MACHINE LEARNING

Preparation Data

Magister Teknik Informatika

Dr. Chairani, S.Kom., M.Eng

IIB DARMAJAYA, 2023/2024

Subject

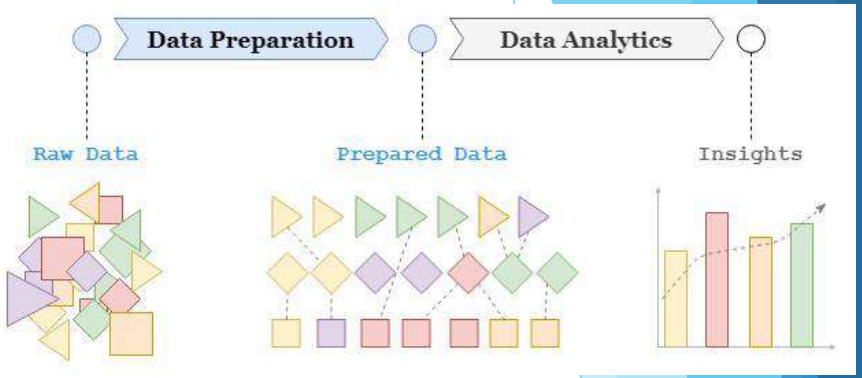
- Bagian Pertama dari Tiga, materi Data Preparation
- Berfokus pada Penentuan Objek Data dan Pembersihan Data
- Konteksnya yaitu:
  - strategi pembersihan data kotor (noise, bias, missing value, outlier, dll)
  - pengecekan kualitas dan tingkat kecukupan data
- Dilanjutkan dengan:
  - transformasi data (modul 9) dan
  - konstruksi data (modul 10)
- Pengetahuan dan pemahaman akan data preparation menjadi syarat mutlak utk mengahasilkan model prediksi yang optimal.

Data Preparation dalam CRISP-DM

- Akronim dari: **CR**oss **I**ndustry **S**tandard **P**rocess **D**ata **M**ining
- Metodologi umum untuk data mining, analitik, dan proyek data sains, berfungsi menstandarkan proses data mining lintas industri
- Digunakan untuk semua level dari pemula hingga pakar

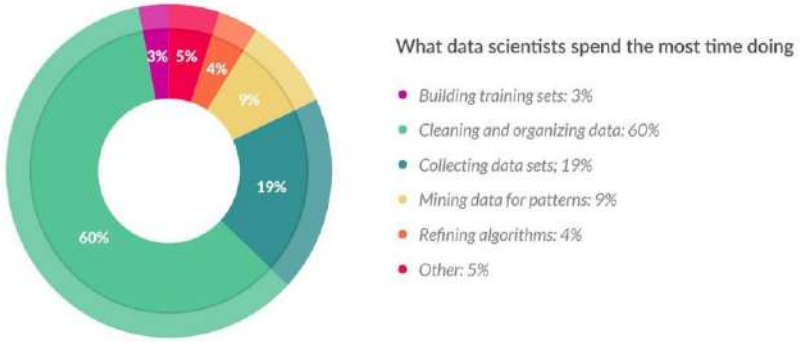
Terminologi dan Definisi

- Istilah lain: Data Pre-processing, Data Manipulation, Data Cleansing/ Normalization
- De;nisi:
  - *transformasi data mentah menjadi format yang mudah dipahami*
  - menemukan data yang relevan untuk disertakan dalam aplikasi analitik shg memberikan informasi yang dicari oleh analis atau pengguna bisnis
  - *langkah pra-pemrosesan yang melibatkan pembersihan, transformasi, dan konsolidasi data.*

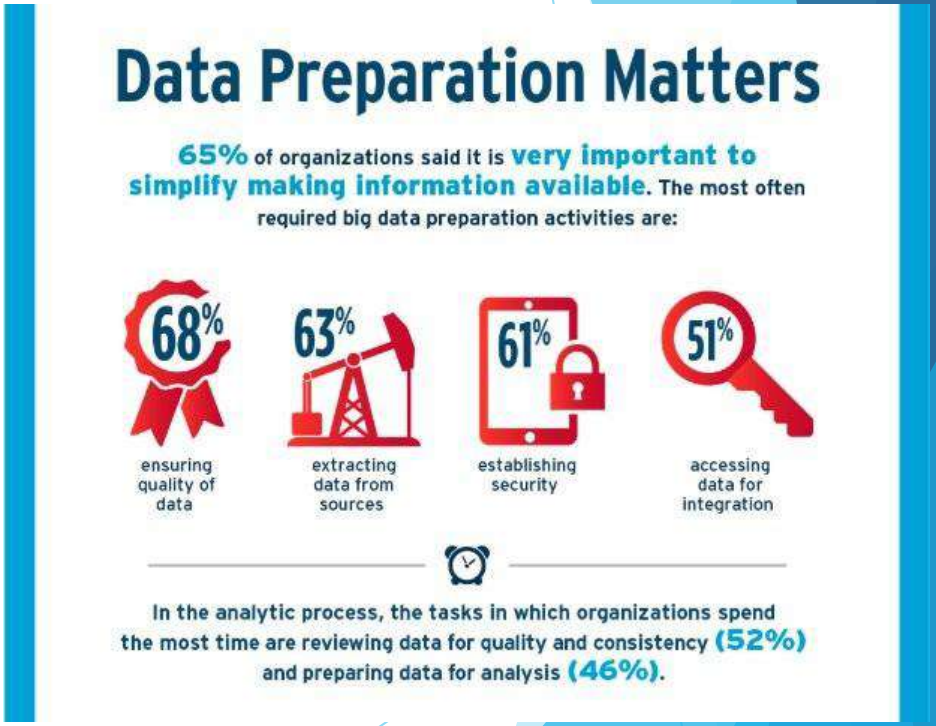


- Definisi:
  - proses yang melibatkan koneksi ke satu atau banyak sumber data yang berbeda, membersihkan data kotor, memformat ulang atau merestrukturisasi data, dan akhirnya menggabungkan data ini untuk digunakan untuk analisis.

Fakta Terkait Data Preparation

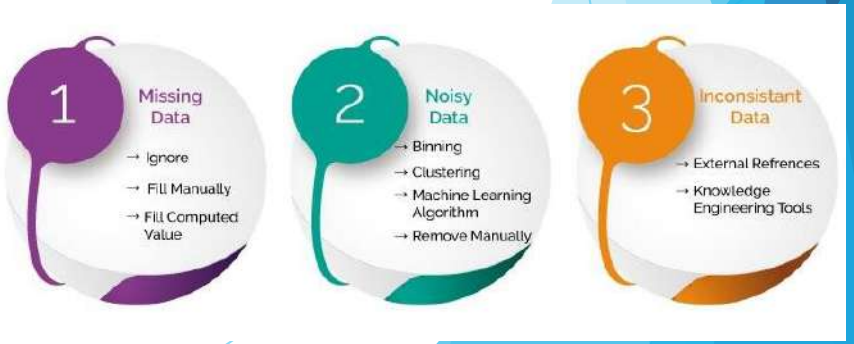


- 60-80% porsi kegiatan data saintis (forbes, crowdflower 2016)
  - data yang ada saat ini dari banyak sumber data dan format yang beragam (terstruktur, semi, dan tidak terstruktur)
  - kualitas model prediktif bergantung pada kualitas data (GIGO)



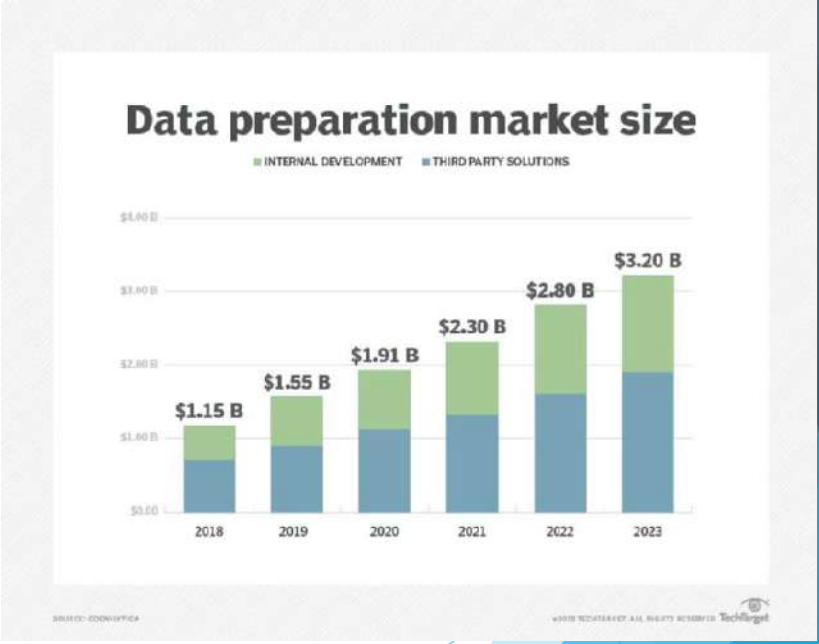
Pentingnya Data Preparation

- data perlu diformat sesuai dengan software yang digunakan
- data perlu disesuaikan dengan metode data sain yang digunakan
- data real-world cenderung 'kotor':
  - tidak komplit: kurangnya nilai attribute, kurangnya atribut tertentu/penting, hanya berisi data agregate. misal: pekerjaan="" (tidak ada isian)
  - *noisy*: memiliki error atau outlier. misal: Gaji="-10", Usia="222"
- data real-world cenderung 'kotor':
  - tidak konsisten: memliki perbedaan dalam kode dan nama. misal : Usia= "32" TglLahir="03/07/2000"; rating "1,2,3" -- > rating "A, B, C"
- kolom dan baris yang saling bertukar
- banyak variabel dalam satu kolom yang sama

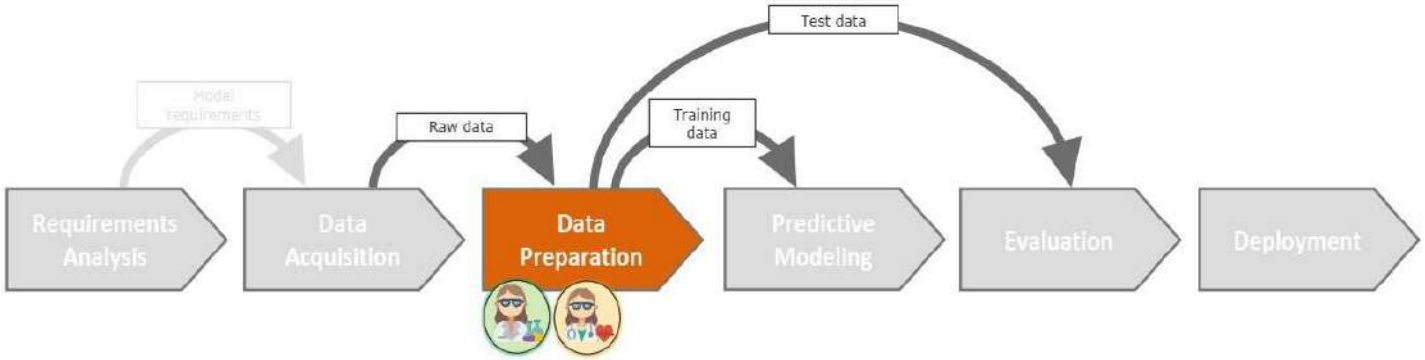


Manfaat Data Preparation

- Kompilasi Data menjadi E;sien dan Efektif (menghindari duplikasi)
- Identi;kasi dan Memperbaiki Error
- Mudah Perubahan Secara Global
- Menghasilkan Informasi yang Akurat utk Pengambilan Keputusan
- Nilai Bisnis dan ROI (Return on Ivestment) akan Meningkat

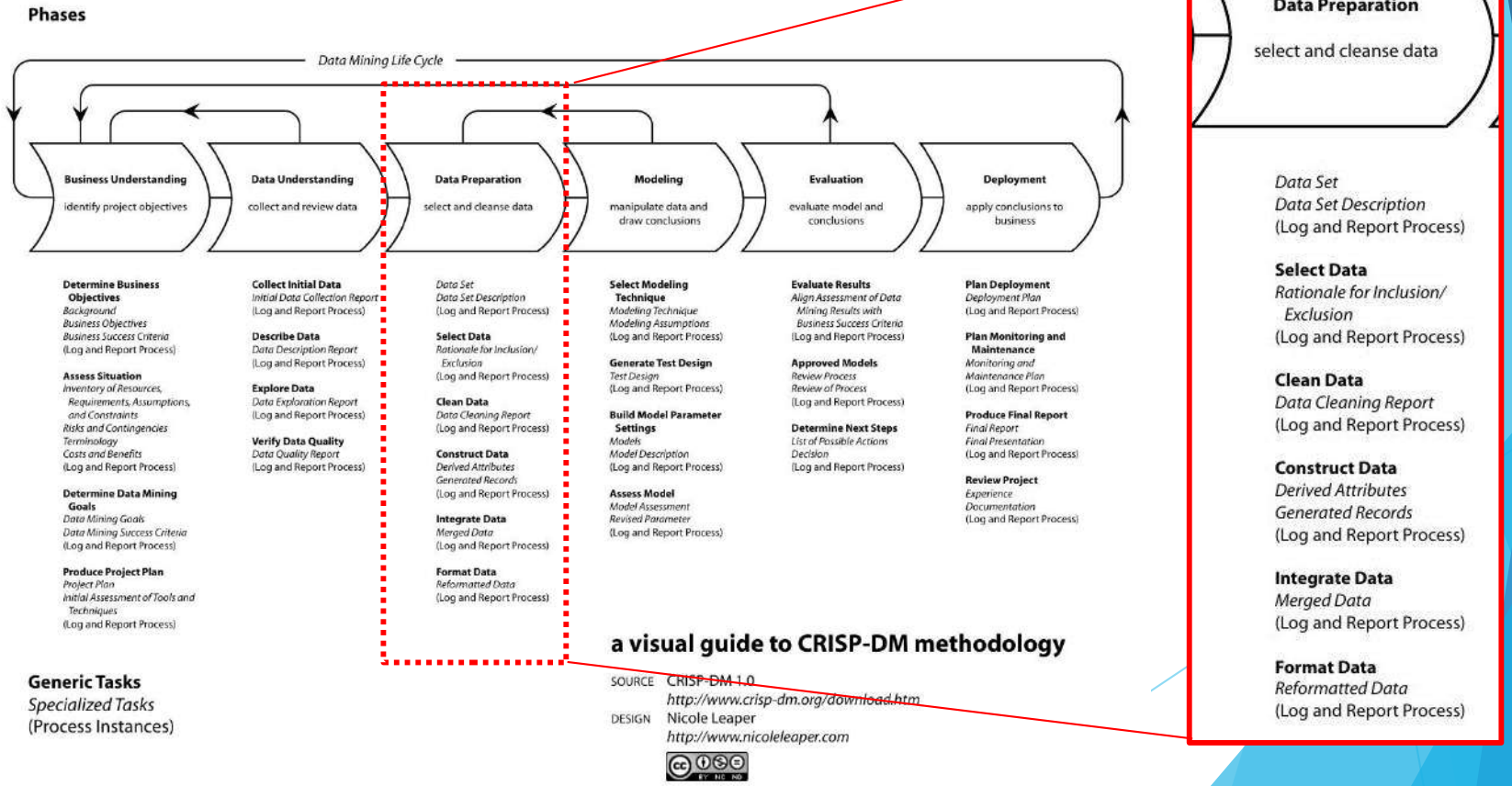


Tahapan dan Tantangan Data Preparation



- **Memakan Waktu Lama**
- **Porsi Teknis yang Dominan**
- **Data yang Tersedia Tidak Akurat atau Jelas/Tidak Langsung Pakai**
- Data tidak Balance Saat Pengambilan Sampel
- Rentan akan Error

Data Preparation dalam CRISP-DM



Tahapan Data Preparation: Pemilihan, Pembersihan & Validasi

Modul 8

1. **Pilih/ Select Data**
   - Pertimbangkan apemilihan data
   - Tentukan dataset yang akan digunakan
   - Kumpulkan data tambahan yang sesuai (internal atau eksternal)
   - Pertimbangkan penggunaan teknik pengambilan sampel
   - Jelaskan mengapa data tertentu dimasukkan atau dikecualikan
2. **Bersihkan/ Clean Data**
   - Perbaiki, hapus atau abaikan noise
   - Putuskan bagaimana menangani nilai-nilai khusus dan maknanya
   - Tingkat agregasi, nilai yang hilang (missing value), dll
   - Bersihkan atau manipulasi outlier
3. **Validasi Data**
   - Periksa/Nilai Kualitas Data
   - Periksa/Nilai Tingkat Kecukupan Data

Paramater/Daftar Isi Dokumentasi Data Cleaning

Laporan dokumentasi data cleaning, setidaknya memiliki parameter berikut:

- Data Set Description
- Data Set yang digunakan
- Jenis noise yang terjadi pada data (diantaranya: Missing data, Data errors; Coding inconsistencies; Missing/ bad metadata
- Pendekatan yang dilakukan untuk menghilangkan noise tersebut
- Teknik mana yang digunakan sehingga berhasil untuk menghilangkan noise tersebut
- Apakah ada kasus atau atribut yang tak dapat diselamatkan
- Pastikan data yang dikecualikan karena kondisi noisenya

Paramater/Daftar Isi Dokumentasi Data Validation

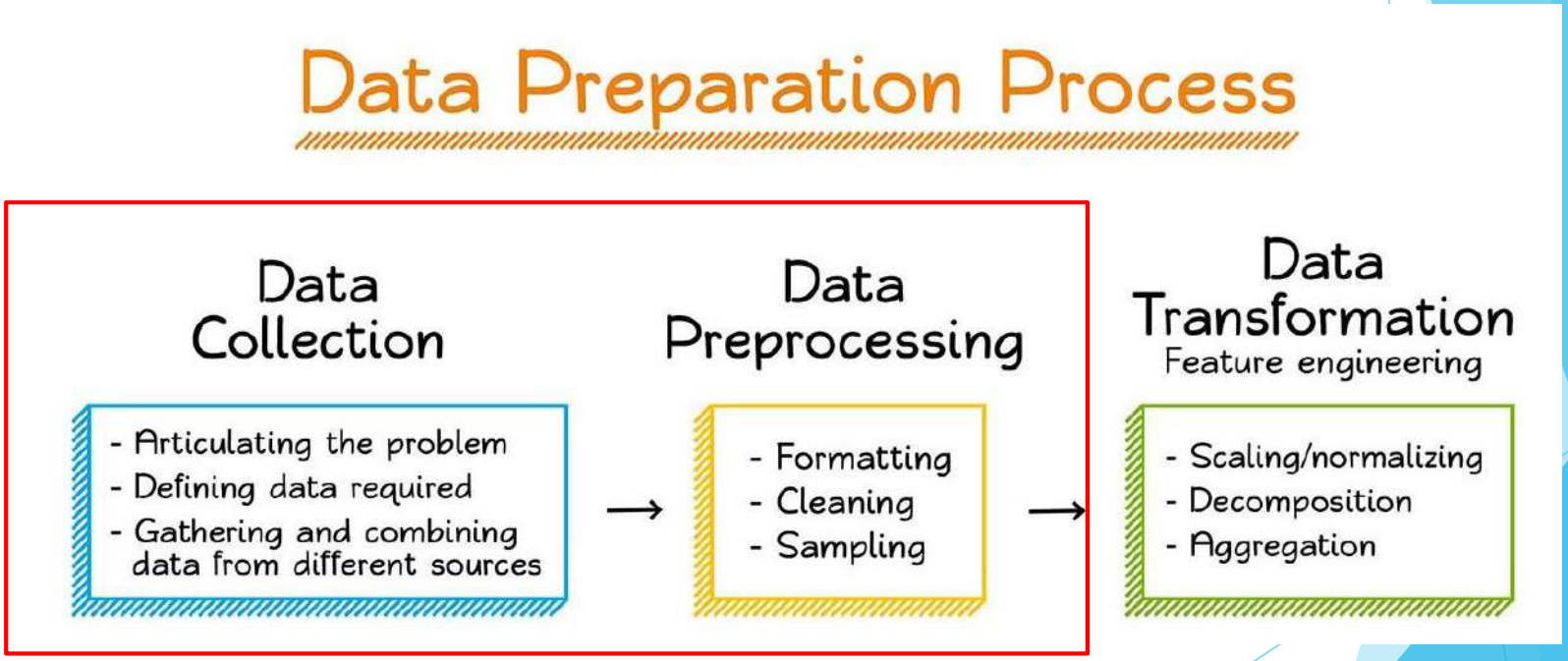
Laporan dokumentasi data cleaning, setidaknya memiliki parameter berikut:

- Validasi data
  - Kebenaran, misal di Indonesia isian Gender yang diakui hanya 2 P/W; Agama hanya 6 (Islam, Protestan, Katholik, Hindu, Budha, Konghucu)
  - Kelengkapan, misal data propinsi seluruh Indonesia (34 prov), namun hanya sebagian yg ada
  - Konsistensi, misal penulisan STM atau SMK;
- Kecukupan data → Perlukan diulang berikan justi;kasi (Resampling)

Rincian Tahapan Data Preparation

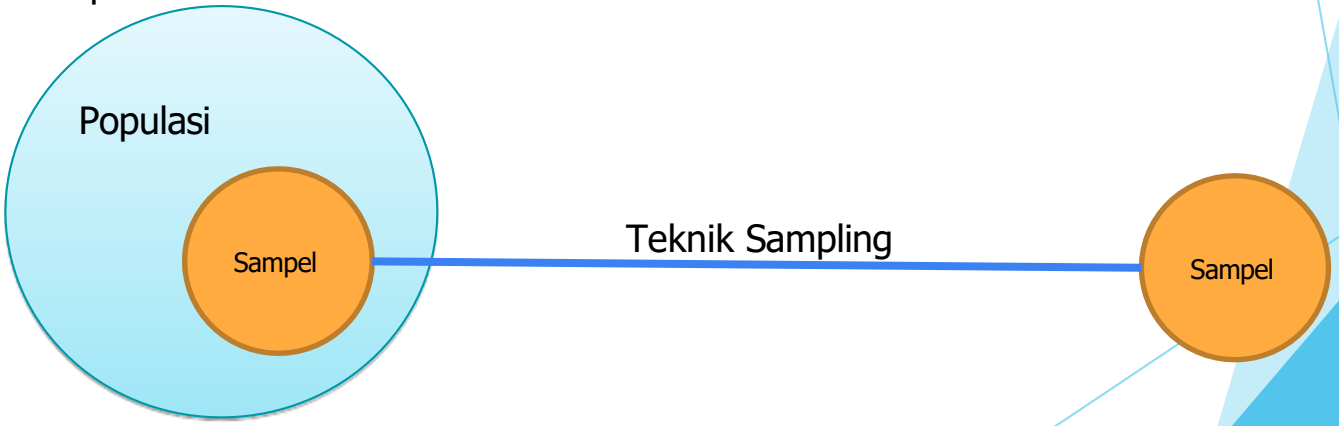
3. **Bangun/ Construct Data**
   - Atribut turunan.
   - Latar belakang pengetahuan.
   - Bagaimana atribut yang hilang dapat dibangun atau diperhitungkan
4. **Integrasi/ Integrate Data**
   - Mengintegrasikan sumber dan menyimpan hasil (tabel dan catatan baru)
5. **Bentuk/ Format Data**

Tahapan Data Preparation: Versi Simple



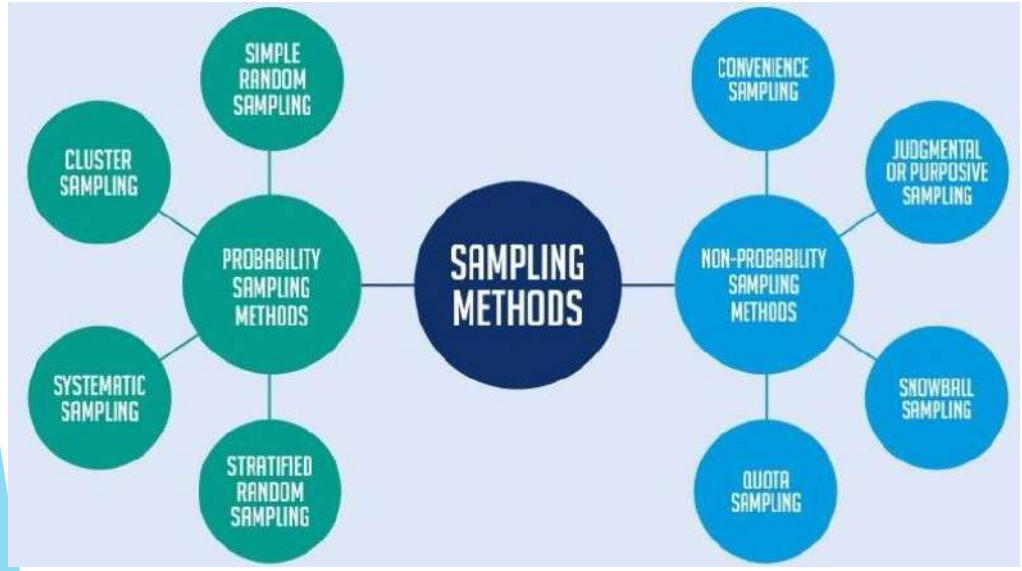
Sampling Data: Pengertian Sampling

- Sebelum melakukan tahapan dalam data preparation, terlebih dahulu adalah pemilihan/penentuan objek yang dapat dilakukan dengan menggunakan penentuan:
  - Populasi
  - Sampel



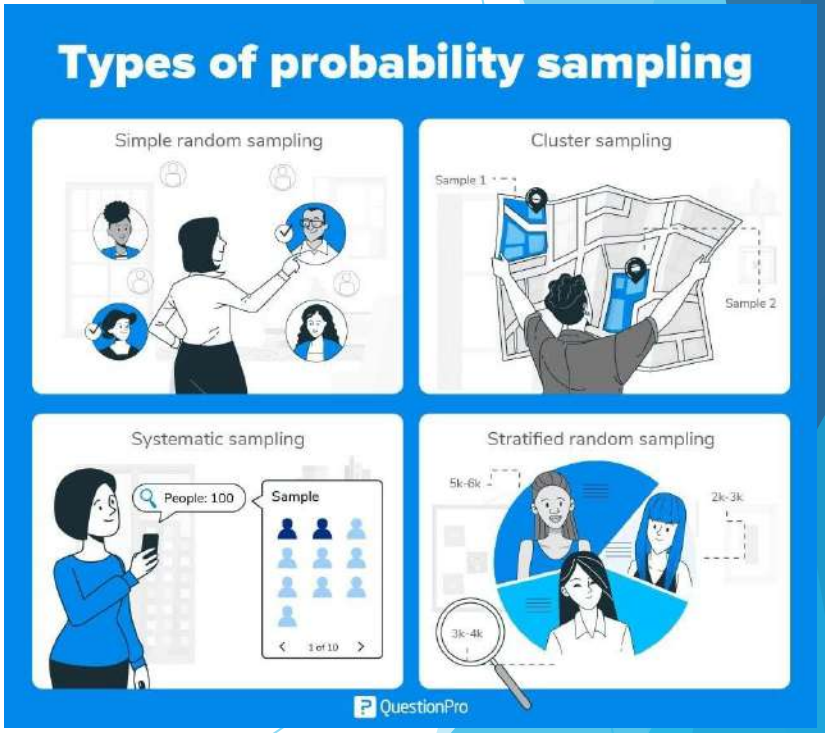
Sampling Data: Metode Sampling

- Kategori Metode Sampling

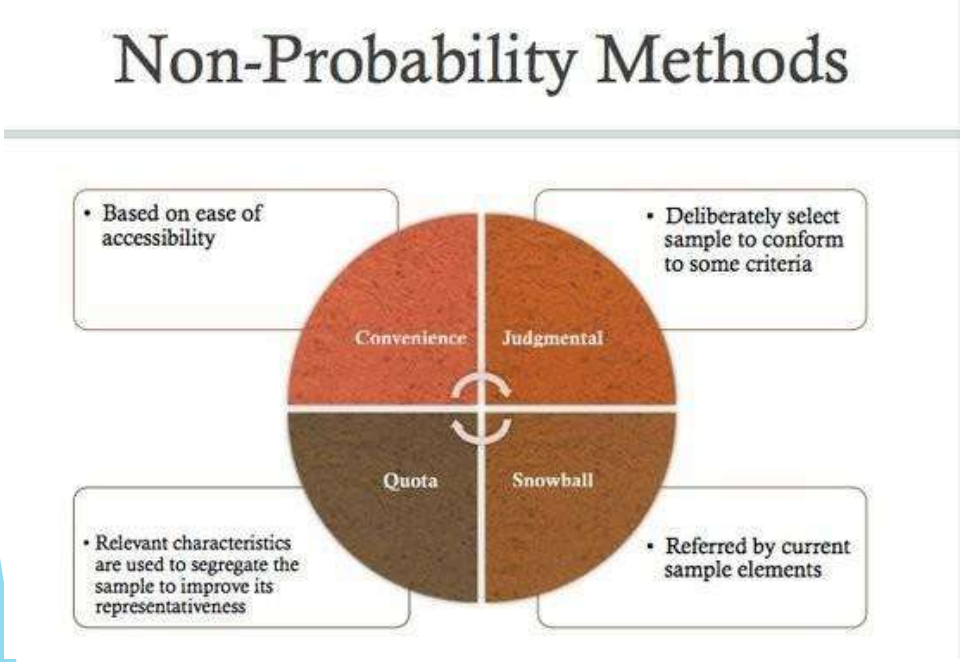


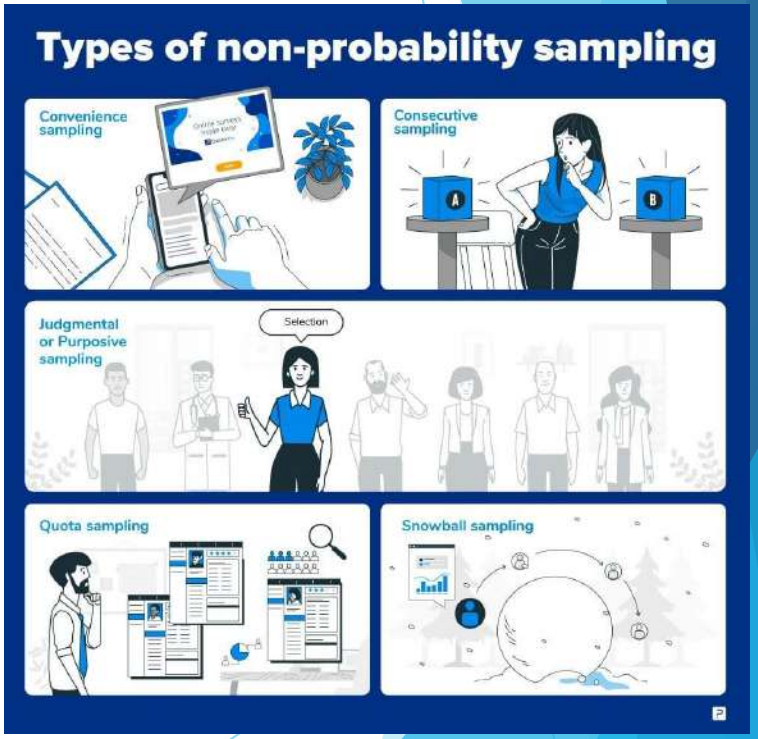
- Probability Sampling:
  - Populasi diketahui
  - Randomisasi/keteracakan: Ya
  - Conclusiver
  - Hasil: Unbiased
  - Kesimpulan: Statistik
- Non-Probability Samplimg
  - Populasi tidak diketahui
  - Keterbatasan penelitian
  - Randomisasi/keteracakan: Tidak
  - Exploratory
  - Hasil: Biased
  - Kesimpulan: Analitik

Sampling Data: Metode Sampling

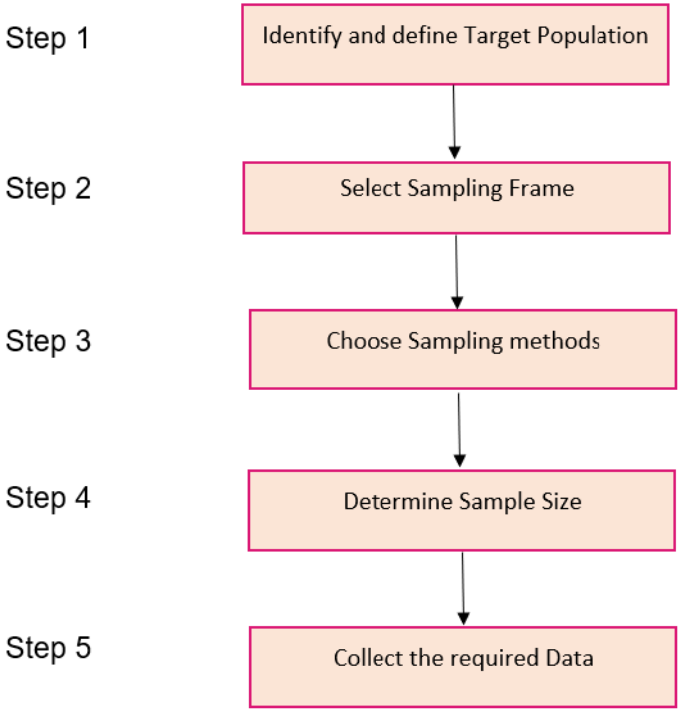


Sampling Data: Teknik Sampling



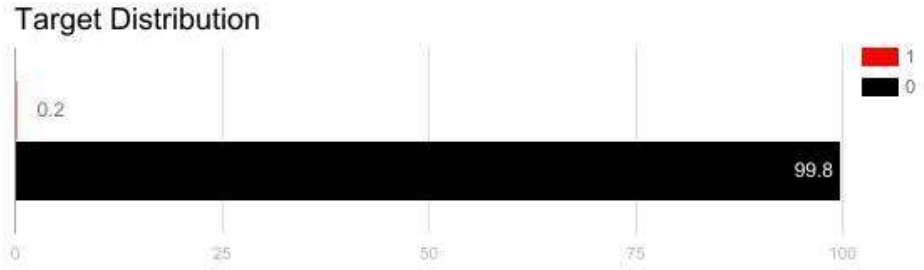


Sampling Data: Tahapan Sampling



Imbalance Dataset: Resampling

- Ini dilakukan setelah proses pemilihan, pembersihan dan rekayasa ;tur dilakuan atas pertanyaan:
  - Tanya: apakah kelas target data yang kita inginkan telah secara sama terdistribusi di seluruh dataset?
  - Jawab: Di banyak kasus tidak/belum tentu. Biasanya terjadi imbalance (ketidakseimbangan) antara dua kelas. Misal utk dataset tentang detekis fraud di perbankan, lelang real-time, atau deteksi intrusi di network! Biasanya data dari dataset tersebut berukuran sangat kecil atau kurang dari 1%, namun sangat signi;kan. Kebanyakan algoritma ML tidak bekerja baik utk dataset imbalance tsb.

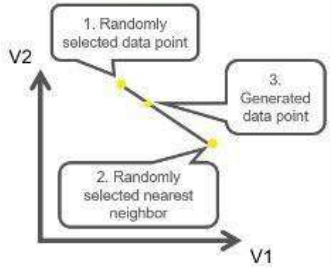
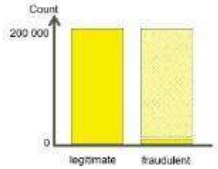
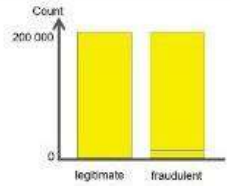
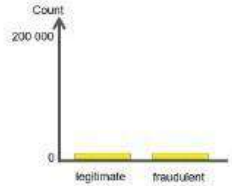


Imbalance Dataset: Resampling

- Berikut adalah bbrp cara utk mengatasi imbalance dataset:
  - Gunakan pengukuran (metrik) yang tepat, misal dengan menggunakan:
    - **Precision/Spesikasi**: berapa banyak instance yang relevan
    - **Recall/Sensitifitas**: berapa banyak instance yang dipilih
    - **F1 score**: harmonisasi mean dari precision dan recall
    - **Matthews correlation coefficient (MCC)**: koe;sien korelasi antara klasi;kasi biner antara observasi vs prediksi
    - **Area under the ROC curve (AUC)**: relasi antara tingkat true-positive vs false-positive
  - Resample data training, dengan dua metode:
    - **Undersampling**: menyeimbangkan dataset dengan mereduksi ukuran kelas yang melimpah. Dilakukan jika kuantitas data mencukupi
    - **Oversampling**: Kebalikan dari undersampling, dilakukan jika kuantitas data tidak mencukupi

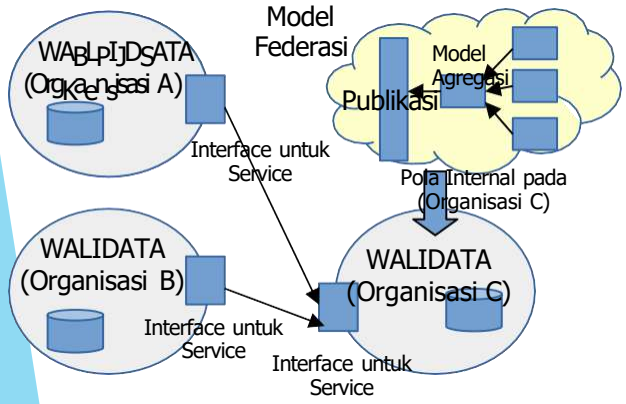
Imbalance Dataset: Resampling

- Teknik Resampling:
  - oversampling (SMOTE)
  - oversampling (Bootstrap)
  - undersampling (Bootstrap)

Resampling method	Description	Target class distribution after resampling
Oversampling (SMOTE)	Generate new synthetic fraudulent transactions until the number of fraudulent transactions is ca. equal to the number of legitimate transactions: 1. Select one of the fraudulent transactions in the training data randomly 2. Select one of its *n* nearest neighbors in the same fraudulent class randomly 3. Select a random point between the existing fraudulent transaction and its nearest neighbor 	• Original data in yellow • New synthetic data in light patterned yellow 
Oversampling (Bootstrap)	Randomly draw with replacement a sample of fraudulent transactions until the number of fraudulent transactions is ca equal to the number of legitimate transactions	
Undersampling (Bootstrap)	Randomly draw with replacement as many legitimate transactions as there are fraudulent transactions	

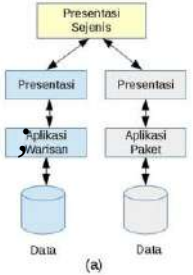
Keberadaan data dari berbagai sumber (stakeholder)

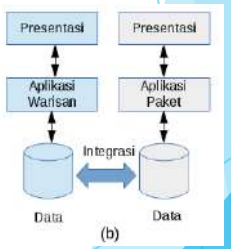
GIGO

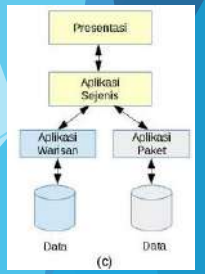


- Sistem informasi di masing-masing organisasi tidak bisa bertukar data/informasi pada lingkungan heterogen
- Interoperabilitas data akan menge;sienkan kerja serta dapat melakukan **prediksi dan analisis berbasiskan AI**
- Mendukung knowledge discovery dan decision making

- **Integrasi Presentasi.** User interface yang menyediakan akses pada suatu aplikasi. kinerja, persepsi, dan tidak adanya interkoneksi antara aplikasi dan data.
- **Integrasi Data**. Dilakukan langsung pada basis data atau struktur data. Jika terjadi perubahan model data, maka integrasinya perlu direvisi atau dilakukan ulang.
- **Integrasi Fungsional** Proses integrasi dilakukan pada level logika bisnis pada beberapa aplikasi.







Pekerjaan Unit Kerja → Pekerjaan Lembaga → Pekerjaan Pemerintah → Pekerjaan Publik

Kelengkapan Data sesuai Tujuan

Data Perencanaan

Data Pelaksanaan

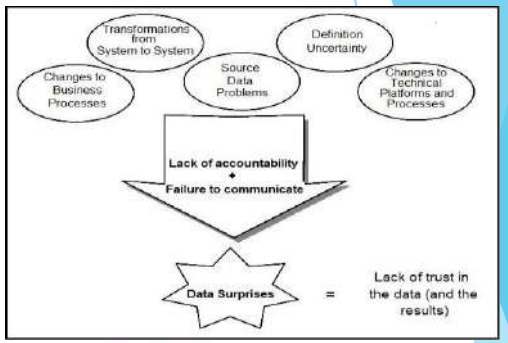
Data Pengawasan

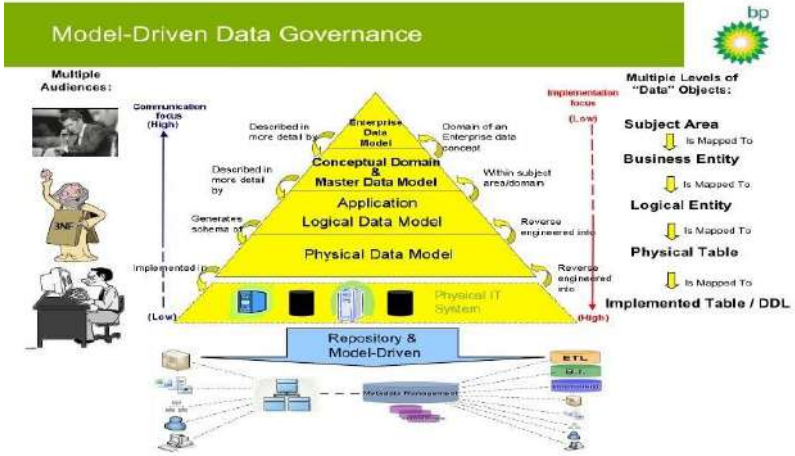
Data Penindakan

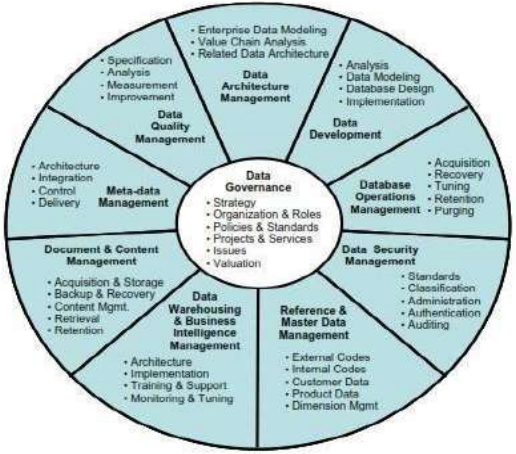
Penggunaan data/fungsi bersama - interoperabilitas

- Menaikkan kualitas data di lingkungan organisasi
- Konsistensi data dan update dijaga
- Mengupayakan agar memiliki skema yang sama ataupun pemetaan skema yang terbuka → skema data diketahui umum
- Mengupayakan data referensi sama → data referensi diketahui

Kualitas Data bergantung Governance







Federated Database

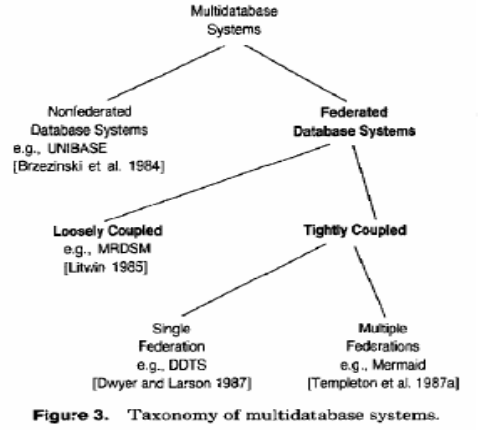
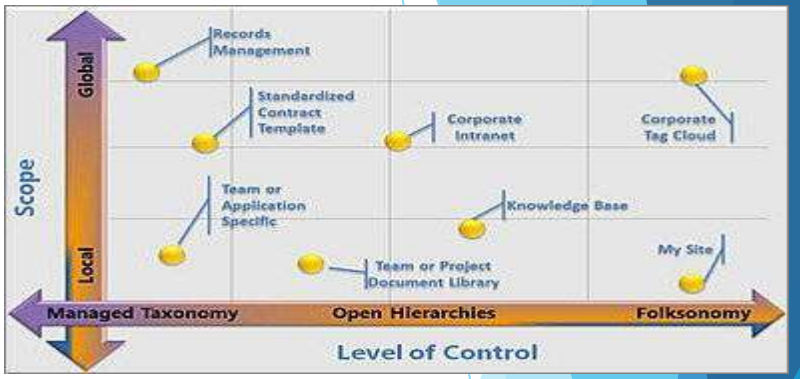


Figure 3. Taxonomy of multidatabase systems.

- Tidak mungkin memaksa setiap pihak "menyerahkan" datanya
- Setiap pihak memiliki teknologi dan sistem masing-masing
- Transparency
- Heterogeneity
- Functionality
- Autonomy of underlying federated sources,
- Extensibility & Openness
- Optimized performance



Sistem A

Metadata

XSD1

XML1

Data1

XSD dari Metadata

XML Translator

XSLT

De;nisi
XSD1 → XSD2

Metadata

XML2

Data1

XSD2

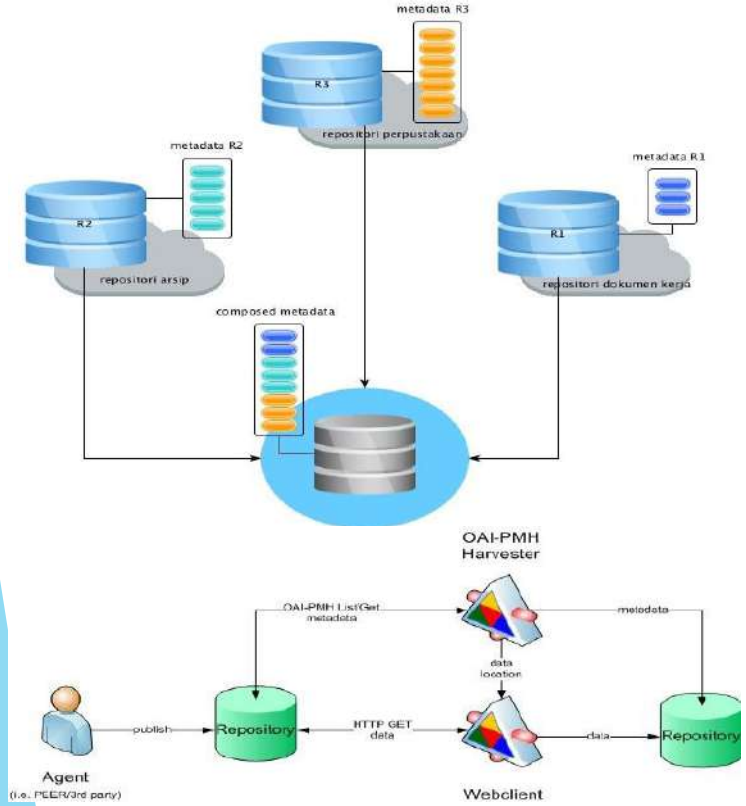
XSD dari Metadata

Sistem B

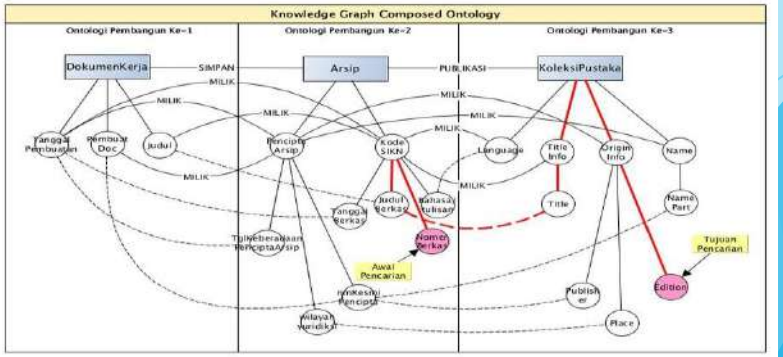
DocID	FieldA	FieldB	FieldC

ID	FieldX	FieldY

Ontologi dan Database



DATABASE RELASIONAL	ONTOLOGI
Close World Assumption (CWA), Fokus pada data	Open World Assumption (OWA), Fokus pada makna
Adanya Constraint untuk mencapai data integritas, namun mungkin menyembunyikan makna	Adanya Ontology axioms untuk menspesi;kasikan makna, dapat digunakan untuk pencapaian integritas
Tidak menggunakan hirarki ISA	Hirarki ISA merupakan backbone
Skema lebih sederhan, belum tentu dapat digunakan kembali	Skema lebih kompleks, dapat digunakan kembali



Pemilihan (Seleksi Fitur) Data

	K1	K2	K3	K4	K5	K6
R1						
R2						
R3						
R4						

Fitur:
Kolom yang dipilih
Untuk sesuai tujuan

Name of the statistical features	Formula/description
Standard error	$\sqrt{\frac{1}{n-2}\left[\sum(y-\bar{y})^2-\frac{\sum(x-\bar{x})(y-\bar{y})^2}{\sum(x-\bar{x})^2}\right]}$
Standard deviation	$\sqrt{\frac{n\sum x^2-(\sum x)^2}{n(n-1)}}$
Sample variance	$\sqrt{\frac{n\sum x^2-(\sum x)^2}{n(n-1)}}$
Kurtosis	$\left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_i-\bar{x}}{s_d}\right)^4\right\}-\frac{3(n-1)^2}{(n-2)(n-3)}$
Skewness	$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i-\bar{x}}{s_d}\right)^3$
Maximum value	Maximum signal point value in a given signal.
Minimum value	Minimum signal point value in a given signal.
Range	Difference in maximum and minimum signal point values for a given signal.
Sum	Sum of all feature values for each sample.
Mean	The arithmetic average of a set of values or distribution.
Median	Middle value separating the greater and lesser halves of a data set.
Mode	A statistical term that refers to the most frequently occurring number found in a set of numbers. (i.e.) The

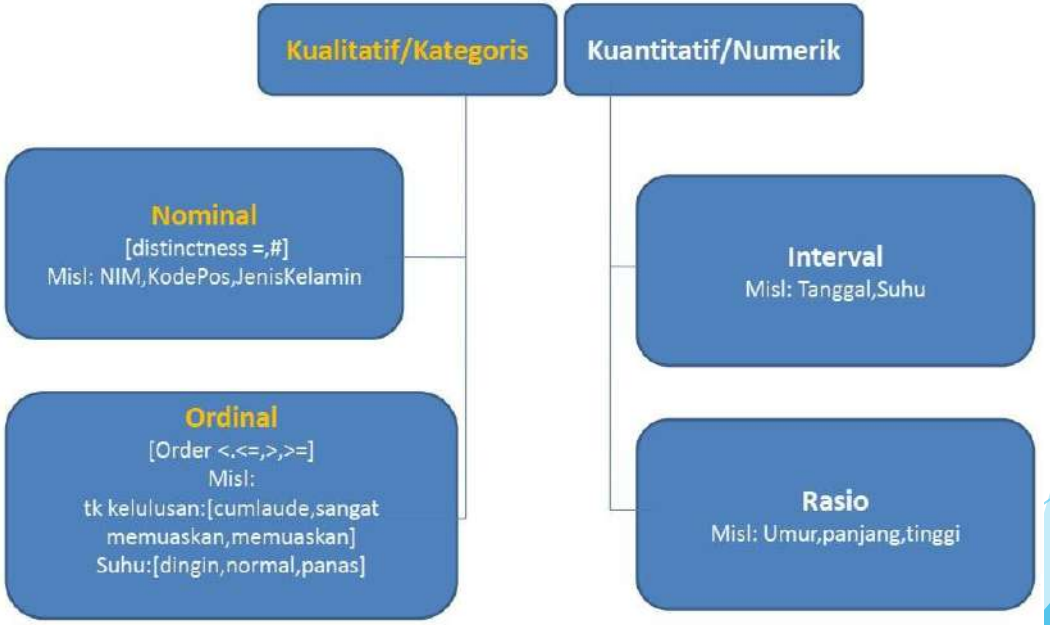
- Setelah menentukan sampling atas data yang akan diambil nanti, selanjutnya adalah melakukan seleksi **fitur** (feature selection) atas data sampling tsb --> Memilih Kolom/Atribut/Variabel yang akan diolah lebih lanjut
- Terminologi **fitur** di Data Science atau Machine Learning adalah Kolom/Atribut/Variabel yang dianggap & dihitung sebagai prioritas (sedikit berbeda dengan terminologi fitur di Statitistika)
- Seleksi fitur merupakan konsep inti dalam ML yang berdampak besar bagi kinerja model prediksi,
- Fitur data yang tidak/sebagian saja relevan dampak berdampak negatif thdp kinerja model
- Definisi Seleksi Fitur: proses otomatis atau manual memilih fitur data yang **paling berkontribusi** thdp variabel prediksi atau output yang diinginkan.

Seleksi Fitur Data

- Manfaat:
  - Reduksi *Overfirring*: semakin kecil data redundant maka keputusan berdasarkan noise semakin berkurang
  - Meningkatkan Akurasi: semakin kecil data misleading maka akurasi model lebih baik
  - Reduksi Waktu Training: semakin kecil titik data (data point) maka kompleksitas algortima berkurang dan latih algoritma lebih cepat
- Jenis:
  - **Unsupervised**: metode yang **mengabaikan variabel target,** seperti menghapus variabel yang berlebihan menggunakan *korelasi*
  - **Supervised**: metode yang **menggunakan variabel target**, seperti menghapus variabel yang tidak relevan

Seleksi Fitur

- Membedakan jenis data: Numerik vs Kategorik (lihat modul 6 utk penjabaran)



Validasi Data

- Veri;kasi vs Validasi
  - Veri;kasi: Benar vs Salah (sesuai prosedur)
  - Validasi: Kuat vs Lemah (sesuai kenyataan)
- Validasi merupakan tahapan kritikal yang sering diabaikan DS-tist pemula, karena memeriksa, diantaranya sbb:
  - Tipe Data (mis. integer, float, string)
  - Range Data
  - Uniqueness (mis. Kode Pos)
  - Consisten expression (mis. Jalan, Jl., Jln.)
  - Format Data (mis. utk tgl "YYYY-MM-DD" VS "DD-MM-YYYY.") → tmt (terhitung mulai tanggal)
  - Nilai Null/Missing Values
  - Misspelling/Type
  - Invalid Data (gender: L/P: L; Laki-laki; P: Pria/Perempuan?)

- Teknik Validasi Data dan Model:
  - Akurasi
  - Kelengkapan
  - Konsistensi
  - Ketepatan Waktu
  - Kepercayaan
  - Nilai Tambah
  - Penafsiran
  - Kemudahan Akses

Pandas: DataFrame

Syntax – Creating DataFrames

	a	b	c
1	4	7	10
2	5	8	11
3	6	9	12

```
df = pd.DataFrame(
        {"a" : [4 ,5, 6],
         "b" : [7, 8, 9],
         "c" : [10, 11, 12]},
       index = [1, 2, 3])
```
Specify values for each column.

```
df = pd.DataFrame(
    [[4, 7, 10],
     [5, 8, 11],
     [6, 9, 12]],
    index=[1, 2, 3],
    columns=['a', 'b', 'c'])
```
Specify values for each row.

		a	b	c
n	v			
d	1	4	7	10
	2	5	8	11
e	2	6	9	12

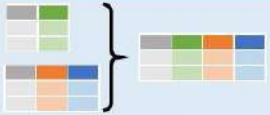
```
df = pd.DataFrame(
        {"a" : [4 ,5, 6],
         "b" : [7, 8, 9],
         "c" : [10, 11, 12]},
index = pd.MultiIndex.from_tuples(
        [('d',1),('d',2),('e',2)],
          names=['n','v']))
```
Create DataFrame with a MultiIndex

Reshaping Data – Change the layout of a data set

`pd.melt(df)`
Gather columns into rows.

`df.pivot(columns='var', values='val')`
Spread rows into columns.

`pd.concat([df1,df2])`
Append rows of DataFrames



`pd.concat([df1,df2], axis=1)`
Append columns of DataFrames

`df.sort_values('mpg')`
Order rows by values of a column (low to high).

`df.sort_values('mpg',ascending=False)`
Order rows by values of a column (high to low).

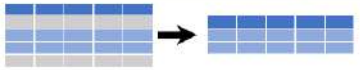
`df.rename(columns = {'y':'year'})`
Rename the columns of a DataFrame

`df.sort_index()`
Sort the index of a DataFrame

`df.reset_index()`
Reset index of DataFrame to row numbers, moving index to columns.

`df.drop(columns=['Length','Height'])`
Drop columns from DataFrame

Subset Observations (Rows)



`df[df.Length > 7]`
Extract rows that meet logical criteria.

`df.drop_duplicates()`
Remove duplicate rows (only considers columns).

`df.head(n)`
Select first n rows.

`df.tail(n)`
Select last n rows.

`df.sample(frac=0.5)`
Randomly select fraction of rows.

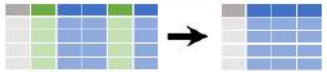
`df.sample(n=10)`
Randomly select n rows.

`df.iloc[10:20]`
Select rows by position.

`df.nlargest(n, 'value')`
Select and order top n entries.

`df.nsmallest(n, 'value')`
Select and order bottom n entries.

Subset Variables (Columns)



`df[['width','length','species']]`
Select multiple columns with specific names.

`df['width']` *or* `df.width`
Select single column with specific name.

`df.filter(regex='regex')`
Select columns whose name matches regular expression *regex*.

Hands On: Seleksi Fitur

- Dalam praktek kali ini akan digunakan 3 teknik seleksi ;tur yang mudah dan memberikan hasil yang baik:
  - Seleksi Univariat (Univariate Selection)
  - Pentingnya Fitur (Feature Importance)
  - Matriks Korelasi (Correlation Matrix) dengan *Hearmap*
- Sumber dataset: https://www.kaggle.com/iabhishekofficial/mobile-price-classification#train.csv

- Deskripsi variabel dari dataset:
  - *battery_power:* Total energy a battery can store in one time measured in mAh
  - *blue*: Has Bluetooth or not
  - *clock_speed:* the speed at which microprocessor executes instructions
  - *dual_sim*: Has dual sim support or not
  - *fc:* Front Camera megapixels
  - *four_g:* Has 4G or not
  - *int_memory:* Internal Memory in Gigabytes
  - *m_dep:* Mobile Depth in cm
  - *mobile_wt:* Weight of mobile phone
  - *n_cores:* Number of cores of the processor
  - *pc:* Primary Camera megapixels
  - *px_height:* Pixel Resolution Height

Hands On: Seleksi Fitur

- **Deskripsi variabel dari dataset** (lanjutan):
  - Seleksi Univariate
    - Uji statistik dapat digunakan utk memilih ;tur-
      - ;tur tsb yang memiliki relasi paling kuat dengan variabel output
    - Library scikit-learn menyediakan class
      - SelectKBest yang digunakan utk serangkaian uji statistik berbeda utk memilih angka spesi;k dari ;tur
    - Berikut ini adalah uji statistik chi-square utk
      - ;tur non-negatif utk memilih 10 ;tur terbaik dari dataset *Mobile Price Range Predicrion*.

`px_width:` Pixel Resolution Width
`ram:` Random Access Memory in MegaBytes
`sc_h:` Screen Height of mobile in cm
`sc_w:` Screen Width of mobile in cm
`talk_time:` the longest time that a single battery charge will last when you are

- `three_g:` Has 3G or not
- `touch_screen:` Has touch screen or not
- `wifi:` Has wifi or not
- `price_range:` This is the target variable with a value of 0(low cost), 1(medium cost), 2(high cost) and 3(very high cost).

Hands On: Seleksi Fitur

- Seleksi Univariat (lanjutan):

```
import pandas as pd
import numpy as np
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

data = pd.read_csv("C:/Users/Bayu/Documents/DTS 2021/Datasets/train.csv")

X = data.iloc[:,0:20]  #independent columns
y = data.iloc[:,-1]    # target colum i.e price range

# apply SelectKBest class to extract

bestfeatures = SelectKBest(score_func=chi2, k=10)
fit = bestfeatures.fit(X,y)
dfscores = pd.DataFrame(fit.scores_)
dfcolumns = pd.DataFrame(X.columns)

#concat two dataframes for better visualization

featureScores = pd.concat([dfcolumns,dfscores],axis=1)
featureScores.columns = ['Specs','Score']  #naming the dataframe columns
print(featureScores.nlargest(10,'Score'))  #print 10 best features
```

```
           Specs          Score
13           ram  931267.519053
11     px_height   17363.569536
0   battery_power   14129.866576
12      px_width    9810.586750
8       mobile_wt      95.972863
6       int_memory      89.839124
15          sc_w       16.480319
16     talk_time       13.236400
4              fc       10.135166
14          sc_h        9.614878
```

Import library / modul yang dibutuhkan

Load datasets, sesuaikan dengan path direktori masing-masing

`Iloc[]`, digunakan untuk untuk seleksi/ slicing data dengan parameter index menggunakan bilangan bulat.

Output

Hands On: Seleksi Fitur

- **Feature Importance (FT)**
  - FT berfungsi memberi skor untuk setiap ;tur data, semakin tinggi skor semakin penting atau relevan ;tur tersebut terhadap variabel output
  - FT merupakan kelas inbuilt yang dilengkapi dengan Pengklasi;kasi Berbasis Pohon (Tree Based Classi;er), kita akan menggunakan Pengklasi;kasi Pohon Ekstra untuk mengekstraksi 10 ;tur teratas untuk kumpulan data

```
import pandas as pd
import numpy as np

data = pd.read_csv("C:/Users/Bayu/Documents/DTS 2021/Datasets/train.csv")
X = data.iloc[:,0:20]  #independent columns
y = data.iloc[:,-1]    #target column i.e price range

from sklearn.ensemble import ExtraTreesClassifier
import matplotlib.pyplot as plt
model = ExtraTreesClassifier()
model.fit(X,y)

print(model.feature_importances_) #use inbuilt class feature_importances of tree based classifiers

#plot graph of feature importances for better visualization
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(10).plot(kind='barh')
plt.show()
```

Mende;nisikan model yang akan digunakan yaitu menggunakan algoritma **ExtraTreesClassifier**.

`model. t()` untuk melatih model diikuti oleh parameter variable data

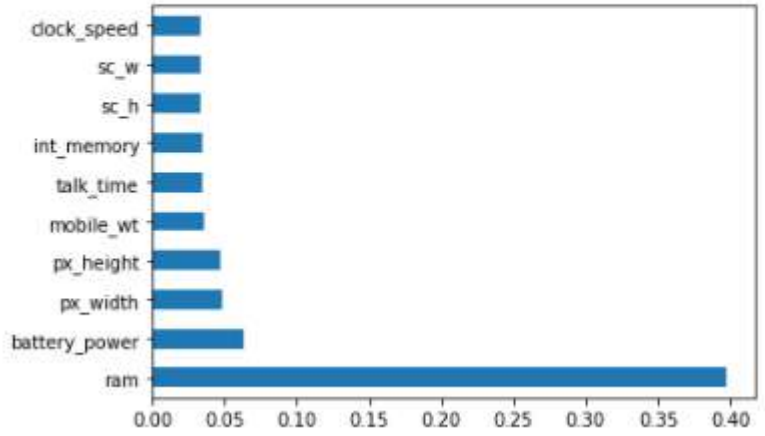
`.largest(10)`: membuat ploting 10 data teratas.

`.plot(kind='barh')`: untuk membuat jenis plot diagram batang horizontal

Hands On: Seleksi Fitur

- Output:

```
[0.06329642 0.0193987  0.03334552 0.0188696  0.03144026 0.01622896
 0.03468226 0.03269537 0.03574171 0.03269081 0.03317167 0.04704737
 0.04849356 0.39695054 0.03392805 0.03372551 0.03512574 0.01359888
 0.01910327 0.02046578]
```



Hands On: Seleksi Fitur

- **Matriks Korelasi dengan *Heatmap***
  - Korelasi menyatakan bagaimana ;tur terkait satu sama lain atau variabel target.
  - Korelasi bisa positif (kenaikan satu nilai ;tur meningkatkan nilai variabel target) atau negatif (kenaikan satu nilai ;tur menurunkan nilai variabel target)
  - *Hearmap* memudahkan untuk mengidenti;kasi ;tur mana yang paling terkait dengan variabel target, kami akan memplot peta panas ;tur yang berkorelasi menggunakan `seaborn` library

```python
import pandas as pd
import numpy as np
import seaborn as sns

data = pd.read_csv("C:/Users/Bayu/Documents/DTS 2021/Datasets/train.csv")

X = data.iloc[:,0:20]  #independent columns
y = data.iloc[:,-1]    #target column i.e price range

#get correlations of each features in dataset
corrmat = data.corr()
top_corr_features = corrmat.index
plt.figure(figsize=(20,20))

#plot heat map
g=sns.heatmap(data[top_corr_features].corr(),annot=True,cmap="RdYlGn")
```

Figure: adalah window atau page atau halaman dalam objek visual. kalau kita ngegambar di kertas, maka kertas tersebutlah yang di namakan ;gure.

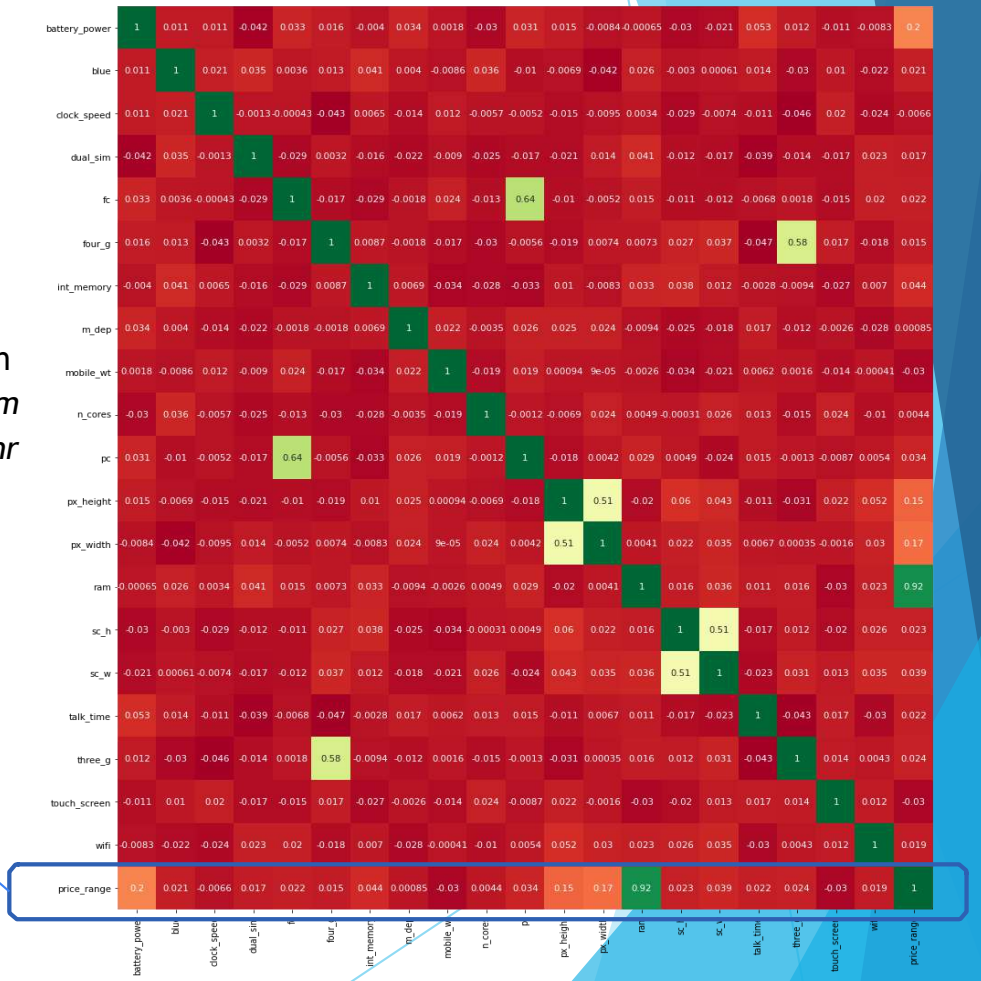
Ágsize(): *ukuran dari figure, mengambil dua paramerer lebar dan ringgi (dalam inci)*

cmap: Colormap digunakan untuk memetakan nilai data yang dinormalisasi ke warna RGBA.

annot=***True*** untuk menampilkan korelasi antar atribut. Jika nilai korelasi mendekati 1 maka hubungan antar atribut semakin tinggi

Hands On: Seleksi Fitur

- **Matriks Korelasi dengan *Heatmap* (lanjutan)**
  - lihat pada baris terakhir yaitu *price range*, korelasi antara price range dengan ;tur lain dimana ada relasi kuat dengan variabel *ram* dan diikuti oleh var *barrery power* , *px heighr* and *px widrh*.
  - sedangkan utk var *clock_speed* dan *n_cores* berkorelasi lemah dengan *price range*



Hands-on Data Cleaning

- Data cleaning atas data berantakan (messy data), seperti:
  - missing value,
  - format tidak konsisten
  - record tidak berbentuk baik (malformed record)
  - outlier yang berlebihan
- Lingkup hands-on:
  - Membuang kolom-kolom tidak penting dalam suatu DataFrame
  - Mengubah indeks di DataFrame
  - Membersihkan kolom dengan metode `.str()`
  - Membersihkan semua dataset dengan fungsi `DataFrame.applymap()`
  - Merubah nama kolom sehingga kolom lebih mudah dikenali
  - Melewatkan baris-baris tidak penting dalam ;le CSV