DATA SET

Nugroho D.S.

Jenis Dataset

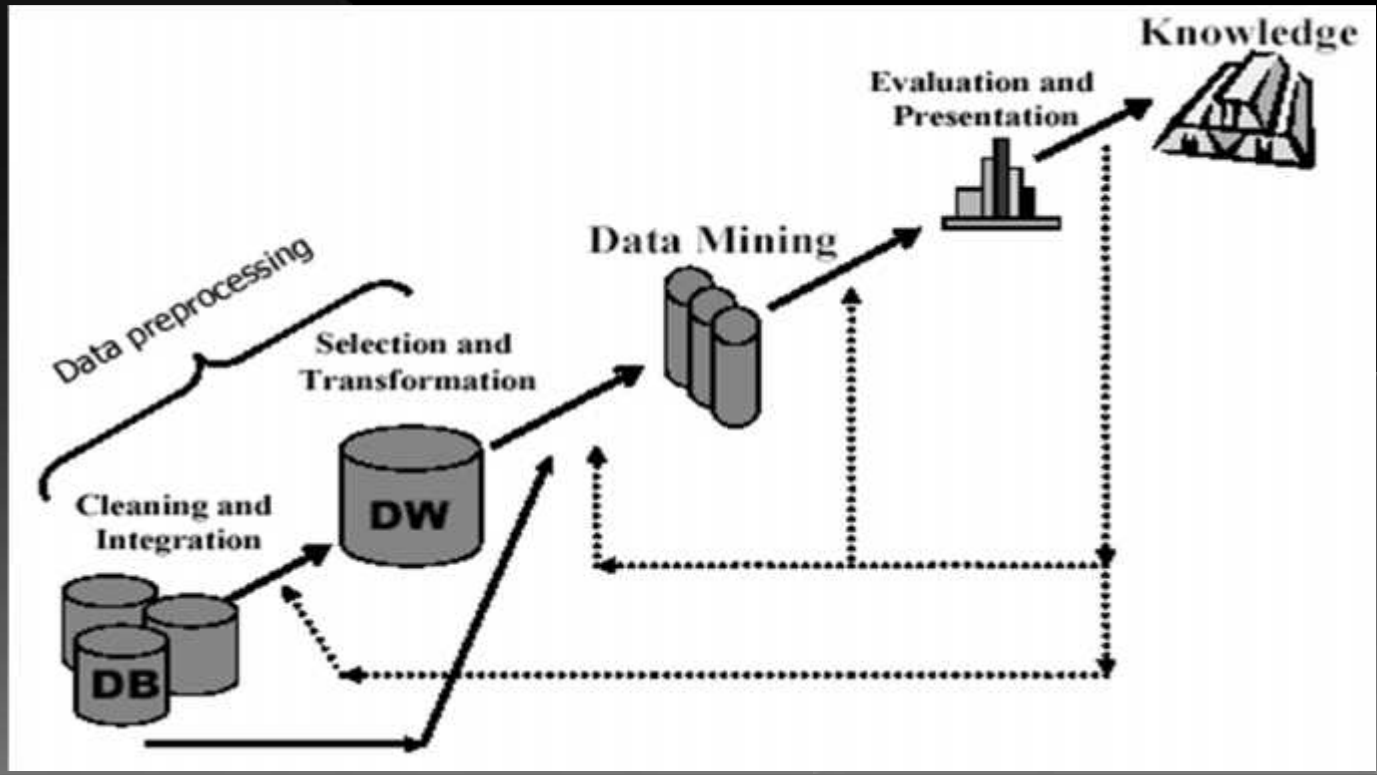
- **Private**
  - Bank, Rumah Sakit, Industri, Pabrik, Perusahaan Jasa, etc
- **Public**
  - **UCI Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html)**
  - **ACM KDD Cup (http://www.sigkdd.org/kddcup/)**
- Trend penelitian data mining saat ini adalah menguji metode yang dikembangkan oleh peneliti dengan public dataset, sehingga penelitian dapat bersifat: **comparable**, **repeatable** dan **verifiable**

Atribut, Class dan Tipe Data

- Atribut adalah **faktor atau parameter yang menyebabkan** class/label/target terjadi
- Class adalah atribut yang akan dijadikan **target**, sering juga disebut dengan **label**
- Tipe data untuk variabel pada statistik terbagi menjadi empat: nominal, ordinal, interval, ratio
- Tapi secara praktis, tipe data untuk atribut pada data mining hanya menggunakan dua:
  1. **Nominal** (Diskrit)
  2. **Numeric** (Kontinyu atau Ordinal)

Data preparation

- Ada 7 (tujuh) tahapan proses data mining, dimana 4 (empat) tahap pertama disebut juga dengan data preprocessing (terdiri dari *data selection, data cleaning, data integration, dan data transformation)*



Data Selection

- pemilihan himpunan data, atau memfokuskan pada sampel data, Di mana data yang relevan.

Data Cleaning

Dalam data cleaning yang akan kita lakukan antara lain mengisi missing value, mengidentifikasi outlier, menangani data noise, mengoreksi data yang tidak konsisten.

Data Integration

- Data integration adalah suatu langkah untuk menggabungkan data dari beberapa sumber. Data integration hanya dilakukan jika data berasal dari tempat yang berbeda-beda (sumber data tidak hanya dari 1 tempat).

Data Transformation

Data transformation yaitu mengubah suatu data supaya diperoleh data yang lebih berkualitas. Yang akan dilakukan antara lain menghilangkan noise dari data (smoothing), meng-agregasi data, generalisasi data, normalisasi data, dan pembentukan atribut/fitur.