Pengantar Data Mining

oleh:
Entin Martiana

Pengantar

- Mengapa data mining?
- Apa data mining?
- Data Mining: data apa saja?
- Fungsi data mining
- Model dalam data mining
- Fungsi dalam data mining
- Permasalahan dalam data mining
- Aplikasi pada data mining
- 10 algoritma data mining yang paling umum

Mengapa Datamining

We are drowning in data,

but starving for knowledge!

Mengapa DM:
Banjir Data

- Twitter: 8000an tweet per detik → 600 juta tweet per hari.
- Facebook: 30 milyar item (link, status, note, foto dst) per bulan. 500 juta user menghabiskan 700 milyar menit per bulan di situs FB.
- Indomaret: 4500an gerai, asumsikan 3 transaksi per menit = 12 juta transaksi per hari se Indonesia.
- Kartu kredit visa: berlaku di 200 negara. 10 ribu transaksi per detik → 850 juta transaksi per hari.

Mengapa data mining?

- Digitalisasi, kemajuan sistem informasi → data, data, data (Tera → Peta)
- Web → berita, blog, twitter, forum, flickr, fb, youtube
- Streaming data → twitter, f4, sensor (satelit)

Evolusi DB

- 60-an: koleksi data (file system primitif)
- 70-80: MIS (Sistem Informasi Management)
- 80-sekarang: OO, Deductive, Spatial, Multimedia
- 90-sekarang: Web based (XML, web mining), Datawarehouse, OLAP, Text Database, Text + Data mining
- 05-sekarang: Stream data management and mining, Cloud, Web

Apa Data Mining?

- Data mining (pencarian pengetahuan dari data)
  - Mengekstrak secara otomatis pola atau pengetahuan yang menarik (tidak sederhana, tersembunyi, tidak diketahui sebelumnya, berpotensi berguna) dari data dalam jumlah sangat besar.

Apa Datamining? (lanj)

- Nama alternatif: Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence dsb
- Keuntungan bagi organisasi yang menerapkan data mining?

Keuntungan Datamining

- Perusahaan fokus ke informasi yg berharga di datawarehouse/databasenya.
- Meramalkan masa depan → perusahaan dapat mempersiapkan diri

Contoh:

Midwest grocery chain menggunakan DM untuk menganalisisi pola pembelian: saat pria membeli popok di hari Kamis dan Sabtu, mereka juga membeli minuman.

Analisis lebih lanjut: pembeli ini belanja di hari kamis dan sabtu, tapi di hari kamis jumlah item lebih sedikit. Kesimpulan yang diambil: pembeli membeli minuman untuk dihabiskan saat weekend.

Tindak lanjut: menjual minuman dengan harga full di hari Kamis dan Sabtu. Mendekatkan posisi popok dan minuman.

Contoh Aplikasi

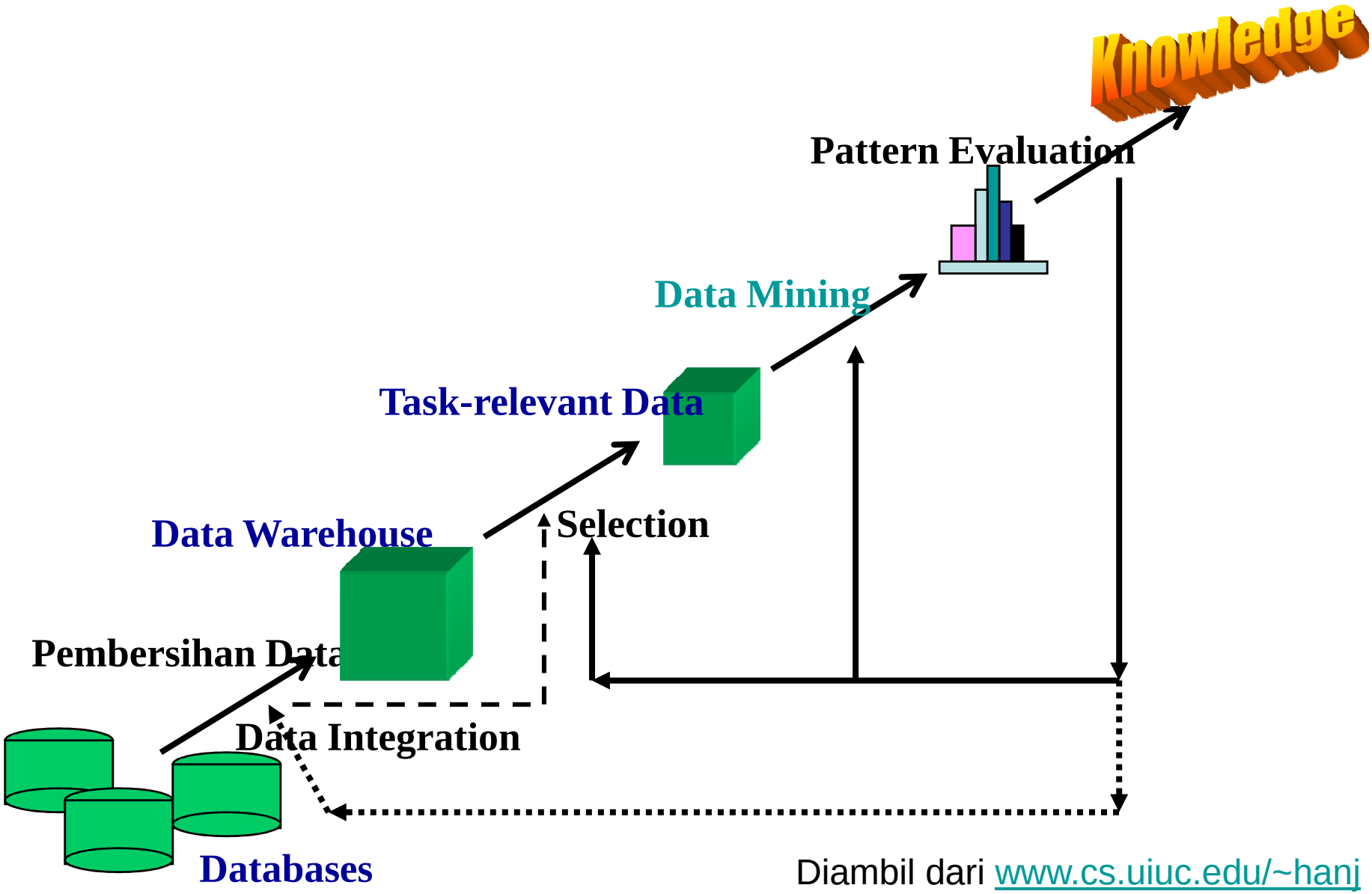
Bank me-mining transaksi customer untuk mengidentifikasi customer yang kemungkinan besar tertarik terhadap produk baru.

Setelah teknik ini digunakan, terjadi peningkatan **20 kali lipat penurunan biaya** dibandingkan dengan cara biaya.

Contoh Aplikasi

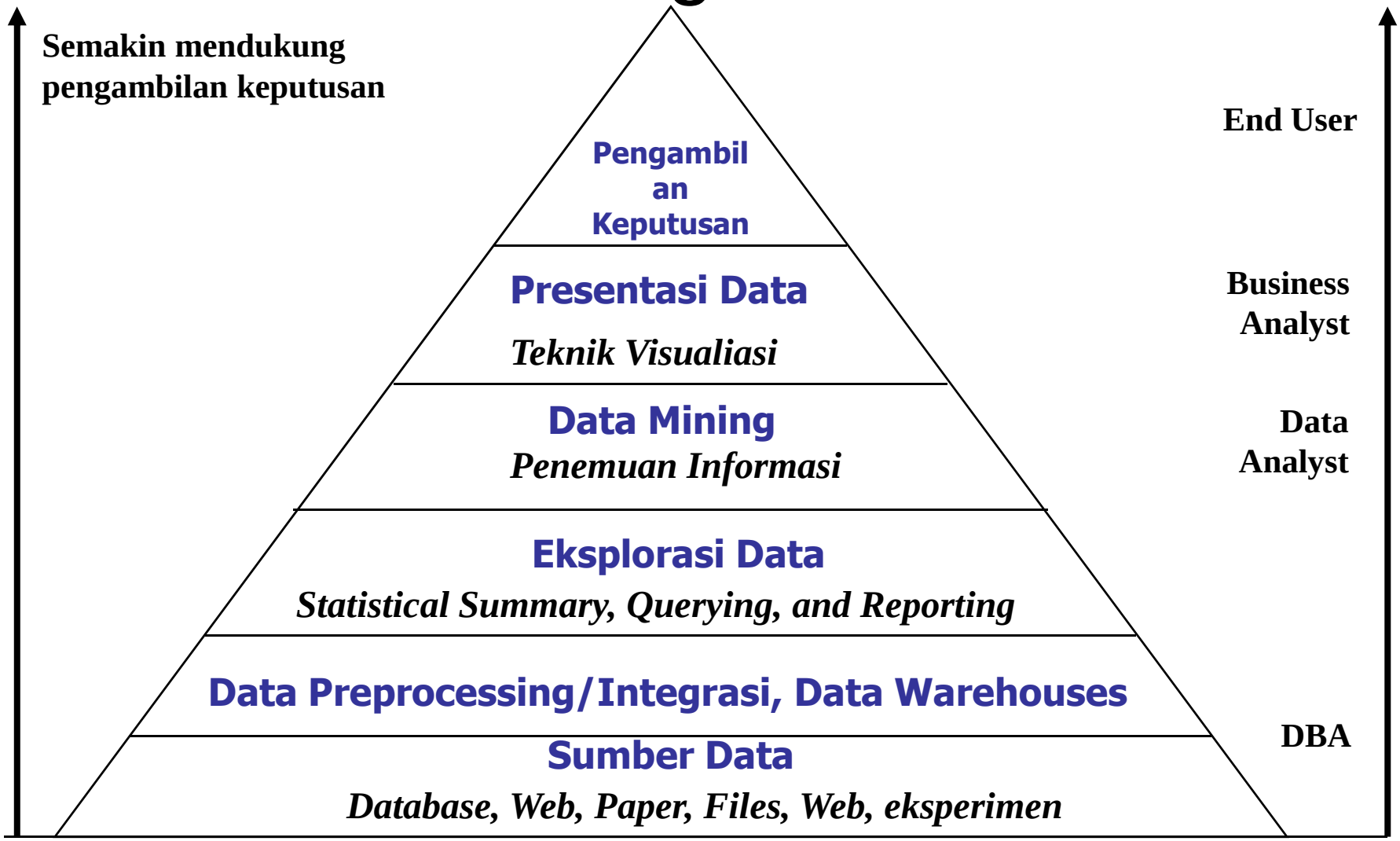
Perusahaan transportasi memining data customer untuk mengelompokan customer yang memiliki nilai tinggi yang perlu diprioritaskan.

Proses Datamining



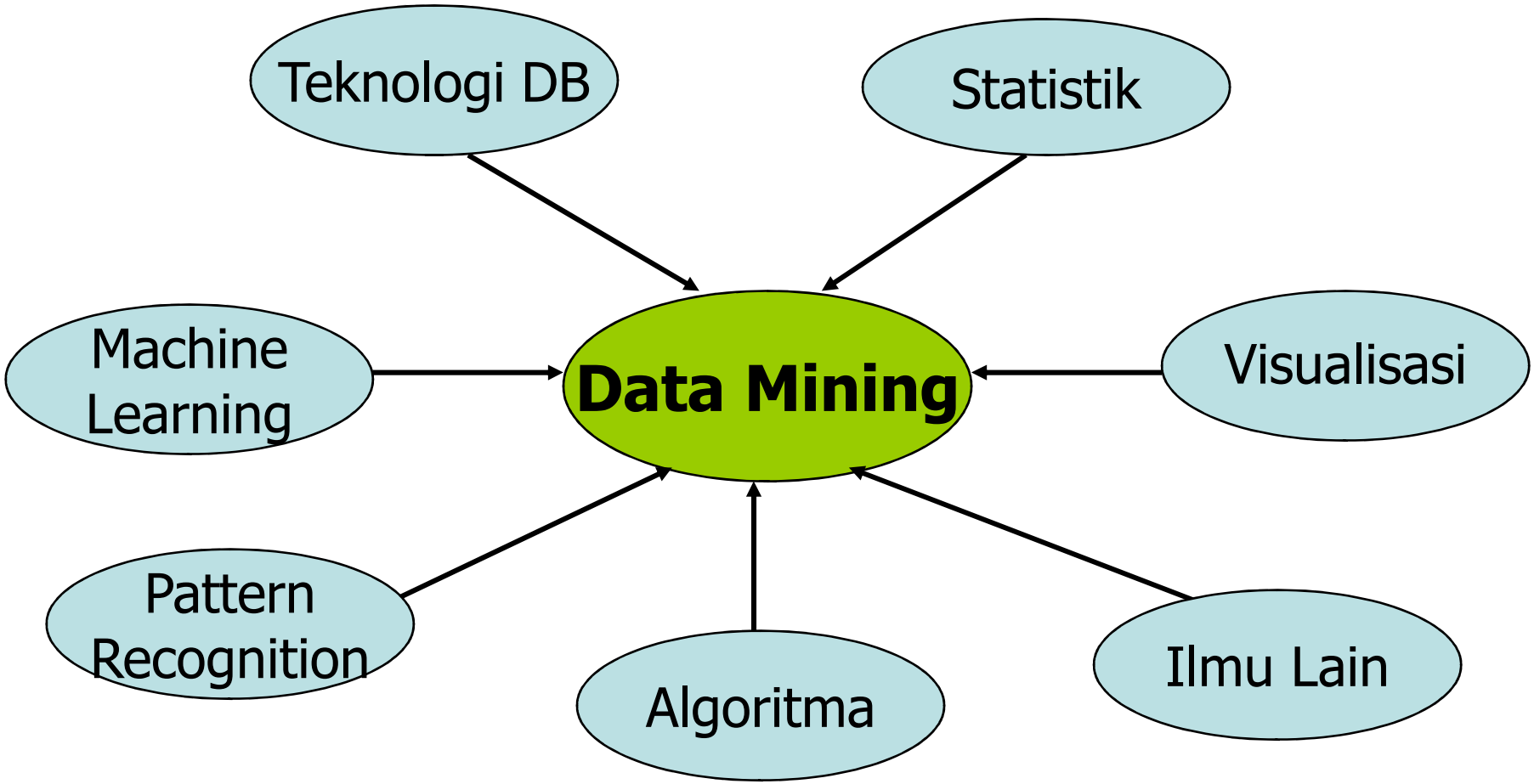
Diambil dari www.cs.uiuc.edu/~hanj

Data Mining dan Business Intelligence



Diambil dari www.cs.uiuc.edu/~hanj

Data Mining: Multi Disiplin Ilmu



Mengapa tidak analisis data biasa?

- Jumlah data yang sangat besar
  - Algoritma harus scalable untuk menangani data yang sangat besar (tera)
- Dimensi yang sangat besar: ribuan field
- Data Kompleks
  - Aliran data dan sensor
  - Data terstruktur, graph, social networddk, multi-linked data
  - Database dari berbagai sumber, database lama
  - Spasial (peta), multimedia, text, web
  - Software Simulator

Data Mining dari berbagai sudut pandang

- **Data**
  - Relational, datawarehouse,web, transaksional, stream, OO, spasial, text, multimedia
- **Pengetahuan yang akan ditambang**
  - Karakterisitik, diskriminasi, asosiasi, klasifikasi, clustering, trend, outlier
- **Teknik**
  - Database, OLAP, machine learning, statistik, visualiasi
- **Penerapan**
  - Retail, telekomunikasi, banking, analisis kejahatan, bio-data mining, saham, text mining, web mining

Klasifikasi sistem Data Mining

- Fungsi
  - Deskriptif
  - Prediktif
- Sudut pandang:
  - Data : Jenis data yang akan ditambang
  - Pengetahuan view: Pengetahuan yang akan ditemukan
  - Teknik: Teknik yang akan digunakan
  - Aplikasi

Data Mining: Data apa saja?

- **Database Tradisional**
  - Relational database, data warehouse, transactional database
- **Advanced Database**
  - Data streams dan data sensor
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases dan legacy databases
  - Spatial data dan spatiotemporal data
  - Multimedia database
  - Text databases
  - World-Wide Web

Model dalam Data Mining

- **<u>Verification Model</u>**
  - Model ini menggunakan (hypothesis) dari pengguna, dan melakukan test terhadap perkiraan yang diambil sebelumnya dengan menggunakan data-data yang ada.
  - Model *verifikasi* menggunakan pendekatan *top down* dengan mengambil hipotesa dari user dan memeriksa validitasnya dengan data sehingga bisa dibuktikan kebenaran hipotesa tersebut.

Model dalam Data Mining

- **Discovery Model**
  - Sistem secara langsung menemukan informasi-informasi penting yang tersembunyi dalam suatu data yang besar. Data-data yang ada kemudian dipilah-pilah untuk menemukan suatu pola, trend yang ada, dan keadaan umum pada saat itu tanpa adanya campur tangan dan tuntutan dari pengguna.
  - Model *knowledge discovery* menggunakan pendekatan *bottom up* untuk mendapatkan informasi yang sebelumnya tidak diketahui. Model ini terbagi menjadi dua *directed knowledge discovery* dan *undirected knowledge discovery*.

Model dalam Data Mining

- **Discovery Model**
  - Pada *directed knowledge discovery*, data mining akan mencoba mencari penjelasan nilai target field tertentu (seperti pengahasilan, respons, usia, dan lain-lain) terhadap field-field yang lain.
  - Pada *undirected knowledge discovery* tidak ada target field karena komputer akan mecari pola yang ada pada data. Jadi *undirected knowledge discovery* digunakan untuk mengenali hubungan/relasi yang ada pada data sedangkan directed discovery akan menjelaskan hubungan/relasi tersebut.

Fungsi dalam Data Mining

- Fungsi atau sub kegiatan yang ada dalam data mining dalam rangka menemukan, menggali, atau menambang pengetahuan, mengacu pada Larose (2005), terdapat enam fungsi dalam data mining, yaitu:
  - Fungsi deskripsi (description)
  - Fungsi estimasi (estimation)
  - Fungsi prediksi (prediction)
  - Fungsi klasifikasi (classification)
  - Fungsi pengelompokan (classification),
  - Fungsi asosiasi (association).

Fungsi dalam Data Mining

- Mengacu pada Berry dan Browne (2006), keenam fungsi data mining tersebut dapat dipilah menjadi:
  - Fungsi minor atau fungsi tambahan, yang meliputi ketiga fungsi pertama, yaitu *deskripsi*, estimasi, dan prediksi
  - Fungsi mayor atau fungsi utama, yang meliputi ketiga fungsi berikutnya, yaitu klasifikasi, pengelompokkan, dan asosiasi.

Data Mining: Data apa saja?

- **Database Tradisional**
  - Relational database, data warehouse, transactional database
- **Advanced Database**
  - Data streams dan data sensor
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases dan legacy databases
  - Spatial data dan spatiotemporal data
  - Multimedia database
  - Text databases
  - World-Wide Web

Aplikasi Data Mining

Pemasaran/ Penyewaan

- Identifikasi pola pembayaran pelanggan
- Menemukan asosiasi diantara karakteristik demografik pelanggan
- Analisis keranjang pemasaran

Perbankan

- Mendeteksi pola penyalahgunaan kartu kredit
- Identifikasi pelanggan yang loyal
- Mendeteksi kartu kredit yang dihabiskan oleh kelompok pelanggan

Asuransi & Pelayanan Kesehatan

- Analisis dari klaim
- Memprediksi pelanggan yang akan membeli polis baru
- Identifikasi pola perilaku pelanggan yang berbahaya

Aplikasi Data Mining

- Analisa Perusahaan dan Manajemen Resiko
  - Perencanaan Keuangan dan Evaluasi Aset
  - Perencanaan Sumber Daya (Resource Planning)
  - Persaingan (competition) → Competitive Intelligence
- Telekomunication
  - menerapkan data mining untuk melihat dari jutaan transaksi yang masuk, transaksi mana saja yang masih harus ditangani secara manual (dilayani oleh orang).

Permasalahan Pada DM

- Metodologi
  - Mining beragam pengetahuan dari beragam sumber data
  - Kinerja: efesiensi, efektivitas dan skalabilitas
  - Evaluasi pola
  - Background knowledge
  - Noise (gangguan) dan data yang tidak lengkap
  - Distributed dan paralel method.
  - knowledge fusion (penggabungan)

Permasalahan DM (lanj)

- Interaksi pengguna
  - Data mining query languages dan ad-hoc mining
  - Visualisasi
  - Interactive mining
- Aplikasi
  - Domain spesifiik
  - Perlindungan data

Top-10 Algorithm di ICDM’06

- **#1: C4.5 (61 votes)**
- **#2: K-Means (60 votes)**
- **#3: SVM (58 votes)**
- **#4: Apriori (52 votes)**
- **#5: EM (48 votes)**
- **#6: PageRank (46 votes)**
- **#7: AdaBoost (45 votes)**
- **#7: kNN (45 votes)**
- **#7: Naive Bayes (45 votes)**
- **#10: CART (34 votes)**

Seputar Perkuliahan

- Sistem Penilaian: 20 % tugas, 40% UTS, 40% UAS
- Referensi : Data Mining: Concepts and Techiques, Jiawei Han
  (bab 1 sd bab 8)

Materi Kuliah

Tentative main topics for Data Mining

- Introduction to Data Mining → 1
- Preprocessing → 2
- Association rule (apriori) → 3
- Classification (Decision Tree) → 4
- Clustering → 5-7
  - Introduction to clustering
  - Clustering algorithms
  - Cluster analysis
- Text search & Mining → 8
- Multimedia Data Mining → 9
- Visualization → 10