

Chapter 2. Data

- Tipe Data**
- Statika Data**
- Data Quality, Data Cleaning dan Data Integration**
- Data Transformation**
- Reduksi Dimensi**
- Simpulan**

Tipe Data Sets: (1) Record Data

- Relational records
 - Tabel relasional, sangat terstruktur
- Data matrix, Misalnya Matriks numerik, crosstabs

	China	England	France	Japan	USA	Total
Active Outdoors Crochet Glove		12.00	4.00	1.00	240.00	257.00
Active Outdoors Lycra Glove		10.00	6.00		323.00	339.00
InFlux Crochet Glove	3.00	6.00	8.00		132.00	149.00
InFlux Lycra Glove		2.00			143.00	145.00
Triumph Pro Helmet	3.00	1.00	7.00		333.00	344.00
Triumph Vertigo Helmet		3.00	22.00		474.00	499.00
Xtreme Adult Helmet	8.00	8.00	7.00	2.00	251.00	276.00
Xtreme Youth Helmet		1.00			76.00	77.00
Total	14.00	43.00	54.00	3.00	1,972.00	2,086.00

Person:

Pers_ID	Surname	First_Name	City
0	Miller	Paul	London
1	Ortega	Alvaro	Valencia
2	Huber	Urs	Zurich
3	Blanc	Gaston	Paris
4	Bertolini	Fabrizio	Rom

— no relation

Car:

Car_ID	Model	Year	Value	Pers_ID
101	Bentley	1973	100000	0
102	Rolls Royce	1965	330000	0
103	Peugeot	1993	500	3
104	Ferrari	2005	150000	4
105	Renault	1998	2000	3
106	Renault	2001	7000	3
107	Smart	1999	2000	2

- Data transaksi

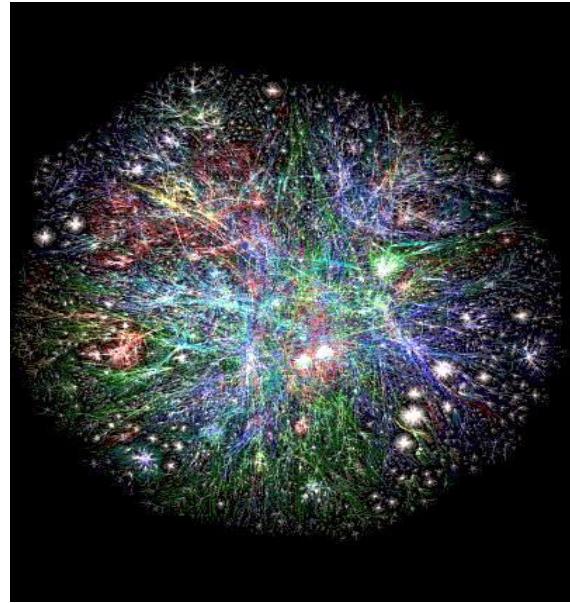
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

	team	coach	pla y	ball	score	game	n wi	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

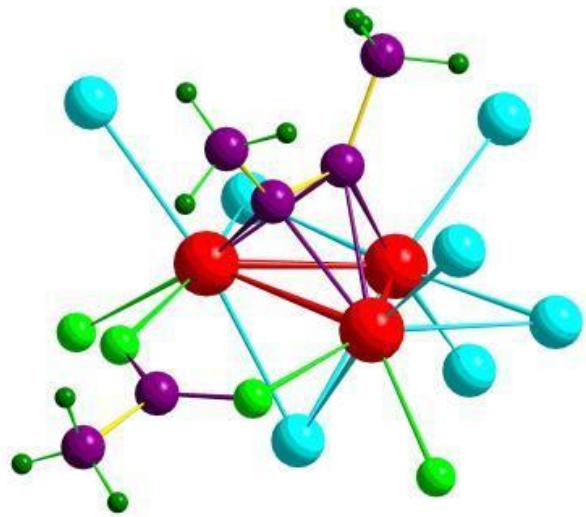
- Data dokumen: matrix dari document teks

Tipe Data Sets: (2) Graphs dan Networks

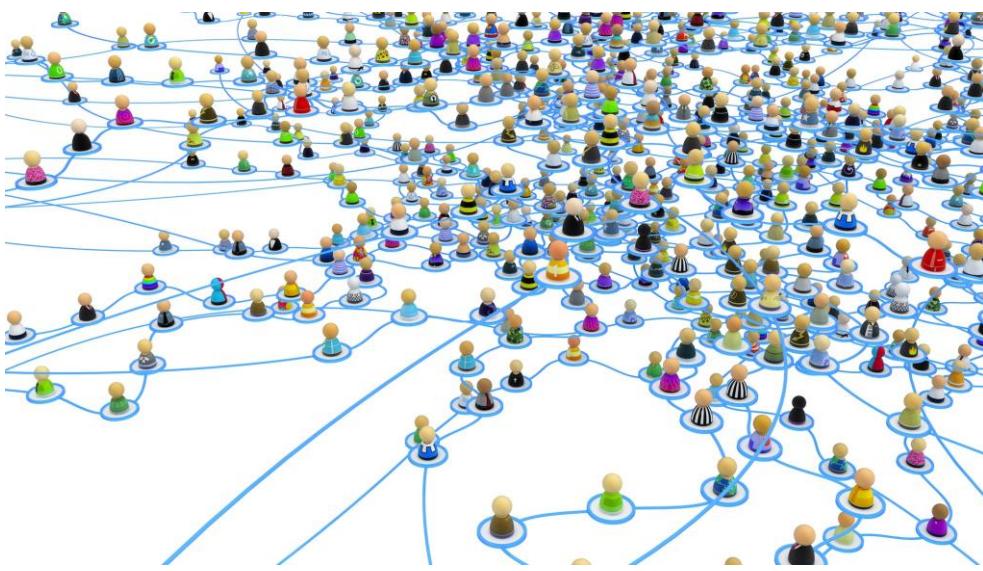
□ Jaringan transportasi



□ World Wide Web



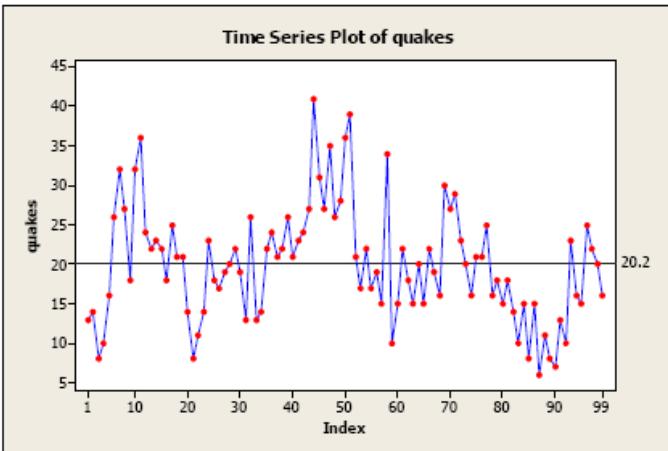
□ Struktur Molekul



□ Jaringan sosial atau informasi

Tipe Data Sets: (3) Ordered Data

- Data video: urutan gambar
- Data temporal: time-series

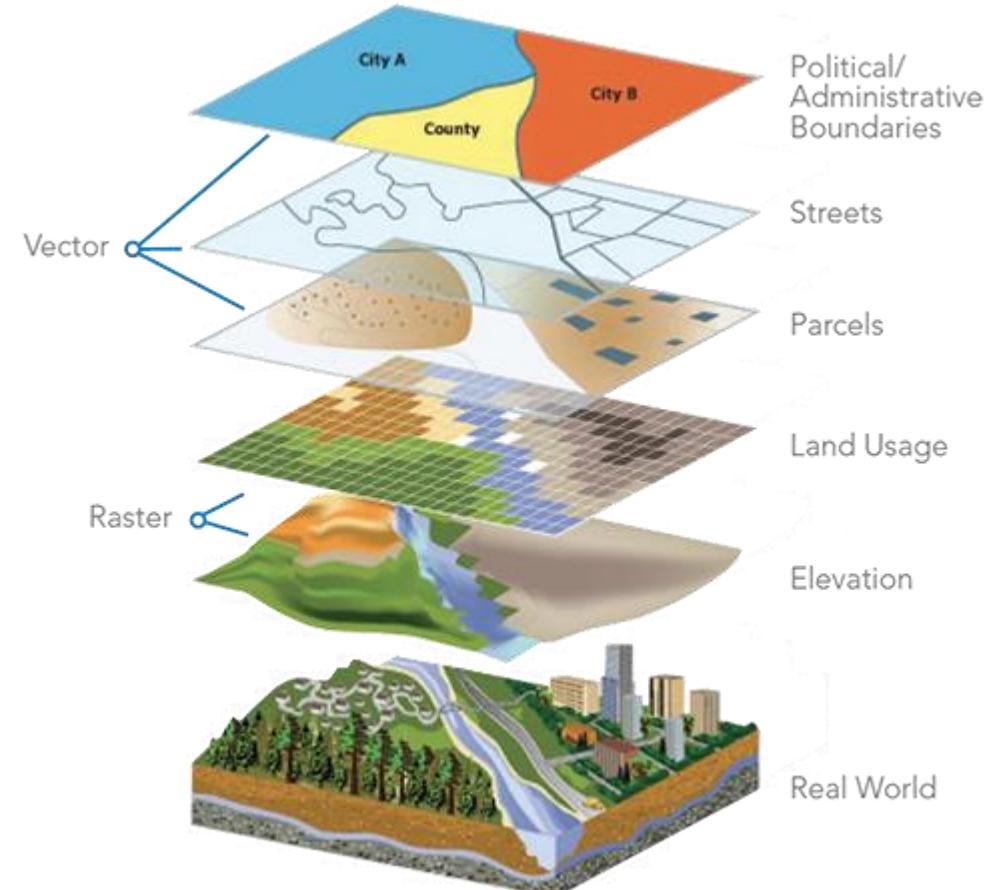


- Sequential Data: urutan transaksi
- Data urutan genetik

Human	GTTTGAGG	-	ATGTTCAACAAATGCTCCTTCATTCCCTATTTACAGACCTGCCGCA
Chimpanzee	GTTTGAGG	-	ATGTTCAATAATGCTGCTTCACTCCCTATTTACAGACCTGCCGCA
Macaque	GTTTGAGG	-	ATGCTCAATAATGCTCCTTCATTCCCTCATTACAACCTGCCGCA
Human	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTAGTAATTGAGTGT	Start	
Chimpanzee	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTAGTAATTGAGTGT		
Macaque	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTAGTAATTGAGTGT		
Human	GATCTGGAGACTAACTCTGAAATAAAAGCTGATTATTTATTTCTCAAAACAA		
Chimpanzee	GATCTGGAGACTAAACTCTGAAATAAAAGCTGATTATTTATTTCTCAAAACAA		
Macaque	TATCTGGAGACTAAACTCTGAAATAAAAGCTGATTATTTATTTCTCAAAACAA		
Human	CAGAACACGATTTAGCAAATTACTCTTAAGATAATTATTTACATTCTATATTCTCTA		
Chimpanzee	CAGAACACGATTTAGCAAATTACTCTTAAGATAACTATTTCACATTCTATATTCTCTA		
Macaque	CAGAACATGATTTAGCAAATTACCTCTTAAGATAATTATTTGCACCTCTATATTCTCTA		
Human	CCCTGAGTTGATGTTGAGCAATATGTCACCTTCATAAAGCCAGGTATACAC-----TTATG		
Chimpanzee	CCCTGAGTTGATGTTGAGCCGATGTCACCTTCATAAAGCCAGGTATACAC-----TTATG		
Macaque	CCCTGAGTTGATGTTGAGCAATATGTCACCTCCACAAAGCCAGGTATATACATTACG		
Human	GACAGGTAAGTAAAAACATATTATTTACGTTTGTCCAAGAATTAAATTTC	H I Y S T F L S K	
Chimpanzee	GACAGGTAAGTAAAAACATATTATTTACGTTTGTCCAAGAATTAAATTTC		
Macaque	GACAGGTAAGTAAAAACATATTATTTACGTTTGTCCAAGAATTAAATTTC		
Human	AACTGTTGCGCGTGTGGTAA---TGTAAAAACAAACTCAGTACA		
Chimpanzee	AACTGTTGCGCGTGTGGTAA---TGTAAAAACAAACTCAGTACA		
Macaque	AACTGTTGCGCGTGTGGTAA---CBTAAAAACAAACTCAGTACA		

Tipe Data Sets: (4) Spatial, image dan multimedia Data

- Data spasial: maps



- Image data:

- Video data:

Karakteristik Penting Data Terstruktur

- ❑ Dimensi
 - ❑ Curse of dimensionality (Kutukan Dimensi) => fenomena yang terjadi ketika jumlah dimensi (atau fitur) dalam dataset meningkat
- ❑ Kelangkaan
 - ❑ Hanya kehadiran yang diperhitungkan => dalam beberapa analisis data, yang dianggap penting adalah apakah suatu fitur ada atau tidak, tanpa memedulikan nilainya
- ❑ Resolusi
 - ❑ Pola tergantung pada skala => pola atau struktur yang ditemukan dalam data dapat bervariasi tergantung pada skala atau tingkat pengamatan yang digunakan.
- ❑ Distribusi
 - ❑ Sentralitas dan dispersi
 - ❑ Sentralitas => nilai-nilai yang mewakili titik pusat atau "rata-rata" dari data.
 - ❑ Dispersi => seberapa bervariasi atau tersebar data di sekitar nilai pusatnya

Objek Data

- ❑ Data sets terdiri dari objek data
- ❑ Objek data mewakili entitas
- ❑ Contoh:
 - ❑ Database penjualan : Penlanggan, Item, Penjualan
 - ❑ Database Medis : pasien, perawatan
 - ❑ Database universitas: mahasiswa, profesor, kursus
- ❑ Juga disebut sampel, *Contoh, instances, data points, objects, tuples*
- ❑ Objek data dijelaskan berdasarkan **atribut**
- ❑ Baris database → Objek data; Kolom → Atribut

Atribut

- **Atribut (atau dimensi, fitur, variabel)**
 - data field, mewakili karakteristik atau fitur objek data.
 - *Misalnya, _ID pelanggan, nama, alamat*
- Types:
 - Nominal (misalnya, merah, biru)
 - Biner (misalnya, {true, false})
 - Ordinal (misalnya, {mahasiswa baru,mahasiswa Tingkat dua, junior, senior})
 - Numerik: kuantitatif
 - Skala interval: 100°C adalah skala interval, tahun 2024
 - Skala rasio: 100 K adalah skala rasio, usia 37
 - Atribut Diskrit vs. Kontinu

Jenis Atribut

- ❑ **Nominal:** kategori, negara, atau "nama benda"
 - ❑ *Warna Rambut= {coklat kemerahan, hitam, pirang, coklat, abu-abu, merah, putih}*
 - ❑ status perkawinan, pekerjaan, nomor ID, kode pos
- ❑ **Binary**
 - ❑ Atribut nominal dengan hanya 2 status (0 dan 1)
 - ❑ Biner simetris: kedua hasil sama pentingnya
 - ❑ misalnya, jenis kelamin
 - ❑ Biner asimetris: hasil tidak sama pentingnya.
 - ❑ misalnya, tes medis (positif vs. negatif)
 - ❑ Konvensi: menetapkan 1 untuk hasil yang paling penting (misalnya, HIV positif)
- ❑ **Ordinal**
 - ❑ Values have a meaningful order (ranking) but magnitude between successive values is not known
 - ❑ *Size = {small, medium, large}*, grades, army rankings

Jenis Atribut Numerik

- ❑ Quantity (integer atau Bilangan Real)
- ❑ **Interval**
 - ❑ Diukur pada skala satuan berukuran sama
 - ❑ Nilai memiliki urutan
 - ❑ Misalnya, suhu dalam C° atau F°, tanggal kalender
 - ❑ Tidak ada titik nol sejati
- ❑ **Ratio**
 - ❑ Titik nol yang melekat
 - ❑ Kita dapat berbicara tentang nilai sebagai urutan besarnya yang lebih besar dari satuan pengukuran (10 K° dua kali lebih tinggi dari 5 K°).
 - ❑ misalnya, suhu dalam Kelvin, panjang, hitungan, jumlah moneter

Atribut Diskrit vs. Kontinu

❑ Atribut Diskrit

- ❑ Hanya memiliki kumpulan nilai terbatas atau tak terbatas yang dapat dihitung
 - ❑ Misalnya, kode pos, atau kumpulan kata dalam kumpulan dokumen
- ❑ Terkadang, direpresentasikan sebagai variabel bilangan bulat dan diperoleh dari perhitungan
- ❑ Catatan: Atribut biner adalah kasus khusus dari atribut diskrit

❑ Atribut Kontinu

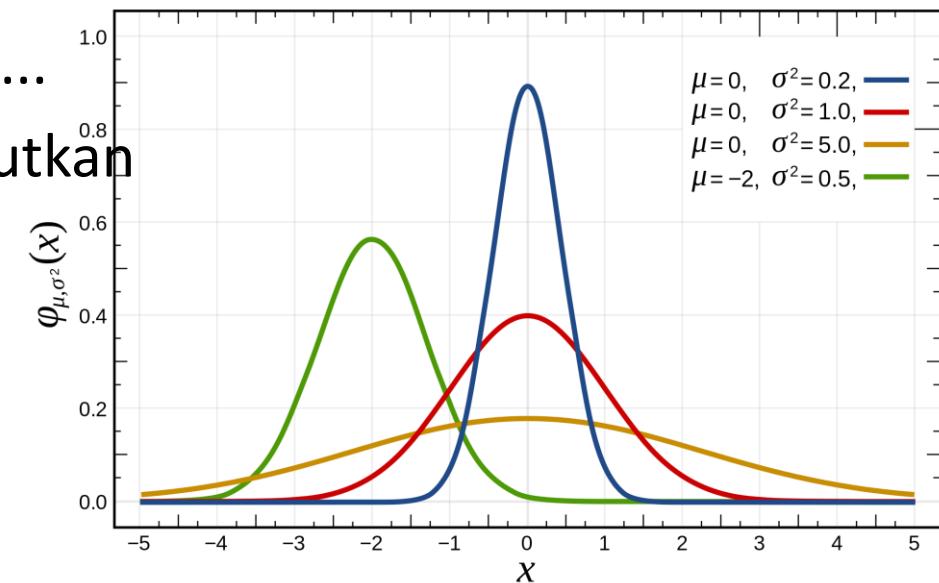
- ❑ Memiliki bilangan real sebagai nilai atribut
 - ❑ Misalnya, suhu, tinggi badan, atau berat badan
- ❑ Praktis, bilangan real biasanya diperoleh dari pengukuran dan direpresentasikan dengan menggunakan jumlah digit yang terbatas
- ❑ Atribut kontinu biasanya direpresentasikan sebagai variabel floating-point

Statika Data

- Mengukur Central Tendency
- Mengukur Penyebaran Data
- Analisis Kovarian dan Korelasi
- Tampilan grafis statis dasar data

Deskripsi Statistik Dasar Data

- Motivasi
 - Untuk lebih memahami data: central tendency, variasi dan penyebaran
- Karakteristik dispersi data
 - Median, max, min, quantiles, outliers, variance, ...
- Dimensi numerik sesuai dengan interval yang diurutkan
 - Dispersi data:
 - Dianalisis dengan beberapa granularitas presisi
 - Boxplot atau analisis kuantil pada interval yang diurutkan
- Analisis dispersi pada ukuran yang dihitung
 - Mengubah ukuran ke dalam dimensi numerik
 - Boxplot atau analisis kuantil pada kubus yang diubah



Mengukur Central Tendency: (1) Mean

□ Mean (Ukuran aljabar) (Sampel vs. Populasi):

Catatan: n adalah ukuran sampel dan N adalah ukuran populasi.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{\sum x}{N}$$

□ Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

□ Trimmed mean:

- Memotong nilai ekstrem(e.g., Olympics gymnastics score computation)

Mengukur Central Tendency: (2) Median

- Median:

- Nilai tengah jika jumlah nilai ganjil, atau rata-rata dari dua nilai tengah
- Estimasi dengan Interpolasi (untuk *grouped data*):

age	frequency
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Perkiraan
median

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum \text{freq})_l}{\text{freq}_{\text{median}}} \right) \text{width}$$

Batas interval rendah

Jumlah sebelum interval median

Lebar interval($L_2 - L_1$)

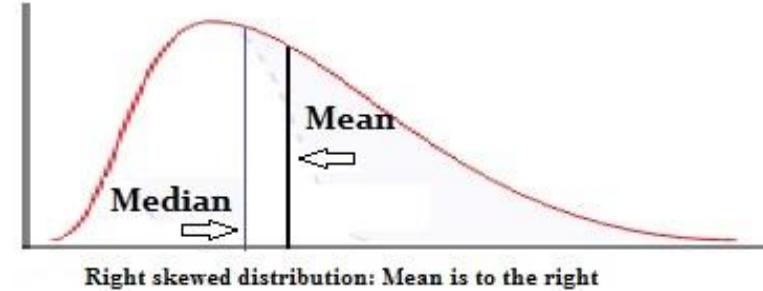
Mengukur Central Tendency: (3) Modus

- Modus: Nilai yang paling sering muncul dalam data

- Unimodal = hanya memiliki satu modus

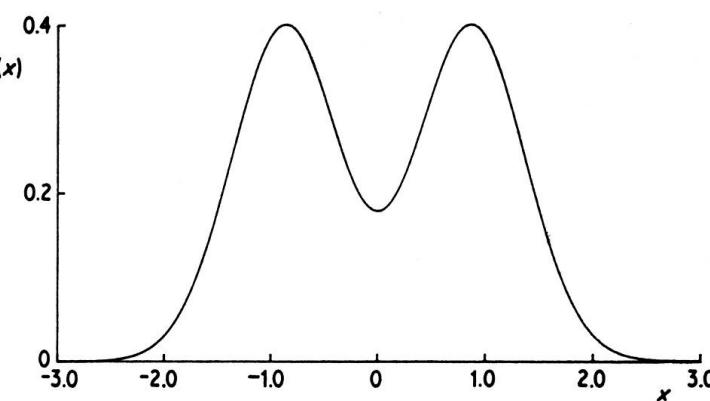
- Empirical formula:

$$\text{mean} - \text{modus} = 3 \times (\text{mean} - \text{median})$$

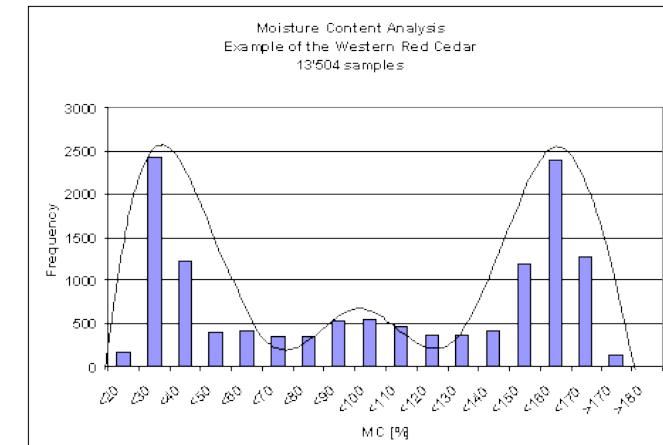
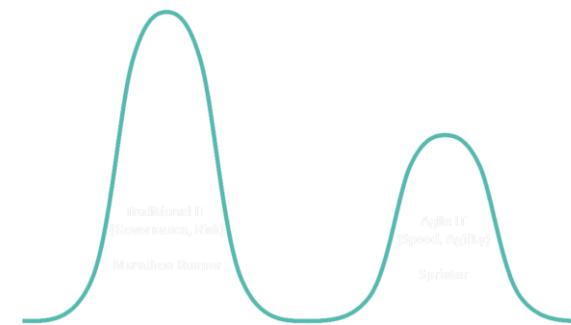


- Multi-modal = memiliki lebih dari satu modus

- Bimodal



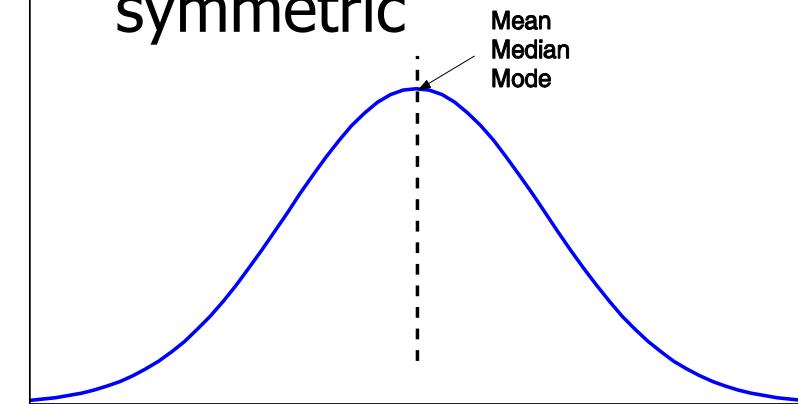
- Trimodal



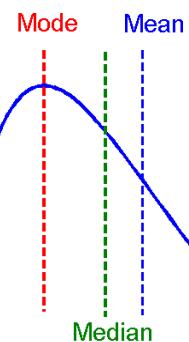
Simetris vs. Data Skewed

- Median, mean and modus simetris, data skewed secara positif dan negatif

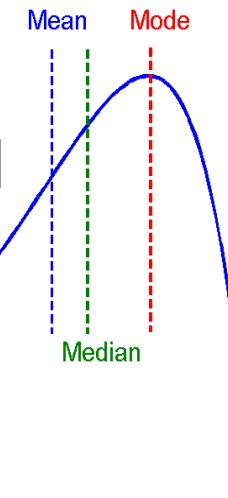
symmetric



positively skewed

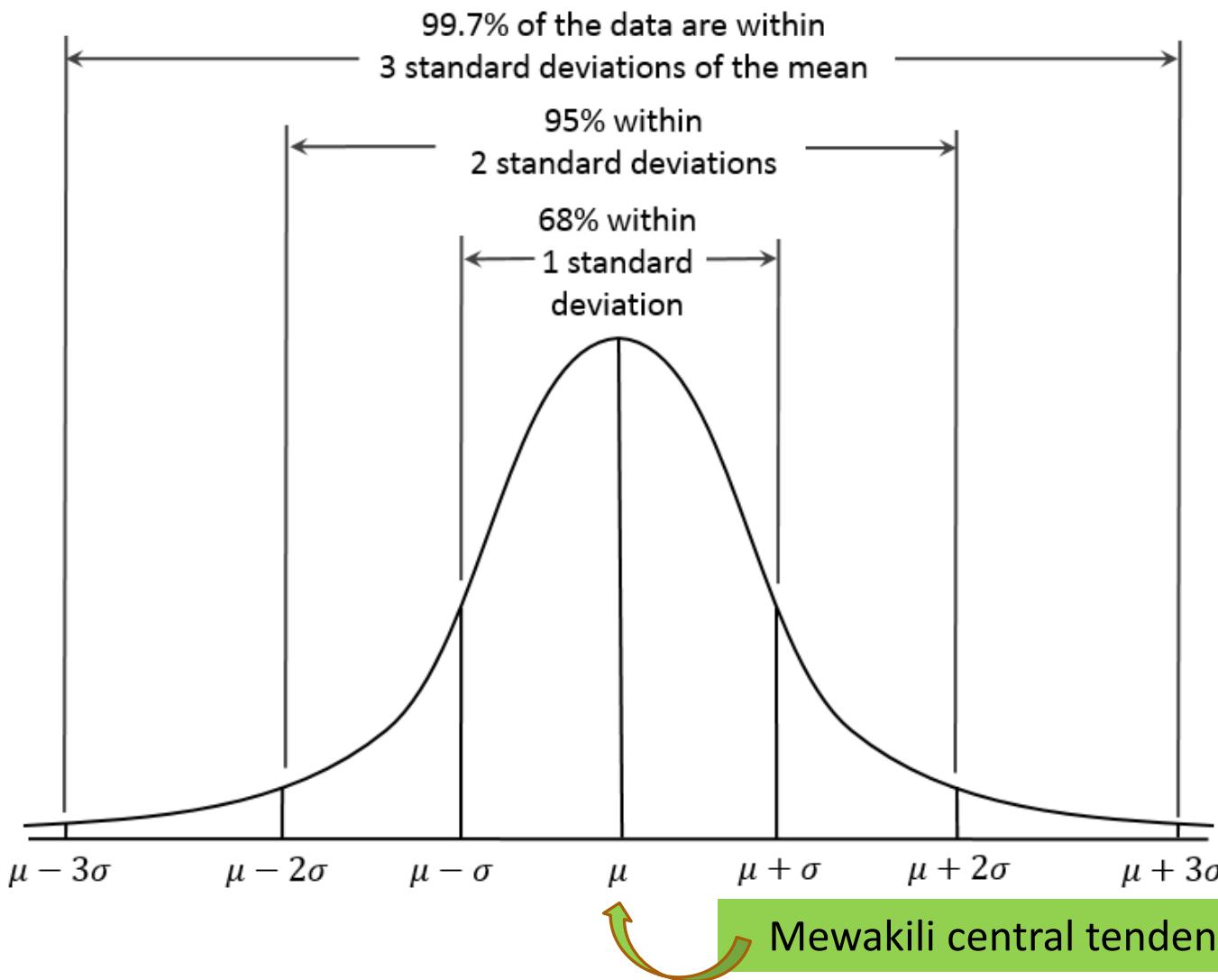


negatively skewed



Sifat Kurva Distribusi Normal

← ————— Mewakili data dispersion, spread ————— →



Mengukur Distribusi Data: Varians dan Standar Deviasi

- Varians dan standar deviasi (sampel: s , populasi: σ)
 - Varians: (aljabar, komputasi yang dapat diskalakan)
 - Q: Bisakah Anda menghitungnya secara bertahap dan efisien?

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2 \right]$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2 = \frac{1}{N} \sum_{i=1}^n x_i^2 - \mu^2$$

Note: Perbedaan rumus untuk sampel vs. populasi

- n : ukuran sampel
- N : ukuran populasi

- Standar deviasi s (atau σ) adalah akar kuadrat dari varians s^2 (atau σ^2)

Analisis Korelasi (untuk Data Kategoris)

- **X² (chi-square) test:**

$$\chi^2 = \sum_i^n \frac{(O_i - E_i)^2}{E_i}$$

observed
↓
expected

- Hipotesis nol: Kedua distribusi bersifat independen
- Sel yang berkontribusi paling besar pada nilai X² adalah sel yang jumlah aktualnya sangat berbeda dari jumlah yang diharapkan
 - Semakin besar nilai X², semakin besar kemungkinan variabel terkait
- Note: Korelasi tidak menyiratkan kausalitas
 - # rumah sakit dan # pencurian mobil di kota berkorelasi
 - Keduanya terkait secara kausal dengan variabel ketiga: populasi

Contoh Perhitungan Chi-Square

	Bermain Catur	Tidak Bermain Catur	Sum (row)
Suka Fiksi Ilmiah	250 (X1)	200 (X2)	450
Tidak Suka Fiksi Ilmiah	50 (X3)	1000 (X4)	1050
Sum(col.)	300	1200	1500

- Hipotesis nol: Kedua distribusi tersebut independen
 - Rasio antara orang yang bermain catur vs tidak bermain catur adalah sama untuk kedua kelompok suka fiksi ilmiah dan tidak suka fiksi ilmiah
 - $X_1:X_2=X_3:X_4=300:1200$
 - $X_1:X_3=X_2:X_4=450:1050$
 - $X_1+X_2=450 \quad X_3+X_4=1050$
 - $X_1+X_3=300 \quad X_2+X_4=1200$

Contoh Perhitungan Chi-Square

	Bermain Catur	Tidak Bermain Catur	Sum (row)
Suka Fiksi Ilmiah	250 (90)	200 (360)	450
Tidak Suka Fiksi Ilmiah	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

Bagaimana cara mendapatkan 90?
450/1500 * 300 = 90

- χ^2 perhitungan (chi-square) (angka dalam tanda kurung adalah hitungan yang diharapkan yang dihitung berdasarkan distribusi data dalam dua kategori)

Kita dapat menolak hipotesis nol dari independence pada tingkat kepercayaan 0.001

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- Ini menunjukkan bahwa suka_fiksi_ilmiah dan bermain_catur berkorelasi dalam kelompok

Contoh Perhitungan Chi-Square

	A	B	C	D	Sum (row)
1					200
0					1000
Sum(col.)	300	300	300	300	1200

- Derajat Kebebasan
 - $(\# \text{categories_in_variable_A} - 1)(\# \text{categories_in_variable_B} - 1)$
 - jumlah nilai yang bebas untuk bervariasi

Contoh Perhitungan Chi-Square

	Bermain Catur	Tidak Bermain Catur	Sum (row)
Suka Fiksi Ilmiah	250 (90)	200 (360)	450
Tidak Suka Fiksi Ilmiah	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

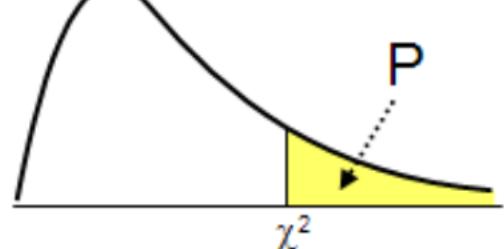
$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

□ Drajat Kebebasan =?



Kita dapat menolak hipotesis nol dari independence pada tingkat kepercayaan 0,001

Values of the Chi-squared distribution



DF	P										
	0.995	0.975	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	0.0000393	0.000982	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.550	10.828
2	0.0100	0.0506	3.219	4.605	5.991	7.378	7.824	9.210	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.860	16.924	18.467
5	0.412	0.831	7.289	9.236	11.070	12.833	13.388	15.086	16.750	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458

Varians untuk variabel tunggal (data numerik)

- Varians variabel acak X memberikan ukuran seberapa besar nilai X menyimpang dari mean atau nilai yang diharapkan dari X :

$$\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = \begin{cases} \sum_x (x - \mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- di mana σ^2 adalah varians X , σ disebut standar deviasi
 - μ adalah mean, dan $\mu = E[X]$ adalah nilai yang diharapkan dari X
- Artinya, varians adalah nilai yang diharapkan dari deviasi kuadrat dari mean
- Ini juga dapat ditulis sebagai: $\sigma^2 = \text{var}(X) = E[(X - \mu)^2] = E[X^2] - \mu^2 = E[X^2] - [E(x)]^2$
- Varians sampel

$$s^2 = \frac{1}{n} \sum_i^n (x_i - \hat{\mu})^2$$

$$s^2 = \frac{1}{n-1} \sum_i^n (x_i - \hat{\mu})^2$$

Kovarian untuk Dua Variabel

- Kovarian antara dua variable X_1 dan X_2

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

dimana $\mu_1 = E[X_1]$ adalah masing-masing mean atau **Nilai yang diharapkan** dari X_1 ;
Demikian pula untuk μ_2

- Kovarian sampel antara X_1 dan X_2 : $\hat{\sigma}_{12} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)$
- Kovarian sampel adalah generalisasi dari varians sampel:

$$\hat{\sigma}_{11} = \frac{1}{n} \sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i1} - \hat{\mu}_1)$$

- **Kovarian positif:** jika $\sigma_{12} > 0$
- **Kovarian negatif:** Jika $\sigma_{12} < 0$

Kovarian untuk Dua Variabel

- ❑ **Independence:** jika X_1 dan X_2 adalah independent, $\sigma_{12} = 0$ tetapi tidak berlaku sebaliknya
 - ❑ Beberapa pasangan variabel acak mungkin memiliki kovarian 0 tetapi tidak independen
 - ❑ Hanya di bawah beberapa asumsi tambahan (misalnya, data mengikuti distribusi normal multivariat) kovarian 0 menyiratkan independensi
- ❑ Example:

X_1	1	-1
X_2	0	1

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

$$E(X_1) = ?$$

$$E(X_2) = ?$$

$$E(X_1 X_2) = ?$$

Contoh: Perhitungan Kovarian

- Misalkan dua saham X₁ dan X₂ memiliki nilai berikut dalam satu minggu:
(2, 5), (3, 8), (5, 10), (4, 11), (6, 14)
- Pertanyaan: Jika saham dipengaruhi oleh tren industri yang sama, apakah harganya naik atau turun bersama?
- Rumus kovarians

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1 X_2] - \mu_1 \mu_2 = E[X_1 X_2] - E[X_1]E[X_2]$$

- Perhitungannya dapat disederhanakan sebagai: $\sigma_{12} = E[X_1 X_2] - E[X_1]E[X_2]$
 - $E(X_1) = (2 + 3 + 5 + 4 + 6)/5 = 20/5 = 4$
 - $E(X_2) = (5 + 8 + 10 + 11 + 14)/5 = 48/5 = 9.6$
 - $\sigma_{12} = (2 \times 5 + 3 \times 8 + 5 \times 10 + 4 \times 11 + 6 \times 14)/5 - 4 \times 9.6 = 4$
- sehingga, X₁ and X₂ naik bersama karena $\sigma_{12} > 0$

Korelasi antara Dua Variabel Numerik

- Korelasi antara dua variabel X₁ dan X₂ adalah kovarian standar, diperoleh dengan menormalkan kovarian dengan standar deviasi masing-masing variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

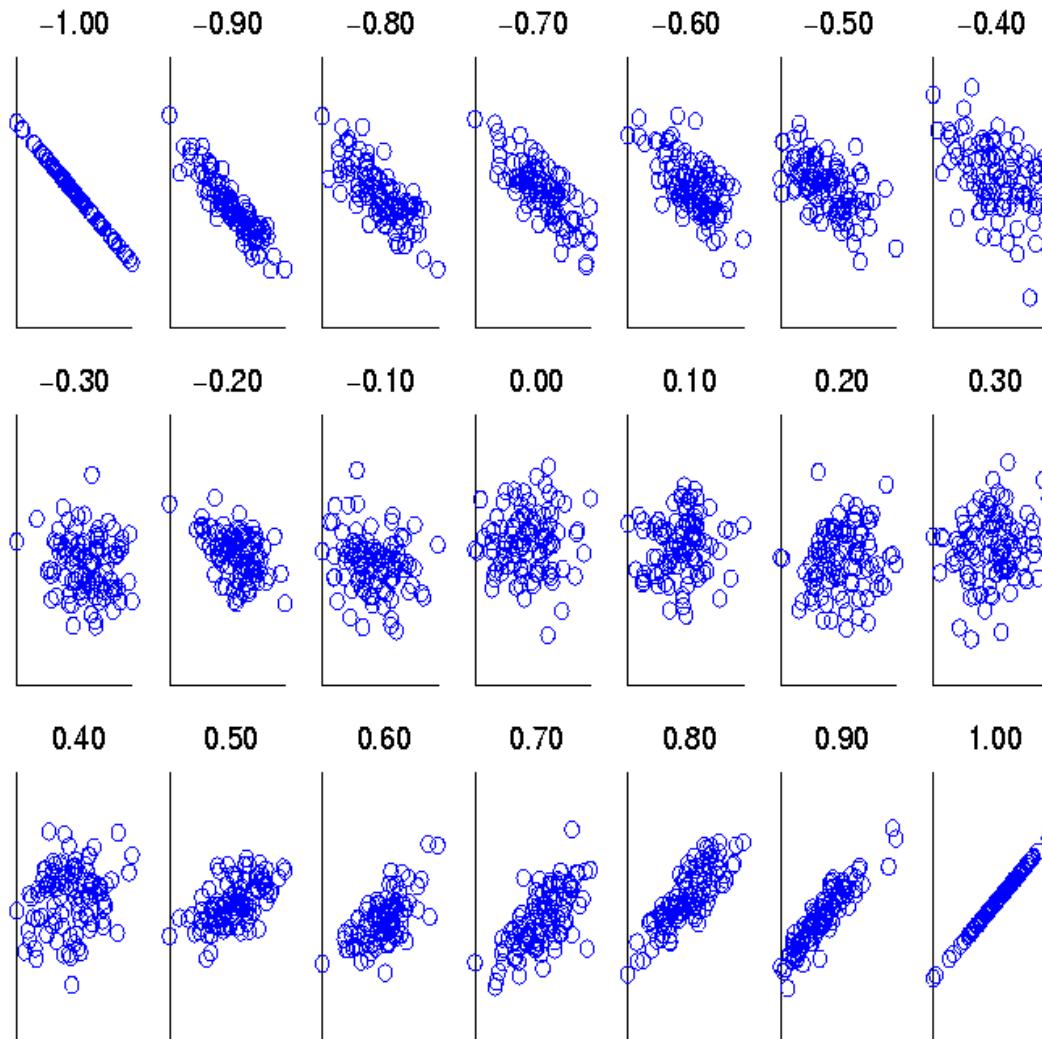
- Contoh korelasi untuk dua atribut X₁ dan X₂:

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^n (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^n (x_{i2} - \hat{\mu}_2)^2}}$$

Dimana n adalah jumlah tuple, μ_1 dan μ_2 adalah mean dari X₁ dan X₂, σ_1 dan σ_2 adalah standar deviasi dari X₁ dan X₂

- jika $\rho_{12} > 0$: A dan B berkorelasi positif (Nilai X₁ meningkat seiring dengan meningkatnya nilai X₂)
 - Semakin tinggi, semakin kuat korelasi
- jika $\rho_{12} = 0$: independent (di bawah asumsi yang sama seperti yang dibahas dalam ko-varians)
- jika $\rho_{12} < 0$: berkorelasi negatif

Memvisualisasikan Perubahan Koefisien Korelasi



- Rentang nilai koefisien korelasi:
[-1, 1]
- Satu set scatter plots menunjukkan himpunan titik dan koefisien korelasinya yang berubah dari -1 to 1

Matriks Kovarian

- Informasi varians dan kovarian untuk kedua variabel X_1 dan X_2 dapat diringkas sebagai matriks kovarian 2×2

$$\begin{aligned}\Sigma &= E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = E[(\begin{matrix} X_1 - \mu_1 \\ X_2 - \mu_2 \end{matrix})(\begin{matrix} X_1 - \mu_1 & X_2 - \mu_2 \end{matrix})] \\ &= \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & E[(X_2 - \mu_2)(X_2 - \mu_2)] \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\end{aligned}$$

- Menggeneralisasikannya ke dimensi d , menjadi

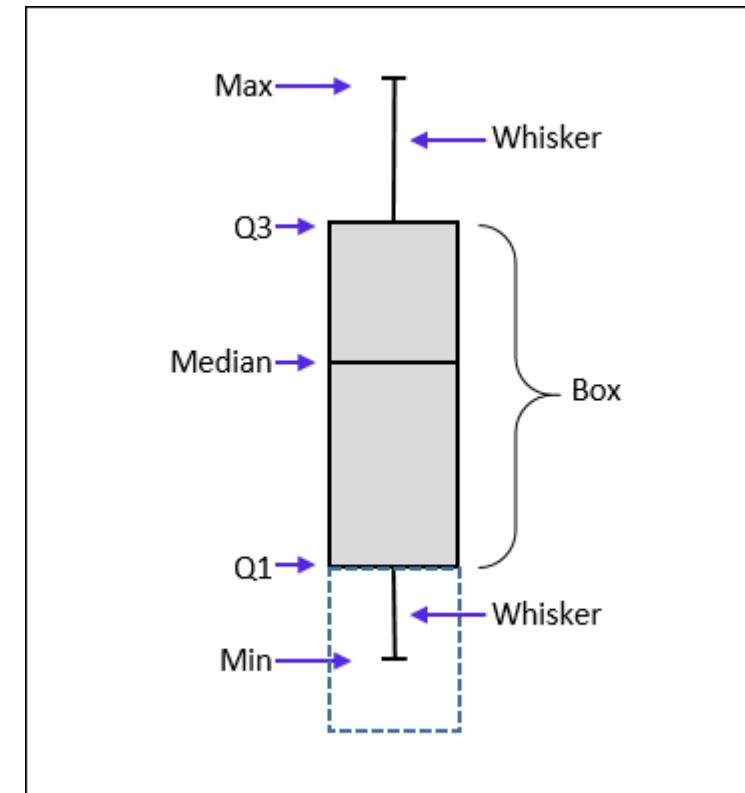
$$D = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{pmatrix} \quad \Sigma = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Tampilan Grafis Deskripsi Statistik Dasar

- **Boxplot:** Tampilan grafis distribusi data lima angka penting
- **Histogram:** sumbu-x adalah values, sumbu-y adalah representasi Frekuensi
- **Quantile plot:** setiap nilai x_i dipasangkan dengan f_i menunjukkan bahwa kira-kira $100 f_i \%$ data adalah $\leq x_i$
- **Quantile-quantile (q-q) plot:** grafik kuantil dari satu distribusi univarian terhadap kuantil yang sesuai dari yang lain
- **Scatter plot:** Setiap pasangan nilai adalah sepasang koordinat dan diplot sebagai titik dalam bidang

Mengukur Dispersi Data: Kuartil & Boxplots

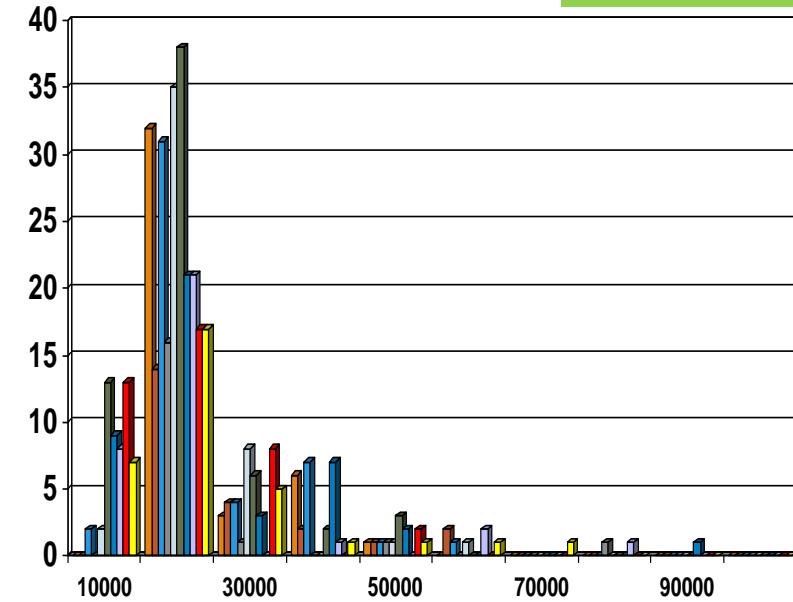
- **Quartiles:** Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
- **Inter-quartile range:** $IQR = Q_3 - Q_1$
- **Five number summary:** min, Q_1 , median, Q_3 , max
- **Boxplot:** Data diwakili dengan kotak
 - Q_1 , Q_3 , IQR: Ujung kotak berada di kuartil pertama dan ketiga, yaitu tinggi kotak adalah IQR
 - Median (Q_2) ditandai dengan garis di dalam kotak
 - Whiskers: dua batas di luar kotak diperpanjang ke Minimum dan Maximum
- Outliers: titik di luar ambang batas outlier yang ditentukan, diplot secara individual
- **Outlier:** biasanya, nilai lebih tinggi/lebih rendah dari $1,5 \times IQR$



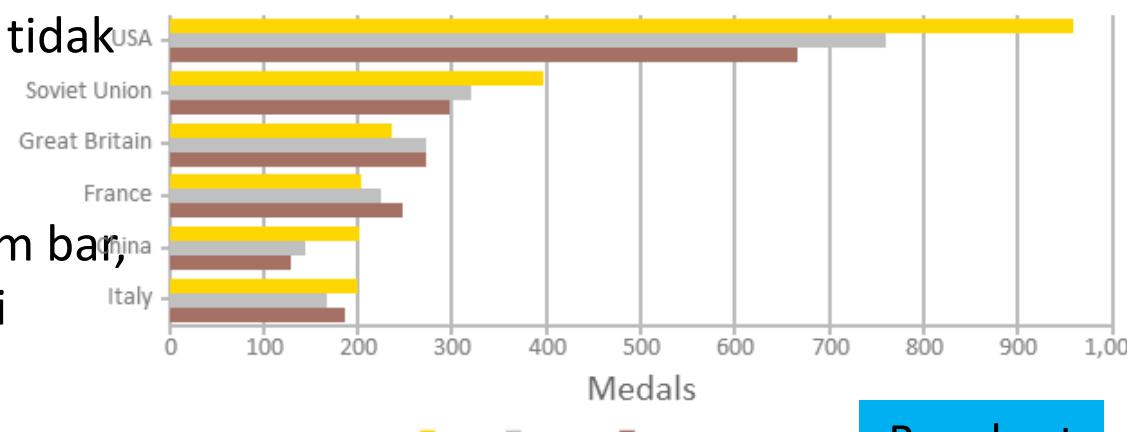
Analisis Histogram

- ❑ Histogram: Tampilan grafik frekuensi tabulasi, ditampilkan sebagai bar
- ❑ Perbedaan antara histogram dan diagram bar
 - ❑ Histogram digunakan untuk menunjukkan distribusi variabel sedangkan diagram bar digunakan untuk membandingkan variabel
 - ❑ Histogram memplot data kuantitatif binned (data numerik yang telah dikelompokkan ke dalam interval-interval tertentu mis: pengelompokan usia) sementara diagram bar memplot data kategoris
 - ❑ Bar dapat diurutkan ulang dalam diagram bar tetapi tidak dalam histogram
 - ❑ Berbeda dari diagram bar karena luas bar yang menunjukkan nilai, bukan tinggi seperti pada diagram bar, perbedaan penting ketika kategorinya tidak memiliki lebar yang seragam

Histogram

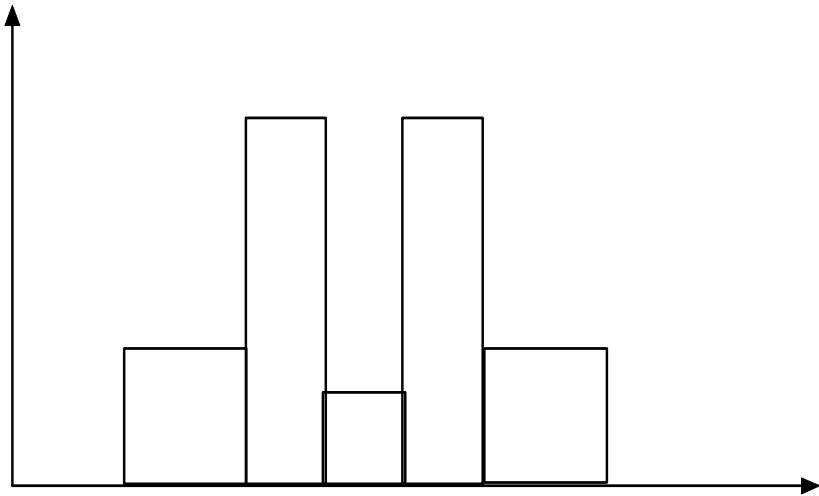


Olympic Medals of all Times (till 2012 Olympics)

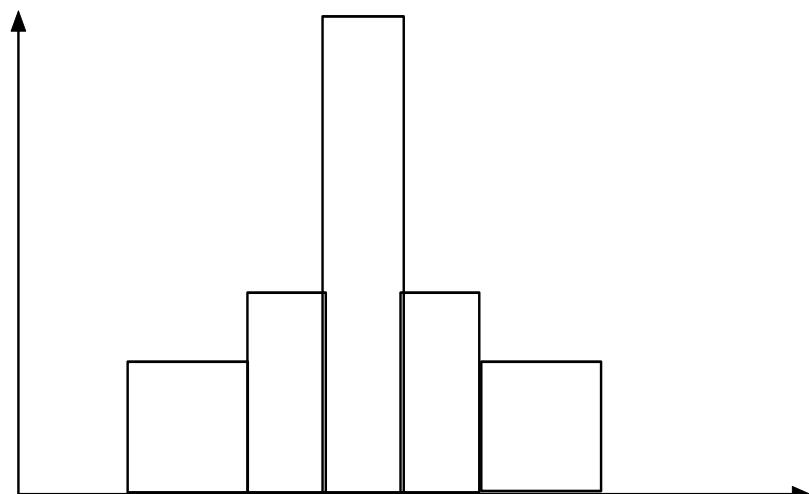


Bar chart

Histogram Sering Memberitahukan Lebih Banyak Hal daripada Boxplot

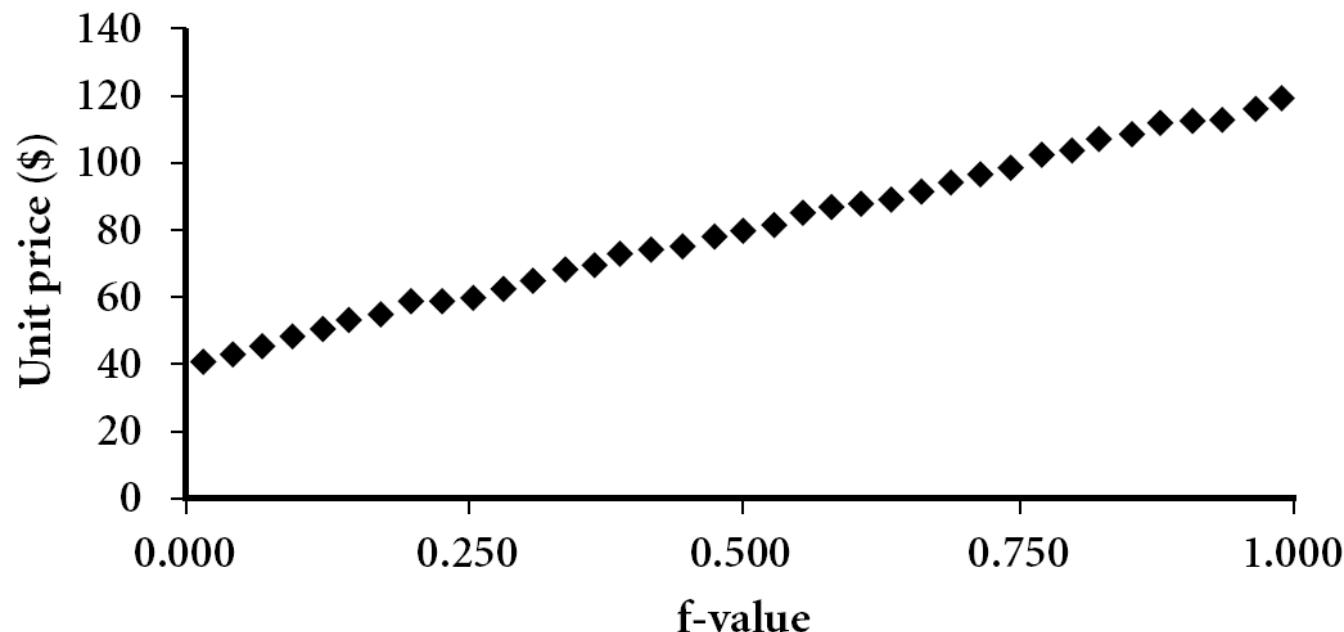


- Dua histogram yang ditampilkan di sebelah kiri mungkin memiliki representasi boxplot yang sama
- Nilai yang sama untuk: min, Q1, median, Q3, max
- Tetapi mereka memiliki distribusi data yang agak berbeda



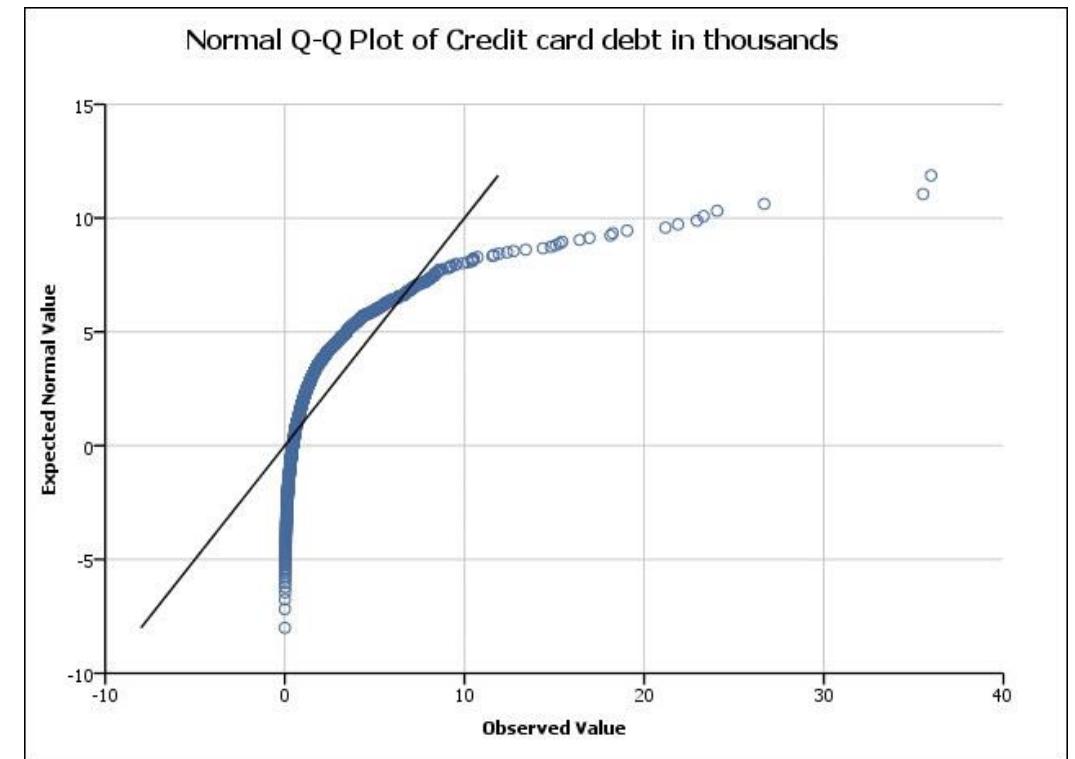
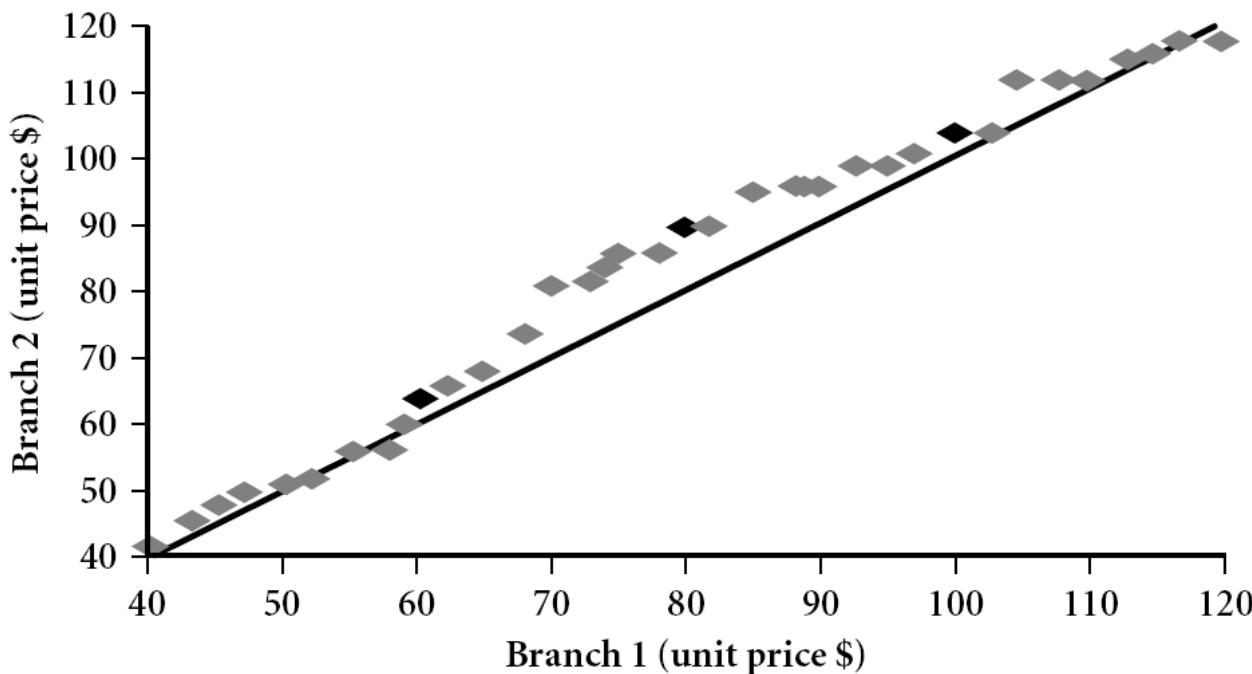
Quantile Plot

- ❑ Menampilkan semua data (memungkinkan pengguna untuk menilai perilaku keseluruhan dan kejadian yang tidak biasa)
- ❑ Plot informasi kuantil
 - ❑ Untuk data x_i , data diurutkan dalam urutan yang meningkat, f_i menunjukkan bahwa kira-kira $100 f_i\%$ data di bawah atau sama dengan nilai x_i



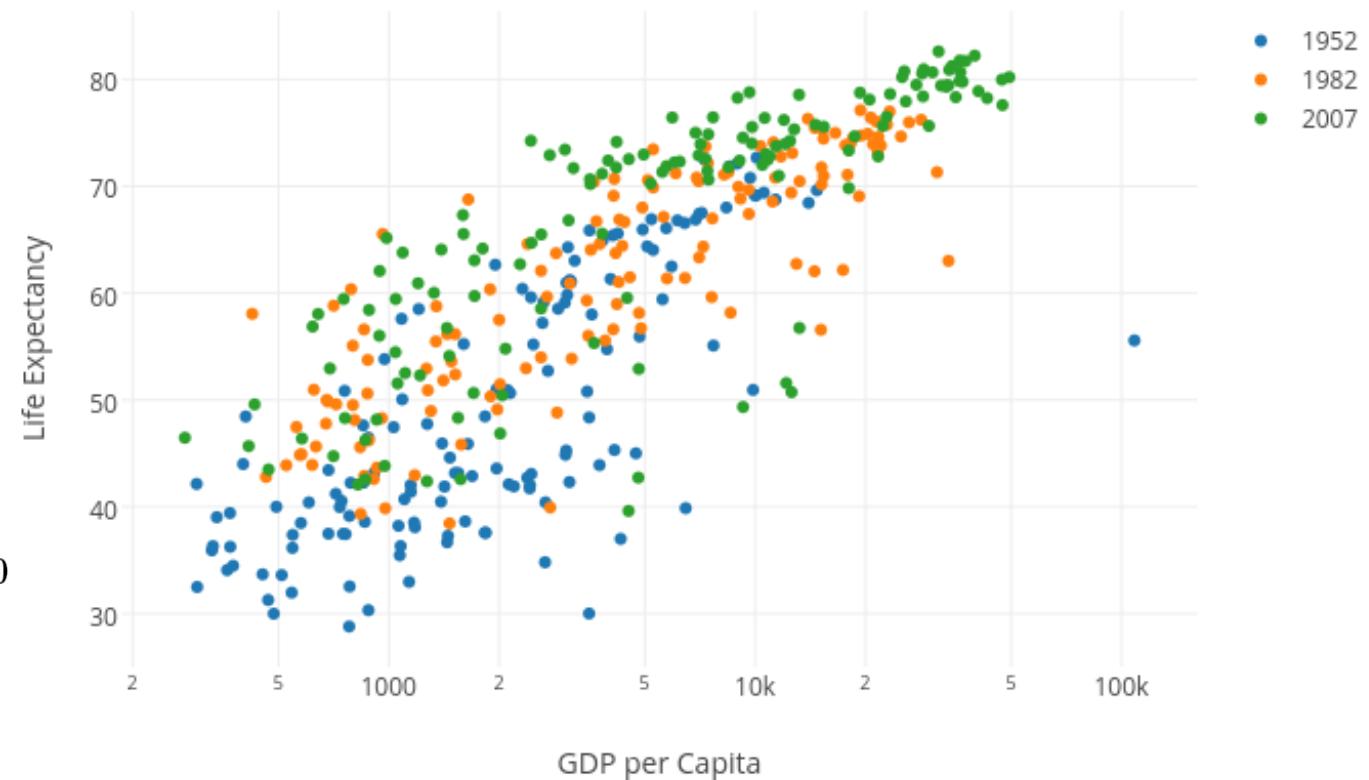
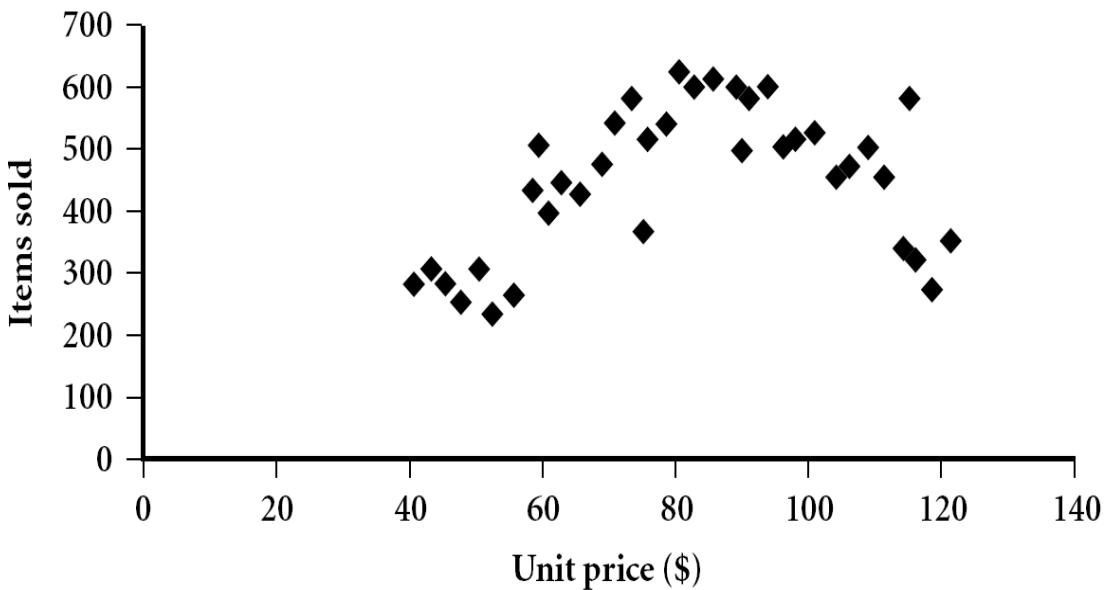
Plot Kuantil-Kuantil (Q-Q)

- Grafik kuantil dari satu distribusi univariat terhadap kuantil yang sesuai dari yang lain
- View: Apakah ada pergeseran dalam peralihan dari satu distribusi ke distribusi lainnya?
- Contoh menunjukkan harga satuan barang yang dijual di Cabang 1 vs. Cabang 2 untuk setiap kuantil. Harga satuan barang yang dijual di Cabang 1 cenderung lebih rendah daripada harga di Cabang 2

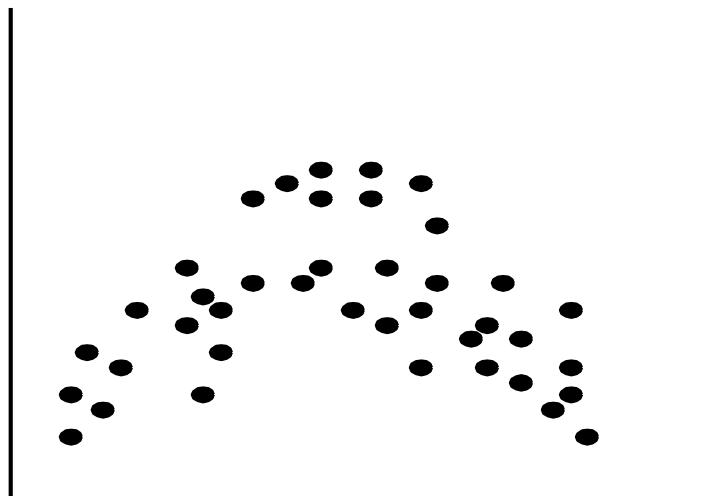
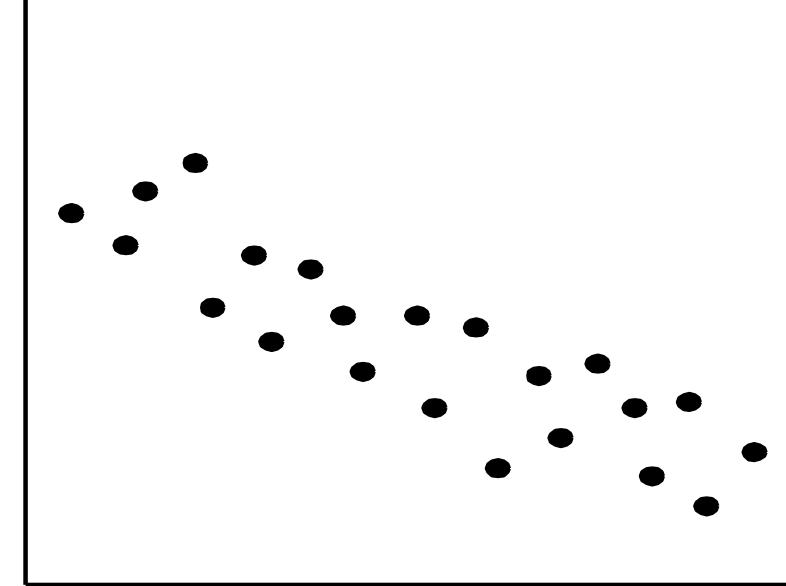
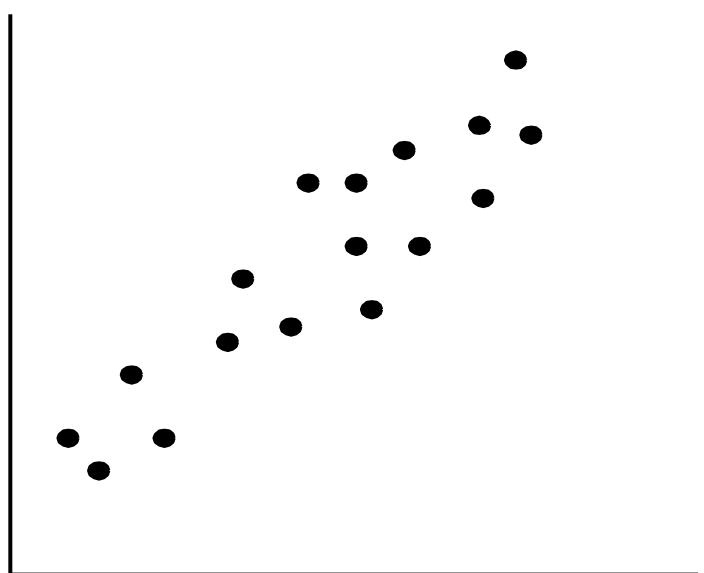


Scatter plot

- Memberikan pandangan pertama pada data bivariat untuk melihat kelompok titik, outlier, dll.
- Setiap pasangan nilai diperlakukan sebagai sepasang koordinat dan diplot sebagai titik di bidang

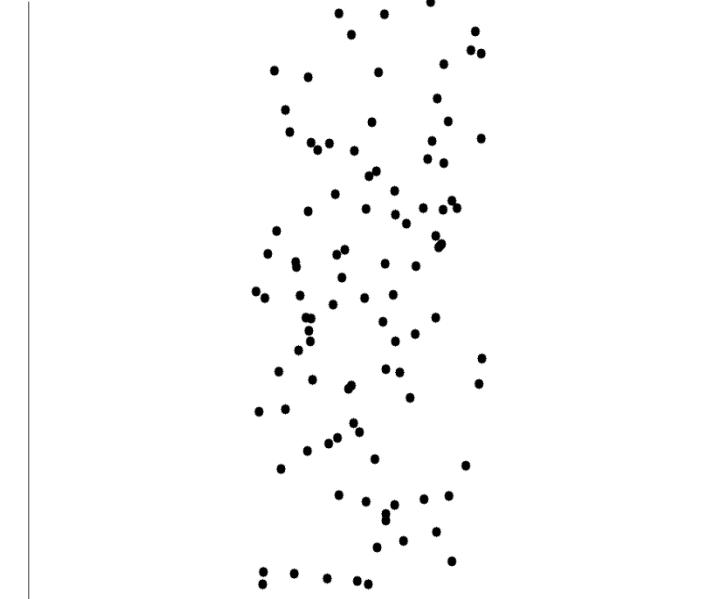
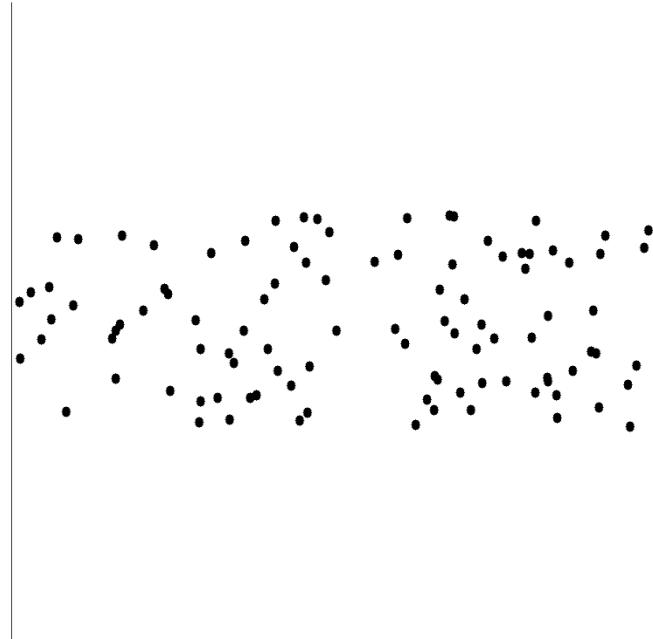


Data Berkorelasi Positif dan Negatif



- Kiri = berkorelasi positif
- Kanan = berkorelasi negatif

Data yang tidak berkorelasi



Data Quality, Data Cleaning and Data Integration

- Data Quality
- Data Cleaning
- Data Integration

Data Preprocessing?

- ❑ **Data cleaning**
 - ❑ Menangani missing data, smooth noisy data, identifikasi atau menghapus outliers, dan Mengatasi inkonsistensi
- ❑ **Data integration**
 - ❑ Integrasi beberapa databases, data cubes, atau files
- ❑ **Data reduction**
 - ❑ Dimensionality reduction
 - ❑ Numerosity reduction
 - ❑ Data compression
- ❑ **Data transformation dan data discretization**
 - ❑ Normalization
 - ❑ Concept hierarchy generation

Why Preprocess the Data? – Data Quality Issues

- Ukuran untuk data quality
 - Akurasi: benar atau salah, akurat atau tidak
 - Kelengkapan: tidak tercatat, tidak tersedia, ...
 - Konsistensi: beberapa diubah tetapi beberapa tidak, menggantung, ...
 - Ketepatan Waktu: apakah pembaruan tepat waktu?
 - Kepercayaan: seberapa dapat dipercaya data tersebut benar?
 - Keterpahaman: seberapa mudah data tersebut dapat dipahami?

Data Cleaning

- ❑ Data di Dunia Nyata Banyak yang berpotensi salah, misalnya, instrumen rusak, kesalahan manusia atau komputer, dan kesalahan transmisi
 - ❑ Incomplete: tidak memiliki nilai atribut, tidak memiliki atribut tertentu yang menarik, atau hanya berisi data agregat
 - ❑ Misalnya *Pekerjaan* = “ ” (missing data)
 - ❑ Noisy: Berisi noise, errors, atau outliers
 - ❑ Misalnya *Gaji* = “-10” (errors)
 - ❑ Inconsistent: Berisi perbedaan kode atau nama, e.g.,
 - ❑ *Age* = “42”, *Birthday* = “03/07/2010”
 - ❑ Was rating “1, 2, 3”, now rating “A, B, C”
 - ❑ perbedaan antara duplicate records
 - ❑ Kesengajaan (misalnya, data yang hilang yang disamarkan)

Incomplete (Missing) Data

- ❑ Data tidak selalu available
 - ❑ Contoh: banyak tuple tidak memiliki nilai yang tercatat untuk beberapa atribut, seperti pendapatan pelanggan dalam data penjualan
- ❑ Missing data disebabkan oleh
 - ❑ Kerusakan peralatan
 - ❑ Tidak konsisten dengan data lain yang tercatat sehingga dihapus
 - ❑ Data tidak dimasukkan karena kesalahpahaman
 - ❑ Data tertentu mungkin tidak dianggap penting pada saat entri
 - ❑ Tidak mencatat riwayat atau perubahan data

Cara Menangani Missing Data?

- Abaikan tuple: biasanya dilakukan ketika label kelas hilang (saat melakukan klasifikasi)—tidak efektif ketika % nilai yang hilang per atribut sangat bervariasi
- Isi nilai yang hilang secara manual
- Isi secara otomatis dengan
 - konstanta global: misalnya, "tidak diketahui", kelas baru?!
 - rata-rata atribut
 - rata-rata atribut untuk semua sampel yang termasuk dalam kelas yang sama
 - **nilai yang paling mungkin: berbasis inferensi seperti rumus Bayesian atau pohon keputusan**

Noisy Data

- **Noise:** kesalahan acak atau varians dalam variabel terukur
- **Incorrect attribute values** mungkin karena
 - Instrumen pengumpulan data yang rusak
 - Masalah entri data
 - Masalah transmisi data
 - Keterbatasan teknologi
 - Inkonsistensi dalam konvensi penamaan
- **Masalah data lainnya**
 - Catatan duplikat
 - Data tidak lengkap
 - Data yang tidak konsisten

Cara Menangani Noisy Data?

- Binning
 - Pertama-tama urutkan data dan partisi ke dalam bin (frekuensi yang sama)
 - Kemudian lakukan smooth dengan rata-rata bin , median bin, boundaries bin
- Regression
 - Smooth dengan Menyesuaikan data ke dalam fungsi regresi
- Clustering
 - Mendeteksi dan menghapus outlier
- Semi-supervised: kombinasi komputer dan inspeksi manusia
 - Mendeteksi nilai yang mencurigakan dan pengecekan oleh manusia (misalnya, menangani kemungkinan outlier)

Data Cleaning sebagai Proses

- ❑ **Deteksi Data discrepancy**
 - ❑ Gunakan metadata (misalnya, domain, rentang, dependensi, distribusi)
 - ❑ Periksa bidang overloading
 - ❑ Periksa keunikan rule, rule berurutan dan rule null
 - ❑ Gunakan alat komersial
 - ❑ Data scrubbing: menggunakan pengetahuan domain sederhana (misalnya, kode pos, pemeriksa ejaan) untuk mendeteksi kesalahan dan melakukan koreksi
 - ❑ Data auditing: dengan menganalisis data untuk menemukan aturan dan hubungan untuk mendeteksi pelanggar (misalnya, korelasi dan pengelompokan untuk menemukan outlier)
- ❑ **Data migration and integration**
 - ❑ Data migration tools: menentukan transformasi
 - ❑ ETL (Extraction/Transformation>Loading) tools: Memungkinkan pengguna untuk menentukan transformasi melalui antarmuka pengguna grafis
- ❑ Integrasi kedua proses
 - ❑ Iteratif dan interaktif

Data Integration

- ❑ Data integration
 - ❑ Menggabungkan data dari berbagai sumber ke dalam penyimpanan yang koheren
- ❑ Mengapa dilakukan data integration?
 - ❑ Membantu mengurangi/menghindari noise
 - ❑ Mendapatkan gambaran yang lebih lengkap
 - ❑ Meningkatkan kecepatan dan kualitas penambangan
- ❑ **Skema Integrasi :**
 - ❑ misalnya, A.cust-id B.cust- #
 - ❑ Mengintegrasikan metadata dari berbagai sumber
- ❑ **Identifikasi entitas:**
 - ❑ Mengidentifikasi entitas dari beberapa sumber data, misalnya, Bill Clinton = William Clinton

Menangani Noise dalam Integrasi Data

- Mendeteksi konflik data value
 - Untuk entitas yang sama, nilai atribut dari sumber yang berbeda representasi yang berbeda, skala yang berbeda
- Menyelesaikan informasi konflik
 - Gunakan mean/median/mode/max/min
 - Gunakan yang terbaru
 - Pertimbangkan kualitas sumber
- Pembersihan data + integrasi data

Menangani Redundansi dalam Integrasi Data

- ❑ Data redundan sering terjadi saat integrasi beberapa database
 - ❑ *Identifikasi objek: Atribut atau objek yang sama mungkin memiliki nama yang berbeda dalam database yang berbeda*
 - ❑ *Data yang dapat diturunkan: Satu atribut mungkin merupakan atribut "turunan" di tabel lain*
- ❑ Apa masalahnya?
 - ❑
$$Y = 2X \rightarrow Y = X_1 + X_2 \quad Y = 3X_1 - X_2 \quad Y = -1291X_1 + 1293X_2$$
- ❑ Redundant atribut dapat dideteksi dengan analisis korelasi dan analisis kovarian

Data Transformation

- Normalisasi
- Diskretisasi
- Kompresi Data
- Sampling

Data Transformation

- ❑ Fungsi yang memetakan seluruh rangkaian nilai dari atribut yang diberikan ke rangkaian nilai pengganti yang baru, yaitu setiap nilai lama dapat diidentifikasi dengan salah satu nilai baru
- ❑ Metode
 - ❑ Smoothing: Menghapus noise dari data
 - ❑ Attribute/feature construction
 - ❑ Atribut-atribut baru yang dibangun dari atribut-atribut yang diberikan
 - ❑ Aggregation: Summarization, data cube construction
 - ❑ Normalization: diskalakan dalam rentang yg lebih kecil, rentang yg ditentukan
 - ❑ min-max normalization
 - ❑ z-score normalization
 - ❑ normalization dengan penskalaan desimal
 - ❑ Discretization: Concept hierarchy climbing

Normalization

- **Min-max normalization:** to $[new_min_A, new_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new_max_A - new_min_A) + new_min_A$$

- Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]

- Then \$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ : mean, σ : standard deviation): **Standarisasi**

$$v' = \frac{v - \mu_A}{\sigma_A}$$

Z-score: menunjukkan seberapa jauh suatu data menjauh dari rata-rata (mean) dalam satuan standar deviasi

- Ex. Let $\mu = 54,000$, $\sigma = 16,000$. Then $\frac{73,600 - 54,000}{16,000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j} \quad \text{Di mana } j \text{ adalah bilangan bulat terkecil sehingga } \text{Max}(|v'|) < 1$$

Diskretisasi

- ❑ Tiga jenis atribut
 - ❑ Nominal—nilai dari himpunan yang tidak berurutan, misalnya, warna, profesi
 - ❑ Ordinal—nilai dari himpunan yang diurutkan, misalnya, pangkat militer atau akademik
 - ❑ Numeric—bilangan real, misalnya, bilangan bulat atau bilangan real
- ❑ Diskretisasi: Membagi rentang atribut kontinu menjadi interval
 - ❑ Label interval kemudian dapat digunakan untuk mengganti nilai data actual
 - ❑ Kurangi ukuran data dengan diskritisasi
 - ❑ Supervised vs. unsupervised
 - ❑ Split (top-down) vs. merge (bottom-up)
 - ❑ Diskretisasi dapat dilakukan secara rekursif pada atribut
 - ❑ Persiapan untuk analisis lebih lanjut, misalnya, klasifikasi

Metode Diskretisasi Data

- ❑ Binning
 - ❑ Top-down split, unsupervised
- ❑ Histogram analysis
 - ❑ Top-down split, unsupervised
- ❑ Clustering analysis
 - ❑ Unsupervised, top-down split or bottom-up merge
- ❑ Decision-tree analysis
 - ❑ Supervised, top-down split
- ❑ Correlation (e.g., χ^2) analysis
 - ❑ Unsupervised, bottom-up merge
- ❑ Note: Semua metode dapat diterapkan secara rekursif

Diskretisasi Sederhana: Binning

- Partisi equal-distance
 - Membagi rentang menjadi N interval dengan ukuran yang sama: grid seragam
 - jika A dan B adalah nilai terendah dan tertinggi dari atribut, lebar interval adalah: $W = (B - A)/N$.
 - Yang paling mudah, tetapi outlier mungkin mendominasi presentasi
 - Skewed data tidak ditangani dengan baik
- Partisi equal-frequency
 - Membagi rentang menjadi N interval, masing-masing berisi jumlah sampel yang kira-kira sama
 - Penskalaan data yang baik
 - Mengelola atribut kategoris bisa jadi rumit

Contoh: Metode Binning untuk Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- * Partition into equal-frequency (**equal-depth**) bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

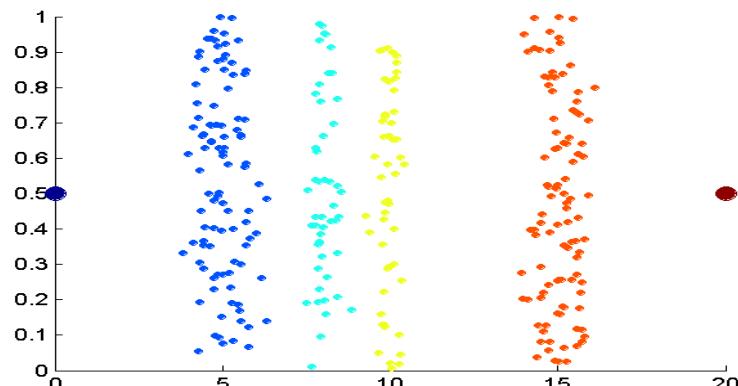
- * Smoothing by **bin means**:

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

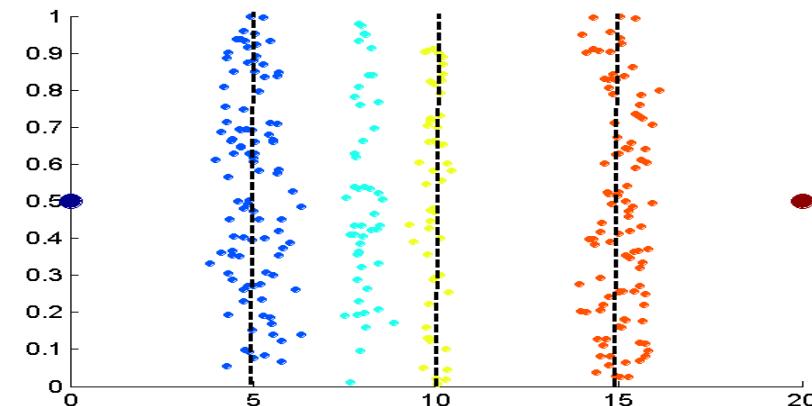
- * Smoothing by **bin boundaries**:

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

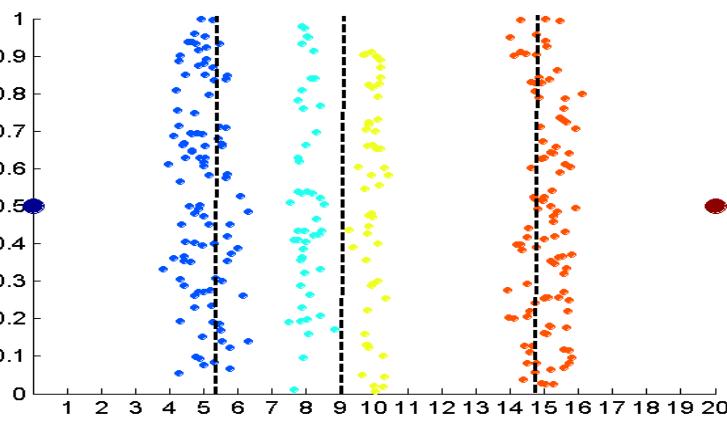
Diskretisasi Tanpa Supervision: Binning vs. Clustering



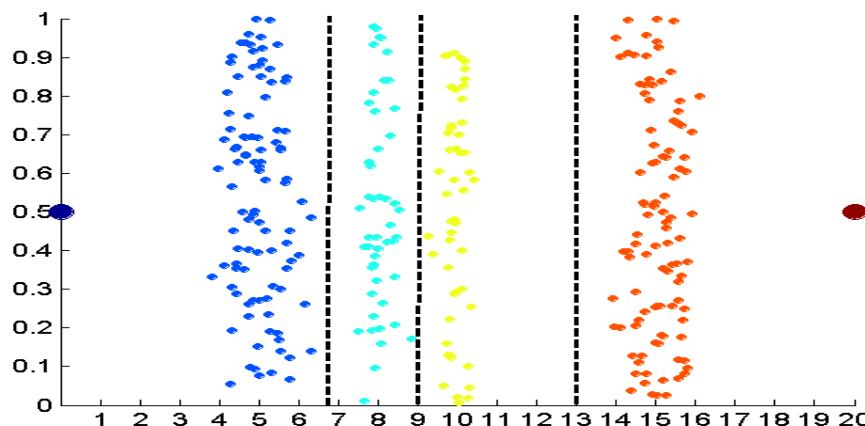
Data



Equal width (distance) binning



Equal depth (frequency) (binning)



K-means clustering leads to better results

Diskretisasi dengan Klasifikasi & Analisis Korelasi

- ❑ Classification (e.g., decision tree analysis)
 - ❑ Supervised: Diberi label kelas, misalnya, kanker vs. jinak
 - ❑ Menggunakan entropi untuk menentukan titik pemisahan (titik diskretisasi)
 - ❑ Top-down, recursive split
- ❑ Correlation analysis (e.g., Chi-merge: χ^2 -based discretization)
 - ❑ Supervised: use class information
 - ❑ Bottom-up merge: temukan interval neighbor terbaik (yang memiliki distribusi kelas yang serupa, yaitu nilai rendah χ^2) untuk di merger
 - ❑ Merger dilakukan secara rekursif, hingga kondisi berhenti yang telah ditentukan sebelumnya

Concept Hierarchy Generation

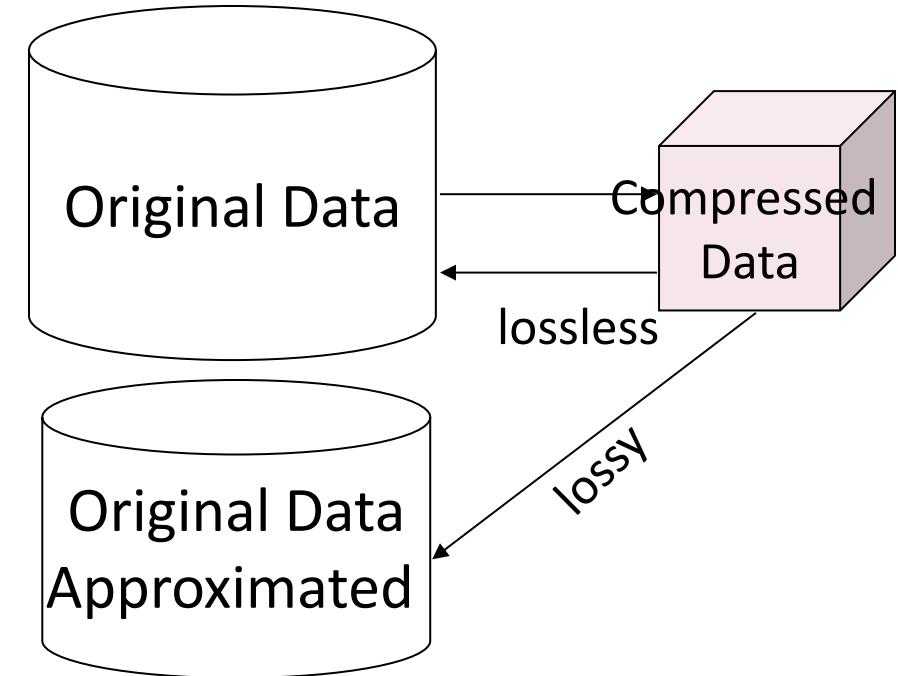
- Concept hierarchy mengatur konsep (yaitu, nilai atribut) secara hierarkis dan biasanya dikaitkan dengan setiap dimensi dalam data warehouse
- Concept hierarchy formation: Secara rekursif mengurangi data dengan mengumpulkan dan mengganti low level concepts (seperti nilai numerik untuk usia) dengan higher level concepts (seperti remaja, dewasa, atau senior)
- Concept hierarchies dapat ditentukan secara eksplisit oleh pakar domain dan/atau data warehouse designers
- Concept hierarchy dapat dibentuk secara otomatis untuk data numerik dan nominal—Untuk data numerik, gunakan metode diskretisasi

Concept Hierarchy Generation untuk Nominal Data

- ❑ Spesifikasi urutan atribut parsial/total secara eksplisit pada tingkat skema oleh pengguna atau ahli
 - ❑ *Jalan < Kecamatan < Kota < Provinsi*
- ❑ Spesifikasi hierarki untuk sekumpulan nilai dengan pengelompokan data eksplisit
 - ❑ {balikpapan, samarinda, penajam} < kaltim
- ❑ Spesifikasi hanya sebagian set atribut
 - ❑ Misalnya, hanya jalan < kecamatan, bukan yang lain
- ❑ Pembuatan hierarki otomatis (atau tingkat atribut) dengan analisis jumlah nilai yang berbeda
 - ❑ Misalnya, untuk sekumpulan atribut: {jalan, kecamatan, kota, provinsi}

Data Compression

- ❑ String compression
 - ❑ Biasanya tanpa kehilangan informasi, tetapi hanya manipulasi terbatas yang mungkin tanpa perlu ekspansi
- ❑ Audio/video compression
 - ❑ Biasanya lossy compression, dengan penyempurnaan progresif
 - ❑ Kadang-kadang fragmen kecil dari sinyal dapat direkonstruksi tanpa perlu merekonstruksi keseluruhan
- ❑ Time sequence is not audio
 - ❑ Biasanya pendek dan bervariasi perlahan seiring waktu
- ❑ Data reduction dan dimensionality reduction juga dapat dianggap sebagai bentuk kompresi data



Lossy vs. lossless compression

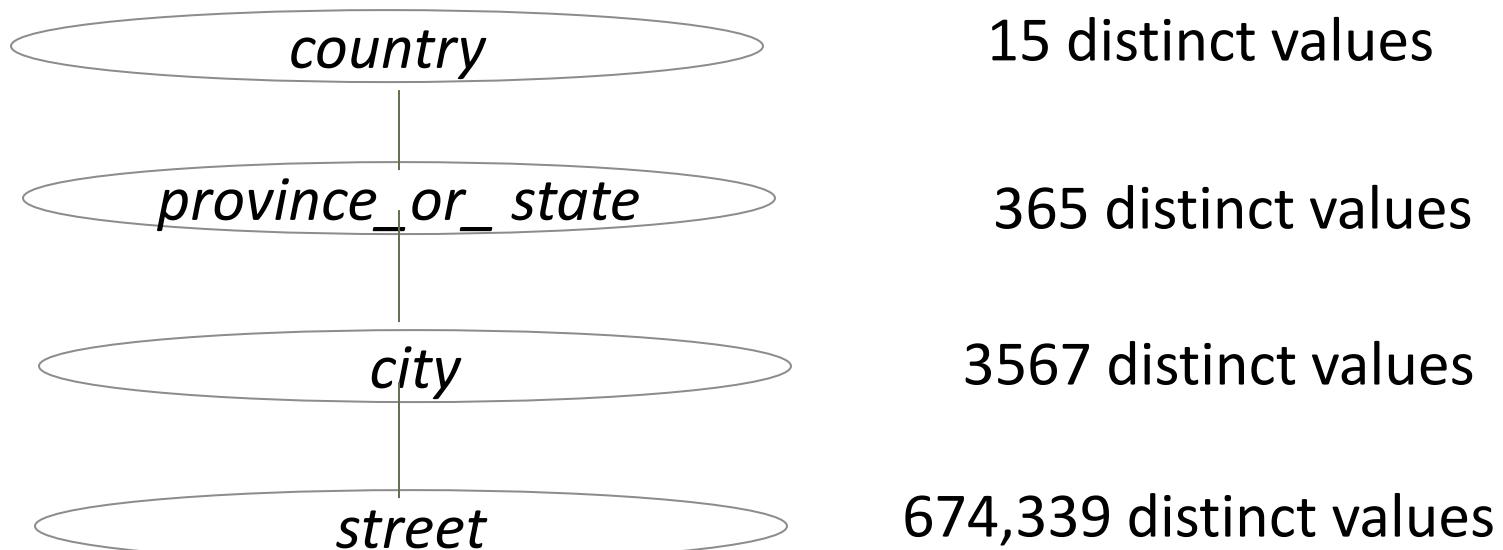
Data Cube Aggregation

- Tingkat terendah dari data cube (base cuboid)
 - Data agregat untuk entitas individu yang diminati
 - Misalnya seorang customer dalam data warehouse panggilan telepon
- Multiple levels agregasi dalam data cubes
 - Lebih mengurangi ukuran data yang akan ditangani
- Referensi tingkat yang sesuai
 - Gunakan representasi terkecil yang cukup untuk menyelesaikan tugas
- Kueri mengenai informasi teraggregasi sebaiknya dijawab menggunakan data cube, jika memungkinkan



Automatic Concept Hierarchy Generation

- ❑ Beberapa hierarki dapat dibuat secara otomatis berdasarkan analisis jumlah nilai yang berbeda per atribut dalam kumpulan data
 - ❑ Atribut dengan nilai paling unik ditempatkan di tingkat terendah hierarki
 - ❑ Pengecualian, misalnya, hari kerja, bulan, kuartal, tahun

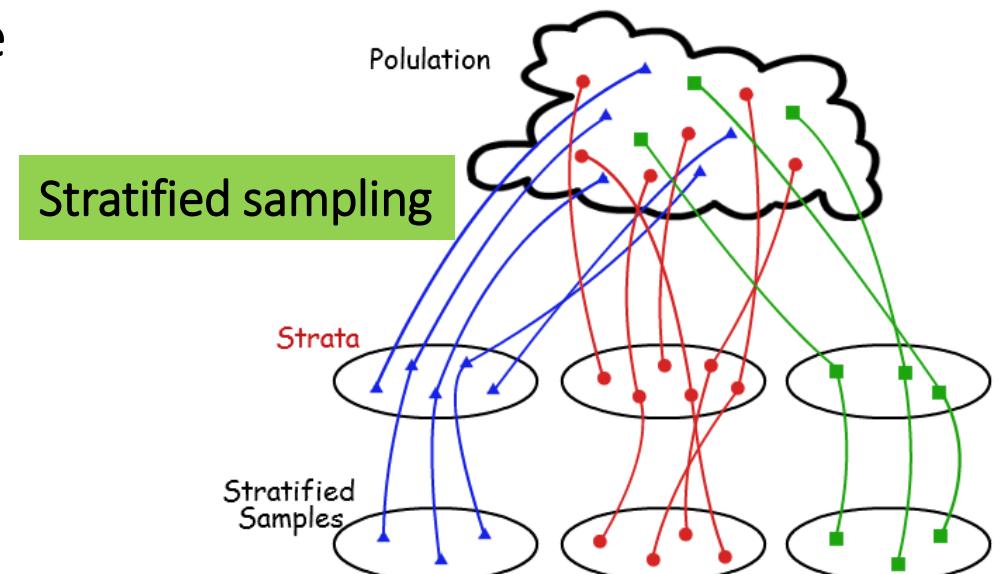
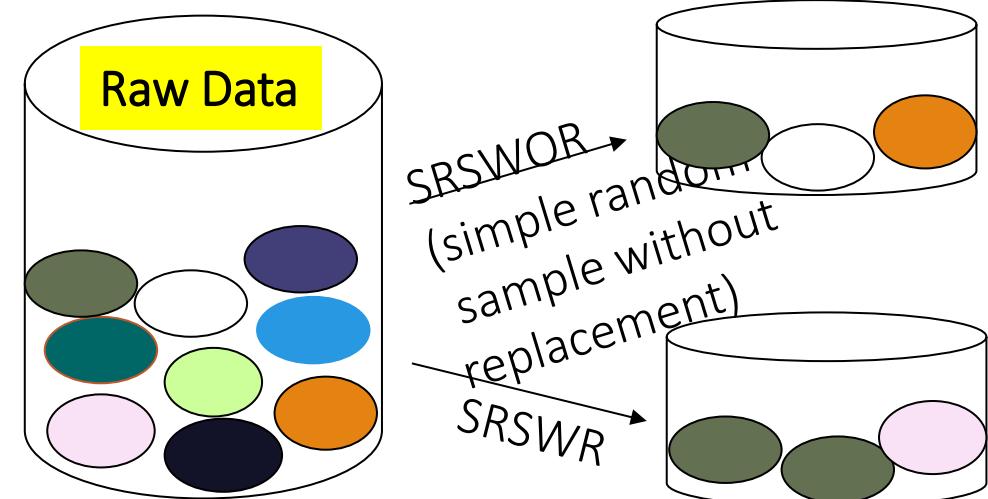


Sampling

- Sampling: obtaining a small sample s to represent the whole data set N
- Memungkinkan algoritma mining berjalan dengan kompleksitas yang berpotensi sub-linear terhadap ukuran data
- Key principle: Memilih subset data yang representative
 - Pengambilan sampel acak sederhana mungkin memiliki kinerja yang sangat buruk dengan adanya skew
 - Mengembangkan metode pengambilan sampel adaptif, misalnya, pengambilan sampel bertingkat

Types of Sampling

- ❑ **Simple random sampling:** probabilitas yang sama untuk memilih item tertentu
- ❑ **Sampling without replacement**
 - ❑ Setelah objek dipilih, objek tersebut dihapus dari populasi
- ❑ **Sampling with replacement**
 - ❑ A selected object is not removed from the population
- ❑ **Stratified sampling**
 - ❑ Partisi (atau mengelompokkan) kumpulan data, dan menarik sampel dari setiap partisi (secara proporsional, yaitu, persentase data yang kira-kira sama)

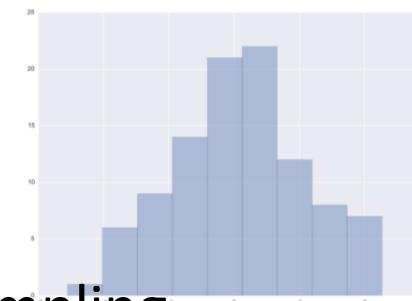
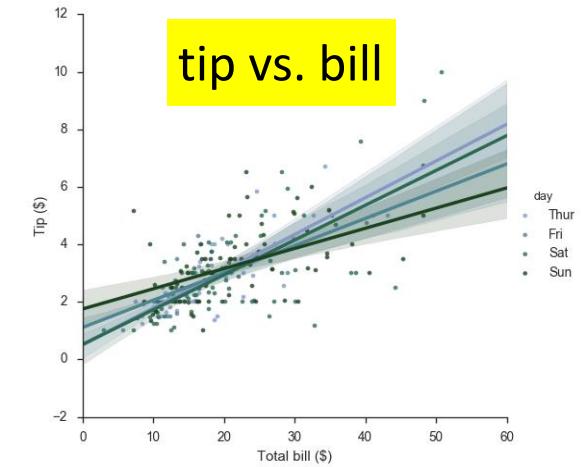


Data Reduction

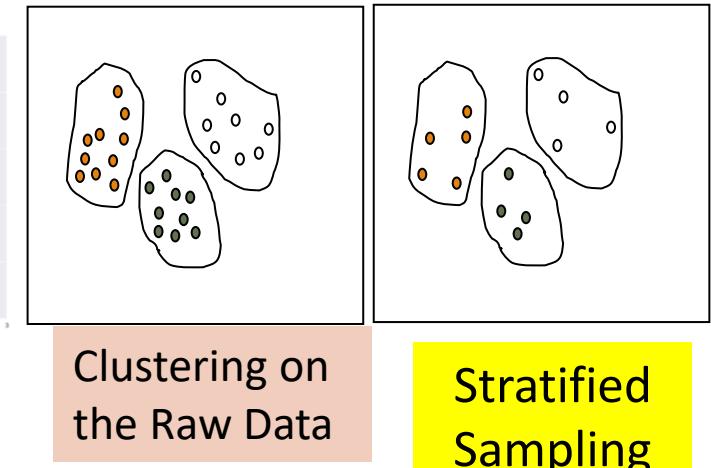
- ❑ Data reduction:
 - ❑ Mendapatkan representasi yang diperkecil dari himpunan data
 - ❑ Jauh lebih kecil volumenya tetapi menghasilkan hasil analitik yang hampir sama
- ❑ Mengapa data reduction?—database/data warehouse dapat menyimpan terabyte data
 - ❑ Analisis kompleks mungkin membutuhkan waktu yang sangat lama untuk dijalankan pada kumpulan data lengkap
- ❑ Metode untuk data reduction (*data size reduction* atau *numerosity reduction*)
 - ❑ Regression and Log-Linear Models
 - ❑ Histograms, clustering, sampling
 - ❑ Data cube aggregation
 - ❑ Data compression

Data Reduction: Parametric vs. Non-Parametric Methods

- Reduce data volume dengan memilih alternatif bentuk representasi data yang lebih kecil
- **Parametric methods** (misalnya, regresi)
 - Asumsikan data sesuai dengan model tertentu, estimasi parameter model, simpan hanya parameter tersebut, dan buang data (kecuali kemungkinan outlier)
 - Contoh: Model log-linear
- **Non-parametric methods**
 - Jangan mengasumsikan model
 - Major families: histograms, clustering, sampling, ...



Histogram

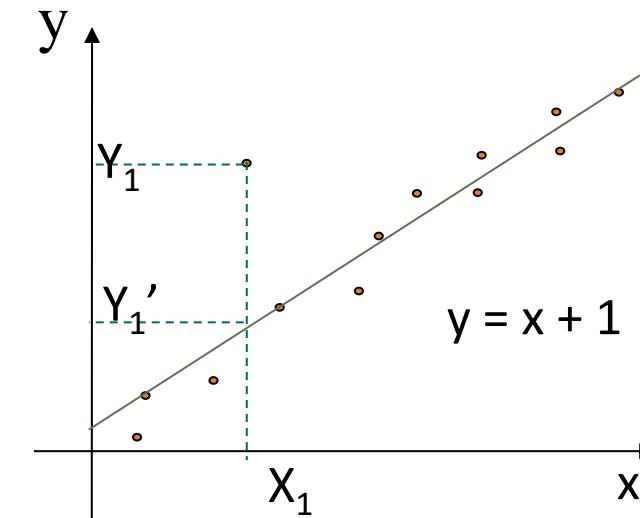


Clustering on
the Raw Data

Stratified
Sampling

Parametric Data Reduction: Regression Analysis

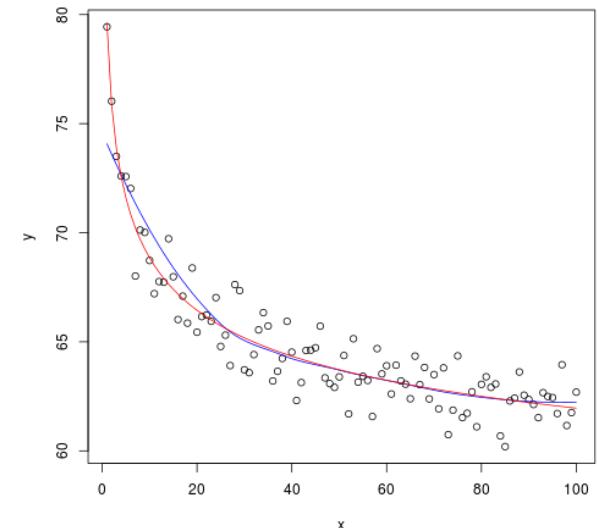
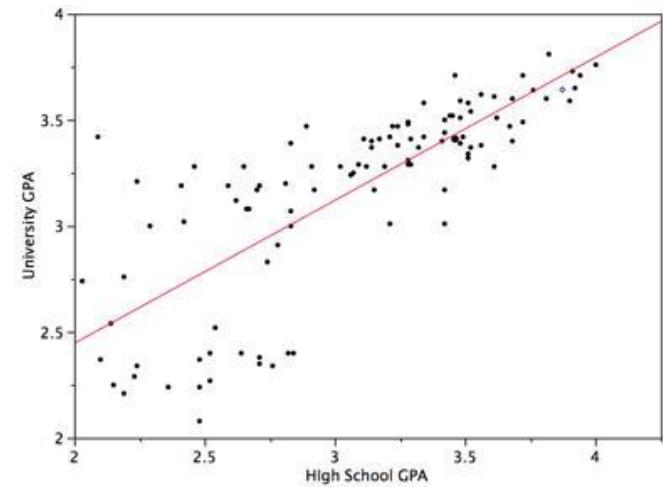
- Regression analysis: teknik pemodelan dan analisis data numerik yang terdiri dari nilai-nilai variabel dependen (juga disebut variabel respons atau pengukuran) dan satu atau lebih variabel independen (juga dikenal sebagai variabel penjelasan atau prediktor)
- Parameter diperkirakan untuk memberikan "kecocokan" data
- Pada umumnya, kecocokan terbaik dievaluasi dengan menggunakan metode kuadrat terkecil, tetapi kriteria lain juga telah digunakan



- Digunakan untuk prediksi (termasuk perkiraan data deret waktu), inferensi, pengujian hipotesis, dan pemodelan hubungan kausal

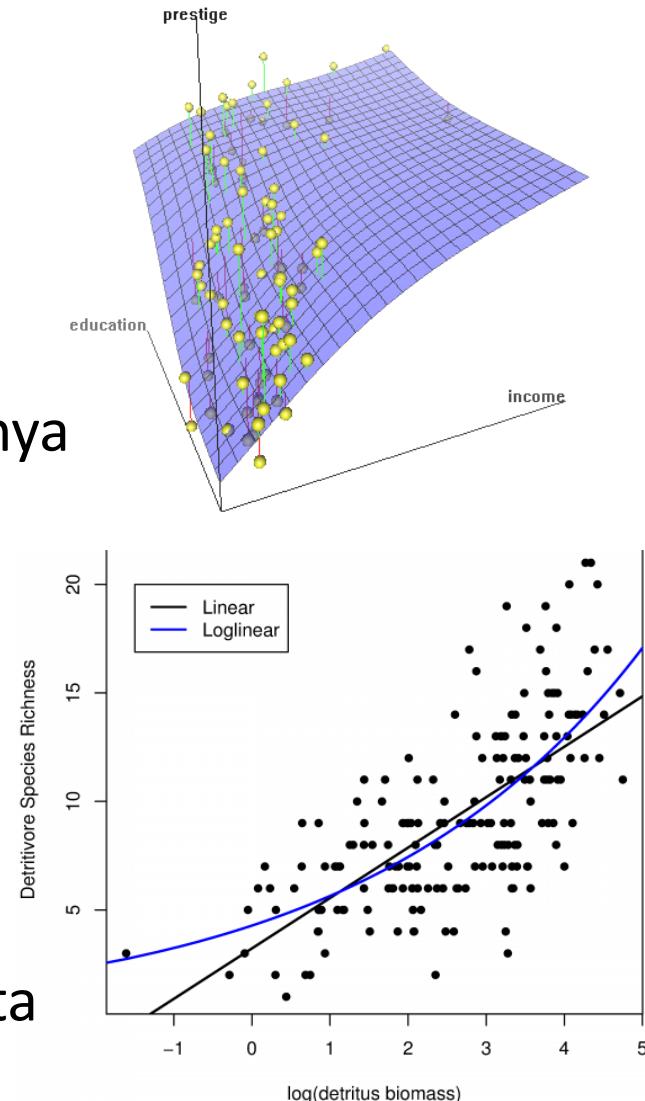
Linear and Multiple Regression

- ❑ Linear regression: $Y = w X + b$
 - ❑ Data yang dimodelkan agar sesuai dengan garis lurus
 - ❑ Sering menggunakan metode kuadrat terkecil untuk menyesuaikan garis
 - ❑ Dua koefisien regresi, w dan b , menentukan garis dan akan diperkirakan dengan menggunakan data yang ada
 - ❑ Menggunakan kriteria kuadrat terkecil ke nilai yang diketahui $Y_1, Y_2, \dots, X_1, X_2, \dots$
- ❑ Nonlinear regression:
 - ❑ Data dimodelkan oleh fungsi yang merupakan kombinasi nonlinier dari parameter model dan bergantung pada satu atau lebih variabel independent
 - ❑ Data dipasang dengan metode perkiraan berturut-turut



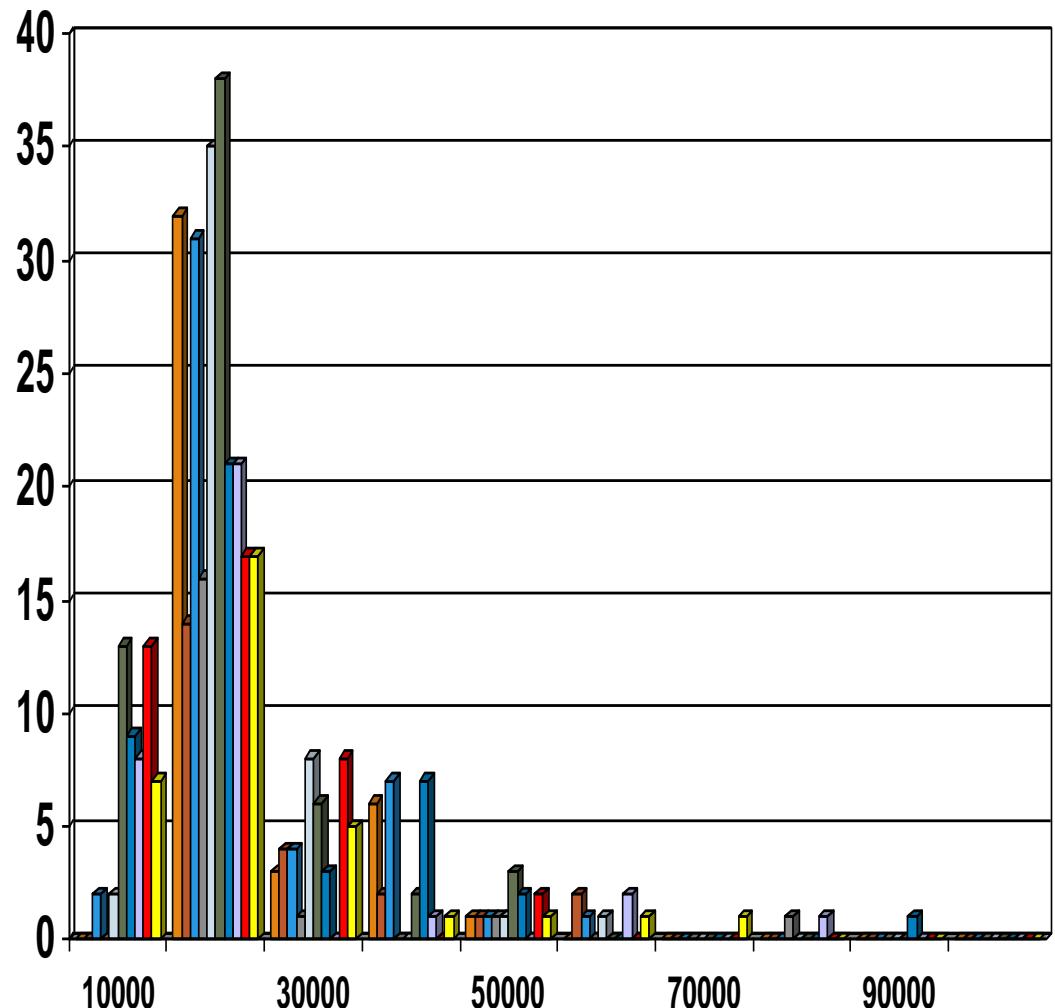
Multiple Regression and Log-Linear Models

- Multiple regression: $Y = b_0 + b_1 X_1 + b_2 X_2$
 - Memungkinkan variabel respons Y dimodelkan sebagai fungsi linier vektor fitur multidimensi
 - Banyak fungsi nonlinier dapat diubah menjadi di atas
- Log-linear model:
 - Model matematika yang berbentuk fungsi yang logaritmanya merupakan kombinasi linier dari parameter model, yang memungkinkan untuk menerapkan regresi linier (mungkin multivariat)
 - Perkirakan probabilitas setiap titik (tuple) dalam multi-dimen. ruang untuk sekumpulan atribut diskretisasi, berdasarkan subset kombinasi dimensi yang lebih kecil
 - Berguna untuk pengurangan dimensi dan penghalusan data



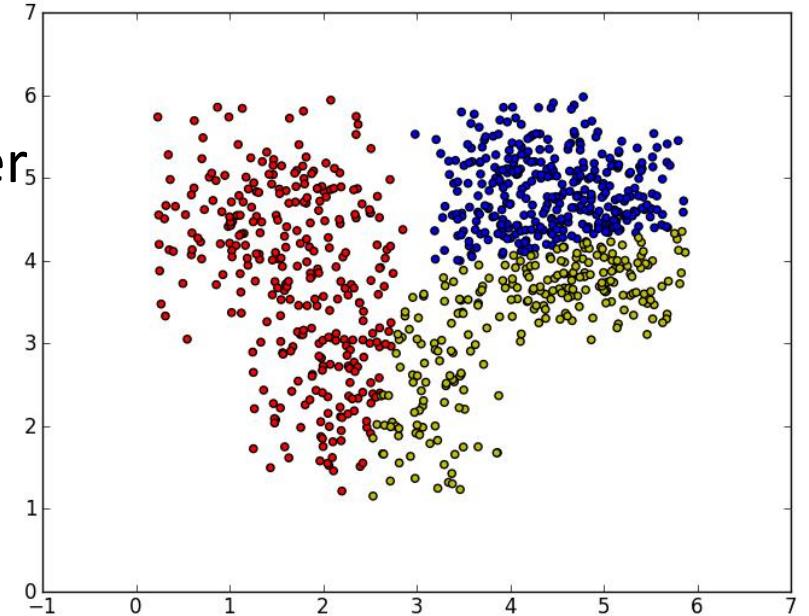
Histogram Analysis

- ❑ Membagi data ke dalam bucket dan simpan rata-rata (jumlah) untuk setiap bucket
- ❑ Partitioning rules:
 - ❑ Equal-width: equal bucket range
 - ❑ Equal-frequency (or equal-depth)



Clustering

- Mempartisi kumpulan data ke dalam kluster berdasarkan kesamaan, dan simpan representasi kluster (misalnya, centroid dan diameter) saja
- Dapat sangat efektif jika data terkelompok, tetapi tidak jika data "tersebar"
- Dapat memiliki klasterisasi hierarkis dan disimpan dalam struktur pohon indeks multi-dimensional



Dimensionality Reduction

- What Is Dimensionality Reduction?
- Dimensionality Reduction Methods
 - Principal Component Analysis
 - Attribute Subset Selection
 - Nonlinear Dimensionality Reduction Methods

What Is Dimensionality Reduction?

❑ Curse of dimensionality (Kutukan Dimensi)

- ❑ Saat dimensi meningkat, data menjadi semakin jarang
- ❑ Kepadatan dan jarak antara titik, yang penting untuk klasterisasi dan analisis outlier, menjadi kurang berarti
- ❑ Kombinasi subruang yang mungkin akan berkembang secara eksponensial

❑ Dimensionality reduction

- ❑ Mengurangi jumlah variabel acak yang sedang dipertimbangkan, dengan mendapatkan sekumpulan variabel utama

❑ Keuntungan dari dimensionality reduction

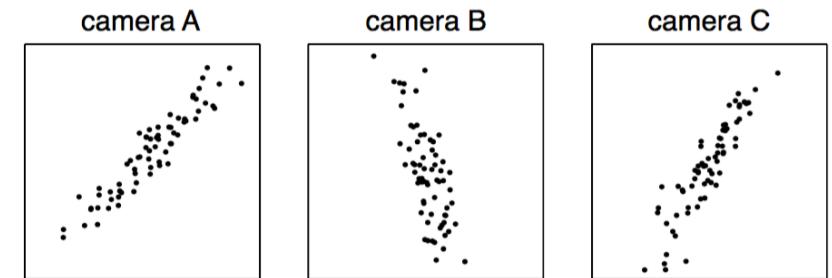
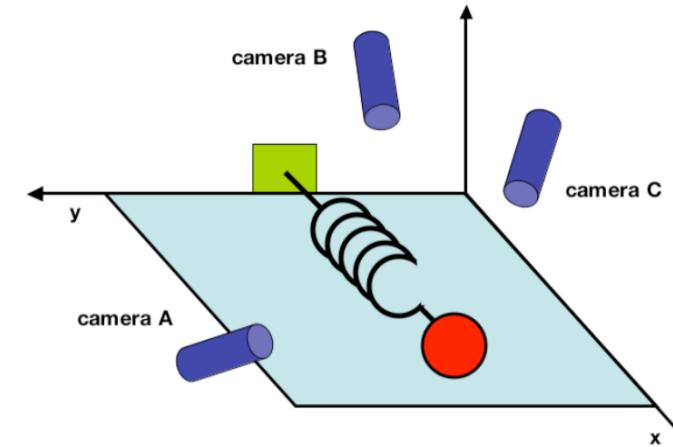
- ❑ Menghindari curse of dimensionality
- ❑ Membantu menghilangkan fitur yang tidak relevan dan reduce noise
- ❑ Kurangi waktu dan ruang yang dibutuhkan di data mining
- ❑ Memungkinkan visualisasi yang lebih mudah

Dimensionality Reduction Methods

- Dimensionality reduction methodologies
 - **Feature selection:** Menemukan subset variabel asli (atau fitur, atribut)
 - **Feature extraction:** Ubah data di ruang dimensi tinggi ke ruang dengan dimensi yang lebih kecil
- Beberapa metode dimensionality reduction
 - Principal Component Analysis
 - Attribute Subset Selection
 - Nonlinear Dimensionality Reduction

Principal Component Analysis (PCA)

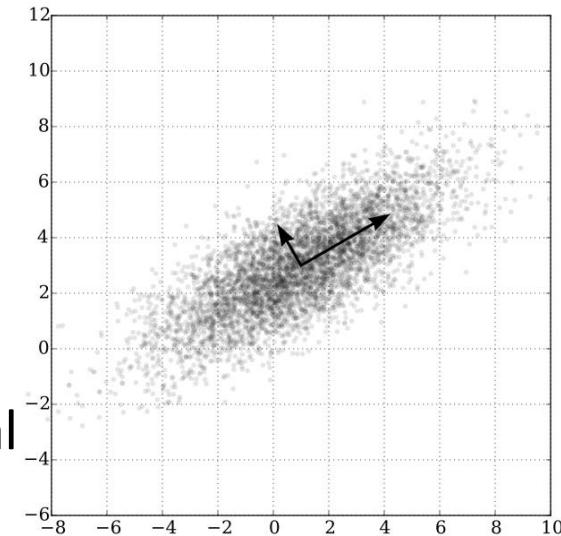
- PCA: Prosedur statistik yang menggunakan transformasi ortogonal untuk mengubah sekumpulan pengamatan variabel yang mungkin berkorelasi menjadi sekumpulan nilai variabel yang tidak berkorelasi linier yang disebut ***principal components***
- Data asli diproyeksikan ke ruang yang jauh lebih kecil, menghasilkan dimensionality reduction
- Metode: Temukan vektor eigen dari matriks kovarians, dan vektor eigen ini mendefinisikan ruang baru



Bola bergerak dalam garis lurus. Data dari tiga kamera mengandung banyak redundansi

Principal Component Analysis (Method)

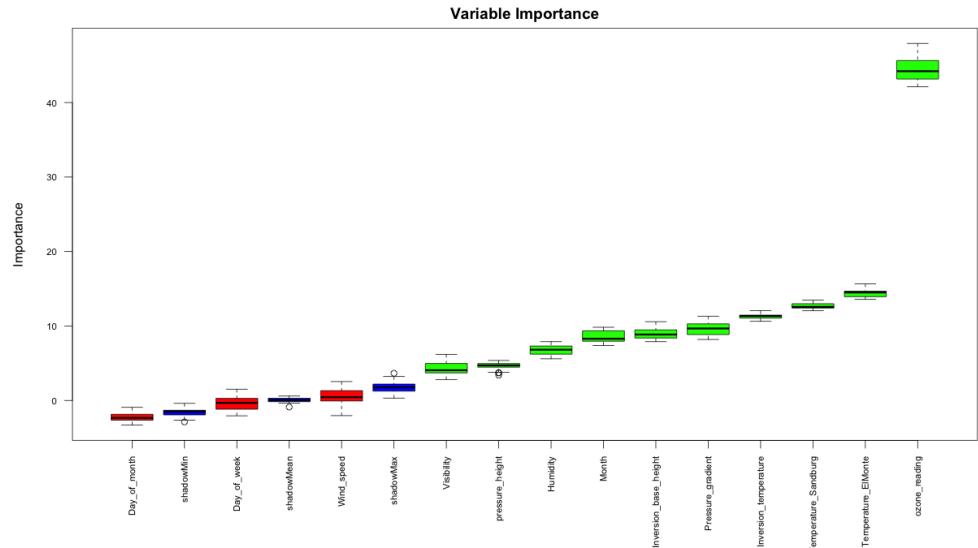
- Diberikan N vektor data dari n -dimensi, temukan $k \leq n$ vektor ortogonal (*principal components*) paling baik digunakan untuk merepresentasikan data
- Menormalkan data input: Setiap atribut berada dalam rentang yang sama
- Hitung k vektor ortonormal (satuan), yaitu, *principal components*
- Setiap data input (vektor) adalah kombinasi linier dari vector k principal component
- *principal components* diurutkan dalam urutan decreasing “significance” atau strength
- Karena komponen diurutkan, ukuran data dapat dikurangi dengan menghilangkan komponen yang lemah, yaitu, komponen dengan varians rendah (yaitu, menggunakan *principal components* terkuat, untuk merekonstruksi perkiraan yang baik dari data asli) Hanya berfungsi untuk data numerik



Ack. Wikipedia: Principal Component Analysis

Attribute Subset Selection

- ❑ Cara lain untuk mengurangi dimensi data
- ❑ Redundant attributes
 - ❑ Menduplikasi banyak atau semua informasi yang terkandung dalam satu atau beberapa atribut lainnya
 - ❑ Misalnya, harga pembelian suatu produk dan jumlah pajak penjualan yang dibayarkan
- ❑ Irrelevant attributes
 - ❑ Tidak berisi informasi yang berguna untuk tugas data mining yang ada
 - ❑ Misalnya. ID siswa seringkali tidak relevan dengan tugas memprediksi IPK-nya



Heuristic Search in Attribute Selection

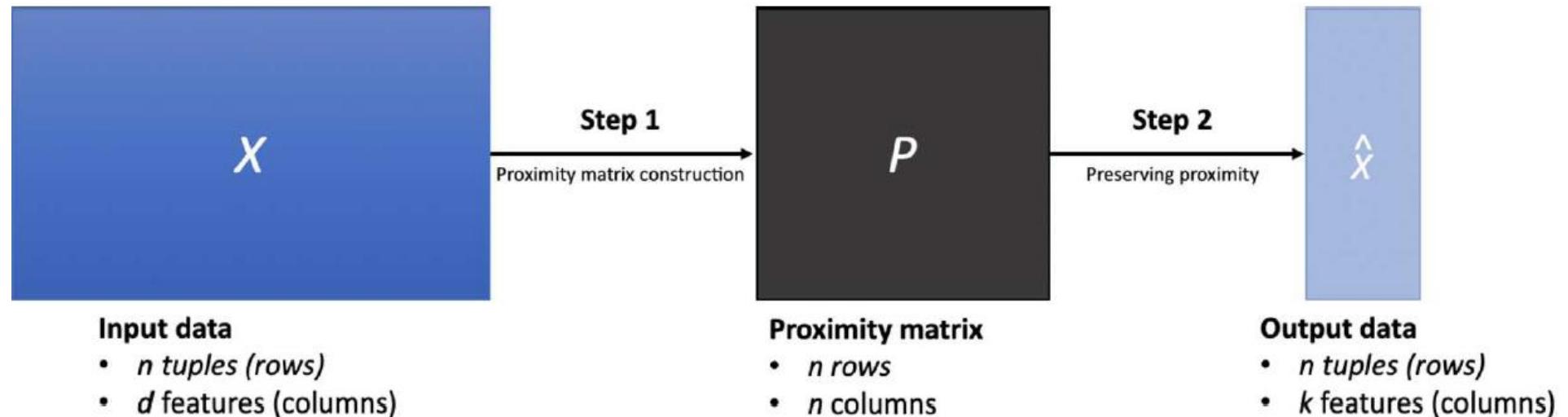
- ❑ Ada kemungkinan kombinasi atribut 2^d dari atribut d
- ❑ Metode Typical heuristic attribute selection :
 - ❑ Atribut terbaik di bawah asumsi independensi atribut: pilih berdasarkan uji signifikansi
 - ❑ Pemilihan fitur langkah demi langkah terbaik:
 - ❑ Atribut tunggal terbaik dipilih terlebih dahulu Kemudian atribut terbaik berikutnya berdasarkan kondisi pada atribut yang pertama, dan seterusnya
 - ❑ Step-wise attribute elimination:
 - ❑ Berulang kali menghilangkan atribut terburuk
 - ❑ Pemilihan dan eliminasi atribut gabungan terbaik
 - ❑ Optimal branch and bound:
 - ❑ menggunakan attribute elimination dan backtracking

Attribute Creation (Feature Generation)

- ❑ Buat atribut (fitur) baru yang dapat menangkap informasi penting dalam dataset lebih efektif daripada yang asli
- ❑ Tiga metodologi umum
 - ❑ Attribute extraction
 - ❑ Domain-specific
 - ❑ Mapping data ke new space
 - ❑ Misalnya Fourier transformation, wavelet transformation
 - ❑ Attribute construction
 - ❑ Combining features
 - ❑ Data discretization

Nonlinear Dimensionality Reduction Methods

- ❑ PCA adalah metode linier untuk dimensionality reduction
 - ❑ Setiap principal component adalah kombinasi linier dari atribut input asli
 - ❑ Bekerja dengan baik jika data input kira-kira mengikuti distribusi Gaussian atau membentuk beberapa klaster yang dapat dipisahkan secara linier
- ❑ Ketika data input tidak dapat dipisahkan secara linier, kita perlu membangun matriks kedekatan (P) dan mempelajari matriks baru dengan k fitur ($k \ll d$) yang mempertahankan kedekatan



Nonlinear Dimensionality Reduction (I): Kernel PCA (KPCA)

- ❑ Gunakan fungsi kernel $\kappa(\cdot)$ untuk membangun matriks kernel: $P(i, j) = \kappa(x_i, x_j)$, dan pelajari representasi dimensi rendah terbaik sehingga perkiraan matriks kedekatan \hat{P} sedekat mungkin dengan matriks kernel P
- ❑ Ini dapat diperoleh dengan menggunakan vektor eigen top-k dan nilai eigen dari matriks kernel P
- ❑ Typical kernel functions:
 - ❑ (1) polynomial kernel: $\kappa(x_i, x_j) = (1 + x_i \cdot x_j)^p$
 - ❑ (2) radial basis function (RBF): $\kappa(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$
- ❑ Jika kita memilih kernel linier: $\kappa(x_i, x_j) = x_i \cdot x_j$, KPCA degenerates to the standard PCA
- ❑ Rumus utama Kernel PCA vs. SNE (Stochastic neighborhood embedding)

	Step 1: Proximity Construction	Step 2: Preserving Proximity
KPCA	$P(i, j) = \kappa(x_i, x_j)$	$\min \sum_{i,j=1}^n (P(i, j) - \hat{P}(i, j))^2 = \ P - \hat{P}\ _{fro}^2$
SNE	$P(i, j) = \frac{e^{-d_{ij}^2}}{\sum_{l=1, l \neq i}^n e^{-d_{il}^2}}$	$\min \sum_{i=1}^n \text{KL}(P_i \hat{P}_i)$

Nonlinear Dimensionality Reduction (II): t-SNE

- SNE (Stochastic neighborhood embedding)
 - Construct a proximity matrix P using the formula: $P(i, j) = \frac{e^{-d_{ij}^2}}{\sum_{l=1, l \neq i}^n e^{-d_{il}^2}}$ where $d_{ij}^2 = \frac{\|x_i - x_j\|^2}{2\sigma^2}$
 - rep. the probability that x_j is the neighbor of x_i
 - Suppose we have learned the low-dimensional representations \hat{x}_i , we can compute another estimated proximity matrix in the similar way: $\hat{P}(i, j)$
 - We want to make the estimated proximity matrix \hat{P} to be as close as possible to P
 - That is, we want to minimize the overall K-L divergence, that is,

$$\hat{x}_i = \arg \min_{\hat{x}_i, (i=1, \dots, n)} \sum_{i=1}^n D_{KL}(P_i || \hat{P}_i)$$

Step 1: Proximity Construction

$$\text{KPCA} \quad P(i, j) = \kappa(x_i, x_j)$$

$$\text{SNE} \quad P(i, j) = \frac{e^{-d_{ij}^2}}{\sum_{l=1, l \neq i}^n e^{-d_{il}^2}}$$

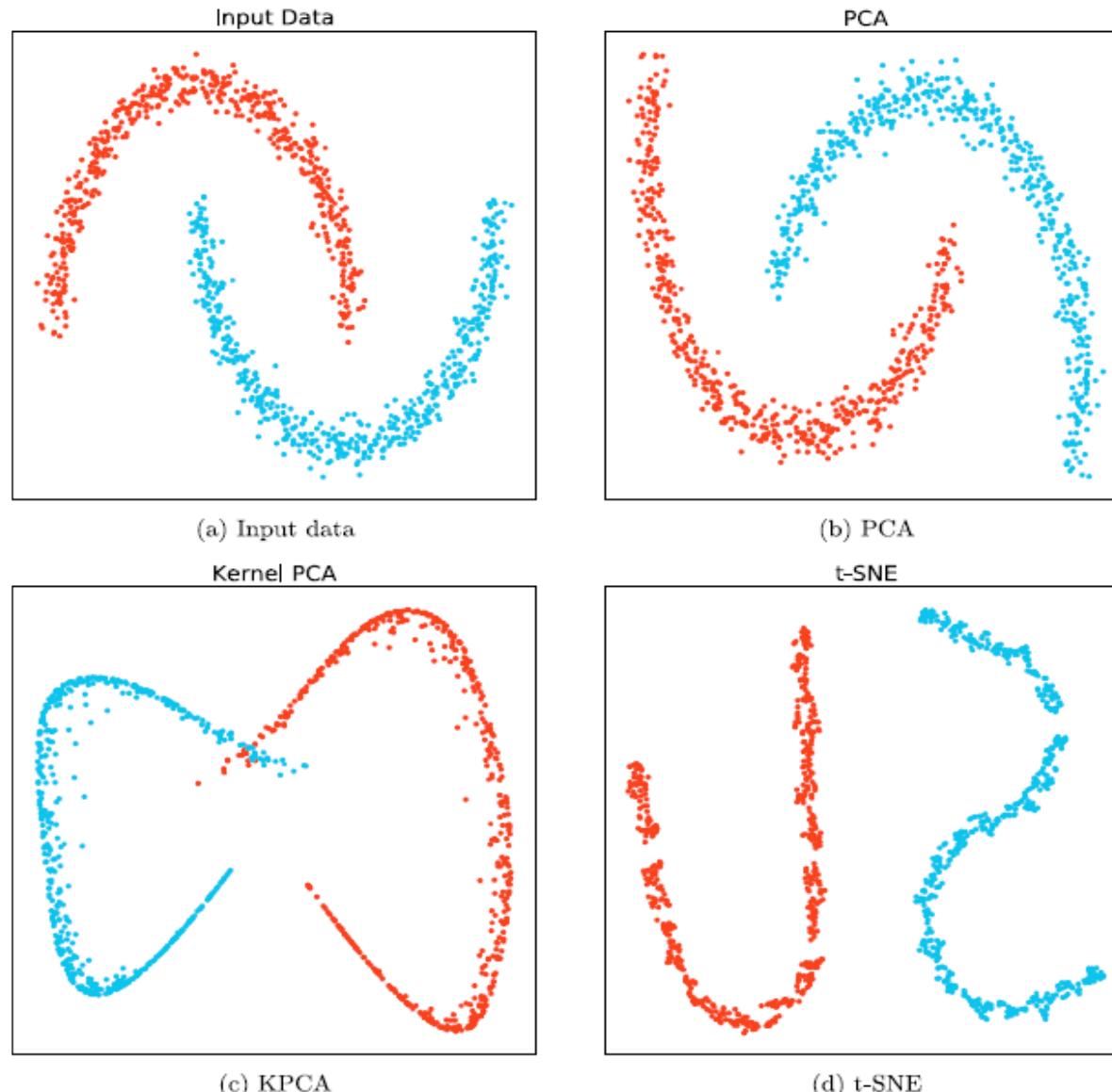
Step 2: Preserving Proximity

$$\min \sum_{i,j=1}^n (P(i, j) - \hat{P}(i, j))^2 = \|P - \hat{P}\|_{fro}^2$$

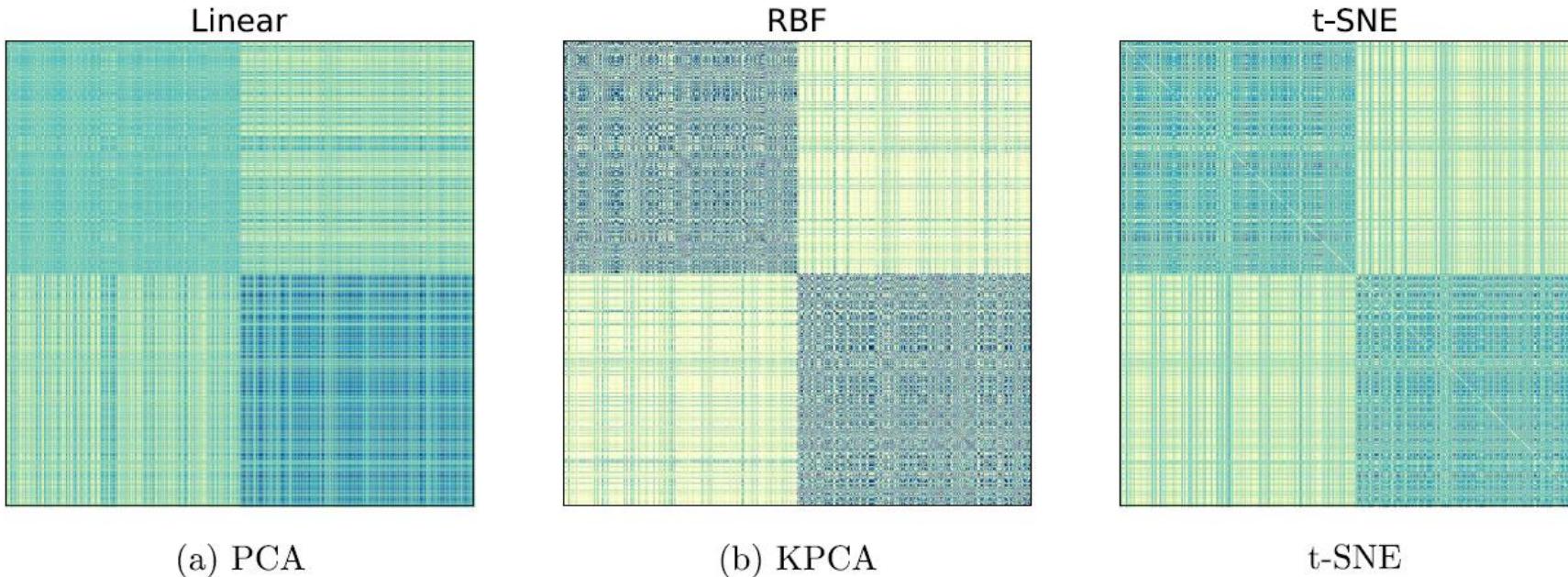
$$\min \sum_{i=1}^n \text{KL}(P_i || \hat{P}_i)$$

Example: Comparison on Nonlinear Data Points: Linear vs. Nonlinear Dimensional Reduction Methods

- Visualisasi: Contoh metode dimensionality reduction linear vs. nonlinear
- Diberikan kumpulan data input dalam ruang 2-D (Gbr. (a)): Titik data merah dan biru tidak dapat dipisahkan secara linier
- Transformasi PCA tidak dapat membuatnya dapat dipisahkan secara linier
- KPCA dapat membuat poin dapat dipisahkan secara linier
- t-SNE (NSE distribusi-t) dapat membuatnya dapat dipisahkan secara linier



Heatmap of the Proximity Matrices: Linear vs. Nonlinear Dimensional Reduction Methods



The heatmaps of the proximity matrices in PCA (a), KPCA (b), and t-SNE(c)

- ❑ Dua blok diagonal menunjukkan kedekatan dalam dua kelompok masing-masing
- ❑ Dua blok off-diagonal menunjukkan kedekatan antara data dari dua kluster
- ❑ Dengan metode nonlinier (KPCA dan t-SNE), kedekatan antara tuple data dari kluster yang sama jauh lebih tinggi daripada kedekatan antara tuple data dari kluster yang berbeda

Simpulan

- ❑ Data types and attribute types
 - ❑ Nominal, binary, ordinal, numerical, discrete vs. continuous attributes
- ❑ Statistics of data
 - ❑ Central tendency, dispersion, covariance and correlation, graphical displays
- ❑ Data quality measures, data cleaning, and data integration
- ❑ Data transformation: normalization, discretization, data compression and sampling
- ❑ Dimensionality reduction methodologies
 - ❑ Principal Component Analysis (PCA), attribute subset selection, and nonlinear dimensionality reduction