

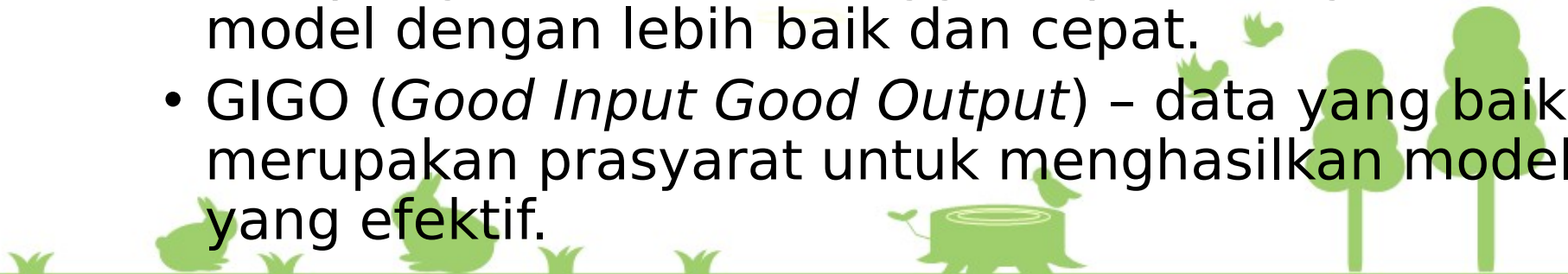
Data Preparation

Big Data Analytics



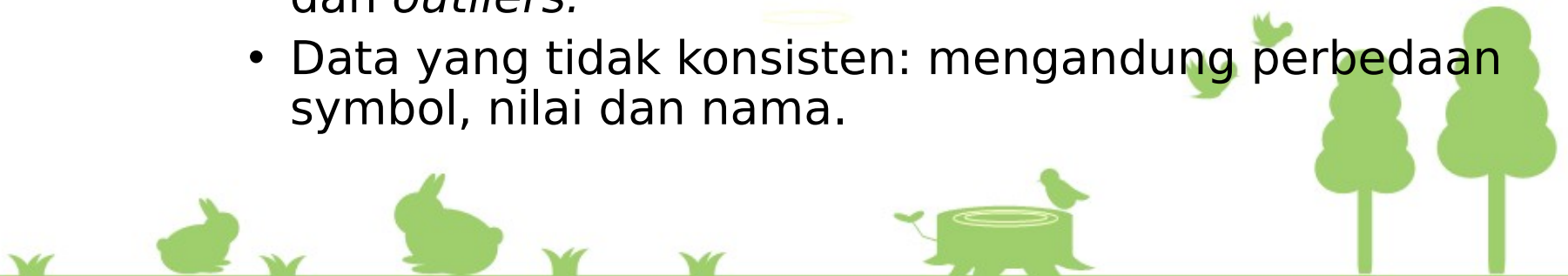
Pengenalan *Data Preparation* (1)

- Kenapa *data preparation* itu dibutuhkan?
 - Untuk mengurangi kesalahan data atau mendeteksi anomali data sedini mungkin.
 - Kesalahan data dan anomali data yang minimal akan meningkatkan *correctness* dan akurasi hasil pengolahan data.
 - *Data preparation* juga berarti mempersiapkan alat pengolah data sehingga dapat menghasilkan model dengan lebih baik dan cepat.
 - GIGO (*Good Input Good Output*) – data yang baik merupakan prasyarat untuk menghasilkan model yang efektif.



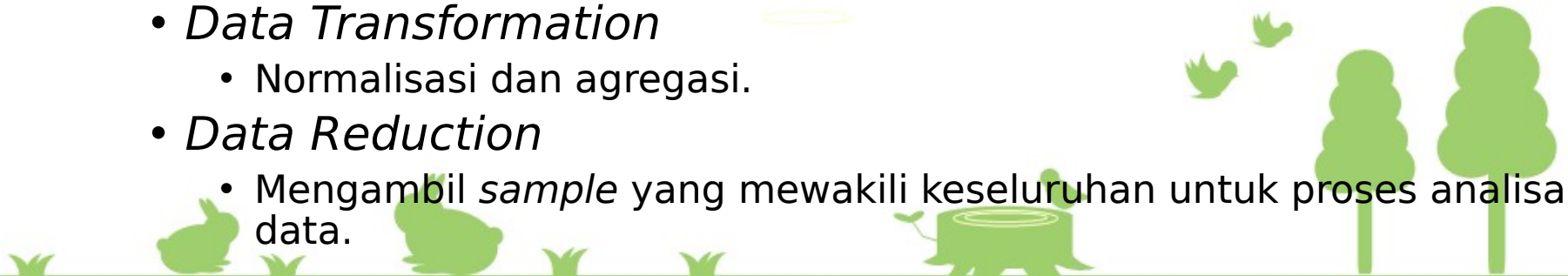
Pengenalan *Data Preparation* (2)

- *Data preparation* juga diperlukan karena:
 - Suatu alat atau aplikasi pengolah data membutuhkan data dalam format tertentu.
 - Tipikal data dari dunia nyata yang mengandung:
 - Data yang tidak lengkap: Adanya nilai kosong, kekurangan atribut yang penting, atau hanya memiliki data agregat.
 - Data yang “ribut”: mengandung banyak kesalahan data dan *outliers*.
 - Data yang tidak konsisten: mengandung perbedaan symbol, nilai dan nama.



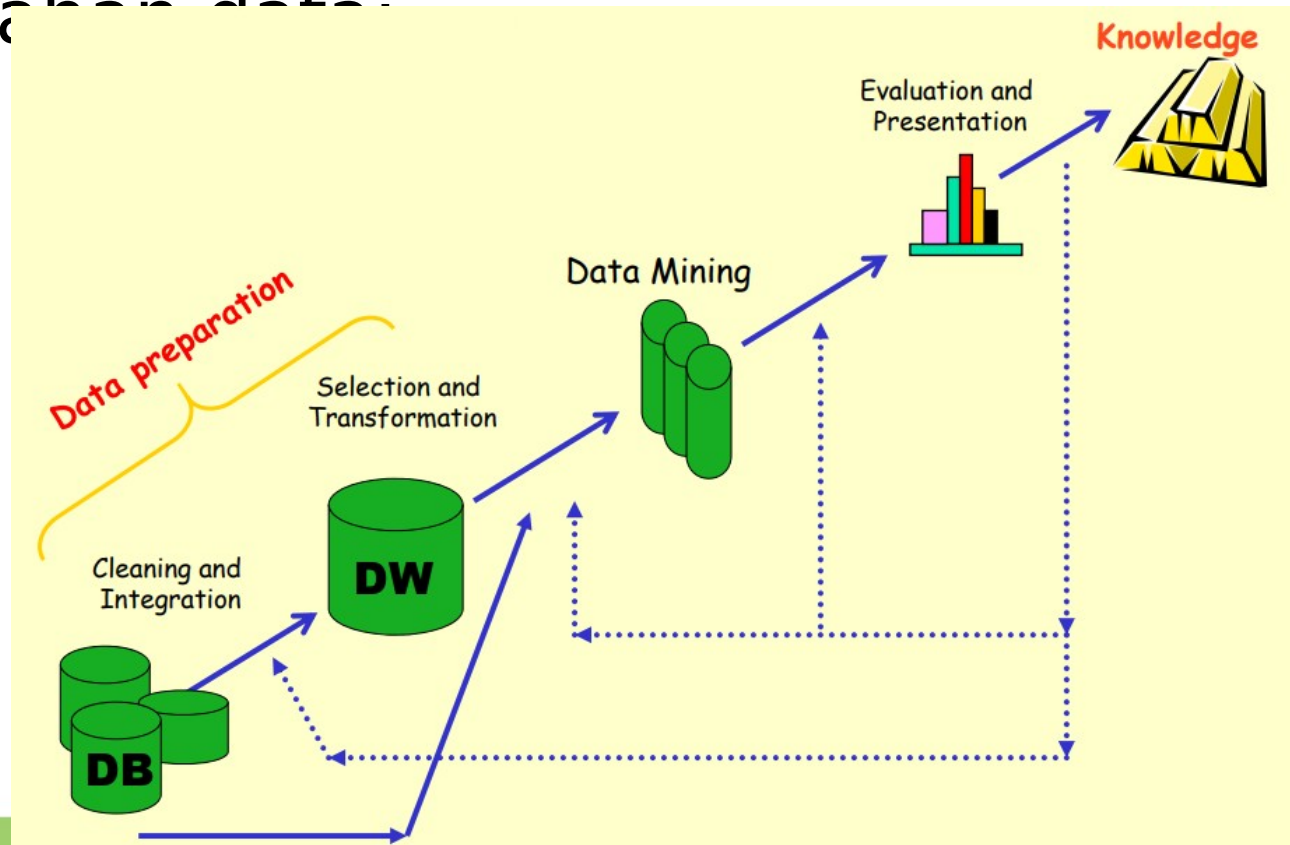
Pengenalan *Data Preparation* (3)

- Tugas-tugas utama pada *data preparation*:
 - *Data discretization*
 - Pengurangan fitur data hingga bagian terpenting, khususnya untuk data angka.
 - *Data Cleaning*
 - Mengisi nilai-nilai yang hilang, membersihkan data-data “noisy”, mendeteksi atau menghilangkan *outliers*, dan mengatasi inkonsistensi data.
 - *Data Integration*
 - Integrasi data dari berbagai sumber.
 - *Data Transformation*
 - Normalisasi dan agregasi.
 - *Data Reduction*
 - Mengambil *sample* yang mewakili keseluruhan untuk proses analisa data.



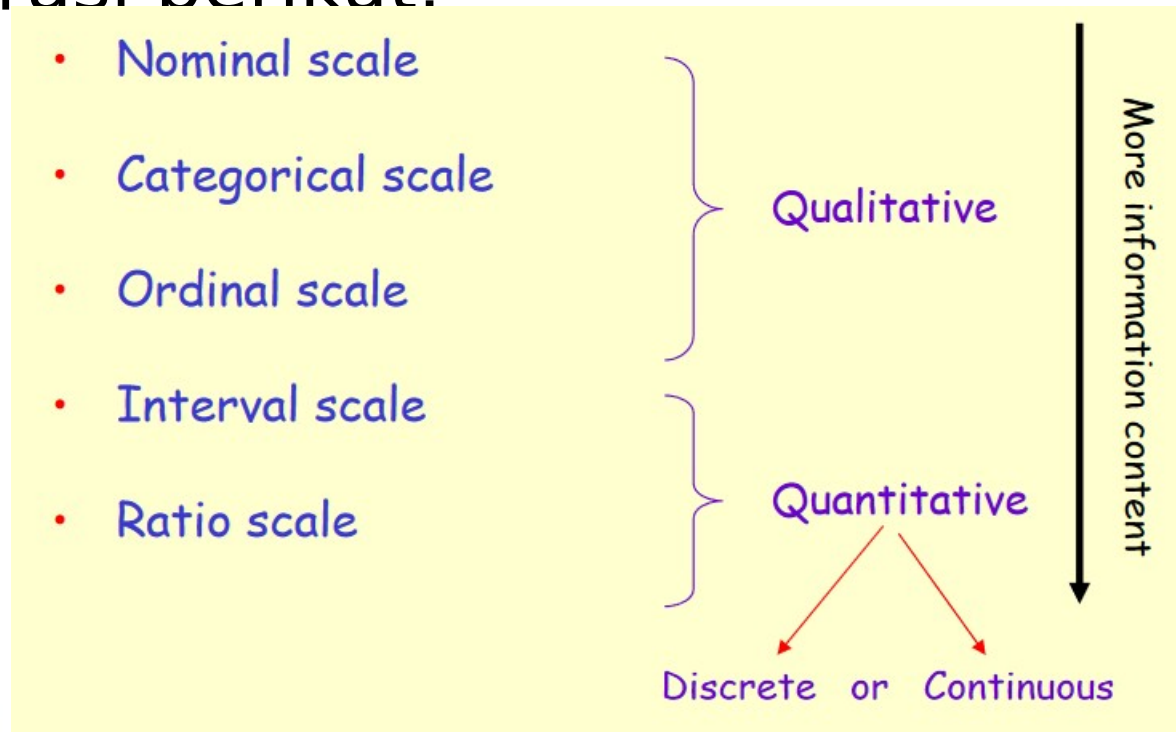
Pengenalan *Data Preparation* (4)

- Posisi *data preparation* dalam tahapan pengolahan data

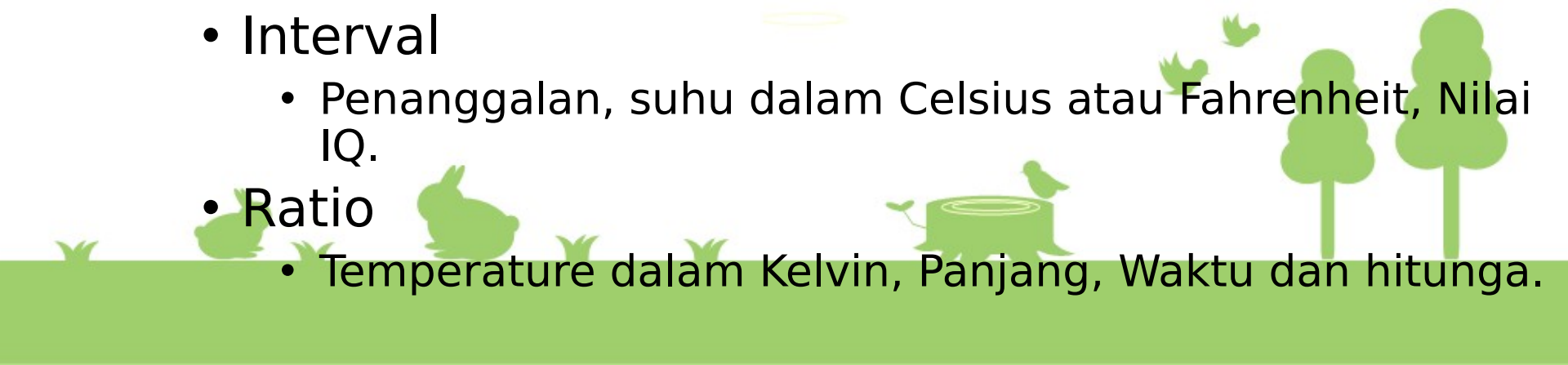


Tipe Data (1)

- Tipe-tipe pengukuran dapat dilihat dari ilustrasi berikut:



Tipe Data (2)

- Contoh-contoh tipe pengukuran:
 - Nominal:
 - ID, Nama
 - Categorical
 - Warna mata, kode pos, propinsi
 - Ordinal
 - Ranking, peringkat, tinggi dalam satuan (tinggi, pendek)
 - Interval
 - Penanggalan, suhu dalam Celsius atau Fahrenheit, Nilai IQ.
 - Ratio
 - Temperature dalam Kelvin, Panjang, Waktu dan hitunga.
- 

Tipe Data (3)

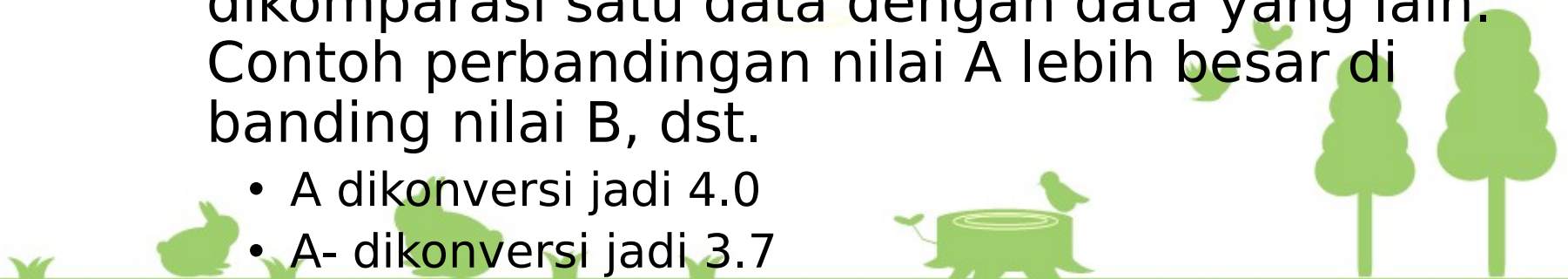
- Contoh tipe-tipe pengukuran:

Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
1	Sunny	85	85	Light	No
2	Sunny	80	90	Strong	No
3	Overcast	83	86	Light	Yes
4	Rain	70	96	Light	Yes

Day	Outlook	Temperature	Humidity	Wind	PlayTennis?
5	Rain	Hot	High	Light	No
6	Rain	Hot	High	Strong	No
7	Overcast	Hot	High	Light	Yes
8	Sunny	Hot	High	Light	Yes
9	Sunny	Mild	High	Light	Yes
10	Rain	Cool	Normal	Light	Yes
11	Sunny	Cool	Normal	Strong	No
12	Overcast	Cool	Normal	Strong	Yes
13	Overcast	Mild	High	Light	No
14	Rain	Cool	Normal	Light	Yes
		Normal	Light	Yes	
		Mild	Normal	Strong	Yes
		Mild	High	Strong	Yes
		Hot	Normal	Light	Yes
		Mild	High	Strong	No

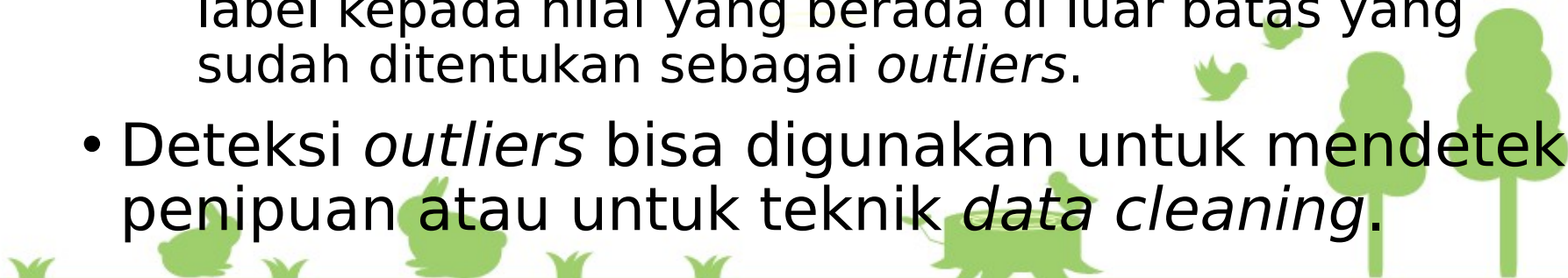
Tipe Data (4)

- Konversi Data
 - Diperlukan bila suatu data dibutuhkan oleh aplikasi yang berbeda dalam tipe pengukuran yang berbeda.
 - Contoh, aplikasi A membutuhkan data C dalam bentuk numeric, aplikasi B membutuhkan data C dalam bentuk ordinal.
 - Diperlukan bila suatu data ordinal akan dikomparasi satu data dengan data yang lain. Contoh perbandingan nilai A lebih besar dibanding nilai B, dst.
 - A dikonversi jadi 4.0
 - A- dikonversi jadi 3.7
 - B+ dikonversi jadi 3.3
 - B dikonversi jadi 3.0



Outliers (1)

- *Outliers* adalah nilai-nilai yang berada di luar *range* data.
 - *Outliers* adalah sebuah data yang berada jauh dari kelompok data sehingga menimbulkan kecurigaan bahwa data tersebut berasal dari metode atau sumber data yang berbeda.
- *Outlier* bisa dideteksi dengan cara:
 - Membuat standarisasi observasi dan memberikan label kepada nilai yang berada di luar batas yang sudah ditentukan sebagai *outliers*.
- Deteksi *outliers* bisa digunakan untuk mendeteksi penipuan atau untuk teknik *data cleaning*.



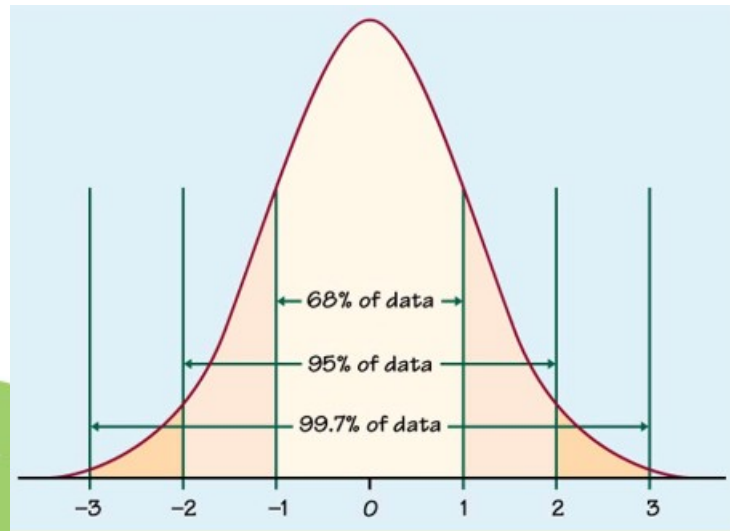
Outliers (2)

- Solusi untuk mengatasi *outliers* adalah:
 - Tidak melakukan apa-apa.
 - Menerapkan batas atas dan batas bawah nilai dari suatu observasi.
 - Mengatasi dengan teknik binning.
 - Teknik binning adalah teknik mengubah data yang *continuous* menjadi data diskrit.



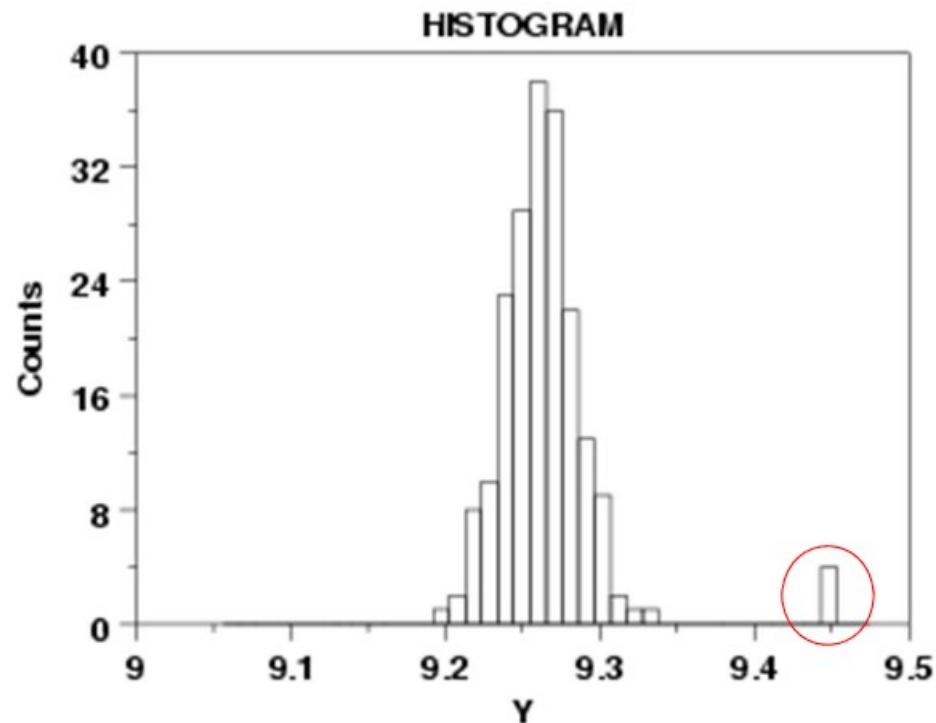
Outliers (3)

- Cara mendeteksi *outliers*:
 - Univariate
 - Hitung mean dan standard deviation dari sekumpulan data. Untuk $k=2$ dan 3 , data x adalah *outlier* bila berada di luar batas (asumsi distribusi normal).



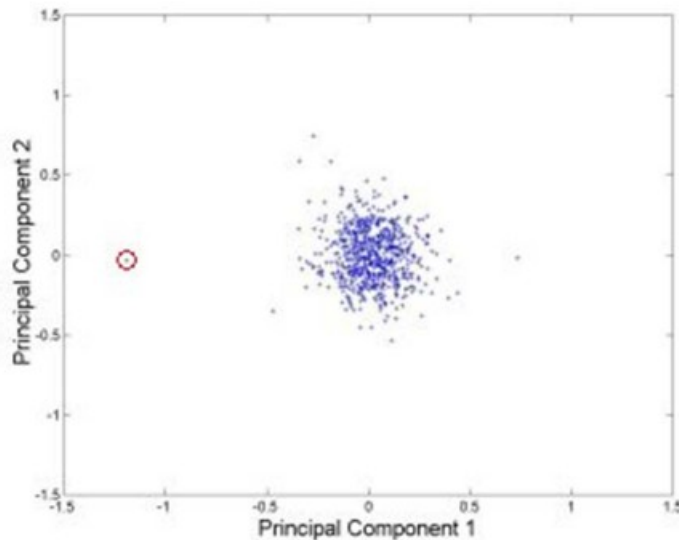
Outliers (4)

- Ilustrasi deteksi *outlier* dengan data Univariate

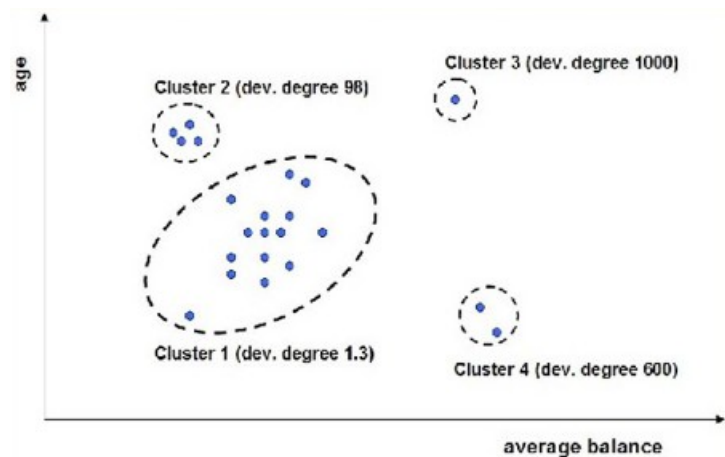


Outliers (5)

- Mendeteksi *outlier* untuk data *Multivariate*:
 - Menggunakan teknik *clustering*, dimana *cluster* dengan jumlah data yang kecil adalah *outliers*.



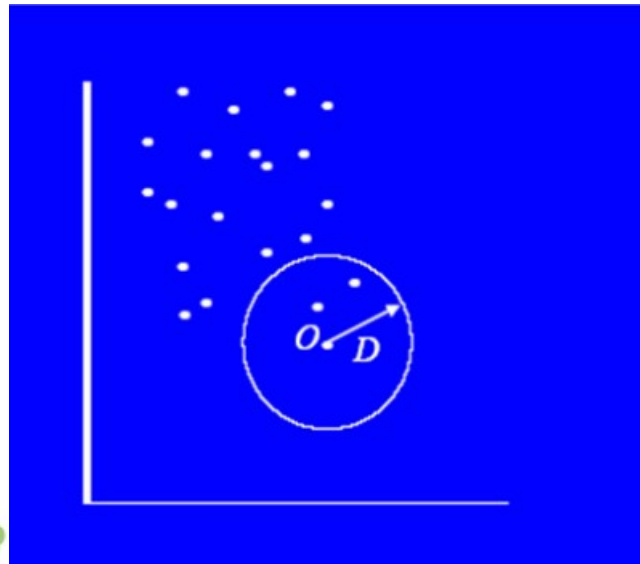
Data Awal



Grafik dengan cluster outlier

Outliers (5)

- Mendeteksi *outlier* untuk data *Multivariate*:
 - Menggunakan teknik berdasarkan jarak.
 - Sebuah data dengan sedikit data di sekitarnya (dalam himpunan D) dikategorikan sebagai *outliers*.



Transformasi Data (1)

- Transformasi data yang paling umum adalah normalisasi.
- Pada metode berbasis jarak (*distance-based method*), normalisasi mencegah atribut dengan rentang yang besar menyebabkan atribut dengan rentang kecil menjadi tidak “terlihat”.
- Metode normalisasi:
 - Min-max normalization
 - Z-score normalization
 - Normalization dengan *decimal scaling*



Transformasi Data (2)

- Formula normalisasi untuk:

- Min-max normalization:

$$v' = \frac{v - \min_v}{\max_v - \min_v} (\text{new_max}_v - \text{new_min}_v) + \text{new_min}_v$$

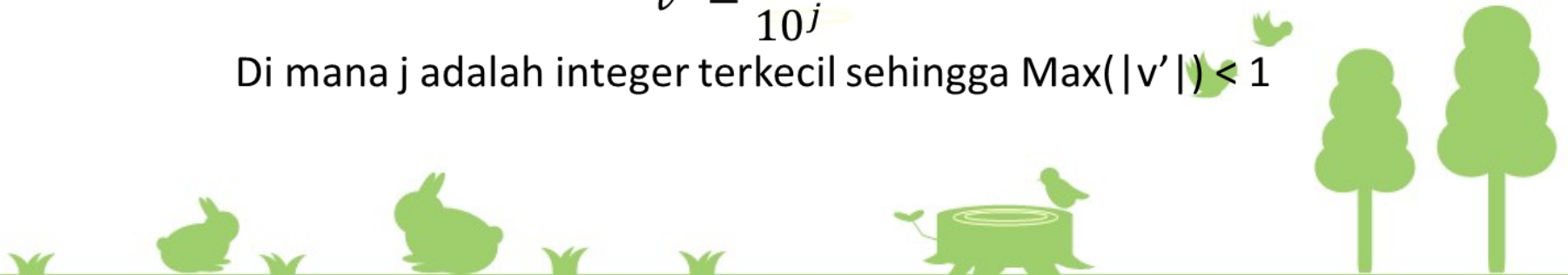
- Z-score normalization:

$$v' = \frac{v - \bar{v}}{\sigma_v}$$

- Normalisasi dengan *decimal scaling*:

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

Di mana j adalah integer terkecil sehingga $\text{Max}(|v'|) < 1$

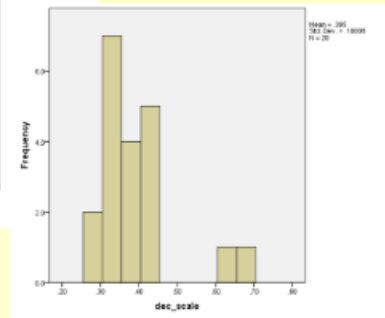
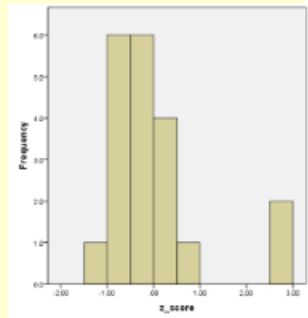
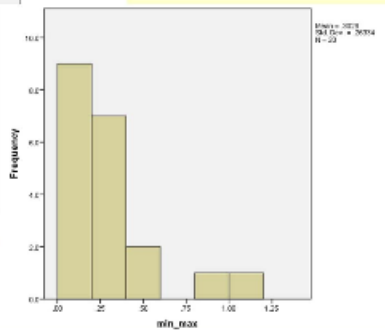
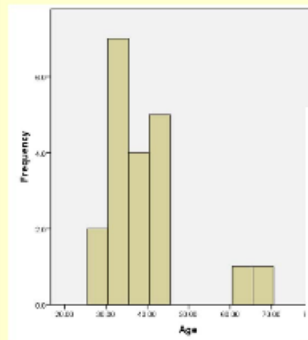


Transformasi Data (3)

- Contoh normalisasi

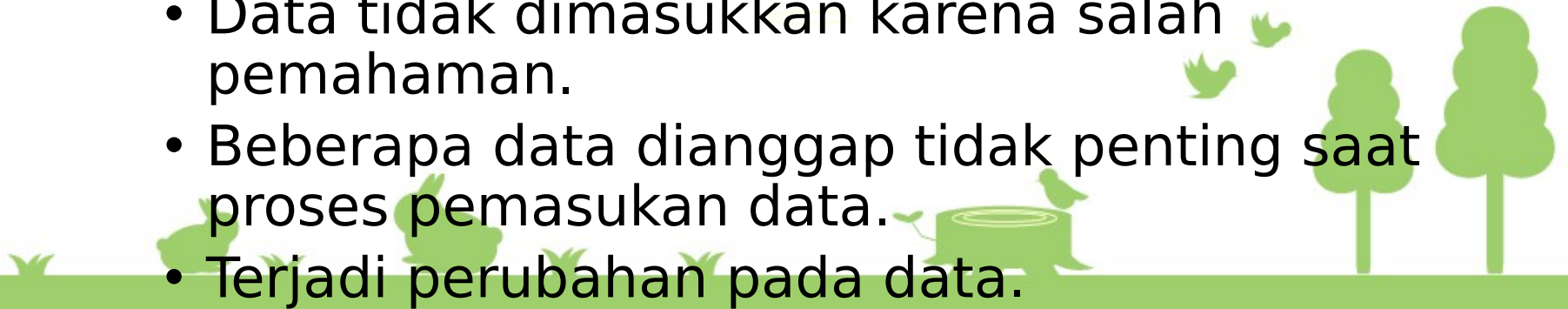
Age	min-max (0-1)	z-score	dec. scaling
44	0.421	0.450	0.44
35	0.184	-0.450	0.35
34	0.158	-0.550	0.34
34	0.158	-0.550	0.34
39	0.289	-0.050	0.39
41	0.342	0.150	0.41
42	0.368	0.250	0.42
31	0.079	-0.849	0.31
28	0.000	-1.149	0.28
30	0.053	-0.949	0.3
38	0.263	-0.150	0.38
36	0.211	-0.350	0.36
42	0.368	0.250	0.42
35	0.184	-0.450	0.35
33	0.132	-0.649	0.33
45	0.447	0.550	0.45
34	0.158	-0.550	0.34
65	0.974	2.548	0.65
66	1.000	2.648	0.66
38	0.263	-0.150	0.38

28	minimun
66	maximum
39.50	avgerage
10.01	standard deviation



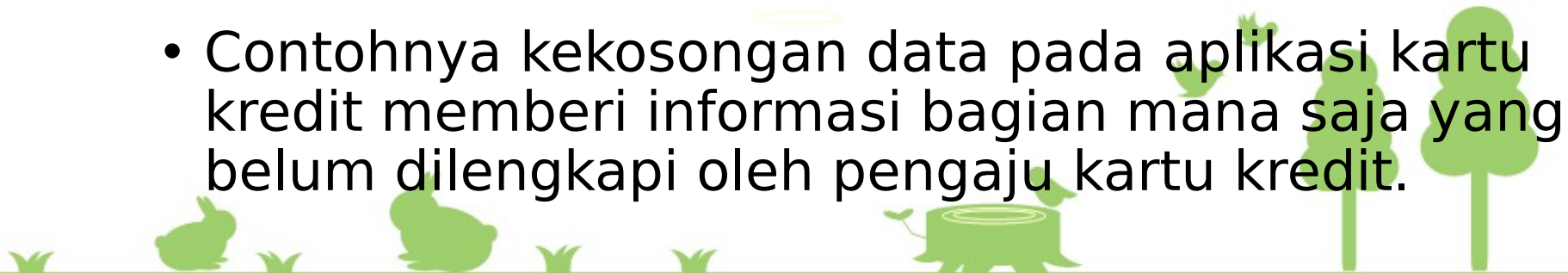
Kekosongan Data (1)

- Data tidak selalu tersedia.
 - Contoh, data pelanggan yang tidak mencantumkan data gaji pelanggan.
- Kekosongan data bisa disebabkan oleh:
 - Kesalahan alat
 - Tidak konsisten dengan data yang lain sehingga terhapus saat proses penyimpanan.
 - Data tidak dimasukkan karena salah pemahaman.
 - Beberapa data dianggap tidak penting saat proses pemasukan data.
 - Terjadi perubahan pada data.

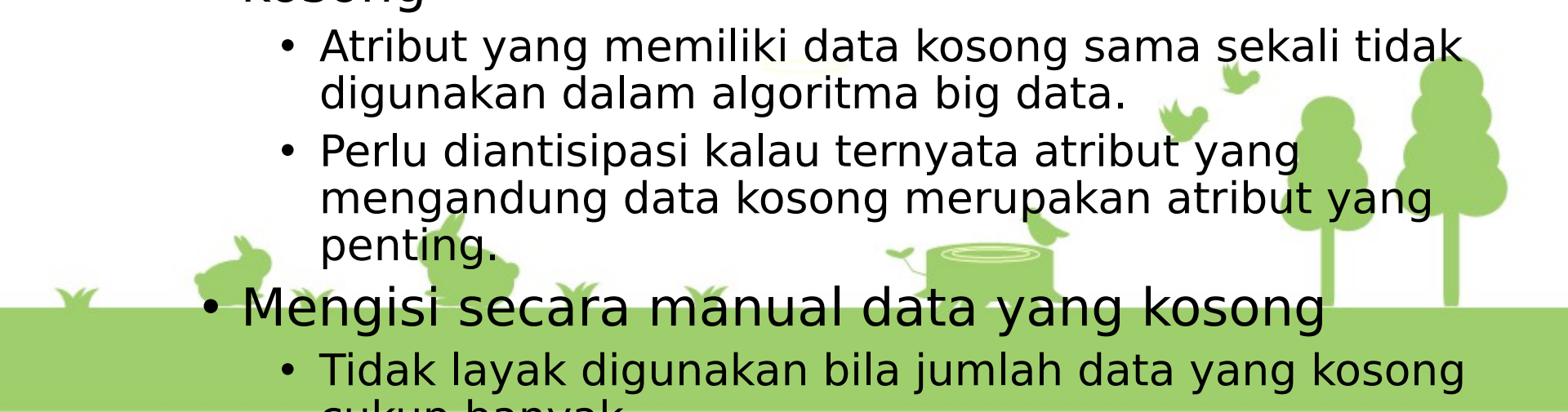


Kekosongan Data (2)

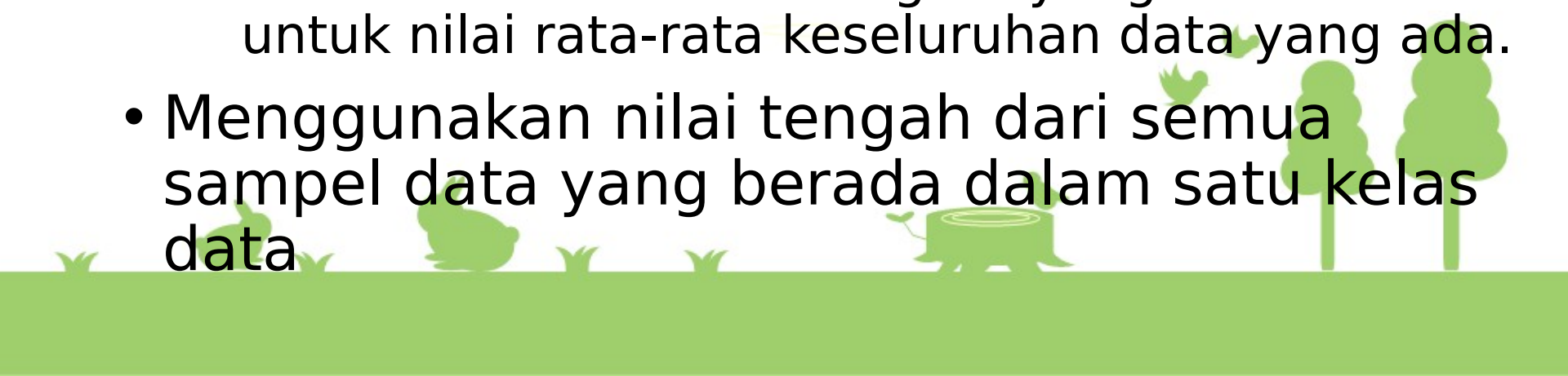
- Kekosongan data harus diantisipasi.
- Bagi beberapa metode big data, ada yang mengabaikan kekosongan data, ada juga yang menggunakan metric untuk mengganti nilai data yang kosong.
- Di lain pihak, kekosongan data bisa memberi informasi tertentu.
 - Contohnya kekosongan data pada aplikasi kartu kredit memberi informasi bagian mana saja yang belum dilengkapi oleh pengaju kartu kredit.



Kekosongan Data (3)

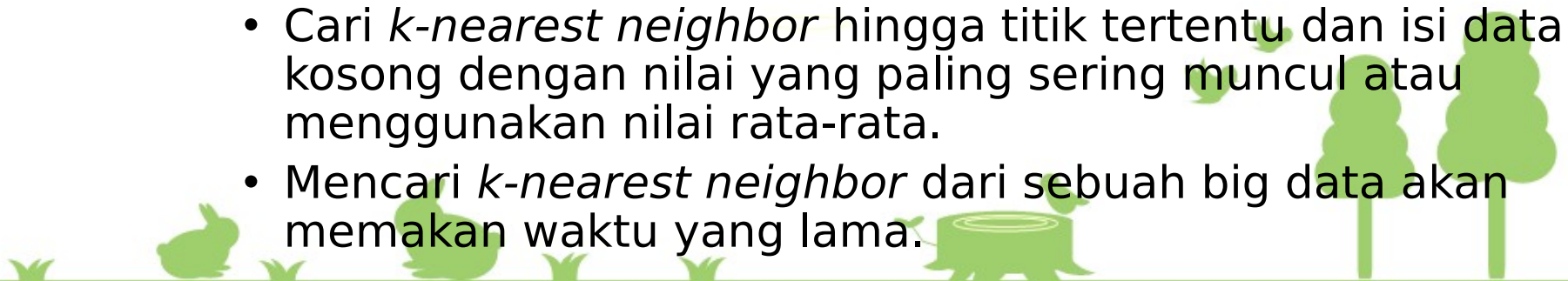
- Cara menangani kekosongan data:
 - Mengabaikan data yang kosong
 - Tidak efektif di saat persentase data yang hilang pada setiap atribut memiliki variasi yang besar sehingga bisa mengarah kepada ketidakcukupan data atau pengambilan sampel yang bias.
 - Mengabaikan atribut yang mengandung data kosong
 - Atribut yang memiliki data kosong sama sekali tidak digunakan dalam algoritma big data.
 - Perlu diantisipasi kalau ternyata atribut yang mengandung data kosong merupakan atribut yang penting.
 - Mengisi secara manual data yang kosong
 - Tidak layak digunakan bila jumlah data yang kosong cukup banyak.
- 

Kekosongan Data (4)

- Menggunakan konstanta untuk mengisi data kosong
 - Contohnya konstanta “unknown”
 - Cara ini bisa membuat kategori data yang baru.
 - Menggunakan nilai tengah dari suatu atribut untuk mengisi data kosong
 - Cara ini memiliki efek negatif yang minimum untuk nilai rata-rata keseluruhan data yang ada.
 - Menggunakan nilai tengah dari semua sampel data yang berada dalam satu kelas data
- 

Kekosongan Data (5)

- Menggunakan *most probable value* untuk mengisi kekosongan data.
 - Menggunakan teknik *inference-based* seperti formula Bayesian atau *decision tree*
 - Identifikasi hubungan diantara variabel
 - Teknik Linear regression, multiple linear regression, dan nonlinear regression.
- Teknik estimasi Nearest-Neighbour
 - Cari *k-nearest neighbor* hingga titik tertentu dan isi data kosong dengan nilai yang paling sering muncul atau menggunakan nilai rata-rata.
 - Mencari *k-nearest neighbor* dari sebuah big data akan memakan waktu yang lama.



Kekosongan Data (6)

- Langkah-langkah menangani kekosongan data perlu memperhatikan:
 - Hindari pengisian data kosong yang menyebabkan penambahan bias atau distorsi bagi data yang sudah ada.



Redudansi Data

- Redudansi data bisa terjadi di saat integrase database
 - Atribut yang sama bisa memiliki nama yang berbeda di database yang berbeda.
 - Satu atribut merupakan atribut yang didapat dari hasil komputasi atribut yang lain. Contohnya: perhitungan gaji bulanan.
- Redudansi data untuk atribut numeric bisa dideteksi menggunakan analisis korelasi.

$$r_{xy} = \frac{\frac{1}{N-1} \cdot \sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^N (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^N (y_i - \bar{y})^2}} \quad (-1 \leq r_{xy} \leq 1)$$