Unsupervised Learning

Tien-Lam Pham

lam.phamtien@phenikaa-uni.edu.vn

PIAS
VIỆN NGHIÊN CỨU
TIÊN TIẾN PHENIKAA

PHENIKAA
UNIVERSITY

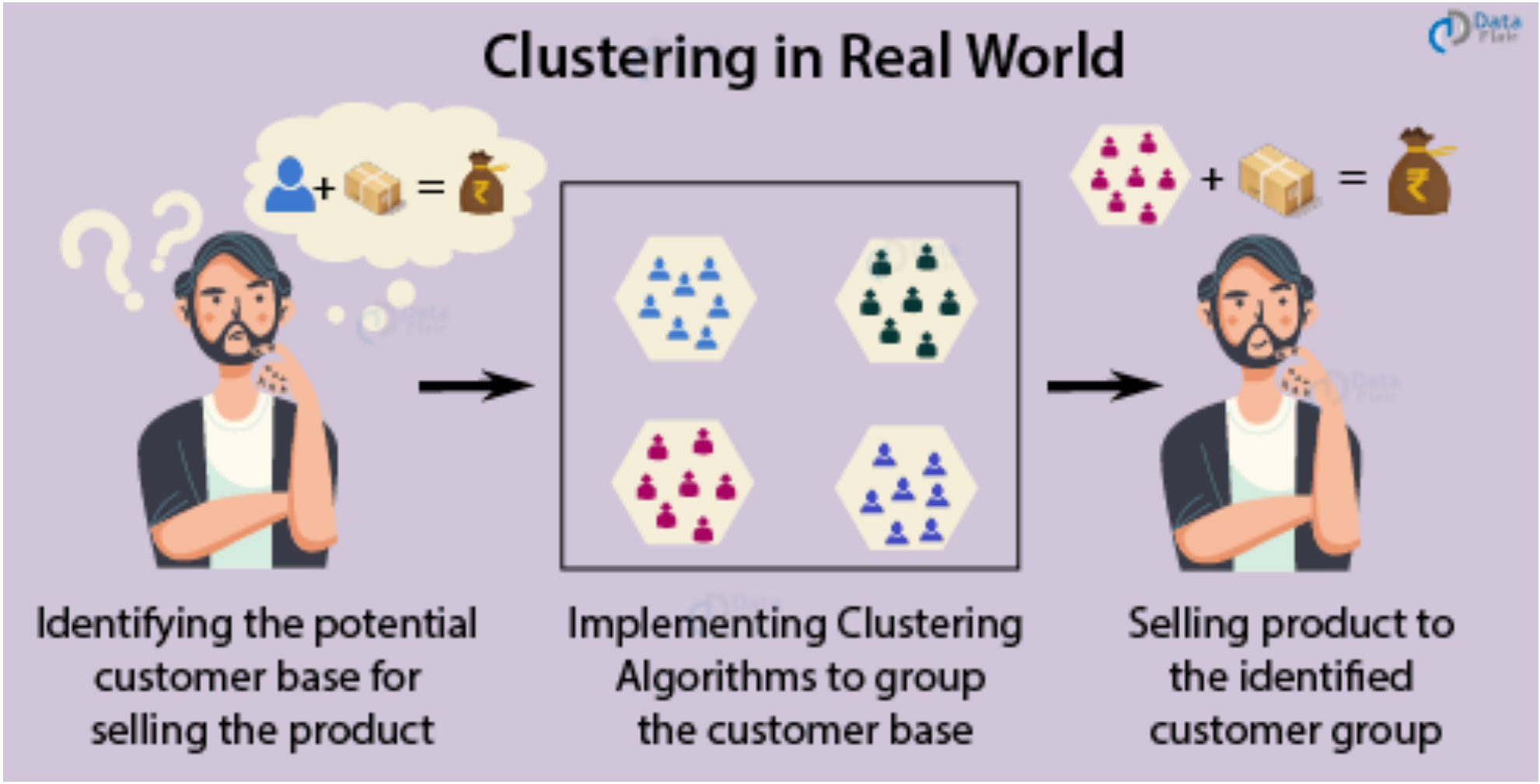
Data analysis and machine learning

- Population (quần thể)
- Tập mẫu
- Tập dữ liệu
- Feature vectors
- Đặc trưng thống kê: min, max, mean, variance, standard deviation
- Hệ số tương quan
- Phân tích biểu đồ: scatter plot, histogram
- Phân tích dự báo: regression (hồi quy) and classification (phân loại)
- **Phân cụm (clustering) và giảm chiều (dimensionality reduction)**

Unsupervised learning

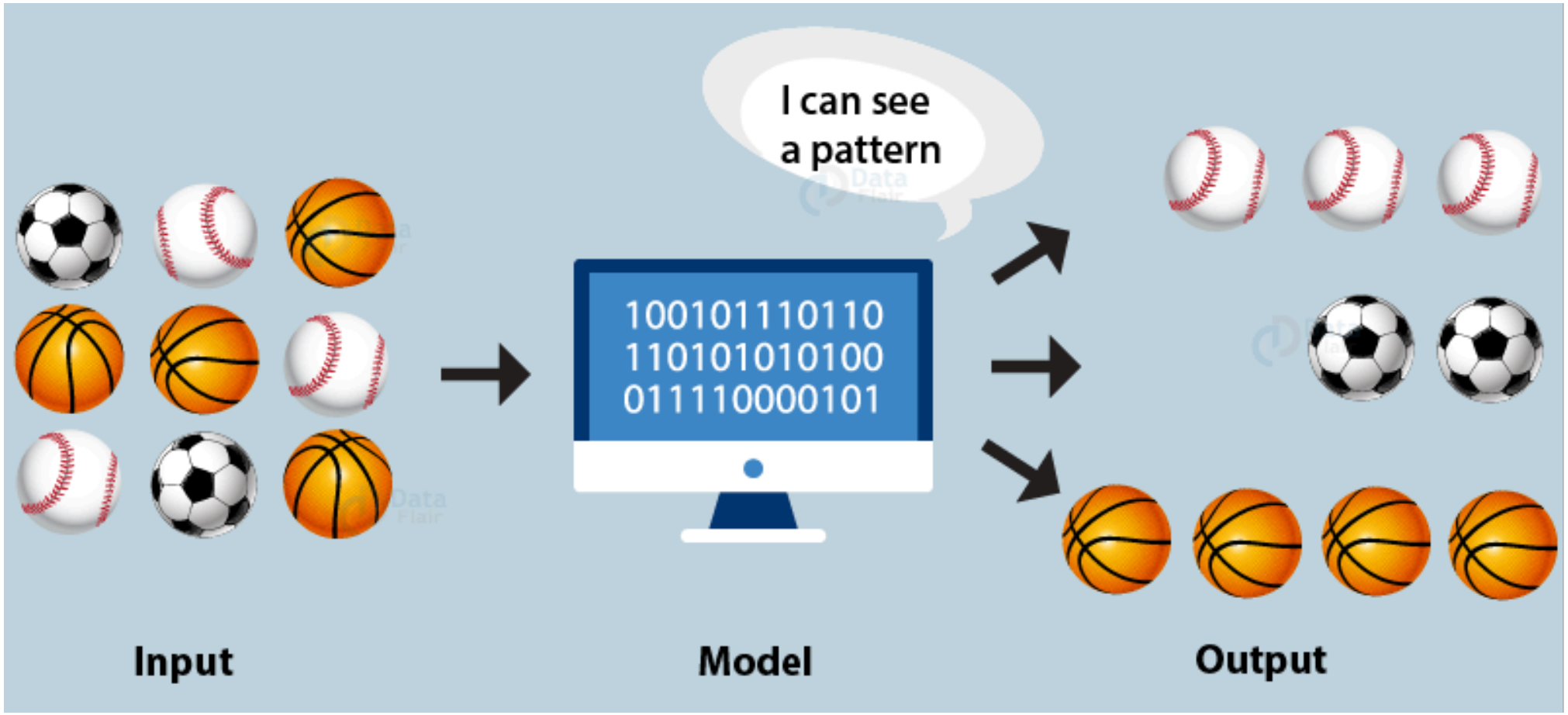
- Dữ liệu không có labels?
- Phân cụm: tìm kiếm các patterns trong dữ liệu
  - ✓ Nhóm các thư điện tử, nhóm các tìm kiếm, nhocms các văn bản, etc.
  - ✓ Nhóm các genes, nhóm các phân tử
  - ✓ Nhóm các loại khách hành
  - ✓ Phân chia các vùng trong ảnh
- Giảm chiều: visualize data, nén dữ liệu

Phân cụm



https://data-flair.training/blogs/clustering-in-machine-learning/

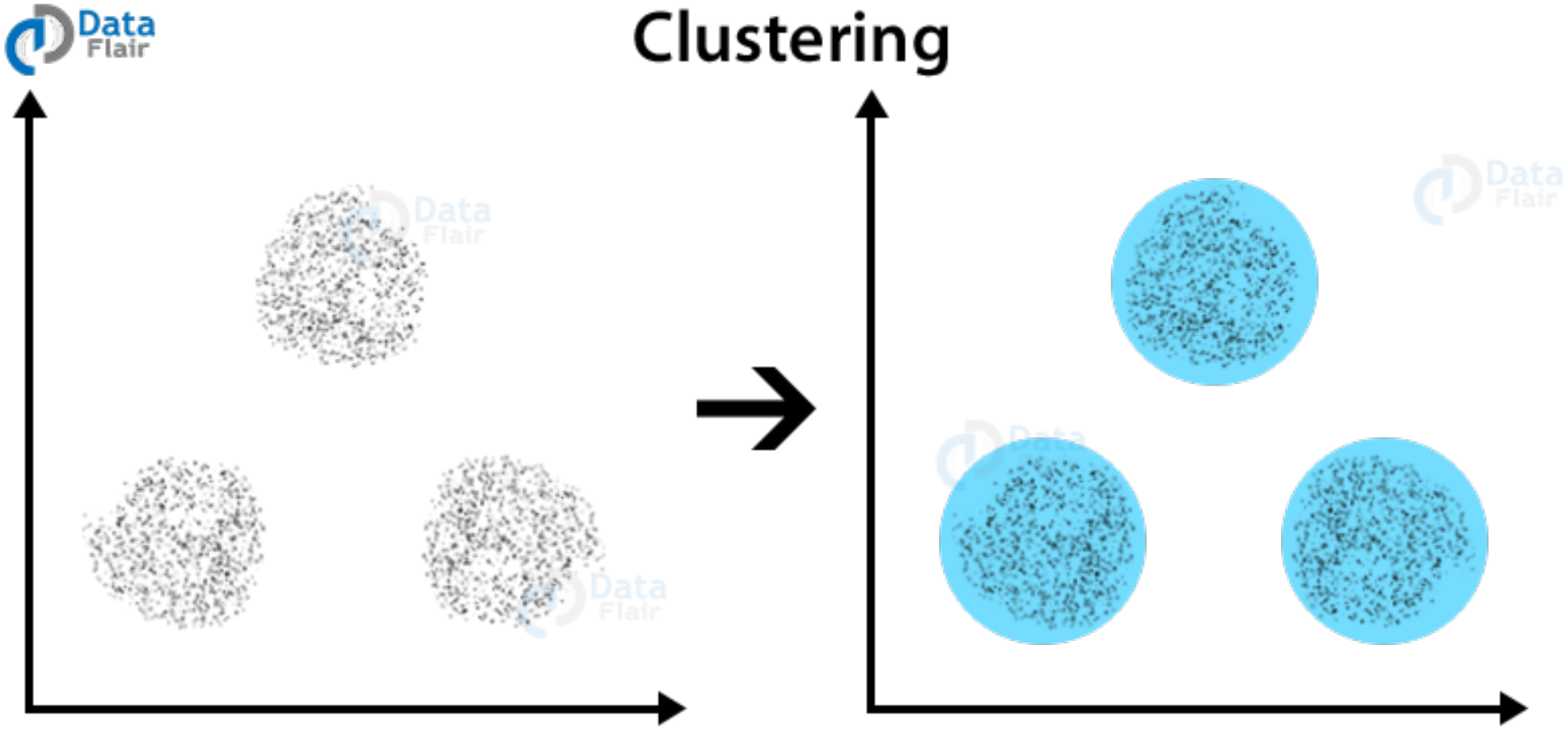
Phân cụm



https://data-flair.training/blogs/clustering-in-machine-learning/

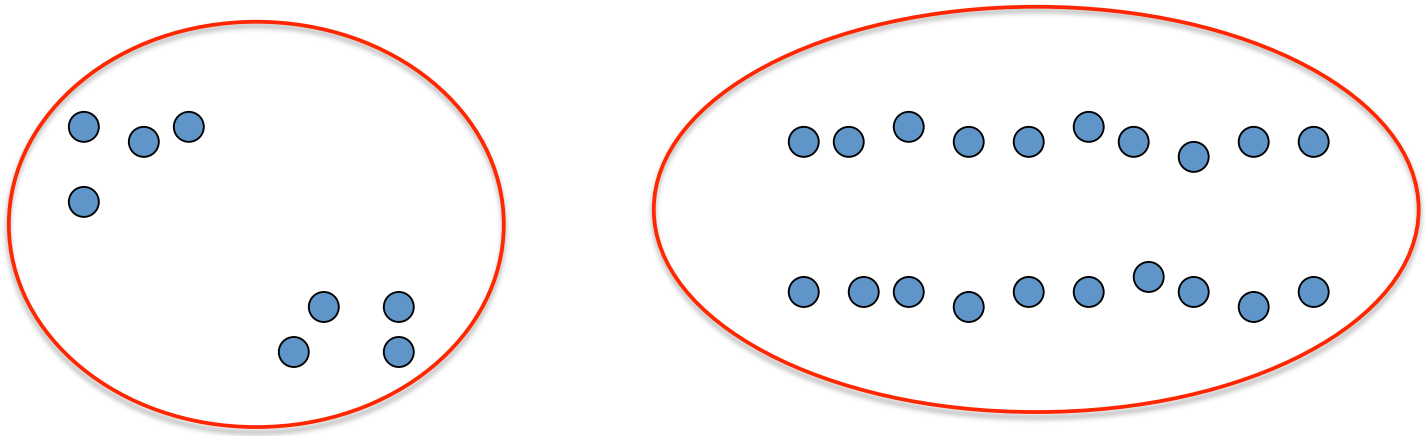
Phân cụm

Phân cụm



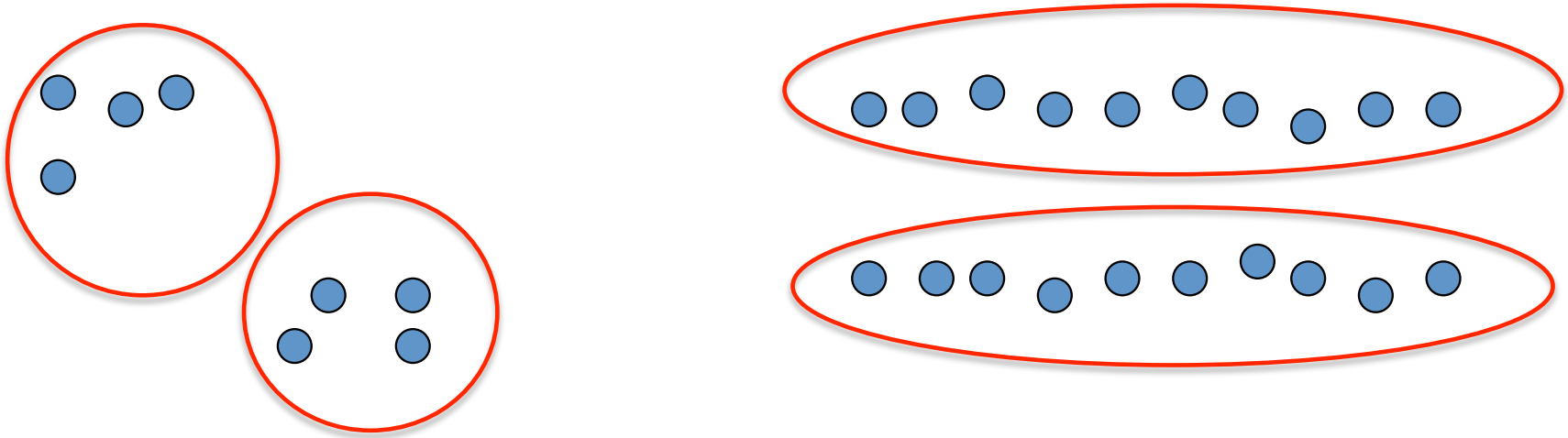
Phân cụm

- Mục tiêu nhóm các điểm dữ liệu có những thuộc tính chung



Phân cụm

- Mục tiêu nhóm các điểm dữ liệu có những thuộc tính chung



Thế nào là những điểm có thuộc tính chung?

Similarity measurement

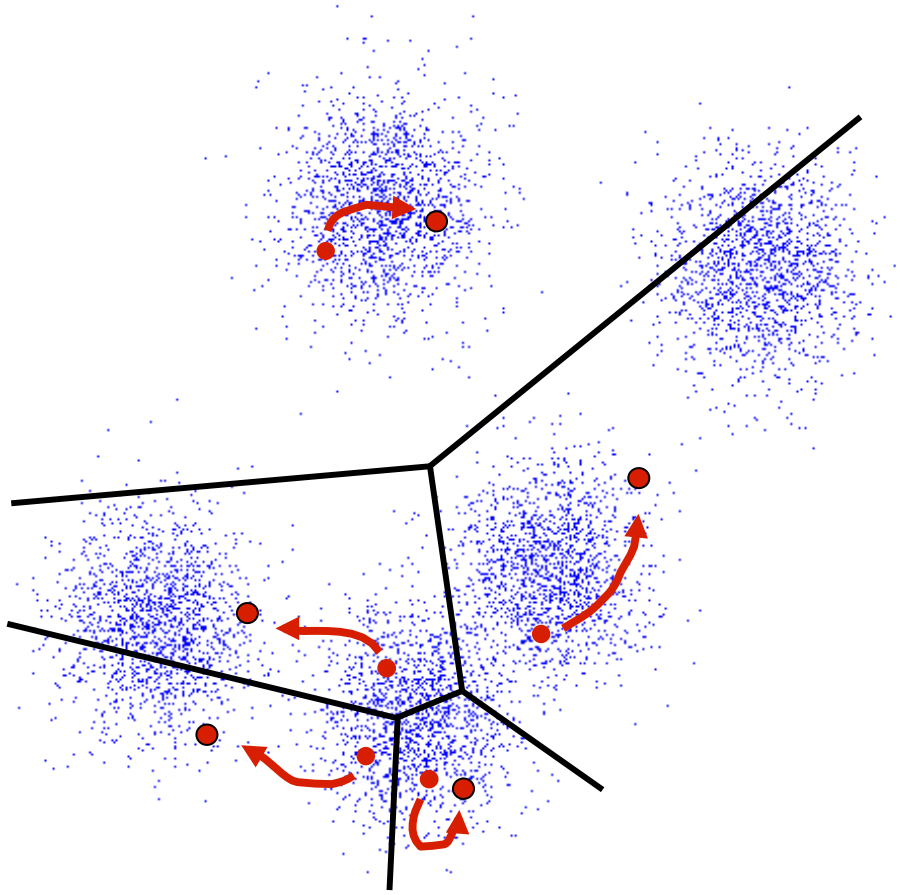
Euclidean distance:

$$\text{dist}(\vec{x}, \vec{y}) = ||\vec{x} - \vec{y}||_2^2$$

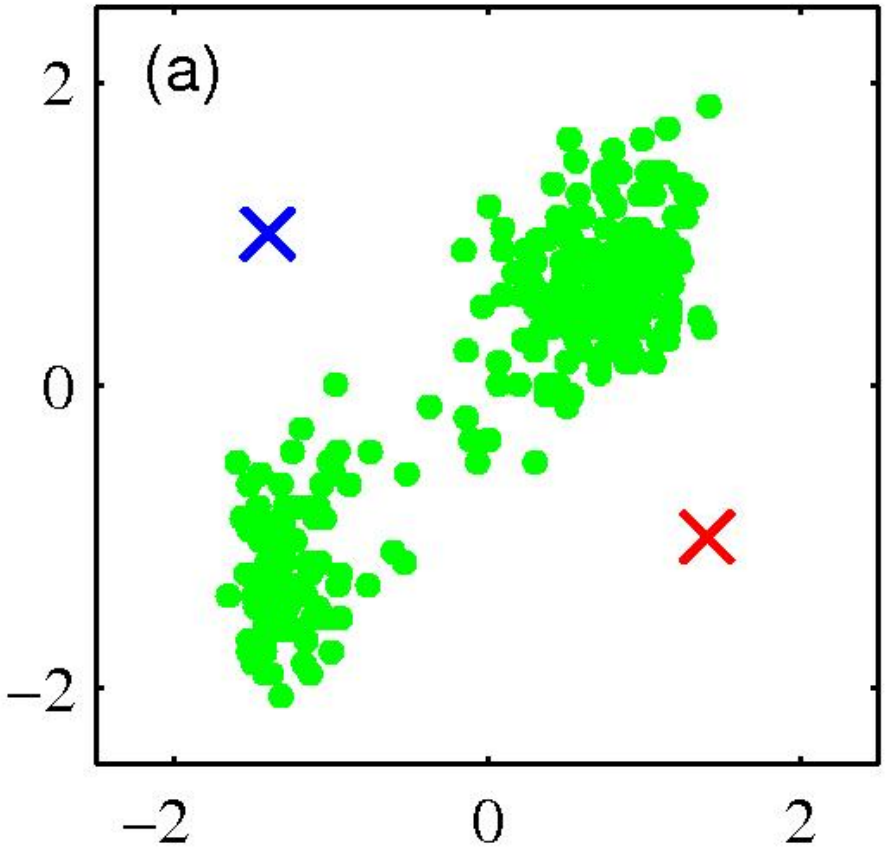
Các thuật toán phân cum phụ thuộc và phép đó độ tương tự (hoặc khoảng cách)

Kmean clustering

- An iterative clustering algorithm
  - Initialize: Pick *K* random points as cluster centers
  - Alternate:
    1. Assign data points to closest cluster center
    2. Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change



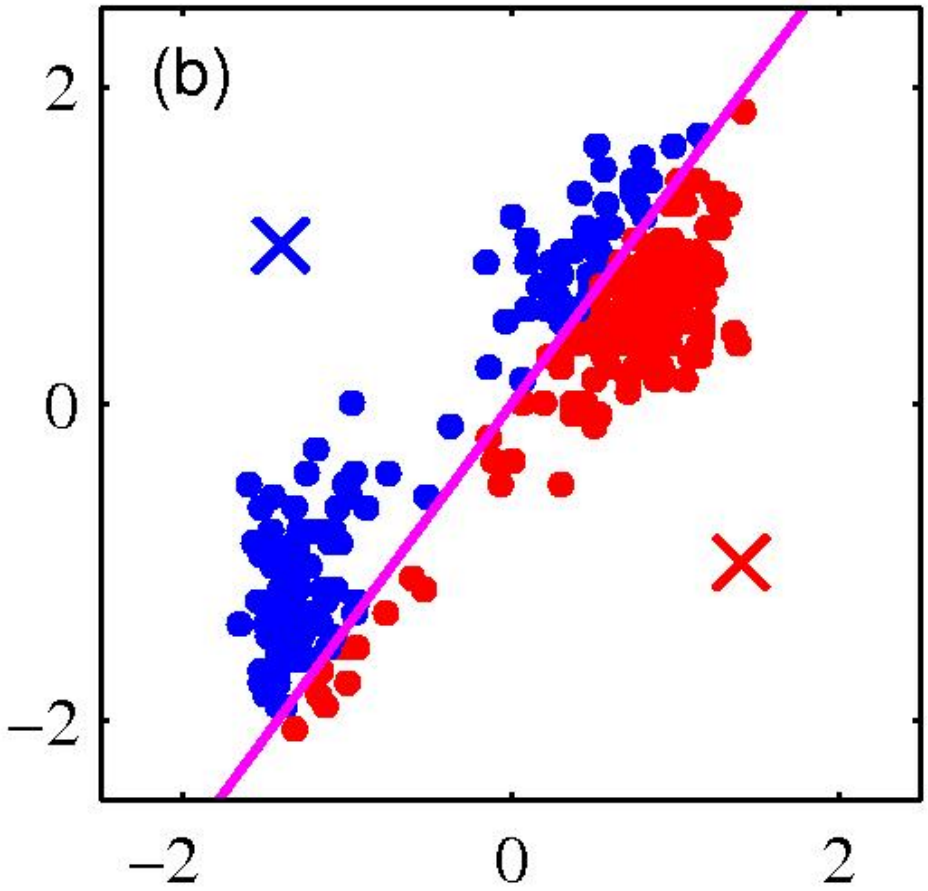
Kmean clustering



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

Shown here for *K*=2

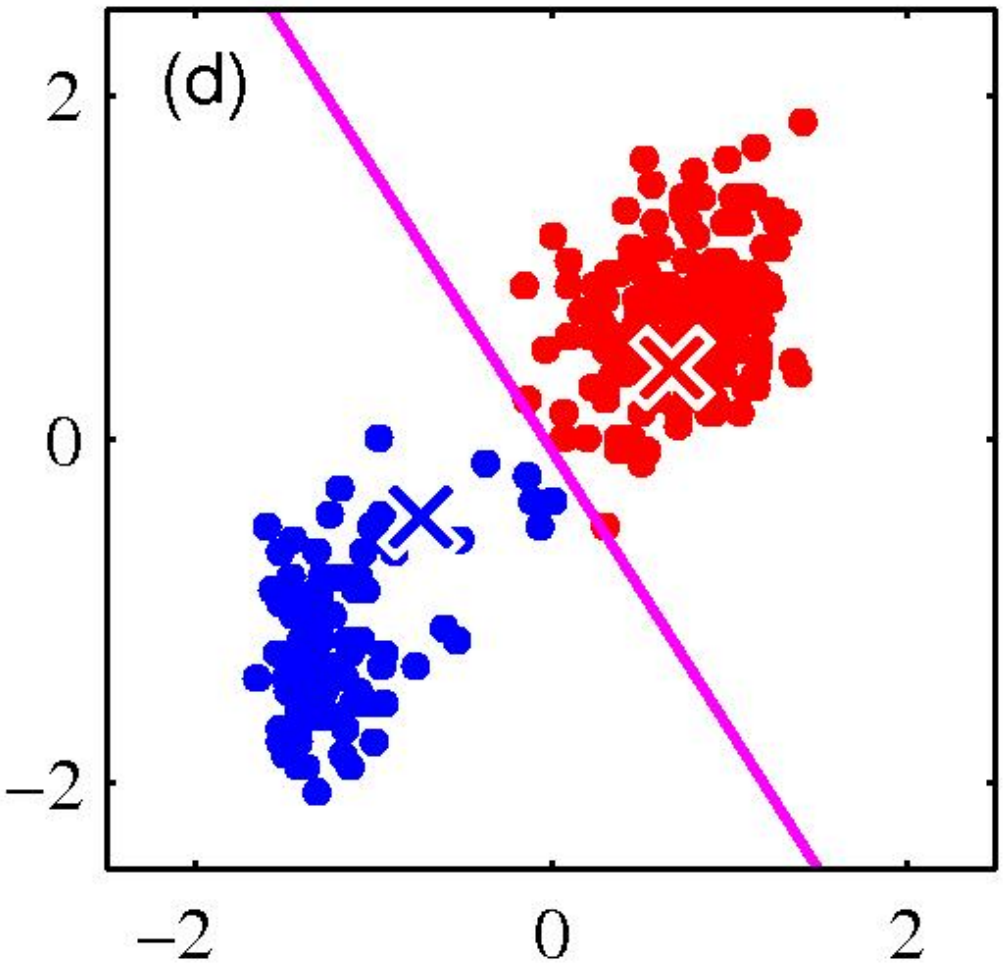
Kmean clustering



Iterative Step 1

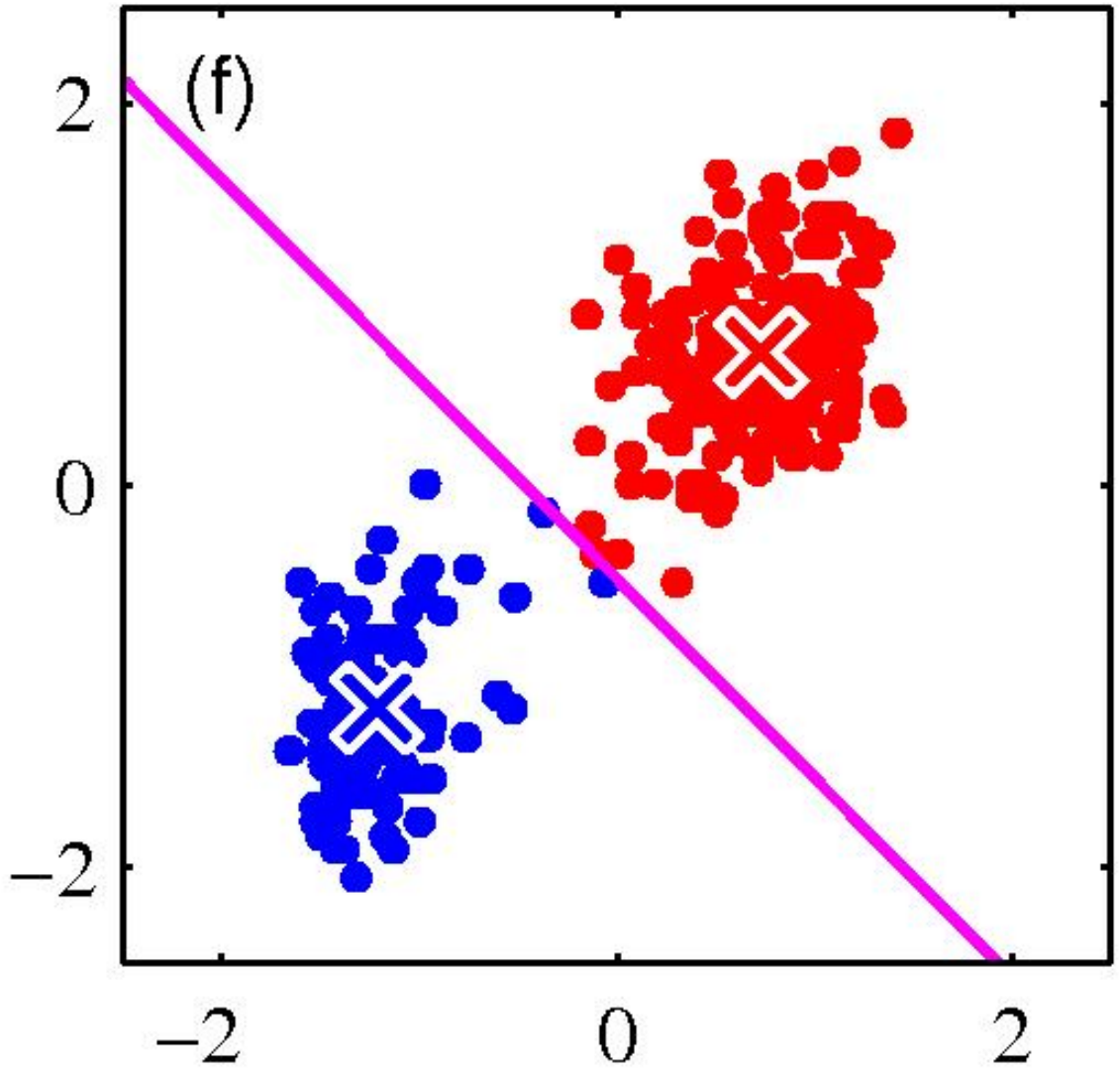
- Assign data points to closest cluster center

Kmean clustering



- Repeat until convergence

Kmean clustering



Kmean clustering

Objective

$$\min_{\mu}\min_{C}\sum_{i=1}^{k}\sum_{x\in C_i}|x-\mu_i|^2$$

1. Fix μ, optimize *C:*

$$\min_{C}\sum_{i=1}^{k}\sum_{x\in C_i}|x-\mu_i|^2=\min_{c}\sum_{i}^{n}\left|x_i-\mu_{x_i}\right|^2$$

Step 1 of kmeans

2. Fix *C,* optimize μ:

$$\min_{\mu}\sum_{i=1}^{k}\sum_{x\in C_i}|x-\mu_i|^2$$

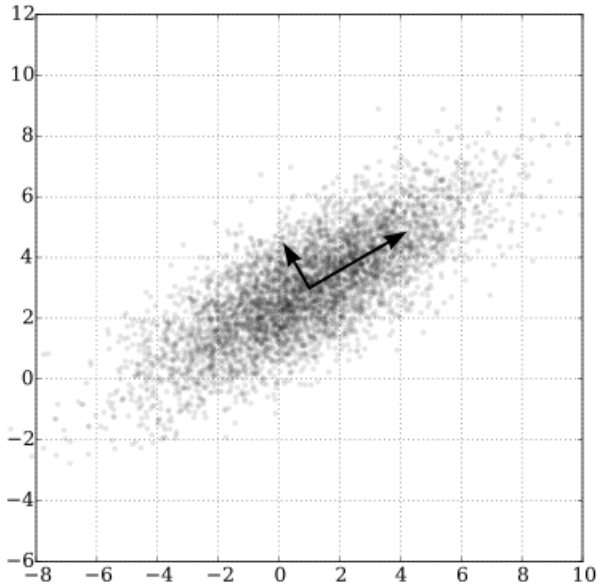
   – Take partial derivative of μ_i and set to zero, we have

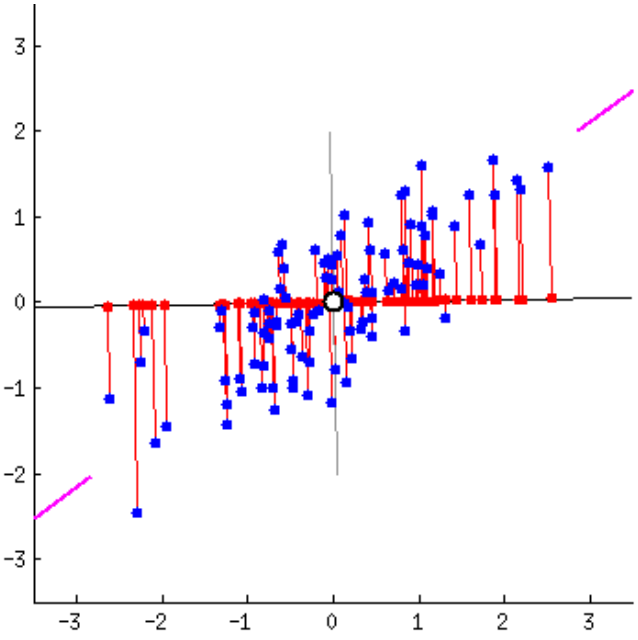
$$\mu_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$$

Step 2 of kmeans

Giảm chiều

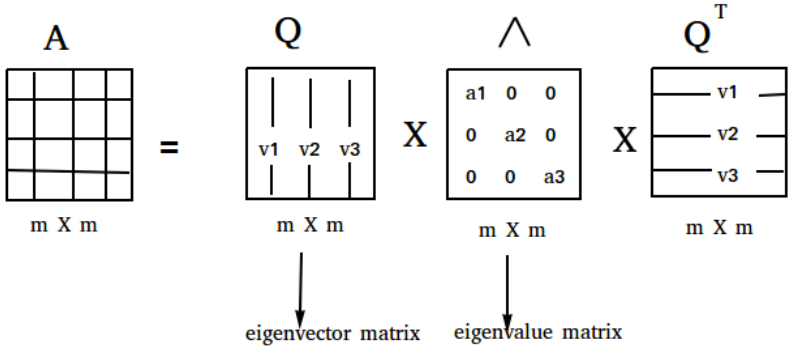
Principal Component Analysis





$$\mathbf{A} = \begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

$$\begin{aligned} \operatorname{cov}(X,Y) &= \operatorname{E}[(X - \operatorname{E}[X])\,(Y - \operatorname{E}[Y])] \\ &= \operatorname{E}[XY - X\operatorname{E}[Y] - \operatorname{E}[X]Y + \operatorname{E}[X]\operatorname{E}[Y]] \\ &= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y] - \operatorname{E}[X]\operatorname{E}[Y] + \operatorname{E}[X]\operatorname{E}[Y] \\ &= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y], \end{aligned}$$



Data analysis and machine learning

- Population (quần thể)
- Tập mẫu
- Tập dữ liệu
- Feature vectors
- Đặc trưng thống kê: min, max, mean, variance, standard deviation
- Hệ số tương quan
- Phân tích biểu đồ: scatter plot, histogram
- Phân tích dự báo: regression (hồi quy) and classification (phân loại)
- **Phân cụm (clustering) và giảm chiều (dimensionality reduction)**