

**Materi III**

# Data Preparation

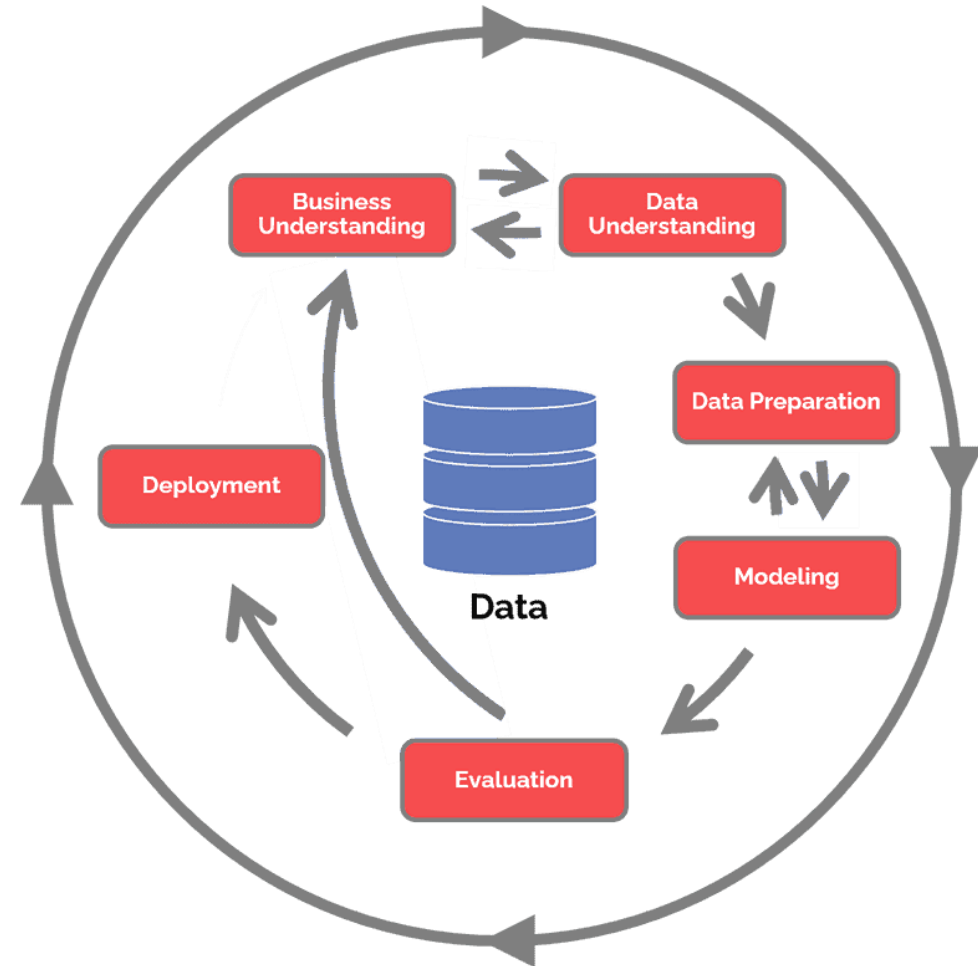


Reza R Pratama  
Ade Satya Wahana

# Data Mining Lifecycle

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

*Notice the iteration!*



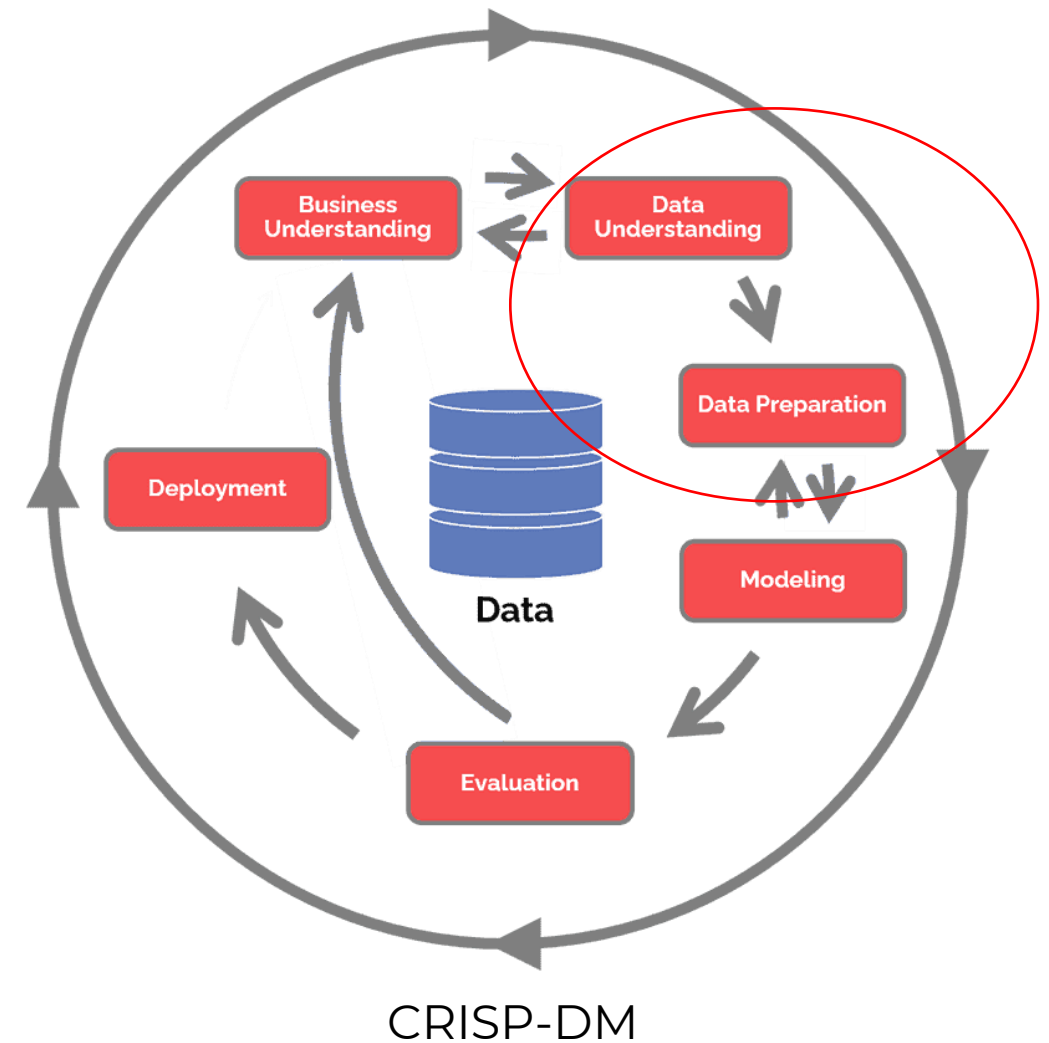
THE FAMOUS CRISP-DM

# Data Preparation

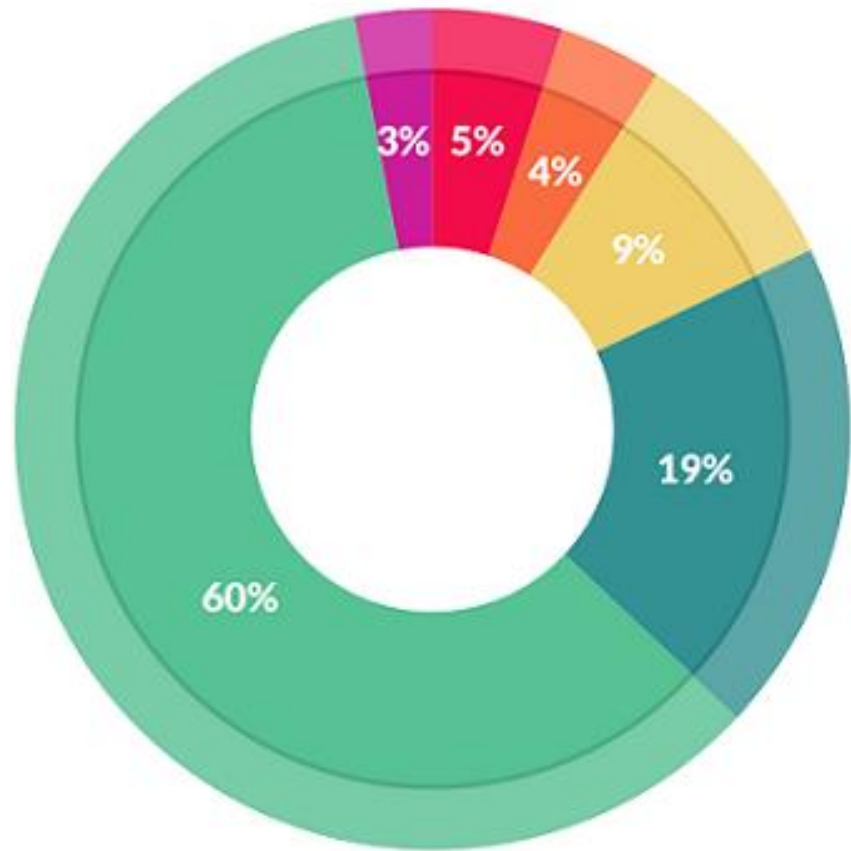
Suatu proses yang dilakukan untuk membuat **data mentah** menjadi **data yang berkualitas dan siap untuk digunakan** dalam proses analisis atau permodelan.

- Semua aktivitas untuk membuat data final (raw to final dataset)
- Kemungkinan dikerjakan beberapa kali dan bisa tidak berurutan
- Pemilihan table, record, dan atribut serta transformasi dan cleaning data

Chapman et al (2000)



# Data Preparation is Very Time Consuming!



Sumber: Forbes

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%

# Why we need preprocessing?

## Real World Data are Dirty

- Tidak Lengkap (banyak data kosong)
- Noisy / banyak outlier
- Tidak Berkualitas (tidak konsisten, tidak akurat, dll)

## Some Information are hidden within data

- Informasi dapat diekstrak dari data yang ada (umur dapat dihitung dari tanggal lahir)
- Kadang informasi pada data harus disajikan secara eksplisit untuk meningkatkan performa model

## Machine learning model's performance depends on data

- Beberapa hanya bisa memproses data dalam bentuk numerical
- Sensitif terhadap outlier
- Beberapa model memiliki persyaratan (4 asumsi klasik linear model, NN butuh data pada range 0-1)

# Outline!

## Data Cleansing

- Memperbaiki data dan tabel
- Menghapus duplicate dan missing data
- Data Outlier

## Feature Engineering

- Domain specific (Derived Features)
- Binning / Encoding
- Vectorizer
- \*Text and Date can generate many feature

## Feature Transformation

- Standardization (0,1)
- Scaling

### PS:

Hari ini kita hanya akan mempelajari sebagian metode data preprocessing  
Penggunaan di case lain bisa jadi akan berbeda, apabila ada masalah – Google is our friend!

# Memperbaiki Data & Label

Seringkali dataset yang kita peroleh format datanya masih tidak sesuai dengan yang diharapkan.

## **Misalnya:**

- Data yang masih tersebar dalam banyak file.
- Kolom yang berisi nilai uang namun dianggap sebagai string karena ada simbol mata uang di depannya.
- Data yang disajikan tidak konsisten.

# Handling Missing Value

## Remove Rows / Columns

remove column if n  
missing rows  $\gg$  n rows

remove row if n  
missing rows  $\ll$  n rows

## Value Imputation

mode / most frequent  
(categorical)

mean / median  
(numerical)

Random / defined  
value

## Model based Imputation

Use other features to  
predict missing rows



# Outliers

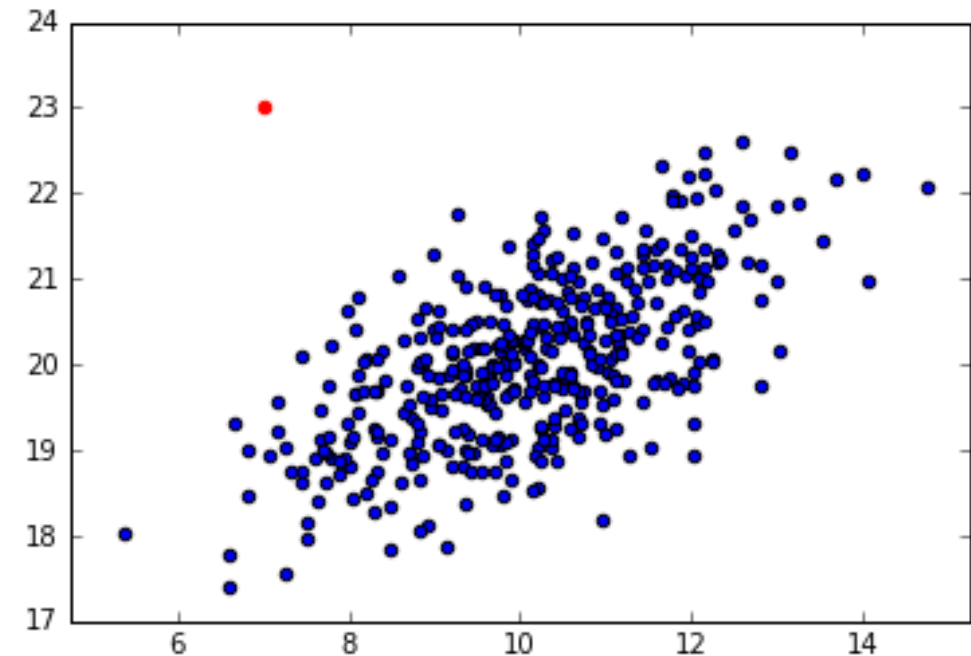
Merupakan suatu observasi yang berada di tempat yang jauh berbeda dibandingkan observasi lainnya dalam suatu populasi.

**Contoh:** hasil survei kekayaan

Definisi/Batasan outlier berbeda-beda, tergantung bagaimana analisis menyikapi datanya.

## Pengecekan outlier:

Memeriksa data menyeluruh melalui gambar/grafik dan/atau mencari data yang berbeda jauh dengan titik data secara umum



# Feature Engineering

Menggunakan domain knowledge.

Proses untuk mengekstraksi karakteristik, properti, dan atribut dari raw data

ai.stanford.edu

## **Intinya:**

- Menambah Fitur
- Mengurangi Fitur

## **How?**






- Brainstorming or testing features;
- Deciding what features to create;
- Creating features;
- Testing the impact of the identified features on the task;
- Improving your features if needed;
- Repeat.

# Encoding






- Mengubah data categorical menjadi numerical
- Sebagian besar model tidak bisa menerima data categorical

## Metode Encoding

- Label encoding – data ordinal atau tidak?
- One hot encoding – hati-hati curse of dimensionality

Gender	Is_Male	Is_Female
	0	1
	0	1
	1	0
	0	1
	1	0

**One Hot Encoding**

Tree	Type
	1
	2
	1
	2
	3

**Label Encoding**

# Feature Transformation

## Perbaikan data

- String -- huruf besar/kecil
- Numerical -- unit pengukuran sama

## Scaling

- Scaling ke dalam ukuran tertentu
- Misal 0 s.d. 100 atau 0 s.d. 1
- Beberapa fitur memiliki scaling yang berbeda – samakan!

Standardization

Normalization

## Transformasi Data

- Mentransformasi data agar terdistribusi normal
- Menggunakan log, kuadrat (power), atau akar (squareroot)

# Praktik Data Preparation

