

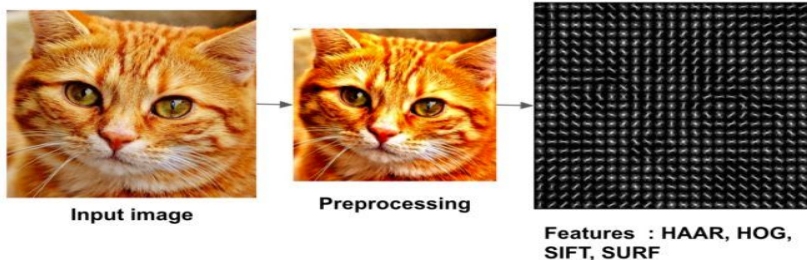
Feature Engineering & Machine Learning 101: Linear Regression

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

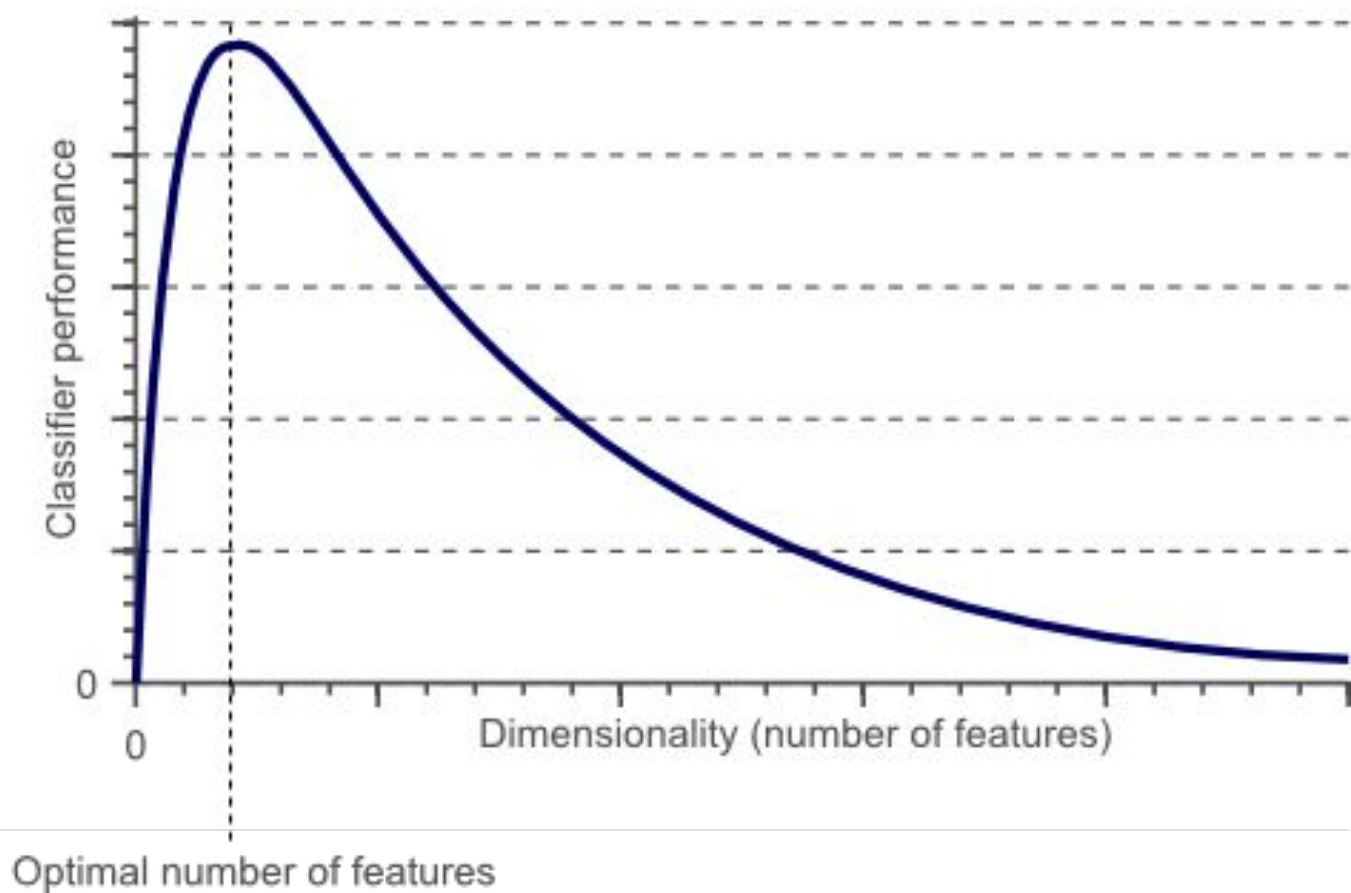


Curse of Dimensionality

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

Encoding classes for sequence labeling

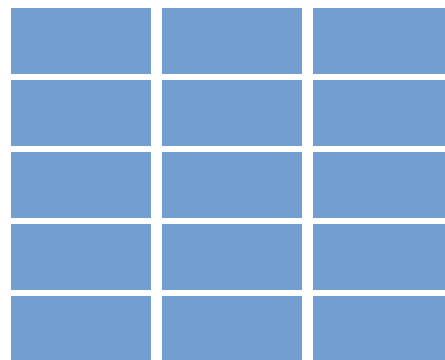
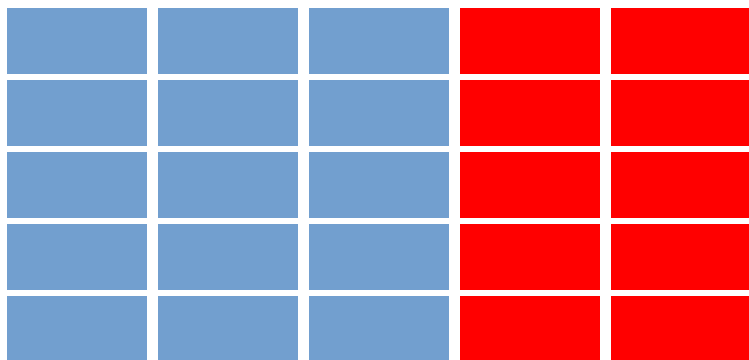
	IO encoding	IOB encoding
Fred	PER	B-PER
showed	O	O
Sue	PER	B-PER
Mengqiu	PER	B-PER
Huang	PER	I-PER
's	O	O
new	O	O
painting	O	O



Cure the Curse

- **Menggunakan *domain knowledge***

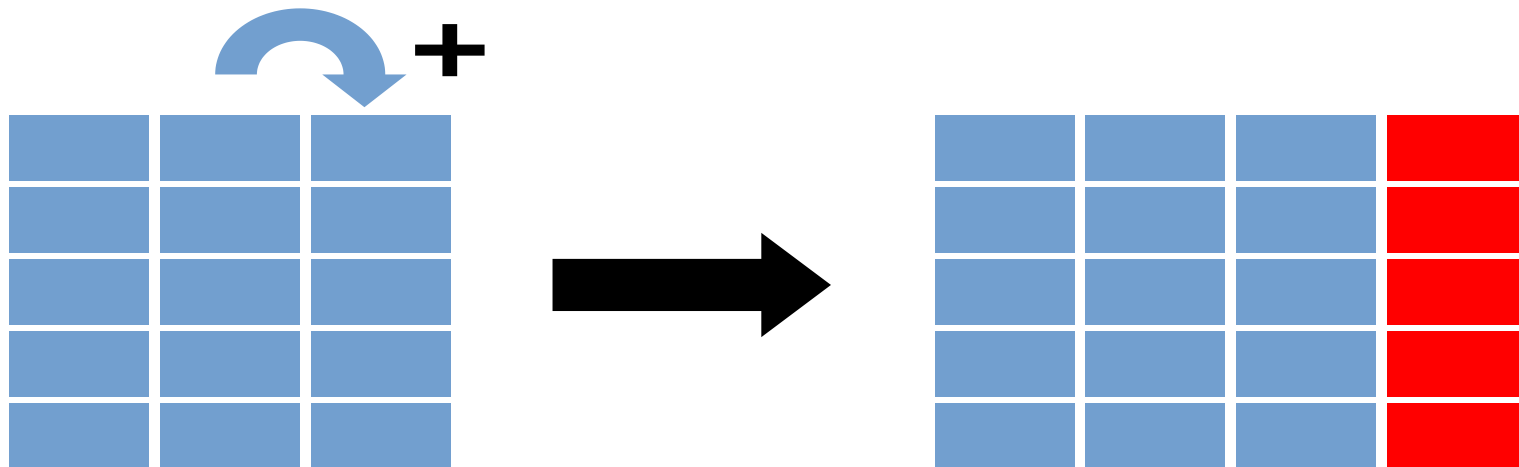
- Seleksi fitur
- *Feature importance*
- *PCA*



Cure the Curse

- **Ekstraksi Fitur**

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

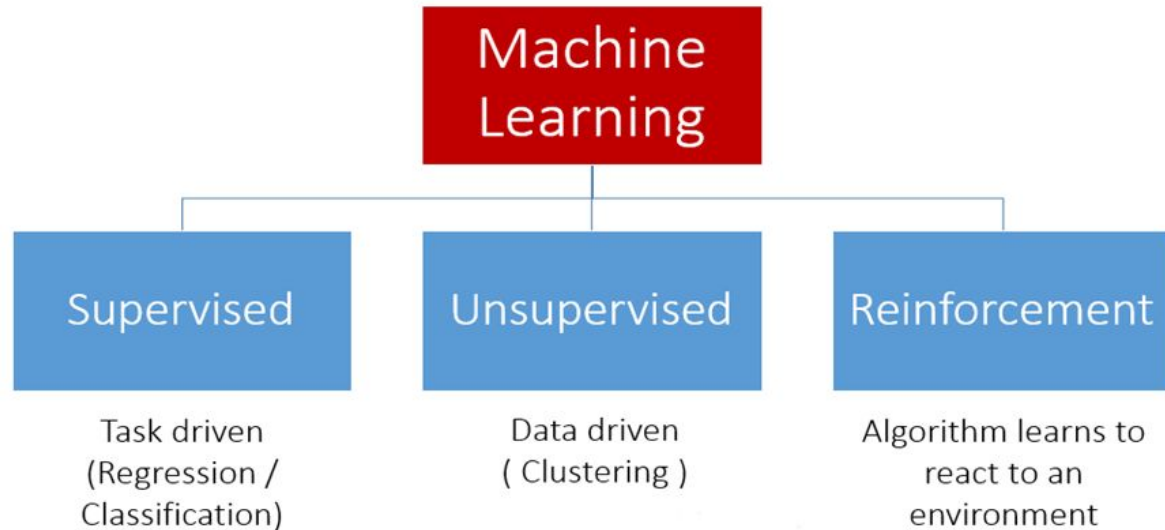


Memperoleh fitur yang baik itu **sulit, membutuhkan waktu serta pengetahuan ahli**. "Pembelajaran Mesin Terapan" merupakan arti lain dari rekayasa fitur.

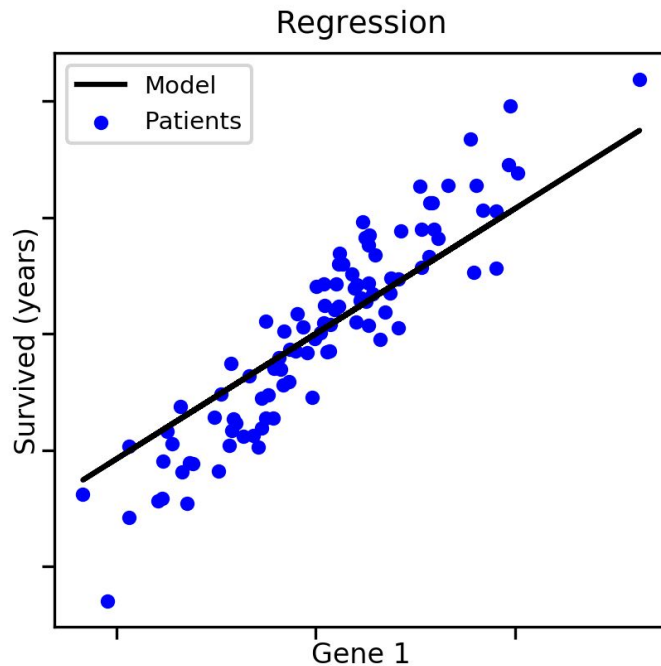
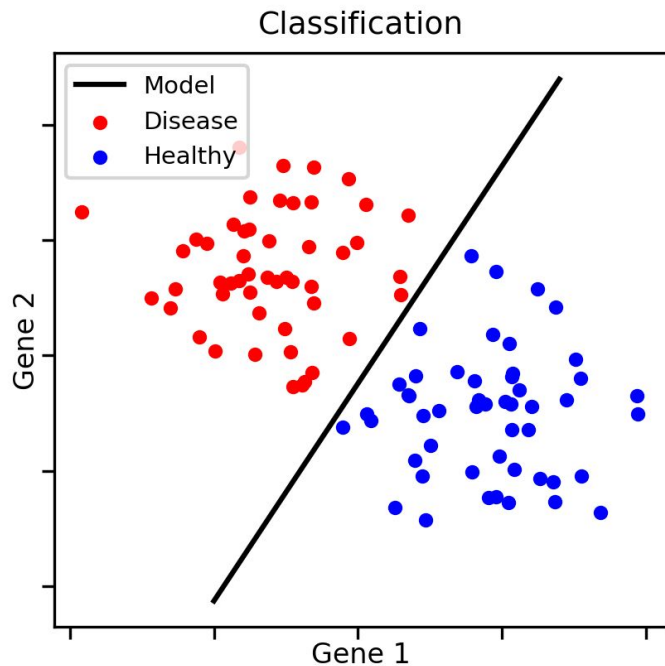
— [Andrew Ng](#), *Machine Learning and AI via Brain simulations*^[1]

Machine Learning

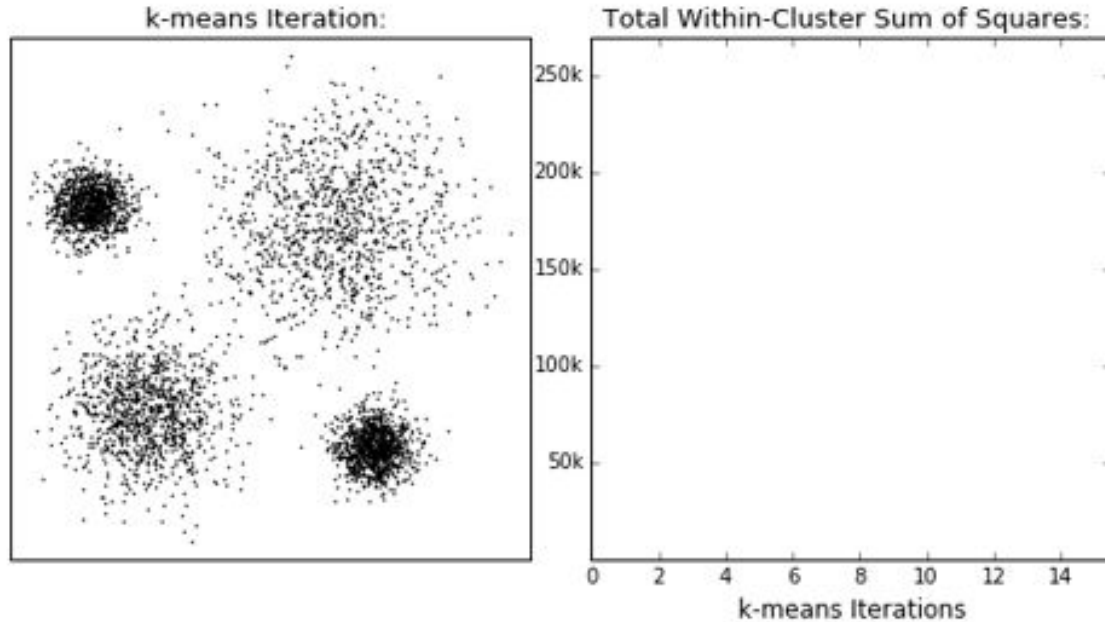
Types of Machine Learning



Supervised Machine Learning



Unsupervised Learning



Bagaimana Caranya Satu dari Sekian Banyak Model?

Start here!

Is there a target variable?



algorithm

Yes

Supervised learning

Is your target variable a numeric or categorical value?

Numeric

Regression

Categorical

Classification

Binary or multiclass classifier?

Binary

Multiclass

Is one of the class massively under-represented?

Yes

One-class SVM
anomaly detection

No

Do you need a model to describe existing patterns within the data?

Yes

Choose one:

decision trees
classification

rule learners
classification

SVM
classification

random forest
classification

No

More than 20 features?

Yes

SVM
classification

random forest
classification

No

Do you need a model to describe existing relationships within the data?

Yes

logistic regression
classification

No

Explainable rules / class boundaries?

Yes

decision trees
classification

No

Choose one:

SVM
classification

neural network
classification

random forest
classification

Do you need a model to describe existing patterns within the data?

Yes

association rules
pattern detection

No

k-means
clustering

Supplement with prior probability distribution?

Yes

Bayesian linear regression
regression

No

OLS regression
regression

Avoidance of strong assumptions or model interpretability?

Interpretability

Robustness

More performance or less complexity?

Performance

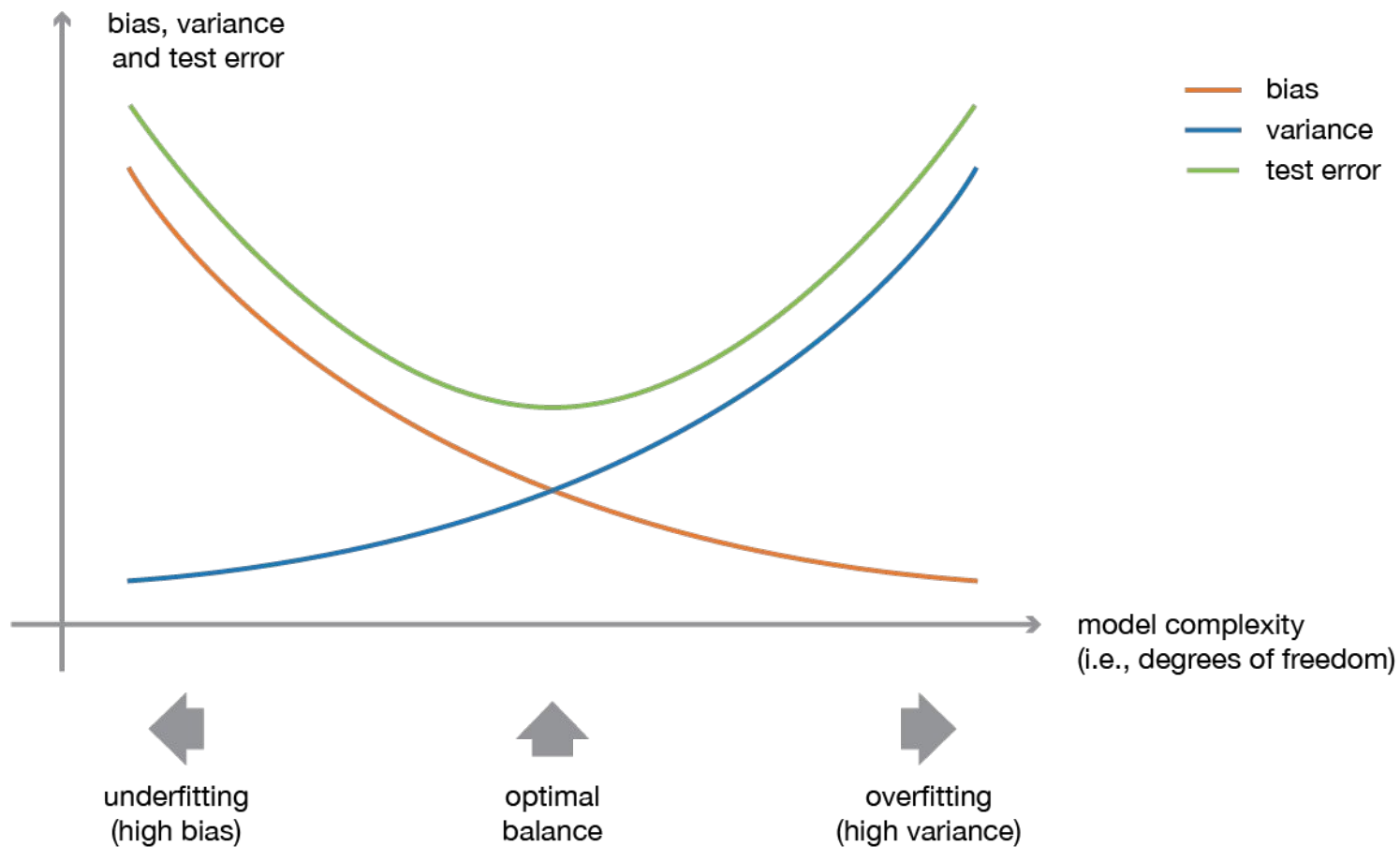
random forest regression
regression

Less complexity

regression trees
regression

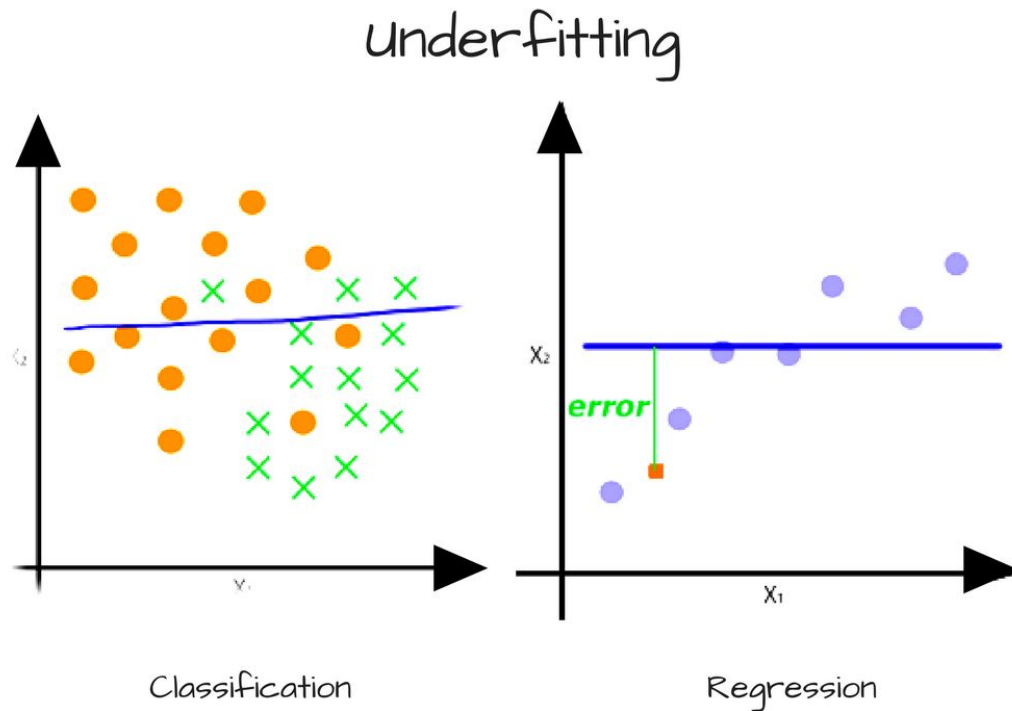
Choose one:

model trees
regression

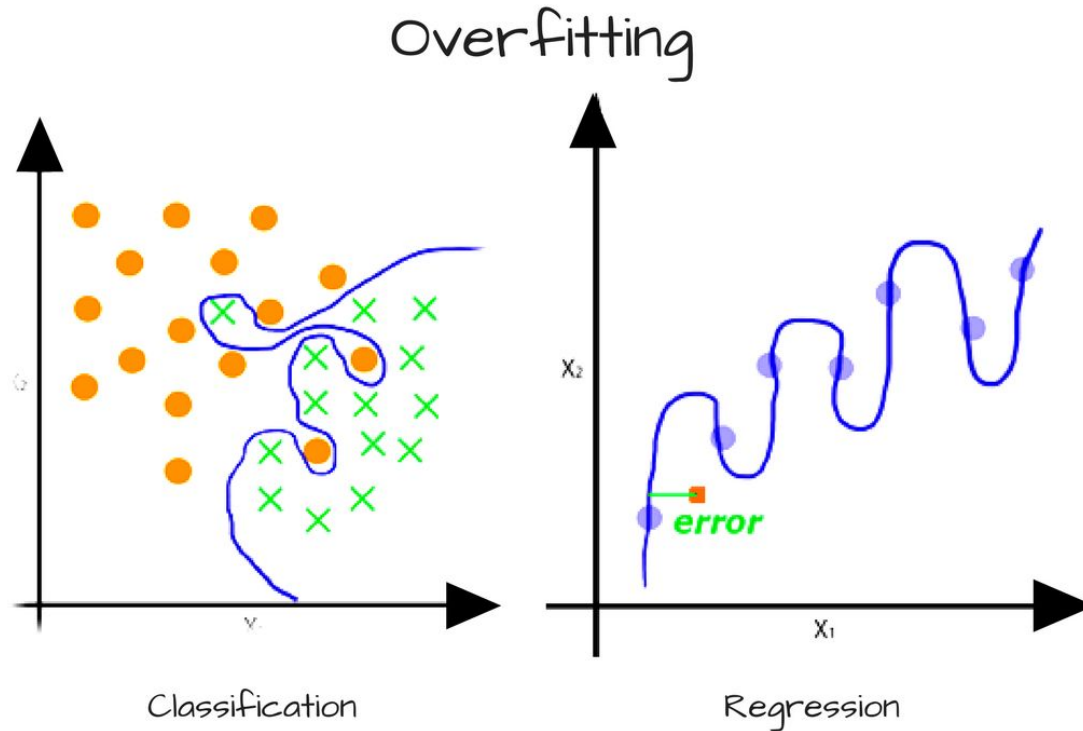


Evaluasi

Mengapa Perlu Evaluasi?

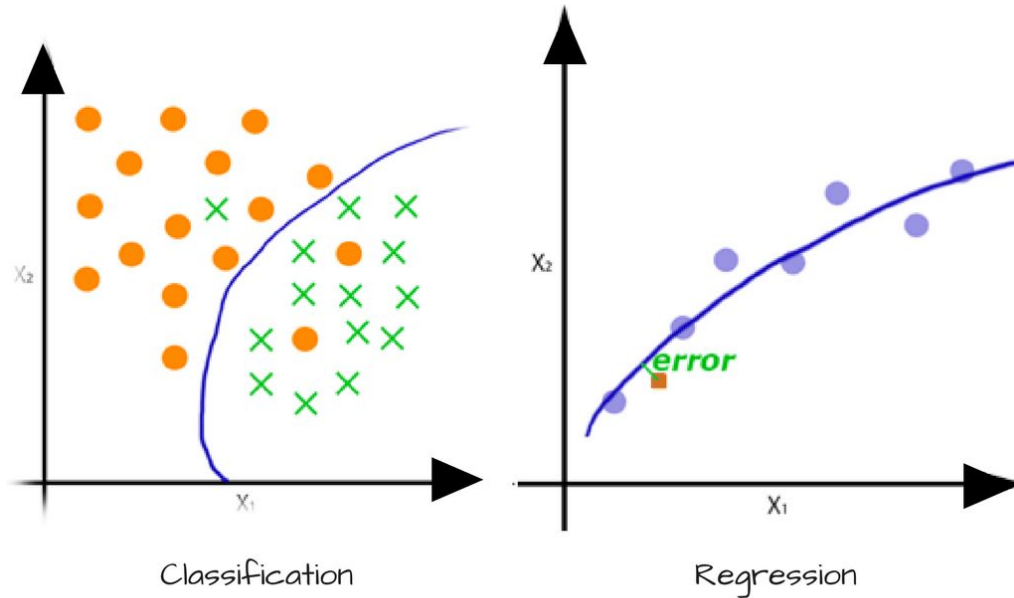


Mengapa Perlu Evaluasi?



Mengapa Perlu Evaluasi?

Optimum model



Linear Regression

Linear regression equation

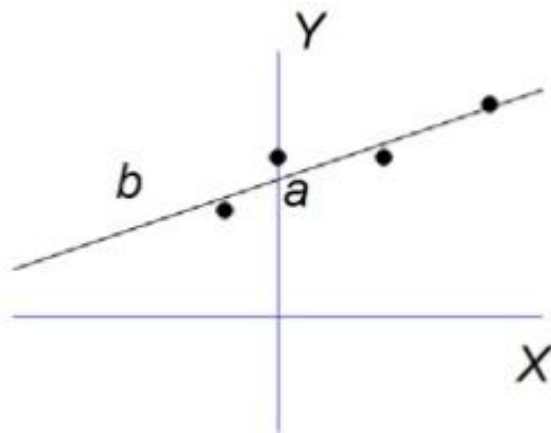
(without error)

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

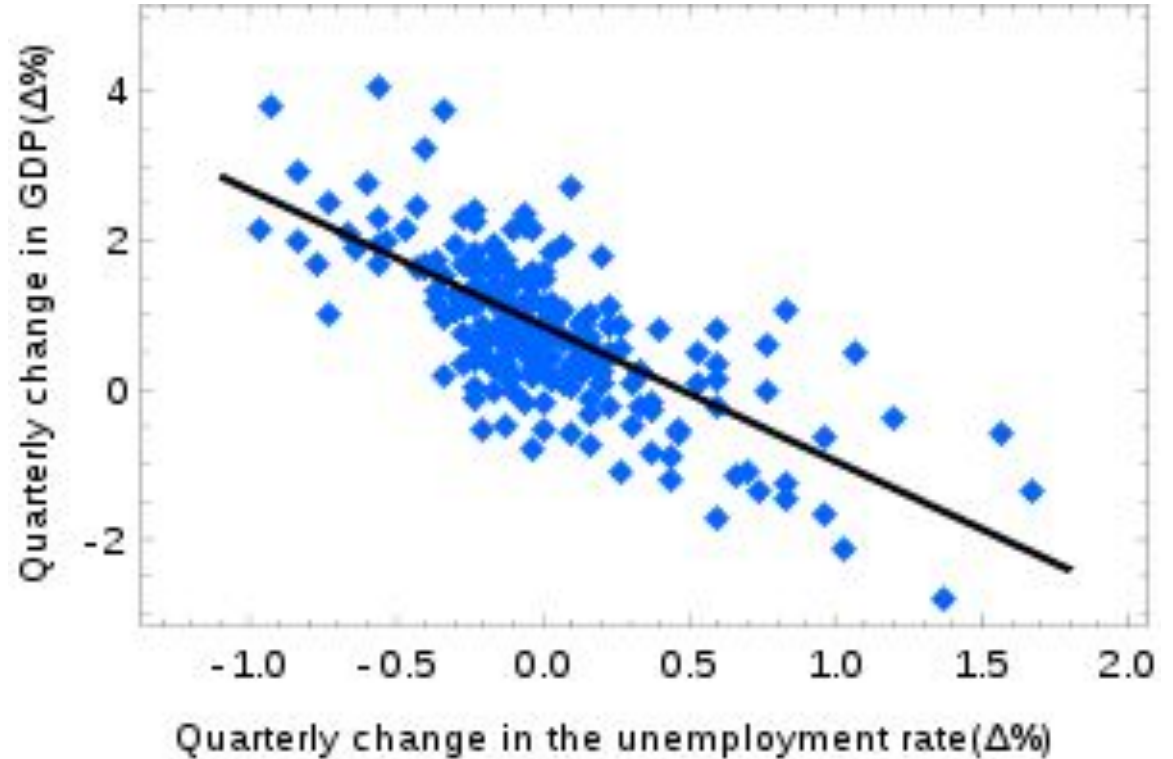
predicted
values of Y

b = slope = rate of
predicted \uparrow/\downarrow for Y
scores for each unit
increase in X

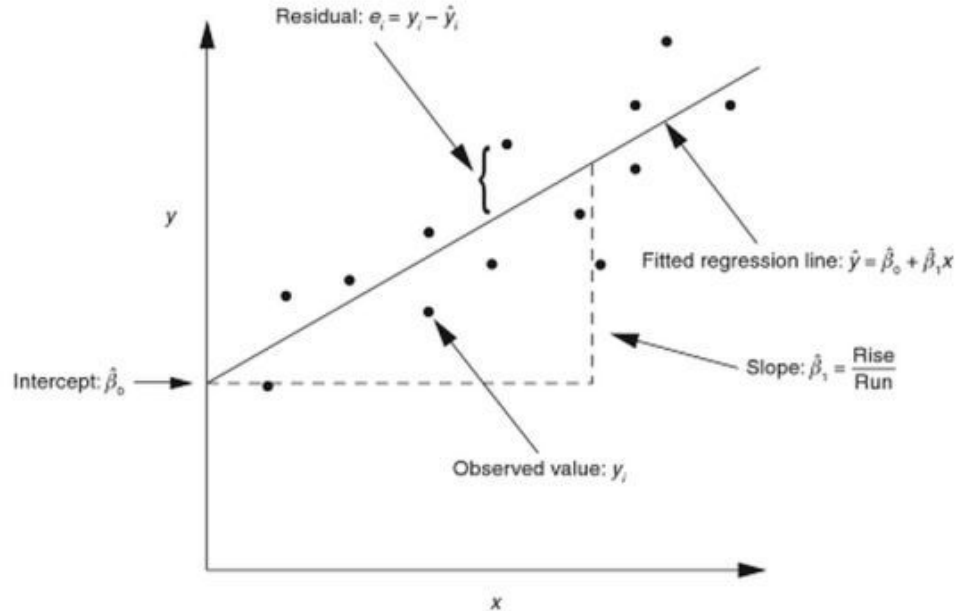
Y -intercept =
level of Y
when X is 0



Contoh Regresi Linear Sederhana



Metrik Regresi: RMSE



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



Smaller is better