



Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ



Giải thuật gom cụm Clustering algorithms

Đỗ Thanh Nghi
dtngghi@cit.ctu.edu.vn

Cần Thơ
12-02-2019

Nội dung

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

Nội dung

- **Giới thiệu về clustering**
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

Clustering

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

■ gom nhóm

- nature của dữ liệu thường không có nhiều thông tin sẵn có như lớp (nhãn)
- gom nhóm : mô hình gom cụm dữ liệu (không có nhãn) sao cho các dữ liệu cùng nhóm có các tính chất tương tự nhau và dữ liệu của 2 nhóm khác nhau sẽ có các tính chất khác nhau
- có nhiều nhóm giải thuật khác nhau : hierarchical clustering, partitioning, density-based, model-based, etc.
- được sử dụng nhiều : K-Means, Dendrogram, SOM, EM
- được ứng dụng thành công trong hầu hết các lĩnh vực tìm kiếm thông tin, phân tích dữ liệu, etc.

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

Top 10 DM algorithms (2015)



Here are the algorithms:

- 1. C4.5
- 2. k-means
- 3. Support vector machines
- 4. Apriori
- 5. EM
- 6. PageRank
- 7. AdaBoost
- 8. kNN
- 9. Naive Bayes
- 10. CART

Clustering

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

■ gom nhóm

- thường dựa trên cơ sở khoảng cách
- nên chuẩn hóa dữ liệu
- khoảng cách được tính theo từng kiểu của dữ liệu : số thực, nhị phân, rời rạc

Kiểu số thực

- khoảng cách *Minkowski*

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

$i = (x_{i1}, x_{i2}, \dots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \dots, x_{jp})$ là 2 phần tử dữ liệu trong p -dimensional, q là số nguyên dương

- nếu $q = 1$, d là khoảng cách Manhattan
- nếu $q = 2$, d là khoảng cách Euclid
- khoảng cách cosine : $d_{\cos}(i, j) = i^T j / (\|i\| \|j\|)$

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

Kiểu nhị phân

		Object <i>j</i>		
		1	0	<i>sum</i>
Object <i>i</i>	1	<i>a</i>	<i>b</i>	<i>a+b</i>
	0	<i>c</i>	<i>d</i>	<i>c+d</i>
	<i>sum</i>	<i>a+c</i>	<i>b+d</i>	<i>p</i>

■ khoảng cách đối xứng :
$$d(i, j) = \frac{b+c}{a+b+c+d}$$

■ khoảng cách bất đối xứng :
$$d(i, j) = \frac{b+c}{a+b+c}$$

■ hệ số Jaccard bất đối xứng :
$$sim_{Jaccard}(i, j) = \frac{a}{a+b+c}$$

Kiểu rời rạc (nominal type)

- ví dụ : thuộc tính color có giá trị là red, green, blue, etc.
 - phương pháp matching đơn giản, m là số lượng matches và p là tổng số biến (thuộc tính), khoảng cách được định nghĩa :

$$d(i, j) = \frac{p - m}{p}$$

Nội dung

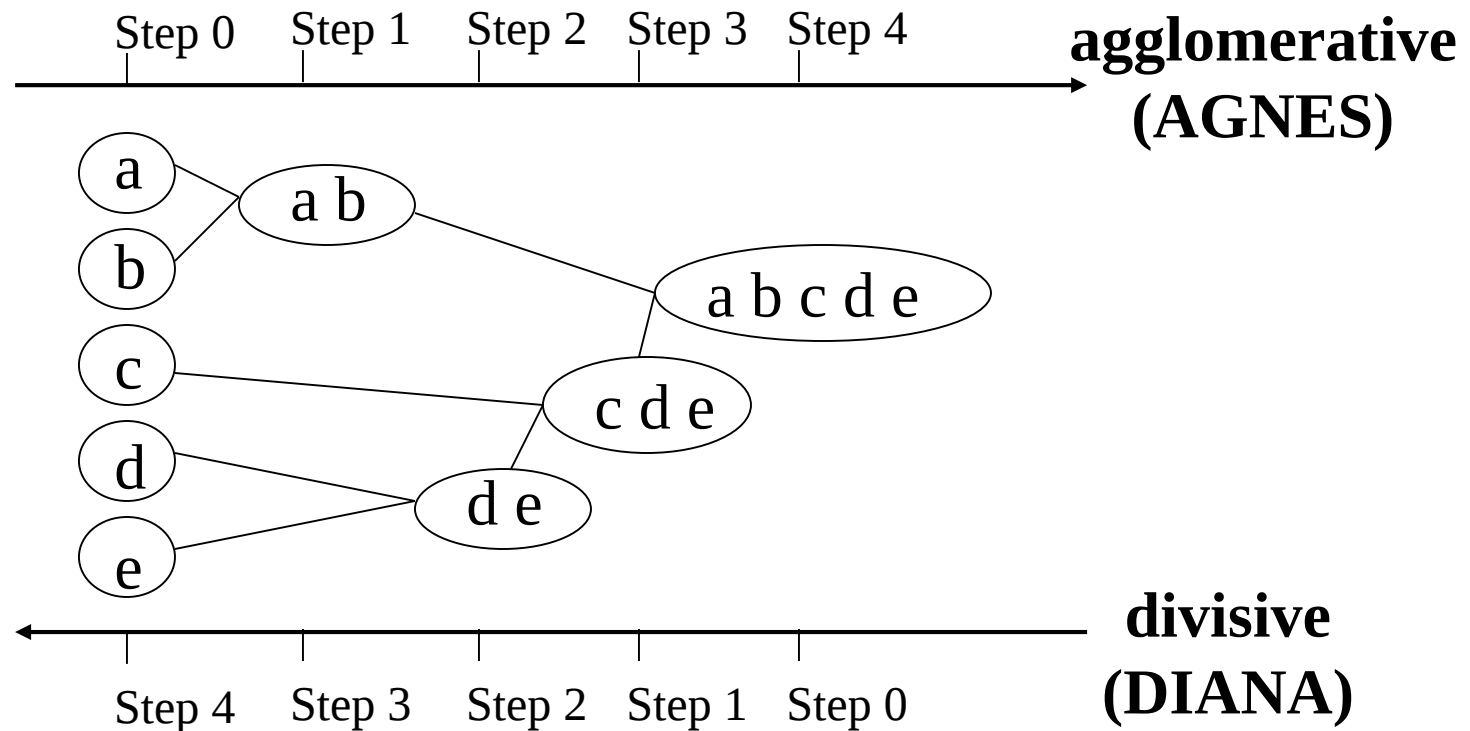
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering

- bottom up
 - bắt đầu với những clusters chỉ là 1 phần tử
 - ở mỗi bước, merge 2 clusters gần nhau thành 1
 - khoảng cách giữa 2 clusters : 2 điểm gần nhất từ 2 clusters, hoặc khoảng cách trung bình, etc.
- top down
 - bắt đầu với 1 cluster là tất cả dữ liệu
 - tìm 2 clusters con
 - tiếp tục đệ quy trên 2 clusters con
- kết quả sinh ra dendrogram

Hierarchical clustering

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



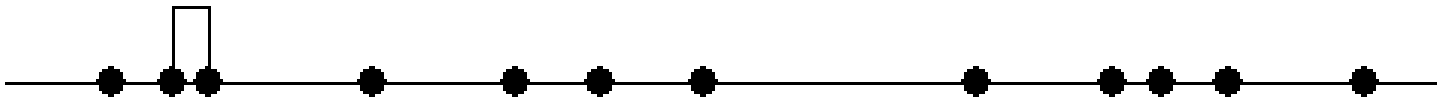
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



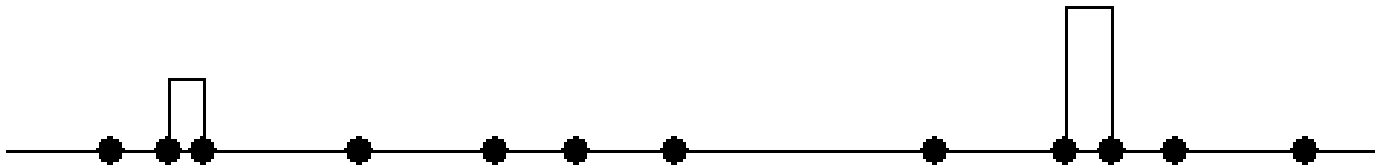
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



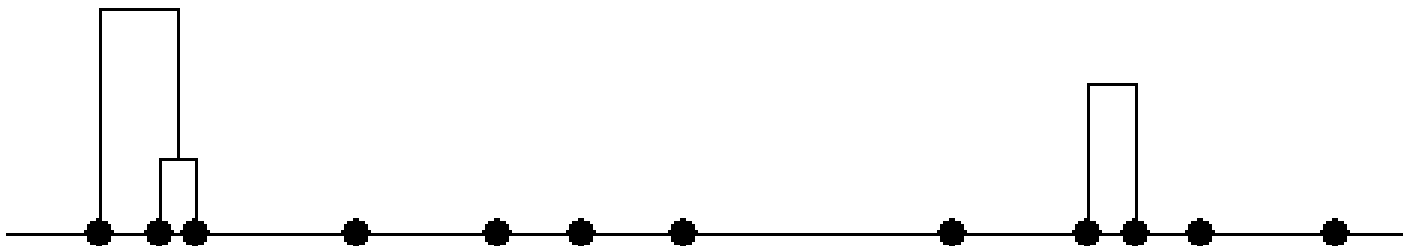
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



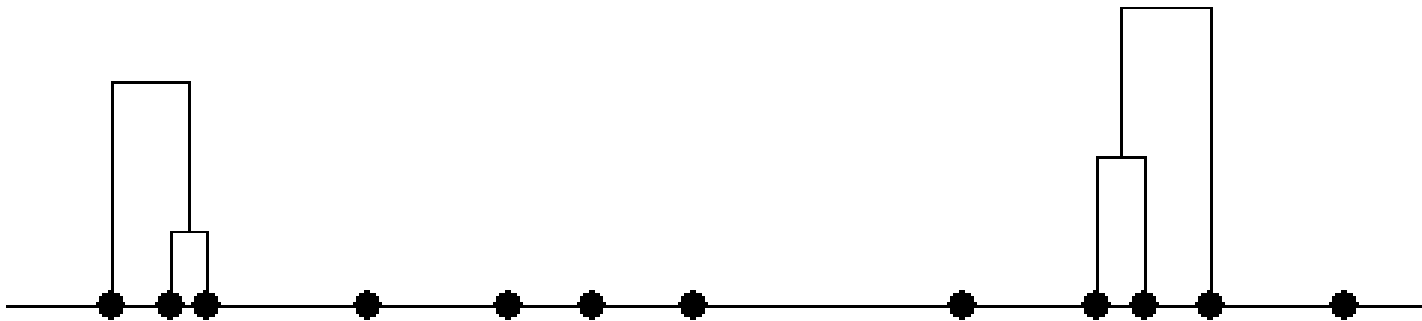
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



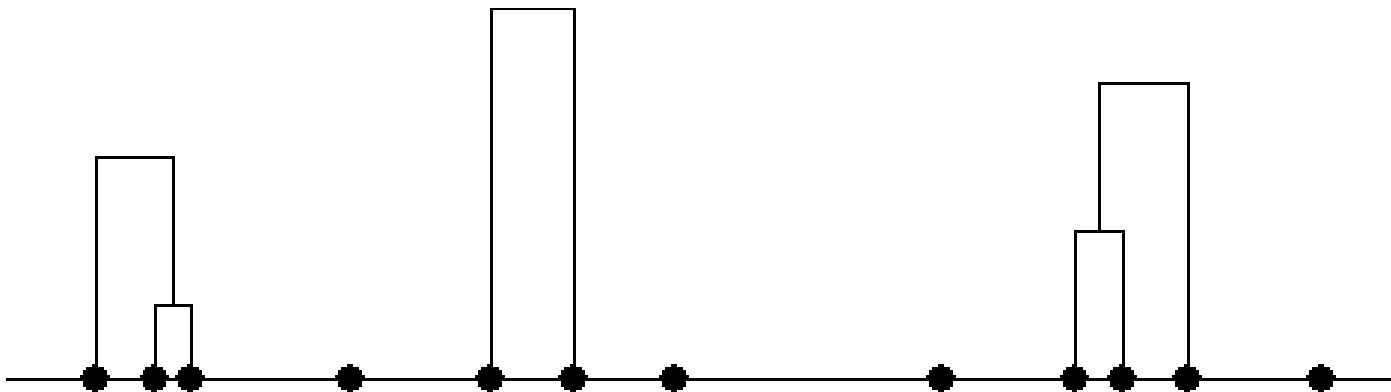
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



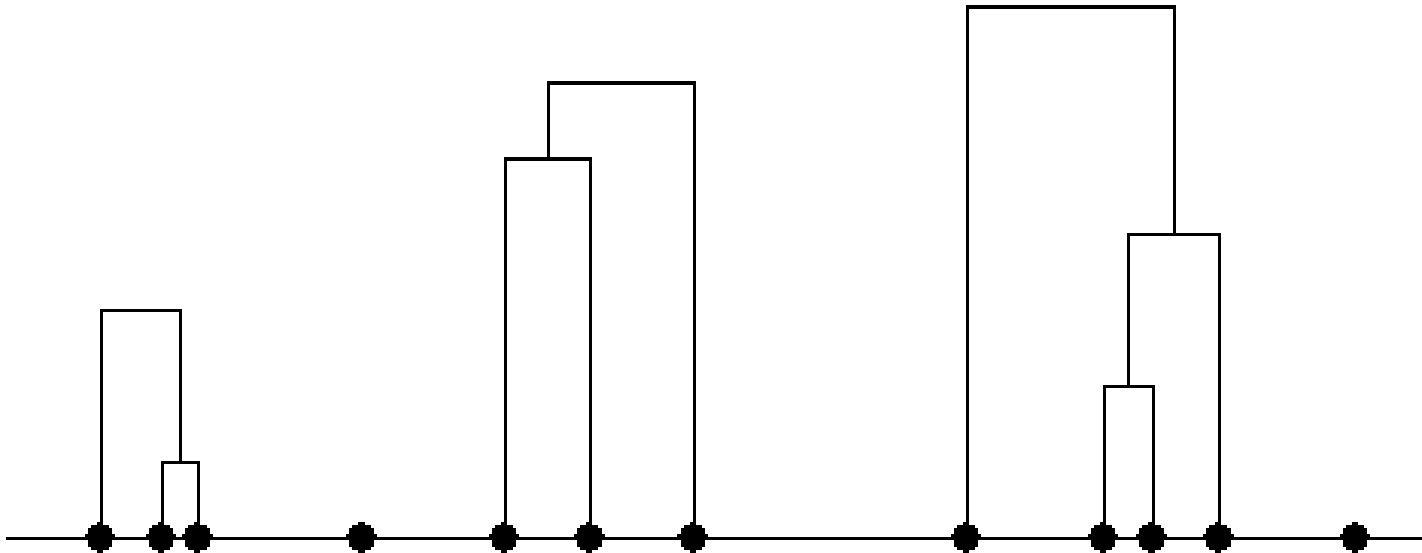
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



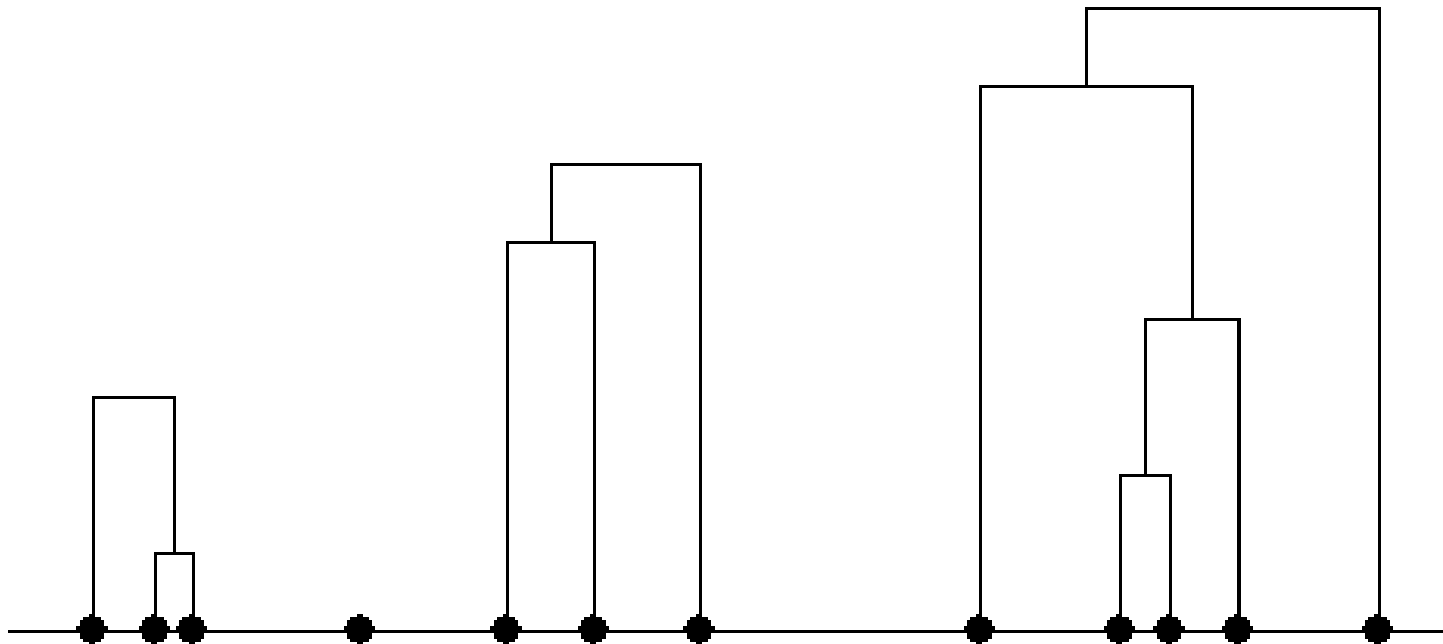
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



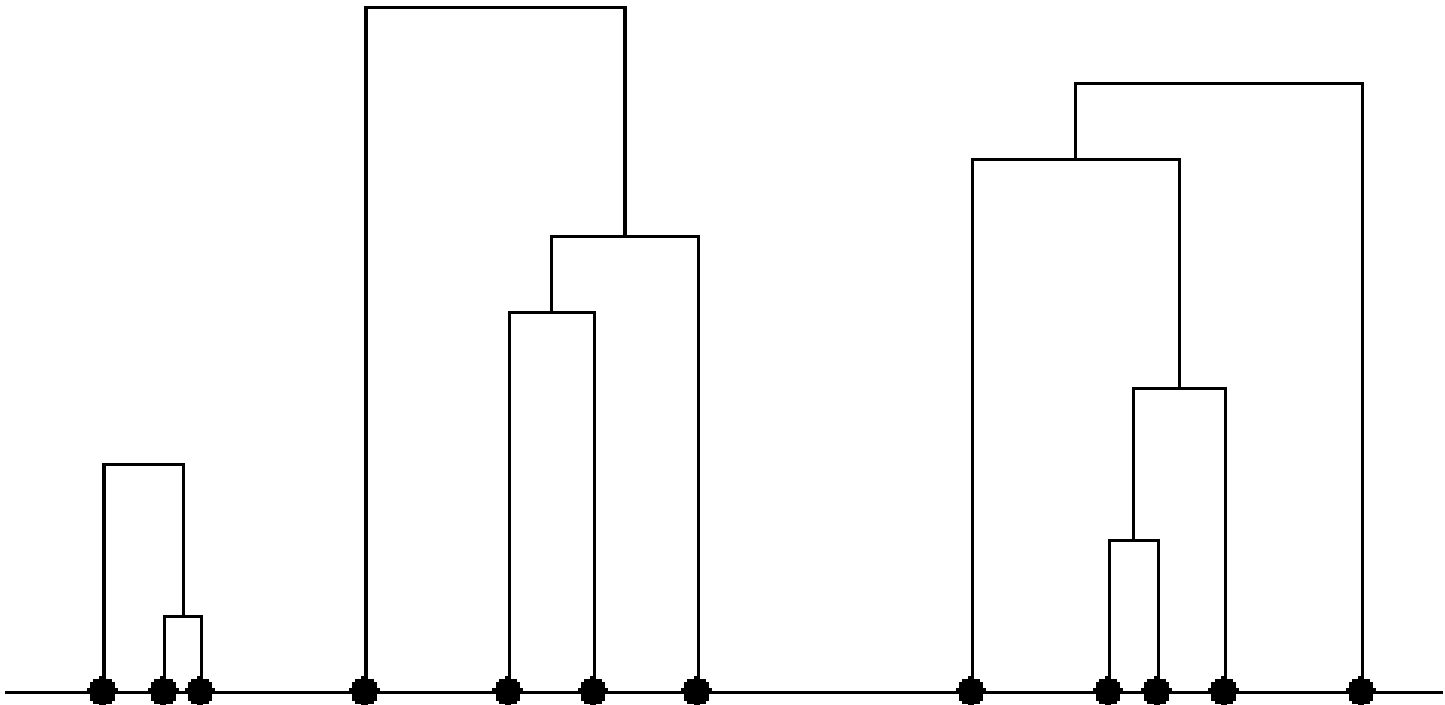
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



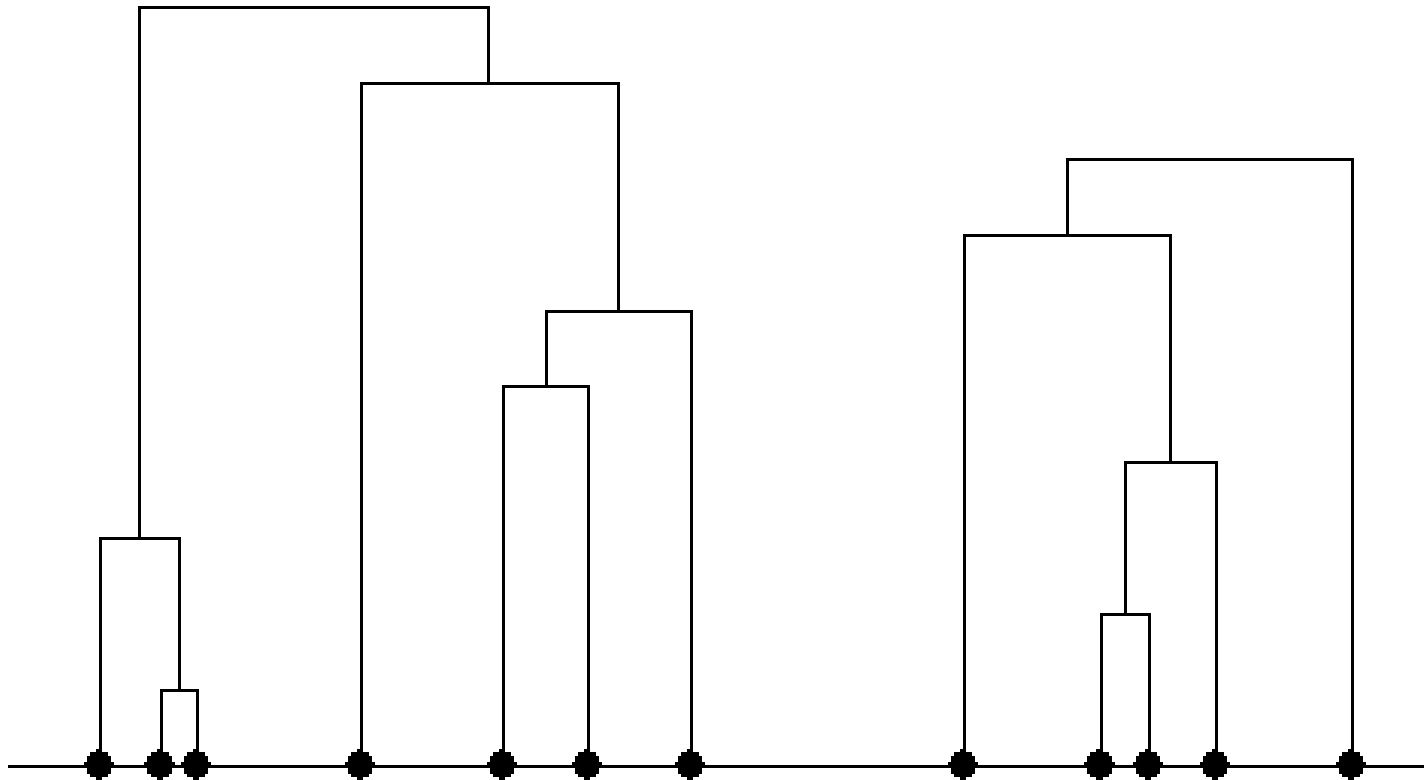
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



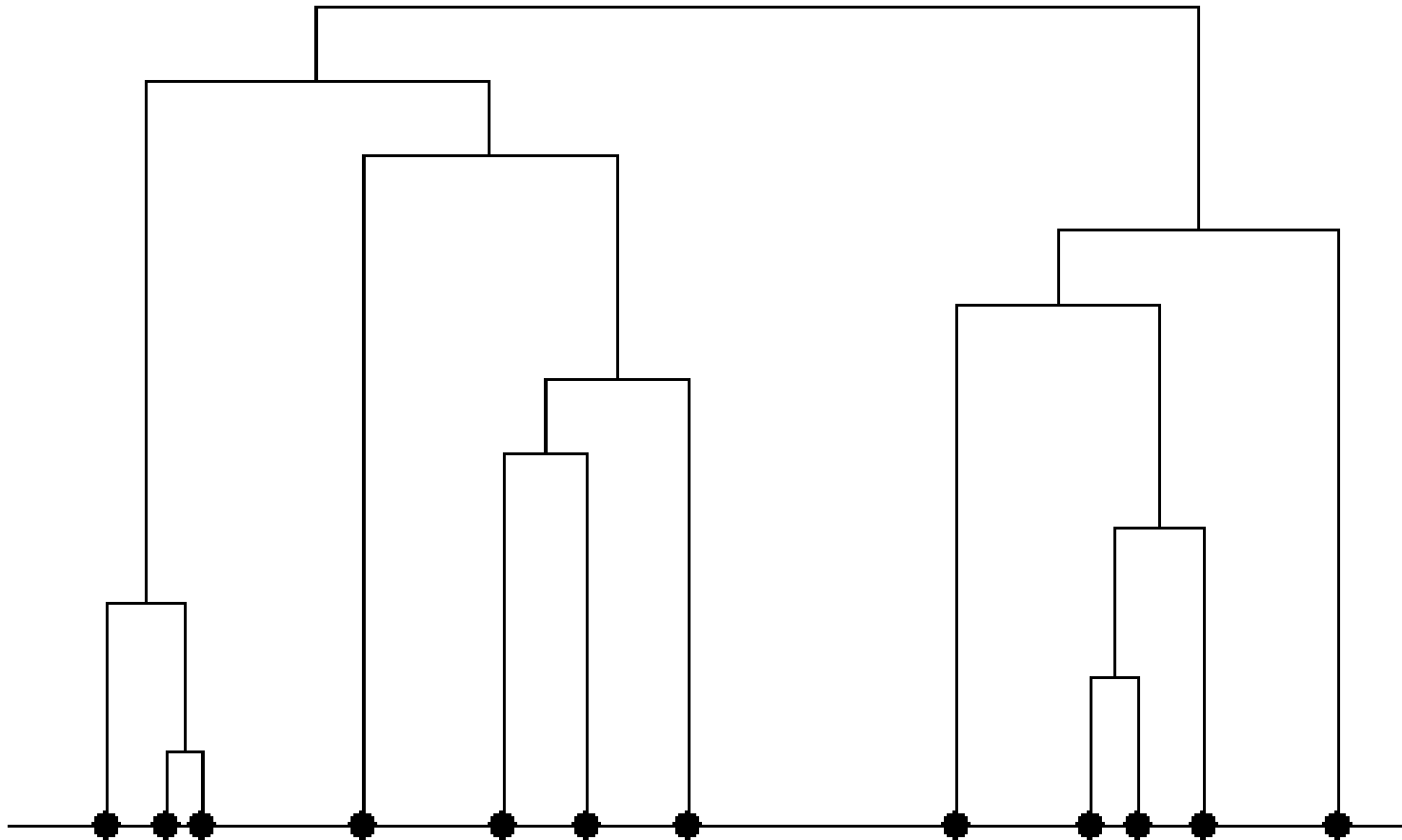
Hierarchical clustering (Single link)

- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



Hierarchical clustering (Single link)

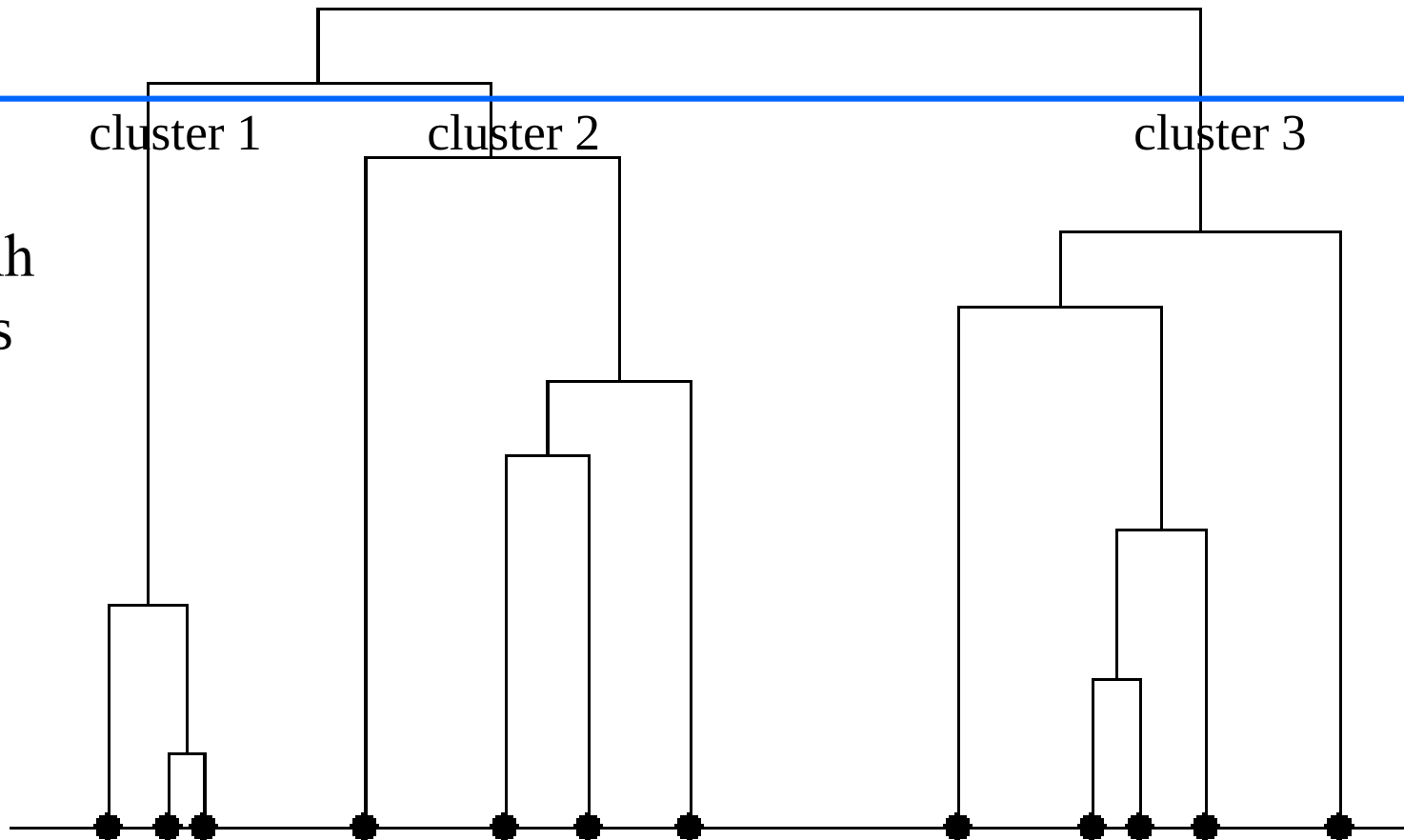
- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển



- Giới thiệu về clustering
- **Hierarchical clustering**
- K-Means
- Kết luận và hướng phát triển

Hierarchical clustering (Single link)

cắt =>
xác định
clusters



Hierarchical clustering

- nhận xét
 1. giải thuật đơn giản
 2. cho kết quả dễ hiểu
 3. không cần tham số
 4. chạy chậm
 5. BIRCH (Zhang et al., 1996) sử dụng cấu trúc index để xử lý dữ liệu lớn

Nội dung

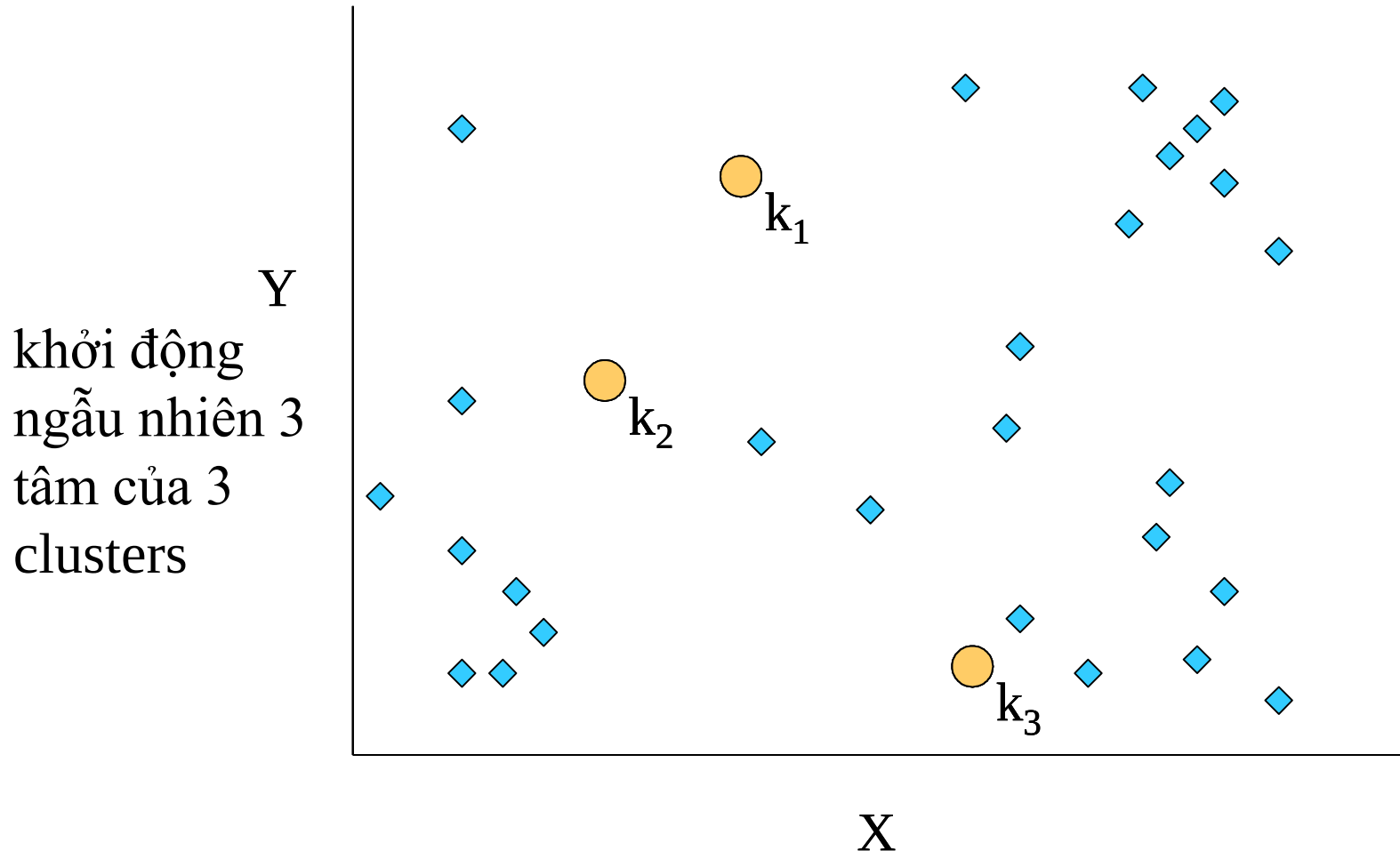
- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

Giải thuật K-Means

- giải thuật
 1. khởi động ngẫu nhiên K tâm (center) của K clusters
 2. mỗi phần tử được gán cho tâm gần nhất với phần tử dựa vào khoảng cách (e.g. khoảng cách Euclid)
 3. cập nhật lại các tâm của K clusters, mỗi tâm là giá trị trung bình (mean) của các phần tử trong cluster của nó
 4. lặp lại bước 2,3 cho đến khi hội tụ

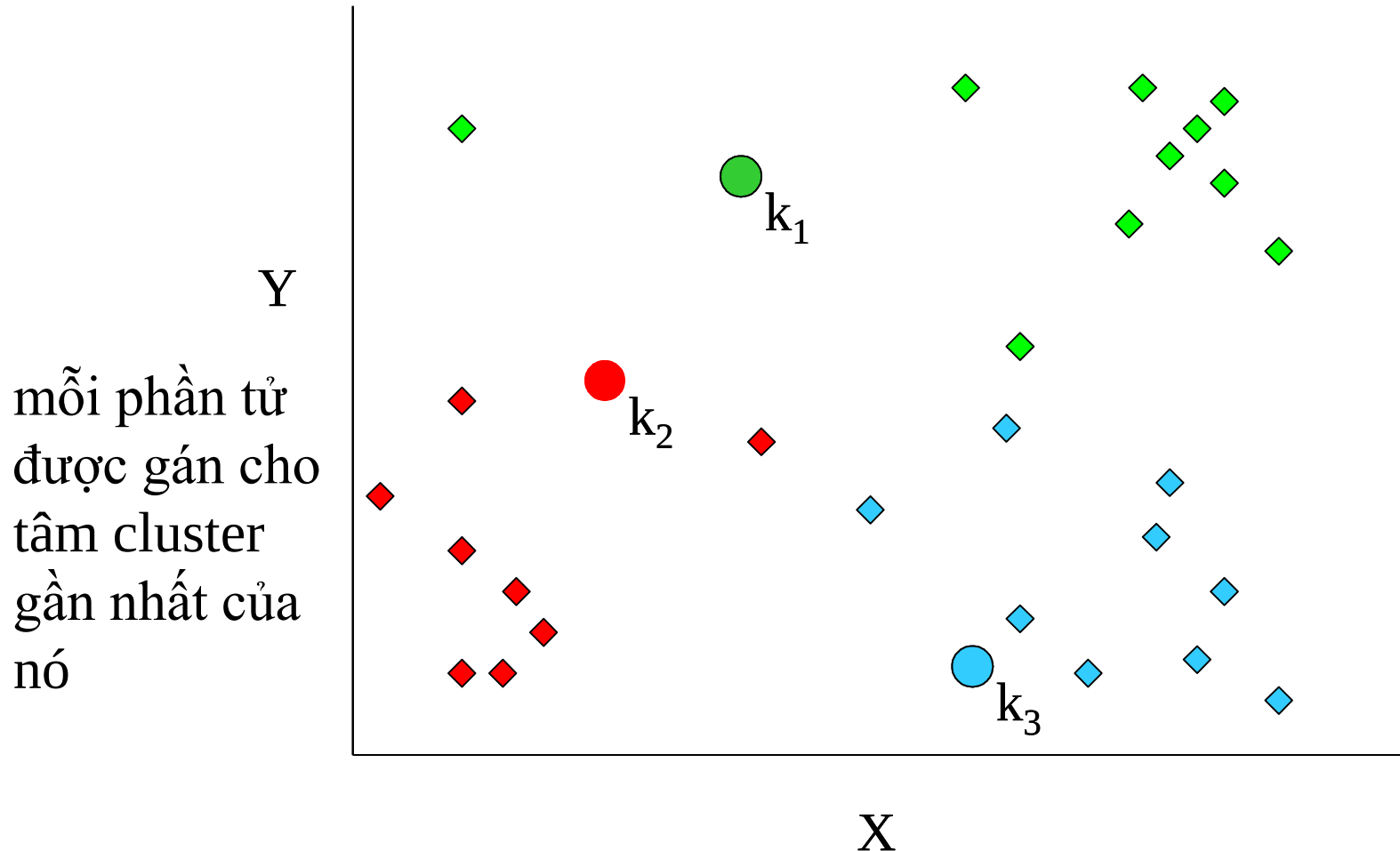
Giải thuật K-Means

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển



Giải thuật K-Means

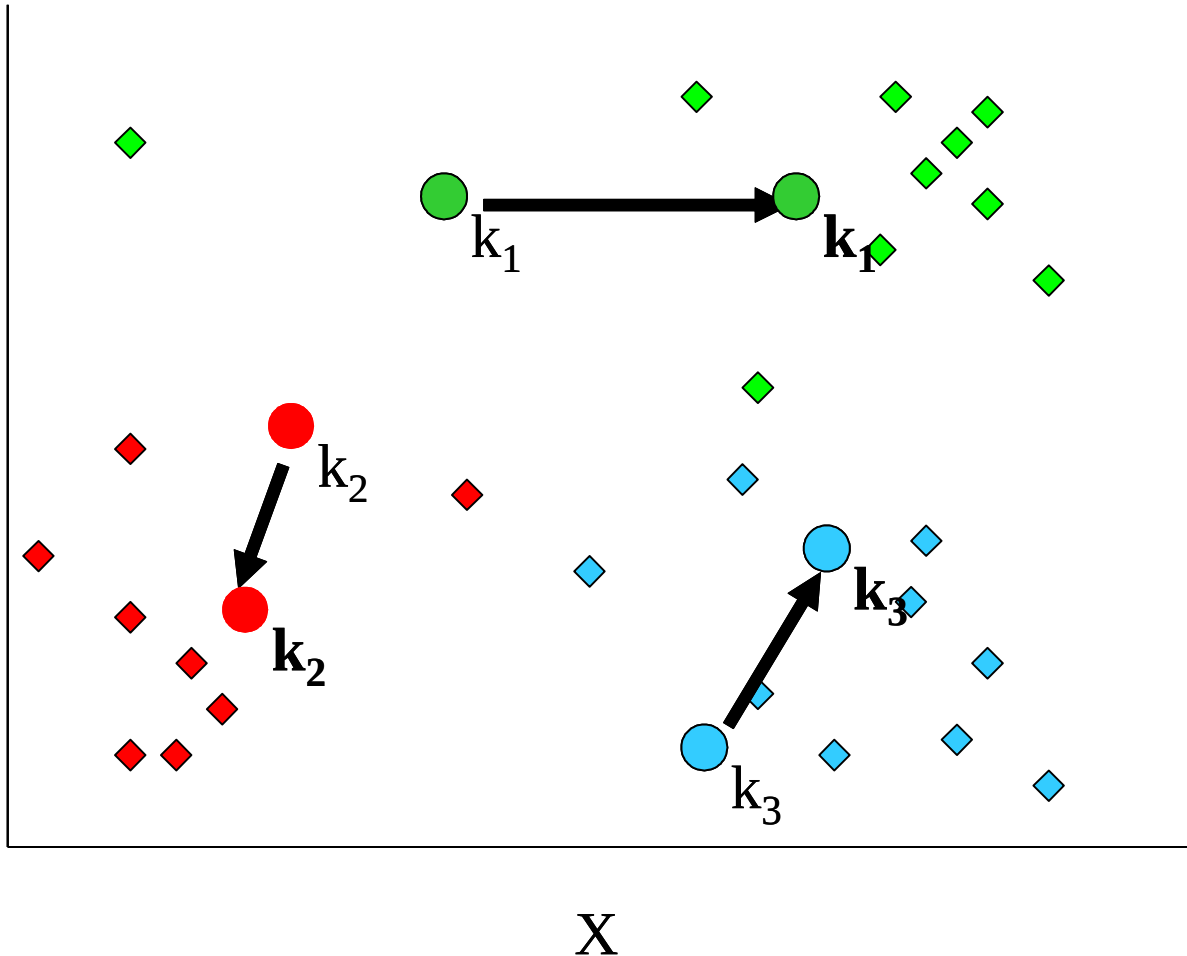
- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển



Giải thuật K-Means

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

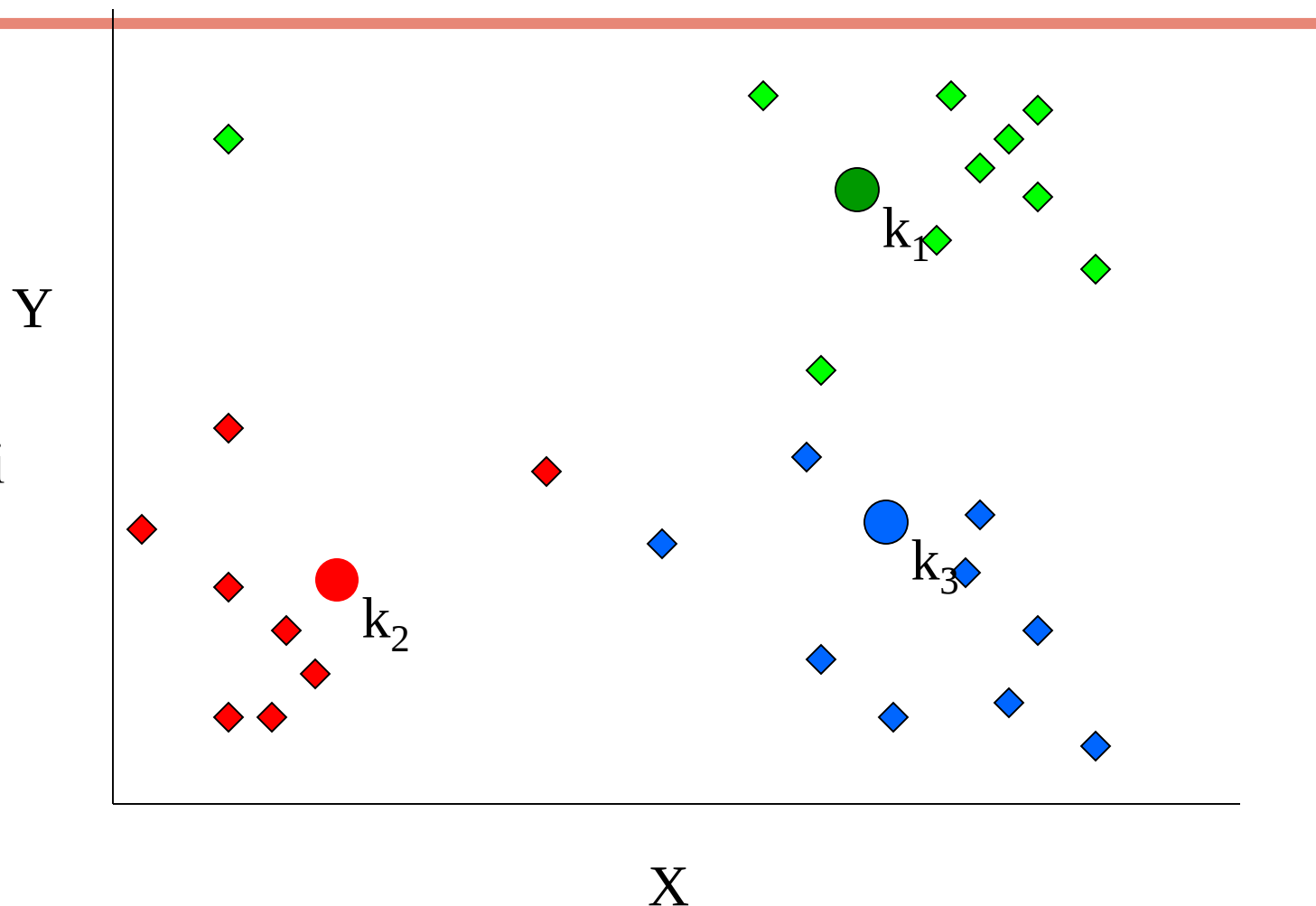
cập nhật lại
tâm của các
cluster (giá trị
trung bình của
các phần tử
trong cluster)



Giải thuật K-Means

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

cấu hình mới
của lần lặp
tiếp theo

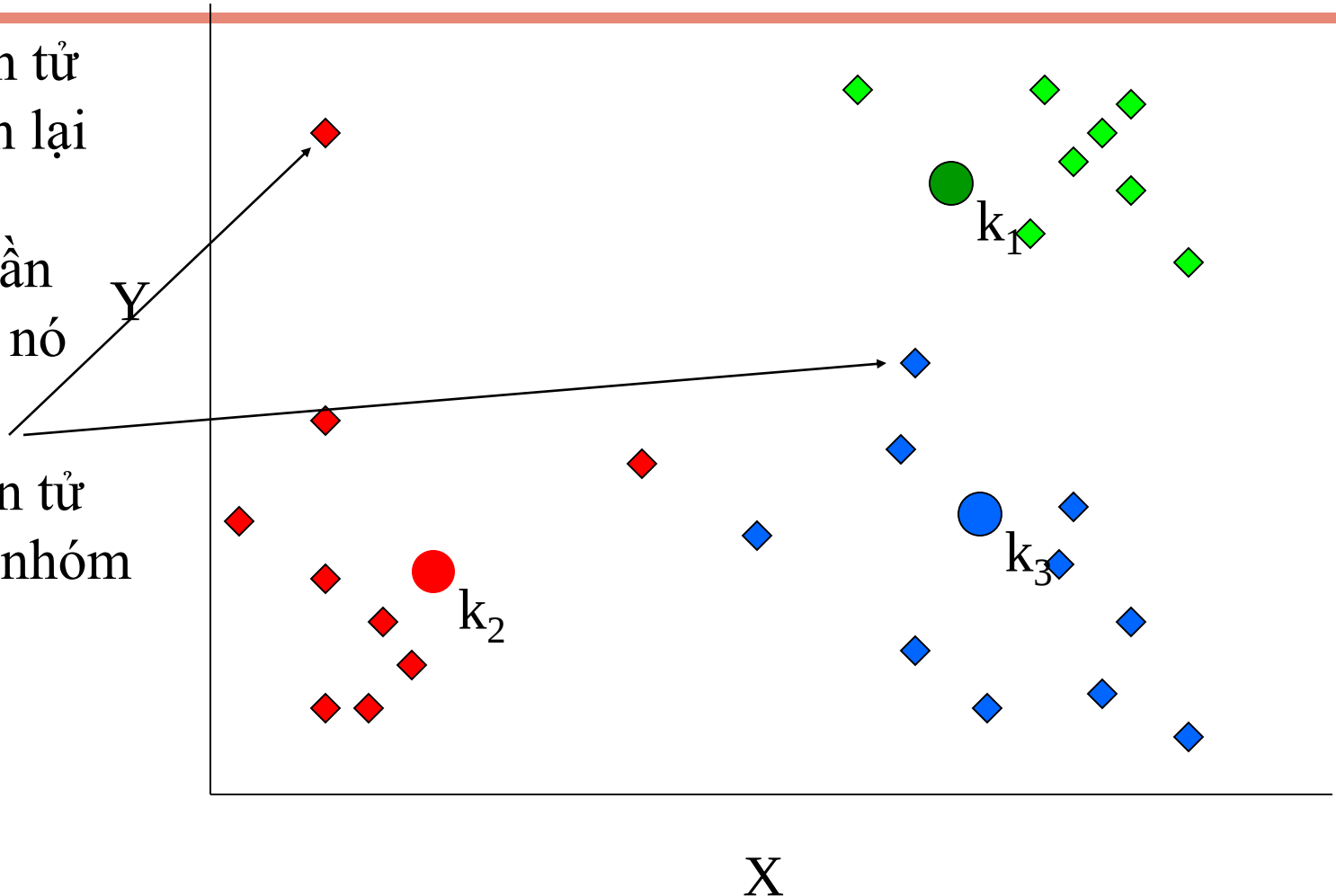


Giải thuật K-Means

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

mỗi phần tử
được gán lại
cho tâm
cluster gần
nhất của nó

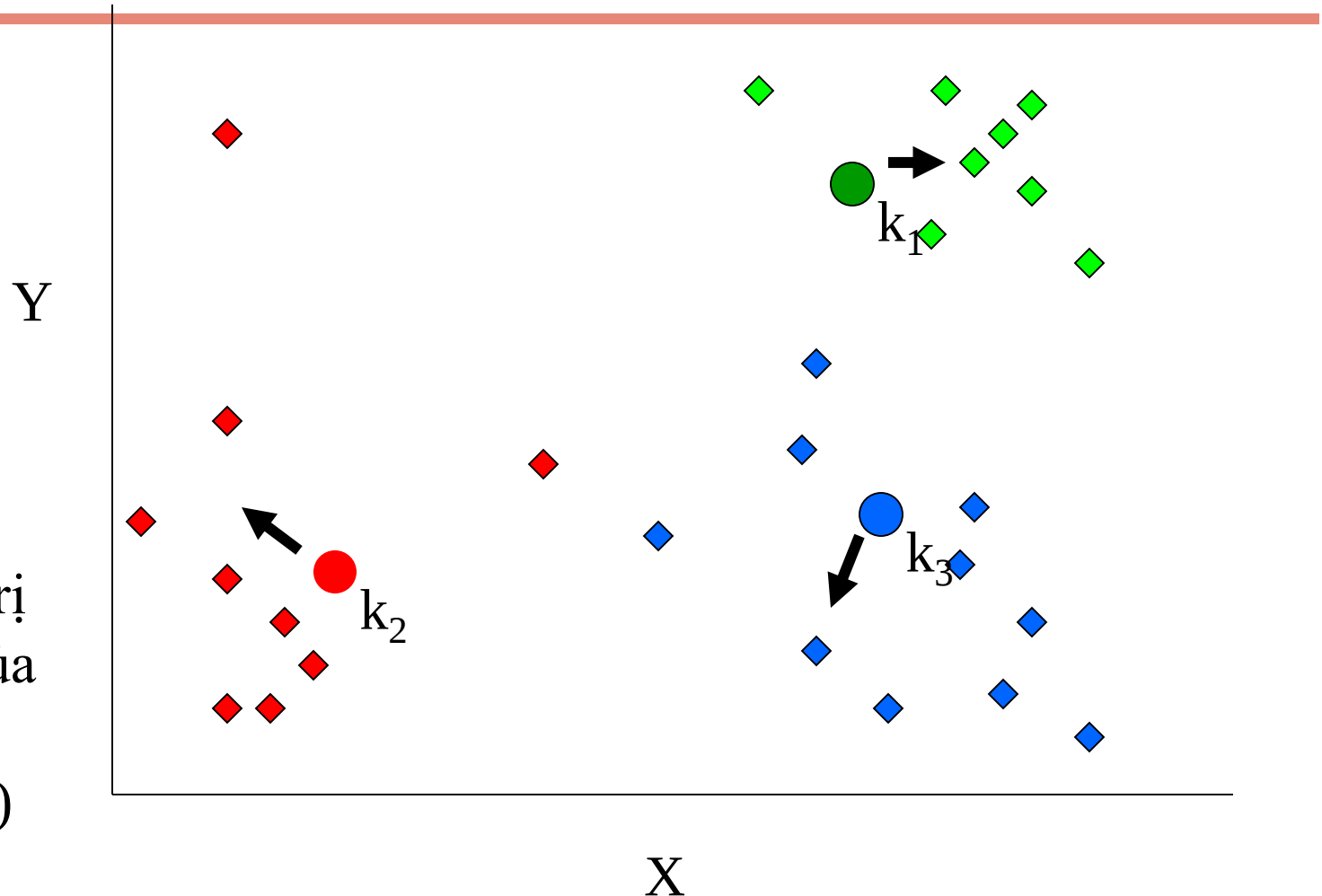
có 2 phần tử
thay đổi nhóm



Giải thuật K-Means

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

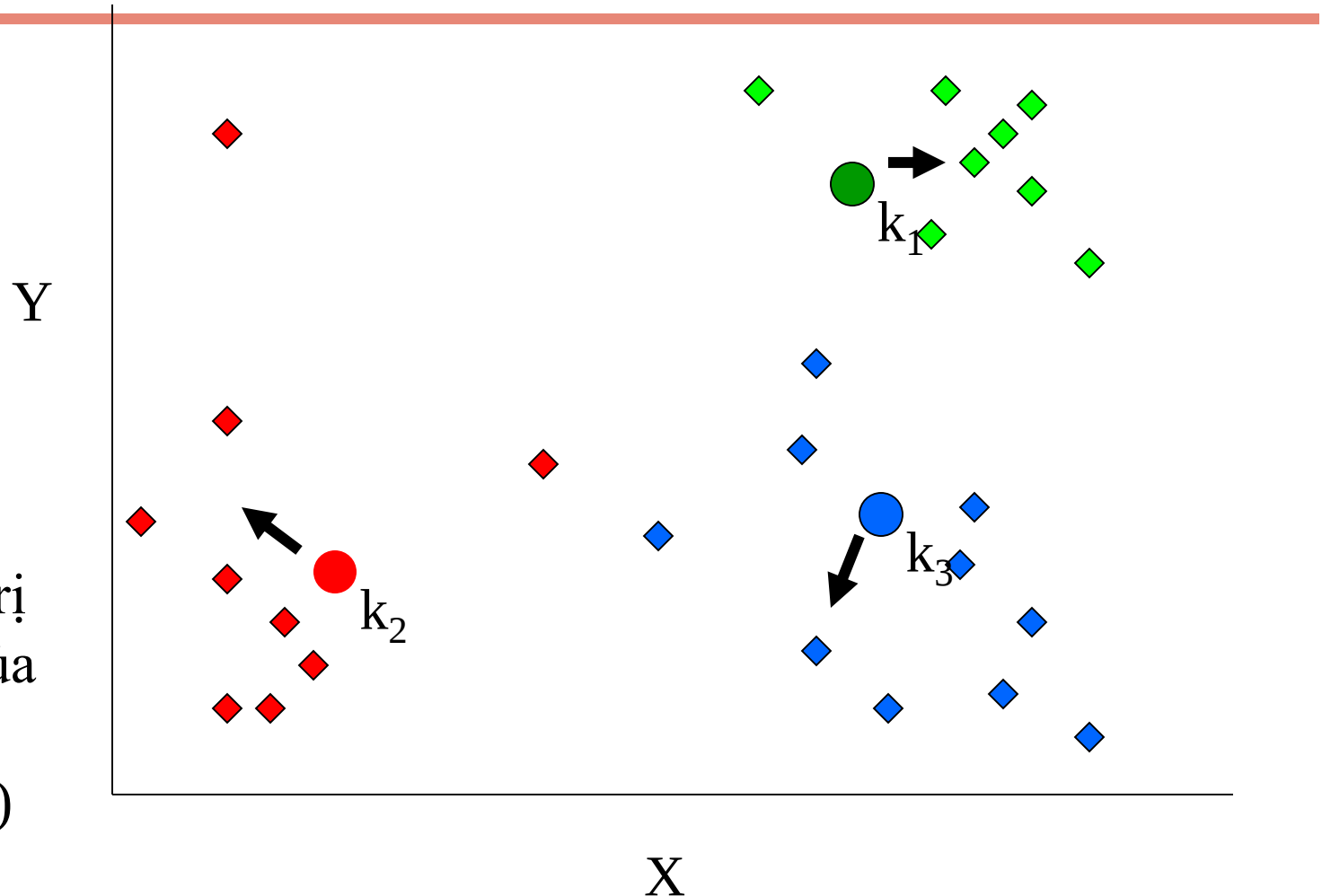
cập nhật lại
tâm của các
cluster (giá trị
trung bình của
các phần tử
trong cluster)



Giải thuật K-Means

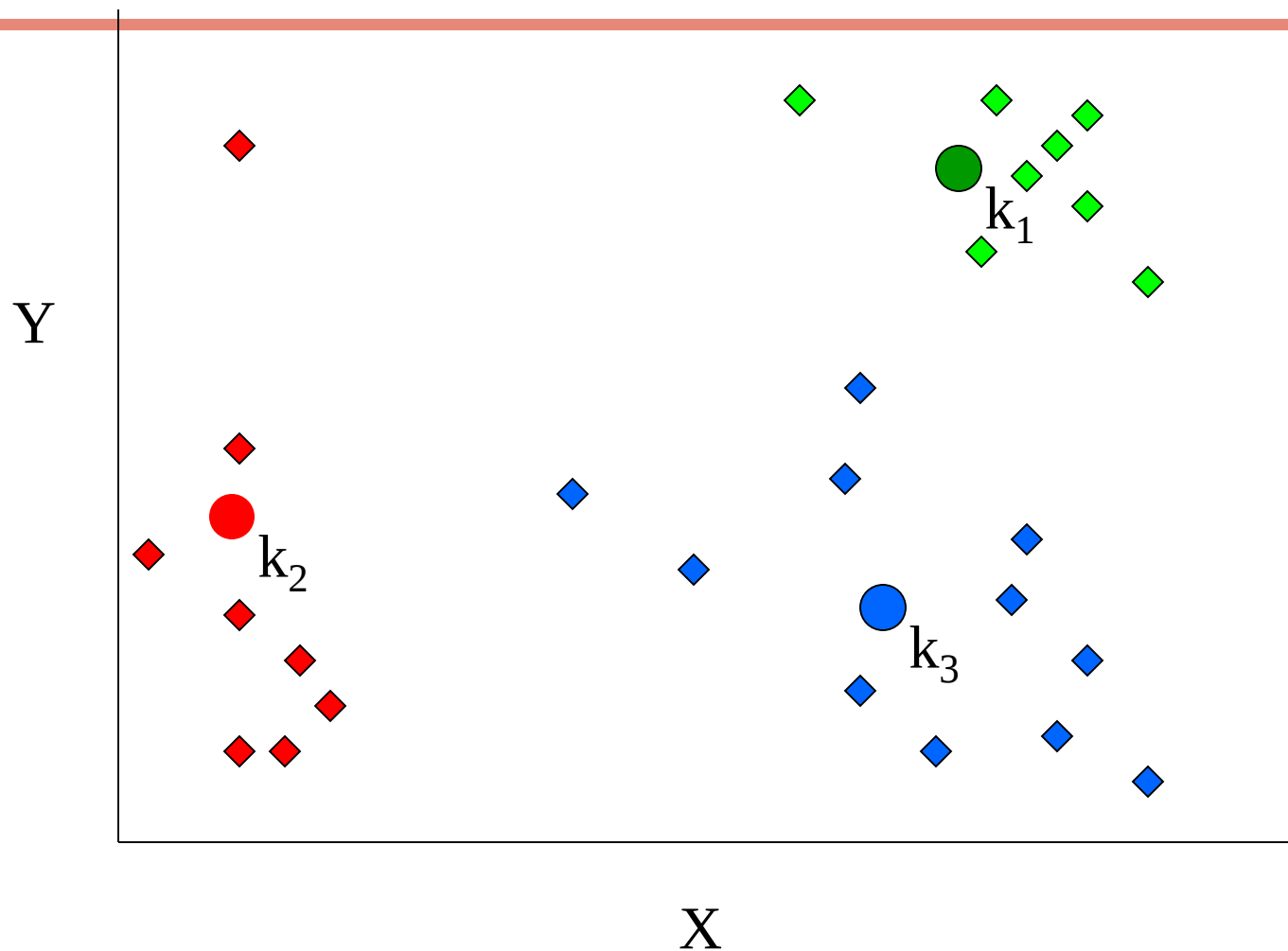
- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển

cập nhật lại
tâm của các
cluster (giá trị
trung bình của
các phần tử
trong cluster)



Giải thuật K-Means

- Giới thiệu về clustering
- Hierarchical clustering
- **K-Means**
- Kết luận và hướng phát triển



Giải thuật K-Means

- nhận xét
 1. giải thuật đơn giản
 2. cho kết quả dễ hiểu
 3. cần cho tham số K (số lượng clusters)
 4. kết quả phụ thuộc vào việc khởi động ngẫu nhiên K tâm (center) của K clusters : có thể khắc phục bằng cách khởi động lại nhiều lần.
 5. khả năng chịu đựng nhiễu không tốt (ảnh hưởng bởi các phần tử outliers) : có thể khắc phục bằng K-Medoids, không sử dụng giá trị trung bình, nhưng sử dụng phần tử ngay giữa

Nội dung

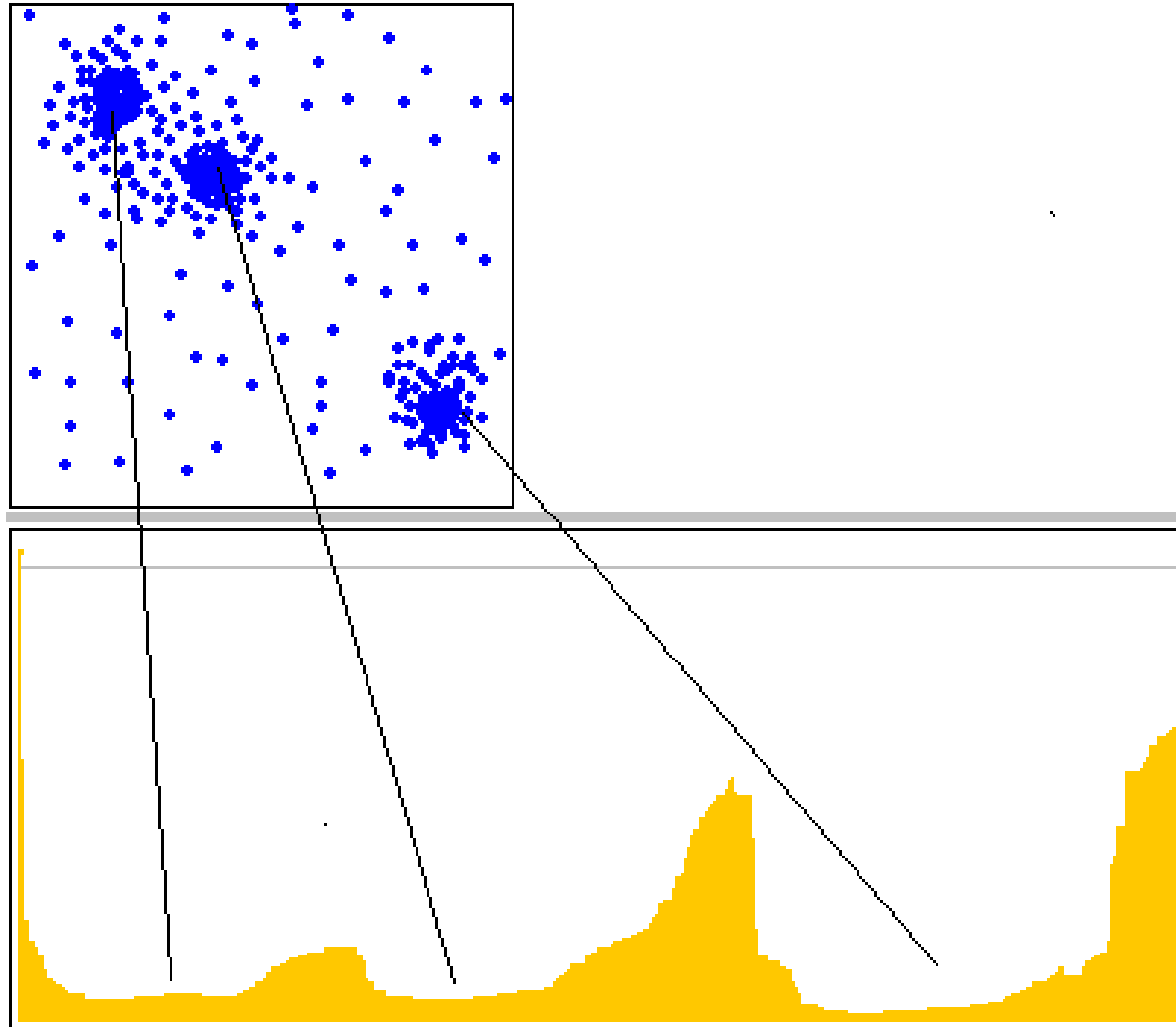
- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- **Kết luận và hướng phát triển**

Giải thuật clustering

- còn nhiều phương pháp khác
 - density-based : DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), DENCLUE (Hinneburg & Keim, 1998)
 - model-based : EM (Expected maximization), SOM (Kohonen, 1995)

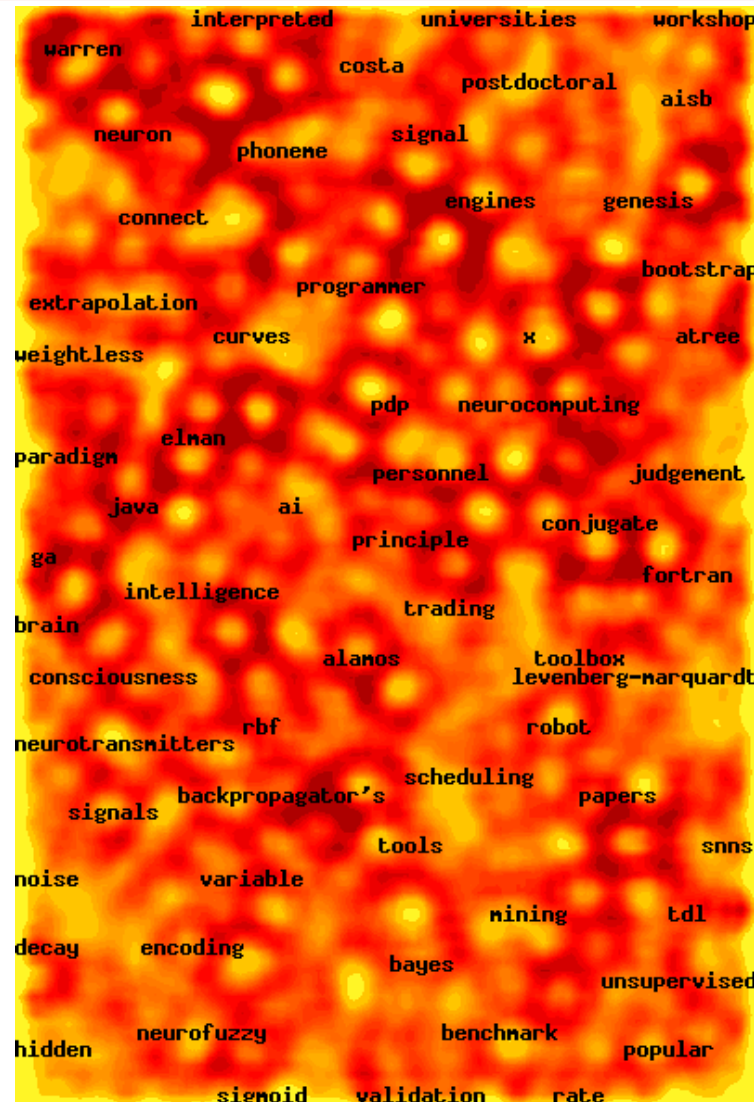
Clustering với OPTICS

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển



Clustering 12088 web articles với SOM

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển



Hướng phát triển²

- Giới thiệu về clustering
- Hierarchical clustering
- K-Means
- Kết luận và hướng phát triển

- các kiểu dữ liệu phức tạp
- tăng tốc độ xử lý
- các tham số đầu vào của giải thuật
- diễn dịch kết quả sinh ra
- phương pháp kiểm chứng chất lượng mô hình



Cám ơn !