3050571 Practical Clin Data Sci

Session 8: Clustering

February 15, 2024

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

The heart of clustering

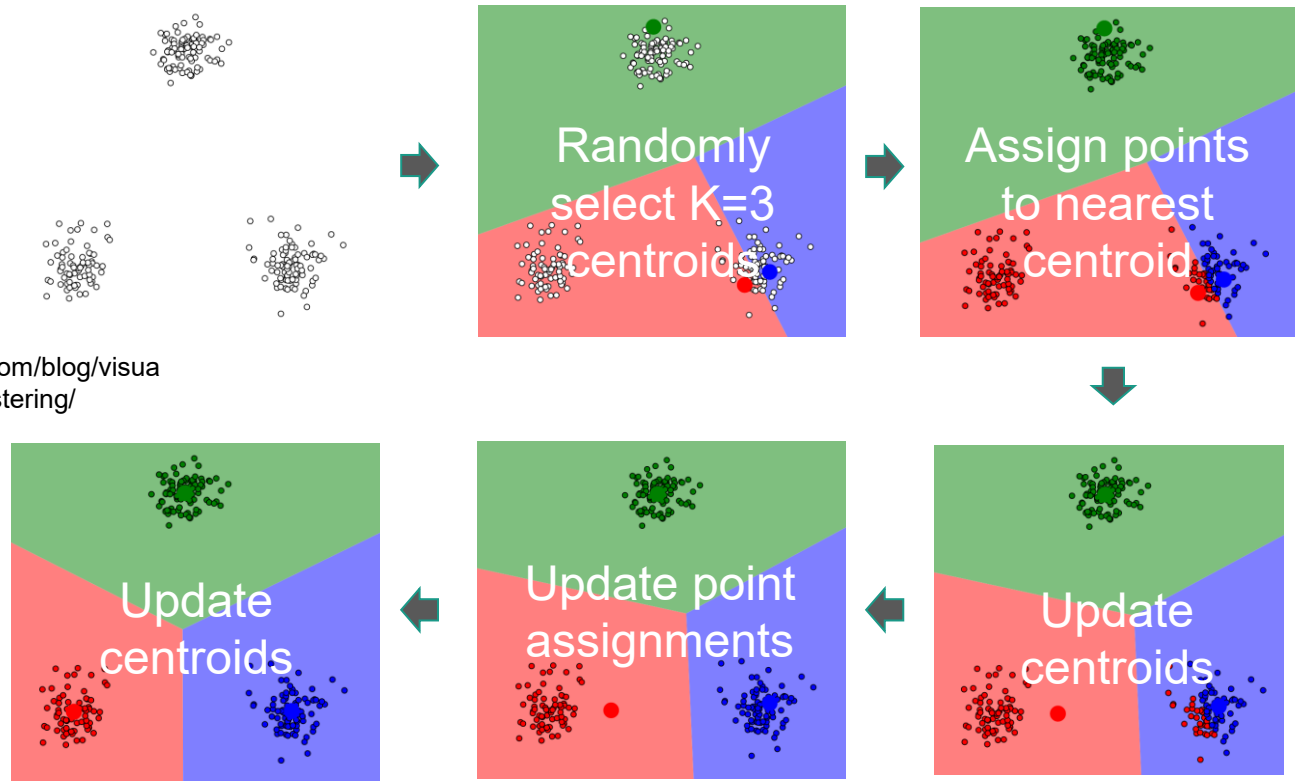
- **Goal**: Group similar data point together

- How to define similarity?
  - Distance: Between two data points
  - Linkage: Between groups of data points

- How many clusters is appropriate?
  - Within-cluster (small) versus between-cluster (large) distance

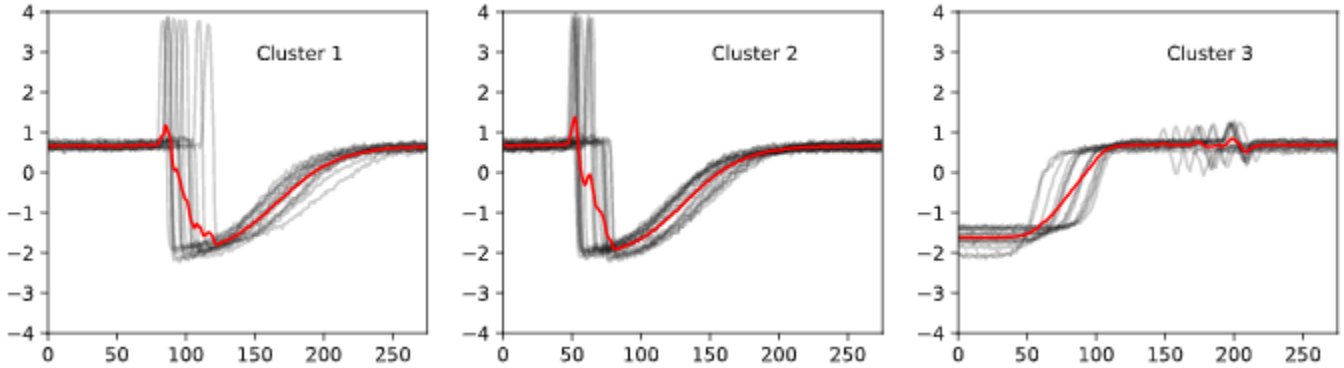
k-means, *k*-medoids clustering

k-means algorithm



www.naftaliharris.com/blog/visualizing-k-means-clustering/

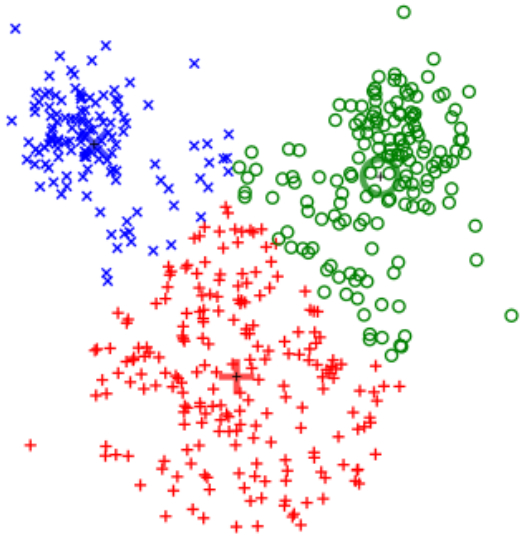
k-mean clustering of signal data



https://towardsdatascience.com/how-to-apply-k-means-clustering-to-time-series-data-28d04a8f7da3

- Normalize y-axis signals to focus on trend, not magnitude
- Expect a small number of major trends, each with many observations
  - Such as upward, downward, etc.

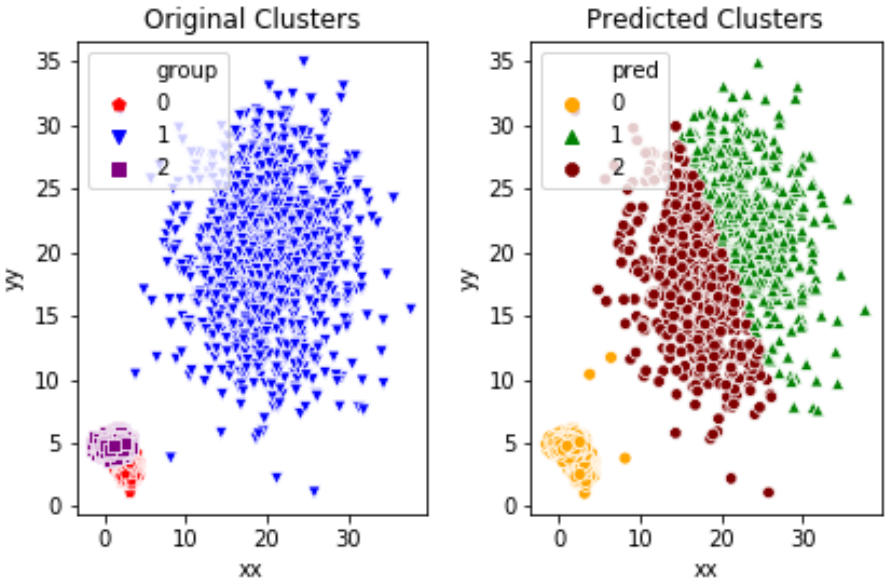
Limitation of *k*-mean



- **Euclidean distance**
- **Clusters are of equal radius**
- The **initial guess of the locations of *k* means** can affect the final clusters
  - Need to be repeated

https://en.wikipedia.org/wiki/K-means_clustering

Visual inspection is key

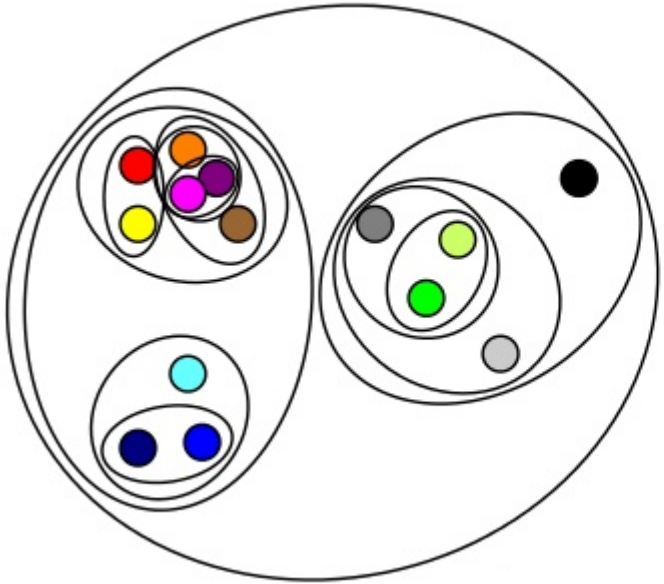


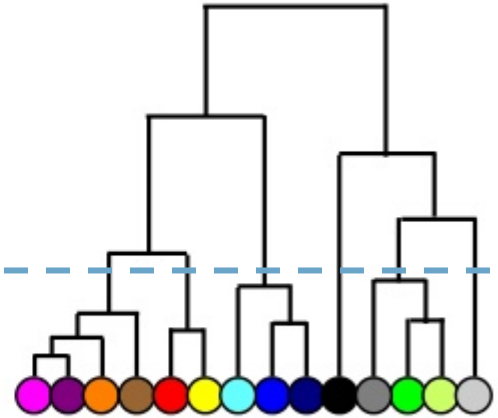
- Dimensionality reduction plot
- Distribution of feature values across groups
  - Explain subpopulations

https://towardsdatascience.com/kmeans-hyper-parameters-explained-with-examples-c93505820cd3

Agglomerative, hierarchical clustering

Agglomerative / Hierarchical clustering

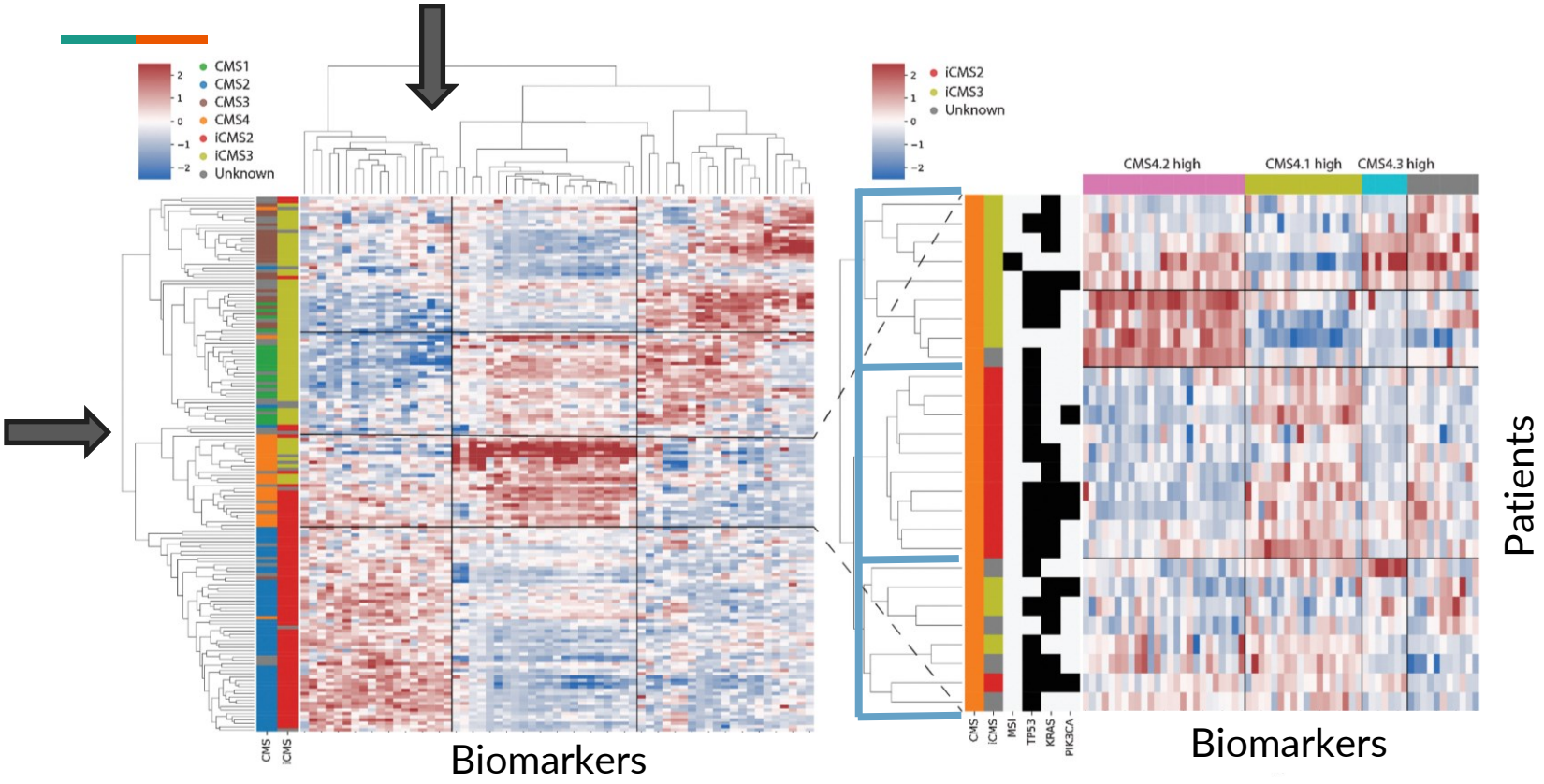




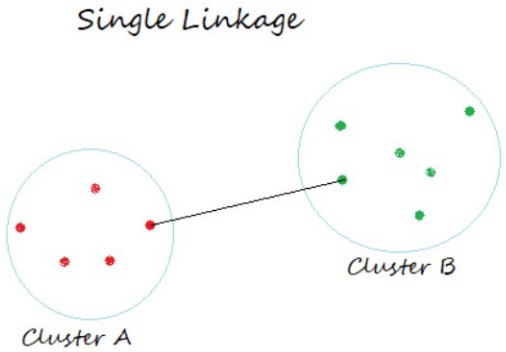
Select the **number of cluster** or **maximum distance** within group

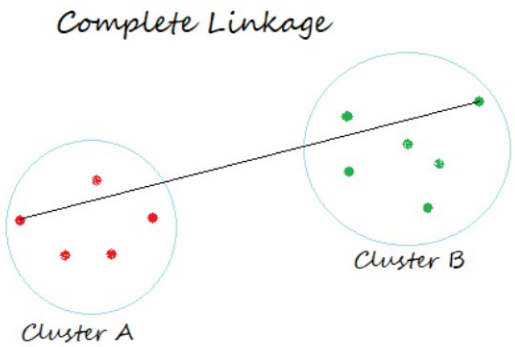
Source: www.slideshare.net/ElenaSgis/data-preprocessing-and-unsupervised-learning-methods-in-bioinformatics

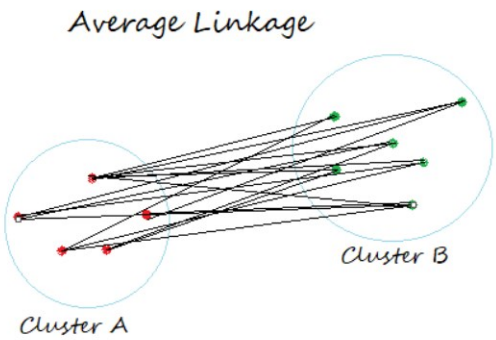
Hierarchical clustering of samples & features

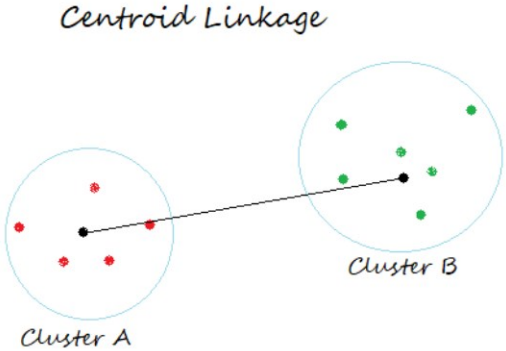


Linkage = distance metric for groups of data points









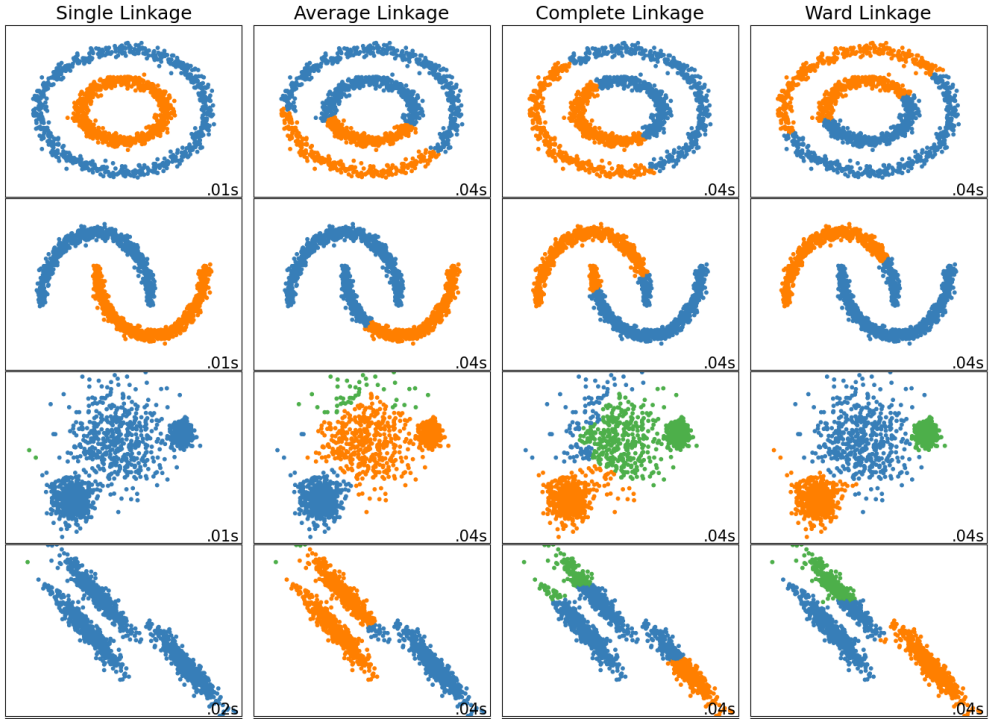
Ward Linkage

Reduction in variance after merging. Assume Euclidean distance.

https://www.analyticsvidhya.com/blog/2021/06/single-link-hierarchical-clustering-clearly-explained/

Impact of linkage

Single Linkage allows the formation of irregular cluster shapes

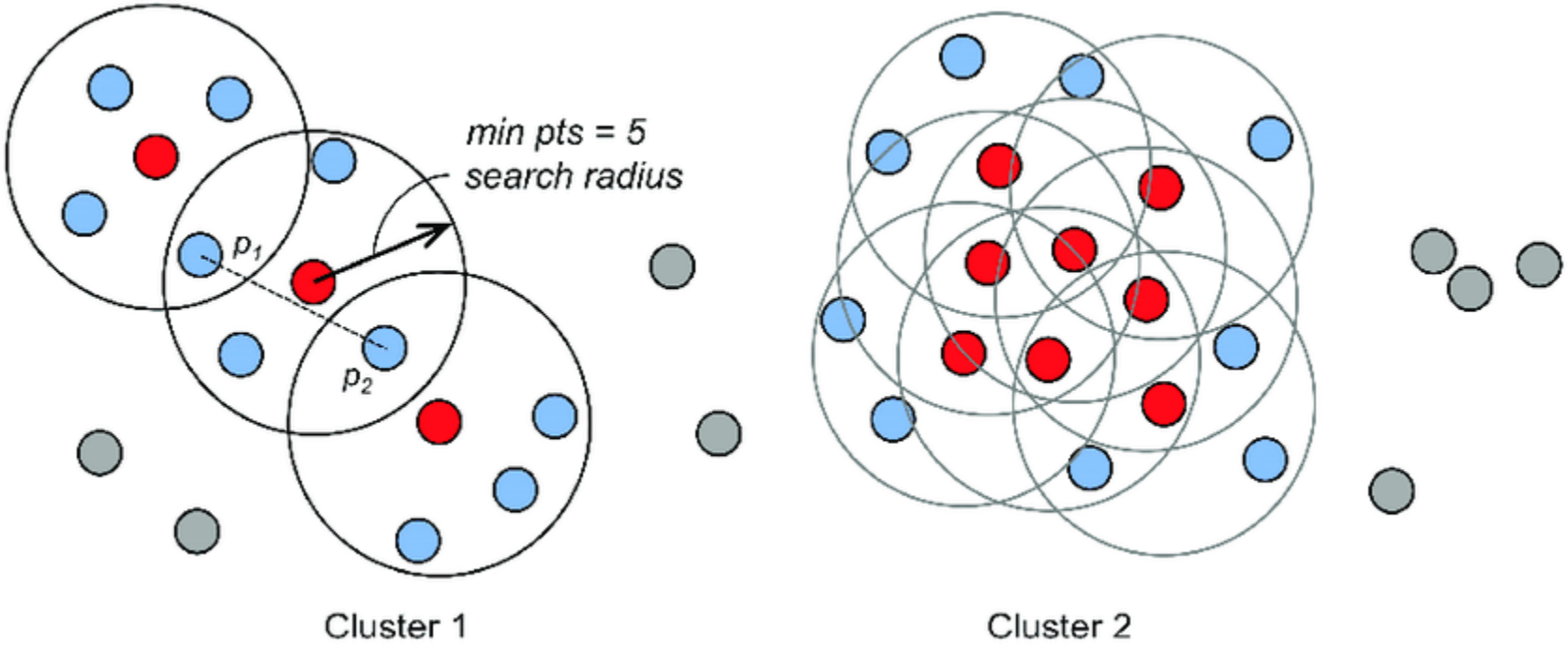


The choices of linkage alone is not enough to handle every case

https://scikit-learn.org/stable/auto_examples/cluster/plot_linkage_comparison.html

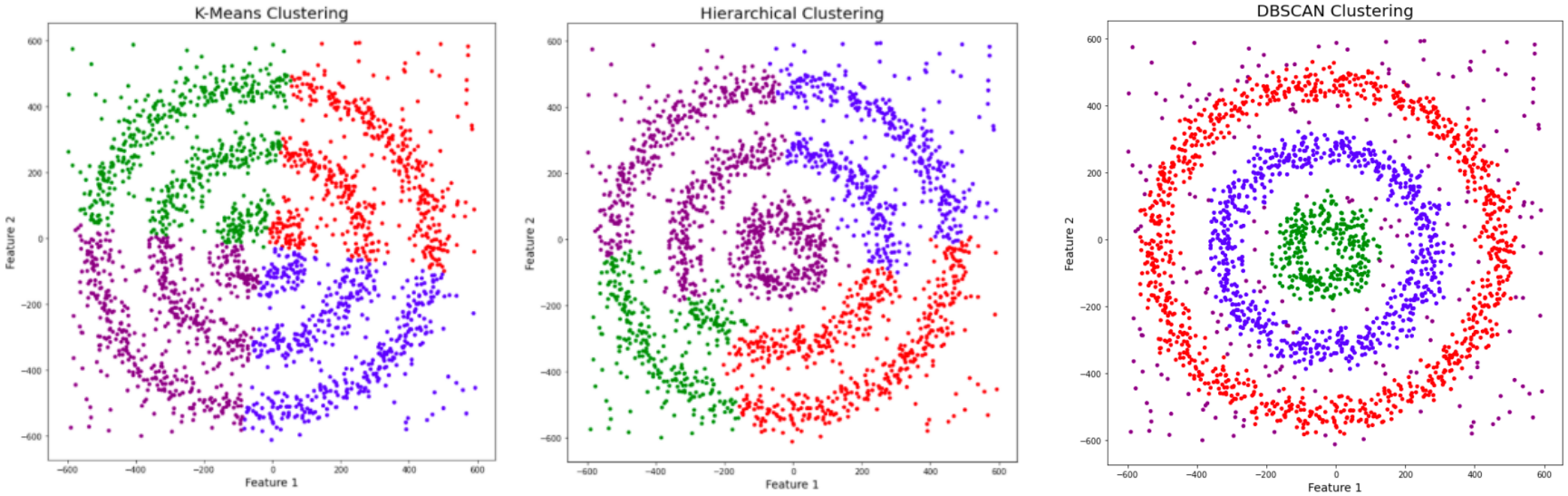
Density-Based Spatial Clustering of Applications with Noise

DBSCAN: A density-based technique



Difrancesco, P.-M. Remote Sensing 12:1885 (2020)

DBSCAN can handle complex cluster shape

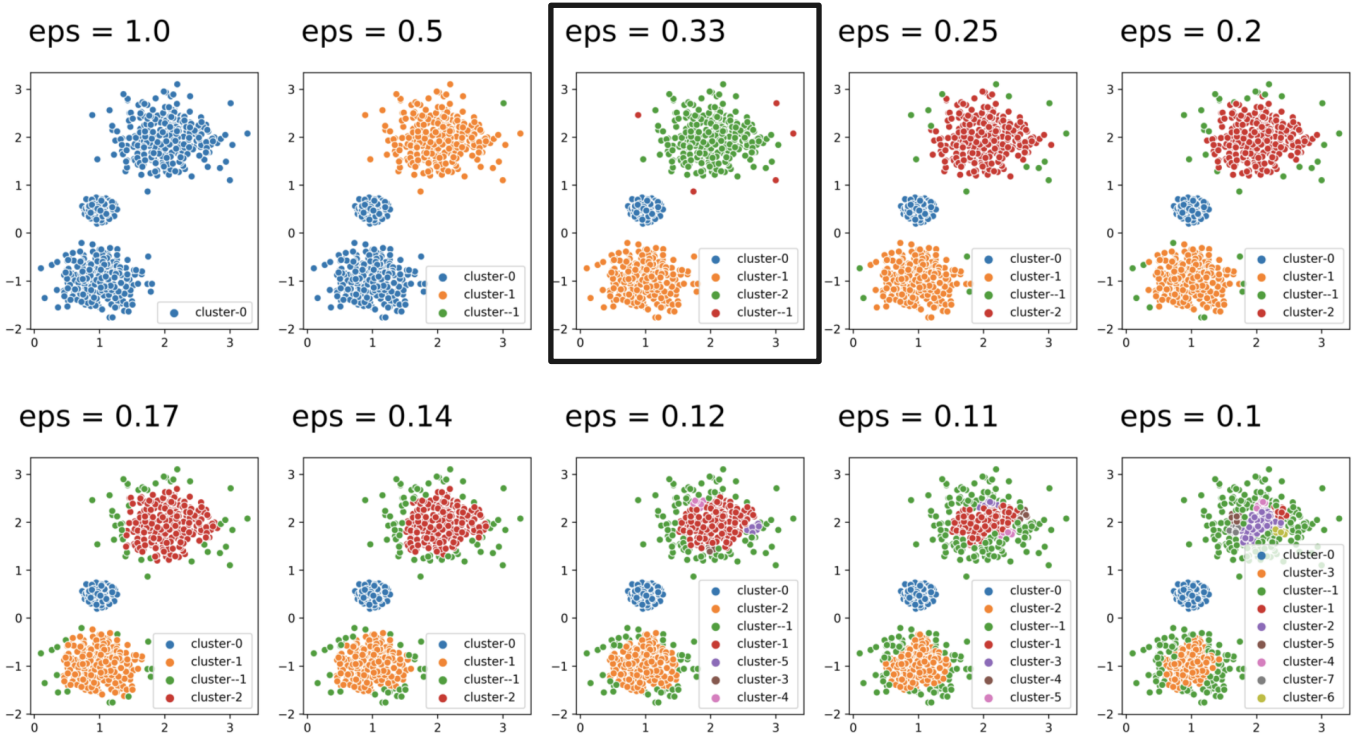


https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

- Distance-based techniques assume that data are spread in all directions

But may be difficult to tune

High eps, too few clusters

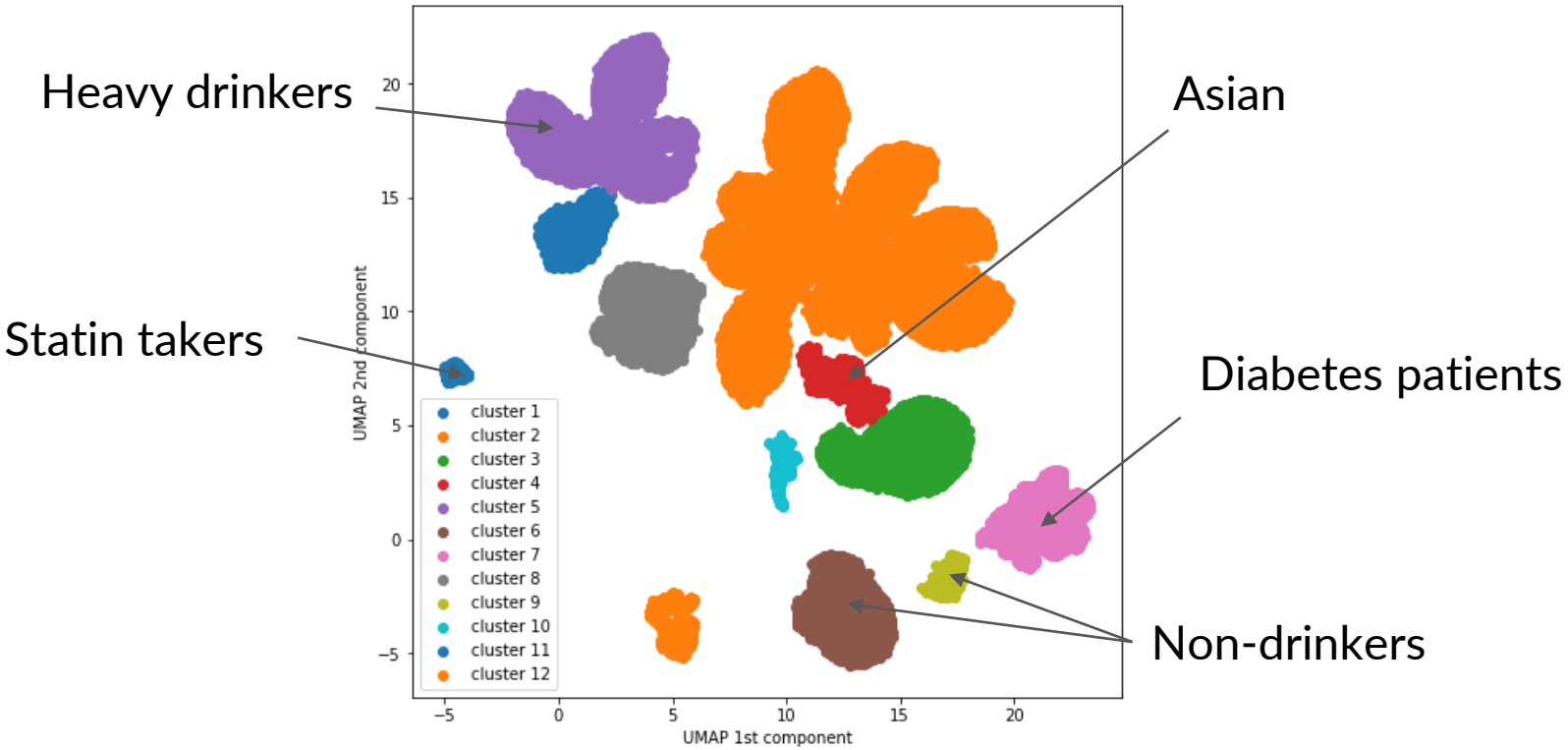


Low eps, too many outliers

Very low eps, too many clusters

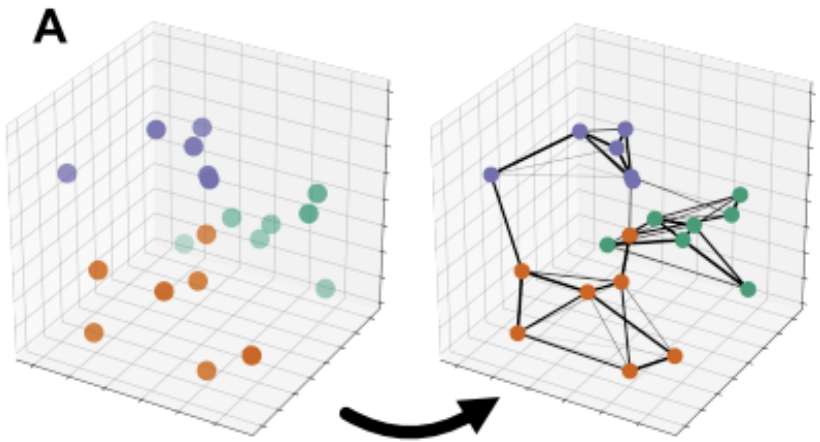
https://towardsdatascience.com/

DBSCAN on patient data

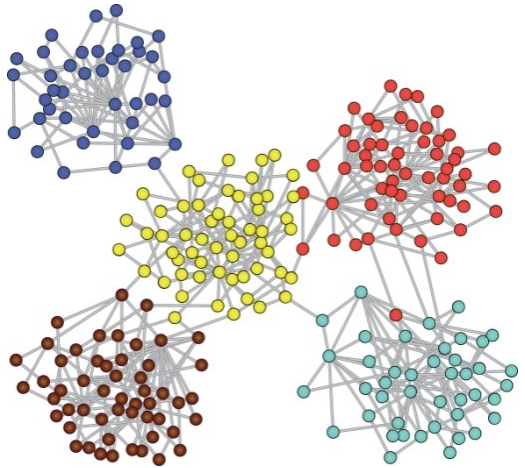


Network clustering

Network clustering



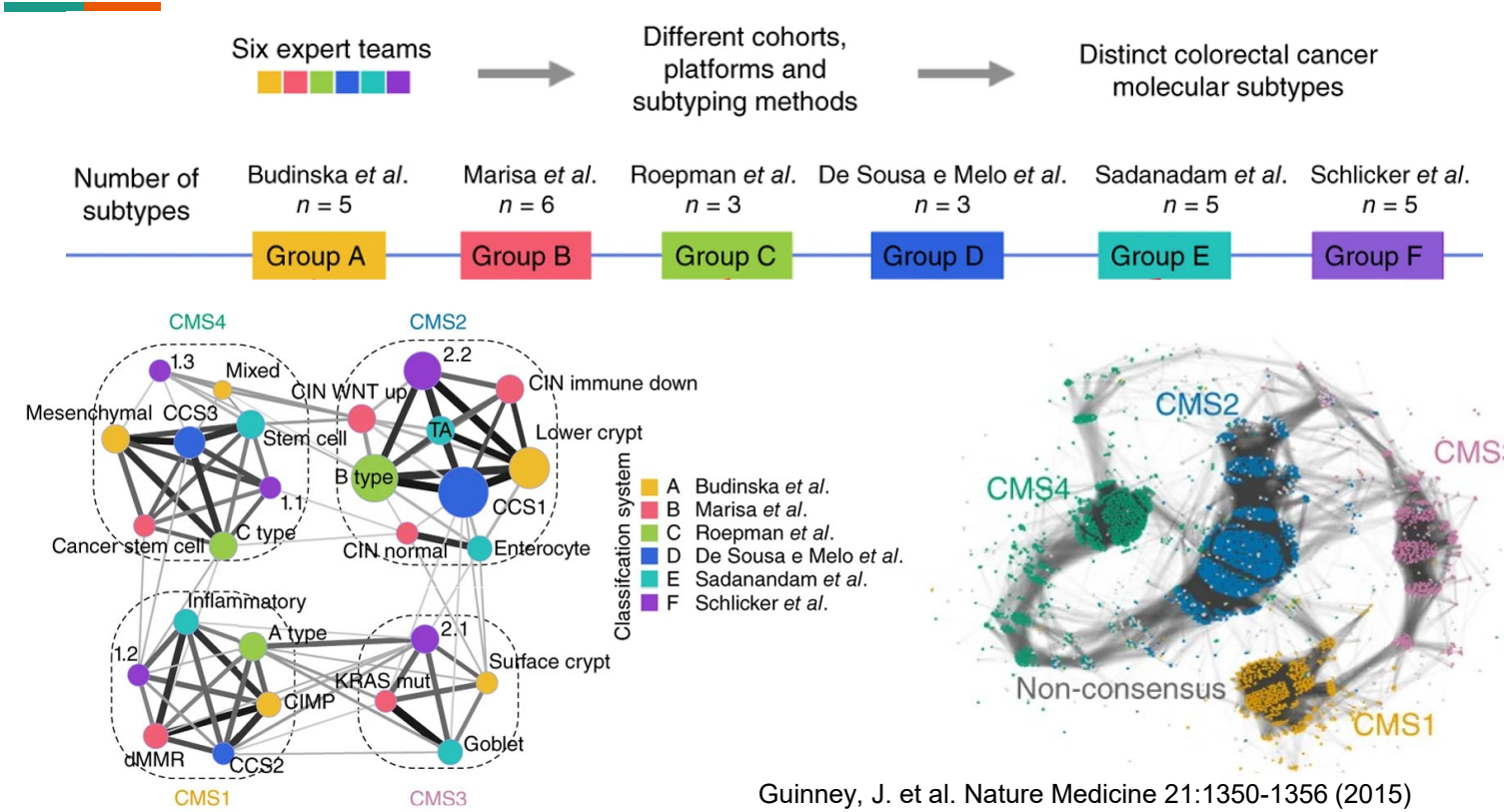
Sainburg, T. *et al.*, Neural Comput 33(11):2881-2907 (2021)



https://github.com/topics/graph-clustering

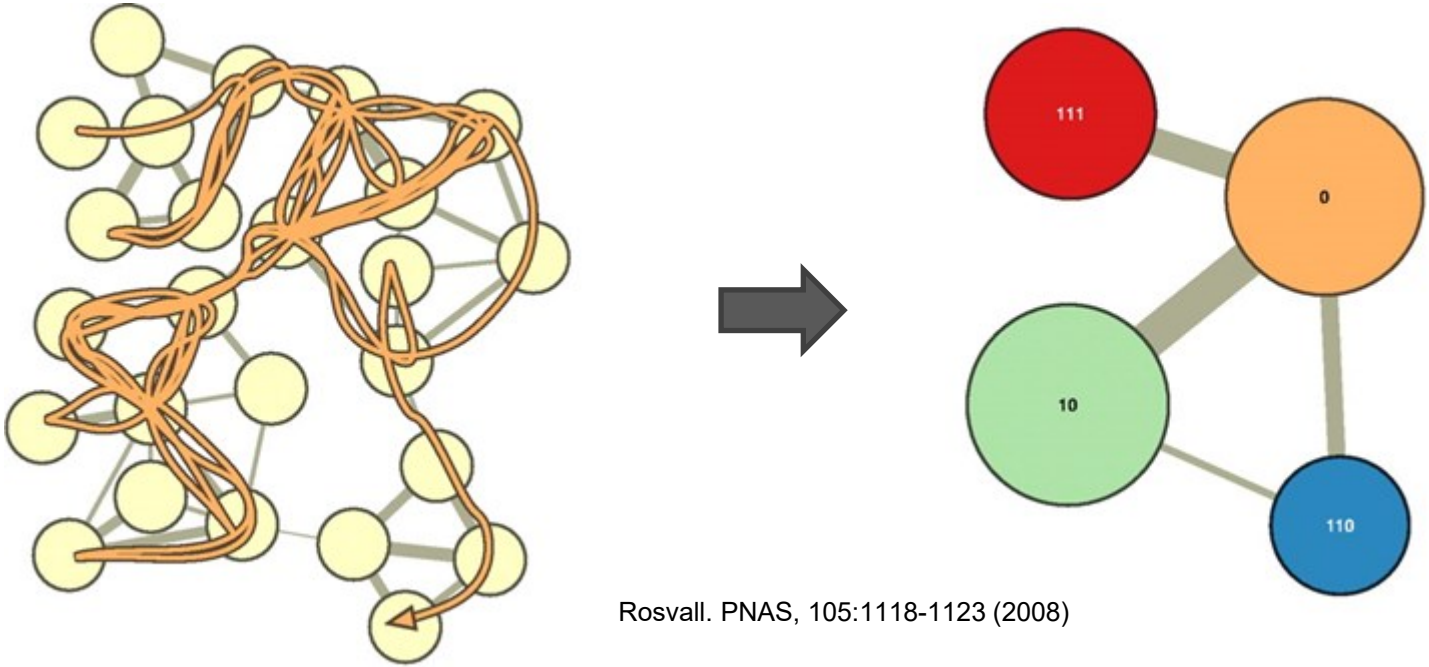
- View similarity between samples as a network
- [Optional] Remove edges with low score
- Split network into modules with dense edges or high-score edges

Consensus colorectal cancer subtyping



Guinney, J. et al. Nature Medicine 21:1350-1356 (2015)

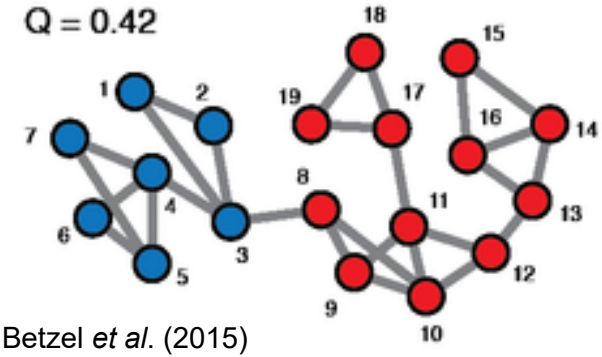
Identifying network module via random walks



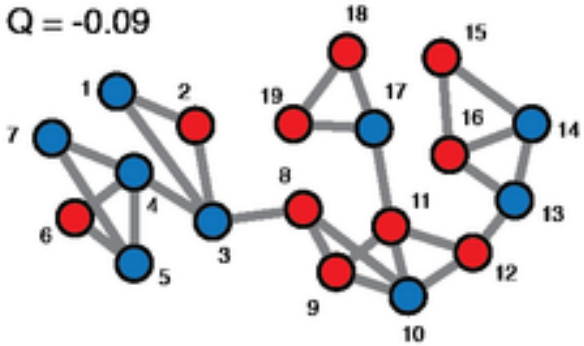
Rosvall. PNAS, 105:1118-1123 (2008)

- Nodes within a module should be visited together more often

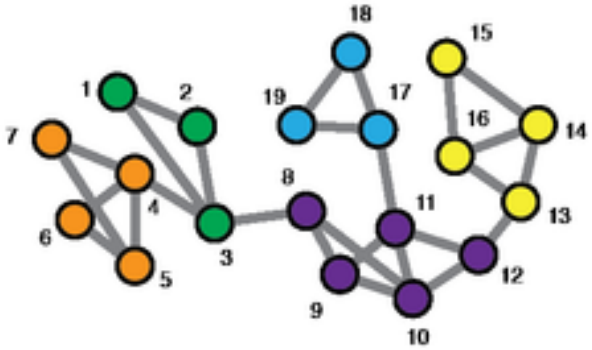
Modularity score



Betzel *et al*. (2015)



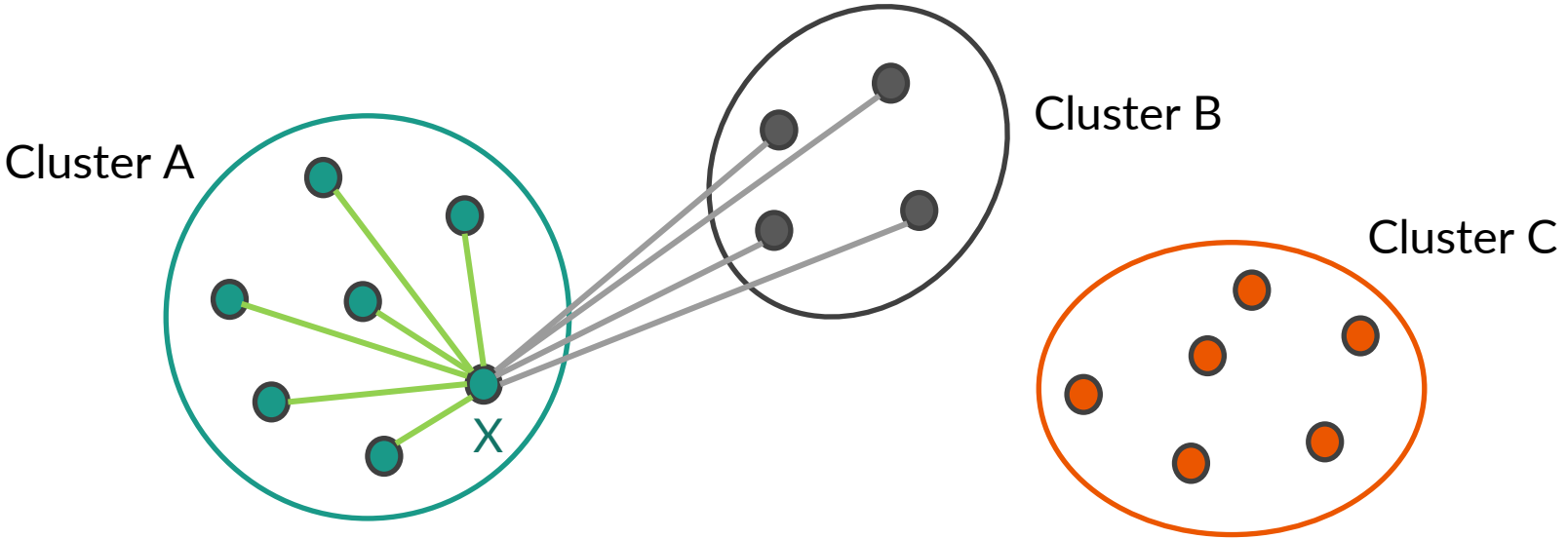
Which is correct?



- Number of within-cluster edges compared to expectation (based on number of nodes and global number of edges)
- Multiple resolutions

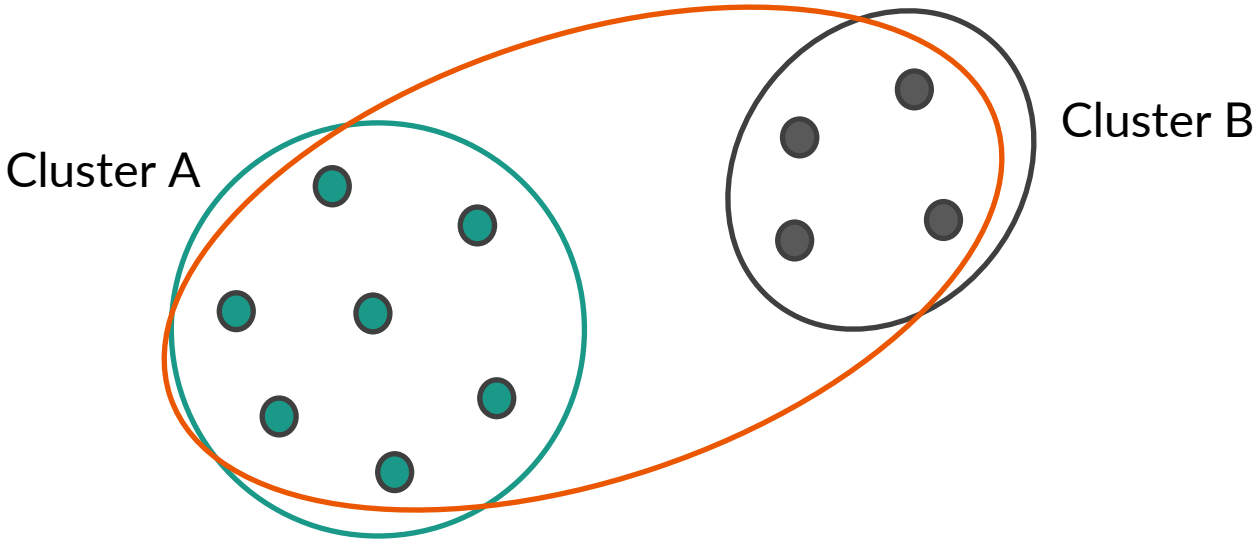
Finding the right number of clusters (tuning clustering parameters)

What is a good clustering result?



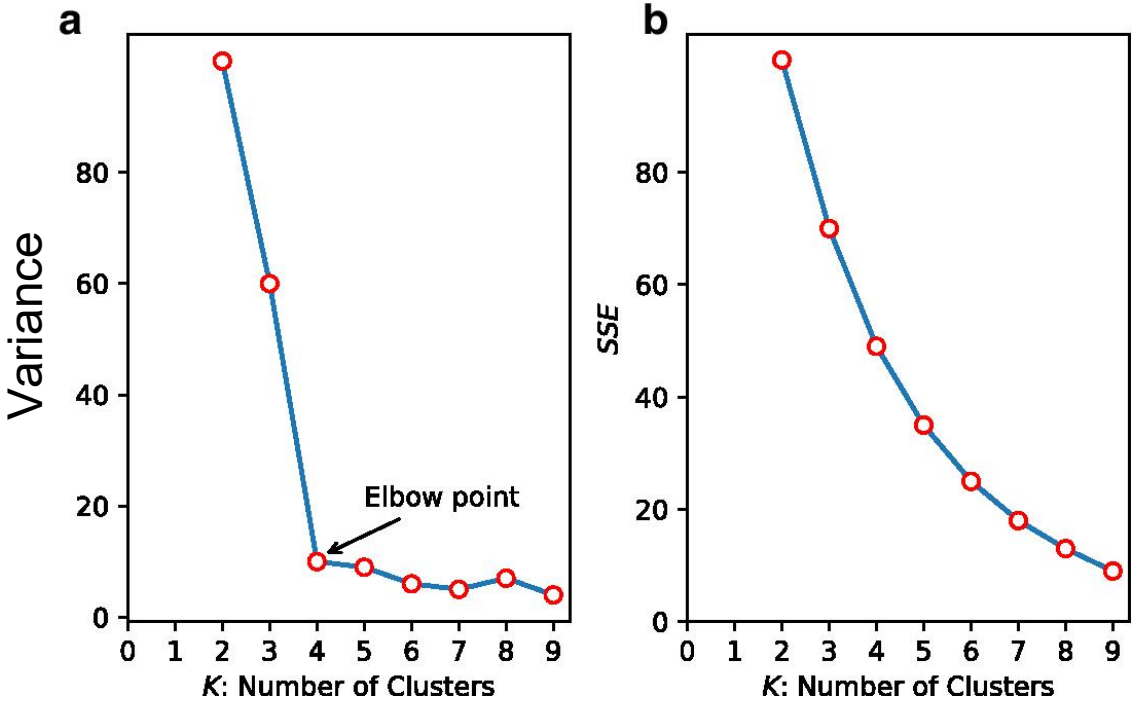
- Distances from **X** to other members in **A** should be smaller than the distances from **X** to members of other clusters (**B** is the closest)

What is a good clustering result?



- The average of variances of A (green) and B (black) should be smaller than the variance of the un-clustered data (orange)

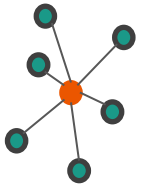
The (obsolete) elbow method



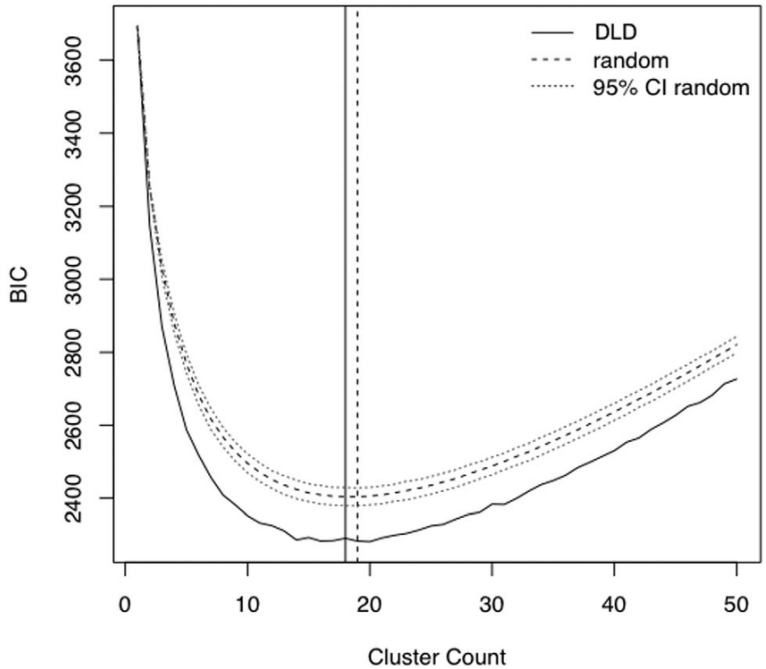
Shi, C. et al. EURASIP J on Wireless Commu and Networking (2021)

- Having more clusters always reduces variance
- Find the number of clusters where further clustering does not provide clearly better result (small reduction in variance)

Using statistical testing



P(data | mean, variance)

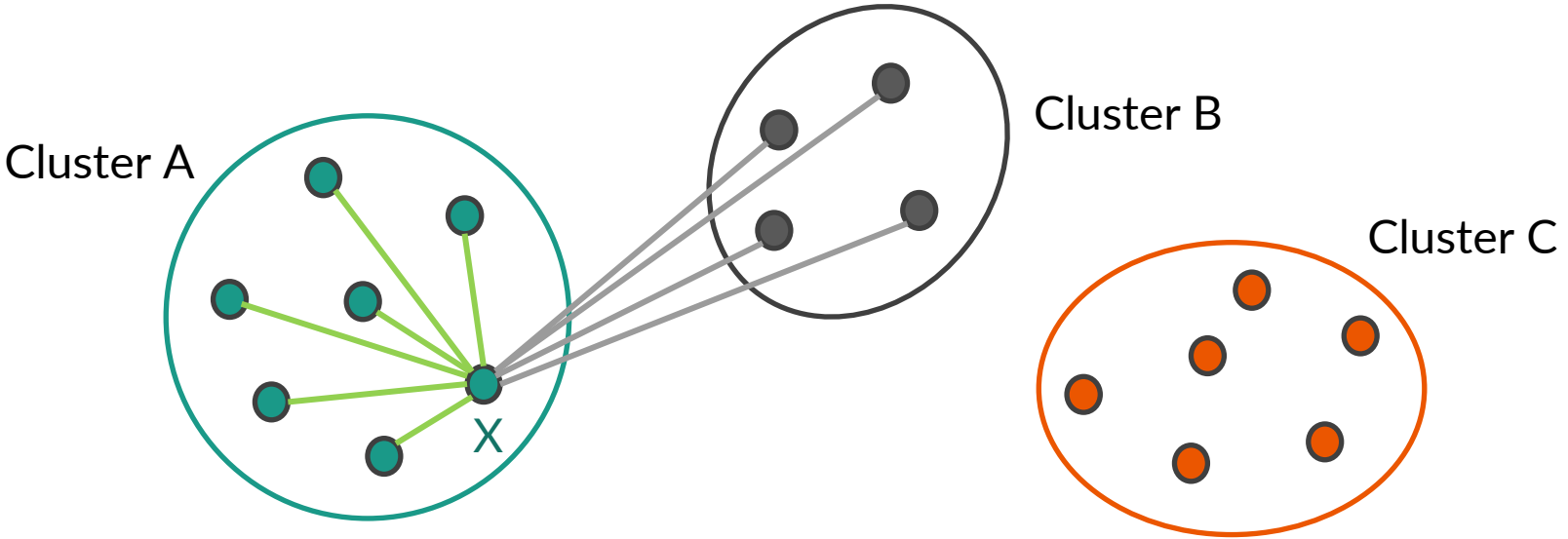


- Each clustering result is a model
- Number of parameters (k) = 2 x number of clusters

- BIC = k ln (N) – 2 ln (likelihood)
- N = number of data points

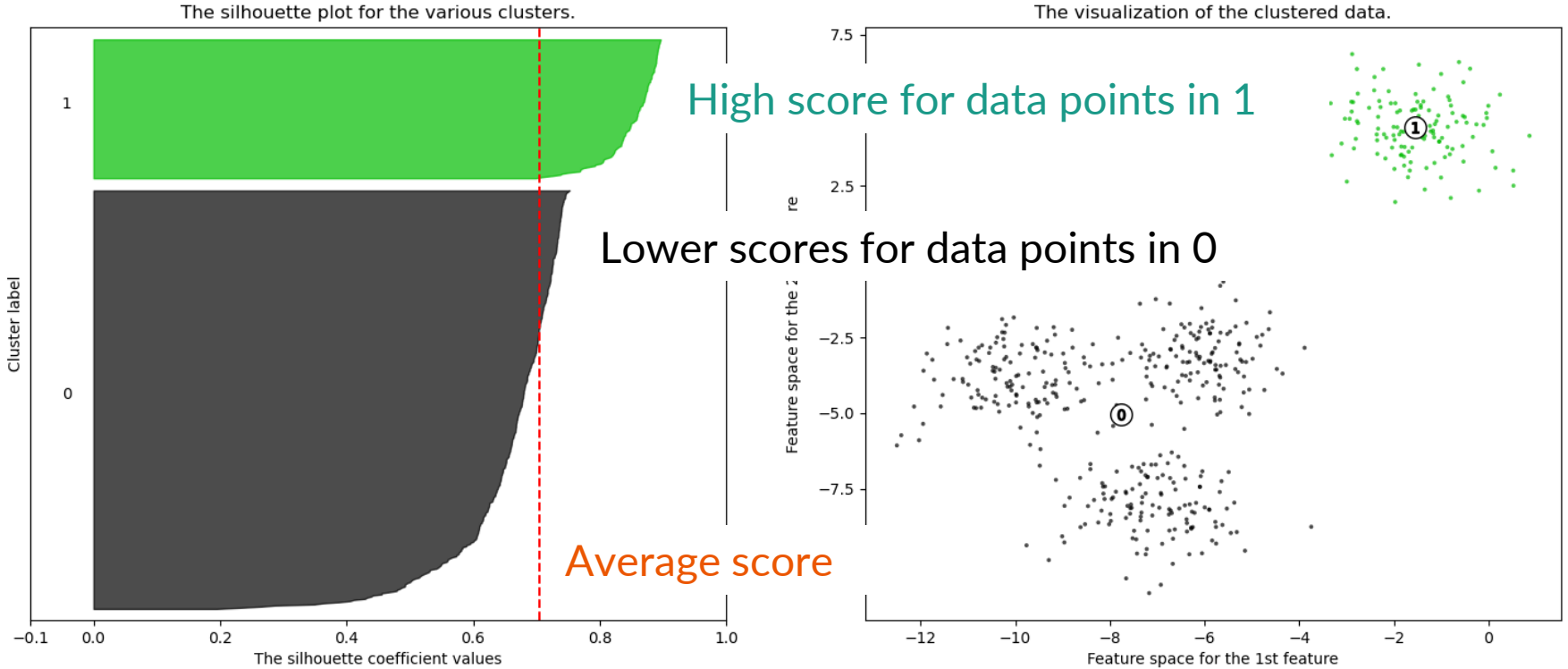
Lancaster, H.S. and Camarata, S. Int J of Lang and Communicable Disorders (2018)

Silhouette score

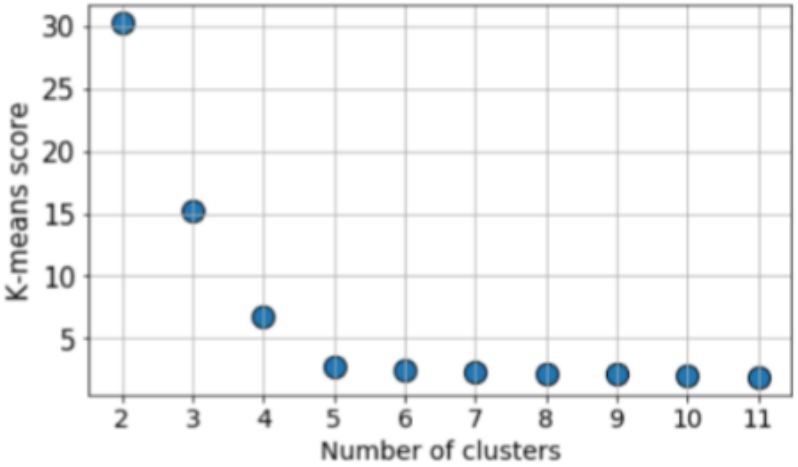


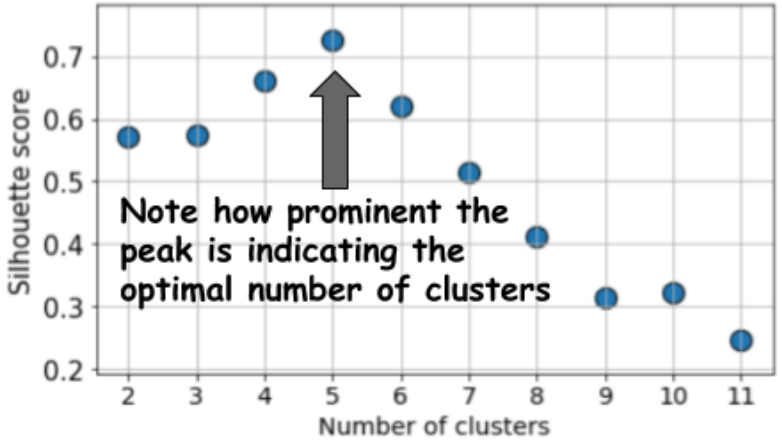
- Silhouette coefficient for X $= \frac{(\text{Avg distance to B} - \text{Avg distance within A})}{\max(\text{Avg distance within A}, \text{Avg distance to B})}$
- High when A and B are far apart, and each is small

Visualization of Silhouette score



Silhouette score is more objective

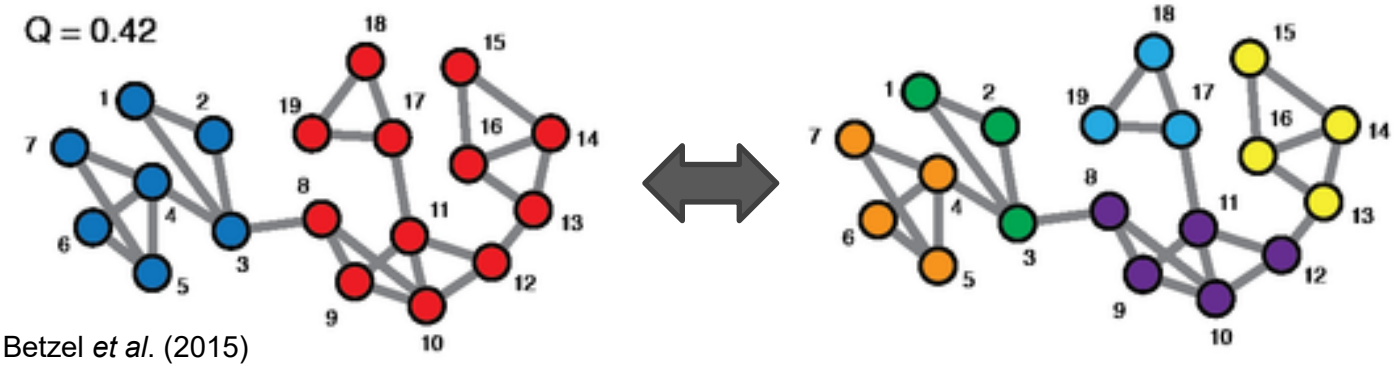




https://www.kdnuggets.com/2019/10/clustering-metrics-better-elbow-method.html

- Still need visual inspection

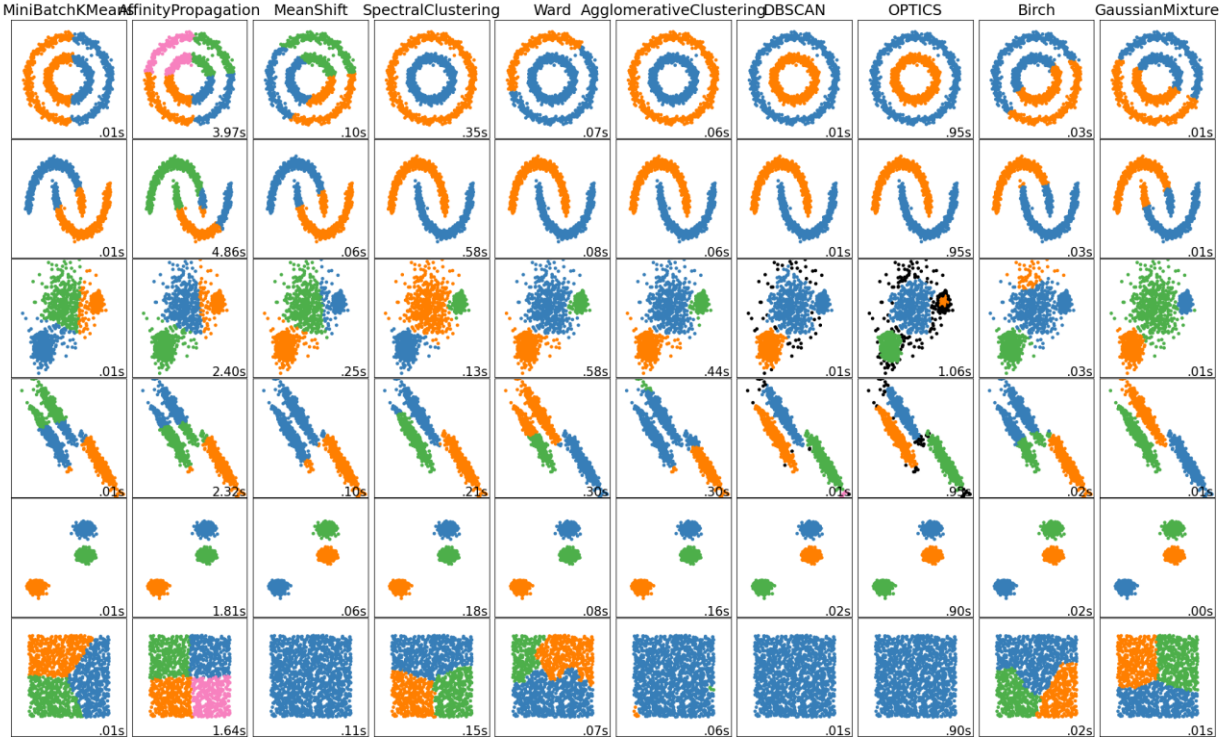
The resolution issue



Betzel *et al*. (2015)

- Algorithms cannot identify clusters of differing size and radii
- Iterative or multi-stage clustering
  - Perform a preliminary clustering
  - Repeat the clustering on each cluster
  - Stop when the optimal number of cluster is 1

No universal solution



https://scikit-learn.org/0.23/auto_examples/cluster/plot_cluster_comparison.html

Any questions?

See you on February 16th