AI Mastery Course

orbit FUTURE ACADEMY | Skills For Future Jobs

Format Nama:
NamaLengkap_NamaKelas

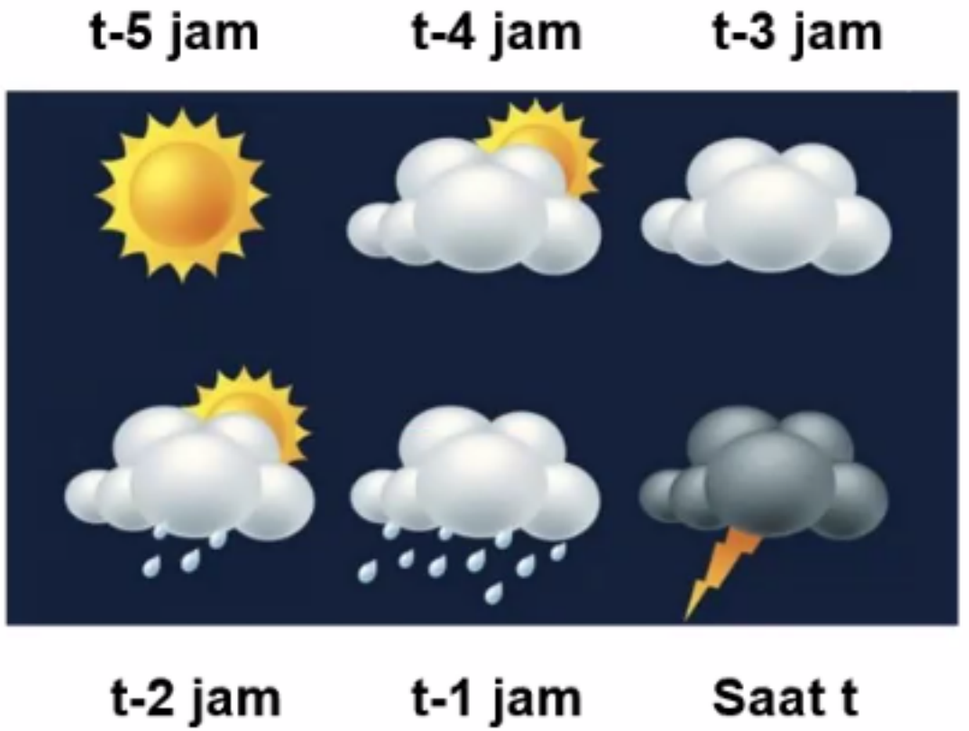
Module
Data Science

Section
Time Series
(Statistical Approach)

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.



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)
10	9	8
9	9	8
8	8	7
8	7	8
7	7	7
10	8	9

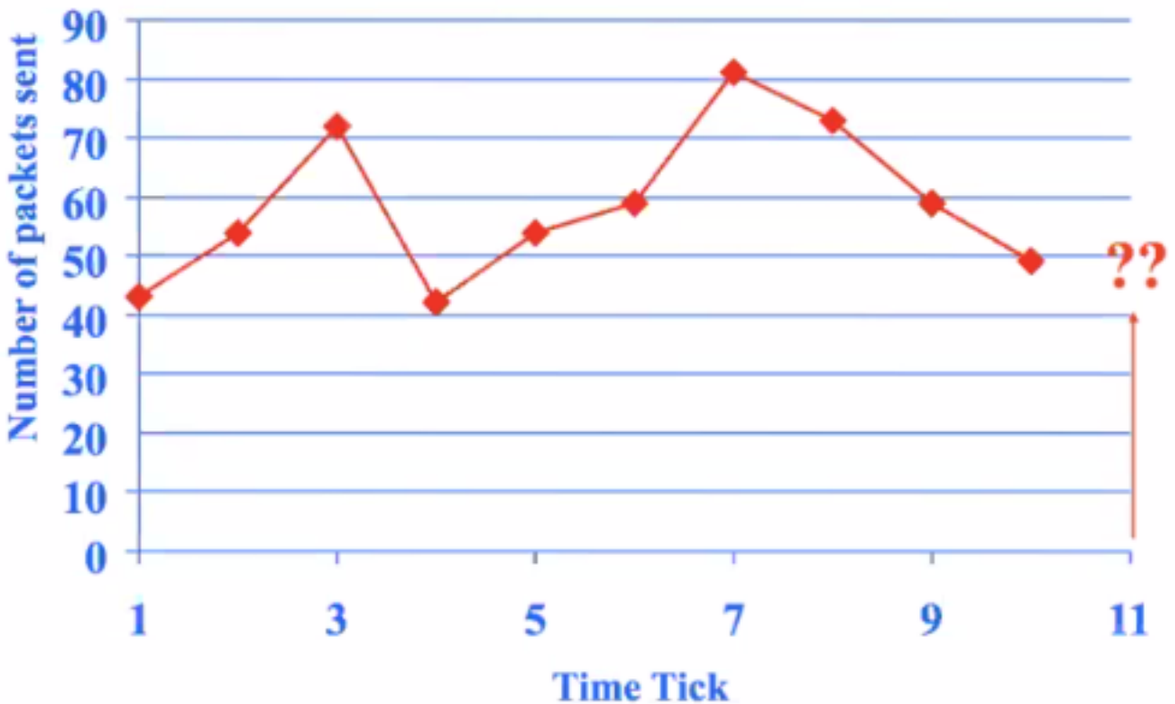
Jam(t)	Cuaca(Y1t)	Suhu(Y2t)
13	berawan	27
14	mendung	26
15	hujan	25
16	hujan	25
17	hujan	24
18	berawan	24

Brand	Tahun	Tingkat promo(Xt)	Penjualan(Yt)
A	2011	9	8000
A	2012	8	10000
A	2013	7	10000
B	2011	8	8000
B	2012	9	9000
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 ...*

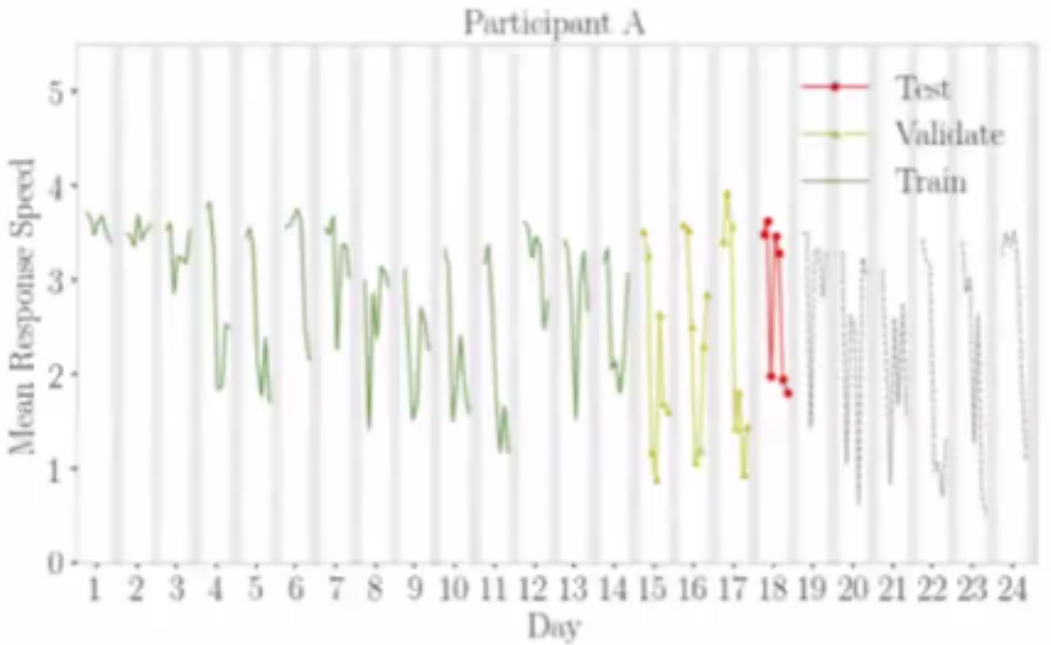
A Time-Series Analysis of my Girlfriends Mood Swings

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

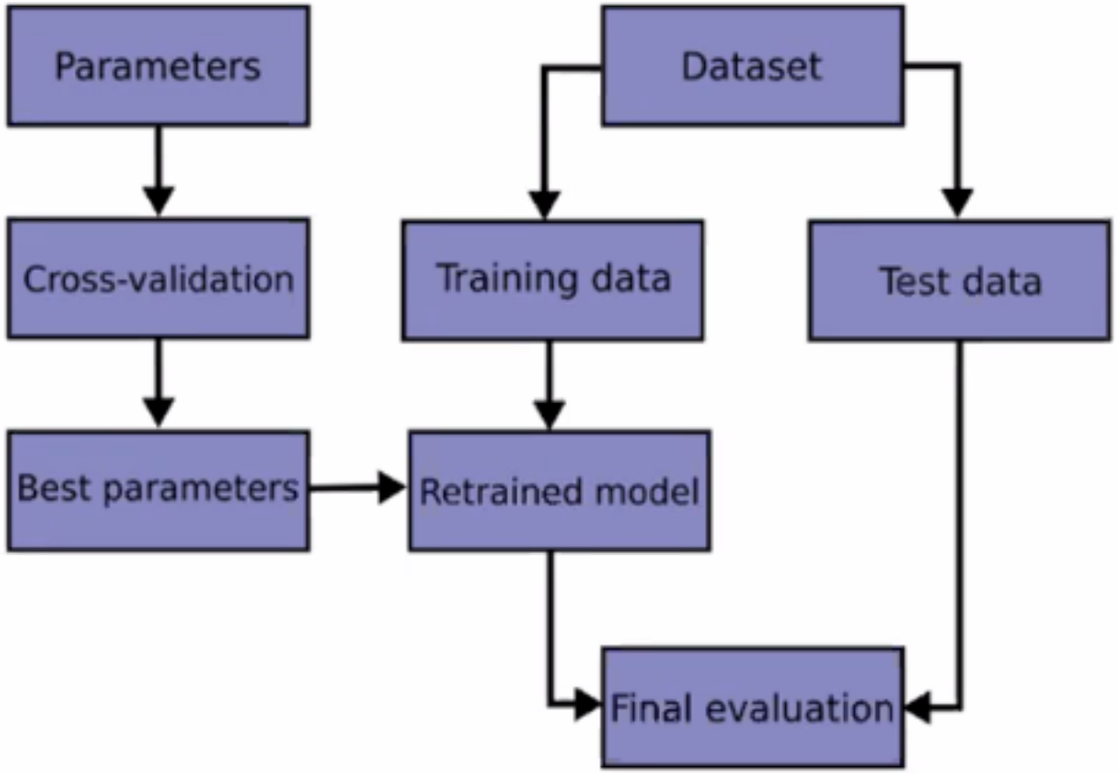
*penelitian dari Department of Applied Psychological Machine Learning, University, Pittsburgh, PA, USA (https://jabde.com/2021/05/23/girlfriends-mood-time-series-analysis/)

Data

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.



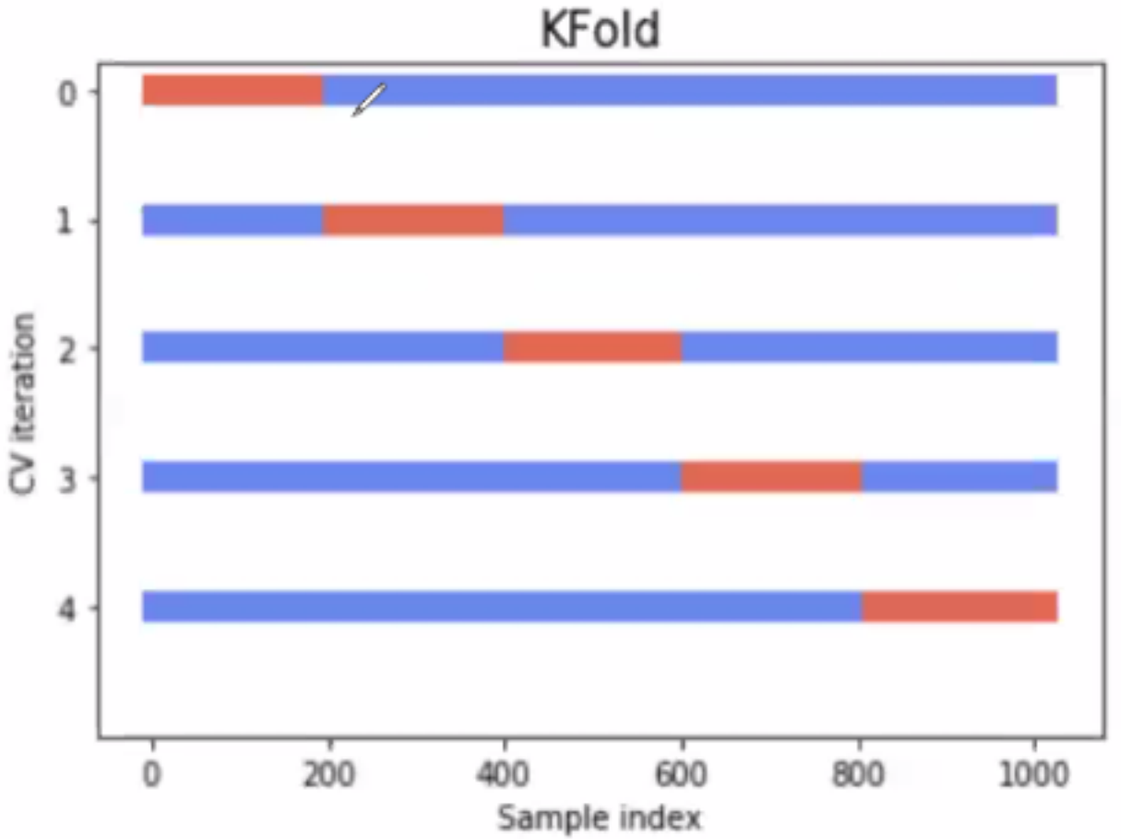
Cross Validation



- **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/)

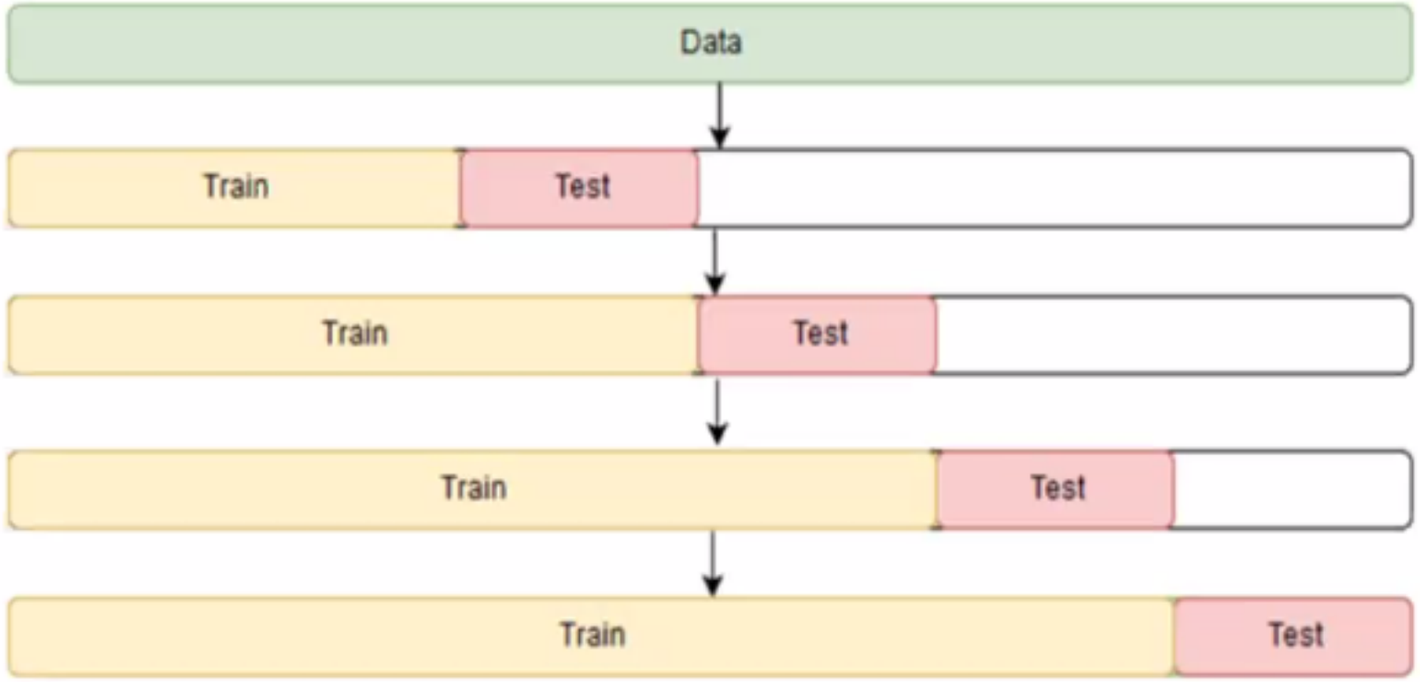
Cross Validation: K-Fold



- 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/)

Cross Validation



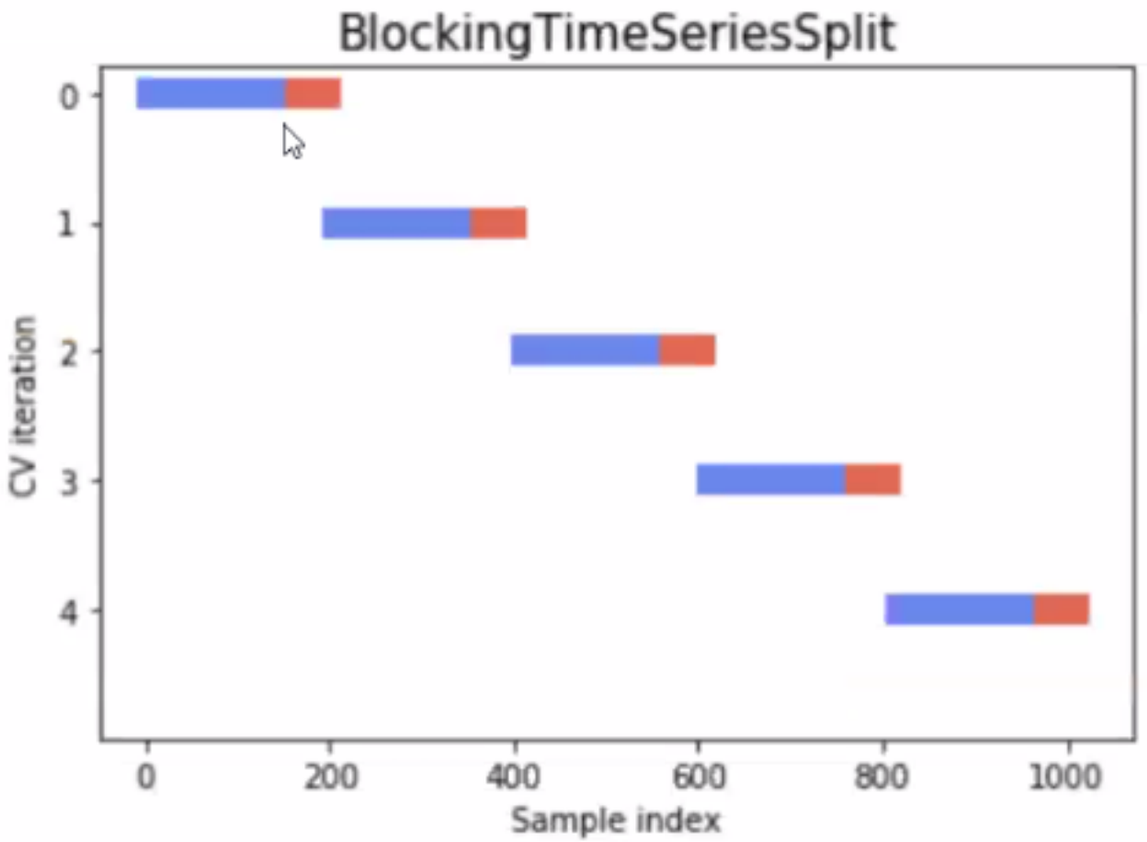
Teknik Cross Validation untuk **Time Series**:

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 Blocked



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

Cross Validation

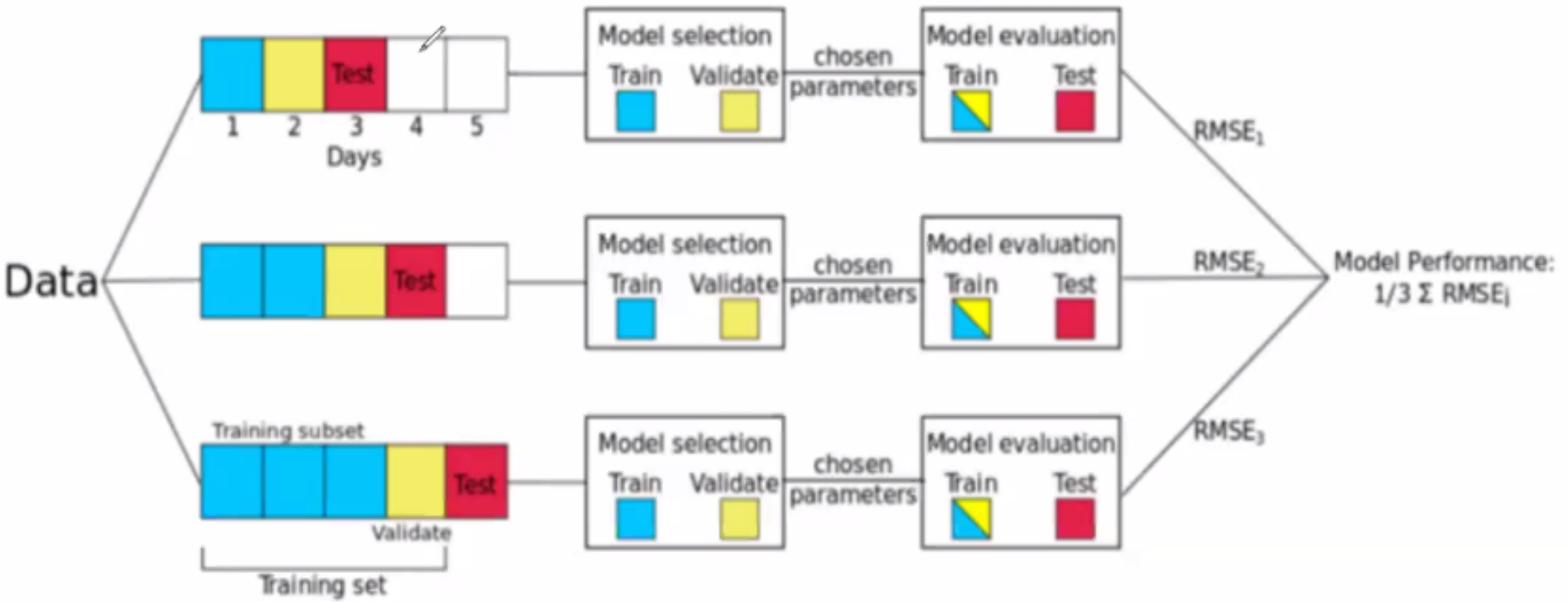
Teknik Predict Second Half

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

orbit FUTURE ACADEMY | Skills For Future Jobs

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

Statistics Model

Secara Umum, ada 2 jenis model di time series:

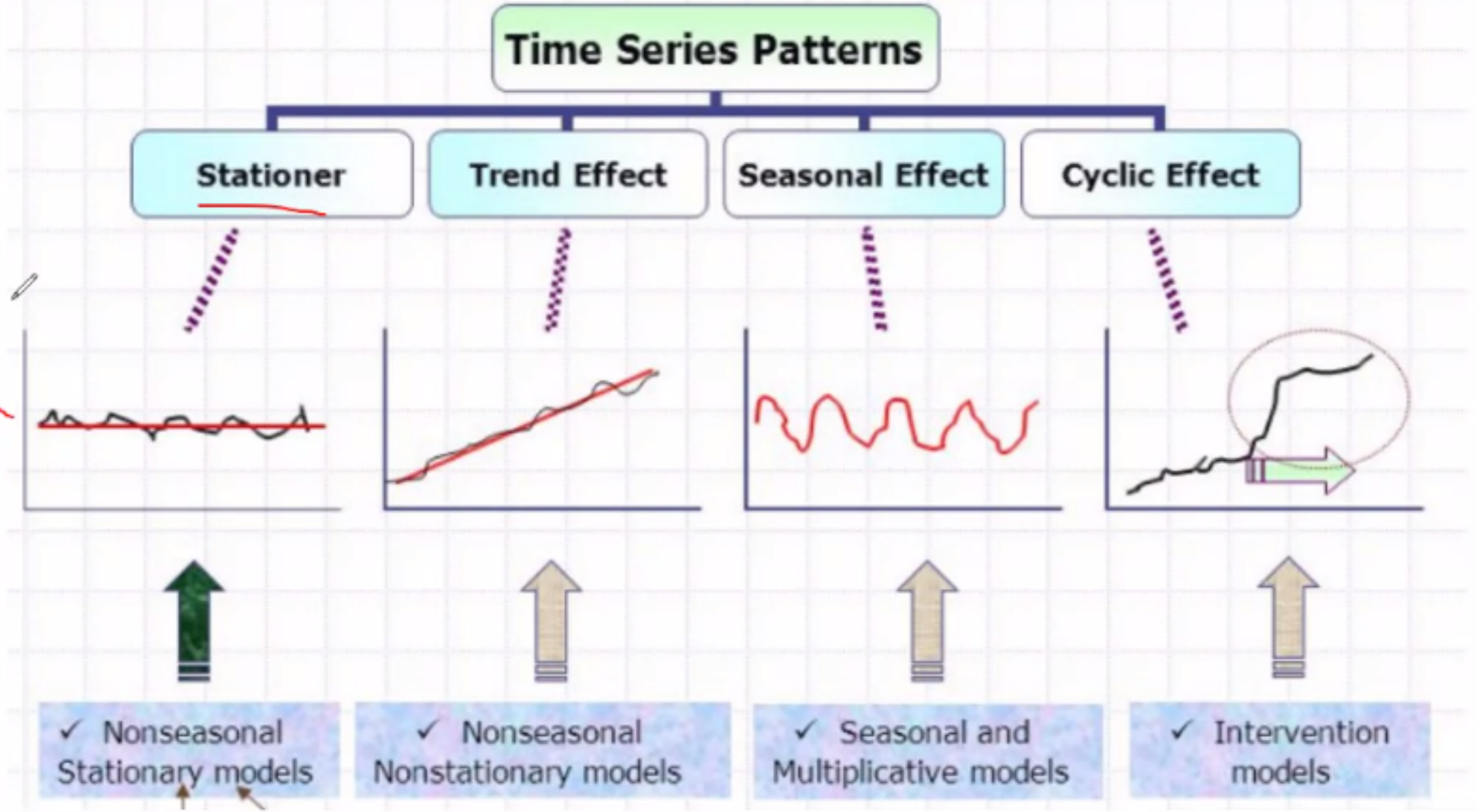
Statistical Approach	Deep Learning Approach
• Metode Box-jenkins (ARIMA) • Model musiman • Model trend • Vector Autoregressive (VAR) • Hidden Markov Model (HMM) • dll	• Recurrent Neural Network (RNN) • Long Short-Term Memory (LSTM) • Wavenet • Multi-Layer Perceptron • dll

Statistics Model

Secara Umum, ada 2 jenis model di time series:

Statistical Approach	Deep Learning Approach
• Metode Box-jenkins (ARIMA) • Model musiman • Model trend • Vector Autoregressive (VAR) • Hidden Markov Model (HMM) • dll	• Recurrent Neural Network (RNN) • Long Short-Term Memory (LSTM) • Wavenet • Multi-Layer Perceptron • dll

Statistics Model



Box-Jenkins

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 :

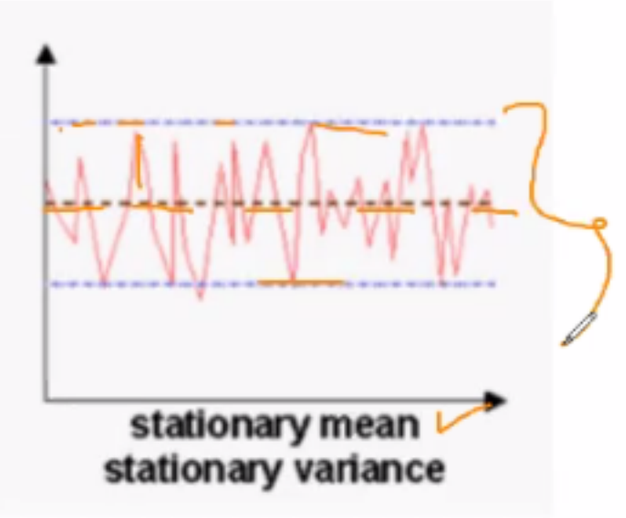
1. Uji Stasioneritas
2. Jika tidak stasioner lakukan transformasi atau differencing
3. Uji ACF & PACF
4. Pilih model ARIMA(p,d,q) yang tepat
5. Evaluasi model dengan RMSE dll
6. Lakukan forecasting

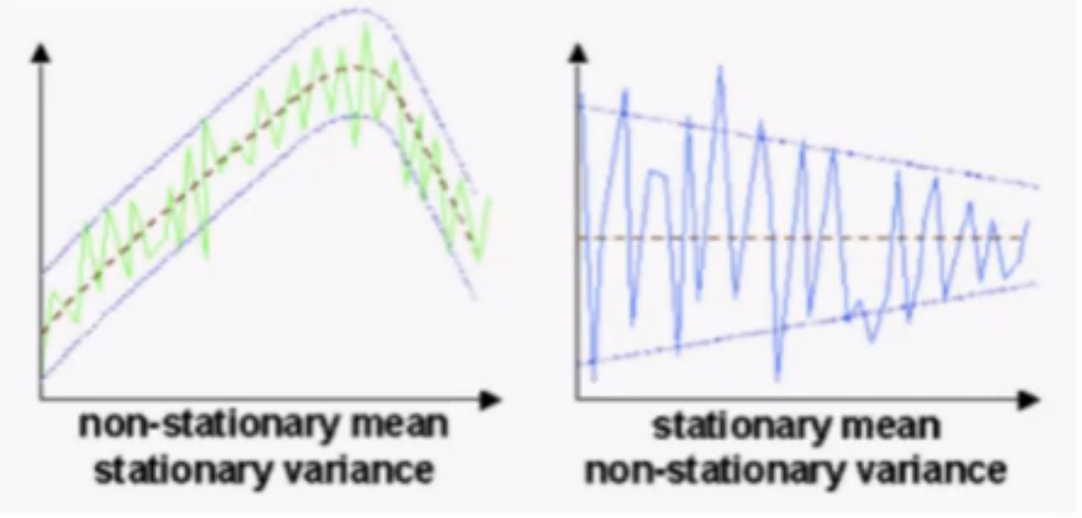
Box-Jenkins

Uji Stasioneritas

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

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

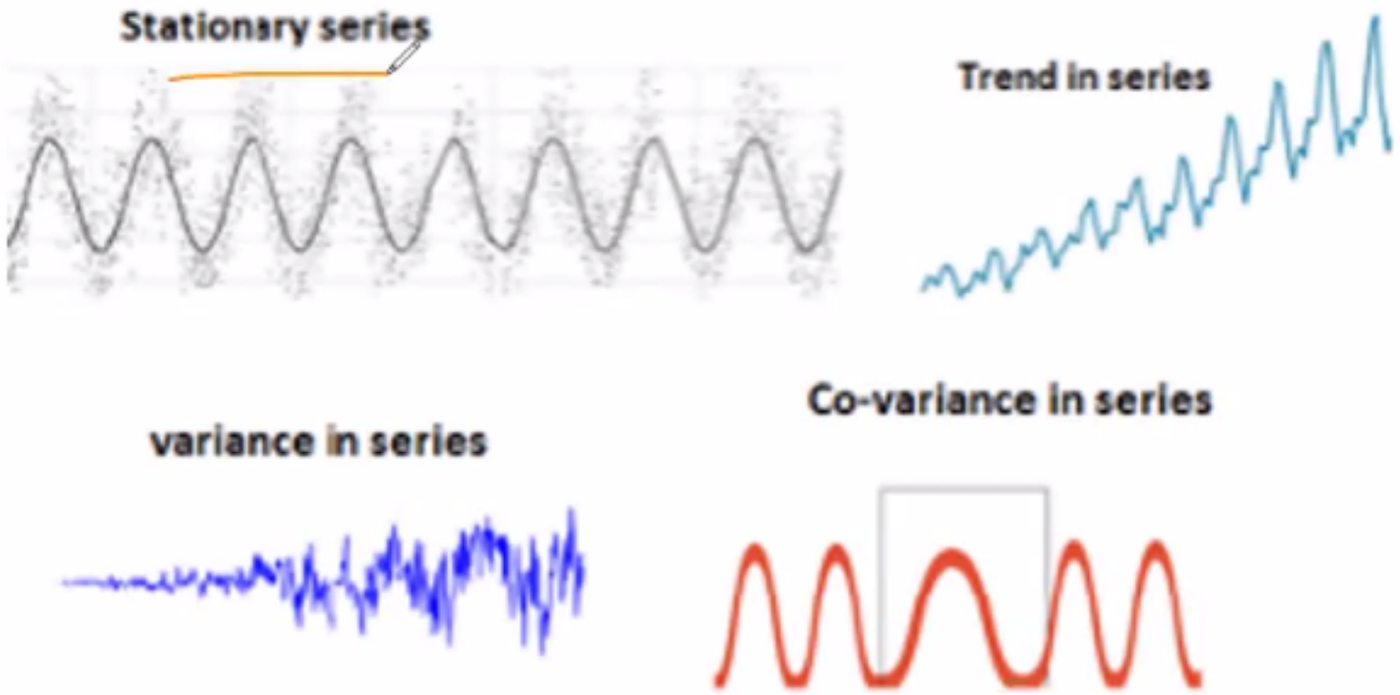




Box-Jenkins

Uji Stasioneritas

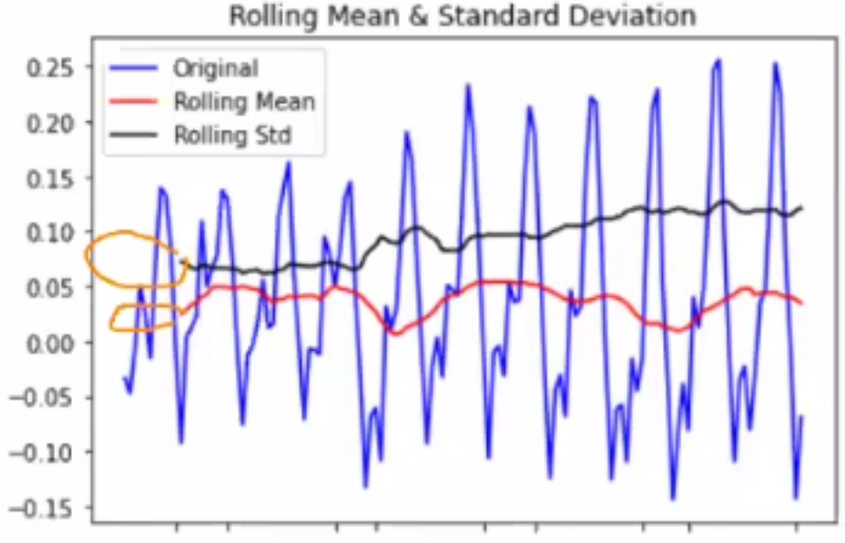
1. Plotting line graph of the data

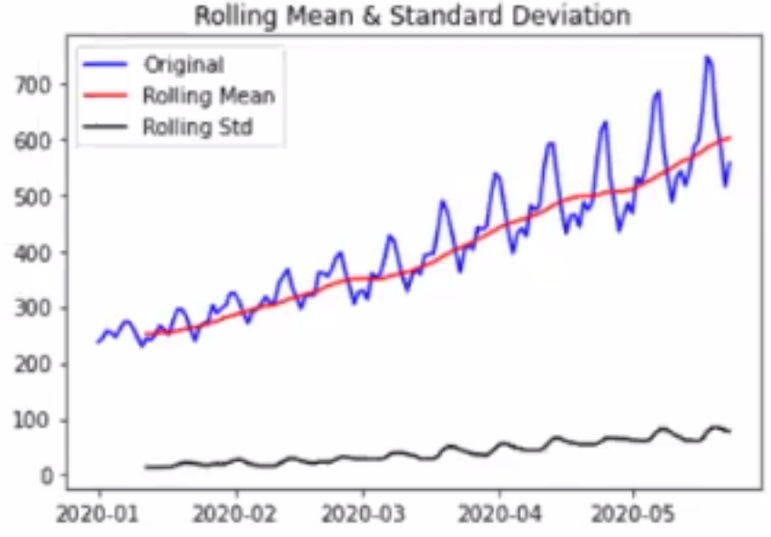


Data yang stasioner terlihat dari grafik garisnya yang naik turun di tempat yang sama

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

Uji Stasioneritas

3. Dickey-fuller Test

Tes ini menggunakan hipotesis:

Ho : Data tidak stasioner

Ha : Data Stasioner

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

adf

```
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(data))
2. Differencing, differencing adalah mengubah data menjadi:

$$d^{(1)}(t) = x(t) - x(t-1)$$ → difference 1 kali

$$d^{(2)}(t) = d^{(1)}(t) - d^{(1)}(t-1)$$ → difference 2 kali

$$d^{(m)}(t) = d^{(m-1)}(t) - d^{(m-1)}(t-1)$$ → difference m kali

Setelah transformasi dan differencing, lakukan uji stasioner lagi

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 : $\Phi(1) = \rho(1)$.

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),** subtitusikan $d^{(m)}(t) = d^{(m-1)}(t) - d^{(m-1)}(t-1)$ ke x_t

Box-Jenkins

Pemilihan Model

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

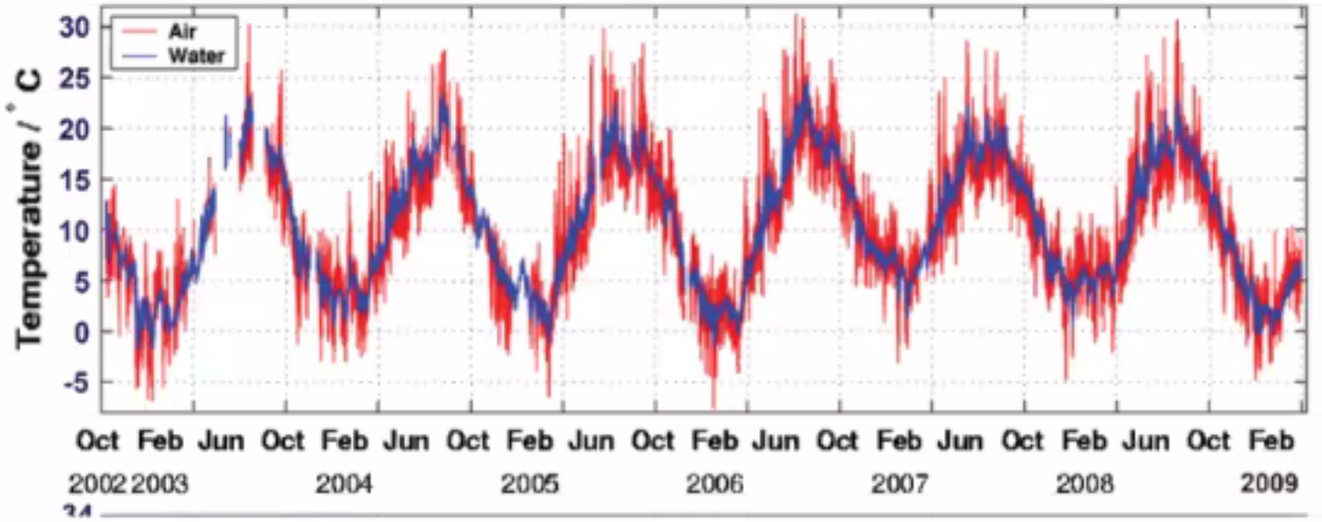
ARIMA(p, d, q) :
p = cuts offnya AR
d = berapa kali differencing
q = cuts offnya MA

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)

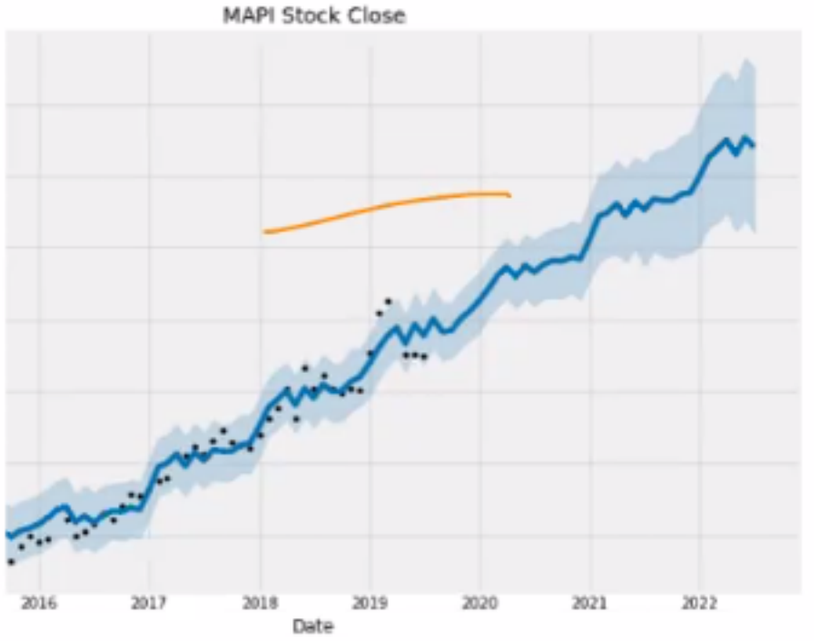
Seasonal ARIMA (SARIMA)

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.**



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 = base + trend + seasonality + residual

Multiplicative Decomposition

Mengkombinasikan model ARIMA, trend, dan seasonal dengan mengalikan :

y = base x trend x seasonality x residual