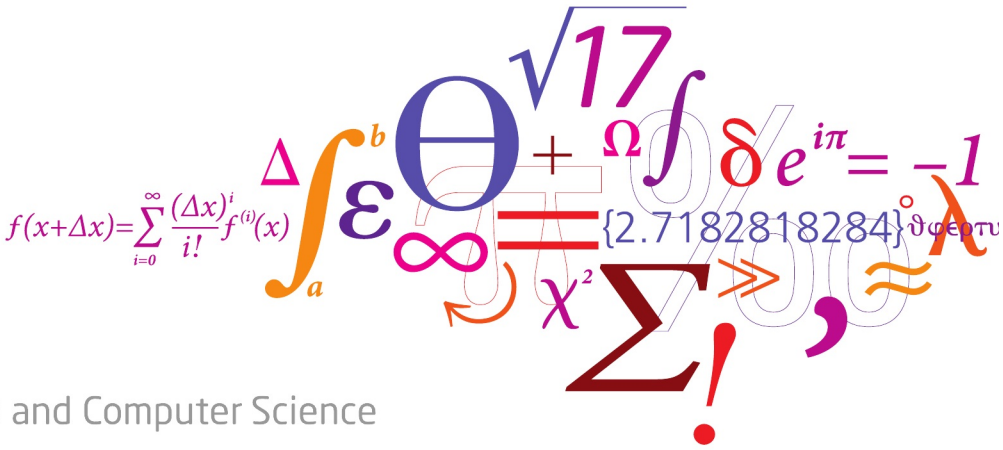
02450: Introduction to Machine Learning and Data Mining

K-means and hierarchical clustering

Tue Herlau

DTU Compute, Technical University of Denmark (DTU)



DTU Compute
Department of Applied Mathematics and Computer Science

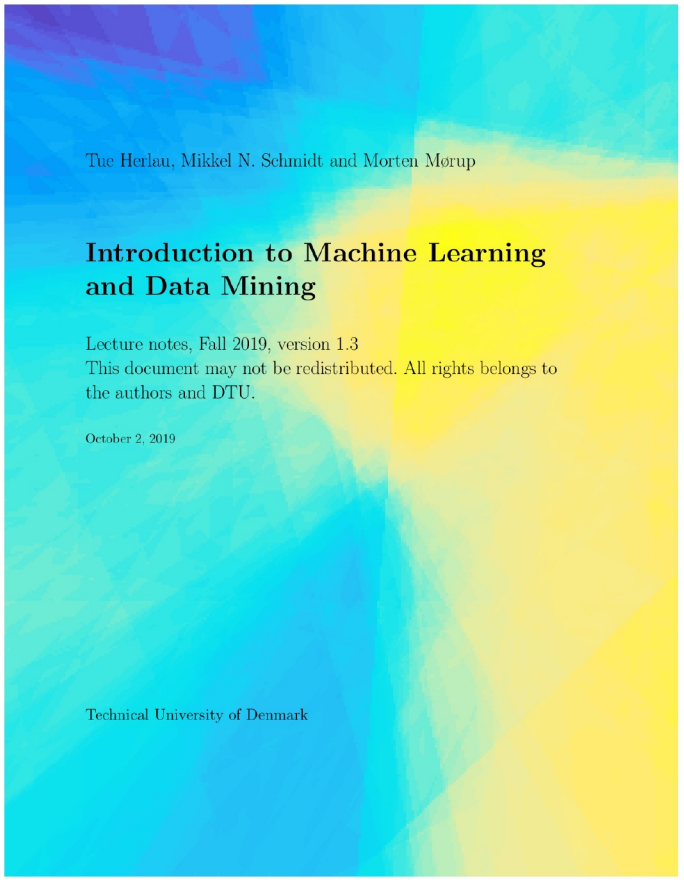
Today

Feedback Groups of the day:

Dmitri Khlebutin, Marie De Vignes De Puylaroque, Katrine Rimmen Thomsen, Thomas Alexander Storm, Anna Maria Clara Schibelle, Nikolaj Abakumov, Frederik Giediesak Probst, Victor Bang Malmkjær, Thomas Wiklund Jørgensen, Linam Eleonora Larsen, Jón Jacobsen, Xingguo Sun, Kaiyue Xu, Fredrik Mette Myklebust, Lars Christian Jacobsen, Marie Højmark Pedersen, Kirstine Birk Petersen, Kyle Hutto, Yuchen Li, Baixun Wang, Johan-Frederik Nielsen, Shuting Yang, Piriya Sureshkumar, Malou Meisen Lokdam, Tor Krøjer Toudahl, Serdar Tunçkol, Adrian-Nicolae Stefan, David Bilde Lang, Jens Håkon Visbech Christensen, Mikkel Mondrup Lynggaard, César Alejandro Torrealba Vázquez, Thomas Halkier Nicolajsen, Magnus Hindborg Hovmann, Bernadette Kofoed Christiansen, Martin Hemmingsen, Jeppe Nygård Samuelsen, Victor Juarez Guajardo-Fajardo, Sayuri Tais Miyamoto Magnabosco, Nikolaj Søndergaard Povlsen, Rasmus Juul Pedersen, Ellen Winther Mortensen, Julius Pulvertaft Rasmussen, Jasmina Tokmic, Jonne Kaunisto, Saujanya Mulukutla, Stefan Skrydstrup Pedersen, Spardha Virendra Jhamb, Markus Hans Kristofer Johansson, Thomas Nygaard Nilsson, Nicolai Nykvist, Steffen Rasch Olsen, Kseniya

Reading material:

Chapter 18



Lecture Schedule

1. Introduction
   3 September: C1

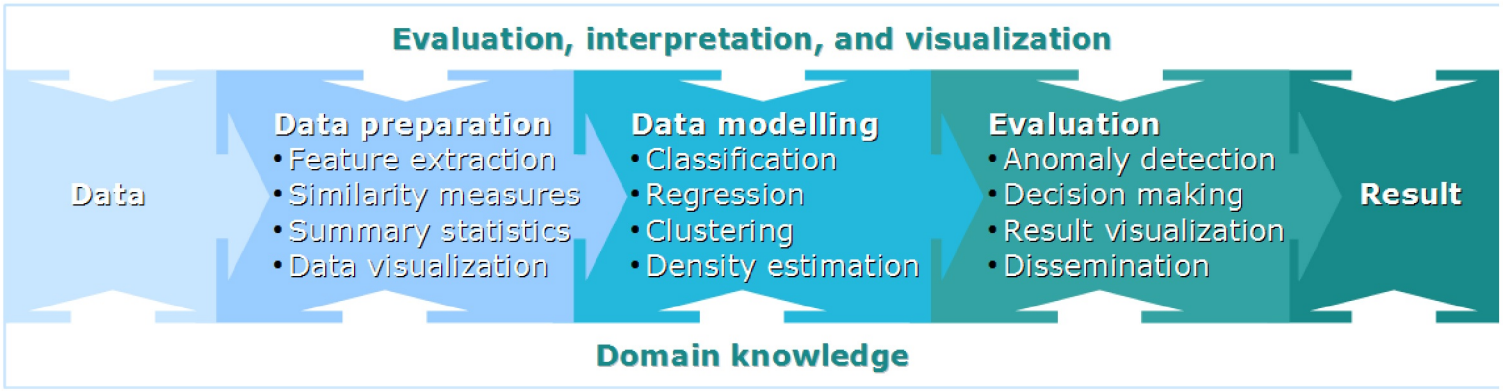
   Data: Feature extraction, and visualization
2. Data, feature extraction and PCA
   10 September: C2, C3
3. Measures of similarity, summary statistics and probabilities
   17 September: C4, C5
4. Probability densities and data Visualization
   24 September: C6, C7

   Supervised learning: Classification and regression
5. Decision trees and linear regression
   1 October: C8, C9 **(Project 1 due before 13:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   8 October: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   22 October: C11, C13
8. Artificial Neural Networks and Bias/Variance
   29 October: C14, C15
9. AUC and ensemble methods
   5 November: C16, C17

   Unsupervised learning: Clustering and density estimation
10. **K-means and hierarchical clustering**
    12 November: C18 **(Project 2 due before 13:00)**
11. Mixture models and density estimation
    19 November: C19, C20
12. Association mining
    26 November: C21

    Recap
13. Recap and discussion of the exam
    3 December: C1-C21 **(Project 3 due before 13:00)**

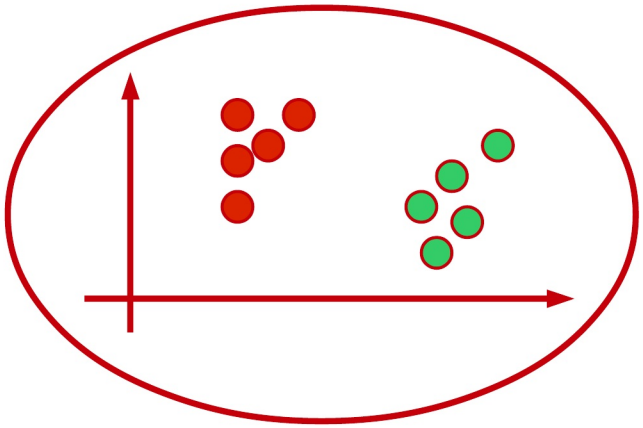
Piazza online help: `https://piazza.com/dtu.dk/fall2019/02450`
Videos/streaming of lectures: `https://video.dtu.dk`



Learning Objectives

- Understand the principles behind K-means and hierarchical clustering
- Understand how different linkage functions affects clustering types
- Evaluate clustering quality using class label information

Supervised and Unsupervised learning

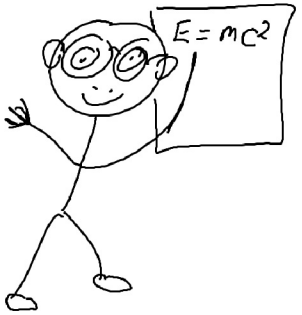


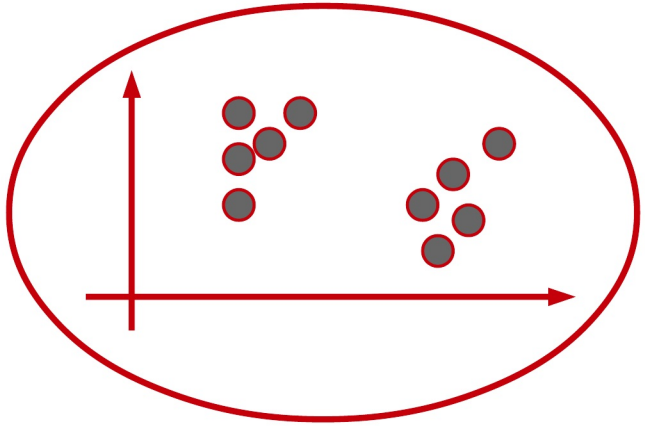
Supervised Learning

Input data $\mathbf{x}_n$ and output y_n

$f(x) \rightarrow y$.

(Classification and Regression)



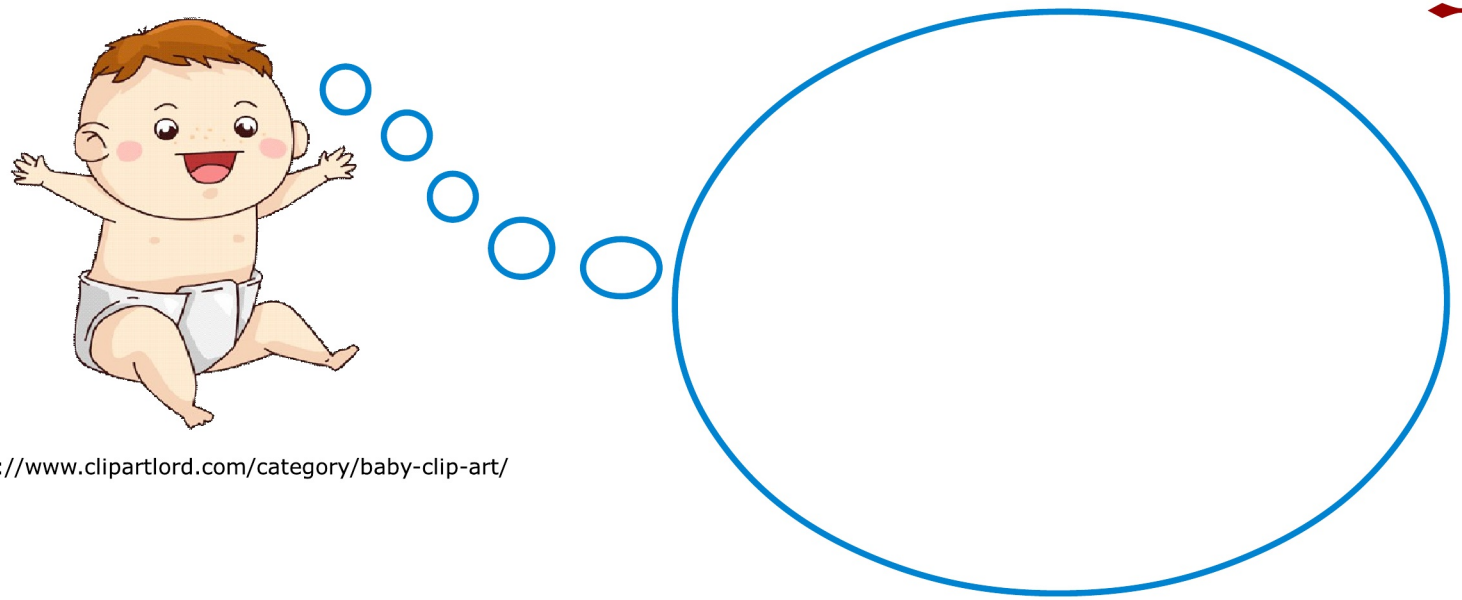


Unsupervised Learning

Input data $\mathbf{x}_n$ alone

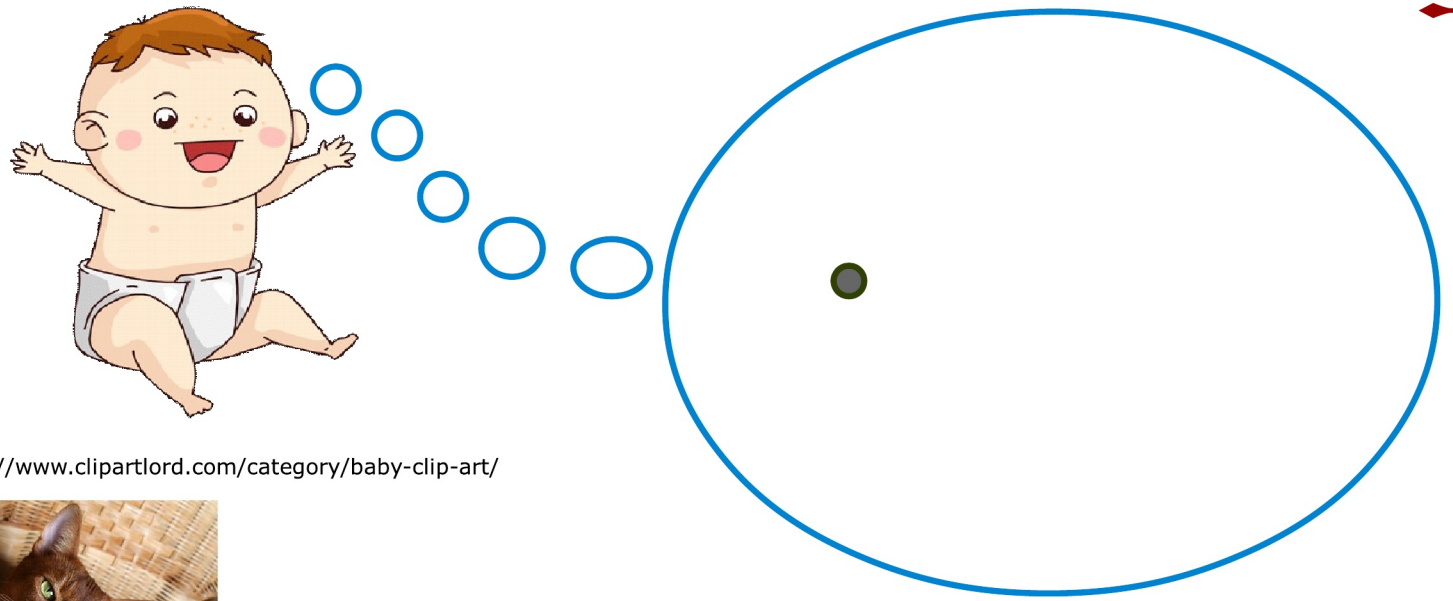
(Exploratory analysis)

Imagine you observe the world for the first time!



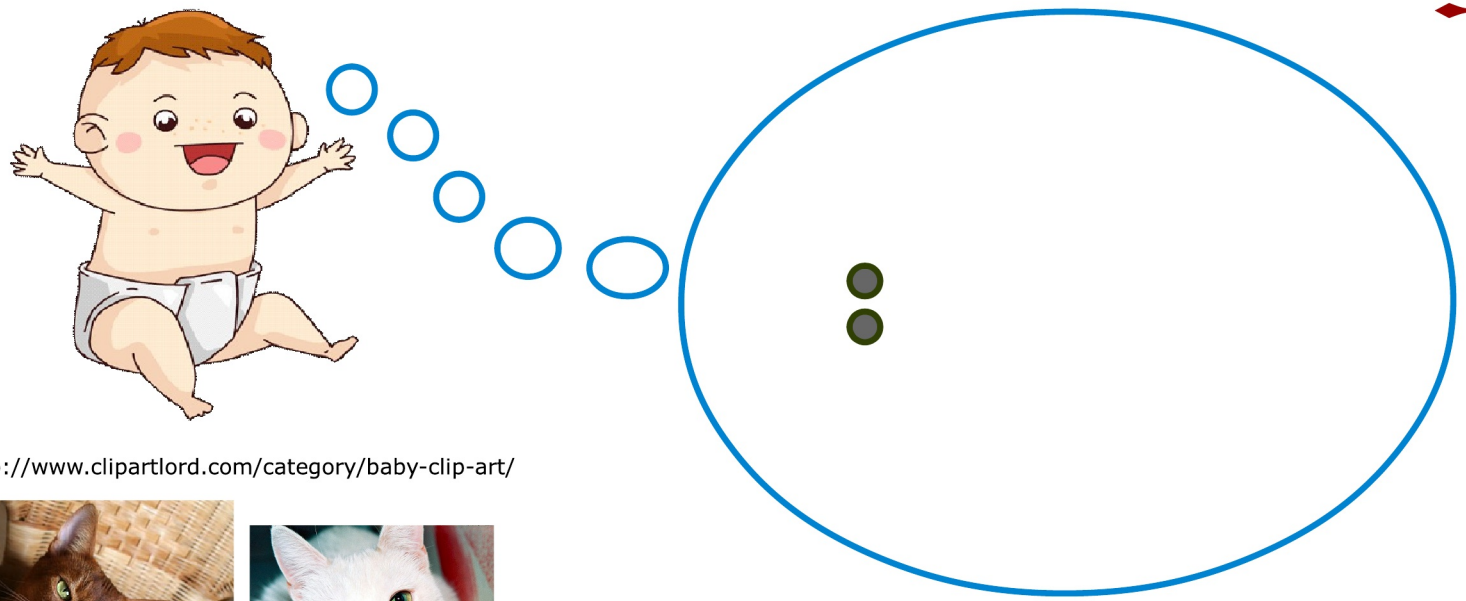
http://www.clipartlord.com/category/baby-clip-art/

Imagine you observe the world for the first time!



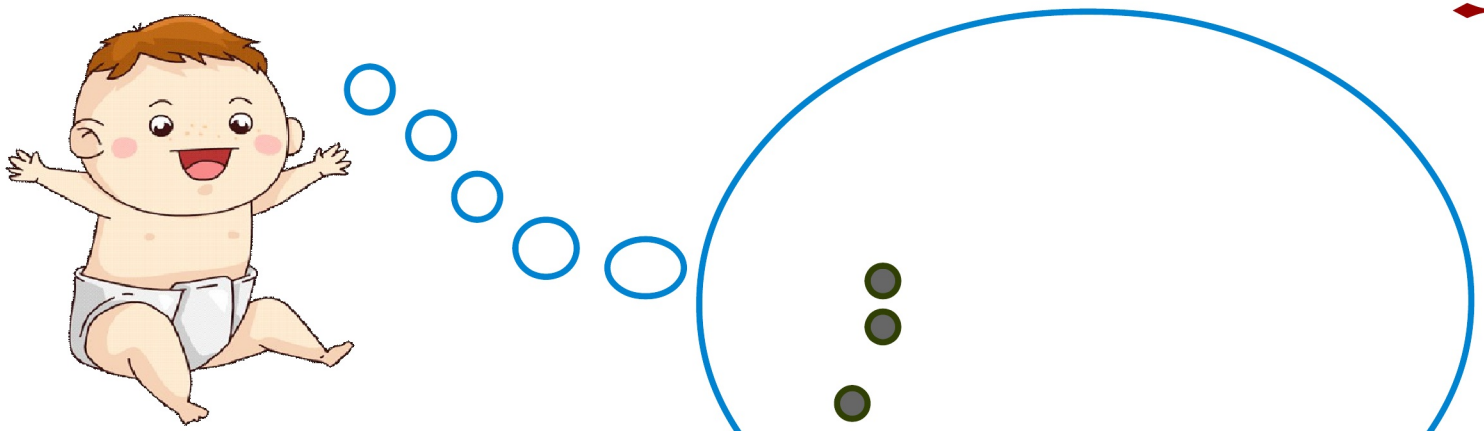
http://www.clipartlord.com/category/baby-clip-art/

Imagine you observe the world for the first time!

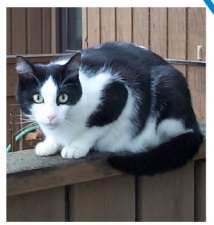


http://www.clipartlord.com/category/baby-clip-art/

Imagine you observe the world for the first time!

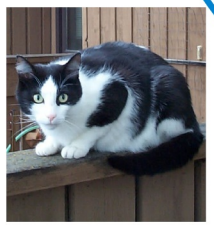


http://www.clipartlord.com/category/baby-clip-art/

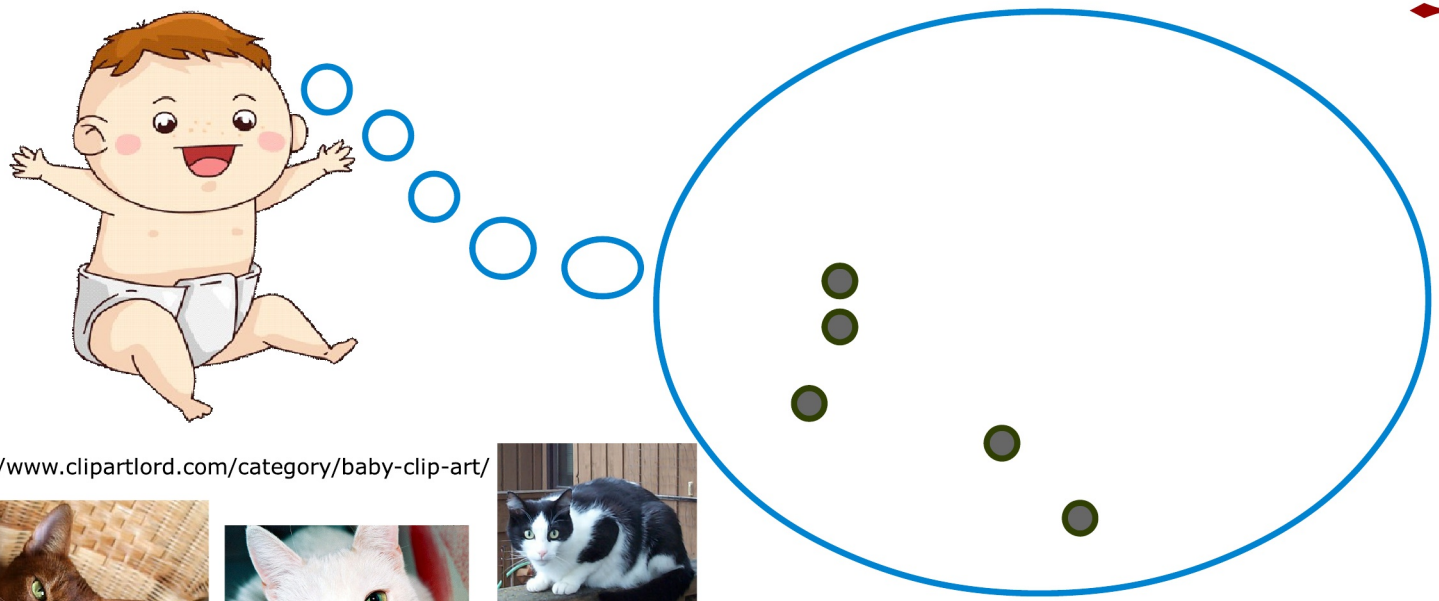


Imagine you observe the world for the first time!

http://www.clipartlord.com/category/baby-clip-art/

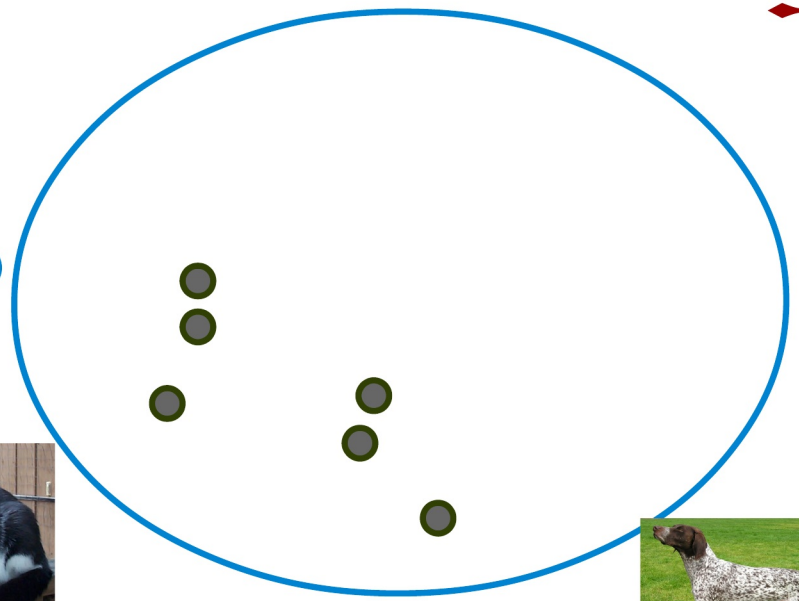


Imagine you observe the world for the first time!

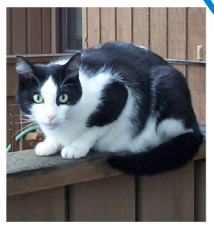


http://www.clipartlord.com/category/baby-clip-art/

Imagine you observe the world for the first time!

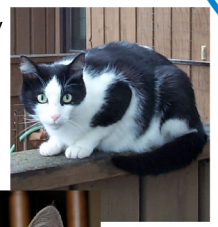


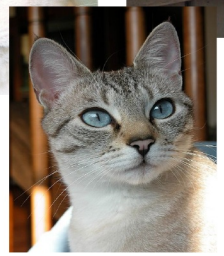
http://www.clipartlord.com/category/baby-clip-art/



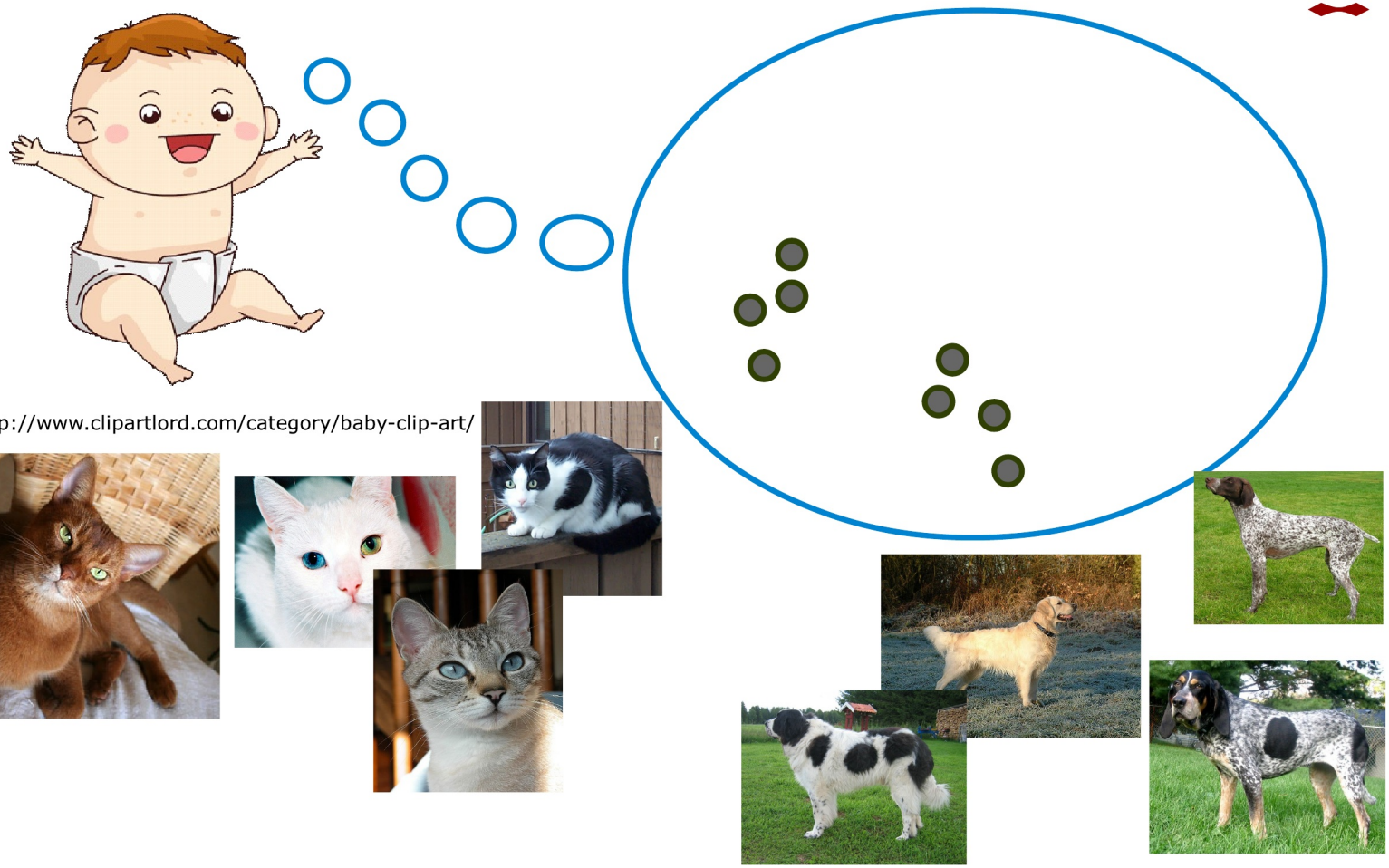
Imagine you observe the world for the first time!

http://www.clipartlord.com/category/baby-clip-art/





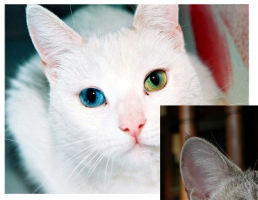
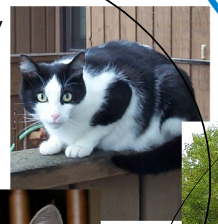
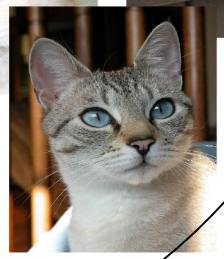
Imagine you observe the world for the first time!

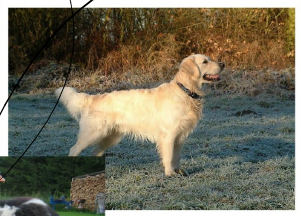


http://www.clipartlord.com/category/baby-clip-art/

Imagine you observe the world for the first time!

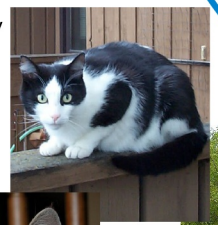
http://www.clipartlord.com/category/baby-clip-art/

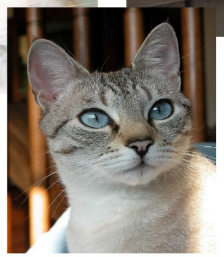







Imagine you observe the world for the first time!

http://www.clipartlord.com/category/baby-clip-art/





We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
http://commons.wikimedia.org/wiki/File:MasPiri-Astro-SVE.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg
http://commons.wikimedia.org/wiki/File:Cat002.jpg
https://commons.wikimedia.org/wiki/Dog#/media/File:BluetickCoonhound.jpg
http://commons.wikimedia.org/wiki/File:Saurier2.jpg

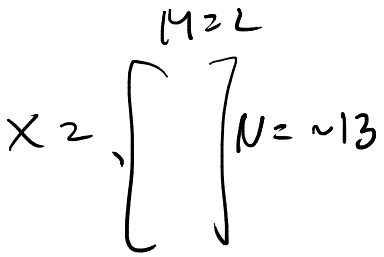
Unsupervised learning

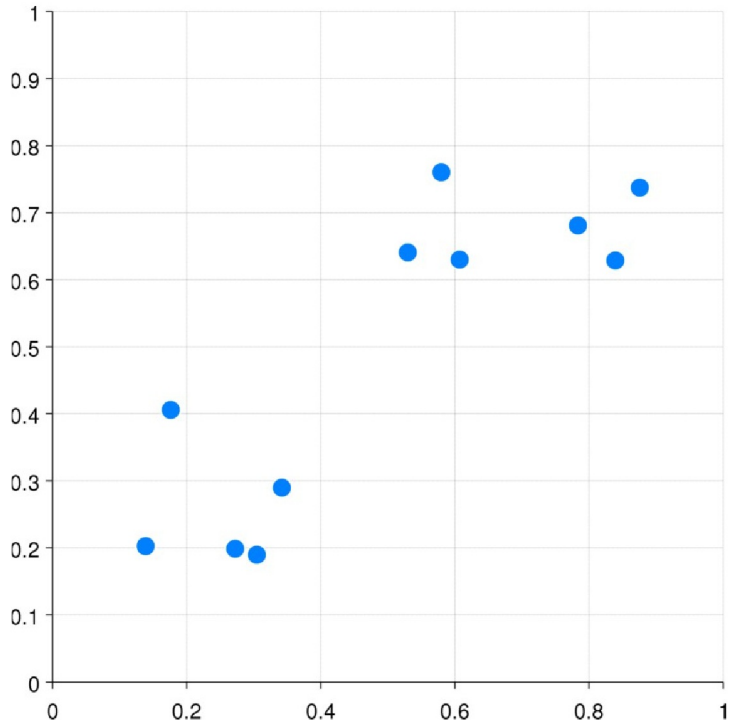
- **Supervised learning**
  - Use the data to learn the output values
- **Unsupervised learning**
  - No output variables available
  - Sometimes called exploratory analysis
  - What to learn from the data?
    - Structure
    - Regularities
    - Hidden information
    - Etc.

Clustering

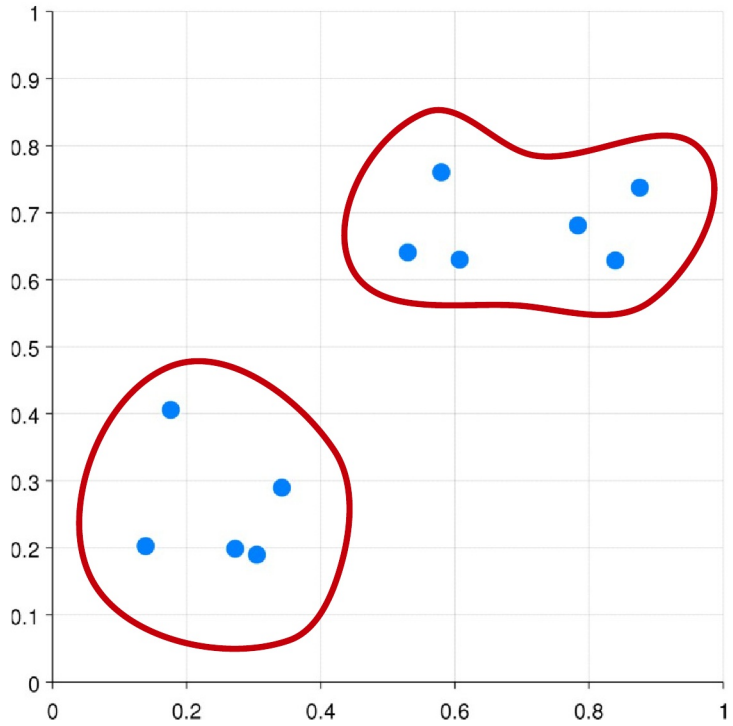
- Divide data into groups (subsets/clusters) that are
  - **Meaningful**: Capture the natural structure of the data
  - **Useful**: Depends on purpose
- Observations in the same cluster are **similar in some sense**
- Unsupervised classification

Clustering

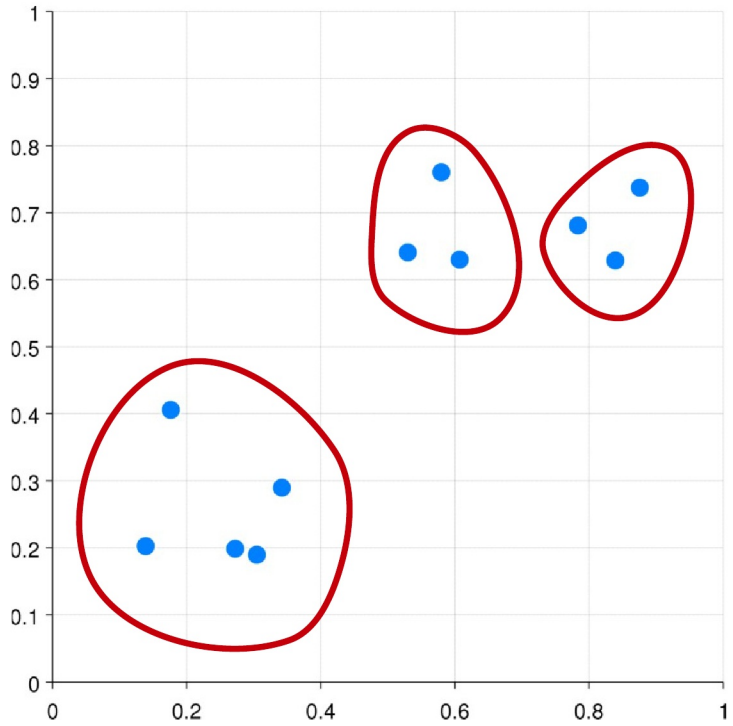




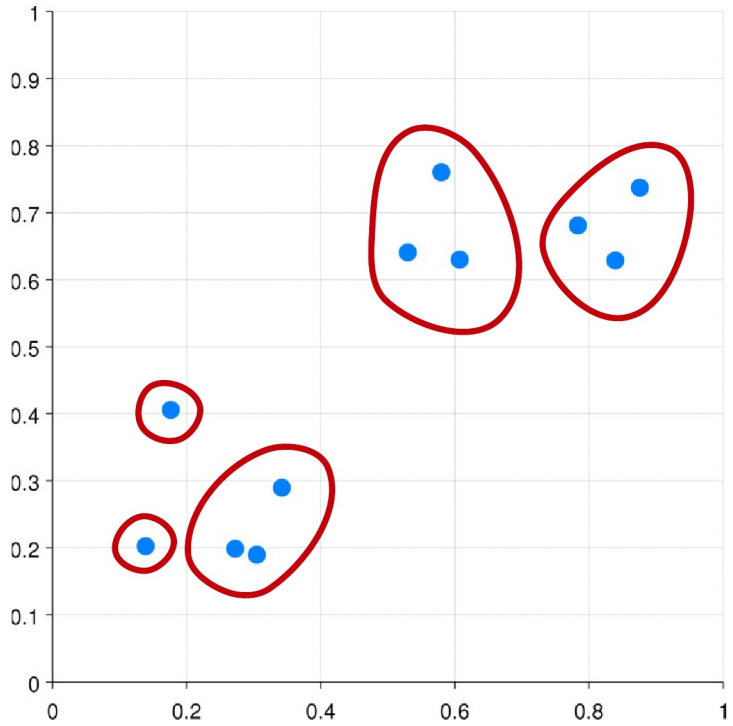
Clustering



Clustering

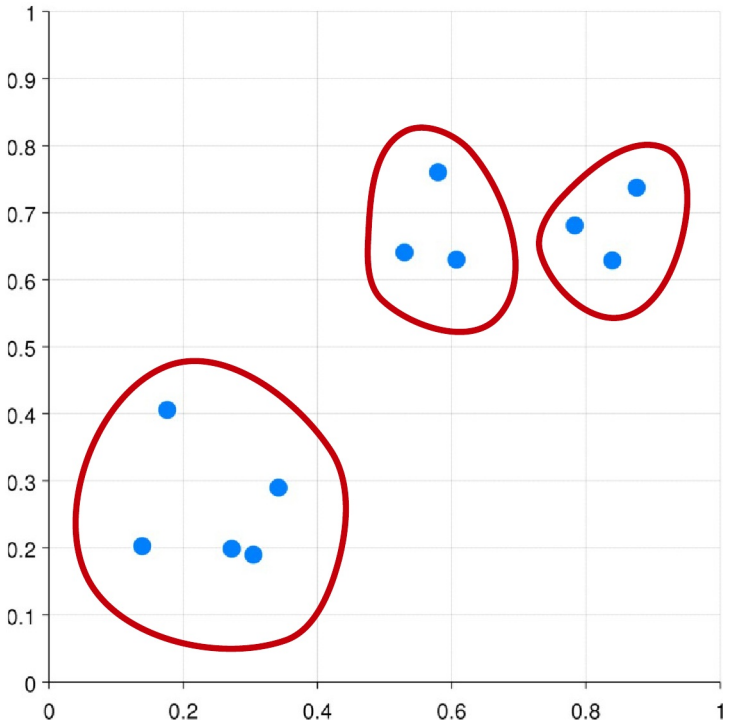


Clustering

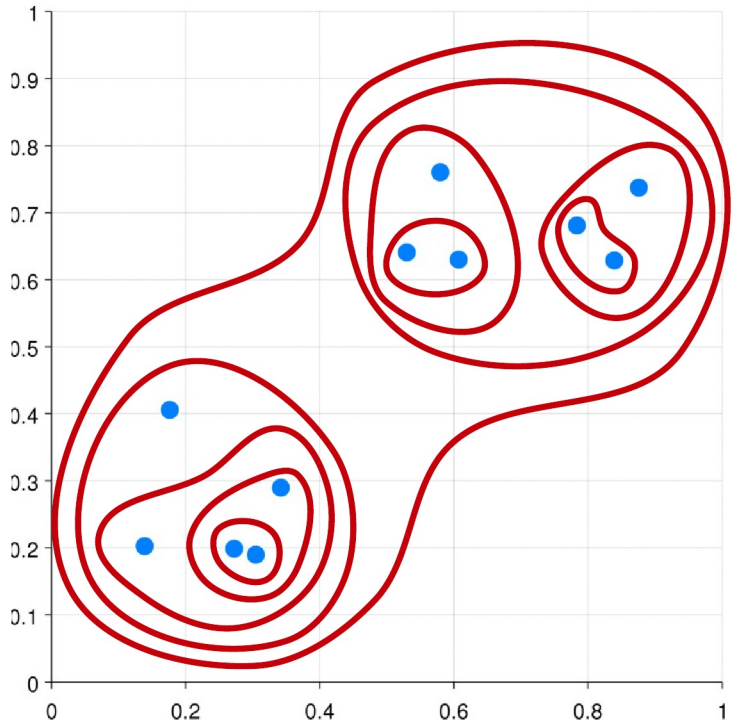


Partitional / hierarchical clustering

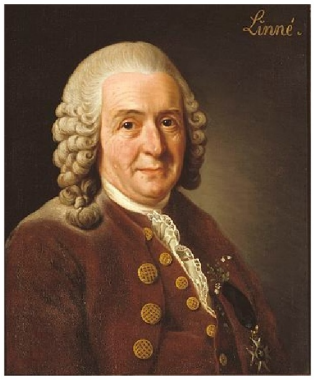
Partitional



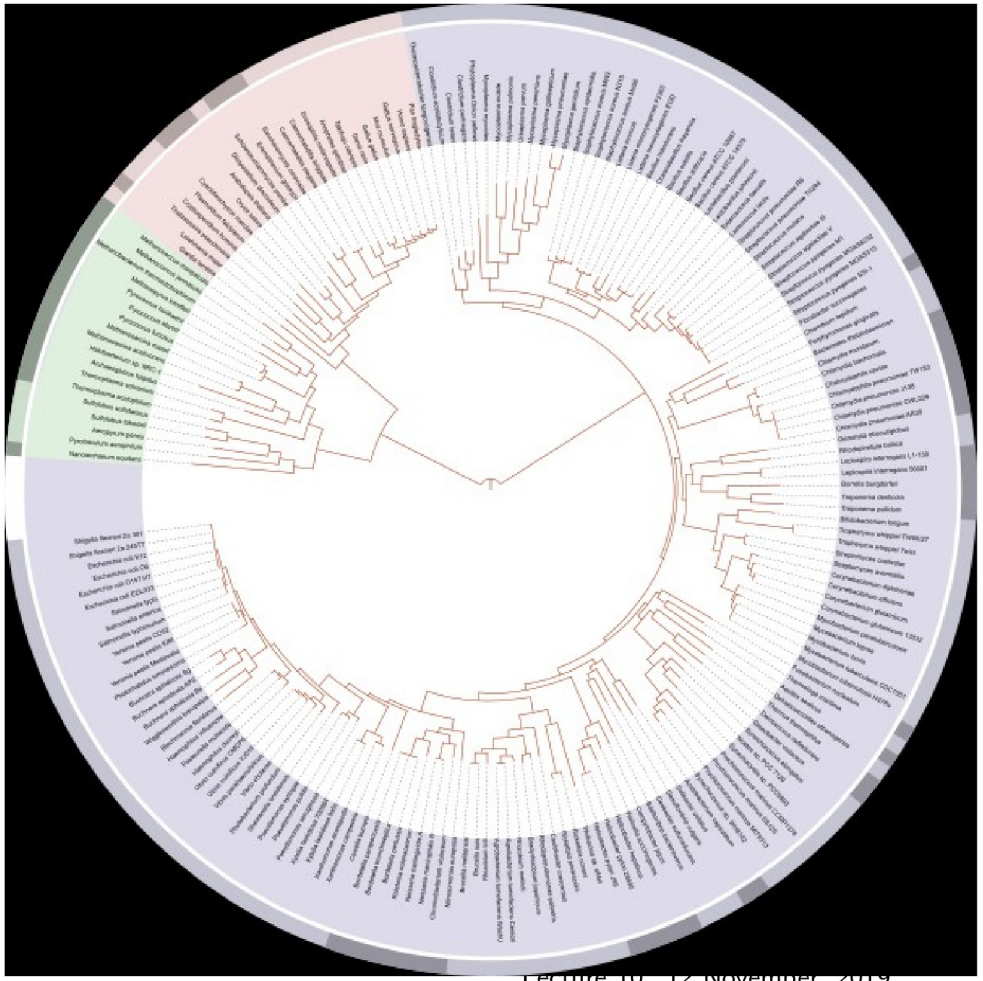
Hierarchical



Phylogenetic trees may be considered a type of hierarchical clustering



Carl Linnaeus
(1707 – 1778)
http://en.wikipedia.org/wiki/Carl_Linnaeus

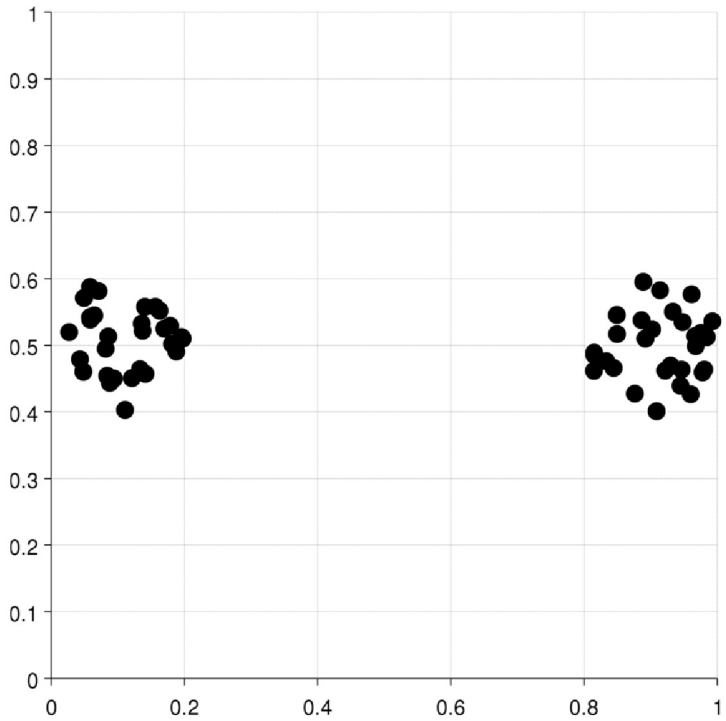


http://en.wikipedia.org/wiki/File:Tree_of_life_SVG.svg

Types of clustering

Well-separated

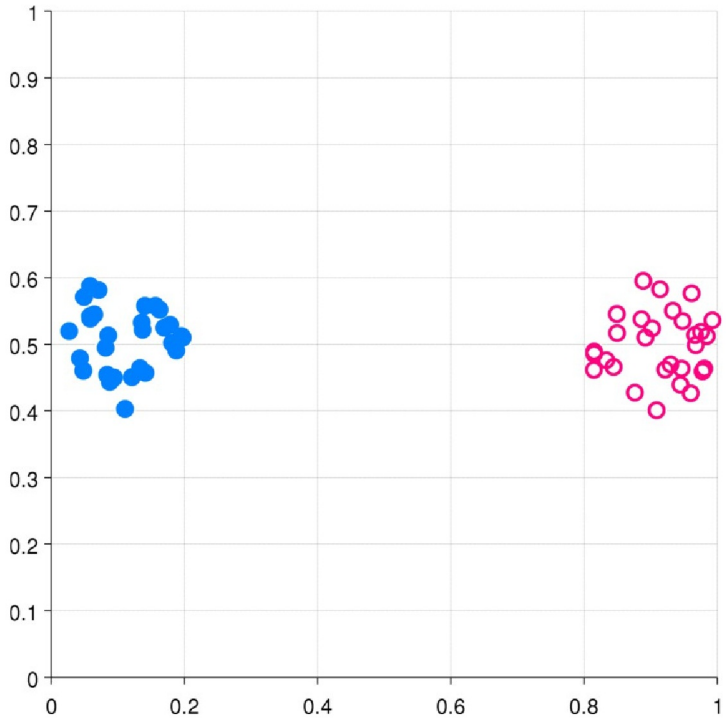
- Each point is closer to all points in its cluster than any point in another cluster



Types of clustering

Well-separated

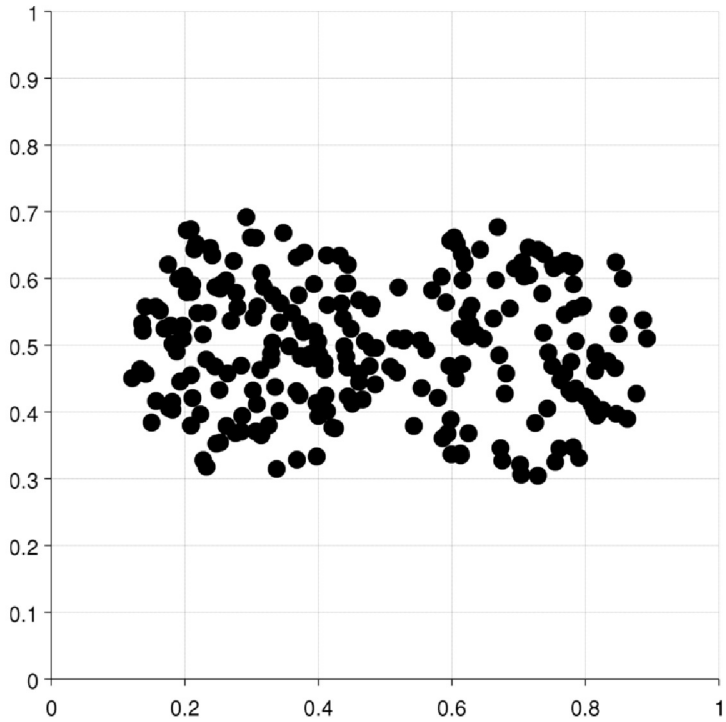
- Each point is closer to all points in its cluster than any point in another cluster



Types of clustering

Center-based

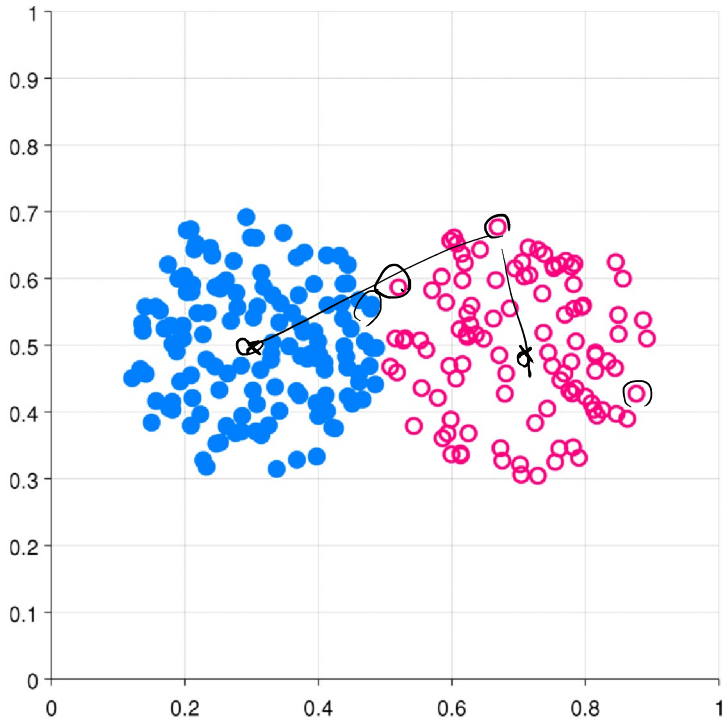
- Each point is closer to the center of its cluster than to the center of any other cluster



Types of clustering

Center-based

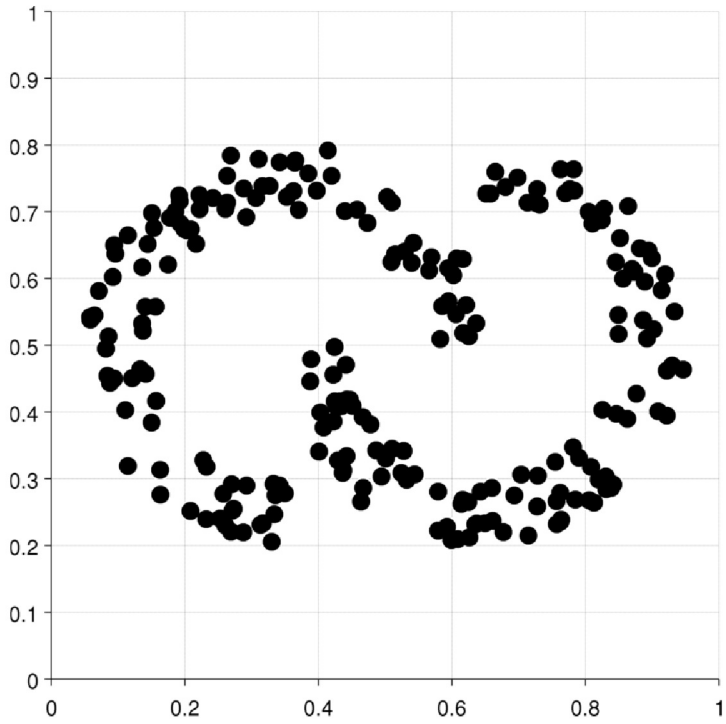
- Each point is closer to the center of its cluster than to the center of any other cluster



Types of clustering

Contiguity-based

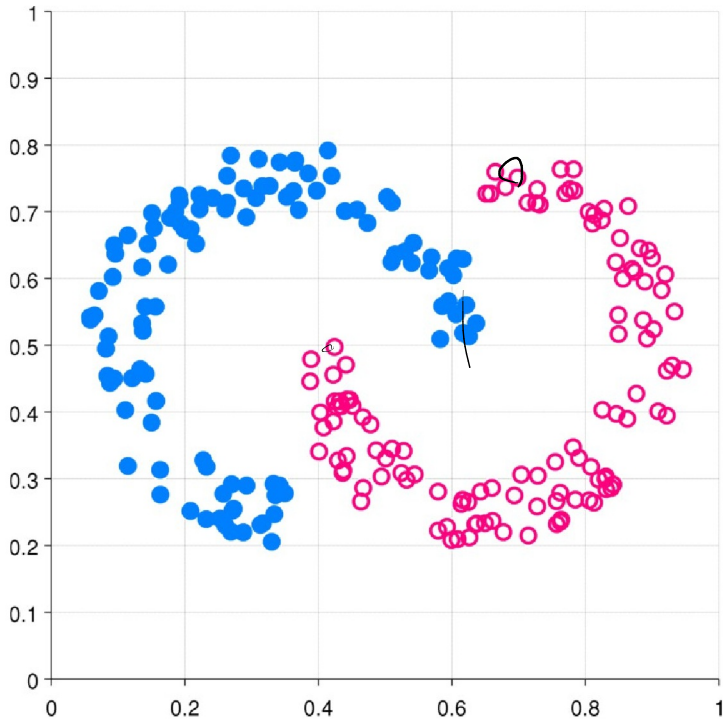
- Each point is closer to at least one point in its cluster than to any point in another cluster



Types of clustering

Contiguity-based

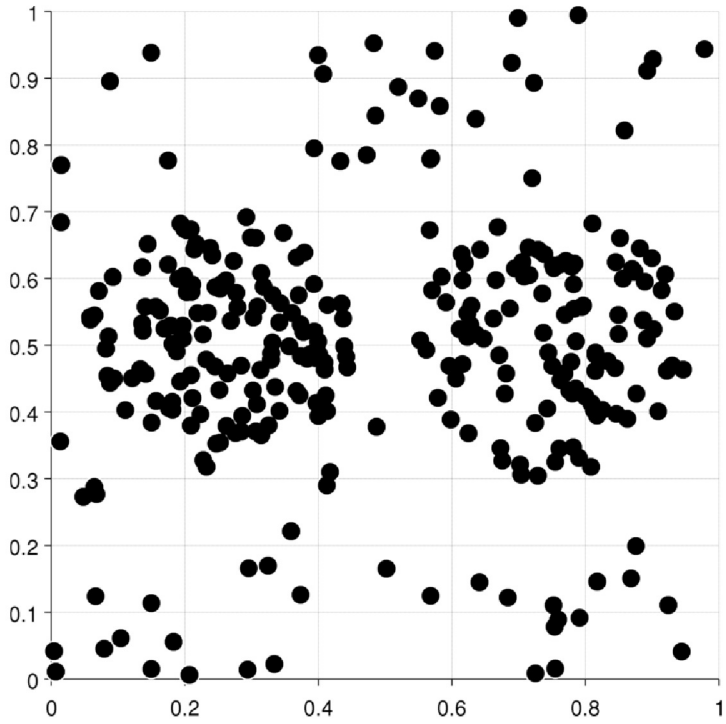
- Each point is closer to at least one point in its cluster than to any point in another cluster



Types of clustering

Density-based

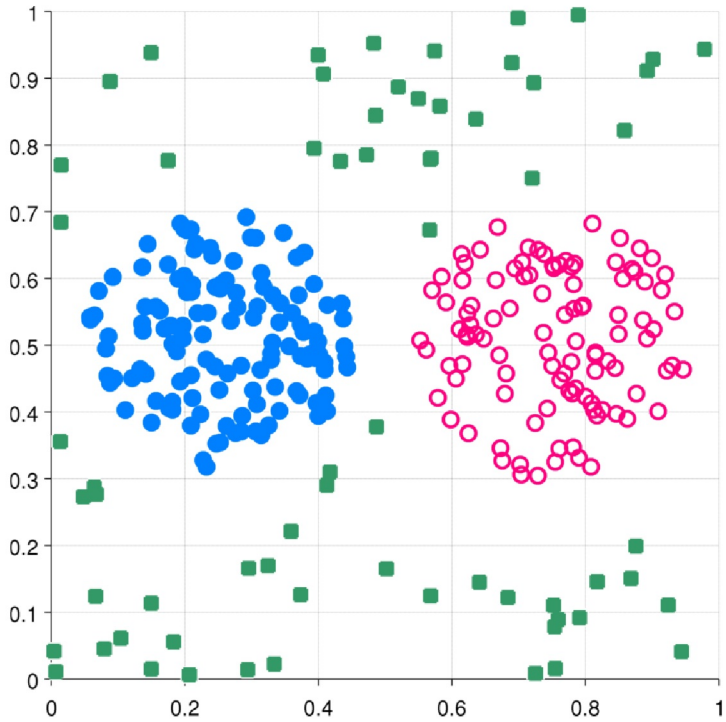
- Clusters are regions of high density separated by regions of low density



Types of clustering

Density-based

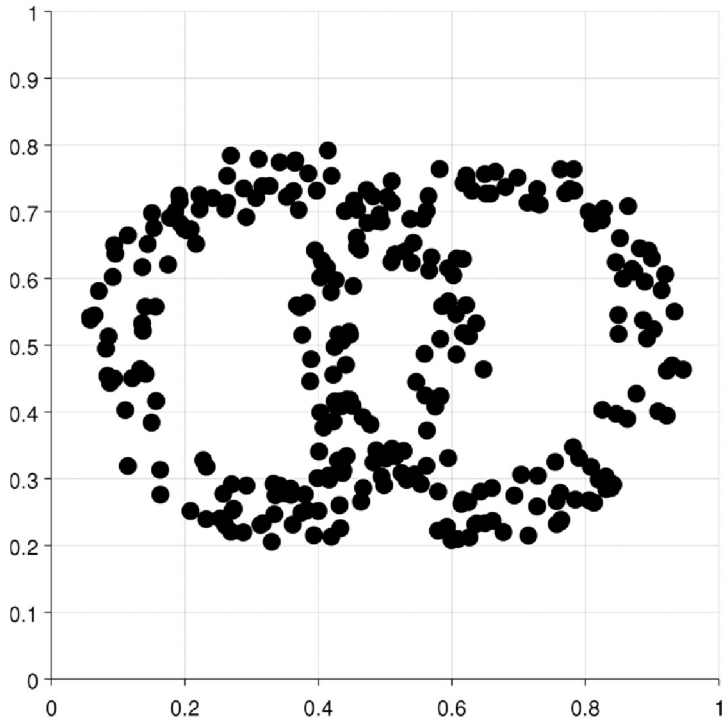
- Clusters are regions of high density separated by regions of low density



Types of clustering

Conceptual clusters

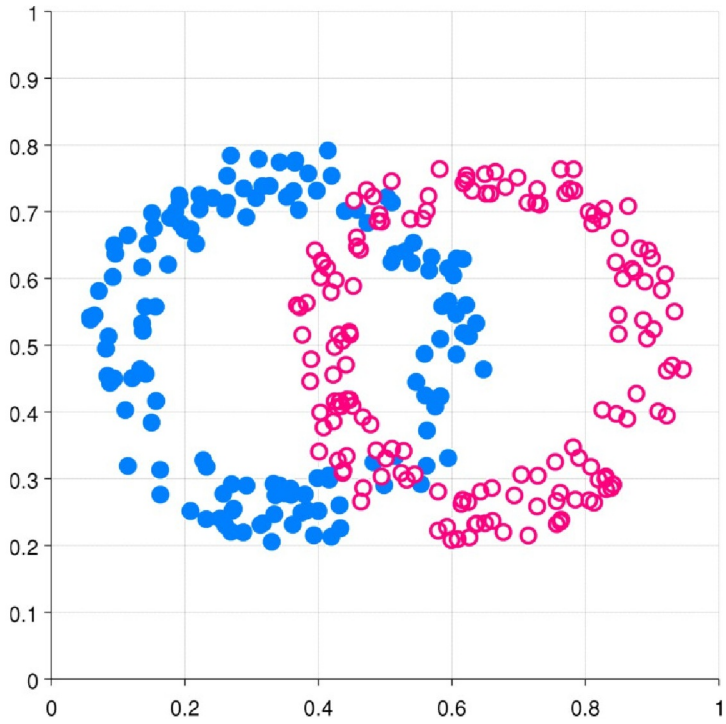
- Points in a cluster share some general property that derives from the entire set of points



Types of clustering

Conceptual clusters

- Points in a cluster share some general property that derives from the entire set of points



Quiz 01 (please answer on Piazza): Clustering types

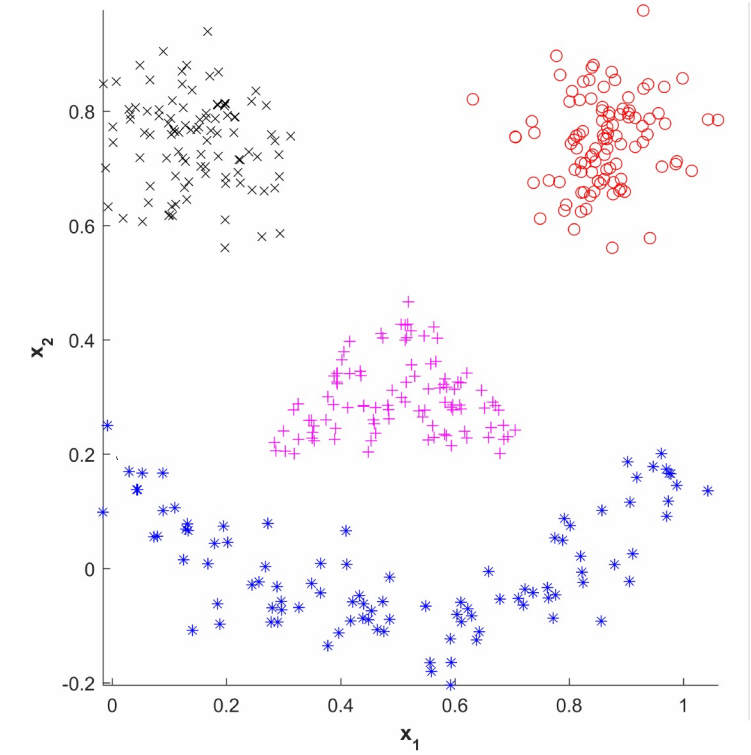


Figure 1: A clustering problem containing four clusters indicated by black crosses, red circles, magenta plusses and blue stars.

Consider the clustering problem given in Figure 1. Which clustering approach is *most* suited for correctly separating the data into the four groups indicated by black crosses, red circles, magenta plusses, and blue asterics?

A. A well-separated clustering approach.

✓ B. A contiguity-based clustering approach.

C. A center-based clustering approach.

D. A conceptual clustering approach.

E. Don't know.

Solution:

As the observation in each cluster is at least closest to one other observation in its cluster than to an observation in another cluster a contiguity based approach is most suited.

K-means clustering

$X = N \times M$

$\mu_1 \dots \mu_k \in \mathbb{R}^M$.

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change

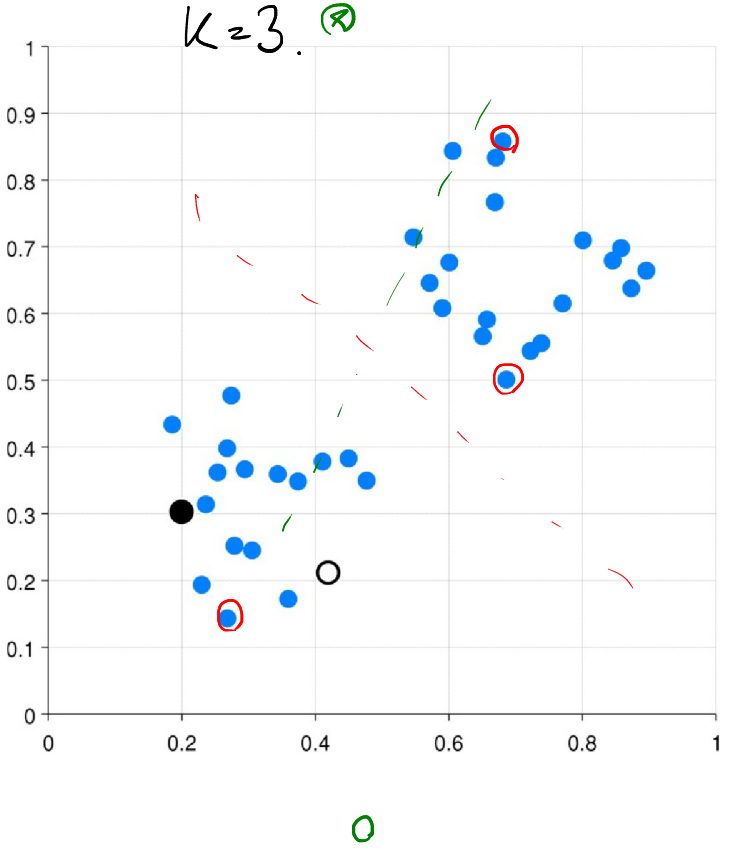
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



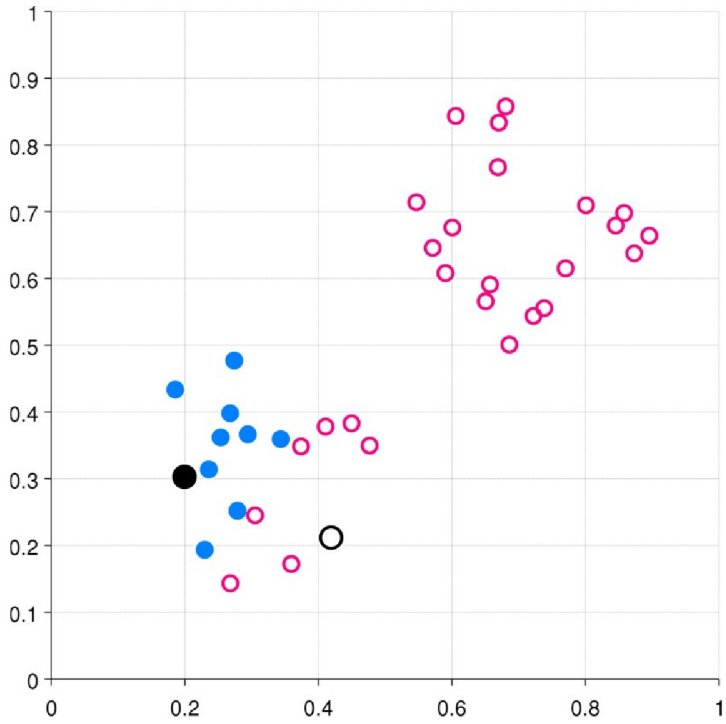
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



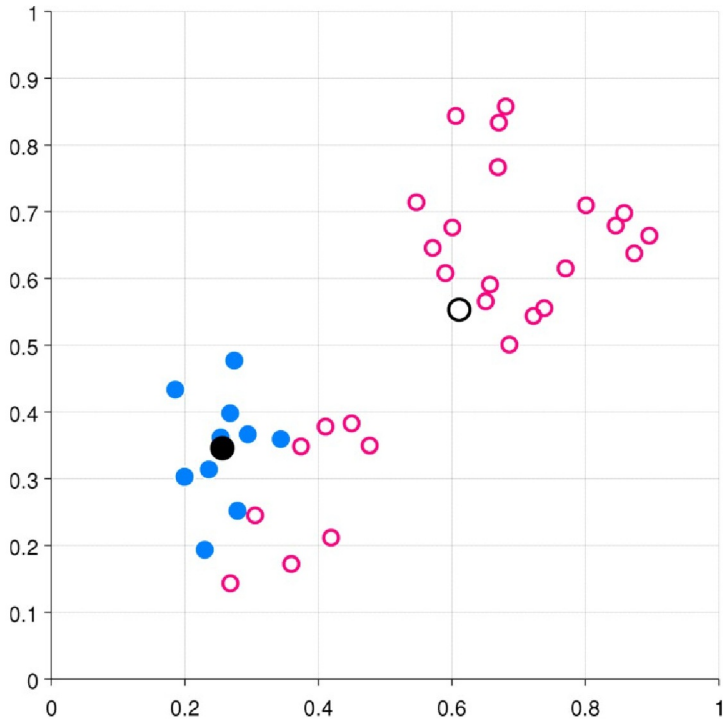
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



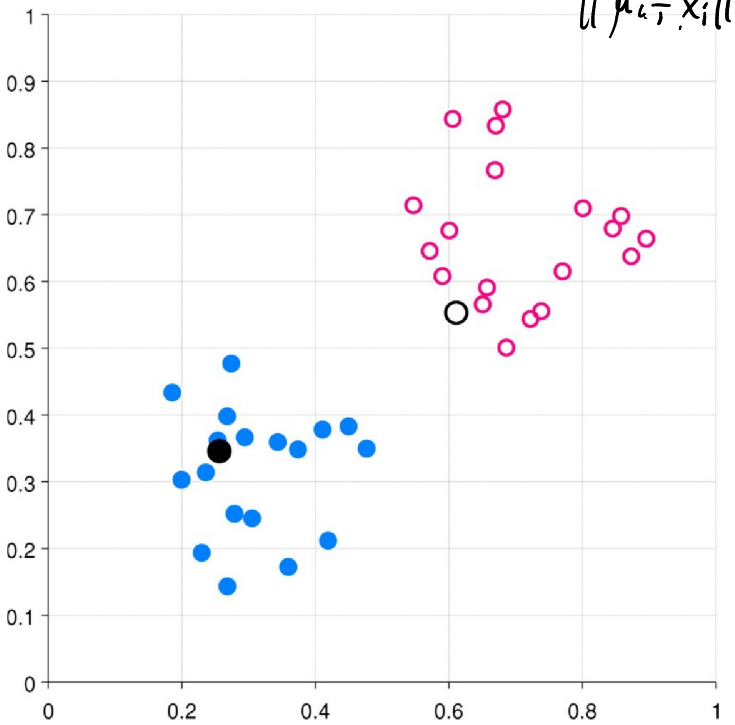
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change



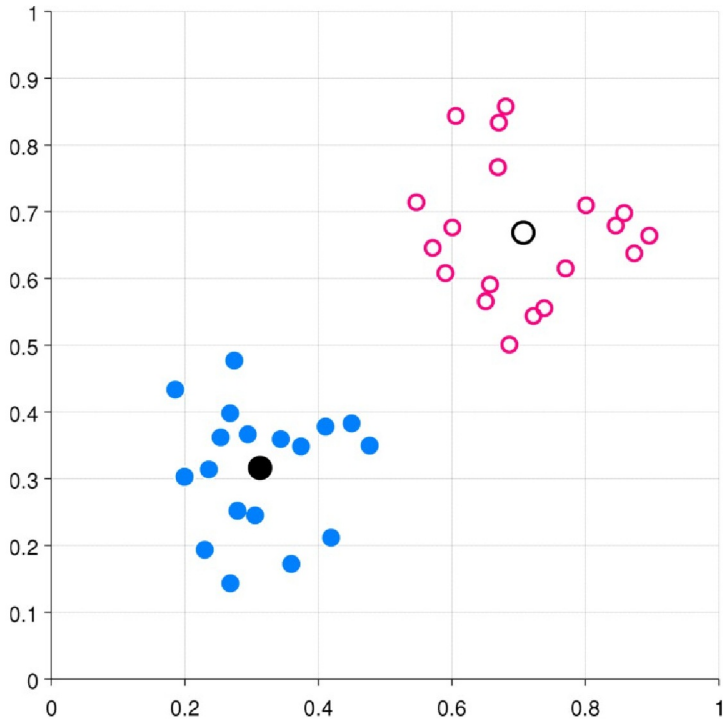
K-means clustering

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change

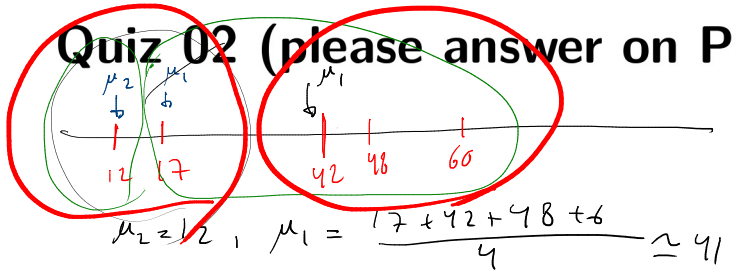


K-means clustering

How do I

- Find the closest centroid? $\|\mu_k - x_i\| \longrightarrow d(\mu_k, x_i)$
  - Use a suitable **dissimilarity/similarity measure**
- Compute the cluster centroids
  - Depends on dissimilarity/similarity measure
  - For example, for Euclidean distance the mean is optimal

Quiz 02 (please answer on Piazza): K-means



Consider the following dataset

$$\boldsymbol{X} = \{42, 60, 17, 48, 12\}$$

$N = 5$, $M = 1$

Select K points as initial centroids

Repeat

- Form K clusters by assigning each point to its closest centroid
- Recompute the centroids of each cluster

Until centroids do not change

We wish to apply the K-means algorithm with $K = 2$ clusters to this dataset and we initialize with cluster centroids at $\mu_1 = 17$ and $\mu_2 = 12$. Carefully, using pen and paper, go through each step of the K-means algorithm until it converge. What is the final clustering?

A. $\{60, 48\}, \{12, 17, 42\}$

B. $\{42, 60, 48, 17\}, \{12\}$

C. $\{60\}, \{12, 17, 42, 48\}$

D. $\{42, 60, 48\}, \{12, 17\}$ ✓

E. Don't know.

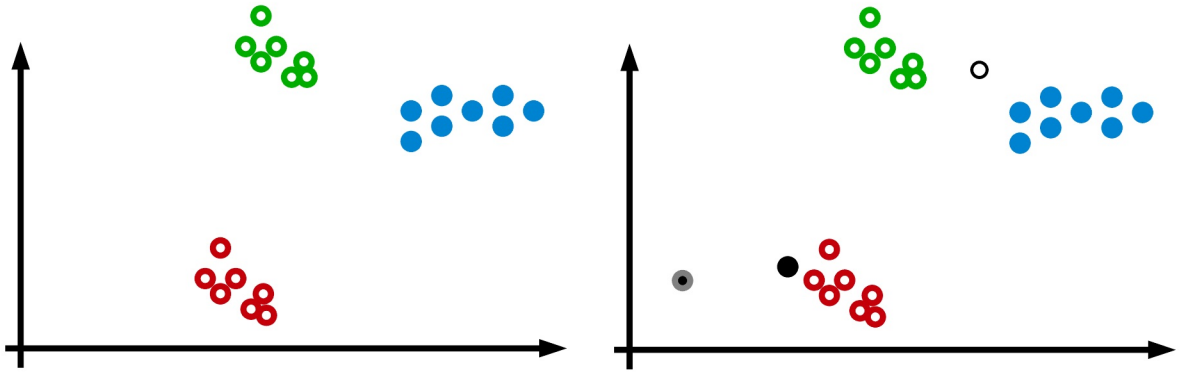
Solution:

The correct answer is D. We will verify this by listing the intermediate steps of the K-means algorithm:

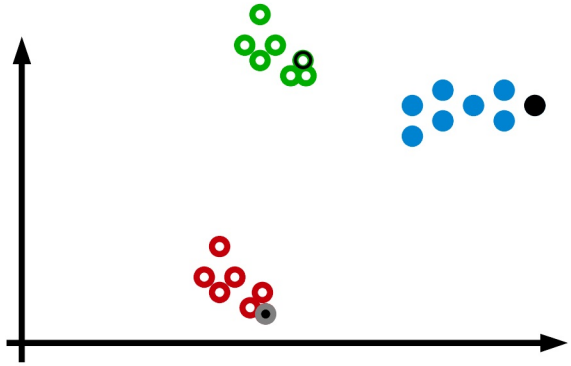
1. The initial clustering will be
$$\{42, 60, 48, 17\}, \quad \{12\}.$$
2. The new centroids will be
$$\mu_1 = \frac{42 + 60 + 48 + 17}{4} = 41.75, \mu_2 = \frac{12}{1} = 12.$$
3. The new clusters will then be
$$\{42, 60, 48\}, \quad \{12, 17\}.$$
4. The centroids then become:
$$\mu_1 = \frac{42 + 60 + 48}{3} = 50, \mu_2 = \frac{12 + 17}{2} = 14.5.$$

It is easy to verify the cluster assignment/centroids will no longer be updated and the method therefore stops.

How will the data (top-left diagram) be clustered given the initialization of the three centroids shown at the right and at the bottom?



- What could we do if we have an empty cluster?
- What could be a good initialization procedure? (Farthest First)



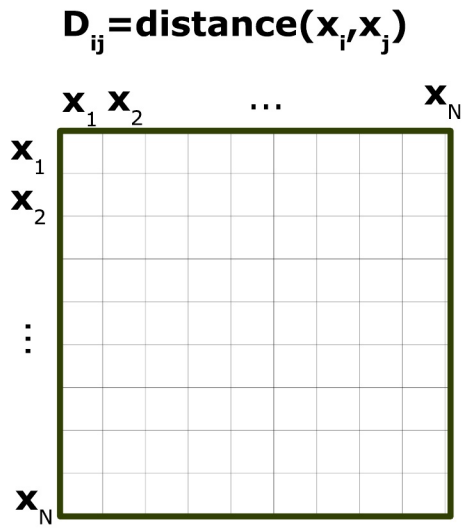
Agglomerative hierarchical clustering

Initialize the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



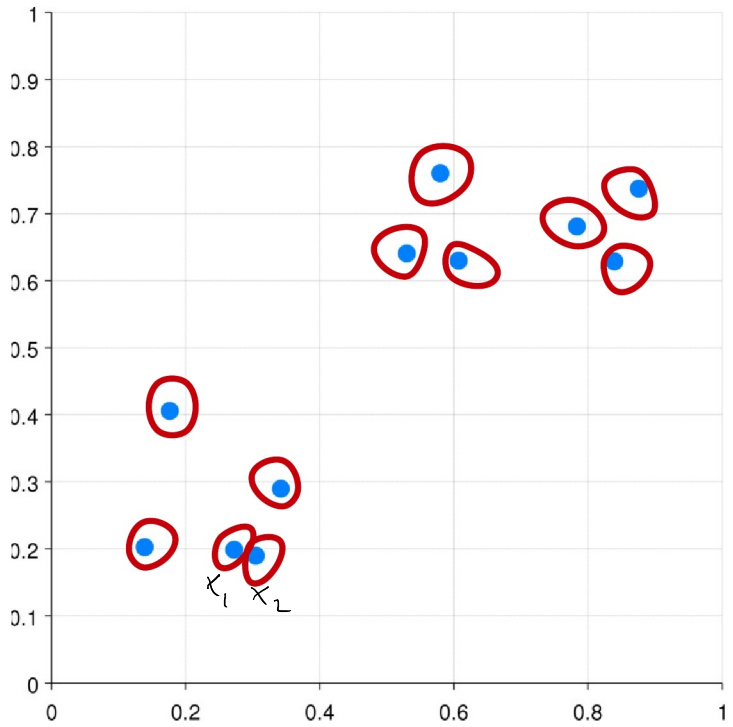
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



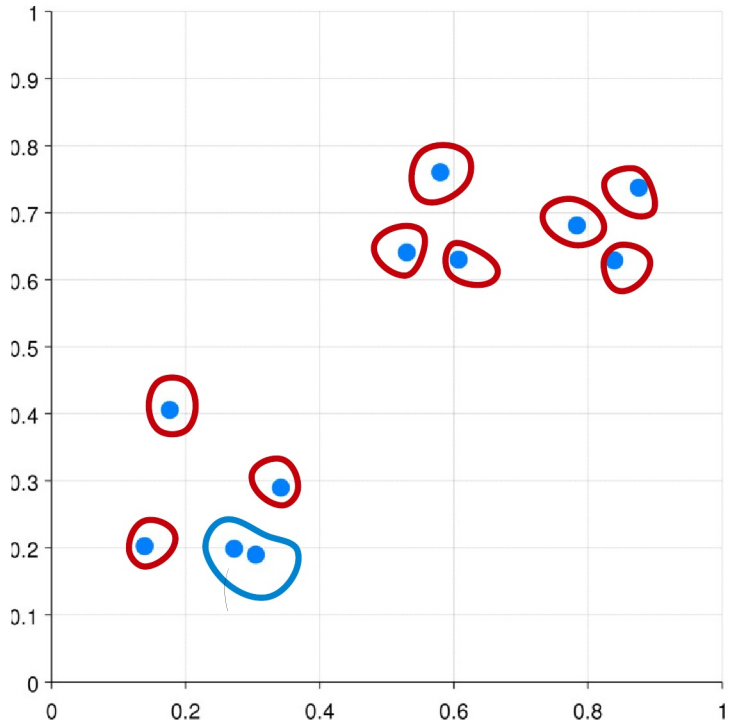
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



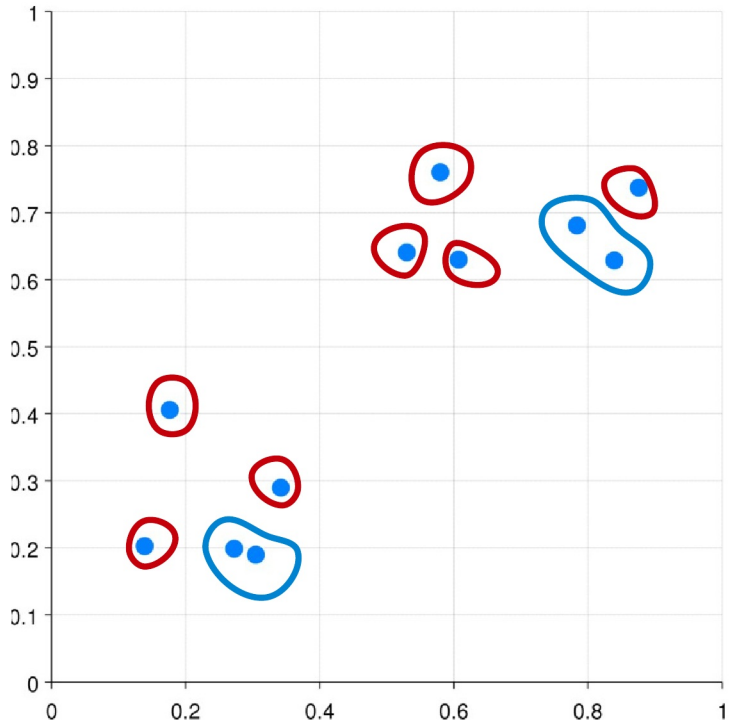
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



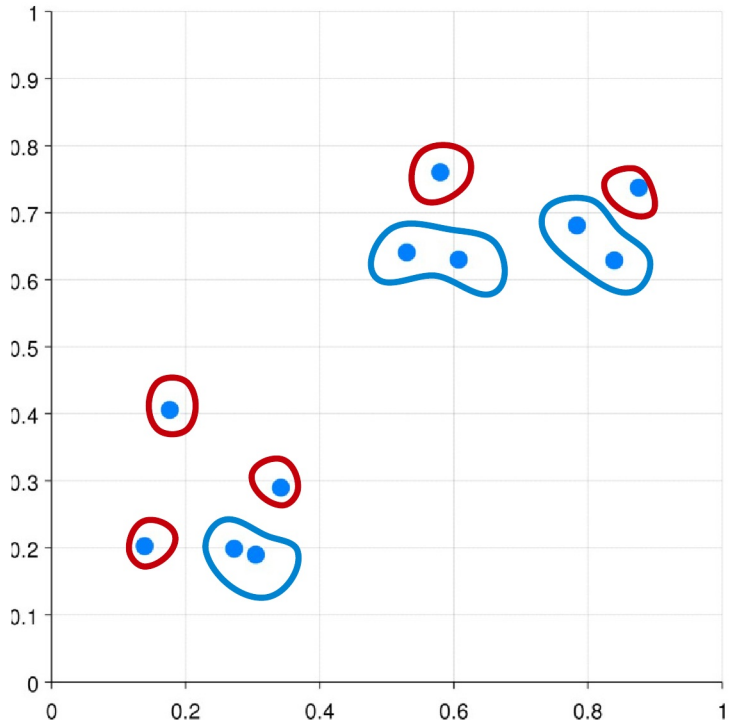
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



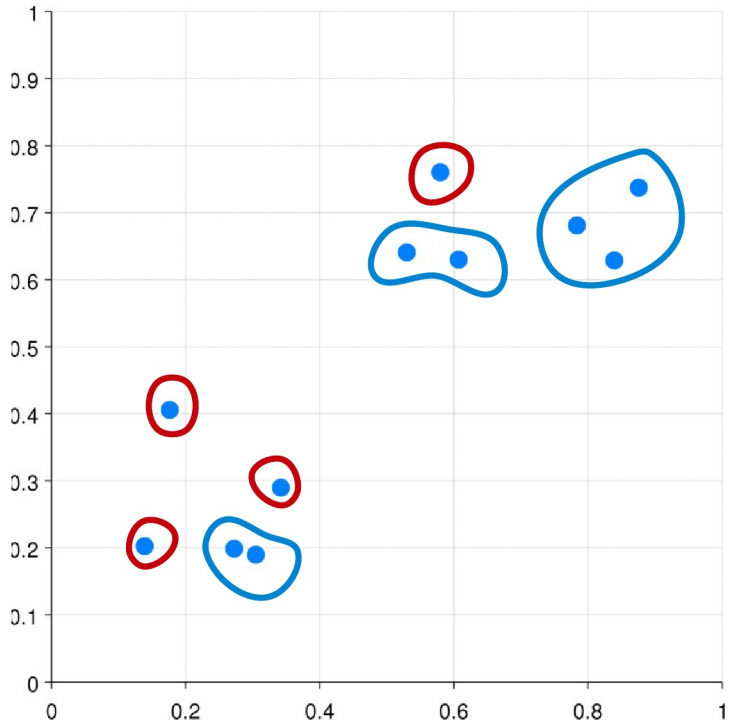
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



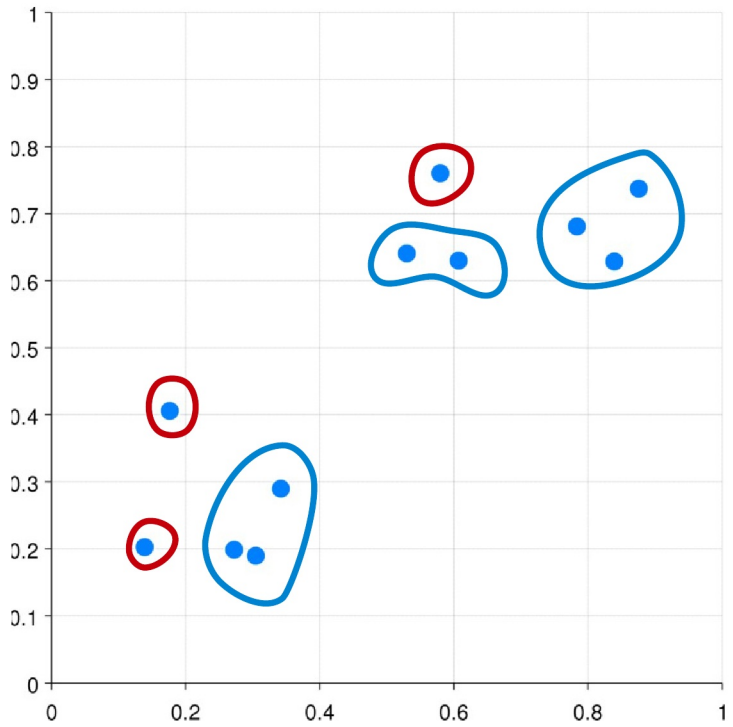
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



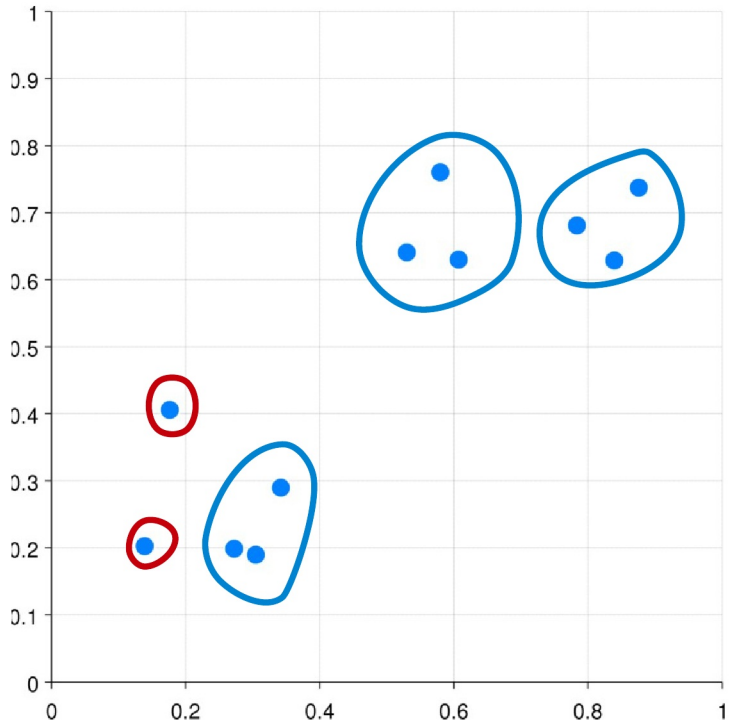
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



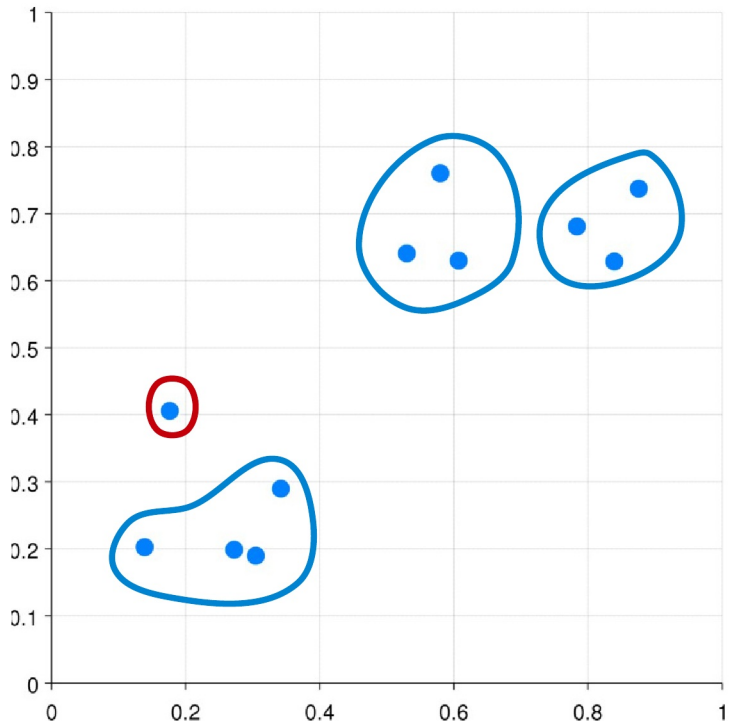
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



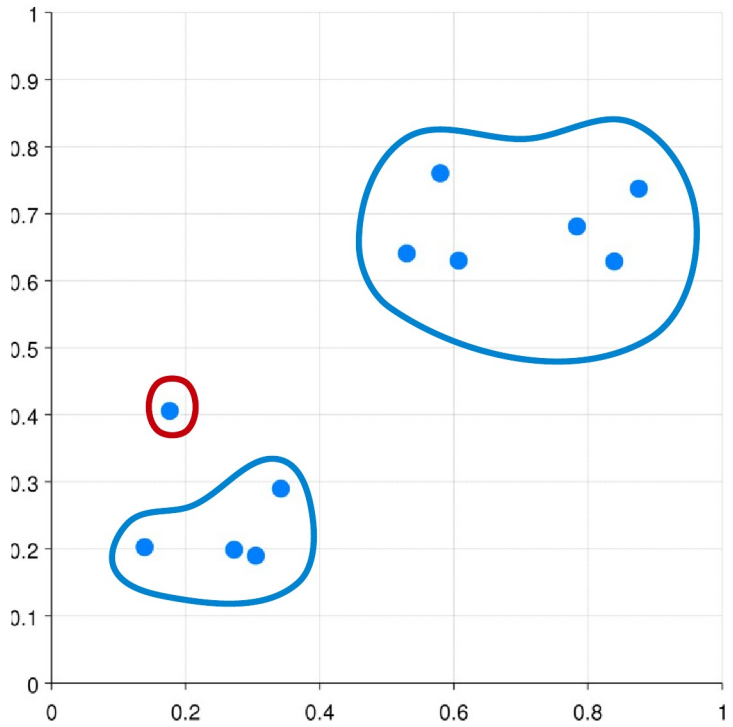
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



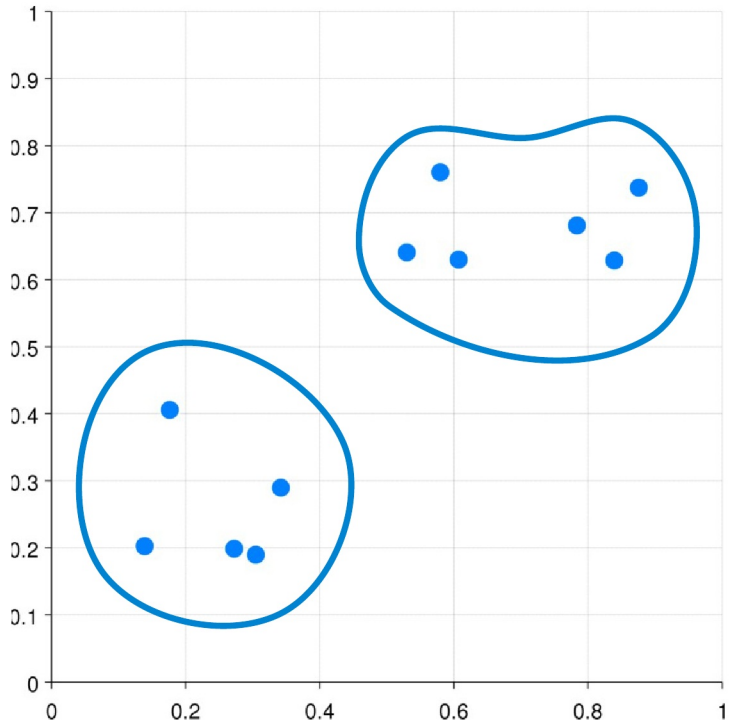
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



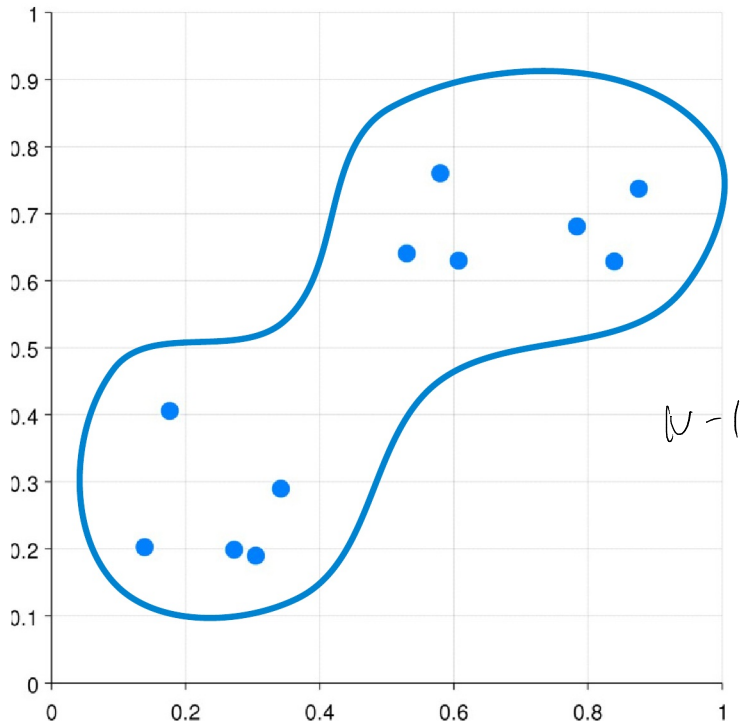
Agglomerative hierarchical clustering

Compute the proximity matrix

Repeat

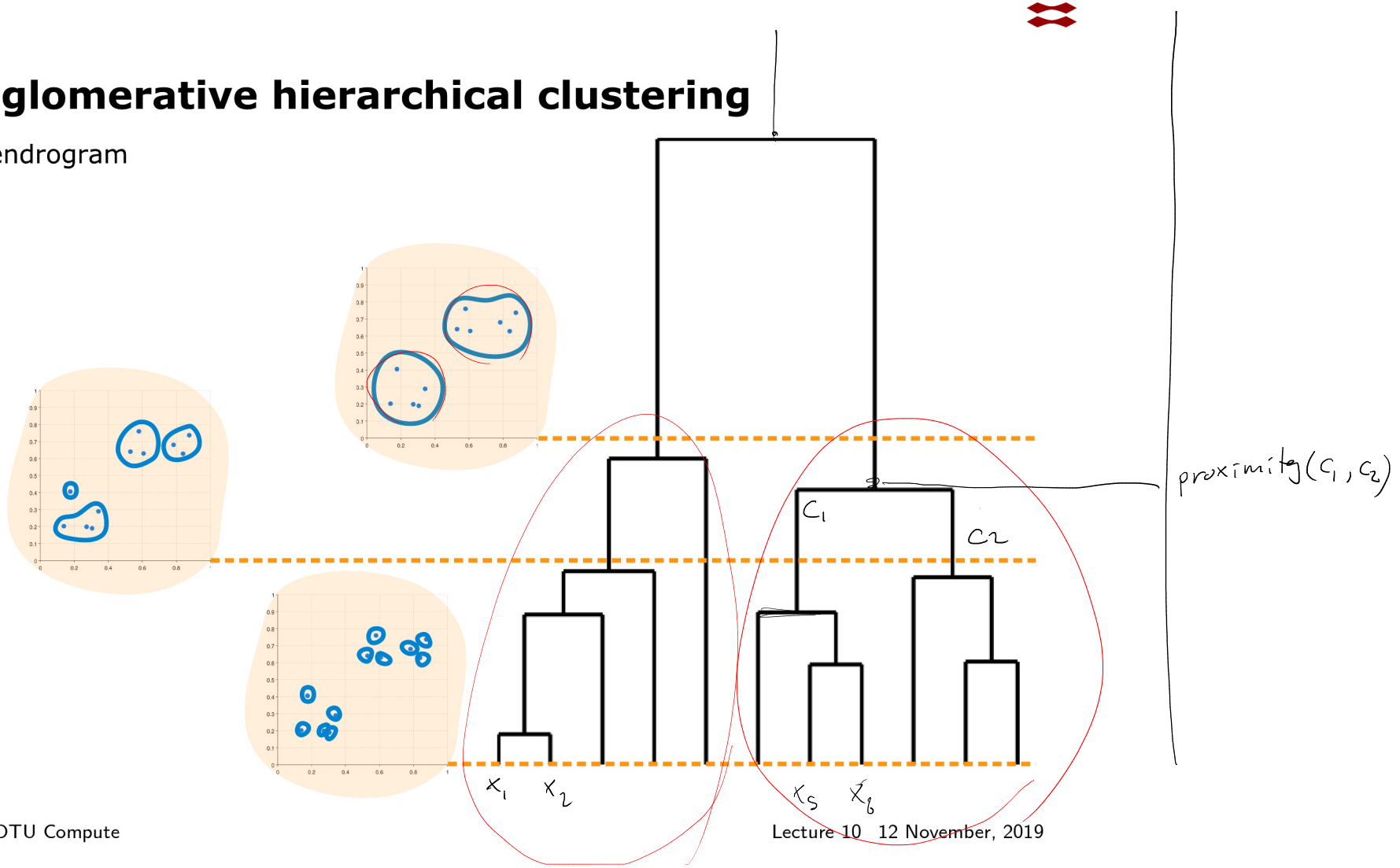
- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

Until only one cluster remains



Agglomerative hierarchical clustering

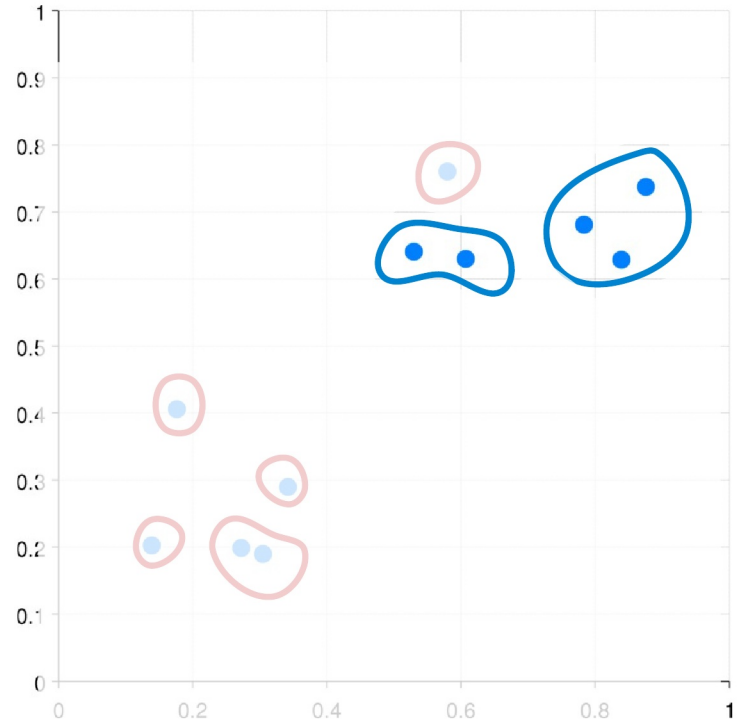
- Dendrogram



Similarity between clusters

- The **key operation** in agglomerative hierarchical clustering is measuring **distance (dissimilarity) between clusters**

$d(x_i, x_j)$

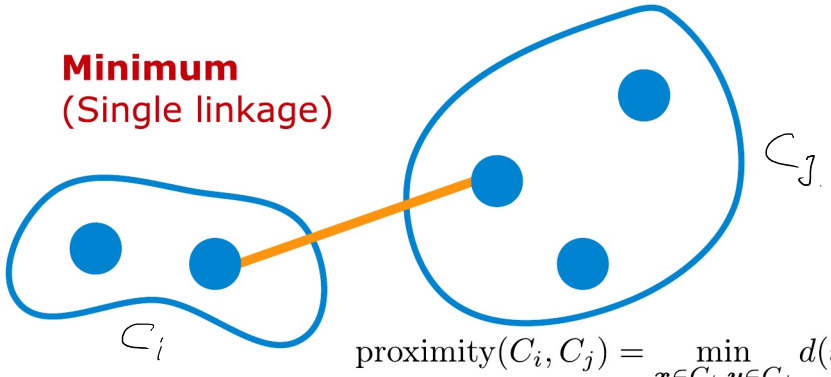


Proximity between clusters

- Can be computed using **proximity between objects**

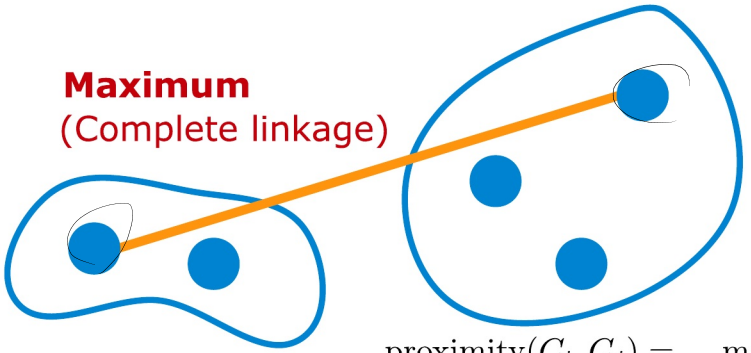
- In our example before we used Euclidian distance as proximity measure

Minimum
(Single linkage)



$$\text{proximity}(C_i, C_j) = \min_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})$$

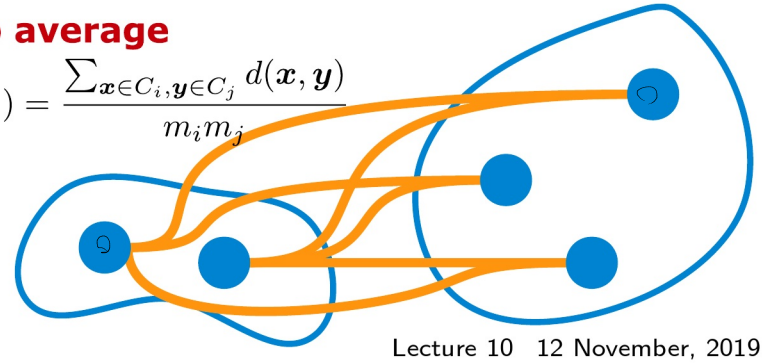
Maximum
(Complete linkage)



$$\text{proximity}(C_i, C_j) = \max_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})$$

Group average

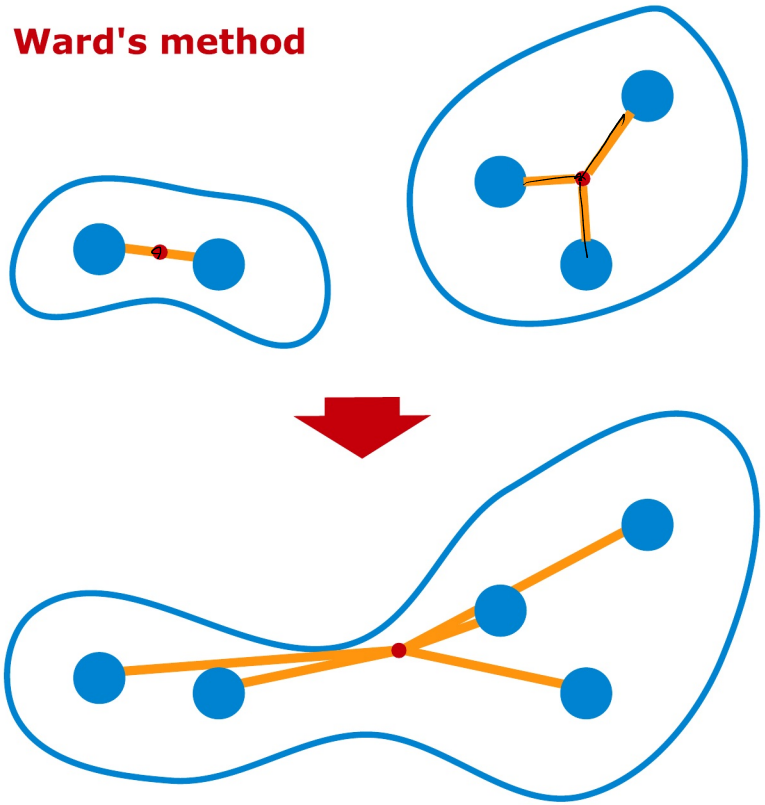
$$\text{proximity}(C_i, C_j) = \frac{\sum_{\boldsymbol{x} \in C_i, \boldsymbol{y} \in C_j} d(\boldsymbol{x}, \boldsymbol{y})}{m_i m_j}$$



C_i: Observations in cluster i
C_j: Observations in cluster j
m_i: Number of observations in cluster i
m_j: Number of observations in cluster j

Similarity between clusters

- Increase in sum of squared error after merging the two clusters should be as small as possible



Quiz 03 (please answer on Piazza): Dendrograms

14:00

Consider once more the dataset:

$$\boldsymbol{X} = \{42, 60, 17, 48, 12\}$$

Using pen-and-paper, carefully build a dendrogram from X one step at a time using Euclidean distance and *minimum* (single) linkage. What will the dendrogram look like?

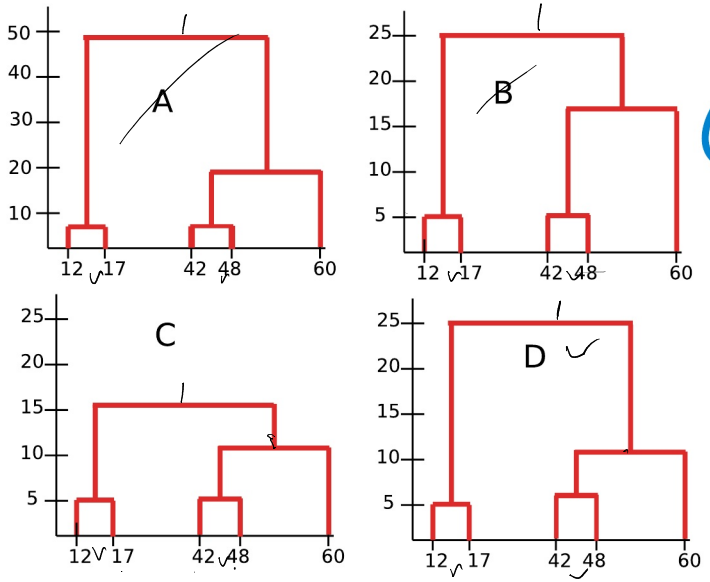
A. Dendrogram A

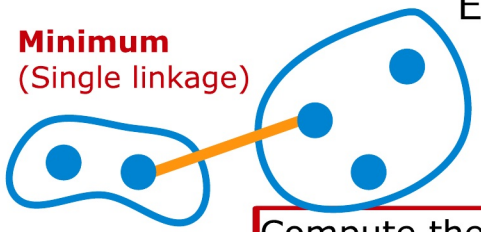
B. Dendrogram B

C. Dendrogram C

D. Dendrogram D

E. Don't know.





Compute the proximity matrix

Repeat

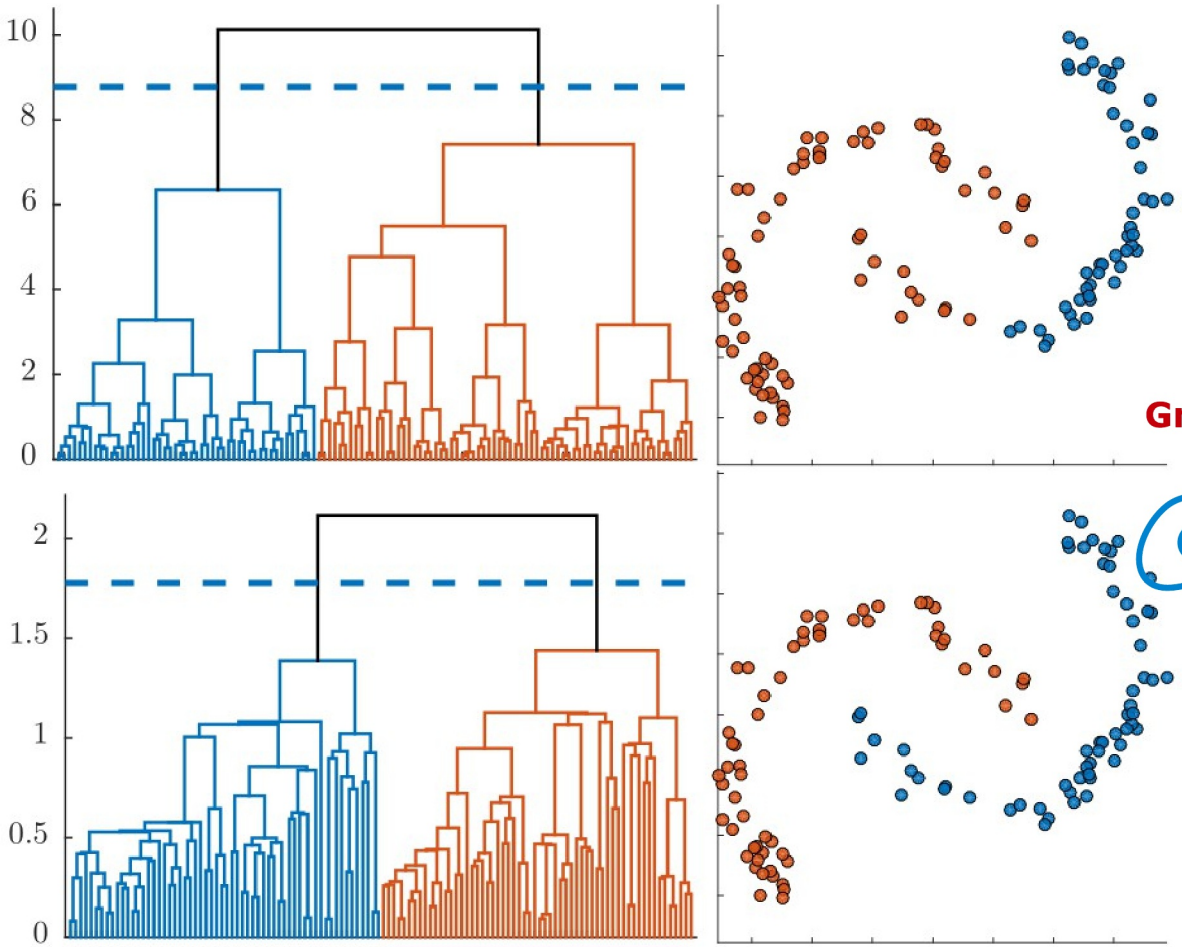
- Merge the two closest clusters
- Update the proximity matrix to reflect the proximity between the new cluster and the original clusters

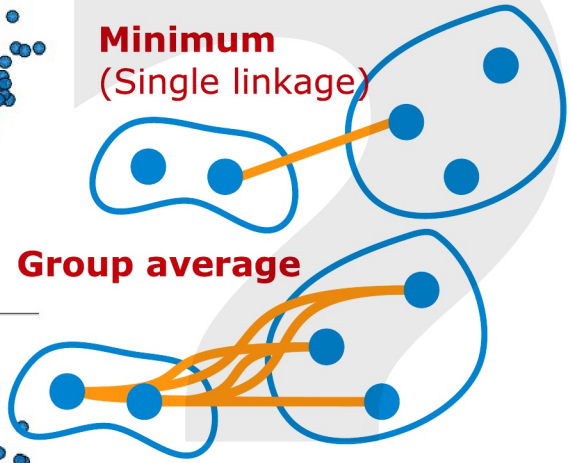
Until only one cluster remains

Solution:

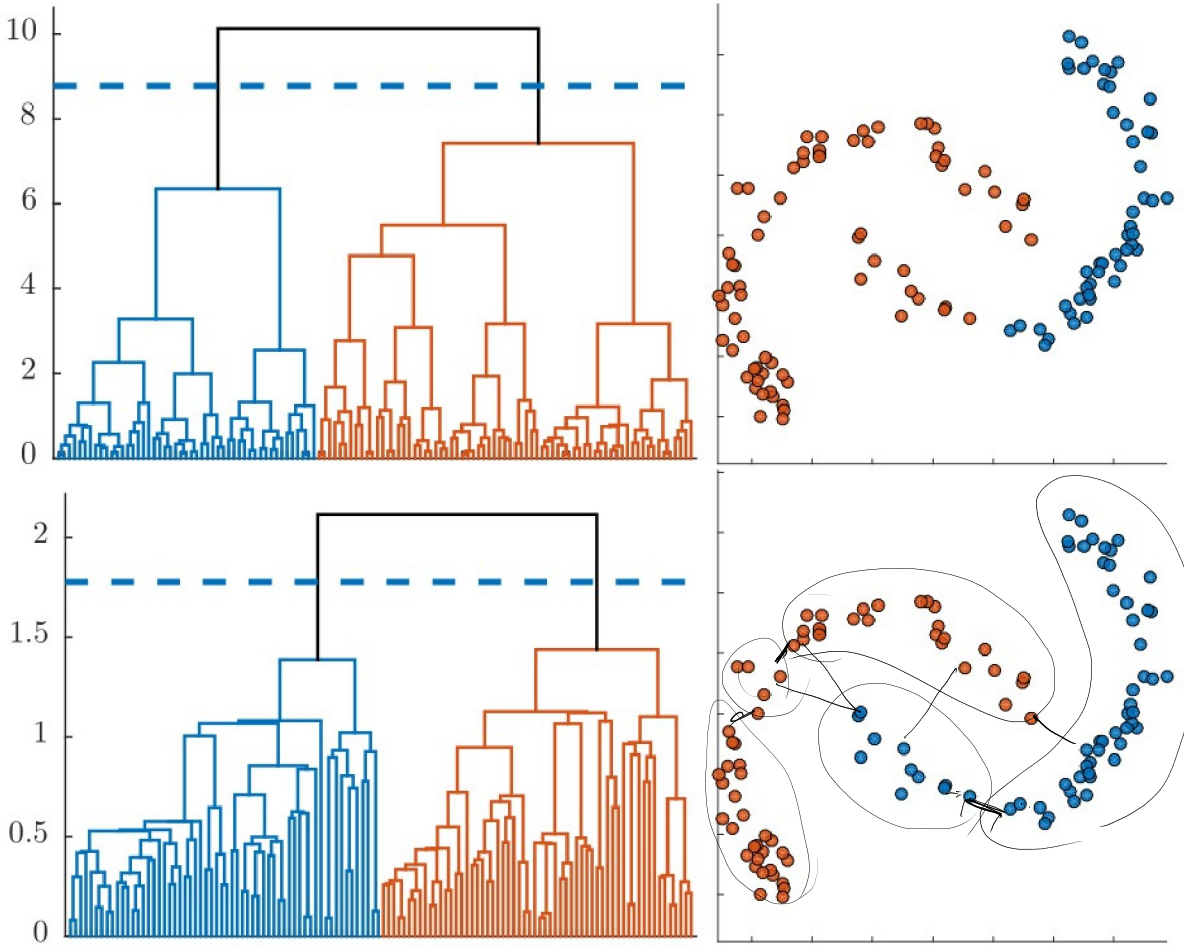
The correct answer is D. The clusters will merge at height 5, 6, 12 and 25.

Clusterings and linkage function

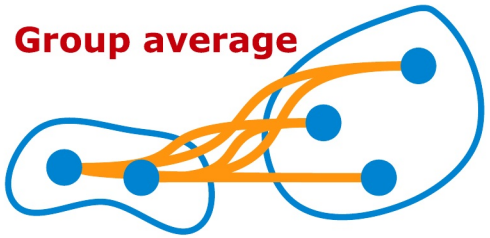




Clusterings and linkage function

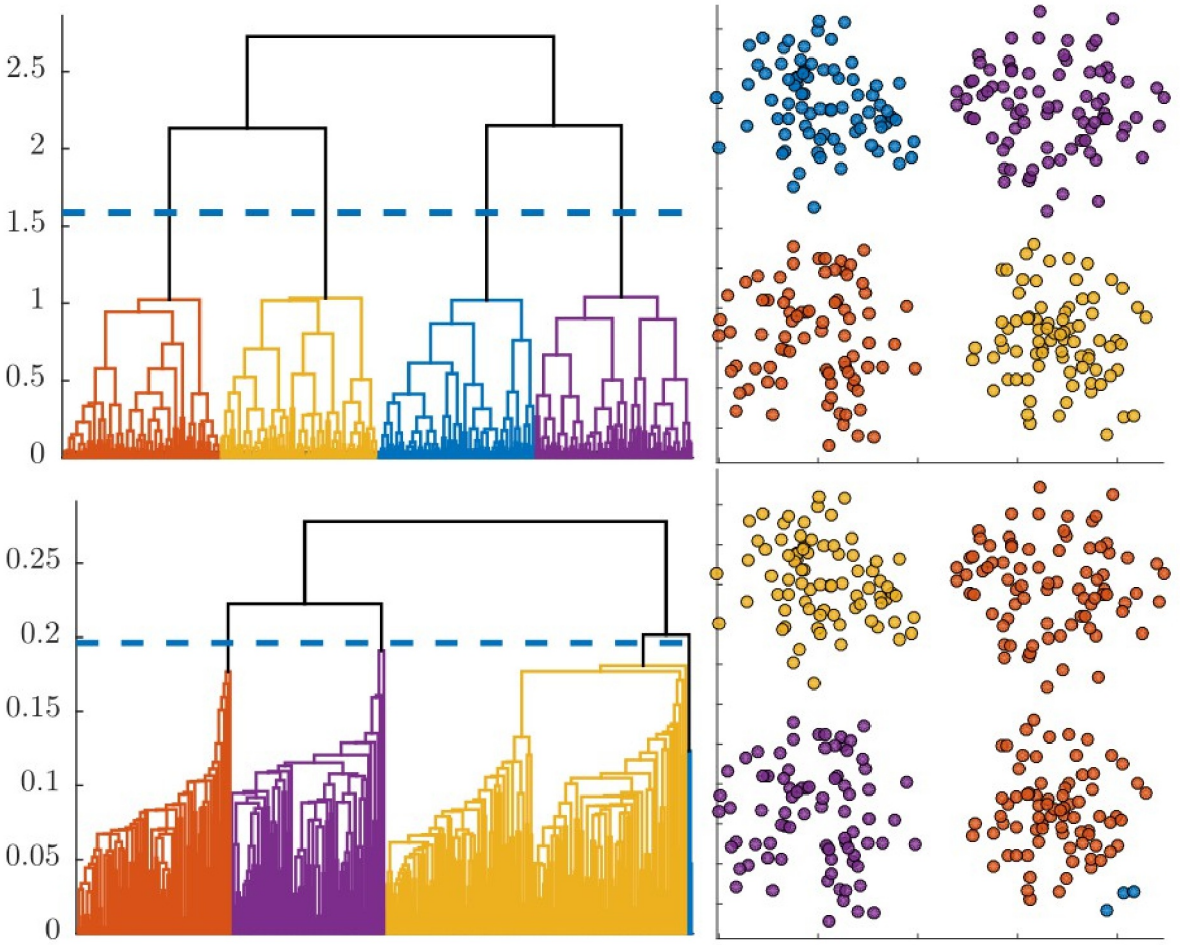


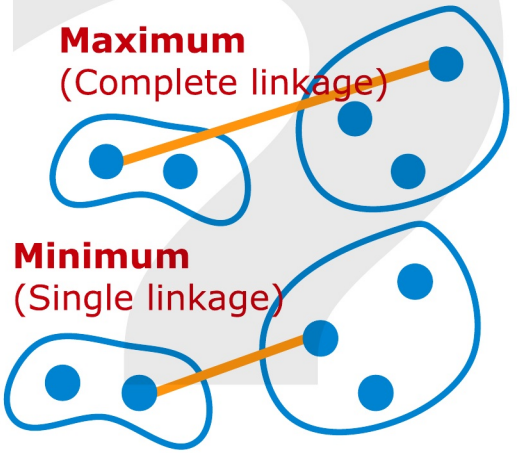
Group average



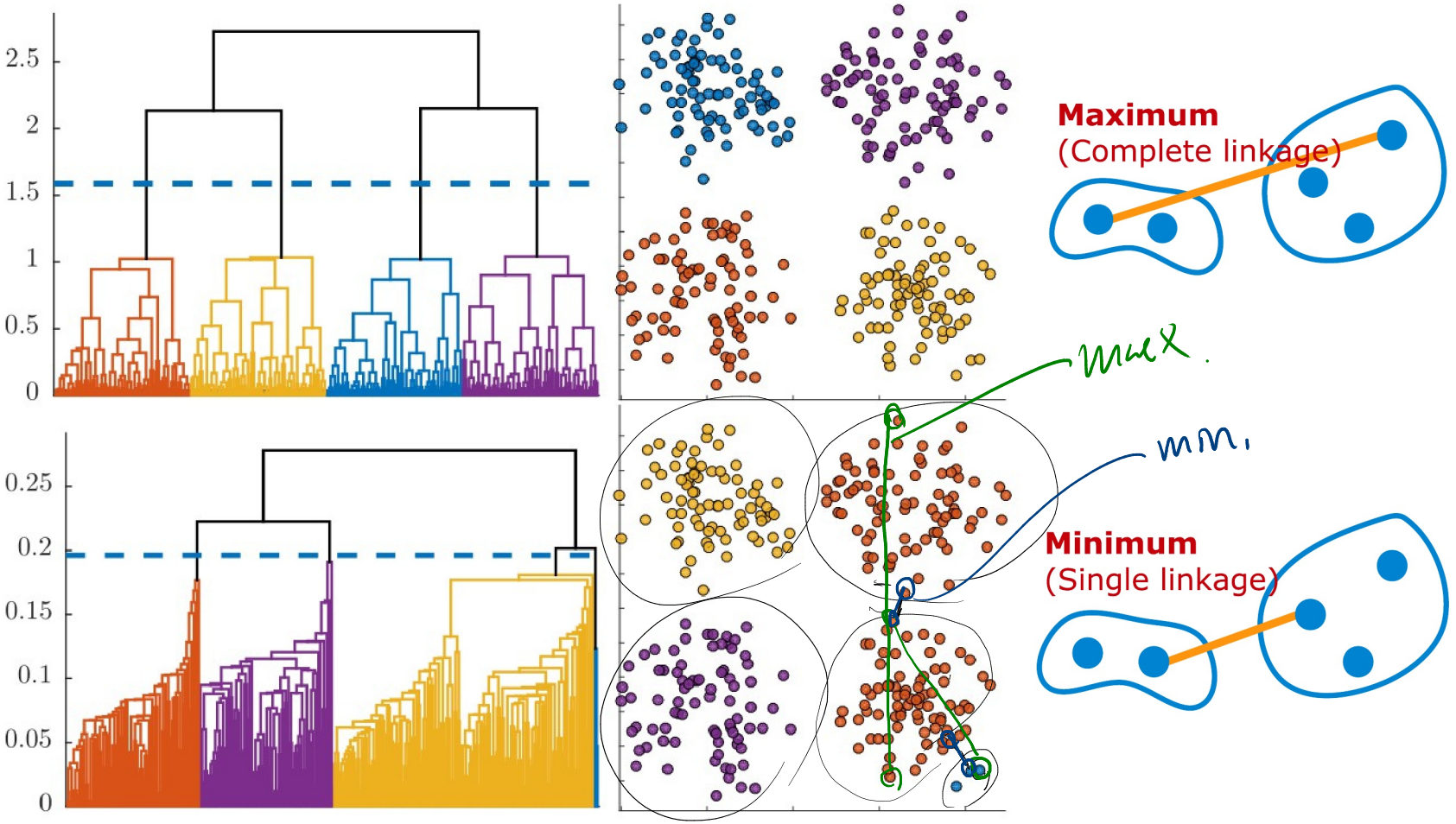
Minimum
(Single linkage)

Clusterings and linkage function

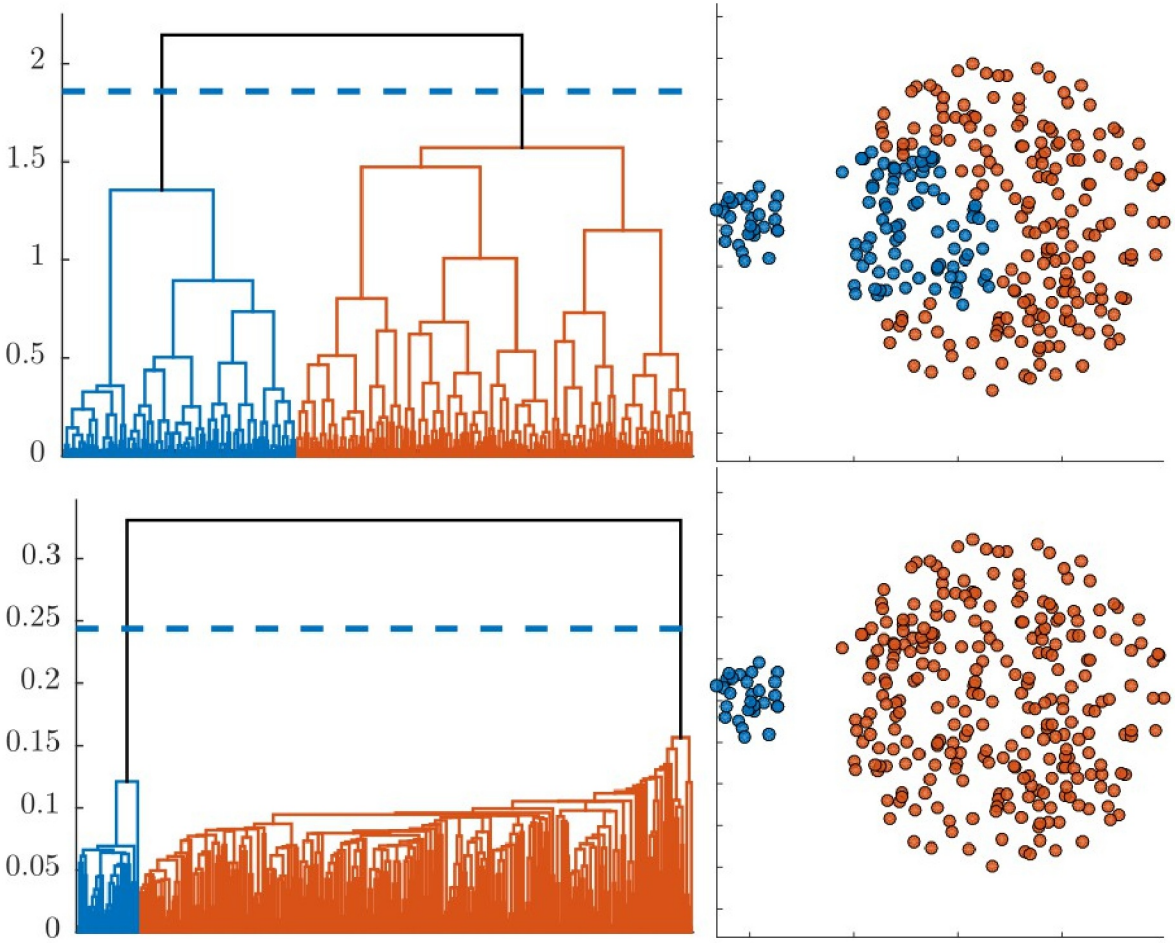


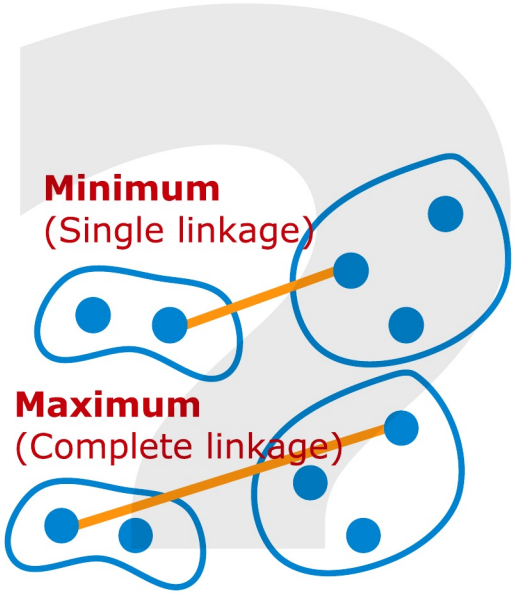


Clusterings and linkage function

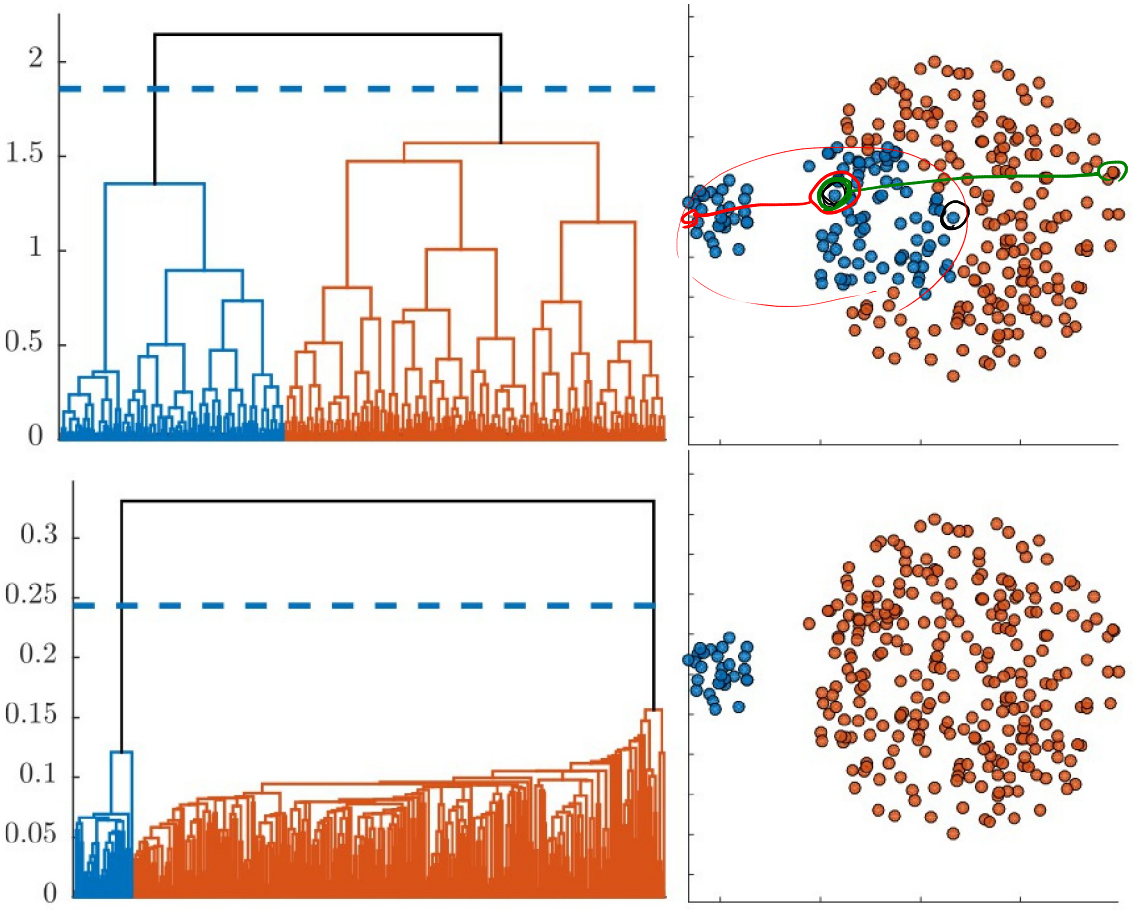


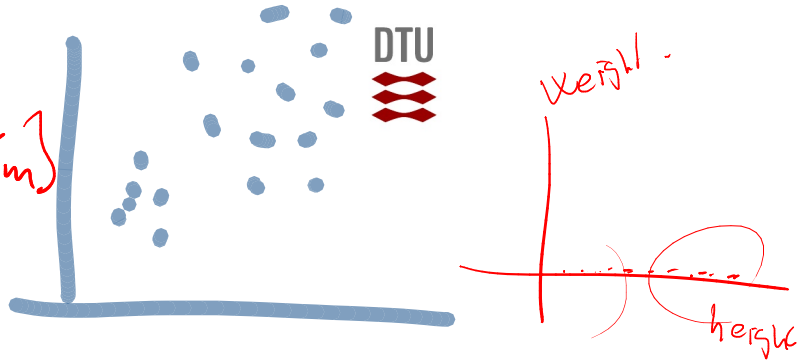
Clusterings and linkage function





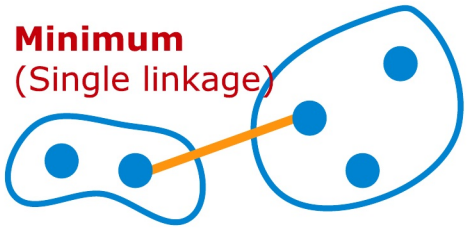
Clusterings and linkage function





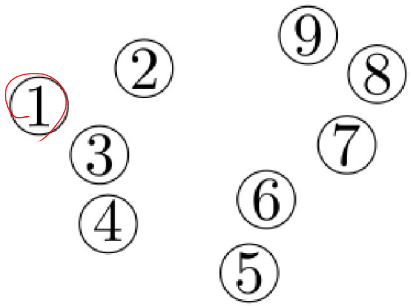
Maximum
(Complete linkage)

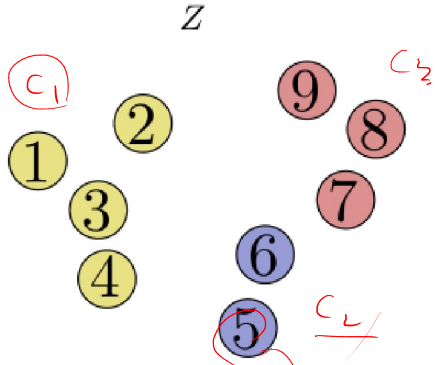
Minimum
(Single linkage)

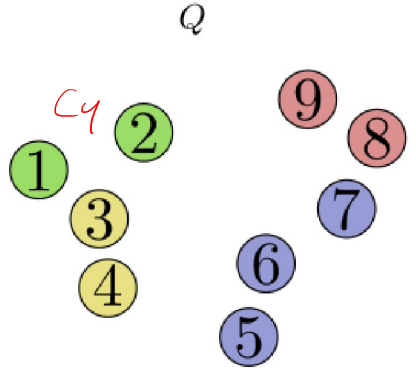


Comparing partitions

- How similar are Q and Z





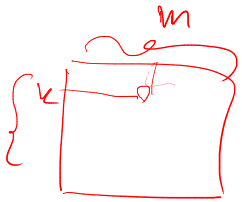


$$Z = \begin{bmatrix} 1 & 1 & 1 & 1 & 2 & 2 & 3 & 3 & 3 \end{bmatrix}$$
$$Q = \begin{bmatrix} 4 & 4 & 1 & 1 & 2 & 2 & 2 & 3 & 3 \end{bmatrix}$$

- Note encoding is (and should be!) arbitrary

$$Q' = \begin{bmatrix} 10 & 10 & 3 & 3 & 8 & 8 & 8 & 1 & 1 \end{bmatrix}$$

Encoding



$$n_{km} = \{\text{Observations assigned to cluster } k \text{ in } Z \text{ and } m \text{ in } Q\} = \sum_{i=1}^{N}\sum_{j=1}^{N} \delta_{z_i,k}\delta_{z_j,m}$$

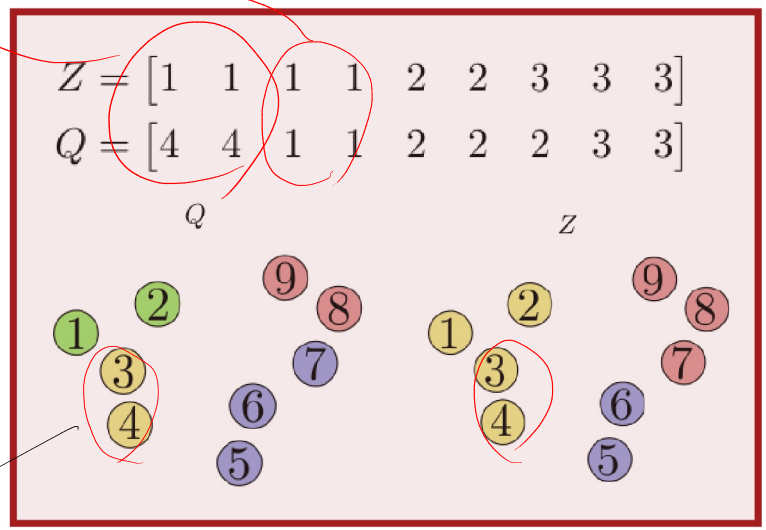
$$\boldsymbol{n}^Z = \{\text{Number of observations assigned to cluster } k \text{ in } Z\} = \sum_{m=1}^{M} n_{km}$$

$$\boldsymbol{n}^Q = \{\text{Number of observations assigned to cluster } m \text{ in } Q\} = \sum_{k=1}^{K} n_{km}$$

$$\boldsymbol{n} = \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}$$

Note the horizontal/vertical sums of $\boldsymbol{n}$:

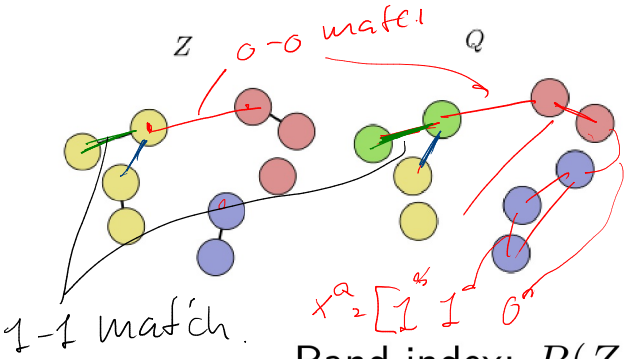
$$\boldsymbol{n}^Z = \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{n}^Q = \begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

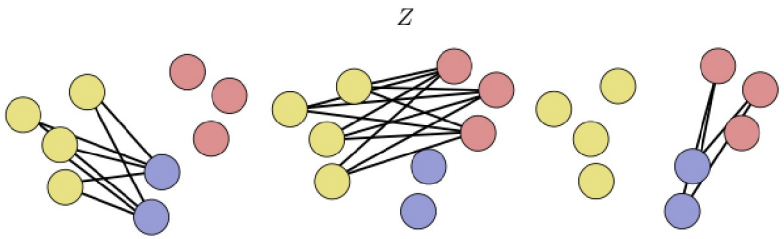


Jaccard and SMC

- Any two observations i, j can either be in the same cluster, or in different clusters
- There are $\frac{1}{2}N(N-1)$ pairs total
- We get two $\frac{1}{2}N(N-1)$-long binary vectors corresponding to each pair i, j

$$S = \{ \text{ Number of pairs } i, j \text{ in the same cluster in } Z, Q \}$$
$$D = \{ \text{ Number of pairs } i, j \text{ in different clusters in } Z, Q \}$$





$$\text{Rand index: } R(Z, Q) = \frac{S + D}{\frac{1}{2}N(N-1)} = \frac{4 + 24}{\frac{1}{2}9 \cdot 8}$$

$$\text{Jaccard similarity: } J(Z, Q) = \frac{S}{\frac{1}{2}N(N-1) - D} = \frac{4}{\frac{1}{2}9 \cdot 8 - 24} = \frac{1}{3}$$

Jaccard and rand index in general

Recall

$$\boldsymbol{n} = \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad \boldsymbol{n}^Z = \begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{n}^Q = \begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

$S = \{$ Number of pairs i, j in the same cluster in $Z, Q\}$

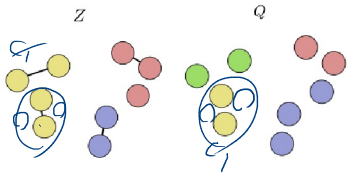
$$= \sum_{k=1}^{K} \sum_{m=1}^{M} \frac{n_{km}(n_{km}-1)}{2}$$

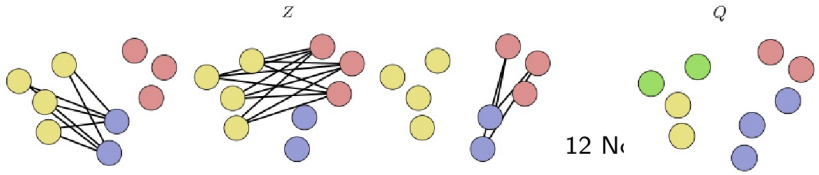
$$= \frac{2(2-1)}{2} + \frac{2(2-1)}{2} + \frac{2(2-1)}{2} + \frac{1(1-1)}{2} + \frac{2(2-1)}{2} = 4$$

$D = \{$ Number of pairs i, j in different clusters in $Z, Q\}$

$$= \frac{N(N-1)}{2} - \sum_{k=1}^{K} \frac{n_k^Z(n_k^Z-1)}{2} - \sum_{m=1}^{M} \frac{n_m^Q(n_m^Q-1)}{2} + S$$

$$= 36 - 10 - 6 + 4 = 24$$





Quiz 04: Cluster overlap

	o_1	o_2	o_3	o_4	o_5	o_6	o_7	o_8	o_9	o_{10}
o_1	0.0	2.0	5.7	0.9	2.9	1.8	2.7	3.7	5.3	5.1
o_2	2.0	0.0	5.6	2.4	2.5	3.0	3.5	4.3	6.0	6.2
o_3	5.7	5.6	0.0	5.0	5.1	4.0	3.3	5.4	1.2	1.8
o_4	0.9	2.4	5.0	0.0	2.7	2.1	2.2	3.5	4.6	4.4
o_5	2.9	2.5	5.1	2.7	0.0	3.5	3.7	4.0	5.8	5.7
o_6	1.8	3.0	4.0	2.1	3.5	0.0	1.7	5.3	3.8	3.7
o_7	2.7	3.5	3.3	2.2	3.7	1.7	0.0	4.2	3.1	3.2
o_8	3.7	4.3	5.4	3.5	4.0	5.3	4.2	0.0	5.5	6.0
o_9	5.3	6.0	1.2	4.6	5.8	3.8	3.1	5.5	0.0	2.1
o_{10}	5.1	6.2	1.8	4.4	5.7	3.7	3.2	6.0	2.1	0.0

Table 1: The pairwise distances between $N = 10$ observations from the travel review dataset. the colors indicate classes

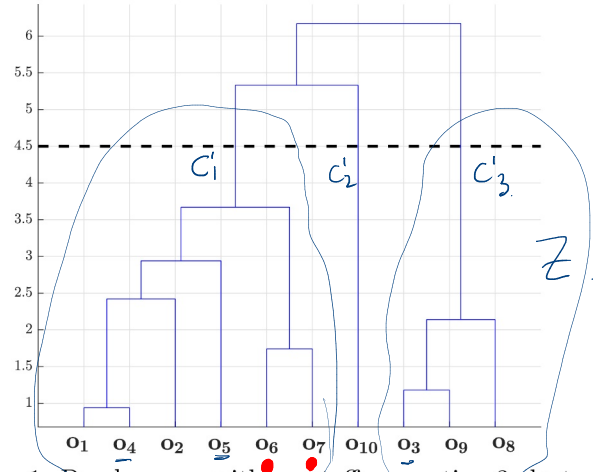


Figure 1: Dendrogram with a cutoff generating 3 clusters.

Consider the dendrogram in Figure 1. Suppose we apply a cutoff (indicated by the black line) thereby generating three clusters. We wish to compare the quality of this clustering, Q, to the ground-truth clustering, Z, indicated by the colors in Table 1. Recall the *Jaccard similarity* of the two clusters is

$$\mathrm{J}[Z,Q] = \frac{S}{\frac{1}{2}N(N-1) - D}$$

in the notation of the lecture notes. What is the Jaccard similarity of the two clusterings?

A. $\mathrm{J}[Z,Q] \approx 0.104$

B. $\mathrm{J}[Z,Q] \approx 0.143$

C. $\mathrm{J}[Z,Q] \approx 0.174$

D. $\mathrm{J}[Z,Q] \approx 0.153$

E. Don't know.

Solution:

To compute J$[Z, Q]$, note Z is the clustering corresponding to the colors in **??** and Q the clustering obtained by cutting the dendrogram in **??** given as:

$$\{10\}, \{1, 2, 4, 5, 6, 7\}, \{3, 8, 9\}$$

From this information we can define the counting matrix $\boldsymbol{n}$ as

$$\boldsymbol{n} = \begin{bmatrix} 0 & 2 & 0 \\ 0 & 2 & 1 \\ 1 & 2 & 2 \end{bmatrix}$$

It is then a simple matter of using the definitions in the lecture notes (see chapter 17.4) to compute

$$S = 4, D = 17$$

From this the answer by simply plugging the values into the formula given in the text and answer B is correct.

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i$, $i = 1, \ldots, n$

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i$, $i = 1, \ldots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i$, $i = 1, \ldots, n$
- **Information** obtained from observing x_i is

$$I_i = -\log p_i$$

- Average information obtained is called the **entropy**

$$H[p_X] = \mathbb{E}[I] = -\sum_{i=1}^{n} p_i \log p_i$$

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i$, $i = 1, \ldots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

- Average information obtained is called the **entropy**

$$H[p_X] = \mathbb{E}[I] = -\sum_{i=1}^{n} p_i \log p_i$$

- Entropy is defined for general densities $P(X = x_i, Y = y_j) = p_{ij}$

$$H[p_{XY}] = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log p_{ij}$$

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i$, $i = 1, \ldots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

$\sum_j p(x_i, y_j) = p(x_i)$

- Average information obtained is called the **entropy**

$$H[p_X] = \mathbb{E}[I] = -\sum_{i=1}^{n} p_i \log p_i$$

- Entropy is defined for general densities $P(X = x_i, Y = y_j) = p_{ij}$

$$H[p_{XY}] = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log p_{ij}$$

- The **Mutual information** is defined as

$$\mathrm{MI}[X, Y] = H[P_X] + H[P_Y] - H[P_{XY}]$$

Entropy and mutual information recap

- Consider a probability distribution $P(X = x_i) = p_i$, $i = 1, \ldots, n$
- **Information** obtained from observing x_i is

$$I = -\log p_i$$

- Average information obtained is called the **entropy**

$$H[p_X] = \mathbb{E}[I] = -\sum_{i=1}^{n} p_i \log p_i$$

- Entropy is defined for general densities $P(X = x_i, Y = y_j) = p_{ij}$

$$H[p_{XY}] = -\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \log p_{ij}$$

- The **Mutual information** is defined as

$$\mathrm{MI}[X, Y] = H[P_X] + H[P_Y] - H[P_{XY}]$$

- The **Normalized mutual information** is defined as

$$\mathrm{NMI}[X, Y] = \frac{\mathrm{MI}[X, Y]}{\sqrt{H[P_X]}\sqrt{H[P_Y]}}$$

Comparing using mutual information

We simply define $P_{ZQ}(i,j) = \frac{1}{N} n_{ij}$. Example:

$$P_{QZ} = \frac{1}{9}\begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \; \boldsymbol{n}^Z = \frac{1}{9}\begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \; \boldsymbol{n}^Q = \frac{1}{9}\begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

Comparing using mutual information

We simply define $P_{ZQ}(i,j) = \frac{1}{N} n_{ij}$. Example:

$$P_{QZ} = \frac{1}{9}\begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad \boldsymbol{n}^Z = \frac{1}{9}\begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{n}^Q = \frac{1}{9}\begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

- **Entropy** computed as $H[p_X] = -\sum_{i=1}^{n} p_i \log p_i$:

$$\text{Entropy of } Z\text{:} \quad H[Z] = -\frac{4}{9}\log\frac{4}{9} - \frac{1}{3}\log\frac{1}{3} - \frac{2}{9}\log\frac{2}{9} \approx 1.06$$

$$\text{Entropy of } Q\text{:} \quad H[Q] = -\frac{2}{9}\log\frac{2}{9} - \frac{2}{9}\log\frac{2}{9} - \frac{2}{9}\log\frac{2}{9} - \frac{1}{3}\log\frac{1}{3} \approx 1.37$$

$$\text{Entropy of } Z \text{ and } Q\text{:} \quad H[ZQ] = -4 \times \frac{2}{9}\log\frac{2}{9} - \frac{1}{9}\log\frac{1}{9} = 1.58.$$

Comparing using mutual information

We simply define $P_{ZQ}(i,j) = \frac{1}{N} n_{ij}$. Example:

$$P_{QZ} = \frac{1}{9}\begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad \boldsymbol{n}^Z = \frac{1}{9}\begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{n}^Q = \frac{1}{9}\begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

- **Entropy** computed as $H[p_X] = -\sum_{i=1}^{n} p_i \log p_i$:

$$\text{Entropy of } Z: \quad H[Z] = -\frac{4}{9}\log\frac{4}{9} - \frac{1}{3}\log\frac{1}{3} - \frac{2}{9}\log\frac{2}{9} \approx 1.06$$

$$\text{Entropy of } Q: \quad H[Q] = -\frac{2}{9}\log\frac{2}{9} - \frac{2}{9}\log\frac{2}{9} - \frac{2}{9}\log\frac{2}{9} - \frac{1}{3}\log\frac{1}{3} \approx 1.37$$

$$\text{Entropy of } Z \text{ and } Q: \quad H[ZQ] = -4 \times \frac{2}{9}\log\frac{2}{9} - \frac{1}{9}\log\frac{1}{9} = 1.58.$$

- **Mutual information**:

$$\mathrm{MI}[Z,Q] = H[Z] + H[Q] - H[Z,Q] \approx 1.06 + 1.37 - 1.58 \approx 0.85.$$

Comparing using mutual information

We simply define $P_{ZQ}(i,j) = \frac{1}{N} n_{ij}$. Example:

$$P_{QZ} = \frac{1}{9}\begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 0 & 0 \\ 0 & 1 & 2 & 0 \end{bmatrix}, \quad \boldsymbol{n}^Z = \frac{1}{9}\begin{bmatrix} 4 \\ 2 \\ 3 \end{bmatrix}, \quad \boldsymbol{n}^Q = \frac{1}{9}\begin{bmatrix} 2 & 3 & 2 & 2 \end{bmatrix}$$

- **Entropy** computed as $H[p_X] = -\sum_{i=1}^{n} p_i \log p_i$:

$$\text{Entropy of } Z: \quad H[Z] = -\frac{4}{9}\log\frac{4}{9} - \frac{1}{3}\log\frac{1}{3} - \frac{2}{9}\log\frac{2}{9} \approx 1.06$$

$$\text{Entropy of } Q: \quad H[Q] = -\frac{2}{9}\log\frac{2}{9} - \frac{2}{9}\log\frac{2}{9} - \frac{2}{9}\log\frac{2}{9} - \frac{1}{3}\log\frac{1}{3} \approx 1.37$$

$$\text{Entropy of } Z \text{ and } Q: \quad H[ZQ] = -4 \times \frac{2}{9}\log\frac{2}{9} - \frac{1}{9}\log\frac{1}{9} = 1.58.$$

- **Mutual information**:

$$\mathrm{MI}[Z,Q] = H[Z] + H[Q] - H[Z,Q] \approx 1.06 + 1.37 - 1.58 \approx 0.85.$$

- **Normalized mutual information**:

$$\mathrm{NMI}[Z,Q] = \frac{\mathrm{MI}[Z,Q]}{\sqrt{H[Z]}\sqrt{H[Q]}} \approx \frac{0.85}{\sqrt{1.06}\sqrt{1.37}} \approx 0.70.$$

feedback on project 2 in 2 weeks from today.