Pengertian Data, Jenis dan Tipe Data, Deskripsi Data, Data Preprocessing

Dr. Eng. Chastine Fatichah, S.Kom, M.Kom

Departemen Teknik Informatika

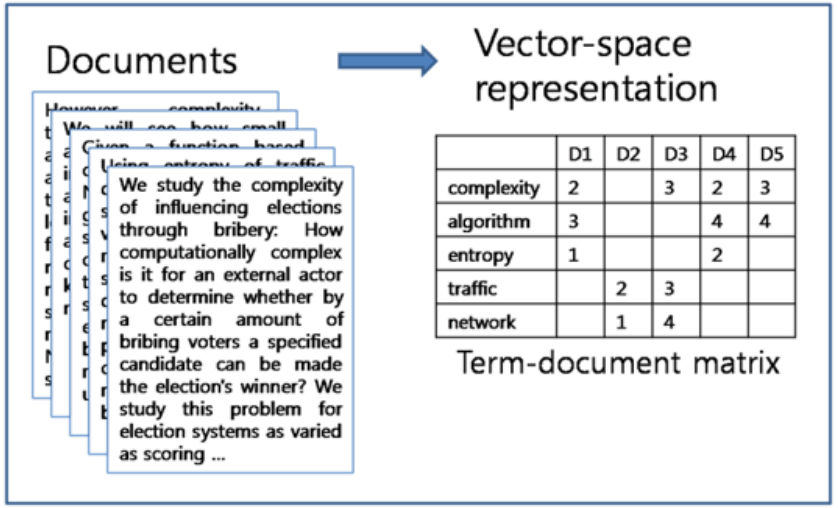
Institut Teknologi Sepuluh Nopember

chastine@if.its.ac.id

Jenis Data

- Record
  - Data matrix
  - Document data (document-term matrix)
  - Transaction data

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin
0	1	1	Allen, Miss. Elisabeth Walton	female	29.00	0	0	24160	211.3375	B5
1	1	1	Allison, Master. Hudson Trevor	male	0.92	1	2	113781	151.5500	C22 C26
2	1	0	Allison, Miss. Helen Loraine	female	2.00	1	2	113781	151.5500	C22 C26
3	1	0	Allison, Mr. Hudson Joshua Creighton	male	30.00	1	2	113781	151.5500	C22 C26

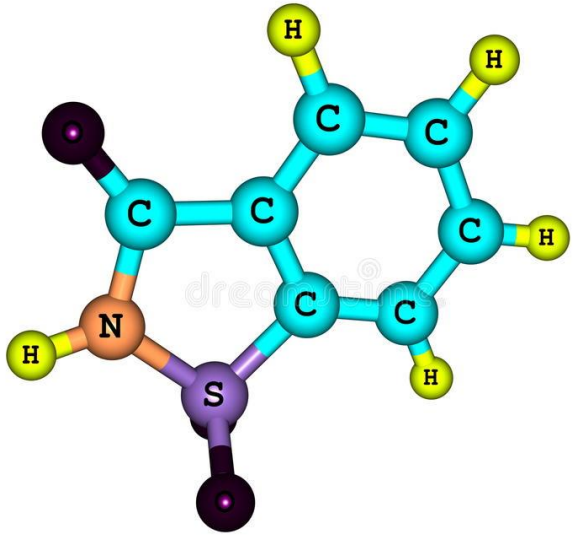


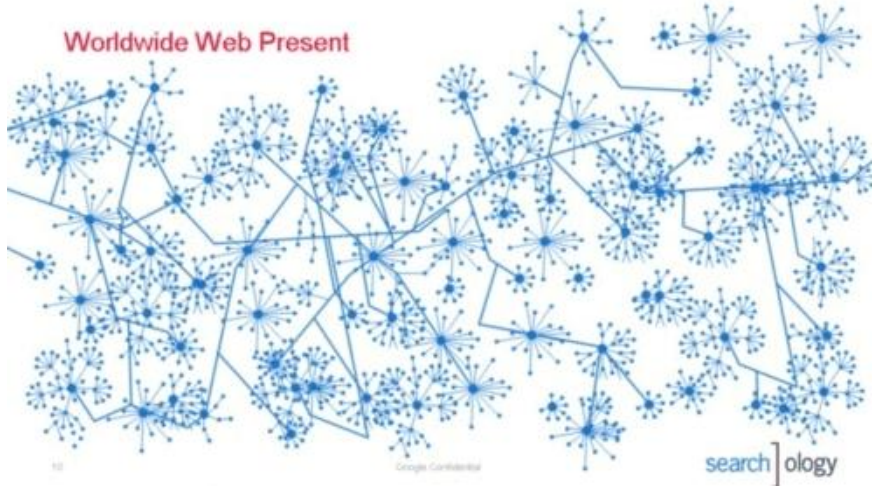
	D1	D2	D3	D4	D5
complexity	2		3	2	3
algorithm	3			4	4
entropy	1			2	
traffic		2	3		
network		1	4		

TID	Items
1	Bread, Coke, Milk
2	Bread, Jam
3	Coke, Milk, Chips
4	Bread, Jam, Chip, Milk
5	Coke, Jam, Chip, Milk

Jenis Data

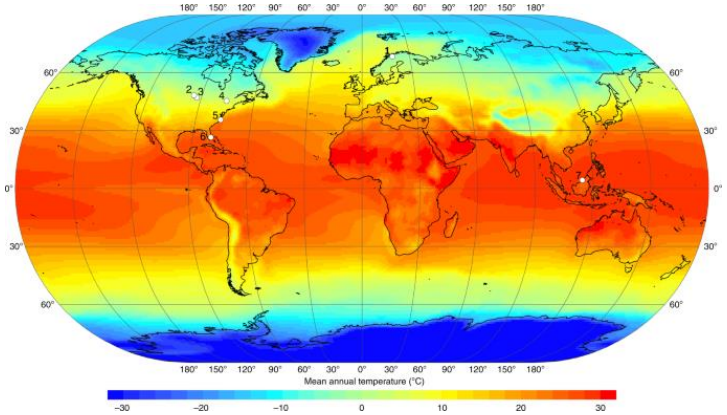
- Graph and network
  - World Wide Web
  - Social networks
  - Molecular Structures





Jenis Data

- Ordered
  - Video data
  - Spatio-Temporal data
  - Sequential Data
  - Genetic sequence data
- Spatial, image, and multimedia
  - Spatial data (maps)
  - Image data
  - Voice data
  - Video data

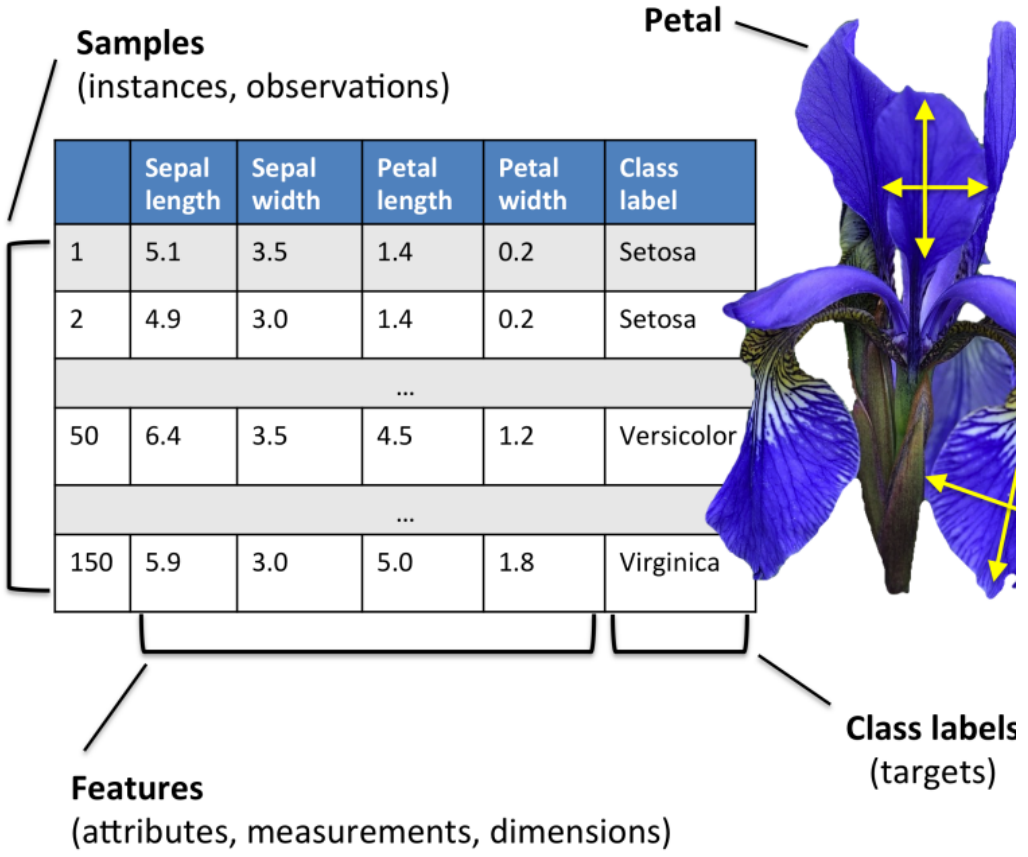


(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Pengertian Data

- **Data (Dataset)** merupakan kumpulan dari data obyek yang merepresentasikan sebuah entitas (atribut)
- **Data obyek** disebut juga sebagai *record, point, sample, instance*
- **Atribut** merepresentasikan karakteristik sebuah data obyek
  - Misalnya: tinggi badan, berat badan, usia, jenis kelamin.
  - Atribut disebut juga sebagai variabel, fitur
- Contoh Dataset:
  - Penjualan: tanggal penjualan, nama pelanggan, nama barang, jumlah penjualan



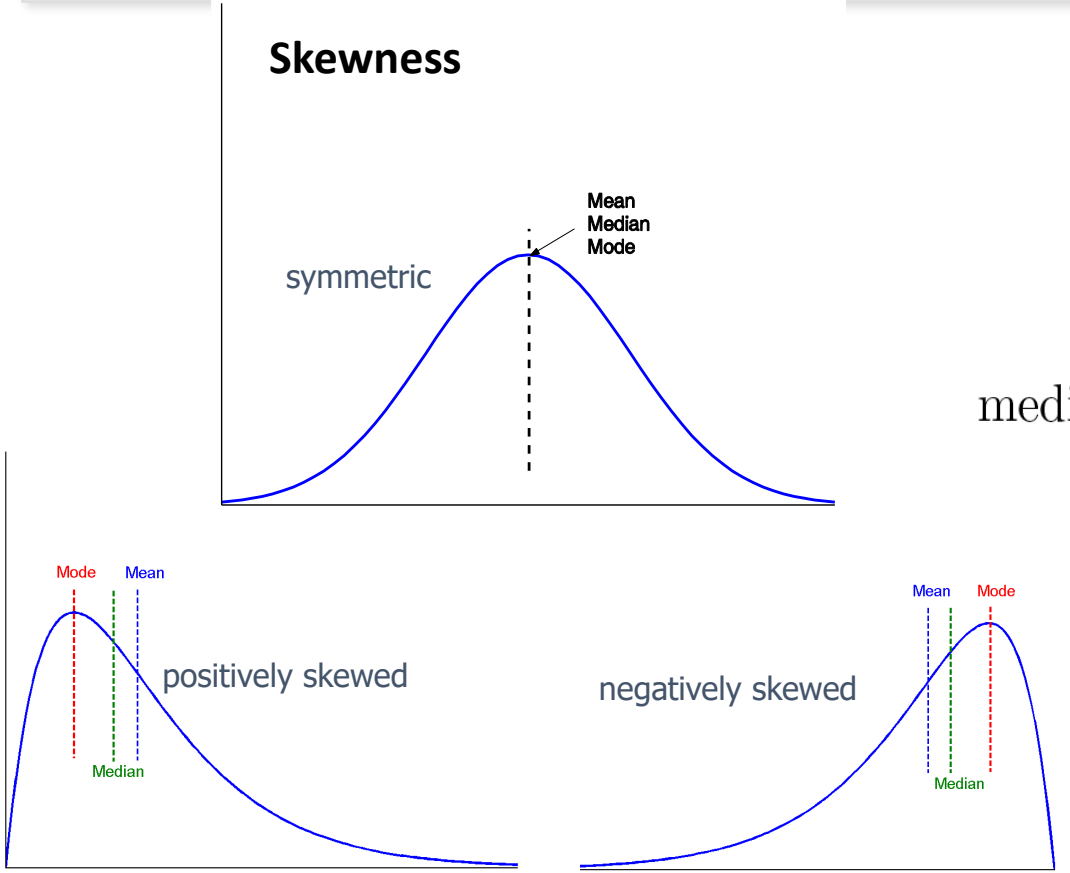
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

Tipe Atribut

- **Nominal**
  - Berupa kategori, contoh: jenis kelamin, status perkawinan,…
- **Binary**
  - Atribut nominal dengan hanya 2 nilai yaitu 0 dan 1
- **Ordinal**
  - Nilai yang merepresentasikan urutan, contoh: ukuran, nilai matakuliah, …
- **Numeric**
  - Quantity (integer atau real-valued)
  - Interval
    - Ukuran skala unit, contoh: suhu, tanggal
  - Ratio
    - Panjang, harga, umur

Statistik Dasar untuk Deskripsi Data

- Agar dapat memahami data terkait pusat distribusi, variasi, dan sebaran data
- Pengukuran tendensi sentral: mean, median, mode



$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

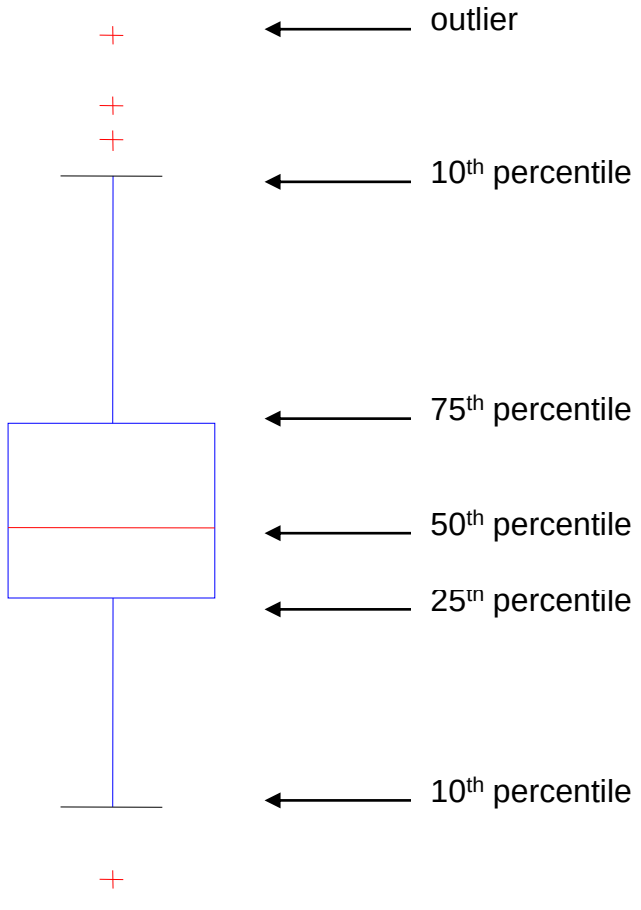
$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Statistik Dasar untuk Deskripsi Data

- Pengukuran sebaran data
  - Variance and Standard Deviation

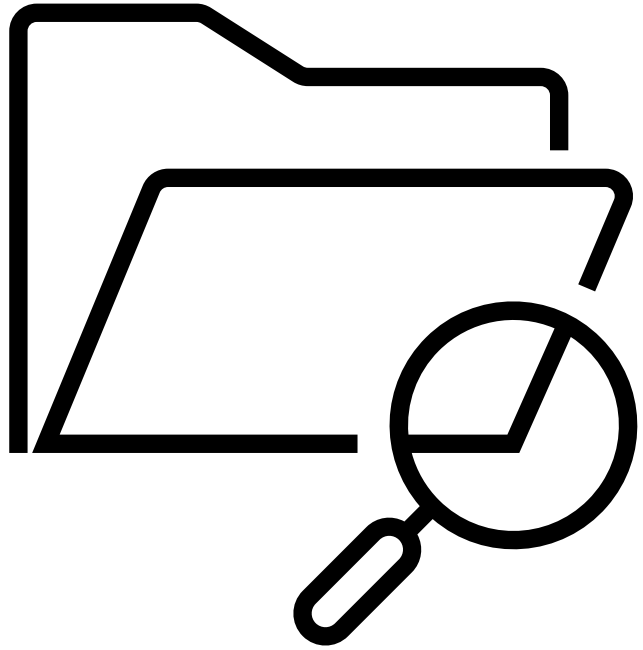
$$\text{variance}(x) = s_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \overline{x})^2$$

  - Quartiles, outliers and boxplots
    - Quartiles: Q1 (25th percentile), Q3 (75th percentile)
    - Inter-quartile range: IQR = Q3 – Q1
    - Boxplot
    - Outlier: nilai yang lebih tinggi/rendah dari 1.5 x IQR



Data Quality

- ***Accuracy:*** benar atau salah, akurat atau tidak
- ***Completeness***: ada yang tidak tercatat, tidak tersedia, …
- ***Consistency***: tidak konsisten
- ***Timeliness***: apakah terupdate?
- ***Believability***: seberapa dipercaya data itu benar?
- ***Interpretability***: seberapa mudah data dapat dipahami?



Data Quality

- Noise and outliers
- Missing values
- Duplicate data

Data Preprocessing

Data cleaning	Imputasi (missing values), smoothing (noisy data), identifikasi atau penghapusan (outliers), dan penanganan (data inconsistencies)
Data integration	Integrasi dari multiple databases, data cubes, atau files
Data reduction	Dimensionality reduction Numerosity reduction Data compression
Data transformation and data discretization	Normalisasi Diskritasi data

Data Cleaning

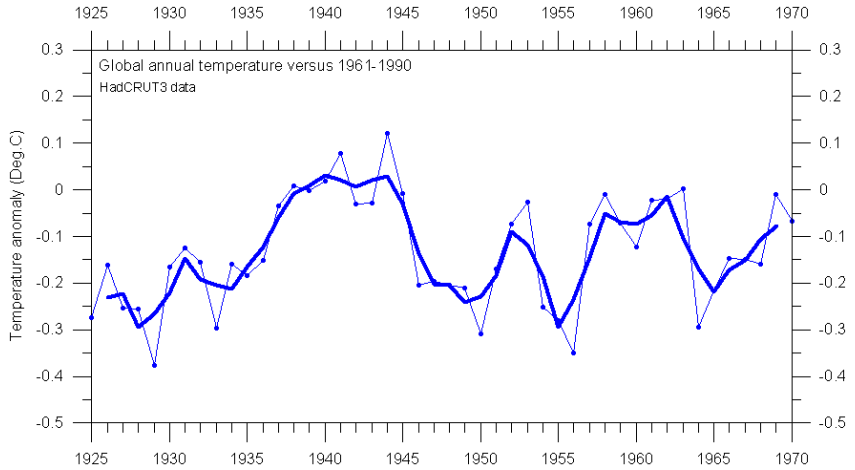
- Data riil umumnya kotor dan potensi adanya *incorrect* data karena *instrument faulty*, *human or computer error*, dan *transmission error*
  - *Incomplete* (*missing value*)
    - Contoh: pekerjaan = “ “
  - *Noisy*: berisi noise, error, atau outlier
    - Contoh: gaji = “−20” (error)
  - *Inconsistent*: ada perbedaan pada kode dan nama
    - Contoh: usia = “31”, tanggal lahir =“24/02/2000”
    - Contoh: sebelumnya rating “1, 2, 3”, sekarang rating “A, B, C”
  - *Intentional*
    - Jan. 1 default tanggal lahir

Data Cleaning

- Bagaimana menangani incomplete data (*missing value*)?
  - Eliminasi data object
  - Dibiarkan (*ignore*)
  - Imputasi nilai secara manual
  - Imputasi nilai secara otomatis menggunakan
    - Nilai mean, median, mode dari atribut yang ada missing value
    - Nilai mean, median, mode untuk semua sampel yang memiliki kelas yang sama
    - Nilai estimasi menggunakan metode Bayesian, Decision tree, Regresi, k-Nearest Neighbor, Expectation Maximization,…

Data Cleaning

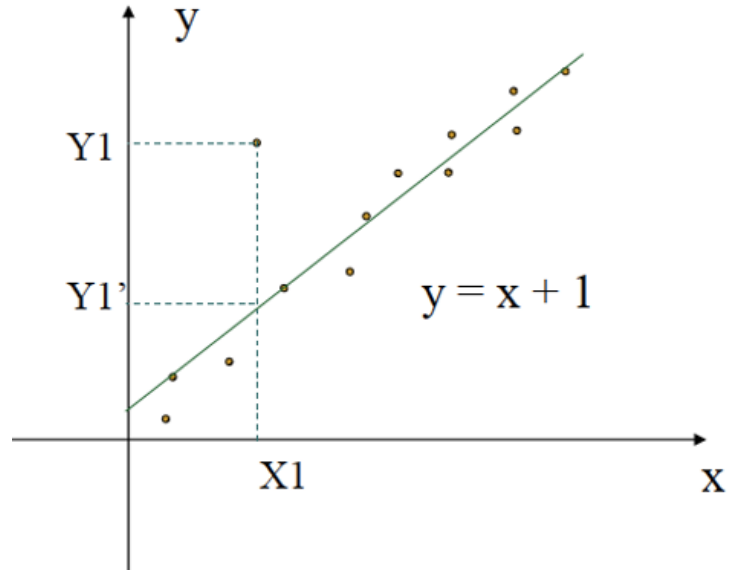
- Bagaimana menangani noisy data?
  - Binning
    - Mengurutkan data dan membagi menjadi beberapa bins berdasarkan frequency (*equal-frequency*)
    - Kemudian melakukan *smooth by bin means, smooth by bin median, smooth by bin boundaries*, dsb.

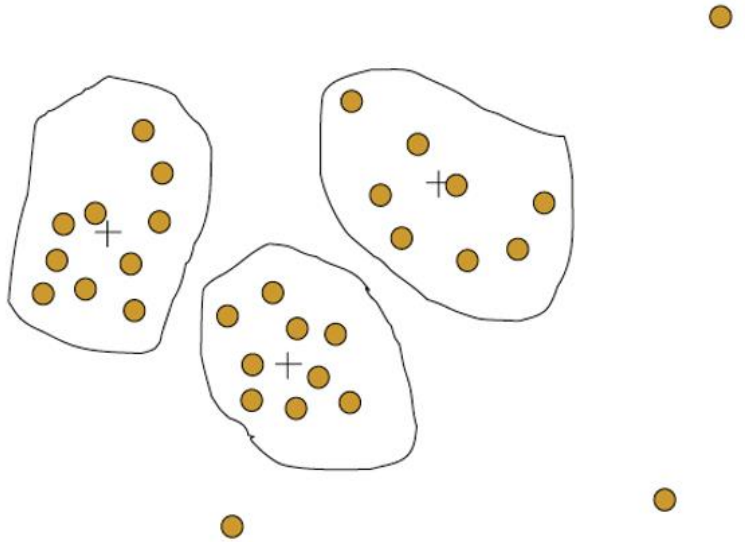


* Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Data Cleaning

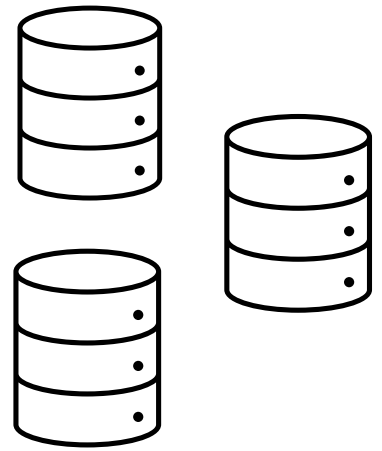
- Bagaimana menangani noisy data?
  - Regression
    - Melakukan *smooth by fitting* data ke fungsi regresi
  - Clustering
    - Mendeteksi dan menghapus outliers
  - *Combined computer and human inspection*
    - Mendeteksi nilai yang meragukan dan dicek secara manual





Data Integration

- Bagaimana menangani redundansi data ketika melakukan data intergasi dari multiple databases
  - Identifikasi: Atribut sama namun mempunya nama yang berbeda
  - Derivasi: Satu atribut bisa diderivasi dari atribut lain pada database lain, misalnya: total gaji setahun
- Redundansi attribut bisa dideteksi menggunakan analisis korelasi dan kovarian
  - *Chi square test* untuk tipe nominal
  - *Pearson correlation* untuk tipe numerik



Data Reduction

- Mengurangi representasi data menjadi lebih kecil
- Motivasi: Analisis data yang kompleks memerlukan waktu komputasi yang lama
- Strategi
  - ***Dimensionality reduction*** **(menghapus atribut tidak penting)**
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - ***Numerosity reduction (Data Reduction)***
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - ***Data compression***

Data Transformation

- Proses yang mentransformasi nilai asli ke nilai baru
- Metode atau Pendekatan
  - ***Smoothing***: menghapus noise
  - ***Attribute/feature construction***
    - Membuat atribut baru dari atribut yang sudah ada
  - ***Aggregation***
  - ***Normalization***
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling

Sumber referensi

- Data Mining: Concepts and Techniques (3rd Edition), Jiawei Han, Micheline, Kamber, and Jian Pei, University of Illinois at Urbana-Champaign & Simon Fraser University, 2011
- Introduction to Data Mining, Tan, Steinbach, Kumar, 2004

Terimakasih