

# Pengantar Data Mining #3: Praproses Data [1]

Isnan Mulia, S.Komp, M.Kom

# Kualitas Data

- Data yang akan digunakan dalam *data mining* tidak selalu dalam kondisi yang siap pakai
- Kemungkinan yang bisa terjadi:
  - Data tidak lengkap; ada atribut yang tidak ada nilainya
  - Data mengandung nilai yang tidak seharusnya diisikan
  - Data tidak konsisten, khususnya yang didapatkan dari berbagai sumber
  - Ukuran data terlalu besar

Faktor yang mempengaruhi kualitas data

- Akurasi → data sesuai dengan fakta yang ada

Kendala: data yang didapat bukan merupakan data asli responden

- Kelengkapan → setiap *field* data diisi

Kendala: ada *field* yang tidak dianggap penting saat pemasukan data

- Konsistensi → keseragaman format data

Kendala: penamaan tidak seragam

- Ketepatan waktu → waktu input data sesuai dengan waktu data didapatkan

Kendala: ada data yang terlambat diinput

- Dapat dipercaya → bagaimana data dapat dipercaya oleh pengguna

- Dapat diinterpretasikan → seberapa mudah data dapat dipahami

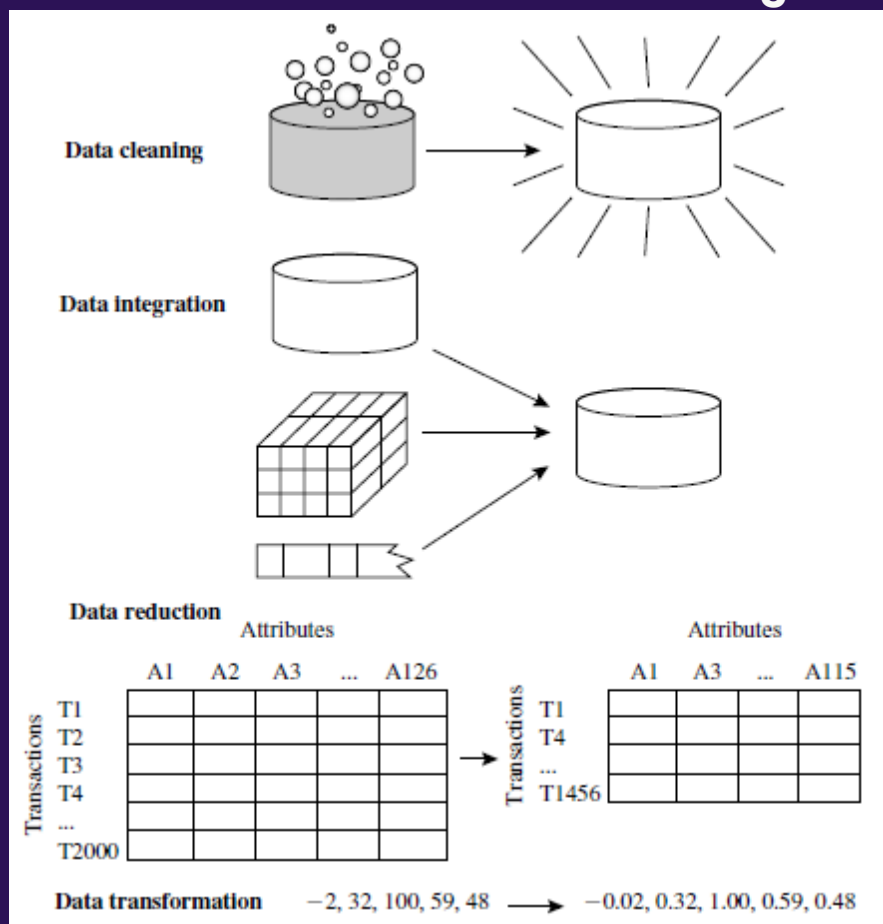
# Tujuan Praproses Data

- Meningkatkan kualitas data yang digunakan
- Meningkatkan hasil *data mining*
- Meningkatkan efisiensi dan kemudahan proses *data mining*

# Tugas dalam Praproses Data

- *Data cleaning* → membersihkan data dari nilai kosong, derau (*noise*), pencilan
- *Data integration* → menggabungkan banyak *database*, *file*, dll menjadi 1 sumber
- *Data reduction* → mengurangi ukuran data
- *Data transformation* → mengubah bentuk data yang ada menjadi bentuk yang lain

# Tugas dalam Praproses Data



# ***Data Cleaning***

- Membersihkan data-data yang mengandung:
  - Nilai kosong (*missing values*) → atribut objek yang tidak memiliki nilai
  - Derau (*noise*) → *error*/variasi acak pada variabel terukur; nilai atribut yang tidak bermakna
  - Pencilan (*outlier*) → nilai atribut yang terlalu besar/kecil & sangat berbeda dengan nilai atribut yang lain
  - Inkonsistensi → ada label kategori yang tidak sesuai

## Data Cleaning – Missing Values

No	Nama	JK	Pendidikan	Lama Bekerja	Grade	Tunjangan
1	Ani	P	S2		Supervisor	Ya
2	Budi	L	S1	5	Manager	Ya
3	Cipta	L	S2	2	Supervisor	Ya
4	Doddy		SMA	3	Staf	Tidak
5	Endah	P	SMA	2	Staf	Tidak
6	Farid	L		3	Staf	Tidak
7	Ginanjari	L	S1	4	Supervisor	Ya
8	Hanum	P	S1	2		Ya
9	Intan	P	S2	3	Supervisor	Ya
10	Joko	L	S1		Staf	Tidak



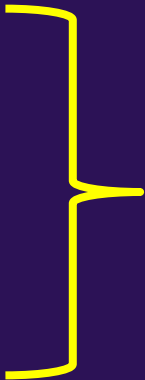
*Missing values*



## Data Cleaning – Missing Values

Menangani *missing values*:

- Menghapus baris/*tuple* data
- Isi *missing value* secara manual → cukup melelahkan
- Isi *missing value* menggunakan:
  - ukuran pemusatan data (rata-rata/median)
  - rata-rata/median atribut dari sampel dengan label kelas yang sama dengan baris/*tuple* yang mengandung *missing value*
  - nilai yang paling mungkin, yang ditentukan dari regresi atau induksi *decision tree*
- Mendesain *database* & prosedur input data dengan sebaik-baiknya, untuk meminimalkan kemungkinan munculnya *missing values*



Dapat membuat data menjadi *bias*

Data Cleaning – Missing Values

No	Nama	JK	Pendidikan	Lama Bekerja	Grade	Tunjangan
1	Ani	P	S2	1	Supervisor	Ya
2	Budi	L	S1	5	Manager	Ya
3	Cipta	L	S2	2	Supervisor	Ya
4	Doddy	2	SMA	3	Staf	Tidak
5	Endah	P	SMA	2	Staf	Tidak
6	Farid	L	3	3	Staf	Tidak
7	Ginanmar	L	S1	4	Supervisor	Ya
8	Hanum	P	S1	2	4	Ya
9	Intan	P	S2	3	Supervisor	Ya
10	Joko	L	S1	5	Staf	Tidak

1 => "3"

2 => "L"

3 => "S1"

4 => "Staf"

5 => "3"

Menangani *noise: data smoothing* melalui *binning*

- Mengurutkan data, kemudian membagi data tsb ke dalam sejumlah *bin*/ember, kemudian dihaluskan berdasarkan:
  - Rata-rata *bin*
  - Median *bin*
  - Nilai batas minimum & maksimum *bin*

Kelebihan *data smoothing*:

- Dapat digunakan untuk memahami tren pada data
- Membantu dalam mendapatkan hasil yang akurat dari data

Kekurangan *data smoothing*:

- Tidak selalu memberikan penjelasan yang jelas mengenai pola pada data
- Ada kemungkinan mengabaikan suatu titik data

Data: 8, 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34

Setelah diurutkan: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

Pembagian data menggunakan *bin* dengan frekuensi sama:

- *Bin* 1: 8, 9, 15, 16
- *Bin* 2: 21, 21, 24, 26
- *Bin* 3: 27, 30, 30, 34

*Data smoothing* menggunakan rata-rata *bin*:

- *Bin* 1: rata-rata =  $\frac{8+9+15+16}{4} = 12 \rightarrow$  data *bin* 1: 12, 12, 12, 12
- *Bin* 2: rata-rata =  $\frac{21+21+24+26}{4} = 23 \rightarrow$  data *bin* 2: 23, 23, 23, 23
- *Bin* 3: rata-rata =  $\frac{27+30+30+34}{4} = 30 \rightarrow$  data *bin* 3: 30, 30, 30, 30

Data: 8, 16, 9, 15, 21, 21, 24, 30, 26, 27, 30, 34

Setelah diurutkan: 8, 9, 15, 16, 21, 21, 24, 26, 27, 30, 30, 34

Pembagian data menggunakan *bin* dengan frekuensi sama:

- *Bin 1*: 8, 9, 15, 16
- *Bin 2*: 21, 21, 24, 26
- *Bin 3*: 27, 30, 30, 34

*Data smoothing* menggunakan batas minimum & maksimum *bin*:

- *Bin 1*: minimum = 8, maksimum = 16 → data *bin 1*: 8, 8, 16, 16
- *Bin 2*: minimum = 21, maksimum = 26 → data *bin 2*: 21, 21, 26, 26
- *Bin 3*: minimum = 27, maksimum = 34 → data *bin 3*: 27, 27, 27, 34

# ***Data Integration***

- Menggabungkan data dari berbagai sumber
- Harus memperhatikan struktur data yang digunakan
- Masalah yang mungkin muncul:
  - Masalah identifikasi entitas → ada *field* customer\_id & cust\_number
  - Perbedaan tipe data yang digunakan
  - Redundansi data → dapat dideteksi menggunakan analisis korelasi
- Hasil: *data warehouse*

Data Integration

Nama	Pekerjaan	Lokasi Rumah	Gender	Kartu	Rumah	Menikah	Pulsa (Ribuan)	Internet (Ribuan)	Jumlah Anak	Kategori Pelanggan
Andi	Analisis	A	Pria	Prabayar	Kontrak	Tidak	100	150	0	Silver
Budi	Dokter	A	Pria	Pascabayar	Pribadi	Ya	500	300	2	Platinum
Citra	Guru	B	Wanita	Prabayar	Kontrak	Tidak	100	100	0	Silver
Dedi	Analisis	A	Pria	Prabayar	Kontrak	Ya	150	200	3	Gold
Evan	Dokter	C	Pria	Pascabayar	Pribadi	Ya	700	400	4	Platinum

+

Nama	Pekerjaan	Alamat	Jenis Kelamin	Prabayar	Kontrak	Menikah	Pulsa (Ribuan)	Internet (Ribuan)	Jumlah Anak	Kelompok
Feni	Dokter	2	W	0	0	1	600	380	1	1
Gito	Guru	1	P	1	1	0	100	70	0	3
Hani	Analisis	3	W	1	1	0	200	250	0	2
Jodi	Dokter	1	P	0	0	1	450	270	2	1

Data Integration

=

Nama	Pekerjaan	Lokasi Rumah	Gender	Kartu	Rumah	Menikah	Pulsa (Ribu)	Internet (Ribu)	Jumlah Anak	Kategori Pelanggan
Andi	Analisis	A	Pria	Prabayar	Kontrak	Tidak	100	150	0	Silver
Budi	Dokter	A	Pria	Pascabayar	Pribadi	Ya	500	300	2	Platinum
Citra	Guru	B	Wanita	Prabayar	Kontrak	Tidak	100	100	0	Silver
Dedi	Analisis	A	Pria	Prabayar	Kontrak	Ya	150	200	3	Gold
Evan	Dokter	C	Pria	Pascabayar	Pribadi	Ya	700	400	4	Platinum
Feni	Dokter	B	Wanita	Pascabayar	Pribadi	Ya	600	380	1	Platinum
Gito	Guru	A	Pria	Prabayar	Kontrak	Tidak	100	70	0	Silver
Hani	Analisis	C	Wanita	Prabayar	Kontrak	Tidak	200	250	0	Gold
Jodi	Dokter	A	Pria	Pascabayar	Pribadi	Ya	450	270	2	Platinum



# ***Data Reduction***

- Mengurangi dimensi dari data
- Tujuan: mendapatkan representasi data dengan ukuran yang lebih kecil, tetapi masih mempertahankan keutuhan data aslinya
- Strategi:
  - Reduksi dimensional
  - Reduksi jumlah data
  - Kompresi data

### Reduksi dimensional

- Proses mengurangi jumlah variabel acak atau atribut yang dipertimbangkan
- Contoh: transformasi wavelet, PCA, seleksi atribut

### Reduksi jumlah data

- Mengganti data asli dengan representasi data yang berukuran lebih kecil
- Bisa berupa teknik parametrik atau nonparametrik
- Contoh:
  - Parametrik: regresi, model log-linear
  - Nonparametrik: histogram, clustering, *sampling*, *data cube aggregation*

## Kompresi data

- Data dipadatkan/dikompres
- Jika data asli bisa direkonstruksi dari data terkompres tanpa kehilangan informasi, maka disebut metode kompresi *lossless*
- Jika hanya bisa didapatkan data perkiraan dari rekonstruksi data terkompres, maka disebut metode kompresi *lossy*

## Data Reduction – Reduksi Dimensional

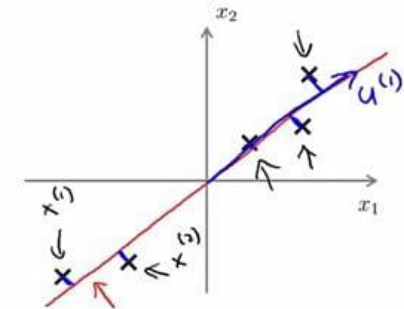
### Principal Component Analysis:

- Mencari  $k$  vektor orthogonal berdimensi  $n$  yang dapat merepresentasikan data dengan baik, dengan  $k \leq n$   
=> Menggabungkan semua atribut yang tersedia menjadi atribut baru
- Langkah-langkah:
  1. Normalisasi data input yang digunakan
  2. Hitung vektor orthonormal yang menjadi basis untuk data input ternormalisasi → *principal component*
  3. Urutkan *principal component* berdasarkan signifikansinya, dari yang paling signifikan sampai yang paling tidak signifikan
  4. Hapus beberapa *principal component* yang paling tidak signifikan

# Data Reduction – Reduksi Dimensional

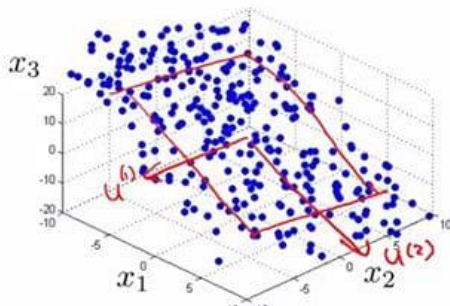
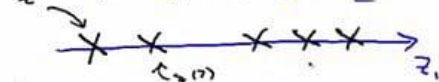
## Principal Component Analysis:

### Principal Component Analysis (PCA) algorithm



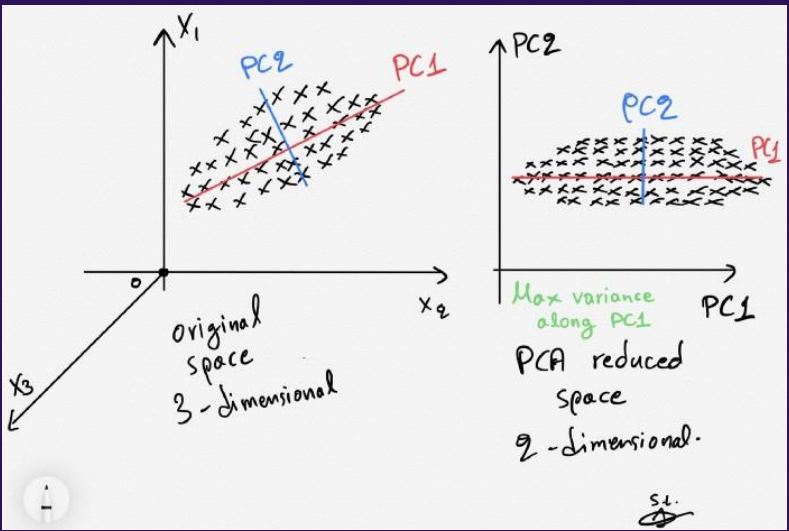
Reduce data from 2D to 1D

$x^{(i)} \in \mathbb{R}^2 \rightarrow z^{(i)} \in \mathbb{R}$



Reduce data from 3D to 2D

Sumber:  
<https://dezyre.gumlet.io/files.dezyre.com/images/Tutorials/Principal+Component+Analysis.jpg>



Sumber:  
[https://miro.medium.com/max/700/1\\*ba0XpZtJrgh7UpzWclgZ1Q.jpeg](https://miro.medium.com/max/700/1*ba0XpZtJrgh7UpzWclgZ1Q.jpeg)

## **Data Reduction – Reduksi Dimensional**

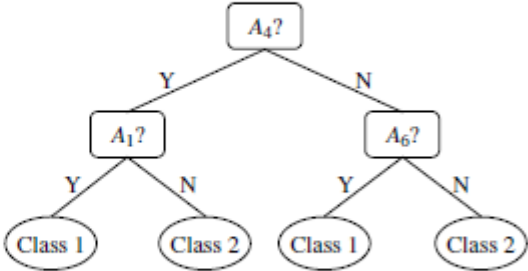
### Seleksi *subset* atribut:

- Menghapus atribut yang tidak relevan/redundan
- Untuk  $n$  atribut, terdapat  $2^n$  *subset* yang mungkin
  - Pencarian manual dapat melelahkan
  - Penentuan *subset* atribut dapat dilakukan menggunakan metode heuristik
- Tujuan: menemukan himpunan atribut minimum sehingga distribusi peluang data kelas yang dihasilkan dapat sedekat mungkin dengan distribusi yang didapatkan menggunakan data asli
- Atribut “terbaik” & “terburuk” ditentukan menggunakan uji signifikansi statistik, dengan asumsi setiap atribut saling bebas satu dengan lainnya

## Data Reduction – Reduksi Dimensional

### Seleksi *subset* atribut:

- Metode heuristik untuk seleksi *subset* atribut:
  - *Stepwise forward selection*
  - *Stepwise backward selection*
  - Kombinasi *stepwise forward selection* & *stepwise backward selection*
  - Induksi *decision tree*

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$   <pre> graph TD     A4["A4?"] -- Y --&gt; A1["A1?"]     A4 -- N --&gt; A6["A6?"]     A1 -- Y --&gt; C1_1((Class 1))     A1 -- N --&gt; C2_1((Class 2))     A6 -- Y --&gt; C1_2((Class 1))     A6 -- N --&gt; C2_2((Class 2))           </pre> $\Rightarrow$ Reduced attribute set: $\{A_1, A_4, A_6\}$

## Data Reduction – Reduksi Dimensional

Nama	Pekerjaan	Lokasi Rumah	Gender	Kartu	Rumah	Menikah	Pulsa (Ribuan)	Internet (Ribuan)	Jumlah Anak	Kategori Pelanggan
Andi	Analisis	A	Pria	Prabayar	Kontrak	Tidak	100	150	0	Silver
Budi	Dokter	A	Pria	Pascabayar	Pribadi	Ya	500	300	2	Platinum
Citra	Guru	B	Wanita	Prabayar	Kontrak	Tidak	100	100	0	Silver
Dedi	Analisis	A	Pria	Prabayar	Kontrak	Ya	150	200	3	Gold
Evan	Dokter	C	Pria	Pascabayar	Pribadi	Ya	700	400	4	Platinum
Feni	Dokter	B	Wanita	Pascabayar	Pribadi	Ya	600	380	1	Platinum
Gito	Guru	A	Pria	Prabayar	Kontrak	Tidak	100	70	0	Silver
Hani	Analisis	C	Wanita	Prabayar	Kontrak	Tidak	200	250	0	Gold
Jodi	Dokter	A	Pria	Pascabayar	Pribadi	Ya	450	270	2	Platinum

➔ Atribut Nama dapat dihapus, karena nama melekat pada objek karyawan, & tidak memiliki pola tertentu yang dapat digunakan untuk menentukan kategori pelanggan, seperti atribut lainnya



## Recap

- Kualitas data
- Tugas dalam praproses data
- *Data cleaning*
- *Data integration*
- *Data reduction* (PCA, seleksi atribut)

Next: bagaimana cara kerja metode praproses diskretisasi?

つづく