

Data Mining

Pre-Test

1. Jelaskan perbedaan antara **data**, **informasi** dan **pengetahuan**!
2. Jelaskan apa yang anda ketahui tentang **data mining**!
3. Sebutkan **peran utama data mining**!
4. Sebutkan **pemanfaatan dari data mining** di berbagai bidang!
5. **Pengetahuan atau pola apa yang bisa kita dapatkan dari data di bawah?**

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMAN 7	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

Course Outline



1. Pengantar Data Mining

1.1 Apa itu Data Mining?

1.2 Peran Utama dan Metode Data Mining

1.3 Sejarah dan Penerapan Data Mining

1.1 Apa itu Data Mining?

Manusia Memproduksi Data

Manusia memproduksi beragam data yang **jumlah dan ukurannya sangat besar**

- Astronomi
- Bisnis
- Kedokteran
- Ekonomi
- Olahraga
- Cuaca
- Financial
- ...



Pertumbuhan Data

Astronomi

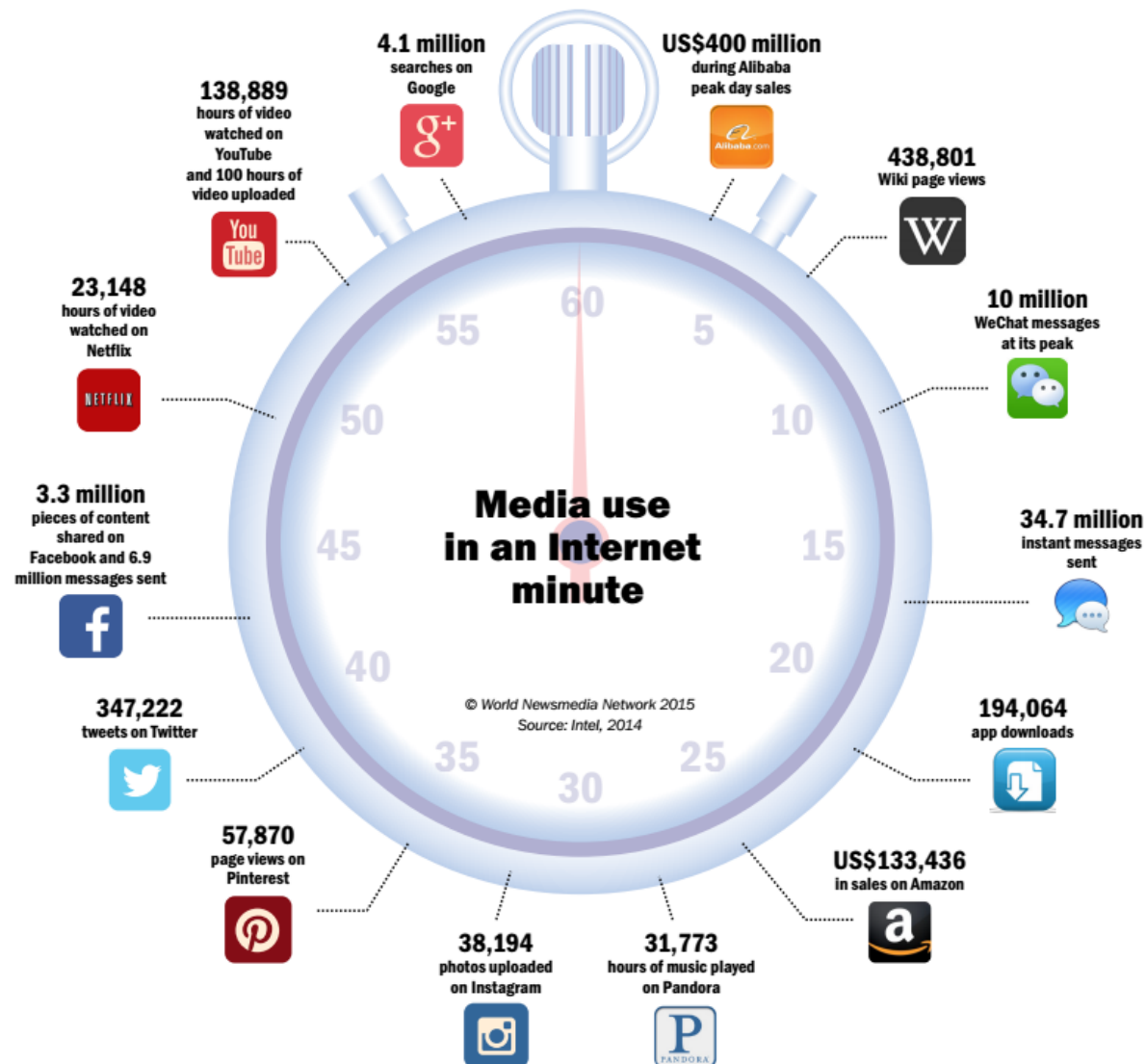
- **Sloan Digital Sky Survey**
 - New Mexico, 2000
 - **140TB** over 10 years
- **Large Synoptic Survey Telescope**
 - Chile, 2016
 - Will acquire **140TB every five days**

kilobyte (kB)	10^3
megabyte (MB)	10^6
gigabyte (GB)	10^9
terabyte (TB)	10^{12}
petabyte (PB)	10^{15}
exabyte (EB)	10^{18}
zettabyte (ZB)	10^{21}
yottabyte (YB)	10^{24}

Biologi dan Kedokteran

- European Bioinformatics Institute (**EBI**)
 - **20PB of data** (genomic data doubles in size each year)
 - A single sequenced human genome can be around **140GB** in size

Perubahan Kultur dan Perilaku



*g Data Trends
dia, 2015)*

Datangnya Tsunami Data

- **Mobile Electronics market**

- 5B mobile phones in use in 2010
- 150M tablets was sold in 2012 (IDC)
- 200M is global notebooks shipments in 2012 (Digitimes Research)

- **Web and Social Networks generate**
amount of data

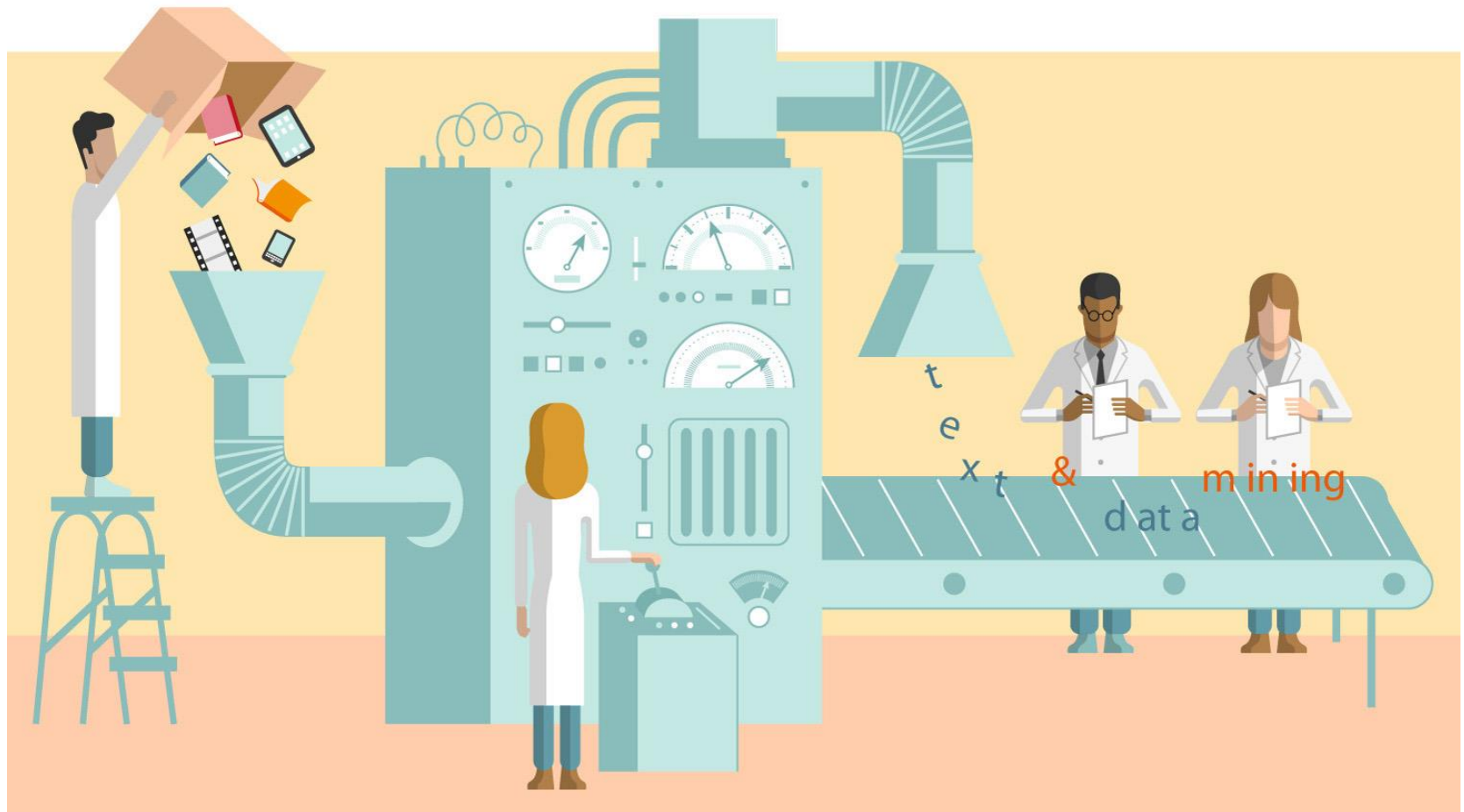
- Google processes 100 PB per day, 3 million servers
- Facebook has 300 PB of user data per day
- Youtube has 1000PB video storage
- 235 TBs data collected by the US Library of Congress
- 15 out of 17 sectors in the US have more data stored per company than the US Library of Congress

kilobyte (kB)	10^3
megabyte (MB)	10^6
gigabyte (GB)	10^9
terabyte (TB)	10^{12}
petabyte (PB)	10^{15}
exabyte (EB)	10^{18}
zettabyte (ZB)	10^{21}
yottabyte (YB)	10^{24}

Mengapa Data Mining?

We are drowning in data, but
starving for knowledge!

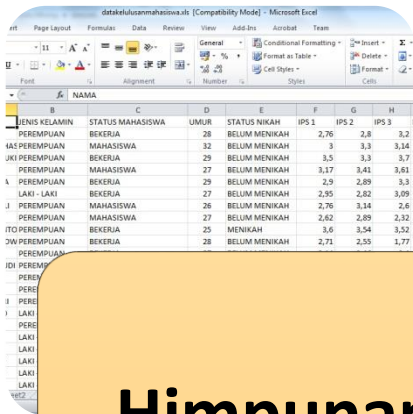
Apa itu Data Mining?



Apa itu Data Mining?

- Disiplin ilmu yang mempelajari **metode** untuk **mengekstrak pengetahuan** atau **menemukan pola** dari suatu data yang besar
- Ekstraksi dari **data** ke **pengetahuan**:
 1. **Data**: **fakta yang terekam** dan tidak membawa arti
 2. **Pengetahuan**: **pola, rumus**, aturan atau model yang muncul dari data
- Nama lain data mining:
 - **Knowledge Discovery in Database (KDD)**
 - Knowledge extraction
 - Pattern analysis
 - Information harvesting
 - Business intelligence

Apa Itu Data Mining?



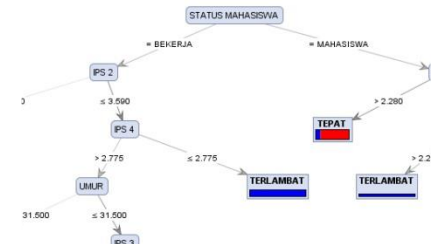
	B	C	D	E	F	G	H
	NAMA	STATUS MAHASISWA	UMUR	STATUS NIKAH	IPS 1	IPS 2	IPS 3
1	LENIS KELAMN	MAHASISWA	28	BELUM MENIKAH	2,76	2,8	3,2
2	PEREMPUAN	MAHASISWA	32	BELUM MENIKAH	3	3,3	3,14
3	UKI PEREMPUAN	MAHASISWA	29	BELUM MENIKAH	3,5	3,3	3,7
4	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	3,17	3,41	3,61
5	PEREMPUAN	MAHASISWA	29	BELUM MENIKAH	2,9	2,89	3,3
6	LAKI - LAKI	MAHASISWA	27	BELUM MENIKAH	2,95	2,82	3,09
7	PEREMPUAN	MAHASISWA	28	BELUM MENIKAH	2,76	3,14	2,6
8	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	2,62	2,89	2,12
9	ITO PEREMPUAN	MAHASISWA	25	MENIKAH	3,6	3,54	3,52
10	PEREMPUAN	MAHASISWA	28	BELUM MENIKAH	2,71	2,55	1,77



$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \frac{b-a}{n} \sum_{k=1}^n f\left(a + \frac{b-a}{n} \cdot k\right)$$

$$= \left(-m_2 \tilde{x} \tan(\Phi) \right) \left[l - \frac{r^2}{4l} + r \left(\cos(\omega t) + \frac{r}{4l} \cos(2\omega t) \right) \right]$$

$$= F_1 \cdot e^{\left(-\zeta + \sqrt{\zeta^2 - 1} \right) \omega t} \quad \left(-\zeta - \sqrt{\zeta^2 - 1} \right) \omega t$$



**Himpunan
Data**

**Metode Data
Mining**

Pengetahuan

Definisi Data Mining

- Melakukan **ekstraksi** untuk mendapatkan **informasi penting** yang sifatnya **implisit** dan sebelumnya tidak diketahui, dari suatu data (*Witten et al., 2011*)
- Kegiatan yang meliputi pengumpulan, pemakaian data historis untuk **menemukan keteraturan, pola dan hubungan** dalam set data berukuran besar (*Santosa, 2007*)
- **Extraction of interesting** (non-trivial, **implicit**, **previously unknown** and potentially useful) **patterns or knowledge** from huge amount of data (*Han et al., 2011*)

Data - Informasi – Pengetahuan

NIP	TGL	DATANG	PULANG
1103	02/12/2004	07:20	15:40
1142	02/12/2004	07:45	15:33
1156	02/12/2004	07:51	16:00
1173	02/12/2004	08:00	15:15
1180	02/12/2004	07:01	16:31
1183	02/12/2004	07:49	17:00

Data Kehadiran Pegawai

Data - Informasi – Pengetahuan

NIP	Masuk	Alpa	Cuti	Sakit	Telat
1103	22				
1142	18	2		2	
1156	10	1	11		
1173	12	5			5
1180	10			12	

Informasi Akumulasi Bulanan Kehadiran Pegawai

Data - Informasi – Pengetahuan

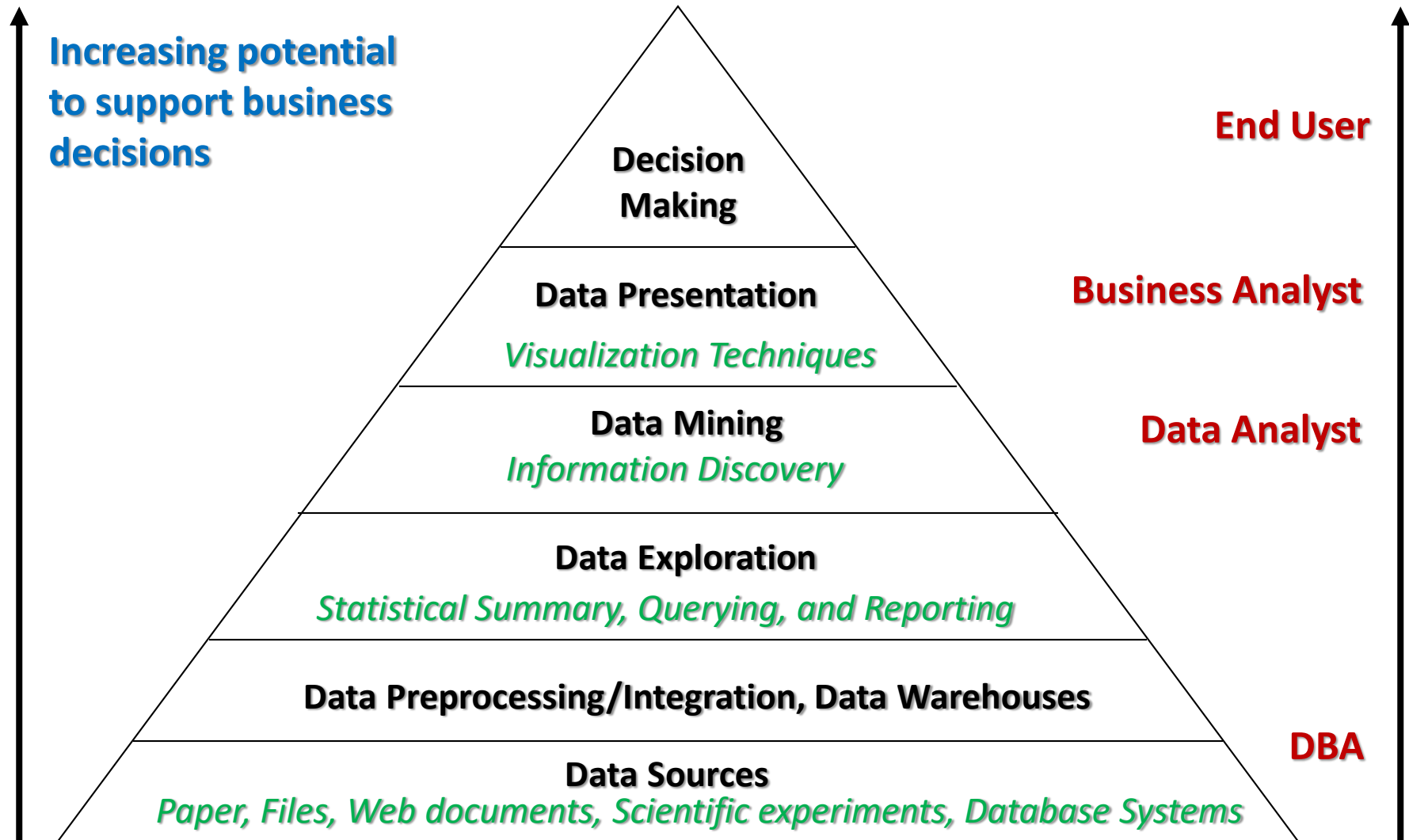
	Senin	Selasa	Rabu	Kamis	Jumat
Terlambat	7	0	1	0	5
Pulang Cepat	0	1	1	1	8
Izin	3	0	0	1	4
Alpa	1	0	2	0	2

Pola Kebiasaan Kehadiran Mingguan Pegawai

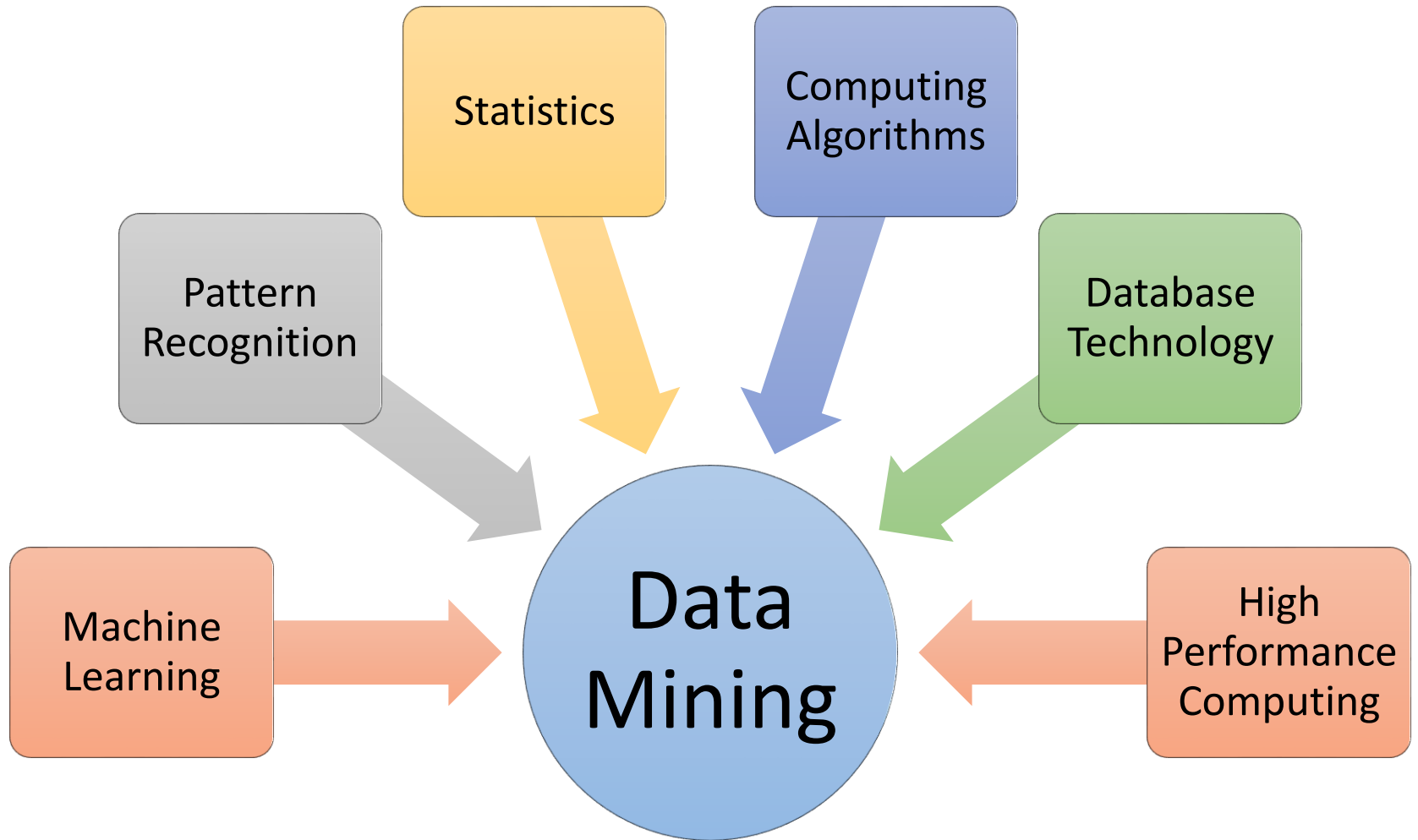
Data - Informasi – Pengetahuan - Kebijakan

- Kebijakan **penataan jam kerja karyawan** khusus untuk hari senin dan jumat
- Peraturan jam kerja:
 - Hari **Senin** dimulai jam 10:00
 - Hari **Jumat** diakhiri jam 14:00
 - Sisa jam kerja **dikompensasi ke hari lain**

Data Mining pada Business Intelligence



Hubungan dengan Berbagai Bidang



Masalah-Masalah di Data Mining

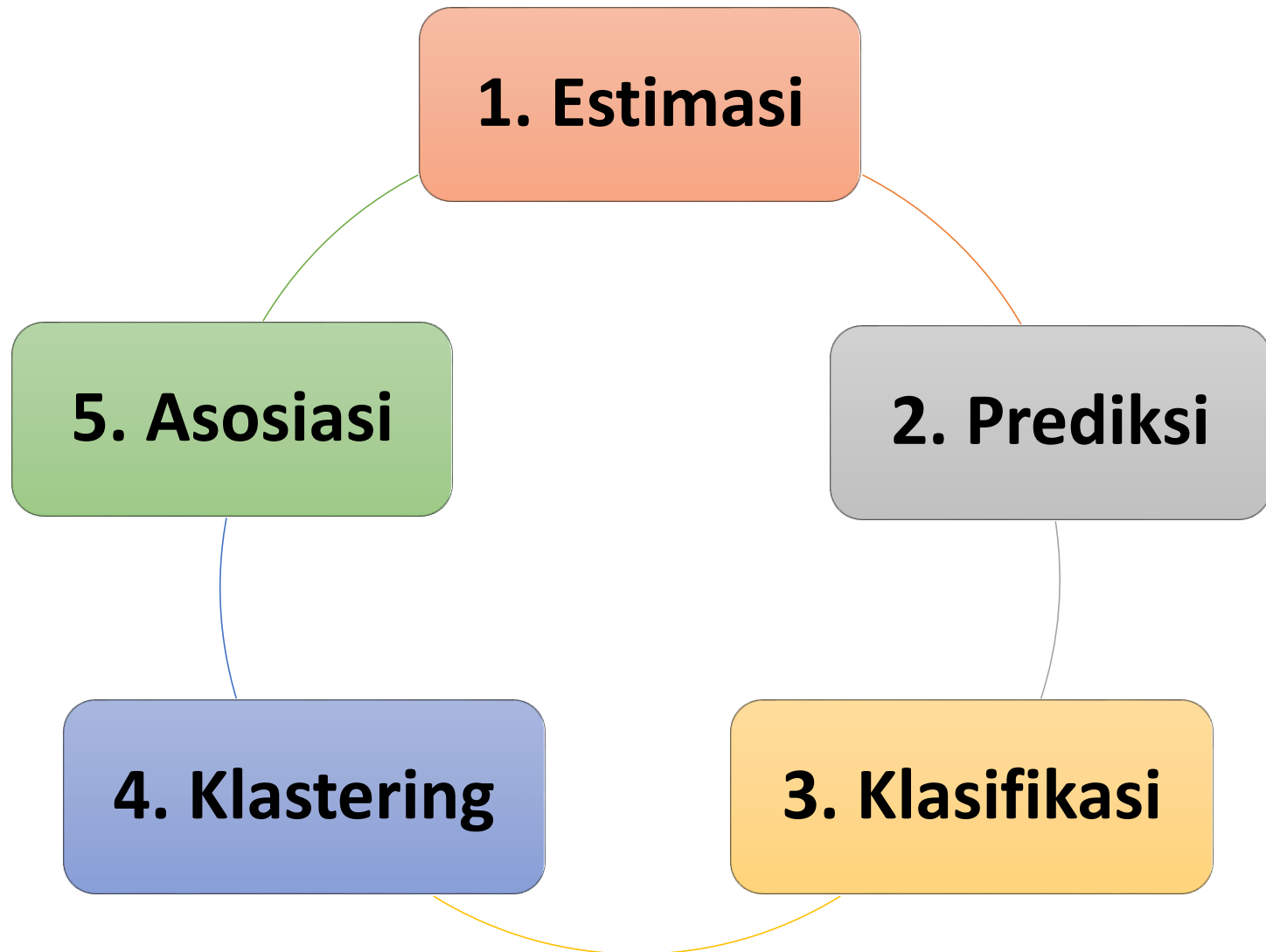
- Tremendous **amount** of data
 - Algorithms must be **highly scalable** to handle such as tera-bytes of data
- **High-dimensionality** of data
 - Micro-array may have tens of **thousands of dimensions**
- High **complexity** of data
 - **Data streams** and sensor data
 - **Time-series data**, temporal data, sequence data
 - Structure data, graphs, **social networks** and multi-linked data
 - Heterogeneous **databases** and legacy databases
 - Spatial, spatiotemporal, **multimedia**, text and **Web data**
 - **Software programs**, scientific simulations
- New and sophisticated **applications**

Latihan

1. Jelaskan dengan kalimat sendiri apa yang dimaksud dengan **data mining**?
2. Sebutkan **sudut pandang multidimensi** dari data mining!

1.2 Peran Utama Data Mining

Peran Utama Data Mining



Dataset (Himpunan Data)

Attribute/Feature

Class/Label/Target

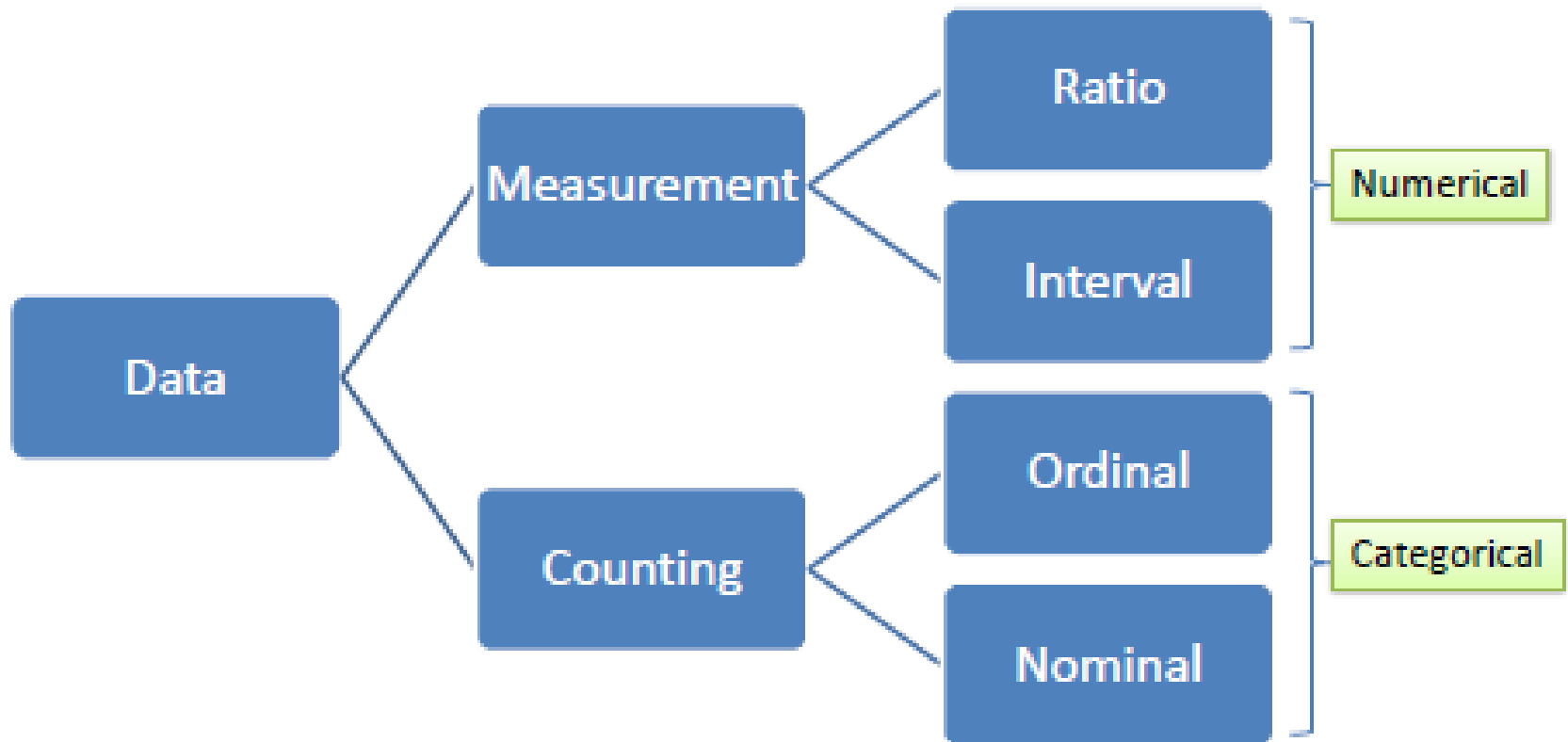
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	<i>Iris setosa</i>
2	4.9	3.0	1.4	0.2	<i>Iris setosa</i>
3	4.7	3.2	1.3	0.2	<i>Iris setosa</i>
4	4.6	3.1	1.5	0.2	<i>Iris setosa</i>
5	5.0	3.6	1.4	0.2	<i>Iris setosa</i>
...					
51	7.0	3.2	4.7	1.4	<i>Iris versicolor</i>
52	6.4	3.2	4.5	1.5	<i>Iris versicolor</i>
53	6.9	3.1	4.9	1.5	<i>Iris versicolor</i>
54	5.5	2.3	4.0	1.3	<i>Iris versicolor</i>
55	6.5	2.8	4.6	1.5	<i>Iris versicolor</i>
...					
101	6.3	3.3	6.0	2.5	<i>Iris virginica</i>
102	5.8	2.7	5.1	1.9	<i>Iris virginica</i>
103	7.1	3.0	5.9	2.1	<i>Iris virginica</i>

Record/
Object/
Sample/
Tuple

Nominal

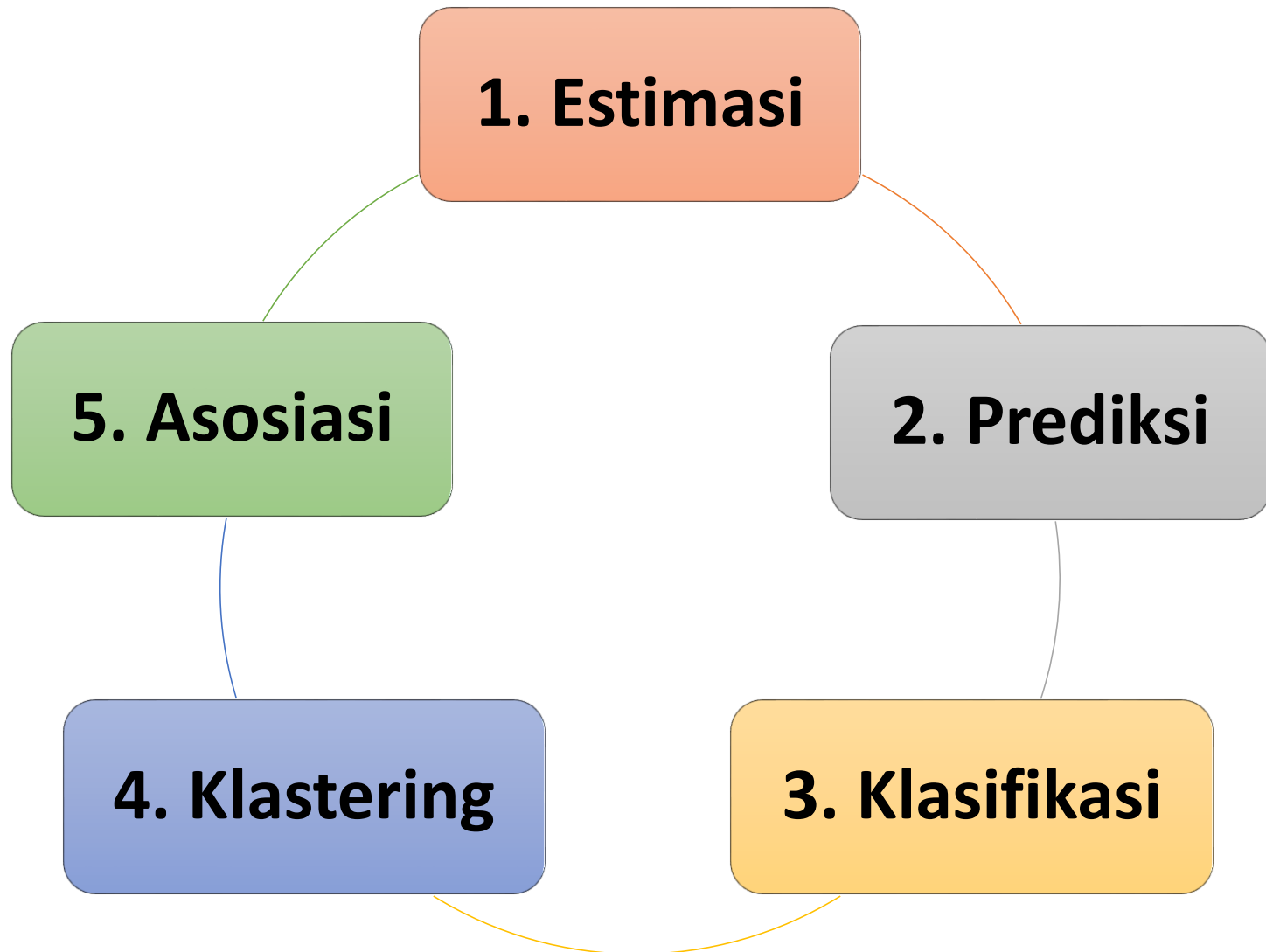
Numerik

Jenis Atribut



Jenis Atribut	Deskripsi	Contoh	Operasi
Ratio (Mutlak)	<ul style="list-style-type: none"> Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui Mempunyai titik nol yang absolut (*, /) 	<ul style="list-style-type: none"> Umur Berat badan Tinggi badan Jumlah uang 	geometric mean, harmonic mean, percent variation
Interval (Jarak)	<ul style="list-style-type: none"> Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui Tidak mempunyai titik nol yang absolut (+, -) 	<ul style="list-style-type: none"> Suhu 0°c-100°c, Umur 20-30 tahun 	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
Ordinal (Peringkat)	<ul style="list-style-type: none"> Data yang diperoleh dengan cara kategorisasi atau klasifikasi Tetapi diantara data tersebut terdapat hubungan atau berurutan (<, >) 	<ul style="list-style-type: none"> Tingkat kepuasan pelanggan (puas, sedang, tidak puas) 	median, percentiles, rank correlation, run tests, sign tests
Nominal (Label)	<ul style="list-style-type: none"> Data yang diperoleh dengan cara kategorisasi atau klasifikasi Menunjukkan beberapa object yang berbeda (=, ≠) 	<ul style="list-style-type: none"> Kode pos Jenis kelamin Nomer id karyawan Nama kota 	mode, entropy, contingency correlation, χ^2 test

Peran Utama Data Mining



1. Estimasi Waktu Pengiriman Pizza

Customer	Jumlah Pesanan (P)	Jumlah Traffic Light (TL)	Jarak (J)	Waktu Tempuh (T)
1	3	3	3	16
2	1	7	4	20
3	2	4	6	18
4	4	6	8	36
...				
1000	2	4	2	12

Label

Pembelajaran dengan
Metode Estimasi (*Regresi Linier*)

$$\text{Waktu Tempuh (T)} = 0.48P + 0.23TL + 0.5J$$

Pengetahuan

Contoh: Estimasi Performansi CPU

- **Example** • 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
...							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- **Linear regression** function

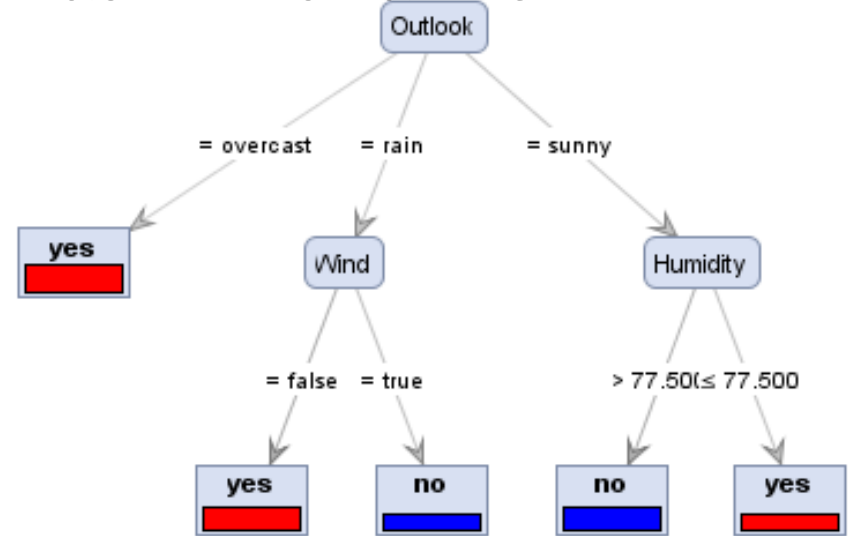
$$\begin{aligned} \text{PRP} = & -55.9 + 0.0489 \text{ MYCT} + 0.0153 \text{ MMIN} + 0.0056 \text{ MMAX} \\ & + 0.6410 \text{ CACH} - 0.2700 \text{ CHMIN} + 1.480 \text{ CHMAX} \end{aligned}$$

Output/Pola/Model/Knowledge

1. Formula/**Function** (Rumus atau Fungsi Regresi)

- $\text{WAKTU TEMPUH} = 0.48 + 0.6 \text{ JARAK} + 0.34 \text{ LAMPU} + 0.2 \text{ PESANAN}$

2. Decision **Tree** (Pohon Keputusan)



3. Korelasi dan **Asosiasi**

4. **Rule** (Aturan)

- IF $\text{ips3}=2.8$ THEN lulustepatwaktu

5. **Cluster** (Klaster)

2. Prediksi Harga Saham

Label



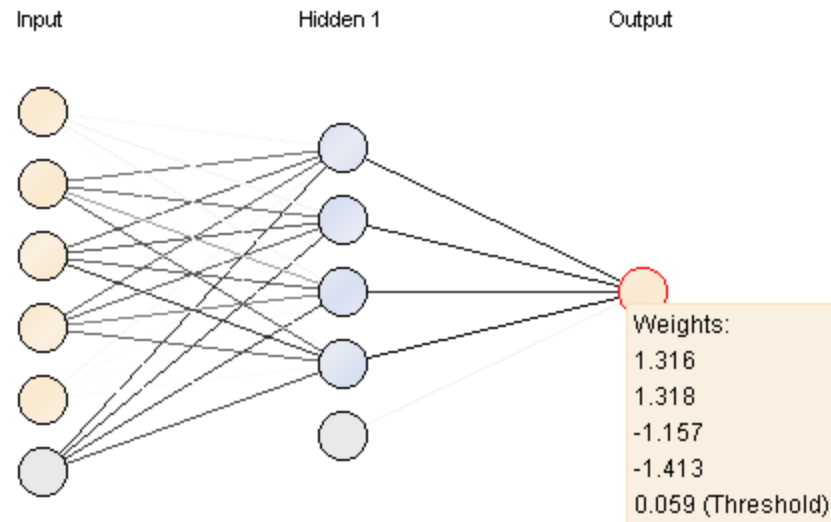
Row No.	Close	Date	Open	High	Low	Volume
1	1286.570	Apr 11, 2006	1296.600	1300.710	1282.960	2232880000
2	1288.120	Apr 12, 2006	1286.570	1290.930	1286.450	1938100000
3	1289.120	Apr 13, 2006	1288.120	1292.090	1283.370	1891940000
4	1285.330	Apr 17, 2006	1289.120	1292.450	1280.740	1794650000
5	1307.280	Apr 18, 2006	1285.330	1309.020	1285.330	2595440000
6	1309.930	Apr 19, 2006	1307.650	1310.390	1302.790	2447310000
7	1311.460	Apr 20, 2006	1309.930	1318.160	1306.380	2512920000
8	1311.280	Apr 21, 2006	1311.460	1317.670	1306.590	2392630000
9	1308.110	Apr 24, 2006	1311.280	1311.280	1303.790	2117330000
10	1301.740	Apr 25, 2006	1308.110	1310.790	1299.170	2366380000
11	1305.410	Apr 26, 2006	1301.740	1310.970	1301.740	2502690000
12	1309.720	Apr 27, 2006	1305.410	1315	1295.570	2772010000
13	1310.610	Apr 28, 2006	1309.720	1316.040	1306.160	2419920000

Dataset harga saham dalam bentuk **time series** (rentet waktu)



Pembelajaran dengan
Metode Prediksi (*Neural Network*)

Pengetahuan berupa Rumus
Neural Network



Prediction Plot



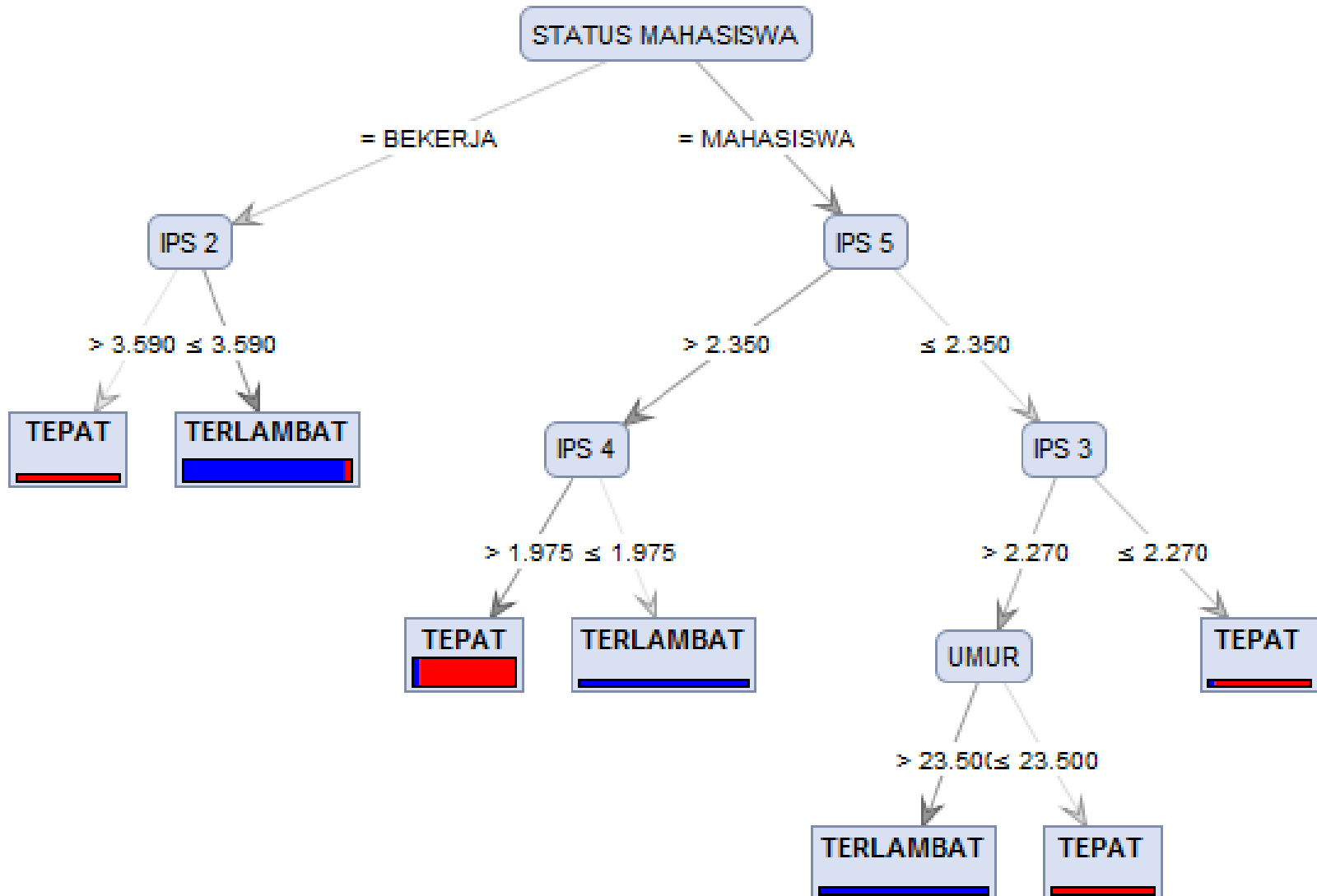
3. Klasifikasi Kelulusan Mahasiswa

Label
↓

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

↓
Pembelajaran dengan
Metode Klasifikasi (*C4.5*)
↓

Pengetahuan Berupa Pohon Keputusan



Contoh: Rekomendasi Main Golf

- **Input:**

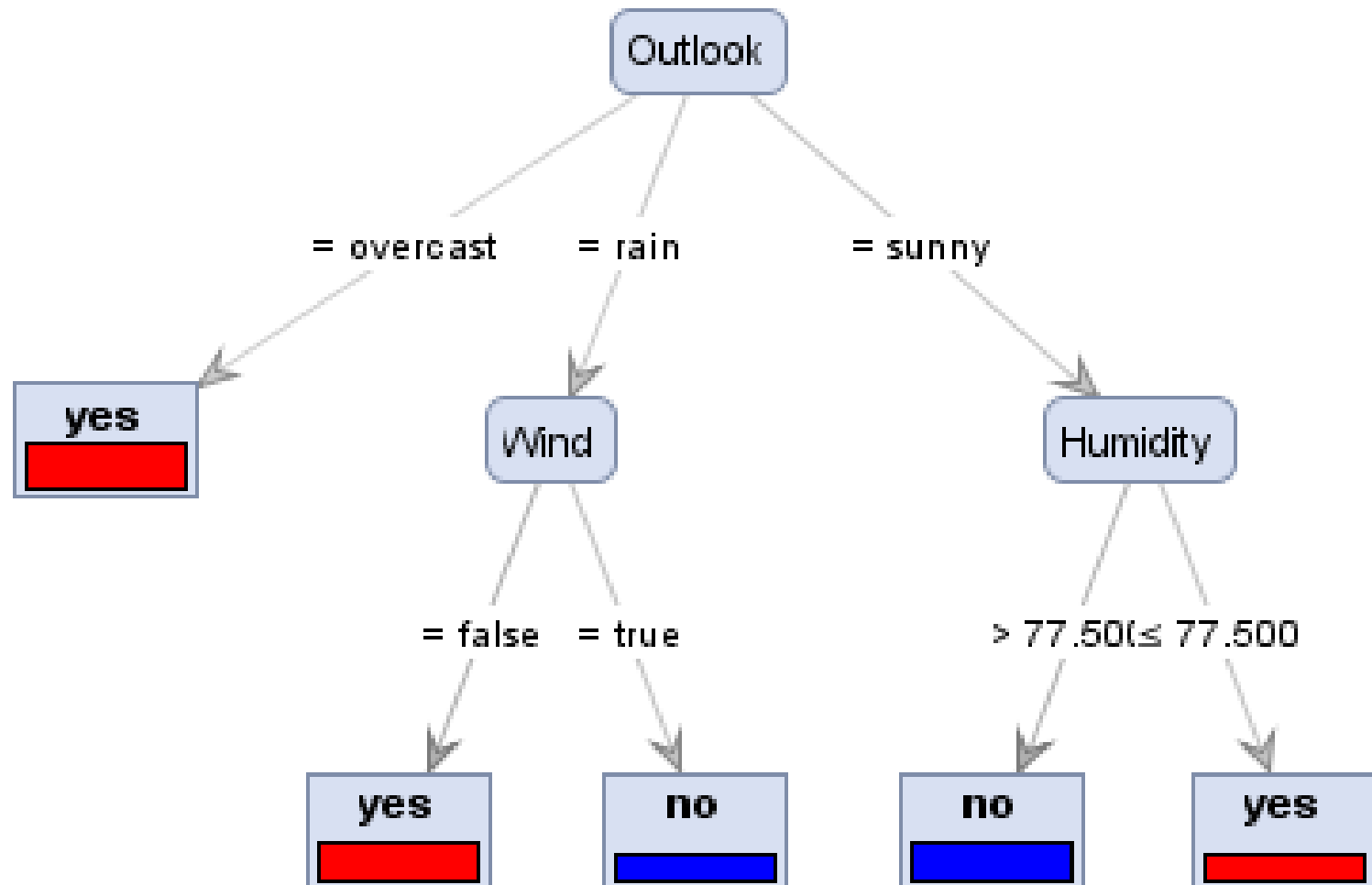
Outlook	Temperature	Humidity	Windy	Play
Sunny	hot	high	false	no
Sunny	hot	high	true	no
Overcast	hot	high	false	yes
Rainy	mild	high	false	yes
Rainy	cool	normal	false	yes
Rainy	cool	normal	true	no
Overcast	cool	normal	true	yes
Sunny	mild	high	false	no
Sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
Sunny	mild	normal	true	yes
Overcast	mild	high	true	yes
Overcast	hot	normal	false	yes
Rainy	mild	high	true	no

- **Output (**

If outlook = sunny and humidity = high then play = no
If outlook = rainy and windy = true then play = no
If outlook = overcast then play = yes
If humidity = normal then play = yes
If none of the above then play = yes

Contoh: Rekomendasi Main Golf

• (



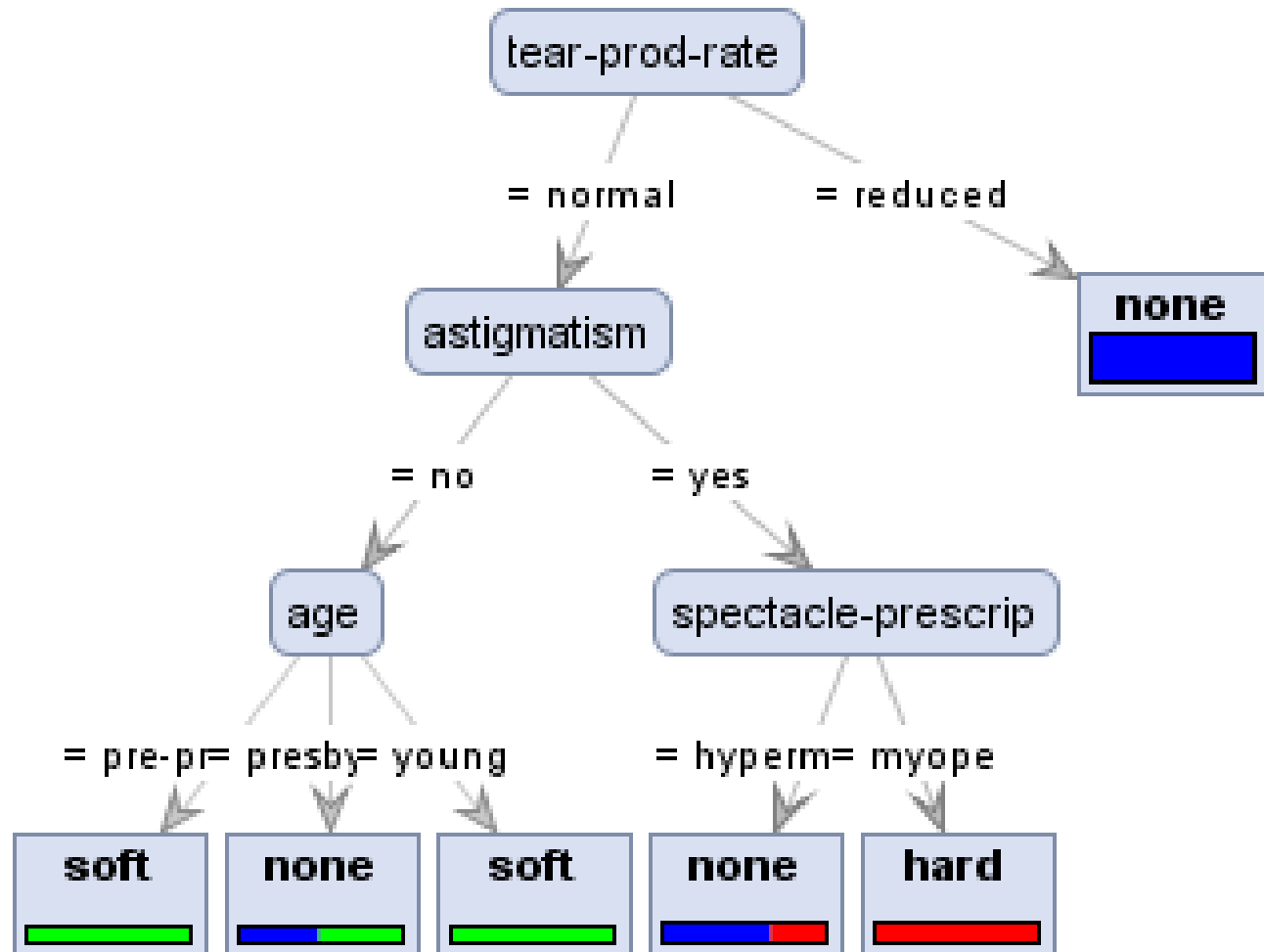
Contoh: Rekomendasi Contact Lens

- **Input:**

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Recommended Lenses
young	myope	no	reduced	none
young	myope	no	normal	soft
young	myope	yes	reduced	none
young	myope	yes	normal	hard
young	hypermetrope	no	reduced	none
young	hypermetrope	no	normal	soft
young	hypermetrope	yes	reduced	none
young	hypermetrope	yes	normal	hard
pre-presbyopic	myope	no	reduced	none
pre-presbyopic	myope	no	normal	soft
pre-presbyopic	myope	yes	reduced	none
pre-presbyopic	myope	yes	normal	hard
pre-presbyopic	hypermetrope	no	reduced	none
pre-presbyopic	hypermetrope	no	normal	soft

Contoh: Rekomendasi Contact Lens

- **Output/Model** (Tree):



4. Klastering Bunga Iris

Dataset Tanpa Label

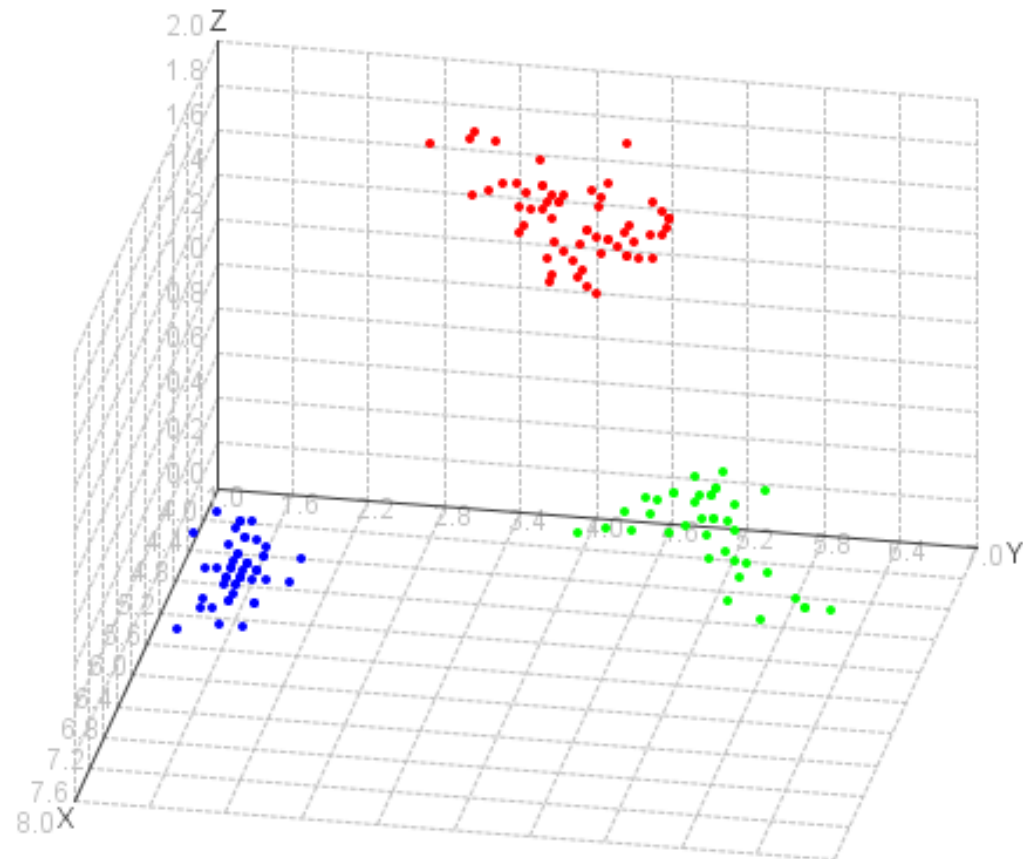
Row No.	id	a1	a2	a3	a4
1	id_1	5.100	3.500	1.400	0.200
2	id_2	4.900	3	1.400	0.200
3	id_3	4.700	3.200	1.300	0.200
4	id_4	4.600	3.100	1.500	0.200
5	id_5	5	3.600	1.400	0.200
6	id_6	5.400	3.900	1.700	0.400
7	id_7	4.600	3.400	1.400	0.300
8	id_8	5	3.400	1.500	0.200
9	id_9	4.400	2.900	1.400	0.200
10	id_10	4.900	3.100	1.500	0.100
11	id_11	5.400	3.700	1.500	0.200



Pembelajaran dengan
Metode Klastering (*K-Means*)


Pengetahuan Berupa Klaster

cluster ● cluster_0 ● cluster_1 ● cluster_2



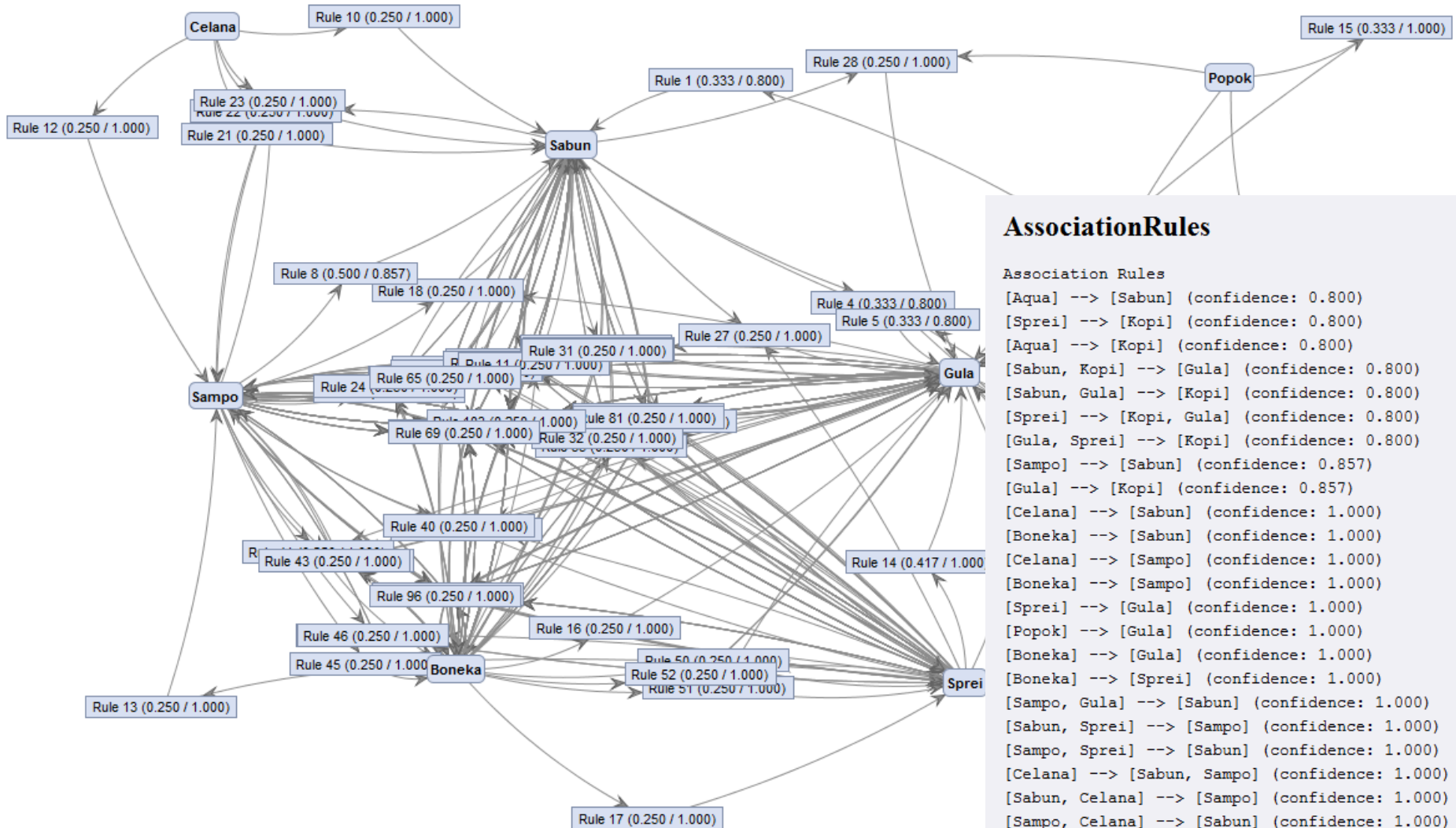
5. Aturan Asosiasi Pembelian Barang

Row No.	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
1	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
2	0.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0
3	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
4	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
6	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
8	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
9	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
10	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
11	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0



Pembelajaran dengan
Metode Asosiasi (*FP-Growth*)

Pengetahuan Berupa Aturan Asosiasi



AssociationRules

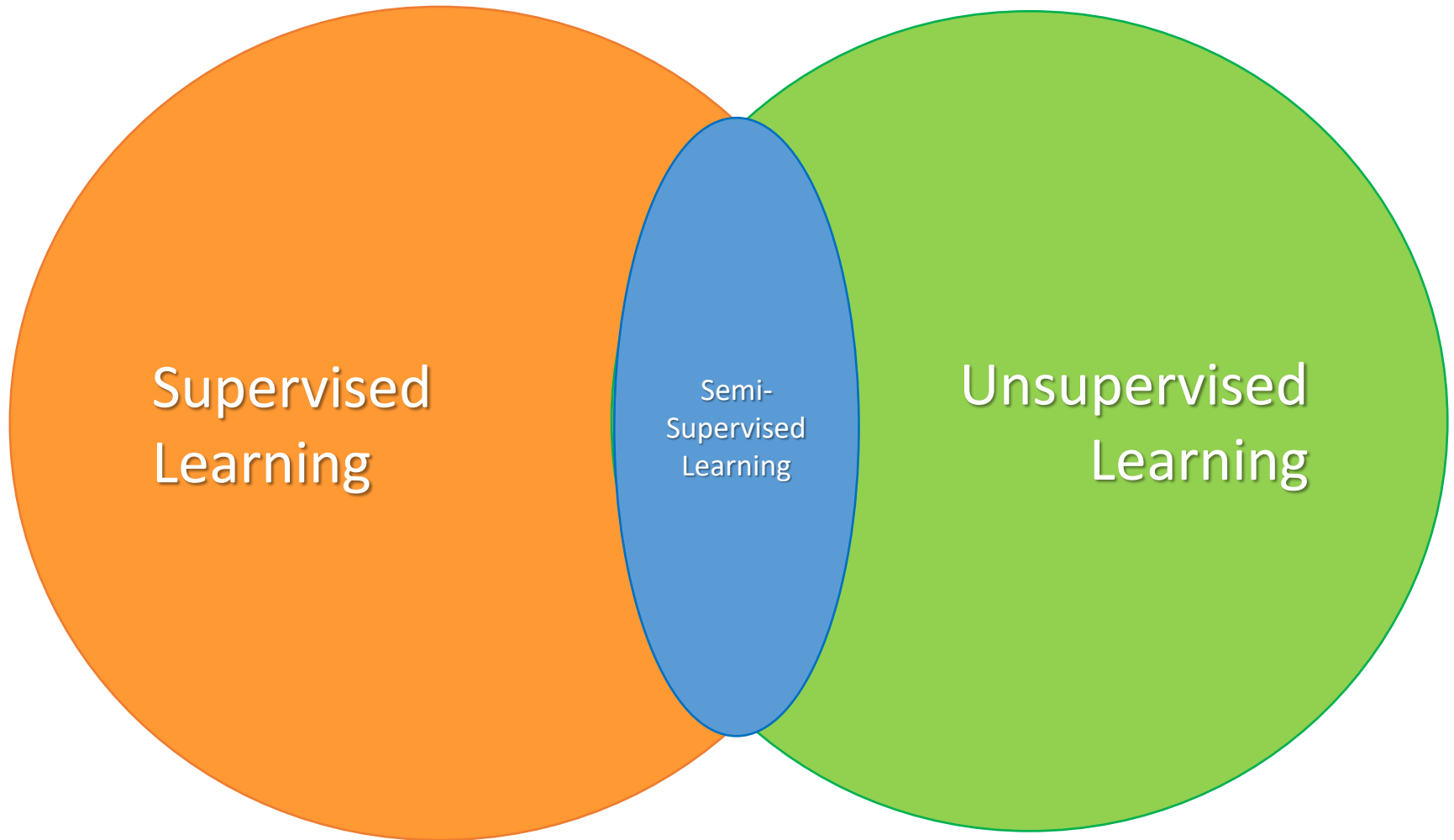
Association Rules

```
[Aqua] --> [Sabun] (confidence: 0.800)
[Sprei] --> [Kopi] (confidence: 0.800)
[Aqua] --> [Kopi] (confidence: 0.800)
[Sabun, Kopi] --> [Gula] (confidence: 0.800)
[Sabun, Gula] --> [Kopi] (confidence: 0.800)
[Sprei] --> [Kopi, Gula] (confidence: 0.800)
[Gula, Sprei] --> [Kopi] (confidence: 0.800)
[Sampo] --> [Sabun] (confidence: 0.857)
[Gula] --> [Kopi] (confidence: 0.857)
[Celana] --> [Sabun] (confidence: 1.000)
[Boneka] --> [Sabun] (confidence: 1.000)
[Celana] --> [Sampo] (confidence: 1.000)
[Boneka] --> [Sampo] (confidence: 1.000)
[Sprei] --> [Gula] (confidence: 1.000)
[Popok] --> [Gula] (confidence: 1.000)
[Boneka] --> [Gula] (confidence: 1.000)
[Boneka] --> [Sprei] (confidence: 1.000)
[Sampo, Gula] --> [Sabun] (confidence: 1.000)
[Sabun, Sprei] --> [Sampo] (confidence: 1.000)
[Sampo, Sprei] --> [Sabun] (confidence: 1.000)
[Celana] --> [Sabun, Sampo] (confidence: 1.000)
[Sabun, Celana] --> [Sampo] (confidence: 1.000)
[Sampo, Celana] --> [Sabun] (confidence: 1.000)
[Boneka] --> [Sabun, Sampo] (confidence: 1.000)
[Sabun, Boneka] --> [Sampo] (confidence: 1.000)
[Sampo, Boneka] --> [Sabun] (confidence: 1.000)
[Sabun, Sprei] --> [Gula] (confidence: 1.000)
```

Contoh Aturan Asosiasi

- Algoritma *association rule* (aturan asosiasi) adalah algoritma yang menemukan atribut yang “**muncul bersamaan**”
- Contoh, pada hari Kamis malam, 1000 pelanggan telah melakukan belanja di supermarket ABC, dimana:
 - 200 orang membeli **Sabun Mandi**
 - dari 200 orang yang membeli sabun mandi, 50 orangnya membeli **Fanta**
- Jadi, association rule menjadi, “**Jika membeli sabun mandi, maka membeli Fanta**”, dengan nilai **support** = $200/1000 = 20\%$ dan nilai **confidence** = $50/200 = 25\%$
- Algoritma association rule diantaranya adalah: **A priori algorithm, FP-Growth algorithm, GRI algorithm**

Metode Learning Pada Algoritma DM




1. Supervised Learning

- Pembelajaran dengan **guru**, data set memiliki **target/label/class**
- **Sebagian besar** algoritma data mining (estimation, prediction/forecasting, classification) adalah supervised learning
- Algoritma melakukan proses belajar berdasarkan **nilai dari variabel target** yang terasosiasi dengan nilai dari variable prediktor

Dataset dengan Class

Attribute/Feature

Class/Label/Target



	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	<i>Iris setosa</i>
2	4.9	3.0	1.4	0.2	<i>Iris setosa</i>
3	4.7	3.2	1.3	0.2	<i>Iris setosa</i>
4	4.6	3.1	1.5	0.2	<i>Iris setosa</i>
5	5.0	3.6	1.4	0.2	<i>Iris setosa</i>
...					
51	7.0	3.2	4.7	1.4	<i>Iris versicolor</i>
52	6.4	3.2	4.5	1.5	<i>Iris versicolor</i>
53	6.9	3.1	4.9	1.5	<i>Iris versicolor</i>
54	5.5	2.3	4.0	1.3	<i>Iris versicolor</i>
55	6.5	2.8	4.6	1.5	<i>Iris versicolor</i>
...					
101	6.3	3.3	6.0	2.5	<i>Iris virginica</i>
102	5.8	2.7	5.1	1.9	<i>Iris virginica</i>
103	7.1	3.0	5.9	2.1	<i>Iris virginica</i>

Nominal

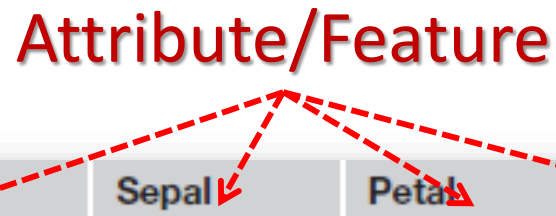
Numerik

2. Unsupervised Learning

- Algoritma data mining mencari pola dari **semua variable (atribut)**
- Variable (atribut) yang menjadi **target/label/class tidak ditentukan (tidak ada)**
- Algoritma **clustering** adalah algoritma unsupervised learning

Dataset tanpa Class

Attribute/Feature



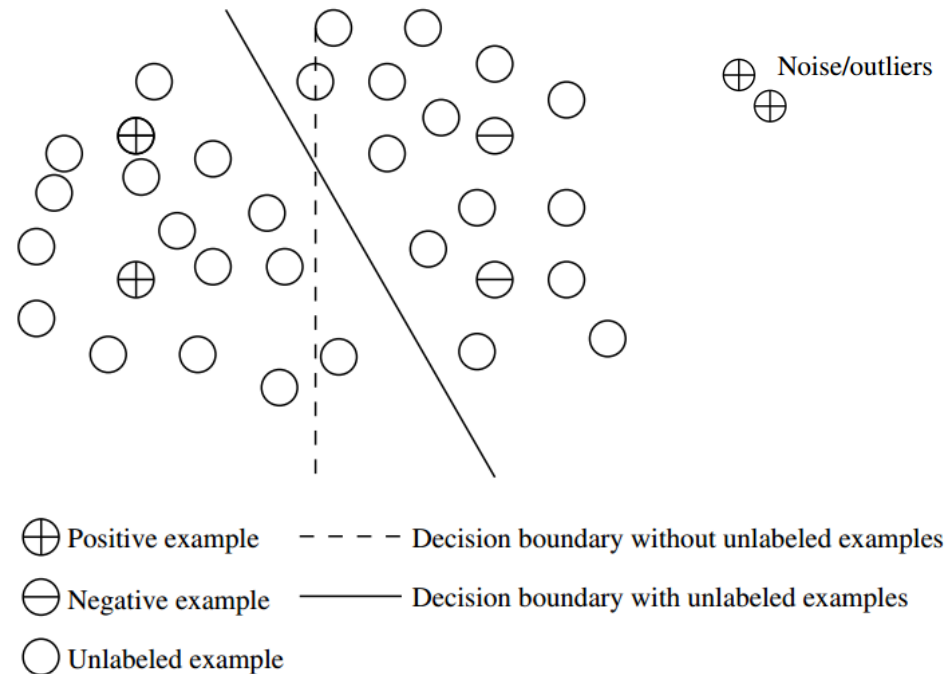
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
...				
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
...				
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1

3. Semi-Supervised Learning

- Semi-supervised learning adalah metode data mining yang menggunakan **data dengan label dan tidak berlabel sekaligus** dalam proses pembelajarannya
- Data yang memiliki kelas digunakan untuk **membentuk model** (pengetahuan), data tanpa label digunakan untuk **membuat batasan** antara kelas

3. Semi-Supervised Learning

- If we consider the **labeled examples**, the **dashed line** is the decision boundary that best partitions the positive examples from the negative examples
- Using the **unlabeled examples**, we can refine the decision boundary to the **solid line**
- Moreover, we can detect that the **two positive examples** at the top right corner, though labeled, are likely **noise or outliers**



Algoritma Data Mining (DM)

1. Estimation (Estimasi):

- Linear Regression, Neural Network, Support Vector Machine, etc

2. Prediction/Forecasting (Prediksi/Peramalan):

- Linear Regression, Neural Network, Support Vector Machine, etc

3. Classification (Klasifikasi):

- Naive Bayes, K-Nearest Neighbor, C4.5, ID3, CART, Linear Discriminant Analysis, Logistic Regression, etc

4. Clustering (Klastering):

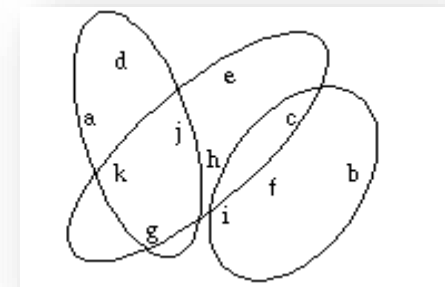
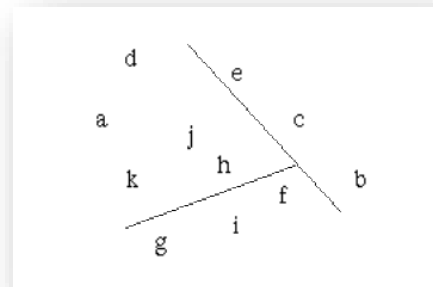
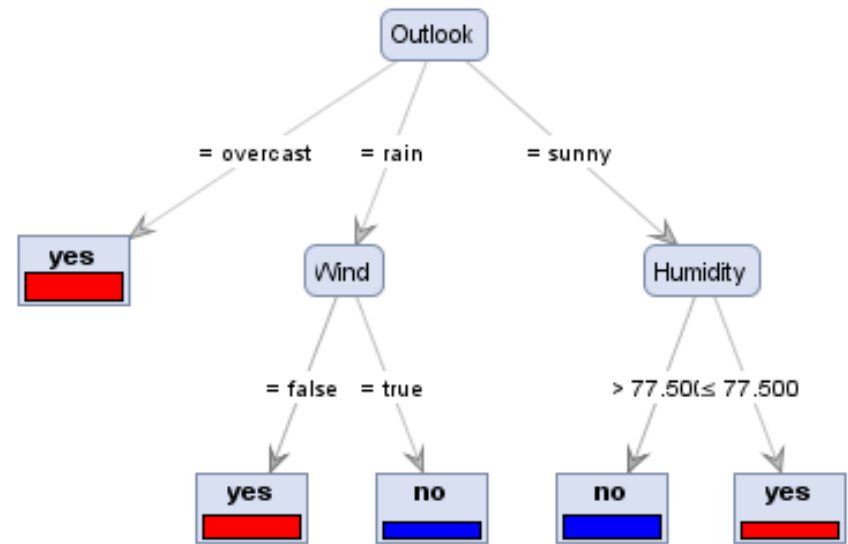
- K-Means, K-Medoids, Self-Organizing Map (SOM), Fuzzy C-Means, etc

5. Association (Asosiasi):

- FP-Growth, A Priori, Coefficient of Correlation, Chi Square, etc

Output/Pola/Model/Knowledge

1. Formula/**Function** (Rumus atau Fungsi Regresi)
 - $\text{WAKTU TEMPUH} = 0.48 + 0.6 \text{ JARAK} + 0.34 \text{ LAMPU} + 0.2 \text{ PESANAN}$
2. Decision **Tree** (Pohon Keputusan)
3. Tingkat **Korelasi**
4. **Rule** (Aturan)
 - IF $\text{ips3}=2.8$ THEN lulustepatwaktu
5. **Cluster** (Klaster)



Latihan

1. Sebutkan **5 peran utama** data mining!
2. Jelaskan perbedaan **estimasi** dan **prediksi**!
3. Jelaskan perbedaan **prediksi** dan **klasifikasi**!
4. Jelaskan perbedaan **klasifikasi** dan **klustering**!
5. Jelaskan perbedaan **klustering** dan **association**!
6. Jelaskan perbedaan **estimasi** dan **klasifikasi**!
7. Jelaskan perbedaan **estimasi** dan **klustering**!
8. Jelaskan perbedaan **supervised** dan **unsupervised** learning!
9. Sebutkan **tahapan utama proses** data mining!

1.3 Sejarah dan Penerapan Data Mining

Evolution of Sciences

- Before 1600: **Empirical science**
- 1600-1950s: **Theoretical science**
 - Each discipline has grown a *theoretical component*
 - Theoretical models *motivate experiments* and generalize understanding
- 1950s-1990s: **Computational science**
 - Most disciplines have grown a third, *computational branch* (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
 - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form *solutions for complex mathematical models*
- 1990-now: **Data science**
 - The *flood of data* from new scientific instruments and simulations
 - The ability to economically store and manage petabytes of data online
 - The Internet makes all these archives universally accessible
 - *Data mining is a major new challenge!*

*Jim Gray and Alex Szalay, The World Wide Telescope:
An Archetype for Online Science, Comm. ACM, 45(11): 50-54, Nov. 2002*

Contoh Penerapan Data Mining

- Penentuan kelayakan aplikasi peminjaman uang di bank
- Penentuan pasokan listrik PLN untuk wilayah Jakarta
- Prediksi profile tersangka koruptor dari data pengadilan
- Perkiraan harga saham dan tingkat inflasi
- Analisis pola belanja pelanggan
- Memisahkan minyak mentah dan gas alam
- Menentukan kelayakan seseorang dalam kredit KPR
- Penentuan pola pelanggan yang loyal pada perusahaan operator telepon
- Deteksi pencucian uang dari transaksi perbankan
- Deteksi serangan (*intrusion*) pada suatu jaringan