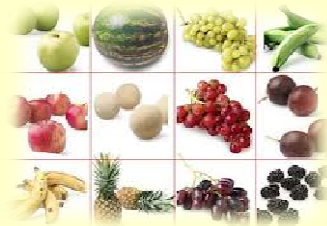




Introduction to Classification



Ali Ridho Barakbah

Knowledge Engineering Research Group
Soft Computing Laboratory
Department of Information and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya



Electronic Engineering
Polytechnic Institute of Surabaya

Ali Ridho Barakbah

Knowledge Engineering
(knoWing) Research Group



Classification: Definition

- Given a collection of records (*training set*)
 - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
 - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Tan, Steinbach, Kumar, *Introduction to Data Mining*



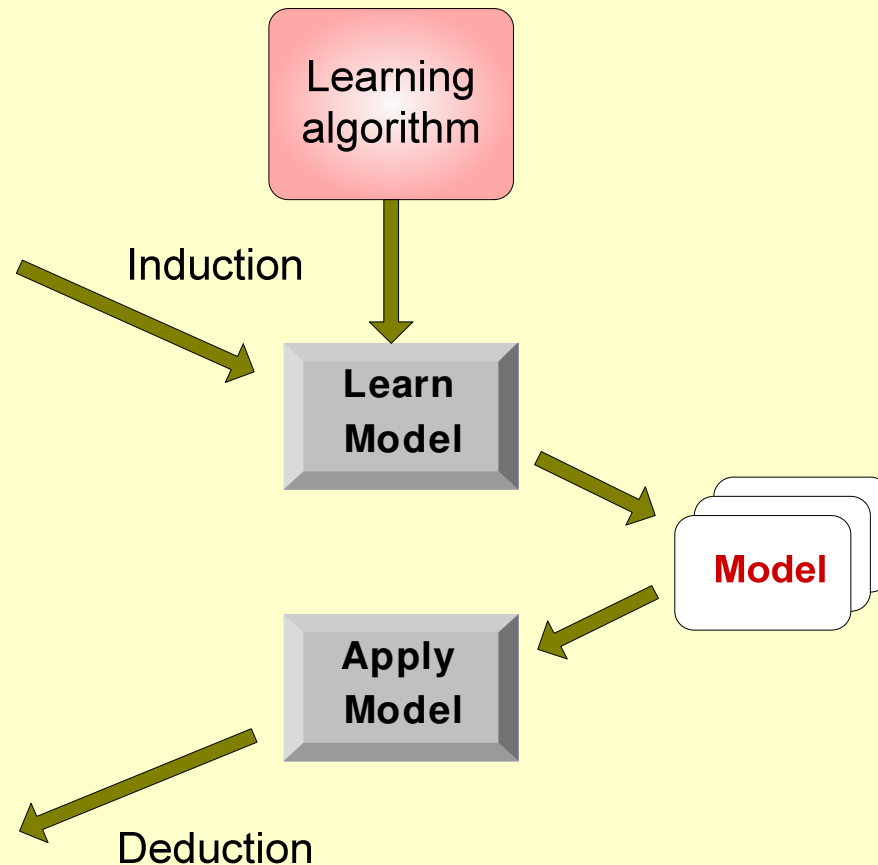
Illustrating Classification Task

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

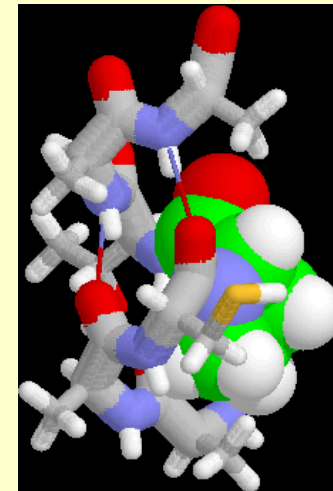
Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Examples of Classification Task

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil
- Categorizing news stories as finance, weather, entertainment, sports, etc



Fase klasifikasi?

- Proses klasifikasi biasanya dibagi menjadi dua fase : learning dan test.
 - Fase learning → sebagian data yang telah diketahui kelas datanya diumpankan untuk membentuk model perkiraan.
 - Fase test → model yang sudah terbentuk diuji dengan sebagian data lainnya untuk mengetahui akurasi dari model tsb.
- Bila akurasi mencukupi model ini dapat dipakai untuk prediksi kelas data yang belum diketahui.
- Klasifikasi dicirikan dengan data training mempunyai label, berdasarkan label ini proses klasifikasi memperoleh pola attribut dari suatu data.

Ide Mesin Pembelajaran

Fakta harian dalam 6 hari dan keputusan untuk berolah-raga sebagai berikut:

#	Cuaca	Temperatur	Kecepatan Angin	Berolah-raga
1	Cerah	Normal	Pelan	Ya
2	Cerah	Normal	Pelan	Ya
3	Hujan	Tinggi	Pelan	Tidak
4	Cerah	Normal	Kencang	Ya
5	Hujan	Tinggi	Kencang	Tidak
6	Cerah	Normal	Pelan	Ya

(1) Ketika cuaca cerah, apakah akan berolah-raga?

(2) Ketika cuaca cerah dan temperatur normal, apakah akan berolah-raga?

Penyajian keputusan berdasarkan fakta
iniilah yang mengilhami konsep dari mesin
pembelajaran

Data Training



Key	Attribut			Target
Day	Cuaca	Temperatur	Kecepatan Angin	Berolah-raga
D1	Cerah	Normal	Pelan	Ya
D2	Cerah	Normal	Pelan	Ya
D3	Hujan	Tinggi	Pelan	Tidak
D4	Cerah	Normal	Kencang	Ya
D5	Hujan	Tinggi	Kencang	Tidak
D6	Cerah	Normal	Pelan	Ya

- Attribut adalah kolom data, ada atribut dan target
- Instance adalah isi dari attribut sebagai contoh atribut cuaca mempunyai instance “cerah” dan “hujan”, sering ditulis dengan cuaca={cerah,hujan}
- Record/tuple adalah baris data

Ide Mesin Pembelajaran

Pada dasarnya semua algoritma yang dikembangkan dalam mesin pembelajaran adalah algoritma yang menghasilkan hipotesa dari suatu keputusan berdasarkan data pembelajaran yang diberikan.



Fact

Data	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

Problem description

<?, Cold, High, ?, ?, ?>

Data 3

No

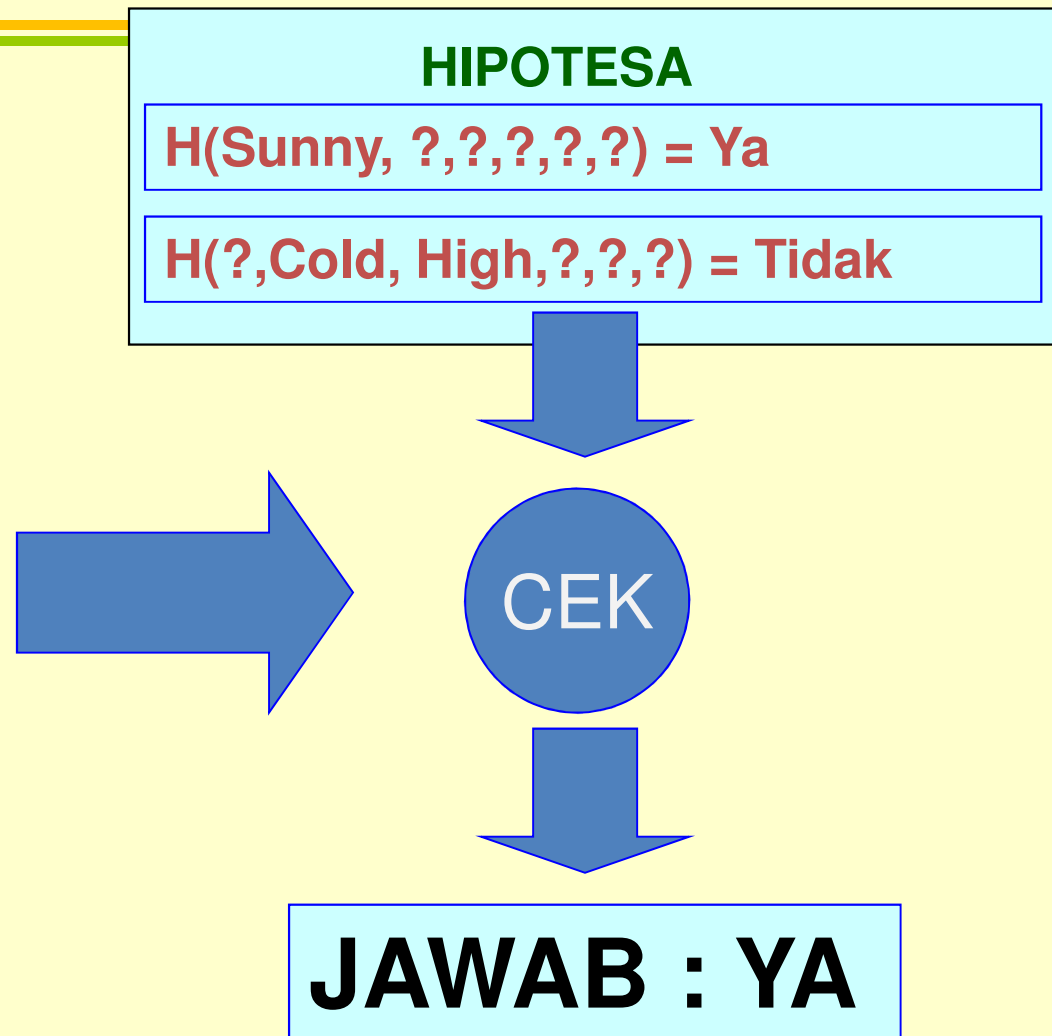
<Sunny, Warm, ?, ?, ?, ?>

Data 1, 2, 4

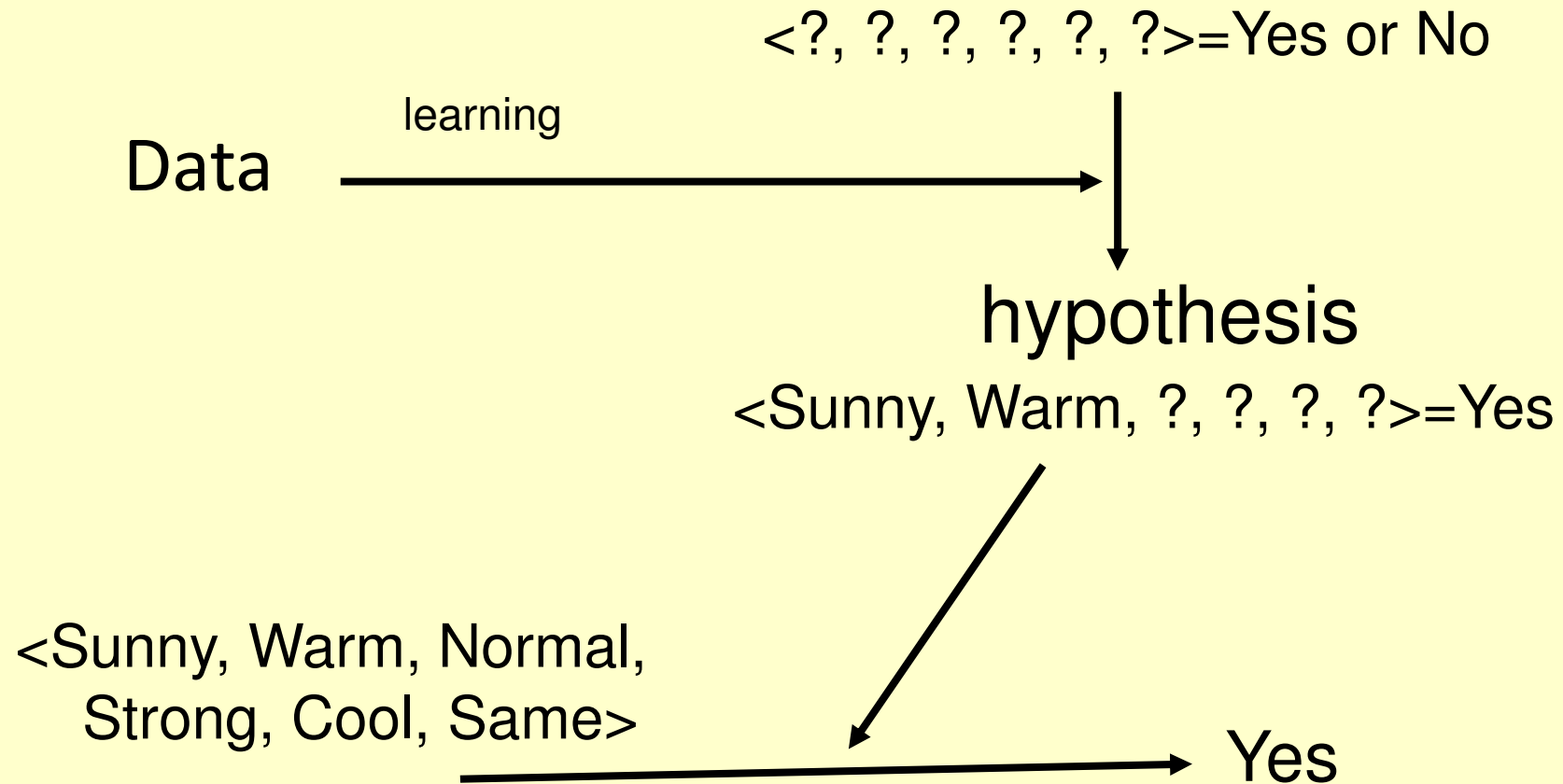
Yes

**Our human brain can answer these questions.
But how the machine can answer?**

Contoh Keputusan Dari Hipotesa



Learning Process



Klasifikasi dengan Find-S

- Find-S adalah suatu metode paling sederhana yang dapat digunakan untuk mendapatkan suatu hipotesa berdasarkan data.
- Find-S mencari kesamaan nilai atribut untuk memperoleh suatu hipotesa
- Kelemahan dari Find-S adalah data yang digunakan harus bersifat konsisten dan tidak bias ??? (Terlalu sulit untuk dapat memperoleh data semacam ini pada persoalan nyata)

Find-S

< ϕ , ϕ , ϕ , ϕ , ϕ , ϕ >

<Sunny, Warm, Normal, Strong, Warm, Same> \longrightarrow <Sunny, Warm, Normal, Strong, Warm, Same>

<Sunny, Warm, High, Strong, Warm, Same> \longrightarrow <Sunny, Warm, ?, Strong, Warm, Same>

<Sunny, Warm, High, Strong, Cool, Change> \longrightarrow <Sunny, Warm, ?, Strong, ?, ? >

Kelebihan dan Kelemahan Find-S

- Advantage
 - Very simple
- Disadvantage
 - Ignores the negative data

Candidate-Elimination

$$S_0 \quad \boxed{\langle \phi, \phi, \phi, \phi, \phi, \phi \rangle}$$

?

$$G_0 \quad \boxed{\langle ?, ?, ?, ?, ?, ? \rangle}$$

Candidate-Elimination

S_0 $\langle \phi, \phi, \phi, \phi, \phi, \phi \rangle$



S_1 $\langle \text{Sunny, Warm, Normal, Strong, Warm, Same} \rangle$

$\langle \text{Sunny, Warm, Normal, Strong, Warm, Same} \rangle$
=Yes

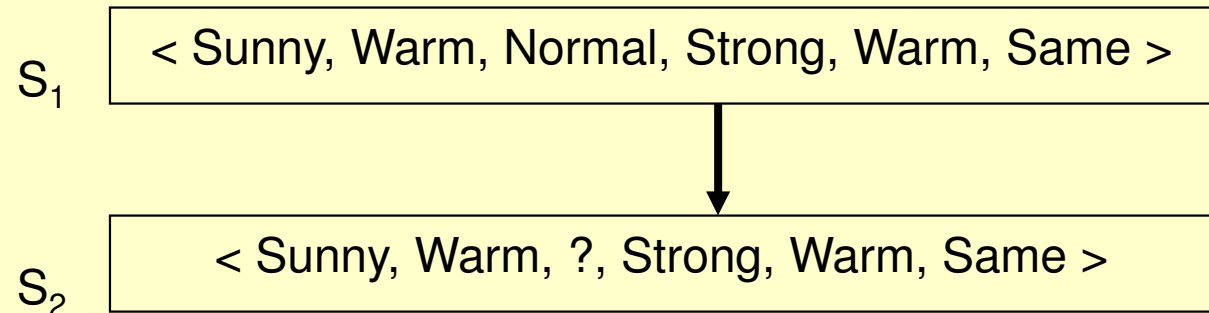
?

G_1 $\langle ?, ?, ?, ?, ?, ? \rangle$



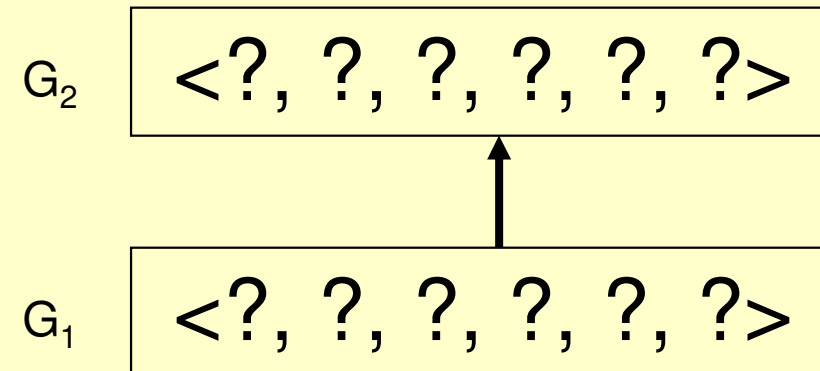
G_0 $\langle ?, ?, ?, ?, ?, ? \rangle$

Candidate-Elimination

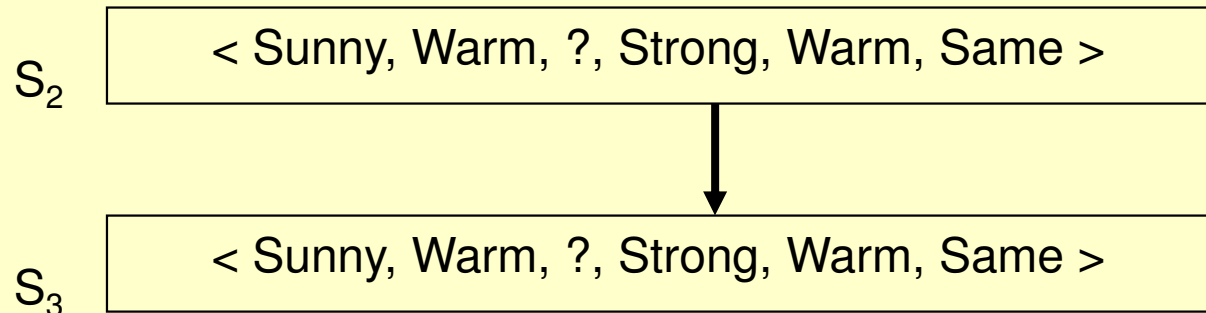


$\langle \text{Sunny, Warm, High, Strong, Warm, Same} \rangle$
=Yes

?

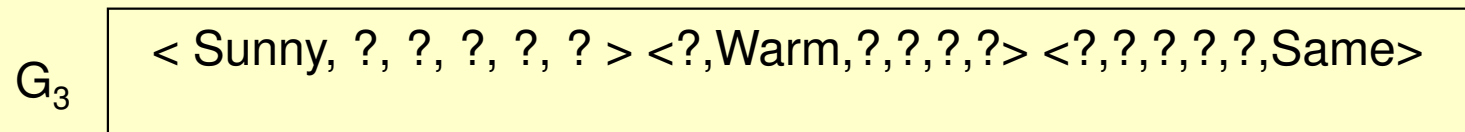


Candidate-Elimination

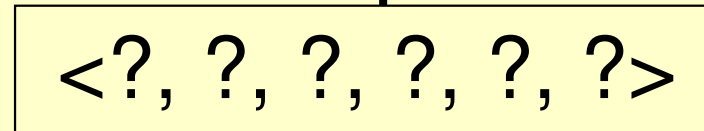


$\langle \text{Rainy, Cold, High, Strong, Warm, Change} \rangle$
=No

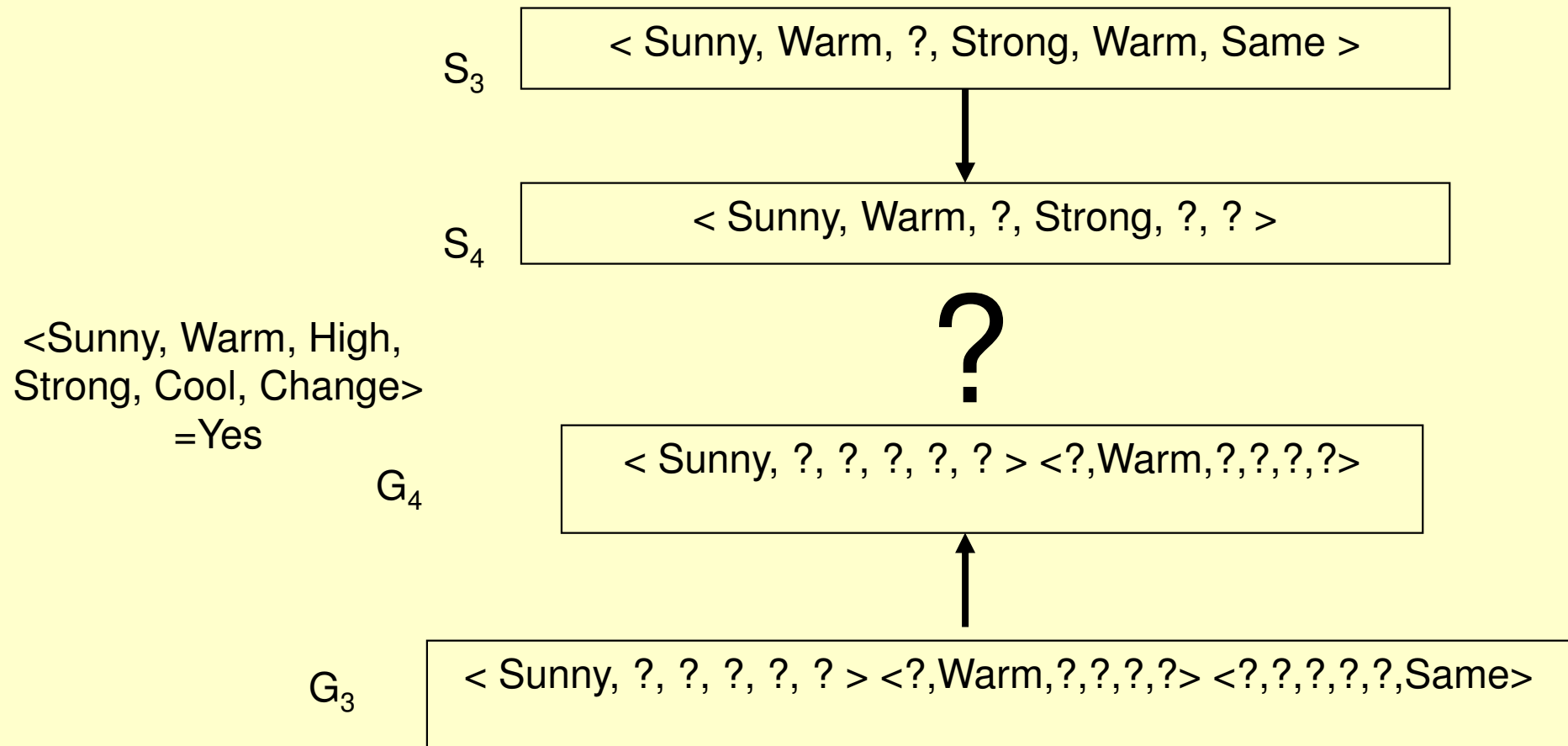
?



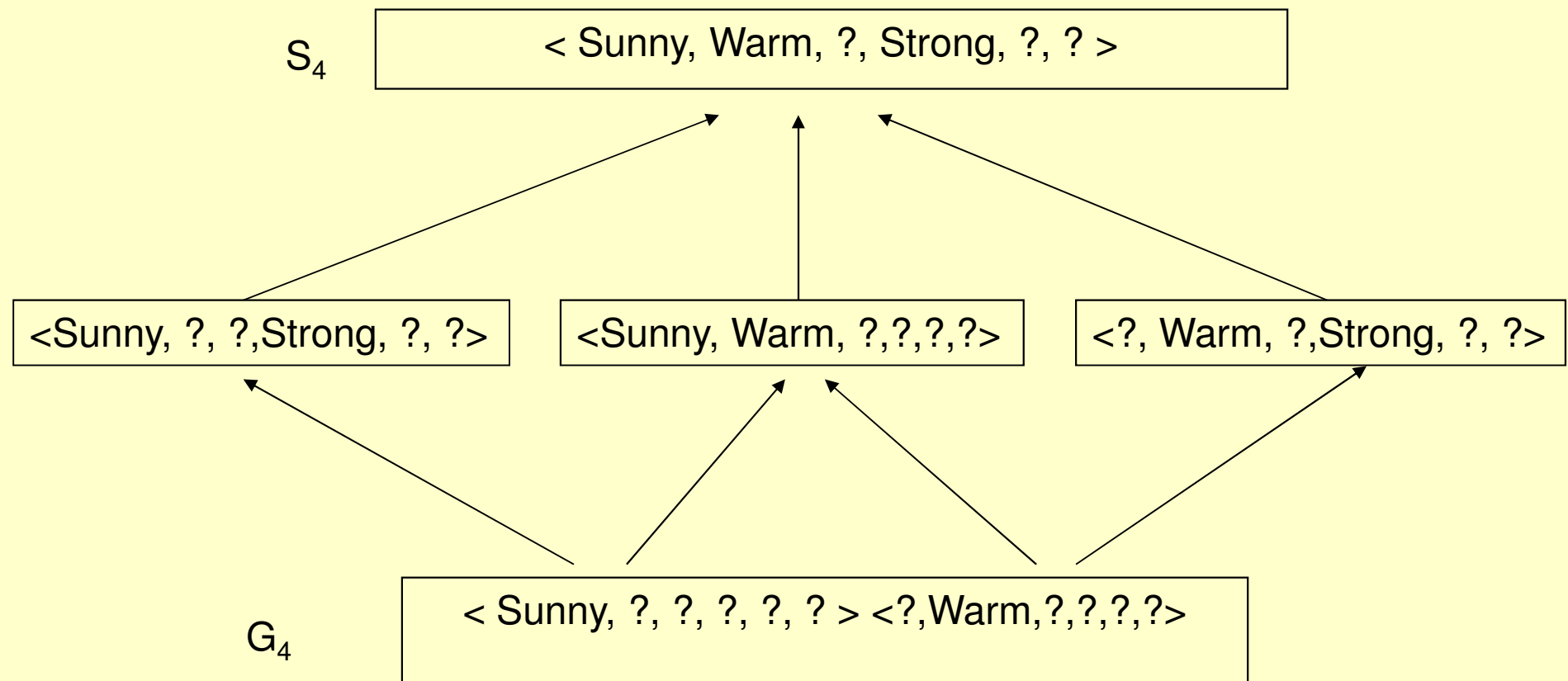
G_2



Candidate-Elimination



Candidate-Elimination



-
- Advantage
 - Consider the negative data to strengthen the hypothesis
 - Disadvantage
 - If the data is not consistent, S and G can not match
 - Difficult to implement in the programming

Klasifikasi dengan 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
 - Pengklasifikasian dilakukan terhadap 1 labeled data terdekat
- k-NN
 - Pengklasifikasian 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

Contoh kasus 2:

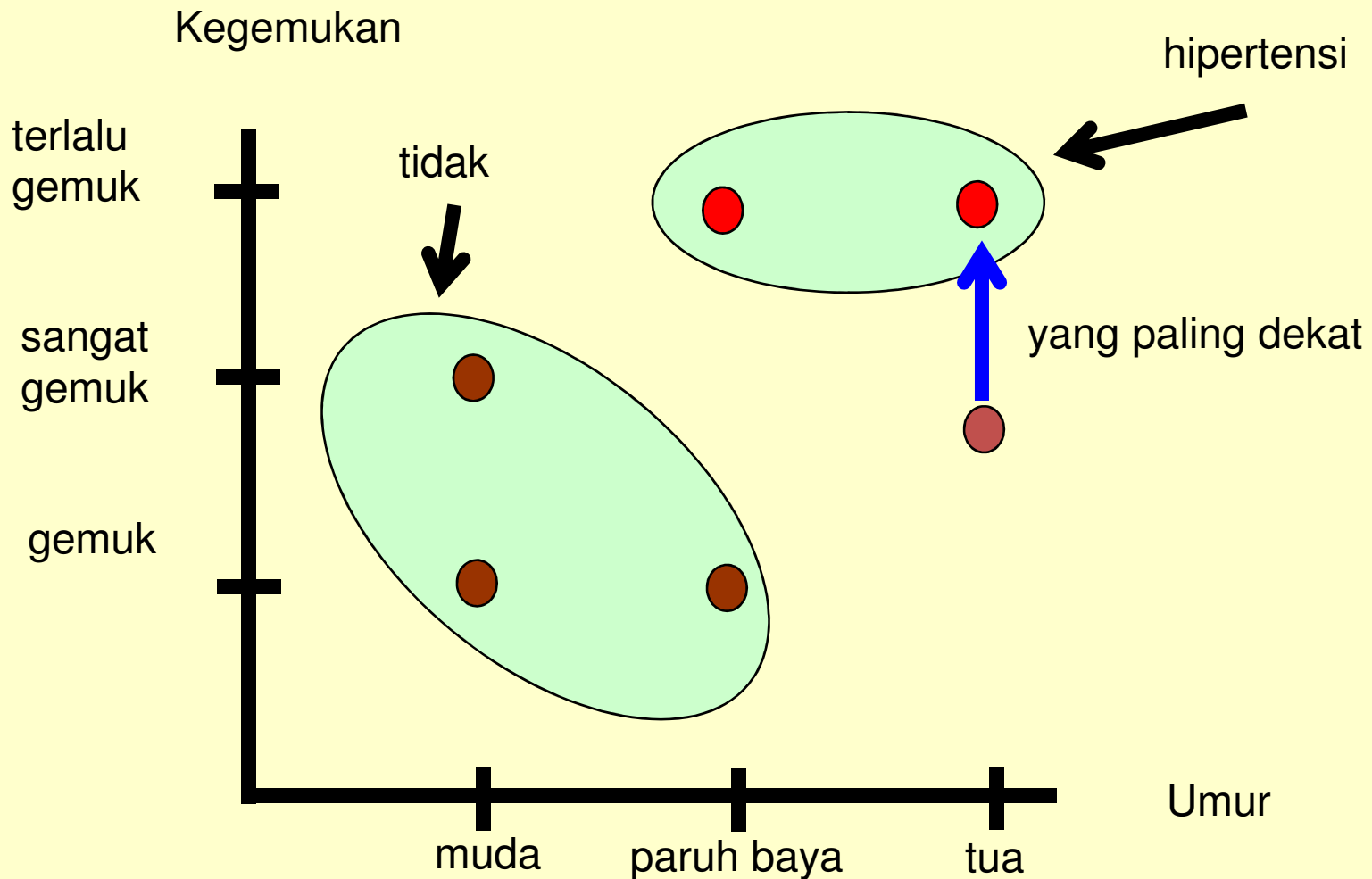
Pengenalan untuk menentukan
seseorang itu mempunyai hipertensi atau tidak

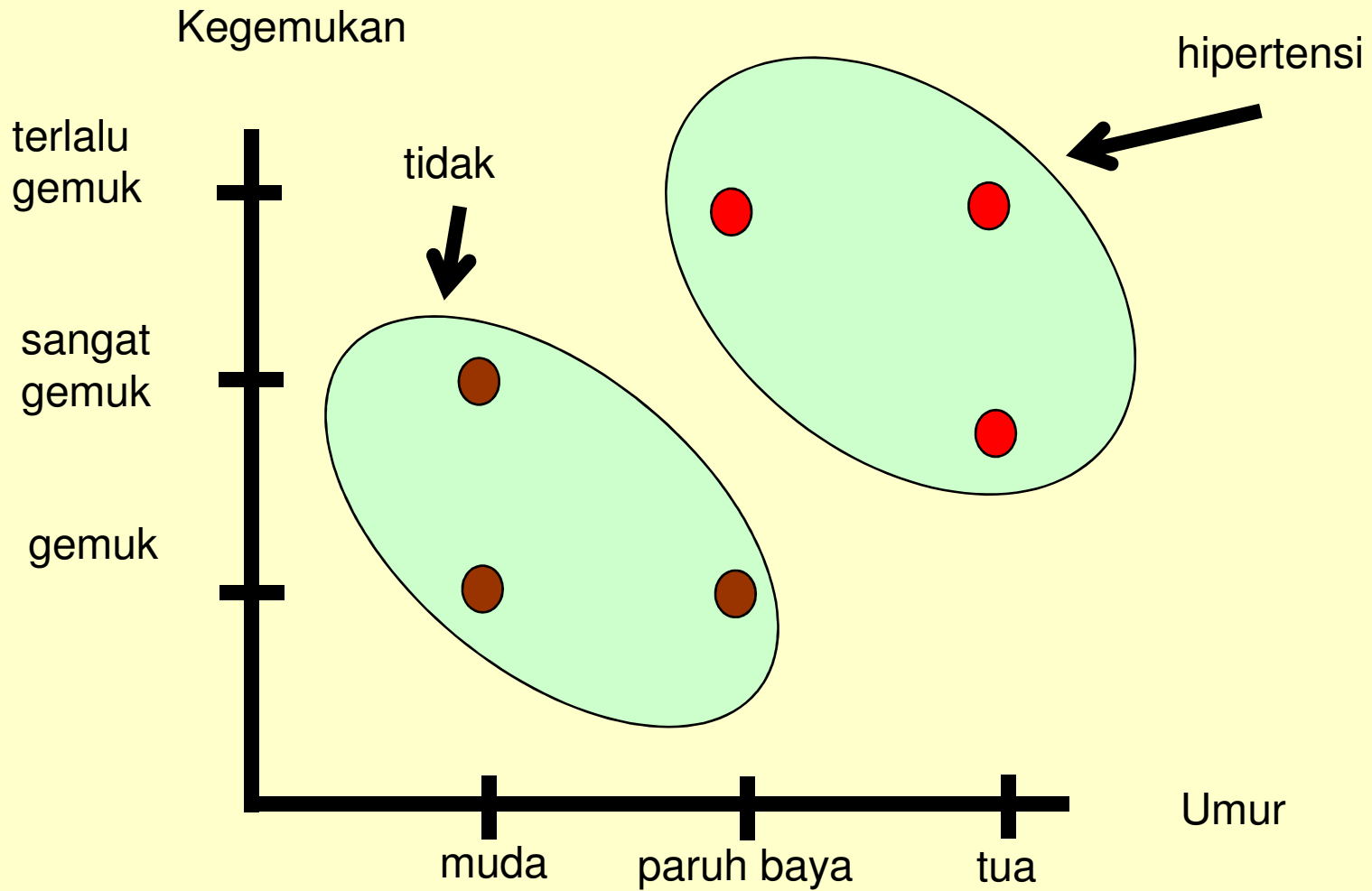
Umur	Kegemukan	Hipertensi
muda	gemuk	Tidak
muda	sangat gemuk	Tidak
paruh baya	gemuk	Tidak
paruh baya	terlalu gemuk	Ya
tua	terlalu gemuk	Ya
tua	sangat gemuk	?

data
baru



Penyelesaian dengan 1-NN

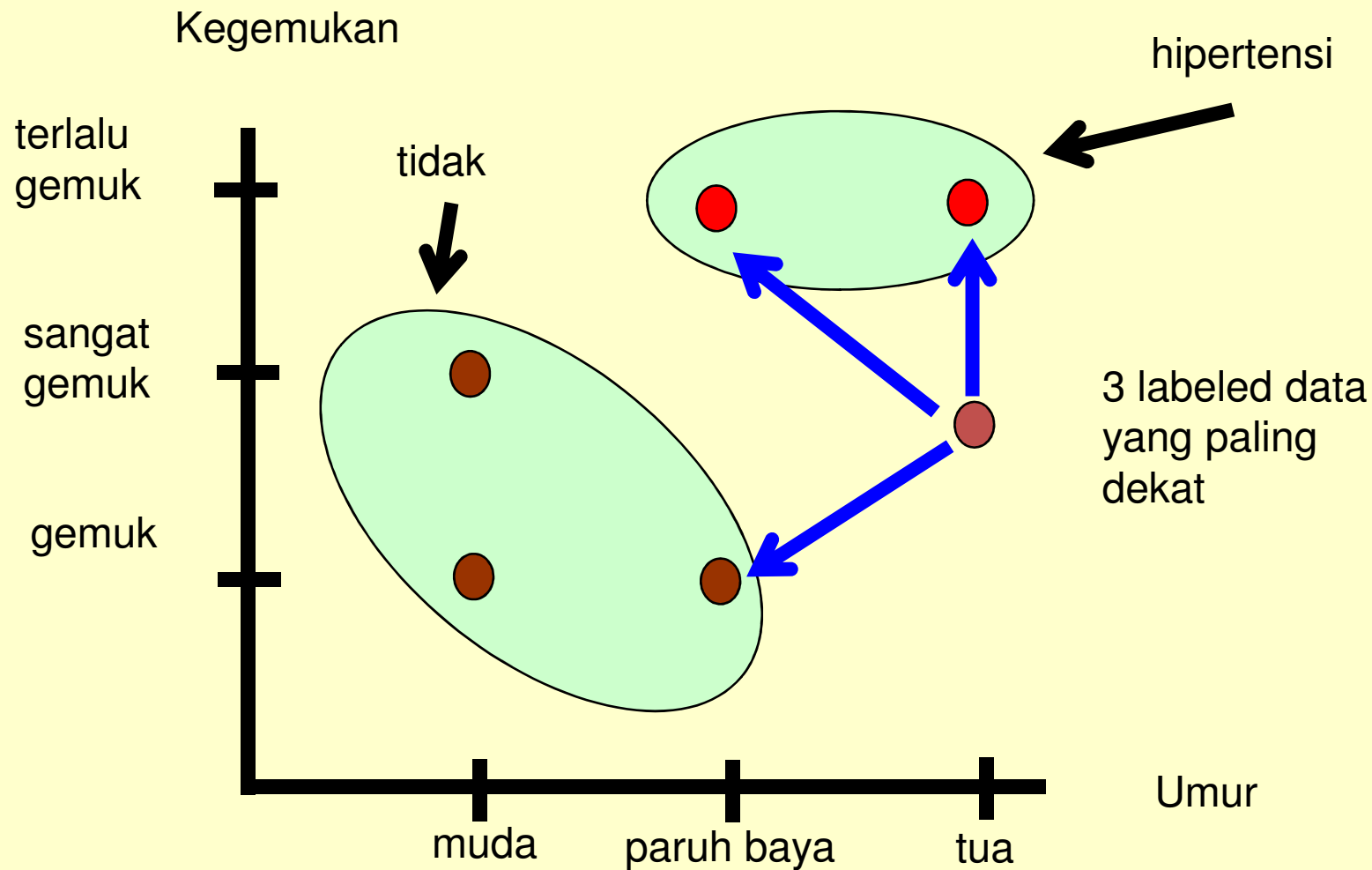


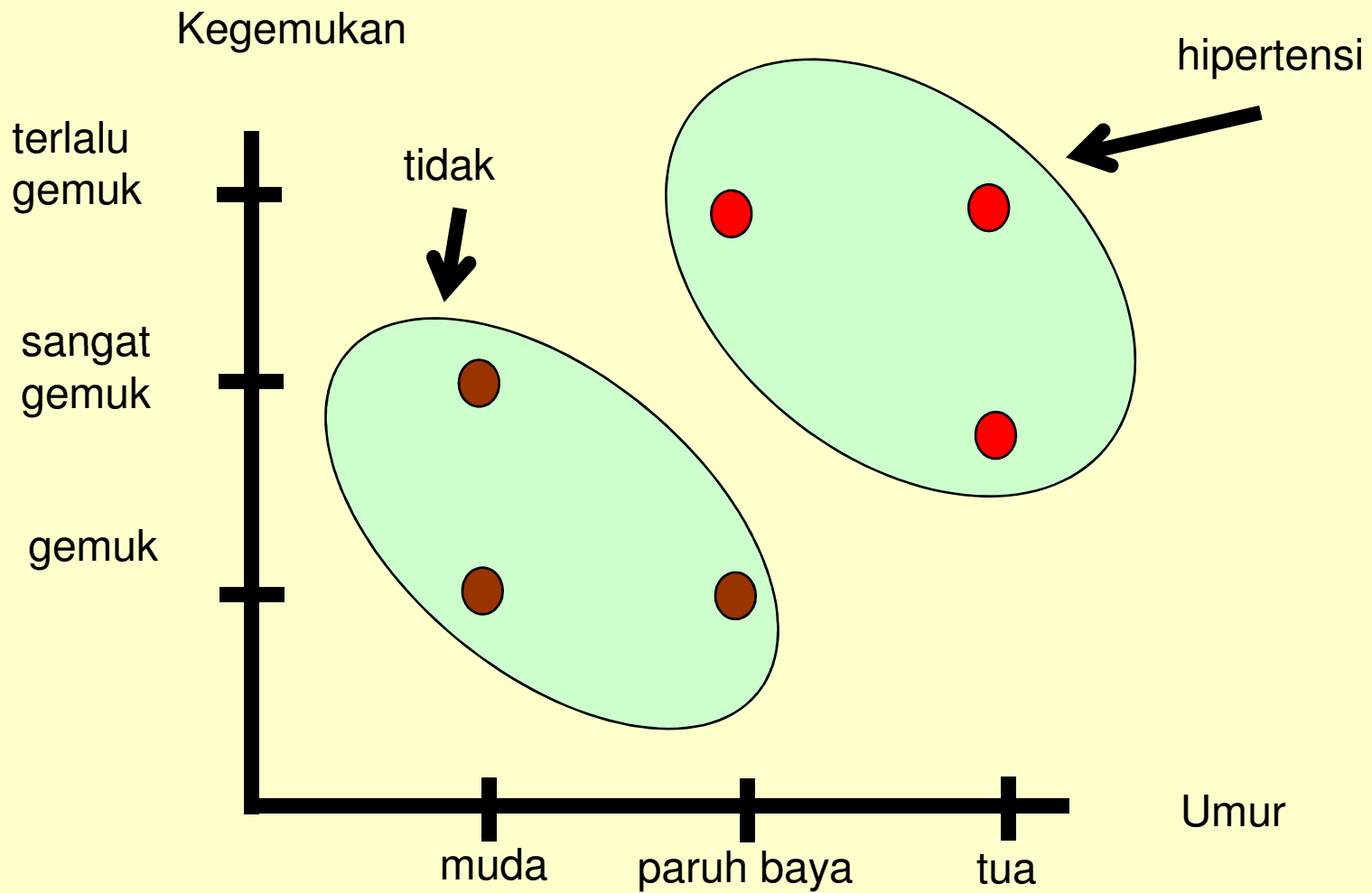


Algoritma k-NN

- 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)





Keuntungan

- Analytically tractable
- Implementasi sangat sederhana
- Tingkat error $>$ bayesian, $<$ $2 \times$ bayesian
- Memungkinkan parallel implementation

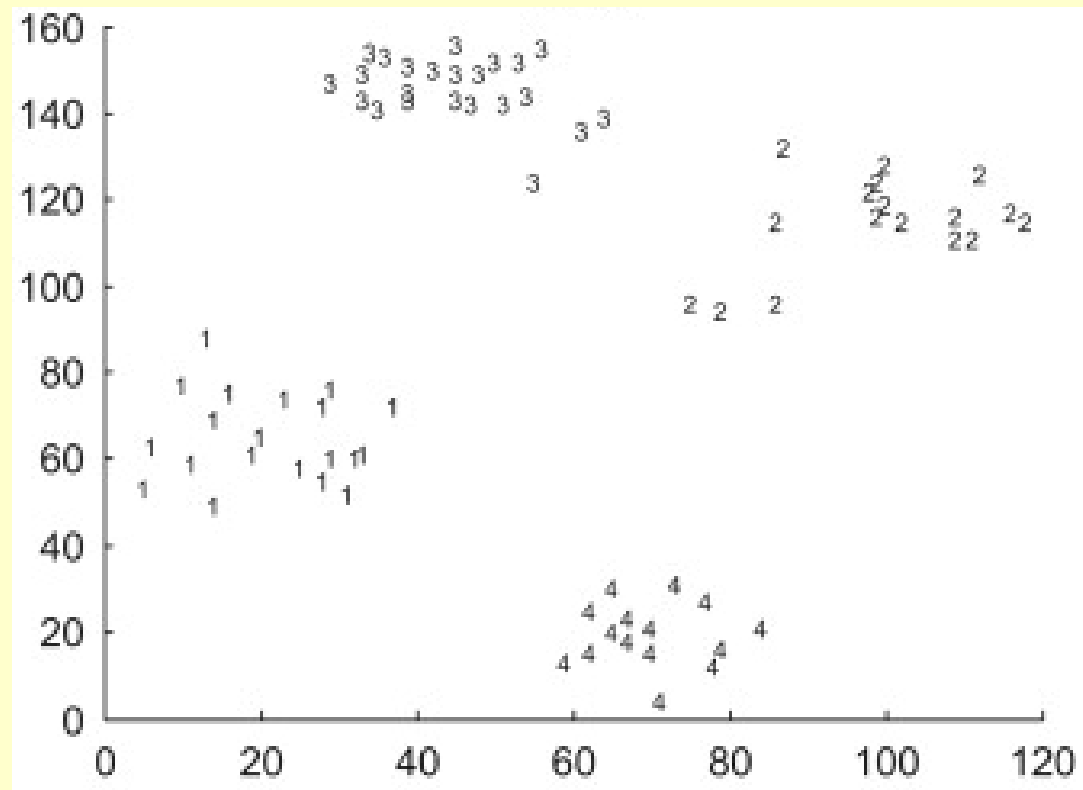
Kelemahan

- Butuh memori besar
- Komputasi besar

Tugas

- Case : Classification with Ruspini Dataset
- Represents a simple, well-known example that is commonly used as a benchmark problem in evaluating classification and clustering methods and is widely available, incorporated as a built-in data object in both R and S-plus statistics packages.
- Number of attributes : 2
- Number of data: 75
- Number of classes : 4
 - Class 1 → 20 data
 - Class 2 → 17 data
 - Class 3 → 23 data
 - Class 4 → 15 data

Ruspini Dataset



Task

- Ambillah 80% data pertama pada masing-masing class sebagai training data.
- Pakailah 20% data sisanya pada masing-masing class sebagai data untuk uji coba
- Lakukan klasifikasi masing-masing data uji coba dan bandingkan hasilnya pada hasil sesungguhnya.
- Catatlah berapa jumlah kesalahan yang terjadi pada semua data uji coba (dalam persen).
- Buat laporan

-
- Lakukan percobaan dengan melibatkan beberapa metode klasifikasi:
 - 1-NN
 - 3-NN
 - 5-NN

References

- Tom Michael, Machine Learning, McGraw-Hill publisher, 1997.
- Ali Ridho Barakbah, *Machine Learning*, Lecture Handout, Electronic Engineering Polytechnic Institute of Surabaya.

