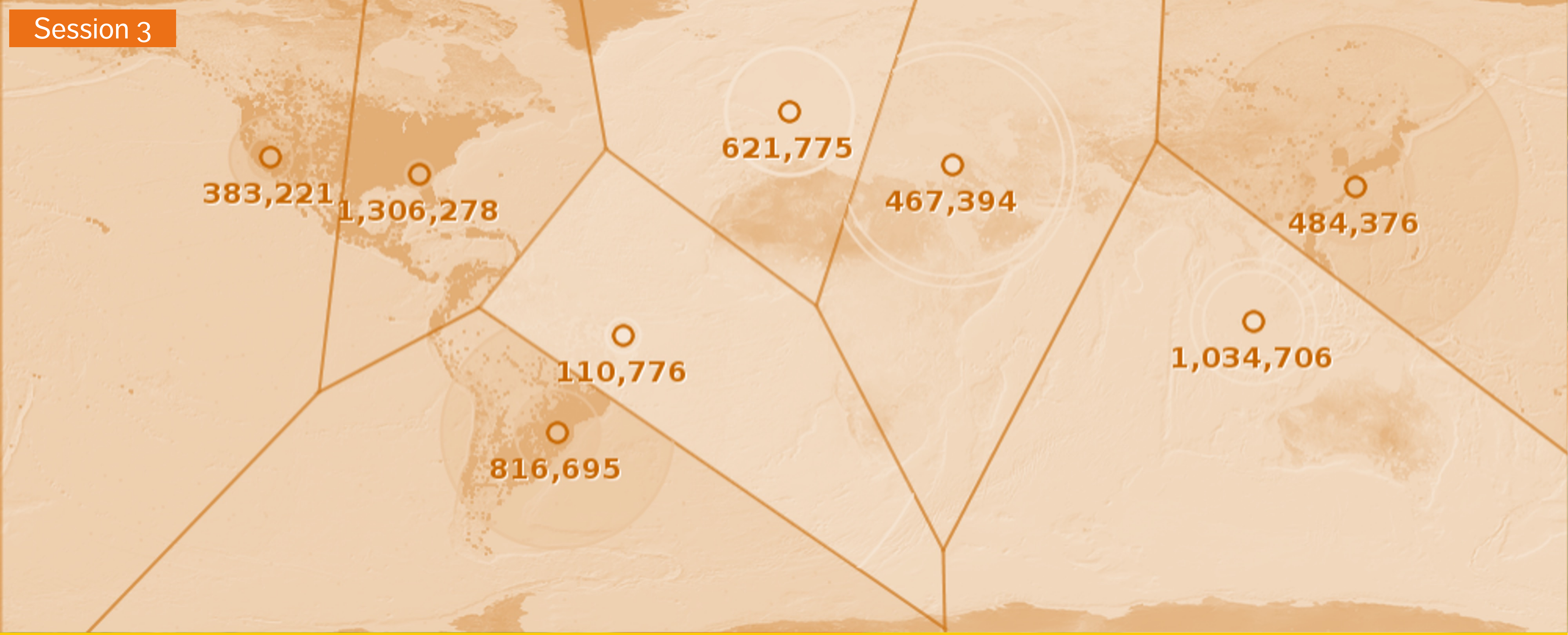


8. k-Means and Other Algorithms

Clustering Algorithms

k-Means
- classical (and over-used) model
- assumptions made about the shape of clusters

Hierarchical Clustering
- easy to interpret, deterministic

Cluster Ensembles

Latent Dirichlet Allocation
- used for topic modeling

Expectation Maximization

Clustering Algorithms

Balanced Iterative Reducing and Clustering using Hierarchies

Density-Based Spatial Clustering of Applications with Noise

- graph-based

Affinity Propagation

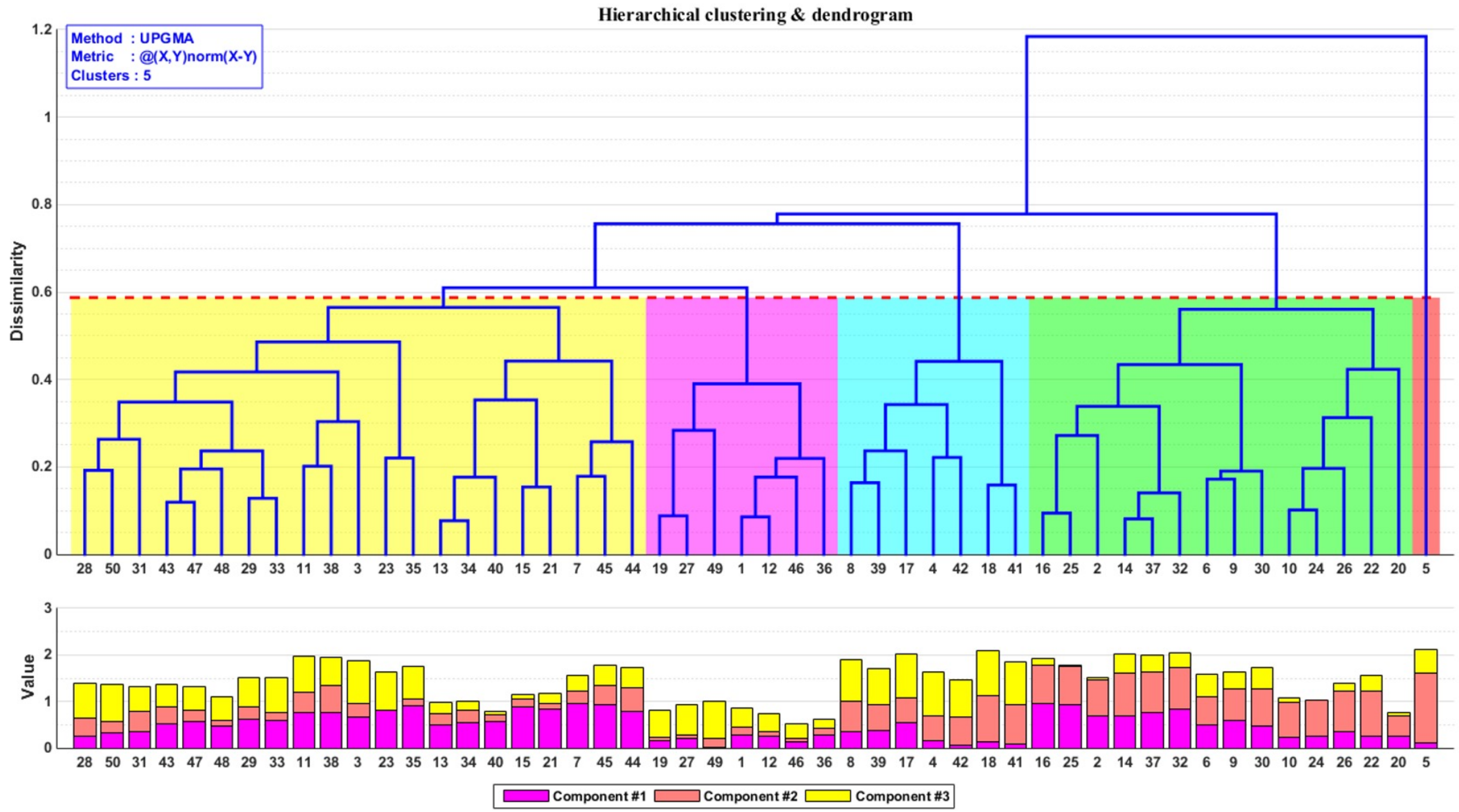
- selects the optimal number of clusters automatically

Spectral Clustering

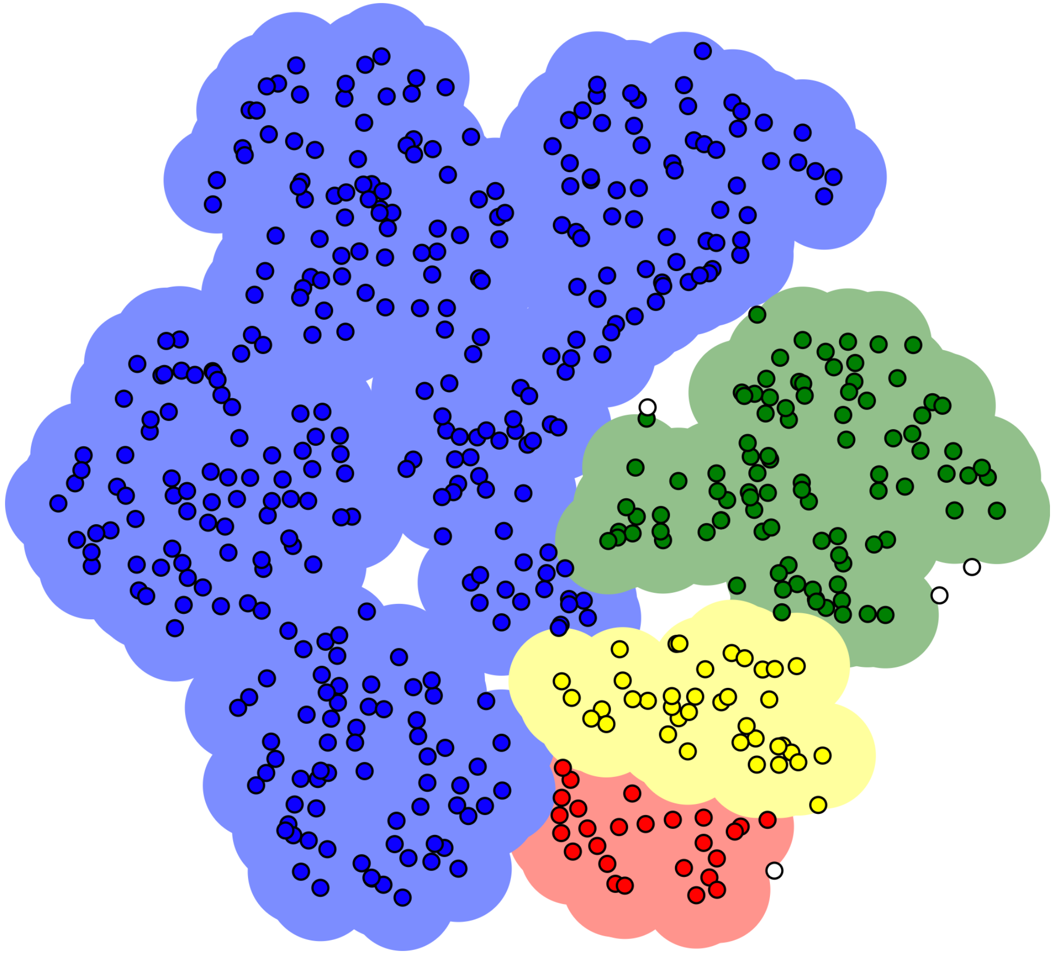
- recognizes non-blob clusters

Fuzzy Clustering

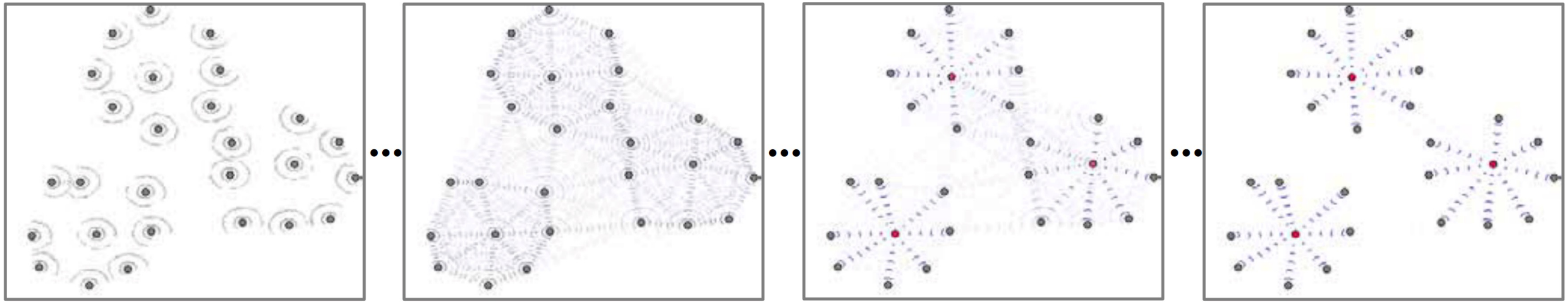
Hierarchical Clustering



DBSCAN



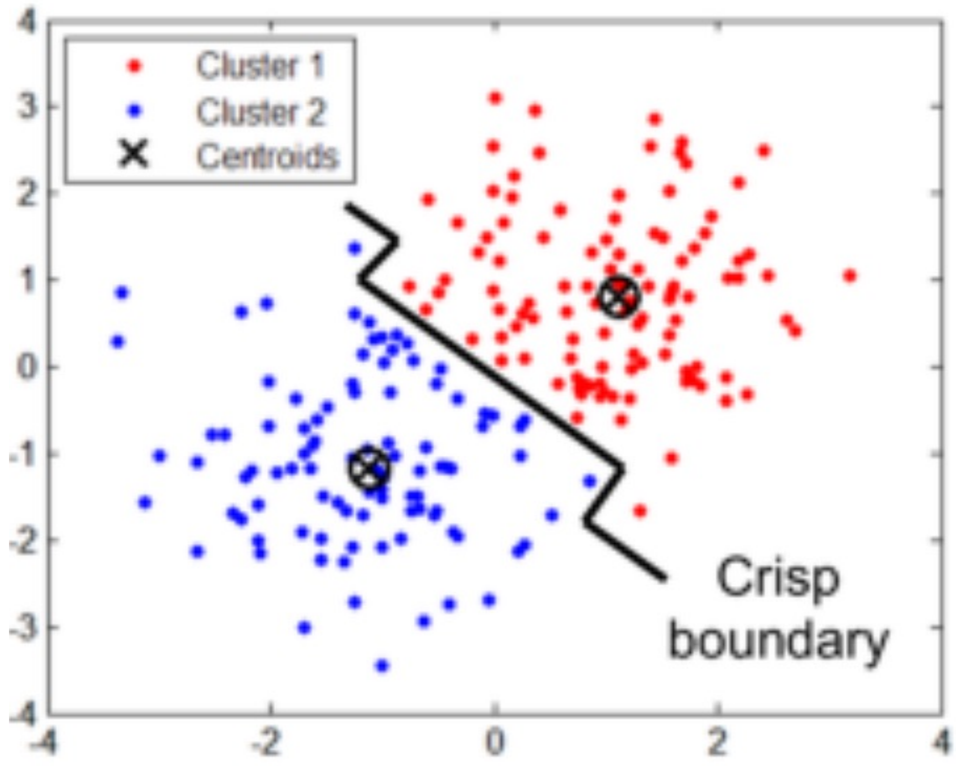
Affinity Propagation

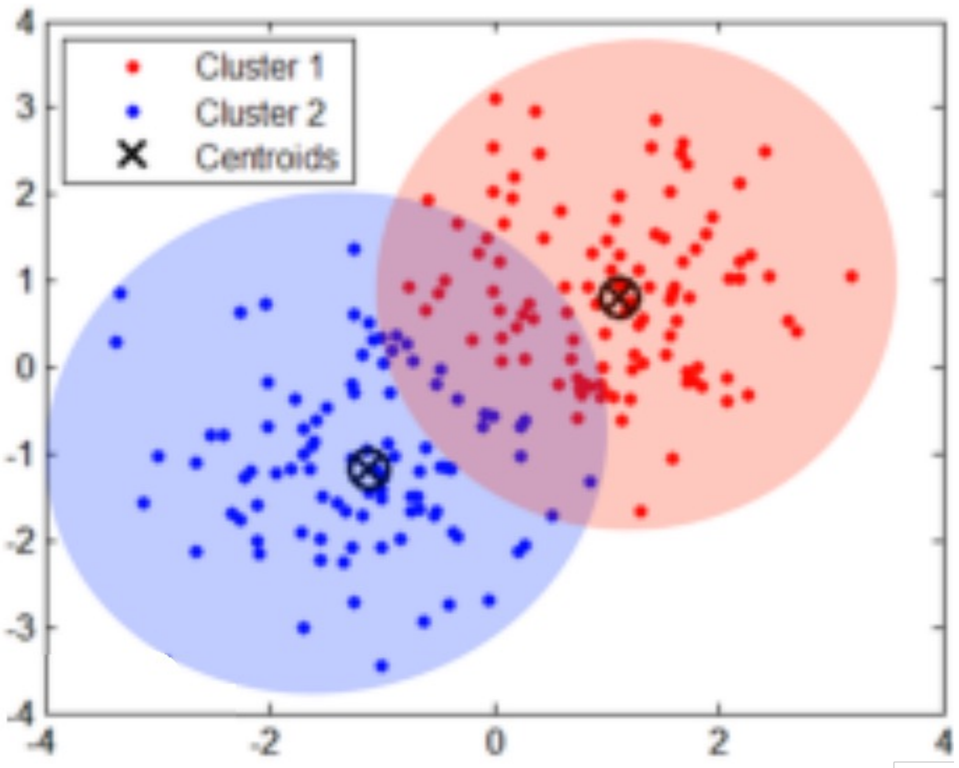


initialization | iteration X (uncertainty) | iteration Y (emergence) | final iteration (convergence)

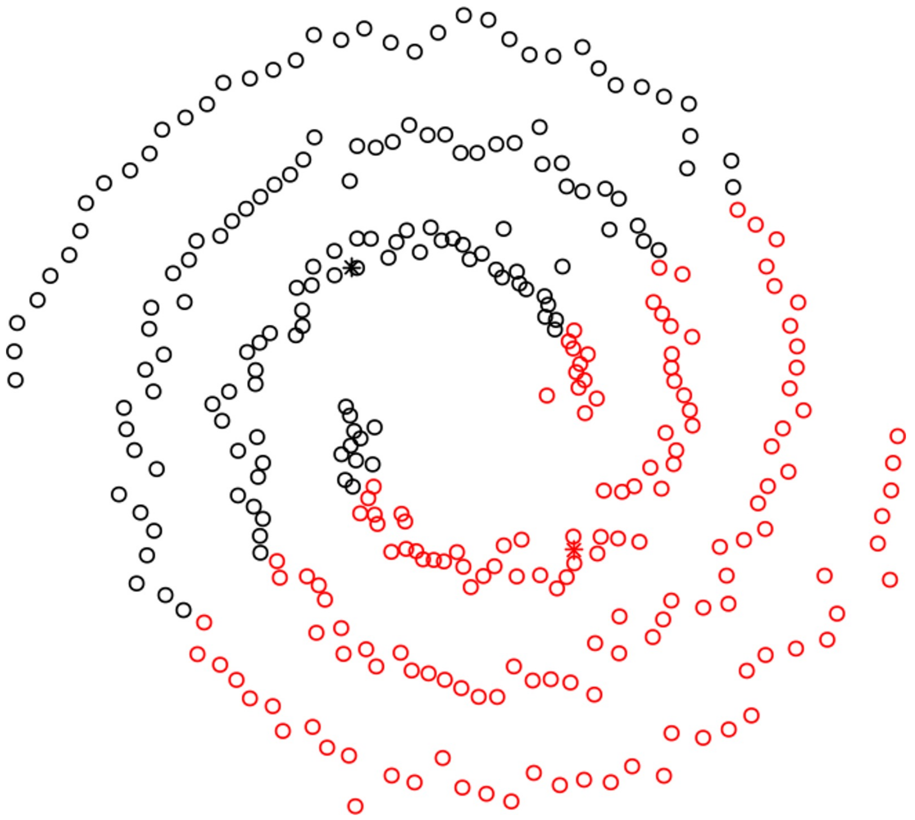
[A. Bagnaro, F. Baltar, and G. Brownstein]

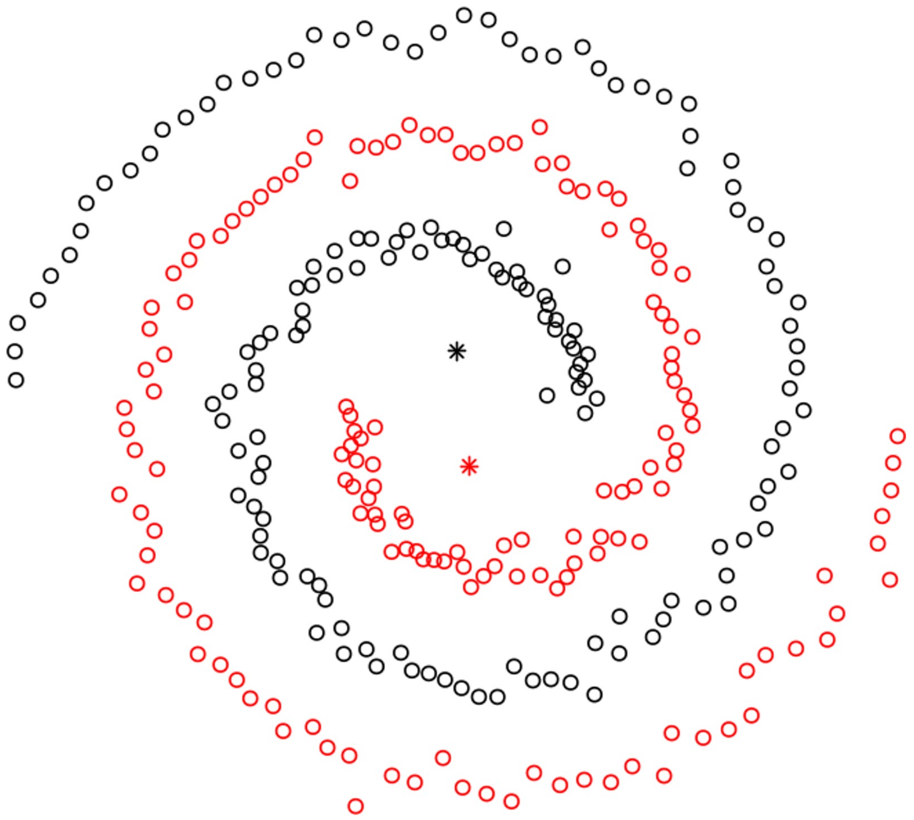
k-Means and Fuzzy c-Means





k-Means and Spectral Clustering

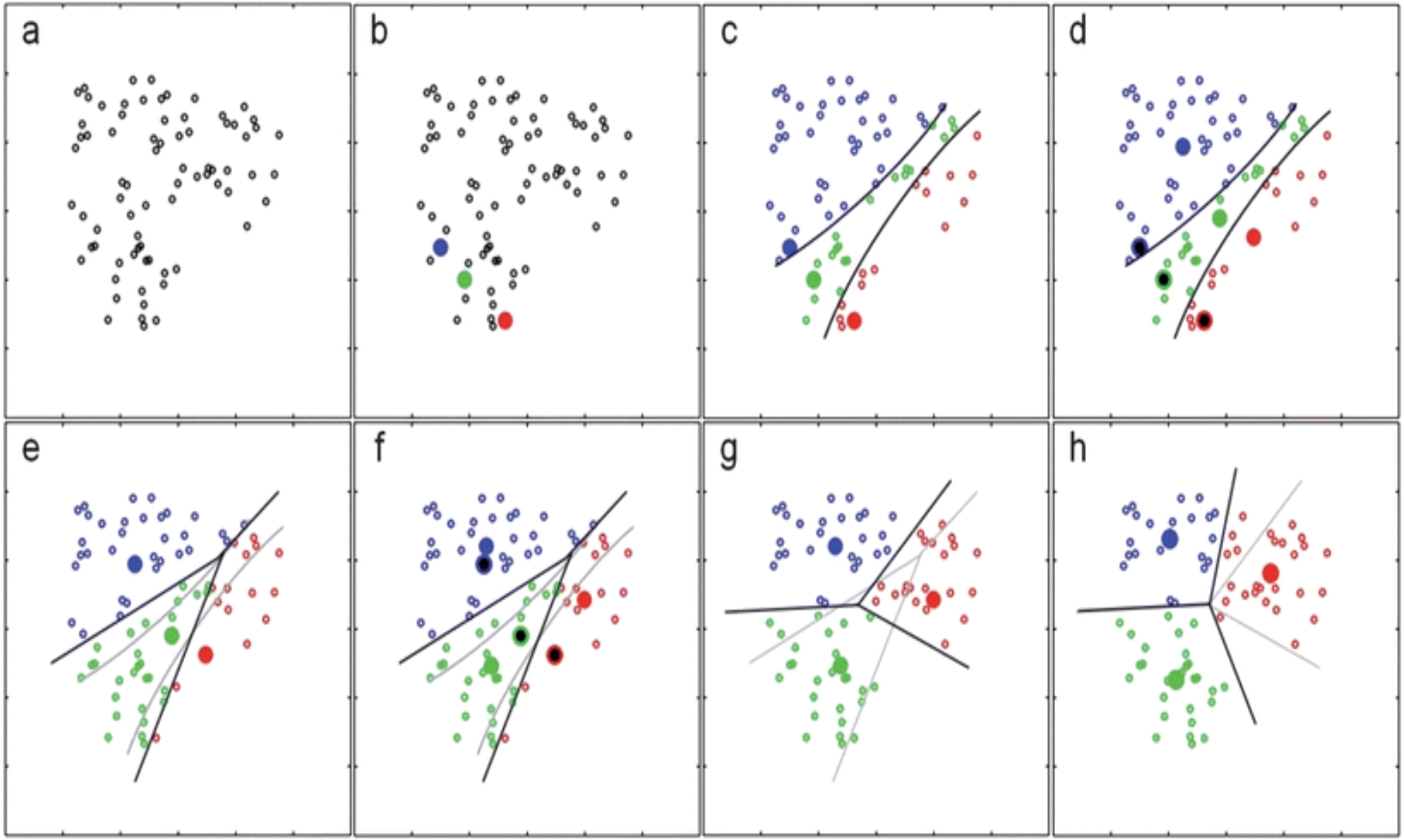




k-Means Algorithm

1. Select the desired **number of clusters**, say k
2. Randomly choose k instances as initial **cluster centres**
3. Calculate the **distance** from each observation to each centre
4. Place each instance in the cluster whose centre it is **nearest** to
5. Compute the **centroid** for each cluster
6. Repeat steps 3 – 5 with the new centroids
7. Repeat step 6 until the clusters are **stable**

[A.H. Du]



k-Means Strengths

Easy to implement (without having to actually compute pairwise distances).

- extremely common as a consequence
- elegant and simple

In many contexts, k-means is a **natural** way to look at grouping observations.

Helps provide a **basic understanding of the data structure** in a first pass.

k-Means Limitations

Data points can only be assigned to **one** cluster

- this can lead to overfitting
- robust solution: consider the probability of belonging to each cluster

Underlying clusters are assumed to be **blob-shaped**

- k-means will fail to produce useful clusters if that assumption is not met in practice

Clusters are assumed to be separate (discrete)

- k-means does not allow for **overlapping** or **hierarchical** groupings

k-Means Limitations

There are many ways to pick the **optimal number** of clusters k.

One problem is that the algorithm is stochastic: different initial configurations may yield **different outcomes**, which may yield a different optimal number.

It may also depend on the **size** of data, the choice of **distance**, the choice of **cluster quality metric**, etc.

Suggested Reading

k-Means and Other Algorithms

Data Understanding, Data Analysis, Data Science
Machine Learning 101

Clustering

- Clustering Algorithms
- k-Means
- Toy Example: Iris Dataset

R Examples

- Clustering: Iris Dataset

Spotlight on Clustering

*Simple Clustering Methods (advanced)
*Advanced Clustering Approaches (advanced)

Exercises

k-Means and Other Algorithms

1. Go over the iris clustering example found in DUDADS (see suggested reading). Repeat the process with the UniversalBank dataset (you may wish to visualize the dataset first) in order to build a clustering scheme. Determine the optimal number of clusters using the Davies-Bouldin index.