Clustering

Why cluster?

- Information extraction
- Creating hierarchy of objects
- Data simplification set for further analysis
- Data compression
- Anomaly detection
- Feature generation

Clustering is unsupervised

- No specific task
- No specific number of cluster
- No specific quality criterion

K-Means

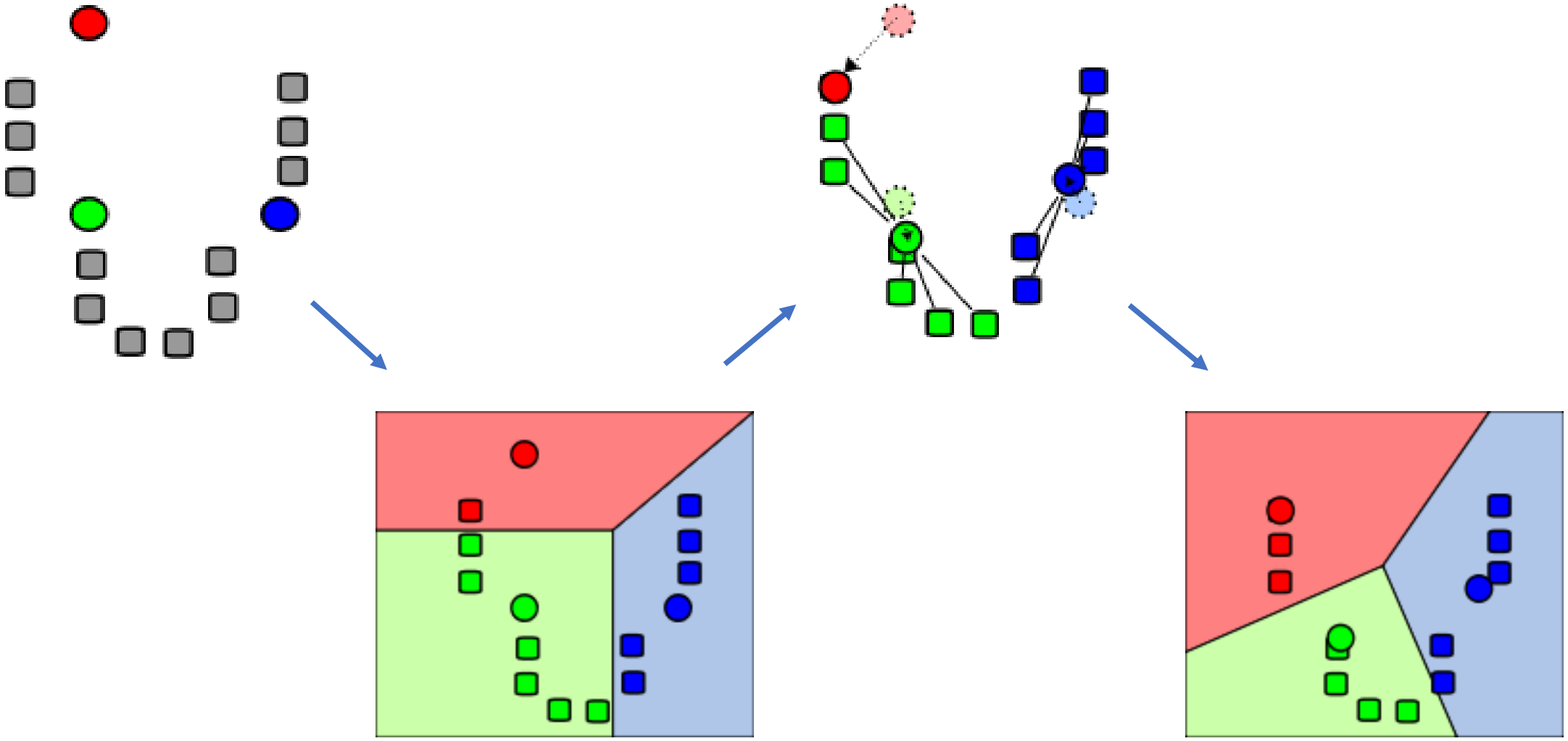
The number of clusters is predefined.

μ_i is the center of C_i cluster.

The aim is to minimize the variance:

$$\sum_{x_j \in X} \min_{\mu_i} \left|\left| x_j - \mu_i \right|\right|_2^2$$

K-Means



K-Means Algorithm

1. Initialize clusters at random.

2. Assign datapoint to cluster with the closest center.

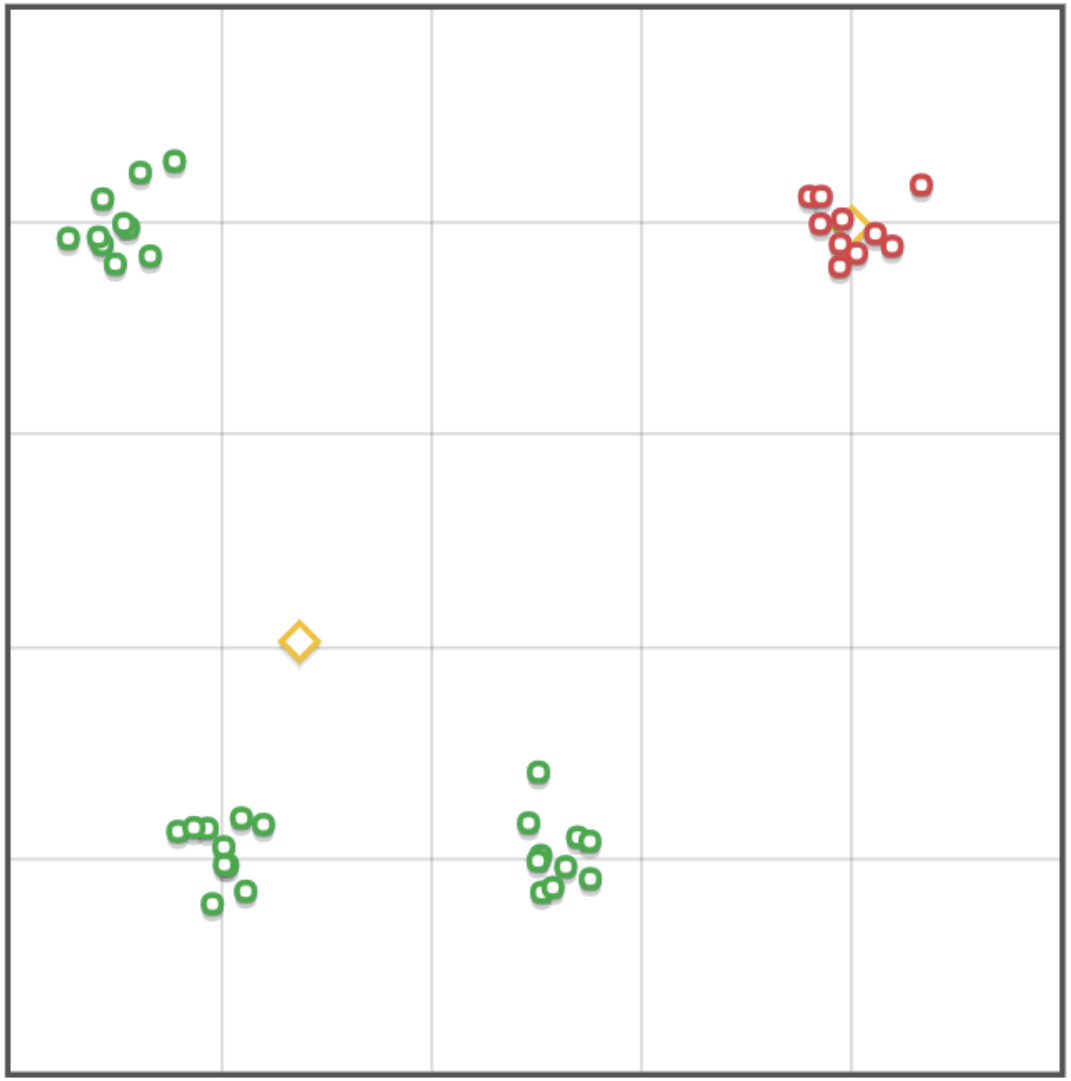
3. Shift cluster's center to the center of mass:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

4. Repeat 2-3. until convergence.

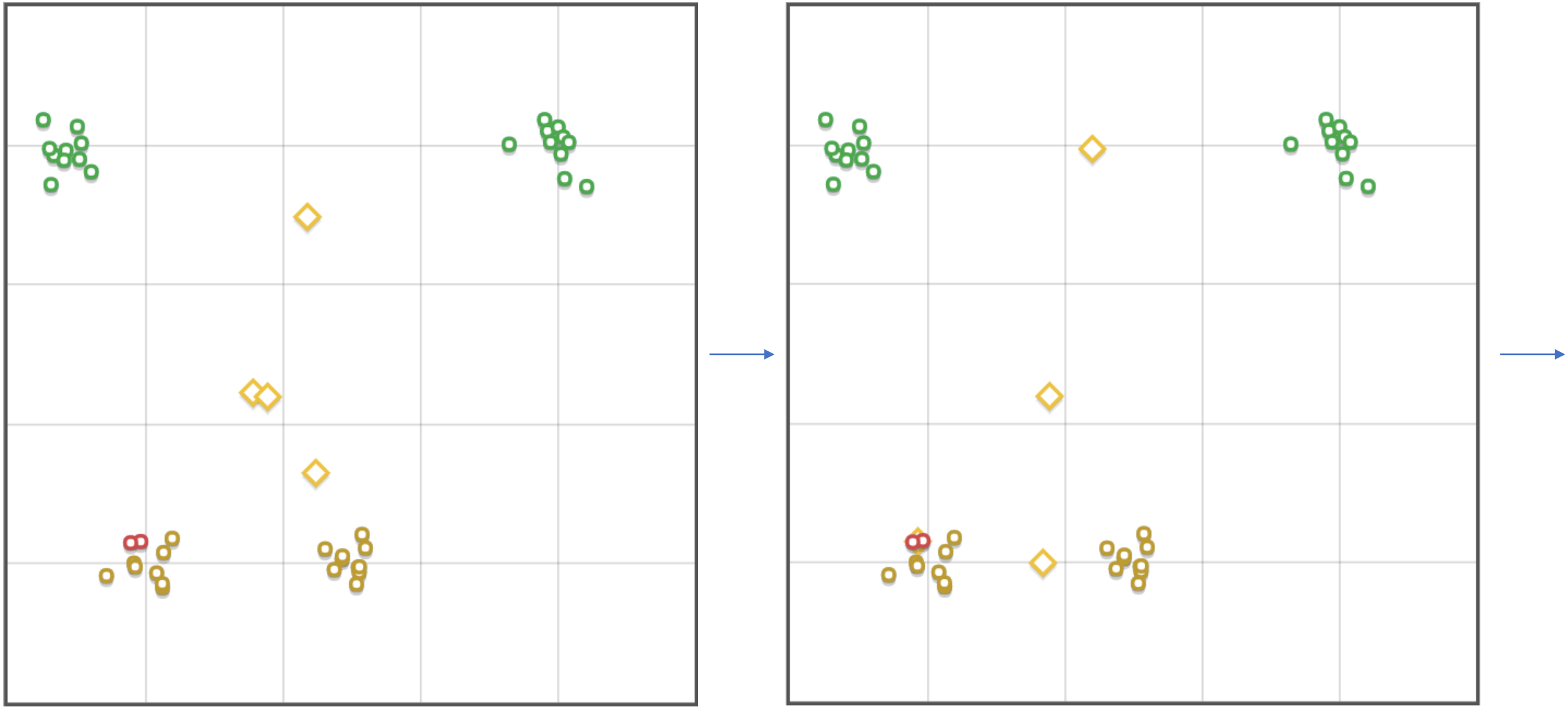
Problems with K-Means

Selecting the correct number of clusters



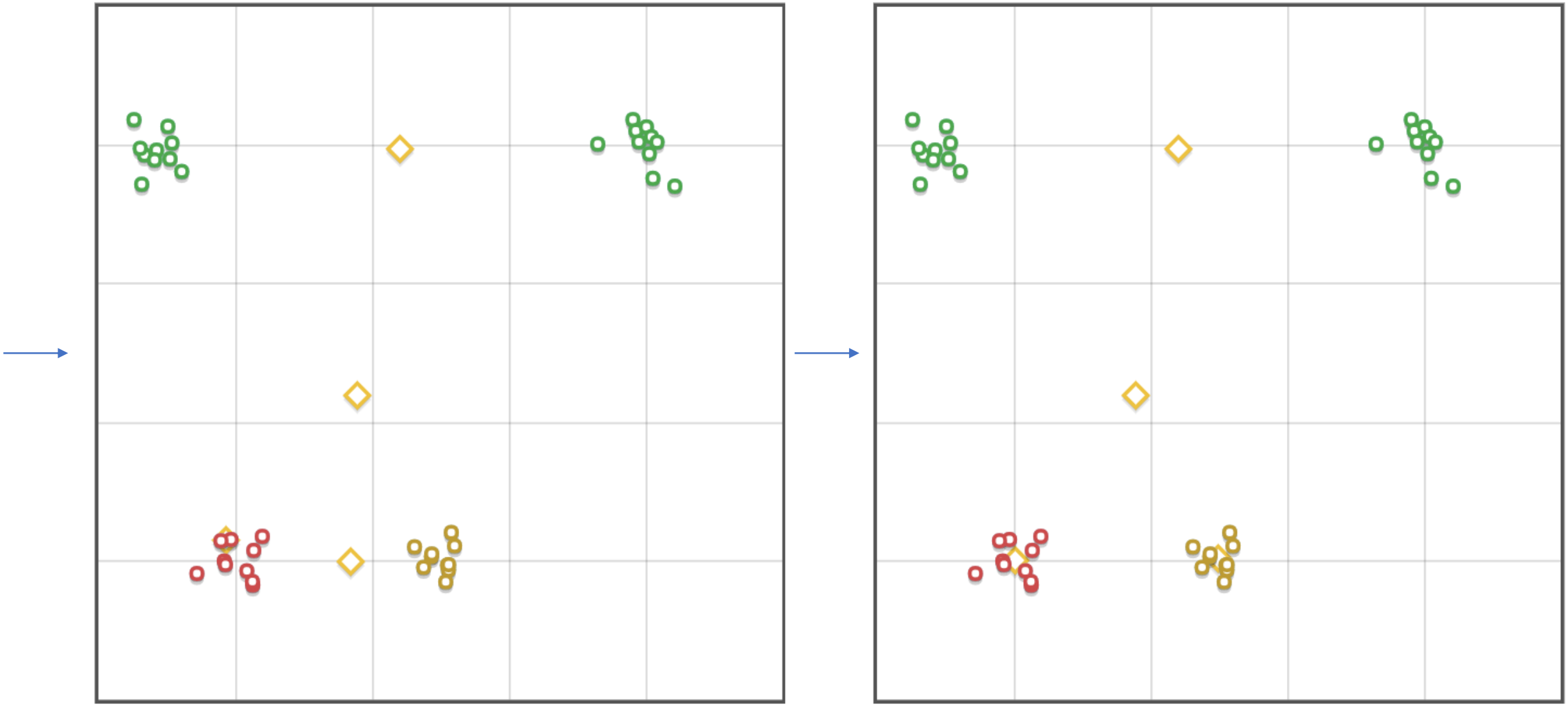
Problems with K-Means

Initializing the centers



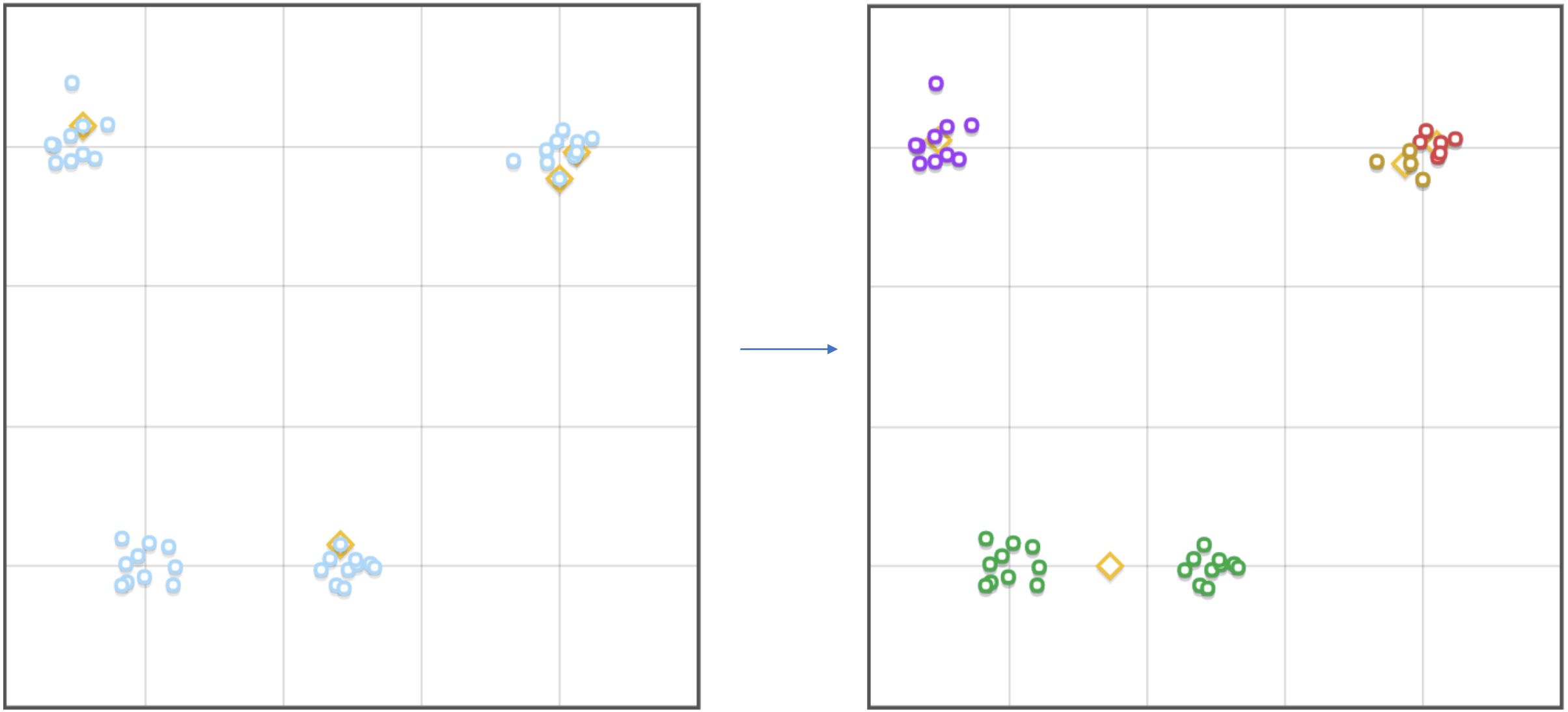
Problems with K-Means

Initializing the centers



Problems with K-Means

Initializing the centers



k-means++

1) Select the first center randomly among the data points.

2) For every x calculate the distance to the nearest center - $\mathrm{M}(\mathrm{x})$.

3) Select the next center with probability that is proportional to $\mathrm{M}^2(\mathrm{x})$.

4) Repeat 2-3 $(k-1)$ times.

5) Launch k-means.

Mean Shift

The number of clusters is not predefined.

$$\mu_i^0 = x_i$$

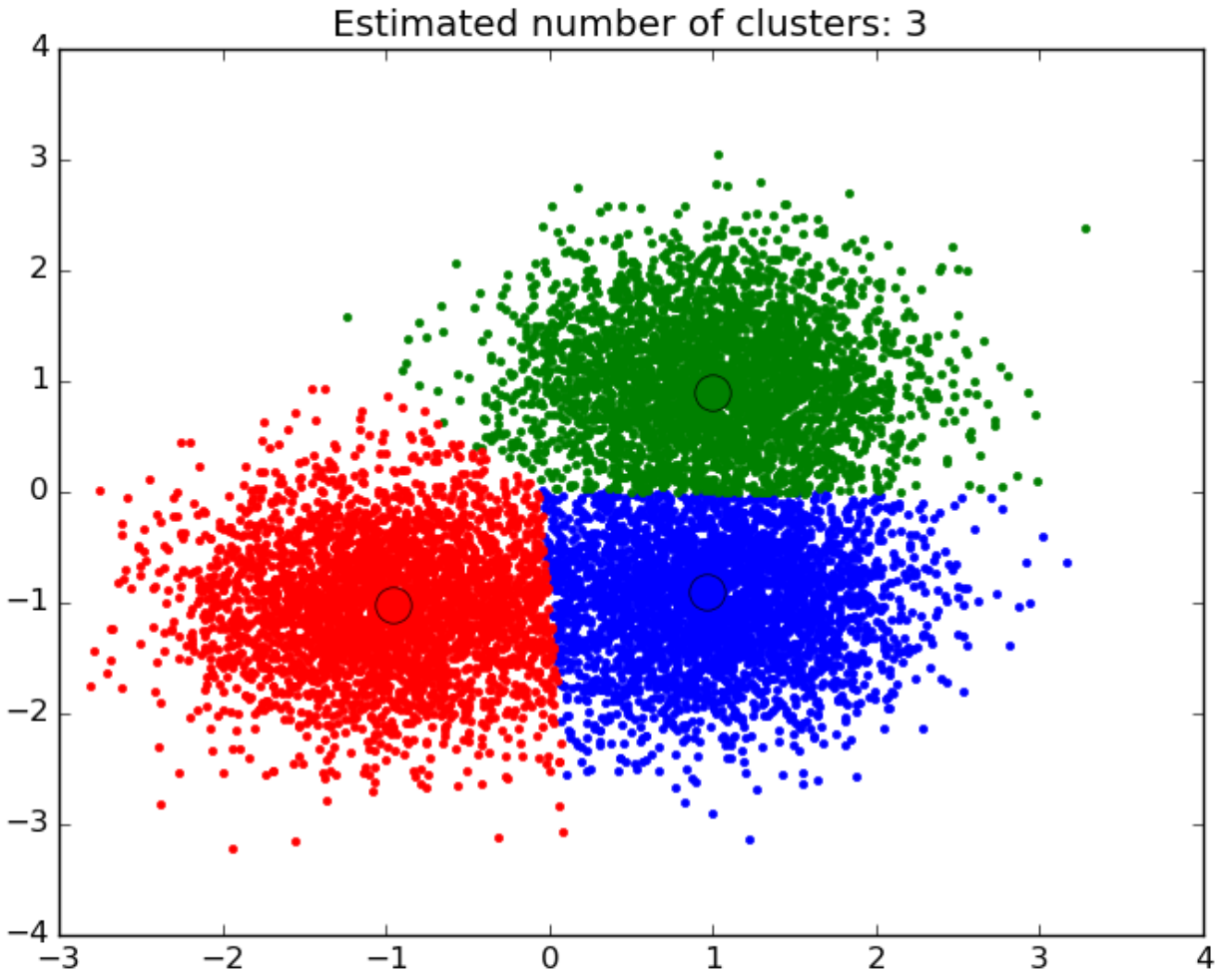
$$\mu_i^{t+1} = \frac{\sum_{x_j \in N(\mu_i^t)} RBF(x_j - \mu_i^t) x_j}{\sum_{x_j \in N(\mu_i^t)} RBF(x_j - \mu_i^t)},$$

$N(\mu_i^t)$ – is some neighborhood of μ_i.

$$RBF^*(x_j - \mu_i^t) = e^{-c\left\|x_j - \mu_i^t\right\|_2^2}$$

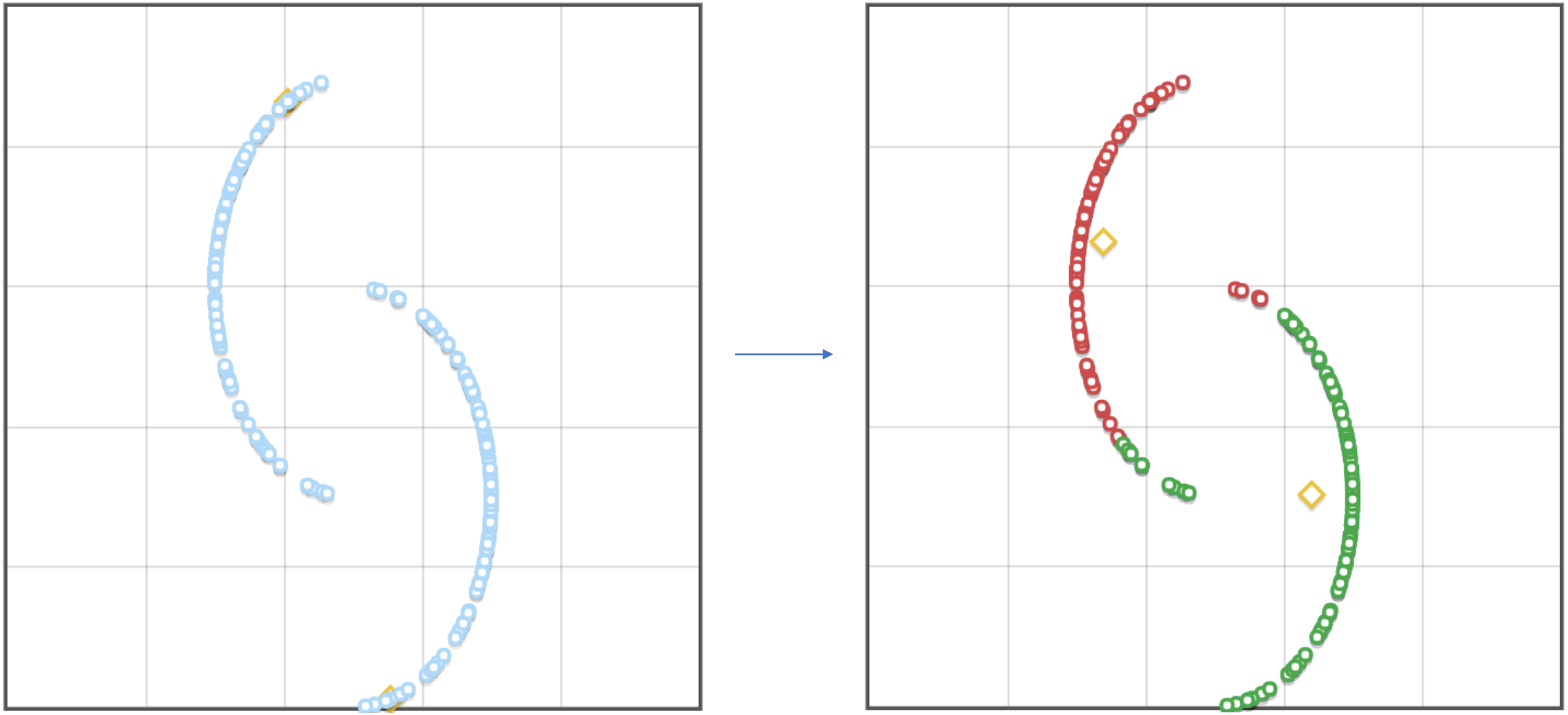
Repeat until convergence, then merge close centers.

*Radial Basis Function



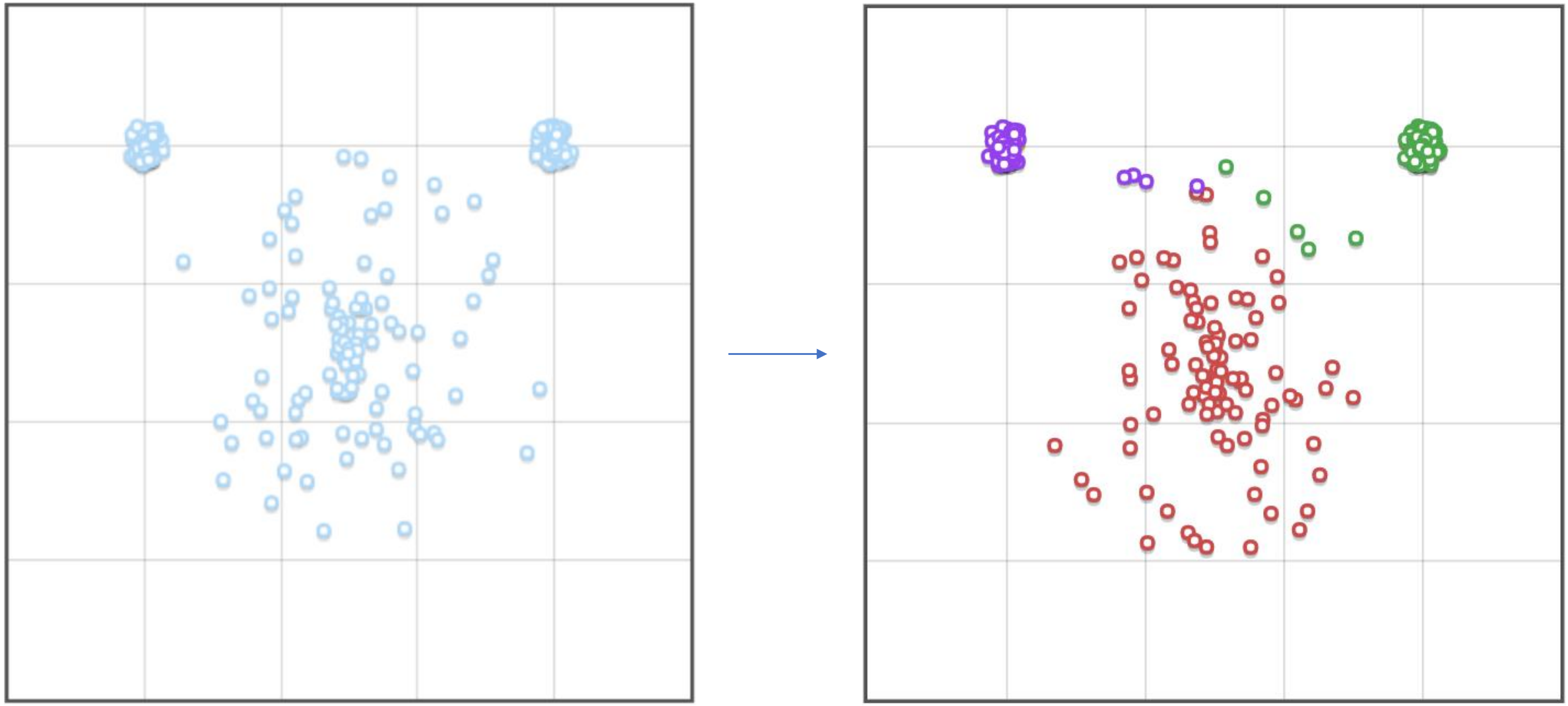
Problems with K-Means

Non-spherical clusters



Problems with K-Means

Clusters of different sizes

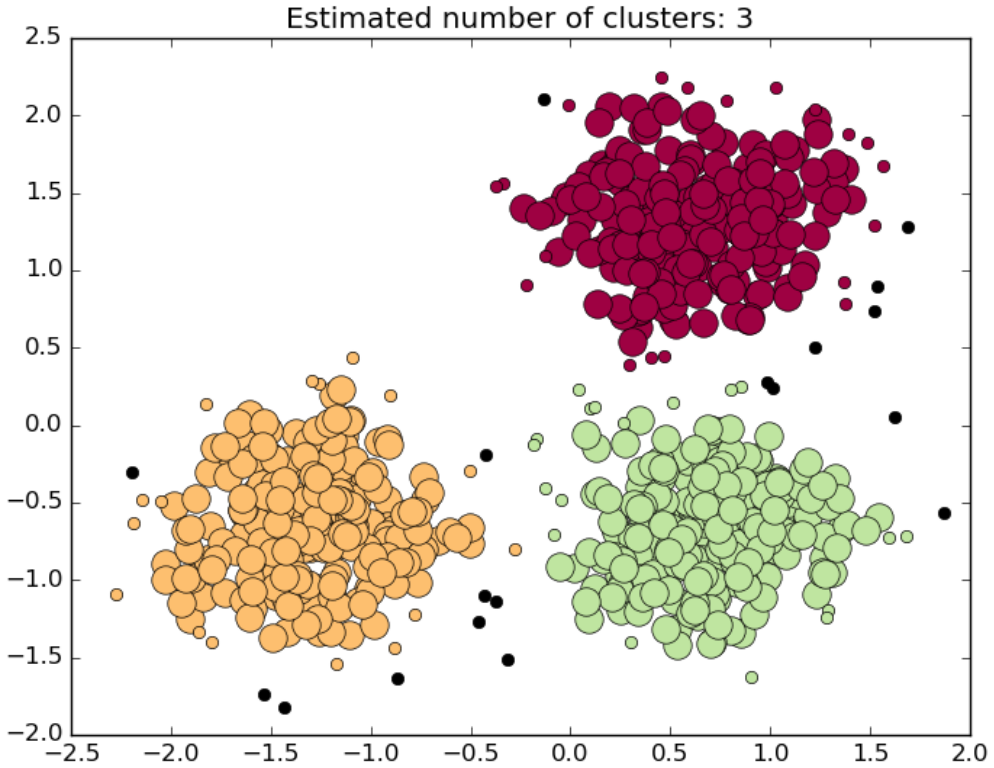


DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

The number of clusters is not predefined.

Define core samples – data points that contain at least m data points in their ϵ-neighborhood.

Merge core samples (if they are within ϵ of each other) and their neighborhoods.



Hierarchical clustering

Agglomerative Clustering

Hierarchical bottom-to-top clustering.

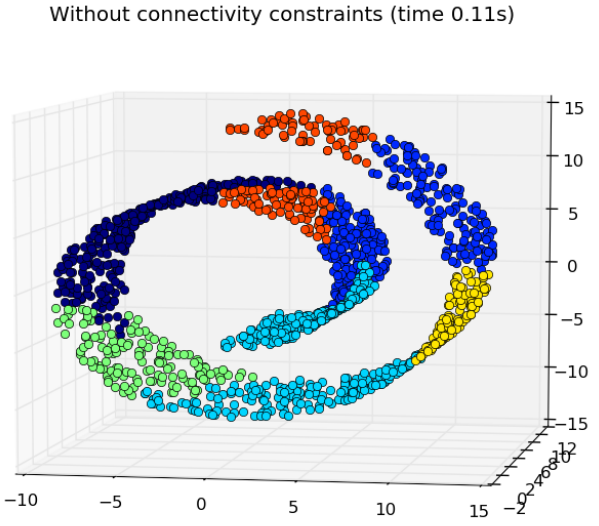
Initialize every point as its own cluster.

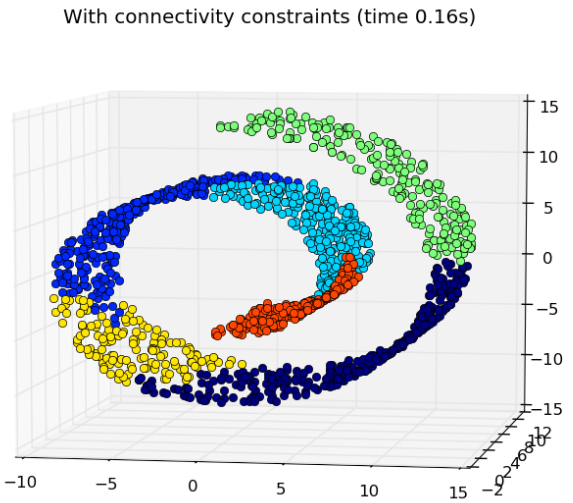
Three linkage strategies. Each strategy aims to minimize on of the following metrics:

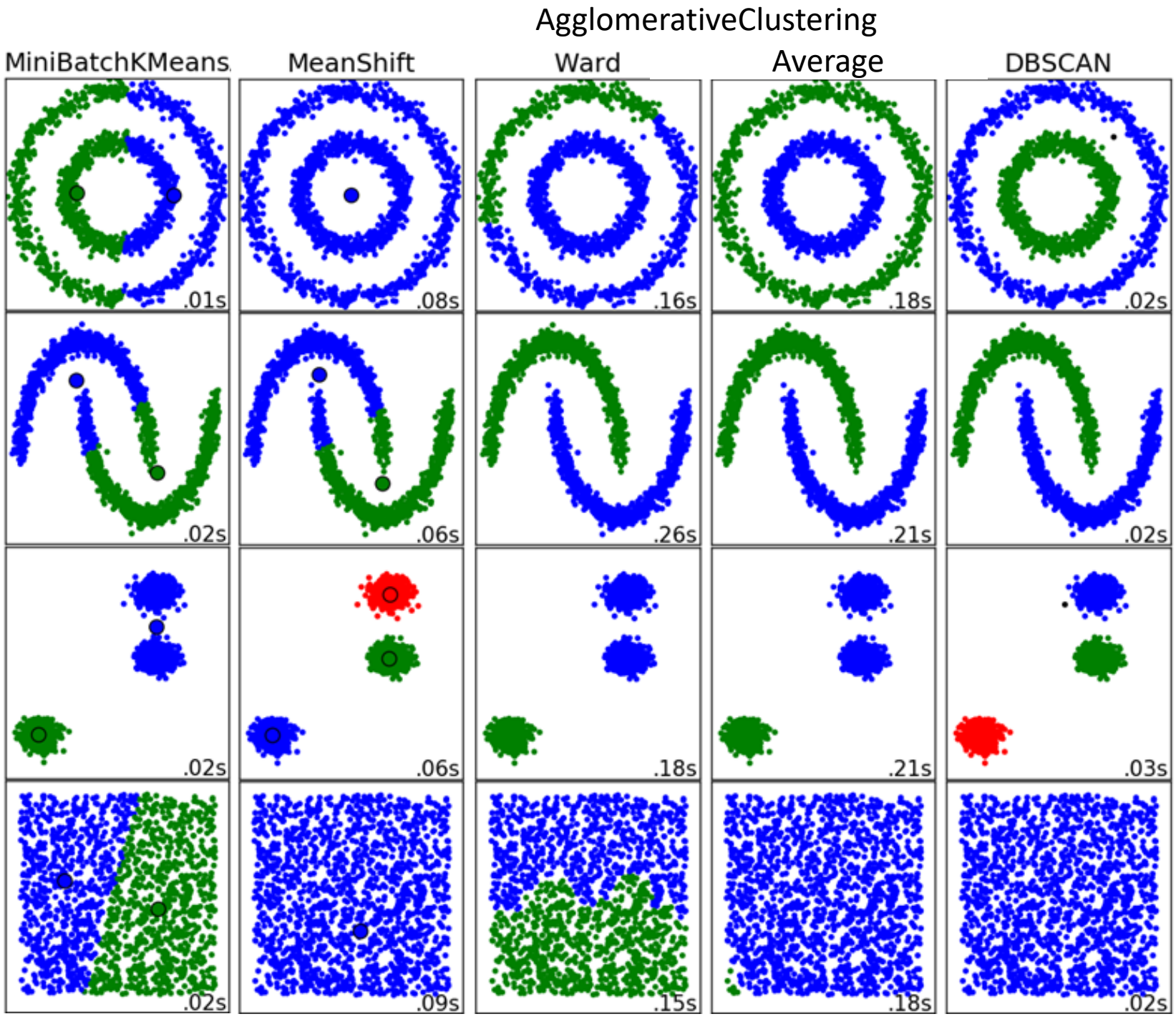
- ward – variance of joined clusters
- average – average distance between points in clusters
- maximum (complete) – maximum distance between points in clusters

Join until just one cluster remains.

As an additional condition you can add connectivity constraint – meaning we only join clusters if the minimum distance between them is less then some predetermined constant.






AgglomerativeClustering
MiniBatchKMeans
MeanShift
Ward
Average
DBSCAN
.01s
.08s
.16s
.18s
.02s
.02s
.06s
.26s
.21s
.02s
.02s
.06s
.18s
.21s
.03s
.02s
.09s
.15s
.18s
.02s

Clustering metrics

External:

- End metric
- Rand score
- Mutual information

- Homogeneity score:

$$\frac{1}{|D|}\sum_i \max_y |\mathrm{x}_j \in C_i, y_j = y|$$

Clustering metrics

Internal:

Silhouette coefficient:

$$s = \frac{b - a}{\max(a, b)}$$

- a - The mean distance between a sample and all other points in the same cluster.
- b - The mean distance between a sample and all other points in the *next nearest cluster*.

Dunn index:

$$D = \frac{\min_{i \neq j} \rho(\mu_i, \mu_j)}{\max_{\mathrm{x}_i, \mathrm{x}_j \in \mu} \rho(\mathrm{x}_i, \mathrm{x}_j)}$$

Davies-Bouldin index:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\overline{\rho(\mu_i, \mathrm{x}^i)} + \overline{\rho(\mu_j, \mathrm{x}^j)}}{\rho(\mu_i, \mu_j)} \right)$$