Fall 2004, CIS, Temple University

CIS527: Data Warehousing, Filtering, and Mining

Lecture 6

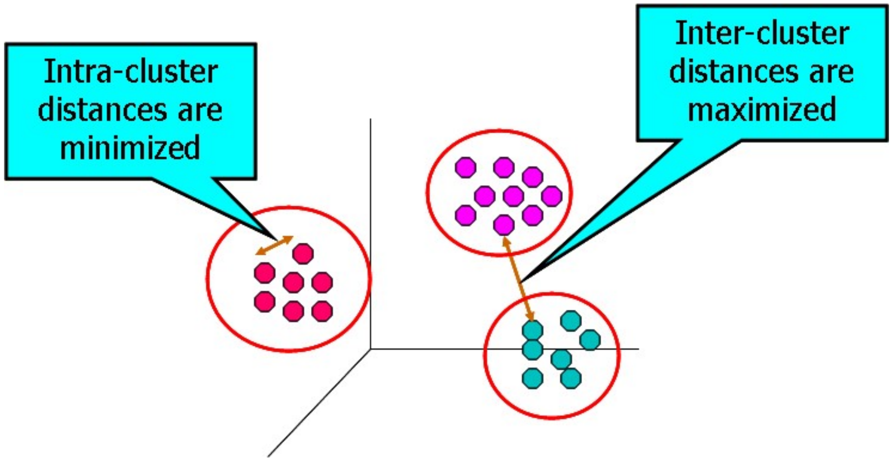
- Clustering

Lecture slides taken/modified from:

- Jiawei Han (http://www-sal.cs.uiuc.edu/~hanj/DM_Book.html)
- Vipin Kumar (http://www-users.cs.umn.edu/~kumar/csci5980/index.html)

What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
- Clustering is unsupervised classification: no predefined classes
- Typical applications
  - to get insight into data
  - as a preprocessing step


Intra-cluster distances are minimized
Inter-cluster distances are maximized

General Applications of Clustering

- Pattern Recognition
- Spatial Data Analysis
  - create thematic maps in GIS by clustering feature spaces
  - detect spatial clusters and explain them in spatial data mining
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Land use: Identification of areas of similar land use in an earth observation database
- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability

Data Structures in Clustering

- Data matrix
  - (two modes)

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

- Dissimilarity matrix
  - (one mode)

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

Measuring Similarity

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric:

  d(*i, j*)

- There is a separate “quality” function that measures the “goodness” of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
- Weights should be associated with different variables based on applications and data semantics.
- It is hard to define “similar enough” or “good enough”
  - the answer is typically highly subjective.

Interval-valued variables (Biến phạm vi)

- Standardize data
  - Calculate the mean squared deviation:

$$s_f = \frac{1}{n}(|x_{1f} - m_f|^2 + |x_{2f} - m_f|^2 + ... + |x_{nf} - m_f|^2)$$

where

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + ... + x_{nf}).$$

  - Calculate the standardized measurement (*z-score*)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using mean absolute deviation (chỉ có trị tuyệt đối chứ không bình phương như công thức trên) could be more robust than using standard deviation

Similarity and Dissimilarity Between Objects

- Distances are normally used to measure the similarity or dissimilarity between two data objects
- Some popular ones include: *Minkowski distance*:

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

where *i* = (x_{i1}, x_{i2}, …, x_{ip}) and *j* = (x_{j1}, x_{j2}, …, x_{jp}) are two *p*-dimensional data objects, and *q* is a positive integer

- If *q = 1*, *d* is Manhattan distance

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + ... + |x_{i_p} - x_{j_p}|$$

Similarity and Dissimilarity Between Objects

- *If q* = 2, *d* is Euclidean distance:

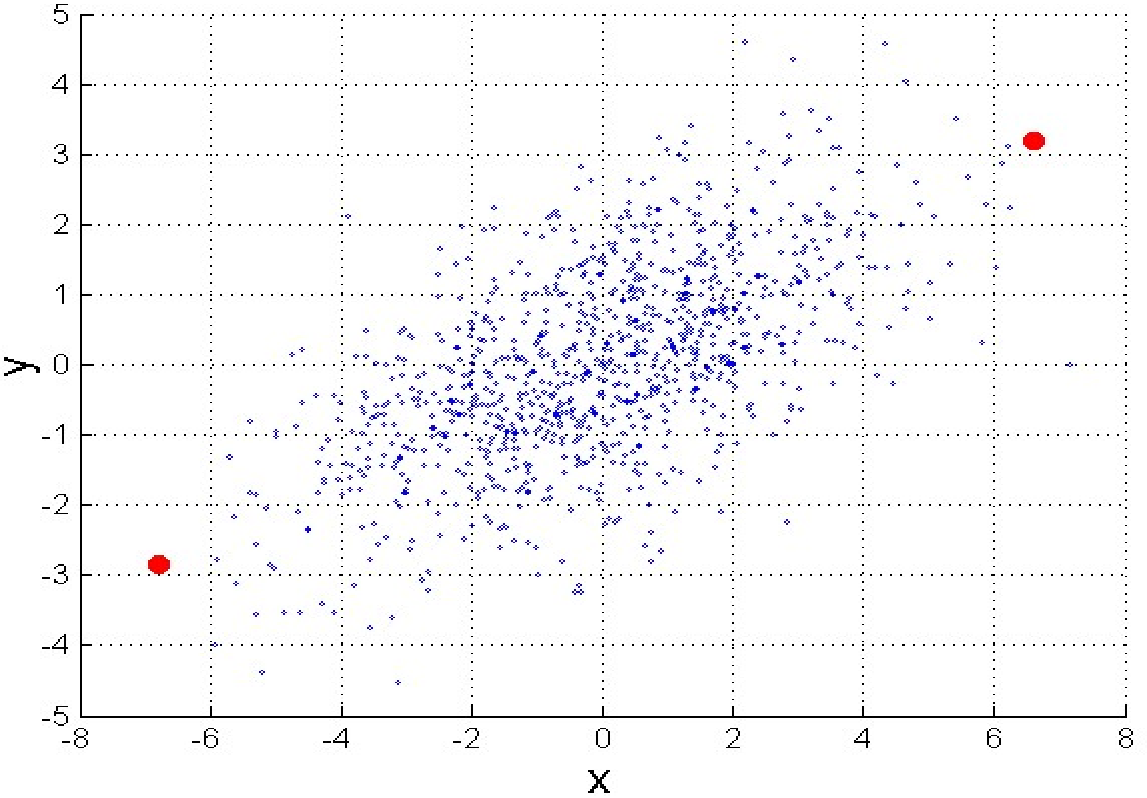
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \ldots + |x_{i_p} - x_{j_p}|^2)}$$

  - Properties
    - $d(i,j) \geq 0$
    - $d(i,i) = 0$
    - $d(i,j) = d(j,i)$
    - $d(i,j) \leq d(i,k) + d(k,j)$

- Also one can use weighted distance, parametric Pearson product moment correlation, or other disimilarity measures.

Mahalanobis Distance

$$\mathit{mahalanobis}(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$

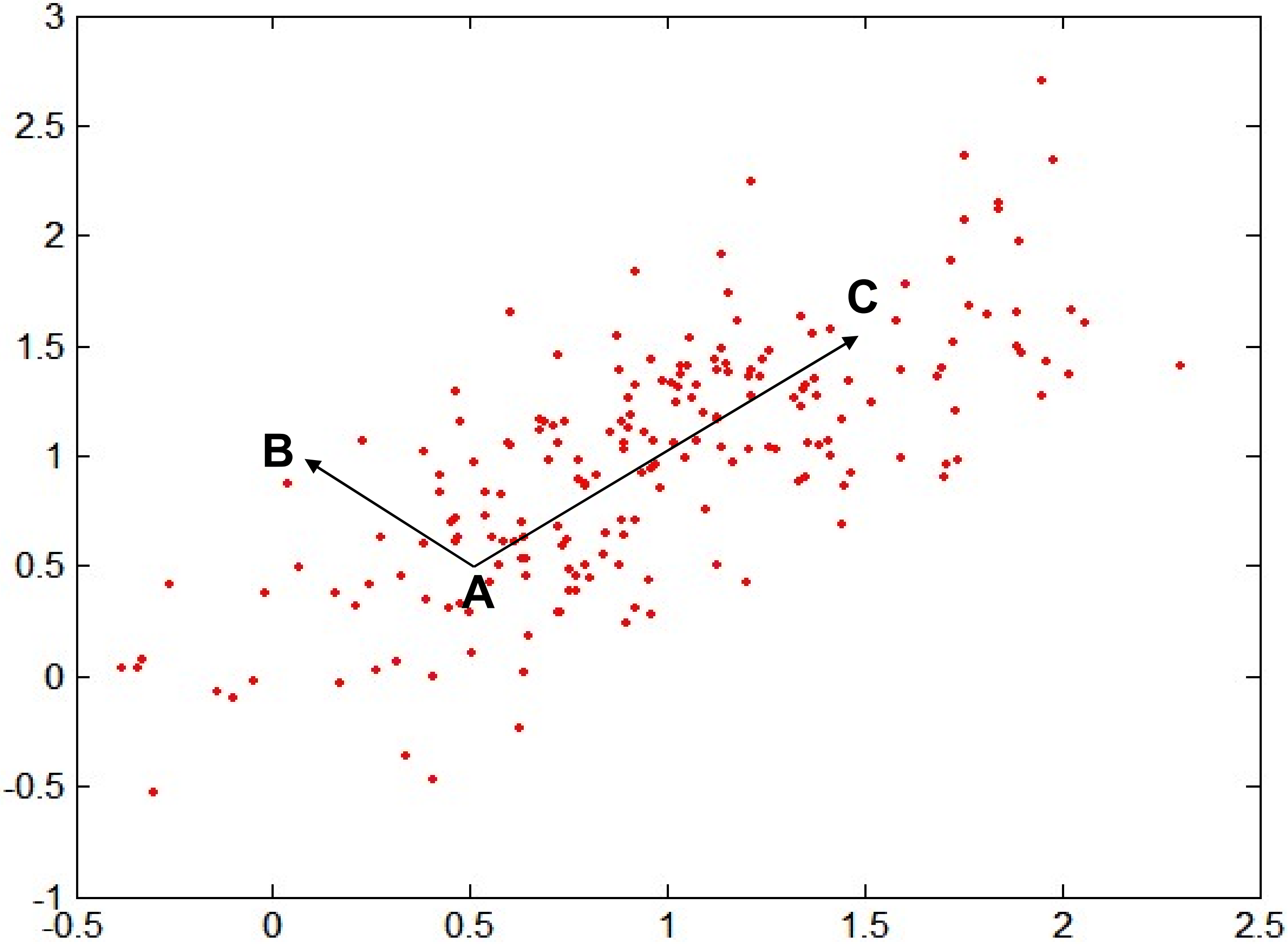


Σ is the covariance matrix of the input data *X*

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$$

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2|| ,$$

where $\bullet$ indicates vector dot product and $||\ d\ ||$ is the length of vector d.

Giá trị từ -1 đến 1, 1 là tương đồng hoàn toàn

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**
d_2 = **1 0 0 0 0 0 0 1 0 2**

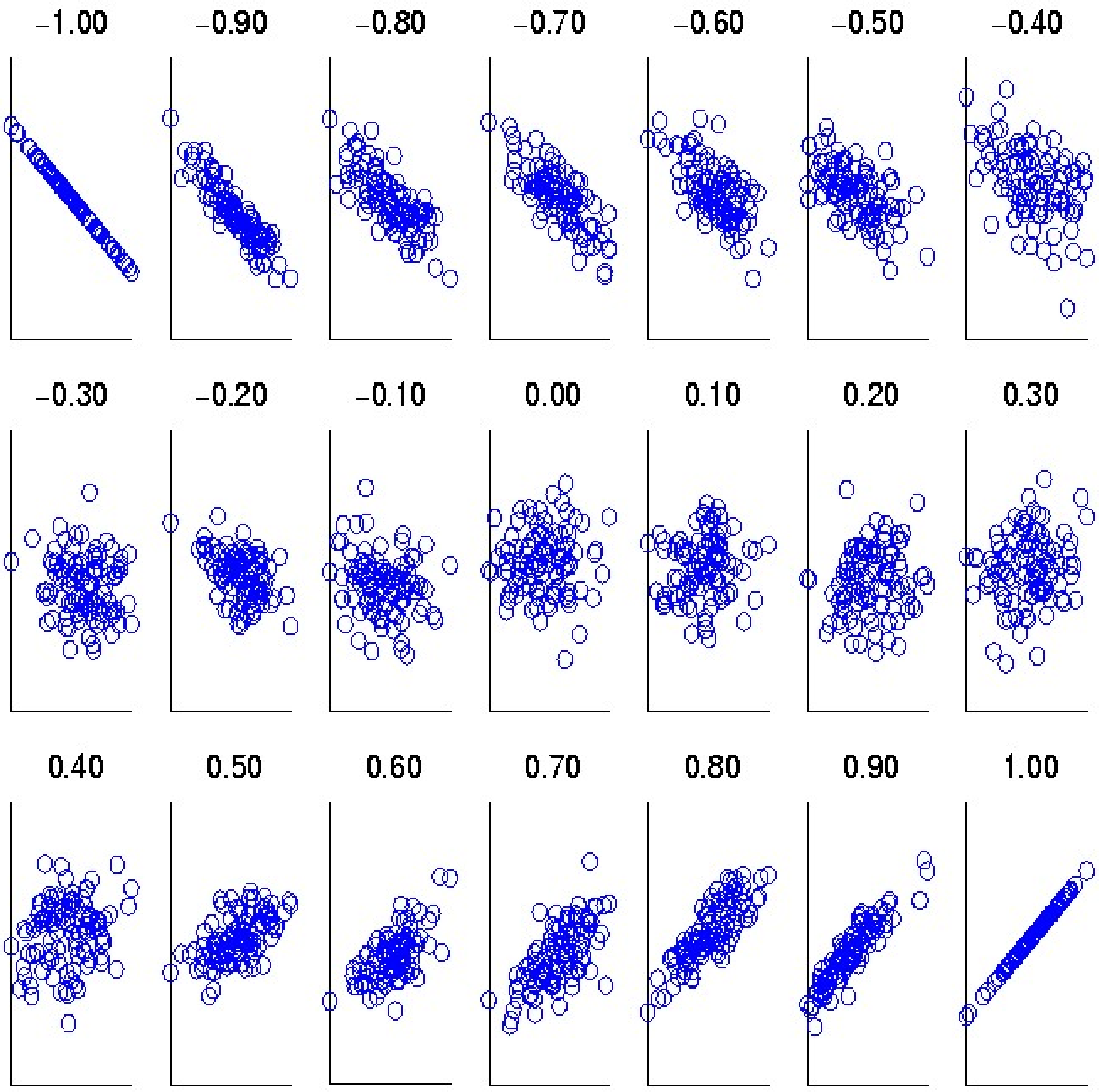
$d_1 \bullet d_2$= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5

$||d_1||$ = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{\mathbf{0.5}}$ = (42) $^{\mathbf{0.5}}$ = 6.481

$||d_2||$ = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{\mathbf{0.5}}$ = (6) $^{\mathbf{0.5}}$ = 2.245

$\cos(d_1, d_2)$ = .3150

Correlation Measure



Scatter plots showing the similarity from –1 to 1.

Biến nhị phân

- A contingency table for binary data

		Object *j*		
		1	0	*sum*
	1	a	b	$a+b$
Object *i*	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- Độ tương đồng giữa 2 đối tượng i, j có thể được tính bằng khoảng cách đơn giản (bất biến nếu biến nhị phân là đối xứng (0 và 1 ý nghĩa như nhau)): $d(i,j) = \dfrac{b+c}{a+b+c+d}$

- Độ tương đồng Jaccard (không bất biến nếu biến nhị phân là không đối xứng):

$$d(i,j) = \frac{b+c}{a+b+c}$$

Dissimilarity between Binary Variables

- Example

Dấu hiệu bệnh (Fever, Cough) và các Test là bất đối xứng vì Y (có dấu hiệu) hay P (có ho, test dương tính) có ý nghĩa hơn N
Và vì bộ thuộc tính nhiều bất đối xứng => coi bộ này là bất đối xứng => dùng công thức Jaccard

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

- gender is a symmetric attribute
- the remaining attributes are asymmetric binary
- let the values Y and P be set to 1, and the value N be set to 0

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

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

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

Ở đây không bao gồm Gender

Nominal Variables (Biến tên)

- Là một sự tổng quát hóa của biến nhị phân trong đó biến có thể có nhiều hơn 2 trạng thái, e.g., red, yellow, blue,…
- Method 1: Simple matching
  - *m*: số thuộc tính có cùng giá trị của 2 đối tượng i và j
  - *p*: tổng số thuộc tính
  - Khoảng cách có thể được tính bằng: $d(i, j) = \frac{p - m}{p}$
- Method 2: use a large number of binary variables
  - creating a new binary variable for each of the *M* nominal states

Ordinal Variables (Biến trật tự)

- 1 biến có trật tự có thể là rời rạc hoặc liên tục
- Trật tự quan trọng vì thể hiện sự phân bậc đối tượng
- Can be treated like interval-scaled
  - replacing x_{if} (thuộc tính thứ *f* của đối tượng x_i) by their rank

$$r_{if} \in \{1, \dots, M_f\}$$

  - 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 (Biến tỉ lệ)

- Ratio-scaled variable: Biến tỉ lệ là một đơn vị đo lường dương trên một phạm vi phi tuyến hoặc dạng lũy thừa xấp xỉ ví dụ như Ae^{Bt} or Ae^{-Bt}
- Methods:
  - Coi chúng như biến phạm vi (mặc dù không tốt)
  - Áp dụng chuyển đổi logarithmic

$$y_{if} = log(x_{if})$$

  - Coi chúng như dữ liệu có trật tự và liên tục và coi cấp bậc của chúng như khoảng phạm vi

Variables of Mixed Types (Dữ liệu hỗn hợp)

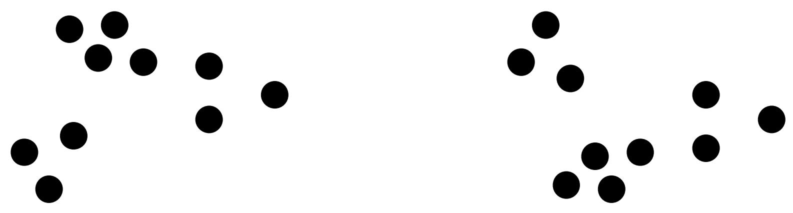
- A database may contain all the six types of variables
  - symmetric binary, asymmetric binary, nominal, ordinal, interval and ratio.
- One may use a weighted formula to combine their effects.

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

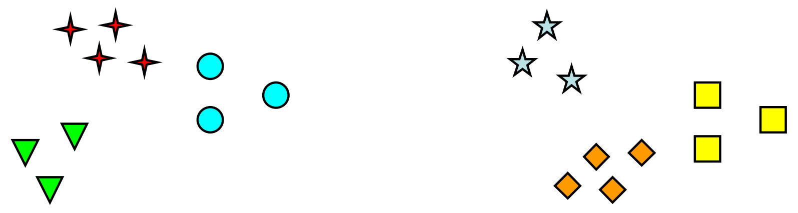
  - *f* là dạng nhị phân hoặc dạng tên:

    $d_{ij}^{(f)} = 0$ nếu $x_{if} = x_{jf}$, không thì $d_{ij}^{(f)} = 1$.
  - *f* là dạng phạm vi: use the normalized distance
  - *f* dạng trật tự hoặc tỉ lệ:
    - compute ranks r_{if} and
    - and treat z_{if} as interval-scaled $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

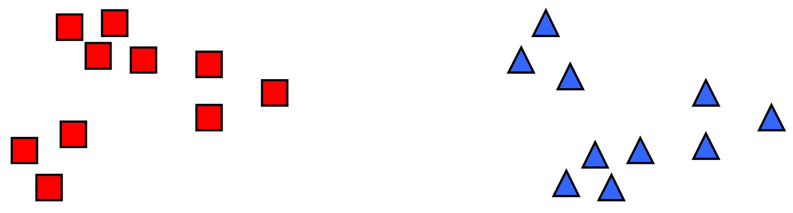
Notion of a Cluster can be Ambiguous



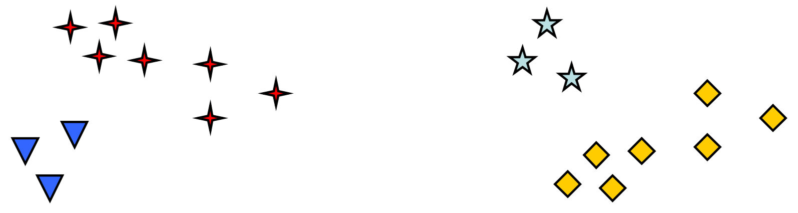
How many clusters?



Six Clusters



Two Clusters



Four Clusters

Other Distinctions Between Sets of Clusters

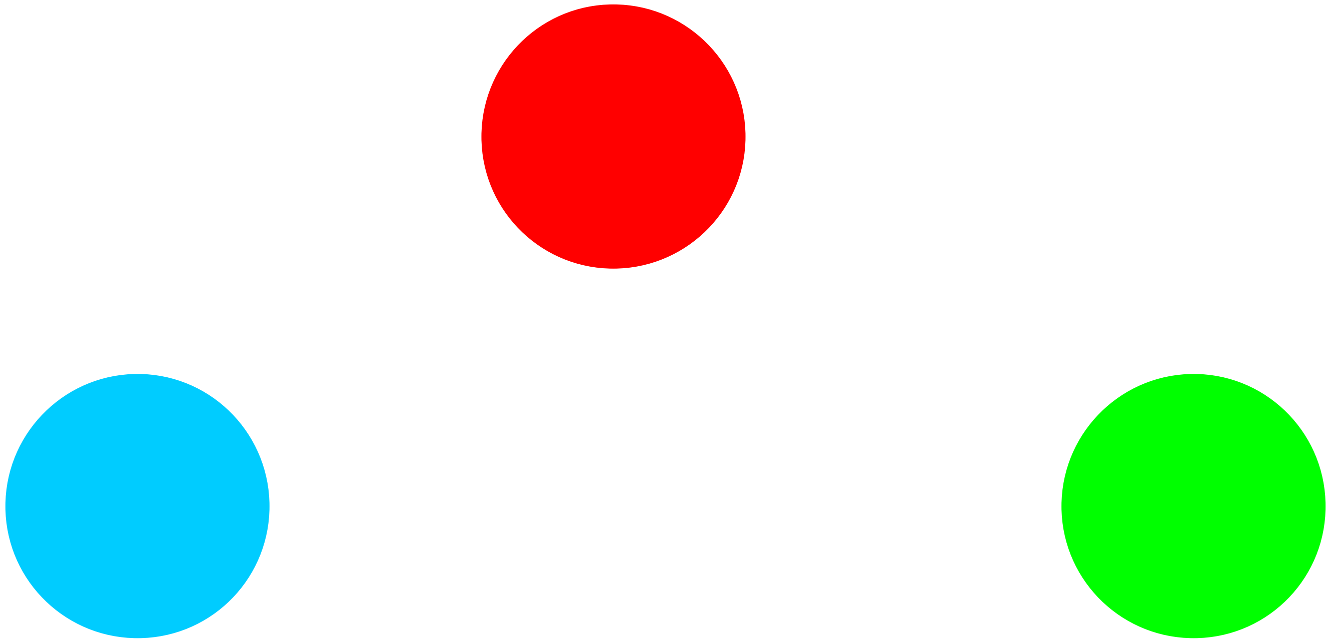
- Exclusive versus non-exclusive
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or ‘border’ points
- Fuzzy versus non-fuzzy
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- Partial versus complete
  - In some cases, we only want to cluster some of the data
- Heterogeneous versus homogeneous
  - Cluster of widely different sizes, shapes, and densities

Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

Types of Clusters: Well-Separated

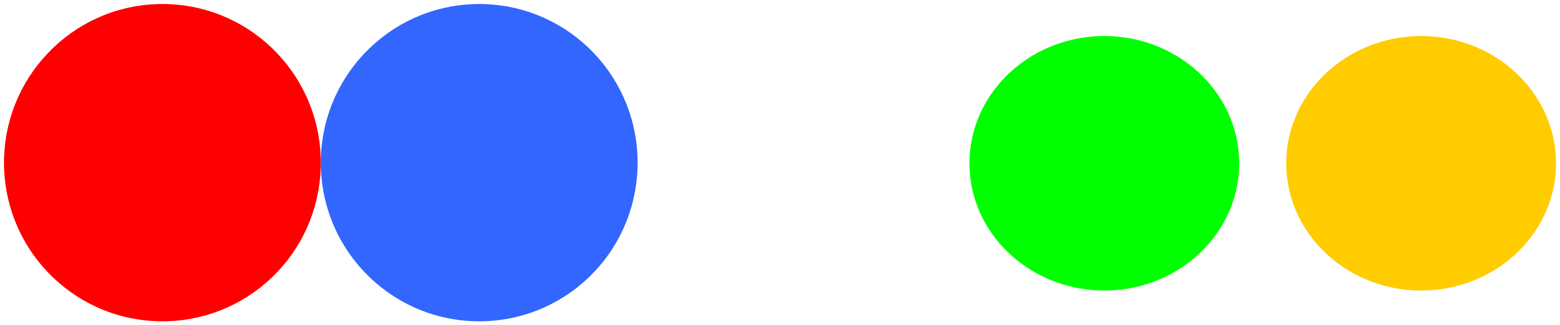
- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Types of Clusters: Center-Based

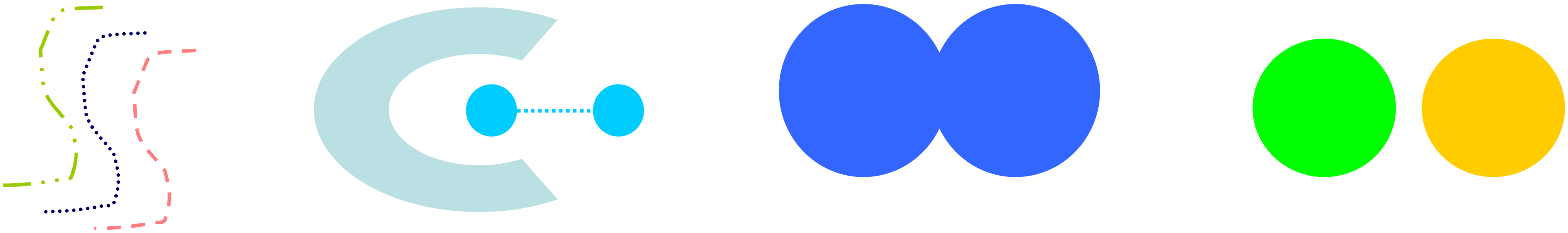
- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster



4 center-based clusters

Types of Clusters: Contiguity-Based

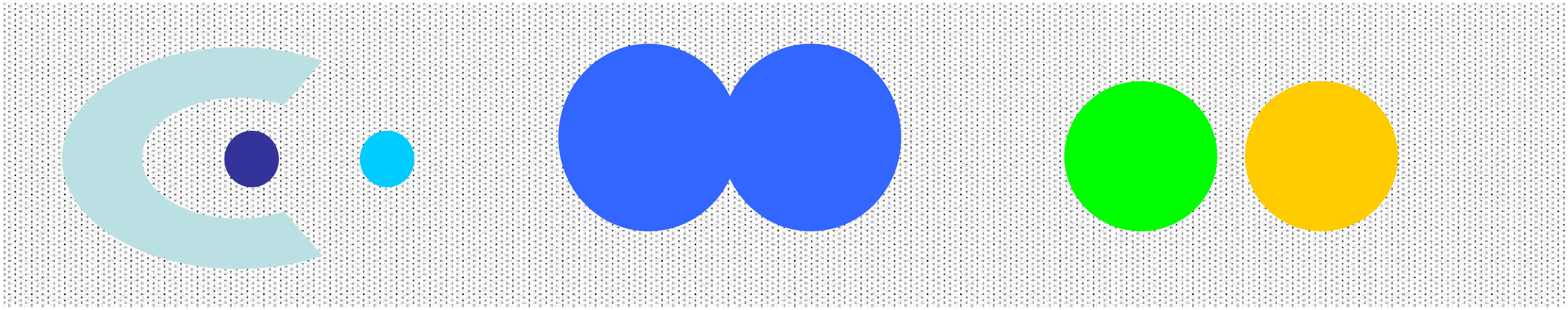
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



8 contiguous clusters

Types of Clusters: Density-Based

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.

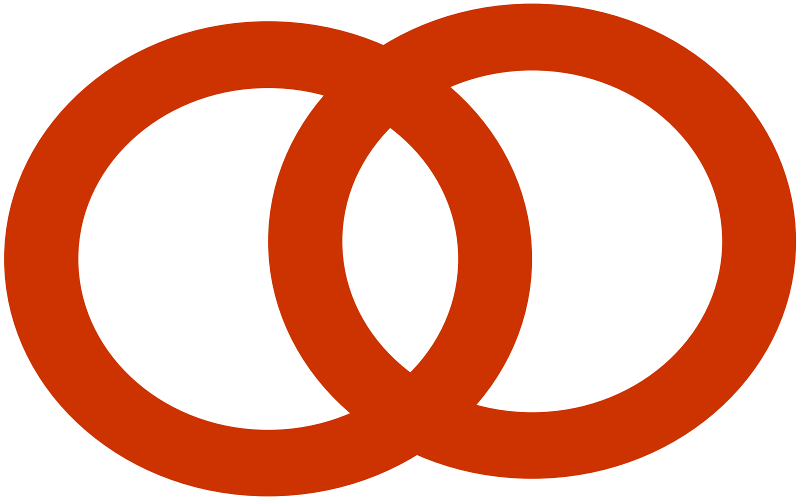


6 density-based clusters

Types of Clusters: Conceptual Clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.

.



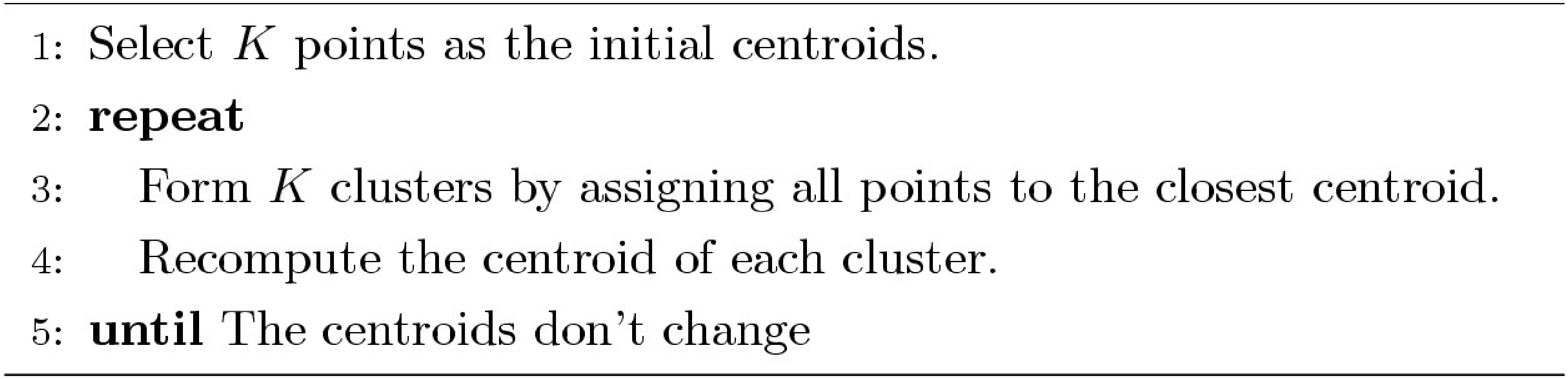
2 Overlapping Circles

Major Clustering Approaches

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion
- Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based: based on connectivity and density functions
- Grid-based: based on a multiple-level granularity structure
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

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

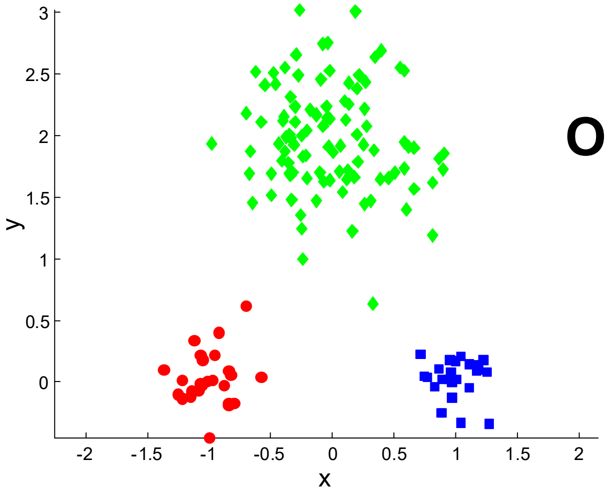


```
Select $K$ points as the initial centroids.
repeat
    Form $K$ clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

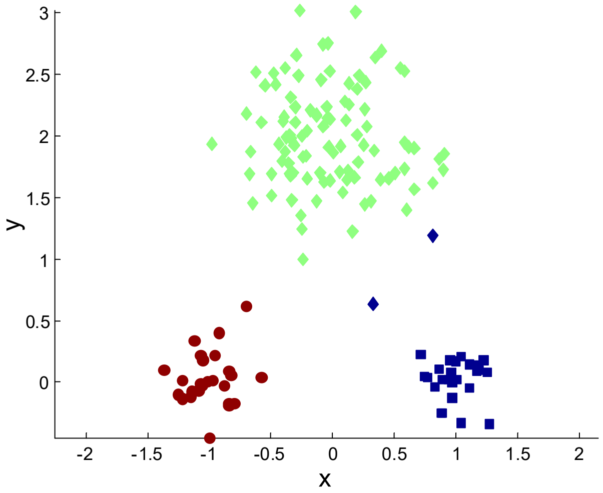
K-means Clustering – Details

- Initial centroids are often chosen randomly .
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- ‘Closeness’ is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to ‘Until relatively few points change clusters’
- Complexity is O(n * K * I * d)
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of attributes

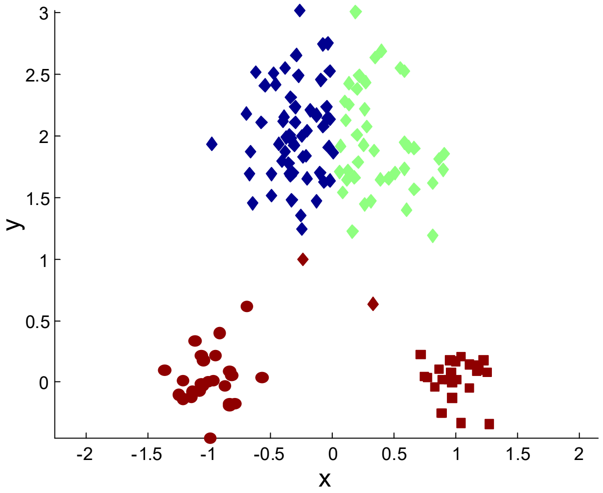
Two different K-means Clusterings



Original Points



Optimal Clustering



Sub-optimal Clustering

- Importance of choosing initial centroids

Evaluating K-means Clusters

- Tổng bình phương lỗi (Sum of Squared Error (SSE)):
  - Đối với mỗi điểm, lỗi được tính là khoảng cách tới cụm gần nhất
  - Các lỗi tính được ở trên bình phương lên cộng hết lại.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x là một điểm dữ liệu trong cụm Ci và mi là điểm đại diện cho cụm Ci
    - can show that m_i corresponds to the center (mean) of the cluster
  - Given two clusters, we can choose the one with the smallest error
  - One easy way to reduce SSE is to increase K, the number of clusters
    - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing
- Bisecting K-means
  - Not as susceptible to initialization issues

Handling Empty Clusters

- Basic K-means algorithm can yield empty clusters

- Several strategies
  - Lựa chọn các điểm có đóng góp nhiều nhất tới tổng bình phương lỗi SSE và đưa điểm đó vào cụm dữ liệu rỗng
  - Lựa chọn một điểm trong cụm có SSE cao nhất và đưa vào cụm rỗng đó để giảm SSE nhiều nhất có thể đồng thời làm cụm rỗng có phần tử
  - Nếu có nhiều cụm rỗng thì công việc trên được lặp lại nhiều lần

Pre-processing and Post-processing

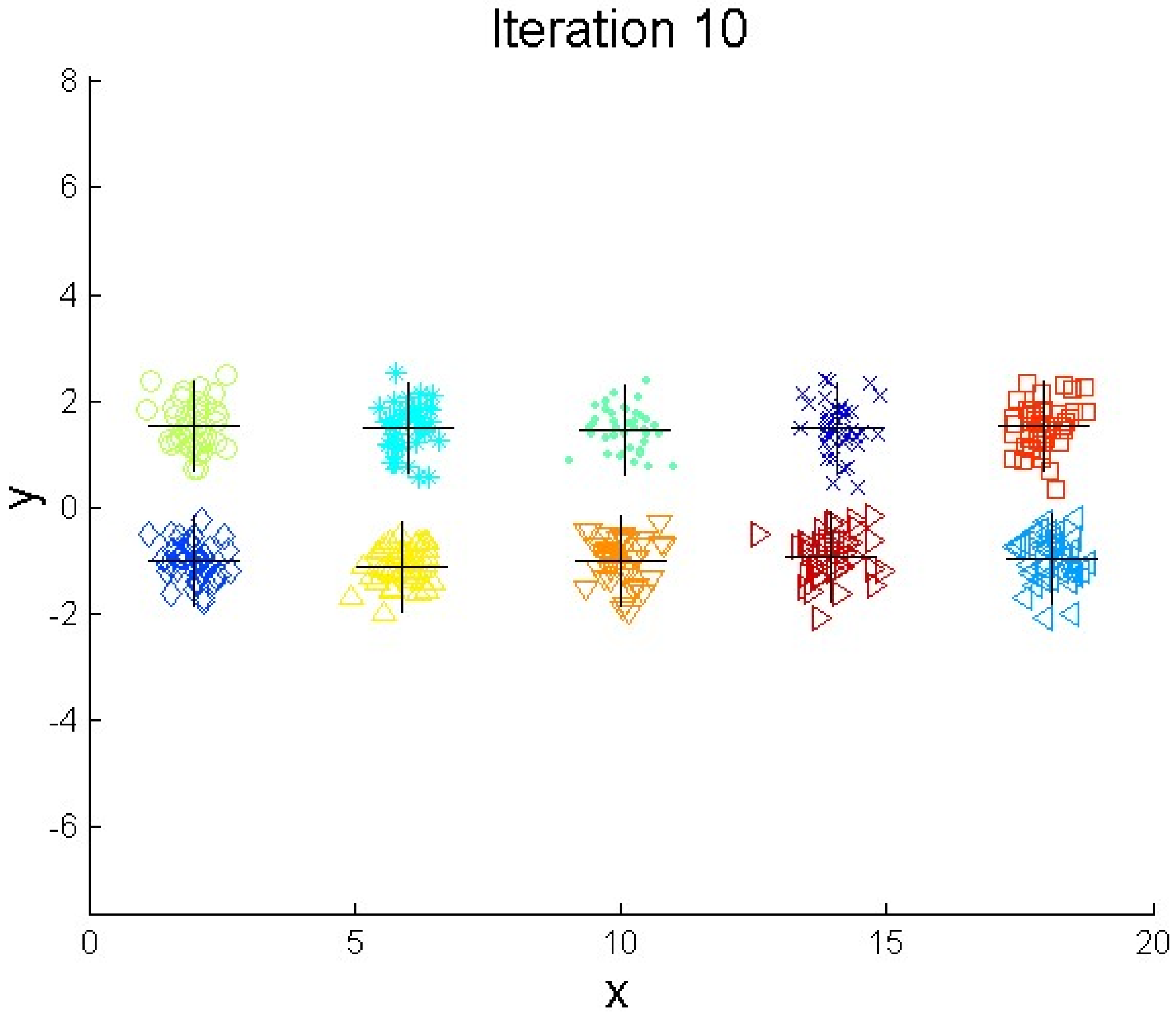
- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers
  - Phân chia những cụm lỏng lẻo (hay mật độ các phần tử trong cụm không đồng đều, chỗ dày đặc, chỗ thưa thớt), hay nói cách khác là các cụm có tổng bình phương lỗi lớn thành các cụm nhỏ
  - Merge clusters that are ‘close’ and that have relatively low SSE
  - Can use these steps during the clustering process
    - ISODATA

Bisecting K-means (K-means phân đôi)

- Bisecting K-means algorithm
  - một biến đổi của K-mean mà có thể sinh ra một sự phân cụm có phân cấp hoặc phân cụm dạng phân mảnh

```
1.  Khởi tạo danh sách L các cụm để chứa các cụm tìm được, ban đầu chỉ chứa có một cụm bao gồm tất cả các điểm
2.  Lặp các bước sau
3.  Chọn một cụm trong danh sách L các cụm trên
4.  For i=1 to số lượng vòng lặp định trước do
5.      Phân đôi cụm được lựa chọn thành 2 phân cụm bằng K-mean
6.  End for
7.  Thêm hai phân cụm kết quả của những lần phân đôi cụm trên với tổng bình phương lỗi SSE nhỏ nhất vào danh sách các cụm
8.  Cho đến khi danh sách các cụm chứa K cụm thì dừng.
```

Bisecting K-means Example

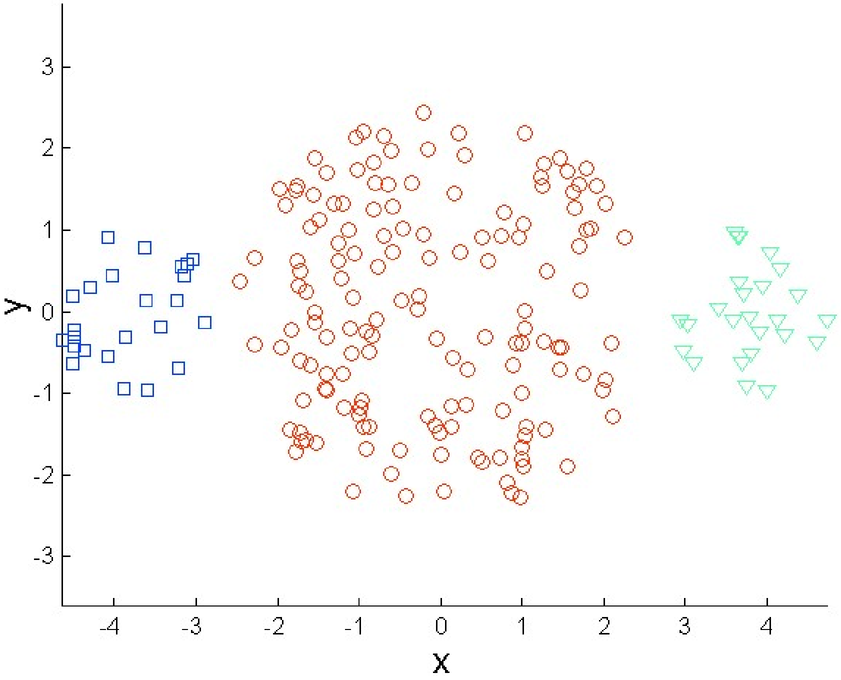


Limitations of K-means

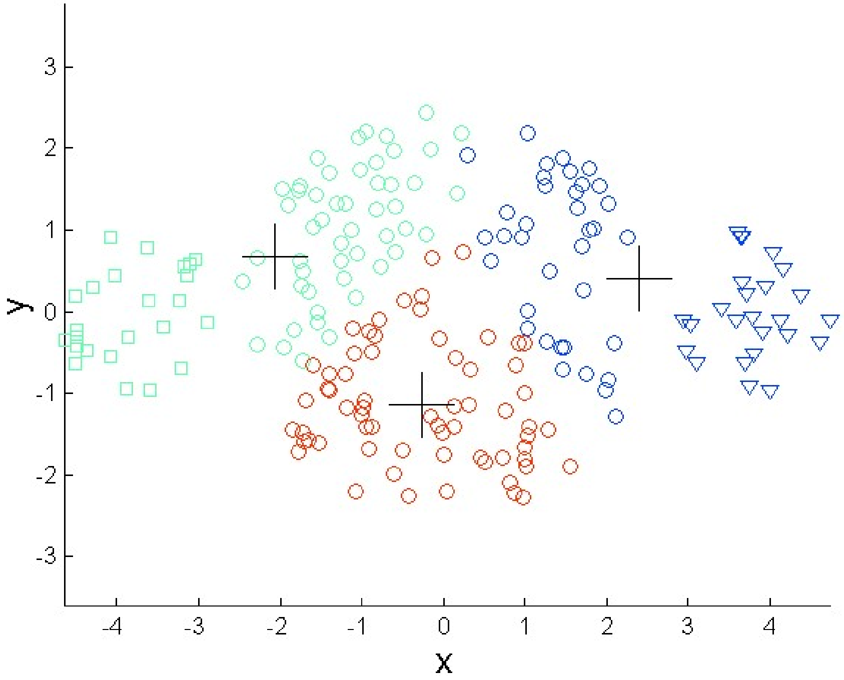
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- K-means has problems when the data contains outliers.

Limitations of K-means: Differing Sizes

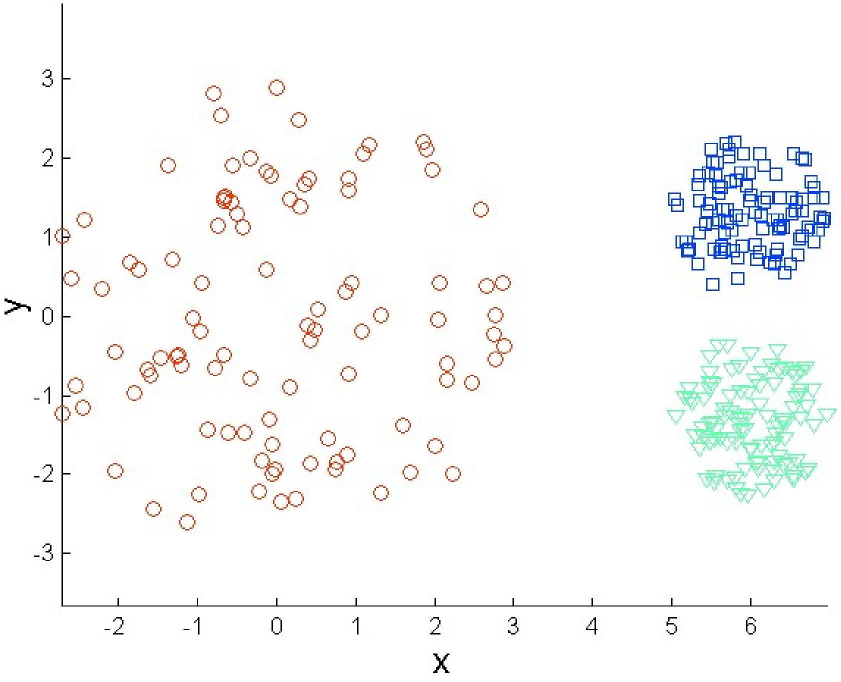


Original Points

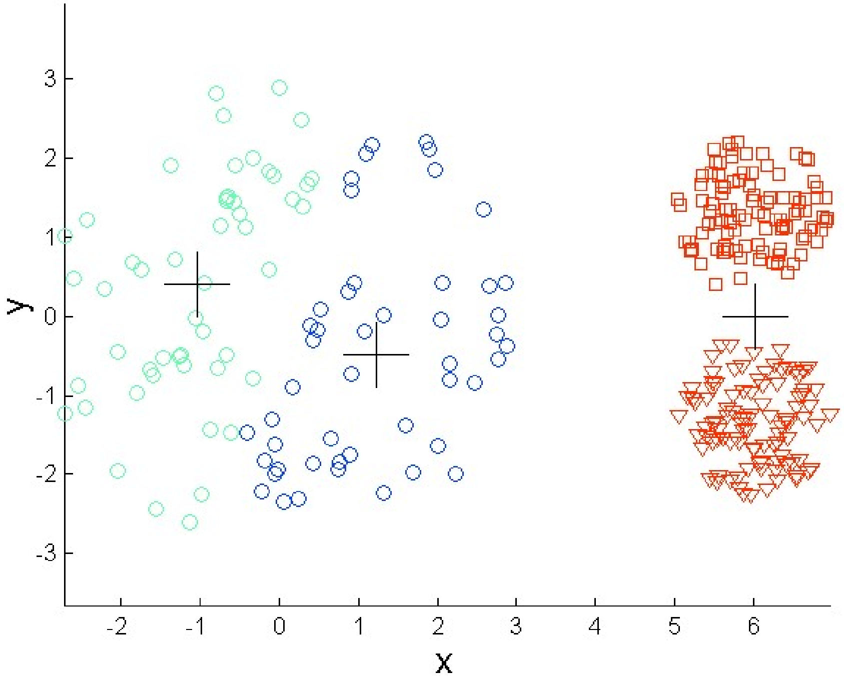


K-means (3 Clusters)

Limitations of K-means: Differing Density

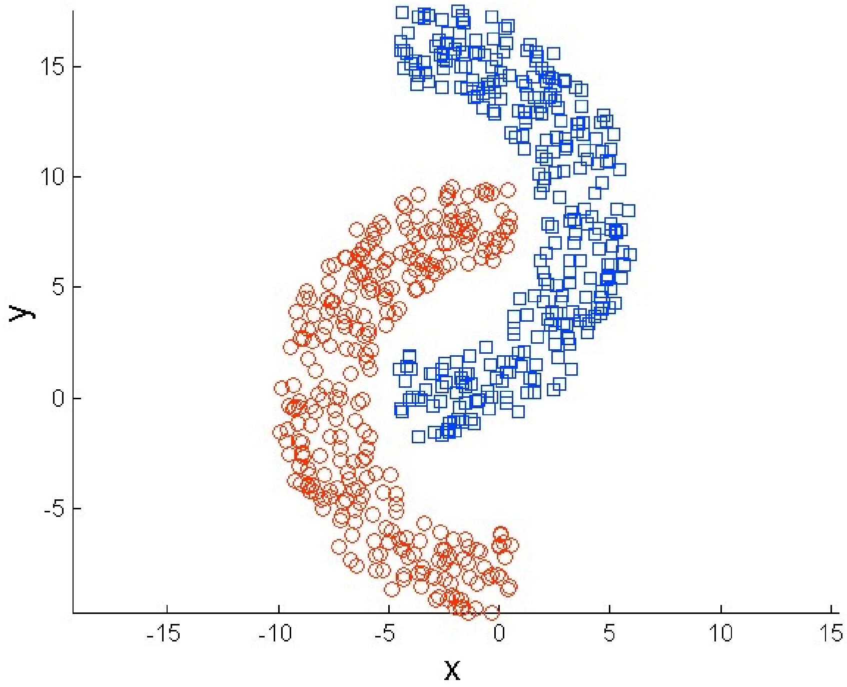


Original Points

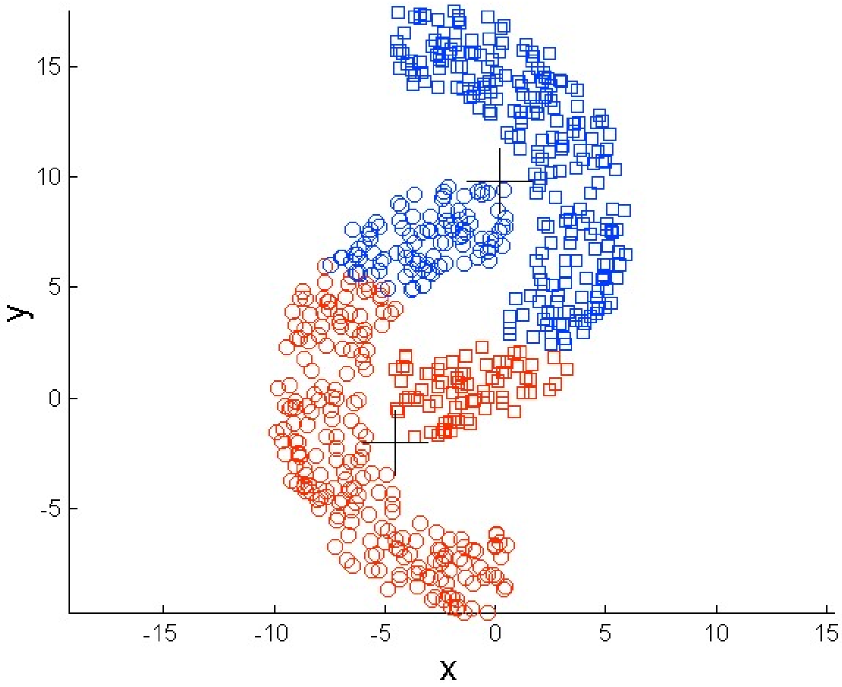


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

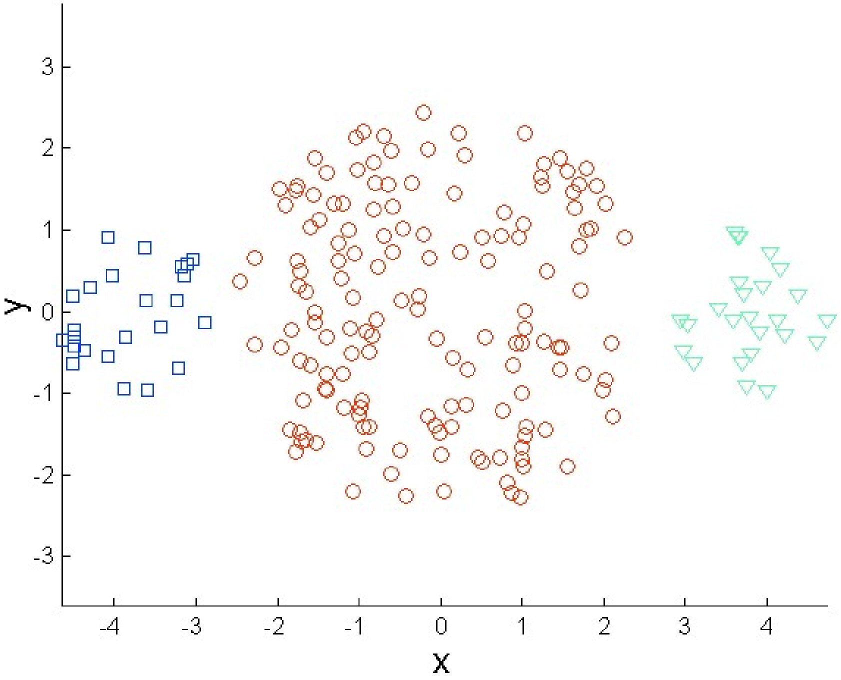


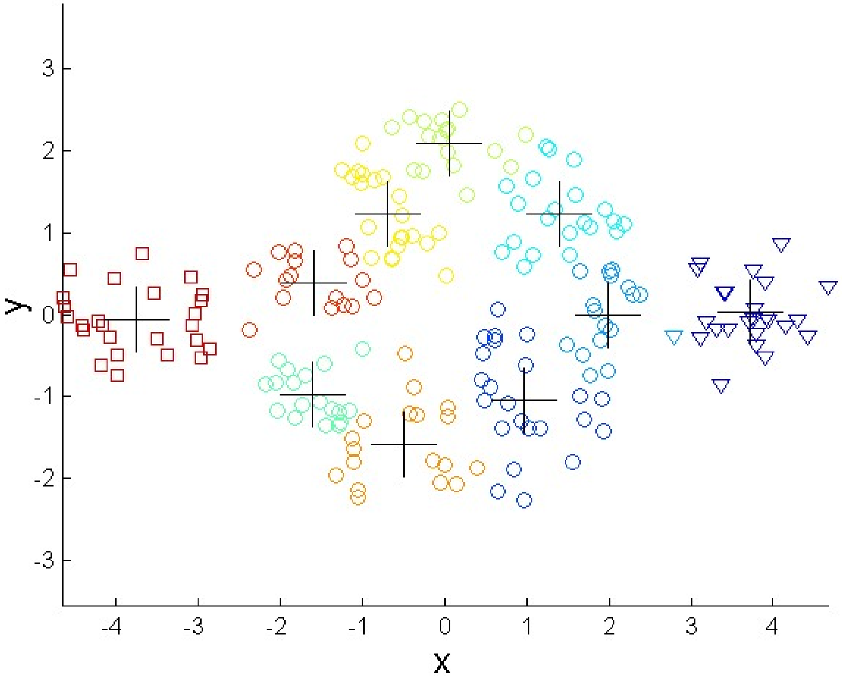
Original Points



K-means (2 Clusters)

Overcoming K-means Limitations





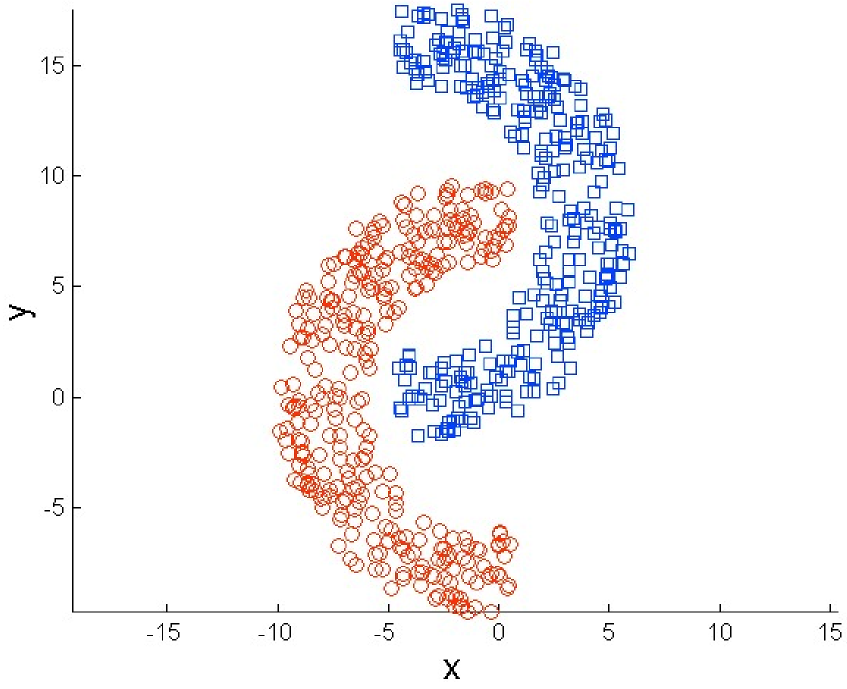
Original Points

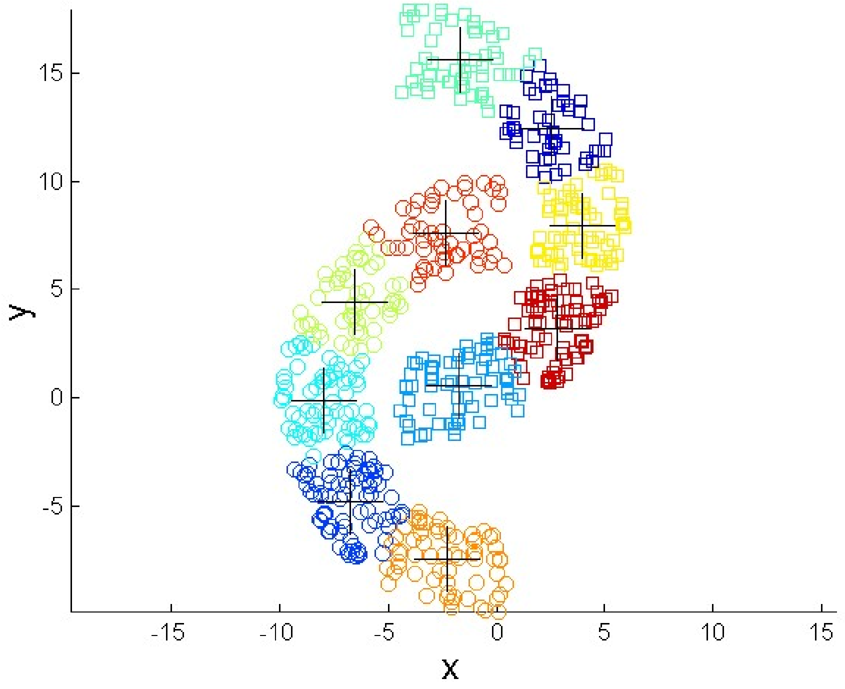
K-means Clusters

One solution is to use many clusters.

Find parts of clusters, but need to put together.

Overcoming K-means Limitations





Original Points

K-means Clusters

Variations of the *K-Means* Method

- A few variants of the *k-means* which differ in
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- Handling categorical data: *k-modes* (Huang'98)
  - Replacing means of clusters with <u>modes</u>
  - Using new dissimilarity measures to deal with categorical objects
  - Using a <u>frequency</u>-based method to update modes of clusters
- Handling a mixture of categorical and numerical data: *k-prototype* method

The *K-Medoids* Clustering Method

- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - *PAM* works effectively for small data sets, but does not scale well for large data sets
- *CLARA* (Kaufmann & Rousseeuw, 1990)
  - draws *multiple samples* of the data set, applies *PAM* on each sample, and gives the best clustering as the output
- *CLARANS* (Ng & Han, 1994): Randomized sampling
- Focusing + spatial data structure (Ester et al., 1995)