



Bài giảng môn học:

Học Máy (Machine Learning)

CHƯƠNG 4: HỌC KHÔNG GIÁM SÁT (Unsupervised Learning)

Giảng viên: Đặng Văn Nam

Email: dangvannam@hmg.edu.vn

Nội dung chương 4

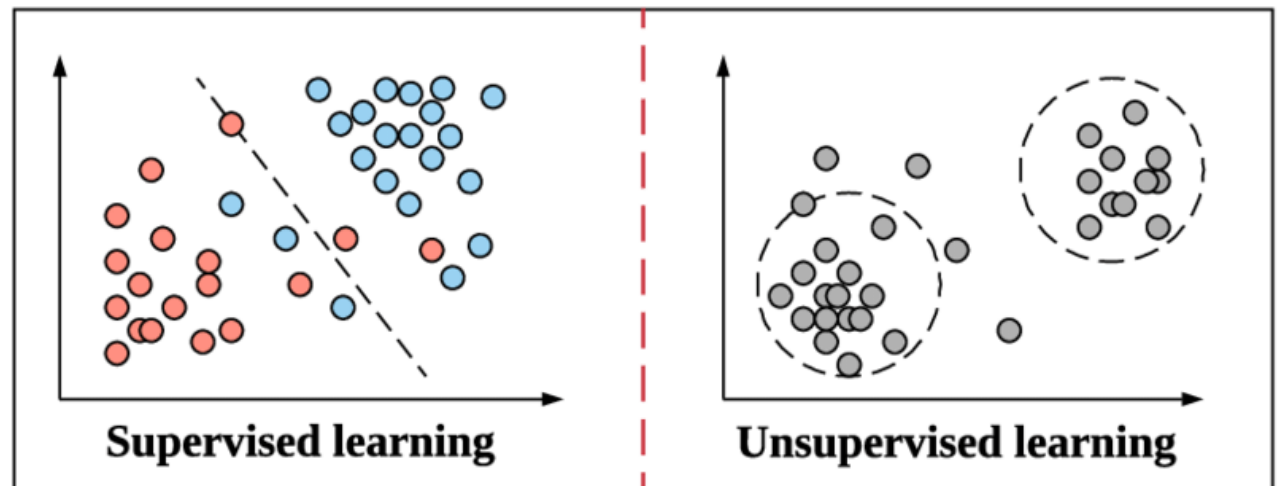
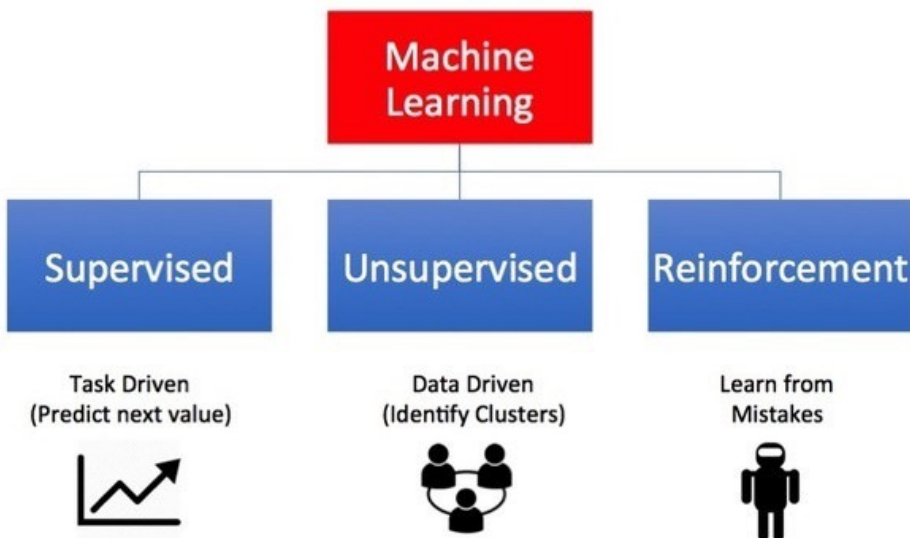
- 1. Tổng quan về học không giám sát**
- 2. Phân cụm dữ liệu (Clustering)**
- 3. Một số thuật toán phân cụm cơ bản**
- 4. Phân cụm khách hàng với KMeans**

1. Học không giám sát

Học không giám sát

- Nếu dữ liệu huấn luyện chỉ bao gồm các dữ liệu đầu vào (Biến độc lập X) mà không có đầu ra (Biến phụ thuộc y) tương ứng. Các thuật toán Học máy có thể không dự đoán được đầu ra nhưng vẫn **trích xuất được những thông tin quan trọng** dựa trên mối liên quan giữa các điểm dữ liệu. Các thuật toán trong nhóm này được gọi là học không giám sát (unsupervised learning).

Types of Machine Learning



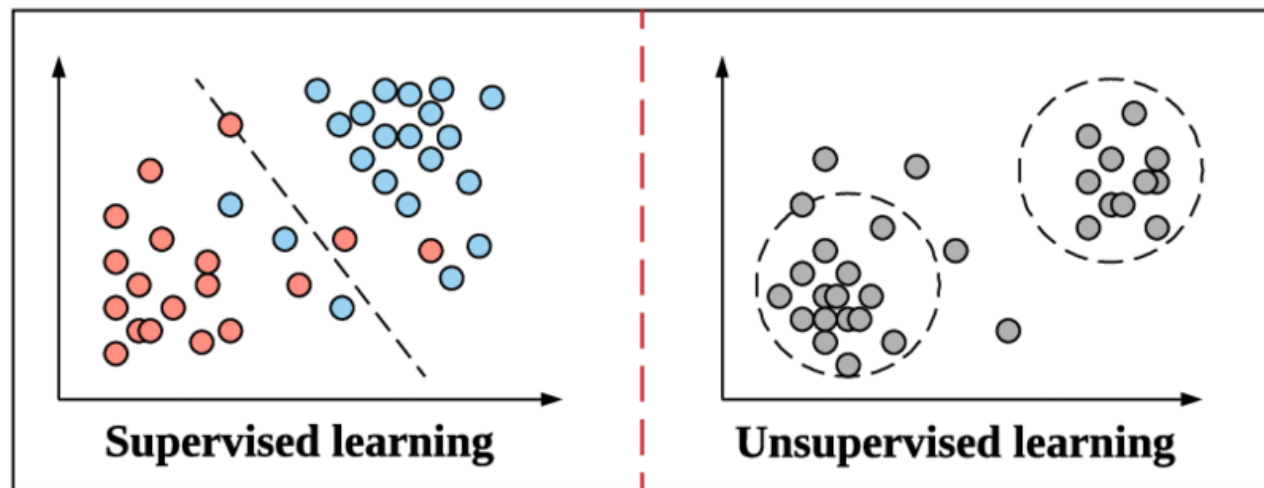
Học không giám sát

- Học máy có giám sát và không có giám sát:

Học máy có giám sát (Supervised learning): Dự đoán đầu ra (label - y) của một dữ liệu mới (new sample) dựa trên các cặp (X, y) đã biết từ trước.

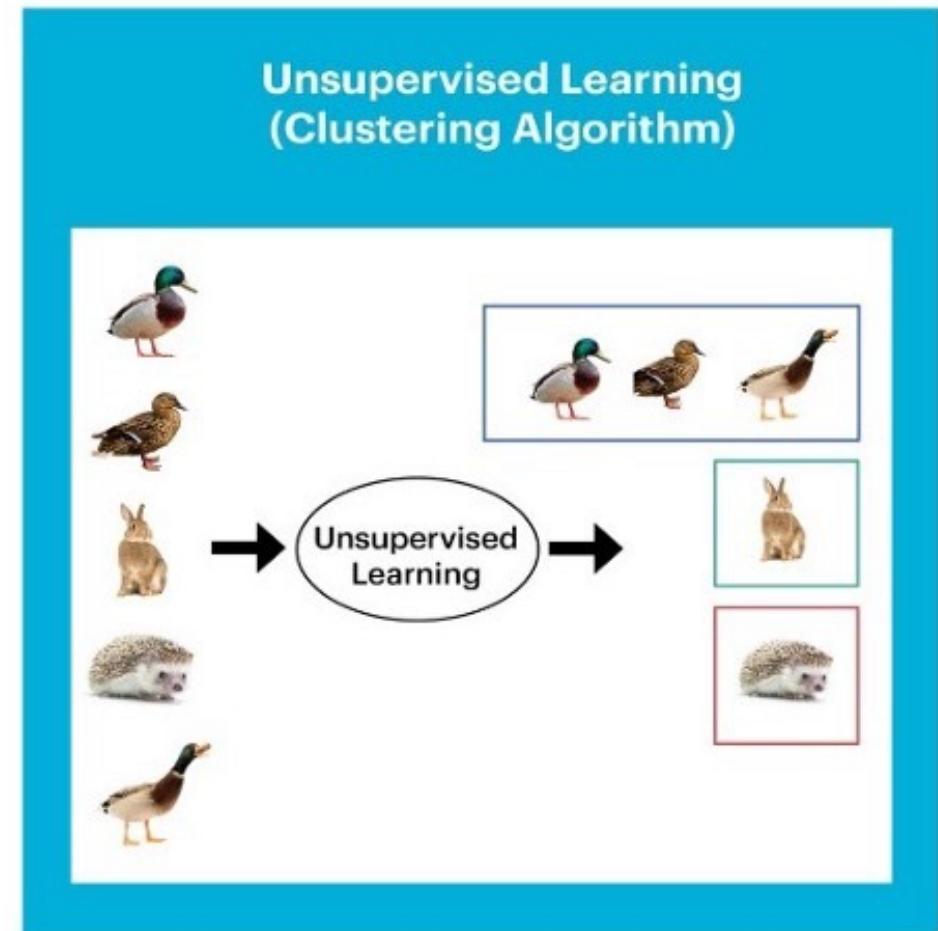
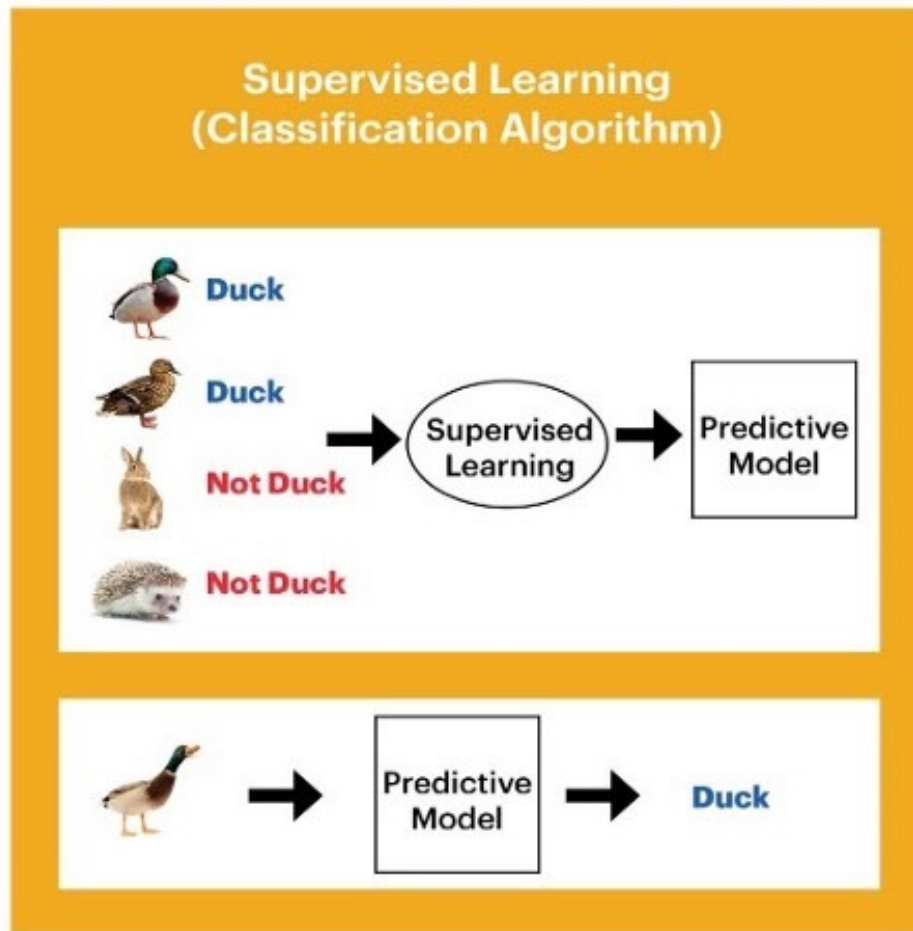
Học máy không giám sát (Unsupervised learning): chỉ có dữ liệu đầu vào (X) mà không có nhãn (label - y).

– Mục đích là **khai phá dữ liệu** để tìm ra các cấu trúc nội tại trong dữ liệu để thực hiện một công việc nào đó, ví dụ như phân nhóm (clustering) hoặc giảm số chiều của dữ liệu (dimension reduction) để thuận tiện trong việc lưu trữ và tính toán.

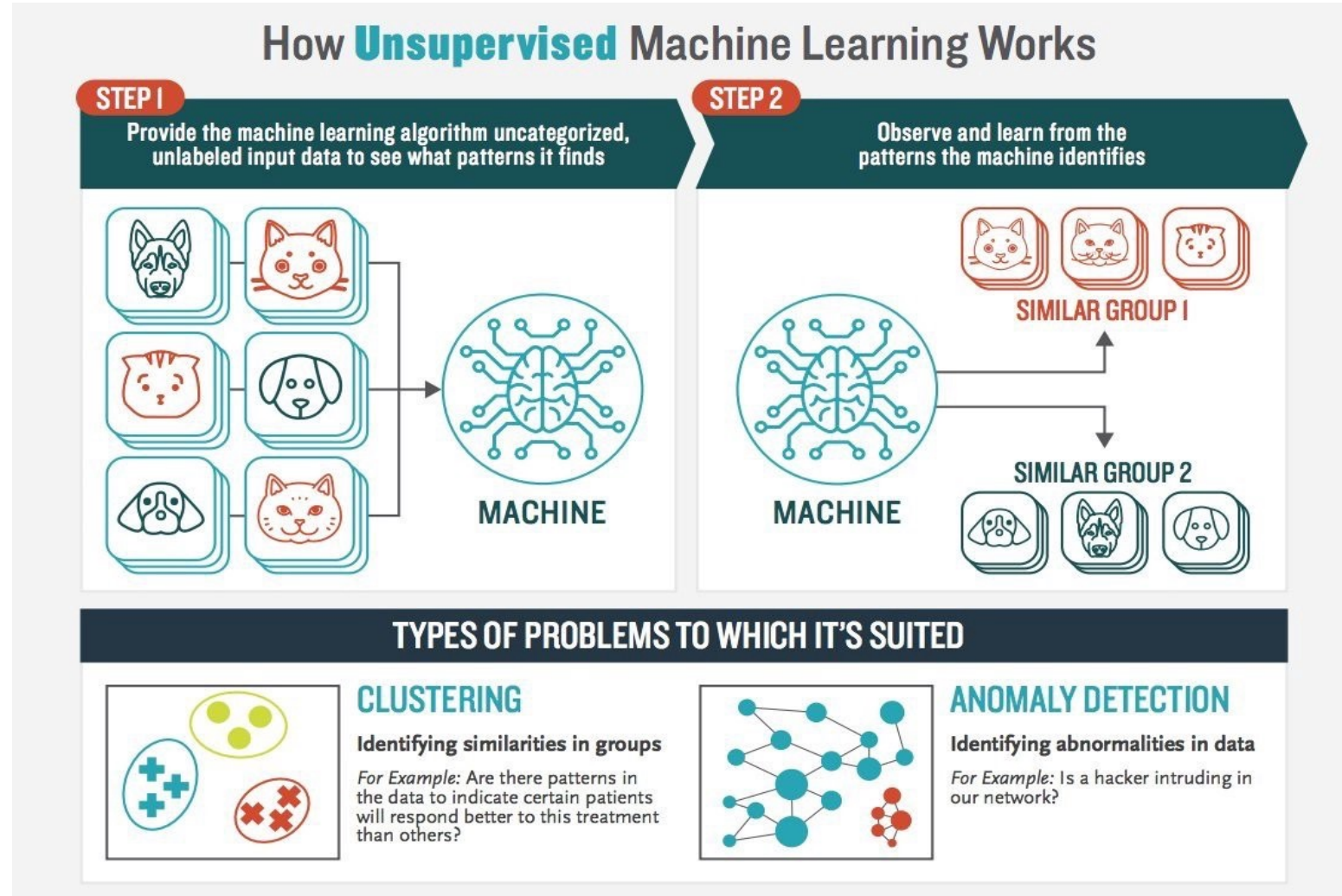


Học không giám sát

Phân biệt giữa Supervised Learning với Unsupervised Learning



Học không giám sát



Các thuật toán phân cụm và phát hiện bất thường là hai ví dụ điển hình trong học không giám sát.

2. Phân cụm dữ liệu

Phân cụm dữ liệu (Clustering)

Phân cụm dữ liệu là một phương pháp xử lý thông tin quan trọng và phổ biến, nó nhằm **khám phá mối liên hệ giữa các mẫu dữ liệu** bằng cách **tổ chức chúng thành các cụm** tương tự.

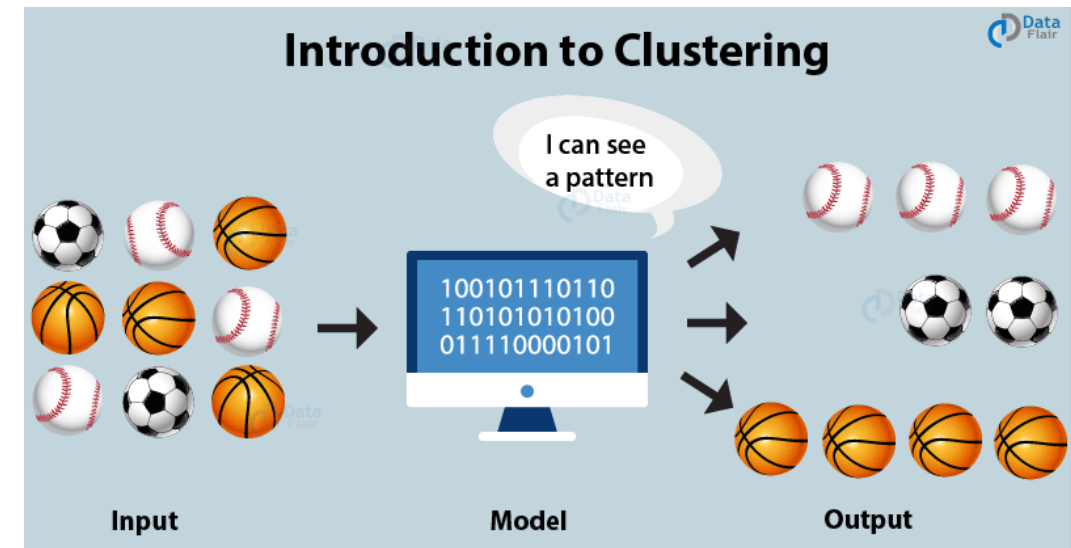
Phân cụm dữ liệu nhằm mục đích chính là khai phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó, cho phép người ta đi sâu vào phân tích và nghiên cứu cho từng cụm dữ liệu này nhằm khám phá và tìm kiếm các thông tin tiềm ẩn, hữu ích phục vụ cho ra quyết định.



Phân cụm dữ liệu (Clustering)

Mục tiêu của phân cụm dữ liệu đó là chia các đối tượng thành các cụm thuần nhất và phân biệt với nhau, tức là thành các nhóm đối tượng **thỏa mãn 2 điều kiện** sau:

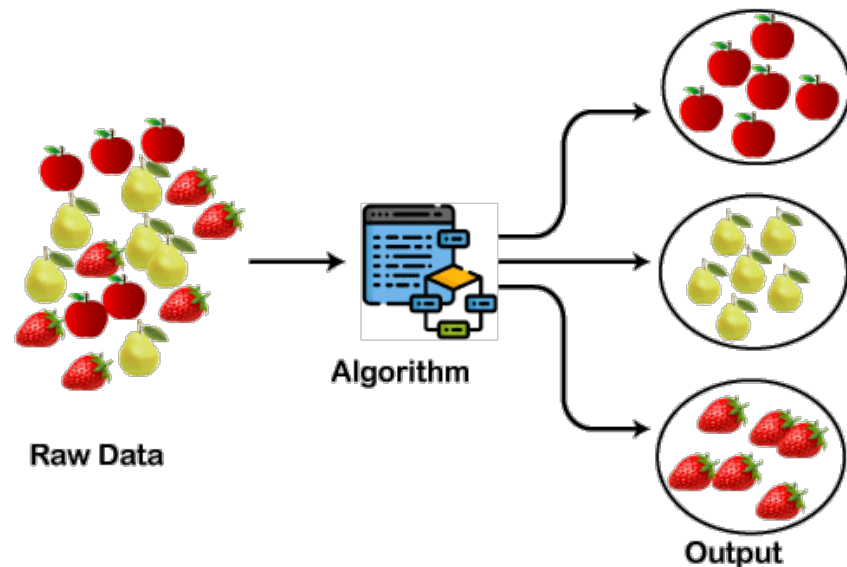
- Độ tương tự của các đối tượng trong mỗi nhóm cao nhất có thể (**tiêu chuẩn liên kết chặt**)
- Các đối tượng trong các nhóm khác nhau phân biệt nhất có thể (**tiêu chuẩn tách rời**)



Phân cụm dữ liệu (Clustering)

Phân cụm được ứng dụng trong nhiều bài toán:

- Market Segmentation
- Statistical data analysis
- Social network analysis
- Image segmentation
- Anomaly detection, etc.



Original image



K-means



C-ACO



C-Bat



C-Cuckoo



C-Firefly

Phân cụm dữ liệu (Clustering)

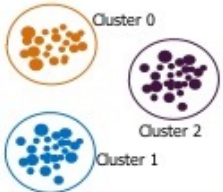
Các loại phân cụm

Types of Clustering

edureka!

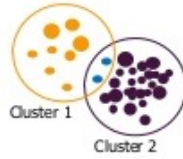
Exclusive Clustering

- An item belongs exclusively to one cluster, not several.
- K-means does this sort of exclusive clustering.



Overlapping Clustering

- An item can belong to multiple clusters
- Its degree of association with each cluster is known
- Fuzzy/ C-means does this sort of exclusive clustering.

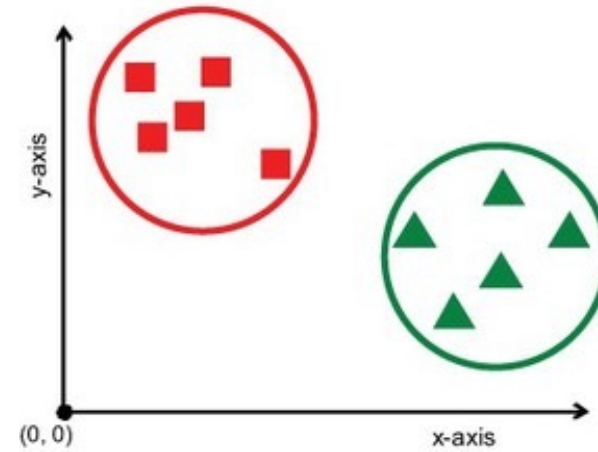


Hierarchical Clustering

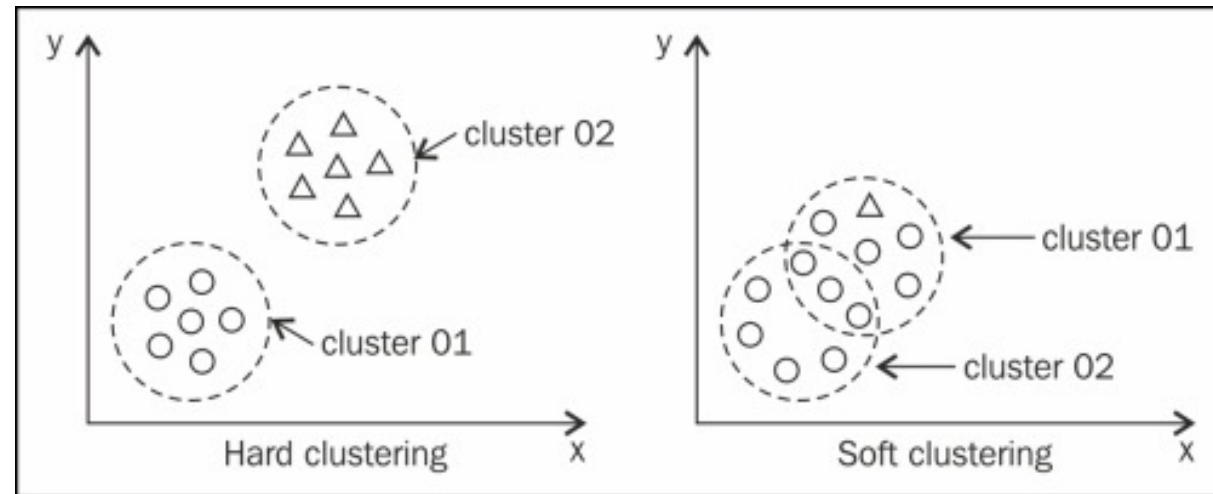
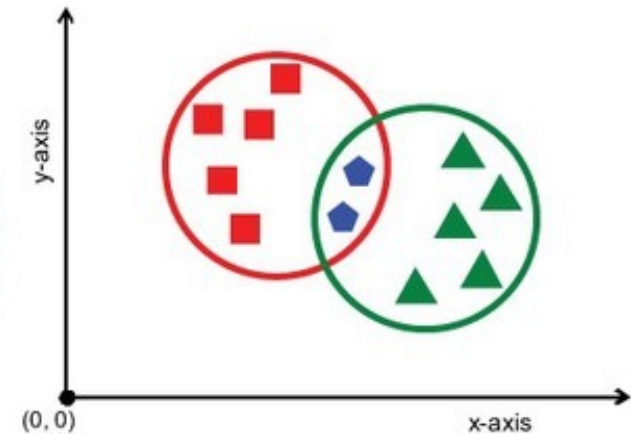
- When two clusters have a parent-child relationship or a tree-like structure then it is Hierarchical clustering



Exclusive clustering



Overlapping clustering

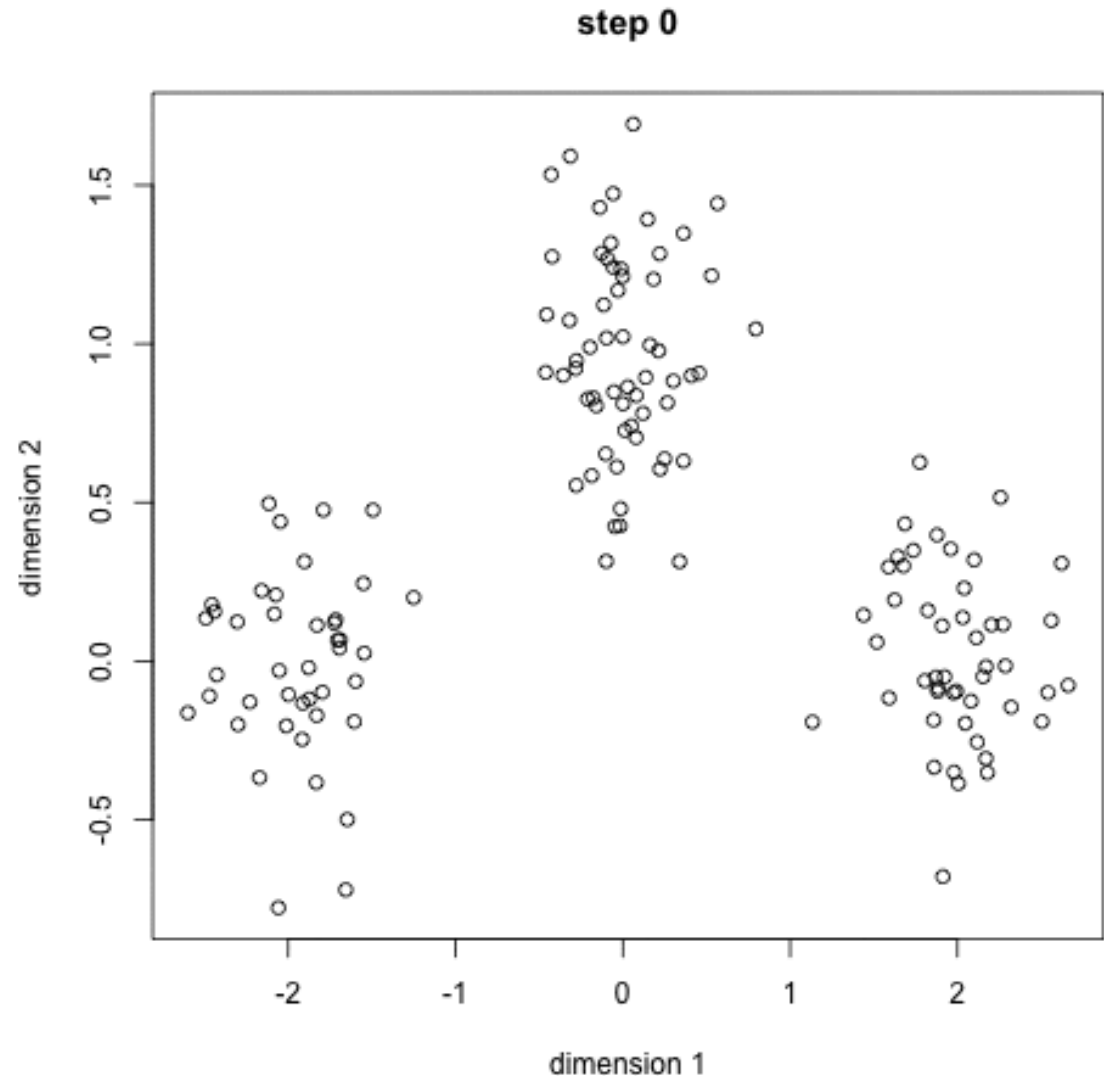


3. Một số thuật toán phân cụm

<https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>

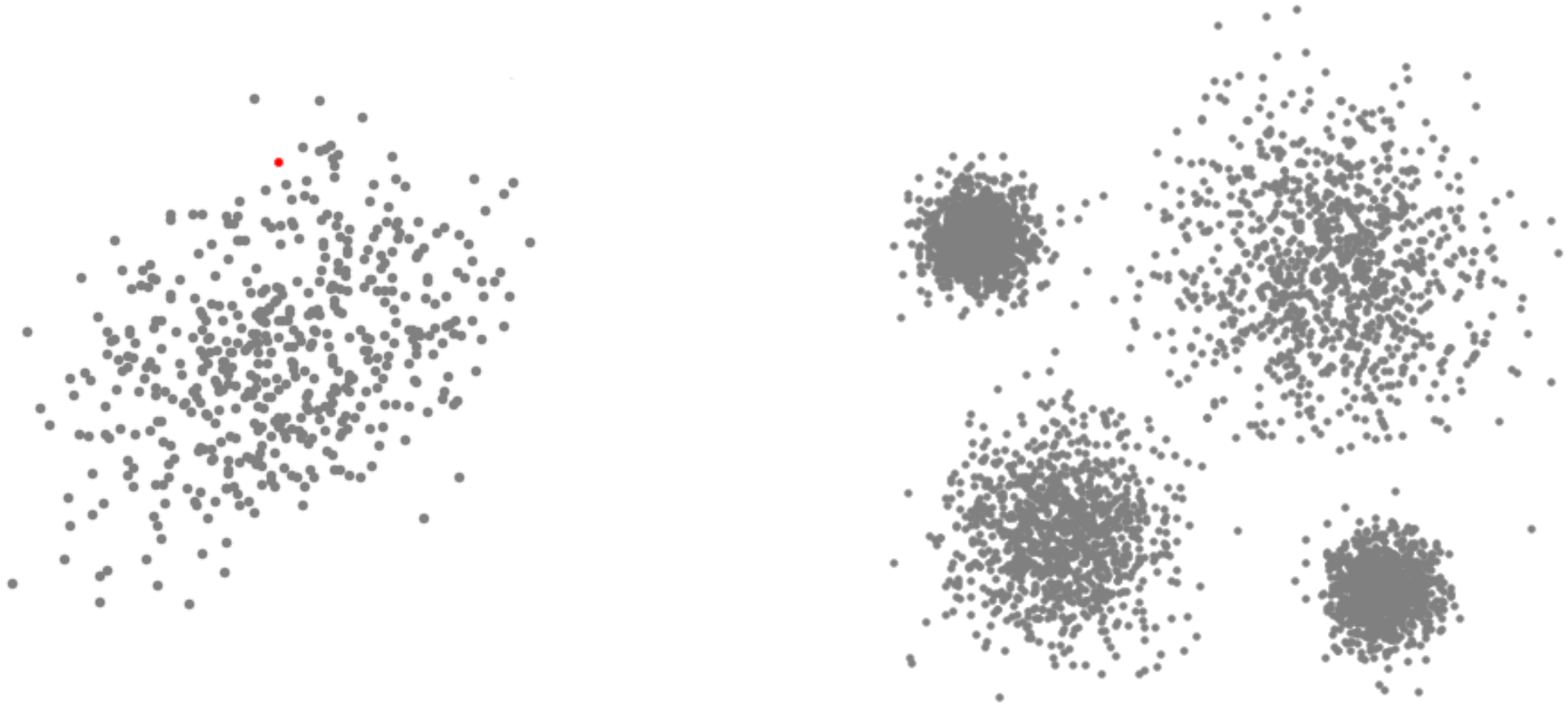
Thuật toán Phân cụm dữ liệu (Clustering)

1. Thuật toán K-Means clustering



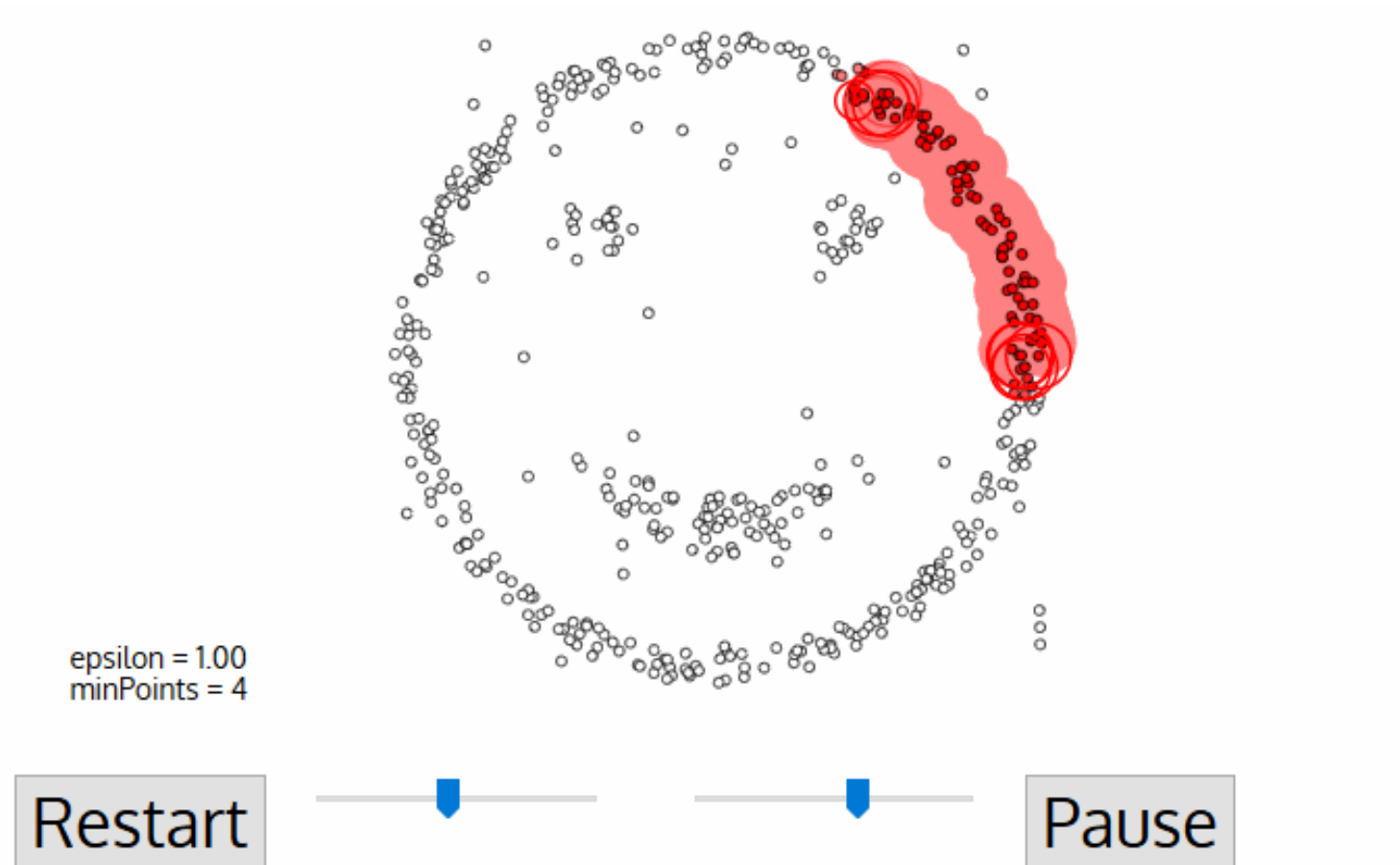
Thuật toán Phân cụm dữ liệu (Clustering)

2. Thuật toán Mean-Shift clustering



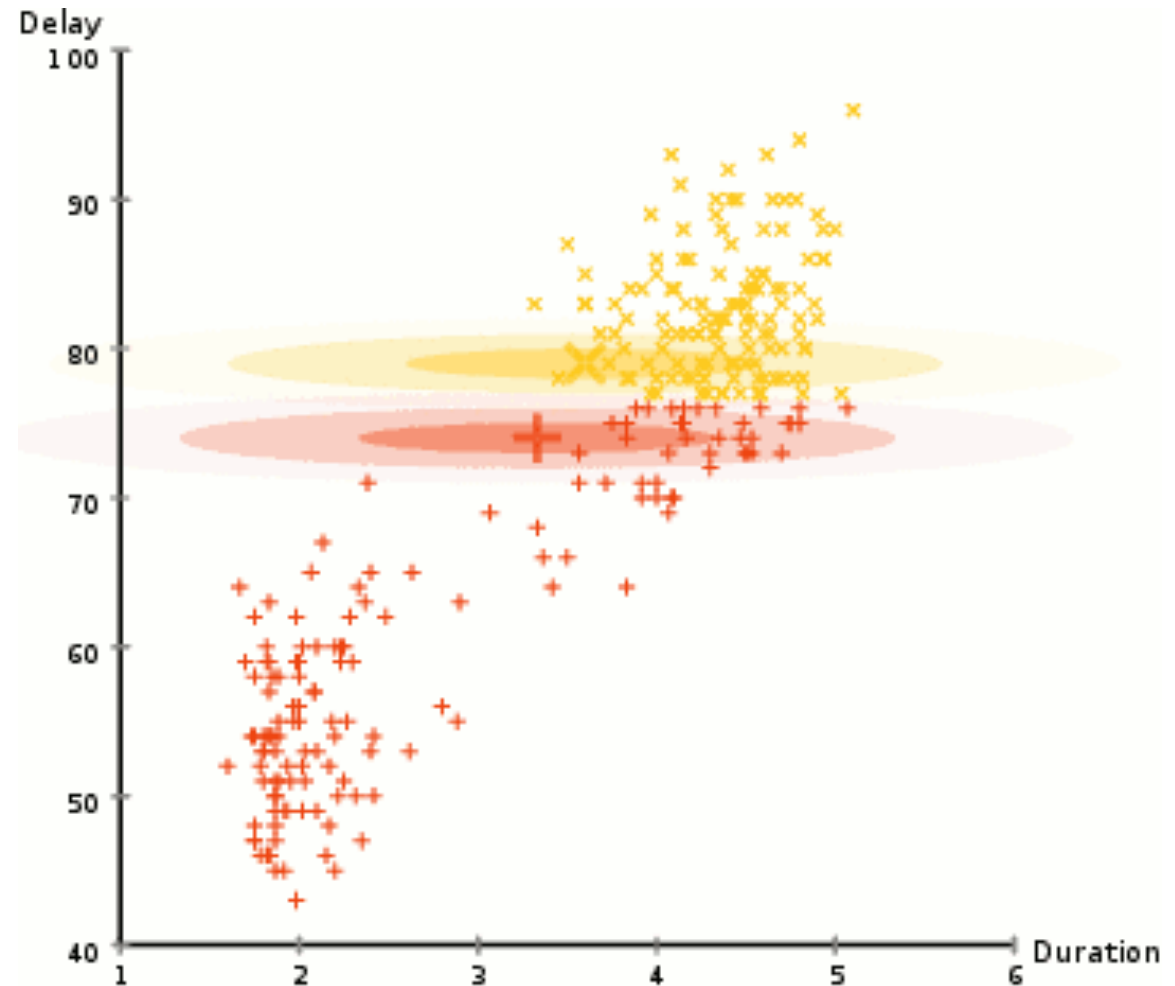
Thuật toán Phân cụm dữ liệu (Clustering)

3. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)



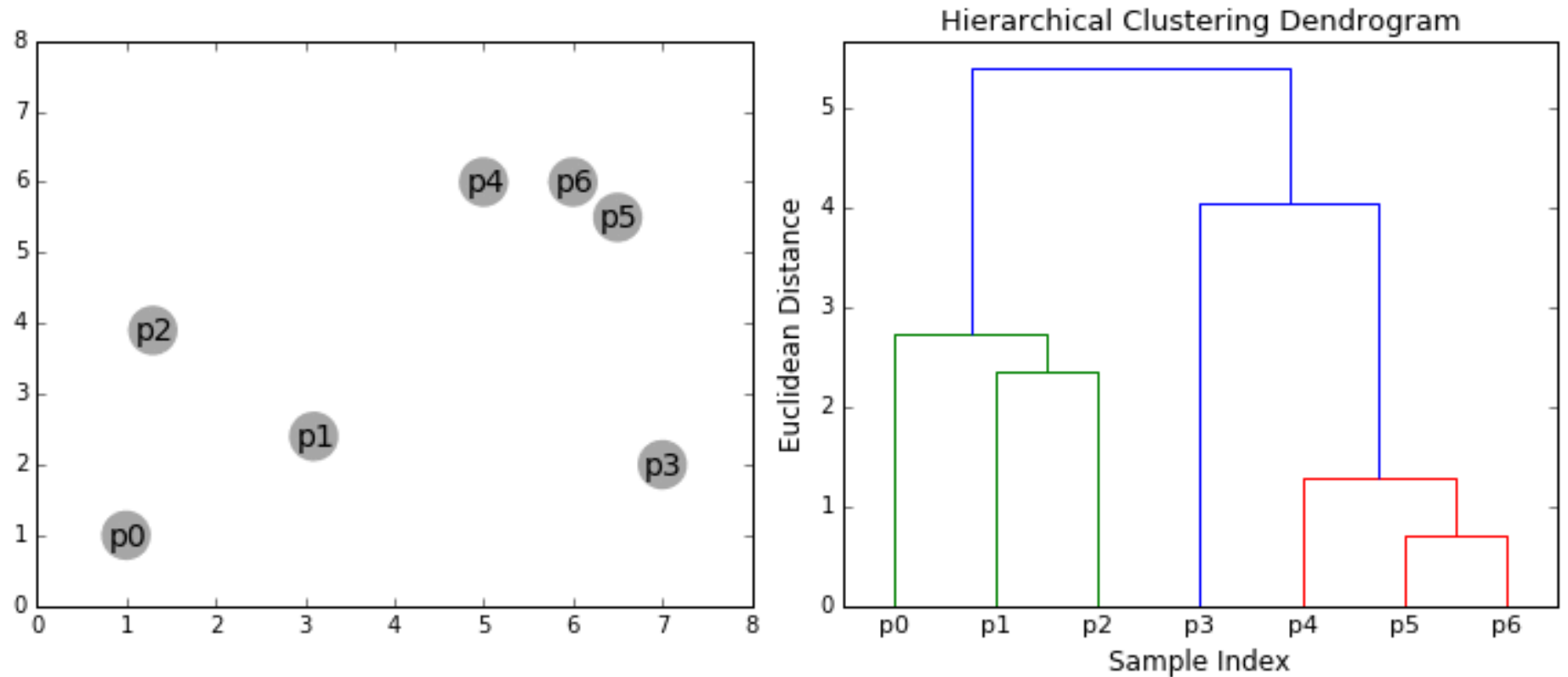
Thuật toán Phân cụm dữ liệu (Clustering)

4. Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)



Thuật toán Phân cụm dữ liệu (Clustering)

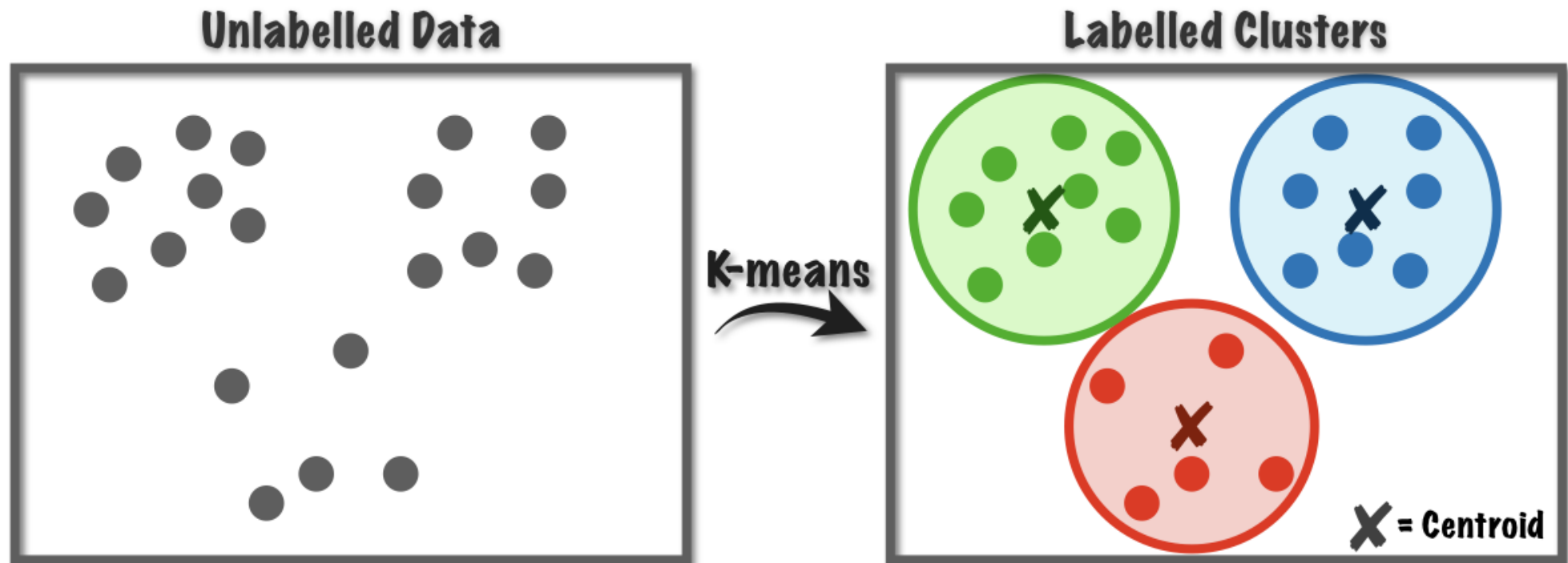
5. Agglomerative Hierarchical Clustering



4. Thuật toán KMeans

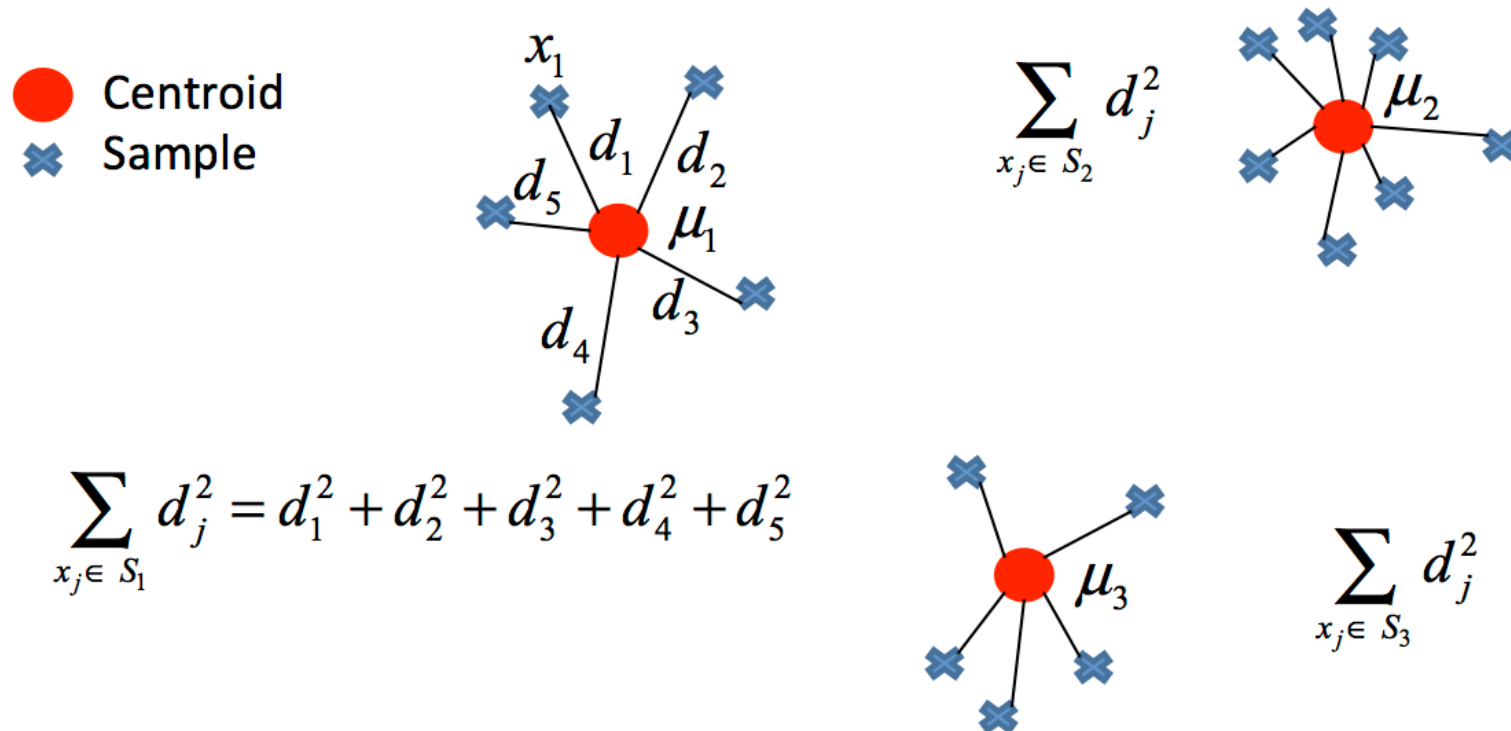
Thuật toán Kmeans

- Kmeans là thuật toán quan trọng và phổ biến trong kỹ thuật phân cụm dữ liệu.
- Ý tưởng chính của thuật toán Kmeans là tìm cách phân nhóm các đối tượng (Objects) đã cho vào k cụm (k là số các cụm được xác định trước, k là số nguyên dương) sao cho tổng bình phương khoảng cách giữa các đối tượng đến tâm nhóm (centroid) là nhỏ nhất.



Thuật toán Kmeans

- Trong các thuật toán gom cụm các điểm được nhóm theo khái niệm “độ gần” hay “độ tương tự”. Với Kmeans, phép đo mặc định cho “độ tương tự” là khoảng cách Euclide.



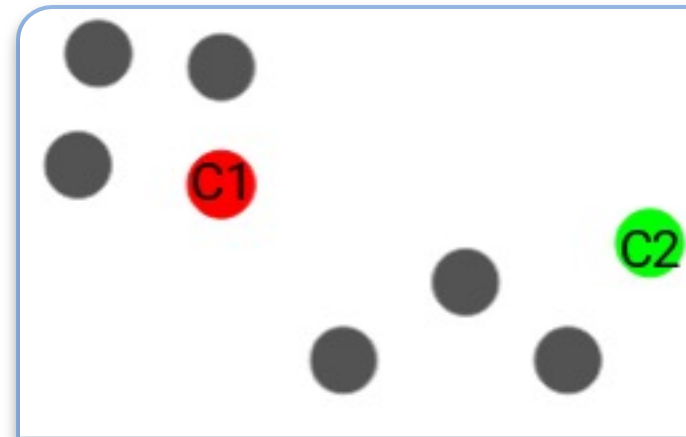
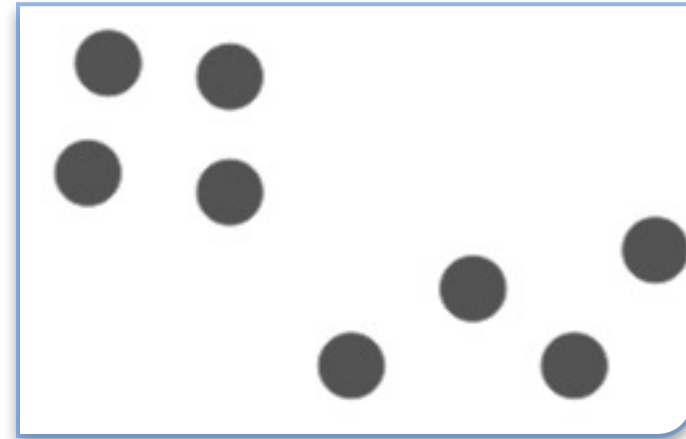
$$\min_S E(\mu_i) = \sum_{x_j \in S_1} d_j^2 + \sum_{x_j \in S_2} d_j^2 + \sum_{x_j \in S_3} d_j^2$$

Mô tả thuật toán Kmeans

- Giả thiết tập dữ liệu gồm 8 điểm, sử dụng thuật toán Kmeans để nhóm các dữ liệu này.

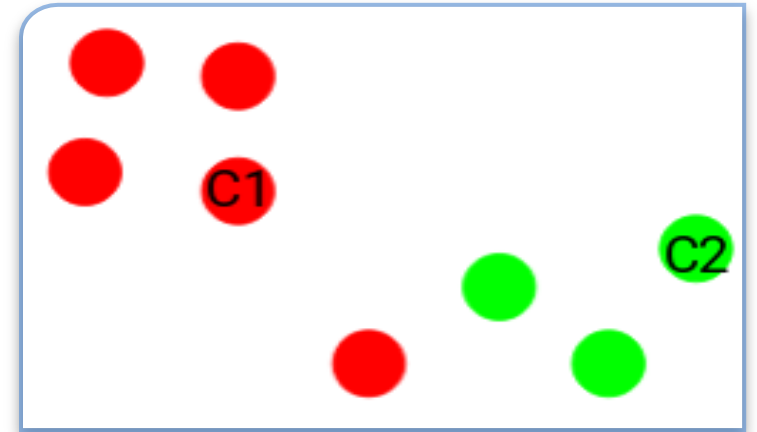
Bước 1: Chọn số cụm k. Giả sử chọn $k=2$

Bước 2: Chọn k điểm ngẫu nhiên làm trọng tâm (Điểm **màu đỏ** và **xanh** là hai trung tâm vừa chọn của 2 cụm)

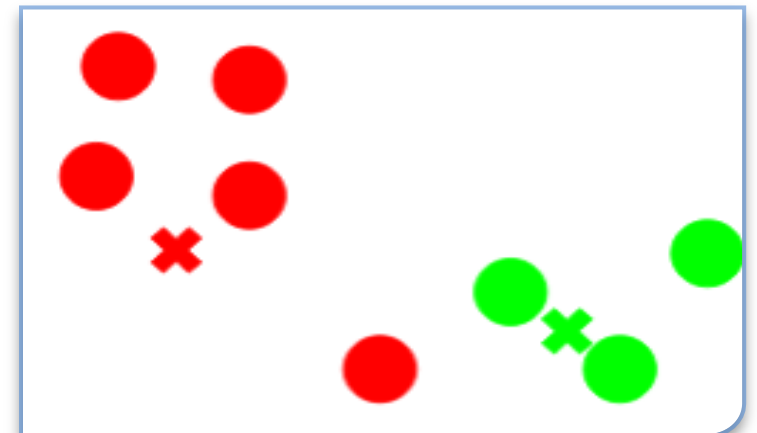


Mô tả thuật toán Kmeans

Bước 3: Gán tất cả các điểm cho trọng tâm gần nhất. Các điểm gần màu đỏ hơn sẽ được chuyển thành màu đỏ, gần màu xanh hơn được chuyển thành màu xanh



Bước 4: Tính toán lại trọng tâm các cụm mới hình thành. (Các dấu nhân màu đỏ và xanh là vị trí trọng tâm mới tương ứng với 2 cụm)

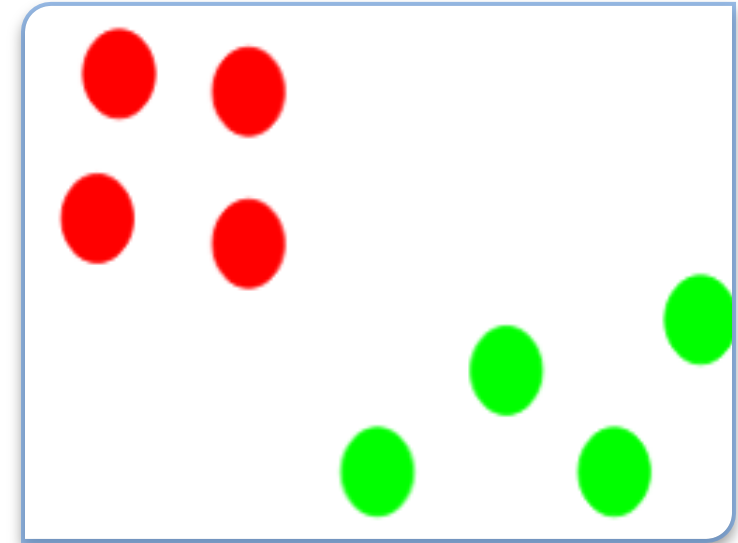


Mô tả thuật toán Kmeans

Bước 5: Lặp lại bước 3 và 4

Vòng lặp sẽ dừng nếu xuất hiện:

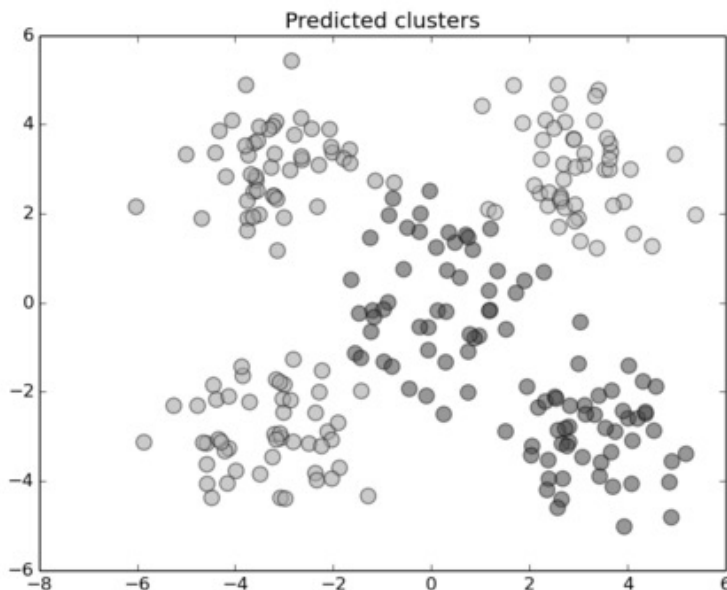
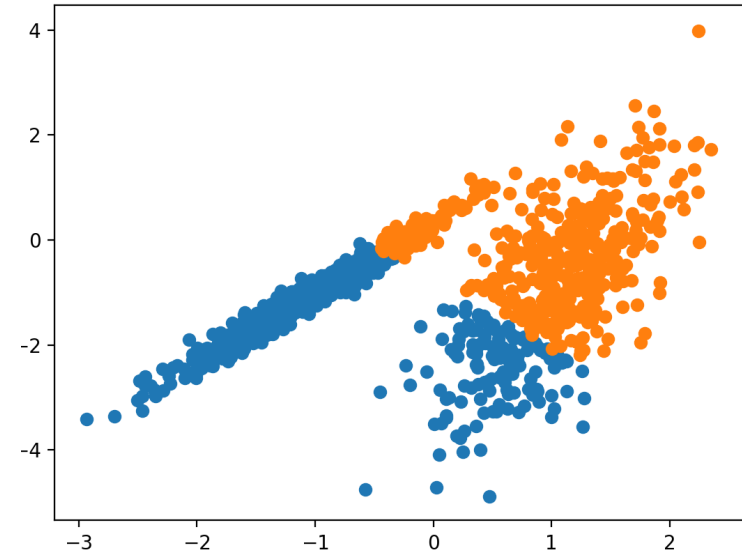
- Các trung tâm của các cụm mới hình thành không thay đổi.
- Các điểm vẫn nằm trong cùng một cụm.
- Đã đạt đến số lần lặp lại tối đa.



Vấn đề với thuật toán Kmeans

Vấn đề 1: Vị trí các điểm trung tâm ban đầu đang được lựa chọn ngẫu nhiên. Các điểm này có thể được chọn quá gần gây ra số vòng lặp tăng lên rất nhiều hoặc có thể bài toán sẽ không có điểm dừng

→ Kmeans++



Vấn đề 2: Với Kmeans số cụm k phải khai báo trước. Vậy làm sao để có thể lựa chọn được số cụm k phù hợp nhất?



5. Ví dụ

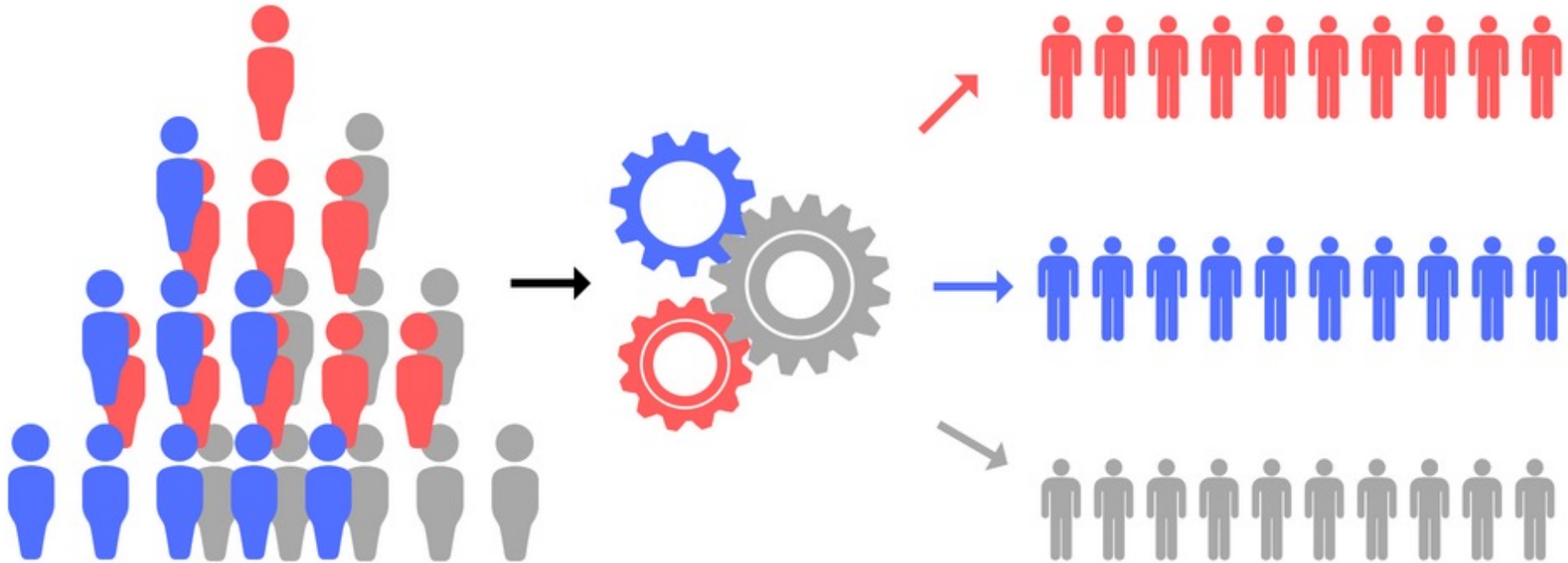
Giới thiệu bài toán

Nhân dịp 10 năm thành lập, một trung tâm thương mại muốn có những ưu đãi cho khách hàng của mình. Tuy nhiên trung tâm thương mại chưa biết lên kế hoạch ưu đãi thế nào cho phù hợp với khách hàng vì mỗi khách hàng có nhu cầu nhận ưu đãi không giống nhau. Và bài toán đặt ra là từ một số dữ liệu khách hàng mà trung tâm thương mại có được đưa ra được những chiến dịch ưu đãi phù hợp cho khách hàng.

Dĩ nhiên không thể lên kế hoạch ưu đãi cho từng khách hàng được vì số lượng khách hàng rất lớn, làm vậy mất rất nhiều thời gian và công sức. **Vậy trung tâm thương mại có thể làm gì?**



Giới thiệu bài toán

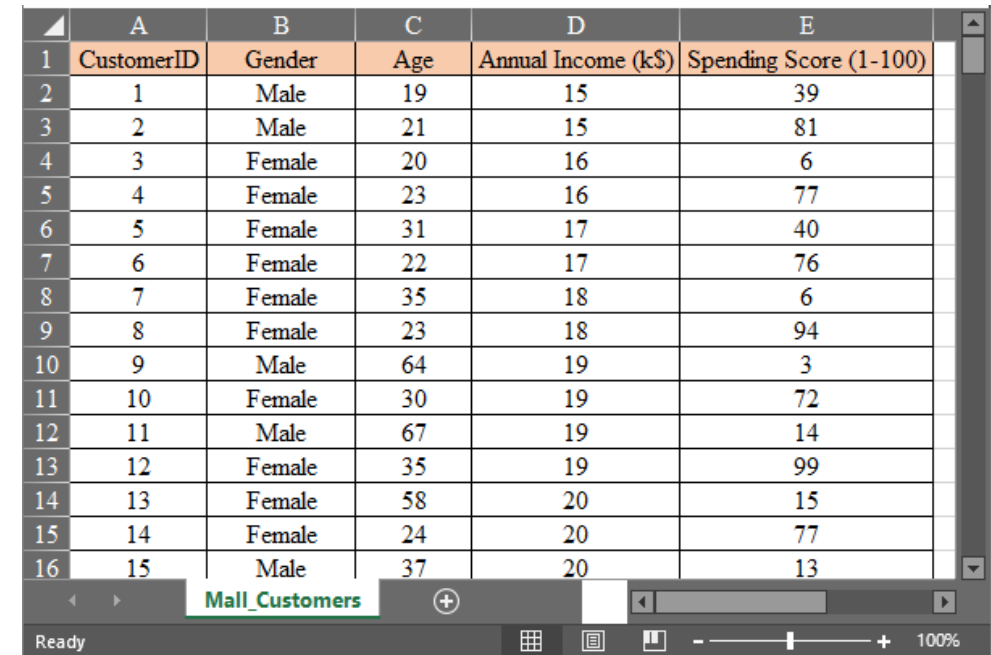


Lựa chọn đưa ra là hãy phân khách hàng thành các **nhóm** khách hàng khác nhau.
→ Thay vì đưa ưu đãi cho từng khách hàng thì bây giờ trung tâm chỉ cần đưa ra chiến lược ưu đãi cho từng nhóm khách hàng đó.

Giới thiệu bài toán

Mô tả tập dữ liệu **Mall_Customers.xlsx**: Tập dữ liệu gồm 215 khách hàng với một số thuộc tính:

- **Gender**: Giới tính của khách hàng
- **Age**: Tuổi của khách hàng
- **Income**: Thu nhập hàng năm của khách hàng (x1000 USD)
- **Spending score**: Điểm chi tiêu do trung tâm mua sắm chỉ định dựa trên hành vi chi tiêu của khách hàng với thang điểm từ 1-100

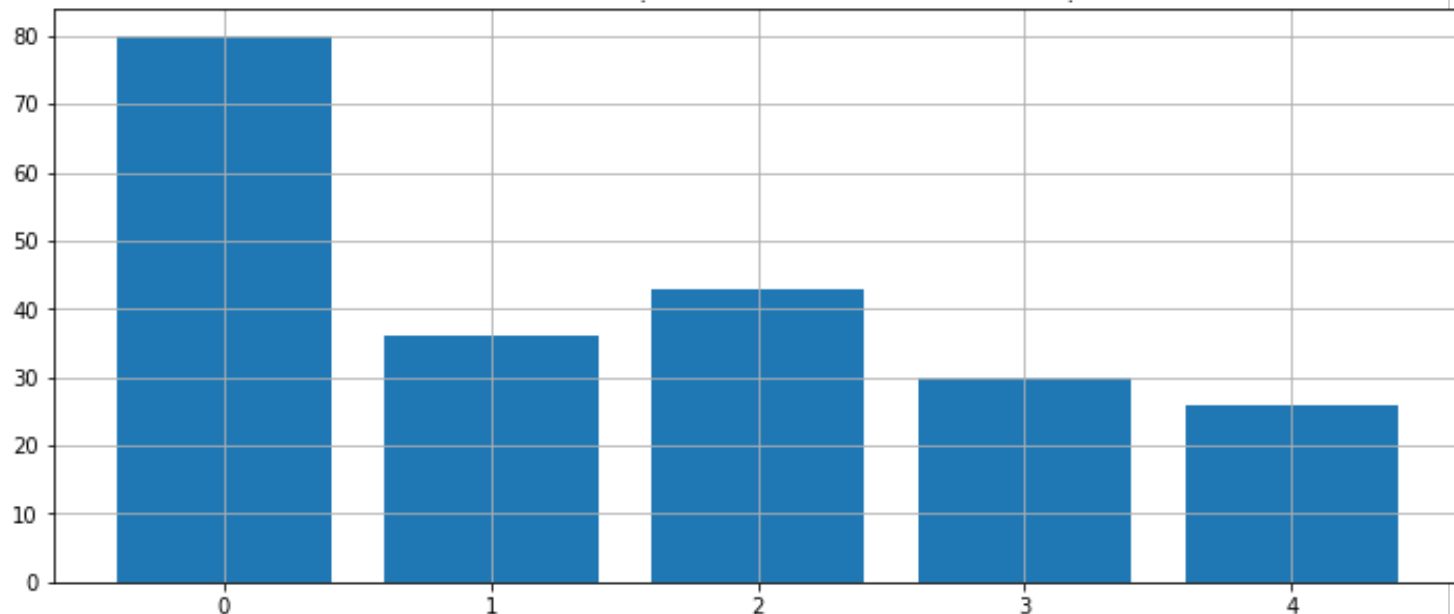


	A	B	C	D	E
1	CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
2	1	Male	19	15	39
3	2	Male	21	15	81
4	3	Female	20	16	6
5	4	Female	23	16	77
6	5	Female	31	17	40
7	6	Female	22	17	76
8	7	Female	35	18	6
9	8	Female	23	18	94
10	9	Male	64	19	3
11	10	Female	30	19	72
12	11	Male	67	19	14
13	12	Female	35	19	99
14	13	Female	58	20	15
15	14	Female	24	20	77
16	15	Male	37	20	13

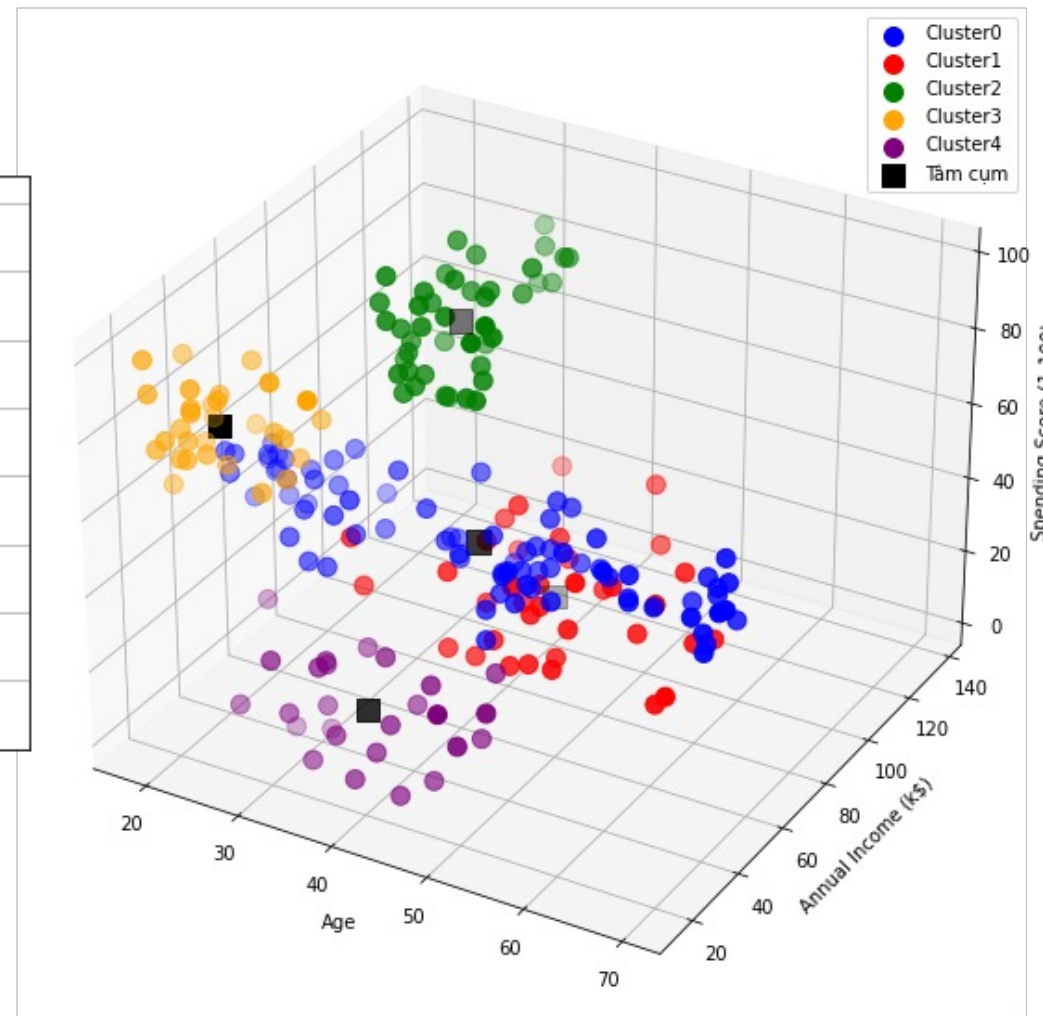
Giới thiệu bài toán

Sử dụng 3 thuộc tính: Age, Income, Score để phân cụm dữ liệu

THỐNG KÊ SỐ LƯỢNG KHÁCH HÀNG THEO TỪNG CỤM



Theo dõi các bước trong file code trên Jupiter Notebook!





THỰC HÀNH

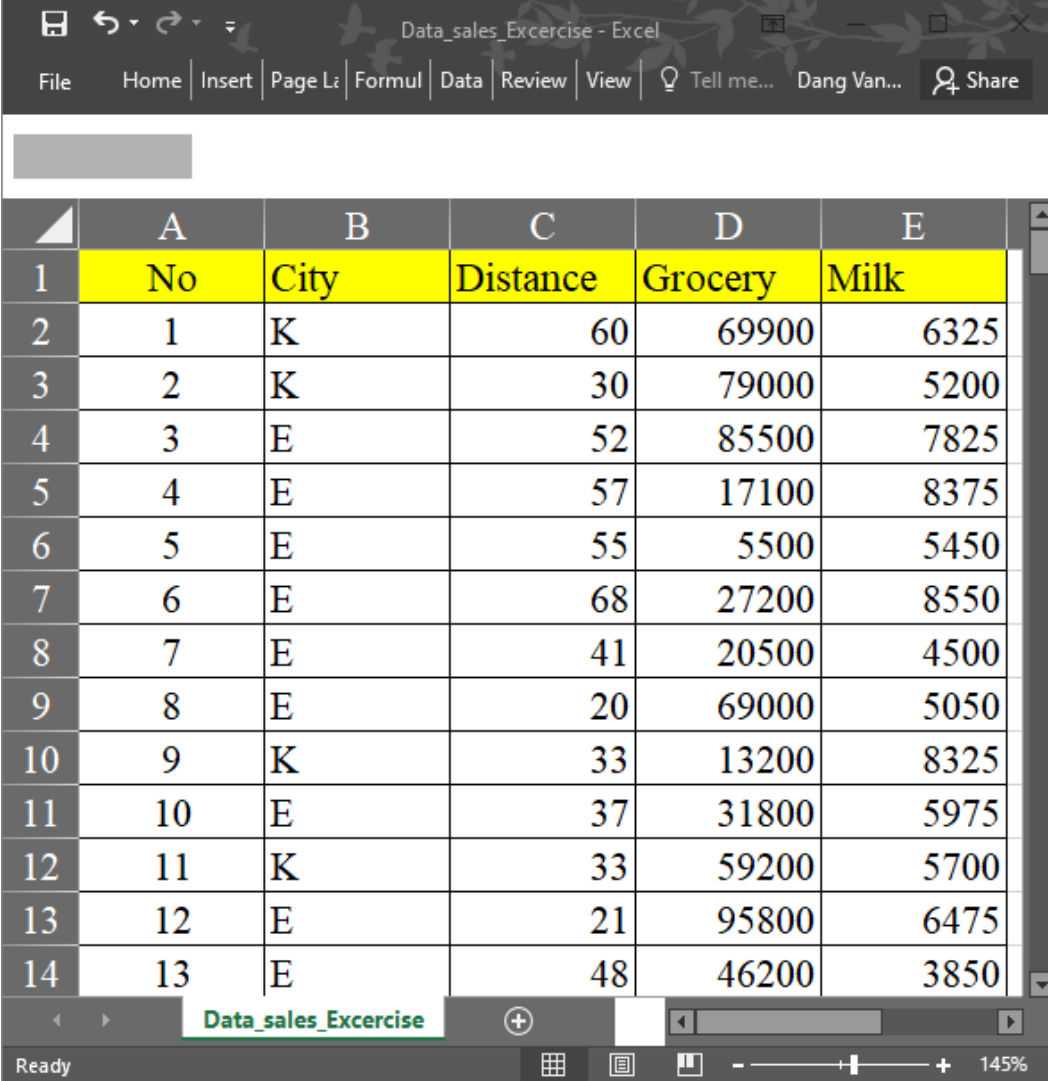
Yêu cầu.

Mô tả tập dữ liệu: **Data_sales_Excercise.csv**

Dữ liệu mua hàng của 200 MiniMart với một nhà phân phối hàng hóa trong năm 2019.

Trong đó:

- **Cột No:** Mã của MiniMart
- **Cột City:** Ký hiệu khu vực đặt MiniMart
- **Cột Distance:** Khoảng cách từ MiniMart tới nhà phân phối.
- **Cột Grocery:** Số tiền MiniMart đã dùng để mua hàng tạp hóa trong năm 2019.
- **Cột Milk:** Số tiền mà MiniMart đã dùng để mua sữa từ nhà phân phối trong năm 2019



The screenshot shows an Excel spreadsheet titled "Data_sales_Excercise - Excel". The spreadsheet contains a table with 6 columns and 14 rows of data. The first row is highlighted in yellow. The columns are labeled: No, City, Distance, Grocery, and Milk. The data rows show the following values:

	A	B	C	D	E
1	No	City	Distance	Grocery	Milk
2	1	K	60	69900	6325
3	2	K	30	79000	5200
4	3	E	52	85500	7825
5	4	E	57	17100	8375
6	5	E	55	5500	5450
7	6	E	68	27200	8550
8	7	E	41	20500	4500
9	8	E	20	69000	5050
10	9	K	33	13200	8325
11	10	E	37	31800	5975
12	11	K	33	59200	5700
13	12	E	21	95800	6475
14	13	E	48	46200	3850

Yêu cầu.

Sử dụng thuật toán phân cụm Kmeans cho tập dữ liệu với 2 thuộc tính phân cụm:
Grocery, Milk:

1. Sử dụng phương pháp khuỷu tay (Elbow) xác định số cụm tối ưu.
2. Thực hiện phân cụm dữ liệu với số cụm tối ưu đã chỉ ra ở yêu cầu 1. Trực quan hóa kết quả phân cụm, Liệt kê danh sách MiniMart theo từng cụm và cho nhận xét.





Thank you!