Data Mining

Pertemuan 1: Pendahuluan

Tentang Dosen Pengampu

- Nama: Intan Yuniar Purbasari
- Kontak:
  - Email: intanyuniar.if@upnjatim.ac.id
  - WA: 083857716113
- Mata kuliah yang diampu:
  - Struktur Data
  - Kecerdasan Buatan
  - Pemrograman Berorientasi Objek
  - Pemrograman Dasar

Tentang Mata Kuliah

- MK Pilihan bidang minat CIS (mulai TA 2014/2015)
- 3 SKS
- MK Prasyarat: Kecerdasan Buatan
- Nilai minimum kelulusan: C-
- Slide materi & pengumuman terbaru → E-Learning UPN

Capaian Pembelajaran

- Pada akhir kuliah, mahasiswa mampu memahami konsep data mining dalam aplikasinya pada komputasi dan sistem cerdas serta diharapkan mampu merencanakan dan mendesain aplikasi pendukung data mining untuk keperluan suatu studi kasus tertentu.

Materi

- Pendahuluan (Knowledge Data Discovery, Definisi, Tinjauan dari sudut komersial dan ilmu pengetahuan, macam-macam model Data Mining)
- Data dan Preprocessing
- Tools dan aplikasi yang digunakan
- Klasifikasi
- Analisis asosiasi
- Analisis kluster
- Deteksi anomali/outlier

Penilaian

- NTS:
  - – 40% UTS
  - – 50% tugas
  - – 10% kehadiran

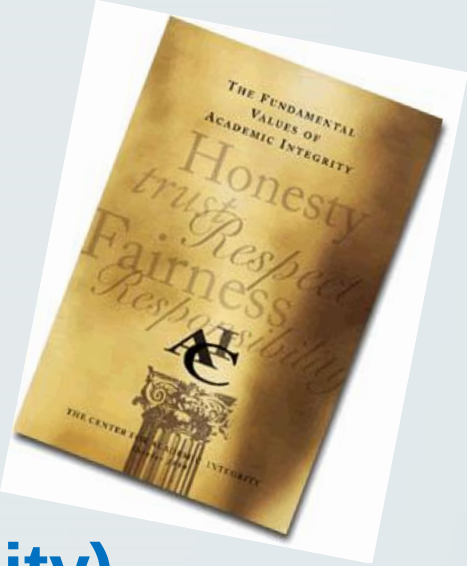
- NAS:
  - – 40% tugas
  - – 50% Final Project
  - – 10% kehadiran

Catatan: Untuk dapat lulus MK ini, semua komponen di atas tidak boleh kosong

Nilai Akhir = (NTS+NAS)/2

Nilai Integritas Akademik

- **Kejujuran (Honesty)**
- **Kepercayaan (Trust)**
- **Keadilan (Fairness)**
- **Penghormatan (Respect)**
- **Tanggung jawab (Responsibility)**


The Fundamental Values of Academic Integrity
Honesty
trust
Respect
Fairness
Responsibility
AI
The Center for Academic Integrity

TEGAKKAN!!

Pustaka

- Pang-Ning Tan, M. Steibach, V. Kumar, *Introduction to Data Mining,* Pearson Education, Inc., 2006
- Witten H. Ian and Frank Eibe, *Data Mining Practical Machine Learning Tools and Techniques 3rd Edition*, Morgan Kaufmann Publishers, 2011
- Chakrabarti et al, *Data Mining: Know It All*, Morgan Kaufmann Publishers, 2009
- Budi Santosa, *Data Mining: Teknik Pemanfaatan Data untuk Keperluan Bisnis*, Graha Ilmu, 2007
- Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*, Informatika, 2017

Latar Belakang Data Mining

- Problem:

BIG DATA

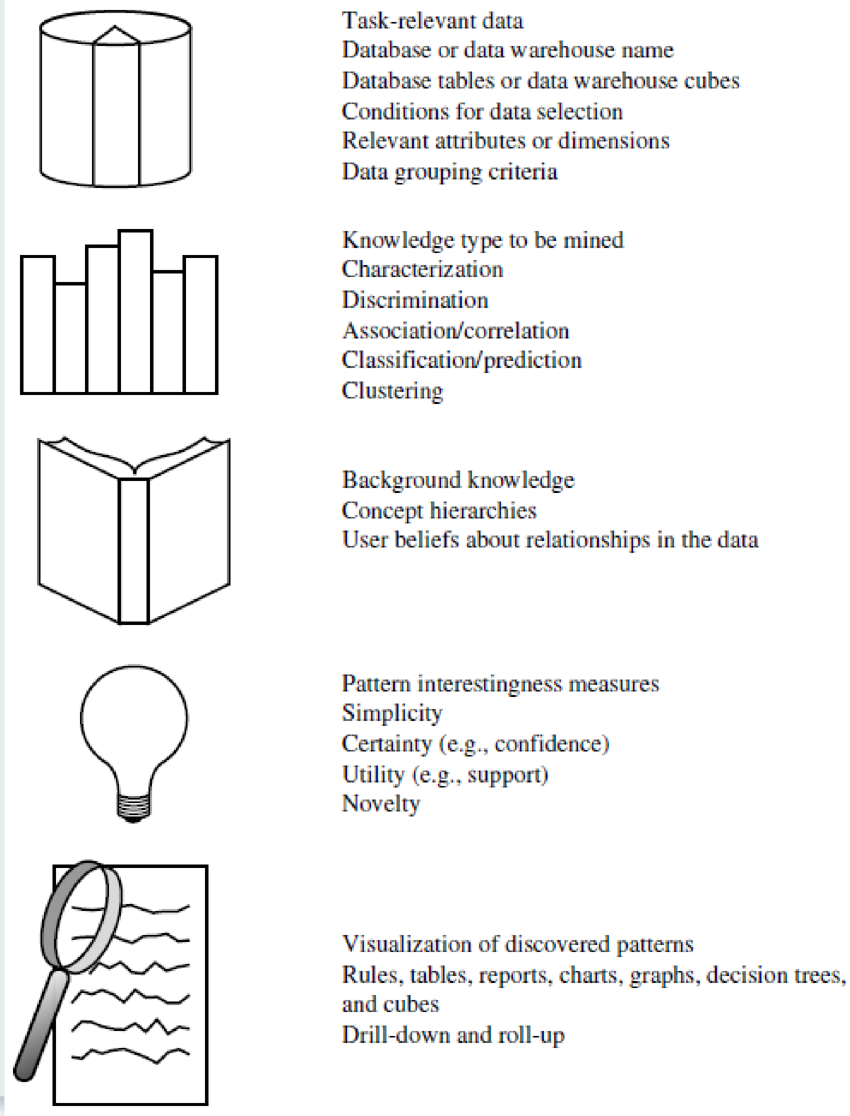
Tiga Karakteristik Big Data 3V

- **Volume (Isi):**
  - Ukuran data sangat besar, lebih besar dari petabytes
- **Velocity (kecepatan):**
  - Data besar harus dapat di-mining dalam periode waktu tertentu
- **Variety (Variasi):**
  - Sumber big data sangat bervariasi, mulai dari yang terstruktur hingga tidak terstruktur

Data Mining

- Sebuah metode/tool/software untuk mencari informasi, pola, trend dalam data yang berjumlah besar
- Video1
- Video2: Facebook Graph Engine-DM App
- Video3: DM App in Banking app
- Video4: Bioinformatics

Tugas Dasar Data Mining



Task-relevant data
Database or data warehouse name
Database tables or data warehouse cubes
Conditions for data selection
Relevant attributes or dimensions
Data grouping criteria

Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering

Background knowledge
Concept hierarchies
User beliefs about relationships in the data

Pattern interestingness measures
Simplicity
Certainty (e.g., confidence)
Utility (e.g., support)
Novelty

Visualization of discovered patterns
Rules, tables, reports, charts, graphs, decision trees, and cubes
Drill-down and roll-up

Model Data Mining

- Metode Prediksi *(Prediction Method):* Menggunakan variabel data untuk memprediksi variabel lain yang belum diketahui nilainya
- Metode Deskripsi *(Description Method):* fokus pada pencarian pola pada data sehingga dapat dipahami oleh manusia

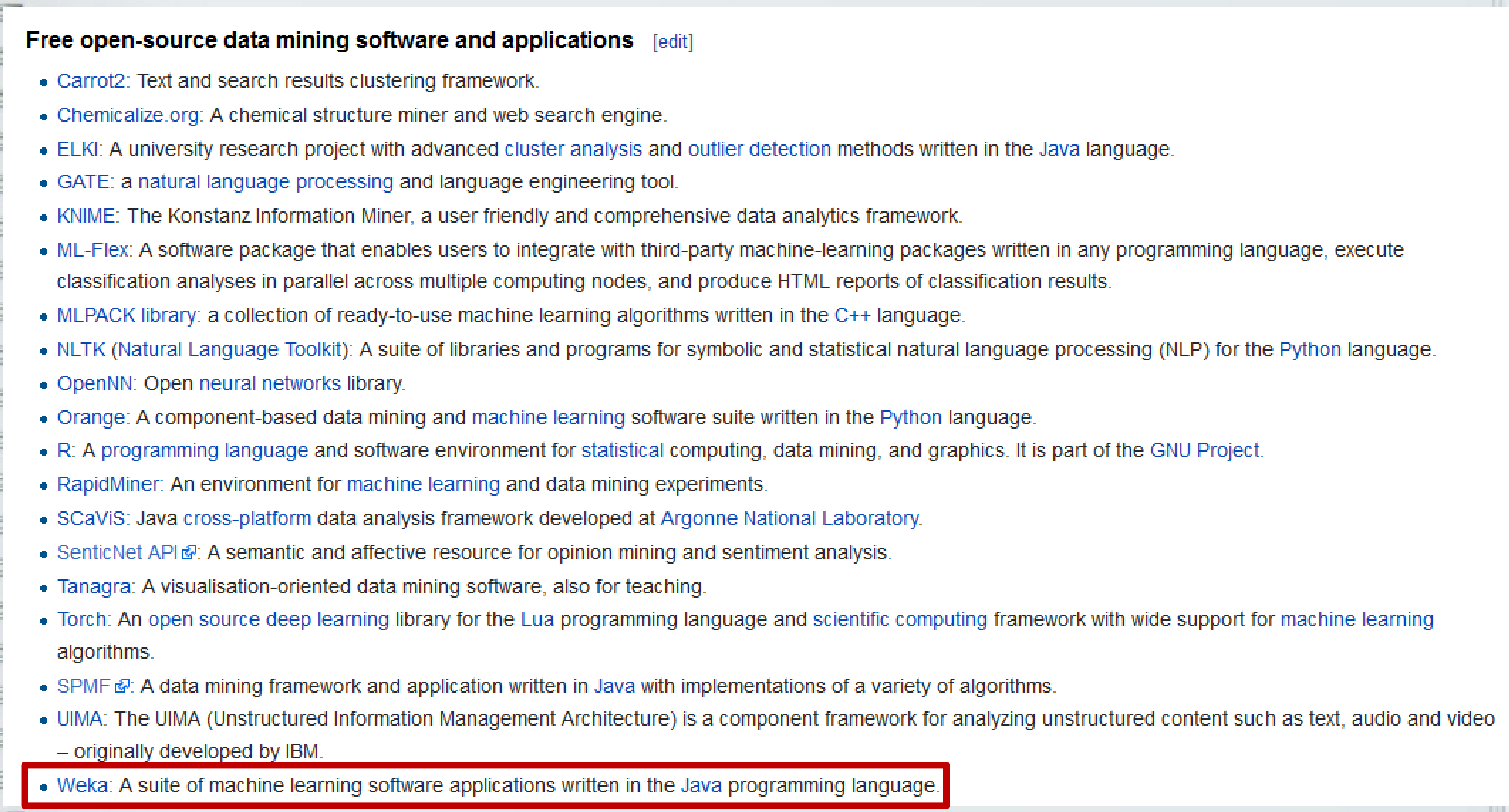
AI vs Machine Learning vs Data Mining

- **AI (Artificial Intelligence/Kecerdasan Buatan):**
  - Fokus bagaimana membuat sebuah sistem yang cerdas
- **Machine Learning (Pembelajaran Mesin):**
  - Bagian dari AI yang fokus kepada pemberian pengetahuan baru kepada sebuah sistem → membuat sistem yang dapat belajar
- **Data Mining:**
  - Fokus kepada penemuan informasi dalam data dengan menggunakan teknik dari machine learning

Aplikasi Data Mining

- Games → game yang menggunakan tabel sebagai penyimpanannya
- Business → analisis trend, CRM, market based analysis
- Science and Engineering → bioinformatika, monitor kondisi untuk peralatan listrik tegangan tinggi
- Medical Data Mining
- Spatial Data Mining → Geographic Information System (GIS)
- Text Mining → Pengelompokan dokumen, deteksi plagiarisme, Opinion & sentiment analysis
- Image Mining → Content-Based Image Retrieval (contoh: Google Image Search)
- Video Mining → Content-Based Video Retrieval
- Audio Mining → Content-Based Audio Retrieval (contoh: Shazam, SoundHound)
- Human rights (hak asasi manusia)….

Software

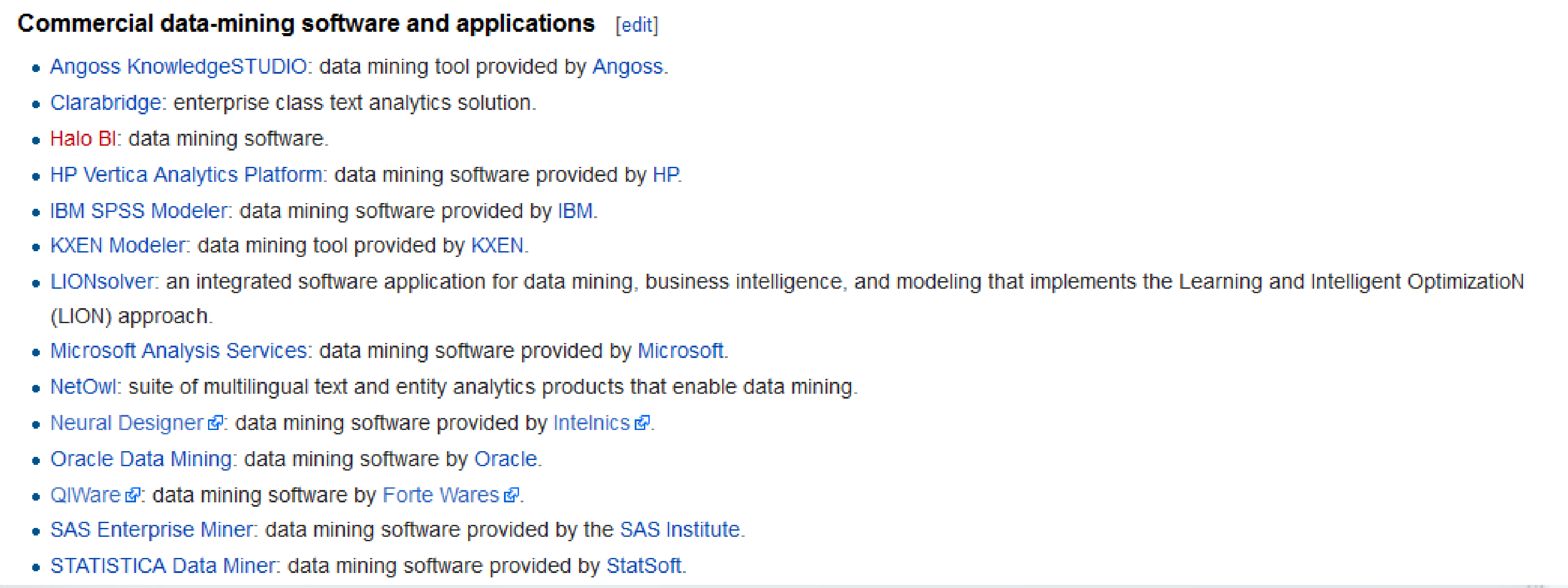


Free open-source data mining software and applications [edit]

- Carrot2: Text and search results clustering framework.
- Chemicalize.org: A chemical structure miner and web search engine.
- ELKI: A university research project with advanced cluster analysis and outlier detection methods written in the Java language.
- GATE: a natural language processing and language engineering tool.
- KNIME: The Konstanz Information Miner, a user friendly and comprehensive data analytics framework.
- ML-Flex: A software package that enables users to integrate with third-party machine-learning packages written in any programming language, execute classification analyses in parallel across multiple computing nodes, and produce HTML reports of classification results.
- MLPACK library: a collection of ready-to-use machine learning algorithms written in the C++ language.
- NLTK (Natural Language Toolkit): A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python language.
- OpenNN: Open neural networks library.
- Orange: A component-based data mining and machine learning software suite written in the Python language.
- R: A programming language and software environment for statistical computing, data mining, and graphics. It is part of the GNU Project.
- RapidMiner: An environment for machine learning and data mining experiments.
- SCaViS: Java cross-platform data analysis framework developed at Argonne National Laboratory.
- SenticNet API: A semantic and affective resource for opinion mining and sentiment analysis.
- Tanagra: A visualisation-oriented data mining software, also for teaching.
- Torch: An open source deep learning library for the Lua programming language and scientific computing framework with wide support for machine learning algorithms.
- SPMF: A data mining framework and application written in Java with implementations of a variety of algorithms.
- UIMA: The UIMA (Unstructured Information Management Architecture) is a component framework for analyzing unstructured content such as text, audio and video – originally developed by IBM.
- Weka: A suite of machine learning software applications written in the Java programming language.

Sumber: http://en.wikipedia.org/wiki/Data_mining

Software



Commercial data-mining software and applications [edit]

- Angoss KnowledgeSTUDIO: data mining tool provided by Angoss.
- Clarabridge: enterprise class text analytics solution.
- Halo BI: data mining software.
- HP Vertica Analytics Platform: data mining software provided by HP.
- IBM SPSS Modeler: data mining software provided by IBM.
- KXEN Modeler: data mining tool provided by KXEN.
- LIONsolver: an integrated software application for data mining, business intelligence, and modeling that implements the Learning and Intelligent OptimizatioN (LION) approach.
- Microsoft Analysis Services: data mining software provided by Microsoft.
- NetOwl: suite of multilingual text and entity analytics products that enable data mining.
- Neural Designer: data mining software provided by Intelnics.
- Oracle Data Mining: data mining software by Oracle.
- QIWare: data mining software by Forte Wares.
- SAS Enterprise Miner: data mining software provided by the SAS Institute.
- STATISTICA Data Miner: data mining software provided by StatSoft.

Sumber: http://en.wikipedia.org/wiki/Data_mining

Data Mining dari sudut pandang etika

- Pengumpulan data untuk proses mining dapat berhubungan dengan privasi, etika, dan legalitas
- Sangat direkomendasikan untuk menginformasikan kepada individu yang akan memberikan datanya tentang:
  - Tujuan pengumpulan data
  - Bagaimana data akan digunakan
  - Siapa saja yang dapat mengakses data
  - Status keamanan terhadap data
  - Bagaimana cara mengupdate data
- Data dapat dimodifikasi agar bersifat *anonim* agar pemilik data tidak dapat diidentifikasi

Data Mining dari sudut pandang hukum: **The Right To Be Forgotten (Hak untuk Dilupakan)**

- Sudah diterapkan di Uni Eropa (terutama Prancis) dan Argentina sejak beberapa tahun terakhir
- Memberikan hak kepada individu untuk menghapus datanya dari catatan manapun sehingga tidak dapat diakses lagi

Next Week

- Teknik-teknik Pre-processing data