Machine Learning

Fundamentals

Pengajar

Alan Nurcahyo

Education:
Akuntansi Pemerintahan **STAN DIII** - 2009
Akuntansi Pemerintahan **STAN DIV** - 2015
Msc. Data Science **American University** Aug 2019 - Dec 2020

Work/internship:
Pelaksana, Biro Perencanaan dan Keuangan
2010 - now
Research Assistant, Lab@DC, OCA, Washington DC *https://thelab.dc.gov/*
Sept 2019 - May 2020
*Data Science Consultant, Seafood globalization Lab http://www.jessicagephart.c*om
Sept 2020 - Dec 2020

Pengajar

Ahmad Rasis Mardhi

Education:
Administrasi Perpajakan **STAN DIII** - 2007
Akuntansi Pemerintahan **STAN DIV** - 2013
Master of Professional Accounting **The University of Sydney** Jul 2016 - Jul 2018

Work/internship:
Pelaksana, Direktorat Jenderal Pajak
2007 - now
Accounting Tutor, International House, Sydney
Mar 2017 - Jul 2018

outline

- What, why, how, example and some fundamental concepts ...
- Basic methods: Linear/Logistics, Decision tree, clustering, Ensemble methods ...

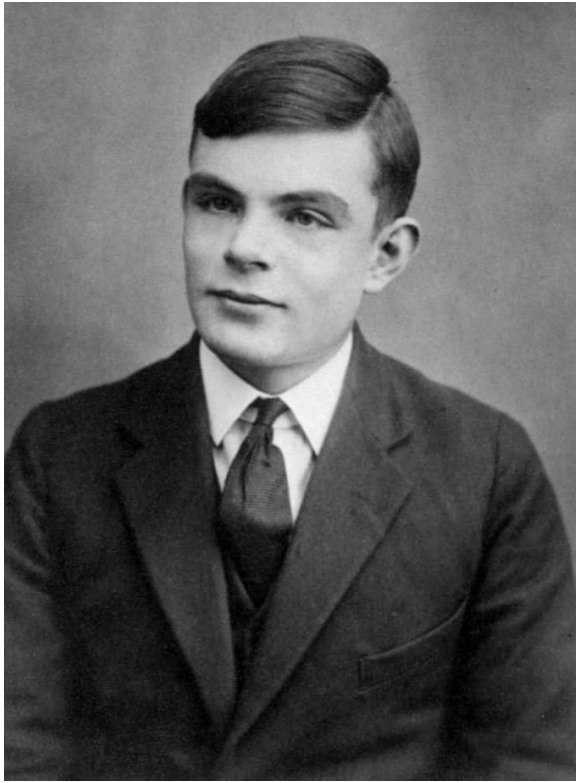
Why study Machine Learning?

- Growing **volumes** and **varieties** of available data, computational processing that is cheaper and more powerful, and affordable data storage leads to Bigger more complex data;
- This data is **too much for human** to handle;
- machine learning has been around for several decades. **It was initially shunned due to its large computational requirements** and **the limitations of computing power** present at the time. However, machine learning has seen a revival in recent years due to the **information explosion.**
- Machine Learning made it's possible to quickly and automatically produce models that can **analyze** this data and deliver **faster, more accurate results** – even on a **very large scale.**

What is Machine Learning?

Dalam sebuah paper sekitar 1950, Turing istilah machine learning pertama kali muncul:

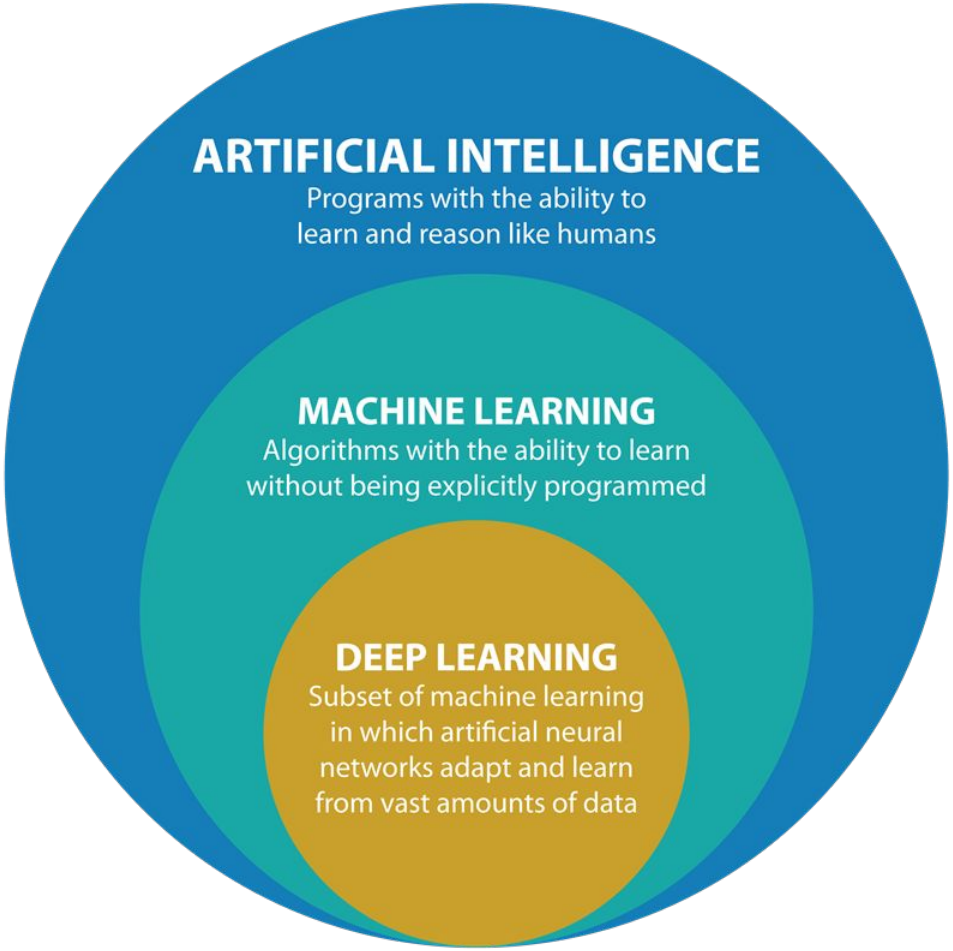
“Machine Learning is an application of artificial intelligence where a computer/machine learns from the past experiences (input data) and makes future predictions. The performance of such a system should be at least human level.”



What is Machine Learning?

Arthur Samuel (1959) mendeskripsikan machine learning sebagai :

“the study that gives computers the ability to learn without being explicitly programmed.”


ARTIFICIAL INTELLIGENCE
Programs with the ability to
learn and reason like humans
MACHINE LEARNING
Algorithms with the ability to learn
without being explicitly programmed
DEEP LEARNING
Subset of machine learning
in which artificial neural
networks adapt and learn
from vast amounts of data

What is Machine Learning?

Tom Mitchell (1998):

"a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

What is Machine Learning?

- Mengerjakan sesuatu (Task- T) :
- Yang diukur berdasarkan suatu metrics (Performance - P) :
- Berdasarkan data (experience - E) :

What is Machine Learning? - example

“Machine Learning untuk mengenali angka pada plat nomor kendaraan“

- (Task- T) : “ mengenali dan mengkategorikan angka dari 0-9”
- (Performance - P) : “persentase angka yang diklasifikasikan dengan tepat (akurasi)
- (experience - E) : “dataset plat nomor kendaraan yang berisi kumpulan angka dan huruf”

Easier example...

Would you like a coffe?

Francise	Price	Cozy	Like
Starbuck	50	8	
Warkop	5	4	
Reman	15	6	
Janji Jiwa	20	6	
Kenangan	25	7	

Exercise! --- P/T/E or None?

“Program komputer yang mengklasifikasikan makanan seafood dan bukan seafood”

A. Jumlah makanan yang dengan benar diklasifikasikan sebagai seafood
B. Mengubah daftar menu menjadi matrix/angka
C. Mengklasifikasikan label makanan sebagai seafood atau bukan seafood
D. Download daftar makanan dari internet
E. Dataset berisi makanan yang telah dilabeli seafood dan bukan seafood

Machine Learning kategori

Supervised Learning: machine learning mempelajari dataset yang memiliki label sebagai target (E) untuk melakukan prediksi atau klasifikasi (T)

Unsupervised Learning: machine learning mempelajari dataset yang tidak memiliki label sebagai target (E) untuk melihat/mempelajari pola yang ada (T)

Reinforcement learning: Machine learning mempelajari data yang diproduksi dari hasil simulasi (E) untuk mencapai tujuan yang ditetapkan (T)

Machine Learning kategori- Supervised

- Machine learning mempelajari dataset yang memiliki label (y~ x)
- Data set dapat memiliki banyak variable (y, x1, x2, ... xn)
- Label/target tidak harus berupa angka

Machine Learning kategori- Supervised

- Prediksi supervised learning :

a. Regresi : Jika label/target berupa angka (numeric/ continuous); memprediksi nilai/jumlah.

    Contoh : toko es krim memprediksi biaya listrik setiap bulannya; Asuransi memprediksi nilai klaim yang akan diminta oleh pengguna asuransi

b. Klasifikasi: Jika label/target berupa kategori; memprediksi kelompok/kategori

    Contoh : google memprediksi apakah email spam atau bukan spam, Youtube memprediksi apakah video melanggar copyright atau tidak

Machine Learning kategori- Supervised

- Naive Bayes
- Linear Regression
- Logistic Regression
- Decision Tree
- Support Vector Machine
- Neural Network.. etc

Supervised - aplikasi

Example Classification:

Face Recognition

Character Recognition

Speech Recognition

Medical Diagnosis

Web Advertising

Example Regression:

Price Prediction

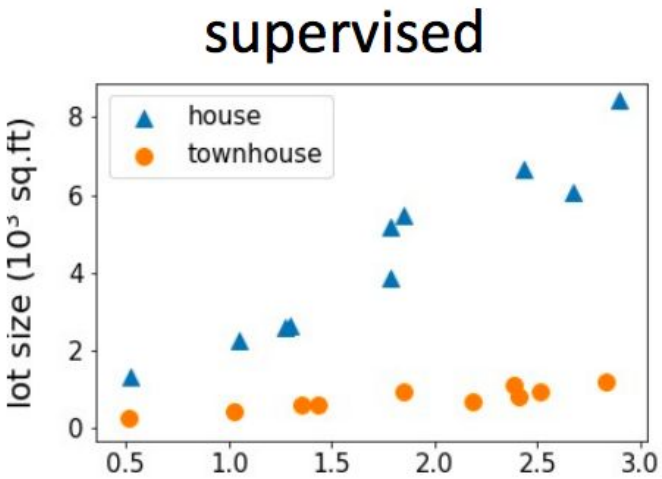
Navigating Car (angle of steering wheel - CMY Nav Lab)

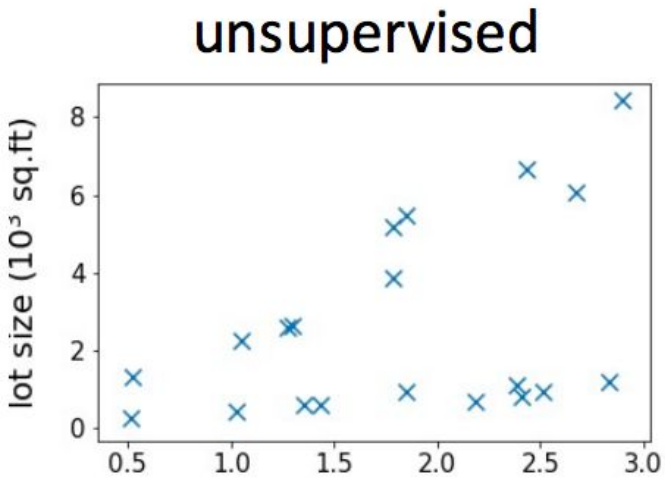
Kinetic Robot Arm

Machine Learning kategori- Unsupervised

Dataset tidak memiliki label (x1, x2, ... xn)

Bertujuan untuk menemukan pola



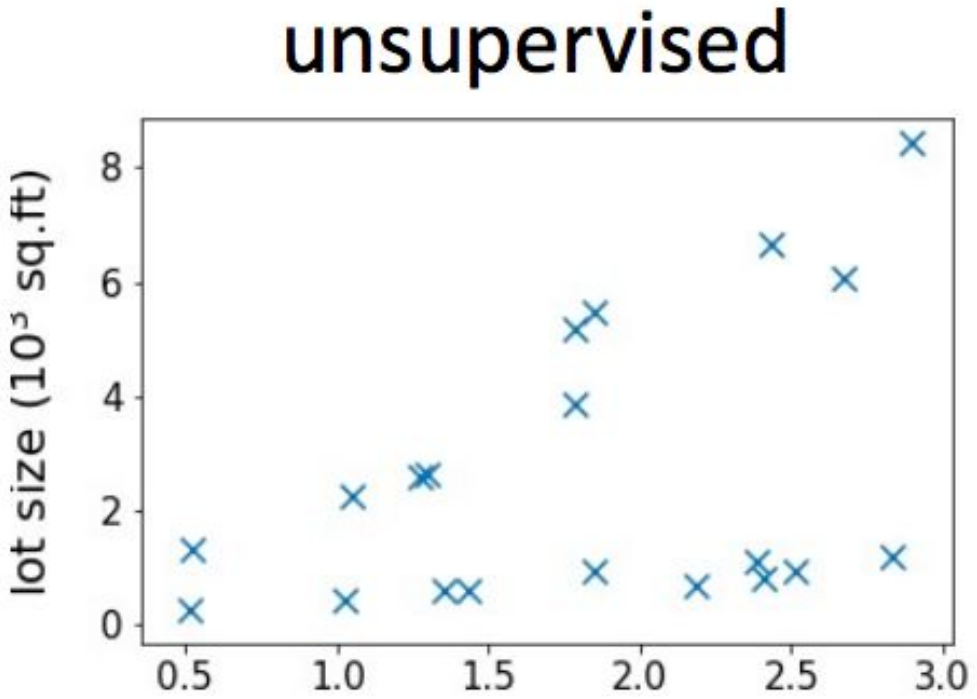


Machine Learning kategori- Unsupervised

Clustering adalah salah satu contoh unsupervised learning:

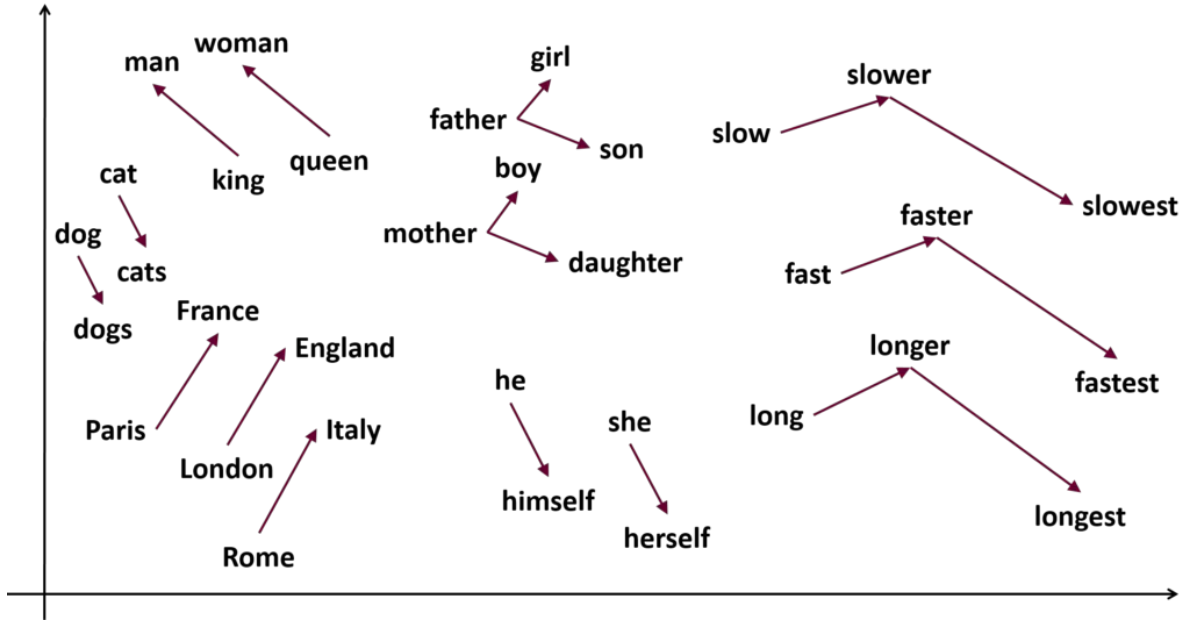
- K-means segmentasi customer,

- Principle Component Analysis (PCA) untuk mengurangi jumlah data



Machine Learning kategori- Unsupervised

Contoh lain dari unsupervised adalah word embedding. Ex: word2vector mencari pola dari kata-kata dalam word repository/Dictionary

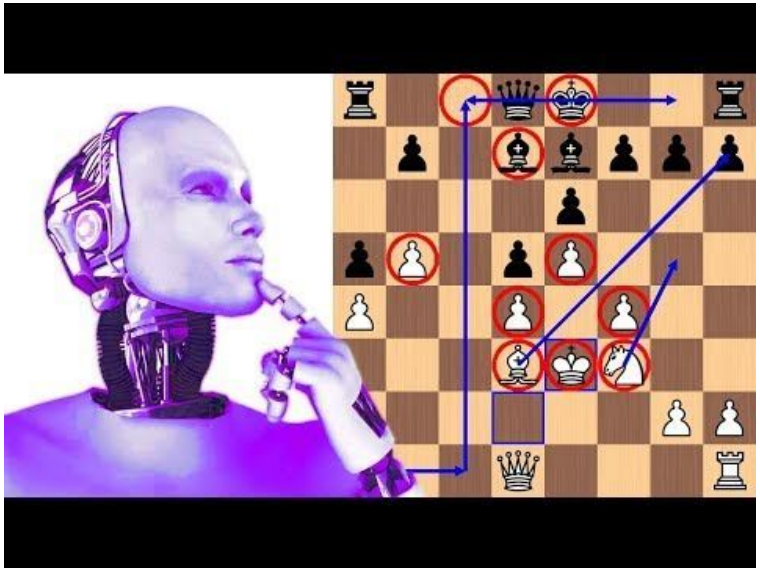


Machine Learning kategori - Reinforced Learning

Machine learning memproduksi data untuk menemukan strategy optimal

Mencoba strategy A -> menyimpan data hasil strategy A -> mencoba strategy B ->menyimpan data hasil strategy B dan memutuskan strategy yang optimal -> mencoba strategy C -> ...dst

contoh: Alpha Zero/Alpha Go, Game AI (tetris),



Exercise!

- KPPN Memprediksi jumlah kas yang ditarik oleh satuan kerja per bulan.
- DJP memprediksi apakah wajib pajak patuh atau tidak.
- DJP mengelompokkan wajib pajak pribadi berdasarkan karakteristiknya.
- KPKNL memprediksi nilai wajar barang milik negara.
- Biro SDM memprediksi pegawai yang akan keluar dari PNS

Alur dasar Machine Learning

1. Mengumpulkan dan mempersiapkan data
2. Memahami data
3. Membagi dataset menjadi 2: training dan test set
4. Menentukan model berdasarkan masalah (T menurut model) dan ukuran keberhasilan model (P)
5. Menjalankan model menggunakan training set (E)
6. Mengevaluasi model menggunakan test (P)
7. Melakukan Prediksi/Deployment

Alur dasar Machine Learning

1. Mengumpulkan, memahami (EDA) - *Gathering the data*
- Tujuan machine learning - Right Question > petabytes of data
- Mengumpulkan data

2. Mempersiapkan data:
- Cleaning and Preprocessing - Garbage in Garbage out
- This two first steps is where you spend most of your time!!

Alur dasar Machine Learning

Suppose you want to create ML software to predict coffe and tea!

Drinks	Caffein	Color	Coffe?
Drink 1	10	500	0
Drink 2	10	400	0
Drink 3	20	100	1
Drink 4	10	50	1
Drink 5	20	150	0
…….	...		
Drink 99	10	300	0

EDA- Exploratory Data Analisis

Analisis yang dilakukan untuk memahami karakteristik data, menemukan outliers, dan pola-pola di dalam data dalam bentuk graphical dan non-graphical;

Bentuk EDA antara lain:

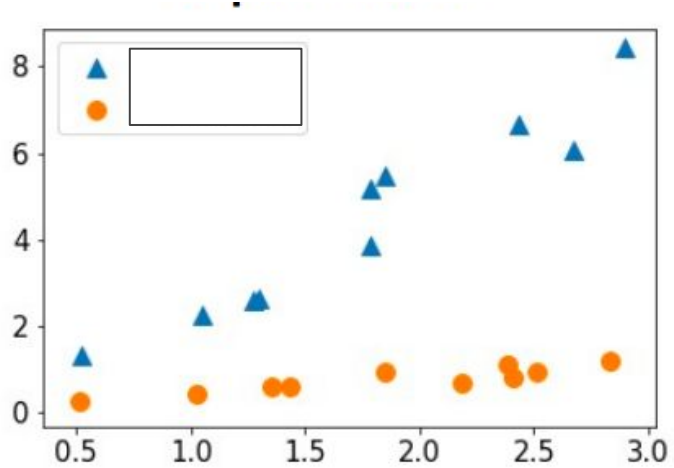
- Univariative: misalnya mean, median, min, max, boxplot, histogram
- Multivariate: korelasi, cross tabulation, scatter plot X dan Y, atau X1 dan X2, dsb
- Clustering: k-means
- Outliers analysis

Output: Is there data problem (ex. Imbalance, multicorrelation)? Can we improve data (ex. Feature engineering)?

Alur dasar Machine Learning

Suppose you want to create ML software to predict coffe and tea!

Drinks	Caffein	Color	Coffe?
Drink 1	10	500	0
Drink 2	10	400	0
Drink 3	20	100	1
Drink 4	10	50	1
Drink 5	20	150	0
…….	...		
Drink 99	10	300	0



Alur dasar Machine Learning

3. Membagi dataset menjadi 2 [umumnya random split]:

- Training akan digunakan oleh model untuk mencari pola (“*learning*” in machine learning)
- Test akan digunakan untuk menguji kebenaran pola yang ditemukan model
- Pembagian training-test umumnya dilakukan secara Acak
- Proporsi umum : 80 persen training : 20 persen testing -> arbitrary
- *More method on this later....*

Alur dasar Machine Learning

Suppose you want to create ML software to predict coffe and tea!

Training	Caffein	Color	Coffe?
Drink 1	10	500	0
Drink 4	10	400	0
Drink 6	20	100	1
Drink 13	10	50	1
Drink 15	20	150	0
…….	...		
Drink 89	10	300	0

Test	Caffein	Color	Coffe?
Drink 1	10	500	0
Drink 2	10	400	0
Drink 3	20	100	1
Drink 5	10	50	1
Drink 7	20	150	0
…….	...		
Drink 99	10	300	0

Alur dasar Machine Learning

4. Menentukan model berdasarkan masalah (T menurut model) dan ukuran keberhasilan model (P) :

- Supervised atau unsupervised? Regresi atau klasifikasi?
- Complex (many variables) or simple model (few variables)? (**beware** : bias vs variance trade off)
- Contoh P regressi: MSE, RMSE.
- Contoh P klasifikasi: Akurasi

Alur dasar Machine Learning

- Mengukur Performance Regression problem

  MSE, MAE

Alur dasar Machine Learning

- Mengukur Performance Classification Problem

  Confusion Matrix (FPR, TPR, accuracy, recall, etc)

	Actual -- 1	Actual -- 0
Predicted -- 1	100 (TP)	5 (FP)
Predicted -- 0	10 (FN)	90 (TN)

Alur dasar Machine Learning

	Actual -- 1	Actual -- 0
Predicted -- 1	100 (TP)	5 (FP)
Predicted -- 0	10 (FN)	90 (TN)

True Positive Rate - TPR = TP/P = 100/110 = %

True Negative Rate - TNR = TN/N = = 90/95 = %

False Positive Rate - FPR = FP/N = 5/90 = %

False Negative Rate - FNR = FN/P = 10/110 = %

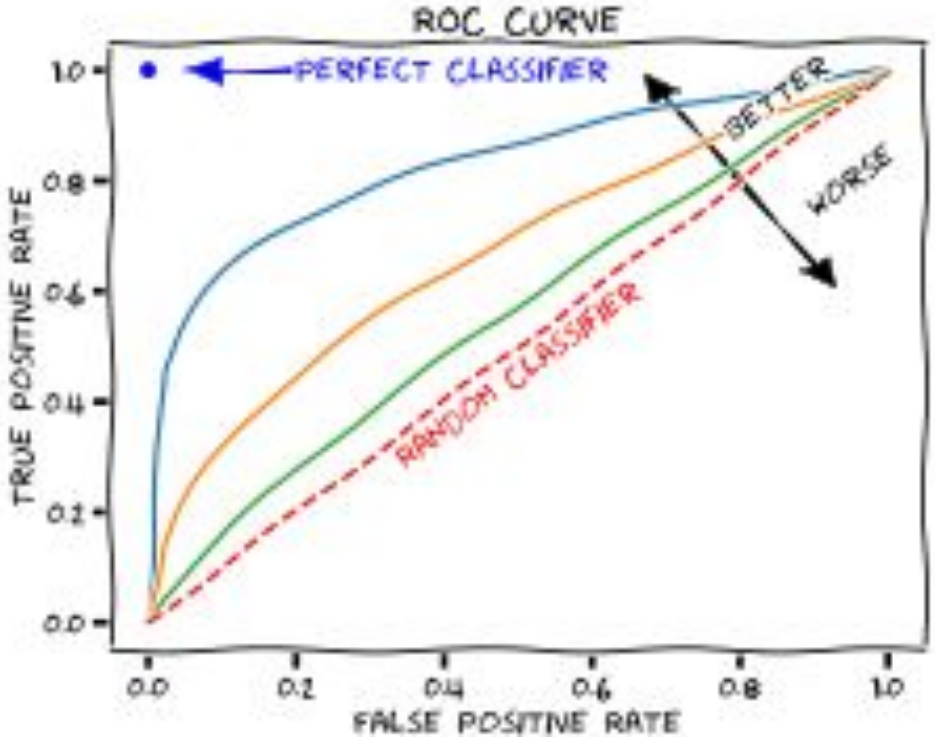
Alur dasar Machine Learning

Accuracy: Nilai yang secara benar diprediksi
(TP+TN) / (P+N)

Precision : nilai positif yang benar diprediksi dari semua yang diprediksi positif
TP/ (TP+FP)

Recall : Nilai positif yang diprediksi benar dari nilai yang benar-benar positif/TPR
TP/P

Alur dasar Machine Learning- ROC Curve



Alur dasar Machine Learning

5. Melakukan training

- Melakukan training untuk menemukan pola.
- Memilih model terbaik berdasarkan training berdasarkan P yang telah ditentukan, untuk selanjutnya di-”test” lagi menggunakan test set.

Alur dasar Machine Learning

6.. Mengevaluasi performa model terhadap test data

- **Underfitting** : Performa buruk pada saat training. Contoh: hasil akurasi 50%.
- **Overfitting** : Performa bagus pada saat training namun buruk pada saat testing
- *More on this later...*

Alur dasar Machine Learning

7.. Predict! (deployment)

Predict drink X is tea or coffe?

Alur dasar Machine Learning

Mengukur Performance Classification Problem: Coffe or Tea?

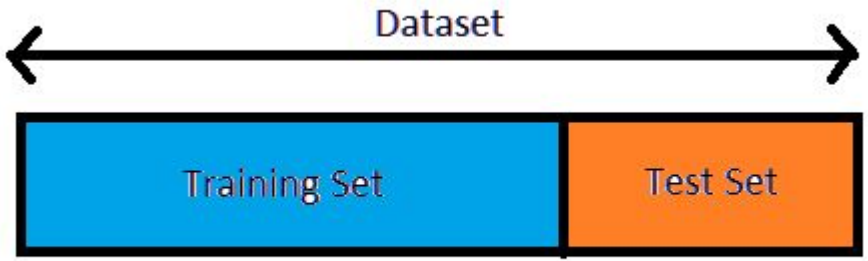
Question: Can our classifier perform worse than coin flip?

Few concept revisit

- *Splitting training and testing - Cross Validation*
- *Bias Variance tradeoff*

Cross validation

Metode validasi model



Terkadang random split tidak cukup:

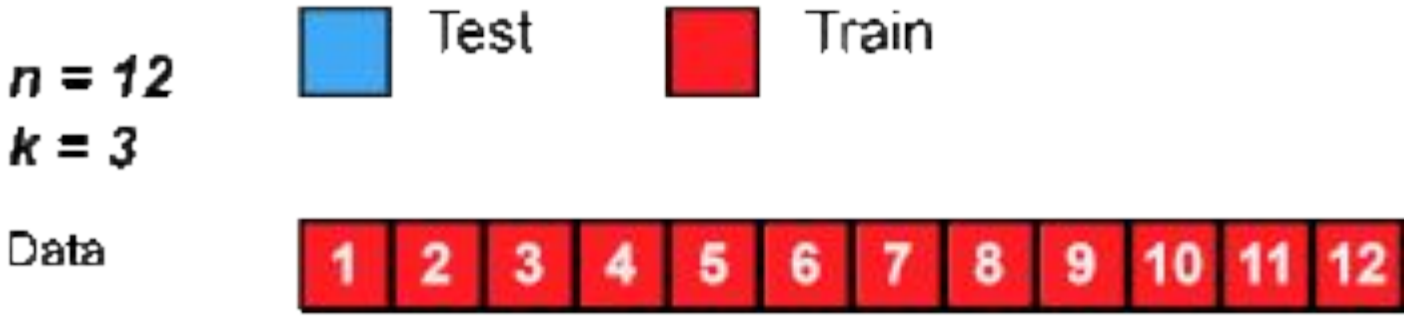
1. Data tidak cocok untuk di split (misalnya time series data)
2. Jumlah training terlalu sedikit untuk di split lagi

*Picture from medium.com

LOOCV- leave one out Cross Validation

*Picture from wikipedia.com

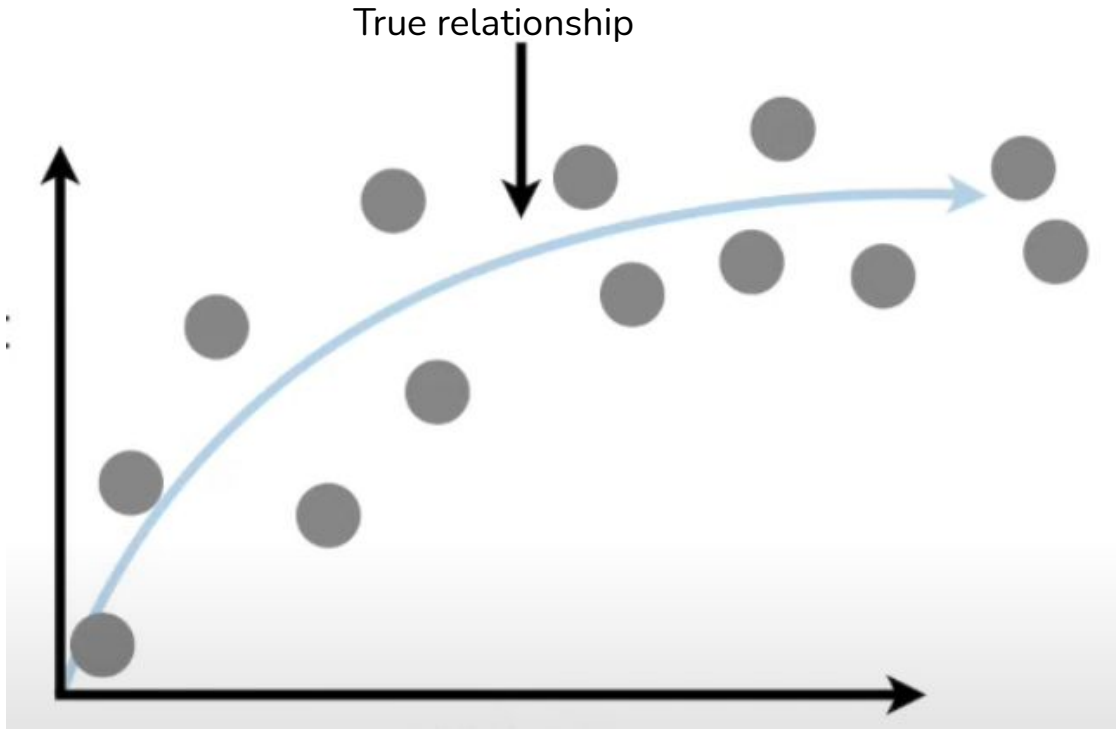
K fold-Cross validation



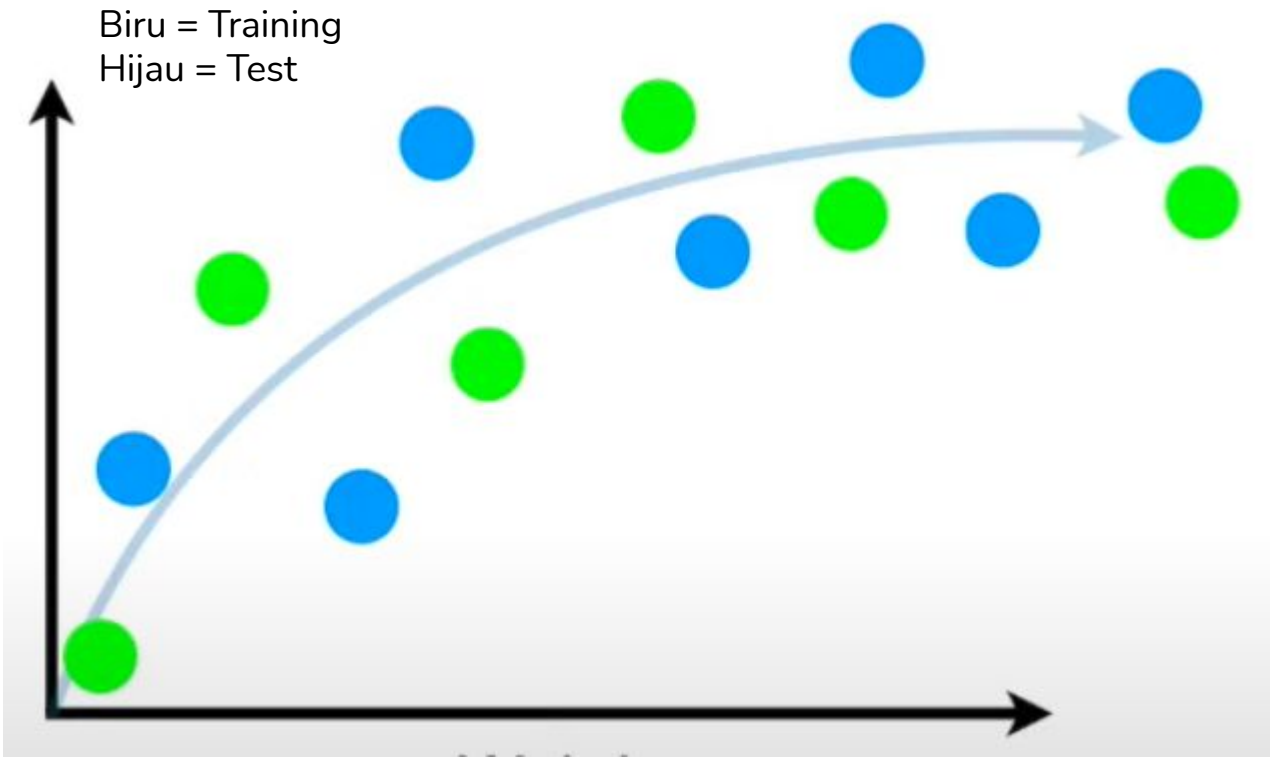
*Picture from wikipedia.com

Bias - Variance Trade off

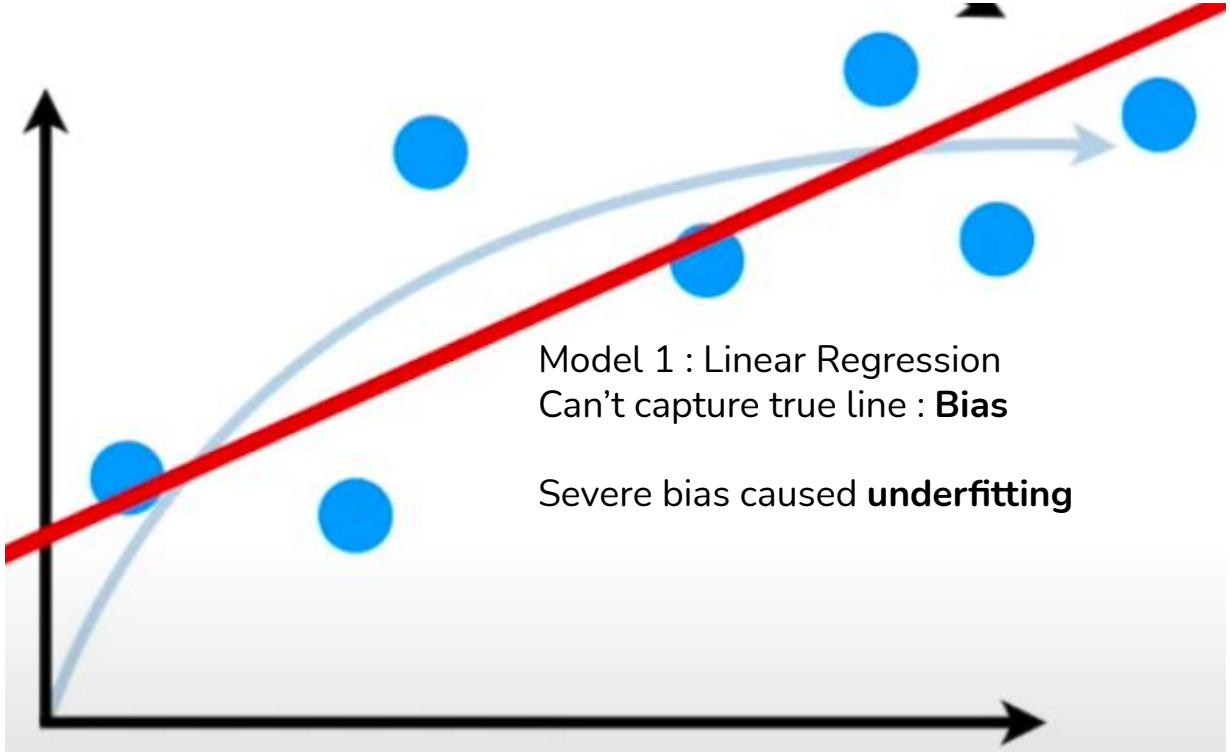
Bias - Variance Trade off



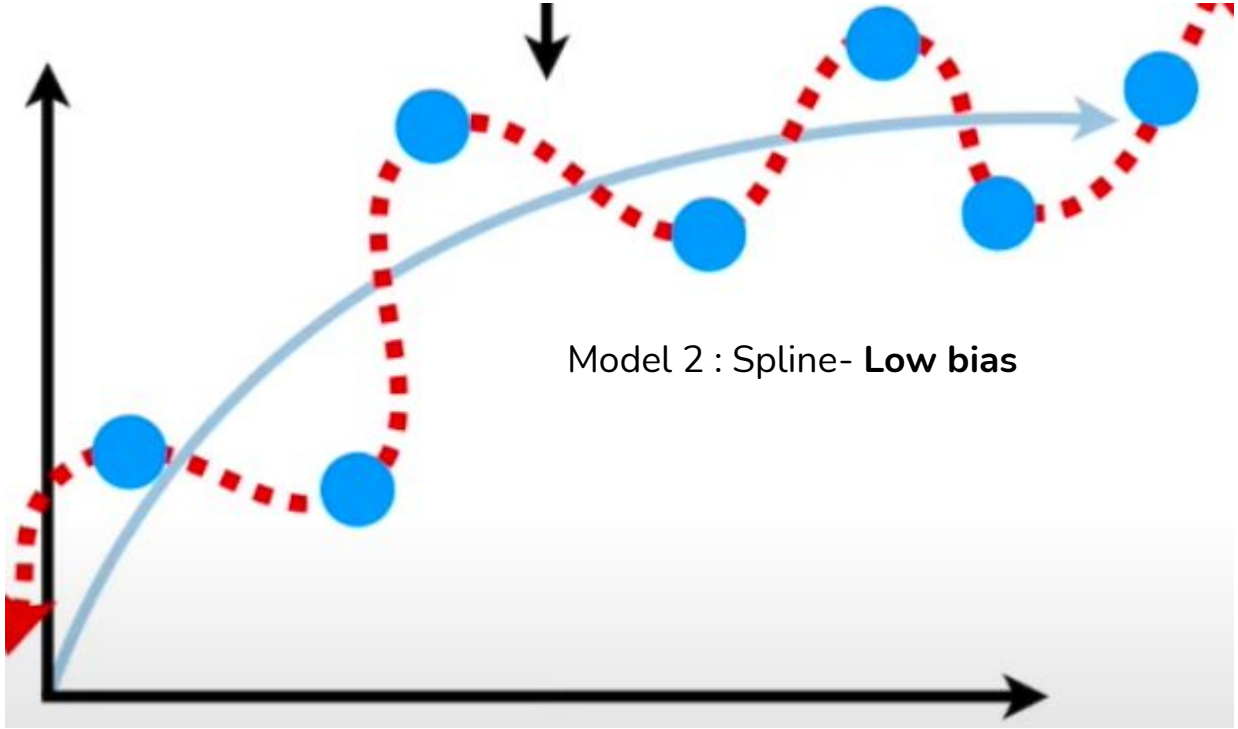
Bias - Variance Trade off



Bias - Variance Trade off

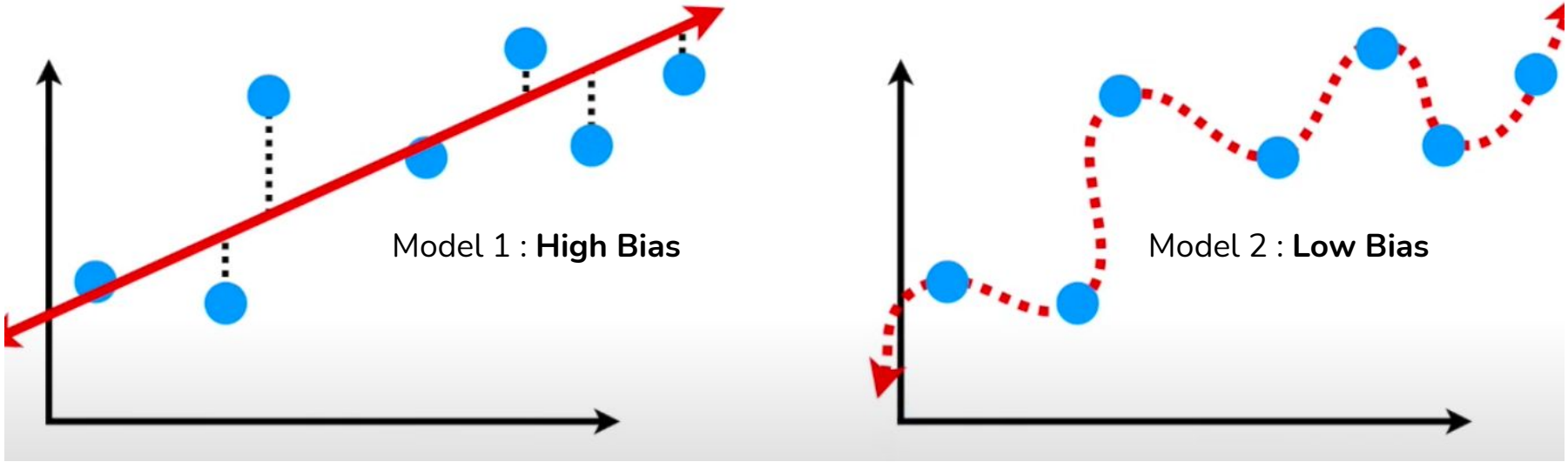


Bias - Variance Trade off

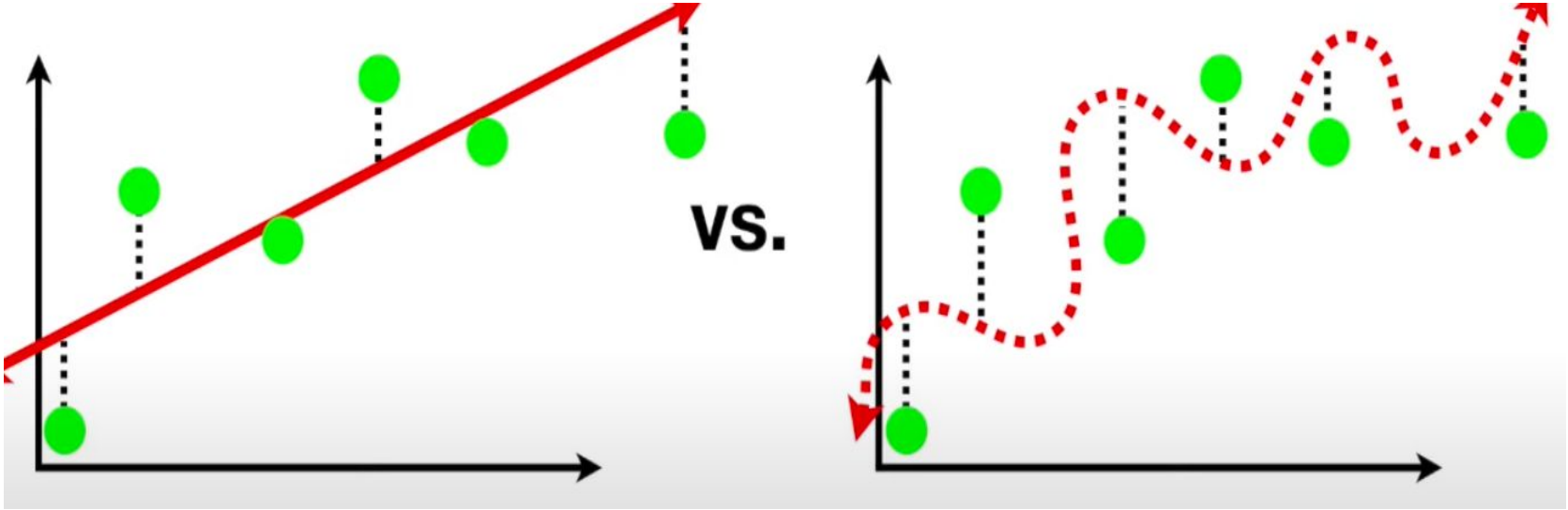


Model 2 : Spline- **Low bias**

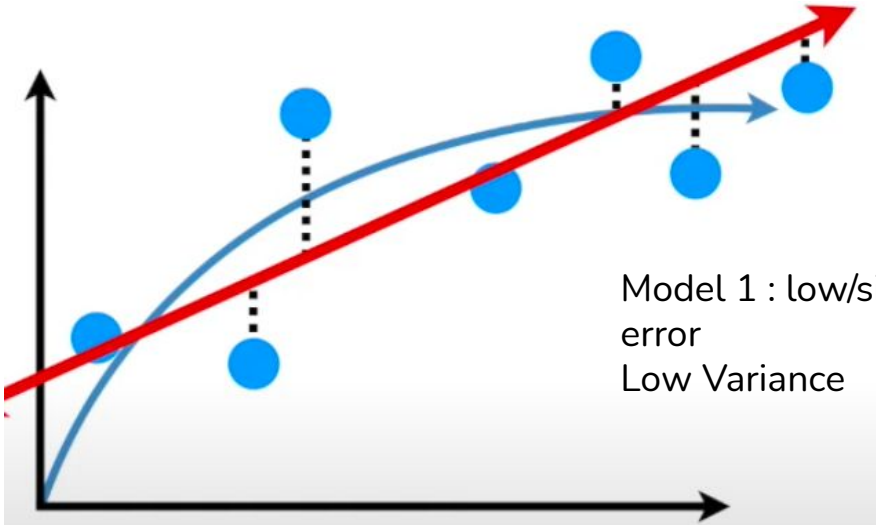
Bias - Variance Trade off



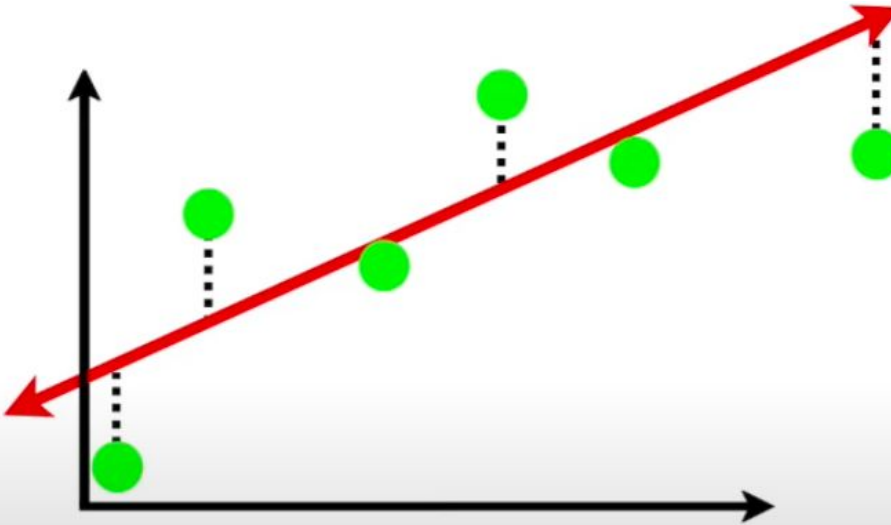
Bias - Variance Trade off



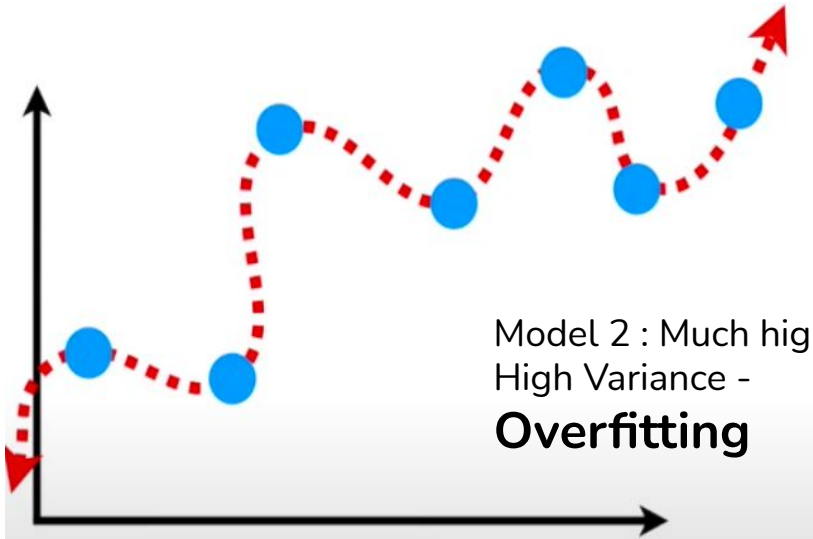
Bias - Variance Trade off



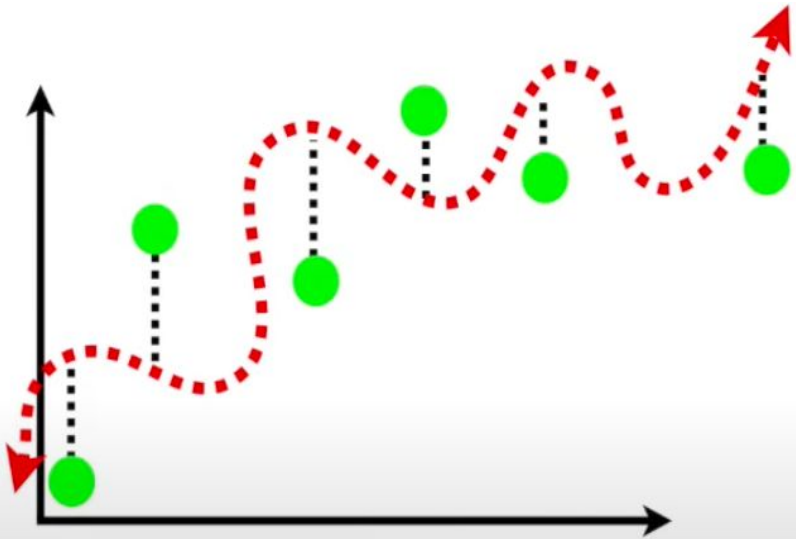
Model 1 : low/similar error
Low Variance



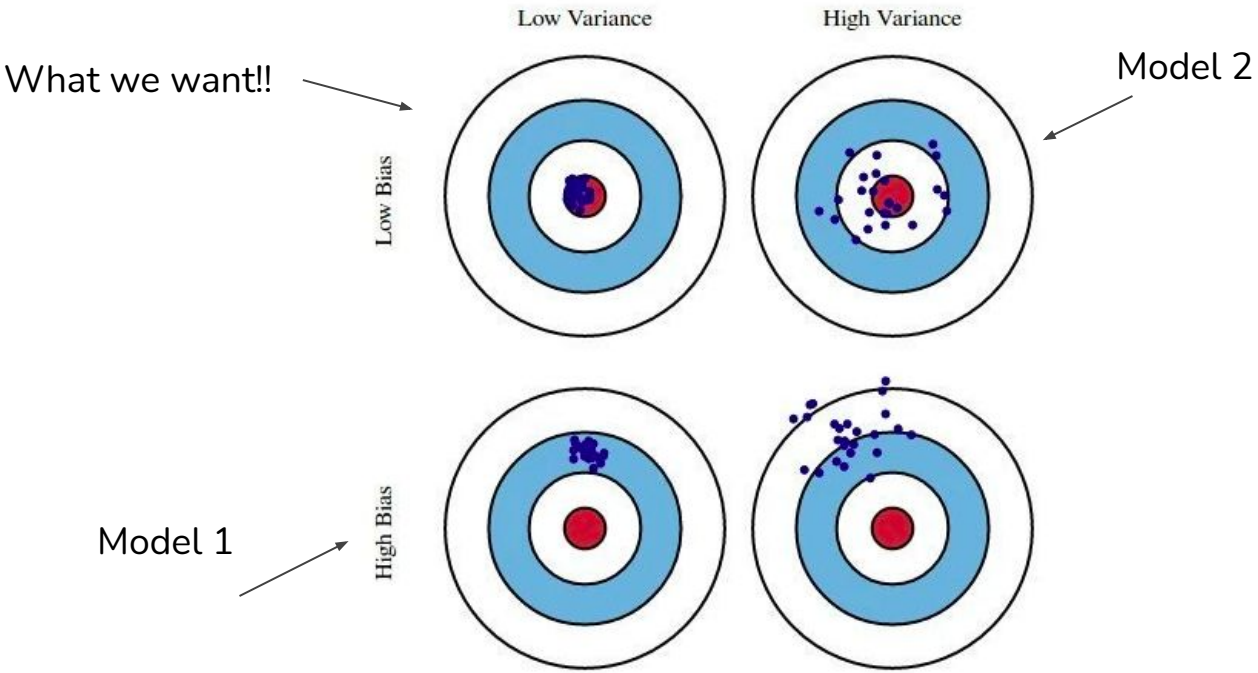
Bias - Variance Trade off



Model 2 : Much higher error
High Variance -
Overfitting

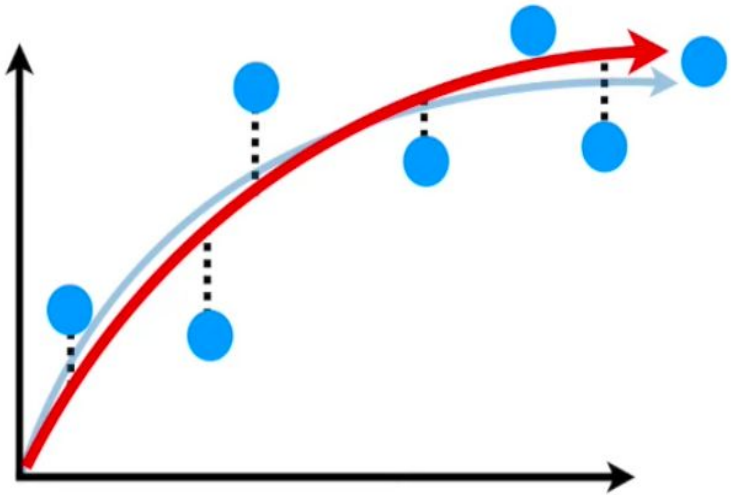


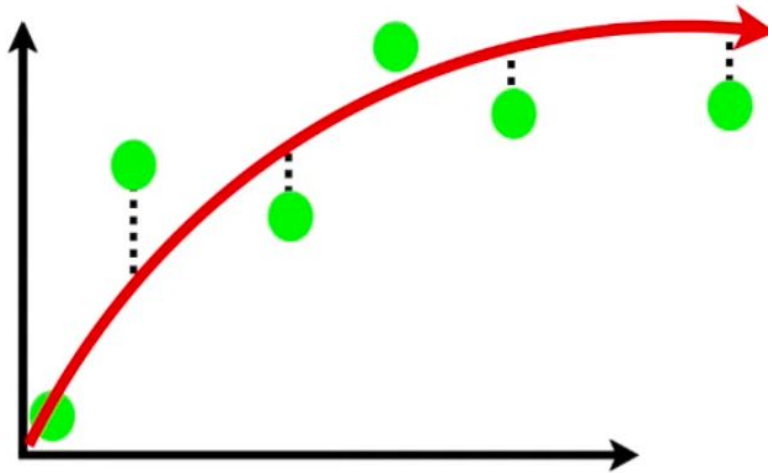
Bias - Variance Trade off



Source: kdnuggets.com

Bias - Variance Trade off





Ideal model: Low bias - low Variance

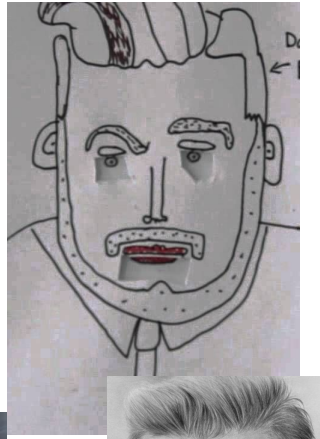
Easier said than Done !

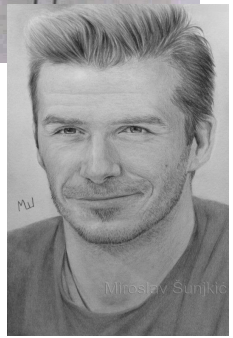
Possible Solution: regularization (Lasso, Ridge), ensemble (bagging/boosting)

Machine Learning is art!

True line

Models





Final note: Machine Learning vs Statistics

*“Perbedaan utama machine learning dan statistik adalah tujuannya. Machine learning didesain untuk membuat **prediksi** seakurat mungkin. Statistik bertujuan untuk **menjelaskan** hubungan atas variabel.”*

Final note: Machine Learning vs Statistics

Contoh: **linear model**

Dalam Statistik linear model didesain untuk dapat menjelaskan hubungan antar variabel menggunakan persamaan linear. **training and test** tidak diperlukan, model dievaluasi dengan signifikansi dan asumsi yang digunakan model.

Dalam Machine learning, linear model digunakan untuk memprediksi outcome dengan persamaan linear. Evaluasi menggunakan training dan test set. *“We don’t actually care about true relationship, we only care because it helps us predict real world outcome.”*

Final note: Machine Learning vs Statistics

Which one is better? Depends on your purpose!

- *I want to explain why this is happen, and maybe make some reasonable prediction: Statistics.*
- *I want to make a good prediction, why is this happening is a second priority: Machine Learning.*