

# DATA MINING

( 3 SKS )

# Data Mining

<b>Kode Matakuliah</b>	<b>: GC314</b>	
<b>Nama Matakuliah</b>	<b>: Data Mining</b>	
<b>Komponen Penilaian</b>	<b>: Kehadiran</b>	<b>20%</b>
	<b>Keaktifan Kelas</b>	<b>15%</b>
	<b>Seminar</b>	<b>20%</b>
	<b>UTS</b>	<b>20%</b>
	<b>UAS</b>	<b>25%</b>

## **KEPUSTAKAAN:**

1. Mundy, Thornwaite, Kimball, “ Introduction to Data Mining, Pang Ning Tan”, International Edition, PEARSON, 2006.
2. Ian H. Witten, Eibe Frank, “Data Mining: Practical Machine
3. Learning Tools and Techniques with Java Implementations”, 2nded., Morgan Kaufmann., 2005.
4. Retno Tri Vulandari, S.Si., M.Si., “Data Mining Teori dan Aplikasi Rapidminer”, Penerbit Gava Media, 2017.
5. Kusrini, Emha Taufiq Luthfi, “Algoritma Data Mining”, STMIK Amikom Yogyakarta, 2009.
6. Eko Prasetyo, “ Data Mining, Mengolah Data menjadi Informasi menggunakan Matlab”, Penerbit Andi, 2014.
7. Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, Third Edition, Morgan Kaufmann Publisher, 2012
8. Sarfaraz M Manik, “Data Mining Making Sense of Data”
9. Akannsha A. Totewar, “Data Mining: Concepts and Techniques”
10. Dari beberapa sumber lain

# Manusia Menghasilkan Data

Manusia setiap hari menghasilkan/ memproduksi bermacam data yang jumlah dan ukurannya sangat besar

Di Bidang:  
Pendidikan  
Kesenian  
Kedokteran  
Rumah Sakit  
Nuklir  
Astronomi  
Bisnis  
Ekonomi  
Olahraga  
Cuaca  
Financial  
dll



# Data ?

- Data adalah catatan kumpulan *fakta*
- Sekumpulan keterangan/fakta yang dibuat dengan simbol, angka, kata-kata, maupun kalimat
- Data dapat diperoleh melalui sebuah proses pencarian serta pengamatan yang tepat berdasarkan sumber-sumber tertentu
- Kumpulan deskripsi/keterangan dasar yang berasal dari obyek maupun kejadian nyata/real
- Kumpulan yang terdiri dari fakta-fakta untuk memberikan gambaran yang luas terkait dengan suatu keadaan
- Melalui data seseorang dapat menganalisis, menggambarkan, atau menjelaskan suatu keadaan





Data



diolah melalui berbagai penelitian dan percobaan

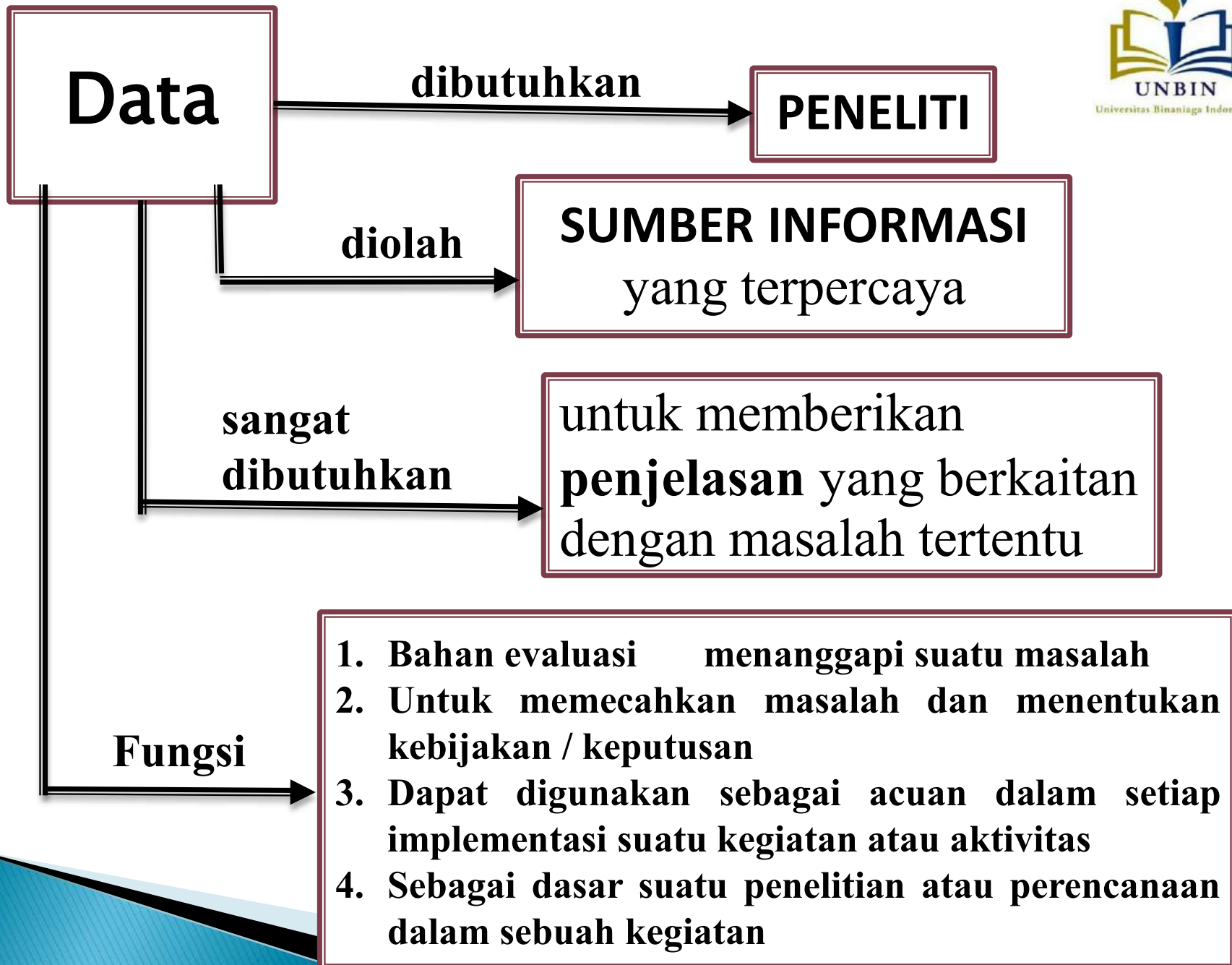


dibentuk menjadi suatu hal yang lebih beragam, seperti **database**



sebagai **solusi** dalam menyelesaikan suatu **masalah**







# Jenis Data

berdasarkan

**Sifatnya**

**Kuantitatif**

**Kualitatif**

**Cara Memperoleh**

**Primer**

**Sekunder**

**Sumbernya**

**Internal**

**Eksternal**

# Jenis Data berdasarkan Sifatnya

## Kuantitatif

- Data yang diperoleh dengan melakukan survei
- jawaban berupa angka (bersifat lebih obyektif)
- Bila seseorang membaca atau memahami data kuantitatif akan menafsirkan dengan sama sesuai nilai angkanya
- Contoh: Berat Tio 68 kg

## Kualitatif

- Merupakan data deskriptif atau data yang tidak berbentuk angka
- Umumnya dinyatakan dalam bentuk simbol, gambar, atau variabel
- Diperoleh melalui kuesioner, wawancara, studi literature atau observasi
- Bersifat obyektif, sehingga setiap orang yang membacanya akan menimbulkan arti serta penafsiran yang berbeda-beda
- Contoh: Kualitas pelayanan RS



# Jenis Data berdasarkan Cara Memperoleh

## Primer

- Data yang diperoleh dari objek yang diteliti oleh orang atau organisasi yang sedang melakukan penelitian
- Contoh dari data primer: data hasil wawancara langsung, hasil survei, dan kuesioner terhadap responden

## Sekunder

- Data yang diperoleh dari sumber lain yang telah ada
- Peneliti tidak mengumpulkan data langsung dari objek yang diteliti
- Contoh jenis data sekunder: data sensus penduduk, data penyakit dan data yang dikeluarkan oleh pemerintah



# Jenis Data berdasarkan Sumbernya

## Internal

- Data yang diperoleh secara langsung dari tempat penelitian
- Contoh data jenis ini: jumlah karyawan, tingkat kepuasan karyawan dalam suatu institusi, dan kebutuhan tenaga kerja di suatu perusahaan.

## Eksternal

- Data yang didapat dari luar lingkup kerja di suatu perusahaan
- Umumnya sebagai data pembandingan
- Contoh data jenis ini: data kependudukan, jumlah mahasiswa di kampus dan data penjualan produk dari perusahaan lain.



# Manfaat dan Fungsi Data

## 1. *Sebagai acuan kegiatan*

Data dapat digunakan sebagai acuan atau tolak ukur untuk membuat sebuah kegiatan tertentu yang diinginkan.

## 2. *Sebagai dasar perencanaan*

- Dalam membuat suatu perencanaan sangat diperlukan parameter yang akurat data adalah parameter yang akurat dapat dijadikan dasar dalam membuat sebuah perencanaan
- Data juga dapat digunakan sebagai dasar melakukan perkiraan keadaan di waktu yang akan datang
- Dengan berdasarkan pada data, sebuah perencanaan akan lebih terarah sehingga dapat diperoleh hasil yang tepat



### 3. *Dasar untuk membuat keputusan*

- Sebuah data juga dapat bermanfaat untuk membuat sebuah keputusan
- Dari data yang ada, seseorang dapat membuat keputusan terbaik terhadap suatu permasalahan yang ada
- Dengan begitu, seseorang akan dengan mudah menentukan keputusan berdasarkan data yang dapat dipertanggung jawabkan

### 4. *Sebagai bahan untuk Evaluasi*

- Data juga dapat dijadikan sebagai bahan evaluasi
- Misalnya dalam sebuah lembaga atau organisasi tertentu pasti dibutuhkan suatu evaluasi dalam rangka meningkatkan kualitasnya



# Data Mining ?



## PENGANTAR DATA MINING

- Setiap hari dalam bidang apapun, seseorang apakah itu pelajar, Mahasiswa, Guru, Dosen, Pegawai/karyawan, sebuah Lembaga atau Perusahaan pasti akan menghasilkan **Data**
- Berhari-hari, berbulan-Bulan, bertahun-tahun, **Data** akan semakin banyak, bertumpuk dan membutuhkan lokasi penyimpanan khusus
- Era saat ini, seiring perkembangan zaman, **Data** yang banyak itu disimpan secara elektronik
- Selama dua decade terakhir telah terjadi peningkatan yang dramatis terhadap jumlah data atau informasi yang disimpan secara elektronik.
- Telah diperkirakan sebelumnya bahwa jumlah informasi di dunia akan berlipat ganda setiap 20 bulan dan jumlah ukuran basis data akan bertambah lebih cepat lagi dari itu.

- Teknologi database saat ini memungkinkan untuk menyimpan sejumlah data dalam jumlah yang sangat besar dan terakumulasi → Volume data menjadi sangat besar
- Disinilah awal timbulnya persoalan **ledakan data** (jumlah data yang tiba-tiba menjadi sangat besar)

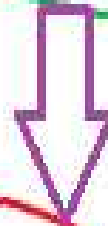
**Data perlu disimpan**

**Penting untuk mengetahui  
proses penemuan  
pengetahuan (knowledge)  
dari data yang disimpan**

**Data yang tersimpan dalam  
sebuah gudang data (data  
warehouse) perlu dianalisa**

# Ledakan Data

Permasalahannya??

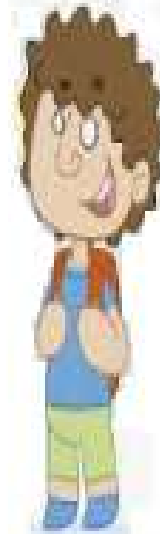


Volume  
of Data

1970 1980 1990 2000

Apa yang harus dilakukan  
dengan data-data tersebut ?

Perkembangan Database (Ledakan Data)



## *Secara umum*

- Informasi merupakan hal penting dalam menunjang operasi bisnis dan membantu pengambil keputusan untuk mendapatkan gambaran lebih tentang bisnis mereka
- Sistem manajemen basis data memberikan akses terhadap data namun hanya sebagian kecil kontribusinya terhadap apa yang seharusnya dapat dihasilkan dari data-data itu
- Sistem pemrosesan transaksi online (OLTP) tradisional sangat baik dalam menyimpan data secara cepat, aman, dan efisien ke dalam basis data namun tidak cukup baik dalam hal kemampuan melakukan analisa terhadap data-data yang ada

Solusi untuk persoalan penemuan pengetahuan dalam database berukuran besar adalah dengan menggunakan


**Data Mining & Data Warehousing**

Peran Data Mining

**Memberikan kontribusi besar bagi setiap perusahaan yang mengimplementasikannya**




## 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*)



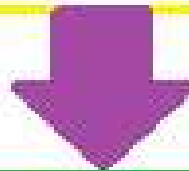
Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data (*Han et al., 2011*)

# PENGERTIAN DATA MINING



- Proses penemuan pola yang menarik dari data yang tersimpan dalam jumlah besar. Merupakan evolusi alami dari teknologi database, dan merupakan metode yang paling banyak dibutuhkan, dengan aplikasi yang sangat luas.
- Ekstraksi dari suatu informasi yang berguna atau menarik (non-trivial, implisit, sebelumnya belum diketahui, potensial kegunaannya) pola atau pengetahuan dari data yang disimpan dalam jumlah besar.
- Ekplorasi dari analisa secara otomatis atau semiotomatis terhadap data-data dalam jumlah besar untuk mencari pola dan aturan yang berarti.

# PENGERTIAN DATA MINING



Disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data yang besar

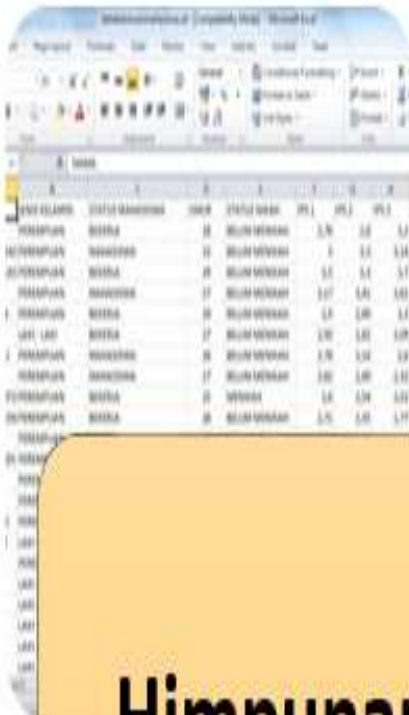
## **Ekstraksi dari data ke pengetahuan:**

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





NO	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
1	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
2	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
3	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
4	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
5	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
6	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
7	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
8	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
9	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
10	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
11	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
12	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
13	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
14	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
15	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
16	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
17	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
18	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
19	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS
20	STATUS	STATUS	STATUS	STATUS	STATUS	STATUS

**Himpunan  
Data**

$$f(x) = \sum_{k=1}^n f_k(x) = f(x)$$

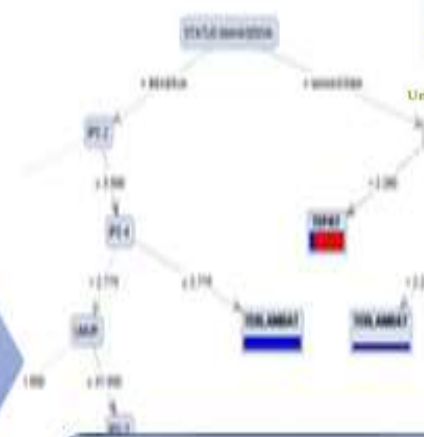
$$\int_a^b f(x) dx = \lim_{n \rightarrow \infty} \frac{b-a}{n} \sum_{k=1}^n f\left(a + \frac{b-a}{n} k\right)$$

$$= \left( -\frac{1}{\omega} \sin(\omega t) \right) \left( \frac{1}{4} - \frac{1}{4} \cos(2\omega t) + \frac{1}{4} \cos(2\omega t) \right)$$

$$\left( -\frac{1}{\omega} \sin(\omega t) \right) \left( \frac{1}{4} - \frac{1}{4} \cos(2\omega t) + \frac{1}{4} \cos(2\omega t) \right)$$

11

**Metode Data  
Mining**



**Pengetahuan**

- Pada dasarnya data mining berhubungan erat dengan analisa data dan penggunaan perangkat lunak untuk **mencari pola dan kesamaan dalam sekumpulan data**. Ide dasarnya sangat menggali sumber yang berharga dari tempat yang sama sekali tidak diduga seperti perangkat lunak data mining mengekstrasi pola yang sebelumnya tidak terlihat atau tidak begitu jelas sehingga tidak seorang pun yang memperhatikan sebelumnya
- Analisa data mining berjalan pada data yang cenderung terus membesar dan teknik terbaik yang digunakan kemudian beorientasi kepada data berukuran sangat besar untuk mendapatkan kesimpulan dan keputusan paling layak
- Meskipun sebagian besar teknik data mining sudah ada sejak lama, namun hanya pada beberapa tahun terakhir ini data mining benar-benar berperan yaitu sejak dilakukan komersialisasi data mining.

## **Alasan data mining dibutuhkan**

### **1. Data telah mencapai jumlah dan ukuran yang sangat besar**

- ❖ Hasil dan proses data mining merupakan suatu informasi yang akan mendasari tindakan tertentu sehingga tingkat kebenaran informasi tersebut menjadi sangat signifikan, dan makin besar serta makin banyak data yang digunakan maka akan semakin valid hasilnya
- ❖ Perkembangan data dalam hal jumlah dan ukuran telah mencapai kecepatan yang sangat cepat, sehingga ukuran basis data yang dimiliki oleh sebuah perusahaan bisa mencapai kisaran gigabyte atau bahkan terabyte.

### **2. Telah dilakukan proses data warehousing**

- ❖ Untuk mencapai hasil yang memuaskan, maka sumber data yang digunakan dalam proses data mining seringkali merupakan data gabungan dari banyak departemen, daerah operasi bahkan dari sumber-sumber lain seperti data kependudukan
- ❖ Oleh karena itu maka disarankan perlunya proses data warehousing untuk menjaga konsistensi, memberikan prespektif yang lebih baik terhadap data dan menjaga integritas data.

### 3. Kemampuan Komputasi yang semakin terjangkau

- ❖ Pada dasarnya proses data mining melakukan banyak akses terhadap data yang sangat besar, juga melakukan proses komputasi yang membutuhkan sumber daya sangat besar
- ❖ Penurunan harga yang cukup cepat terhadap perangkat keras komputer serta semakin tingginya kinerja yang berhasil dicapai oleh perangkat komputer maupun teknologi pengolahan data seperti teknologi paralel proses saat ini, menjadikan proses data mining sudah cukup layak untuk dilakukan secara komersial.

### 4. Persaingan bisnis yang semakin ketat

- ❖ Tekanan persaingan bisnis yang semakin ketat mendorong perusahaan-perusahaan untuk selalu berinovasi agar mampu meningkatkan daya saingnya dipasar global.
- ❖ Beberapa tren yang berkembang saat ini:
  - a. Setiap bisnis adalah bisnis pelayanan
  - b. Adanya fenomena kustomisasi produk oleh masyarakat
  - c. Informasi adalah produk

## MODEL DALAM DATA MINING (ada 2 tipe)

### Verifikasi

Menggunakan pendekatan top down dengan mengambil hipotesa dari user dan memeriksa validitasnya dengan data sehingga bisa dibuktikan kebenaran hipotesa tersebut

### Knowledge Discovery (2)

Menggunakan pendekatan bottom up untuk mendapatkan informasi yang sebelumnya tidak diketahui.

Model ini terbagi menjadi dua

#### Directed Knowledge Discovery

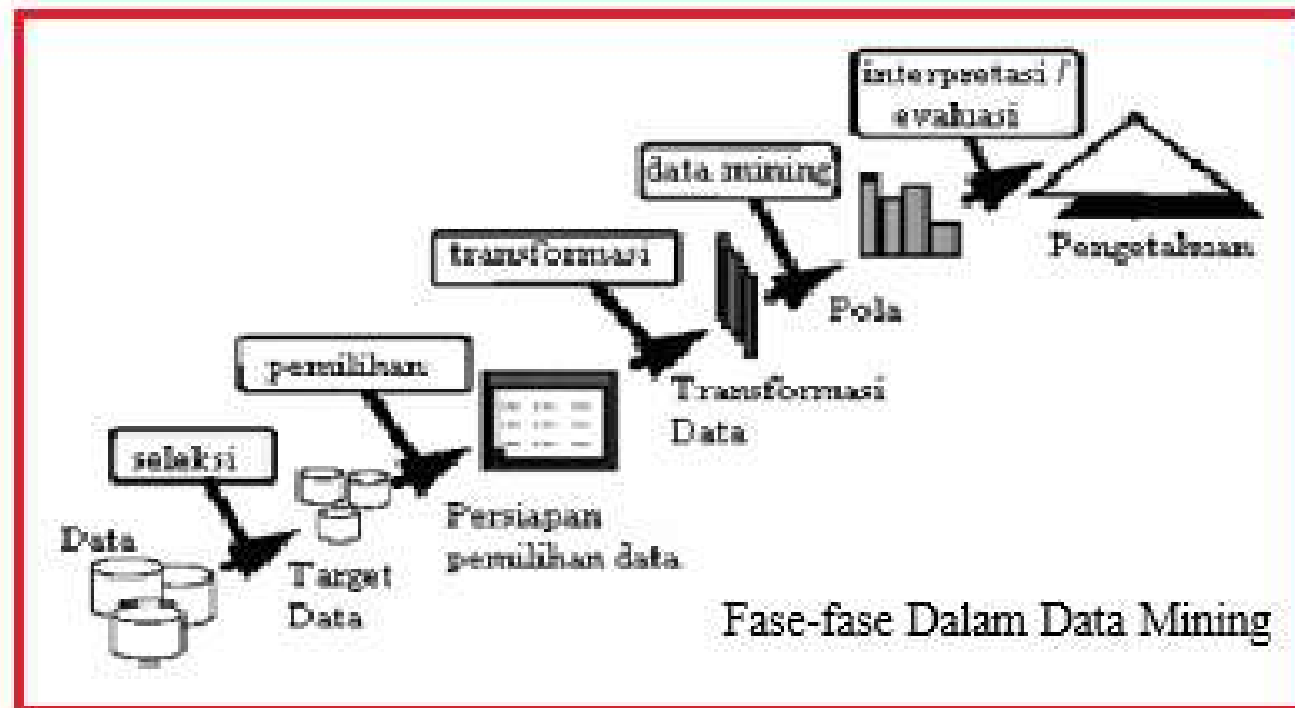
Data Mining akan mencoba mencari penjelasan nilai target field tertentu ( seperti penghasilan, respons, usia, dll) terhadap field-field yang lain

#### Undirected Knowledge Discovery

- ❖ Tidak ada target field karena komputer akan mencari pola yang ada pada data
- ❖ Jadi undirected knowledge discovery digunakan untuk mengenali hubungan / relasi yang ada pada data sedangkan directed knowledge discovery akan menjelaskan hubungan / relasi tersebut.

## TAHAPAN PROSES DALAM DATA MINING

- Ada beberapa tahapan proses dalam data mining
- Gambar dibawah menunjukkan beberapa tahap / proses yang berlangsung dalam data mining
- Fase awal dimulai dari data sumber dan berakhir dengan adanya informasi yang dihasilkan dari beberapa tahapan



# Tahapan proses dalam Data Mining

## 1. Seleksi Data

- Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai
- Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dalam basis data operasional

## 2. Pre-processing/ Cleaning ( pemilihan data )

- Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD
- Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi)
- Juga dilakukan proses enrichment, yaitu proses “memperkaya” data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal



## Tahapan proses dalam Data Mining (Lanjutan...)

### 3. Transformasi

- Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining
- Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data

### 4. Data mining

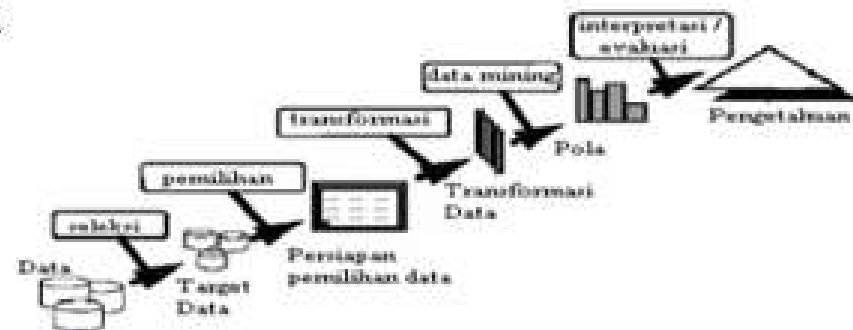
- Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu
- Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan

### 5. Interpretasi / Evaluasi

- Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan
- Tahap ini merupakan bagian dari proses KDD yang disebut dengan **interpretation**
- Tahap ini mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya

# Knowledge discovery in databases (KDD)

- Proses KDD secara garis besar memang terdiri dari 5 tahap seperti yang telah dijelaskan sebelumnya



- Akan tetapi, dalam proses KDD yang sesungguhnya, dapat saja terjadi iterasi atau pengulangan pada tahap-tahap tertentu
- Pada setiap tahap dalam proses KDD, seorang analis dapat saja kembali ke tahap sebelumnya, Sebagai contoh, pada saat coding atau data mining, analis menyadari proses cleaning belum dilakukan dengan sempurna, atau mungkin saja analis menemukan data atau informasi baru untuk “memperkaya” data yang sudah ada.

## Contoh:

# Data - Informasi – Pengetahuan

## Data Kehadiran Pegawai

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

Contoh:

Data - **Informasi** – Pengetahuan  
Informasi Akumulasi Bulanan Kehadiran Pegawai

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

Contoh:



# Data - Informasi – Pengetahuan

## Pola Kebiasaan Kehadiran Mingguan Pegawai

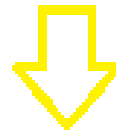
	Senin	Selasa	Rabu	Kamis	Jumat
<u>Terlambat</u>	7	0	1	0	5
Pulang Cepat	0	1	1	1	8
Izin	3	0	0	1	4
Alpa	1	0	2	0	2

# 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**



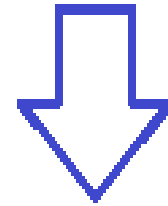
## METODE DATA MINING



- Data mining model dibuat berdasarkan salah satu dari dua jenis pembelajaran: supervised dan unsupervised
- Fungsi **Supervised**: untuk memprediksi suatu nilai
- Fungsi **Unsupervised**: untuk mencari struktur intrinsik, relasi dalam suatu data yang tidak memerlukan class atau label sebelum dilakukan proses pembelajaran
- Contoh dari algoritma **Unsupervised** → K-means clustering, Apriori association rules
- Contoh dari algoritma **Supervised** → NaiveBayes untuk klasifikasi



Klasifikasi Metode data mining berdasarkan fungsi yang dilakukan/ berdasarkan jenis aplikasi yang menggunakannya:



- **Klasifikasi (supervised)**
- **Clustering (unsupervised)**
- **Association Rules (unsupervised)**
- **Attribute Importance (supervised)**



**Terima Kasih.....**