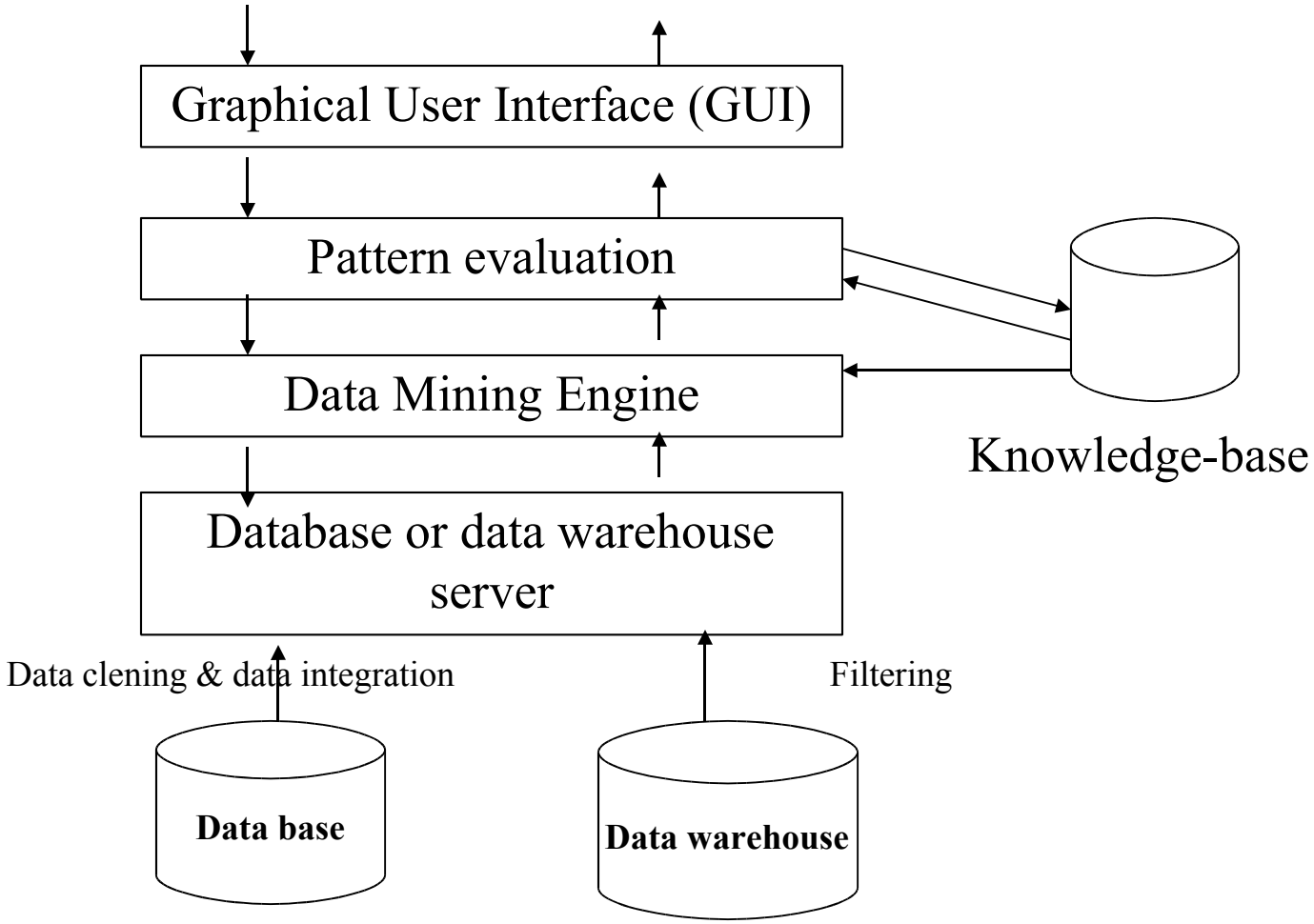
PERTEMUAN 3

ARSITEKTUR & MODEL DATA MINING

Arsitektur : Sistm Data Mining



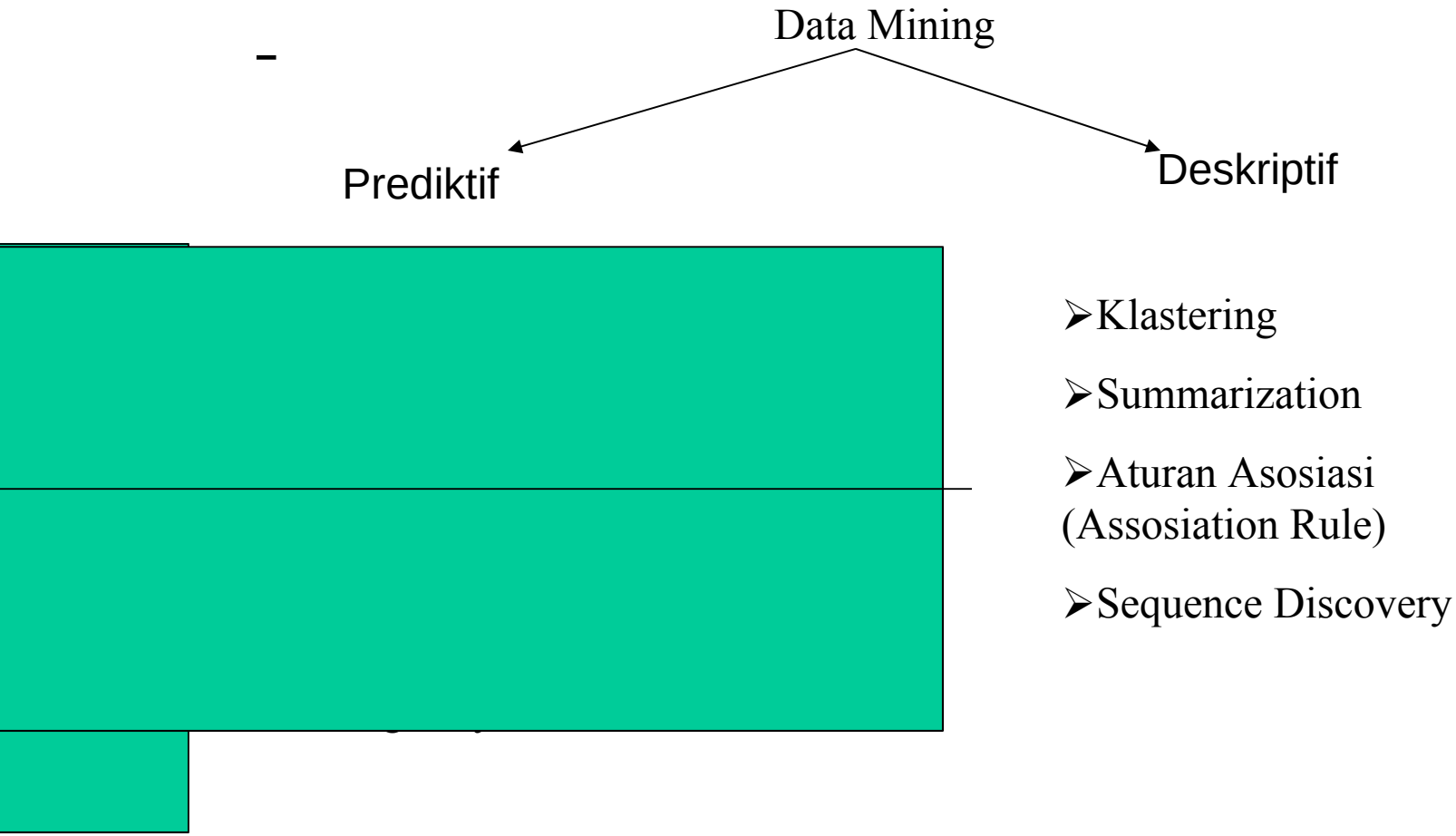
Keterangan :

1. Data cleaning (Pembersihan Data) : untuk membuang data yang tidak konsisten dan noise)
2. Data integration : penggabungan data dari beberapa sumber
3. Data Mining Engine : Mentranformasikan data menjadi bentuk yang sesuai untuk di mining
4. Pattern evaluation : untuk menemukan yang bernilai melalui knowledge base
5. Graphical User Interface (GUI) : untuk end user

Semua tahap bersifat interaktif di mana user terlibat langsung atau dengan perantaraan knowledge base

Model Data Mining

- Prediction Methods
  - –Menggunakan beberapa variabel untuk memprediksi sesuatu atau suatu nilai yang akan datang.
- Description Methods
  - – Mendapatkan pola penafsiran (human-interpretable patterns) untuk menjelaskan data.



- Klastering
- Summarization
- Aturan Asosiasi (Assosiation Rule)
- Sequence Discovery

Klasifikasi

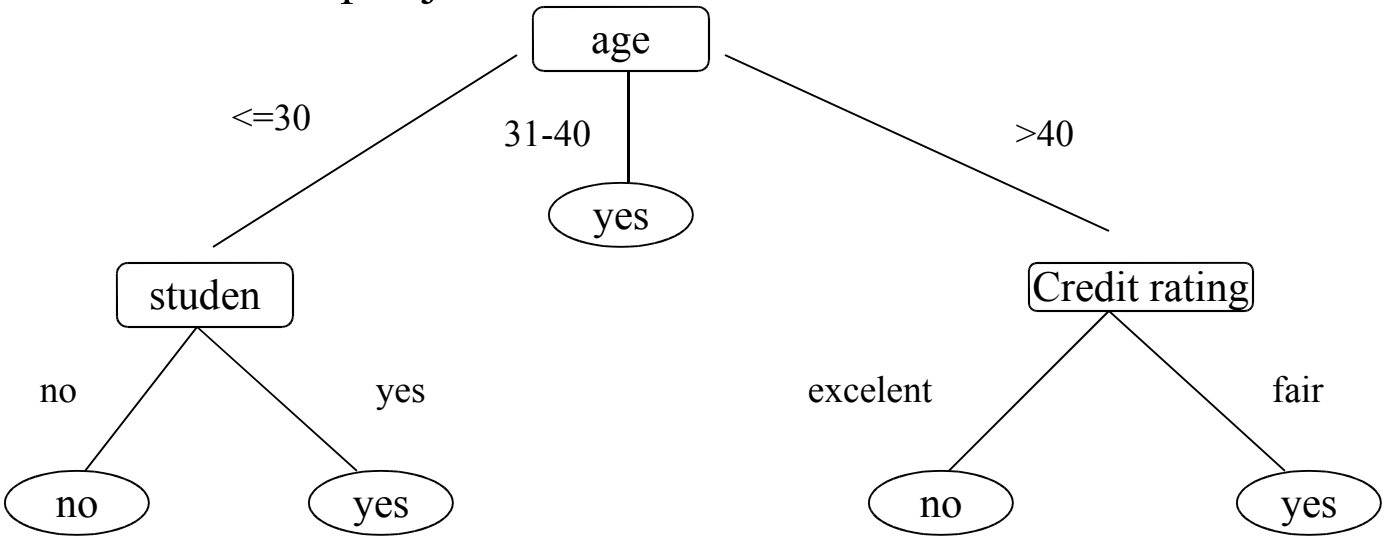
- Proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk dapat memprediksi kelas dari suatu objek yang labelnya tidak diketahui
- **Contoh : Mendeteksi Penipuan**
- **Tujuan : Memprediksi kasus kecurangan transaksi kartu kredit.**
  - **Pendekatan :**
    - **Menggunakan transaksi kartu kredit dan informasi dilihat dari atribut account holder**
      - **Kapan cutomer melakukan pembelian, Dengan cara apa customer membayar, sebarapa sering customer membayar secara tepat waktu, dll**
    - **Beri nama/tanda transaksi yang telah dilaksanakan sebagai transaksi yang curang atau yang baik. Ini sebagai atribut klass (the class attribute.)**
    - **Pelajari model untuk class transaksi**
    - **Gunakan model ini untuk mendetekdi kecurangan dengan meneliti transaksi kartu kredit pada account.**

Regression

- Digunakan untuk memetakan data dengan prediksi atribut bernilai real
- Contoh:
  - Memprediksi jumlah penjualan produk baru pada advertising expenditure.
  - Memprediksi kecepatan memutar (wind velocities) pada fungsi temperatur, tekanan udara , dll

Decision tree (Pohon keputusan)

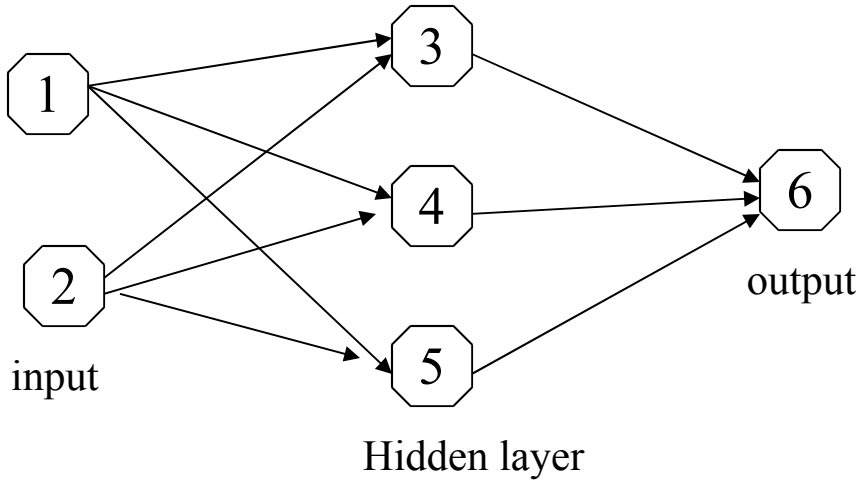
- Salah satu model klasifikasi yang mudah di interpretasikan
- Contoh : identifikasi pembeli komputer (dari decision tree di bawah ini ternyata salah satu kelompok yang potensial adalah orang yang berusia < 30 dan pelajar



Prediksi_

Neural Network (Jaringan syaraf tiruan)

- Jaringan syaraf buatan di mulai dengan layer input, dimana tiap simpul berkorespondensi dengan variabel prediktor.
- Simpul- simpul input ini terhubung kebeberapa simpul dalam hidden layer.
- Dan simpul dalam hidden layer dapat terhubung ke simpul lain dalam hidden layer atau ke output layer.
- Output layer terdiri dari satu atau beberapa variable respon



- Telekomunikasi

  Data mining digunakan untuk melihat jutaan transaksi yang masuk dengan tujuan menambah layanan otomatis

- Keuangan_

  Data mining digunakan untuk mendeteksi transaksi-transaksi keuangan yang mencurigakan dimana akan susah dilakukan jika menggunakan analisis standar.

- Asuransi

  Australian Health Insurance Commision menggunakan data mining untuk mengidentifikasi layanan kesehatan dan berhasil menghemat satu juta dollar pertahun

- Olah raga

  IBM Advanced Scout menggunakan data mining untuk menganalisis statistik permainan NBA dalam rangka competitive advantage untuk tim New York Knicks

- Astronomi

  Jet Propulsion Laboratory (JPL) di Pasadena dan Pulomar Observatory menemukan 22 quasar dengan bantuan data mining.

- Internet Web Surf-Aid_

  IBM Surf-Aid menggunakan algoritma data mining untuk mendata akses halaman Web khususnya berkaitan dengan pemasaran melalui web.

Tools Data Mining

- Karateristik-karateristik penting dari tool data mining meliputi :
  - Data preparation facilities
  - Selection of data mining operation (algorithms)
  - Product scalability and performance
  - Facilities for visualization of result

- Data mining tool, meliputi :
  - Integral Solution Ltd's Clementine
  - DataMind Corp's Data Crusher
  - IBM's Intelligent Miner
  - Silicon Graphics Inc.'s MineSet
  - Informations Discovery Inc.'s Data Mining Suite
  - SAS Institute Inc.'s SAS System and Right Information System'Thought.

Evolusi Database

- Th 1960
  - Pengumpulan data, pembuatan data, IMS dan network DBMS
- Th 1970
  - Model data relasional, Implementasi DBMS relasional
- Th 1980
  - RDBMS, Model data lanjutan (extended-relational, OO, deductive)
- Th 1990
  - Data mining, data warehouse, database multimedia, dan Web database.
- Th 2000
  - Stream data managemen dan mining
  - Data mining dengan berbagai variasi aplikasi
  - Teknologi web dan sistem informasi global

Teknik – teknik Database

Searching

- Searching dilakukan untuk memeriksa serangkaian item yang memiliki sifat-sifat yang diinginkan.
- Tindakan untuk menemukan suatu item tertentu baik yang diketahui keberadaannya maupun tidak.
- Memasukkan kata dalam suatu program komputer untuk membandingkan dengan informasi yang ada dalam database.

Indexing

- Indexing adalah struktur-struktur akses yang digunakan untuk mempercepat respon dalam mendapatkan record-record pada kondisi-kondisi pencarian tertentu.
- Indexing field adalah suatu struktur akses index yang biasanya menjelaskan field tunggal dari suatu file.
- Indexing organization memberikan efisiensi akses ke record-record secara berurut atau random.

Data Reduction_

- *Data reduction* adalah transformasi suatu masalah ke masalah lain dan dapat digunakan untuk mendefinisikan serangkaian masalah yang kompleks.
- *Data reduction* merupakan teknik yang digunakan untuk mentransformasi dari data mentah ke bentuk format data yang lebih berguna. Sebagai contoh *groupping*, *summing* dan *averaging data*.
- *Data reduction* dilakukan untuk mengatasi ukuran data yang terlalu besar. Ukuran data yang terlalu besar dapat menimbulkan ketidakefisienan proses dan peningkatan biaya pemrosesan.
- *Data reduction* dilakukan dalam tahap data preprocessing pada rangkaian proses *Knowledge Discovery Databases* (KDD) sebelum data mining dengan tujuan mengurangi ukuran data yang besar.

OLAP (On-line analytical processing)

- OLAP adalah suatu sistem atau teknologi yang dirancang untuk mendukung proses analisis kompleks dalam rangka mengungkapkan kecenderungan pasar dan faktor-faktor penting dalam bisnis

- OLAP ditandai dengan kemampuannya menaikkan atau menurunkan dimensi data sehingga kita dapat menggali data sampai pada level yang sangat detail dan memperoleh pandangan yang lebih luas mengenai objek yang sedang kita analisis.

- OLAP secara khusus memfokuskan pada pembuatan data agar dapat diakses pada saat pendefinisian kembali dimensi.

- OLAP dapat digunakan membuat rangkuman dari multidimensi data yang berbeda, rangkuman baru dan mendapatkan respon secara online, dan memberikan view dua dimensi pada data cube multidimensi secara interaktif.

Data Warehouse

Definisi :

- Data Warehouse adalah Pusat repositori informasi yang mampu memberikan database berorientasi subyek untuk informasi yang bersifat historis yang mendukung DSS (Decision Suport System) dan EIS (Executive Information System).
- Salinan dari transaksi data yang terstruktur secara spesifik pada query dan analisa.
- Salinan dari transaksi data yang terstruktur spesifik untuk query dan laporan

Tujuan :

- Meningkatkan kualitas dan akurasi informasi bisnis dan mengirimkan informasi ke pemakai dalam bentuk yang dimengerti dan dapat diakses dengan mudah.

Ciri-ciri Data Warehouse

Terdapat 4 karateristik data warehouse

- Subject oriented
  - Data yang disusun menurut subyek berisi hanya informasi yang penting bagi pemprosesan decision support.
  - Database yang semua informasi yang tersimpan di kelompokkan berdasarkan subyek tertentu misalnya: pelanggan, gudang, pasar, dsb.
  - Semua Informasi tersebut disimpan dalam suatu sistem *data warehouse*.
  - Data-data di setiap subyek dirangkum ke dalam dimensi, misalnya : periode waktu, produk, wilayah, dsb, sehingga dapat memberikan nilai sejarah untuk bahan analisa.

- Integrated
  - Jika data terletak pada berbagai aplikasi yang terpisah dalam suatu lingkungan operasional, encoding data sering tidak seragam sehinggga bila data dipindahkan ke data warehouse maka coding akan diasumsikan sama seperti lazimnya.

- Time-variant
  - Data warehouse adalah tempat untuk storing data selama 5 sampai 10 tahun atau lebih, data digunakan untuk perbandingan atau perkiraan dan data ini tidak dapat diperbaharui.

- Non volatile
  - Data tidak dapat diperbaharui atau dirubah tetapi hanya dapat ditambah dan dilihat.

Masalah-masalah dalam menerapkan *Data warehouse* :

- Dokumentasi dan pengelolaan metadata dari *data warehouse.*
- Penentuan aturan dalam proses transformasi untuk memetakan berbagai sumber legacy data yang akan dimasukkan ke dalam *data warehouse*.
- Pencapaian proses pengembangan yang handal, baik dalam membangun, mempimplementasikan, maupun memelihara *data warehouse*.

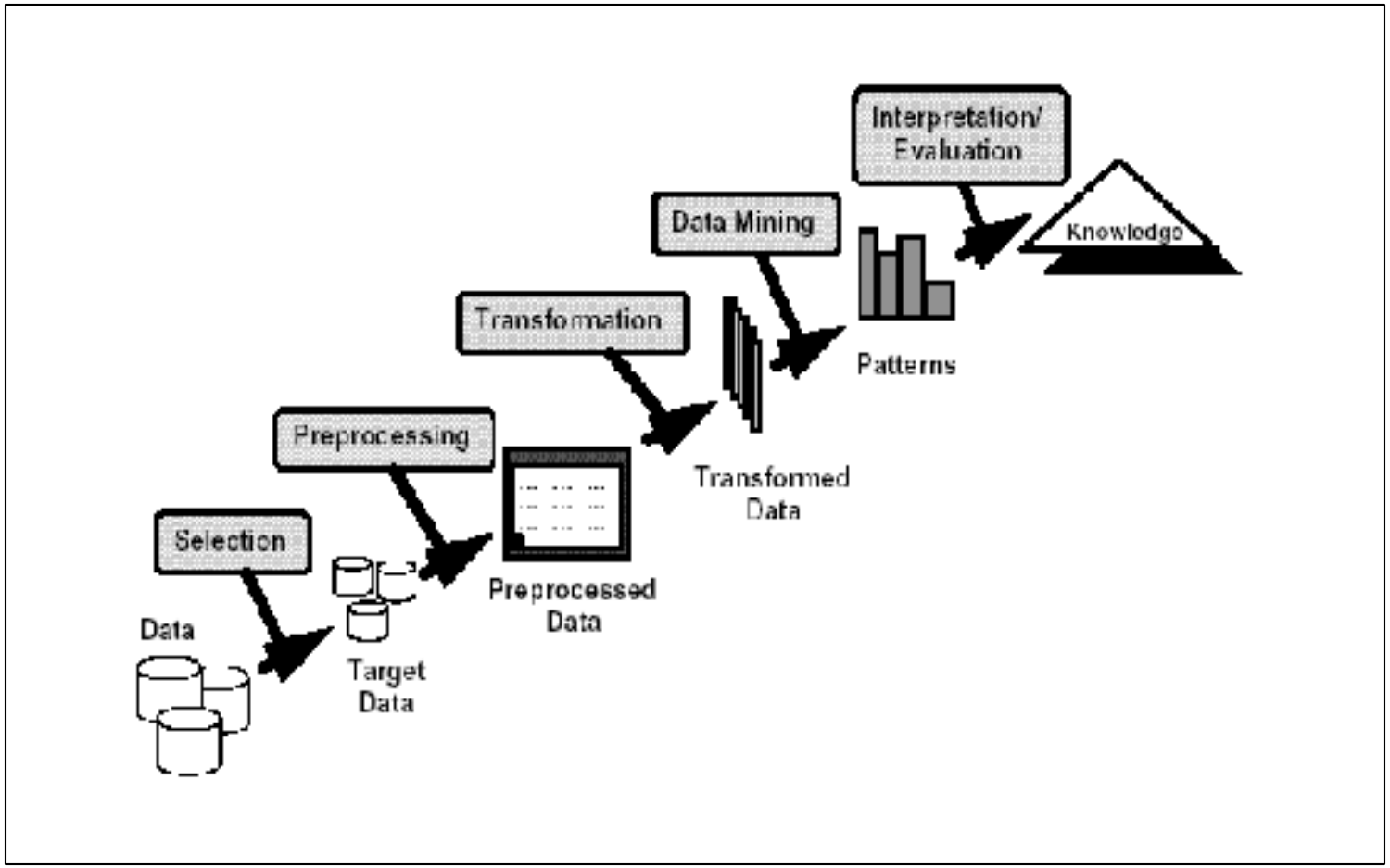
*Data Preprocessing_*

- Data preprocessing menerangkan tipe-tipe proses yang melaksanakan data mentah untuk mempersiapkan proses prosedur yang lainnya.
- Dalam data mining menstrasformasi data ke suatu format yang prosesnya lebih mudah dan efektif untuk kebutuhan pemakai, contohnya Neural Network.
- Terdapat beberapa alat dan metode yang berbeda yang digunakan untuk preprocessing seperti :
  - *Sampling :* menyeleksi subset representatif dari populasi data yang besar.
  - *Transformation :* memanipulasi data mentah untuk menghasilkan input tunggal.
  - *Denoising :* menghilangkan noise dari data
  - *Normalization :* mengorganisasi data untuk pengaksesan yang lebih spesifik
  - *Feature extration :* membuka spesifikasi data yang signifikan dalam konteks tertentu.

Knowledge Discovery In Database (KDD)

- KDD berhubungan dengan teknik integrasi dan penemuan ilmiah, interprestasi dan visualisasi dari pola-pola sejumlah kumpulan data.

- *Knowledge discovery in databases* (KDD) adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti.



Gambar. 1. Tahapan KDD

Tahapan Proses KDD

3. *Data Selection*

- Menciptakan himpunan data target , pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan.

- Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses *data mining*, disimpan dalam suatu berkas, terpisah dari basis data operasional.

2. *Pre-processing/ Cleaning*

- Pemprosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan noise dilakukan.
- Sebelum proses *data mining* dapat dilaksanakan, perlu dilakukan proses *cleaning* pada data yang menjadi fokus KDD.
- Proses *cleaning* mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi).
- Dilakukan proses *enrichment*, yaitu proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

2. *Transformation*

- Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai.

- Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses *data mining*. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data

2. *Data mining*

- Pemilihan tugas data mining; pemilihan goal dari proses KDD misalnya klasifikasi, regresi, clustering, dll.

- Pemilihan algoritma data mining untuk pencarian (searching)

- *Proses Data mining* yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam *data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. *Interpretation/ Evaluation*

- Penerjemahan pola-pola yang dihasilkan dari *data mining*.
- Pola informasi yang dihasilkan dari proses *data mining* perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.
- Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.