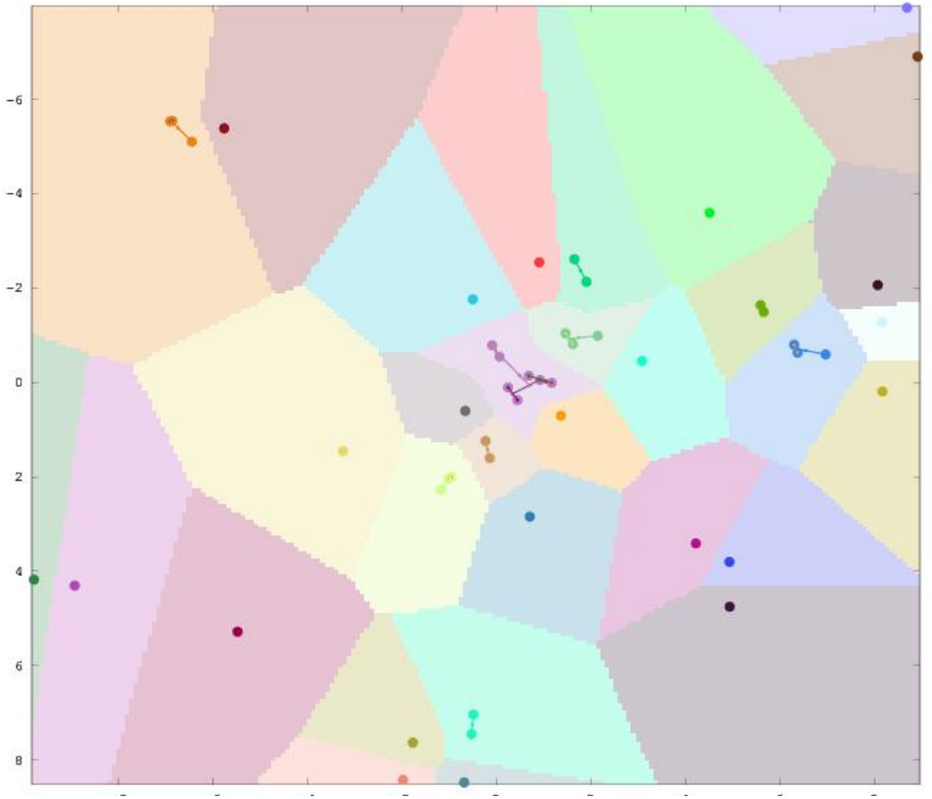
complete linkage



Example

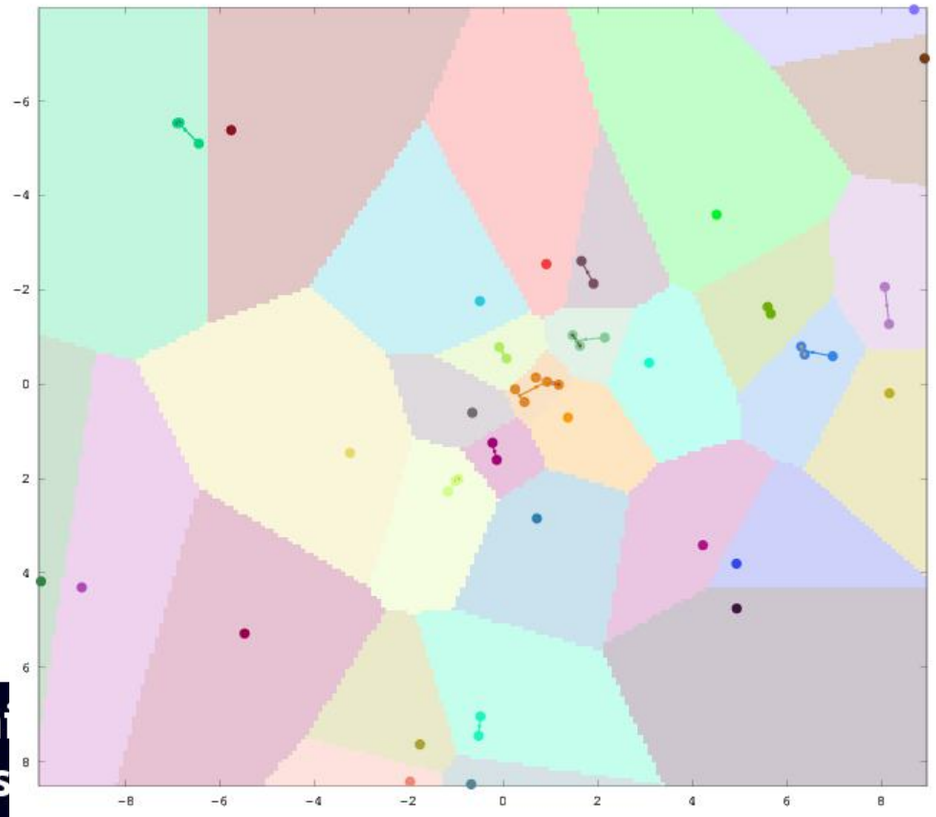
32 clusters

single-linkage has nonconvex Voronoi region near middle

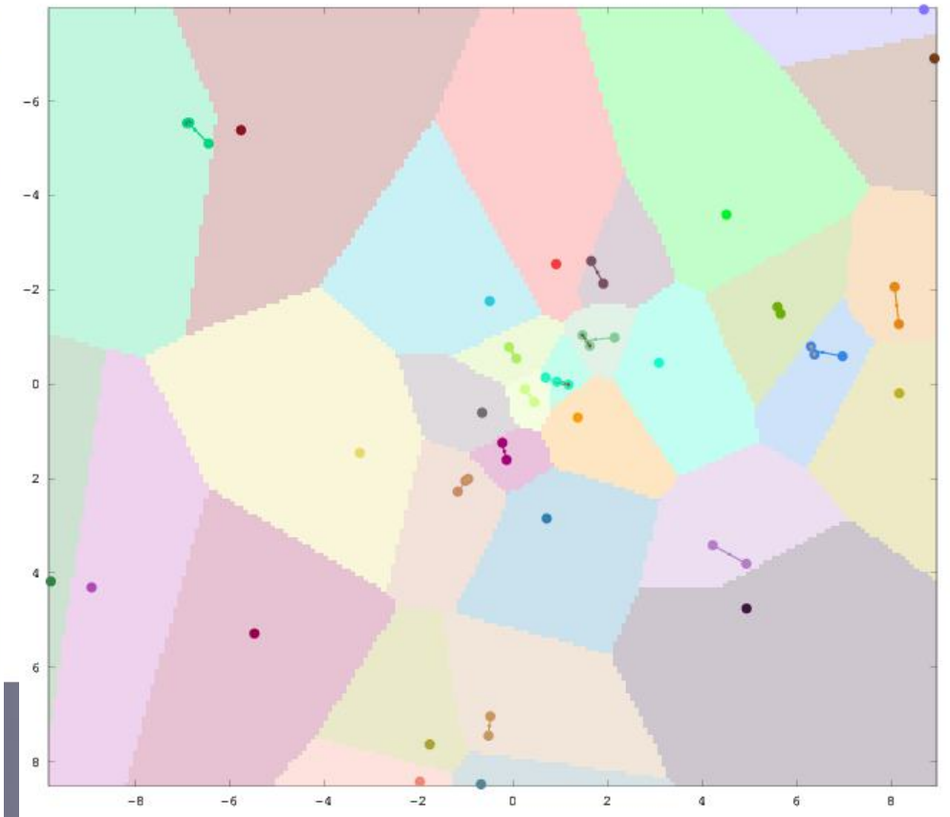


single-linkage

average linkage



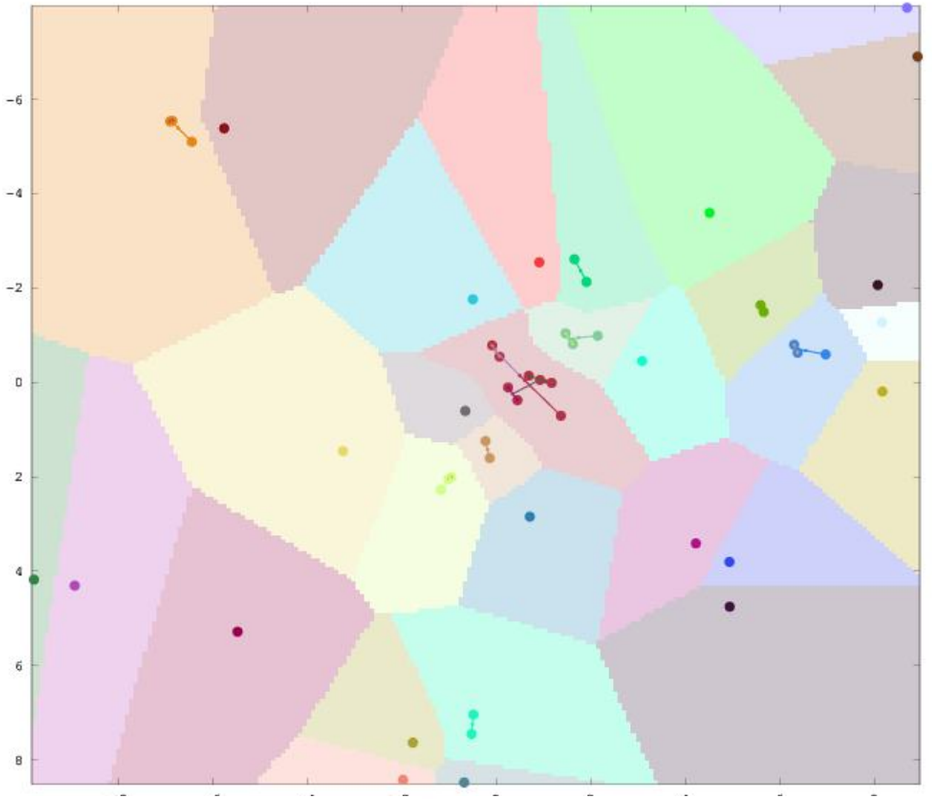
complete linkage



Example

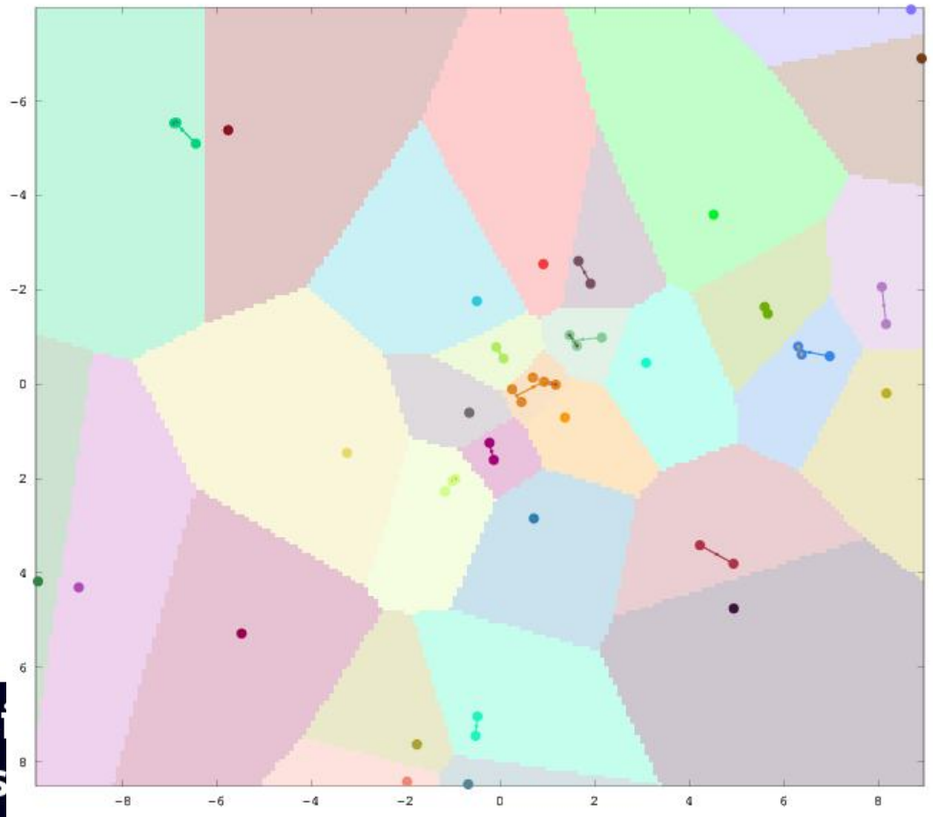
31 clusters

single-linkage has nonconvex Voronoi region near middle

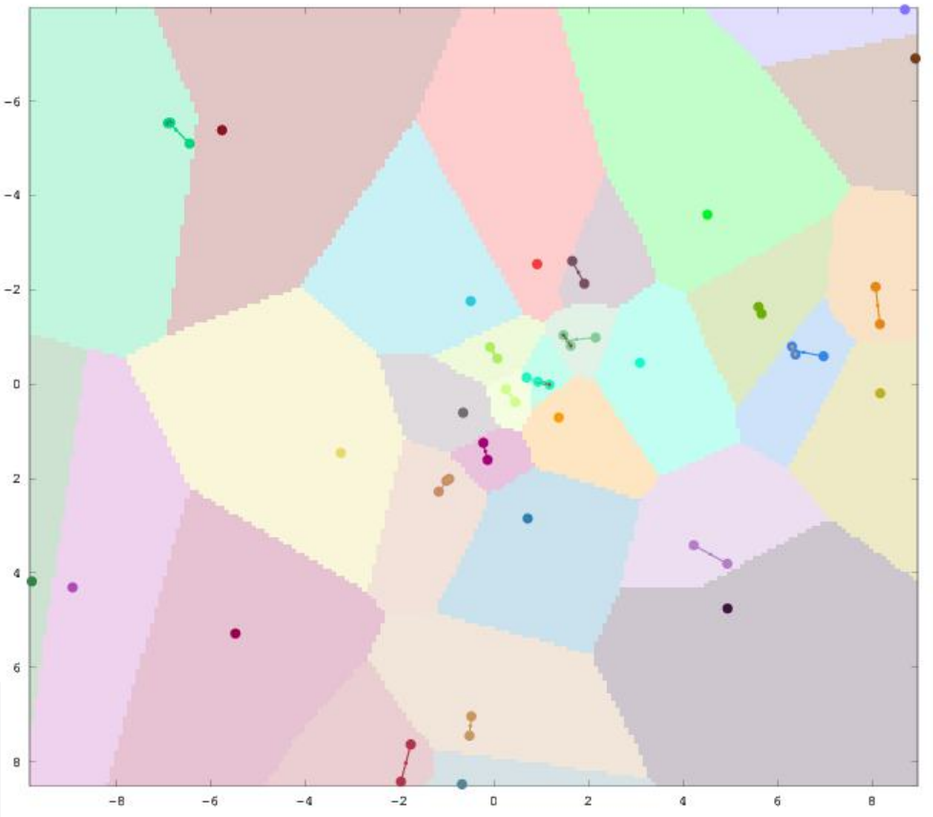


single-linkage

average linkage

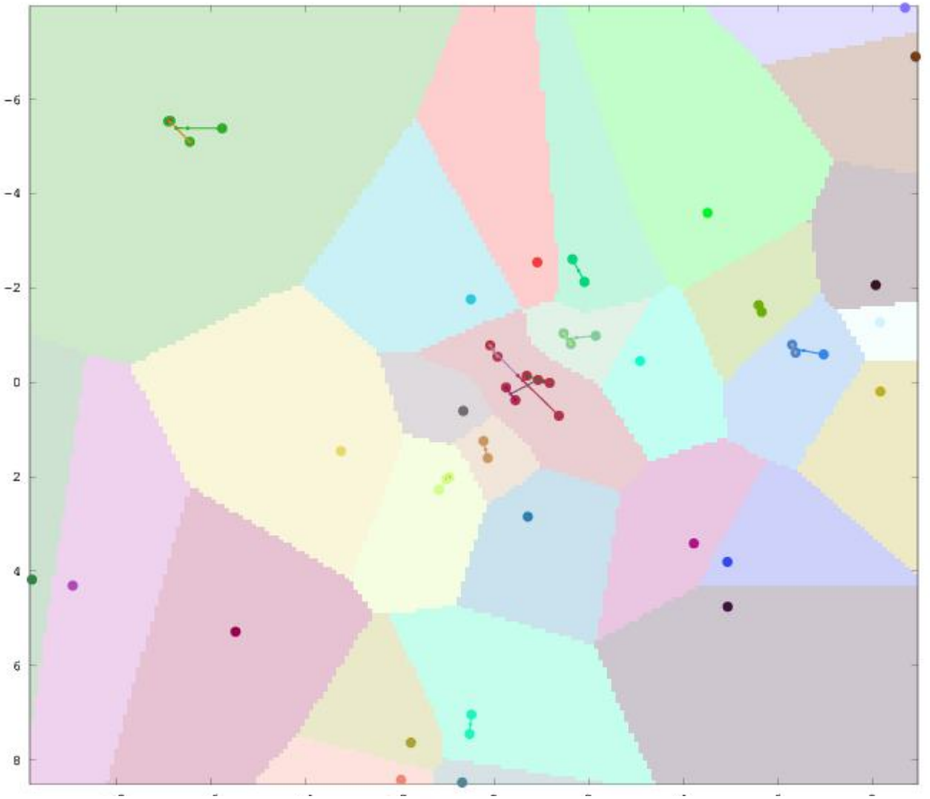


complete linkage



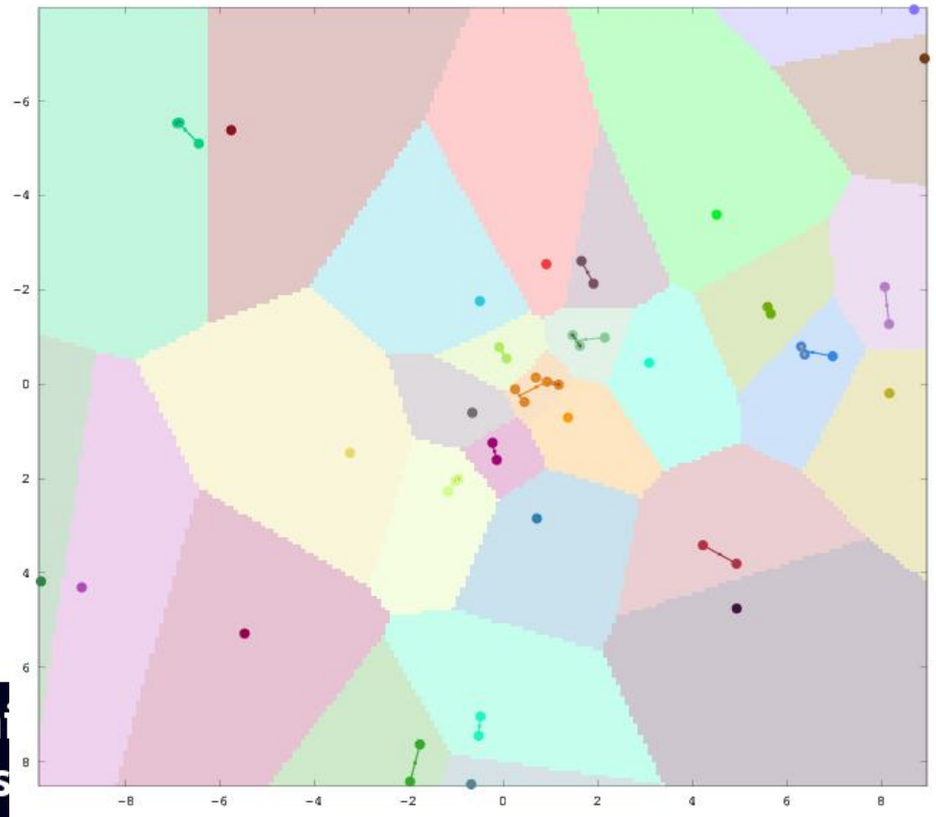
Example

30 clusters

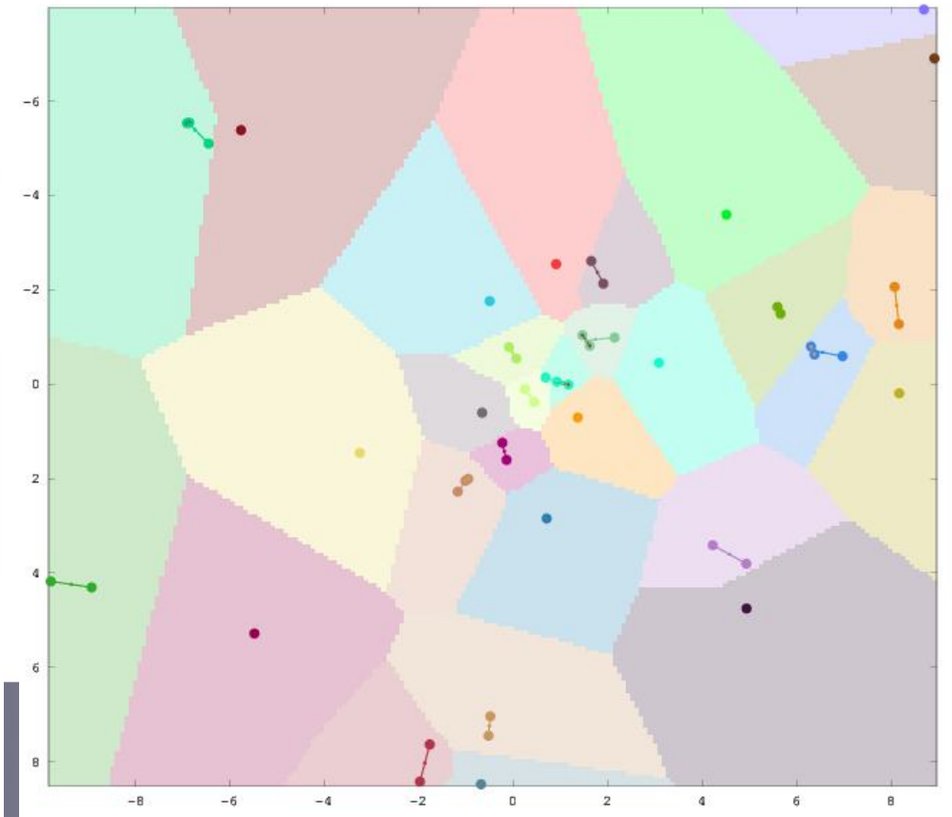


single-linkage

average linkage

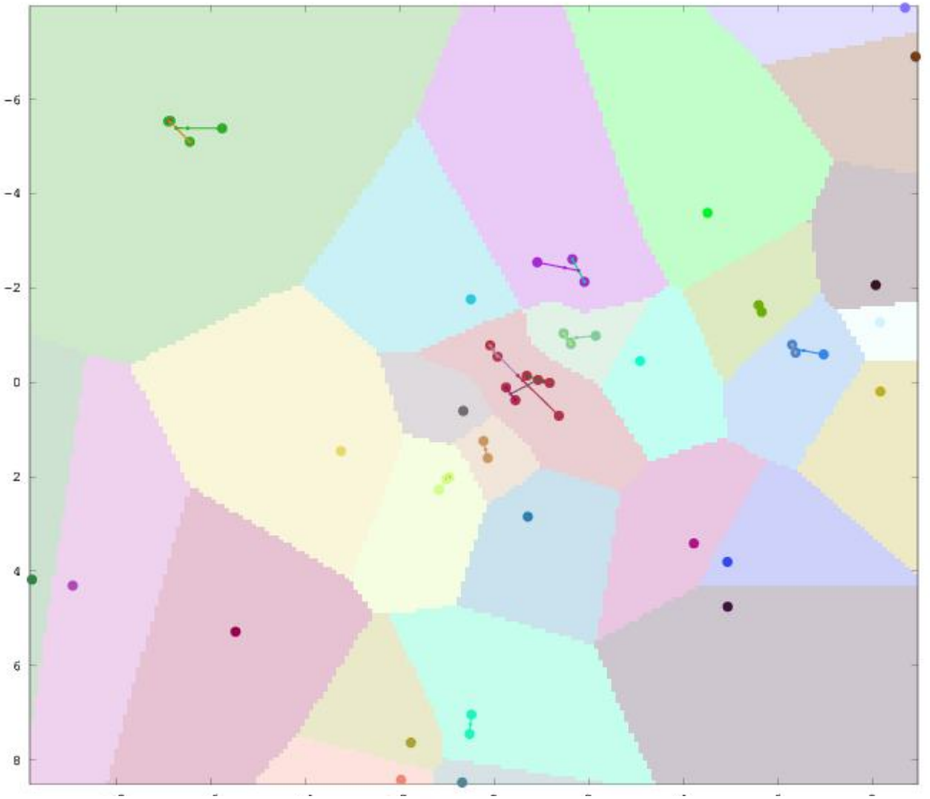


complete linkage



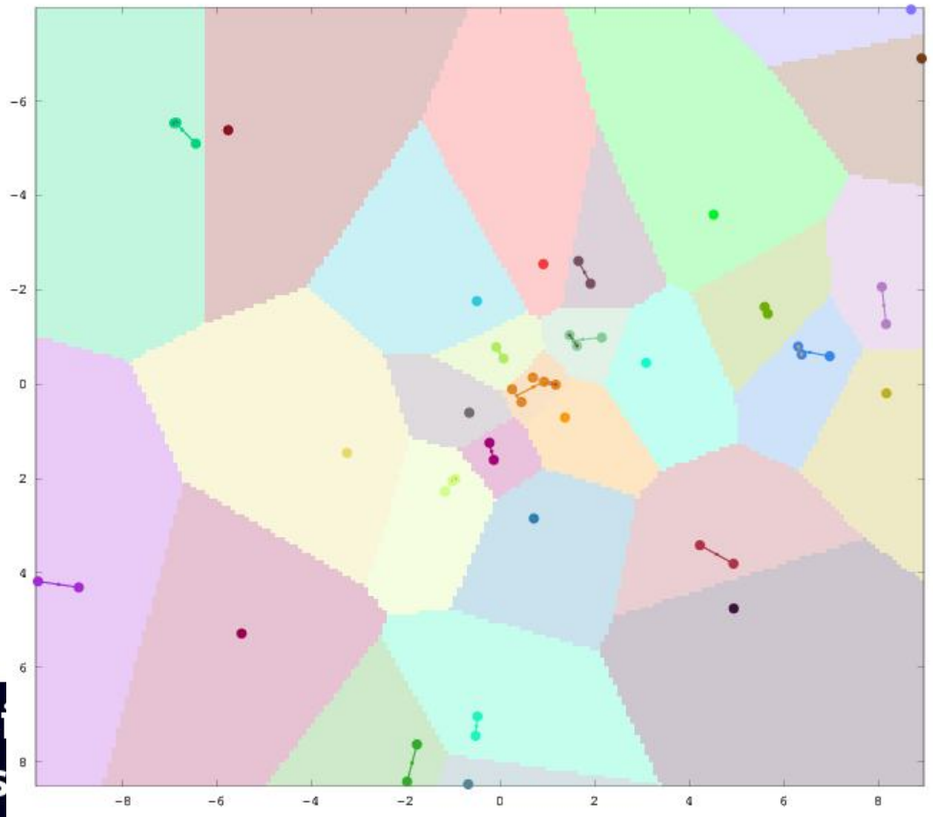
Example

29 clusters

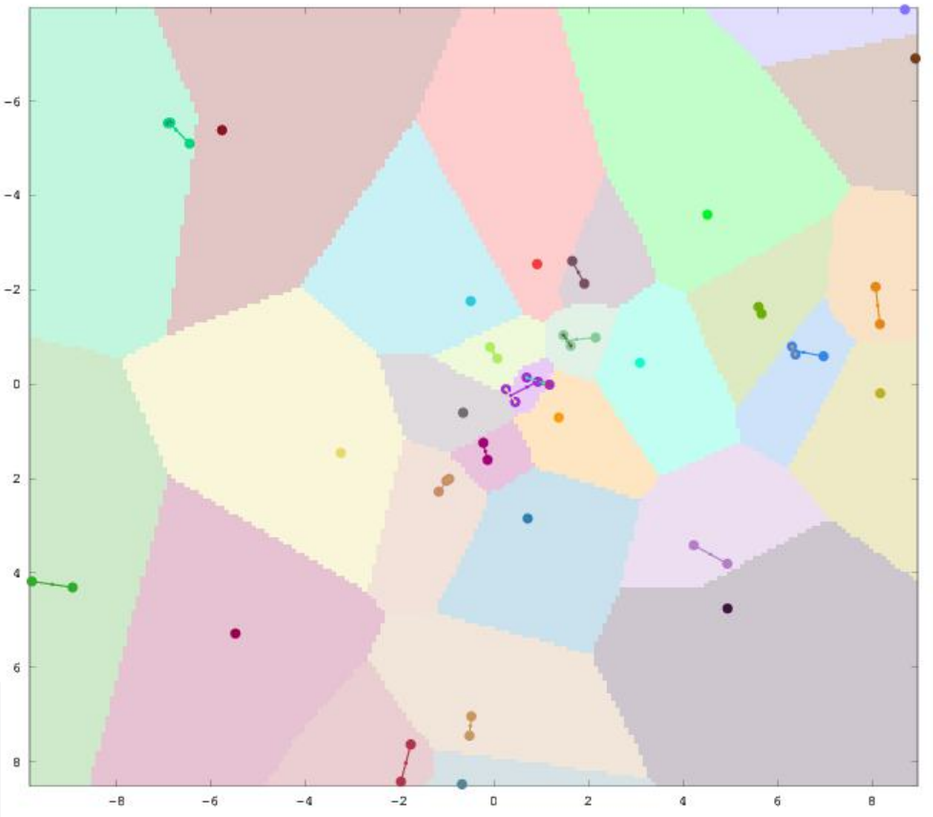


single-linkage

average linkage

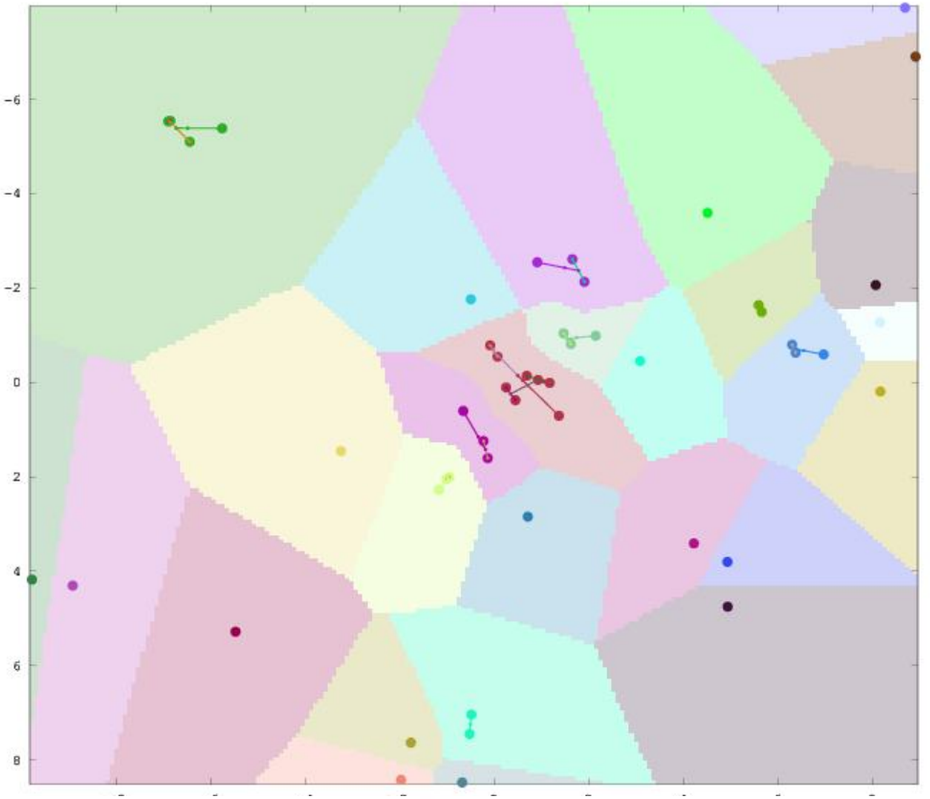


complete linkage



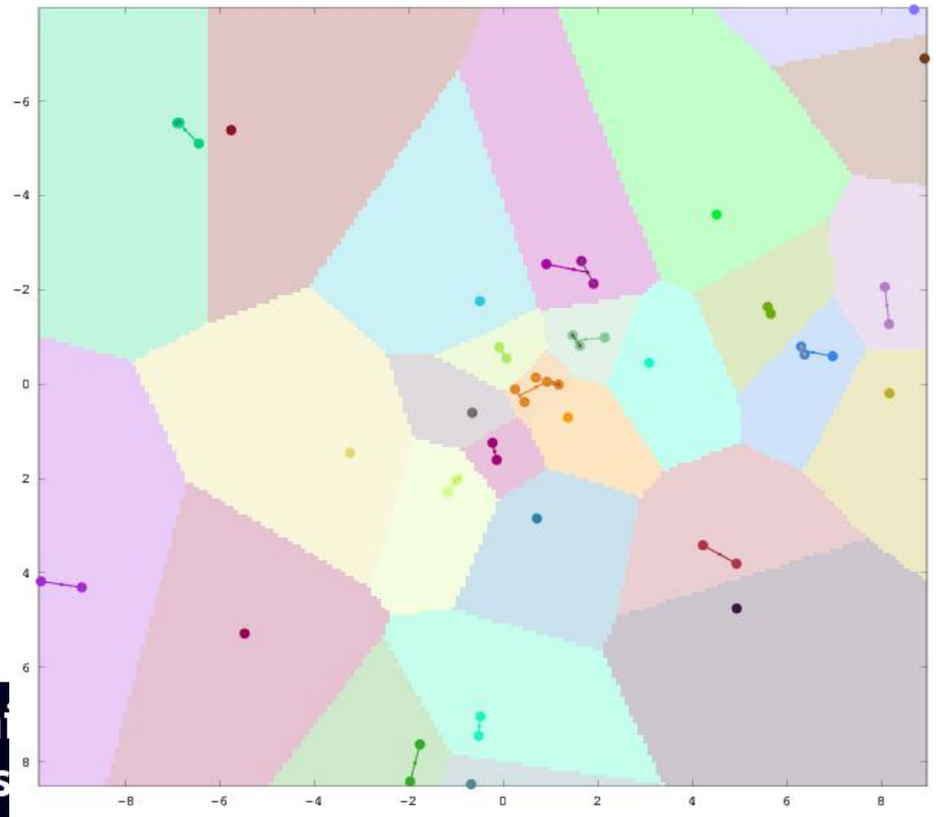
Example

28 clusters

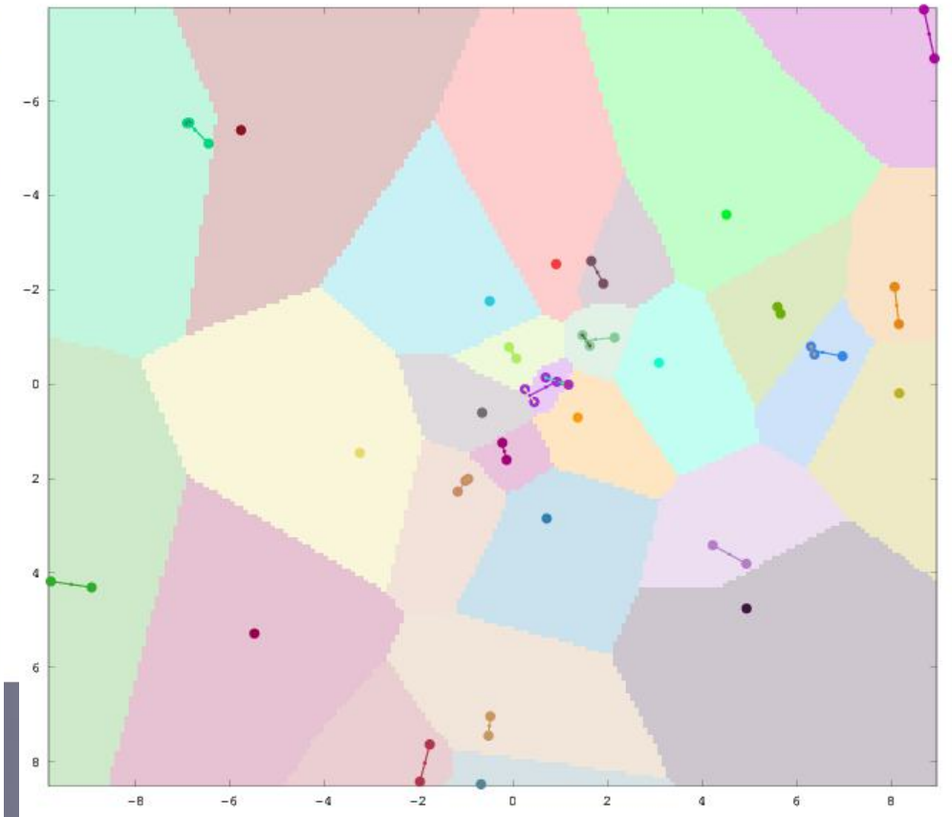


single-linkage

average linkage

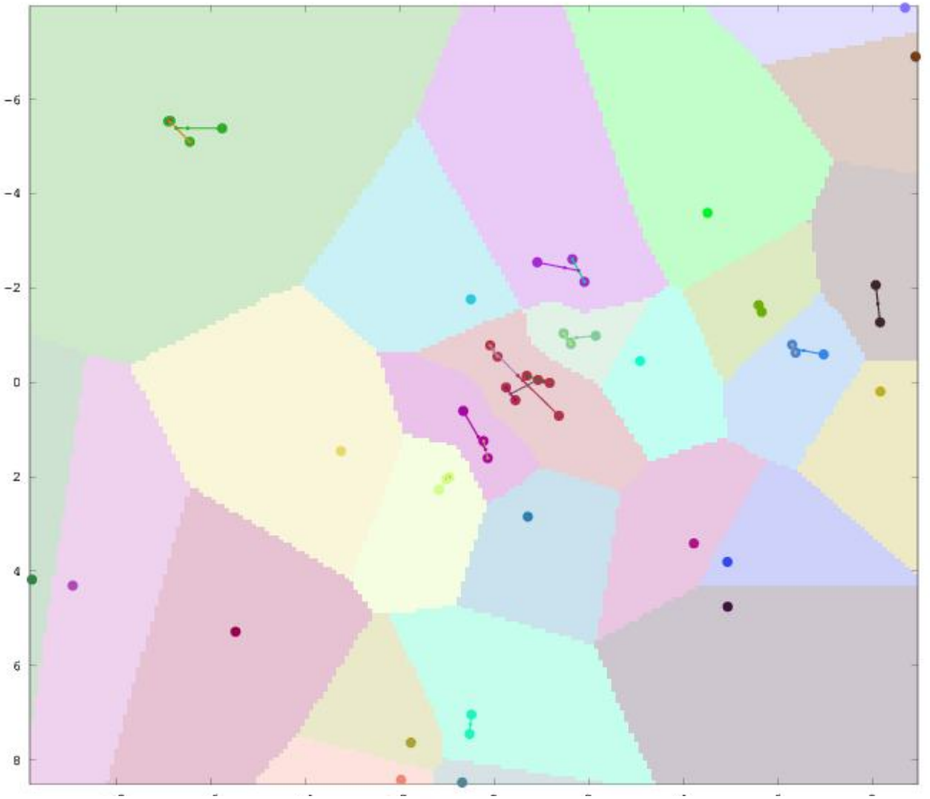


complete linkage



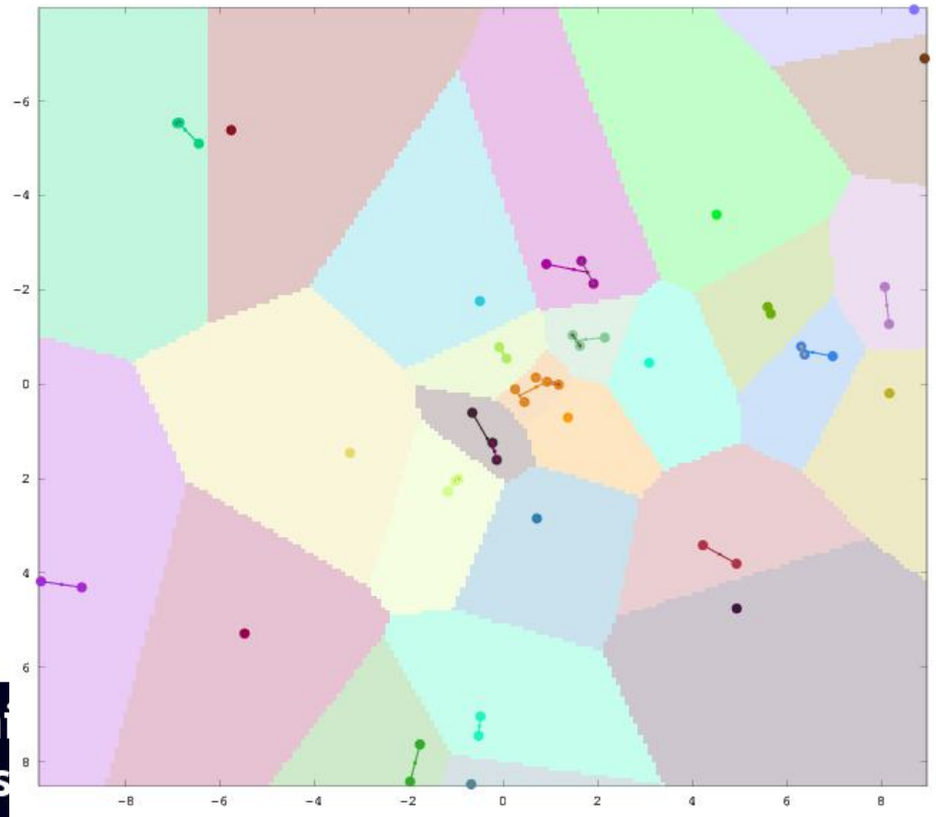
Example

27 clusters

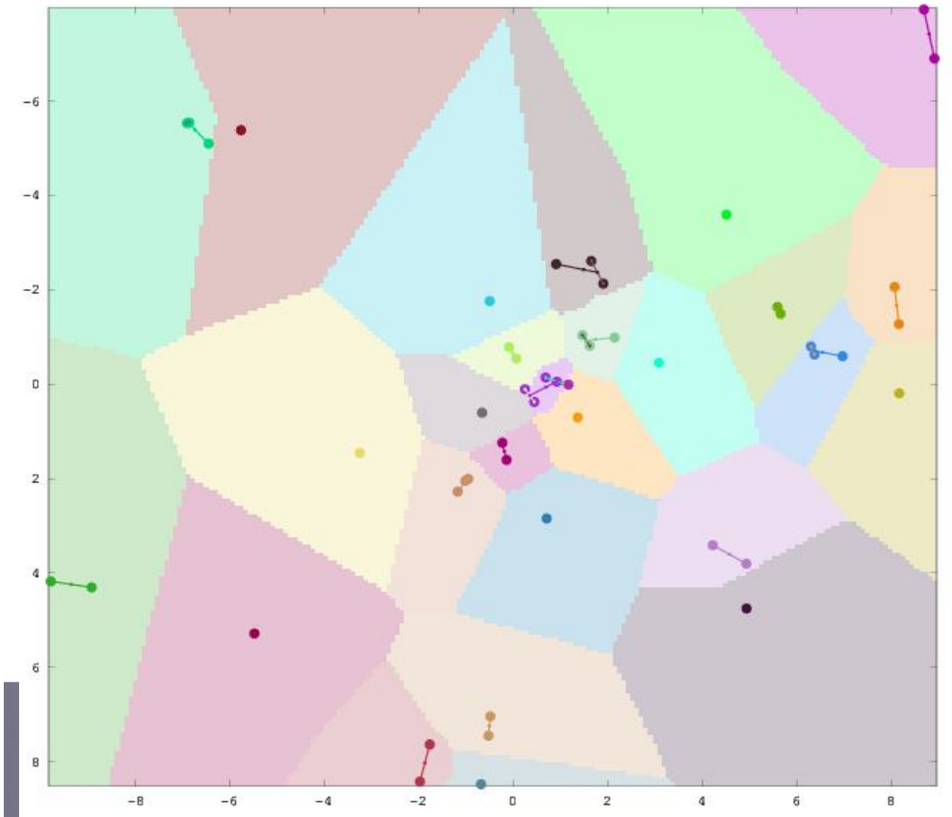


single-linkage

average linkage

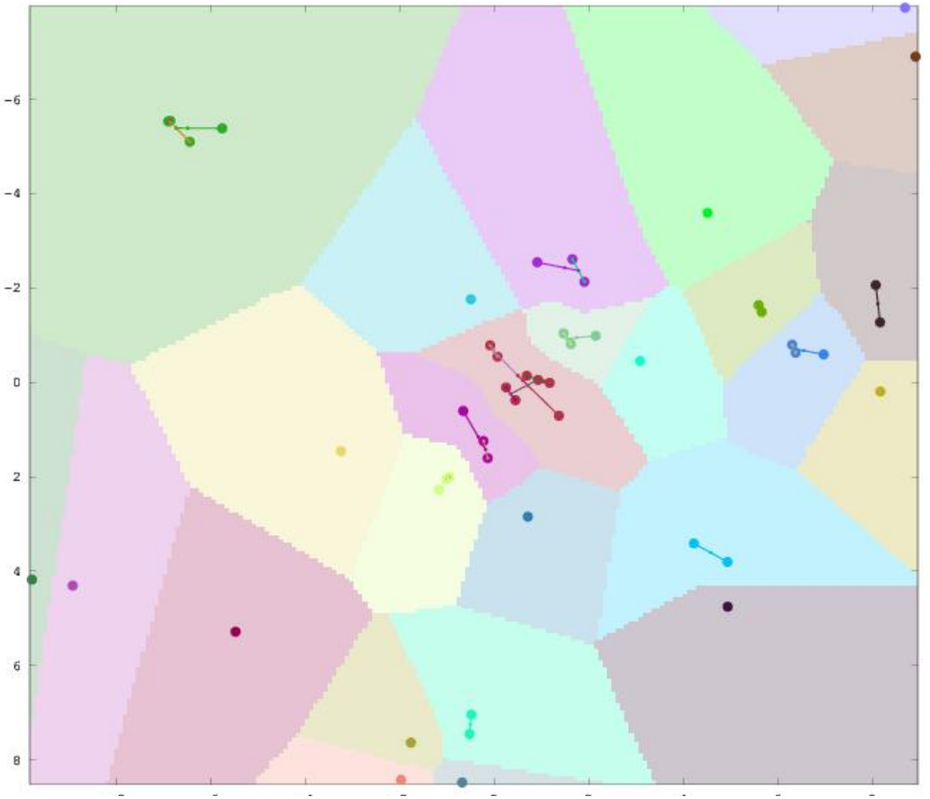


complete linkage



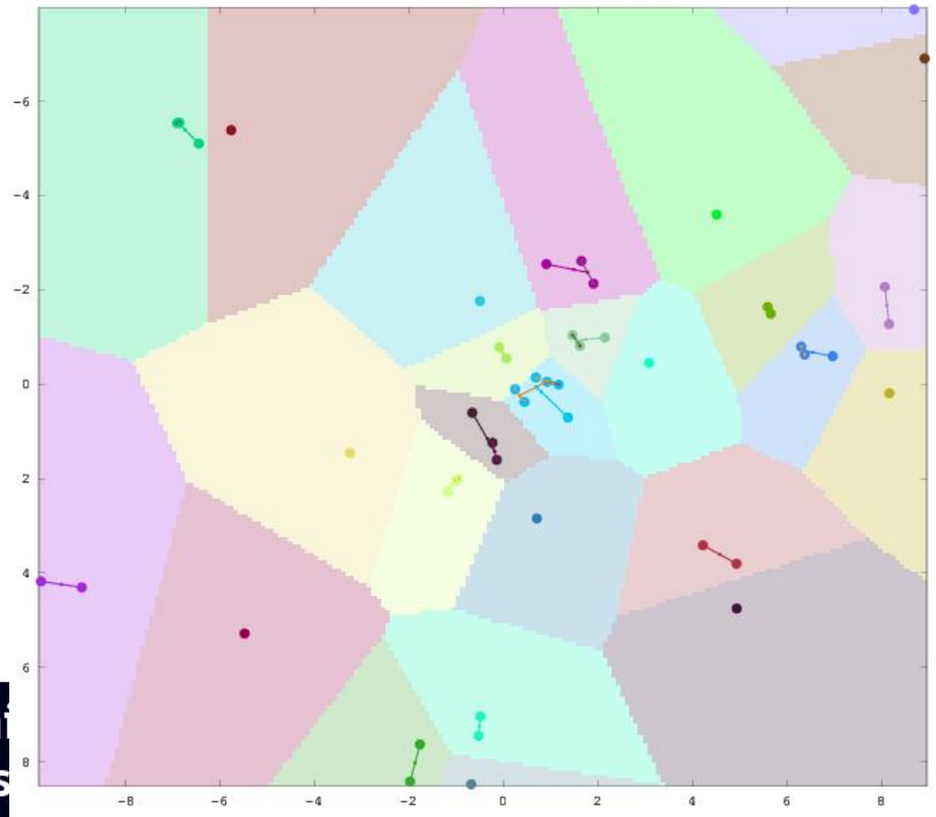
Example

26 clusters

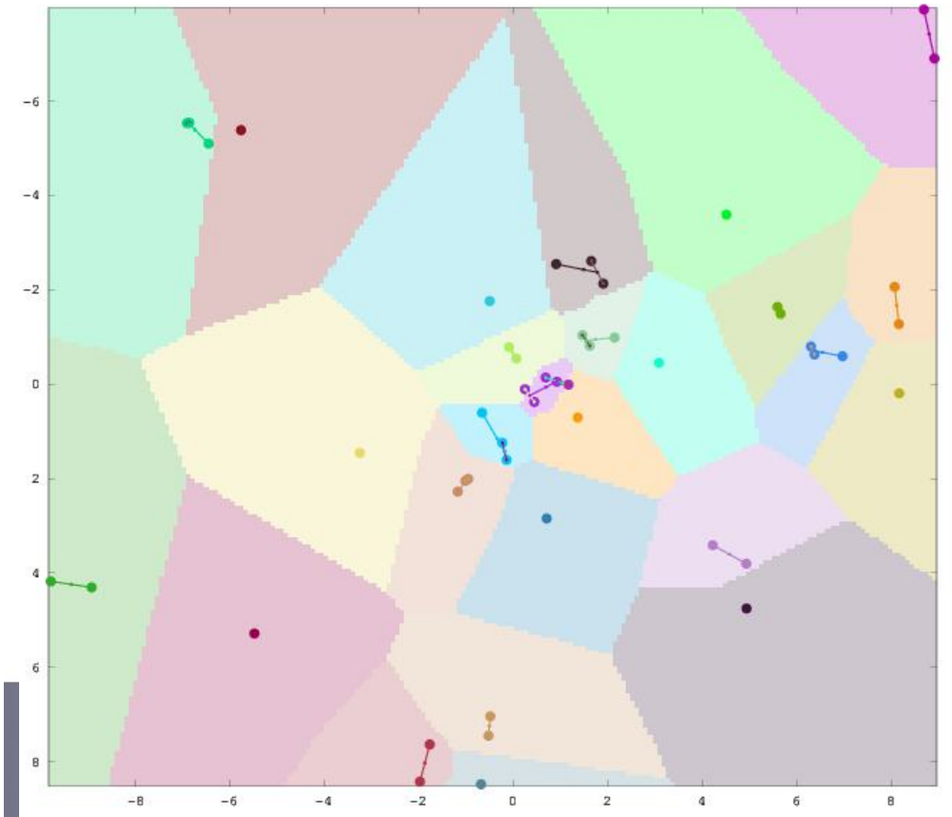


single-linkage

average linkage

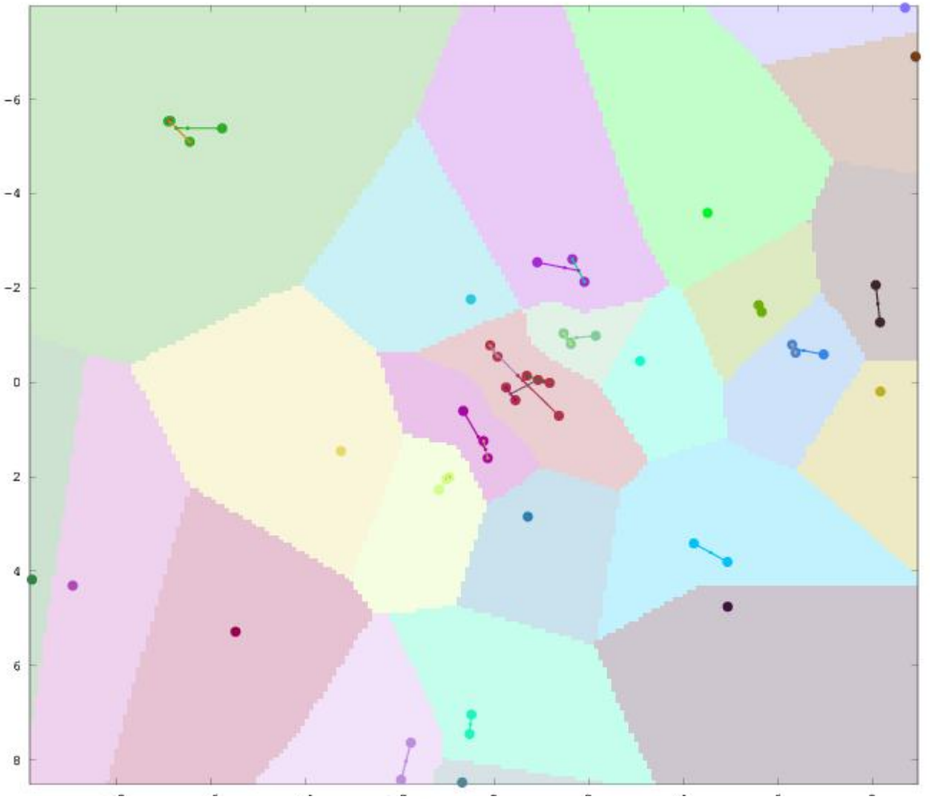


complete linkage



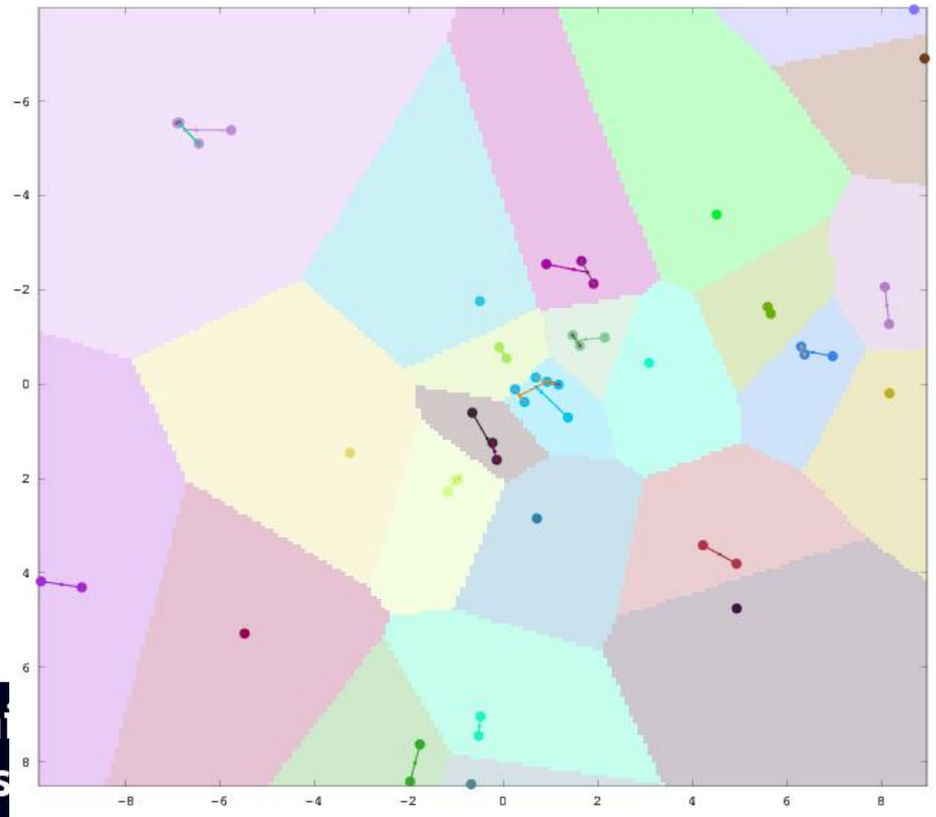
Example

25 clusters

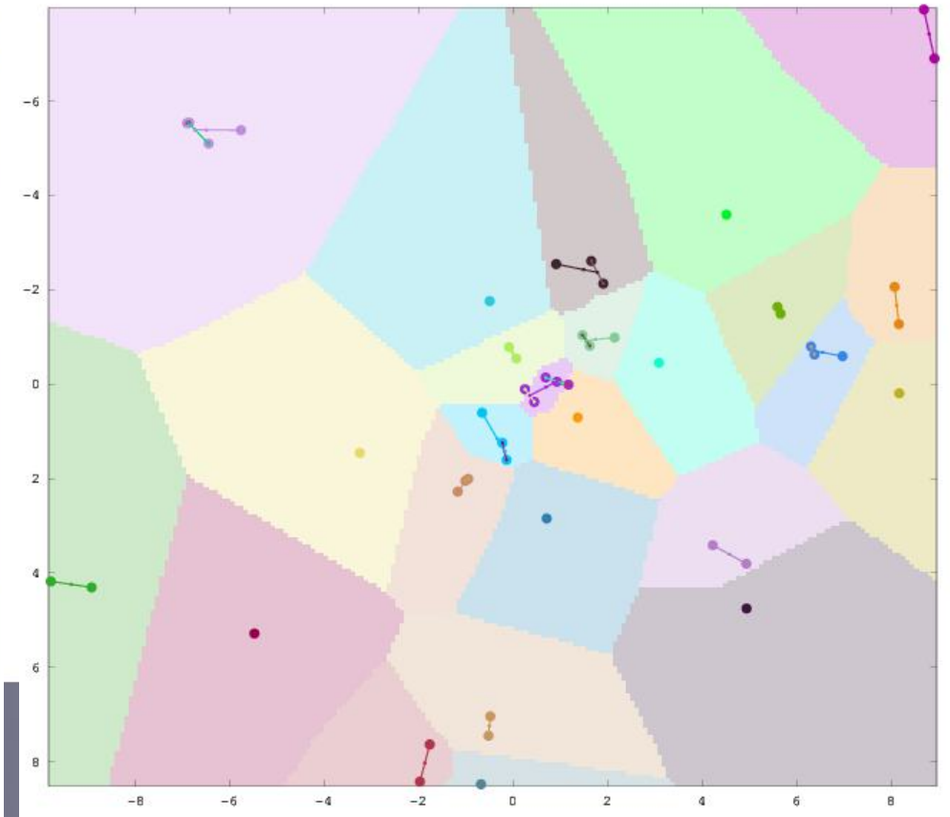


single-linkage

average linkage

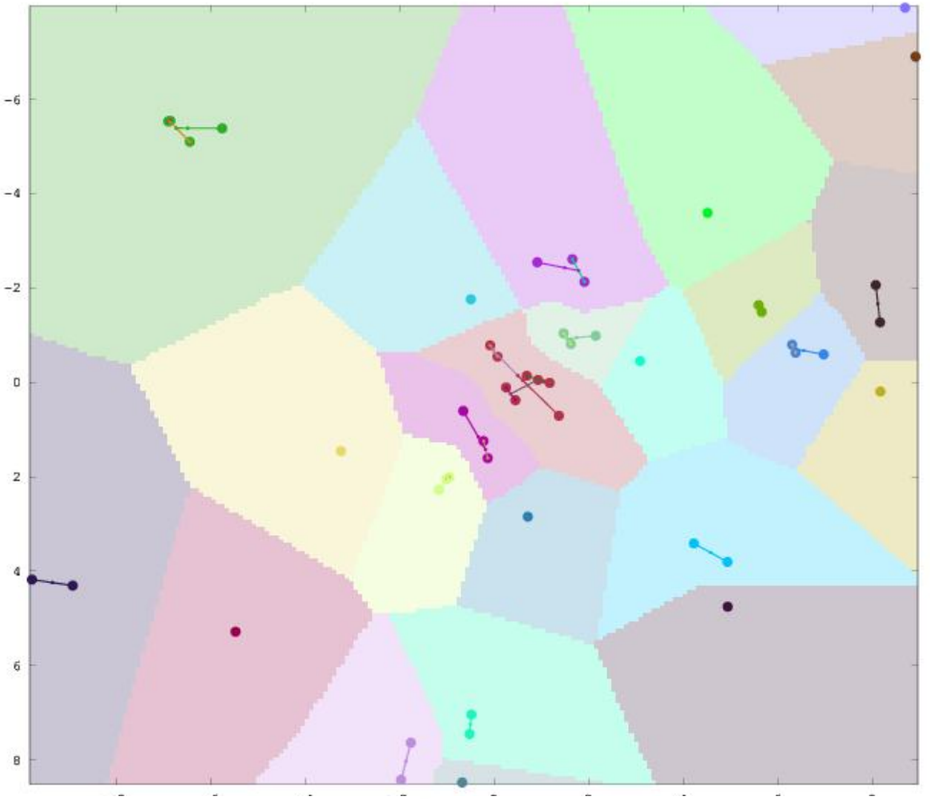


complete linkage



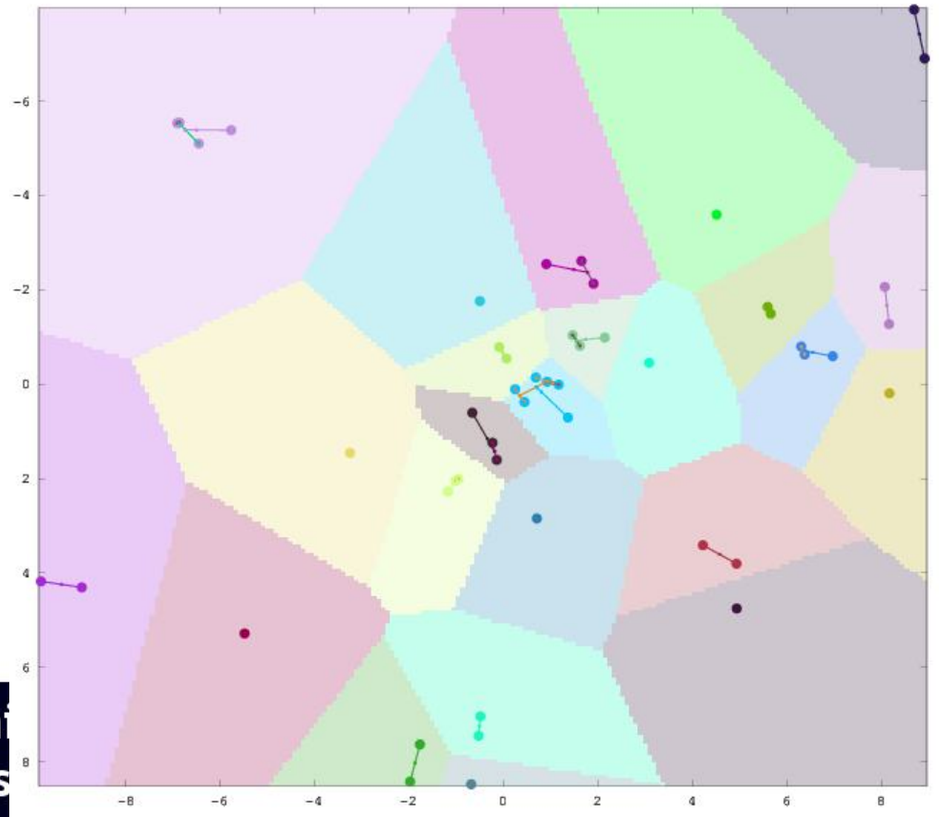
Example

24 clusters

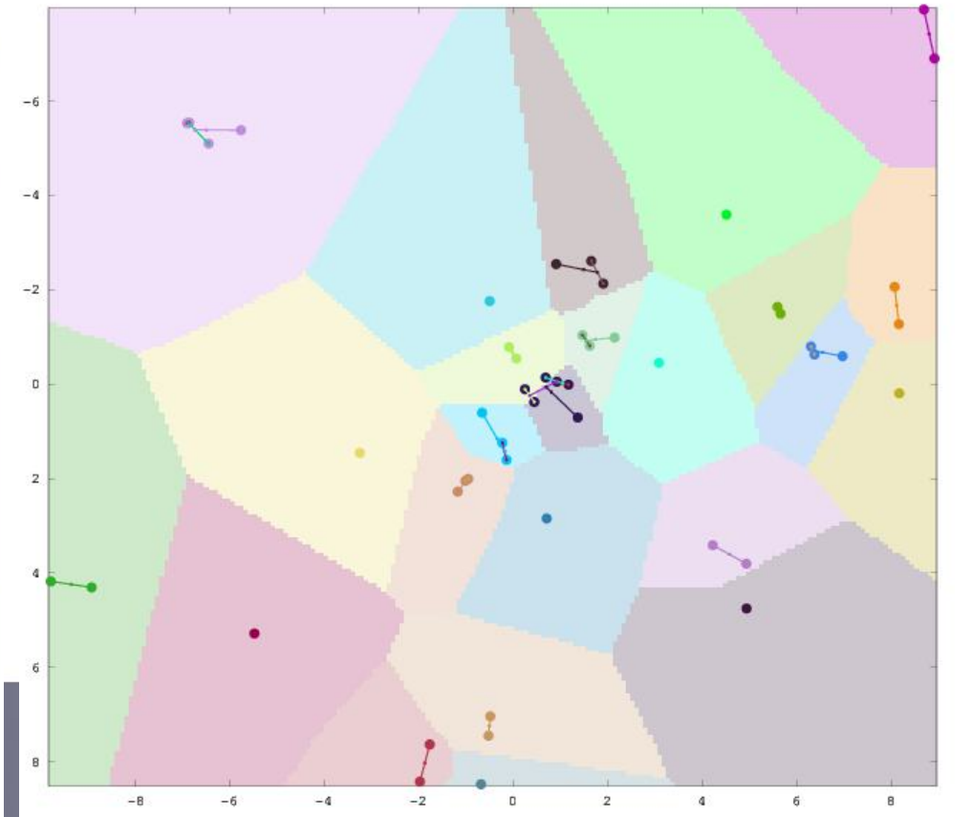


single-linkage

average linkage



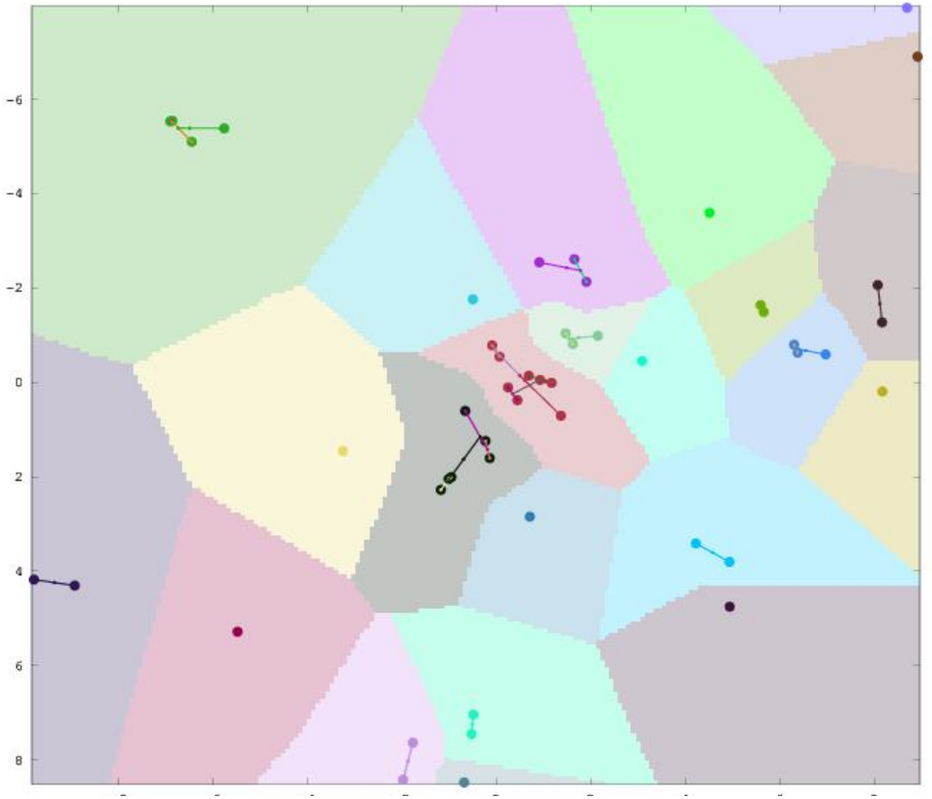
complete linkage



Example

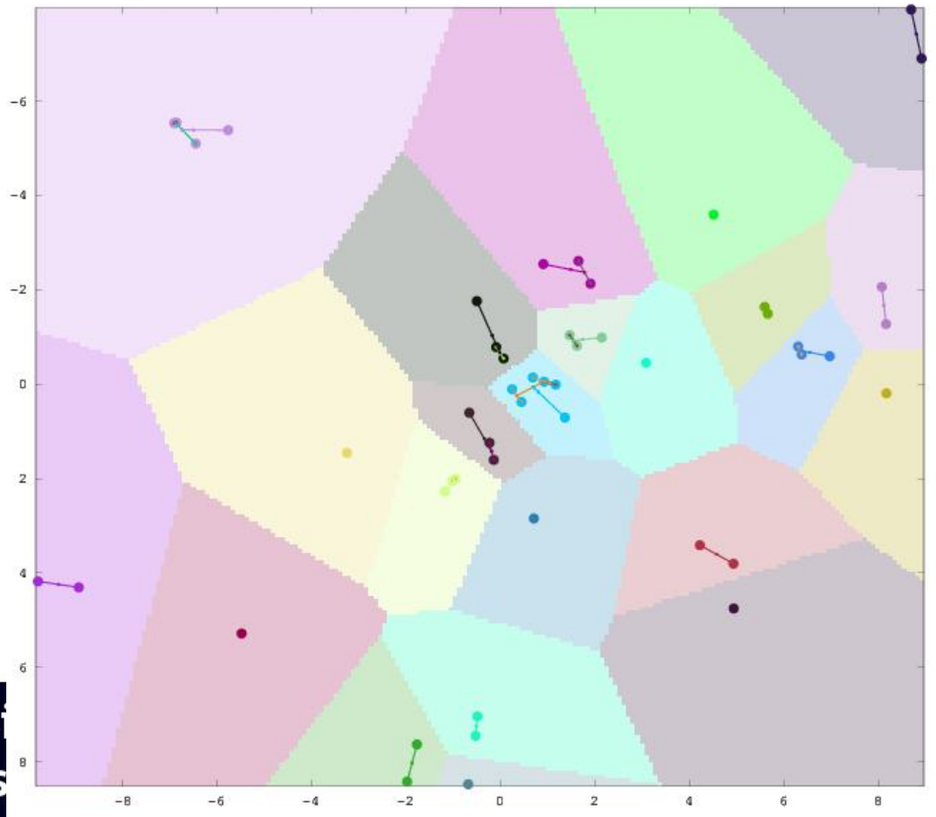
23 clusters

subtrees within clusters start to be visible

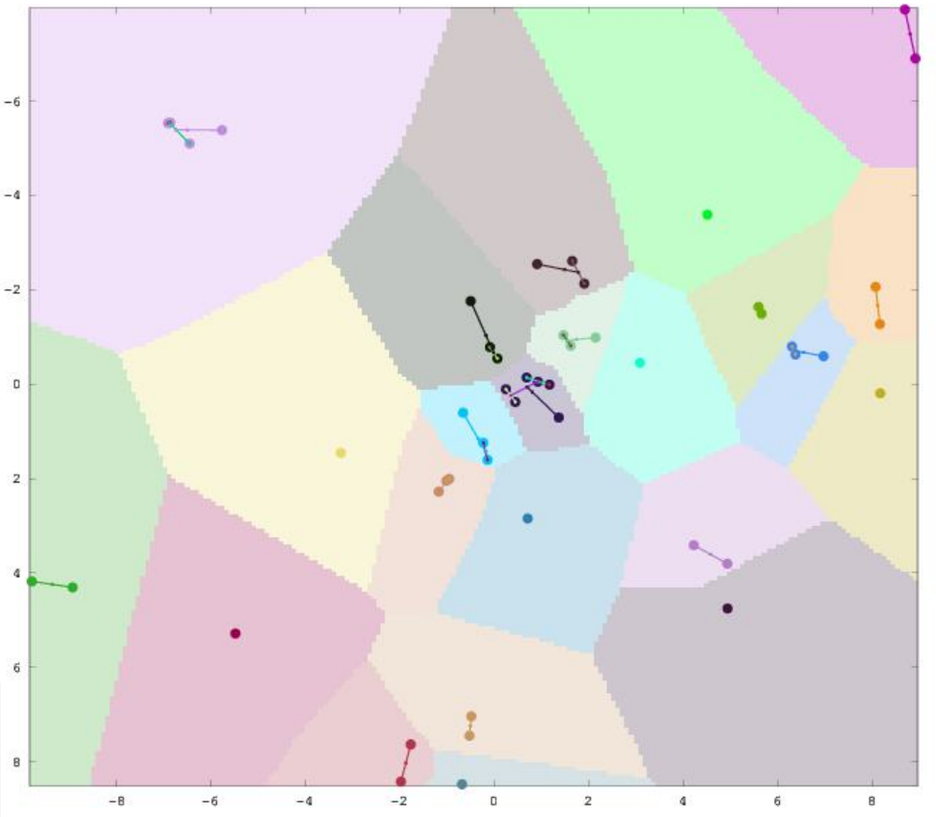


single-linkage

average linkage

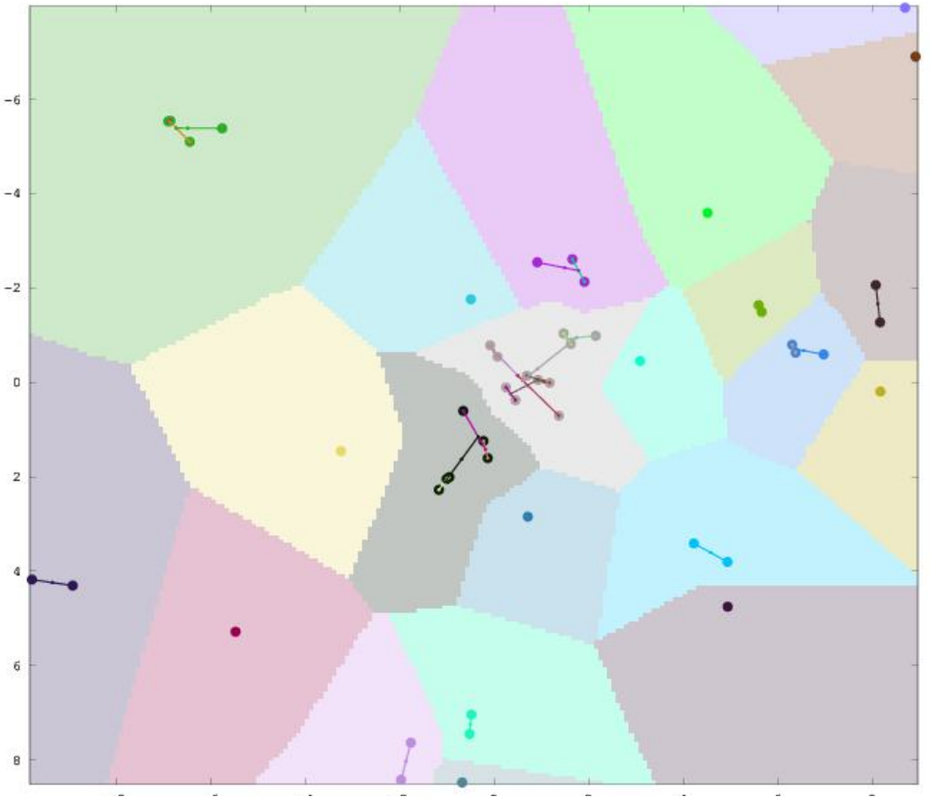


complete linkage



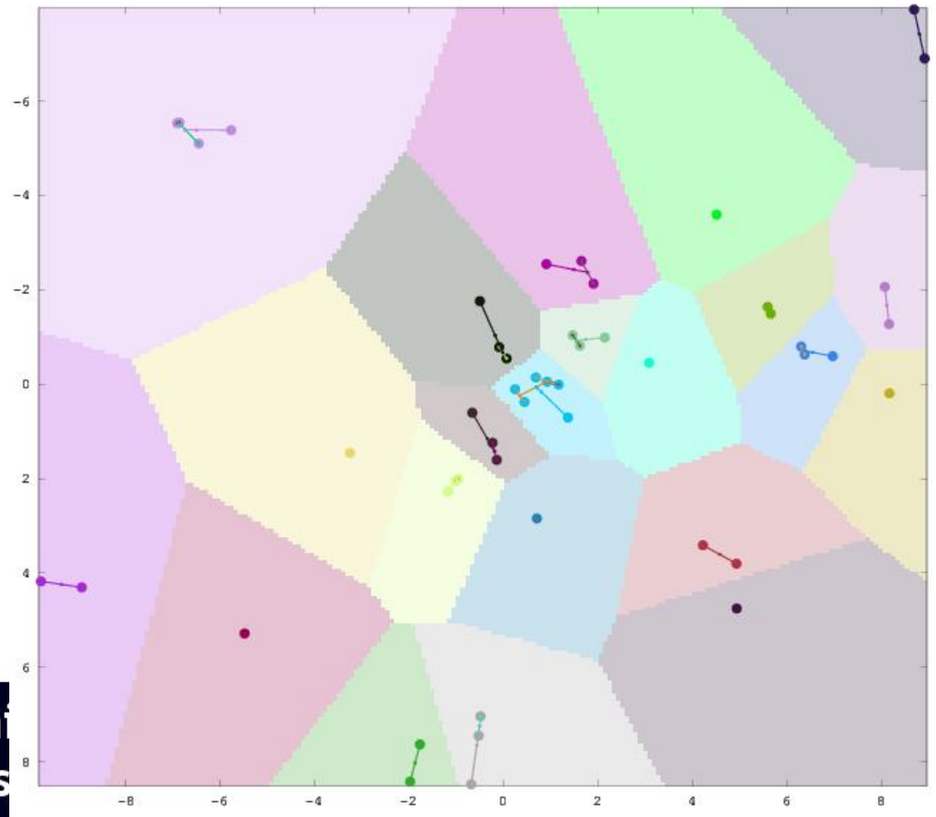
Example

22 clusters

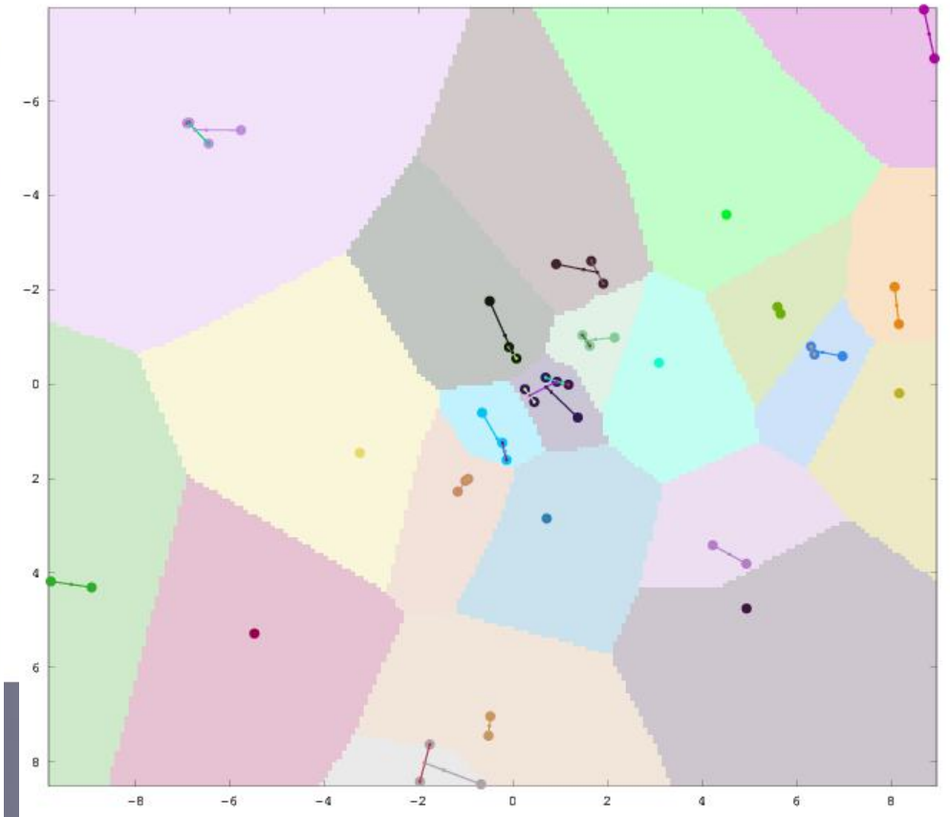


single-linkage

average linkage

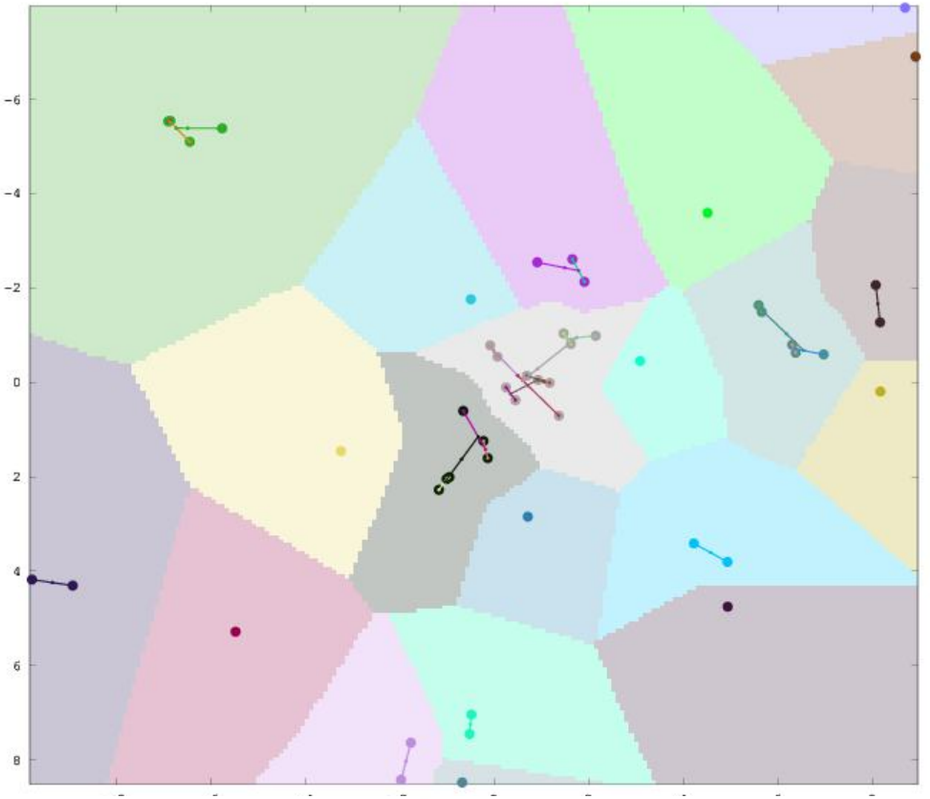


complete linkage



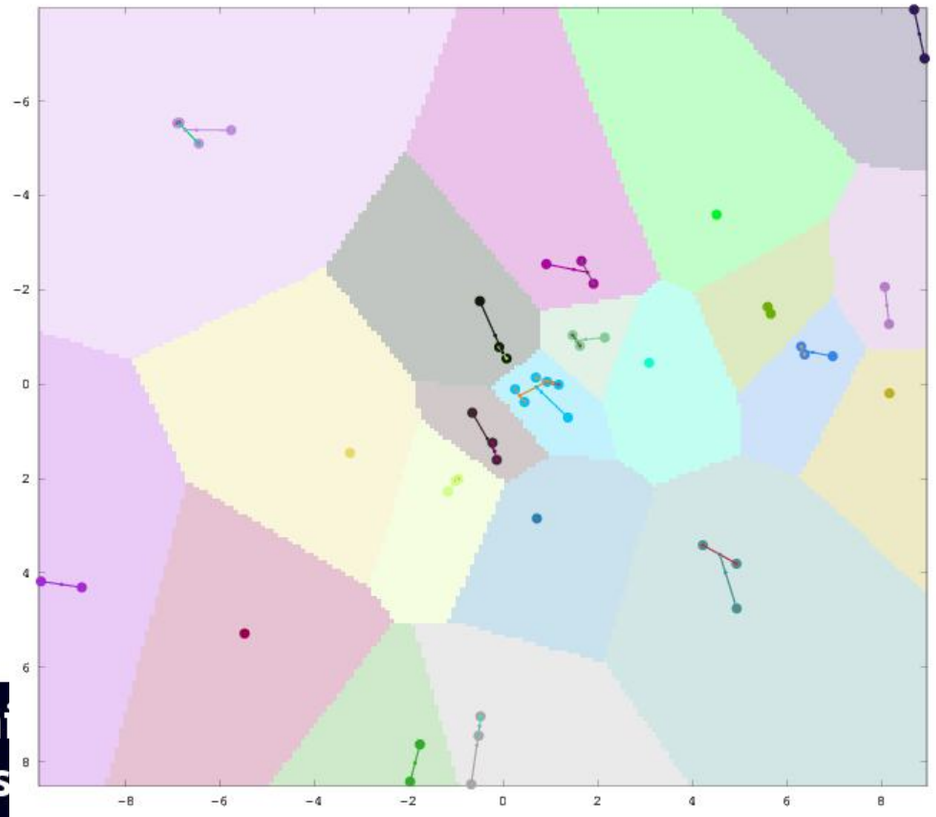
Example

21 clusters

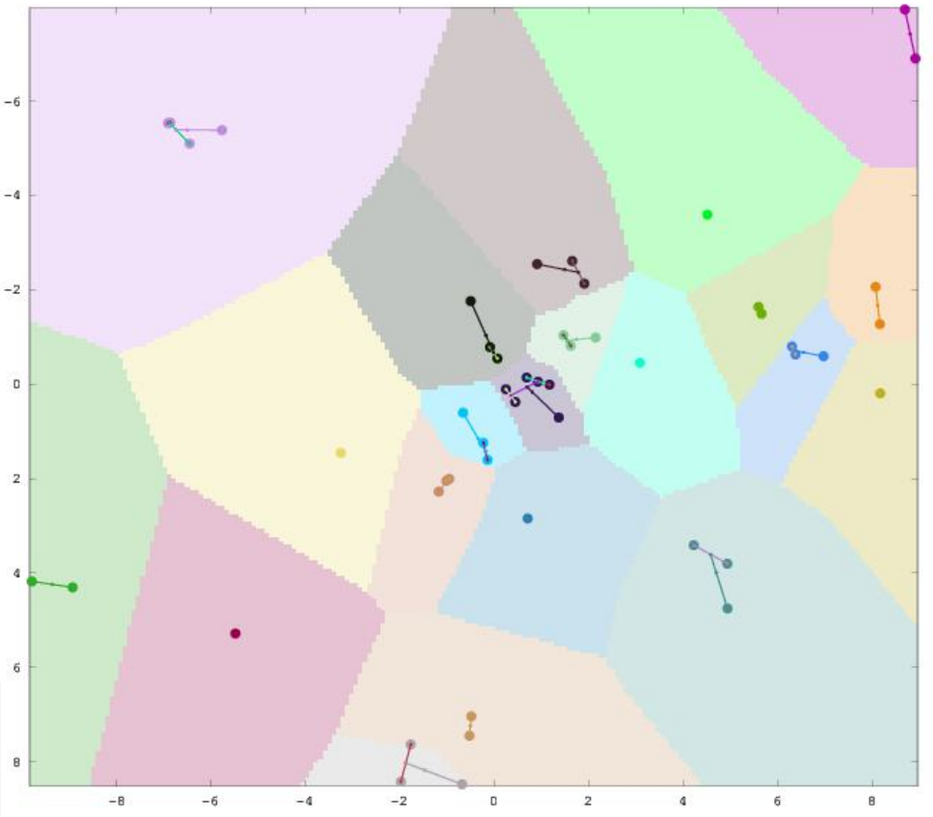


single-linkage

average linkage

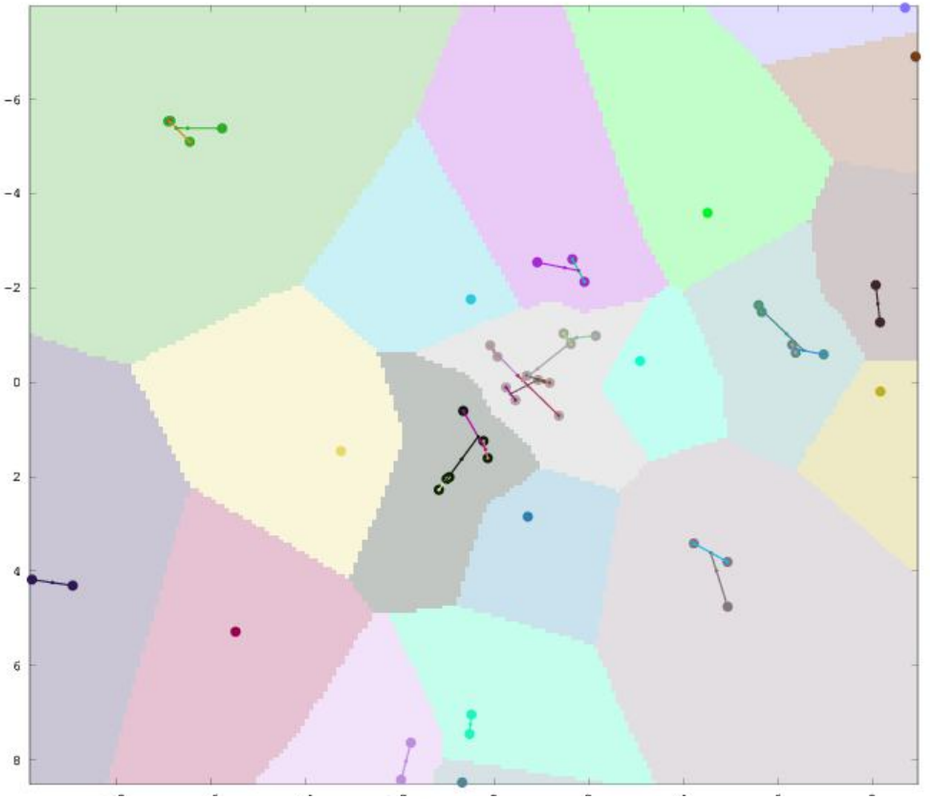


complete linkage



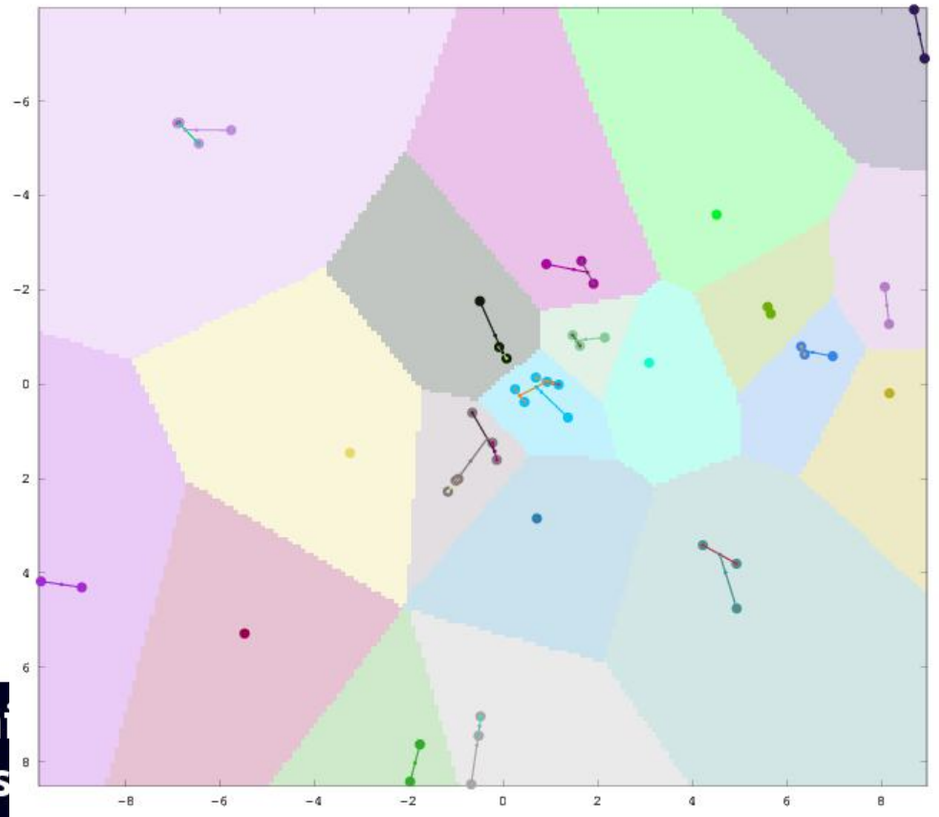
Example

20 clusters

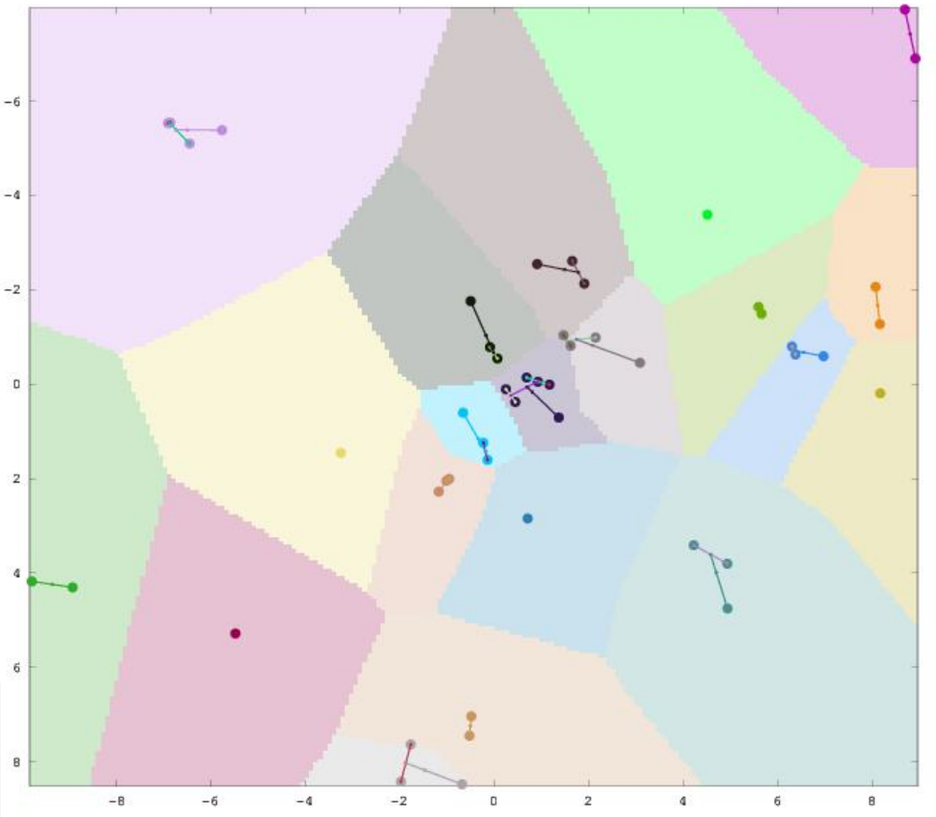


single-linkage

average linkage



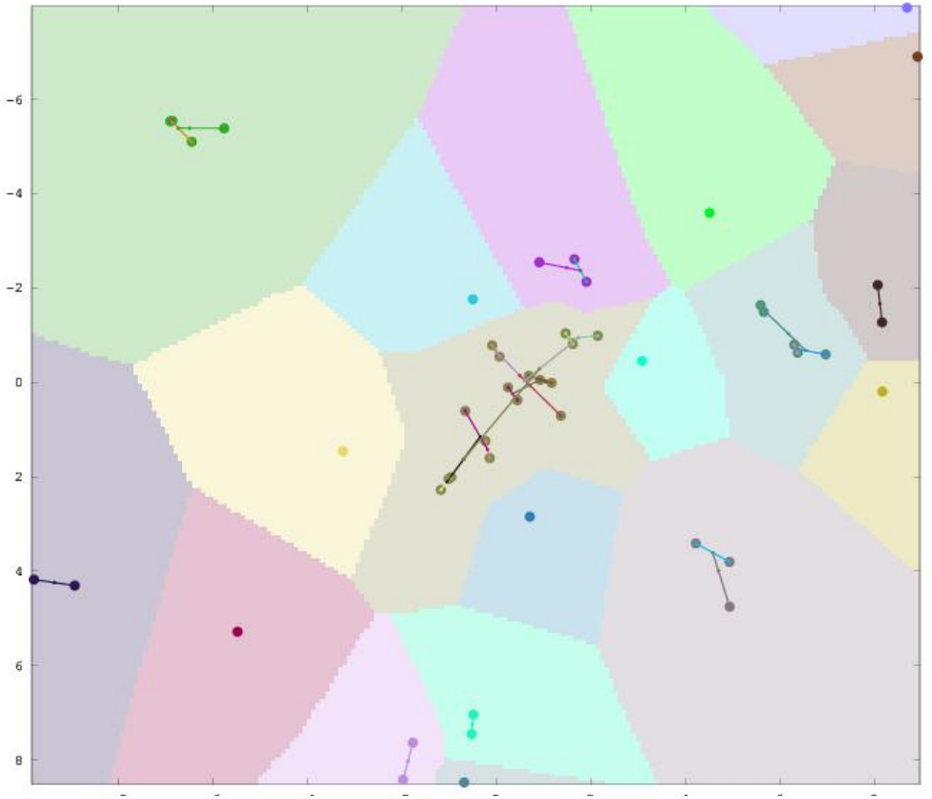
complete linkage



Example

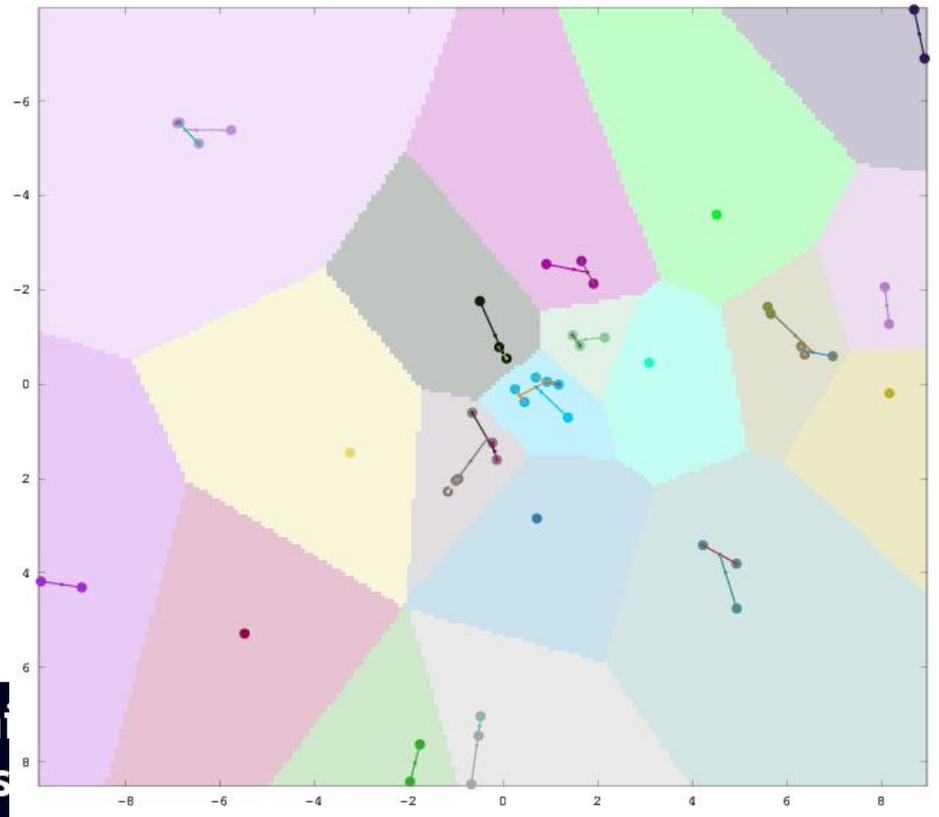
19 clusters

single-linkage again has clearly nonconvex Voronoi region near middle

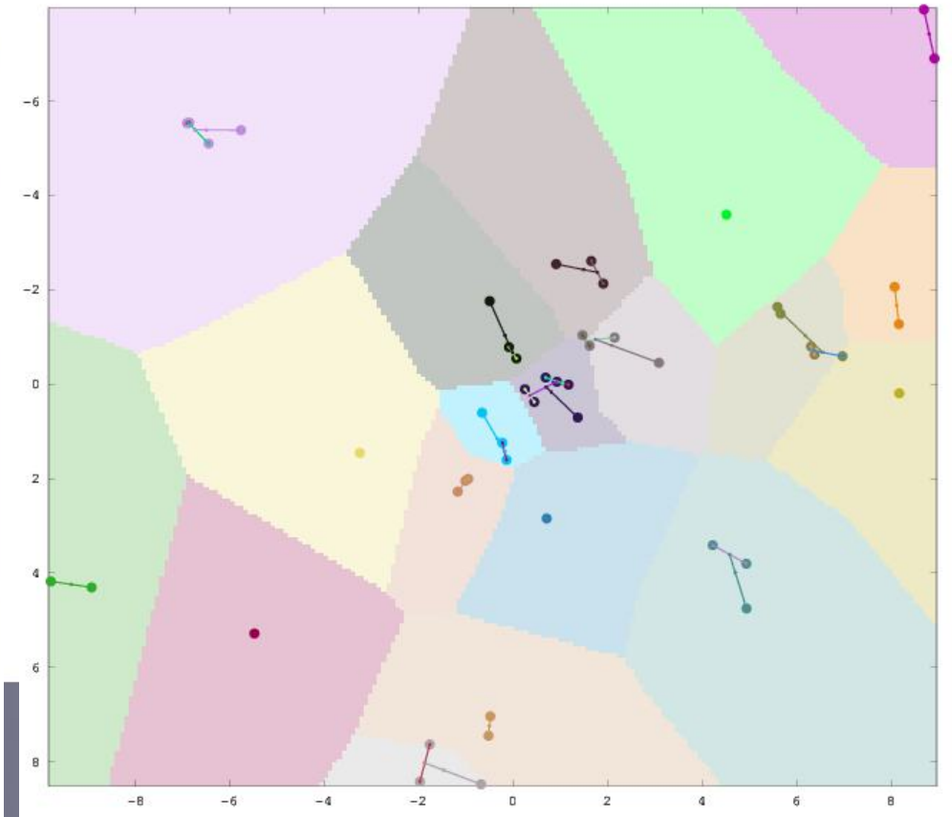


single-linkage

average linkage

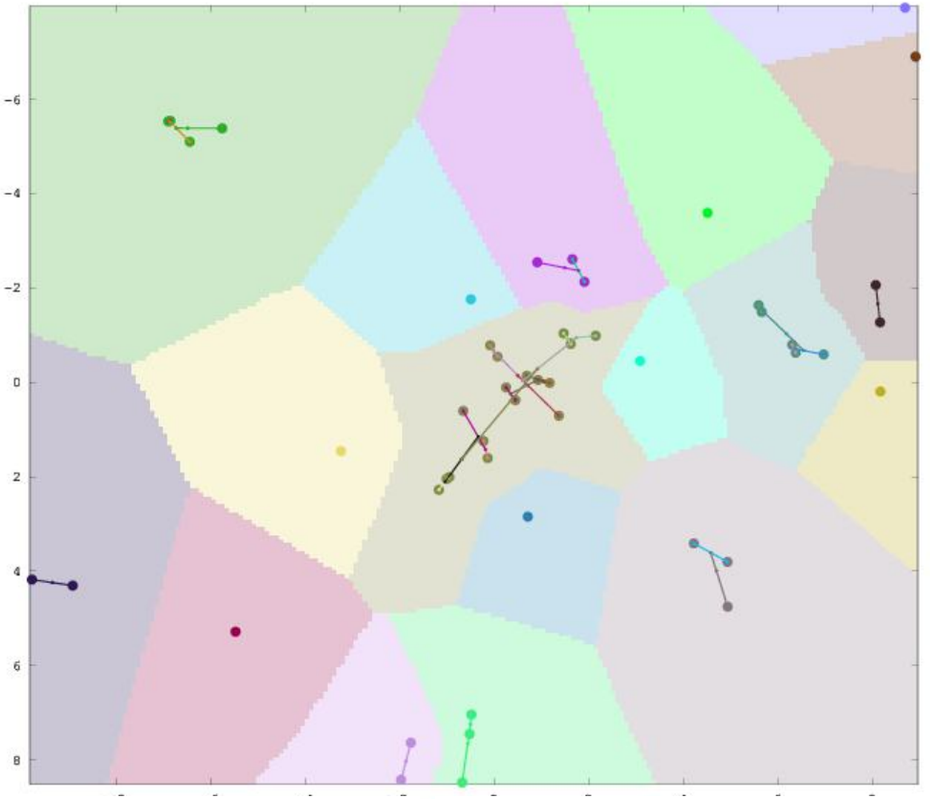


complete linkage



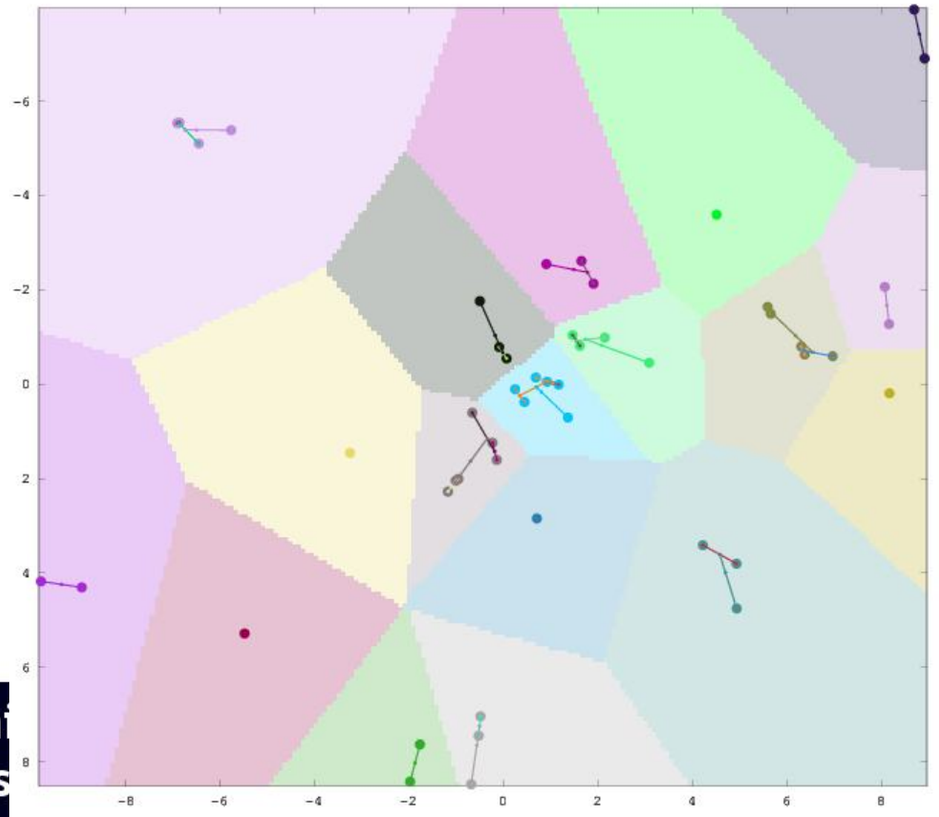
Example

18 clusters

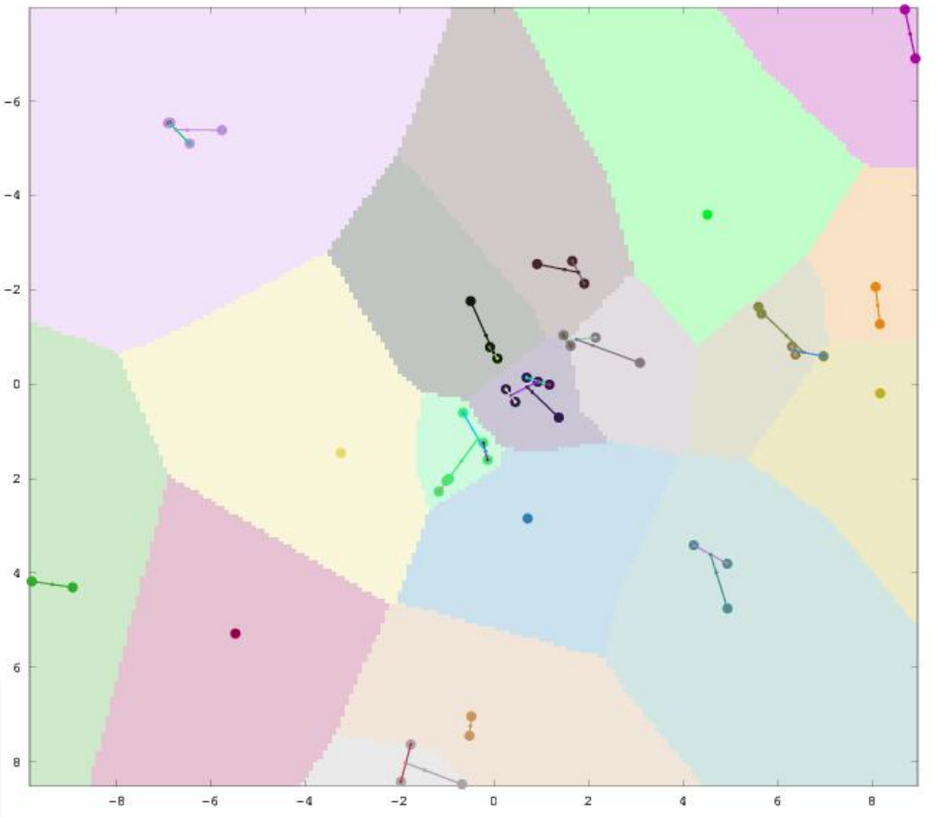


single-linkage

average linkage

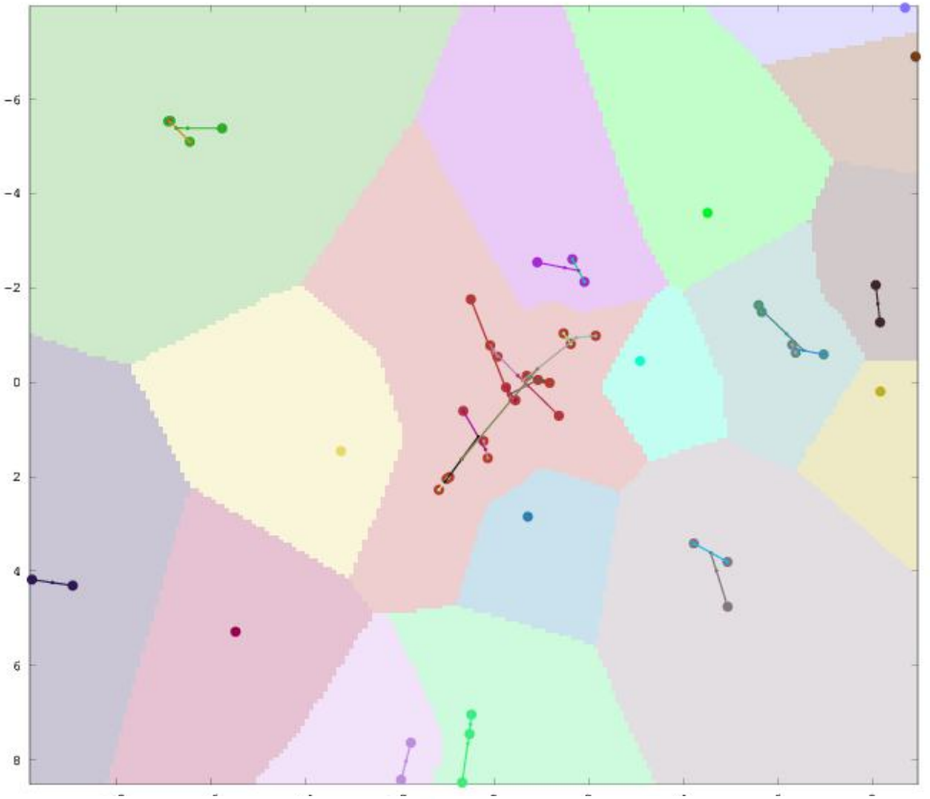


complete linkage



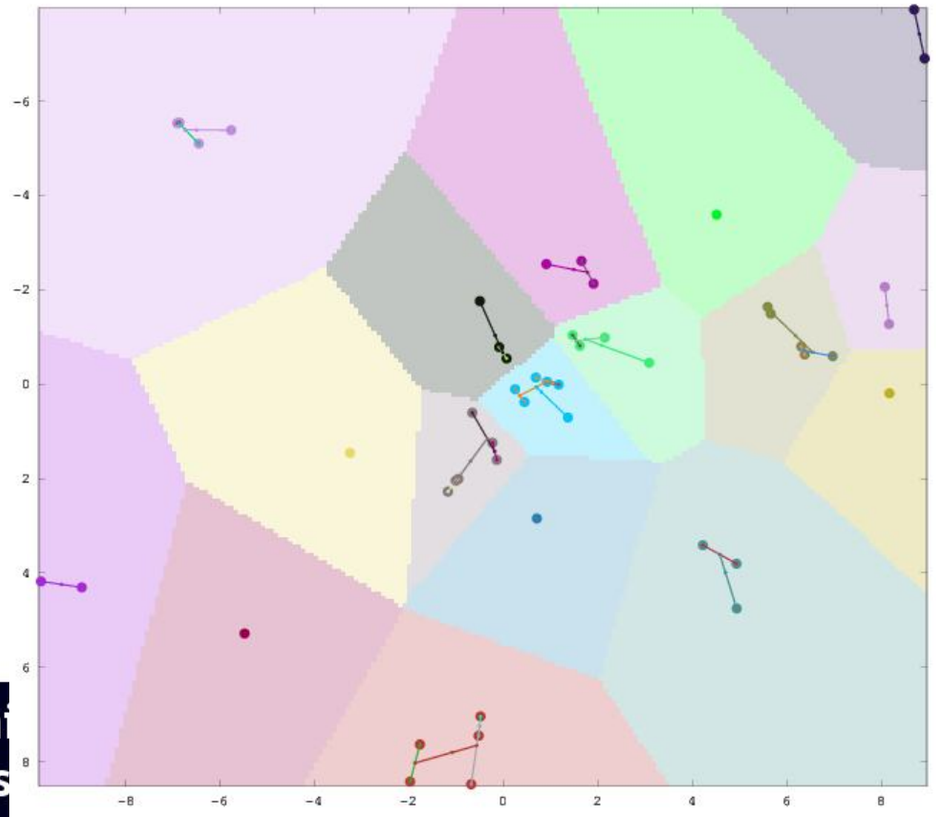
Example

17 clusters

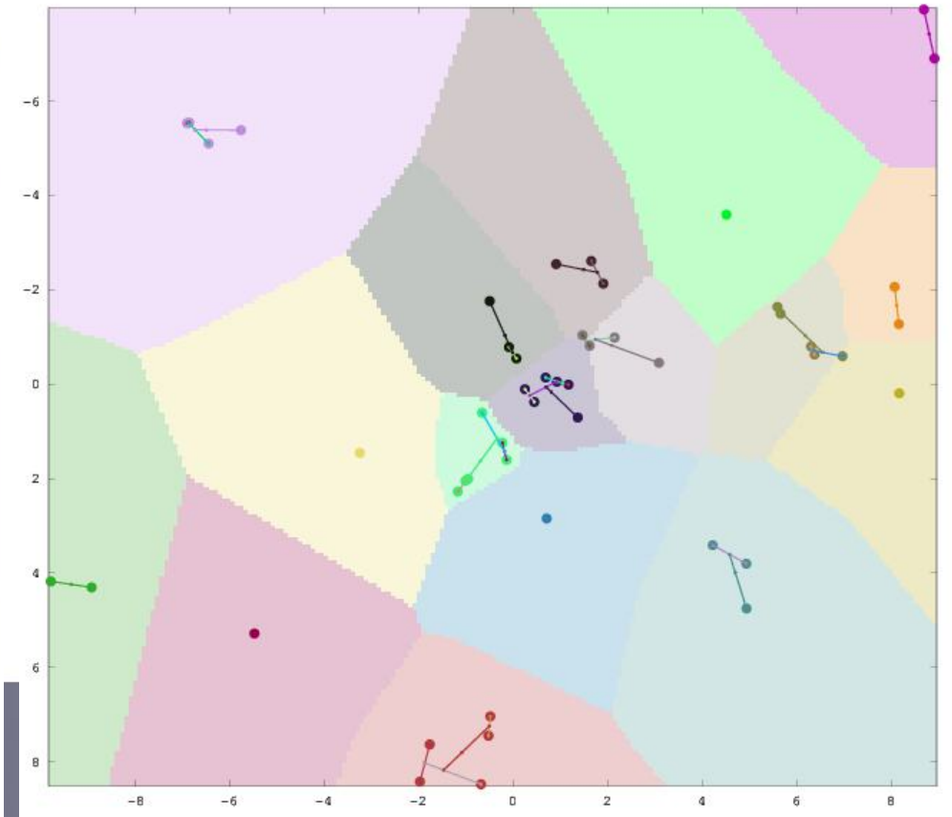


single-linkage

average linkage

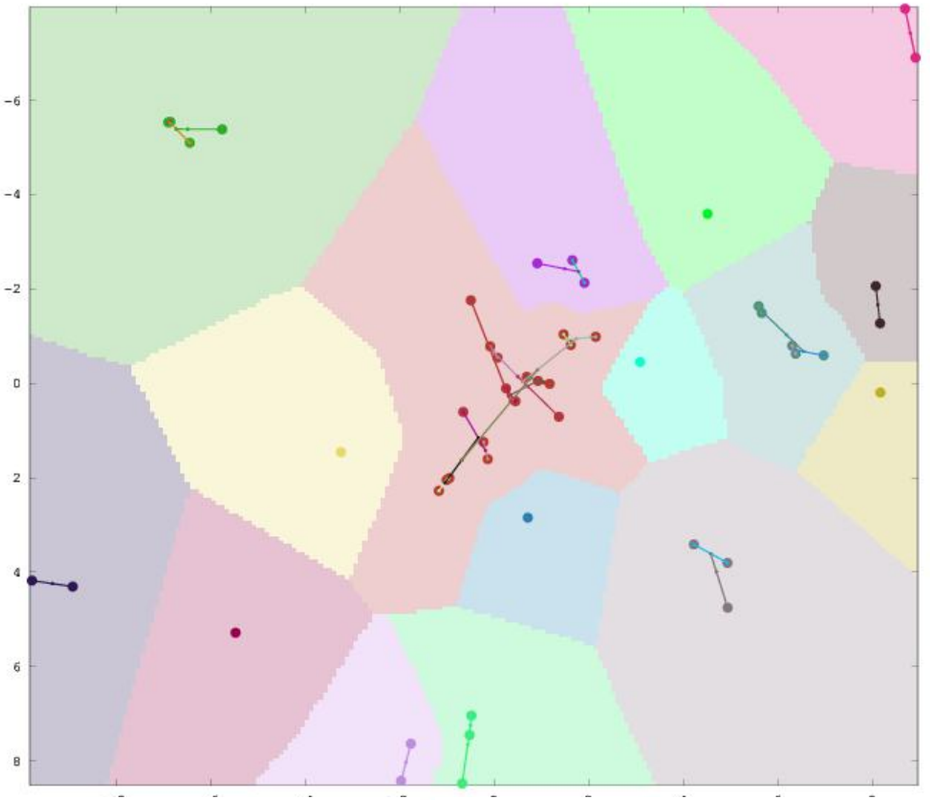


complete linkage



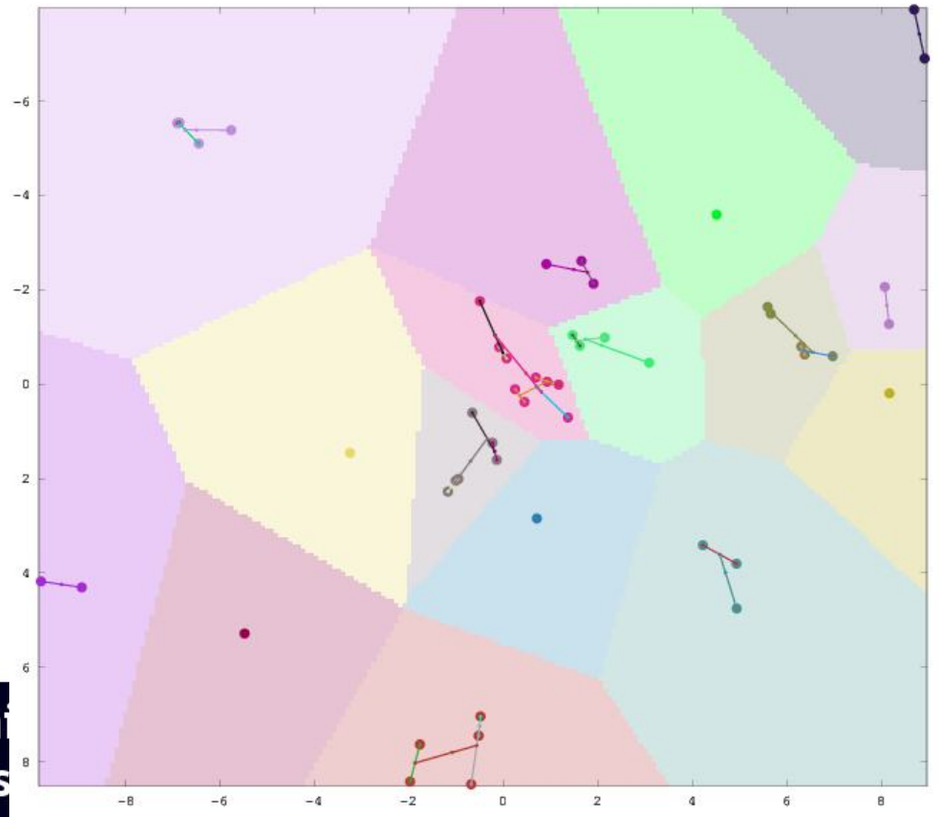
Example

16 clusters

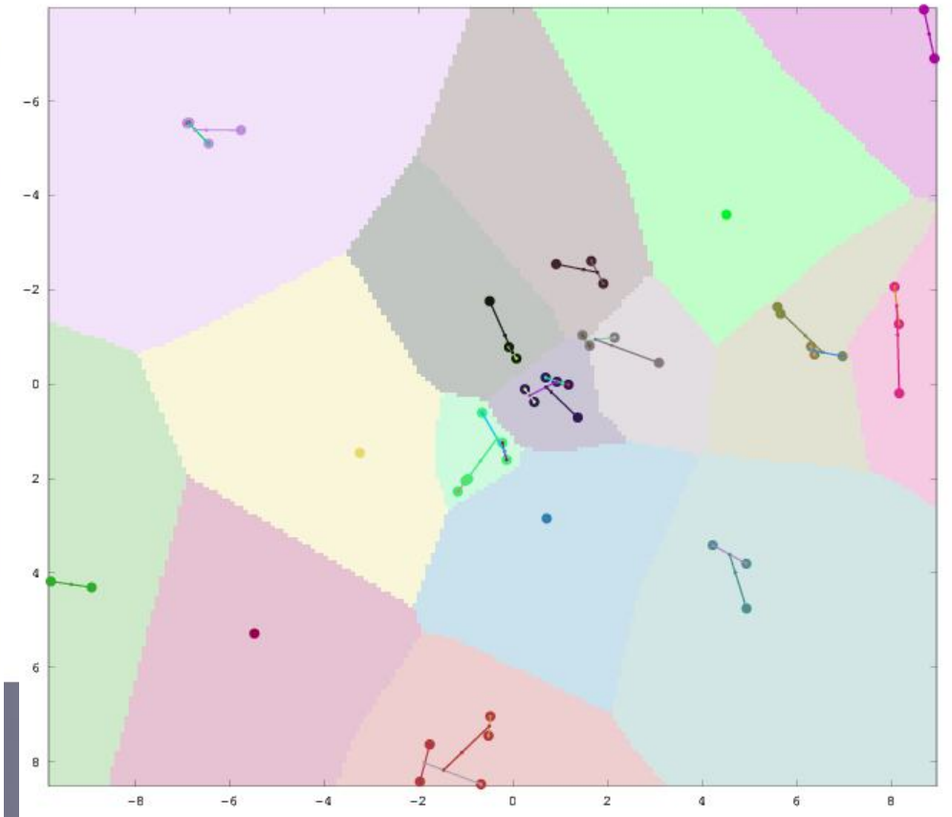


single-linkage

average linkage



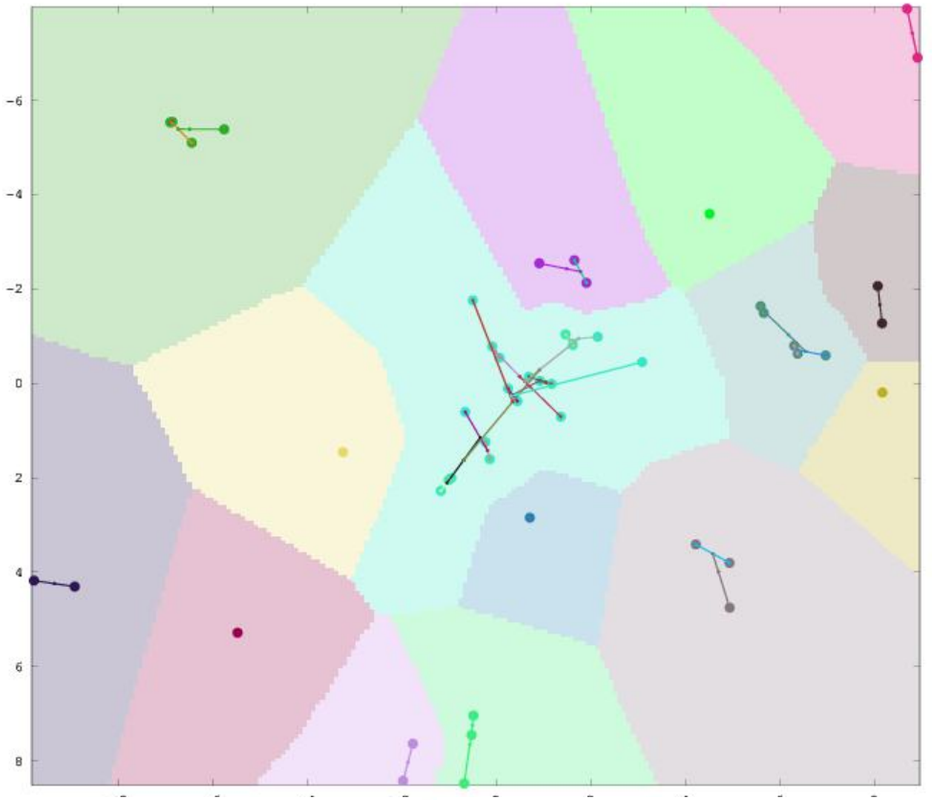
complete linkage



Example

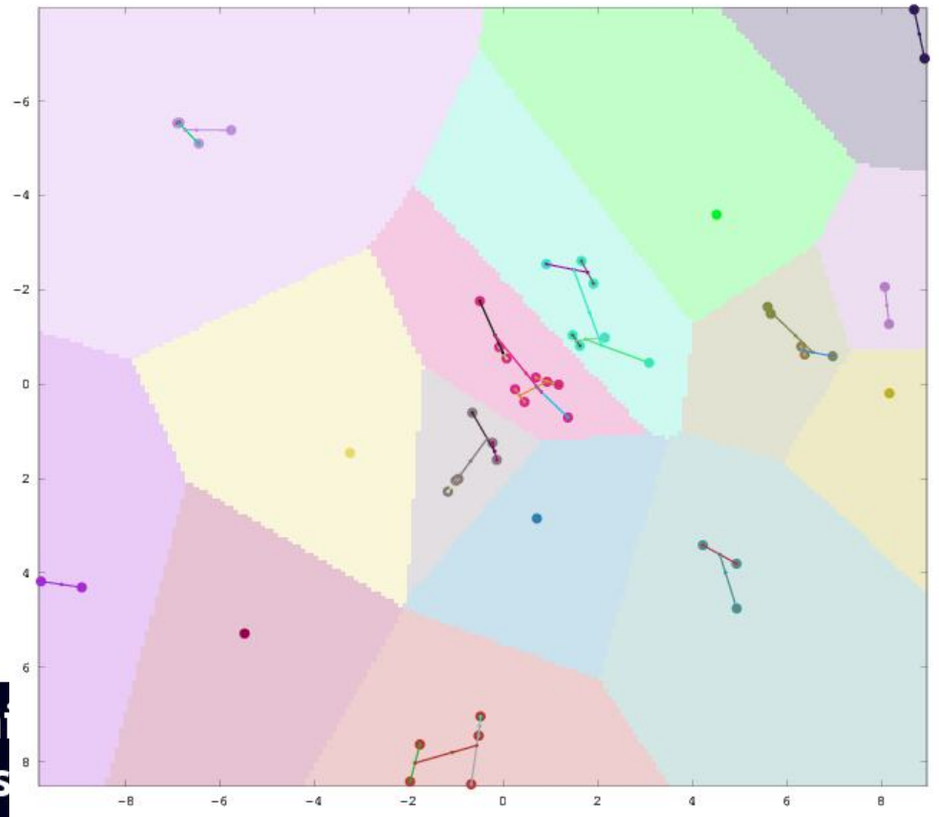
15 clusters

complete-linkage has points outside their cluster's Voronoi region

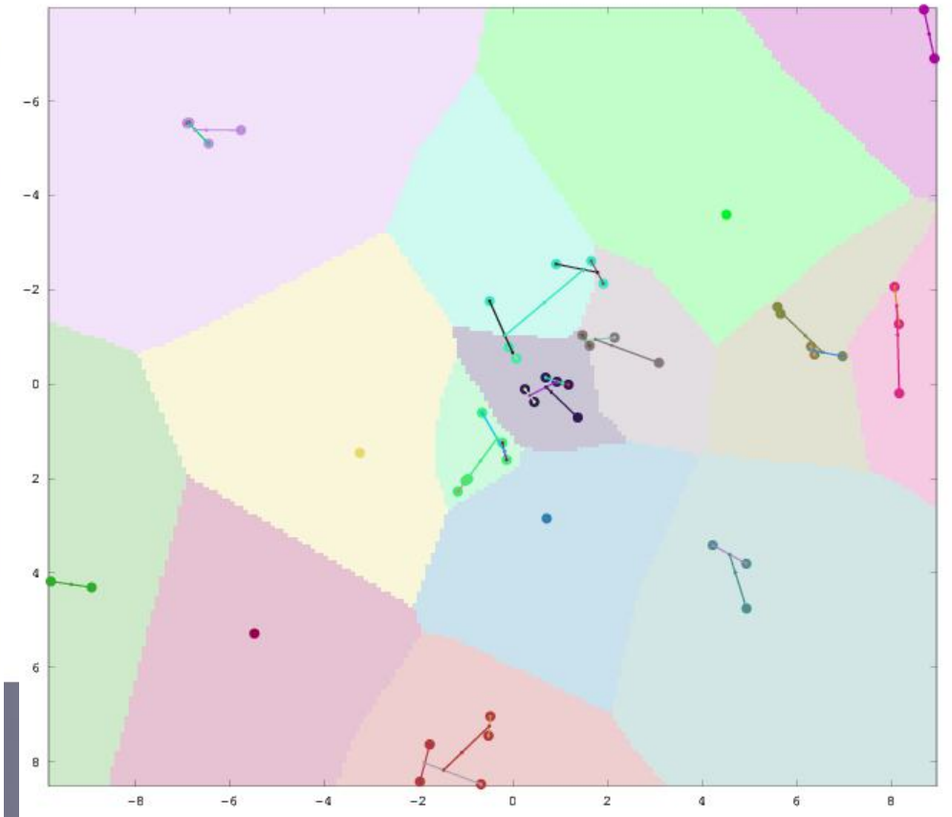


single-linkage

average linkage

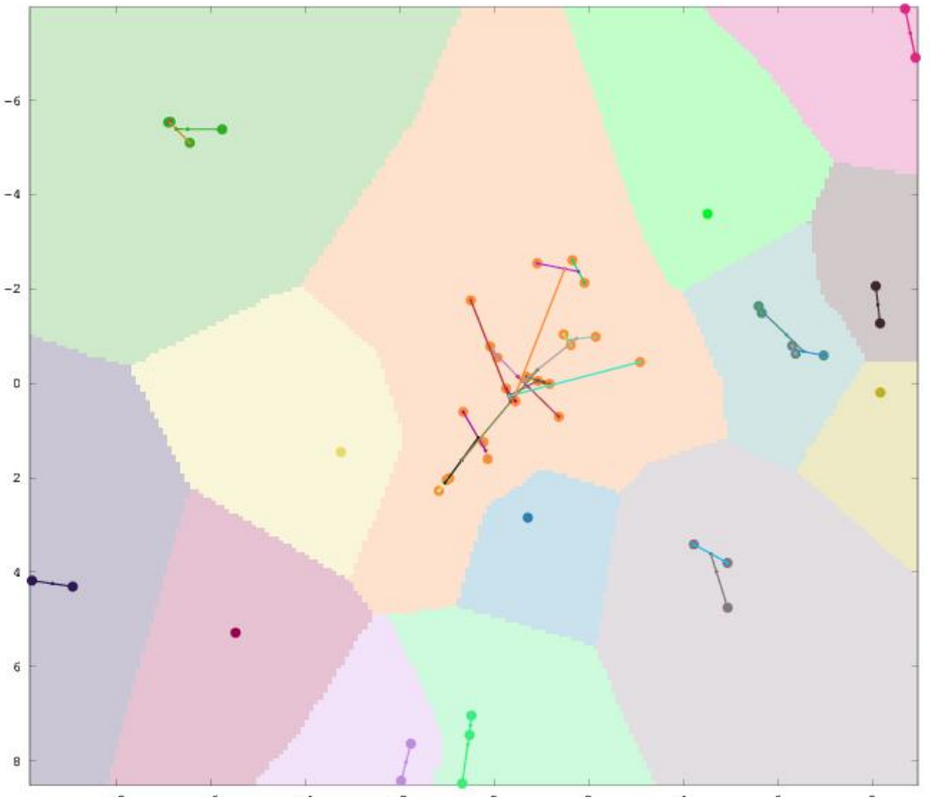


complete linkage



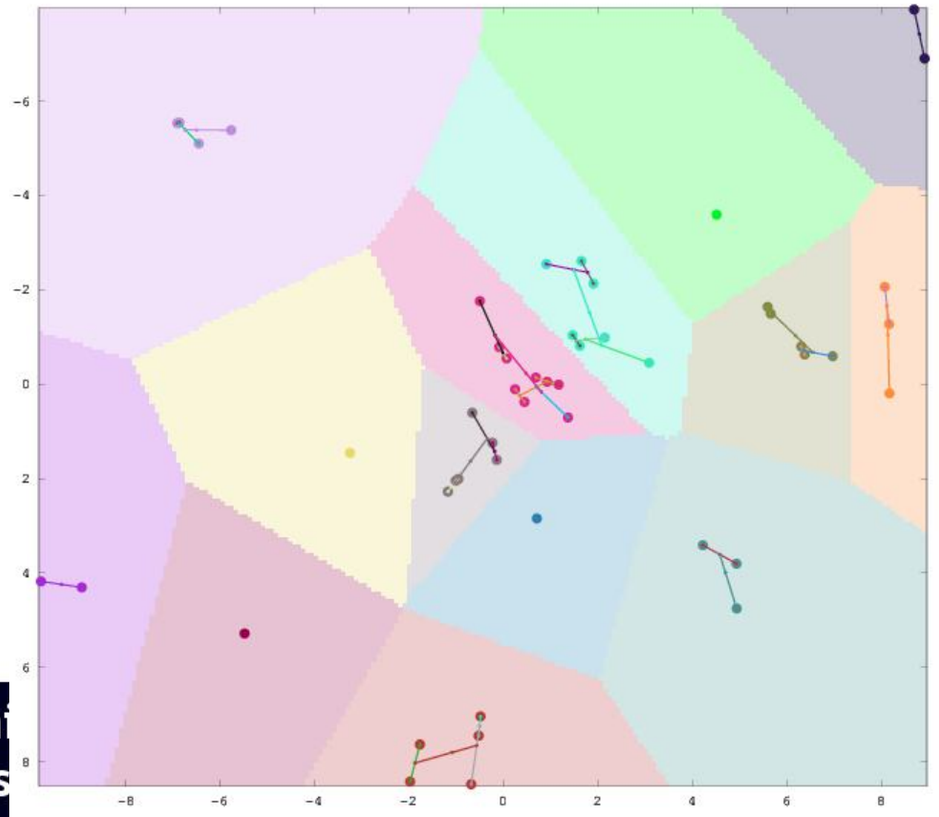
Example

14 clusters

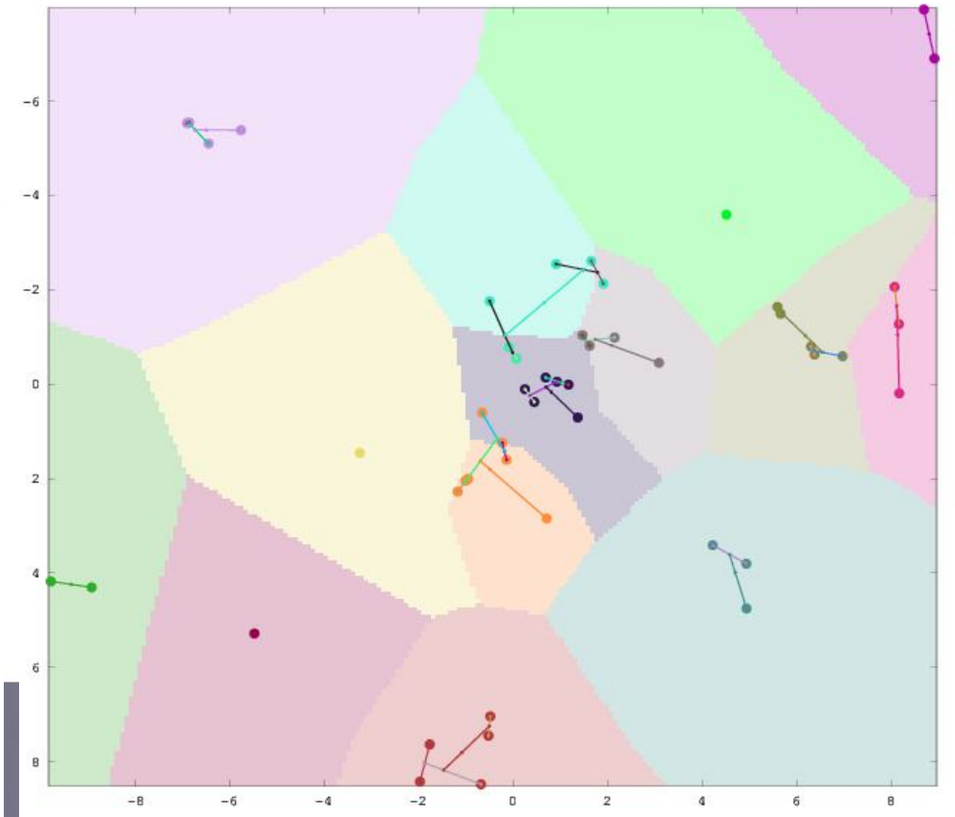


single-linkage

average linkage

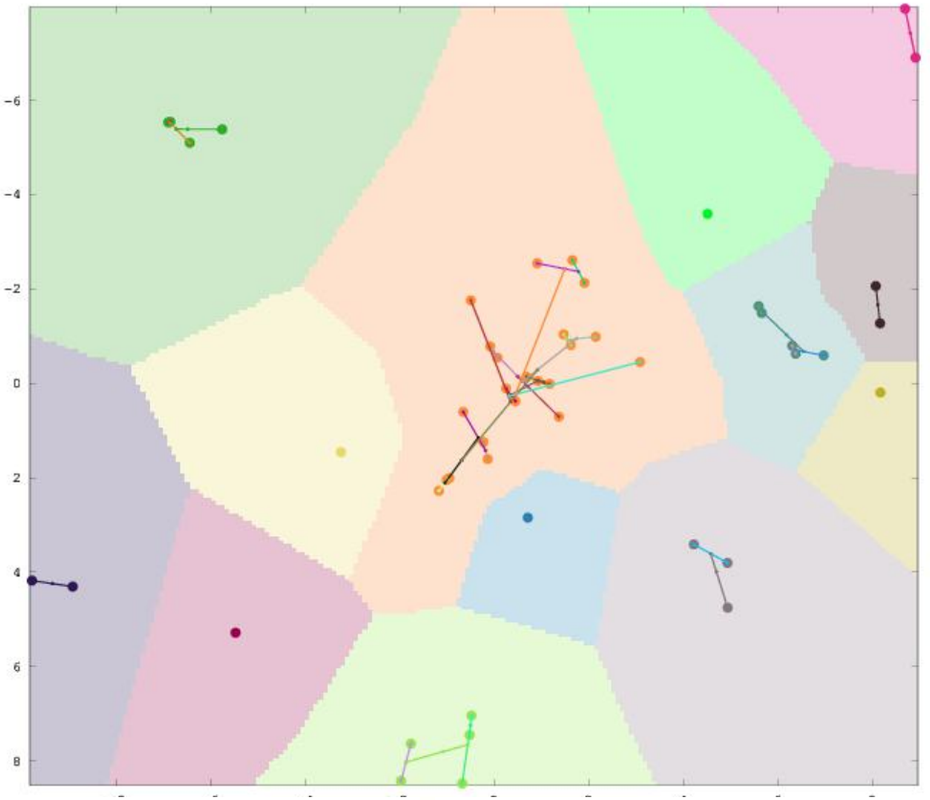


complete linkage



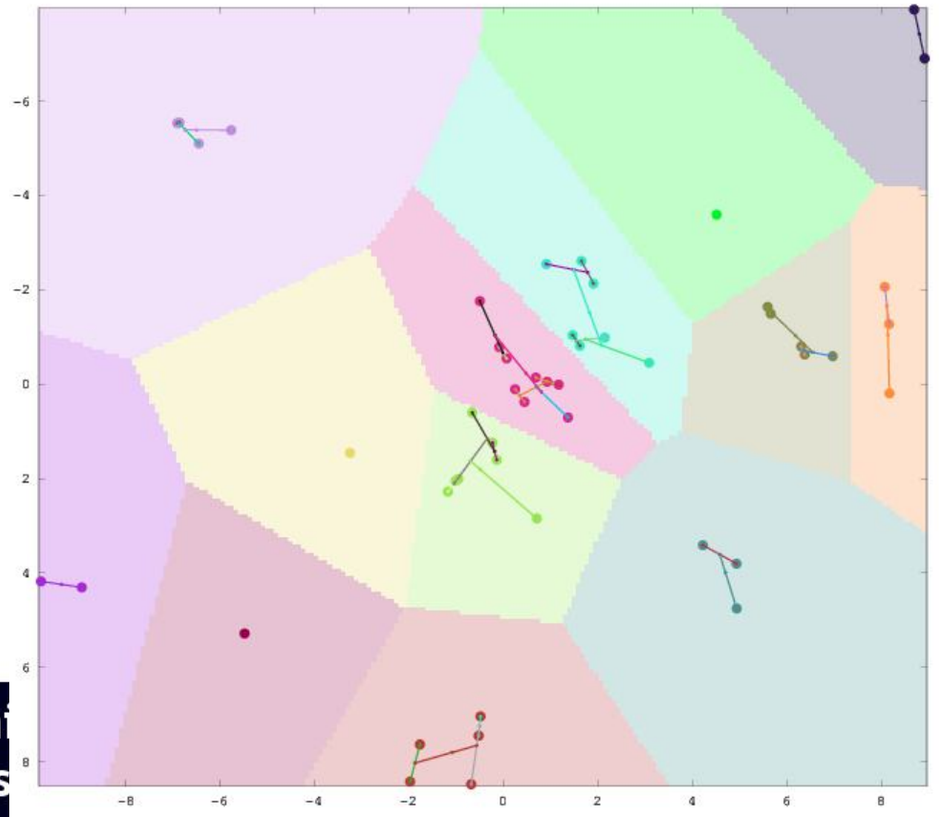
Example

13 clusters

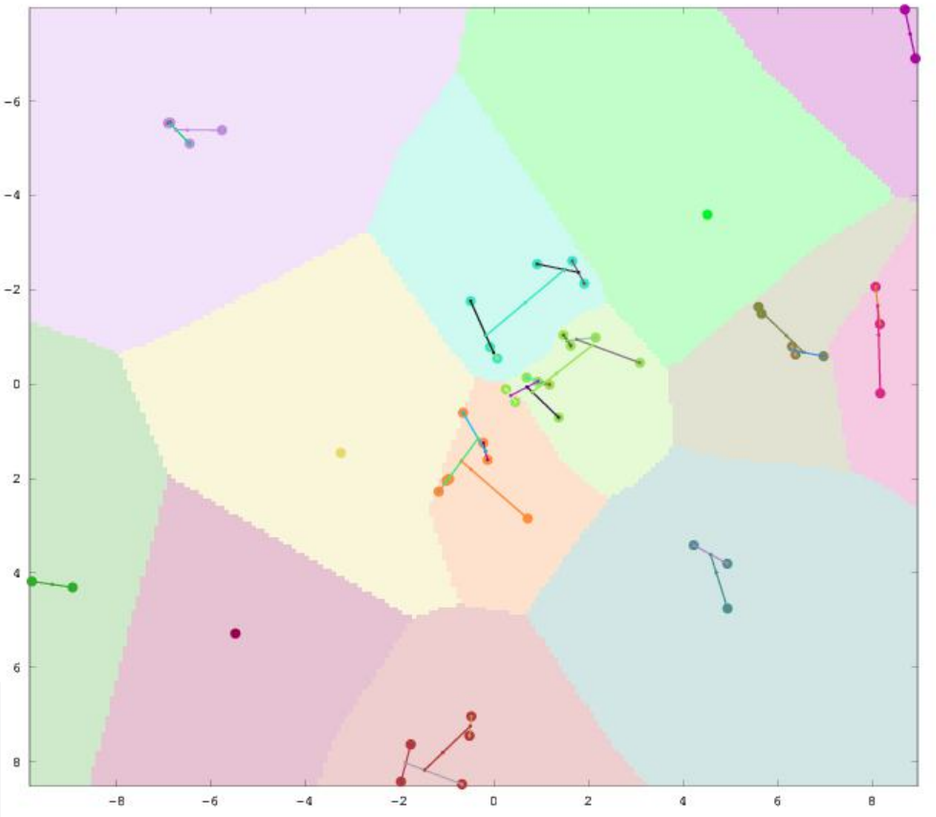


single-linkage

average linkage

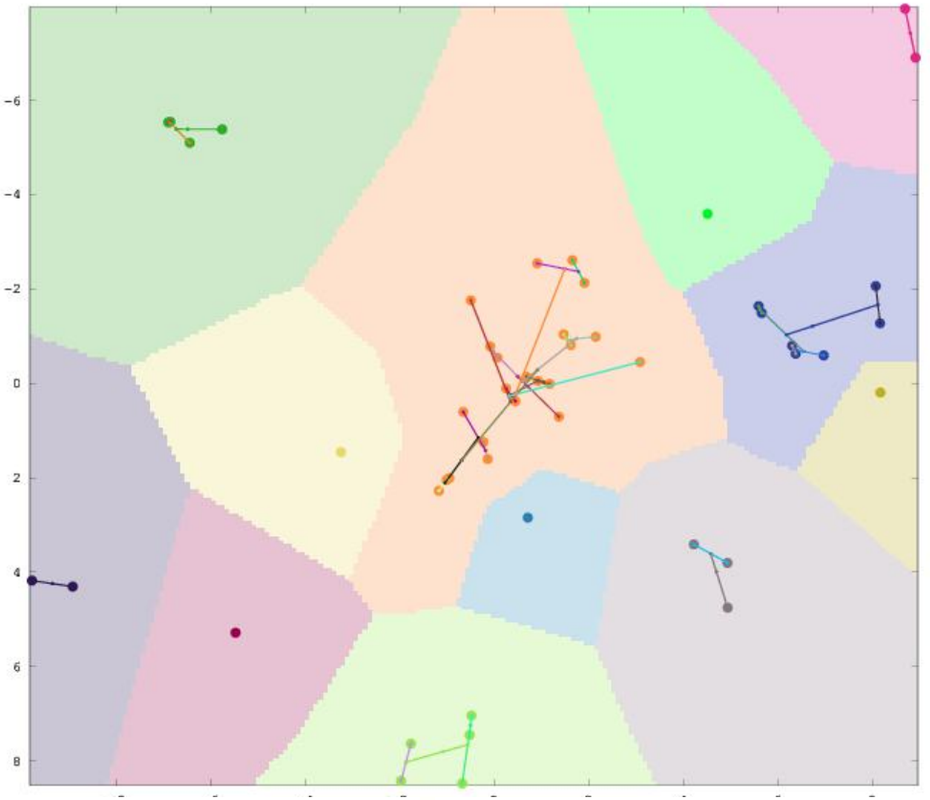


complete linkage



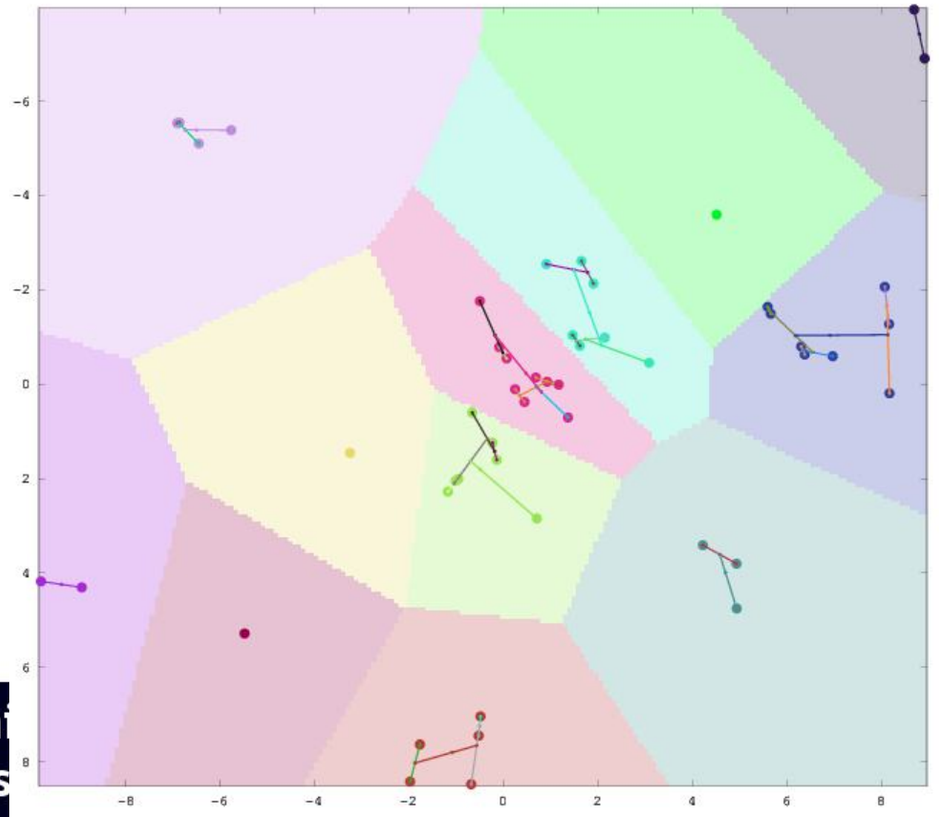
Example

12 clusters

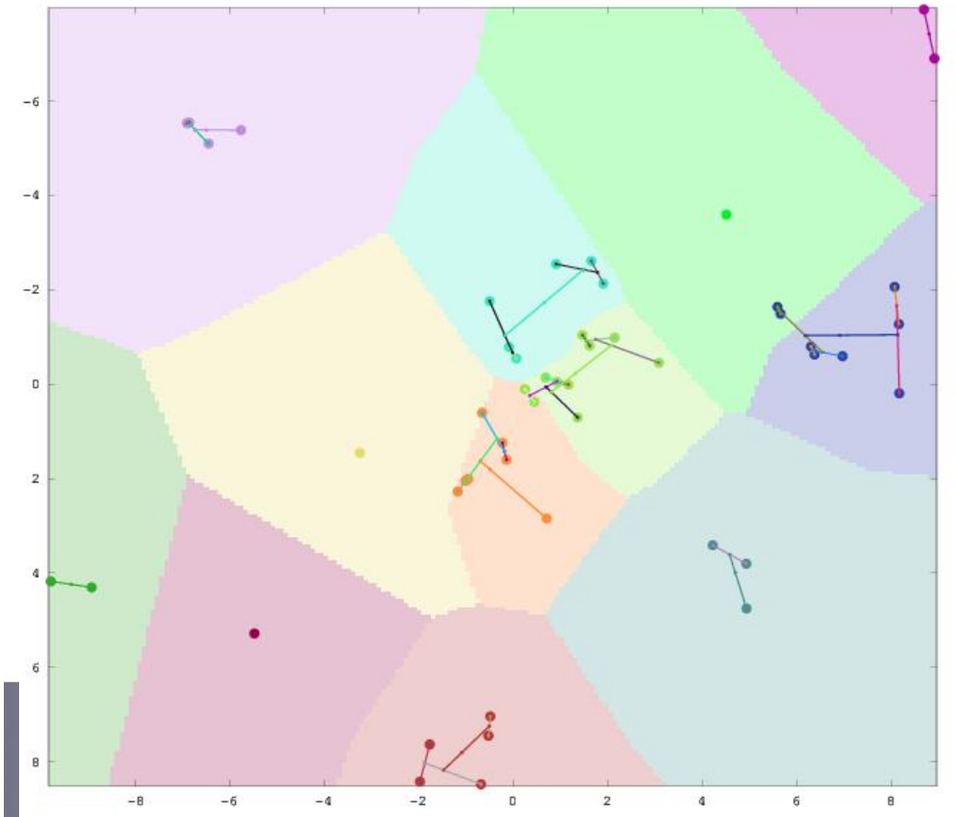


single-linkage

average linkage



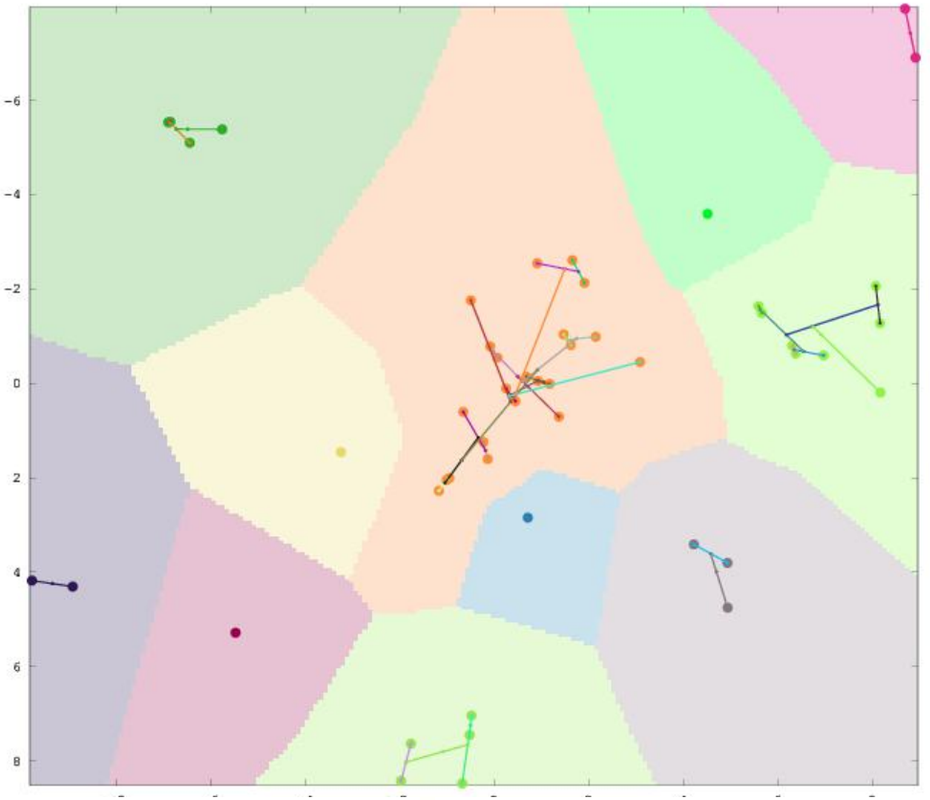
complete linkage



Example

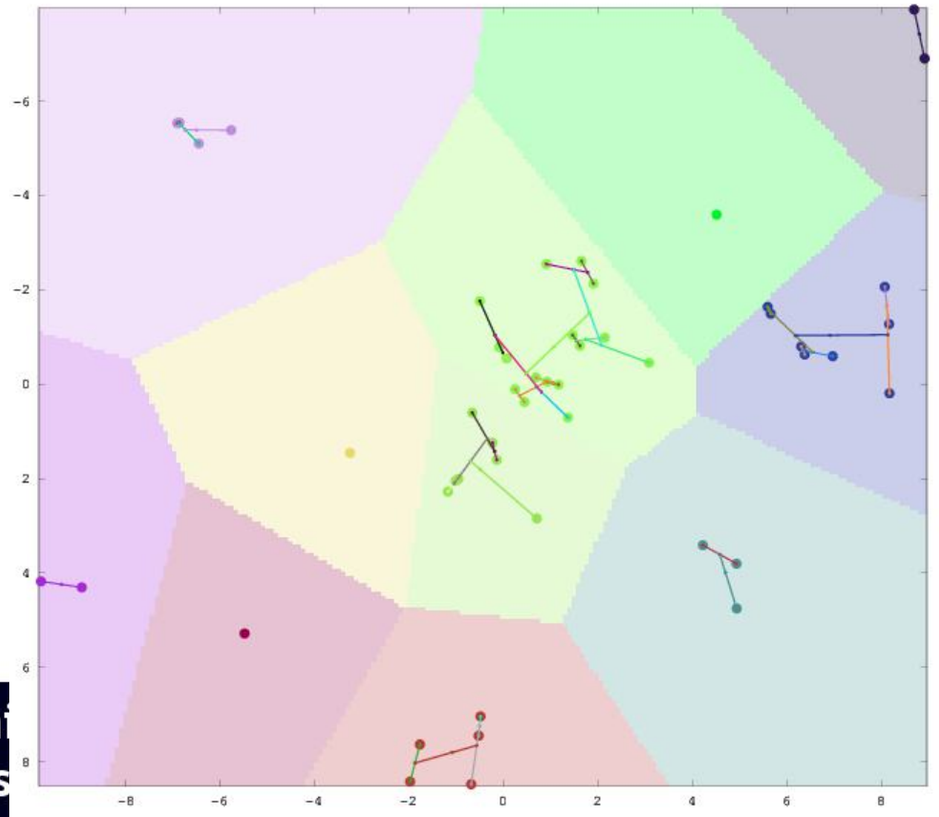
11 clusters

complete-linkage: merging two clusters shrank the Voronoi region

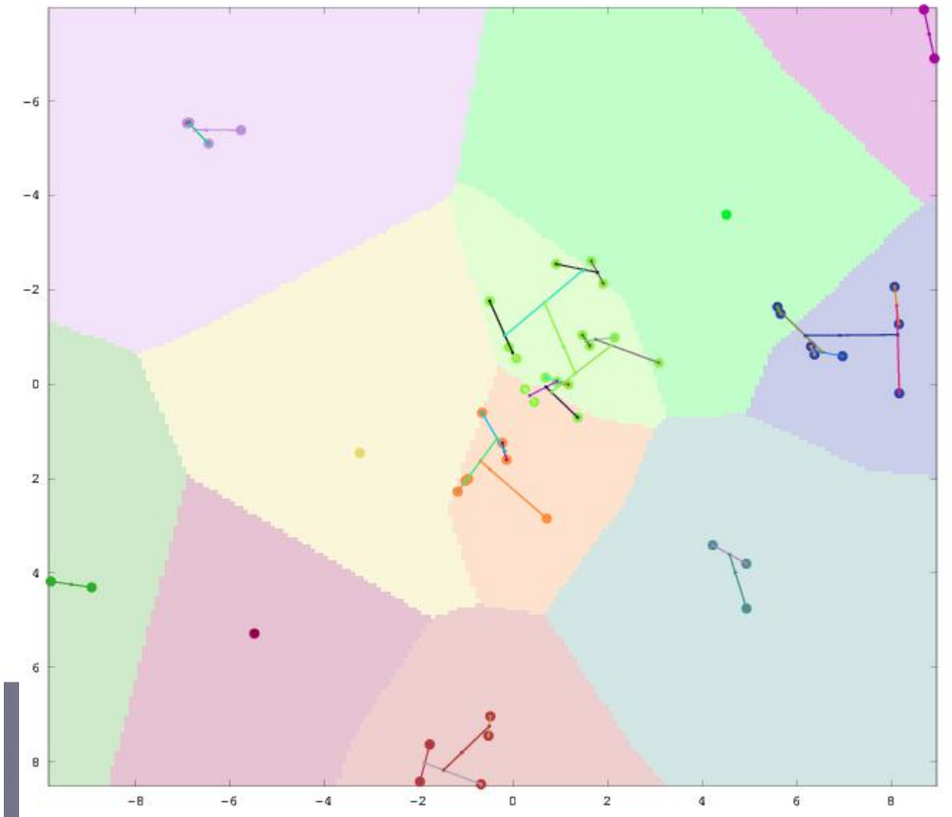


single-linkage

average linkage

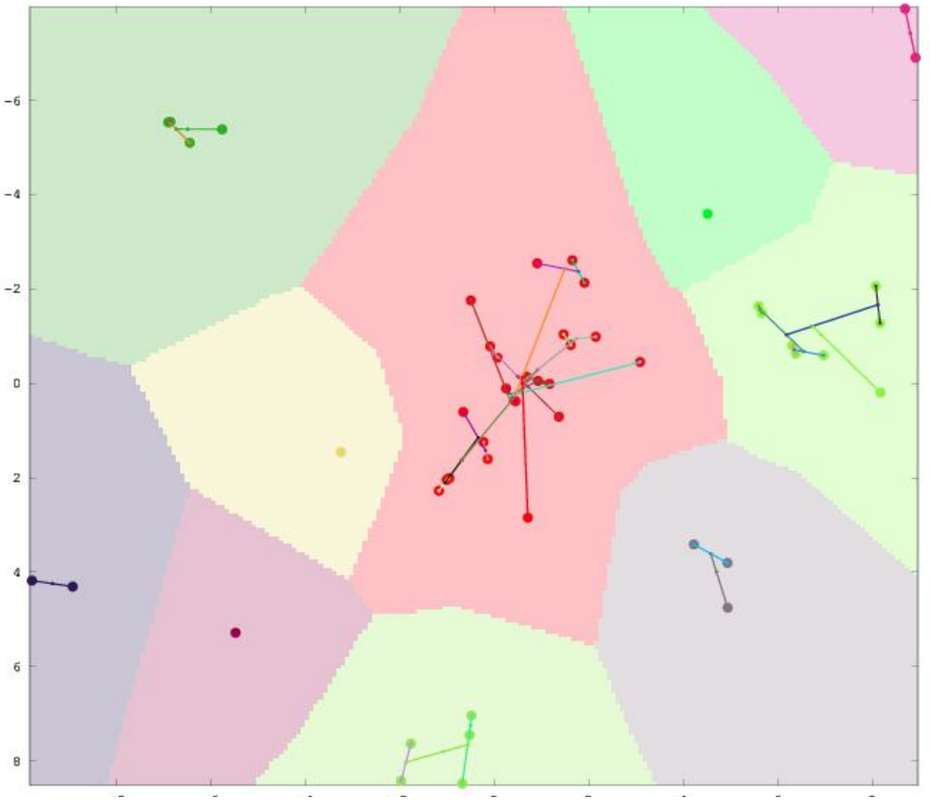


complete linkage



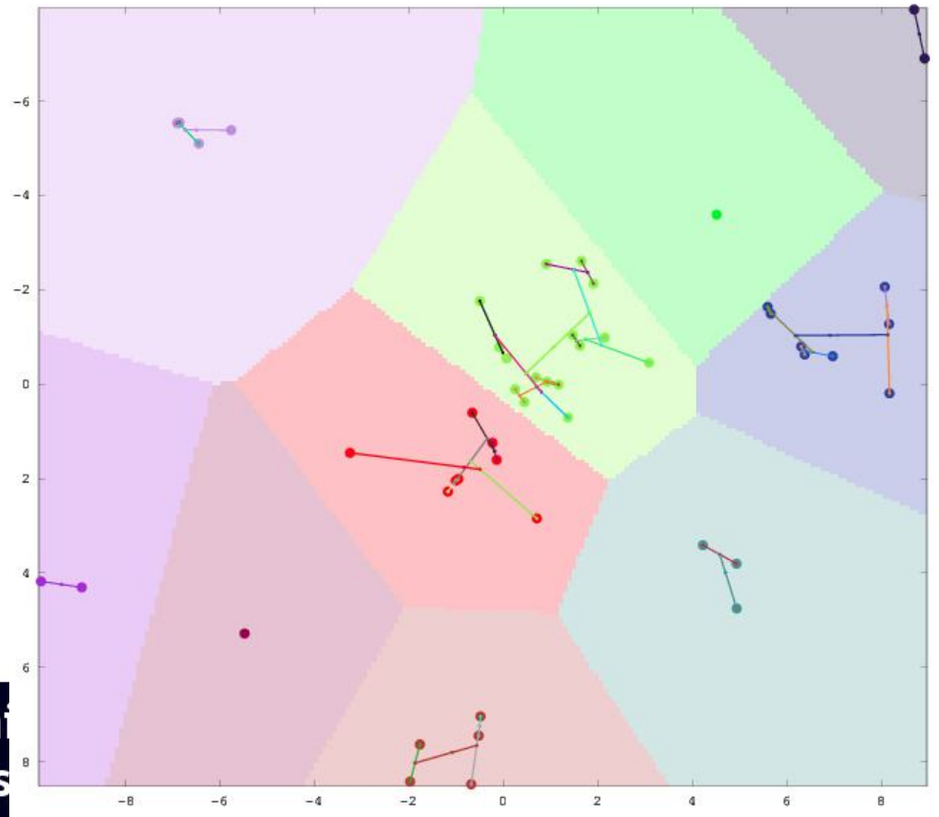
Example

10 clusters

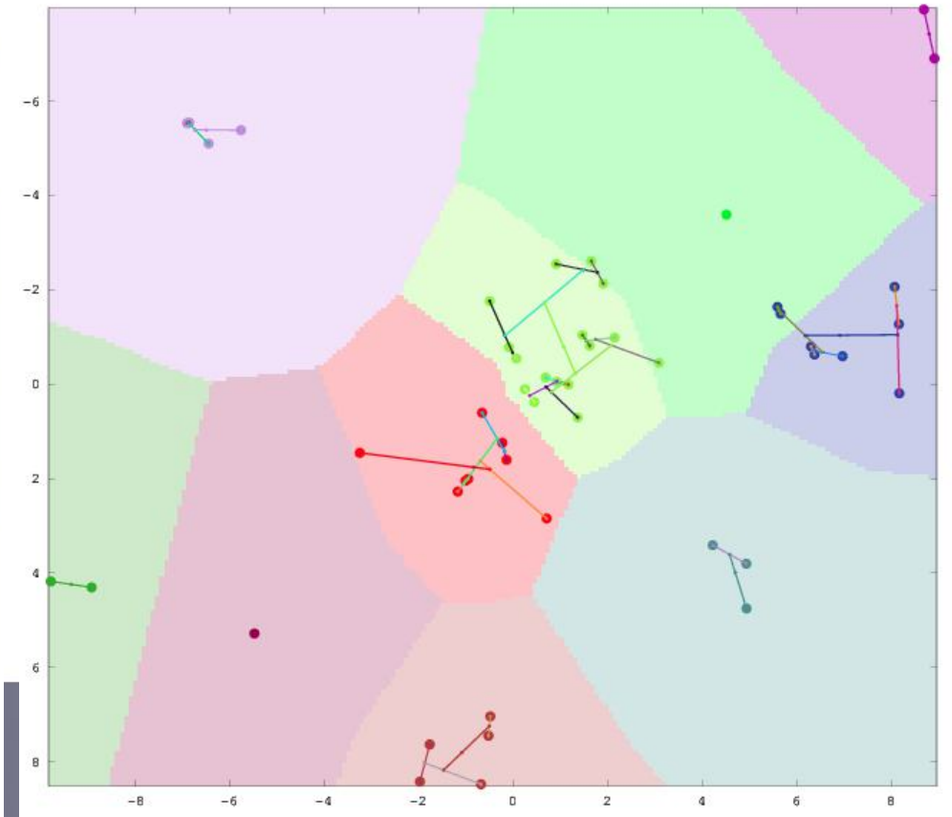


single-linkage

average linkage

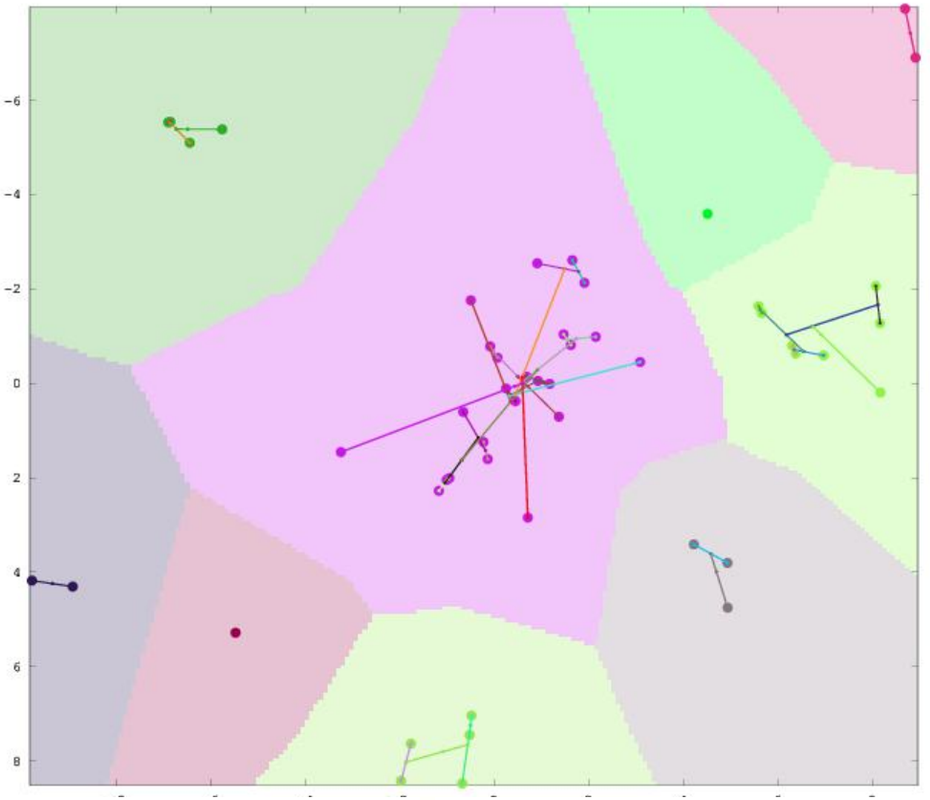


complete linkage



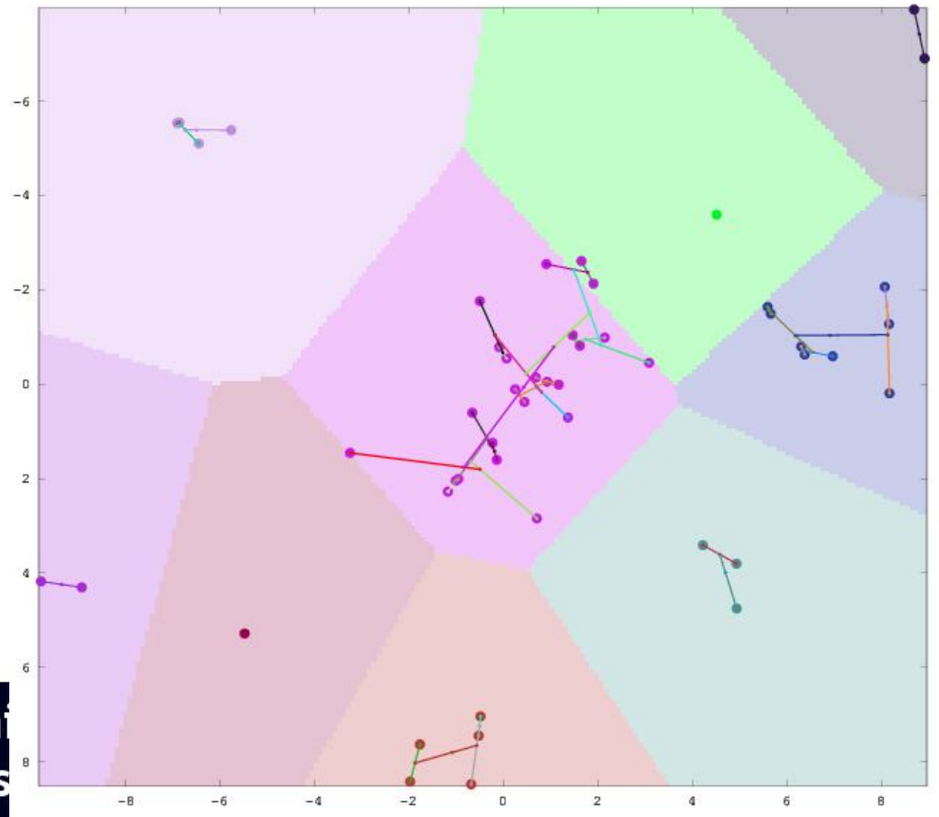
Example

9 clusters

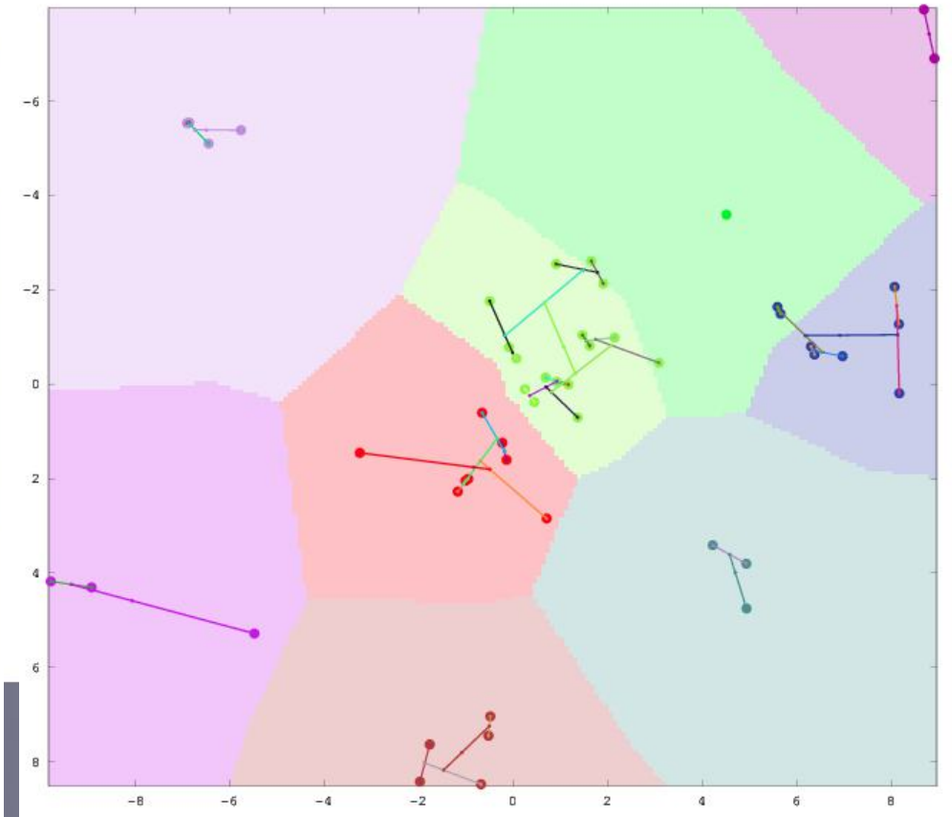


single-linkage

average linkage

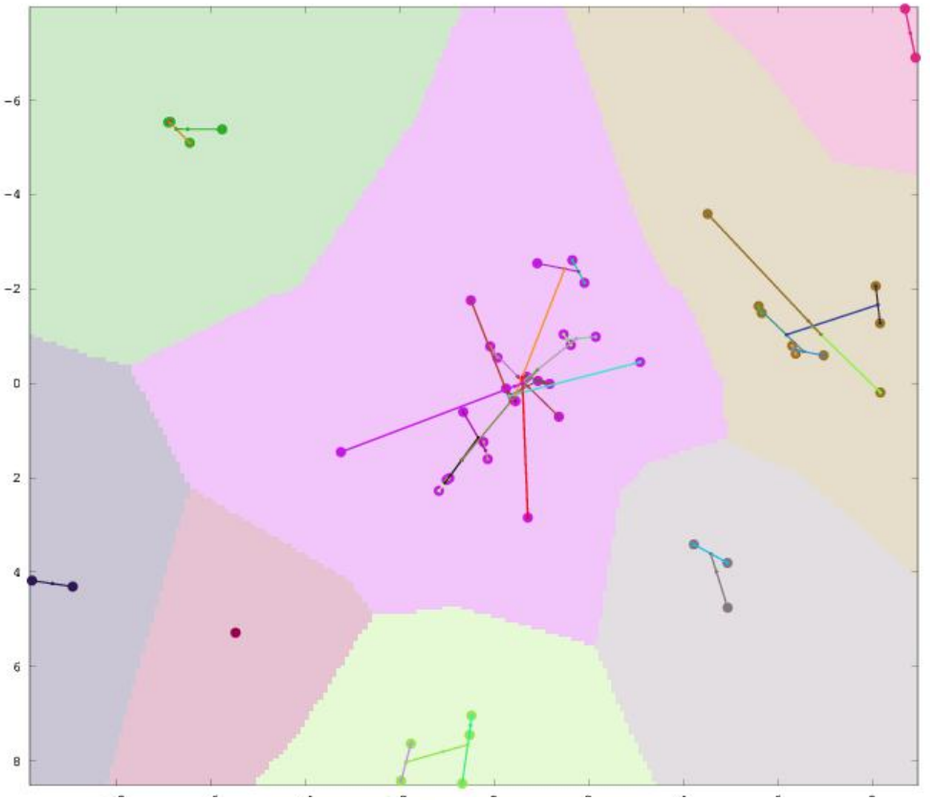


complete linkage



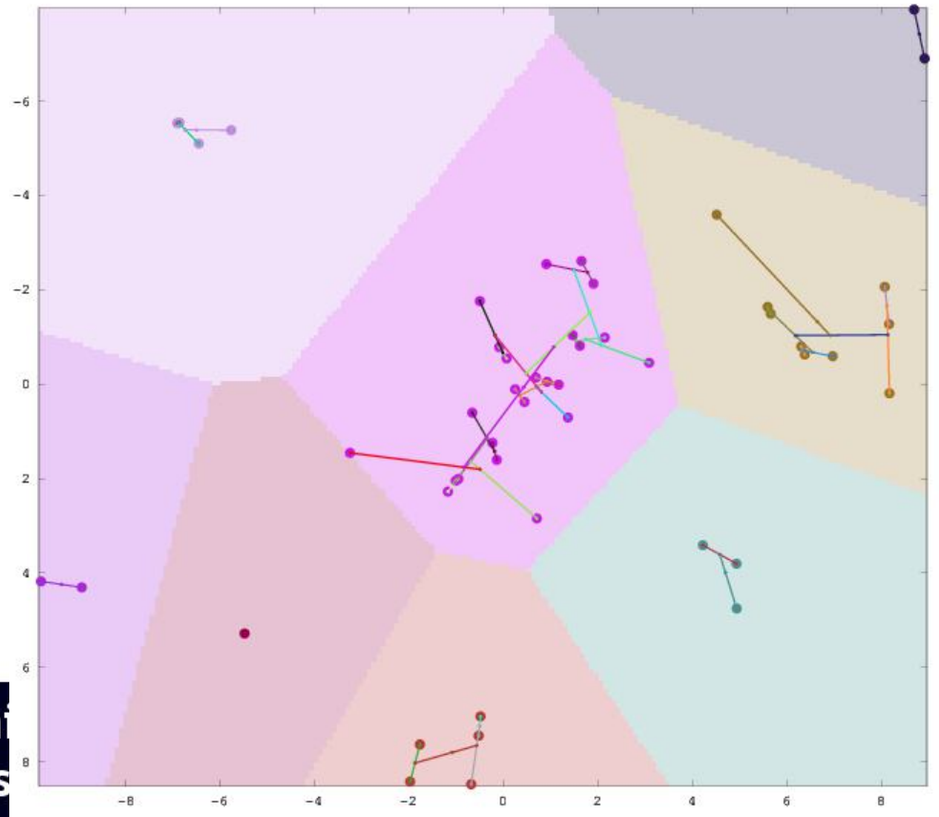
Example

8 clusters

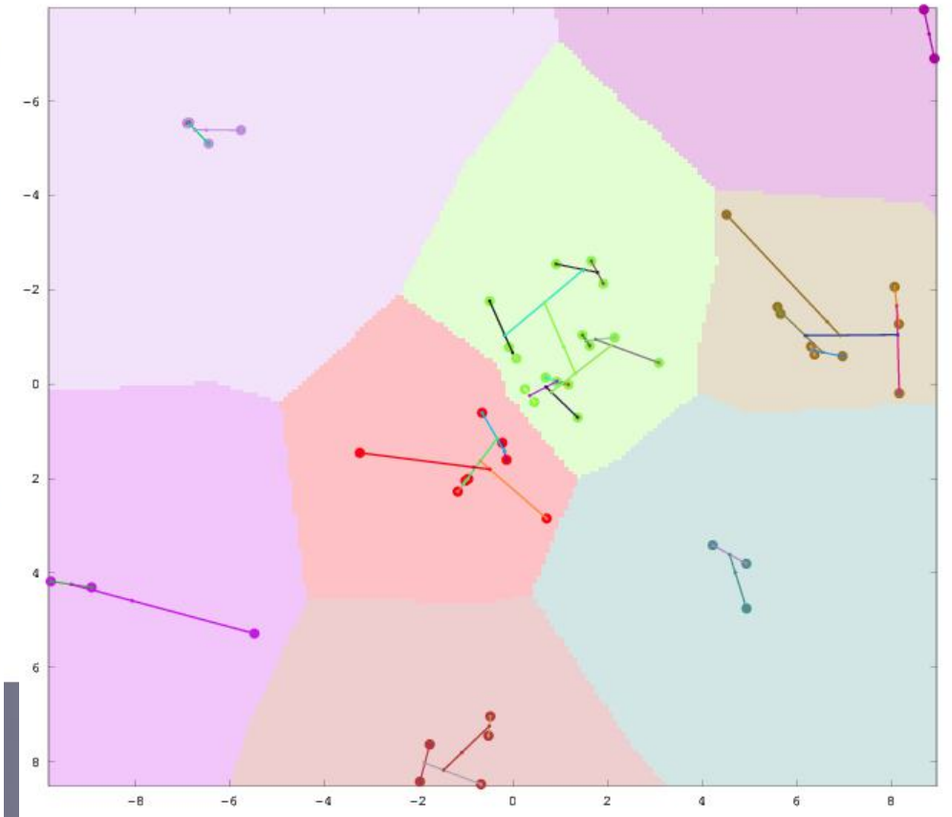


single-linkage

average linkage

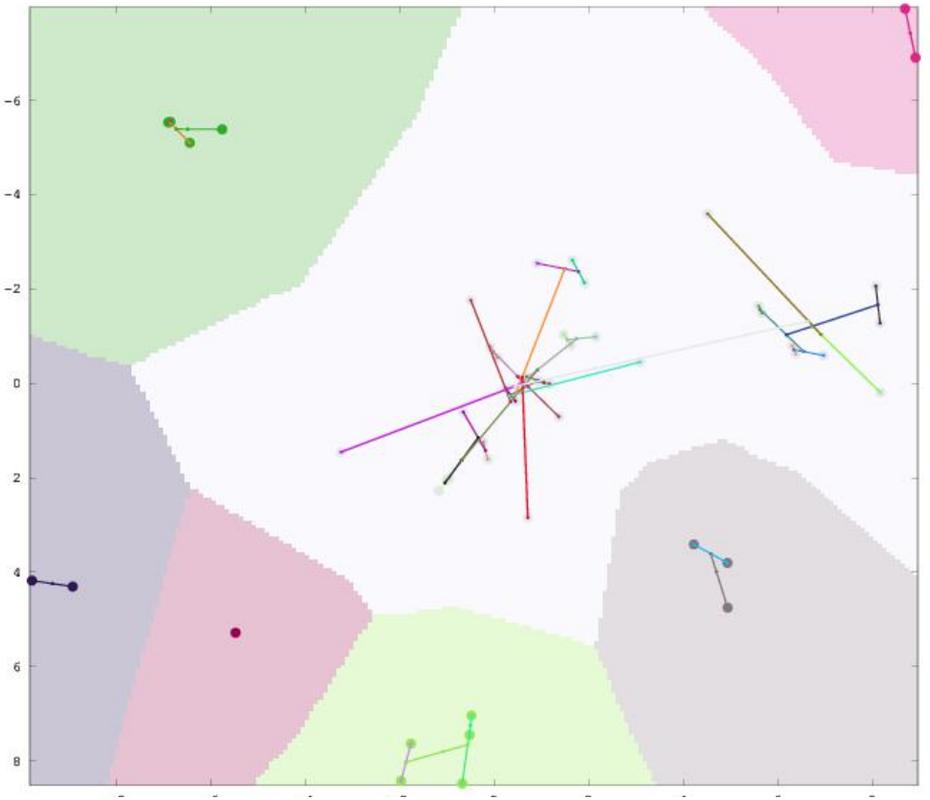


complete linkage



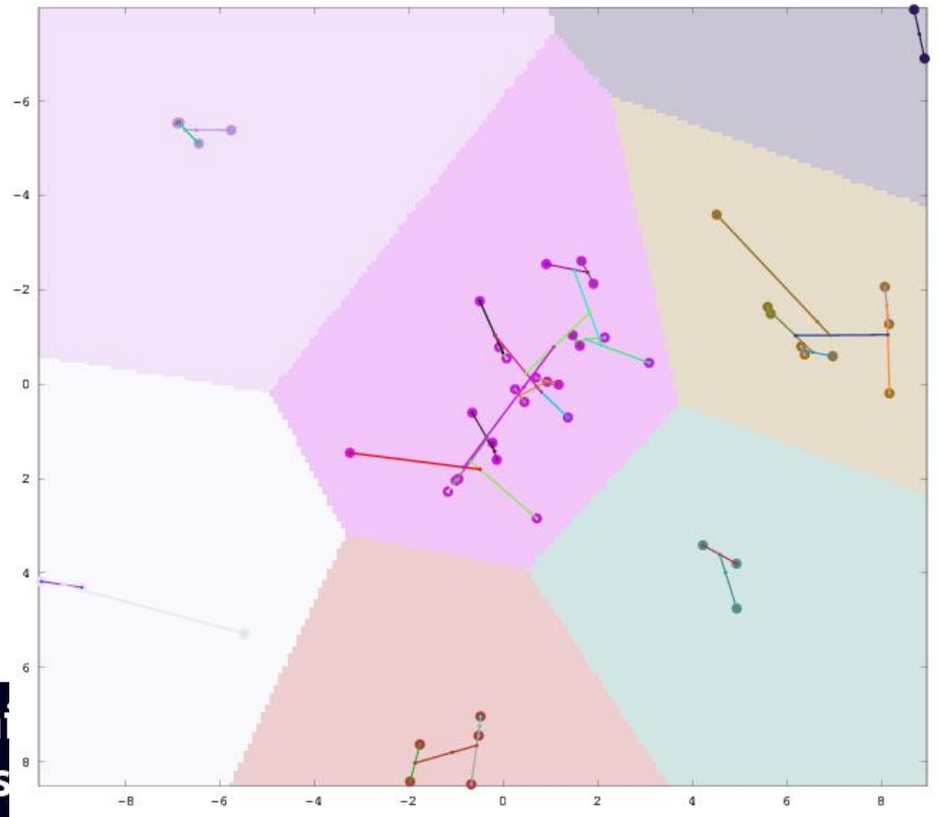
Example

7 clusters

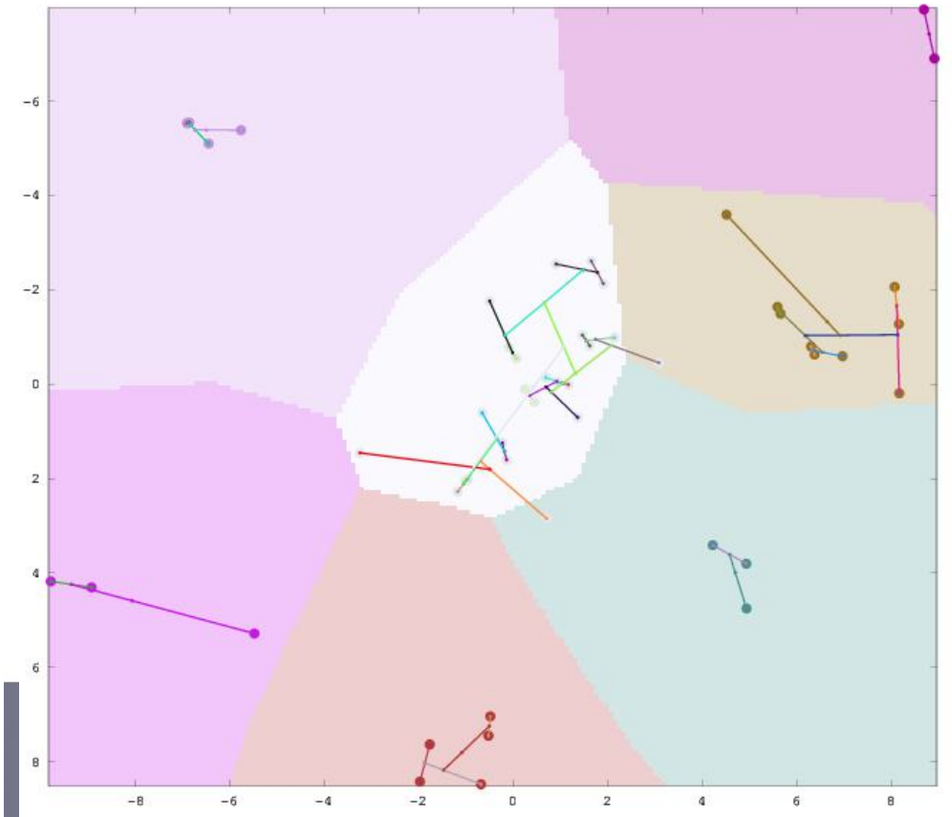


single-linkage

average linkage

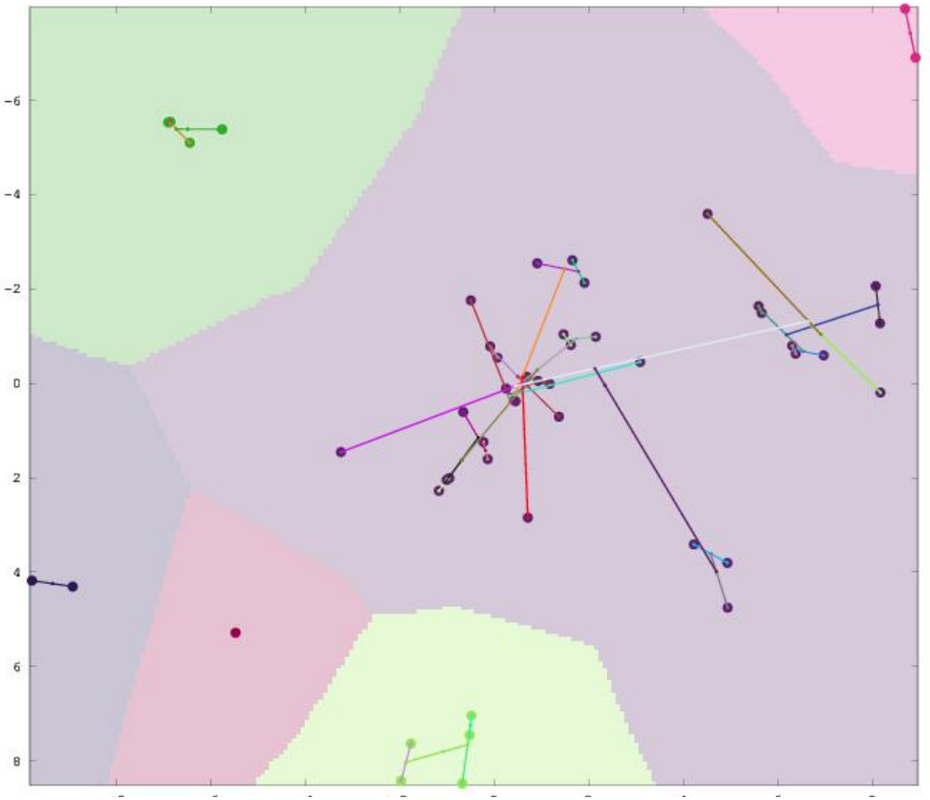


complete linkage



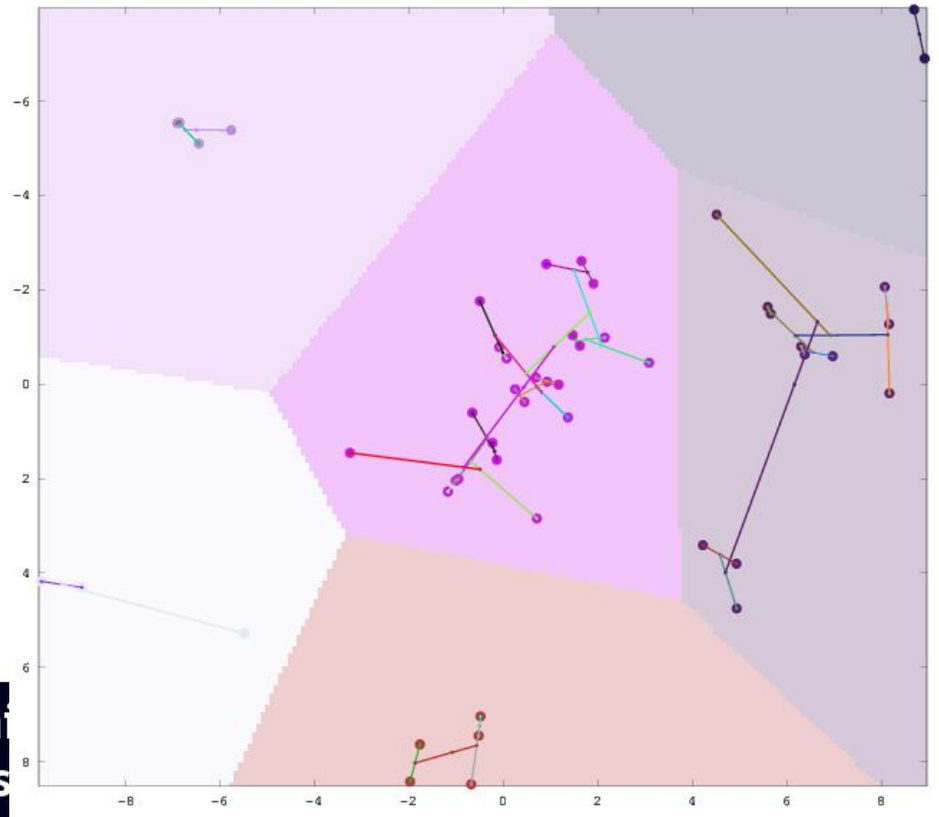
Example

6 clusters

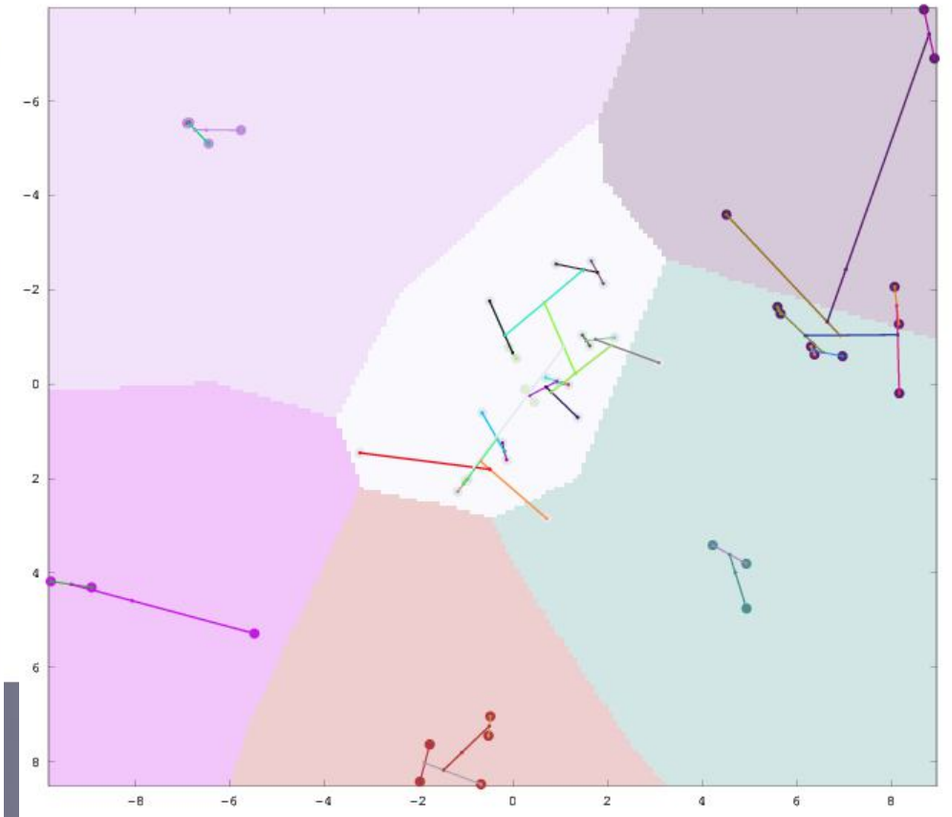


single-linkage

average linkage



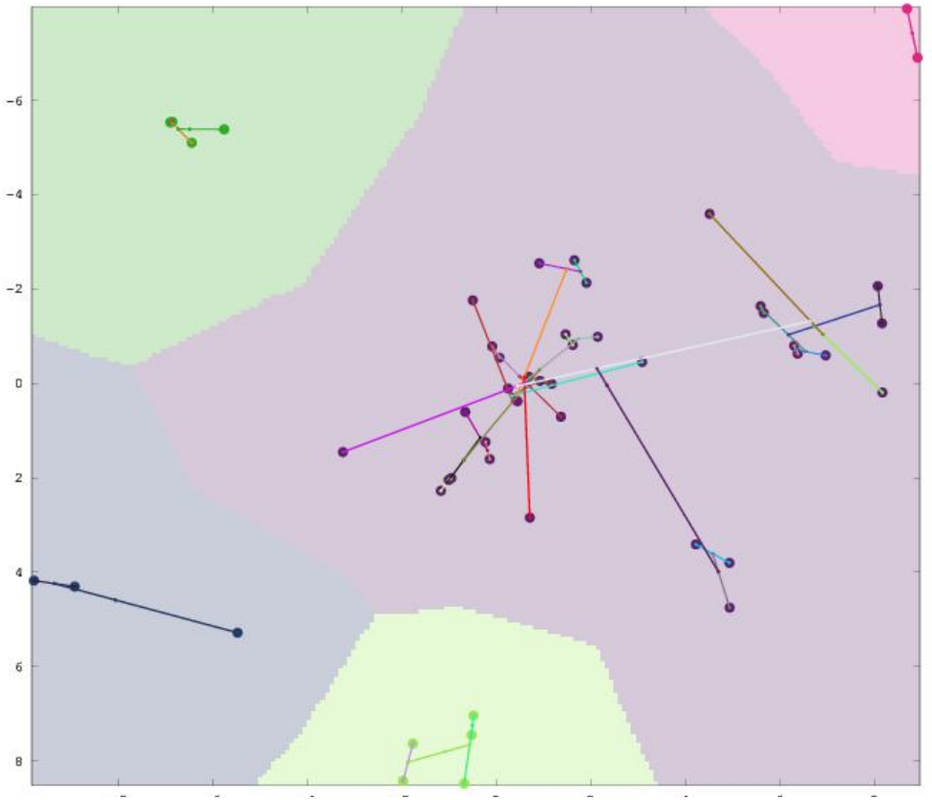
complete linkage



Example

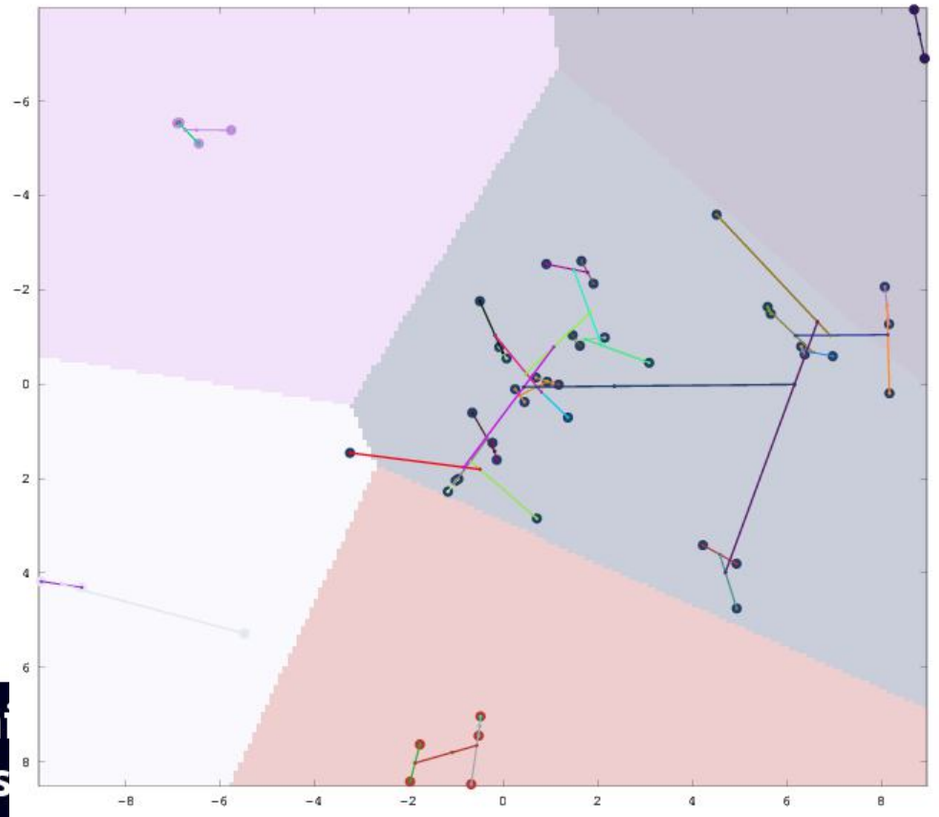
5 clusters

average-linkage has points outside their cluster's Voronoi region

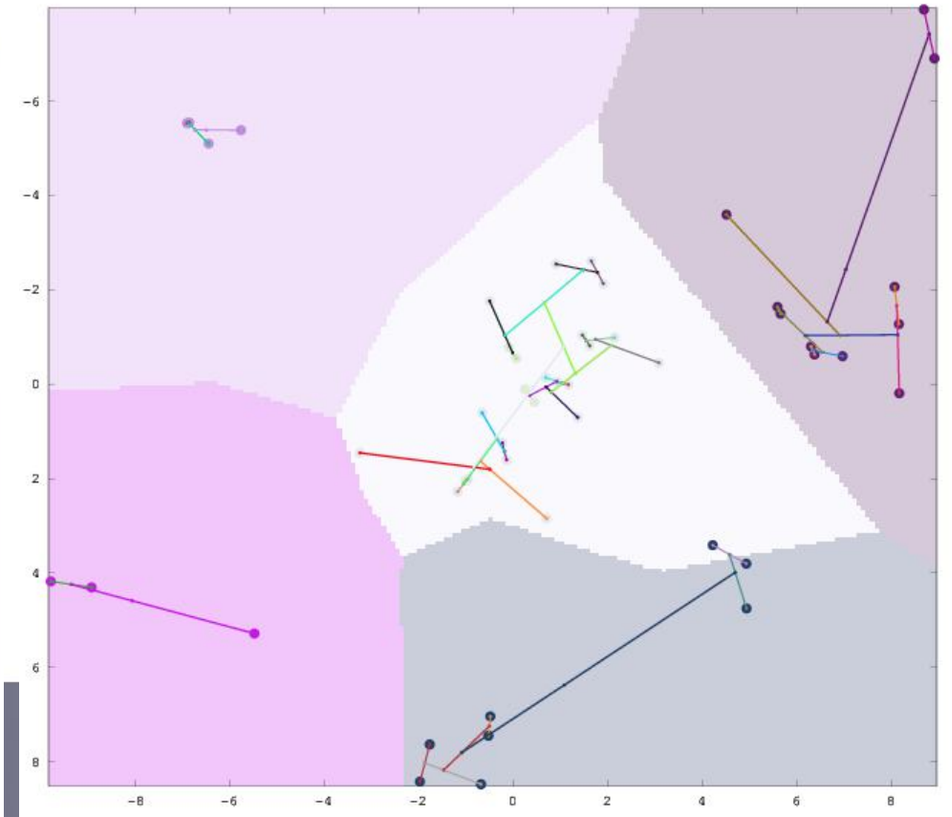


single-linkage

average linkage

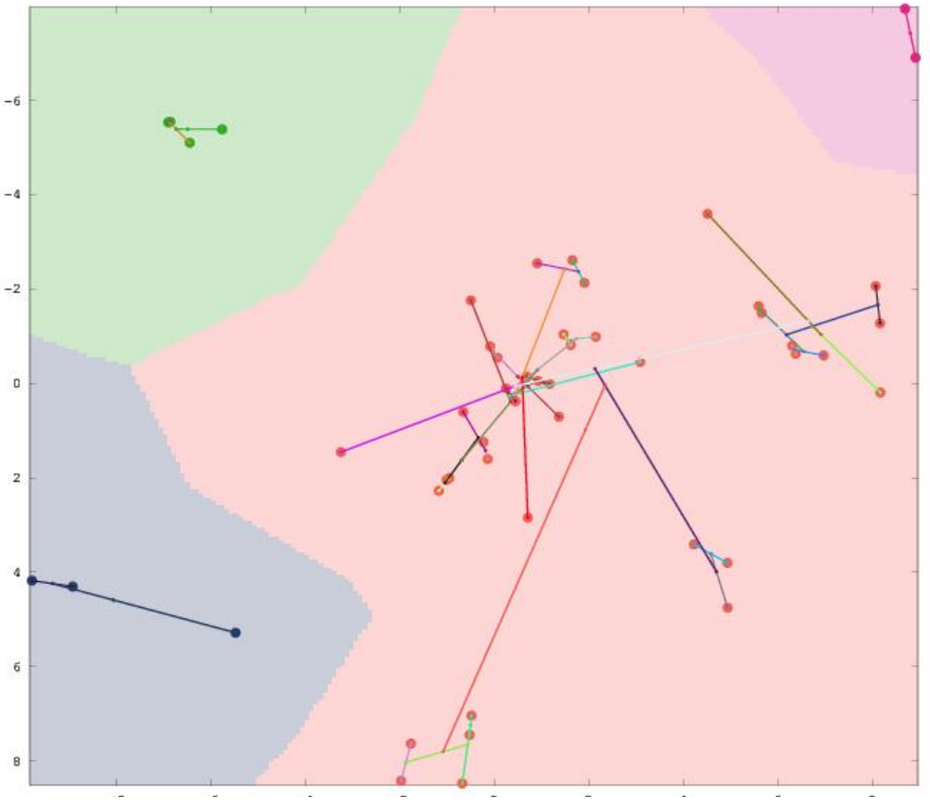


complete linkage



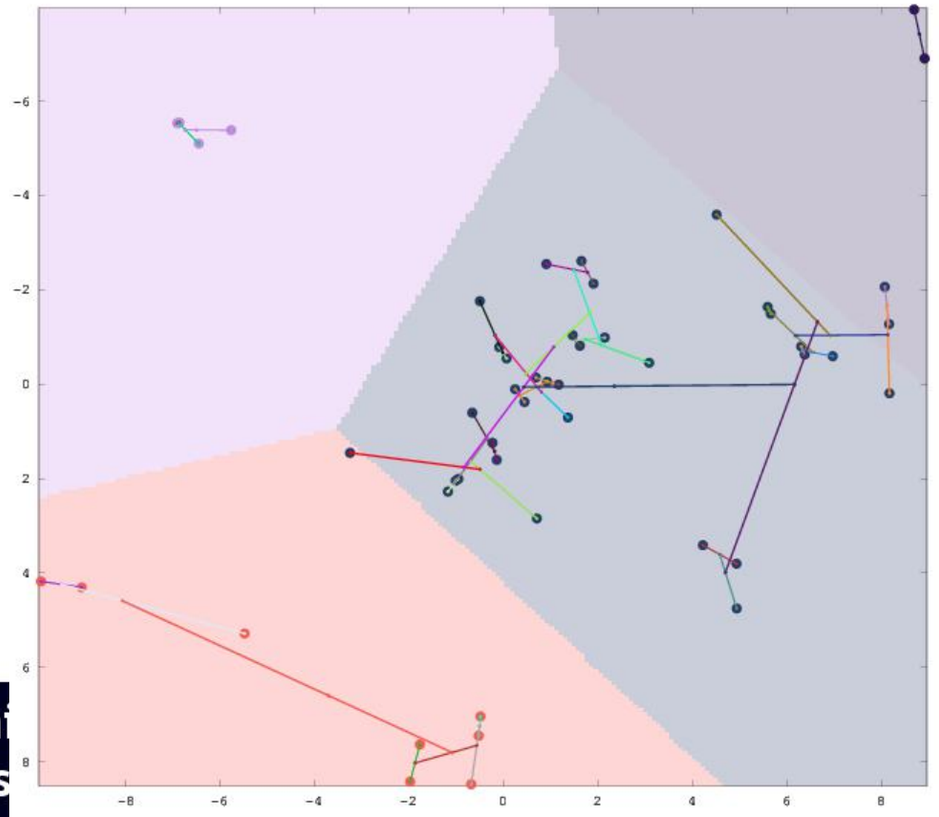
Example

4 clusters

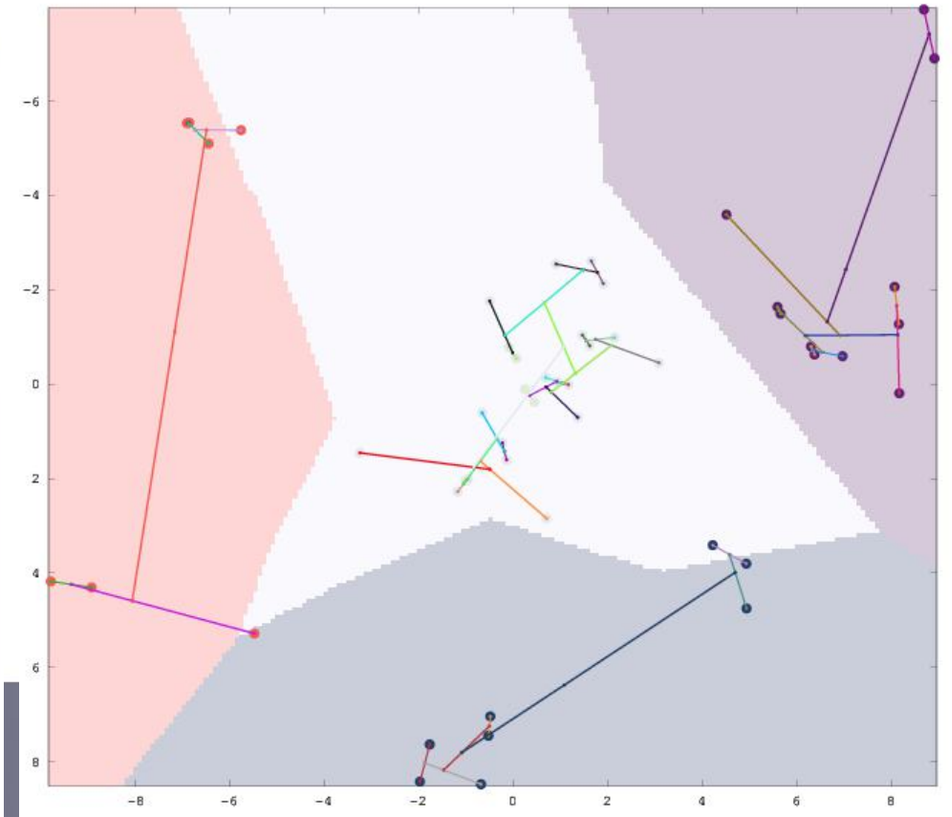


single-linkage

average linkage



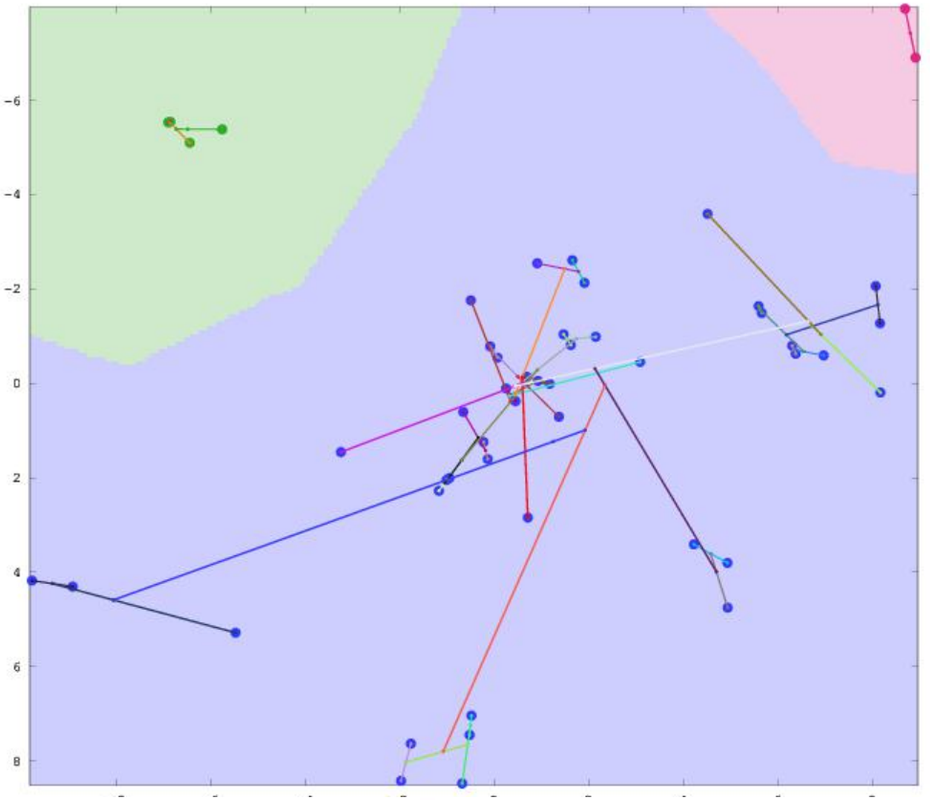
complete linkage



Example

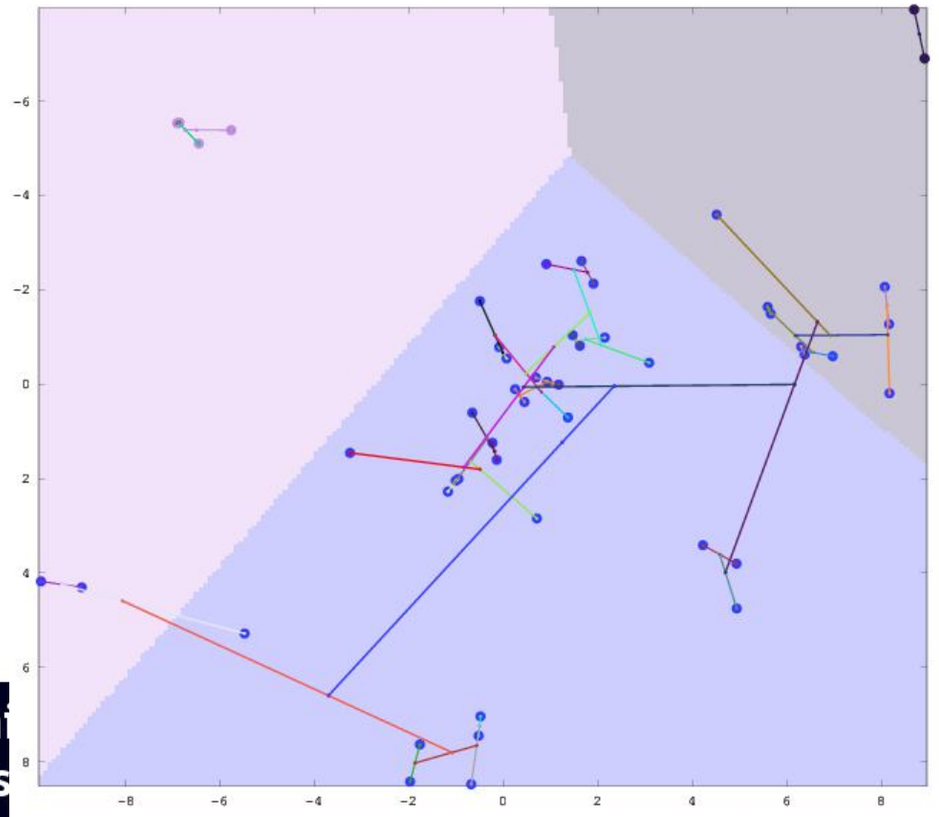
3 clusters

nonconvex Voronoi regions for single-linkage and complete-linkage

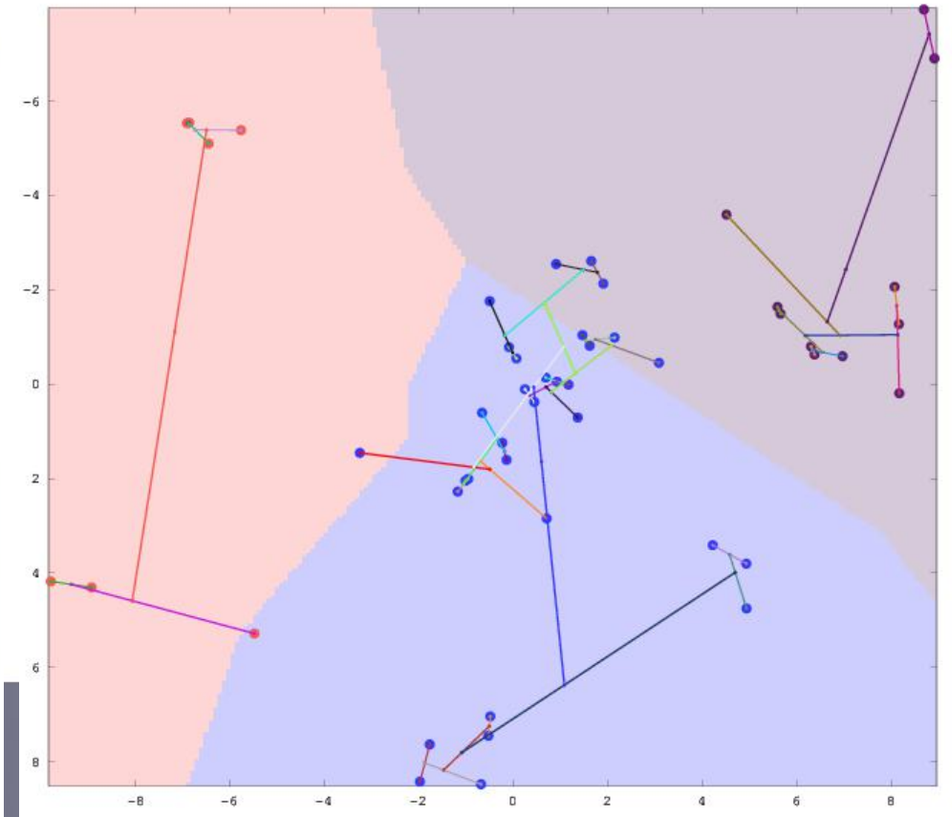


single-linkage

average linkage

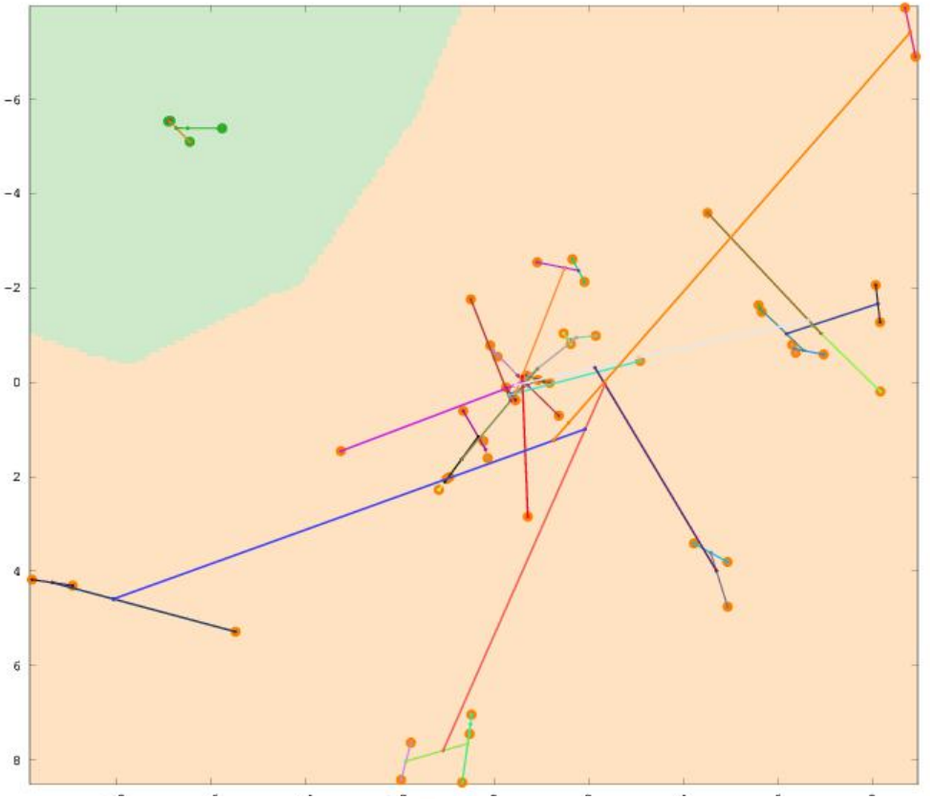


complete linkage



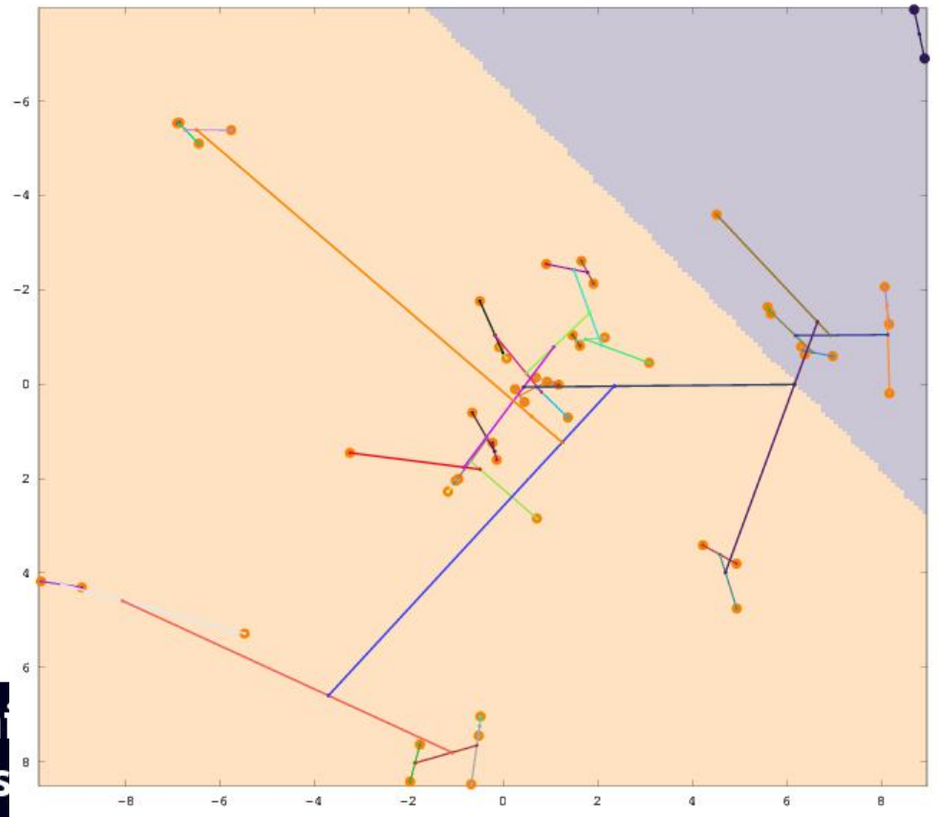
Example

2 clusters

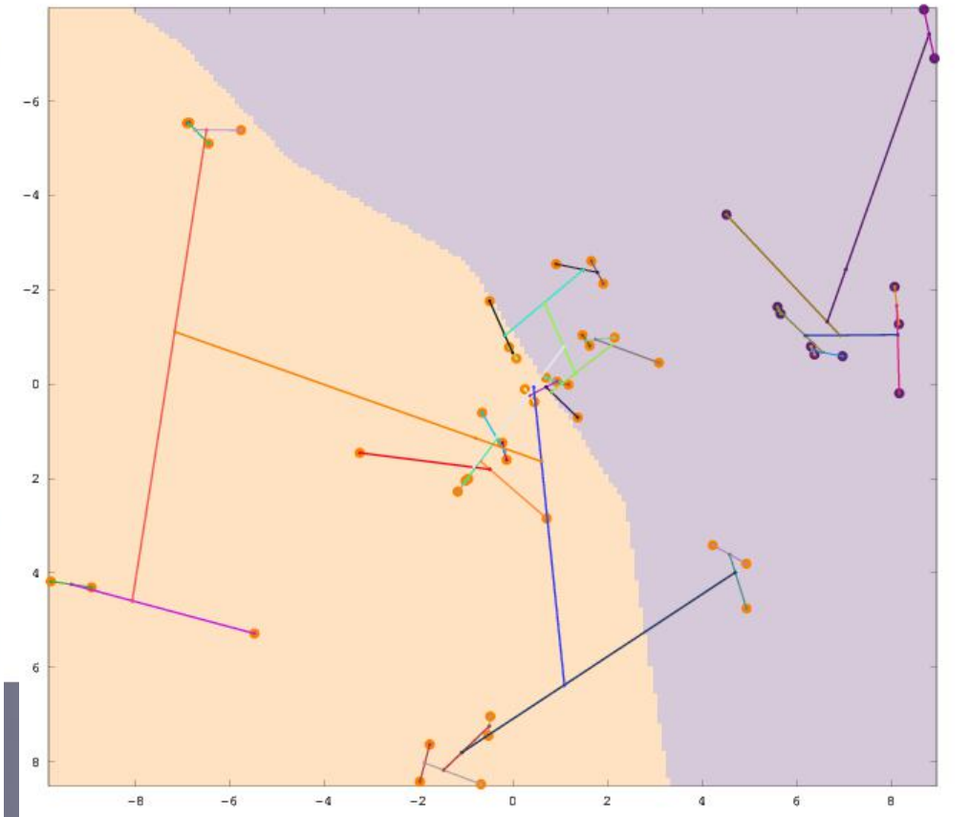


single-linkage

average linkage



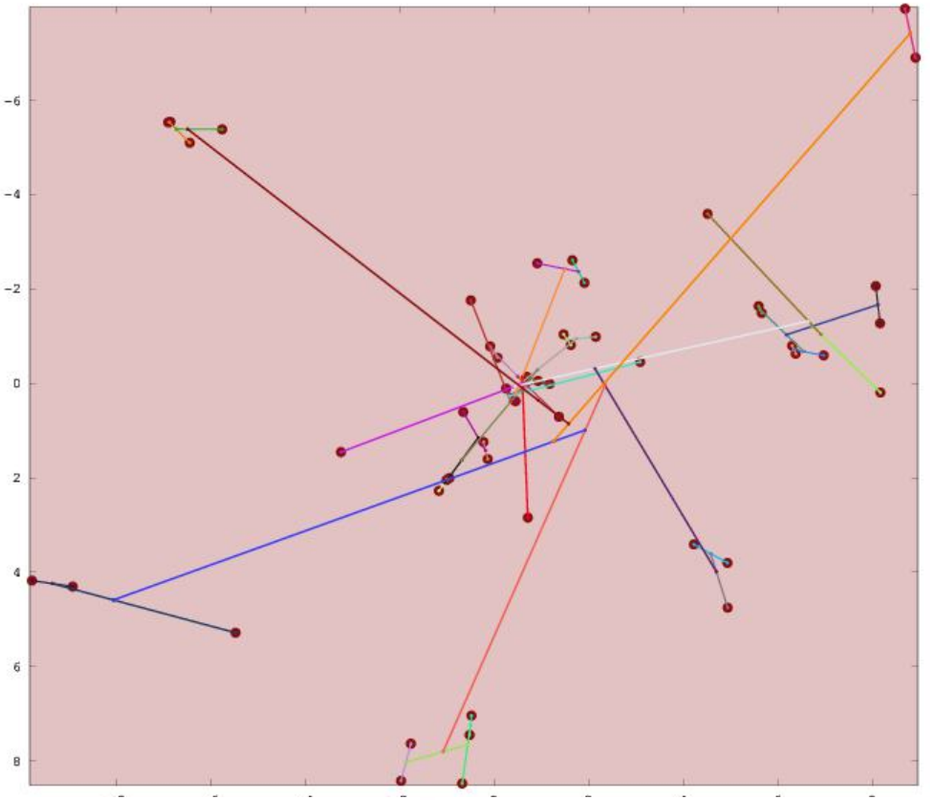
complete linkage



Example

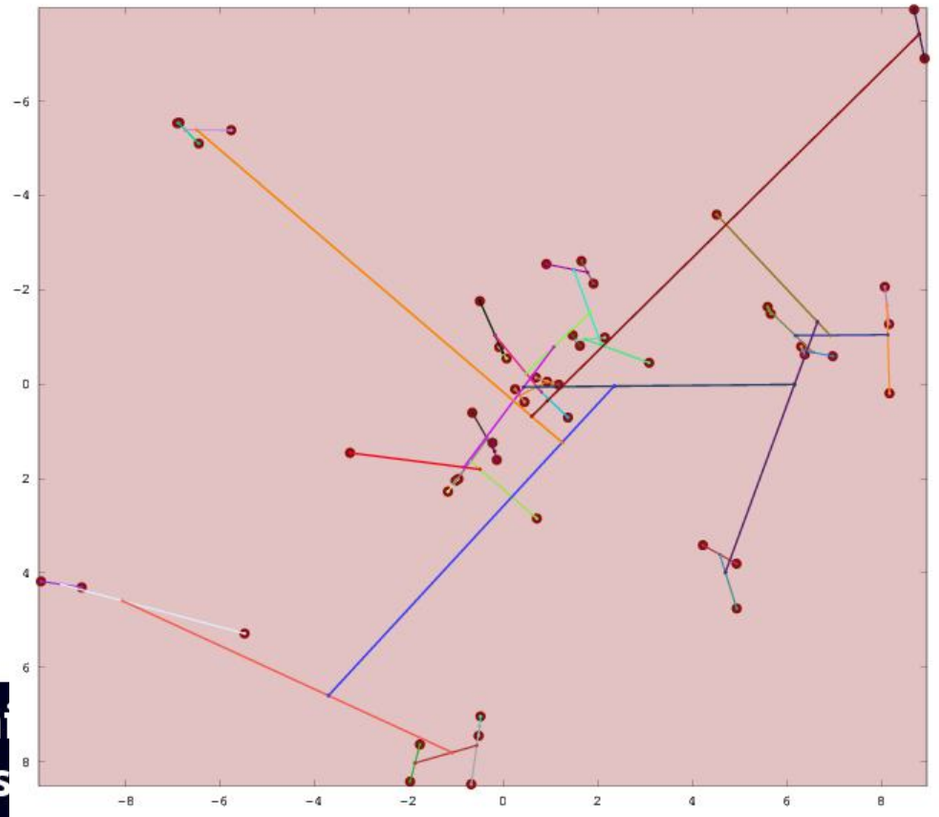
1 cluster

trees show the different merging paths of the algorithms

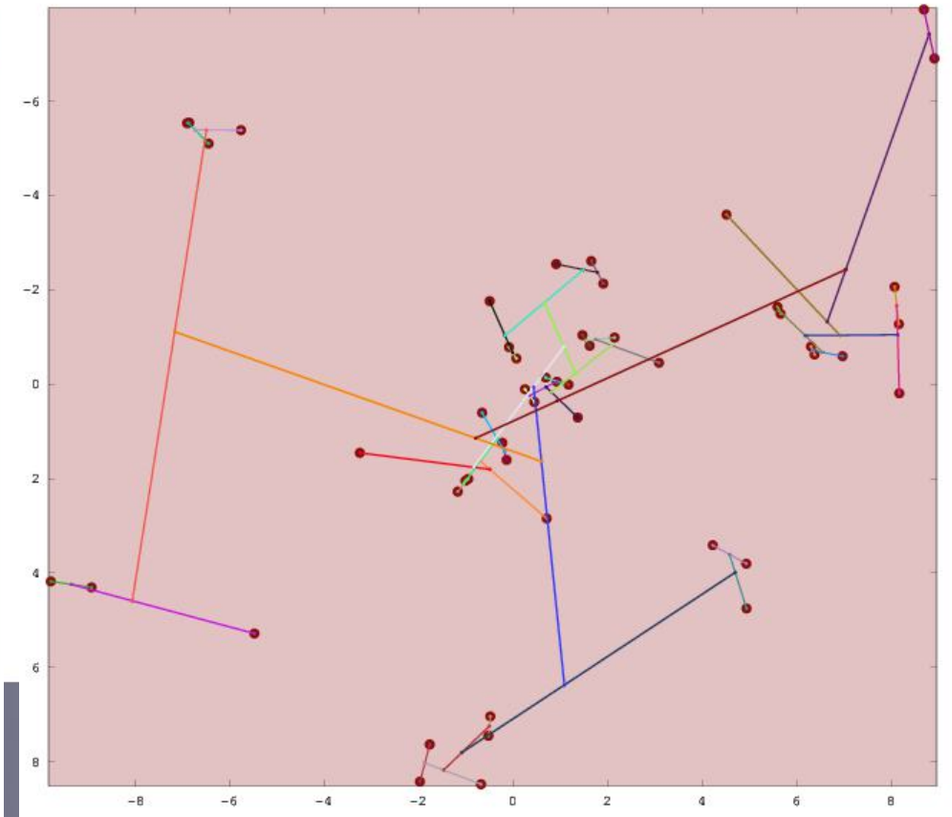


single-linkage

average linkage



complete linkage



Observations

- All three methods produce a hierarchical partitioning clustering but make different choices
- Single-linkage is more ”greedy” and can yield large clusters early on
- Single-linkage iterations result in a union of two previous Voronoi regions. The other two methods change the shape of Voronoi regions in each merge.
- Average- and complete-linkage can yield Voronoi regions that do not contain all training points of that cluster.