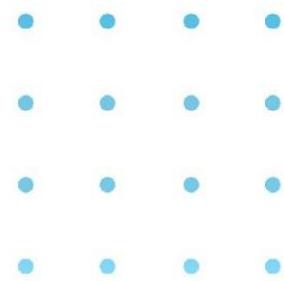


AI Mastery Course

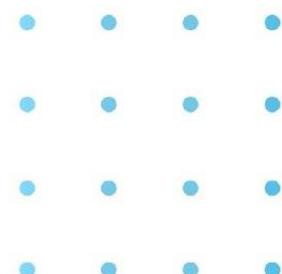


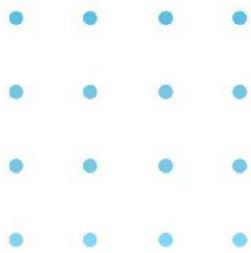
Module 1

Data Science

Section

Introduction Data Science

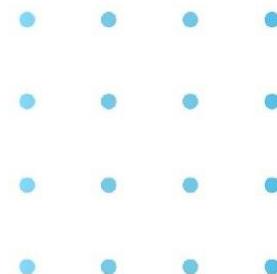




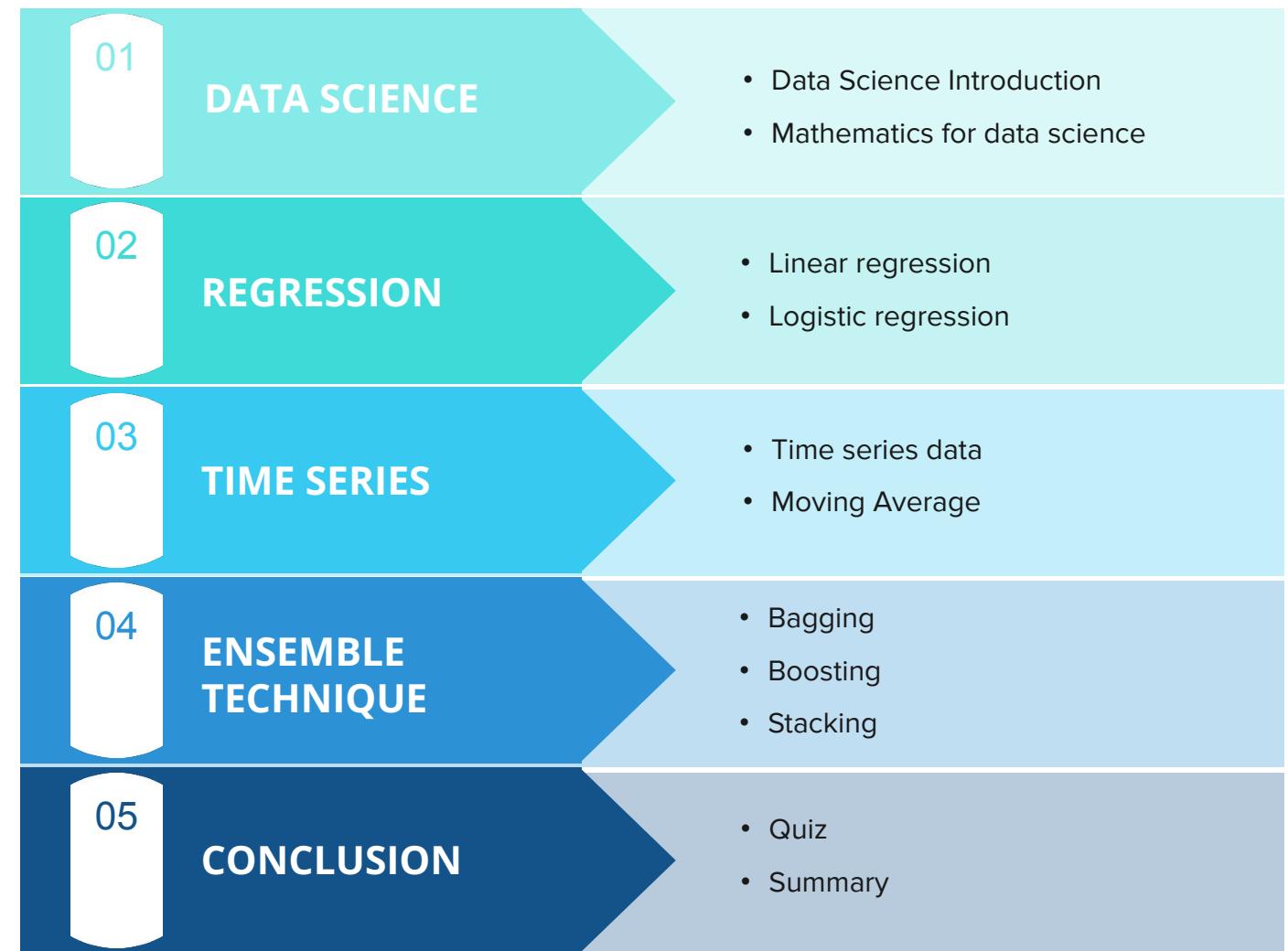
Learning Objectives

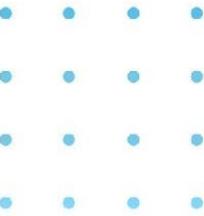
Di akhir modul ini, Anda akan dapat:

- Memahami tentang data science
- Mengerti data yang terdapat pada data science
- Mengerti manfaat data science
- Memahami model / algoritma pada data science



Agenda

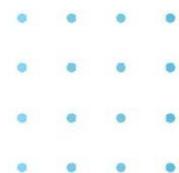




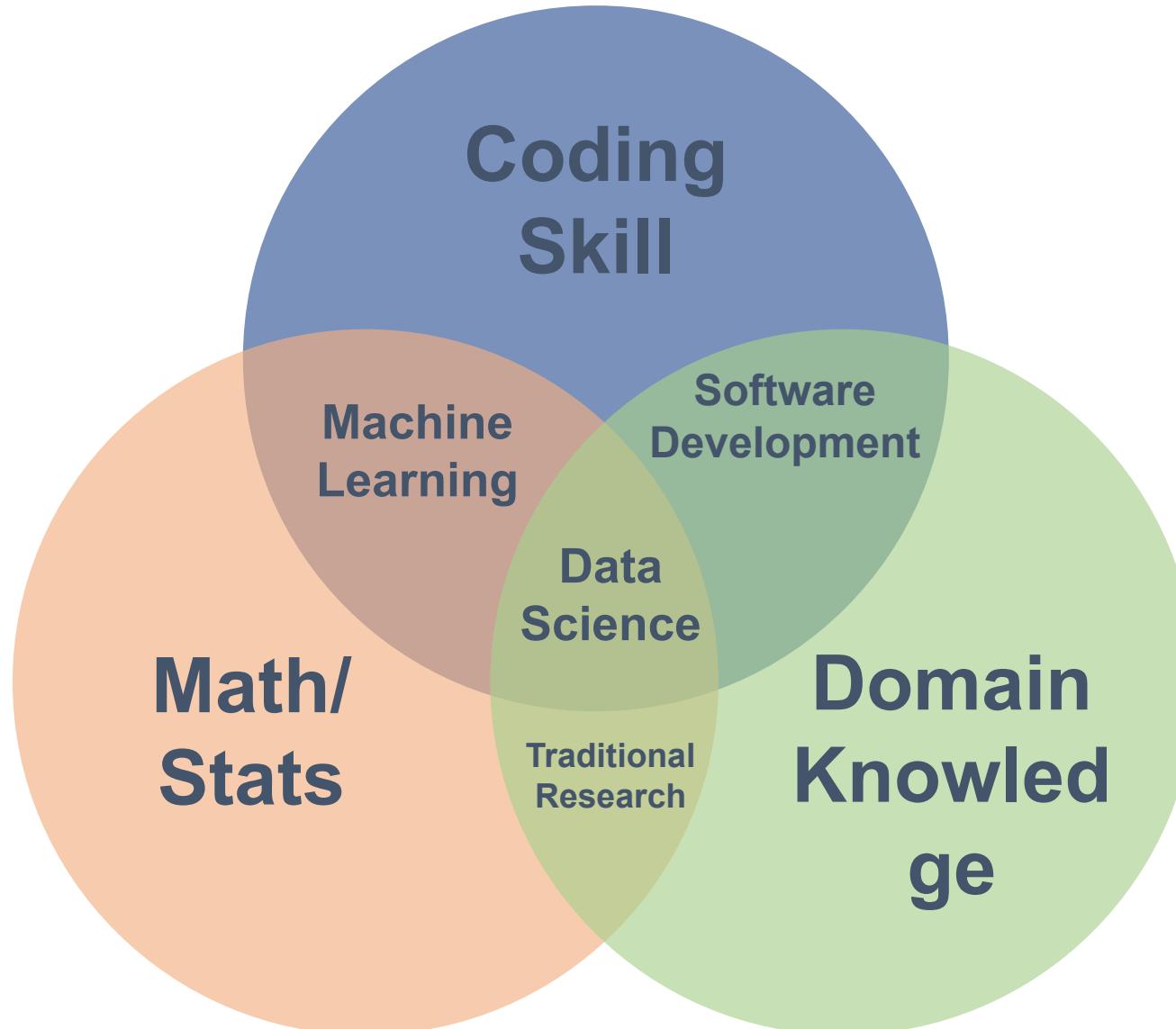
01

DATA SCIENCE

- Data science introduction
- Mathematics for data science



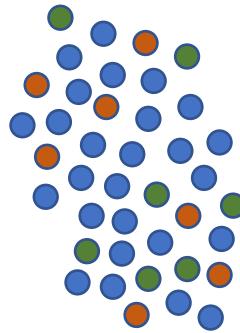
Data Science Introduction



Pilar kemampuan data science:

- Memiliki kemampuan coding pada suatu bahasa pemrograman
- Memiliki kemampuan analisa statistik / matematika
- Memahami case yang sedang diamati

Data Science Introduction



Raw Data

Informasi :

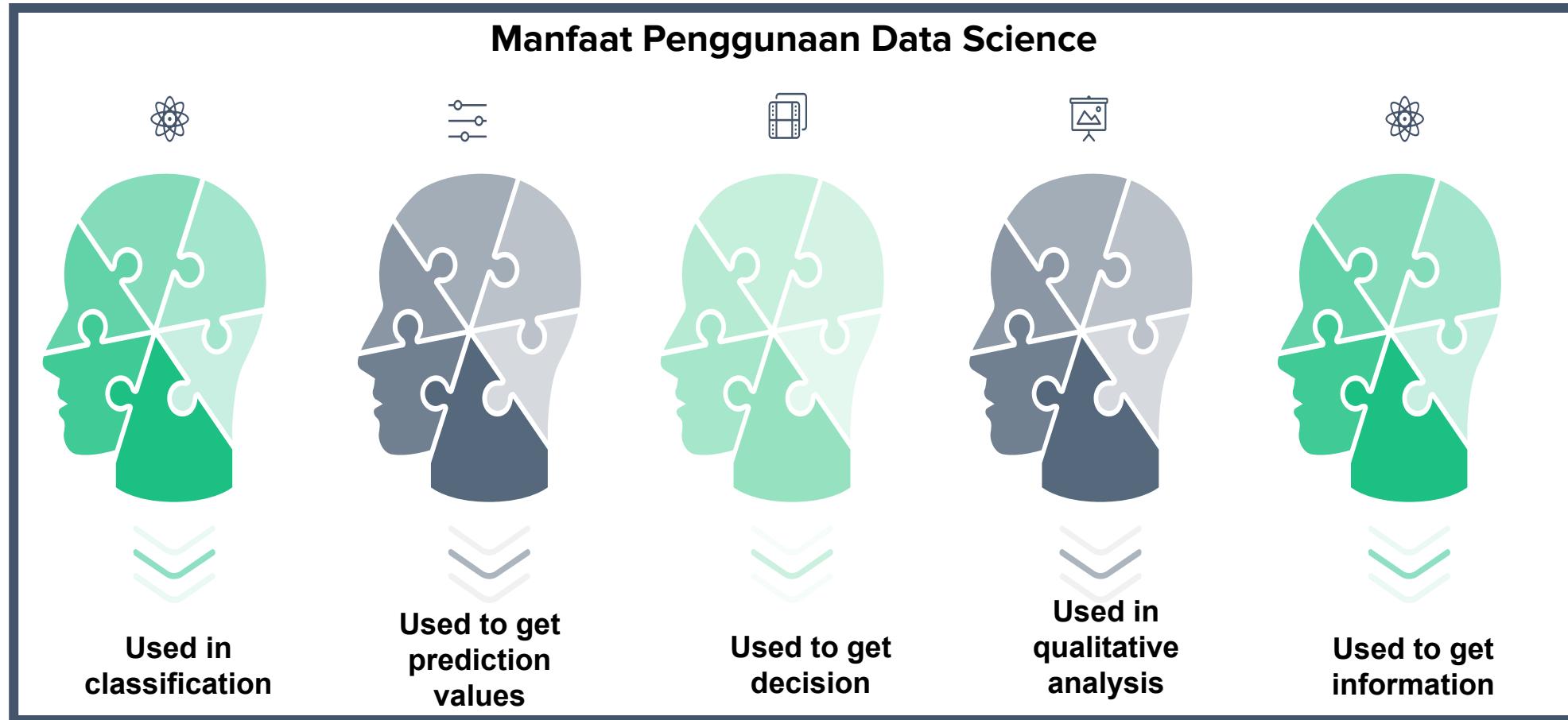
Sosial

Bisnis

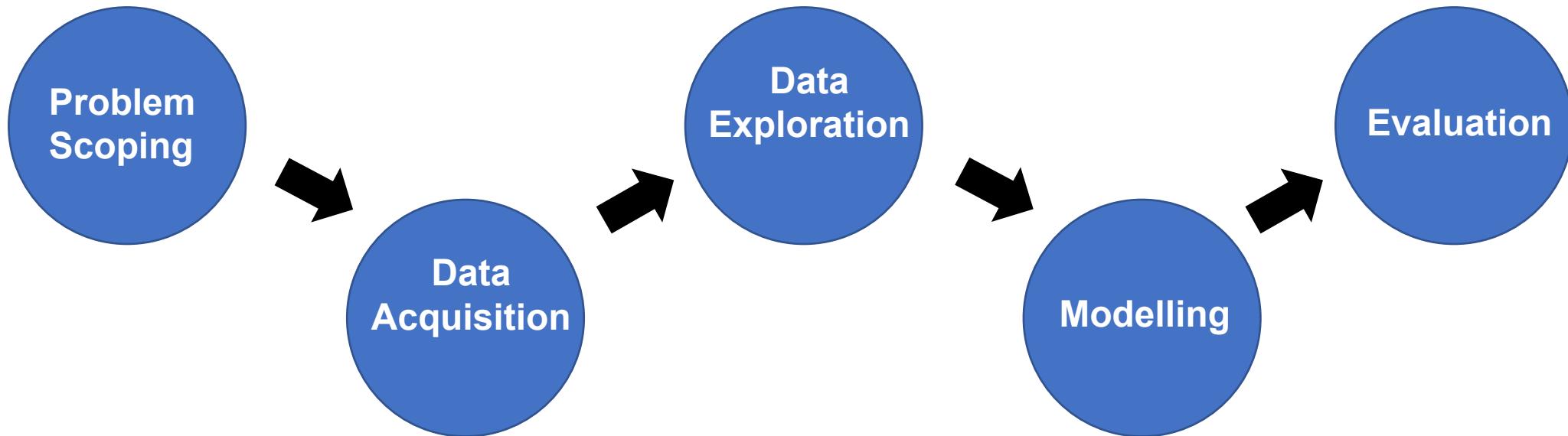
Kesehatan

Kebijakan

Data Science Introduction

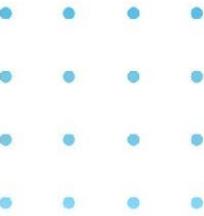


Data Science Project Cycle



Mathematics for Data Science

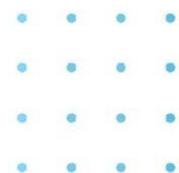
- Teori peluang
- Teori optimasi
- Aljabar linier
- Fungsi diferensial



02

Regression

- Linear Regression
- Logistic Regression



Regression

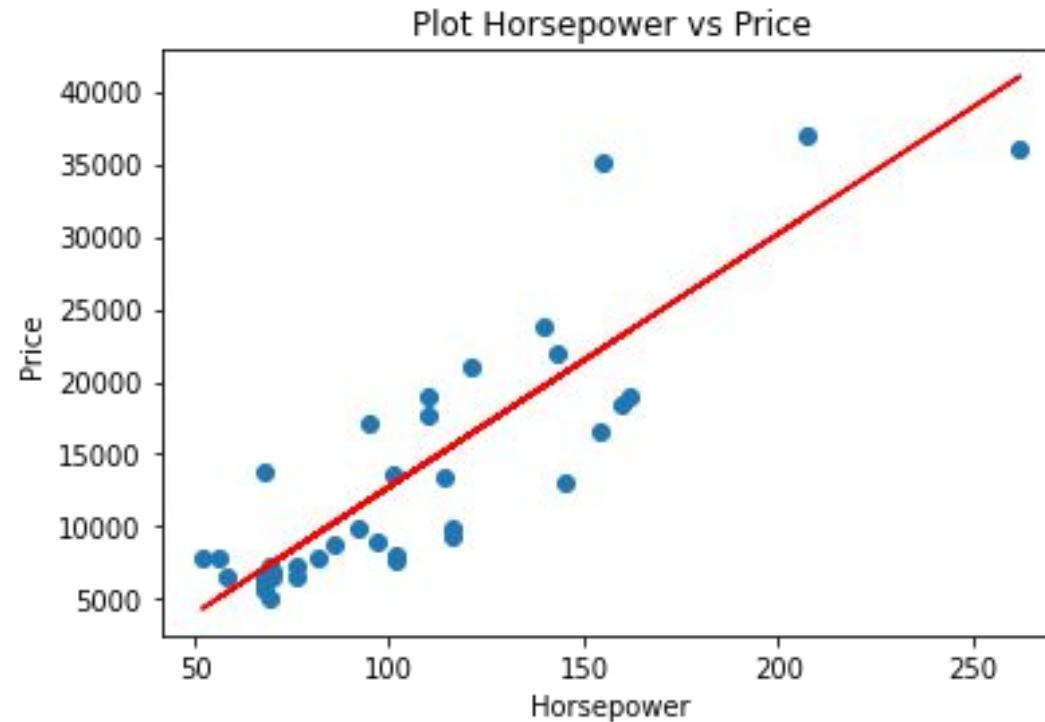


Regression



Representasi hubungan dari suatu variabel/fitur/case yang mempengaruhi variabel/fitur/case yang lain.

Linear Regression



Menganalisa hubungan antara pengaruh besaran horsepower produk mobil terhadap harga jual mobil.

Linear Regression

- Model yang digunakan untuk data dengan variabel target bertipe numerik
- Model yang digunakan untuk prediksi nilai variabel target
- Representasi hubungan dari 1 atau banyak variabel prediktor terhadap variabel target

Linear Regression

Model Persamaan :

$$\hat{y} = \epsilon$$

$$\beta = -$$

$$\alpha = -$$

Linear Regression

Contoh :

Kamera (x)	Harga (y)
8	7
2	3
6	7
4	2
7	8
3	3

Linear Regression

Kamera (x)	Harga (y)
8	7
2	3
6	7
4	2
7	8
3	3

$$\beta = \frac{6(177) - (30 \times 30)}{6(178) - (900)}$$

$$\beta = 0,96$$

$$\alpha = \frac{30(178) - (30 \times 177)}{6(178) - (900)}$$

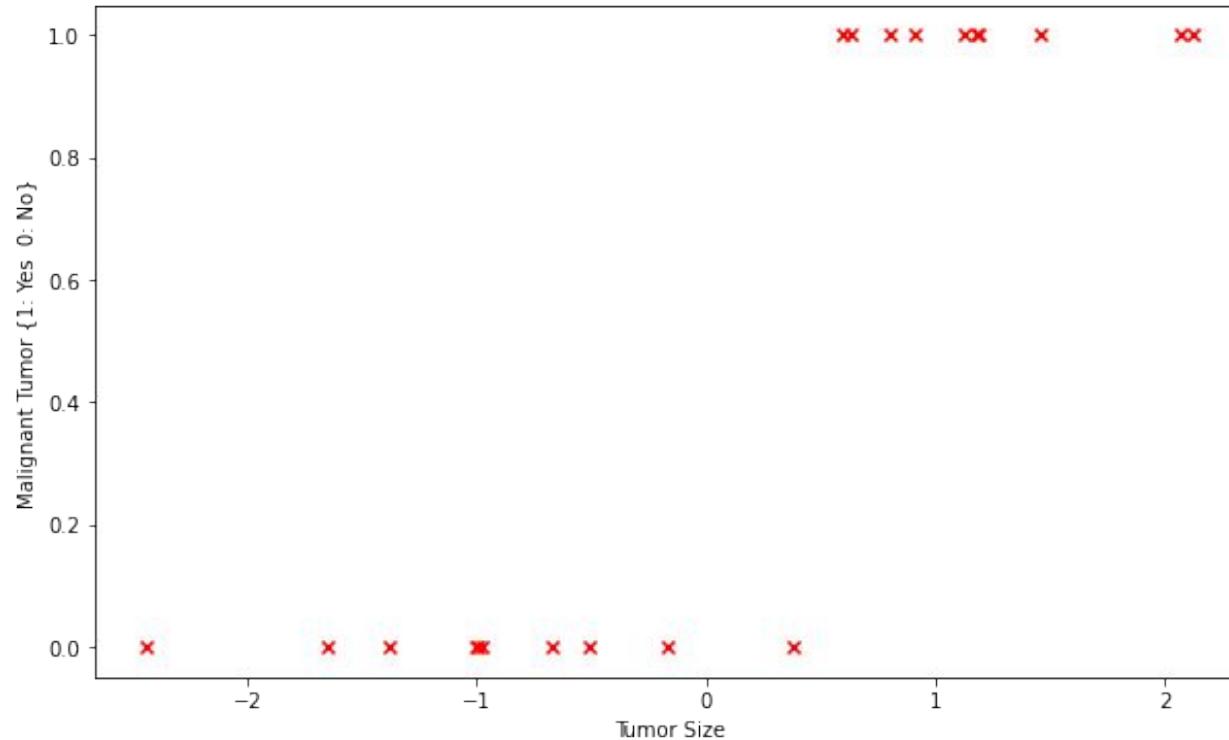
$$\alpha = 0,178$$

$$\hat{y} = ($$

Logistic Regression

- Model yang digunakan untuk data dengan variabel target bertipe kategori
- Model yang digunakan untuk klasifikasi kategori variabel target
- Representasi hubungan dari 1 atau banyak variabel prediktor terhadap variabel target

Logistic Regression



Menganalisa hubungan antara ukuran sel tumor terhadap kategori tumor (ganas/tidak).

Logistic Regression

Model persamaan binary logistic regression:

$$p = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

Nilai α dan β diperoleh dengan

Likel



03 Time Series

- Time series data
- Moving average

Time Series Data

Data yang disusun berdasar urutan waktu, baik detik, menit, jam, harian hingga tahunan. Serta tidak memasukkan data variabel lain sebagai pengaruh.

Contoh:

- **Penambahan penderita covid-19 setiap hari**
- **Pergerakan harga saham setiap menit**
- **Tingkat inflasi negara setiap tahun**

Time Series Data



Source :
indopremier.com

Data menunjukkan
adanya trend naik

Time Series Data



Source :
covid19.
go.id

Data menunjukkan adanya
trend turun

Moving Average

Terdapat data time series, lalu akan dihitung MA dengan ukuran 3.

1	2	3	7	9
---	---	---	---	---

1	2	3	7	9
---	---	---	---	---

Moving Sum Average = 2

1	2	3	7	9
---	---	---	---	---

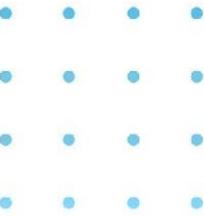
Moving Sum Average = 2, 4

1	2	3	7	9
---	---	---	---	---

Moving Sum Average = 2, 4, 6

Moving Average

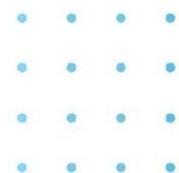
- **Simple Moving Average**
- **Cumulative Moving Average**
- **Exponential Moving Average**



04

Ensemble Technique

- Bagging
- Boosting
- Stacking



Ensemble Technique

Teknik yang dilakukan dengan membangun beberapa model untuk suatu dataset.

Untuk memperoleh hasil yang lebih baik.

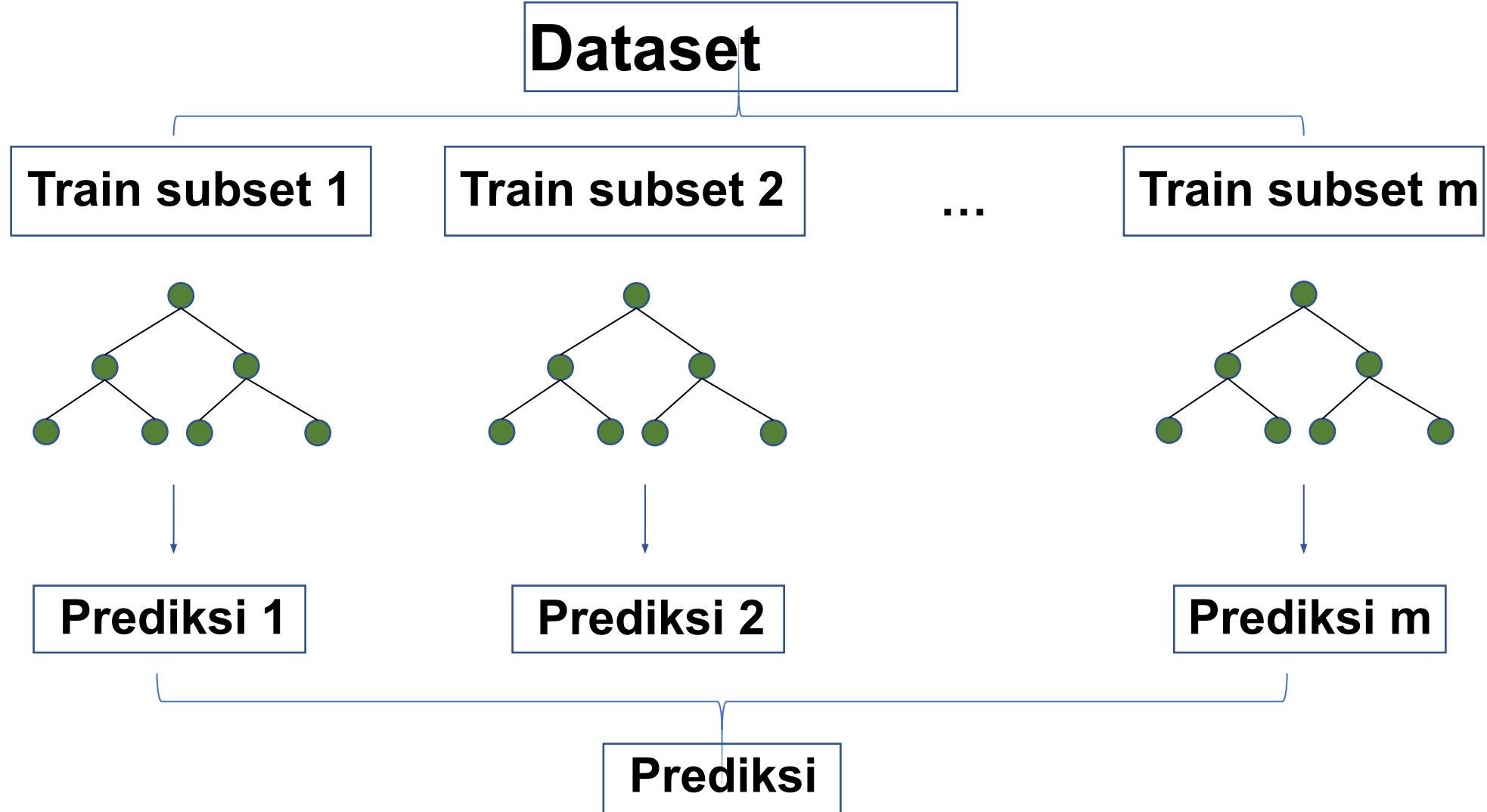
- **Bagging**
- **Boosting**
- **Stacking**

Bagging

Bagging sering disebut dengan bootstrap aggregation. Merupakan algoritme machine learning dimana prosesnya menggunakan beberapa model pada sampel dari dataset yang sama. Lalu menggabungkan modelnya dengan statistik sederhana, seperti voting.

Contoh : algoritme random forest

Bagging

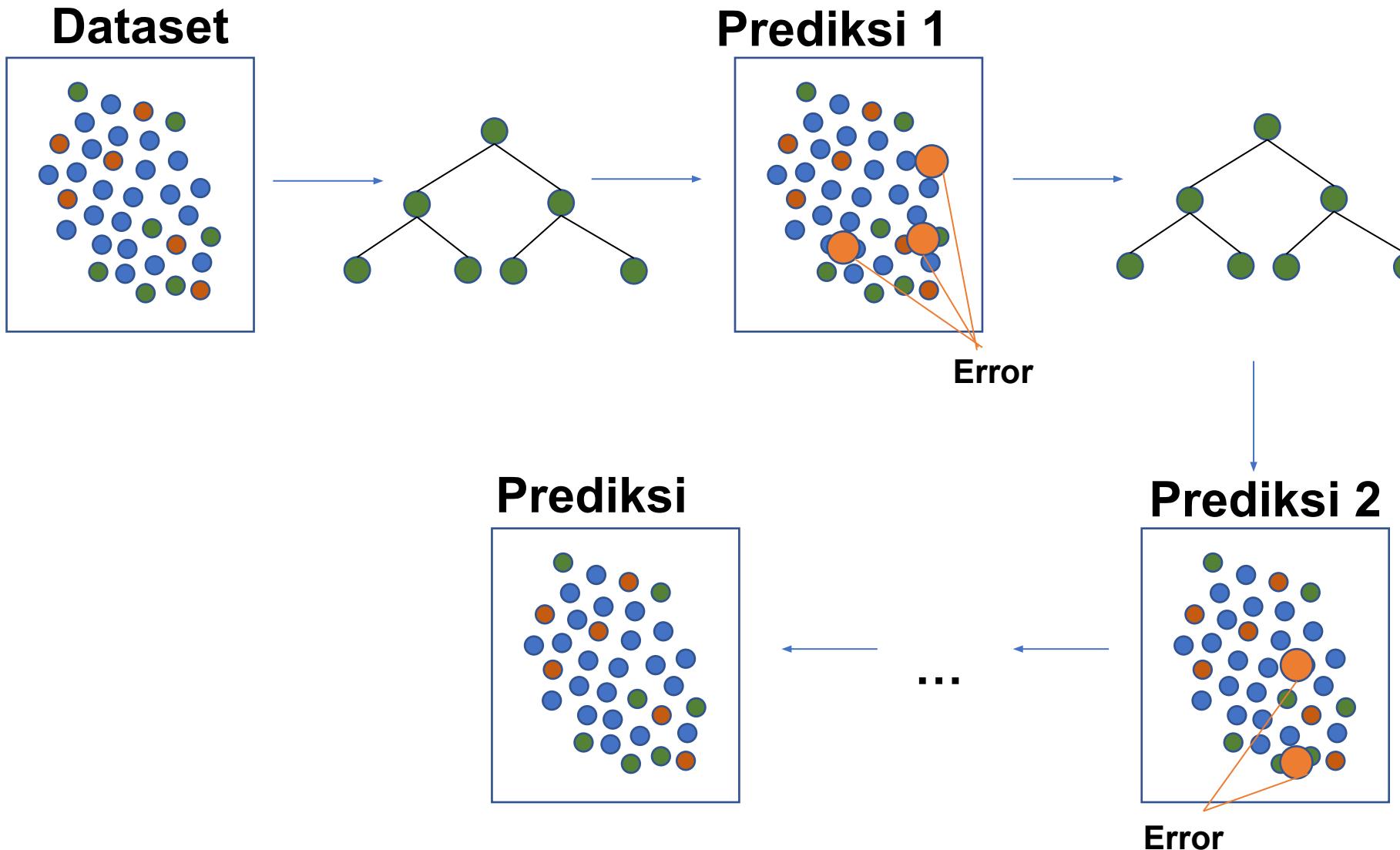


Boosting

Boosting merupakan algoritme yang mengkonversi weak learners menjadi strong learners. Weak learners merupakan model yang memiliki performa sedikit lebih baik terlepas dari penyebaran data train nya. Pada boosting, prediksi dilakukan secara berurutan, dimana setiap prediksi selanjutnya akan memperhatikan eror yang dihasilkan prediksi sebelumnya.

Contoh : Gradient Boosting Tree

Boosting

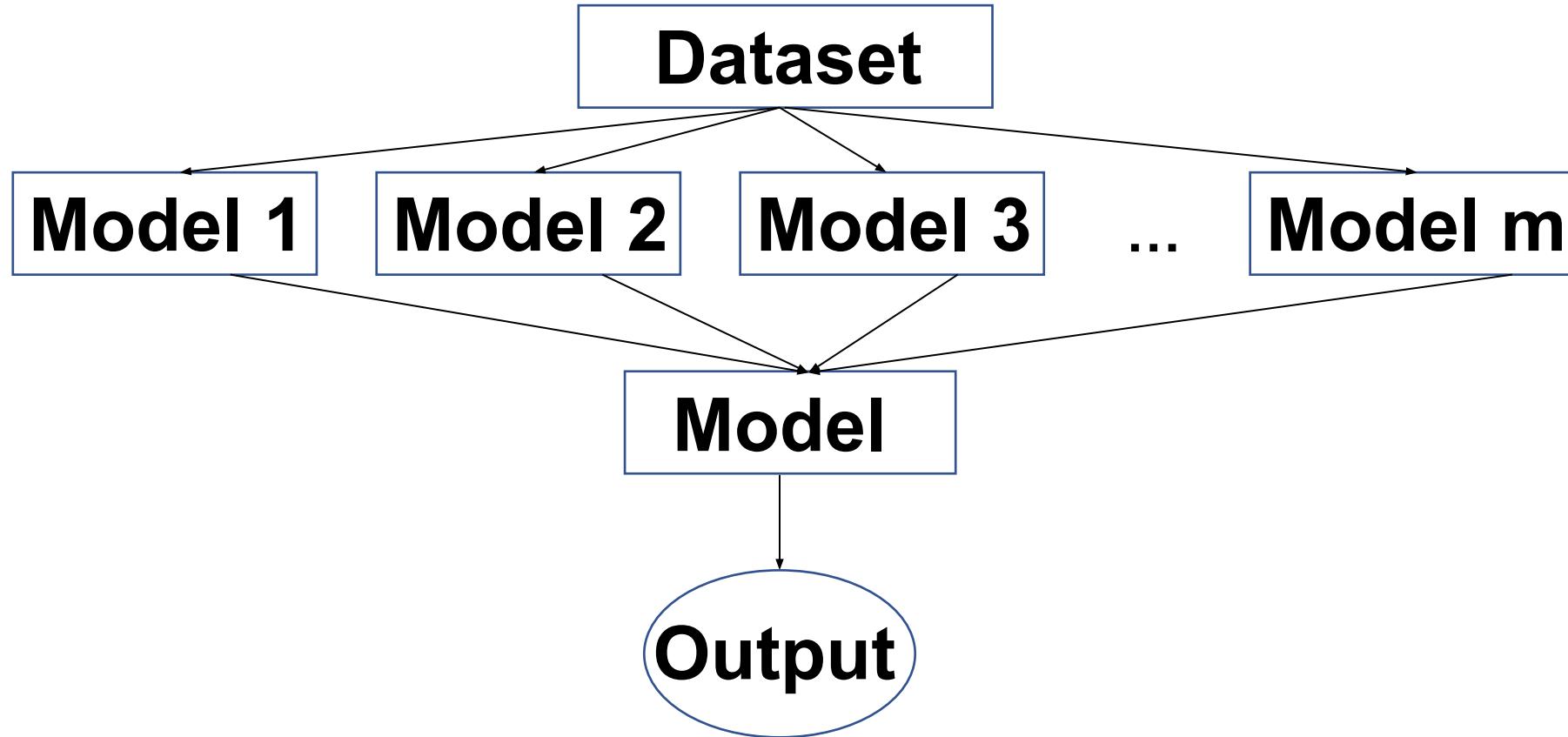


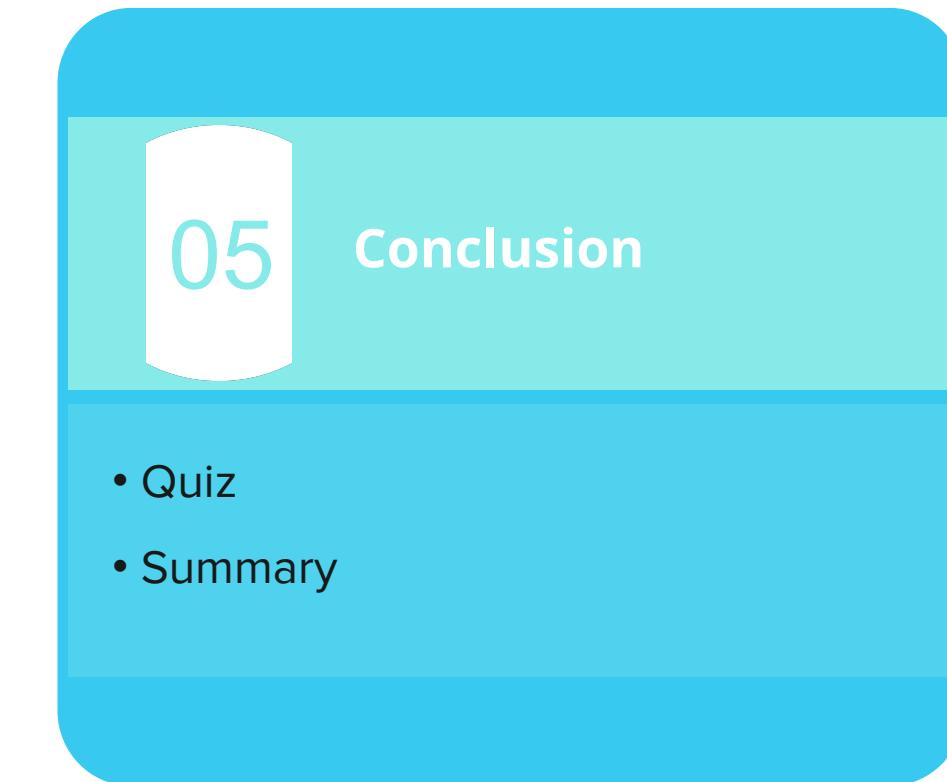
Stacking

Stacking adalah proses learning yang menggabungkan beberapa model prediksi berdasar meta-predictor yang berupa ide sederhana. Stacking merupakan teknik:

- **Training beberapa model learning sederhana**
- **Menggabungkan model dengan model meta, yang diharapkan memberikan nilai akurasi yang lebih baik**
- **Mempertimbangkan learners yang beragam dan menggabungkannya dengan melakukan training model meta untuk output prediksi berdasar prediksi weak learners.**

Stacking





05 Conclusion

- Quiz
- Summary

Quiz

1. Tipe data yang tidak bisa digunakan untuk variabel target model regresi logistic adalah ...

- A. Numerik
- B. Ordinal
- C. Nominal
- D. Kategorik



Quiz

1. Tipe data yang tidak bisa digunakan untuk variabel target model regresi logistic adalah ...

- A. Numerik
- B. Ordinal
- C. Nominal
- D. Kategorik



Jawaban : A

Quiz

2. Data berikut merupakan data time series, kecuali?

- A. Penambahan pasien covid19
- B. Fluktuasi nilai harga saham
- C. Berita / teks



Quiz

2. Data berikut merupakan data time series, kecuali?

- A. Penambahan pasien covid19
- B. Fluktuasi nilai harga saham
- C. Berita / teks

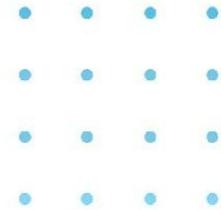


Jawaban : C

Summary

- Data science merupakan proses memahami atau memperoleh informasi dari suatu data mentah yang selanjutnya dapat digunakan untuk prediksi atau menentukan kategori.
- Data yang bisa diproses pada data science adalah data yang berbentuk tabular, baik data dengan 1 kolom maupun multiple kolom.
- Data science bermanfaat untuk memberikan suatu informasi serta melakukan prediksi.
- Model atau algoritme pada data science diantaranya regresi, moving average, dan ensemble technique.





TERIMA KASIH

Orbit Future Academy

PT Orbit Ventura Indonesia
Center of Excellence (Jakarta Selatan)
Gedung Veteran RI, Lt.15
Unit Z15-002, Plaza Semanggi
Jl. Jenderal Sudirman Kav.50, Jakarta
12930, Indonesia

- Jakarta Selatan/Pusat
- Jakarta Barat/BSD
- Kota Bandung
- Kab. Bandung
- Jawa Barat

Hubungi Kami

Director of Sales & Partnership
ira@orbitventura.com
+62 858-9187-7388

Social Media

-  [Orbit Future Academy](#)
-  [@OrbitFutureAcademyIn1](#)
-  [OrbitFutureAcademy](#)
-  [Orbit Future Academy](#)



WELCOME

**Data Science Courses
AI Mastery Program**

Wed, March 23rd 2022

08.00 – 12.00

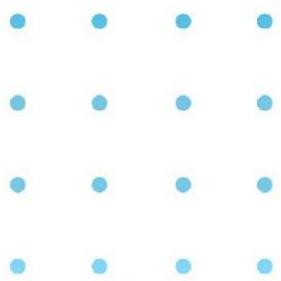
**Group V :
Kaizen-AlphaGo-Devalopa-Khasanah Ilmi-Alfatih**

TODAY'S SCHEDULE

Time	Activities	Duration
08.00 – 08.30	Opening, etc.	30'
08.30 – 09.45	Session 1	75'
09.45 – 10.15	Break	30'
10.15 – 11.30	Session 2	75'
11.30 – 12.00	FGD (conditional)	30'

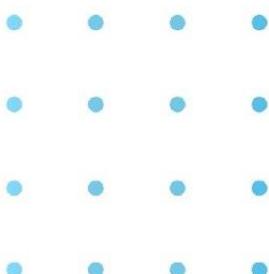
Disusun oleh:
Data Science
Team

Kode Materi:
[XX-XX-XX]



Data Science

Visualization with Tableau





Objective

Memahami apa itu visualisasi data dan betapa pentingnya proses visualisasi data untuk kebutuhan analisis dan pemodelan data menggunakan salah satu *software* visualisasi data Tableau yang populer.

Outline

- Apa itu Visualisasi Data
- Macam-macam Grafik
- Apa itu Tableau
- Demo Tableau
- Latihan Tableau

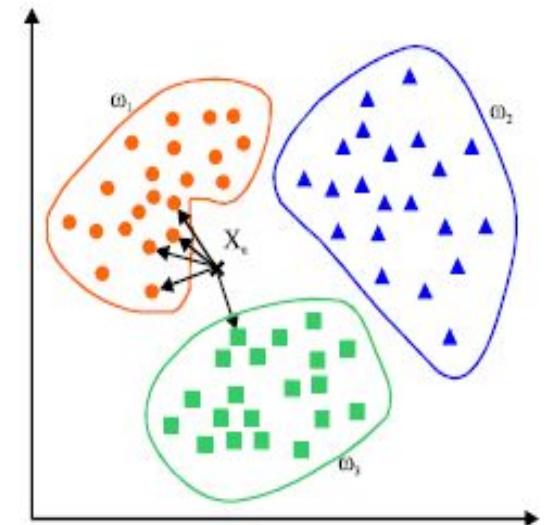
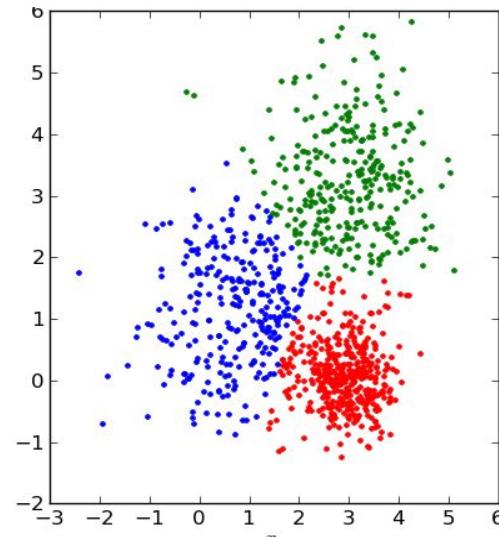
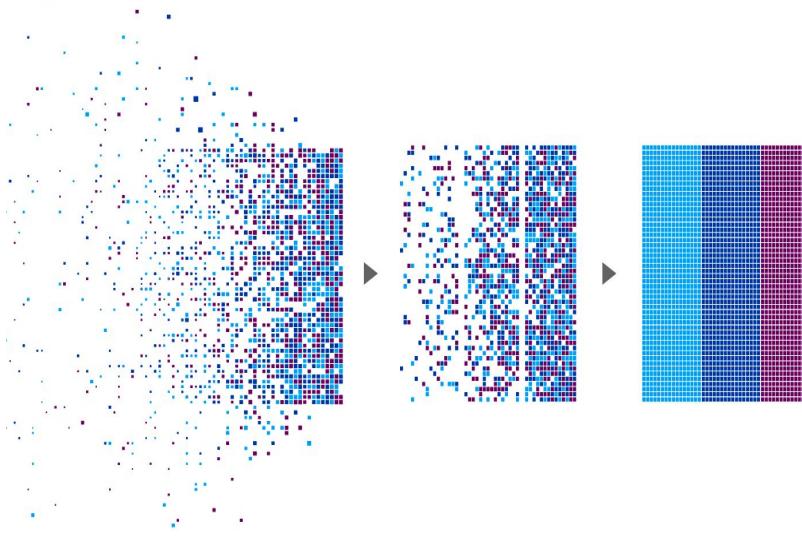


What is Data Visualization ?

Data Visualization



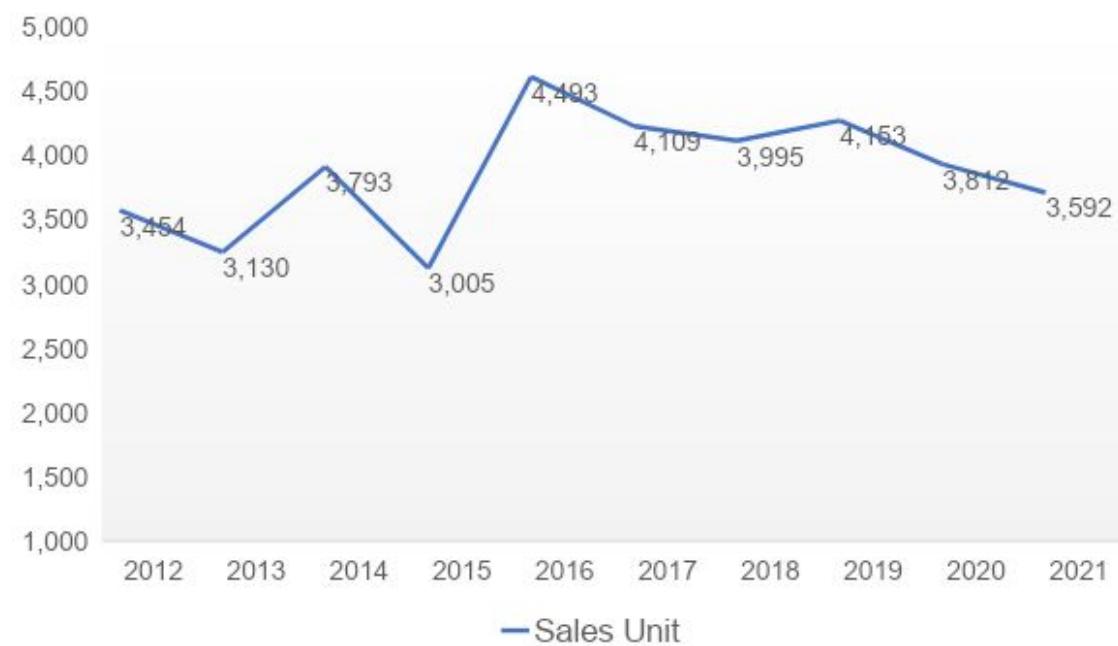
Visualisasi data adalah representasi grafis dari informasi dan data. Dengan menggunakan elemen visual seperti bagan, grafik, dan peta, alat visualisasi data menyediakan cara yang dapat digunakan untuk melihat dan memahami tren, outlier, dan pola dalam data.

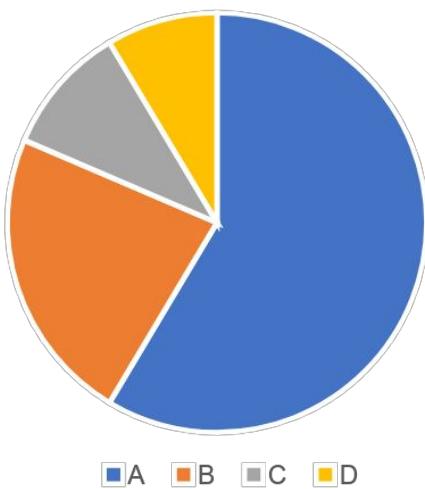
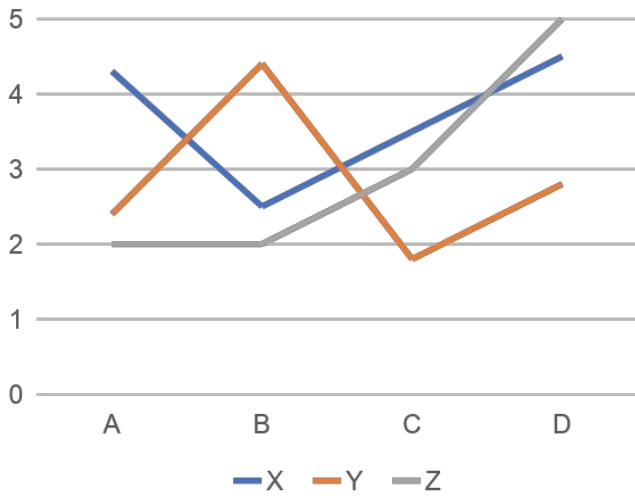
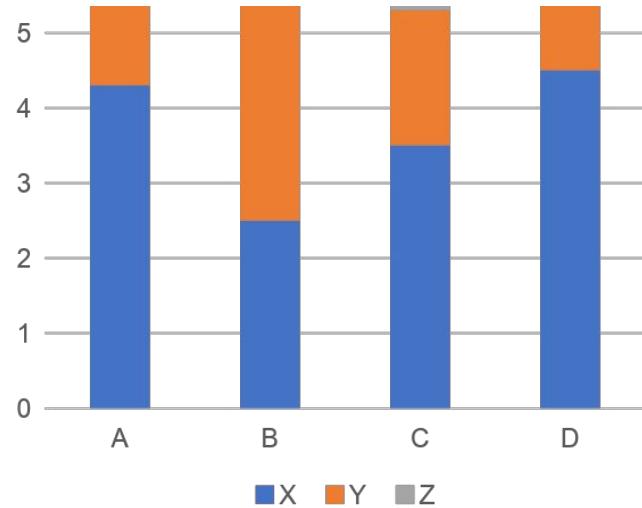


Why Data Visualization is Important ?

Why Data Visualization is Important

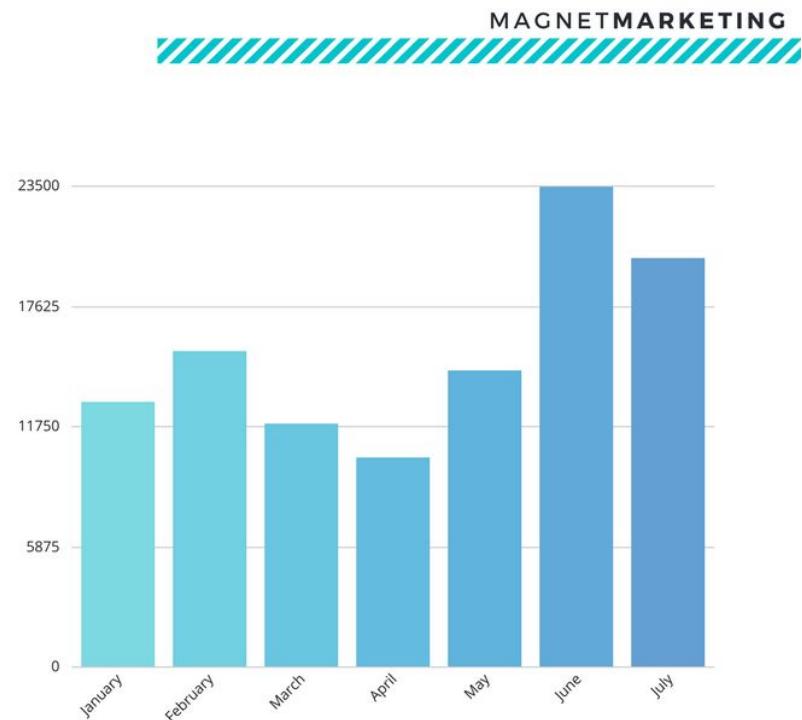
Tahun	Sales (unit)	Profit
2012	3454	Rp6.044.500
2013	3130	Rp5.477.500
2014	3793	Rp6.637.750
2015	3005	Rp5.258.750
2016	4493	Rp7.862.750
2017	4109	Rp7.190.750
2018	3995	Rp6.991.250
2019	4153	Rp7.267.750
2020	3812	Rp6.671.000
2021	3592	Rp6.286.000





What Chart do you know ?

Bar Chart



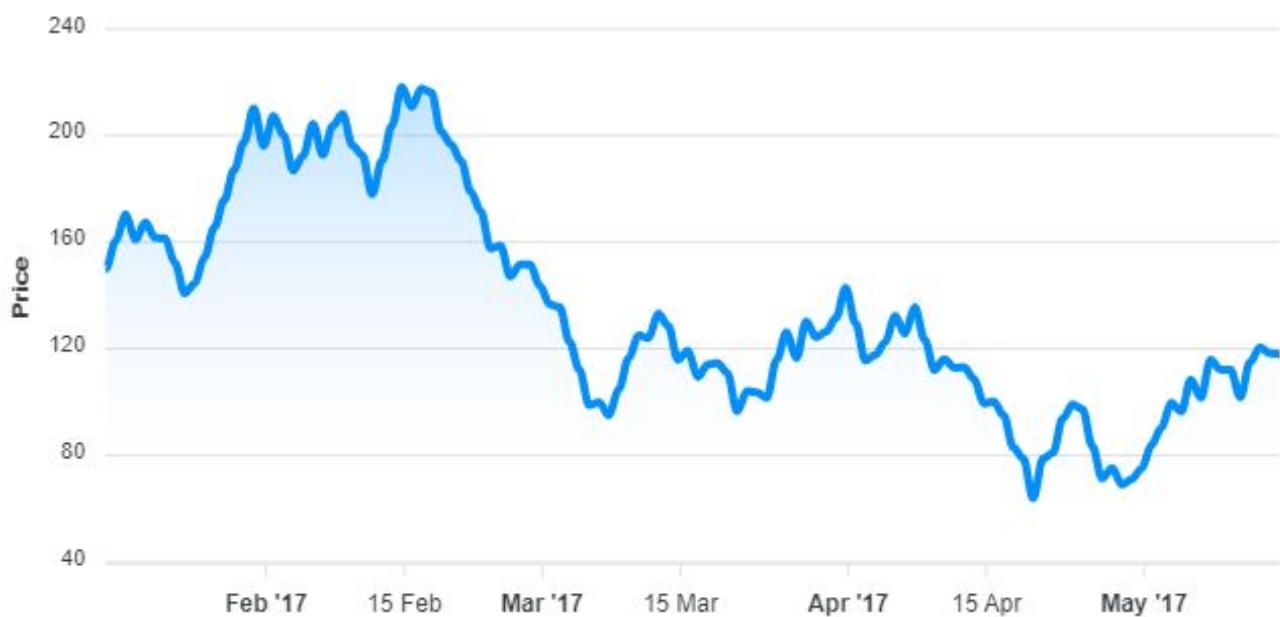
Lorem ipsum dolor sit amet, consectetur
adipiscing elit. Sed varius varius nisi
placerat fringilla.

Bar Chart

Grafik yang menyajikan data kategorikal dalam bentuk batang dengan panjang batang yang mewakili nilainya.

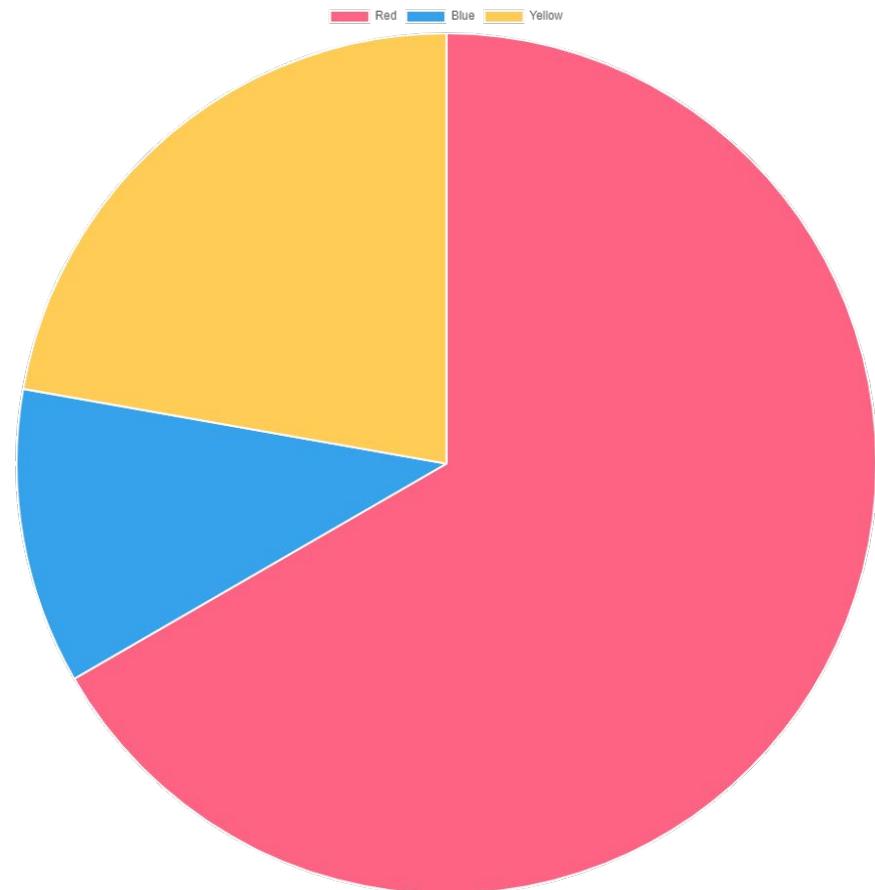
Line Chart

Stock Price Movement



Line Chart

Visualisasi data yang digunakan untuk memperlihatkan perubahan kondisi dari waktu ke waktu.

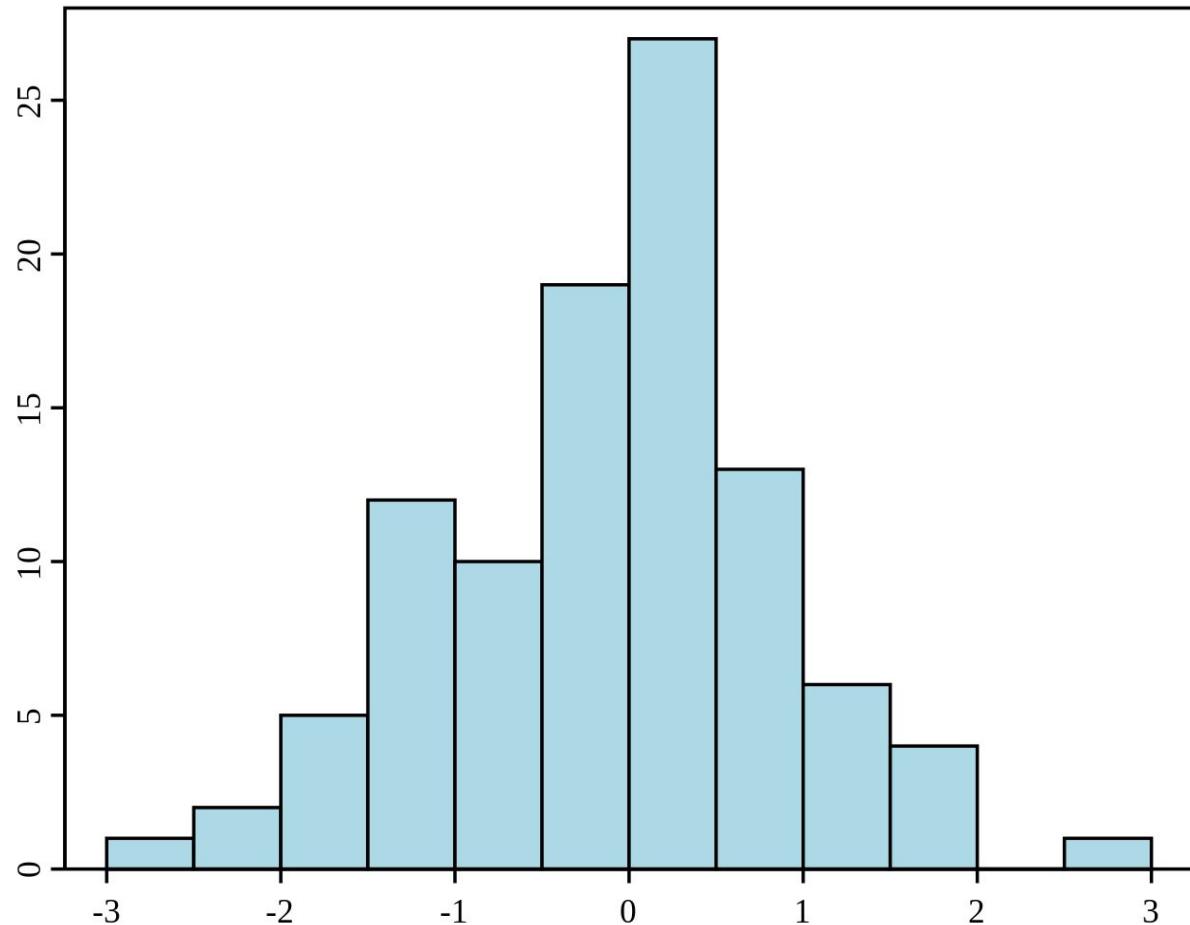


Pie Chart

Visualisasi data untuk menunjukkan sebuah komposisi tertentu dari sebuah data.

Bagian yang ada di dalam pie chart menunjukkan proporsi dari bagian data terhadap keseluruhan data

Histogram Chart

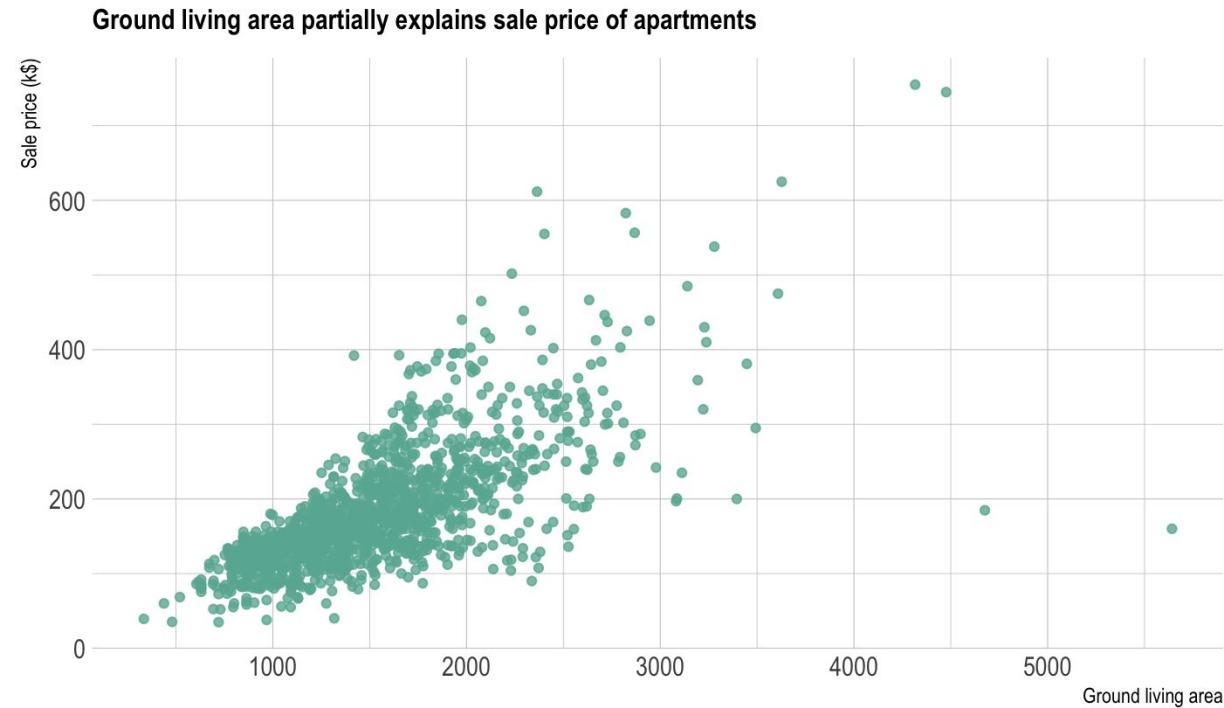


Histogram

Grafik yang biasa digunakan untuk menampilkan distribusi dari sebuah data.

Biasa digunakan untuk melihat estimasi distribusi probabilitas dari variable yang kontinu

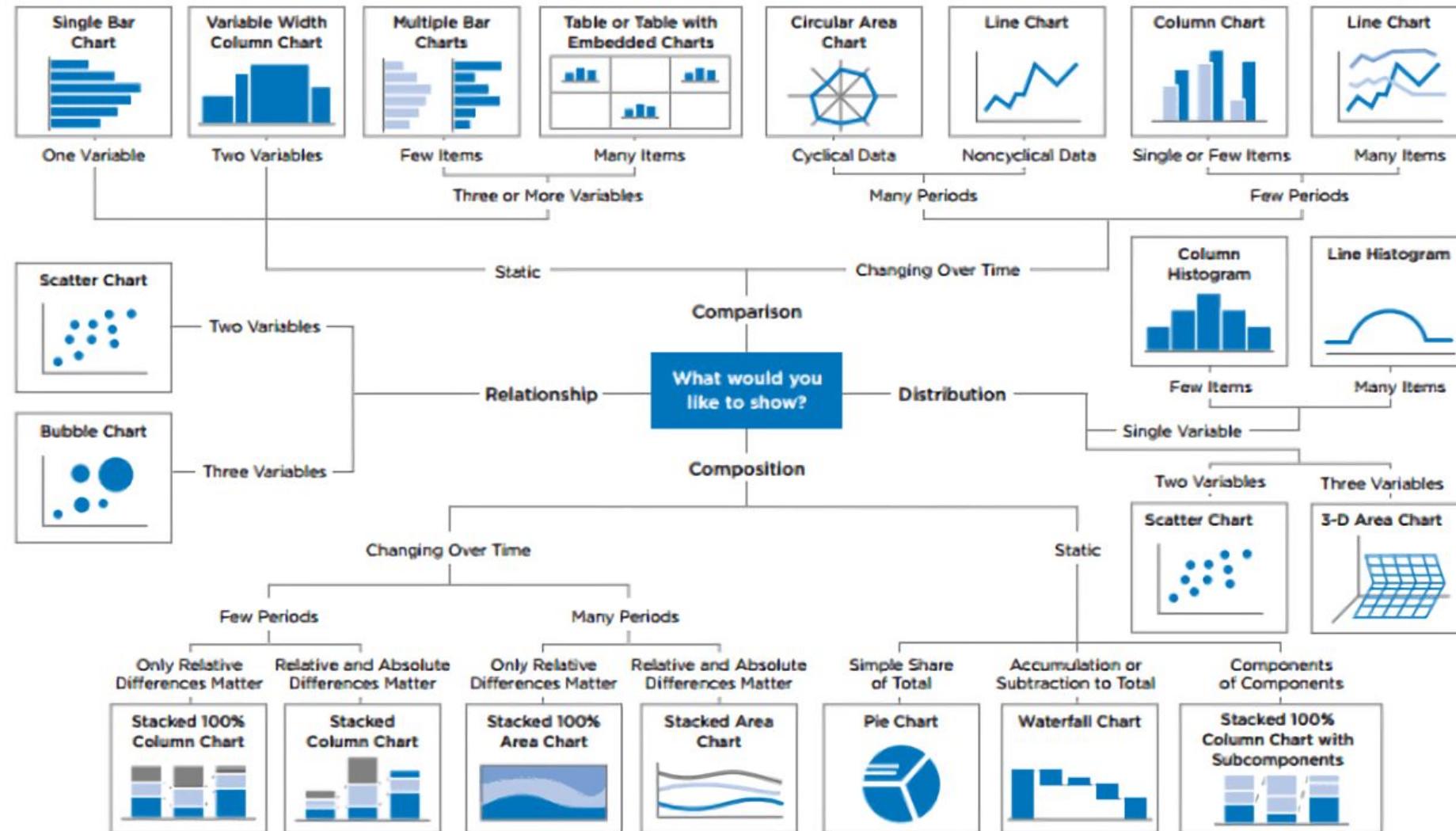
Scatter Chart



Scatter Plot

Grafik ini biasa digunakan untuk melihat suatu pola hubungan antara 2 variable numerik.

SELECTING THE APPROPRIATE CHART FOR STRATEGY PRESENTATIONS



What is Tableau ?

Data A

Data B

Data C

Data D

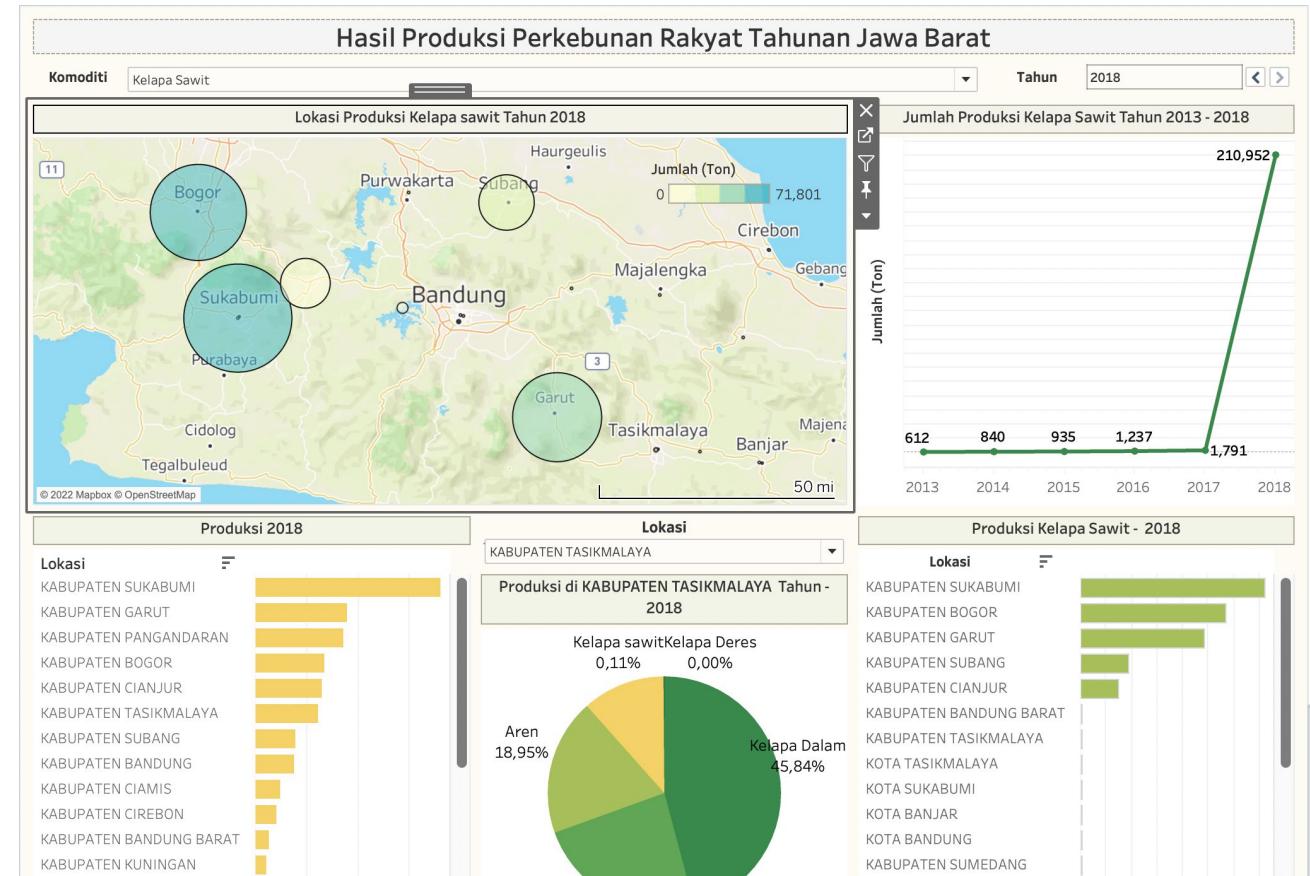
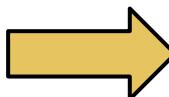




Tableau adalah platform untuk merepresentasikan data dalam bentuk analitik visual melalui *chart* dan *dashboard* untuk memecahkan masalah.

Tableau Dekstop

Tableau Public

Tableau Reader

Jenis Tableau

Sumber Data

Tableau Desktop	Tableau Public	Tableau Reader
Semua sumber data bisa digunakan	Sumber data hanya di Excel dan Text Files	Tidak membutuhkan untuk terhubung dengan sumber data

Keamanan Data

Tableau Desktop

Tidak ada yang bisa melihat
laporan milik kita

Tableau Public

Seluruh laporan akan otomatis
menjadi akses **public** ketika di
publish / di simpan

Tableau Reader

Tidak ada yang bisa melihat
laporan milik kita

Jumlah Data

Tableau Desktop	Tableau Public	Tableau Reader
Tidak terbatas	1 juta baris data dapat di proses dan dibagikan	Data tidak terbatas , namun hanya dapat dilihat pada tampilan static (seperti excel)

Biaya Data

Tableau Desktop	Tableau Public	Tableau Reader
Personal (\$999) & Professional (\$1999)	Gratis !	Gratis !

Pengguna Data

Tableau Desktop

Data Scientist, Business Intelligence

Tableau Public

Junior Data Analyst & Junior Business Intelligence

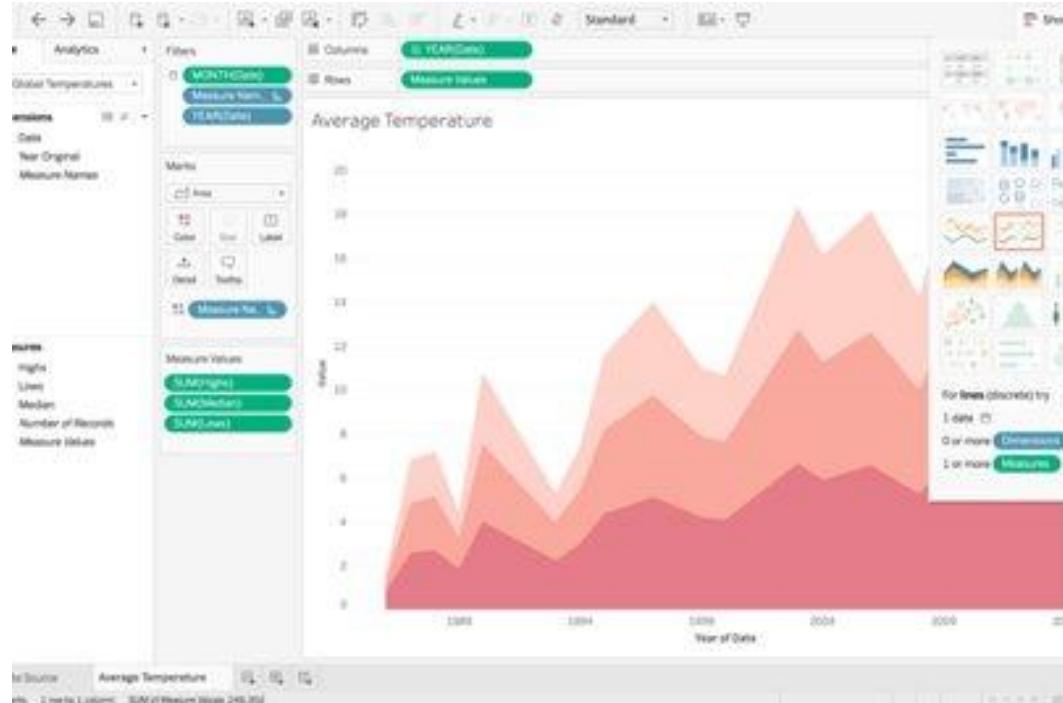
Tableau Reader

Professional yang hanya butuh membaca sebuah data



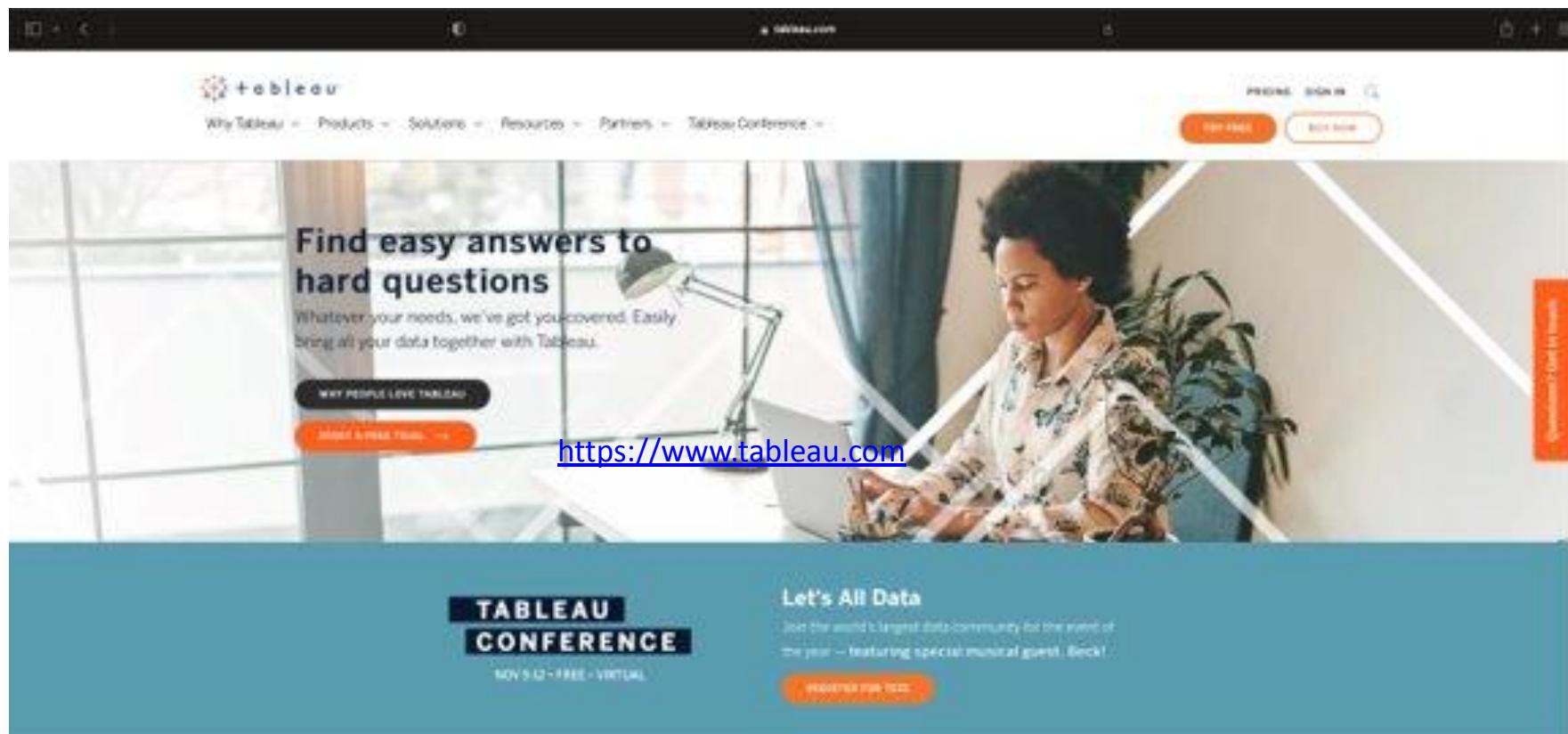
Tableau memiliki banyak fitur dan fungsi yang bisa kamu manfaatkan. Terdapat **4 Fungsi Utama** Tableau adalah sebagai berikut.

- Menerjemahkan data menjadi bentuk visual
- Mengelola *metadata*
- Mengimpor berbagai ukuran dan *range* data
- Membuat visualisasi data tanpa *coding*



Kelebihan Tableau

- Pilihan visual yang interaktif
- *User friendly*
- Dapat mengolah banyak sumber data
- *Dashboard mobile friendly*
- Terintegrasi dengan bahasa skrip

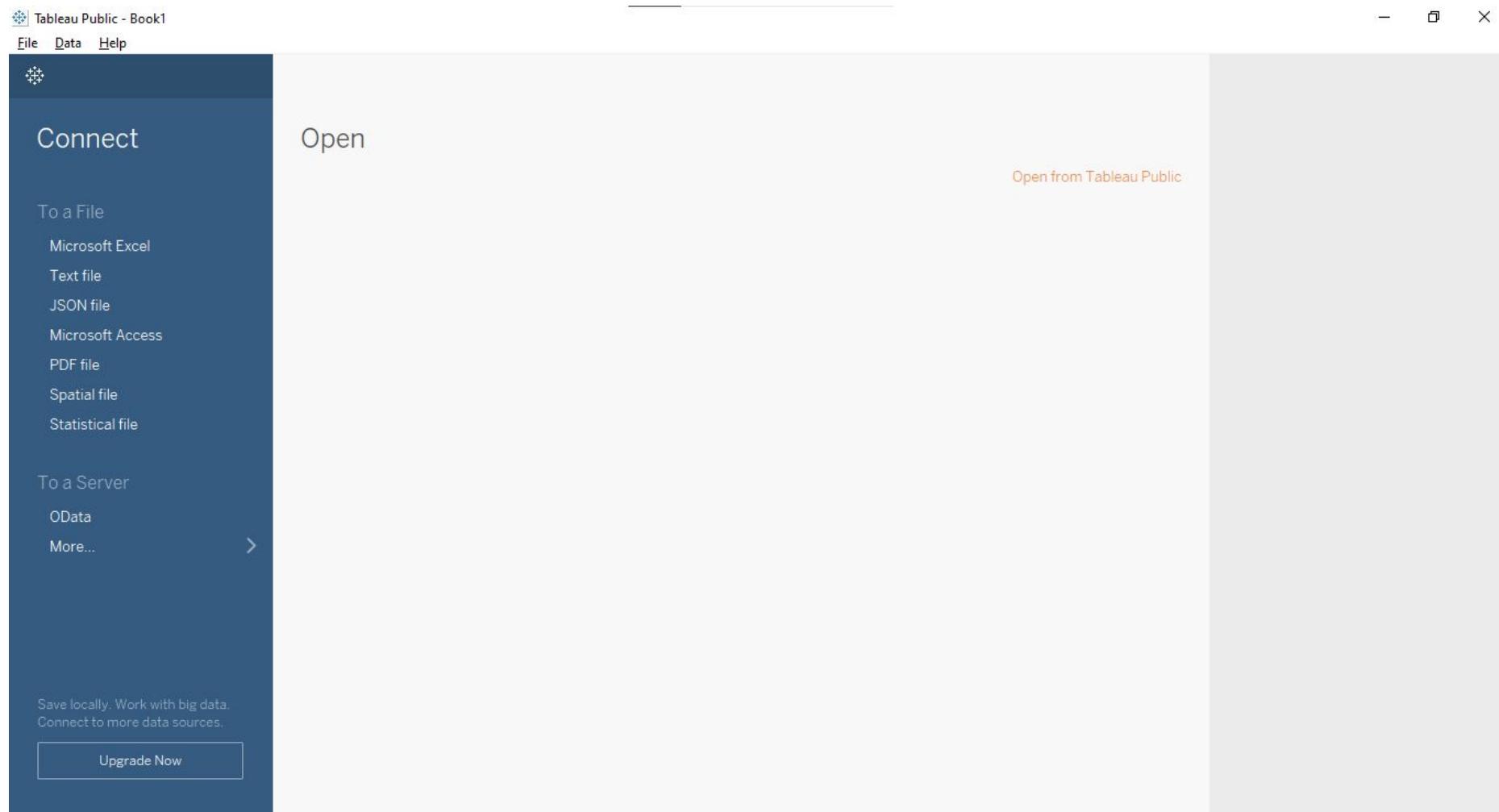


Meet the world's leading analytics platform.

Tableau helps people see and understand data. Our visual analytics platform is transforming the way people use data to

<https://www.tableau.com>

Tableau Interface



Tampilan Start Page Tableau

Tableau Interface

The screenshot shows the Tableau Public interface with the following details:

- File Bar:** File, Data, Window, Help.
- Connections:** Sample - Superstore Sales (Microsoft Excel).
- Sheets:** Orders, Returns, Users.
- Data Interpreter:** A checkbox labeled "Use Data Interpreter" is checked, with a note: "Data Interpreter might be able to clean your Microsoft Excel workbook."
- Table Preview:** The Orders sheet is selected, showing 21 fields and 8390 rows. The columns include Row ID, Order ID, Order Date, Order Priority, Order Quantity, Sales, Discount, Ship Mode, Profit, and Unit Price.
- Bottom Navigation:** Data Source, Sheet 1, and various icons for saving and sharing.
- Bottom Footer:** Muhammad ... and navigation icons.

Tampilan Data Source Tableau

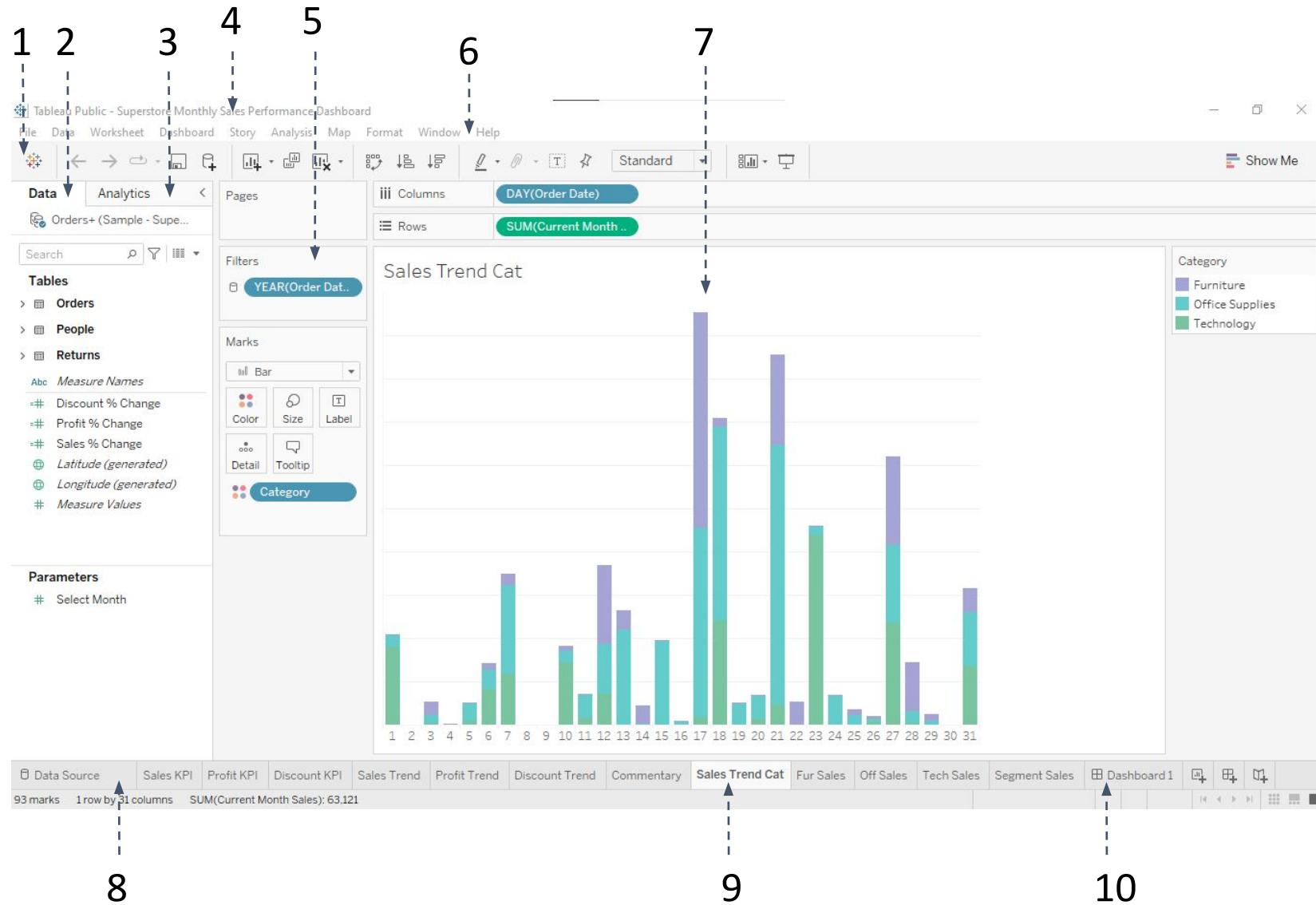
Tableau Interface

The screenshot shows the Tableau Public interface with the following components:

- Top Bar:** File, Data, Worksheet, Dashboard, Story, Analysis, Map, Format, Window, Help.
- Data Pane:** Shows a list of tables and measures:
 - Tables: Customer Name, Customer Segment, Order Date, Order ID, Order Priority, Product Category, Product Container, Product Name, Product Sub-Category, Province, Region, Row ID, Ship Date, Ship Mode.
 - Measures: Discount, Order Quantity, Product Base Margin, Profit, Sales, Shipping Cost, Unit Price, Latitude (generated).
- Marks Card:** Displays options for Color, Size, Text, Detail, and Tooltip.
- Worksheet Area:** Labeled "Sheet 1". It has three empty drop zones:
 - Top-left: "Drop field here".
 - Top-right: "Drop field here".
 - Bottom-left: "Drop field here".
- Show Me:** A panel on the right showing various chart and map types.
- Bottom Navigation:** Data Source, Sheet 1, and other worksheet icons.
- Bottom Status Bar:** Shows the user's name (Muhammad ...) and navigation controls.

Tampilan Worksheet Tableau

Tableau Interface



1. Go to Start Page
2. Data Pane
3. Analytics Pane
4. Workbook Name
5. View Cards
6. Toolbar
7. Worksheet View
8. Go to Data Source
9. Worksheet Tabs
10. Dashboard Tabs

Tableau Interface

Term	Deskripsi
Go to Start Page	Tombol untuk kembali ke halaman Desktop awal Tableau
Data Pane	Berisi dimensions dan measures dari data yang diupload ke dalam halaman utama
Analytics Pane	Berisi pilihan analisis yang dapat digunakan untuk membuat garis regresi, prediksi, trend lines, crosstab dan boxplot, dan lainnya
Workbook Name	Nama yang diberikan kepada workbook
View Cards	Digunakan untuk melakukan modifikasi tampilan worksheet
Toolbar	Simbol yang digunakan untuk mendapatkan akses cepat ke fitur favorit
Worksheet View	Sheet yang digunakan untuk membuat visualisasi
Go to Data Source	Tanda yang digunakan untuk kembali ke sumber data
Worksheet Tabs	Tombol untuk melihat sheet visualisasi yang telah atau akan dibuat
Dashboard Tabs	Tombol untuk melihat sheet dashboard yang telah atau akan dibuat

Data Pane

Data Analytics <
Orders+ (Sample - Supe...
Search ⌂ | ⌂ ▾
Tables
Sales
= # Shipping Time (Days)
Orders (Count)
▼ # People
 Abc Person
 Abc Region (People)
 # People (Count)
▼ # Returns
 Abc Order ID (Returns)
 Abc Returned
 # Returns (Count)
 Abc Measure Names
 :# Discount % Change
 :# Profit % Change
 :# Sales % Change
 🌐 Latitude (generated)
 🌐 Longitude (generated)
 # Measure Values
▼ Parameters
Select Month

Komponen	Deskripsi
Dimensions	Panel yang berisi data kategorik seperti nama, tanggal, atau data geografis
Measures	Panel yang berisi data numerik yang dapat diukur
Parameters	Variabel yang dapat menggantikan nilai konstan yang ditentukan oleh peneliti
Sets	Subset dari data yang didefinisikan

Data Modifiers

Data Analytics <

Orders+ (Sample - Supe...
Search

Tables

- # Sales
- =# Shipping Time (Days)
- # Orders (Count)

People

- Abc Person
- Abc Region (People)
- # People (Count)

Returns

- Abc Order ID (Returns)
- Abc Returned
- # Returns (Count)

Abc Measure Names

- =# Discount % Change
- =# Profit % Change
- =# Sales % Change
- (@) Latitude (generated)
- (@) Longitude (generated)
- # Measure Values

Parameters

- # Select Month

Ikon	Deskripsi
Abc	Nilai Teks (string)
	Nilai Tanggal
	Nilai Tanggal & Waktu
#	Nilai Numerik
T F	Nilai Boolean (hanya relasional)
	Nilai Geografis (digunakan dengan peta)
Abc	Ikon berwarna biru menandakan <i>field</i> bersifat diskrit
#	Ikon berwarna hijau menandakan <i>field</i> bersifat kontinu
=Abc	Ikon yang didahului dengan tanda sama dengan (=) menunjukkan bahwa <i>field</i> tersebut merupakan hasil dari duplikasi
#!	Ikon yang diakhiri dengan tanda seru (!) menunjukkan bahwa <i>field</i> tersebut tidak valid

Analytics Pane

The screenshot shows the 'Analytics' tab selected in the top navigation bar. The left sidebar contains three main sections: 'Summarize', 'Model', and 'Custom'. Each section lists several analytical tools with corresponding icons.

- Summarize:** Constant Line, Average Line, Median with Quartiles, Box Plot, Totals.
- Model:** Average with 95% CI, Median with 95% CI, Trend Line, Forecast, Cluster.
- Custom:** Reference Line, Reference Band, Distribution Band, Box Plot.

Komponen	Deskripsi
Summarize	<i>Tools</i> yang digunakan untuk menambahkan komponen analytics yang diinginkan pada grafik seperti garis konstan, rata-rata, median dengan kuartil, dan total
Model	<i>Tools</i> yang digunakan untuk menambahkan informasi pemodelan menurut keinginan peneliti seperti garis tren, peramalan, dan distribusi rata-rata
Custom	<i>Tools</i> yang digunakan untuk menambahkan <i>custom lines, bands, and boxplot</i>

Component View

Pages

Filters

- Top Region
- Top Ship Mode
- YEAR(Order Dat..)

Marks

- Automatic
- Color
- Size
- Text
- Detail
- Tooltip

- Top Region
- SUM(Top Regi..)
- Top Ship Mode
- AVG(Shipping ..)

Columns

Rows

Komponen	Deskripsi
Columns and Rows	Tarik <i>dimensions</i> dan <i>measures</i> ke dalam panel ini untuk mendapatkan visualisasi data yang diinginkan
Pages	Menunjukkan perubahan data dari waktu ke waktu pada <i>dimensions</i>
Filters	Tarik <i>fields</i> ke dalam panel ini untuk membatasi jumlah data yang ditampilkan. <i>Filter</i> yang tampil pada <i>dashboard</i> memungkinkan audiens untuk mengatur apa yang ingin dilihat dalam visualisasi data
Marks	Panel untuk mengatur jenis grafik yang ingin digunakan dalam visualisasi data
Marks Card	Panel untuk mengatur tampilan pada <i>mark</i> , seperti warna, ukuran, teks, detil, dan tooltip

Fields dalam Panel Marks

The screenshot shows the Tableau interface with the Marks shelf open. The shelf includes options for Automatic, Color, Size, Text, Detail, Tooltip, and several specific fields like Top Region, SUM(Top Regi..), Top Ship Mode, and AVG(Shipping ..). Above the Marks shelf, there are sections for Pages and Filters, with the YEAR(Order Dat..) filter selected.

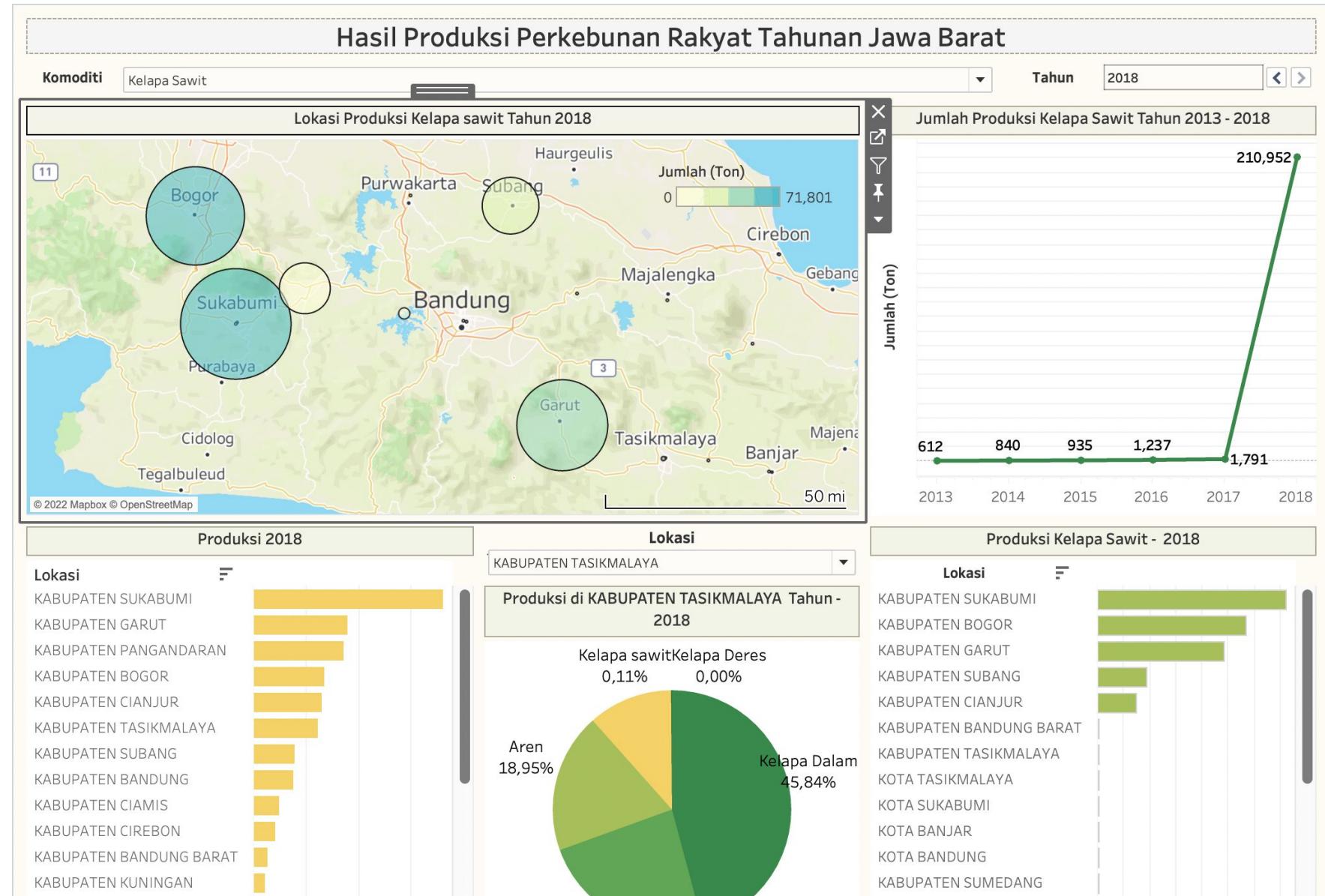
Icon	Deskripsi
Segment	Field berwarna biru menunjukkan bahwa data bersifat diskrit
SUM(Sales)	Field berwarna hijau menunjukkan bahwa data bersifat kontinu
Segment	Simbol sort pada bagian kanan field menunjukkan bahwa data telah diurutkan
SUM(Sales)	Simbol delta pada bagian kanan field menunjukkan bahwa data telah dikalkulasi
+ State	Simbol plus (+) dan minus (-) menunjukkan bahwa field memiliki hierarki dan dapat diperluas atau dipersempit
- Country	

Output File Tableau

Tipe File	Ukuran	Jenis Kebutuhan	Keterangan
Tableau Workbook (twb)	Kecil	Default dari Tableau untuk menyimpan pekerjaan	Informasi hasil visualisasi data tanpa <i>data source</i>
Tableau Package Workbook (twbx)	Potensial Besar	Digunakan untuk <i>sharing</i> dengan pembaca Tableau yang lain tanpa mengakses <i>data source</i>	Mengekstrak data dan informasi dari <i>workbook</i> untuk membuat visualisasi
Tableau Datasource (tds)	Kecil	Digunakan jika sering mengakses <i>data source</i>	Alamat <i>server</i> , <i>password</i> , dan <i>metadata</i> yang berhubungan dengan <i>data source</i>
Tableau Data Extract (tde)	Potensial Besar	Digunakan untuk meningkatkan performa dari hasil visualisasi dengan menambahkan beberapa fungsi	<i>Data source</i> yang terfilter dan data gabungan selama diekstrak
Tableau Bookmark (tbm)	Standar Kecil	Digunakan untuk <i>sharing worksheet</i> satu dengan yang lainnya	Informasi untuk visualisasi dan <i>data source</i> jika datanya dipaketkan



Demo Tableau



Koleksi
Data



Time to Try !!

- Effective Computation in Physics oleh Anthony Scopatz & Kathryn D. Huff



```
from heart import gratitude
```

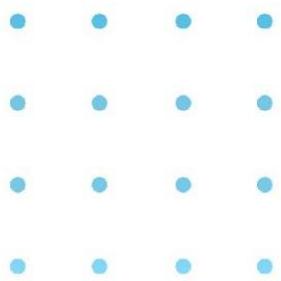
```
x = "bye"
```

```
y = "until we meet again"
```

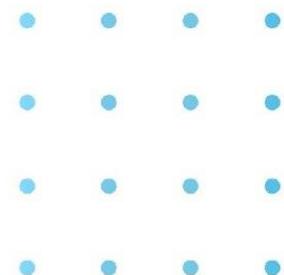
```
Thank_You = gratitude(feeling=[happy, super, hopeful]).fit(x, y)
```

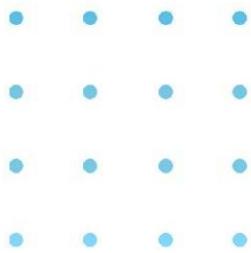
```
print(Thank_You)
```

AI Mastery Course



Dimensionality Reduction

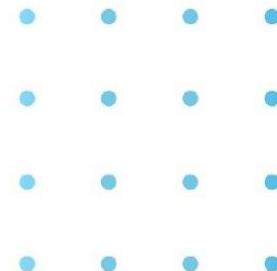




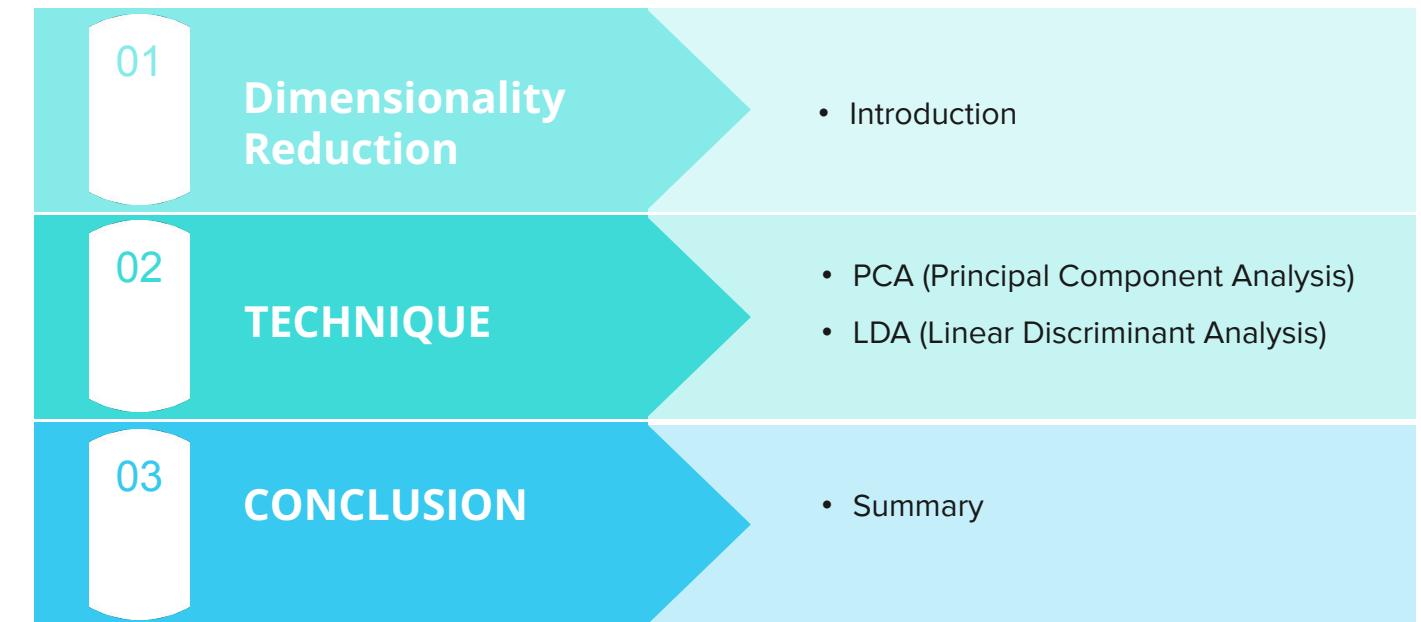
Learning Objectives

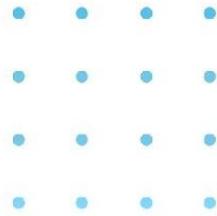
Di akhir modul ini kita diharapkan dapat memahami:

- Dimensionality Reduction di AI
- Tujuan dimensionality reduction
- Teknik dimensionality reduction



Agenda

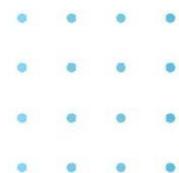




01

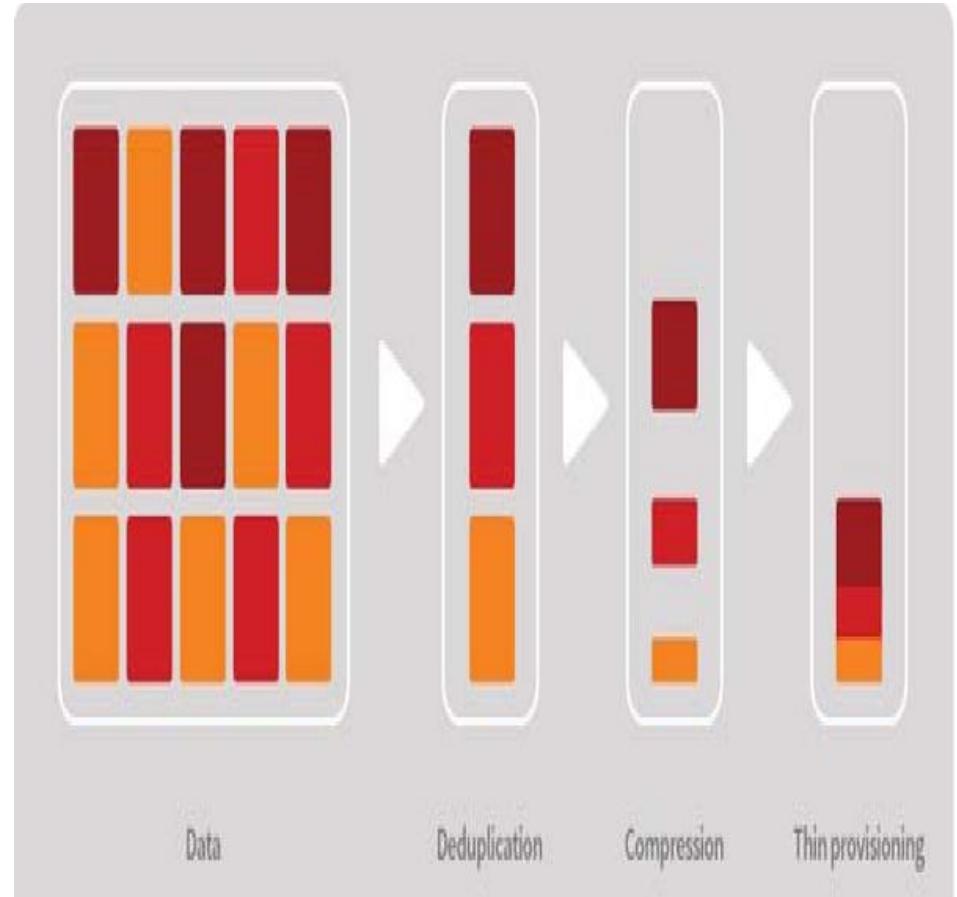
Dimensionality Reduction

- Introduction



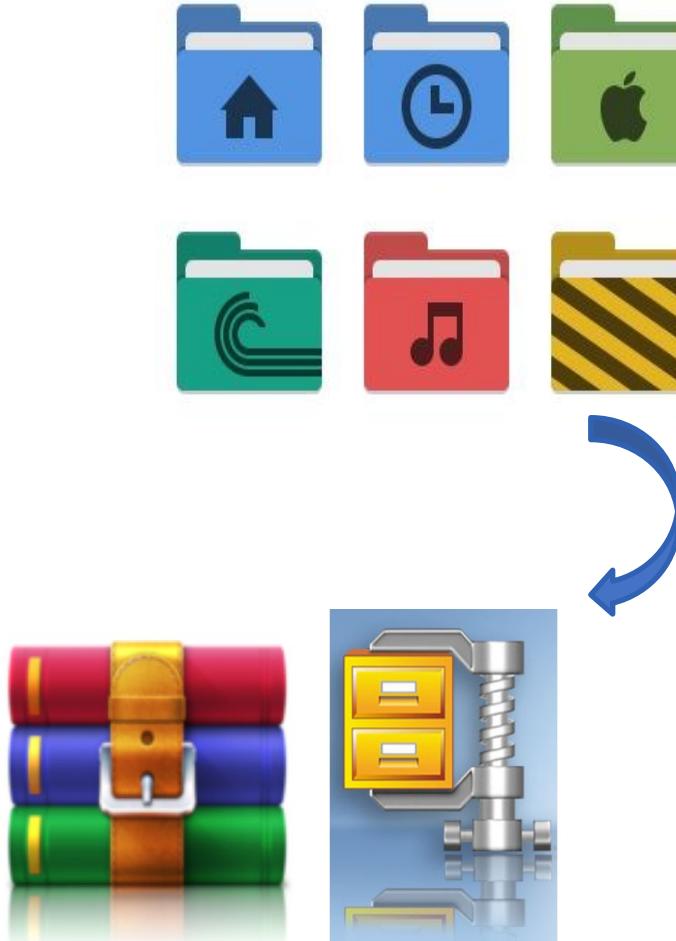
Introduction

- *Dimensionality reduction* atau reduksi dimensi adalah teknik untuk mengurangi dimensi dataset dalam hal ini fitur data.
- Biasanya dataset yang ingin diproses memiliki puluhan bahkan mungkin ratusan fitur atau kolom.
- Dengan reduksi dimensi, kita dapat mengurangi jumlah fitur atau kolom tanpa menghilangkan informasi dari dataset tersebut.



Introduction

- Reduksi dimensi pada prinsipnya sama dengan ketika kita mengompres *file* yang berukuran besar menjadi *zip file*
- Kompresi *file* tidak akan menghilangkan atau mengurangi informasi yang ada di dalam *file* tersebut.
- Membuatnya lebih sederhana sehingga mengurangi ukuran *file* yang dapat mempercepat proses transfer *file*.

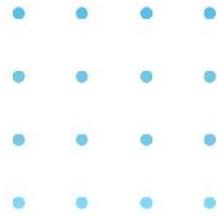


Alasan perlunya melakukan reduksi dimensi :

- Banyaknya variabel input dapat menurunkan performa *machine learning*.
- Dataset yang digunakan pada umumnya direpresentasikan oleh baris dan kolom, sehingga dalam hal ini variabel inputnya adalah kolom atau fiturnya.
- Jumlah fitur yang sangat banyak sering kali dapat mengakibatkan *data point* merepresentasikan sampel yang tidak representatif. Ini dapat sangat mempengaruhi performa algoritma *machine learning*.
- Selain itu, semakin banyak variabel pada dataset, semakin tinggi pula jumlah sampel yang mewakili semua kombinasi kemungkinan nilai fitur. Model akan menjadi lebih kompleks dan akan meningkatkan kemungkinan overfitting.

Tujuan reduksi dimensi :

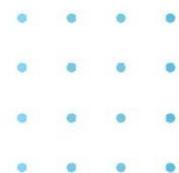
Reduksi dimensi bertujuan untuk menghindari overfitting. *Data training* dengan fitur yang lebih sedikit akan membuat model *machine learning* tetap simpel.



02

Technique

- PCA (Principal Component Analysis)
- LDA (Linear Discriminant Analysis)



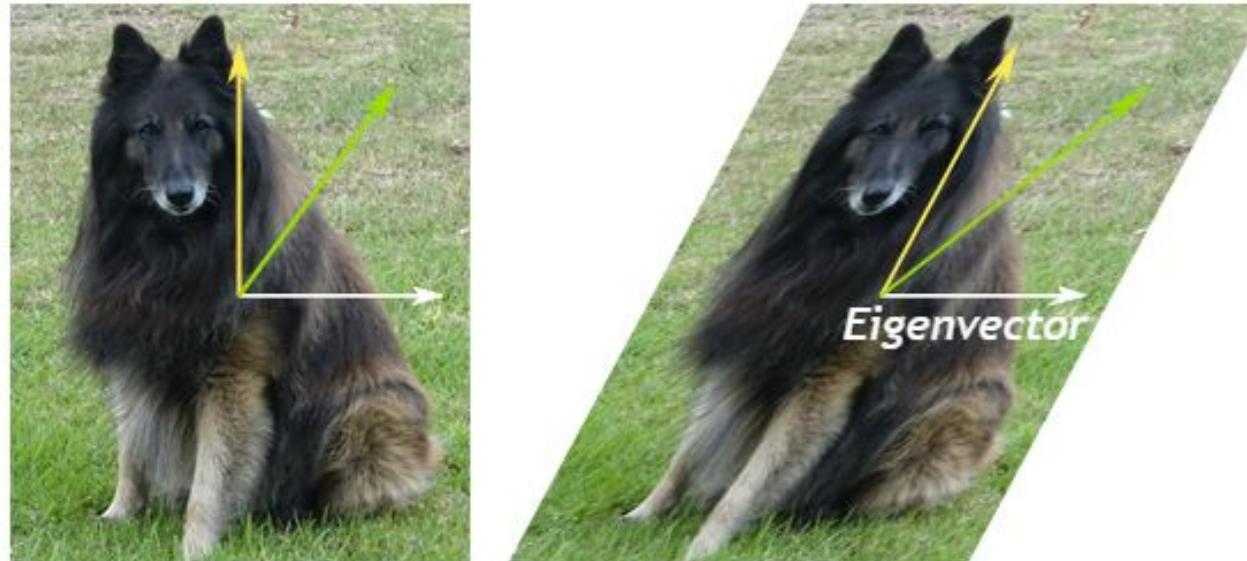
Teknik Dimensionality Reduction

1. PCA (Principal Component Analysis)
2. LDA (Linear Discriminant Analysis)



Review Nilai Eigen dan Vektor Eigen

Sebelum masuk ke PCA & LDA, mari mengingat lagi apa itu nilai eigen dan vektor eigen. Eigen berasal dari bahasa Jerman yang berarti karakteristik. Vektor eigen adalah vektor yang menjadi karakteristik sebuah matriks dimana arahnya tidak berubah meski dilakukan transformasi



Teknik Dimensionality Reduction

Review Nilai Eigen dan Vektor Eigen

Nilai eigen adalah bilangan yang berasosiasi dengan panjang vector. Dapat berubah setelah ditransformasi oleh matriks. Pada gambar berikut, A adalah matriks persegi, x adalah vektor eigen, dan λ adalah nilai eigen.

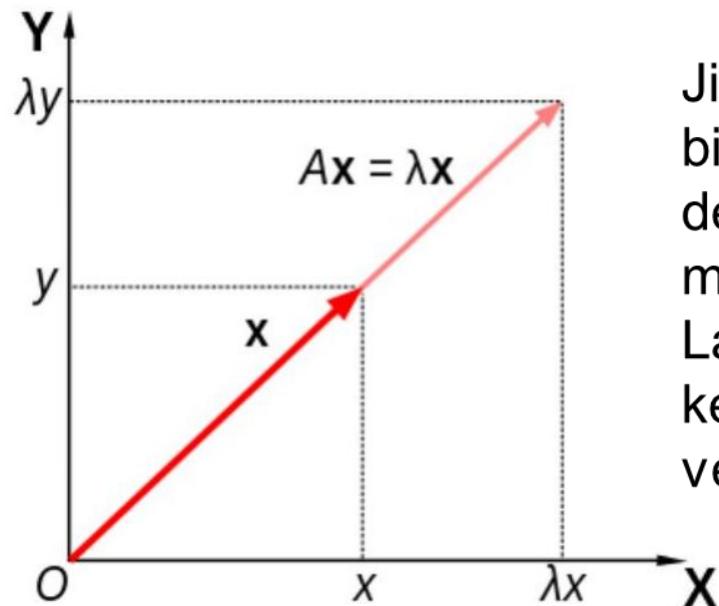
$$Ax = \lambda x$$

$$Ax - \lambda x = 0$$

$$Ax - \lambda Ix = 0$$

I merupakan matriks Identitas

$$(A - \lambda I)x = 0$$



Jika **x bukan vektor 0**, maka kita bisa menemukan nilai eigen dengan menghitung determinant matriks $(A - \lambda I)$.

Lalu **mensubtitusi nilai eigen** tsb ke $Ax = \lambda x$ untuk menemukan vektor eigennya.

Lynne J. Williams, et al.

Principal component analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables. Its goal is to extract the important information from the table, to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity of the observations and of the variables as points in maps.

Source : “Principle Component Analysis”, Journal Wires Computational Statistc, Vol.2, page 433-459, 2010

Alaa Tharwat, et al.

Linear Discriminant Analysis (LDA) is a very common technique for dimensionality reduction problems as a pre-processing step for machine learning and pattern classification applications.

Source : “Linear discriminant analysis: A detailed tutorial”, Journal AI Communications, vol. 30, no. 2, pp. 169-190, 2017

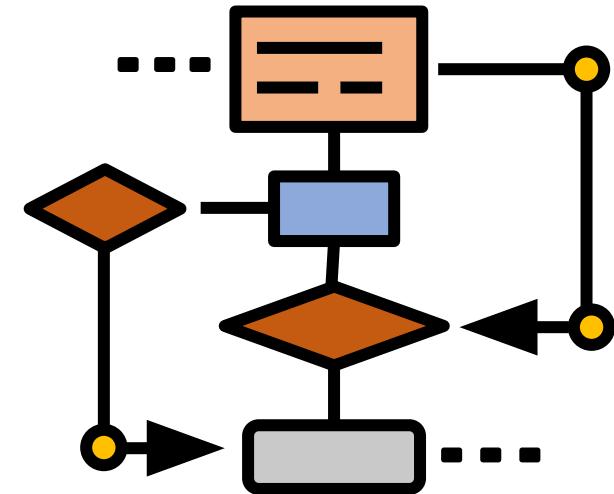
PCA (Principal Component Analysis)

Principal Component Analysis atau PCA adalah teknik reduksi dimensi yang paling populer. Teknik ini menggunakan operasi matriks sederhana dari aljabar linier dan statistik untuk menghitung proyeksi dari data asli ke dalam dimensi dengan jumlah yang sama atau lebih sedikit.

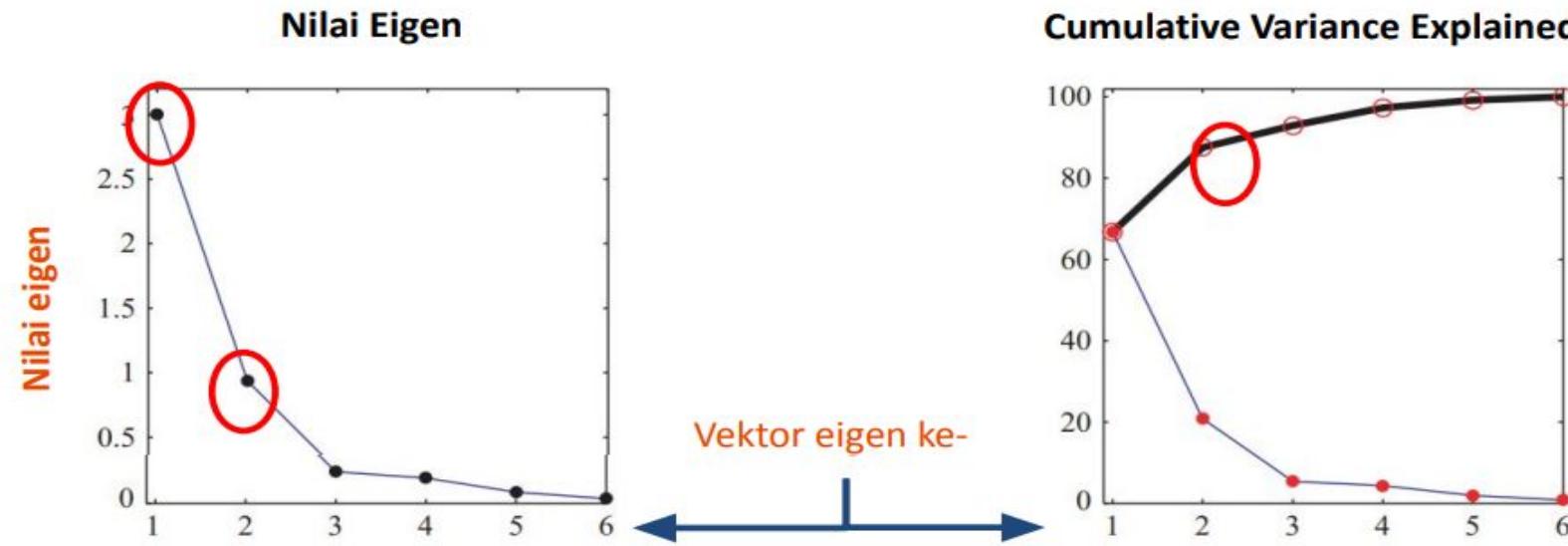


PCA (Principal Component Analysis)

1. Buat matriks dari data featurenya (misal matriks Mx)
2. Standardize data dengan rata-rata dan standar deviasi
3. Buat matriks covariant-nya dan hitung nilai eigen dan vektor eigennya
4. Pilih vektor eigen yang nilai eigennya cukup besar, buat matriks sesuai urutan besar nilai eigennya (misalnya matriks A)
5. Matriks komponen utama = $Mx * Me$

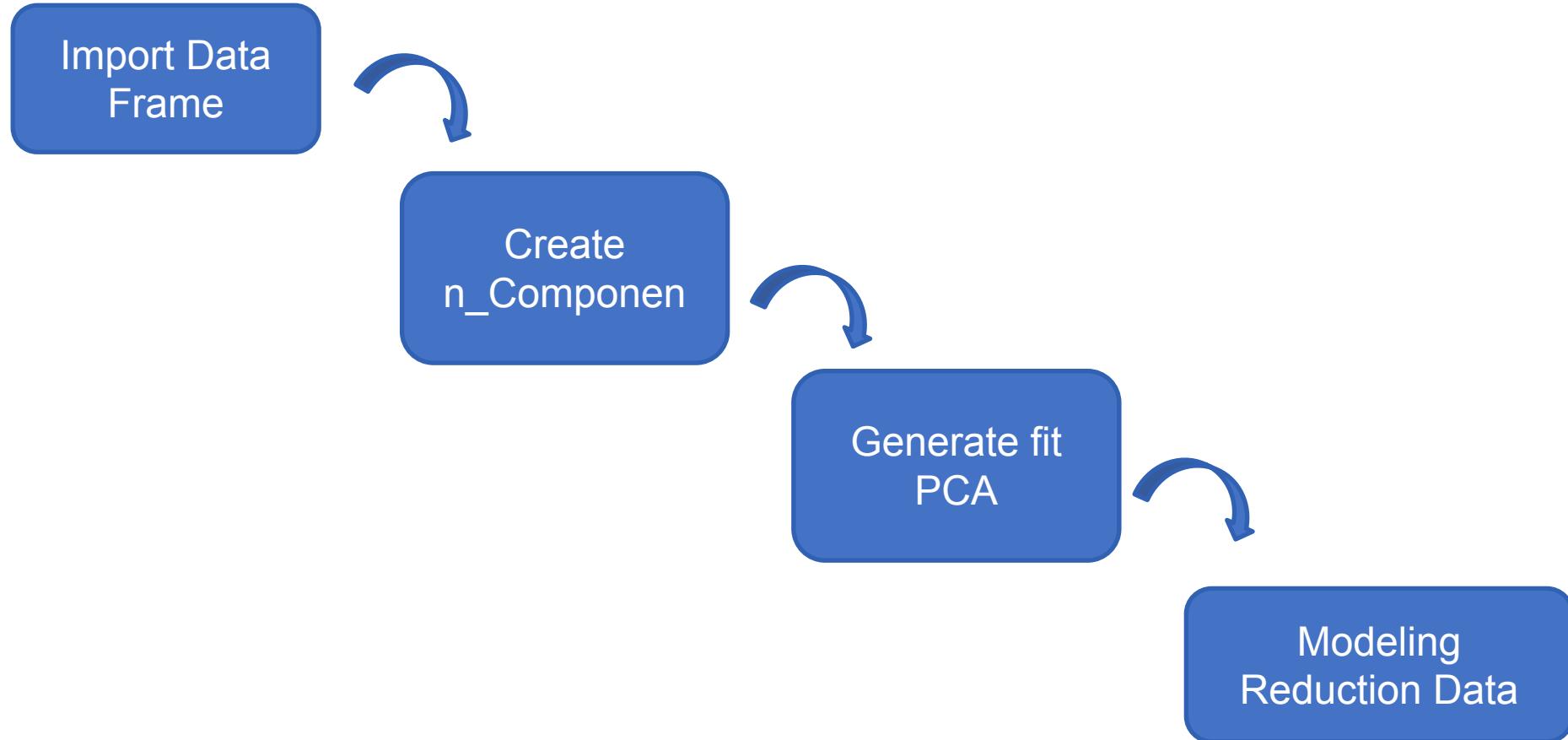


PCA (Principal Component Analysis)



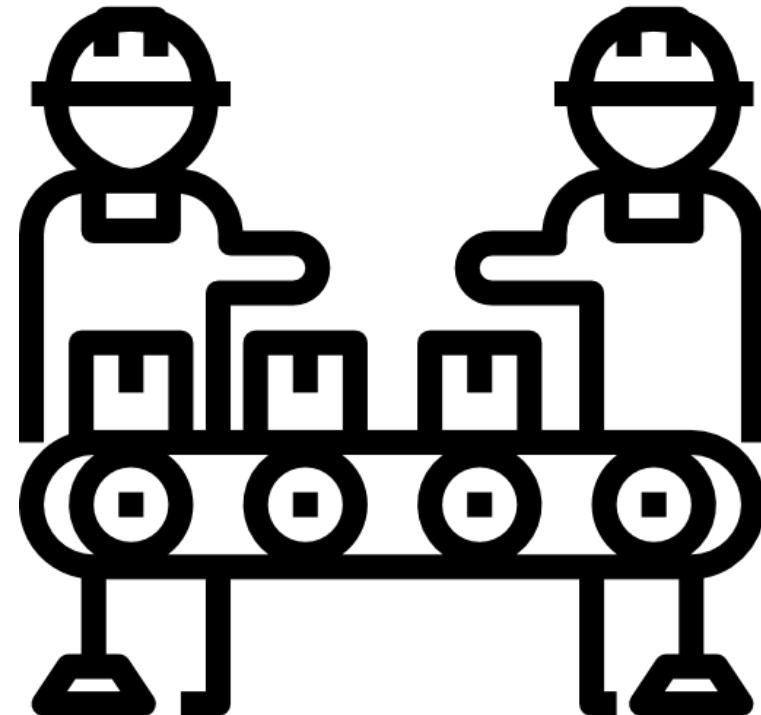
- Vektor eigen yang dipilih adalah yang nilai eigennya ≥ 1 (menurut Dillon & Goldstein (1984)), atau
- Komponen yang memiliki cumulative variance explained $\geq 75\%$ (Marison, 1976).

PCA Process



PCA Process

- Menyiapkan dataframe
- Dataframe dengan men-generate bilangan float random antara 0 dan 1 sebagai fitur-fiturnya (ada 10 fitur)
- Serta membuat label A, B, dan C sebanyak masing-masing 100 data.



PCA Process

```
import pandas as pd
import numpy as np
import random

data = {'feature_1': [random.uniform(0, 1) for i in range(100)],
        'feature_2': [random.uniform(0, 1) for i in range(100)],
        'feature_3': [random.uniform(0, 1) for i in range(100)],
        'feature_4': [random.uniform(0, 1) for i in range(100)],
        'feature_5': [random.uniform(0, 1) for i in range(100)],
        'feature_6': [random.uniform(0, 1) for i in range(100)],
        'feature_7': [random.uniform(0, 1) for i in range(100)],
        'feature_8': [random.uniform(0, 1) for i in range(100)],
        'feature_9': [random.uniform(0, 1) for i in range(100)],
        'feature_10': [random.uniform(0, 1) for i in range(100)],
        'label': [random.choice(['A', 'B', 'C']) for i in range(100)}}

df = pd.DataFrame(data)
df.head()
```

PCA Process

	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	feature_10	label
0	0.943520	0.406432	0.855416	0.446514	0.750474	0.921333	0.560966	0.988022	0.027169	0.505478	C
1	0.976941	0.592664	0.534635	0.195937	0.492972	0.932282	0.316151	0.356252	0.193895	0.967565	A
2	0.013127	0.340734	0.942017	0.793352	0.639681	0.388316	0.143313	0.350705	0.051329	0.972181	B
3	0.621991	0.973175	0.546698	0.117971	0.381381	0.600269	0.558188	0.241915	0.732767	0.814111	A
4	0.568197	0.147490	0.960422	0.354628	0.293954	0.910487	0.448038	0.030934	0.287319	0.142854	C

Info Penting : bahwa jika menggunakan dataset asli, pastikan data fitur sudah di-scaling, bisa dengan **normalisasi** atau **standarisasi** sehingga bentuknya mirip seperti output di atas.

PCA Process

```
x = df.iloc[:, :-1]
x.head()
```

	feature_1	feature_2	feature_3	feature_4	feature_5	feature_6	feature_7	feature_8	feature_9	feature_10
0	0.943520	0.406432	0.855416	0.446514	0.750474	0.921333	0.560966	0.988022	0.027169	0.505478
1	0.976941	0.592664	0.534635	0.195937	0.492972	0.932282	0.316151	0.356252	0.193895	0.967565
2	0.013127	0.340734	0.942017	0.793352	0.639681	0.388316	0.143313	0.350705	0.051329	0.972181
3	0.621991	0.973175	0.546698	0.117971	0.381381	0.600269	0.558188	0.241915	0.732767	0.814111
4	0.568197	0.147490	0.960422	0.354628	0.293954	0.910487	0.448038	0.030934	0.287319	0.142854

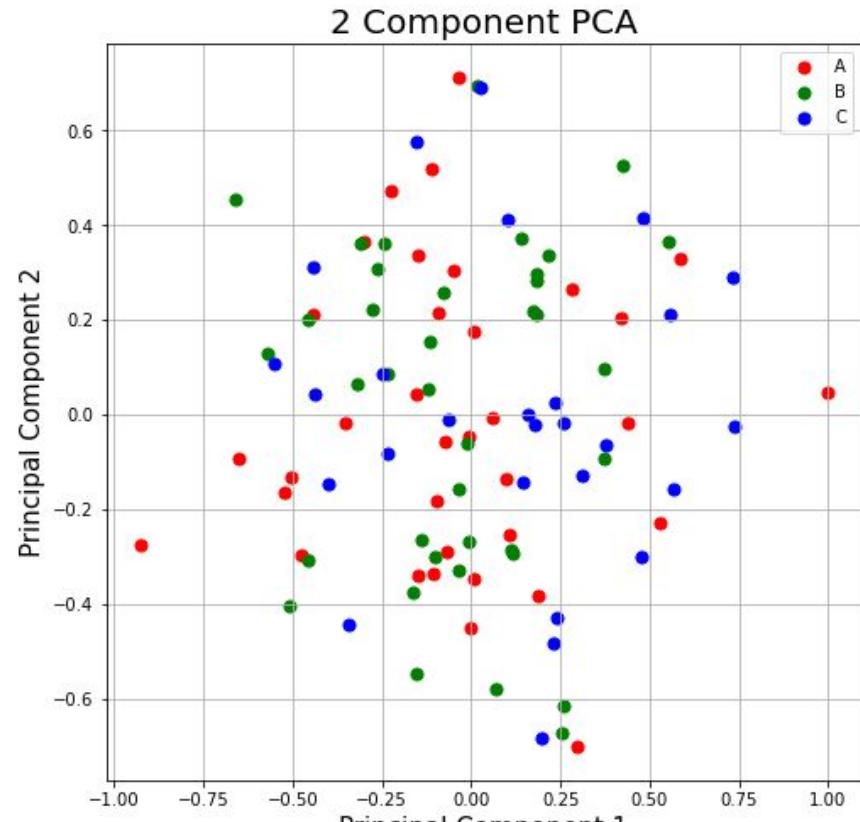
```
y = df.iloc[:, -1]
y.head()
```

```
0    C
1    A
2    B
3    A
4    C
Name: label, dtype: object
```

PCA Process

1. Setelah dataframe siap, selanjutnya adalah import library
2. Membuat Class PCA() dengan memberikan parameter `n_components` yang mendefinisikan jumlah komponen atau kolom baru yang diinginkan
3. Setelah membuat class, kita lakukan `fit_transform` pada data fitur. Jadi, nantinya 10 kolom fitur di atas akan diproses menggunakan teknik PCA sehingga akan mengeluarkan 2 kolom baru hasil reduksi.
4. Selanjutnya kita buat dataframe baru hasil `fit_transform` tersebut

PCA Process



Data Visualization – PCA

LDA (Linear Discriminant Analysis)

Linear Discriminant Analysis merupakan salah satu algoritma machine learning untuk klasifikasi yang juga dapat digunakan untuk reduksi dimensi. Cara kerjanya yaitu dengan menghitung statistik ringkas untuk fitur-fitur input menurut label, seperti *mean* dan standar deviasi.



LDA (Linear Discriminant Analysis)

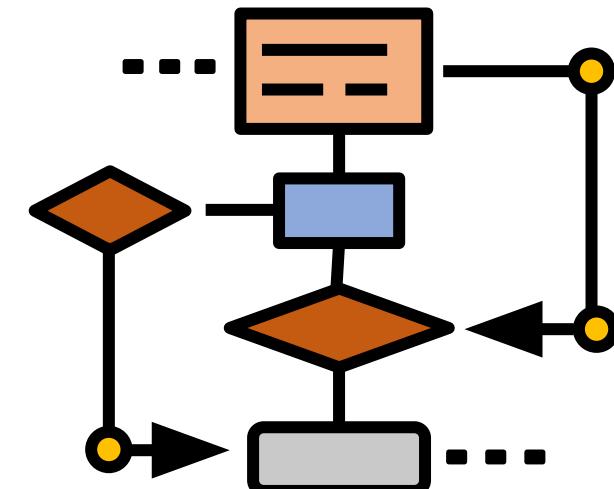
1. Untuk tiap kelas i, hitung rata-rata feature lalu dijadikan vektor \mathbf{m}_i , dan hitung vektor rata-rata feature keseluruhan (\mathbf{m})
2. Hitung matriks scatter within class

$$S_W = \sum_{i=1}^c S_i$$

where
 $S_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i) (\mathbf{x} - \mathbf{m}_i)^T$
(scatter matrix for every class)

3. Hitung matriks scatter antar kelas

$$S_B = \sum_{i=1}^c N_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$$



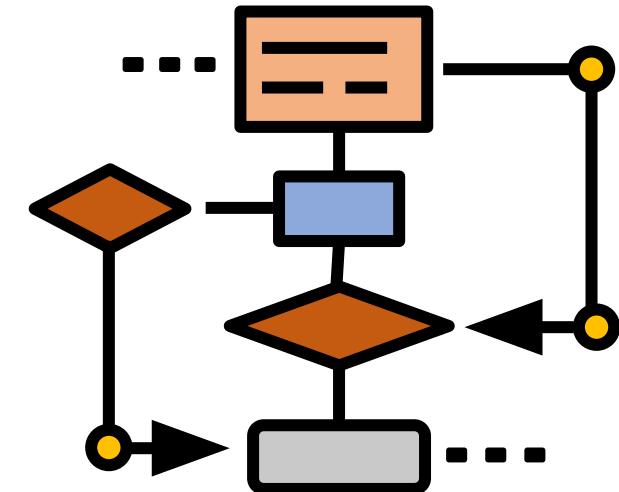
LDA (Linear Discriminant Analysis)

4. Hitung eigen vektor dan eigen value dari matriks:

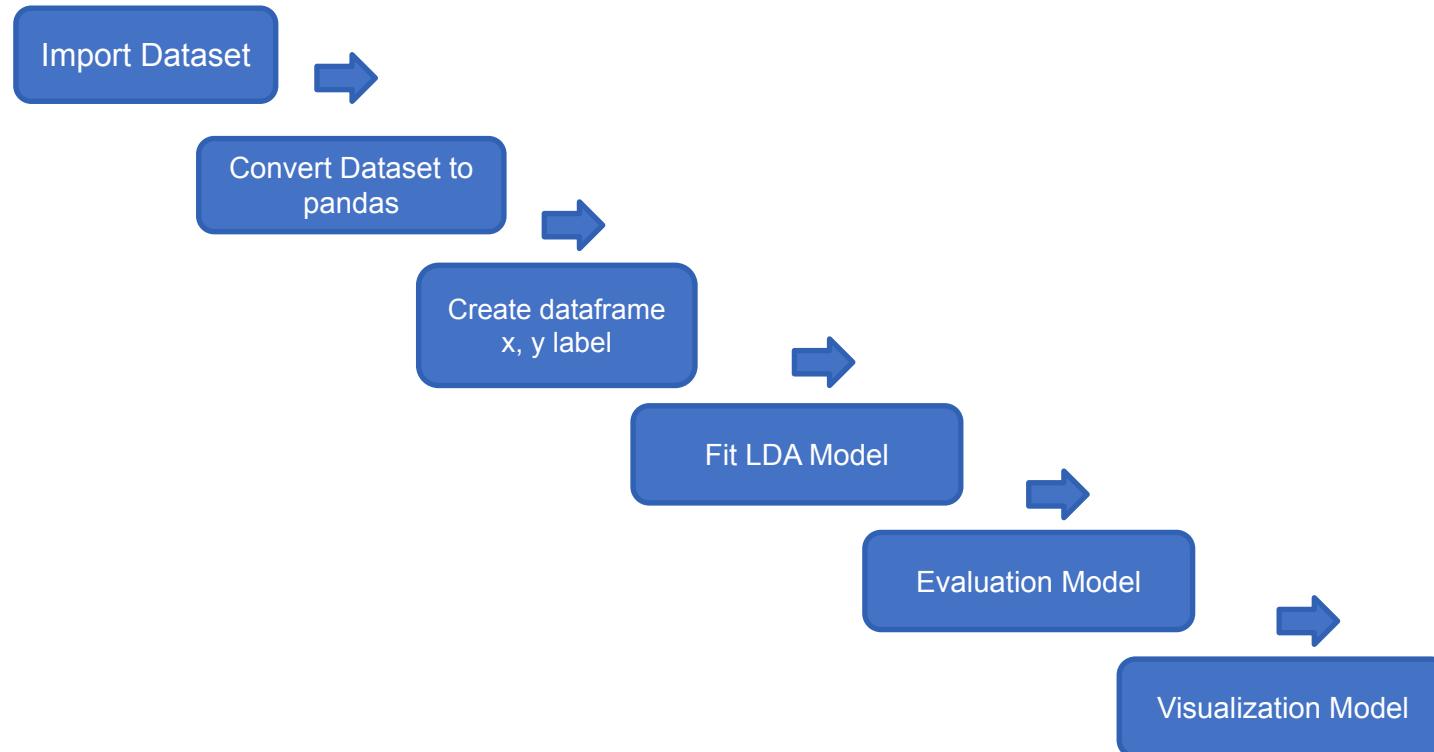
$$S_W^{-1} S_B$$

5. Pilih vektor eigen yang nilai eigennya cukup besar, buat matriks sesuai urutan besar nilai eigennya (misalnya matriks A)

6. Matriks feature terbaru = $X * A$. Proses pemilihan berapa vektor eigen yang dipilih sama dengan PCA



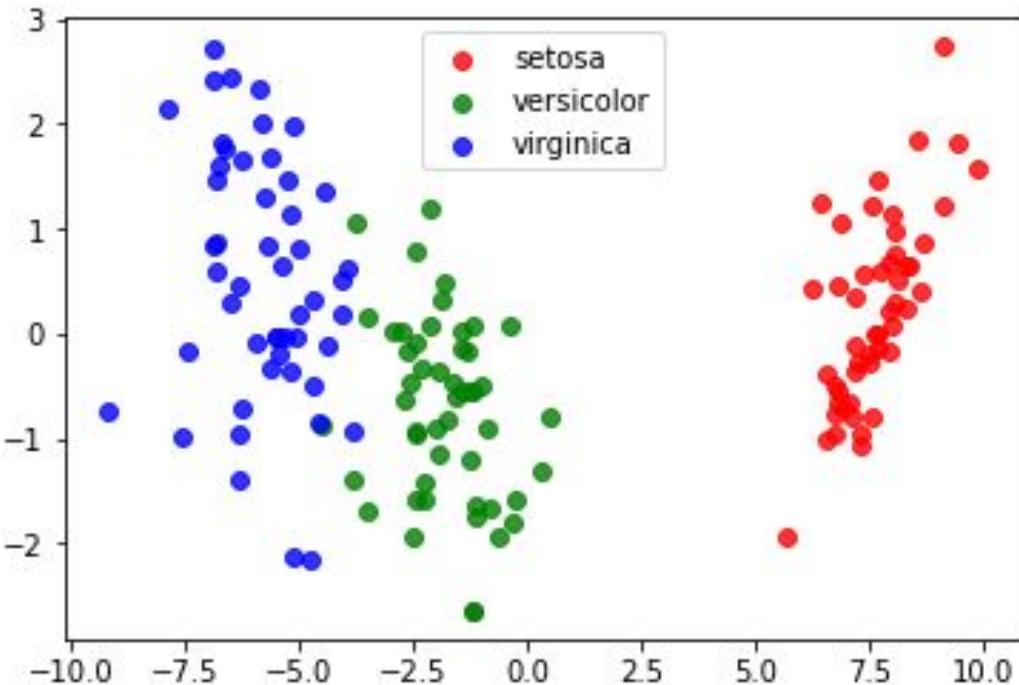
LDA Process



PCA Process

1. Import library yang dibutuhkan untuk support model LDA
2. Import dataset lalu konversi ke format dataframe menggunakan pandas
3. Mendefinisikan variabel x dan variabel y
4. Membuat fit model LDA
5. Melakukan evaluasi model
6. Melakukan prediksi untuk data observasi baru

PCA Process



- Didapatkan hasil plot data model LDA dengan sebaran data pada clusternya masing-masing
- Terlihat hasil plot dengan Batasan nilainya serta sebaran data yang linear (kesamaan data)

Persamaan :

Sama-sama menggunakan nilai eigen dan vektor eigen

Perbedaan :

PCA → unsupervised learning (tidak menggunakan label)

LDA → supervised learning (menggunakan label)

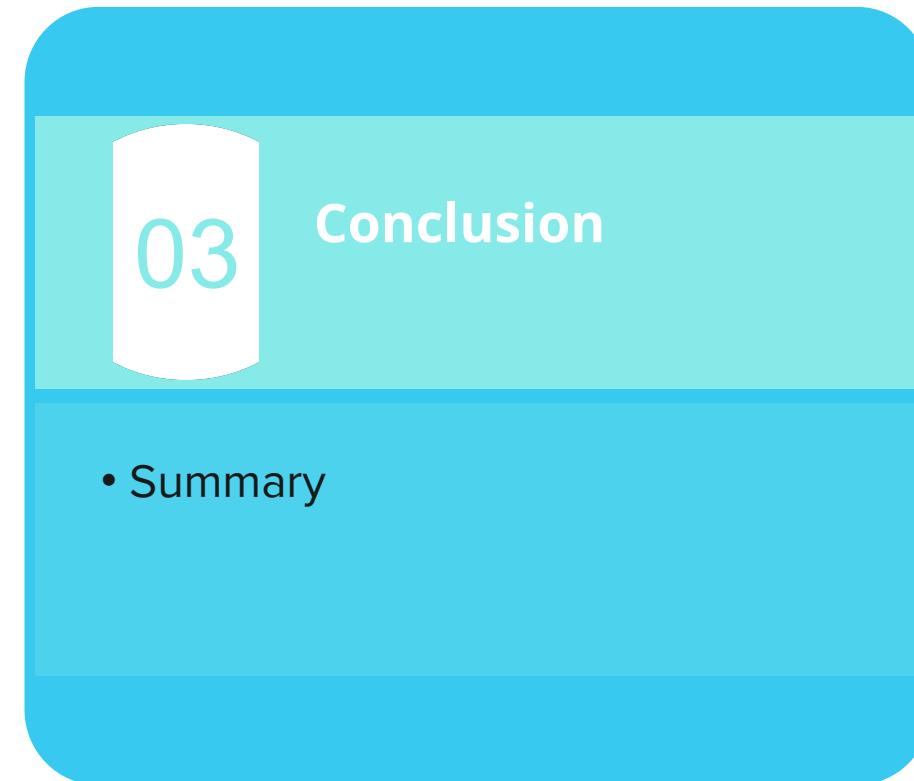
PCA → memaksimalkan variance di tiap feature terbaru

LDA → feature baru yang memaksimalkan jarak antar kelas

Dimensionality Reduction



Let's Code



03 Conclusion

- Summary

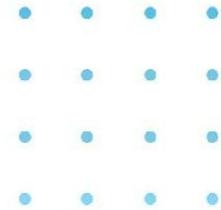
Conclusion

Summary

- Dimensionality reduction berfungsi untuk mengurangi fitur atau variabel predictor yang banyak menjadi beberapa fitur saja. Akan tetapi pengurangan fitur tidak menghilangkan informasi yang dimuat pada data tersebut
- Teknik yang digunakan ada 2, yakni : PCA dan LDA.
- Kedua teknik menggunakan nilai eigen dan vektor eigen dalam penyelesaian.

Reference

1. <https://ilmudatapy.com/teknik-reduksi-dimensi/>
2. <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wics.101>
3. <https://content.iospress.com/articles/ai-communications/aic729>
4. <https://archive.ics.uci.edu/ml/datasets.php>



TERIMA KASIH

Orbit Future Academy

PT Orbit Ventura Indonesia
Center of Excellence (Jakarta Selatan)
Gedung Veteran RI, Lt.15
Unit Z15-002, Plaza Semanggi
Jl. Jenderal Sudirman Kav.50, Jakarta
12930, Indonesia

- Jakarta Selatan/Pusat
- Jakarta Barat/BSD
- Kota Bandung
- Kab. Bandung
- Jawa Barat

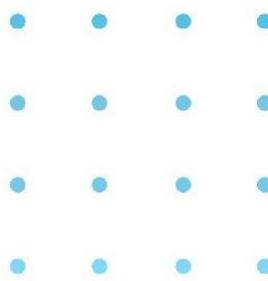
Hubungi Kami

Director of Sales & Partnership
ira@orbitventura.com
+62 858-9187-7388

Social Media

-  [Orbit Future Academy](#)
-  [@OrbitFutureAcademyIn1](#)
-  [OrbitFutureAcademy](#)
-  [Orbit Future Academy](#)

AI Mastery Course

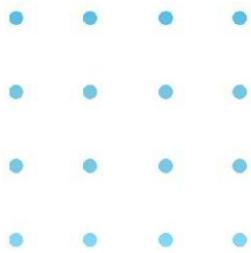


Data Science Section Recommender System



Tim Penyusun :

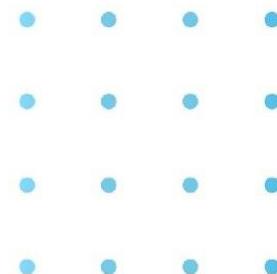
1. Andhini Rahma Santoso
 2. Ely Sudarsono
 3. Hafizah Ilma
 4. Shaifudin Zuhdi
 5. Uswatun Hasanah
-
-
-
-
-



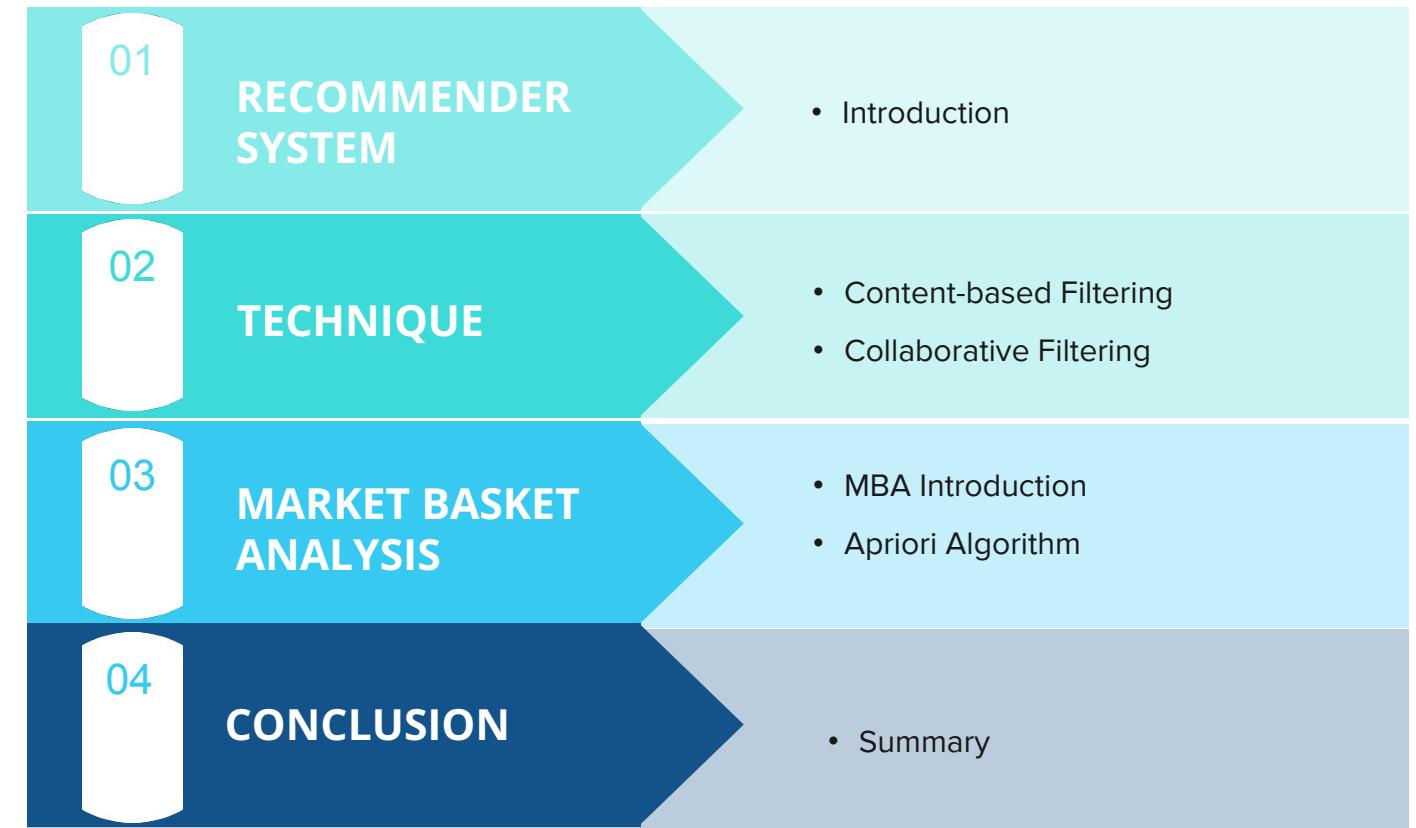
Learning Objectives

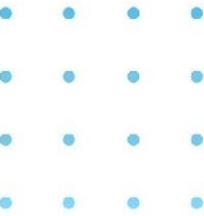
Di akhir modul ini kita diharapkan dapat memahami:

- Rekomendasi System dalam AI
- Market Basket Analysis
- Association Rules Mining
- Apriori Introduction



Agenda

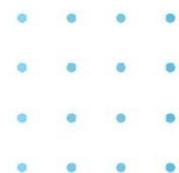




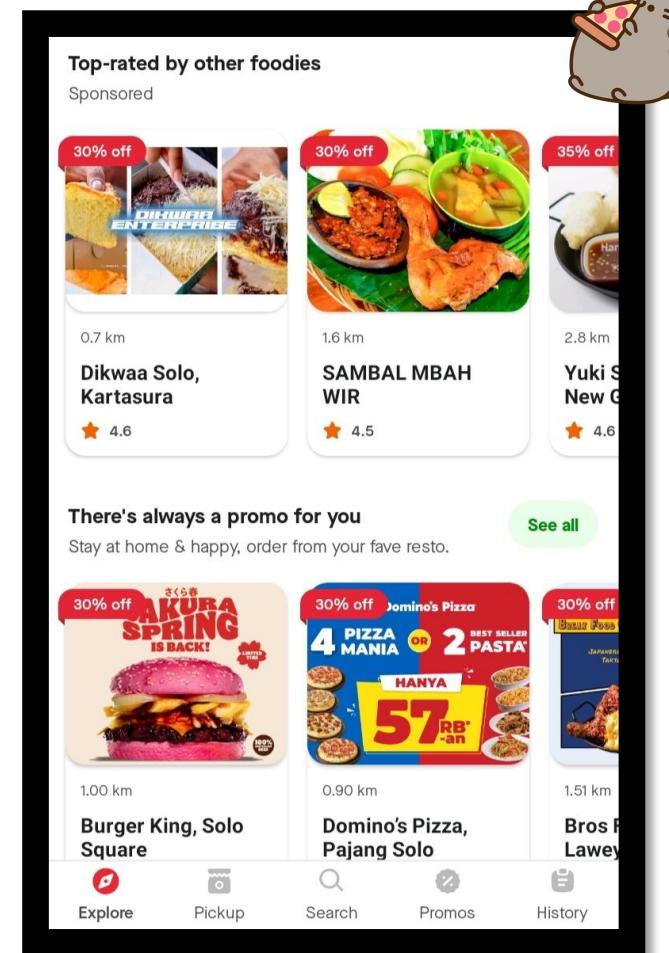
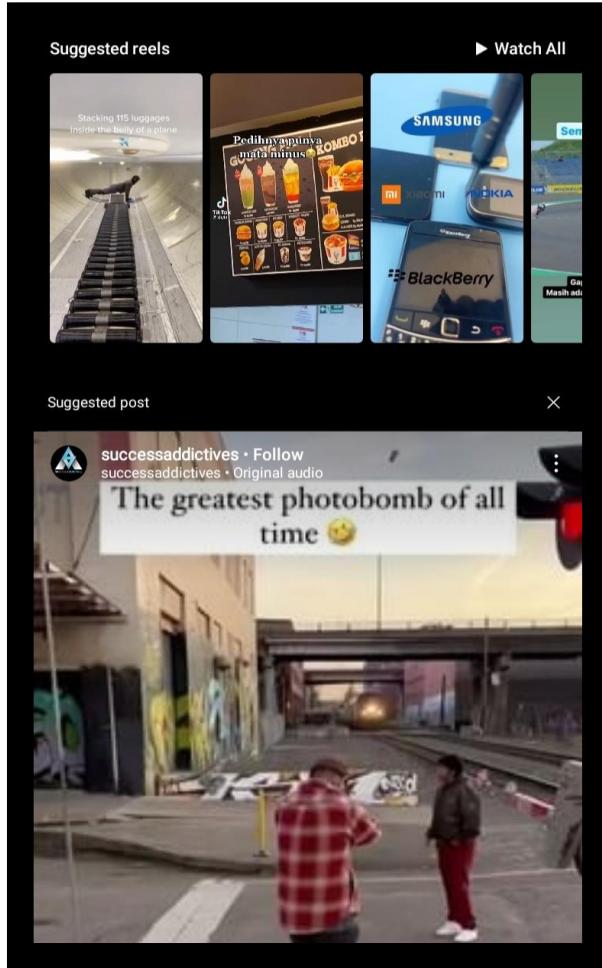
01

Recommender System

- Introduction



Kok bisa ya IG mengetahui konten bermanfaat buat saya? Go-Food mengetahui apa yang saya butuhkan ?



Instagram : Suggested Reels, Suggested Post

Go-Food : Top-rated by other, Promo for you

Recommender System



Source: analyticsindiamag.com

- Memberikan **informasi** berupa saran objek yang kemungkinan diminati / dibutuhkan pengguna.
- Data input yang digunakan berupa **profil** atau **riwayat aktivitas pengguna**.
- **Tujuan** dari sistem rekomendasi adalah untuk **meningkatkan aktivitas pengguna** dengan memberikan daftar item yang "**disarankan**".

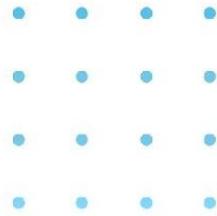
Metode Recommender System



[Source : dbversity.com](https://dbversity.com/)

Terdapat berbagai metode untuk membuat sistem rekomendasi ini, diantaranya:

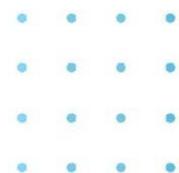
- Collaborative filtering recommendations
- Content-based recommendation
- Hybrid recommendation
- dan lain-lain



02

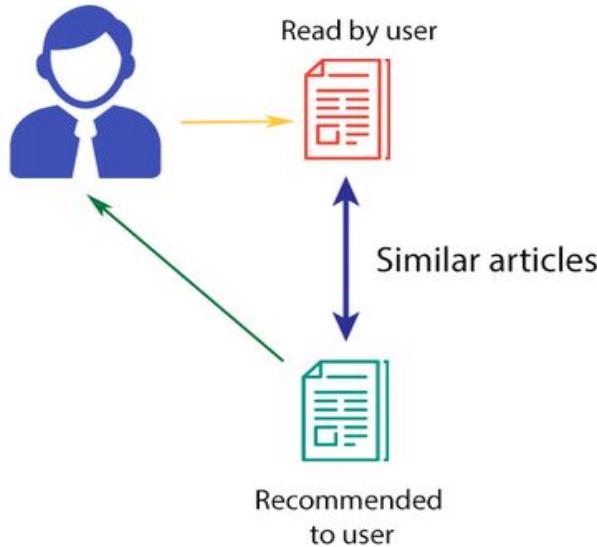
Technique

- Content-based Filtering
- Collaborative Filtering



A. Content-Based Filtering

CONTENT-BASED FILTERING



- Pendekatan ini menggunakan informasi tambahan tentang pengguna atau item.
- Metode pemfilteran ini menggunakan fitur item untuk merekomendasikan item lain yang serupa dengan apa yang disukai pengguna dan juga berdasarkan tindakan mereka sebelumnya atau umpan balik eksplisit.

A. Content-Based Filtering

Bagaimana mengetahui kemiripan suatu user/item?

1. Membuat vektor yang melambangkan tiap item yang disebut sebagai **item profile** dan vektor yang melambangkan user disebut **user profile**.

Item Profile

	Actor1	Actor2	Comic
Movie A	1	0		0	1
Movie B	0	1			
Movie C				1	1

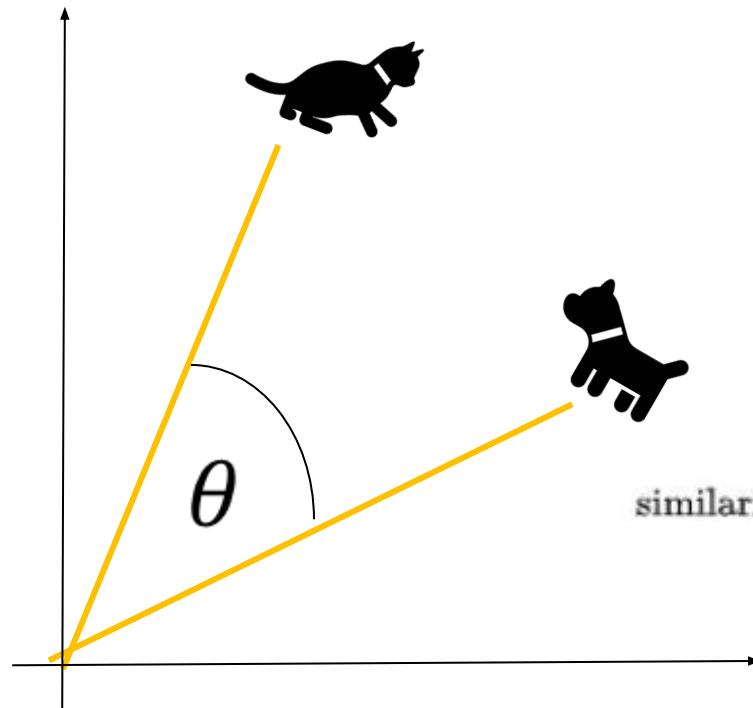
User Profile

	Movie1	Movie2	Drama
User	1	0		0	1

A. Content-Based Filtering

Bagaimana mengetahui kemiripan suatu user/item?

2. Menghitung kemiripan tiap item profile dengan user profile. Terdapat beberapa metode seperti Cosine Similarity, Euclidean Distance, dan Pearson's Correlation



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

A. Content-Based Filtering

Bagaimana mengetahui kemiripan suatu user/item?

- Euclidian Distance

$$sim(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

- Pearson's Correlation

$$sim(x, y) = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n \sum x^2 - (\sum x)^2)(n \sum y^2 - (\sum y)^2)}}$$

- Manhattan Distance

$$sim(x, y) = \sum_{i=1}^n |x_i - y_i|$$

A. Content-Based Filtering

Bagaimana mengetahui kemiripan suatu user/item?

3. Kemudian kita melakukan sorting item berdasarkan **skor similarity**-nya, sehingga:

Item	Similarity	Recommended?
Movie A	0.95	Yes
Movie B	0.88	Yes
Movie C	0.54	No
Movie D	0.30	No



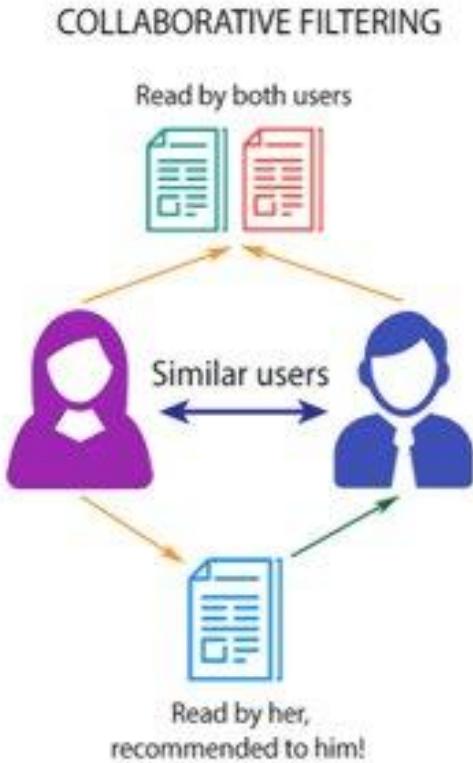
beberapa item dengan **skor tertinggi** muncul sebagai rekomendasi

A. Content-Based Filtering

Contoh lain untuk data artikel/buku/sinopsis film. Diawali dengan membuat bag of words, selanjutnya bentuk vektor dari masing-masing film (kita dapat menggunakan TF IDF).

	1 This	2 movie	3 is	4 very	5 scary	6 and	7 long	8 not	9 slow	10 spooky	11 good	Length of the review(in words)
Review 1	1	1	1	1	1	1	1	0	0	0	0	7
Review 2	1	1	2	0	0	1	1	0	1	0	0	8
Review 3	1	1	1	0	0	0	1	0	0	1	1	6

B. Collaborative Filtering



- Merekendasikan item berdasarkan feedback (rating) masa lalu dari sekelompok pengguna, dengan asumsi:

"Customers who had **similar tastes in the past**, will have **similar tastes in the future.**"

B. Collaborative Filtering

1. Memory Based

- Item yang disarankan adalah item yang disukai (diberi rating tinggi) oleh pengguna target, atau yang mirip dengan item lain yang disukai oleh pengguna target (similarity-based methods)

Metode Naïve Neighborhood

- Menghitung similarity dan membuat prediksi

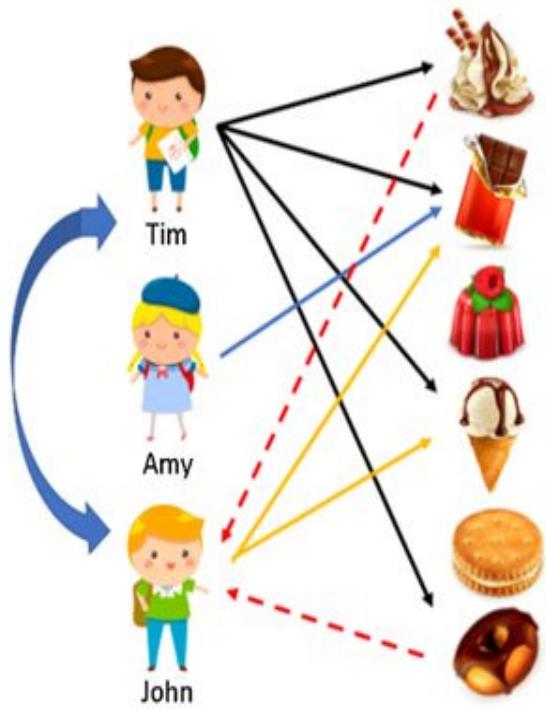
2. Model-Based

- Item yang disarankan dipilih berdasarkan hasil model yang dilatih untuk mengidentifikasi pola pada data input (clustering, dimensionality reduction, diffusion-based methods).

- Matrix Factorization
- Restricted Boltzman Machine

B. Collaborative Filtering

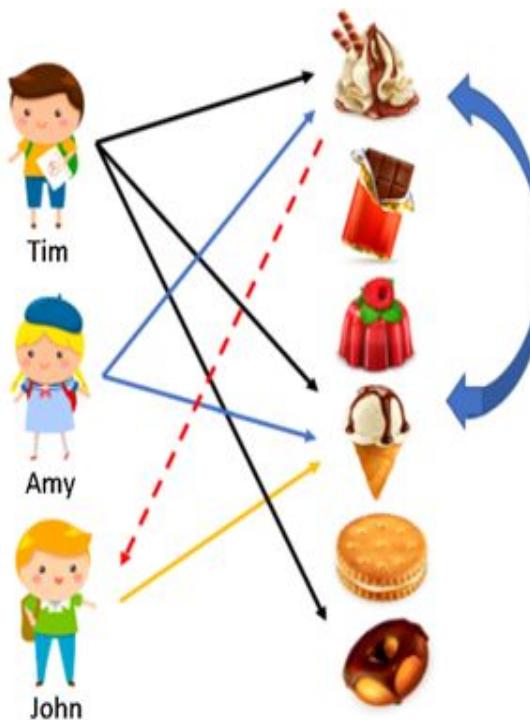
Memory-based



(a) User-based filtering

Source:

<https://predictivehacks.com/item-based-collaborative-filtering-in-python/>



(b) Item-based filtering

Dibagi menjadi dua kelas, yaitu:

- **User-based**

Metode yang merekomendasikan item dengan melihat kemiripan sekelompok pengguna dengan active user (pengguna yang ingin kita coba berikan rekomendasi).

- **Item-based.**

Metode rekomendasi yang didasari dengan melihat kesamaan antar item menggunakan peringkat (rating) oleh pengguna.

B. Collaborative Filtering

Memory (User)-based

Diberikan data rating dalam bentuk tabel di bawah ini. Kemudian menentukan apakah item 5 recommended atau tidak untuk Alice.

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

B. Collaborative Filtering

Memory (User)-based

Menghitung similarity antara Alice dan user lainnya. Hasilnya bahwa Alice terlihat lebih mirip dengan User 1 dan 2

	Item1	Item2	Item3	Item4	Item5
Alice	5	3	4	4	?
User1	3	1	2	3	3
User2	4	3	4	3	5
User3	3	3	1	5	4
User4	1	5	5	2	1

Menggunakan Pearson correlation

sim = 0.853

sim = 0.707

sim = 0.0

sim = - 0.792

B. Collaborative Filtering

Menghitung similarity antar Alice (x) dan user1 (y)

x	y	xy	x^2	y^2
5	3	15	25	9
3	1	3	9	1
4	2	8	16	4
4	3	12	16	9

$$\begin{aligned}\sum x &= 16 \\ \sum y &= 9 \\ \sum xy &= 38\end{aligned}$$

$$\sum x^2 = 66$$

$$\sum y^2 = 23$$

$$sim(x, y) = \frac{4(38) - (16)(9)}{\sqrt{(4(66) - (16^2))(4(23) - (9^2))}} = 0,853$$

Menghitung similarity antar Alice (x) dan user2 (u)

x	u	xu	x^2	u^2
5	4	20	25	16
3	3	9	9	9
4	4	16	16	16
4	3	12	16	9

$$\begin{aligned}\sum x &= 16 \\ \sum u &= 14 \\ \sum xu &= 57\end{aligned}$$

$$\begin{aligned}\sum x^2 &= 66 \\ \sum u^2 &= 50\end{aligned}$$

$$sim(x, u) = \frac{4(57) - (16)(14)}{\sqrt{(4(66) - (16^2))(4(50) - (14^2))}} = 0,707$$

B. Collaborative Filtering

Memory (User)-based

Membuat prediksi rating berdasarkan rating dari user lain yang paling mirip

- Estimasi dengan **rata-rata** dari rating user lain yang mirip:

$$pred(alice, item5) = \frac{\sum_{b \in N} r_{user}}{N} = \frac{3 + 5}{2} = 4$$

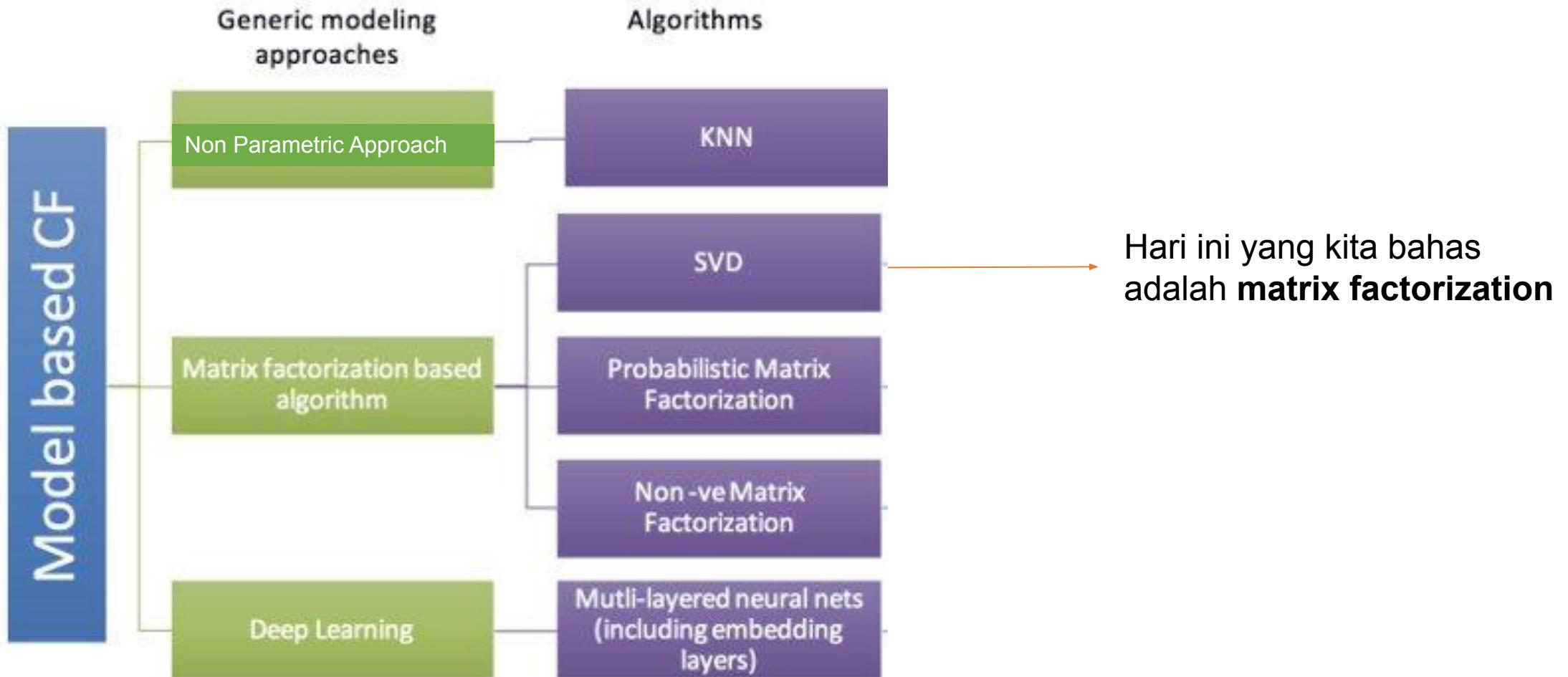
- Estimasi dengan “weighted average”

$$pred(alice, item5) = \bar{r} + \frac{\sum_{b \in N} (sim(alice, b)(r_{(b, item5)} - \bar{r}_b))}{\sum_{b \in N} sim(alice, b)}$$

$$pred(alice, item5) = 4 + \frac{0,853(3 - 2,25) + 0,707(5 - 3,5)}{0,853 + 0,707} = 5,09$$

B. Collaborative Filtering

Model-based



B. Collaborative Filtering

Model-based

Dekomposisi Nilai Singular (Singular Value Decomposition / SVD)

Dekomposisi SVD adalah metode untuk membuat matriks terdekomposisi menjadi menjadi 3 matriks sebagai berikut:

$$A_{n \times d} = U_{n \times r} \begin{pmatrix} D_{r \times r} & V^T_{r \times d} \end{pmatrix}$$

dimana U & V matriks ortogonal

U = vector eigen dari AA^T

V = vektor eigen dari A^TA

r = rank dari matriks

D = matriks diagonal, diagonalnya adalah akar dari nilai eigen A^TA dan AA^T

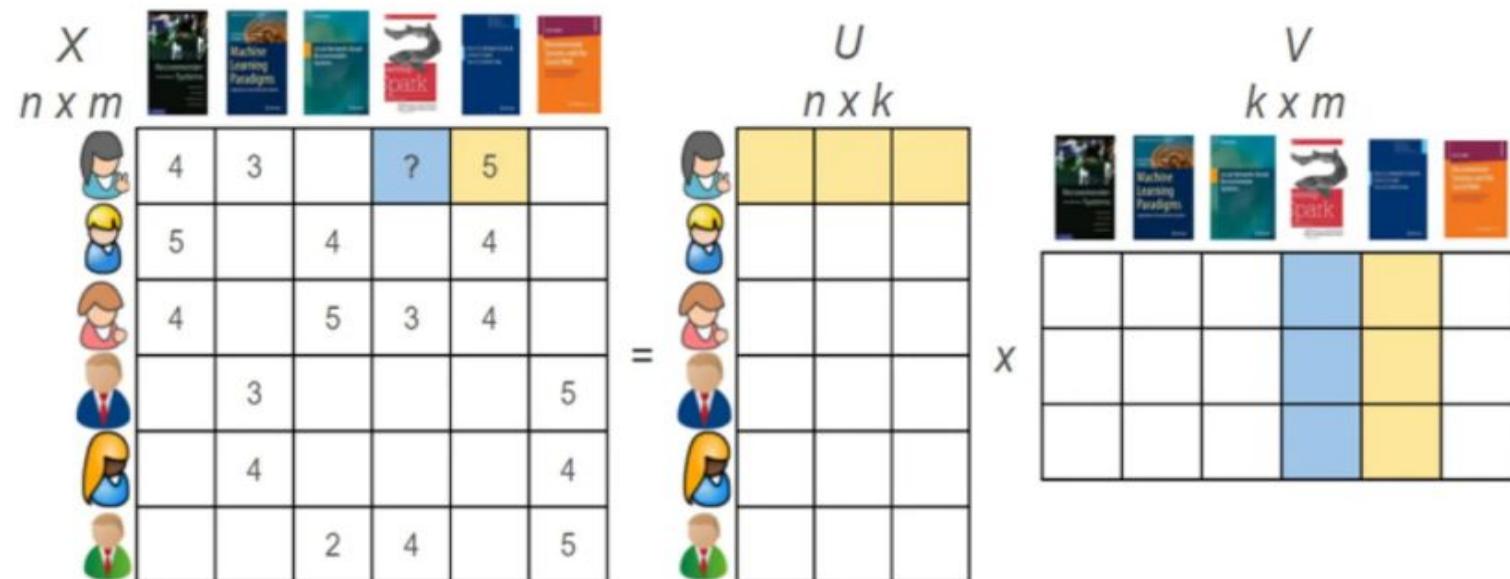
B. Collaborative Filtering

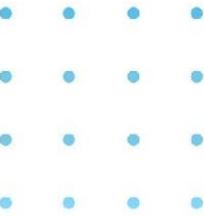
Model-based

Dekomposisi Nilai Singular (Singular Value Decomposition / SVD)

SVD digunakan untuk mengkomposisi nilai matriks rating relative dengan penilaian 1 dan yang lainnya.

Sehingga, user mendapat rekomendasi item berdasarkan rating yang disukai user lain yang punya kesamaan selera

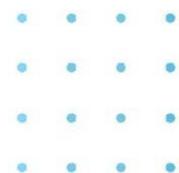




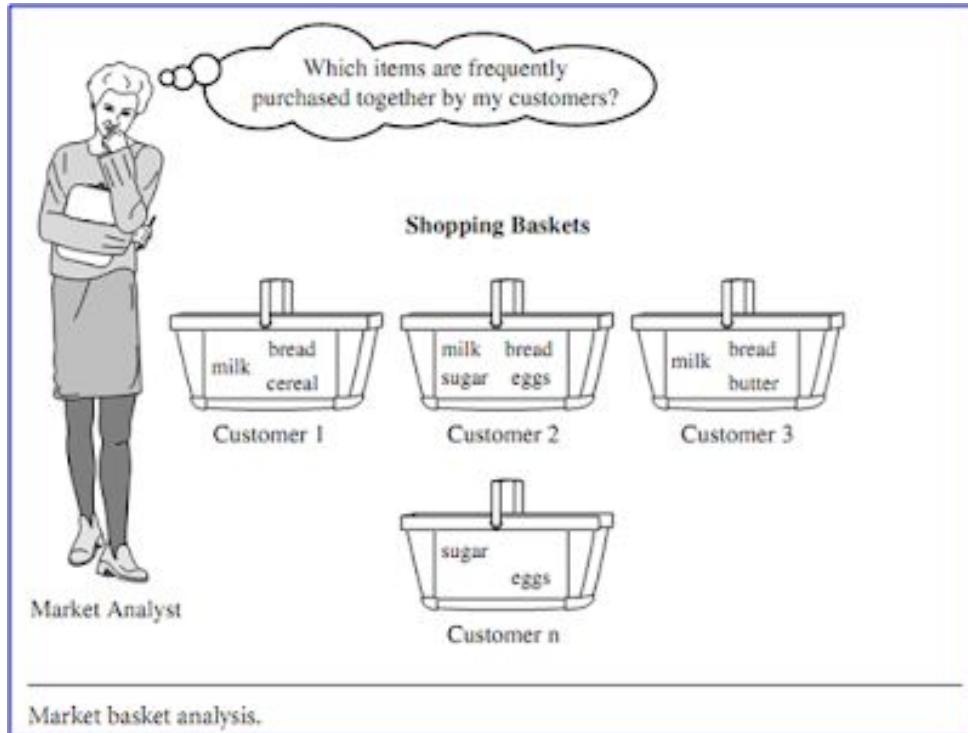
03

Market Basket Analysis

- MBA Introduction
- Algoritma Apriori



Market Basket Analysis



www.ilmuskippsi.com

Market basket analysis adalah pencarian pengetahuan dari suatu **transaksi data** ketika kita tidak mengetahui pola spesifik apa yang kita cari.

Kebutuhan *market basket analysis* berawal dari keakuratan dan manfaat yang dihasilkannya dalam wujud **aturan assosiasi (association rules)**.

Yang dimaksud dengan *association rules* adalah pola-pola keterkaitan data dalam basis data.

Market Basket Analysis



Market basket analysis adalah teknik yang digunakan oleh pengecer besar untuk mengungkap hubungan antar item.

Teknik ini mencari kombinasi item yang sering terjadi bersamaan dalam transaksi, memberikan informasi untuk memahami perilaku pembelian. Hasil dari jenis teknik ini adalah, secara sederhana, seperangkat aturan yang dapat dipahami sebagai "jika ini, maka itu".

Fungsi

Mie Instant ==> Saos Sambal

Rules digunakan dalam marketing untuk membuat berbagai keputusan, contohnya:

- Letakkan kedua barang berdekatan (agar tidak lupa keduanya untuk dibeli)
- Letakkan kedua barang berjauhan (agar konsumen akan melihat-lihat barang yang lain)
- Satukan kedua barang dalam sebuah promo (promo akan jadi lebih menarik karena konsumen memang membutuhkan keduanya)
- Satukan kedua barang dengan barang lain yang kurang laku (Cross selling)
- Naikkan barang yang satu dan turunkan yang lain (teknik kompetisi dengan “toko sebelah”)
- Jangan iklankan kedua barang bersamaan.
- Tawarkan promo saos dalam bentuk sachet gratis setiap membeli mie instan premium.

Market Basket Analysis

$$\text{Support} = \frac{\text{frq}(X, Y)}{N}$$
$$\text{Rule: } X \Rightarrow Y \rightarrow \text{Confidence} = \frac{\text{frq}(X, Y)}{\text{frq}(X)}$$
$$\text{Lift} = \frac{\text{Support}}{\text{Supp}(X) \times \text{Supp}(Y)}$$

<http://www.saedsayad.com/>

- Nilai **Support** merupakan persentase dari semua transaksi yang terjadi yang mengandung itemset tersebut.
- Nilai **Confidence** merupakan perbandingan antara nilai support dari himpunan items yang terdapat di dalam rule dan nilai support himpunan items yang mendahuluinya.



Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Nilai **lift rasio** merupakan suatu ukuran dalam mengetahui kekuatan suatu aturan asosiasi. Lift rule $A \Rightarrow B$ adalah sebuah ukuran seberapa lebih sering A dan B muncul bersamaan dibandingkan jika mereka saling bebas secara statistika. Range nilai lift ratio adalah 0 sampai tak hingga

- Jika lift ratio < 1 , artinya A dan B berkorelasi negatif
- Jika lift ratio $= 1$, artinya A dan B tidak berkorelasi
- Jika lift ratio > 1 , artinya A dan B berkorelasi positif, sering muncul bersama-sama.

Market Basket Analysis



$\{roti\} \rightarrow \{susu\}$
(support = 40%, confidence = 50%)

- Terdapat 40% dari seluruh transaksi memuat roti dan susu”.
- Seorang konsumen yang membeli roti mempunyai kemungkinan 50% untuk membeli susu.

Market Basket Analysis

Studi Kasus - MBA for Bakery Shop



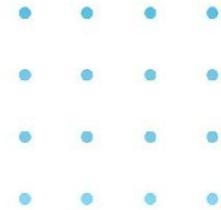
Link Google Colab : <https://bit.ly/MarketBasketAnalysis-Aproriari>

Summary

- Recommender system digunakan berbagai aplikasi, seperti Go-Jek, Instagram, dll.
- Recommender system bekerja berdasar karakteristik user dan/atau item.
- Item yang memiliki karakter sama maka akan direkomendasikan.
- Teknik recommender system yakni market basket juga diterapkan untuk teknik marketing produk.

Referensi

- Recommender systems oleh LinyuanLüa,
MatúšMedob,ChiHoYeungb,Yi-ChengZhangb,Zi-KeZhanga, TaoZhoud
(<https://doi.org/10.1016/j.physrep.2012.02.006>)
- SISTEM REKOMENDASI PRODUK SEPATU DENGAN MENGGUNAKAN METODE COLLABORATIVE FILTERING oleh Arif Kurniawan (ISSN: 2089-9815)
- User-based Collaborative Filtering dengan memanfaatkan Pearson Correlation Untuk Mencari Neighbor Terdekat Dalam Sistem Rekomendasi oleh Arvid Theodorus, Djoko Budiyanto Setyohadi, Ernawati
- Sistem Rekomendasi Mobil Berdasarkan Demographic dan Content-based Filtering oleh Herastia Maharani, Ferry Gunawan. Jurnal Telematika



TERIMA KASIH

Orbit Future Academy

PT Orbit Ventura Indonesia
Center of Excellence (Jakarta Selatan)
Gedung Veteran RI, Lt.15
Unit Z15-002, Plaza Semanggi
Jl. Jenderal Sudirman Kav.50, Jakarta
12930, Indonesia

- Jakarta Selatan/Pusat
- Jakarta Barat/BSD
- Kota Bandung
- Kab. Bandung
- Jawa Barat

Hubungi Kami

Director of Sales & Partnership
ira@orbitventura.com
+62 858-9187-7388

Social Media

-  [Orbit Future Academy](#)
-  [@OrbitFutureAcademyIn1](#)
-  [OrbitFutureAcademy](#)
-  [Orbit Future Academy](#)

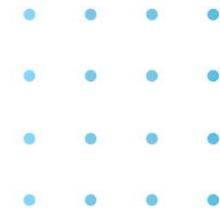
AI Mastery Course



Module
Data Science

Section
Time Series

(Statistical Approach)





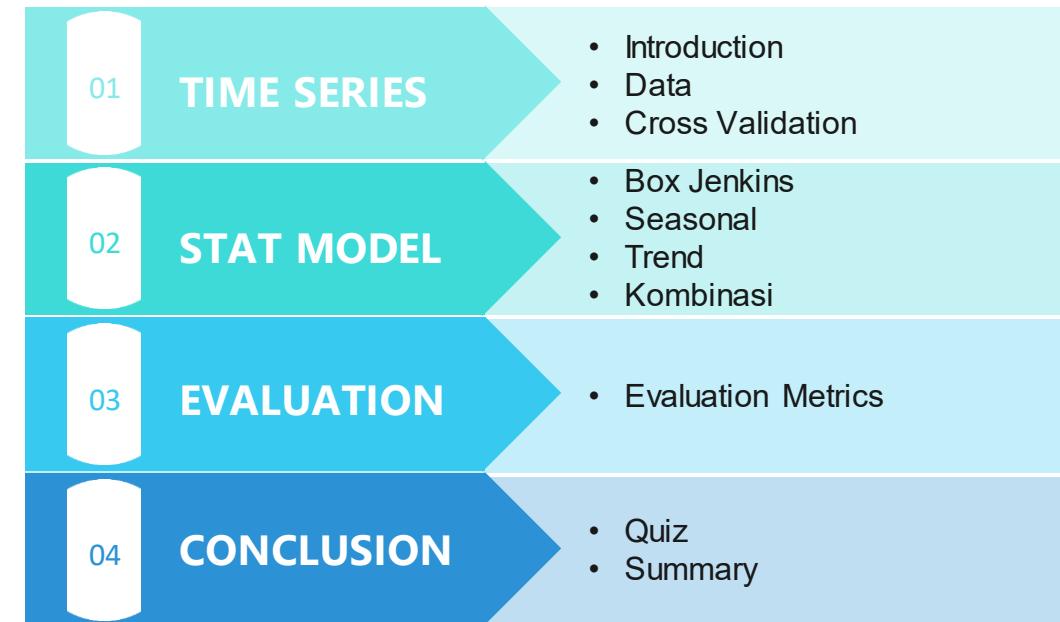
Learning Objectives

Di akhir modul ini kita diharapkan dapat memahami:

- Pengertian data time series dan contoh data
- Bagaimana membagi data time series menjadi data train dan data test
- Pola data yang ada pada data time series
- Syarat yang harus dipenuhi data untuk bisa dimodelkan secara time series



Agenda





01

TIME SERIES

- Introduction
- Data
- Cross Validation

Introduction

Data time series atau data runtun waktu adalah suatu himpunan data pengamatan yang dibangun dalam urutan waktu.

Model time series mengasumsikan bahwa kejadian di waktu t berhubungan dengan kejadian di $t-1$, $t-2$, dst.

$t-5$ jam $t-4$ jam $t-3$ jam



$t-2$ jam $t-1$ jam Saat t

Contoh : Data cuaca pada beberapa jam pada periode tertentu

Introduction

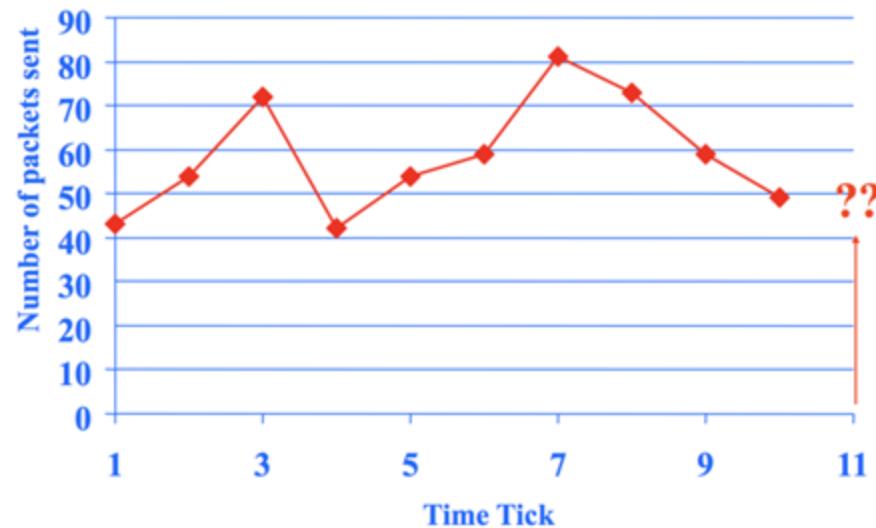
Data

Cross Section	Time Series	Longitudinal Data
Data yang tidak bergantung dengan waktu	Variabel dependent (label) bergantung pada waktu	Variabel independent (feature) bergantung pada waktu
Nilai (Y) Jam belajar (X1) Gizi (X2)	Jam(t) Cuaca(Y1t) Suhu(Y2t)	Brand Tahun Tingkat promo(Xt) Penjualan(Yt)
10 9 8	13 berawan 27	A 2011 9 8000
9 9 8	14 mendung 26	A 2012 8 10000
8 8 7	15 hujan 25	A 2013 7 10000
8 7 8	16 hujan 25	B 2011 8 8000
7 7 7	17 hujan 24	B 2012 9 9000
10 8 9	18 berawan 24	B 2013 7 8000

Introduction

Time series banyak digunakan untuk membuat forecasting, diantaranya :

1. Prediksi cuaca
2. Prediksi jumlah penumpang pesawat di bulan tertentu
3. Manajemen stok barang
4. Prediksi harga barang
5. dsb



Introduction

Bahkan hingga memahami wanita ...*

A Time-Series Analysis of my Girlfriends Mood Swings

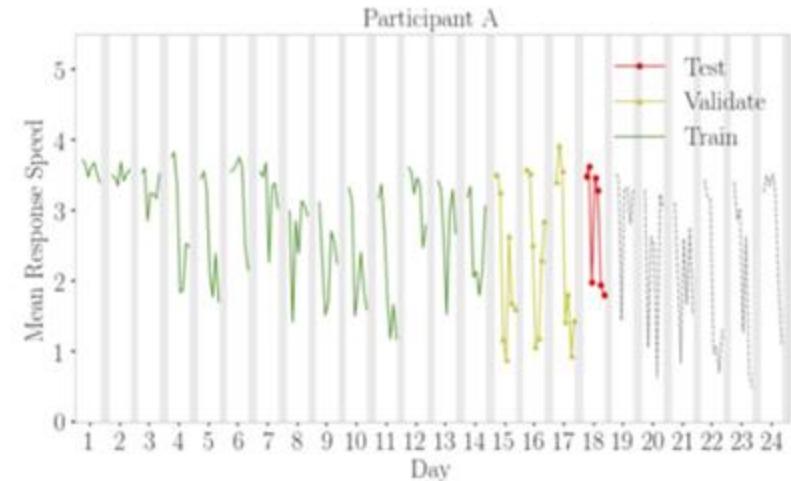
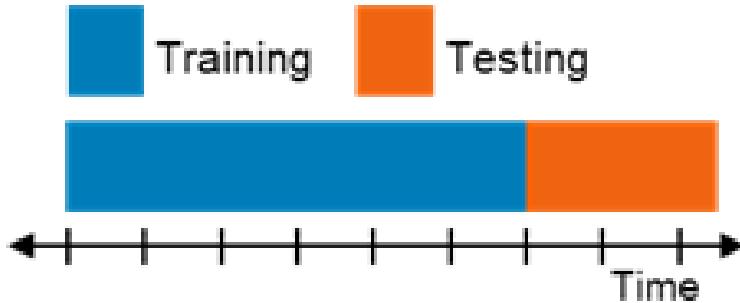
• B McGraw • May 23, 2021 ■ Articles, Computer Science
• Academic Article, Machine Learning, Modeling, Parody, Relationships, Statistics,
Time Series Analysis



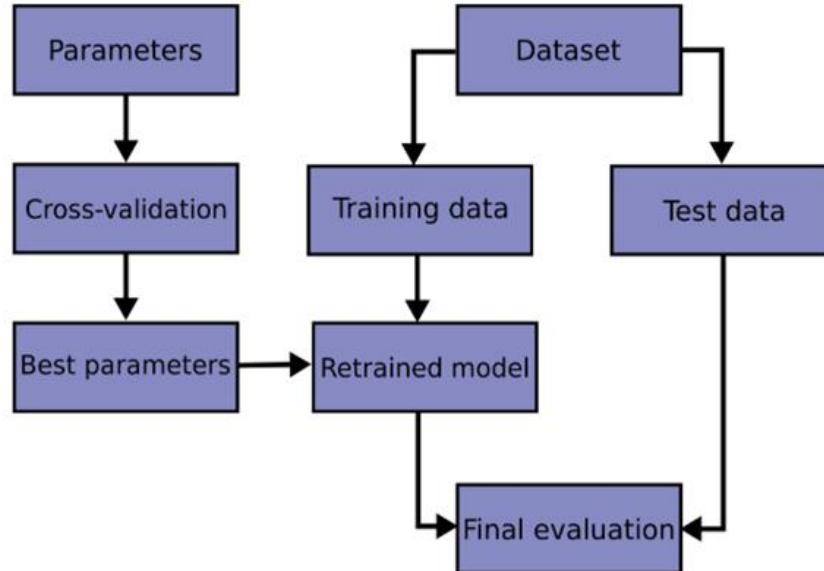
*penelitian dari Department of Applied Psychological Machine Learning, University, Pittsburgh, PA, USA (<https://jabde.com/2021/05/23/girlfriends-mood-time-series-analysis/>)

Berbeda dengan data pada umumnya, data time series tidak bisa dibagi menjadi training dan testing secara acak, karena datanya harus berurutan. Misalkan testing datanya 20% untuk data 10 tahun, maka trainingnya haruslah 8 tahun pertama, dan testingnya 2 tahun selanjutnya.

Time-based Estimation



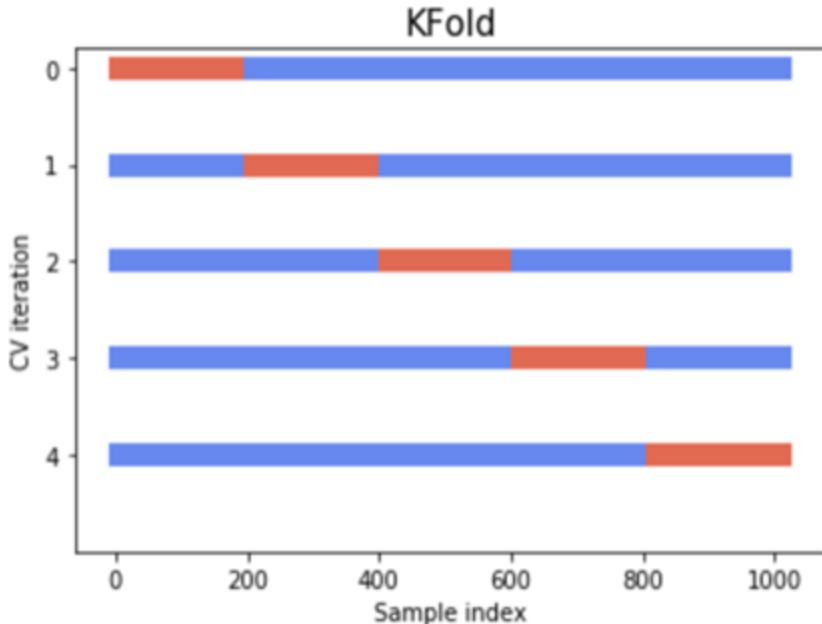
Cross Validation



- **Cross-validation (CV)** adalah metode statistik yang dapat digunakan untuk mengevaluasi kinerja model atau algoritma dimana data dibagi menjadi data training dan testing.
- **Contoh:** K-Fold Cross Validation

Sumber: S. Shrivastava, 2020 (<https://medium.com/>)

Cross Validation: K-Fold



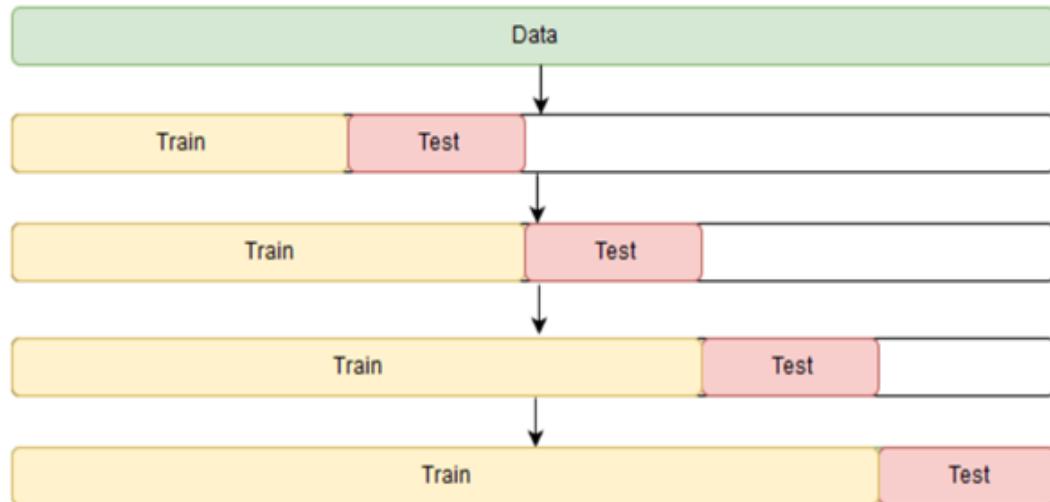
- Salah satu jenis **cross validation** yang berfungsi untuk mengevaluasi kinerja proses sebuah metode algoritma dengan membagi sampel data secara **acak** dan mengelompokkan data sebanyak nilai **K k-fold**.

Sumber: S. Shrivastava, 2020 (<https://medium.com/>)

Apa bisa diterapkan untuk Time Series?



Cross Validation



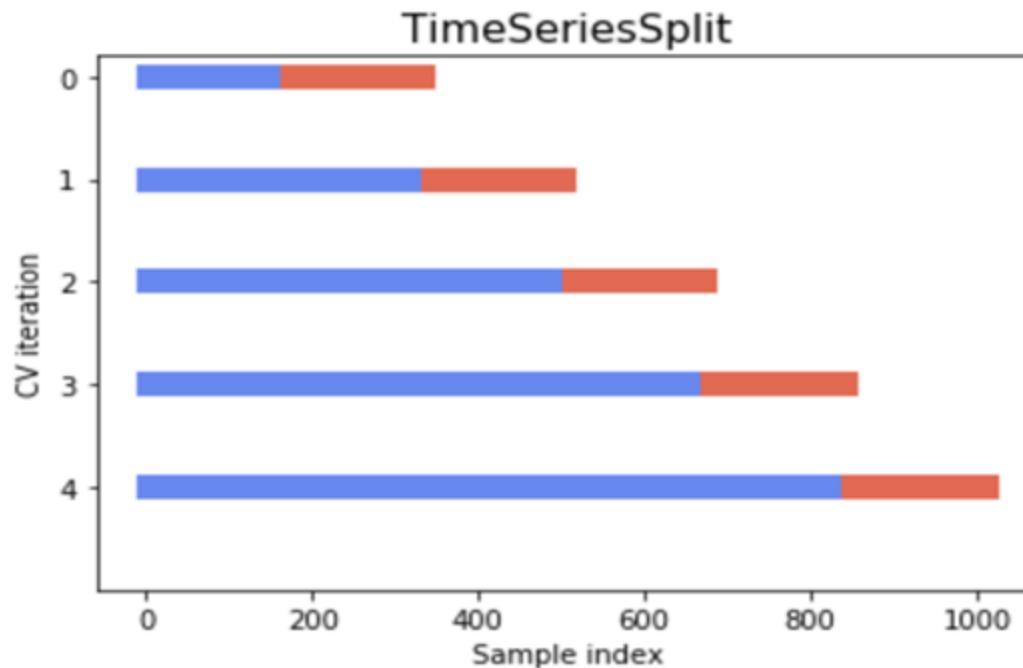
Teknik Cross Validation untuk Time Series:

1. Time Series Split Cross-Validation
2. Blocked Cross-Validation
3. Predict Second Half
4. Day Forward-Chaining

Sumber: S. Shrivastava, 2020 (<https://medium.com/>)

Cross Validation

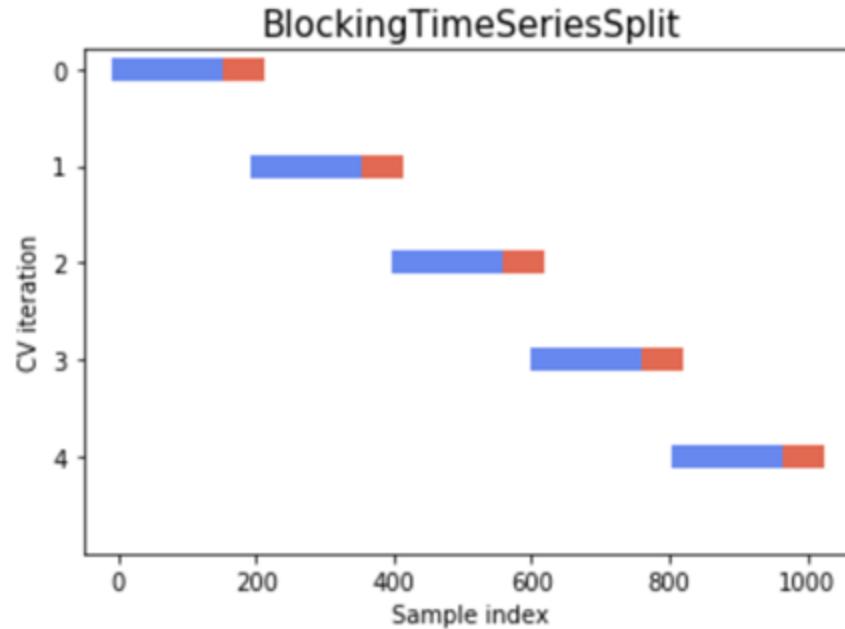
Teknik Time Series Split



Sumber: S. Shrivastava, 2020 (<https://medium.com/>)

Cross Validation

Teknik Blocked

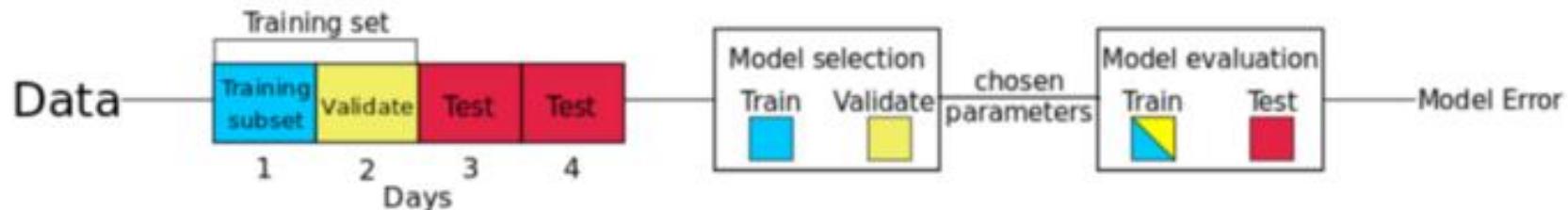


Sumber: S. Srivastava, 2020 (<https://medium.com/>)

Cross Validation

Teknik Predict Second Half

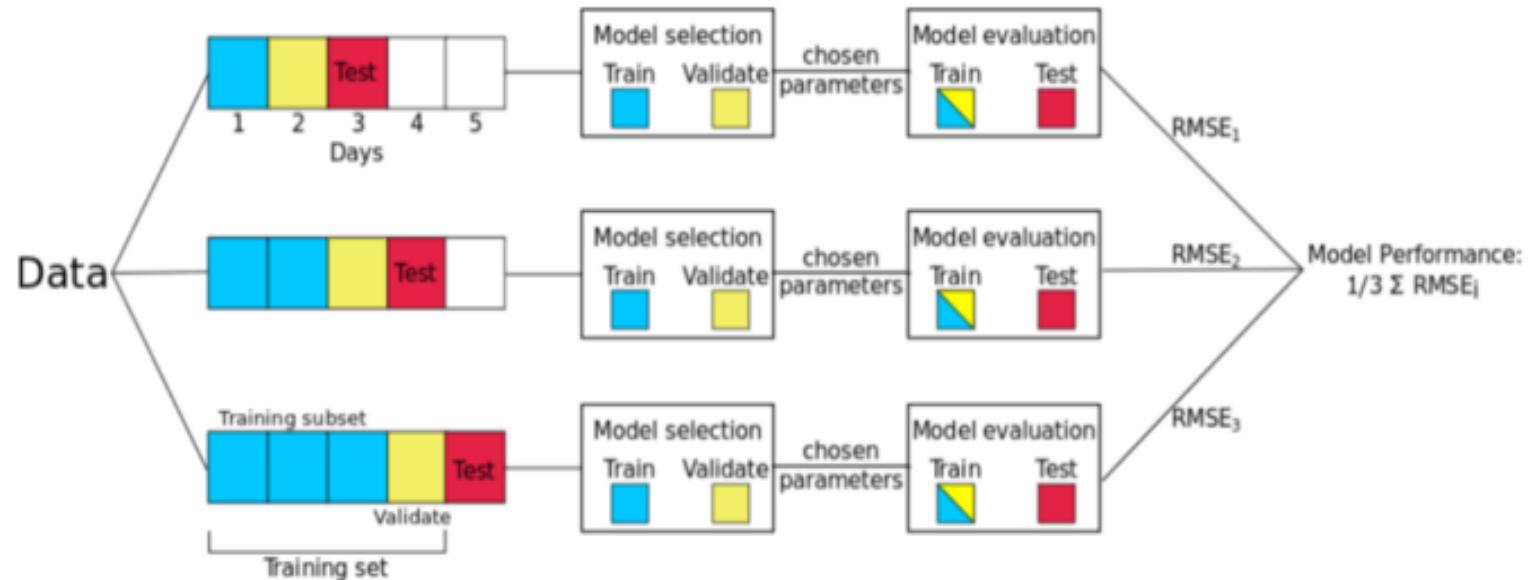
Trainingnya 25% pertama, Validationnya 25% selanjutnya, testing 50% data terakhir



Sumber: S. Shrivastava, 2020 (<https://medium.com/>)

Cross Validation

Teknik Day Forward-Chaining



Sumber: S. Shrivastava, 2020 (<https://medium.com/>)

Cross Validation

Metode	Split	(+)	(-)
KFold Time series split	k	Dapat melihat bagaimana model berubah seiring bertambahnya waktu	Kemungkinan akan memunculkan data leakage
Blocked Cross Validation	k	Mengatasi data leakage	Very computationally expensive
Predict Second Half	1	Mudah diimplementasi dan computationally inexpensive	Kemungkinan akan muncul bias
Day Forward-Chaining	k	Menghindari bias dan melihat perubahan model seiring waktu	Very computationally expensive, multiple model



02

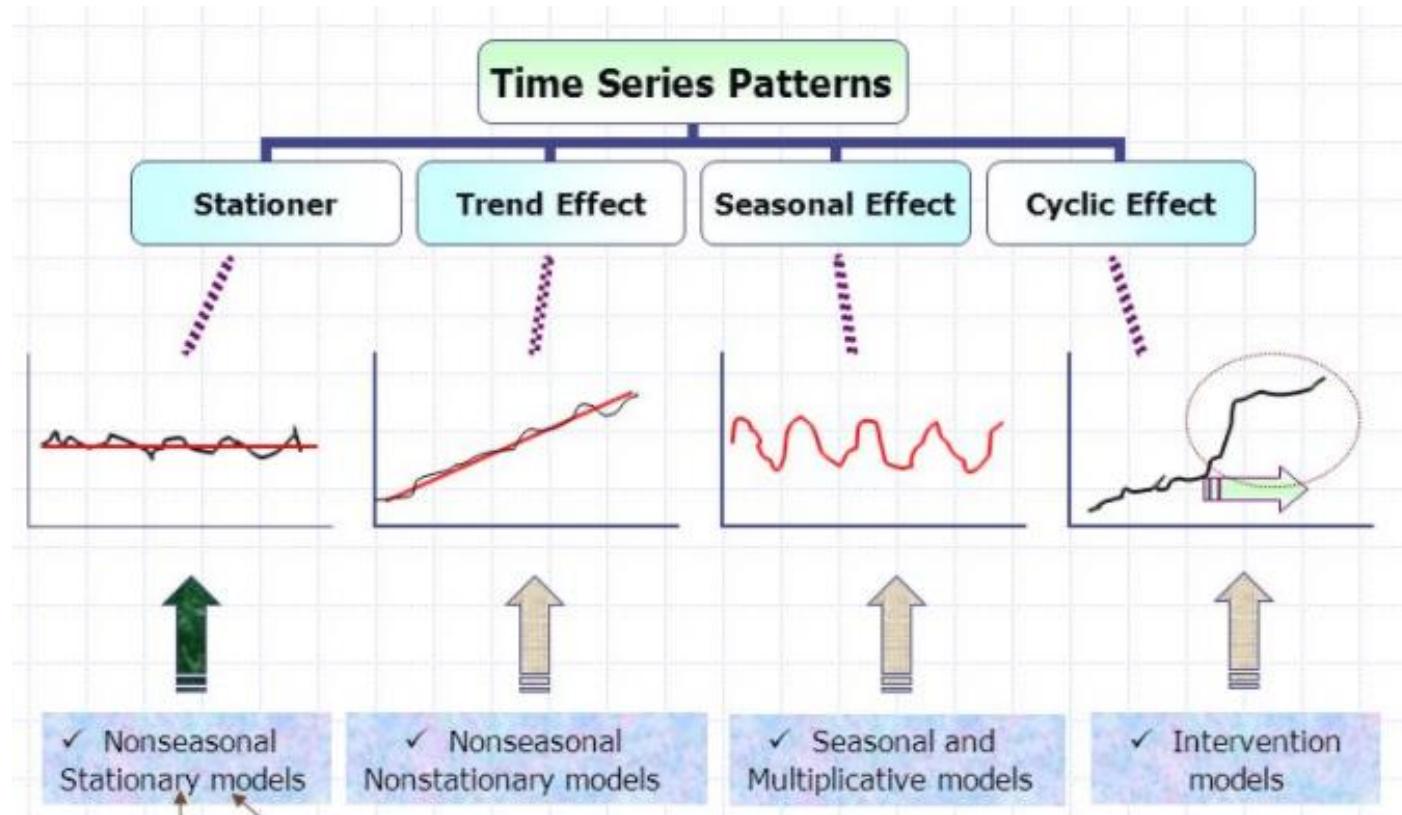
STAT MODEL

- Box Jenkins
- Seasonal
- Trend
- Kombinasi

Secara Umum, ada 2 jenis model di time series:

Statistical Approach	Machine Learning Approach
<ul style="list-style-type: none">• Metode Box-jenkins (ARIMA)• Model musiman• Model trend• Vector Autoregressive (VAR)• Hidden Markov Model (HMM)• dll	<ul style="list-style-type: none">• Recurrent Neural Network (RNN)• Long Short-Term Memory (LSTM)• Wavenet• Multi-Layer Perceptron• dll

Statistics Model



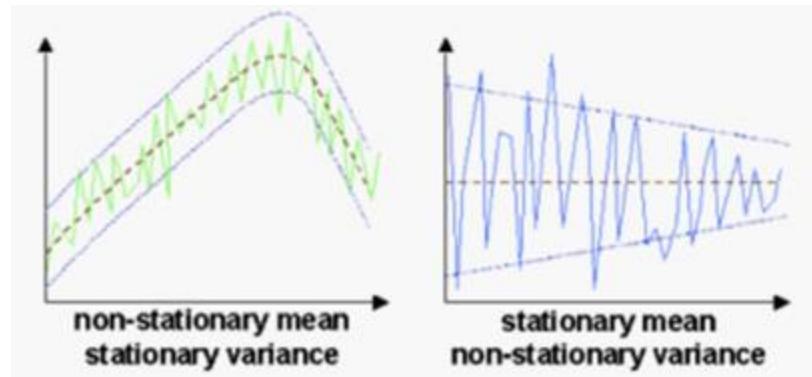
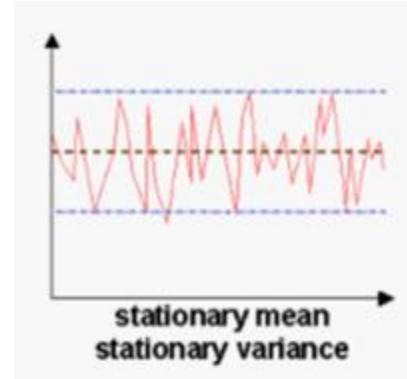
Metode ARIMA (Autoregressive Integrated Moving Average)/Box-Jenkins. Metode ini dikenalkan oleh 2 statistikawan George Box dan Gwilym Jenkins yang mengembangkan **metode pemilihan model dari melihat stasioneritasnya**. Langkah-langkah dalam metode ini :

1. Uji Stasioneritas
2. Jika tidak stasioner lakukan transformasi atau differencing
3. Uji ACF & PACF
4. Pilih model ARIMA(p,d,q) yang tepat
5. Evaluasi model dengan RMSE dll
6. Lakukan forecasting

Uji Stasioneritas

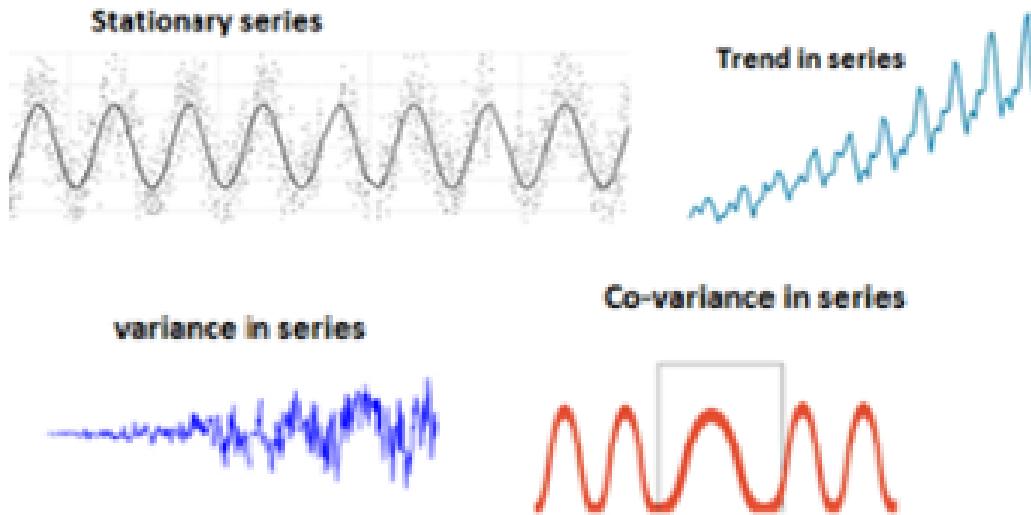
Stasioner adalah kondisi dimana mean dan varians data tidak mengalami perubahan secara sistematis. Terdapat 3 jenis uji stasioneritas:

1. Plotting line graph of the data
2. Plotting Rolling Statistics
3. Dickey-fuller Test



Uji Stasioneritas

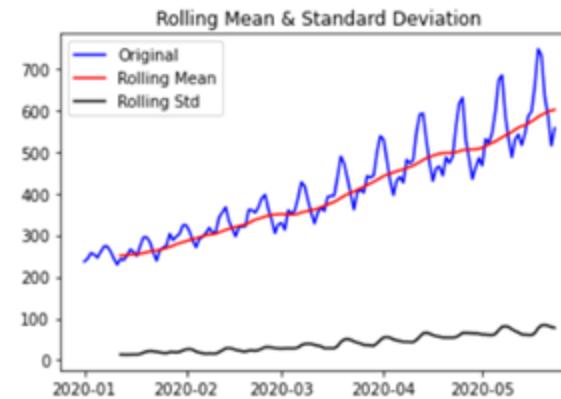
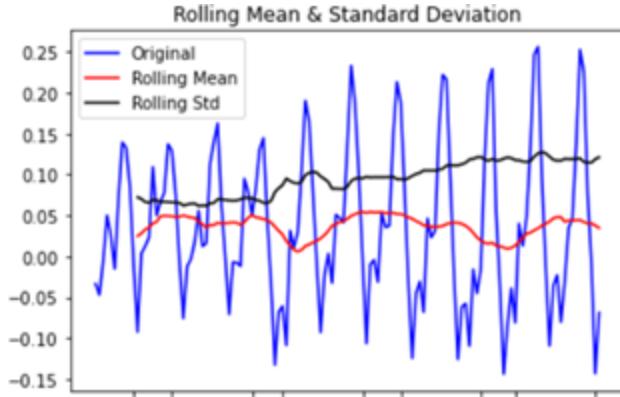
1. Plotting line graph of the data



Data yang stasioner terlihat dari grafik garisnya yang naik turun di tempat yang sama

Uji Stasioneritas

2. Plotting Rolling Statistics



Plot ini untuk menunjukkan apakah mean dan variansnya stabil atau tidak. Jika grafik rolling mean dan variansnya cenderung lurus atau tidak banyak perubahan maka dikatakan stasioner

Box-Jenkins

Uji Stasioneritas

3. Dickey-fuller Test

Tes ini menggunakan hipotesis:

H_0 : Data tidak stasioner

H_a : Data Stasioner

H_0 ditolak jika pvalue < α atau test statistics < critical value

Results of Dickey-Fuller Test:

Test Statistic	-3.234394
p-value	0.018078
#Lags Used	13.000000
Number of Observations Used	130.000000
Critical Value (1%)	-3.481682
Critical Value (5%)	-2.884042
Critical Value (10%)	-2.578770

Terlihat pvalue 0.018 < **0.05**, dan test statistics = -3.23 < -2.88. Sehingga dengan kepercayaan **95%** data ini stasioner

Jika Data Tidak Stasioner

Lakukan ini:

1. Transformasi data (misalnya semua data dihitung jadi $\log(\text{data})$)
2. Differencing, differencing adalah mengubah data menjadi:

$$d^{(1)}(t) = x(t) - x(t - 1) \quad \rightarrow \text{difference 1 kali}$$

$$d^{(2)}(t) = d^{(1)}(t) - d^{(1)}(t - 1) \quad \rightarrow \text{difference 2 kali}$$

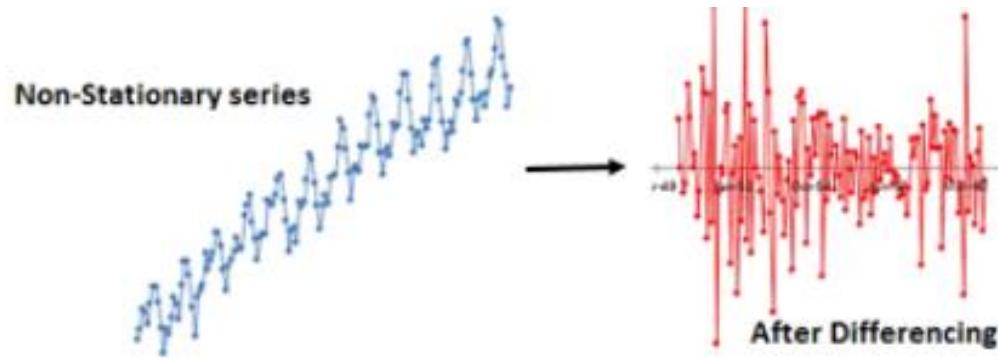
$$d^{(m)}(t) = d^{(m-1)}(t) - d^{(m-1)}(t - 1) \quad \rightarrow \text{difference m kali}$$

Setelah transformasi dan differencing, lakukan uji stasioner lagi

Jika Data Tidak Stasioner

Setelah transformasi dan differencing, lakukan uji stasioner lagi.

Jika 1x differencing belum stasioner, coba lakukan yang kedua dst



Uji ACF & PACF

- AutoCorrelation Function (ACF) mengukur korelasi antara x_t dan x_{t-h} . ACF pada lag ke-h dihitung dengan rumus :

$$\rho(h) = \frac{\text{Covariance}(x_t, x_{t-h})}{\text{Variance}(x_t)}$$

- Partial AutoCorrelation Function (PACF) mengukur korelasi parsial antara x_t dan x_{t-h} . PACF pada lag ke-1 dihitung sama dengan ACF lag ke-1 : $\Phi(1) = \rho(1)$.

Uji ACF & PACF

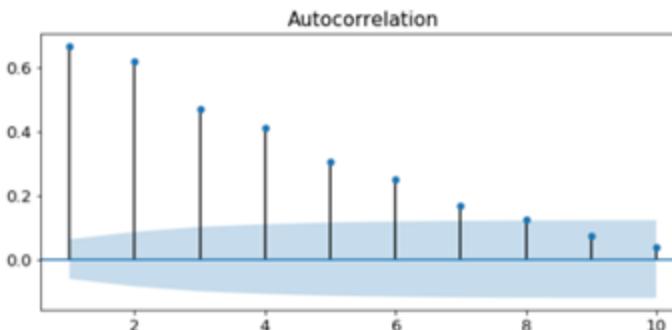
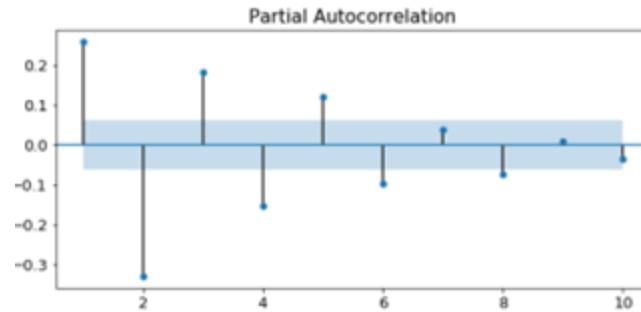
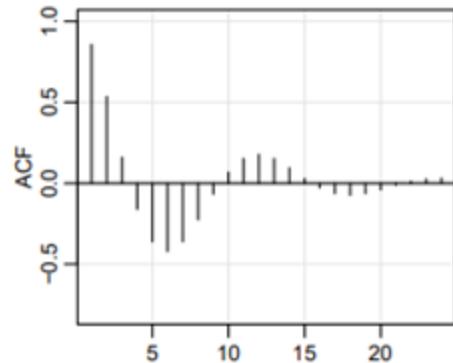
$$\Phi(2) = \frac{\text{Covariance}(x_t, x_{t-2}|x_{t-1})}{\sqrt{\text{Variance}(x_t|x_{t-1})\text{Variance}(x_{t-2}|x_{t-1})}}$$

$$\Phi(3) = \frac{\text{Covariance}(x_t, x_{t-3}|x_{t-1}, x_{t-2})}{\sqrt{\text{Variance}(x_t|x_{t-1}, x_{t-2})\text{Variance}(x_{t-3}|x_{t-1}, x_{t-2})}}$$

Begitu seterusnya, PACF untuk lag ke-h ditulis $\Phi(h)$.

Uji ACF & PACF dilakukan dengan membuat plot. Hasil plotnya antara 2 :
Tails off/ Dies down (menyusut) atau **Cuts off (muncul tiang pancang)**
pada lag ke-k

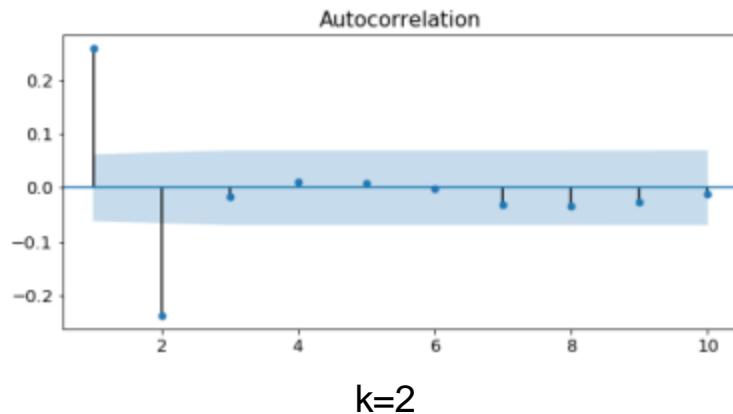
Contoh hasil yang tails off / Dies Down (menyusut)



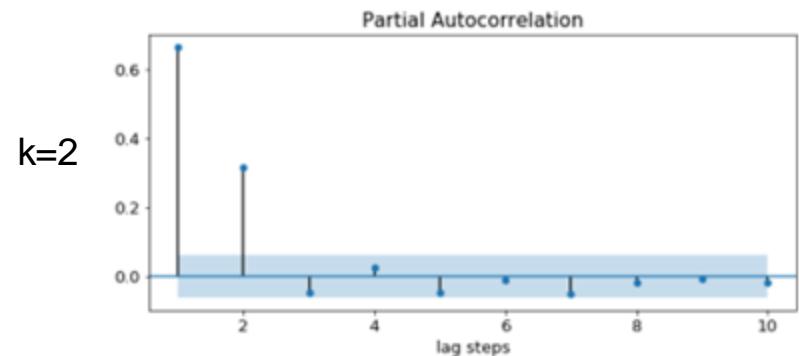
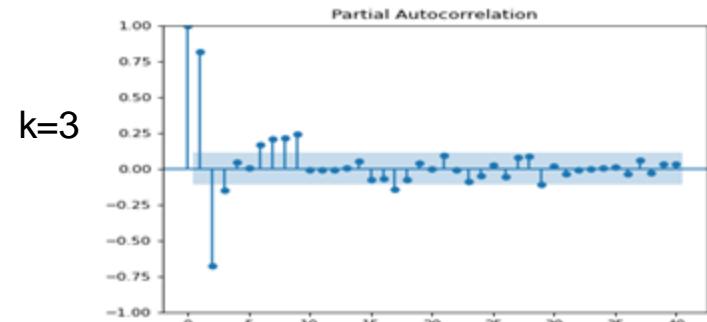
Grafik agak lama masuk ke wilayah signifikansi

Box-Jenkins

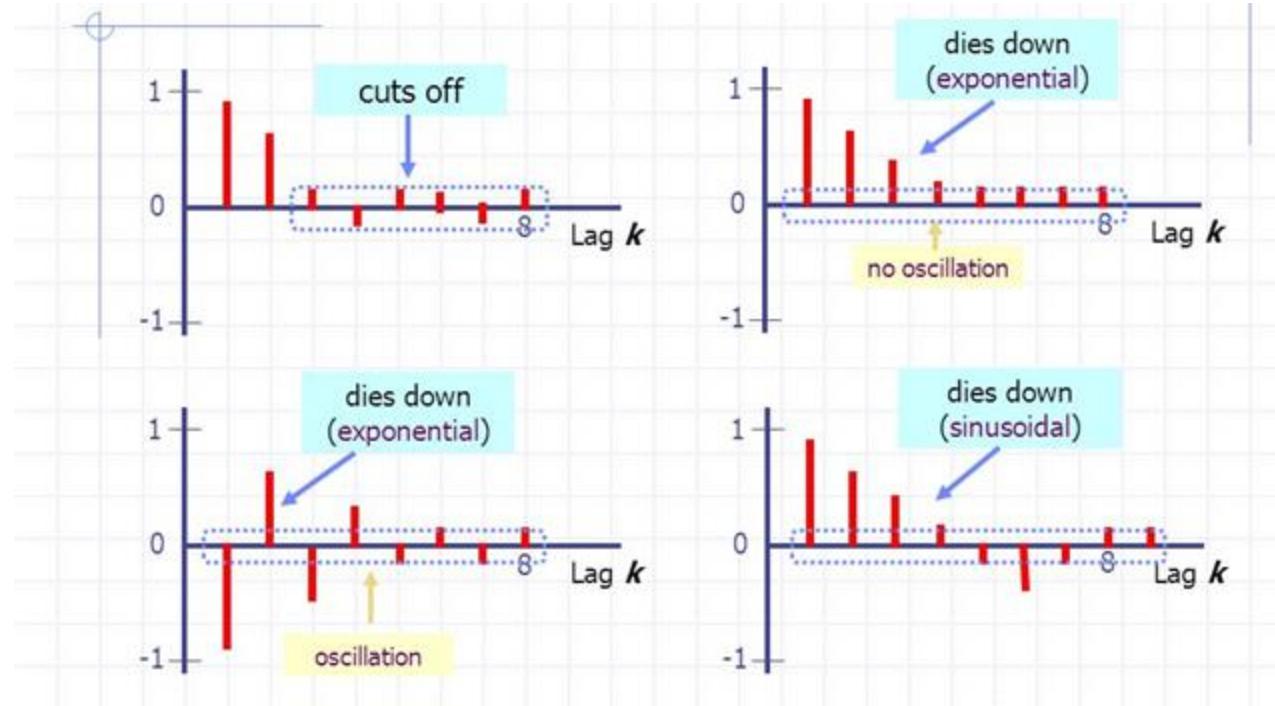
Contoh hasil yang Cuts off (ada tiang pancang)

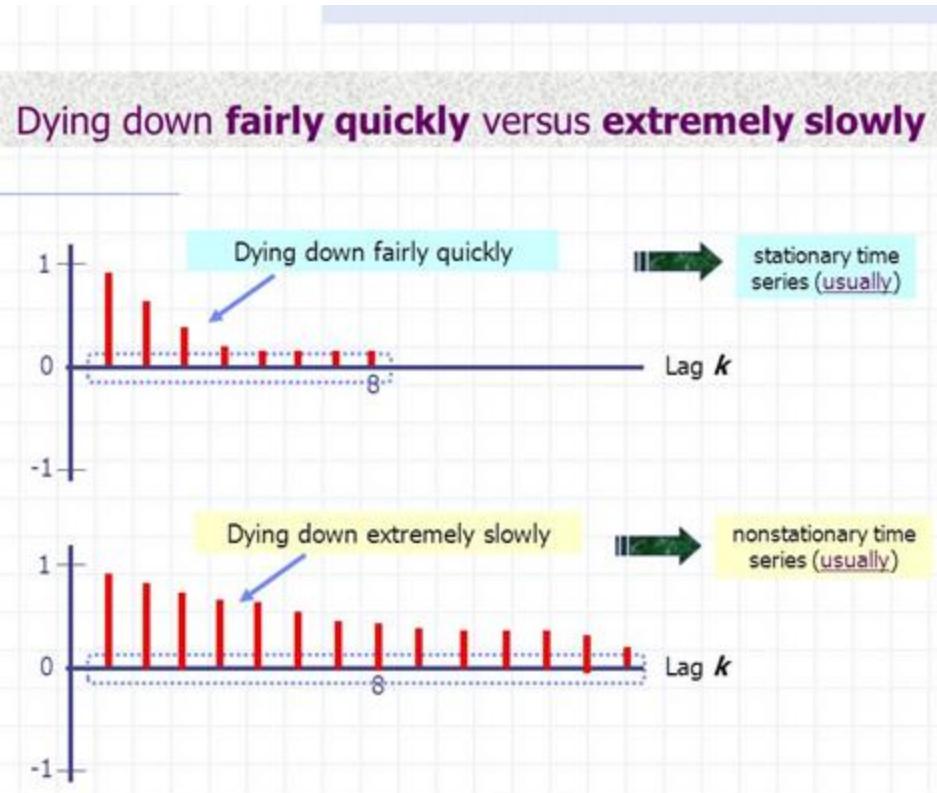


Grafik masuk area signifikansi pada lag awal2



Box-Jenkins





Model ARIMA

AR(p)

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + w_t,$$

w=white noise

$$w_t \sim wn(0, \sigma_w^2),$$

MA(q)

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \cdots + \theta_q w_{t-q},$$

ARMA(p, q)

$$x_t = \phi_1 x_{t-1} + \cdots + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q},$$

Untuk **ARIMA (p,d=m,q)**, substitusikan $d^{(m)}(t) = d^{(m-1)}(t) - d^{(m-1)}(t-1)$ ke x_t

Box-Jenkins

Pemilihan Model

	ACF	PACF
AR(p)	Dies Down	Cuts off after lag - p
MA(q)	Cuts off after lag - q	Dies Down
ARMA	Dies Down	Dies Down

ARIMA(p, d, q) :

p = cuts offnya AR

d = berapa kali differencing

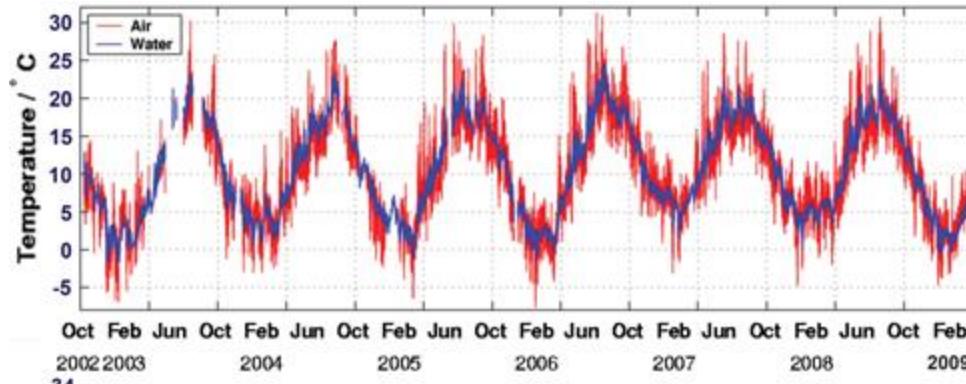
q = cuts offnya MA

Jadi misalnya ARIMA(1,0,0) itu sama aja dengan AR(1)

ARIMA (1,1,1) artinya differencing 1 kali, model gabungan AR(1) dan MA(1)

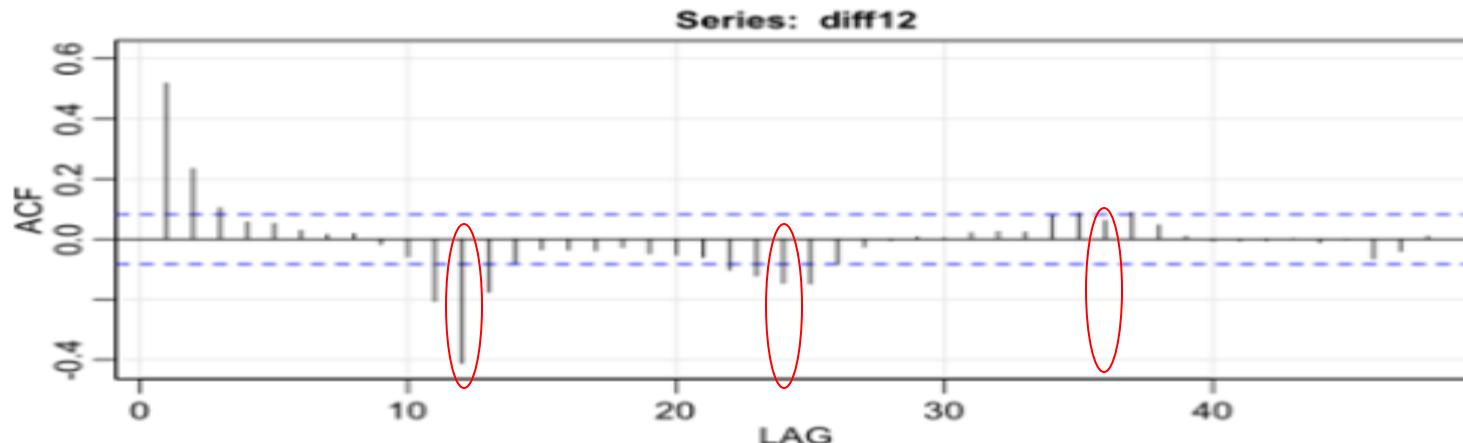
Seasonal ARIMA (SARIMA)

Jika plot data terlihat berulang setiap kurun waktu tertentu, maka bisa jadi ini adalah data musiman. Misalnya data rata-rata temperatur harian di suatu wilayah, akan berulang setiap 12 bulan. Model yang bisa dibentuk adalah kombinasi ARIMA dan Musiman (Seasonal) yang disebut **SARIMA**.



Seasonal ARIMA (SARIMA)

Data musiman juga terlihat dari kurva ACF dan PACF yang mencapai titik maksimum/minimum di tiap periode yang sama. Seperti contoh berikut yang berulang tiap 12 periode



Model Trend

Data trend terlihat dari kurvanya yang konsisten naik atau konsisten turun seperti berikut:



Model trend dapat diperoleh dengan regresi linier dengan waktu sebagai feature atau mengkombinasikan dengan SARIMA/ ARIMA

Kombinasi

Secara umum ada 2 cara kombinasi model :

Additive Decomposition

Mengkombinasikan model ARIMA, trend, dan seasonal dengan menjumlahkan :

$$y = \text{base} + \text{trend} + \text{seasonality} + \text{residual}$$

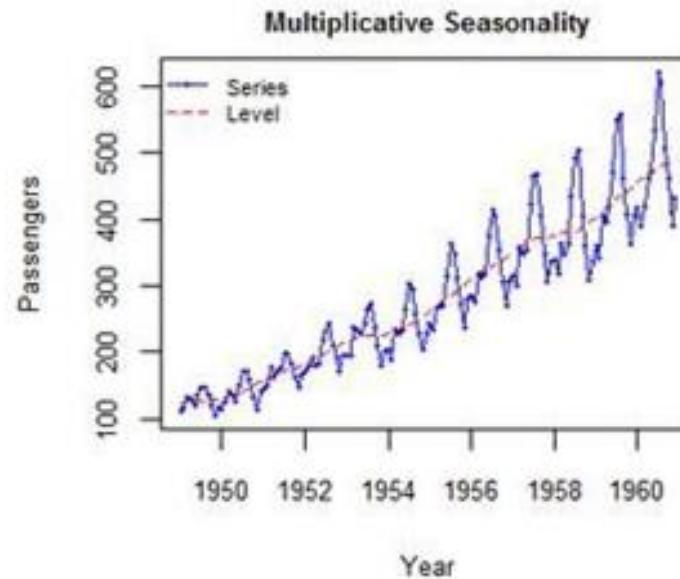
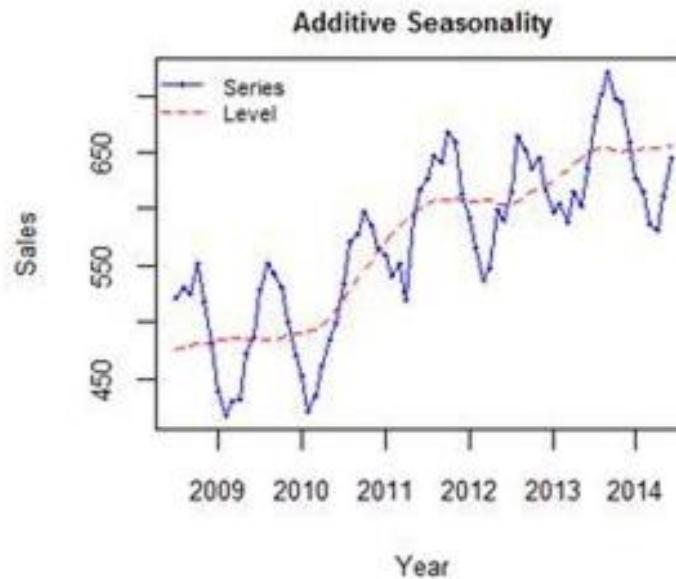
Multiplicative Decomposition

Mengkombinasikan model ARIMA, trend, dan seasonal dengan mengalikan :

$$y = \text{base} \times \text{trend} \times \text{seasonality} \times \text{residual}$$

Kombinasi

Contoh perbedaannya :





03 EVALUATION

- Evaluation Metrics

Evaluation Metrics

Beberapa metrics untuk mengukur performance dari model time series:

1. Mean Absolute Error (MAE)

$$\frac{\sum_{i=1}^n |\hat{y}_i - y_i|}{n}$$

2. Mean Absolute Percent Error (MAPE)

$$\frac{100}{n} \times \sum_{i=1}^n \frac{|\hat{y}_i - y_i|}{y_i}$$

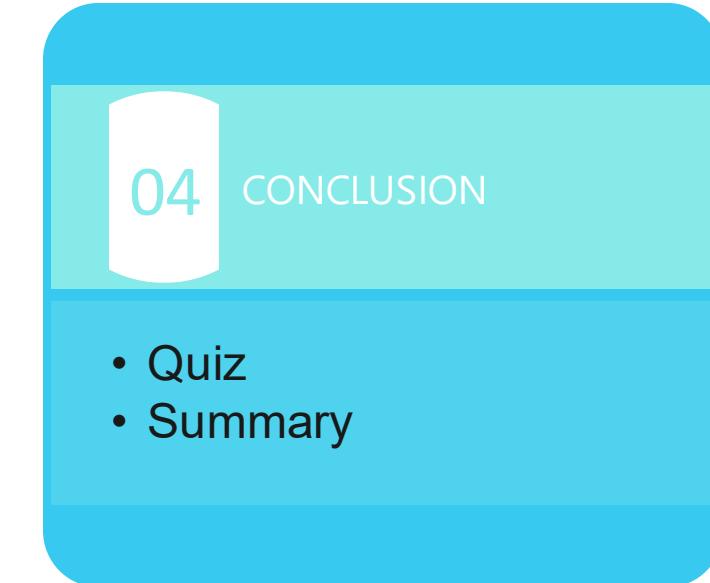
3. Mean Squared Error (MSE)

$$\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}$$

4. Root Mean Squared Error(RMSE)

$$\sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

5. Lainnya : AIC, BIC, MAD, MSD, dll



04 CONCLUSION

- Quiz
- Summary

1. Berikut bukan merupakan data time series...

- a. Data inflasi
- b. Data pasien terkonfirmasi positif covid19
- c. Data curah hujan
- d. Data sensus penduduk

1. Berikut bukan merupakan data time series...

- a. Data inflasi
- b. Data pasien terkonfirmasi positif covid19
- c. Data curah hujan
- d. Data sensus penduduk

Jawaban : d

- Data runtun waktu/time series adalah suatu himpunan data pengamatan yang dibangun dalam urutan waktu.
- Split Data time series tidak bisa dibagi secara acak (random) namun pembagian datanya harus berurutan
- Pola data time series secara umum ada 4 jenis: stasioner, trend, seasonal, dan cyclic effect.
- Analisis Box Jenkins ARIMA hanya berlaku pada data runtun waktu yang stasioner. Namun data runtun waktu yang tidak stasioner dapat ditransformasi menjadi runtun waktu yang stasioner, sehingga ARIMA juga dapat digunakan untuk data runtun waktu yang tidak stasioner.

References

1. **Time Series Analysis and Its Applications** by Robert H. Shumway & David S. Stoffer
2. **Time Series Analysis with R**, by Jonathan D. Cryer & Kung-Sik Chan
3. **Basic Econometrics**, by Damodar N. Gujarati
4. **Introduction to Modern Time Series Analysis**, Gebhard Kirchgässner & Jürgen Wolters
5. **Ekonometrika untuk Analisis Ekonomi dan Keuangan**, Fakultas Ekonomi Universitas Indonesia
6. **Diktat - Time Series Analysis**, by Siana Halim
7. <https://www.statsmodels.org/dev/tsa.html>
8. <https://medium.com/@stallonejacob/time-series-forecast-a-basic-introduction-using-python-414fcb963000>
9. <https://medium.com/purwadhikaconnect/mengenal-time-series-dan-struktur-yang-membentuknya-2e74252178c2>
10. <https://people.duke.edu/~rnau/411arim.htm>
11. <https://id1lib.org/s/time%20series%20analysis%20python>
12. <https://cmemorys.medium.com/k-fold-cross-validation-secarak-singkat-30f8e5188f46>

Let's Code!

<https://colab.research.google.com/drive/1mVBIYZWMiwu5YaoquRC96-xz0zLnNxFm?usp=sharing>



TERIMA KASIH

Orbit Future Academy

PT Orbit Ventura Indonesia
Center of Excellence (Jakarta Selatan)
Gedung Veteran RI, Lt.15
Unit Z15-002, Plaza Semanggi
Jl. Jenderal Sudirman Kav.50, Jakarta
12930, Indonesia

- Jakarta Selatan/Pusat
- Jakarta Barat/BSD
- Kota Bandung
- Kab. Bandung
- Jawa Barat

Hubungi Kami

Director of Sales & Partnership
ira@orbitventura.com
+62 858-9187-7388

Social Media

-  Orbit Future Academy
-  @OrbitFutureAcademyInd
-  OrbitFutureAcademy
-  Orbit Future Academy

AI Mastery Course

Module
Data Science
Section
Time Series
(Deep Learning Approach)





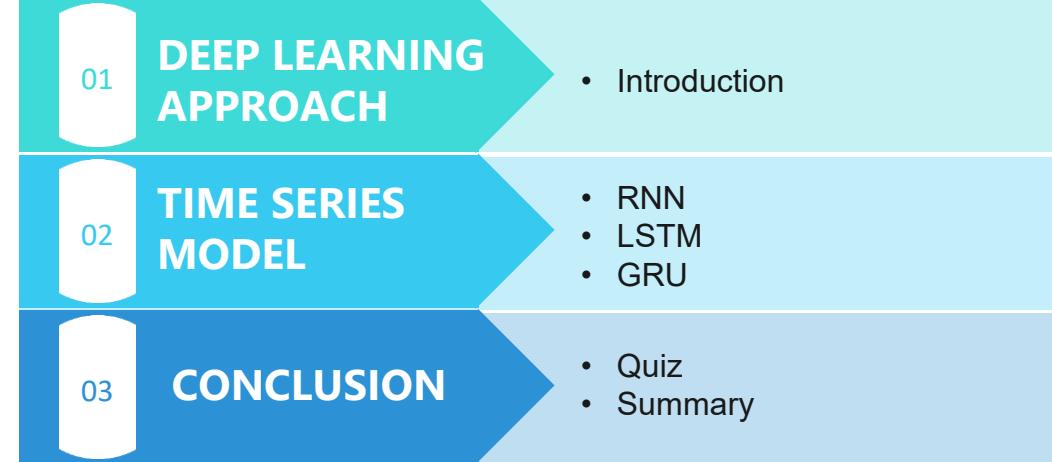
Learning Objectives

Di akhir modul ini kita diharapkan memahami:

- Arsitektur jaringan RNN, LSTM, GRU
- Dapat melakukan time series forecasting menggunakan deep learning



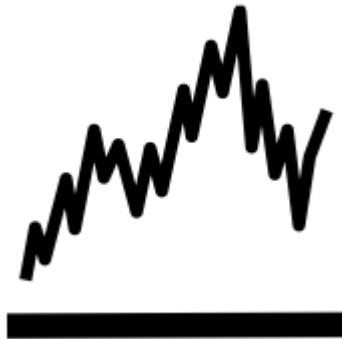
Agenda





01 DEEP LEARNING APPROACH

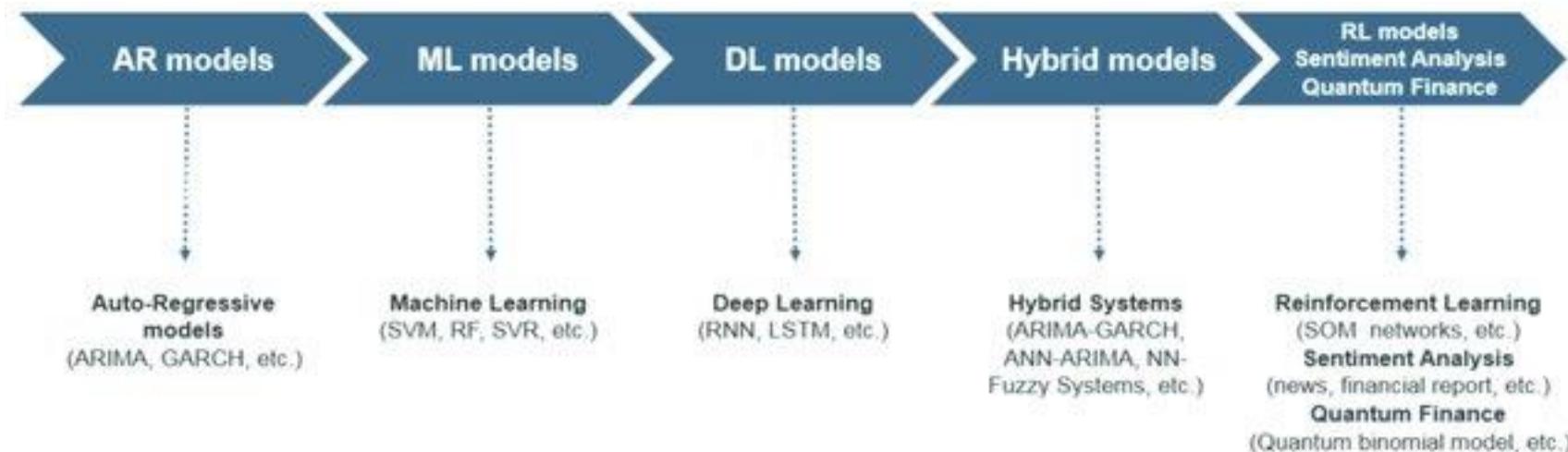
- Introduction



Time Series Forecasting selalu menjadi area penelitian yang sangat penting di banyak domain karena berbagai jenis data disimpan dalam waktu time series. Misalnya prakiraan cuaca, manajemen rantai pasokan dan peramalan harga saham, dll.

Pada model seperti model autoregressive (AR) atau exponential smoothing – rekayasa fitur dilakukan secara manual dan seringkali beberapa parameter dioptimalkan juga dengan mempertimbangkan pengetahuan domain.

Model Deep Learning hanya mempelajari fitur, dan dinamika dari data langsung (tanpa mempertimbangkan pengetahuan domain). Berkat ini, Deep learning dapat **mempercepat proses** *data preparation* dan dapat **mempelajari pola data** yang **lebih kompleks** dengan cara yang lebih lengkap.



Sumber : [10.3390/app9245574](https://doi.org/10.3390/app9245574)

Statistical Approach vs Deep Learning Approach

Statistical Approach	Deep Learning Approach
<ul style="list-style-type: none">+ Memberikan hasil yang lebih precise dengan nilai evaluasi error yang kecil- Membutuhkan pengetahuan terkait domain data saat mengaplikasikan (proses ini bisa mengambil waktu yang lama dan juga dapat terjadi kesalahan dalam mengambil asumsi)	<ul style="list-style-type: none">+ Lebih mudah diaplikasikan untuk tugas terkait time series forecasting+ Memberikan hasil yang hampir serupa dengan menggunakan statistical approach+ Mampu menangani dataset yang besar, dan juga hasilnya dapat meningkat, sedangkan statistical approach cenderung tidak mengalami peningkatan jika datasetnya ditambah+ Tidak membutuhkan pengetahuan tentang data domain

Mengapa Deep Learning?

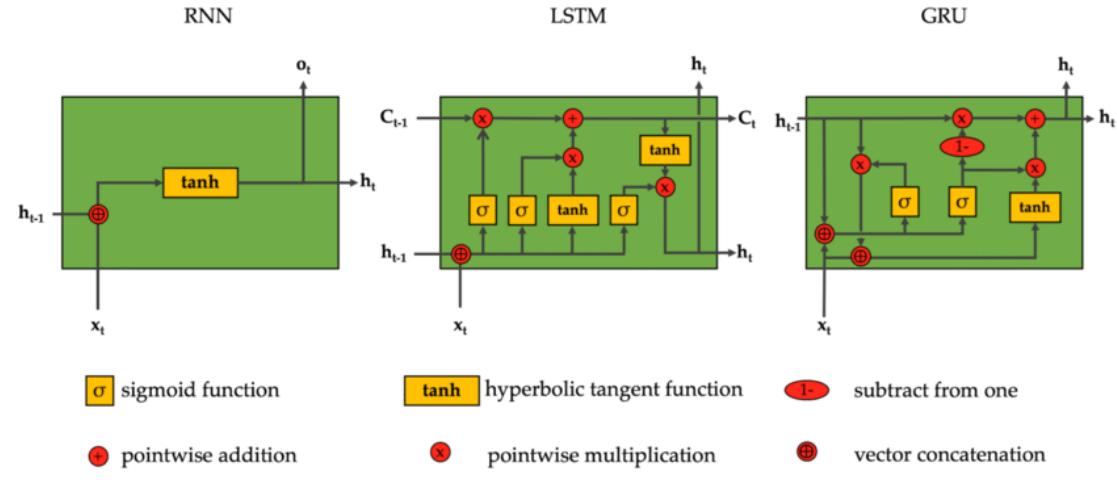
Deep Learning muncul untuk mengatasi keterbatasan model Machine Learning tradisional yang memiliki keterbatasan, berupa :

- Bila ada *missing value* dapat mempengaruhi kinerja model;
- Tidak mampu mengenali pola kompleks dalam data;
- Biasanya bekerja dengan baik hanya dalam beberapa langkah *forecasting*, dan tidak dalam *forecasting* jangka panjang.

Deep Learning Architecture

Ada beberapa arsitektur deep learning yang dapat digunakan untuk menganalisa data time series, diantaranya :

- MLP
- CNN
- RNN
- LSTM
- GRU



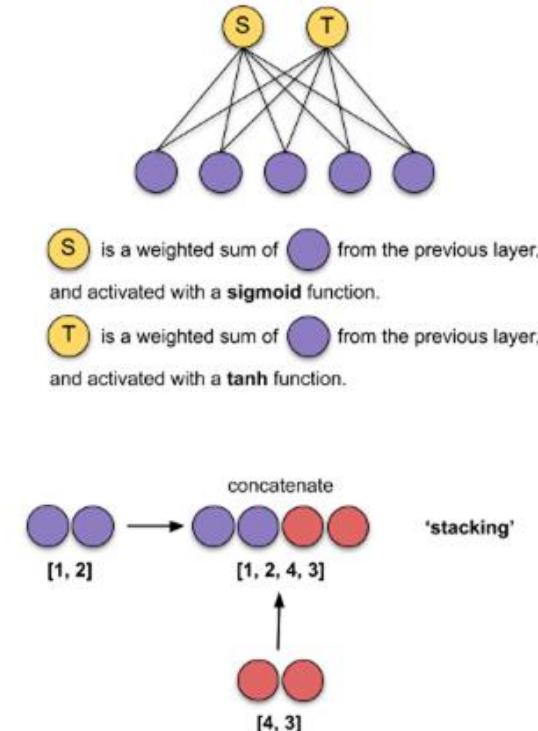
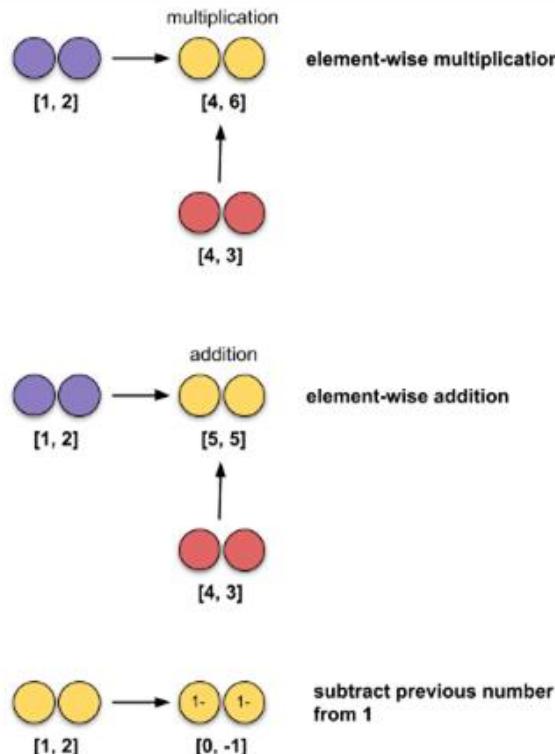
sumber : [10.31223/osf.io/xs36g](https://doi.org/10.31223/osf.io/xs36g)



02 TIME SERIES MODEL

- RNN
- LSTM
- GRU

Operasi RNN, LSTM, GRU



<https://towardsdatascience.com/animated-rnn-lstm-and-gru-ef124d06cf45>

Operasi Element Wise Product



Element Wise Product atau Hadamard product, entrywise product atau Schur product, adalah operasi perkalian dengan input 2 matriks berukuran sama dengan hasil operasinya menghasilkan matriks berukuran sama. Secara matematis dapat dituliskan sebagai :

$$(A \circ B)_{ij} = (A \odot B)_{ij} = (A)_{ij}(B)_{ij}.$$

Nama operasi ini diberi nama dari matematikawan Perancis Jacques Hadamard atau matematikawan Jerman Issai Schur. Hasil kali Hadamard bersifat asosiatif dan distributif. Berbeda dengan perkalian matriks pada umumnya Hadamard Product bersifat **komutatif**.

$$\begin{array}{c|c|c} \begin{matrix} \textcolor{red}{\square} & \textcolor{red}{\square} & \dots & \textcolor{red}{\square} \\ \vdots & \vdots & & \vdots \\ \textcolor{red}{\square} & \textcolor{red}{\square} & \dots & \textcolor{red}{\square} \end{matrix} & \circ & \begin{matrix} \textcolor{green}{\square} & \textcolor{green}{\square} & \dots & \textcolor{green}{\square} \\ \vdots & \vdots & & \vdots \\ \textcolor{green}{\square} & \textcolor{green}{\square} & \dots & \textcolor{green}{\square} \end{matrix} \\ m & m & m \end{array} = \begin{matrix} \textcolor{blue}{\square} & \textcolor{blue}{\square} & \dots & \textcolor{blue}{\square} \\ \vdots & \vdots & & \vdots \\ \textcolor{blue}{\square} & \textcolor{blue}{\square} & \dots & \textcolor{blue}{\square} \end{matrix} \quad A \quad \circ \quad B \quad = \quad C$$

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \circ \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & a_{13}b_{13} \\ a_{21}b_{21} & a_{22}b_{22} & a_{23}b_{23} \\ a_{31}b_{31} & a_{32}b_{32} & a_{33}b_{33} \end{bmatrix}.$$

[https://en.wikipedia.org/wiki/Hadamard_product_\(matrices\)](https://en.wikipedia.org/wiki/Hadamard_product_(matrices))

Operasi Element Wise Addition

Element wise addition disebut juga matrix addition adalah operasi penjumlahan pada matriks. Pada paper, dan sering digunakan dalam *machine learning*. operasi ini sering menggunakan simbol \oplus .

$$\begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} + \begin{bmatrix} j & m & p \\ k & n & q \\ l & o & r \end{bmatrix}$$

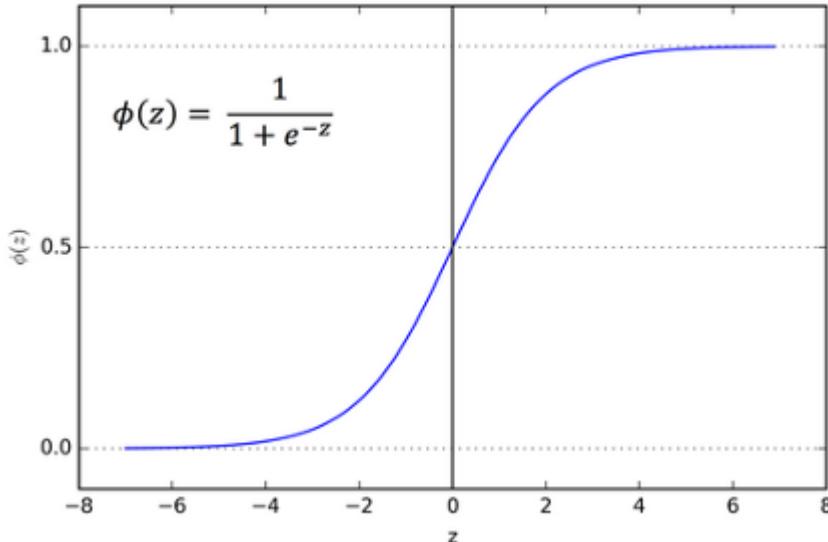
$$\begin{bmatrix} a+j & d+m & g+p \\ b+k & e+n & h+q \\ c+l & f+o & i+r \end{bmatrix}$$

$$[\mu, \sigma] = \text{FC}(\text{Conv}(I) \oplus \sum_i^{N(i)} M_{t_1:t_{obs}}^i), \quad z \sim \mathcal{N}(\mu, \sigma),$$

where \oplus denotes element-wise addition. The latent variable z enables modal predictions by going into the LSTM decoder with the last hidden state h_{obs} during observation. During the forecasting phase, the predicted next

<https://www.tutorialspoint.com;element-wise-addition-explained-a-beginner-guide-machine-learning-tutorial/>

Sigmoid Activation Function

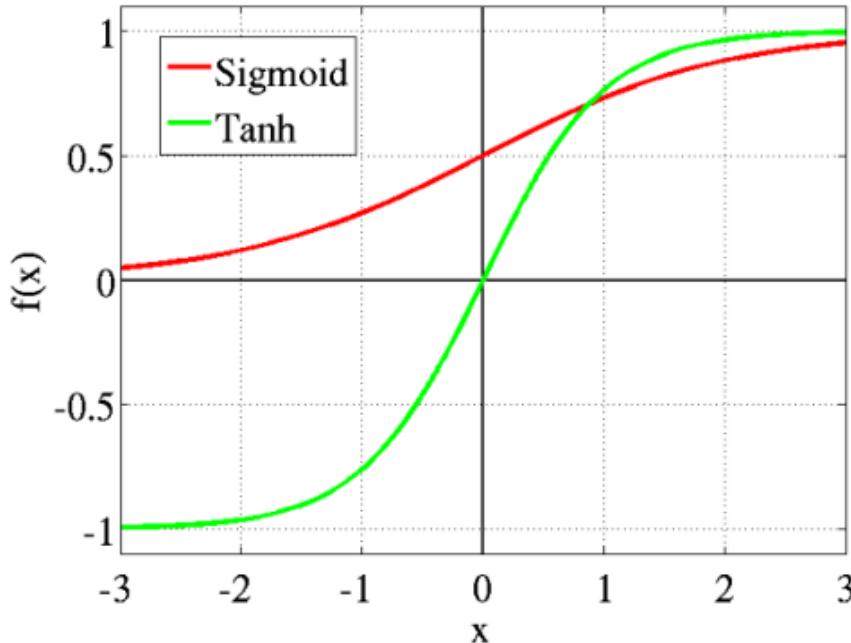


Sigmoid akan menerima angka tunggal dan mengubah nilai x menjadi sebuah nilai yang memiliki range mulai dari 0 sampai 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

<https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>

Tanh Activation Function



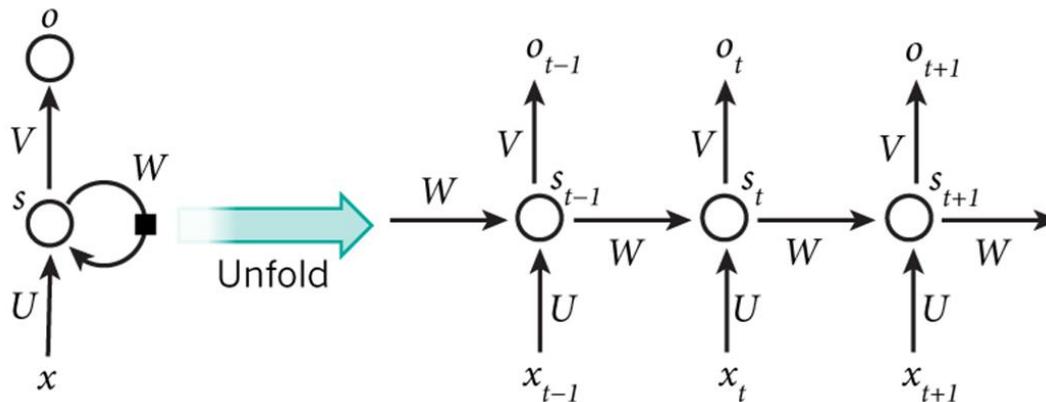
Tanh akan mengubah nilai input x nya menjadi sebuah nilai yang memiliki range mulai dari -1 sampai 1.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

<https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>

Recurrent Neural Network

Recurrent neural network (RNN) yang juga disebut jaringan umpan balik adalah jenis jaringan pada neural network dimana terdapat loop sebagai koneksi umpan balik dalam jaringan (Lin & Lee, 1996: 340). RNN memiliki tiga lapisan yaitu layer input, layer tersembunyi yang berulang, dan layer output (Salehinejad et al., 2017).

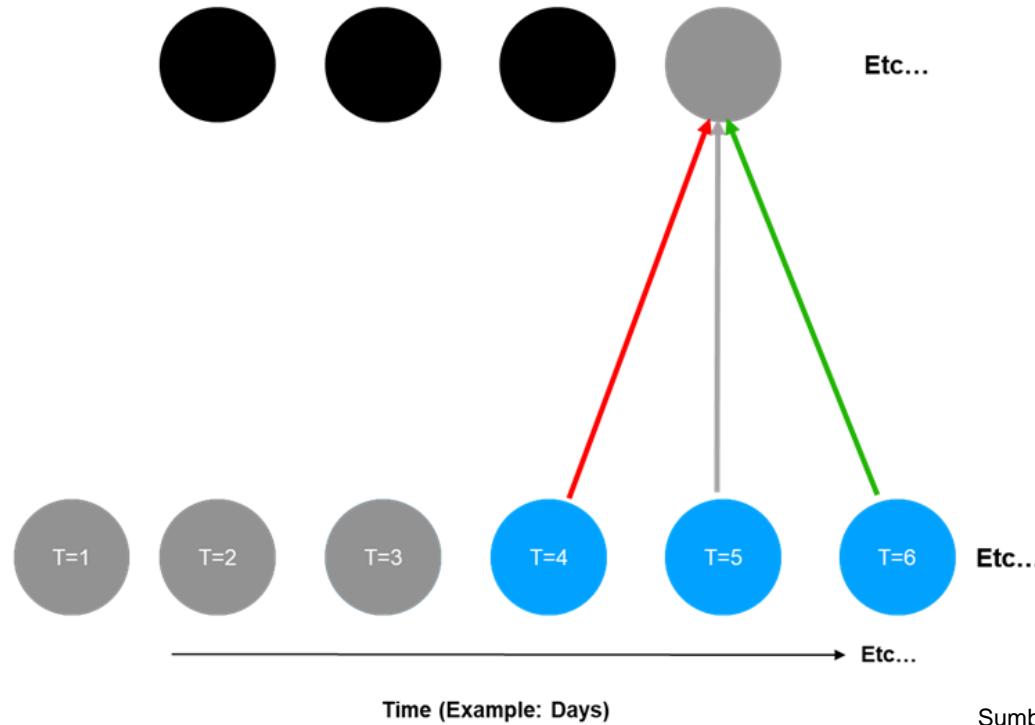


Sumber Gambar : Nvidia Deep Learning Institute

Recurrent Neural Network

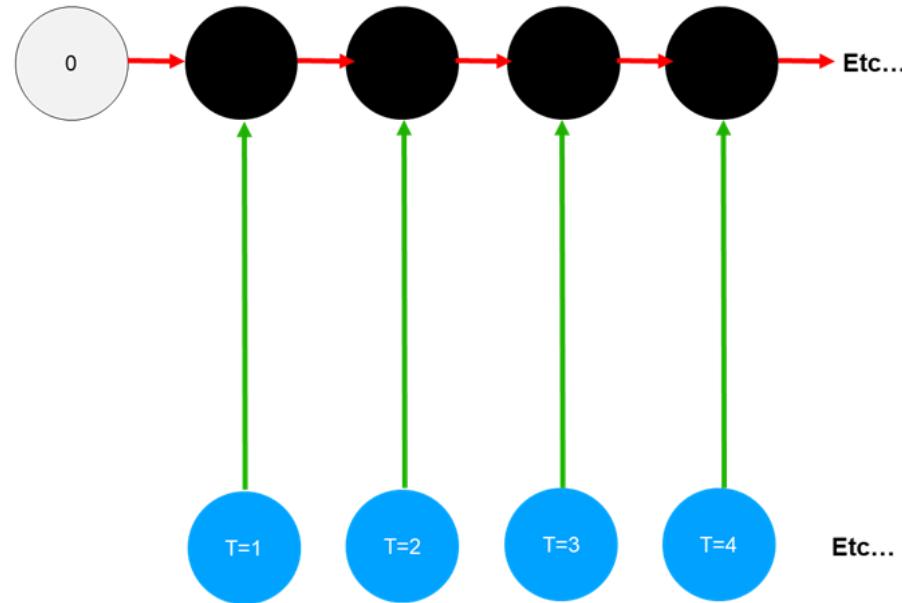
- **Layer input** memiliki unit input, terkoneksi penuh ke unit tersembunyi yang ada di layer tersembunyi.
- Unit tersembunyi itu terhubung satu sama lain secara berulang.
- **Layer tersembunyi** bisa didefinisikan sebagai “memori” atau ruang keadaan yang berdimensi tinggi dengan dinamika non-linier untuk mengingat dan memproses informasi masa lalu.
- Keadaan tersembunyi akan merangkum semua informasi unik yang diperlukan sebagai keadaan terakhir dari jaringan, melalui serangkaian langkah waktu.
- Informasi itu lalu terintegrasi, sehingga mampu menentukan perilaku jaringan di masa depan dan melakukan prediksi yang akurat di **layer output**.

1D CNN:



Sumber Gambar : Nvidia Deep Learning Institute

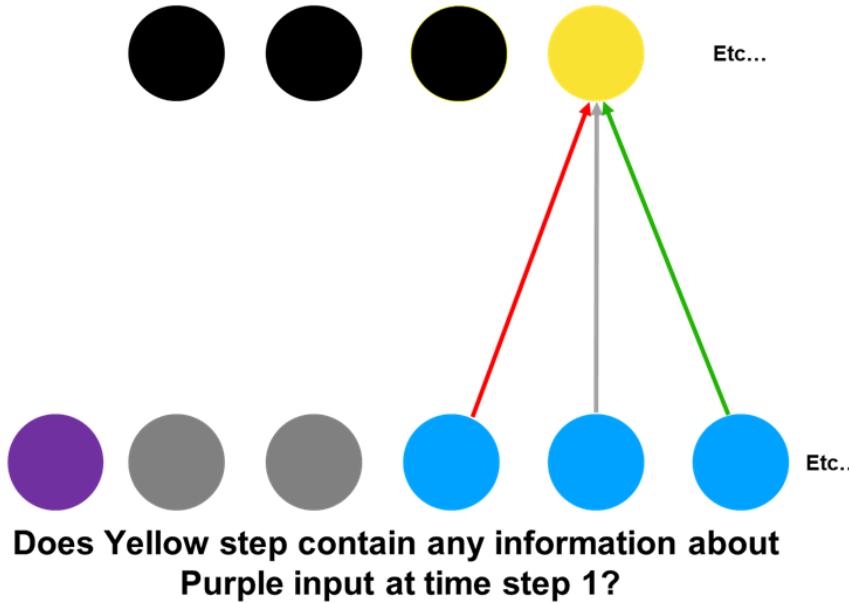
RNN Unrolling:



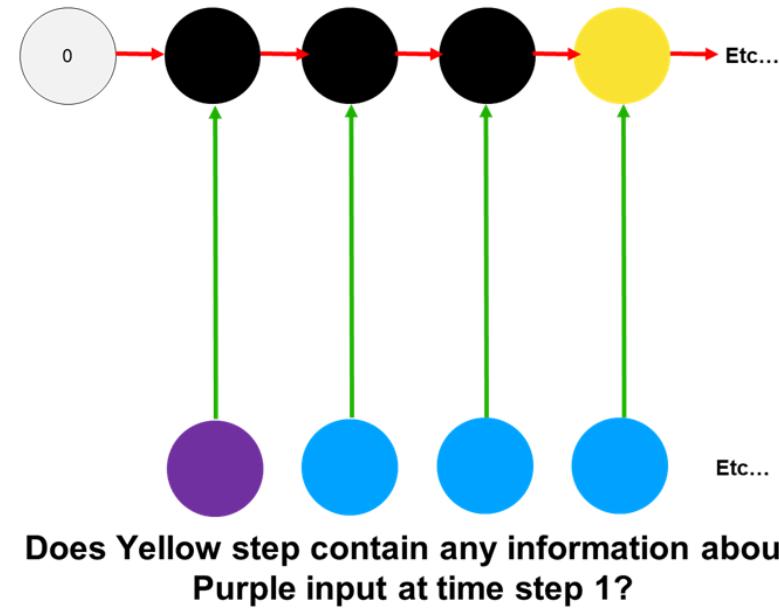
So what is different?

Sumber Gambar : Nvidia Deep Learning Institute

1D CNN:



RNN By Contrast:



Sumber Gambar : Nvidia Deep Learning Institute

CNN vs RNN

Data 1

X_1	X_2	X_3	X_4	y
1	-32	12	4	21
12	45	14	3	8
-4	2	-13	54	3
17	-6	-22	24	31
-10	23	64	15	37

Sampel

Data 2

X_1	X_2	X_3	y
[1,2,3]	[10,20,30]	[5,10,15]	64
[3,4,5]	[100,110,120]	[45,50,55]	196
[7,8,9]	[70,80,90]	[10,15,20]	135
[13,14,15]	[60,70,80]	[15,20,25]	136
[21,22,23]	[30,40,50]	[5,10,15]	104

Sampel

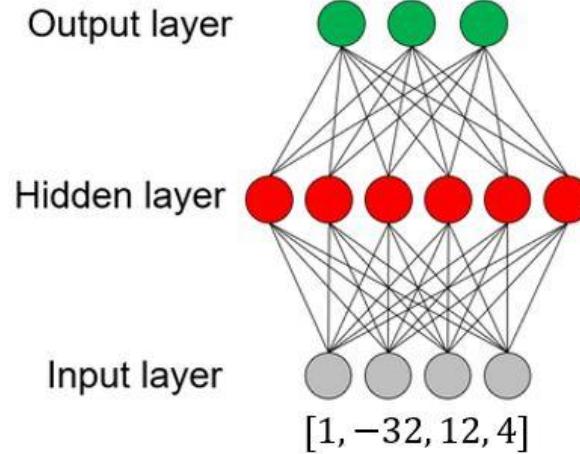
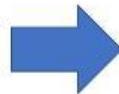
Data Sekuensial

Contoh data FFNN (CNN) dan RNN

Convolution Neural Network

Data 1

<i>input</i>	<i>output</i>
[1, -32, 12, 4]	21
[12, 45, 14, 3]	8
[-4, 2, -13, 54]	3
[17, -6, -22, 24]	31
[-10, 23, 64, 15]	37



Input berdimensi 2
 $[sample, feature] = [5, 4]$

Setiap sampelnya merupakan vektor yang terdiri dari beberapa variabel. vektor tersebut langsung diumpulkan ke neuron pada layer berikutnya sebagai kombinasi linear dan dimasukan ke fungsi aktivasi.

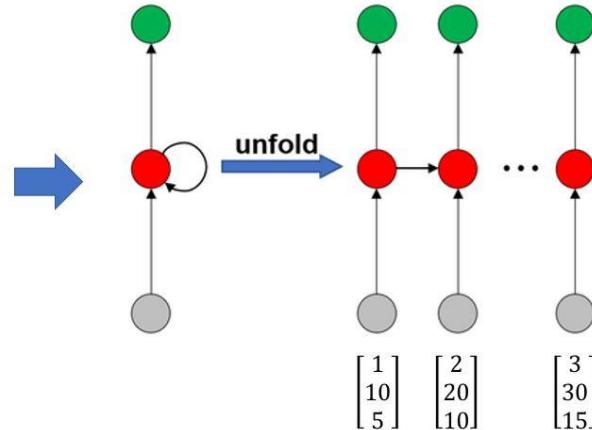
Sumber Gambar : taudata.id

Recurrent Neural Network

Data 2

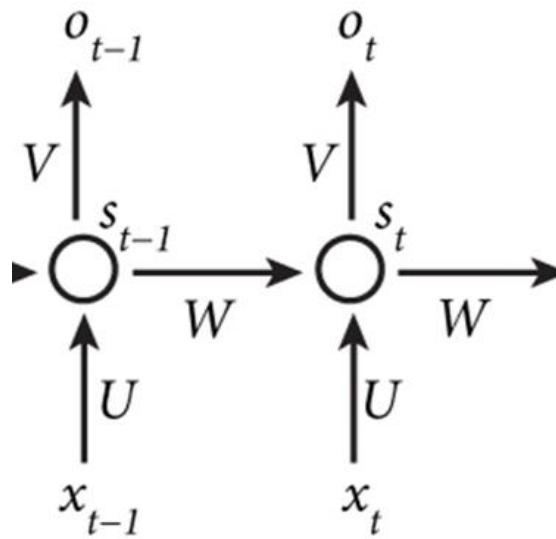
input	output
$\begin{bmatrix} 1 \\ 10 \\ 5 \end{bmatrix} \quad \begin{bmatrix} 2 \\ 20 \\ 10 \end{bmatrix} \quad \begin{bmatrix} 3 \\ 30 \\ 15 \end{bmatrix}$	64
$\begin{bmatrix} 3 \\ 100 \\ 45 \end{bmatrix} \quad \begin{bmatrix} 4 \\ 110 \\ 50 \end{bmatrix} \quad \begin{bmatrix} 5 \\ 120 \\ 55 \end{bmatrix}$	196
$\begin{bmatrix} 7 \\ 70 \\ 10 \end{bmatrix} \quad \begin{bmatrix} 8 \\ 80 \\ 15 \end{bmatrix} \quad \begin{bmatrix} 9 \\ 90 \\ 20 \end{bmatrix}$	135
$\begin{bmatrix} 13 \\ 60 \\ 15 \end{bmatrix} \quad \begin{bmatrix} 14 \\ 70 \\ 20 \end{bmatrix} \quad \begin{bmatrix} 15 \\ 80 \\ 25 \end{bmatrix}$	136
$\begin{bmatrix} 21 \\ 30 \\ 5 \end{bmatrix} \quad \begin{bmatrix} 22 \\ 40 \\ 10 \end{bmatrix} \quad \begin{bmatrix} 23 \\ 50 \\ 15 \end{bmatrix}$	104

Input berdimensi 3
[sample, time step, feature] = [5,3,3]



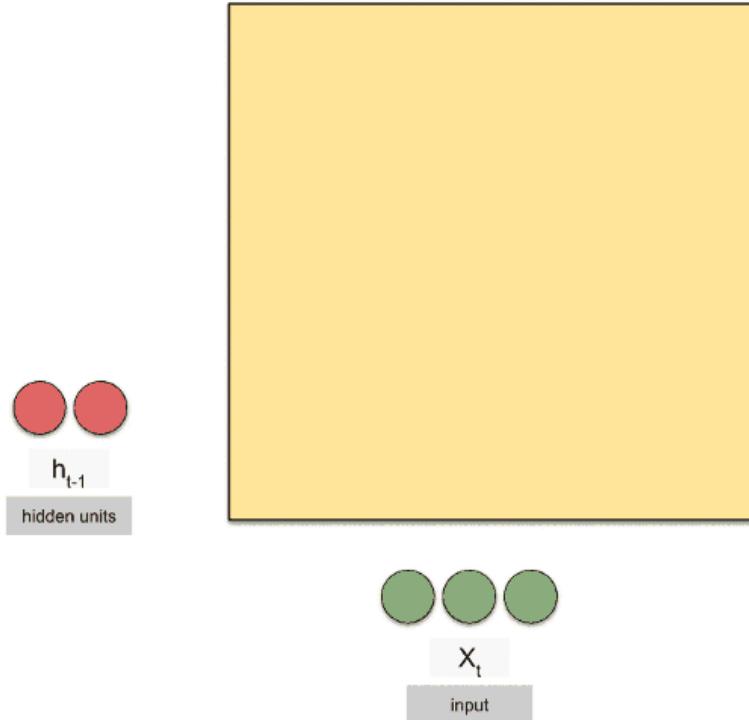
Setiap sampelnya terdiri dari data sekuensial yaitu data terurut yang tiap urutannya merupakan vektor yang terdiri dari beberapa variabel. vektor-vektor tersebut diumpulkan ke tiap-tiap blok memori secara berurutan.

Recurrent Neural Network



- x_t** → input pada time-t
 U → coefficient matrix untuk x_t
 W → coefficient matrix untuk s_{t-1}
 s_t → hidden state pada time-t, dimana :
 $s_t = f(Ux_t + Ws_{t-1} + b)$,
 f biasanya fungsi non-linear seperti tanh atau ReLU,
 b = vektor konstanta
 V → coefficient matrix untuk s_t
 o_t → output state pada time-t, hasil activation function, misal :
 $o_t = \text{Softmax}(Vs_t + c)$,
 c = vektor konstanta

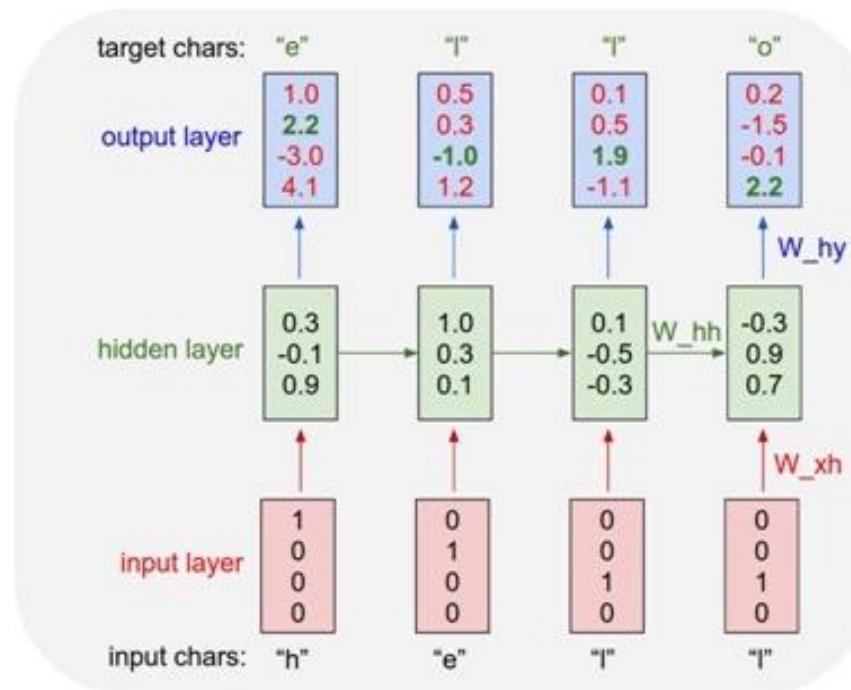
Recurrent Neural Network



Recurrent Neural Network

contoh di domain NLP

input berupa vektor
tiap huruf



nilai tertinggi keluar sebagai hasil

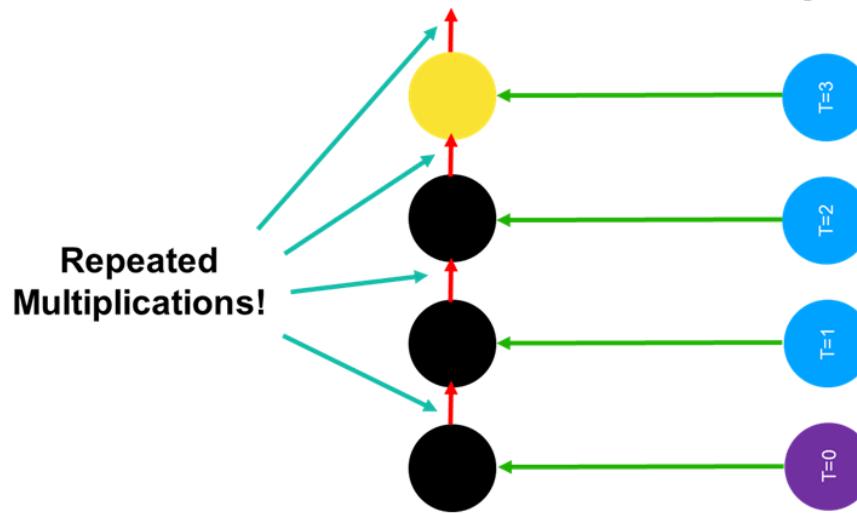
Sumber Gambar : <https://socs.binus.ac.id/2017/02/13/rnn-dan-gru/>

Recurrent Neural Network

Arsitektur dari RNN mirip dengan model Feed-Forward Network yang selama ini kita kenal. Tetapi ada beberapa perbedaan, yakni :

- Pada RNN output state sebelumnya juga diikutkan, dan cocok untuk memproses data sekuensial.
- Ada ***vanishing gradient***, sehingga terbatas hanya bisa melihat beberapa step sebelumnya.

The Problem with Depth



“The Exploding Gradient Problem”
 $2 \times 2 \times 2 \times 2 \times 2 \times \dots =$ too large for computer to work with

Sumber Gambar : Nvidia Deep Learning Institute

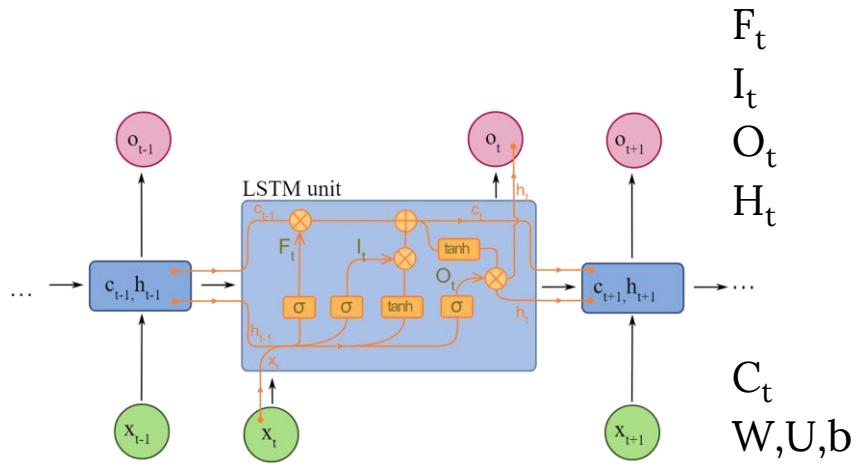
Gradient Descent Problem

Salah satu kelemahan RNN adalah pembelajaran jangka panjang dengan gradient descent bisa menghasilkan masalah menghilang atau meledaknya gradien (Salehinejad et al., 2017).

- nilai gradient descent bisa “menghilang” bila memilih bobot yang lebih kecil dari 1 (< 1) dikenal dengan **vanishing gradient problem**.
- nilai gradient descent bisa “meledak” secara eksponensial bila memilih bobot lebih besar dari 1 (> 1)

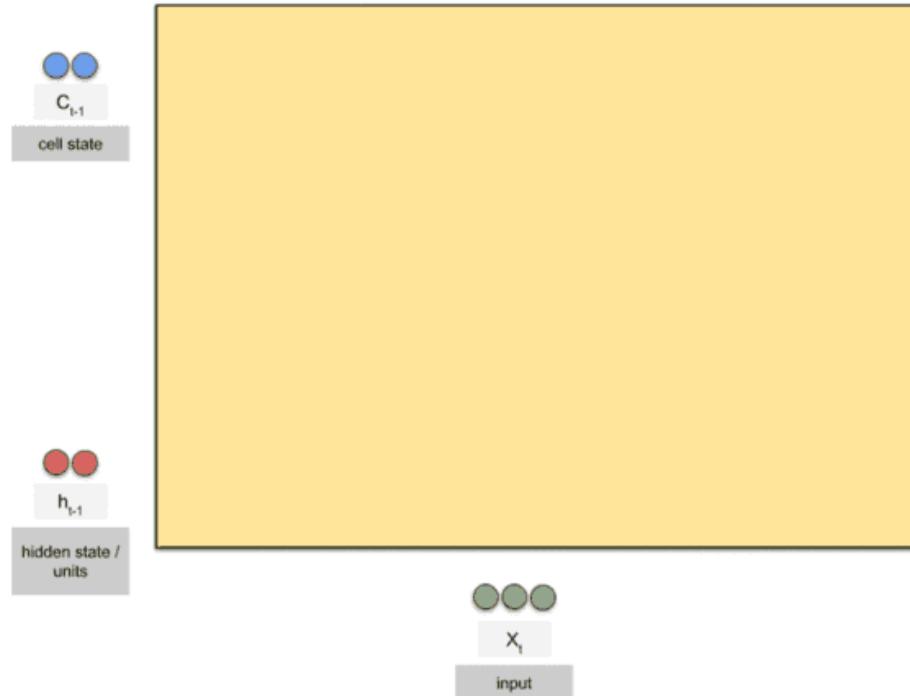
Salah satu cara mengatasi masalah menghilang atau meledaknya gradien adalah memodifikasi arsitektur model dengan memasukkan unit gerbang yang dirancang khusus untuk menyimpan informasi selama waktu periode yang lama. Mekanisme gerbang yang paling dikenal saat ini adalah Long-Short Term Memory (LSTM) dan Gated Recurrent Unit (GRU).

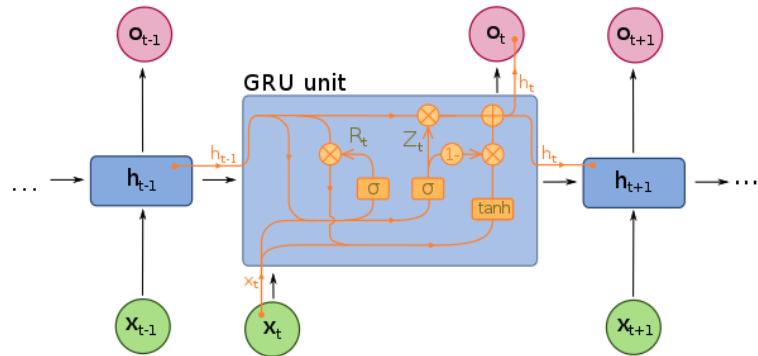
- LSTM mampu menangani penghafalan dan pengingatan kembali untuk **jangka panjang**, khususnya **data yang sangat besar**. LSTM pada prinsipnya dapat menggunakan unit memorinya untuk mengingat informasi yang jaraknya jauh dan melacak berbagai atribut teks yang sedang diproses pada penerapan NLP. (Karpathy et al., 2016).
- GRU memiliki **parameter yang lebih sedikit** dari LSTM, sehingga **cocok untuk data yang sedikit**, agar tidak terjadi overfitting. Selain itu, GRU memberikan konvergensi yang **lebih cepat** dan hasilnya **bisa disandingkan** dengan LSTM (Chung et al., 2014)



- vektor aktivasi *forget gate*
- vektor aktivasi *Input gate*
- vektor aktivasi *output gate*
- H_t sebagai vektor hidden state atau dikenal sebagai vektor keluaran dari unit LSTM
- vektor cell state,
- matriks bobot dan parameter vektor bias yang perlu dipelajari selama pelatihan.

LSTM



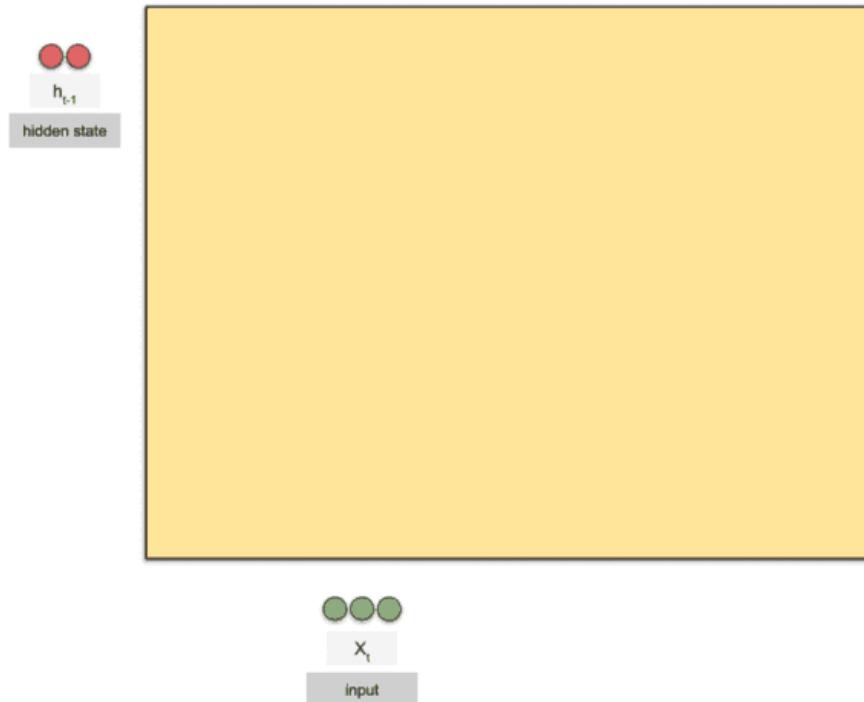


x_t
 Z_t
 R_t
 O_t
gate
 H_t
 W, U, b

- input vektor ke GRU
- vektor update gate
- vektor reset gate
- vektor aktivasi output

- vektor hidden state
- matriks bobot dan parameter vektor bias yang perlu dipelajari selama pelatihan.

Recurrent Neural Network



Deep Learning for Forecasting

Misalkan kita memiliki data time series sebagai berikut:

- Untuk melakukan forecasting data tersebut, hal yang perlu dilakukan adalah menjadikan data tersebut menjadi data sub sekuensial menggunakan sliding window.
- Sliding Window: Misalkan window size = w , data ke t diprediksi dengan melihat **w data sebelumnya**.

Sliding Window: $w = 3$

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
-------	-------	-------	-------	-------	-------	-------	-------	-------	----------

x	y

timestamp (t)	values (x_t)
1	13
2	11
3	14
4	16
5	17
6	15
7	18
8	20
9	21
10	19

Deep Learning for Forecasting

Apabila kita memilih window size = 5, setelah dilakukan sliding window pada data time series yang diperlihatkan di awal, maka akan kita dapatkan data sebagai berikut:

<i>timestamp (t)</i>	<i>values (x_t)</i>
1	13
2	11
3	14
4	16
5	17
6	15
7	18
8	20
9	21
10	19



<i>X</i>	<i>y</i>
[13,11,14,16,17]	15
[11,14,16,17,15]	18
[14,16,17,15,18]	20
[16,17,15,18,20]	21
[17,15,18,20,21]	19

<https://tau-data.id/lstm/>

Evaluation Metrics

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

\hat{y} – predicted value of y
 \bar{y} – mean value of y



03 CONCLUSION

- Summary

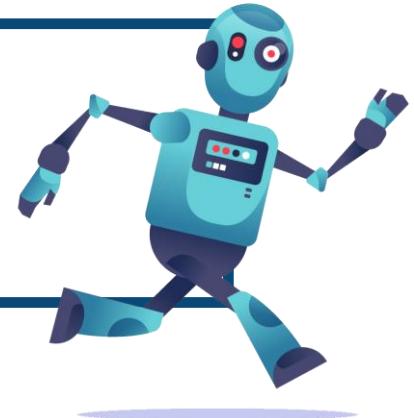
- **Metode Deep Learning yang biasa digunakan untuk time series forecasting adalah RNN, LSTM, dan GRU**
- **Ketiga metode memiliki arsitektur model yang berbeda, sehingga kebaikan model yang dihasilkan akan berbeda.**
- **Untuk evaluasi menggunakan teknik yang sama dengan metode pendekatan statistik.**

References

- Machine Learning for Quantitative Finance Applications: A Survey ([10.3390/app9245574](https://doi.org/10.3390/app9245574))
- Deep Learning for Time Series Forecasting: A Survey (<https://www.liebertpub.com/doi/10.1089/big.2020.0159>)
- A Comprehensive Review of Deep Learning Applications in Hydrology and Water Resources ([10.31223/osf.io/xs36g](https://osf.io/xs36g))
- Nvidia Deep Learning Institute (courses.nvidia.com)
- <https://socs.binus.ac.id/2017/02/13/rnn-dan-gru/>
- [https://en.wikipedia.org/wiki/Hadamard_product_\(matrices\)](https://en.wikipedia.org/wiki/Hadamard_product_(matrices))
- <https://towardsdatascience.com/animated-rnn-lstm-and-gru-ef124d06cf45>
- <https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>
- <http://eprints.uny.ac.id/13402/1/SKRIPSI%20NANANG%20HERMAWAN-NIM%201010305141012.pdf>
- <https://www.tutorialexample.com/element-wise-addition-explained-a-beginner-guide-machine-learning-tutorial/>

Let's Code!

<https://colab.research.google.com/drive/1nXp6gIYfT3aoExNMZRK3kWNPs2i4Zdap?usp=sharing>





Kampus
Merdeka
INDONESIA JAYA

orbit
FUTURE ACADEMY
Skills
For
Future
Jobs

TERIMA KASIH

Orbit Future Academy

PT Orbit Ventura Indonesia
Center of Excellence (Jakarta Selatan)
Gedung Veteran RI, Lt.15
Unit Z15-002, Plaza Semanggi
Jl. Jenderal Sudirman Kav.50, Jakarta
12930, Indonesia

- Jakarta Selatan/Pusat
- Jakarta Barat/BSD
- Kota Bandung
- Kab. Bandung
- Jawa Barat

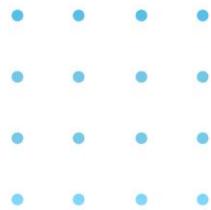
Hubungi Kami

Director of Sales & Partnership
ira@orbitventura.com
+62 858-9187-7388

Social Media

- Orbit Future Academy
- @OrbitFutureAcademyIn1
- OrbitFutureAcademy
- Orbit Future Academy

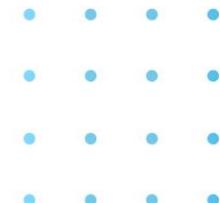
AI Mastery Course



Kampus
Merdeka
INDONESIA JAYA

Orbit
FUTURE ACADEMY | Skills
For Future Jobs

Module
Data Science
Section
K-Means Clustering



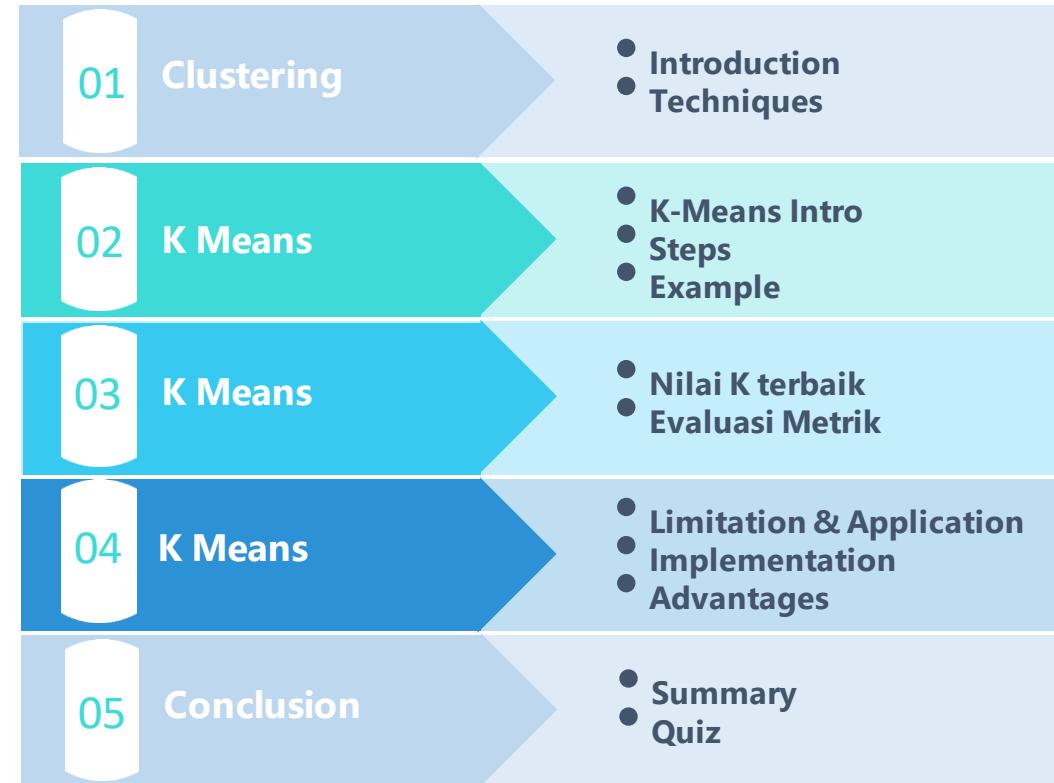


Learning Objectives

- Memahami langkah dalam clustering
- Memahami teknik dalam menentukan banyak cluster
- Dapat mengimplementasikan clustering
- Mengetahui manfaat clustering



Agenda



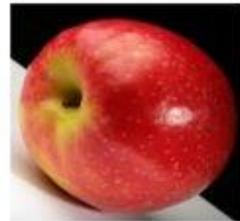


01

CLUSTERING

- Introduction
- Techniques

Clustering - Introduction

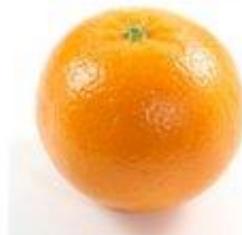
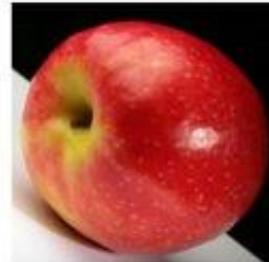


How many clusters do you expect?

Clustering - Introduction

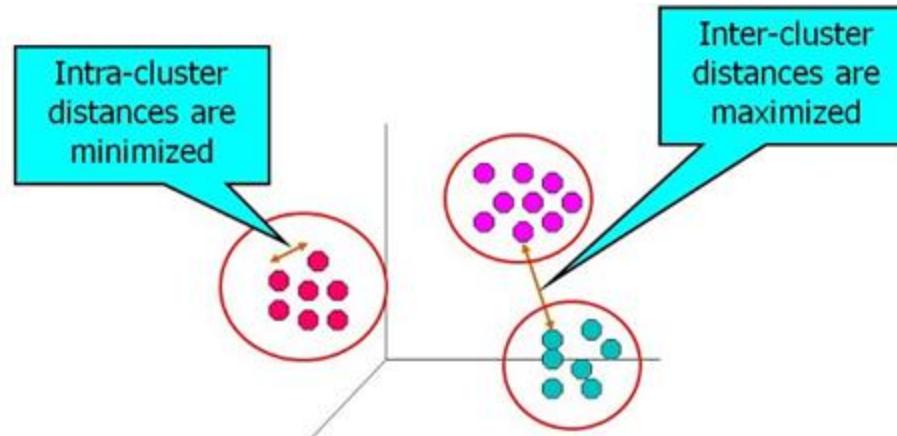
Kampus
Merdeka
INDONESIA JAYA

Orbit
FUTURE ACADEMY
Skills
For
Future
Jobs



Clustering - Introduction

Clustering adalah sebuah proses untuk mengelompokan data ke dalam beberapa *cluster* atau kelompok sehingga data dalam satu *cluster* memiliki tingkat kemiripan yang **maksimum** dan data antar *cluster* memiliki kemiripan yang **minimum** - Tan (2006).

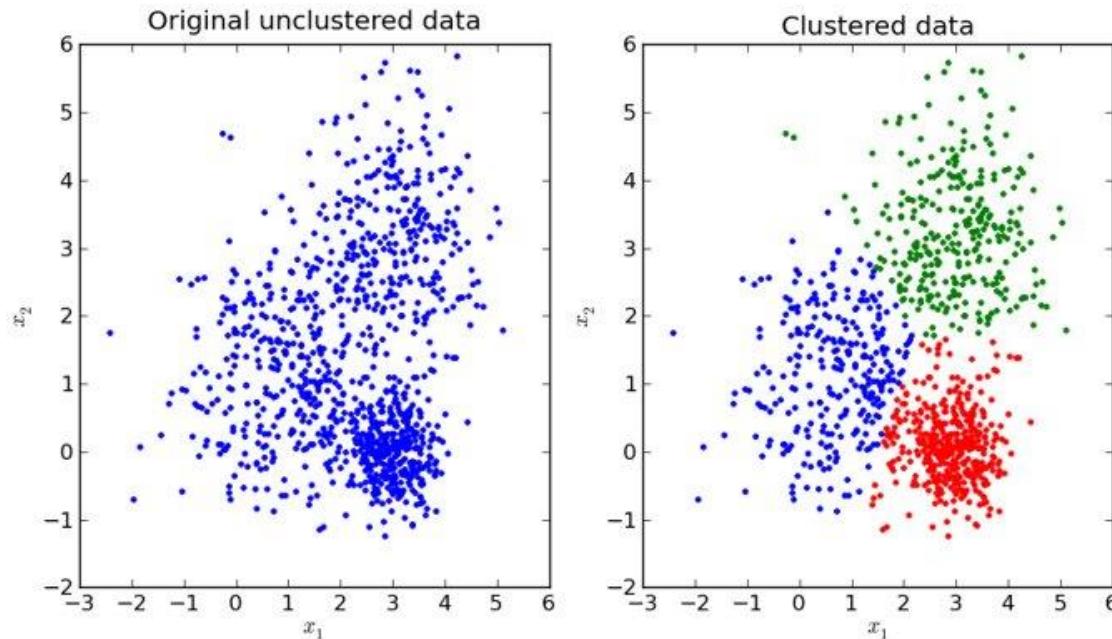


Clustering - Introduction

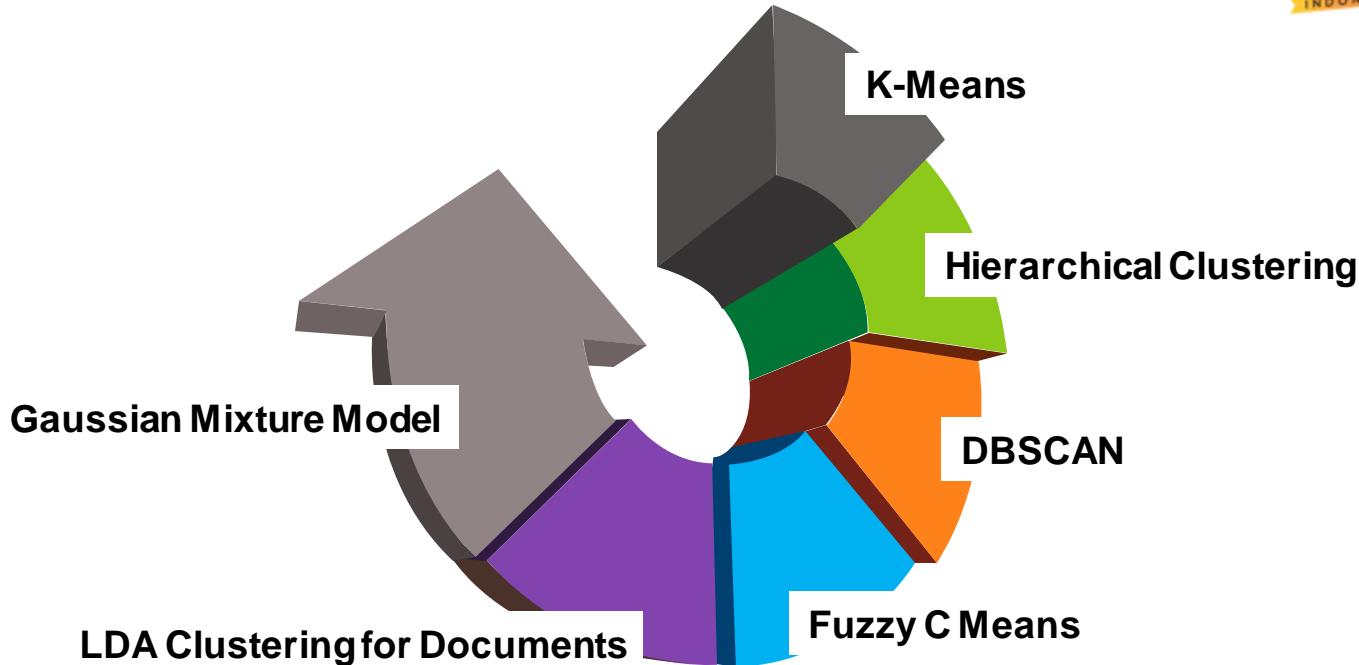
Clustering merupakan proses partisi satu set objek data ke dalam himpunan bagian yang disebut dengan *cluster*.

- Objek di dalam *cluster* yang sama memiliki kemiripan karakteristik antar satu sama lainnya dan berbeda dengan *cluster* yang lain.
- Partisi tidak dilakukan secara manual melainkan dengan suatu algoritma *clustering*.
- Oleh karena itu, *clustering* sangat berguna dan bisa menemukan *group* atau kelompok yang tidak dikenal dalam data.

Clustering - Introduction



Clustering - Introduction





02 K Means

- K-Means Intro
- Steps
- Example

K Means Clustering - Intro

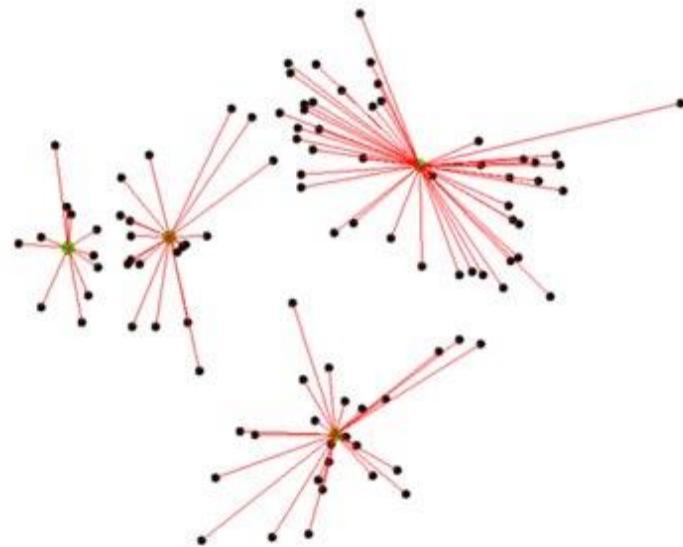
- K-means merupakan salah satu metode clustering non hirarki yang berusaha mempartisi data yang ada ke dalam bentuk satu atau lebih cluster.
- Clustering dimulai dari kelompok pertama centroid yang **dipilih secara acak**. Kelompok centroid ini digunakan sebagai titik awal untuk setiap cluster (kemudian dilakukan perhitungan berulang untuk mengoptimalkan posisi centroid).



K Means Clustering - Intro

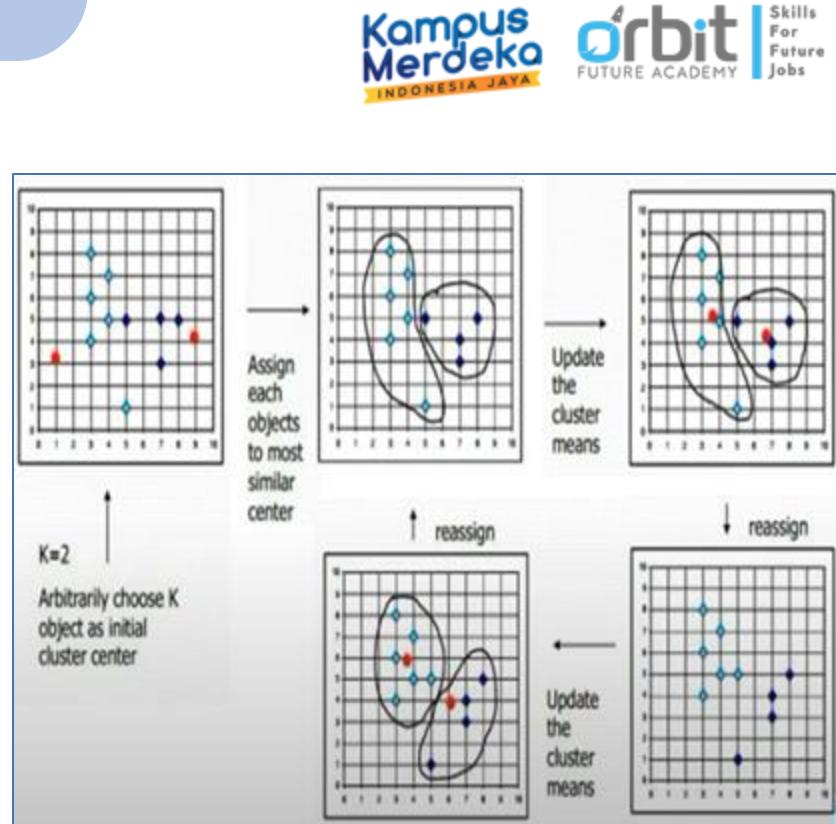
Proses perhitungan berhenti atau selesai ketika:

- Centroid telah stabil atau tidak ada perubahan cluster (konvergen)
- Jumlah iterasi yang ditentukan tercapai.



K Means Clustering – Steps

1. Tentukan jumlah cluster.
2. Alokasikan data secara random ke cluster yang ada sesuai jarak terdekat.
3. Hitung rata-rata setiap cluster dari data yang tergabung di dalamnya. Lalu, geser centroid ke means (M) yang baru
4. Alokasikan kembali semua data ke cluster sesuai jarak terdekat.
5. Ulang proses nomor 3, sampai tidak ada perubahan cluster (konvergen) atau sampai iterasi tertentu



K Means Clustering – Steps

Kita dapat menghitung jarak tiap data dengan centroid dengan menggunakan:

Euclidean Distance:

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2}$$

atau

Manhattan Distance:

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

K-Means Clustering – Contoh Kasus

Diberikan dataset seperti tabel di bawah ini. Kemudian, kita diminta untuk meng-cluster-kan data tersebut

Data ke-	X	Y	Cluster
1	3	9	?
2	9	9	?
3	12	9	?
4	3	6	?
5	3	3	?
6	6	3	?

K-Means Clustering – Penyelesaian Contoh Kasus



Skills
For
Future
Jobs

Langkah 1: Tentukan Jumlah Cluster (K)

- Pertama-tama, kita akan **menentukan jumlah cluster** yang akan dibentuk.
- Pada contoh ini, kita akan meng-cluster-kan data ke dalam **dua cluster (K=2)** berbeda, yakni **C1** dan **C2**.

Data ke-	X	Y	C1	C2
1	3	9	?	?
2	9	9	?	?
3	12	9	?	?
4	3	6	?	?
5	3	3	?	?
6	6	3	?	?

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 2: Tentukan Centroid

- Tentukan **centroid awal** tiap cluster. Pada contoh ini kita gunakan **data ke-5**: “M1: {3, 3}”; dan **data ke-6**: “M2: {6, 3}” (centroid dapat dipilih secara acak)

Data ke-	X	Y	C1	C2
1	3	9	?	?
2	9	9	?	?
3	12	9	?	?
4	3	6	?	?
5	3	3	?	?
6	6	3	?	?

Diagram showing the assignment of data points to clusters M1 and M2. Yellow arrows point from data points 5 and 6 to their respective centroid positions in the C1 and C2 columns.

M1: {3, 3}

M2: {6, 3}

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 3: Tentukan Centroid Terdekat

- Temukan **centroid terdekat** untuk setiap record. Pada contoh ini, kita akan menggunakan persamaan **Euclidean Distance**.

Euclidean Distance:

$$d(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2}$$

Data ke-1: {3, 9}

$$C1 = \sqrt{(3 - 3)^2 + (9 - 3)^2} = 6$$

$$C2 = \sqrt{(3 - 6)^2 + (9 - 3)^2} = 6.7$$

Data ke-2: {9, 9}

$$C1 = \sqrt{(9 - 3)^2 + (9 - 3)^2} = 8.5$$

$$C2 = \sqrt{(9 - 6)^2 + (9 - 3)^2} = 6.7$$

Data ke-3: {12, 9}

$$C1 = \sqrt{(12 - 3)^2 + (9 - 3)^2} = 10.8$$

$$C2 = \sqrt{(12 - 6)^2 + (9 - 3)^2} = 8.5$$

Data ke-4: {3, 6}

$$C1 = \sqrt{(3 - 3)^2 + (6 - 3)^2} = 3$$

$$C2 = \sqrt{(3 - 6)^2 + (6 - 3)^2} = 4.2$$

Data ke-5: {3, 3}

$$C1 = \sqrt{(3 - 3)^2 + (3 - 3)^2} = 0$$

$$C2 = \sqrt{(3 - 6)^2 + (3 - 3)^2} = 3$$

Data ke-6: {6, 3}

$$C1 = \sqrt{(6 - 3)^2 + (3 - 3)^2} = 3$$

$$C2 = \sqrt{(6 - 6)^2 + (3 - 3)^2} = 0$$

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 3: Tentukan Centroid Terdekat

- Selanjutnya, hasil perhitungan centroid tadi, kita susun ke dalam tabel.

Data ke-1: {3, 9}

$$C1 = \sqrt{(3 - 3)^2 + (9 - 3)^2} = 6$$

$$C2 = \sqrt{(3 - 6)^2 + (9 - 3)^2} = 6.7$$

Data ke-2: {9, 9}

$$C1 = \sqrt{(9 - 3)^2 + (9 - 3)^2} = 8.5$$

$$C2 = \sqrt{(9 - 6)^2 + (9 - 3)^2} = 6.7$$

Data ke-3: {12, 9}

$$C1 = \sqrt{(12 - 3)^2 + (9 - 3)^2} = 10.8$$

$$C2 = \sqrt{(12 - 6)^2 + (9 - 3)^2} = 8.5$$

Data ke-4: {3, 6}

$$C1 = \sqrt{(3 - 3)^2 + (6 - 3)^2} = 3$$

$$C2 = \sqrt{(3 - 6)^2 + (6 - 3)^2} = 4.2$$

Data ke-5: {6, 3} 

$$C1 = \sqrt{(6 - 3)^2 + (3 - 3)^2} = 0$$

$$C2 = \sqrt{(6 - 6)^2 + (3 - 3)^2} = 0$$

Data ke-	X	Y	C1	C2
1	3	9	6	6.7
2	9	9	8.5	6.7
3	12	9	10.8	8.5
4	3	6	3	4.2
5	3	3	0	3
6	6	3	3	0

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 3: Tentukan Centroid Terdekat

- Tentukan cluster tiap data dengan mengambil nilai centroid terkecil (terdekat). Sehingga, diperoleh cluster setiap record sebagai berikut:

Data ke-	X	Y	C1	C2	Cluster
1	3	9	6	6.7	C1
2	9	9	8.5	6.7	C2
3	12	9	10.8	8.5	C2
4	3	6	3	4.2	C1
5	3	3	0	3	C1
6	6	3	3	0	C2

Cara Menentukan Cluster (C1 atau C2):

- Contoh pada **data ke-1** $C1 = 6 < C2 = 6.7$, maka data ke-1 termasuk Cluster **C1**
- Begitu juga dengan **data ke-2**, $C1 = 8.5 > C2 = 6.7$, maka data ke-2 termasuk Cluster **C2**

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 4: Update Centroid

- Selanjutnya nilai centroid harus dihitung ulang untuk menentukan jarak minimum yang baru

Cluster 1
 $ClusterCenter(M1(x))$

$$= \frac{x_1 + x_4 + x_5}{3}$$
$$= \frac{3+3+3}{3}$$
$$= \frac{9}{3}$$
$$= 3$$

$ClusterCenter(M1(y))$

$$= \frac{y_1+y_4+y_5}{3}$$
$$= \frac{9+6+3}{3}$$
$$= \frac{18}{3}$$
$$= 6$$

Cluster 2

$ClusterCenter(M2(x))$

$$= \frac{x_2 + x_3 + x_6}{3}$$
$$= \frac{9+12+6}{3}$$
$$= \frac{27}{3}$$
$$= 9$$

$ClusterCenter(M2(y))$

$$= \frac{y_2+y_3+y_6}{3}$$
$$= \frac{9+9+3}{3}$$
$$= \frac{21}{3}$$
$$= 7$$



Update Centroid

Centroid

M1	3	6
M2	9	7

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 5: Tentukan Cluster dengan Centroid yang Baru

- Hitung jarak minimumnya kembali dengan menggunakan centroid yang baru, sehingga di dapat hasilnya sebagai berikut.

Data ke-1: {3, 9}

$$C1 = \sqrt{(3 - 3)^2 + (9 - 6)^2} = 3$$

$$C2 = \sqrt{(3 - 9)^2 + (9 - 7)^2} = 6.3$$

Data ke-4: {3, 6}

$$C1 = \sqrt{(3 - 3)^2 + (6 - 6)^2} = 0$$

$$C2 = \sqrt{(3 - 9)^2 + (6 - 7)^2} = 6.1$$

Data ke-2: {9, 9}

$$C1 = \sqrt{(9 - 3)^2 + (9 - 6)^2} = 6.7$$

$$C2 = \sqrt{(9 - 9)^2 + (9 - 7)^2} = 2$$

Data ke-5: {3, 3}

$$C1 = \sqrt{(3 - 3)^2 + (3 - 6)^2} = 3$$

$$C2 = \sqrt{(3 - 9)^2 + (3 - 7)^2} = 7.2$$

Data ke-3: {12, 9}

$$C1 = \sqrt{(12 - 3)^2 + (9 - 6)^2} = 9.5$$

$$C2 = \sqrt{(12 - 9)^2 + (9 - 7)^2} = 3.6$$

Data ke-6: {6, 3}

$$C1 = \sqrt{(6 - 3)^2 + (3 - 6)^2} = 4.2$$

$$C2 = \sqrt{(6 - 9)^2 + (3 - 7)^2} = 5$$

K-Means Clustering – Penyelesaian Contoh Kasus



Langkah 5: Tentukan Cluster dengan Centroid yang Baru

- Ulangi tahap 3 untuk melihat apakah terjadi perpindahan cluster dengan menggunakan centroid yang telah diupdate. Dari hasil perhitungan pada slide sebelumnya, diperoleh cluster baru sebagai berikut.

Data ke-	Awal				Cluster sebelumnya	Iterasi ke-1		
	X	Y	C1	C2		C1	C2	Cluster baru
1	3	9	6	6.7	C1	3	6.3	C1
2	9	9	8.5	6.7	C2	6.7	2	C2
3	12	9	10.8	8.5	C2	9.5	3.6	C2
4	3	6	3	4.2	C1	0	6.1	C1
5	3	3	0	3	C1	3	7.2	C1
6	6	3	3	0	C2	4.2	5	C1

Dapat dilihat bahwa pada **data ke-6** terjadi perpindahan cluster, artinya pada iterasi ke 1 ini hasil clustering **belum konvergen**, sehingga harus dilakukan clustering kembali (iterasi kedua) dengan nilai centroid yang harus di-update ulang.

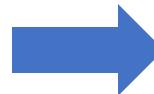
K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 6: Update Centroid Baru (2)

- Selanjutnya nilai centroid harus dihitung ulang untuk menentukan jarak minimum yang baru (2)

$$\begin{array}{ll} \text{Cluster 1} & \text{ClusterCenter}(M1(x)) \\ \text{ClusterCenter}(M1(x)) & \text{ClusterCenter}(M1(y)) \\ = \frac{x_1 + x_4 + x_5 + x_6}{4} & = \frac{y_1 + y_4 + y_5 + y_6}{4} \\ = \frac{3+3+3+6}{4} & = \frac{9+6+3+3}{4} \\ = \frac{15}{4} & = \frac{21}{4} \\ = 3.75 & = 5.25 \end{array}$$

$$\begin{array}{ll} \text{Cluster 2} & \text{ClusterCenter}(M2(x)) \\ \text{ClusterCenter}(M2(x)) & \text{ClusterCenter}(M2(y)) \\ = \frac{x_2 + x_3}{2} & = \frac{y_2 + y_3}{2} \\ = \frac{9+12}{2} & = \frac{9+9}{2} \\ = \frac{21}{2} & = \frac{18}{2} \\ = 10.5 & = 9 \end{array}$$



Update Centroid		
Centroid		
M1	3.75	5.25
M2	10.5	9

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 7: Tentukan Cluster dengan Centroid yang Baru (2)

- Hitung jarak minimumnya kembali dengan menggunakan centroid yang baru, sehingga di dapat hasilnya sebagai berikut.

Data ke-1: {3, 9}

$$C1 = \sqrt{(3 - 3.75)^2 + (9 - 5.25)^2} = 3.8$$

$$C2 = \sqrt{(3 - 10.5)^2 + (9 - 9)^2} = 7.5$$

Data ke-2: {9, 9}

$$C1 = \sqrt{(9 - 3.75)^2 + (9 - 5.25)^2} = 6.4$$

$$C2 = \sqrt{(9 - 10.5)^2 + (9 - 9)^2} = 1.5$$

Data ke-3: {12, 9}

$$C1 = \sqrt{(12 - 3.75)^2 + (9 - 5.25)^2} = 9.1$$

$$C2 = \sqrt{(12 - 10.5)^2 + (9 - 9)^2} = 1.5$$

Data ke-4: {3, 6}

$$C1 = \sqrt{(3 - 3.75)^2 + (6 - 5.25)^2} = 1.1$$

$$C2 = \sqrt{(3 - 10.5)^2 + (6 - 9)^2} = 8.1$$

Data ke-5: {3, 3}

$$C1 = \sqrt{(3 - 3.75)^2 + (3 - 5.25)^2} = 2.4$$

$$C2 = \sqrt{(3 - 10.5)^2 + (3 - 9)^2} = 9.6$$

Data ke-6: {6, 3}

$$C1 = \sqrt{(6 - 3.75)^2 + (3 - 5.25)^2} = 3.2$$

$$C2 = \sqrt{(6 - 10.5)^2 + (3 - 9)^2} = 7.5$$

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 7: Tentukan Cluster dengan Centroid yang Baru (2)

- Ulangi tahap 3 untuk melihat apakah terjadi perpindahan cluster dengan menggunakan centroid yang telah diupdate. Dari hasil perhitungan pada slide sebelumnya, diperoleh cluster baru sebagai berikut.

Data ke-	Awal					Cluster sebelumnya	Iterasi ke-1			Iterasi ke-2		
	X	Y	C1	C2			C1	C2	Cluster baru	C1	C2	Cluster baru
1	3	9	6	6.7		C1	3	6.3	C1	3.8	7.5	C1
2	9	9	8.5	6.7		C2	6.7	2	C2	6.4	1.5	C2
3	12	9	10.8	8.5		C2	9.5	3.6	C2	9.1	1.5	C2
4	3	6	3	4.2		C1	0	6.1	C1	1.1	8.1	C1
5	3	3	0	3		C1	3	7.2	C1	2.4	9.6	C1
6	6	3	3	0		C2	4.2	5	C1	3.2	7.5	C1

Dapat dilihat bahwa pada **iterasi 2 (Cluster baru)** tidak terjadi perpindahan cluster pada setiap record, maka pengclusteran sudah konvergen atau **sudah optimal**.

K-Means Clustering – Penyelesaian Contoh Kasus

Langkah 8: Clustering Selesai

- Dapat dilihat bahwa pada **iterasi 2 (Cluster baru)** tidak terjadi **perpindahan cluster** pada setiap record, maka peng-clusteran sudah konvergen atau **sudah optimal**.

Data ke-	Awal					Cluster sebelumnya	Iterasi ke-1			Iterasi ke-2		
	X	Y	C1	C2	C1		C2	Cluster baru	C1	C2	Cluster baru	
1	3	9	6	6.7	C1	3	6.3	C1	3.8	7.5	C1	
2	9	9	8.5	6.7	C2	6.7	2	C2	6.4	1.5	C2	
3	12	9	10.8	8.5	C2	9.5	3.6	C2	9.1	1.5	C2	
4	3	6	3	4.2	C1	0	6.1	C1	1.1	8.1	C1	
5	3	3	0	3	C1	3	7.2	C1	2.4	9.6	C1	
6	6	3	3	0	C2	4.2	5	C1	3.2	7.5	C1	



03 K Means

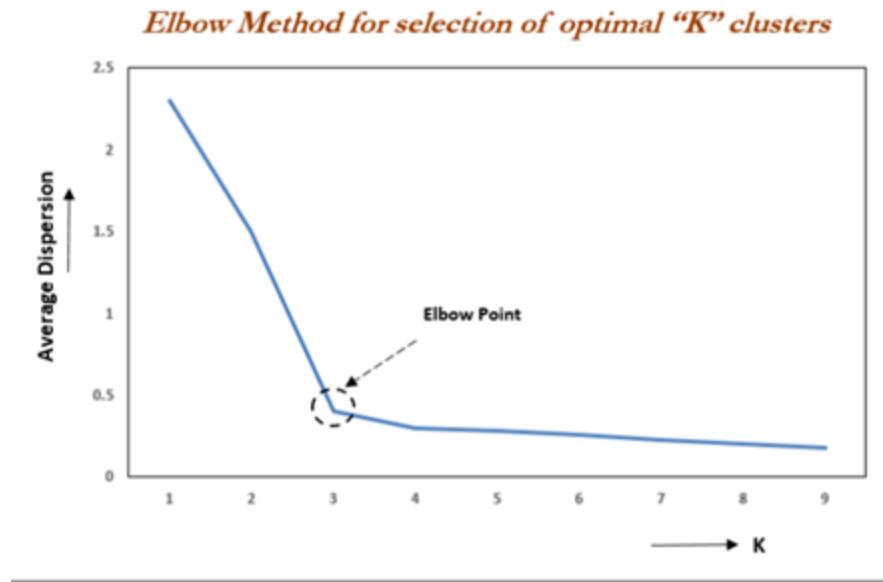
- Nilai K terbaik
- Evaluasi Metrik

Finding the best value of K

Metode Elbow merupakan suatu metode yang digunakan untuk menghasilkan informasi dalam menentukan jumlah cluster terbaik dengan cara melihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik.

Berikut ini tahapan algoritma metode Elbow dalam menentukan nilai k pada K-Means:

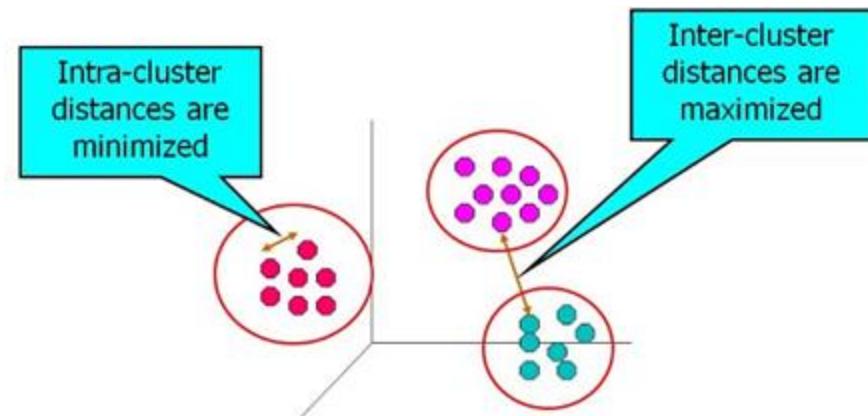
1. Menginisialisasi awal nilai k ;
2. Menaikan nilai k ;
3. Menghitung hasil sum of square error dari tiap nilai k ;
4. Analisis hasil sum of square error dari nilai k yang mengalami penurunan secara drastis ;
5. Cari dan tetapkan nilai k yang berbentuk siku.



Matriks Evaluasi pada Clustering

Terdapat beberapa metode yang dapat digunakan untuk mengukur performance dari clustering :

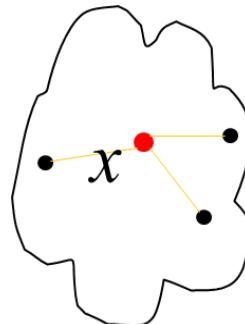
- Silhouette Score
- Rand Index
- Adjusted Rand Index
- Mutual Information
- Calinski-Harabasz Index
- Davies-Bouldin Index



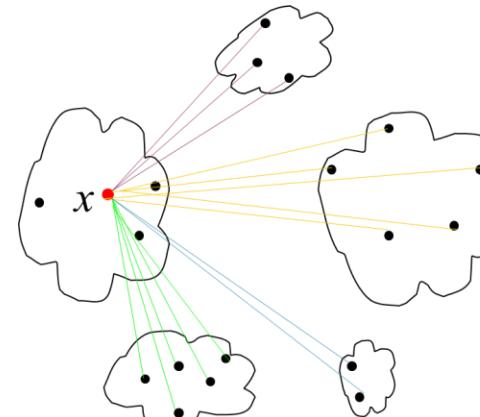
Silhouette Score (SC)

Untuk evaluasi cluster secara umum, kita bisa menggunakan Silhouette Coefficient (SC). SC menggunakan nilai Cohesion dan Separation

- **Cohesion** : mengukur seberapa dekat jarak antar objek di dalam cluster
- **Separation** : mengukur seberapa terpisah jarak antar cluster

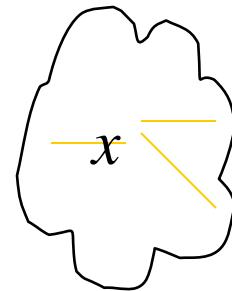


cohesion



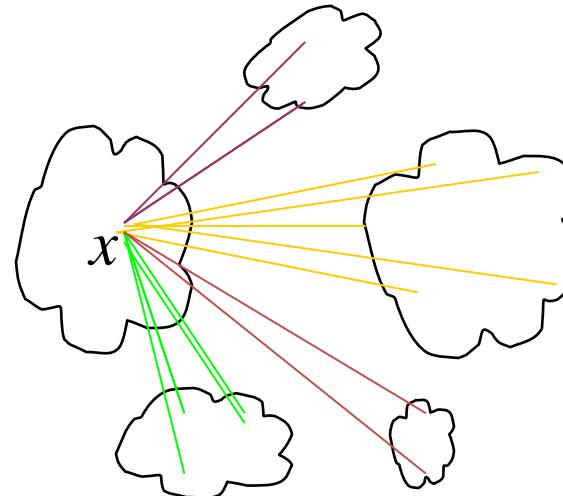
separation

Silhouette Score (SC)



cohesion

$a(x)$: jarak rata-rata x ke semua vektor lain dalam cluster yang sama



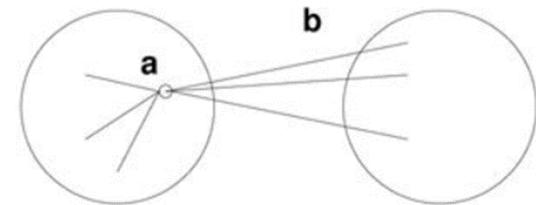
separation

$b(x)$: jarak rata-rata x ke vektor di cluster lain. Lalu pilih yang minimum di antara cluster.

Silhouette Score (SC)

- X merupakan nilai centroid dari setiap cluster
- Hitung a = jarak rata-rata x ke semua vektor lain dalam cluster yang sama
- Hitung b = jarak rata-rata minimum x ke vektor di cluster lain
- silhouette $s(x)$:

$$s(x) = \frac{b(x) - a(x)}{\max \{a(x), b(x)\}}$$



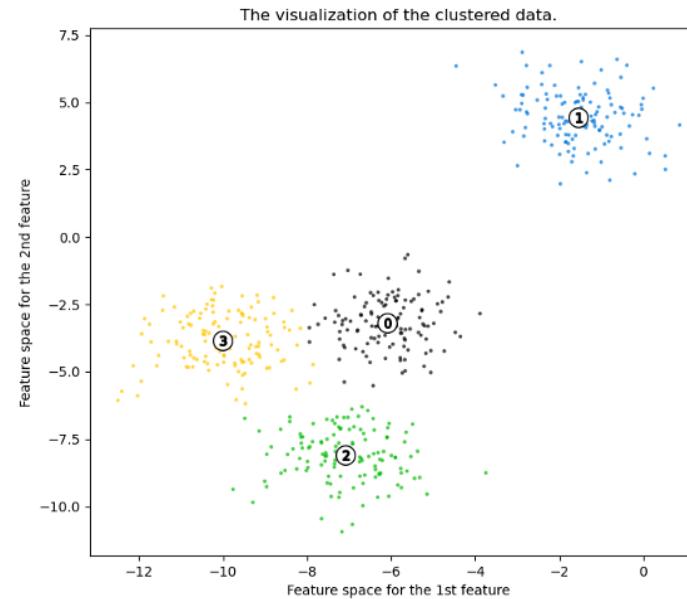
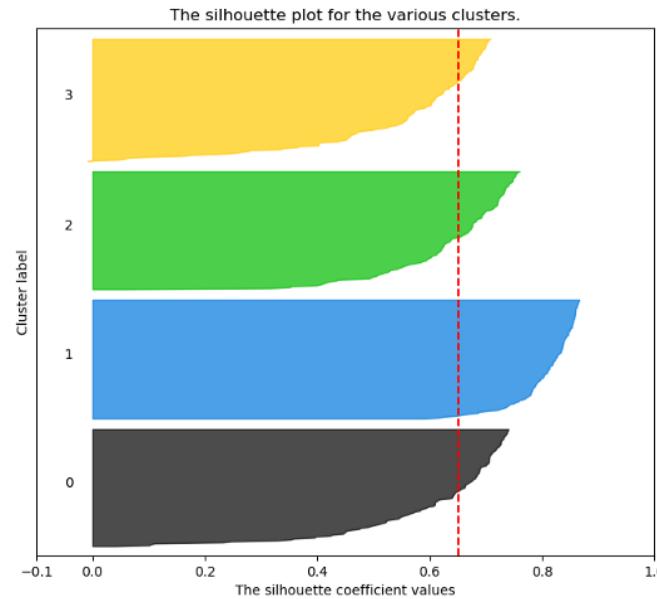
$s(x) = [-1, +1]$: -1 = buruk, 1 = bagus

- Silhouette Coefficient (SC):

$$SC = \frac{1}{N} \sum_{i=1}^N s(x)$$

Silhouette Score (SC)

Contoh visualisasi hasil evaluasi dengan Silhouette Score dengan nilai k = 4



- Rata-rata nilai silhouette pada k = 4 adalah 0.65



04 K Means

- Limitation & Application
- Advantages & Disadvantages
- Implementation

Limitation

Jenis data mudah divisualisasikan sebagai data 2-D dan juga dalam bentuk ini K-Means dapat bekerja dengan baik. Dimana bentuk yang lebih kompleks seperti square atau sphere akan sulit bagi K-Means untuk bekerja dengan baik pada dataset tersebut.

Bentuk 2-D sederhana akan diclusterkan dengan baik menggunakan algoritma K-Means.

Application

- Relatif efisien dan cepat, dengan hasil komputasinya $O(tkn)$.
- K-Means dapat diaplikasikan pada machine learning atau data mining
- Digunakan untuk pemahaman data akustik dalam memahami ucapan untuk mengkonversi dari bentuk gelombang menjadi salah satu kategori.
- Digunakan untuk memilih color palettes pada devices old fashioned graphical display dan image quantization.

Kelebihan dan Kekurangan K-Means

Kelebihan

- Sederhana
- Umum digunakan
- Mudah diimplementasikan
- Waktu training relatif cepat

Kekurangan

- Centroid di-inisialisasi secara random sehingga proses pencarian cluster dapat berbeda-beda (waktu dan kompleksitasnya)
- Biaya komputasi cukup tinggi karena diperlukan perhitungan jarak dari tiap data ke centroid



05 CONCLUSION

- Summary
- Quiz

Summary

- Clustering hampir sama dengan klasifikasi, akan tetapi klasifikasi sudah diketahui variabel kelas/kategorinya, sedangkan clustering harus ditentukan sendiri dari data yang ada.
- Cluster terbentuk berdasarkan data-data yang dekat dengan centroid cluster tersebut serta jauh dengan centroid dari cluster lainnya.
- Clustering akan bekerja dengan baik untuk data yang dapat divisualisasikan dengan bentuk 2-D.
- Semakin banyak data, cluster, serta iterasinya maka akan menyebabkan proses komputasi juga semakin besar.

Quiz

1. Nilai K pada K-means clustering memiliki makna untuk ...

- a. banyak data
- b. banyak cluster
- c. banyak iterasi
- d. banyak eror

1. Nilai K pada K-means clustering memiliki makna untuk ...

- a. banyak data
- b. banyak cluster
- c. banyak iterasi
- d. banyak eror

Jawaban: b

2. Mengukur jarak antara centroid cluster dengan objek yang berada dalam cluster disebut ...

- a. Cohesion
- b. Separation
- c. Coefficient
- d. Support

2. Mengukur jarak antara centroid cluster dengan objek yang berada dalam cluster disebut ...

- a. Cohesion
- b. Separation
- c. Coefficient
- d. Support

Jawaban: a

Let's Code!

<https://colab.research.google.com/drive/1naxosIBFzXSXz5gsZCLozjglOr8O895v?usp=sharing>



TERIMA KASIH

Orbit Future Academy

PT Orbit Ventura Indonesia
Center of Excellence (Jakarta Selatan)
Gedung Veteran RI, Lt.15
Unit Z15-002, Plaza Semanggi
Jl. Jenderal Sudirman Kav.50, Jakarta
12930, Indonesia

- Jakarta Selatan/Pusat
- Jakarta Barat/BSD
- Kota Bandung
- Kab. Bandung
- Jawa Barat

Hubungi Kami

Director of Sales & Partnership
ira@orbitventura.com
+62 858-9187-7388

Social Media

- Orbit Future Academy
- @OrbitFutureAcademyInd
- OrbitFutureAcademy
- Orbit Future Academy