Clustering

Applications of Machine Learning in Education by Chun-Shu Wei, 108-1

Institute of Education, National Chiao Tung University

Outline

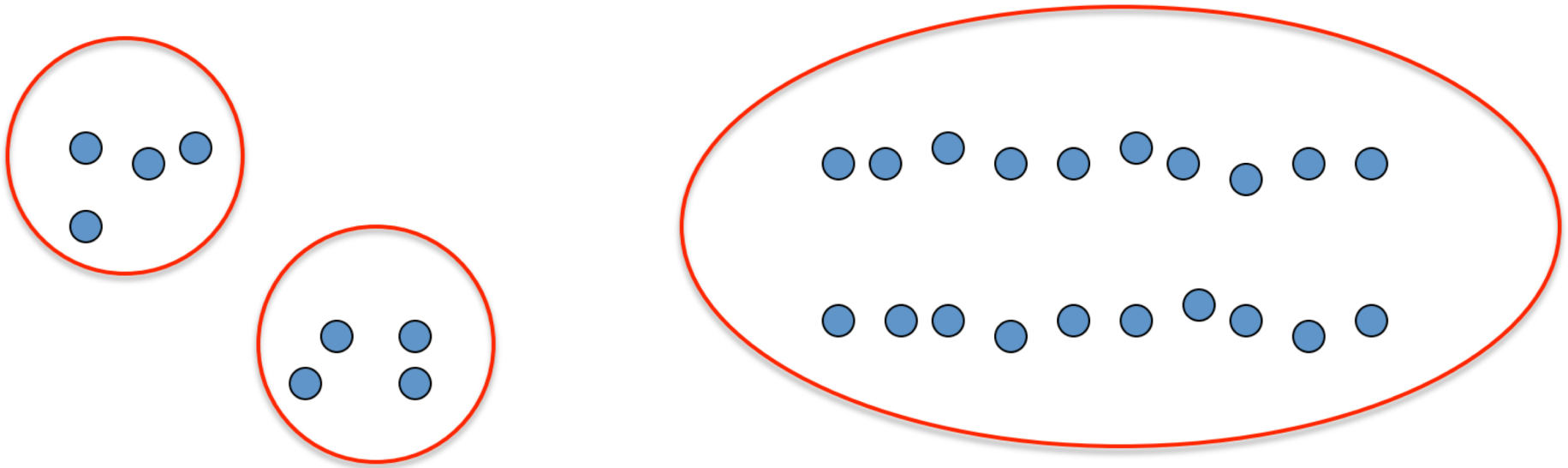
- *k*-means algorithm (*k*-平均演算法)
- Hierarchical clustering (階層式分群)

References

- Wikipedia
- http://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf
- https://medium.com/@chih.sheng.huang821/機器學習-集群分析-k-means-clustering-e608a7fe1b43

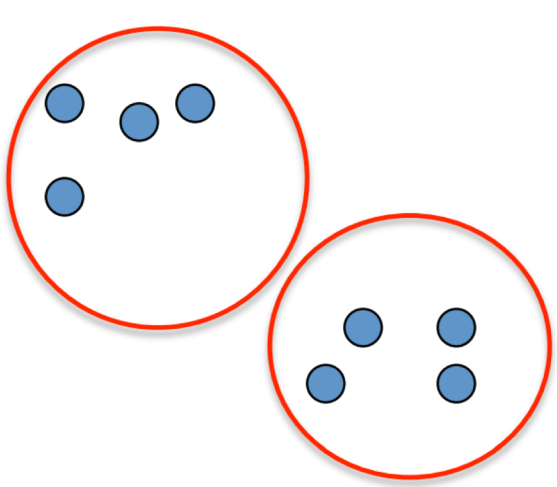
What is Clustering (分群)

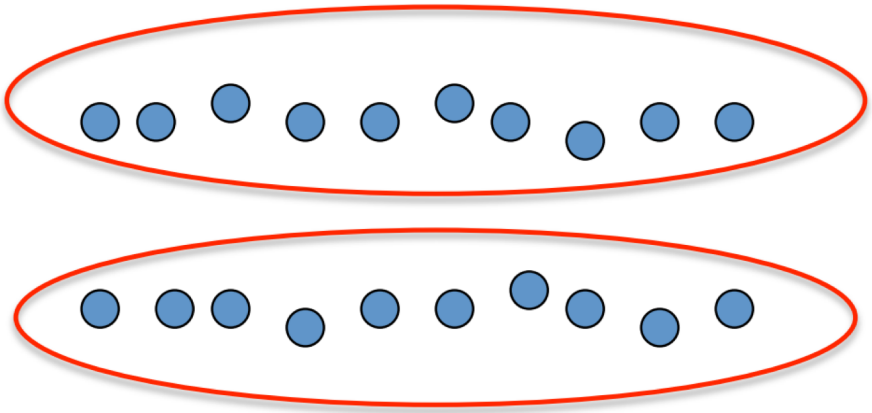
- Basic idea: group together **similar** instances
- Clustering=分群=聚類=群聚=集群=叢集…



What is Clustering (分群)

- Basic idea: group together **similar** instances



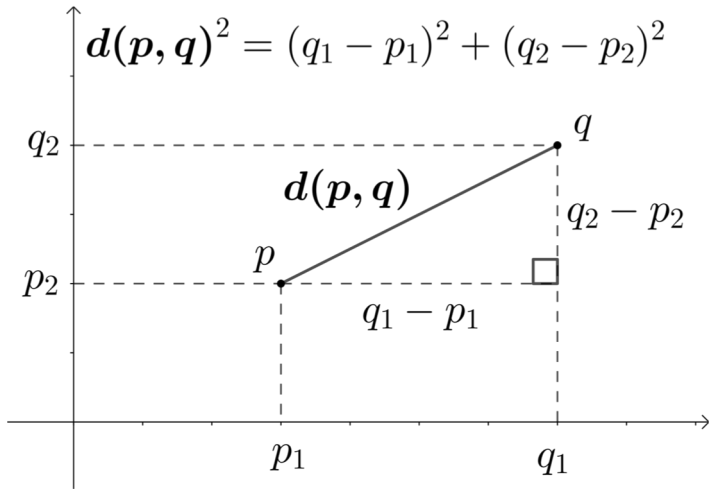


What is “similar”

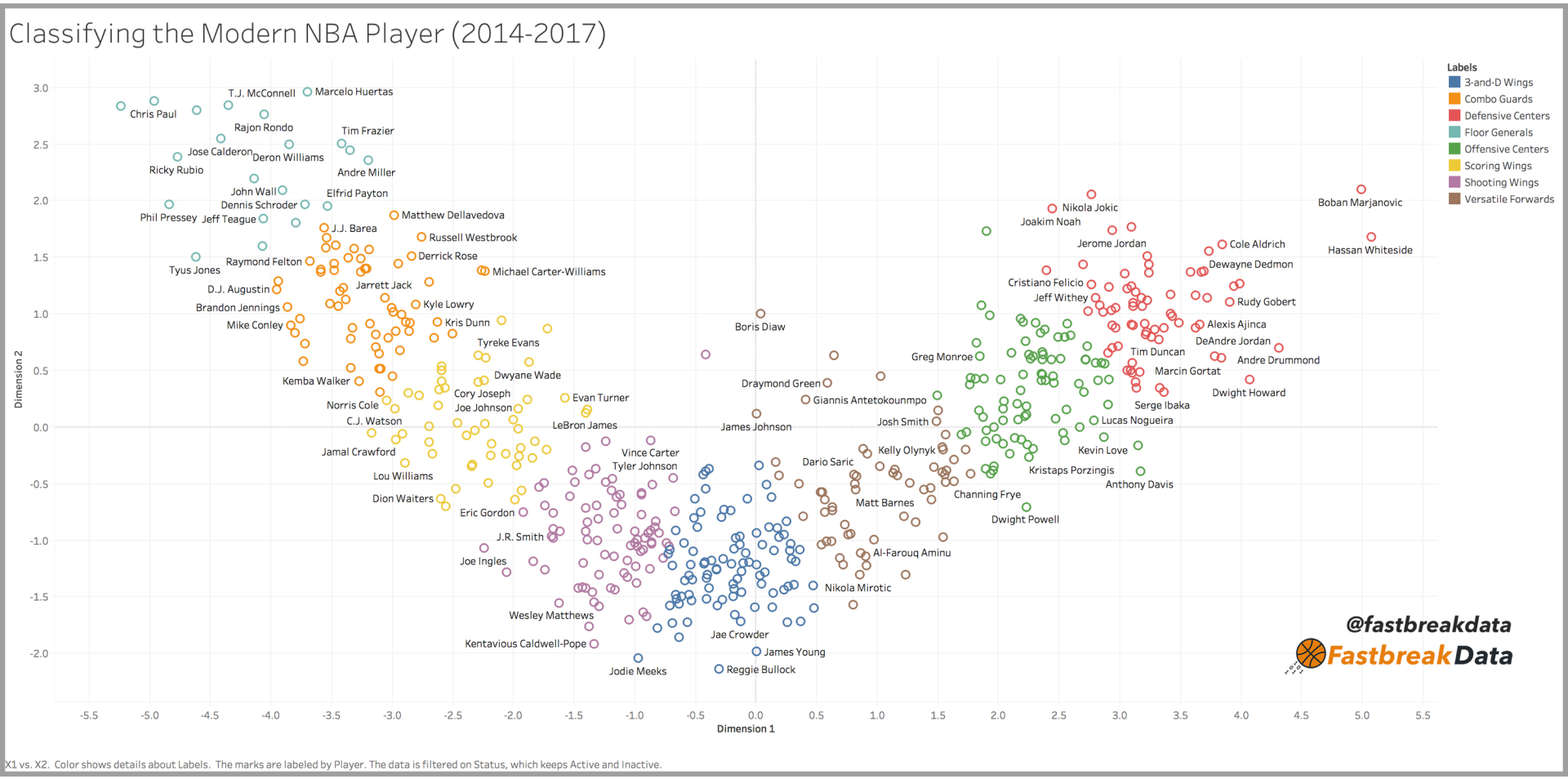
- Can be measured by a certain type of distance metric (距離函數)
- Ex: Euclidean distance (歐式距離)

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$



Applications of Clustering



k-Means Algorithm

An iterative (疊代) clustering algorithm

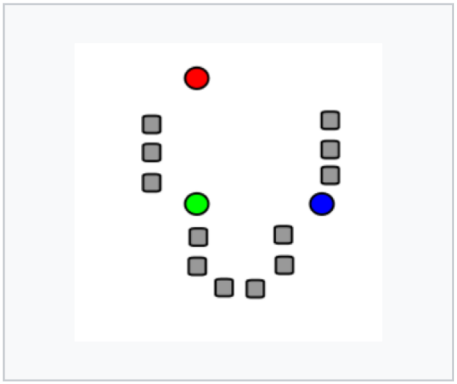
- Initialization (起始): Pick k random points as cluster centers.

1. Assignment (分配): assign data points to closest cluster center
2. Update (更新): change the cluster center to the centroid (幾何中心) of its assigned points

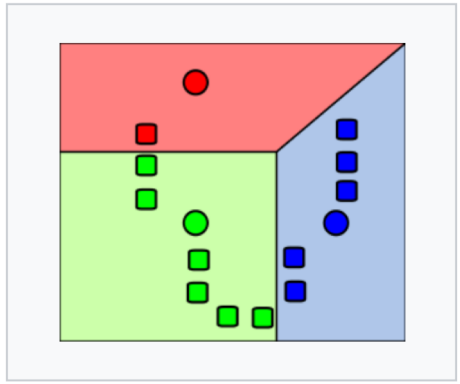
- Stop when no points' assignments change.

k-Means Algorithm

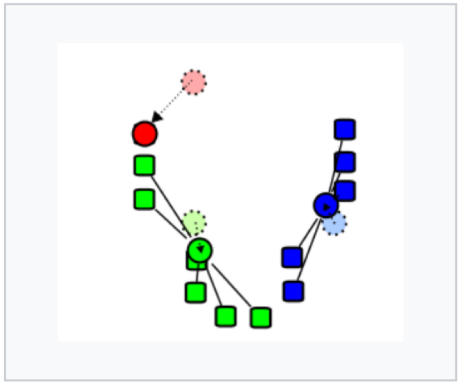
Demonstration of the standard algorithm



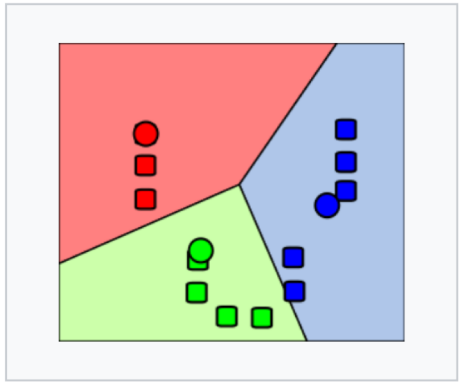
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

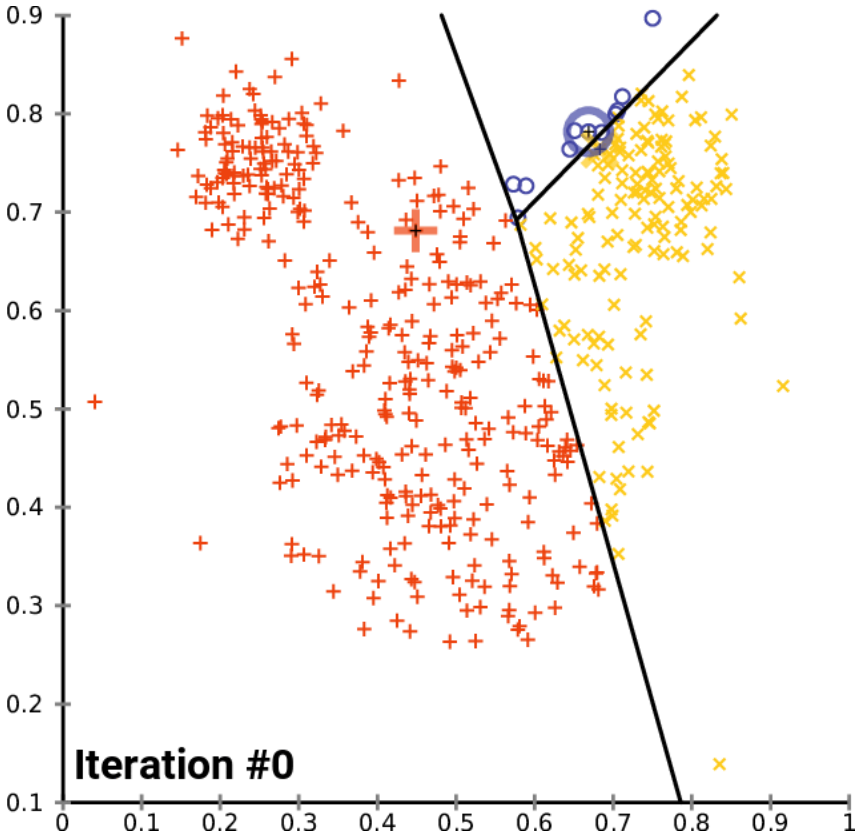


3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

k-Means Algorithm



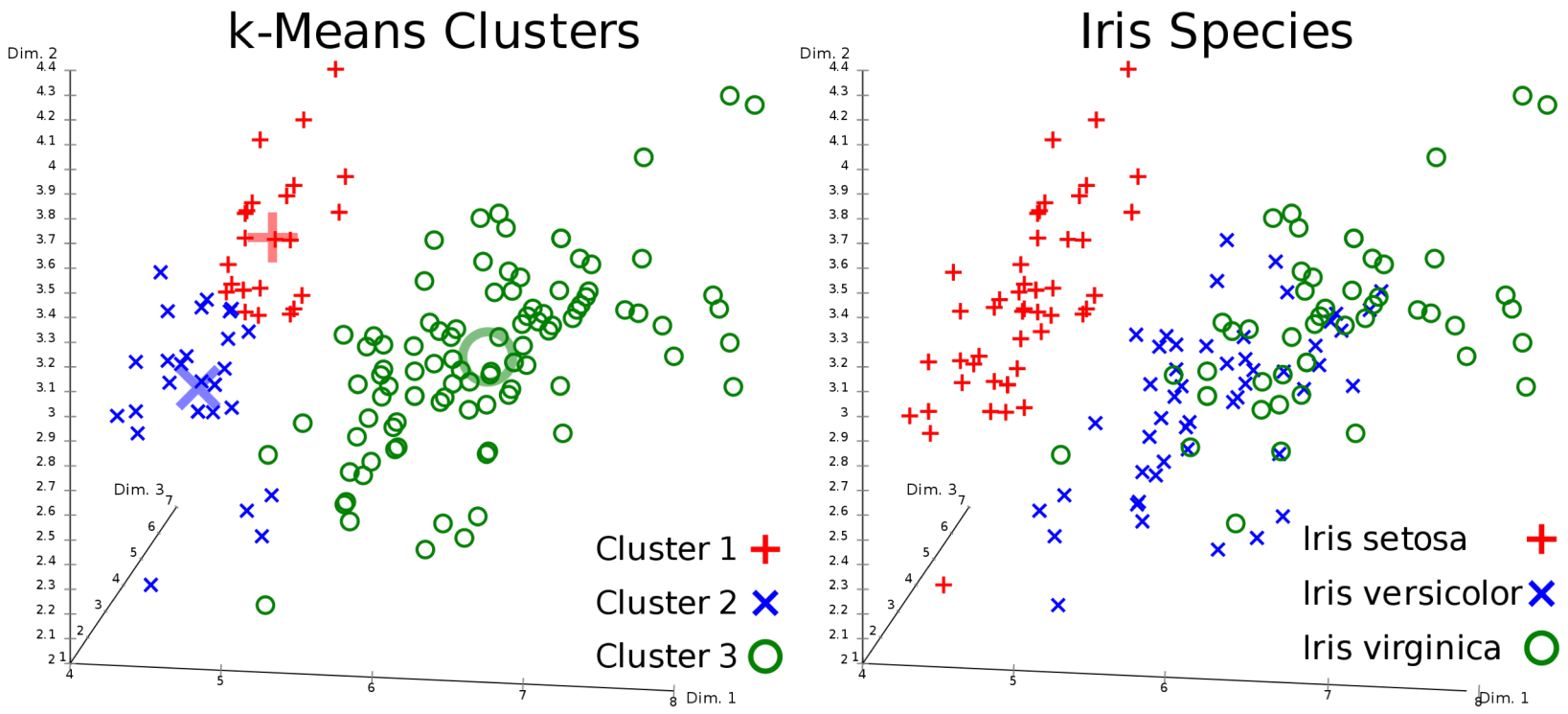
Not a Quiz 5-1

http://tiny.cc/q9c4dz

- **Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means? 連續跑多次k-means分群是否能得到不變的分群結果？**
  - A. Yes 是
  - B. No 否
  - C. Can't say 無從判斷
  - D. None of these 以上皆非

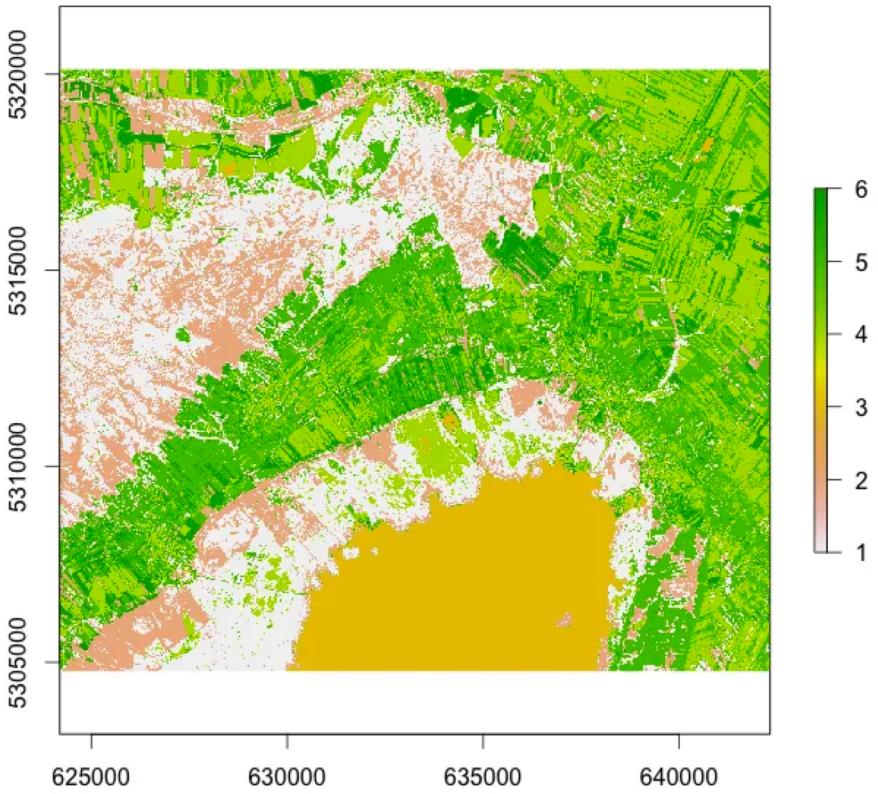
Iris Data Set

Unsatisfactory k-means clustering



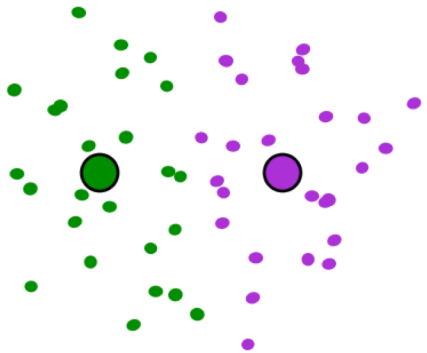
(Wikipedia)

k-Means Clustering of Satellite Imaging

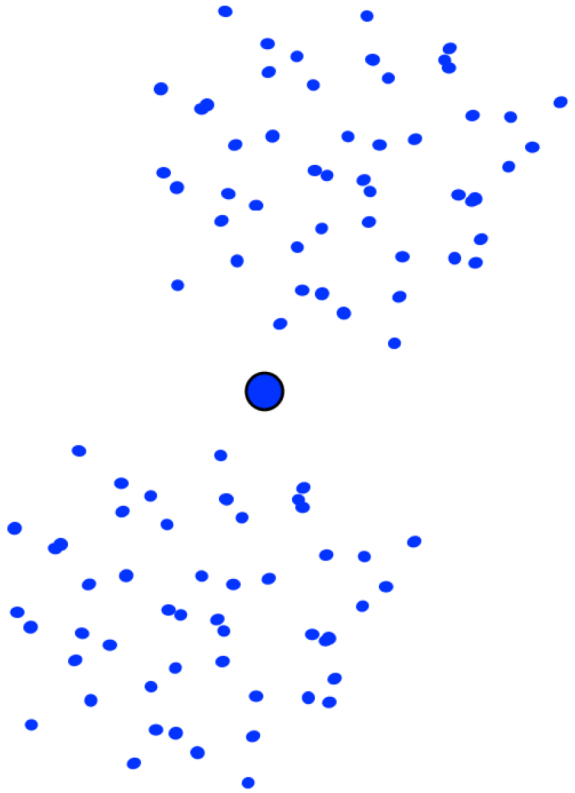


https://www.gis-blog.com/unsupervised-kmeans-classification-of-satellite-imagery-using-r/

k-Means Algorithm

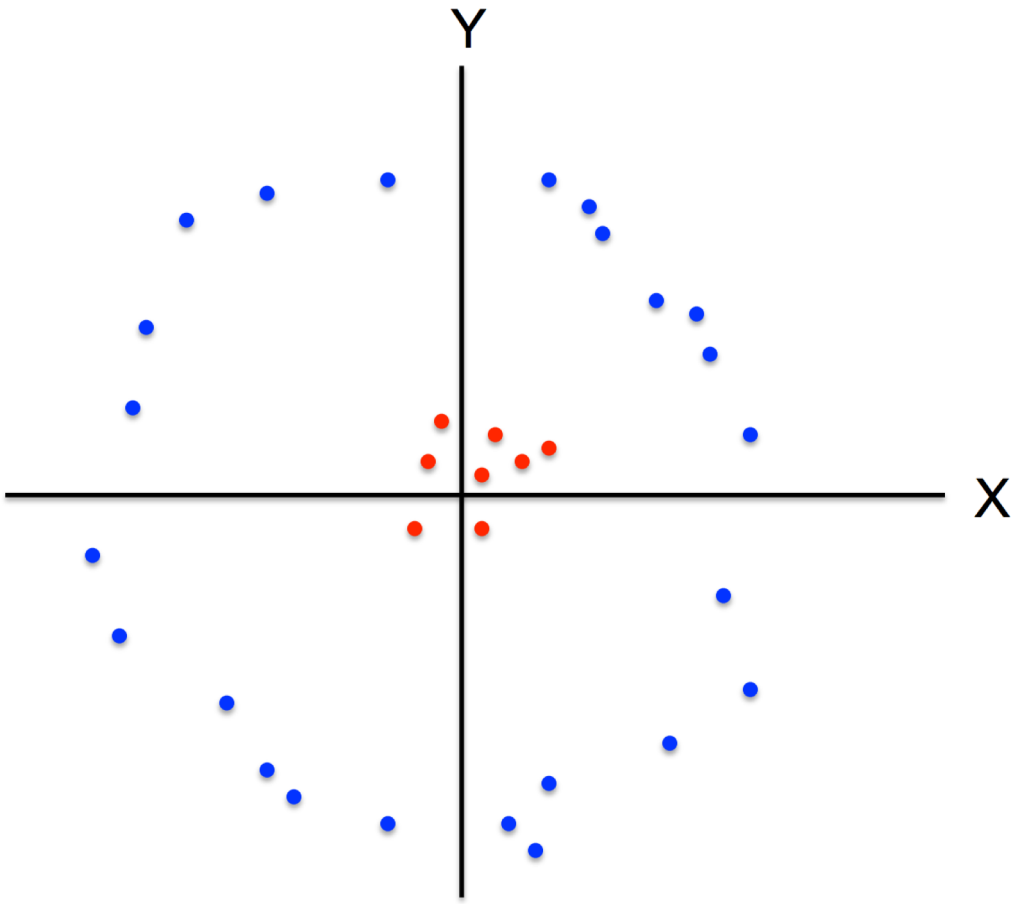


Would be better to have
one cluster here



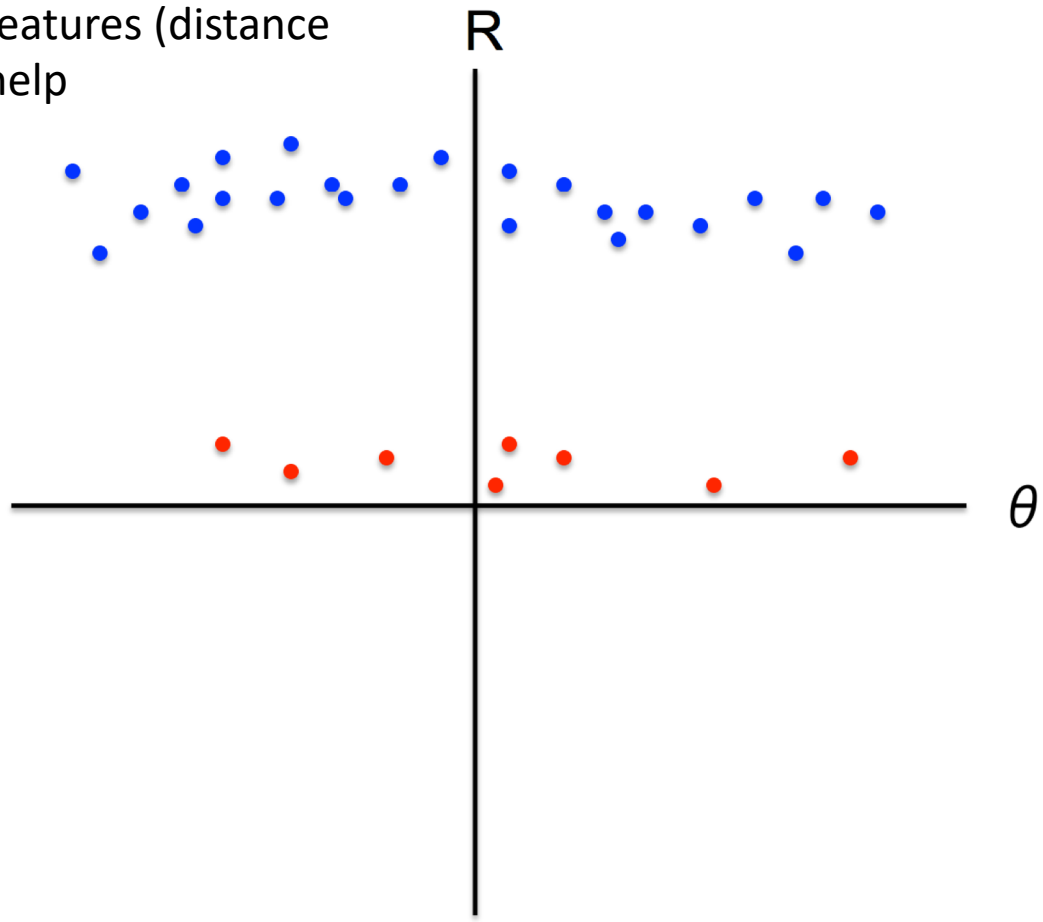
… and two clusters here

k-Means Algorithm



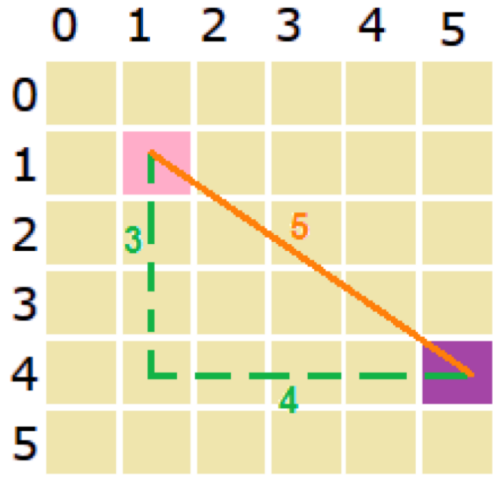
k-Means Algorithm

Changing the features (distance function) can help



Other Distance Metrics

- Manhattan distance

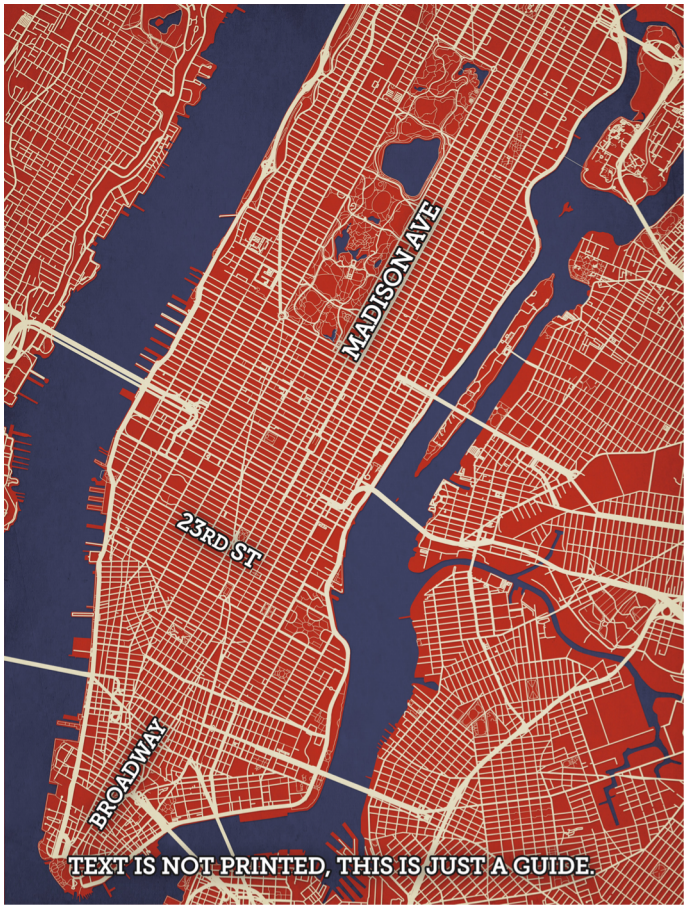


Point P1 = (1,1)

Point P2 = (5,4)

Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

Manhattan distance = $|5-1| + |4-1| = 7$



Other Distance Metrics

- Mahalanobis distance

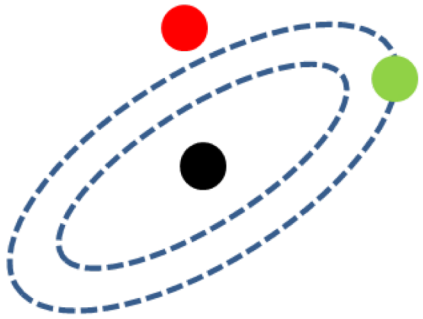
$$d_M(x,y) = \sqrt{(x-y)^T S^{-1}(x-y)}$$

$$= \sqrt{\begin{bmatrix} x_1 - y_1 & x_2 - y_2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}}$$

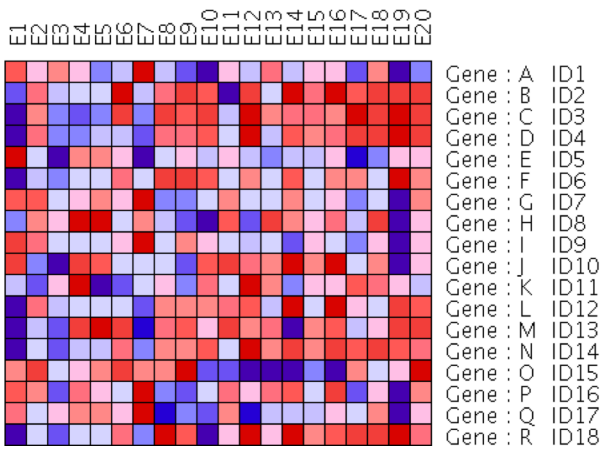
$$= \sqrt{\begin{bmatrix} \frac{x_1 - y_1}{\sigma_1^2} & \frac{x_2 - y_2}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}}$$

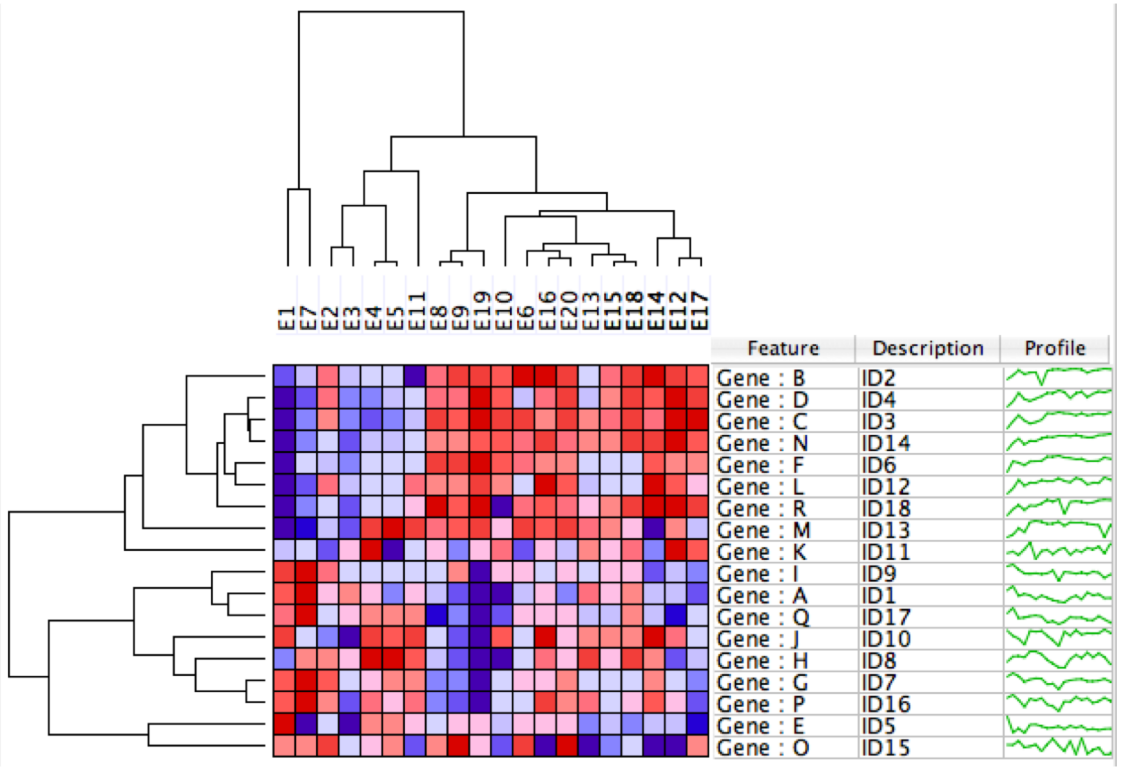
$$= \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2} + \frac{(x_2 - y_2)^2}{\sigma_2^2}}$$

distance(red, black) > distance(green, black)



Hierarchical Clustering

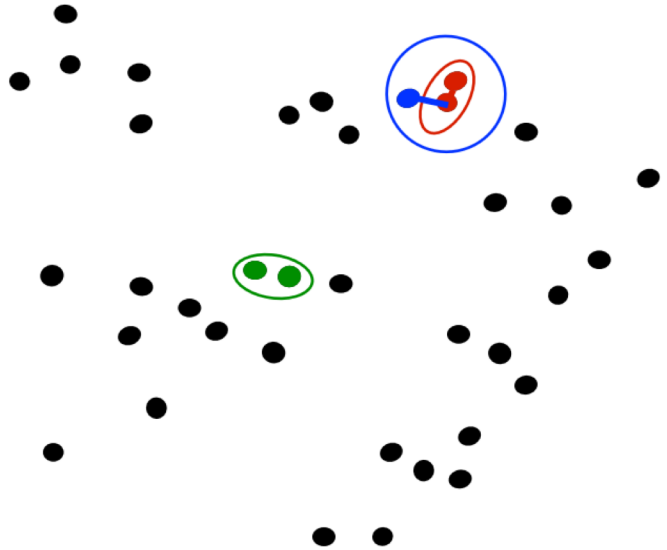


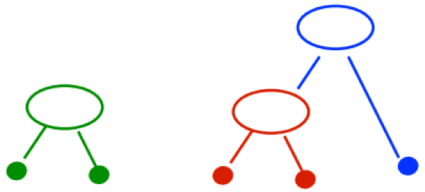


http://labs.mcdb.lsa.umich.edu/labs/schiefel/rootepinet/software.php

Hierarchical Clustering

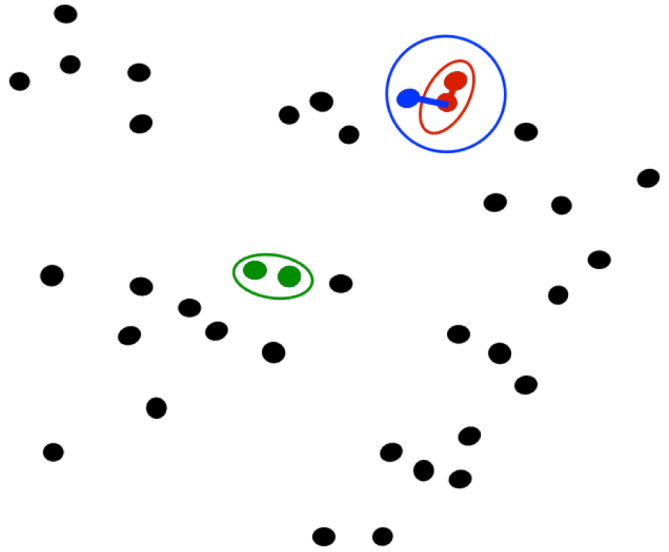
- Start from merging very similar instances
- Incrementally build larger clusters out of smaller clusters

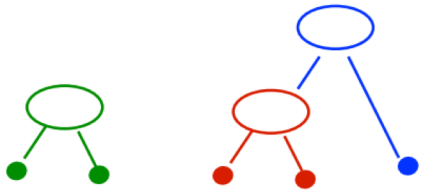




Hierarchical Clustering

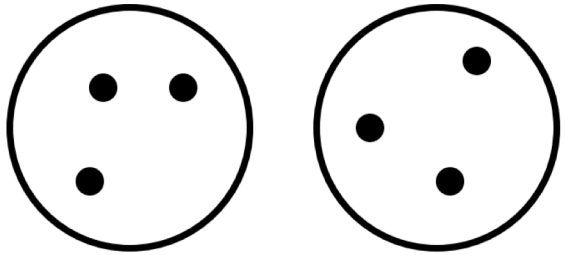
- Algorithm:
  - Initially, each instance in its own cluster
  - Repeat:
    - Pick the two closest clusters
    - Merge them into a new cluster
    - Stop when there is only one cluster left
- Produces not one clustering, but a family of clusters represented by a **dendrogram (樹狀圖)**





Hierarchical Clustering

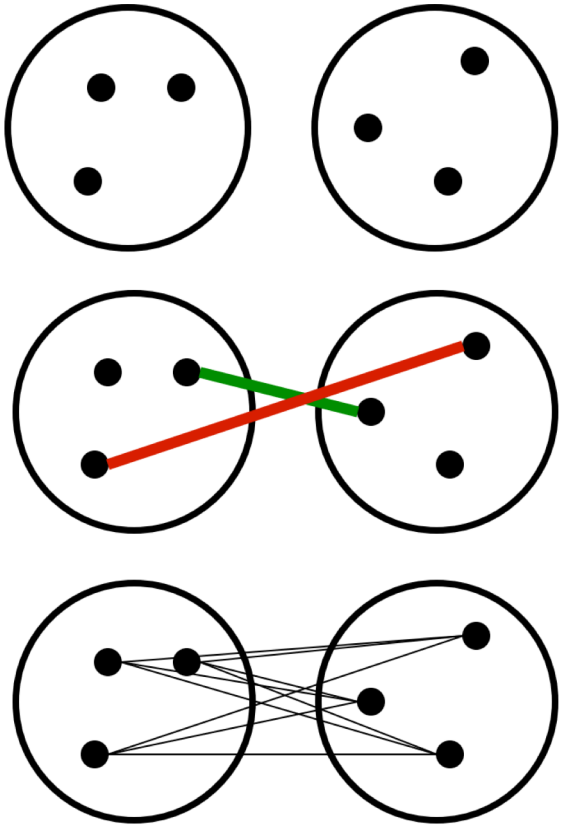
- How to define “closest” for clusters with multiple elements?



Hierarchical Clustering

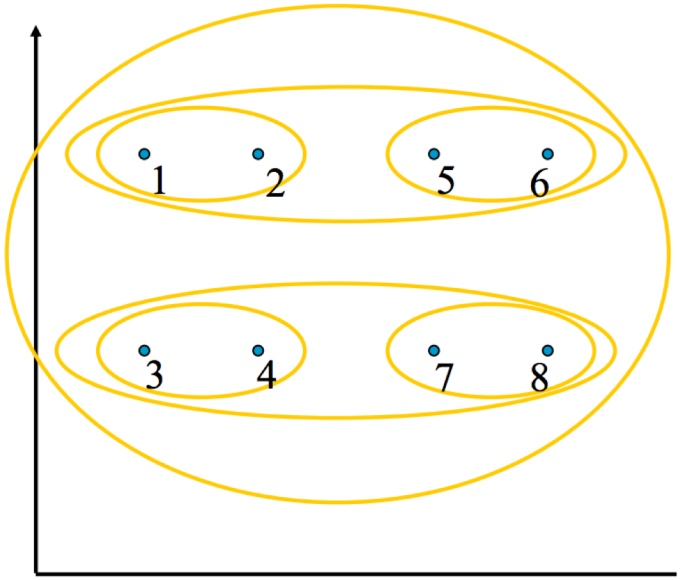
- How to define “closest” for clusters with multiple elements?

- Closest pair (single-link clustering)
- Farthest pair (complete-link clustering)
- Average of all pairs

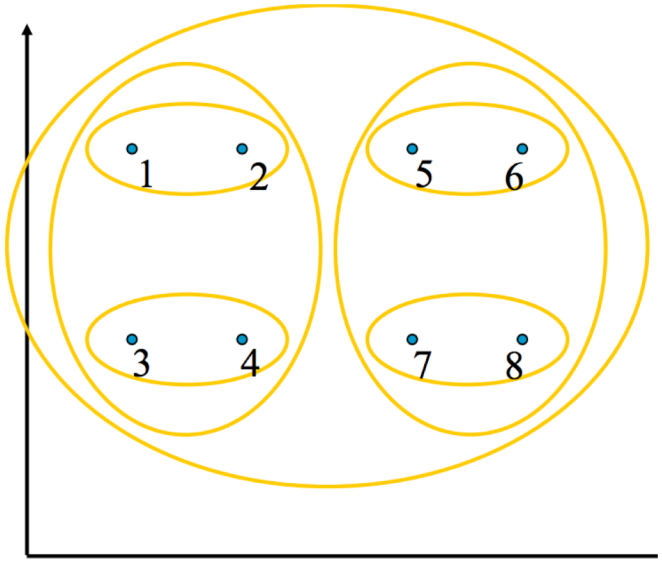


Hierarchical Clustering

Closest pair
(single-link clustering)

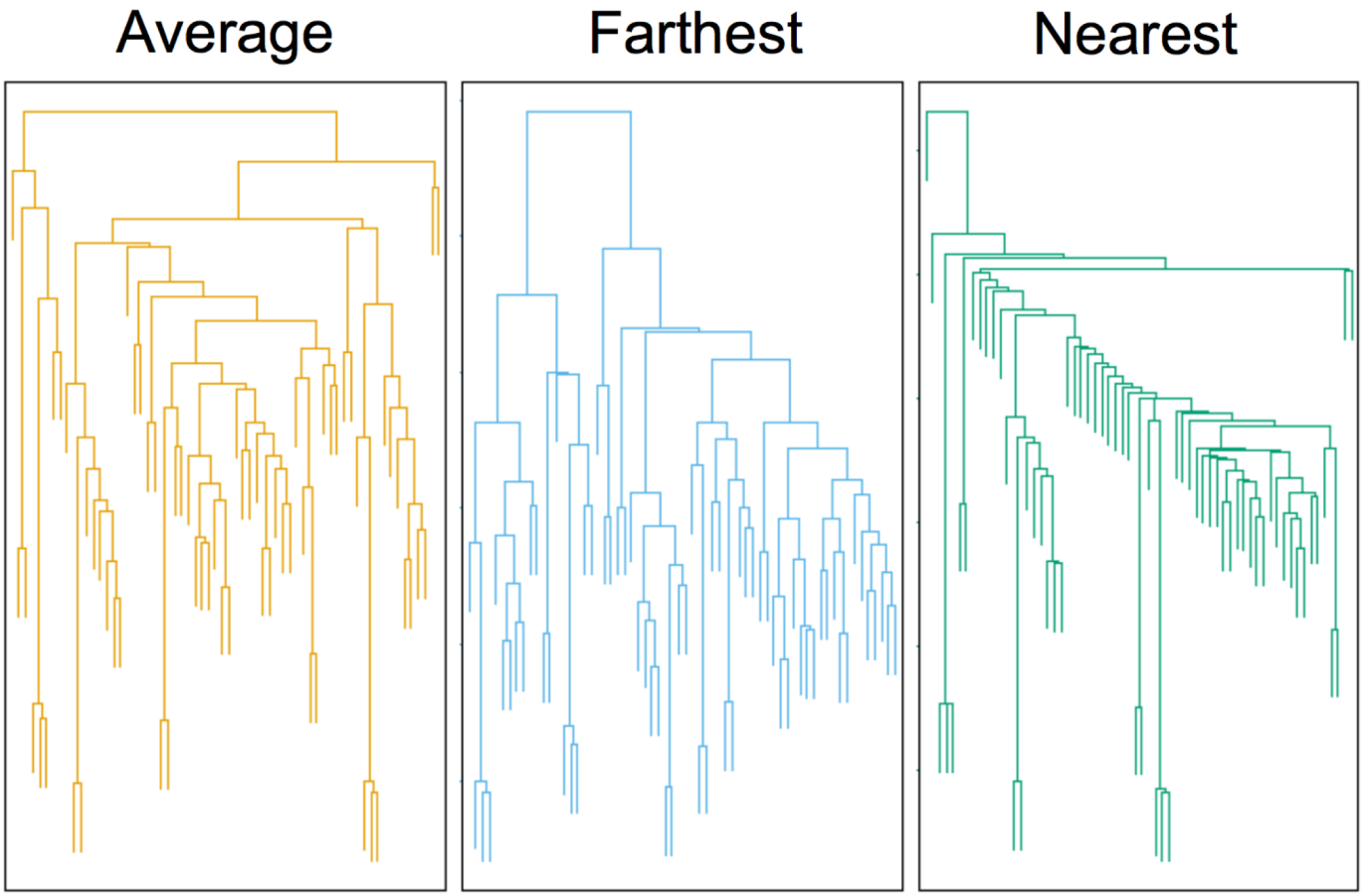


Farthest pair
(complete-link clustering)



[Pictures from Thorsten Joachims]

Hierarchical Clustering



Mouse tumor data from [Hastie *et al.*]

Not a Quiz 5-2

http://tiny.cc/ojd4dz

- **What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithm for the same dataset? 階層式分群時，同一資料集會因何者因素產生不同的樹狀圖？**
  - A. Distance function used 所使用的距離函數
  - B. of data points used 所使用的資料樣本
  - C. of variables used 所使用的變數
  - D. B and C only 僅B和C
  - E. All of the above 以上皆是