

# Chapter 1. Introduction

---

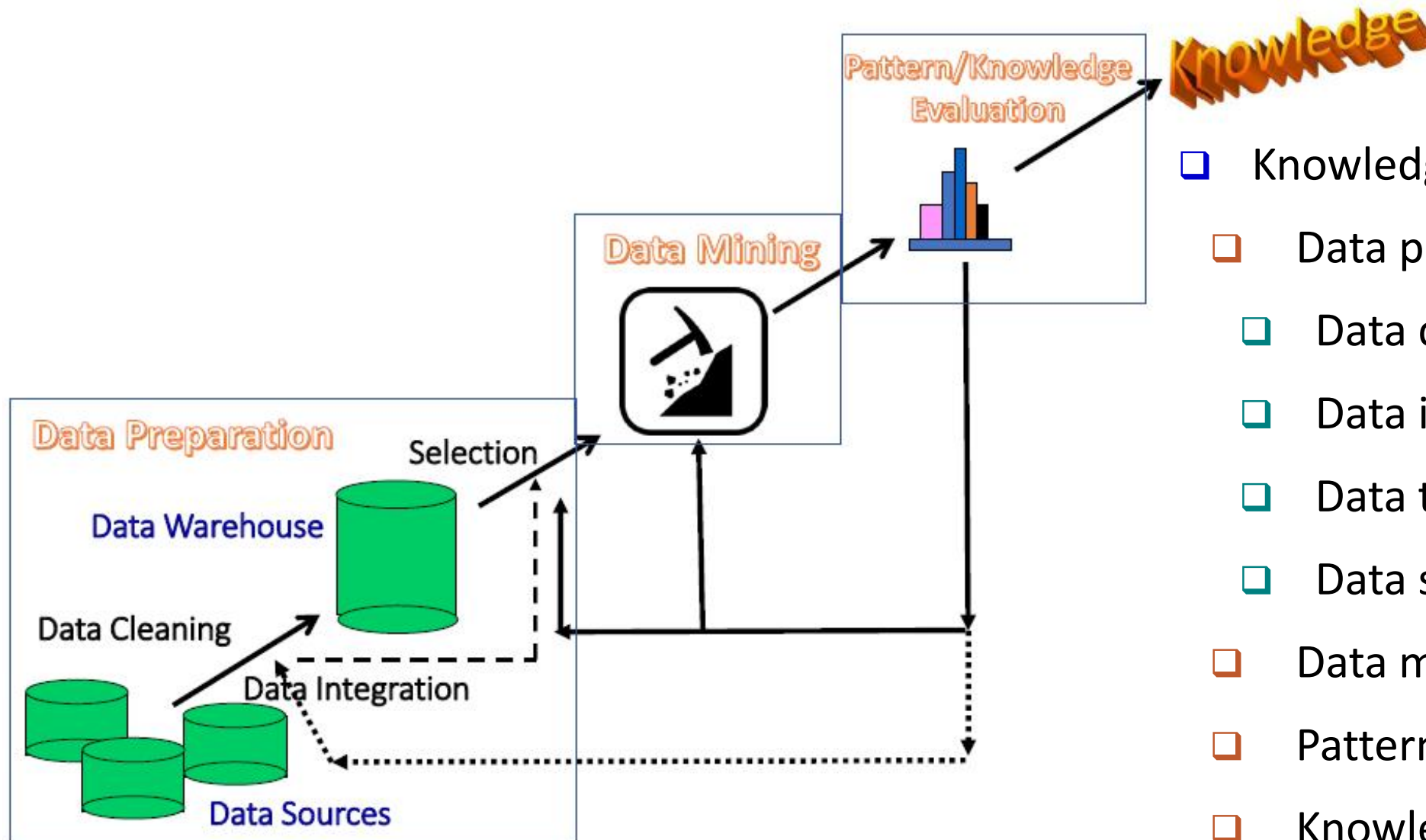
- ❑ **Apa itu Data Mining?**
- ❑ **Data Mining: Langkah Penting dalam Knowledge Discovery**
- ❑ **Keragaman Jenis Data untuk Data Mining**
- ❑ **Mining Berbagai macam Knowledge**
- ❑ **Data Mining: Pertemuan dari Multiple Disciplines**
- ❑ **Data Mining dan Aplikasi**
- ❑ **Data Mining dan Masyarakat**
- ❑ **Simpulan**

# Apa itu Data Mining?

---

- ❑ Kita hidup di dunia di mana sejumlah besar data dihasilkan secara konstan dan cepat
- ❑ **Data mining** adalah proses menemukan pola, model, dan jenis pengetahuan lain yang menarik dalam kumpulan data besar
  - ❑ “Data mining”: nama yang salah? Seharusnya “knowledge mining from data”
  - ❑ Istilah Lain: *Knowledge mining from data, KDD (Knowledge Discovery from Data), penemuan pola, ekstraksi pengetahuan, analisis data, pengumpulan informasi*
- ❑ Data mining adalah bidang yang masih muda, dinamis, dan menjanjikan
- ❑ Contoh: Data mining mengubah kumpulan data yang besar menjadi pengetahuan
  - ❑ Google Flu Tren (2008-2015) menemukan hubungan erat antara jumlah orang yang mencari info terkait flu. dan jumlah orang yang memiliki gejala flu
  - ❑ Sistem ini dapat memperkirakan aktivitas flu hingga dua minggu lebih cepat daripada sistem tradisional

# Data Mining: Langkah Penting dalam Knowledge Discovery



## □ Knowledge Discovery Process

□ Data preparation

□ Data cleaning

□ Data integration

□ Data transformation

□ Data selection

□ Data mining

□ Pattern/model evaluation

□ Knowledge presentation

# Keragaman Jenis Data untuk Data Mining (I)

---

## ❑ Data terstruktur vs. tidak terstruktur

- ❑ *Terstruktur: struktur seragam, seperti catatan atau tabel, didefinisikan oleh kamus data, dengan sekumpulan atribut tetap, masing-masing dengan serangkaian rentang nilai tetap dan makna semantic*
- ❑ Misalnya Data yang disimpan dalam database relasional, *data cubes*, *data matrices*, and many *data warehouses*
- ❑ *Semi-terstruktur: memungkinkan objek data berisi nilai yang ditetapkan, sekumpulan kecil nilai yang diketik heterogen, atau struktur nested, atau untuk memungkinkan struktur objek atau sub-objek didefinisikan secara fleksibel dan dinamis*
- ❑ Data yang memiliki struktur tertentu dengan makna semantik yang jelas, seperti kumpulan data transaksional, *sequence data set* (misalnya, data deret waktu, data gen atau protein, atau data Weblog)
- ❑ *Graph atau network data*: Jenis kumpulan data semi-terstruktur yang lebih canggih
- ❑ *Data tidak terstruktur*: Data teks dan data multimedia (misalnya, audio, gambar, video)
- ❑ Data dunia nyata seringkali merupakan campuran dari data terstruktur, semi-terstruktur dan data tidak terstruktur

# Keragaman Jenis Data untuk Data Mining (II)

---

## ❑ Data yang terkait dengan aplikasi yang berbeda

- ❑ Aplikasi yang berbeda: kumpulan data yang berbeda dan memerlukan metode analisis data yang berbeda
  - ❑ Sequence data: *Urutan biologis vs. urutan transaksi belanja*
  - ❑ Time-series: Kumpulan nilai numerik yang diurutkan dengan interval waktu yang sama
  - ❑ Data spasial, temporal, dan spatiotemporal
  - ❑ Graph and network data: Jejaring sosial, jaringan komunikasi komputer, jaringan biologis, dan jaringan informasi mungkin membawa semantik yang agak berbeda
- ❑ Pada kumpulan data yang sama, menemukan berbagai jenis pola: memerlukan metode penambangan yang berbeda
  - ❑ Misalnya program perangkat lunak: menemukan modul yang plagiat vs. menemukan bug copy-and-paste

## ❑ Stored vs. streaming data

- ❑ Stored data: Terbatas, disimpan dalam berbagai jenis repositori data besar
- ❑ Streaming data (Misalnya video surveillance atau remote sensing): Respons yang dinamis, terus datang, tak terbatas, dan real-time—menimbulkan tantangan pada penambangan data yang efektif

# Mining Berbagai Macam Knowledge

---

- ☐ **Peringkasan Data Multidimensi**
- ☐ **Pola, Asosiasi, dan Korelasi yang Sering Ditambang**
- ☐ **Klasifikasi dan Regresi untuk Analisis Prediktif**
- ☐ **Analisis Cluster**
- ☐ **Deep Learning**
- ☐ **Analisis Outlier**

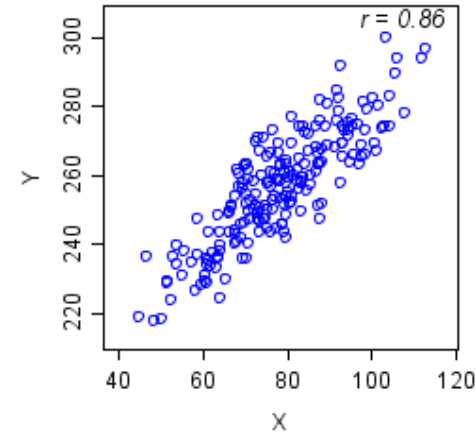
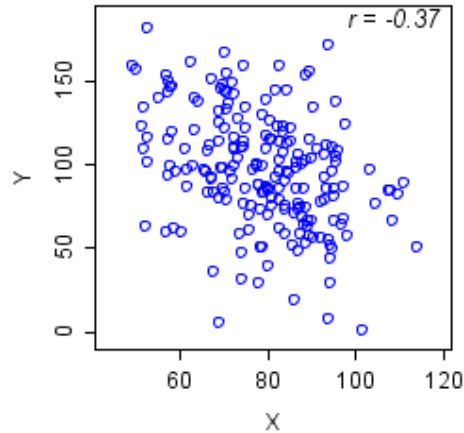
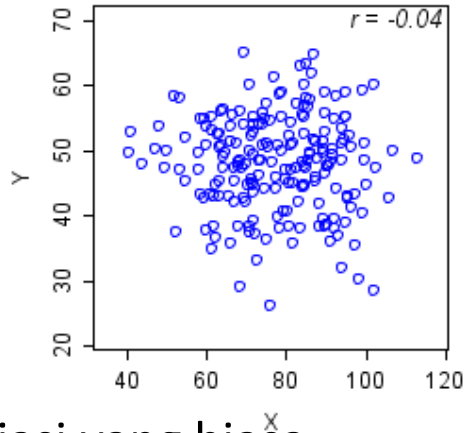
# Peringkasan Data Multidimensi

- ❑ Integrasi informasi dan kontruksi data warehouse
  - ❑ Data cleaning, transformation, integration, dan multidimensi data model
- ❑ Teknologi Data cube
  - ❑ Metode komputasi yang dapat diskalakan agregat multidimensi
  - ❑ OLAP (online analytical processing)
- ❑ Deskripsi konsep multidimensi: Karakterisasi dan diskriminasi
  - ❑ Menggeneralisasikan, meringkas, dan membedakan karakteristik data, misalnya, daerah kering vs. basah



# Penemuan Pola: Mining Frequent Patterns, Asosiasi, dan Korelasi

- Frequent patterns (atau frequent itemsets)
  - Item apa yang sering dibeli bersama di minimarket?
- Analisis Asosiasi dan Korelasi

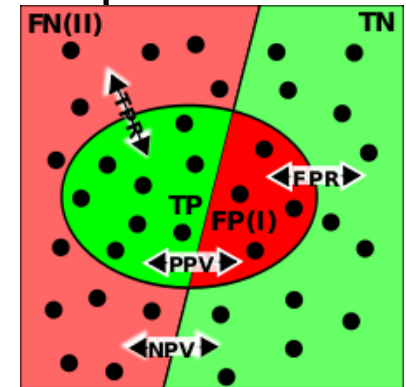


- Aturan asosiasi yang biasa
  - Popok → Bir [0.5%, 75%] (support, confidence)
  - Apakah item yang sangat terkait juga berkorelasi kuat?
- Bagaimana cara menambang pola dan aturan tersebut secara efisien dalam dataset besar?
- Bagaimana cara menggunakan pola seperti itu untuk klasifikasi, pengelompokan, dan aplikasi lainnya?



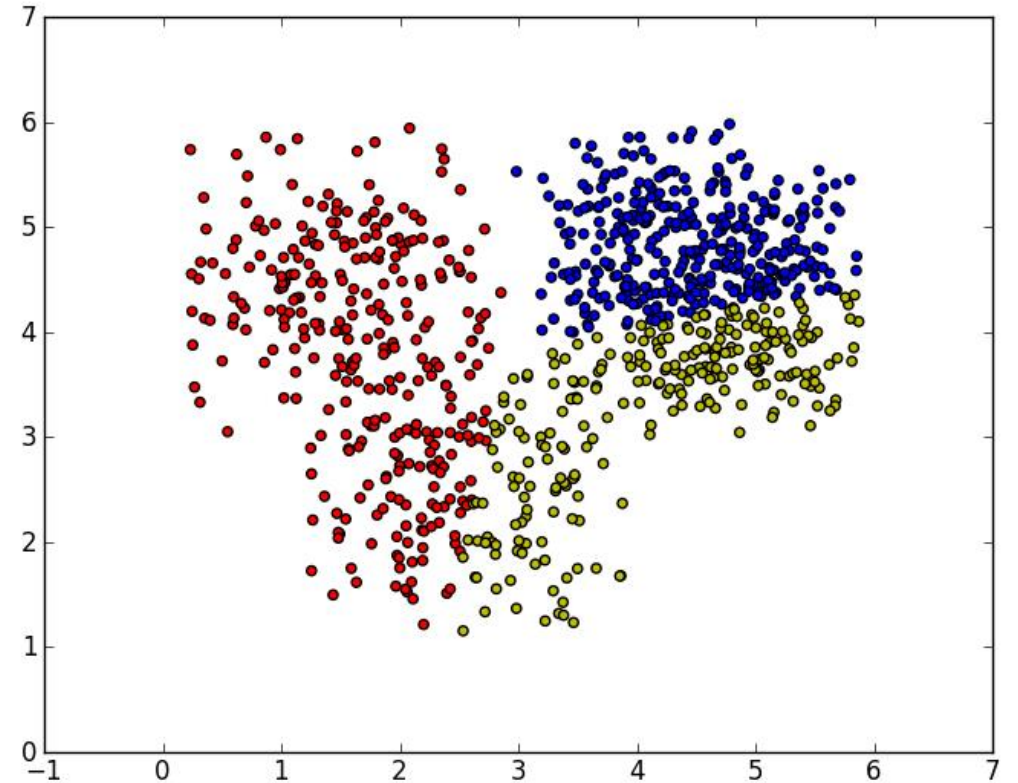
# Klasifikasi dan Regresi untuk Analisis Prediktif

- Klasifikasi dan prediksi label
  - Membangun model (fungsi) berdasarkan beberapa contoh pelatihan
  - Menjelaskan dan membedakan kelas atau konsep untuk prediksi masa depan
    - Kel. 1. Mengklasifikasikan negara berdasarkan (iklim)
    - Kel. 2. Mengklasifikasikan mobil berdasarkan (jarak tempuh bensin)
  - Memprediksi beberapa label kelas yang tidak diketahui
- Metode yang biasa digunakan
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- Pengaplikasian:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...



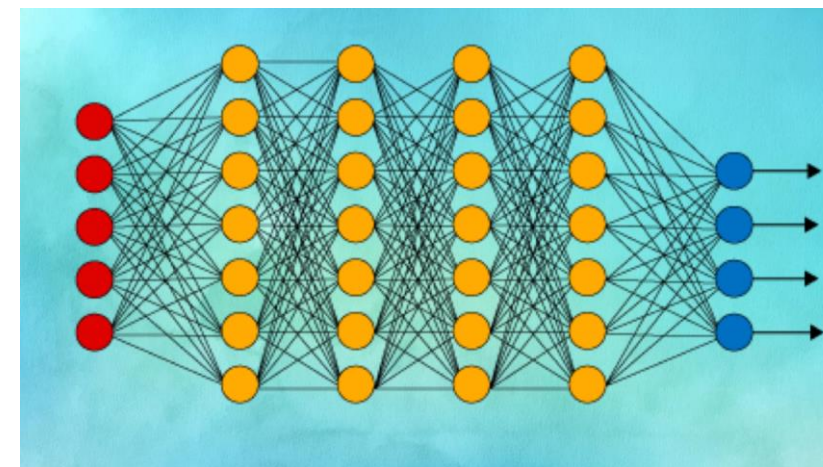
# Analisis Cluster

- ❑ Unsupervised learning (yaitu, Label Kelas tidak diketahui)
- ❑ Mengelompokkan data untuk membentuk kategori baru (yaitu, cluster), misalnya, rumah cluster untuk menemukan pola distribusi
- ❑ Prinsip: Memaksimalkan kesamaan intra-kelas & meminimalkan kesamaan antar kelas
- ❑ Banyak metode dan aplikasi



# Deep Learning

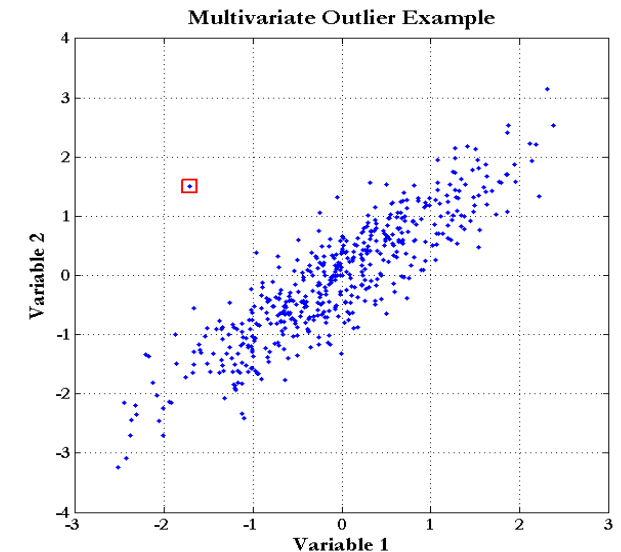
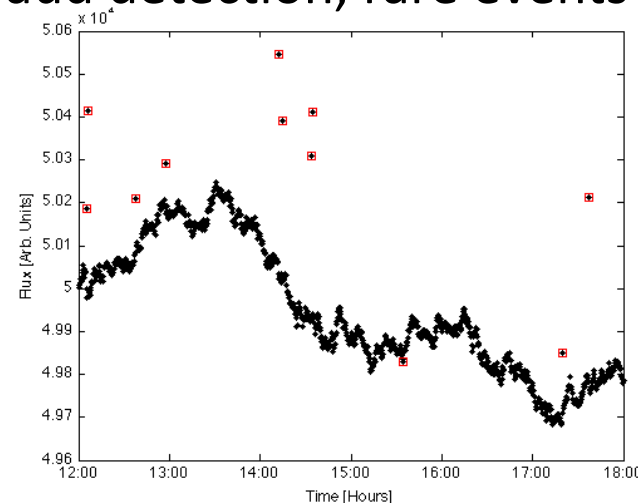
- ❑ Deep learning: Perbatasan dinamis yang berkembang pesat dalam pembelajaran mesin
- ❑ Deep learning telah mengembangkan berbagai arsitektur jaringan saraf
  - ❑ Feed-forward neural networks
  - ❑ Convolutional neural networks
  - ❑ Recurrent neural networks
  - ❑ Graph neural networks
  - ❑ Transformer
- ❑ Deep learning memiliki aplikasi yang luas di computer vision, natural language processing, machine translation, social network analysis, dan lain-lain
- ❑ Deep learning telah membentuk kembali berbagai tugas data mining
  - ❑ Misalnya classification, clustering, outlier detection, and reinforcement learning



# Outlier Analysis

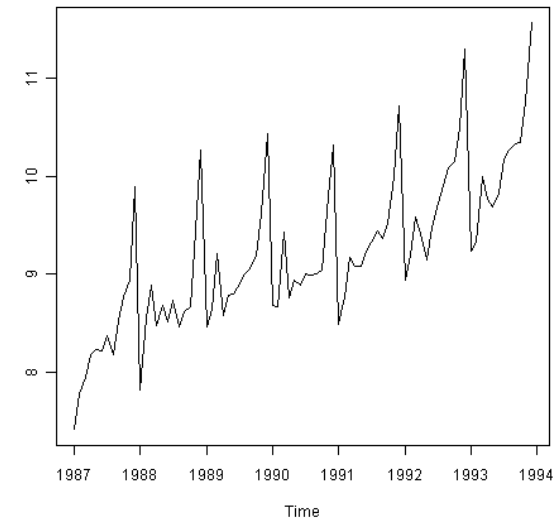
## Outlier Analysis

- Outlier: Objek data yang tidak sesuai dengan perilaku umum data
- Noise atau exception?—Sampah satu orang bisa menjadi harta karun bagi orang lain
- Methods: berdasarkan produk clustering atau regression analysis, ...
- Berguna dalam fraud detection, rare events analysis



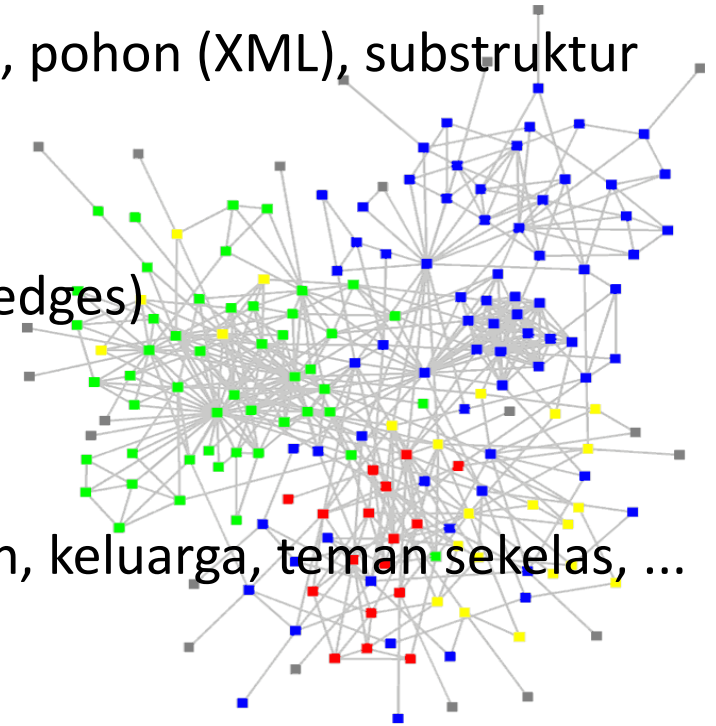
# Fungsi Data Mining Lainnya: Time dan Ordering: Sequential Pattern, Trend dan Evolution Analysis

- ❑ Sequence, trend dan evolution analysis
  - ❑ Trend, time-series, dan analisis deviasi
    - ❑ misalnya, regresi dan prediksi nilai
  - ❑ Sequential pattern mining
    - ❑ Misalnya membeli kamera digital, kemudian membeli kartu memori besar
  - ❑ Periodicity analysis
  - ❑ Motifs dan biological sequence analysis
    - ❑ Approximate dan consecutive motifs
  - ❑ Similarity-based analysis
- ❑ Mining data streams
  - ❑ Aliran data yang teratur, bervariasi terhadap waktu, dan berpotensi tidak terbatas



# Fungsi Data Mining Lainnya : Struktur dan Network Analysis

- Graph mining
  - Menemukan subgraf yang sering (misalnya, senyawa kimia), pohon (XML), substruktur (fragmen web)
- Information network analysis
  - Social networks: actors (objects, nodes) and relationships (edges)
    - misalnya, jaringan penulis di CS, jaringan teroris
  - Multiple heterogeneous networks
    - Seseorang bisa berupa beberapa jaringan informasi: teman, keluarga, teman sekelas, ...
  - Links membawa banyak informasi semantik: Link mining
- Web mining
  - Web adalah jaringan informasi yang besar: dari PageRank hingga Google
  - Analisis jaringan informasi Web
    - Penemuan komunitas web, penambangan opini, penambangan penggunaan, ...





# Evaluasi Pengetahuan

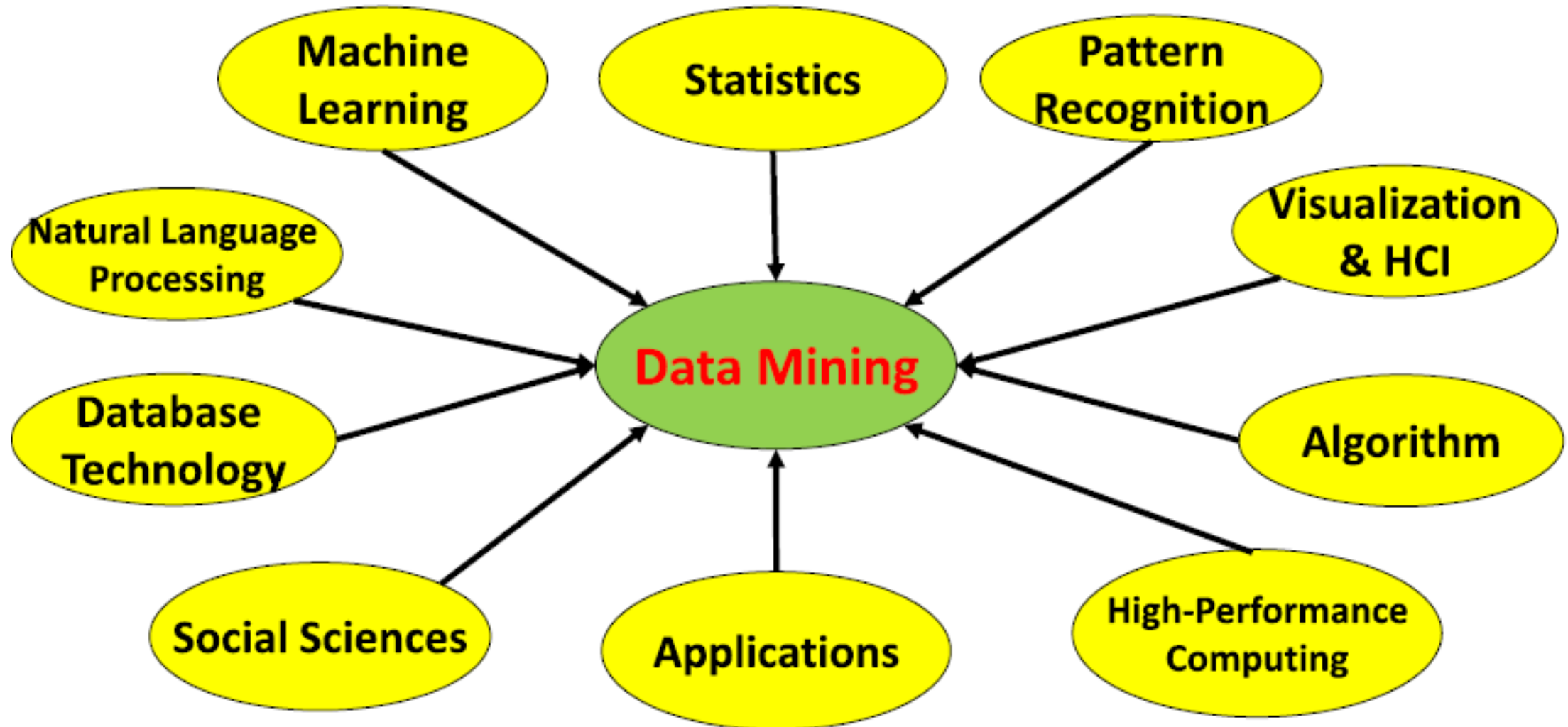
---

- ❑ Apakah semua pengetahuan yang ditambang menarik?
  - ❑ Seseorang dapat menambang sejumlah besar "pola"
  - ❑ Beberapa mungkin hanya muat ruang dimensi tertentu (waktu, lokasi, ...)
  - ❑ Beberapa mungkin tidak representatif, mungkin sementara, ...
- ❑ Evaluasi pengetahuan yang ditambang → langsung menambang hanya pengetahuan yang menarik?
  - ❑ Descriptive vs. predictive
  - ❑ Coverage
  - ❑ Typicality vs. novelty
  - ❑ Accuracy
  - ❑ Timeliness
  - ❑ ...



# Data Mining: Pertemuan Berbagai Disiplin Ilmu

---






# Mengapa Pertemuan Berbagai Disiplin Ilmu?

---

- ❑ Jumlah data yang luar biasa
  - ❑ Algoritma harus dapat diskalakan untuk menangani big data
- ❑ Data dimensi tinggi
  - ❑ Micro-array mungkin memiliki puluhan ribu dimensi
- ❑ Kompleksitas data yang tinggi
  - ❑ Data streams dan Data sensor
  - ❑ Time-series data, temporal data, sequence data
  - ❑ Structure data, graphs, social dan information networks
  - ❑ Spasial, spatiotemporal, multimedia, Teks and Data web
  - ❑ Software programs, scientific simulations
- ❑ Aplikasi baru dan canggih

# Penambangan Data dan Aplikasi

- ❑ Web page analysis: classification, clustering, ranking
  - ❑ Collaborative analysis & recommender systems
  - ❑ Basket data analysis untuk pemasaran yang ditargetkan
  - ❑ Analisis data Biologis dan Medis
  - ❑ Data mining dan software engineering
  - ❑ Data mining dan text analysis
  - ❑ Data mining dan social dan information network analysis
  - ❑ Built-in (invisible data mining) fungsi dalam Google, Microsoft, LinkedIn, Meta, ...
  - ❑ Sistem/alat penambangan data khusus utama
    - ❑ SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools)
- 



# Data Mining dan Masyarakat

---

- ❑ Teknologi Data mining technology dapat bermanfaat bagi masyarakat
  - ❑ Misalnya: Membantu penemuan ilmiah, manajemen bisnis, pemulihan ekonomi, dan perlindungan keamanan (misalnya, penemuan penyusup dan serangan siber secara real-time)
- ❑ Perlu menjaga dari penyalahgunaan data mining
  - ❑ Data mining juga menimbulkan risiko secara tidak sengaja mengungkapkan beberapa informasi rahasia bisnis atau pemerintah dan mengungkapkan informasi pribadi individu
- ❑ Studi tentang keamanan data di data mining and Menjaga privasi data publishing dan data mining adalah tema penelitian yang penting dan berkelanjutan
  - ❑ Filosofinya adalah untuk mengamati sensitivitas data dan menjaga keamanan data dan privasi orang sambil melakukan penambangan data yang sukses

# Simpulan

---

- ❑ Data mining: Menemukan pola dan pengetahuan menarik dari data dalam jumlah besar
- ❑ Proses KDD meliputi data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, dan knowledge presentation
- ❑ Metode data mining yang berbeda pada berbagai macam data
- ❑ Fungsi Data mining : summarization, pattern discovery, classification, clustering, deep learning, outlier analysis, trend and outlier analysis, ...
- ❑ Data mining adalah pertemuan dari berbagai disiplin ilmu
- ❑ Data mining memiliki aplikasi yang luas
- ❑ Mempromosikan penambangan data yang aman untuk bermanfaat bagi masyarakat