DATA MINING

3 SKS | Semester 6 | S1 Sistem Informasi

Pertemuan 3

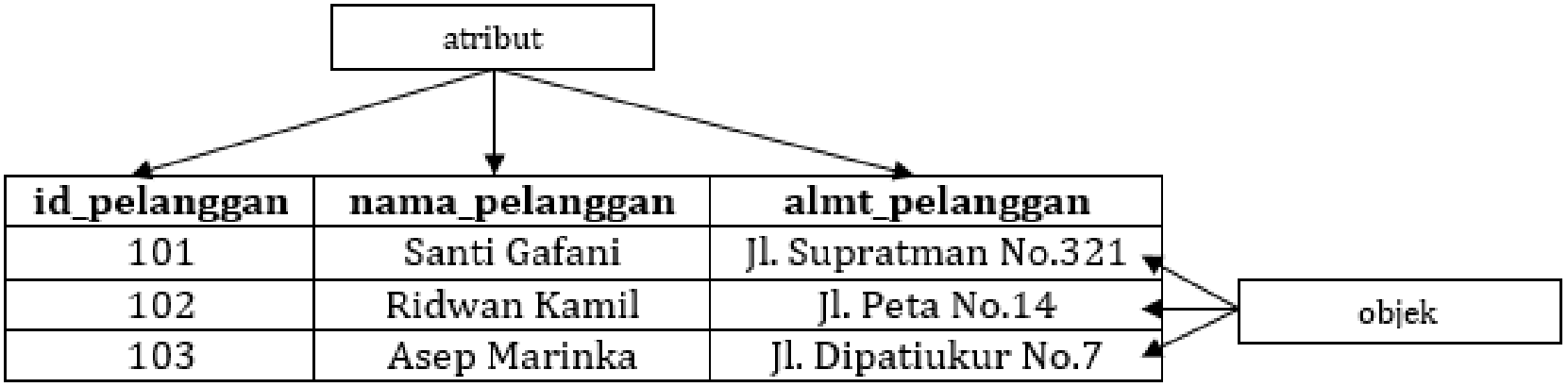
Nizar Rabbi Radliya
nizar.radliya@yahoo.com

Definisi Set Data

Set Data / *Data Set* / Himpunan Data → Kumpulan objek dan atributnya.

Objek = record, point, vector, pattern, event, observation, case, sample, instance, entitas.

Atribut = variabel, field, fitur, atau dimensi.



Tipe Data

Empat sifat yang dimiliki atribut secara umum, yaitu:

1. Pembeda (distinctness): = dan ≠
2. Urutan (order): <, >, ≤, ≥
3. Penjumlahan, Pengurangan (addition): + dan –
4. Perkalian, Pembagian (multiplication): * dan /

Tipe Data

Tipe Atribut		Penjelasan	Contoh
Kategoris (Kualitatif)	Nominal	Nilai atribut berupa nominal memberikan nilai berupa nama. Dengan nama inilah sebuah atribut membedakan dirinya pada data yang satu dengan yang lain (=, ≠).	Kode Pos, NIM, Jenis Kelamin.
	Ordinal	Nilai atribut bertipe ordinal mempunyai nilai berupa nama yang mempunyai arti informasi terurut (<, >, ≤, ≥).	Indek Nilai (A, B, C, D, E)
Numerik (Kuantitatif)	Interval	Nilai atribut dimana perbedaan diantara dua nilai mempunyai makna yang berarti (+, -).	Tanggal
	Rasio	Nilai atribut dimana perbedaan diantara dua nilai dan rasio dua nilai mempunyai makna yang berarti (*, /)	Panjang, berat, tinggi

Tipe Data

Sementara berdasarkan jumlah nilainya, atribut dapat dibedakan menjadi:

1. Diskret

Mempunyai nilai dalam himpunan jumlah yang terbatas atau domainnya terbatas. Contoh: indek nilai (A, B, C, D, E), jenis kelamin (pria, wanita), benar/salah, ya/tidak, 0/1.

2. Kontinu

Mempunyai jangkauan nilai real. Biasanya menggunakan representasi floating point (desimal).

Contoh: panjang, tinggi, berat.

Karakteristik Set Data

1. **Dimensionalitas (dimensionality)**
2. **Sparsitas (sparsity)**
3. **Resolusi (resolution)**

Karakteristik Set Data

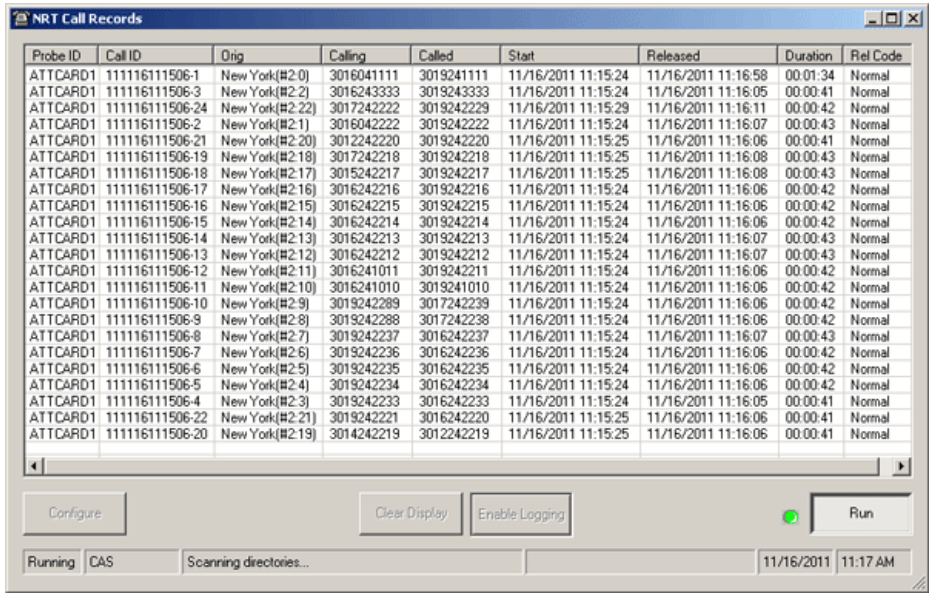
1. **Dimensionalitas (dimensionality)**

a) Dimensionalitas → jumlah atribut yang dimiliki oleh objek-objek dalam data set.

b) Tinggi rendahnya dimensi menentukan perbedaan secara kualitatif.

c) Curse of dimensionality.

d) Pengurangan dimensi (*dimensionality reduction*)



Karakteristik Set Data

2. Sparsitas (sparsity)

a) Untuk set data dengan fitur asimetrik banyak atribut data mempunyai nilai 0 di dalamnya; dalam banyak kasus, kurang dari 1% mempunyain nilai bukan 0.

b) Komputasi menjadi lebih ringan (cepat) dan kapasitas penyimpanan juga lebih sedikit.

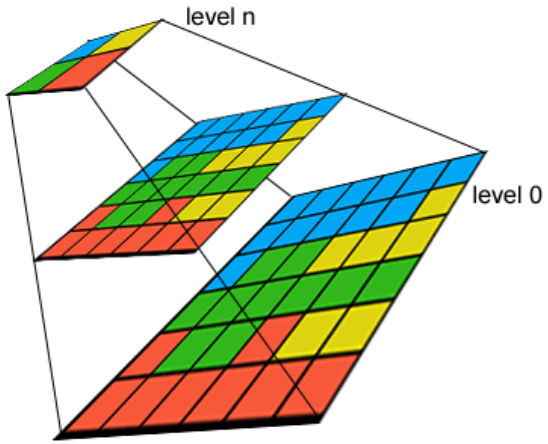
Karakteristik Set Data

3. Resolusi (resolution)

a) Data yang digambarkan dalam bentuk grafik (memerlukan koordinat spasial) karakteristik resolusi yang digunakan juga akan berpengaruh.

b) Pola dalam data bergantung pada level resolusi.

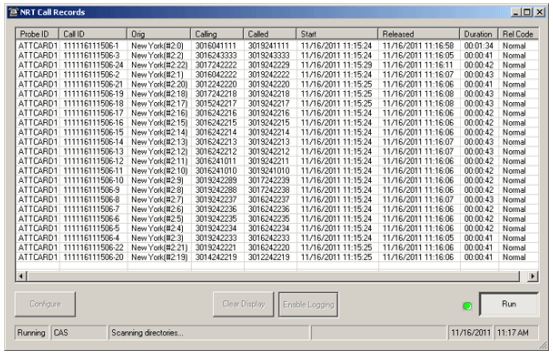
c) Jika resolusi terlalu baik (tidak ada perbedaan/halus), pola mungkin tidak akan kelihatan, jika resolusi terlalu kasar atau sempit, pola juga akan hilang.

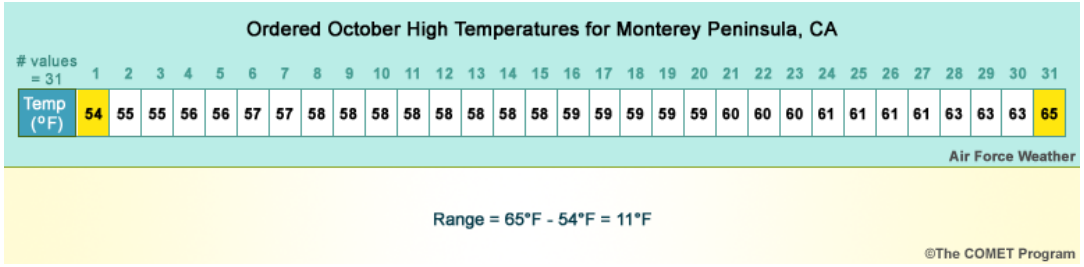


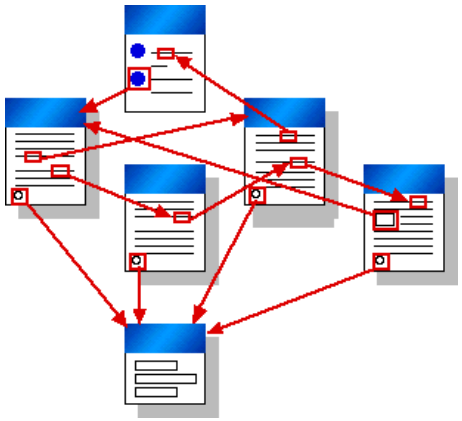
3 levels of a pyramid

Jenis Data

1. **data record, : (data matrik (matrix data), data keranjang belanja (market basket data), dan data dokumen),**
2. data berbasis grafik (graph data),
3. dan data terurut (ordered data),
4. dll







Jenis Data

1. Data Matrik

Mempunyai sejumlah atribut (fitur) numerik yang sama.

Sekumpulan data matrik dapat diinterpretasikan sebagai matrik M x N.

Tinggi	Berat	Sepatu	Celana
168	60	38	30
175	85	42	35
170	77	39	39

Jenis Data

2. Data Keranjang Belanja/Transaksi

Setiap recordnya berisi sejumlah item.

Jumlah item untuk sebuah transaksi bisa berbeda dengan transaksi yang lain.

TID	Item
1	Susu, Bedak, Sabun
2	Susu, Mentega
3	Bedak, Gula, Sabun, Roti

Jenis Data

3. Data Dokumen

Setiap dokumen merupakan satu vektor ‘term’.

Tiap term merupakan satu komponen (atribut) dari vektor tersebut.

Nilai dari setiap komponen menyatakan berapa kali kemunculan term tersebut dalam suatu dokumen.

	Sistem	Informasi	Data	Proses
Makalah 1	18	19	15	9
Makalah 2	6	6	4	8
Makalah 3	14	0	2	12

Kualitas Data

1. Kesalahan Pengukuran (Measurement Error)

a. Noise

Merupakan komponen random dari suatu error pengukuran. Noise berkaitan dengan modifikasi dari nilai asli.

b. Bias

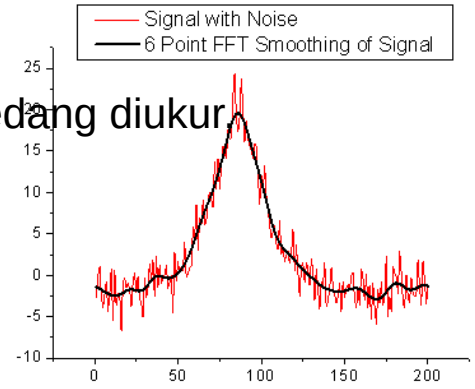
Suatu variasi pengukuran dari kuantitas yang sedang diukur dengan pengurangan antara mean dan nilai kuantitas yang diketahui.

c. Precision

Kedekatan dari pengukuran berulang (dari kuantitas yang sama) satu dengan yang lainnya. Diukur dengan standar deviasi.

d. Accuracy

Kedekatan pengukuran terhadap nilai sebenarnya dari kuantitas yang sedang diukur.



Kualitas Data

Contoh kasus:

Terdapat berat standar laboratorium suatu benda adalah 1 gram dan kita akan menghitung precision dan bias dari skala benda dari hasil pengukuran yang baru. Kita melakukan pengukuran sebanyak lima kali dan memperoleh {1.015, 0.990, 1.013, 1.001, 0.986}.

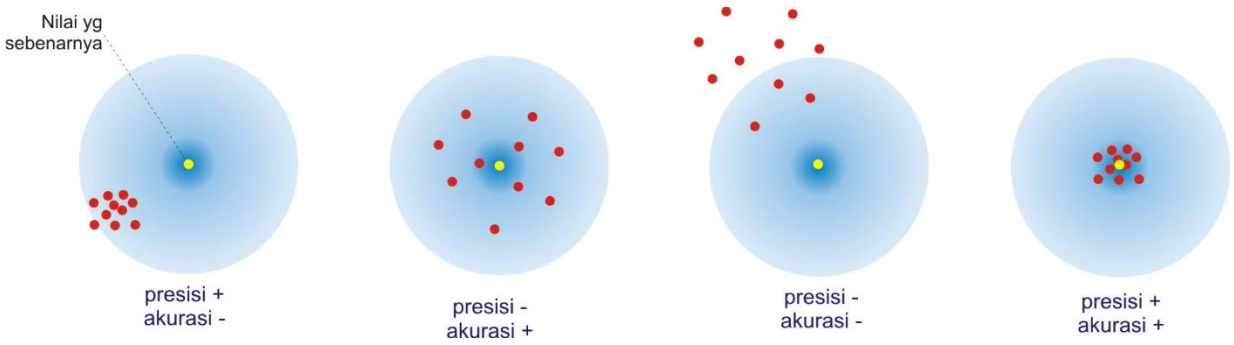
Jawab:

Mean = (1.015 + 0.990 + 1.013 + 1.001 + 0.986) / 5 = 1.001

Bias = 1.001 – 1 = 0.001

Precision=

$$\sqrt{\frac{(1.015 - 1.001)^2 + (0.990 - 1.001)^2 + (1.013 - 1.001)^2 + (1.001 - 1.001)^2 + (0.986 - 1.001)^2}{4}}$$



Kualitas Data

2. Kesalahan Pengumpulan (Collection Error)

Mengacu pada kesalahan-kesalahan (error) seperti hilangnya objek data atau nilai atribut, atau lingkup objek data yang tidak tepat.

a. Outliers
b. Missing value
c. Duplicate data

Kualitas Data

Outliers

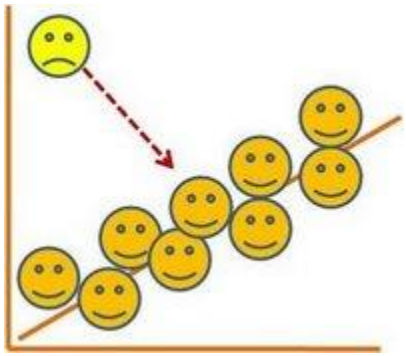
Merupakan objek data dengan sifat yang berbeda sekali dari kebanyakan objek data dalam data-set.

Terdapat beberapa hal yang mempengaruhi munculnya data outlier antara lain:

1) Kesalahan dalam pemasukan data

2) Kesalahan dalam pengambilan sample

3) Memang ada data-data ekstrim yang tidak dapat dihindarkan keberadaannya.



Kualitas Data

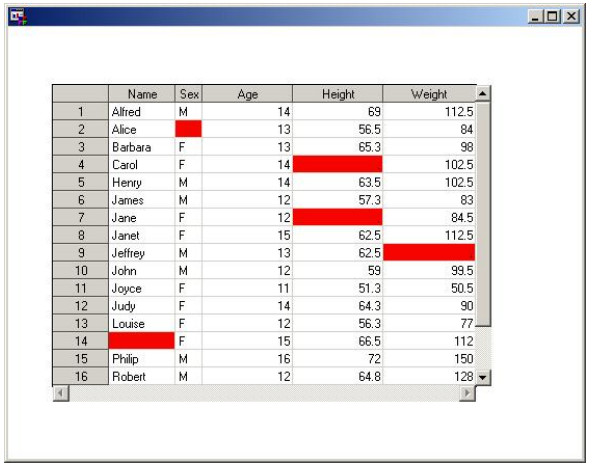
Missing value

Merupakan nilai dari suatu atribut yang tidak ditemukan.

Asalannya terjadinya missing value adalah:

1) Informasi tidak diperoleh

2) Atribut yang mungkin tidak bisa diterapkan ke semua kasus



	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69	112.5
2	Alice		13	56.5	84
3	Barbara	F	13	65.3	98
4	Carol	F	14		102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83
7	Jane	F	12		84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	
10	John	M	12	59	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90
13	Louise	F	12	56.3	77
14		F	15	66.5	112
15	Philip	M	16	72	150
16	Robert	M	12	64.8	128

Penanganan missing values adalah dengan:

1) Mengurangi objek data

2) Memperkirakan missing values

3) Mengabaikan missing values pada saat analisis

4) Mengganti dengan semua nilai yang mungkin (tergantung probabilitasnya)

Kualitas Data

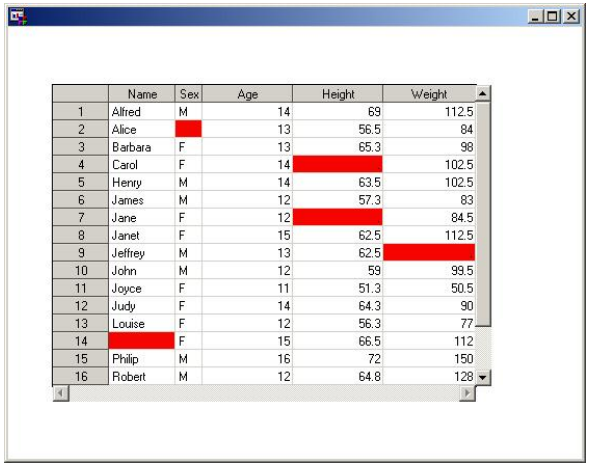
Missing value

Merupakan nilai dari suatu atribut yang tidak ditemukan.

Asalannya terjadinya missing value adalah:

1) Informasi tidak diperoleh

2) Atribut yang mungkin tidak bisa diterapkan ke semua kasus



	Name	Sex	Age	Height	Weight
1	Alfred	M	14	69	112.5
2	Alice		13	56.5	84
3	Barbara	F	13	65.3	98
4	Carol	F	14		102.5
5	Henry	M	14	63.5	102.5
6	James	M	12	57.3	83
7	Jane	F	12		84.5
8	Janet	F	15	62.5	112.5
9	Jeffrey	M	13	62.5	
10	John	M	12	59	99.5
11	Joyce	F	11	51.3	50.5
12	Judy	F	14	64.3	90
13	Louise	F	12	56.3	77
14		F	15	66.5	112
15	Philip	M	16	72	150
16	Robert	M	12	64.8	128

Penanganan missing values adalah dengan:

1) Mengurangi objek data

2) Memperkirakan missing values

3) Mengabaikan missing values pada saat analisis

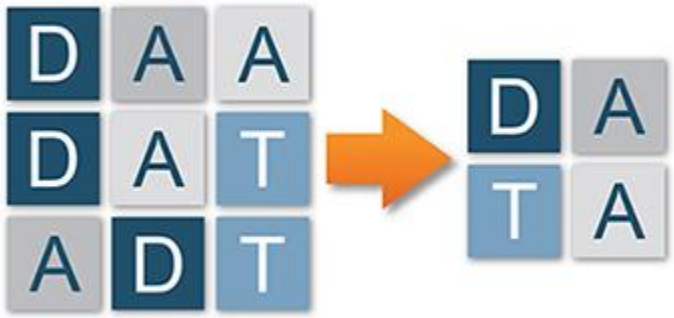
4) Mengganti dengan semua nilai yang mungkin (tergantung probabilitasnya)

Kualitas Data

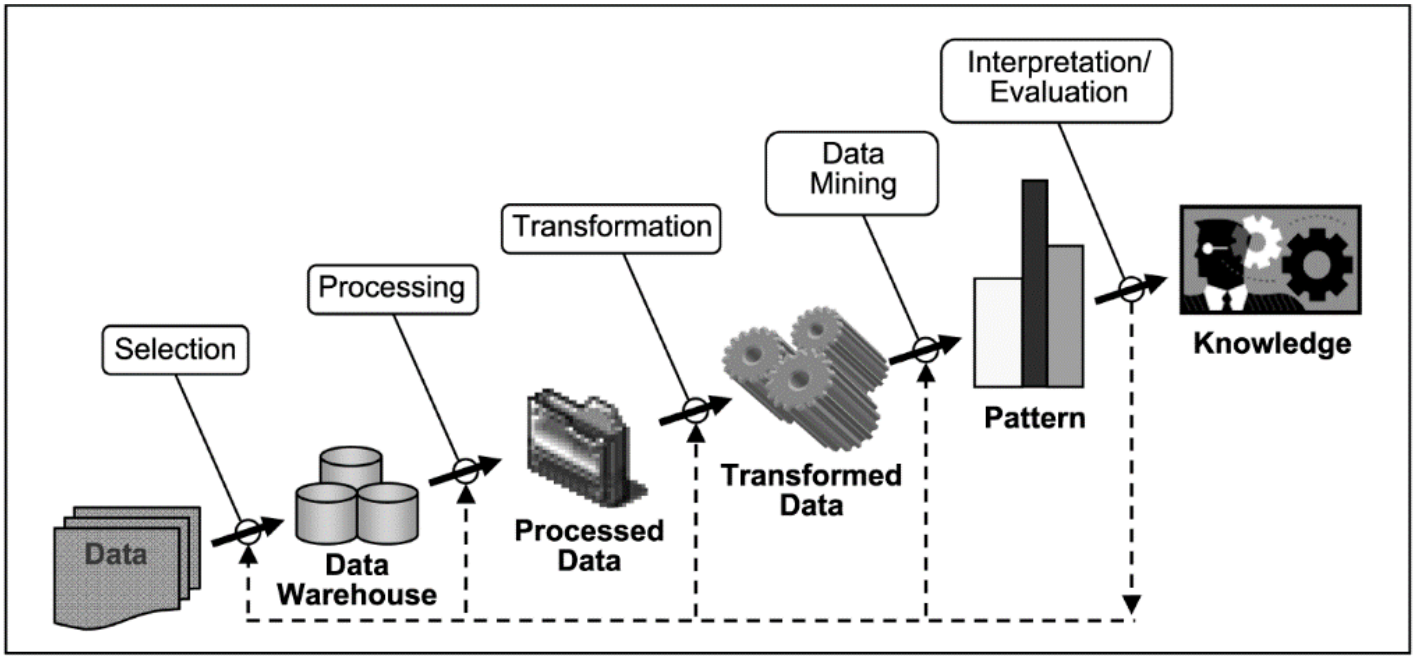
Duplicate data

Set data mungkin terdiri dari objek data yang ganda (duplikat), atau hampir selalu terjadi duplikasi antara satu dengan yang lainnya.

Persoalan utama ketika menggabungkan data dari sumber-sumber yang bervariasi (heterogen).



Pre-Processing



- **Agregasi (aggregation)**
- **Penarikan contoh (sampling)**
- **Diskretisasi dan binerisasi (discretization and binarization)**
- **Pemilihan fitur (feature subset selection)**
- **Transformasi atribut (attribute transformation)**

Agregasi (aggregation)

- ✓ Proses mengkombinasikan dua atau lebih objek ke dalam sebuah objek tunggal;
- ✓ Sangat berguna ketika pada set data ada sejumlah nilai dalam satu fitur yang sebenarnya satu kelompok;
- ✓ Tidak akan menyimpang dari deskripsi fitur tersebut jika nilainya digabungkan.

Agregasi yang dapat dilakukan adalah sum (jumlah), average (rata-rata), min (terkecil), max (terbesar).

Agregasi (aggregation)

Tabel 1. Set Data Transaksi Pembelian Oleh Pelanggan

Cabang	IDT	Tanggal	Total
Bandung	B01001	30-01-2015	250.000
Bandung	B01002	30-01-2015	300.000
Tasikmalaya	T01001	30-01-2015	500.000
Tasikmalaya	T01002	30-01-2015	450.000
Tasikmalaya	T01003	31-01-2015	350.000

Tabel 2. Set Data Transaksi Pembelian Oleh Pelanggan Setelah Agregasi

Cabang	Tanggal	Total
Bandung	30-01-2015	550.000
Tasikmalaya	30-01-2015	950.000
Tasikmalaya	31-01-2015	350.000

Agregasi (aggregation)

Beberapa alasan melakukan agregasi:

- ✓ Set data yang lebih kecil akan membutuhkan memori penyimpanan yang lebih sedikit (pengurangan data atau perubahan skala).
- ✓ Waktu pemrosesan dalam algoritma data mining menjadi lebih cepat.
- ✓ Agregasi bertindak untuk mengubah cara pandang terhadap data dari level rendah menjadi level tinggi.
- ✓ Perilaku pengelompokan objek atau atribut sering kali lebih stabil dari pada objek individu itu sendiri (lebih sedikit variasinya).

Diskretisasi dan binerisasi (discretization and binarization)

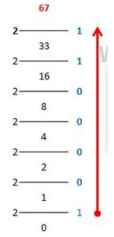
- ✓ Transformasi data dari tipe kontinu dan diskret ke atribut biner disebut binerisasi (binarization).
- ✓ Transformasi data dari atribut kontinu ke atribut kategoris disebut diskretisasi (discretization).

Binerisasi (binarization)

- ✓ M macam nilai kategoris, masing-masing diberikan nilai yang unik dengan nilai integer dalam jangkauan [0,M-1]
- ✓ Jumlah bit yang dibutuhkan untuk binerisasi adalah N=$\lceil log_2(M) \rceil$

Tabel 3. Konversi Atribut Kategoris ke Tiga Atribut Biner

Nilai Kategoris	Nilai Integer	Nilai Biner		
		X1	X2	X3
Rusak	0	0	0	0
Jelek	1	0	0	1
Sedang	2	0	1	0
Bagus	3	0	1	1
Sempurna	4	1	0	0



Diskretisasi (discretization)

- ✓ Pertama, memutuskan berapa jumlah kategori yang harus digunakan.
- ✓ Kedua, menentukan bagaimana memetakan nilai-nilai dari atribut kontinyu ke nilai kategoris.

Contoh nilai yang ada pada tabel 4 diubah menjadi atribut katarogikal dengan nilai: rendah, sedang, tinggi.

Tabel 4. Contoh Atribut Kontinu Yang Akan Didiskretisasi

Atribut Kontinu
125
100
70
120
95
60
220
85
75
90

Pendekatan equal width:

Range data [60 - 220]

Rendah: range [60-113]

Sedang: range [114-167]

Tinggi: range [168-220]

Transformasi atribut (attribute transformation)

- ✓ Sebagain fungsi dari transformasi atribut adalah standarisasi dan normalisasi.
- ✓ Tujuan dari standarisasi dan normalisasi adalah untuk membuat keseluruhan nilai mempunyai suatu sifat khusus.

Transformasi atribut (attribute transformation)

Salah satu contoh transformasi standarisasi adalah dengan:

1. Hitung nilai tengah dengan median;
2. Hidung absolute standard deviation dengan persamaan.

Rumus persamaan yang akan digunakan:

$$\sigma_A = \sum_{i=1}^{m} |x_i - \mu|$$

$$x' = \frac{(x - \mu)}{\sigma_A}$$

Median untuk **jumlah data (*n*) ganjil**

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

Median untuk **jumlah data (*n*) genap**

$$Me = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$$

Transformasi atribut (attribute transformation)

Sebagai contoh lakukan standarisasi dari data set berikut x = {2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2, 1, 1.5, 1.1}. Dari data tersebut dihitung median = μ=(1.9+2)/2=1.95.

Tabel 5. Contoh Standarisasi

x	x-μ	| x-μ |	x′
0.5	-1.45	1.45	-0.24
1.0	-0.95	0.95	-0.16
1.1	-0.85	0.85	-0.14
1.5	-0.45	0.45	-0.08
1.9	-0.05	0.05	-0.01
2.0	0.05	0.05	0.01
2.2	0.25	0.25	0.05
2.3	0.35	0.35	0.06
2.5	0.55	0.55	0.1
3.1	1.15	1.15	0.19
		$\sigma_A = 6.1$	

Transformasi atribut (attribute transformation)

Transformasi atribut menggunakan normalisasi menggunakan pendekatan linear, yang pertama kita terlebih dahulu menghitung rata-rata (persamaan 1) dan varian (persamaan 2) dengan rumus:

$$x_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik} \qquad \text{(persamaan 1)}$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - x_k) \qquad \text{(persamaan 2)}$$

Data hasil normalisasi dapat dihitung menggunakan cara pertama dengan persamaan berikut:

$$x_{ik} = \frac{x_{ik} - x_k}{\sigma_k} \qquad \text{(persamaan 3)}$$

Hasil normalisasi dengan cara persamaan 3 didapatkan fitur yang mempunyai sifat ***zero-mean dan unit variance*.**

Transformasi atribut (attribute transformation)

Sebagai contoh ada data $X = \{x_1, x_2, x_3, x_4, x_5\}^T$, dimana untuk $x_1 = \{0, 2, 1\}$, $x_2 = \{1, 7, 1\}$, $x_3 = \{2, 6, 3\}$, $x_4 = \{5, 1, 4\}$, $x_5 = \{3, 3, 4\}$.

Jangkauan nilai untuk fitur pertama adalah [0,5], fitur kedua [1,7], fitur ketiga [1,4]. Masing-masing fitur memiliki jangkauan yang tidak sama.

Tabel 6. Contoh Data Belum Normal

Fitur 1	Fitur 2	Fitur 3
0	2	1
1	7	1
2	6	3
5	1	4
3	3	4

Transformasi atribut (attribute transformation)

Jika dilakukan normalisasi menggunakan pendekatan linear yang pertama, dihitung terlebih dahulu rata-rata dan standar deviasi. Untuk fitur pertama, didapatkan:

$$x_1 = \frac{1}{5} \times (0 + 1 + 2 + 5 + 3) = 2.2$$

$$\sigma_1^2 = \frac{1}{5-1} \times ((0 - 2.2)^2 + (1 - 2.2)^2 + (2 - 2.2)^2 + (5 - 2.2)^2 + (3 - 2.2)^2) = 3.7$$

$$\sigma_1 = 1.9235$$

$$x_{11} = \frac{0 - 2.2}{1.9235} = -1.1437$$

$$x_{21} = \frac{1 - 2.2}{1.9235} = -0.6239$$

$$x_{31} = \frac{2 - 2.2}{1.9235} = -0.1040$$

$$x_{41} = \frac{5 - 2.2}{1.9235} = -1.4557$$

$$x_{51} = \frac{3 - 2.2}{1.9235} = -0.4159$$

Fitur 1
0
1
2
5
3

Transformasi atribut (attribute transformation)

Tabel 7. Hasil Normalisasi *Zero-Mean* dan *Unit Variance*

Fitur 1	Fitur 2	Fitur 3
-1.1437	-0.6954	-1.0550
-0.6239	1.236	-1.0550
-0.1040	0.8499	0.2638
1.4557	-1.0817	0.9231
0.4159	-0.3091	0.9231

Transformasi atribut (attribute transformation)

Teknik linear yang lain adalah dengan menskalakan jangkauan setiap fitur dalam jangkauan [0,1]:

$$x_{ik} = \frac{x_{ik} - \min(x_k)}{\max(x_k) - \min(x_k)}$$

Tabel 8. Hasil Normalisasi Linear [0,1]

Fitur 1	Fitur 2	Fitur 3
0	0.1667	0
0.2000	1.0000	0
0.4000	0.8333	0.6667
1.0000	0	1.0000
0.6000	0.3333	1.0000

Tabel 6. Contoh Data Belum Normal

Fitur 1	Fitur 2	Fitur 3
0	2	1
1	7	1
2	6	3
5	1	4
3	3	4

Transformasi atribut (attribute transformation)

Teknik linear yang lain adalah dengan menskalakan jangkauan setiap fitur dalam jangkauan [-1,1] :

$$x_{ik} = \frac{2x_{ik} - (\max(x_k) + \min(x_k))}{\max(x_k) - \min(x_k)}$$

Tabel 9. Hasil Normalisasi Linear [-1,1]

Fitur 1	Fitur 2	Fitur 3
-1.0000	-0.6667	-1.0000
-0.6000	1.0000	-1.0000
-0.2000	0.6667	0.3333
1.0000	-1.0000	1.0000
0.2000	-0.3333	1.0000

Tabel 6. Contoh Data Belum Normal

Fitur 1	Fitur 2	Fitur 3
0	2	1
1	7	1
2	6	3
5	1	4
3	3	4

Materi Minggu Ke 4

Similaritas dan Dissimilaritas

1. Similaritas dan dissimilaritas data satu atribut
2. Similaritas dan dissimilaritas data multiatribut