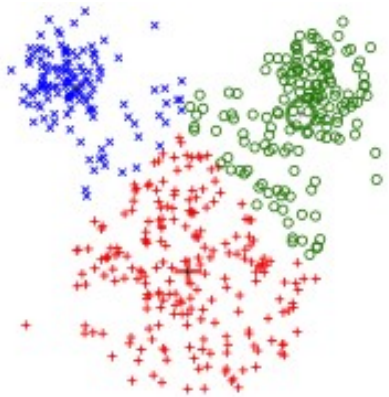


Clustering & Cluster Analysis

Ali Ridho Barakbah

Knowledge Engineering Research Group

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

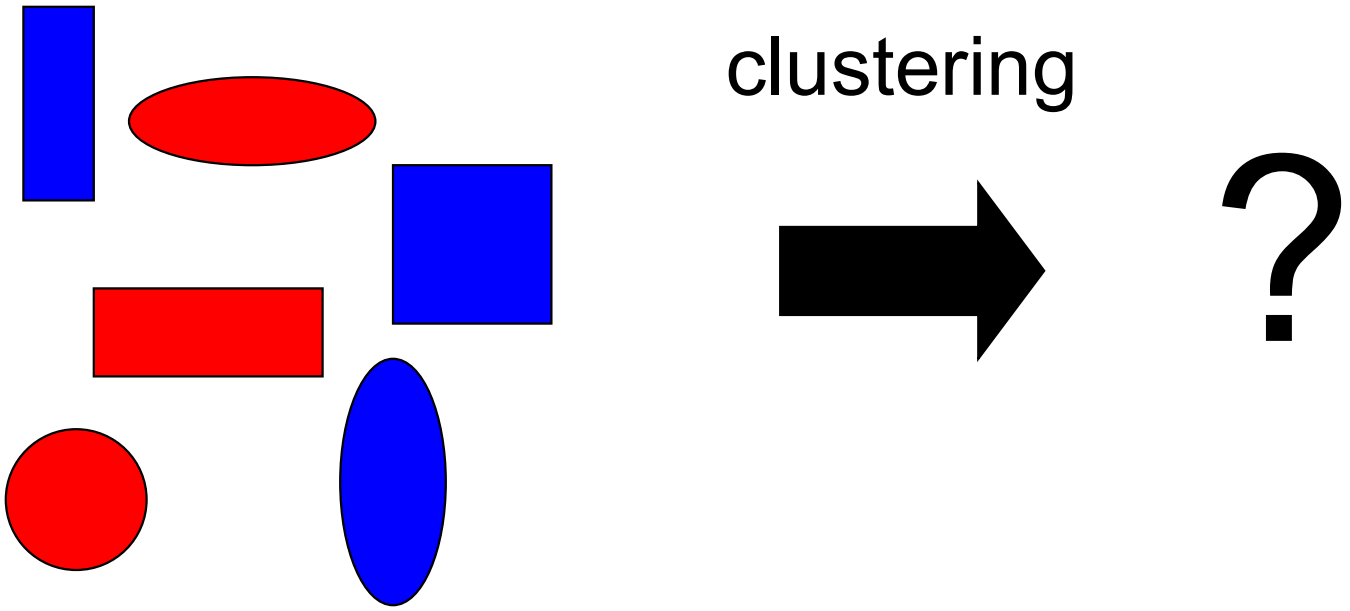
http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/index.html

What is clustering?

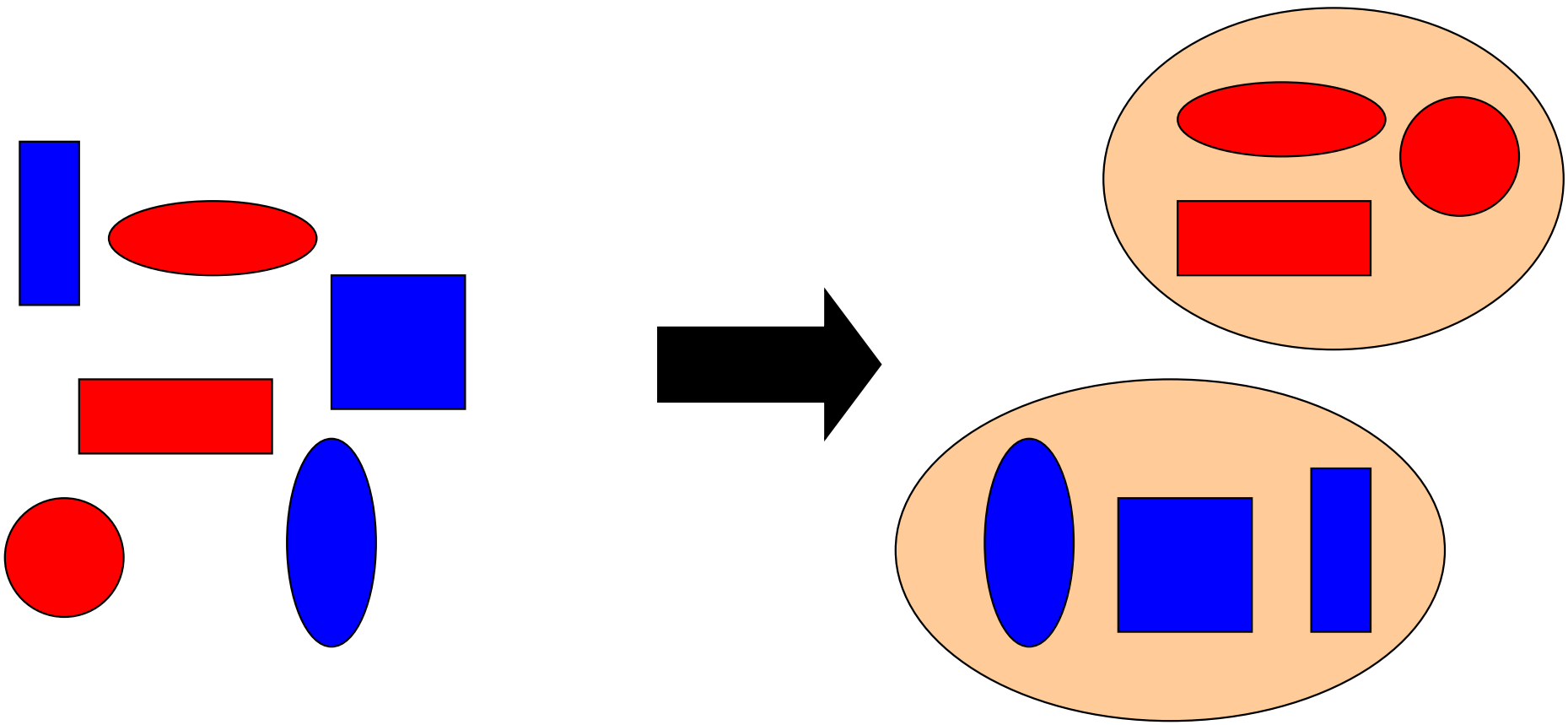
the process of organizing objects into groups whose members are similar in some way

http://www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html/index.html

Ilustrasi clustering

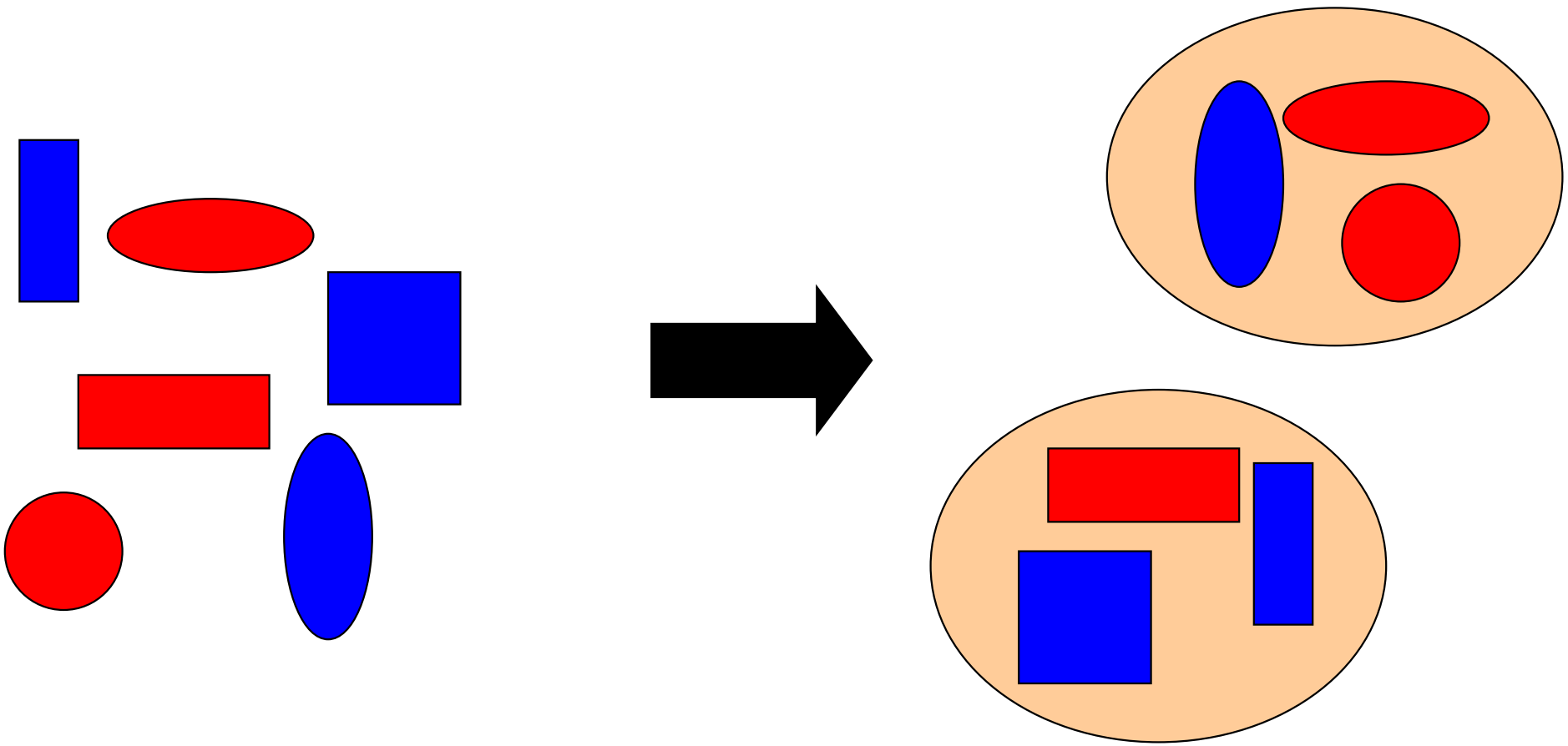


Ilustrasi clustering



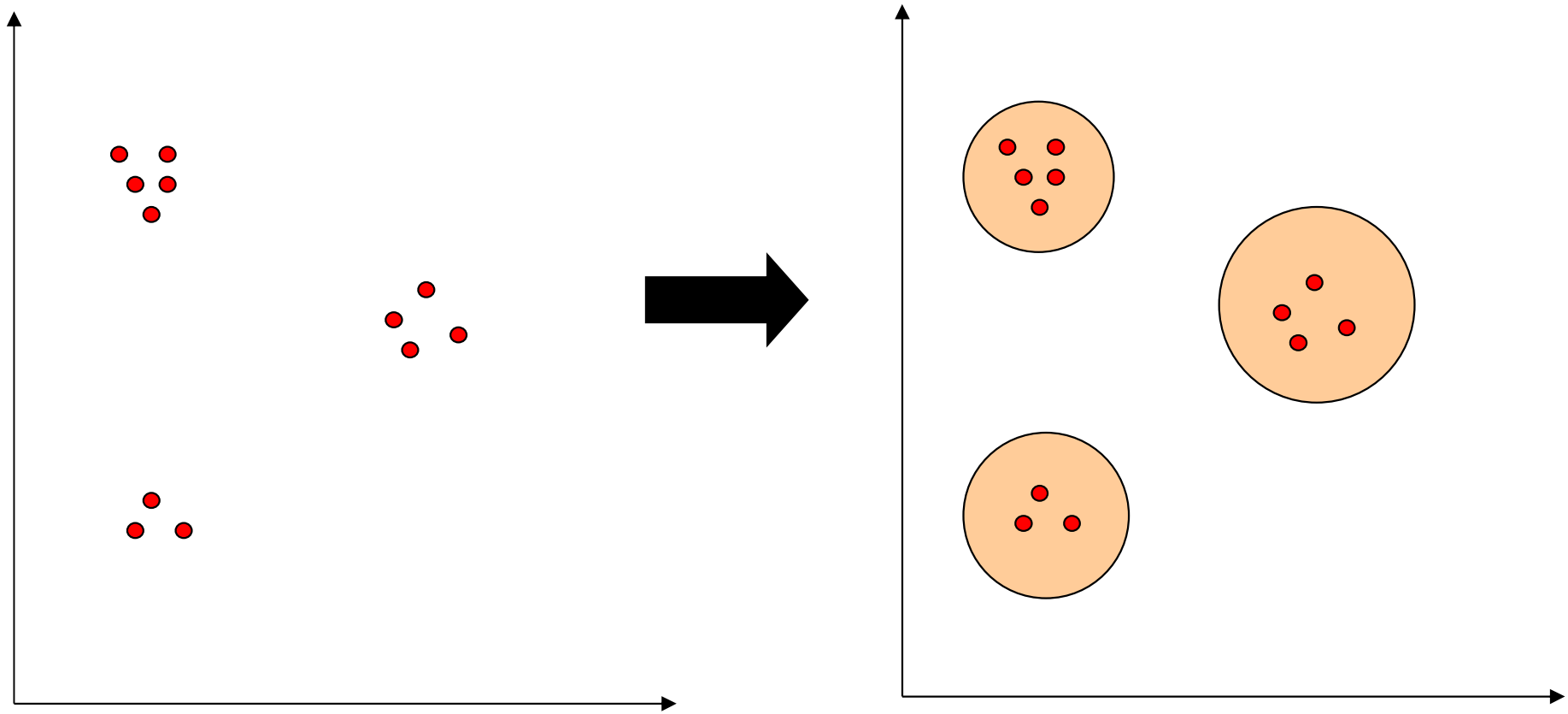
Similaritas berdasarkan warna

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)

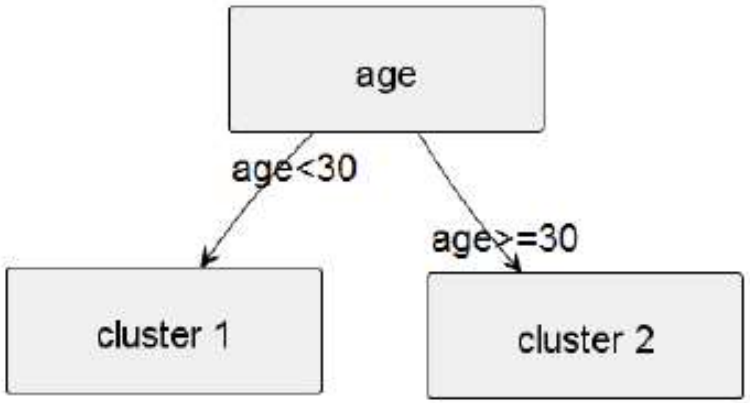
Data penyakit hipertensi

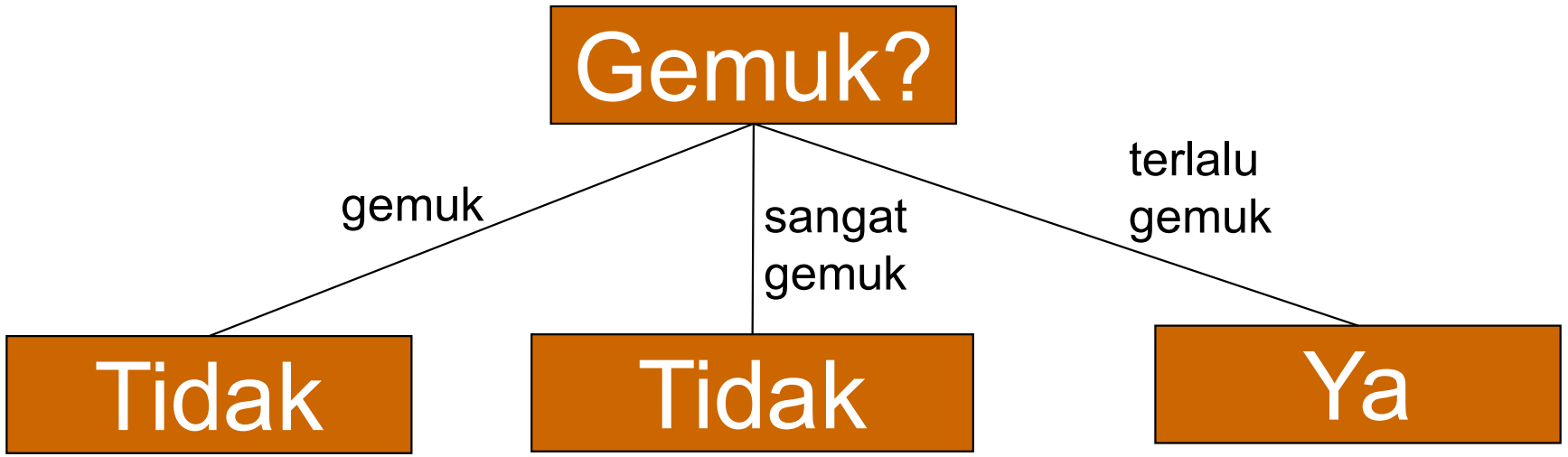
Umur	Kegemukan	Hipertensi
muda	gemuk	Tidak
muda	sangat gemuk	Tidak
paruh baya	gemuk	Tidak
paruh baya	terlalu gemuk	Ya
tua	terlalu gemuk	Ya

label

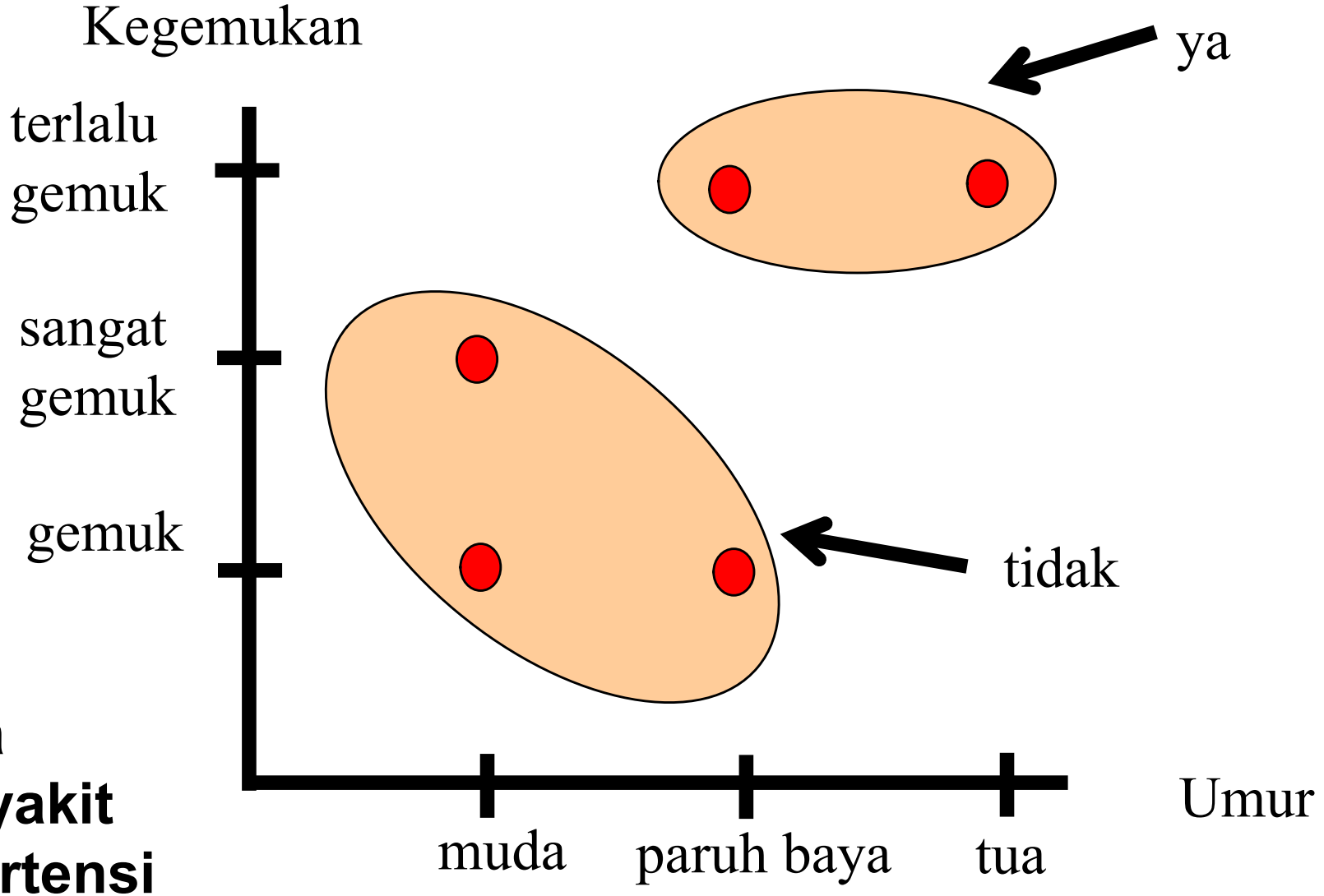
Supervised data

Penyelesaian dengan Decision Tree





Classification
(kasus sederhana)



Data penyakit hipertensi

Clustering
(kasus sederhana)

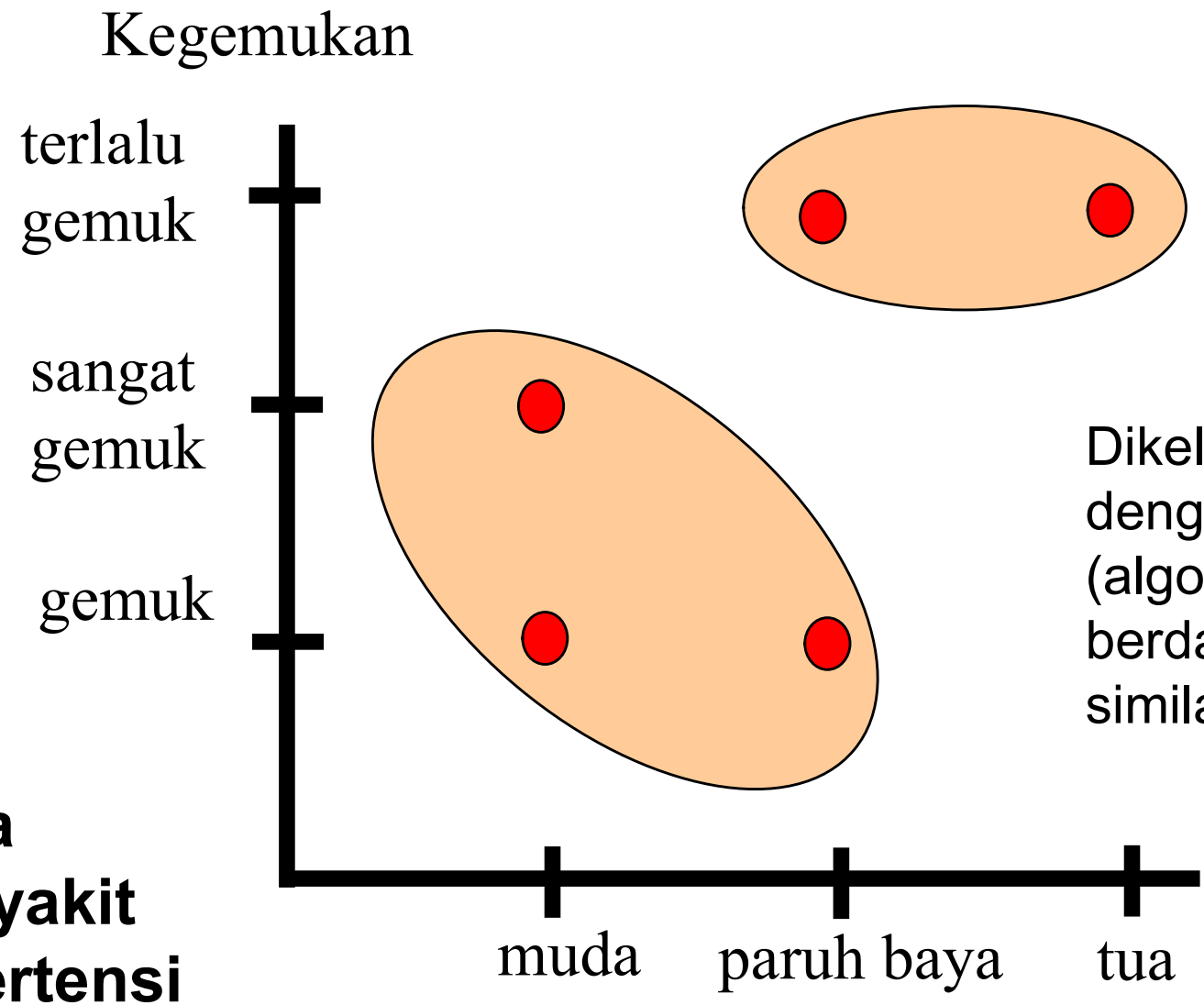
Data penyakit hipertensi

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)



Dikelompokkan dengan cara (algoritma) tertentu berdasarkan similaritas tertentu

Data penyakit hipertensi

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 Clustering

- Developed by Mac Queen (1967)
- A partitioning clustering method that separates data into *k* mutually excessive groups.
- K-means clustering minimizes the sum of distance from each data point to its clusters.
- The most well known, widely used and fast method for clustering because of its ability to cluster a kind of huge data and also outliers quickly and efficiently.

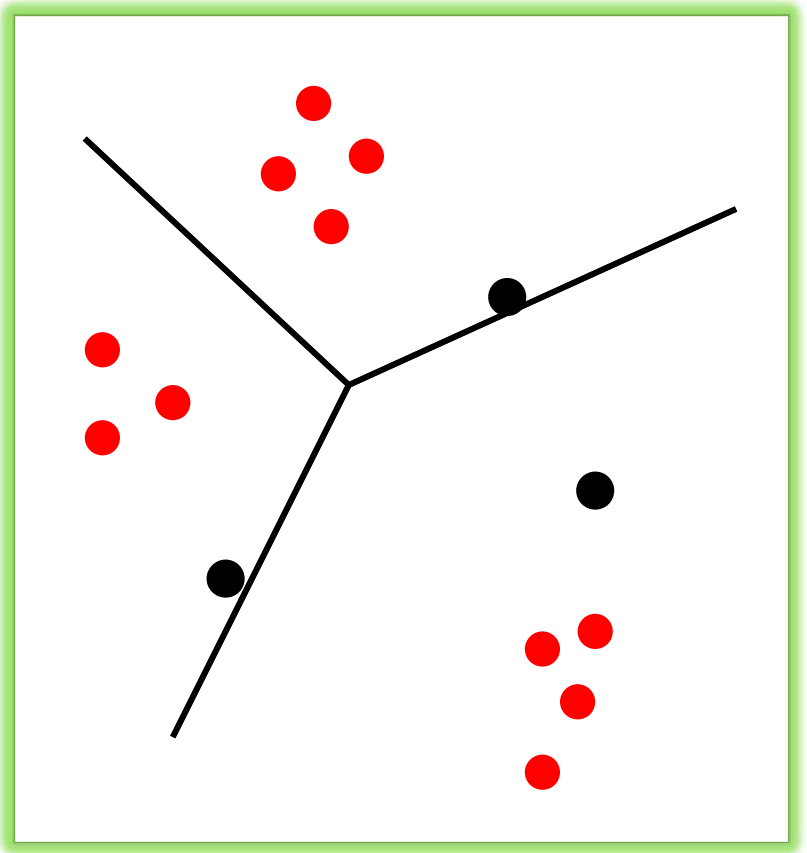
K-means Algorithm

- Let $A=\{a_i \mid i=1,\ldots,f\}$ be attributes of f-dimensional vectors and $X=\{x_i \mid i=1,\ldots,N\}$ be each data of A.
- The K-means clustering separates X into k partitions called clusters $S=\{s_i \mid i=1,\ldots,k\}$ where $M \in X$ is $M_i=\{m_{ij} \mid j=1,\ldots, n(s_i)\}$ as members of s_i, where $n(s_i)$ is number of members for s_i. Each cluster has cluster center of $C=\{c_i \mid i=1,\ldots,k\}$.
- K-means clustering algorithm can be described as follows:

1. Initiate its algorithm by generating random starting points of initial centroids C.
2. Calculate the distance d between X to cluster center C. Euclidean distance is commonly used to express the distance.
3. Separate x_i for $i=1..N$ into S in which it has minimum $d(x_i,C)$.
4. Determine the new cluster centers c_i for $i=1..k$ defined as:

$$c_i = \frac{1}{n_i} \sum_{j=1}^{n(s_i)} m_{ij} \in s_i$$

5. Go back to step 2 until all centroids are convergent.

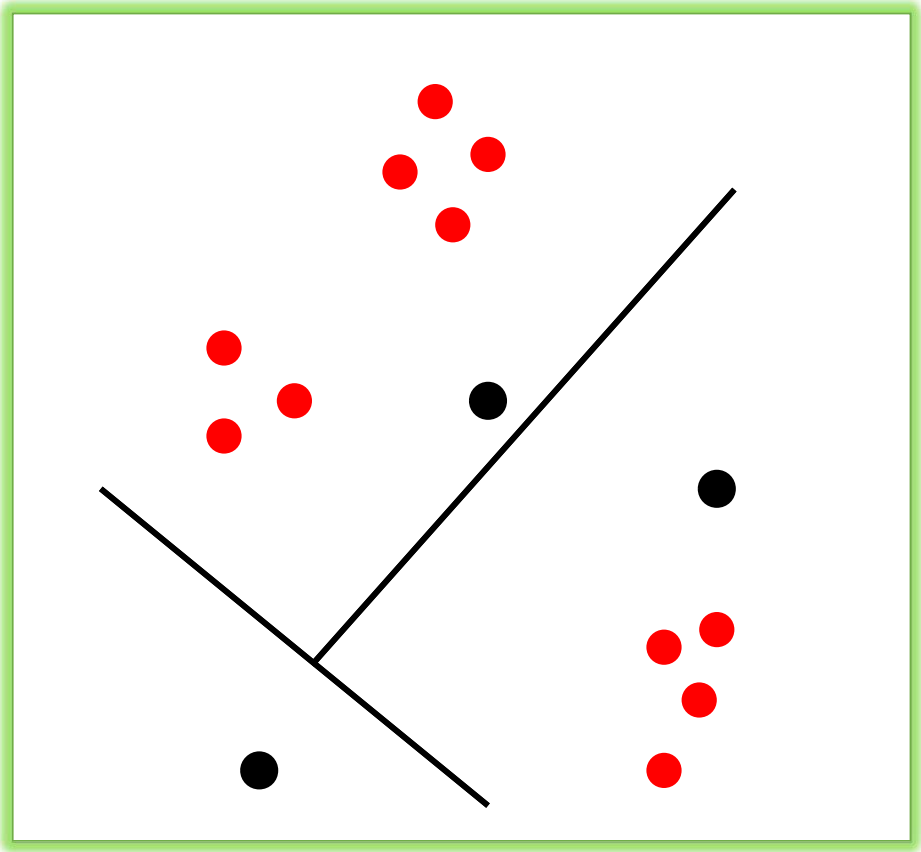


K-means drawback:
cluster centers far from the final cluster

K-means clustering generates initial starting points randomly.

If designated initial starting points randomly close to final cluster centers, then K-means clustering can find the final cluster centers.

If they are far from the final cluster centers, they will lead to incorrect clustering results

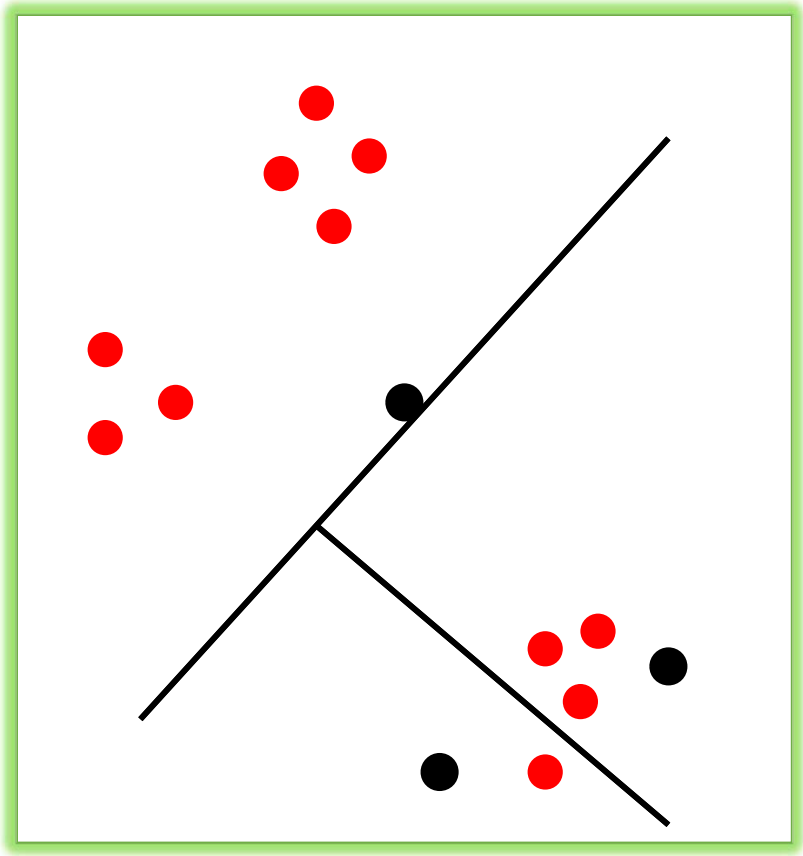


K-means drawback:
may be trapped in the local optima

The K-means algorithm generates the initial centroids randomly and fails to consider a spread out placement of them spreading within the feature space.

In this case, the initial centroids may be placed so close together that some become inconsequential.

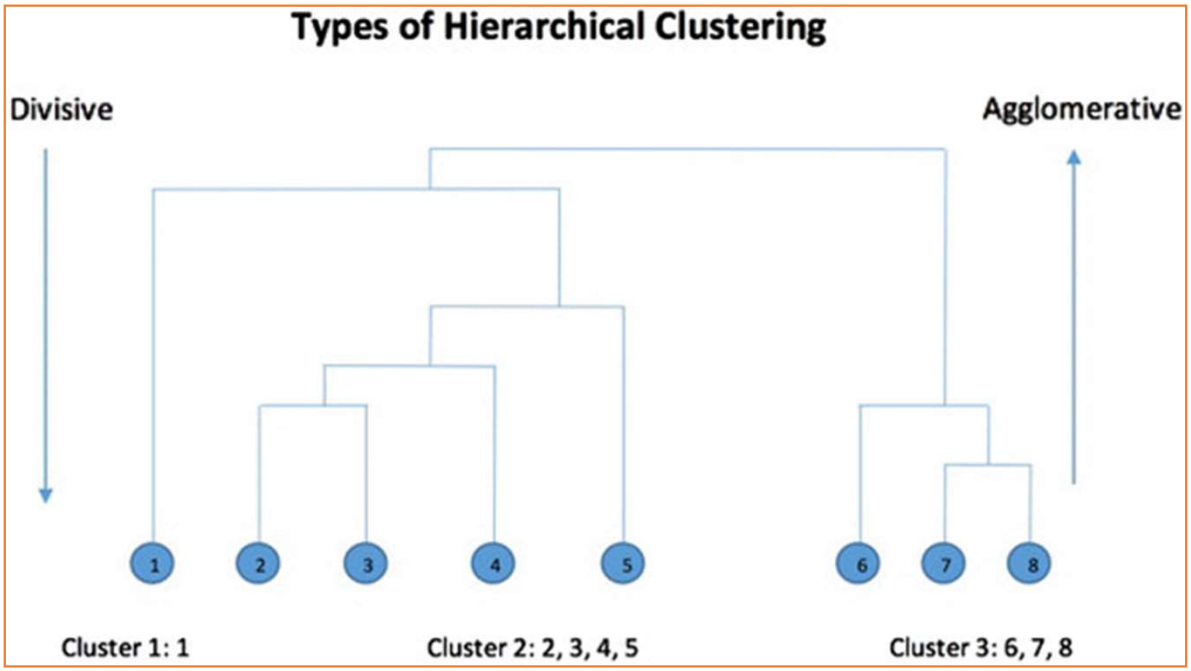
Because of this, the initial centroids generated by K-means may be trapped in the local optima.



Hierarchical Clustering

Divisive/partitioning
- 1 cluster to k clusters
- Top to down division

Agglomerative/Hierarchical
- N clusters to k clusters
- Down to top merge



Clustering stage

- Single Linkage
  - Step by step clustering
- Multi Linkage
  - Multi step clustering

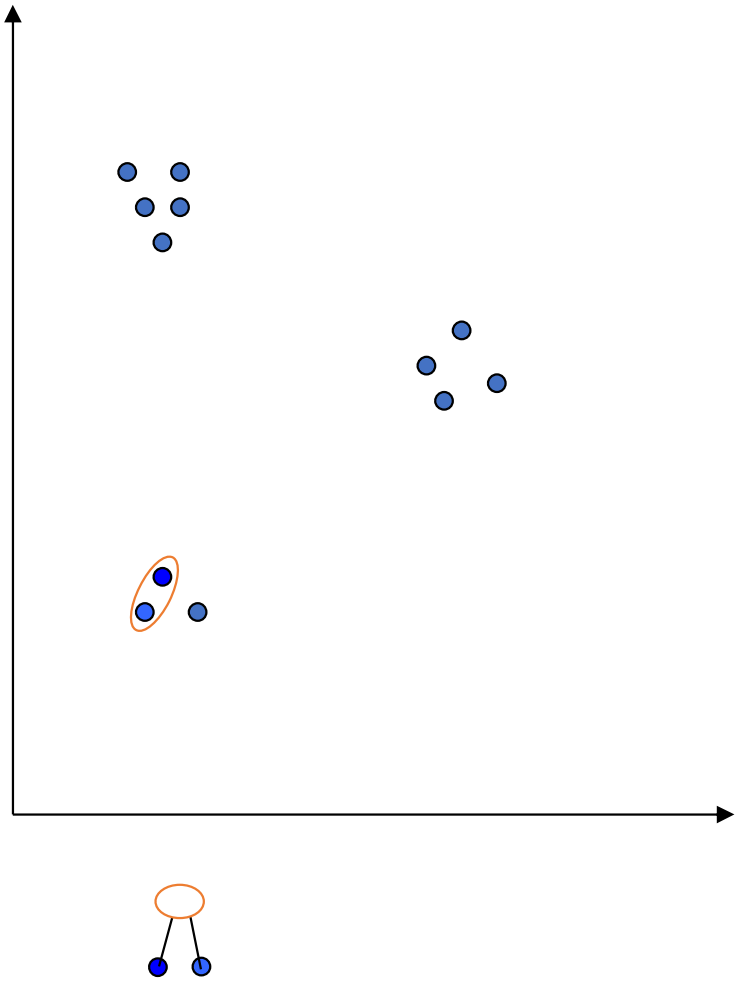
How it works: Aglomerative

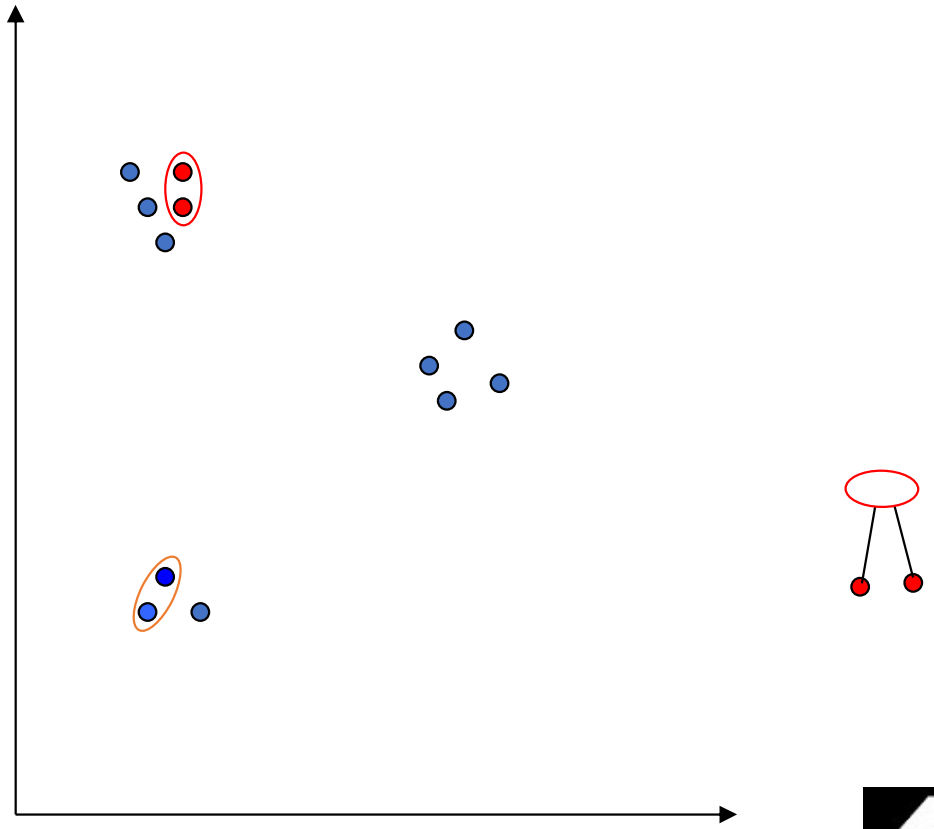
Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering (defined by S.C. Johnson in 1967) is this:

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

Single Linkage algorithm:

1. Every point is it's own cluster
2. Find most nearest cluster
3. Merge it into a parent cluster
4. Repeat

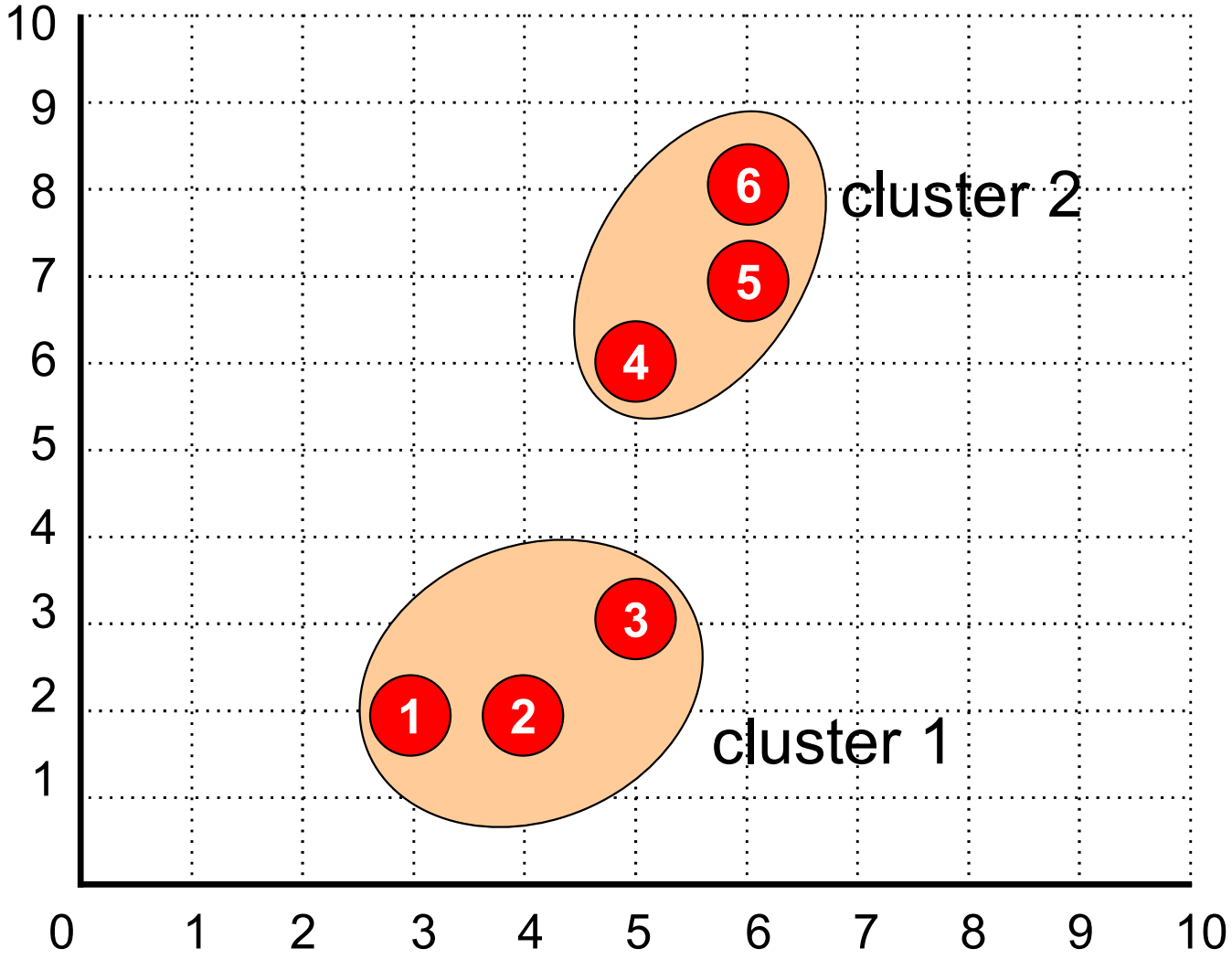




How do we define similarity between cluster?

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

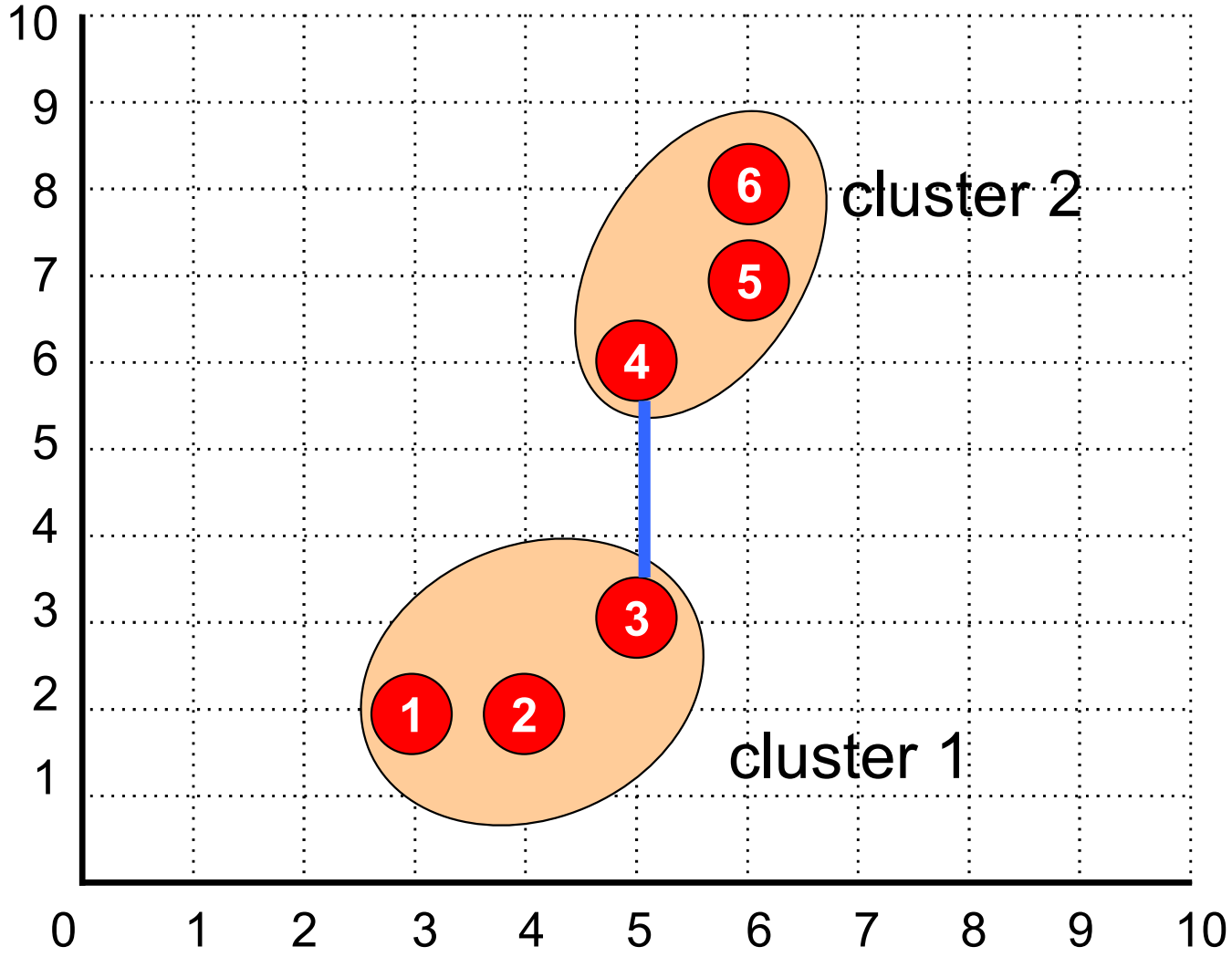
Distance Measurement



Distance from cluster 1 to cluster 2

?

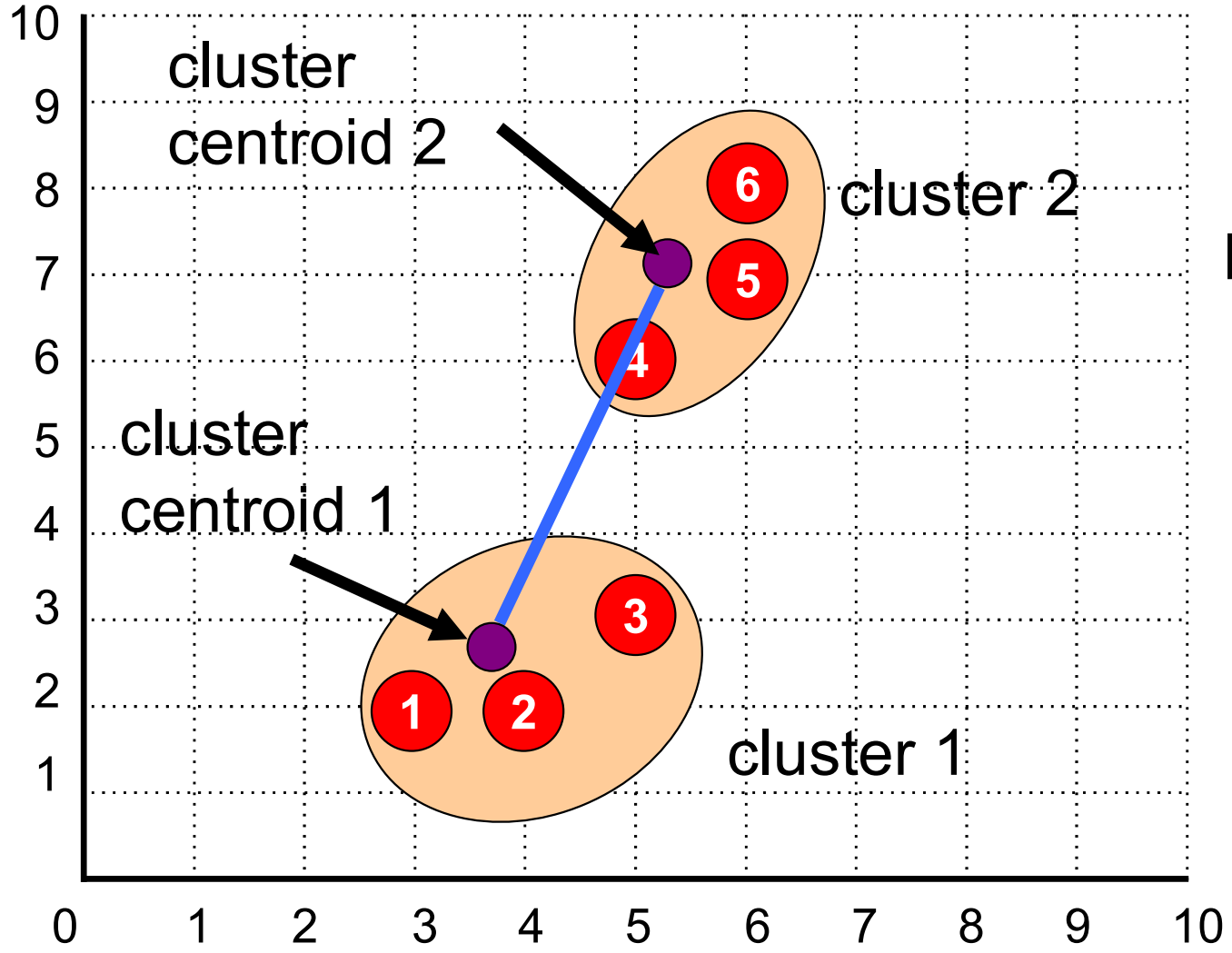
Single Linkage
Minimum distance between cluster



Distance from cluster 1 to cluster 2
=
Distance of data 3 to data 4

Centroid Linkage

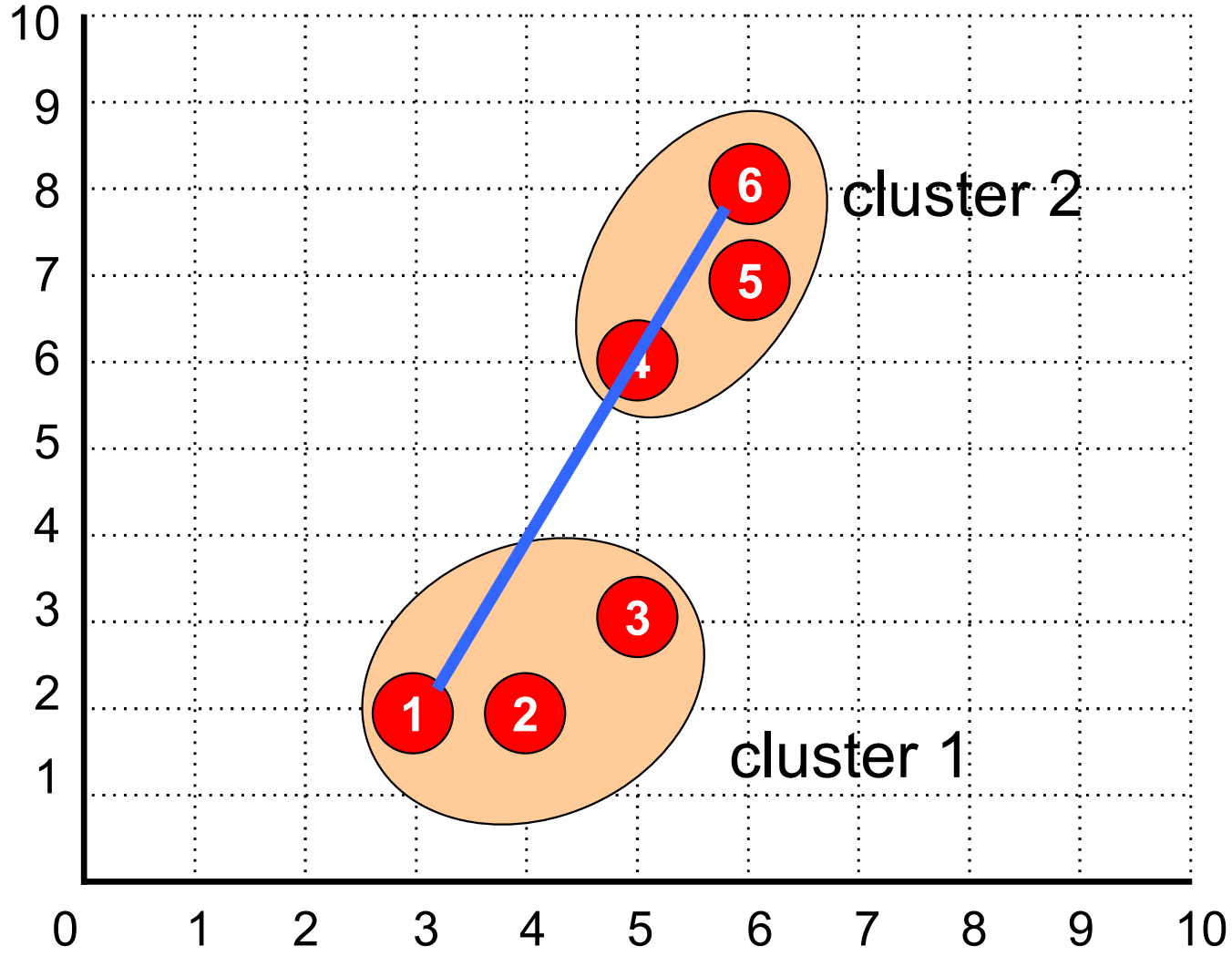
Centroid distance between cluster



Distance from cluster 1 to cluster 2
=
Distance of cluster centroid 3 to cluster centroid 4

Complete Linkage

Maximum distance between cluster

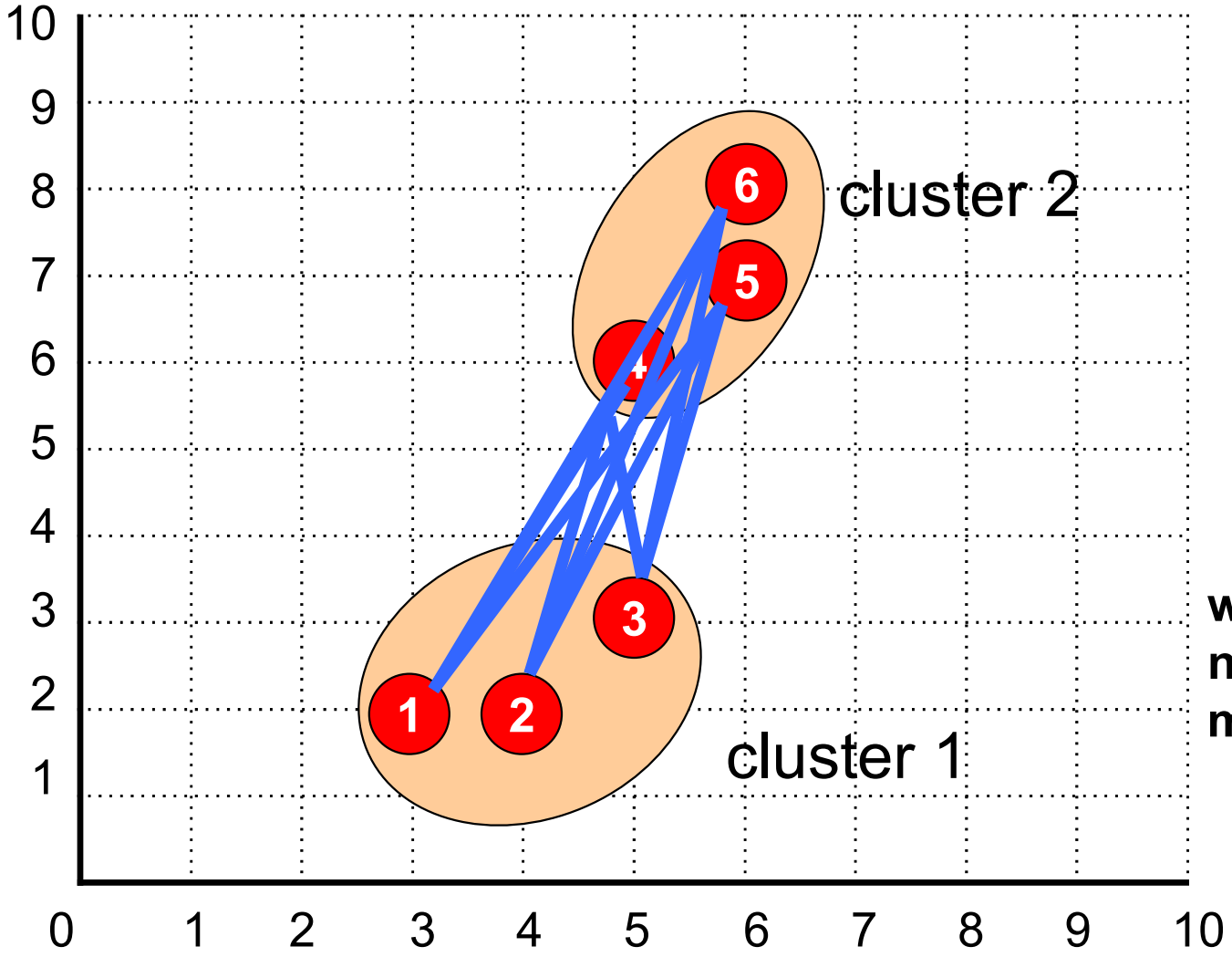


Distance from cluster 1 to cluster 2

=

Distance of data 1 to data 6

Average Linkage
Average distance between cluster



Distance of cluster 1 to cluster 2
=
∑Distance among data
----
n x m

where:
n=number of data in cluster 1
m=number of data in 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.

Cluster Analysis

- Variance
- Sum of Squared Error
- Centroid Proximity Index
- Error ratio

Variance

- Digunakan untuk mengukur nilai penyebaran dari data-data hasil clustering
- Dipakai untuk data yang bertipe unsupervised
- Variance pada clustering ada 2 macam:
  - Variance within cluster
  - Variance between clusters

Good cluster

is when the members of a cluster have a high degree of similarity to each other (internal homogeneity) and are not like members of other clusters (external homogeneity)

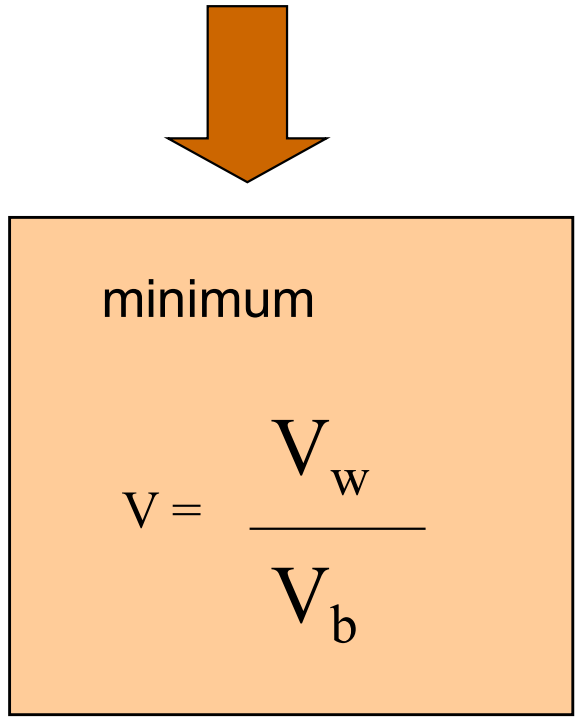
Variance & homogeneity

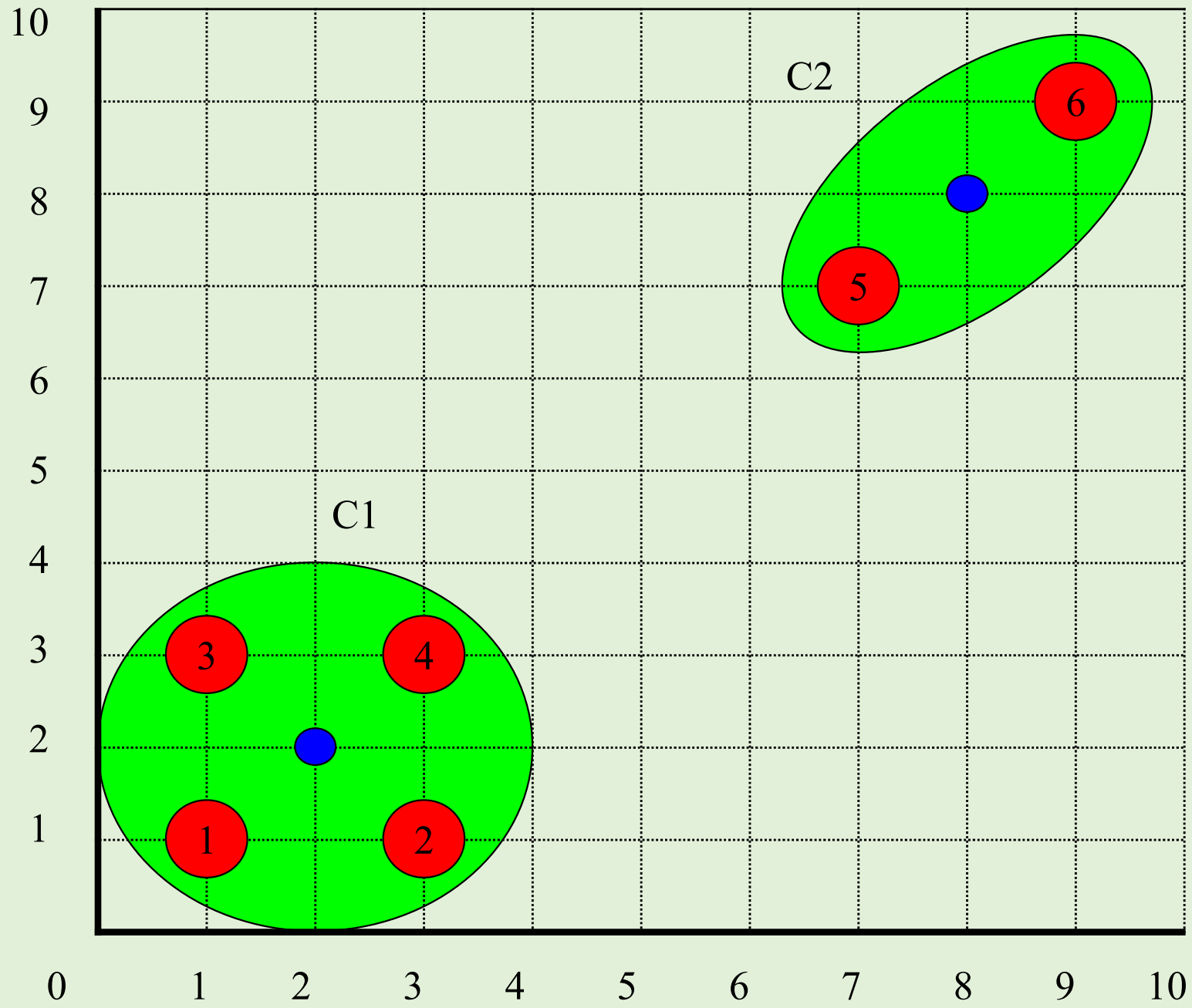
internal homogeneity → Variance within cluster (V_w)

external homogeneity → Variance between clusters (V_b)

Ideal cluster

- The ideal cluster has minimum V_w to express internal homogeneity and maximum V_b to express external homogeneity.




C1
C2
1
2
3
4
5
6
0
1
2
3
4
5
6
7
8
9
10

Cluster Variance

$$v_c^{\,2} = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} \left(d_i - \overline{d}_i \right)^2$$

$v_c^{\,2}$ = variance pada cluster c
$c = 1..k$, dimana k = jumlah cluster
n_c = jumlah data pada cluster c
d_i = data ke-i pada suatu cluster
$\overline{d}_i$ = rata-rata dari data pada suatu cluster

Variance within cluster

$$v_w = \frac{1}{N-k} \sum_{i=1}^{k} (n_i - 1) \cdot v_i^2$$

v_w = variance within cluster
N = jumlah semua data

Variance between clusters

$$v_b = \frac{1}{k-1}\sum_{i=1}^{k} n_i \left(\overline{d_i} - \overline{\overline{d}}\right)^2$$

$\overline{\overline{d}}$ = rata-rata dari $\overline{d_i}$

Variance dari semua cluster

$$\nu = \frac{\nu_w}{\nu_b}$$

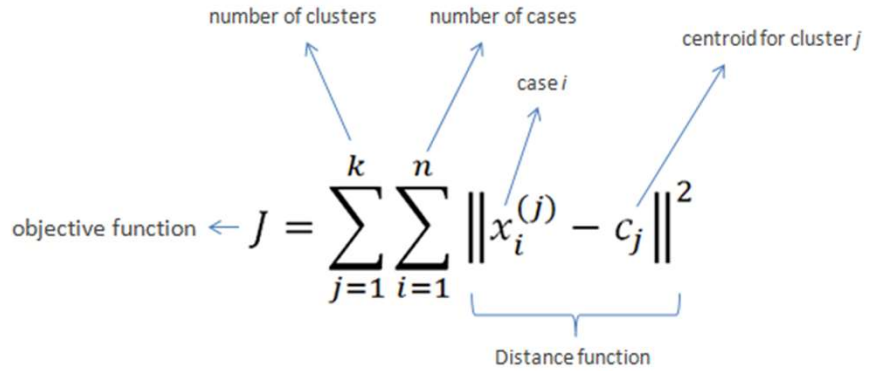
Sum of Squared Error

The most widely used criterion to quantify cluster homogeneity is the Sum of Squared Error (SSE) criterion

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n(s_i)} \left\| m_{ij} - \bar{s}_i \right\|^2$$

$$MSE = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n(s_i)} \left\| m_{ij} - \bar{s}_i \right\|^2}{N}$$

SSE



Centriod Proximity Index

- Proposed by Barakbah and Kiyoki[1]
- To analyze the closeness of the final centroids of the clustering result to the centroids of the real data sets.

$$CPI = \min \sum_{i=1}^{k} (\|c_i - r_i\|)$$

where c_i is i-th final centroid of clustering result and r_i is i-th real centroid of datasets.

[1]Ali Ridho Barakbah, Yasushi Kiyoki, "A Fast Algorithm for K-Means Optimization using Pillar Algorithm", The 2nd International Workshop with Mentors on Databases, Web and Information Management for Young Researchers, August 2-4, 2010, Tokyo, Japan.

Error ratio

- Dipakai jika dataset yang digunakan adalah supervised
- Biasanya digunakan untuk mengukur tingkat presisi dari metode clustering
- Rumus:

$$Error = \frac{missclassified}{jumlahdata} \times 100\%$$

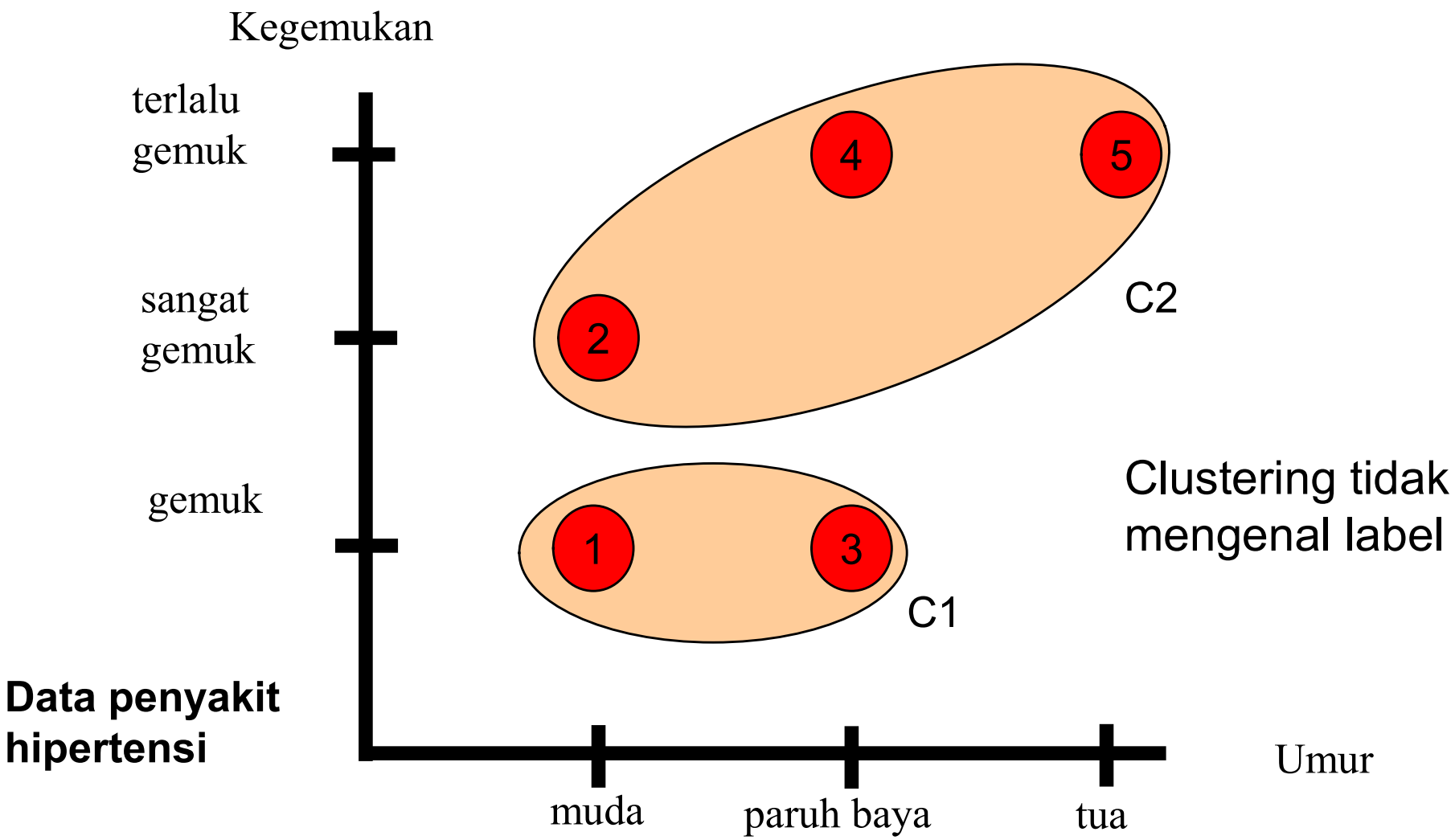
Contoh sederhana

Data penyakit hipertensi

Data ke-	Umur	Kegemukan	Hipertensi
1	muda	gemuk	Tidak
2	muda	sangat gemuk	Tidak
3	paruh baya	gemuk	Tidak
4	paruh baya	terlalu gemuk	Ya
5	tua	terlalu gemuk	Ya

Supervised data

Contoh Hasil Clustering



Menghitung error ratio

	Label	Kombinasi 1 C1 → Tidak C2 → Ya	Kombinasi 2 C1 → Ya C2 → Tidak
Data 1	Tidak	Tidak	Ya
Data 2	Tidak	Ya	Tidak
Data 3	Tidak	Tidak	Ya
Data 4	Ya	Ya	Tidak
Data 5	Ya	Ya	Tidak
Misclassified		1	4
Error ratio		20%	80%

Semua kemungkinan label dicoba sehingga ada n! kombinasi

Ambil error ratio yang terkecil