Chapter 7: Clustering

Yunming Ye, Baoquan Zhang
School of Computer Science
Harbin Institute of Technology, Shenzhen

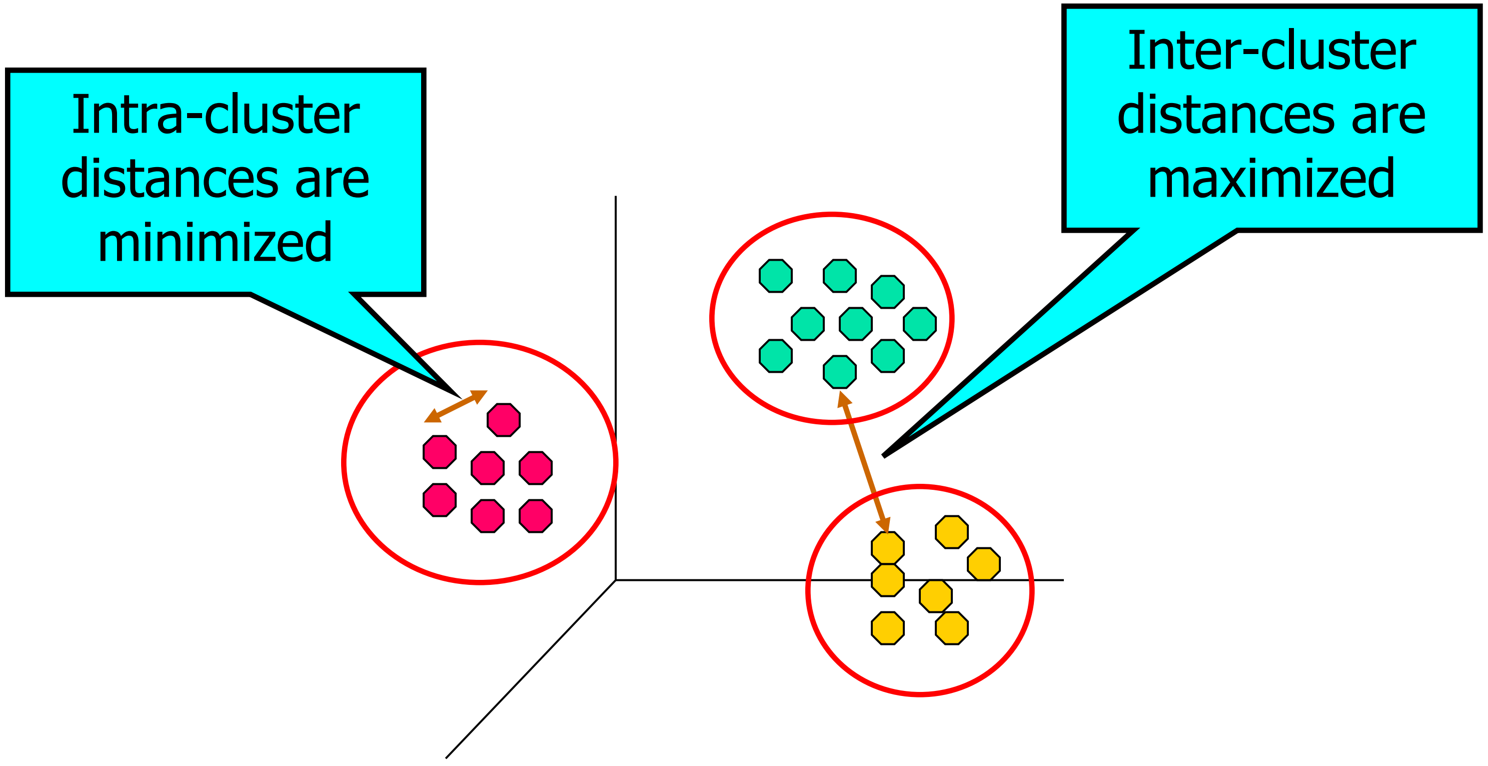
Agenda

- Introduction to Cluster Analysis
- Distance Metrics of Different Data
- Basic Clustering Algorithms
- Clustering with Deep Learning

7.1 Introduction to Cluster Analysis

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



General Applications of Clustering

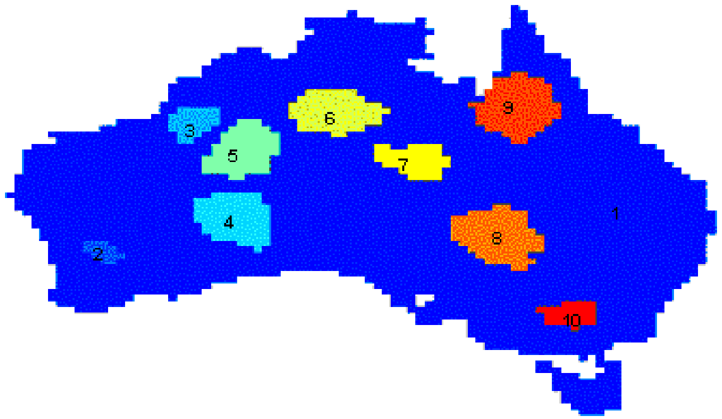
- Business Intelligence
  - Cluster analysis of data
  - Customer segmentation
  - Fraud detection
  - Missing value prediction
- WWW Applications
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns
- Pattern Recognition
- Spatial Data Analysis

Functions of Cluster Analysis

- **Understanding**

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

- **Summarization**
  - Reduce the size of large data sets
- **Preprocessing**
  - A preprocessing step for other data mining algorithms



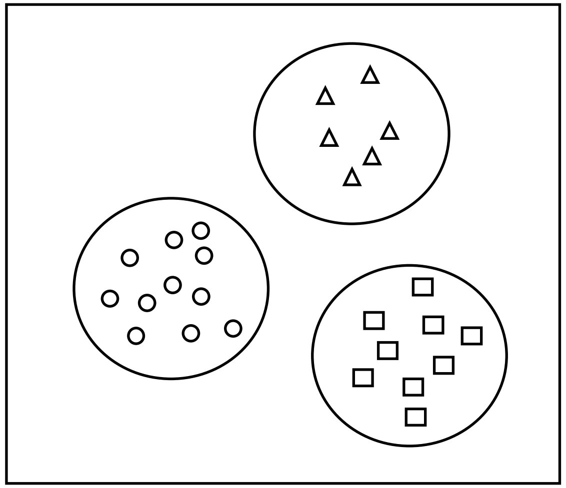
Cluster Definition

- **Cluster Definition**

A cluster is a subset of objects in data which are similar to each other in the cluster according to some similarity measure and dissimilar to other objects outside of the cluster.

Some concepts:

- **Cluster center**
- **Cluster size**
- **Cluster density**
- **Cluster descriptions**



Notion of a Cluster can be Ambiguous

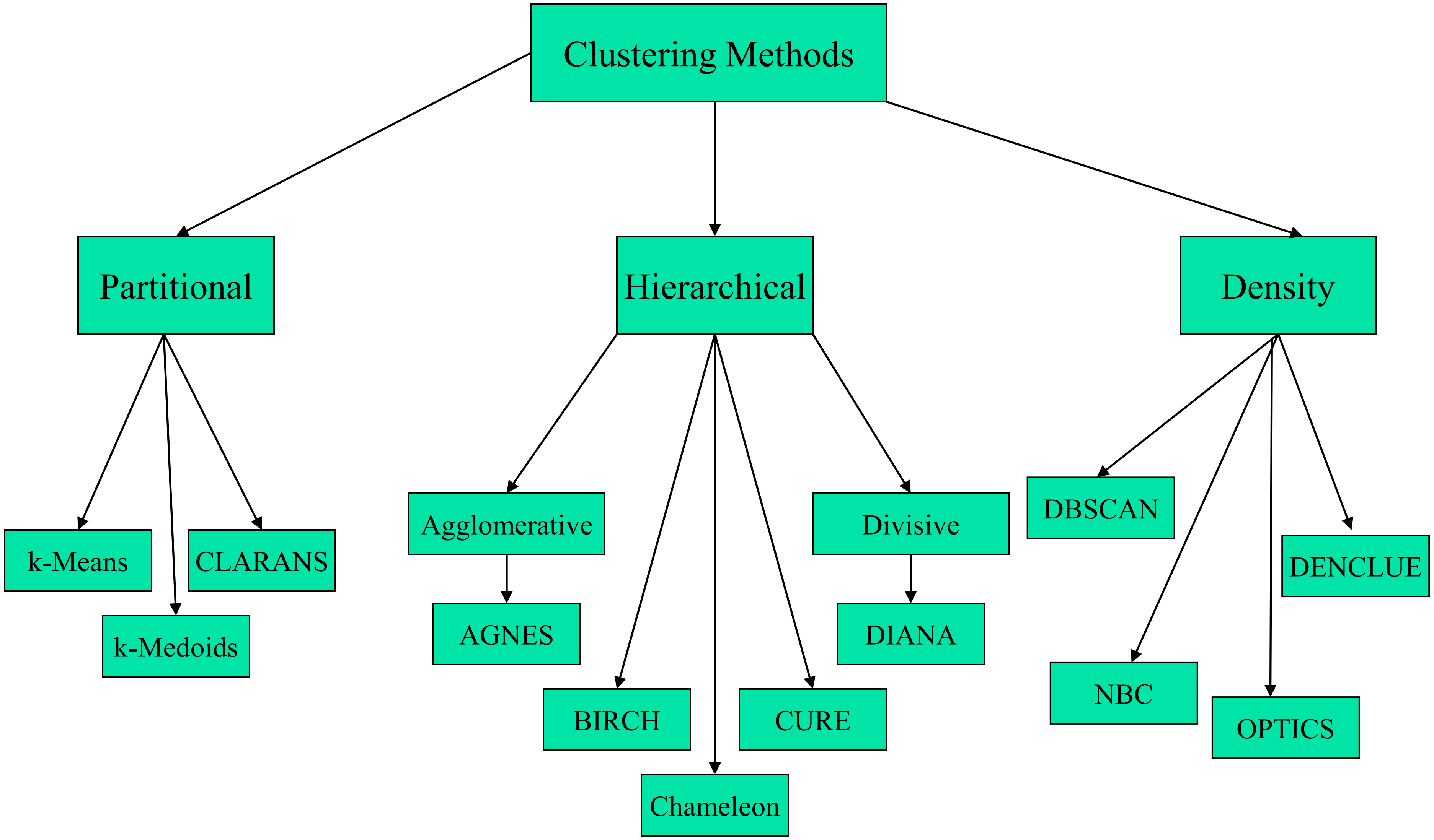
How many clusters?

Six Clusters

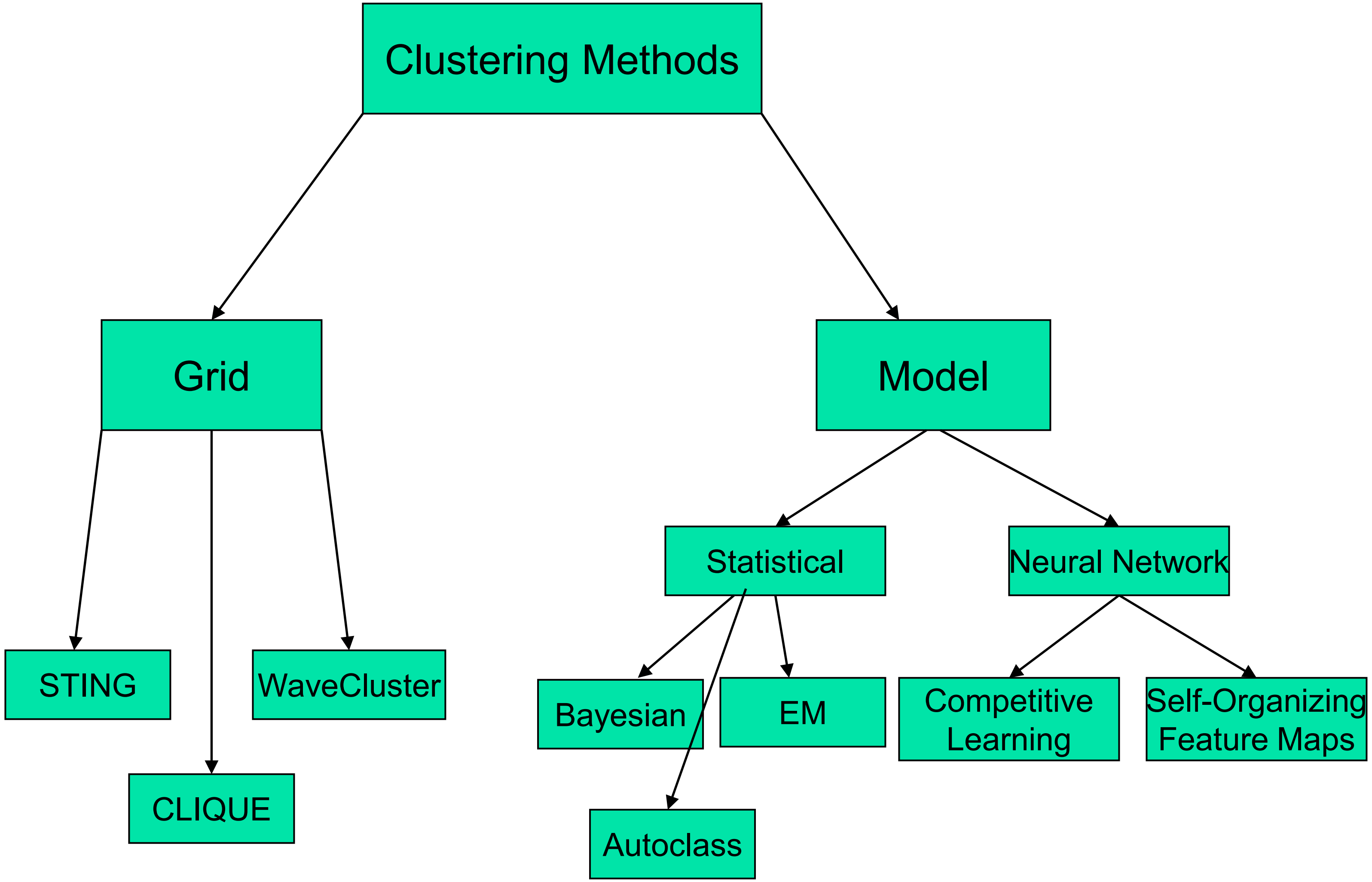
Two Clusters

Four Clusters

Clustering Methods



Clustering Methods



7.2 Distance Metrics of Different Data

Data Structures

- Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Type of Data in Clustering Analysis

- **Interval-scaled variables:**
- **Binary variables:**
- **Nominal, ordinal, and ratio variables:**
- **Variables of mixed types:**

Interval-valued Variables

- Continuous measurements of a roughly linear scale
  - Weight, height, latitude and longitude coordinates, temperature, etc.
- Effect of measurement units in attributes
  - Smaller unit → larger variable range → larger effect to the result
  - Standardization + background knowledge

Similarity and Dissimilarity Between Objects

- Distances are normally used measures
- Minkowski distance: a generalization

$$d(i,j) = \sqrt[q]{|x_{i_1} - x_{j_1}|^q + |x_{i_2} - x_{j_2}|^q + \ldots + |x_{i_p} - x_{j_p}|^q} \quad (q > 0)$$

- If q = 2, d is Euclidean distance
- If q = 1, d is Manhattan distance
- Weighed distance

$$d(i,j) = \sqrt[q]{w_1 |x_{i_1} - x_{j_1}|^q + w_2 |x_{i_2} - x_{j_2}|^q + \ldots + w_p |x_{i_p} - x_{j_p}|^q)} \quad (q > 0)$$

Binary Variables

- A contingency table for binary data

		Object j		
		1	0	Sum
Object i	1	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	p

- Symmetric variable: each state carries the same weight
  - Invariant dissimilarity

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Asymmetric variable: the positive value carries more weight
  - Noninvariant dissimilarity (Jacard)

$$d(i,j) = \frac{r+s}{q+r+s}$$

Noninvariant Dissimilarity between Binary Variables

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Nominal Variables

- A generalization of the binary variable in that it can take more than 2 states, e.g., Red, yellow, blue, green
- Method 1: simple matching
  - m: # of matches, p: total # of variables

$$d(i,j) = \frac{p-m}{p}$$

- Method 2: use a large number of binary variables
  - Creating a new binary variable for each of the M nominal states

Ordinal Variables

- An ordinal variable can be discrete or continuous
- Order is important, e.g., rank
- Can be treated like interval-scaled
  - Replace x_{if} by their rank $r_{if} \in \{1, \ldots, M_f\}$
  - Map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Compute the dissimilarity using methods for interval-scaled variables

Ratio-scaled Variables

- Ratio-scaled variable: a positive measurement on a nonlinear scale
  - E.g., approximately at exponential scale, such as Ae^{Bt}
- Treat them like interval-scaled variables?
  - Not a good choice: the scale can be distorted!
- Apply logarithmic transformation, $y_{if} = \log(x_{if})$
- Treat them as continuous ordinal data, treat their rank as interval-scaled

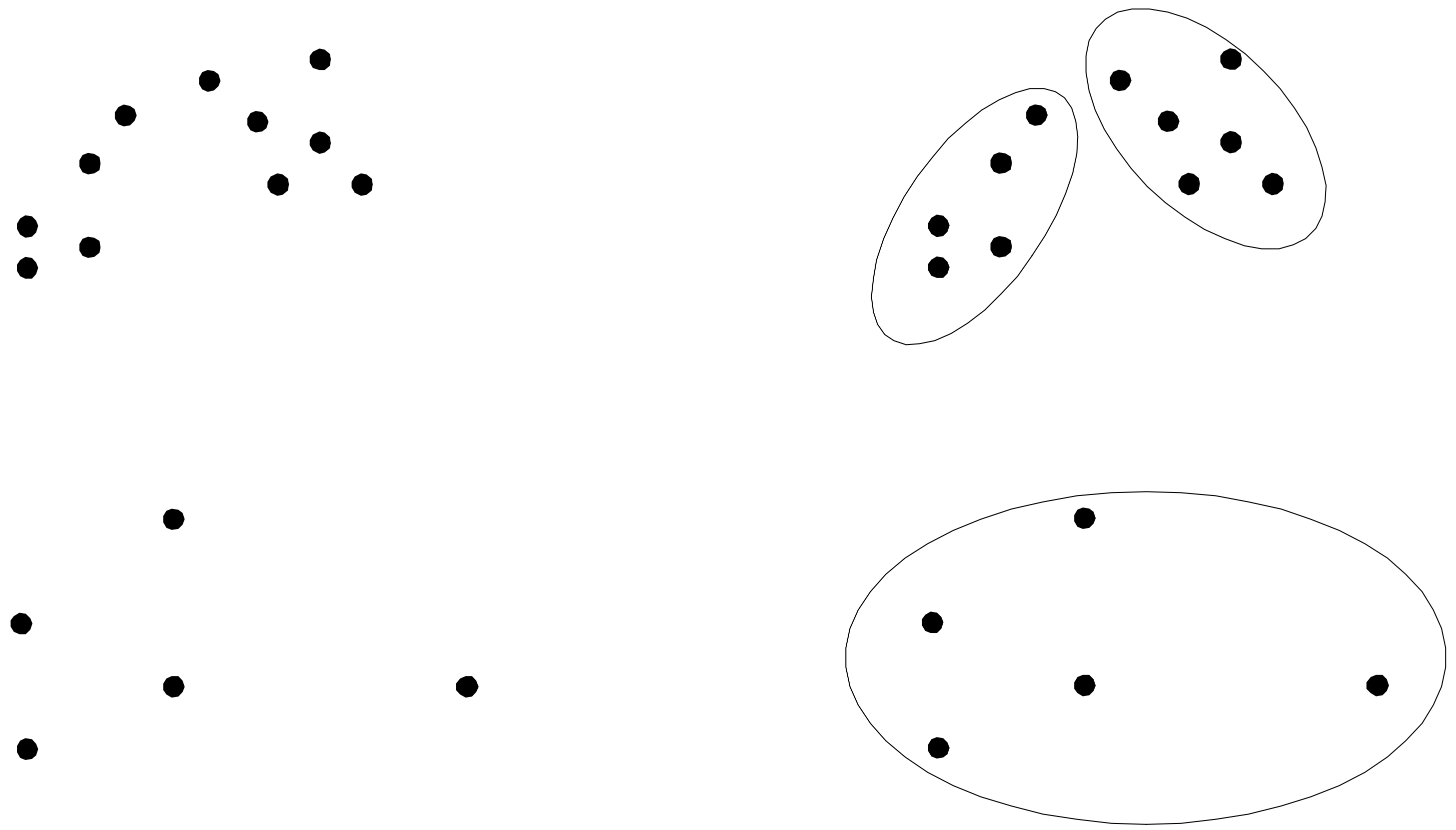
Variables of Mixed Types

- A database may contain all the six types of variables
  - Symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio
- One may use a weighted formula to combine their effects

$$d(i,j) = \frac{\Sigma_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\Sigma_{f=1}^{p} \delta_{ij}^{(f)}}$$

7.3 Basic Clustering Algorithms

Partitional Clustering



Original Points **A Partitional Clustering**

Partitioning Algorithms: Basic Concepts

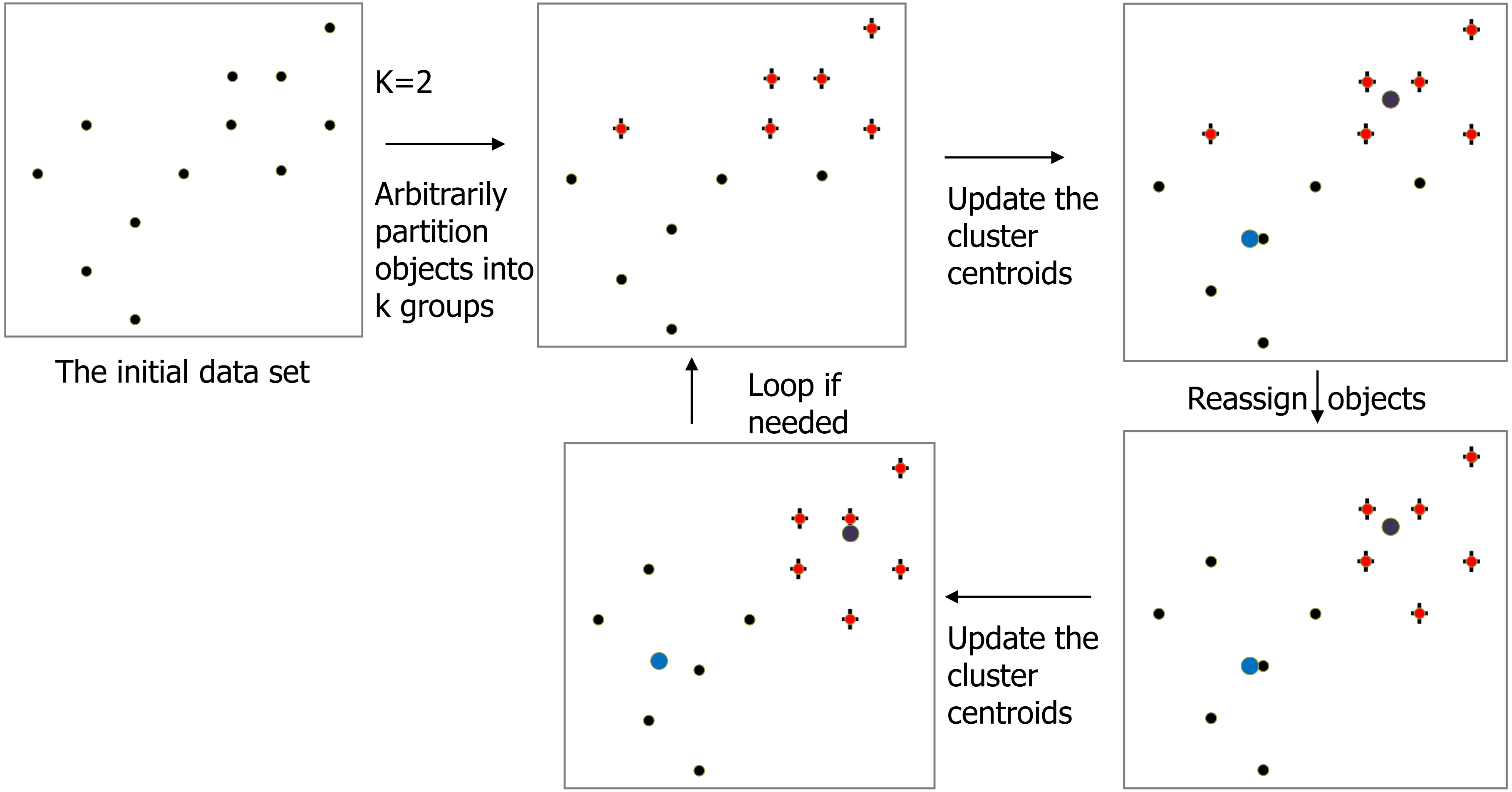
- Partition n objects into k clusters
  - Optimize the chosen partitioning criterion
- Global optimal: examine all possible partitions
  - $(k^n-(k-1)^n-\ldots-1)$ possible partitions, too expensive!
- Heuristic methods: k-means
  - K-means: a cluster is represented by the center

K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change

K-Means: Example



K-means: A Mathematical Programming Problem

- ***Minimize***

$$P(W,Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{i,l}\, d(X_i, Q_l)$$

- ***Subject to***

$$\sum_{l=1}^{k} w_{i,l} = 1$$ ***1≤i≤n***

$$w_{i,l} \in \{0,1\}$$

- ***1≤i≤n, 1≤l≤k***

An Iterative Solution

- **Problem P can be solved by iteratively solving the following two sub problems:**

- **Problem $P1$:**
  - Fix $Q = \hat{Q}$ and solve the reduced problem

$$P(W, \hat{Q})$$

- **Problem $P2$:**
  - Fix $W = \hat{W}$ and solve the reduced problem

$$P(\hat{W}, Q)$$

Sub Problem Solutions

$$Minimize: P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}\, d(X_i, Q_l)$$

- ***Solution to P1:***

1. $w_{i,l} = 1$ *If* $d(X_i, Q_l) \leq d(X_i, Q_t)$ *,* *for* $1 \leq t \leq k$

2. $w_{i,l} = 0$ *for* $t \neq l$

- ***Solution to P2:***

$$q_{l,i} = \frac{\sum_{i=1}^{n} w_{i,l}\, x_{i,j}}{\sum_{i=1}^{n} w_{i,l}}$$ *for 1 ≤ l ≤ k, 1 ≤ j ≤ m*

Derivation of Solution to P2

$$Minimize: P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}\, d(X_i, Q_l)$$

$$\frac{\partial P(W,Q)}{\partial Q_l} = \frac{\partial}{\partial Q_l}\left(\sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}\, d(X_i,Q_l)\right)$$

$$= \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l} \frac{\partial}{\partial Q_l}(Q_l - X_i)^2 = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l}\, 2*(Q_l - X_i) = 0$$

$$\Rightarrow \sum_{X_i \in C_l} 2*(Q_l - X_i) = 0$$

C_l — a cluster l

$$\Rightarrow m_l Q_l = \sum_{X_i \in C_l} X_i \Rightarrow Q_l = \frac{1}{m_l}\sum_{X_i \in C_l} X_i$$

m_l — size of cluster l

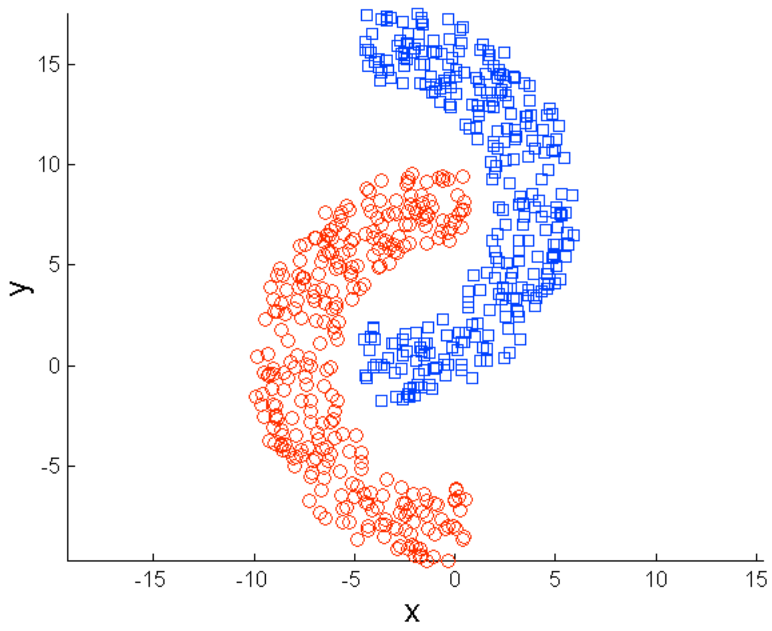
Properties of K-means Algorithm

- *Efficient in clustering large data*
- *Solution depends on initial means*
- *Sensitive to outliers*
- *Spherical clusters*
- *Numeric data*

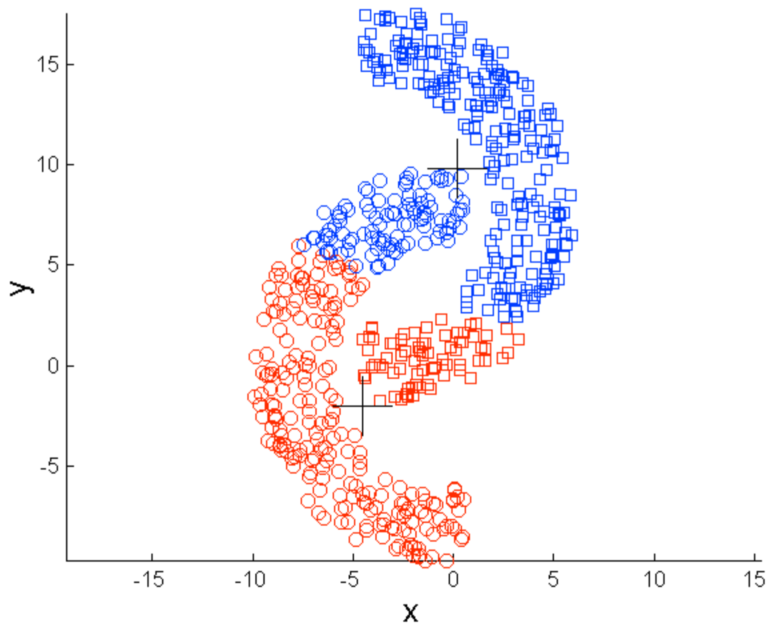
Comments on the *K-Means* Method

- Strength
  - *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, k, $t << n$.
  - Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*
- Weakness
  - Sensitive to initial centroids
  - Need to specify k, the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*
  - Not suitable to discover clusters with *non-convex shapes*

Limitations of K-means: Non-globular Shapes

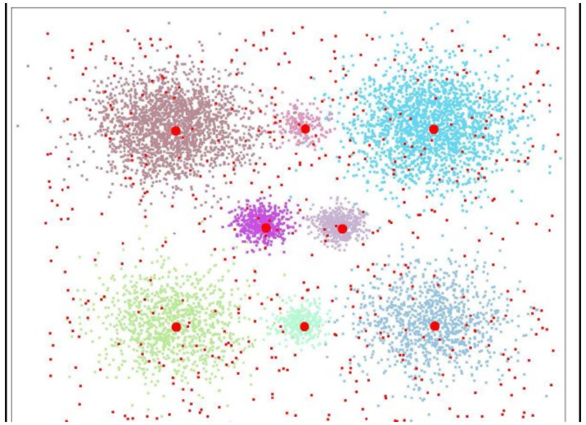


Original Points



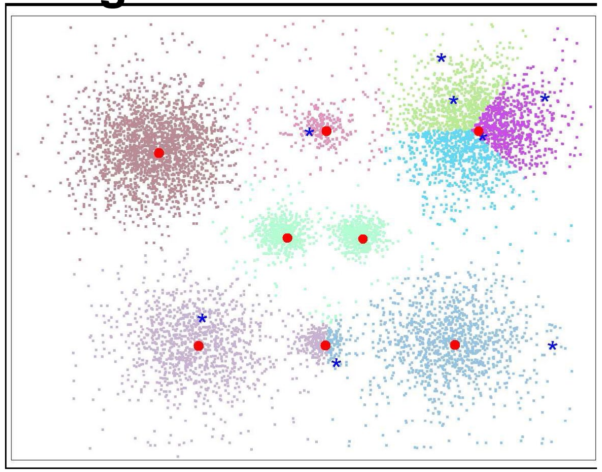
K-means (2 Clusters)

Problems of Selecting Initial Centers in *k-means* Clustering

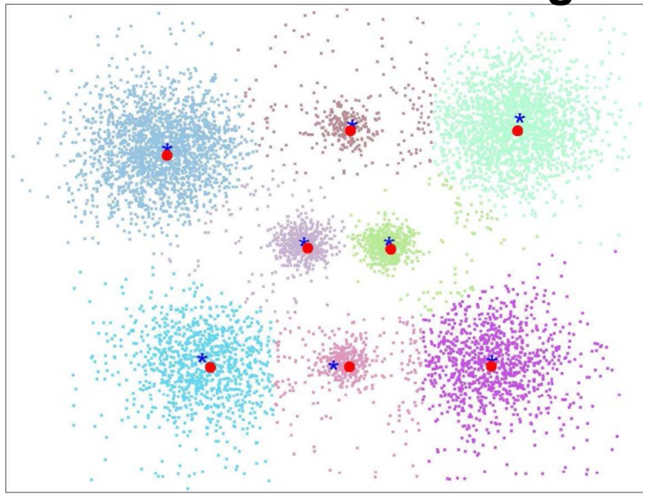


Original Data Set

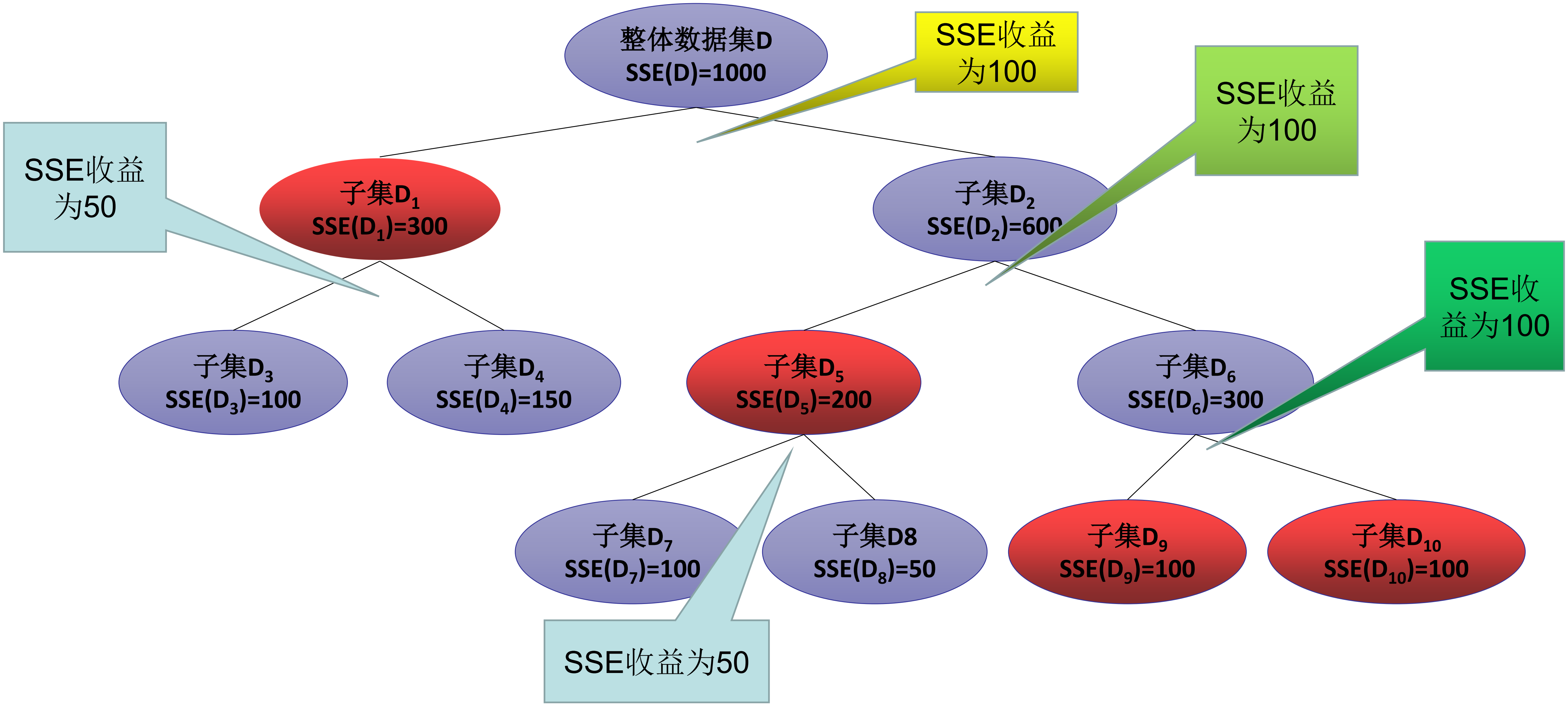
Clustering Result 1



Clustering Result 2

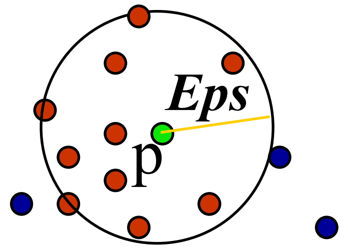


Bisecting K-means



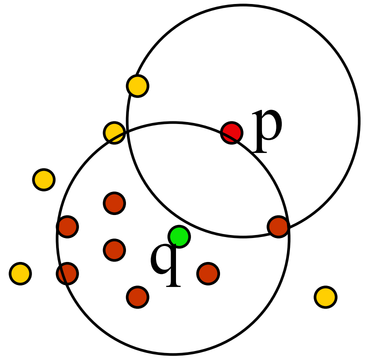
Density-Based Clustering: Definitions

- **Two parameters:**
  - *Eps*: Maximum radius of the neighborhood
  - *MinPts*: Minimum number of points in an Eps-neighborhood of that point
- $N_{Eps}(p)$: $\{q$ *belongs to D* $\mid dist(p,q) <= Eps\}$



Density-Based Clustering: Definitions

- **<u>Directly density-reachable</u>: A point *p* is directly density-reachable from a point *q* wrt. *Eps*, *MinPts* if**
  - 1) *p* belongs to $N_{Eps}(q)$
  - 2) core point condition: $|N_{Eps}(q)| >= MinPts$

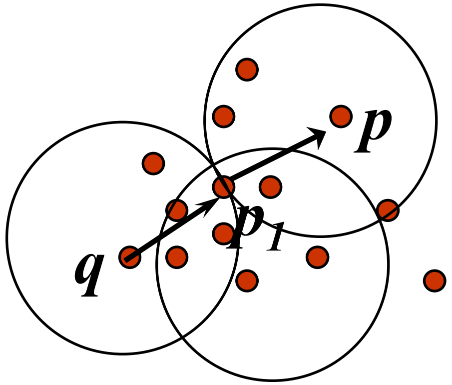


MinPts = 5

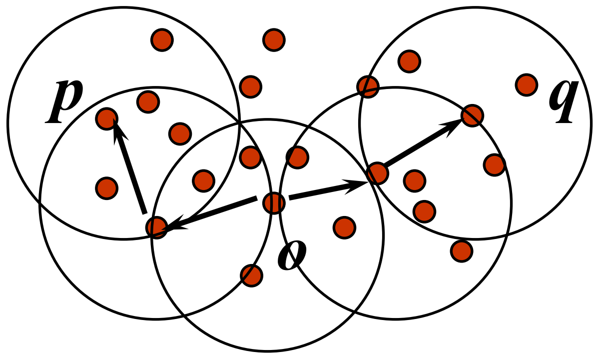
Eps = 1 cm

Density-Based Clustering: Definitions

- **<u>Density-reachable</u>:**
  - A point *p* is density-reachable from a point *q* wrt. *Eps*, *MinPts* if there is a chain of points $p_1, \ldots, p_n$, $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i

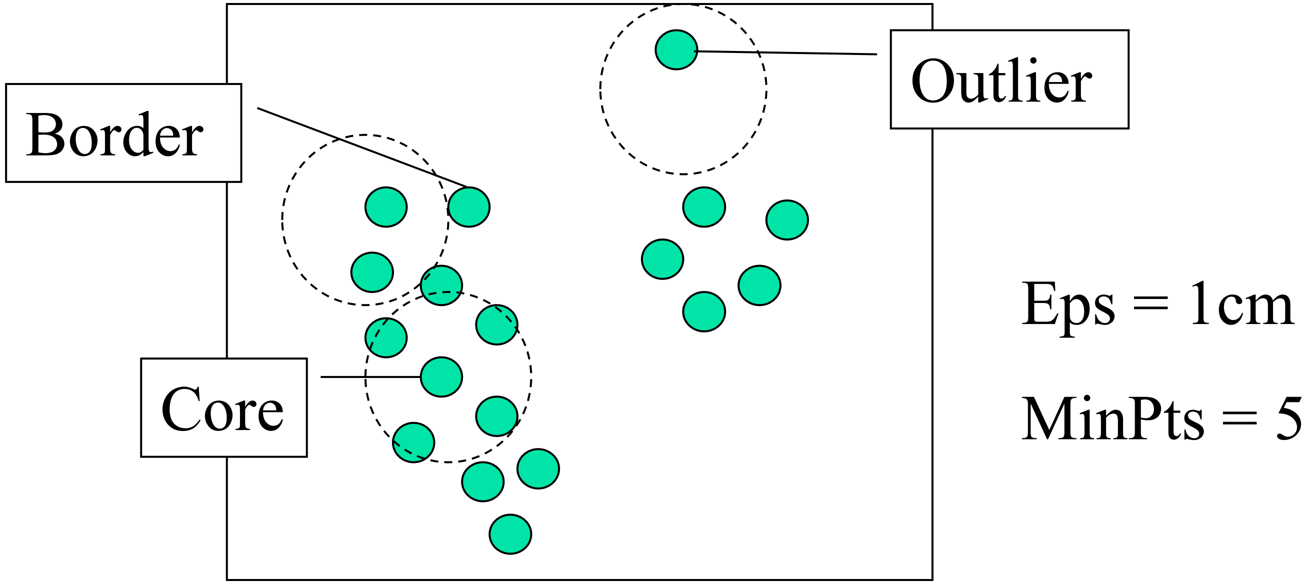


- **<u>Density-connected:</u>**
  - A point *p* is density-connected to a point *q* wrt. *Eps*, *MinPts* if there is a point *o* such that both, *p* and *q* are density-reachable from *o* wrt. *Eps* and *MinPts*.



Density Based Cluster: Definition

- **Relies on a *density-based* notion of cluster:**
  - **A *cluster* is defined as a maximal set of density-connected points**
- **A <u>cluster</u> C is a subset of D satisfying**
  - For all p,q if p is in C, and q is density reachable from p, then q is also in C
  - For all p,q in C: p is density connected to q



Density Based Cluster: Definition

- **<u>Lemma 1</u>:** If *p* is a core point, and *O* is the set of points density reachable from *p*, then *O* is a cluster
- **<u>Lemma 2</u>:** Let *C* be a cluster and *p* be any core point of *C*, then *C* equals the set of density reachable points from p

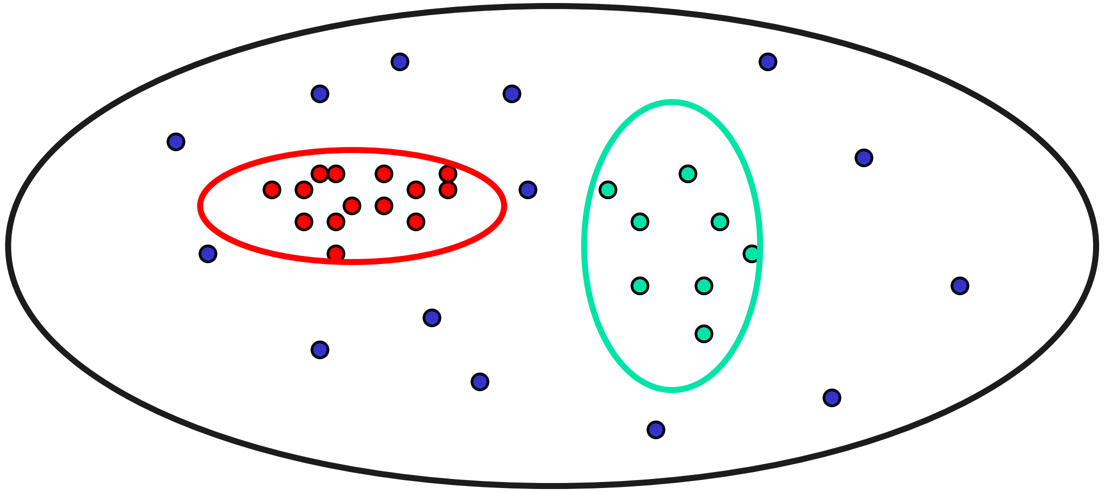
- **Implication:** Finding density reachable point of an arbitrary point generates a cluster. A cluster is unique determined by *any* of its core points

DBSCAN Algorithm

- Arbitrary select a point *p*
- Retrieve all points density-reachable from *p* wrt *Eps* and *MinPts*.
- If *p* is a core point, a cluster is formed.
- If *p* is a border point, no points are density-reachable from *p* and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

Problems of DBSCAN

- Different clusters may have very different densities

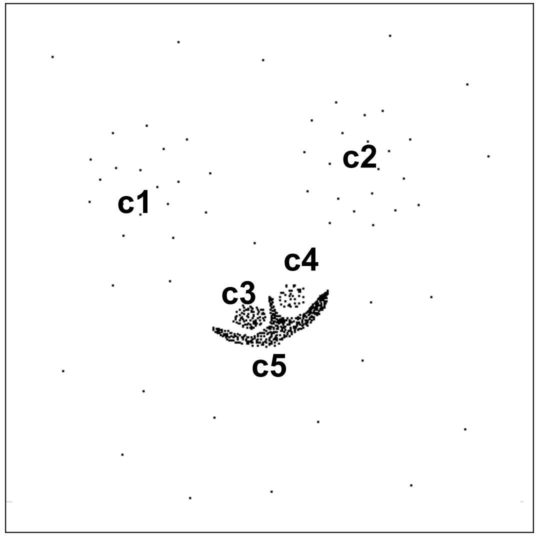


Neighborhood-Based Clustering (NBC)

Density-based clustering algorithms

- **DBSCAN: Not very effective to discover clusters of different local-densities and multi-granularities**

- **Neighborhood-Based Clustering (NBC)**
  - Automatically discover clusters of arbitrary distributions



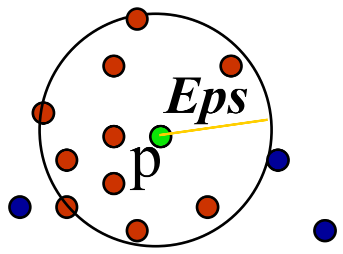
e.g., in this dataset,

DBSCAN puts clusters C3, C4, C5 into one cluster

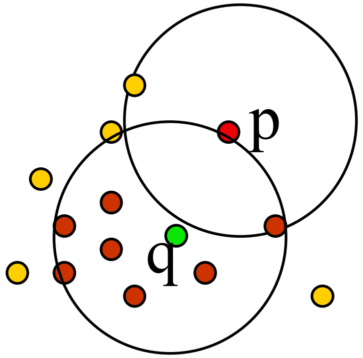
NBC discovers all of the five clusters

Basic Concepts

- ***K Neighborhood:***



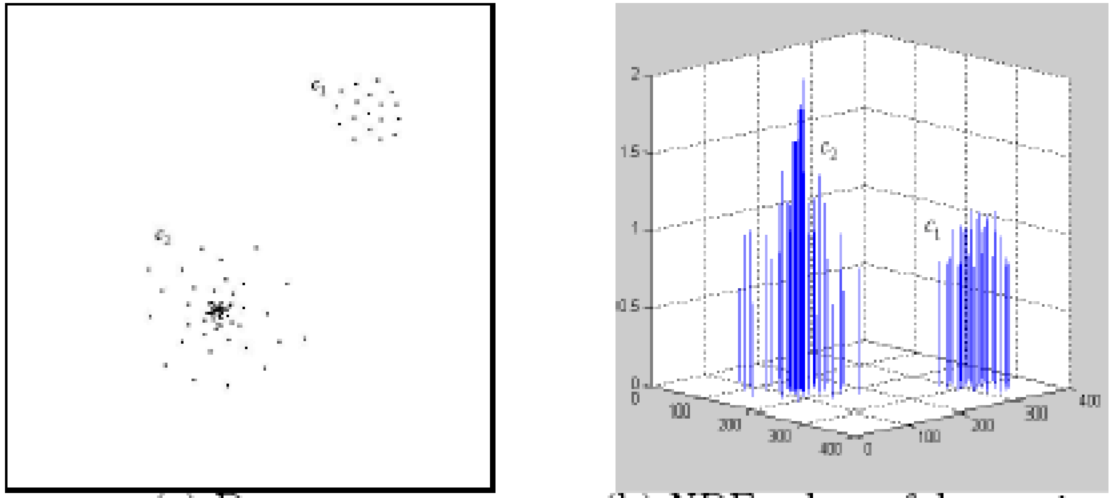
- Reverse K Neighborhood:



Basic Concepts

- **Neighborhood-based Density Factor (NDF)**

$$NDF(p) = \frac{|R-kNB(p)|}{|kNB(p)|}$$



(a) Dataset (b) NDF values of data points

Fig. 1. An illustration of NDF

Define local density for one point

Basic Concepts

- **Local Dense Point (DP)**
  - NDF(p) > 1
- **Local Sparse Point (SP)**
  - NDF(p) < 1
- **Local Even Point (EP)**
  - NDF(p) is equal (or approximately equal) to 1

Basic Concepts

- **Directly Neighborhood-based density reachable (directly ND-reachable)**

$$p \text{ is directly ND-reachable from } q \text{ iff}$$
$$\text{(a) } q \text{ is a DP or EP, and}$$
$$\text{(b) } p \in kNB(q)$$

Basic Concepts

- **ND-reachable**

$$p \text{ is ND-reachable from } q, \text{ iff}$$
$$\text{there is a chain of objects } p_i, \ldots, p_n, p_1 = p, p_n = q,$$
$$p_i \text{ is directly ND-reachable from } p_{i+1}$$

- **ND-connected**

$$p \text{ and } q \text{ are ND-connected if}$$
$$p \text{ and } q \text{ are both ND-reachable from a third object } o$$

Basic Concepts

- Neighborhood-based cluster

Given a dataset D,

a cluster C is a non-empty subset of D such that

(a) for two objects p and q in C, p and q are ND-connected

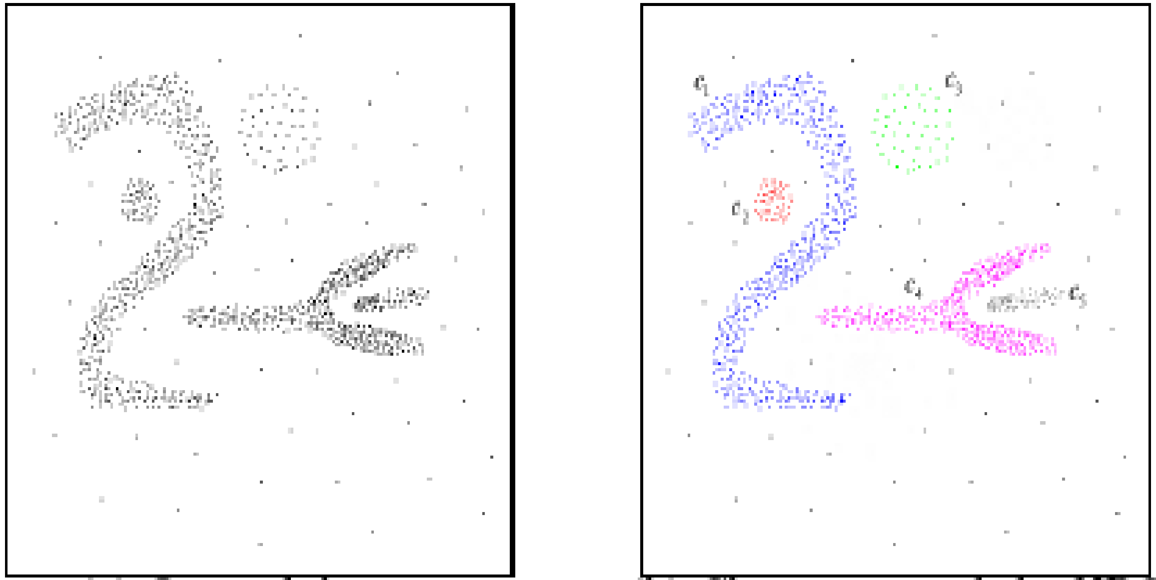
(b) if $p \in C$ and q is ND-connected from p, then $q \in C$

The NBC Algorithm

- **Evaluating NDF values**
  - Using VA-file to support high-dimensional access
  - Search *k*NB and R-*k*NB for each object
  - Calculate NDF
- **Clustering the dataset**
  - Fetch a new DP or EP
  - Create a new cluster
  - Extend the cluster (find all ND-connected objects

Performance Evaluation

- **Discover clusters of arbitrary shapes**

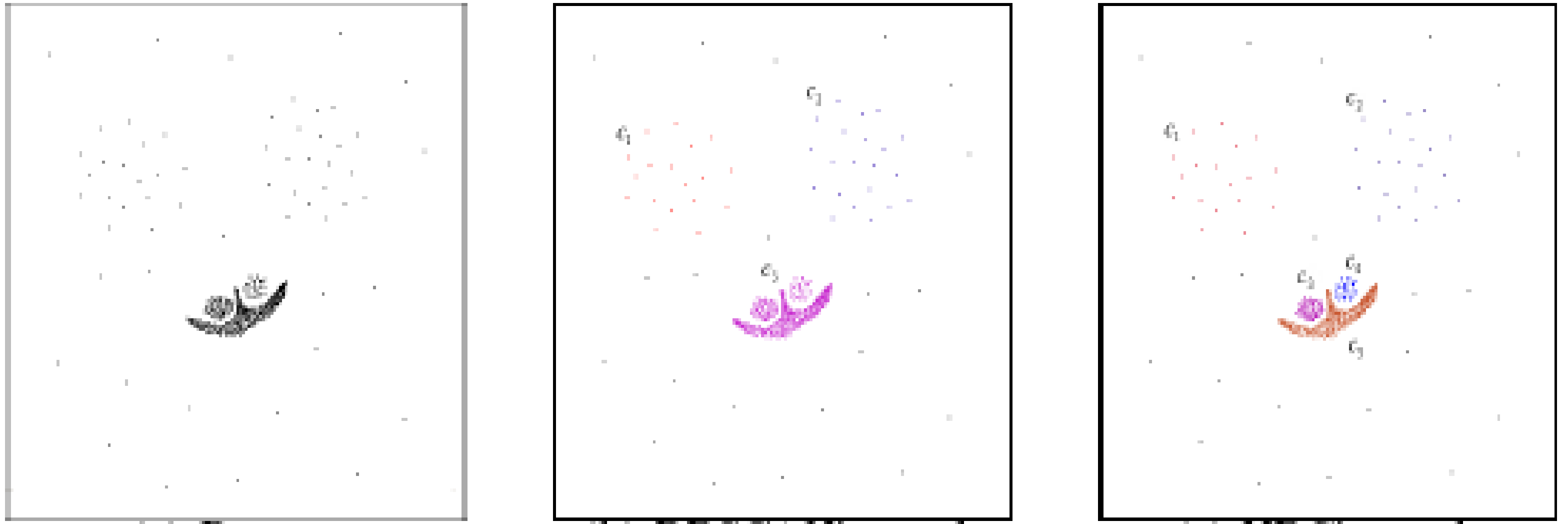


(a) Original dataset (b) Clustering result by NBC

Fig. 3. Discoverying clusters of arbitrary shape

Performance Evaluation

- **Discover clusters of different densities**

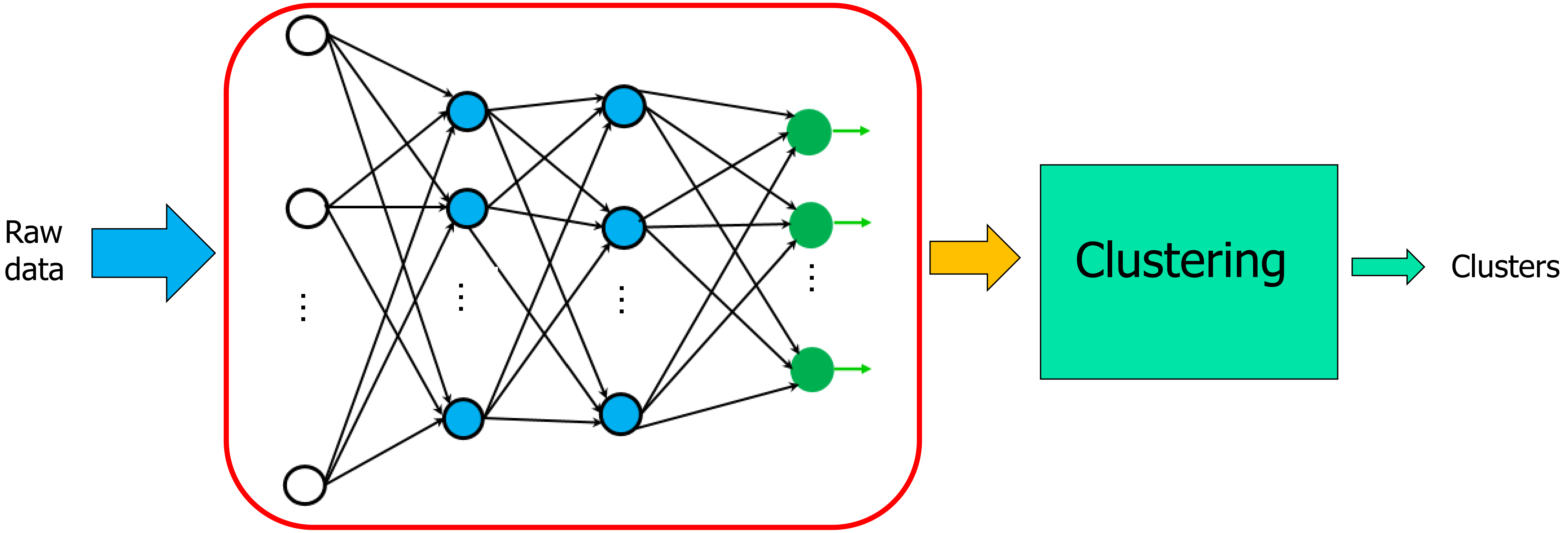


(a) Dataset (b) DBSCAN's result (c) NBC's result

Fig. 4. Discoverying clusters of different densities(NBC *vs.* DBCSAN)

7.4 Clustering with Deep Learning

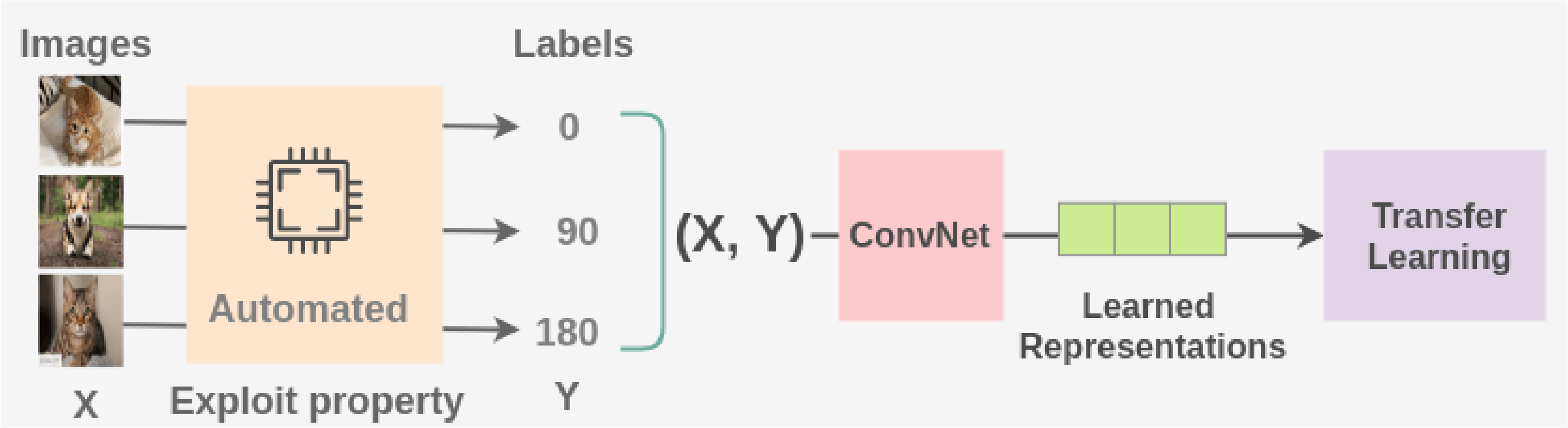
Representation Learning+clustering

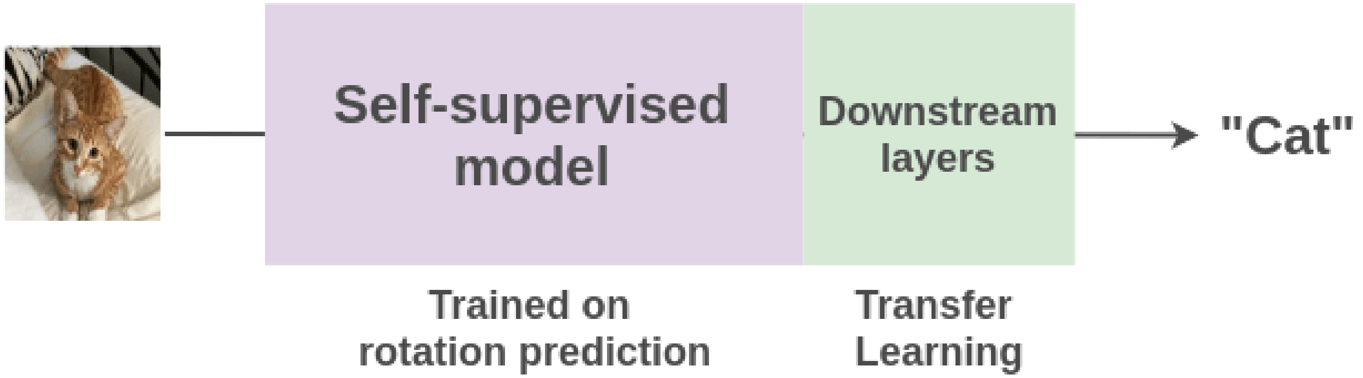


Self-supervised Learning

- *Supervised learning* – learning with labeled data
- *Unsupervised learning* – learning with unlabeled data
- *Self-supervised learning* – representation learning with unlabeled data
  - Learn useful feature representations from unlabeled data through pretext tasks
  - The term "self-supervised" refers to creating its own supervision (without supervision&labels)
  - Self-supervised learning is one category of unsupervised learning

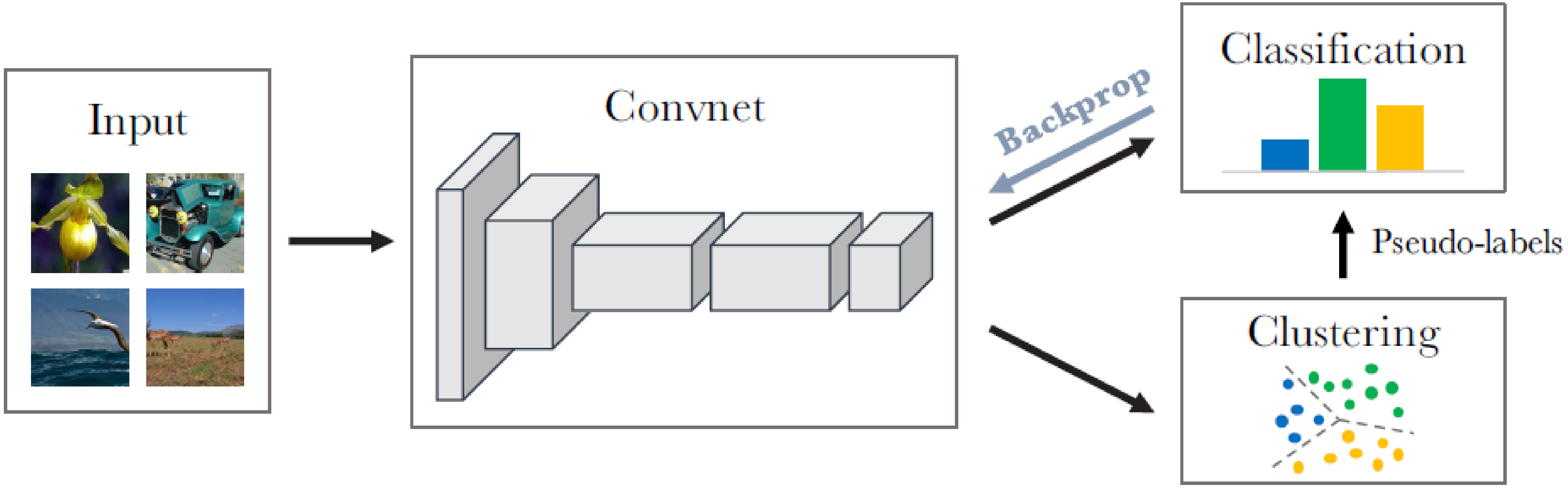
Self-Supervised Learning: an example





Picture from: Amit Chaudhary – The Illustrated Self-Supervised Learning

Deep Clustering



Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV), 2018.

Acknowledgements

- Some text, figures and formulations are from WWW. Thanks for their sharing. If you have copyright claim please contact with me at yym@hit.edu.cn.
- This lecture is distributed for nonprofit purpose.

Thank You for Your Attention

Contact me at: yym@hit.edu.cn

Tel: 26033008, 13760196623

Address: Rm.1402, H# Building