**Khoa Khoa Học & Kỹ Thuật Máy Tính**
**Trường Đại Học Bách Khoa Tp. Hồ Chí Minh**

# Chương 5: Gom cụm dữ liệu

Khai phá dữ liệu

(Data mining)



Học kỳ 1 – 2017-2018

# Nội dung

- 5.1. Tổng quan về gom cụm dữ liệu
- 5.2. Gom cụm dữ liệu bằng phân hoạch
- 5.3. Gom cụm dữ liệu bằng phân cấp
- 5.4. Gom cụm dữ liệu dựa trên mật độ
- 5.5. Gom cụm dữ liệu dựa trên mô hình
- 5.6. Các phương pháp gom cụm dữ liệu khác
- 5.7. Tóm tắt

# Tài liệu tham khảo

**[1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers, 2006.**

**7.1** What is cluster analysis? (p 383 -> 386)

**7.4.1** Classical Partitioning Methods: k-Means and k-Medoids (p 402 -> p 407)

**7.5.1** Agglomerative and Divisive Hierarchical Clustering (p 408 -> 411)

**7.6.1** DBSCAN: A Density-Based Clustering Method (p 418 -> 420)

**7.11** Outlier analysis (p 451 -> 458)

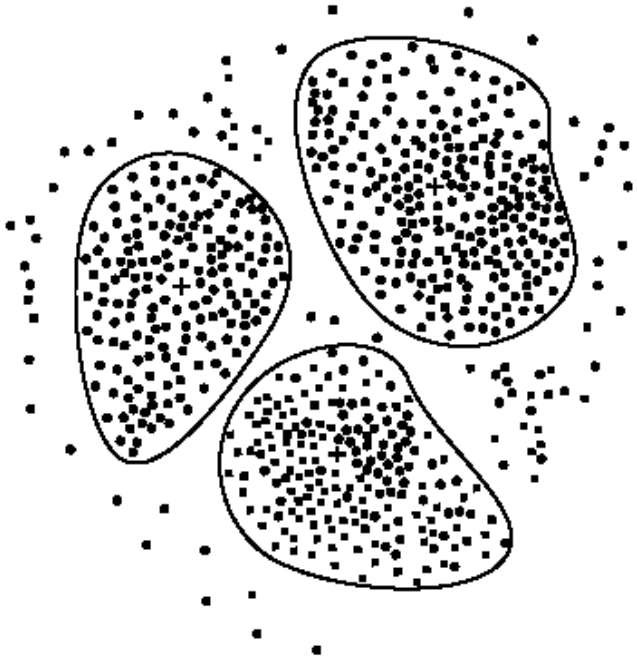
# 5.0. Tình huống 1 – Outlier detection



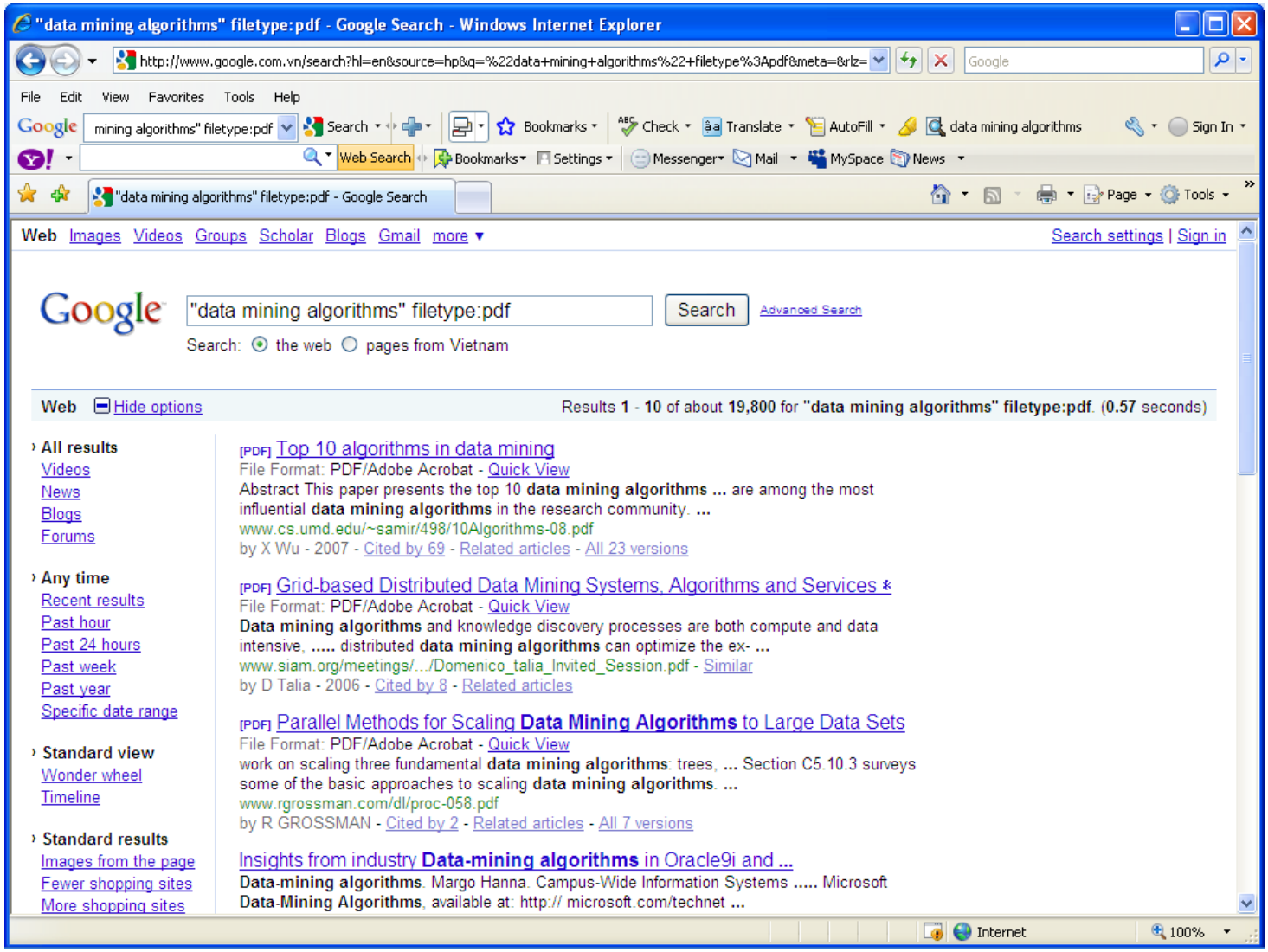

Người đang sử dụng thẻ ID = 1234 thật sự là chủ nhân của thẻ hay là một tên trộm?

# 5.0. Tình huống 2 - Làm sạch dữ liệu

- Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data)
  - Giải pháp giảm thiểu nhiễu
    - Phân tích cụm (cluster analysis)

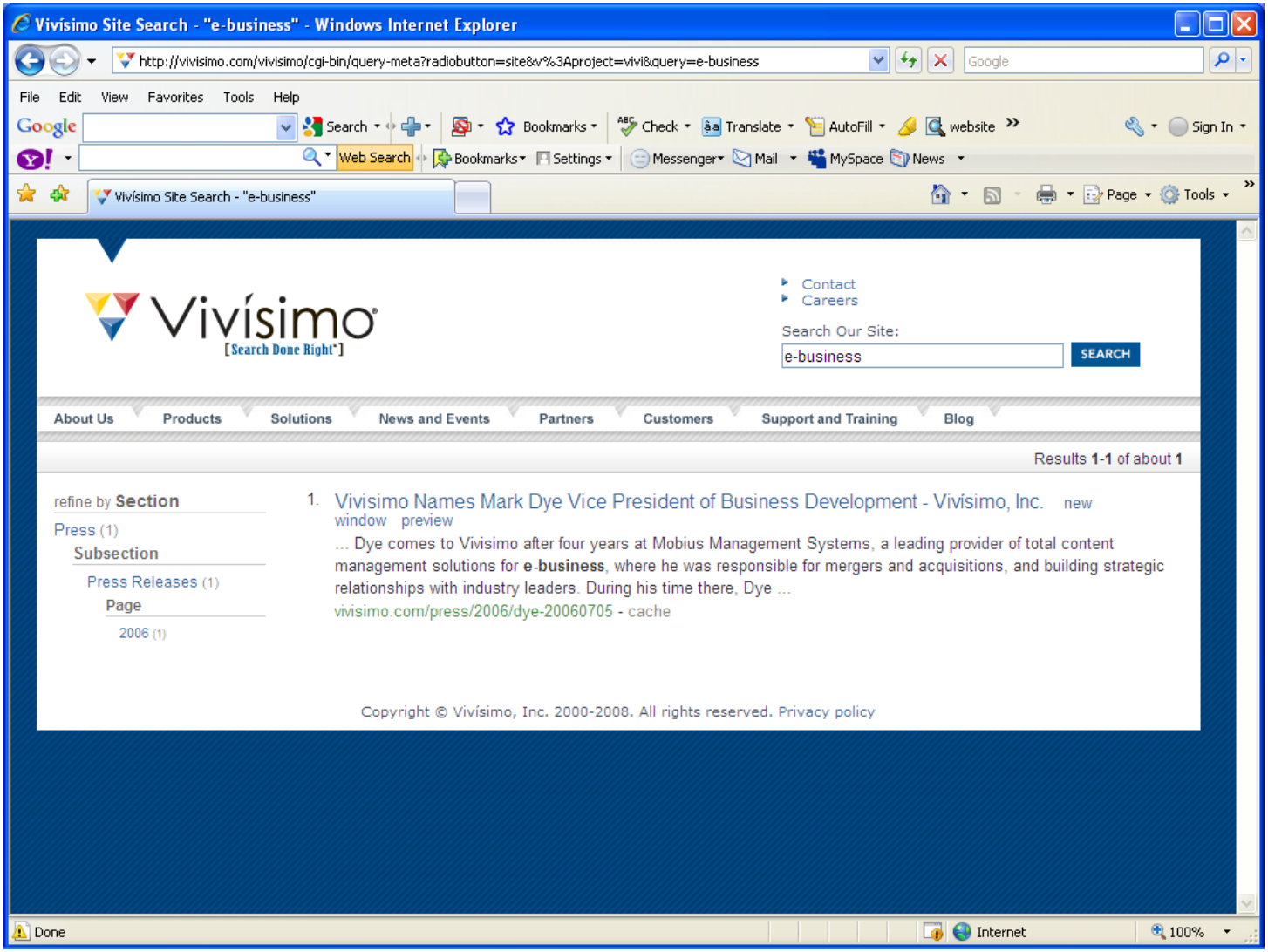


# 5.0. Tình huống 3

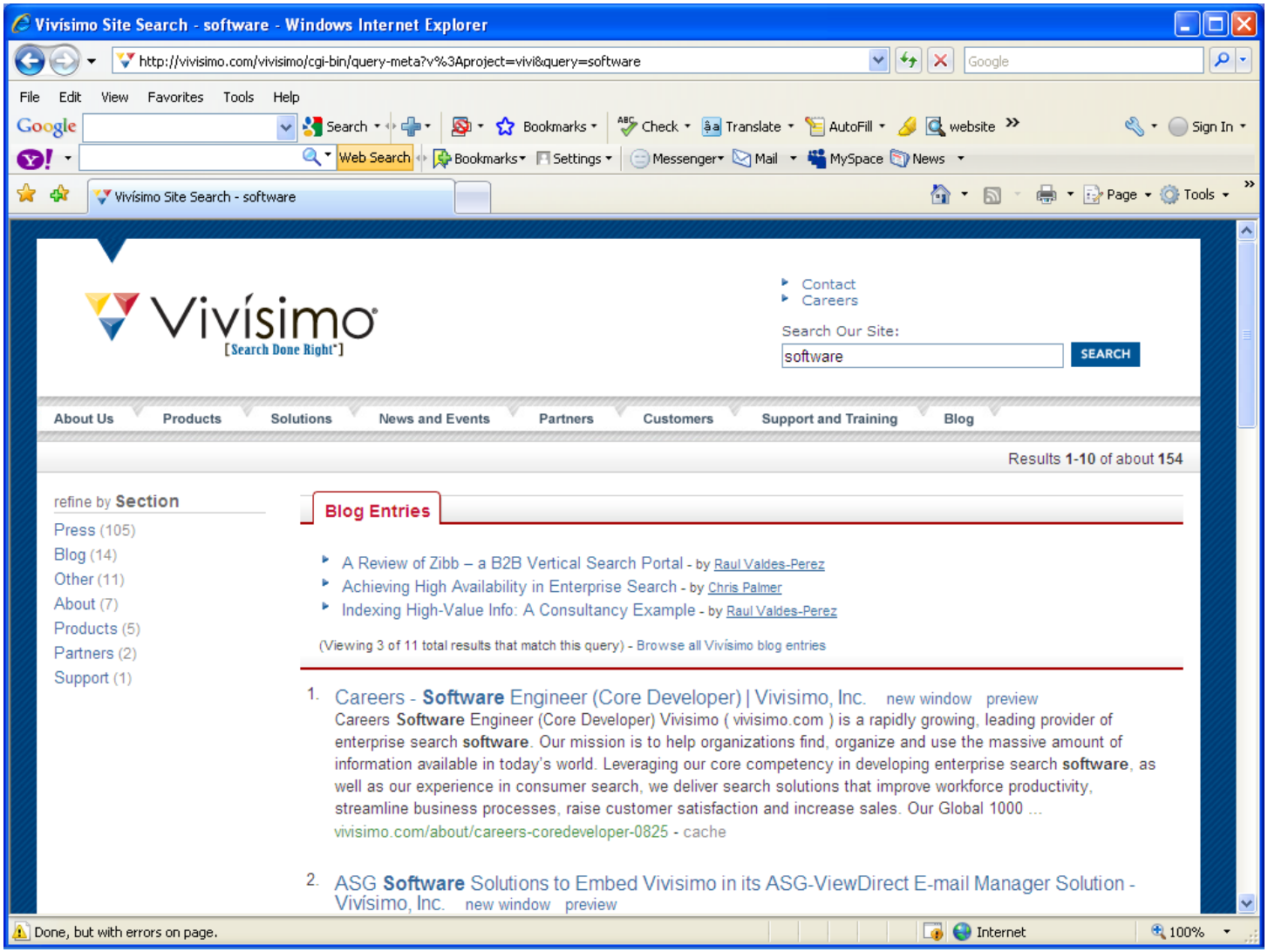


# 5.0. Tình huống 3

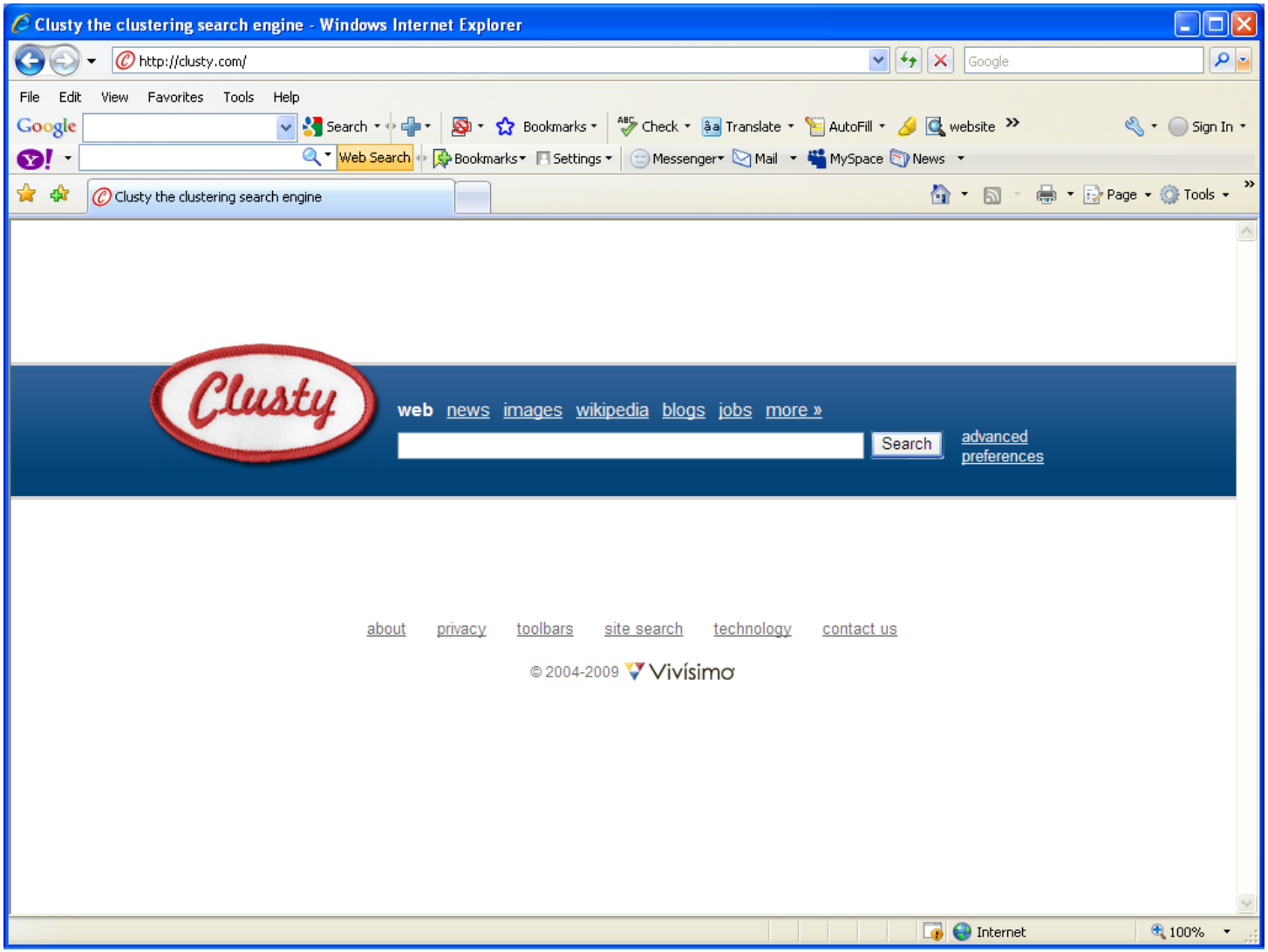
# 5.0. Tình huống 3



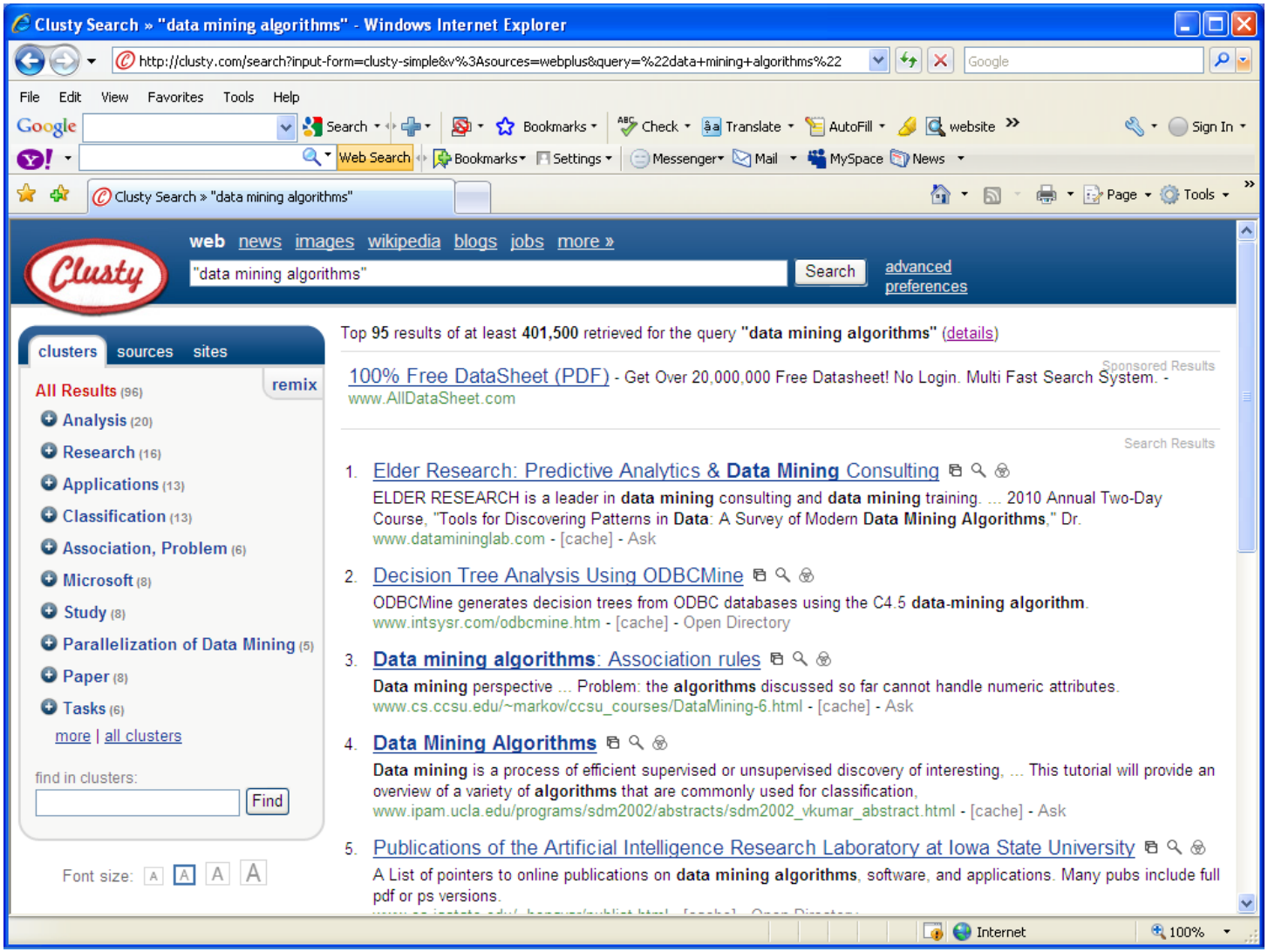
# 5.0. Tình huống 3



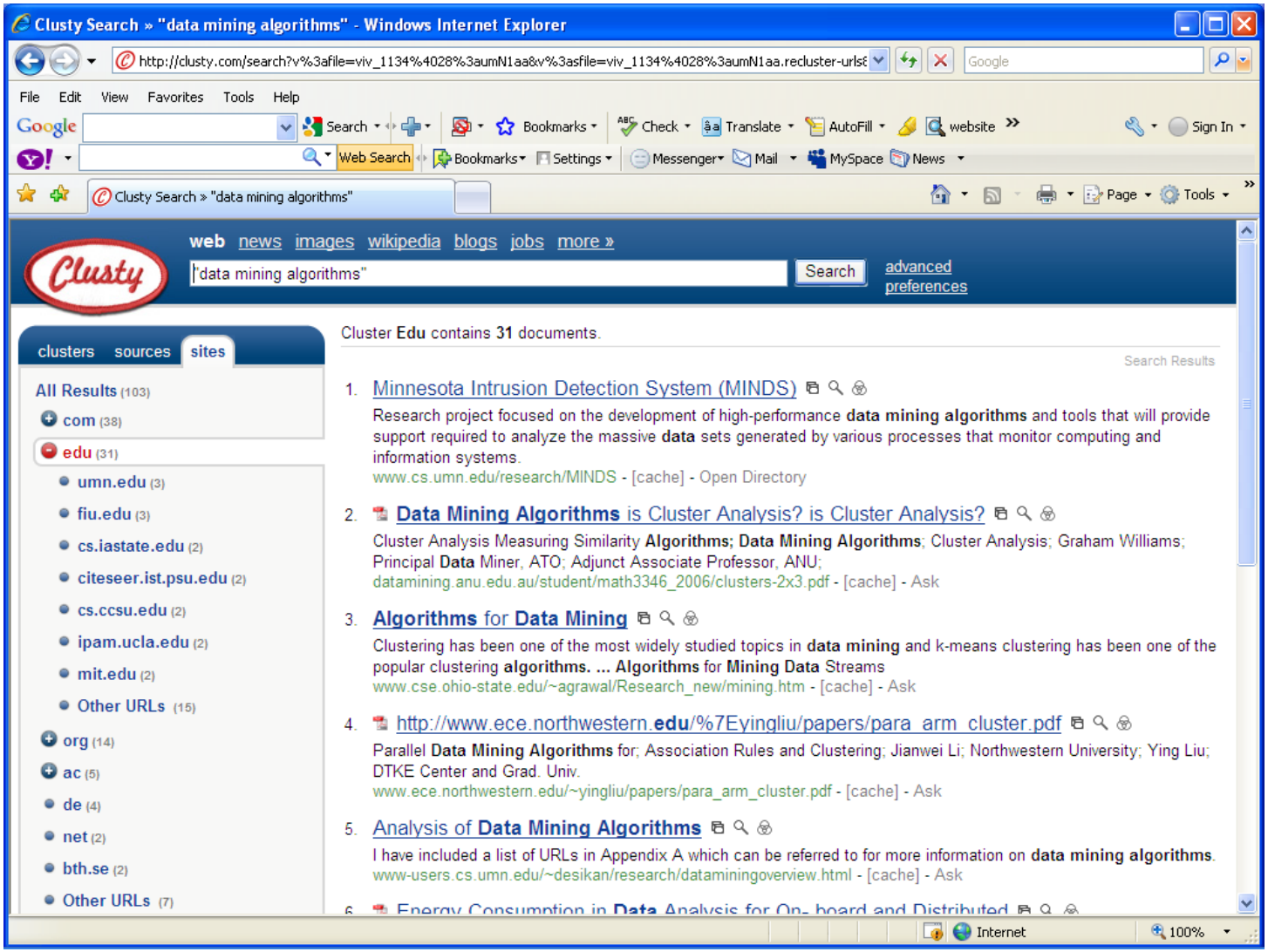
# 5.0. Tình huống 3



# 5.0. Tình huống 3



# 5.0. Tình huống 3

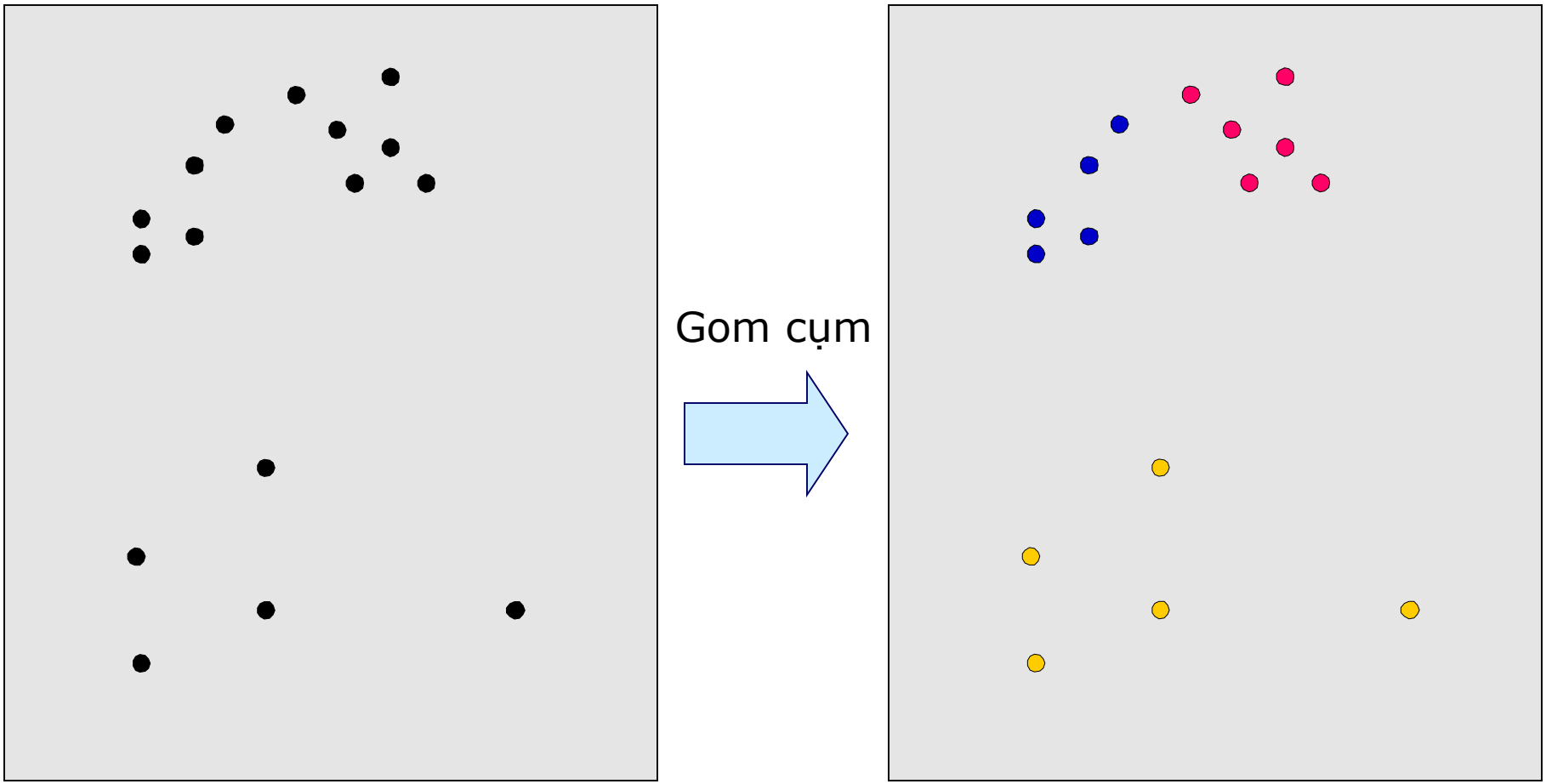


# 5.0. Tình huống 4

http://kdd.ics.uci.edu/databases/CorelFeatures/CorelFeatures.data.html

Gom cụm ảnh

# 5.0. Tình huống …

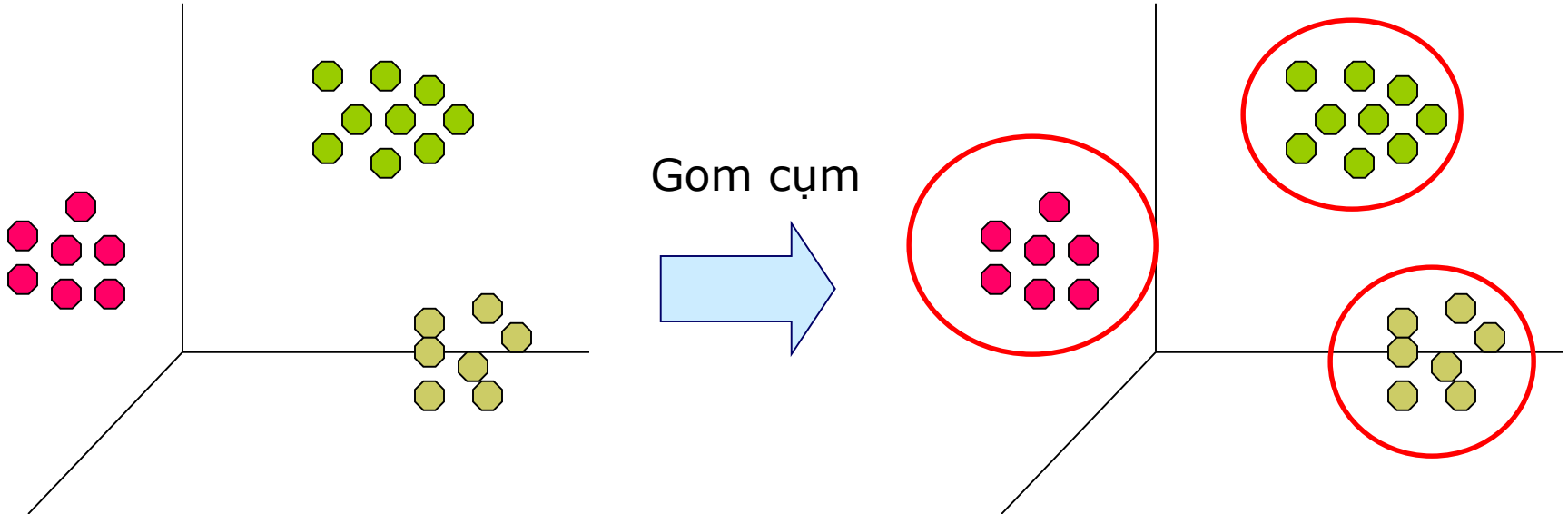


# 5.0. Tình huống …

- Hỗ trợ giai đoạn tiền xử lý dữ liệu (data preprocessing)
- Mô tả sự phân bố dữ liệu/đối tượng (data distribution)
- Nhận dạng mẫu (pattern recognition)
- Phân tích dữ liệu không gian (spatial data analysis)
- Xử lý ảnh (image processing)
- Phân mảnh thị trường (market segmentation)
- Gom cụm tài liệu ((WWW) document clustering)
- …

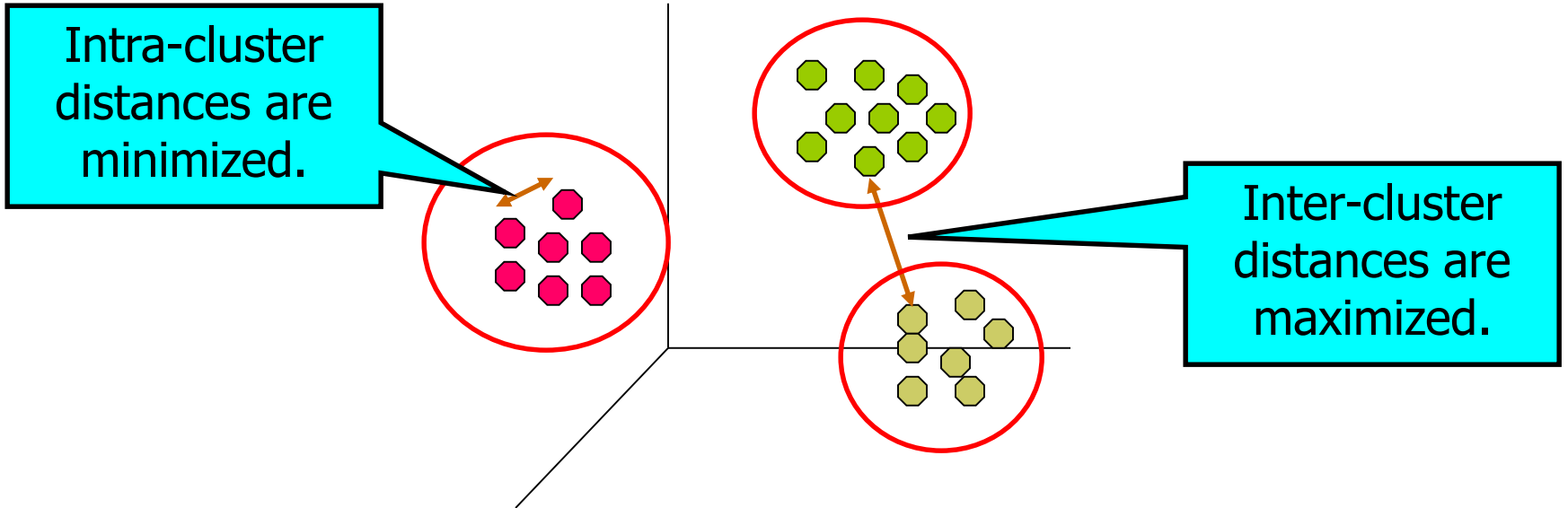
# 5.1. Tổng quan về gom cụm dữ liệu

- Gom cụm
  - Quá trình gom nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm
  - Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.
    - *Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 tương tự Obj2 hơn so với tương tự Obj3.*



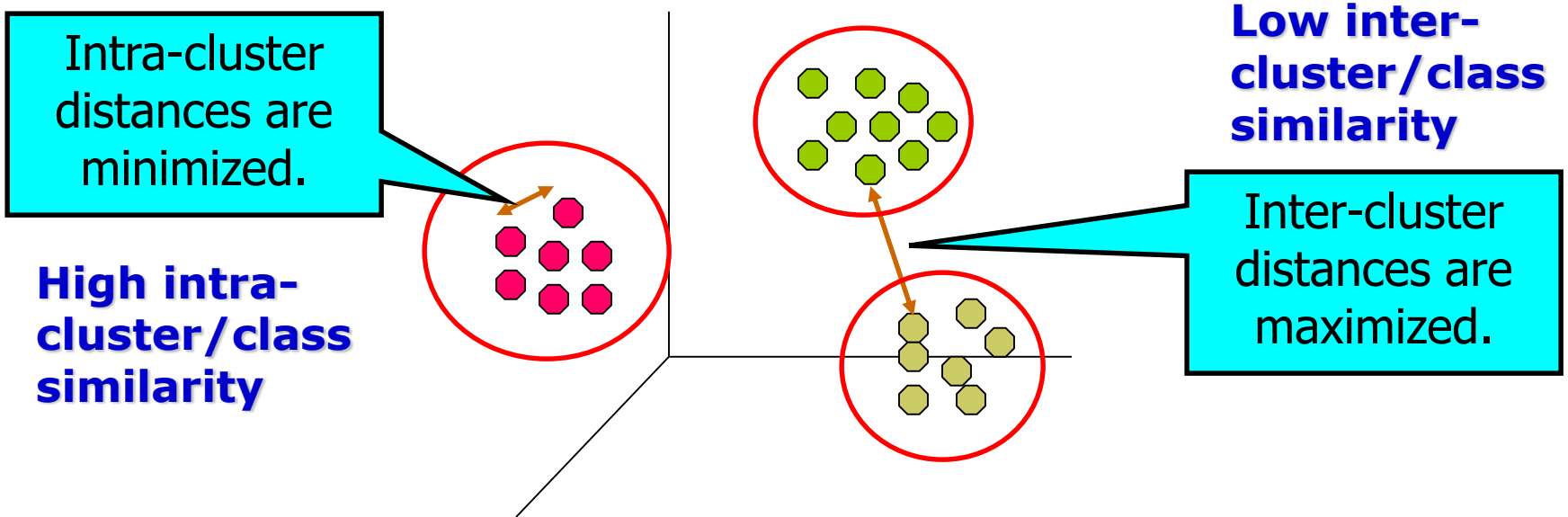
# 5.1. Tổng quan về gom cụm dữ liệu

- Gom cụm
  - Quá trình gom nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm
  - Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.
    - *Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 tương tự Obj2 hơn so với tương tự Obj3.*



# 5.1. Tổng quan về gom cụm dữ liệu

- Gom cụm
  - Quá trình gom nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm
  - Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.
    - *Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 tương tự Obj2 hơn so với tương tự Obj3.*



# 5.1. Tổng quan về gom cụm dữ liệu

- Vấn đề kiểu dữ liệu/đối tượng được gom cụm
  - Ma trận dữ liệu (data matrix)

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

-n đối tượng (objects)

-p biến/thuộc tính (variables/attributes)

# 5.1. Tổng quan về gom cụm dữ liệu

- Vấn đề kiểu dữ liệu/đối tượng được gom cụm
  - Ma trận sai biệt (dissimilarity matrix)

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

d(i, j) là khoảng cách giữa đối tượng i và j; thể hiện sự khác biệt giữa đối tượng i và j; được tính tuỳ thuộc vào kiểu của các biến/thuộc tính.

# 5.1. Tổng quan về gom cụm dữ liệu

- Vấn đề kiểu dữ liệu/đối tượng được gom cụm

d(i, j) là khoảng cách giữa đối tượng i và j; thể hiện sự khác biệt giữa đối tượng i và j; được tính tuỳ thuộc vào kiểu của các biến/thuộc tính.

$$d(i,j) \geq 0$$

$$d(i,i) = 0$$

$$d(i,j) = d(j,i)$$

$$d(i,j) \leq d(i,k) + d(k,j)$$

# 5.1. Tổng quan về gom cụm dữ liệu

- Vấn đề kiểu dữ liệu/đối tượng được gom cụm
  - Đối tượng vector (vector objects)
    - Đối tượng i và j được biểu diễn tương ứng bởi vector x và y.
    - Độ tương tự (similarity) giữa i và j được tính bởi độ đo cosine:

$$s(x, y) = \frac{x^t \cdot y}{||x|| ||y||}$$

x = (x1, …, xp)

y = (y1, …, yp)

s(x, y) = (x1*y1 + … + xp*yp)/((x1$^2$ + … + xp$^2$)$^{1/2}$*(y1$^2$+ … + yp$^2$)$^{1/2}$)

# 5.1. Tổng quan về gom cụm dữ liệu

- Vấn đề kiểu dữ liệu/đối tượng được gom cụm
  - Interval-scaled variables/attributes
  - Binary variables/attributes
  - Categorical variables/attributes
  - Ordinal variables/attributes
  - Ratio-scaled variables/attributes
  - Variables/attributes of mixed types

- A sample data table containing variables of mixed type.

| *object identifier* | *test-1 (categorical)* | *test-2 (ordinal)* | *test-3 (ratio-scaled)* |
|---|---|---|---|
| 1 | code-A | excellent | 445 |
| 2 | code-B | fair | 22 |
| 3 | code-C | good | 164 |
| 4 | code-A | excellent | 1,210 |

# 5.1. Tổng quan về gom cụm dữ liệu

- Interval-scaled variables/attributes

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

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

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

# 5.1. Tổng quan về gom cụm dữ liệu

- Độ đo khoảng cách Minkowski

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

- Độ đo khoảng cách Manhattan

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

- Độ đo khoảng cách Euclidean

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

# 5.1. Tổng quan về gom cụm dữ liệu

- Binary variables/attributes

| | | Object $j$ | | |
|---|---|---|---|---|
| | | 1 | 0 | *sum* |
| Object $i$ | 1 | $a$ | $b$ | $a+b$ |
| | 0 | $c$ | $d$ | $c+d$ |
| | *sum* | $a+c$ | $b+d$ | $p$ (= a + b + c + d) |

Hệ số so trùng đơn giản (nếu symmetric): $d(i,j) = \frac{b+c}{a+b+c+d}$

Hệ số so trùng Jaccard (nếu asymmetric): $d(i,j) = \frac{b+c}{a+b+c}$

# 5.1. Tổng quan về gom cụm dữ liệu

- Binary variables/attributes
  - Ví dụ

| 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: symmetric
  - Binary attributes còn lại: asymmetric
  - Y, P → 1, N → 0

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

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

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

# 5.1. Tổng quan về gom cụm dữ liệu

- Variables/attributes of mixed types
  - Tổng quát

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

    - Nếu $x_{if}$ hoặc $x_{jf}$ bị thiếu (missing) thì $\delta_{ij}^{(f)} = 0$
    - *f (variable/attribute)*: binary (nominal)

      $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$ , or $d_{ij}^{(f)} = 1$ otherwise
    - *f* : interval-scaled (Minkowski, Manhattan, Euclidean)
    - *f* : ordinal or ratio-scaled
      - tính ranks $r_{if}$ và $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
      - $z_{if}$ trở thành interval-scaled

# 5.1. Tổng quan về gom cụm dữ liệu

| Measures | Forms | Comments | Examples and Applications |
|---|---|---|---|
| Minkowski distance | $D_{ij} = \left( \sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert^{1/n} \right)^{n}$ | Metric. Invariant to any translation and rotation only for $n = 2$ (Euclidean distance). Features with large values and variances tend to dominate over other features. | Fuzzy $c$-means with measures based on Minkowski family [130]. |
| Euclidean distance | $D_{ij} = \left( \sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert^{1/2} \right)^{2}$ | The most commonly used metric. Special case of Minkowski metric at $n = 2$. Tend to form hyperspherical clusters. | $K$-means algorithm [191] |
| City-block distance | $D_{ij} = \sum_{l=1}^{d} \lvert x_{il} - x_{jl} \rvert$ | Special case of Minkowski metric at $n = 1$. Tend to form hyperrecutangular clusters. | Fuzzy ART [57] |
| Sup distance | $D_{ij} = \max_{1 \le l \le d} \lvert x_{il} - x_{jl} \rvert$ | Special case of Minkowski metric at $n \to \infty$. | Fuzzy $c$-means with sup norm [39]. |
| Mahalanobis distance | $D_{ij} = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j)$, where $\mathbf{S}$ is the within-group covariance matrix. | Invariant to any nonsingular linear transformation. $\mathbf{S}$ is calculated based on all objects. Tend to form hyperellipsoidal clusters. When features are not correlated, squared Mahalanobis distance is equivalent to squared Euclidean distance. May cause some computational burden. | Ellipsoidal ART [13], Hyperellipsoidal clustering algorithm [194]. |
| Pearson correlation | $D_{ij} = (1 - r_{ij})/2$, where $r_{ij} = \frac{\sum_{l=1}^{d} (x_{il} - \overline{x_i})(x_{jl} - \overline{x_j})}{\sqrt{\sum_{l=1}^{d} (x_{il} - \overline{x_i})^2 \sum_{l=1}^{d} (x_{jl} - \overline{x_j})^2}}$ | Not a metric. Derived from correlation coefficient. Unable to detect the magnitude of differences of two variables. | Widely used as the measure for analyzing gene expression data [80]. |
| Point symmetry distance | $D_{ir} = \min_{\substack{j=1,\ldots,N \\ j \ne i}} \frac{\lVert (\mathbf{x}_i - \mathbf{x}_r) + (\mathbf{x}_j - \mathbf{x}_r) \rVert}{\lVert (\mathbf{x}_i - \mathbf{x}_r) \rVert + \lVert (\mathbf{x}_j - \mathbf{x}_r) \rVert}$ | Not a metric. Compute the distance between an object $\mathbf{x}_i$ and a reference point $\mathbf{x}_r$. $D_{ir}$ is minimized when a symmetric pattern exists. | SBKM (Symmetry-based $K$-means) [264]. |
| Cosine similarity | $S_{ij} = \cos\alpha = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\lVert \mathbf{x}_i \rVert \lVert \mathbf{x}_j \rVert}$ | Independent of vector length. Invariant to rotation, but not to linear transformations. | The most commonly used measure in document clustering [261]. |

# Question

- Briefly outline how to compute the *dissimilarity* between objects described by the following types of variables:

(a) Numerical (interval-scaled) variables

(b) Asymmetric binary variables

(c) Categorical variables

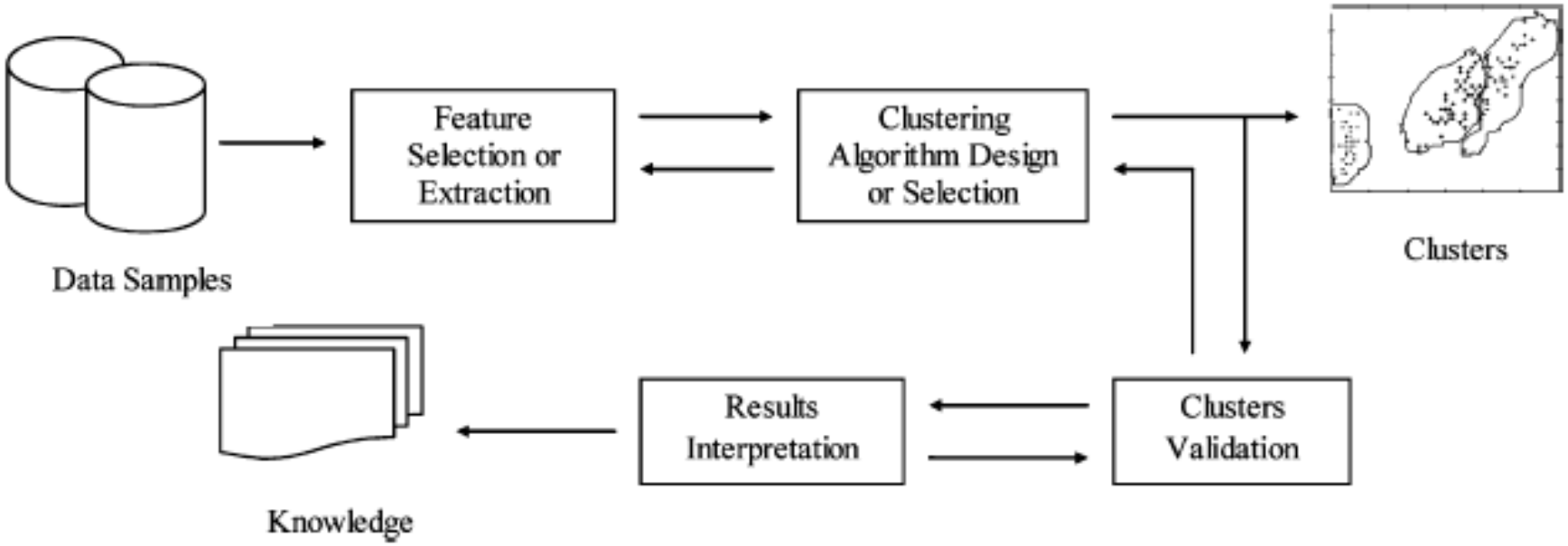
(d) Ratio-scaled variables

(e) Ordinal variables

| *object identifier* | *test-1 (categorical)* | *test-2 (ordinal)* | *test-3 (ratio-scaled)* |
|---|---|---|---|
| 1 | code-A | excellent | 445 |
| 2 | code-B | fair | 22 |
| 3 | code-C | good | 164 |
| 4 | code-A | excellent | 1,210 |

# 5.1. Tổng quan về gom cụm dữ liệu

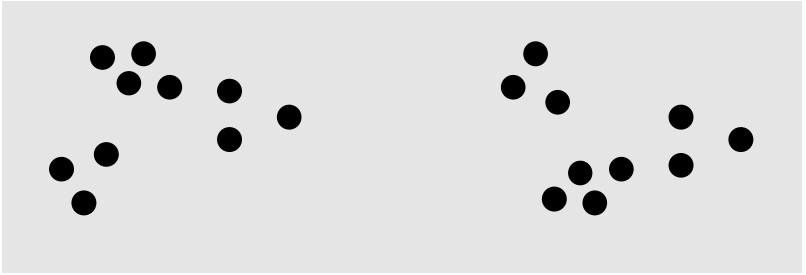
- Quá trình gom cụm dữ liệu



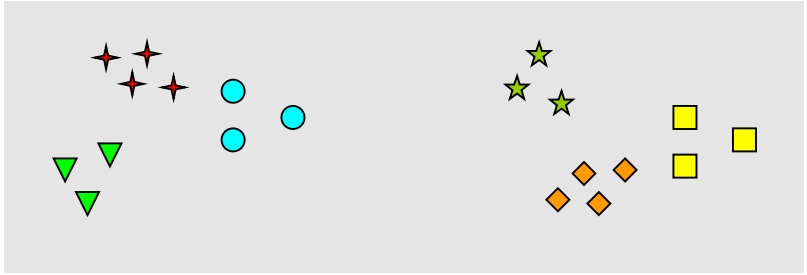

R. Xu, D. Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Networks, 16(3), May 2005, pp. 645-678.

# 5.1. Tổng quan về gom cụm dữ liệu

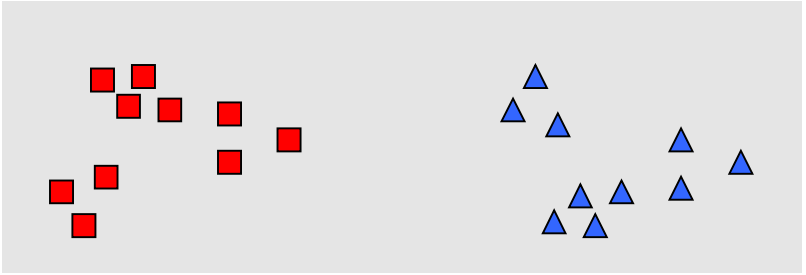
- Mỗi cụm nên có bao nhiêu phần tử?
- Các phân tử nên được gom vào bao nhiêu cụm?
- Bao nhiêu cụm nên được tạo ra?



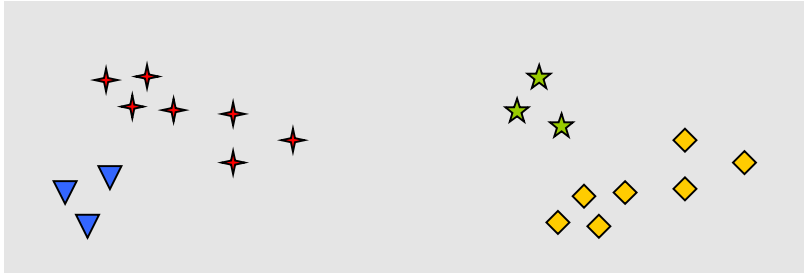
Bao nhiêu cụm?



6 cụm?



2 cụm?



4 cụm?

# 5.1. Tổng quan về gom cụm dữ liệu

- Các yêu cầu tiêu biểu về việc gom cụm dữ liệu
  - Khả năng co giãn về tập dữ liệu (scalability)
  - Khả năng xử lý nhiều kiểu thuộc tính khác nhau (different types of attributes)
  - Khả năng khám phá các cụm với hình dạng tùy ý (clusters with arbitrary shape)
  - Tối thiểu hóa yêu cầu về tri thức miền trong việc xác định các thông số nhập (domain knowledge for input parameters)
  - Khả năng xử lý dữ liệu có nhiễu (noisy data)

# 5.1. Tổng quan về gom cụm dữ liệu

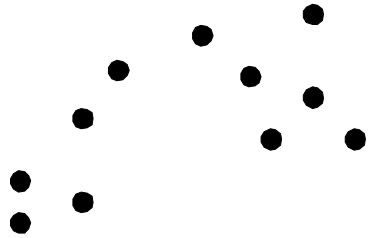
- Các yêu cầu tiêu biểu về việc gom cụm dữ liệu
  - Khả năng gom cụm tăng dần và độc lập với thứ tự của dữ liệu nhập (incremental clustering and insensitivity to the order of input records)
  - Khả năng xử lý dữ liệu đa chiều (high dimensionality)
  - Khả năng gom cụm dựa trên ràng buộc (constraint-based clustering)
  - Khả diễn và khả dụng (interpretability and usability)

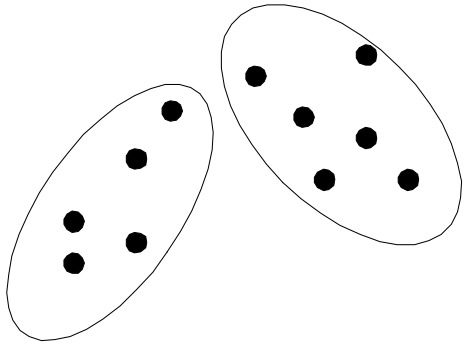
# 5.1. Tổng quan về gom cụm dữ liệu

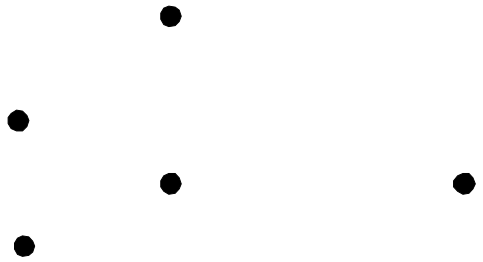
- Phân loại các phương pháp gom cụm dữ liệu tiêu biểu
  - Phân hoạch (partitioning): các phân hoạch được tạo ra và đánh giá theo một tiêu chí nào đó.
  - Phân cấp (hierarchical): phân rã tập dữ liệu/đối tượng có thứ tự phân cấp theo một tiêu chí nào đó.
  - Dựa trên mật độ (density-based): dựa trên connectivity and density functions.
  - Dựa trên lưới (grid-based): dựa trên a multiple-level granularity structure.
  - Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất.
  - …

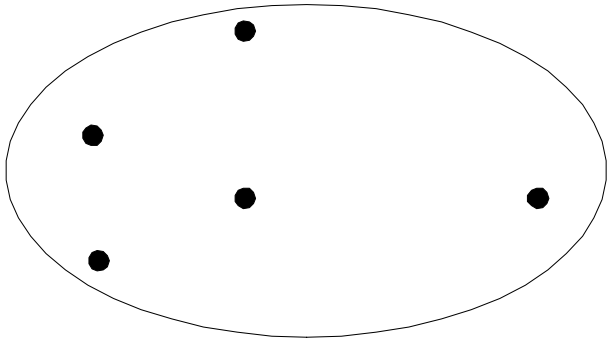
# 5.1. Tổng quan về gom cụm dữ liệu

- Phân loại các phương pháp gom cụm dữ liệu tiêu biểu







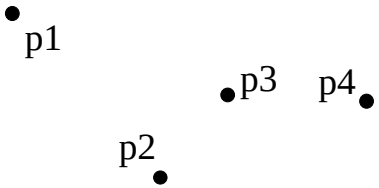


**Original Points**

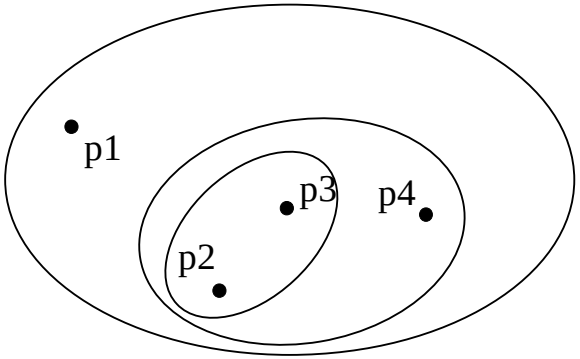
**Partitioning**

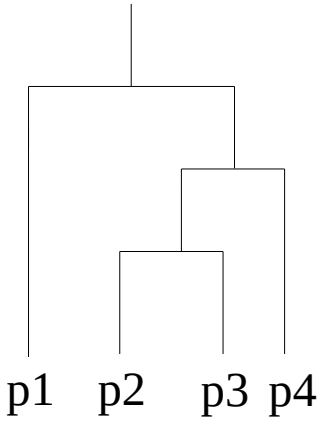
# 5.1. Tổng quan về gom cụm dữ liệu

- Phân loại các phương pháp gom cụm dữ liệu tiêu biểu




**Original Points**




**Hierarchical**

# 5.1. Tổng quan về gom cụm dữ liệu

- Các phương pháp đánh giá việc gom cụm dữ liệu
  - Đánh giá ngoại (external validation)
    - Đánh giá kết quả gom cụm dựa vào cấu trúc được chỉ định trước cho tập dữ liệu
  - Đánh giá nội (internal validation)
    - Đánh giá kết quả gom cụm theo số lượng các vector của chính tập dữ liệu (ma trận gần – proximity matrix)
  - Đánh giá tương đối (relative validation)
    - Đánh giá kết quả gom cụm bằng việc so sánh các kết quả gom cụm khác ứng với các bộ trị thông số khác nhau

→ Tiêu chí cho việc đánh giá và chọn kết quả gom cụm tối ưu

- Độ nén (compactness): các đối tượng trong cụm nên gần nhau.
- Độ phân tách (separation): các cụm nên xa nhau.

# 5.1. Tổng quan về gom cụm dữ liệu

- Các phương pháp đánh giá việc gom cụm dữ liệu
  - Đánh giá ngoại (external validation)
    - Độ đo: Rand statistic, Jaccard coefficient, Folkes and Mallows index, …
  - Đánh giá nội (internal validation)
    - Độ đo: Hubert's $\Gamma$ statistic, Silhouette index, Dunn's index, …
  - Đánh giá tương đối (relative validation)

# 5.1. Tổng quan về gom cụm dữ liệu

- Các phương pháp đánh giá việc gom cụm dữ liệu
  - Các độ đo đánh giá ngoại (external validation measures – contingency matrix)

| | Measure | Notation | Definition | Range |
|---|---|---|---|---|
| 1 | Entropy | $E$ | $-\sum_i p_i(\sum_j \frac{p_{ij}}{p_i}\log\frac{p_{ij}}{p_i})$ | $[0, \log K']$ |
| 2 | Purity | $P$ | $\sum_i p_i(\max_j \frac{p_{ij}}{p_j})$ | (0,1] |
| 3 | F-measure | $F$ | $\sum_j p_j \max_i[2\frac{p_{ij}}{p_i}\frac{p_{ij}}{p_j}/(\frac{p_{ij}}{p_i}+\frac{p_{ij}}{p_j})]$ | (0,1] |
| 4 | Variation of Information | $VI$ | $-\sum_i p_i \log p_i - \sum_j p_j \log p_j - 2\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_i p_j}$ | $[0, 2\log\max(K, K')]$ |
| 5 | Mutual Information | $MI$ | $\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{p_i p_j}$ | $(0, \log K']$ |
| 6 | Rand statistic | $R$ | $[\binom{n}{2} - \sum_i\binom{n_{i\cdot}}{2} - \sum_j\binom{n_{\cdot j}}{2} + 2\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$ | (0,1] |
| 7 | Jaccard coefficient | $J$ | $\sum_{ij}\binom{n_{ij}}{2}/[\sum_i\binom{n_{i\cdot}}{2} + \sum_j\binom{n_{\cdot j}}{2} - \sum_{ij}\binom{n_{ij}}{2}]$ | [0,1] |
| 8 | Fowlkes and Mallows index | $FM$ | $\sum_{ij}\binom{n_{ij}}{2}/\sqrt{\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}}$ | [0,1] |
| 9 | Hubert $\Gamma$ statistic I | $\Gamma$ | $\frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}}{\sqrt{\sum_i\binom{n_{i\cdot}}{2}\sum_j\binom{n_{\cdot j}}{2}[\binom{n}{2}-\sum_i\binom{n_{i\cdot}}{2}][\binom{n}{2}-\sum_j\binom{n_{\cdot j}}{2}]}}$ | (-1,1] |
| 10 | Hubert $\Gamma$ statistic II | $\Gamma'$ | $[\binom{n}{2} - 2\sum_i\binom{n_{i\cdot}}{2} - 2\sum_j\binom{n_{\cdot j}}{2} + 4\sum_{ij}\binom{n_{ij}}{2}]/\binom{n}{2}$ | [0,1] |
| 11 | Minkowski score | $MS$ | $\sqrt{\sum_i\binom{n_{i\cdot}}{2} + \sum_j\binom{n_{\cdot j}}{2} - 2\sum_{ij}\binom{n_{ij}}{2}}/\sqrt{\sum_j\binom{n_{\cdot j}}{2}}$ | $[0, +\infty)$ |
| 12 | classification error | $\varepsilon$ | $1 - \frac{1}{n}\max_\sigma\sum_j n_{\sigma(j),j}$ | [0,1) |
| 13 | van Dongen criterion | $VD$ | $(2n - \sum_i\max_j n_{ij} - \sum_j\max_i n_{ij})/2n$ | [0, 1) |
| 14 | micro-average precision | $MAP$ | $\sum_i p_i(\max_j \frac{p_{ij}}{p_i})$ | (0,1] |
| 15 | Goodman-Kruskal coefficient | $GK$ | $\sum_i p_i(1 - \max_j \frac{p_{ij}}{p_i})$ | [0,1) |
| 16 | Mirkin metric | $M$ | $\sum_i n_{i\cdot}^2 + \sum_j n_{\cdot j}^2 - 2\sum_i\sum_j n_{ij}^2$ | $[0, 2\binom{n}{2})$ |

Note: $p_{ij} = n_{ij}/n,\ p_i = n_{i\cdot}/n,\ p_j = n_{\cdot j}/n.$



J. Wu and et al. Adapting the Right Measures for K-means Clustering. KDD'09, Paris, France, July 2009.

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đánh giá kết quả gom cụm

Partition C

| Partition P | $C_1$ | $C_2$ | $\cdots$ | $C_{K'}$ | $\sum$ |
|---|---|---|---|---|---|
| $P_1$ | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1K'}$ | $n_{1\cdot}$ |
| $P_2$ | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2K'}$ | $n_{2\cdot}$ |
| $\cdot$ | $\cdot$ | $\cdot$ | $\cdots$ | $\cdot$ | $\cdot$ |
| $P_K$ | $n_{K1}$ | $n_{K2}$ | $\cdots$ | $n_{KK'}$ | $n_{K\cdot}$ |
| $\sum$ | $n_{\cdot 1}$ | $n_{\cdot 2}$ | $\cdots$ | $n_{\cdot K'}$ | $n$ |

Contingency matrix

-Partition P: kết quả gom cụm trên n đối tượng

-Partition C: các cụm thật sự của n đối tượng

-$n_{ij} = |P_i \cap C_j|$: số đối tượng trong $P_i$ từ $C_j$

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đánh giá kết quả gom cụm

| I | $C_1$ | $C_2$ | $C_3$ | $\sum$ |
|---|---|---|---|---|
| $P_1$ | 3 | 4 | 12 | 19 |
| $P_2$ | 8 | 3 | 12 | 23 |
| $P_3$ | 12 | 12 | 0 | 24 |
| $\sum$ | 23 | 19 | 24 | 66 |

| II | $C_1$ | $C_2$ | $C_3$ | $\sum$ |
|---|---|---|---|---|
| $P_1$ | 0 | 7 | 12 | 19 |
| $P_2$ | 11 | 0 | 12 | 23 |
| $P_3$ | 12 | 12 | 0 | 24 |
| $\sum$ | 23 | 19 | 24 | 66 |

Kết quả gom cụm theo phương án I và II

-Partition P: kết quả gom cụm trên n (=66) đối tượng

-Partition C: các cụm thật sự của n (=66) đối tượng

-$n_{ij} = |P_i \cap C_j|$: số đối tượng trong $P_i$ từ $C_j$

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đánh giá kết quả gom cụm
  - Entropy (trị nhỏ khi chất lượng gom cụm tốt)

$$Entropy(I) = -\sum_i p_i (\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$$

$$= -\sum_i \frac{n_i}{n} (\sum_j \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i})$$

$$= -\frac{19}{66}(\frac{3}{19}\log\frac{3}{19} + \frac{4}{19}\log\frac{4}{19} + \frac{12}{19}\log\frac{12}{19})$$

$$-\frac{23}{66}(\frac{8}{23}\log\frac{8}{23} + \frac{3}{23}\log\frac{3}{23} + \frac{12}{23}\log\frac{12}{23})$$

$$-\frac{24}{66}(\frac{12}{24}\log\frac{12}{24} + \frac{12}{24}\log\frac{12}{24} + \frac{0}{24}\log\frac{0}{24})$$

$$= ???$$

$$Entropy(II) = -\sum_i p_i (\sum_j \frac{p_{ij}}{p_i} \log \frac{p_{ij}}{p_i})$$

$$= -\sum_i \frac{n_i}{n} (\sum_j \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i})$$

$$= -\frac{19}{66}(\frac{0}{19}\log\frac{0}{19} + \frac{7}{19}\log\frac{7}{19} + \frac{12}{19}\log\frac{12}{19})$$

$$-\frac{23}{66}(\frac{11}{23}\log\frac{11}{23} + \frac{0}{23}\log\frac{0}{23} + \frac{12}{23}\log\frac{12}{23})$$

$$-\frac{24}{66}(\frac{12}{24}\log\frac{12}{24} + \frac{12}{24}\log\frac{12}{24} + \frac{0}{24}\log\frac{0}{24})$$

$$= ???$$

→ Gom cụm theo phương án I hay phương án II tốt???

# 5.2. Gom cụm dữ liệu bằng phân hoạch

**Algorithm: *k*-means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

**Input:**

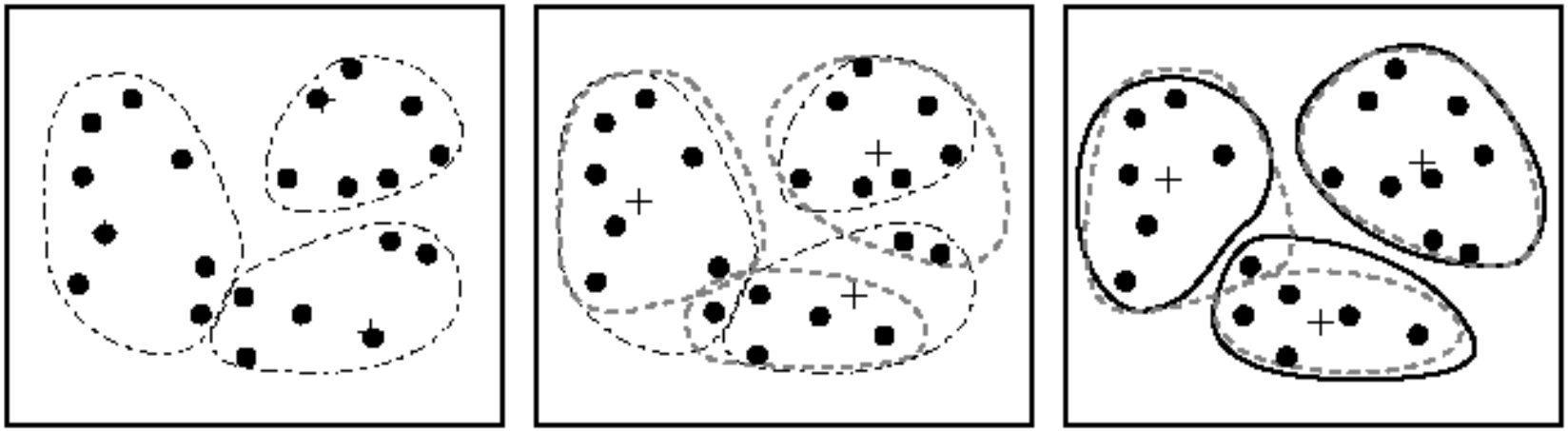
- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

**Output:** A set of *k* clusters.

**Method:**

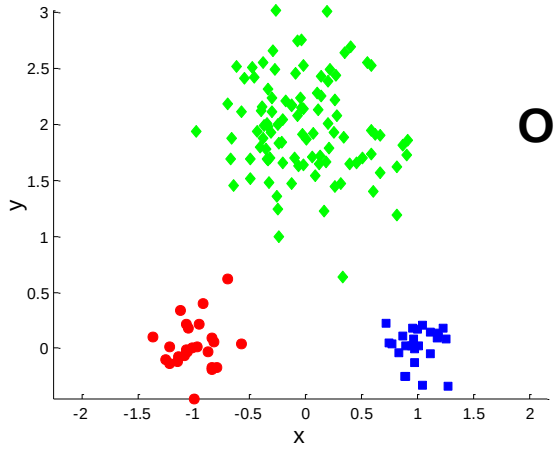
(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) **until** no change;

# 5.2. Gom cụm dữ liệu bằng phân hoạch

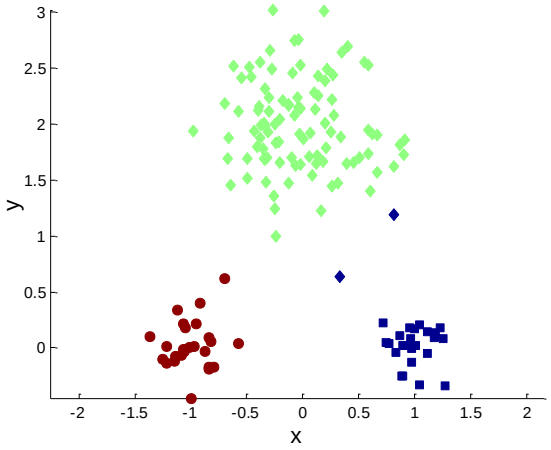



Clustering of a set of objects based on the *k*-means method. (The mean of each cluster is marked by a "+".)

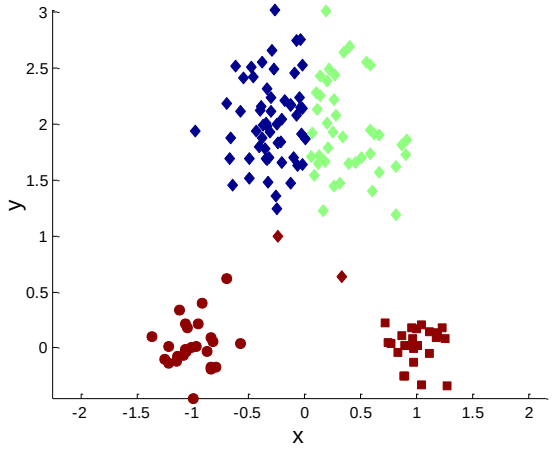
# 5.2. Gom cụm dữ liệu bằng phân hoạch




**Original Points**




**Optimal Clustering**




**Sub-optimal Clustering**

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đặc điểm của giải thuật k-means?

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đặc điểm của giải thuật k-means
  - Bài toán tối ưu hóa
    - Cực trị cục bộ
  - Mỗi cụm được đặc trưng hóa bởi trung tâm của cụm (i.e. đối tượng trung bình (mean)).
    - Không thể xác định được đối tượng trung bình???
    - Số cụm k nên là bao nhiêu?
  - Độ phức tạp: O(nkt)
    - n là số đối tượng, k là số cụm, t là số lần lặp
      - k << n, t << n

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đặc điểm của giải thuật k-means
  - Ảnh hưởng bởi nhiễu (các phần tử kì dị/biên)
  - Không phù hợp cho việc khai phá ra các cụm có dạng không lồi (nonconvex) hay các cụm có kích thước rất khác nhau
    - Kết quả gom cụm có dạng siêu cầu (hyperspherial)
    - Kích thước các cụm kết quả thường đồng đều (relatively uniform sizes)

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đặc điểm của giải thuật k-means
  - Đánh giá kết quả gom cụm của giải thuật k-means với hai trị k1 (phương án I) và k2 (phương án II) khác nhau trên cùng tập dữ liệu mẫu cho trước
    - Entropy (trị nhỏ khi chất lượng gom cụm tốt)
      - Entropy (I) = ???
      - Entropy (II) = ???

      → Gom cụm theo phương án I hay phương án II tốt?

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đặc điểm của giải thuật k-means
  - Đánh giá kết quả gom cụm của giải thuật k-means với hai trị k1 (phương án I) và k2 (phương án II) khác nhau trên cùng tập dữ liệu mẫu cho trước
    - F-measure (trị lớn khi chất lượng gom cụm tốt)
      - F-measure (I) = ???
      - F-measure (II) = ???

      → Gom cụm theo phương án I hay phương án II tốt?

      → Kết quả đánh giá trùng với kết quả đánh giá dựa trên độ đo Entropy?

# 5.2. Gom cụm dữ liệu bằng phân hoạch

**Algorithm: *k*-medoids.** PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

**Input:**

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

**Output:** A set of *k* clusters.

**Method:**

(1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, $o_{random}$;
(5) compute the total cost, *S*, of swapping representative object, $o_j$, with $o_{random}$;
(6) **if** $S < 0$ **then** swap $o_j$ with $o_{random}$ to form the new set of *k* representative objects;
(7) **until** no change;

# Exercise

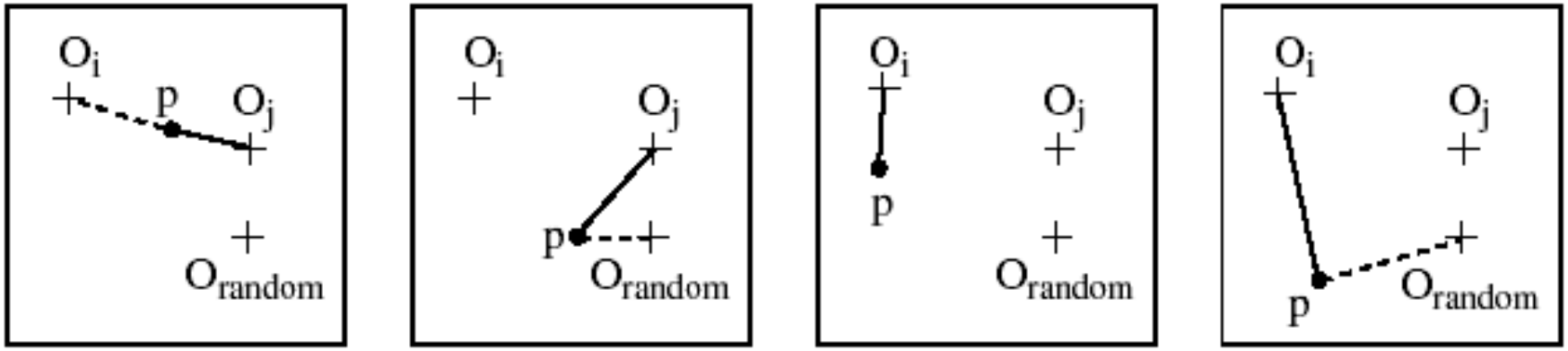
- Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters:

  A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9).

  The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the k-*means* algorithm to show *only*
- (a) The three cluster centers after the first round execution
- (b) The final three clusters

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Tính "total cost S of swapping $O_j$ và $O_{random}$" = $\Sigma_p C_{p/O_i O_{random}}$

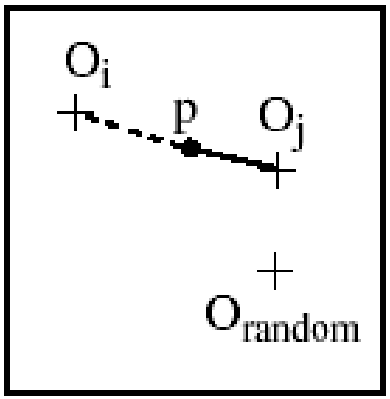



1. Reassigned to $O_i$
2. Reassigned to $O_{random}$
3. No change
4. Reassigned to $O_{random}$

• data object
\+ cluster center
— before swapping
--- after swapping

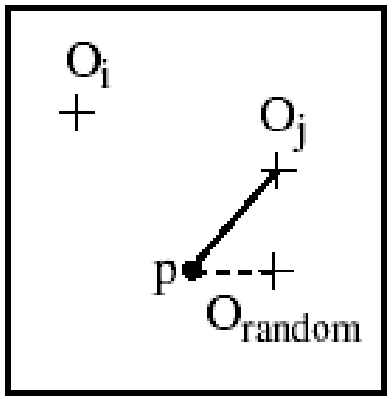
# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Tính "total cost S of swapping $O_j$ và $O_{random}$" = $\Sigma_p C_{p/O_i O_{random}}$



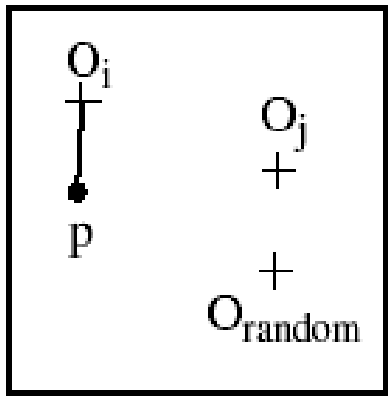

1. Reassigned to $O_i$

$C_{p/O_i O_{random}} = d(p,O_j) - d(p,O_i)$



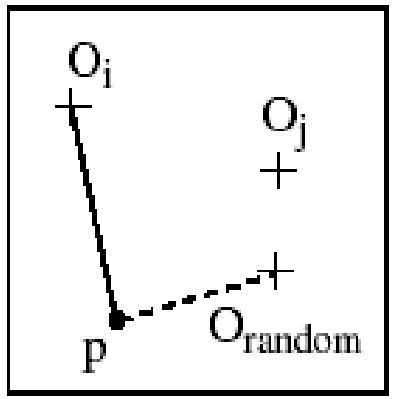

2. Reassigned to $O_{random}$

$C_{p/O_i O_{random}} = d(p,O_j) - d(p,O_{random})$




3. No change

$C_{p/O_i O_{random}} = 0$




4. Reassigned to $O_{random}$

$C_{p/O_i O_{random}} = d(p,O_i) - d(p,O_{random})$

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đặc điểm của giải thuật PAM (k-medoids)
  - ???

# 5.2. Gom cụm dữ liệu bằng phân hoạch

- Đặc điểm của giải thuật PAM (k-medoids)
  - Mỗi cụm được đại diện bởi phần tử chính giữa cụm (centroid).
    - Giảm thiểu sự ảnh hưởng của nhiễu (phần tử biên/kì dị/cực trị).
    - Số cụm k cần được xác định trước.
  - Độ phức tạp cho mỗi vòng lặp $O(k(n-k)^2)$
    - Giải thuật bị ảnh hưởng bởi kích thước tập dữ liệu.

# 5.3. Gom cụm dữ liệu bằng phân cấp

- Gom cụm dữ liệu bằng phân cấp (hierarchical clustering): nhóm các đối tượng vào cây phân cấp của các cụm
  - Agglomerative: bottom-up (trộn các cụm)
  - Divisive: top-down (phân tách các cụm)

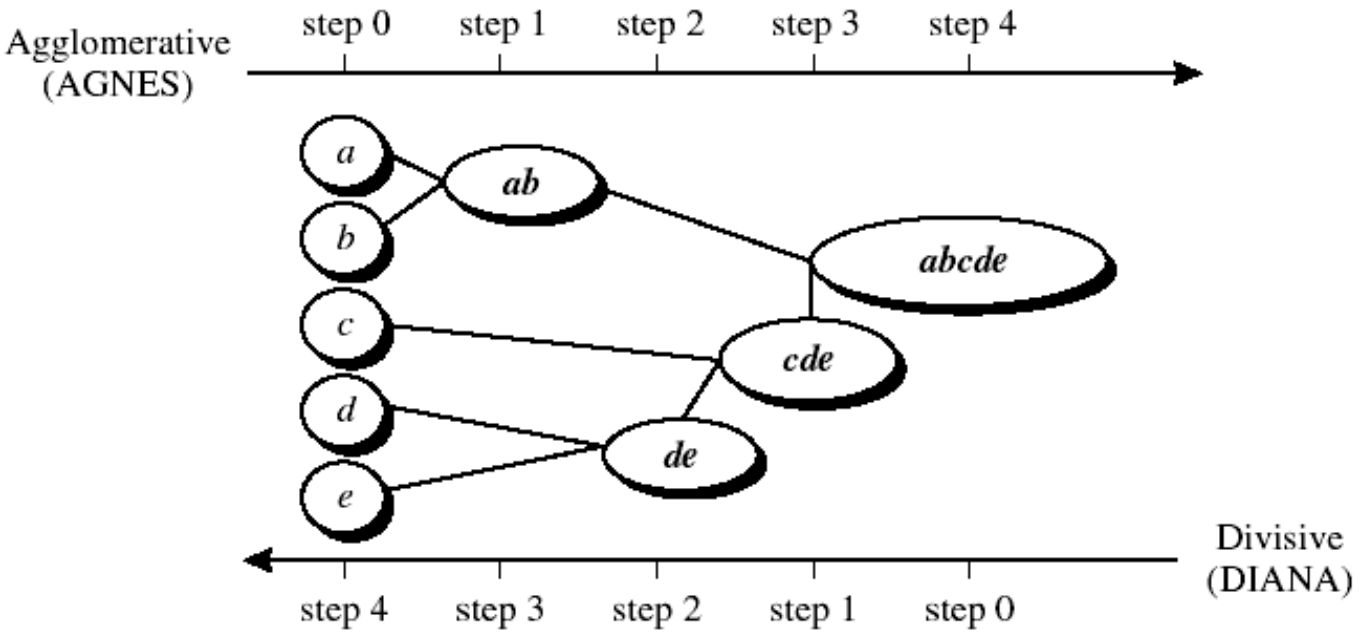
→ Không yêu cầu thông số nhập k (số cụm)

→ Yêu cầu điều kiện dừng

→ Không thể quay lui ở mỗi bước trộn/phân tách

# 5.3. Gom cụm dữ liệu bằng phân cấp

- An agglomerative hierarchical clustering method: AGNES (Agglomerative NESting) → bottom-up
- A divisive hierarchical clustering method: DIANA (Divisive ANAlysis) → top-down



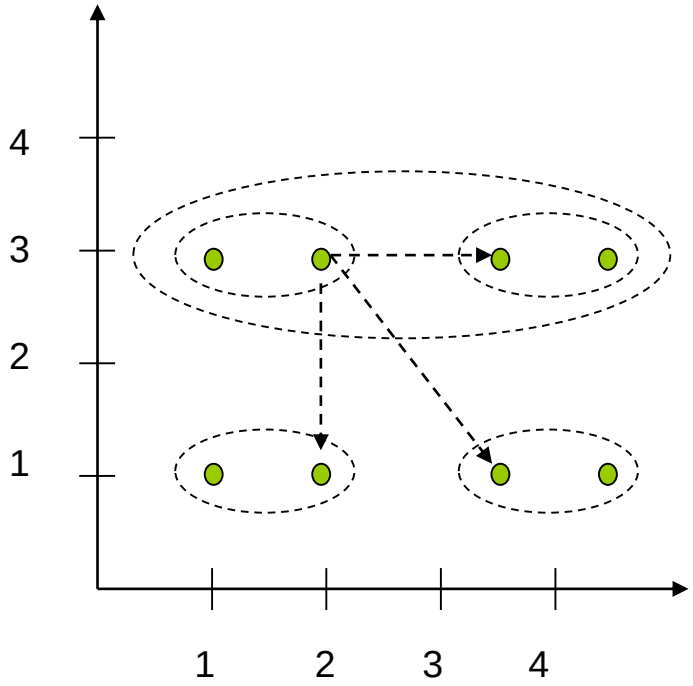

Agglomerative and divisive hierarchical clustering on data objects $\{a, b, c, d, e\}$.

# 5.3. Gom cụm dữ liệu bằng phân cấp

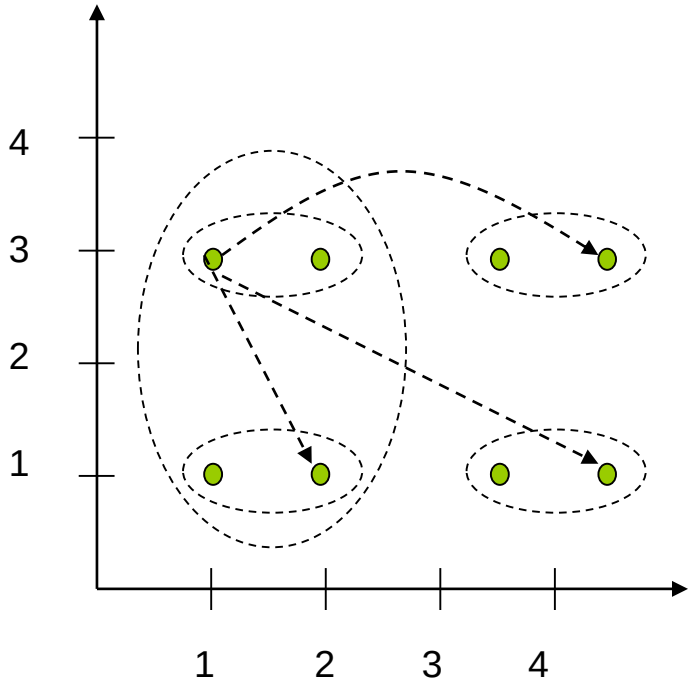
- An agglomerative hierarchical clustering method: AGNES (Agglomerative NESting)
  - Khởi đầu, mỗi đối tượng tạo thành một cụm.
  - Các cụm sau đó được trộn lại theo một tiêu chí nào đó.
    - Cách tiếp cận single-linkage: cụm C1 và C2 được trộn lại nếu khoảng cách giữa 2 đối tượng từ C1 và C2 là ngắn nhất.
  - Quá trình trộn các cụm được lặp lại đến khi tất cả các đối tượng tạo thành một cụm duy nhất.
- A divisive hierarchical clustering method: DIANA (Divisive ANAlysis)
  - Khởi đầu, tất cả các đối tượng tạo thành một cụm duy nhất.
  - Một cụm được phân tách theo một tiêu chí nào đó đến khi mỗi cụm chỉ có một đối tượng.
    - Khoảng cách lớn nhất giữa các đối tượng cận nhau nhất.

# 5.3. Gom cụm dữ liệu bằng phân cấp

**Single-linkage**



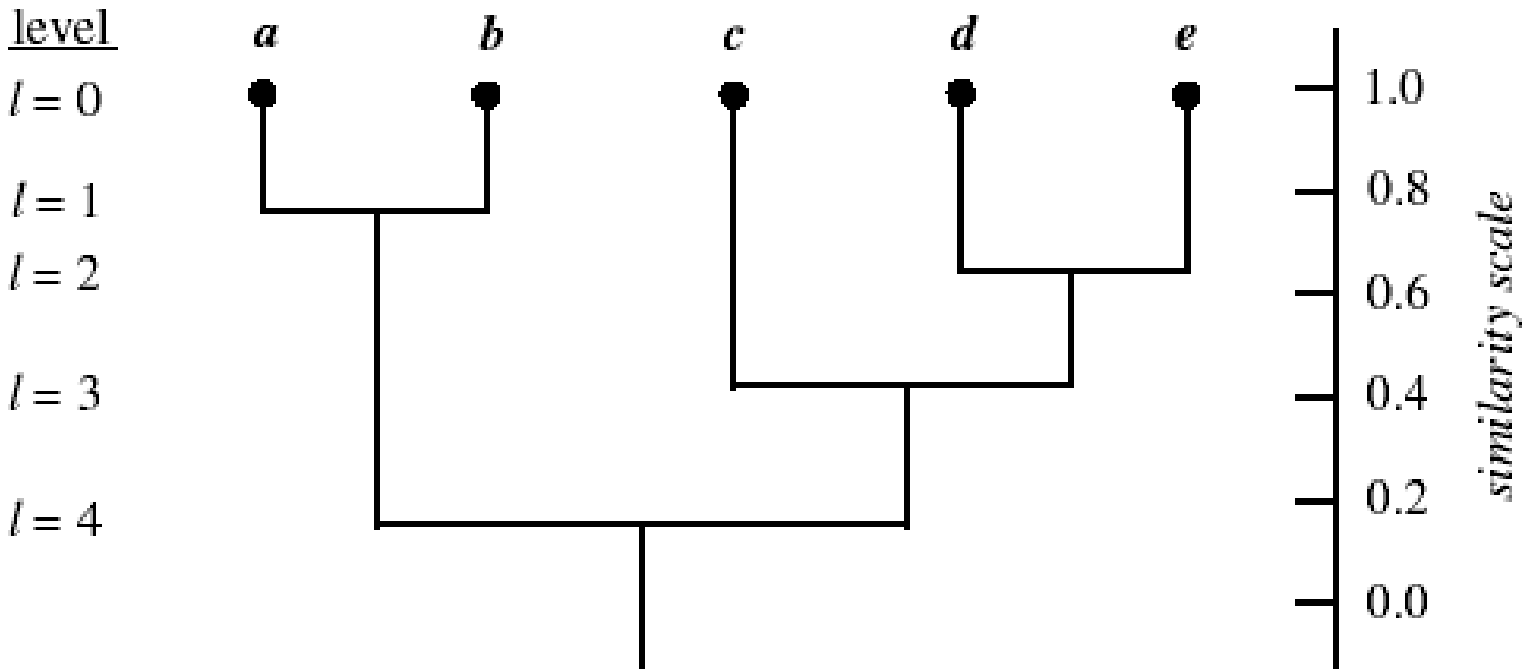

**Complete-linkage**




Tiêu chí trộn các cụm: single-linkage và complete-linkage

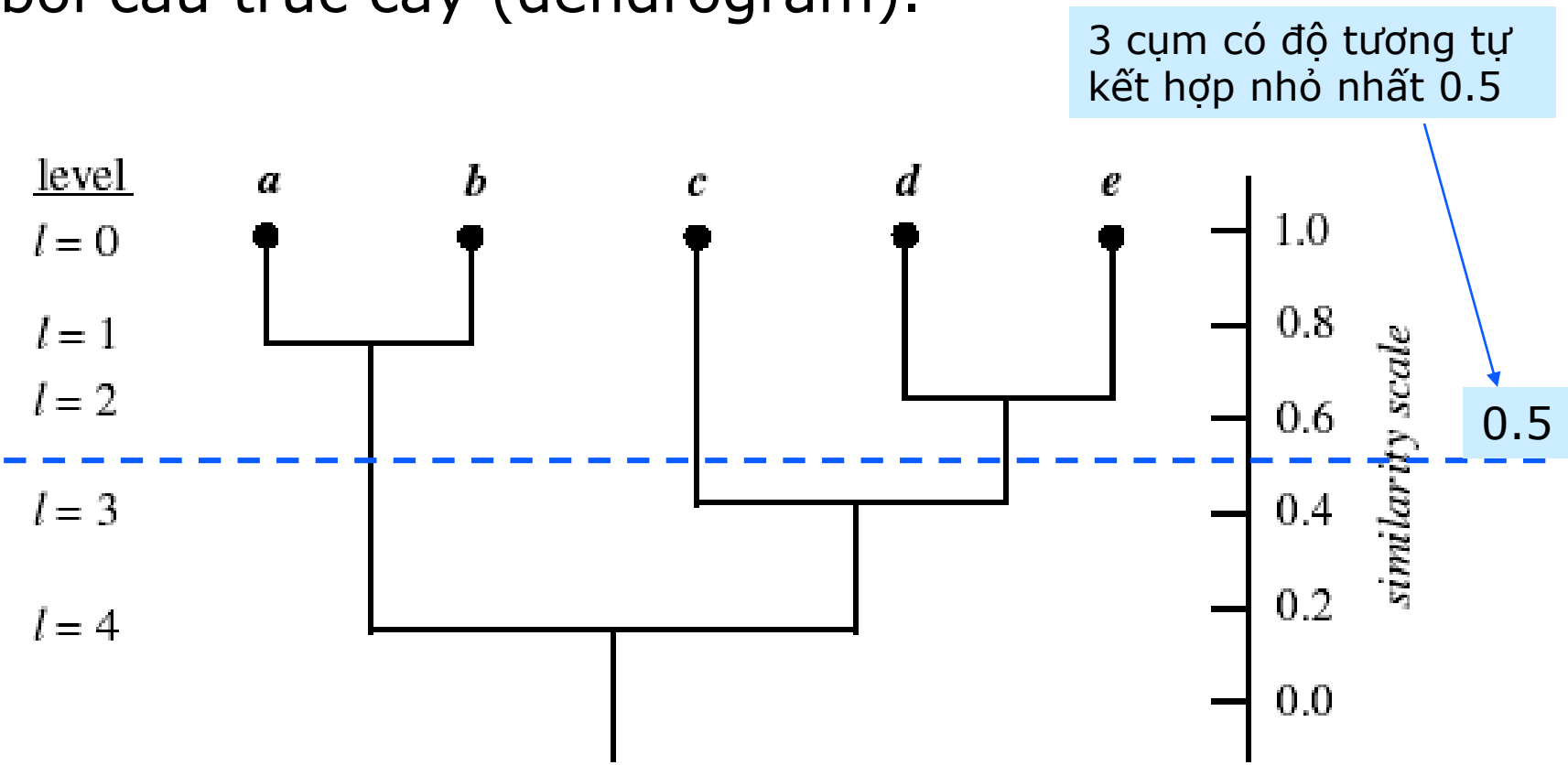
# 5.3. Gom cụm dữ liệu bằng phân cấp

- Quá trình gom cụm bằng phân cấp được biểu diễn bởi cấu trúc cây (dendrogram).



# 5.3. Gom cụm dữ liệu bằng phân cấp

- Quá trình gom cụm bằng phân cấp được biểu diễn bởi cấu trúc cây (dendrogram).



# 5.3. Gom cụm dữ liệu bằng phân cấp

- Các độ đo dùng đo khoảng cách giữa các cụm $C_i$ và $C_j$

$$\text{Minimum distance}: \quad d_{min}(C_i, C_j) = min_{p \in C_i,\, p' \in C_j} |p - p'|$$

$$\text{Maximum distance}: \quad d_{max}(C_i, C_j) = max_{p \in C_i,\, p' \in C_j} |p - p'|$$

$$\text{Mean distance}: \quad d_{mean}(C_i, C_j) = |m_i - m_j|$$

$$\text{Average distance}: \quad d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

p, p': các đối tượng

|p-p'|: khoảng cách giữa p và p'

$m_i$, $m_j$: đối tượng trung bình của $C_i$, $C_j$, tương ứng

$n_i$, $n_j$: số lượng đối tượng của $C_i$, $C_j$, tương ứng

# 5.3. Gom cụm dữ liệu bằng phân cấp

- Một số giải thuật gom cụm dữ liệu bằng phân cấp
  - BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): phân hoạch các đối tượng dùng cấu trúc cây theo độ co giãn của phân giải (scale of resolution)
  - ROCK (Robust Clustering using linKs): gom cụm dành cho các thuộc tính rời rạc (categorical/discrete attributes), trộn các cụm dựa vào sự kết nối lẫn nhau giữa các cụm
  - Chameleon: mô hình động để xác định sự tương tự giữa các cặp cụm

# 5.3. Gom cụm dữ liệu bằng phân cấp

- Một số vấn đề với gom cụm dữ liệu bằng phân cấp
  - Chọn điểm trộn/phân tách phù hợp
  - Khả năng co giãn (scalability)
    - Mỗi quyết định trộn/phân tách yêu cầu kiểm tra/đánh giá nhiều đối tượng/cụm.

  → Tích hợp gom cụm dữ liệu bằng phân cấp với các kỹ thuật gom cụm khác

  → Gom cụm nhiều giai đoạn (multiple-phase clustering)

# 5.4. Gom cụm dữ liệu dựa trên mật độ

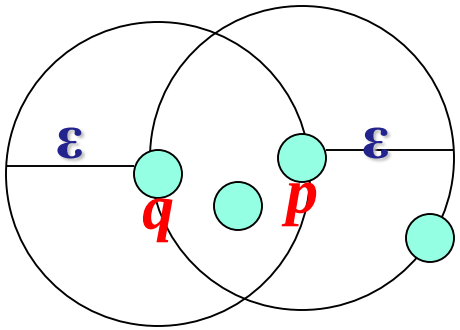
- Gom cụm dữ liệu dựa trên mật độ
  - Mỗi cụm là một vùng dày đặc (dense region) gồm các đối tượng.
    - Các đối tượng trong vùng thưa hơn được xem là nhiễu.
  - Mỗi cụm có dạng tùy ý.
- Giải thuật
  - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - OPTICS (Ordering Points To Identify the Clustering Structure)
  - DENCLUE (DENsity-based CLUstEring)

# 5.4. Gom cụm dữ liệu dựa trên mật độ

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - Phân tích các điểm kết nối nhau dựa vào mật độ
- OPTICS (Ordering Points To Identify the Clustering Structure)
  - Tạo ra thứ tự các điểm dữ liệu tùy vào cấu trúc gom cụm dựa vào mật độ của tập dữ liệu
- DENCLUE (DENsity-based CLUstEring)
  - Gom cụm dựa vào các hàm phân bố mật độ

# 5.4. Gom cụm dữ liệu dựa trên mật độ

- Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
  - ***ε***: bán kính của vùng láng giềng của một đối tượng, gọi là ***ε***-neighborhood.
  - ***MinPts***: số lượng đối tượng ít nhất được yêu cầu trong ***ε***-neighborhood của một đối tượng.
    - Nếu đối tượng có ***ε***-neighborhood với ***MinPts*** thì đối tượng này được gọi là đối tượng lõi (core object).

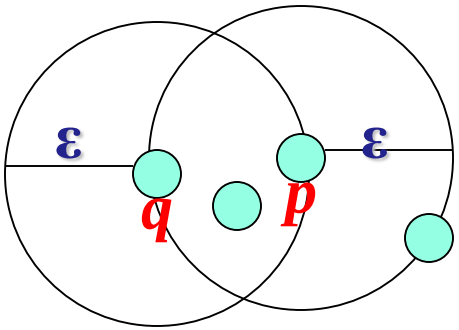



*p:* core object (***MinPts*** = 3)

*q:* không là core object

# 5.4. Gom cụm dữ liệu dựa trên mật độ

- Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
  - ***Directly density-reachable*** (khả năng đạt được trực tiếp): *q* có thể đạt được trực tiếp từ *p* nếu *q* trong vùng láng giềng ***ε***-neighborhood của *p* và *p* phải là core object.




*p:* directly density-reachable đối với *q*?

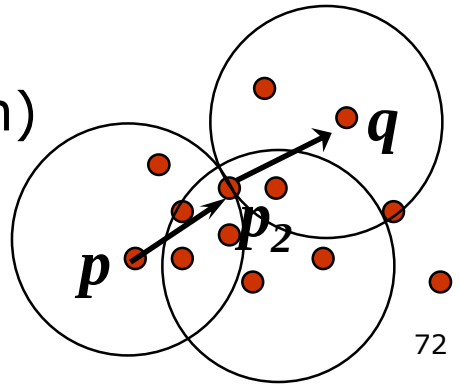
*q:* directly density-reachable đối với *p*?

*p:* directly density-reachable đối với *q*? X

*q:* directly density-reachable đối với *p*? √

# 5.4. Gom cụm dữ liệu dựa trên mật độ

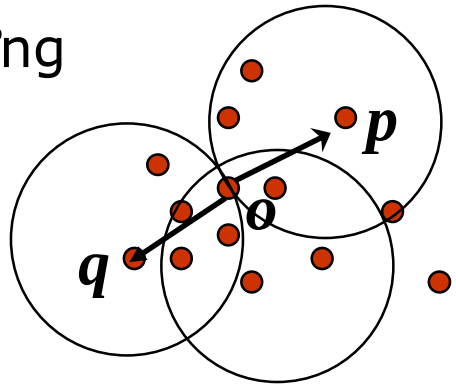
- Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
  - ***Density-reachable*** (khả năng đạt được):
    - Cho trước tập đối tượng ***D***, ***ε*** và ***MinPts***
    - *q* ***density-reachable*** từ *p* nếu ∃ chuỗi các đối tượng $p_1$, ..., $p_n \in$ ***D*** với $p_1 = p$ và $p_n = q$ sao cho $p_{i+1}$ ***directly density-reachable*** từ $p_i$ theo các thông số ***ε*** và ***MinPts***, $1 \le i \le n$.
    - Bao đóng truyền (transitive closure) của directly density-reachable
    - Quan hệ bất đối xứng (asymmetric relation)

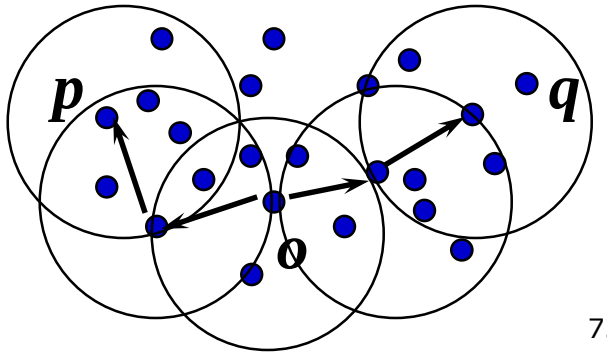



***MinPts*** = 5

# 5.4. Gom cụm dữ liệu dựa trên mật độ

- Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
  - ***Density-connected*** (nối kết dựa trên mật độ):
    - Cho trước tập các đối tượng ***D***, ***ε*** và ***MinPts***
    - $p, q \in \boldsymbol{D}$
    - $q$ ***density-connected*** với $p$ nếu $\exists\ o \in \boldsymbol{D}$ sao cho cả $q$ và $p$ đều ***density-reachable*** từ $o$ theo các thông số ***ε*** và ***MinPts***.
    - Quan hệ đối xứng

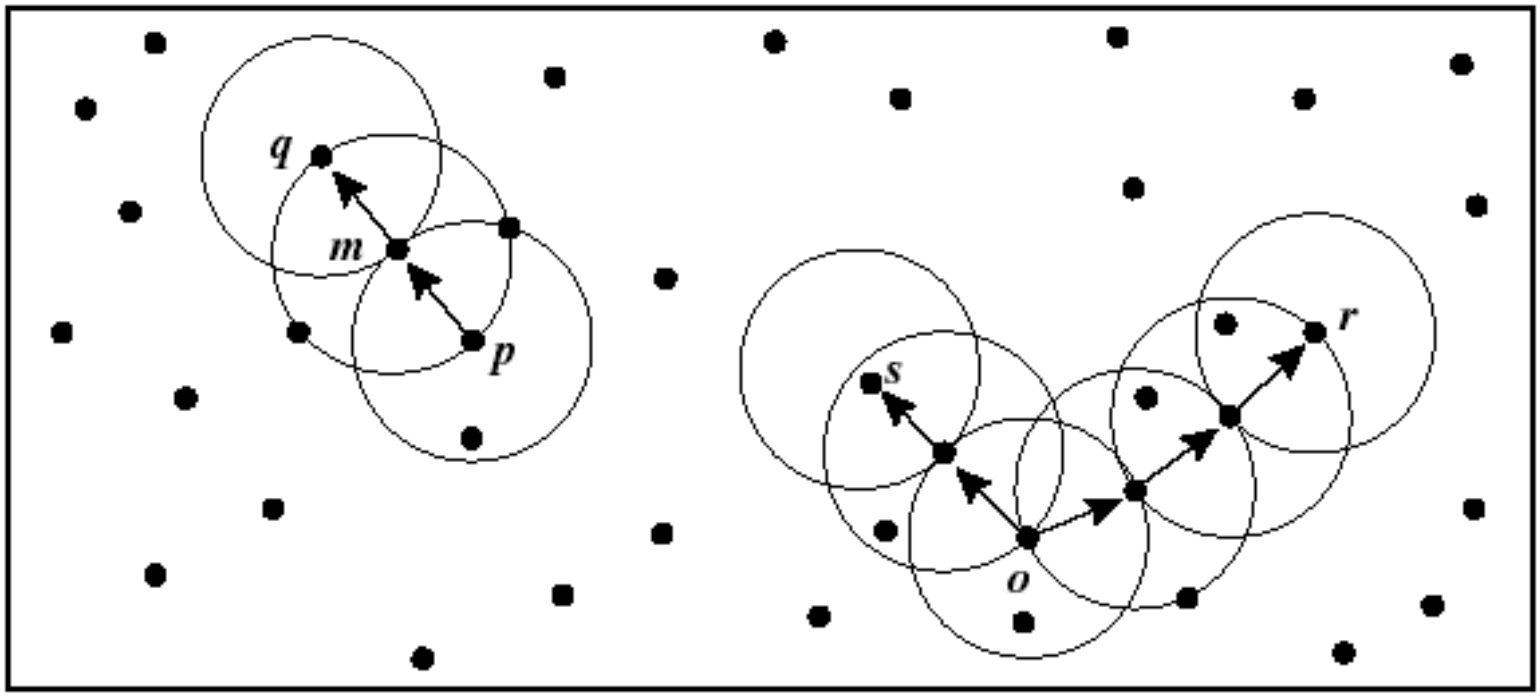


# 5.4. Gom cụm dữ liệu dựa trên mật độ

- Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ

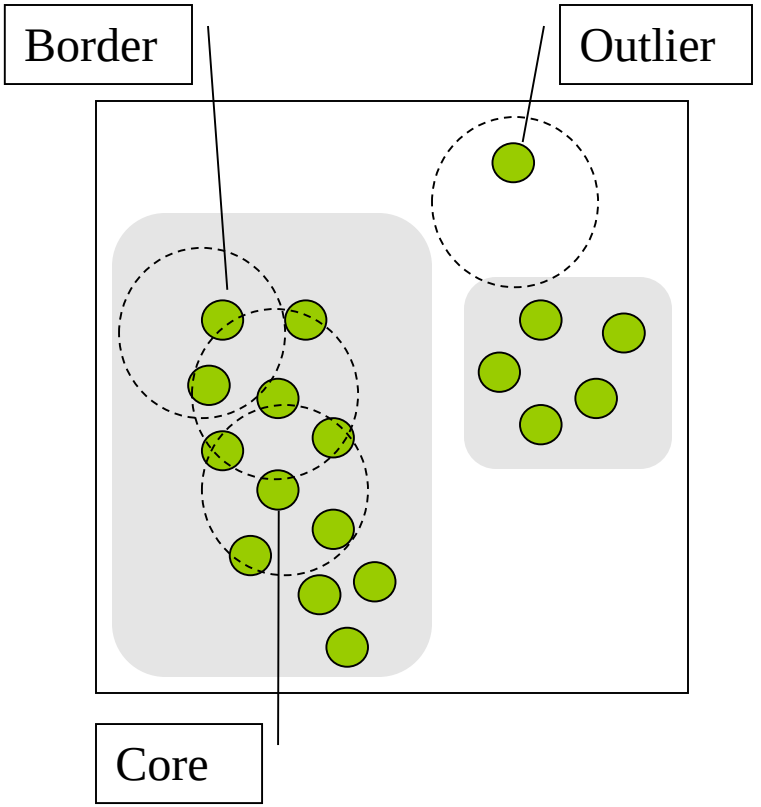
***MinPts*** = 3




Density reachability and density connectivity in density-based clustering

# 5.4. Gom cụm dữ liệu dựa trên mật độ

- Các khái niệm dùng trong gom cụm dữ liệu dựa trên mật độ
  - Cụm dựa trên mật độ (density based cluster): tập tất cả các đối tượng được nối kết với nhau dựa trên mật độ.
    - Đối tượng thuộc về cụm có thể là core object.
      - Nếu đối tượng đó không là core object thì gọi là đối tượng ranh giới (border object).
    - Đối tượng không thuộc về cụm nào được xem là nhiễu (noise/outlier).



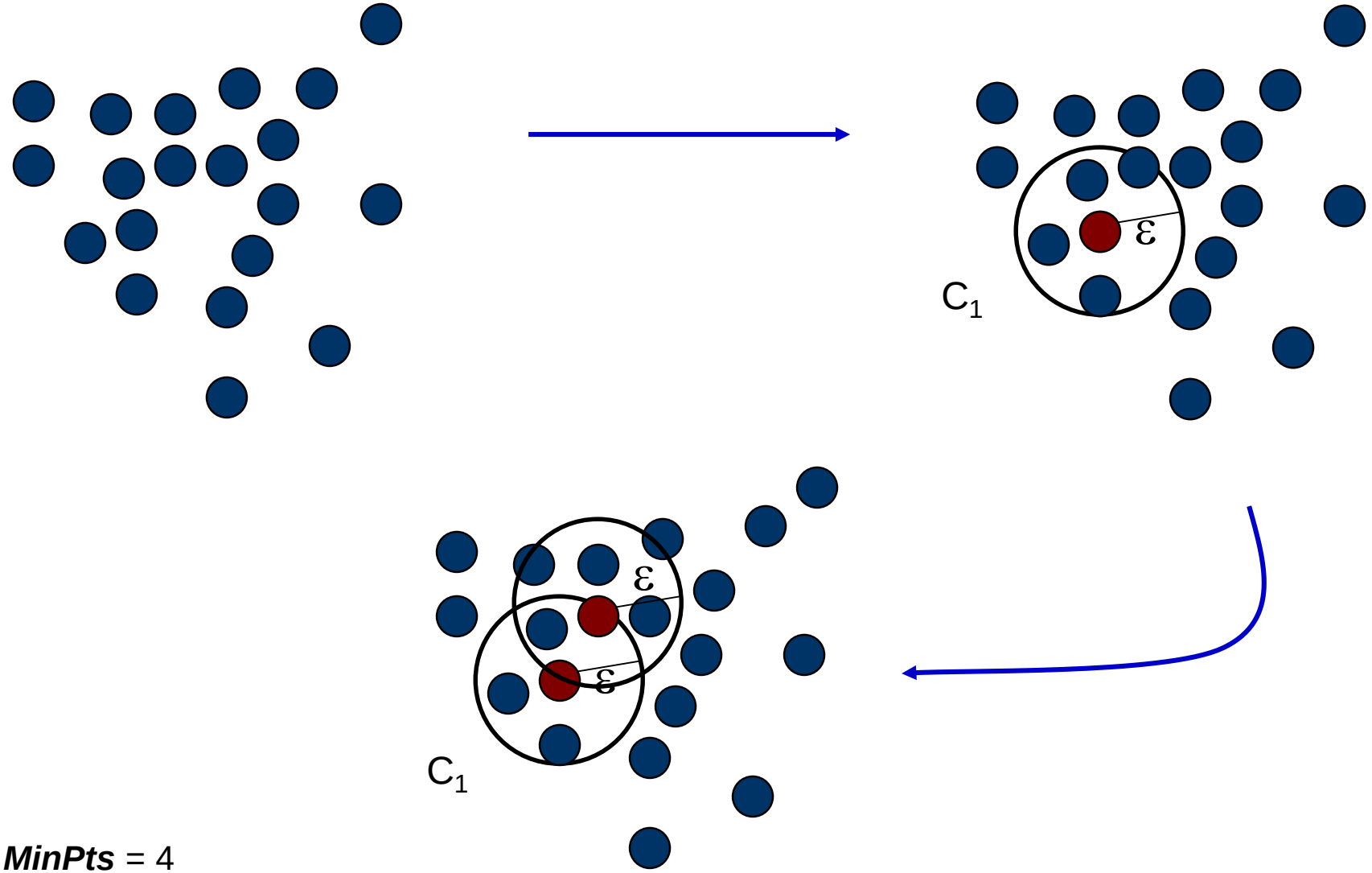

$\varepsilon$ = 1cm

***MinPts*** = 5

# 5.4. Gom cụm dữ liệu dựa trên mật độ

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - Input: tập đối tượng ***D***, ***ε***, ***MinPts***
  - Output: density-based clusters (và noise/outliers)
  - Giải thuật
    - 1. Xác định ***ε***–neighborhood của mỗi đối tượng $p \in \boldsymbol{D}$.
    - 2. If *p* là core object, tạo được một cluster.
    - 3. Từ bất kì core object *p*, tìm tất cả các đối tượng ***density-reachable*** và đưa các đối tượng này (hoặc các cluster) vào cùng cluster ứng với *p*.
      - 3.1. Các cluster đạt được (density-reachable cluster) có thể được trộn lại với nhau.
      - 3.2. Dừng khi không có đối tượng mới nào được thêm vào.

# 5.4. Gom cụm dữ liệu dựa trên mật độ




***MinPts*** = 4

# 5.4. Gom cụm dữ liệu dựa trên mật độ

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
  - Đặc điểm ???
    - Các cụm có dạng và kích thước khác nhau.
      - Không có giả định về phân bố của các đối tượng dữ liệu
        - Không yêu cầu về số cụm
      - Không phụ thuộc vào cách khởi động (initialization)
    - Xử lý nhiễu (noise) và các phần tử biên (outliers)
    - Yêu cầu trị cho thông số nhập
      - Yêu cầu định nghĩa của mật độ (density)
      - ***ε*** và ***MinPts***
    - Độ phức tạp
      - O(nlogn) → $O(n^2)$

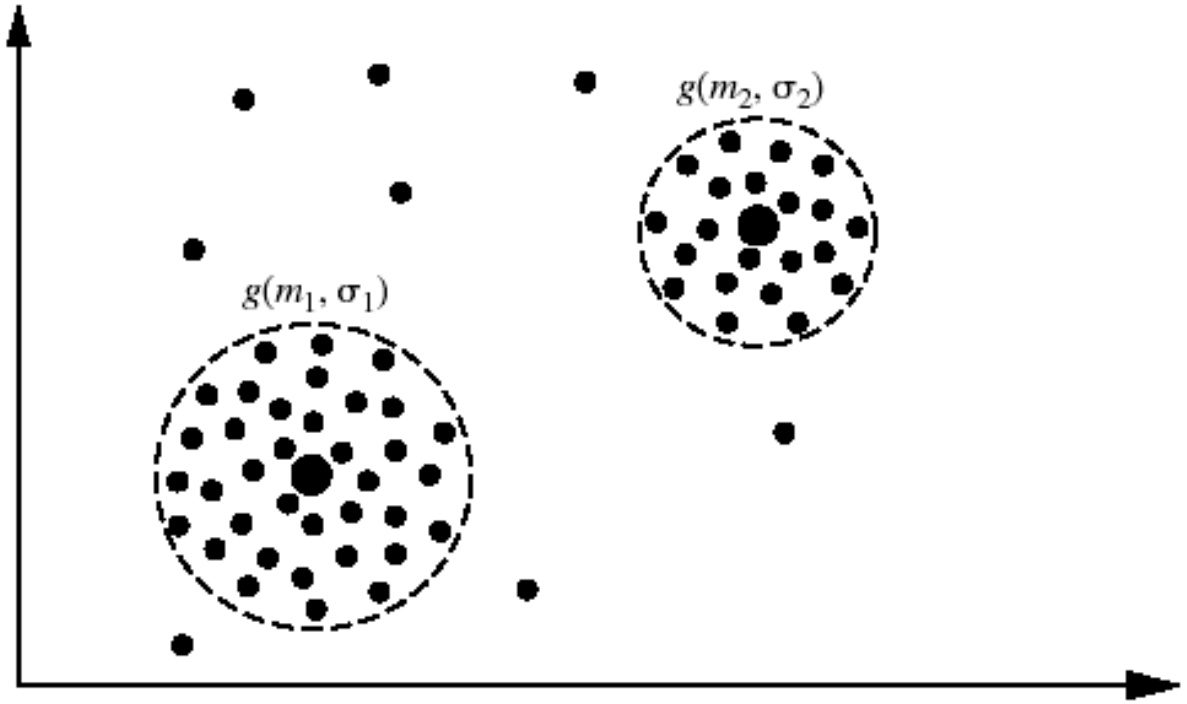
# 5.5. Gom cụm dữ liệu dựa trên mô hình

- Tối ưu hóa sự phù hợp giữa dữ liệu và mô hình toán nào đó
  - Giả định về quá trình tạo dữ liệu
    - Dữ liệu được tạo ra với nhiều sự phân bố xác suất khác nhau.
- Các phương pháp
  - Tiếp cận thống kê
    - Mở rộng của giải thuật gom cụm dựa trên phân hoạch k-means: Expectation-Maximization (EM)
  - Tiếp cận học máy: gom cụm ý niệm (conceptual clustering)
  - Tiếp cận mạng neural: Self-Organizing Feature Map (SOM)

# 5.5. Gom cụm dữ liệu dựa trên mô hình

- Gom cụm Expectation-Maximization (EM)
  - Giải thuật tinh chỉnh lặp để gán các đối tượng vào các cụm (bước kỳ vọng) và ước lượng trị thông số (bước cực đại hoá).
- Gom cụm ý niệm (conceptual clustering)
  - Tạo ra cách phân lớp các đối tượng chưa được gán nhãn dựa vào các mô tả đặc trưng cho mỗi nhóm đối tượng ứng với mỗi khái niệm (concept).
- Gom cụm với mạng neural
  - Biểu diễn mỗi cụm là một ví dụ tiêu biểu (exemplar).
  - Exemplar đóng vai trò của một prototype của cụm.
  - Các đối tượng mới được phân bố vào một cụm nếu tương tự với exemplar của cụm đó nhất dựa trên độ đo khoảng cách.

# 5.5. Gom cụm dữ liệu dựa trên mô hình




Each cluster can be represented by a probability distribution, centered at a mean, and with a standard deviation. Here, we have two clusters, corresponding to the Gaussian distributions $g(m_1, \sigma_1)$ and $g(m_2, \sigma_2)$, respectively, where the dashed circles represent the first standard deviation of the distributions.

# 5.5. Gom cụm dữ liệu dựa trên mô hình

- Giải thuật Expectation-Maximization (EM)
  - Gán một đối tượng vào một cụm nếu tương tự trung tâm (mean) của cụm đó nhất
    - Dựa vào trọng số (weight) của đối tượng đối với mỗi cụm
      - Xác suất thành viên (probability of membership)
    - Không có ranh giới giữa các cụm
    - Trung tâm của mỗi cụm được tính dựa vào các độ đo có trọng số (weighted measures).
  - Hội tụ nhanh nhưng có thể tối ưu cục bộ

# 5.5. Gom cụm dữ liệu dựa trên mô hình

- Giải thuật Expectation-Maximization (EM)
  - Input: tập n đối tượng, **K** (số cụm)
  - Output: trị tối ưu cho các thông số của mô hình
  - Giải thuật:
    - 1. Khởi trị
      - 1.1. Chọn ngẫu nhiên **K** đối tượng làm trung tâm của **K** cụm
      - 1.2. Ước lượng trị ban đầu cho các thông số (nếu cần)
    - 2. Lặp tinh chỉnh các thông số (cụm):
      - 2.1. Bước kỳ vọng (expectation step): gán mỗi đối tượng $x_i$ đến cụm $C_k$ với xác suất $P(x_i \in C_k)$ với k=1..**K**
      - 2.2. Bước cực đại hóa (maximization step): ước lượng trị các thông số
      - 2.3. Dừng khi thỏa điều kiện định trước.

# 5.5. Gom cụm dữ liệu dựa trên mô hình

- Giải thuật Expectation-Maximization (EM)
  - Giải thuật:
    - 1. Khởi trị
    - 2. Lặp tinh chỉnh các thông số (cụm):
      - 2.1. Bước kỳ vọng (expectation step): gán mỗi đối tượng $x_i$ đến cụm $C_k$ với xác suất $P(x_i \in C_k)$

$$P(\boldsymbol{x_i} \in C_k) = p(C_k|\boldsymbol{x_i}) = \frac{p(C_k)p(\boldsymbol{x_i}|C_k)}{p(\boldsymbol{x_i})}$$

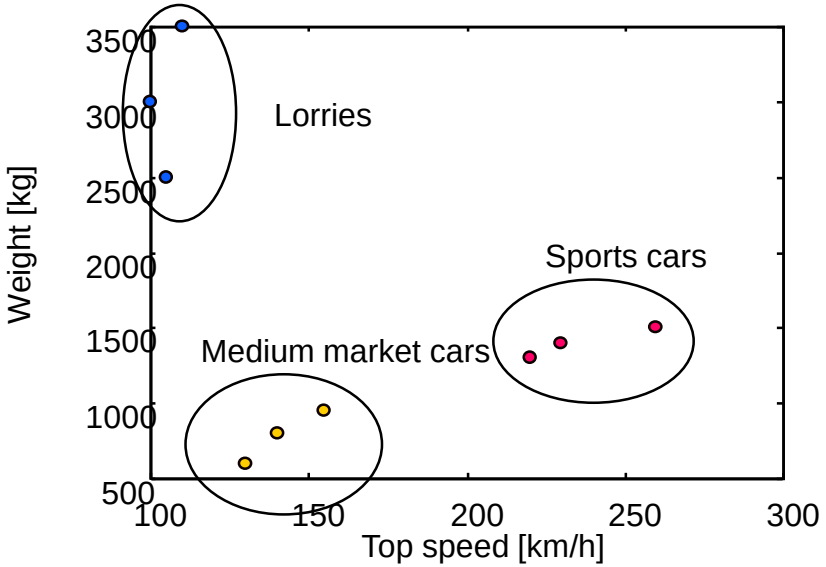
      - 2.2. Bước cực đại hóa (maximization step): ước lượng trị các thông số

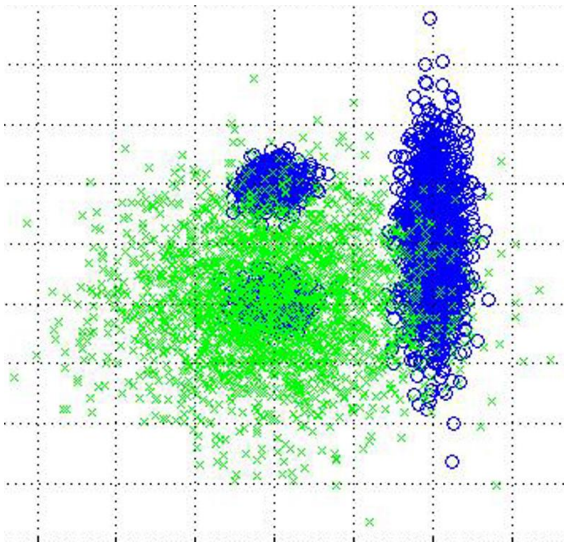
$$m_k = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{x_i}P(\boldsymbol{x_i} \in C_k)}{\sum_j P(\boldsymbol{x_i} \in C_j)}$$

($m_k$ là trung tâm của cụm $C_k$, j = 1..**K**, k = 1..**K**)

# 5.6. Các phương pháp gom cụm dữ liệu khác

- Gom cụm cứng (hard clustering)
  - Mỗi đối tượng chỉ thuộc về một cụm.
  - Mức thành viên (degree of membership) của mỗi đối tượng với một cụm hoặc là 0 hoặc là 1.
  - Ranh giới (boundary) giữa các cụm rõ ràng.
- Gom cụm mờ (fuzzy clustering)
  - Mỗi đối tượng thuộc về nhiều hơn một cụm với mức thành viên nào đó từ 0 đến 1.
  - Ranh giới giữa các cụm không rõ ràng (mờ - vague/fuzzy).



# 5.7. Tóm tắt

- Gom cụm nhóm các đối tượng vào các cụm dựa trên sự tương tự giữa các đối tượng.
- Độ đo đo sự tương tự tùy thuộc vào kiểu dữ liệu/đối tượng cụ thể.
- Các giải thuật gom cụm được phân loại thành: nhóm phân hoạch, nhóm phân cấp, nhóm dựa trên mật độ, nhóm dựa trên lưới, nhóm dựa trên mô hình, …

# 5.7. Tóm tắt

| Cluster algorithm | Complexity | Capability of tackling high dimensional data |
|---|---|---|
| $K$-means | $O(NKd)$ (time)<br>$O(N+K)$ (space) | No |
| Fuzzy $c$-means | Near $O(N)$ | No |
| Hierarchical clustering* | $O(N^2)$ (time)<br>$O(N^2)$ (space) | No |
| CLARA | $O(K(40+K)^2+K(N-K))^+$ (time) | No |
| CLARANS | Quadratic in total performance | No |
| BIRCH | $O(N)$ (time) | No |
| DBSCAN | $O(N \log N)$ (time) | No |
| CURE | $O(N_{sample}^2 \log N_{sample})$ (time)<br>$O(N_{sample})$ (space) | Yes |
| WaveCluster | $O(N)$ (time) | No |
| DENCLUE | $O(N \log N)$ (time) | Yes |
| FC | $O(N)$ (time) | Yes |
| CLIQUE | Linear with the number of objects, Quadratic with the number of dimensions | Yes |
| OptiGrid | Between $O(Nd)$ and $O(Nd \log N)$ | Yes |
| ORCLUS | $O(K_0^3+K_0Nd+K_0^2d^3)$ (time)<br>$O(K_0d^2)$ (space) | Yes |

R. Xu, D. Wunsch II. Survey of Clustering Algorithms. IEEE Transactions on Neural Netwo 16(3), May 2005, pp. 645-678.