



Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi
Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi
Republik Indonesia



MICROCREDENTIAL: ASSOCIATE DATA SCIENTIST

01 November – 10 Desember 2021

Pertemuan ke-7

Data Preparation 1: Menentukan Objek atau Memilih Data



[ditjen.dikti](#)



[@ditjendikti](#)



[ditjen.dikti](#)



Ditjen Diktiristek



<https://dikti.kemdikbud.go.id/>

Profil Pengajar: Erwin Eko Wahyudi, S.Kom., M.Cs.



Jabatan Akademik: Tenaga Pengajar

Latar Belakang Pendidikan:

- S1: Ilmu Komputer UGM, 2012-2017
- S2: Ilmu Komputer UGM, 2017-2019

Riwayat/Pengalaman Pekerjaan:

- Dosen, UGM, 2021-sekarang
- AI Engineer Recommender System, Bukalapak, 2019-2021

Contact Pengajar:

Ponsel:

0812 1195 4011

Email:

erwin.eko.w@ugm.ac.id

Referensi: SKKNI Data Science

KODE UNIT : J.62DMI00.007.1

JUDUL UNIT : Menentukan Objek Data

DESKRIPSI UNIT: Unit kompetensi ini berhubungan dengan pengetahuan, keterampilan, dan sikap kerja yang dibutuhkan dalam memilih dan memilih data yang sesuai permintaan atau kebutuhan.

ELEMEN KOMPETENSI	KRITERIA UNJUK KERJA
1. Memutuskan kriteria dan teknik pemilihan data	1.1 Kriteria pemilihan data diidentifikasi sesuai dengan tujuan teknis dan aturan yang berlaku 1.2 Teknik pemilihan data ditetapkan sesuai dengan kriteria pemilihan data.
2. Menentukan <i>attributes (columns)</i> dan <i>records (row)</i> data	2.1 Attributes (columns) data diidentifikasi sesuai dengan kriteria pemilihan data. 2.2 Records (row) data diidentifikasi sesuai dengan kriteria pemilihan data.

1. Konteks variabel

- 1.1 Kriteria pemilihan data mencakup kuantitas data (mencakup volume data yang menggambarkan ukuran data misalkan dalam *terabyte, petabyte atau jumlah record*) dan kualitas data (penilaian terhadap nilai mencurigakan, kosong, inkonsisten, duplikasi maupun ambigu). Kriteria bisa berbentuk ketentuan mengenai penciran, korelasi antar atribut, data yang kosong dan sebagainya.
- 1.2 Aturan yang berlaku termasuk di dalamnya prosedur dan otorisasi mengakses data.
- 1.3 Teknik pemilihan data adalah teknik dalam pengambilan sampel, namun secara garis besar dapat dibagi menjadi dua: *probability sampling* atau *random sampling* dan *non-probability sampling*.
- 1.4 *Attributes (columns)* data adalah bagian data, yang mewakili karakteristik atau *feature* dari objek data.
- 1.5 *Records (row)* data adalah mengembalikan hasil *query* sebagai satu baris objek saja dimana baris yang diambil adalah baris pertama.

Course Definition

- Bagian Pertama dari Tiga materi Data Preparation
- Berfokus pada Penentuan Objek Data atau Memilih Data
- Pengetahuan dan pemahaman akan data preparation menjadi syarat mutlak untuk menghasilkan model prediksi yang optimal



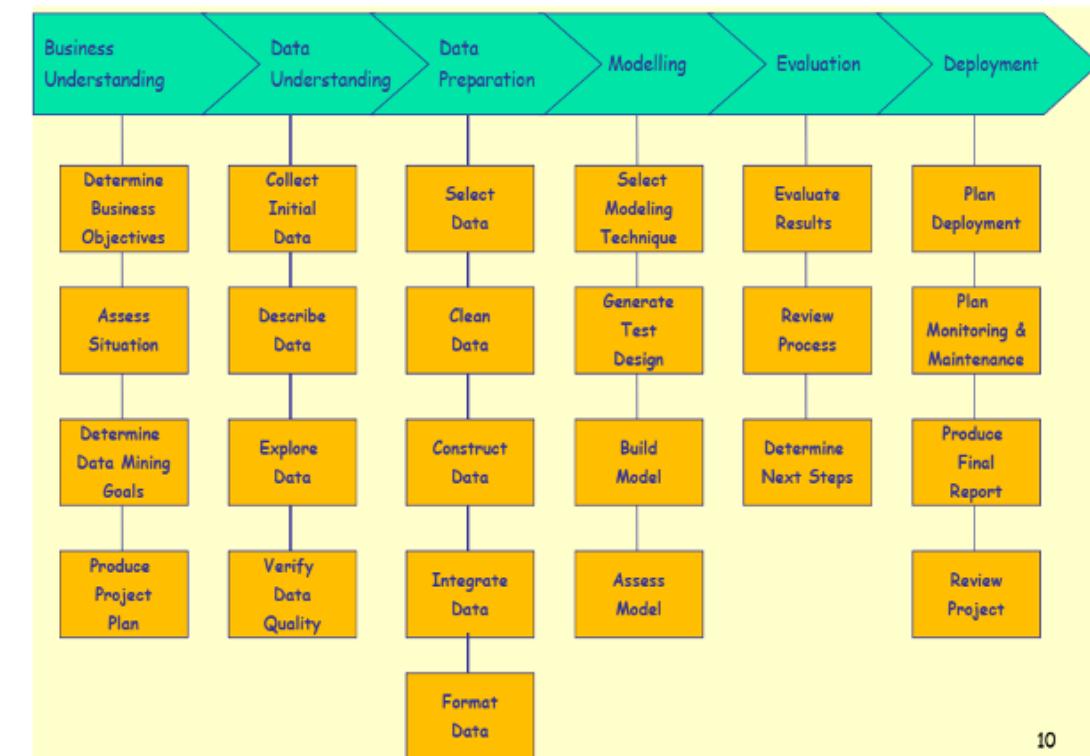
Learning Objective

Dalam pelatihan ini diharapkan:

Peserta memiliki kemampuan untuk memilih dan memilah data sesuai kebutuhan dan sumber daya yang dimiliki

Data Preparation dalam CRISP-DM

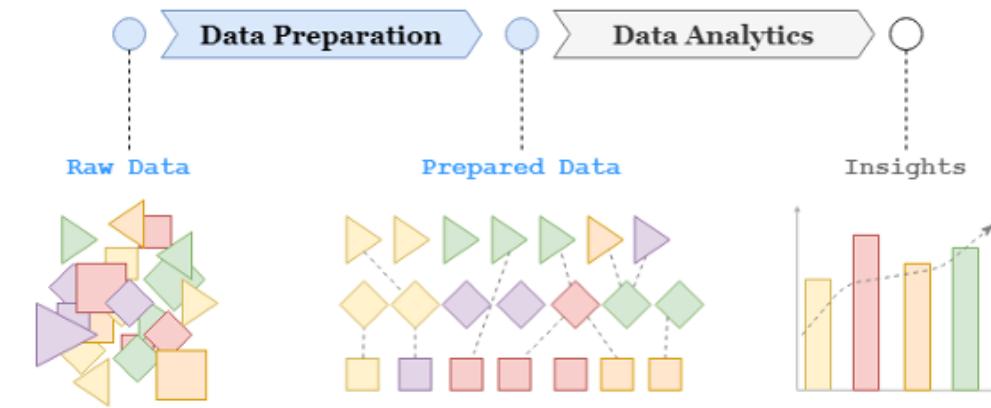
- Akronim dari: **CRoss Industry Standard Process Data Mining**
- Metodologi umum untuk data mining, analitik, dan proyek data sains, berfungsi menstandarkan proses data mining lintas industri
- Digunakan untuk semua level dari pemula hingga pakar



10

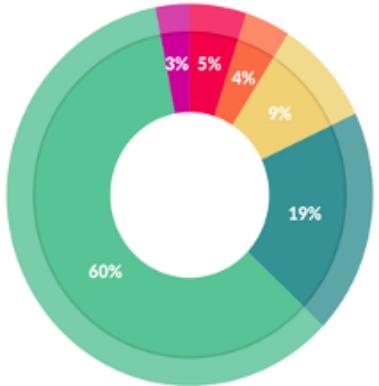
Terminologi dan Definisi

- Istilah lain: **Data Pre-processing**, **Data Manipulation**, **Data Cleansing/ Normalization**
- Definisi:
 - *transformasi data mentah menjadi format yang mudah dipahami*
 - menemukan data yang relevan untuk disertakan dalam aplikasi analitik sehingga 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



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets: 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

- 60-80% porsi kegiatan data scientist (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)

Data Preparation Matters

65% of organizations said it is **very important to simplify making information available**. The most often required big data preparation activities are:



In the analytic process, the tasks in which organizations spend the most time are reviewing data for quality and consistency (52%) and preparing data for analysis (46%).

Pentingnya Data Preparation

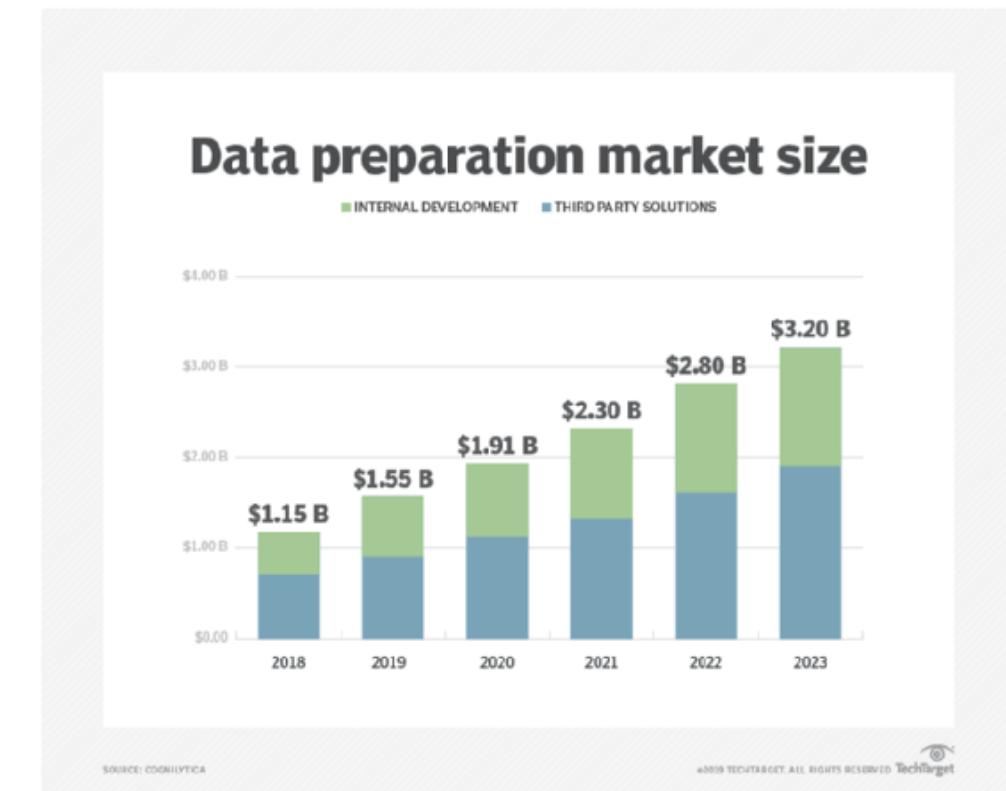
- data perlu diformat sesuai dengan software yang digunakan
- data perlu disesuaikan dengan metode data science yang digunakan
- data real-world cenderung ‘kotor’:
 - tidak komplit: kurangnya nilai attribute, kurangnya atribut tertentu/penting, hanya berisi data aggregate. misal: pekerjaan="" (tidak ada isian)
 - noisy: memiliki error atau outlier. misal: Gaji="-10", Usia="222"
- data real-world cenderung ‘kotor’:
 - tidak konsisten: memiliki 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





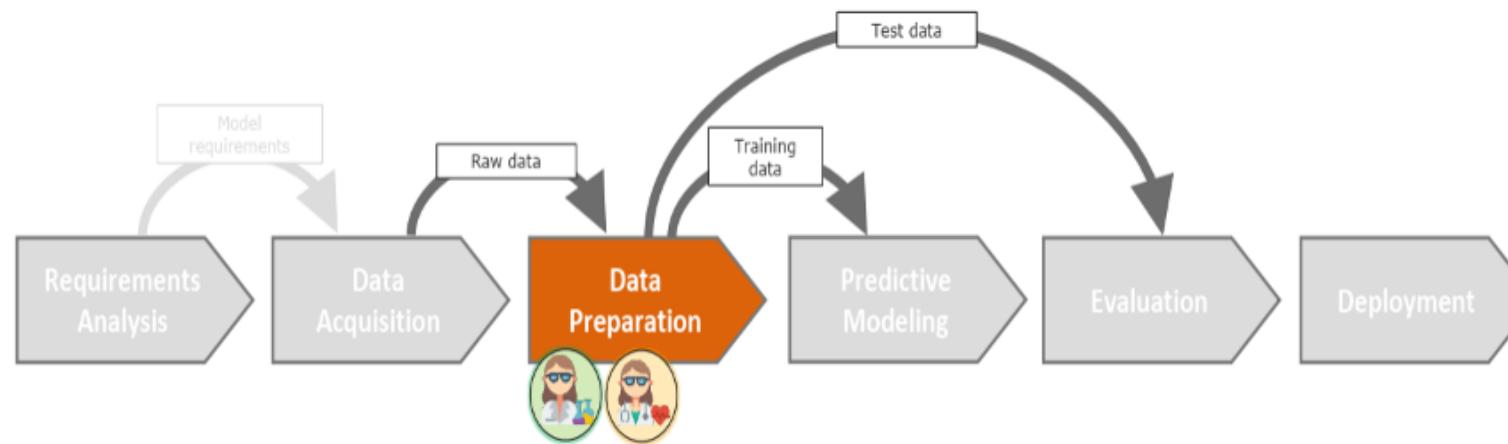
Manfaat Data Preparation

- Kompilasi Data menjadi Efisien dan Efektif (menghindari duplikasi)
- Identifikasi dan Memperbaiki Error
- Mudah Perubahan Secara Global
- Menghasilkan Informasi yang Akurat untuk Pengambilan Keputusan
- Nilai Bisnis dan ROI (Return on Investment) akan Meningkat



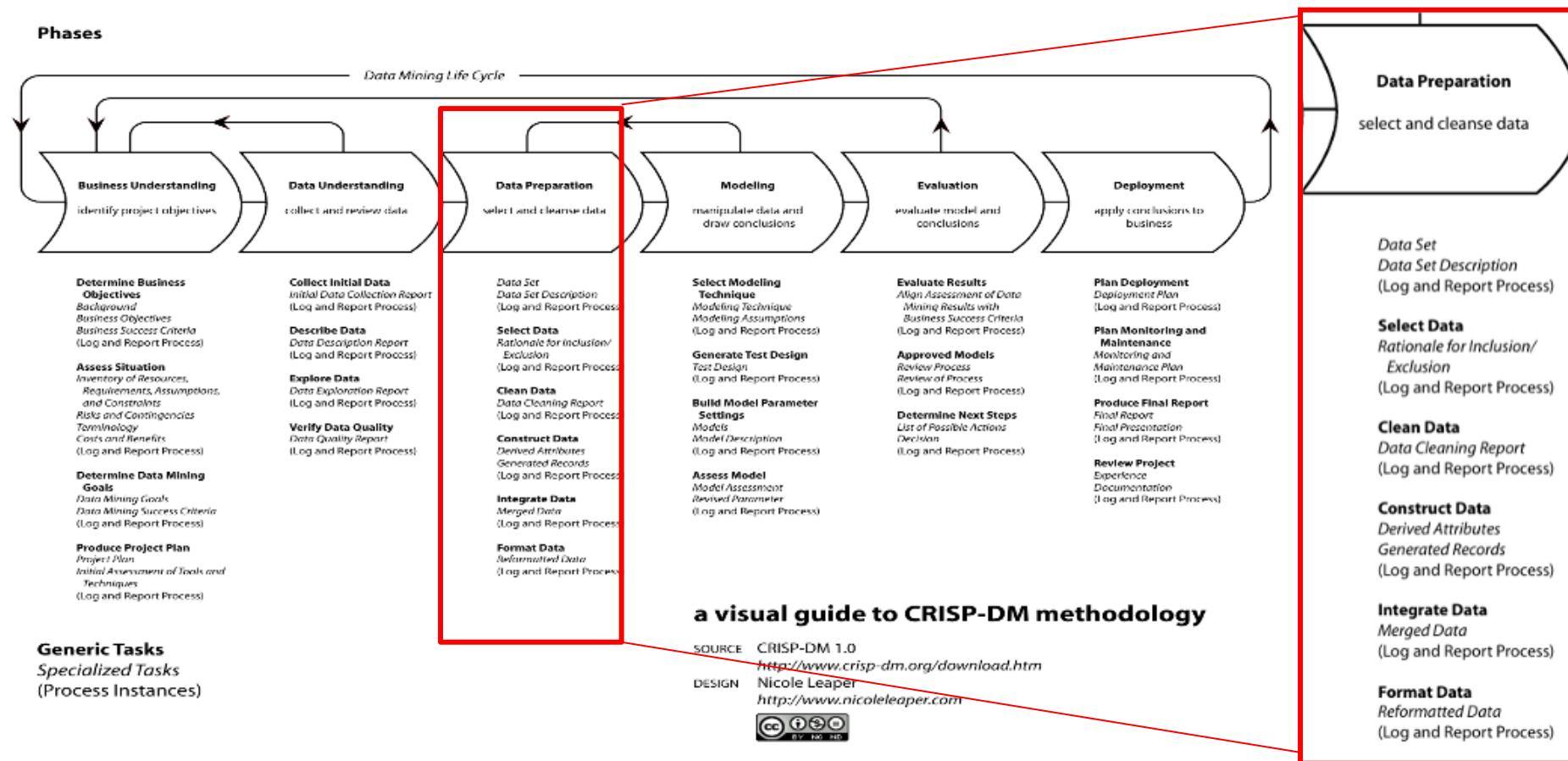


Tahapan dan Tantangan Data Preparation



- **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



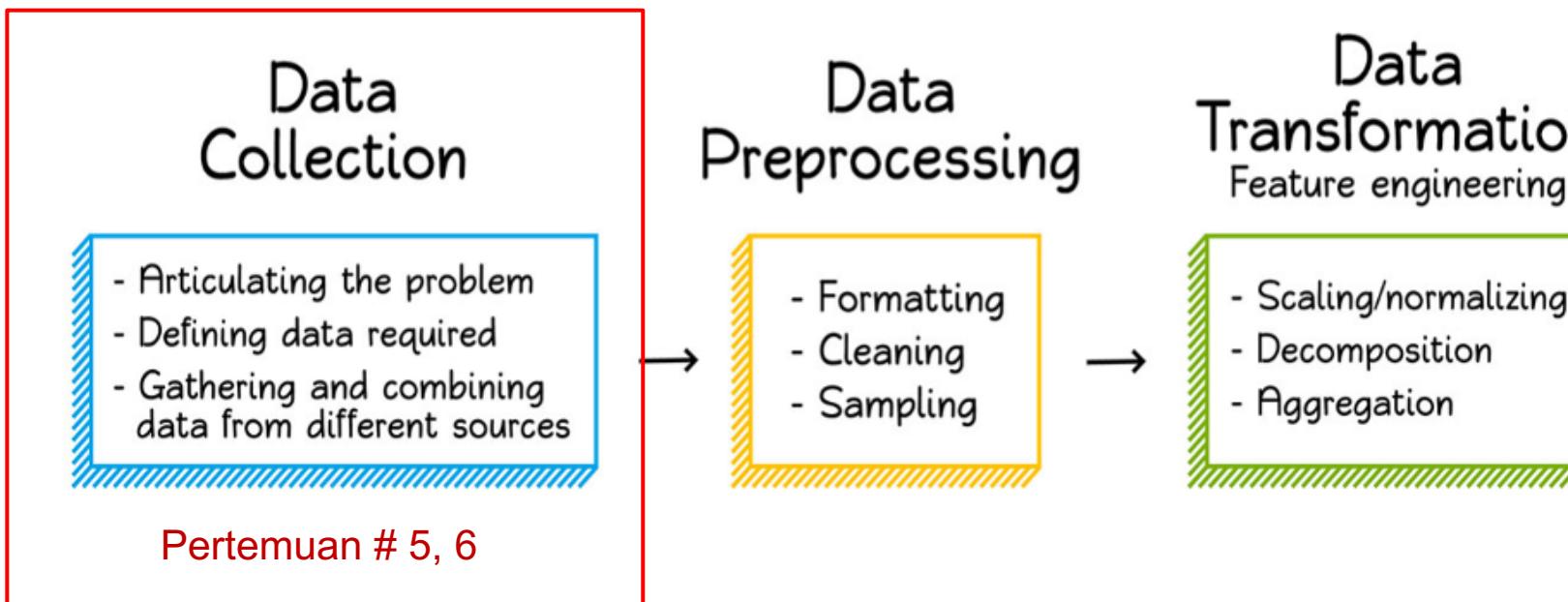
Tahapan Data Preparation: Penentuan Objek atau Memilih Data

- Pertimbangkan pemilihan 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



Tahapan Data Preparation: Versi Simple

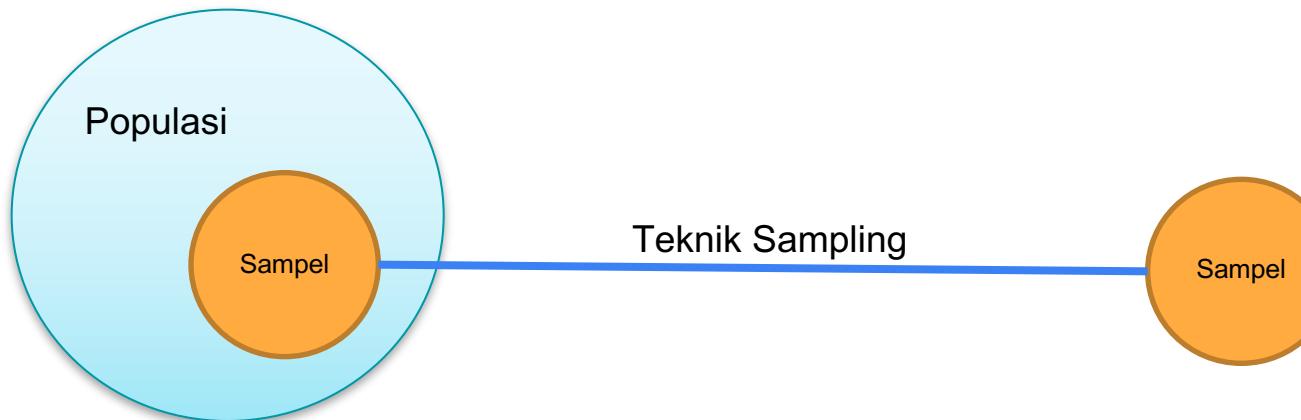
Data Preparation Process





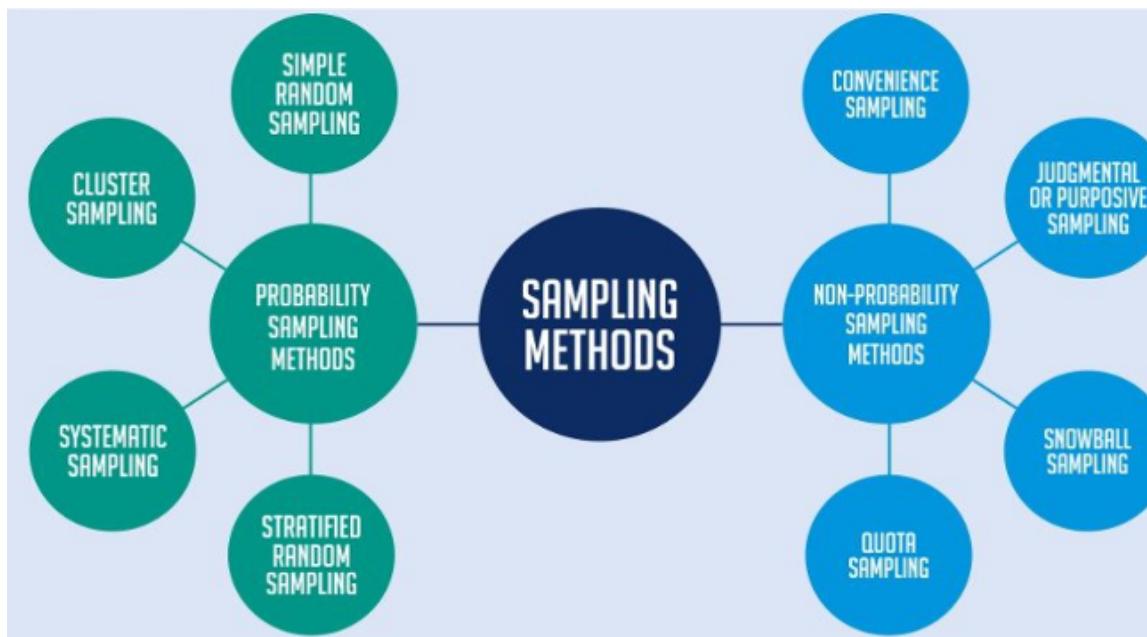
Sampling Data: Pengertian Sampling

- 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



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

Metode Sampling: Probability Sampling

- Probability Sampling adalah penarikan contoh dengan metode peluang yang dilakukan secara acak dan dapat dilakukan dengan cara undian atau tabel bilangan random.
- Sering disebut random sampling, yaitu pengambilan sampel penelitian secara random
- Teknik sampling ini cocok dipilih untuk populasi yang bersifat finit, artinya besaran anggota populasi dapat ditentukan lebih dahulu
- Setiap populasi mempunyai kesempatan yang sama untuk terpilih terutama digunakan dalam penelitian kuantitatif

Sampling Data: Metode Sampling

When to use probability sampling?

1 When you want to reduce the sampling bias
Probability sampling leads to higher quality findings because it provides an unbiased representation of the population.

2 When the population is usually diverse
This sampling method will help pick samples from various socio-economic strata, background, etc. to represent the broader population.

3 To create an accurate sample
Researchers use proven statistical methods to draw a precise sample size to obtain well-defined data.

Learn more:
www.questionpro.com/blog/probability-sampling/

QuestionPro

Types of probability sampling

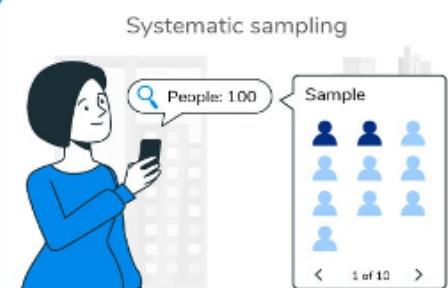
Simple random sampling



Cluster sampling



Systematic sampling



Stratified random sampling





Metode Sampling: Probability Sampling

Probability sampling dapat digunakan ketika:

- Ingin mengurangi bias pada sampel
- Populasi biasanya beragam
- Untuk membuat sampel yang akurat

Tipe-tipe Probability Sampling:

- Simple Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster Sampling

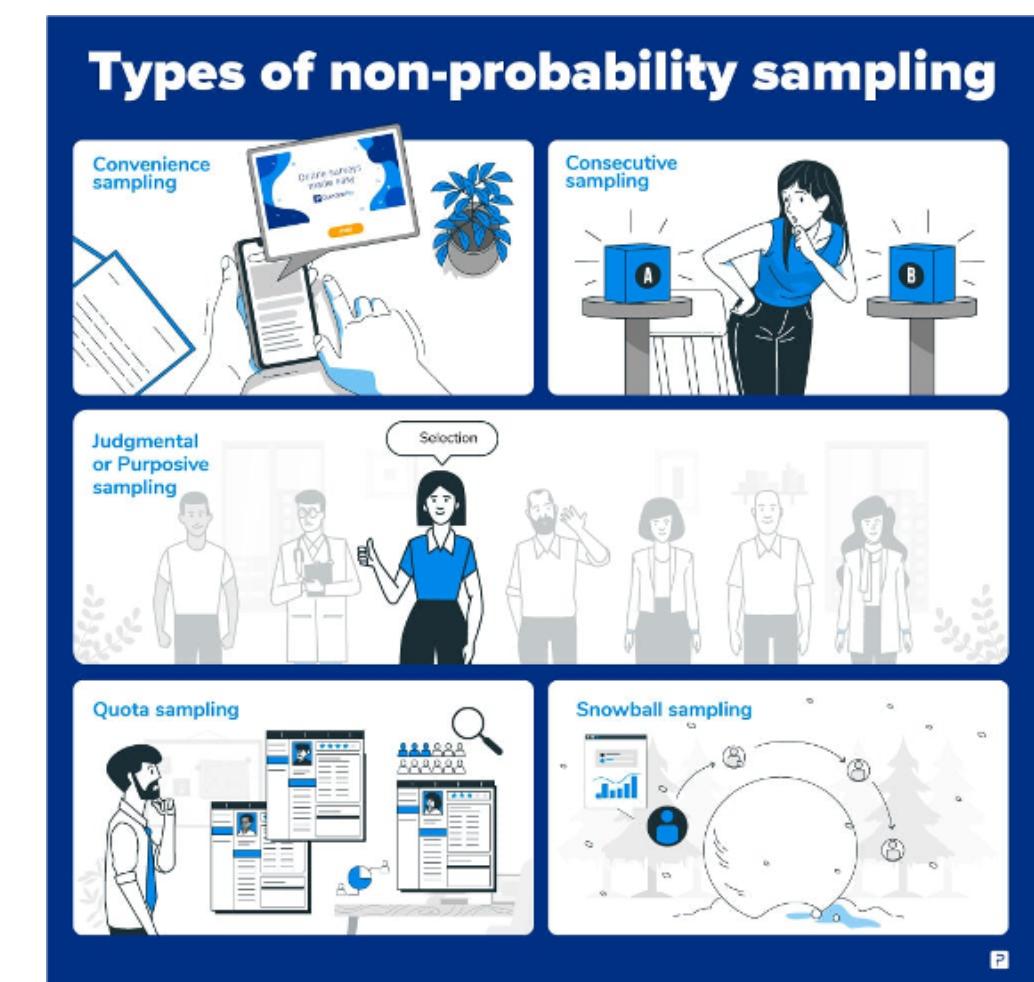
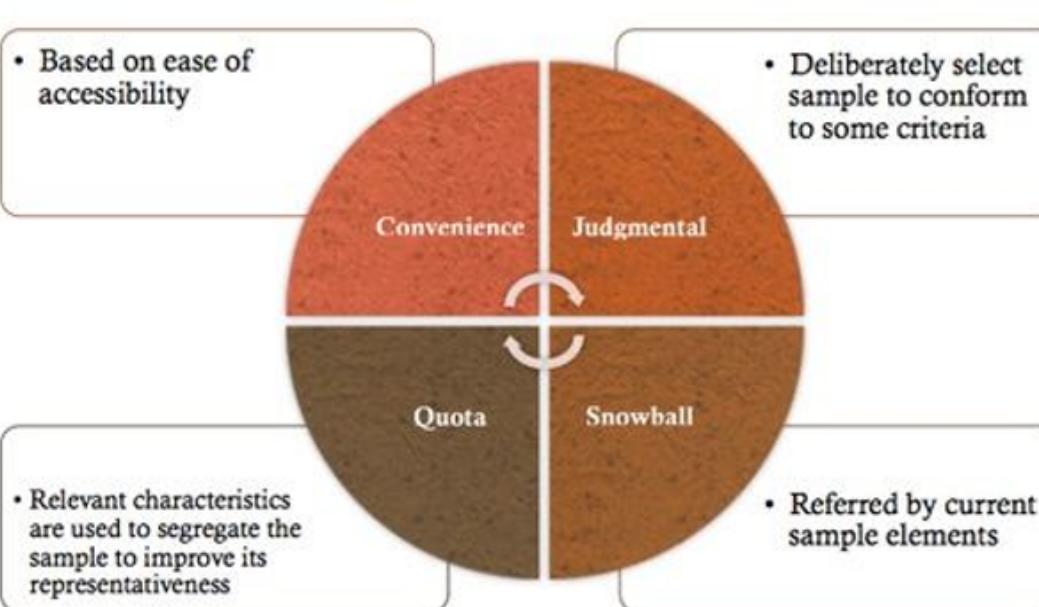
Metode Sampling: Non-Probability Sampling

- Teknik pengambilan sampel yang tidak memberi peluang atau kesempatan sama bagi setiap unsur atau anggota populasi yang dipilih menjadi sampel.
- Tidak bisa digunakan untuk membuat generalisasi.

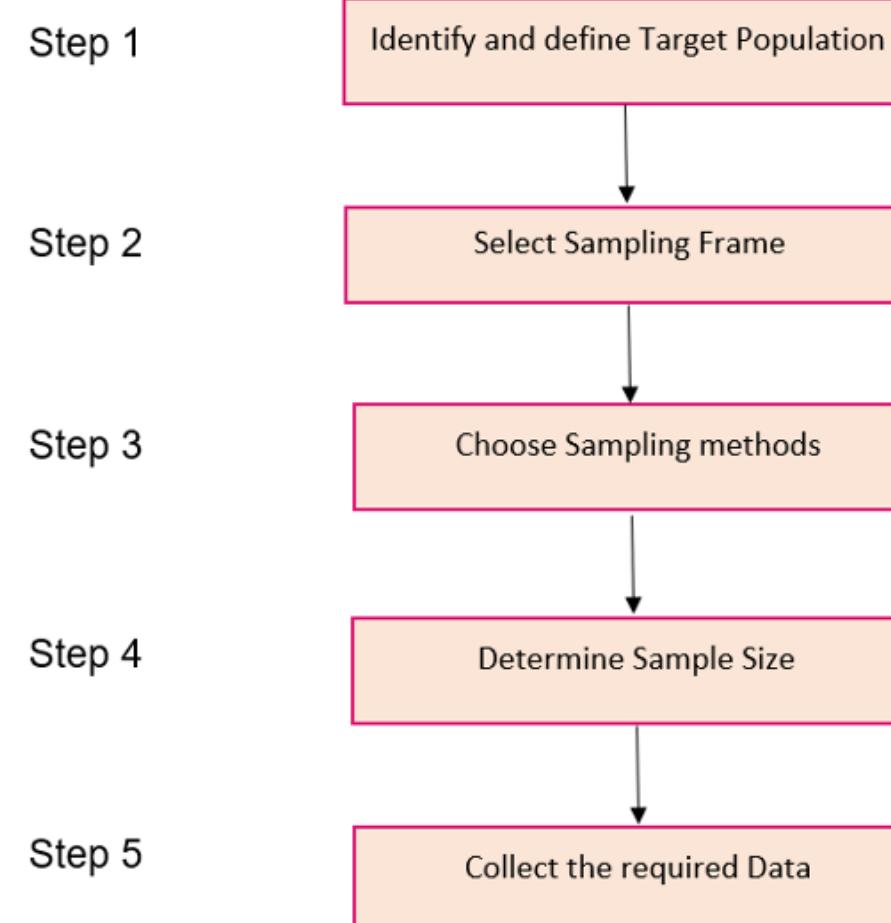


Sampling Data: Teknik Sampling

Non-Probability Methods



Sampling Data: Tahapan Sampling

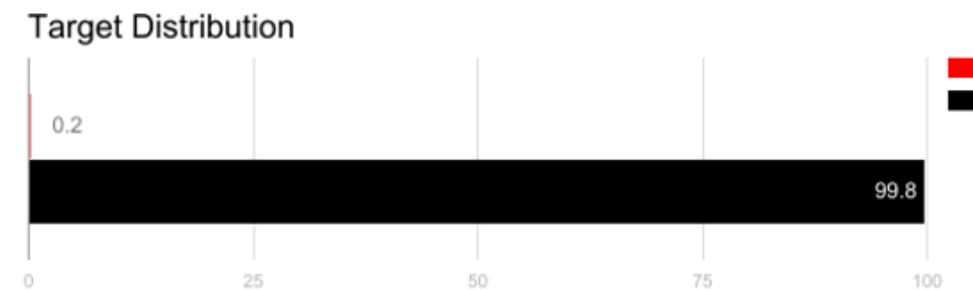


Imbalance Dataset

- Imbalanced Dataset merupakan data yang biasanya diolah secara klasifikasi dengan salah satu kelas/label pada datanya mempunyai nilai yang sangat jauh berbeda dengan jumlahnya dari kelas lainnya
- Pada imbalanced dataset biasanya memiliki data dengan kelas yang sedikit dan data dengan kelas yang banyak (abundant class)
- Contoh kasus yang sering terjadi pada imbalanced dataset: credit scoring

Imbalance Dataset: Resampling

- Ini dilakukan setelah proses pemilihan, pembersihan dan rekayasa fitur dilakukan 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 untuk dataset tentang deteksi 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 signifikan. Kebanyakan algoritma ML tidak bekerja baik untuk dataset imbalance tersebut

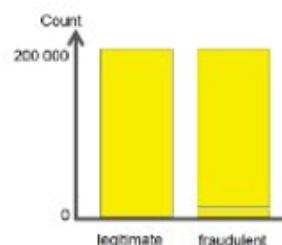
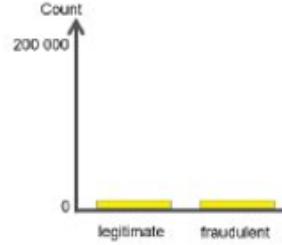


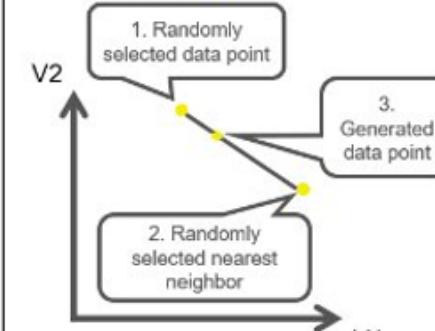
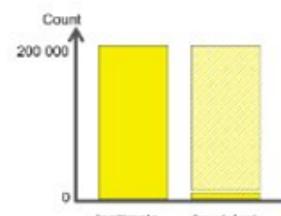
Imbalance Dataset: Resampling

- Berikut adalah beberapa cara untuk mengatasi imbalance dataset:
 - Gunakan pengukuran (metrik) yang tepat, misal dengan menggunakan:
 - **Precision:** berapa banyak instance yang relevan
 - **Recall/Sensitifitas:** berapa banyak instance yang dipilih
 - **F1 score:** harmonisasi mean dari precision dan recall
 - **Matthews correlation coefficient (MCC):** koefisien korelasi antara klasifikasi 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: Teknik Resampling

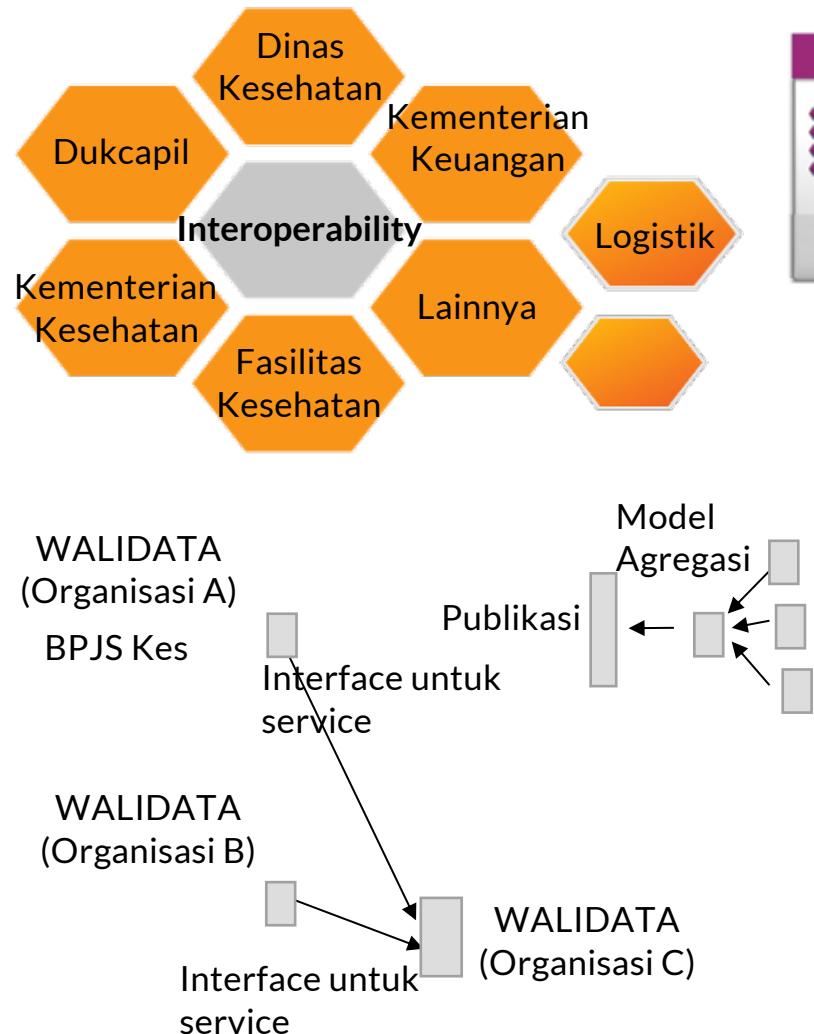
- Teknik Resampling:
 - oversampling (SMOTE)
 - oversampling (Bootstrap)
 - undersampling (Bootstrap)

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	 <table border="1"> <thead> <tr> <th>Category</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>legitimate</td> <td>~200,000</td> </tr> <tr> <td>fraudulent</td> <td>~100</td> </tr> </tbody> </table>	Category	Count	legitimate	~200,000	fraudulent	~100
Category	Count							
legitimate	~200,000							
fraudulent	~100							
Undersampling (Bootstrap)	Randomly draw with replacement as many legitimate transactions as there are fraudulent transactions	 <table border="1"> <thead> <tr> <th>Category</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>legitimate</td> <td>~100</td> </tr> <tr> <td>fraudulent</td> <td>~100</td> </tr> </tbody> </table>	Category	Count	legitimate	~100	fraudulent	~100
Category	Count							
legitimate	~100							
fraudulent	~100							

Resampling method	Description	Target class distribution after resampling						
Oversampling (SMOTE)	<p>Generate new synthetic fraudulent transactions until the number of fraudulent transactions is ca. equal to the number of legitimate transactions:</p> <ol style="list-style-type: none"> 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 	<p>Original data in yellow New synthetic data in light patterned yellow</p>  <table border="1"> <thead> <tr> <th>Category</th> <th>Count</th> </tr> </thead> <tbody> <tr> <td>legitimate</td> <td>~200,000</td> </tr> <tr> <td>fraudulent</td> <td>~200,000</td> </tr> </tbody> </table>	Category	Count	legitimate	~200,000	fraudulent	~200,000
Category	Count							
legitimate	~200,000							
fraudulent	~200,000							

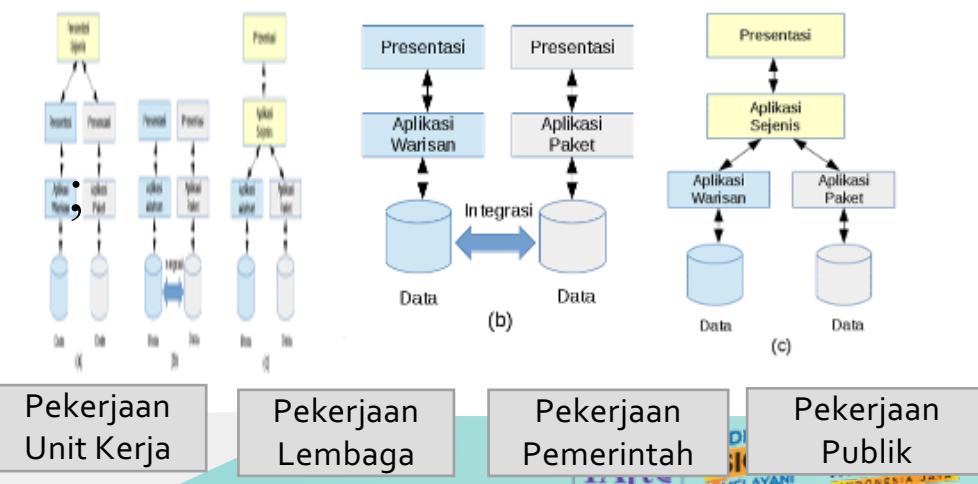


Keberadaan data dari berbagai sumber (stakeholder)



- Sistem informasi di masing-masing organisasi tidak bisa bertukar data/informasi pada lingkungan heterogen
- Interoperabilitas data akan mengefisienkan kerja

- 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.





Data
Perencanaan

Data
Pelaksanaan

Data
Pengawasan

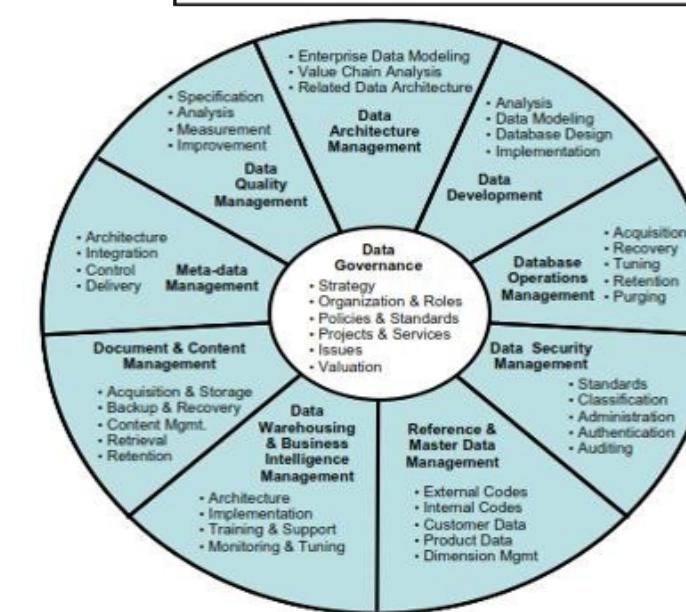
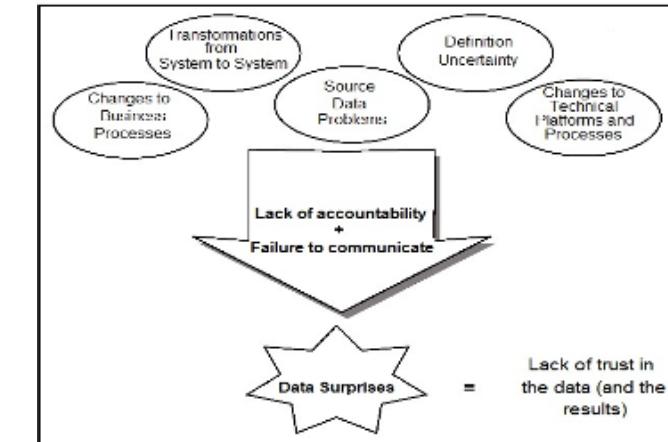
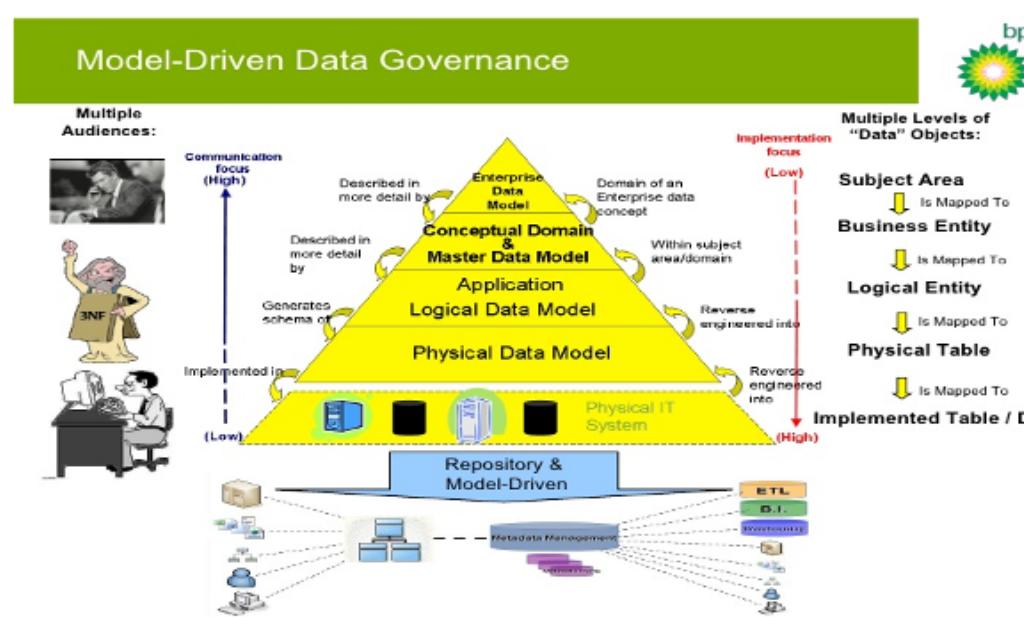
Data
Penindakan

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

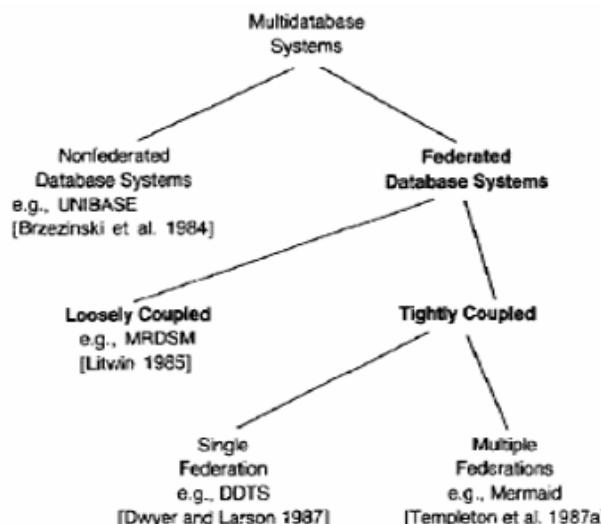
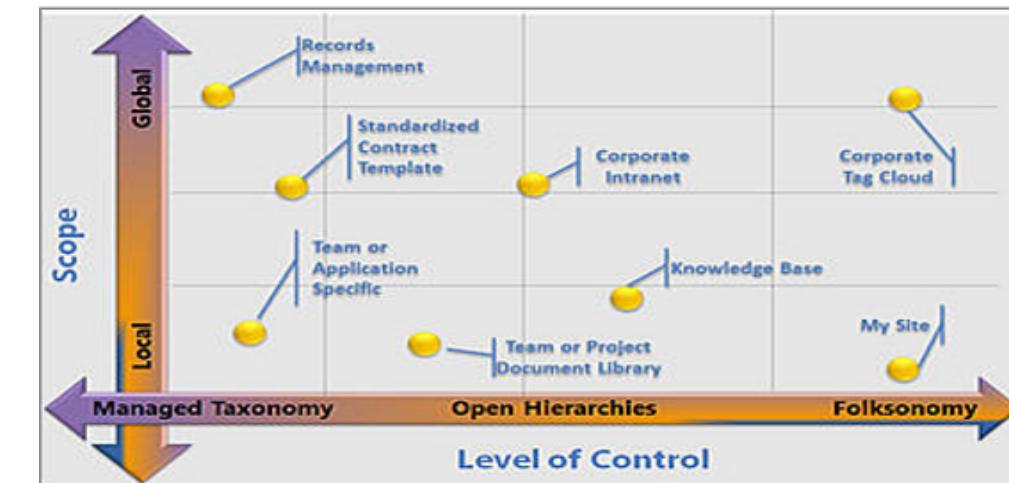


Figure 3. Taxonomy of multidatabase systems.

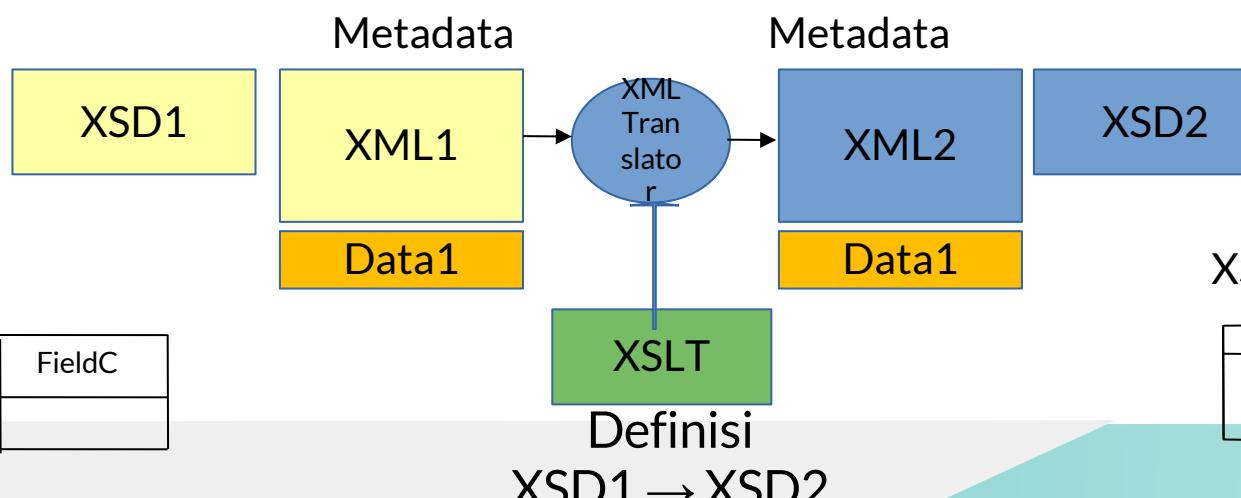
- Tidak mungkin memaksa setiap pihak “menyerahkan” datanya
- Setiap pihak memiliki teknologi dan sistem masing-masing
- Transparency
- Heterogeneity
- Functionality
- Autonomy of underlying federated sources
- Extensibility & Openness
- Optimized performance



Sistem A

XSD dari Metadata

DocID	FieldA	FieldB	FieldC

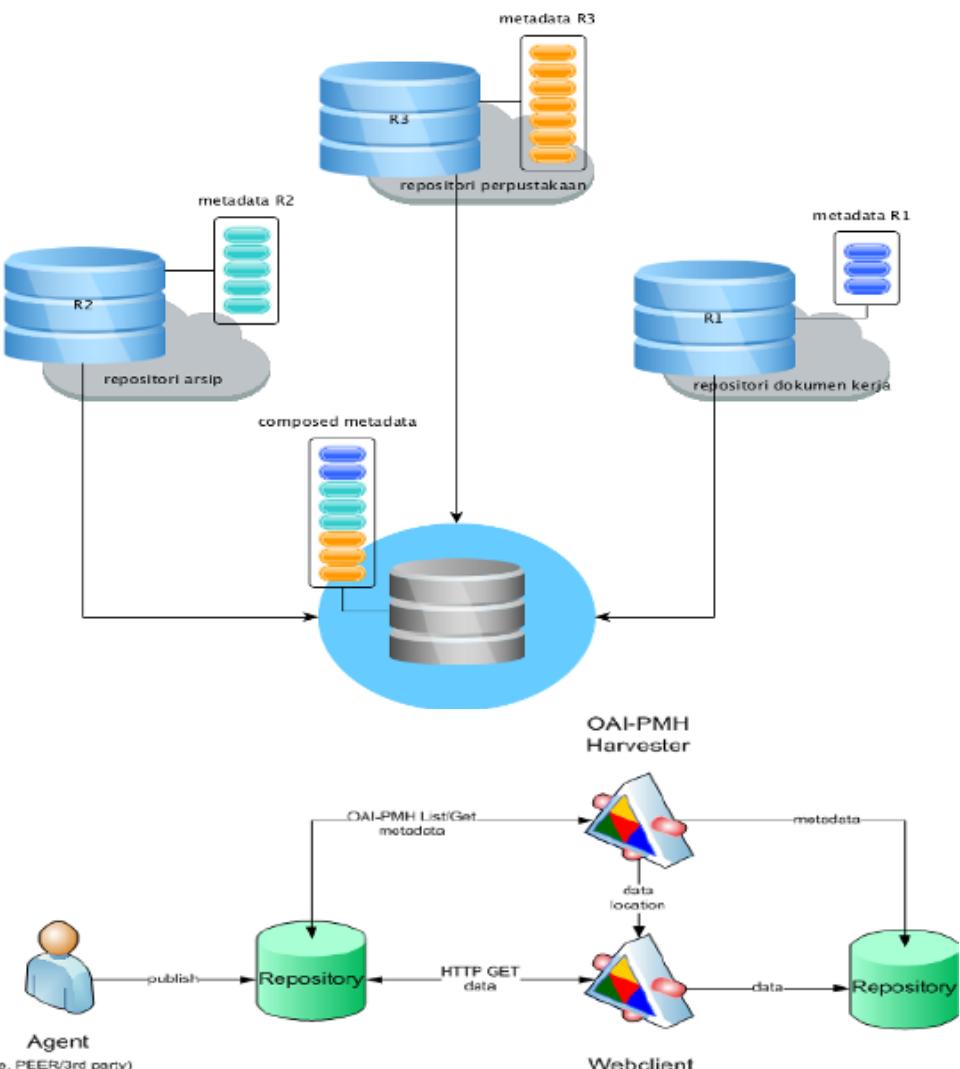


Sistem B

XSD dari Metadata

ID	FieldX	FieldY

Ontologi dan Database

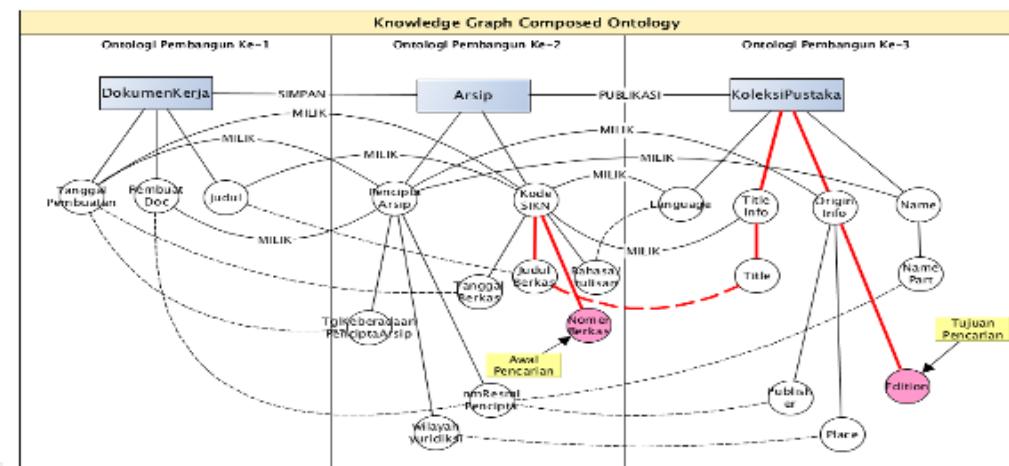


Database Relasional

- Close World Assumption (CWA), focus pada data
- Adanya Constraint untuk mencapai data integritas, namun mungkin menyembunyikan makna
- Tidak menggunakan hirarki ISA
- Skema lebih sederhana, belum tentu dapat digunakan kembali

Ontologi

- Open World Assumption (OWA), focus pada makna
- Adanya Ontology axioms untuk menspesifikasi makna, dapat digunakan untuk pencapaian integritas
- Hirarki ISA merupakan backbone
- Skema lebih kompleks dapat digunakan kembali



Pemilihan (Seleksi Fitur) Data

	K1	K2	K3	K4	K5	K6
R1						
R2						
R3						
R4						

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	$\left\{ \frac{n(n-1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s_x} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$
Skewness	$\frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s_x} \right)^3$
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

Fitur:
 Kolom yang dipilih
 Untuk sesuai
 tujuan

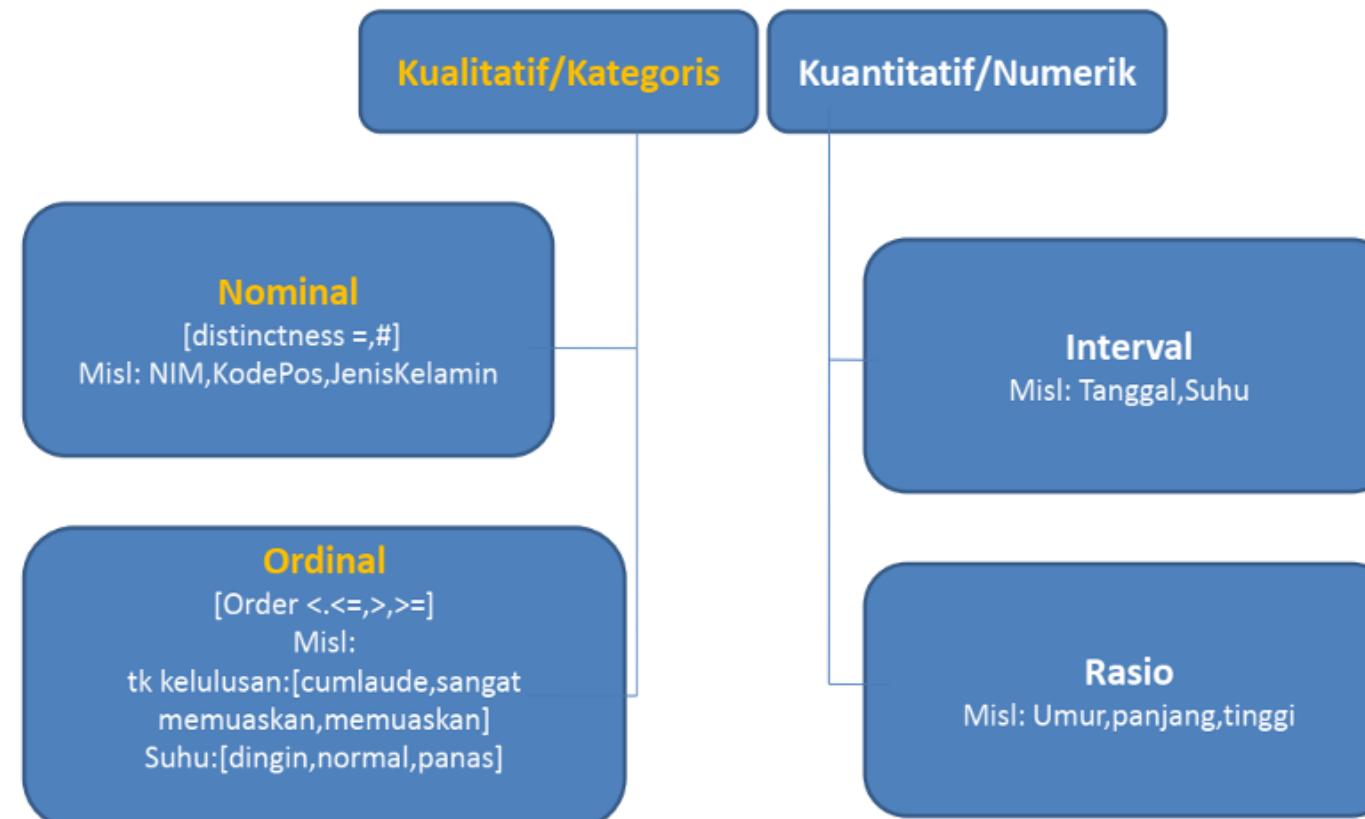
- Setelah menentukan sampling atas data yang akan diambil nanti, selanjutnya adalah melakukan seleksi **fitur** (feature selection) atas data sampling tsb --> Memilih Kolom/Atribut/Variabel yang akan diolah lebih lanjut
- Terminologi **fitur** di Data Science atau Machine Learning adalah Kolom/Atribut/Variabel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

Seleksi Fitur Data

- Manfaat:
 - Reduksi *Overfitting*: 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 algoritma 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



Hands On: Seleksi Fitur

- Dalam praktik kali ini akan digunakan 3 teknik seleksi fitur yang mudah dan memberikan hasil yang baik:
 - Seleksi Univariat (Univariate Selection)
 - Pentingnya Fitur (Feature Importance)
 - Matriks Korelasi (Correlation Matrix) dengan *Heatmap*
- 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):**
 - *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).

- **Seleksi Univariate**

Uji statistik dapat digunakan untuk memilih fitur-fitur tersebut yang memiliki relasi paling kuat dengan variabel output. Library scikit-learn menyediakan class SelectKBest yang digunakan untuk serangkaian uji statistik berbeda untuk memilih angka spesifik dari fitur. Berikut ini adalah uji statistik **chi-square** untuk fitur non-negatif untuk memilih 10 fitur terbaik dari dataset *Mobile Price Range Prediction*.

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 colums
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
```

Import library / modul yang dibutuhkan

Load datasets, sesuaikan dengan path direktori masing-masing

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

	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

Output

Hands On: Seleksi Fitur

- **Feature Importance (FT)**
 - FT berfungsi memberi skor untuk setiap fitur data, semakin tinggi skor semakin penting atau relevan fitur tersebut terhadap variabel output
 - FT merupakan kelas inbuilt yang dilengkapi dengan Pengklasifikasi Berbasis Pohon (Tree Based Classifier), kita akan menggunakan Pengklasifikasi Pohon Ekstra untuk mengekstraksi 10 fitur 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()
```

Mendefinisikan model yang akan digunakan yaitu menggunakan algoritma **ExtraTreesClassifier**.

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

`.nlargest(10)`: membuat plotting 10 data teratas.

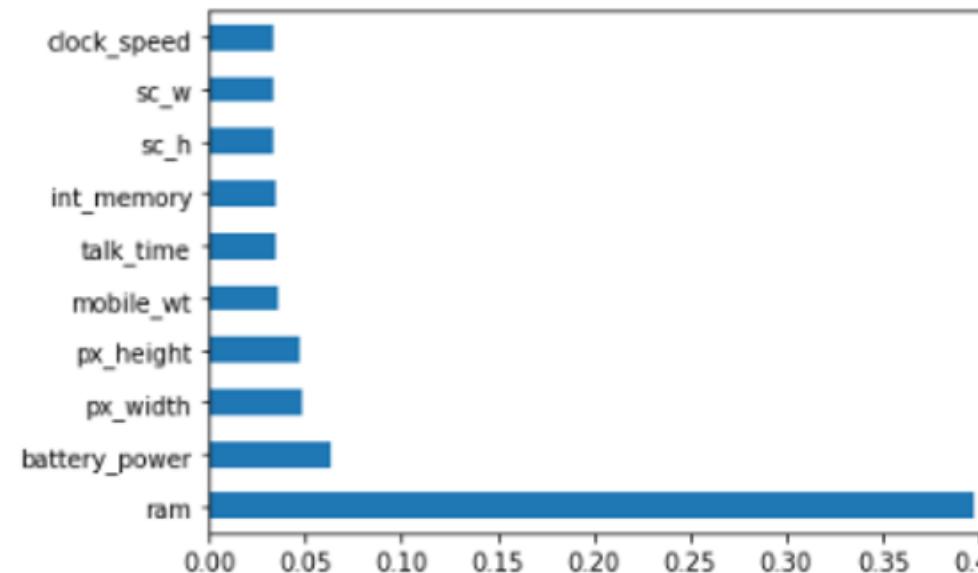
`.plot(kind='barh')` : untuk membuat jenis plot diagram batang horizontal



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 fitur terkait satu sama lain atau variabel target.
- Korelasi bisa positif (kenaikan satu nilai fitur meningkatkan nilai variabel target) atau negatif (kenaikan satu nilai fitur menurunkan nilai variabel target)
- *Heatmap* memudahkan untuk mengidentifikasi fitur mana yang paling terkait dengan variabel target, kami akan memplot peta panas fitur yang berkorelasi menggunakan seaborn library

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

figsize(): ukuran dari figure, mengambil dua parameter lebar dan tinggi (dalam inci)

```
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")
```

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

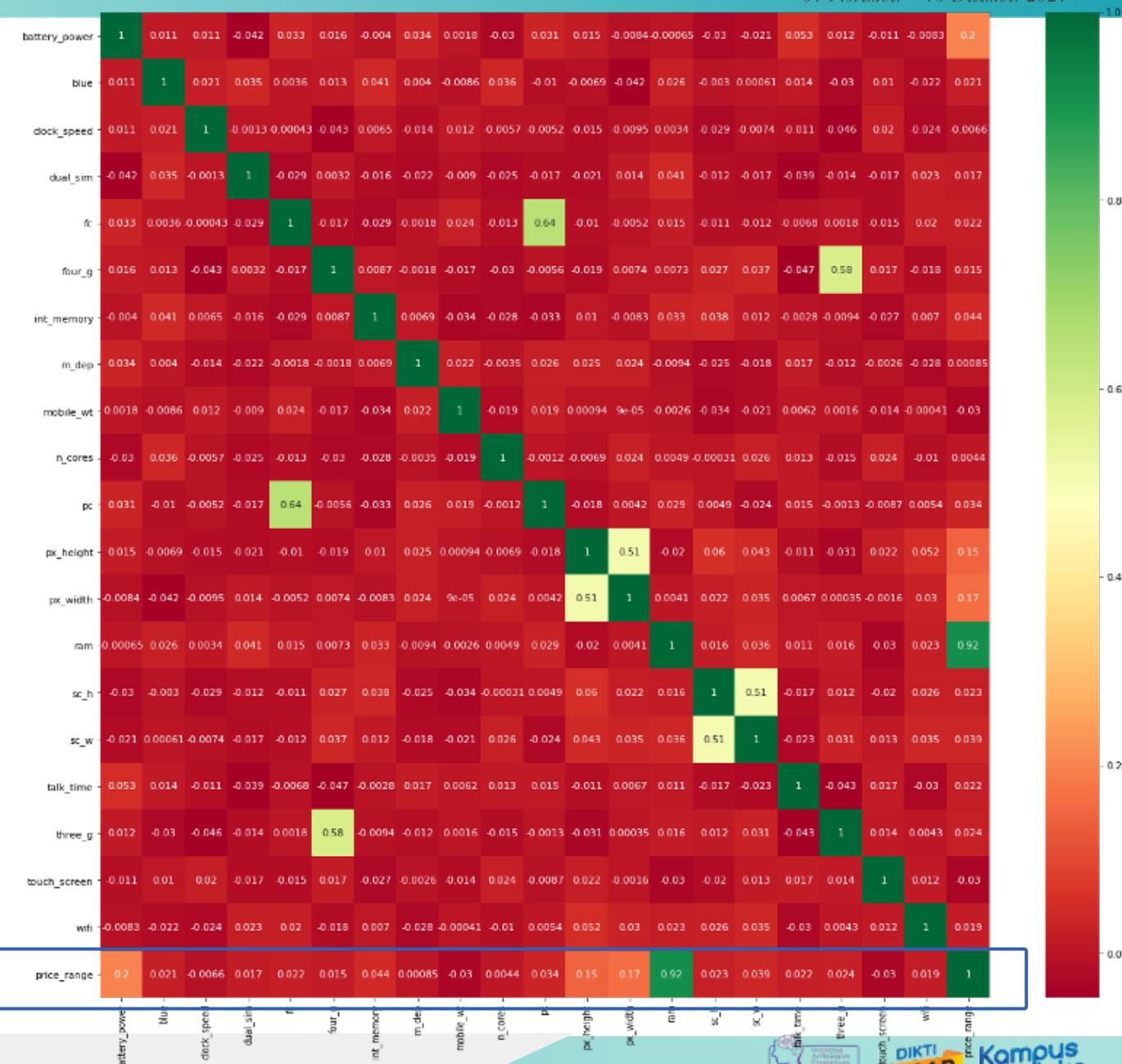
annot=True untuk menampilkan korelasi antar atribut. Jika nilai korelasi mendekati 1 maka hubungan antar atribut semakin tinggi



Hands On: Seleksi Fitur

- Matriks Korelasi dengan Heatmap (lanjutan)**

- lihat pada baris terakhir yaitu *price range*, korelasi antara *price range* dengan fitur lain dimana ada relasi kuat dengan variabel *ram* dan diikuti oleh var *battery power* , *px height* and *px width*.
- sedangkan untuk var *clock speed* dan *n_cores* berkorelasi lemah dengan *price range*



Referensi

- Krensky P. Data PreTools: Goals, Benefits, and The Advantage of Hadoop. Aberdeen Group Report. July 2015
- SAS. Data Preparation Challenges Facing Every Enterprise. ebook. December 2017
- <https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=6e9aa0e36f63>
- <https://improvado.io/blog/what-is-data-preparation>
- <https://searchenterpriseai.techtarget.com/feature/Data-preparation-for-machine-learning-still-requires-humans?>
- <https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>
- CRISP-DM



Tools Lab Online

- jupyter notebook
- scikit-learn
- pandas
- numpy

Ringkasan

- Data preparation memiliki sebutan lain, diantaranya data pre-processing, data cleaning, data manipulation.
- Data preparation mengambil porsi kerja terbanyak dalam data science 60-80%
- Data preparation membutuhkan ketelitian dan kesabaran/kerajinan dari peneliti DS, terutama pemula
- Seleksi Fitur harus dilakukan di awal tahapan data preparation setelah melakukan penentuan metode/teknik sampling



Quiz / Tugas

Quiz dapat diakses melalui <https://spadadikti.id/>

Terima kasih