Aprendizado de Máquina e Reconhecimento de Padrões 2021.2

The Bias-Variance Trade-off

Based on videos from StatQuest, the course 'Machine Learning' from Andrew Ng, and the book 'Hands-on machine learning with Scikit-Learn, Keras and TensorFlow' from A. Géron

Prof. Dr. Samuel Martins (Samuka)

samuel.martins@ifsp.edu.br

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
São Paulo
Campus Campinas

Generalization Error

The **Generalization Error** for any machine learning algorithm can be broken down into three parts:

- Bias Error
- Variance Error
- Irreducible Error
    - This part is due to the noisiness of the data itself.
    - The only way to reduce this part of the error is to **clean up the data**
        - Fix the data sources (e.g., broken sensors), or
        - Detect and remove outliers.

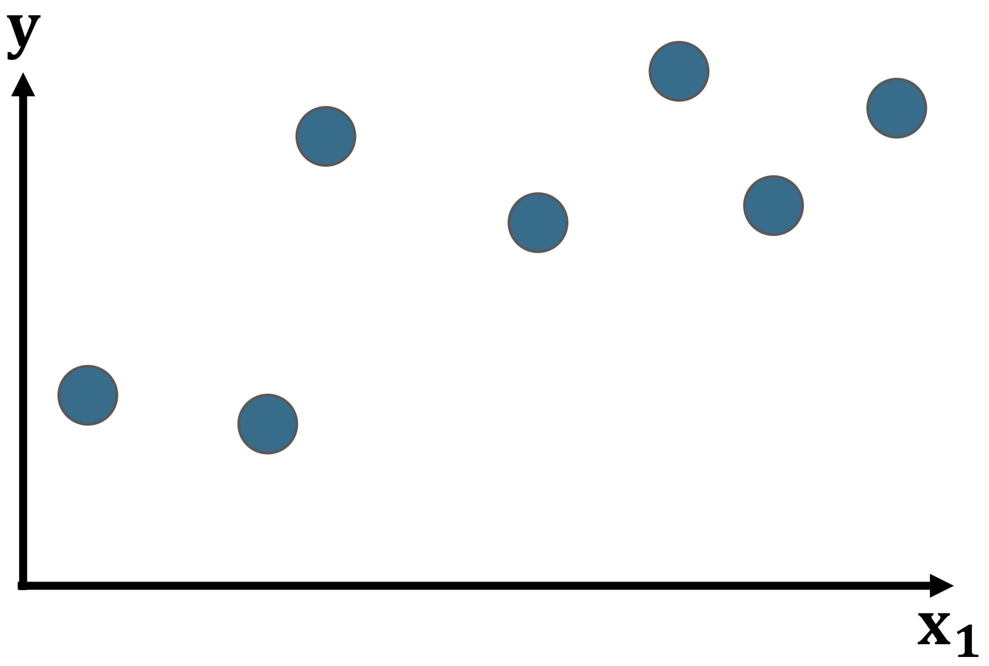
Generalization Error

The **Generalization Error** for any machine learning algorithm can be broken down into three parts:

- **Bias Error**
- **Variance Error**

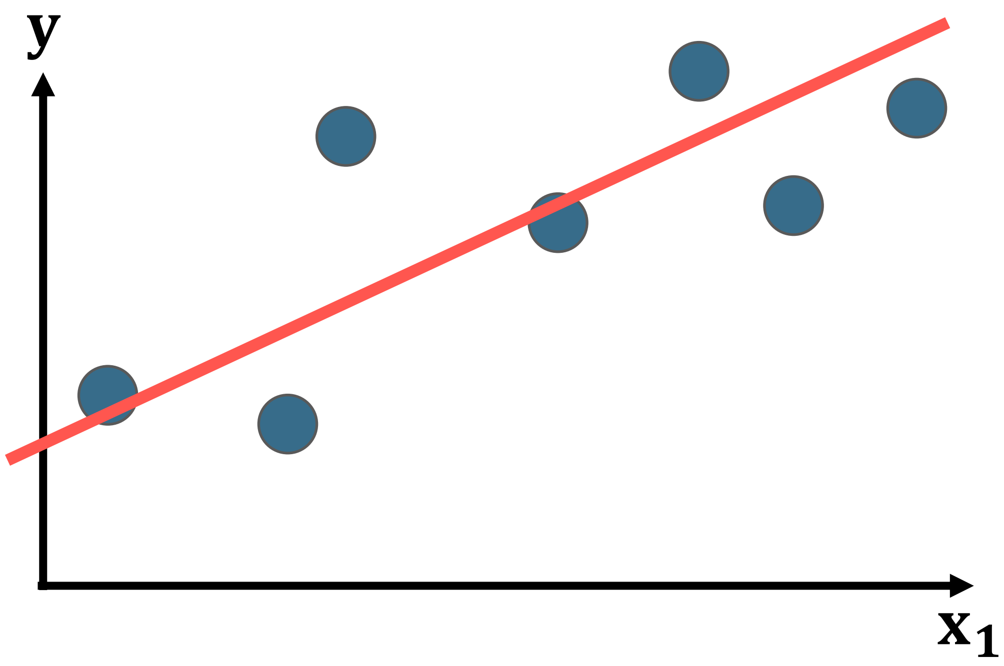
Let’s see these errors

- Irreducible Error
  - This part is due to the noisiness of the data itself.
  - The only way to reduce this part of the error is to **clean up the data**
    - Fix the data sources (e.g., broken sensors), or
    - Detect and remove outliers.


y
x1

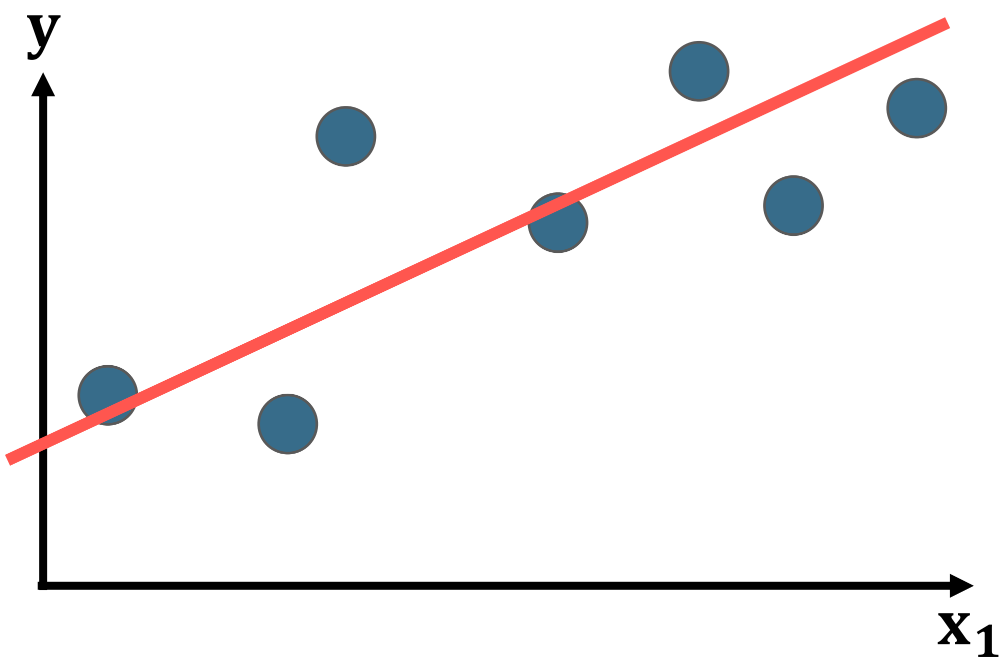
e.g., Linear Regression

Straight Line



e.g., Linear Regression

Straight Line

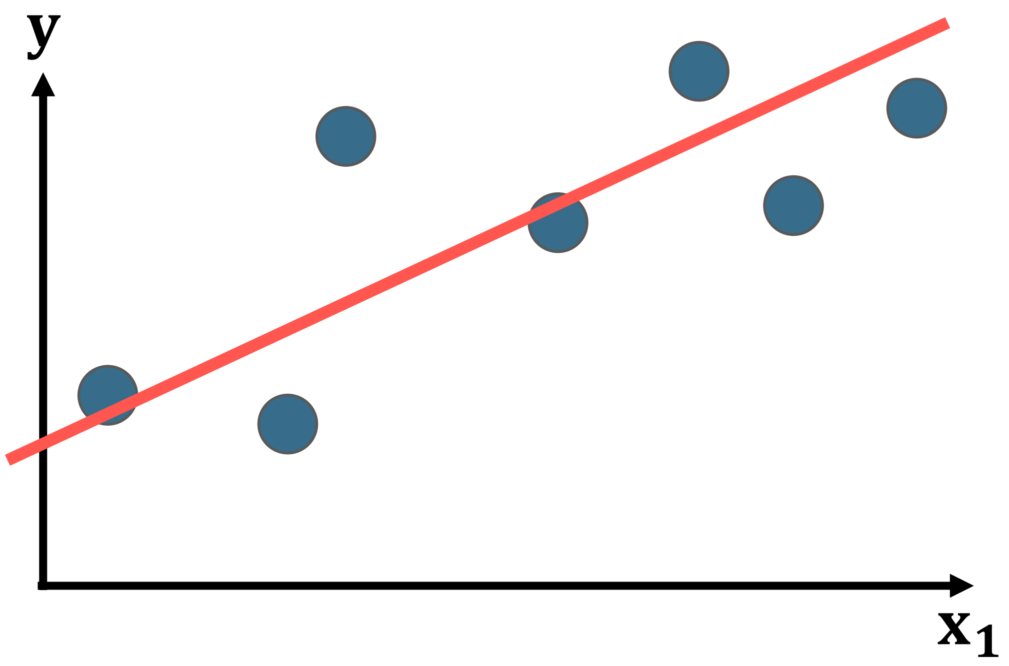


The **linear model** (straight line) **cannot** capture the **true relationship** between $\mathbf{x_1}$ and $\mathbf{y}$.

In ML, this inability is called **bias**.

e.g., Linear Regression

Straight Line



The **linear model** (straight line) **cannot** capture the **true relationship** between $\mathbf{x_1}$ and $\mathbf{y}$.

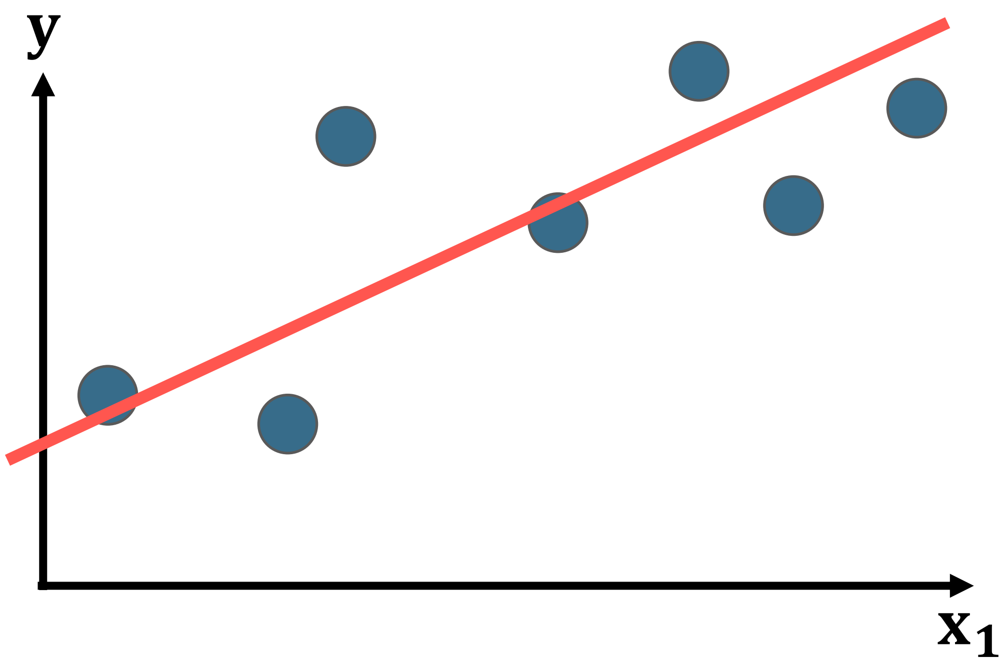
In ML, this inability is called **bias**.

Bias
Error associated to **wrong assumptions (simplifications)** made by a model (e.g., assuming that the data is linear when it is quadratic) to make it **easier to learn**.

Bias
'Average distance' between **predictions** and the **truth**.

e.g., Linear Regression

Straight Line



(relatively) **high** bias

⚠ The **linear model** (straight line) **cannot** capture the **true relationship** between $\mathbf{x_1}$ and $\mathbf{y}$.

In ML, this inability is called bias.

Bias

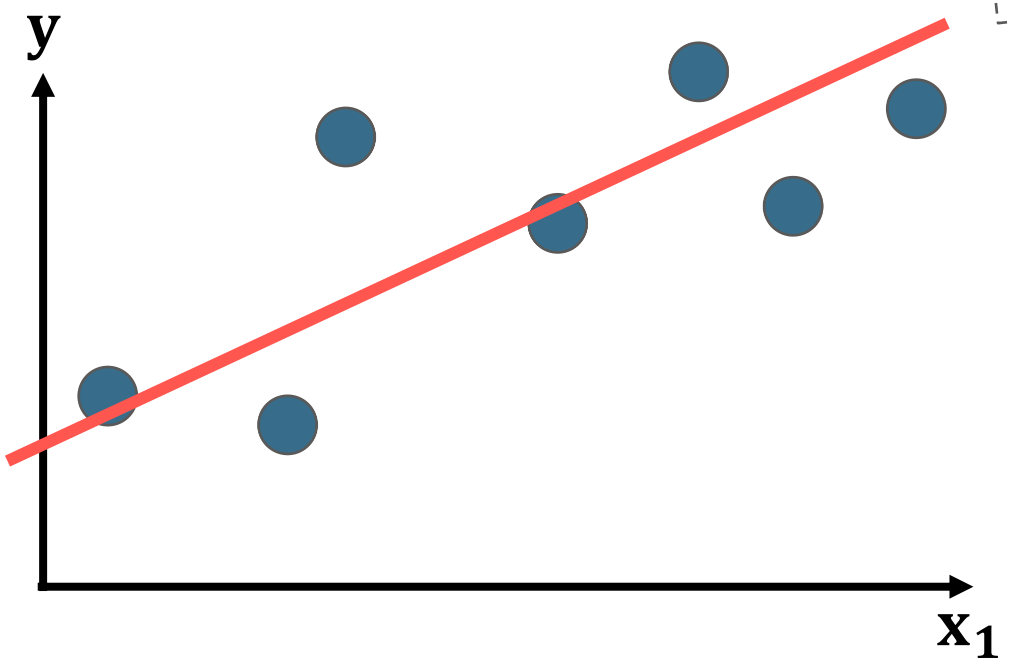
Error associated to **wrong assumptions (simplifications)** made by a model (e.g., assuming that the data is linear when it is quadratic) to make it **easier to learn**.

Bias

'Average distance' between predictions and the **truth**.

e.g., Linear Regression

Straight Line



A **high-bias model** is most likely to **underfit** the training data.

(relatively) **high bias**

The **linear model** (straight line) **cannot** capture the **true relationship** between $\mathbf{x_1}$ and $\mathbf{y}$.

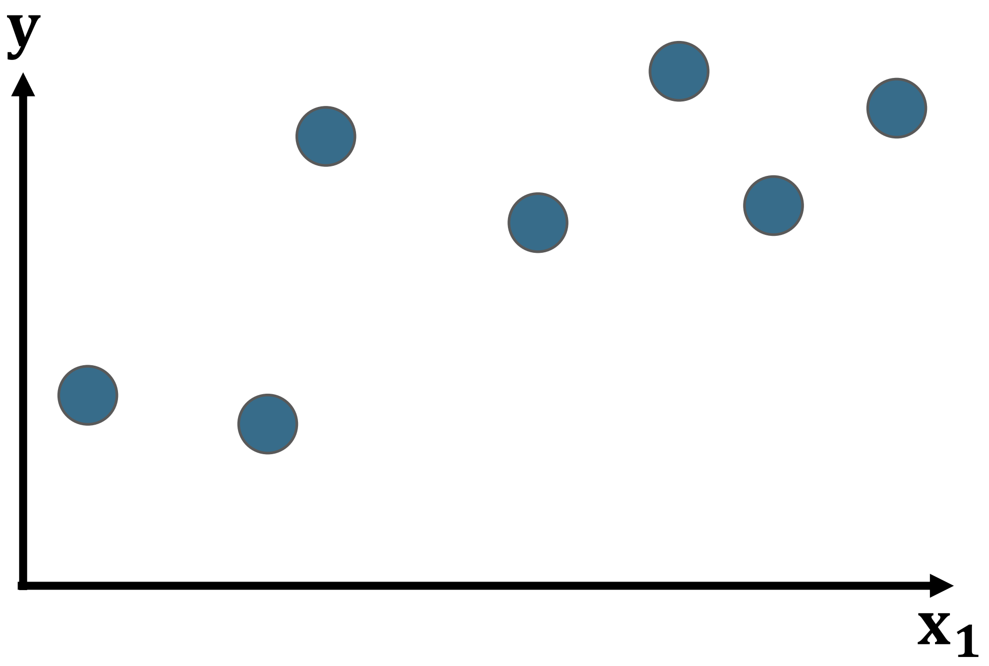
In ML, this inability is called **bias**.

Bias

Error associated to **wrong assumptions (simplifications)** made by a model (e.g., assuming that the data is linear when it is quadratic) to make it **easier to learn**.

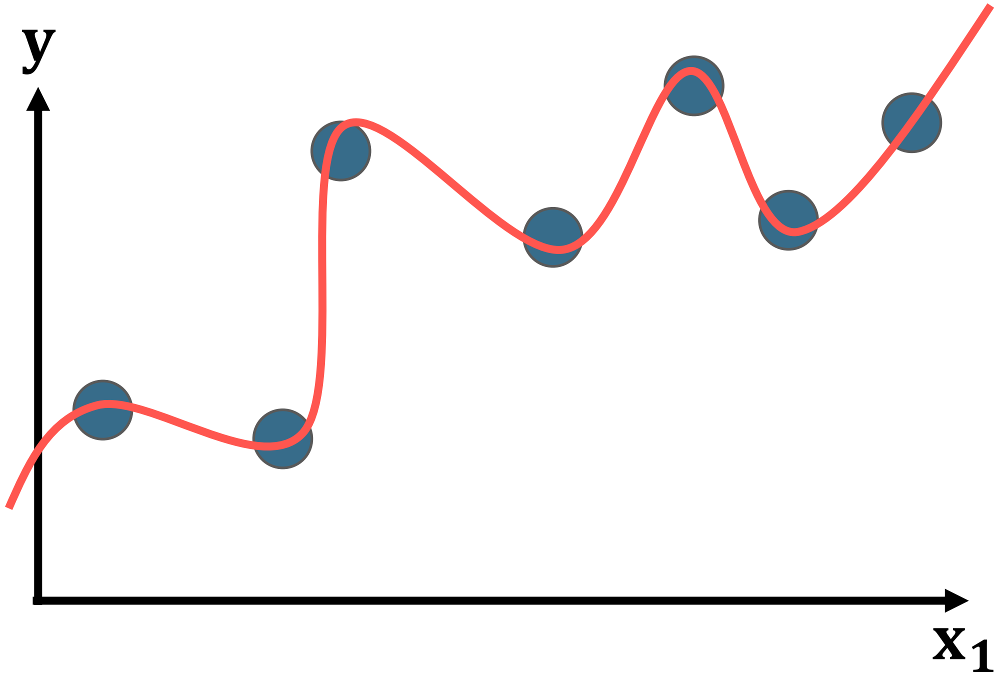
Bias

'Average distance' between **predictions** and the **truth**.


y
x1

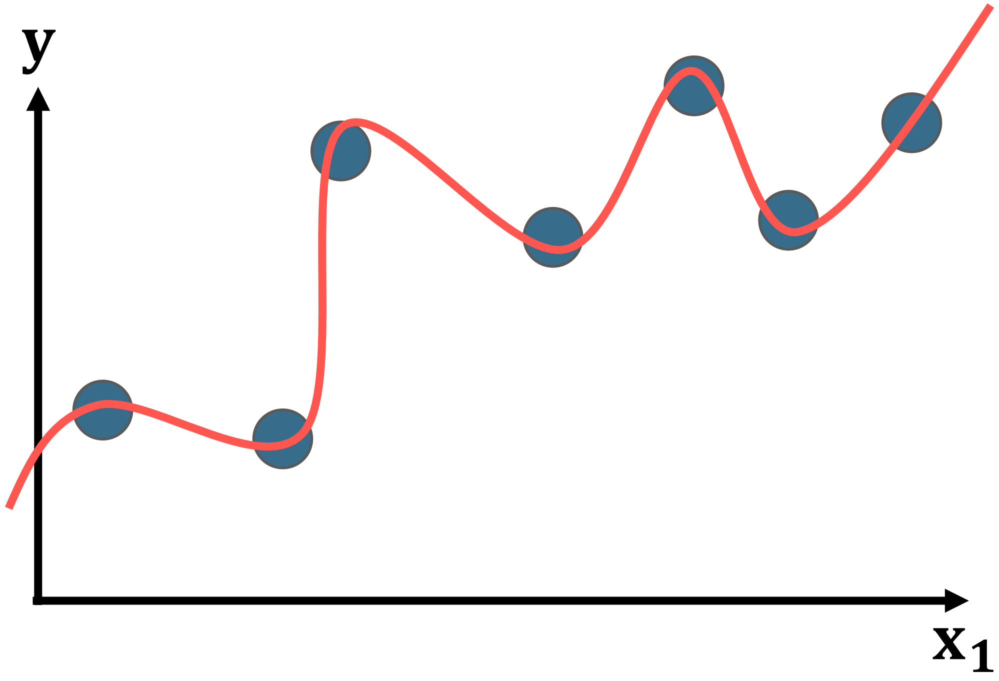
a more complex model ➔ *e.g.,* high-degree polynomial model

Squiggly Line



a more complex model ➔ *e.g.*, high-degree polynomial model

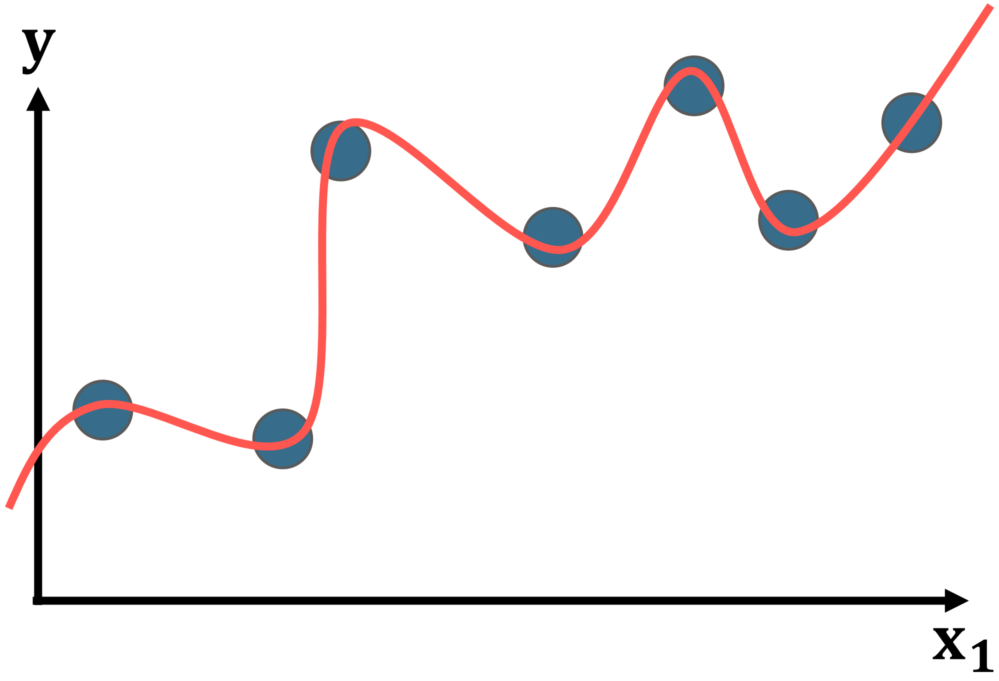
Squiggly Line



The **model** (squiggly line) **can** capture the **true relationship** between $\mathbf{x_1}$ and $\mathbf{y}$.

a more complex model ➔ *e.g.*, high-degree polynomial model

Squiggly Line

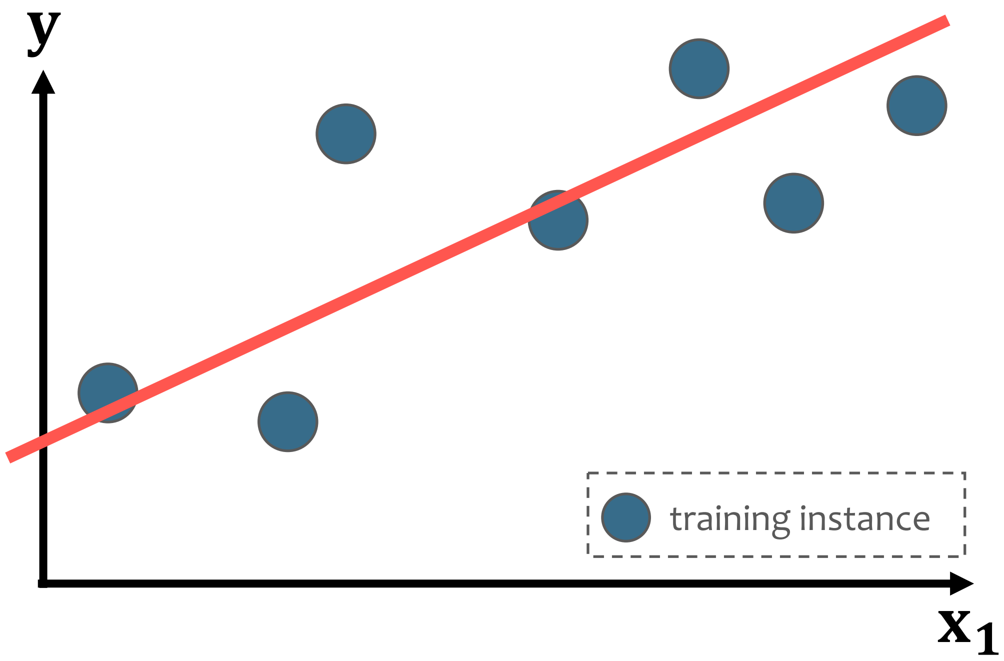


(very) **low bias**

'Average distance' between predictions and the **truth** is **close to zero**.

The **model** (squiggly line) **can** capture the **true relationship** between $\mathbf{x_1}$ and $\mathbf{y}$.

Straight Line



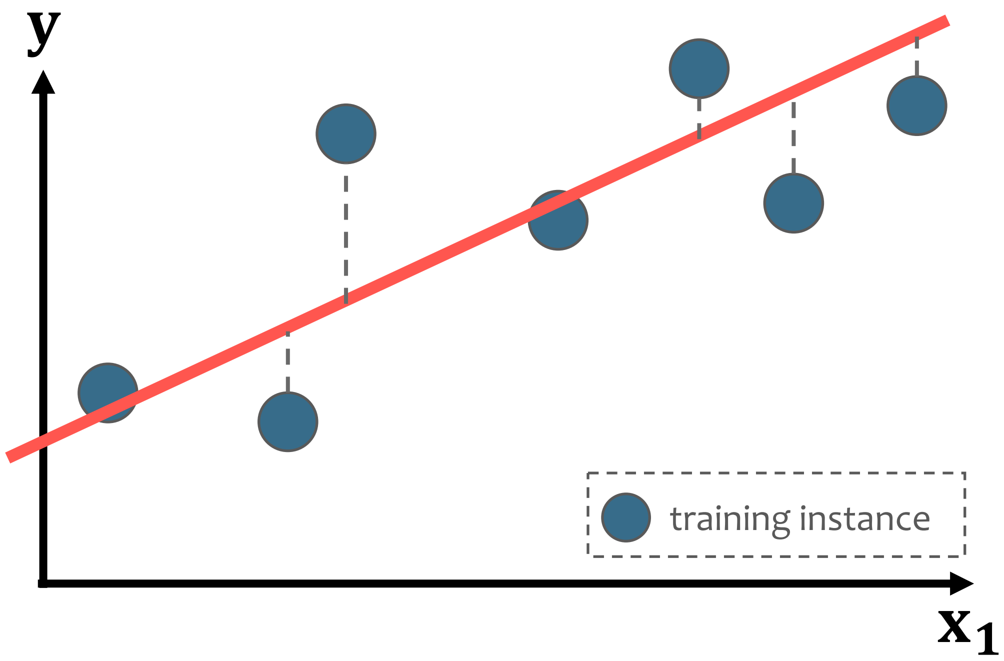
(relatively) **high bias**

Squiggly Line

(very) **low bias**

By considering *only* the **training set errors**, we would pick the **squiggly line** below.

Straight Line

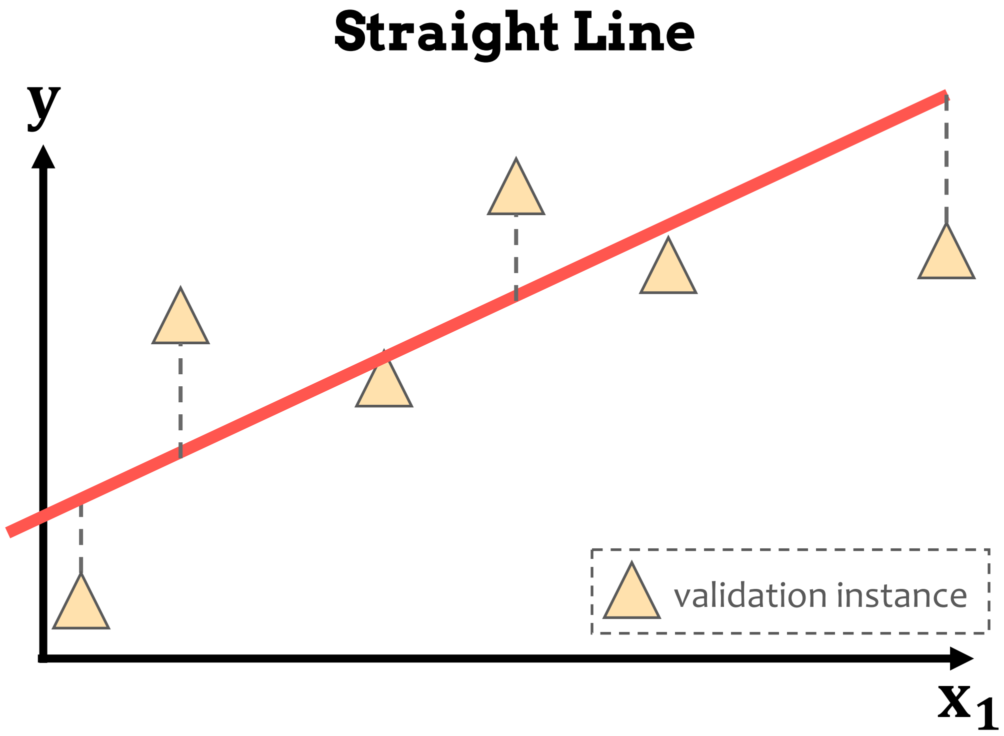


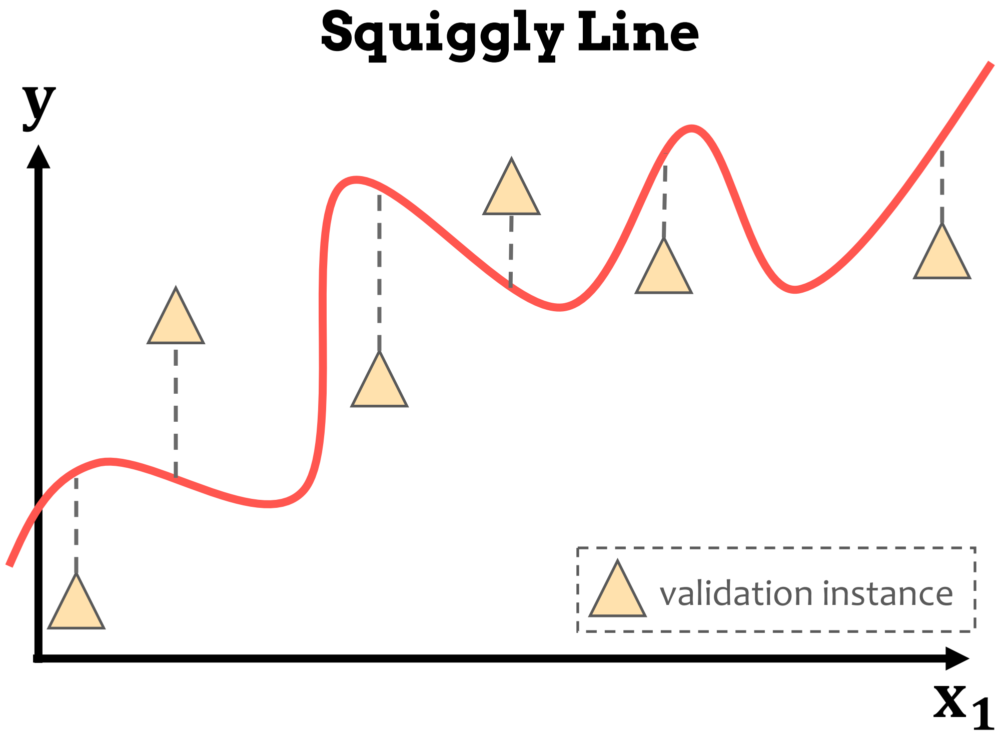
(relatively) **high bias**

Squiggly Line

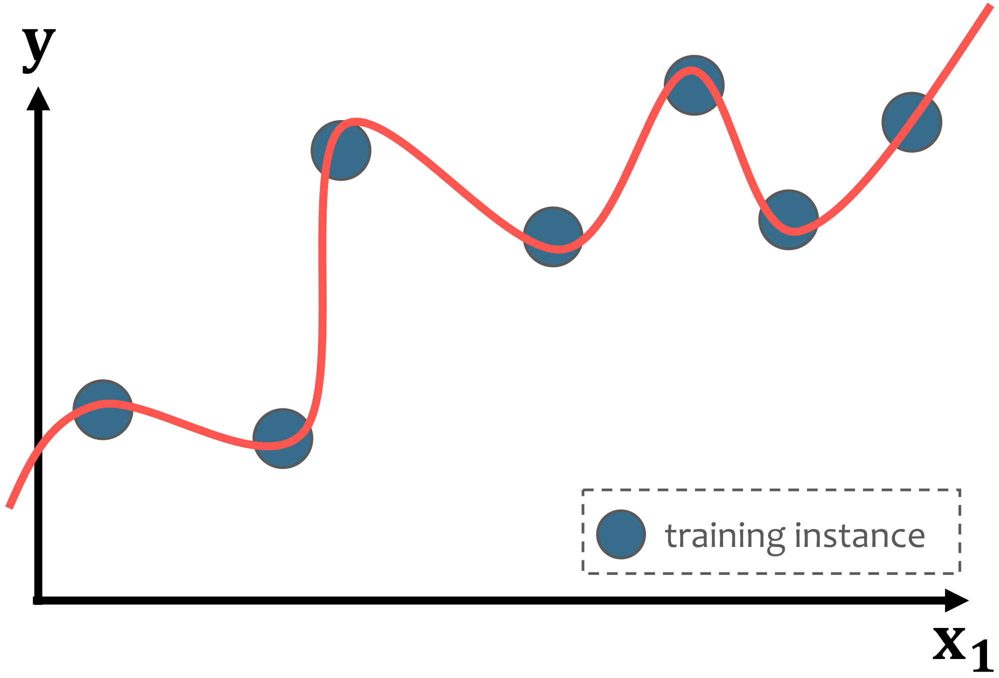
(very) **low bias**

In contrast, the considered **straight line** fits the validation set (unseen data) **better** than squiggly line → better generalization

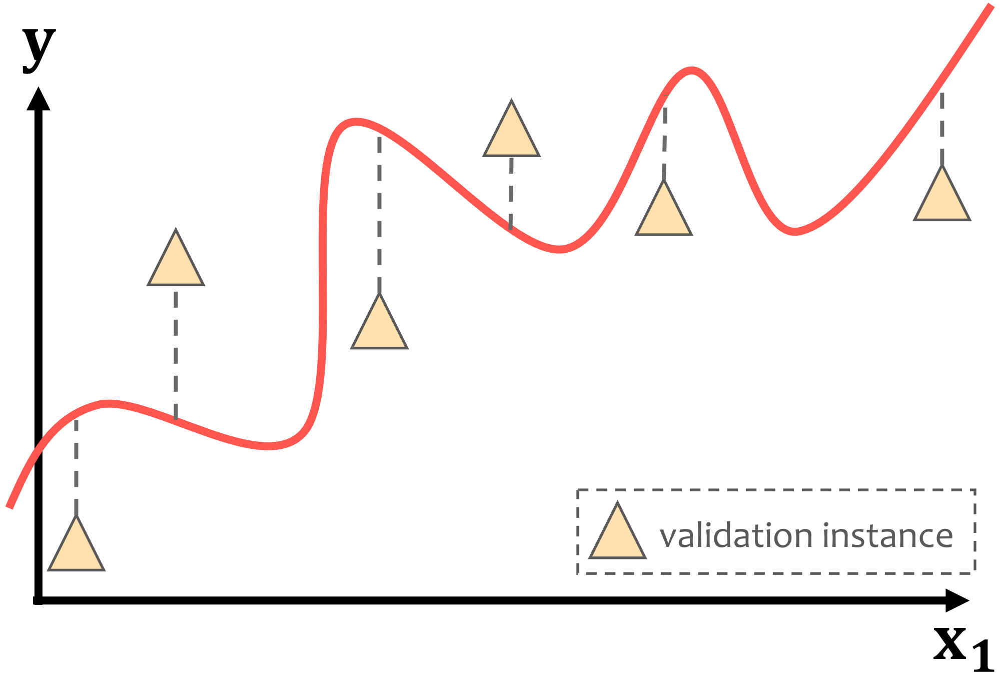




Great job fitting the **training set**



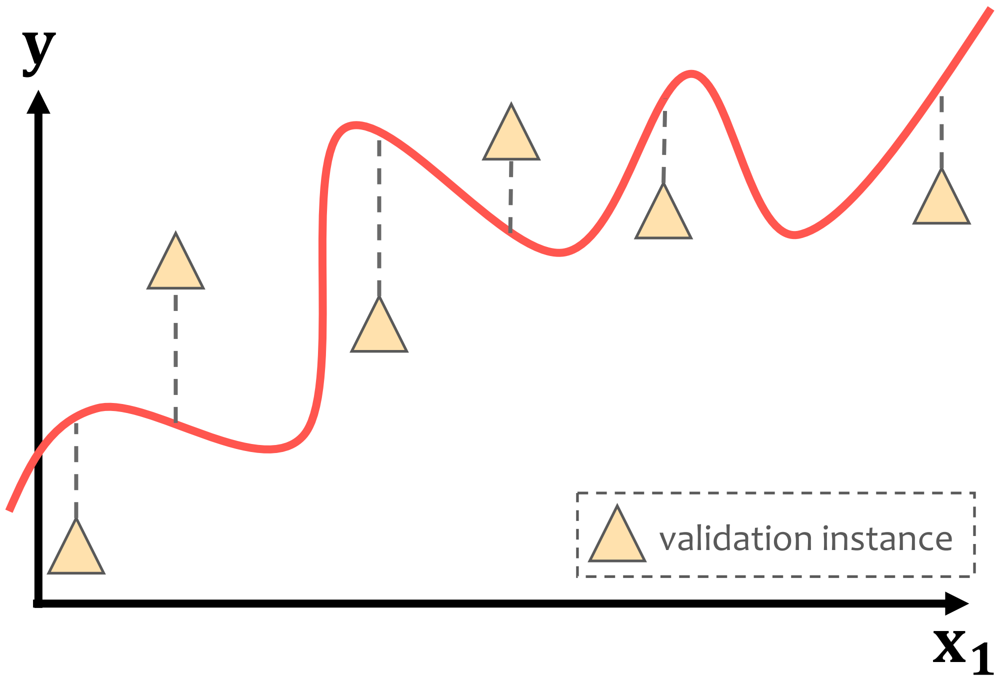
Terrible job fitting the **validation set** → **not generalize well**



In ML, the **difference in fits between datasets** is called **variance.**

Great job fitting the **training set**

Terrible job fitting the **validation set** → **not generalize well**



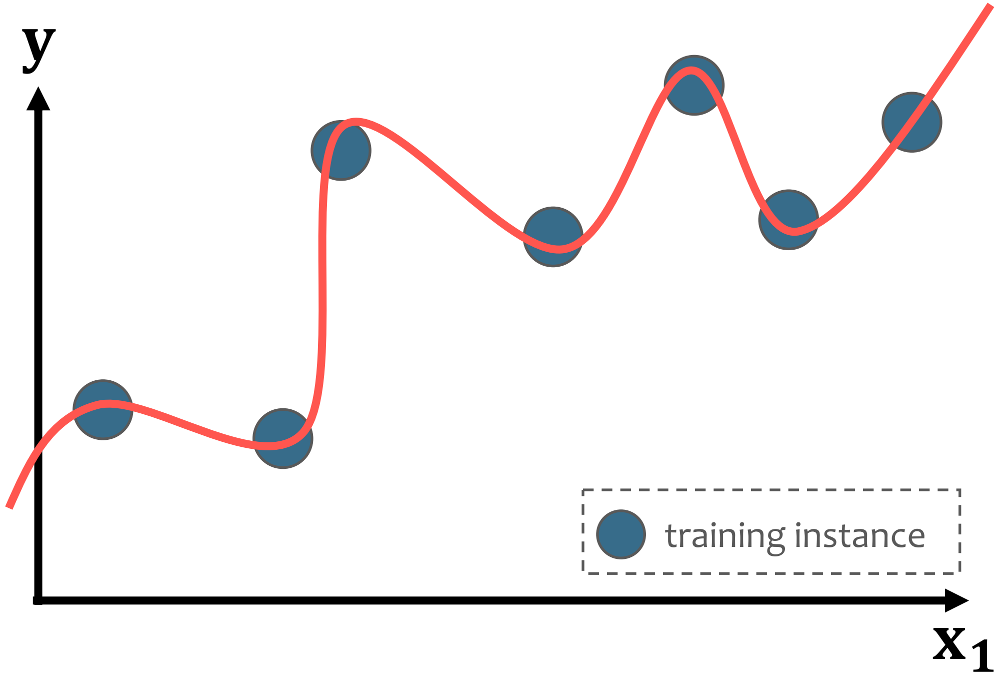
In ML, the **difference in fits between datasets** is called **variance.**

also

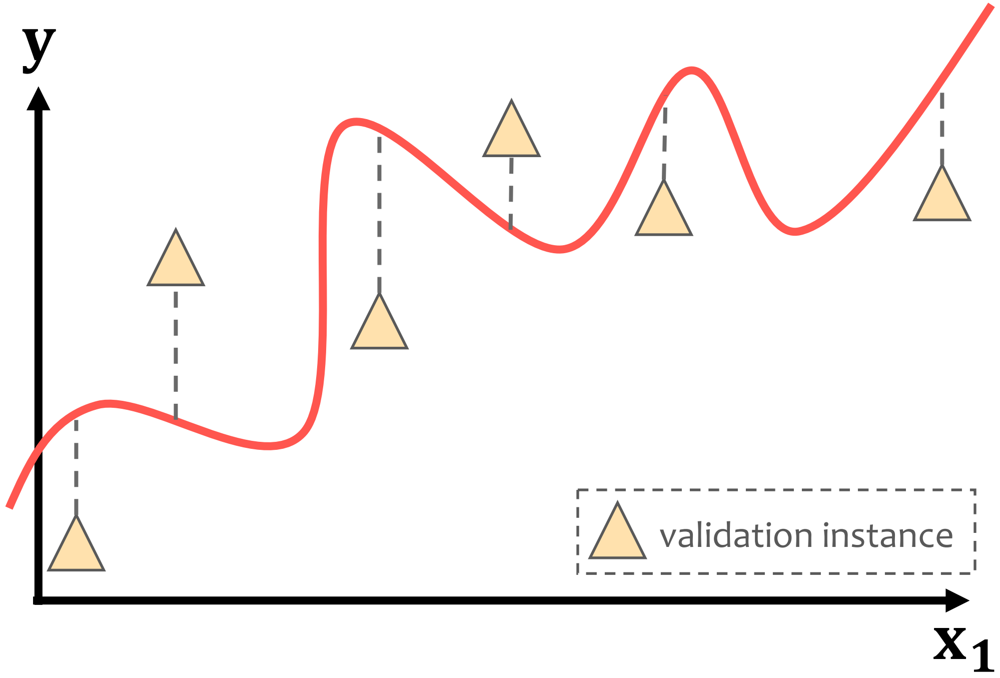
Variance

The amount that the **estimate** of the model will **change** if **different training data** was used.

Great job fitting the **training set**



Terrible job fitting the **validation set** → **not generalize well**



In ML, the **difference in fits between datasets** is called **variance.**

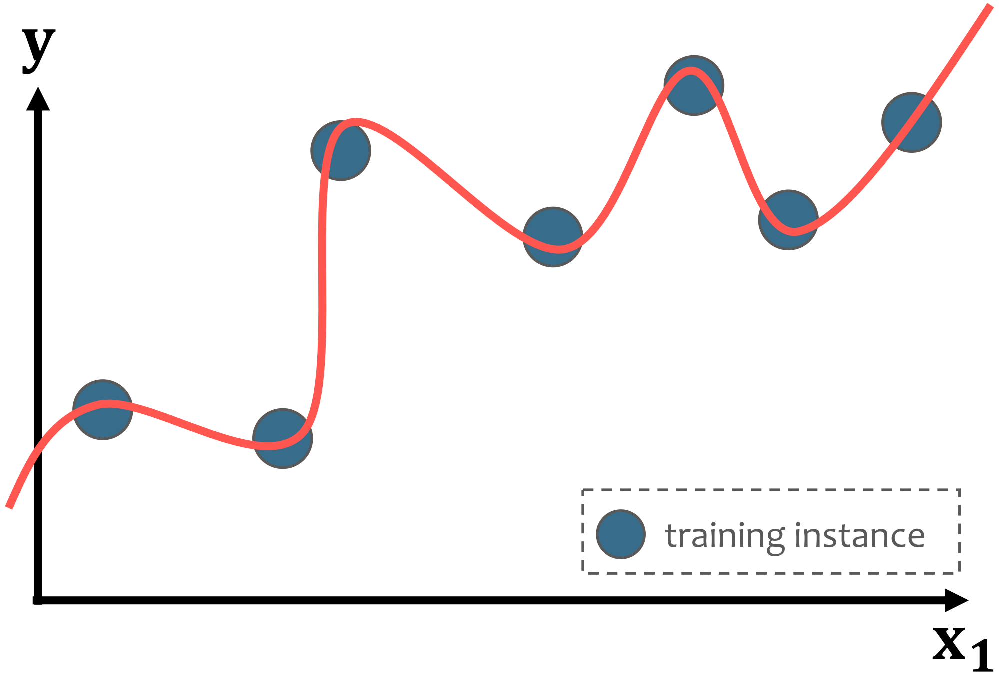
also

Variance

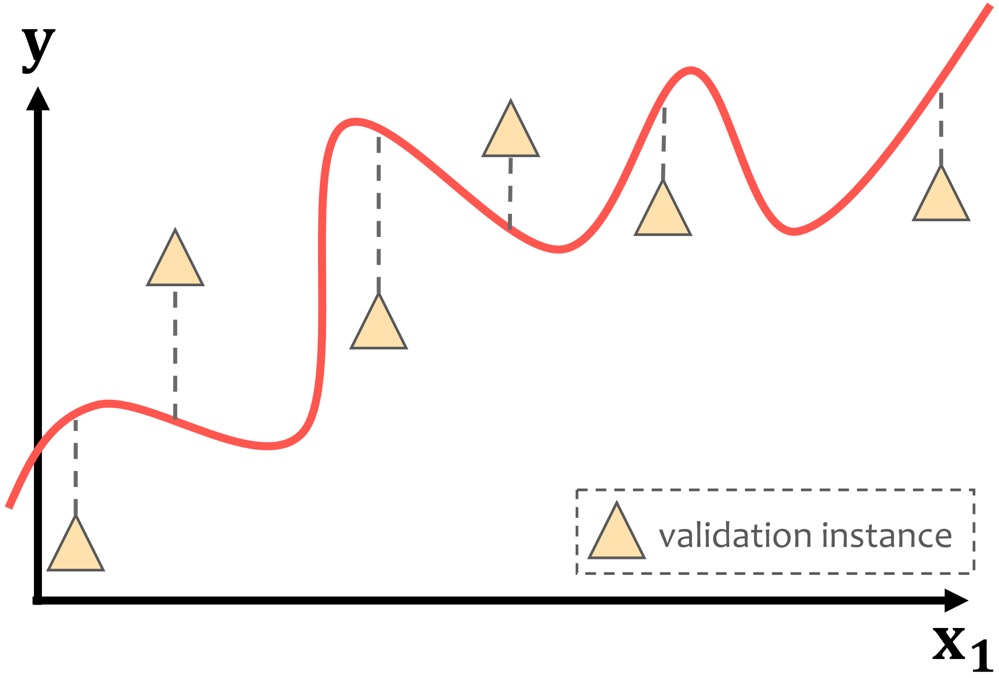
The amount that the **estimate** of the model will **change** if **different training data** was used.

A **high-variance model** is most likely to **overfit** the training data.

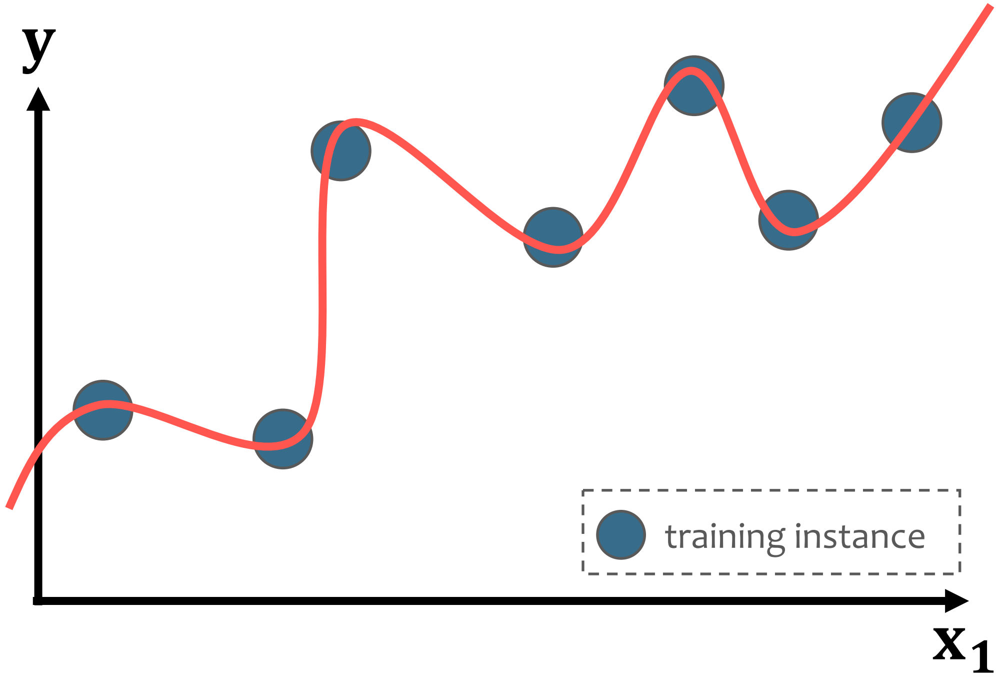
Great job fitting the **training set**

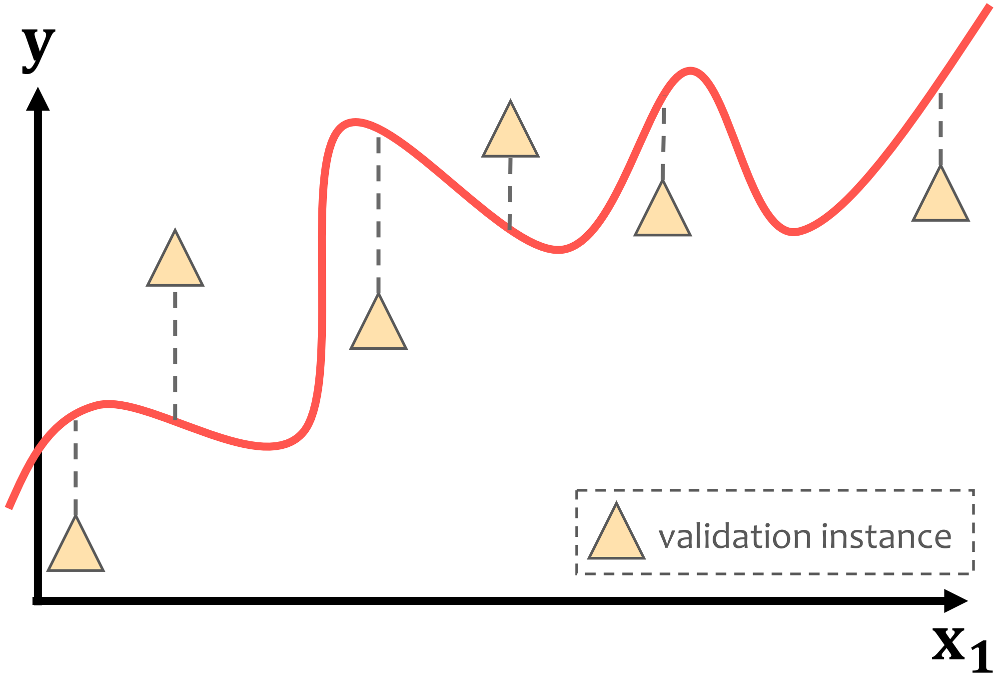


Terrible job fitting the **validation set** → **not generalize well**

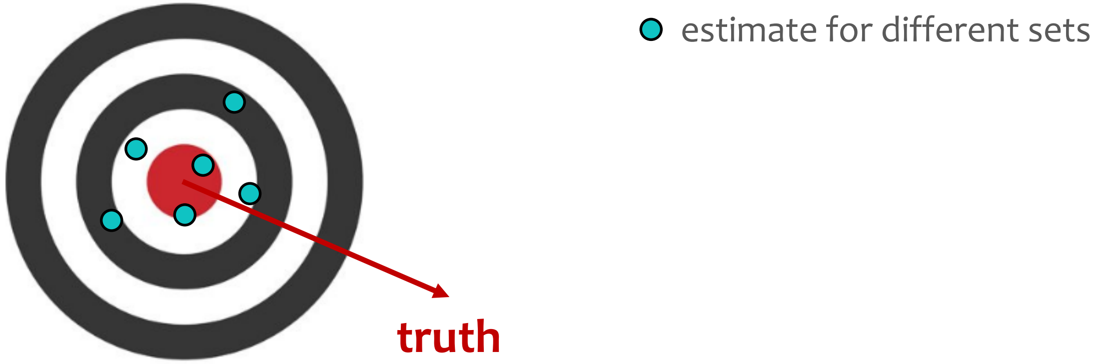


Squiggly Line



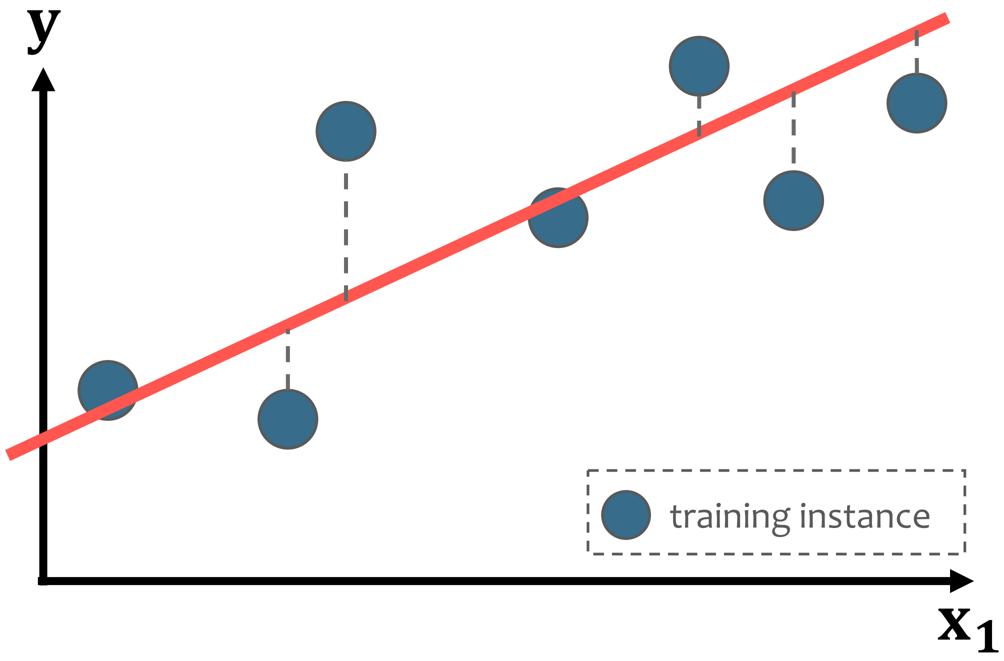


low bias
high variance

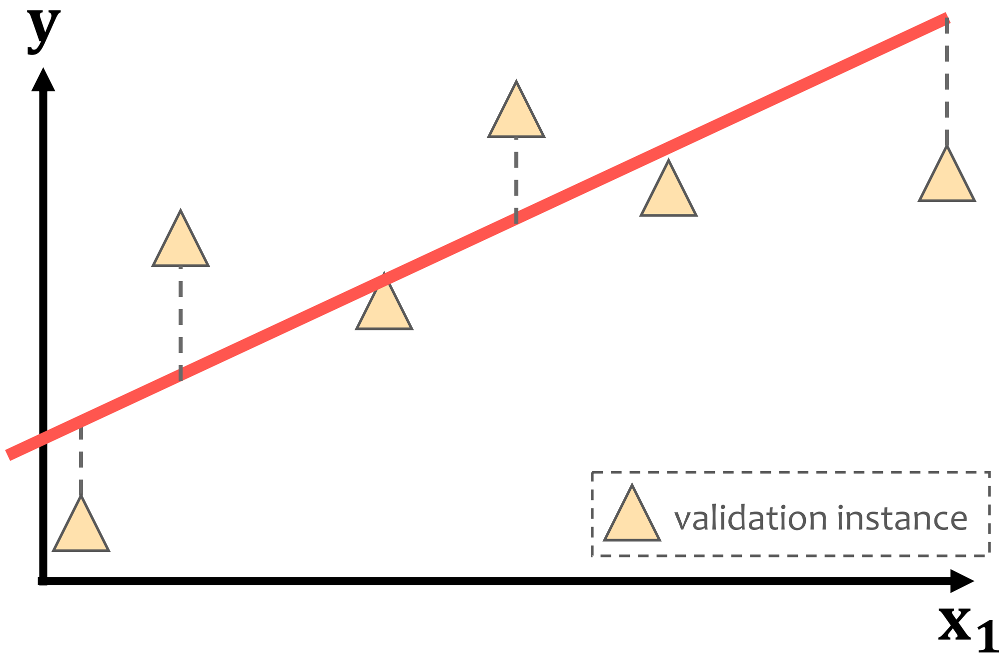


Straight Line

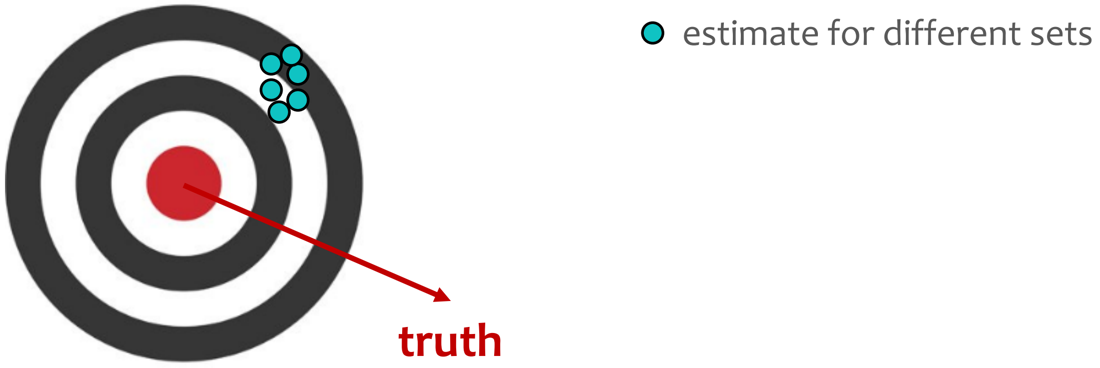
It cannot capture the **true relationship** between $\mathbf{x_1}$ and $\mathbf{y}$.

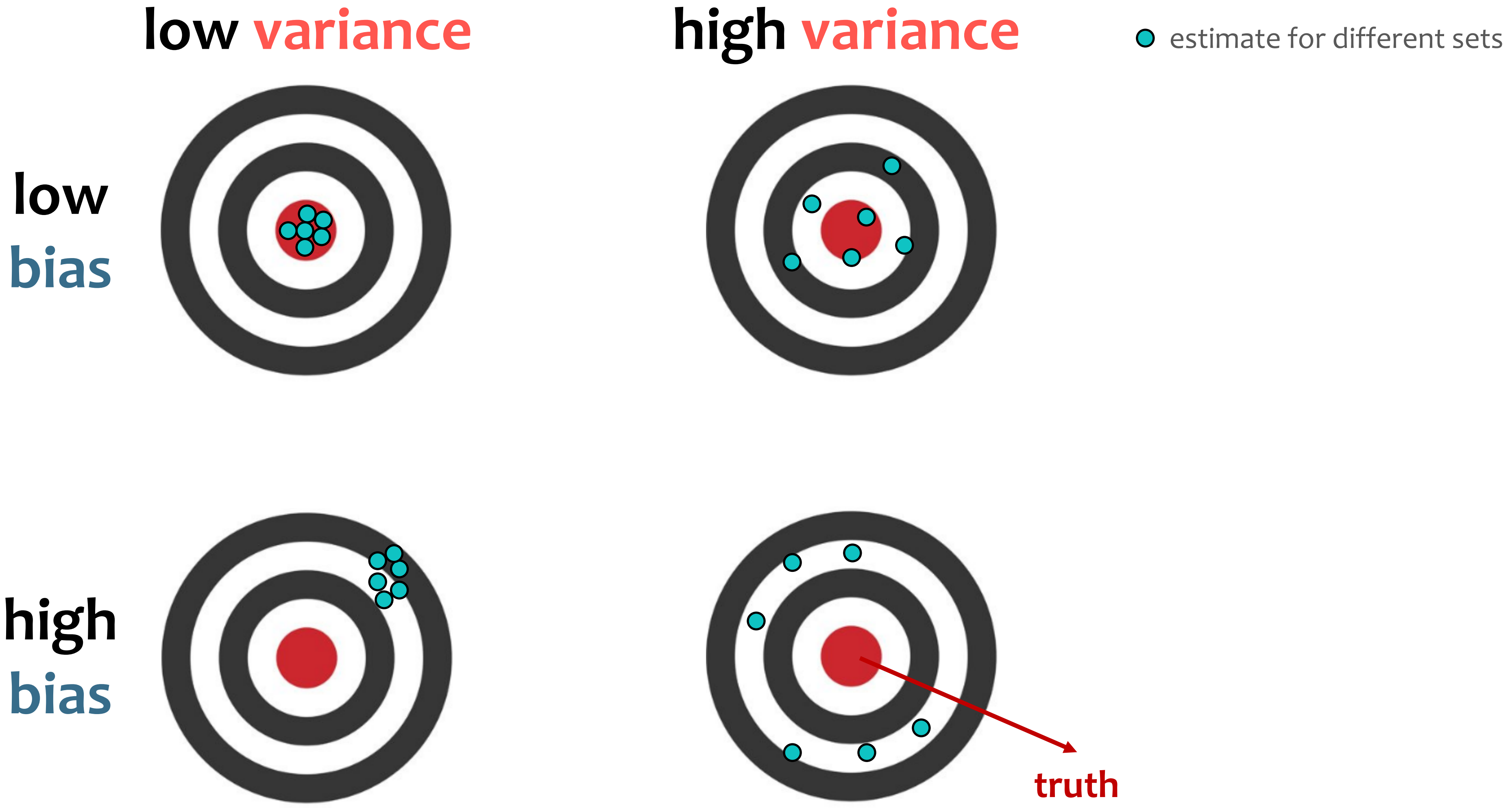


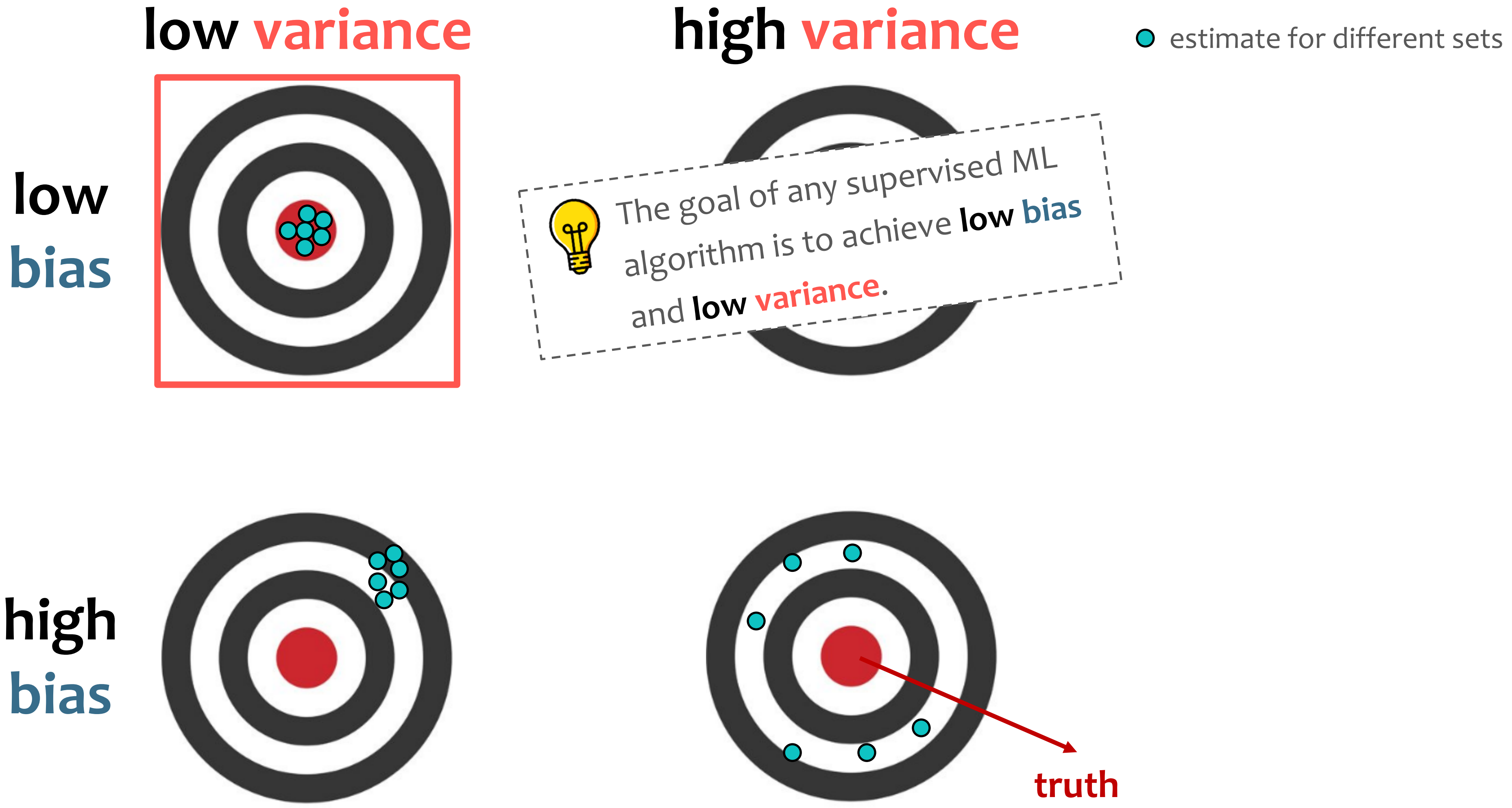
But the **error is similar** for different sets.



(relatively)
high bias
low variance




low variance
high variance
estimate for different sets
low
bias
high
bias
truth


low variance
high variance
estimate for different sets
low bias
high bias
The goal of any supervised ML algorithm is to achieve low bias and low variance.
truth

The Bias-Variance Trade-off

- **Increasing ↑ a model's complexity** will typically:
  - increase ↑ its **variance**
  - reduce ↓ its **bias**.
- **Reducing ↓ a model's complexity** will:
  - increase ↑ its **bias**
  - reduce ↓ its **variance**.

The Bias-Variance Trade-off

- **Increasing ↑ a model's complexity** will typically:
  - increase ↑ its **variance**
  - reduce ↓ its **bias**.
- **Reducing ↓ a model's complexity** will:
  - increase ↑ its **bias**
  - reduce ↓ its **variance**.

or simply

- Increasing ↑ **variance** reduces ↓ **bias**, and vice versa.
- Reducing ↓ **variance** increases ↑ **bias**, and vice versa.

The Bias-Variance Trade-off

- **Increasing ↑ a model's complexity** will typically:
  - increase ↑ its variance
  - reduce ↓ its bias.
- **Reducing ↓ a model's complexity** will:
  - increase ↑ its bias
  - reduce ↓ its variance.

or simply

- Increasing ↑ variance reduces ↓ bias, and vice versa.
- Reducing ↓ variance increases ↑ bias, and vice versa.

- This is why it is called a **trade-off**.

The Bias-Variance Trade-off

- **Increasing ↑ a model's complexity** will typically:
  - increase ↑ its **variance**
  - reduce ↓ its **bias**.
- **Reducing ↓ a model's complexity** will:
  - increase ↑ its **bias**
  - reduce ↓ its **variance**.

or simply

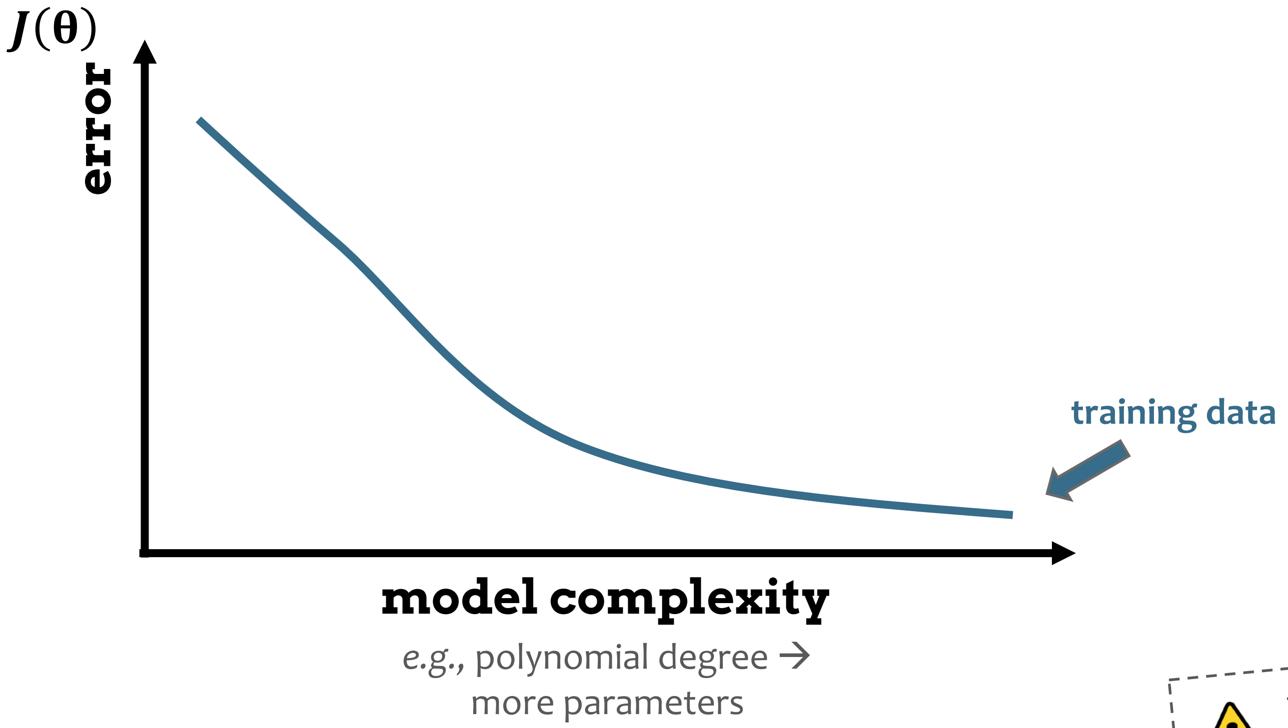
- Increasing ↑ **variance** reduces ↓ **bias**, and vice versa.
- Reducing ↓ **variance** increases ↑ **bias**, and vice versa.

- This is why it is called a **trade-off**.

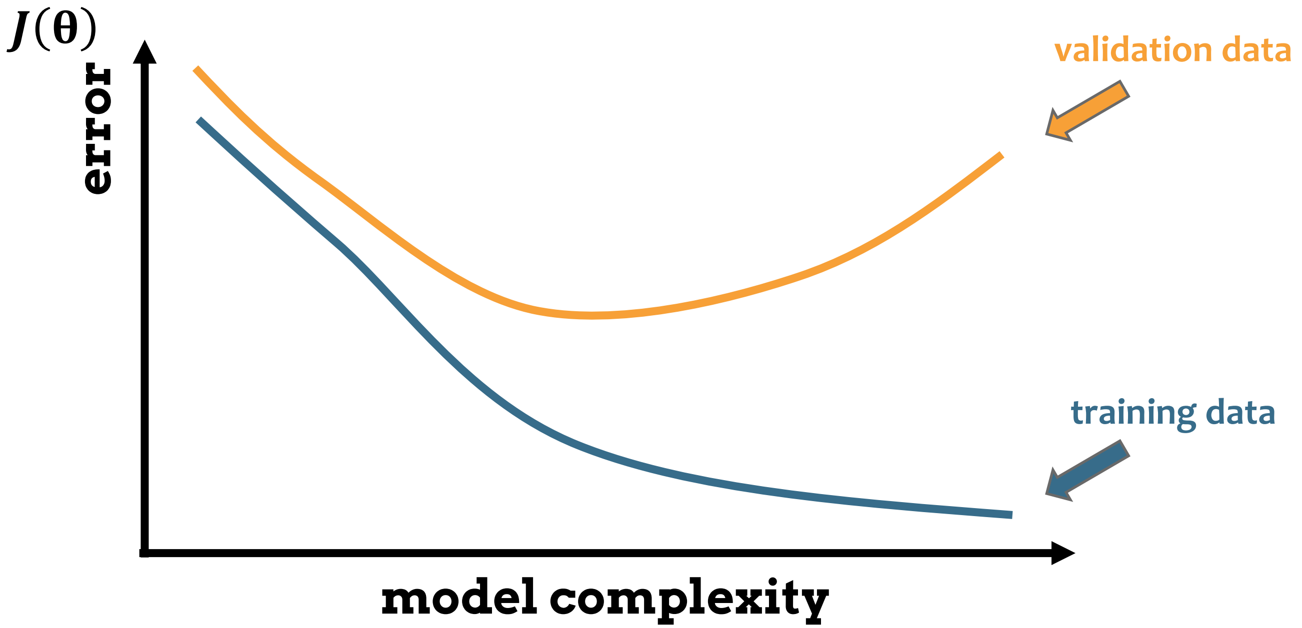
General trend:

- **Linear** ML algorithms often have a high **bias** but a low **variance**.
- **Nonlinear** ML algorithms often have a low **bias** but a high **variance**.

Diagnosing Bias vs Variance



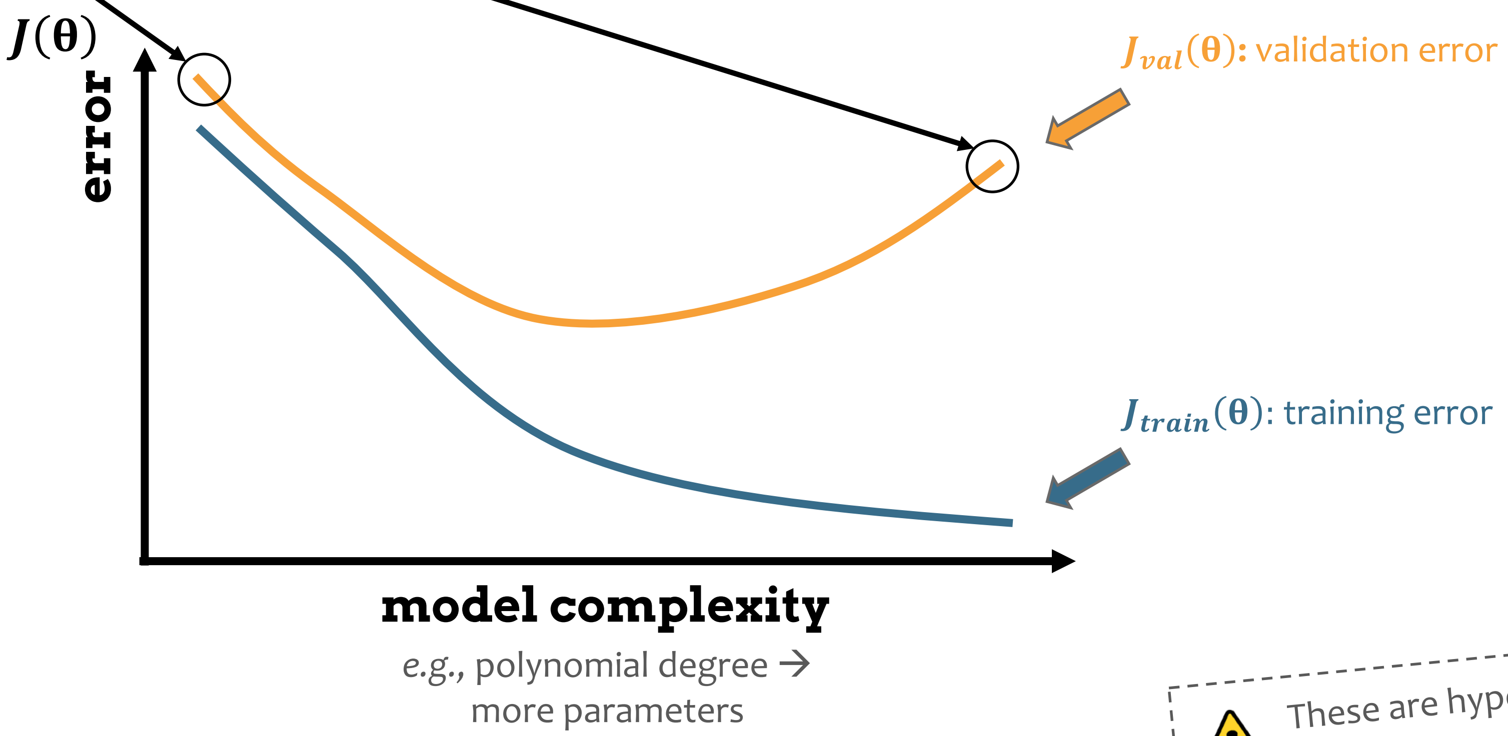
These are hypothetical curves for hypothetical data.



e.g., polynomial degree → more parameters

These are hypothetical curves for hypothetical data.

- Suppose your learning algorithm is performing less well than you were hoping.
- If $J_{val}(\theta)$ is **high**, is it a **bias** problem or a **variance** problem?



These are hypothetical curves for hypothetical data.

- Suppose your learning algorithm is performing less well than you were hoping.
- If $J_{val}(\theta)$ is **high**, is it a **bias** problem or a **variance** problem?

Variance (overfit):

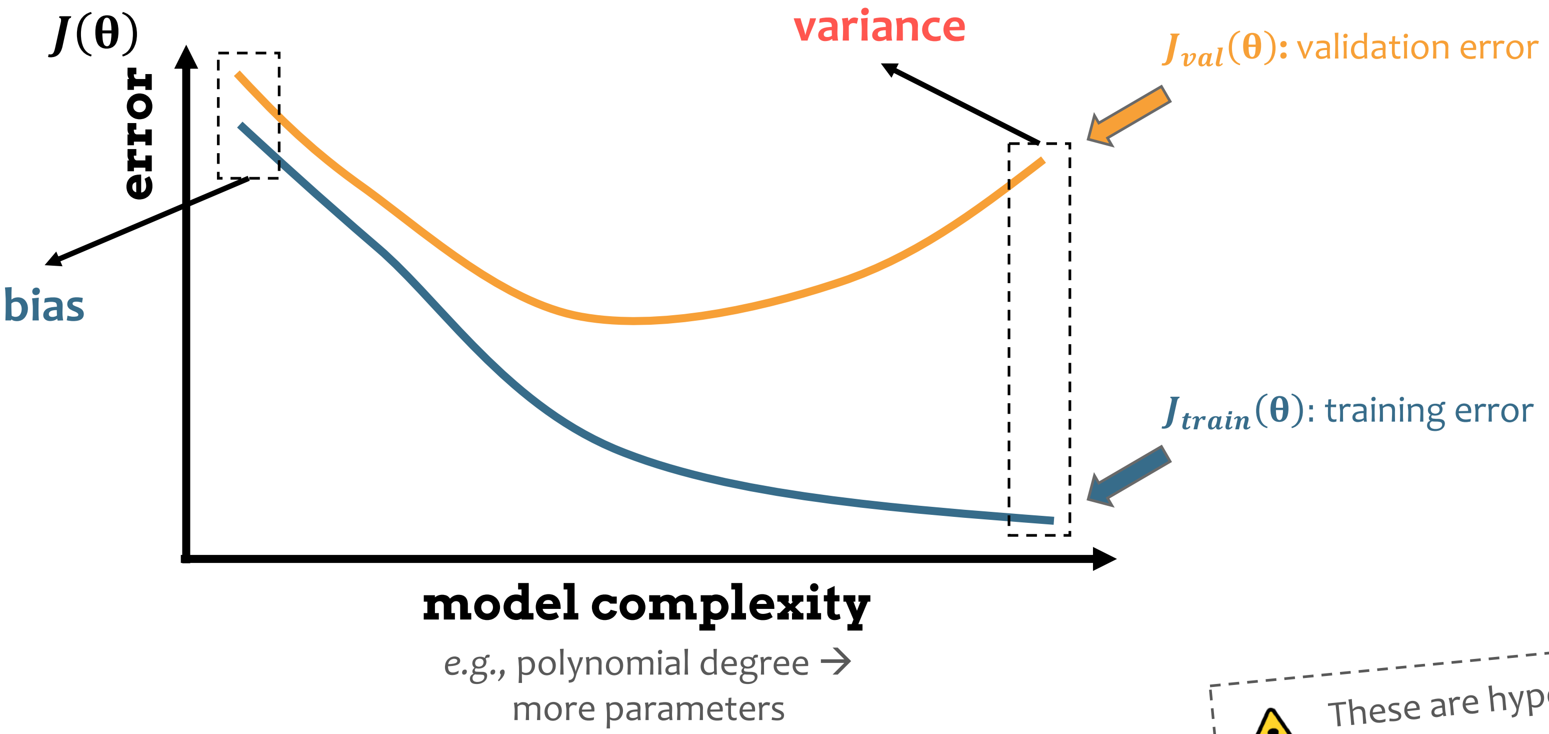
$J_{train}(\theta)$ is low

$J_{train}(\theta) \gg J_{val}(\theta)$

Bias (underfit):

$J_{train}(\theta)$ is high

$J_{train}(\theta) \approx J_{val}(\theta)$

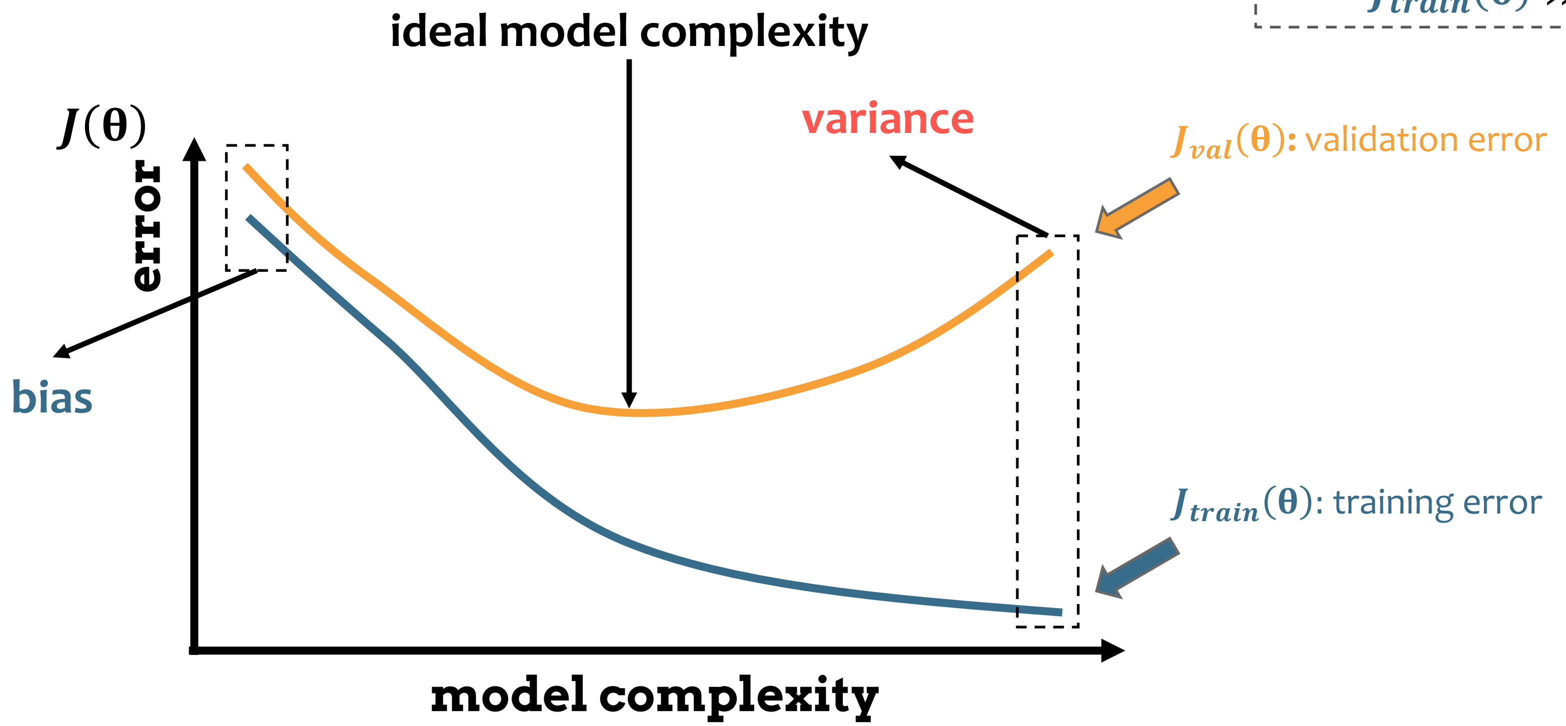


⚠ These are hypothetical curves for hypothetical data.

- Suppose your learning algorithm is performing less well than you were hoping.
- If $J_{val}(\boldsymbol{\theta})$ is **high**, is it a **bias** problem or a **variance** problem?

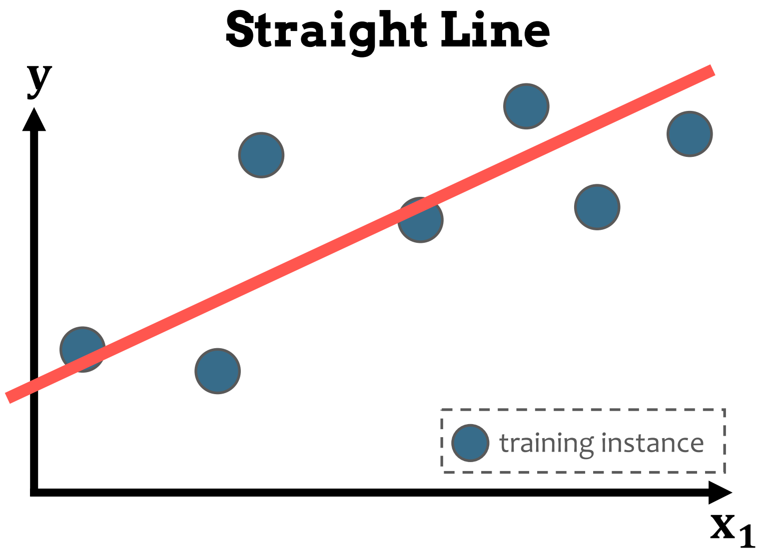
Variance (overfit):
$J_{train}(\boldsymbol{\theta})$ is low
$J_{train}(\boldsymbol{\theta}) \gg J_{val}(\boldsymbol{\theta})$

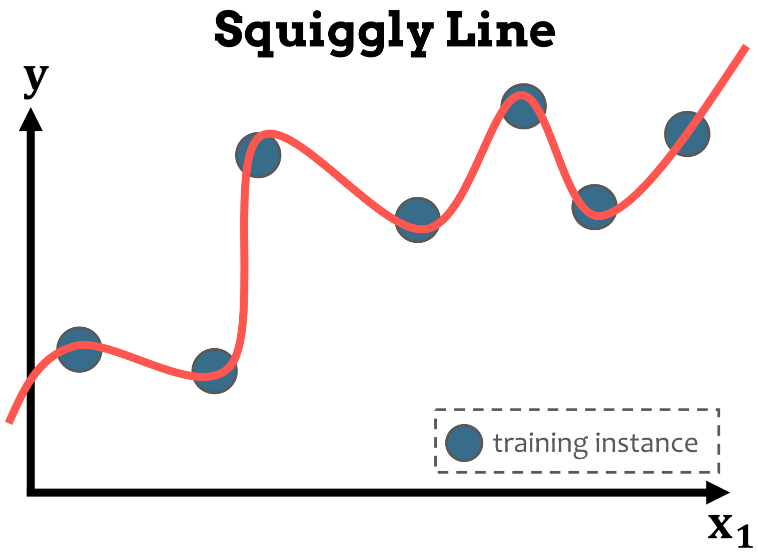
Bias (underfit):
$J_{train}(\boldsymbol{\theta})$ is high
$J_{train}(\boldsymbol{\theta}) \approx J_{val}(\boldsymbol{\theta})$

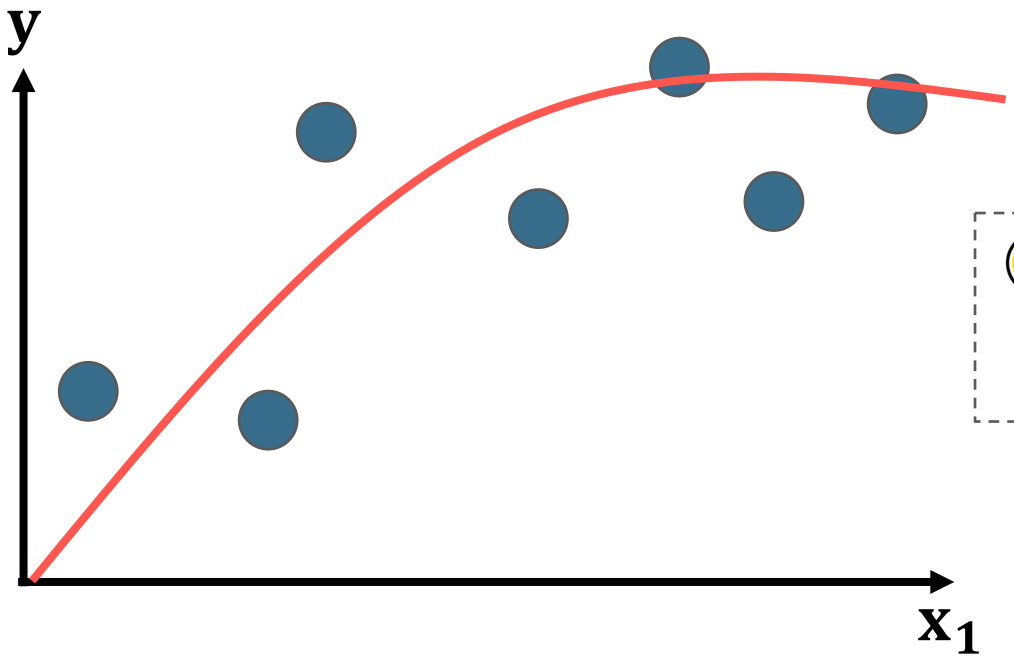


e.g., polynomial degree → more parameters

These are hypothetical curves for hypothetical data.

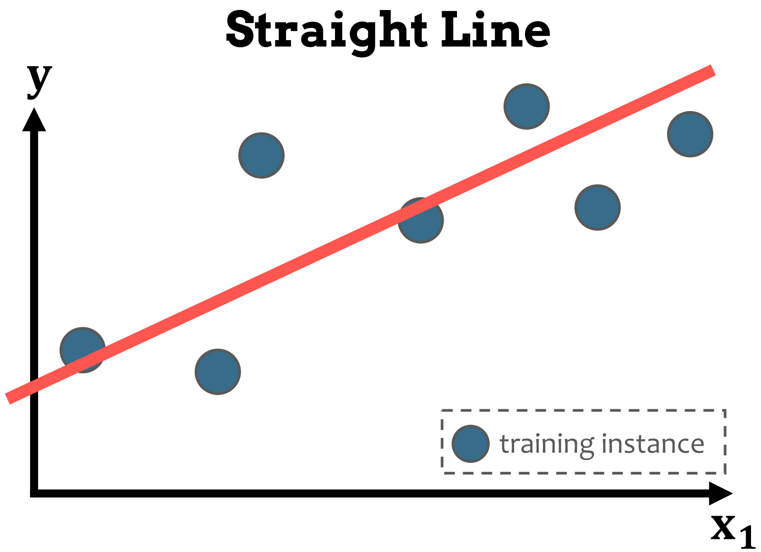


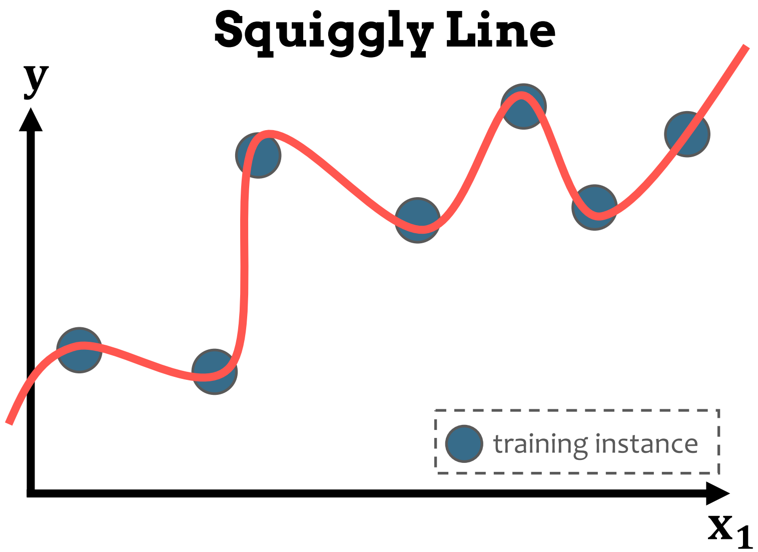




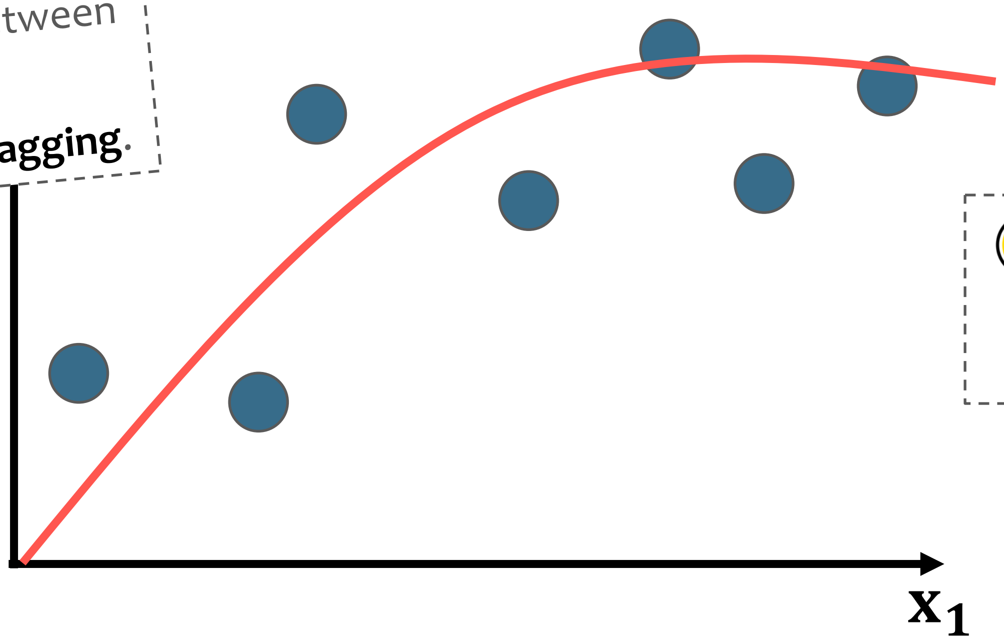
We want an **intermediate solution** between simple and overly complex one.

(relatively)
low bias
low variance





Three commonly strategies used to find the **intermediate model** between simple and complex are: **regularization, boosting** and **bagging**.



We want an **intermediate solution** between simple and overly complex one.

(relatively)

low bias

low variance

Aprendizado de Máquina e Reconhecimento de Padrões 2021.2

The Bias-Variance Trade-off

Prof. Dr. Samuel Martins (Samuka)
samuel.martins@ifsp.edu.br

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
São Paulo
Campus Campinas