

# DATA MINING

## 01 Introduction

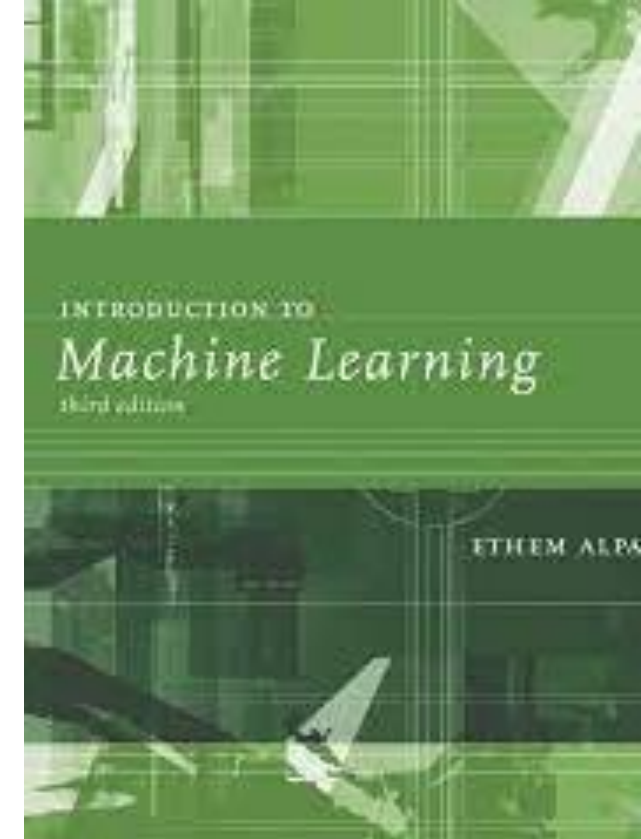
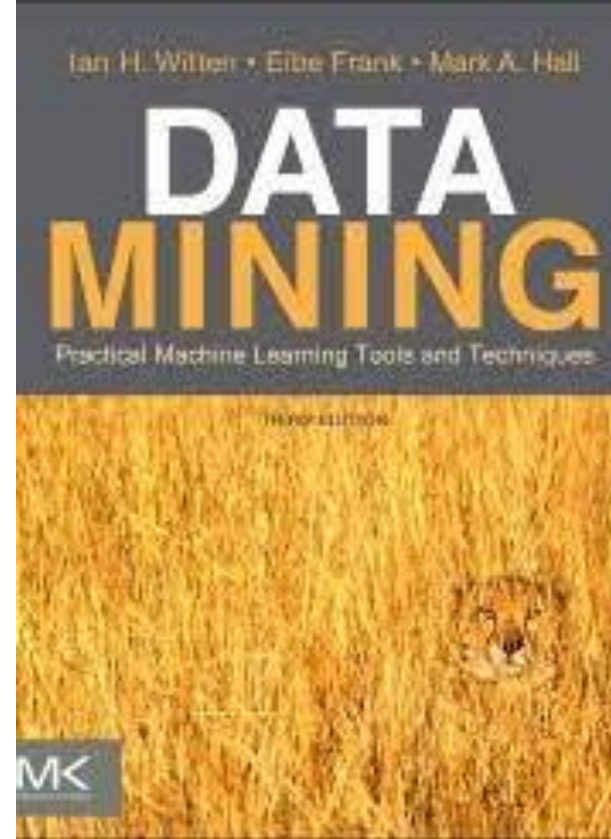
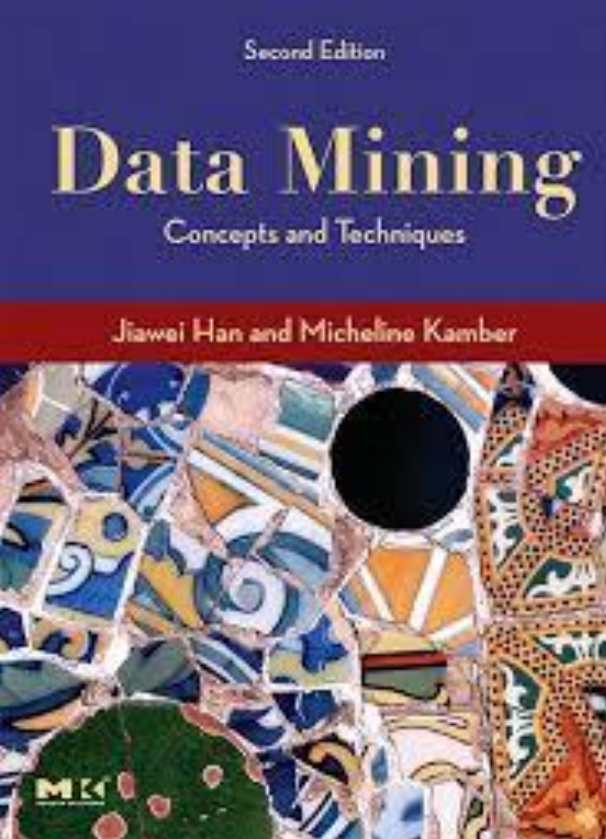
DEWI RAHMAWATI, S.KOM., M.KOM., MOS.

SOFTWARE ENGINEERING - SISTEM INFORMASI – S1

[dewirahmawati@ittelkom-sby.ac.id](mailto:dewirahmawati@ittelkom-sby.ac.id)

# Daftar Pustaka

- Jiawei Han and Micheline Kamber, **Data Mining: Concepts and Techniques Third Edition**, *Elsevier*, 2012
- Ian H. Witten, Frank Eibe, Mark A. Hall, **Data mining: Practical Machine Learning Tools and Techniques 3rd Edition**, *Elsevier*, 2011
- Ethem Alpaydin, **Introduction to Machine Learning**, 3rd ed., *MIT Press*, 2014
- Florin Gorunescu, **Data Mining: Concepts, Models and Techniques**, *Springer*, 2011



## Daftar Pustaka

- Jiawei Han and Micheline Kamber, **Data Mining: Concepts and Techniques Third Edition**, Elsevier, 2012
- Ian H. Witten, Frank Eibe, Mark A. Hall, **Data mining: Practical Machine Learning Tools and Techniques 3rd Edition**, Elsevier, 2011
- Ethem Alpaydm, **Introduction to Machine Learning**, 3rd ed., MIT Press, 2014
- Florin Gorunescu, **Data Mining: Concepts, Models and Techniques**, Springer, 2011

# Tugas, Latihan Soal dan Quiz

## Tugas :

- - Tugas Mandiri
- - Tugas Berkelompok dan di Presentasikan

## Latihan Soal :

- Akan diberikan pada penjelasan materi tertentu

## Quiz

- - Mendadak
- - Ada pemberitahuan sebelumnya

# Sistematika Penilaian

- 10 % Tugas
- 20 % Quiz
- 35 % UTS
- 35 % UAS

Diluar Penilaian Sebenarnya :

- 80 % kehadiran (Minimal 12 x hadir)

Nilai Akhir (Grade) Akan dipertimbangkan

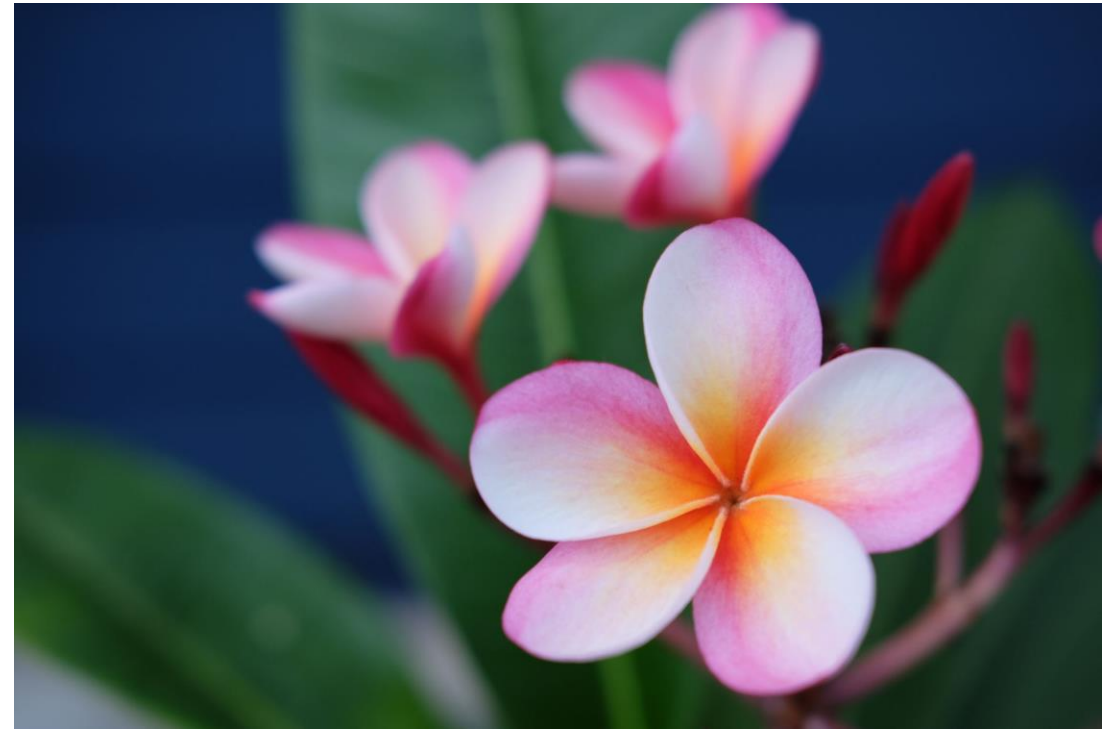
## Batas Nilai Akhir :

<b>Nilai Skor Matakuliah (NSM)</b>	<b>Nilai Mata Kuliah (NMK)</b>
<b><math>NSM &gt; 80</math></b>	<b>A</b>
<b><math>70 &lt; NSM \leq 80</math></b>	<b>AB</b>
<b><math>65 &lt; NSM \leq 70</math></b>	<b>B</b>
<b><math>60 &lt; NSM \leq 65</math></b>	<b>BC</b>
<b><math>55 &lt; NSM \leq 60</math></b>	<b>C</b>
<b><math>40 &lt; NSM \leq 55</math></b>	<b>D</b>
<b><math>NSM \leq 40</math></b>	<b>E</b>

<https://academic.ittelkom-sby.ac.id/wp-content/uploads/2019/03/SK-Buku-Pedoman-Akademik.pdf>

# Outline mata kuliah: data mining

- Pengenalan data mining
- Pemahaman tentang data



# Capaian Pembelajaran pertemuan ini

- Mengetahui konsep dasar, tujuan dan penerapan data mining
- Ilmu yang terkait dengan data mining
- Pengenalan awal tentang metode data mining yang dapat dipakai di dunia nyata
- Pengenalan awal tentang data
- Pengenalan awal tentang Metode learning pada algoritma data mining



Apa itu data mining

# Manusia memproduksi data



*Finansial*



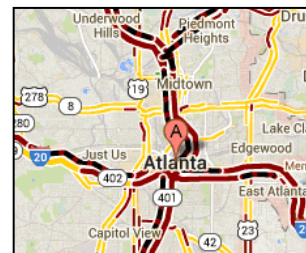
*Kesehatan*



*Pertandingan*



*Cuaca*



*Traffic Patterns*



*Sensor Networks*



*E-Commerce*

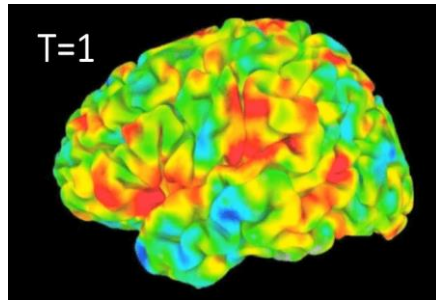
# Kenapa data mining? Commercial viewpoint

- Banyak data dikumpulkan dan di simpan
- Web data
  - Facebook punya miliaran data pengguna
- Pembelian di supermarket (mini market), e-commerce
  - Tokopedia, blibli, shopee, dll menerima jutaan visit perhari
  - alfamart mempunyai data konsumen
- Komputer harganya semakin terjangkau
- Persaingan dari competitor yang semakin kuat

The Google logo, featuring its characteristic multi-colored letters.The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue rectangular background.The Yahoo! logo, with the word "YAHOO!" in a stylized red font.The Amazon.com logo, featuring the text "amazon.com" in a black sans-serif font with a curved orange arrow underneath.

# Kenapa data mining? **scientific** viewpoint

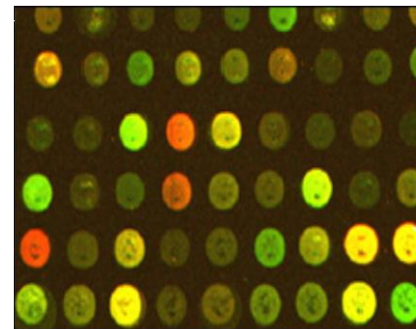
- Data dikumpulkan dan disimpan dengan kecepatan yang sangat tinggi



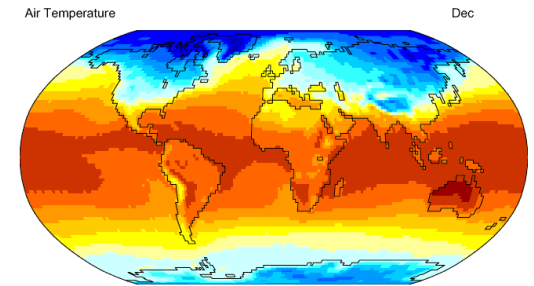
fMRI Data from Brain



Sky Survey Data



Gene Expression Data



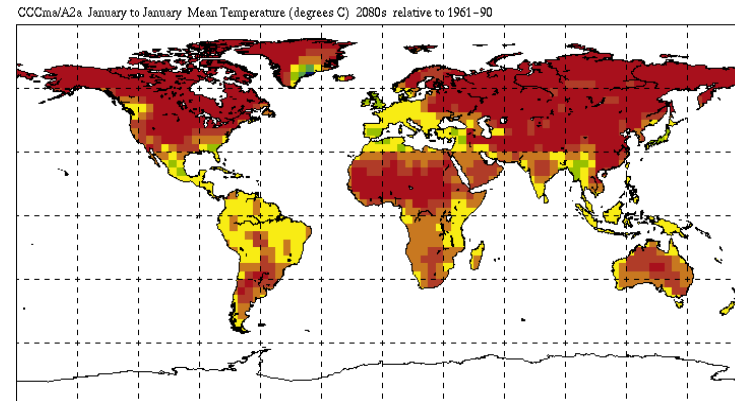
Surface Temperature of Earth

- Data mining membantu ilmuwan
  - dalam menganalisis dataset yang massif secara otomatis
  - Dalam merumuskan dan membuktikan hipotesis

# Data mining bisa membantu menyelesaikan masalah



Improving health care and reducing costs



Predicting the impact of climate change



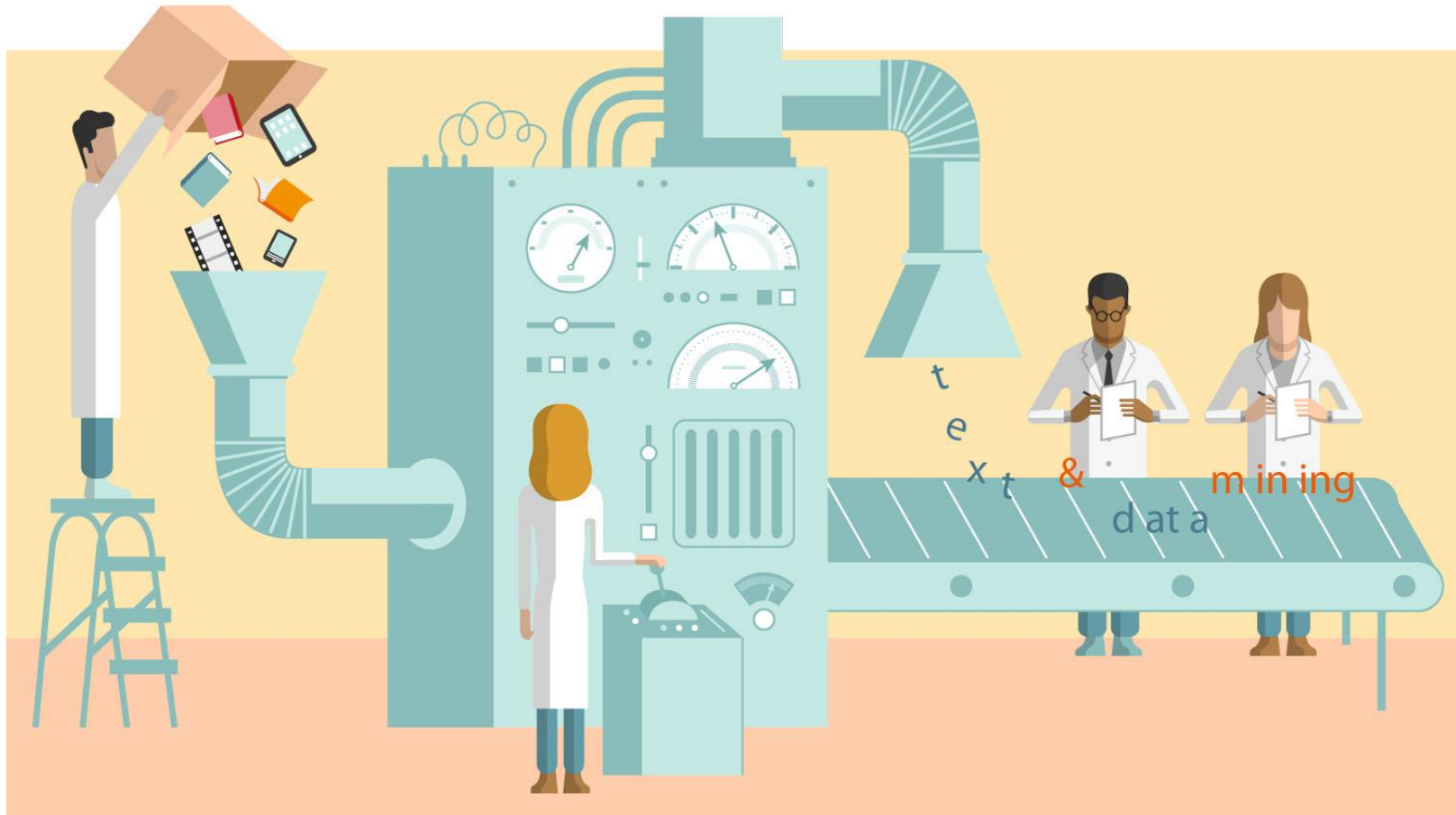
Finding alternative/ green energy sources



Reducing hunger and poverty by increasing agriculture production



# Pengertian data mining



# Pengertian data mining (**secara global**)

- 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

# Pengertian data mining (**menurut pakar**)

- 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*)

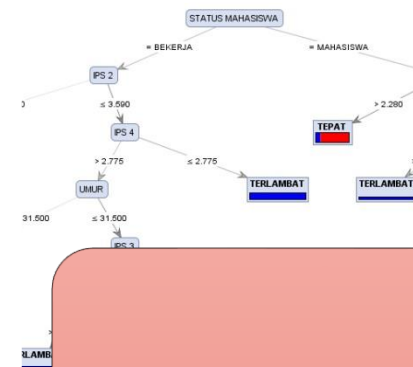


	B	C	D	E	F	G	H
	JENIS KELAMIN	STATUS MAHASISWA	UMUR	STATUS NIKAH	IPS 1	IPS 2	IPS 3
	PEREMPUAN	BEKERJA	28	BELUM MENIKAH	2,76	2,8	3,2
	PEREMPUAN	MAHASISWA	32	BELUM MENIKAH	3	3,3	3,14
	PEREMPUAN	BEKERJA	29	BELUM MENIKAH	3,5	3,3	3,7
	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	3,17	3,41	3,61
	PEREMPUAN	BEKERJA	29	BELUM MENIKAH	2,9	2,89	3,3
	LAKI-LAKI	BEKERJA	27	BELUM MENIKAH	2,35	2,82	3,09
	PEREMPUAN	MAHASISWA	26	BELUM MENIKAH	2,76	3,14	2,6
	PEREMPUAN	MAHASISWA	27	BELUM MENIKAH	2,62	2,89	2,32
	PEREMPUAN	BEKERJA	25	MENIKAH	3,6	3,54	3,52
	PEREMPUAN	BEKERJA	28	BELUM MENIKAH	2,71	2,55	1,77

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

$$-\left(-m \frac{r^2}{4l} \tan(\phi)\right) \left[ l - \frac{r^2}{4l} + r \left( \cos(\omega t) + \frac{r}{4l} \cos(2\omega t) \right) \right]$$

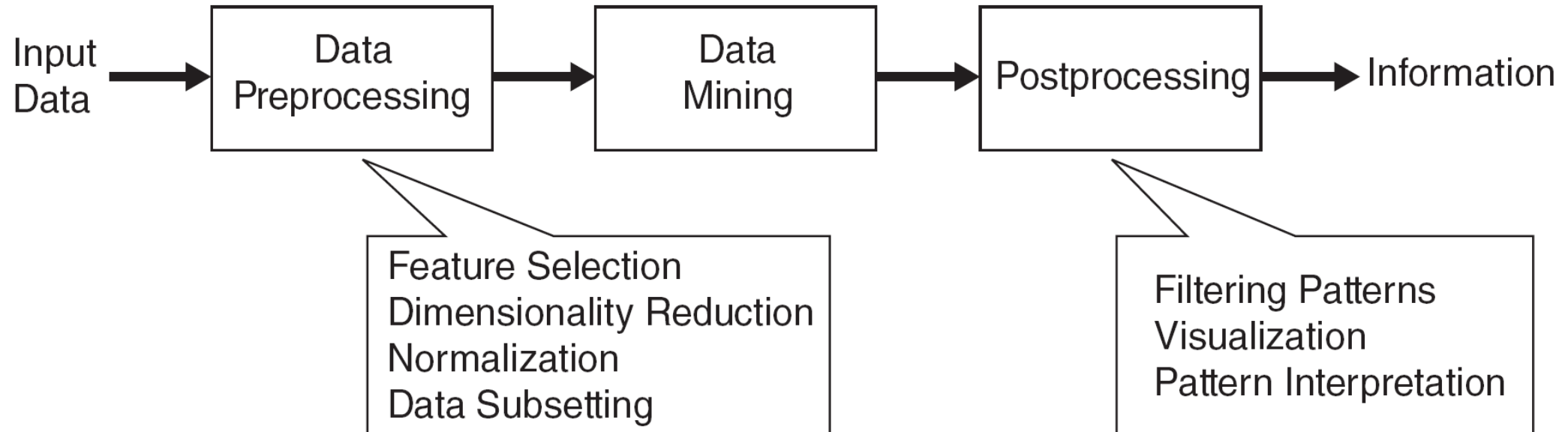
$$= R_1 e^{\left(-\zeta + \sqrt{\zeta^2 - 1}\right) \omega t} - \left(-\zeta + \sqrt{\zeta^2 - 1}\right) \omega t$$



**Himpunan Data**

**Metode Data Mining**

**Pengetahuan**



# 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

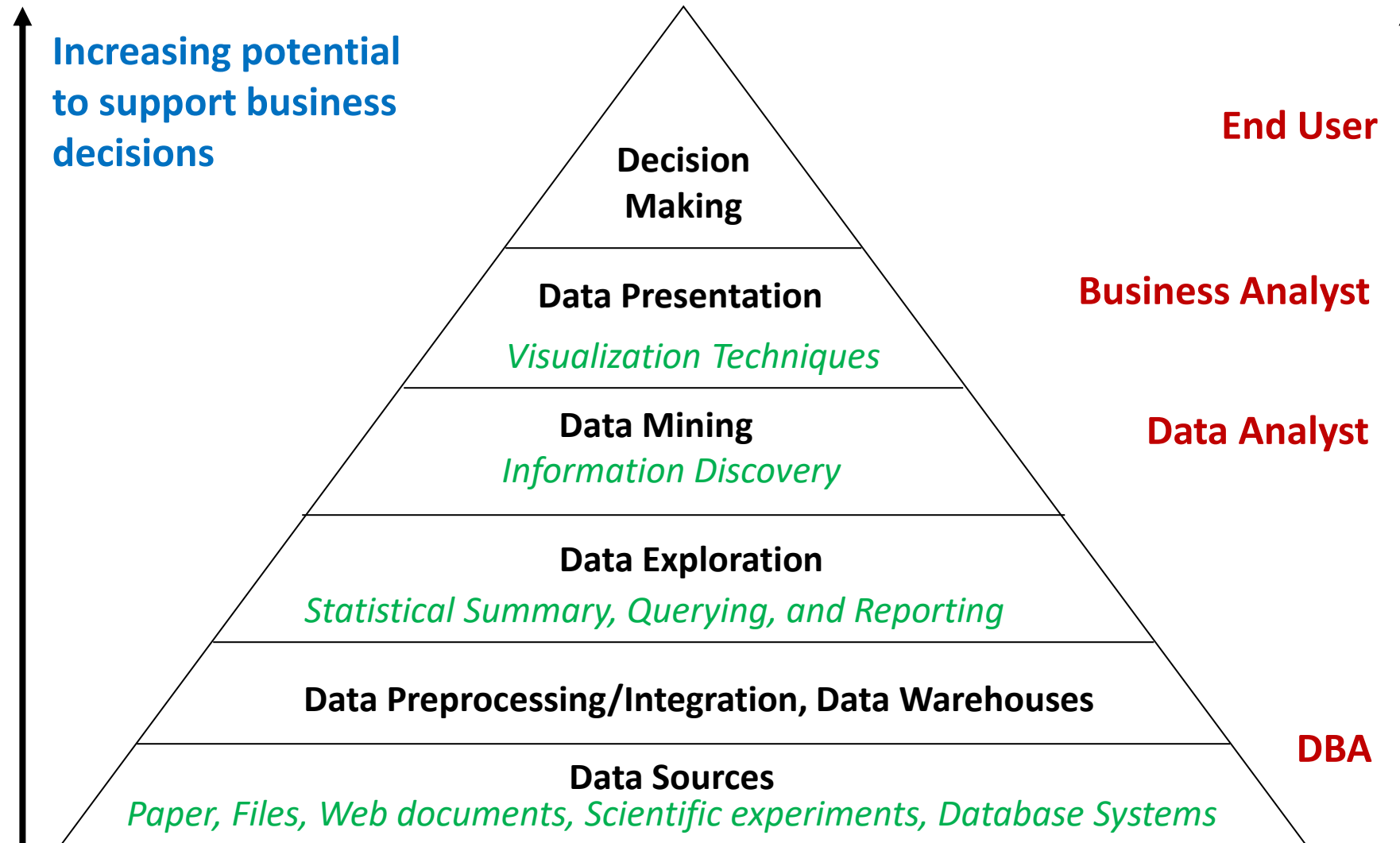
	Senin	Selasa	Rabu	Kamis	Jumat
Terlambat	7	0	1	0	5
Pulang Cepat	0	1	1	1	8
Izin	3	0	0	1	4
Alpa	1	0	2	0	2

Pola Kebiasaan Kehadiran Mingguan Pegawai

# Data – informasi – pengetahuan - kebijakan

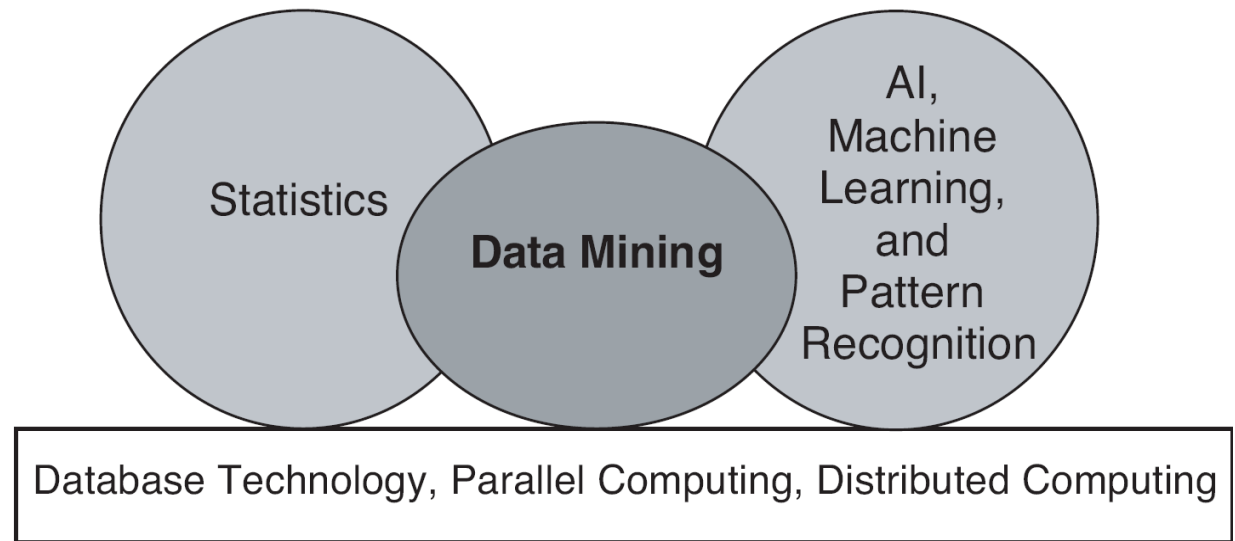
- Kebijakan **penataan jam kerja karyawan** khusus untuk hari senin dan jumat
- Peraturan jam kerja:
  - Hari **Senin** dimulai jam 10:00
  - Hari **Jumat** diakhiri jam 14:00
  - Sisa jam kerja **dikompensasi ke hari lain**

# Data mining pada business intelligence



# Awal mula data mining

- Berasal dari machine learning / AI, pattern recognition, statistic dan database systems
- Data mining cocok untuk data yang berkarakteristik
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - distributed



# 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**



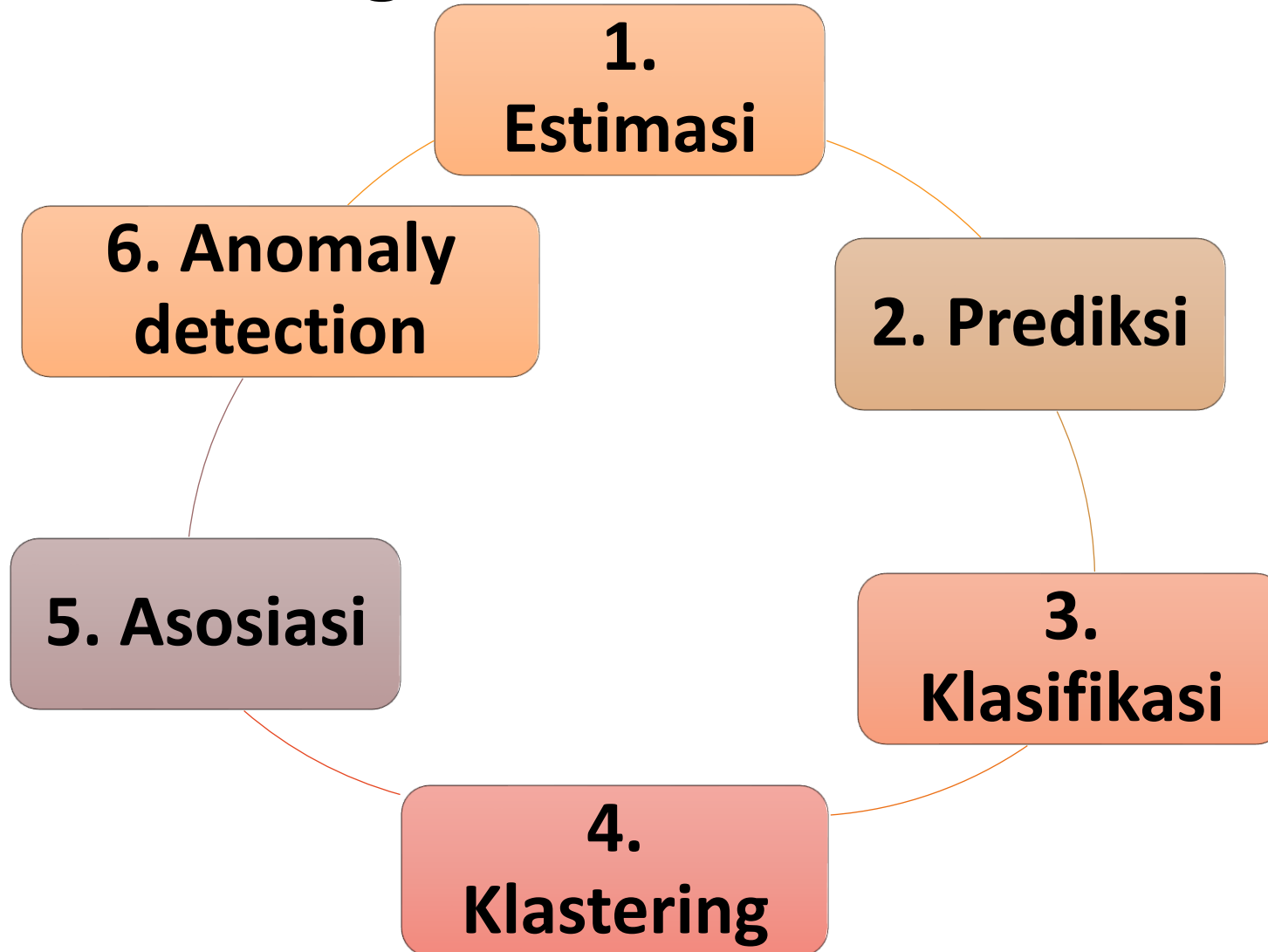


break

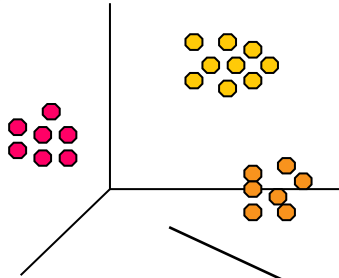
# Tugas Data Mining

- Metode Prediksi
  - Menggunakan beberapa variables untuk memprediksi nilai yang akan datang pada variable lain
    - Prediksi harga rumah
- Metode Deskripsi
  - Menemukan pola dari sebuah data yang bisa dipahami manusia

# Tugas data mining



# Tugas data mining



Clustering

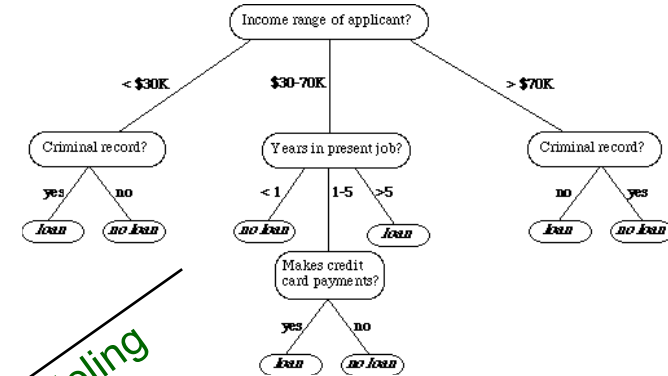
## Data

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes

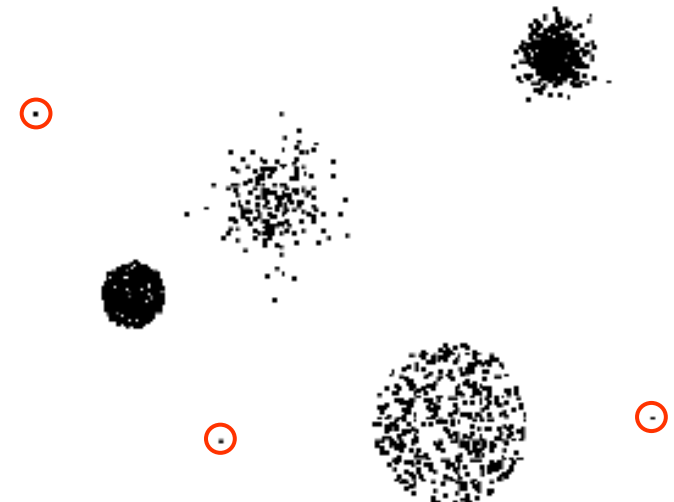
Association Rules



Predictive Modeling



Anomaly Detection



# 1. Estimasi

# Estimasi go-food

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

## 2. Regresi

# Regresi (regression)

- Memprediksi nilai variabel bernilai kontinu yang diberikan berdasarkan nilai-nilai variabel lain, dengan asumsi model dependensi linier atau nonlinier.
- Secara ekstensif dipelajari dalam statistik, bidang jaringan saraf.
- Contoh:
  - Memprediksi jumlah penjualan produk baru berdasarkan pembelanjaan yang menguntungkan.
  - Memprediksi kecepatan angin sebagai fungsi suhu, kelembaban, tekanan udara, dll.
  - Prediksi deret waktu dari indeks pasar saham.

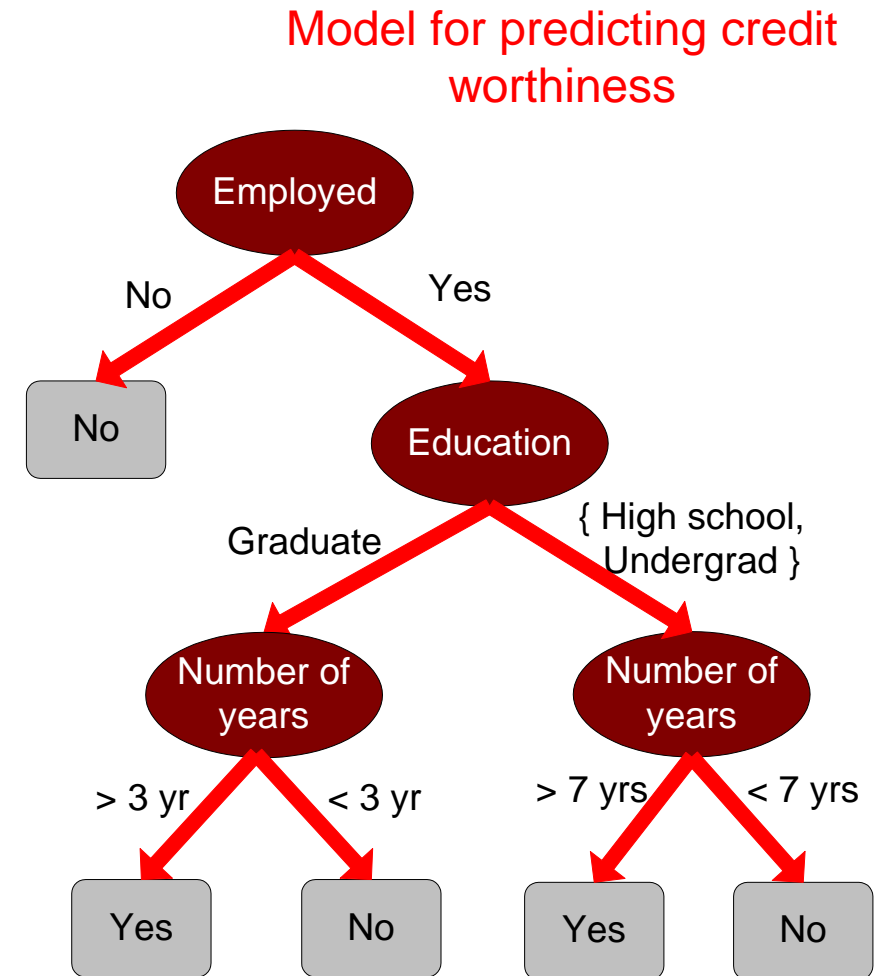


## 3. Klasifikasi

# model prediksi: klasifikasi

Temukan model untuk atribut kelas sebagai fungsi dari nilai atribut lainnya

				Class
<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...	...	...	...	...

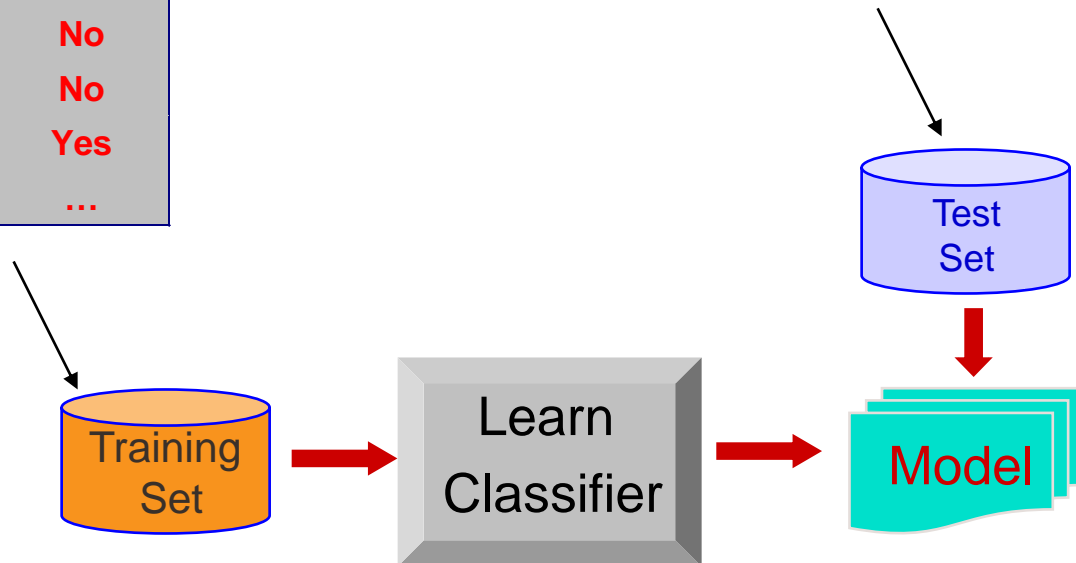


# Contoh klasifikasi

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
...	...	...	...	...

categorical  
categorical  
quantitative  
class

<i>Tid</i>	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
...	...	...	...	...



# Contoh klasifikasi yang lain

- Mengklasifikasikan transaksi kartu kredit sebagai sah atau curang
- Klasifikasi tutupan lahan (badan air, daerah perkotaan, hutan, dll.) Menggunakan data satelit
- Mengkategorikan berita sebagai keuangan, cuaca, hiburan, olahraga, dll
- Mengidentifikasi penyusup di dunia maya
- Memprediksi sel tumor sebagai jinak atau ganas
- Mengklasifikasikan struktur sekunder protein sebagai alfa-helix, beta-sheet, atau koil acak

# Contoh penerapan klasifikasi (1)

## Fraud detection (deteksi penipuan kartu kredit)

- Goal (sasaran)
  - Memprediksi kasus penipuan dalam transaksi kartu kredit.
- Approach (pendekatan)
  - Gunakan transaksi kartu kredit dan informasi pada pemegang akunnya sebagai atribut.
    - Kapan seorang pelanggan membeli, apa yang dia beli, seberapa sering dia membayar tepat waktu, dll
  - Beri label transaksi masa lalu sebagai penipuan atau transaksi wajar. Ini membentuk atribut kelas.
  - Pelajari model untuk kelas transaksi.
  - Gunakan model ini untuk mendeteksi penipuan dengan mengamati transaksi kartu kredit di akun.

# Contoh penerapan klasifikasi (2)

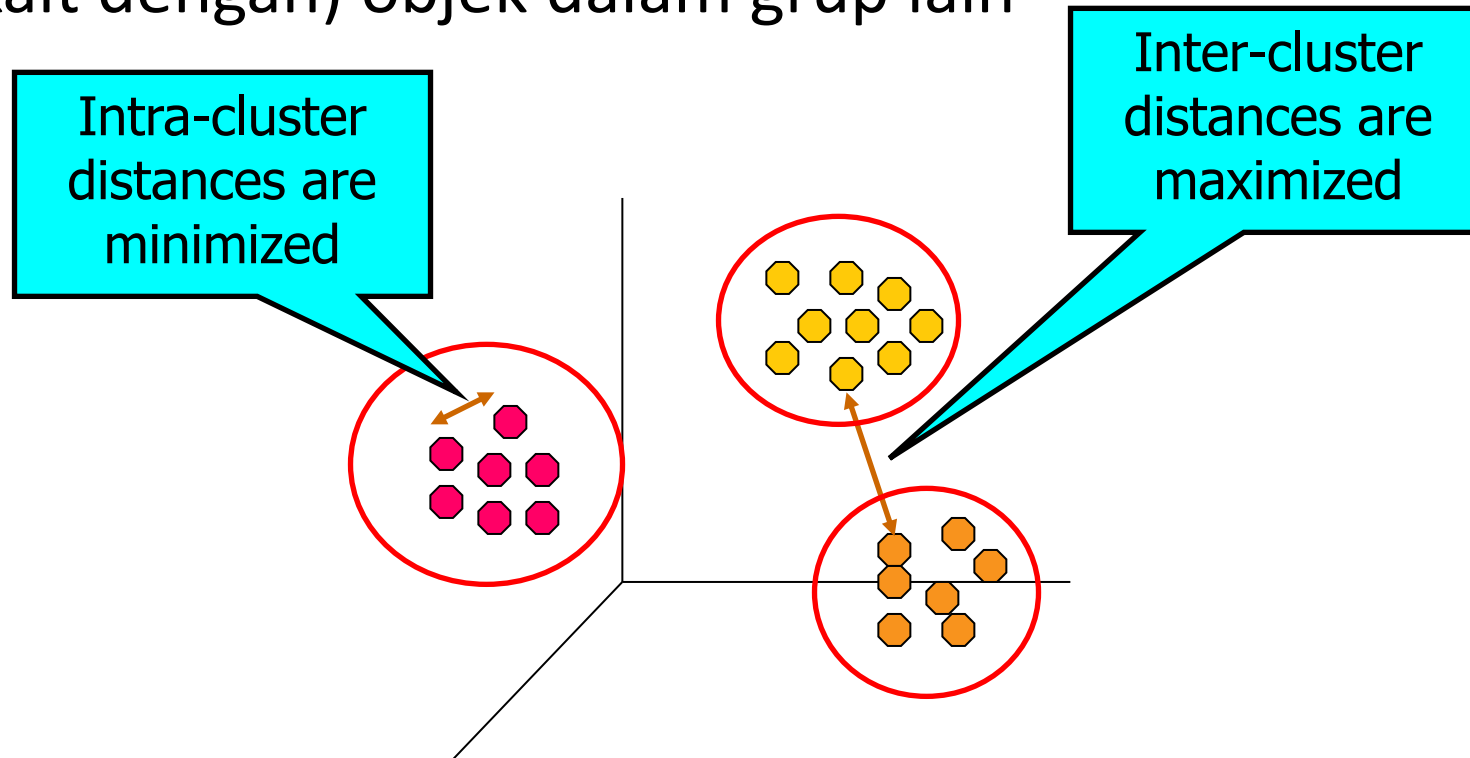
Prediksi churn untuk pelanggan telepon seluler (indosat / telkomsel)

- Goal (sasaran)
  - Memprediksi apakah seorang pelanggan kemungkinan hilang oleh pesaing.
- Approach (pendekatan)
  - Gunakan catatan rinci transaksi dengan masing-masing pelanggan di masa lalu dan sekarang, untuk menemukan atribut.
    - Seberapa sering pelanggan menelepon, di mana ia menelepon, jam berapa ia paling sering menelepon, status keuangannya, status perkawinan, dll.
  - Beri label pada pelanggan sebagai loyal atau tidak loyal.
  - Temukan model untuk loyalitas.

## 4. Clustering

# Clustering

- Menemukan kelompok objek sedemikian rupa sehingga objek dalam grup akan serupa (atau terkait) satu sama lain dan berbeda dari (atau tidak terkait dengan) objek dalam grup lain





# Contoh penerapan clustering

- Understanding
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- Summarization
  - Reduce the size of large data sets

# Contoh penerapan clustering (1)

## Market segmentation

- Goal (sasaran)
  - membagi pasar menjadi subset pelanggan yang berbeda di mana setiap subset mungkin dapat dipilih sebagai target pasar yang akan dicapai dengan bauran pemasaran yang berbeda.
- Approach (pendekatan)
  - Kumpulkan berbagai atribut pelanggan berdasarkan informasi geografis dan gaya hidup mereka.
  - Temukan kelompok pelanggan yang serupa.
  - Mengukur kualitas pengelompokan dengan mengamati pola pembelian pelanggan dalam kelompok yang sama vs yang dari kelompok yang berbeda.

# Contoh penerapan clustering (2)

## Document clustering

- Goal (sasaran)
  - Untuk menemukan kelompok dokumen yang mirip satu sama lain berdasarkan istilah-istilah penting yang muncul di dalamnya..
- Approach (pendekatan)
  - Untuk mengidentifikasi istilah yang sering muncul di setiap dokumen. Bentuk ukuran kesamaan berdasarkan frekuensi istilah yang berbeda. Gunakan untuk mengelompokkan.

## 5. Association Rule Discovery

# Association rule discovery: definisi

- Diberikan satu set records (catatan) yang masing-masing berisi sejumlah item dari koleksi yang diberikan
  - Menghasilkan aturan ketergantungan yang akan memprediksi terjadinya suatu item berdasarkan kemunculan item lainnya.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

$\{\text{Milk}\} \rightarrow \{\text{Coke}\}$

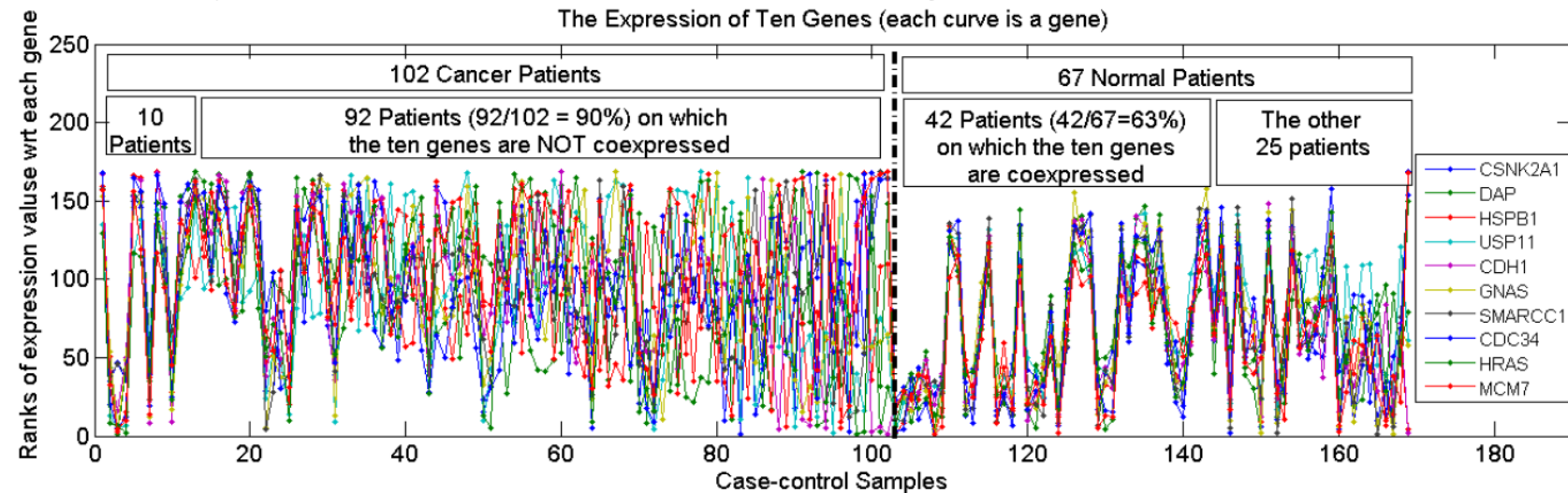
$\{\text{Diaper, Milk}\} \rightarrow \{\text{Beer}\}$

# Association Analysis: Applications

- Analisis pasar-keranjang (Market-basket analysis)
  - Aturan digunakan untuk promosi penjualan, manajemen rak, dan manajemen inventaris
- Diagnosis alarm telekomunikasi
  - Aturan digunakan untuk menemukan kombinasi alarm yang sering muncul bersamaan dalam periode waktu yang sama
- Informatika Medis
  - Aturan digunakan untuk menemukan kombinasi gejala pasien dan hasil tes yang terkait dengan penyakit tertentu

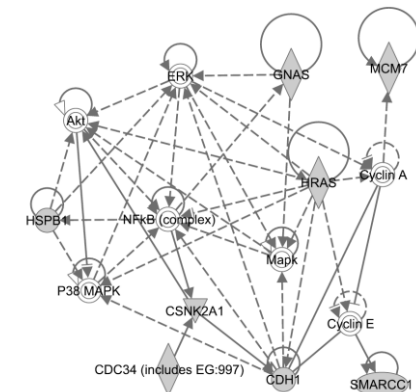
# Association Analysis: Applications

- Contoh Pola Koekspresi Diferensial Subruang dari dataset kanker paru-paru



Three lung cancer datasets [Bhattacharjee et al. 2001], [Stearman et al. 2005], [Su et al. 2007]

Diperkaya dengan jalur pensinyalan TNF / NFB yang dikenal terkait dengan kanker paru-paru



# Association Analysis: Applications

- 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



## 6. Anomaly Detection

# Deteksi anomali

Mendeteksi penyimpangan yang signifikan dari perilaku normal

- Aplikasi:
- Deteksi Penipuan Kartu Kredit
- Deteksi intrusi jaringan
- Identifikasi perilaku anomali dari jaringan sensor untuk pemantauan dan pengawasan.
- Mendeteksi perubahan tutupan hutan global.

# Tes pemahaman

- Pikirkan, kira-kira, data mining bisa diterapkan di area apa saja?

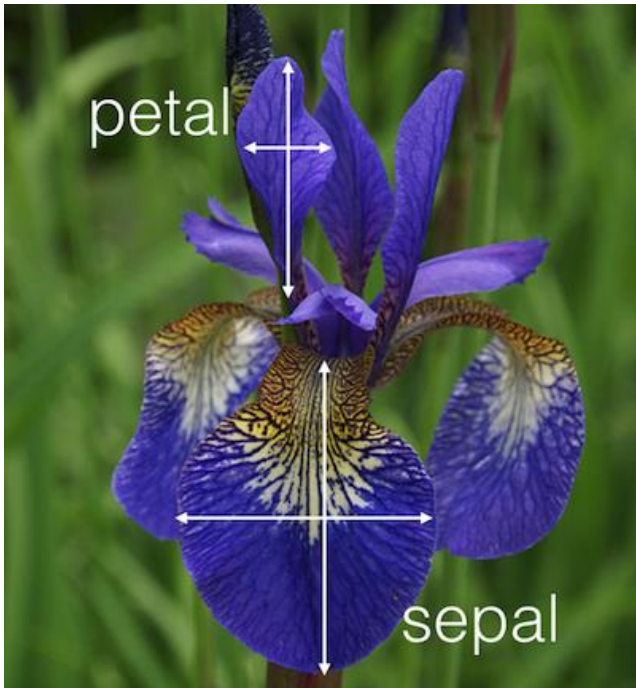


break

# Pengenalan tentang data (dataset)

Akan dijelaskan lebih lanjut di pertemuan berikutnya

# dataset



Petal: daun bunga

sepal: kelopak bunga

Attribute/Feature

Class/Label/Target

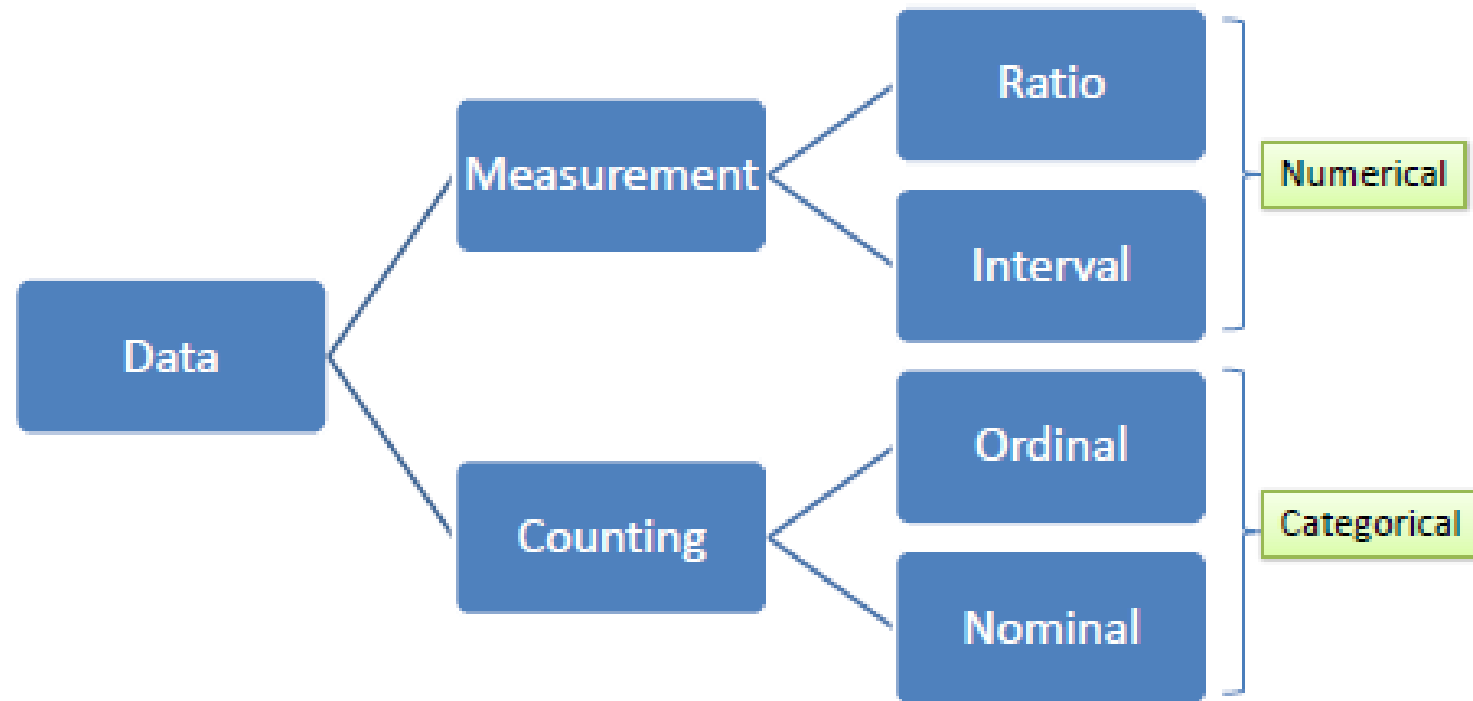
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Type
1	5.1	3.5	1.4	0.2	<i>Iris setosa</i>
2	4.9	3.0	1.4	0.2	<i>Iris setosa</i>
3	4.7	3.2	1.3	0.2	<i>Iris setosa</i>
4	4.6	3.1	1.5	0.2	<i>Iris setosa</i>
5	5.0	3.6	1.4	0.2	<i>Iris setosa</i>
...					
51	7.0	3.2	4.7	1.4	<i>Iris versicolor</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

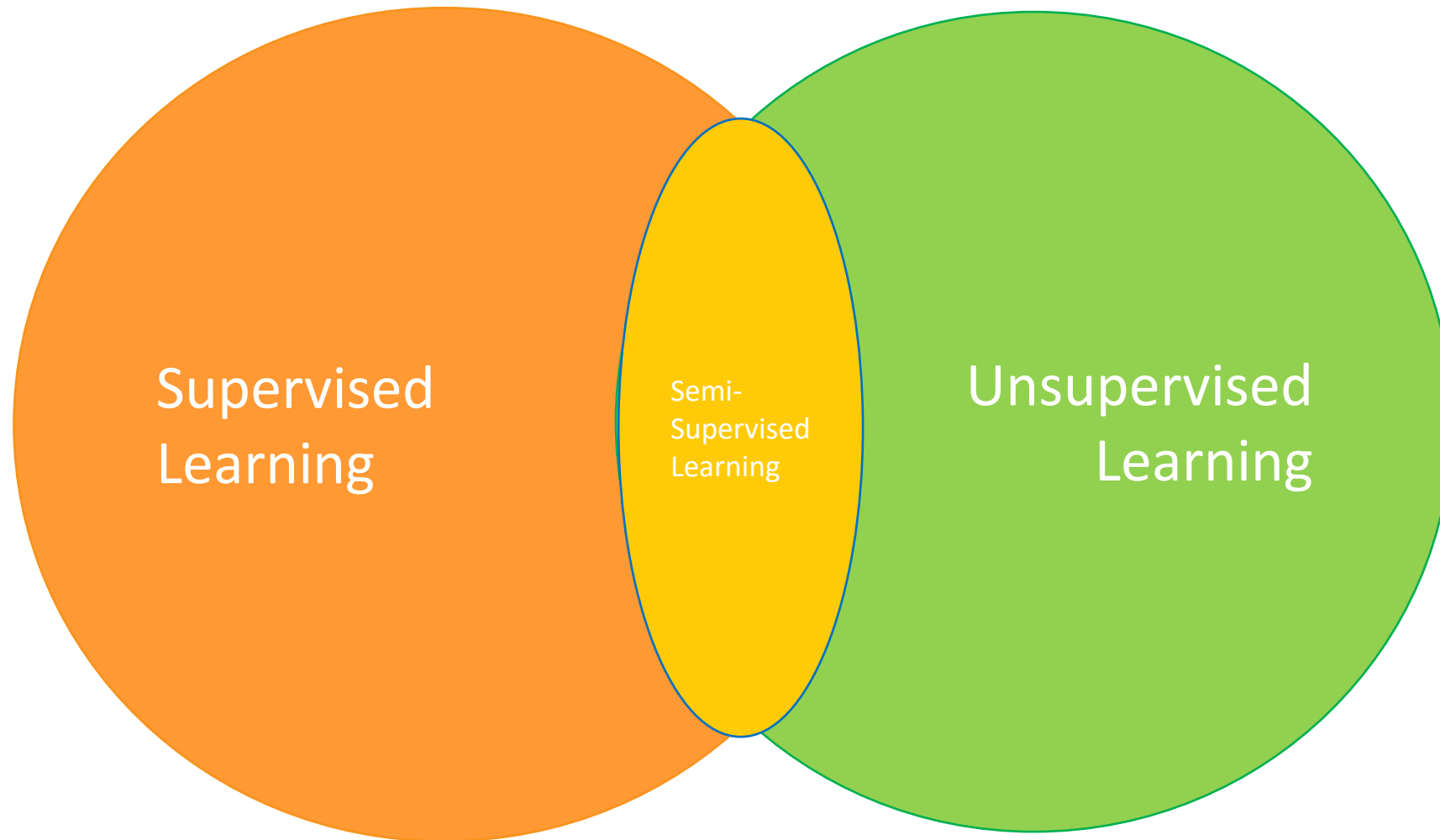


Jenis Atribut	Deskripsi	Contoh	Operasi
Ratio (Mutlak)	<ul style="list-style-type: none"> <li>Data yang diperoleh dengan cara <b>pengukuran</b>, dimana jarak dua titik pada skala sudah diketahui</li> <li>Mempunyai titik <b>nol yang absolut</b> (*, /)</li> </ul>	<ul style="list-style-type: none"> <li>Umur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



Metode learning pada data mining

# Metode learning



# A. 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

## A. Supervised Learning – contoh dataset dengan kelas

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

## B. 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

## B. UnSupervised Learning

Attribute/Feature



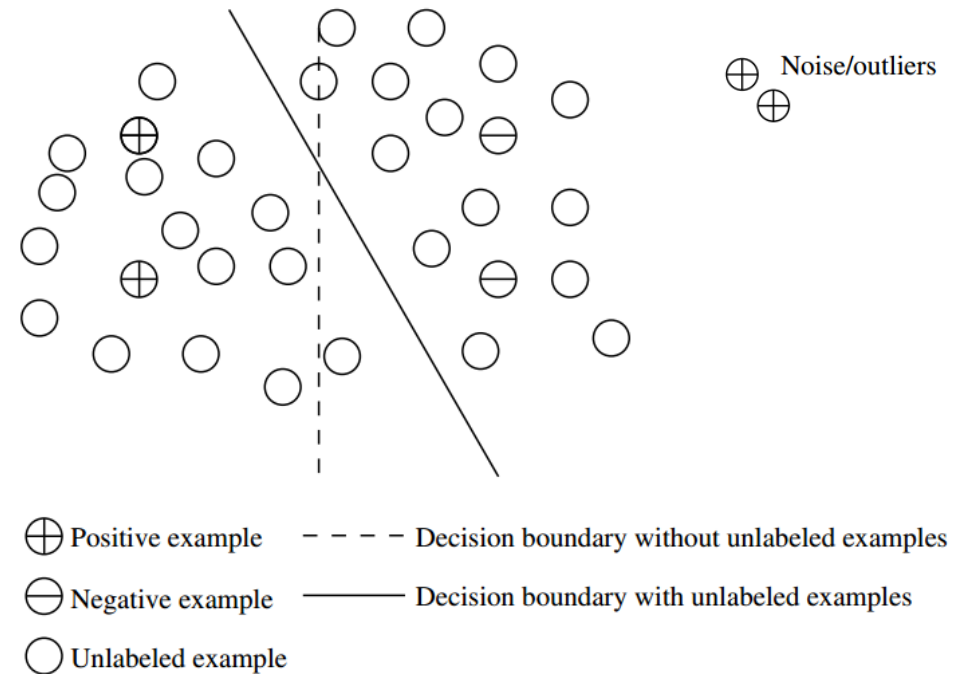
	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
...				
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
...				
101	6.3	3.3	6.0	2.5
102	5.8	2.7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

# A. 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

# C. 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

# Review Sebelum pertemuan selesai

Apa yang sudah kita pelajari pada pertemuan ini?

- Pengenalan data mining
- Kapan memakai klasifikasi
- Kapan memakai clustering
- Kapan memakai association rules
- Kapan memakai anomaly detection
- Pengenalan awal tentang data
- Pengenalan dasar Metode learning pada algoritma data mining

# Penutup

- Pada pertemuan berikutnya, kita akan mempelajari dan membahas tentang “**pemahaman data**”, yang berupa:
  - Review atribut dan object
  - Tipe data
  - Kualitas data
  - Similarity dan distance (jarak)
  - Data preprocessing
  - Data cleansing
  - Data transformation
  - Feature selection

# Tugas

- Dikumpulkan pada:

selesai