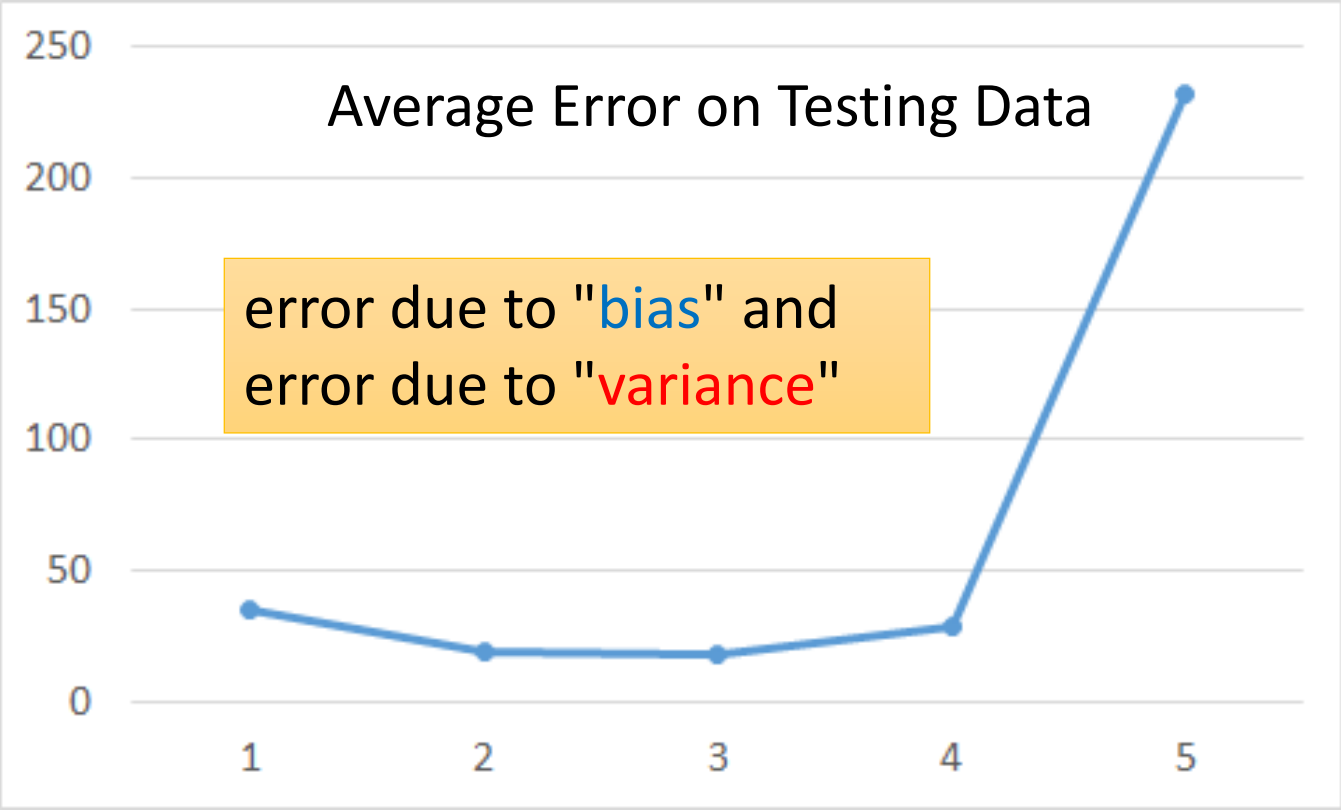
Where does the error come from?

In order to find the appropriate model to improve the performance, we need to diagnose the source of the error.

From bias or variance?

Review



A more complex model does not always lead to better performance on ***testing data***.

Estimator

$$\hat{y} = \hat{f}(\quad)$$

The real function

Only Niantic knows $\hat{f}$

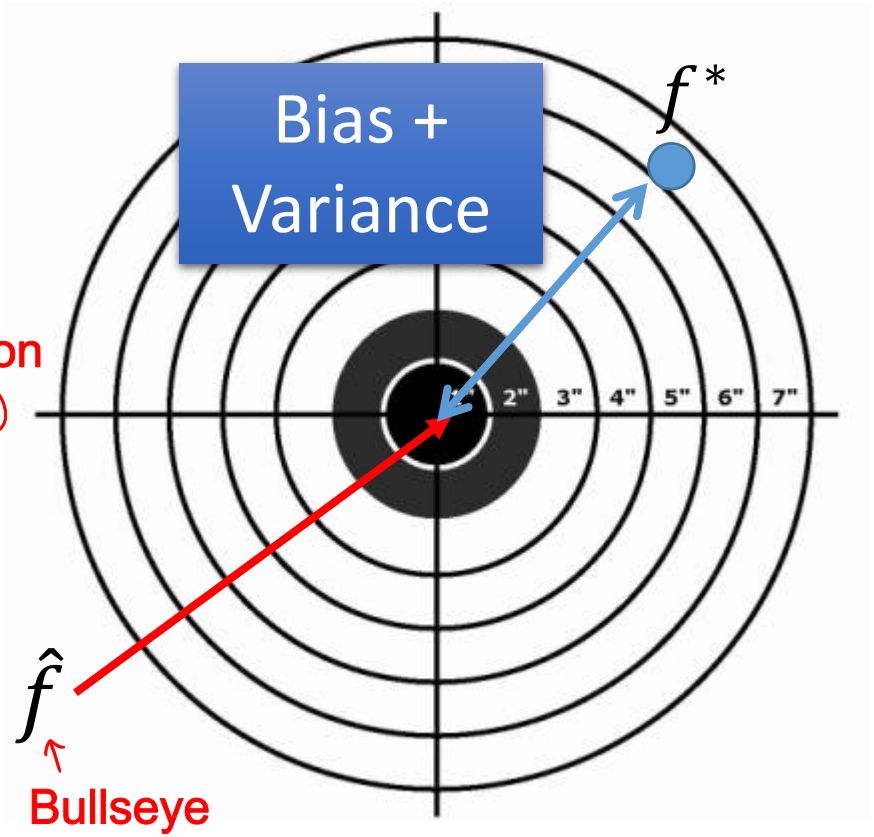
The company of Pokémon.

From training data,

we find f^*

Our model (estimator)

f^* is an estimator of $\hat{f}$



Bias and Variance of Estimator

of mean

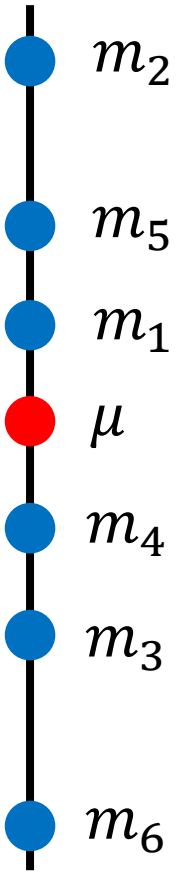
We are aiming μ.
(f* is unbiased.)
unbiased

- Estimate the mean of a variable x
  - assume the mean of x is μ → $\hat{f}_{mean} = \mu$
  - assume the variance of x is σ^2
- Estimator of mean μ $f^*_{mean} = m$
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

Mean of sample (Estimator)

$$m = \frac{1}{N}\sum_n x^n \neq \mu$$

$$E[m] = E\left[\frac{1}{N}\sum_n x^n\right] = \frac{1}{N}\sum_n E[x^n] = \mu$$



Use "E" to evaluate "bias" of estimator of mean.

Bias and Variance of Estimator

of mean

∴ Smaller variance
(More concentrated)

unbiased

- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of mean μ
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

$$m = \frac{1}{N}\sum_n x^n \neq \mu$$

"Bias" of $f^*_{mean} = 0$

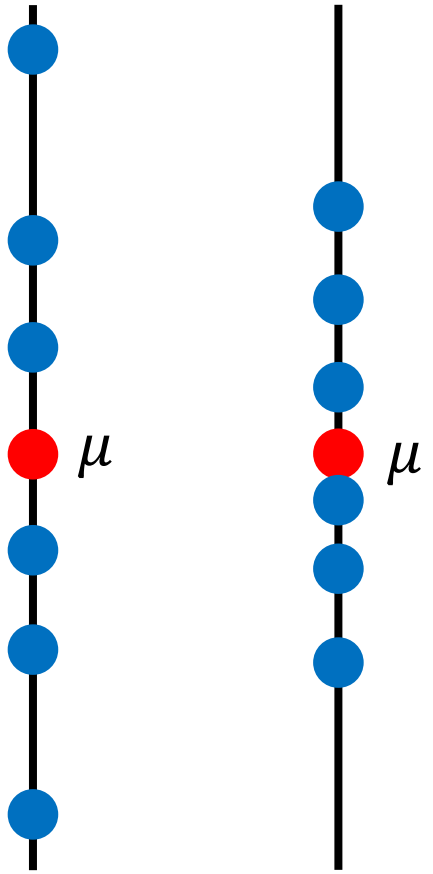
"Variance" of $f^*_{mean} \propto \frac{1}{N}$

$$Var[m] = \frac{\sigma^2}{N}$$

Variance depends on the number of samples

Use "Var" to evaluate "variance" of estimator of mean.

Smaller N ∵ Larger N



of variance

Bias and Variance of Estimator

- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of variance σ^2 $f^*_{variance} = s$
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

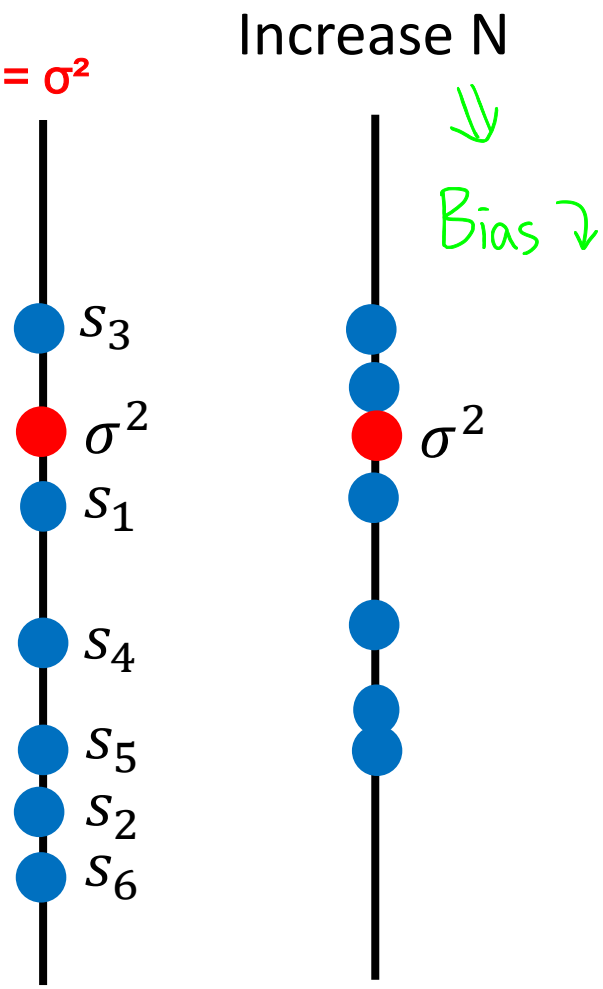
$\hat{f}_{variance} = \sigma^2$

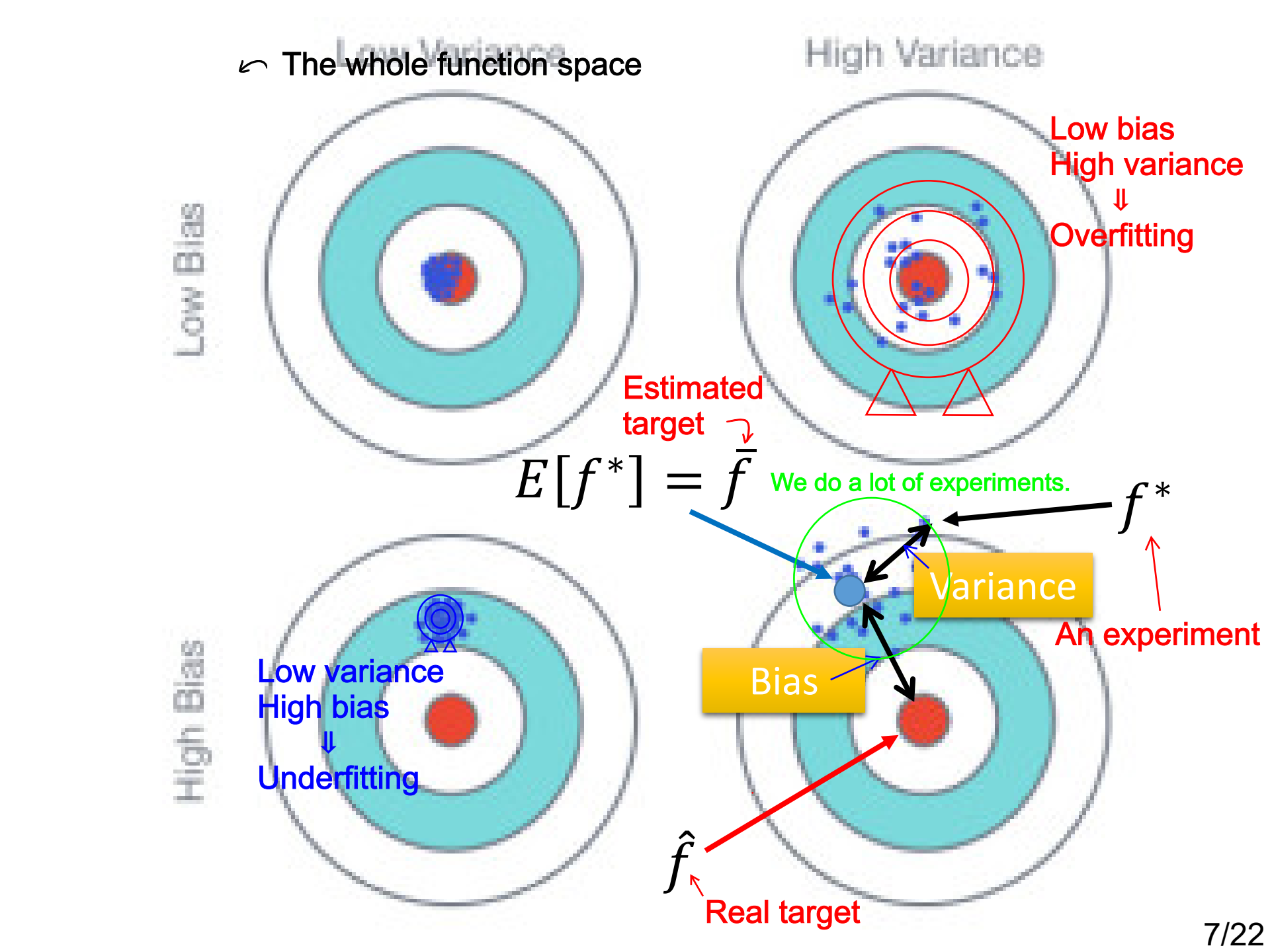
$$m = \frac{1}{N}\sum_n x^n \qquad s = \frac{1}{N}\sum_n (x^n - m)^2$$

Biased estimator

"Bias" of $f^*_{variance}$ = $\frac{\sigma^2}{N}$

$$E[s] = \frac{N-1}{N}\sigma^2 \neq \sigma^2$$

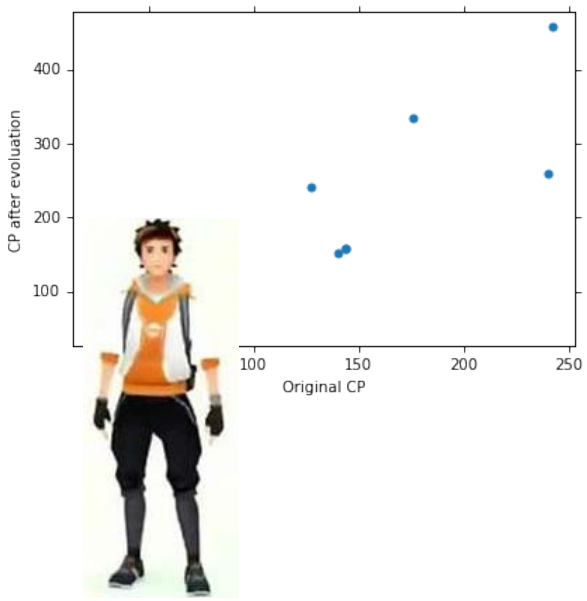



The whole function space
Low Variance
High Variance
Low Bias
High Bias
Low bias
High variance
⇓
Overfitting
Estimated target
$E[f^*] = \bar{f}$
We do a lot of experiments.
f^*
An experiment
Variance
Bias
Low variance
High bias
⇓
Underfitting
$\hat{f}$
Real target

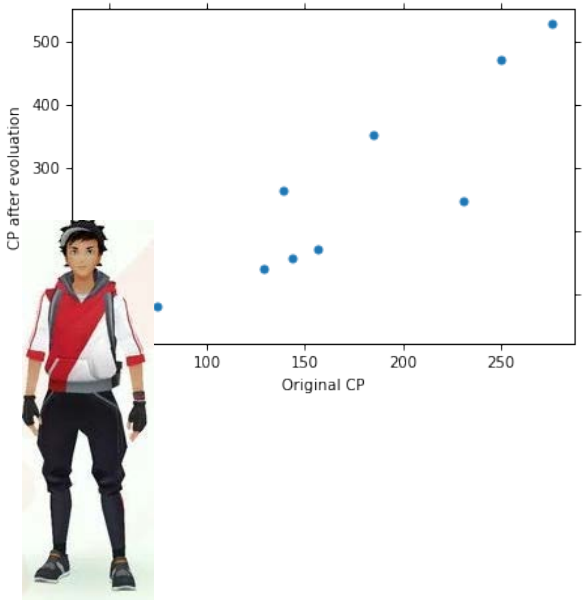
Parallel Universes Different dataset.

- In all the universes, we are collecting (catching) 10 Pokémons as training data to find f^*

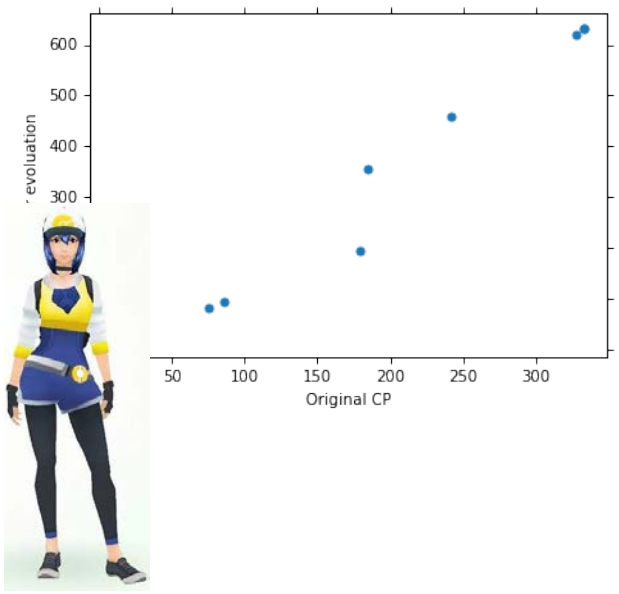
Universe 1



Universe 2



Universe 3



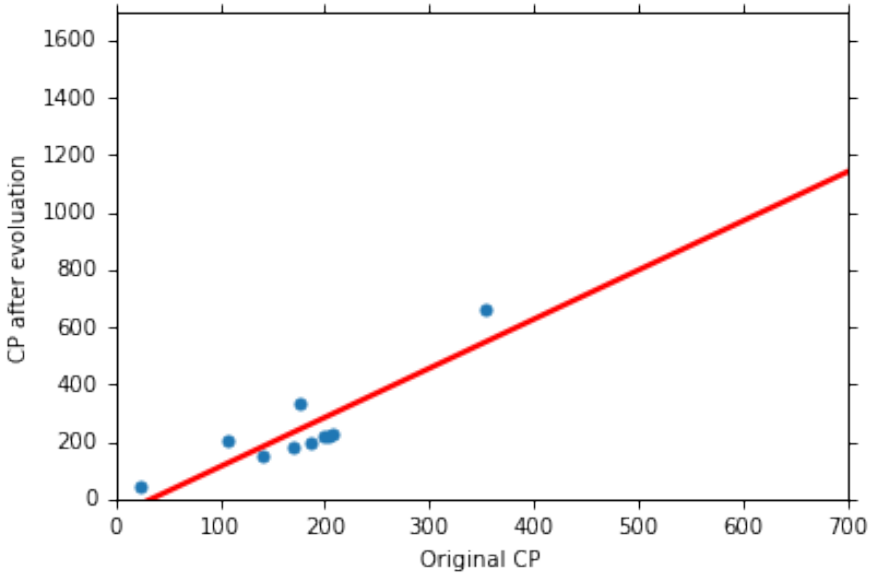
Parallel Universes

- In different universes, we use the same model, but obtain different f^*

Linear regression

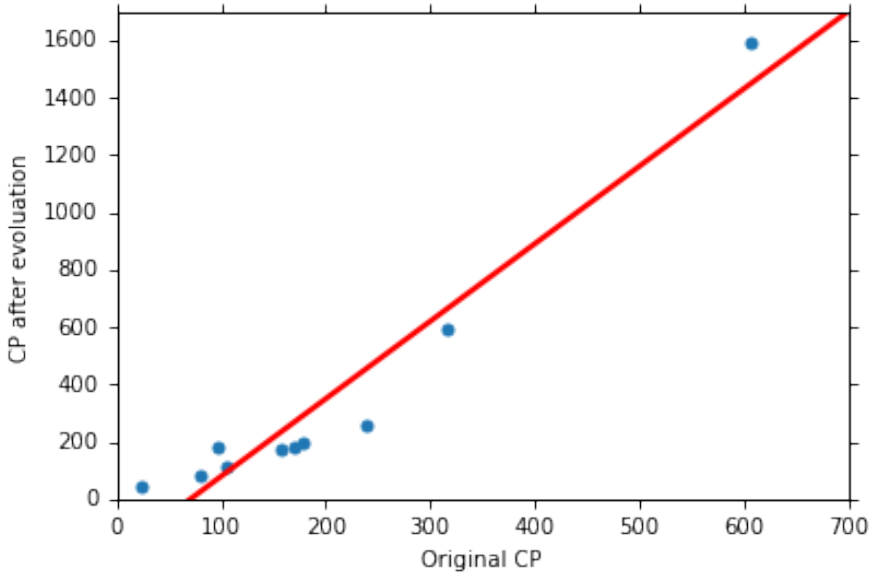
w, b

Universe 123



$y = b + w \cdot x_{cp}$

Universe 345

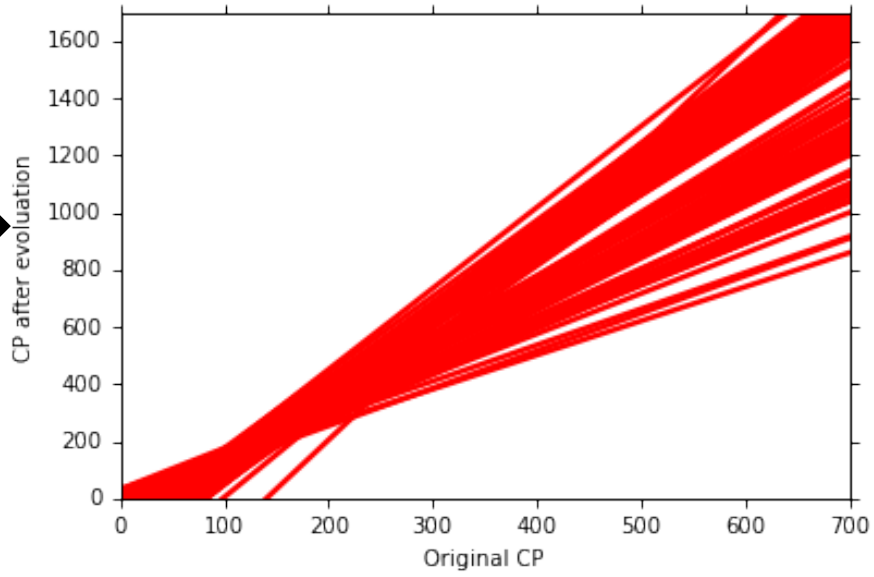


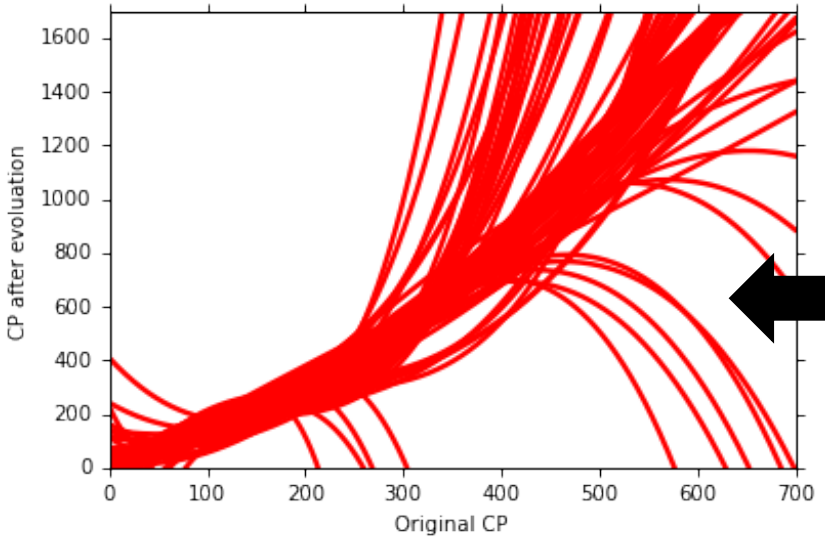
$y = b + w \cdot x_{cp}$

f^* in 100 Universes

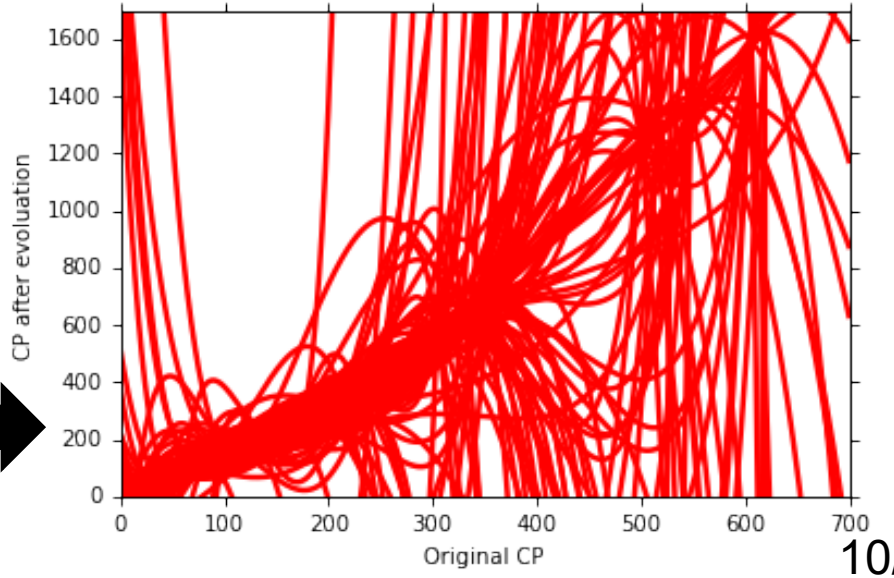
The complexity of the model ↗

$y = b + w \cdot x_{cp}$





$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3$

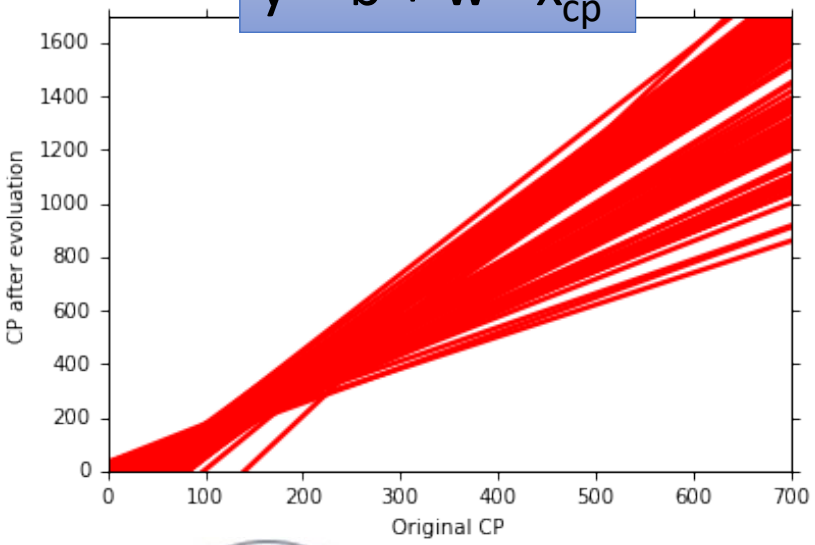


$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$

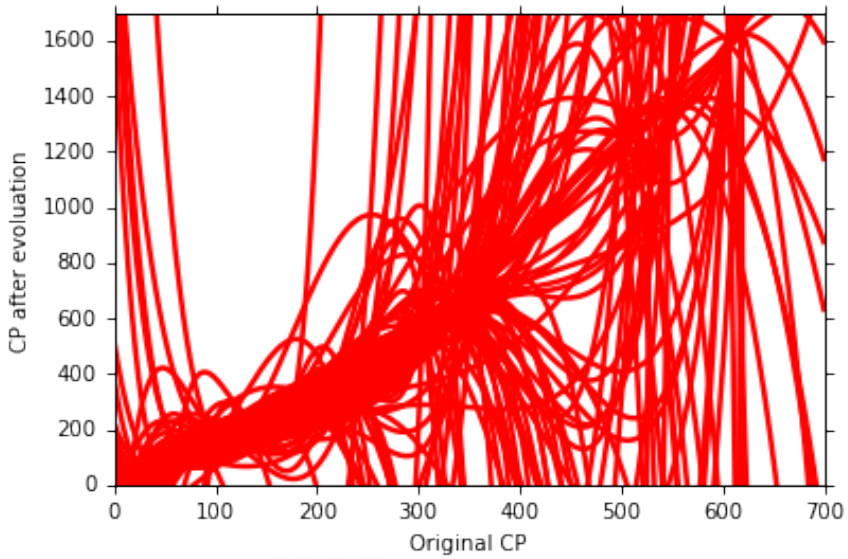
The complexity of the model ↗ ⇒ The variance ↗

Variance

$y = b + w \cdot x_{cp}$



$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$



Small Variance

Large Variance

Simpler model is less influenced by the sampled data

More general

Consider the extreme case f(x) = 5
(Won't be affected at all)

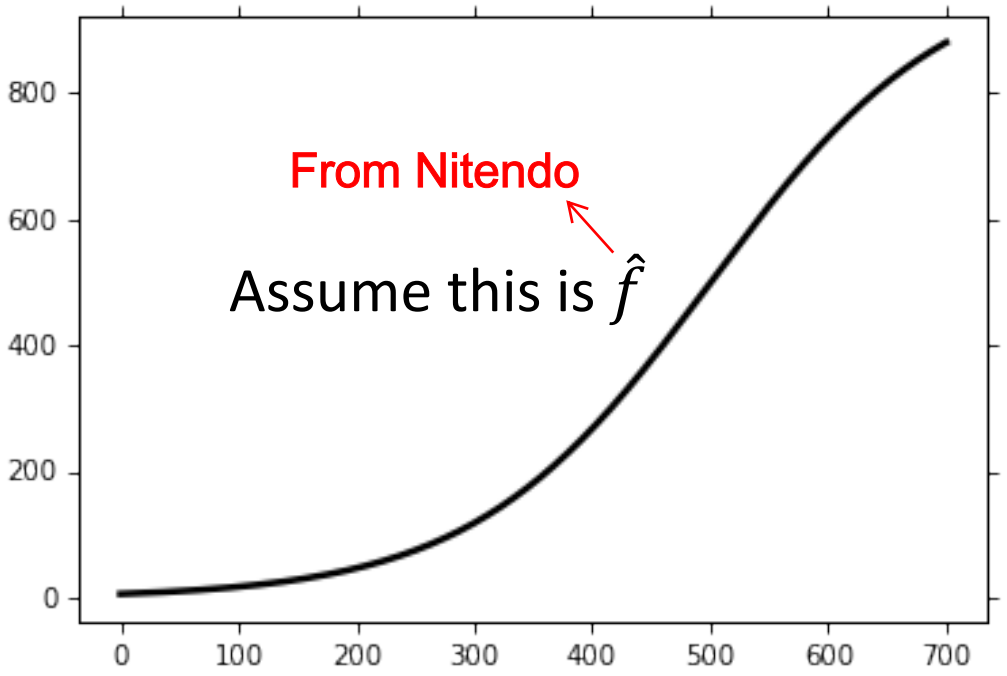
Bias

$$E[f^*] = \bar{f}$$

- Bias: If we average all the f^*, is it close to $\hat{f}$?

Large Bias

Small Bias



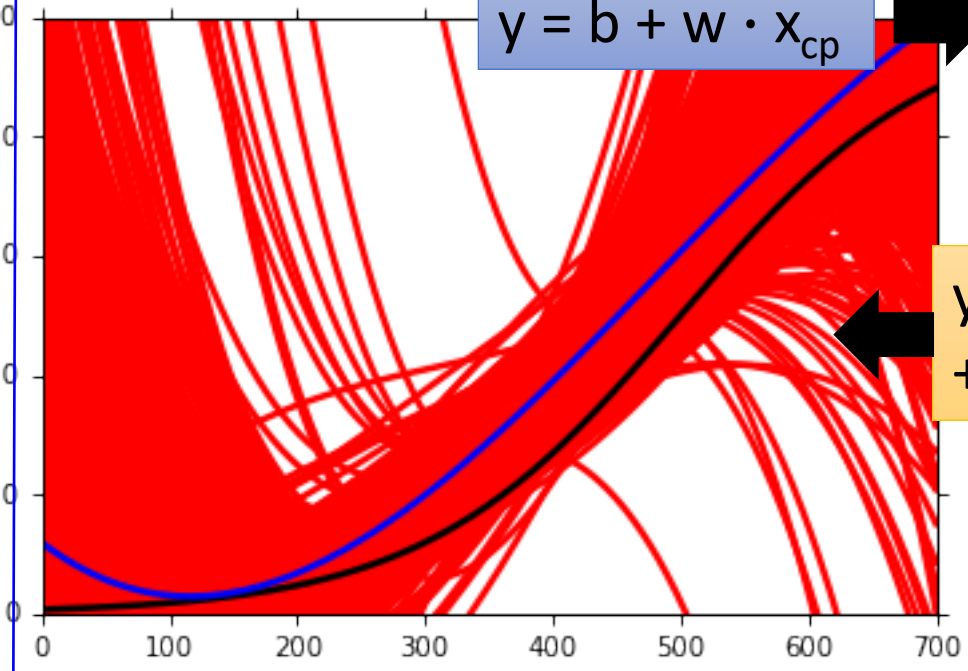
Black curve: the true function $\hat{f}$

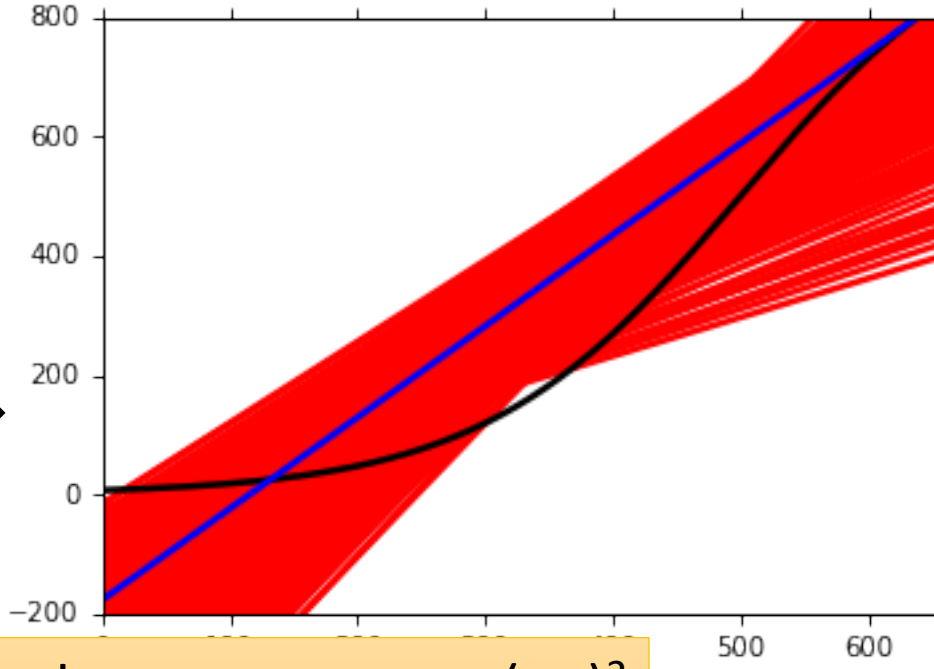
Red curves: 5000 f^*

Blue curve: the average of 5000 f^* $= \bar{f}$

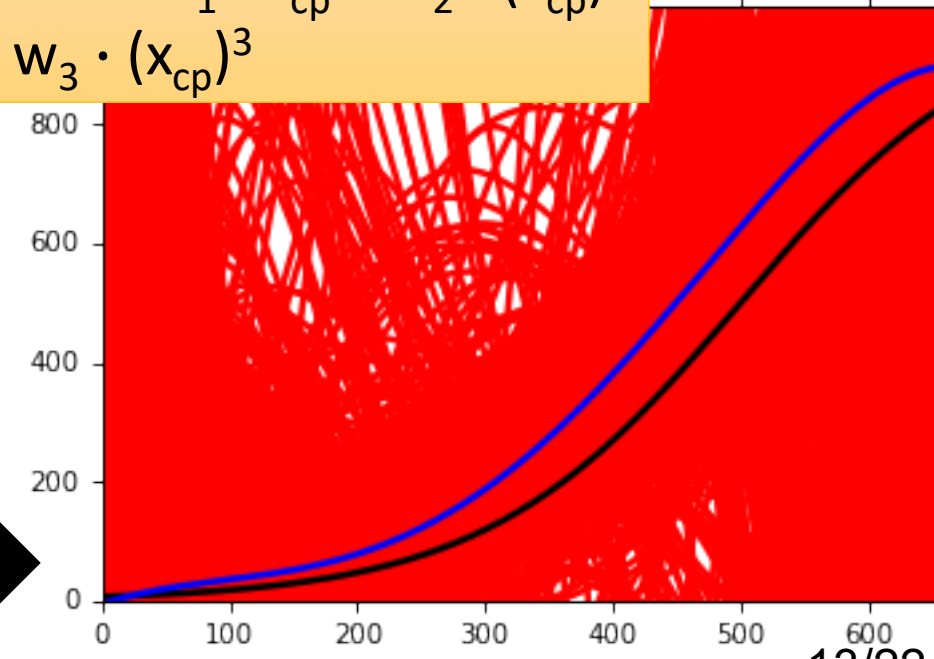
The complexity of the model ↗

$y = b + w \cdot x_{cp}$





$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3$



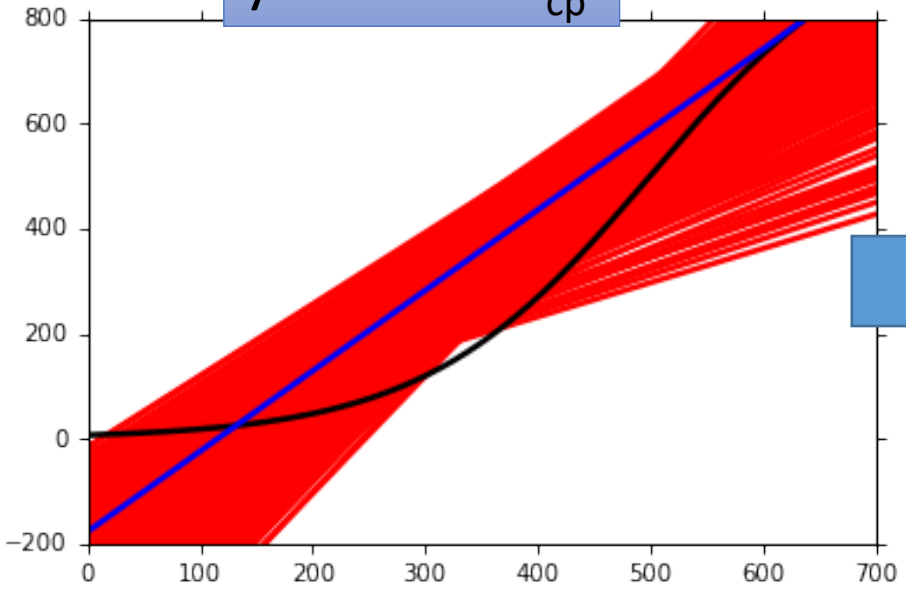
$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$

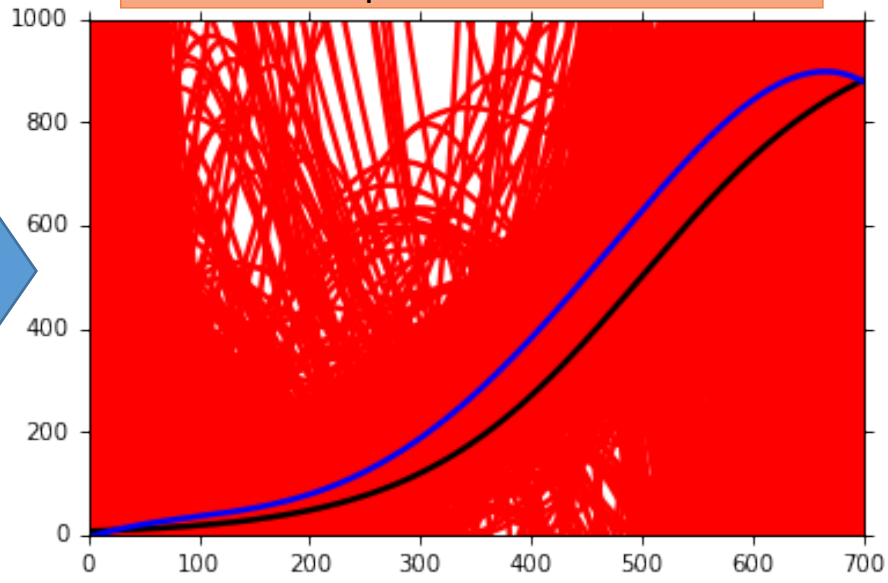
The complexity of the model ↗ ⇒ The bias ↘

Bias

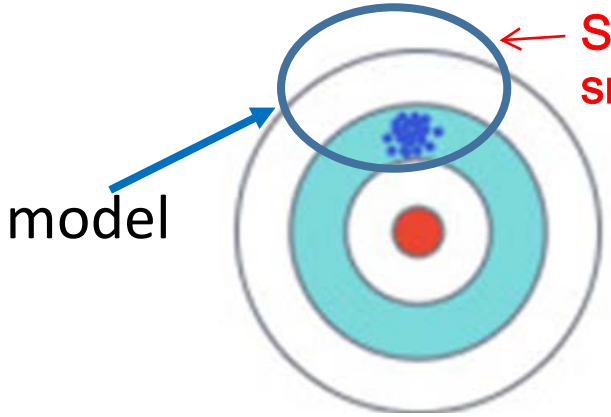
$y = b + w \cdot x_{cp}$

$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$

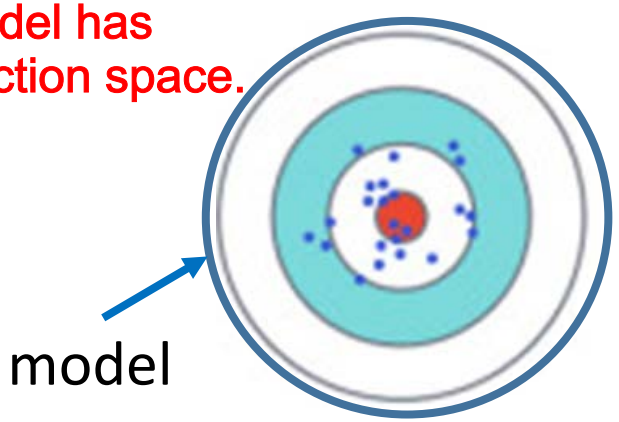




Simpler model has smaller function space.

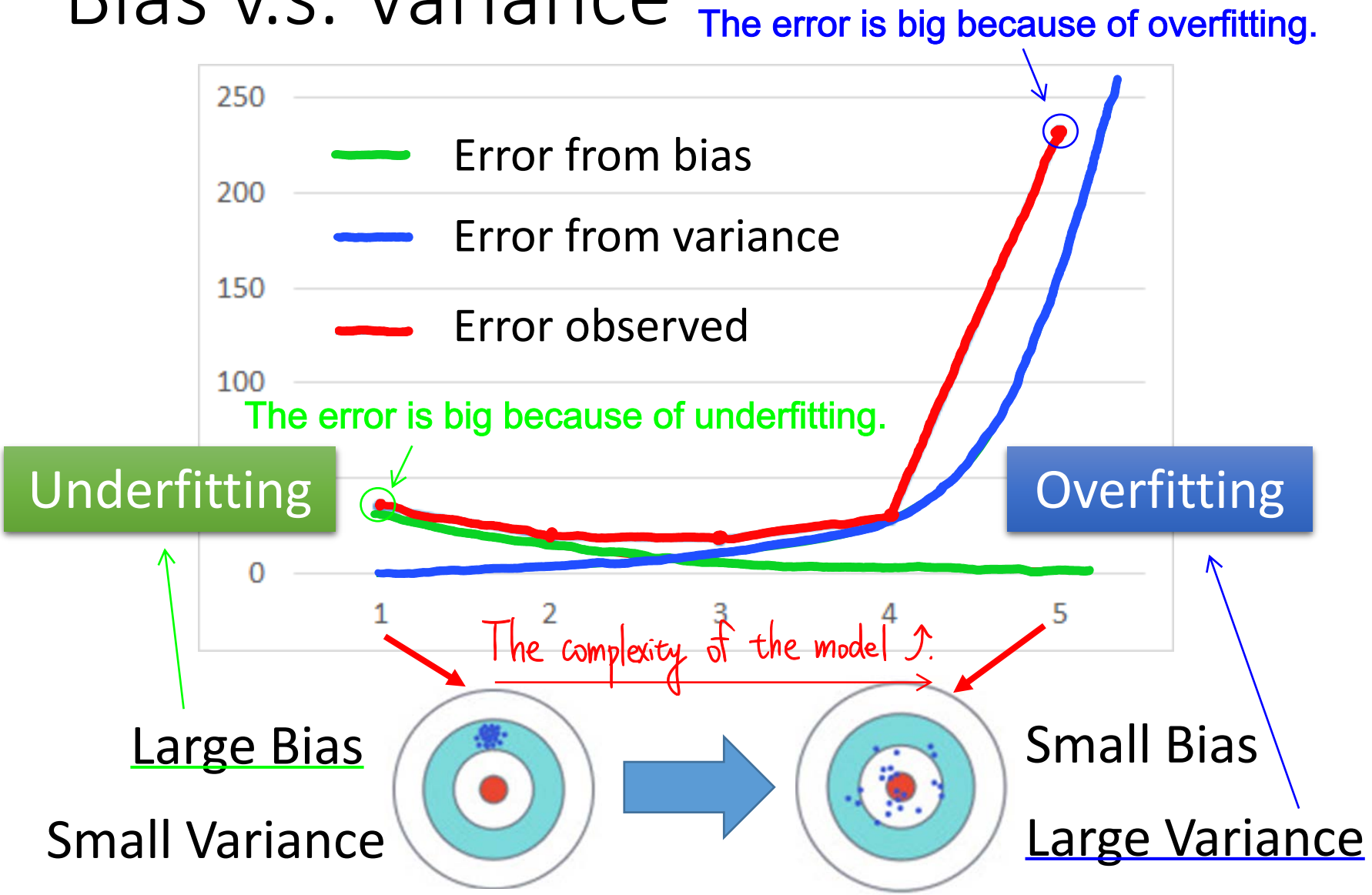


Large Bias



Small Bias

Bias v.s. Variance



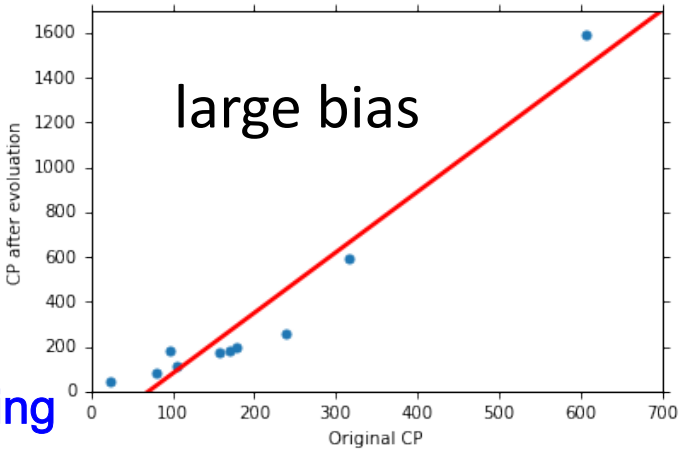
What to do with large bias?

- Diagnosis:
  - If your model cannot even fit the training examples, then you have large bias — Underfitting
  - If you can fit the training data, but large error on testing data, then you probably have large variance — Overfitting
- For bias, redesign your model:
  - Add more features as input
  - A more complex model

large

Underfitting

ex: Feature engineering



Overfitting

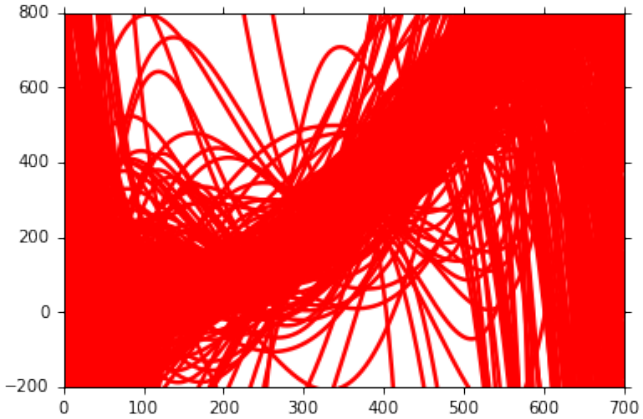
What to do with large variance?

∵ Overfit to training data

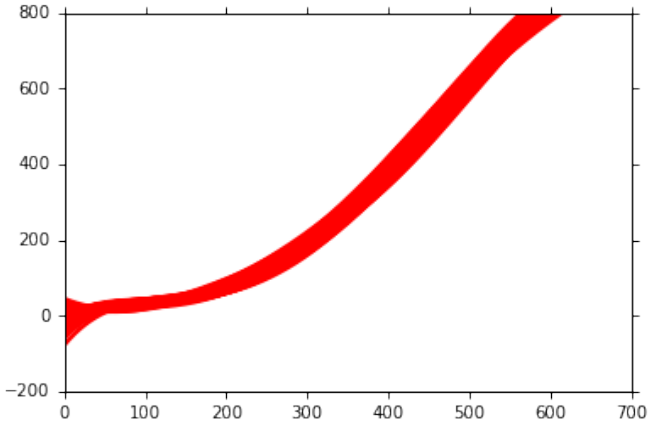
- More data

Very effective, but not always practical

⇒ Create synthetic data

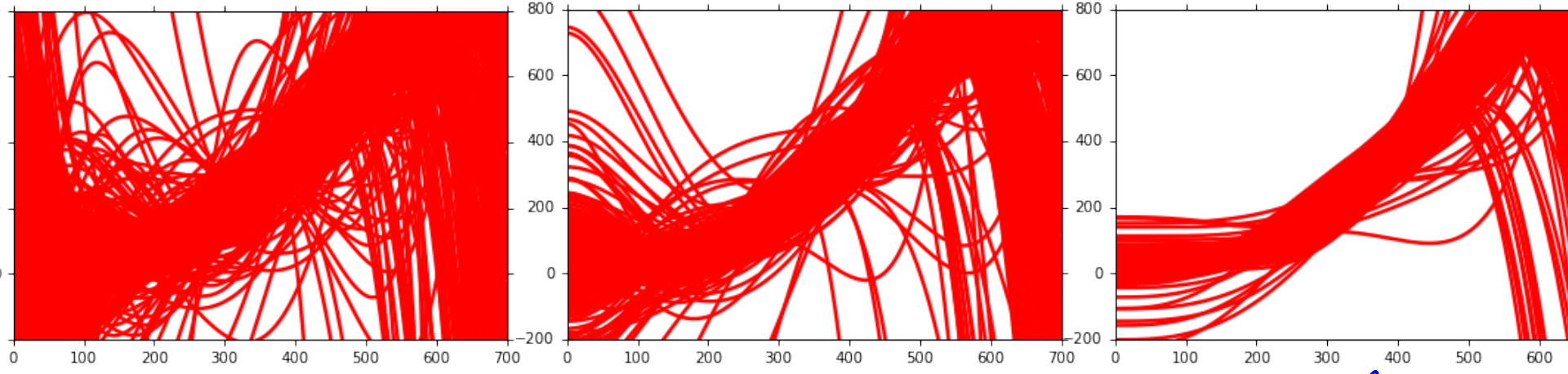


10 examples



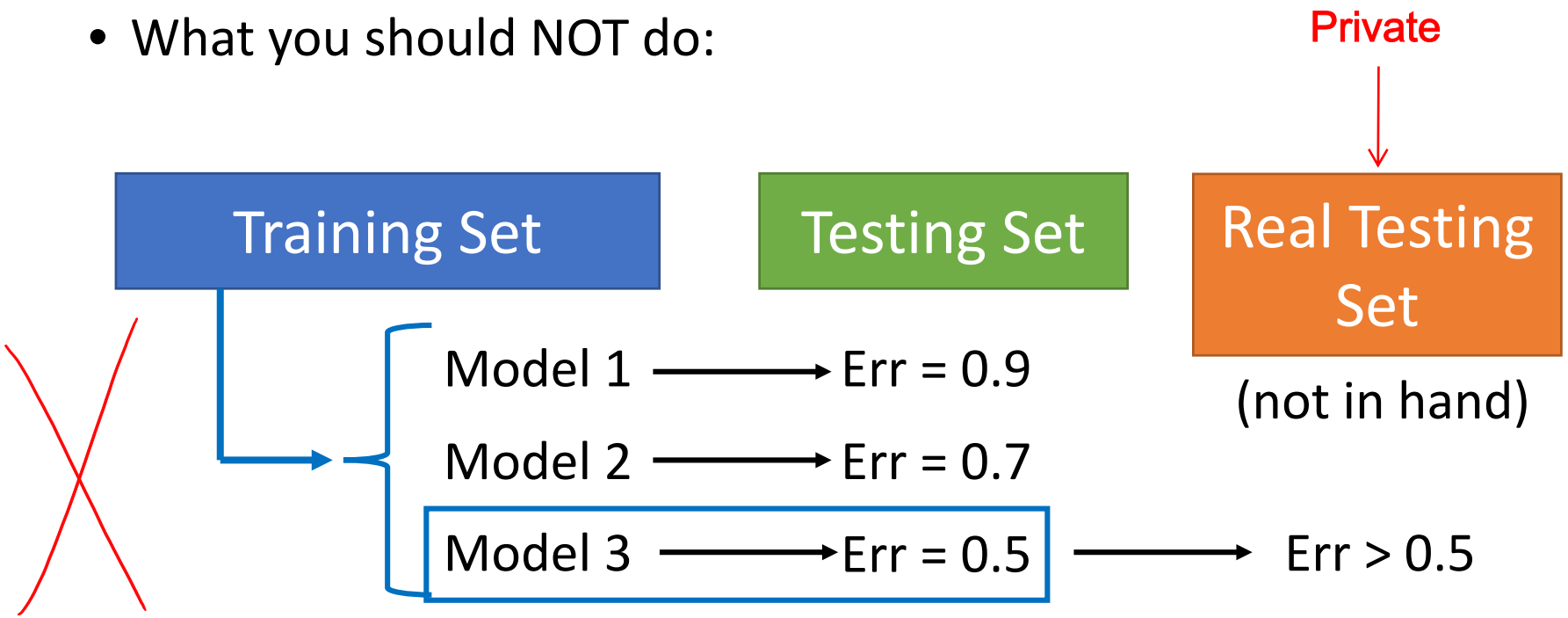
100 examples

- Regularization ➡ May increase bias

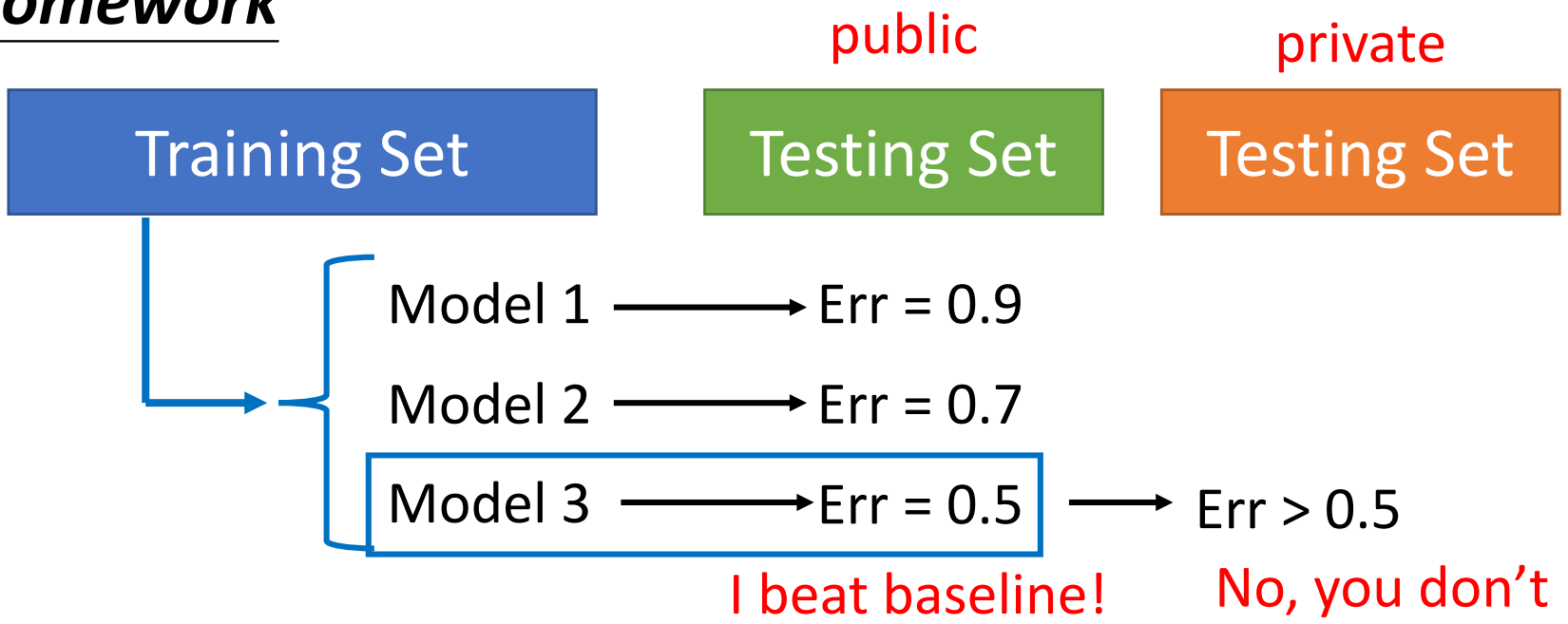


Model Selection

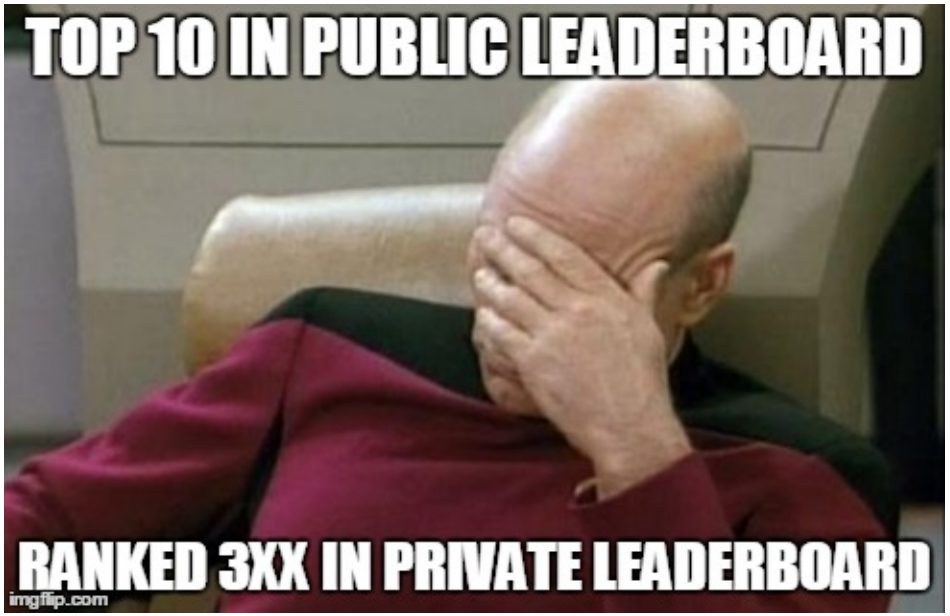
- There is usually a trade-off between bias and variance.
- Select a model that balances two kinds of error to minimize total error Don't want to have overfitting and underfitting.
- What you should NOT do:



Homework



What will happen?



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

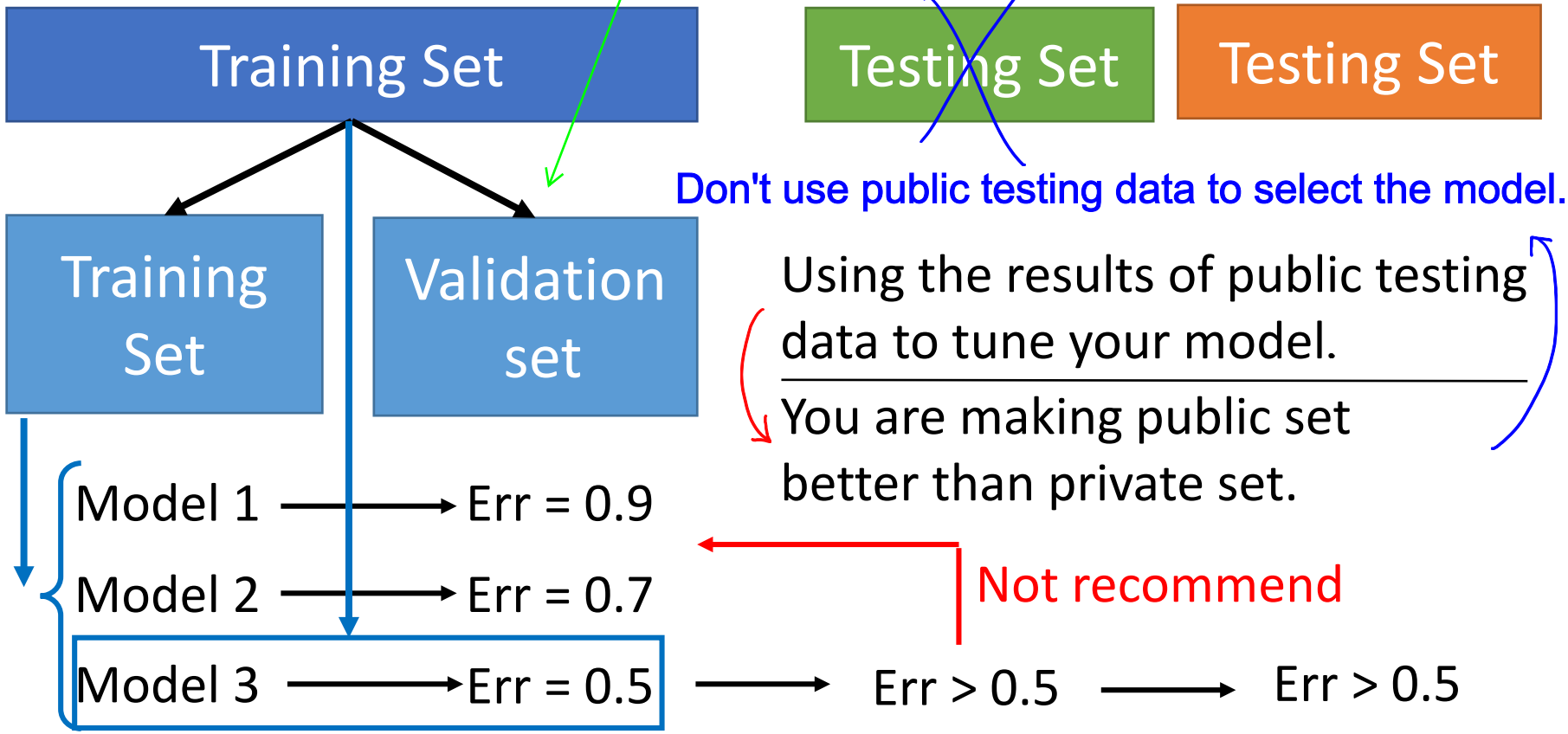
(hyper-parameters)

We use CV to decide the model or training strategy.

Cross Validation (CV)

ex { model A + gradient descent
model A + adam

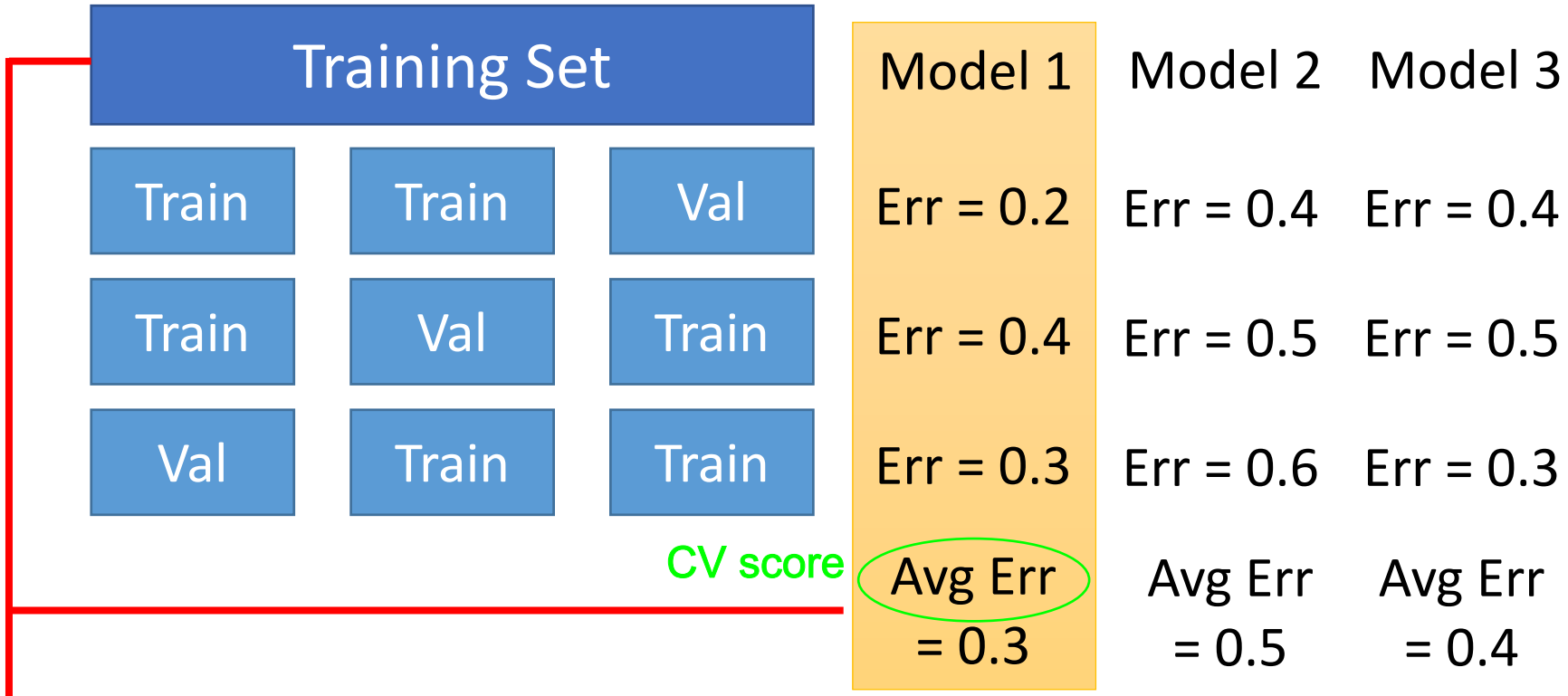
How to prevent the bad division? (Next page)



Divide the training data into N equal parts.

N-fold Cross Validation

Only use training data in CV.



After CV, use the whole training data to train the model, and then send this model for testing.

Testing Set
public

Testing Set
private

Reference

- Bishop: Chapter 3.2