Khoa Công Nghệ Thông Tin
Trường Đại Học Cần Thơ

Giải thuật gom cụm
Clustering algorithms

Đỗ Thanh Nghị
dtnghi@cit.ctu.edu.vn

Cần Thơ
02-12-2008

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

- 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 lãnh vực tìm kiếm thông tin, phân tích dữ liệu, etc.

Clustering

KDnuggets : Polls : Deployed data mining techniques

Poll

Which data mining techniques you used in a successfully deployed application? [173 voters, 509 votes total]

Decision Trees/Rules (88)	17%
Logistic regression (74)	15%
K-means clustering (54)	11%
Association rules (48)	9%
Neural Networks (45)	9%
SVM (33)	6%
Other methods (29)	6%
Self-organizing maps (24)	5%
Other clustering (24)	5%
Hybrid methods (20)	4%
Naive Bayes (19)	4%
Nearest Neighbor (17)	3%
Genetic algorithms (16)	3%
Fuzzy methods (11)	2%
Bayes nets (6)	1%
None (1)	0%

Clustering

- 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, nhị phân, loại, kiểu symbol (interval, histogram, taxonomy

Kiểu số

- khoảng cách *Minkowski*

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

$i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ và $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ 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 : $d_{cos}(i, j) = i^T j/(||i||\ ||j||)$

Kiểu nhị phân

		Object *j*		
		1	0	*sum*
Object *i*	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

- 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 loại (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 :

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

Kiểu symbol

- xem trang publications của Edwin DIDAY và các cộng sự

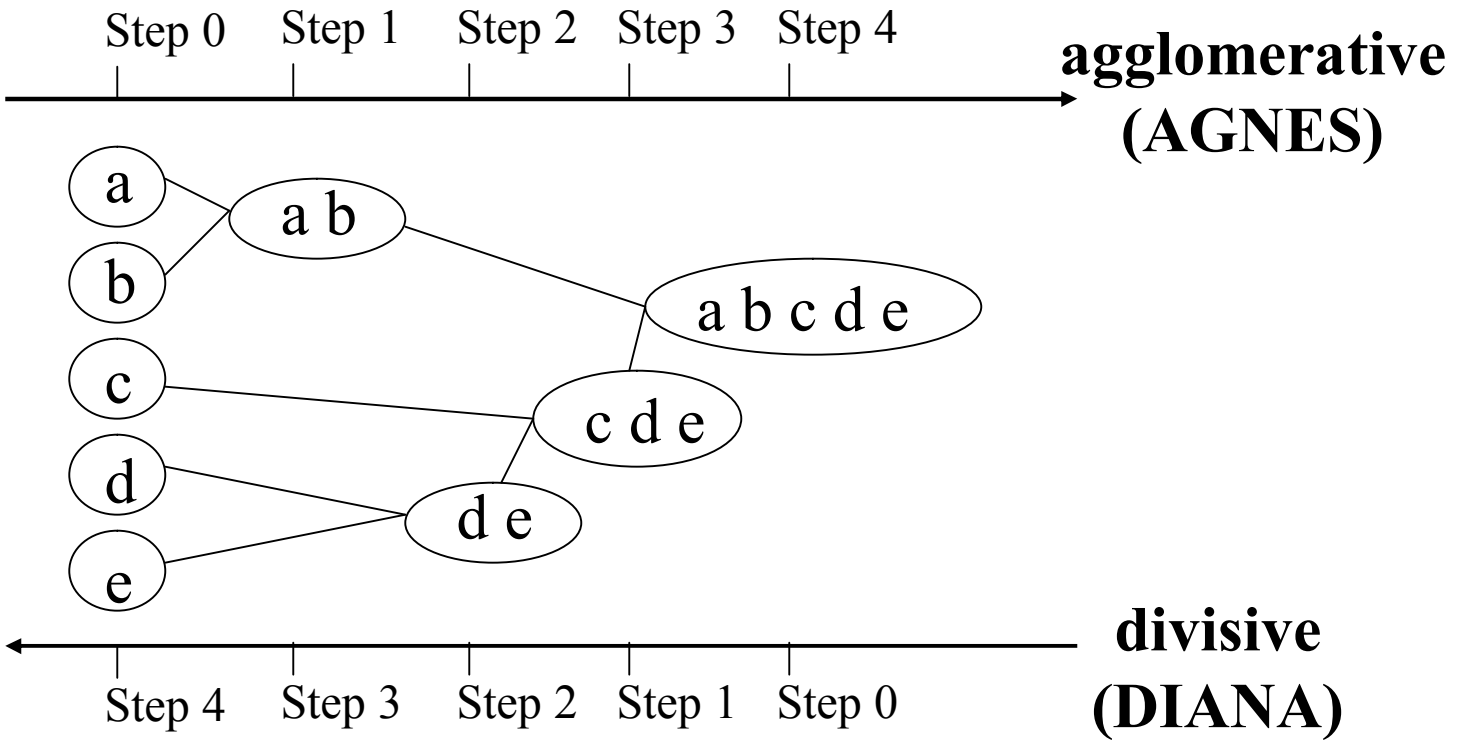
Nội dung

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

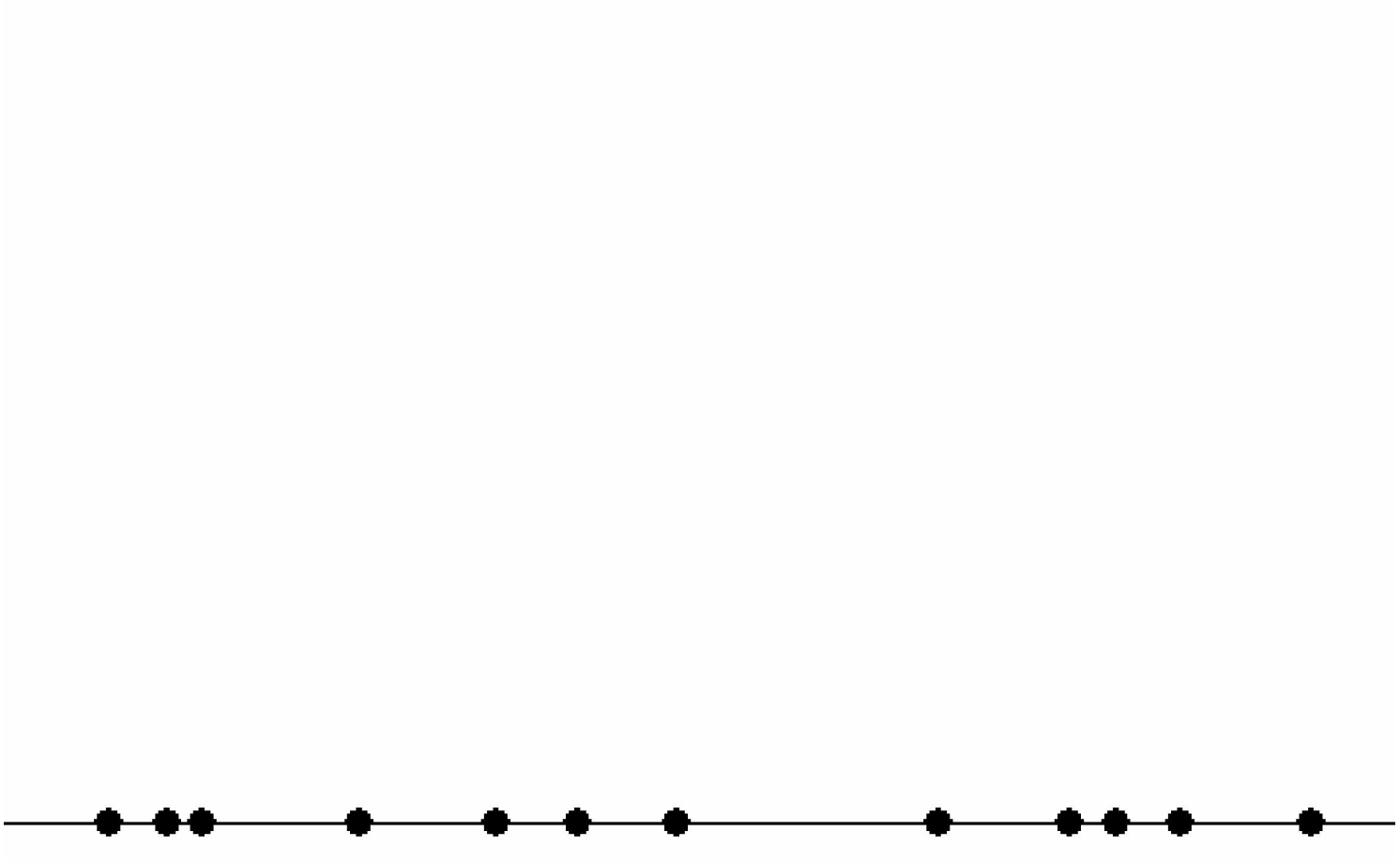
Hierarchical clustering

- 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

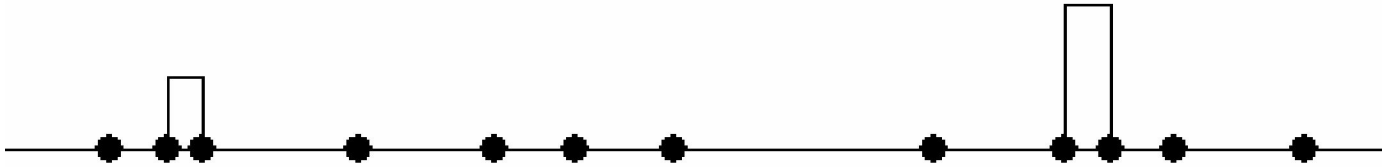


Hierarchical clustering (Single link)

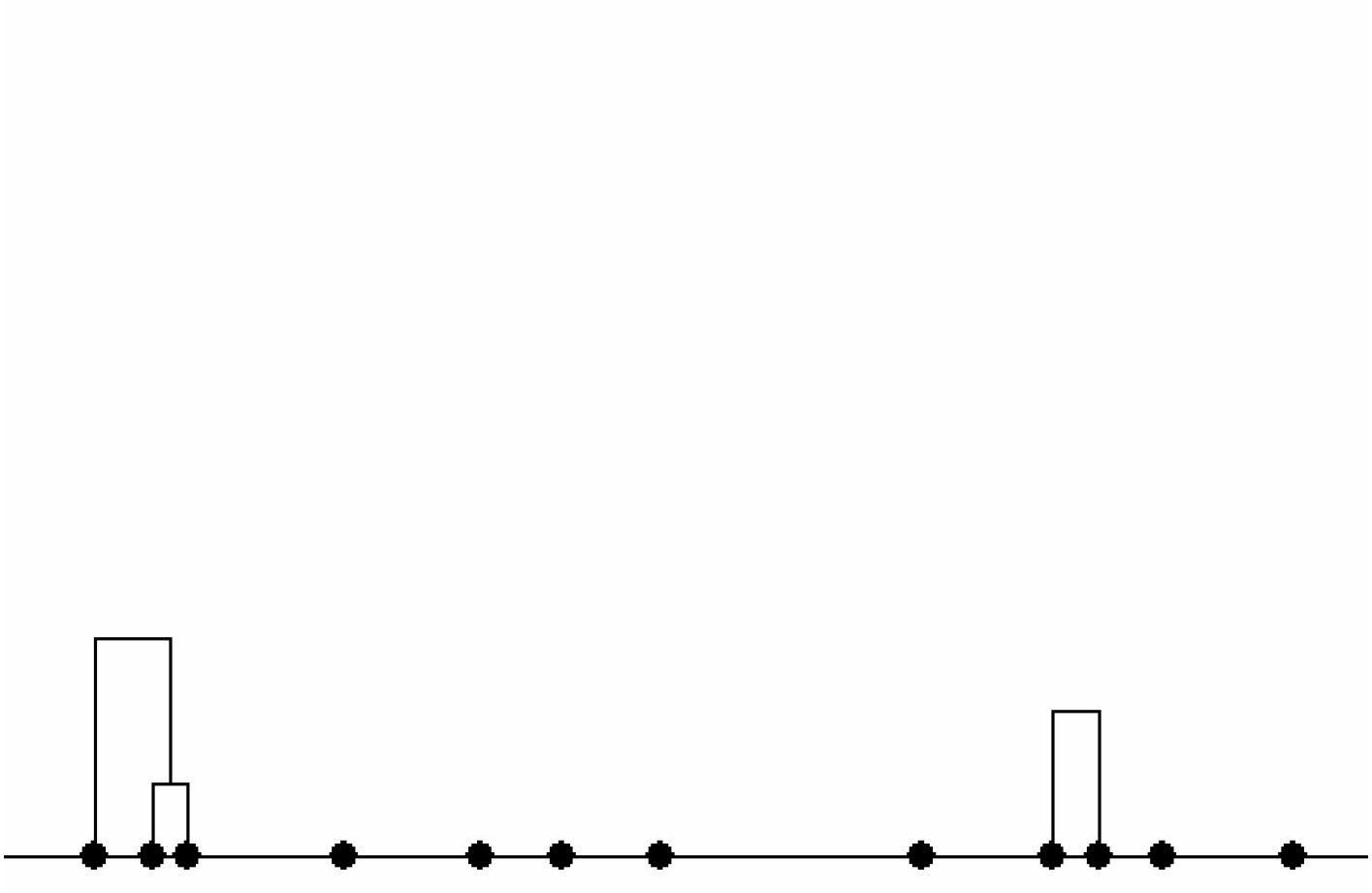


Hierarchical clustering (Single link)

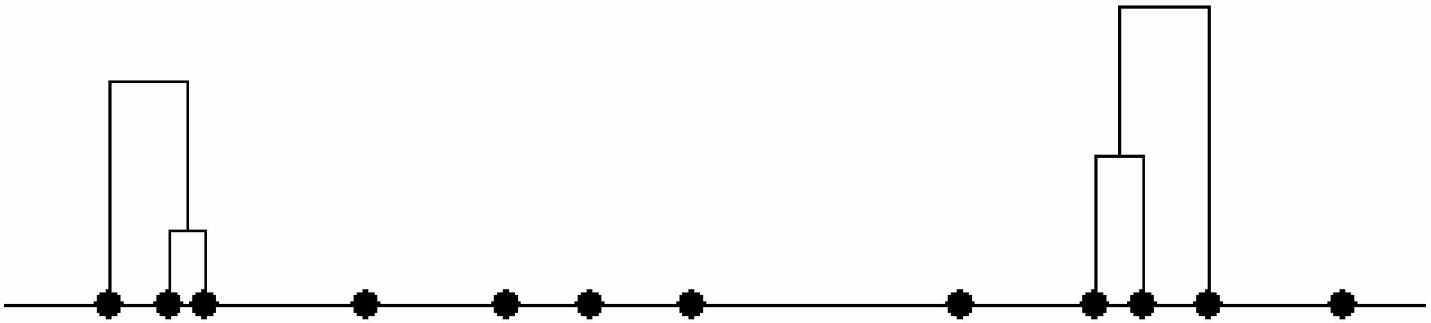
Hierarchical clustering (Single link)



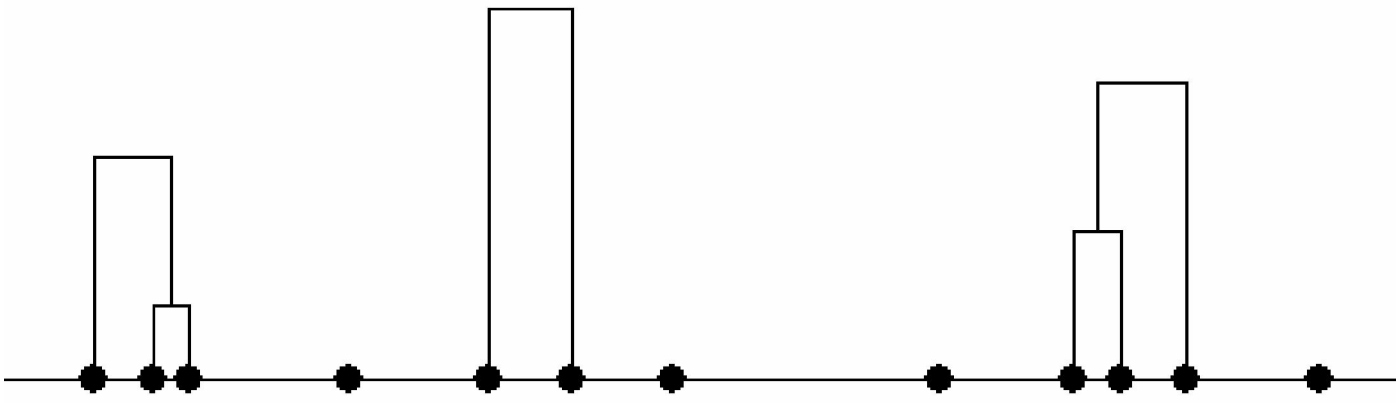
Hierarchical clustering (Single link)



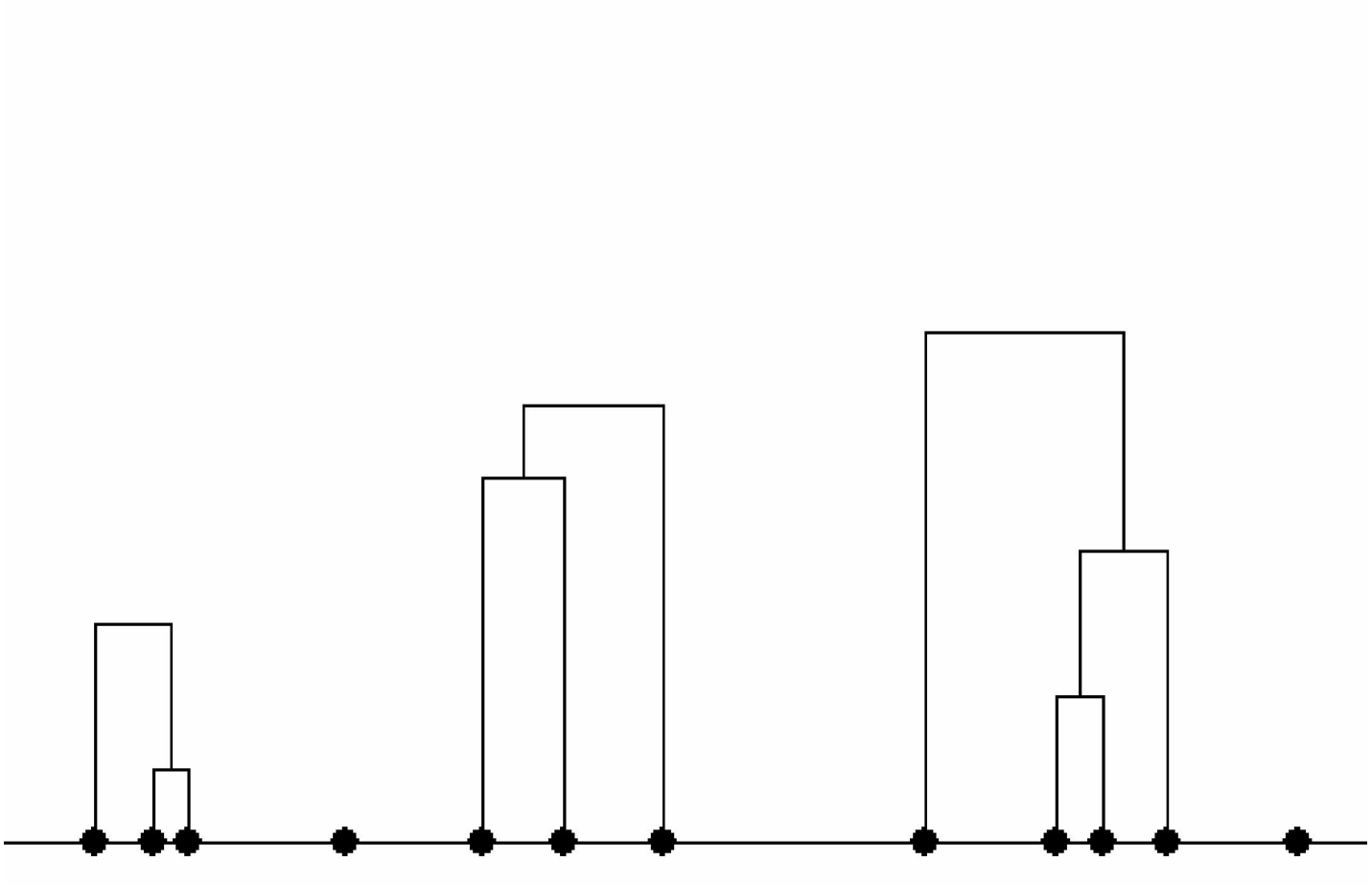
Hierarchical clustering (Single link)



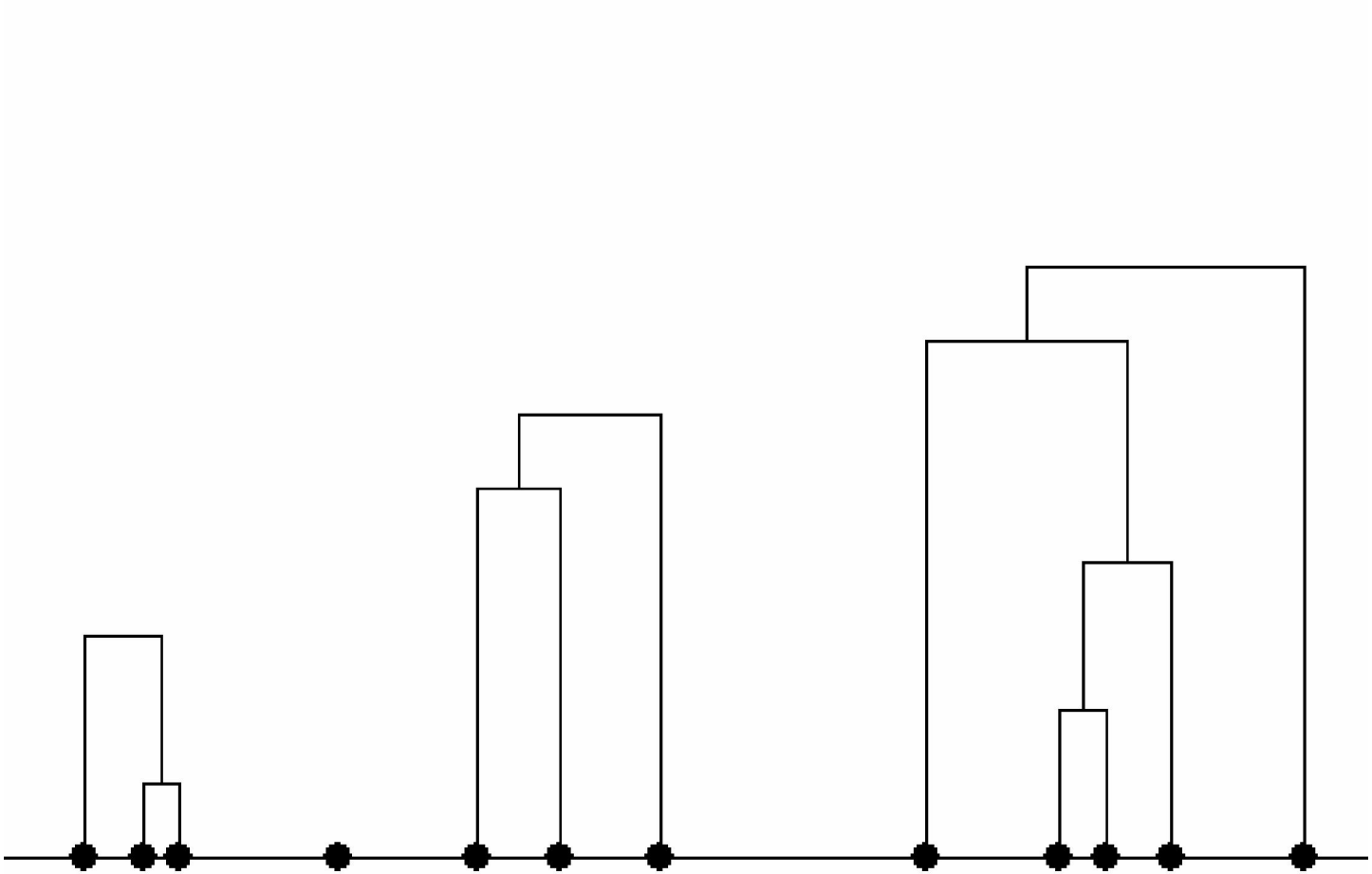
Hierarchical clustering (Single link)



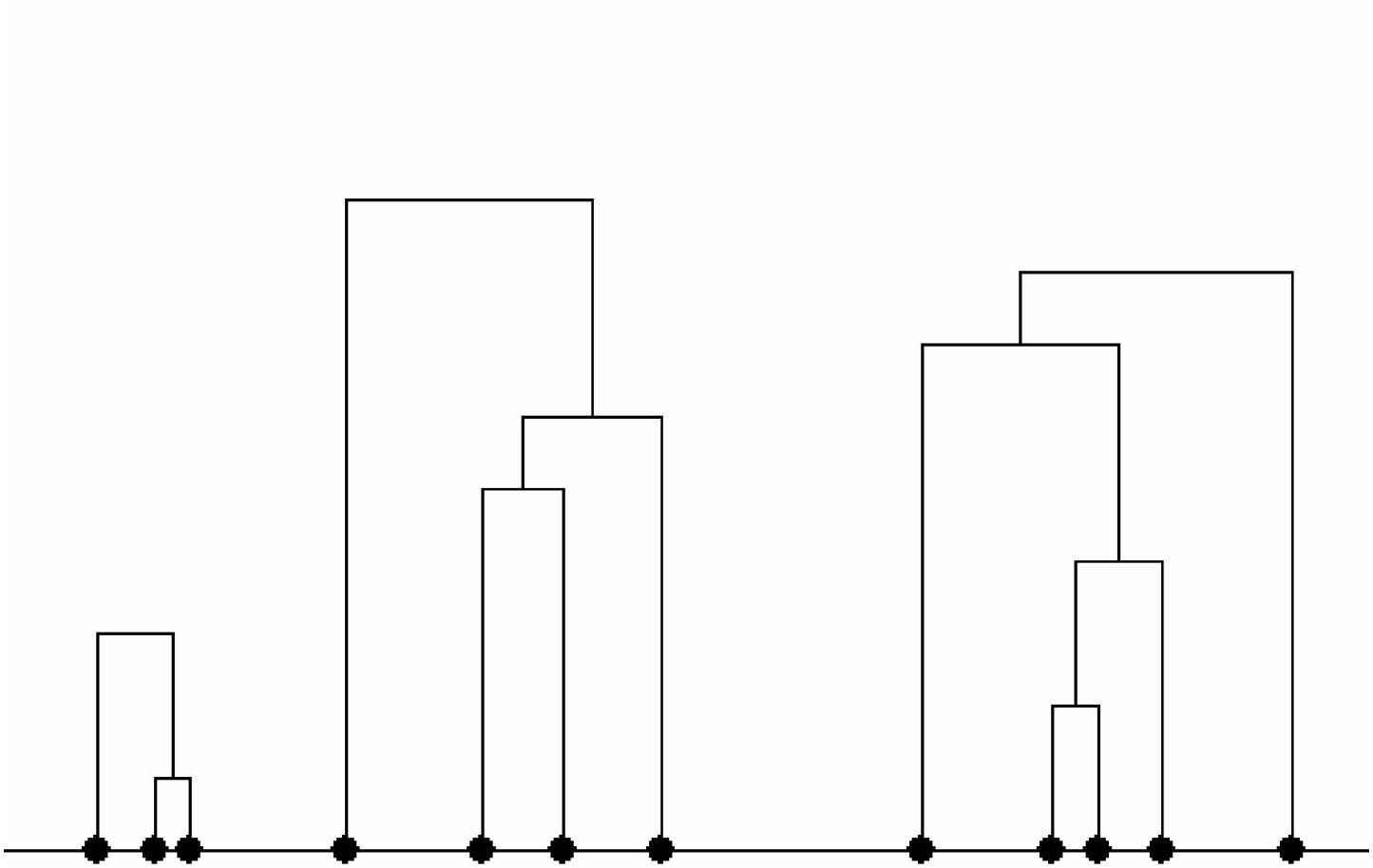
Hierarchical clustering (Single link)



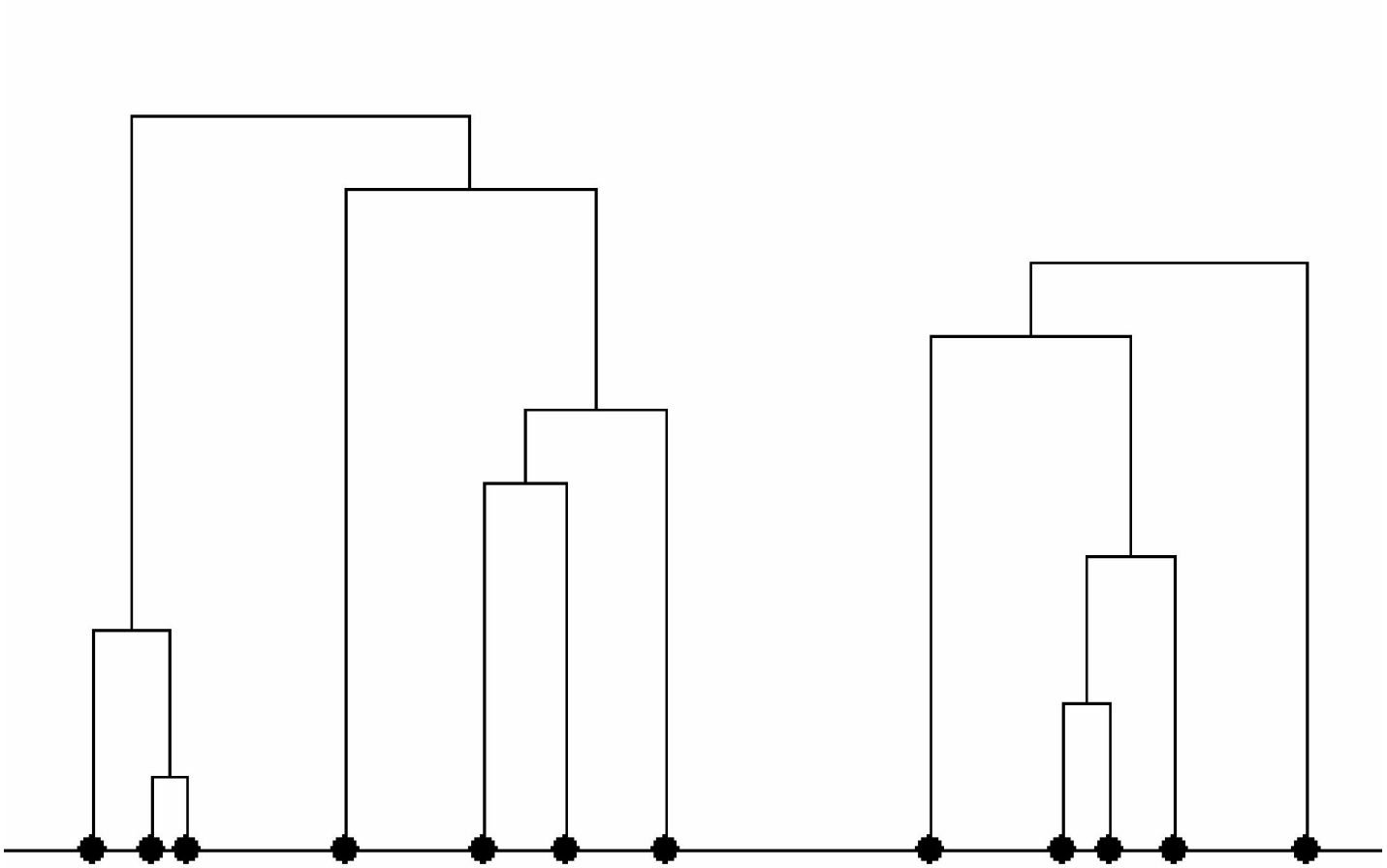
Hierarchical clustering (Single link)



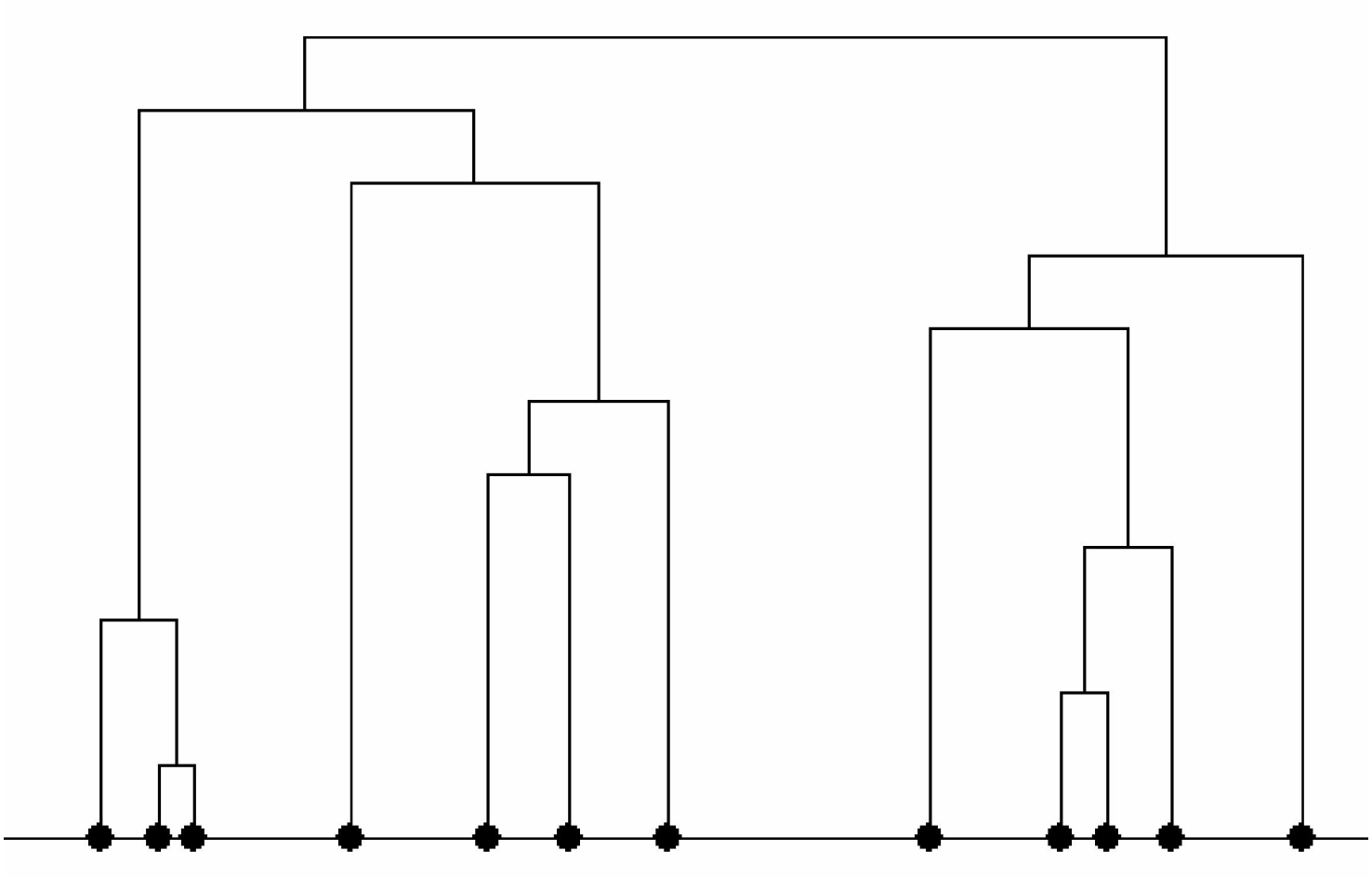
Hierarchical clustering (Single link)



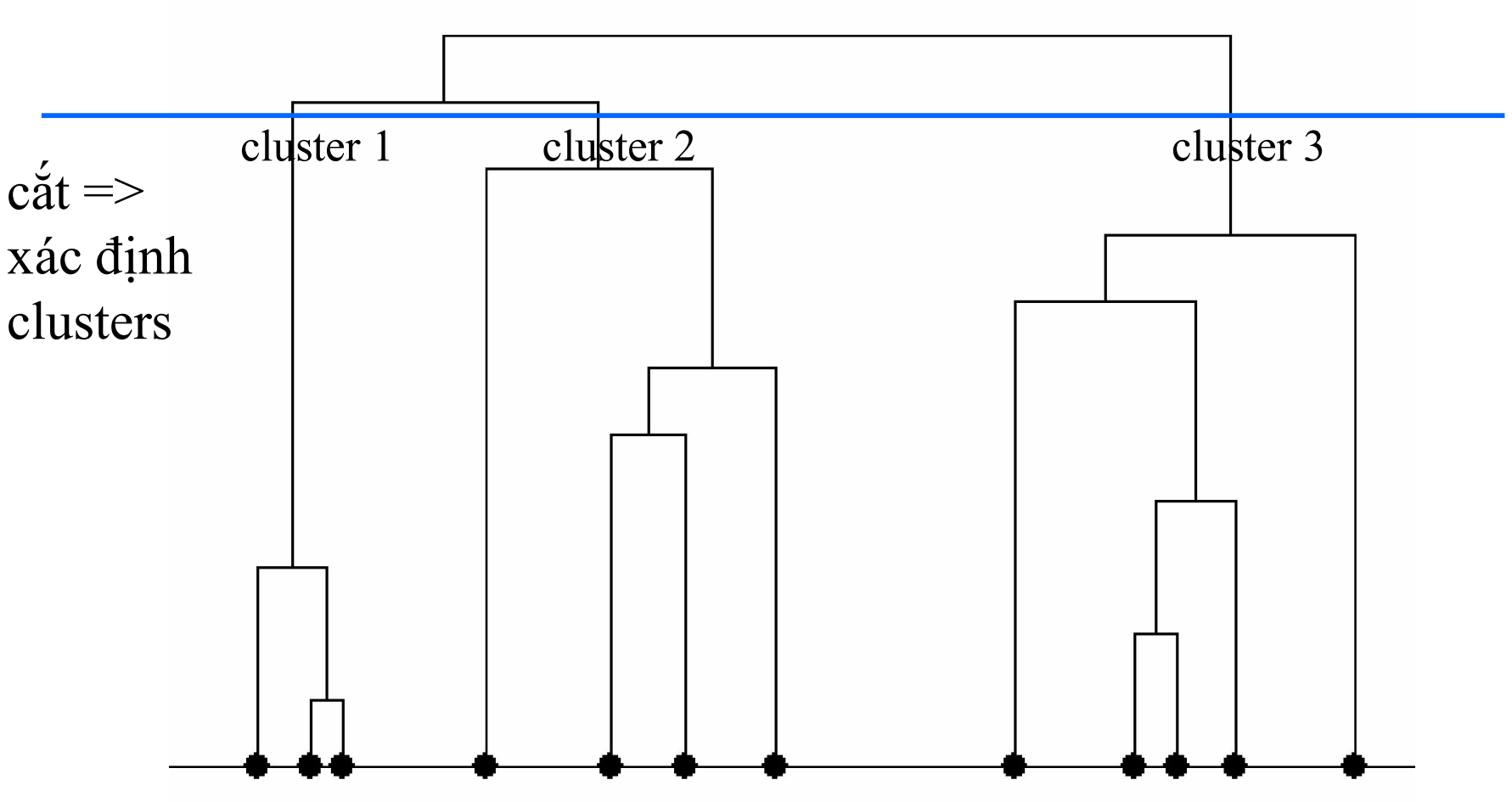
Hierarchical clustering (Single link)



Hierarchical clustering (Single link)



Hierarchical clustering (Single link)



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

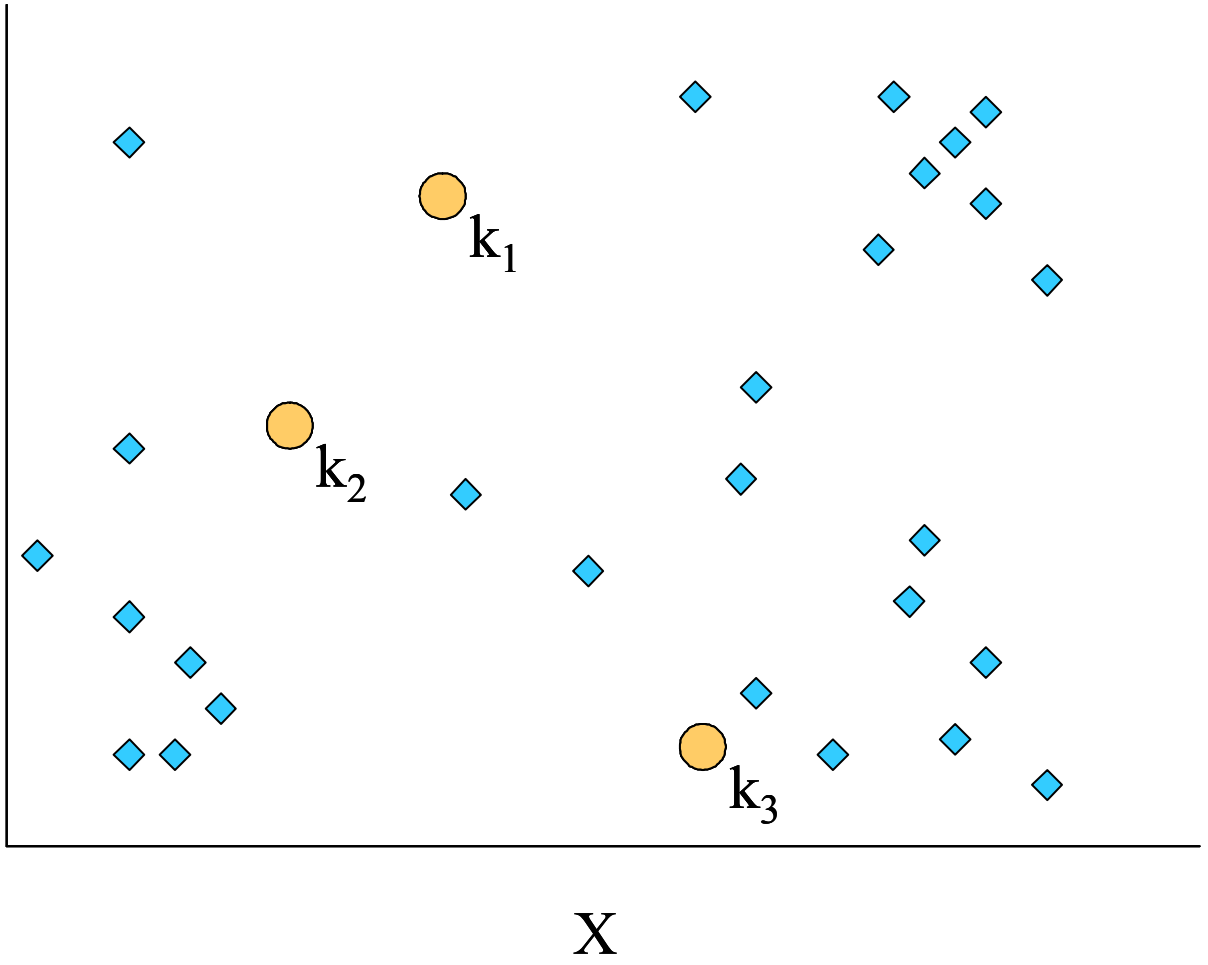
Giải thuật K-Means

- 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

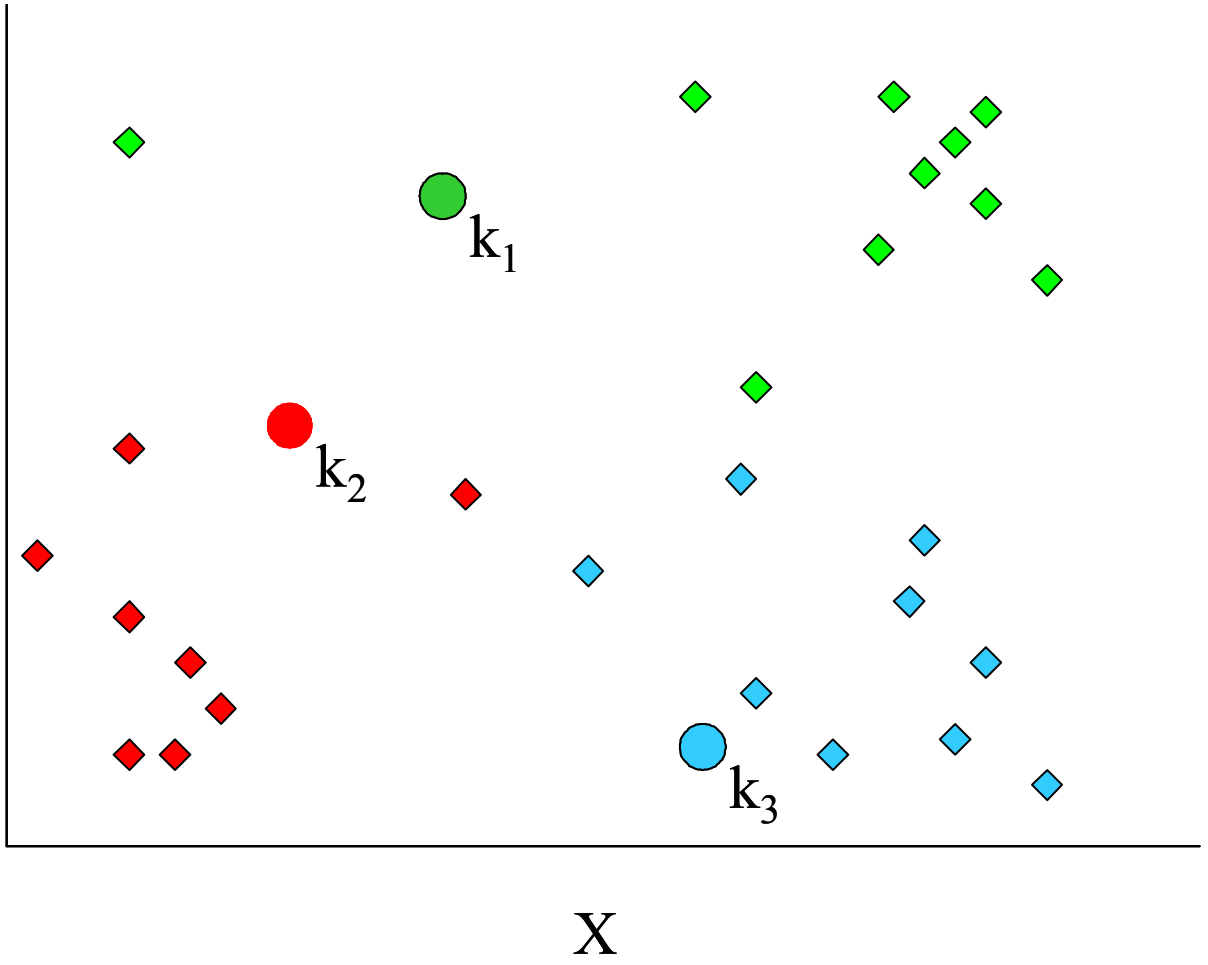
Y

khởi động ngẫu nhiên 3 tâm của 3 clusters



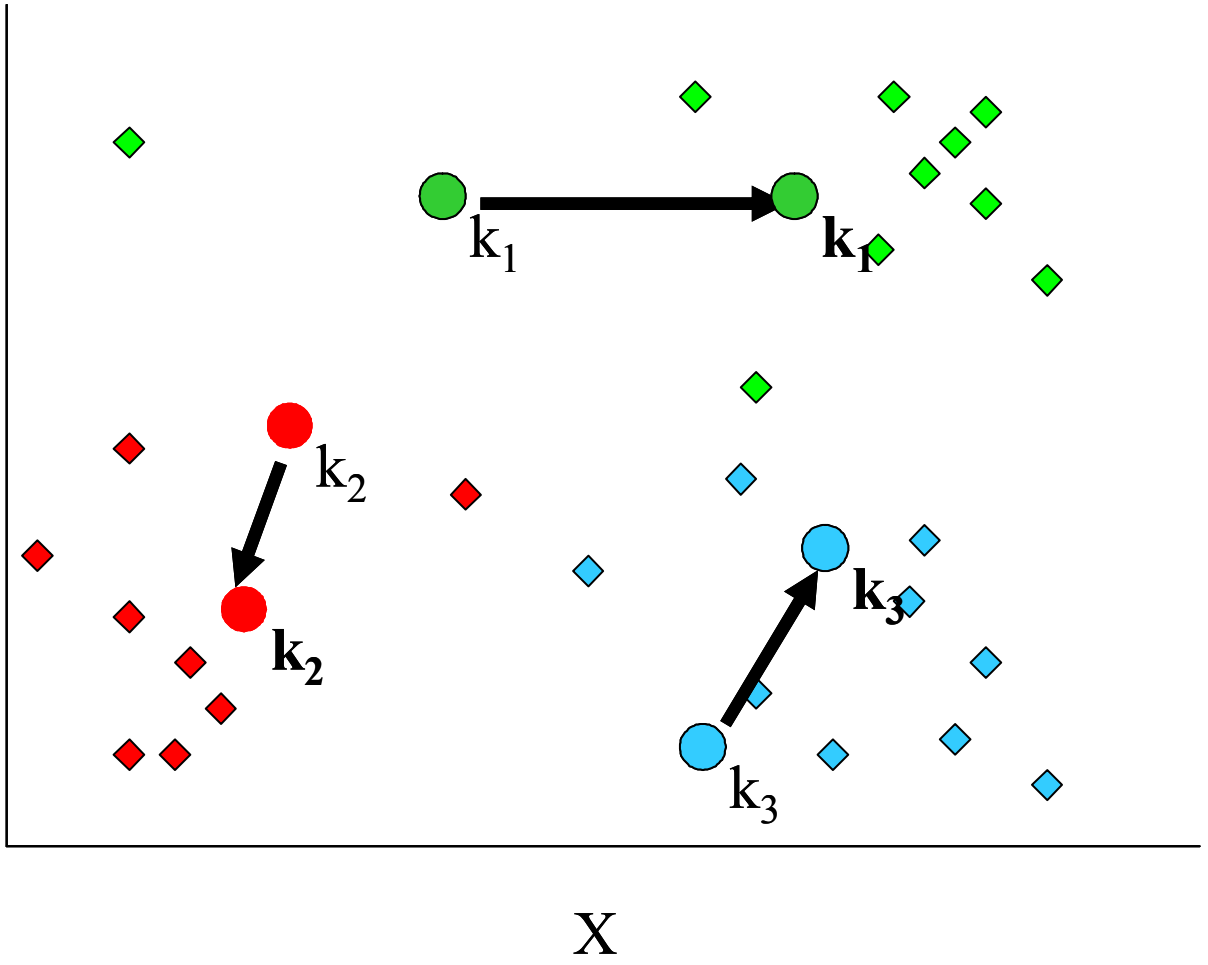
Giải thuật K-Means

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



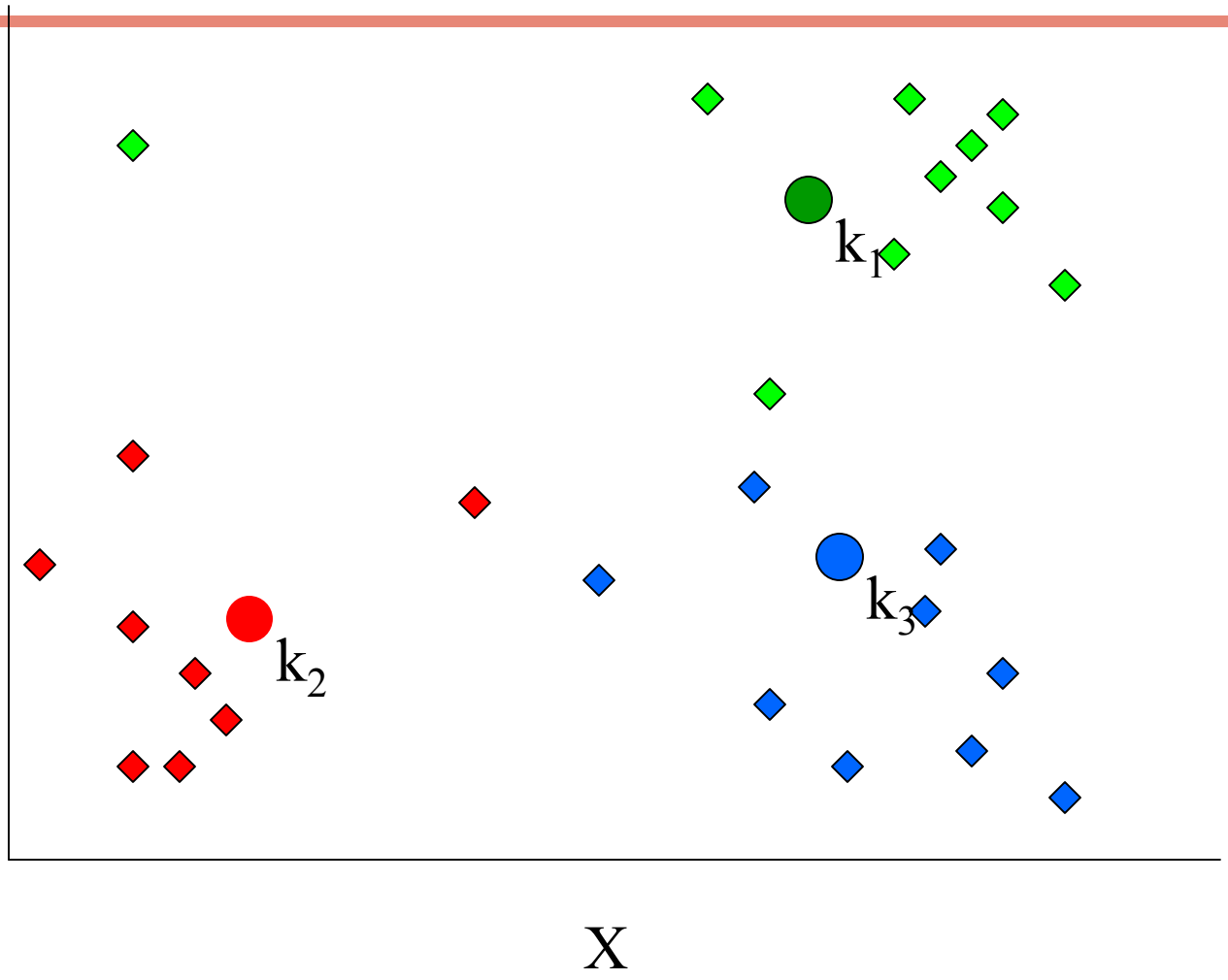
Giải thuật K-Means

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



Giải thuật K-Means

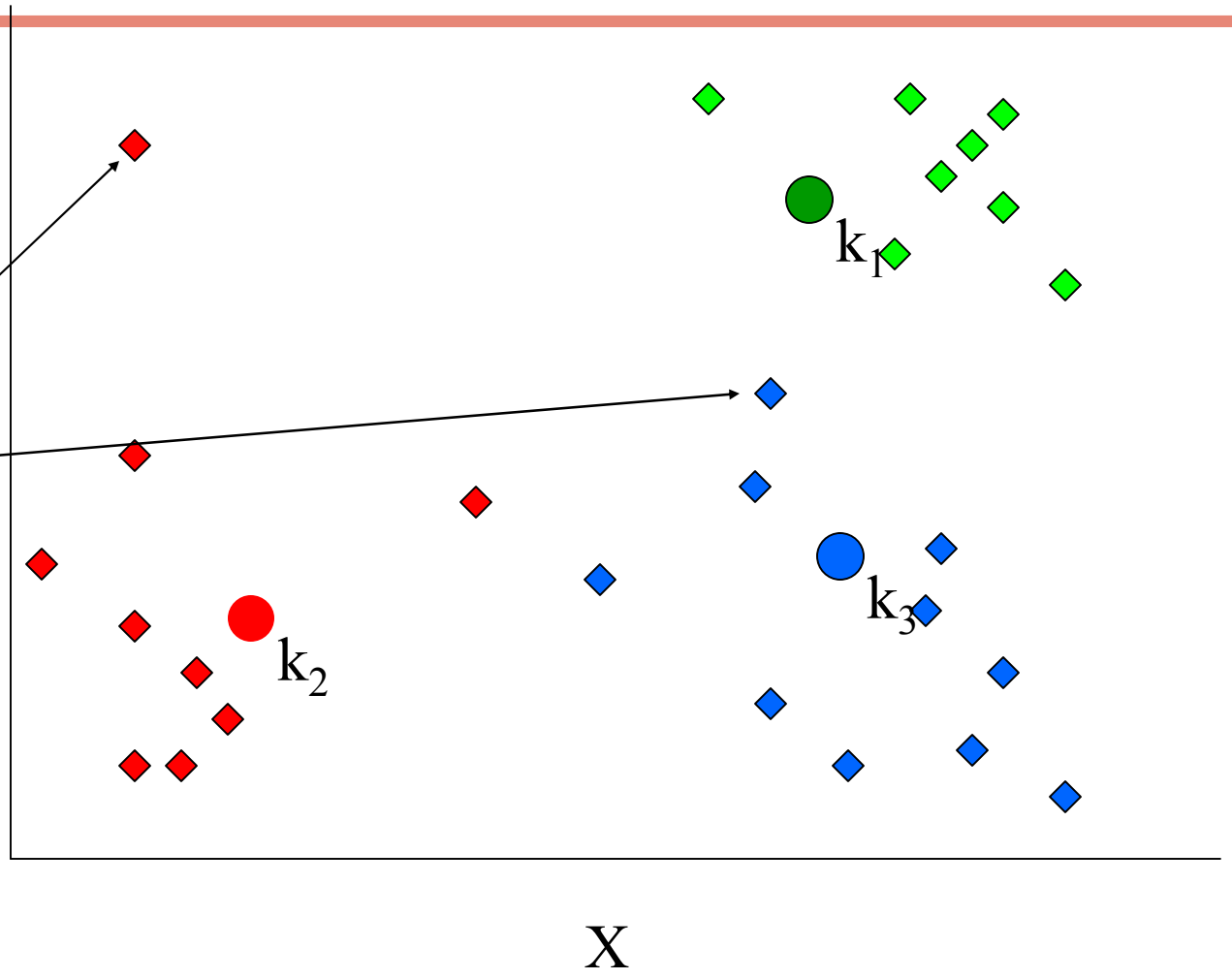
cấu hình mới của lần lặp tiếp theo



Giải thuật K-Means

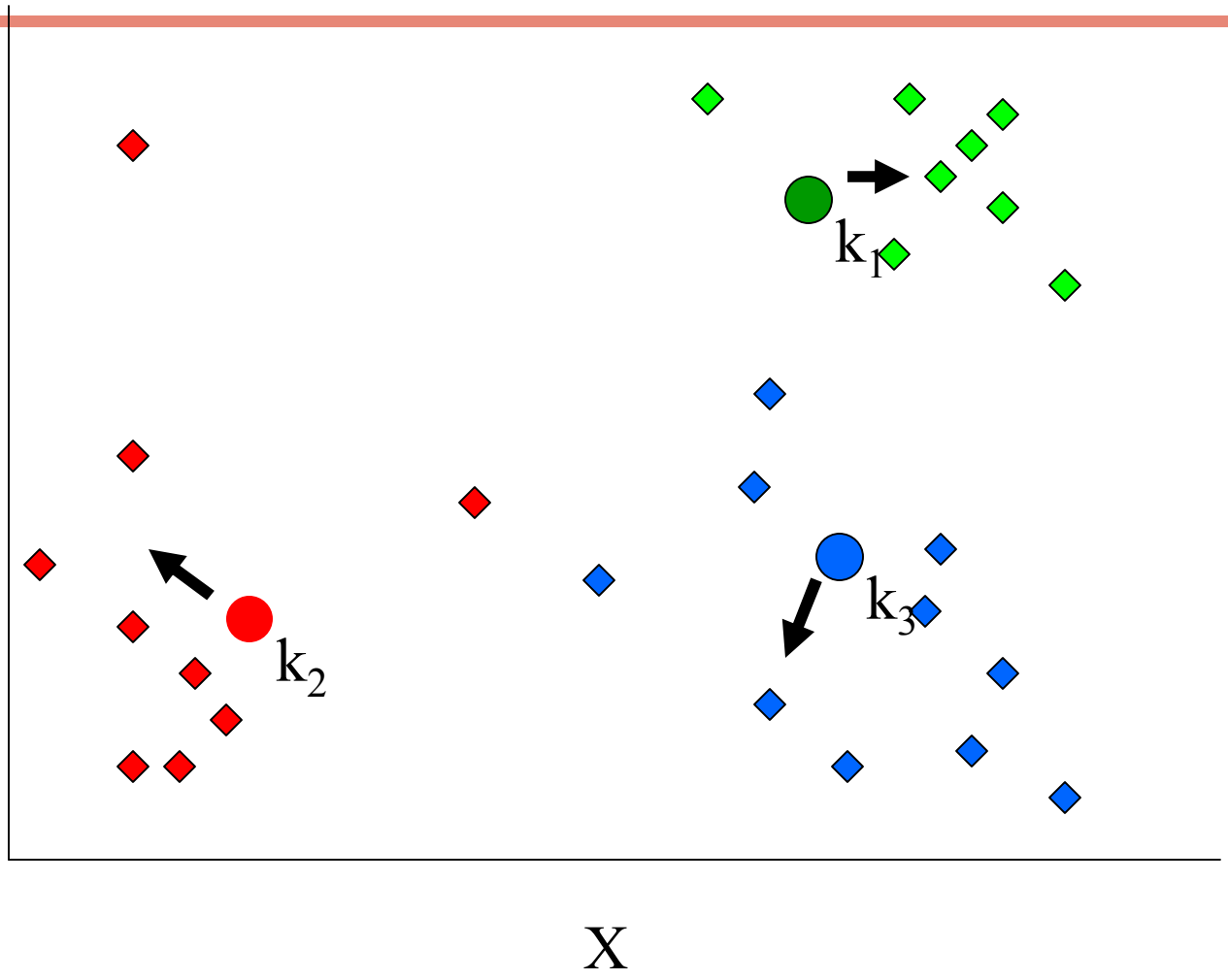
mỗi phần tử được gán lại cho tâm cluster gần nhất của nó

có 2 phần tử thay đổi nhóm



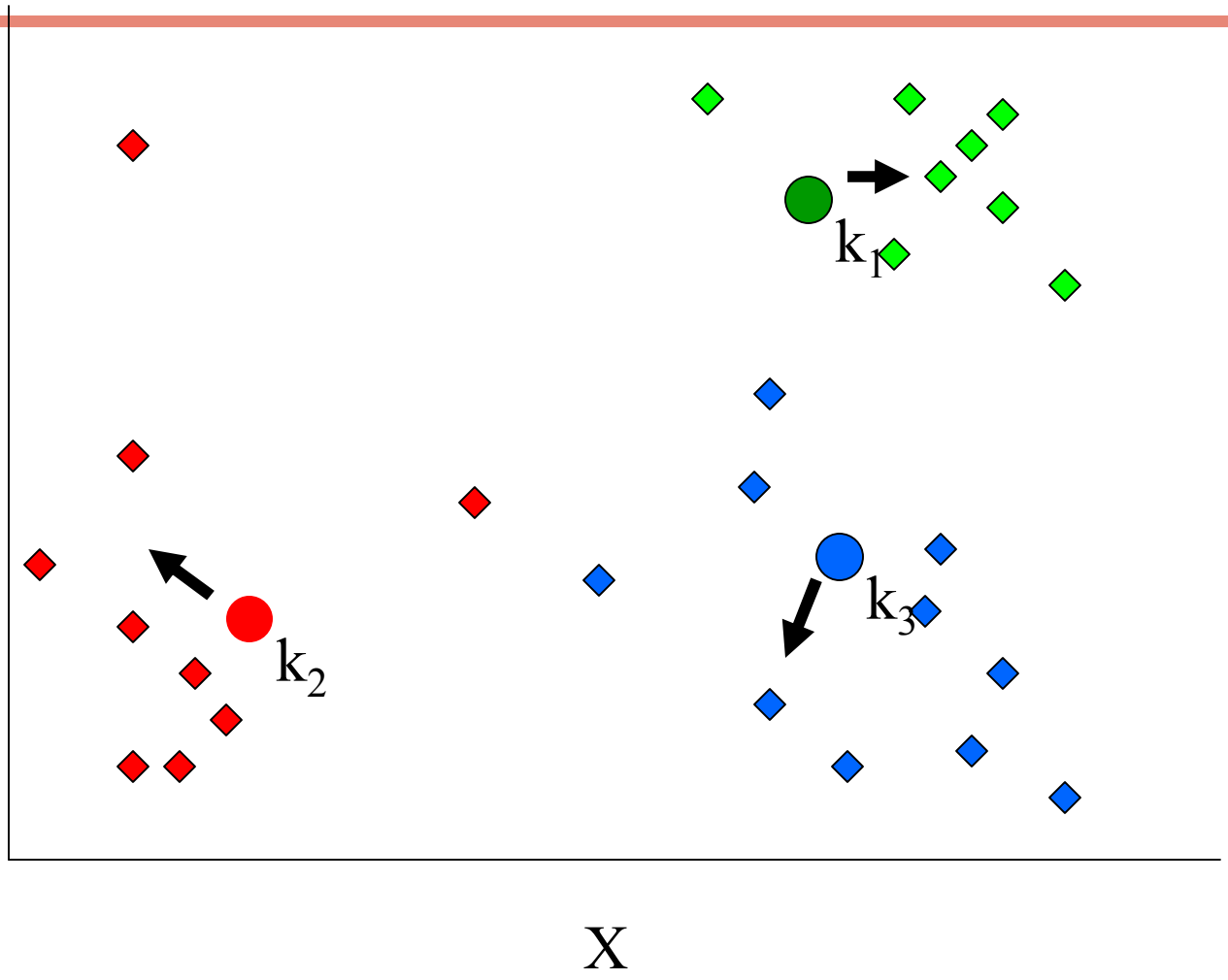
Giải thuật K-Means

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

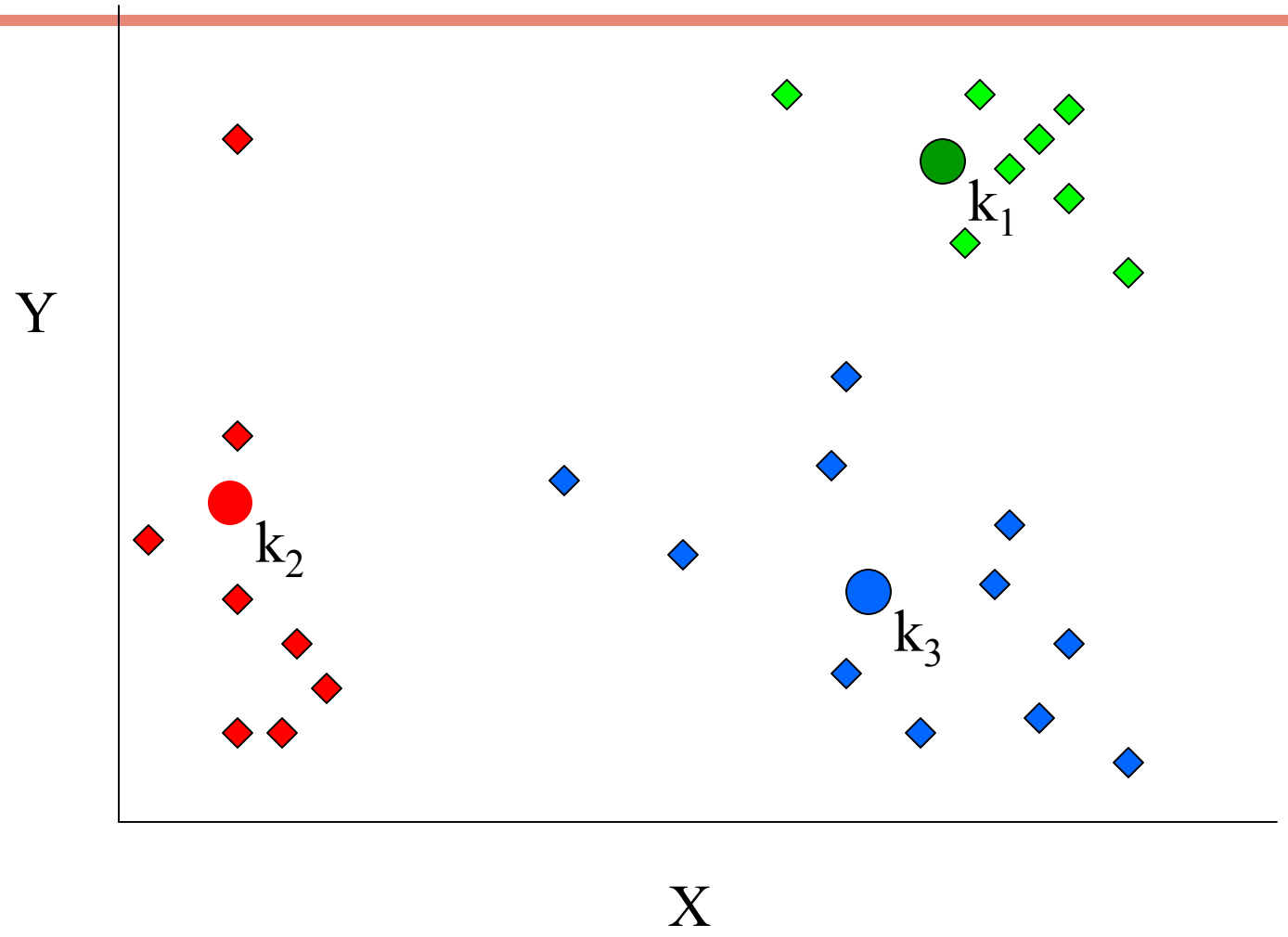


Giải thuật K-Means

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



Giải thuật K-Means



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

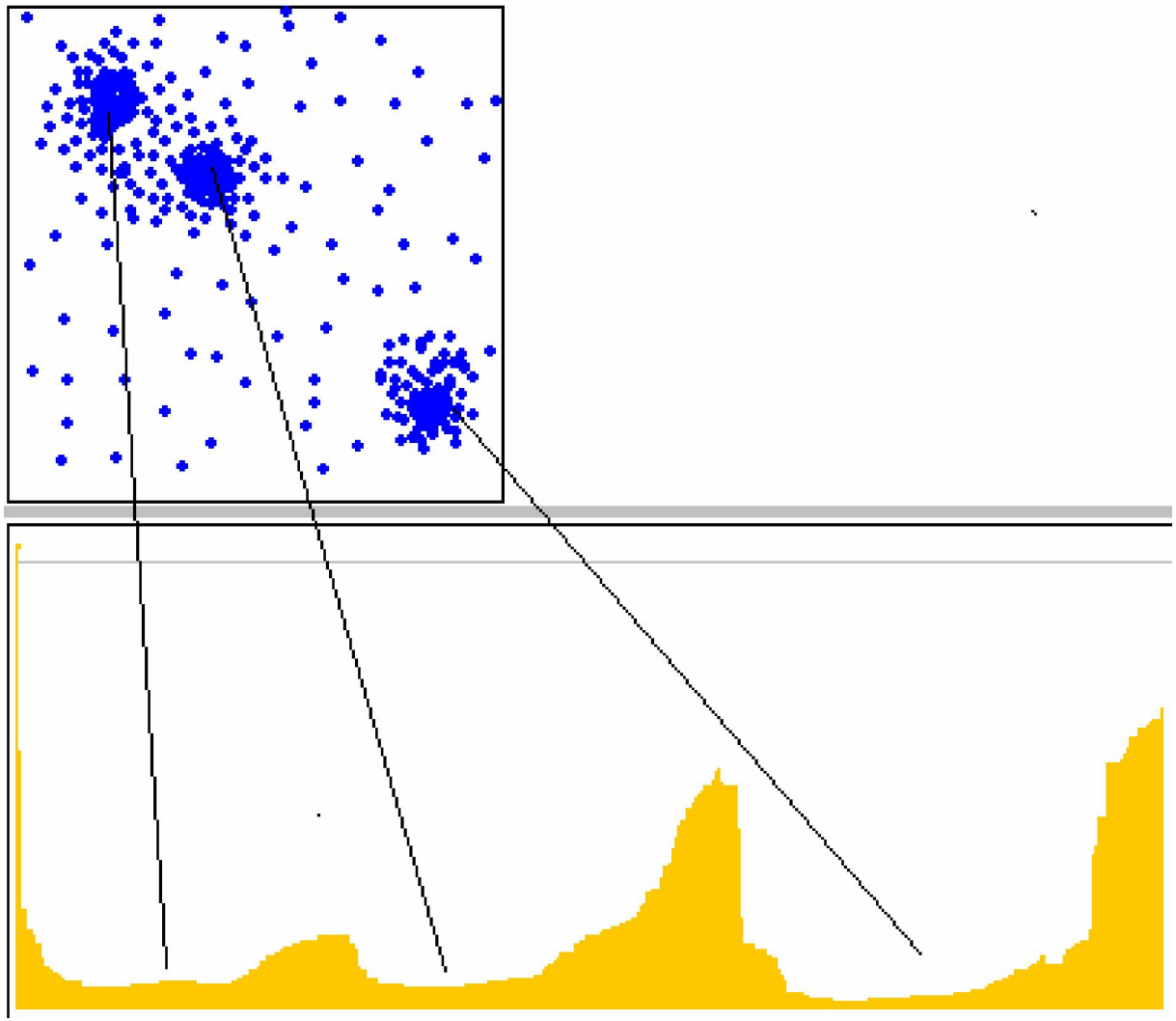
Nội dung

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

Giải thuật clustering

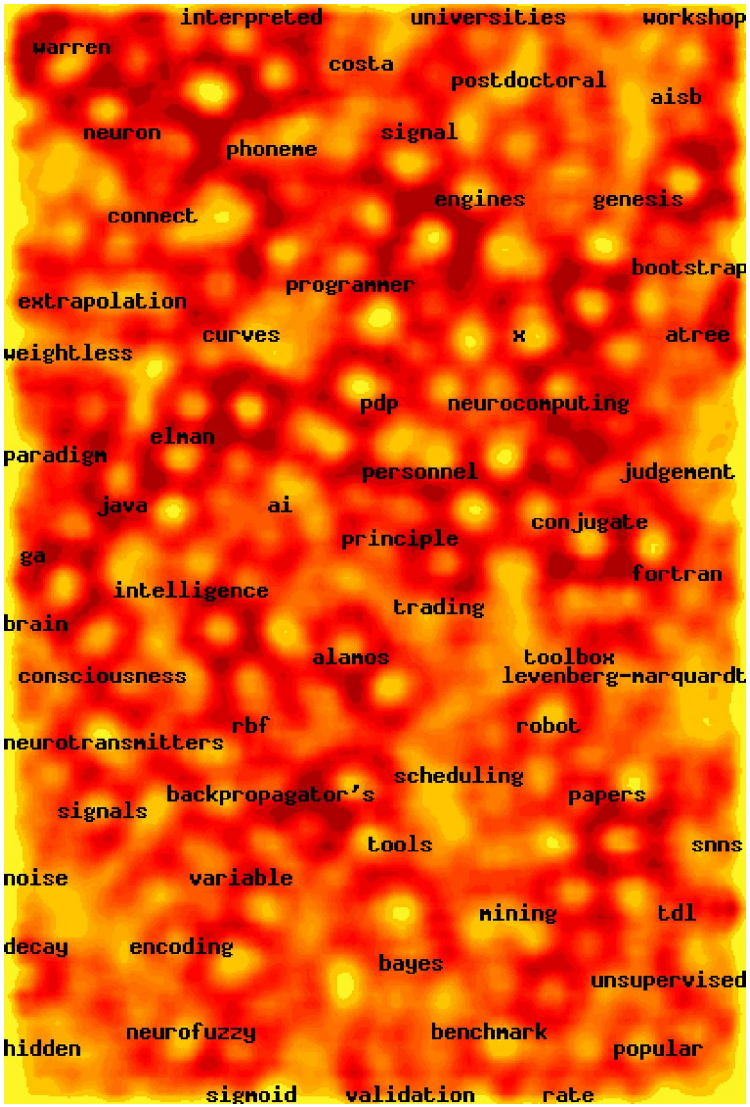
- còn nhiều phương pháp khác
  - density-based : DBSCAN (Ester et al., 1996), OPTICS (Ankerst et al., 1999), DENCLUE (Hinneburg & Keim, 1998)
  - model-based : EM (Expected maximization), SOM (Kohonen, 1995)

Clustering với OPTICS



Clustering 12088 web articles với SOM

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



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 !