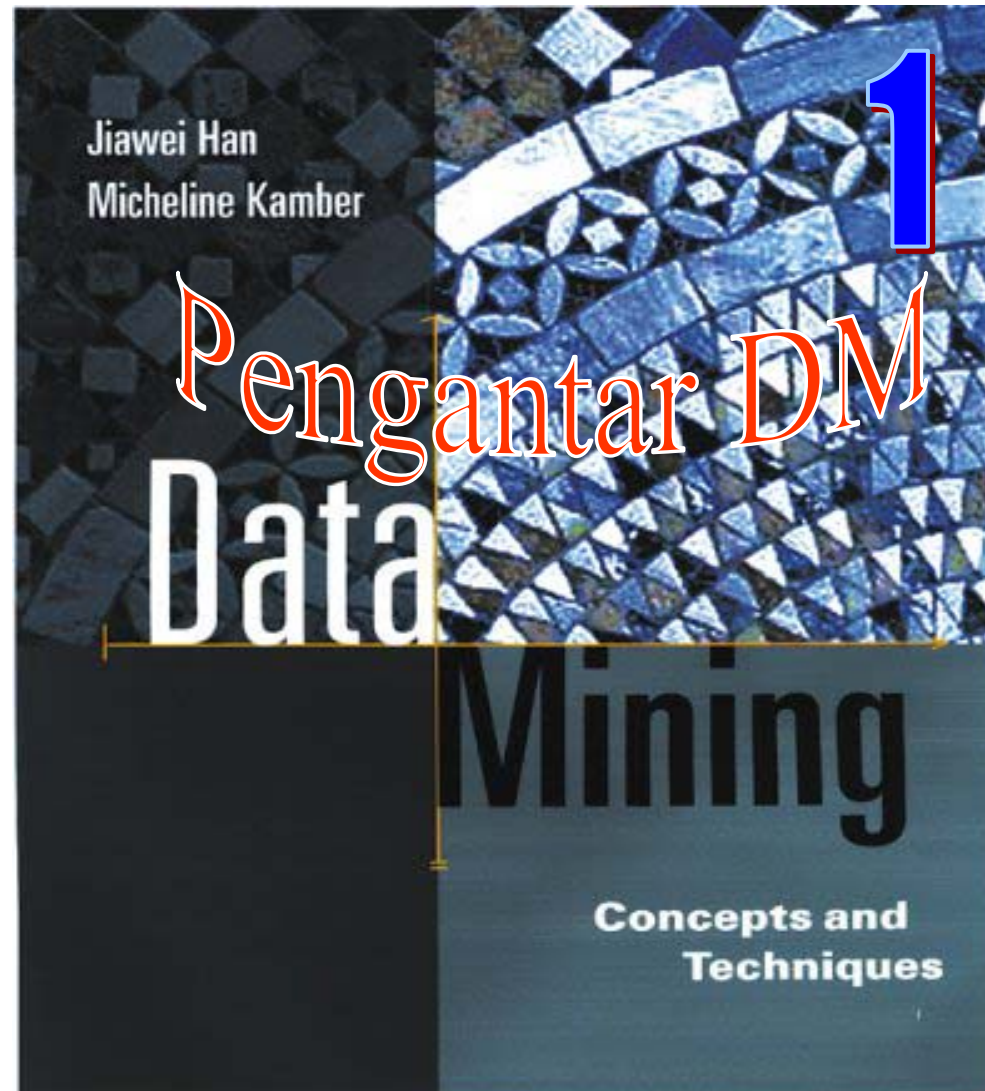
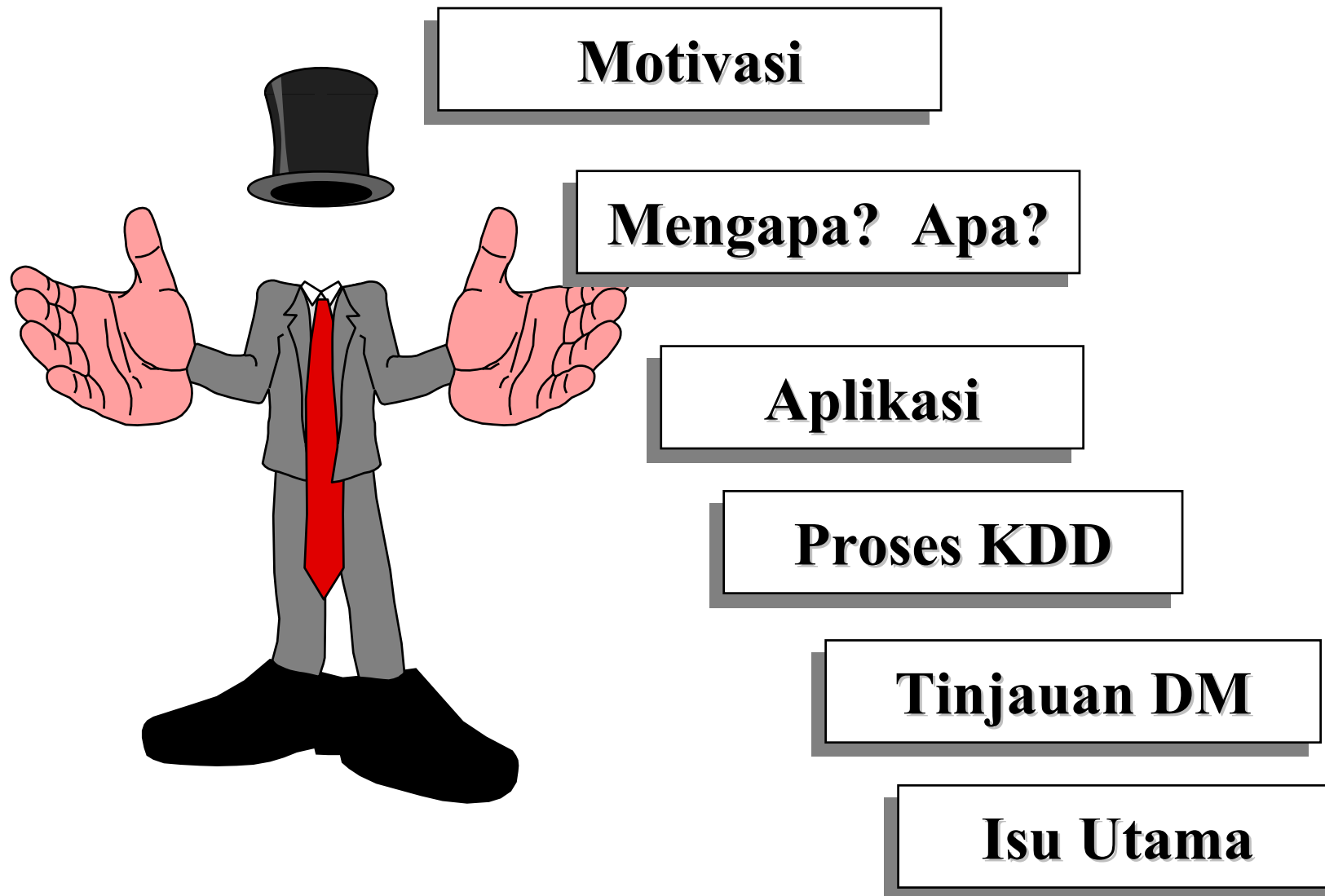


# Konsep dan Teknik Data Mining



# Pengantar Data Mining (DM)



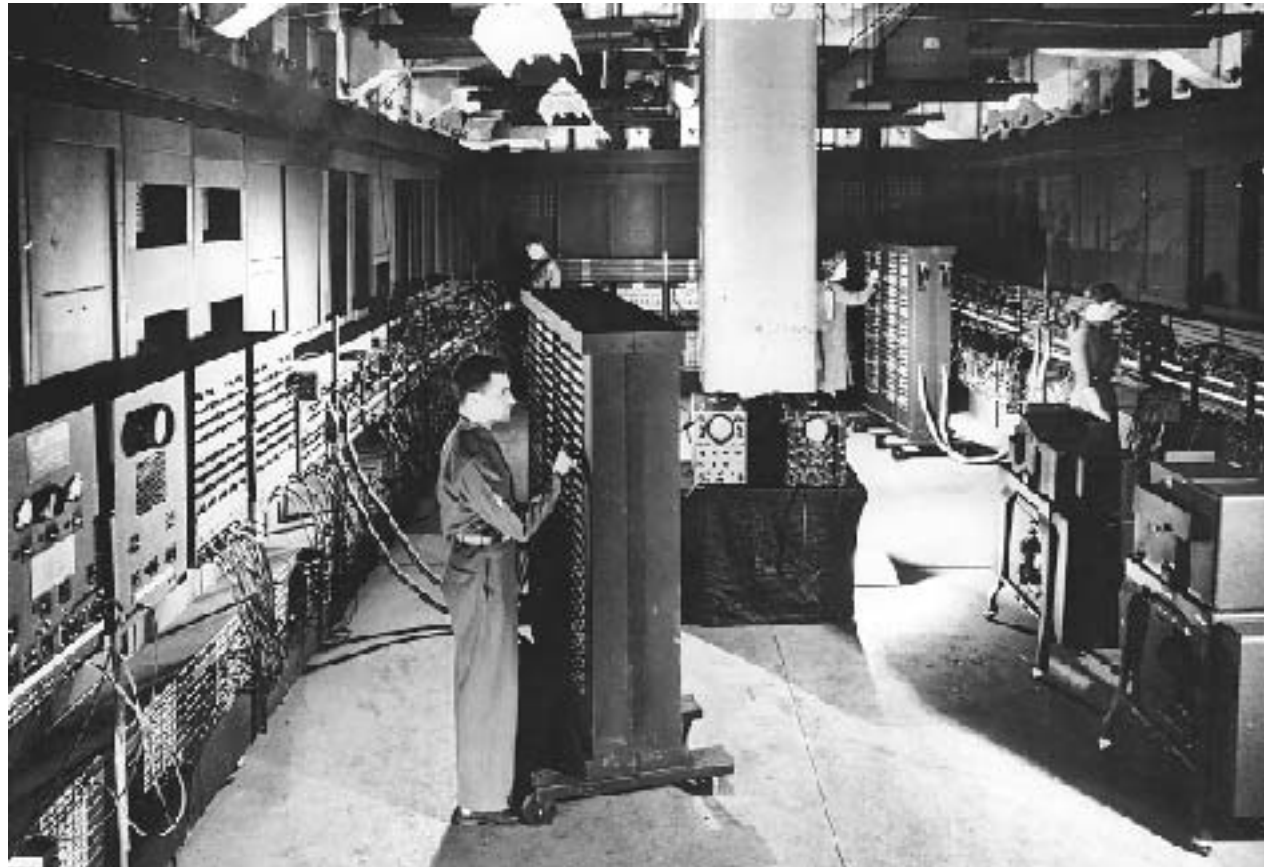
# Motivasi: “Kebutuhan akan Pengetahuan yang ada pada Data”

- Problem ledakan data
  - Tool koleksi data otomatis dan perkembangan teknologi database menyebabkan banyak sekali data yang bisa dikumpulkan di dalam database, data warehouse, dan alat penyimpanan informasi lainnya, untuk dianalisa
- Kita punya banyak data tapi tidak tahu pengetahuan yang tersimpan di dalamnya!

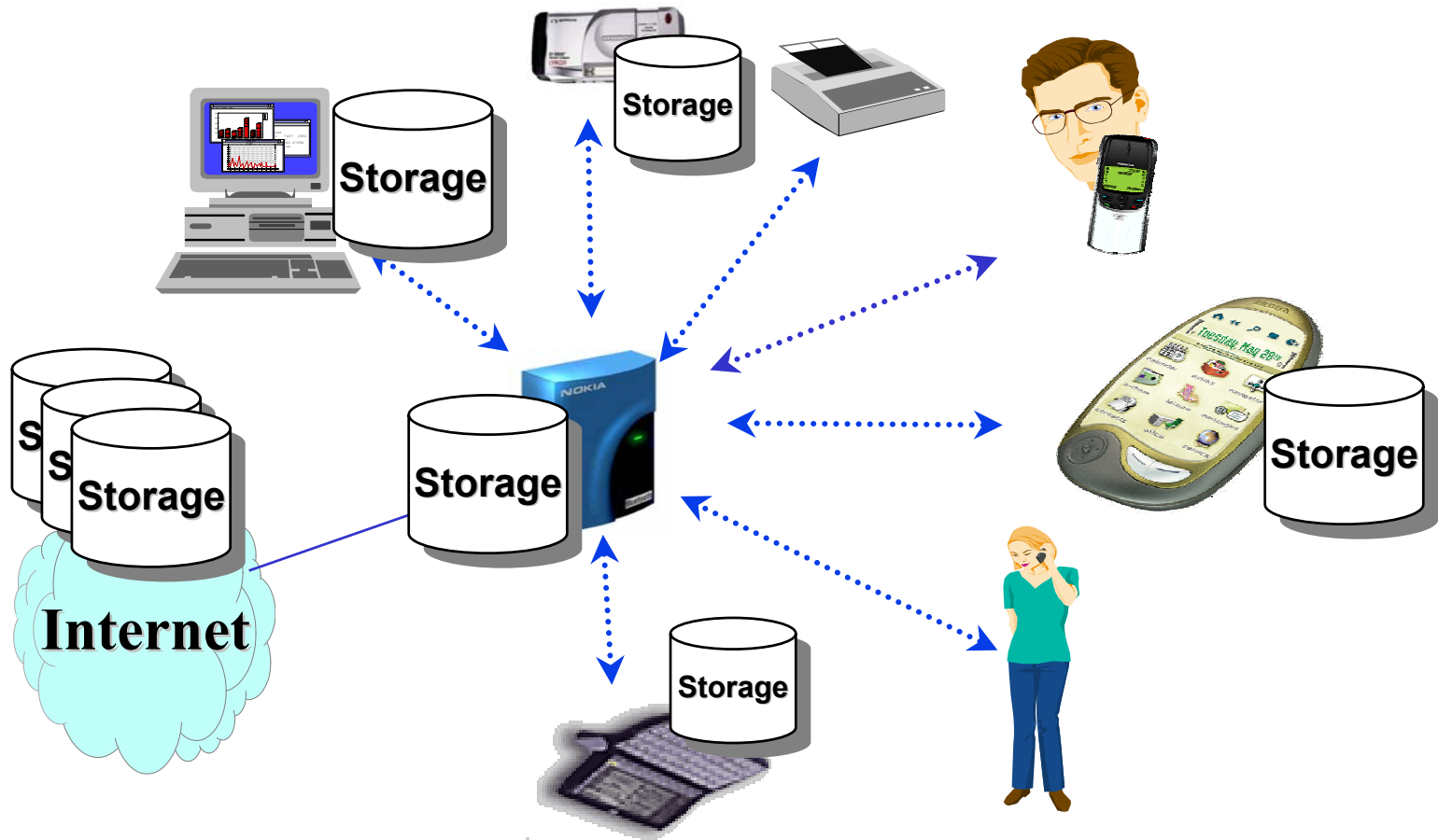
# Motivasi: “Kebutuhan akan Pengetahuan yang ada pada Data”

- Solusi: Penggudangan data dan penambangan data (Data warehousing and data mining)
  - Data warehousing dan on-line analytical processing (OLAP)
  - Penyaringan pengetahuan yang menarik (kaidah, keberaturan, pola, kendala) dari data dalam database yang besar

# Komputer Tahun 1940-an (ENIAC)



# Personal Home Network Tahun 2000-an



# **Evolusi atau Perkembangan Teknologi Database**

- **1960an:** Koleksi data, pembuatan data, IMS dan network DBMS
- **1970an:** Model data relasional dan implementasi DBMS relasional
- **1980an:** RDBMS, model data lanjut (extended-relational, OO, deduktif, dsb.) DBMS berorientasi aplikasi (spasial, saintifik, teknik, dsb.)
- **1990an – 2000an:** Data mining dan data warehousing, database multimedia, teknologi Web

# Apa Itu Data Mining?



- **Pada dasarnya: Perolehan pengetahuan dari data**
  - “Ekstraksi informasi atau pola yang menarik (tidak sepele, implisit, tak-diketahui sebelumnya, mungkin bermanfaat) dari data didalam database yang besar”
- ⇒ **Penyelidikan: analisa data semi-otomatis atas sekumpulan data yang besar**

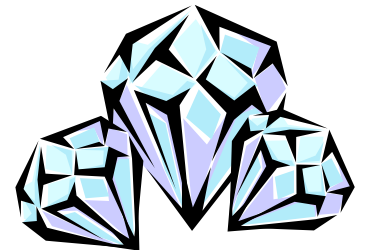




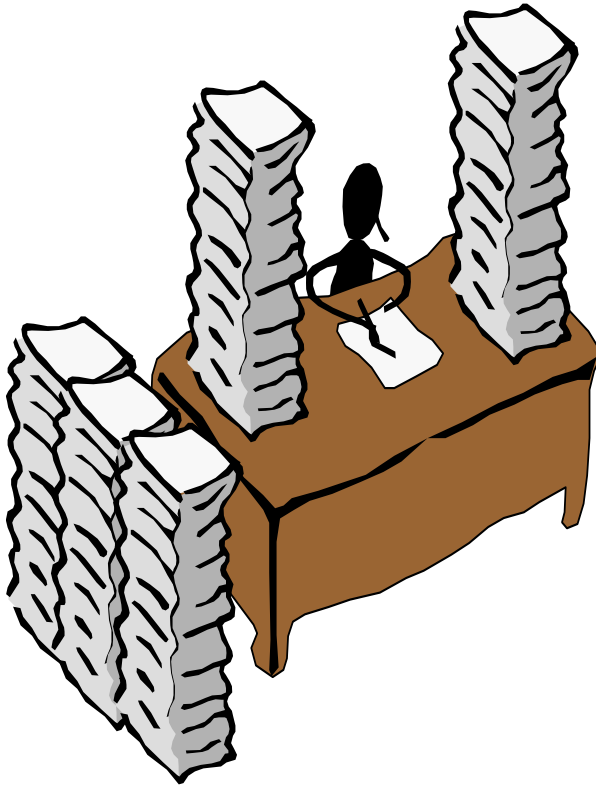
# Apa Itu Data Mining?



- **Istilah yang agak baku:**
  - Data mining
    - Biasanya DM adalah salah satu proses KDD
  - Knowledge discovery in databases (KDD)
    - Istilah umum yang meliputi, preprocessing data, DM, dan postprocessing
- **Istilah yang jarang digunakan:**
  - Ekstraksi pengetahuan, analisa data/pola
- **Publisitas terbaru:**
  - Kecerdasan bisnis (business intelligence), manajemen pengetahuan (knowledge management)



# Mengapa Data Mining?



- **Ketersediaan data dalam jumlah yang sangat besar:**
  - Tool koleksi data otomatis dan perkembangan teknologi database menyebabkan banyak sekali data yang bisa dikumpulkan di dalam database, data warehouse, dan alat penyimpanan informasi lainnya
  - Pemeriksaan data manual adalah membosankan dan terkadang tidak masuk akal

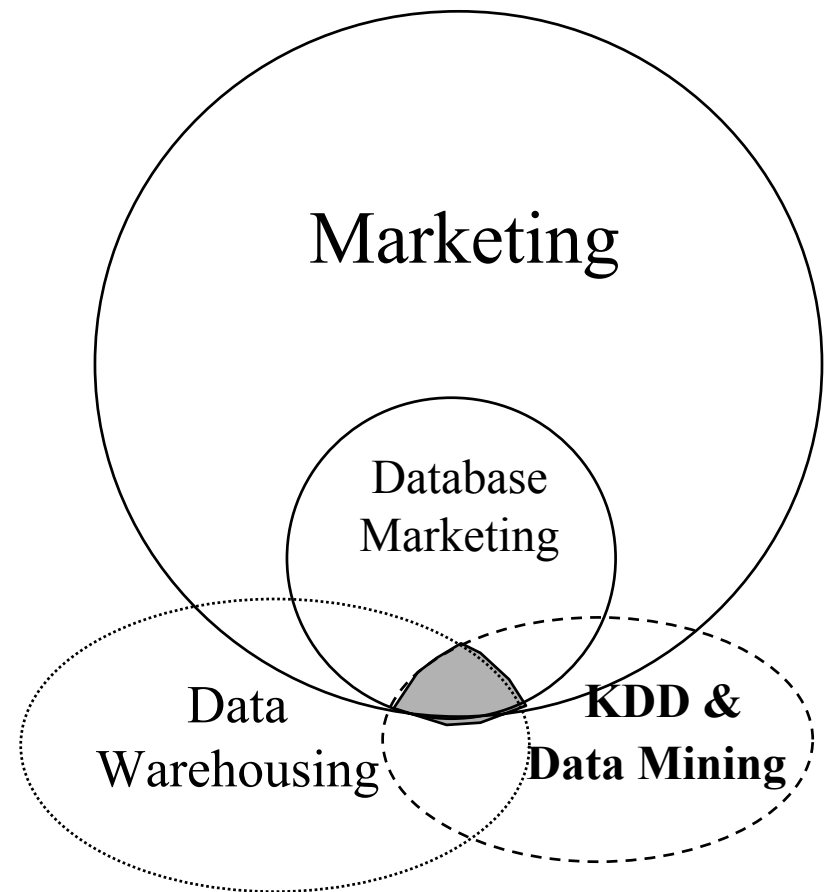
# Apa Manfaat Dari DM?

**Meningkatkan pengetahuan agar bisa membuat keputusan berdasarkan:**

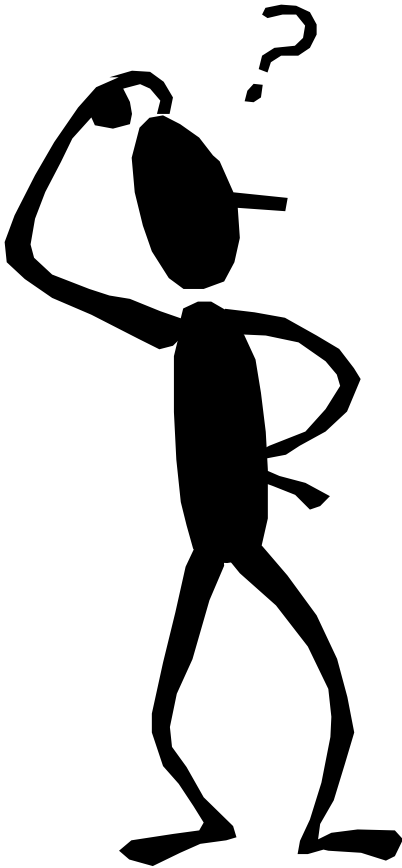
**Misal, pengaruh pada marketing**

**Peran dan pengaruh DM yang bertumbuh pesat dan masih bertumbuh!**

**Tetapi DM tidaklah sekedar marketing...**

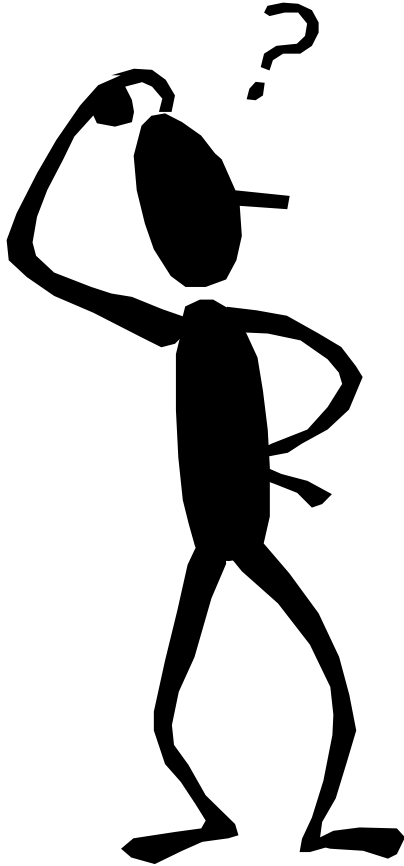


# Potensi Aplikasi?



- **Analisis database dan dukungan keputusan:**
  - Analisis dan manajemen pasar
    - Target pasar, manajemen relasi customer (CRM), analisis keranjang belanja, penjualan silang, segmentasi pasar
  - Analisis dan manajemen resiko
    - Peramalan, tindakan mempertahankan customer, peningkatan asuransi, kontrol kualitas, analisis kompetitif

# Potensi Aplikasi?



- **Aplikasi lain:**
  - Text mining (news group, email, dokumen) dan Web mining
  - Stream data mining
  - Analisis DNA dan bio data

# **Analisis dan Manajemen Pasar**

- **Dari mana data berasal?**
  - Transaksi kartu kredit, loyalty cards, kupon discount, keluhan customer, kajian lifestyle publik
- **Target Pasar**
  - Mendapatkan kelompok model customer yang berbagi karakteristik yang sama: minat, tingkat pendapatan, kebiasaan belanja, dsb.
  - Menentukan pola pembelian customer berdasarkan waktu
- **Analisis lintas pasar**
  - Asosiasi/korelasi antara penjualan produk & taksiran berdasarkan asosiasi demikian

# **Analisis dan Manajemen Pasar**

- **Profil customer**
  - Tipe customer apa membeli produk apa (pengelompokan atau klasifikasi)
- **Analisa kebutuhan customer**
  - Mengenali produk terbaik untuk customer berbeda
  - Meramalkan faktor apa yang akan memikat customer baru
- **Penyediaan rangkuman informasi**
  - Rangkuman laporan multidimensi
  - Rangkuman informasi statistik (kecenderungan data terpusat dan variasi)

# **Analisis dan Manajemen Resiko Perusahaan**

- **Perencanaan keuangan dan evaluasi aset**
  - Analisis dan peramalan cash flow
  - Analisis ganti rugi yang mungkin untuk mengevaluasi aset
  - Analisis cross-sectional dan time series (financial-ratio, analisa trend, dsb.)
- **Perencanaan sumberdaya**
  - Merangkum dan membandingkan sumberdaya dan pengeluaran



# **Analisis dan Manajemen Resiko Perusahaan**

- **Kompetisi**

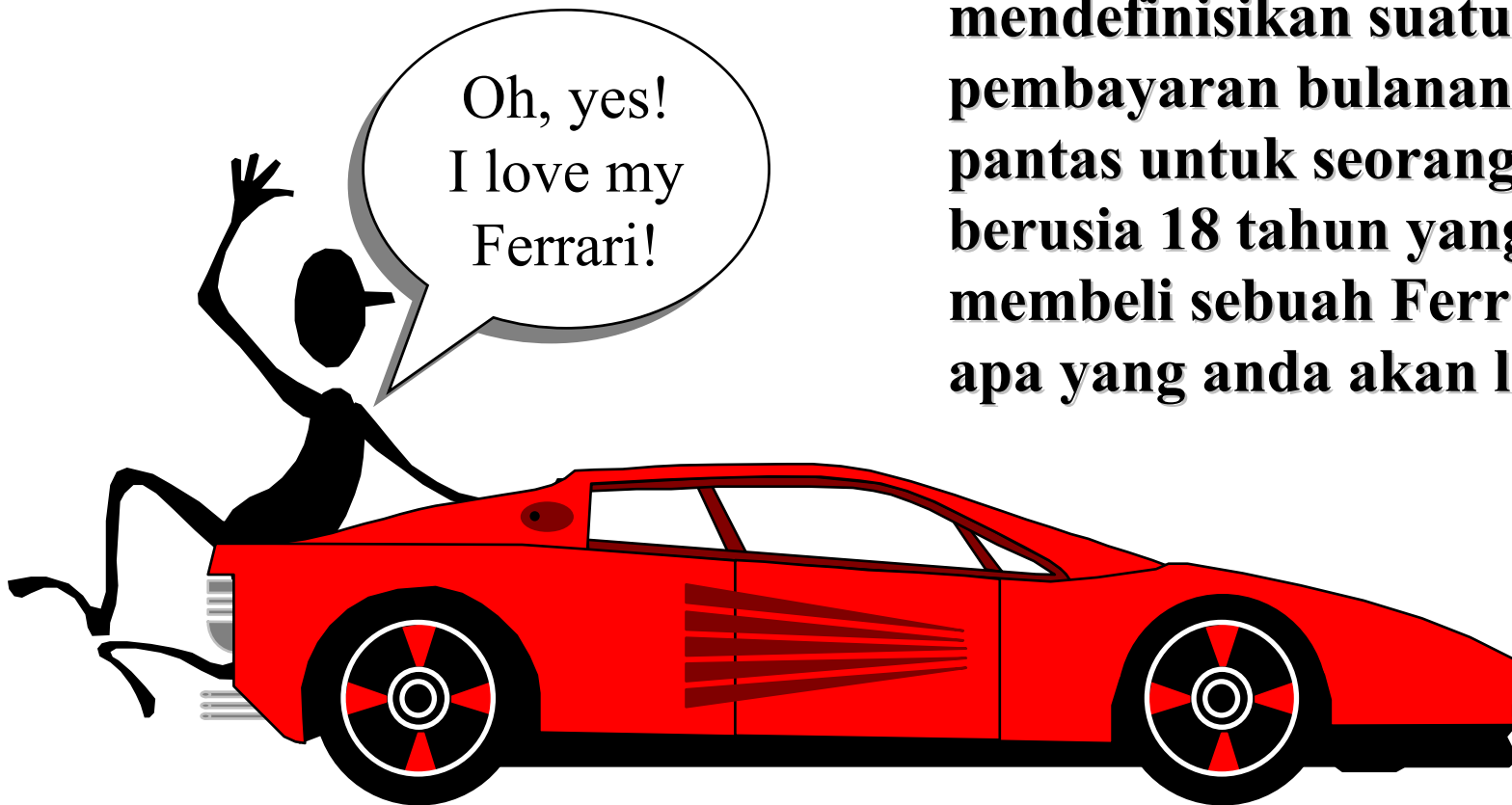
- Memantau pesaing dan arah pasar
- Mengelompokkan customer kedalam kelas dan prosedur harga berbasis kelas
- Menetapkan strategi harga dalam suatu pasar dengan kompetitif tinggi

# Aplikasi Lain

- **Olah raga**
  - IBM Advanced Scout menganalisa statistik (shots blocked, assists, dan fouls) pertandingan NBA untuk mendapatkan keuntungan kompetitif bagi New York Knicks dan Miami Heats
- **Astronomi**
  - Observatory JPL dan Palomar menemukan 22 quasars dengan bantuan data mining
- **Internet Web Surf-Aid**
  - IBM Surf-Aid menerapkan algoritma data mining untuk akses logs halaman Web yang terkait dengan pasar dalam upaya mendapatkan kesukaan dan perilaku customer, menganalisa efektifitas pemasaran Web, perbaikan situs Web organisasi, dsb.

# Contoh (1)

- **Anda seorang petugas asuransi dan anda harus mendefinisikan suatu pembayaran bulanan yang pantas untuk seorang pemuda berusia 18 tahun yang membeli sebuah Ferrari ... apa yang anda akan lakukan?**



# Contoh (1)



- **Kaji seluruh data customer dan data kompensasi pembayaran sebelumnya**
- **Kaji peluang penyebab kecelakaan paling banyak berdasarkan dugaan...**
  - Kelamin pengendara (pria/wanita) dan usia
  - Model dan usia mobil, tempat tinggal
  - dsb.
- **Jika peluang kecelakaan lebih besar dari rata-rata, aturlah pembayaran bulanan yang sesuai!**

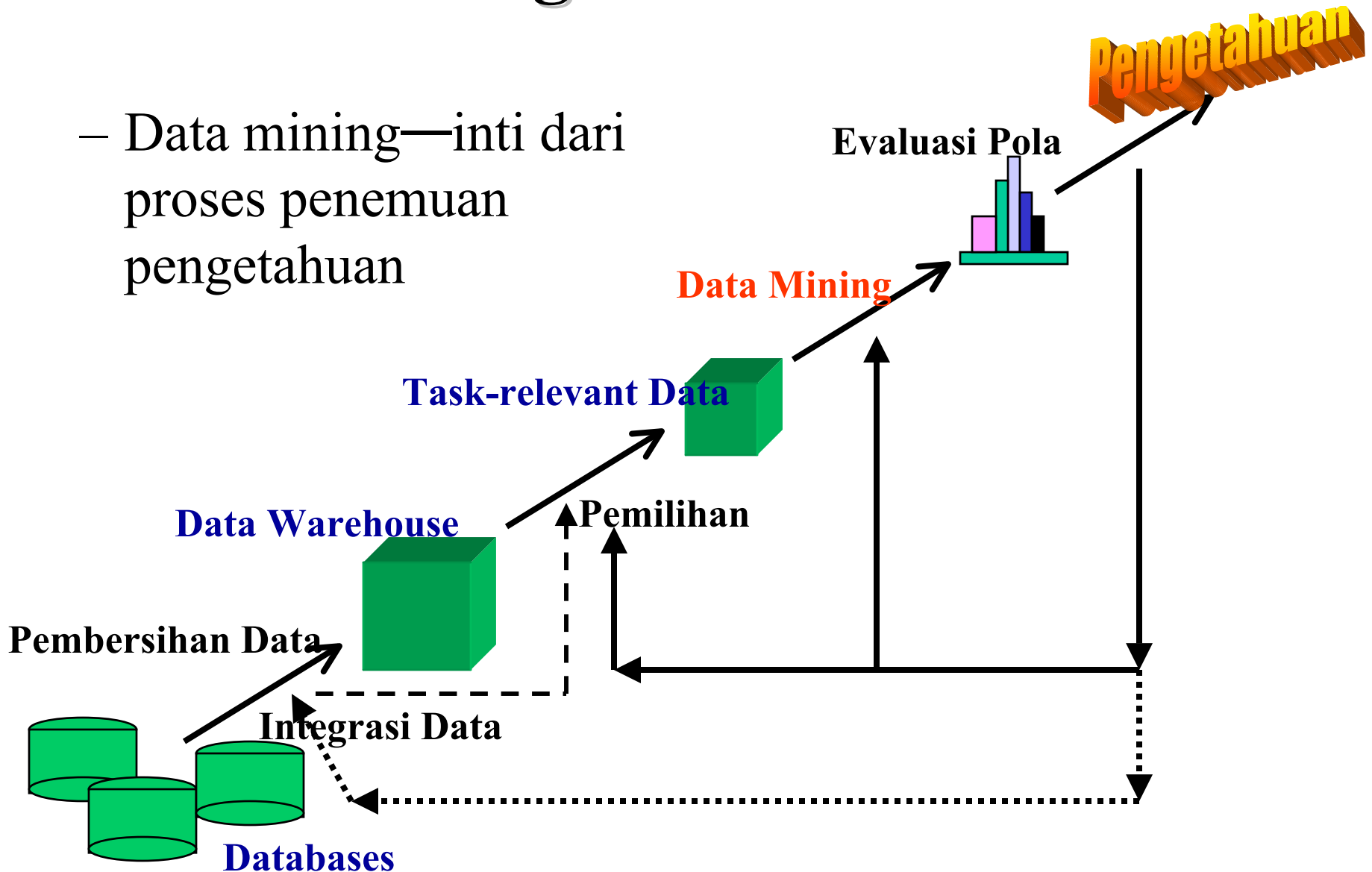
# Contoh (2)

- **Log pengaksesan Web bisa dianalisis untuk ...**
  - Mendapatkan apa kesukaan customer
  - Memperbaiki situs Web organisasi
- **Demikian pula ...**
  - Seluruh jenis analisis log informasi
  - Adaptasi antarmuka/layanan user

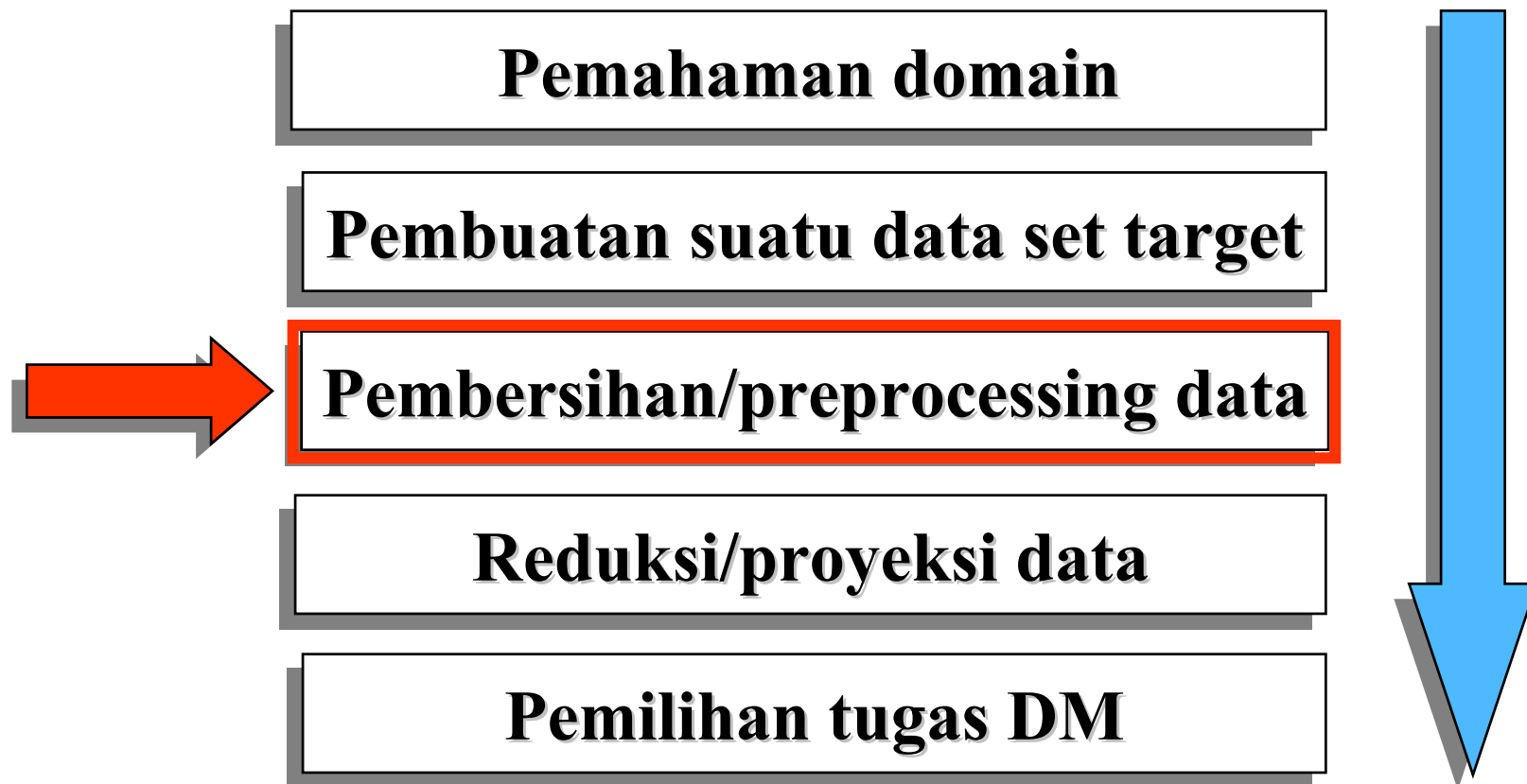


# Data Mining: Suatu Proses KDD

- Data mining—inti dari proses penemuan pengetahuan



# Langkah-Langkah dari Proses KDD (1)



# Langkah-Langkah dari Proses KDD (2)

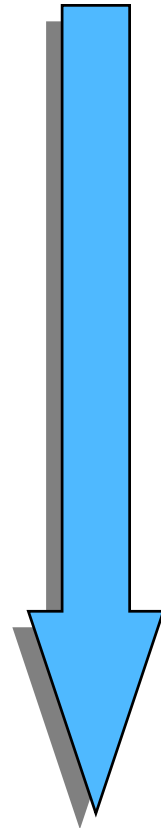
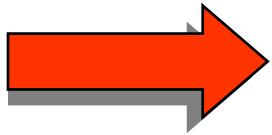
**Pemilihan Algoritma DM**

**Data mining: Pencarian**

**Evaluasi pola**

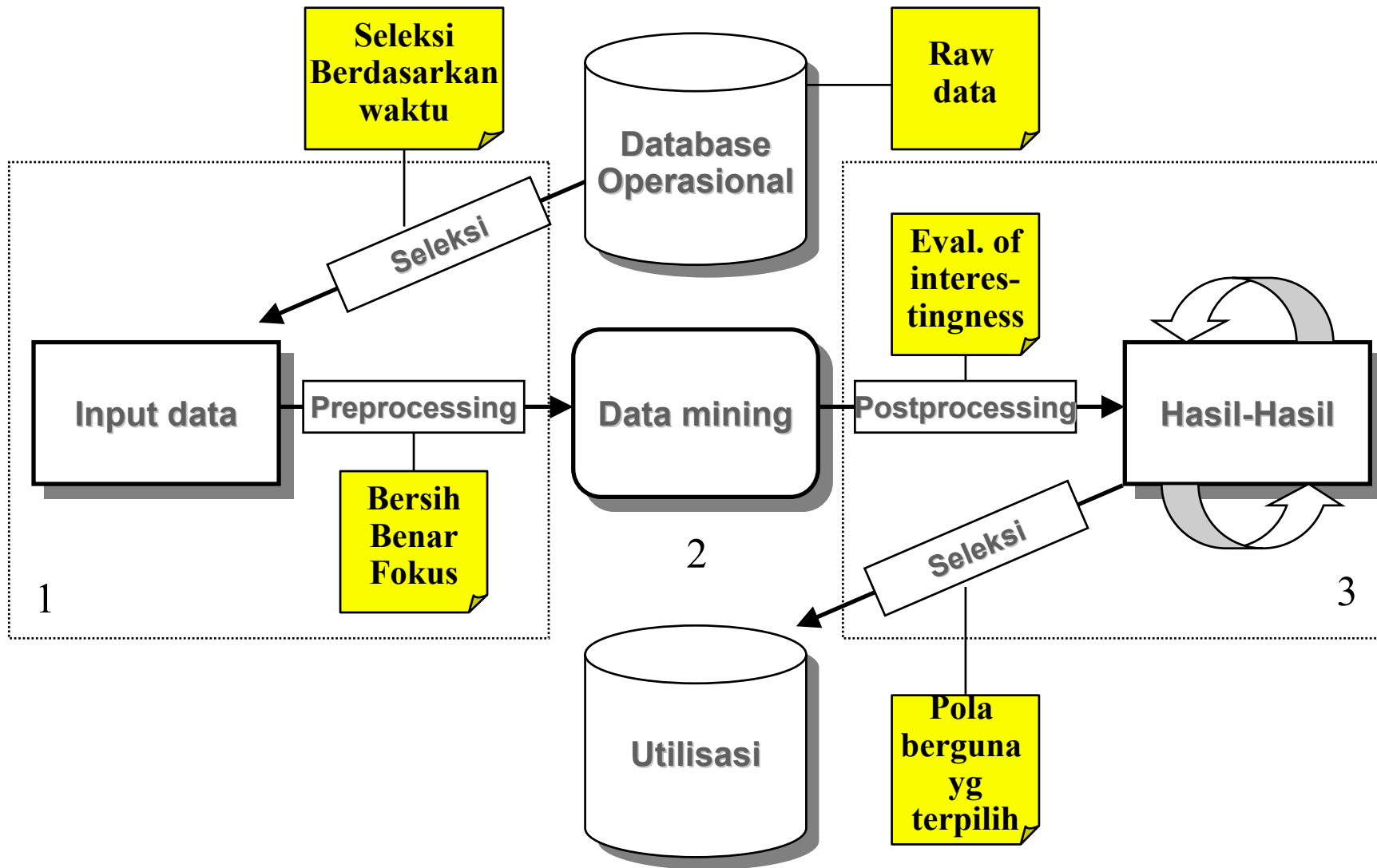
**Penyajian pengetahuan**

**Penggunaan pengetahuan yang diperoleh**

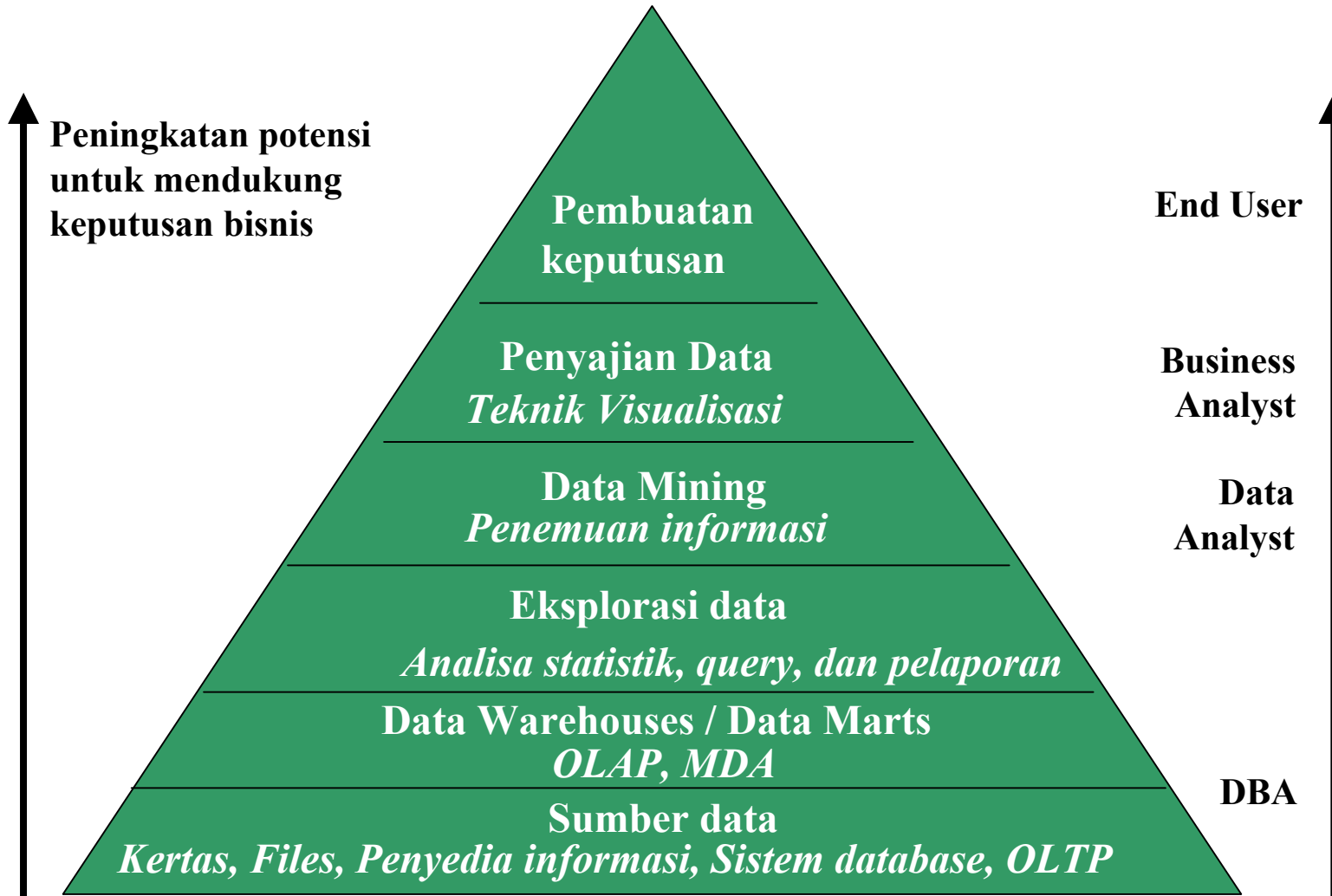




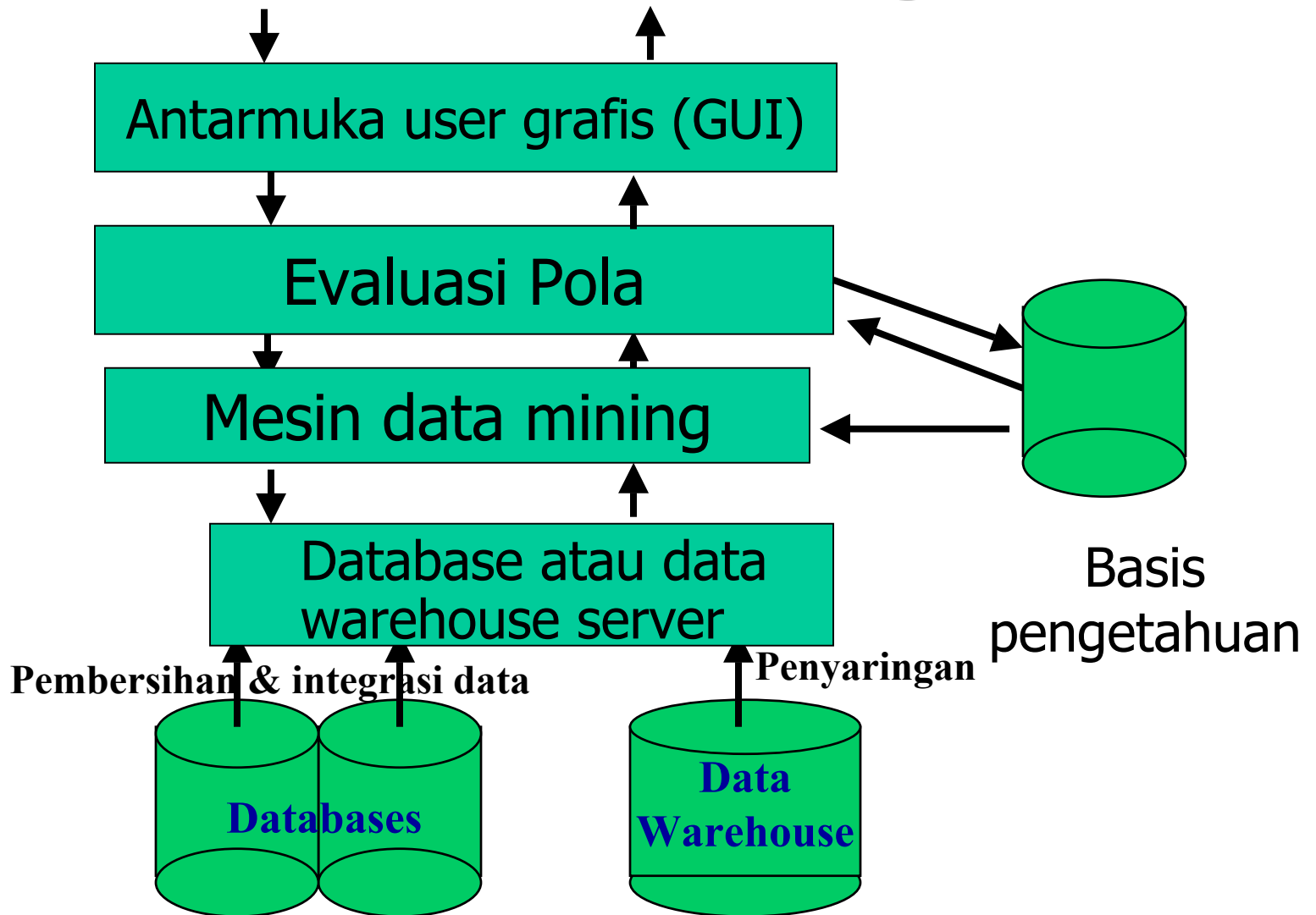
# Ciri Khas Proses KDD



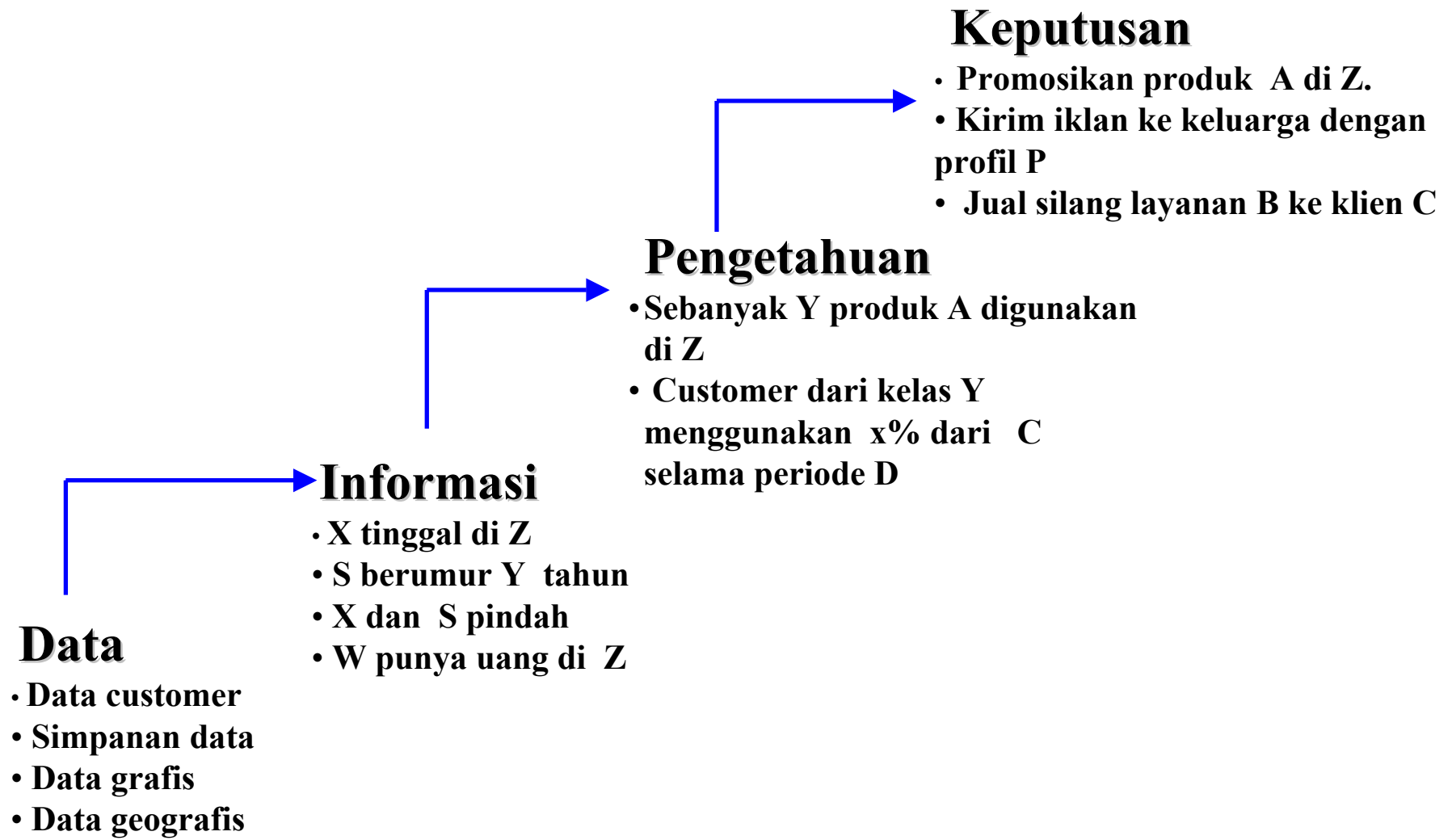
# Utilisasi



# Arsitektur: Sistem Data Mining



# Rantai Nilai



# Fungsionalitas Data Mining

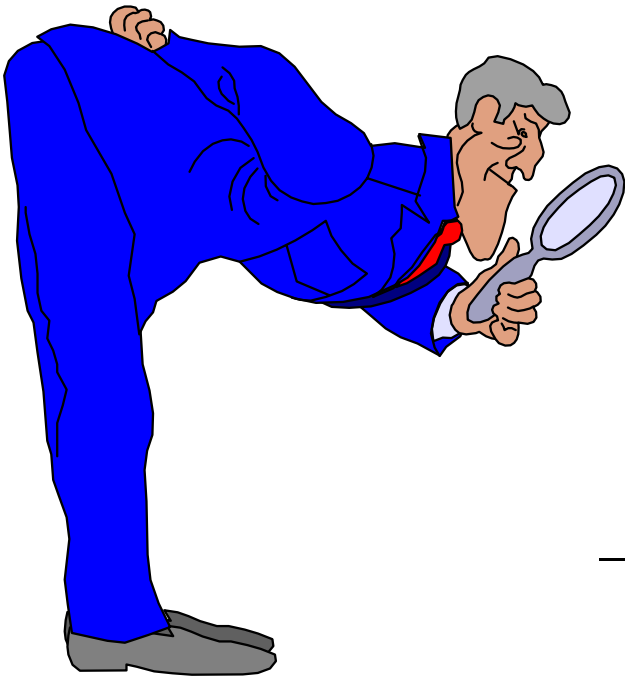
- **Klasifikasi dan prediksi**

- Membangun model (fungsi) yang menguraikan dan membedakan kelas atau konsep untuk peramalan kedepan

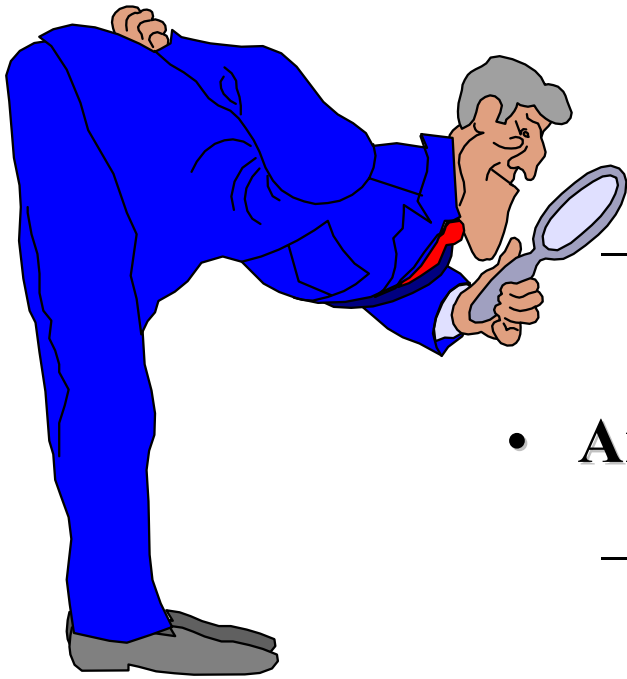
- Misal, mengklasifikasikan negara berdasarkan iklim, atau mengklasifikasikan mobil berdasarkan gas mileage

- Presentasi: pohon-keputusan, kaidah klasifikasi, neural network

- Menaksir beberapa nilai numerik yang tidak diketahui atau hilang



# Fungsionalitas Data Mining



- **Analisis cluster (analisis pengelompokan)**
  - Label kelas tidak diketahui: kelompokkan data untuk membentuk kelas baru, misal mengelompokkan rumah untuk mendapatkan pola distribusi
  - Memaksimalkan kemiripan antar kelas dan meminimumkan kemiripan didalam kelas
- **Analisis outlier**
  - Outlier: suatu objek data yang tidak mengikuti perilaku umum dari data
  - Gangguan atau pengecualian? Tidak! Berguna dalam deteksi kecurangan, analisis peristiwa yang jarang terjadi

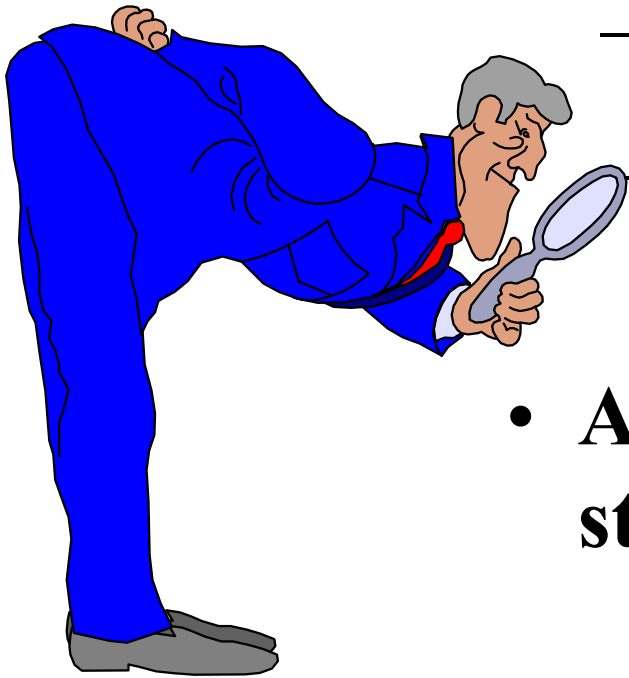
# Fungsionalitas Data Mining

- **Analisis trend dan evolusi**

- Trend dan deviasi: analisis regresi

- Penggalian pola sekuensial, analisis periodisitas

- **Analisis arah pola lain atau statistik**



# **Adakah Seluruh Pola Yang Ditemukan Menarik?**

- **Penambahan data bisa membuat ribuan pola: tidak semua pola tersebut menarik**
  - Pendekatan yang disarankan: berpusatkan kepada manusia, berbasis query, penggalian terfokus
- **Ukuran ketertarikan**
  - Suatu pola adalah menarik jika pola tersebut mudah dipahami oleh manusia, valid (berlaku) pada data baru atau uji dengan suatu derajat kepastian, potensial berguna, baru atau membenarkan suatu hipotesa yang dicari user untuk konfirmasi



# **Adakah Seluruh Pola Yang Ditemukan Menarik?**

- **Ukuran ketertarikan objektif vs. subjektif**
  - Objektif: berdasarkan statistik dan struktur pola, misal, dukungan, kepercayaan, dsb.
  - Subjektif: berdasarkan keyakinan user terhadap data, misal, ketakterdugaan, pengalaman baru, tindakan yang bisa diperkarakan, dsb.

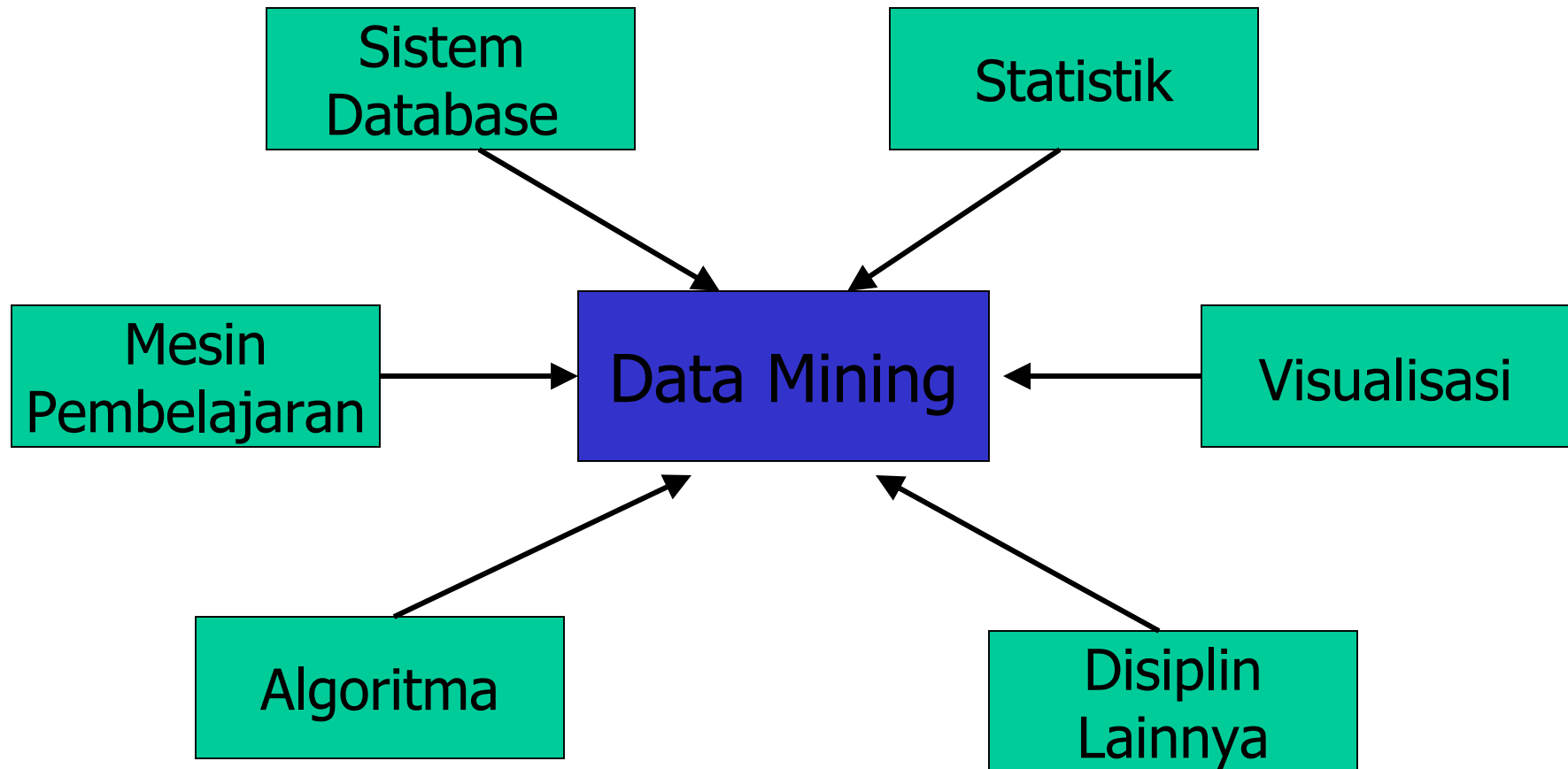
# **Bisakah Memperoleh Seluruh Atau Hanya Pola Yang Menarik?**

- **Dapatkan seluruh pola yang menarik:  
Kelengkapan**
  - Bisakah suatu sistem data mining mendapatkan seluruh pola yang menarik?
  - Pencarian heuristik vs. exhaustive (melelahkan)
  - Asosiasi vs. klasifikasi vs. clustering

# **Bisakah Memperoleh Seluruh Atau Hanya Pola Yang Menarik?**

- **Cari hanya pola yang menarik: suatu problem optimasi**
  - Bisakah suatu sistem data mining mendapatkan hanya pola yang menarik?
  - Pendekatan
    - Pertama dapatkan seluruh pola lalu saring atau keluarkan yang tidak diminati
    - Buatlah hanya pola yang diminati—optimisasi query penggalian

# Data Mining: Tempat Pertemuan Banyak Disiplin



# Tinjauan Data Mining: Skema Klasifikasi

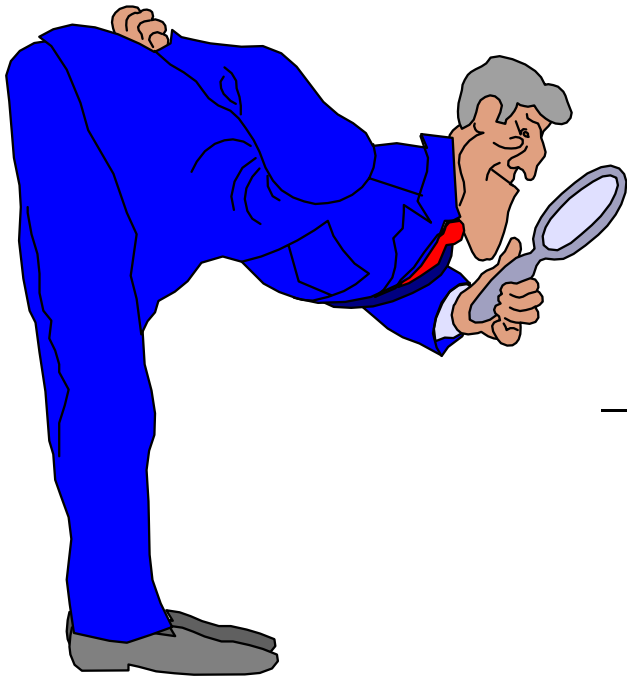
- **Fungsionalitas umum:**

- Uraian data mining:

- Uraikan hal menarik apa yang bisa ditemukan dalam data ini!
- Terangkan data ini ke saya!

- Peramalan data mining:

- Berdasarkan data ini dan sebelumnya, beritahu saya apa yang akan terjadi kemudian!
- Tunjukkan ke saya trend kedepan!



# Tinjauan Data Mining: Skema Klasifikasi



- **Tinjauan multi-dimensi ...**
  - Databases yang akan digali
  - Pengetahuan yang akan dicari
  - Teknik-teknik yang digunakan
  - Aplikasi yang disesuaikan
- **Mari kita lihat lebih dekat pada tinjauan ini ...**

# Tinjauan Data Mining



## Databases yang akan digali

- Relasional
- Data warehouse
- Transaksional
- Object-oriented
- Object-relational
- Aktif
- Spasial
- Time-series
- Teks, XML
- Multi-media
- Heterogen
- Warisan
- Induktif
- WWW
- dsb.

# Tinjauan Data Mining



## Aplikasi yang disesuaikan

- Retail (supermarkets etc.)
- Telecom
- Banking
- Fraud analysis
- DNA mining
- Analisa stock market
- Web mining
- Analisa data log
- dsb.



# **OLAP Mining: Integrasi Dari Data Mining dan Data Warehouse**

- **Sistem data mining, DBMS, Data warehouse systems coupling**
  - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- **Data on-line analytical mining (OLAM)**
  - Integrasi dari penggalian dan teknologi OLAP

# **OLAP Mining: Integrasi Dari Data Mining dan Data Warehouse**

- **Penggalian pengetahuan multi-level secara interaktif**
  - Kebutuhan penggalian pengetahuan dan pola pada suatu level berbeda dari abstraksi dengan drilling/rolling, pivoting, slicing/dicing, dsb.
- **Integrasi dari banyak fungsi penggalian**
  - Klasifikasi berkarakter, pertama clustering dan kemudian asosiasi

# Isu Utama dalam Data Mining

- **Metodologi dan interaksi penambangan**
  - Penambangan berbagai jenis pengetahuan
  - Penambangan pengetahuan secara interaktif
  - Penggabungan latar belakang pengetahuan
  - Bahasa query DM dan DM khusus
  - Visualisasi hasil DM
  - Penanganan gangguan dan data tak lengkap
  - Ketertarikan problem
- **Kinerja dan skalabilitas:**
  - Efisiensi dan skalabilitas dari algoritma DM
  - Metoda penambangan paralel, tersebar dan bertumbuh

# Isu Utama dalam Data Mining

- **Aneka ragam tipe data:**
  - Penanganan tipe data kompleks
  - Penambangan informasi dari database heterogen (Web misalnya)
- **Aplikasi dan integrasi pengetahuan yang didapat:**
  - Tool DM khusus domain
  - Jawaban query cerdas dan pembuatan keputusan
  - Integrasi dari pengetahuan yang didapat dengan pengetahuan yang ada
- **Proteksi data ...**
  - Keamanan
  - Integritas
  - Kerahasiaan



# Historis Aktivitas Data Mining

- **1989 IJCAI Workshop**
- **1991-1994 KDD Workshops**
- **1995-1998 KDD Conferences**
- **1998 ACM SIGKDD**
- **1999- SIGKDD Conferences**
- **dan banyak lagi konferensi kecil / baru dari DM ...**
  - PAKDD, PKDD
  - SIAM-Data Mining, (IEEE) ICDM
  - dsb.

# Rujukan Yang Berguna untuk Data Mining

## “Standards”

- **DM:**
  - Conferences: KDD, PKDD, PAKDD, ...
  - Journals: Data Mining and Knowledge Discovery, CACM
- **DM/DB:**
  - Conferences: ACM-SIGMOD/PODS, VLDB, ...
  - Journals: ACM-TODS, J. ACM, IEEE-TKDE, JIS, ...
- **AI/ML:**
  - Conferences: Machine Learning, AAAI, IJCAI, ...
  - Journals: Machine Learning, Artific. Intell., ...

# Kesimpulan

- **Data mining: penemuan pola menarik dari data set yang besar secara semi-otomatis**
- **Knowledge discovery adalah suatu proses:**
  - Preprocessing
  - Data mining
  - Postprocessing
- **Untuk digali, digunakan atau dimanfaatkan ...**
  - Databases (relasional, object-oriented, spasial, WWW, ...)
  - Pengetahuan (karakterisasi, pengumpulan, asosiasi, ...)
  - Teknik (mesin pembelajaran, statistik, visualisasi, ...)
  - Aplikasi (retail, telecom, Web mining, analisa log, ...)