UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Bias-Variance Tradeoff

Aruni Yasmin Azizah*, Adila Alfa Krisnadhi, Siti Aminah,
Dina Chahyati, Fariz Darari

CSGE603130: Kecerdasan Artifisial dan Sains Data Dasar
Gasal 2022/2023

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Outline

1. Motivasi
2. Bias-Variance Tradeoff

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

Motivasi

Sumber:

- Slides Materi Sistem Cerdas, “Machine Learning: Decision Trees”, Semester Genap 2020/2021
- Stuart Russel & Peter Norvig, “Artificial Intelligence: A Modern Approach”, 4th edition, Pearson, 2020

Recall: Supervised Learning

Ide Dasar

Diberikan sebuah training set (example) N berupa pasangan input-output: $(\mathbf{x}_1, y_1),(\mathbf{x}_2, y_2), \ldots ,(\mathbf{x}_N, y_N)$, di mana y_i dihasilkan dari suatu fungsi yang tidak diketahui $y = f(x)$. Supervised learning mencari fungsi hipotesis *h* yang mengaproksimasi *f*.

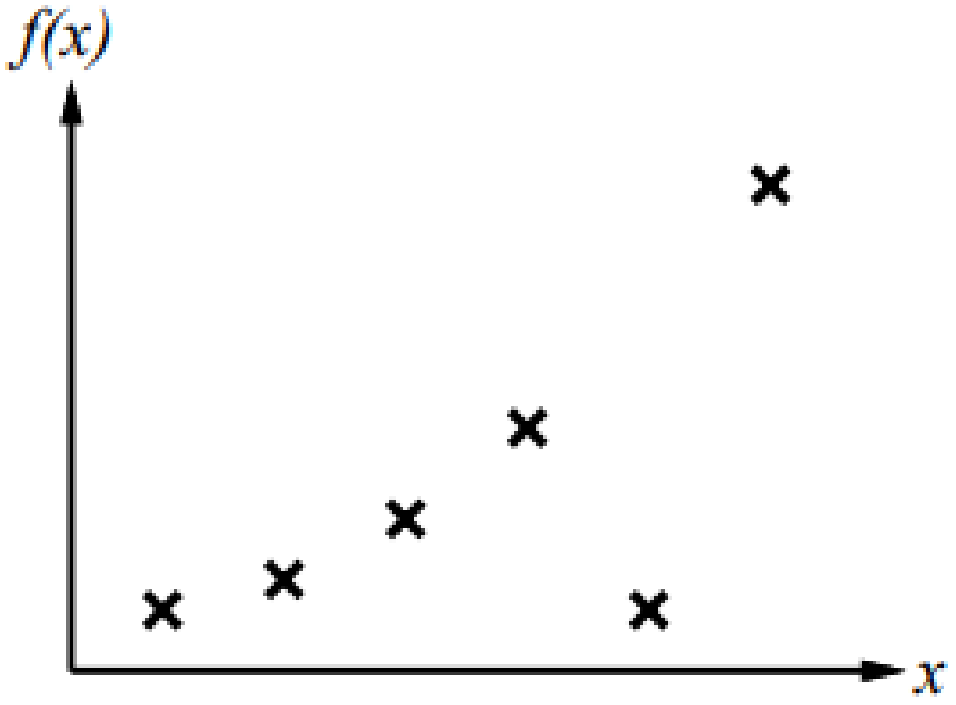
- Jika supervised learning adalah sebuah proses induktif, maka akan menghasilkan sebuah hypotheses space.
- Mis. untuk *curve-fitting*, hypotheses space = fungsi polinomial berderajat n, i.e.

$$f(x) = k_0 + k_1 x + k_2 x^2 + \cdots + k_n x^{n-1} + k_n x^n$$

- Search space terlalu kecil: $f(x)$ yang dicari tidak ada (unrealisable)
- Search space terlalu besar: semakin sulit ditelusuri, semakin banyak hipotesis yang konsisten dengan training example

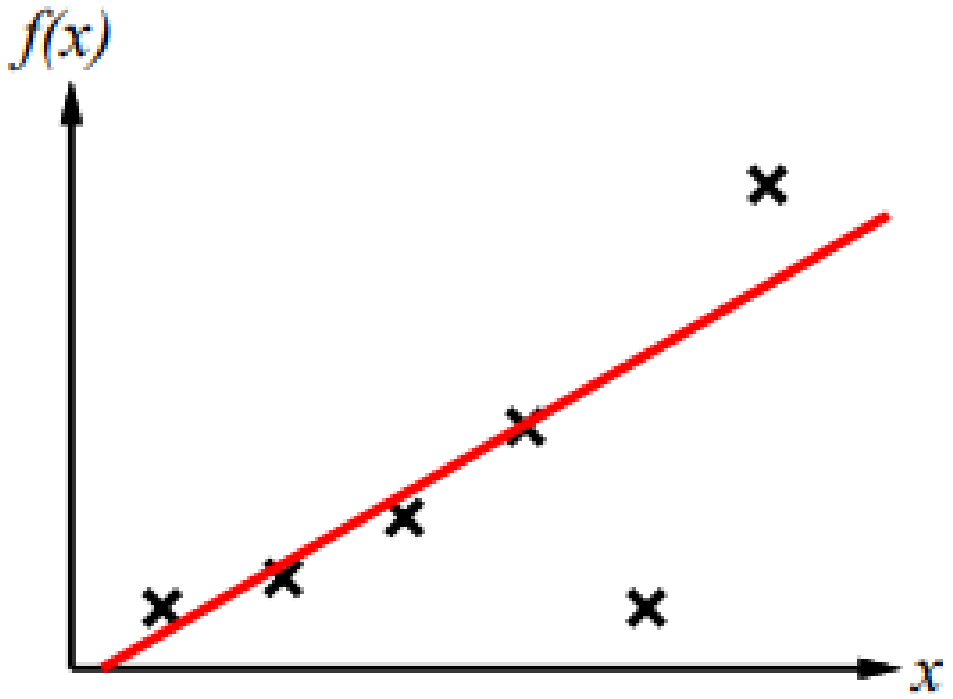
Contoh Curve-Fitting

- Untuk training set yang diberikan di bawah ini, fungsi manakah yang paling “fit” dengan training set dan juga titik-titik lainnya yang tidak diketahui?



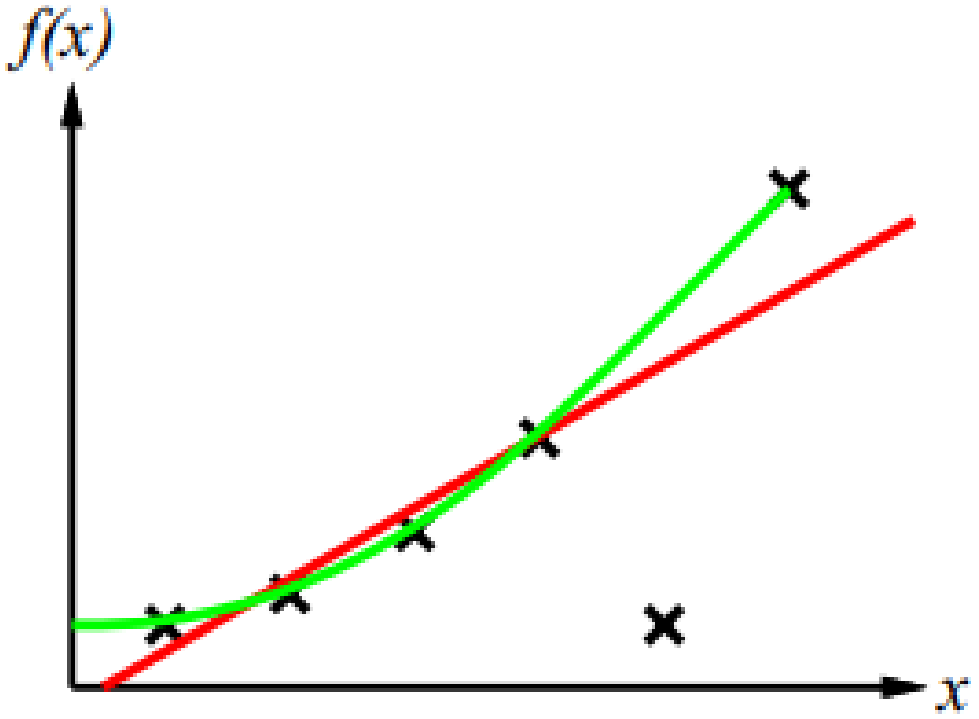
Contoh Curve-Fitting

- Untuk training set yang diberikan di bawah ini, fungsi manakah yang paling “fit” dengan training set dan juga titik-titik lainnya yang tidak diketahui?



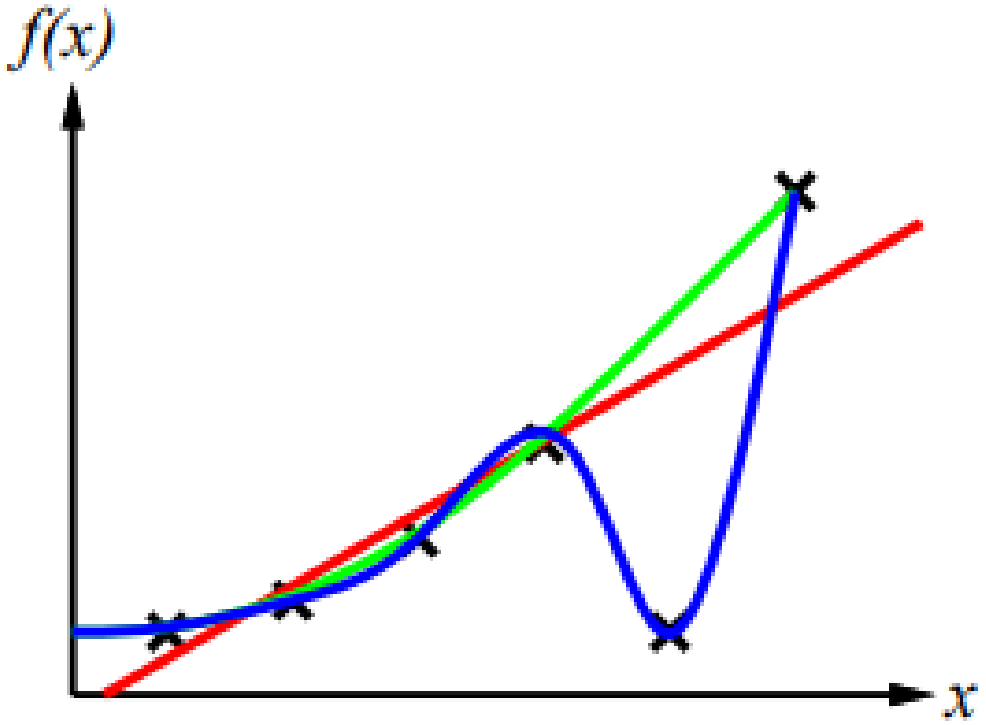
Contoh Curve-Fitting

- Untuk training set yang diberikan di bawah ini, fungsi manakah yang paling “fit” dengan training set dan juga titik-titik lainnya yang tidak diketahui?



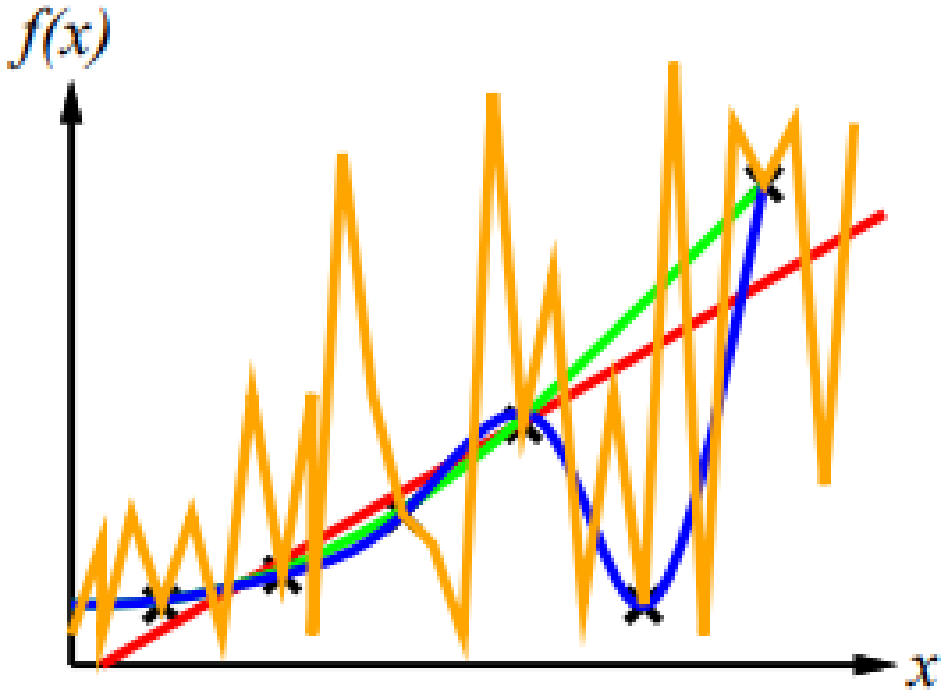
Contoh Curve-Fitting

- Untuk training set yang diberikan di bawah ini, fungsi manakah yang paling “fit” dengan training set dan juga titik-titik lainnya yang tidak diketahui?



Contoh Curve-Fitting

- Untuk training set yang diberikan di bawah ini, fungsi manakah yang paling “fit” dengan training set dan juga titik-titik lainnya yang tidak diketahui?



Consistency vs Simplicity

- Sebuah hipotesis yang konsisten dapat menjelaskan semua training example, i.e. memetakan input ke output yang tepat.
- Ada banyak hipotesis yang konsisten untuk sebuah training set, tergantung dari hypotheses space
- Occam's Razor: pilih yang paling simple!
  - Or more likely... hindari membuat fungsi hipotesis yang terlalu kompleks/konsisten dengan training set
  - Secara intuitif: generalisasi terhadap example baru.
- Biasanya ada trade-off antara consistency dan simplicity.

...But, how do we choose/construct the model?

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

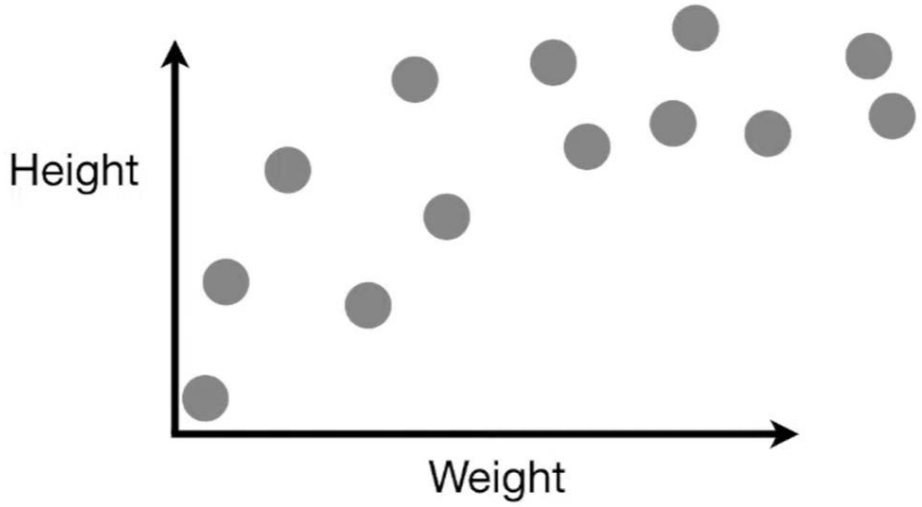
Bias-Variance Tradeoff

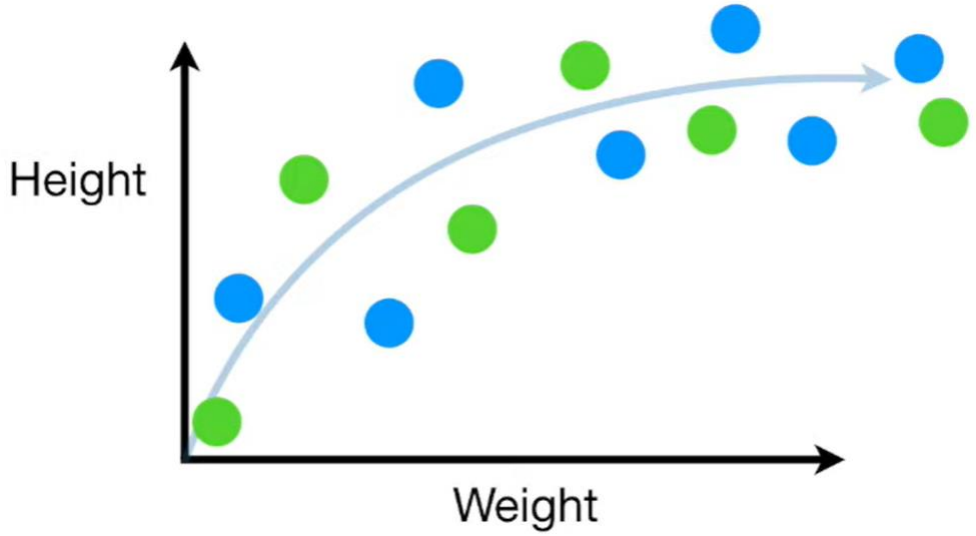
Sumber:

- Bias and Variance Video by StatQuest with Josh Starmer [StatQuest, 2022]
- Adila A. Krisnadhi, Slides Materi Machine Learning “Supervised Learning”, Semester Genap 2020/2021
- Charu C. Aggarwal, “Data Mining: The Textbook”, Springer, 2015
- Kevin P. Murphy, “Probabilistic Machine Learning: An Introduction”, MIT Press, 2021.
- Slides Materi Sistem Cerdas, “Machine Learning: Decision Trees”, Semester Genap 2020/2021
- Stuart Russel & Peter Norvig, “Artificial Intelligence: A Modern Approach”, 4th edition, Pearson, 2020
- https://towardsdatascience.com/ensemble-learning-bagging-boosting-3098079e5422

Case Example

- Latih model untuk prediksi dataset yang diperoleh tentang hubungan antara tinggi (panjang) dan berat tikus
- Dataset dibagi menjadi training data dan testing data

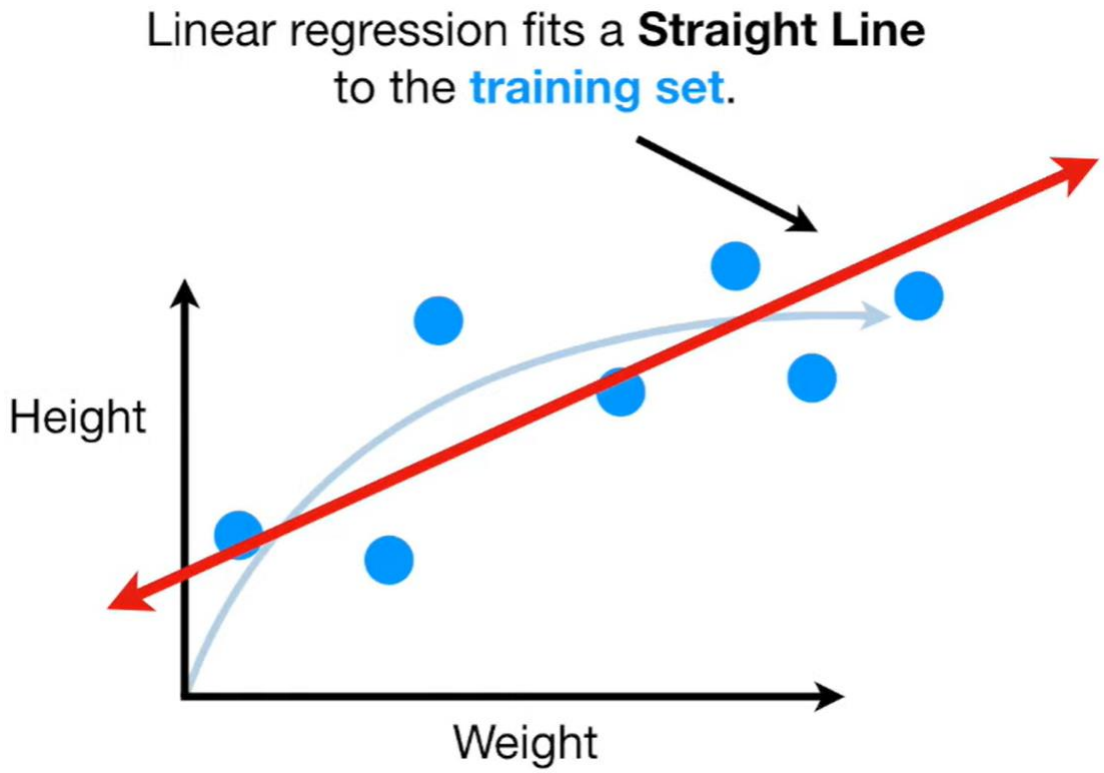




Training Testing

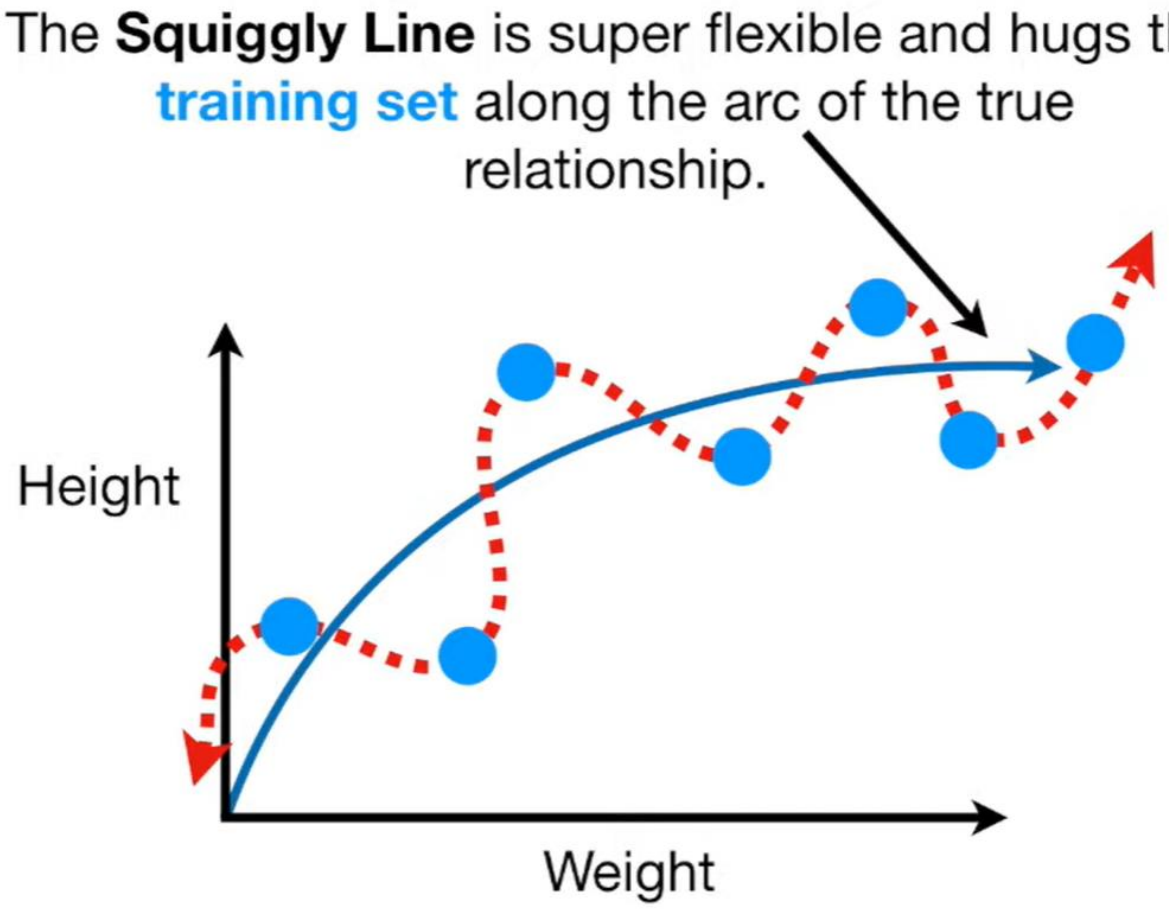
Case Example

- Misalkan kita mencoba Linear Regression (LR) berikut
- Model tidak bisa fit dengan data, tidak bisa menangkap hubungan antara nilai-nilai di dalam dataset, yaitu hubungan antara height dan weight tikus



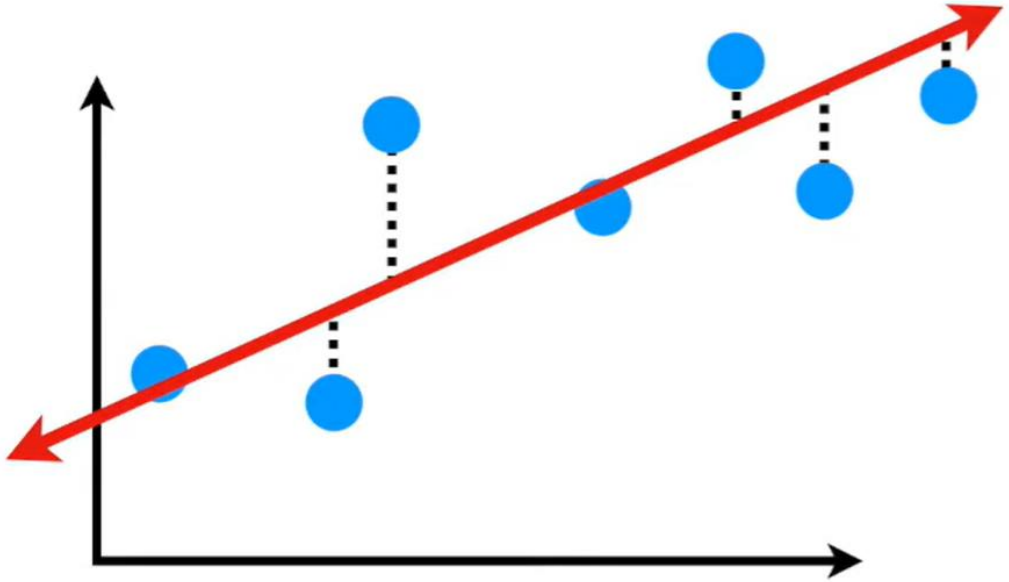
Case Example

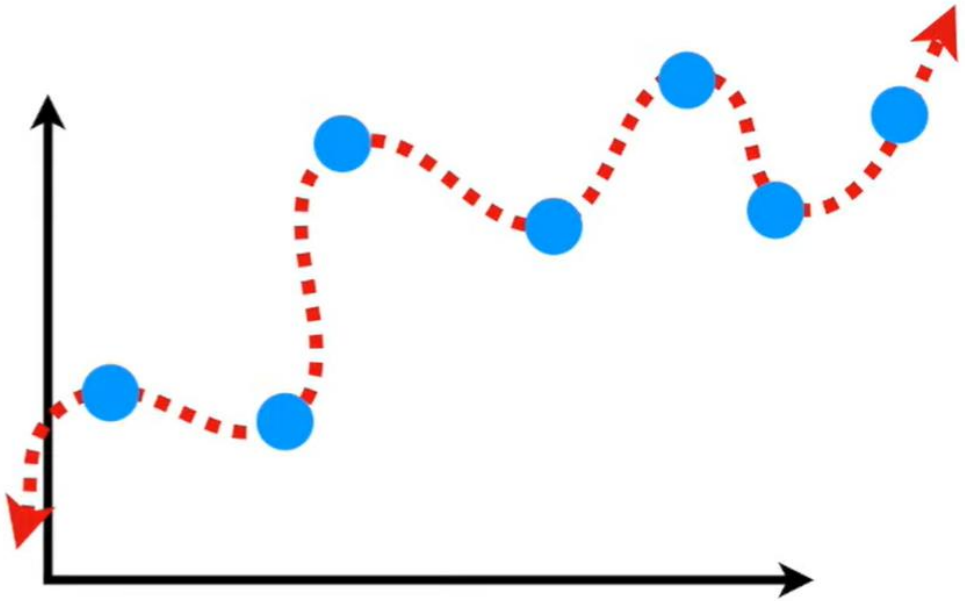
- Coba model yang lebih fleksibel “Squiggly Line”, sangat fit pada training data
- Bandingkan dengan kemampuan Linear Regression sebelumnya



Case Example

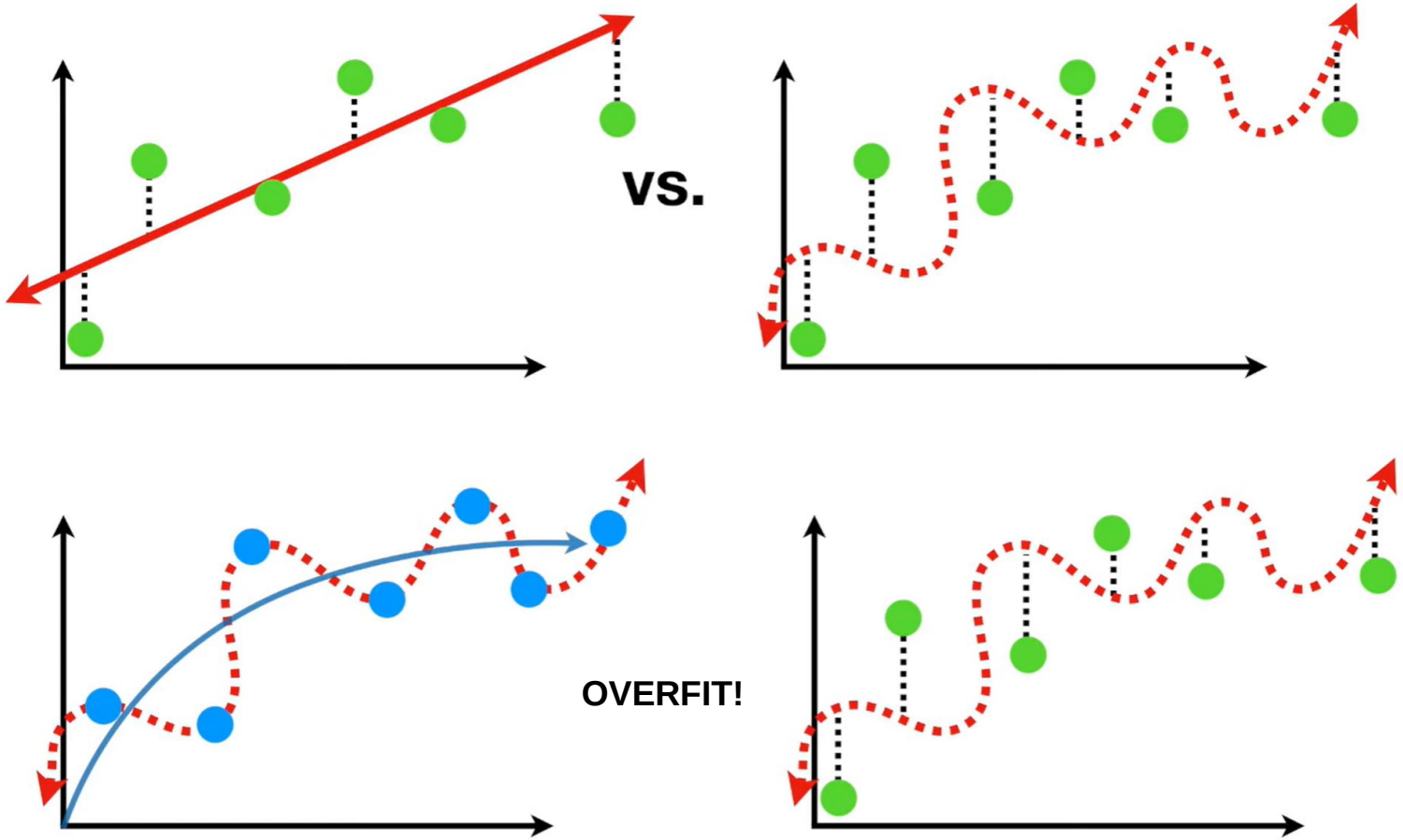
- Lakukan perhitungan error pada training dan bandingkan hasil antara LR dan Squiggly Line, misalkan dengan MSE
- MSE Squiggly Line = 0, MSE LR > 0





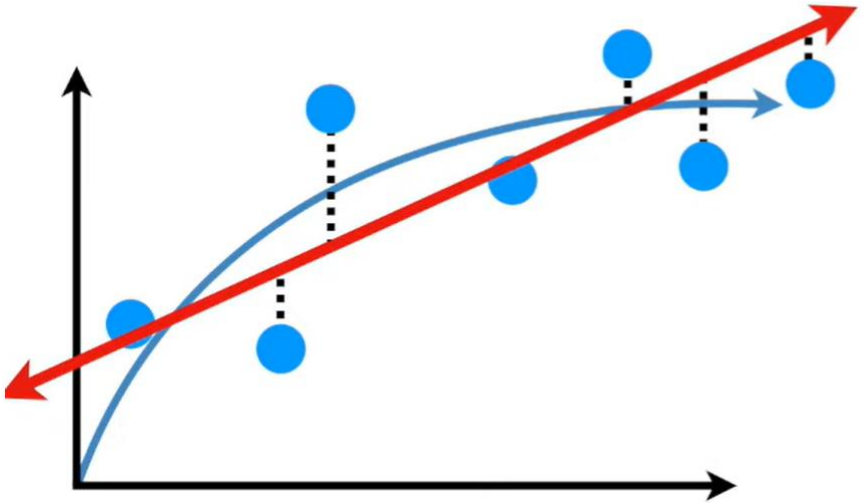
Case Example

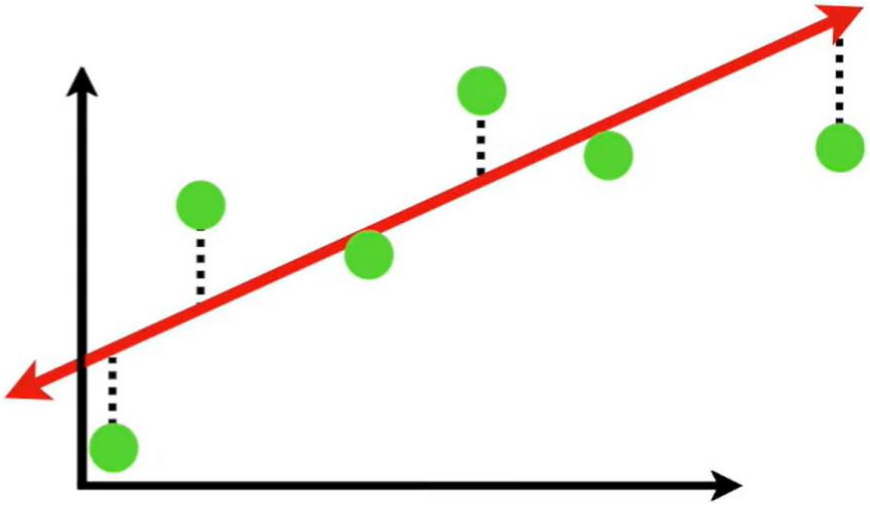
- Kinerja Squiggly Line pada data testing vs Kinerja LR pada data testing
- Pada testing, LR justru bekerja lebih baik!



Case Example

- Bandingkan dengan kinerja LR pada training dan testing

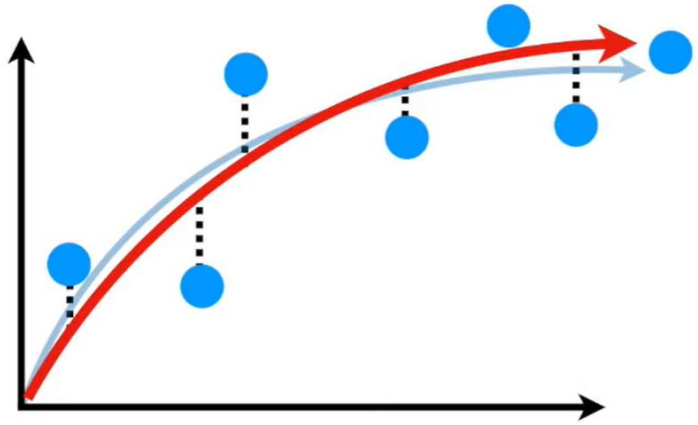


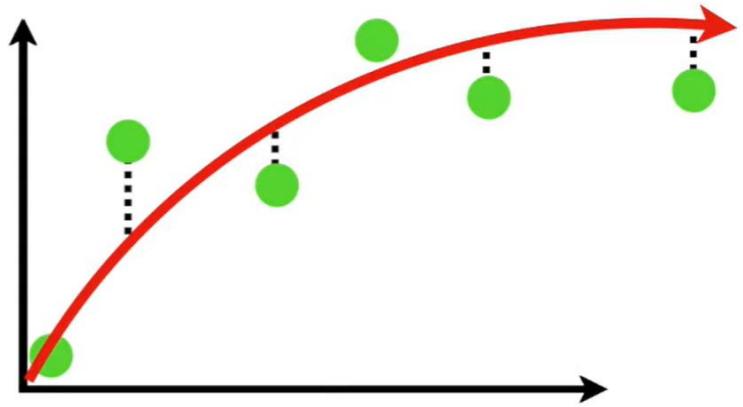


UNDERFIT!

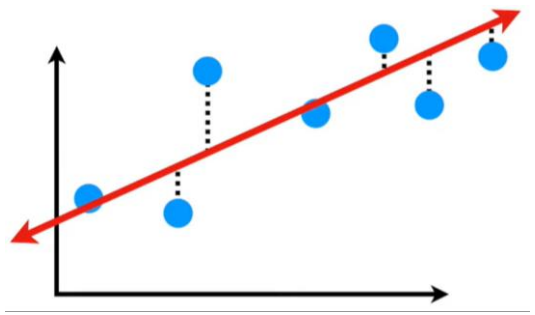
Case Example

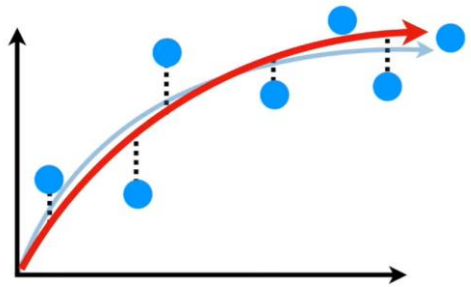
- The ideal model: low bias, low variance

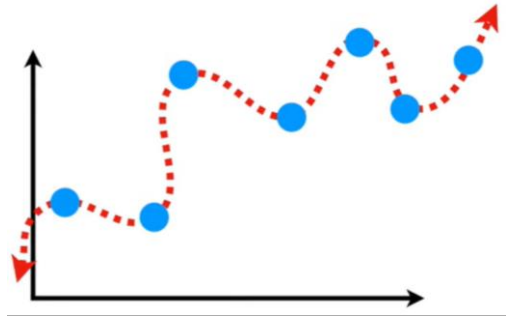




- Find a sweet spot between a simple model and a complex model







Bias & Variance

Berdasarkan contoh, maka:

- Bias adalah error yang merupakan kuadrat dari selisih antara rata-rata prediksi model dengan nilai output sebenarnya pada dataset.
  - ketidakmampuan model untuk mengaproksimasi dengan akurat hubungan pada data
  - Model dengan bias tinggi tidak mempelajari training set dengan baik (*pay little attention*) dan menyederhanakan model, i.e. mengambil asumsi pada sifat model (*simplifying assumptions*)
  - Berimplikasi pada error training dan testing yang tinggi

- Bias rendah biasanya ditemukan pada model yang kompleks

Bias & Variance

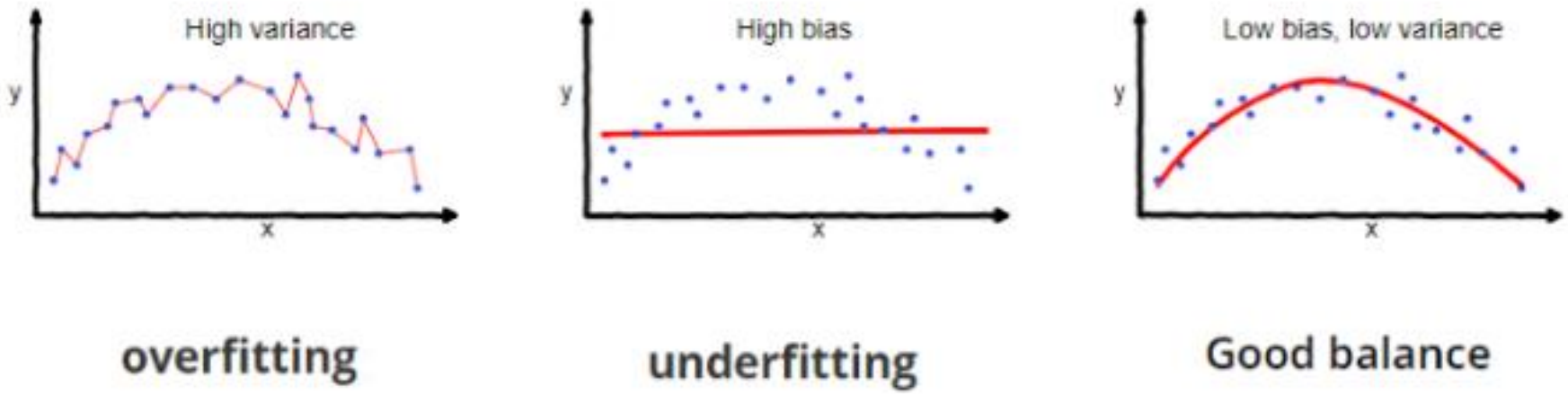
- Variance: *variability of model prediction for a given data point or a value which tells us spread of our data*
  - besarnya perubahan yang terjadi pada model bila training set yang digunakan berubah-ubah*
  - Model dengan variance tinggi akan cenderung lebih konsisten pada training set dan kurang bisa melakukan generalisasi pada data yang belum pernah dilihat/diprediksi.
  - Error rendah pada training, tapi error tinggi saat testing

- Model yang kompleks, terutama bila trainingnya tidak dibatasi, cenderung akan mempunyai variance yang tinggi

- Model yang baik adalah model yang low bias dan low variance:
  - low bias berarti bisa menangkap hubungan data dengan cukup akurat,
  - low variability berarti prediksi cukup konsisten untuk dataset yang berbeda-beda.

Overfitting vs Underfitting

- Underfitting terjadi ketika model kurang dapat menangkap hubungan antar data (*underlying pattern*).
- Overfitting terjadi ketika model mempelajari hubungan antar variable pada data **dan** juga noise pada data.



https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Bias & Variance: Error Modeling

- Bila **X** = (x_1, x_2, x_3, …, x_d) adalah data input, kita dapat mendefinisikan nilai output y sebagai berikut

$$y = f(\mathbf{X}) + \epsilon$$

  di mana f adalah sebuah model dan epsilon adalah error, noise, dsb
- Model f yang baik adalah model yang dapat memprediksi y secara akurat untuk data baru $\hat{\mathbf{x}}$
- Pada model yang ideal, epsilon independen terhadap **X** dan mempunyai mean 0.
- Setiap training example $(\mathbf{x}^{(i)}, y^{(i)})$ memenuhi $y^{(i)} = f(\mathbf{x}^{(i)}) + \epsilon^{(i)}$ di mana f adalah model ideal dan epsilon (i) adalah sebuah variabel acak (RV) dengan mean 0 dan variance σ^2.

Bias & Variance: Error Modeling

- Misalkan g(**X**) adalah aproksimasi dari f (setelah g melalui training). Bila g diberikan data testing $\hat{\mathbf{x}}$, maka prediksinya didefinisikan sebagai $\hat{y} = g(\hat{\mathbf{x}})$
- Asumsikan y adalah nilai output sebenarnya dari $\hat{\mathbf{x}}$. berdasarkan model ideal f, i.e.

$$y = f(\hat{\mathbf{x}}) + \epsilon$$

Maka… mean squared error (MSE) dari testing set adalah ekspektasi dari kuadrat error prediksi pada data testing, i.e.

$$MSE = \mathbb{E}[(y - g(\hat{\mathbf{x}}))^2]$$

Bias & Variance: Error Modeling

Gunakan sifat linearitas dari ekspektasi dan sifat independen antara epsilon, g dan f, sehingga:

$$
\begin{aligned}
MSE &= \mathbb{E}[(y - g(\hat{\mathbf{x}}))^2] = \mathbb{E}[(f(\hat{\mathbf{x}}) + \epsilon - g(\hat{\mathbf{x}}))^2] = \mathbb{E}[(f(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))^2] + \mathbb{E}[\epsilon^2] + \mathbb{E}[2(f(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))\epsilon] \\
&= \mathbb{E}[(f(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))^2] + \mathbb{E}[\epsilon^2] + 2\mathbb{E}[(f(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))\epsilon] \\
&= \mathbb{E}[(f(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))^2] + \underbrace{\mathbb{E}[\epsilon^2]}_{=\sigma^2} + 2\mathbb{E}[(f(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))] \underbrace{\mathbb{E}[\epsilon]}_{=0} = \mathbb{E}[(f(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))^2] + \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[(f(\hat{\mathbf{x}}) - g(\hat{\mathbf{x}}))^2] = \mathbb{E}\big[\big((f(\hat{\mathbf{x}}) - \mathbb{E}[g(\hat{\mathbf{x}})]) - (g(\hat{\mathbf{x}}) - \mathbb{E}[g(\hat{\mathbf{x}})])\big)^2\big] \\
&= \mathbb{E}\big[\underbrace{(\mathbb{E}[g(\hat{\mathbf{x}})] - f(\hat{\mathbf{x}}))^2}_{\text{fixed}}\big] + \mathbb{E}\big[(g(\hat{\mathbf{x}}) - \mathbb{E}[g(\hat{\mathbf{x}})])^2\big] - \mathbb{E}\big[2\underbrace{\big(f(\hat{\mathbf{x}}) - \mathbb{E}[g(\hat{\mathbf{x}})]\big)}_{\text{fixed}}\big(g(\hat{\mathbf{x}}) - \mathbb{E}[g(\hat{\mathbf{x}})]\big)\big] \\
&= \underbrace{(\mathbb{E}[g(\hat{\mathbf{x}})] - f(\hat{\mathbf{x}}))^2}_{\text{Bias}(g(\hat{\mathbf{x}}))} + \underbrace{\mathbb{E}\big[(g(\hat{\mathbf{x}}) - \mathbb{E}[g(\hat{\mathbf{x}})])^2\big]}_{\text{Var}(g(\hat{\mathbf{x}}))} - 2\big(f(\hat{\mathbf{x}}) - \mathbb{E}[g(\hat{\mathbf{x}})]\big)\underbrace{\big(\mathbb{E}[g(\hat{\mathbf{x}})] - \mathbb{E}[g(\hat{\mathbf{x}})]\big)}_{=0}
\end{aligned}
$$

Bias & Variance: Error Modeling

Sehingga, MSE dari prediksi sebuah test data $\hat{\mathbf{x}}$:

$$MSE = \mathbb{E}[(y - g(\hat{\mathbf{x}}))^2] = \big(\text{Bias}(g(\hat{\mathbf{x}}))\big)^2 + \text{Var}(g(\hat{\mathbf{x}})) + \sigma^2$$

- di mana komponen terakhir adalah noise dari distribusi data (*irreducible error*), bias dan variance adalah *reducible error*.

- Bila kita mempunyai testing set, maka error-nya menjadi:

$$\mathbb{E}[\mathbb{E}[(y - g(\hat{\mathbf{x}}))^2]] = \mathbb{E}[\big(\text{Bias}(g(\hat{\mathbf{x}}))\big)^2] + \mathbb{E}[\text{Var}(g(\hat{\mathbf{x}}))] + \sigma^2$$

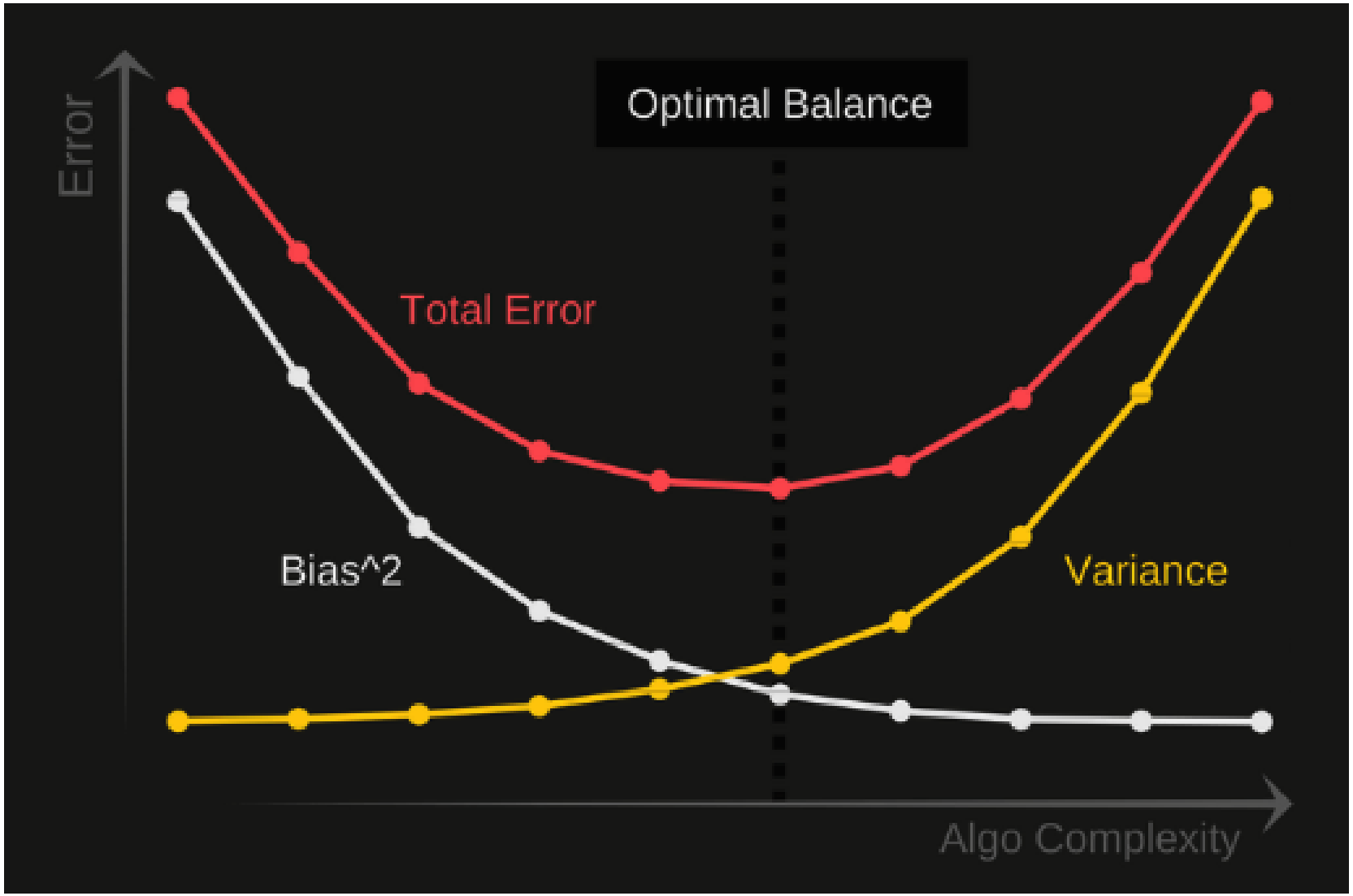
  - ❏ irreducible error tidak bisa diprediksi dan menjadi batas bawah nilai error
  - ❏ target ketika membangun model adalah menurunkan reducible error
  - ❏ Komponen reducible error merepresentasikan bias-variance tradeoff karena ketika bias tinggi, variance biasanya rendah, dan sebaliknya

Bias & Variance: Error Modeling

- Kita dapat membandingkan model-model yang berbeda berdasarkan bias dan variance mereka.
  - Contohnya, garis linear-regression mempunyai bias yang lebih tinggi and variance yang lebih rendah daripada cubic-regression curve.
- Umumnya, ketika menggunakan model yang lebih kompleks/flexible, variance cenderung meningkat dan bias menurun.
  - ketika kita meningkatkan fleksibilitas model, bias cenderung akan menurun lebih cepat daripada meningkatnya variance. Oleh karena itu, ekspektasi MSE testing akan menurun.
  - Namun suatu saat, meningkatnya fleksibilitas tidak berdampak pada menurunnya bias, tetapi justru menyebabkan variance meningkat secara signifikan. Oleh karena itu, ekspektasi MSE testing akan naik setelah titik tersebut.

- Bila model menghasilkan MSE kecil pada training, tetapi MSE besar saat testing, artinya "*the model overfits the training data*". Model yang terlalu fleksibel biasanya mudah overfit.
- Bila model menghasilkan MSE besar baik di training maupun testing, artinya "*the model underfits*".

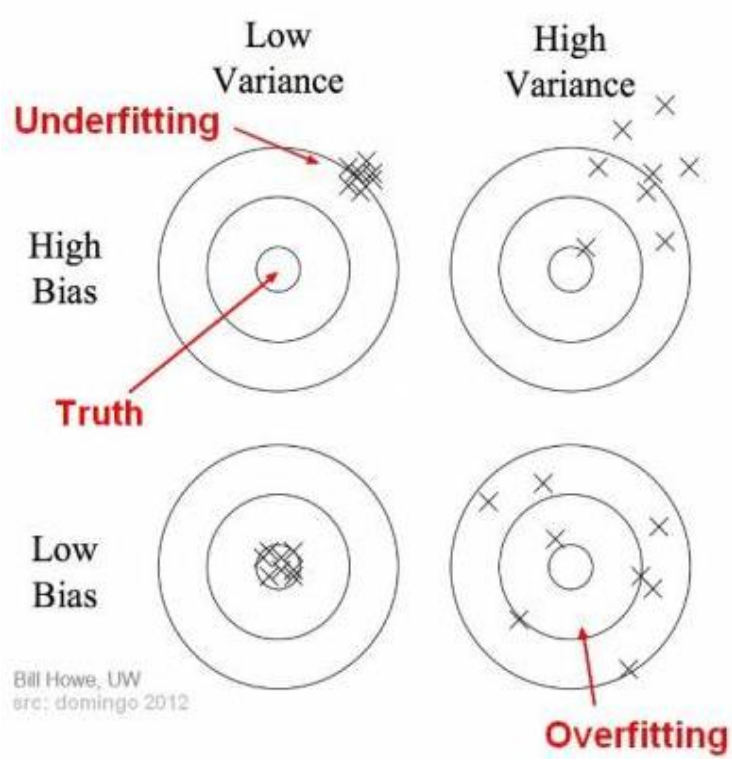
Bias & Variance: Error Modeling



https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Bias & Variance

Bila bias, variance, dan kinerja model dianalogikan sebagai “Game of Darts”.



Agar mendapatkan model dengan bias dan variance yang lebih rendah, kita dapat menerapkan teknik/metode-metode berikut...

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Regularization

- Teknik yang digunakan untuk menurunkan error fitting fungsi pada training set, sekaligus menghindari overfitting.
- Ada 2 strategi yang dapat dilakukan untuk menghindari overfitting pada decision tree:
  - Menghentikan decision tree learning berdasarkan suatu *heuristic-based criteria*.
  - Bangun decision tree seperti biasa sampai kedalaman maximum, lalu lakukan pruning dengan menyatukan (merge) subtree yang terpisah (karena splitting) dengan node parent-nya sampai mendapatkan leaf node yang memenuhi suatu heuristic-based criteria. (lebih lambat daripada strategi pertama)
- Heuristic-based criteria:
  - Jumlah minimum data pada node: setiap leaf node harus mempunyai jumlah examples minimum agar bisa di-split.
  - Jumlah maksimum leaf node: membatasi jumlah leaf node yang diperbolehkan pada tree.
  - Kedalaman maksimum: membatasi kedalaman tree.

Ensemble Learning

Recall

Membangun model prediksi yang terdiri atas M base model. Prediksi dari sebuah ensemble ditentukan dengan mengambil rata-rata dari beberapa model yang dibangun

$$f(y|\mathbf{x}) = \frac{1}{|M|}\sum_{m=1}^{M} f_m(y|\mathbf{x})$$

base model ke-m

- bias akan menyamai base model tetapi dengan variance lebih rendah
- Untuk classification, gunakan **committee method**, yaitu ambil vote mayoritas sebagai hasil prediksi
- … atau gunakan metode *stacking,* i.e. setiap base model diberikan bobot (weight)

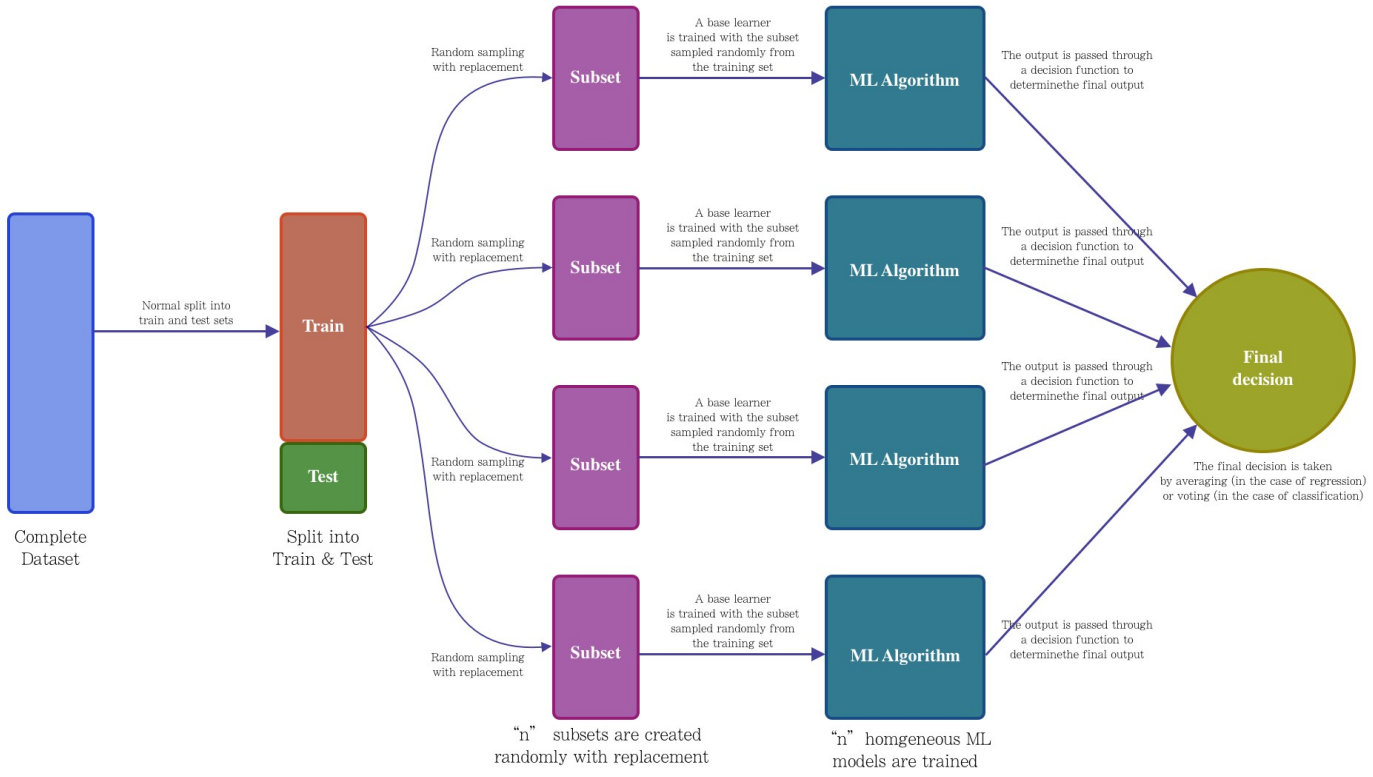
$$f(y|\mathbf{x}) = \sum_{m=1}^{M} w_m f_m(y|\mathbf{x})$$

- Berbagai jenis ensemble learning dirancang untuk meningkatkan kinerja model dengan mengurangi bias, variance atau keduanya.

Ensemble Learning: Bagging

(see slide “Random Forests” ☺)

- Bagging dan Random Forest mengurangi variance dengan cara training N model secara independen (paralel)
- Biasanya menggunakan jenis base model (*weak learners*) yang sama



It is based on the idea that if the variance of a prediction is σ^2, then the variance of the average of k independent and identically distributed (i.i.d.) predictions is reduced to $\frac{\sigma^2}{k}$

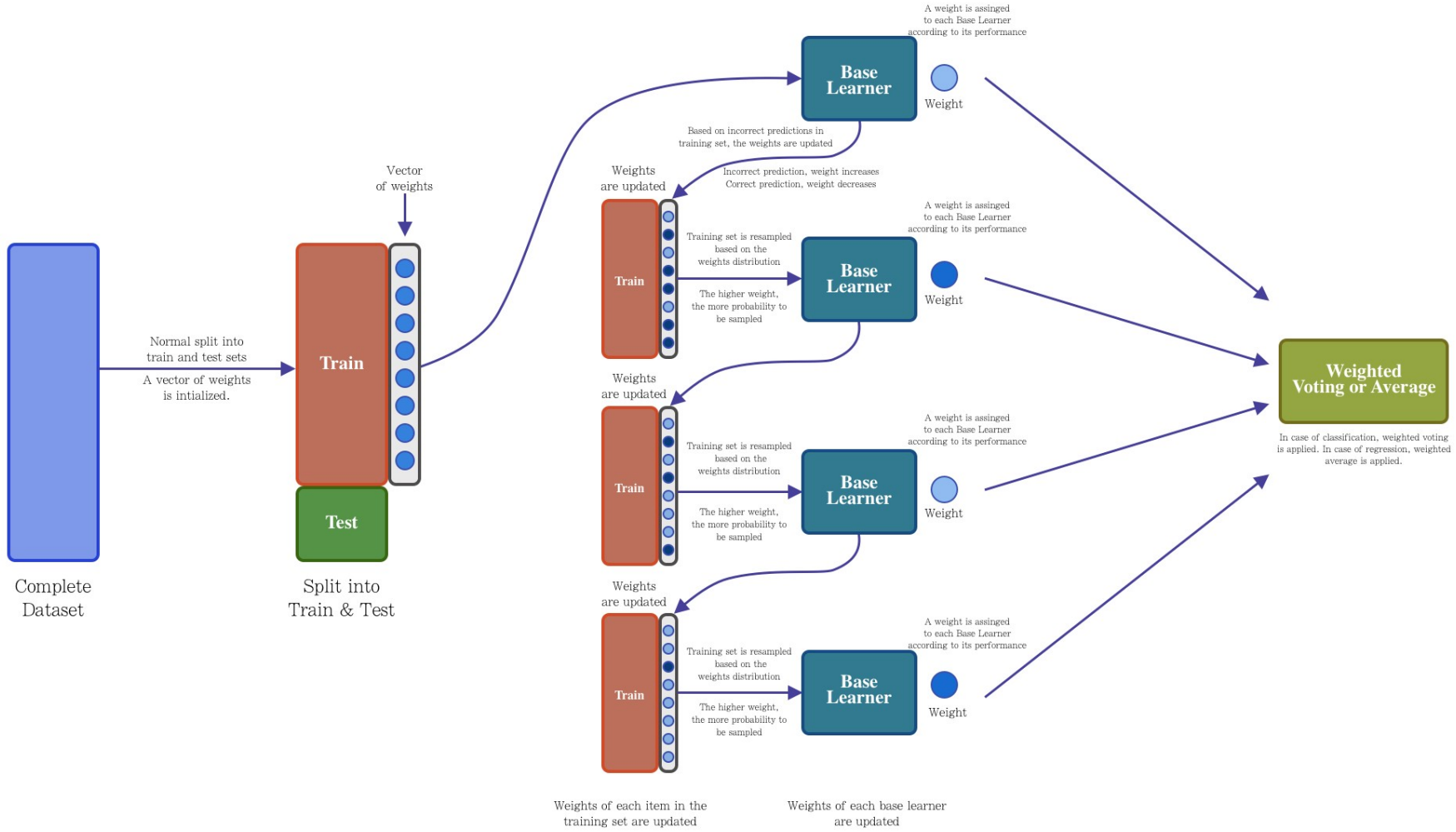
Ensemble Learning: Boosting

- Metode untuk mengurangi bias dengan training N model yang saling dependen, i.e. training dilakukan secara sekuensial. Setiap base model dilatih dengan cara mencoba untuk meningkatkan kinerja atau memperbaiki kesalahan yang dilakukan oleh model sebelumnya
- Perbedaan antara jenis boosting terletak pada bagaimana error setiap base model dipertimbangkan untuk diperbaiki pada base model berikutnya, i.e. bagaimana error diberikan penalti (e.g. memodifikasi bobot/weight atau meminimalkan loss function)
  - AdaBoost (Adaptive Boosting)
  - Gradient Boosting
  - XGBoost (Extreme Gradient Boosting
  - dll

- AdaBoost membuat sebuah vektor bobot, setiap nilai bobot untuk baris data tertentu. Bila ada data yang salah diprediksi, bobotnya ditambah sedangkan data yang diprediksi dengan benar bobotnya dikurangi. Bobot menunjukkan seberapa besar peluang data tersebut dipilih sebagai data training sebuah base model

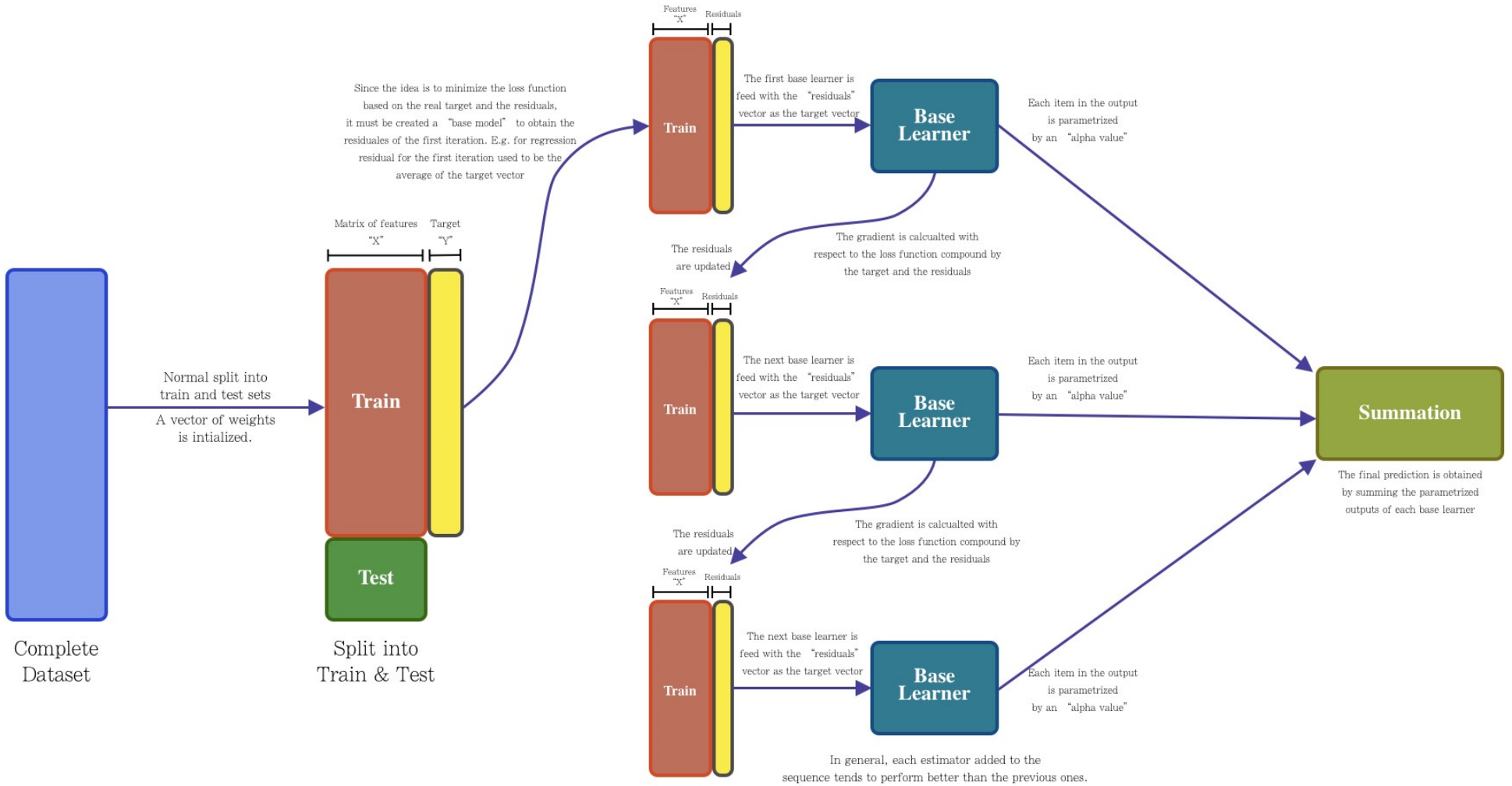
Ensemble Learning: Boosting

- AdaBoost juga memberikan bobot pada setiap base model. Bobot yang lebih besar diberikan pada base model dengan kinerja lebih tinggi



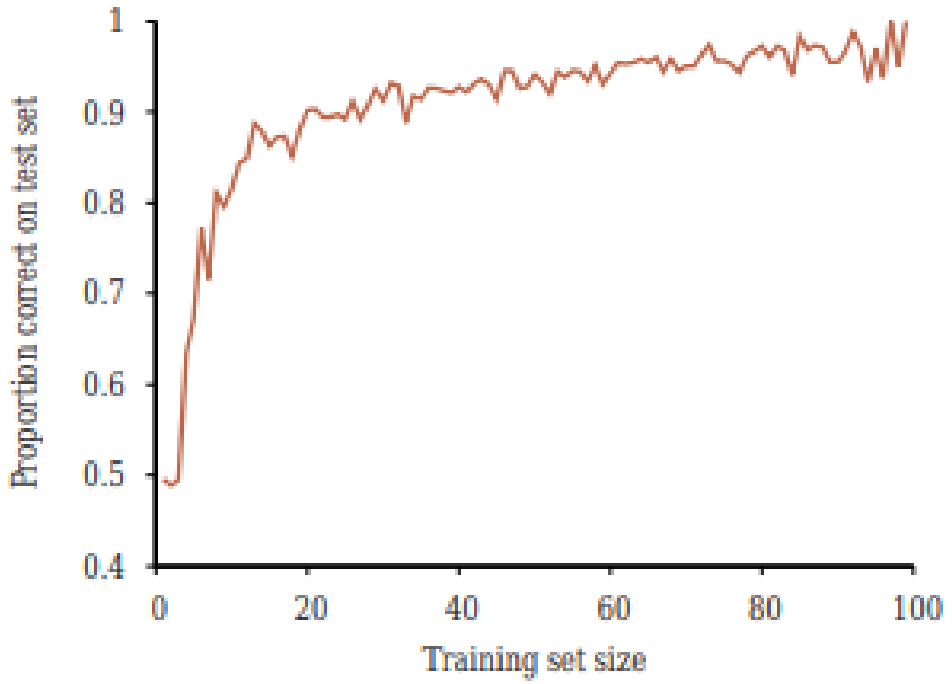
FAKULTAS ILMU KOMPUTER

Ensemble Learning: Boosting



Learning Curve

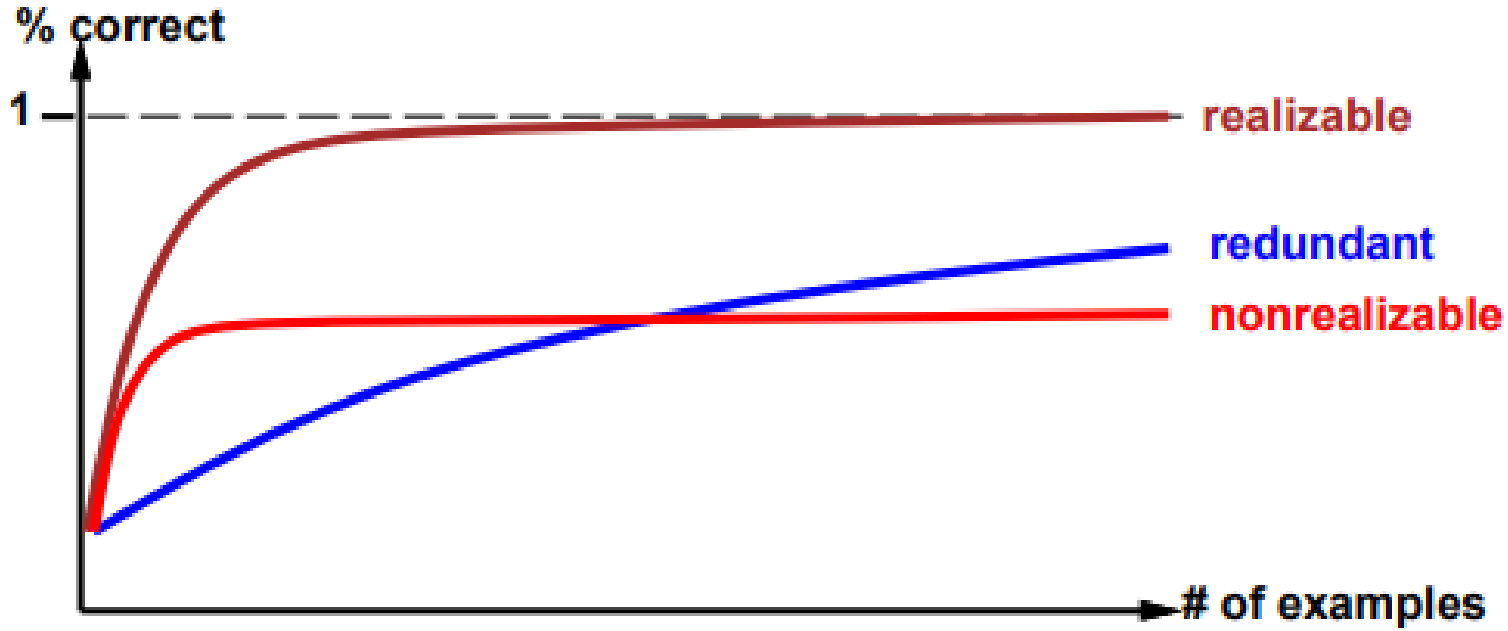
- Cara lain yang bisa dilakukan untuk menurunkan error (variance) adalah menambah data pada training set



Learning Curve

Bentuk learning curve juga tergantung masalah yang dipelajari:

- Realizable: Fungsi target *f*(*x*) bisa dinyatakan
- Non-realizable: Fungsi target *f*(*x*) tidak bisa dinyatakan (kurang atribut?)
- Redundant: Banyak atribut noise yang tidak berguna, menyesatkan (overfitting)!



Bias-Variance Tradeoff

Table 11.1: Impact of different techniques on bias-variance trade-off

Technique	Source/level of bias	Source/level of variance
Simple models	Oversimplification increases bias in decision boundary	Low variance. Simple models do not overfit
Complex models	Generally lower than simple models. Complex boundary can be modeled	High variance. Complex assumptions will be overly sensitive to data variation
Shallow decision trees	High bias. Shallow tree will ignore many relevant split predicates	Low variance. The top split levels do not depend on minor data variations
Deep decision trees	Lower bias than shallow decision tree. Deep levels model complex boundary	High variance because of overfitting at lower levels
Rules	Bias increases with fewer antecedents per rule	Variance increases with more antecedents per rule
Naive Bayes	High bias from simplified model (e.g., Bernoulli) and naive assumption	Variance in estimation of model parameters. More parameters increase variance
Linear models	High bias. Correct boundary may not be linear	Low variance. Linear separator can be modeled robustly
Kernel SVM	Bias lower than linear SVM. Choice of kernel function	Variance higher than linear SVM
k-NN model	Simplified distance function such as Euclidean causes bias. Increases with k	Complex distance function such as local discriminant causes variance. Decreases with k
Regularization	Increases bias	Reduces variance

UNIVERSITAS INDONESIA

FAKULTAS ILMU KOMPUTER

TERIMA KASIH

Disclaimer: Figures and content can be originated from other sources on the Web. The purpose of this slide set is educational only.