Universitas Esa Unggul

Smart, Creative and Entrepreneurial

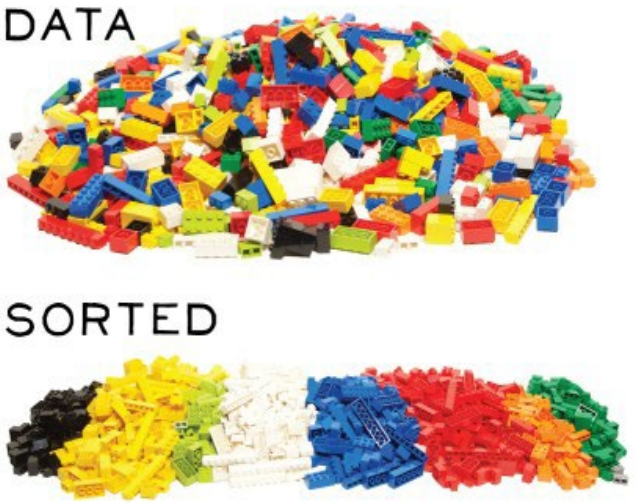
www.esaunggul.ac.id

PRAPROSES DATA
PERTEMUAN - 3

NOVIANDI
PRODI MIK | FAKULTAS ILMU-ILMU KESEHATAN

KEMAMPUAN AKHIR YANG DIHARAPKAN

Mahasiswa mengetahui *preprocessing* data , melakukan proses *cleaning* data, mampu menjelaskan konsep data integras, transformasi, reduksi, dan diskritisasi



LATAR BELAKANG PRAPROSES DATA

Tidak komplit

- Terdapat artribut yang kosong dikarenakan atribut tersebut tidak dapat diaplikasikan untuk semua kasus
- *Human/Hardware/Software problems*

Noisy

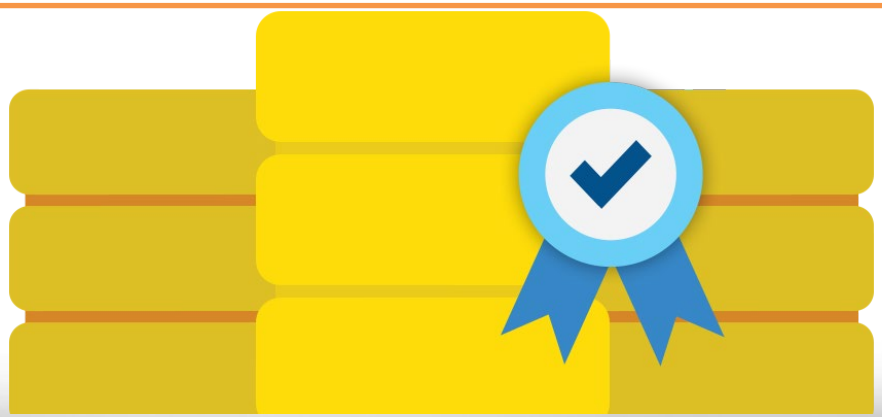
- Data mengandung error atau *outlier* karena terdapat kesalahan dalam penggunaan alat, kesalahan manusia atau komputer pada saat memasukkan data, eror dalam transmisi data

Tidak konsisten

- Format data berubah-ubah dikarenakan berasal dari sumber data yang berbeda. Contoh: Format tanggal

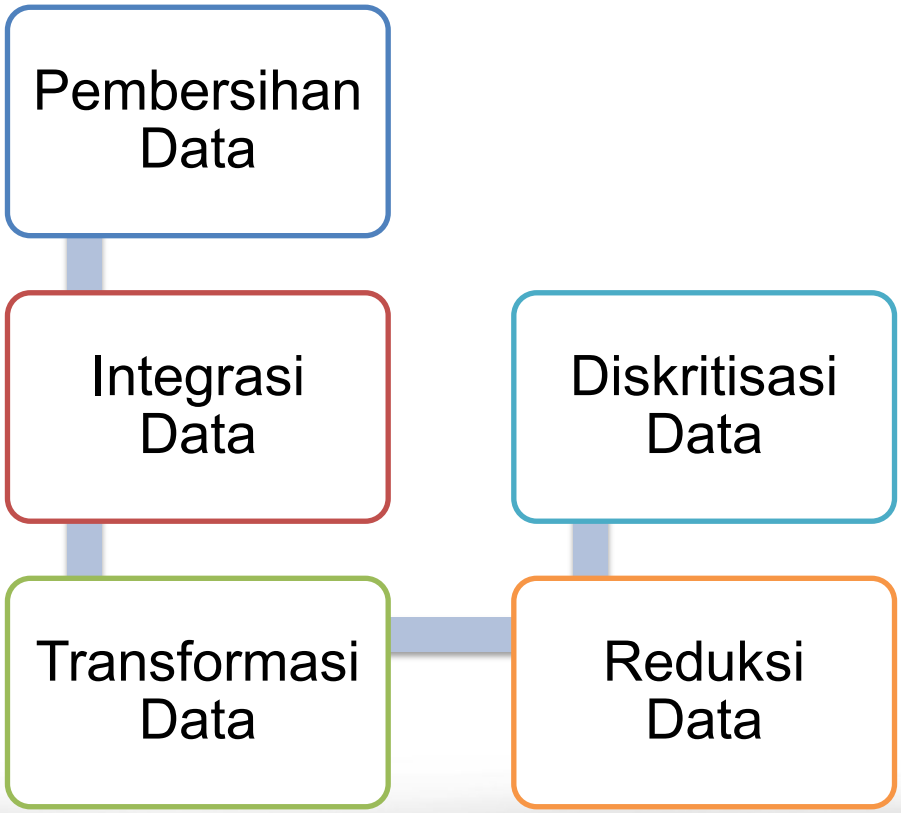
TUJUAN PRAPROSES

- Menghasilkan hasil *mining* yang berkualitas
- Data *warehouse* membutuhkan integrasi yang konsisten
- Data extraction, cleaning, and transformation merupakan salah satu tahapan untuk membangun gudang data



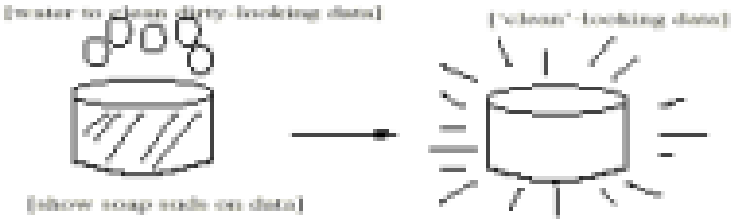
Sumber: www.syncsort.com/Syncsort/media/images/data-quality-hero-mobile.png

TAHAPAN PRAPROSES DATA

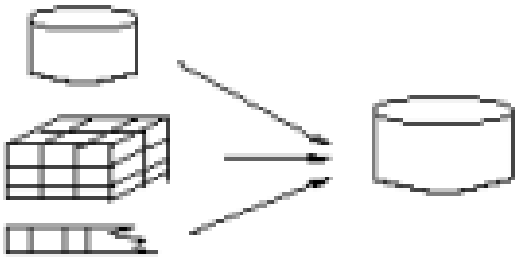


ILUSTRASI PRAPROSES DATA

Data Cleaning

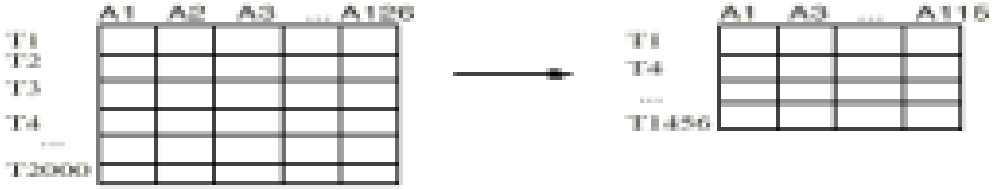


Data Integration



Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



PEMBERSIHAN DATA

Mengisi *missing value*

Meminimumkan *Noise*

Membetulkan data yang tidak konsisten

Mengindentifikasi /membuang outlier

https://developer.salesforce.com/resource/images/trailhead/badges/modules/trailhead_module_data_quality.png

MENGISI *MISSING VALUE*

- Mengabaikan record
- Menggunakan mean/median/modus dari atribut yang mengandung *missing value*
- Menggunakan nilai termungkin (Menerapkan regresi)

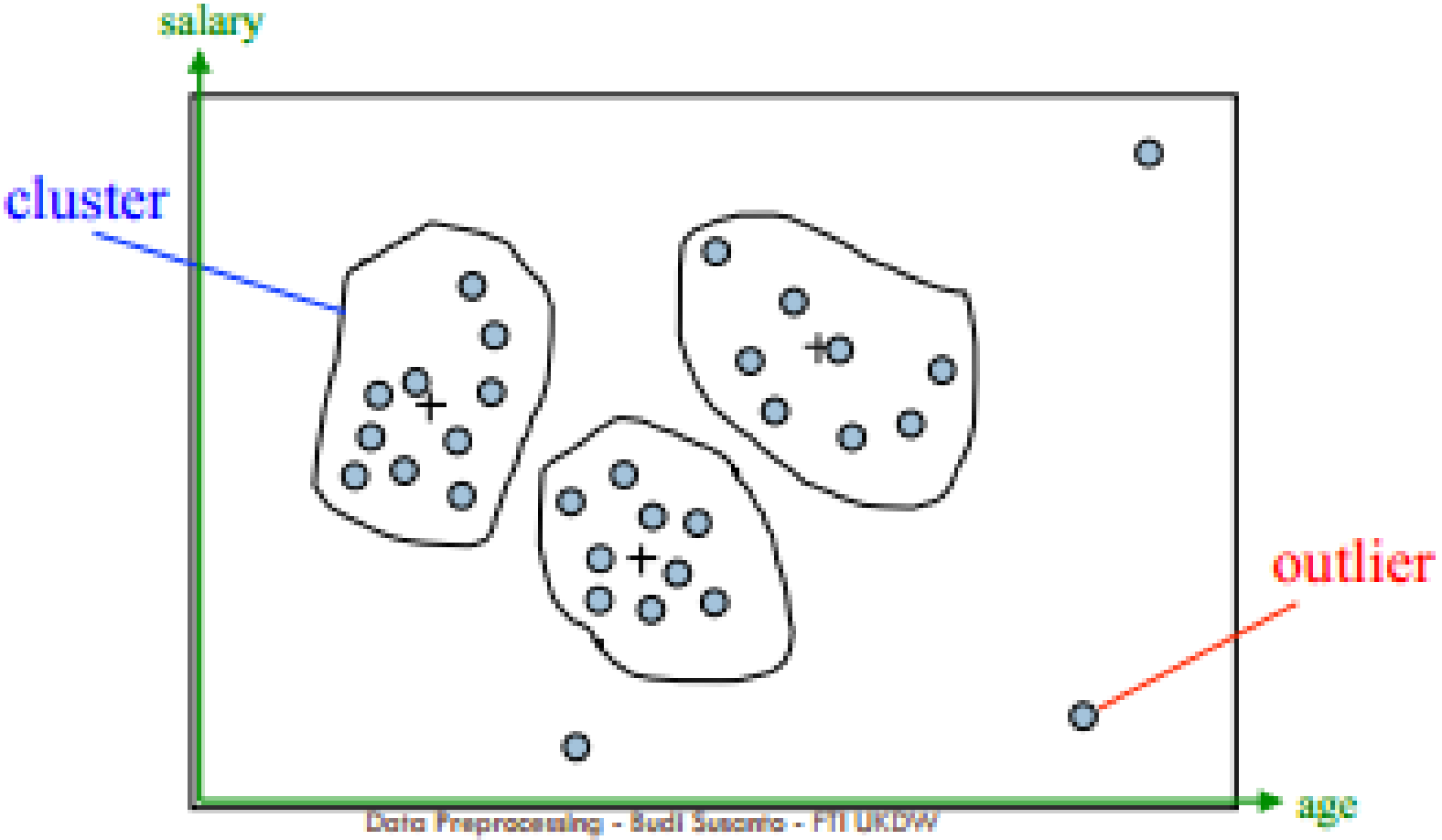
NOISY DATA

Cara mengetahui *outlier* : Clustering, Regresi Linear

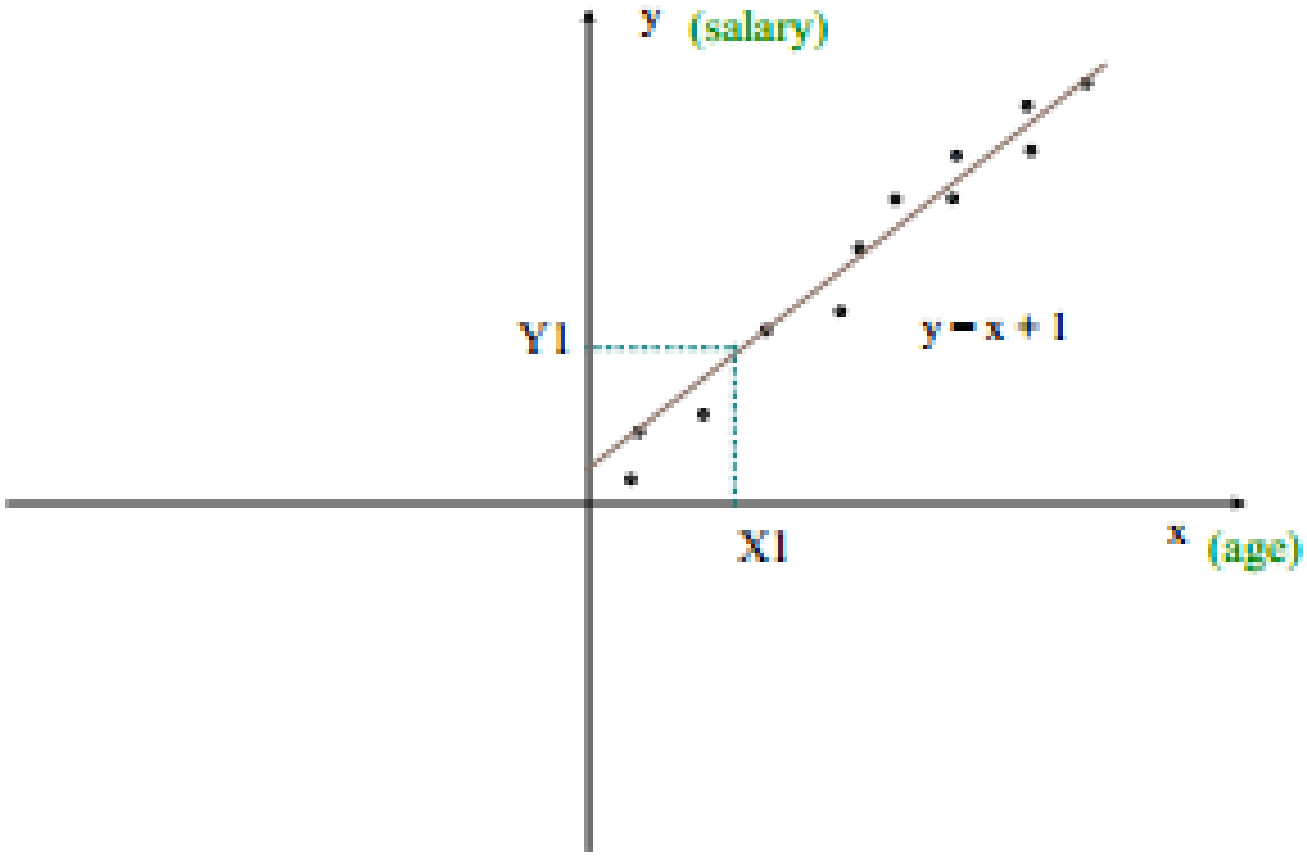
Binning

- Smoothing menggunakan Bin Means
- Smoothing menggunakan Bin Medians
- Smoothing menggunakan Bin Boundaries

MENDETEKSI *OUTLIER* DENGAN CLUSTERING



MENDETEKSI *OUTLIER* DENGAN REGRESI LINEAR



METODE BINING

Metode yang dilakukan untuk mengelompokkan data

Salah satu pendekatan diskritisasi

Urutan proses:

1. Urutkan data dari kecil ke besar (*ascending*)
2. Melakukan partisi data dalam bins menggunakan *equal-width* atau *equal-depth* (frekuensi)
3. Dapat di-*smoothing* menggunakan rata-rata, median, batasan, dsb.

METODE BINING

- Partisi *Equal-Width*

  Langkah-langkah membagi data ke dalam k interval ukuran yang sama. Lebar interval adalah

  w= (max-min)/k

- Partisi *Equal- depth*

  Membagi data ke dalam k kelompok dimana tiap k kelompok berisi jumlah yang sama

CONTOH PARTISI BINNING

Data: 0, 4, 12, 16 16, 18, 24, 26, 28

- Equal Width
  - BIN 1= 0,4
  - BIN 2= 12,16,16,18
  - BIN 3= 24,26,28
- Equal Depth
  - BIN 1= 0, 4, 12
  - BIN 2= 16,16,18
  - BIN 3= 24,26,28

Smoothing berdasarkan **rata-rata**: Semua nilai tiap bin diganti dengan rata-rata nilai tiap bin

Smoothing berdasarkan **batasan**: Setiap nilai bin diganti dengan nilai yang paling dekat dari batasan nilai. Batasan nilai terbentuk dari [min, max] tiap bin

INTEGRASI DATA

- Data dapat bersumber dari beberapa sumber
- Teknik-teknik:

ANALISIS KORELASI

ATRIBUT REDUDAN

DUPLIKASI

MENGATASI REDUNDASI PADA INTEGRASI DATA

PENYEBAB REDUNDANSI

- Atribut yang sama mempunyai nama yang berbeda pada database yang berbeda
- Satu atribut merupakan turunan dari atribut lainnya

Dapat dideteksi menggunakan analisis korelasi

Berhati-hati dalam menggabungkan data dari berbagai sumber untuk mengurangi redundasi

MENGATASI REDUNDASI PADA INTEGRASI DATA

Redudancy/ Duplicate :

Hubungan korelasi antar variabel dapat dilihat menggunakan rumus korelasi. Jika data numerik, hubungan korelasinya seperti dibawah ini:

$$r_{A,B} = \frac{\sum (A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum (AB) - n\,\overline{A}\,\overline{B}}{(n-1)\sigma_A \sigma_B}$$

Semakin besar hasil perhitungan tersebut, semakin tinggi korelasi. Jika hasil perhitungan tersebut =0 berarti independen. Jika kurang dari nol tidak independen

MENGATASI REDUNDASI PADA INTEGRASI DATA

Jika data kategorik, hubungan korelasinya seperti dibawah ini menggunakan chi-square:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Semakin besar chi-square, semakin tinggi korelasi. Jika hasil perhitungan tersebut =0 berarti independen. Jika kurang dari nol tidak independen

CONTOH SOAL MENGGUNAKAN CHI-SQUARE

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

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

TRANSFORMASI DATA

Tujuan diadakan tranformasi data agar data lebih efisien dalam proses data mining dan mungkin juga agar pola yang dihasilkan lebih mudah dipahami.

Hal-hal yang termasuk transformasi data:

- *Smoothing* : Menghapus noise dari data
- *Aggregation* : Ringkasan, Kontruksi data *cube*
- *Normalization* : Min-max, Z-Score, Decimal Scaling

TRANSFORMASI DATA

Normalization

a. *Min-max normalization:* menghasilkan [new_min,new_max]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Contoh soal:

Penghasilan berkisar dari $10,000 sampai $98,000 dinormalisasikan dari [0,1]. Sehingga untuk penghasilan sebesar $73,000 dipetakan ke $\frac{73,000-10,000}{98,000-10,000}$(1-0)+0=0.716

TRANSFORMASI DATA

Normalization

b. Min-max Z-score normalization : μ: mean, σ: standard deviation

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

Contoh soal:

Misal μ = 55,000, σ = 20,000. Maka, $\frac{73{,}000-55000}{20000}$= 0.9

TRANSFORMASI DATA

Normalization

c. Normalisasi pada skala desimal

$$v' = \frac{v}{10^j}$$

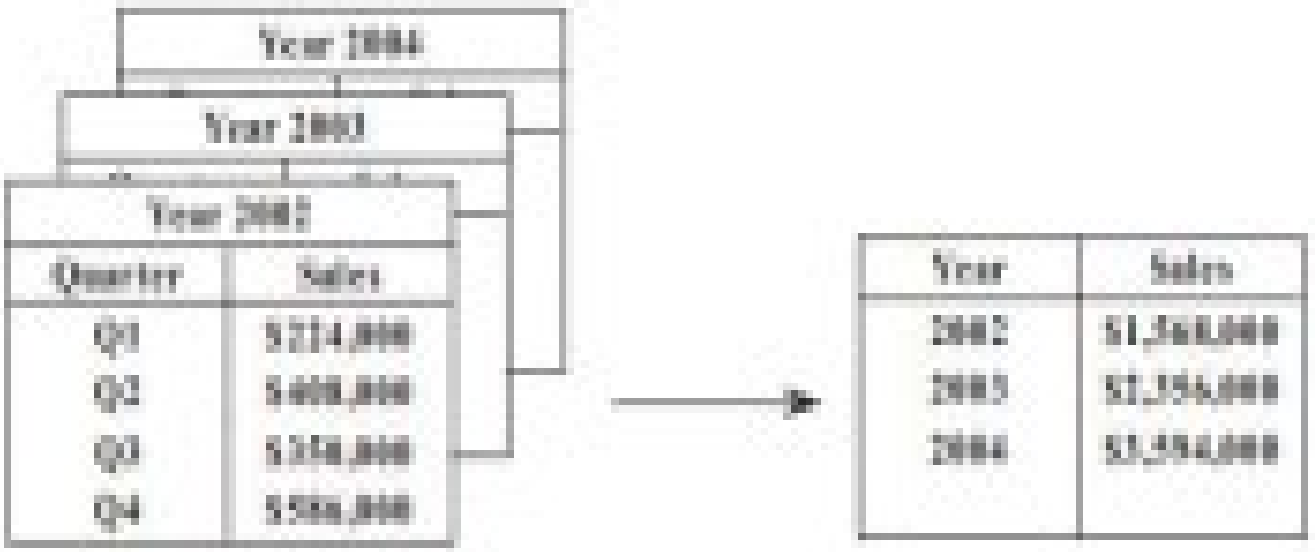
Dimana j adalah bilangan bulat terkecil sehingga Max(|v'|) < 1

REDUKSI DATA

Memperkecil volume tapi menghasilkan analasis data yang sama. Strategi- strategi data reduksi: Data cube aggregation, reduksi dimensi (menghapus atribut yang tidak penting), kompresi data, dsb.

REDUKSI DATA

DATA CUBE AGGREGATION



DISKRITISASI DATA

Terdapat tiga tipe atribut:

- Nominal = Nilai dari sekumpulan data yang tidak beraturan. Contoh: Warna, Profesi
- Ordinal = Nilai dari sekumpulan data yang terurut..
  Contoh: Ip, nomor antrian
- Kontinu = Nilai real seperti integer atau real number

Diskritisasi

Metode disktritisasi bisa dilakukan pada data kontinu.Tahap pertama, kita mengelompokkan nilai ke dalam interval. Setelah itu kita menggantikan nilai atribut dengan label atau interval.

Contoh:
Dataset (age, salary): (26;56,000),(28;70,000),(89;99,000)

TERIMA KASIH ☺