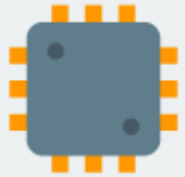
IFK15032– Data Mining



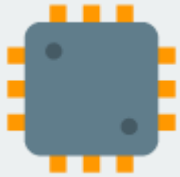
Pengantar Kuliah Data Mining

Rizal Setya Perdana, S.Kom., M.Kom.
Email : rizalespe@ub.ac.id

COMPUTATIONAL AND INTELEGENT SYSTEM LABORATORY
Universitas Brawijaya

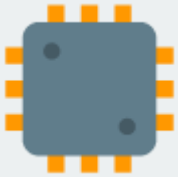
Penjelasan Mata Kuliah

1. Nama Mata Kuliah : Data Mining
2. Kode / SKS : *IFK15032* / 3
3. Semester : Ganjil
4. Prasyarat : -
5. Status mata kuliah : Pilihan



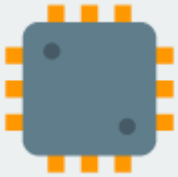
Tujuan (utama) Pembelajaran

- Mahasiswa mampu **memahami konsep, proses, metode dan teknik dasar** data mining
- Mahasiswa mampu **mendeskripsikan dan mendemonstrasikan** konsep, proses, metode dan teknik dasar data mining



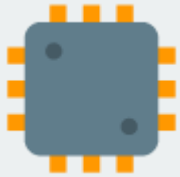
Tujuan Pembelajaran (lanjutan)

- Memahami konsep *data warehousing* dan penyiapan data (**data preprosesing**)
- Memahami konsep pengelompokkan data (**clustering**)
- Memahami konsep prediksi, rekomendasi (**klasifikasi**)
- Algoritma cara menangani **mising value**
- Memahami konsep **Association Rule** (mengetahui keterkaitan antar data)
- Memahami konsep **Squential Pattern** (mengetahui keterkaitan antar data dengan memperhatikan timeline)



Materi Mata Kuliah

1. Pengantar kuliah data mining
2. Data Warehousing dan persiapan data
3. Klastering : Hierarchical Method & K-Means
4. Klasifikasi : Instance Based (KNN)
5. Bayes
6. Algoritma Missing Value
7. Klasifikasi : Rule Based
8. Association Rule Mining
9. Sequential Pattern Mining



Evaluasi & Penilaian

- Point Keaktifan
- **30%** Tugas
- **30%** UTS
- **40%** Projek/UAS

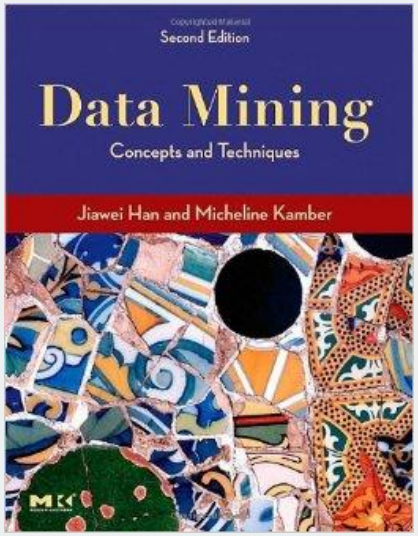
- Pembentukan Kelompok (5 org/per kelompok) + **rencana tugas masing-masing anggota**

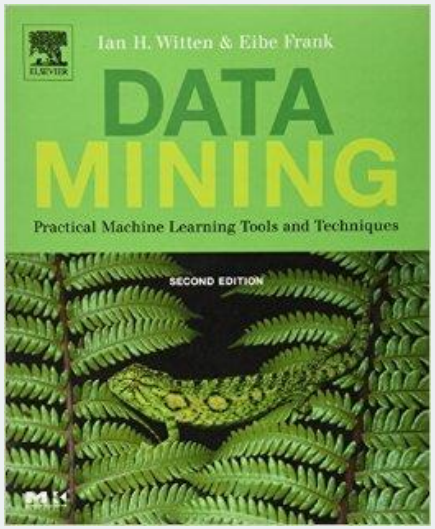
Aturan di Kelas

- Kuliah hadir **tepat waktu**
- Kehadiran minimal **80%**
- **Tidak menggunakan** HP/Smartphone/Laptop sebelum ada instruksi
- **Jujur** dan bersungguh-sungguh dalam mengerjakan setiap Tugas, Kuis maupun Ujian
- Menghindari **plagiasi** dalam setiap tugas
- Pelanggaran plagiasi akan memperoleh sanksi: pekerjaan tidak diakui (**nilai 0**)

Pustaka

- Data Mining: Practical Machine Learning Tools and Techniques (Carlos I. Degregori and I. H. Witten) Elsevier
- Data Mining: Concepts and Techniques (Jiawei Han)
- Gunakan **SCIENCEDIRECT** dan **IEEE EXPLORE** di UB
- **Sumber apapun (Internet, Video dll)**



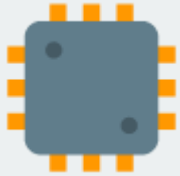


Rencana Perkuliahan

Pertemuan	Materi Kelas	Tugas Projek
1	Pengantar Kuliah	Pembentukkan kelompok tugas
2	Data Warehousing dan persiapan data	Cari Paper Tentang Data Mining (science direct & ieee)
3	Clustering: Hierarchical Method dan K-Means	Review Paper (Problem base, Metode dan Hasil)
4	Klasifikasi: Instance Base (KNN)	Progress paper (untuk persetujuan paper final projek)
5	Bayes	Fix Paper, Topik Final Project & Acc
6	Missing value	Perhitungan Manual
7	Quiz	
8	**UTS**	
9	**UTS**	Presentasi Bedah Paper Base Topik FP & Dataset, Desain Interface
10	Klasifikasi: Rule Based	
11	Association Rule Mining	Progres Koding Preprocessing/Ekstraksi Fitur & Uraian Algoritma
12	Association Rule Mining	Progres Koding Algoritma & Hasil Uji Coba
13	Sequential Pattern Mining	Progres Koding Algoritma & Hasil Uji Coba
14	Quiz	Dok. Langsung berupa Paper
15	Presentasi I	Dok. Langsung berupa Paper
16	Presentasi II (**UAS**)	Dok. Langsung berupa Paper

Sarana Pendukung

- Komputer
  - Laptop sendiri
- *Blog*:
  - http://rizalespe.lecture.ub.ac.id
- *Group Chat / Grup Sosial Media*
- Wakil kelas
  - Untuk komunikasi dengan dosen
  - Secepatnya terpilih dan beritahukan dosen



Data rich, Information Poor

Data ?

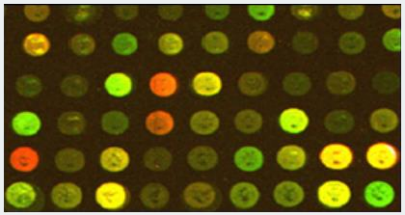
Informasi ?

Why Mine Data? Commercial Viewpoint

- Banyak data telah dikumpulkan dan digudangkan (warehoused)
  - Web data, e-commerce
  - Pembelian pada pusat belanja maupun grosir
  - Transaksi Bank/Credit Card

- Komputer semakin murah dan berkemampuan tinggi
- Tingkat persaingan (*Competitive Pressure*) yang makin kuat
  - Menyediakan layanan yang lebih baik dan sesuai dengan pelanggan

Why Mine Data? Scientific Viewpoint

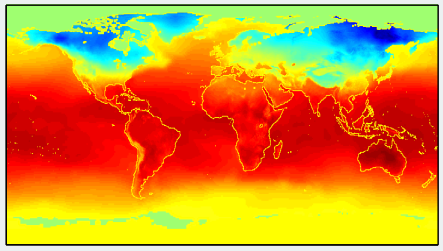


- Data terkumpul dan tersimpan

  pada kecepatan yang luar biasa (GB/hour)
  - Penginderaan jarak jauh pada satelit
  - Pemindaian telescopes angkasa
  - Larik mikro yang membangkitkan data genetik
  - Simulasi ilmiah yang membangkitkan

    data berukurang besar (terabytes of data)
- Cara kuno yang tidak layak untuk data-data mentah (*raw data*)
- Data mining dapat membantu ilmuwan
  - Dalam mengklasifikasikan dan mengelompokkan

    data dalam proses pembentukan hipotesis



Mining Large Data Sets - Motivation

- Kadang terdapat informasi yang "**tersembunyi**" dalam data yang tidak tersedia dengan jelas
- Seorang ahli analisa mungkin membutuhkan waktu berminggu-minggu untuk menemukan informasi yang bermanfaat dari sekumpulan data yang besar
- Kebanyakan data tidak pernah dianalisis secara keseluruhan

Apa Data Mining?

- Beragam definisi:
  - Penguraian (yang tidak sederhana) informasi potensi implicit (tidak nyata/jelas) yang sebelumnya tidak diketahui dari sekumpulan data
  - Penggalian dan analisis, dengan menggunakan peranti otomatis atau semi otomatis, dari sejumlah besar data yang bertujuan untuk menemukan bentuk yang bermanfaat

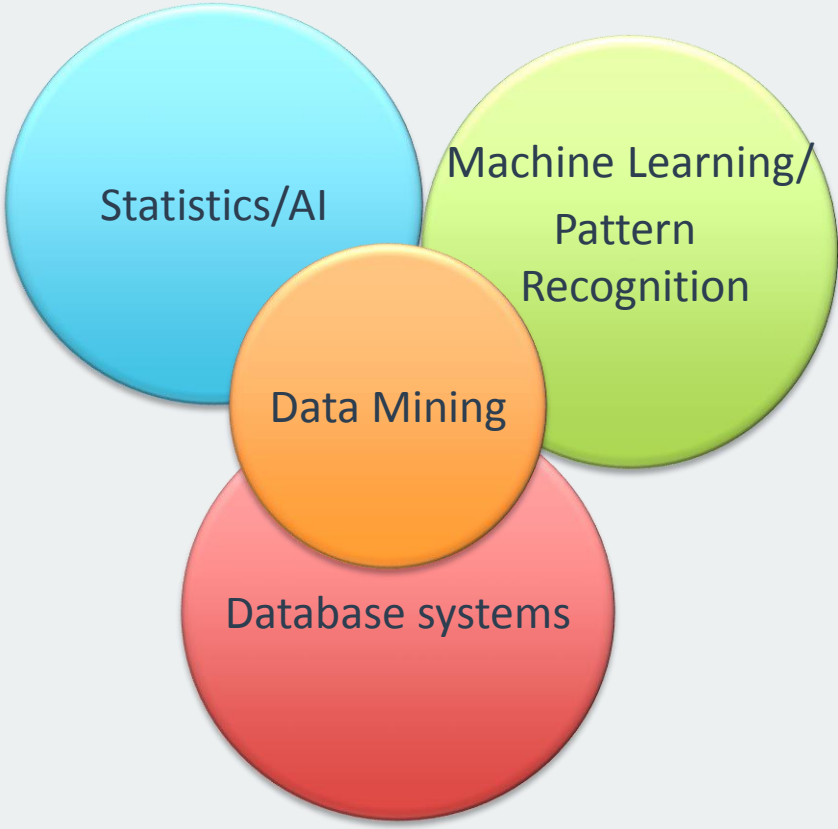
Apa yang (tidak) termasuk Data Mining?

- Apa yang tidak termasuk Data Mining?
  - Mencari nomer telepon pada buku telepon
  - Melalukan query pada suatu search engine untuk informasi tentang “Amazon”

- Apa yang termasuk Data Mining?
  - Nama tertentu lebih lazim dipakai di daerah Jawa (Sutinah, Suliyem, Ngatini, Paijo… di Jawa Tengah)
  - Mengelompokkan secara bersamaan dokumen-dokumen yang dihasilkan oleh search engine menurut hubungan kata-katanya (misal: Amazon rainforest, Amazon.com, etc)

Asal Data Mining

- Menggambarkan ide dari machine learning/AI, pattern recognition, statistics, dan database systems
- Cara tradisional yang sesuai untuk
  - Data yang amat besar
  - Data dengan banyak dimensi
  - Data yang heterogen
    dan tersebarHeterogeneous,



Tugas Data Mining

- Prediction Methods
  - Menggunakan beberapa variabel untuk memprediksi nilai yang tidak diketahui atau nilai di masa mendatang dari variabel lain.

- Description Methods
  - Menemukan bentuk yang mampu diartikan manusia (*human-interpretable patterns*) yang dapat menjelaskan data tertentu.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

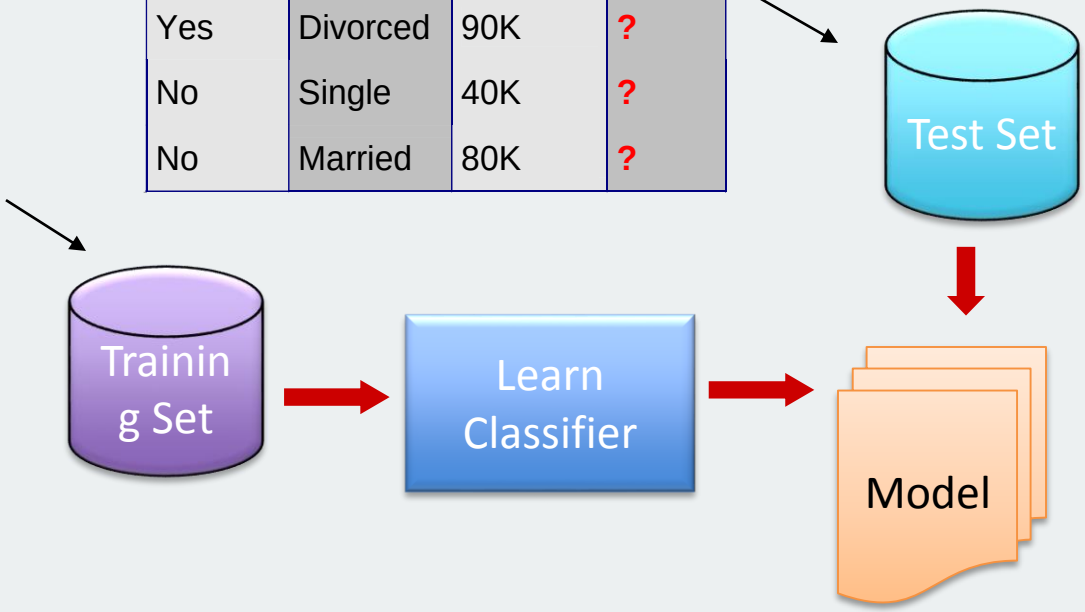
Classification: Definition

- Jika terdapat sekumpulan record (*training set*)
  - Setiap record terdiri dari sekumpulan *attributes*, satu dari atribut bisa merupakan *class*.
- Tentukan suatu *model* untuk atribut class sebagai suatu fungsi nilai dari atribut lain.
- Tujuan: previously unseen records should be assigned a class as accurately as possible.
  - Suatu *test set* digunakan untuk menentukan keakuratan suatu model. Umumnya, data set yang diberikan dibagi ke dalam *training sets* dan *test sets*, *training set* digunakan untuk membentuk model dan *test set* digunakan untuk mengujinya.

Classification Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
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

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application

- Penjualan Langsung (*Direct Marketing*)
  - Tujuan: Mengurangi biaya pengiriman promosi dengan hanya membidik (*targeting*) sejumlah konsumen yang suka membeli produk telepon selular baru.
  - Pendekatan:
    - Gunakan data untuk produk serupa yang telah ditawarkan sebelumnya.
    - Kita tahu konsumen mana yang memutuskan untuk membeli dan yang tidak. Keputusan ***{membeli, tidak membeli}*** membentuk atribut ***class***.
    - Kumpulkan berbagai informasi demografi, gaya hidup, dan semua informasi yang terkait dengan perusahaan (jenis usaha/pekerjaan, di mana mereka tinggal, berapa pendapatann mereka, dsb.) dari konsumen tersebut.
    - Gunakan infomasi ini sebagai atribut masukan pada *learn a classifier model*.

Diambil dari [Berry & Linoff] Data Mining Techniques, 1997

Definisi Clustering

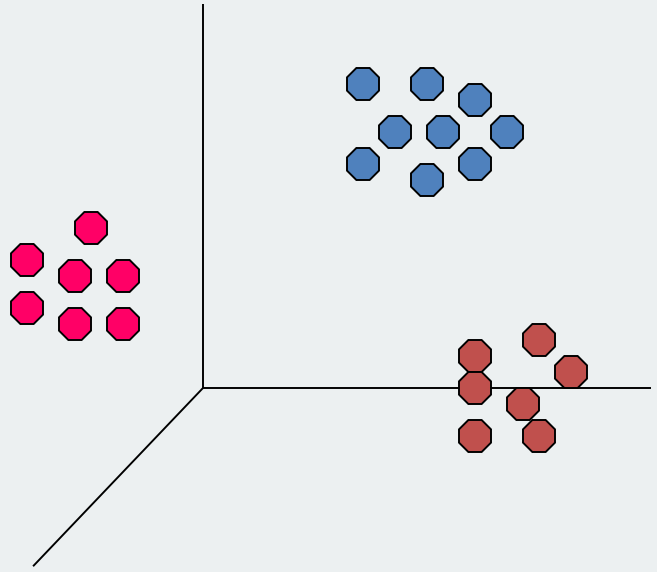
- Diberikan sekumpulan data titik, masing-masing memiliki sekumpulan atribut, dan kesamaan ukuran diantaranya, temukan gugus (*cluster*) sehingga
  - Data titik dalam satu *cluster* lebih serupa kepada yang lain.
  - Data titik dalam satu *cluster* yang berbeda lebih nampak ***kurang serupa*** kepada yang lain.
- Similarity Measures (Ukuran Kesamaan):
  - Euclidean Distance jika atributnya kontinyu.
  - Ukuran kesamaan lain yang khusus untuk problem khusus (Problem-specific Measures).

Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application

- Document Clustering:
  - Tujuan: Untuk menemukan kelompok dokumen yang serupa dengan yang lain berdasar istilah penting yang muncul di kedua dokumen yang dibandingkan.
  - Pendekatan: Mengenali frekuensi kemunculan istilah pada masing-masing dokumen. Membentuk ukuran kesamaan berdasar frekuensi dari istilah yang berbeda. Gunakan ukuran ini sebagai dasar pengelompokkan.
  - Pencapaian: Information Retrieval dapat menggunakan cluster untuk menghubungkan suatu dokumen baru atau mencari istilah pada dokumen yang telah dikelompokkan.

Illustrating Document Clustering

- Titik-titik pengelompokan: 3204 Article dari Kompas.
- Ukuran kesamaan: Seberapa banyak kata yang umum berada dalam dokumen-dokumen ini setelah dilakukan filter.

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Association Rule Discovery: Definition

- Diberikan sekumpulan record di mana masing-masing record terdiri dari sejumlah item dari koleksi yang diberikan;
  - Perlu dibuat dependency rules (aturan ketergantungan) yang akan memprediksikan kemunculan item tersebut berdasarkan kemunculan item yang lain.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

- Supermarket shelf management.
  - Tujuan: Menentukan item yang dibeli secara bersamaan dan cukup oleh banyak konsumen.
  - Approach: Proses data pembelian (point-of-sale) yang dikumpulkan dengan barcode scanners untuk menentukan ketergantungan antar item yang ada.

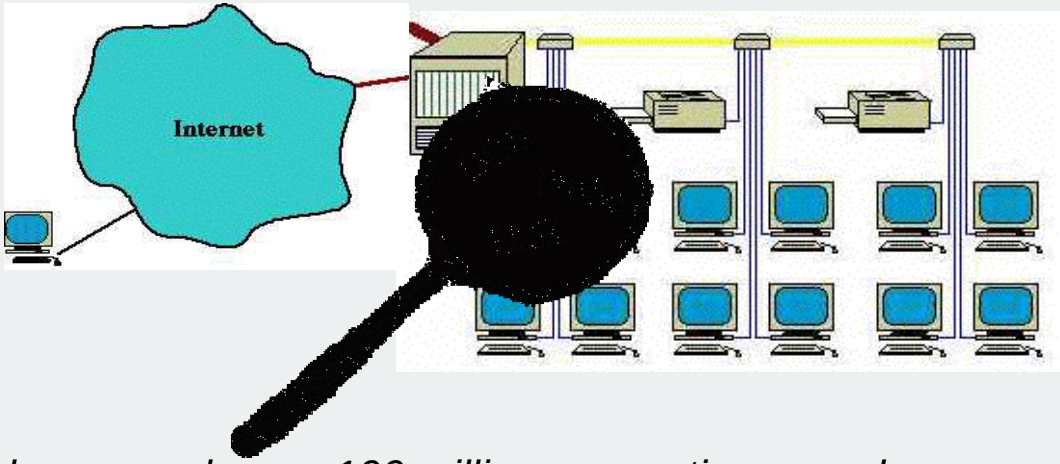
Association Rule Discovery: Application 3

- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection

  - Network Intrusion Detection



Typical network traffic at University level may reach over 100 million connections per day

Tantangan Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Tugas

- Cari informasi mengenai suatu **penerapan data mining** yang telah dikemas menjadi produk perangkat lunak. Uraikan penjelasan mengenai data mining task dari produk tersebut
- Cari informasi mengenai perangkat lunak yang merupakan tools untuk melakukan data mining. Uraikan penjelasan mengenai data mining task yang didukung oleh tools tersebut!

Terimakasih