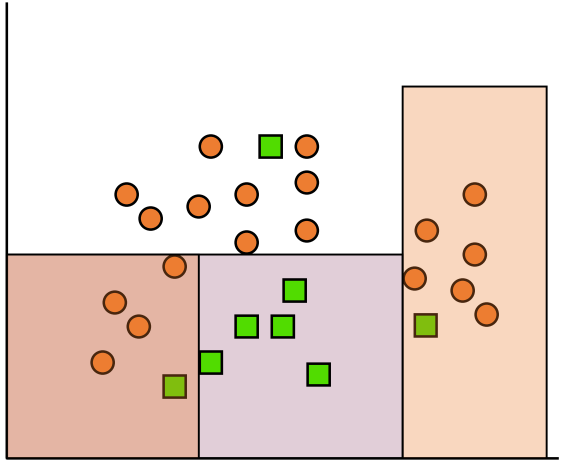
딥러닝 이론 및 소프트웨어 구현
비지도학습 및 강화학습

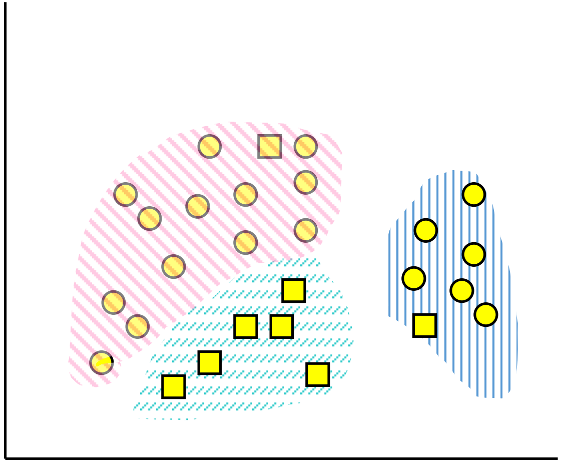
김중헌 교수
중앙대학교 소프트웨어학부
https://sites.google.com/site/joongheonkim/
joongheon@gmail.com

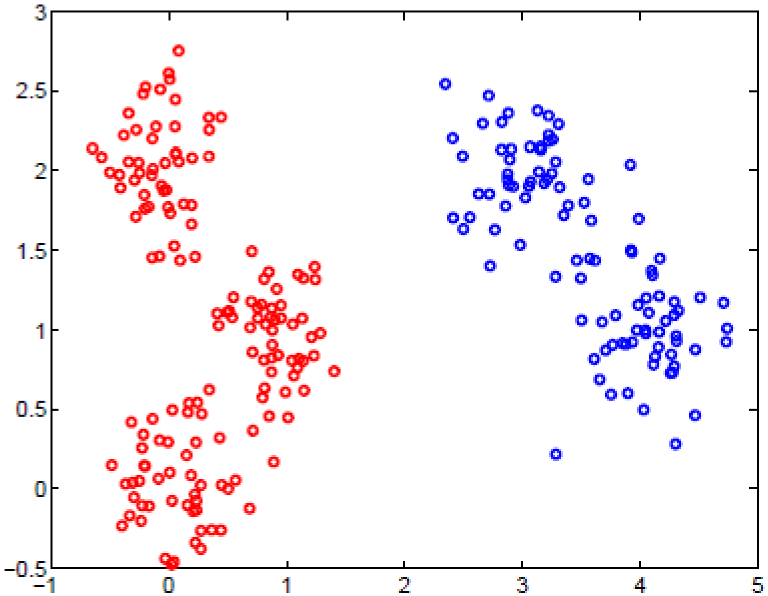
- **Introduction**
- Data Types and Representations
- Distance Measures
- Major Clustering Approaches

- Classification vs. Clustering
  - Classification
    - Supervised Learning
    - Learns a method for predicting the instance class from pre-labeled (classified) instances



- Classification vs. Clustering
  - Clustering
    - Unsupervised Learning
    - Finds "natural" grouping of instances given un-labeled data




3
2.5
2
1.5
1
0.5
0
−0.5
−1
0
1
2
3
4
5

Blue shark, sheep, cat, dog

Lizard, sparrow, viper, seagull, gold fish, frog, red mullet

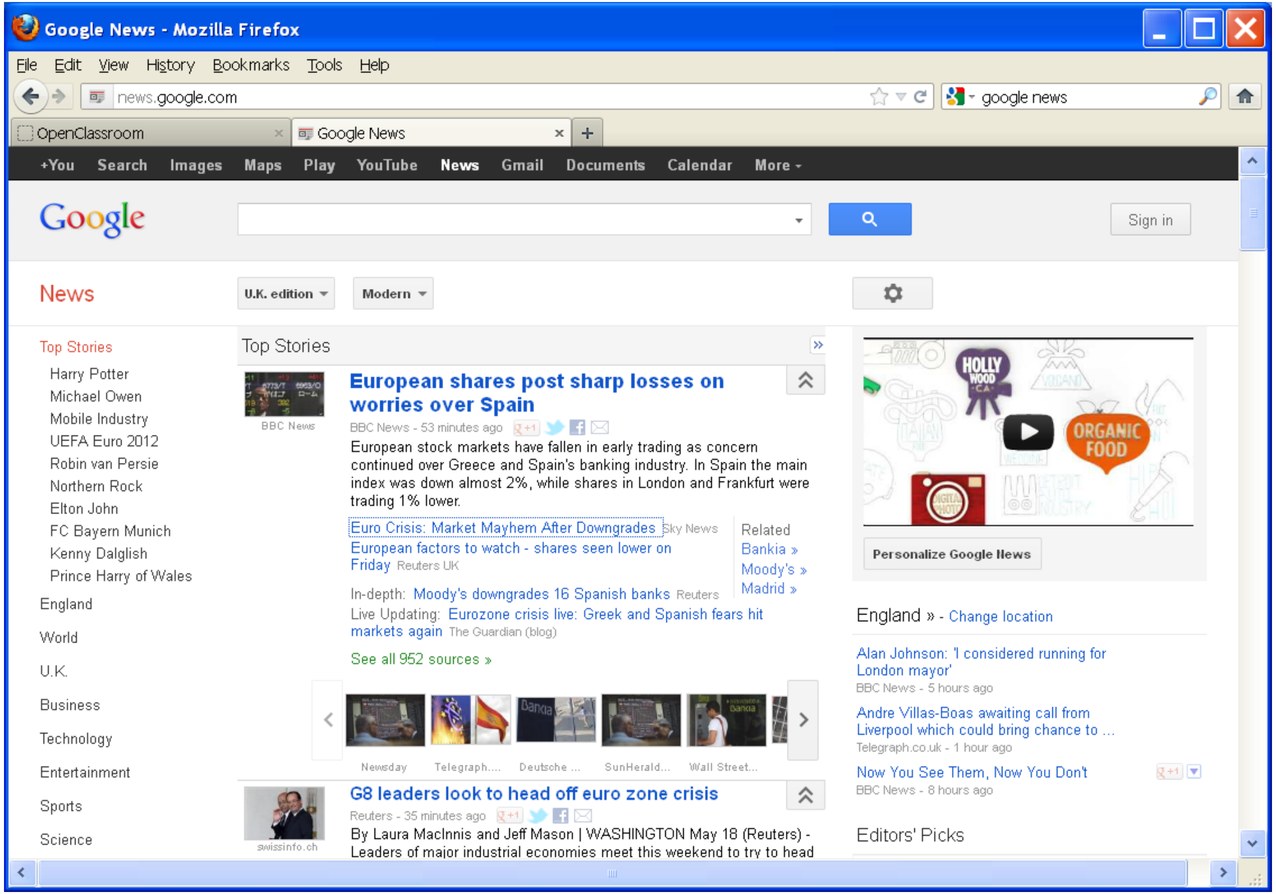
1. Two clusters
2. Clustering criterion:
   How animals bear their progeny

Gold fish, red mullet, blue shark

Sheep, sparrow, dog, cat, seagull, lizard, frog, viper

1. Two clusters
2. Clustering criterion:
   Existence of lungs

Introduction: Real Applications (Google News)



- Introduction
- **Data Types and Representations**
- Distance Measures
- Major Clustering Approaches

- Discrete vs. Continuous
  - **Discrete Feature**
    - Has only a finite set of values
      e.g., zip codes, rank, or the set of words in a collection of documents
    - Sometimes, represented as integer variable
  - **Continuous Feature**
    - Has real numbers as feature values
      e.g., temperature, height, or weight
    - Practically, real values can only be measured and represented using a finite number of digits
    - Continuous features are typically represented as floating-point variables

- Data representations
  - Data matrix (object-by-feature structure)

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

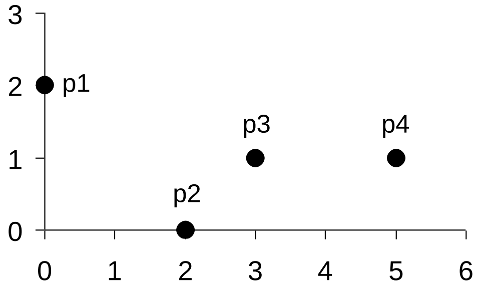
- n data points (objects) with p dimensions (features)
- Two modes: row and column represent different entities

  - Distance/dissimilarity matrix (object-by-object structure)

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

- n data points, but registers only the distance
- A symmetric/triangular matrix
- Single mode: row and column for the same entity (distance)

- Examples



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Data Matrix

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix (i.e., Dissimilarity Matrix) for Euclidean Distance

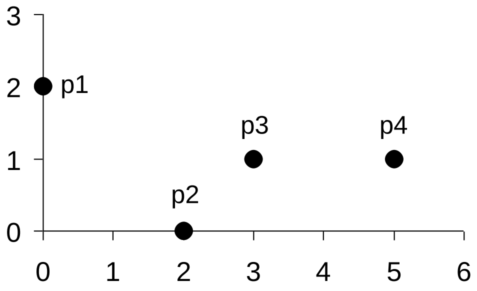
- Introduction
- Data Types and Representations
- **Distance Measures**
- Major Clustering Approaches

- **Minkowski Distance** (http://en.wikipedia.org/wiki/Minkowski_distance)
  - For $\vec{x} = (x_1, \ldots, x_n)$ and $\vec{y} = (y_1, \ldots, y_n)$

$$\boxed{d(\vec{x}, \vec{y}) = \left(|x_1 - y_1|^p + |x_2 - y_2|^p + \cdots + |x_n - y_n|^p\right)^{1/p}}$$

    - $p = 1$: Manhattan (city block) distance
    - $p = 2$: Euclidean distance

  - Do not confuse p with n, i.e., all these distances are defined based on all numbers of features (dimensions).
  - A generic measure: use appropriate p in different applications



L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

Distance Matrix for Manhattan Distance

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

Data Matrix

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix for Euclidean Distance

- **Cosine Measure (Similarity vs. Distance)**
  - For $\vec{x} = (x_1, \dots, x_n)$ and $\vec{y} = (y_1, \dots, y_n)$

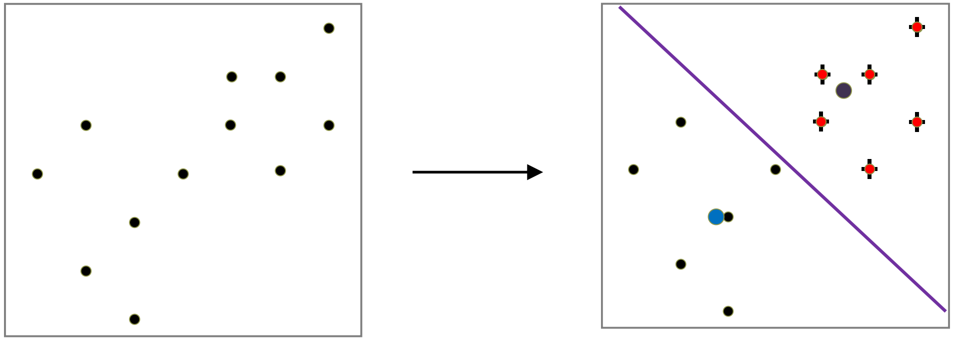
$$d(\vec{x}, \vec{y}) = 1 - \cos(\vec{x}, \vec{y})$$

$$\cos(\vec{x}, \vec{y}) = \frac{x_1 y_1 + \cdots + x_n y_n}{\sqrt{x_1^2 + \cdots + x_n^2}\sqrt{y_1^2 + \cdots + y_n^2}}$$

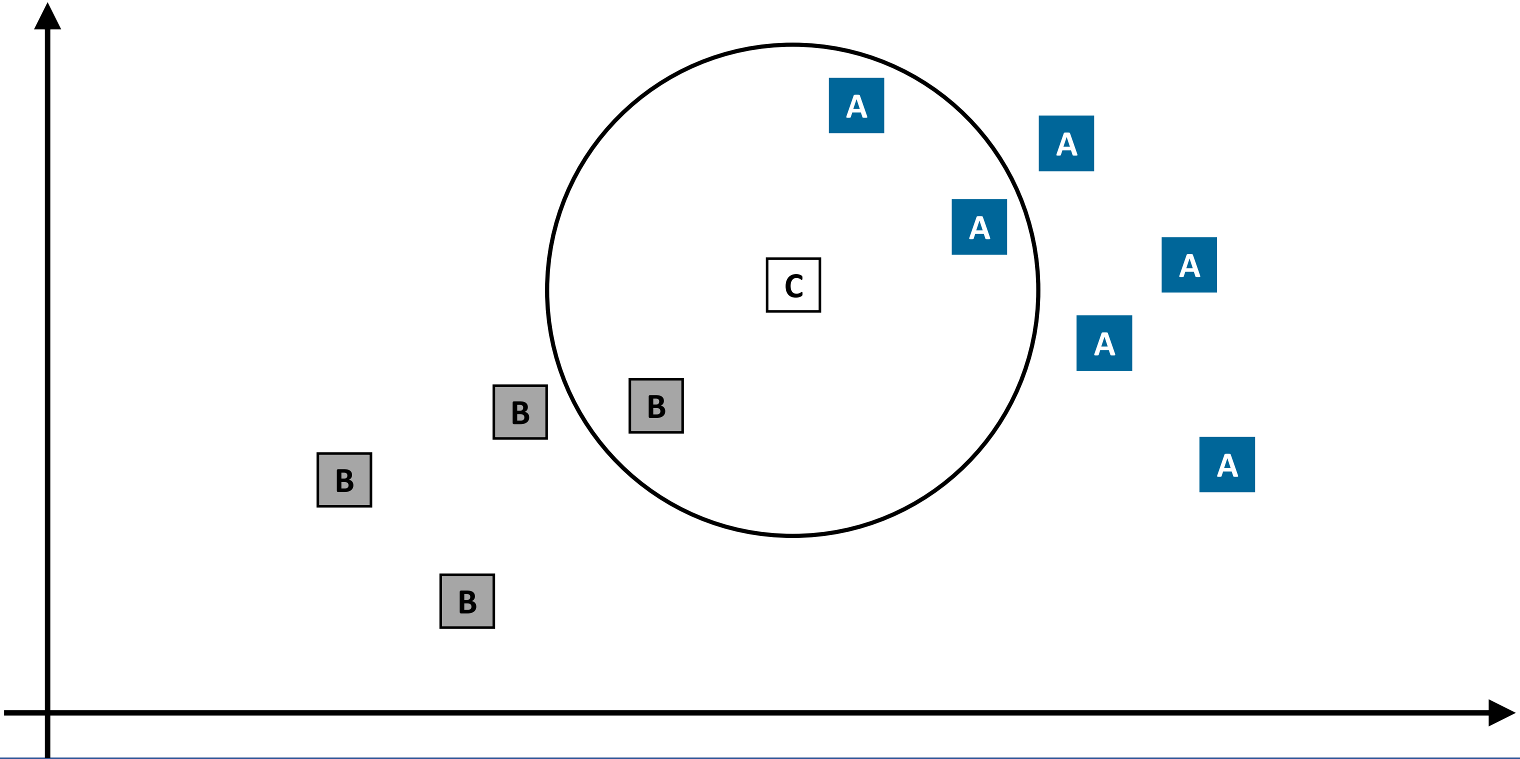
  - Property: $0 \leq d(\vec{x}, \vec{y}) \leq 2$
  - Nonmetric vector objects: keywords in documents, gene features in micro-arrays, …
  - Applications: information retrieval, biologic taxonomy, ...

- Introduction
- Data Types and Representations
- Distance Measures
- **Major Clustering Approaches**

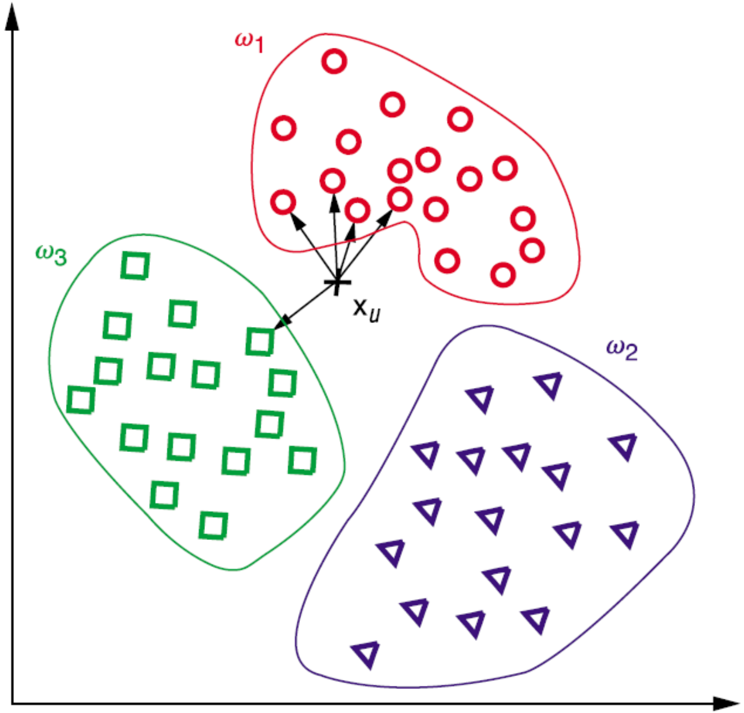
- Partitioning Approach
  - Typical methods: K-means, K-medoids, CLARANS, ……



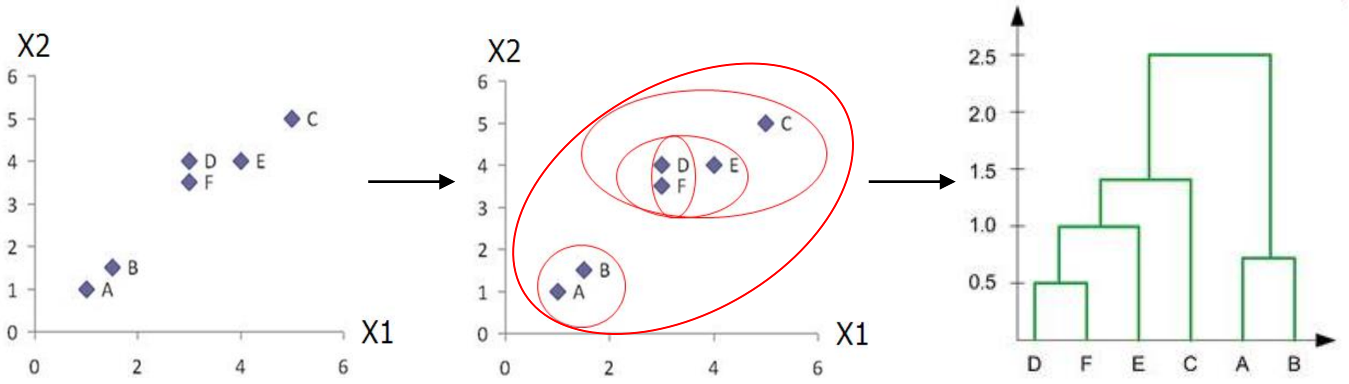
- Partitioning Approach
  - kNN (k Nearest Neighbor: k=3)



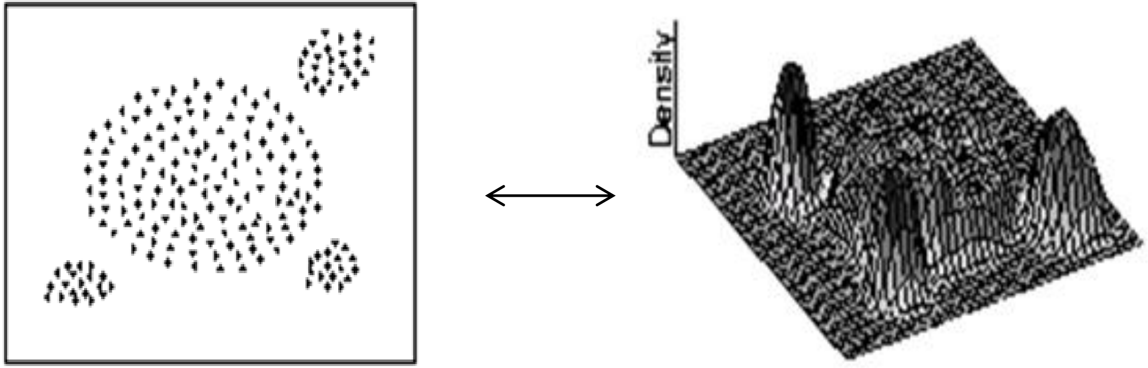
- Partitioning Approach
  - kNN



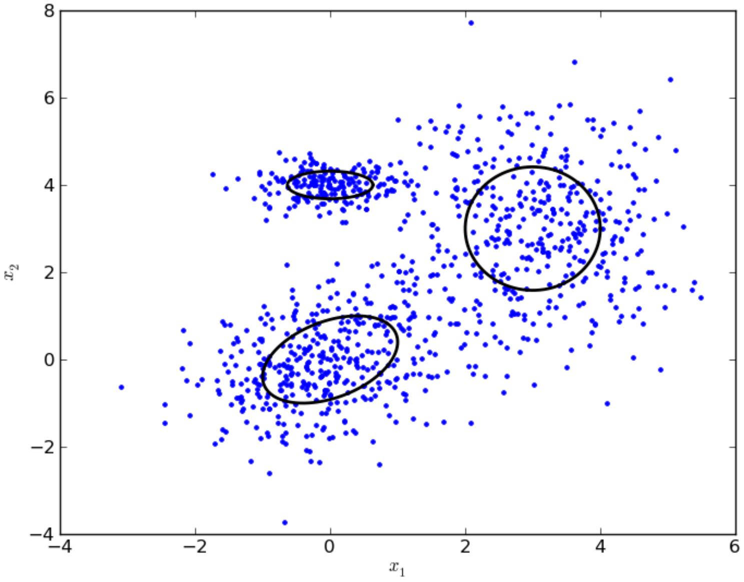
- Hierarchical Approach
  - Typical methods: Agglomerative, Diana, Agnes, BIRCH, ROCK, ……



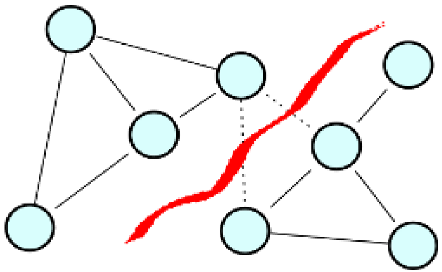
- Density-based Approach
    - Typical methods: DBSACN, OPTICS, DenClue, ……

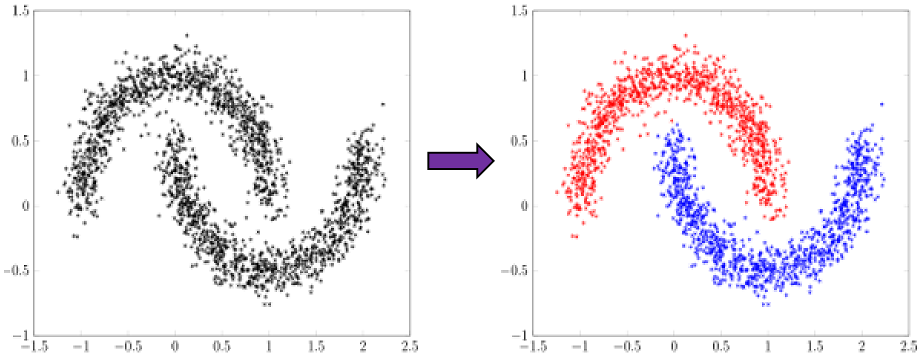


- Model-based Approach
  - Typical methods: Gaussian Mixture Model (GMM), COBWEB, ……



- Spectral Clustering Approach
  - Typical methods: Normalized-Cuts, ……





- Clustering Ensemble Approach
  - Combine multiple clustering results (different partitions)
  - Typical methods: Evidence-accumulation based, graph-based ……