Instance based learning (Nearest Neighbor)

Ali Ridho Barakbah

Nearest Neighbor (NN)

- Merupakan suatu method untuk mengklasifikasikan suatu data baru berdasarkan similaritas dengan labeled data
- Similaritas biasanya memakai metrik jarak
- Satuan jarak umumnya menggunakan euclidian

Nama lain dari NN

- lazy algorithm
- memory-based
- instance-based
- exemplar-based
- case-based
- experience-based

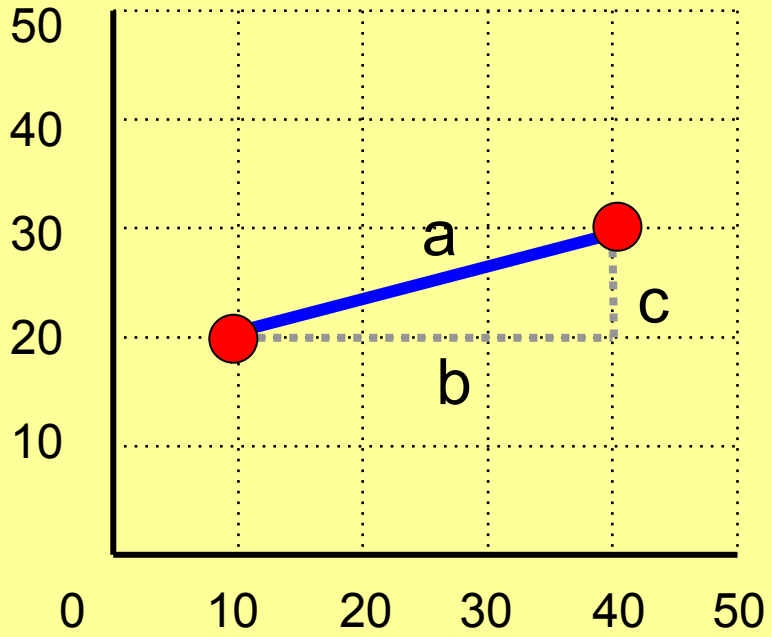
Jenis NN

- 1-NN
  - Pengklasifikasikan dilakukan terhadap 1 labeled data terdekat
- k-NN
  - Pengklasifikasikan dilakukan terhadap k labeled data terdekat
  - $k > 1$

Algoritma 1-NN

- Hitung jarak antara data baru ke setiap labeled data
- Tentukan 1 labeled data yang mempunyai jarak yang paling minimal
- Klasifikasikan data baru ke dalam labeled data tersebut

Penghitungan jarak (Euclidian distance)



$$a^2 = b^2 + c^2$$

$$a = \sqrt{b^2 + c^2}$$

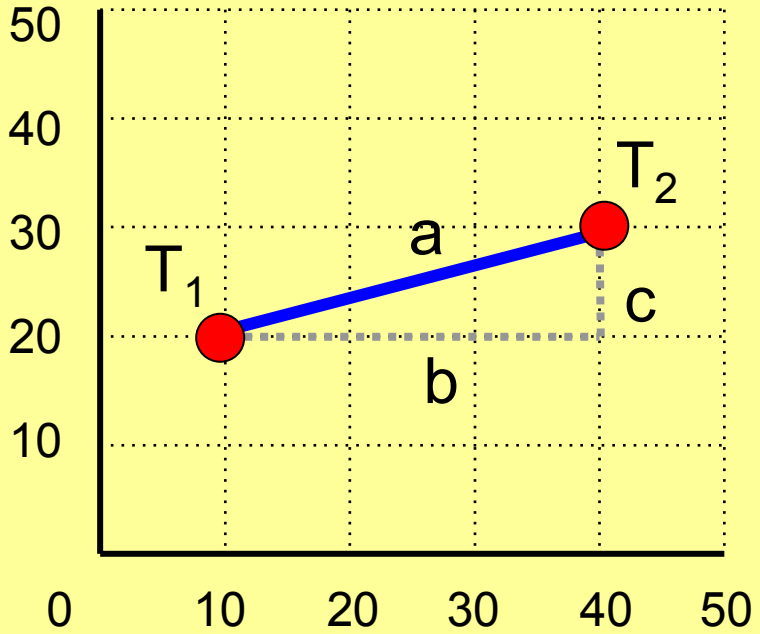
$$= \sqrt{(40-10)^2 + (30-20)^2}$$

$$= \sqrt{(30)^2 + (10)^2}$$

$$= \sqrt{900 + 100}$$

$$= \sqrt{1000} = 31.628$$

Penghitungan jarak dengan vector



$$T_1 = \begin{bmatrix} 10 \\ 20 \end{bmatrix} \qquad T_2 = \begin{bmatrix} 40 \\ 30 \end{bmatrix}$$

$$T = T_2 - T_1 = \begin{bmatrix} 40 \\ 30 \end{bmatrix} - \begin{bmatrix} 10 \\ 20 \end{bmatrix} = \begin{bmatrix} 30 \\ 10 \end{bmatrix}$$

$$D = T' \times T$$

$$= \begin{bmatrix} 30 & 10 \end{bmatrix} \begin{bmatrix} 30 \\ 10 \end{bmatrix} = 900 + 100 = 1000$$

$$a = \sqrt{D} = \sqrt{1000} = 31.628$$

Contoh kasus:

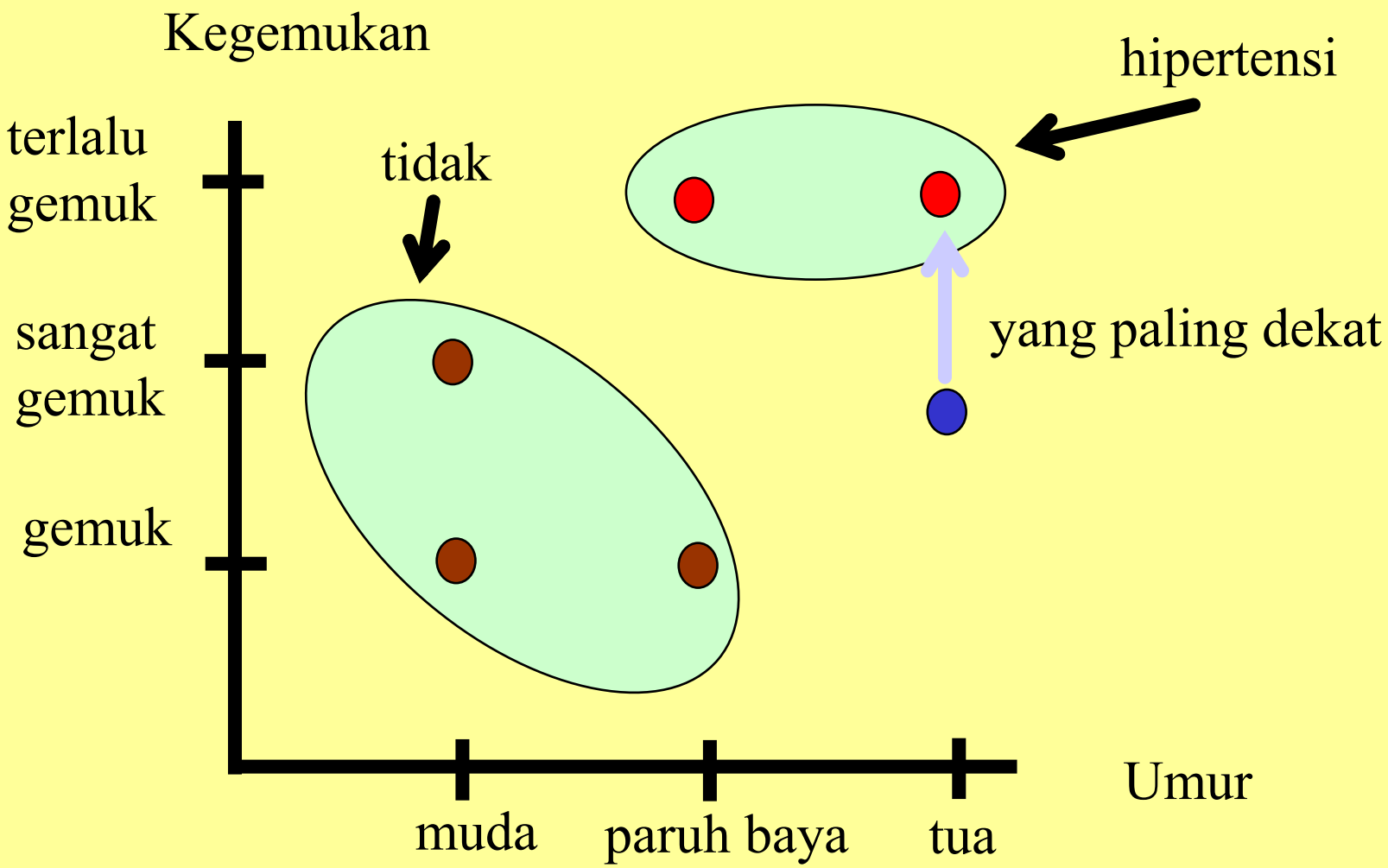
Pengenalan untuk menentukan
seseorang itu mempunyai hipertensi atau tidak

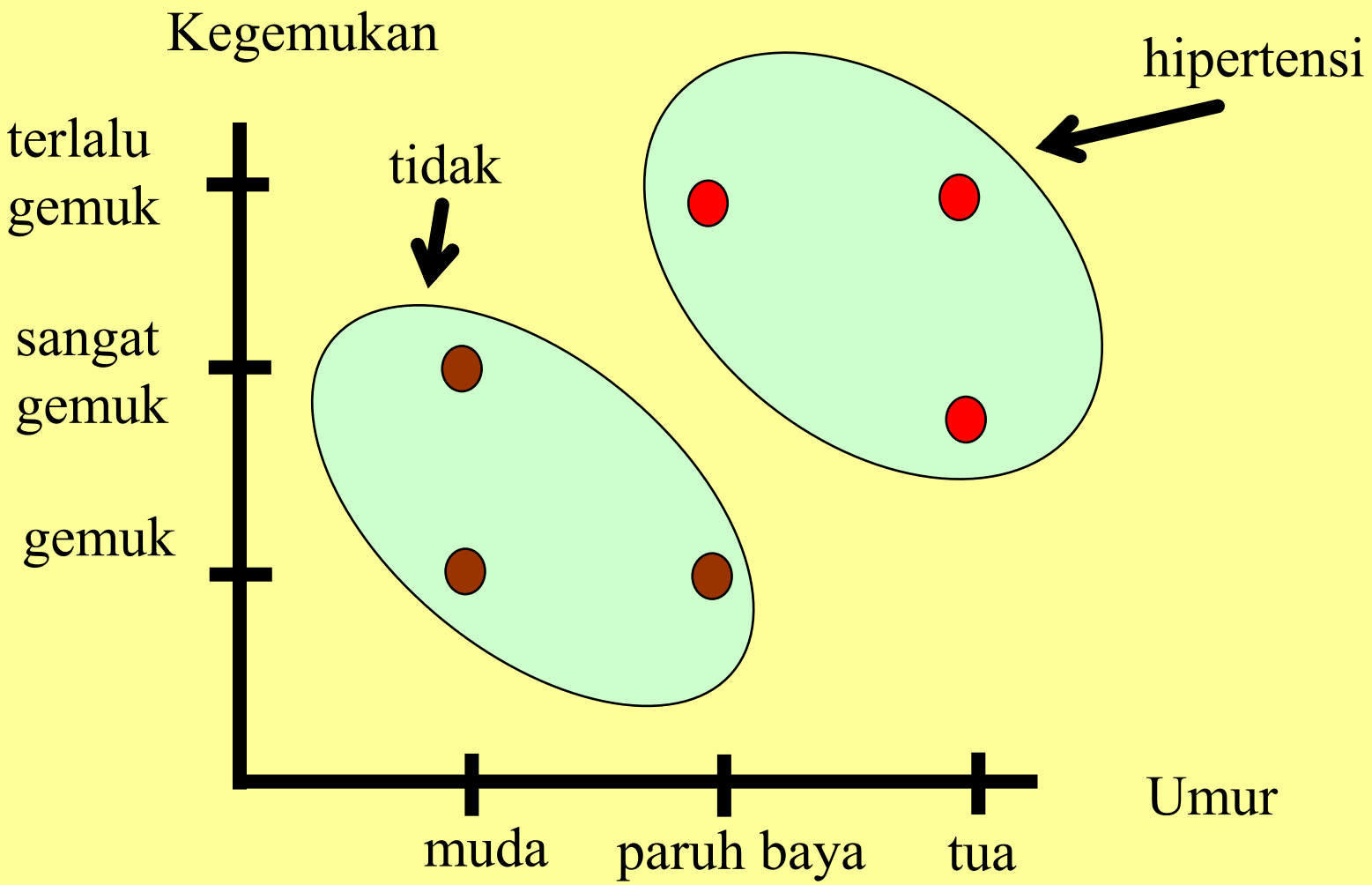
Umur	Kegemukan	Hipertensi
muda	gemuk	Tidak
muda	sangat gemuk	Tidak
paruh baya	gemuk	Tidak
paruh baya	terlalu gemuk	Ya
tua	terlalu gemuk	Ya

tua	sangat gemuk	?

data
baru

Penyelesaian dengan 1-NN

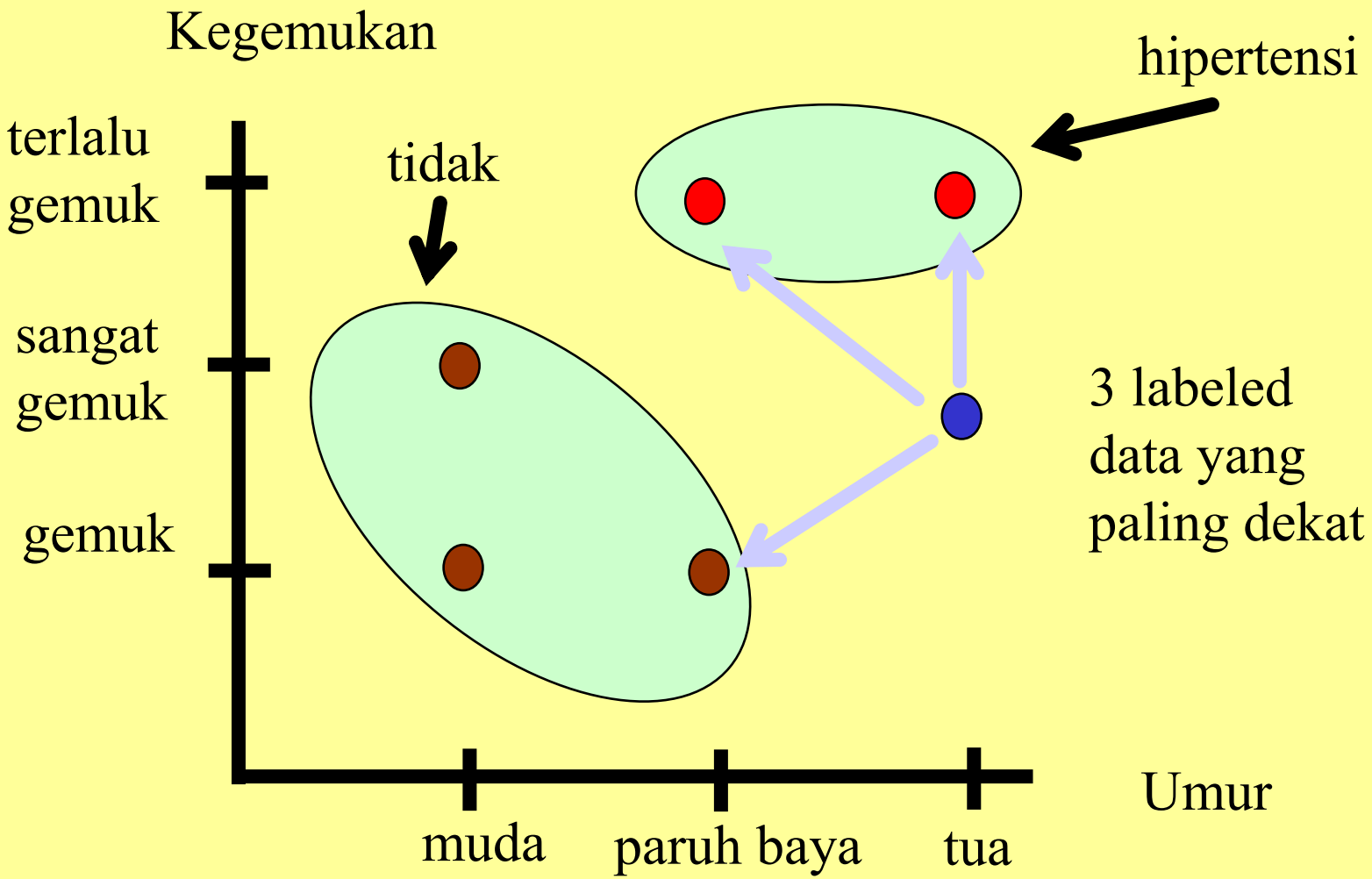


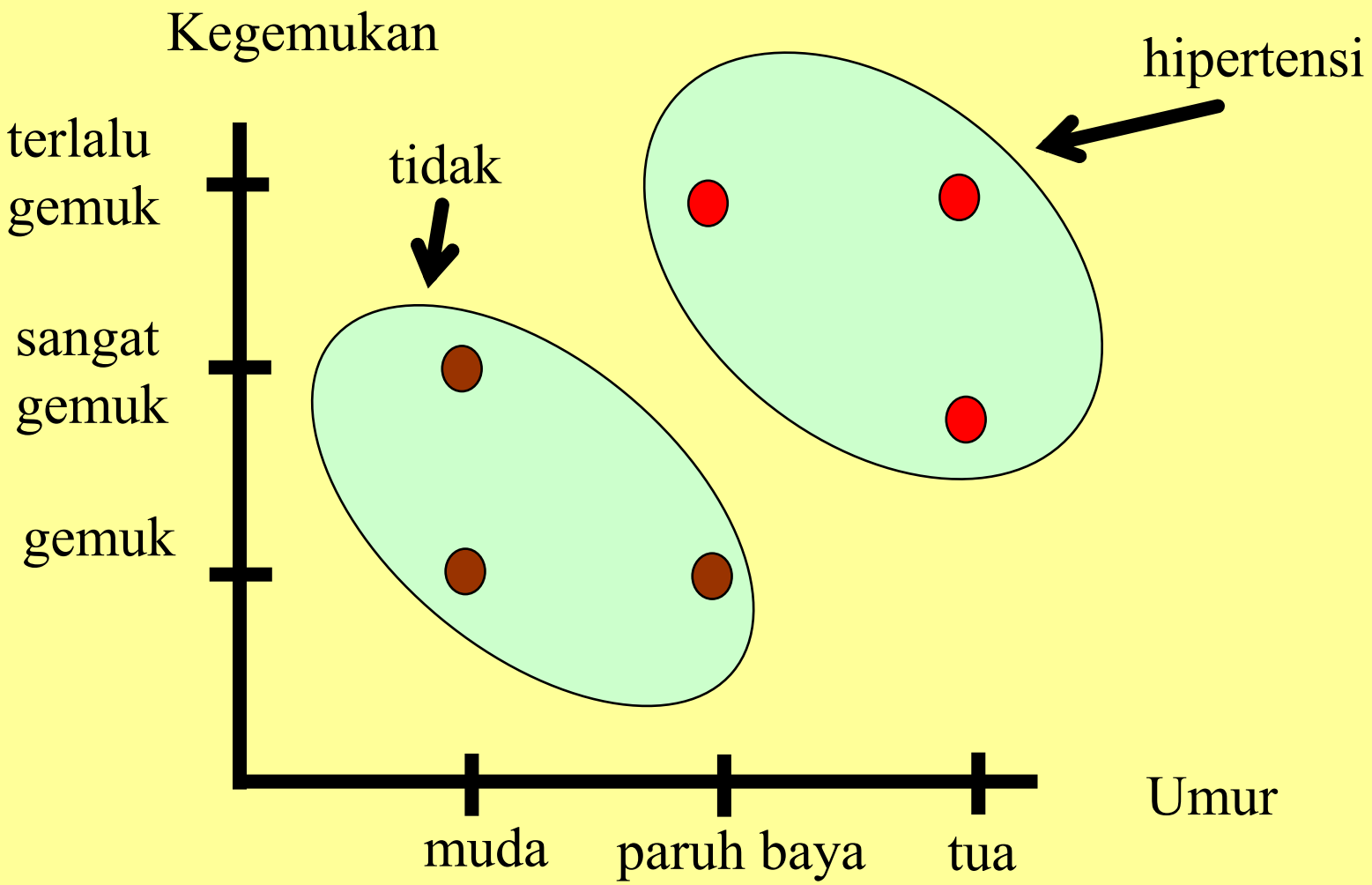

Kegemukan
terlalu gemuk
sangat gemuk
gemuk
tidak
hipertensi
muda
paruh baya
tua
Umur

Algoritma k-NN

- Tentukan k
- Hitung jarak antara data baru ke setiap labeled data
- Tentukan k labeled data yang mempunyai jarak yang paling minimal
- Klasifikasikan data baru ke dalam labeled data yang mayoritas

Penyelesaian dengan k-NN
(misalnya k=3)




Kegemukan
terlalu gemuk
sangat gemuk
gemuk
tidak
hipertensi
muda
paruh baya
tua
Umur

Keuntungan

- Analytically tractable
- Implementasi sangat sederhana
- Memungkinkan parallel implementation

Kelemahan

- Butuh memori besar
- Komputasi besar