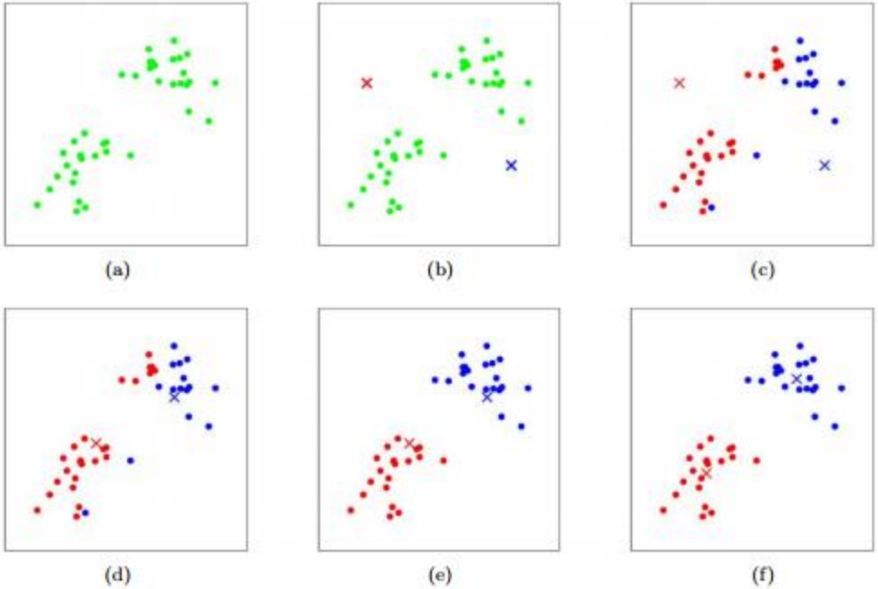
Advanced clustering

B4M36SAN

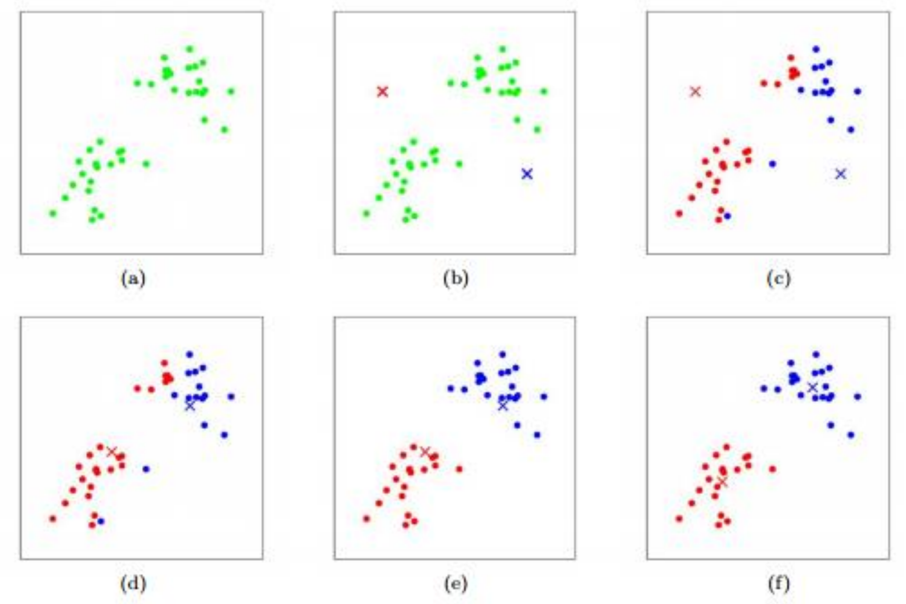
Outline

1. Review of baseline methods
    - K-means, Hierarchical clustering, DBSCAN
2. Spectral clustering
    - Principles and intuition, Showcase
    - DIY implementation
3. K-means on steroids
    - Relation to LDA and PCA
    - Ensemble clustering

K-means



K-means



Advantages:

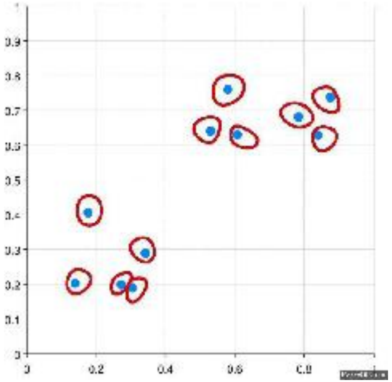
+ Fast, easy, simple

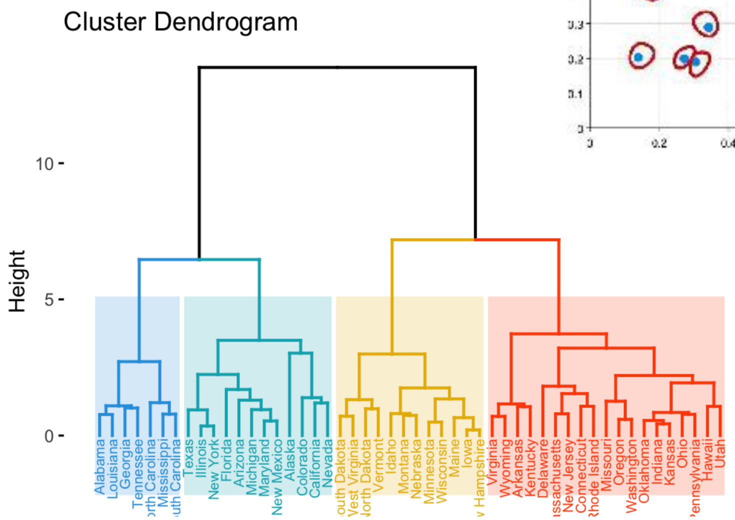
Susceptible to:

- Cluster shapes and densities
- Initialization
- Outliers

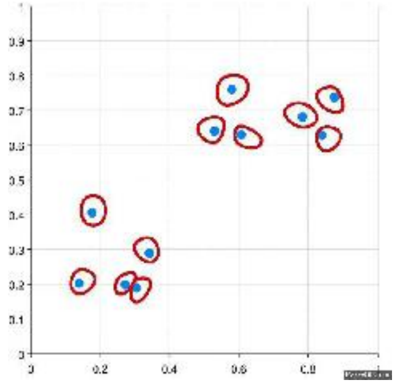
- Predefined number of clusters*

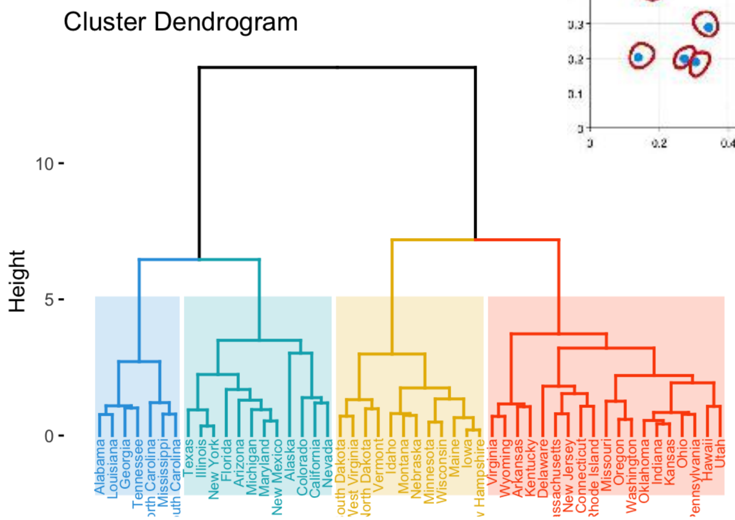
Hierarchical clustering





Hierarchical clustering





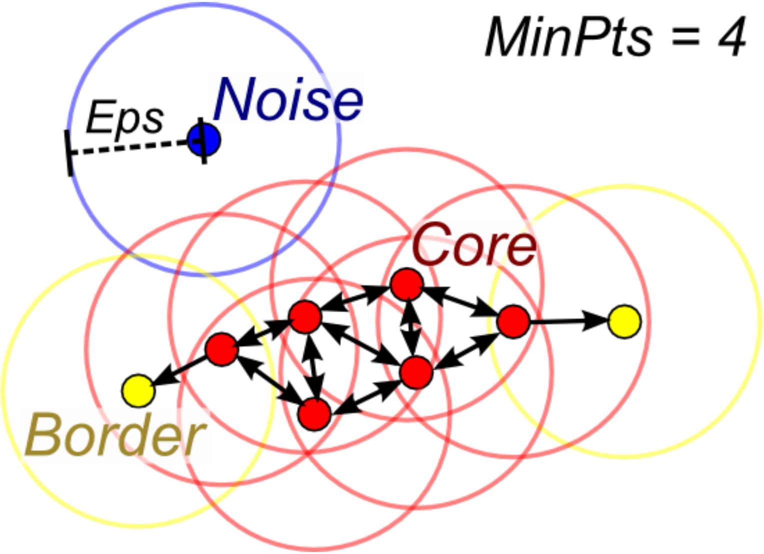
Advantages:

+ More informative hierarchical structure
+ Can vary number of clusters without re-computation

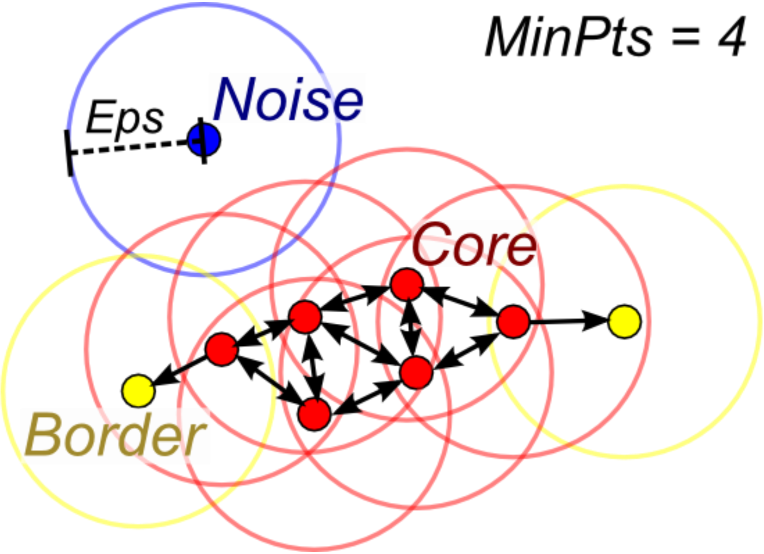
Susceptible to:

- Noise (single link)
- Outliers (complete link)
- Non-spherical clusters (average link)

DBSCAN



DBSCAN



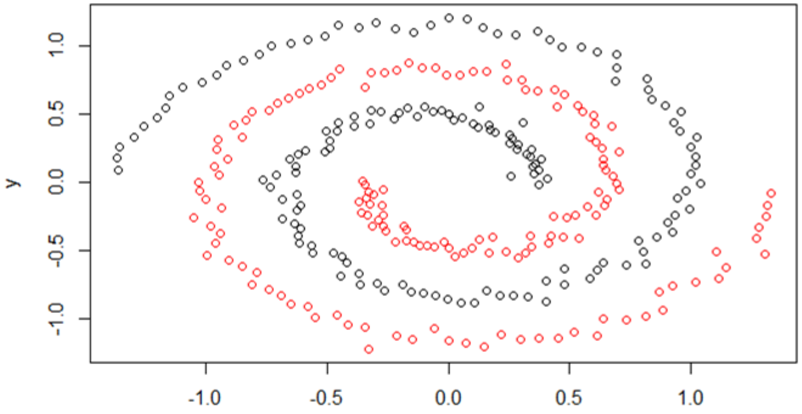
Advantages:

+ Cluster shapes are not an issue
+ Robust towards outliers/noise

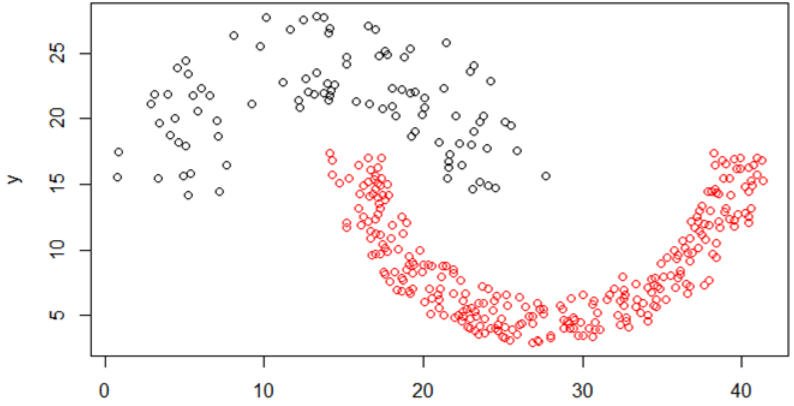
Susceptible to:

- Cluster densities
- Parametrization (*eps, MinPts*)

Datasets



2 spirals dataset

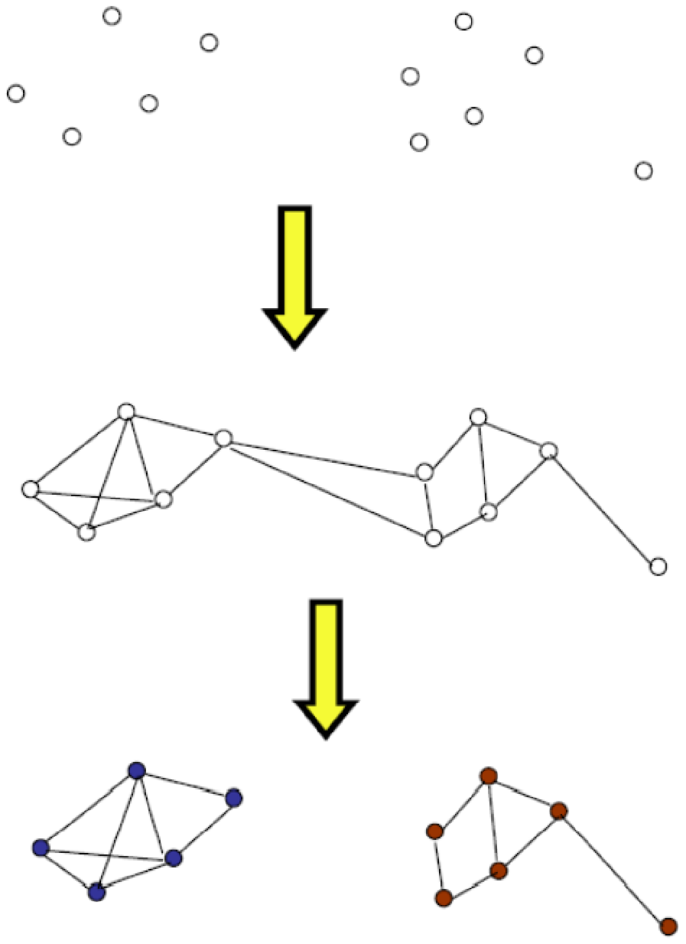


Jain dataset

Experiment yourself

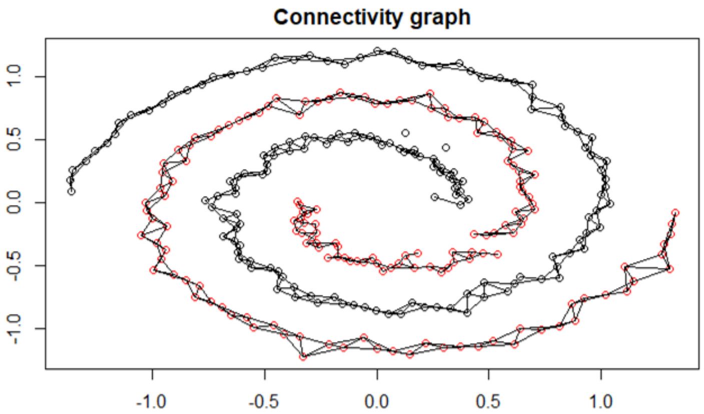
Spectral clustering

- Turns data into a graph
- Finds a *min-cut* of the graph
  - The partition forms the clusters
- Simple idea, not so simple steps



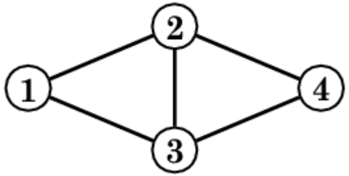
Azran: A Tutorial on Spectral Clustering

Spectral clustering



1. select the similarity function
   - linear, RBF, polynomial, etc.
   - a general rule assigning functions to problems does not exist,
2. compute the similarity (adjacency) matrix $\mathcal{S} = [s_{ij}]_{m\times m}$
   - (a new implicit feature space originates),
3. construct a "reasonable" similarity graph by editing $\mathcal{S}$
   - $\mathcal{S}$ is a complete graph, vertices $\sim$ objects, similarities $\sim$ edges,
   - remove long (improper) edges,
4. derive the Laplace matrix $\mathcal{L}$ out of the similarity matrix $\mathcal{S}$
   - unnormalized: $\mathcal{L} = \mathcal{D} - \mathcal{S}$,
   - normalized: $\mathcal{L}_{rw} = \mathcal{D}^{-1}\mathcal{L} = \mathcal{I} - \mathcal{D}^{-1}\mathcal{S}$,
5. project into an explicit space of k first eigenvectors of $\mathcal{L}$,
   - $\mathcal{V} = [v_{ij}]_{m\times k}$, eigenvectors of $\mathcal{L}$ as columns,
6. k-means clustering in $\mathcal{V}$ matrix
   - $\mathcal{V}$ rows interpreted as new object positions in k-dimensional space.

$$D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$



$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \qquad L = \begin{bmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ 0 & -1 & -1 & 2 \end{bmatrix}$$

Why 2nd eigenvector?

- concern the unnormalized option: $\mathcal{L} = \mathcal{D} - \mathcal{S}$
- then for $\forall f \in \mathbb{R}^m$

$$f'\mathcal{L}f = f'\mathcal{D}f - f'\mathcal{S}f =$$

$$= \sum_{i=1}^{m} d_i f_i^2 - \sum_{i,j=1}^{m} f_i f_j s_{ij} =$$

$$= \frac{1}{2}\Big(\sum_{i=1}^{m}\big(\sum_{j=1}^{m} s_{ij}\big) f_i^2 - 2\sum_{i,j=1}^{m} f_i f_j s_{ij} + \sum_{j=1}^{m}\big(\sum_{i=1}^{m} s_{ij}\big) f_j^2\Big) =$$

$$= \frac{1}{2}\sum_{i,j=1}^{m} s_{ij}(f_i - f_j)^2$$

2nd eigenvector is *f,* that minimizes this function (without proof)

But what is this function telling?

It's a cost function!
If two points are connected i.e $s_{ij}=1$,
it penalizes the difference in their labels

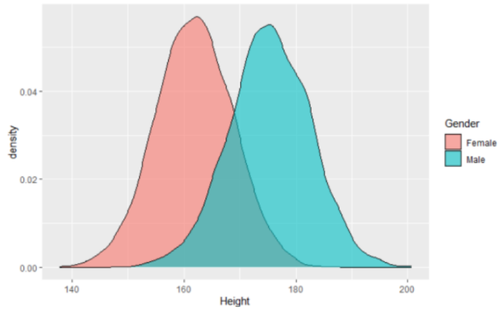
K-means relation to PCA and LDA

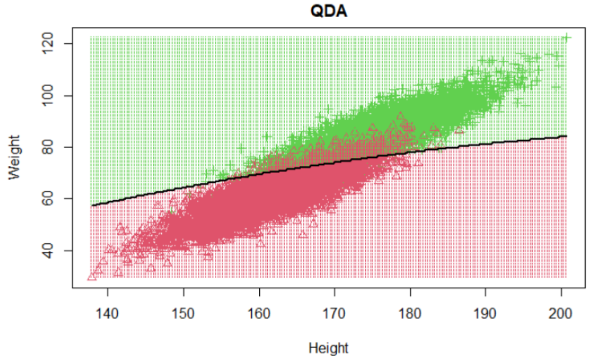
- Initialization issues
  - Repeated starts
  - ***PCA-Part***
    - A divisive hierarchical approach based on **PCA.**
    - Starting from an initial cluster that contains the entire data set, the iteratively select the **cluster with the greatest SSE** and divide it into two subclusters using a **hyperplane** that passes through the cluster centroid and is **orthogonal to the principal eigenvector** of the cluster covariance matrix. This procedure is repeated until K clusters are found

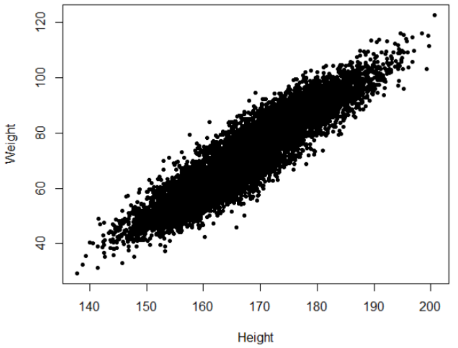
Celebi, M.E., Kingravi, H.A. and Vela, P.A., 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm.

K-means relation to PCA and LDA

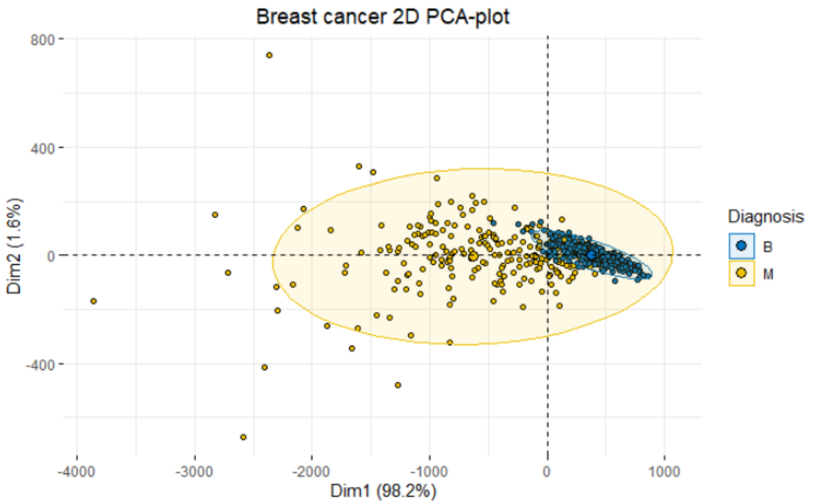
- LDA
  - Assumes data for each **class** come from (mulitvar.) normal distributions
  - Uses Bayes theorem to decide which class a sample belongs to
- EM-GMM clustering
  - Soft version of K-means
  - Also assumes data for each **cluster** come from (mulitvar.) normal distributions
  - The parameters estimated are $\boldsymbol{\mu_c}$,$\boldsymbol{\sigma_c}$ and $\boldsymbol{p_c}$ of the clusters

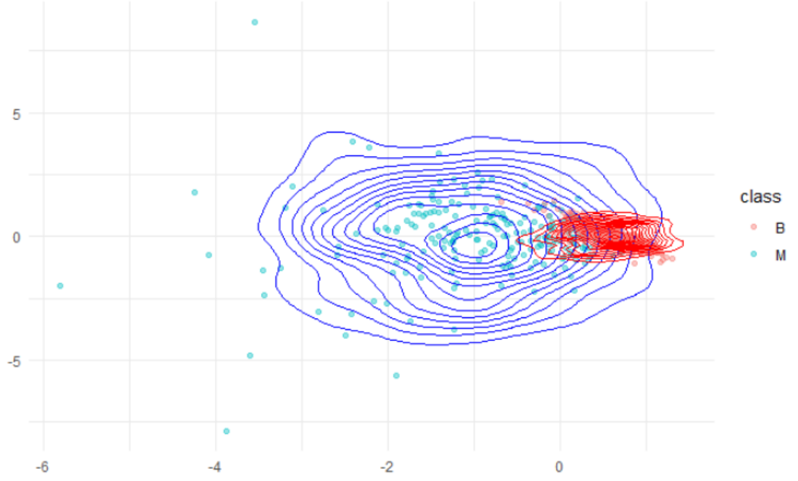






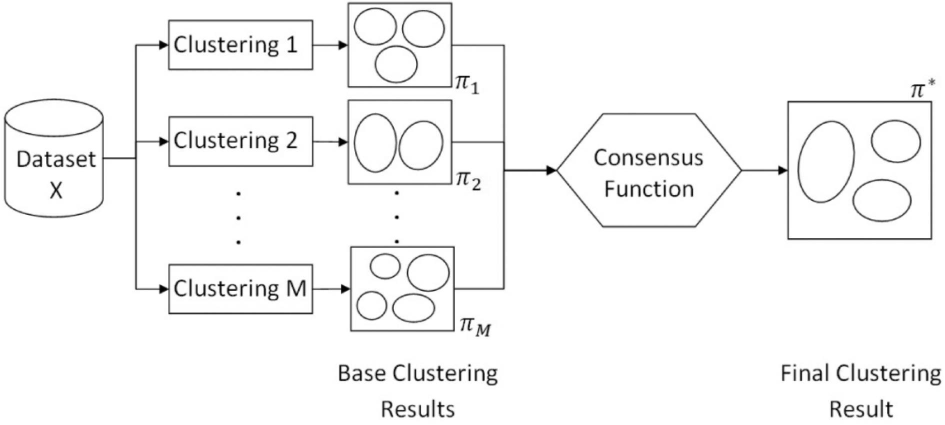
EM clustering on *Breast cancer* dataset

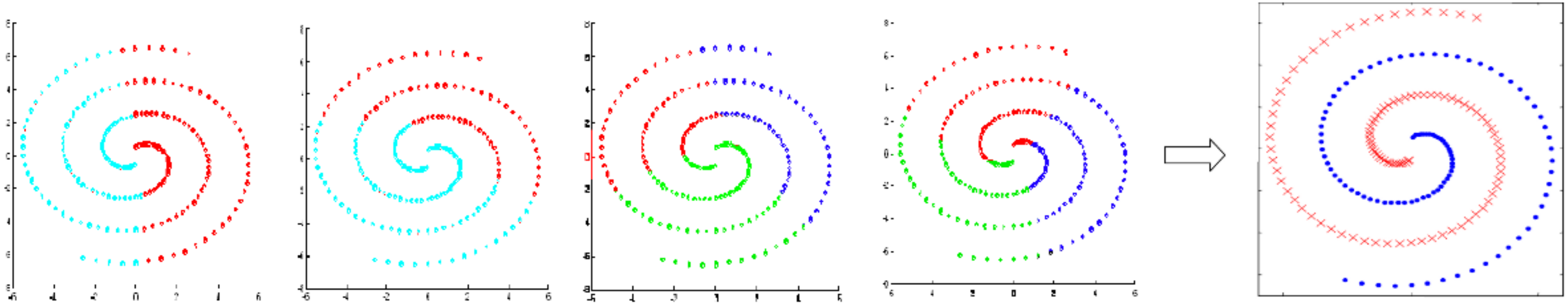




- Demo in the *./extra* folder of the course materials

Ensemble clustering





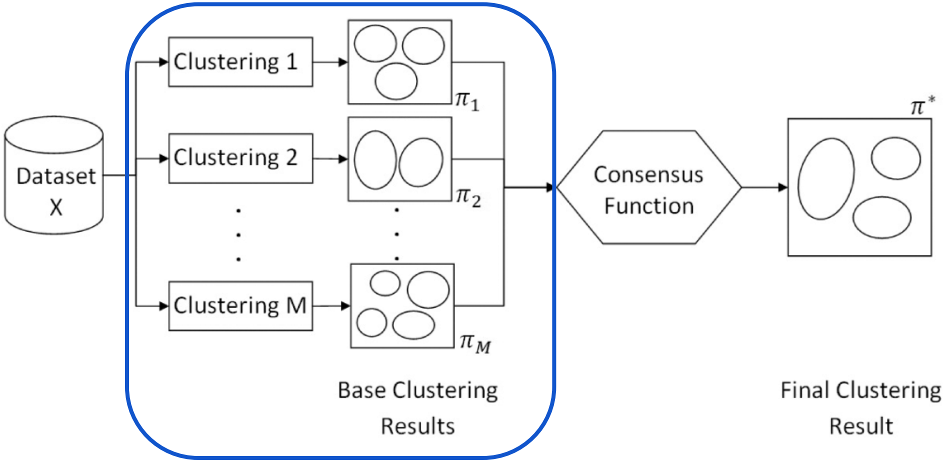
How to generate clusters?

- Using **different clustering algorithms**

  e.g. *K-means, hierarchical clustering, spectral clustering*, …
- Running **the same algorithm with different parameters** or initializations, e.g.,

  use different dissimilarity measures

  use different number of clusters
- Using **different samples of the data**



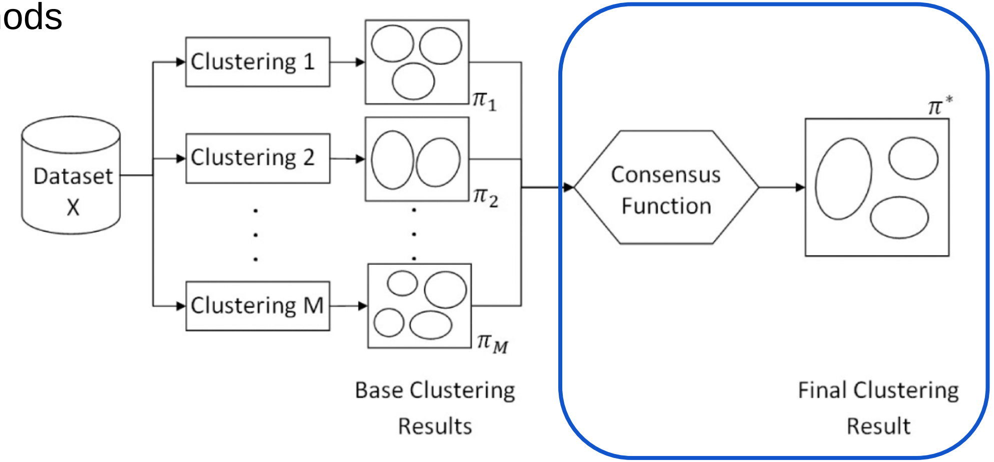
How to combine the partitions?

- **Median partition** based approaches
  - “Averaging” all ensemble partitions

- **Co-occurrence** based approaches
  - Relabeling/voting based methods
  - <u>Co-association matrix </u>based methods
  - Graph based methods

Voting

	P_1	P_2	P_3	P^*
v_1	1	1	1	1
v_2	1	1	1	1
v_3	2	2	1	2
v_4	2	2	2	2
v_5	3	3	3	3
v_6	3	3	3	3



Resources

http://www.cse.msu.edu/~cse802/EnsembleClustering_Jinfeng_jain.pptx

https://stanford.edu/~cpiech/cs221/handouts/kmeans.html

https://csdl-images.computer.org/trans/tk/2012/03/figures/ttk20120304131.gif

https://www.researchgate.net/figure/An-example-of-the-Laplacian-matrix-of-a-simple-network-n-4_fig1_305653264

https://images.amcnetworks.com/ifc.com/wp-content/uploads/2015/03/EnemyAtTheGates_MF.jpg

https://gfycat.com/somelonelycaterpillar

Luxburg07_tutorial_spectral_clustering.pdf (mit.edu)

[1209.1960] A Comparative Study of Efficient Initialization Methods for the K-Means Clustering Algorithm (arxiv.org)

Rajaraman, Anand, and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.