

HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.



ĐẠI HỌC
BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

Một số mô hình phân cụm trong học máy không giám sát

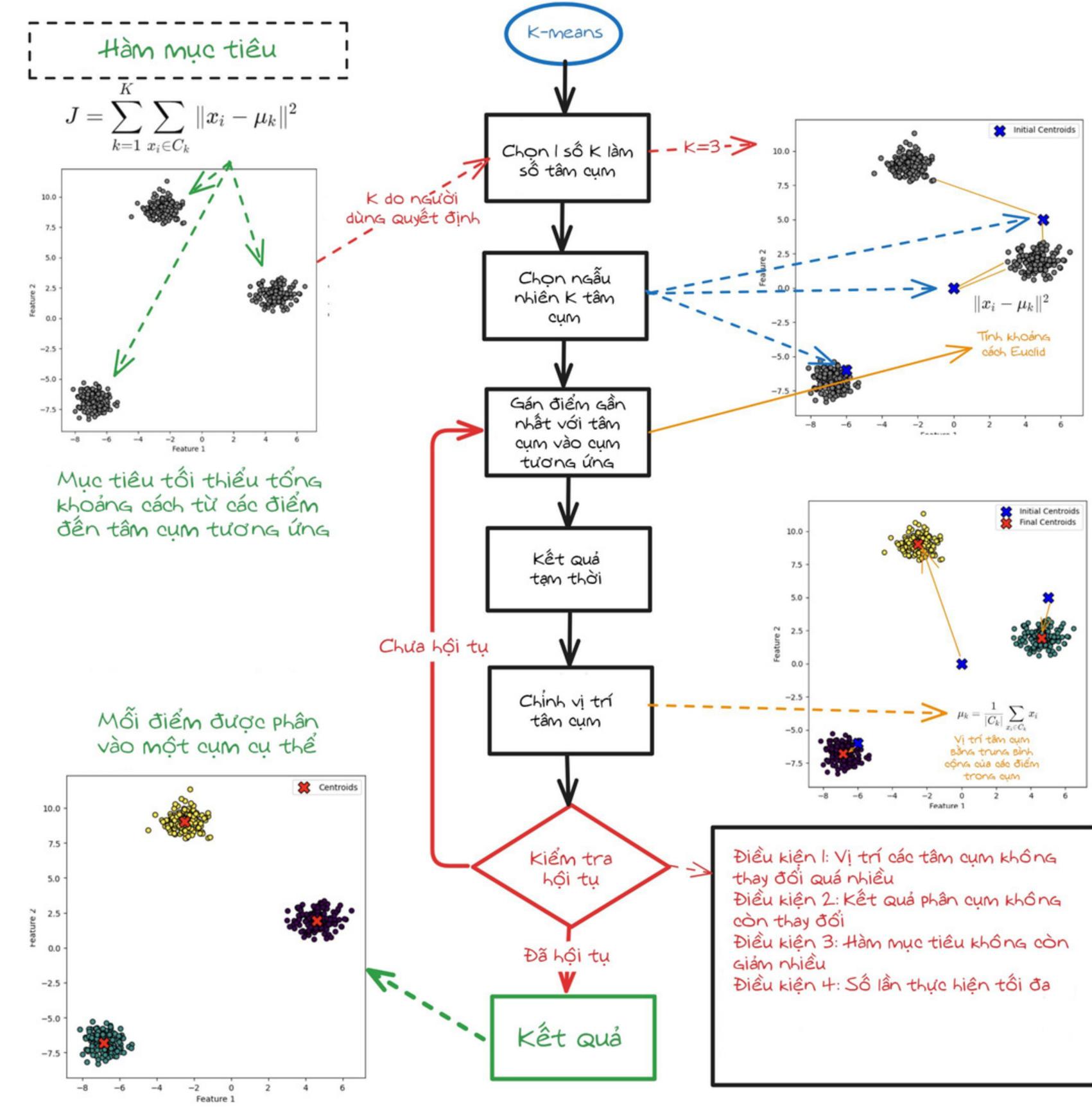
ONE LOVE. ONE FUTURE.

Nội dung chính

- Các mô hình phân cụm:
 - K-means Clustering
 - Agglomerative Clustering
 - DBSCAN
 - GridClus
 - Gaussian Mixture Model
- Ứng dụng trong bài toán thực tế

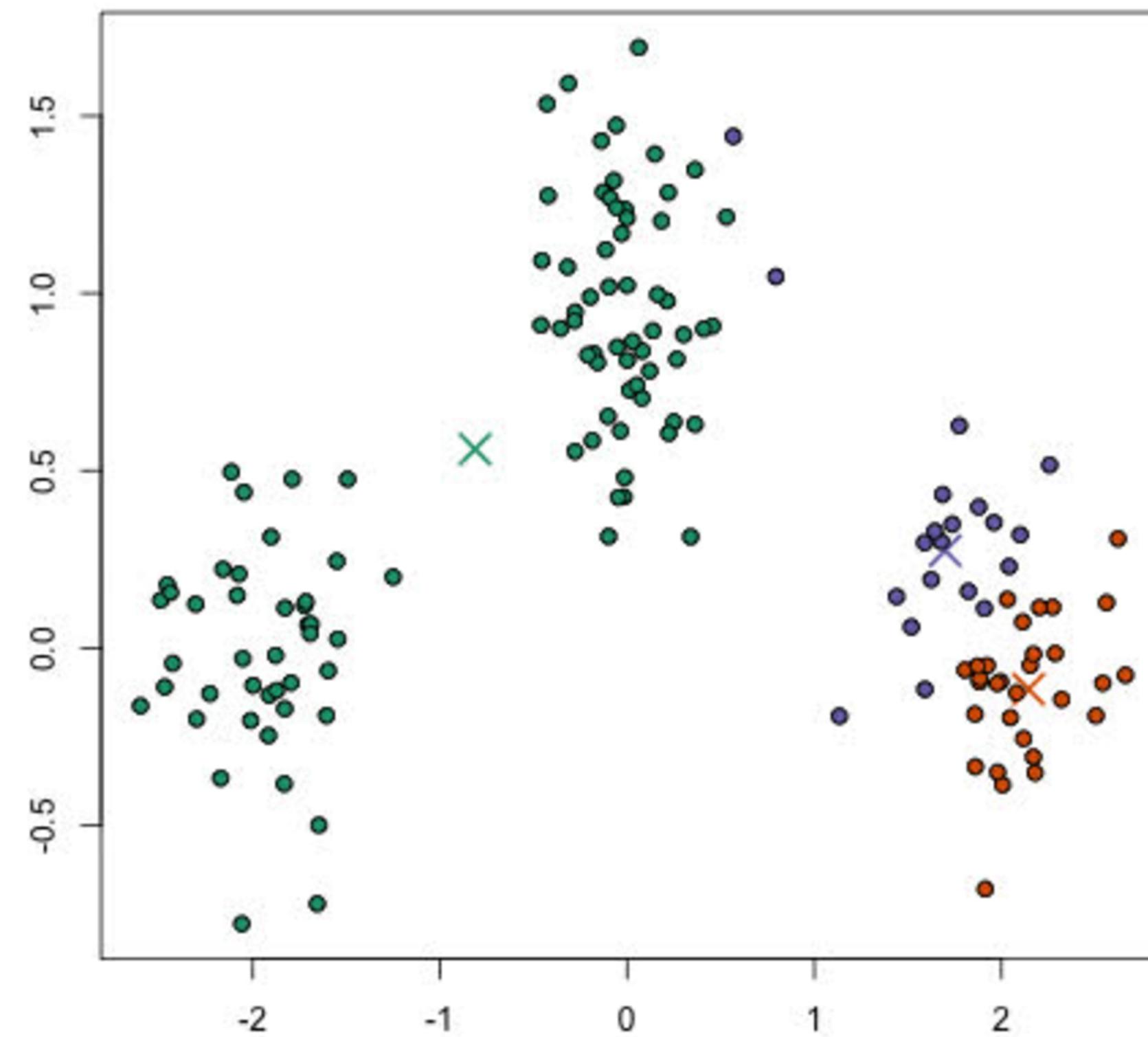


K-Means Clustering



K-Means Clustering

step 6



K-Means Clustering

- Phương pháp xác định số lượng cụm K
 - Elbow Method: Đo tổng bình phương khoảng cách trong cụm. Số cụm càng tăng giá trị càng giảm, nhưng sau một ngưỡng, mức giảm trở nên không đáng kể, tạo nên điểm gãy khúc
 - Silhouette Score: Đo lường chất lượng phân cụm

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$



K-Means Clustering

- Các biến thể tùy theo cách định nghĩa “khoảng cách”

- K-means: $\mu_k = \arg \min_{\mu} \sum_{x_i \in C_k} \|x_i - \mu\|^2$

- K-medians: $\mu_k = \arg \min_{\mu} \sum_{x_i \in C_k} \|x_i - \mu\|$

- K-medoids:

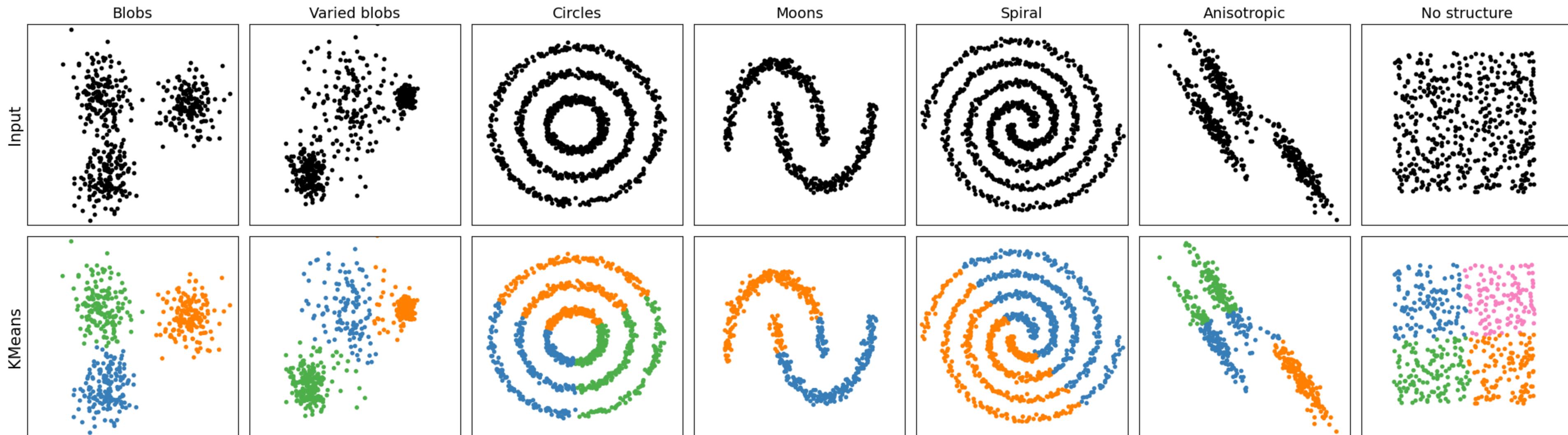


K-Means Clustering

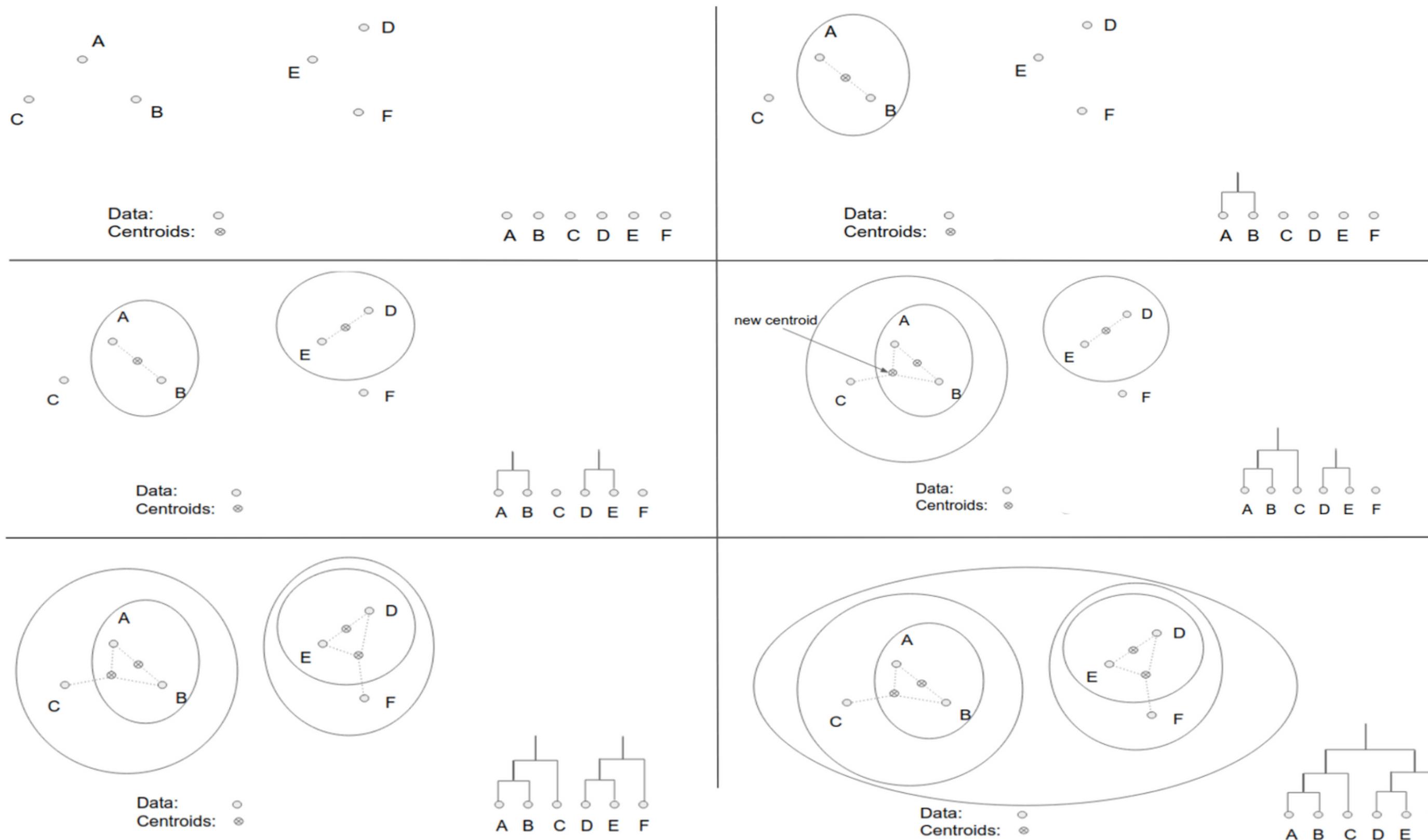
- K-means clustering được ứng dụng trong các bài toán phân khúc khách hàng, xử lý ảnh, phân cụm văn bản,...
- Ưu điểm của K-means:
 - Đơn giản, dễ cài đặt, dễ hiểu
 - Tốc độ nhanh, hiệu quả
 - Hiệu quả trên dữ liệu có cụm dạng cầu
- Nhược điểm của K-means:
 - Phải chọn trước số cụm K, phụ thuộc vào khởi tạo tâm cụm ban đầu
 - Nhạy với nhiễu
 - Chỉ tốt cho cụm dạng cầu



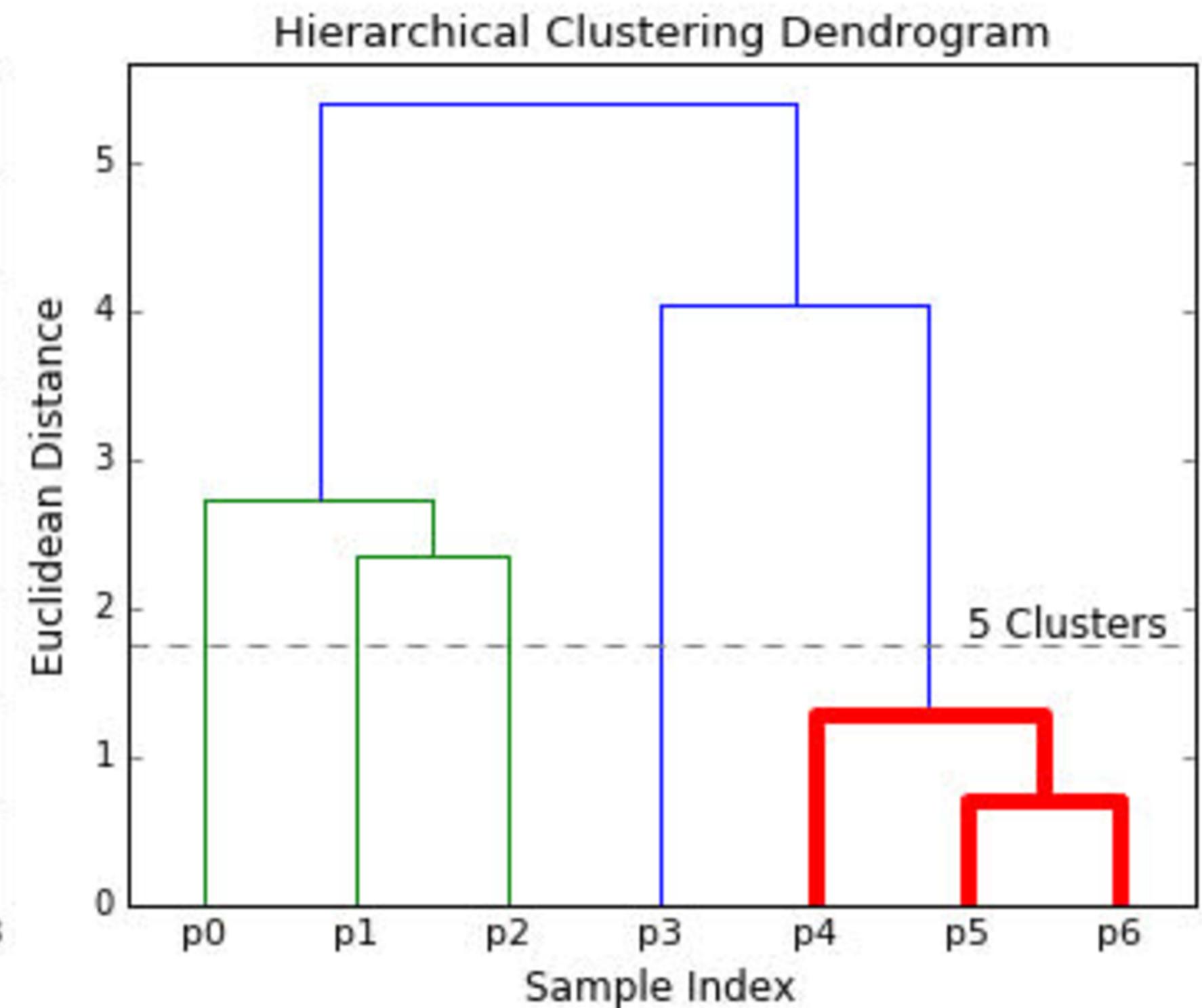
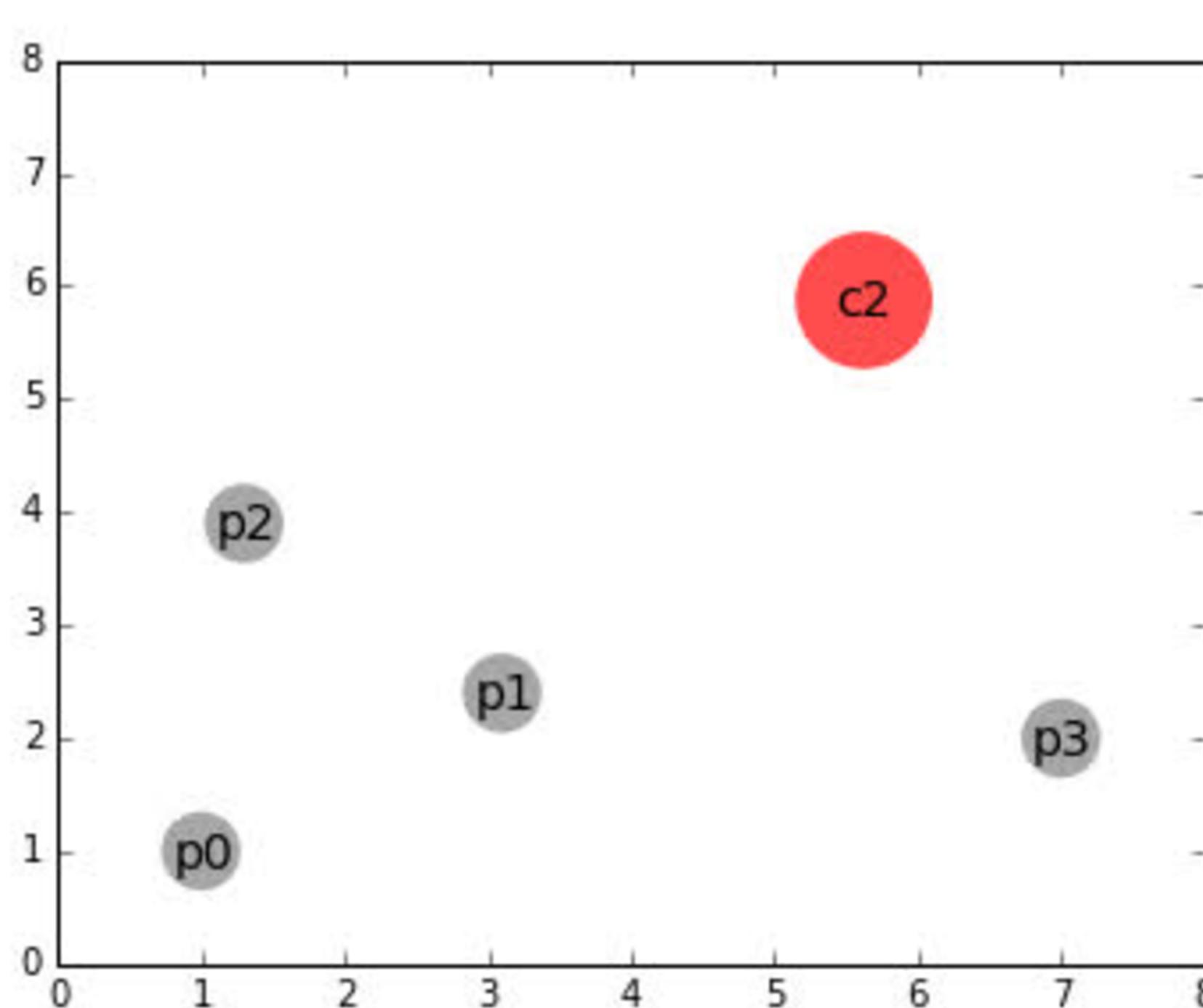
K-Means Clustering



Agglomerative Clustering



Agglomerative Clustering

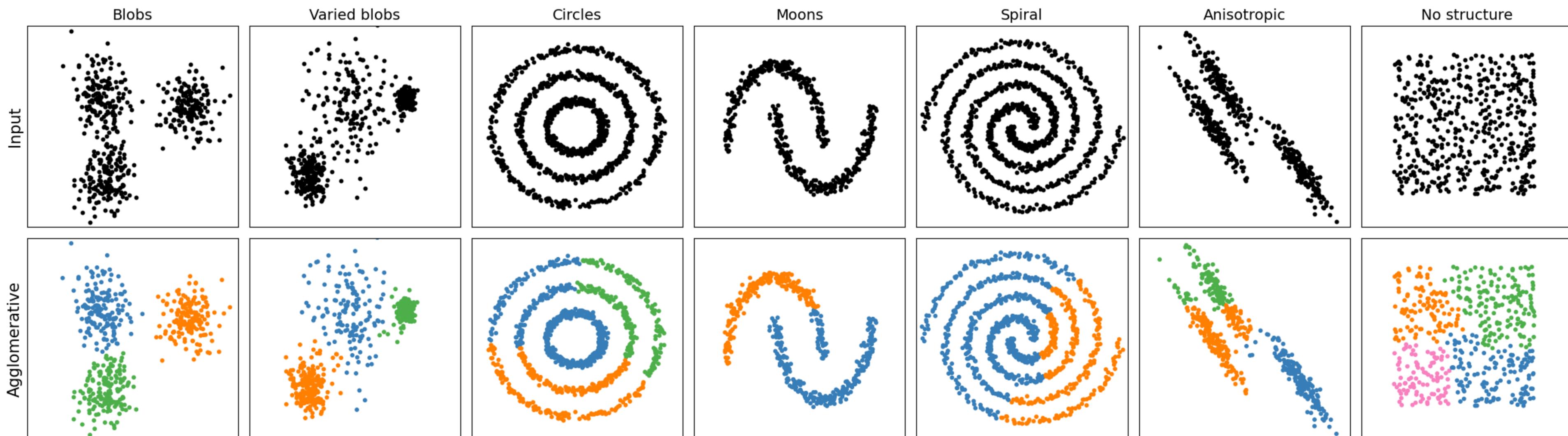


Agglomerative Clustering

- Agglomerative Clustering thường dùng khi muốn khám phá cấu trúc dữ liệu đa cấp, đặc biệt khi không biết số cụm trước
- Ưu điểm của Agglomerative Clustering:
 - Không cần biết số cụm trước
 - Trực quan dễ giải thích, hiển thị cấu trúc phân cấp
- Nhược điểm của K-means:
 - Tốn kém thời gian tính toán
 - Dễ bị kẹt tại tối ưu cục bộ
 - Khó chọn số cụm K chính xác nếu dendrogram không rõ ràng

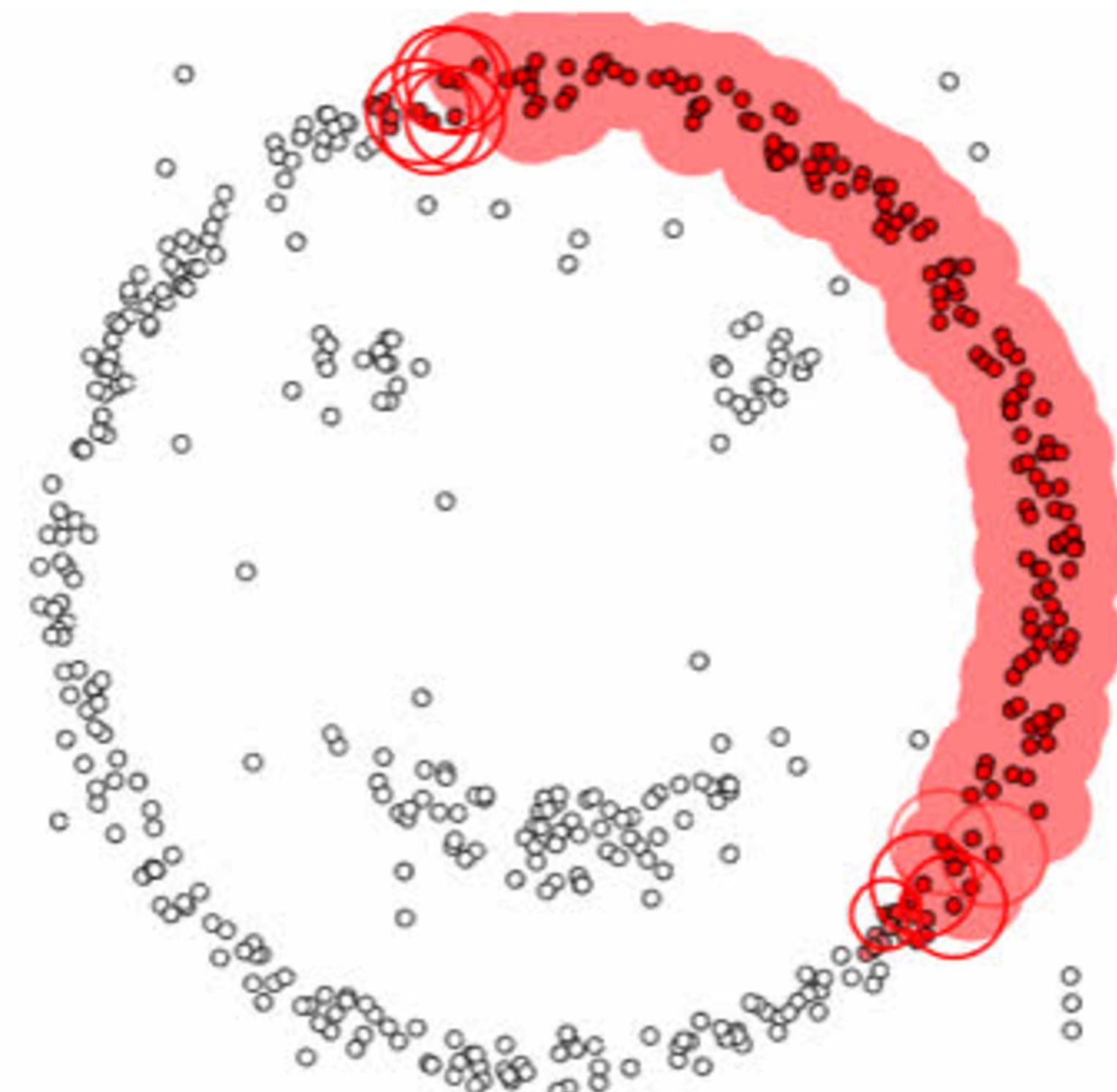


Agglomerative Clustering



- DBSCAN phân cụm dựa trên mật độ điểm
 - Các điểm mật độ cao vào cùng một cụm
 - Điểm nằm ở vùng mật độ thấp được xem là nhiễu
- Một điểm dữ liệu trong DBSCAN được phân làm 3 loại:
 - Điểm lõi: Có ít nhất MinPts điểm (bao gồm chính nó) trong bán kính nhất định
 - Điểm biên: Không phải điểm lõi nhưng nằm trong vùng của một điểm lõi
 - Điểm nhiễu: Vừa không phải điểm lõi, vừa không phải điểm biên

DBSCAN

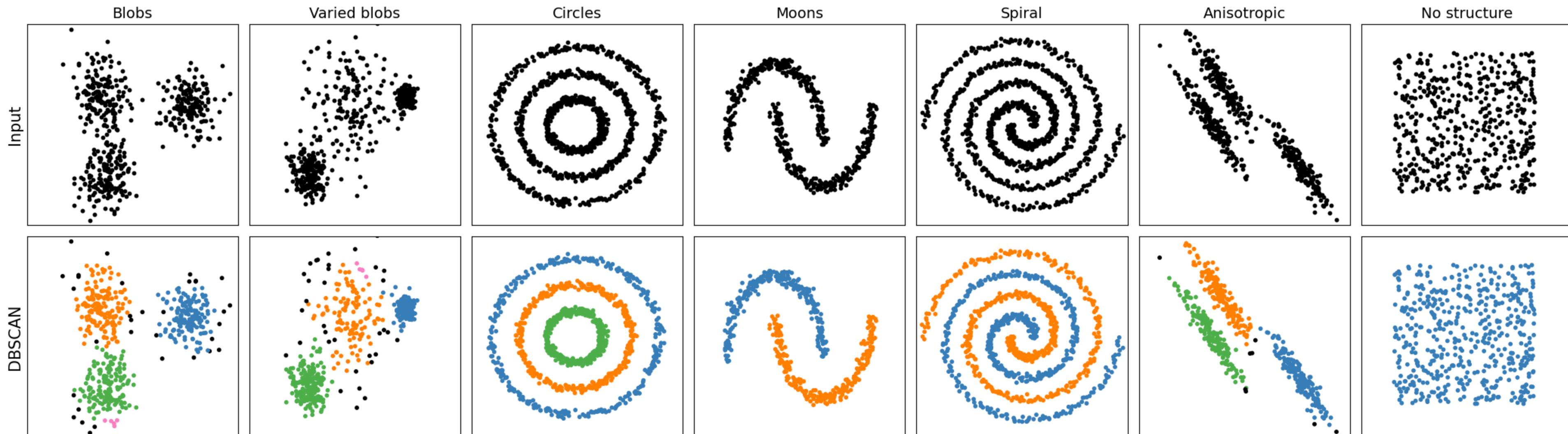


epsilon = 1.00
minPoints = 4

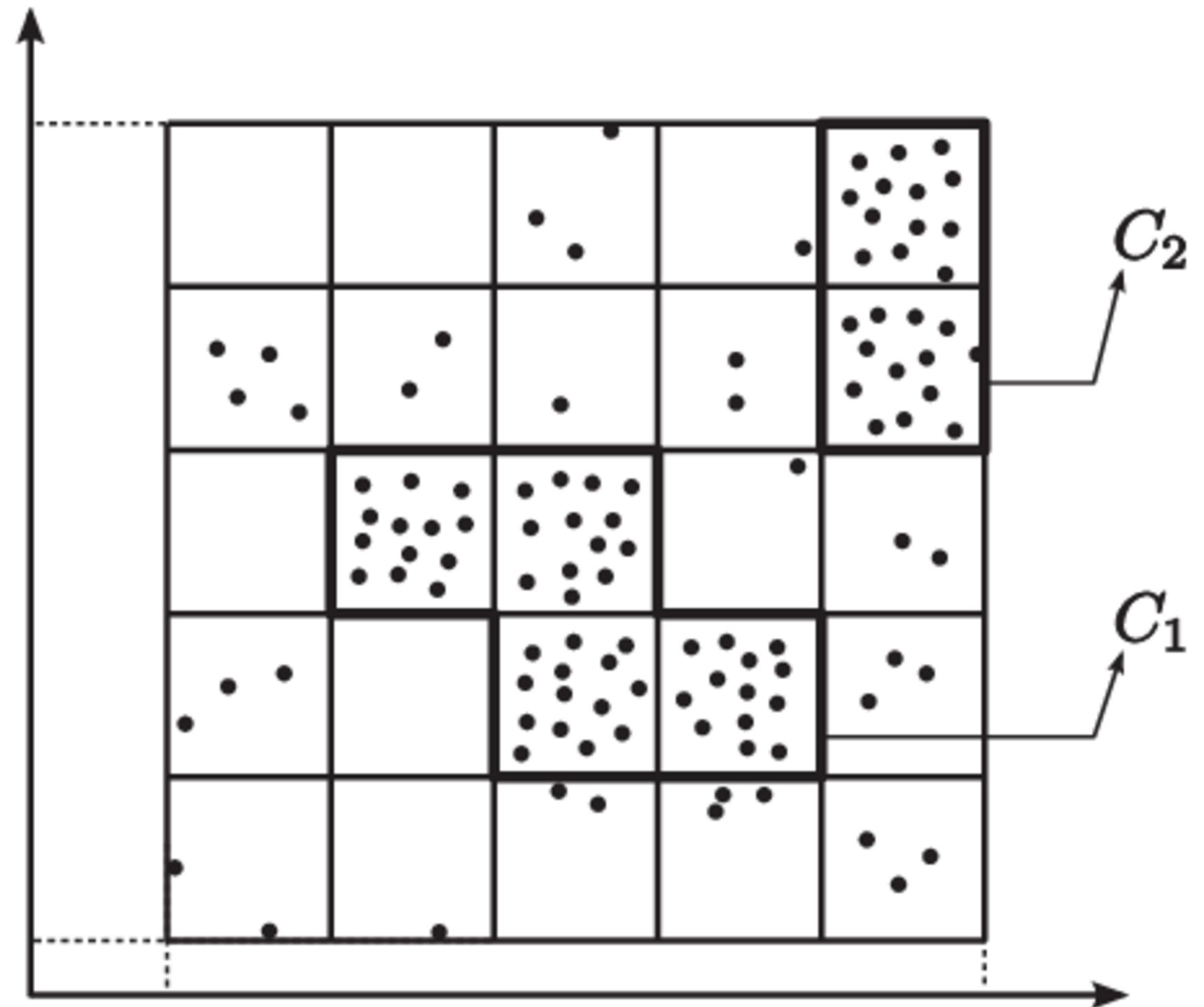


- DBSCAN có tác dụng phát hiện cụm trong dữ liệu mật độ cao cũng như điểm nhiễu, ứng dụng nhiều trong phân tích địa lý, phát hiện gian lận,...
- Ưu điểm của DBSCAN:
 - Không cần biết trước số cụm
 - Không quan trọng hình dạng cụm
 - Phát hiện điểm nhiễu
- Nhược điểm của DBSCAN:
 - Phụ thuộc vào tham số đầu vào
 - Hoạt động kém nếu mật độ các cụm sai khác nhiều
 - Chậm với dữ liệu lớn

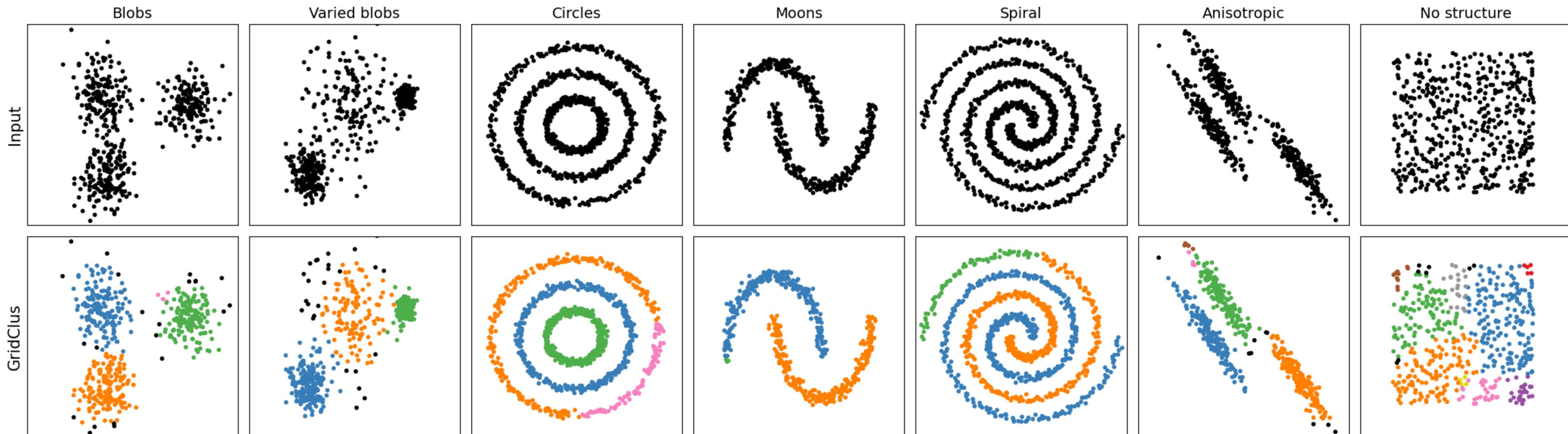
DBSCAN



- GridClus thực hiện chia không gian dữ liệu thành các ô lưới sau đó dựa vào mật độ điểm trong mỗi ô để xác định các cụm

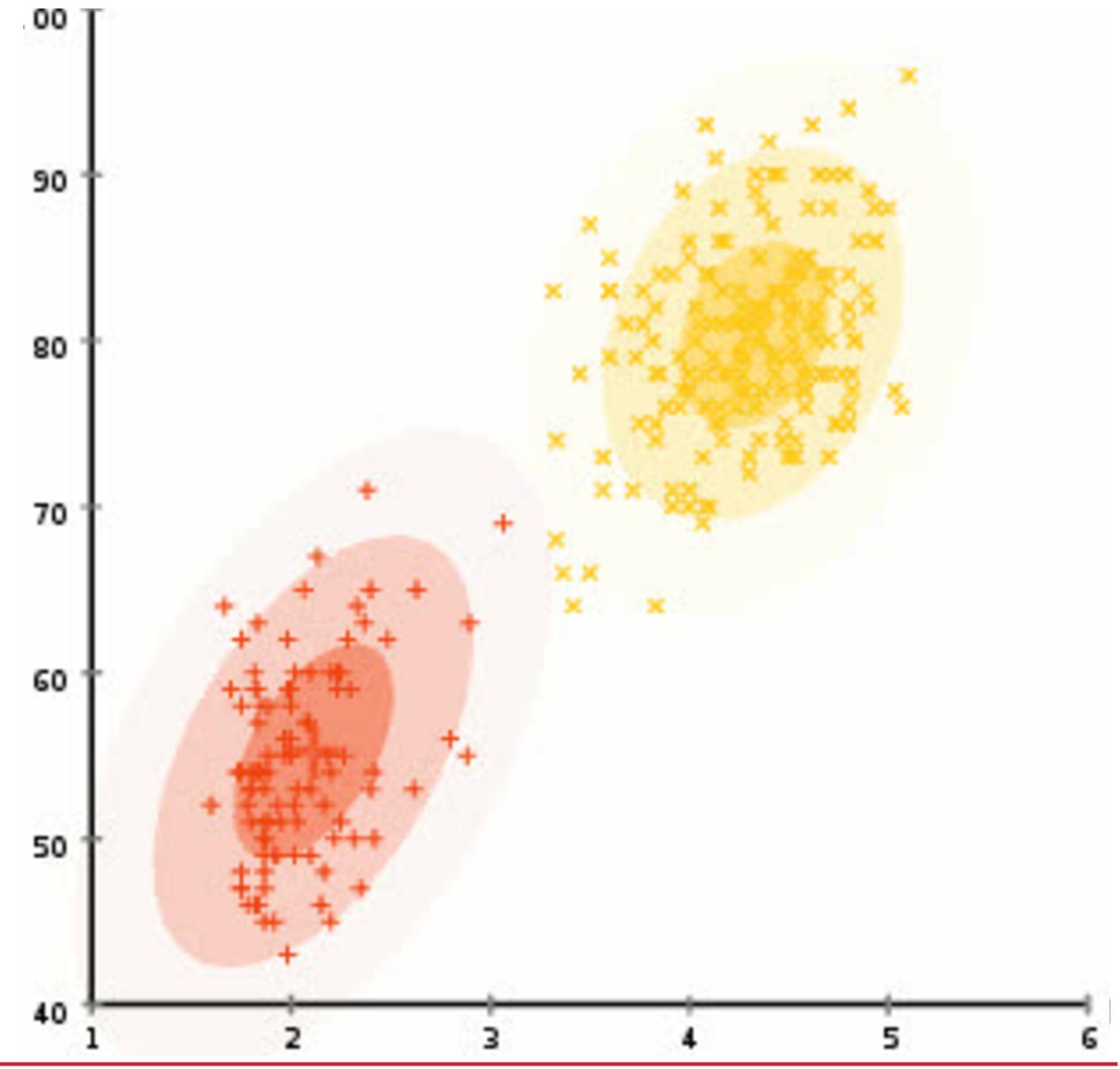


- GridClus ứng dụng trong bài toán dữ liệu lớn, phân tích địa lý,...
- Ưu điểm của GridClus:
 - Tốc độ nhanh, đặc biệt với dữ liệu lớn
 - Kết quả tốt với dữ liệu nhiều chiều
 - Không quan trọng hình dạng cụm
- Nhược điểm của GridClus:
 - Phụ thuộc vào kích thước ô lưới chia ban đầu
 - Hoạt động kém nếu mật độ các cụm sai khác nhiều
 - Phức tạp với dữ liệu nhiều chiều



Gaussian Mixture Model

- Gaussian Mixture Model (GMM) gồm 2 bước chính:
 - E-step: Tính xác suất một điểm thuộc một cụm
 - M-step: Dùng các xác suất đã tính để:
 - Tính trung tâm cụm
 - Tính độ rộng cụm
 - Tính trọng số



Gaussian Mixture Model

- Xác suất một điểm thuộc một cụm:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(x_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_n | \mu_j, \Sigma_j)}$$

- Cập nhật trung tâm cụm:

$$\mu_k = \frac{\sum_{n=1}^N \gamma(z_{nk}) x_n}{\sum_{n=1}^N \gamma(z_{nk})}$$

- Cập nhật độ rộng cụm:

$$\Sigma_k = \frac{\sum_{n=1}^N \gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^N \gamma(z_{nk})}$$

- Cập nhật trọng số:

$$\pi_k = \frac{\sum_{n=1}^N \gamma(z_{nk})}{N}$$

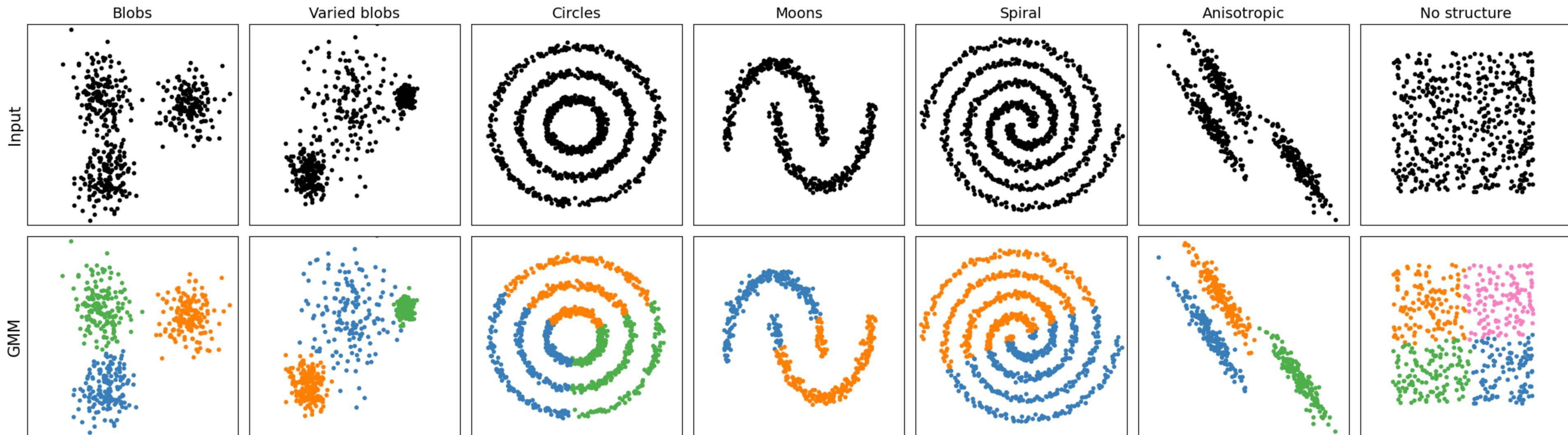


Gaussian Mixture Model

- GMM được ứng dụng trong bài toán phân cụm khi cụm có hình dạng elip, xử lý tín hiệu và âm thanh,...
- Ưu điểm của GMM:
 - Linh hoạt, cho biết xác suất, mức độ thuộc về cụm
 - Mô hình xác suất dễ giải thích, có thể dùng để sinh dữ liệu mới
- Nhược điểm của GMM:
 - Phải chọn trước số cụm K, phụ thuộc khởi tạo ban đầu
 - Phụ thuộc vào giả định phân phối Gaussian của dữ liệu
 - Tốn thời gian, tính toán nặng đặc biệt với dữ liệu nhiều chiều

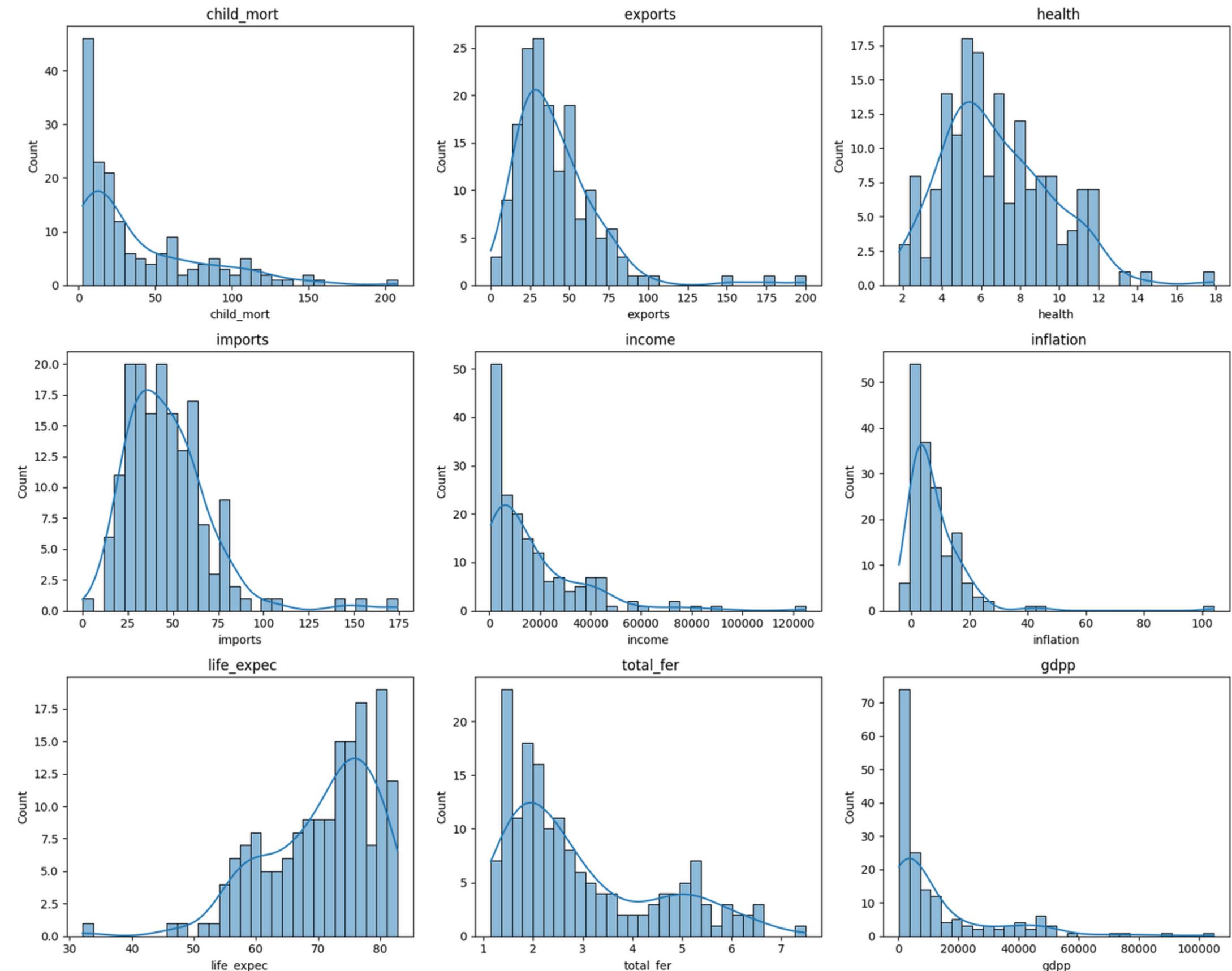


Gaussian Mixture Model



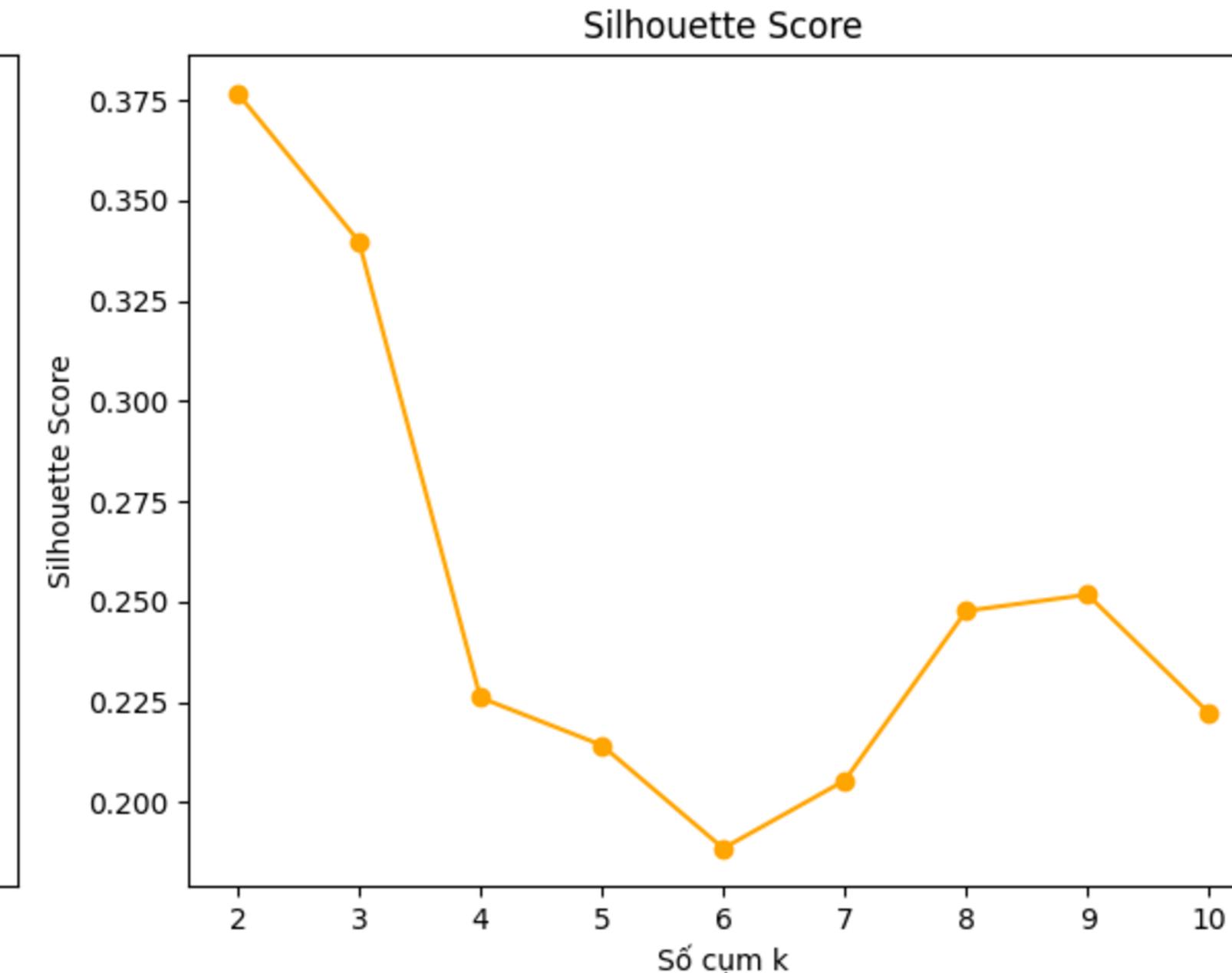
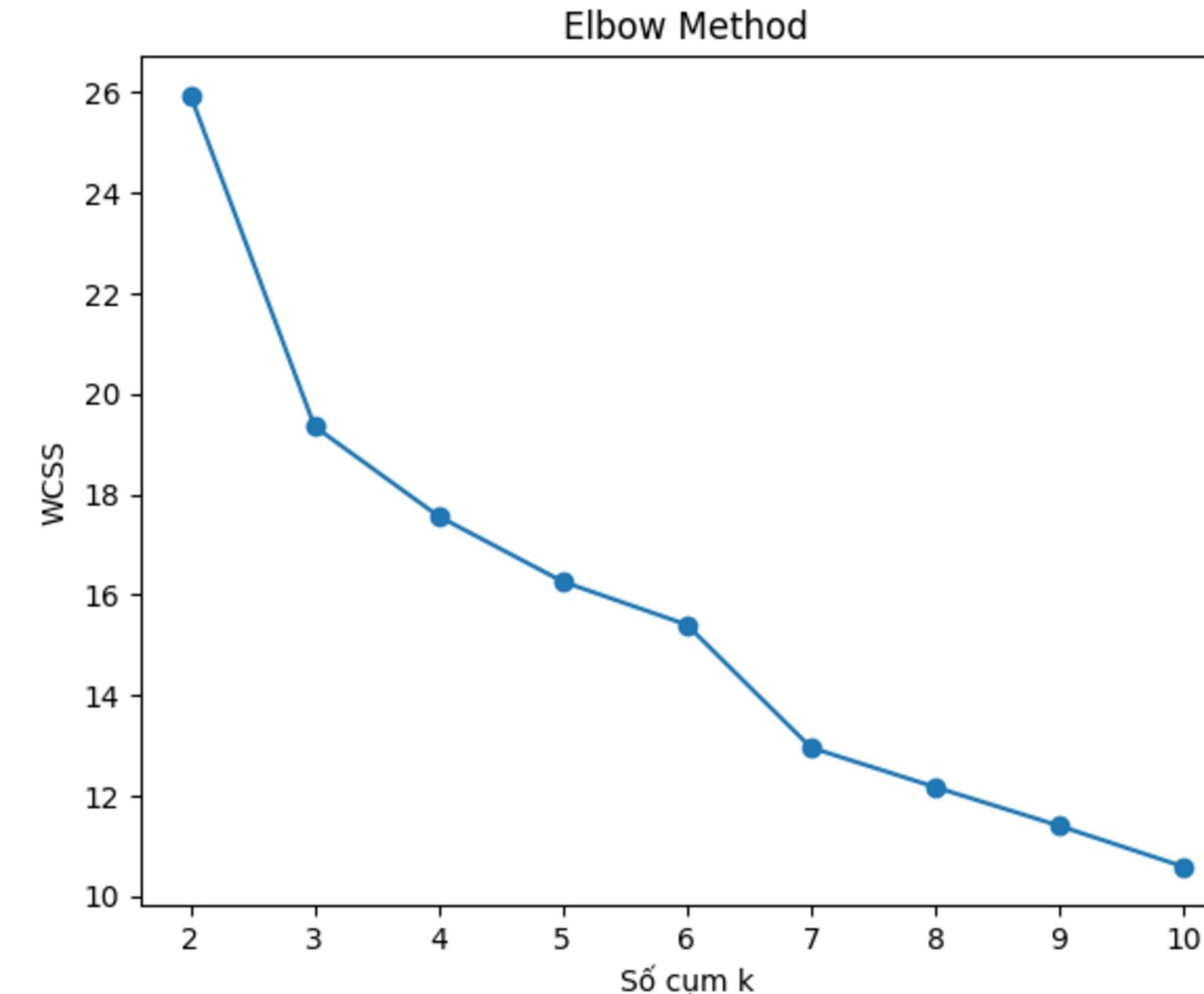
Ứng dụng trong bài toán thực tế

- Bộ dữ liệu:
Unsupervised Learning
on Country Data
- Dữ liệu của 167 nước
gồm các thông tin về:
tỉ lệ sinh, GDP, thu
nhập đầu người, xuất
nhập khẩu,...



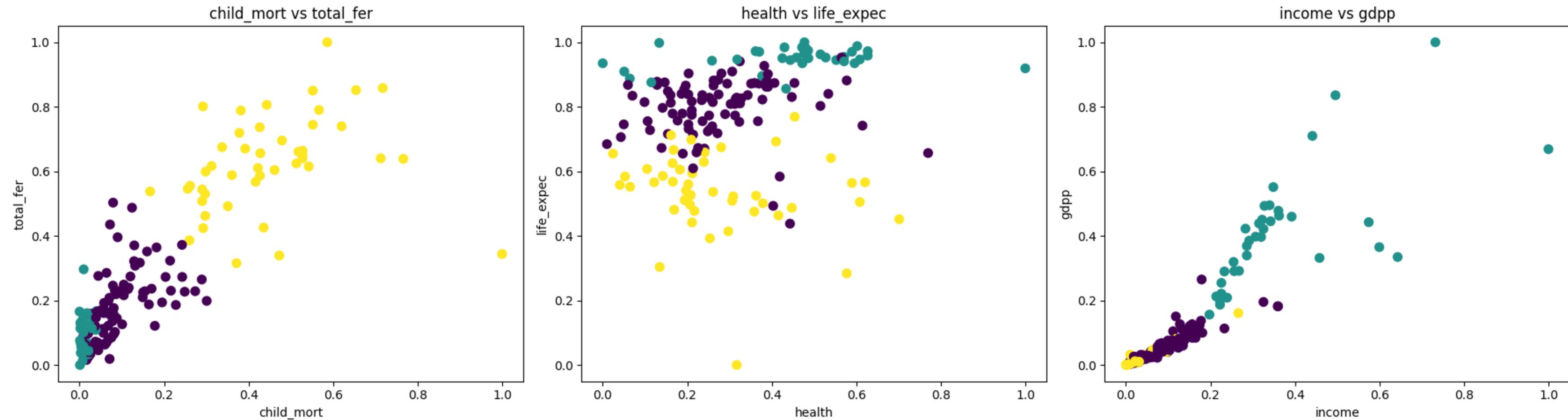
Ứng dụng trong bài toán thực tế

- Xác định số lượng cụm: phương pháp Elbow Method và Silhouette Score



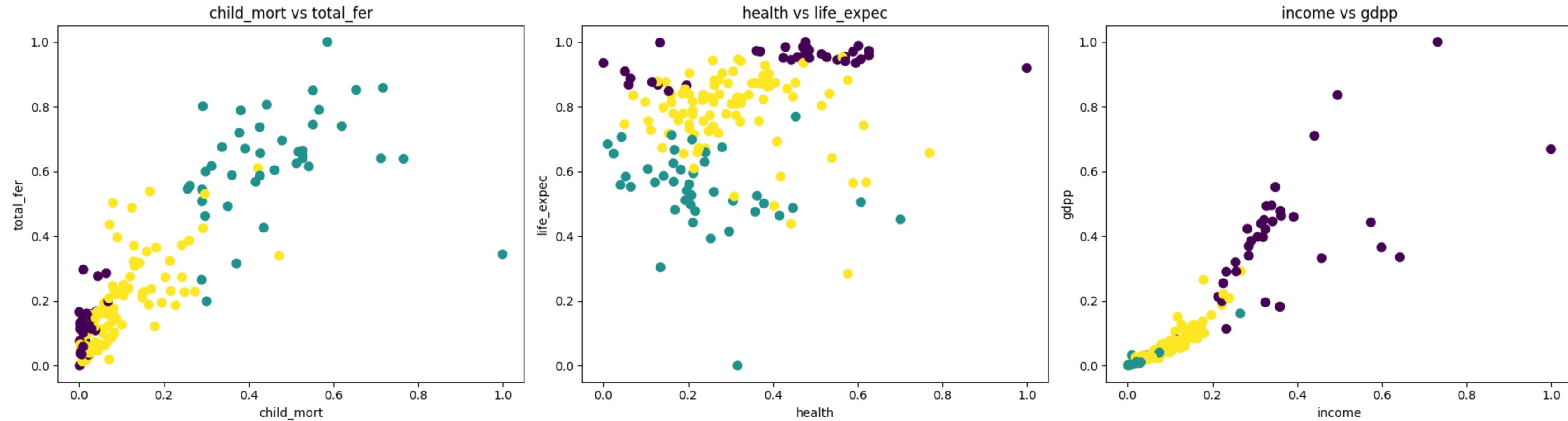
Ứng dụng trong bài toán thực tế

- K-means Clustering



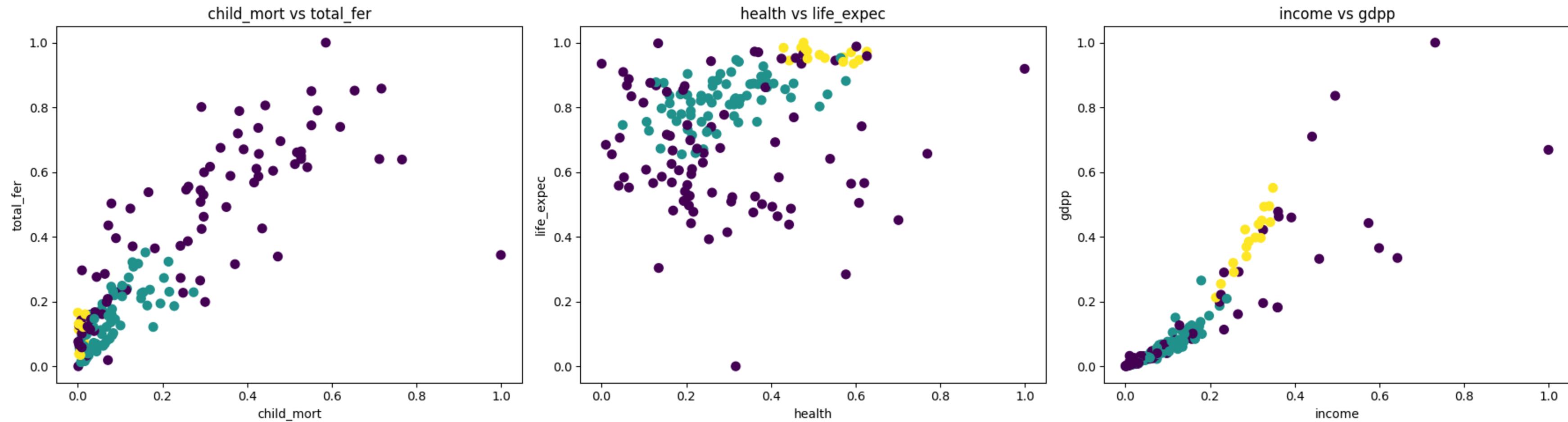
Ứng dụng trong bài toán thực tế

- Agglomerative Clustering



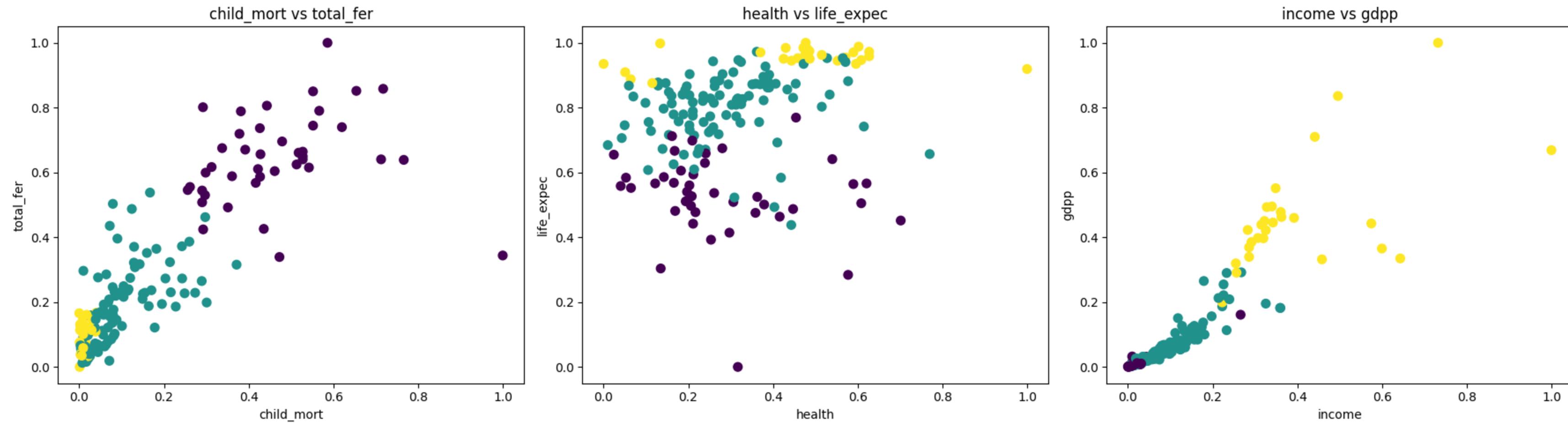
Ứng dụng trong bài toán thực tế

- DBSCAN



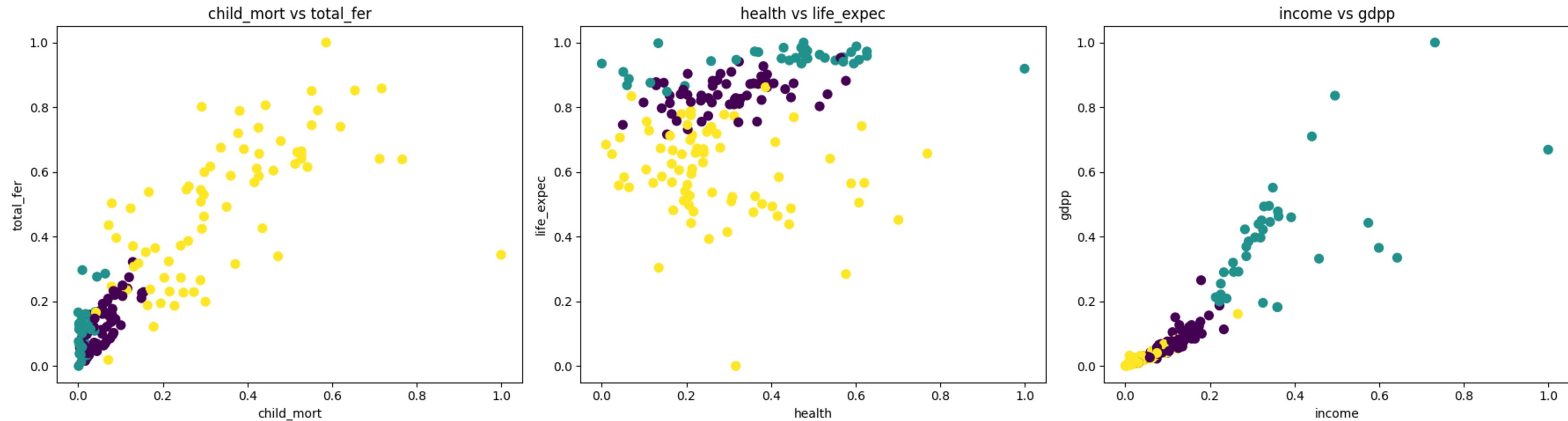
Ứng dụng trong bài toán thực tế

- DBSCAN



Ứng dụng trong bài toán thực tế

- Gaussian Mixture Model





HUST

THANK YOU !