PERTEMUAN 5

Data warehouse and Business Intelligence

Pokok Bahasan

Pertemuan Ke-	Pokok Bahasan
1	Business Intelligence
2	Data Warehousing
3	Business Performance Management
4	Business Performance Management Methodologies
5	Pengantar Data Mining
6	Metode Learning Algoritma Data Mining
7	Review dan Quiz
8	UTS
9	Studi Kasus
10-15	Presentasi Tugas Kelompok
16	UAS

Rencana Pembelajaran

Tugas Kelompok

- ✓ Buat Kelompok maximal 4 orang/kelompok.
- ✓ Pengolahan data menggunakan tools rapidminer
- ✓ Menggunakan salahsatu metode data mining
- ✓ Dataset menggunakan data public atau private, setiap kelompok beda dataset
- ✓ Kumpulkan tugas kelompok tersebut berupa : makalah, dan powerpoint pada pertemuan 10 dan bisa dipresentasikan.
- ✓ Mengumpulkan draft artikel ilmiah.
- ✓ Nilai project & presentasi akan menjadi nilai kelompok, keaktifan dan nilai penguasaan materi.

Data warehouse dan Business Intelligence
(9th Ed., Prentice Hall)

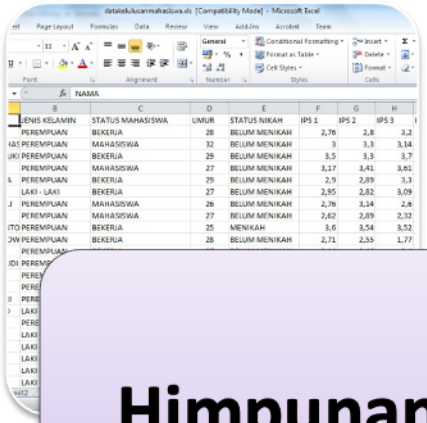
Chapter 5:

Pengantar Data Mining

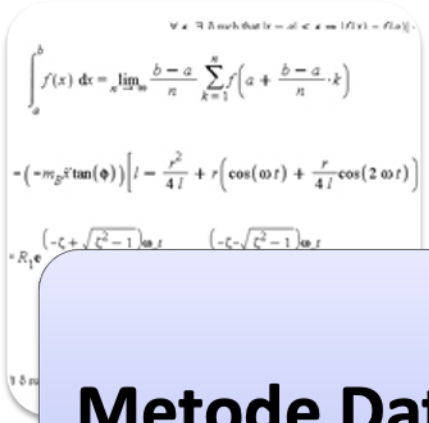
Apa itu Data Mining?

- 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
  - Big data

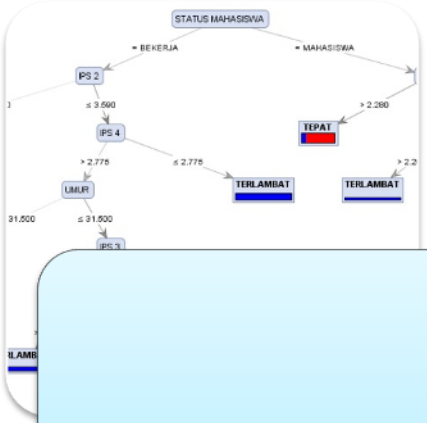
Apa itu Data Mining?



Himpunan Data



Metode Data Mining



Pengetahuan

Contoh Data di Kampus

- Puluhan ribu data mahasiswa di kampus yang diambil dari sistem informasi akademik
- Apakah pernah kita ubah menjadi pengetahuan yang lebih bermanfaat? TIDAK!
- Seperti apa pengetahuan itu? Rumus, Pola, Aturan

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

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	

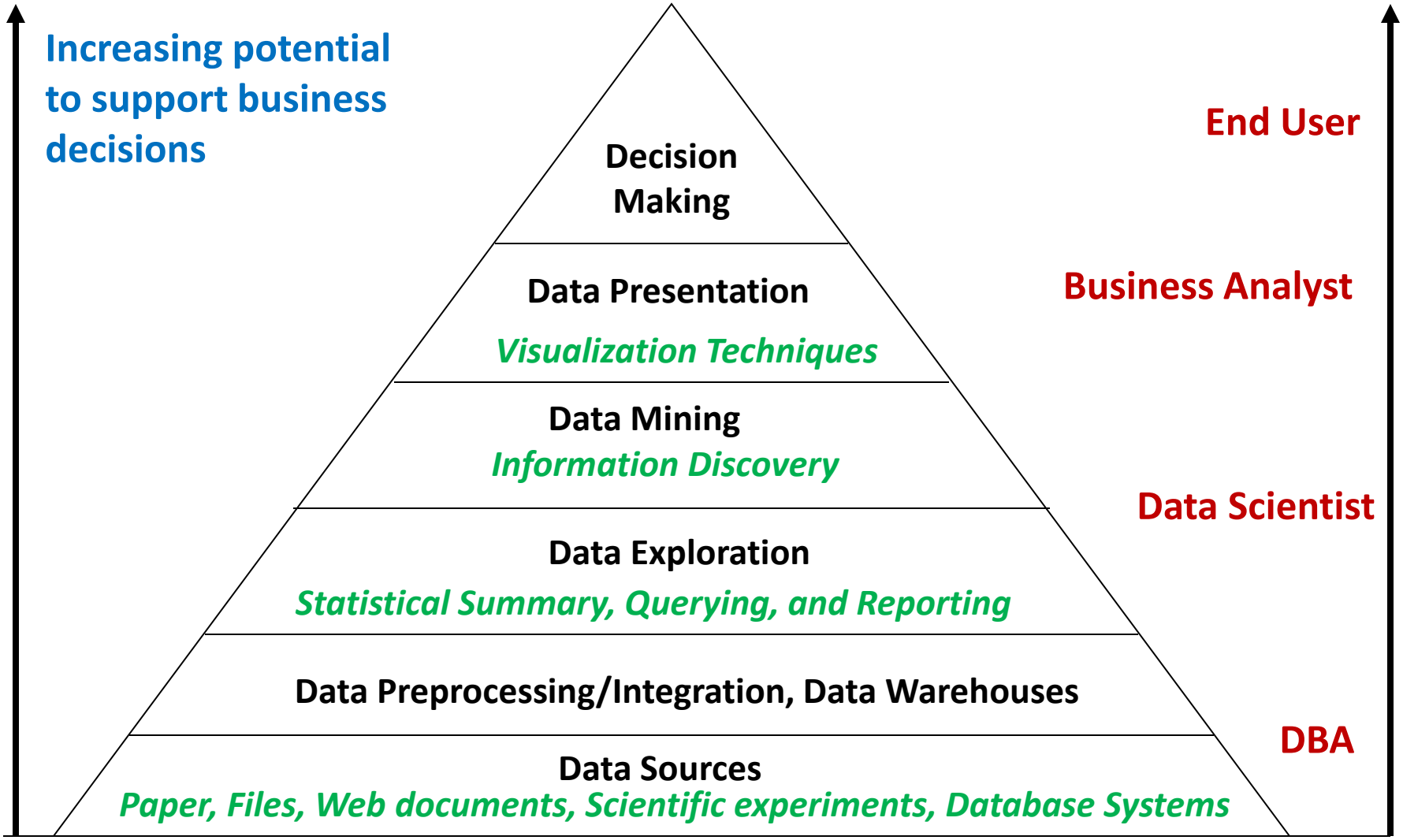
Informasi Akumulasi Bulanan Kehadiran Pegawai

Data - Informasi – Pengetahuan

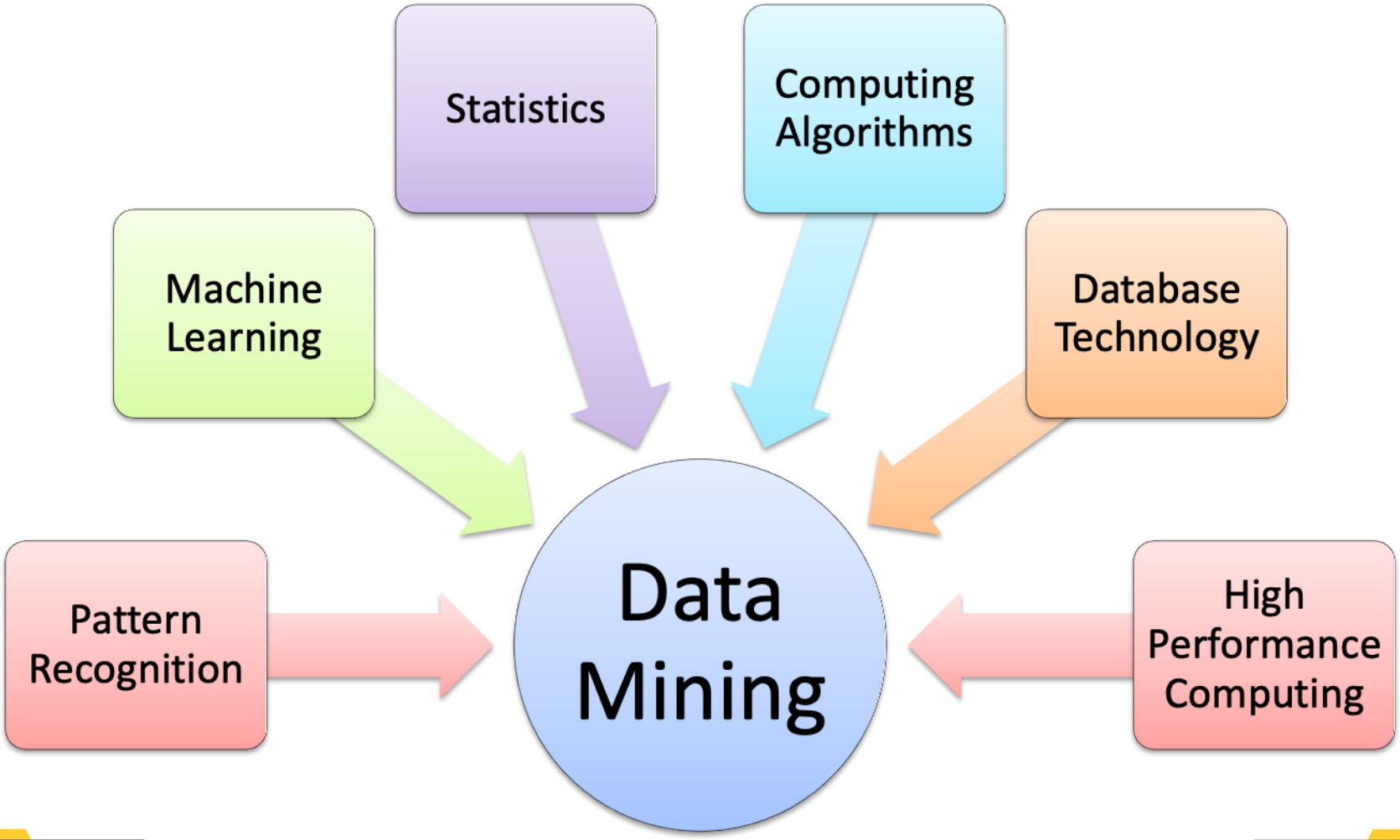
	Senin	Selasa	Rabu	Kamis	Jumat
Terlambat	7	0	1	0	5
Pulang Cepat	0	1	1	1	8
Izin	3	0	0	1	4
Alpa	1	0	2	0	2

Pola Kebiasaan Kehadiran Mingguan Pegawai

Data Mining Tasks and Roles



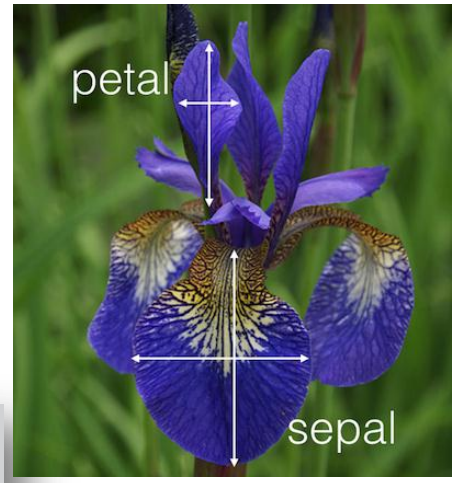
Hubungan Data Mining dan Bidang Lain



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

Dataset (Himpunan Data)



Attribute/Feature/Dimension

Class/Label/Target

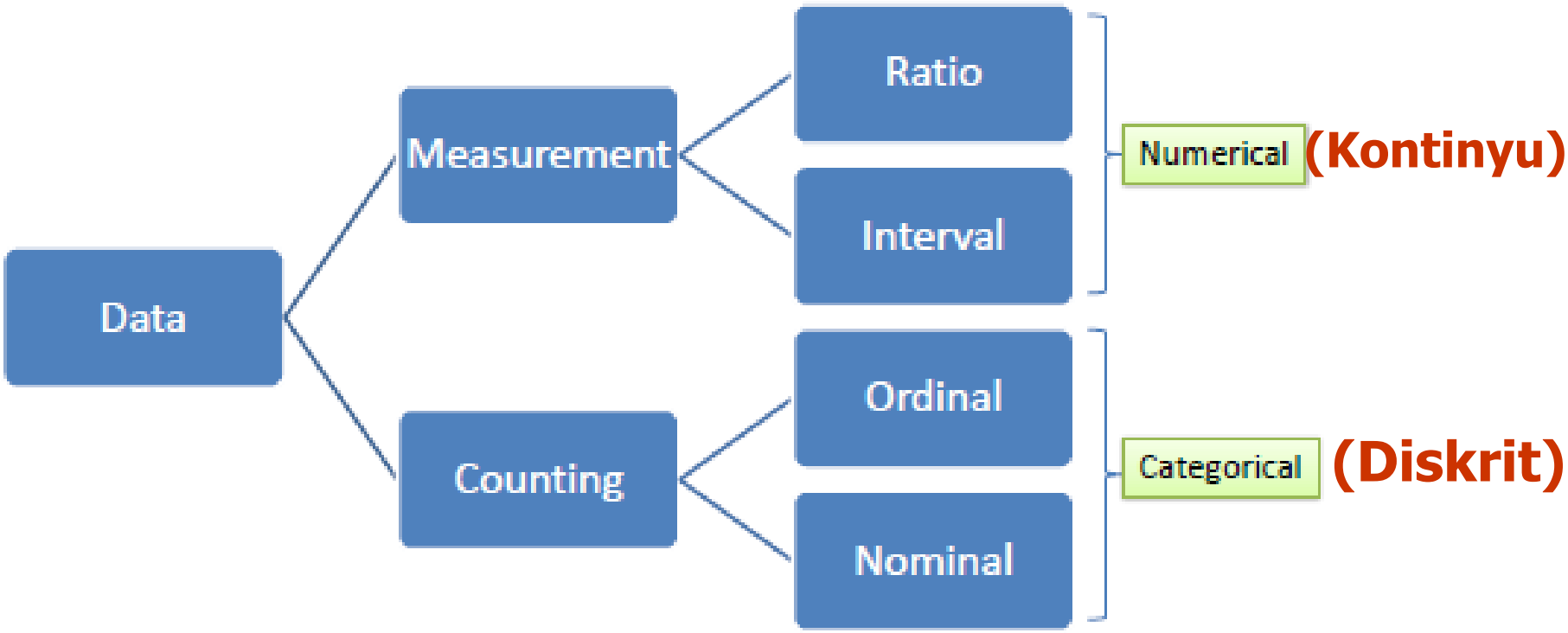
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris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

Record/ Object/ Sample / Tuple/ Data

Nominal

Numerik

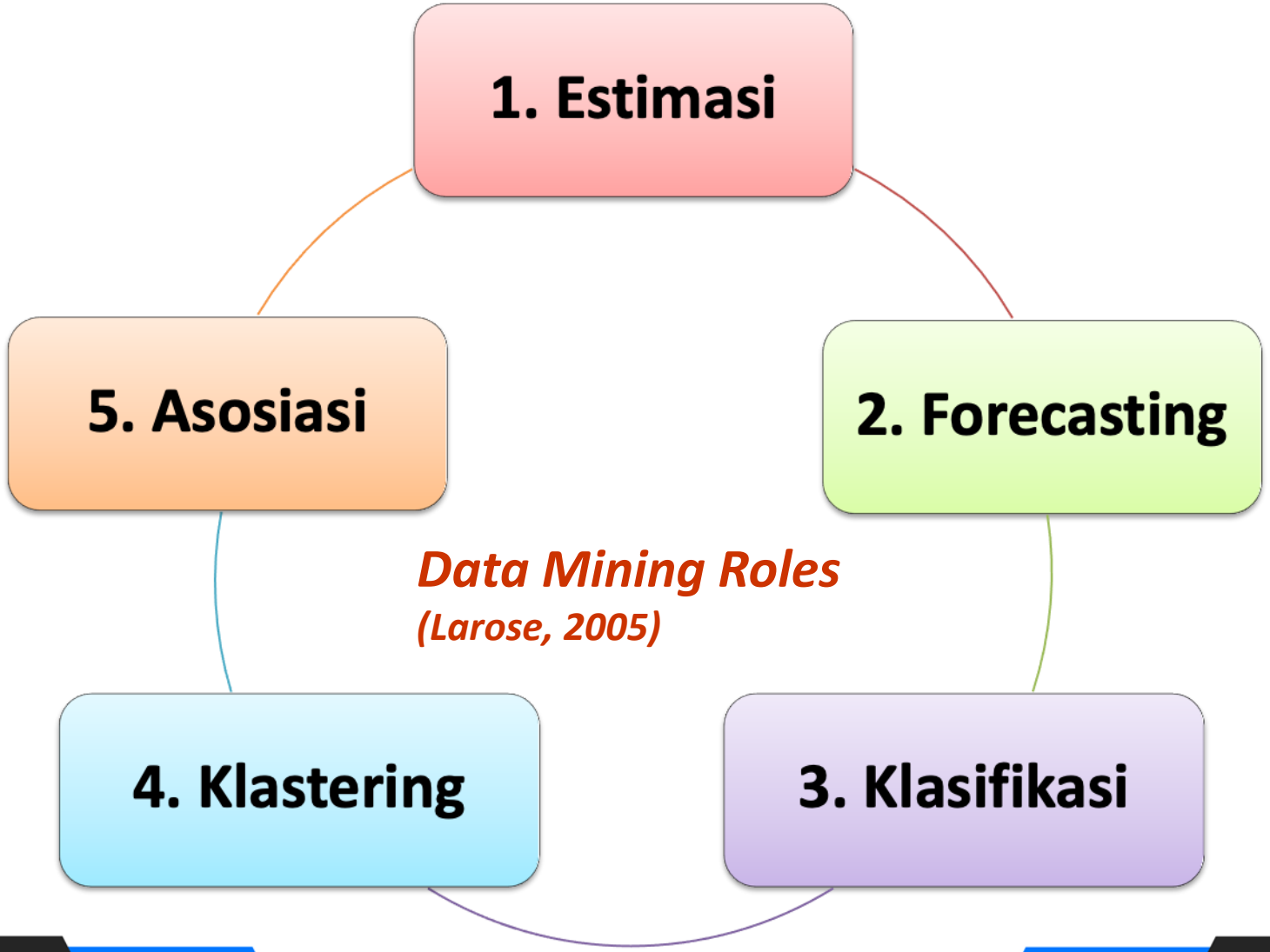
Tipe Data



Tipe Data

Tipe Data	Deskripsi	Contoh	Operasi
Ratio (Mutlak)	• Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui • Mempunyai titik nol yang absolut (*, /)	• Umur • Berat badan • Tinggi badan • Jumlah uang	geometric mean, harmonic mean, percent variation
Interval (Jarak)	• Data yang diperoleh dengan cara pengukuran, dimana jarak dua titik pada skala sudah diketahui • Tidak mempunyai titik nol yang absolut (+, -)	• Suhu 0°c-100°c, • Umur 20-30 tahun	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ordinal (Peringkat)	• Data yang diperoleh dengan cara kategorisasi atau klasifikasi • Tetapi diantara data tersebut terdapat hubungan atau berurutan (<, >)	• Tingkat kepuasan pelanggan (puas, sedang, tidak puas)	median, percentiles, rank correlation, run tests, sign tests
Nominal (Label)	• Data yang diperoleh dengan cara kategorisasi atau klasifikasi • Menunjukkan beberapa object yang berbeda (=, ≠)	• 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

Label

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

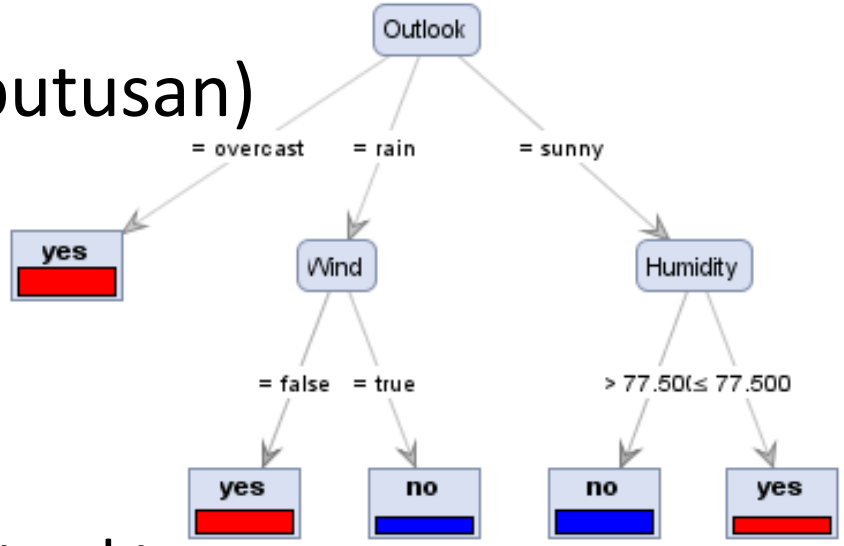
Pembelajaran dengan
Metode Estimasi (*Regresi Linier*)

Waktu Tempuh (T) = 0.48P + 0.23TL + 0.5J

Pengetahuan

Output/Pola/Model/Knowledge

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



2. Forecasting Harga Saham

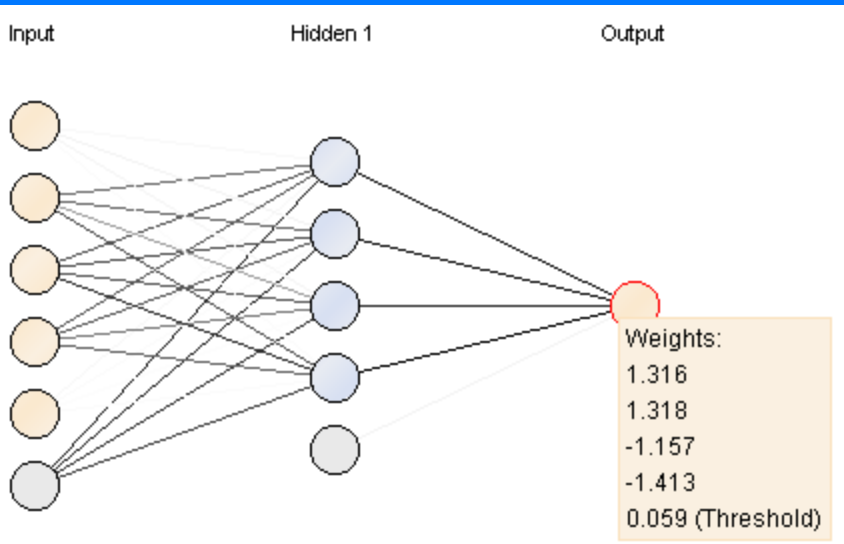
Label Time Series

Row No.	Close	Date	Open	High	Low	Volume
1	1286.570	Apr 11, 2006	1296.600	1300.710	1282.960	223288000(
2	1288.120	Apr 12, 2006	1286.570	1290.930	1286.450	193810000(
3	1289.120	Apr 13, 2006	1288.120	1292.090	1283.370	189194000(
4	1285.330	Apr 17, 2006	1289.120	1292.450	1280.740	179465000(
5	1307.280	Apr 18, 2006	1285.330	1309.020	1285.330	259544000(
6	1309.930	Apr 19, 2006	1307.650	1310.390	1302.790	244731000(
7	1311.460	Apr 20, 2006	1309.930	1318.160	1306.380	251292000(
8	1311.280	Apr 21, 2006	1311.460	1317.670	1306.590	239263000(
9	1308.110	Apr 24, 2006	1311.280	1311.280	1303.790	211733000(
10	1301.740	Apr 25, 2006	1308.110	1310.790	1299.170	236638000(
11	1305.410	Apr 26, 2006	1301.740	1310.970	1301.740	250269000(
12	1309.720	Apr 27, 2006	1305.410	1315	1295.570	277201000(
13	1310.610	Apr 28, 2006	1309.720	1316.040	1306.160	241992000(

Dataset harga saham dalam bentuk time series (rentet waktu)

Pembelajaran dengan Metode Forecasting (*Neural Network*)

Pengetahuan berupa Rumus Neural Network



Prediction Plot

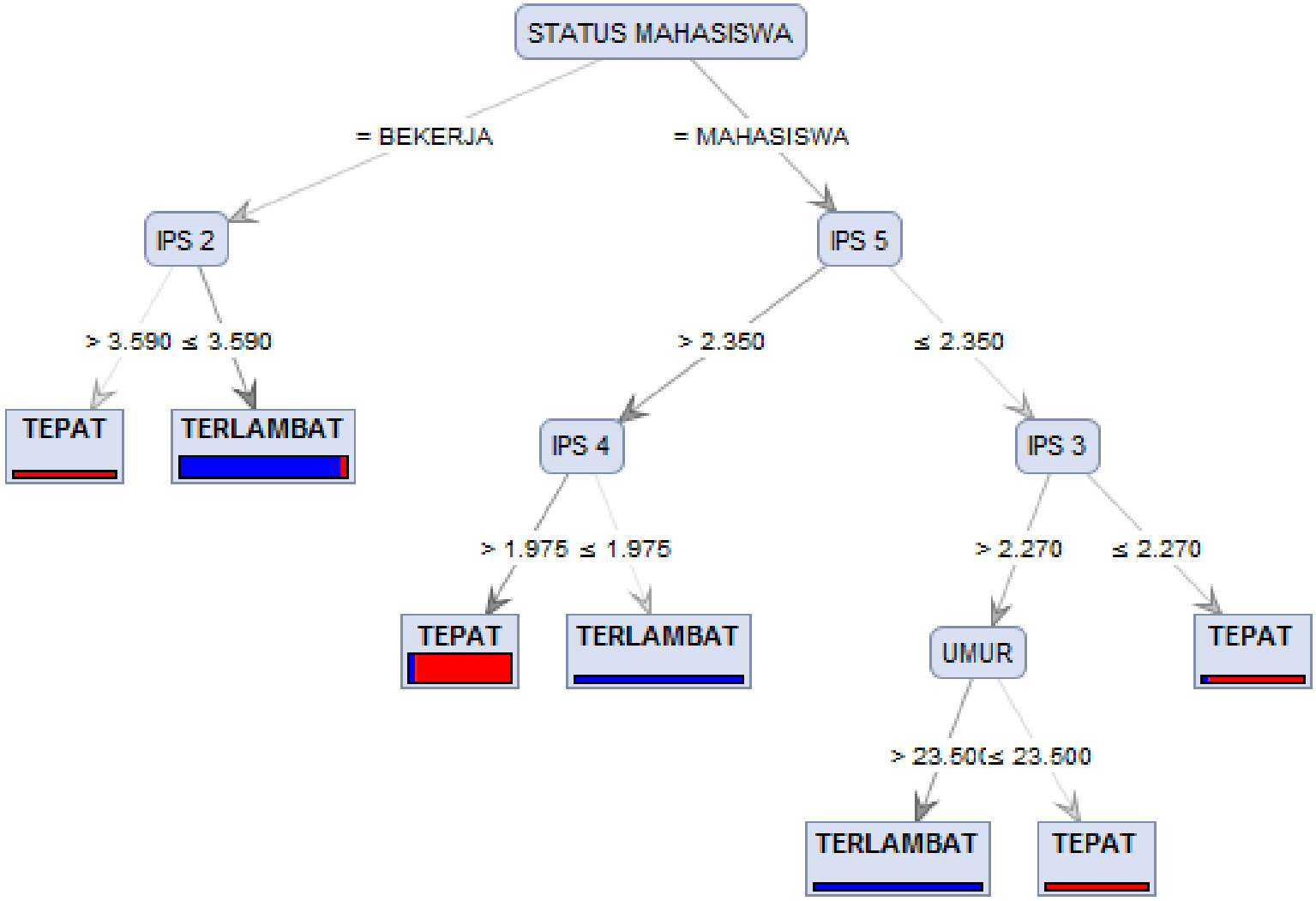
3. Klasifikasi Kelulusan Mahasiswa

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



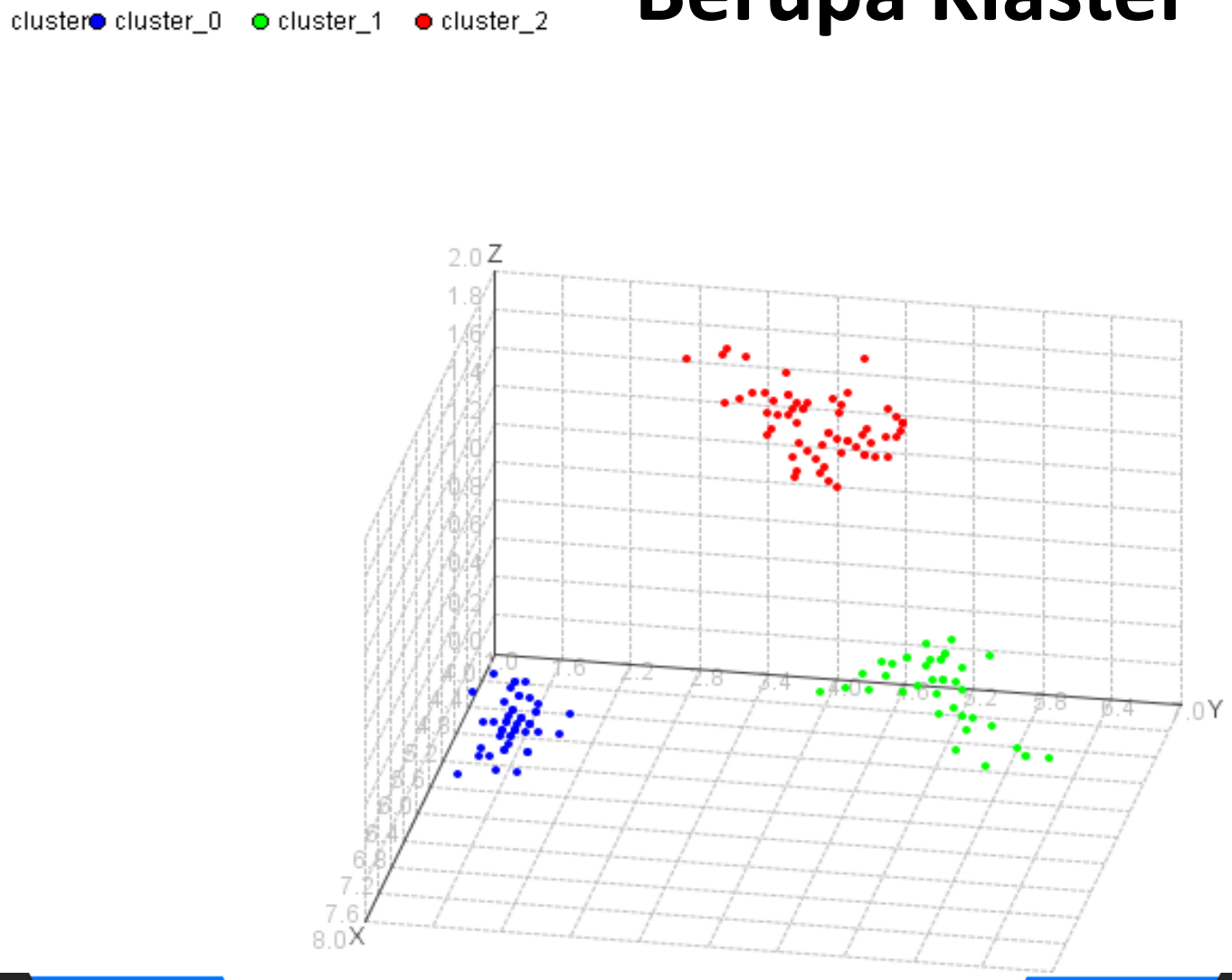
4. Klastering Bunga Iris

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 (Model) Berupa Klaster



Klastering Jenis Pelanggan

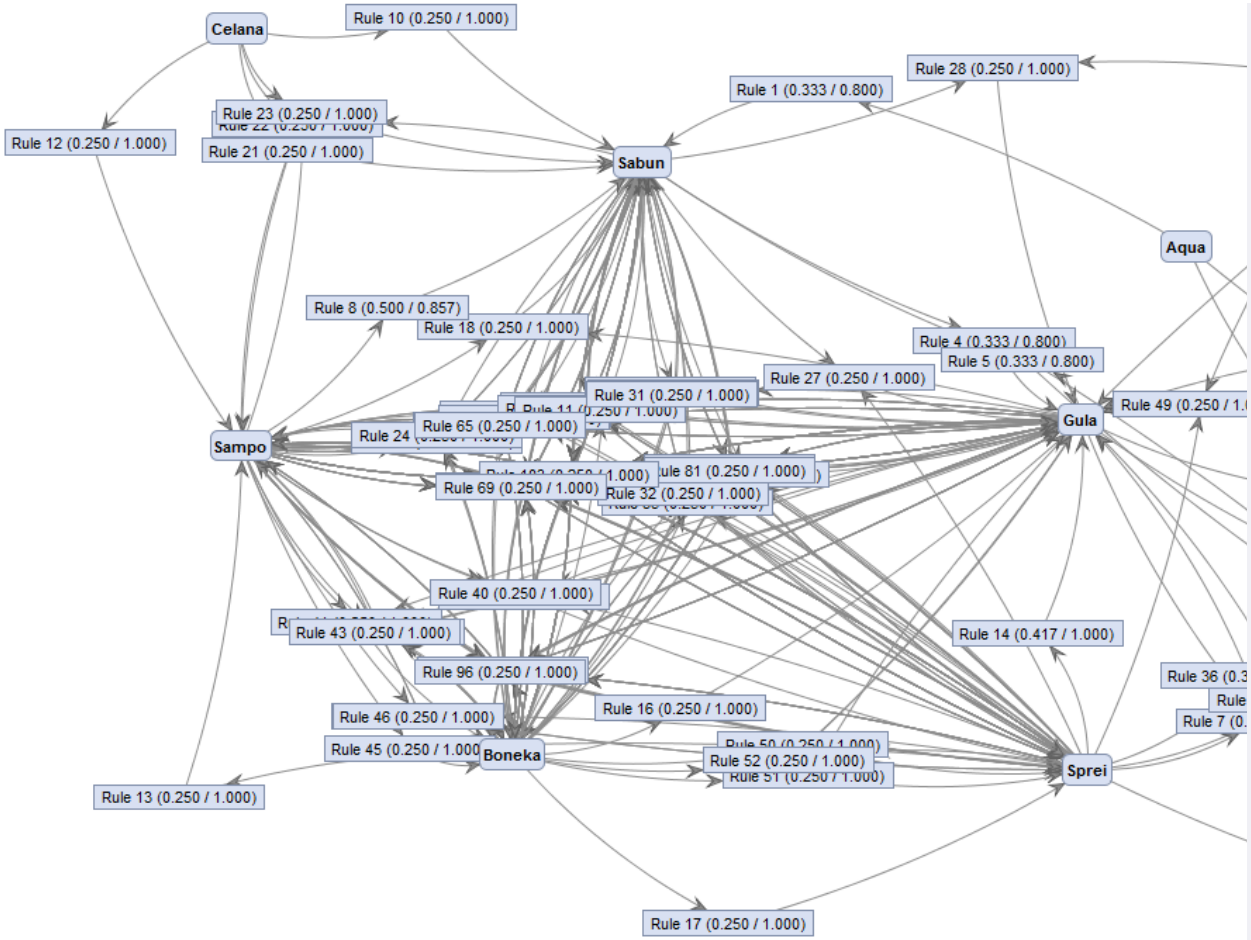
5. Aturan Asosiasi Pembelian Barang

ExampleSet (12 examples, 0 special attributes, 10 regular attributes)

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)
[Sabun, Popok] --> [Gula] (confidence: 1.000)
[Boneka] --> [Sabun, Gula] (confidence: 1.000)
[Sabun, Boneka] --> [Gula] (confidence: 1.000)
[Gula, Boneka] --> [Sabun] (confidence: 1.000)
[Sabun, Sprei] --> [Boneka] (confidence: 1.000)
[Boneka] --> [Sabun, Sprei] (confidence: 1.000)
[Sabun, Boneka] --> [Sprei] (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 supermaket 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