Computer Science Department
Bogor Agricultural University

Pengantar *Data Mining*

Kuliah 1

o 2/11/2017

Agenda

- Pendahuluan
- Pengertian *Data Mining*
- *Knowledge Discovery in Database* (KDD)
- Arsitektur Sistem *Data mining*
- Tugas-tugas dalam *Data mining*
- Aplikasi

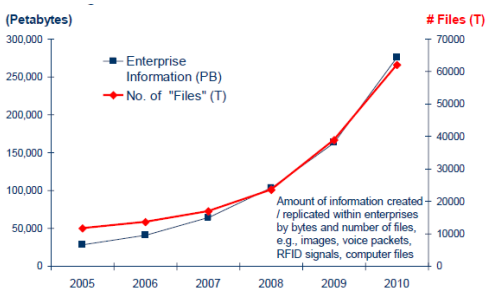
Motivasi

- Masalah eksplorasi data
  - Perkembangan dalam teknologi basis data dan tool terotomasi untuk pengumpulan data telah mengakibatkan menumpuknya data dalam basis data, *data warehouse* dan tempat penyimpanan data lainnya.

    We are drowning in data, but starving for knowledge!
- Solusi: *Data warehousing* dan *data mining*
  - *Data warehousing* dan *on-line analytical processing*
  - Ekstraksi pengetahuan yang menarik (aturan, pola, atau kendala) dari basis data berukuran besar.

Motivasi

- Informasi 'tersembunyi' dalam data
- Analisis secara manual membutuhkan waktu yang cukup lama untuk mencari informasi yang menarik
- Kebanyakan data tidak pernah dianalsisis setelah dikumpulkan



Data Growth Trends (Chute, Manfrediz, Minton, Reinsel, Schlichting, & Toncheva, 2008)

http://beyondplm.com/2012/07/26/plm-whatever-dude-just-get-access-to-data/

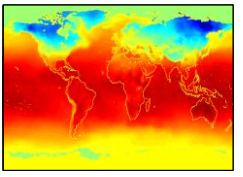
Mengapa Menambang Data?
Sudut pandang komersil

- Telah banyak data yang dikumpukan
  - Web data, e-commerce, e-banking
  - grocery stores
  - Bank/Credit Card transactions
- Teknologi komputer menjadi lebih murah dan powerful
- Tekanan kompetisi semakin kuat
  - Menyediakan layanan yang lebih baik, contoh *Customer Relationship Management*

Mengapa Menambang Data?
Sudut pandang keilmuan

- Data dikumpulkan dan disimpan dengan kecepatan tinggi (GB/hour)
  - Remote sensor pada satellite
  - Microarray pembangkit *gene expression data*
  - Data spasial dalam GIS
  - Simulasi keilmuan

- Teknik tradisional tidak cukup untuk menganalisis data demikian
- Data mining membantu ilmuwan dalam
  - Klasifikasi dan segmentasi data
  - Pemodelan
  - Clustering



Evolution of Sciences: New Data Science Era

- 1990-now: **Data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes
  - Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

Pertumbuhan Pengguna Internet

3,565,088,836
Internet Users in the world

1,148,071,994
Total number of Websites

118,310,419,562
Emails sent today

2,605,955,488
Google searches today

2,432,221
Blog posts written today

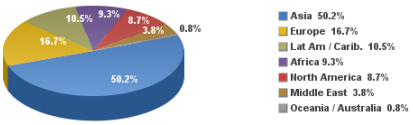
333,244,797
Tweets sent today

3,023,522,118
Videos viewed today on YouTube

1,827,812,839
Facebook active users

Source: http://www.int [5]

Internet Users in the World by Regions June 2016



Source: Internet World Stats - www.internetworldstats.com/stats.htm
Basis: 3,675,824,813 Internet users on June 30, 2016
Copyright © 2016, Miniwatts Marketing Group

Top Sites in Indonesia

	Site	Time on site	Pages per visit	% of traffic from search	Number of links in
1	Google.com Enables users to search the world's information, including webpages, images, and videos. Offers...More	9:05	8.93	2.40%	3,302,600
2	Google.co.id This guide will introduce you to all the different ways you can use Google Talk. ... The Google...More	6:45	5.99	1.70%	20,348
3	Youtube.com User-submitted videos with rating, comments, and contests.	9:45	5.60	8.90%	2,557,498
4	Tribunnews.com TRIBUNNEWS.COM :: Berita Terkini Indonesia Diterbitkan TRIBUN-Network "The National's Loca...More	7:34	4.30	24.90%	32,666
5	Detik.com detikcom is pioneer online media company in Indonesia, provides the most updated & comprehe...More	11:18	5.02	9.40%	81,282
6	Blogspot.co.id	4:26	3.14	46.80%	120

Pengertian Data Mining

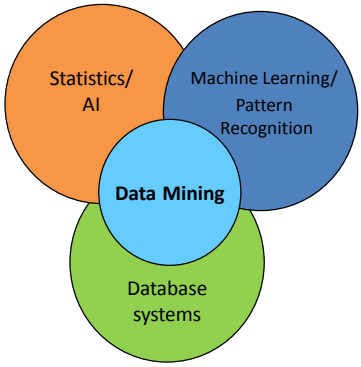
- *Data mining* (*knowledge discovery in databases*):
  - Ekstraksi informasi atau pola yang menarik (non-trivial, implicit, previously unknown dan potentially useful) dalam basis data berukuran besar.
- Istilah lain:
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, dll.
- Yang bukan termasuk *data mining task*?
  - Pemrosesan (deductive) query.
  - Sistem pakar
  - Sistem informasi

Pengertian Data Mining

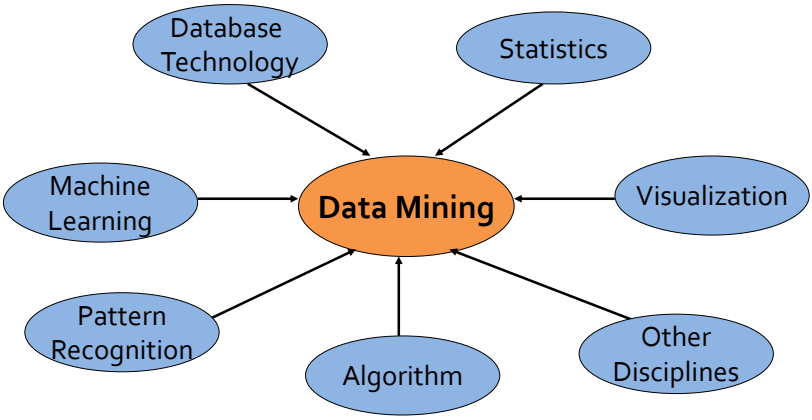
- **Yang bukan *data mining task*?**
  - ✓ Mencari nama mahasiswa S1 Ilmu Komputer pada SIMAK
  - ✓ Mencari nomor telpon dalam direktori telpon
  - ✓ Melakukan kueri pada *search engine* untuk mencari informasi tentang "Amazon"
- **Yang merupakan *data mining task*?**
  - ✓ Mencari nama-nama produk yang sering dibeli bersamaan dengan snack di supermarket
  - ✓ Mengelompokan dokumen-dokumen yang mirip yang dikembalikan oleh *search engine* berdasarkan konteksnya (misalkan Amazon rainforest, Amazon.com

Hubungan *data mining* dengan bidang lain

- Berkaitan erat dengan bidang *machine learning*/AI, *pattern recognition*, statistika, dan sistem basis data
- Teknik tradisonal menjadi tidak sesuai karena
  - Data berukuran besar
  - Tingginya dimensi data
  - Data yang heterogen, dan terdistribusi

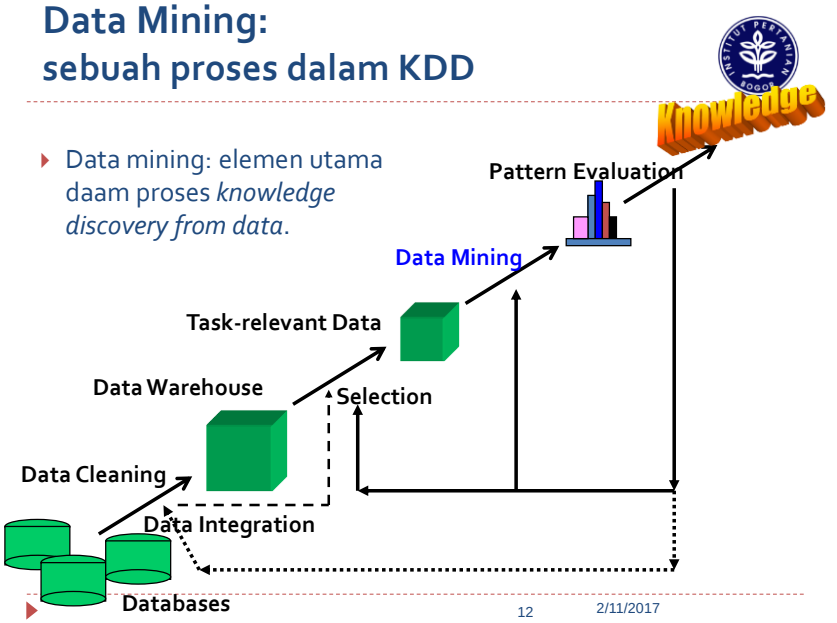


Data Mining: multi disiplin



Data Mining: sebuah proses dalam KDD

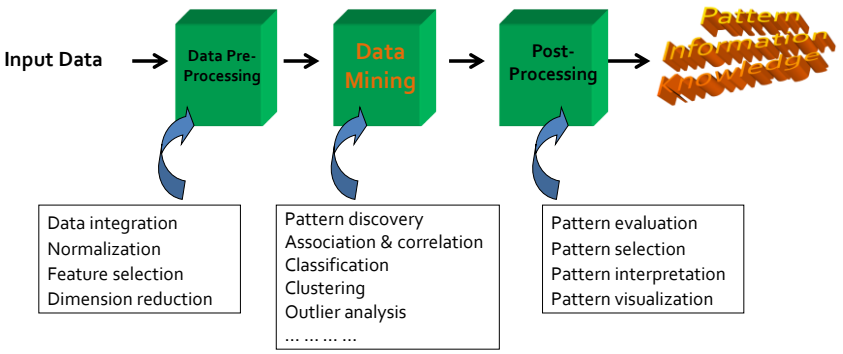
- Data mining: elemen utama daam proses *knowledge discovery from data*.



Data Mining: sebuah proses dalam KDD

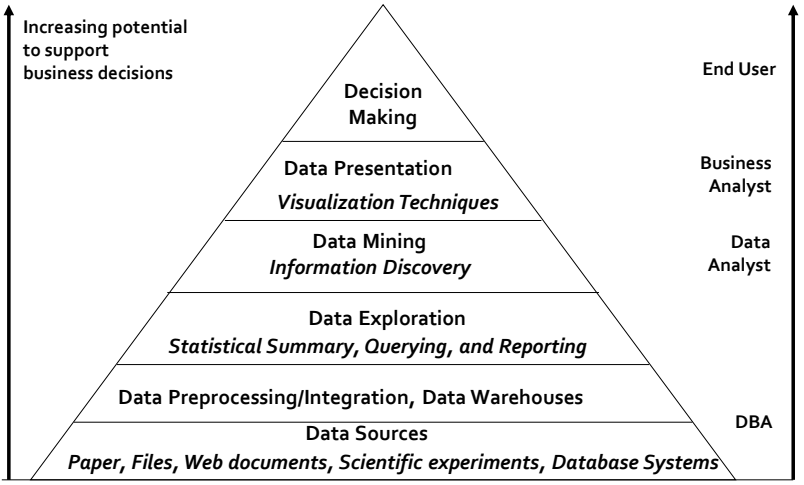
1. **Pembersihan data**: menghilangkan *noise* dan data yang tidak konsisten.
2. **Pengintegrasian data**: data digabungkan dari berbagai sumber.
3. **Seleksi data**: data yang relevan dengan proses analisis diambil dari basis data.
4. **Transformasi data**: data ditransformasikan atau digabungkan ke dalam bentuk yang sesuai untuk di-*mine* dengan cara dilakukan peringkasan atau operasi agregasi.
5. ***Data mining:*** merupakan proses yang penting dalam KDD dimana metode-metode cerdas diaplikasikan untuk mengekstrak pola-pola data.
6. **Evaluasi pola**: untuk mengidentifikasi pola-pola yang menarik yang merepresentasikan pengetahuan berdasarkan suatu ukuran kemenarikan.
7. **Presentasi pengetahuan**: merepresentasikan pengetahuan yang telah digali kepada pengguna.

KDD Process: A Typical View from ML and Statistics



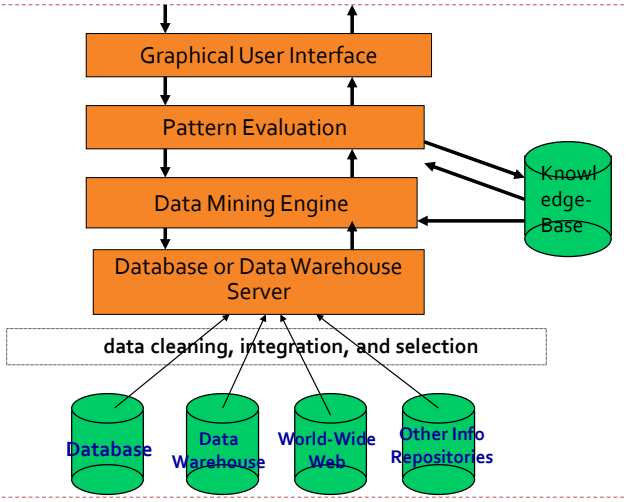
Han and Kamber, 2012

Data Mining dalam Business Intelligence



Han and Kamber, 2012

Arsitektur Sistem Data Mining



Arsitektur Sistem Data Mining

1. **Basis data**, *data warehouse* atau tempat penyimpanan informasi lainnya.
2. **Basis data dan *data warehouse server***. Komponen ini bertanggung jawab dalam pengambilan data yang relevan, berdasarkan permintaan pengguna.
3. **Basis pengetahuan**, merupakan *domain knowledge* yang digunakan untuk memandu pencarian atau mengevaluasi pola-pola yang dihasilkan.
4. ***Data mining engine***, terdiri modul-modul fungsional *data mining* seperti karakterisasi, asosiasi, klasifikasi, dan analisis *cluster*.
5. **Modul evaluasi pola**, menggunakan ukuran-ukuran kemenarikan dan berinteraksi dengan modul *data mining* dalam pencarian pola-pola menarik.
6. **Antarmuka pengguna grafis**, media komunikasi dengan pengguna dan sistem *data mining*.

Data yang ditambang?

- Basis data relasional
- Data warehouse
- Basis data transaksional
- Tempat penyimpanan data lainnya
  - Basis data *object-oriented* dan basis data *object-relational*
  - Basis data spatial
  - Data *time-series* dan data temporal
  - Sequence data (incl. bio-sequences)
  - Data teks dan basis data multimedia
  - WWW

Tugas-tugas dalam Data Mining

- Metode Prediksi
  - Menggunakan beberapa variabel (atribut) untuk memprediksi nilai yang tidak diketahui atau nilai yang akan datang dari variabel (atribut) lain.
- Metode Deskripsi
  - Menemukan pola-pola (korelasi, *trend*, *cluster*, trayektori, dan anomali) yang meringkas hubungan dalam data.

Data Mining Tasks (1)

- Association (*correlation* dan *causality*)
  - Multi-dimensional vs. single-dimensional association
  - Snack → Soft Drink [0.5%, 80%] (support, confidence)
  - contains(T, "computer") → contains(x, "software") [1%, 75%]
- Klasifikasi dan Prediksi
  - Menemukan model (fungsi) yang menjelaskan dan membedakan kelas atau kosep untuk prediksi mendatang
  - Contoh: mengklasifikasikan negara berdasarkan iklim
  - Presentasi: *decision-tree*, *classification rule*, *neural network*
  - Prediksi: mempredikasi nilai numerik yang tidak diketahui atau yang hilang

Data Mining Tasks (2)

- Analisis *cluster*
  - Label kelas tidak diketahui: mengelompokkan data untuk membentuk kelas-kelas yang baru, contoh: mengelompokkan data untuk mencari pola distribusinya
  - *Clustering* berdasarkan prinsip : memaksimumkan kemiripan intra-kelas dan memiminumkan kemiripan interkelas
- Analisis *outlier*
  - Outlier: objek data yang tidak mengikuti perilaku umum dari data
  - Dapat dipandang sebagai *noise* atau eksepsi tetapi berguna dalam *fraud detection*, *rare events analysis*
- Trend dan analisis evolusi
  - Trend dan deviasi: analisis regresi
  - *Sequential pattern mining*, contoh: PC → Anti Virus

Pola yang menarik?

- Data mining dapat membangkitkan ribuan pola : tidak semua pola tersebut menarik
  - Pendekatan: *Human-centered*, *query-based*, *focused mining*
- **Ukuran kemenarikan**
  - Sebuah pola dikatakan menarik jika pola tersebut mudah dimengerti oleh pengguna, valid pada data baru atau data tes dengan derajat kepastian (certainty), berguna, novel, atau memvalidasi hipotesis yang dicari oleh pengguna.

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Aplikasi 1: Analisis dan Manajemen Pasar

- Sumber data?—transaksi kartu kredit, data pengguna kartu yang setia (loyal), kupon *discount*, panggilan keluhan pengguna, studi gaya hidup publik
- *Target marketing*
  - Mencari cluster dari konsumen "model" yang memiliki karakteristik yang sama: minat, level pendapatan, perilaku belanja, dll,
  - Menentukan pola pembelian konsumen pada setiap waktu
- *Cross-market analysis*—Menemukan asosiasi/korelasi antar penjualan produk dan melakukan prediksi berdasarkan asosiasi tersebut
- *Customer profiling*—Tipe konsumen seperti apa yang akan membeli produk tertentu (*clustering* atau klasifikasi)
- *Customer requirement analysis*
  - Mengidentifikasi produk terbaik untuk konsumen-konsumen yang berbeda
  - Memprediksi faktor-faktor yang akan menarik perhatian konsumen baru
- *Summary information*
  - Laporan ringkasan multidimensional
  - Informasi ringkasan statistikal (*data central tendency* dan *variation*)

Aplikasi 1: Klasifikasi

- *Direct Marketing*
  - Tujuan: mengurangi biaya pengiriman surat dengan *mentargetkan* sekelompok konsumen yang mungkin akan membeli produk baru.
- *Fraud Detection*
  - Tujuan: memprediksi kasus-kasus yang mengandung kecurangan dalam transaksi dalam transaksi kartu kredit.

Aplikasi 2: Klasifikasi

- Perpindahan konsumen ke kompetitor (*Customer Attrition /Churn*):
  - Tujuan: memprediksi apakah seorang konsumen akan pindah ke produk kompetitor.
  - Pendekatan:
    - Menggunakan catatan transaksi secara rinci untuk setiap konsumen lampau dan saat ini, untuk menentukan atribut.
      - Seberapa sering konsumen melakukan panggilan, dimana dia melakukan panggilan, kapan konsumen tersebut sering melakukan panggilan, status keuangannya, status pernikahan, dll.
    - Menentukan label konsumen sebagai loyal atau tidak loyal.
    - Tentukan model untuk *loyalty*.

From [Berry & Linoff] Data Mining Techniques, 1997

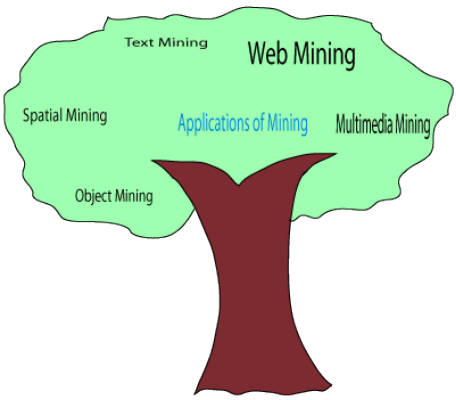
Aplikasi : *Clustering*

- Segmentasi Pasar:
  - Tujuan: membagi pasar ke dalam bagian-bagian konsumen yang berbeda dimana sebuah bagian dapat dipilih sebagai target pasar.
  - Pendekatan:
    - Mengumpulkan atribut-atribut yang berbeda dari konsumen berdasarkan informasi yang terkait gerografisnya dan gaya hidupnya.
    - Menemukan *cluster* dari konsumen-konsumen yang serupa.
    - Mengukur kualitas *clustering* dengan mengobservasi pola pembelian dari konsumen-konsumen dalam kelas yang sama terhadap konsumen-konsumen dari *cluster* yang berbeda.

Aplikasi: Penentuan Aturan Asosiasi

- Manajemen penempatan barang di supermarket.
  - Tujuan: mengidentifikasi item-item yang dibeli secara bersamaan oleh banyak pembeli.
  - Pendekatan: Memproses data *point-of-sale* yang dikumpulkan dengan *barcode scanner* untuk menemukan kebergantungan antar item.
  - Contoh aturan:
    - *{Cola, … } --> {Potato Chips}*

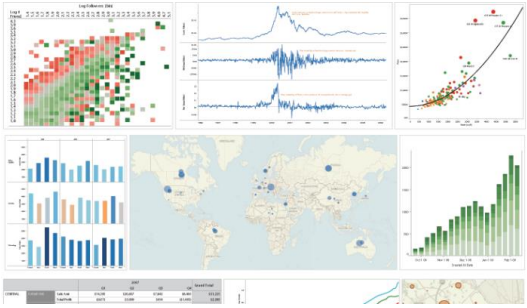
Practial Applications of Data Mining



- http://www.dataminingtools.net/wiki/applications_of_data_mining.php

Rujukan

- Tan P., Michael S., & Vipin K. 2006. *Introduction to Data mining*. Pearson Education, Inc.
- Han J, Kamber M, Pei J. 2011. *Data Mining: Concepts and Techniques*. Ed ke-3. US: Morgan Kaufmann



Topik selanjutnya:
Data dan Eksplorasi Data

https://www.linkedin.com/pulse/data-exploration-discovery-analytics-same-olap-ricky-barron