TS. NGUYỄN QUỐC HÙNG

- Mobile: 0912 251 253
- Email: hungngq@ueh.edu.vn
- Website: https://bit.ueh.edu.vn/nqhung/

KHOA HỌC DỮ LIỆU

- Mã học phần: **25D1INF50905948**
- Thời gian: **11/04/2025 - 16/05/2025**
- Hệ: ĐH, Chính quy
- Số lượng: 48 sinh viên
- Số tín chỉ: 2.00

PHÂN CỤM DỮ LIỆU - CLUSTERING

Thứ 6, thời gian: 12g45-17g05, Giảng đường: N1-303

Phân loại thuật toán khai phá dữ liệu

- Học có giám sát
  - Bài toán phân lớp
  - Có biến target y và phân dữ liệu vào các y phù hợp

- Học không giám sát
  - Bài toán phân cụm
  - Không có target y chỉ có features x

PHÂN CỤM DỮ LIỆU (CLUSTERING)

NỘI DUNG

- **Bài toán phân cụm dữ liệu**
  - Giới thiệu phân cụm dữ liệu
  - Các ứng dụng phân cụm dữ liệu trong kinh tế
- **Một số phương pháp phân cụm**
  - Hierarchical clustering: Agnes, Diana
  - Partitioning clustering: K-means,Fuzzy C-means
- **Đánh giá mô hình phân cụm**
  - Đánh giá ngoài (external validation)
  - Đánh giá nội bộ ((internal validation)
  - Đánh giá tương đối (relative validation)
- **Minh họa bằng công cụ Orange**

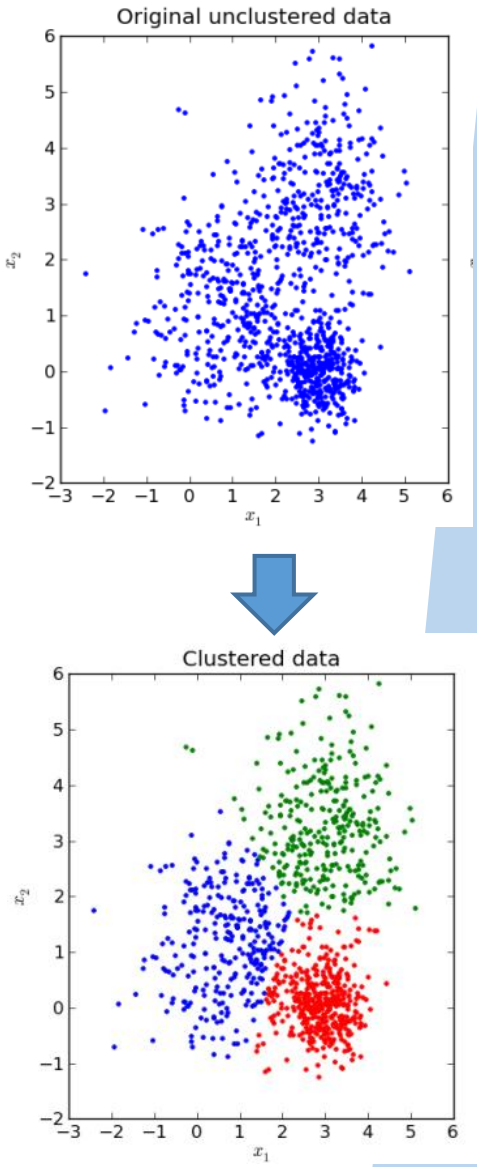
Giới thiệu phân cụm dữ liệu

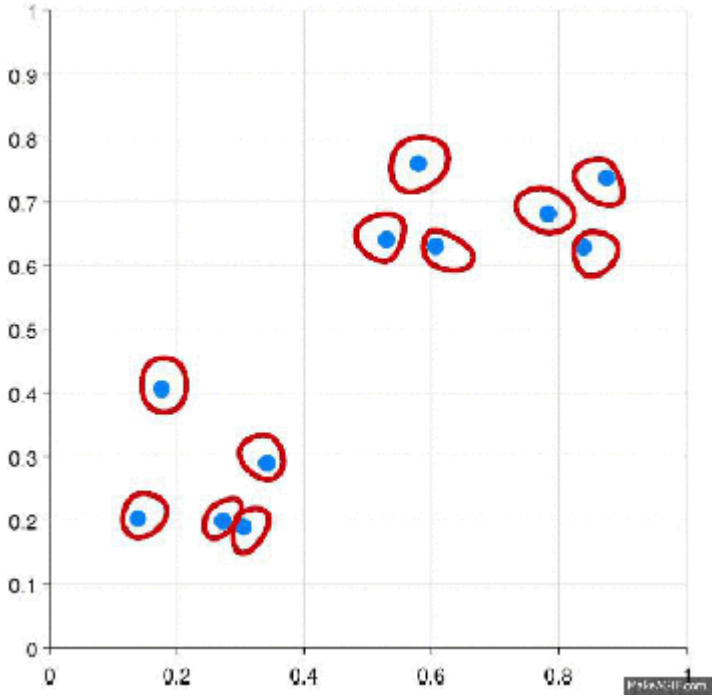
Định nghĩa

Là quá trình gom cụm/nhóm các đối tượng/dữ liệu có đặc điểm tương đồng vào các cụm/nhóm tương ứng. Trong đó:

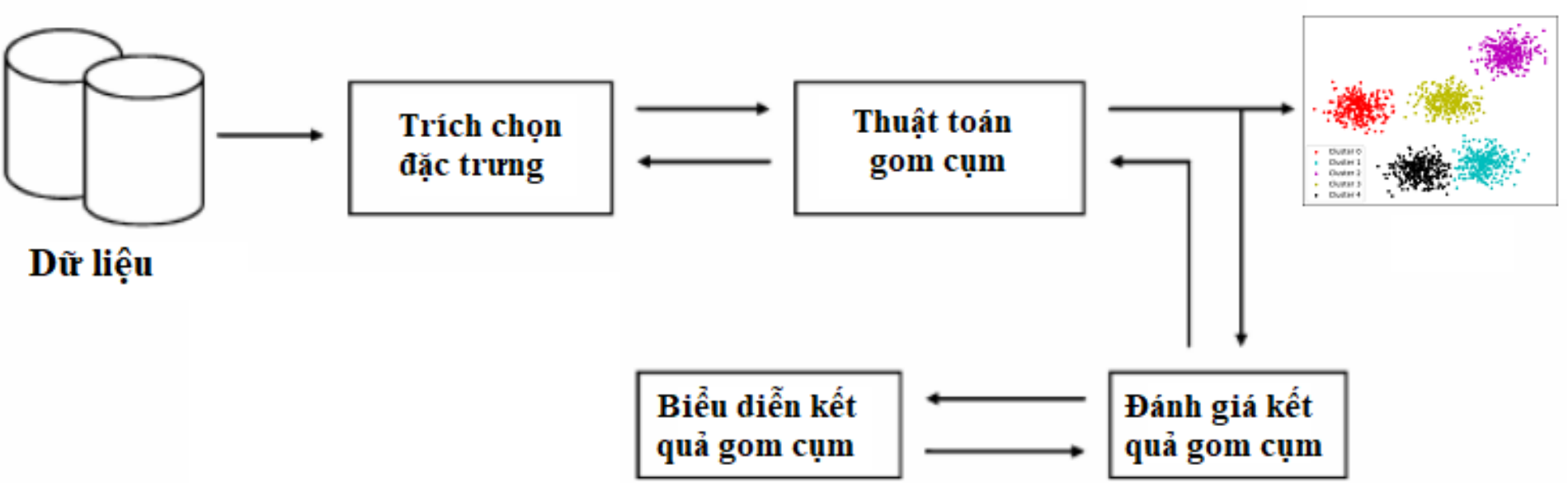
- Các đối tượng trong cùng một cụm sẽ có những tính chất tương tự nhau.
- Các đối tượng thuộc cụm/nhóm khác nhau sẽ có các tính chất khác nhau.

Lưu ý: Dữ liệu của bài toán phân cụm là dữ liệu chưa được gán nhãn. Đây là dữ liệu tự nhiên thường thấy trong thực tế.




1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.2
0.4
0.6
0.8
1

Giới thiệu phân cụm dữ liệu



Mô hình quá trình phân cụm dữ liệu

Giới thiệu phân cụm dữ liệu

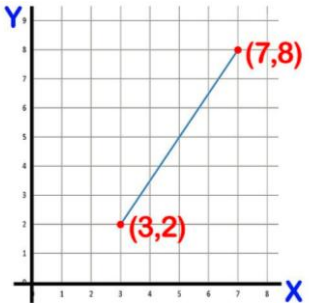
Đặc điểm:

- Nhiệm vụ chính là tìm ra và đo đạc sự khác biệt giữa các đối tượng dữ liệu.
- Phân cụm thuộc nhóm phương pháp học không giám sát (unsupervised learning) vì không biết trước được số nhóm (khác với bài toán phân lớp)
- Một phương pháp phân cụm tốt là phương pháp tạo ra các cụm có chất lượng cao:
  - Độ tương đồng bên trong cụm cao
  - Độ tương tự giữa các cụm thấp (khác biệt cao)
- Các ứng dụng điển hình:
  - Công cụ phân cụm dữ liệu độc lập.
  - Là giai đoạn tiền xử lý cho các thuật toán khác

Giới thiệu phân cụm dữ liệu

- Độ đo phân cụm: được sử dụng làm tiêu chí nhằm tính toán sự tương đồng/sai biệt giữa các đối tượng dữ liệu nhằm phục vụ cho quá trình gom cụm
- Một số độ đo phân cụm:
  - Euclid
  - Cosin $\cos\mu = \frac{v.w}{\|v\|.\|w\|}$
  - Minkowski:

$$\sum_{i=1}^{n}(\|x_i - y_i\|^p)^{\frac{1}{p}}$$



$$\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

horizontal distance

vertical distance

Giới thiệu phân cụm dữ liệu

Phân loại một số phương pháp phân cụm chính

Loại	Đặc điểm	Các phương pháp điển hình
Dựa trên phân cấp (Hierarchical approach)	Phân cấp các đối tượng dựa trên một số tiêu chí	Diana, Agnes, BIRCH, CAMELEON
Dựa trên phân hoạch (Partitioning approach)	Xây dựng các phân hoạch khác nhau và đánh giá chúng. Sau đó, tìm cách tối thiểu hóa tổng bình phương độ lỗi.	**K-means**, k-medoids, **fuzzy C-means**
Dựa trên mật độ (Density-based approach)	Dựa trên các kết nối giữa các đối tượng và hàm mật độ	DBSCAN, OPTICS, DenClue
Dựa trên lưới (Grid-based approach)	Dựa trên cấu trúc độ chi tiết nhiều cấp	STING, WaveCluster, CLIQUE
Dựa trên mô hình (Model-based)	Giả định mỗi cụm có một mô hình và tìm cách fit mô hình đó vào mỗi cụm	EM, SOM, COBWEB

Các ứng dụng phân cụm trong kinh tế

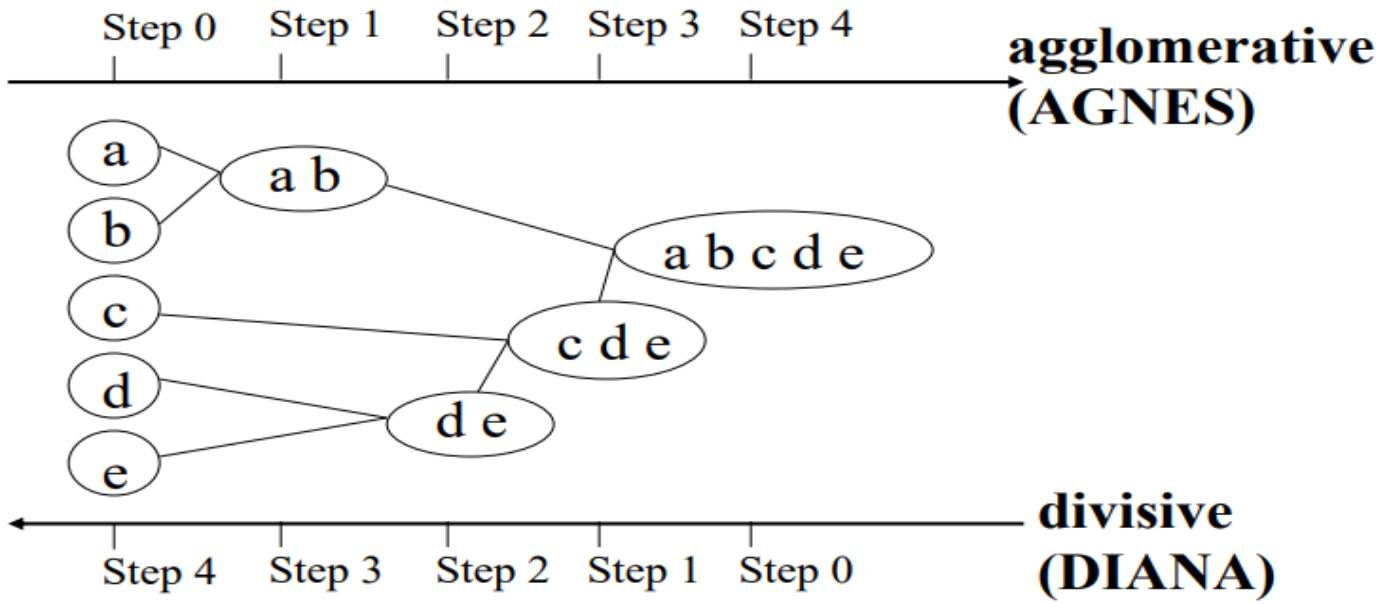
- Dự báo khách hàng tiềm năng
- Phân tích xu hướng hành vi khách hàng
- Phân tích cạnh tranh, xu hướng lựa chọn dịch vụ giữa các nhà cung cấp
- Phân tích đặc tính sản phẩm dịch vụ
- Đánh giá kết quả hoạt động kinh doanh
- Phân tích hành vi người dùng mạng xã hội

Phân cụm phân cấp (Hierarchical clustering)

- Xây dựng một cây phân cấp cho dữ liệu cần gom cụm dựa trên:
  - Ma trận khoảng cách giữa các phần tử (similarity matrix hoặc dissimilarity matrix)
  - Độ đo khoảng cách giữa các cụm (single link, complete link…)
- Phương pháp này không cần xác định trước số cụm nhưng cần xác định điều kiện dừng.
- Các phương pháp điển hình: Diana, Agnes…

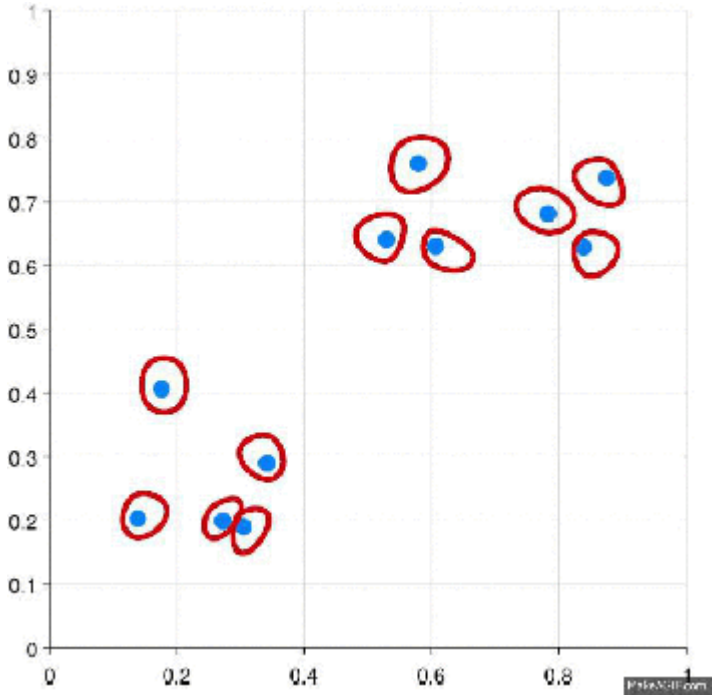
Diana và Agnes

- Được giới thiệu bởi Kaufmann và Rousseeuw năm 1990
- Được cài đặt vào các gói ứng dụng thống kê

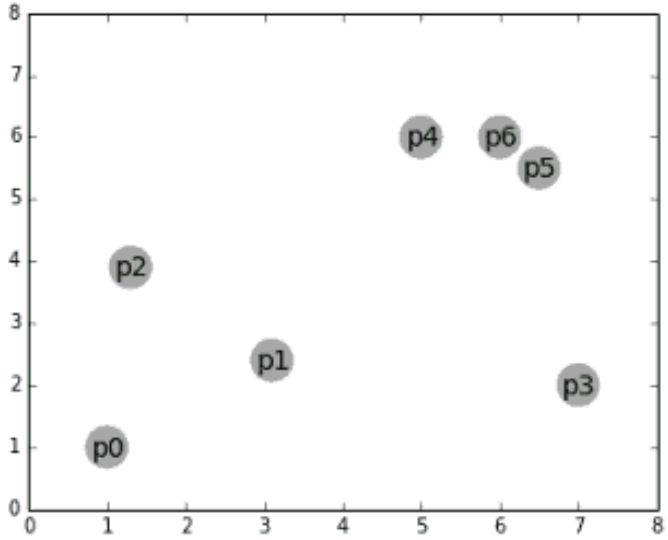


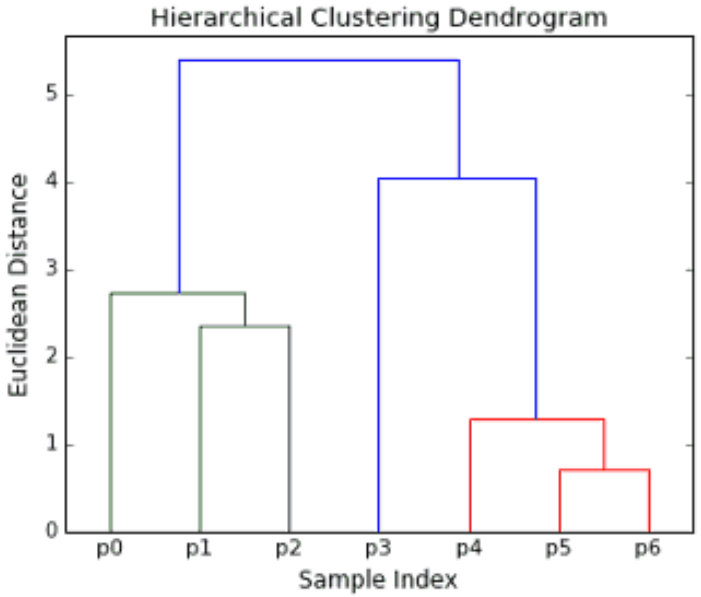
- ✓ Sử dụng ma trận sai khác (dissimilarity matrix) và phương pháp single-link.
- ✓ Là hai phương pháp có thứ tự thực hiện trái ngược nhau

Agnes



Agnes


p0
p1
p2
p3
p4
p5
p6

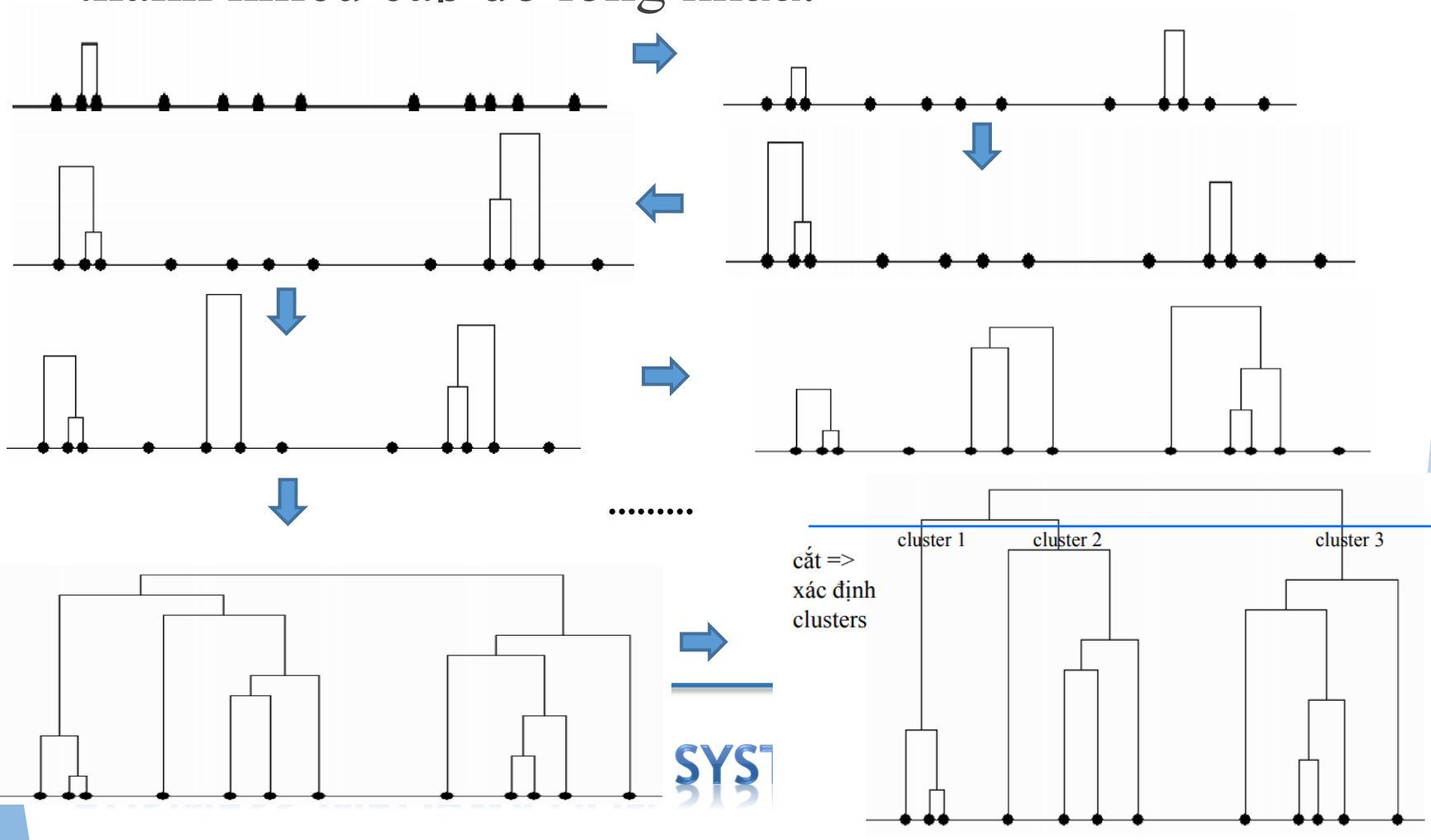

Hierarchical Clustering Dendrogram
Euclidean Distance
Sample Index
p0
p1
p2
p3
p4
p5
p6

AGNES

- Theo chiến lược bottom up:
- Bắt đầu với những cụm chỉ là 1 phần tử.
- Ở mỗi bước, gom 2 cụm gần nhau thành 1 cụm.
  - ✓ Khoảng cách giữa 2 cụm là khoảng cách giữa 2 điểm gần nhất từ hai cụm, hoặc khoảng cách trung bình.
- Quá trình này lặp lại cho đến khi tất cả các phần tử cùng thuộc một cụm lớn.
- Kết quả quá trình phát là một dendrogram (cây phân cấp)

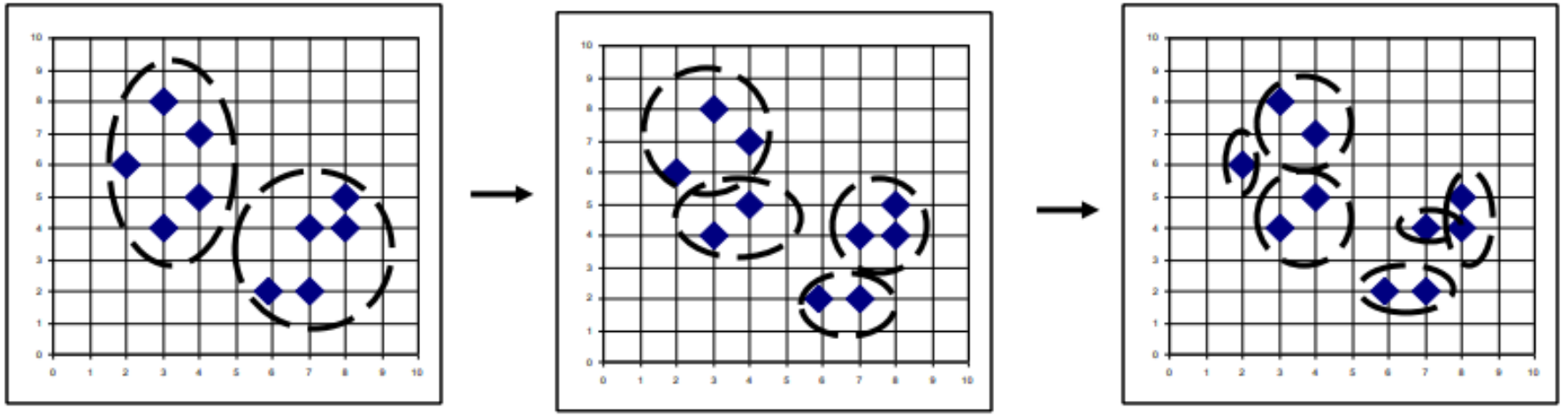
AGNES – DENDROGRAM

- Là sơ đồ/cây biểu diễn sự phân rã các phần tử dữ liệu thành nhiều cấp độ lồng nhau.



DIANA

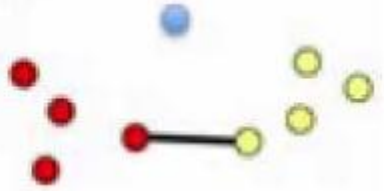
- Theo chiến lược top down:
- Bắt đầu với 1 cụm gồm tất cả phần tử.
- Ở mỗi bước, chia cụm ban đầu thành 2 cụm.
  - ✓ Khoảng cách giữa 2 cụm là khoảng cách giữa 2 điểm gần nhất từ hai cụm, hoặc khoảng cách trung bình.
- Thực hiện đệ quy trên các cụm mới được tách ra và lặp lại cho đến khi mỗi phần tử là 1 cụm.
- Kết quả phát sinh cây phân cấp (dendrogram)



Một số phương pháp tính khoảng cách

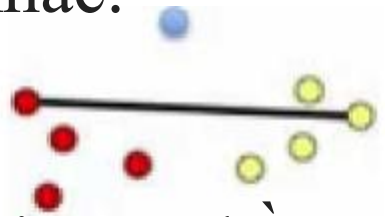
- **Single-link**: khoảng cách nhỏ nhất giữa 1 phần tử trong một cụm với một phần tử ở cụm khác.

$$\text{dist}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \|p - p'\|$$



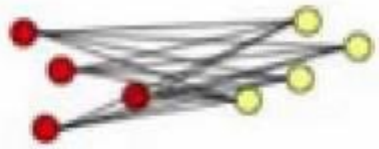
- **Complete-link**: khoảng cách lớn nhất giữa 1 phần tử trong một cụm với một phần tử ở cụm khác.

$$\text{dist}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \|p - p'\|$$



- **Average-link**: khoảng cách trung bình giữa 1 phần tử trong một cụm với một phần tử ở cụm khác.

$$\text{dist}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'|$$



Một số phương pháp tính khoảng cách

- **Mean**: khoảng cách giữa các điểm trung bình (mean) của 2 cụm.

$$\text{dist}(C_i, C_j) = |m_i - m_j|$$

Với m_i và m_j là trung bình của các phần tử trong cụm C_i và C_j

- **Centroid**: khoảng cách giữa các trọng tâm (centroid) của 2 cụm.

$$\text{dist}(C_i, C_j) = \text{dist}(c_i, c_j)$$

Với $\mathbf{c_i}$ và $\mathbf{c_j}$ lần lượt là các trọng tâm của cụm $\mathbf{C_i}$, $\mathbf{C_j}$

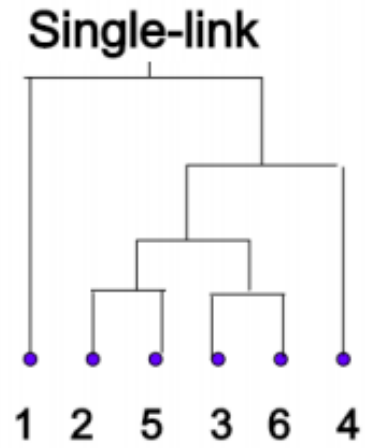
- **Medoid**: khoảng cách giữa các trung tâm cụm (medoid) của 2 cụm.

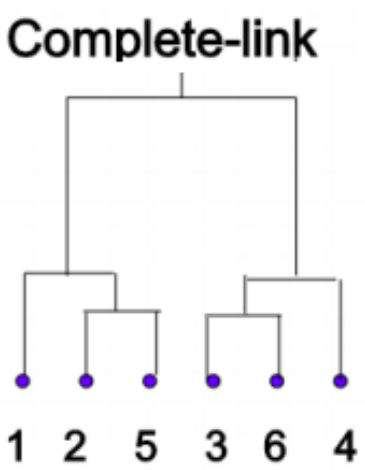
$$\text{dist}(C_i, C_j) = \text{dist}(M_i, M_j)$$

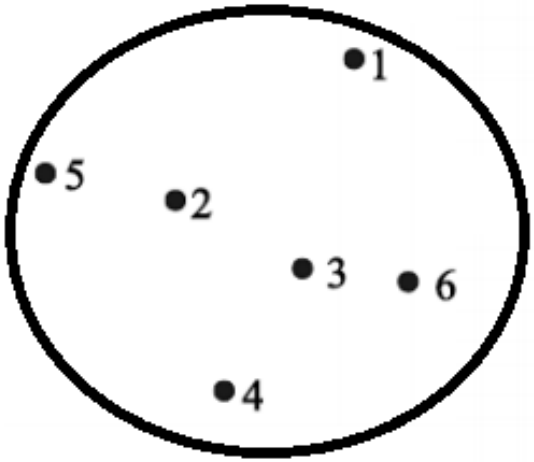
Medoid là phần tử nằm ở trung tâm cụm

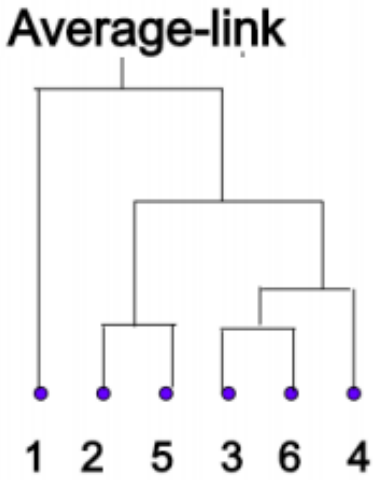
Với M_i và M_j là trung tâm của các phần tử trong cụm C_i và C_j

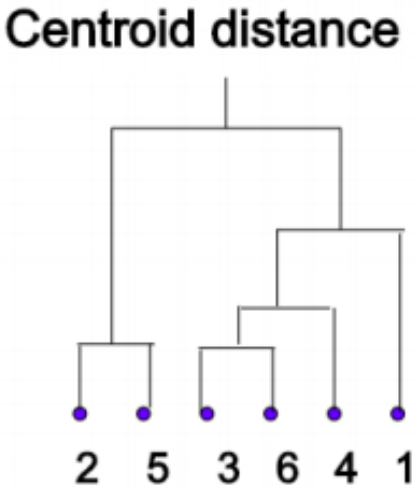
Ví dụ về một số độ đo phổ biến











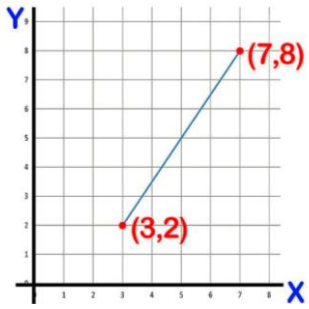
Ví dụ: AGNES

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Min Distance (Single Linkage)

Dist	A	B	C	D, F	E
A	0.00	0.71	5.66	3.20	4.24
B	0.71	0.00	4.95	2.50	3.54
C	5.66	4.95	0.00	2.24	1.41
D, F	3.20	2.50	2.24	0.00	1.00
E	4.24	3.54	1.41	1.00	0.00



$$\sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

horizontal distance

vertical distance

Min Distance (Single Linkage)

Dist	A,B	C	(D, F)	E
A,B	0	4.95	2.50	3.54
C	4.95	0	2.24	1.41
(D, F)	2.50	2.24	0	1.00
E	3.54	1.41	1.00	0

Min Distance (Single Linkage)

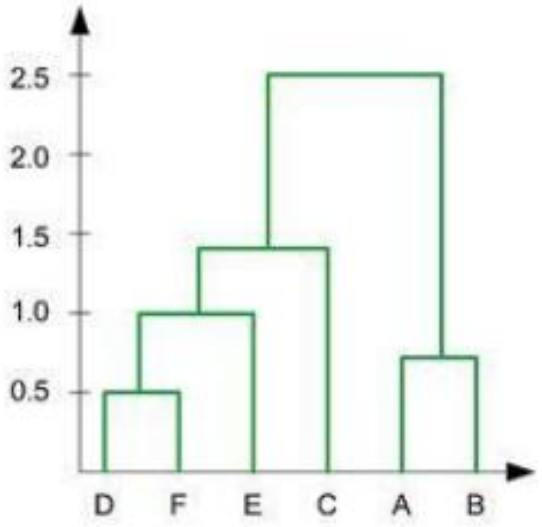
Dist	(A,B)	C	(D, F), E
(A,B)	0.00	4.95	2.50
C	4.95	0.00	1.41
(D, F), E	2.50	1.41	0.00

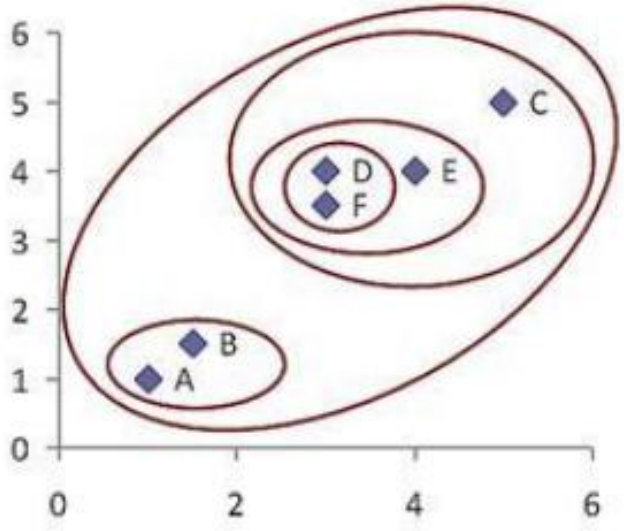
Min Distance (Single Linkage)

Dist	(A,B)	(D, F), E),C
(A,B)	0.00	2.50
((D, F), E),C	2.50	0.00

Ví dụ: AGNES

	X1	X2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5





Nhận xét về phân cụm phân cấp

- Giải thuật đơn giản
- Kết quả dễ hiểu
- Không cần tham số đầu vào
- Không quay lui được
- Tốc độ chậm, không thích hợp trên dữ liệu lớn
- Không xử lý được trên dữ liệu bị thiếu, nhạy cảm với nhiễu

Phân cụm phân hoạch (Partitioning Clustering)

- Phân tập dữ liệu có n phần tử cho trước thành k tập con (k<=n), mỗi tập con biểu diễn một cụm.
- Các cụm hình thành trên cơ sở tối ưu hóa giá trị hàm độ đo tương tự (độ đo phân cụm) sao cho:
  - Mỗi đối tượng thuộc duy nhất 1 cụm, các phần tử trong cụm có sự tương tự nhau.
  - Mỗi cụm có ít nhất 1 phần tử.
- Thuật toán điển hình: K-means, K-mediods, Fuzzy C-means

Thuật toán K-means

- Thuộc nhóm thuật toán phân cụm dựa trên phân hoạch
- Tư tưởng chính:

Ta xem mỗi đối tượng trong tập dữ liệu là một điểm trong không gian *d* chiều (với *d* là số lượng thuộc tính của đối tượng)

- **Bước 1**: Chọn *k* điểm bất kỳ làm các trung tâm ban đầu của *k* cụm.
- **Bước 2**: Phân mỗi điểm dữ liệu vào cụm có trung tâm gần nó nhất. Nếu các điểm dữ liệu ở từng cụm vừa được phân chia không thay đổi so với kết quả của lần phân chia trước nó thì ta dừng thuật toán.
- **Bước 3**: Cập nhật lại trung tâm cho từng cụm bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cụm đó sau khi phân chia ở bước 2.
- **Bước 4**: Quay lại bước 2.

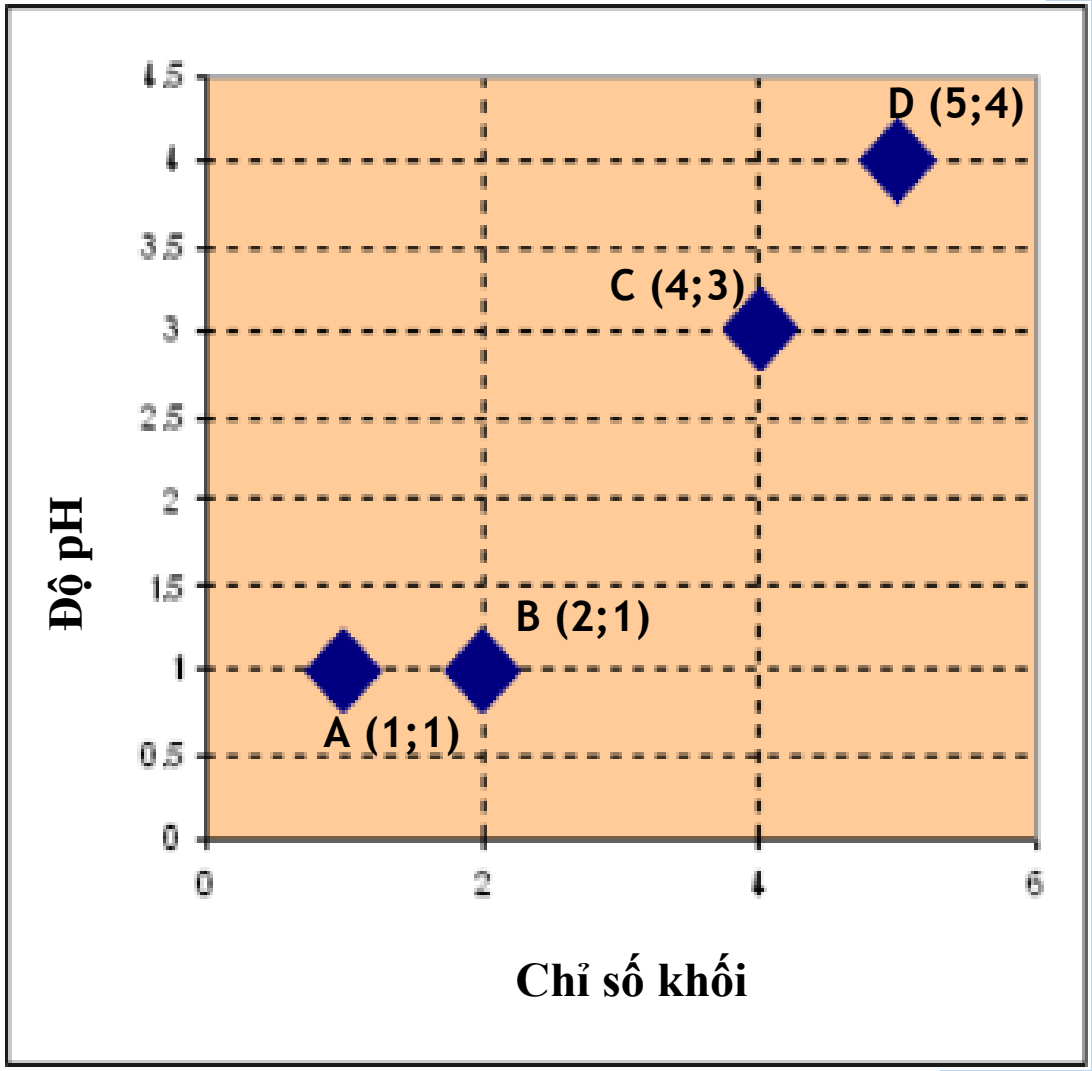
Thuật toán K-means

Ví dụ: Ta có bộ dữ liệu gồm 4 đối tượng là 4 lọ thuốc bị mất nhãn. Biết rằng 4 lọ này thuộc 2 loại khác nhau và mỗi lọ thuốc có 2 thuộc tính là chỉ số khối và độ pH như bảng bên dưới. Ta sẽ sử dụng thuật toán K-means để phân 4 đối tượng này vào 2 cụm

Đối tượng	Chỉ số khối	Độ pH
Thuốc 1	1	1
Thuốc 2	2	1
Thuốc 3	4	3
Thuốc 4	5	4

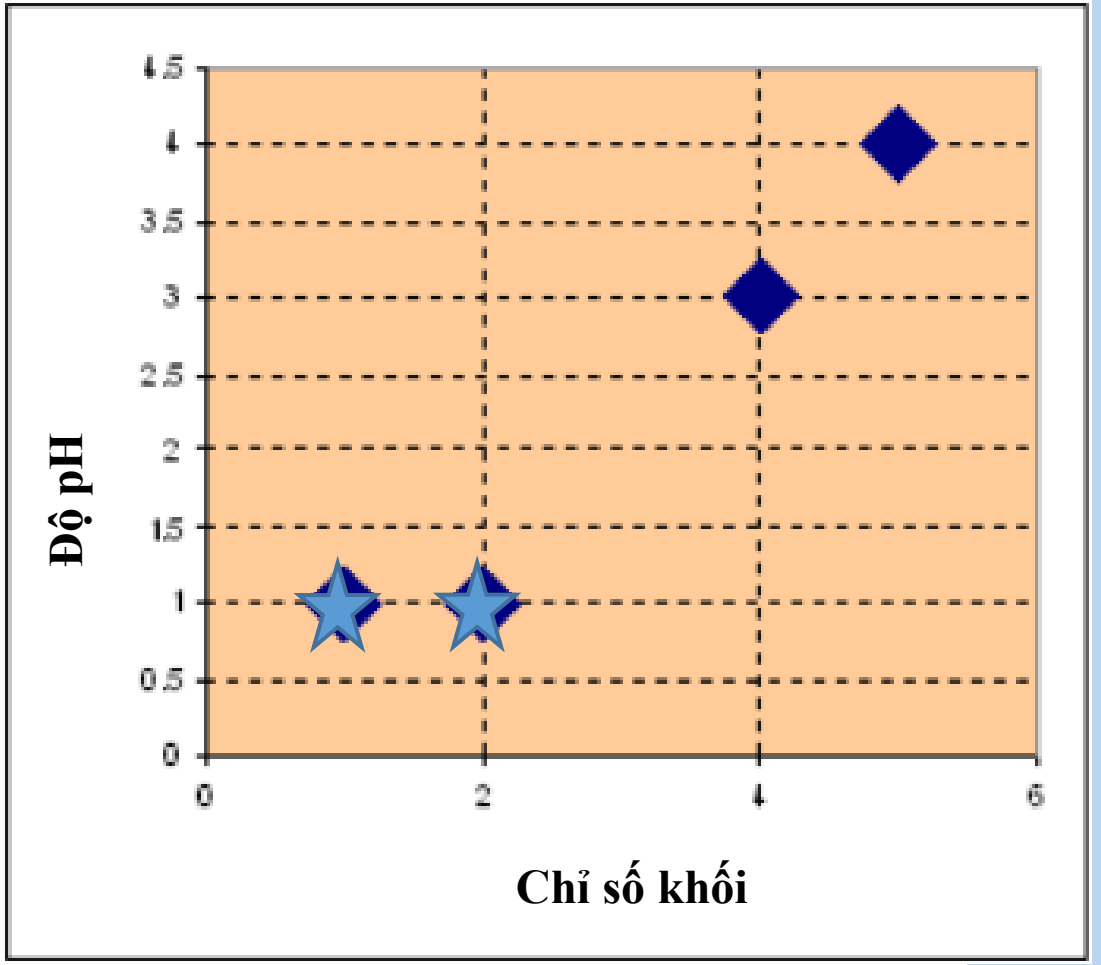
Thuật toán K-means

- ✓ Do đối tượng dữ liệu cho sẵn có 2 thuộc tính nên ta có thể xem mỗi đối tượng là một điểm trong không gian hai chiều với:
  x: chỉ số khối.
  y: độ pH.
- ✓ Các đối tượng có thể được biểu diễn trong không gian hai chiều như hình bên cạnh.



Thuật toán K-means

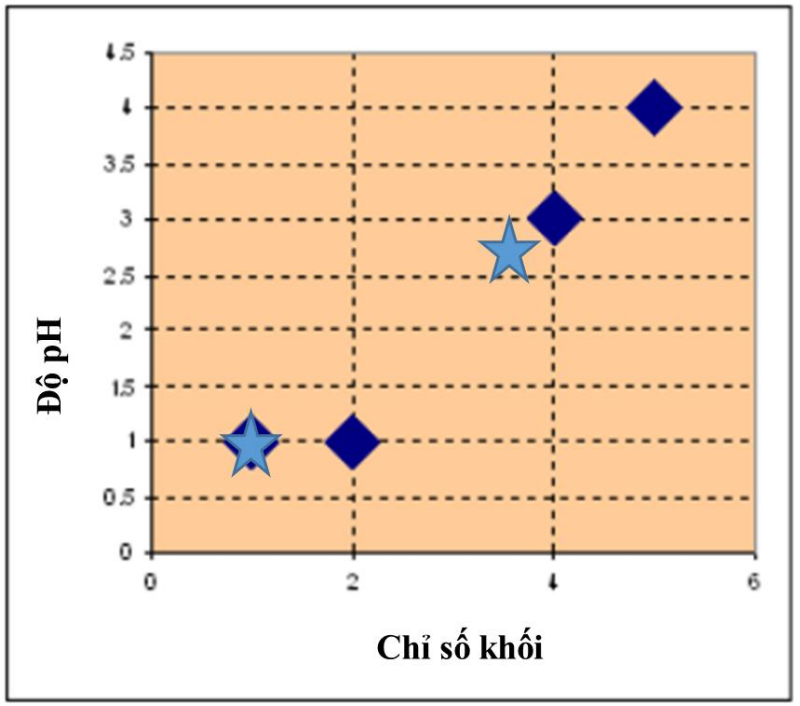
- ✓ Bước 1: Chọn 2 điểm ngẫu nhiên C_1 = A (1;1) và C_2 = B(2;1) làm 2 trung tâm của 2 cụm.
- ✓ Bước 2: Phân cụm cho các điểm trong không gian dữ liệu bằng cách tính khoảng cách Euclid từ mỗi điểm đến từng trung tâm.



Thuật toán K-means

Đối tượng	Khoảng cách tới C1	Khoảng cách tới C2
A(1;1)	0	1
B(2;1)	1	0
C(4;3)	3.606	2.828
D(5;4)	5	4.243

Đối tượng	Khoảng cách tới C1	Khoảng cách tới C2
A(1;1)	1	0
B(2;1)	0	1
C(4;3)	0	1
D(5;4)	0	1



Cụm 1: {A}

Cụm 2: {B,C,D}

✓ **Bước 3**: Cập nhật lại trung tâm của 2 cụm.

C_1 = **(1;1)**

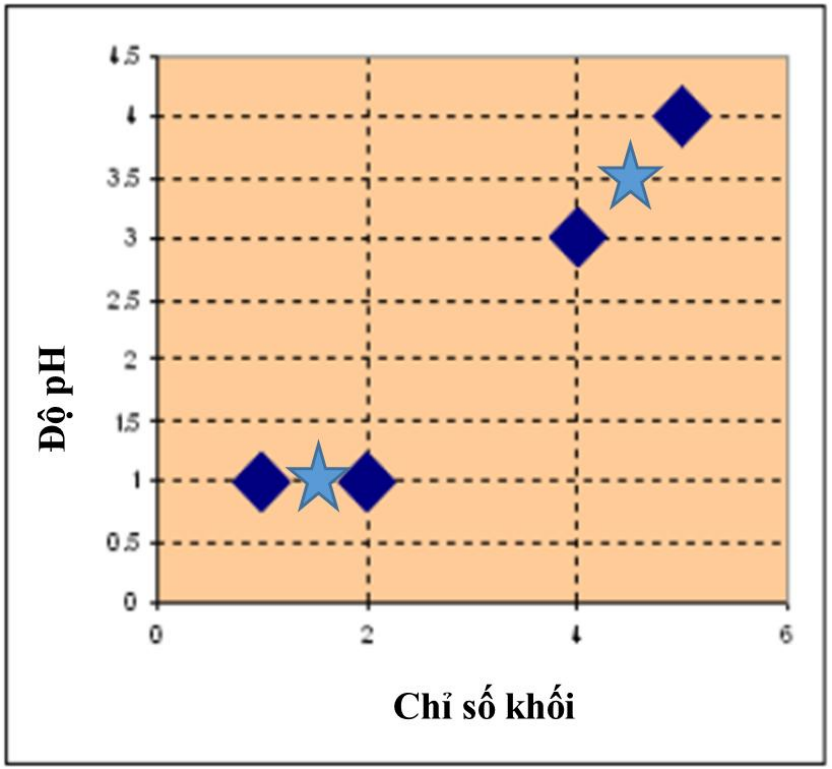
$C_2 = \left(\frac{(x_B+x_C+x_D)}{3}; \frac{(y_B+y_C+y_D)}{3}\right)$ = **(3.67;2.67)**

Thuật toán K-means

Bước 4: lặp lại **bước 2:** phân cụm lại cho các đối tượng dựa theo khoảng cách với 2 trung tâm mới

Đối tượng	Khoảng cách tới C1	Khoảng cách tới C2
A(1;1)	0	3.145
B(2;1)	1	2.357
C(4;3)	4	0.471
D(5;4)	5	1.886

Đối tượng	Khoảng cách tới C1	Khoảng cách tới C2
A(1;1)	1	0
B(2;1)	1	0
C(4;3)	0	1
D(5;4)	0	1

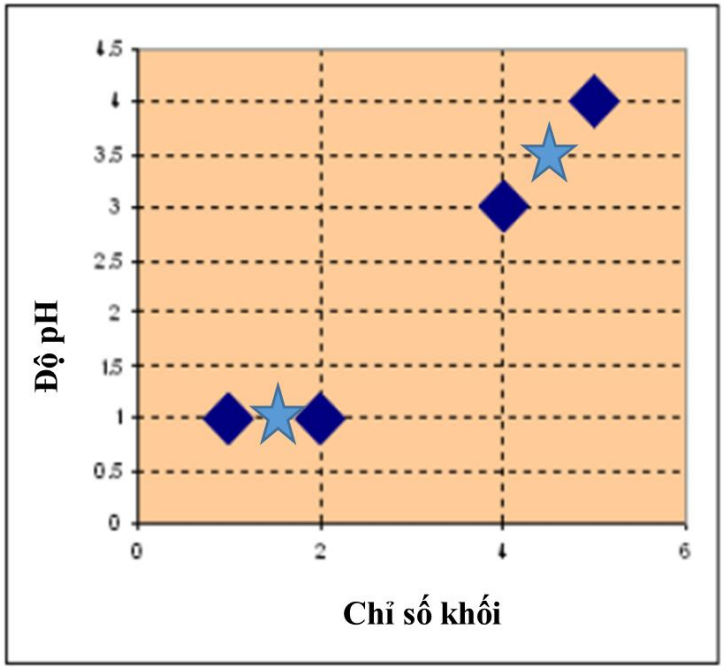


Cụm 1 :{A, B}
Cụm 2 :{C, D}

✓ **Bước 3**: Cập nhật lại trung tâm của 2 cụm.
C_1 = **(1.5;1)** ; C_2 = **(4.5;3.5)**

Thuật toán K-means

Bước 4: lặp lại **bước 2:** phân cụm lại cho các đối tượng dựa theo khoảng cách với 2 trung tâm mới



Đối tượng	Khoảng cách tới C1	Khoảng cách tới C2
A(1;1)	0.5	4.301
B(2;1)	0.5	3.536
C(4;3)	3.202	0.707
D(5;4)	4.61	0.707

Đối tượng	Khoảng cách tới C1	Khoảng cách tới C2
A(1;1)	1	0
B(2;1)	1	0
C(4;3)	0	1
D(5;4)	0	1

Cụm 1 :{A, B}
Cụm 2 :{C, D}

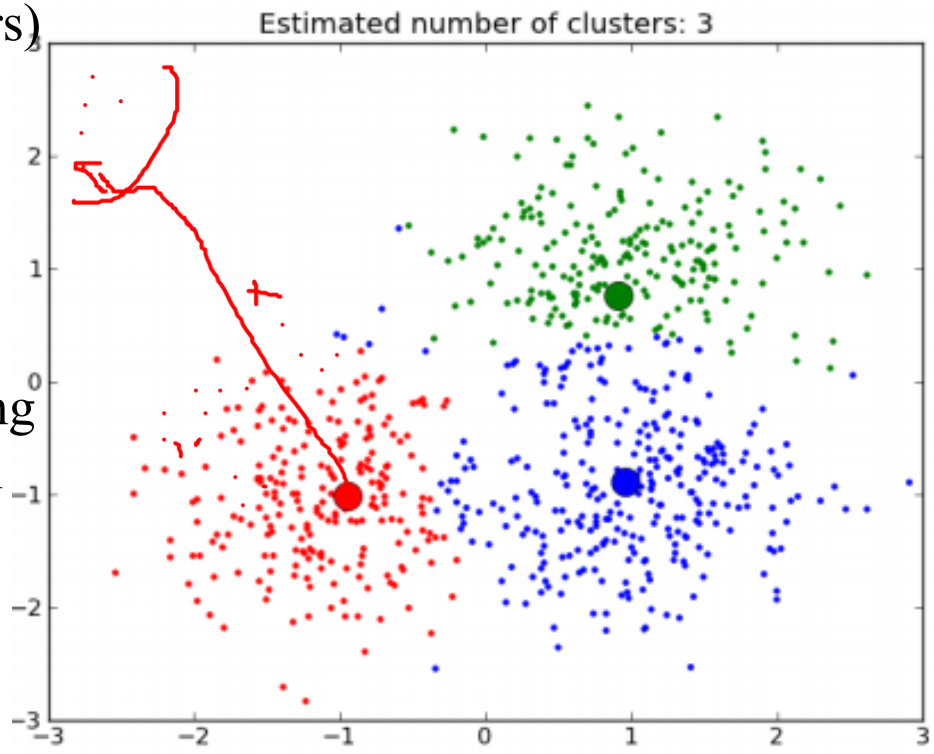
✓ **Bước 3**: Cập nhật lại trung tâm của 2 cụm.

C_1 = **(1.5;1)** ; C_2 = **(4.5;3.5)**

Kết quả phân cụm không đổi => kết thúc

Đánh giá thuật toán K-means

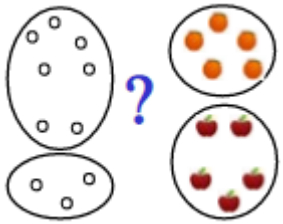
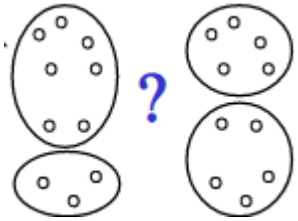
- ✓ Cần biết trước số lượng cụm *k*
- ✓ Nhạy cảm với nhiễu và ngoại biên (outliers)
- ✓ Không phù hợp với phân bố dữ liệu dạng không lồi (non-convex)
- ✓ Kết quả (nghiệm) bài toán phụ thuộc vào cách khởi tạo các trung tâm cụm ban đầu.
  - Trường hợp 1: tốc độ hội tụ chậm
  - Trường hợp 2: kết quả gom cụm không chính xác (do chỉ tìm được các cực trị địa phương chứ không phải toàn cục)
- ✓ Khắc phục:
  - Áp dụng một số phương pháp tính số cụm
  - Chạy thuật toán nhiều lần với các trung tâm khác nhau để tìm giá trị cực tiểu của hàm mất mát



Các phương pháp đánh giá phân cụm dữ liệu

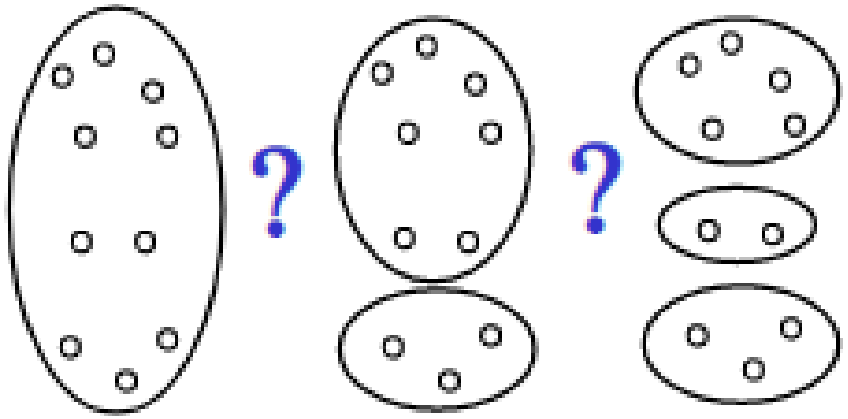
- Là vấn đề khó khăn nhất trong bài toán phân cụm
- Các phương pháp đánh giá việc phân cụm dữ liệu: đánh giá ngoài, đánh giá nội bộ, đánh giá tương đối.
- Một số tiêu chí để đánh giá chất lượng phân cụm là:
  - Độ nén (compactness): các phần tử của cụm phải "gần nhau"
  - Độ phân cách (separation): khoảng cách giữa các cụm nên "xa nhau", phân cách rõ ràng.

Đánh giá ngoài (external validation)

- Là đánh giá kết quả phân cụm dựa vào cấu trúc/ xu hướng phân cụm được chỉ định trước cho tập dữ liệu.
  - So sánh độ sai khác giữa các cụm ⟹ 
  - So sánh với kết quả mẫu (đáp án) ⟹ 
- Các độ đo được sử dụng trong phương pháp này: Rand statistic, Jaccard coefficient, Folkes và Mallows index….

Đánh giá nội bộ (internal validation)

- Là đánh giá kết quả phân cụm mà không có thông tin từ bên ngoài, chủ yếu dựa trên các vector chính của dữ thông qua ma trận xấp xỉ (proximity matrix).
- Tối ưu hóa các chỉ số nội bộ: độ nén, độ phân tách
- Các độ đo được sử dụng trong phương pháp này: Hubert's statistic, **Silhouette index**, Dunn's index, F-ratio, DBI (Davies Bouldin Inde `

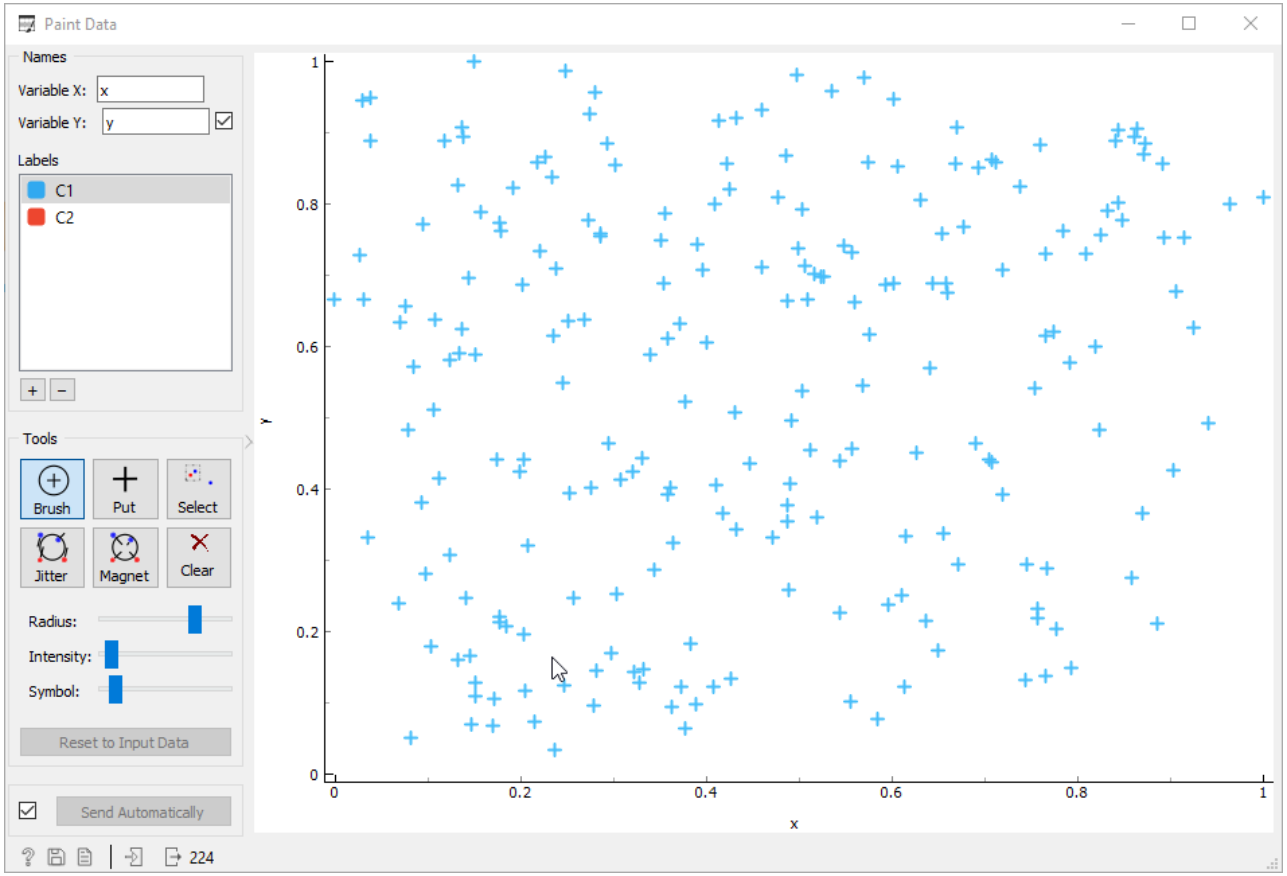


Đánh giá tương đối (relative validation)

- Đánh giá kết quả gom cụm bằng việc so sánh với:
  - Kết quả gom cụm ứng với các bộ trị thông số khác nhau.
  - Kết quả gom cụm của các phương pháp khác

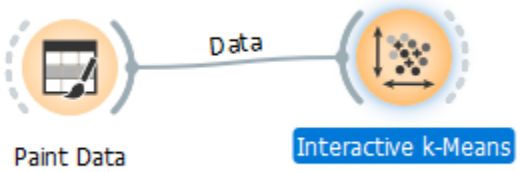
Silhouette index

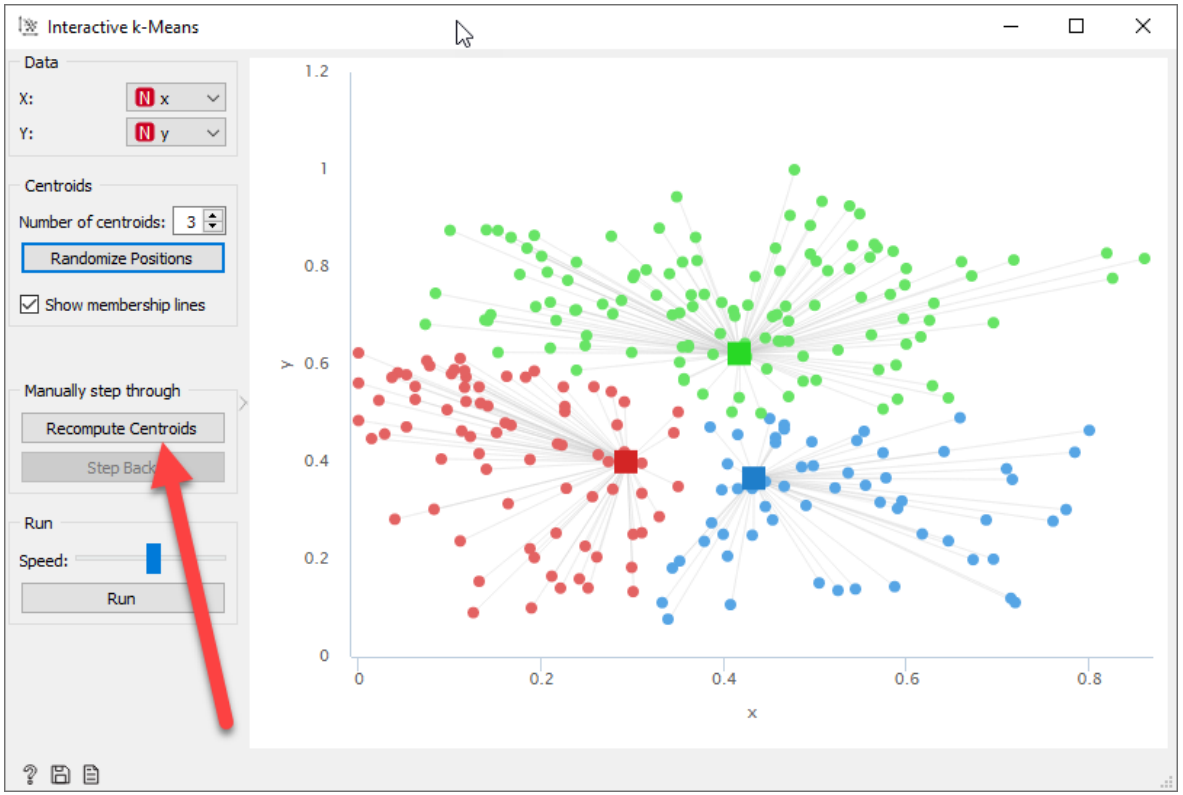
- Paint Data



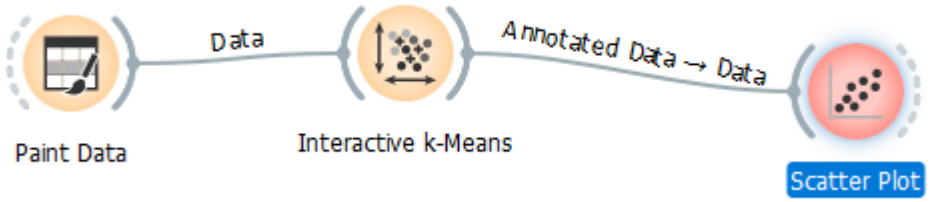
Silhouette index

- Interactive k-Means (Cài đặt bổ sung Addon/ Education)

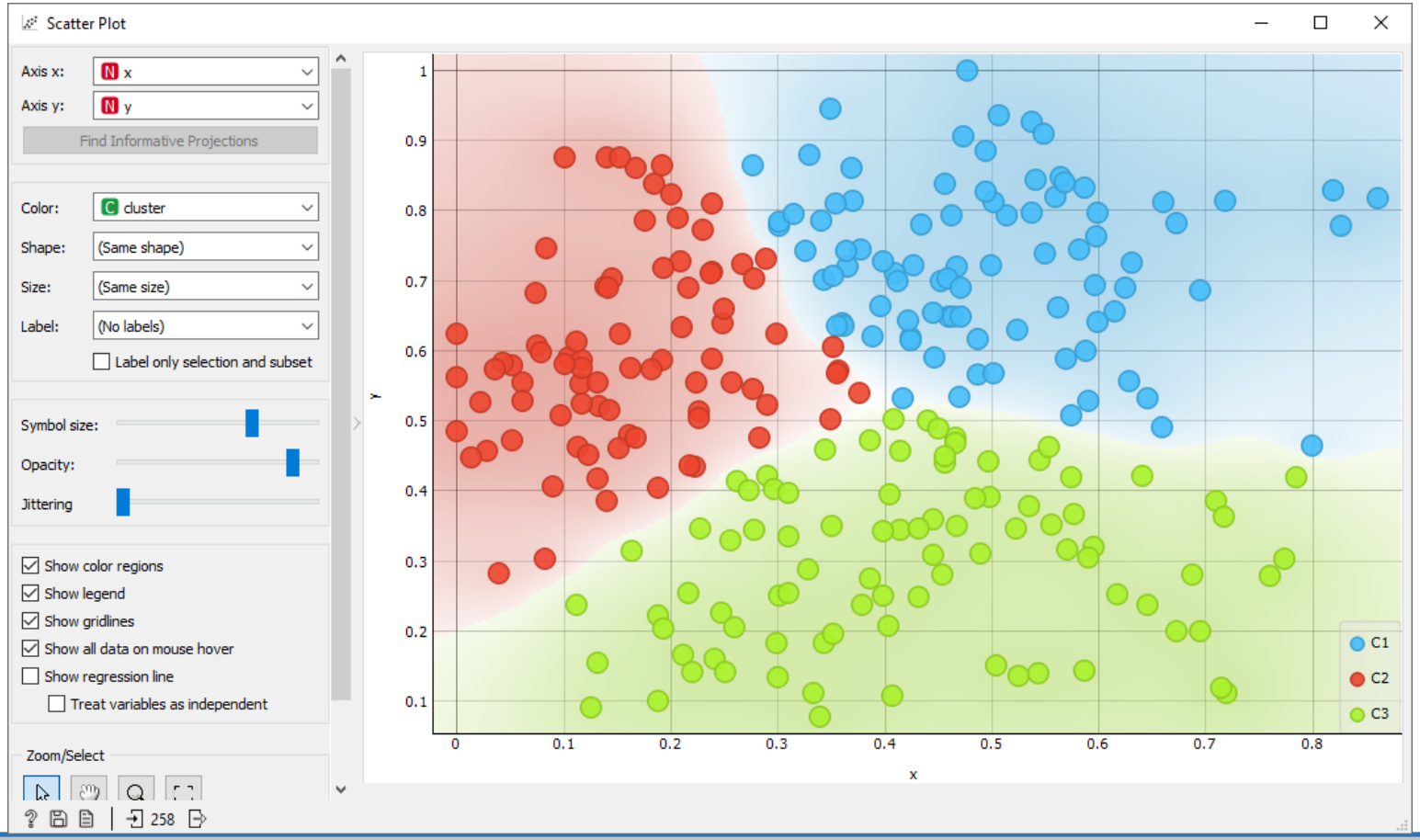




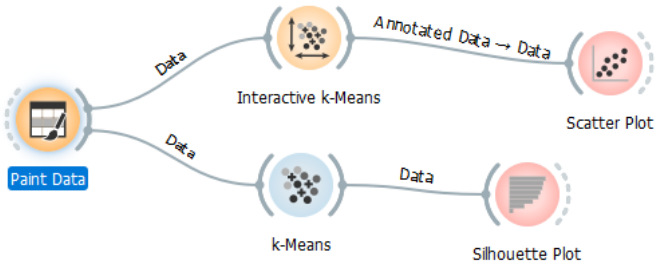
Silhouette index



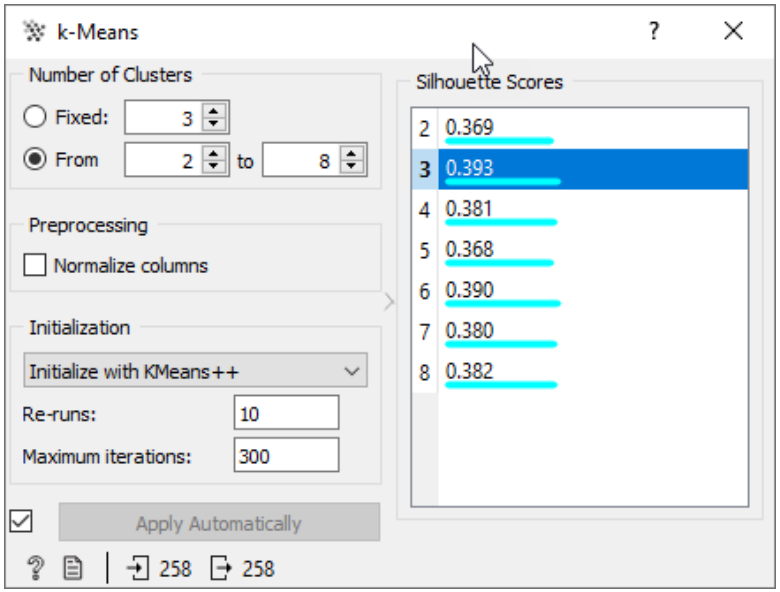
- Scatter Piot

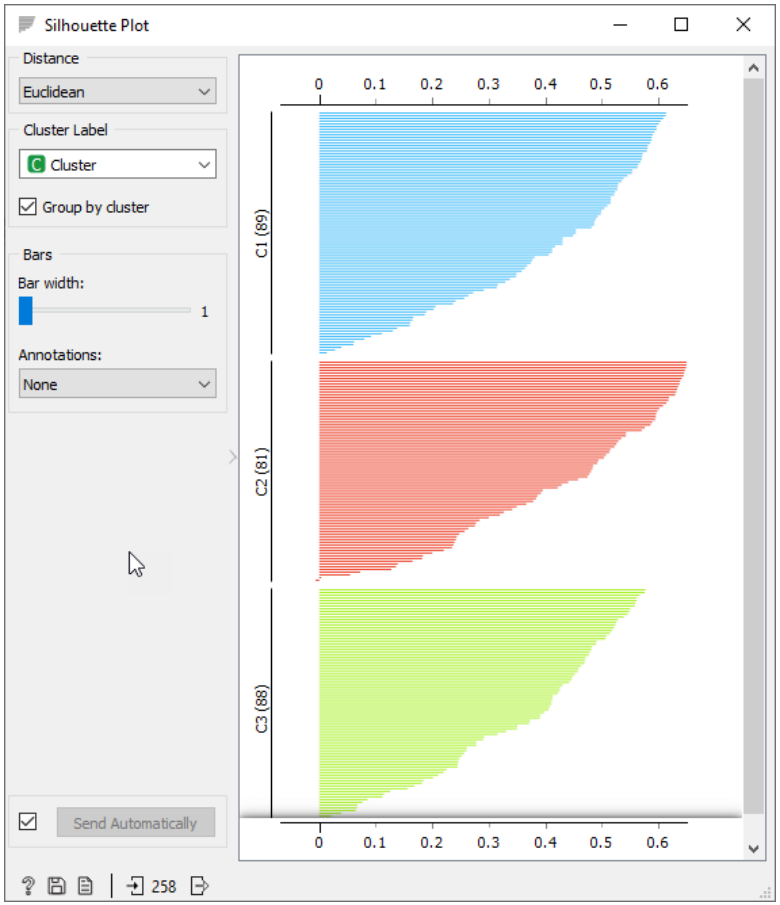


Silhouette index

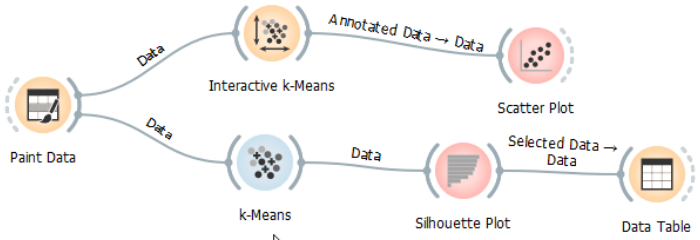


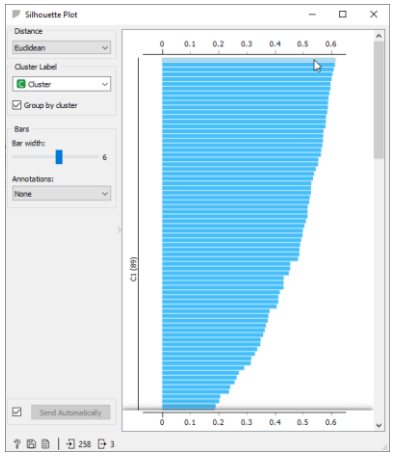
- K-Means - Silhouette Plot

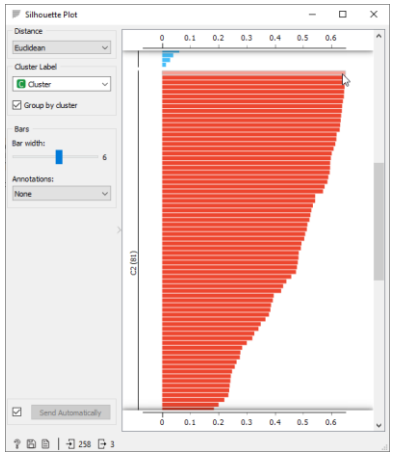


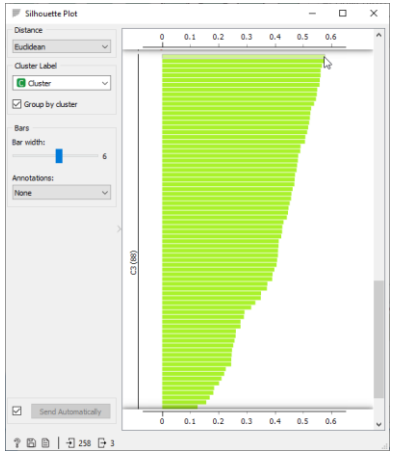


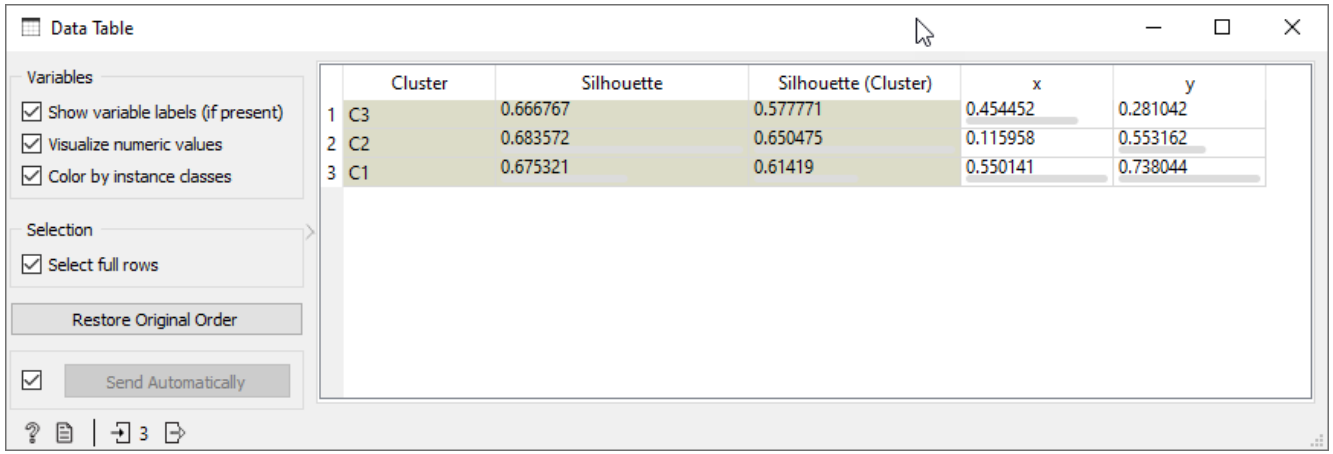
Silhouette Plot >0 là tốt











Silhouette index

- Đưa dữ liệu ra ngoài: Data table

Data Table

Variables

- ☑ Show variable labels (if present)
- ☑ Visualize numeric values
- ☑ Color by instance classes

Selection

- ☑ Select full rows

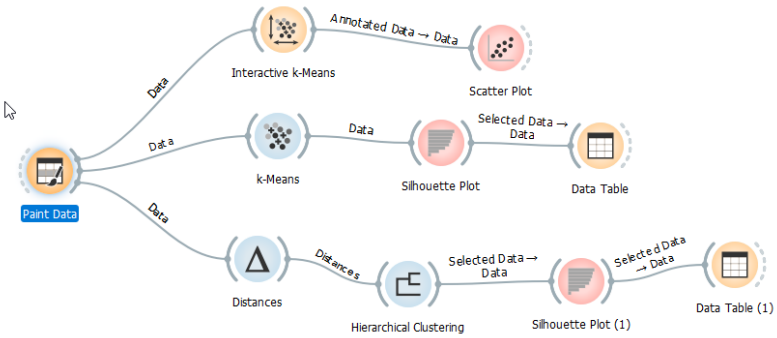
Restore Original Order

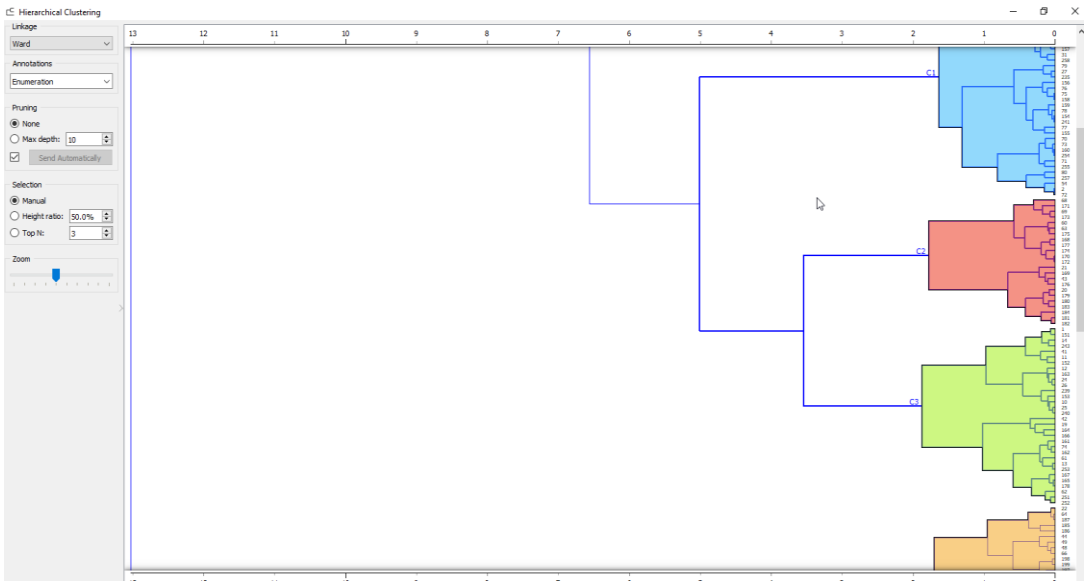
☑ Send Automatically

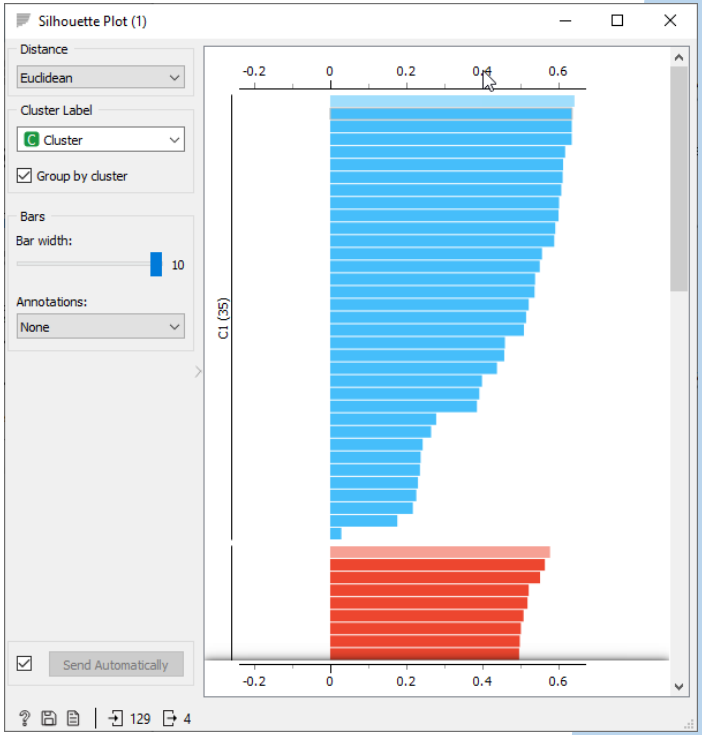
	Cluster	Silhouette	Silhouette (Cluster)	x	y
1	C3	0.666767	0.577771	0.454452	0.281042
2	C2	0.683572	0.650475	0.115958	0.553162
3	C1	0.675321	0.61419	0.550141	0.738044

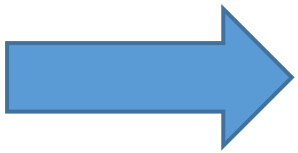
Silhouette index

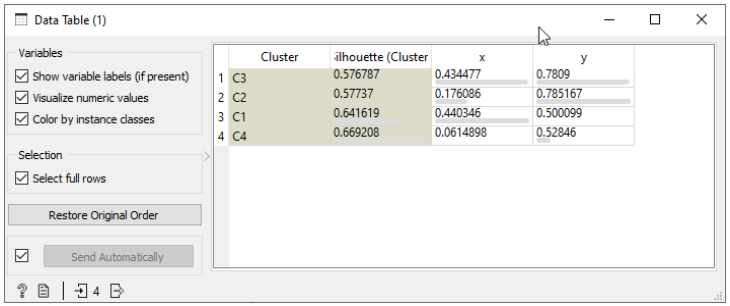
- Distances – Hieracical Clustering



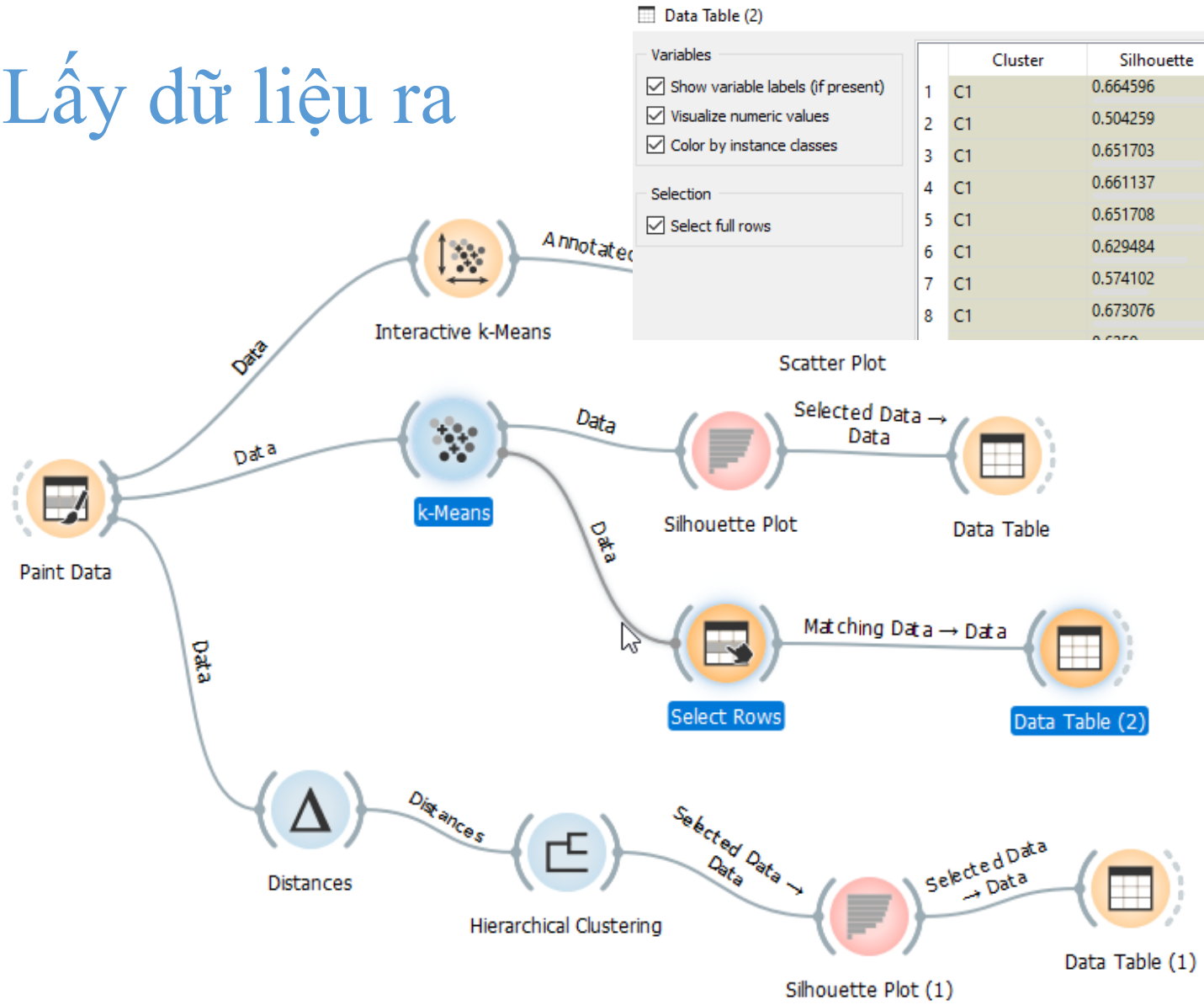








Lấy dữ liệu ra



Demo bằng công cụ Orange

- Quy trình thực hiện