

# Sebelumnya...

---

- Sistem Pakar berbasis Logika Fuzzy

# Kecerdasan Buatan

Pertemuan 10

## Pembelajaran Mesin (*Machine Learning*)

Kelas 10-S1TI-03, 04, 05

Husni

[Lunix96@gmail.com](mailto:Lunix96@gmail.com)

<http://Komputasi.wordpress.com>

# *Outline*

---

- Pembelajaran Mesin
- Arsitektur Sistem
- Pohon Keputusan
- Klasifikasi dengan ID3
- Klasifikasi dengan Bayes

# Pembelajaran Mesin

---

- Kajian mengenai pembuatan program komputer yang secara otomatis meningkatkan atau menyesuaikan kinerjanya melalui pengalaman.
- *“programming by example”*
- Tujuan: menghasilkan algoritma pembelajaran yang belajar secara otomatis tanpa bantuan manusia
- Tidak menyelesaikan masalah secara langsung, program mencari metode terbaik dengan mempelajari contoh yang disediakan.

# Kesulitan Membangun Expert System

---

- Tidak ada pakar (manusia) untuk masalah yang harus diselesaikan
- Ada pakar, tetapi tidak mampu menjelaskan kepakarannya.
- Masalah bersifat dinamis dimana fenomena berubah dengan cepat
- Aplikasi perlu disesuaikan/dipersonalkan untuk setiap pengguna komputer secara terpisah.

# Terapan Pembelajaran Mesin

---

- Pengenalan huruf dan angka (*character*)
- Pengenalan tulisan tangan (*handwriting*)
- Deteksi wajah
- Penyeringan spam
- Pengenalan suara
- Pemahaman bahasa pembicaraan
- Prediksi pasar saham
- Prediksi cuaca
- Diagnosa medis
- Deteksi penipuan
- Pencocokan sidik jari (*fingerprint*)

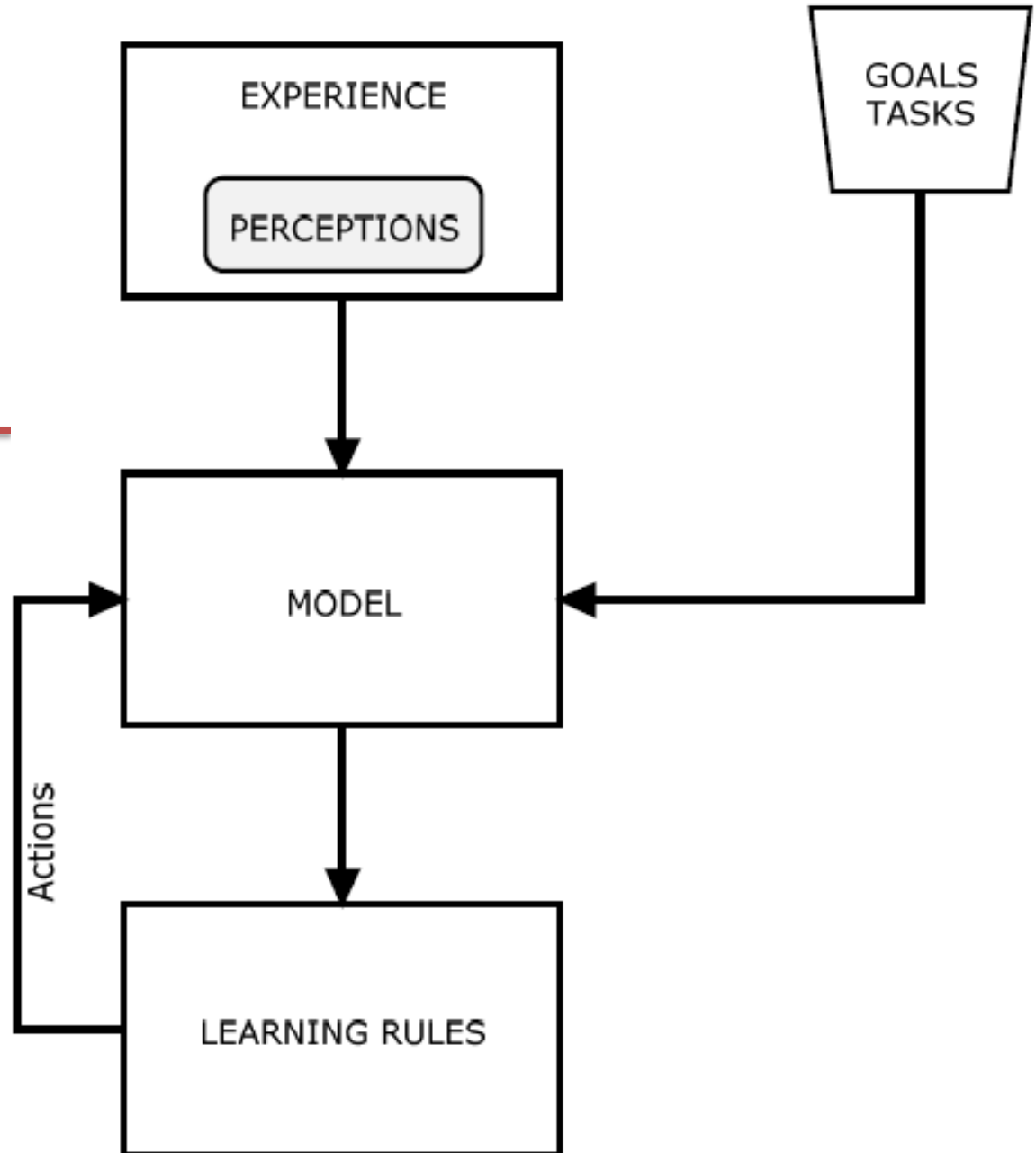
# Komponen Pembelajaran Mesin

---

- **Goal**: tugas yang harus dikerjakan oleh sistem;
- **Model**: fungsi matematika yang memetakan Persepsi (*perception*) menjadi tindakan (*action*);
- **Learning rules**: aturan-aturan untuk mencari model terbaik. Dapat mengupdate parameter model dengan pengalaman baru;
- **Experience**: Himpunan persepsi (dan tindakan-tindakan bersesuaian).

# Arsitektur (Sederhana)

---





# Terminologi

---

- *Sample* (Contoh atau *instance*) adalah obyek yang akan diklasifikasi.
- *Sample* dideskripsikan oleh **atribut-atribut**, dikenal pula sebagai *fitur* atau *variabel*.
- *Label* adalah nama kategori atau kelompok hasil prediksi.
- Selama *training*, algoritma pembelajaran diberikan *sample* berlabel, dan saat pengujian (*test*) disediakan *sample* tak-berlabel.
- **Apakah algoritma dapat memberikan label yang tepat untuk data tak-berlabel?**

# Langkah-Langkah Pembelajaran

---

- Menyiapkan data dan asumsi-asumsi.
- Representasi
- Metode dan estimasi
- Evaluasi
- Pemilihan model

# Klasifikasi Sistem Pembelajaran

---

- Berdasarkan **tujuan** atau **fungsi**: Prediksi, Regresi, klasifikasi, clusterisasi, perencanaan.
- Berdasarkan **model**: Decision trees, Linear separators (perceptron model), Neural networks, Genetic programming, Evolutionary algorithms, Graphical models, Support vector machines, Hidden Markov models
- Berdasarkan **Learning Rules**: gradient descent, least square error, expectation maximization, margin maximization.
- Berdasarkan **Pengalaman**: *Supervised, Unsupervised, Active, Reinforcement learning.*

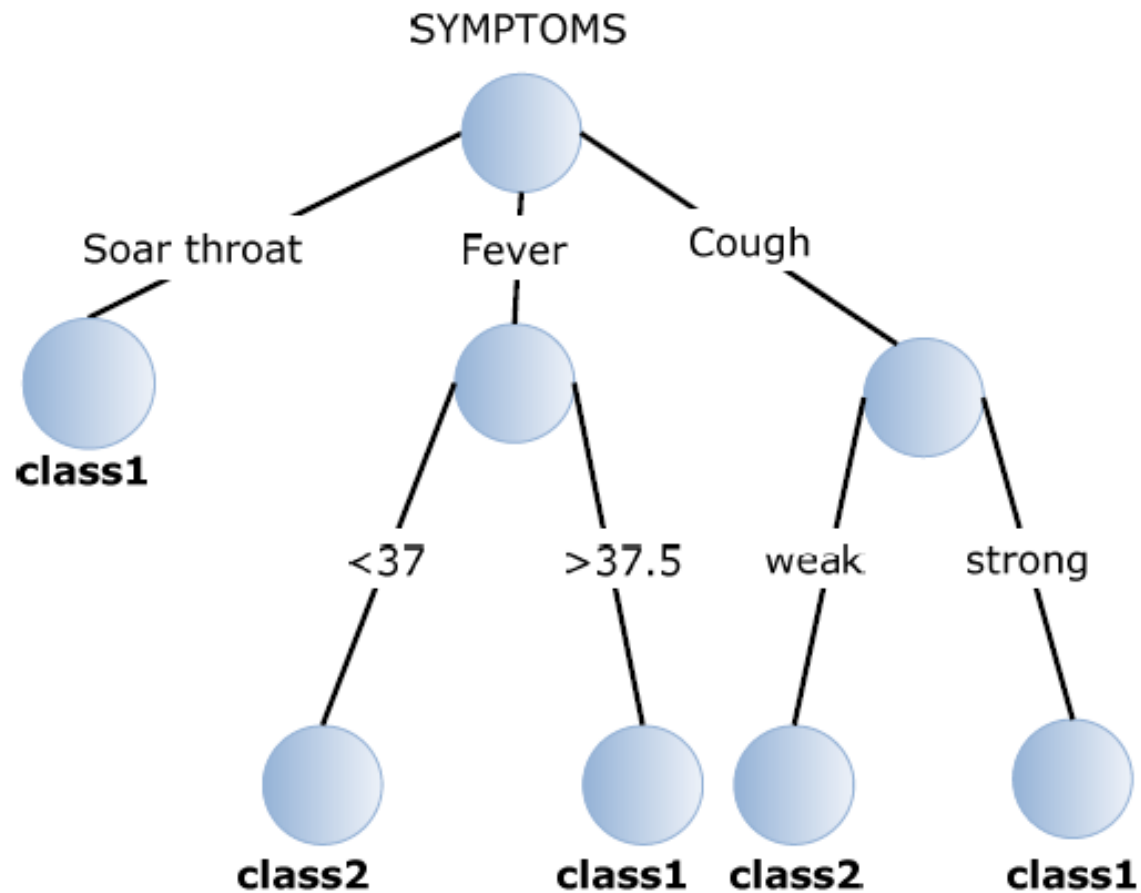
# *Decision Tree*

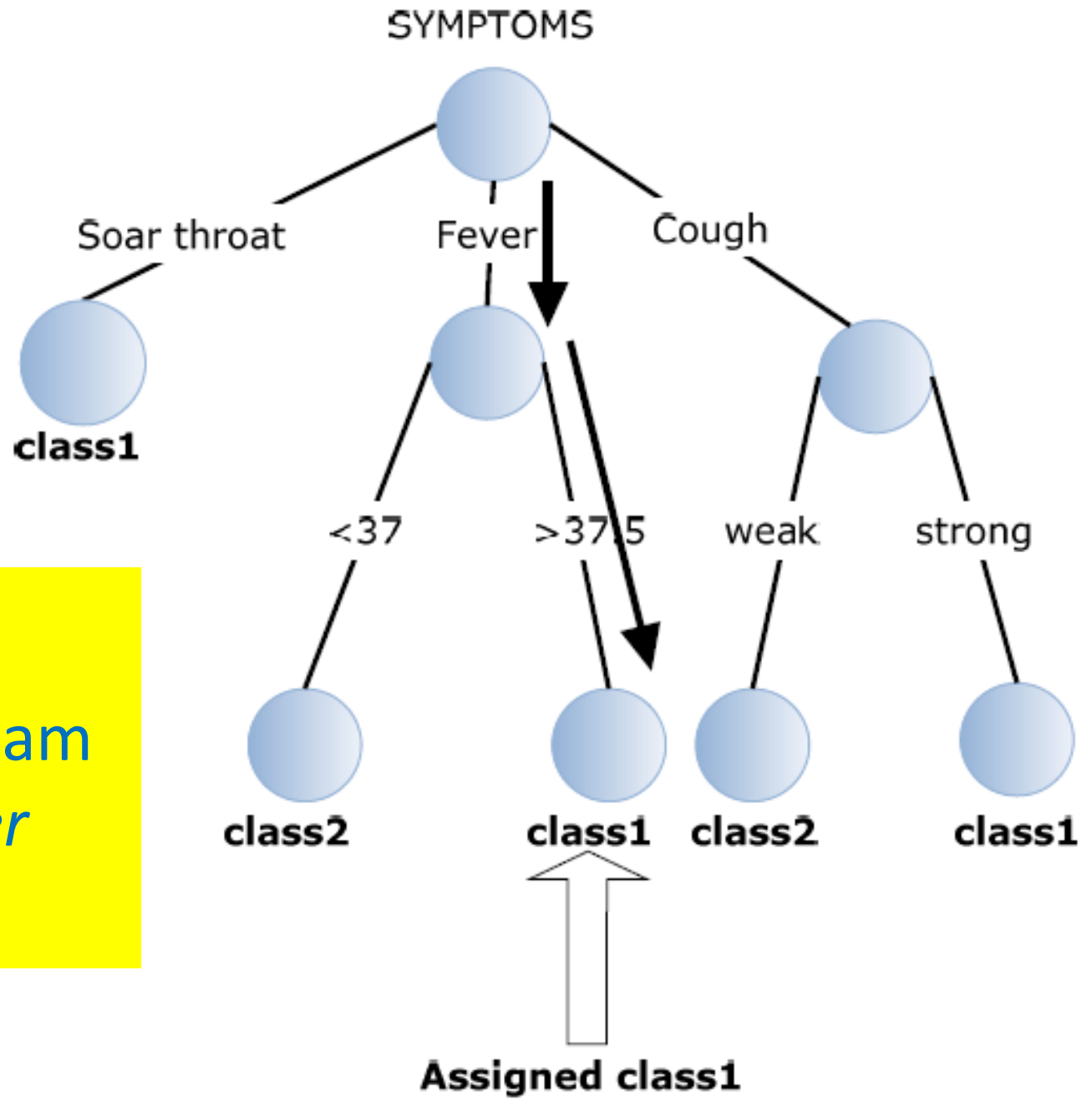
---

- Pembelajaran *Decision Tree* (DT) adalah metode memperkirakan fungsi target bernilai diskret, dimana fungsi tersebut disajikan sebagai pohon keputusan.
- *Representasi Pohon Keputusan:*
  - **Setiap Node internal menguji suatu atribut**
  - **Setiap cabang terkait dengan nilai atribut**
  - **Setiap Node daun menunjukkan suatu klasifikasi**

# Contoh *Tree*/Pohon

- Flue (Class1) atau Tidak (Class2)?

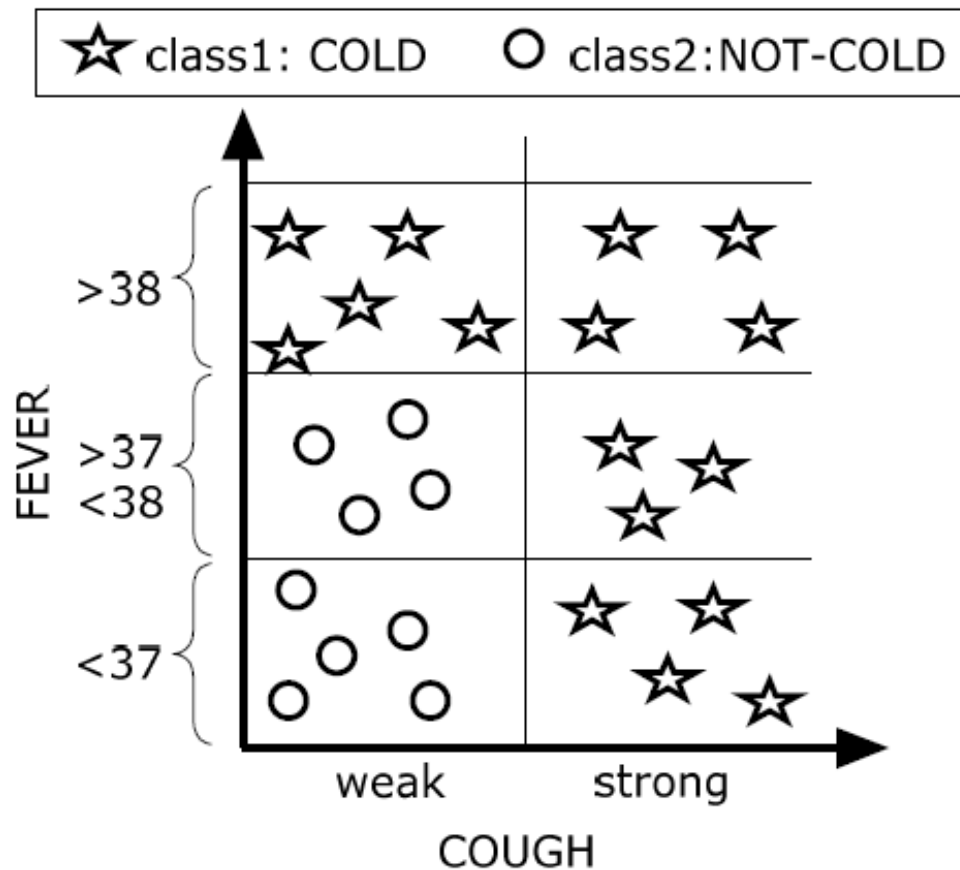




**Test:**  
Pasien Demam  
atau *Fever*  
> 37.5

# Contoh: *Cold* (flue) atau Tidak?

- Data Training (gambar):



*Fever* (demam) terdiri dari 3 nilai:

< 37;

>37 dan < 38;

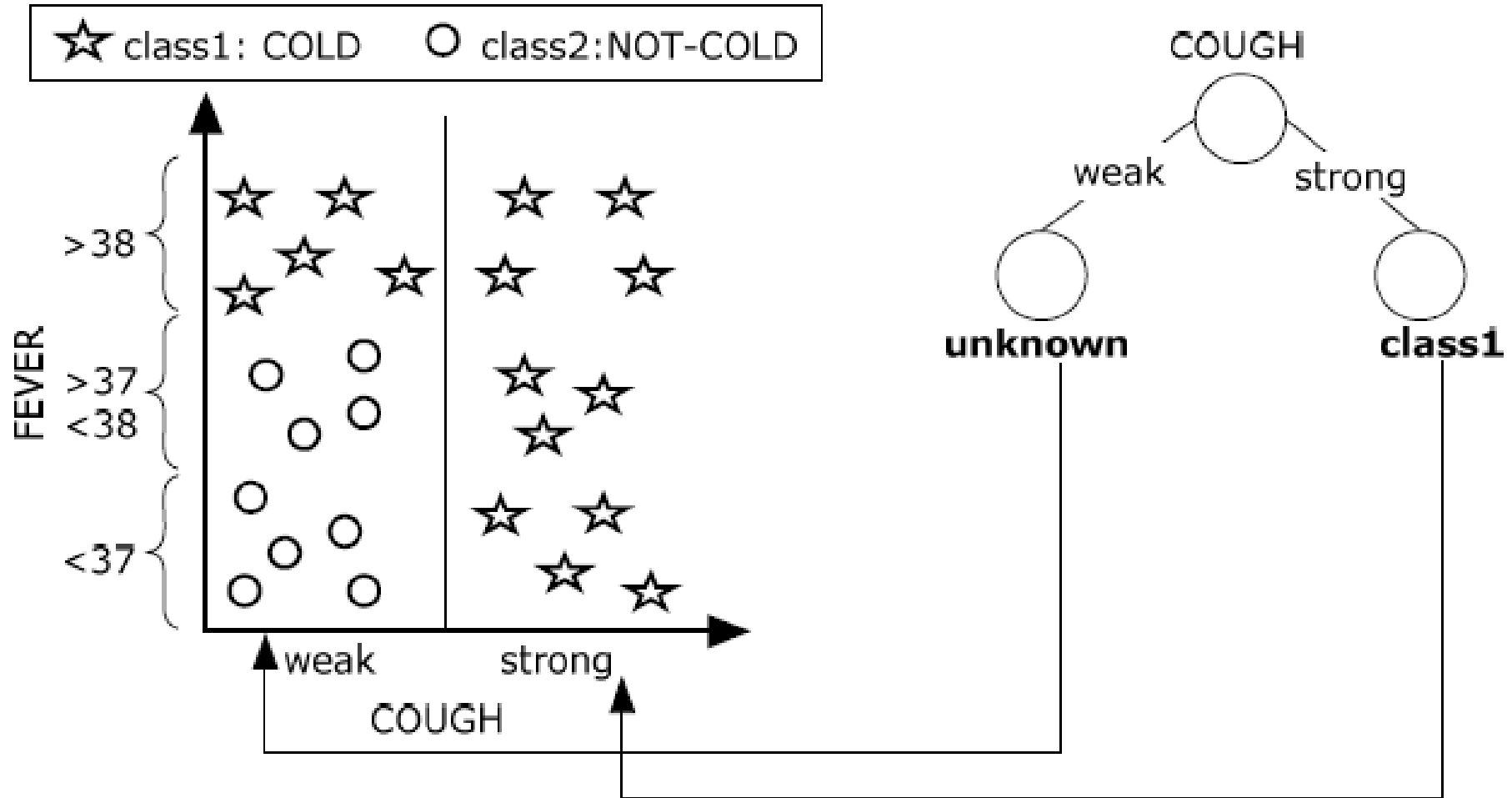
>38.

*Cough* (batuk) terdiri dari 2 nilai:

weak;

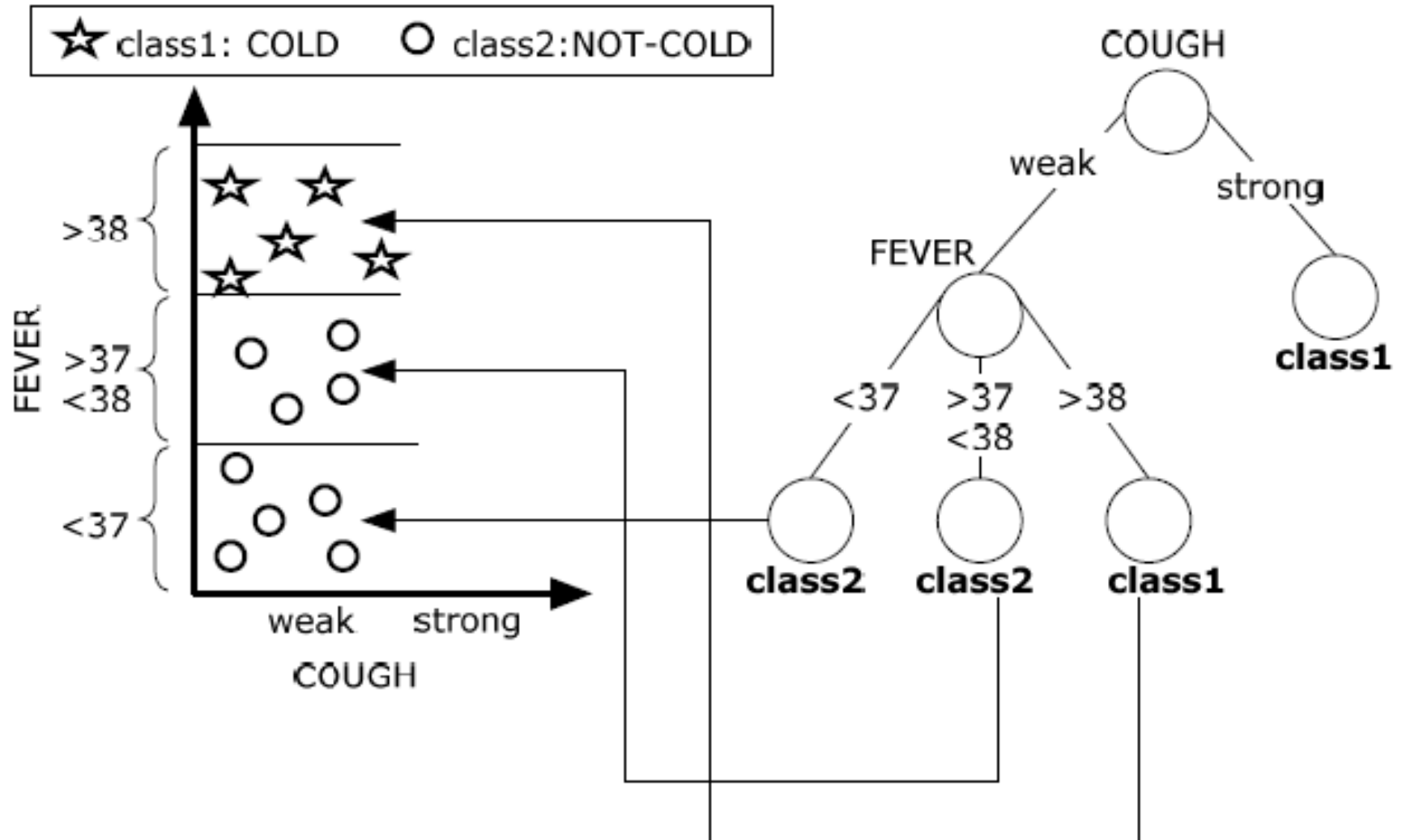
Kencang.

# Pohon: *Cough* dipertimbangkan





# Pohon Akhir, Benarkah?



# *Induction of Decision Trees (ID3)*

---

1. A adalah atribut keputusan “terbaik” bagi node berikutnya
2. Jadikan A sebagai atribut keputusan
3. Untuk setiap nilai A, buat keturunan baru
4. Urutkan *sample* pada node-node daun
5. Jika *sample* telah terklasifikasi secara sempurna, STOP. Jika tidak, lakukan iterasi terhadap node daun baru.

# Algoritma ID3 - Detail

---

Langkah 1. Buat root dari pohon/tree

Langkah 2.

Langkah 2.1. **IF semua sample dari kelas yang sama atau jumlah sample di bawah ambang/threshold**

Then return kelas tersebut

Langkah 2.2. **IF tidak ada atribut yang tersedia return kelas mayoritas**

Langkah 3.

Langkah 3.1. A menjadi atribut terbaik bagi Node berikutnya

Langkah 3.2. Tunjuk A sebagai atribut keputusan bagi Node

Langkah 4. **Untuk setiap nilai v yang mungkin dari A, buat turunan baru dari Node.**

**Tambahkan cabang di bawah label A "A = v"**

Langkah 5.  $S_v$  menjadi subset dari *sample* dimana atribut  $A=v$

Secara rekursif, terapkan algoritma terhadap  $S_v$

Langkah 6. **IF sample *training* terklasifikasi sempurna, SELESAI**

**Else lakukan iterasi pada Node daun baru.**

**End**

# Membangun Tree

---

- Mengapa dihasilkan pohon demikian?
- Mengapa Cough dijadikan sebagai Root?
- Tentukan Atribut Terbaik
- Jangan membangun pohon secara acak/random, dapat menghabiskan banyak *memory* dan waktu proses.
- Salah satu solusi (umum digunakan):  
*Gunakan Entropy dan Information Gain*

# Entropy

---

- $S$  : *sample* (contoh) data training
- $P_+$  : bagian dari contoh positif dalam  $S$
- $P_-$  : bagian dari contoh negatif dalam  $S$
- Ukuran *Entropy* dari  $S$ :

$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

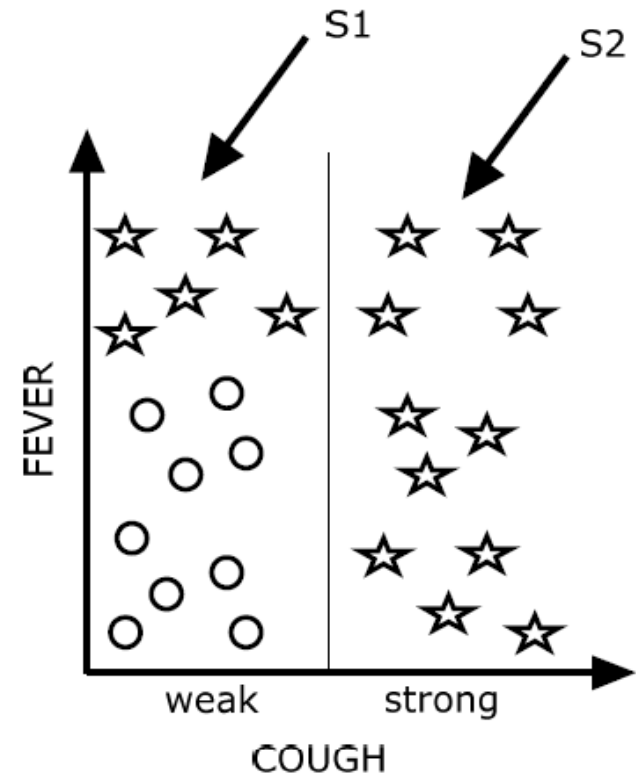
- Disederhakan menjadi:  $\text{Entropy}(S) = \sum_{p \in \{p_+, p_-\}} p(-\log_2 p)$
- $E(S) = 0$  jika *sample* murni (semua + atau semua -),  
 $E(S) = 1$  jika  $p_+ = p_- = 0.5$

# Menghitung Entropy: Basis *Cough*

- Atribut *Cough* membagi sample S menjadi 2, yaitu S1 dan S2

$$S1 = \{5 +, 9 -\}$$

$$S2 = \{11 +, 0 -\}$$

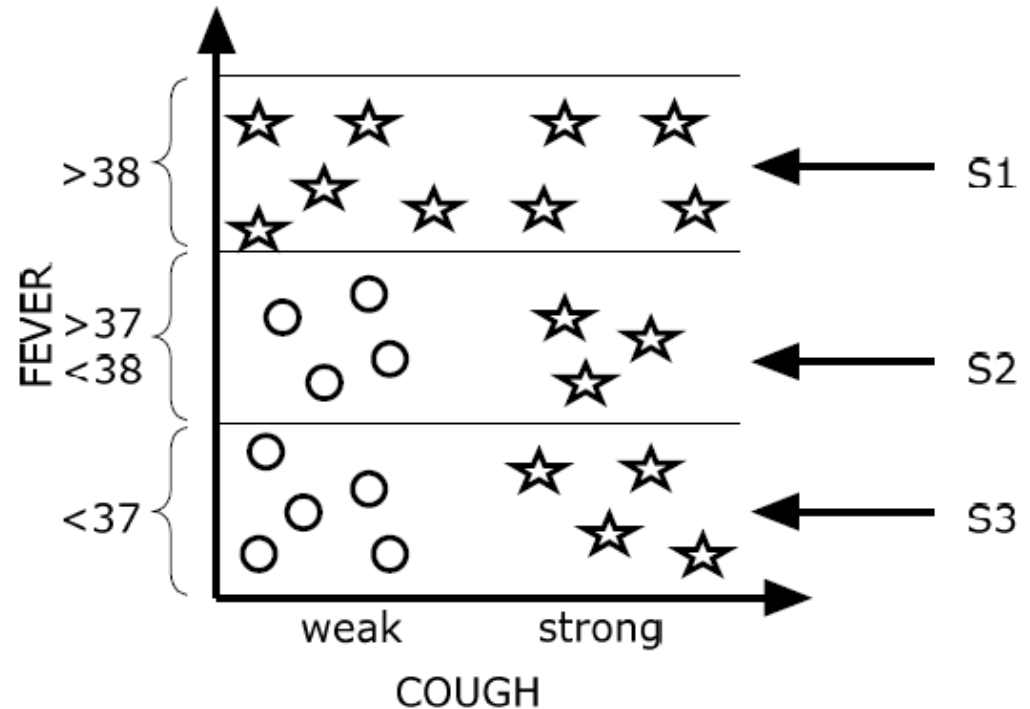


$$Entropy(S1) = -\frac{5}{14} \log_2 \left( \frac{5}{14} \right) - \frac{9}{14} \log_2 \left( \frac{9}{14} \right)$$

$$Entropy(S2) = 0$$

# Menghitung Entropy: Basis *Fever*

- $S1 = \{9 +, 0 -\}$
- $S2 = \{3 +, 4 -\}$
- $S3 = \{4+, 5-\}$



$$Entropy(S1) = 0$$

$$Entropy(S2) = -\frac{3}{7} \log_2 \left( \frac{3}{7} \right) - \frac{4}{7} \log_2 \left( \frac{4}{7} \right)$$

$$Entropy(S3) = -\frac{4}{9} \log_2 \left( \frac{4}{9} \right) - \frac{5}{9} \log_2 \left( \frac{5}{9} \right)$$

# Lanjutkan...

---

- Langkah-langkah:
  - Konversi gambar tersebut ke bentuk tabel atau rules
  - Gunakan *entropy* dan *gain* untuk menentukan Variabel atau Atribut root (Variabel paling penting)
  - Bangun pohon.
  - Jika kelas di daun belum dapat ditentukan, hitung *gain* ulang untuk nilai-nilai terkait.
  - Sampai selesai.



# Information Gain

---

- Disebut juga *mutual information* antara atribut A dan label-label dari *sample S*.

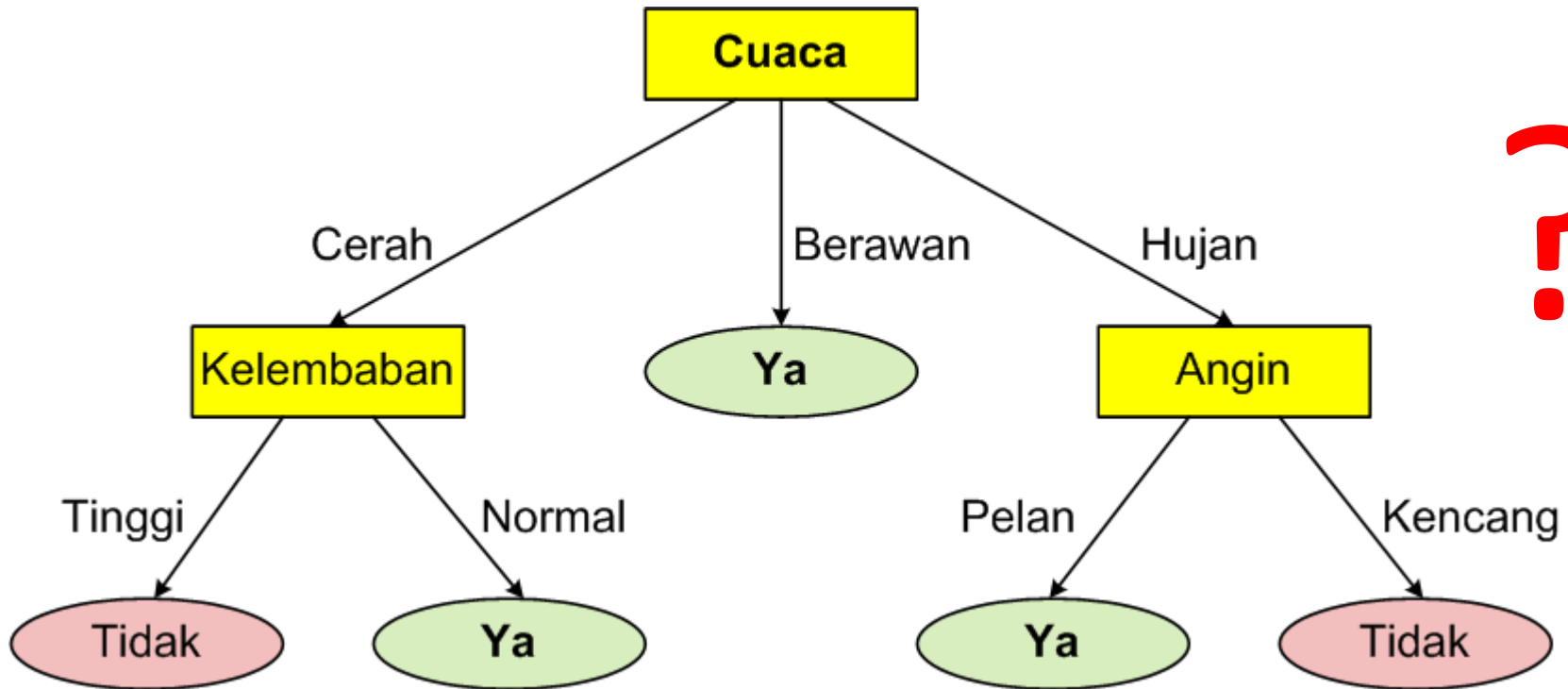
$$gain(S, A) = Entropy(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)$$

- Entropy( $S_v$ ) adalah entropy dari satu sub-sample setelah pembagian S berdasarkan pada semua nilai yang mungkin dari atribut A.
- Nilai *gain* terbesar dianggap sebagai Atribut paling penting.

# Contoh: Main Tennis?

Hari	Cuaca	Suhu	Kelembaban	Angin	Main?
H1	Cerah	Panas	Tinggi	Pelan	Tidak
H2	Cerah	Panas	Tinggi	Kencang	Tidak
H3	Berawan	Panas	Tinggi	Pelan	Ya
H4	Hujan	Sejuk	Tinggi	Pelan	Ya
H5	Hujan	Dingin	Normal	Pelan	Ya
H6	Hujan	Dingin	Normal	Kencang	Tidak
H7	Berawan	Dingin	Normal	Kencang	Ya
H8	Cerah	Sejuk	Tinggi	Pelan	Tidak
H9	Cerah	Dingin	Normal	Pelan	Ya
H10	Hujan	Sejuk	Normal	Pelan	Ya
H11	Cerah	Sejuk	Normal	Kencang	Ya
H12	Berawan	Sejuk	Tinggi	Kencang	Ya
H13	Berawan	Panas	Normal	Pelan	Ya
H14	Hujan	Sejuk	Tinggi	Kencang	Tidak

# Pohon yang dihasilkan



Formulasi Logis:

$(\text{Cuaca} = \text{Cerah} \wedge \text{Kelembaban} = \text{Normal})$

$\vee (\text{Cuaca} = \text{Berawan})$

$\vee (\text{Cuaca} = \text{Hujan} \wedge \text{Angin} = \text{Pelan})$

# Langkah Pertama...

---

- Hitung *Entropy* dari S,  $E(S)$ ?

$$-(9/14) \log_2(9/14) - (5/14) \log_2(5/14) = 0.940$$

- Hitung Gain untuk setiap atribut, yaitu  $\text{Gain}(S, \text{Angin})$ ,  $\text{Gain}(S, \text{Cuaca})$ ,  $\text{Gain}(S, \text{Suhu})$ , dan  $\text{Gain}(S, \text{Kelembaban})$

- Misal:

$$\begin{aligned} \text{Gain}(S, \text{Angin}) &= E(S) - (8/14) * E(S_{\text{Pelan}}) - (6/14) * E(S_{\text{Kencang}}) \\ &= 0.940 - (8/14) * 0.811 - (6/14) * 1.0 \\ &= 0.048 \end{aligned}$$

- Sebelumnya, hitung  $E(S_{\text{Pelan}})$  dan  $E(S_{\text{Kencang}})$ !

# Langkah-langkah

---

- Entropy (SPelan) =  $-6/8 * \text{Log}_2 (6/8) - 2/8 * \text{Log}_2 (2/8) = 0.811$
- Entropy (SKencang) =  $-3/6 * \text{Log}_2 (3/6) - 3/6 * \text{Log}_2 (3/6) = 1.0$

Gain (S, Cuaca) = 0.246

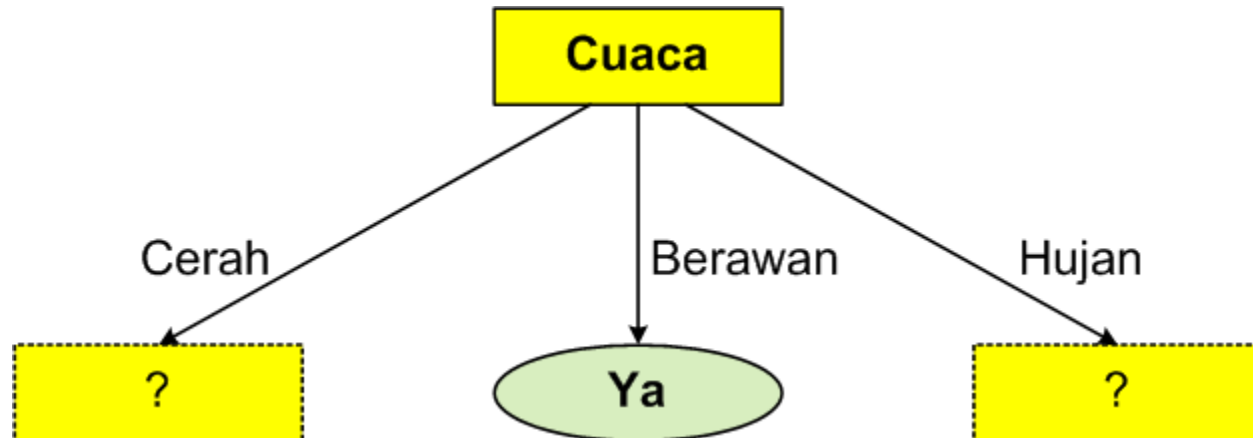
Gain (S, Suhu) = 0.029

Gain (S, Kelembaban) = 0.151

- Nilai Gain (S, Cuaca) paling tinggi. Inilah Variabel terbaik. Jadikan Node ROOT.
- Diperoleh Cuaca dengan 3 Cabang, masing-masing bernilai **Cerah**, **Berawan** dan **Hujan**.

# Langkah-langkah

- Saat Cuaca bernilai Cerah (cabang Cerah), terdapat jawaban Ya atau Tidak. Artinya cabang ini tidak menghasilkan Node Daun.
- Begitu pula saat Cuaca bernilai Hujan (cabang Hujan), masih dapat bernilai Ya atau Tidak.
- Saat Cuaca bernilai Berawan, semuanya mengembalikan label Ya. Berarti diperoleh Node daun bernilai Ya.



# Menentukan Node berikutnya

---

- Pada cabang bernilai Cerah, tentukan variabel baru yang akan menjadi node baru.
- Atribut/Variabel mana yang akan diuji lebih dulu?
- Hitung *Gain* semua variabel, selain Cuaca, yang atribut cuacanya bernilai Cerah.
- $SCerah = \{H1, H2, H8, H9 \text{ dan } H11\}$ , ada 5 *sample* Cuaca bernilai Cerah.

$$\text{Gain}(SCerah, \text{Suhu}) = 0.570$$

$$\text{Gain}(SCerah, \text{Kelembaban}) = 0.970$$

$$\text{Gain}(SCerah, \text{Angin}) = 0.019$$

- Jelas, Gain dari Kelembaban paling tinggi.
- Variabel Kelembaban menjadi node baru di bawah Cuaca (cabang Cerah).

# Contoh: *Gain (SCerah, Suhu)*

- Entropy (SCerah) =  $-2/5 * \text{Log}_2(2/5) - 3/5 * \text{Log}_2(3/5)$   
= 0.970

H1	Cerah	Panas	Tinggi	Pelan	Tidak
H2	Cerah	Panas	Tinggi	Kencang	Tidak
H8	Cerah	Sejuk	Tinggi	Pelan	Tidak
H9	Cerah	Dingin	Normal	Pelan	Ya
H11	Cerah	Sejuk	Normal	Kencang	Ya

$$\text{Gain}(\text{SCerah}, \text{Suhu}) = E(\text{SCerah}) - (2/5 * E(\text{SPanas}) + 2/5 * E(\text{SSejuk}) + 1/5 * E(\text{SDingin}))$$

$E(\text{SPanas}) = 0$  (Semuanya berlabel Tidak saat Suhu bernilai Panas)

$E(\text{SSejuk}) = 1$  (baris H8 berlabel Tidak, H11 bernilai Ya)

$E(\text{SDingin}) = 0$  (semua berlabel Ya)

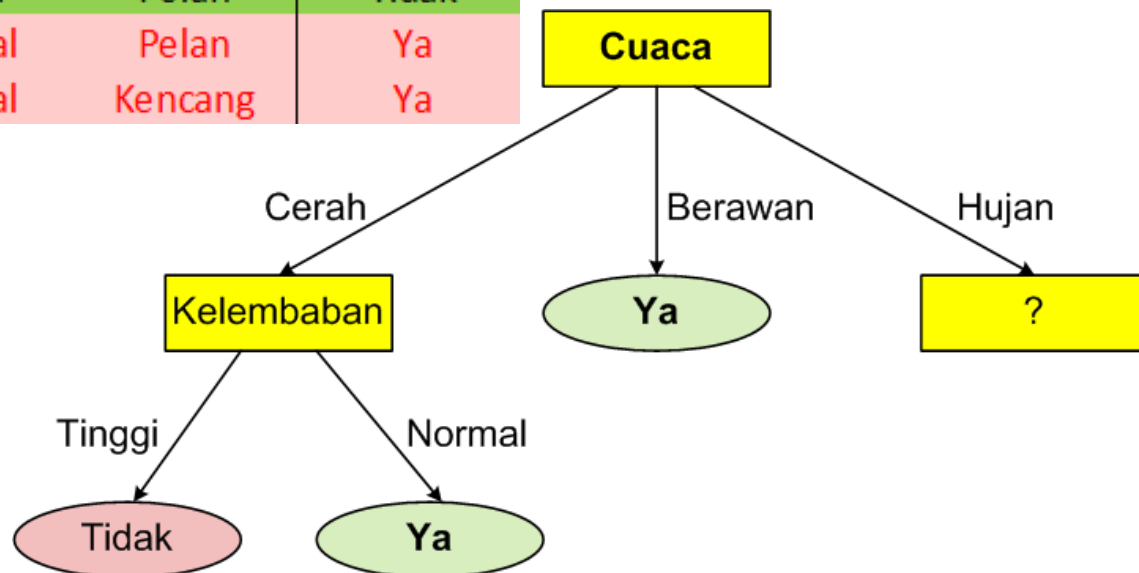
$$\text{Gain}(\text{SCerah}, \text{Suhu}) = 0.970 - (2/5 * 0 + 2/5 * 1 + 1/5 * 0) = 0,570$$



# Cuaca=Cerah, Kelembaban?

- Saat Kelembaban bernilai Tinggi, diperoleh daun berlabel Tidak. Sedangkan saat bernilai Normal, diperoleh daun berlabel Ya.
- Cabang ini selesai.

H1	Cerah	Panas	Tinggi	Pelan	Tidak
H2	Cerah	Panas	Tinggi	Kencang	Tidak
H8	Cerah	Sejuk	Tinggi	Pelan	Tidak
H9	Cerah	Dingin	Normal	Pelan	Ya
H11	Cerah	Sejuk	Normal	Kencang	Ya



# Cabang “Hujan”?

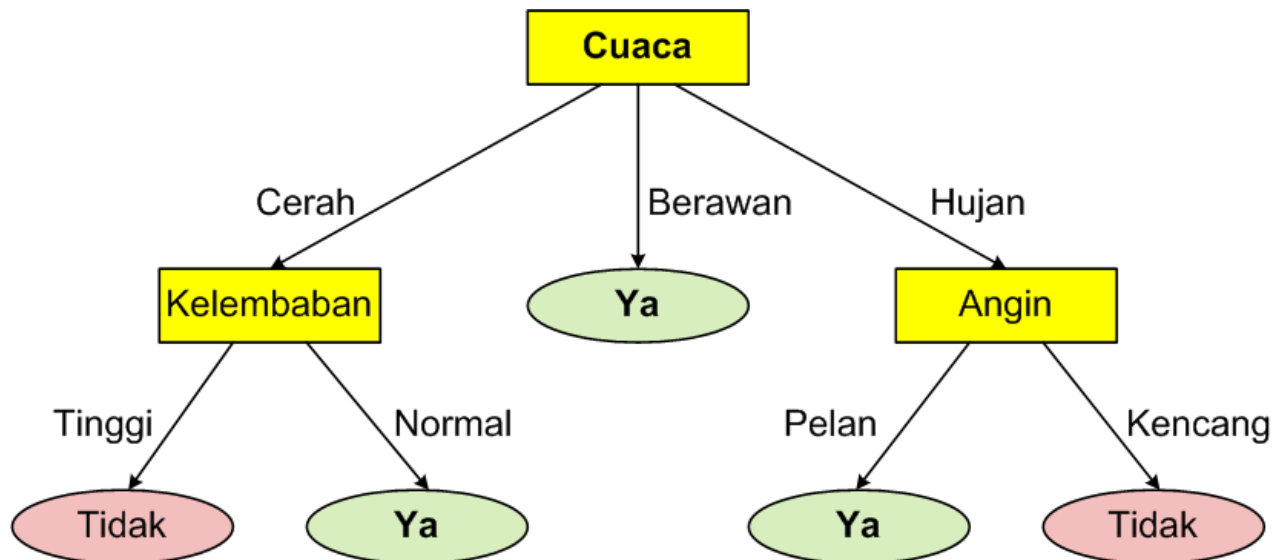
---

- Dengan cara yang sama, hitung  $\text{Gain}(\text{SHujan}, \text{Angin})$ ,  $\text{Gain}(\text{SHujan}, \text{Suhu})$ , dan  $\text{Gain}(\text{SHujan}, \text{Kelembaban})$ .
- Dari hasil perhitungan,  $\text{Gain}(\text{SHujan}, \text{Angin})$  memberikan nilai tertinggi. Artinya, Atribut Angin menjadi node di bawah cabang Hujan tersebut.
- Saat Variabel Angin bernilai Kencang, diperoleh node daun berlabel Tidak, sedangkan saat variabel Angin bernilai Pelan, diperoleh node daun berlabel Ya.
- Selesai.

# Pohon Akhir

- Semua node daun menunjukkan label sesuai nama kategori pada data training.

H4	Hujan	Sejuk	Tinggi	Pelan	Ya
H5	Hujan	Dingin	Normal	Pelan	Ya
H6	Hujan	Dingin	Normal	Kencang	Tidak
H10	Hujan	Sejuk	Normal	Pelan	Ya
H14	Hujan	Sejuk	Tinggi	Kencang	Tidak



# Contoh: Resiko Kredit

No.	Sejarah Kredit	Hutang	Jaminan	Pendapatan	Resiko
1	Buruk	Banyak	Tidak	5 jt -10 jt	Tinggi
2	Tidak Diketahui	Banyak	Tidak	10 jt - 20 jt	Tinggi
3	Tidak Diketahui	Sedikit	Tidak	11 jt - 20 jt	Sedang
4	Tidak Diketahui	Sedikit	Tidak	5 jt -10 jt	Tinggi
5	Tidak Diketahui	Sedikit	Tidak	Di atas 20 jt	Rendah
6	Tidak Diketahui	Sedikit	Ya	Di atas 20 jt	Rendah
7	Buruk	Sedikit	Tidak	5 jt -10 jt	Tinggi
8	Buruk	Sedikit	Tidak	Di atas 20 jt	Sedang
9	Baik	Sedikit	Tidak	Di atas 20 jt	Rendah
10	Baik	Banyak	Ya	Di atas 20 jt	Rendah

# Contoh: Resiko Kredit

No.	Sejarah Kredit	Hutang	Jaminan	Pendapatan	Resiko
11	Baik	Banyak	Tidak	10 jt - 20 jt	Tinggi
12	Baik	Banyak	Tidak	11 jt - 20 jt	Sedang
13	Baik	Banyak	Tidak	Di atas 20 jt	Rendah
14	Buruk	Banyak	Tidak	11 jt - 20 jt	Tinggi
15	Tidak Diketahui	Banyak	Tidak	12 jt - 20 jt	Tinggi
16	Tidak Diketahui	Sedikit	Tidak	13 jt - 20 jt	Sedang
17	Buruk	Sedikit	Ya	14 jt - 20 jt	Sedang
18	Tidak Diketahui	Sedikit	Ya	Di atas 20 jt	Rendah
19	Baik	Sedikit	Ya	5 jt -10 jt	Rendah
20	Buruk	Banyak	Tidak	6 jt -10 jt	Tinggi

- Bagaimana membuat pohon keputusan?
- Coba gunakan **Gain<sub>Split</sub>** sebagai alternatif menentukan variabel terbaik.

# Latihan

---

- Gunakan *Entropy* dan *Gain* untuk membangun *decision tree* dari masalah Resiko Kredit.
- Gunakan *Entropy* dan *Gain<sub>Split</sub>* untuk membangun *decision tree* dari masalah Main Tennis dan Resiko Kredit.
- *Catatan*: Variabel terbaik mempunyai nilai *Gain<sub>Split</sub>* terkecil, berlawanan dengan *Gain*.

$$Gain_{Split} = \sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v)$$

# Solusi dengan Teori BaYa?

Hari	Cuaca	Suhu	Kelembaban	Angin	Main?
H1	Cerah	Panas	Tinggi	Pelan	Tidak
H2	Cerah	Panas	Tinggi	Kencang	Tidak
H3	Berawan	Panas	Tinggi	Pelan	Ya
H4	Hujan	Sejuk	Tinggi	Pelan	Ya
H5	Hujan	Dingin	Normal	Pelan	Ya
H6	Hujan	Dingin	Normal	Kencang	Tidak
H7	Berawan	Dingin	Normal	Kencang	Ya
H8	Cerah	Sejuk	Tinggi	Pelan	Tidak
H9	Cerah	Dingin	Normal	Pelan	Ya
H10	Hujan	Sejuk	Normal	Pelan	Ya
H11	Cerah	Sejuk	Normal	Kencang	Ya
H12	Berawan	Sejuk	Tinggi	Kencang	Ya
H13	Berawan	Panas	Normal	Pelan	Ya
H14	Hujan	Sejuk	Tinggi	Kencang	Tidak

# Fase Pembelajaran (*Learning*)

<b>Cuaca</b>	Main=Ya	Main=Tidak
<i>Cerah</i>	2/9	3/5
<i>Berawan</i>	4/9	0/5
<i>Hujan</i>	3/9	2/5

<b>Suhu</b>	Main=Ya	Main=Tidak
<i>Panas</i>	2/9	2/5
<i>Sejuk</i>	4/9	2/5
<i>Dingin</i>	3/9	1/5

<b>Kelembaban</b>	Main=Ya	Main=Tidak
<i>Tinggi</i>	3/9	4/5
<i>Tidakrmal</i>	6/9	1/5

<b>Angin</b>	Main=Ya	Main=Tidak
<i>Kencang</i>	3/9	3/5
<i>Pelan</i>	6/9	2/5

$$P(\text{Main=Ya}) = 9/14$$

$$P(\text{Main=Tidak}) = 5/14$$



# Fase Pengujian (*Test*)

- **Misal data baru:**

$\mathbf{x}' = (\text{Cuaca}=\text{Cerah}, \text{Suhu}=\text{Dingin}, \text{Kelembaban}=\text{Tinggi}, \text{Angin}=\text{Kencang})$

- **Perhatikan tabel:**

$$P(\text{Cuaca}=\text{Cerah} \mid \text{Main}=\text{Ya}) = 2/9$$

$$P(\text{Suhu}=\text{Dingin} \mid \text{Main}=\text{Ya}) = 3/9$$

$$P(\text{Kelembaban}=\text{Tinggi} \mid \text{Main}=\text{Ya}) = 3/9$$

$$P(\text{Angin}=\text{Kencang} \mid \text{Main}=\text{Ya}) = 3/9$$

$$P(\text{Main}=\text{Ya}) = 9/14$$

$$P(\text{Cuaca}=\text{Cerah} \mid \text{Main}=\text{Tidak}) = 3/5$$

$$P(\text{Suhu}=\text{Dingin} \mid \text{Main}=\text{Tidak}) = 1/5$$

$$P(\text{Kelembaban}=\text{Tinggi} \mid \text{Main}=\text{Tidak}) = 4/5$$

$$P(\text{Angin}=\text{Kencang} \mid \text{Main}=\text{Tidak}) = 3/5$$

$$P(\text{Main}=\text{Tidak}) = 5/14$$

- **Petakan Rule:**

$$P(\text{Ya} \mid \mathbf{x}'): [P(\text{Cerah} \mid \text{Ya})P(\text{Dingin} \mid \text{Ya})P(\text{Tinggi} \mid \text{Ya})P(\text{Kencang} \mid \text{Ya})]P(\text{Main}=\text{Ya}) = 0.0053$$

$$P(\text{Tidak} \mid \mathbf{x}'): [P(\text{Cerah} \mid \text{Tidak})P(\text{Dingin} \mid \text{Tidak})P(\text{Tinggi} \mid \text{Tidak})P(\text{Kencang} \mid \text{Tidak})]P(\text{Main}=\text{Tidak}) = 0.0206$$

**Diperoleh  $P(\text{Ya} \mid \mathbf{x}') < P(\text{Tidak} \mid \mathbf{x}')$ , sehingga  $\mathbf{x}'$  dilabelkan sebagai “*Tidak*”.**