QUANTUS
telematika indonesia

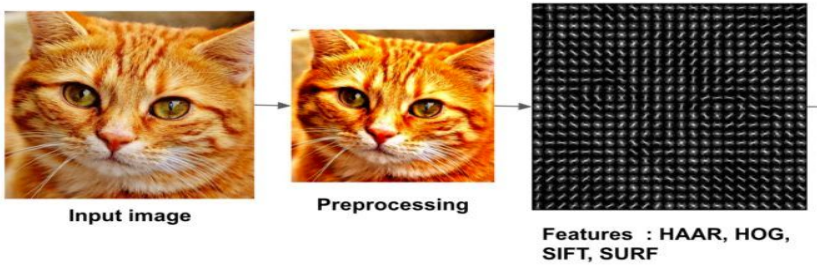
Feature Engineering & Machine Learning 101: Linear Regression

Raden Muhammad Hadi

Feature Engineering
dan
Machine Learning

Curse of Dimensionality

- Dataset terkadang memiliki **dimensi yang sangat besar**!
  - **Fraud Data**
  - **Gambar**: 28 x 28 pixels
  - **Teks**: 1000000 kata = 1000000 fitur!
- **Contoh**:
  - Gambar kucing

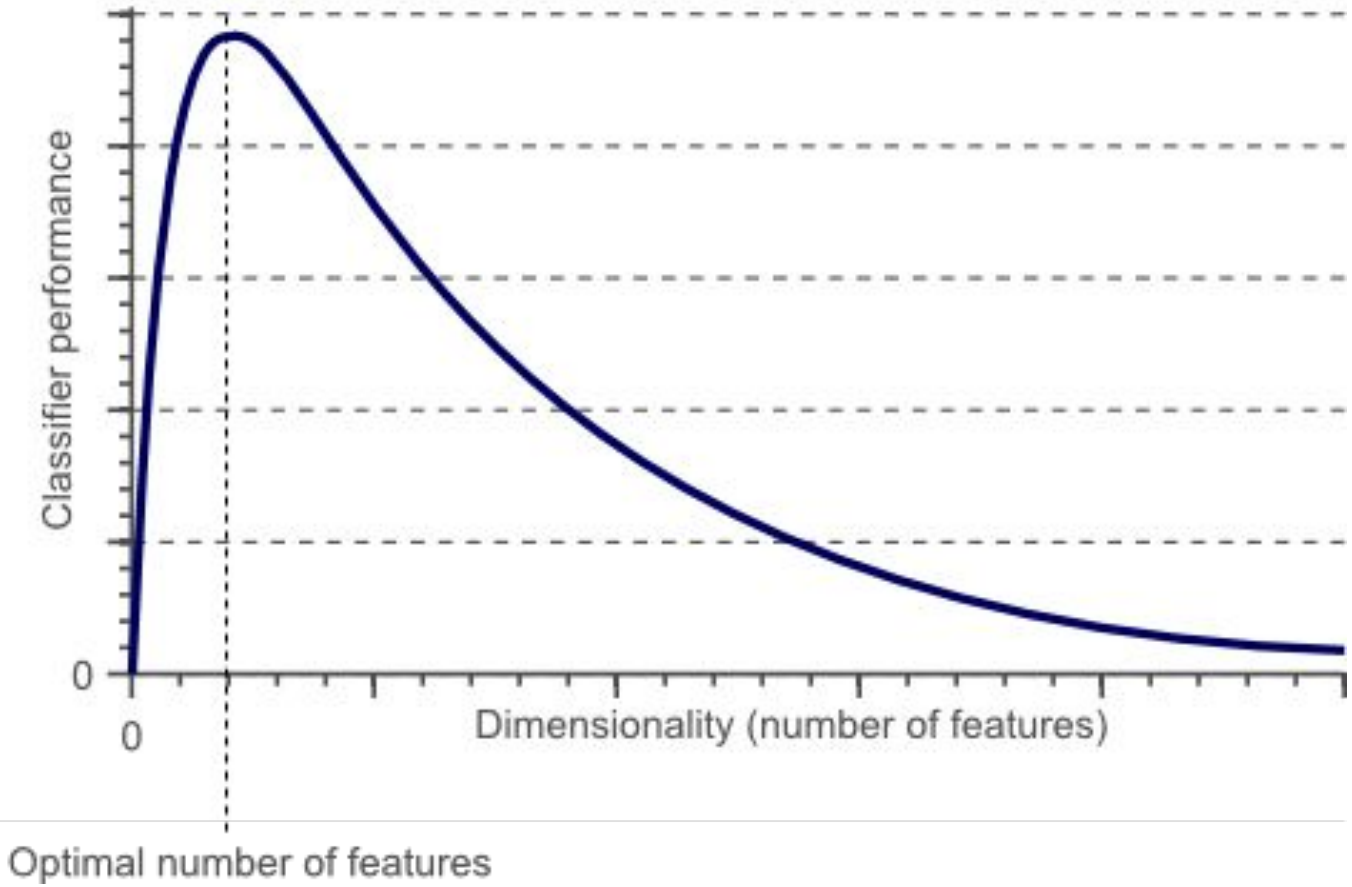


Curse of Dimensionality

- Biaya komputasional tinggi
- Beberapa fitur kadang tidak diperlukan
- *Labeling data* sangat memboroskan waktu dan tenaga

Encoding classes for sequence labeling

	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
's	O	O
new	O	O
painting	O	O

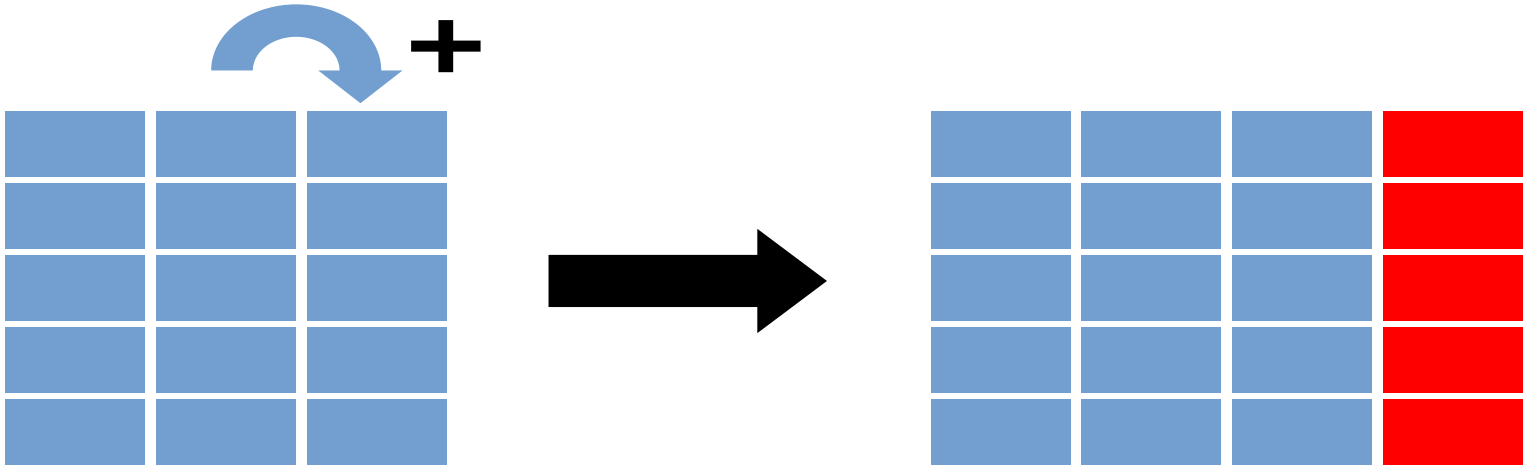

Classifier performance
0
0
Dimensionality (number of features)
Optimal number of features

Cure the Curse

- **Menggunakan *domain knowledge***
  - Seleksi fitur
  - *Feature importance*
  - *PCA*

***C**ure the **C**urse*

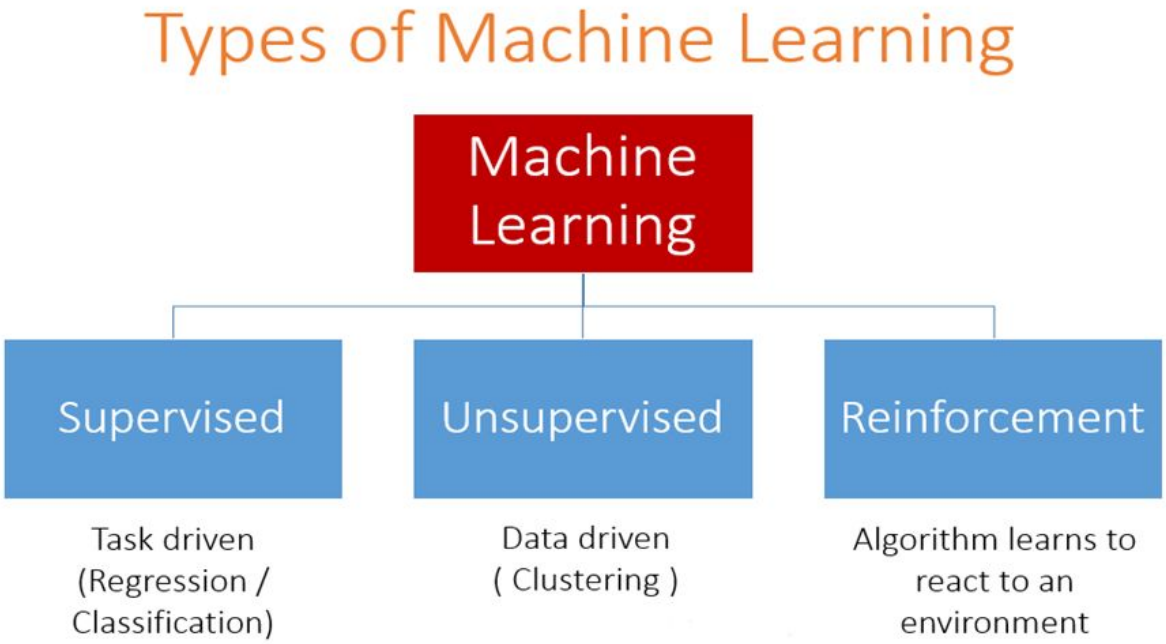
- **Ekstraksi Fitur**
  - Menambah dimensi dari data
  - Dapat **diotomatisasi** dengan algoritma:
    - *Ensemble*, *Feature Synthesis*, etc



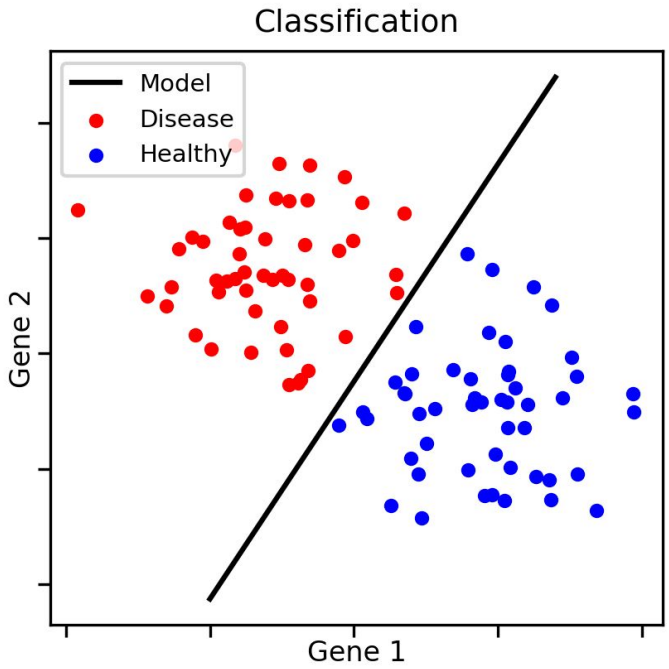
Memperoleh fitur yang baik itu **sulit, membutuhkan waktu serta pengetahuan ahli**. "Pembelajaran Mesin Terapan" merupakan arti lain dari rekayasa fitur.

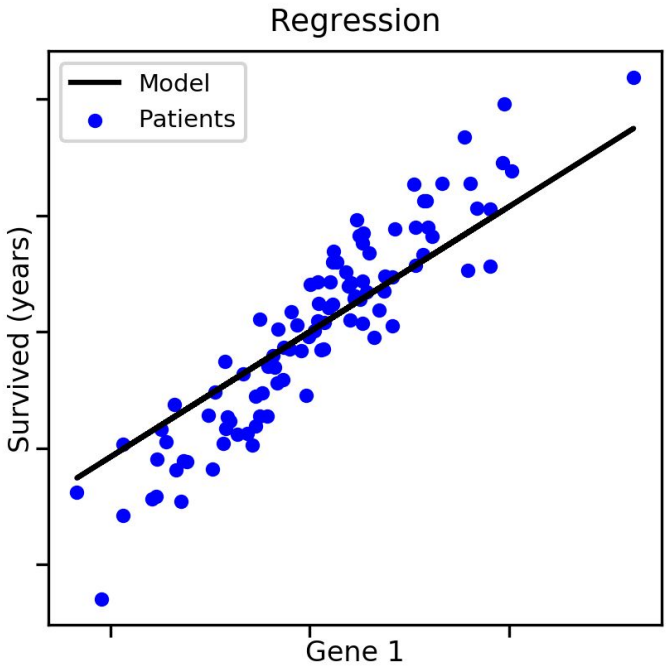
— Andrew Ng, *Machine Learning and AI via Brain simulations*[1]

Machine Learning

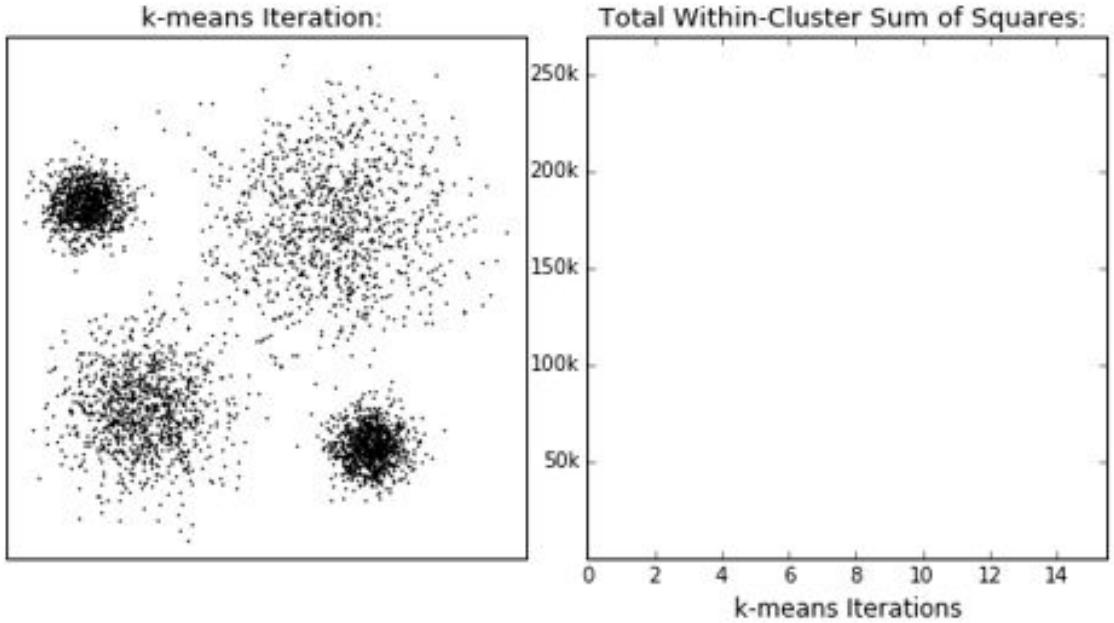


***S**upervised **M**achine **L**earning*

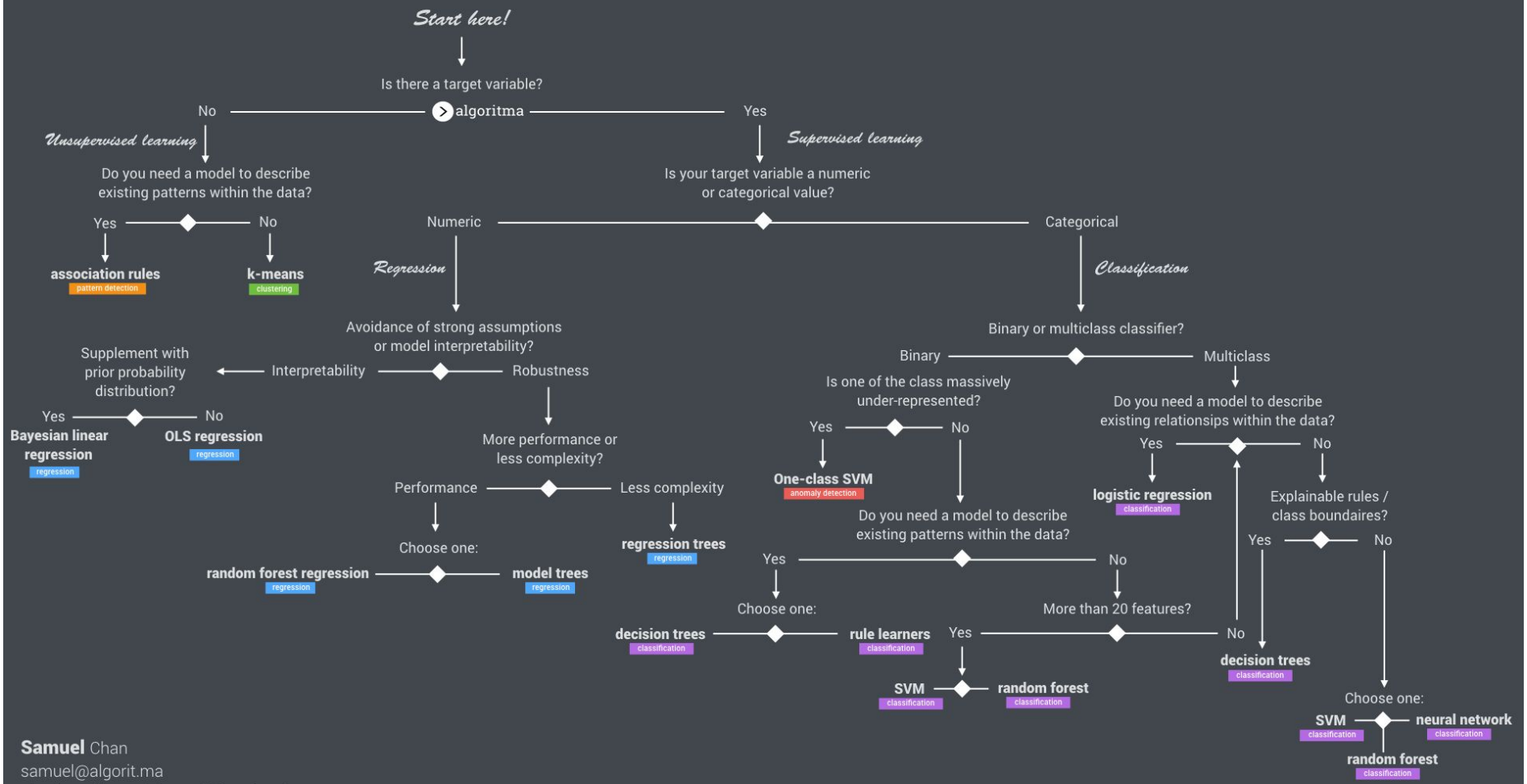




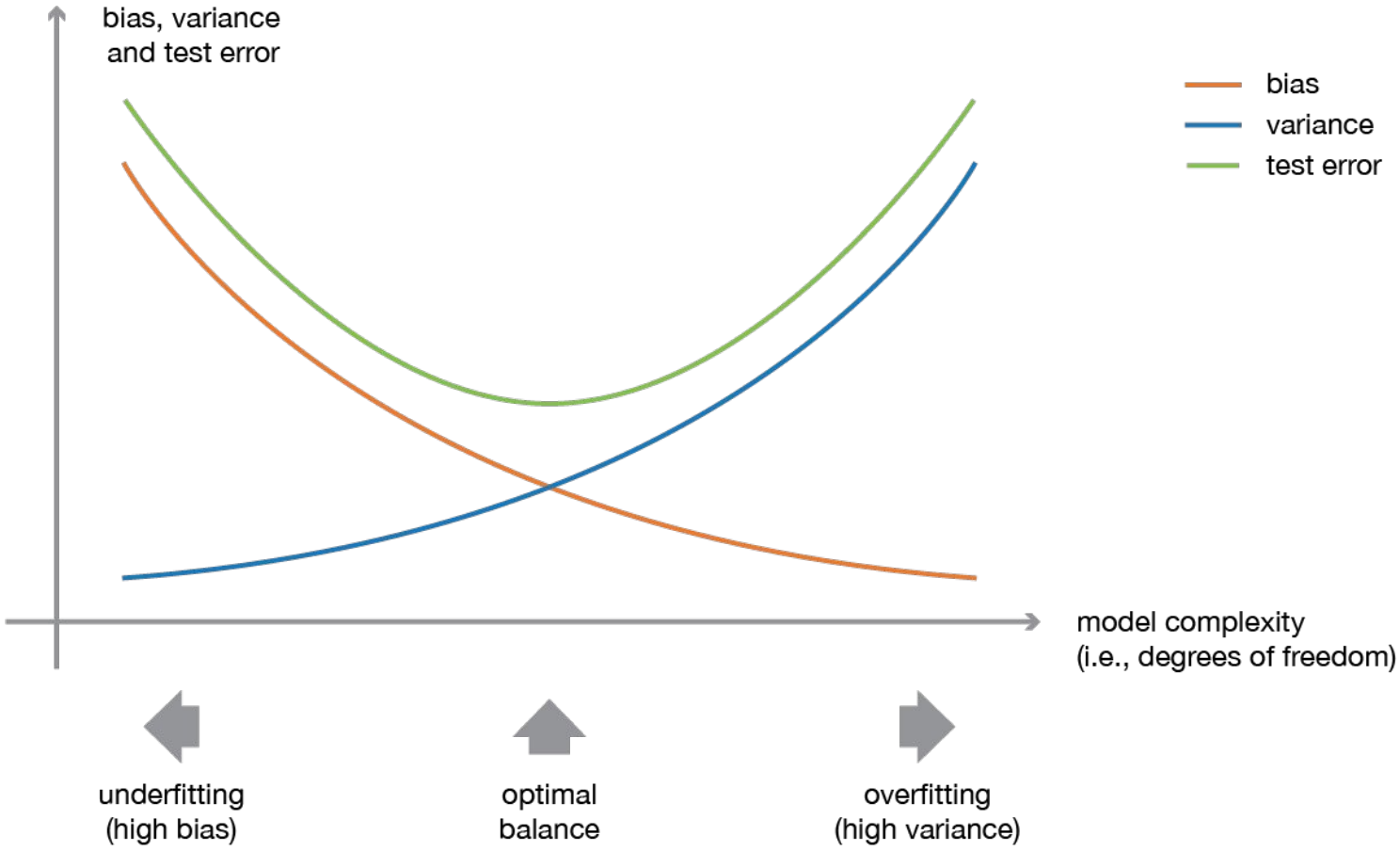
Unsupervised Learning



Bagaimana **C**aranya **S**atu dari **S**ekian **B**anyak **M**odel?

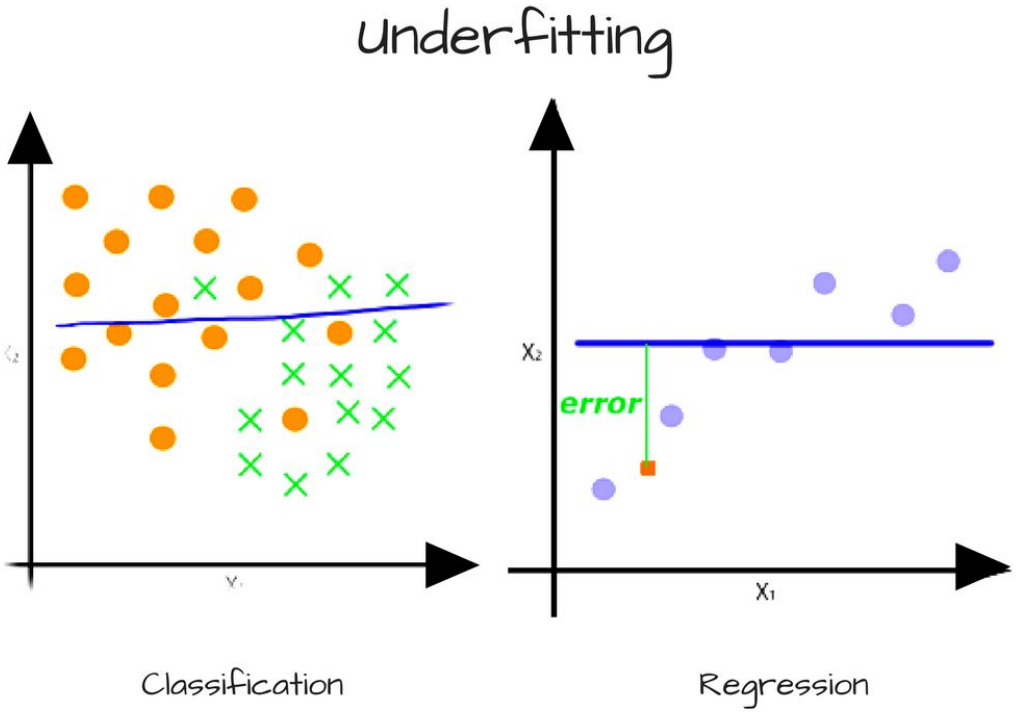


Samuel Chan
samuel@algorit.ma
algoritma | Data Science Education Centre

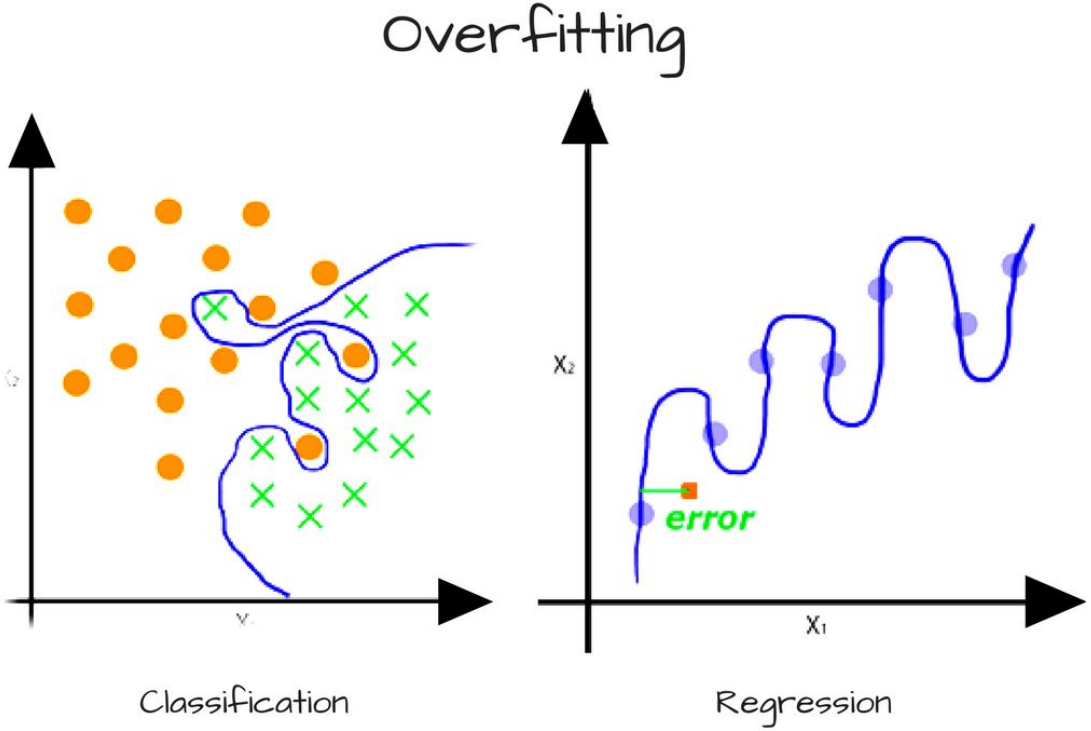

bias, variance
and test error
bias
variance
test error
model complexity
(i.e., degrees of freedom)
underfitting
(high bias)
optimal
balance
overfitting
(high variance)

Evaluasi

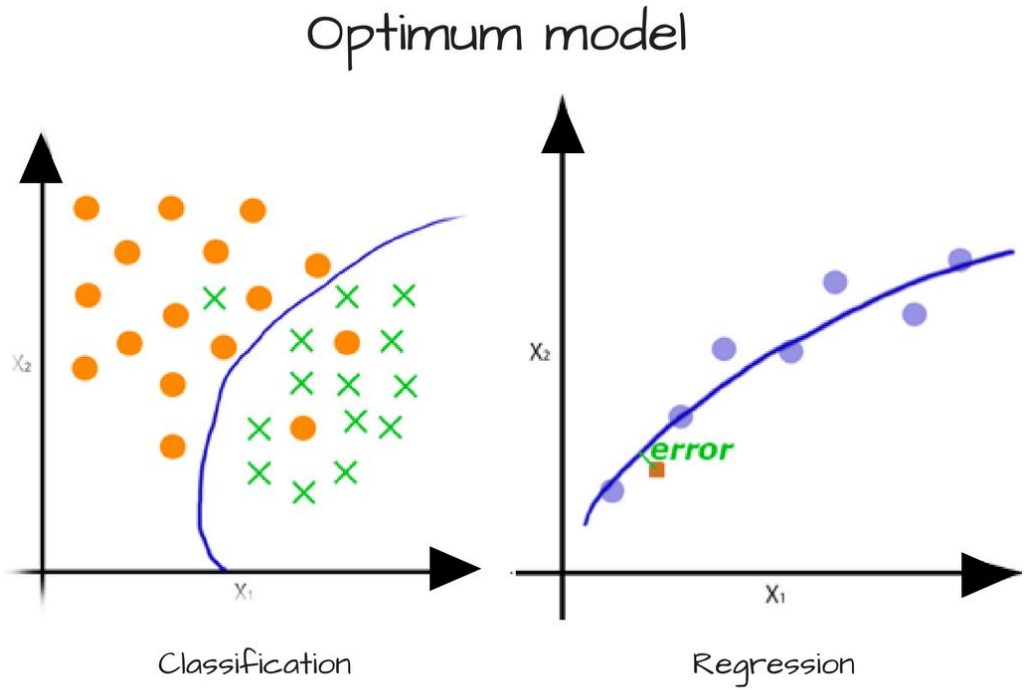
Mengapa **P**erlu **E**valuasi?



Mengapa **P**erlu **E**valuasi?



Mengapa **P**erlu **E**valuasi?



Linear **R**egression

Linear regression equation

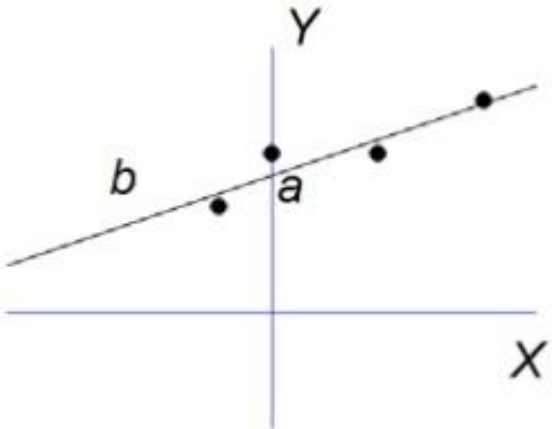
(without error)

$$\hat{Y} = bX + a$$

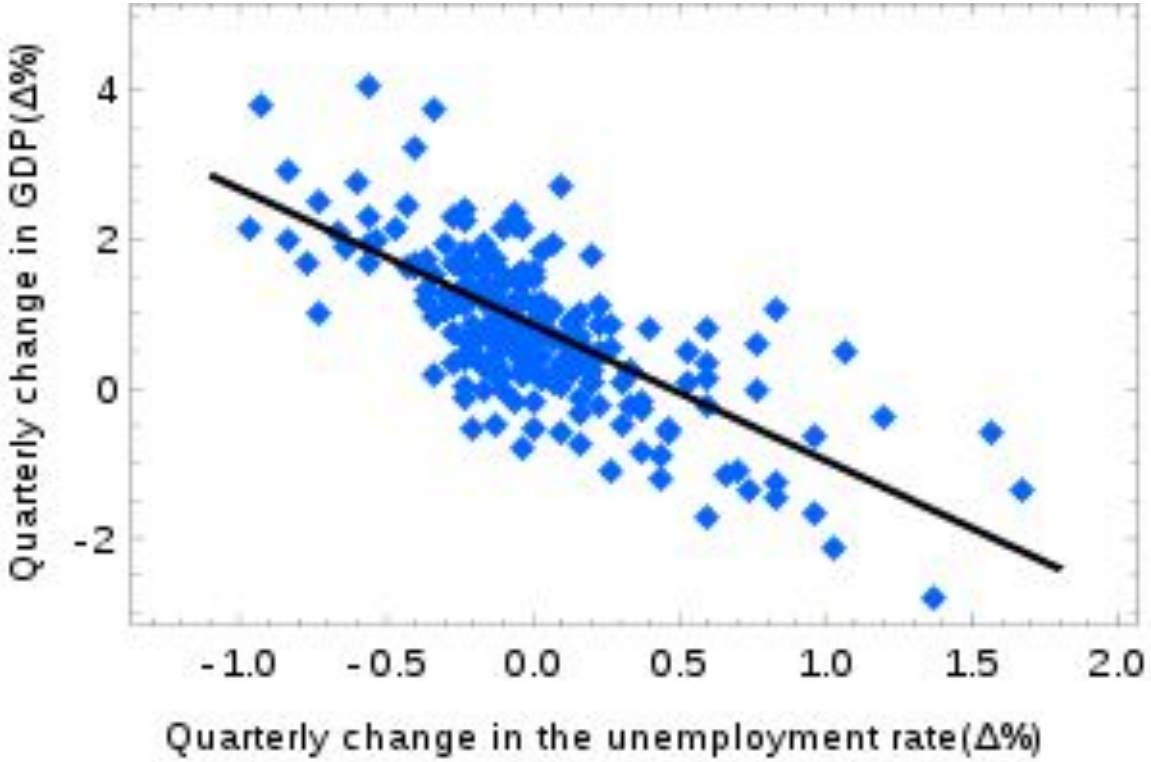
predicted values of *Y*

b = slope = rate of predicted ↑/↓ for Y scores for each unit increase in *X*

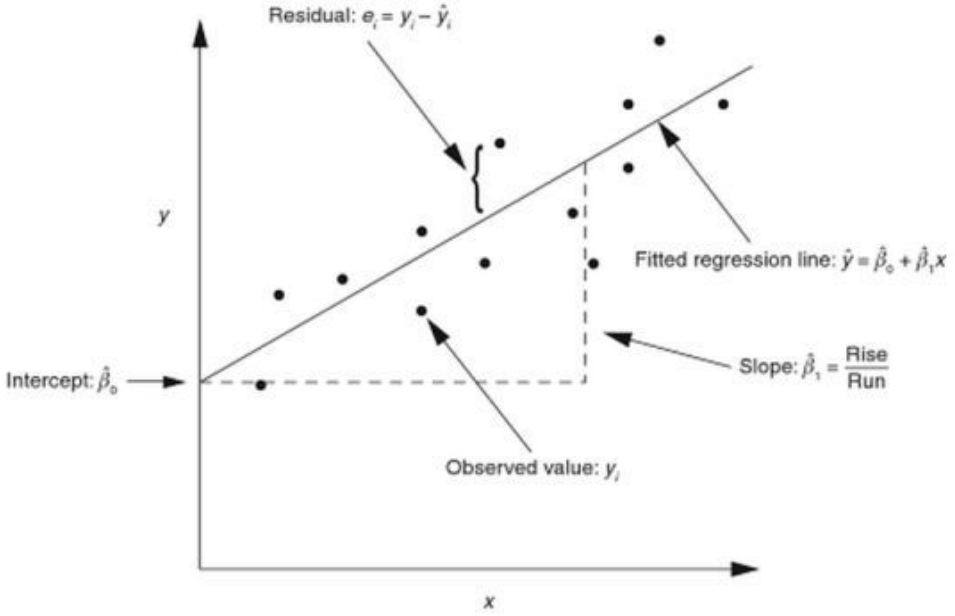
Y-intercept = level of *Y* when X is 0



Contoh **R**egresi
Linear **S**ederhana



Metrik **R**egresi: **R**MSE



$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Smaller is better