General Guidance

Hung-yi Lee 李宏毅

Framework of ML

Training data: $\left\{\left(\boldsymbol{x}^{\mathbf{1}}, \hat{y}^{1}\right),\left(\boldsymbol{x}^{\mathbf{2}}, \hat{y}^{2}\right), \ldots,\left(\boldsymbol{x}^{\boldsymbol{N}}, \hat{y}^{N}\right)\right\}$

Testing data: $\left\{\boldsymbol{x}^{\boldsymbol{N}+\mathbf{1}}, \boldsymbol{x}^{\boldsymbol{N}+\mathbf{2}}, \ldots, \boldsymbol{x}^{\boldsymbol{N}+\boldsymbol{M}}\right\}$

Speech Recognition

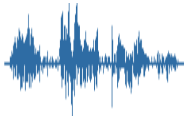
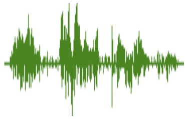
$\boldsymbol{x}$:  $\hat{y}$: phoneme

Image Recognition

$\boldsymbol{x}$: $\hat{y}$: soup

Speaker Recognition

$\boldsymbol{x}$:  $\hat{y}$: John (speaker)

Machine Translation

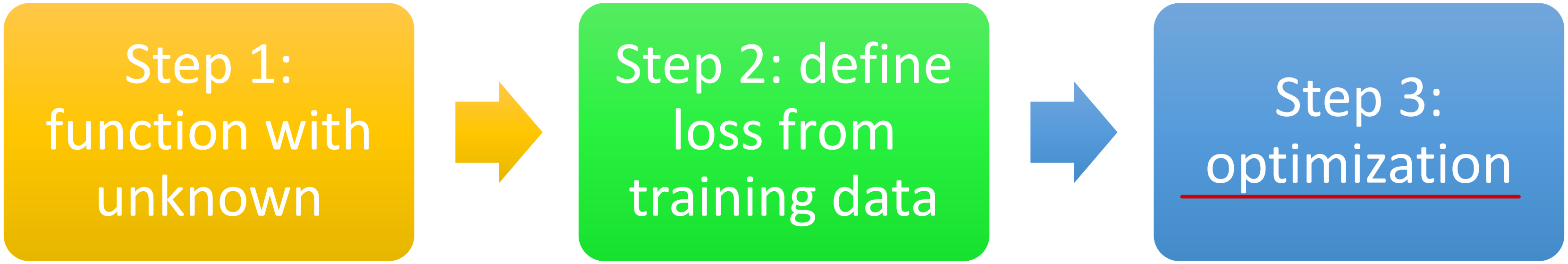
$\boldsymbol{x}$: 痛みを知れ

$\hat{y}$: 了解痛苦吧

Framework of ML

Training data: $\{(\boldsymbol{x}^1, \hat{y}^1), (\boldsymbol{x}^2, \hat{y}^2), \dots, (\boldsymbol{x}^N, \hat{y}^N)\}$

Training:



Step 1: function with unknown

Step 2: define loss from training data

Step 3: optimization

$y = f_{\boldsymbol{\theta}}(\boldsymbol{x})$

$L(\boldsymbol{\theta})$

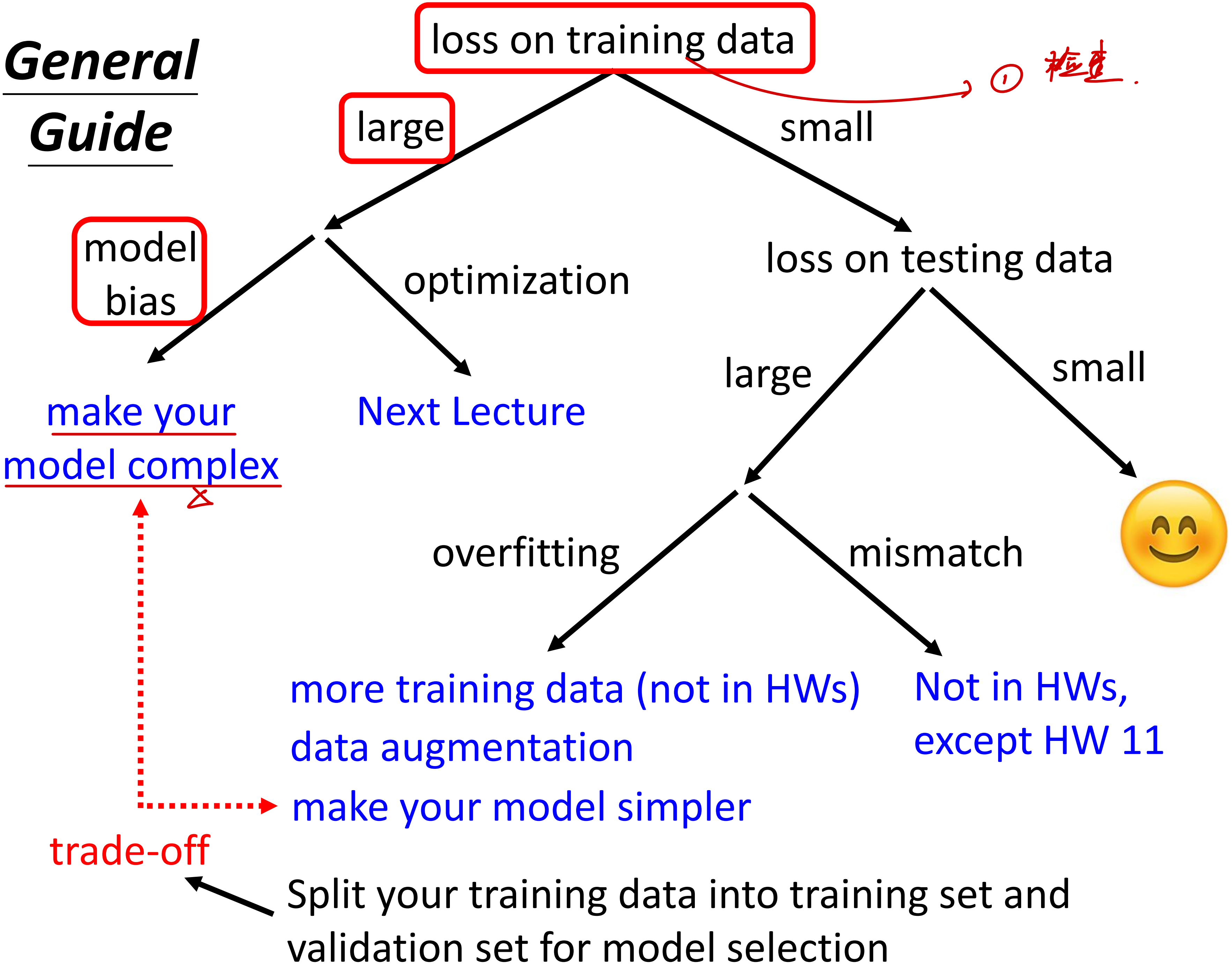
$\boldsymbol{\theta}^* = arg \min_{\boldsymbol{\theta}} L$

Testing data: $\{\boldsymbol{x}^{N+1}, \boldsymbol{x}^{N+2}, \dots, \boldsymbol{x}^{N+M}\}$

Use $y = f_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ to label the testing data

$\{y^{N+1}, y^{N+2}, \dots, y^{N+M}\}$ → Upload to Kaggle

General Guide



loss on training data

① 检查.

large

model bias

make your model complex

optimization

Next Lecture

small

loss on testing data

large

overfitting

more training data (not in HWs)

data augmentation

make your model simpler

mismatch

Not in HWs, except HW 11

small

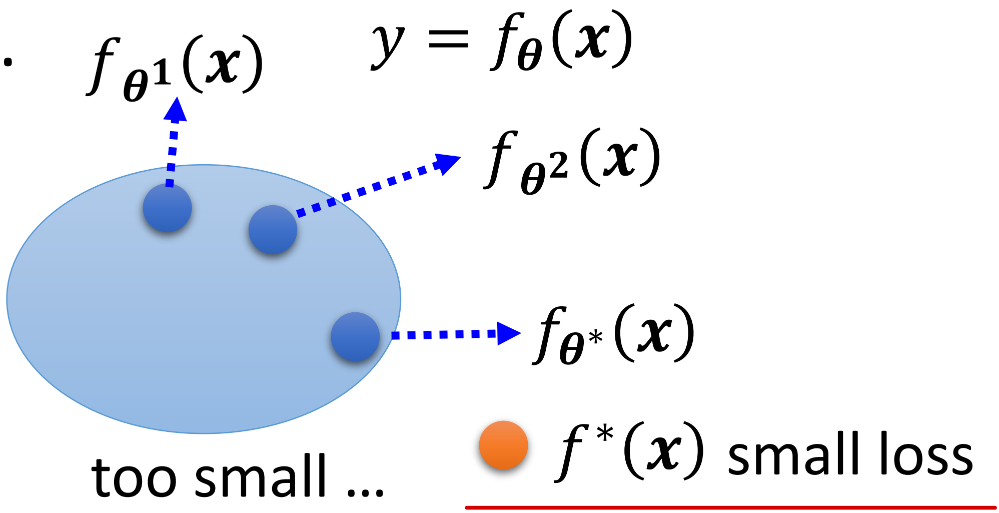
😊

trade-off

Split your training data into training set and validation set for model selection

Model Bias

- The model is too simple.



$y = f_{\boldsymbol{\theta}}(\boldsymbol{x})$
$f_{\boldsymbol{\theta}^1}(\boldsymbol{x})$
$f_{\boldsymbol{\theta}^2}(\boldsymbol{x})$
$f_{\boldsymbol{\theta}^*}(\boldsymbol{x})$
$f^*(\boldsymbol{x})$ small loss
too small ...

find a needle in a haystack ...

... but there is no needle

- Solution: redesign your model to make it more flexible

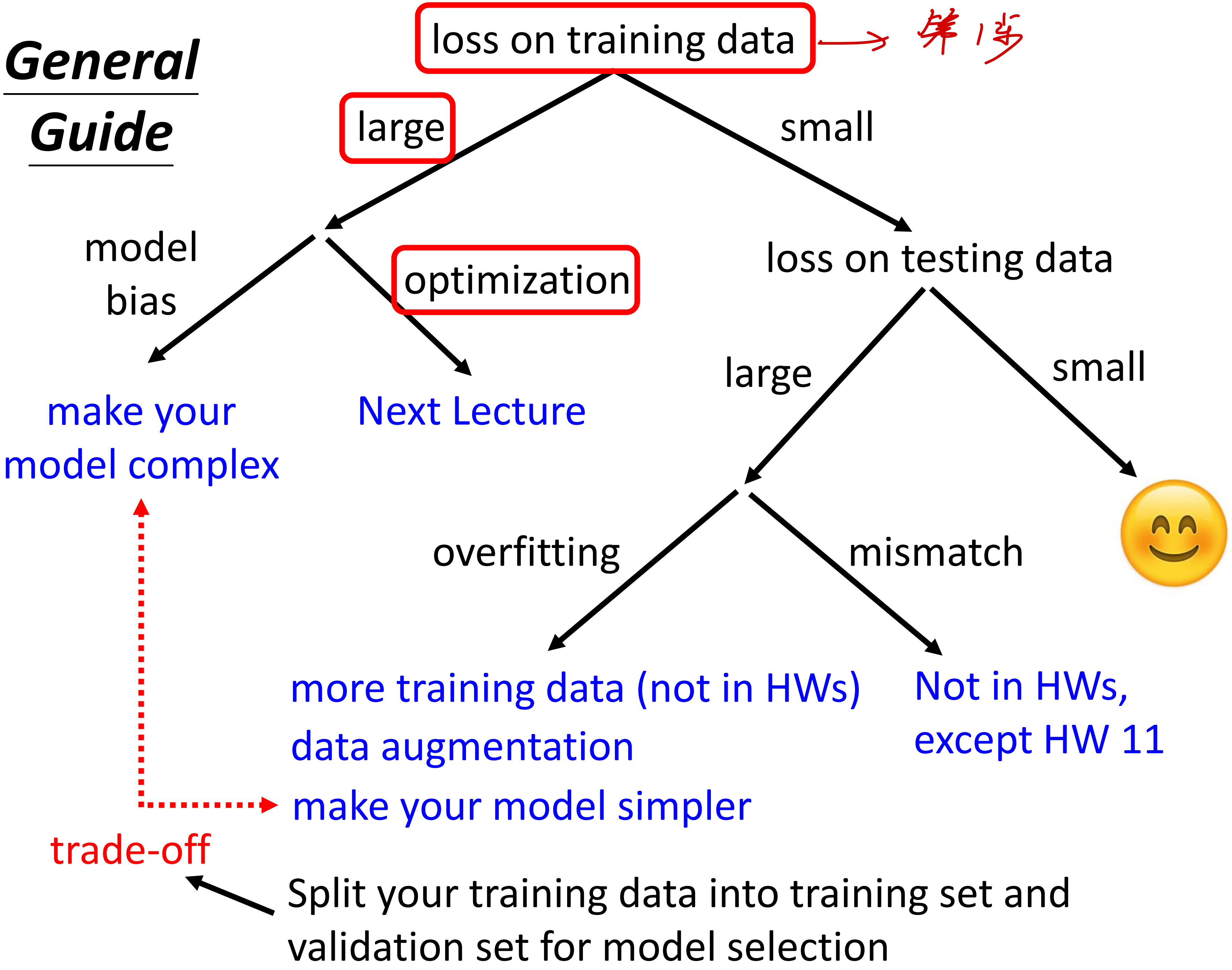
$$y = b + w x_1 \xrightarrow{\text{More features}} y = b + \sum_{j=1}^{56} w_j x_j$$

Deep Learning

(more neurons, layers)

$$y = b + \sum_i c_i \, sigmoid\left(b_i + \sum_j w_{ij} x_j\right)$$

General Guide



loss on training data → 第1步

large

model bias

make your model complex

optimization

Next Lecture

small

loss on testing data

large

overfitting

more training data (not in HWs)

data augmentation

make your model simpler

mismatch

Not in HWs, except HW 11

small

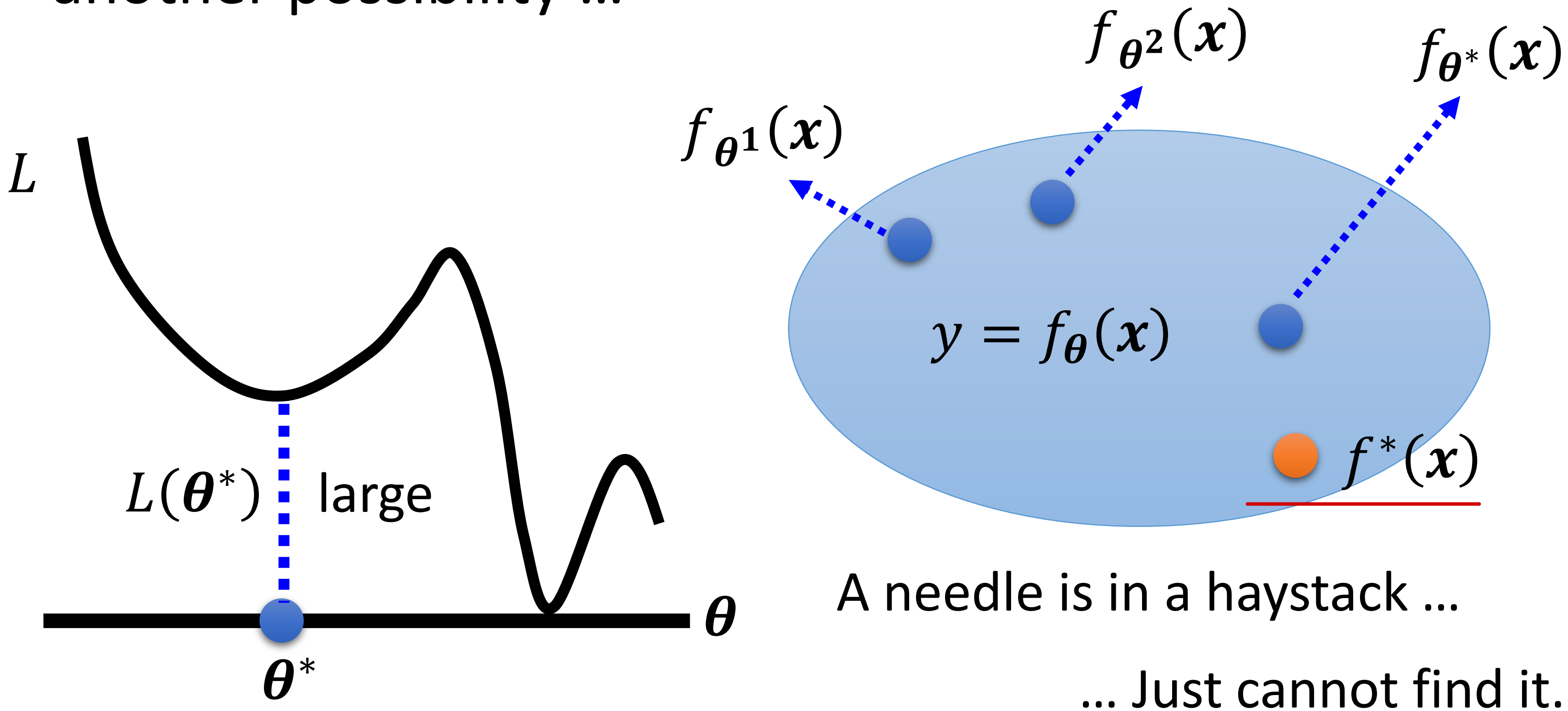
😊

trade-off

Split your training data into training set and validation set for model selection

Optimization Issue

- Large loss not always imply model bias. There is another possibility …


L
$L(\boldsymbol{\theta}^*)$
large
$\boldsymbol{\theta}$
$\boldsymbol{\theta}^*$
$f_{\boldsymbol{\theta}^1}(\boldsymbol{x})$
$f_{\boldsymbol{\theta}^2}(\boldsymbol{x})$
$f_{\boldsymbol{\theta}^*}(\boldsymbol{x})$
$y = f_{\boldsymbol{\theta}}(\boldsymbol{x})$
$f^*(\boldsymbol{x})$

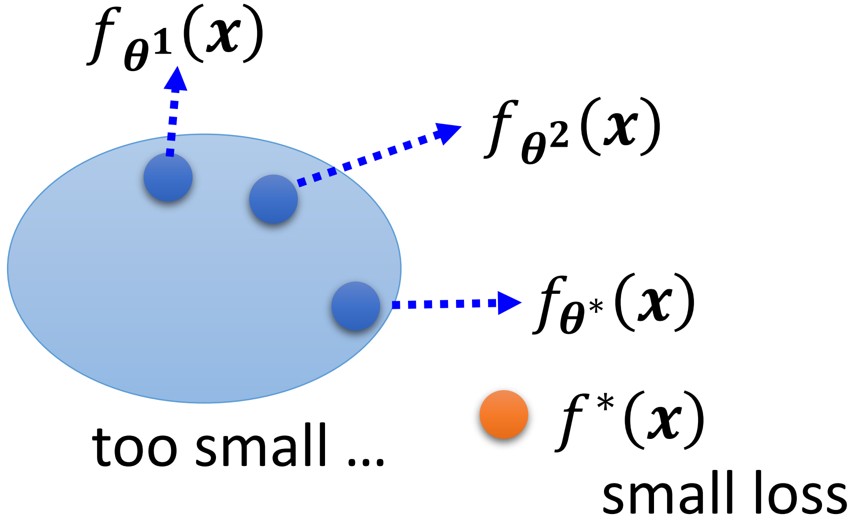
A needle is in a haystack …

… Just cannot find it.

Model Bias

find a needle in a haystack …

… but there is no needle



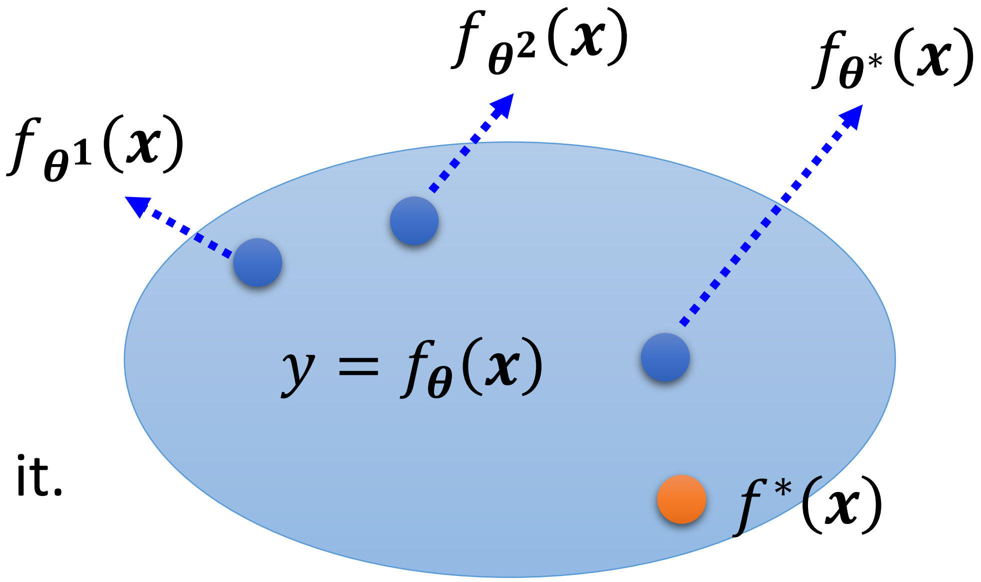
too small …

Which one???

Optimization Issue

A needle is in a haystack …

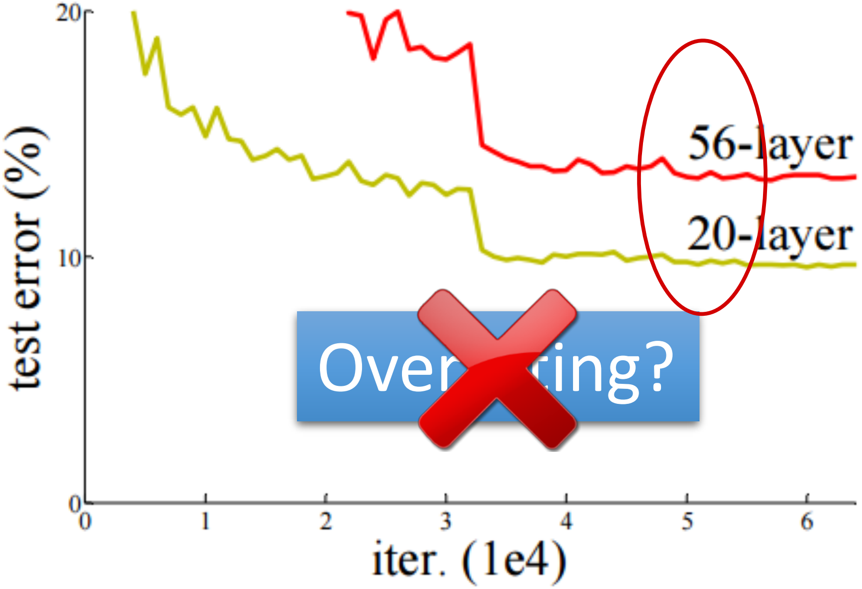
… Just cannot find it.



Ref: http://arxiv.org/abs/1512.03385

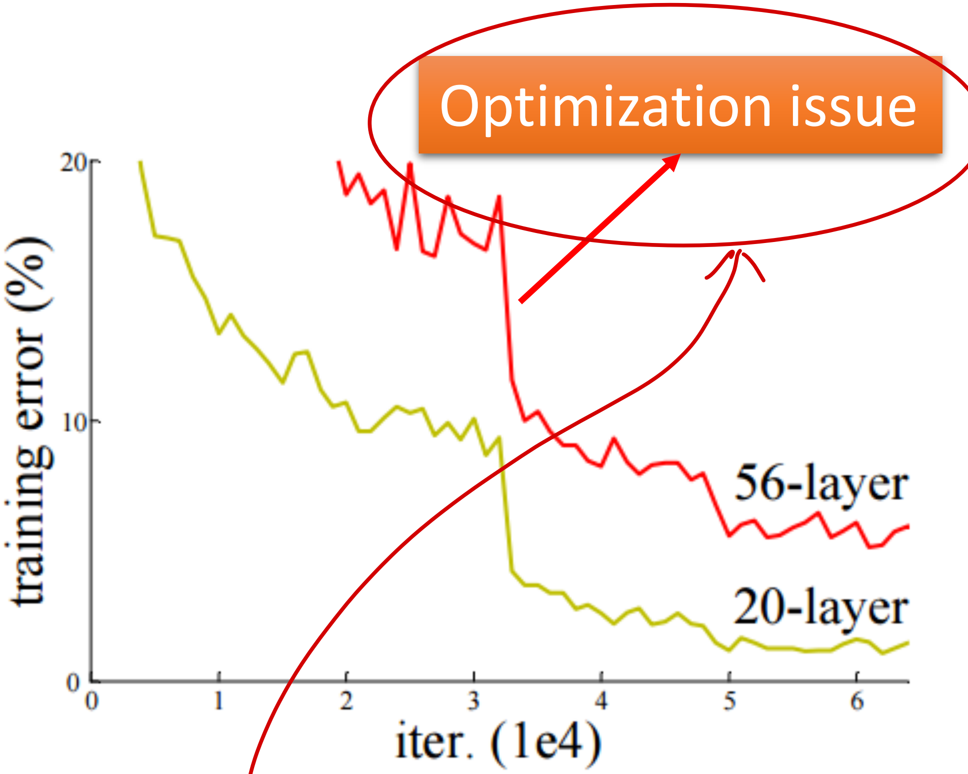
Model Bias v.s. Optimization Issue

- Gaining the insights from comparison



Overfitting?

Testing Data



Optimization issue

Training Data

Ref: http://arxiv.org/abs/1512.03385

Optimization Issue

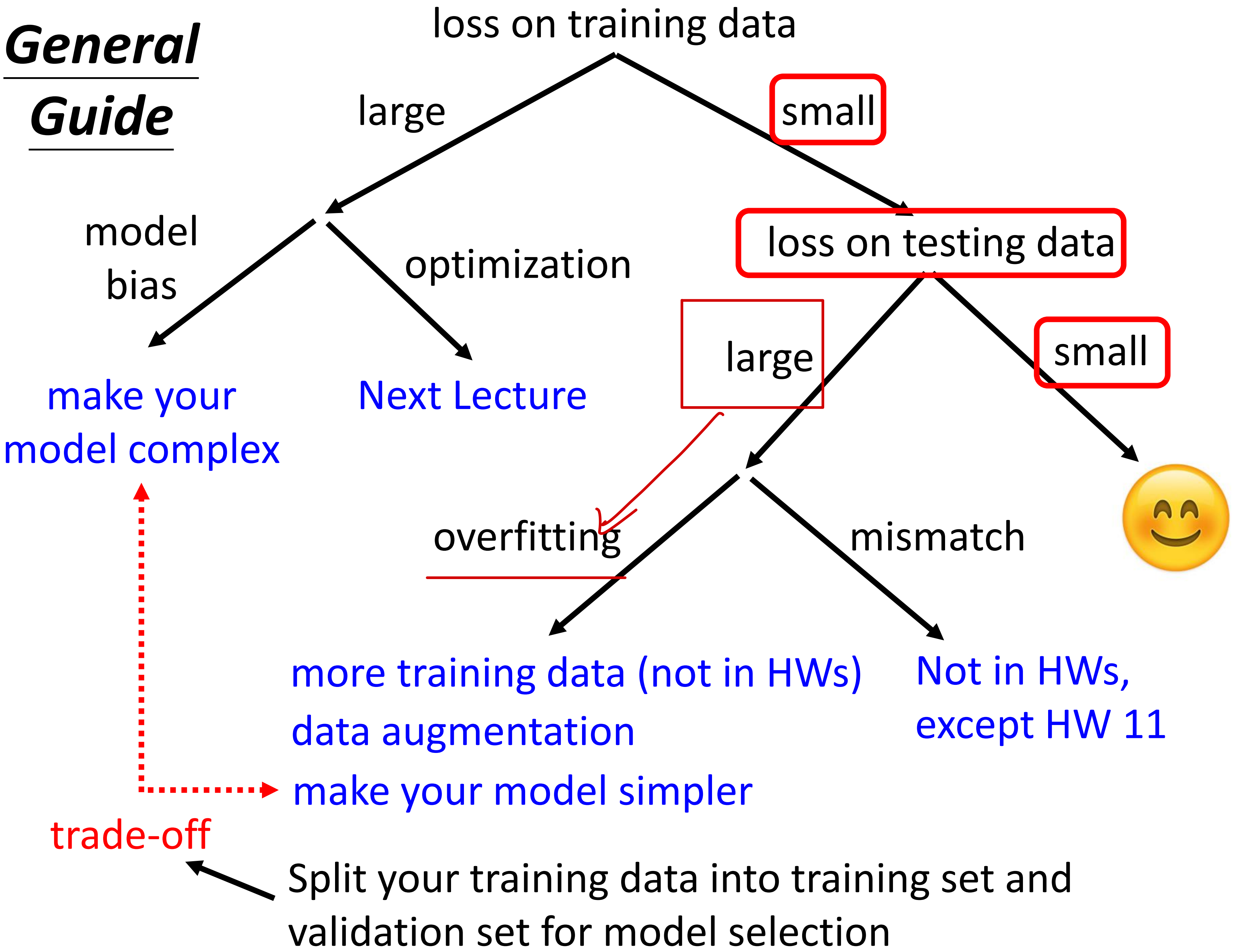
从浅的model
开始.

- Gaining the insights from comparison
- Start from shallower networks (or other models), which are easier to optimize.
- If deeper networks do not obtain smaller loss on **training data**, then there is optimization issue.

	1 layer	2 layer	3 layer	4 layer	5 layer
2017 – 2020	0.28k	0.18k	0.14k	0.10k	0.34k

- Solution: More powerful optimization technology (next lecture)

General Guide



loss on training data

large

model bias

make your model complex

optimization

Next Lecture

small

loss on testing data

large

overfitting

more training data (not in HWs)

data augmentation

make your model simpler

trade-off

Split your training data into training set and validation set for model selection

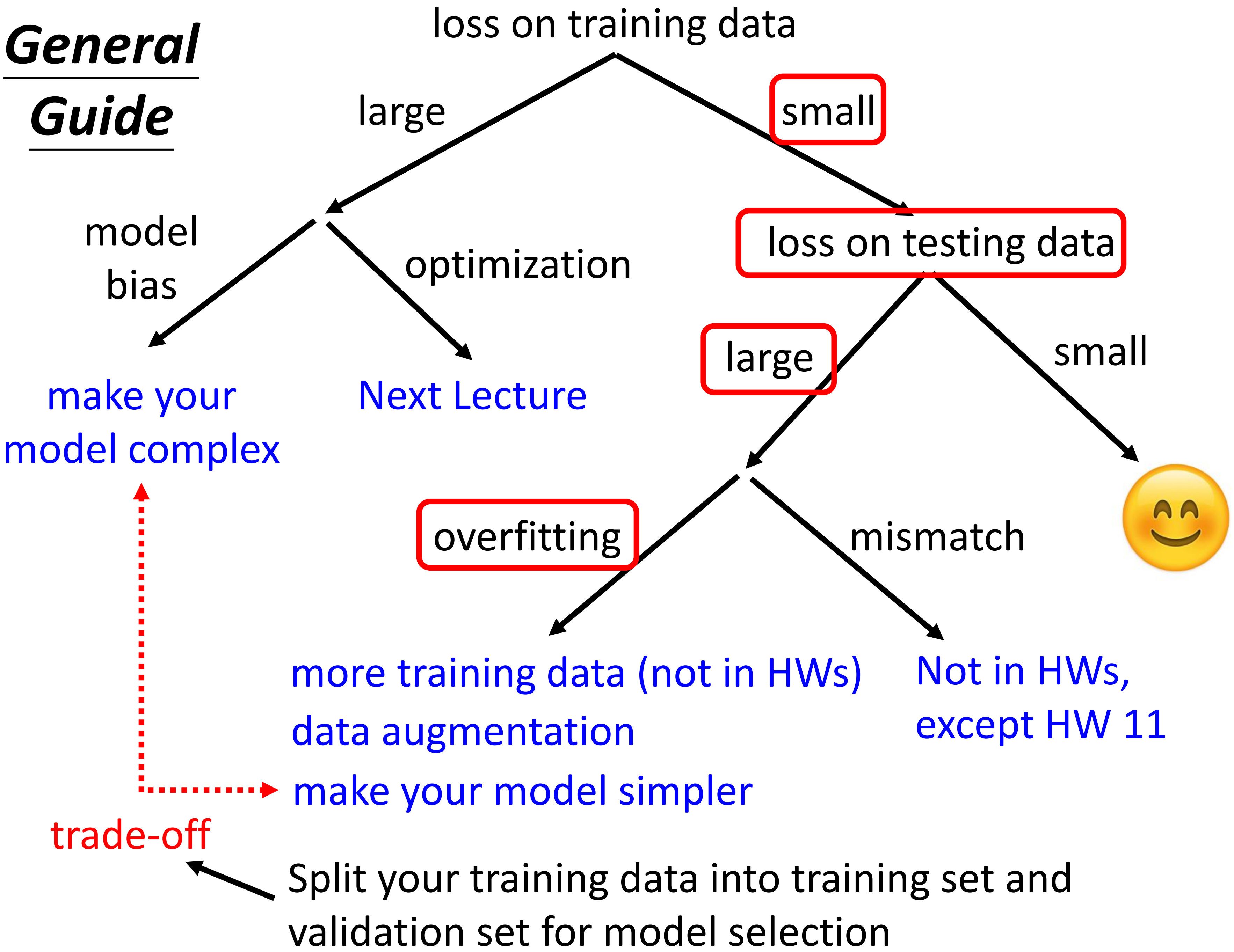
mismatch

Not in HWs, except HW 11

small

😊

General Guide



loss on training data

large

model bias

make your model complex

optimization

Next Lecture

small

loss on testing data

large

overfitting

more training data (not in HWs)

data augmentation

make your model simpler

mismatch

Not in HWs, except HW 11

small

trade-off

Split your training data into training set and validation set for model selection

Overfitting

- Small loss on training data, large loss on testing data. Why?

An extreme example

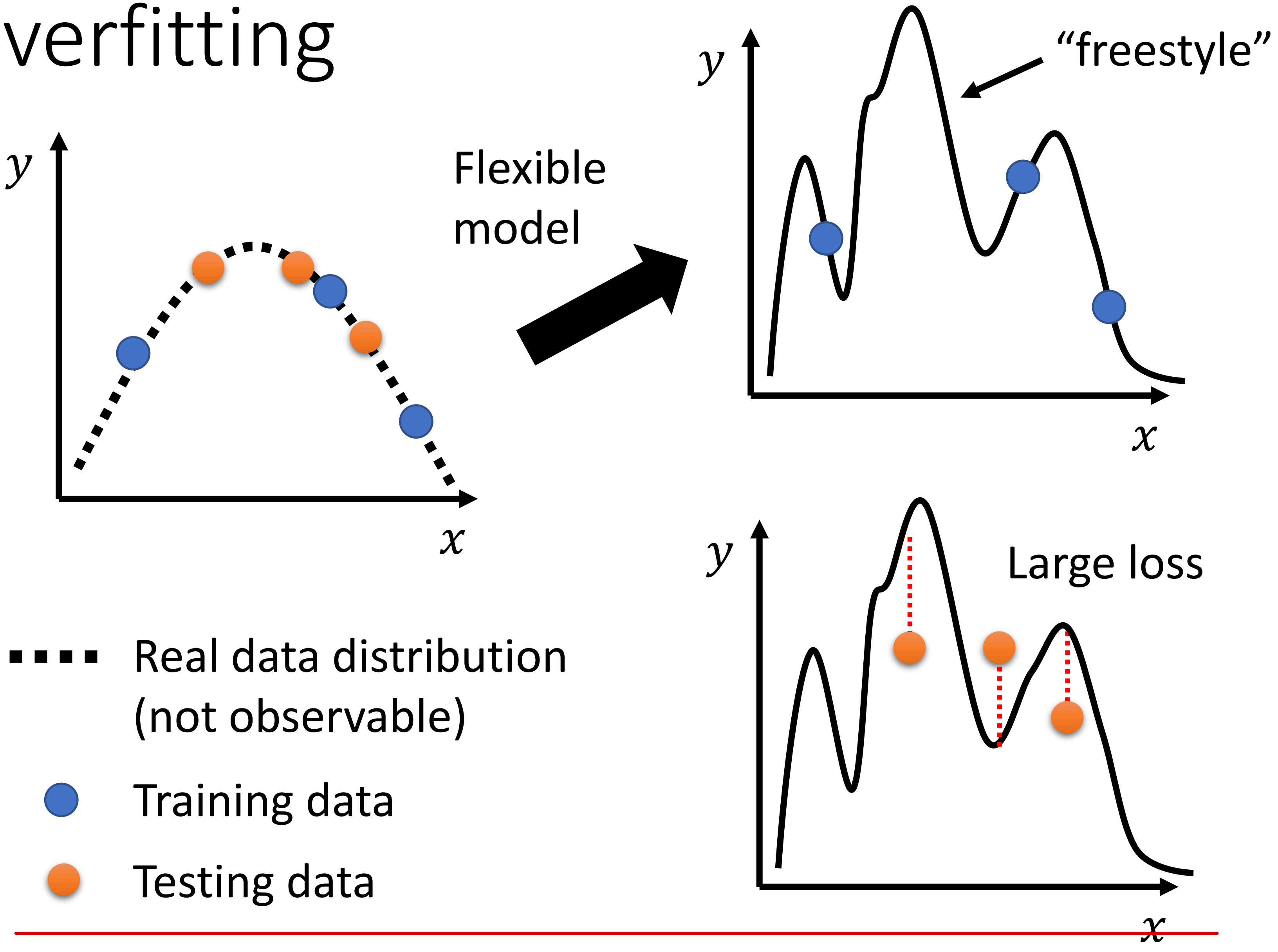
Training data: $\left\{\left(\boldsymbol{x}^{\mathbf{1}}, \hat{y}^{1}\right),\left(\boldsymbol{x}^{\mathbf{2}}, \hat{y}^{2}\right), \ldots,\left(\boldsymbol{x}^{\boldsymbol{N}}, \hat{y}^{N}\right)\right\}$

$$f(\boldsymbol{x})=\begin{cases}\hat{y}^{i} & \underline{\exists \boldsymbol{x}^{\boldsymbol{i}}=\boldsymbol{x}} \\ random & otherwise\end{cases}$$

Less than useless …

This function obtains **zero training loss**, but **large testing loss**.

Overfitting



Flexible model

"freestyle"

Large loss

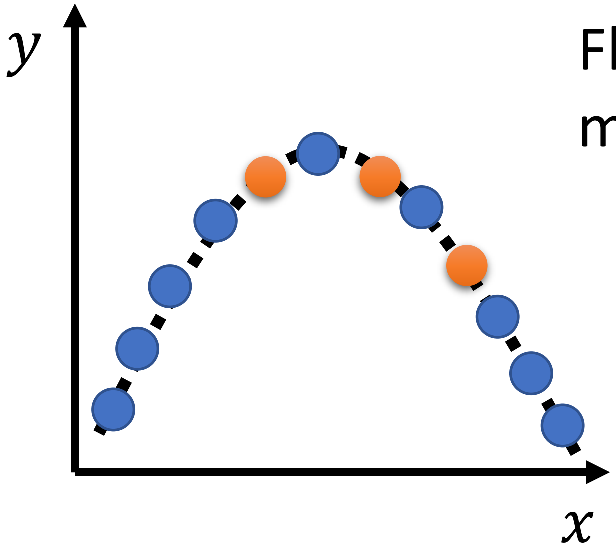
▪▪▪ Real data distribution (not observable)

● Training data

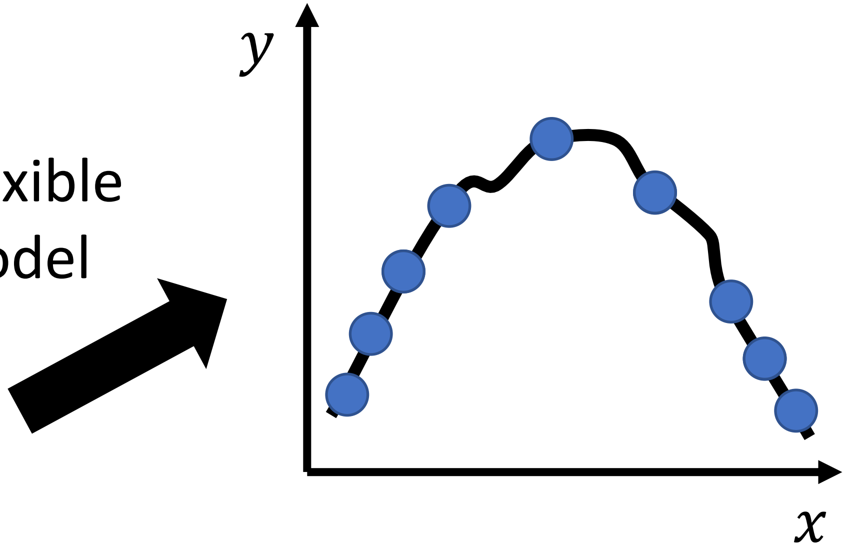
● Testing data

增加feature.

Overfitting



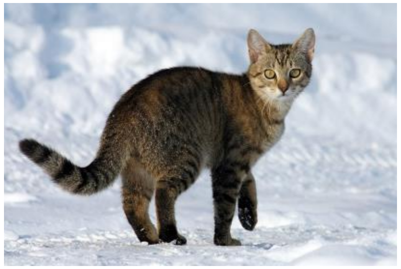
Flexible model

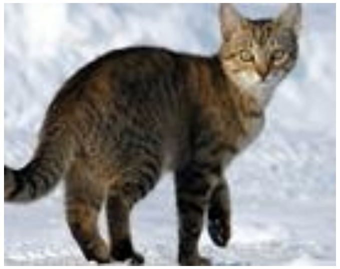


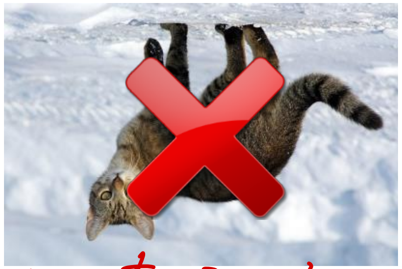
More training data

(cannot do it in HWs)

Data augmentation (you can do that in HWs)

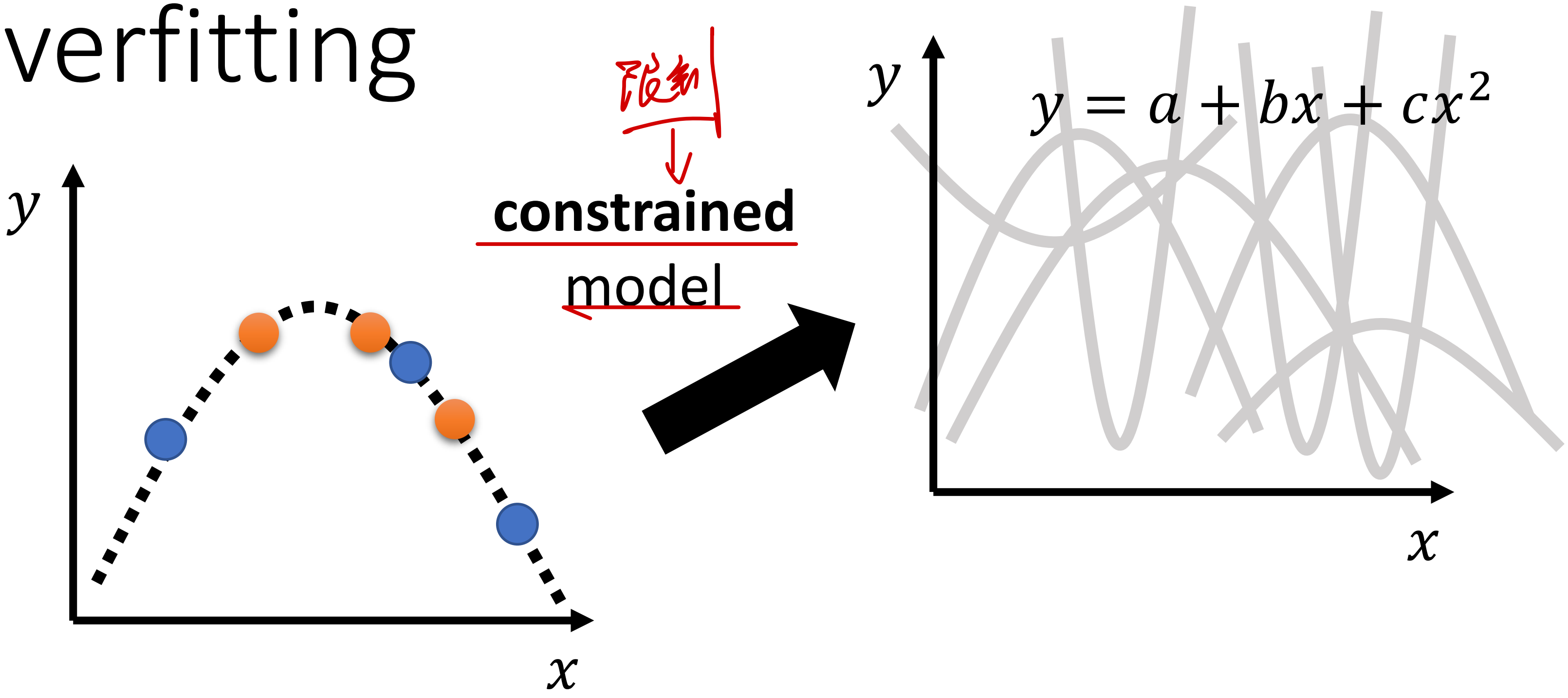






奇怪的东西.

Overfitting

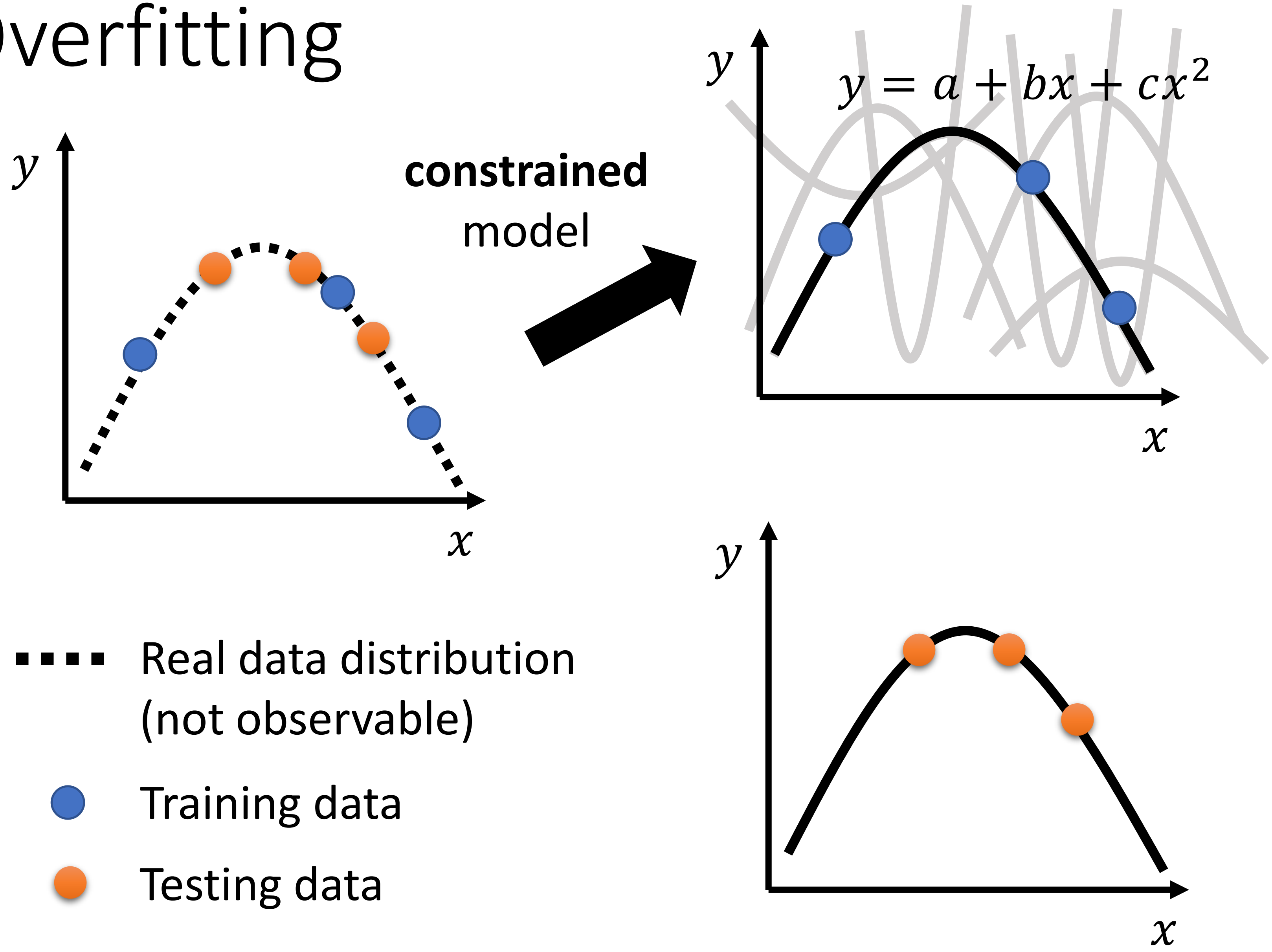


■■■■ Real data distribution
(not observable)

● Training data

● Testing data

Overfitting

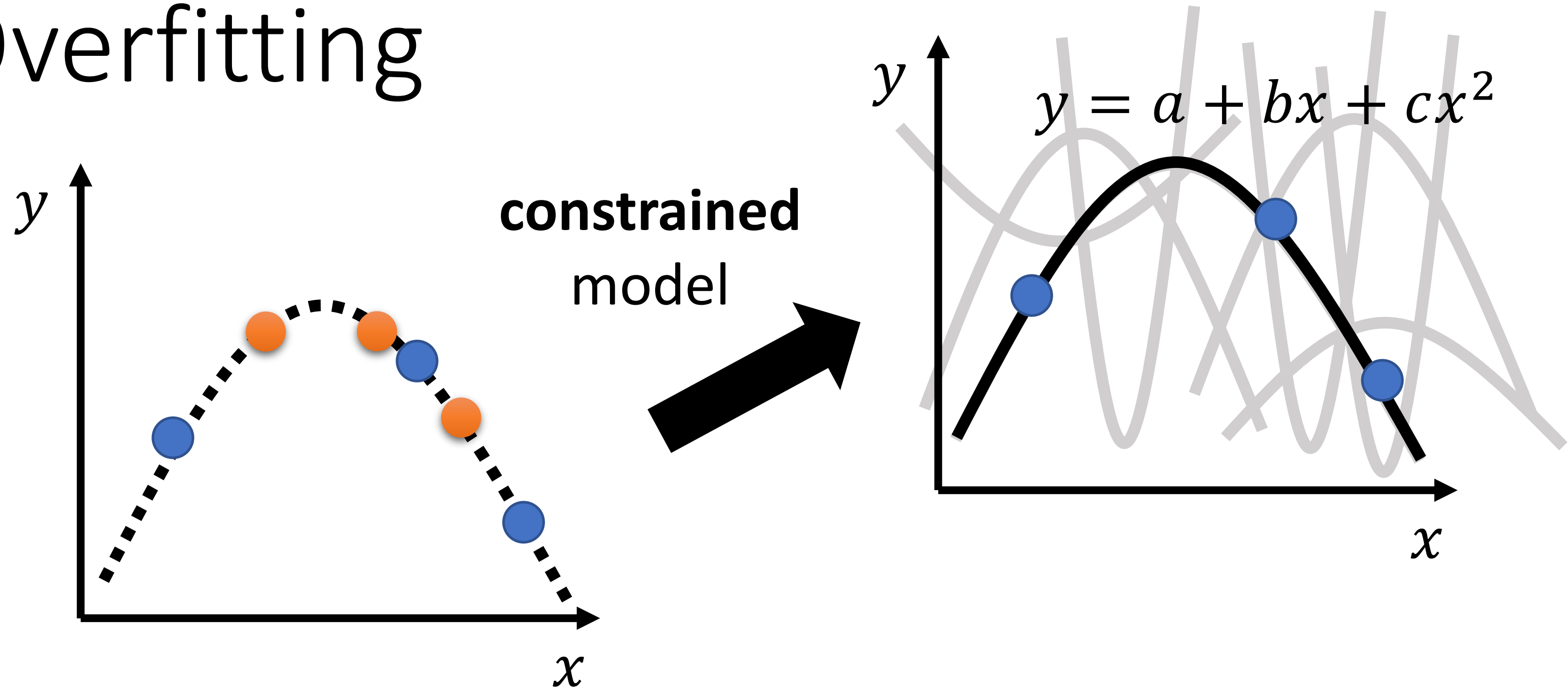


▪▪▪▪ Real data distribution (not observable)

● Training data

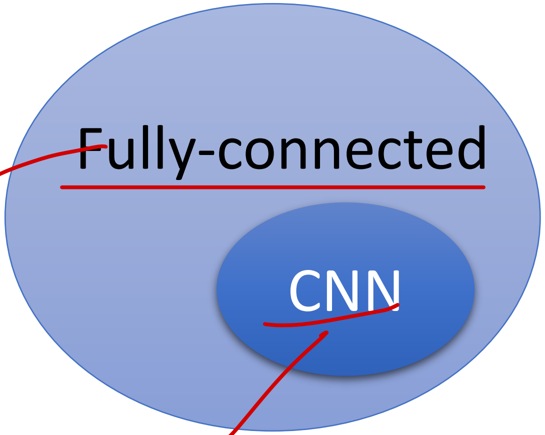
● Testing data

Overfitting



constrained model

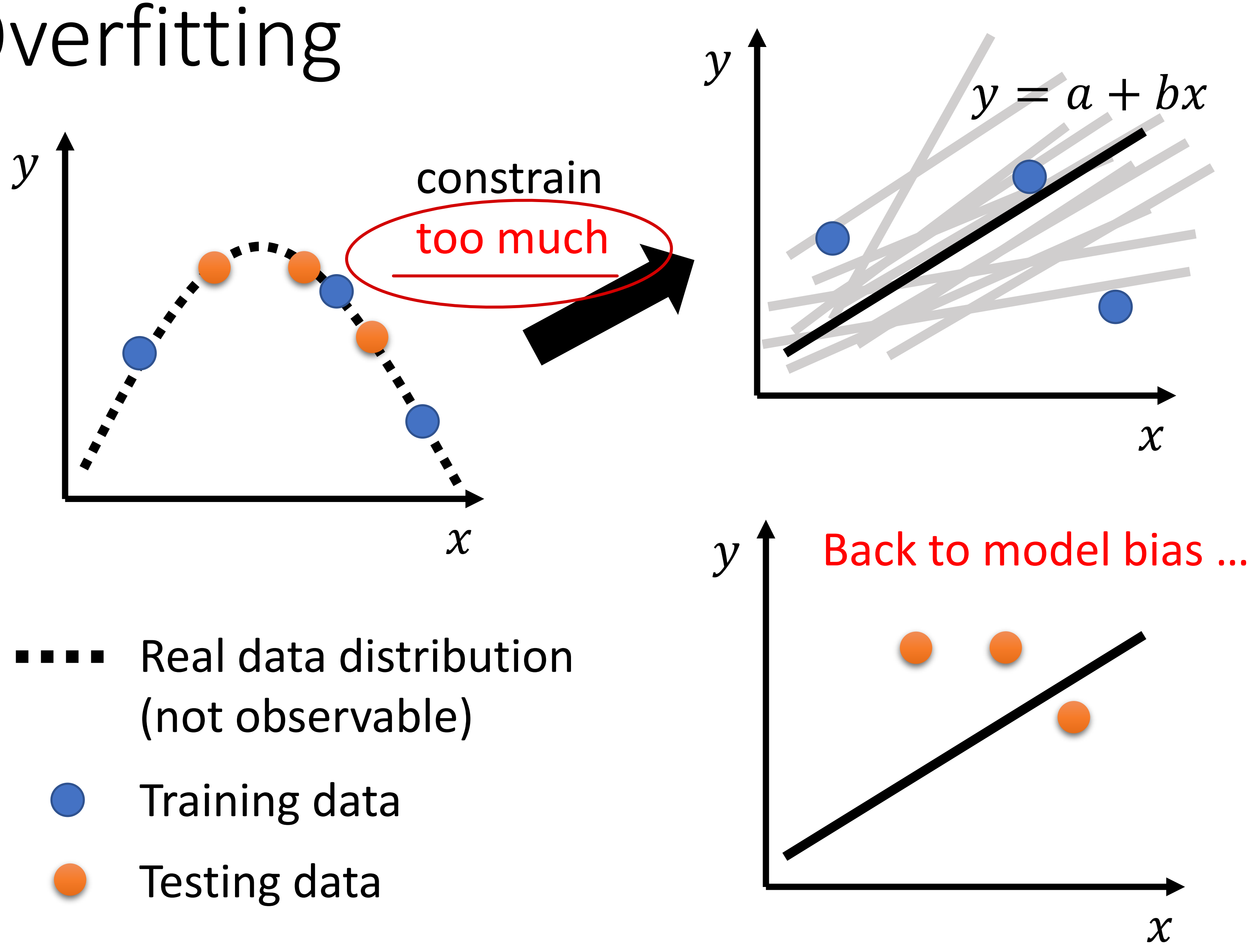
- Less parameters, sharing parameters
- ~~Less features~~
- Early stopping
- Regularization
- Dropout



弹性

比较没有弹性.

Overfitting



constrain ~~too much~~

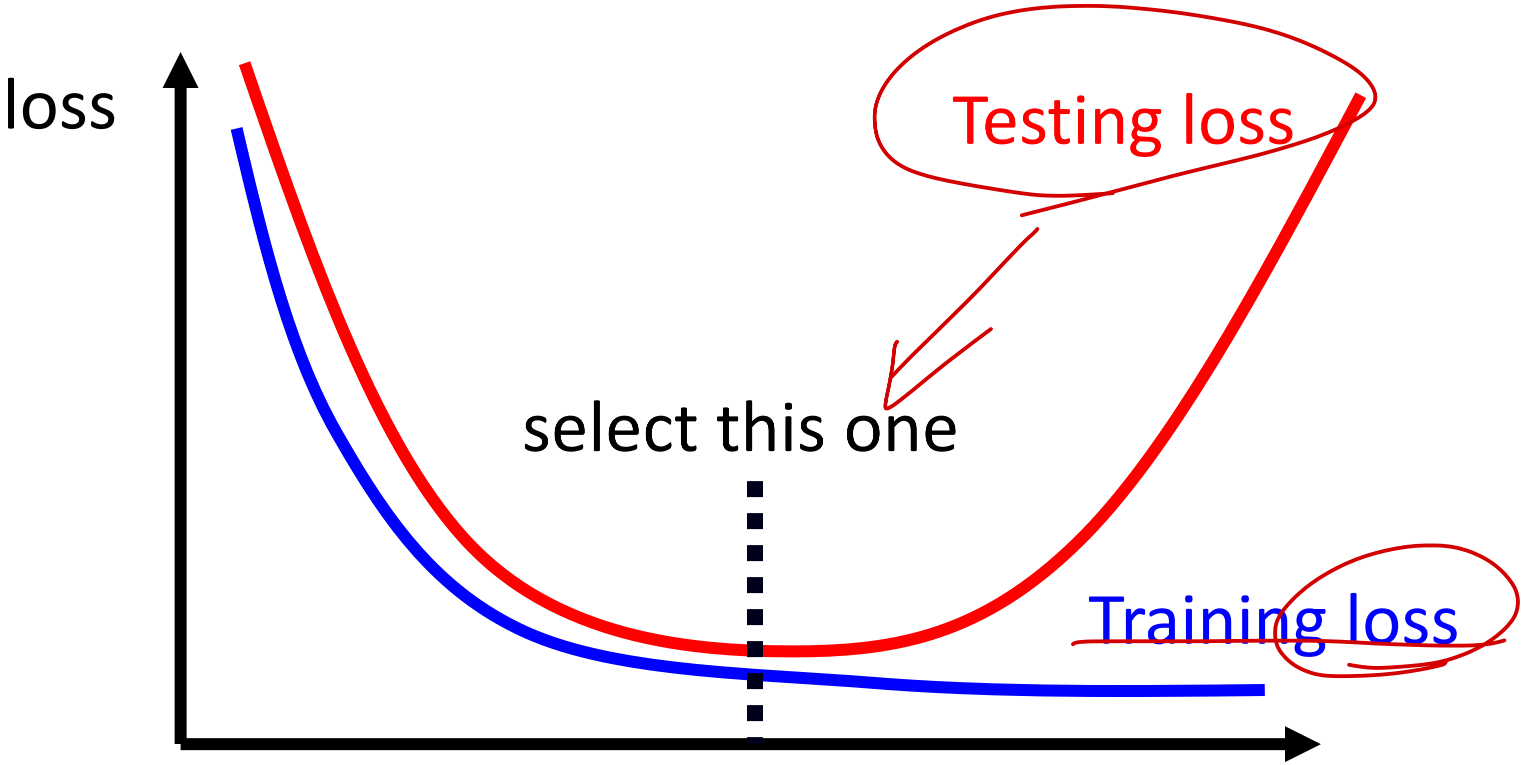
Back to model bias …

■■■■ Real data distribution (not observable)

● Training data

● Testing data

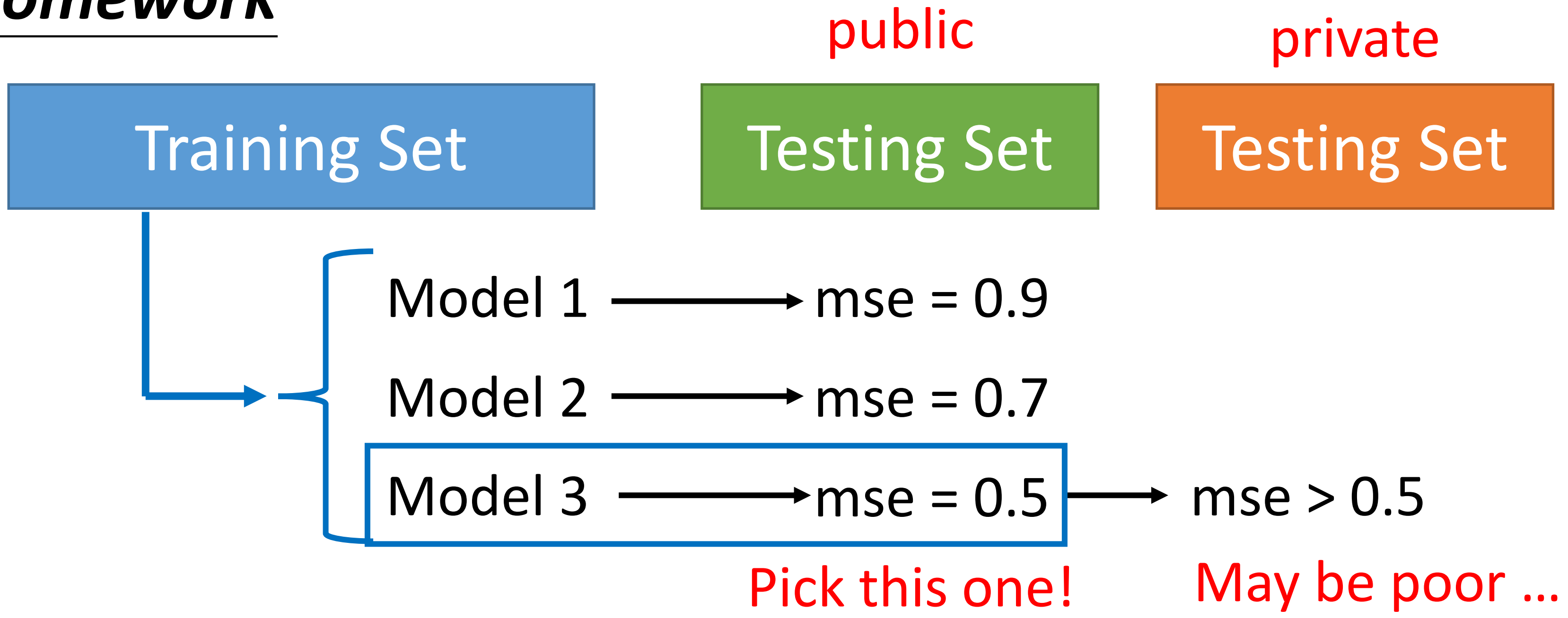
Bias-Complexity Trade-off



Model becomes complex 有多

(e.g. more features, more parameters)

Homework



public

private

Training Set

Testing Set

Testing Set

Model 1 → mse = 0.9

Model 2 → mse = 0.7

Model 3 → mse = 0.5 → mse > 0.5

Pick this one!

May be poor …

The extreme example again

错误中选择没那么错的 X

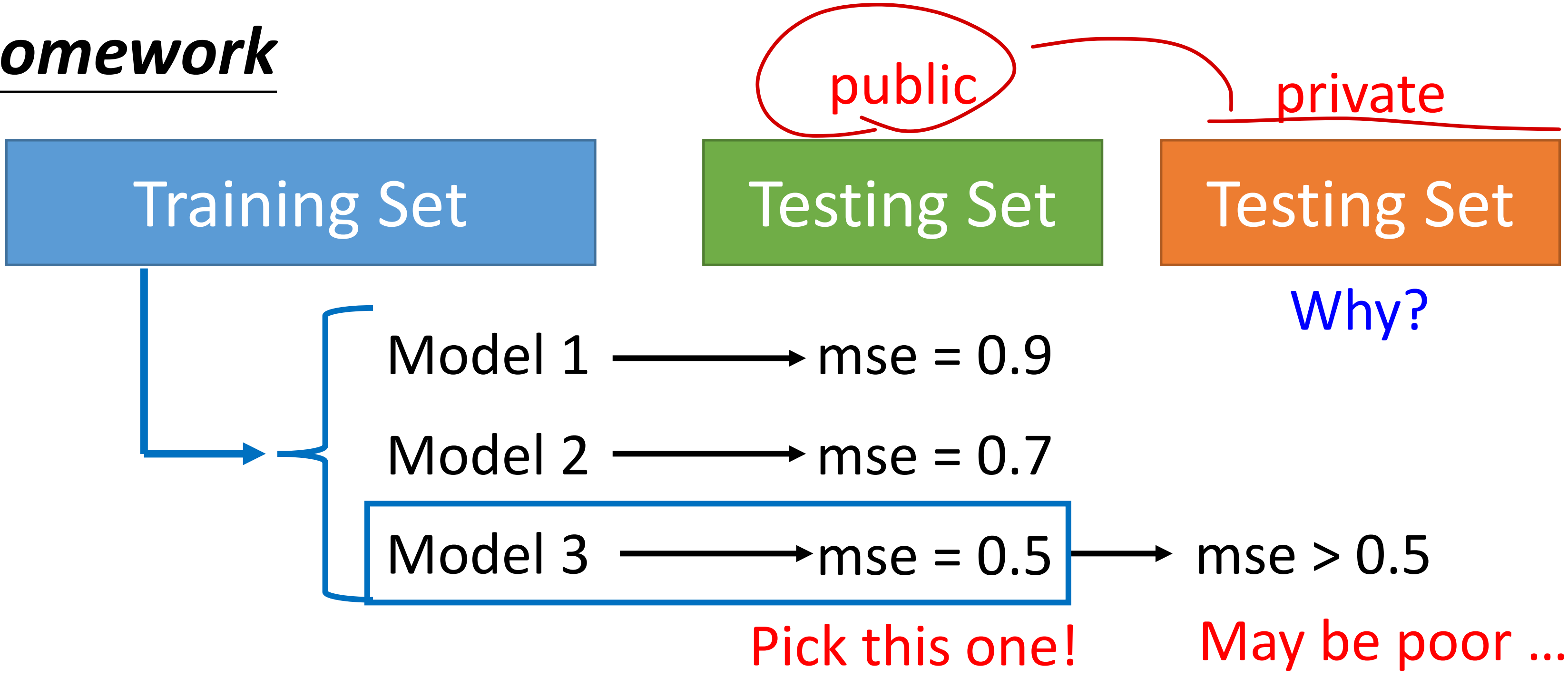
$$f_k(\boldsymbol{x}) = \begin{cases} \hat{y}^i & \exists \boldsymbol{x}^i = \boldsymbol{x} \\ random & otherwise \end{cases}$$

k: 1 - 1000000000000000000000

It is possible that $f_{56789}(\boldsymbol{x})$ **happens** to get good performance on public testing set.

So you select $f_{56789}(\boldsymbol{x})$ …… Random on private testing set

Homework



public

private

Training Set

Testing Set

Testing Set

Why?

Model 1 → mse = 0.9

Model 2 → mse = 0.7

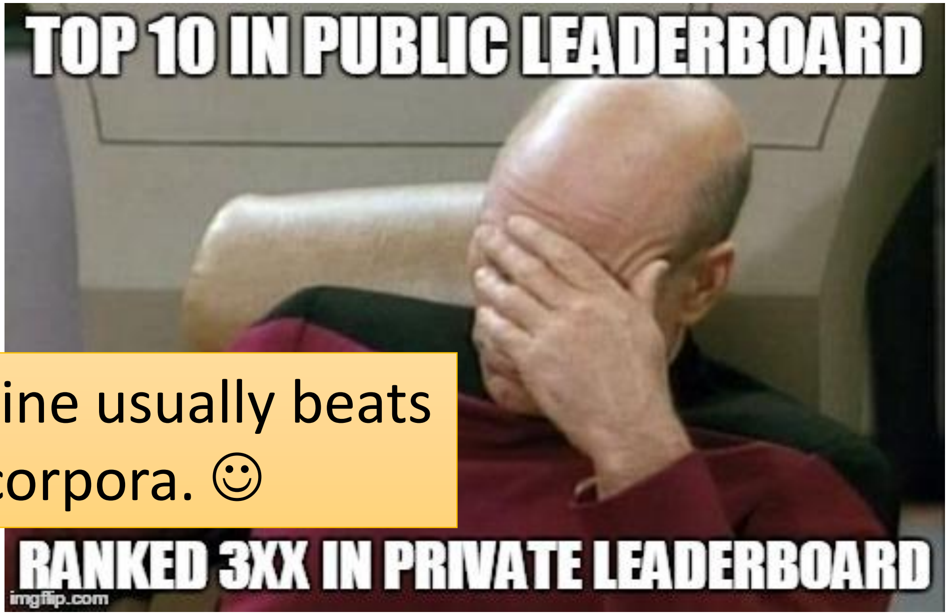
Model 3 → mse = 0.5 → mse > 0.5

Pick this one!

May be poor …

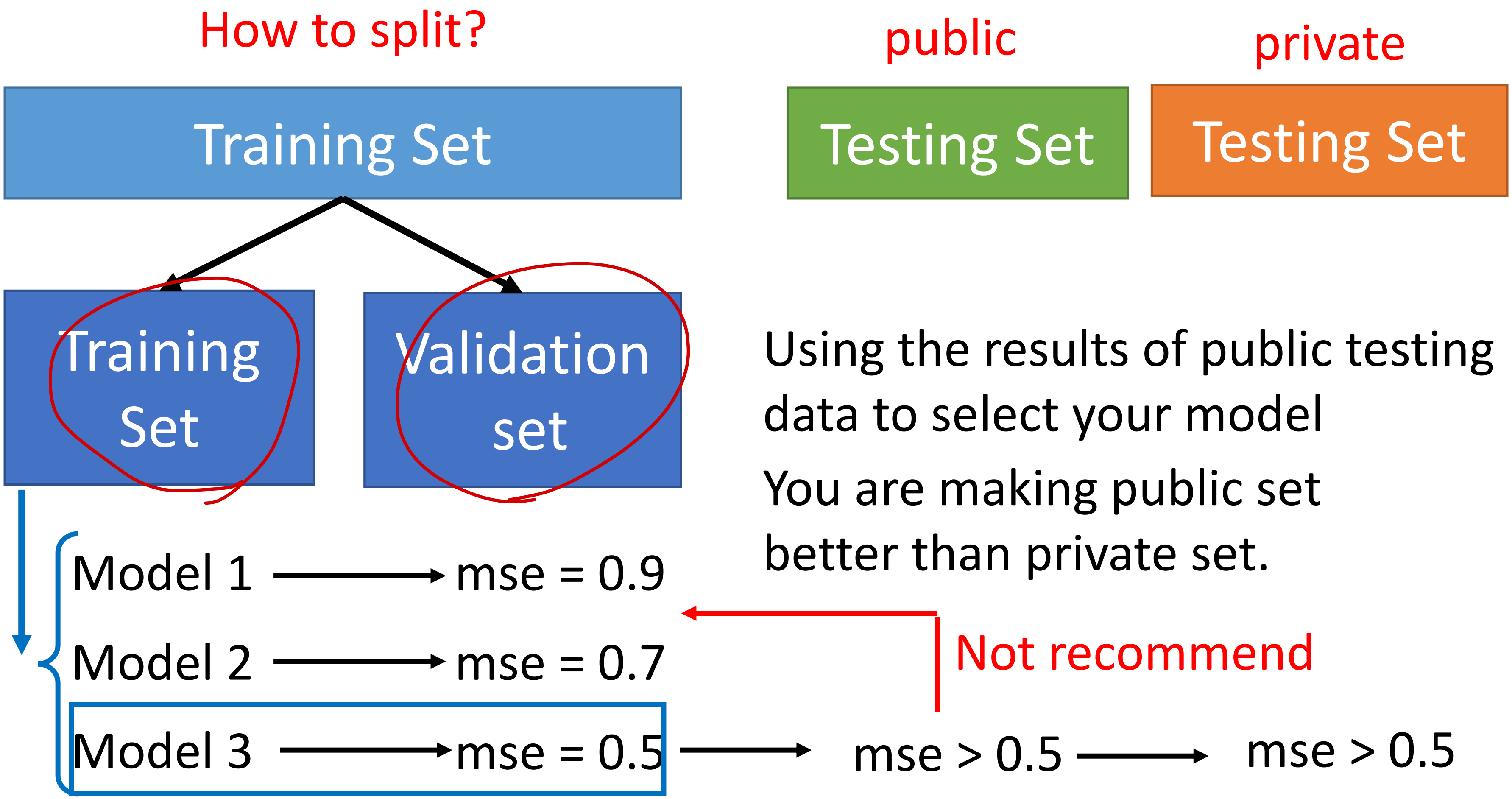
What will happen?

http://www.chioka.in/how-to-select-your-final-models-in-a-kaggle-competitio/


TOP 10 IN PUBLIC LEADERBOARD
RANKED 3XX IN PRIVATE LEADERBOARD
imgflip.com

This explains why machine usually beats human on benchmark corpora. ☺

Cross Validation



How to split?

Training Set

Training Set

Validation set

Model 1 → mse = 0.9

Model 2 → mse = 0.7

Model 3 → mse = 0.5 → mse > 0.5 → mse > 0.5

public

Testing Set

private

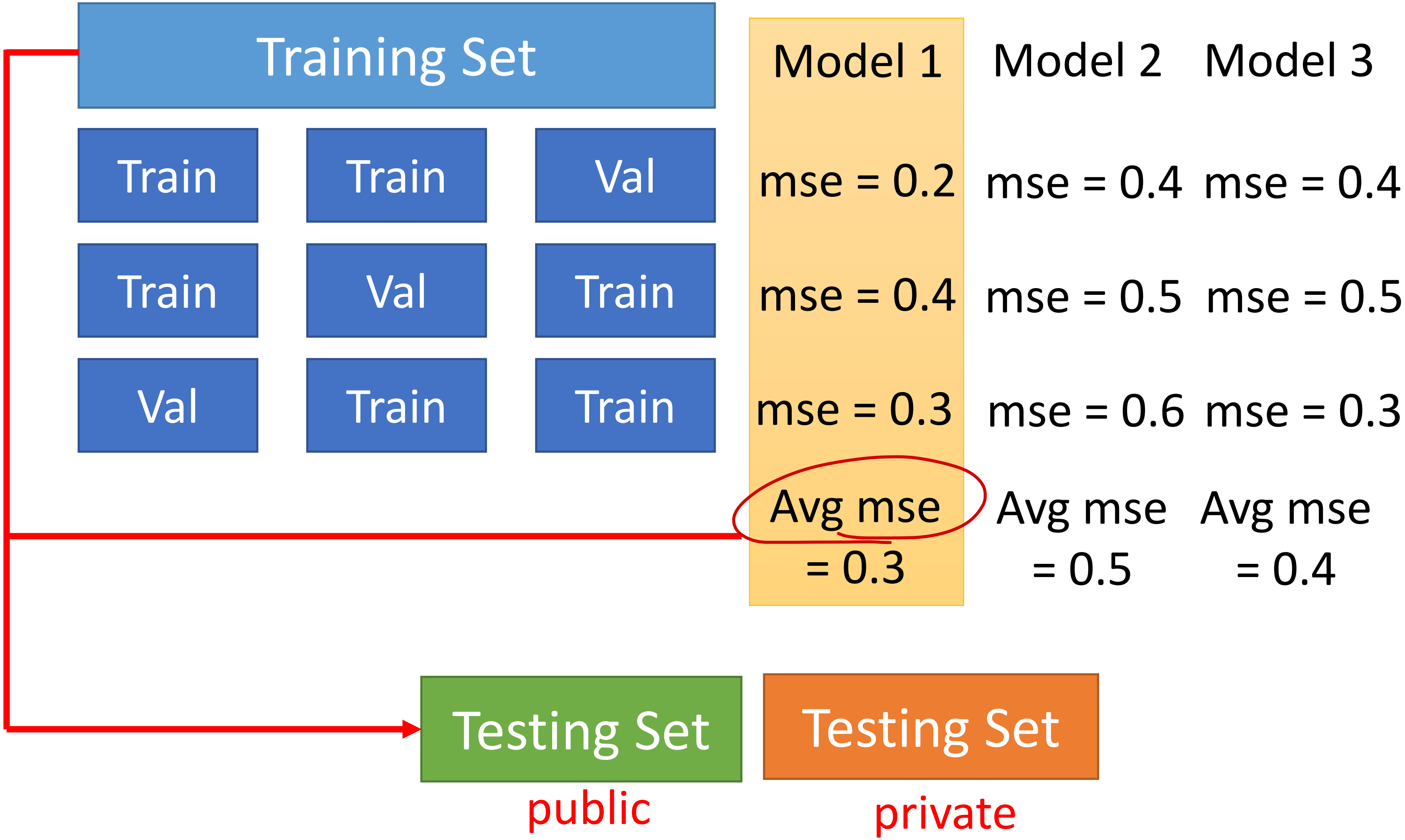
Testing Set

Using the results of public testing data to select your model

You are making public set better than private set.

Not recommend

N-fold Cross Validation



Training Set			Model 1	Model 2	Model 3
Train	Train	Val	mse = 0.2	mse = 0.4	mse = 0.4
Train	Val	Train	mse = 0.4	mse = 0.5	mse = 0.5
Val	Train	Train	mse = 0.3	mse = 0.6	mse = 0.3
			Avg mse = 0.3	Avg mse = 0.5	Avg mse = 0.4

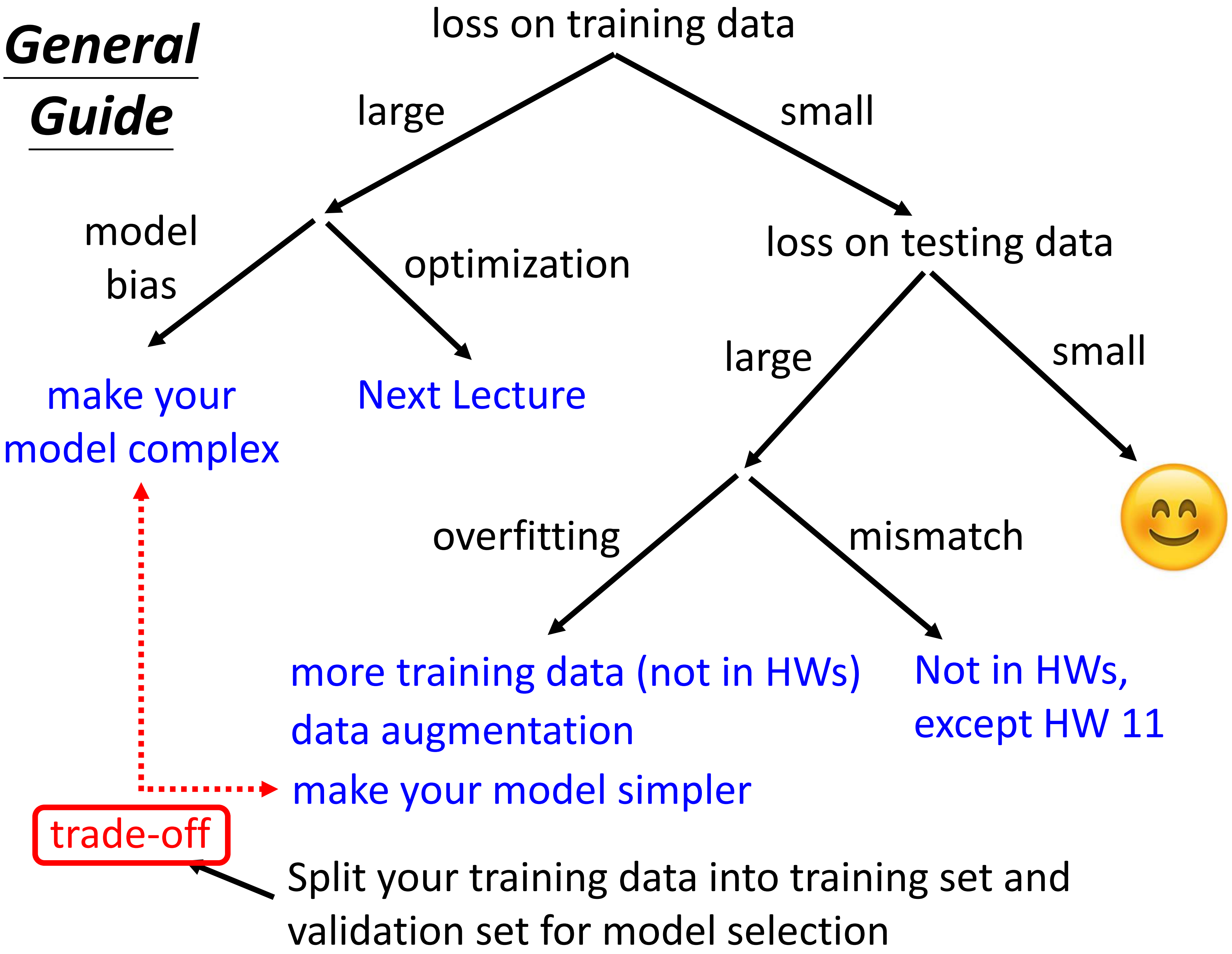
Testing Set

public

Testing Set

private

General Guide



loss on training data

large

model bias

make your model complex

optimization

Next Lecture

small

loss on testing data

large

overfitting

more training data (not in HWs)

data augmentation

make your model simpler

mismatch

Not in HWs, except HW 11

small

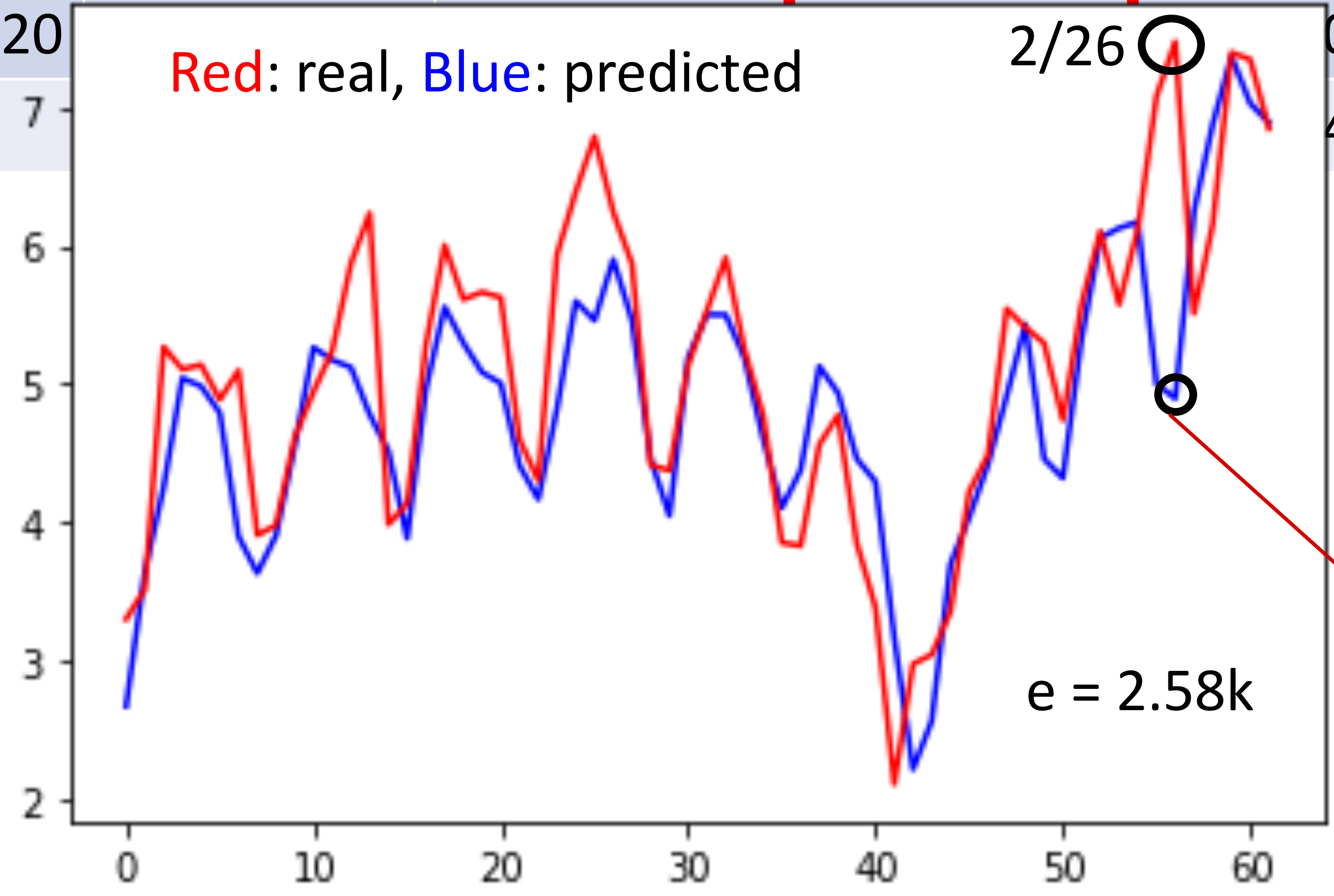
😊

trade-off

Split your training data into training set and validation set for model selection

Let’s predict no. of views of 2/26!

	1 layer	2 layer	3 layer	4 layer
2017 – 2020				0k
2021				4k

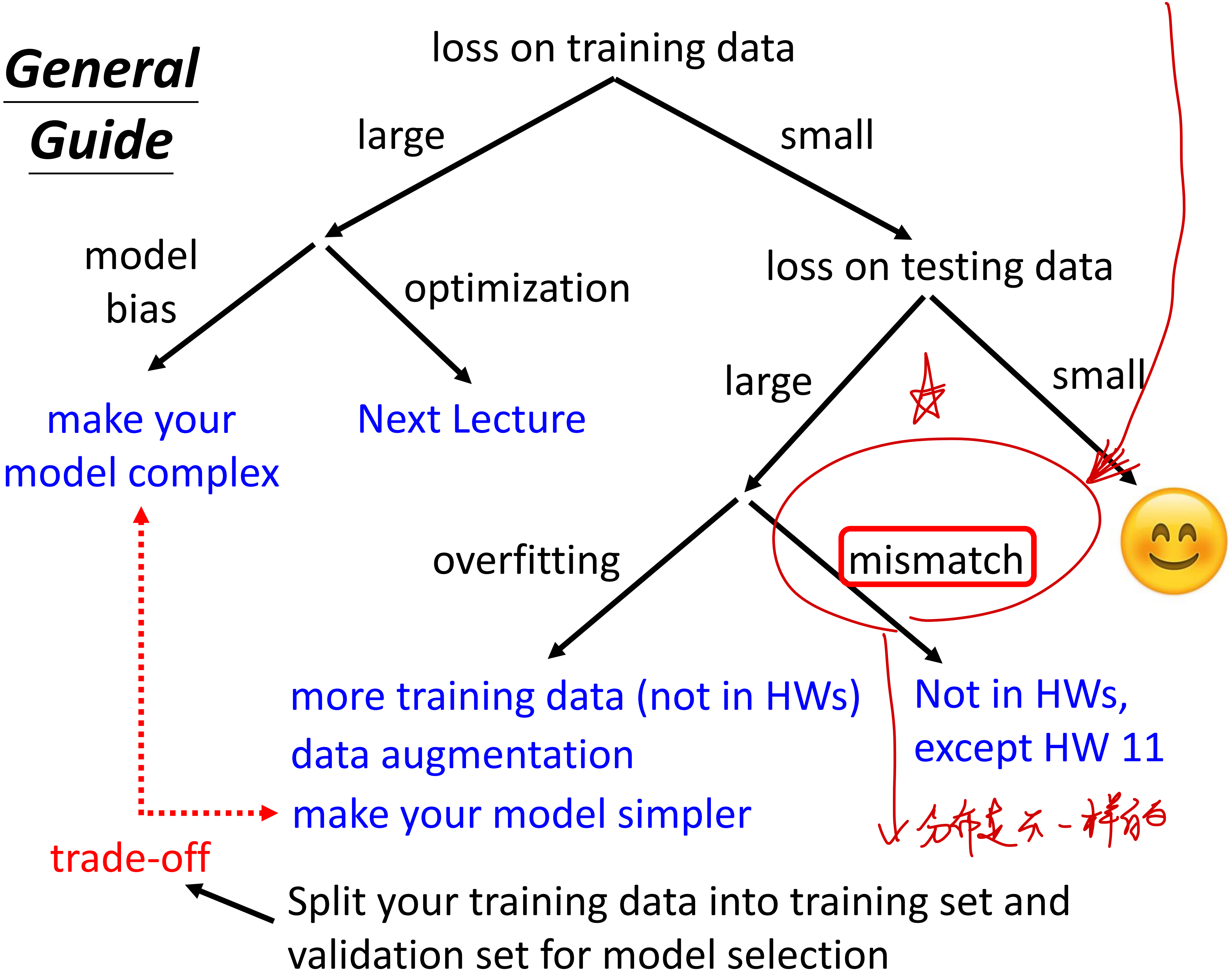


Red: real, Blue: predicted

2/26

e = 2.58k

General Guide



loss on training data

large

small

model bias

make your model complex

optimization

Next Lecture

loss on testing data

large

small

overfitting

more training data (not in HWs)

data augmentation

make your model simpler

trade-off

mismatch

Not in HWs, except HW 11

分布是不一样的

😊

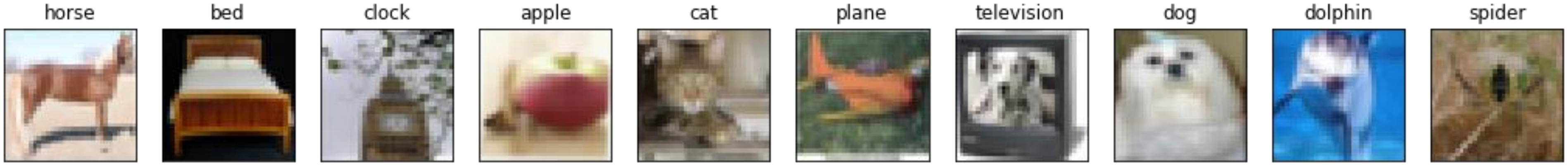
Split your training data into training set and validation set for model selection

Mismatch

- Your training and testing data have different distributions. Be aware of how data is generated.

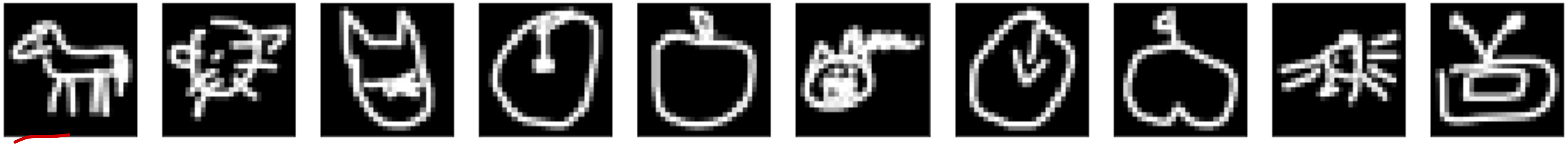
Most HWs do not have this problem, except HW11

Training Data

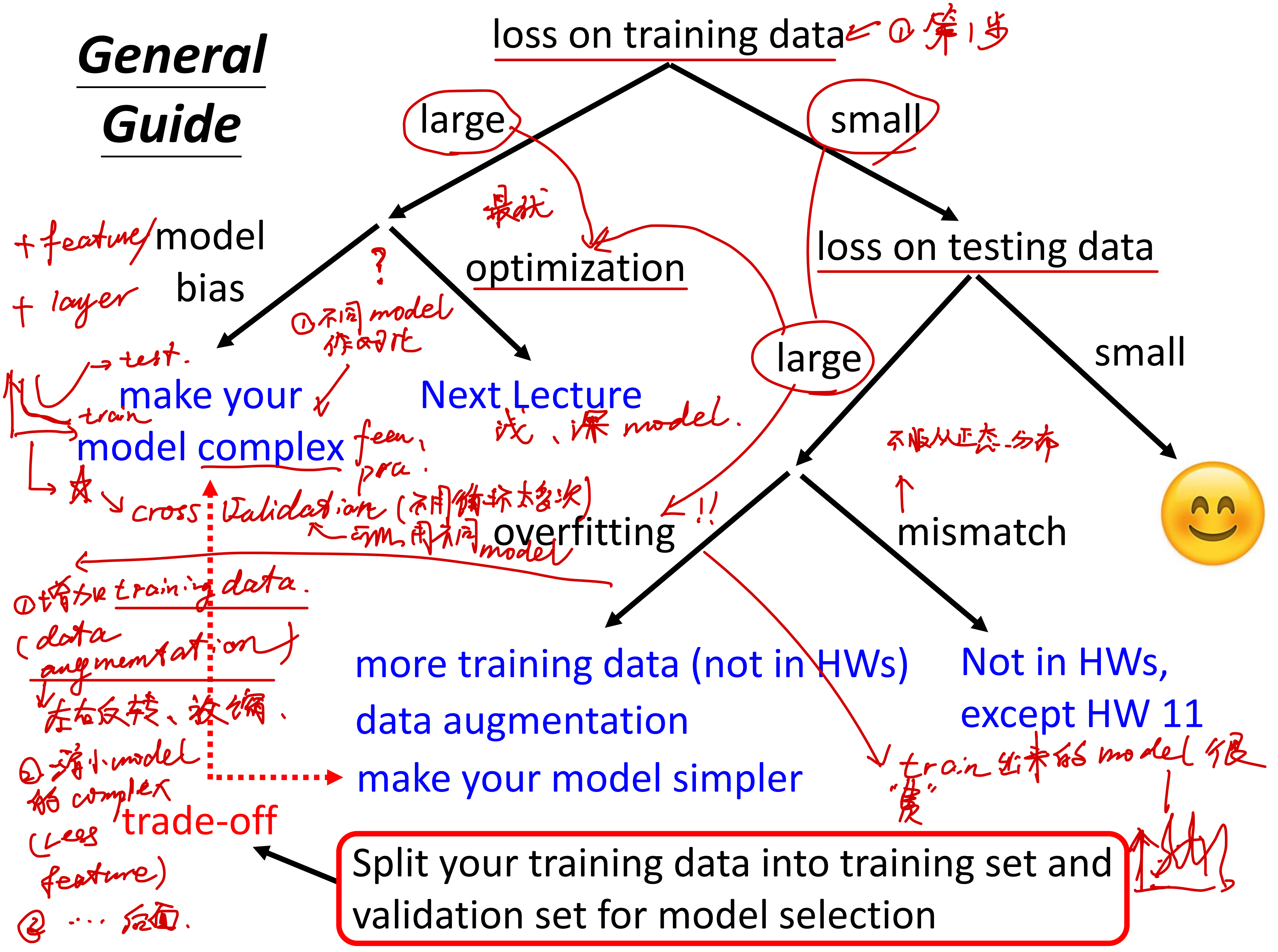


Simply increasing the training data will not help.

Testing Data



General Guide



loss on training data ← ①第1步

large

small

model bias

+feature/

+ layer

?

①不同model 作对比

最优

optimization

make your model complex

→ test.

train

feen, pra.

Next Lecture

浅、深 model.

★ ↘ cross Validation (不用循环太多次)

← 训练，用不同model

loss on testing data

large

small

overfitting !!

mismatch

不服从正态分布

①增加training data.

(data augmentation)

↓ 左右反转、放缩.

②减小model的complex (less feature)

③… 后面.

more training data (not in HWs)

data augmentation

make your model simpler

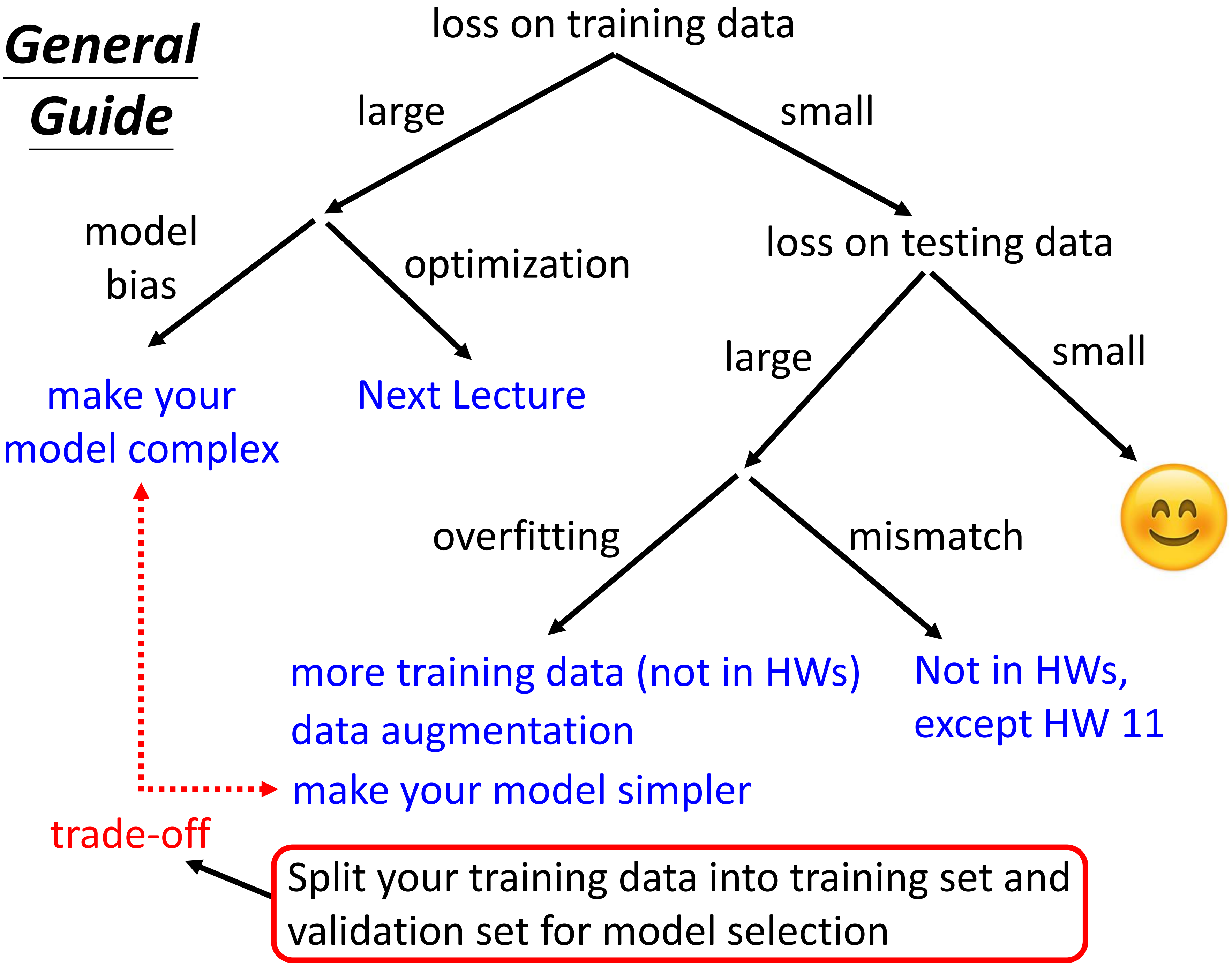
trade-off

train出来的model很"贵"

Not in HWs, except HW 11

Split your training data into training set and validation set for model selection

General Guide



loss on training data

large

model bias

make your model complex

optimization

Next Lecture

small

loss on testing data

large

overfitting

more training data (not in HWs)

data augmentation

make your model simpler

mismatch

Not in HWs, except HW 11

small

😊

trade-off

Split your training data into training set and validation set for model selection