数据挖掘导论

Introduction to Data Mining

Basic Clustering

数据智能实验室
DATA INTELLIGENCE LABORATORY
DI-LAB
浙江大学
Zhejiang University

Agenda

- Clustering: Basic Concepts
- K-means Clustering
- Density-Based Clustering
- Hierarchical Clustering
- Cluster Evaluation
- Summary

Clustering

Basic Idea

- ✓ Group together similar instances

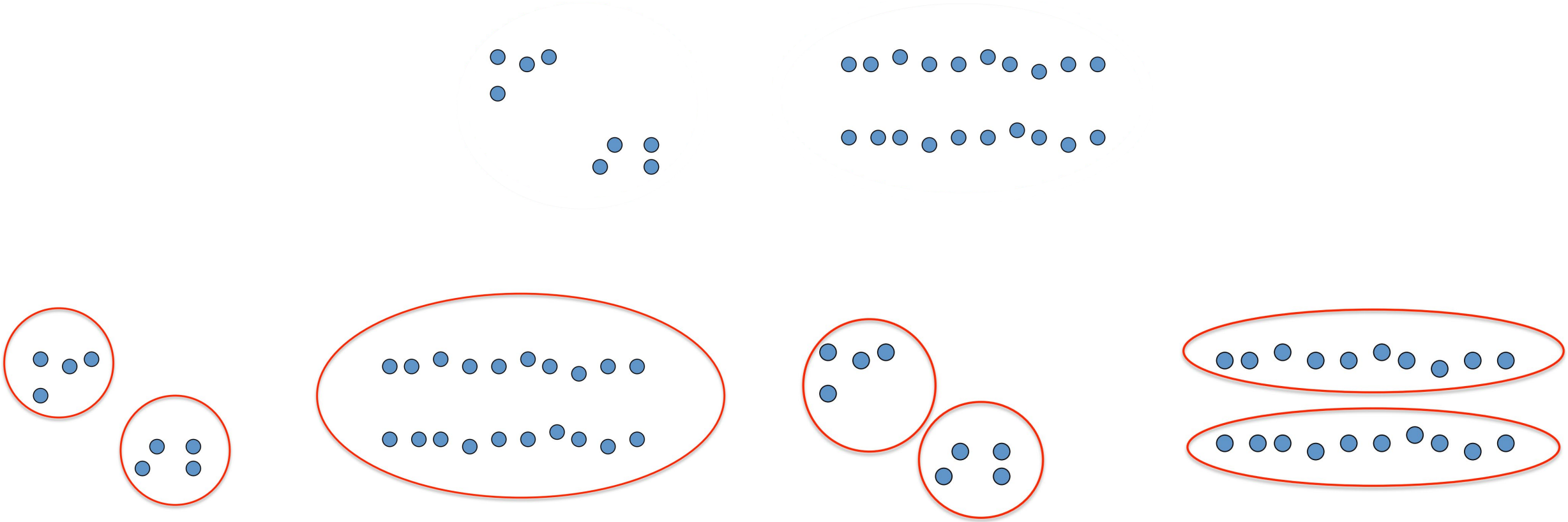
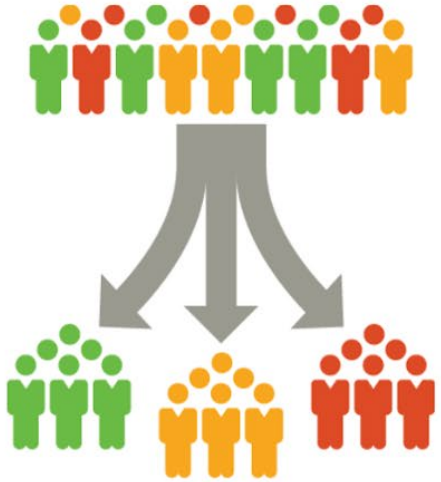


Image source: https://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf

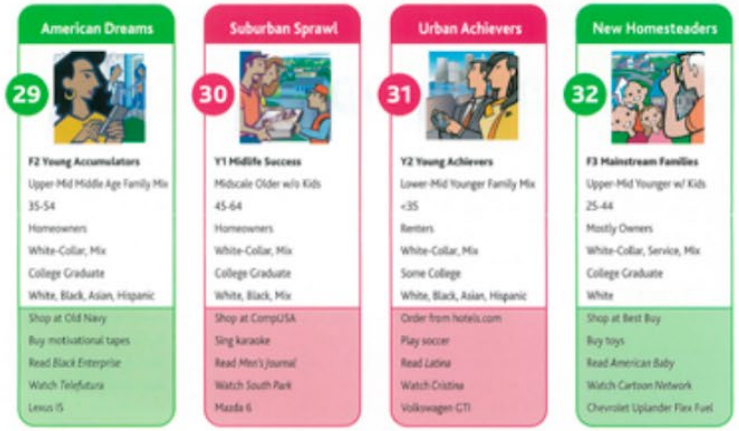
Applications -- Marketing

Auto Segmentation



Interpretable

Data Exploration



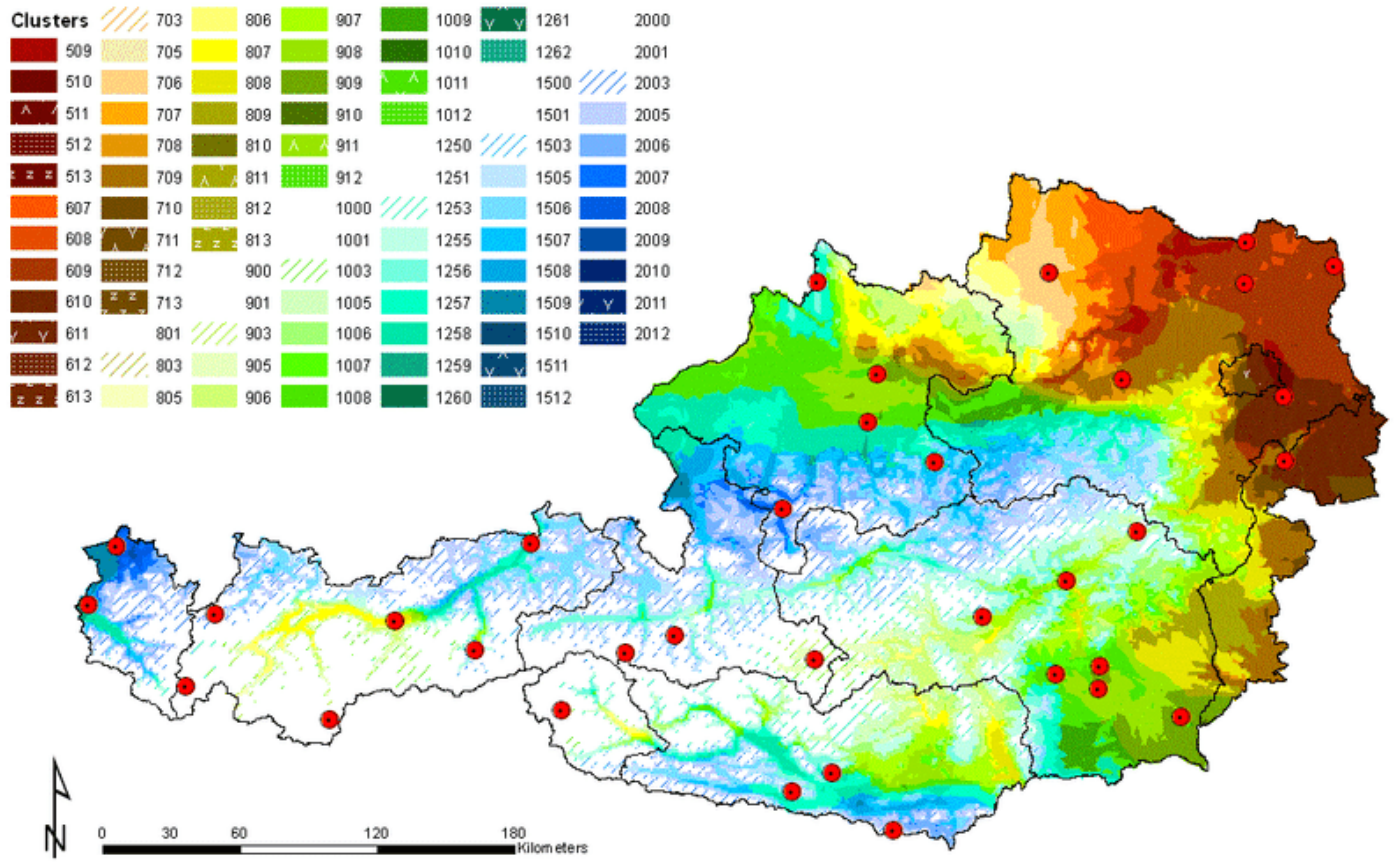
Explainable

Customer clustering for targeted advertising

Image source: https://miro.medium.com/max/1400/0*dIl6S6hqG3M7dYoL

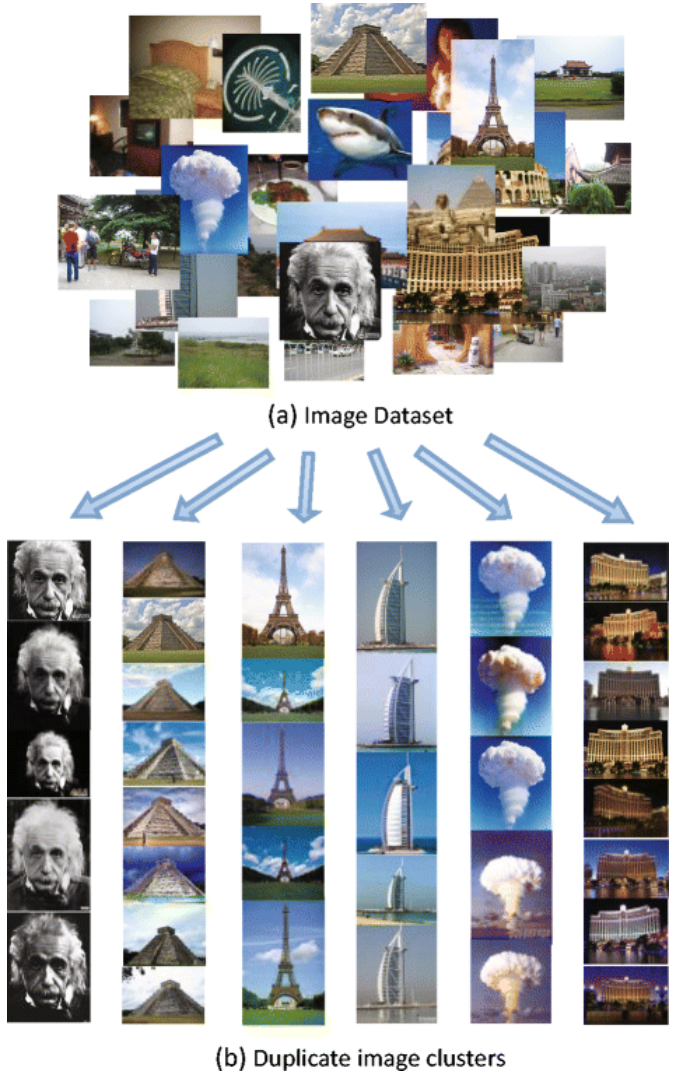
Applications – Image Segmentation

Applications – Heat Map



Climate clusters based on precipitation and temperature classes

Applications – Near-Duplicate Identification



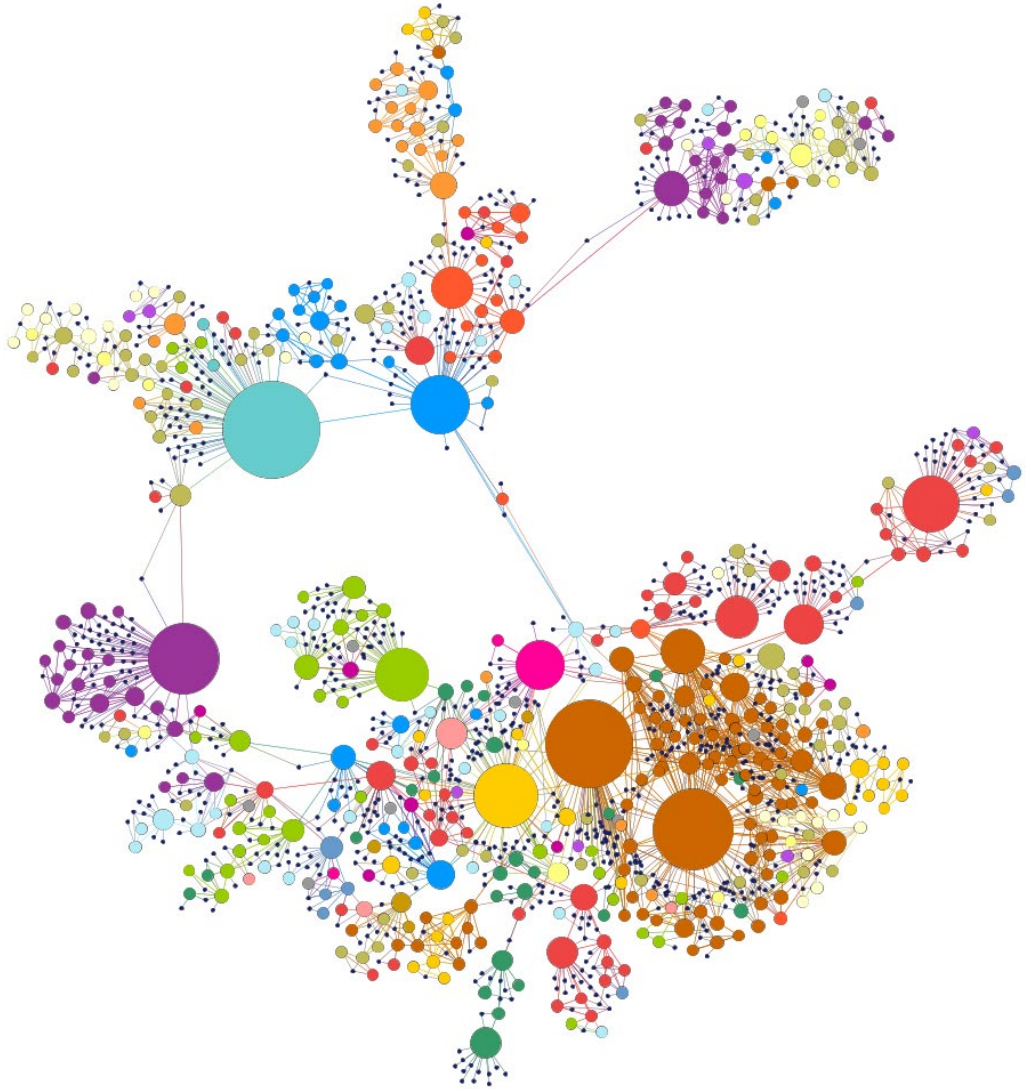
(a) Image Dataset

(b) Duplicate image clusters

Applications – Album Management

Image source: https://miro.medium.com/max/1400/0*dIl6S6hqG3M7dYoL

Applications – Human Disease Network



Nodes represent diseases and two diseases are connected to each other if they share at least one gene in which mutations are associated with both diseases

The resulting network is naturally and visibly clustered according to major disease classes (e.g., bone, cancer, cardiovascular, skeletal, or metabolic diseases; each disease class is represented by a different color).

Example from: http://snap.stanford.edu/deepnetbio-ismb/ipynb/Human+Disease+Network.html

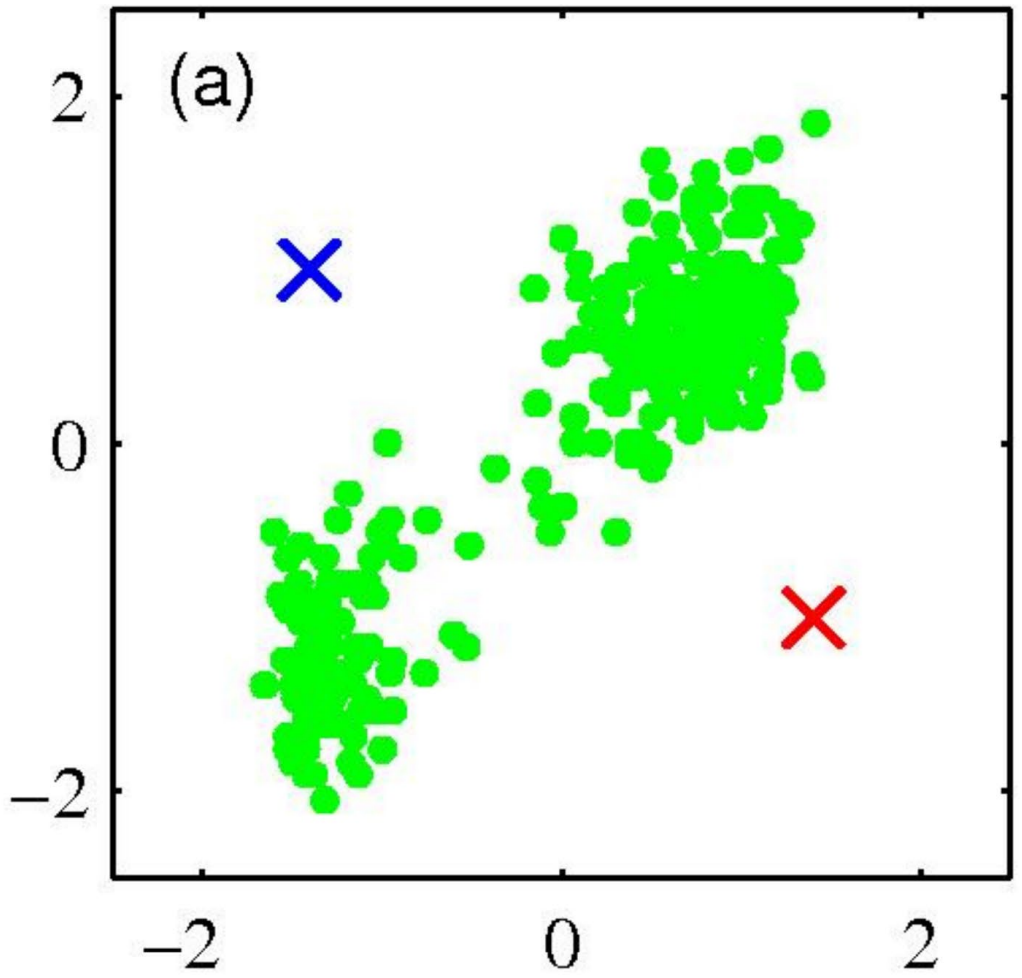
Agenda

- Clustering: Basic Concepts
- K-means Clustering
- Density-Based Clustering
- Hierarchical Clustering
- Cluster Evaluation
- Summary

K-means Clustering

- **Input**
  - ✓ A set of points S to be clustered
  - ✓ A distance measure
  - ✓ A parameter k
- **Output**
  - ✓ k clusters, each is a subset of S
- **K-means is an iterative clustering algorithm**
  - ✓ Initialize: Pick k random points as clusters
  - ✓ Iteration:
    - Assign data points to the closest cluster center
    - Change the cluster center to the average of its assigned points
  - ✓ Stop: when no points' assignments change

K-means Clustering

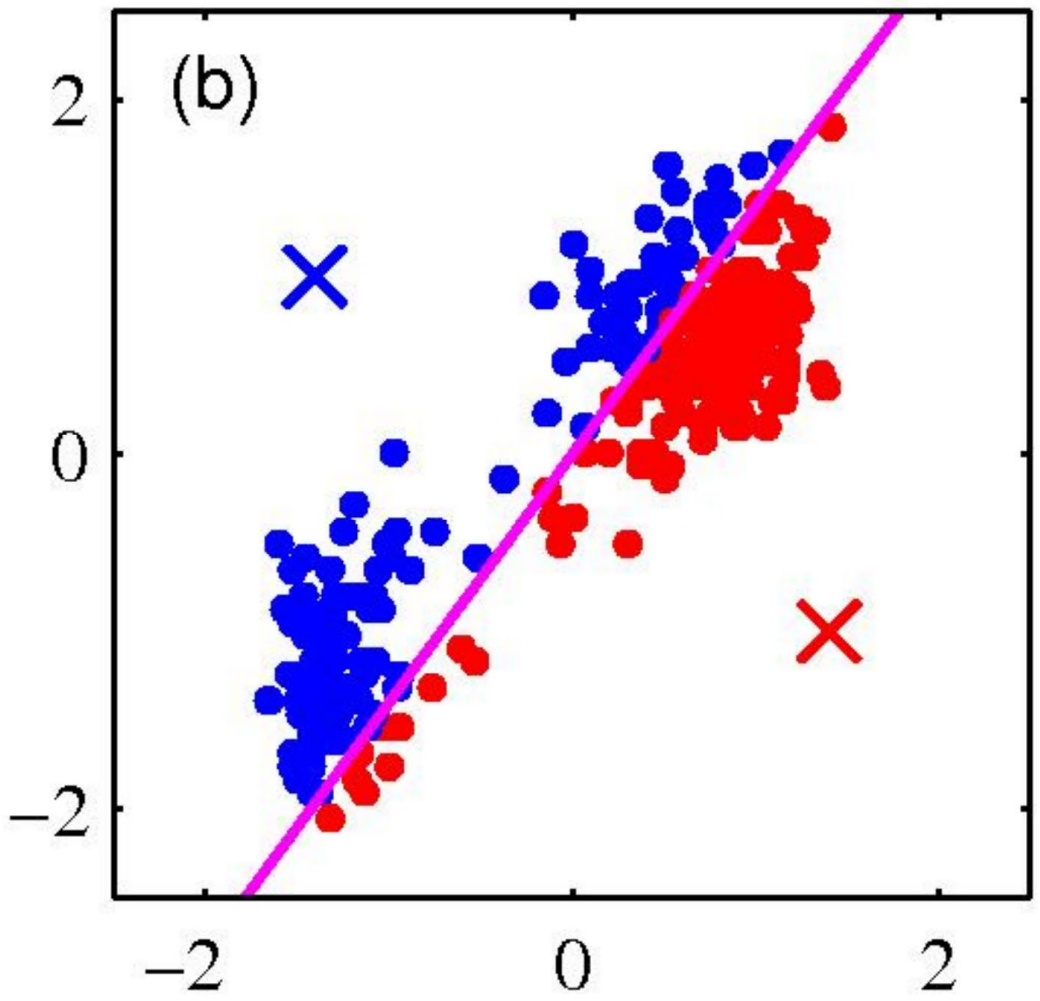


- Pick *K* random points as cluster centers (means)

Shown here for *K*=2

Example source: https://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture14.pdf

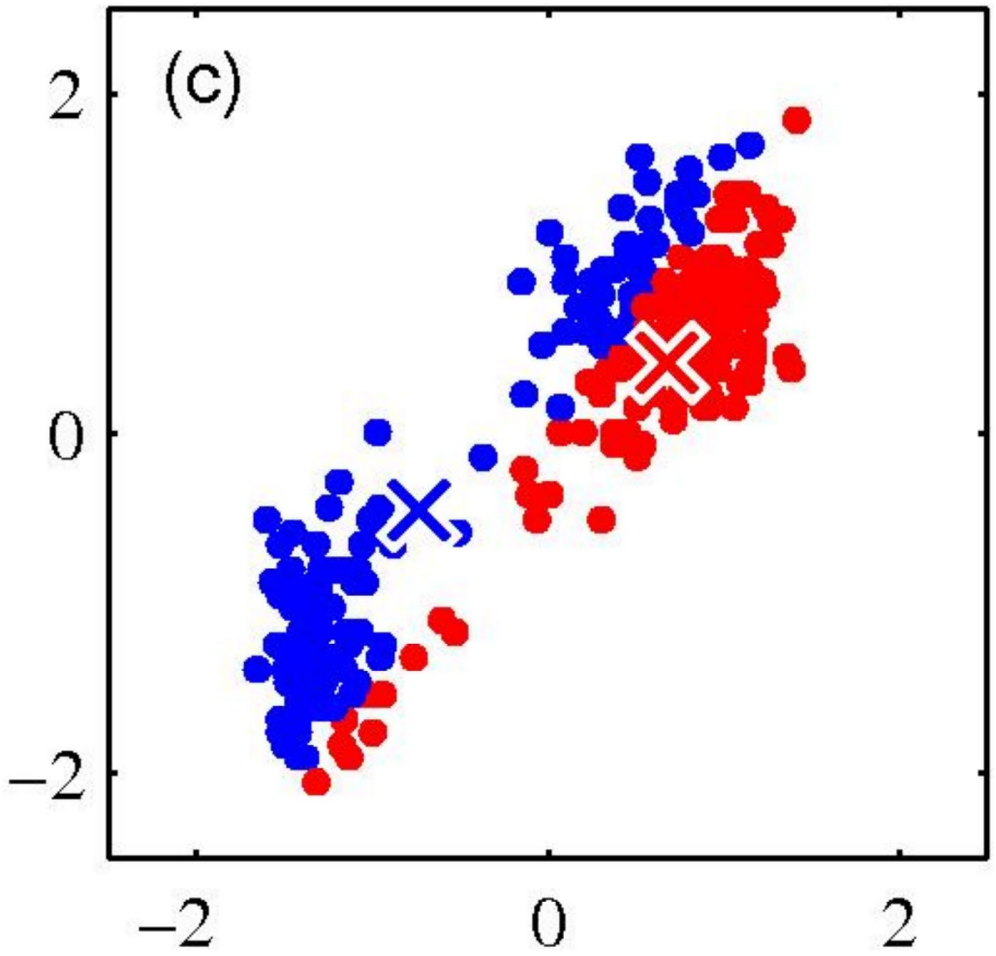
K-means Clustering



Iterative Step 1

- Assign data points to closest cluster center

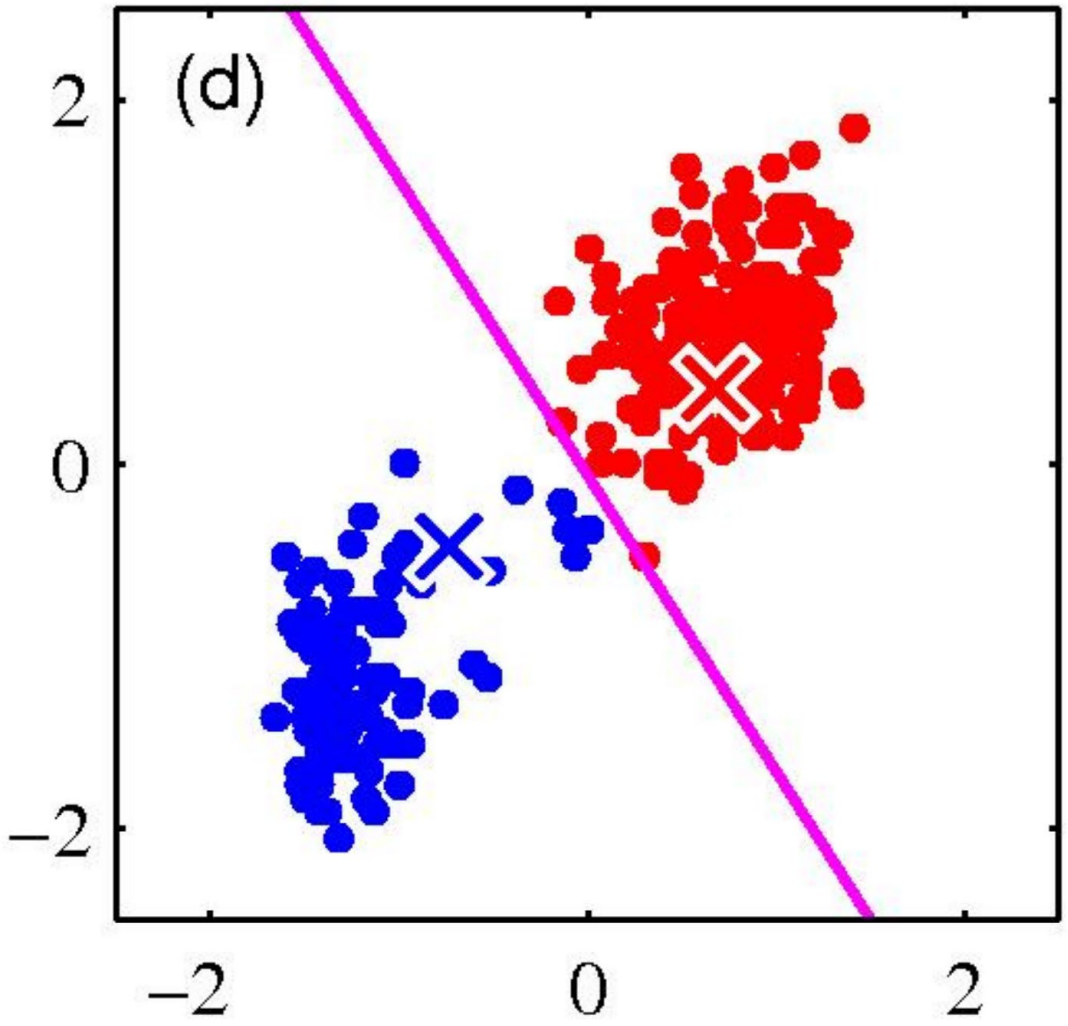
K-means Clustering



Iterative Step 2

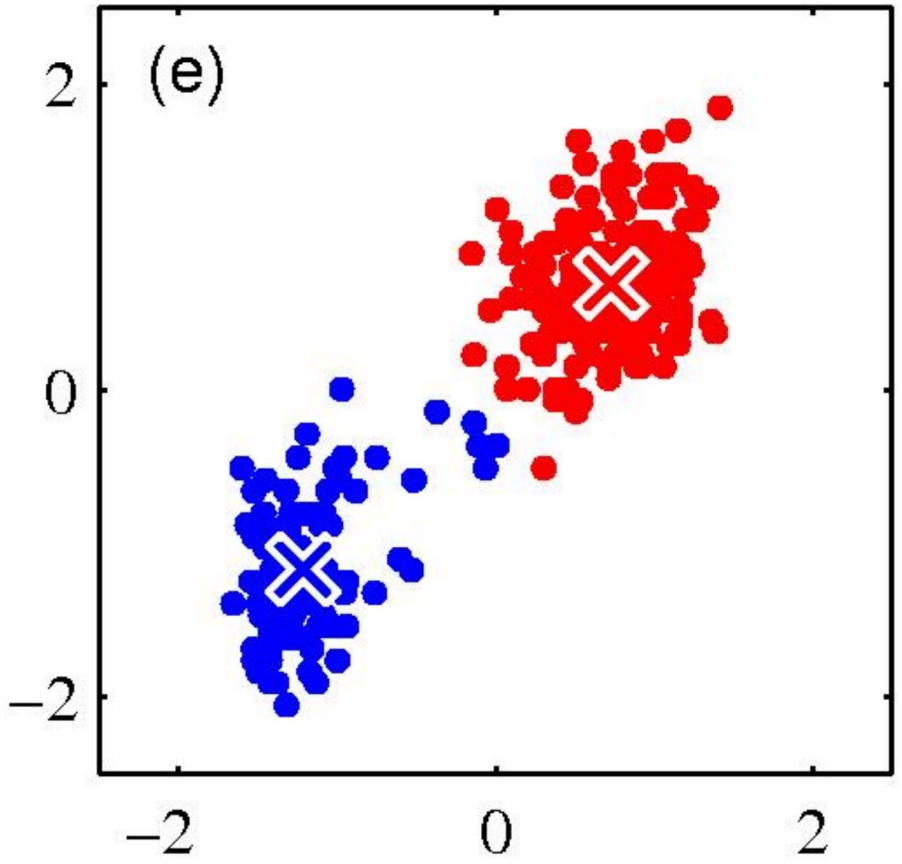
- Change the cluster center to the average of the assigned points

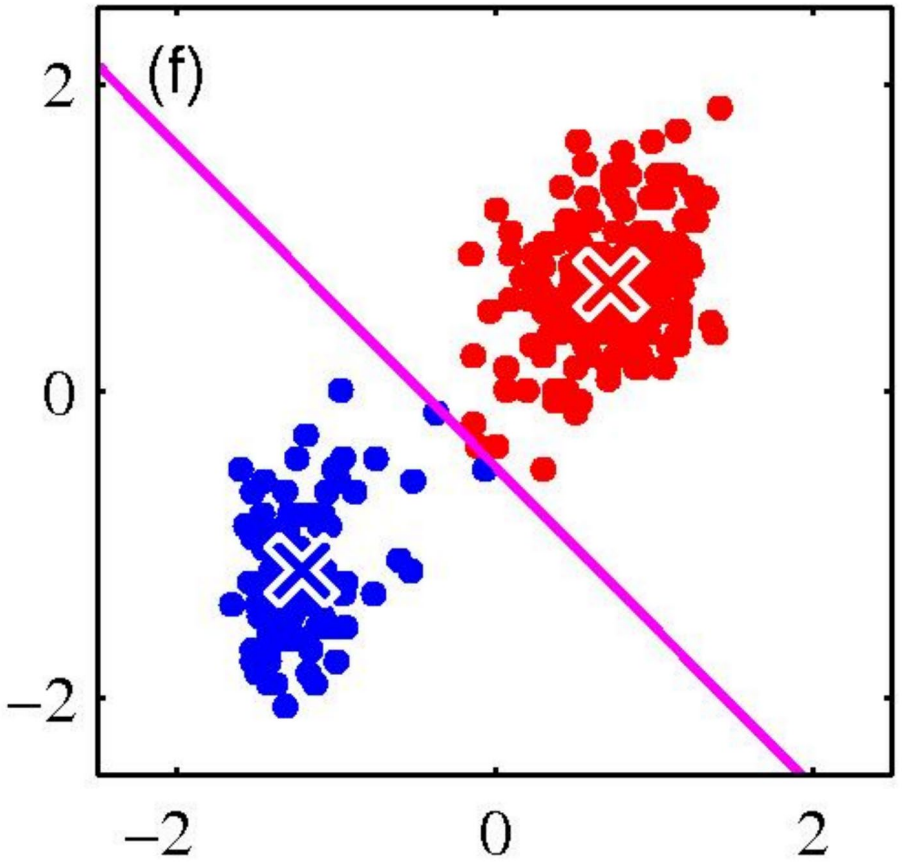
K-means Clustering



- Repeat until convergence

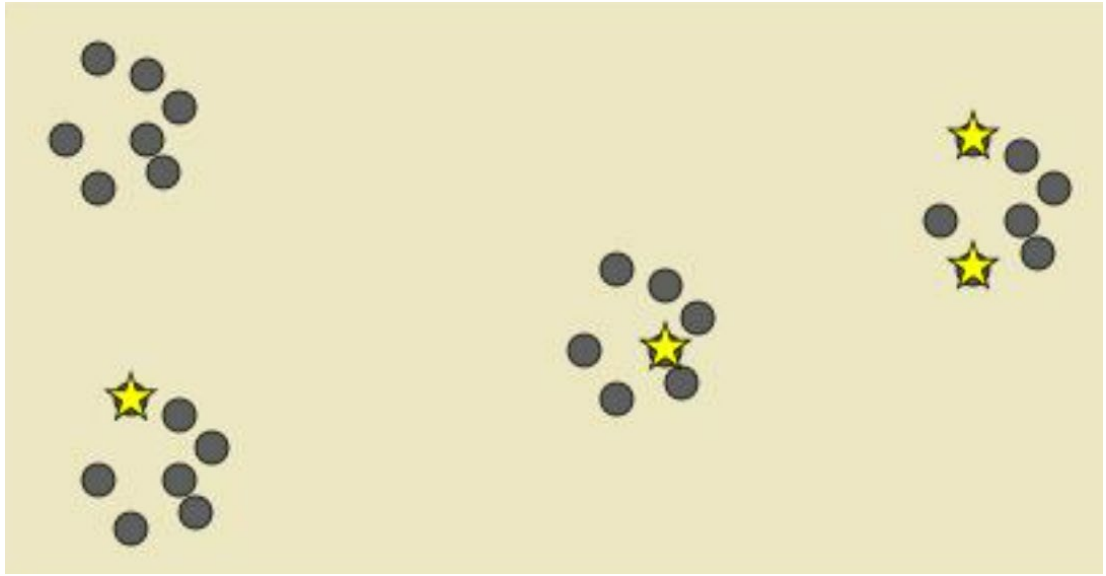
K-means Clustering

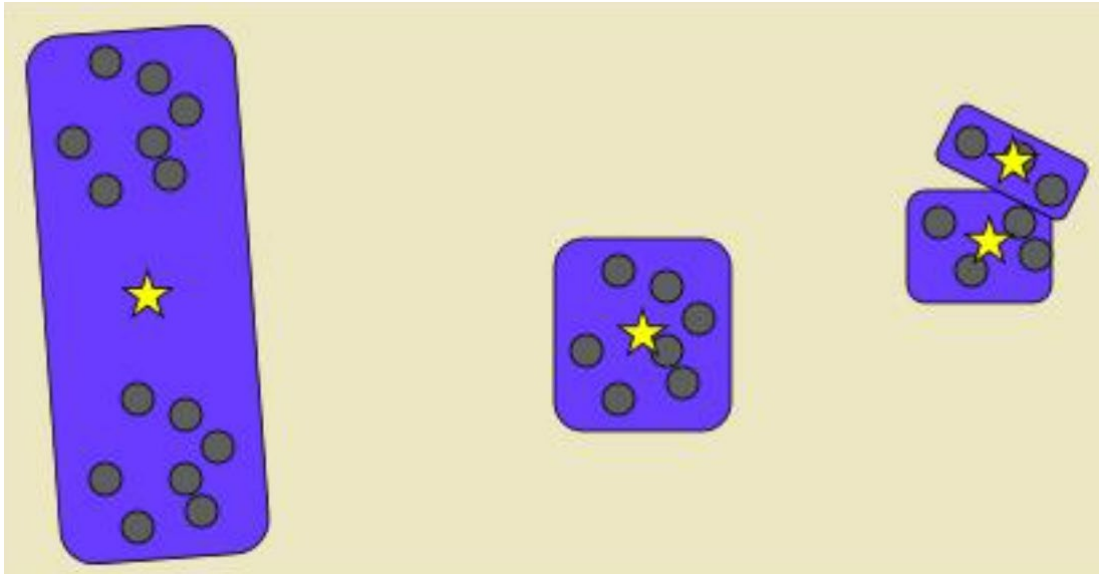




K-means Clustering

- Guaranteed to converge in a finite number of iterations
- A heuristic algorithm with local optimal and the initialization matters



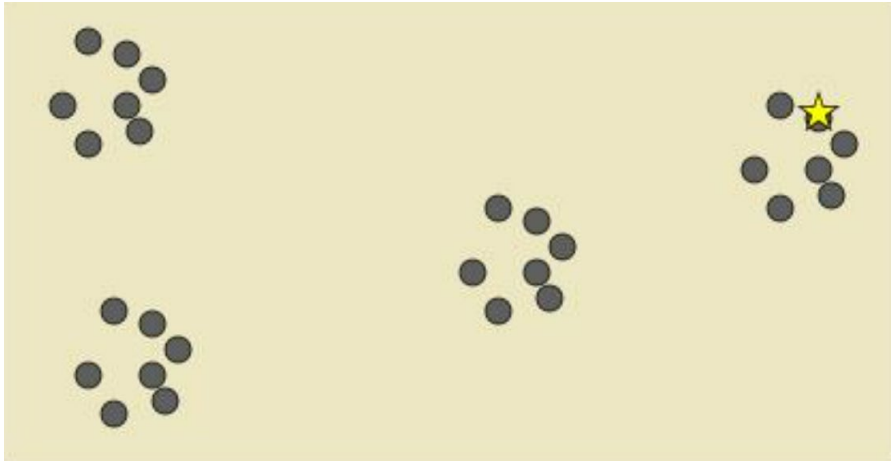


K-means++ (2007 SODA from Stanford)

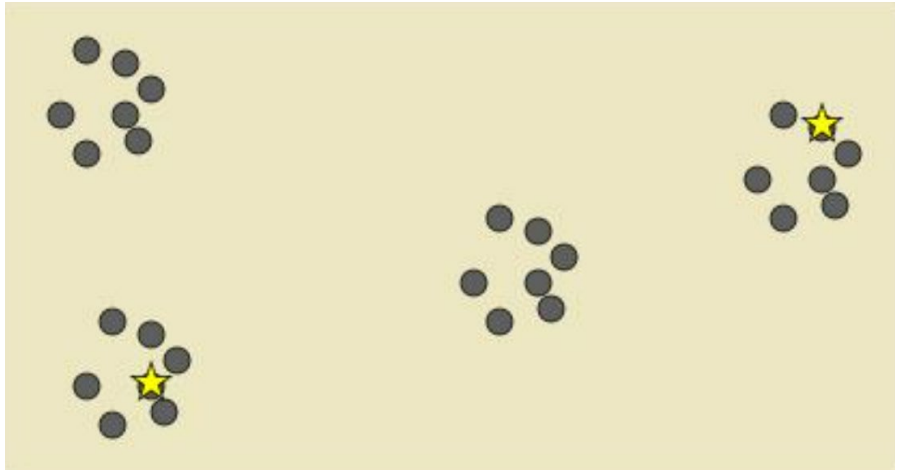
- **A smarter initialization of the centroids**
  - ✓ Choose one center uniformly at random from among the data points
  - ✓ For each data point x, compute D(x), the distance between x and the nearest center that has already been chosen
  - ✓ Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to square of D(x).
  - ✓ Repeat the above two steps until k centers have been chosen
- **The rest of k-means++ algorithm is the same as the standard k-means algorithm**

David Arthur, Sergei Vassilvitskii: k-means++: the advantages of careful seeding. SODA 2007: 1027-1035

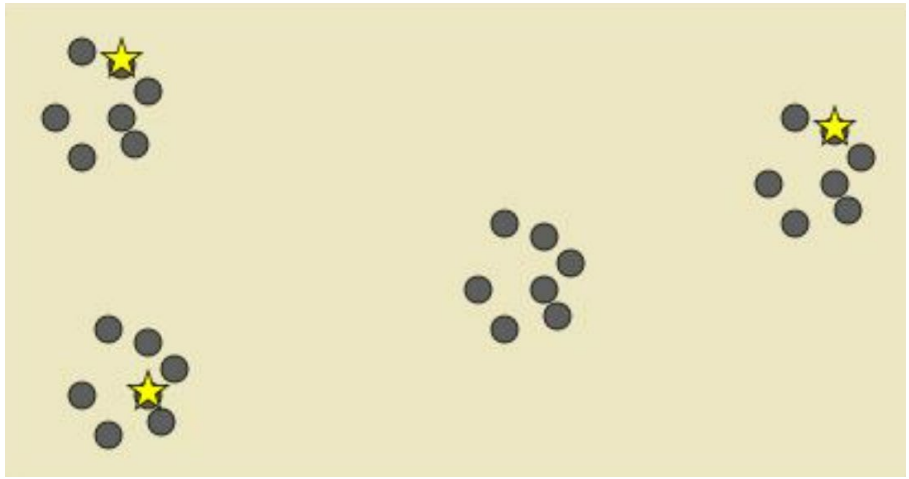
Initialization with k-means++



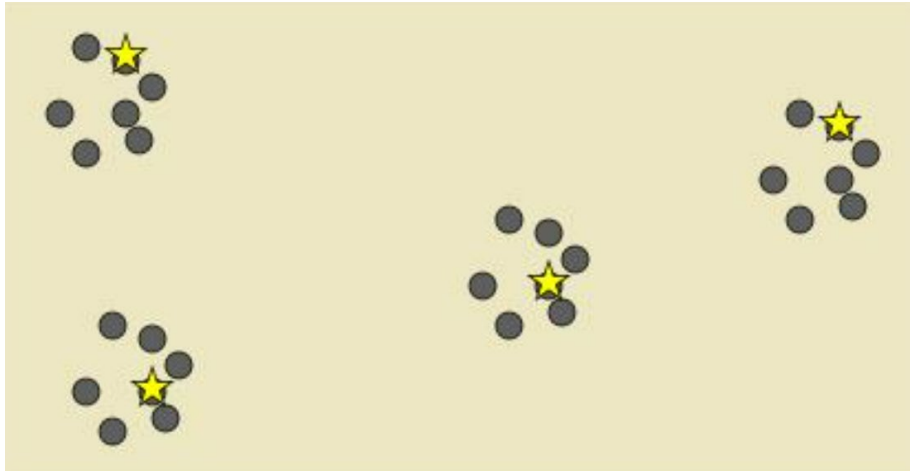
①



②



③



④

Compare k-means++ with k-means

- Experimental results on Intrusion dataset (n=494019, d=35)

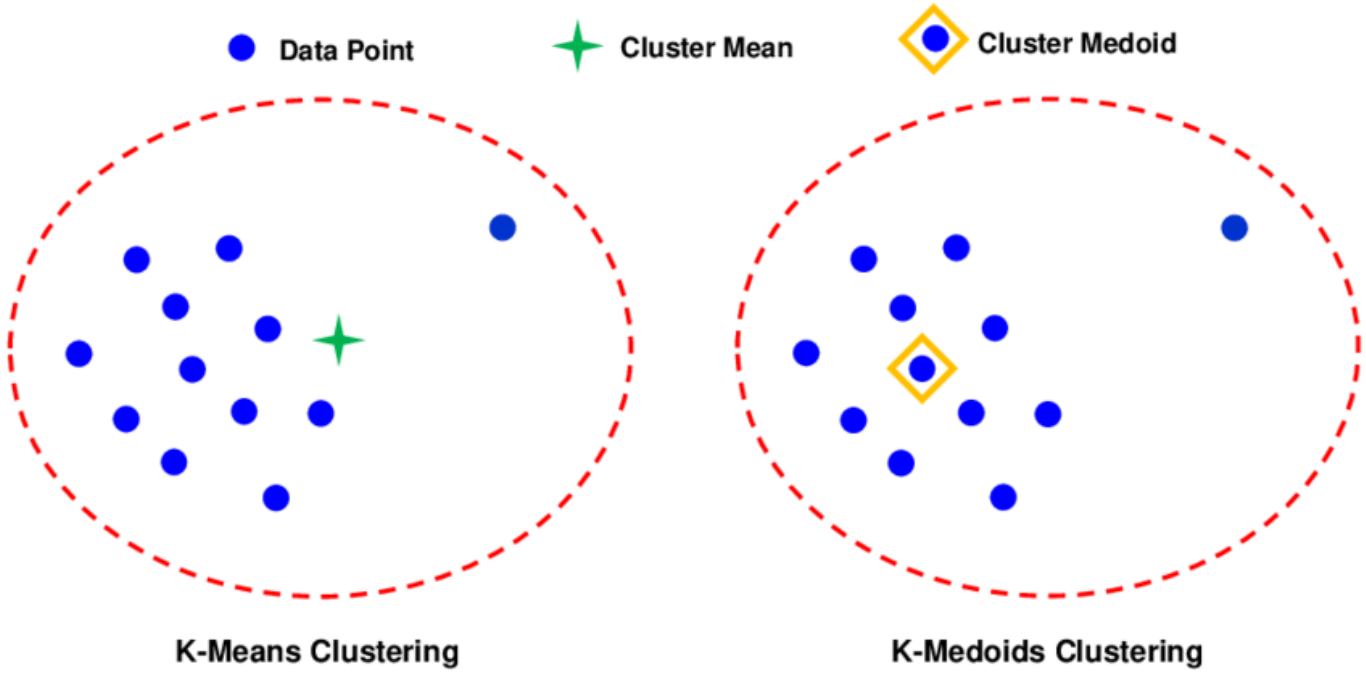
$$\phi = \sum_{x \in \mathcal{X}} \min_{c \in \mathcal{C}} \|x - c\|^2. \qquad 100\% \cdot \left(1 - \frac{\texttt{k-means++ value}}{\texttt{k-means value}}\right)$$

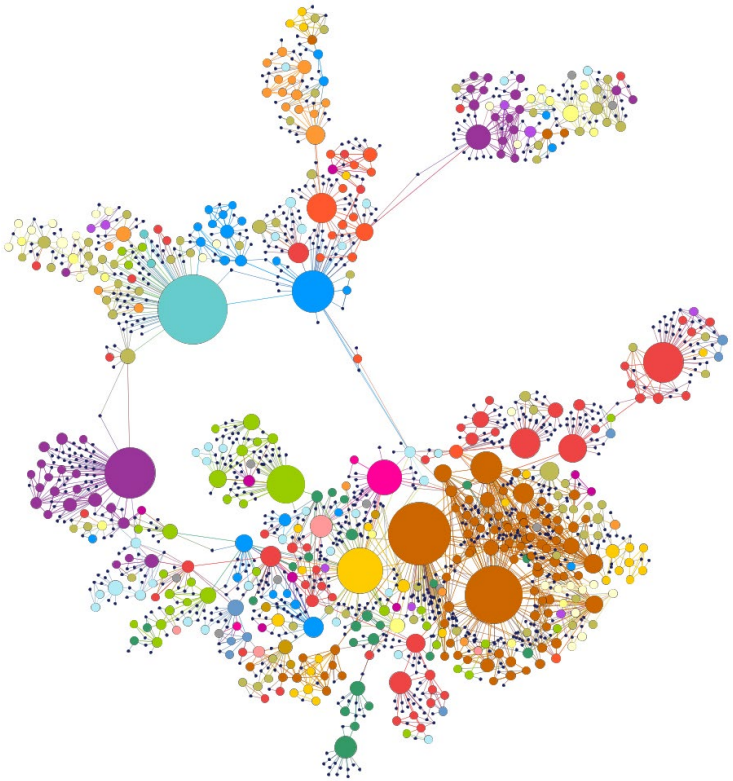
k	Average ϕ		Minimum ϕ		Average T	
	k-means	k-means++	k-means	k-means++	k-means	k-means++
10	$3.387 \cdot 10^8$	93.37%	$3.206 \cdot 10^8$	94.40%	63.94	44.49%
25	$3.149 \cdot 10^8$	99.20%	$3.100 \cdot 10^8$	99.32%	257.34	49.19%
50	$3.079 \cdot 10^8$	99.84%	$3.076 \cdot 10^8$	99.87%	917.00	66.70%

- Better clustering results
- 2-3 times faster

K-medoids Clustering

- Medoid is defined as the point with the minimum sum of distances to other points in the same cluster
- Advantages
  - ✓ The final centroids of k-means are not interpretable
  - ✓ K-means clustering is sensitive to outlier





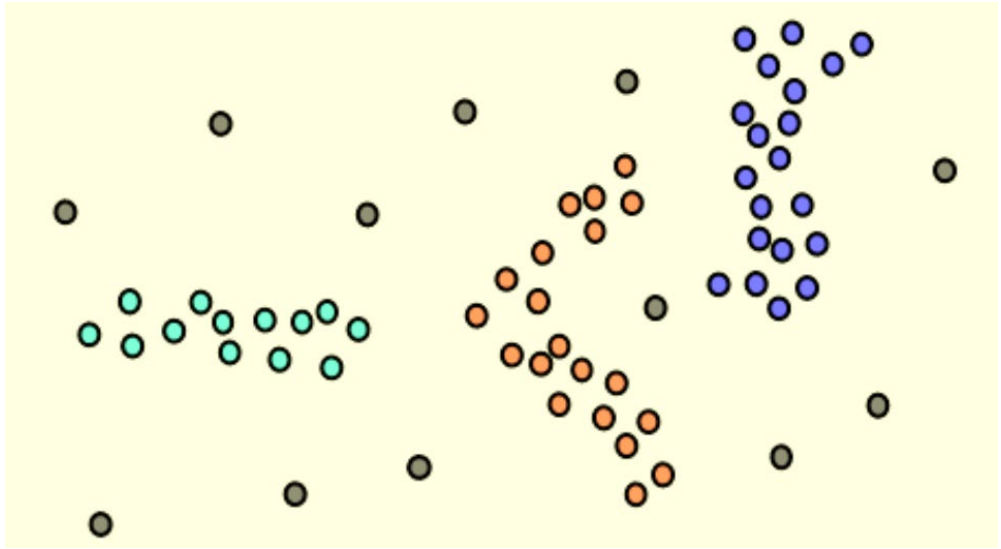
Agenda

- Clustering: Basic Concepts
- K-means Clustering
- Density-Based Clustering
- Hierarchical Clustering
- Cluster Evaluation
- Summary

Density-Based Clustering

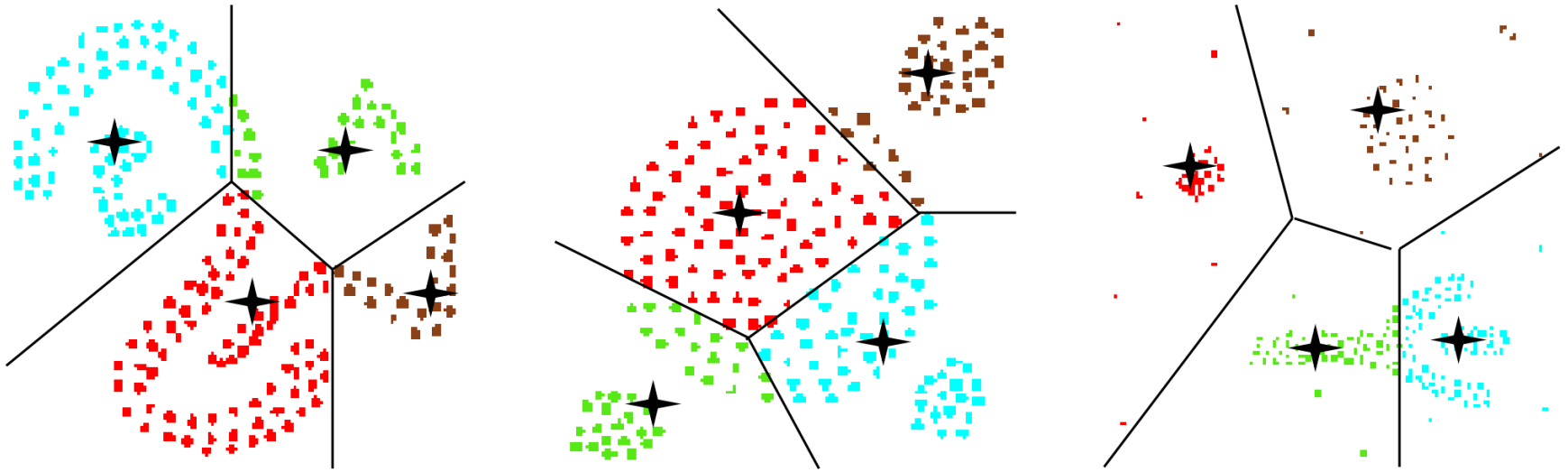
Basic Idea

- ✓ Clusters are generated by their density
- ✓ For any point in a cluster, the local point density around that point has to exceed some threshold



Density-Based Clustering

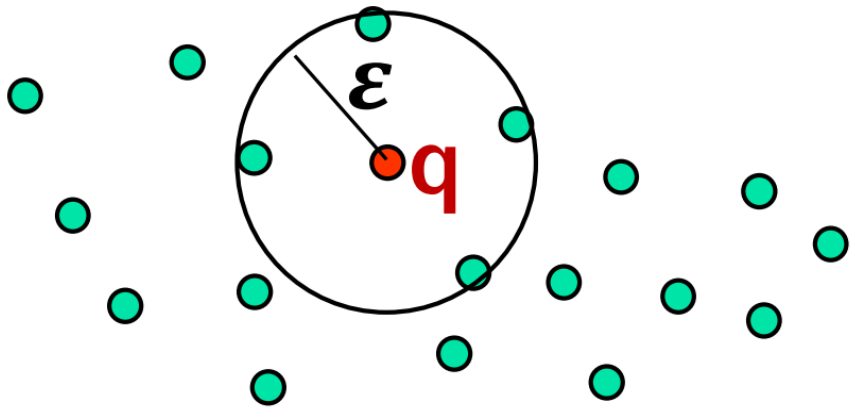
- More robust to noise
- Can handle cases in which k-means or k-medoid fails



Results of a *k*-medoid algorithm for *k*=4

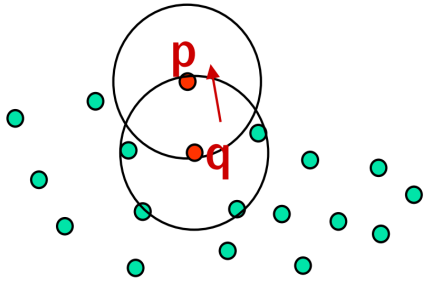
DBSCAN Concepts

- **Density**
  - ✓ Number of points within a specified radius
- **ε-neighborhood**
  - ✓ $N_\varepsilon(q) = \{p \in D | \ dist(p,q) < \varepsilon\}$
- **Core point**
  - ✓ q is called a core point if $| N_\varepsilon(q) | \geq$ MinPts

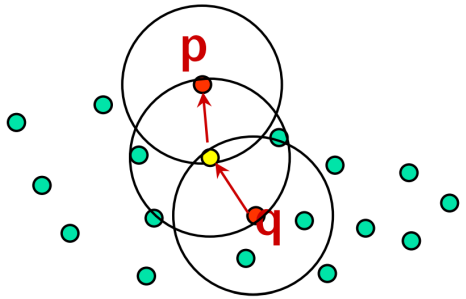


DBSCAN Concepts

- **Directly density-reachable**
  - ✓ Point p is directly density-reachable from q if $p \in N_\varepsilon(q)$ and q is a core point

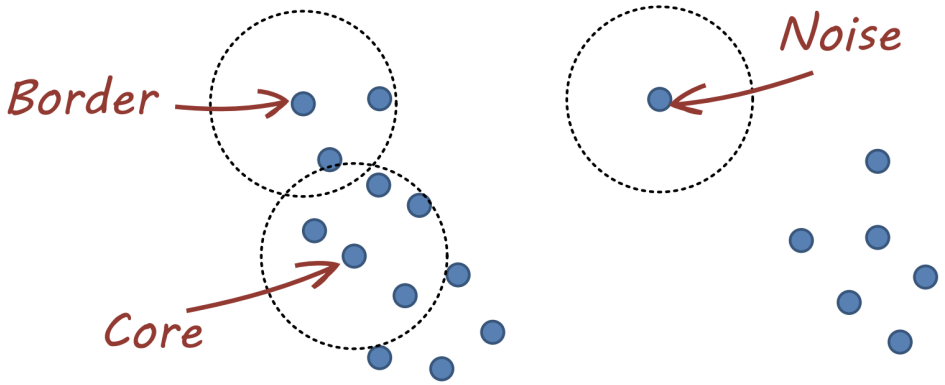


- **Density-reachable**
  - ✓ A point is called density reachable from another point if they are connected through a series of core points.



- **Border point**
  - ✓ $|N_\varepsilon(q)| <$ MinPts and q is density-reachable from a core point

- **Noise**
  - ✓ Remaining points

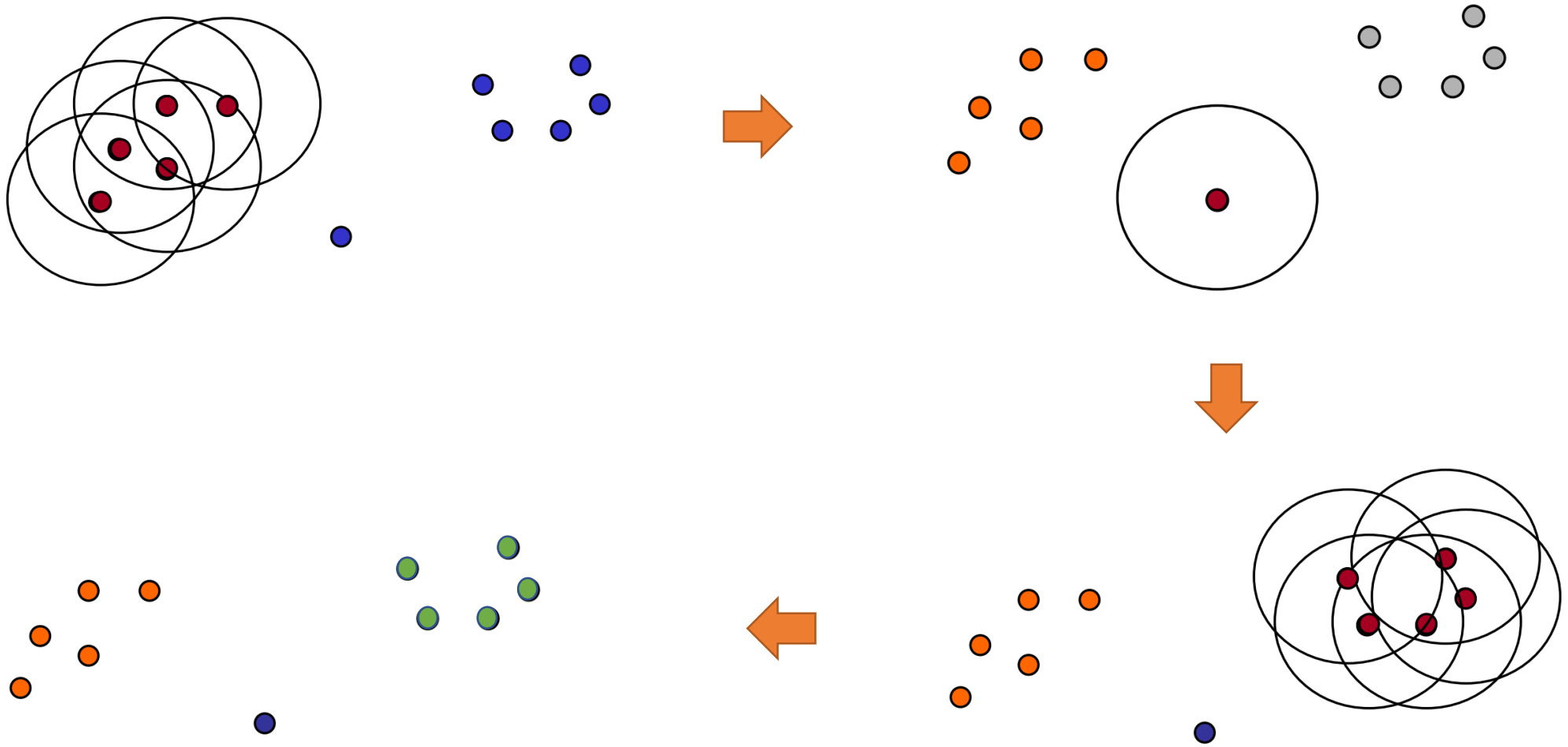


$$\varepsilon = 1.0$$
$$MinPts = 5$$

DBSCAN Algorithm

```
for each o ∈ D do
    if o is not yet classified then
        if o is a core-object then
            collect all objects density-reachable from o
            and assign them to a new cluster.
        else
            assign o to NOISE
```

Example of DBSCAN



Example of DBSCAN

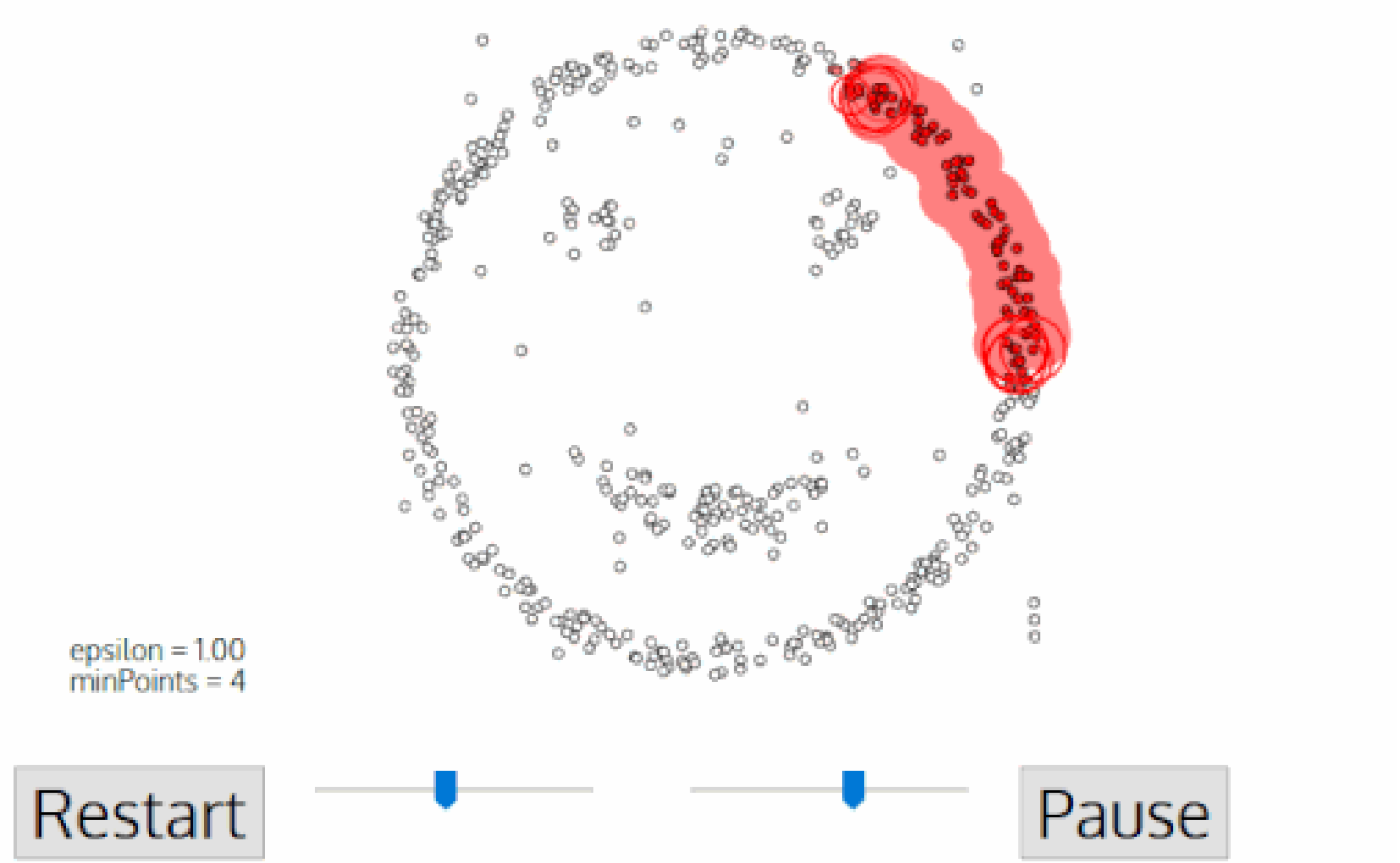


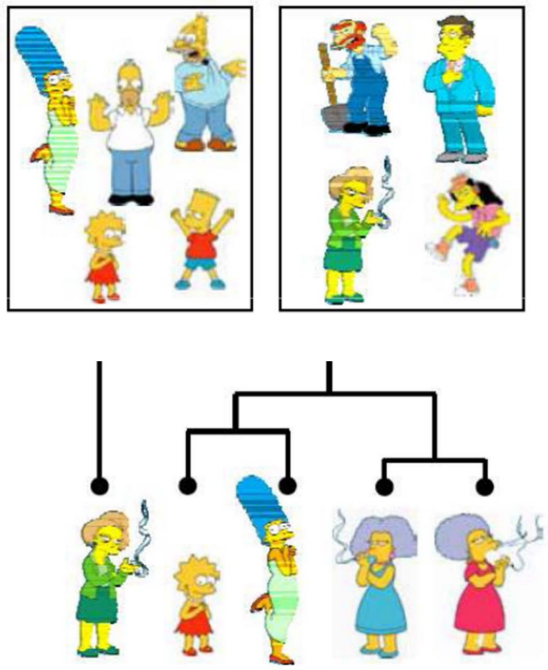
Image source: https://cloud.tencent.com/developer/news/172037

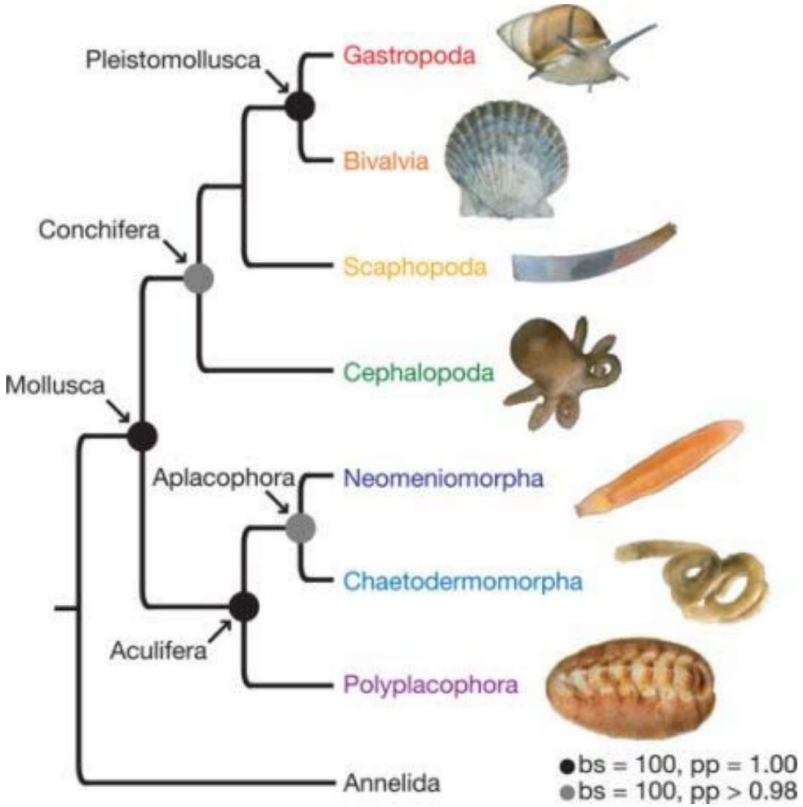
Agenda

- Clustering: Basic Concepts
- K-means Clustering
- Density-Based Clustering
- Hierarchical Clustering
- Cluster Evaluation
- Summary

Hierarchical Clustering

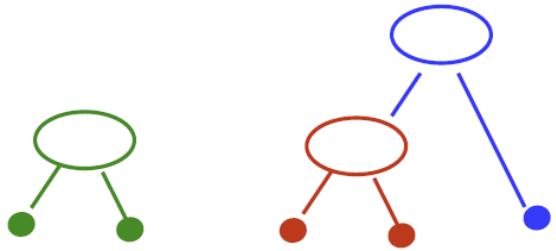
- Produce a set of nested clusters organized as a hierarchical tree
- We do not have to assume any particular number of clusters
- May be useful to discover meaningful taxonomies

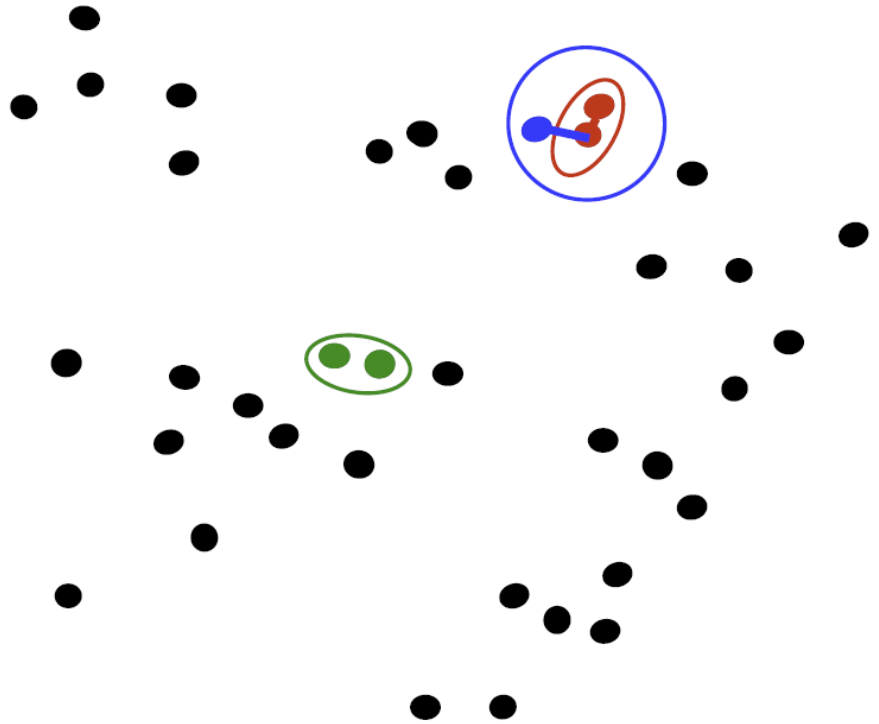




Agglomerative Clustering (Bottom-Up)

- **Basic idea**
  - ✓ First merge very similar instances
  - ✓ Incrementally build larger clusters from the smaller clusters
- **Algorithm**
  - ✓ Initially each data point is treated as a cluster
  - ✓ Repeat
    - Pick the two closest clusters
    - Merge them into a new cluster
  - ✓ Stop when there is only one cluster left

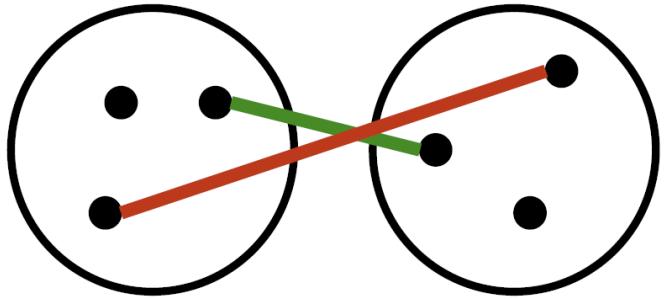




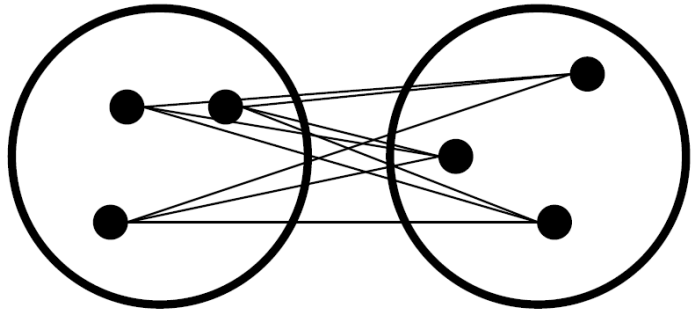
Agglomerative Clustering (Bottom-Up)

- How to define the distance between two clusters?

  - ✓ Closest pair
  - ✓ Farthest pair distance

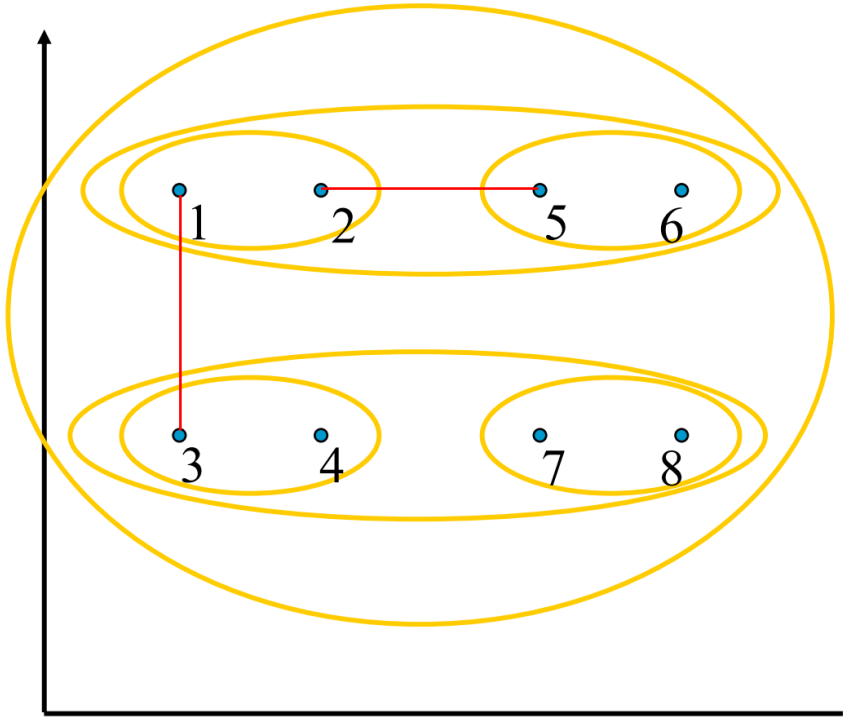


  - ✓ Average pair distance

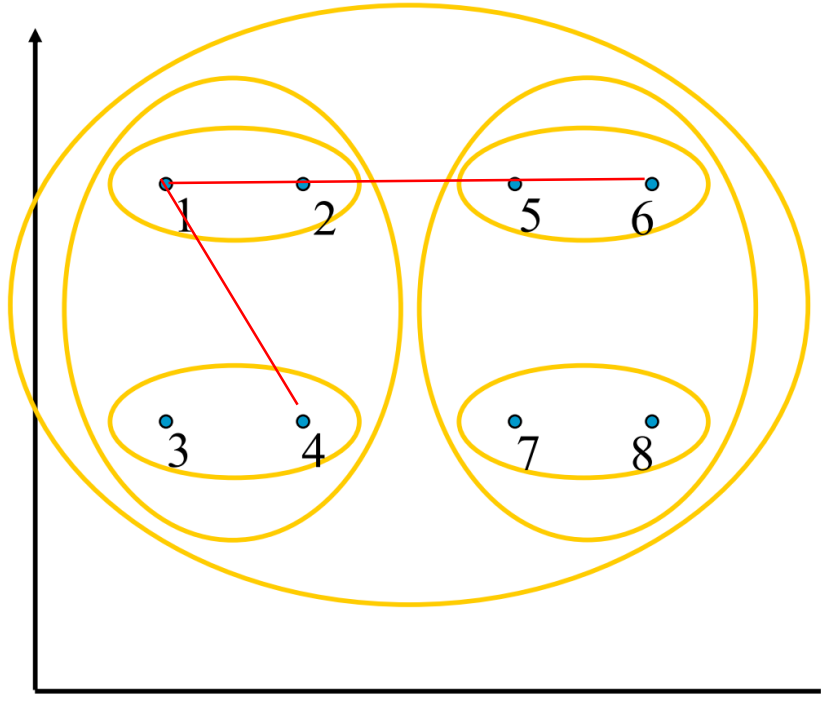


Agglomerative Clustering (Bottom-Up)

- Different choices create different clustering behaviors



Closest Pair



Farthest Pair

Agglomerative Clustering (Bottom-Up)

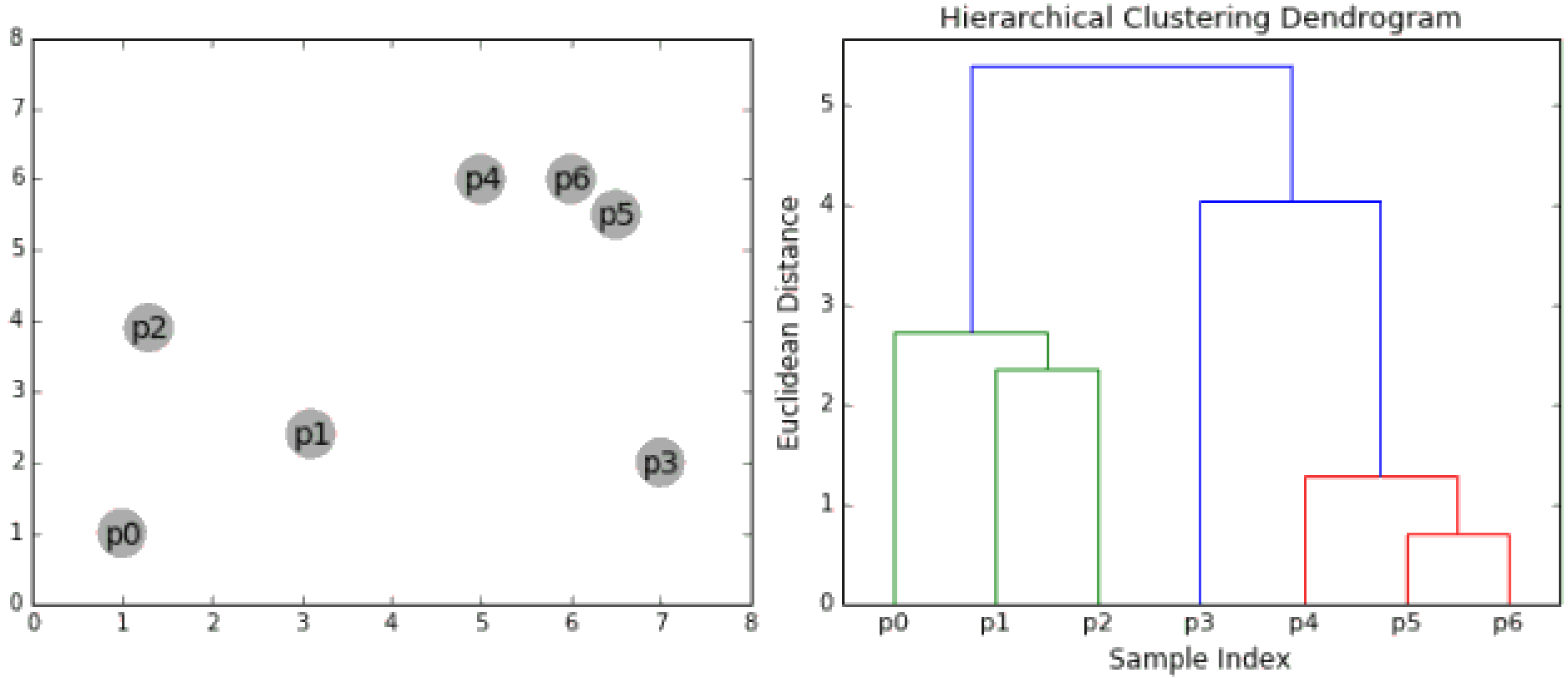
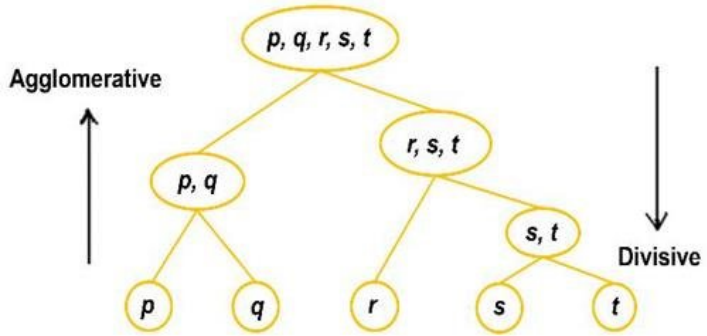


Image source: https://cloud.tencent.com/developer/news/172037

Divisive Clustering (Top-Down)

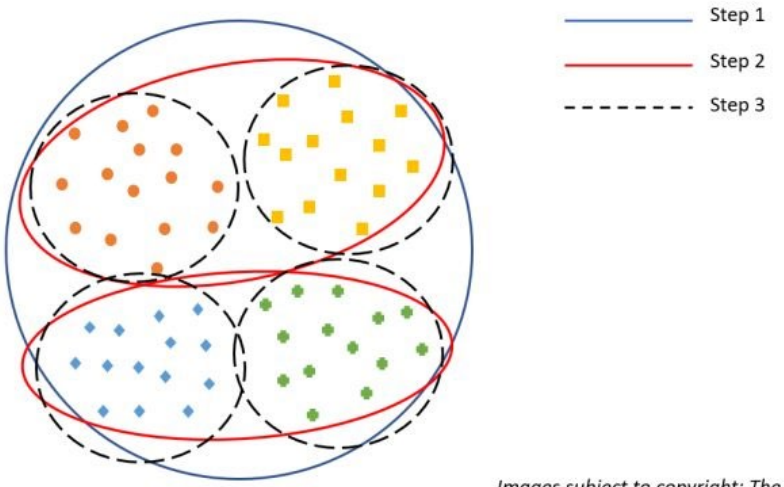
Basic idea

- ✓ Initially, all objects form one cluster
- ✓ Repeat
  - Choose a cluster to split
  - Replace the chosen cluster with sub-clusters
- Stop until all objects are singletons



DIANA Algorithm

- ✓ Select the cluster with largest diameter for splitting
- ✓ Select *q* with the highest average distance to other points
- ✓ Create a new cluster Q={q}
- ✓ Iteratively add p with the highest D(p) to cluster Q
- ✓ D(p)= dist(p, D/Q) - dist(p, Q)
- ✓ Stop when no point with D(p)>0 can be found



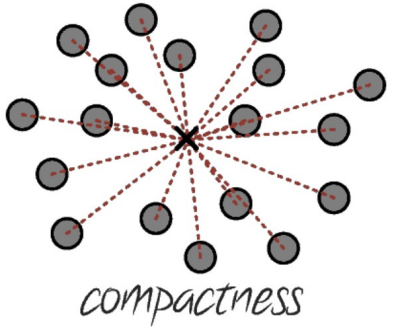
Agenda

- Clustering: Basic Concepts
- K-means Clustering
- Density-Based Clustering
- Hierarchical Clustering
- Cluster Evaluation
- Summary

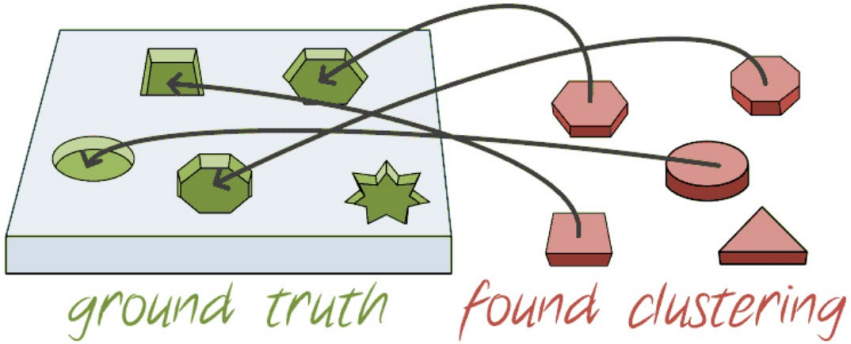
Evaluation of Clustering Results

- **Evaluation based on expert's opinion**
  - ✓ May reveal new insight into the data
  - ✓ Expensive and subjective

- **Evaluation based on internal measures**
  - ✓ No additional information needed
  - ✓ May not be decisive



- **Evaluation based on external measures**
  - ✓ Evaluation is reliable
  - ✓ Needs groundtruth data



Evaluation Based on Internal Measures

- Sum of square distances: $SSD(\mathcal{C}) = \frac{1}{|DB|}\sum_{C_i \in \mathcal{C}} \sum_{p \in C_i} dist\left(p, \mu(C_i)\right)^2$
- Cohesion: measures the similarity of objects within a cluster
- Separation: measures the dissimilarity of one cluster to another one
- Silhouette Coefficient: combines cohesion and separation

Evaluation Based on External Measures

Given clustering $\mathcal{C} = (C_1, \dots, C_k)$ and ground truth $\mathcal{G} = (G_1, \dots, G_l)$ for dataset DB

- Recall: $rec(C_i, G_j) = \frac{|C_i \cap G_j|}{|G_j|}$ Precision: $prec(C_i, G_j) = \frac{|C_i \cap G_j|}{|C_i|}$
- F-Measure: $F(C_i, G_j) = \frac{2 * rec(C_i, G_j) * prec(C_i, G_j)}{rec(C_i, G_j) + prec(C_i, G_j)}$
- Purity (P): $P(\mathcal{C}, \mathcal{G}) = \sum_{C_i \in \mathcal{C}} \frac{|C_i|}{|DB|} pur(C_i, \mathcal{G})$ $pur(C_i, \mathcal{G}) = \max_{G_j \in \mathcal{G}} prec(C_i, G_j)$
- Mutual Entropy: $H(\mathcal{C}|\mathcal{G}) = -\sum_{C_i \in \mathcal{C}} p(C_i) \sum_{G_j \in \mathcal{G}} p(G_j|C_i) log\ p(G_j|C_i)$

$$= -\sum_{C_i \in \mathcal{C}} \frac{|C_i|}{|DB|} \sum_{G_j \in \mathcal{G}} \frac{|C_i \cap G_j|}{|C_i|} * \log_2 \left(\frac{|C_i \cap G_j|}{|C_i|}\right)$$

Agenda

- Clustering: Basic Concepts
- K-means Clustering
- Density-Based Clustering
- Hierarchical Clustering
- Cluster Evaluation
- Summary