Fundamentals of AI

Clustering

Assessment of clustering quality*

*Some materials in this lecture are used from

https://towardsdatascience.com/unsupervised-machine-learning-clustering-analysis-d40f2b34ae7e

https://gdcoder.com/silhouette-analysis-vs-elbow-method-vs-davies-bouldin-index-selecting-the-optimal-number-of-clusters-for-kmeans-clustering/

Problem with clustering

- Any algorithm will deliver some clusters
- Some data are not naturally ‘clusterable’
- No clustering algorithm exists without parameters affecting the number of detected clusters
- We would like clusters *not to be* the results of statistical fluctuations in the point density
- We should always test how far the discovered structure in the data is from ‘random’ data
- Using built-in optimization criteria is not a good idea

Basic questions

- Did we capture any *non-random* structure in the data?
- Did we choose the parameters (e.g., number of clusters) in the most optimal way?
- Do our clusters make sense in the application domain?

Standard protocol for testing cluster validity

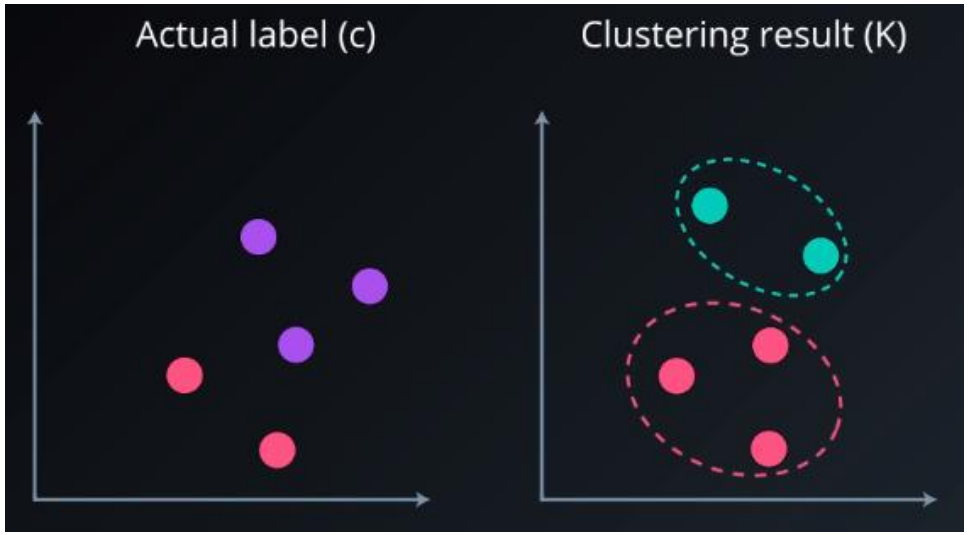
1. Identify the clustering structure and validation type
2. Define a validation index
3. Define a null hypothesis of no structure (for example, *all the locations of data points in the specific regions are equally likely*)
4. Establish the baseline distribution under the null hypothesis (for example, by bootstrapping, resampling or permutations)
5. Calculate the index
6. Test the hypothesis of no structure

‘Supervised’ vs ‘Unsupervised’ clustering validation

- Supervised: How similar is partitioning of the data into *clusters* and *classes*?

- Unsupervised: How ‘compact’ are the obtained clusters?

Rand Index (accuracy of clustering with respect to ‘ground truth’)



- Rand index is the most basic measure of the percentage of correct decisions made by the algorithm
- **a**: is the number of points that are in the same cluster both in C and in K
- **b**: is the number of points that are in the different cluster both in C and in K.
- **n** : is the total number of samples

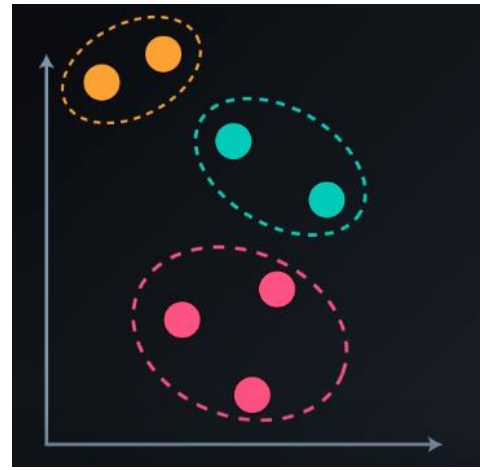
$$\text{Rand Index} = \frac{a + b}{n(n-1)/2}$$

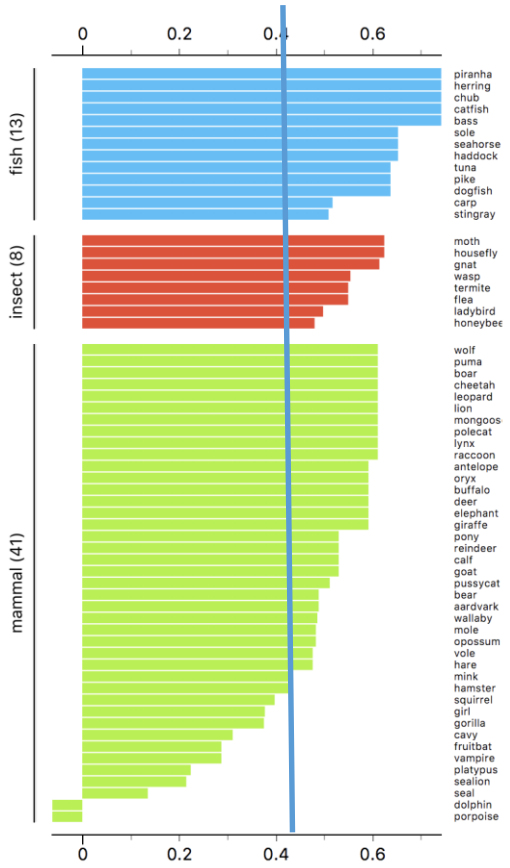
Adjusted Rand Index (ARI) corrects for the case of random label reshuffling

$$\text{ARI} = \frac{\text{RI} - \text{Expected Index}}{\text{Max(RI)} - \text{Expected Index}}$$

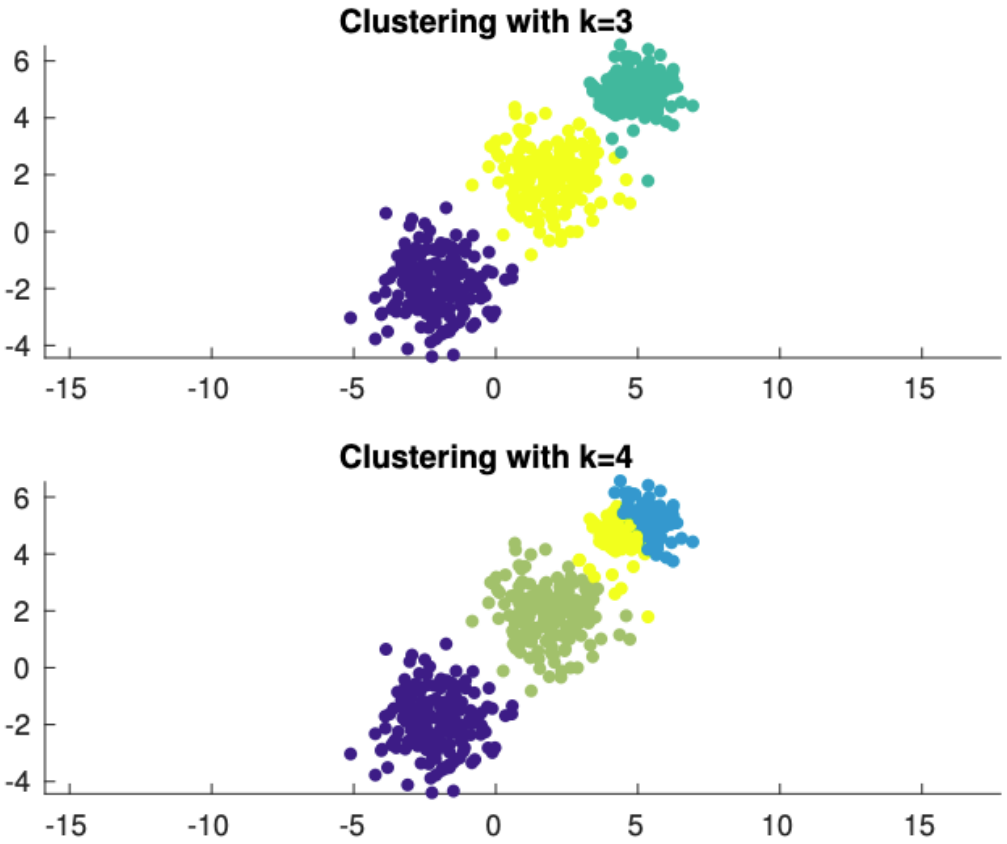
Silhouette

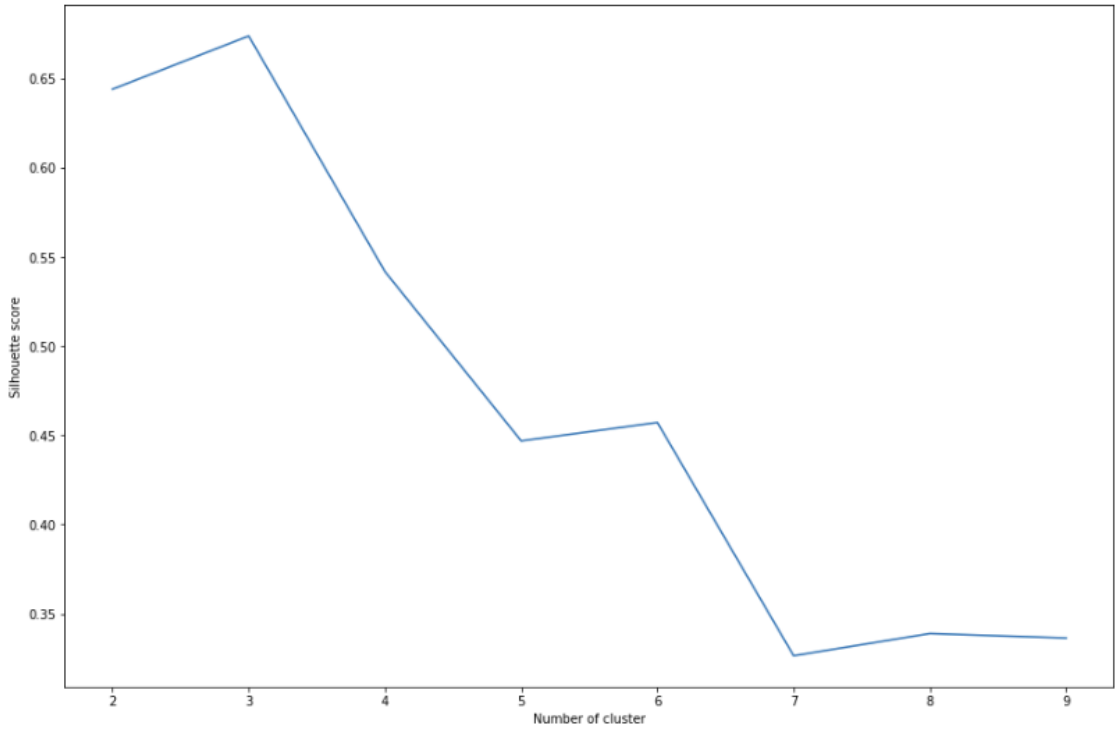
- Suitable for K-means and hierarchical clustering
- a = average distance to other sample in the same cluster
- b = average distance to other sample in closest neighbouring cluster
- Silhouette for a point i: Si = (ai-bi)/max(bi-ai)
- Global silhouette coefficient = average of the sum of Si for each point





Using silhouette to determine the right number of clusters





Other indices (several tens)

- Calinski-Harabasz index
- Davies-Bouldin index
- Dunn index
- Density-Based Clustering Validation index

Consensus clustering and resampling

must do for small number of points!

- How to compare two clustering results?
- Way to find ‘averaged’ clustering, if there are several clustering results available
- Typical application for cluster validation: remove k% of data points, cluster, repeat (resampling)
- Let us generate *H* datasets, then for each *h* = 1…*H*
- Connectivity matrix

$$M^h(i,j) = \begin{cases} 1, & \text{if points i and j belong to the same cluster} \\ 0, & \text{otherwise} \end{cases}$$

- Identicator matrix

$$I^h(i,j) = \begin{Bmatrix} 1, \text{if } i \text{ and } j \text{ are in the same dataset} \\ 0, \text{otherwise} \end{Bmatrix}$$

- Consensus matrix

$$C(i,j) = \left(\frac{\sum_{h=1}^{H} M^h(i,j)}{\sum_{h=1}^{H} I^h(i,j)} \right)$$

Consensus clustering and resampling

must do for small number of points!

- How to compare two clustering results?
- Way to find ‘averaged’ clustering, if there are several clustering results available
- Typical application for cluster validation: remove k% of data points, cluster, repeat (resampling)
- Let us generate *H* datasets, then for each *h* = 1…*H*
- Connectivity matrix

$$M^h(i,j) = \begin{cases} 1, & \text{if points i and j belong to the same cluster} \\ 0, & \text{otherwise} \end{cases}$$

- Identicator matrix

$$I^h(i,j) = \begin{Bmatrix} 1, \text{if } i \text{ and } j \text{ are in the same dataset} \\ 0, \text{otherwise} \end{Bmatrix}$$

- Consensus matrix

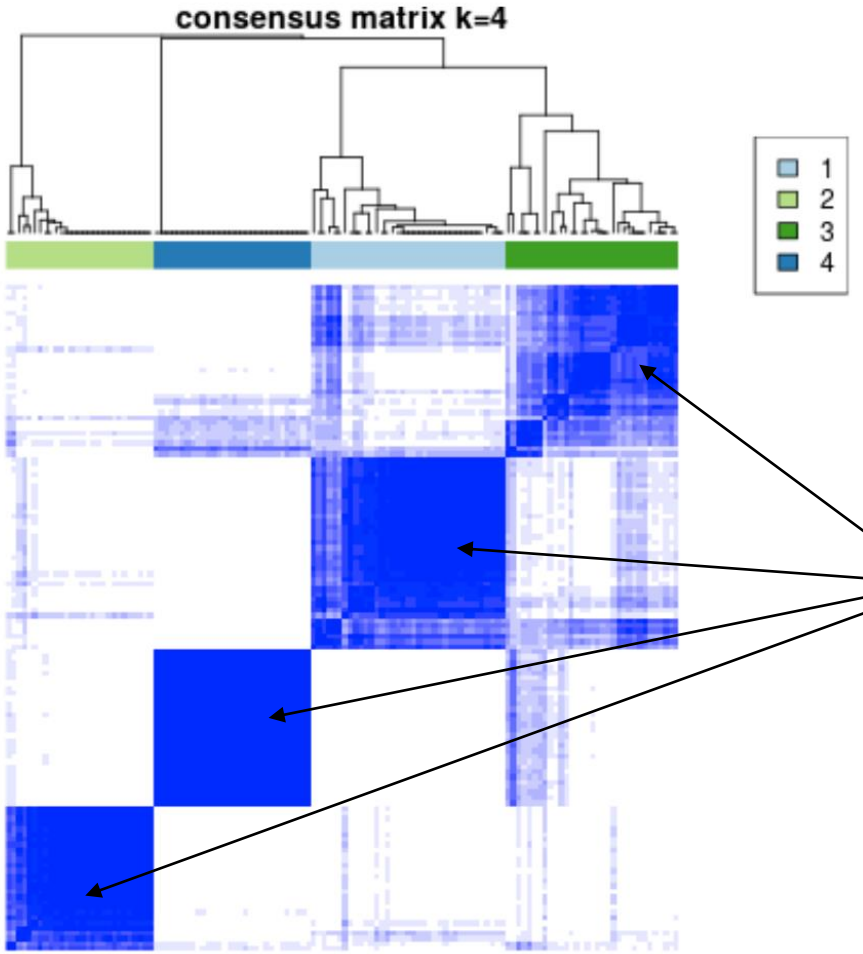
$$C(i,j) = \left(\frac{\sum_{h=1}^{H} M^h(i,j)}{\sum_{h=1}^{H} I^h(i,j)} \right)$$

Consensus clustering and resampling

consensus clusters

$$C(i,j) = \left(\frac{\sum_{h=1}^{H} M^h(i,j)}{\sum_{h=1}^{H} I^h(i,j)} \right)$$

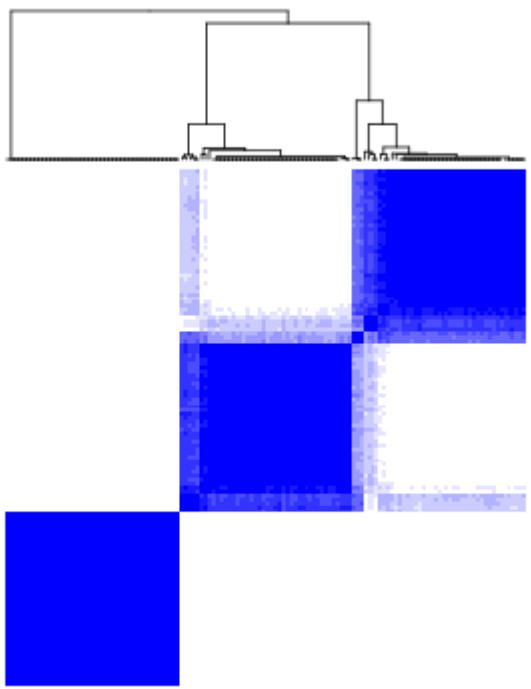
Consensus matrix can be used as a new distance matrix for clustering (e.g., hierarchical clustering)



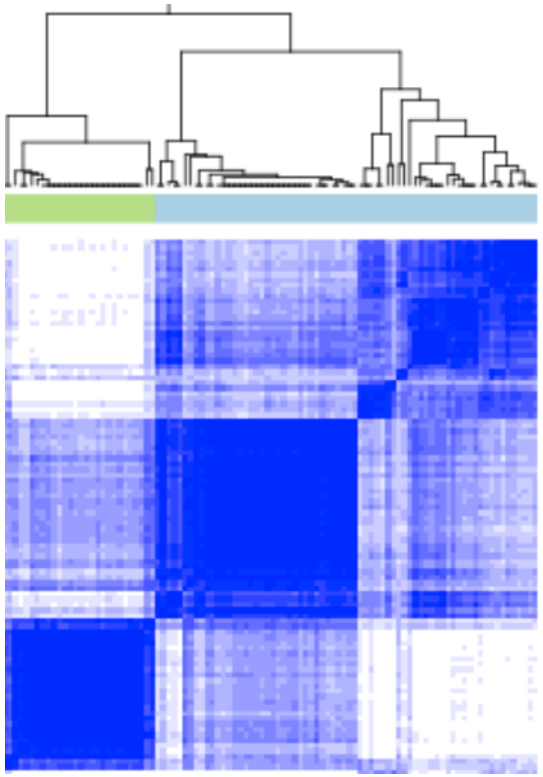
Consensus clusters (their number does not have to be the same as the initial one)

Consensus clustering and resampling

consensus matrix



Good consensus matrix



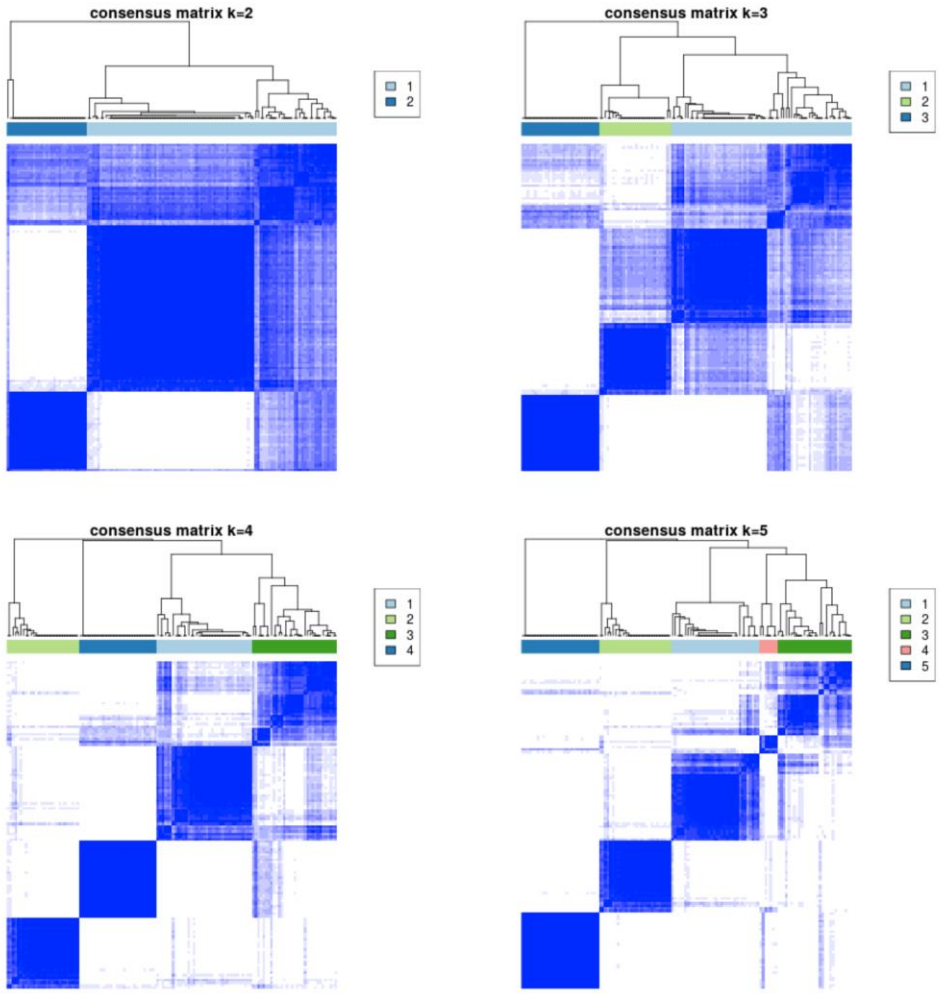
Not that good consensus matrix

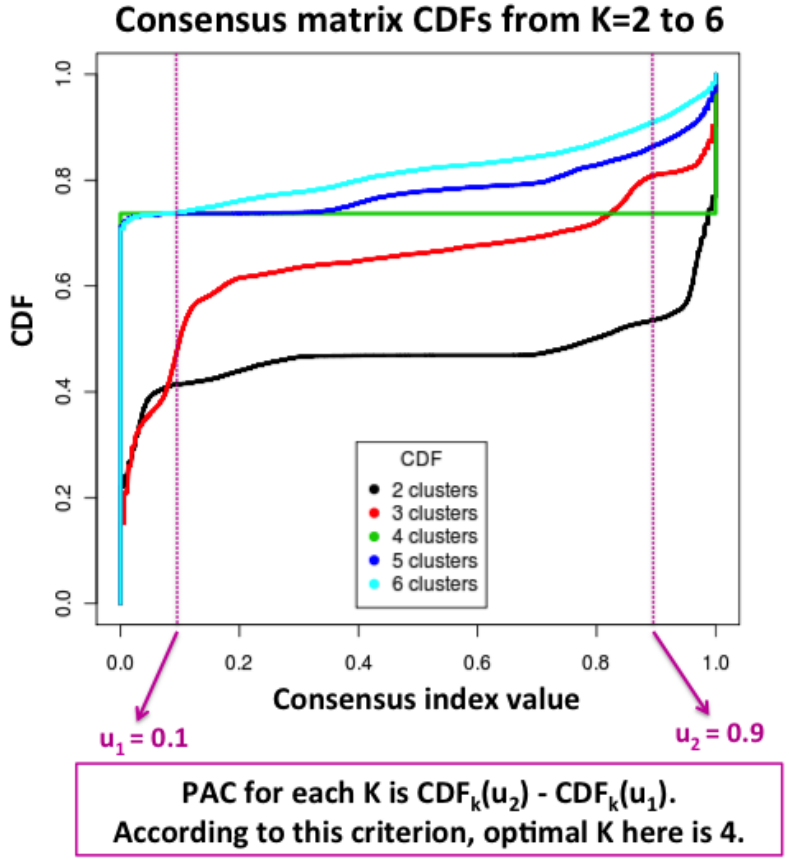
How to quantify the goodness?

Consensus clustering and resampling

Cumulative Density Function (CDF) of consensus matrix

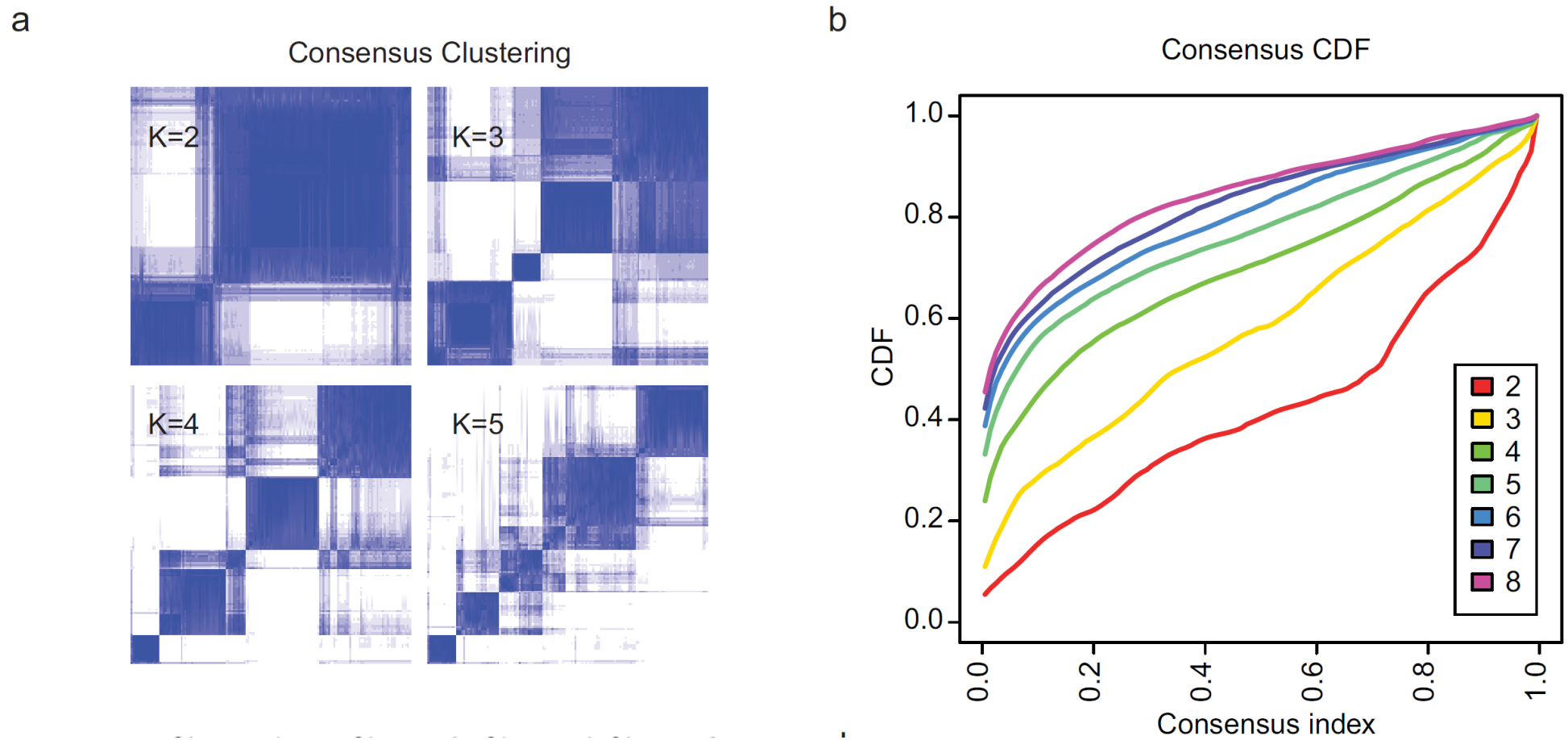
Proportion of ambiguous clustering (PAC)





PAC measures the ‘contrast’of the consensus matrix

Consensus clustering and resampling



Li et al, Scientific Reports, 2014

What you should take with you from ‘Clustering’ lecture

- Distinction between clusters and classes
- Main characteristics of a clustering algorithm
- Major clustering algorithm types
- Good understanding of what k-means, hierarchical clustering and DBSCAN do
- What is the purpose of cluster validation
- Understanding Rand, silhouette index and the principle of consensus clustering