Statistical Databases and Registers with some datamining

a course in

Survey Methodology and Official Statistics

Pages in the book: 501-528

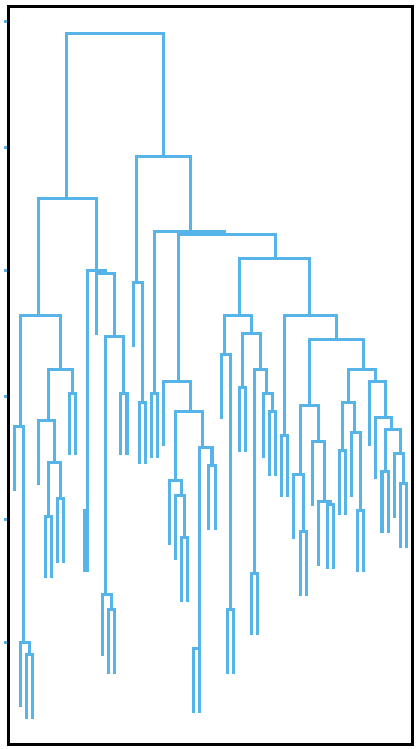
Department of Statistics
Stockholm University

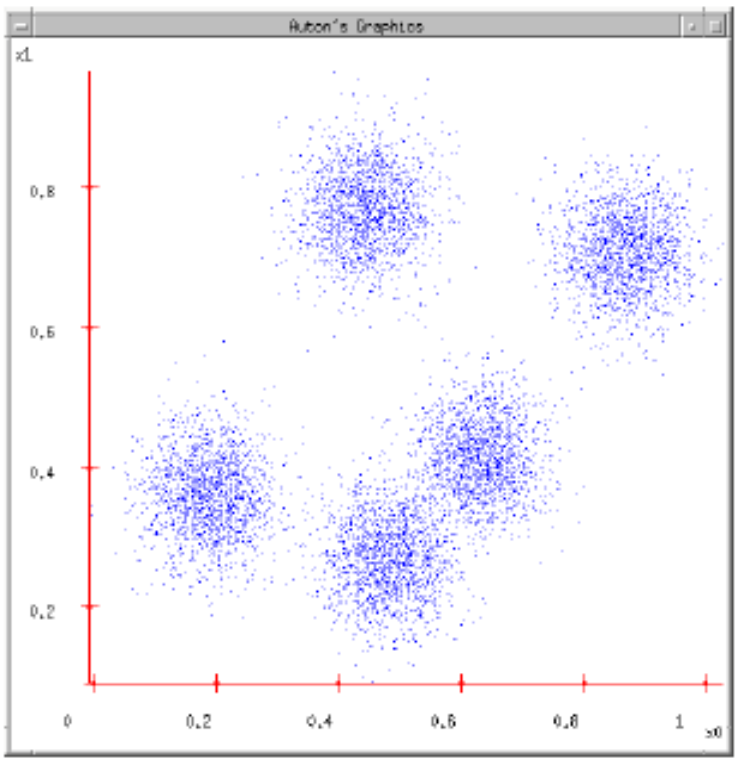
October 2010

Cluster analysis

Goals for cluster analysis (also called data segmentation)

- arrange into a natural hierarchy
- group with respect to similarities





When is cluster analysis useful

- You need to identify people with similar patterns of past purchases so that you can tailor your marketing strategies.
- You've been assigned to group television shows into homogeneous categories based on viewer characteristics. This can be used for market segmentation.
- You can use it in biology, to derive plant and animal taxonomies, to categorize genes with similar functionality, and to gain insight into structures inherent in populations.
- You're trying to examine patients with a diagnosis of depression to determine if distinct subgroups can be identified, based on a symptom checklist.

Tools for cluster analysis

Definition (Data matrix)

Set

$$x_{ij} = \text{measurement object } i \text{ variable } j$$

where $i = 1, 2, \ldots, n, j = 1, 2, \ldots, p$. We then have the $n \times p$ data matrix

$$\mathbb{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

Definition (Dissimilarity matrix)

Set

$$\mathbb{D} = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots \\ d(2,1) & 0 & \cdots & \cdots & \cdots \\ d(3,1) & d(3,2) & 0 & \cdots & \cdots \\ \cdots & \cdots & \cdots & \ddots & \cdots \\ d(n,1) & d(n,2) & \cdots & d(n,n-1) & 0 \end{bmatrix}$$

where $d(i,j)$ is the measured **difference** or **dissimilarity** between object i and j, $i,j = 1,2,\ldots,n$.

Here $d(i,j)$ is a distance function ie

1. $d(i,j) \geq 0$
2. $d(i,i) = 0$
3. $\mathrm{d}(i,j) = d(j,i)$
4. $d(i,j) \leq d(i,k) + d(k,j)$

K-means

K-means is a natural and easy method to find clusters, when you have given the number of clusters.

It is based on Euclidean distance and the following observation

$$
\begin{aligned}
T &= \sum_{j=1}^{p} \sum_{i=1}^{n} \sum_{i'=1}^{n} \left(x_{ij} - x_{i'j}\right)^2 \\
&= \sum_{j} \sum_{i} \sum_{i'} \left(\left(x_{ij} - \bar{x}_j\right)^2 + 2\left(x_{ij} - \bar{x}_j\right)\left(x_{i'j} - \bar{x}_j\right) + \left(x_{i'j} - \bar{x}_j\right)^2 \right) \\
&= \sum_{j} \left(\sum_{i} \sum_{i'} \left(x_{ij} - \bar{x}_j\right)^2 + \sum_{i} \sum_{i'} \left(x_{i'j} - \bar{x}_j\right)^2 \right) \\
&= 2 \times n^2 \times \text{variance}
\end{aligned}
$$

K-means (forts)

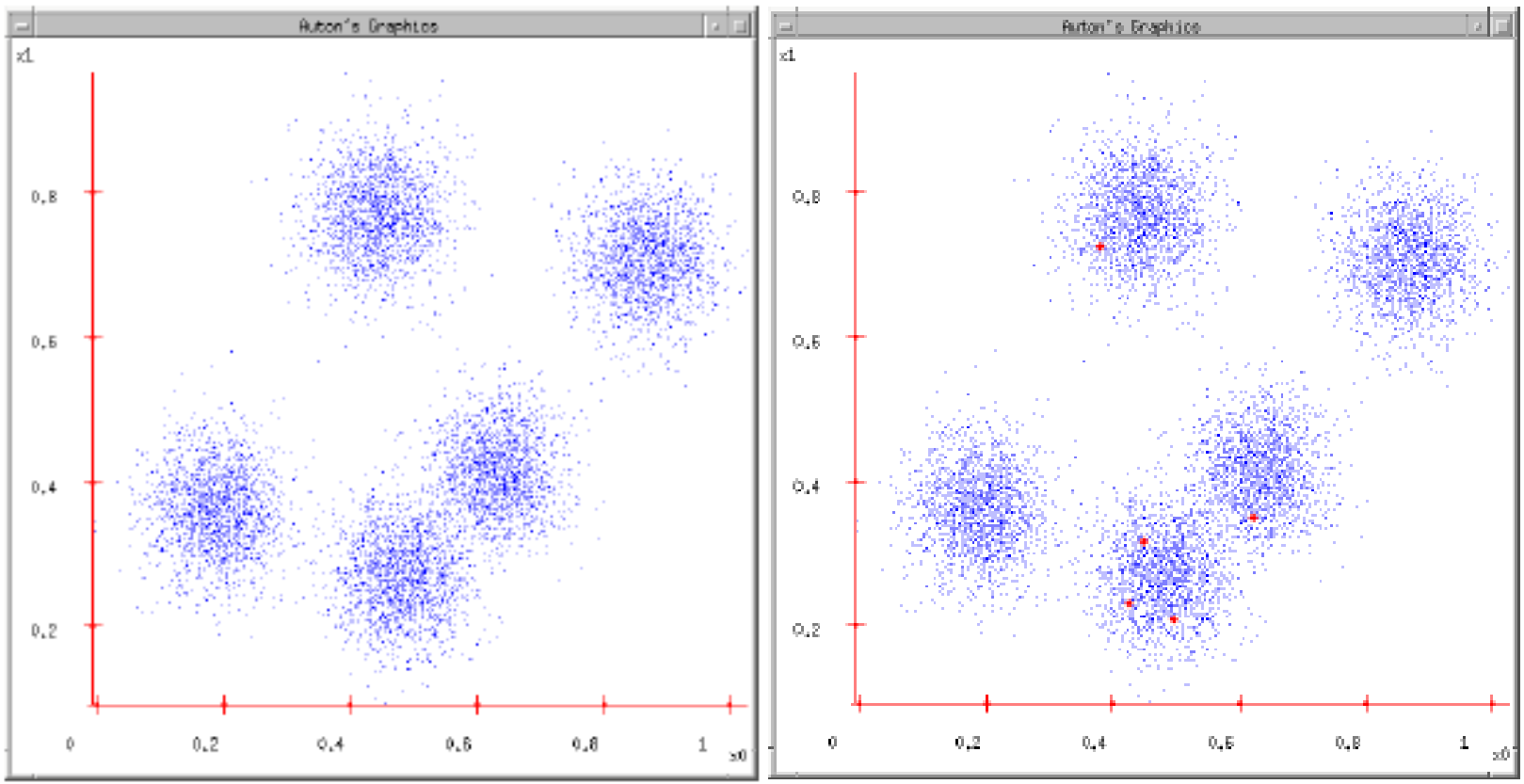
We know that the total variation may be split into two parts: Within cluster variation (W) and between cluster variation (B).

The method of K-means try to find a partition of clusters such that W is minimized. The algorithm to do this is:

ALGORITHM

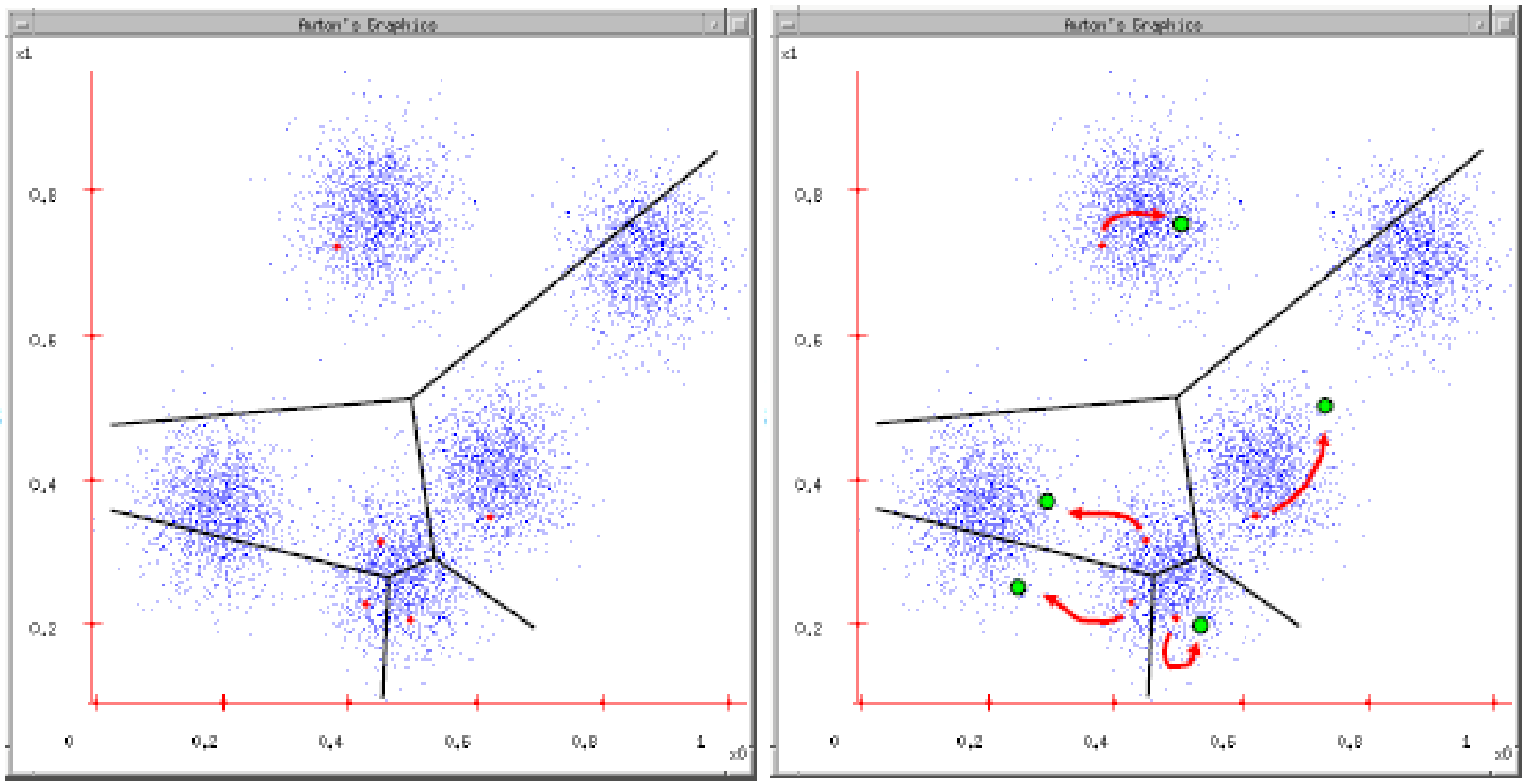
1. Choose points $m_1, \ldots, m_K$ and call them centroids
2. Partition the population into K subsets so each point x_{ij} is attached to the m_k which is closest, Euclidean distance.
3. For subset k, $k = 1, 2, \ldots, K$, compute the center of gravity and set m_k equal to this center.
4. If at least one m_k changes then go to 2, otherwise we stop.

Example: K-means (start)



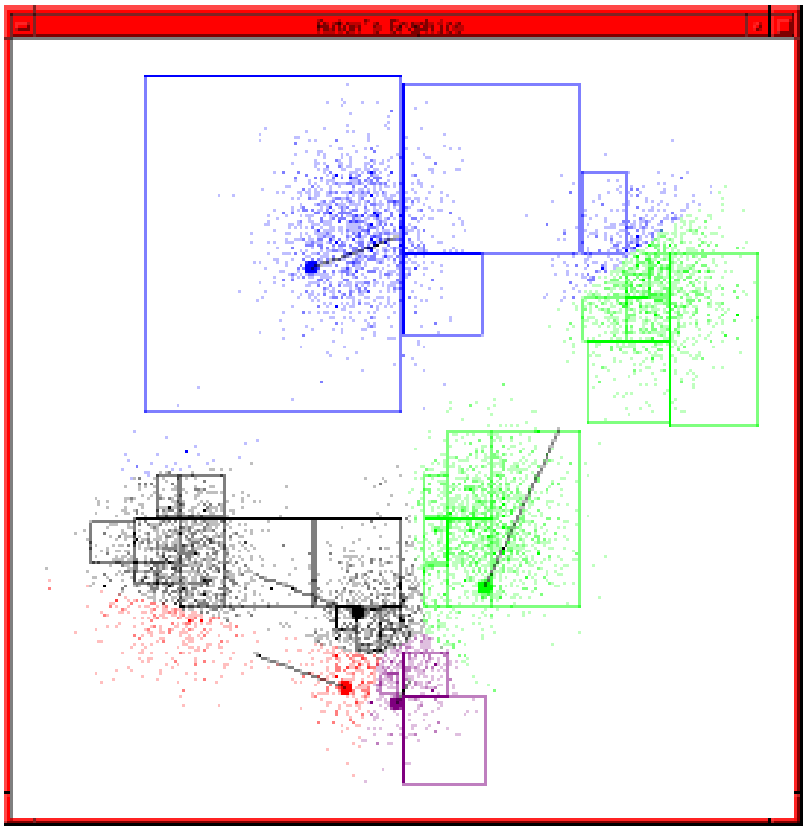
The choosen centroids (5) are in red (not a particular good choice).

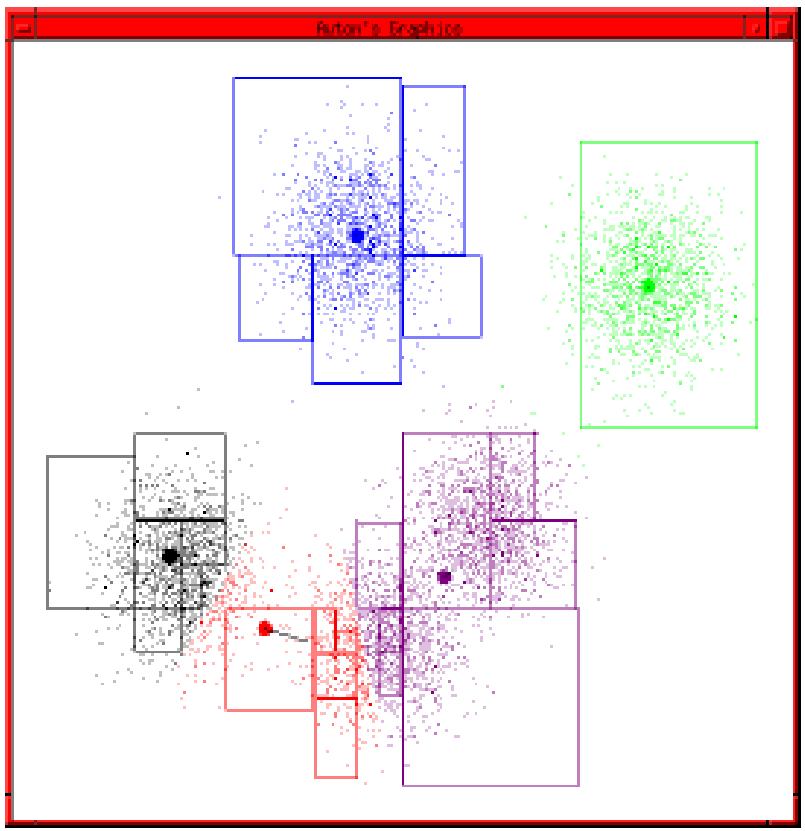
Example: K-means (cont)



The first picture shows the induced partition, by the centroids, of subsets. Second picture shows the new centroids.

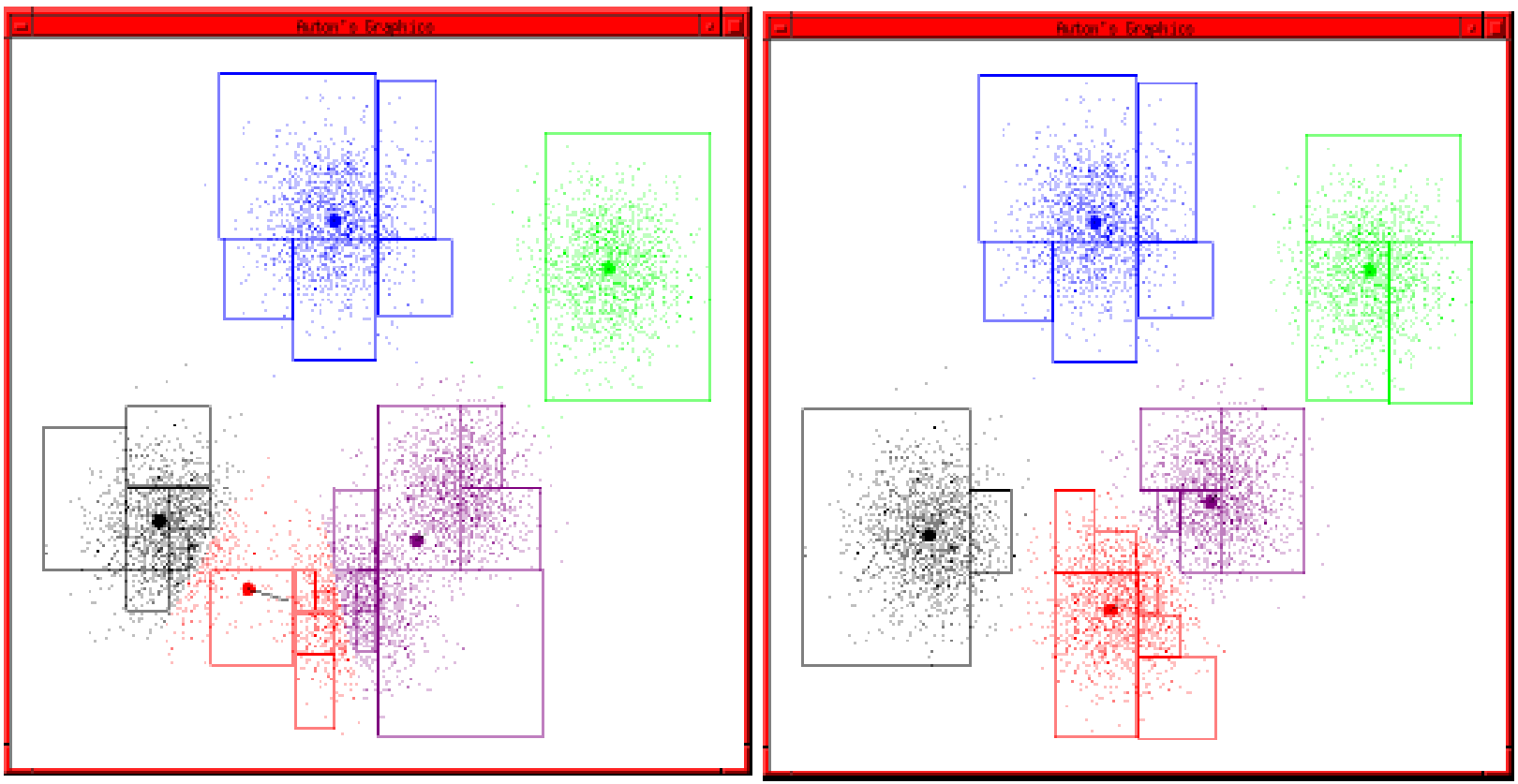
Example: K-means (cont)





In the first and second picture the direction, of the centroids movements, is indicated.

Example: K-means (stop)



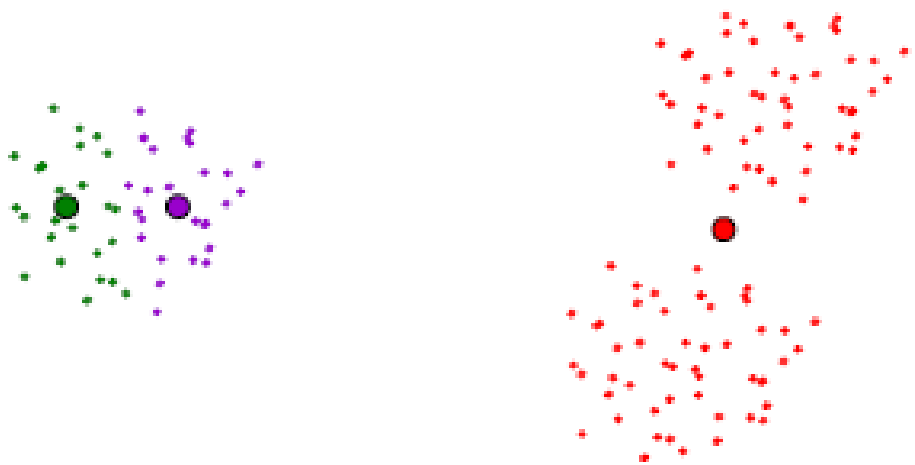
The left picture shows the final result. Also take a look at autonlab.com and the animation **Kmeans**.

Example: K-means (cont)

In our example we had 5 clusters and five centroids and we found the five best cluster.

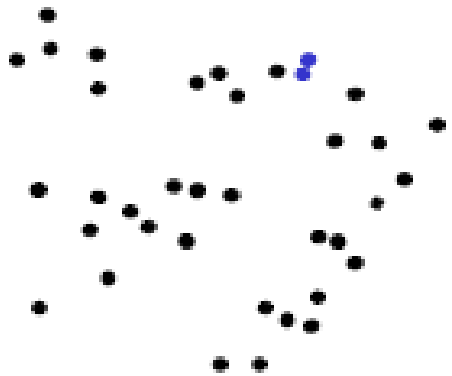
It is not always true that we will find the best partition even though we have the proper amount of centroids.

Contemplate the following figure

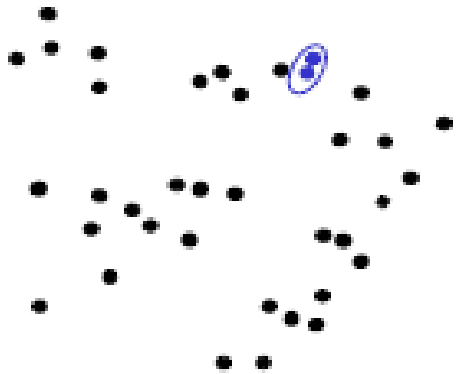


Will the given centroids ever move?

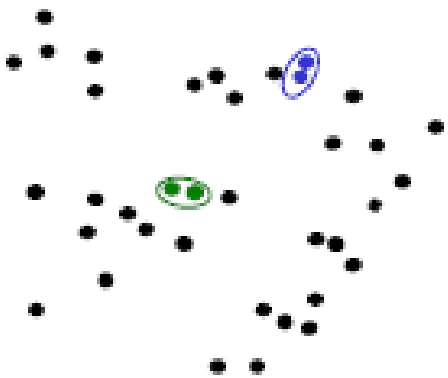
Example: Hierarchical clustering



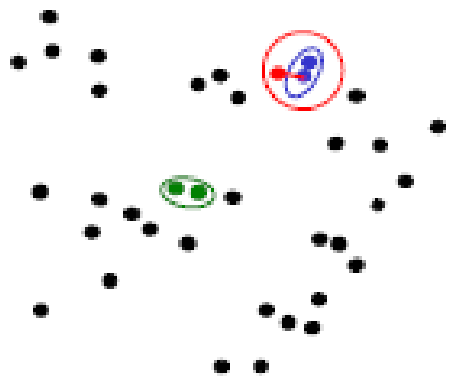
Step 1)



Step 2)



Step 3)



Step 4)

Example: Hierarchical clustering

Such clustering technique give rise to dendrograms