DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi

DATA MINING INTRODUCTION

Nizar Rabbi Radliya
nizar@email.unikom.ac.id

SILABUS

Data Mining Introduction

Preprocessing

Similarity

Association

UTS

Classification & Prediction

Clustering

UAS

REVIEW

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

REVIEW

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 Kehadiran Pegawai Per Bulan

REVIEW

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 Kehadiran Mingguan Pegawai

REVIEW

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 ?

“We are drowning in data, but starving for knowledge"

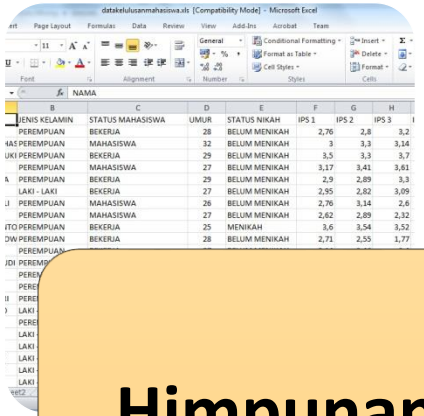
DATA MINING ?

Tan (2006) mendefinisikan data mining sebagai proses untuk mendapatkan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi baru yang diambil dari bongkahan data besar yang membantu dalam pengambilan keputusan.

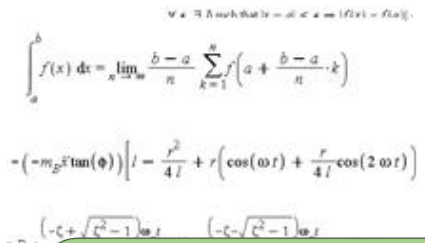
Darly Pregibon (2011) menyatakan bahwa data mining adalah campuran dari statistik, kecerdasan buatan, dan riset basis data.

Pramudiono (2006) mengartikan data mining sebagai serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual.

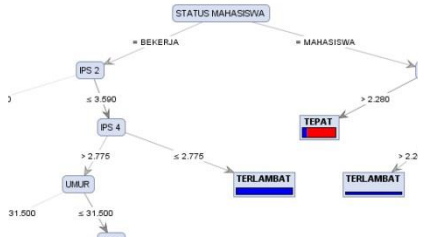
DATA MINING ?



Himpunan Data



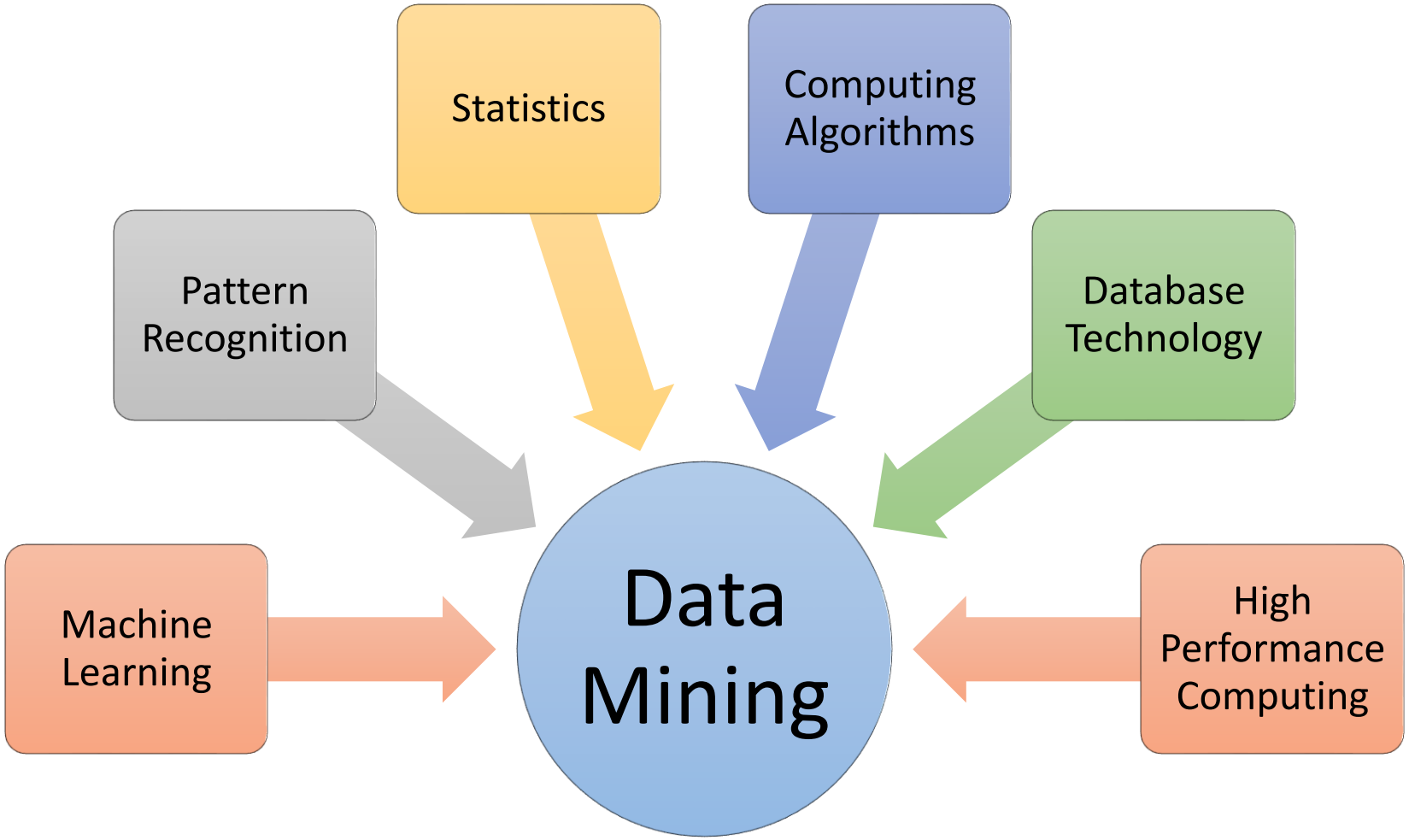
Metode Data Mining



Pengetahuan

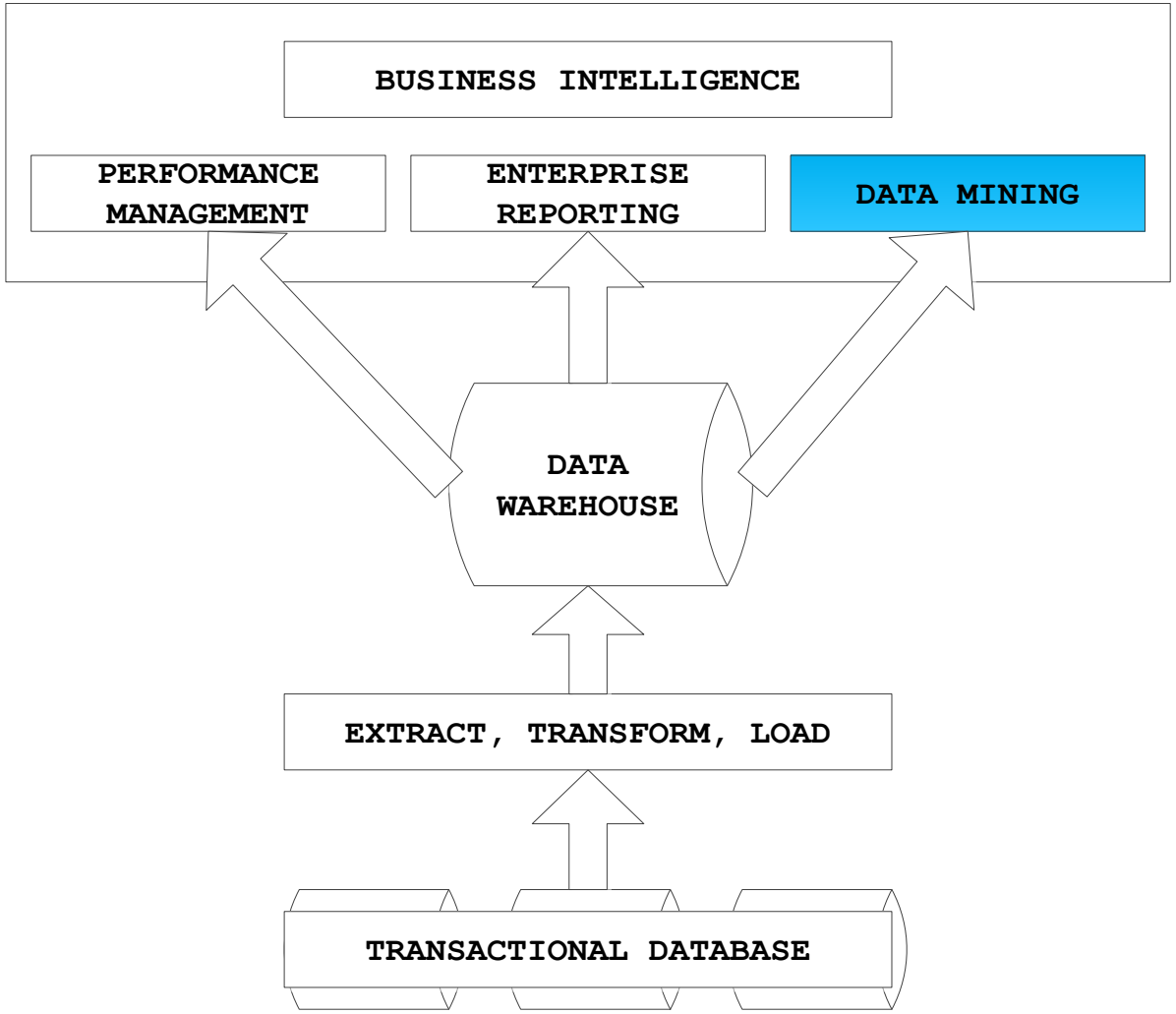
Proses ekstraksi dari DATA ke PENGETAHUAN (pola, rumus, aturan, model) dengan beberapa teknik dari kumpulan data besar.

DATA MINING ?



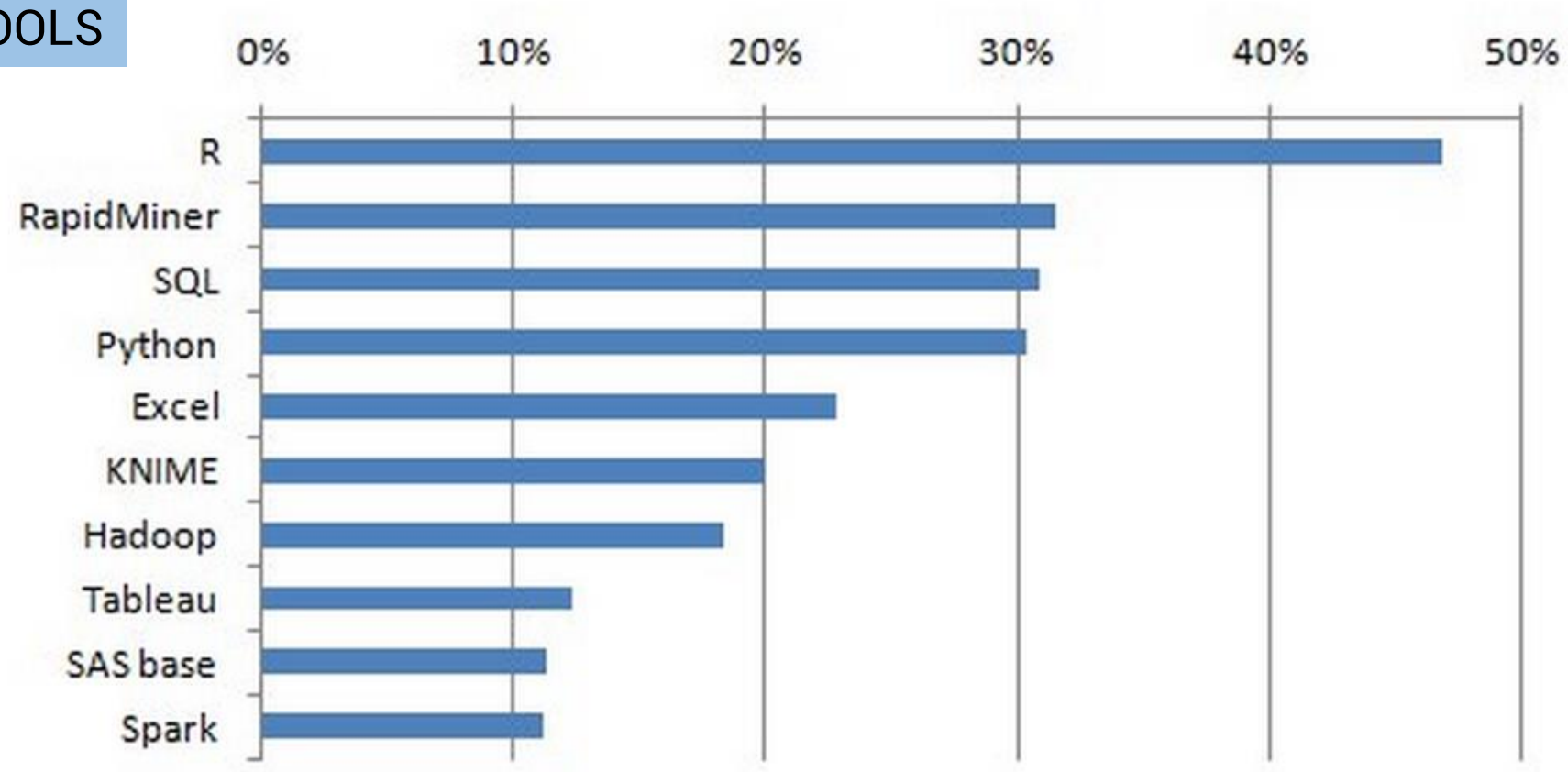
Hubungan dengan Bidang lainnya

DATA MINING ?



Hubungan dengan Bidang lainnya

DATA MINING TOOLS



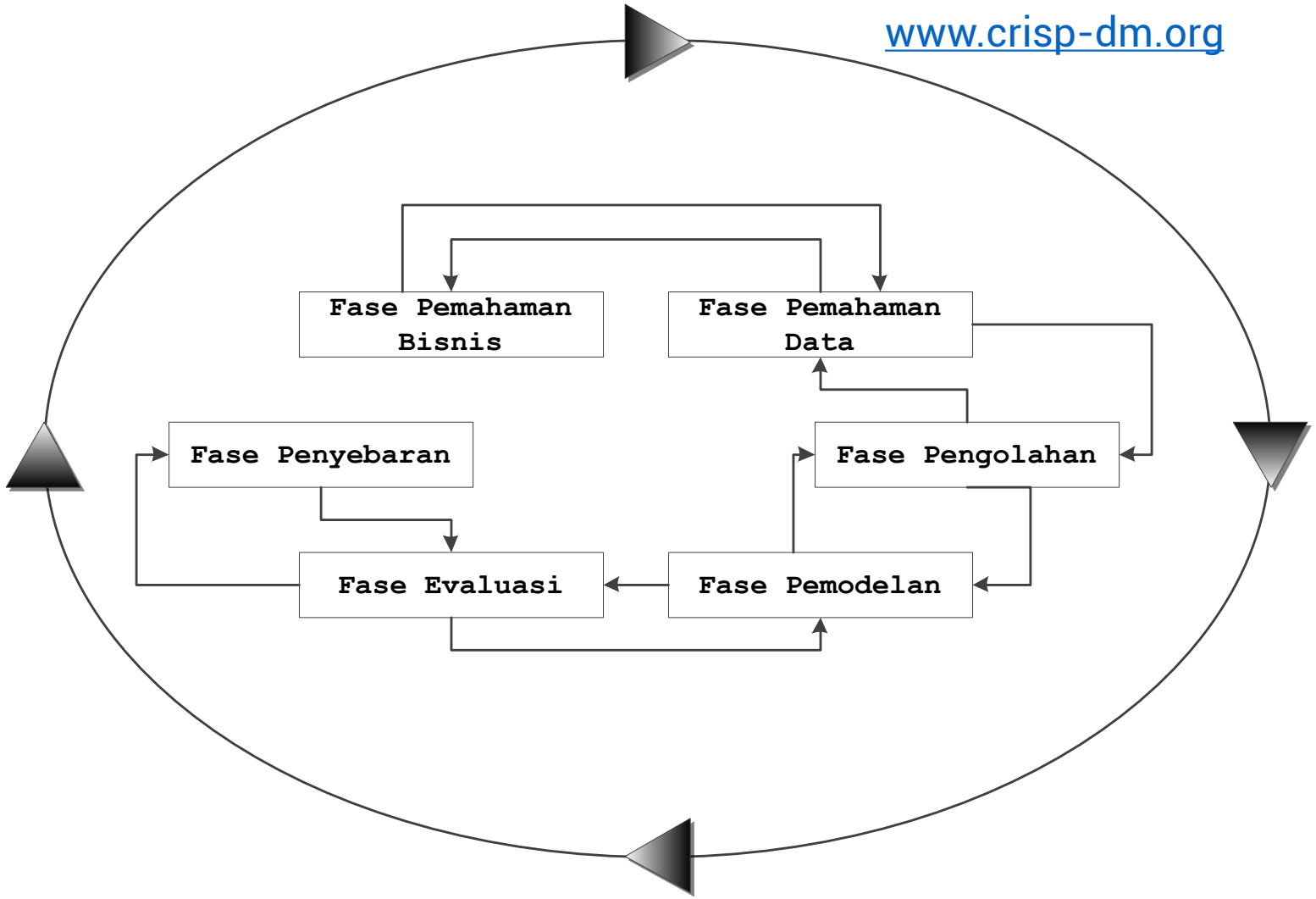
http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html

Penggunaan Software Data Mining

DATA MINING PROCESSES

Cross-Industry Standard Process for Data Mining (CRISP-DM)

www.crisp-dm.org

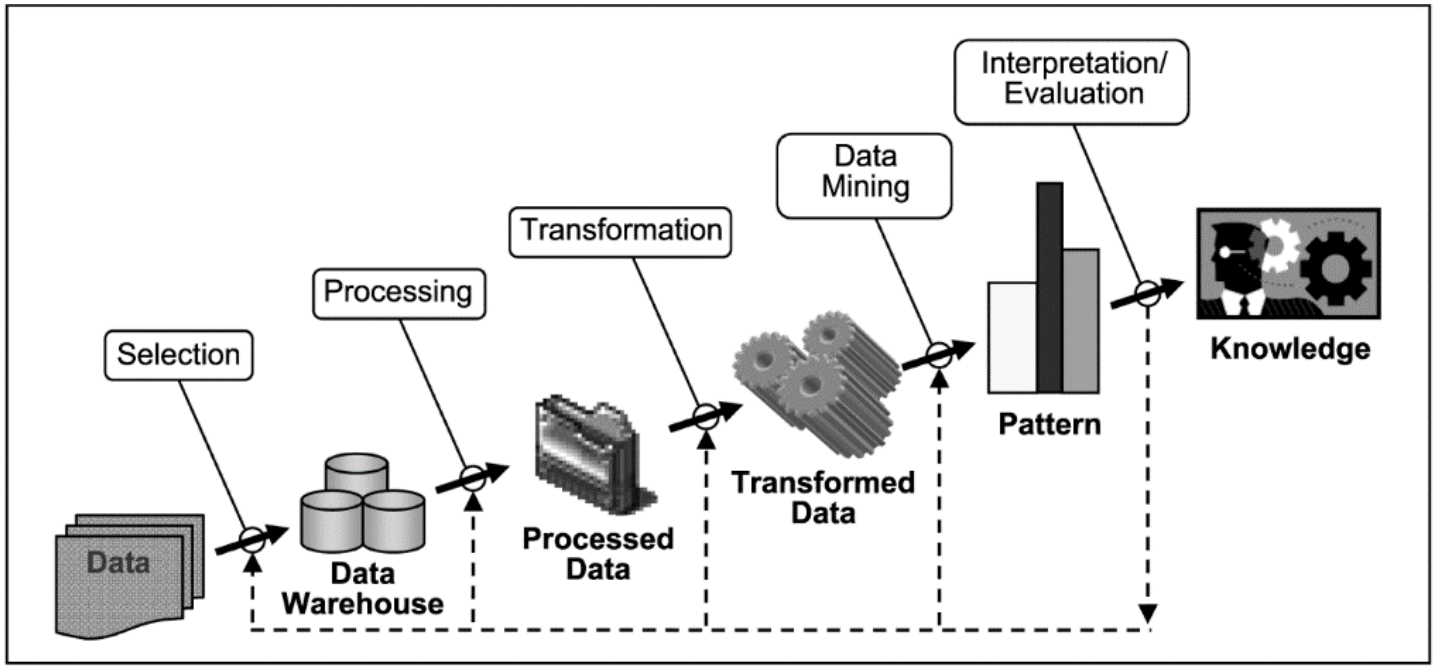


DATA MINING PROCESSES

Tiga langkah utama dalam proses data mining (Gonunescu, 2011)

1. Ekplorasi/pemrosesan awal data
2. Membangun model dan melakukan validasi terhadapnya
3. Penerapan

Peran Data Mining dalam Knowledge Discovery in Database (KDD)



DATA MINING TECHNIQUES

Classification,

Clustering,

Association,

Anomaly,

Prediction,

Estimation

Regression,

Suquatial Pattern,

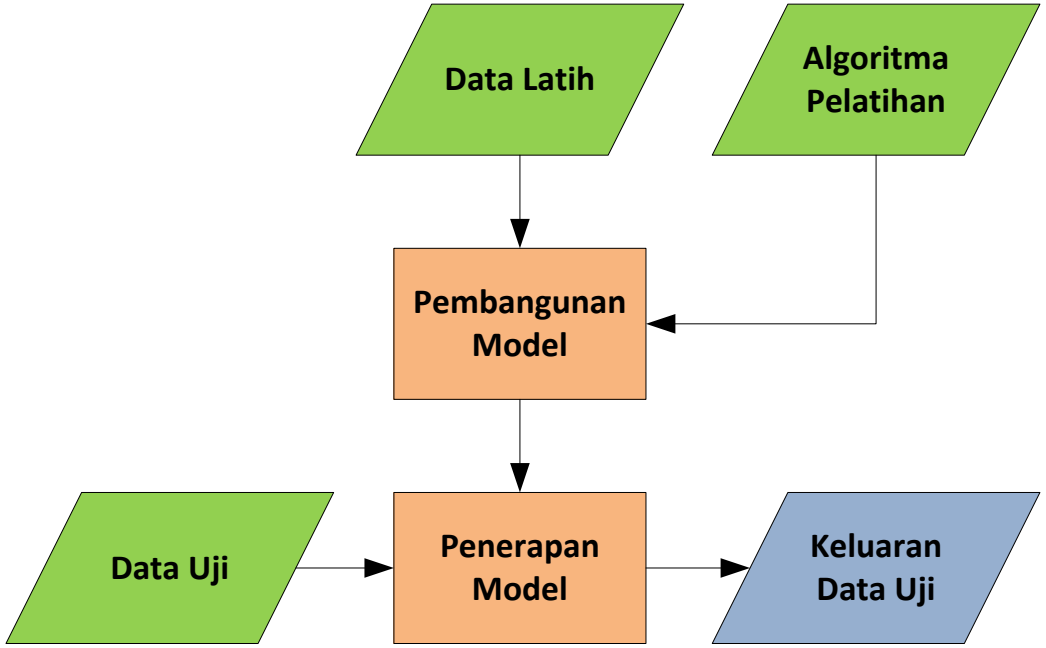
Deviation Detection

Dll

C4.5, Nearest Neighbor, A Priori, Fuzzy C-Means, Bayesian Classification, C4.5, K-Means, SVM, EM, PageRank, AdaBoost, kNN, CART, dll

CLASSIFICATION

Kasifikasi (classification) digunakan untuk pembuatan model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan variabel yang baru.



Algoritma:

Decision Tree Induction (C4.5)

Nearest-Neighbor

Bayesian Classification

Neural Network

Model Evaluation and Selection

Techniques to Improve Classification Accuracy: Ensemble Methods

dll

CLASSIFICATION

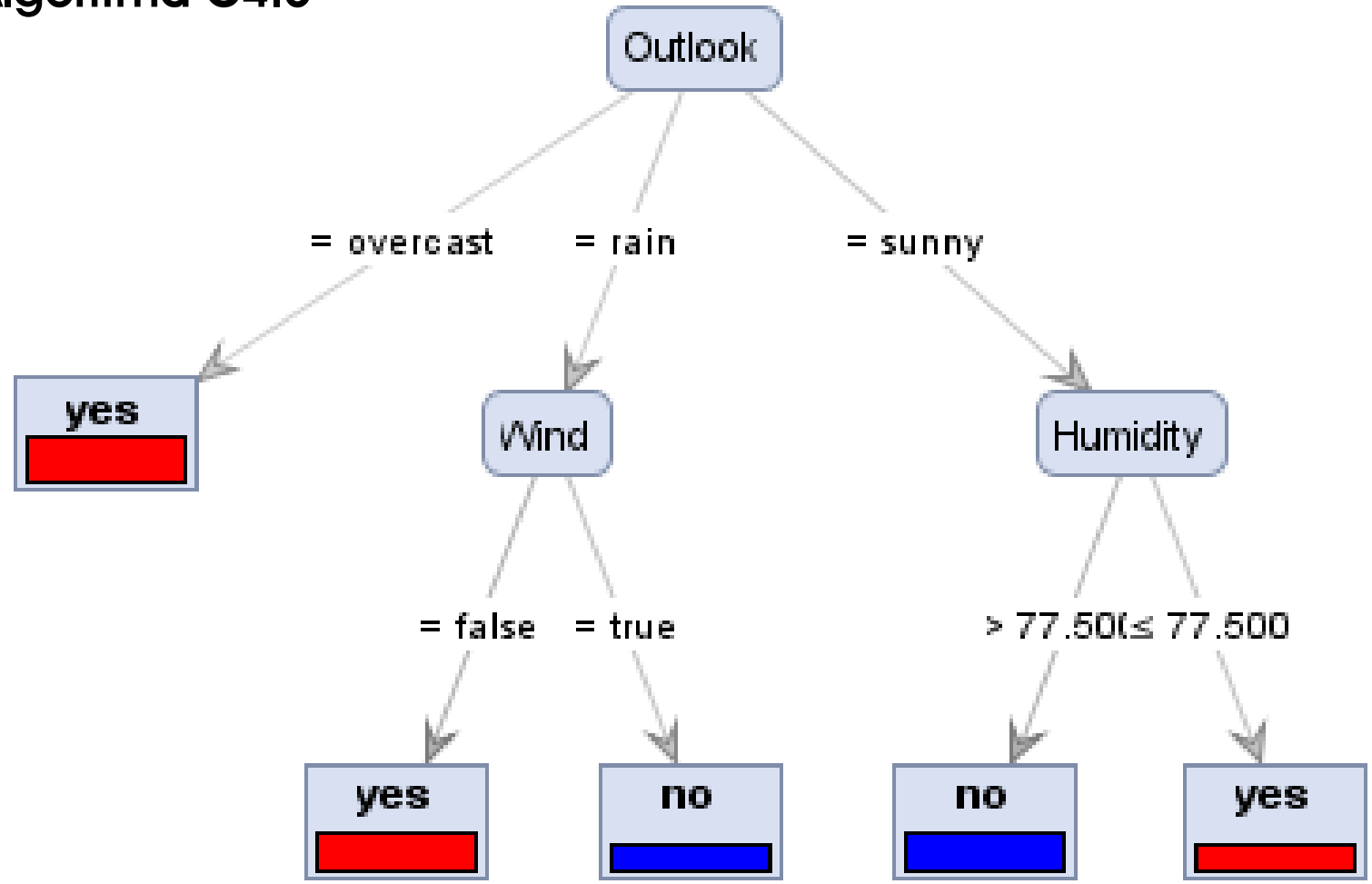
Data Keputusan Bermain Tenis (data set)

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

CLASSIFICATION

Pohon Keputusan Bermain Tenis (model)

Algoritma C4.5



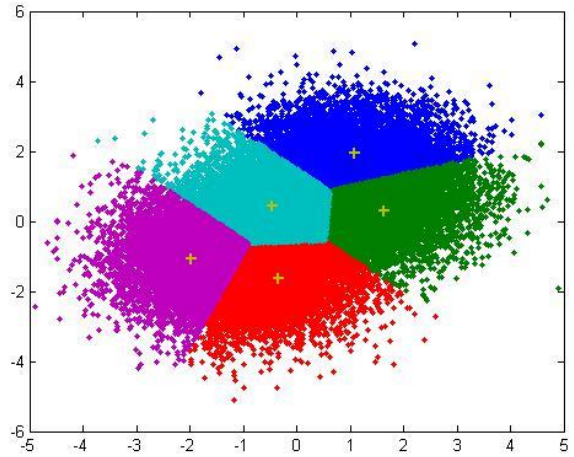
CLASSIFICATION

Seleksi Kondisi untuk Rekomendasi Bermain Tenis (rules)

```
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
```

CLUSTERING

Penklusteran (clustering) digunakan untuk melakukan pengelompokan data-data ke dalam sejumlah kelompok (cluster) berdasarkan karakteristik masing-masing data pada kelompok-kelompok yang ada.



Algoritma:

K-Means

Cluster Analysis: Basic Concepts

Partitioning Methods

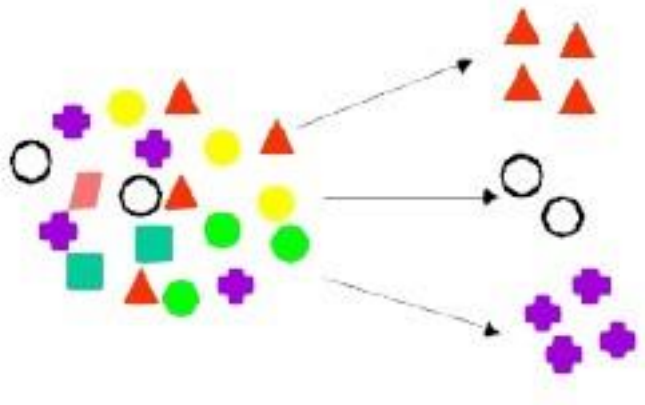
Hierarchical Methods

Density-Based Methods

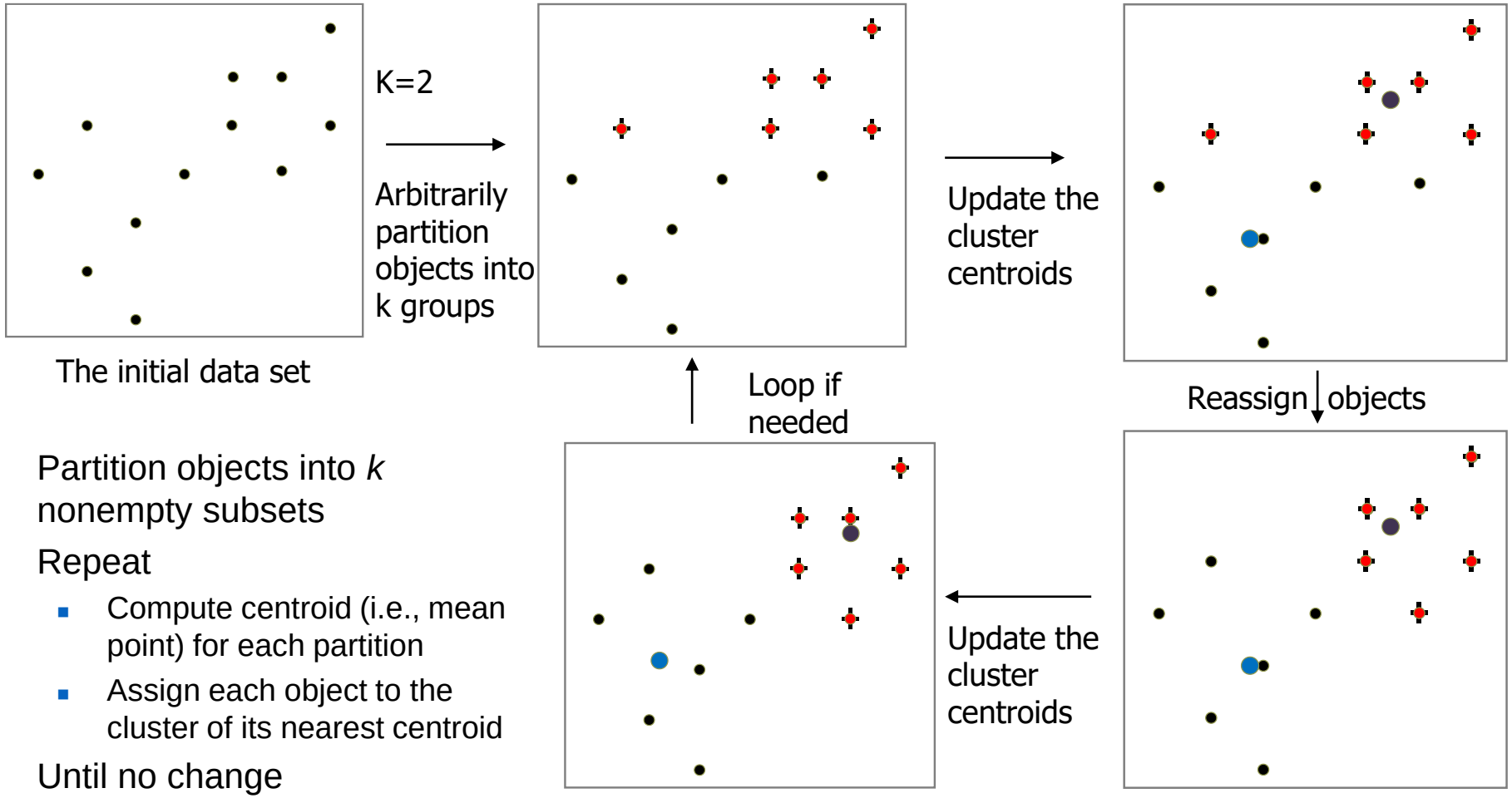
Grid-Based Methods

Evaluation of Clustering

dll



CLUSTERING



- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

CLUSTERING

Menentukan Wilayah Pengembangan Akses Air Bersih

Himpunan Data yang digunakan …

	A	B	C	D
1	AREA	CWA1	CWA2	CWA3
2	Campaka	46.52	45.52	22.76
3	Ciroyom	34.57	33.57	16.78
4	Dungus Cariang	63.32	62.32	31.16
5	Garuda	51.51	50.51	25.25
6	Kebon Jeruk	95.11	94.11	47.05
7	Maleber	37.02	36.02	18.01
8	Antapani Kidul	71.90	70.90	35.45
9	Antapani Kulon	95.47	94.47	47.23
10	Antapani Tengah	48.40	47.40	23.70
11	Antapani Wetan	99.78	98.78	49.39
12	Cisaranten Bina Harapan	88.31	87.31	43.65
13	Sukamiskin			
14	Cibadak	76.45	75.45	37.73
15	Karang Anyar	98.88	97.88	48.94
16	Nyengseret	58.78	57.78	28.89
17	Pelindung Hewan	72.69	71.69	35.84
18	Babakan	39.83	38.83	19.42
19	Babakan Ciparay	90.64	89.64	44.82
20	Cirangrang	77.67	76.67	38.34

CLUSTERING

Menentukan Wilayah Pengembangan Akses Air Bersih

Pemodelan menggunakan algoritma K-Means …

1. Menentukan jumlah cluster = 3

2. Menentukan nilai centroid dari setiap cluster

3. Petakan setiap data pada centroid cluster (cari yang terdekat)

$$D(i,j) = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}$$

dimana:

$D(i,j)$ = Jarak data ke i ke pusat cluster j

X_{ki} = Data ke i pada atribut data ke k

X_{kj} = Titik pusat ke j pada atribut ke k

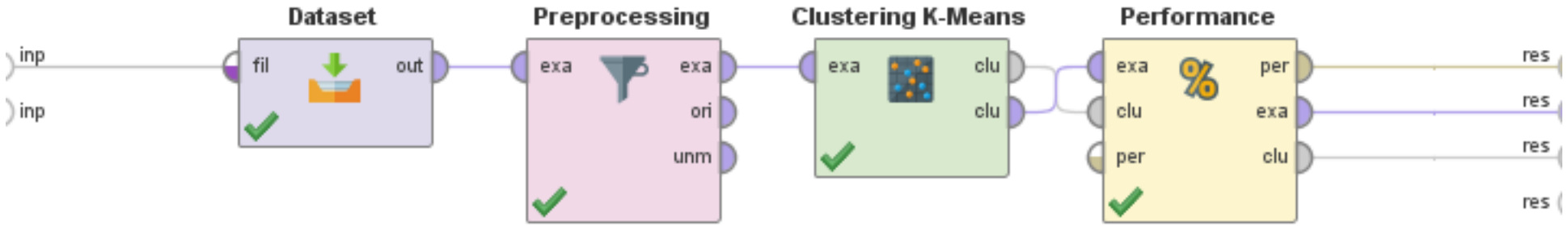
4. Hitung kembali pusat cluster yang baru berdasarkan rata-rata angggota yang ada pada cluster tersebut

5. Setelah didapatkan centroid yang baru dari setiap cluster, lakukan kembali dari langkah ketiga hingga centroid dari setiap cluster tidak berubah lagi dan tidak ada lagi data yang berpindah dari satu cluster ke cluster yang lain

CLUSTERING

Menentukan Wilayah Pengembangan Akses Air Bersih

Pemodelan menggunakan algoritma K-Means …



Cluster	Centroid 1	Centroid 2	Centroid 3	Number of Items	Rekomendasi
Cluster 1	88.34	87.39	43.67	64	Rendah
Cluster 2	61.39	60.39	30.19	12	Sedang
Cluster 3	29.96	28.96	14.48	37	Tinggi

CLUSTERING

Menentukan Strategi Marketing Universitas

Hasil Analisis Clustering

Seluruh wilayah kelurahan yang masuk ke dalam cluster tiga merupakan wilayah prioritas yang direkomendasikan untuk mendapatkan program pengembangan akses air bersih.

Adapun yang wilayah kelurahan yang masuk kedalam cluster tiga diantanya adalah Ciroyom, Maleber, Babakan, Cibangkong, Kebon Waru, Samoja, Jamika, Suka Asih, Mekar Wangi, Cisurupan, Rancabolang, Sindang Jaya.

ASSOCIATION

Asosiasi (association) digunakan untuk menemukan pola yang mendeteksi kumpuluan atribut-atribut yang muncul bersamaan dalam frekuensi yang sering, dan membentuk sejumlah kaidah dari kumpulan-kumpulan tersebut.

Biasa disebut dengan **affinity analysis atau market basket analysis**.

Algoritma:

A Priori

FP-Growth

GRI

dll

Customers who bought this item ...
also bought ...

ASSOCIATION

Data Transaksi (Format Tabular)

Algoritma A Priori

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

ASSOCIATION

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)
[Sprei, Boneka] --> [Sabun] (confidence: 1.000)
```

Contoh, pada hari kamis malam, 1000 pelanggan telah melakukan belanja di supermaket ABC, dimana: 200 orang membeli Teh, dan dari 200 orang yang membeli Teh, 50 orangnya membeli Gula.

Jadi, association rule menjadi, "Jika membeli Teh, maka membeli Gula", dengan nilai **support** = 200/1000*100% = 20% dan nilai **confidence** = 50/200*100% = 25%

ESTIMATION

Estimasi Waktu Pengiriman Pizza

Regresi Linier

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

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

Pengetahuan (Rumus)

NEXT

PREPROCESSING