

# Clustering

Ali Ridho Barakbah, Entin Martiana

Knowledge Engineering Laboratory

Department of Information and Computer Engineering

Politeknik Elektronika Negeri Surabaya

# What is cluster?

---

a collection of objects which are  
“similar” between them and are  
“dissimilar” to the objects belonging to  
other clusters

[http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial\\_html/index.html](http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/index.html)

# What is clustering?

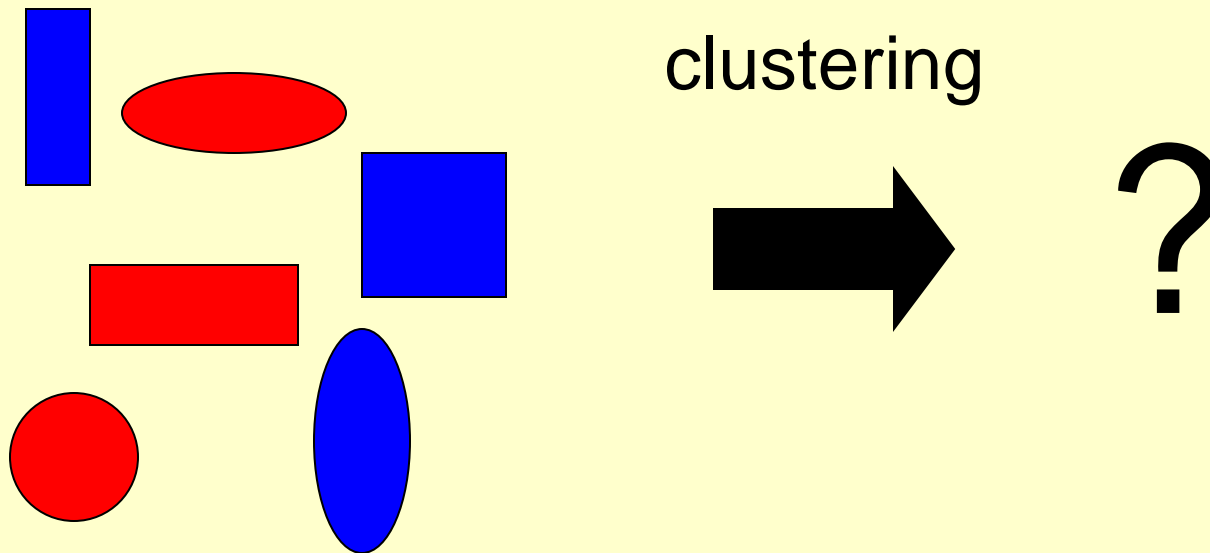
---

the process of organizing objects into groups whose members are similar in some way

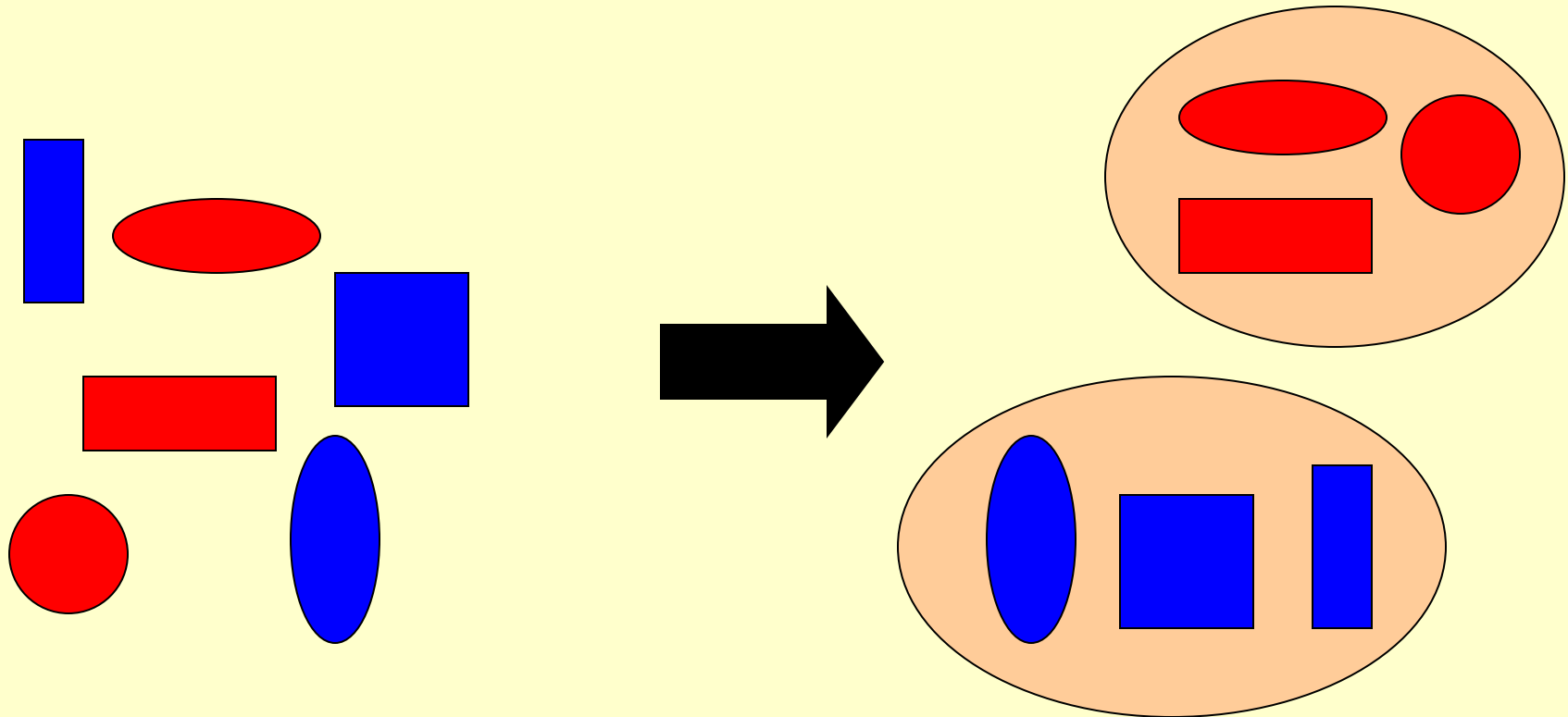
[http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial\\_html/index.html](http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/index.html)

# Ilustrasi clustering

---

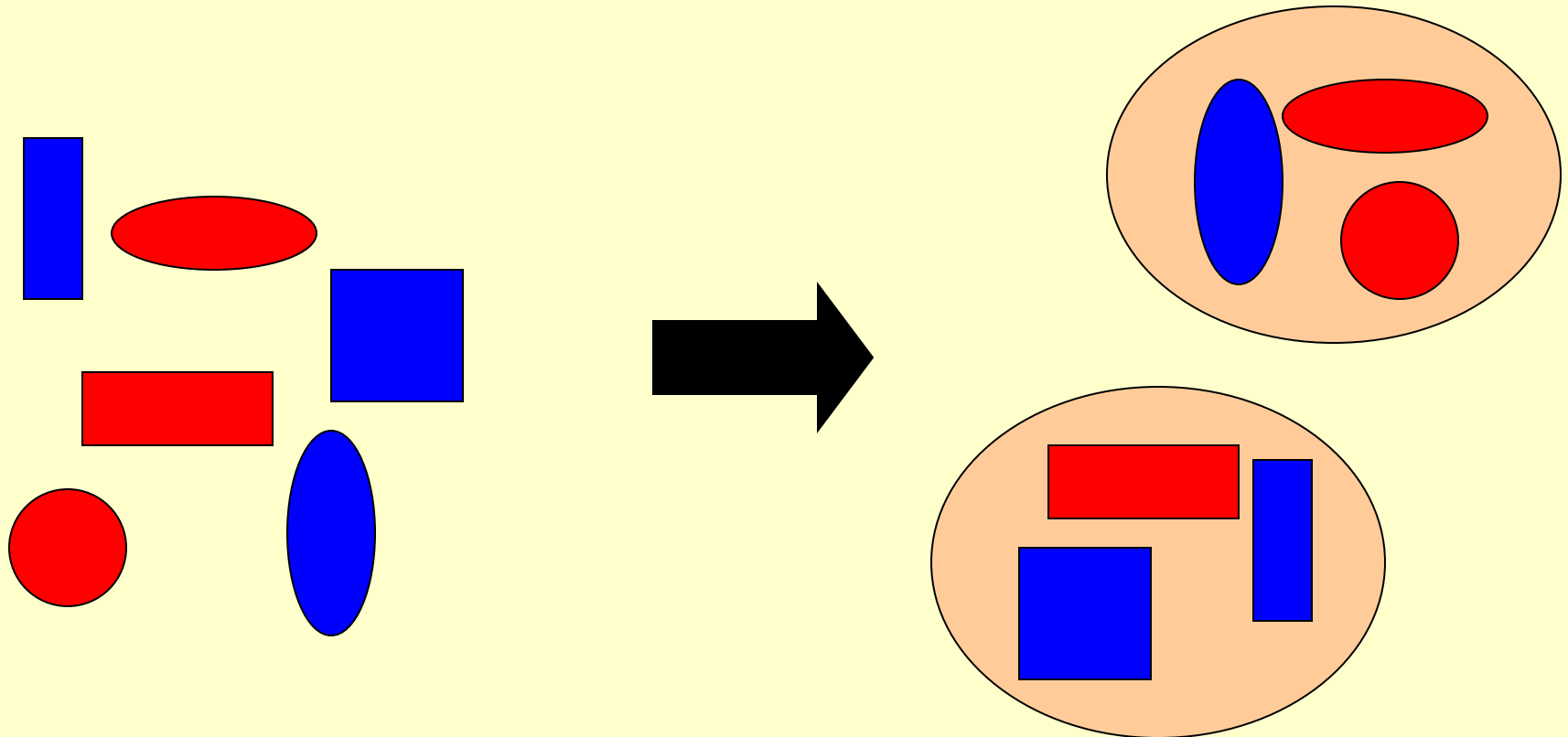


# Ilustrasi clustering



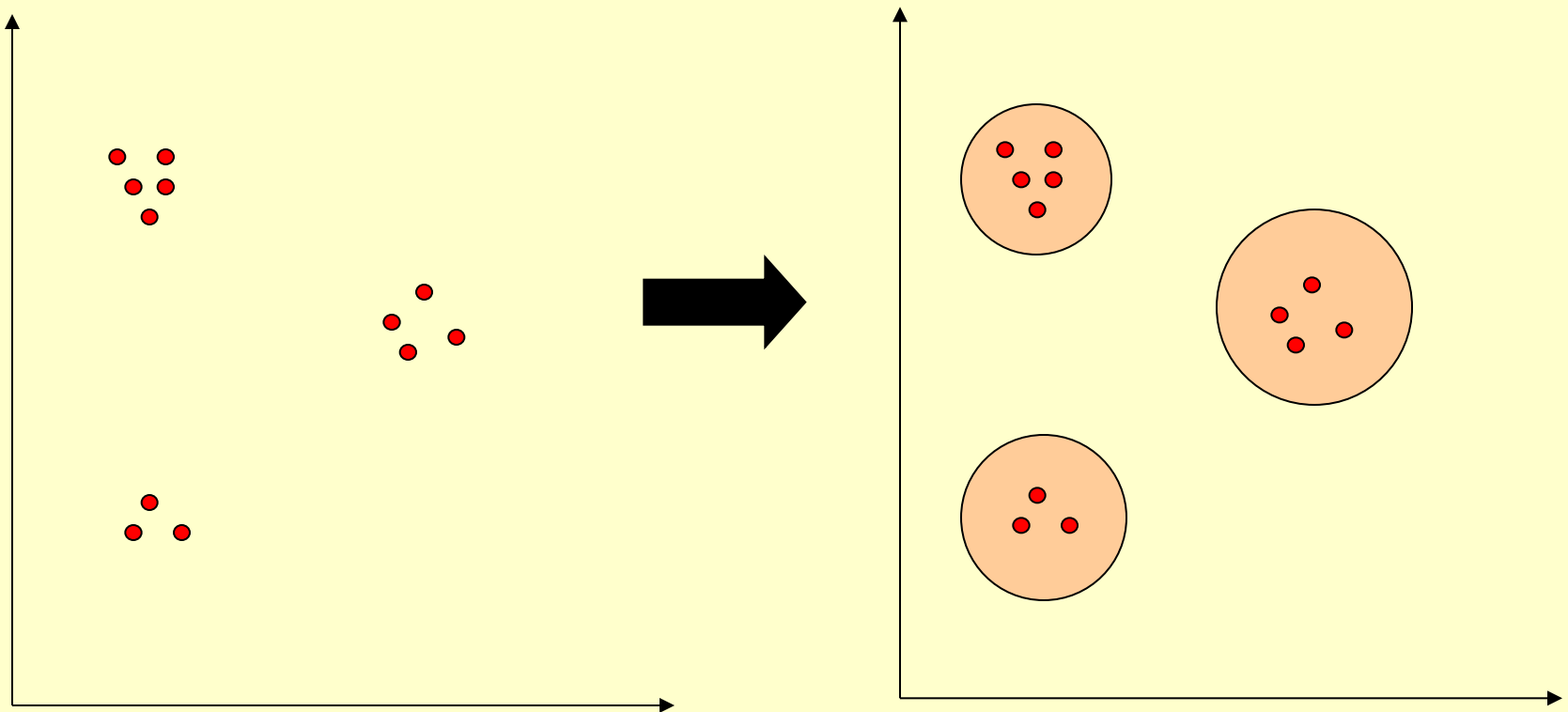
Similaritas berdasarkan warna

# Ilustrasi clustering



Similaritas berdasarkan bentuk

# Ilustrasi clustering



Similaritas berdasarkan jarak

# Clustering vs Classification

|               | Classification | Clustering   |
|---------------|----------------|--------------|
| Data          | supervised     | unsupervised |
| Label         | Ya             | Tidak        |
| Analisa hasil | Error ratio    | Variance     |

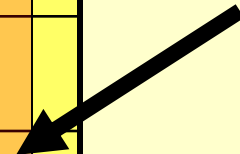


# Classification (kasus sederhana)

## Data penyakit hipertensi

| Umur       | Kegemukan     | Hipertensi |
|------------|---------------|------------|
| muda       | gemuk         | Tidak      |
| muda       | sangat gemuk  | Tidak      |
| paruh baya | gemuk         | Tidak      |
| paruh baya | terlalu gemuk | Ya         |
| tua        | terlalu gemuk | Ya         |

label



Supervised data

# Clustering

## (kasus sederhana)

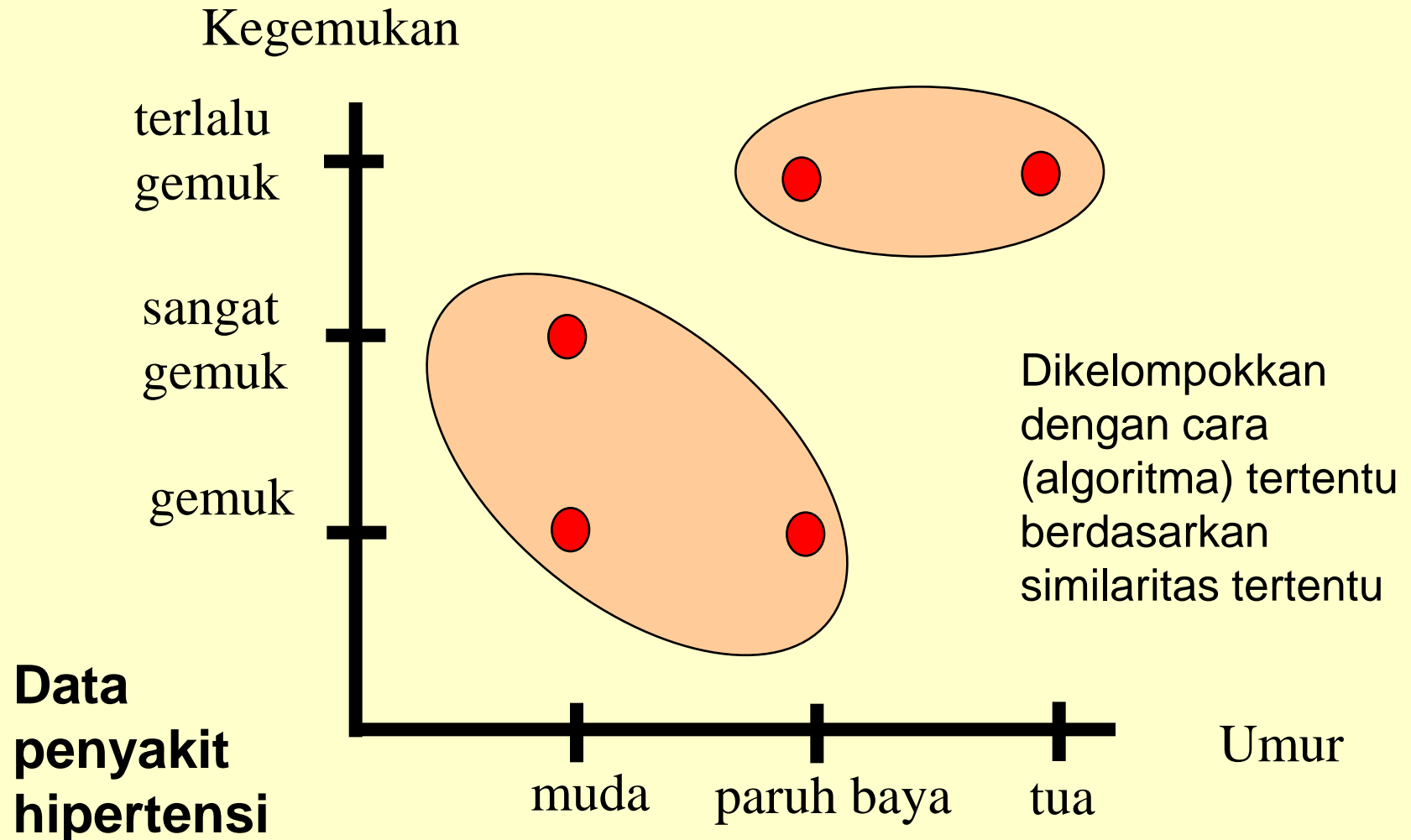
### Data penyakit hipertensi

| Umur       | Kegemukan     |
|------------|---------------|
| muda       | gemuk         |
| muda       | sangat gemuk  |
| paruh baya | gemuk         |
| paruh baya | terlalu gemuk |
| tua        | terlalu gemuk |

tidak ada  
label

Unsupervised data

# Clustering (kasus sederhana)



# Karakteristik clustering

---

- Partitioning clustering
- Hierarchical clustering
- Overlapping clustering
- Hybrid

# Partitioning clustering

---

- Disebut juga exclusive clustering
- Setiap data harus termasuk ke cluster tertentu
- Memungkinkan bagi setiap data yang termasuk cluster tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke cluster yang lain
- Contoh: K-means, residual analysis



# Hierarchical clustering

---

- Setiap data harus termasuk ke cluster tertentu
- Suatu data yang termasuk ke cluster tertentu pada suatu tahapan proses, tidak dapat berpindah ke cluster lain
- Contoh: Single Linkage, Centroid Linkage, Complete Linkage, Average Centroid

# Overlapping clustering

---

- Setiap data memungkinkan termasuk ke beberapa cluster
- Data mempunyai nilai keanggotaan (membership) pada beberapa cluster
- Contoh: Fuzzy C-means, Gaussian Mixture

# Hybrid

---

Mengawinkan karakteristik dari  
partitioning, overlapping dan  
hierarchical



# Algoritma-algoritma clustering

---

- K-means
- Single Linkage
- Centroid Linkage
- Complete Linkage
- Average Linkage
- dll

# K-means

---

- Termasuk partitioning clustering yang memisahkan data ke  $k$  daerah bagian yang terpisah
- K-means algorithm sangat terkenal karena kemudahan dan kemampuannya untuk mengklaster data besar dan data outlier dengan sangat cepat
- Setiap data harus termasuk ke cluster tertentu
- Memungkinkan bagi setiap data yang termasuk cluster tertentu pada suatu tahapan proses, pada tahapan berikutnya berpindah ke cluster yang lain

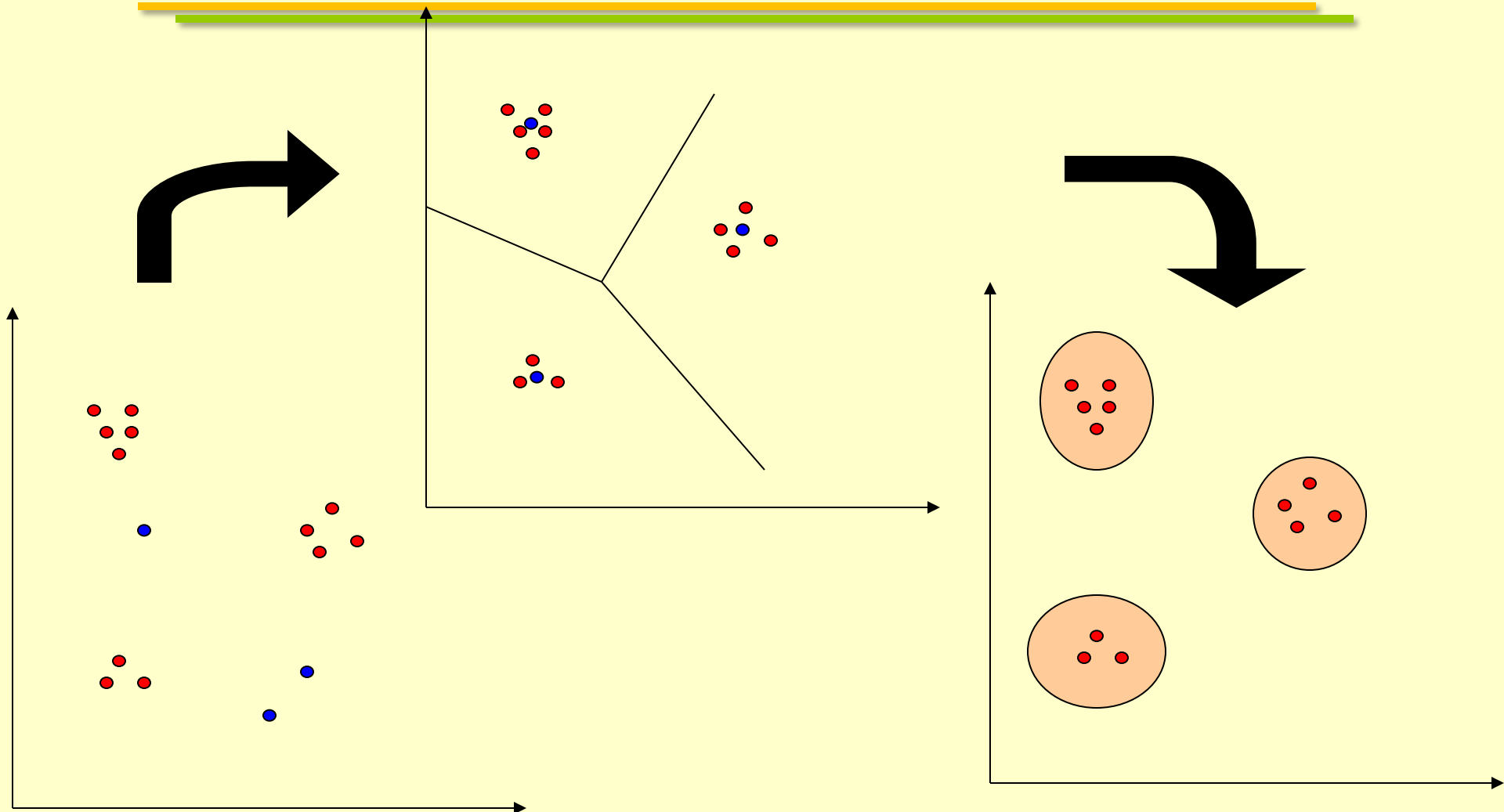
# Algoritma K-means

---

1. Tentukan  $k$  sebagai jumlah cluster yang ingin dibentuk
2. Bangkitkan  $k$  centroids (titik pusat cluster) awal secara random
3. Hitung jarak setiap data ke masing-masing centroids
4. Setiap data memilih centroids yang terdekat
5. Tentukan posisi centroids baru dengan cara menghitung nilai rata-rata dari data-data yang memilih pada centroid yang sama
6. Kembali ke langkah 3 jika posisi centroids baru dengan centroids lama tidak sama.

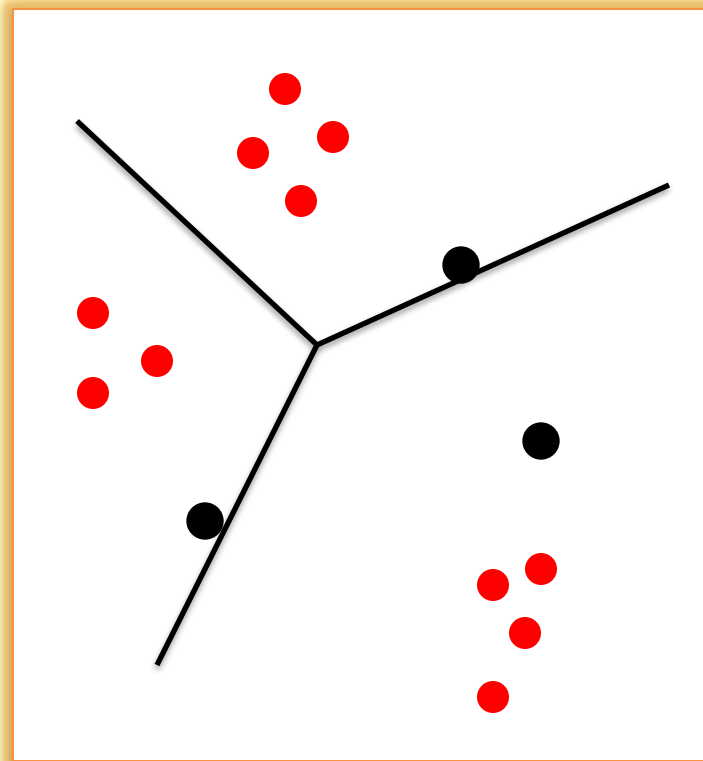


# Algoritma K-means



# K-means Algorithm

---

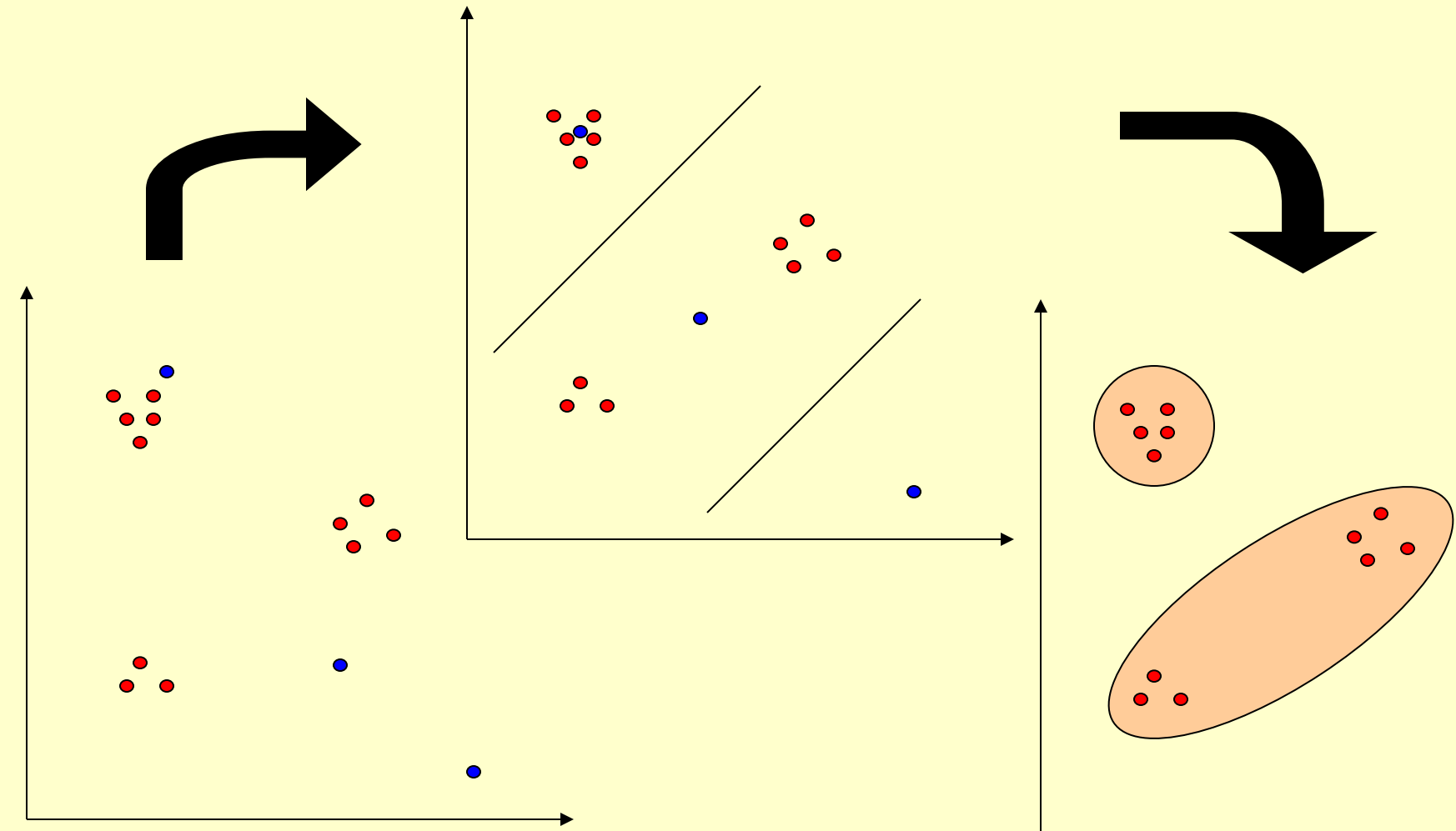


# Karakteristik K-means

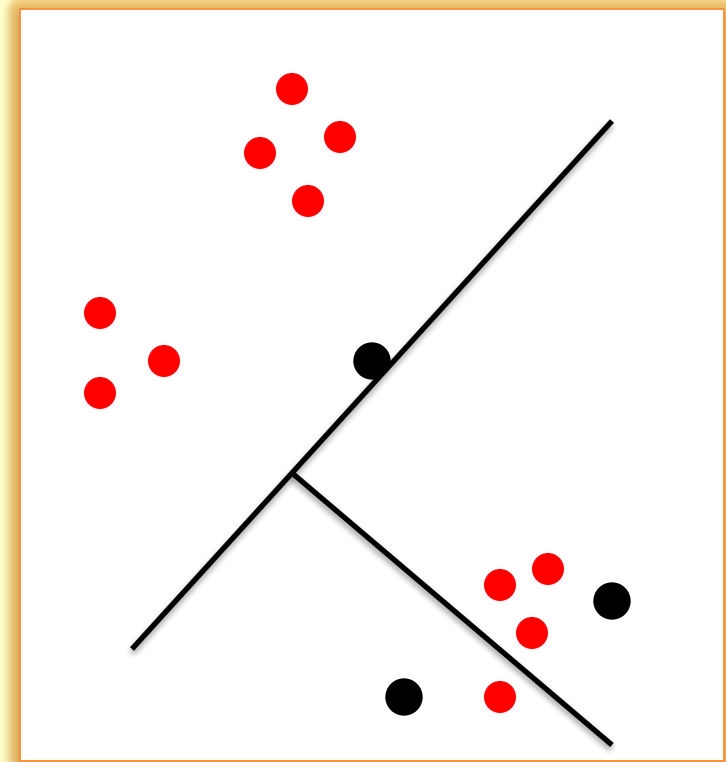
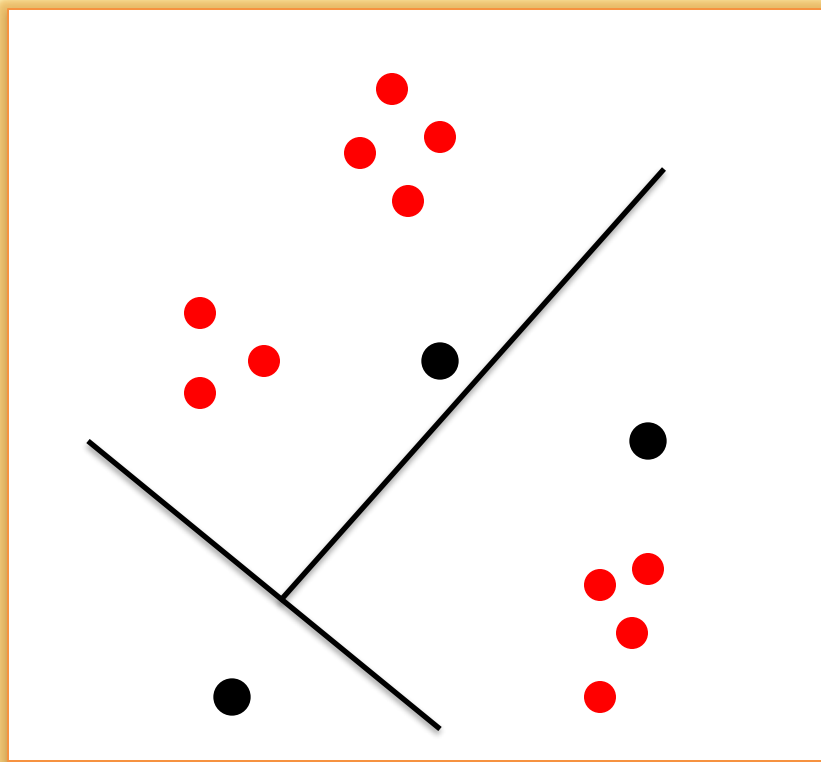
---

- K-means sangat cepat dalam proses clustering
- K-means sangat sensitif pada pembangkitan centroids awal secara random
- Memungkinkan suatu cluster tidak mempunyai anggota
- Hasil clustering dengan K-means bersifat tidak unik (selalu berubah-ubah) - terkadang baik, terkadang jelek.
- K-means sangat sulit untuk mencapai global optimum

# Ilustasi kelemahan K-means



# Kelemahan K-means





# Hierarchical clustering

---

- Single Linkage
- Centroid Linkage
- Complete Linkage
- Average Linkage

# Direction of hierarchy

---

- Divisive
  - 1 cluster to  $k$  clusters
  - Top to down division
- Agglomerative
  - $n$  clusters to  $k$  clusters
  - Down to top merge

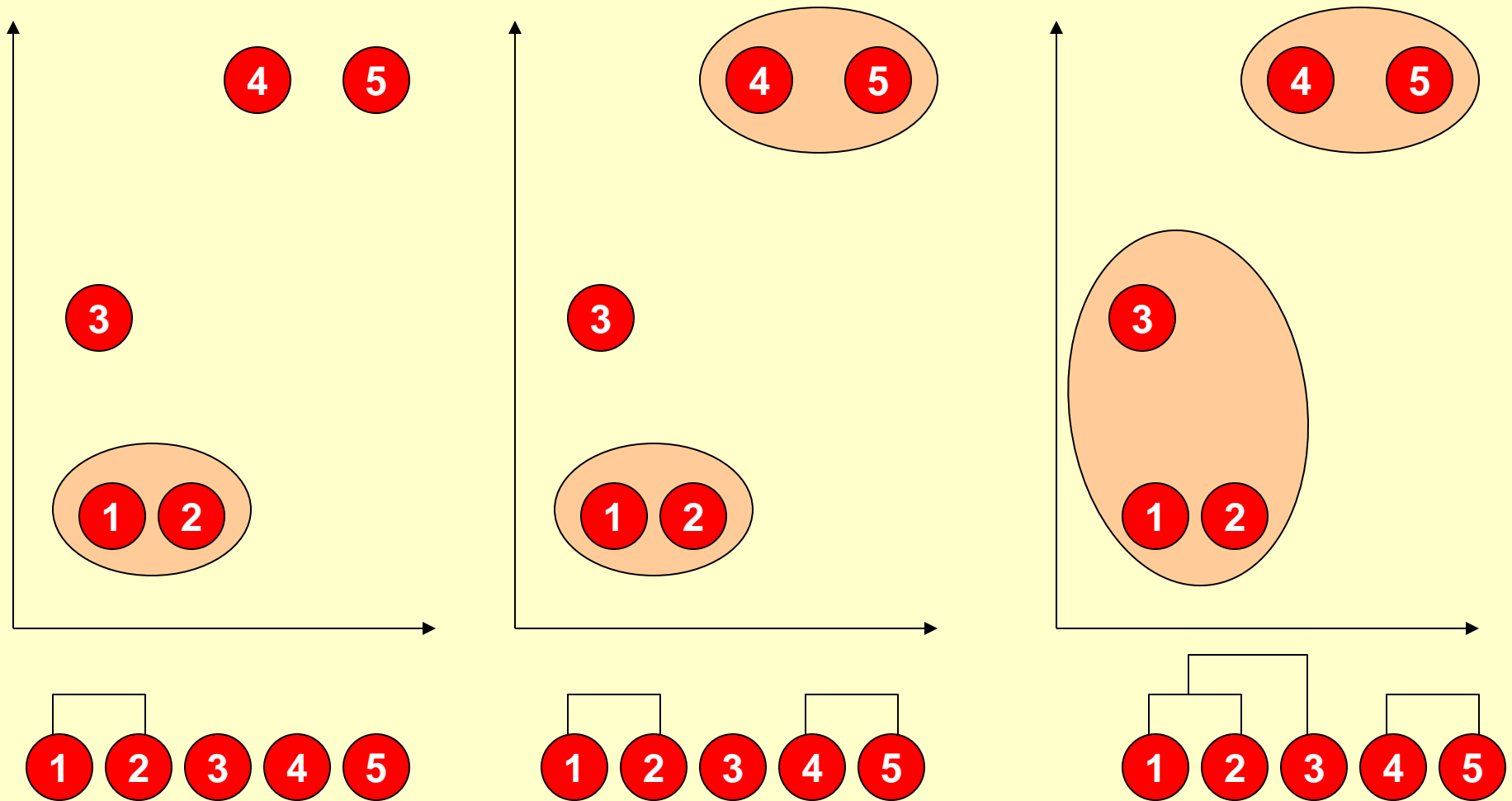
# Algoritma Hierarchical clustering

---

1. Tentukan  $k$  sebagai jumlah cluster yang ingin dibentuk
2. Setiap data dianggap sebagai cluster. Kalau  $n$ =jumlah data dan  $nc$ =jumlah cluster, berarti ada  $nc=n$ .
3. Hitung jarak antar cluster
4. Cari 2 cluster yang mempunyai jarak antar cluster yang paling minimal dan gabungkan (berarti  $nc$  berkurang)
5. Jika  $nc > k$ , kembali ke langkah 3



# Algoritma Hierarchical clustering

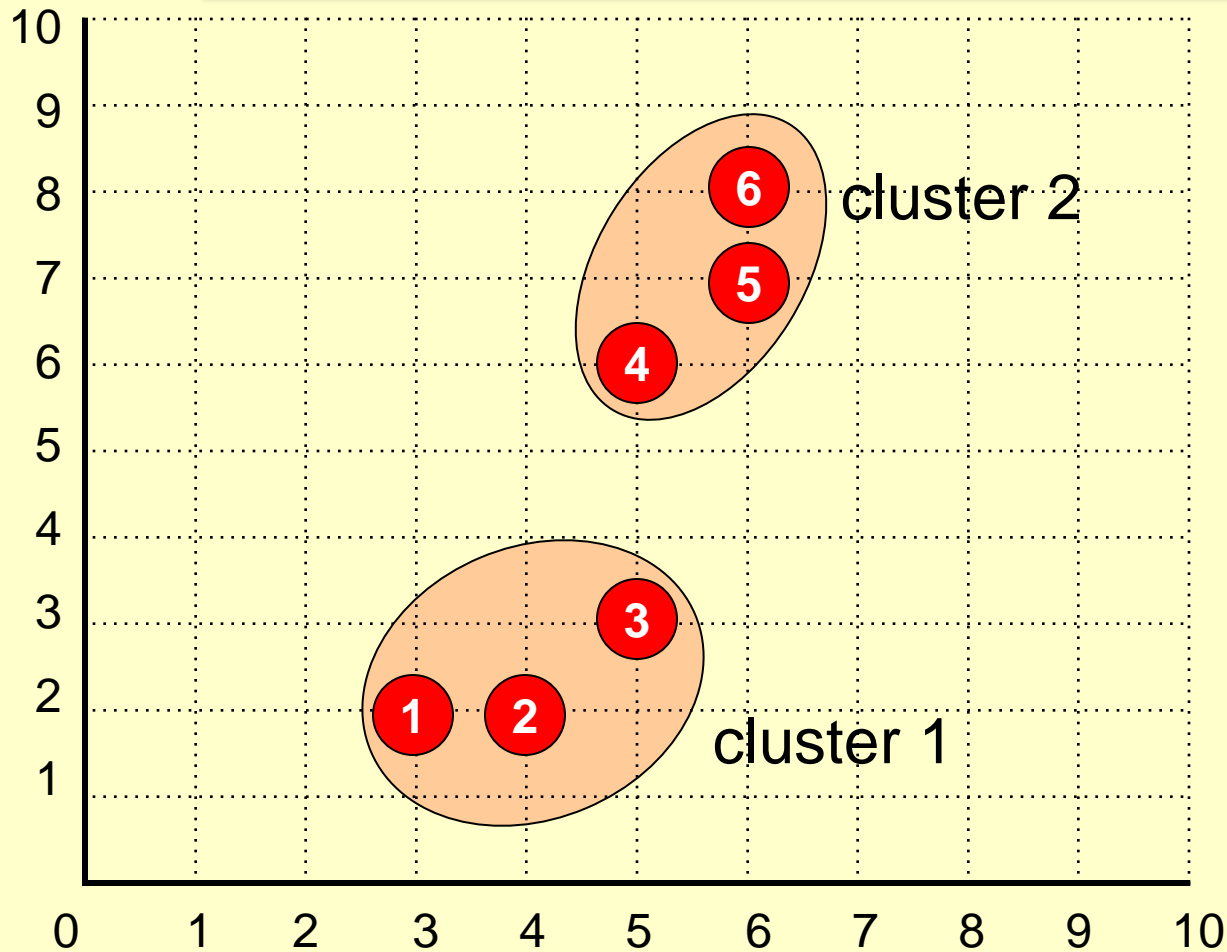


# Similarity between clusters?

---

- Single Linkage
  - Minimum distance between cluster
- Centroid Linkage
  - Centroid distance between cluster
- Complete Linkage
  - Maximum distance between cluster
- Average Linkage
  - Average distance between cluster

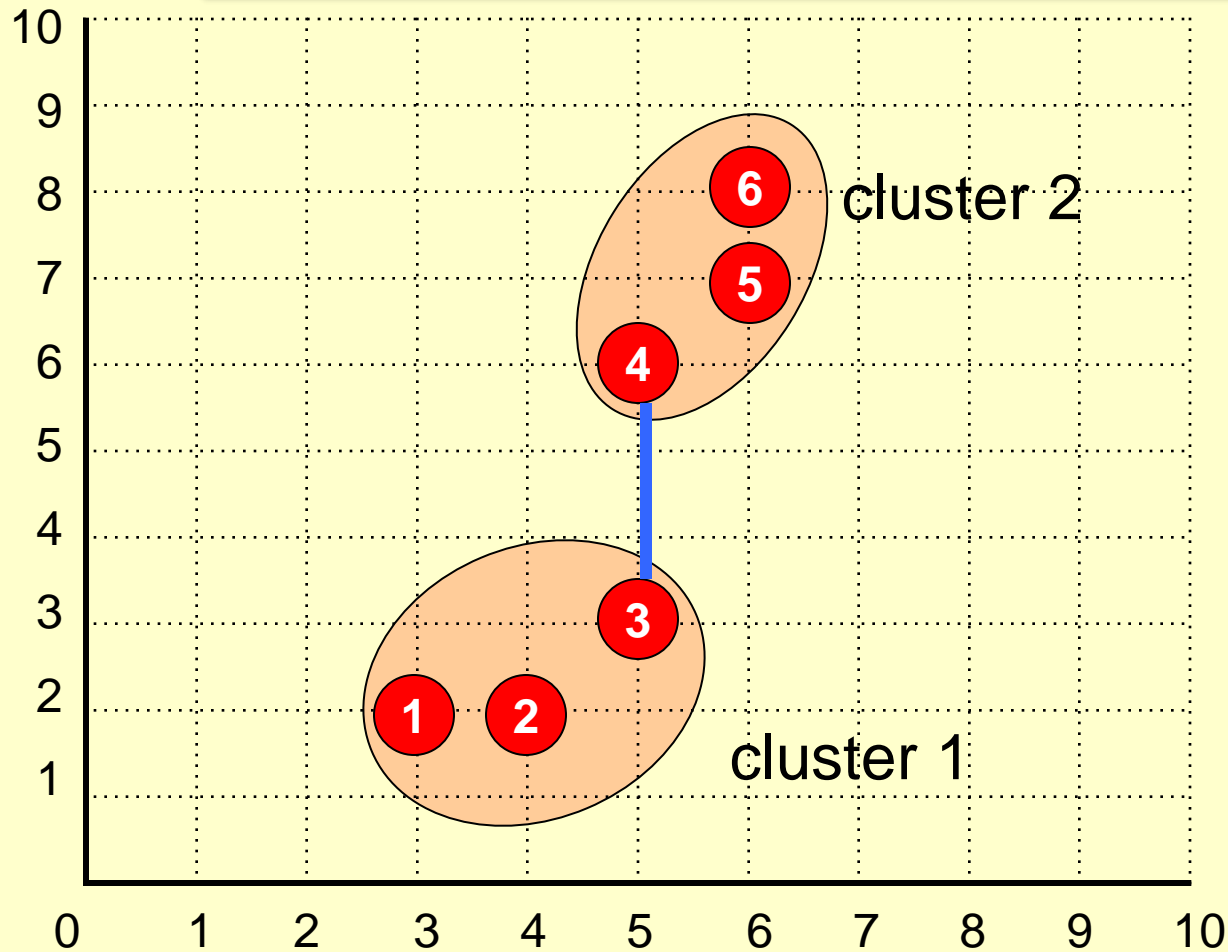
# Pengukuran jarak



Berapa jarak  
cluster 1 ke  
cluster 2

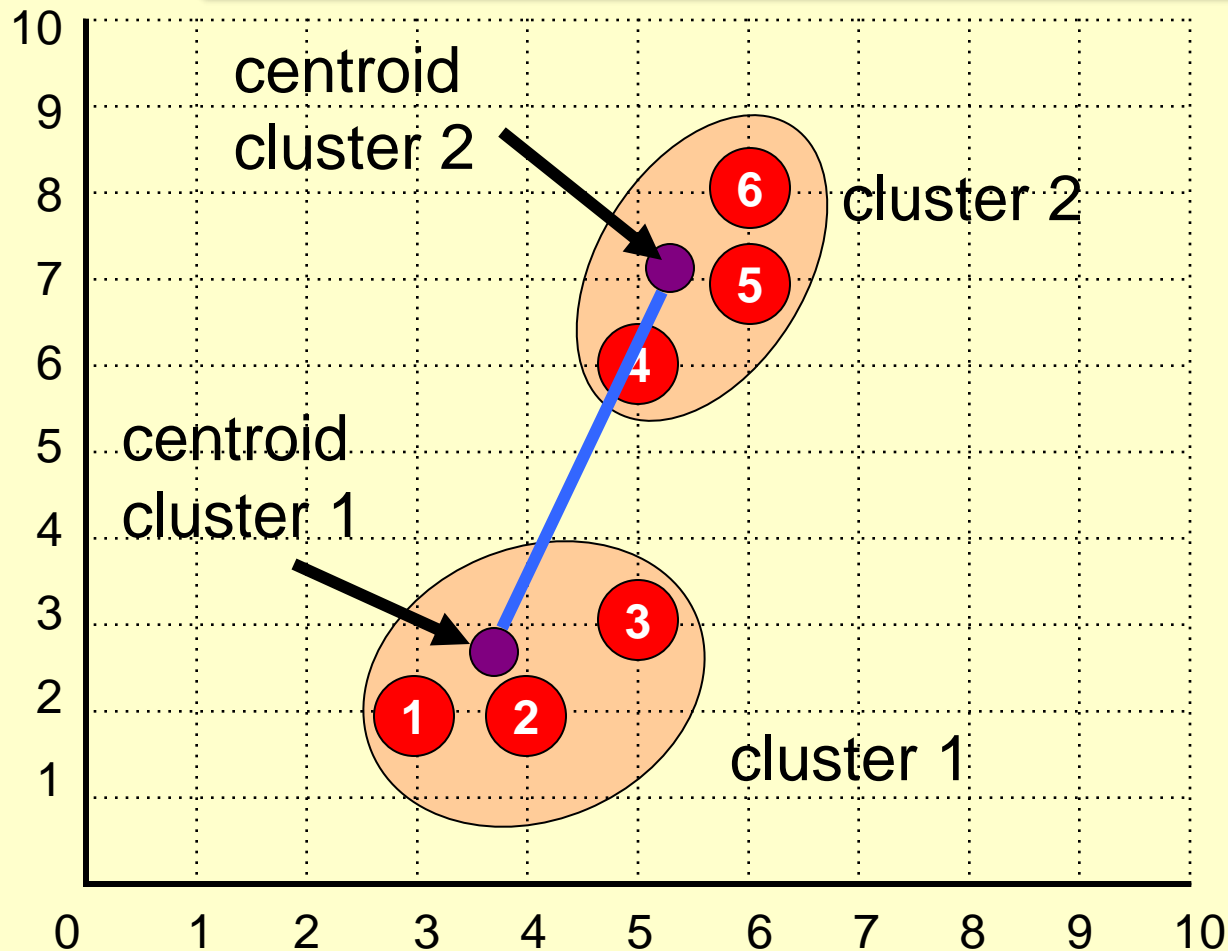
?

# Single Linkage



Jarak cluster 1  
ke cluster 2  
=  
Jarak data 3 ke  
data 4

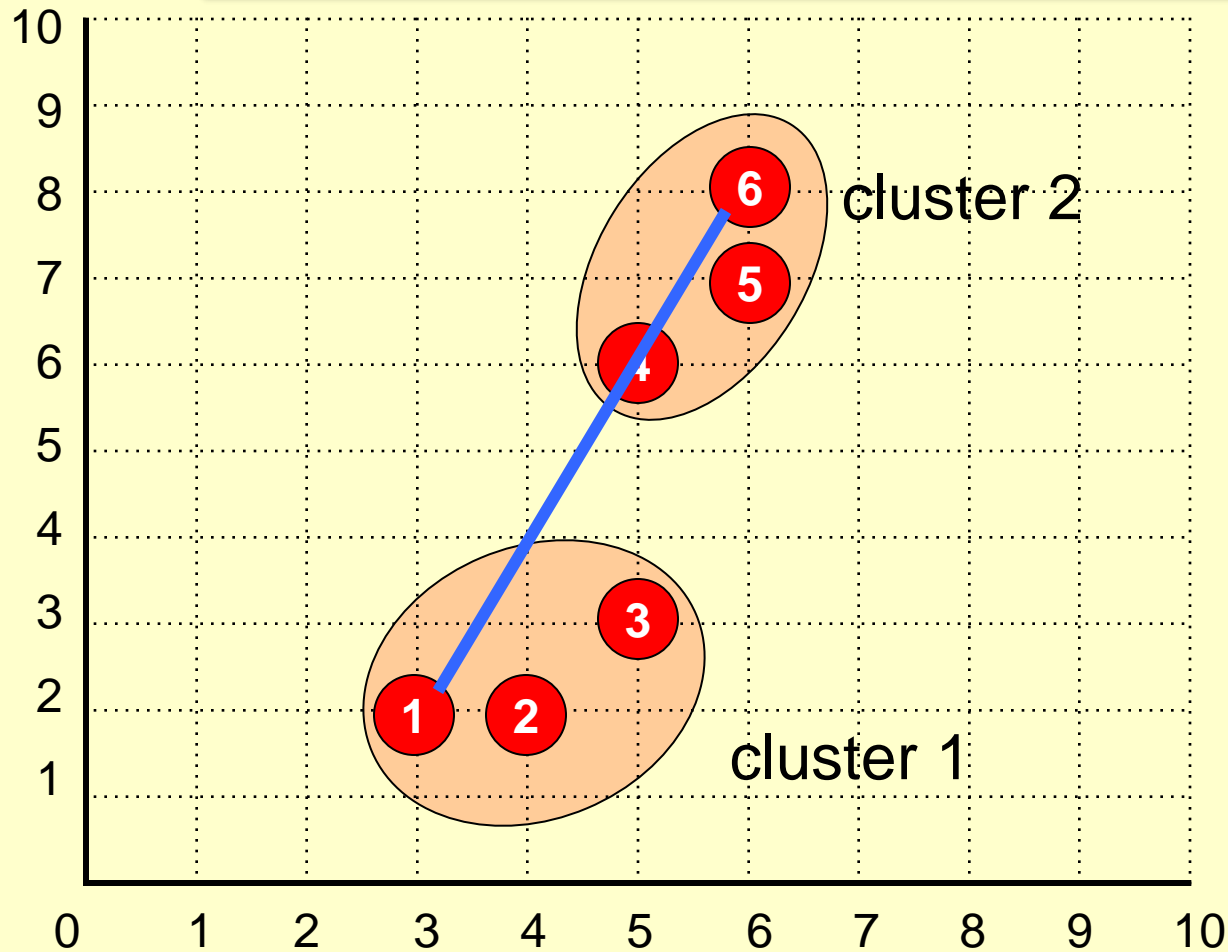
# Centroid Linkage



Jarak cluster 1 ke  
cluster 2  
=  
Jarak centroid  
cluster 1 ke  
centroid cluster 2

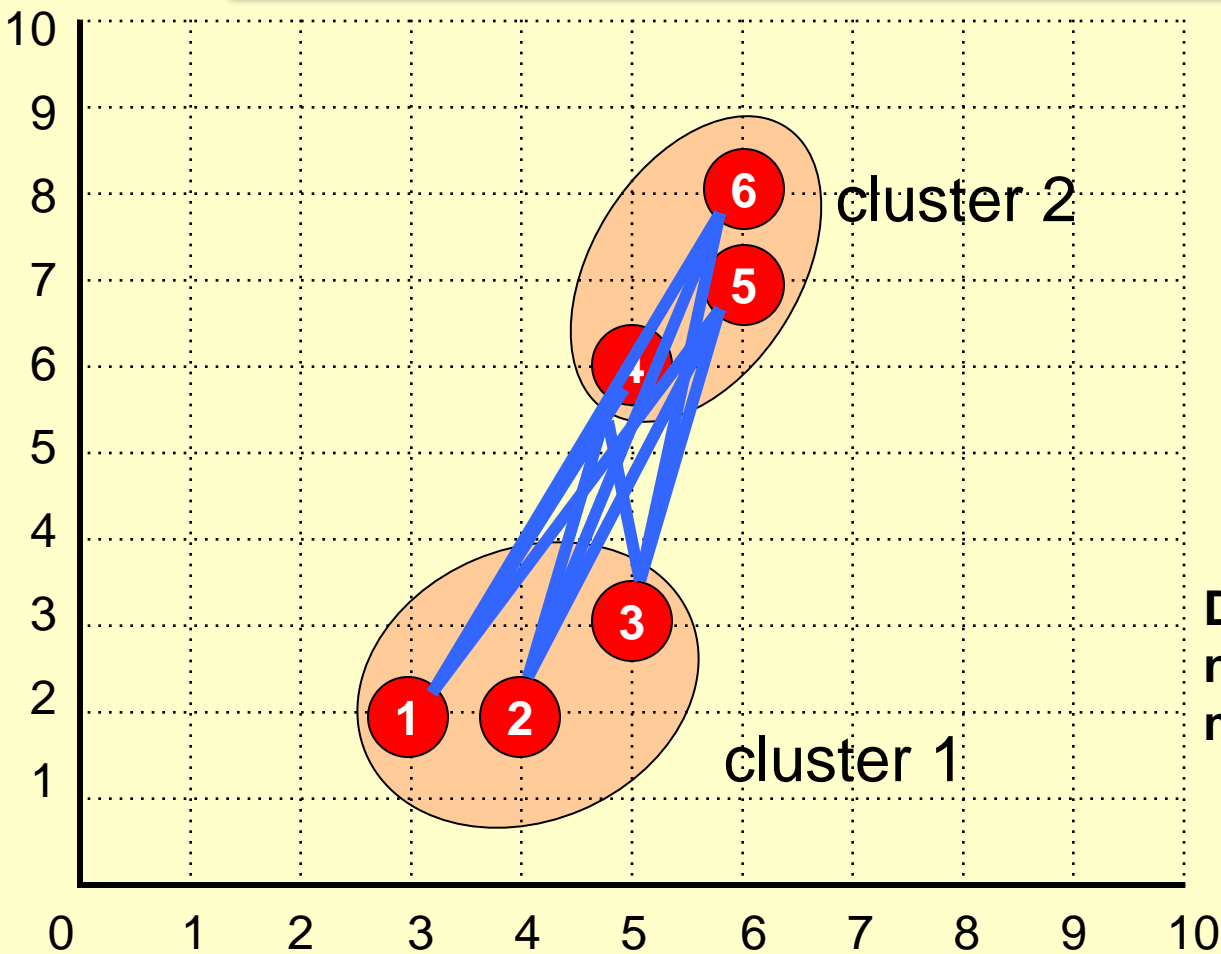


# Complete Linkage



Jarak cluster 1  
ke cluster 2  
=  
Jarak data 1 ke  
data 6

# Average Linkage



$$\begin{aligned} &\text{Jarak cluster 1} \\ &\text{ke cluster 2} \\ &= \\ &\frac{\sum \text{Jarak antar data}}{n \times m} \end{aligned}$$

Dimana:  
n=jumlah data pada cluster 1  
m=jumlah data pada cluster 2

# Hierarchical Clustering & Dataset

---

- **Single Linkage**

Metode ini sangat cocok untuk dipakai pada kasus shape independent clustering, karena kemampuannya untuk membentuk pattern tertentu dari cluster. Untuk kasus condensed clustering, metode ini tidak bagus.

- **Centroid Linkage**

Metode ini baik untuk kasus clustering dengan normal data set distribution. Akan tetapi, metode ini tidak cocok untuk data yang mengandung outlier.

- **Complete Linkage**

Metode ini sangat ampuh untuk memperkecil variance within cluster karena melibatkan centroid pada saat penggabungan antar cluster. Metode ini juga baik untuk data yang mengandung outlier.

- **Average Linkage**

Metode ini relatif yang terbaik dari metode-metode hierarchical. Namun, ini harus dibayar dengan waktu komputasi yang paling tinggi dibandingkan dengan metode-metode hierarchical yang lain.