k-Means Clustering

부산대학교 정보·의생명공학대학
정보컴퓨터공학부

개요

- Clustering
- k-means clustering
- Example code

CLUSTERING (군집화)

What is Clustering?

❖ 군집의 예

- 백만장자 거주지 데이터 → 베버리힐스 또는 맨하튼 쪽에 군집
- 유권자 데이터 → 사커맘, 은퇴자, 젊은백수 등 으로 군집

❖ 군집화에 정답은 없다

- 대학원생을 젊은 백수와 같은 군집으로 묶는 모델
- 또는 부모님 집에 서식하는 기생충과 같은 군집으로 묶는 모델

(출처: 밑바닥부터 시작하는 데이터 과학)

❖ 왜 군집화?

- (유권자의 예) 전체 유권자 집단을 몇 개의 그룹을 나눠, 각 집단에 특성화된 선거전략을 세울 수 있다.
- Labeling is expensive
- Gain insight into the structure of the data

What is Clustering?

❖ 지도학습(supervised learning):

- 학습 데이터에 예측 목표가 되는 속성이 포함된 경우 → Labels provided
- 분류 (예측값: 카테고리), 회귀분석(예측값: 수치) 등

❖ 비지도학습 (unsupervised learning):

- 학습 데이터에 예측 목표값이 주어지지 않음→ Labels not provided
- 데이터에 바로 드러나지 않는 숨은 구조를 찾는 기술: e.g., 군집화, 차원축소
- 군집화(Clustering): 데이터 항목 간의 유사도를 기반으로 비슷한 항목들을 묶어 그룹을 찾는 기법

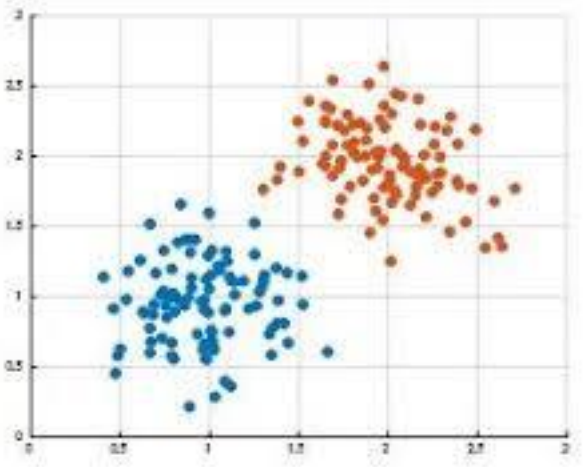
출처: https://github.com/jeonghunyoon/machine-learning-lecture-notes

What is Clustering?

- a.k.a. 군집분석 (Cluster Analysis)
- The process of partitioning a set of data objects (or observations, or instances) into subsets.
- Each subset is a **cluster**, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters.
  - A cluster is a collection of data objects which are
    - Similar (or related) to one another within the same group (i.e., cluster)
    - Dissimilar (or unrelated) to the objects in other groups (i.e., clusters)
- The partitioning is performed by the clustering algorithm
  - Can lead to the discovery of previously unknown groups within the data

What is Clustering?

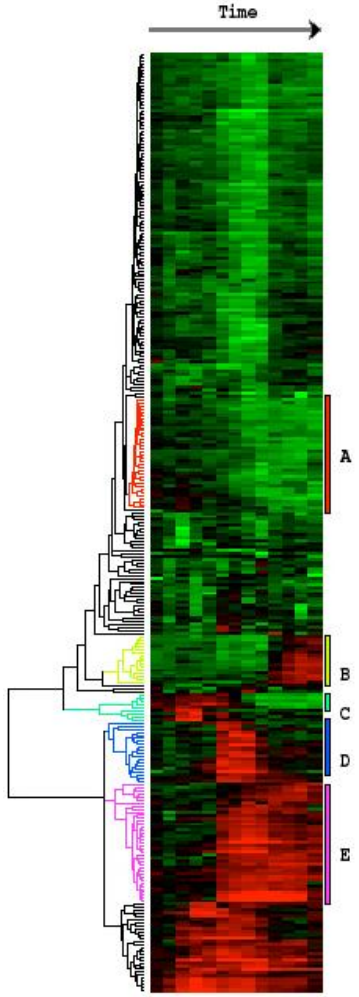
- Cluster analysis (or clustring, data segmentation, …)
  - Given a set of data points, partition them into a set of group(i.e., clustsers) which are as similar as possible
- Cluster analysis is unspervised learning (i.e., no predfined classes)
  - This constrasts with classification (i.e., supervised learning)
- Typical ways to use/apply cluster analysis
  - As a stand-alone tool to get insight into data distribution, or
  - As a preprocessing (or intermediate) step for other algorithms



- A good clustering method will produce high quality clusters which should have
  - High intra-class similarity: Cohesive within clusters
  - Low inter-class similarity: Distinctive between clsters

Clustering examples

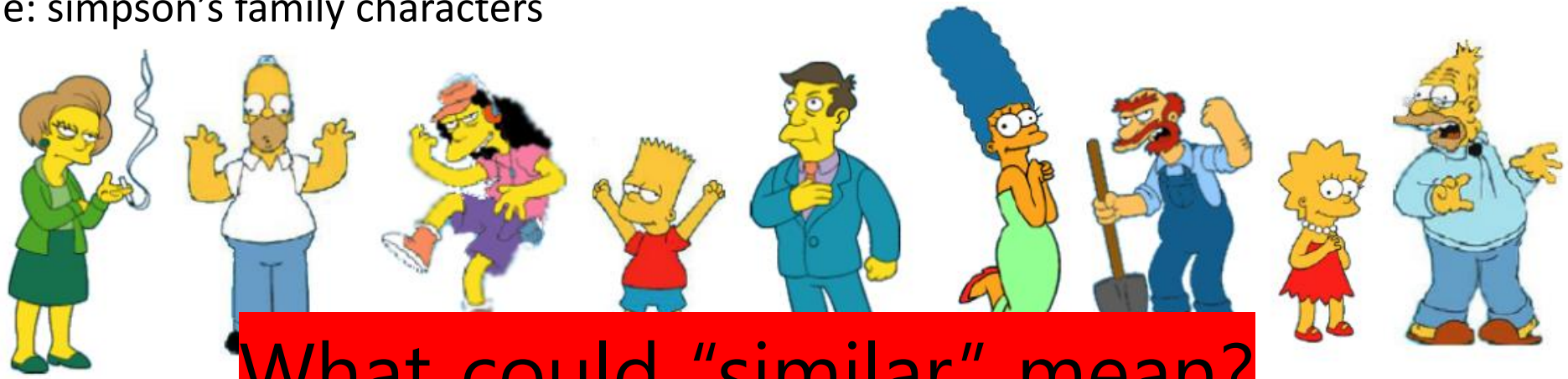
Image segmentation:
- break up the image into meaningful or perceptually similar regions



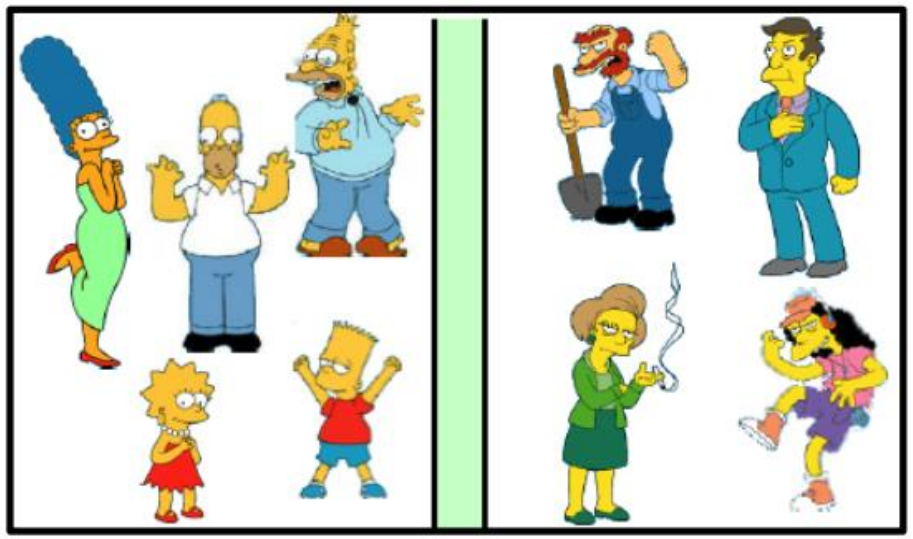
Gene expression data

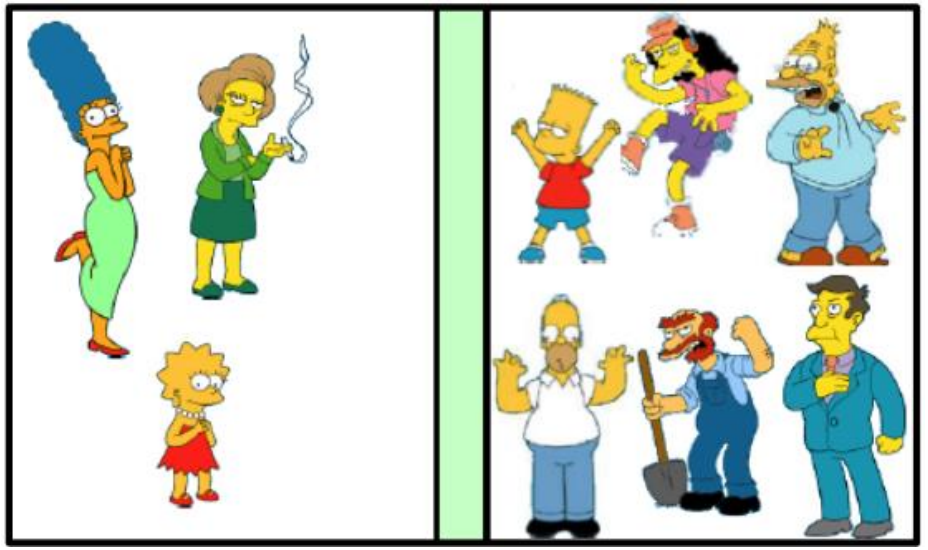
Clustering

- Basic idea: group together similar objects
- Example: simpson's family characters



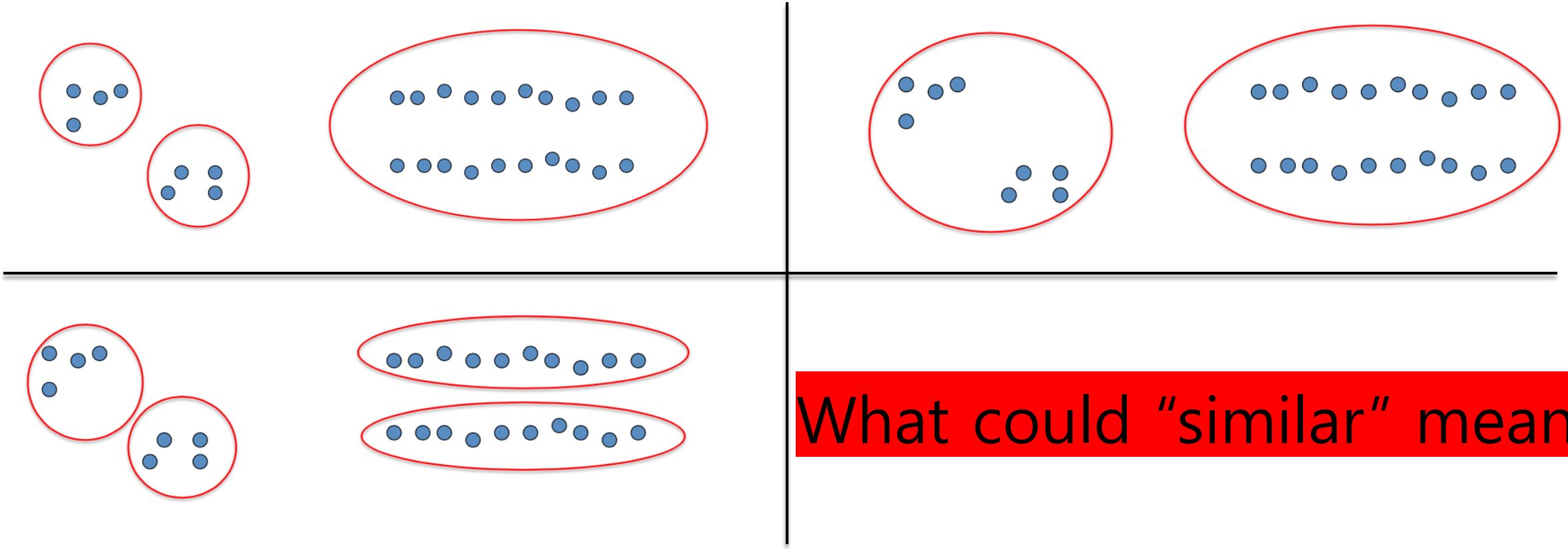
What could "similar" mean?





Clustering

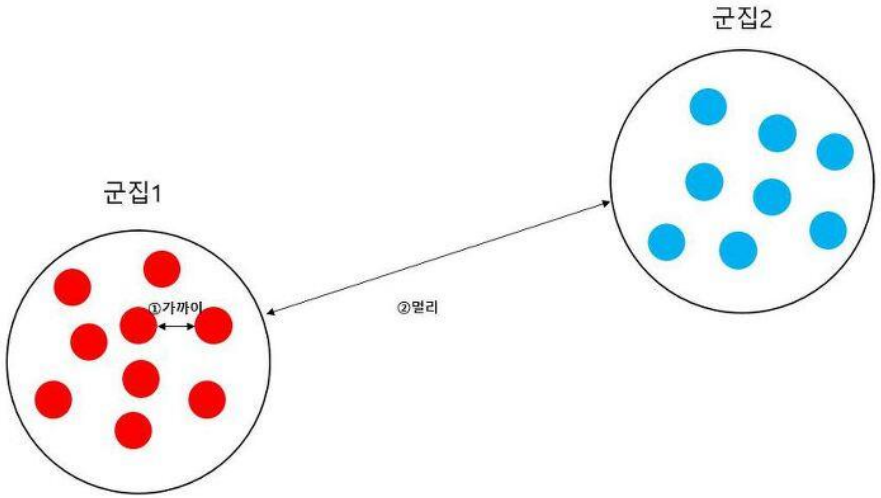
- Basic idea: group together similar objects
- Example: 2D point patterns

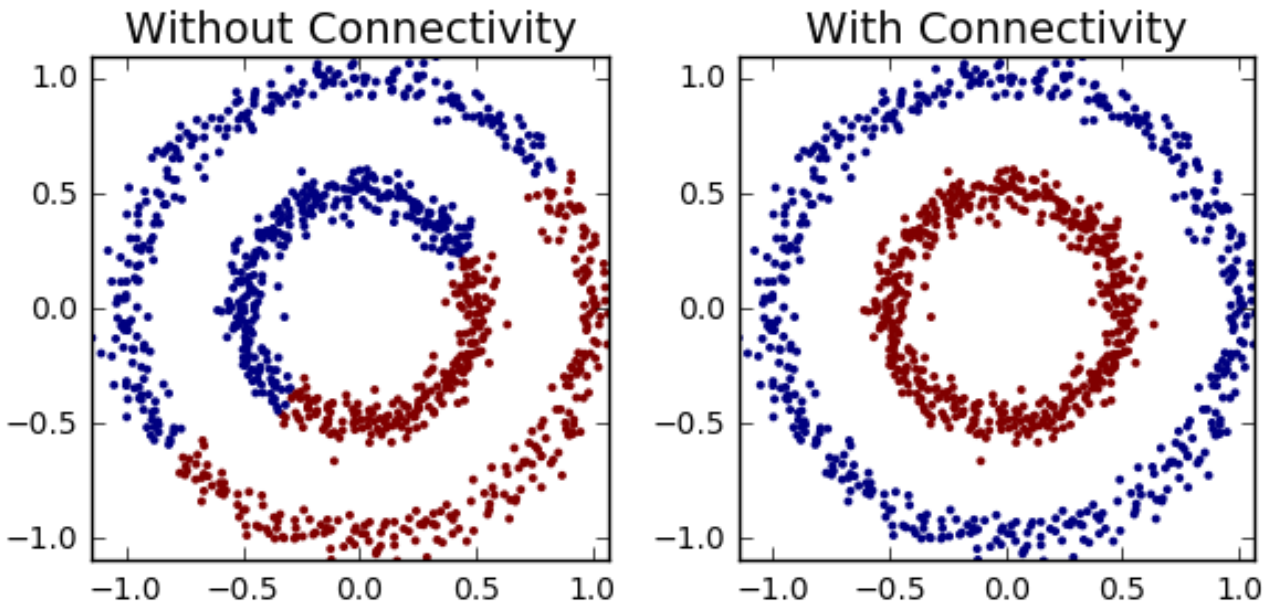


What could “similar” mean?

Similarity (or dissimilrity) measure

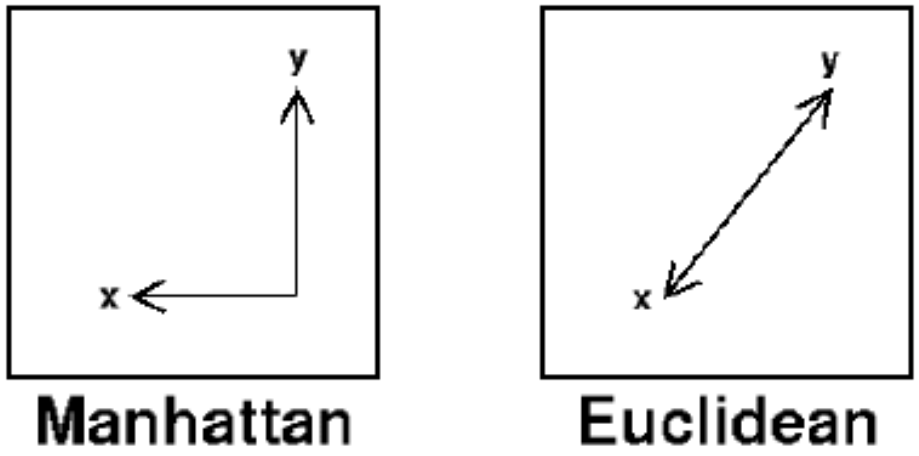
- Distance-based (e.g., Euclidian, road network, vector)
  - Tend to find spherical clusters
- Connectivity-based (e.g., density or contiguity)
  - Can find clusters of arbitrary shape

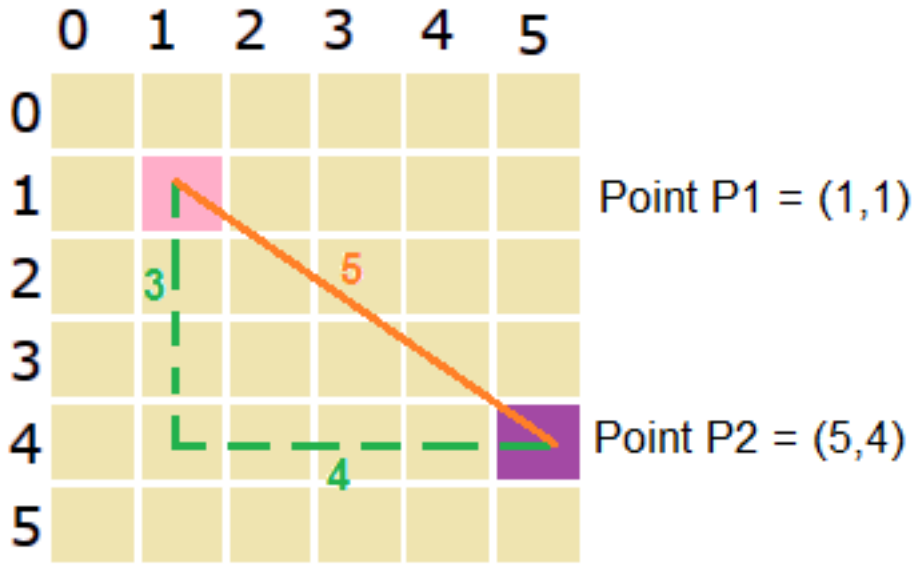




Similarity (or dissimilrity) measure

❖ Euclidean distance vs. Manhattan distance





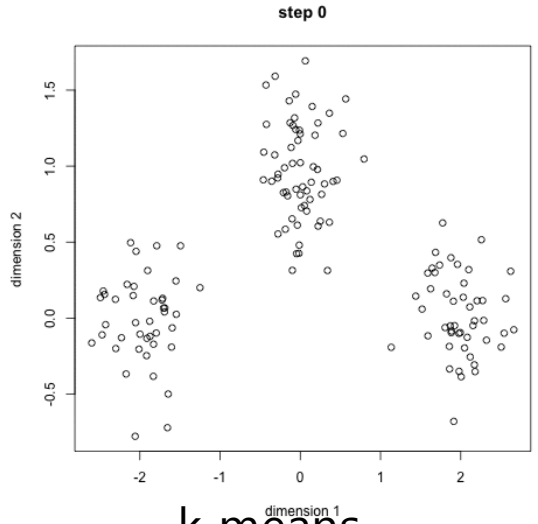
What properties should a distance measure have?

- Symmetric:
  $D(A,B) = D(B,A)$
- Positivity, and self-similarity:
  $D(A,B) \geq 0$, $D(A,B) = 0$ iff $A=B$
- Triangle inequality:
  $D(A,B) + D(B,C) \geq D(A,C)$

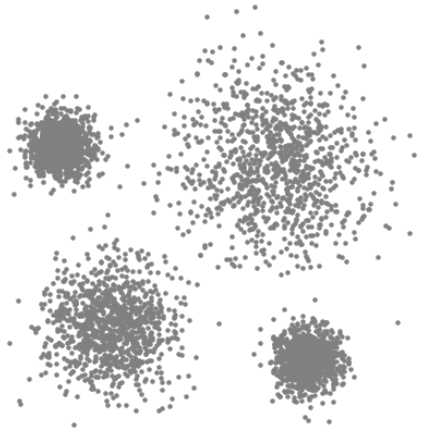
Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2} = 5$

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

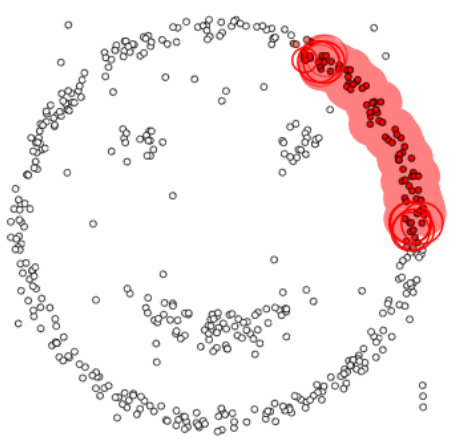
Clustering methods



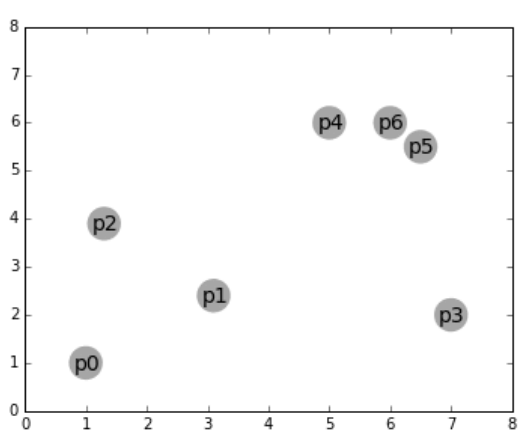
k-means
(Partitioning method)

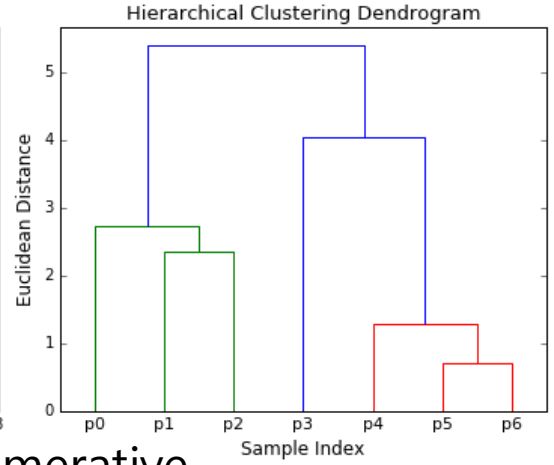


Mean-Shift
(Grid-based)



DBSCAN
(Density-based)





Agglomerative
(Hierarchical)