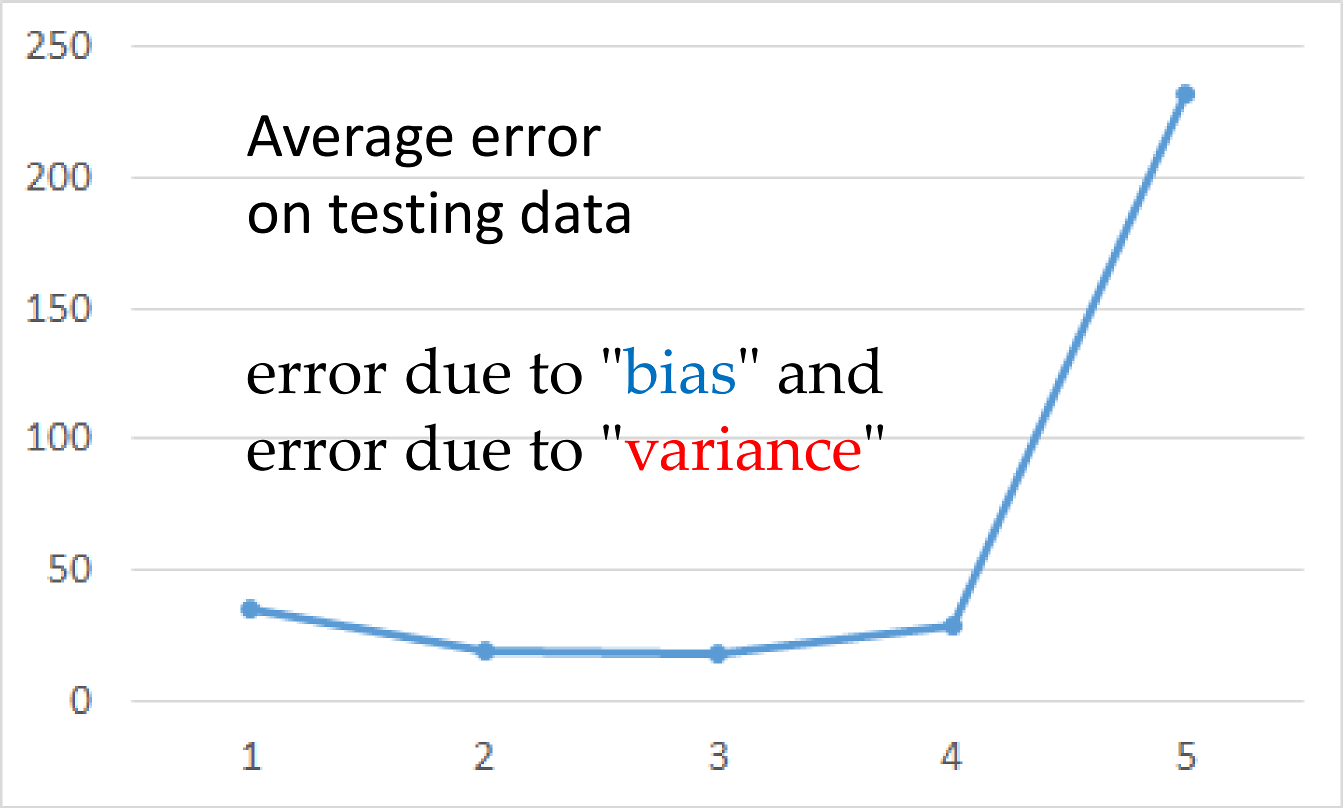
Where does the error come from?

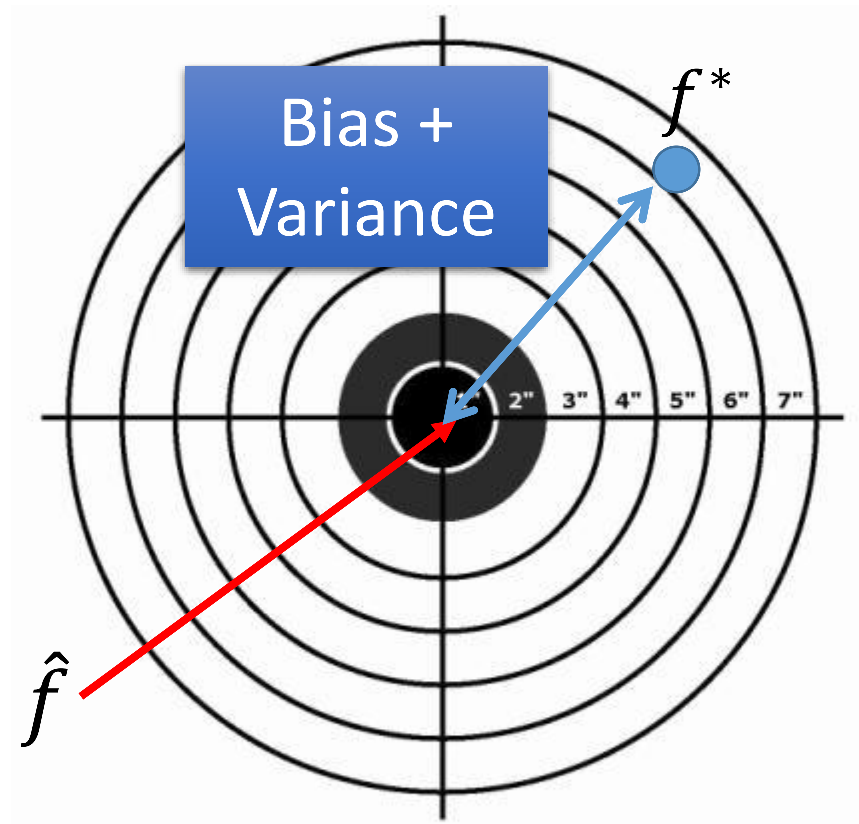
Review



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

Estimator

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



Only Niantic knows $\hat{f}$

From training data,
we find f^*

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

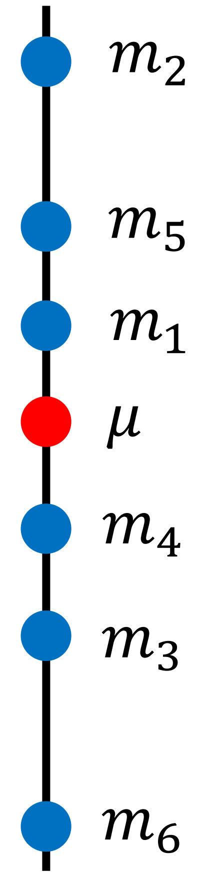
Bias and Variance of Estimator

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

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

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

unbiased



Bias and Variance of Estimator

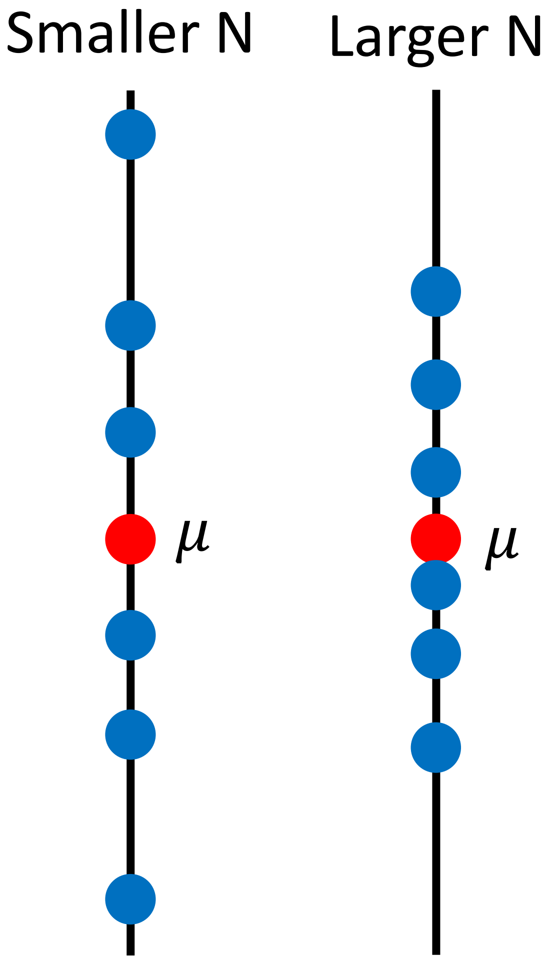
- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of mean μ
  - Sample N points: $\{x^1, x^2, \dots, x^N\}$

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

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

Variance depends on the number of samples

unbiased



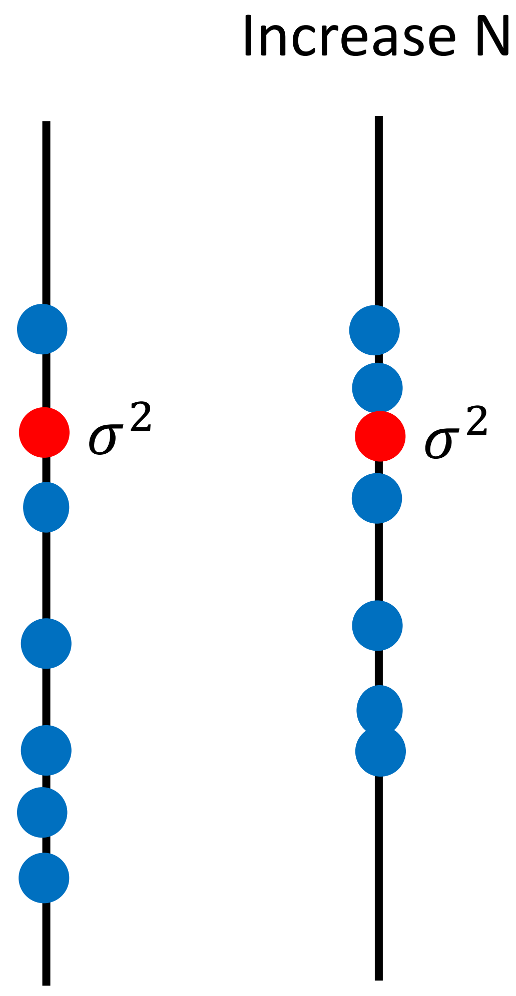
Bias and Variance of Estimator

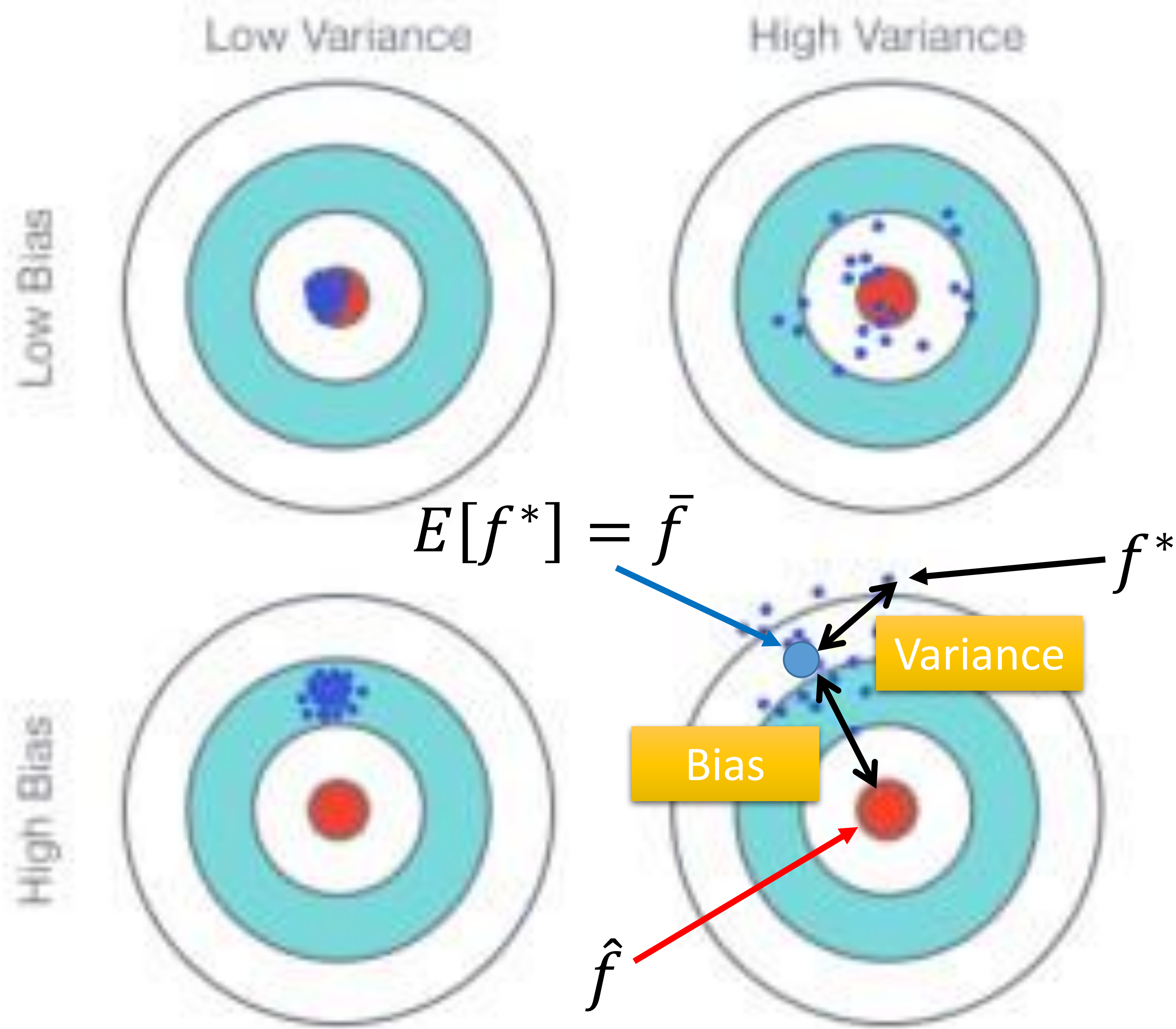
- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of variance σ^2
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

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

Biased estimator

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

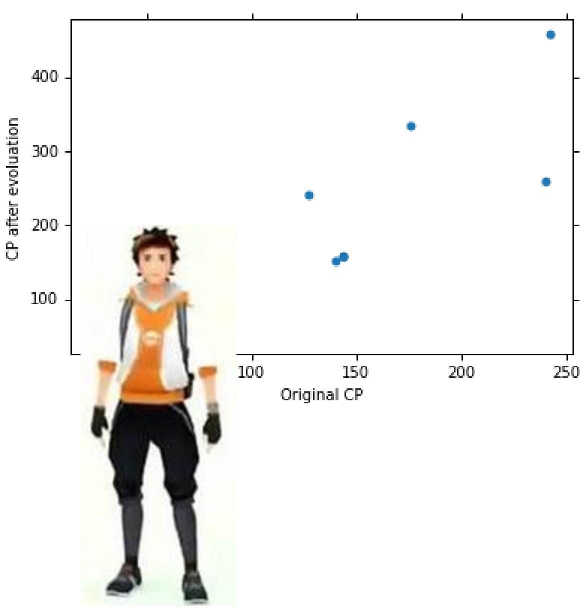



Low Variance
High Variance
Low Bias
High Bias
$E[f^*] = \bar{f}$
f^*
Variance
Bias
$\hat{f}$

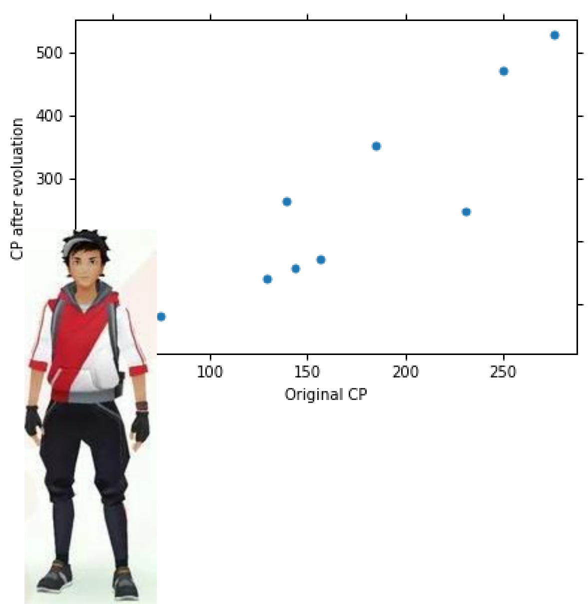
Parallel Universes

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

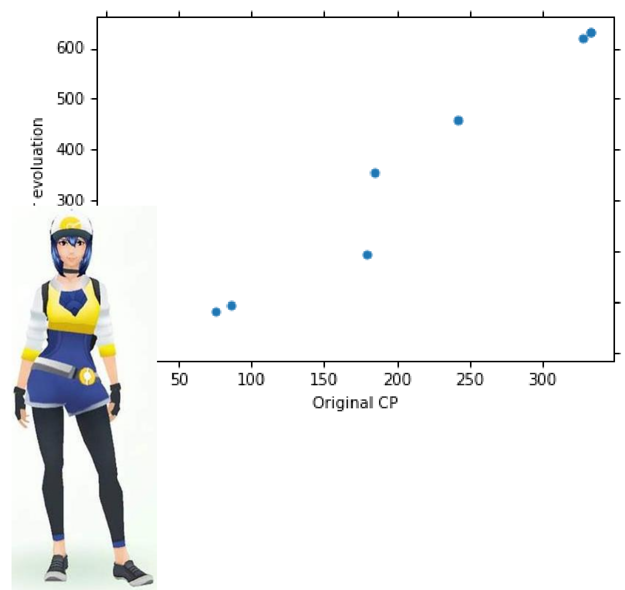
Universe 1



Universe 2



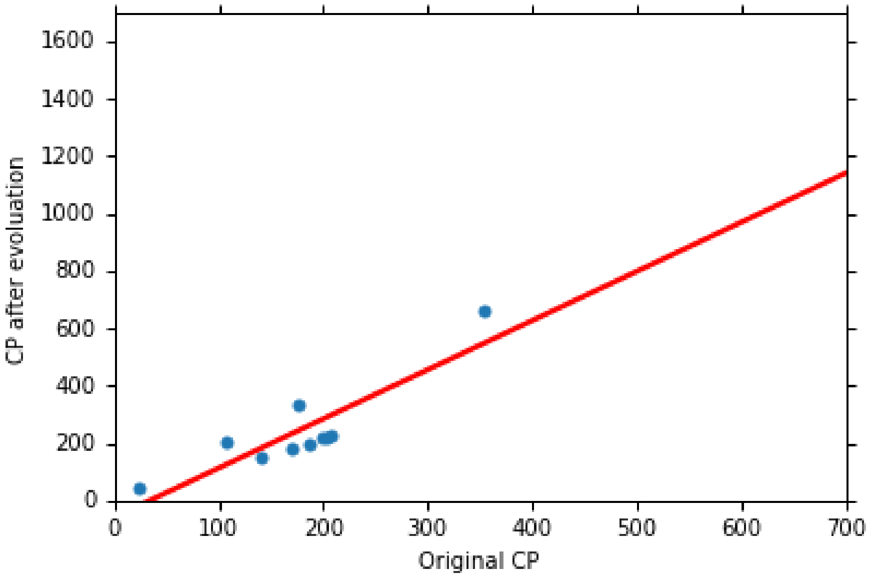
Universe 3



Parallel Universes

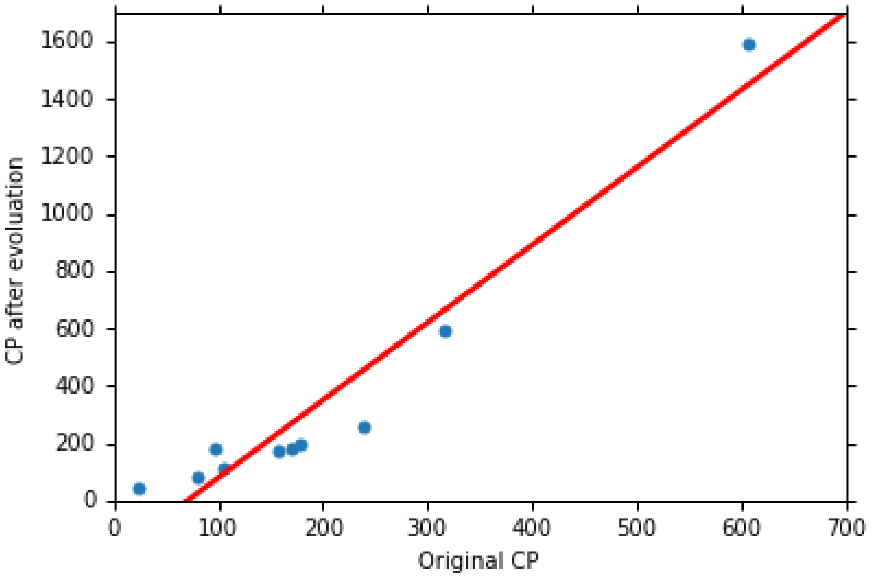
- In different universes, we use the same model, but obtain different f^*

Universe 123



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

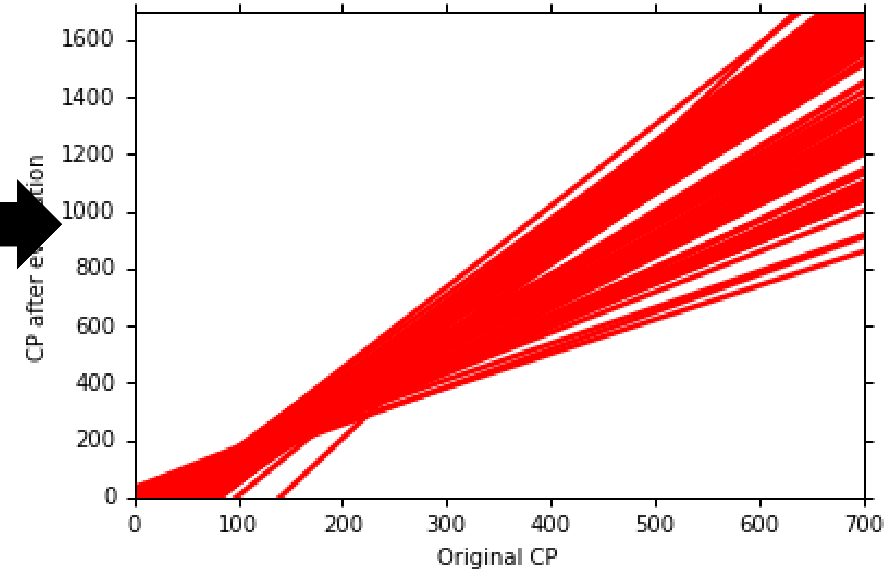
Universe 345



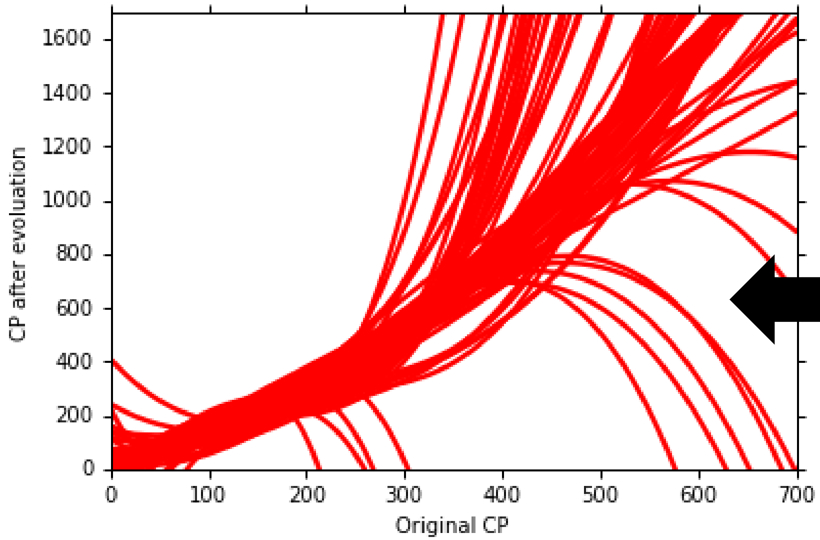
$$y = b + w \cdot x_{cp}$$

f^* in 100 Universes

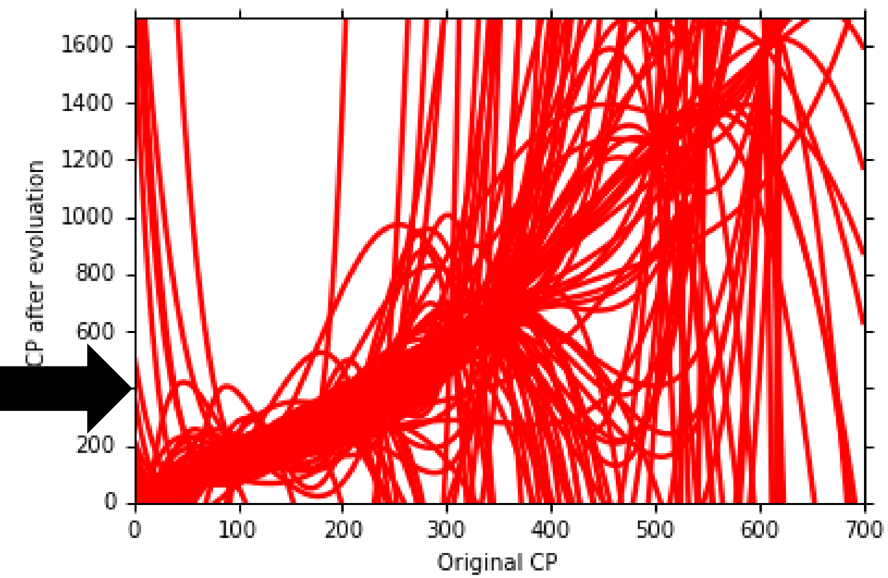
$$y = b + w \cdot x_{cp}$$



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



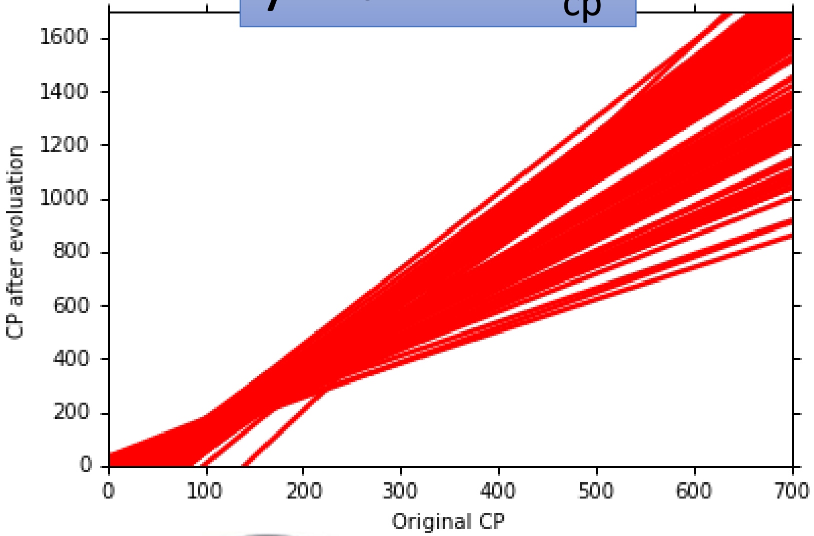
$$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$$

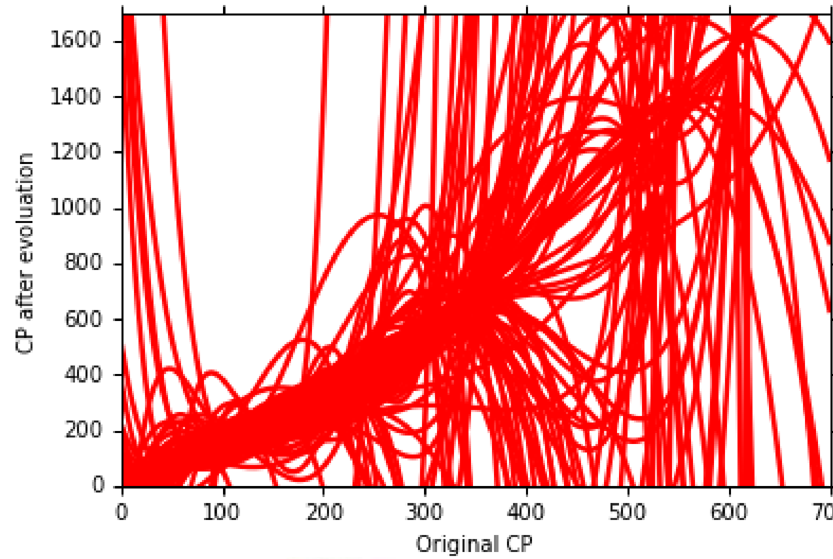


Variance

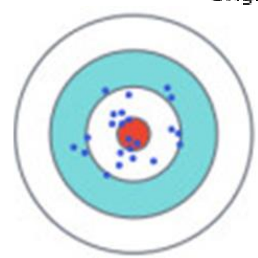
$$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$$

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





Small Variance



Large Variance

Simpler model is less influenced by the sampled data

Consider the extreme case f(x) = c

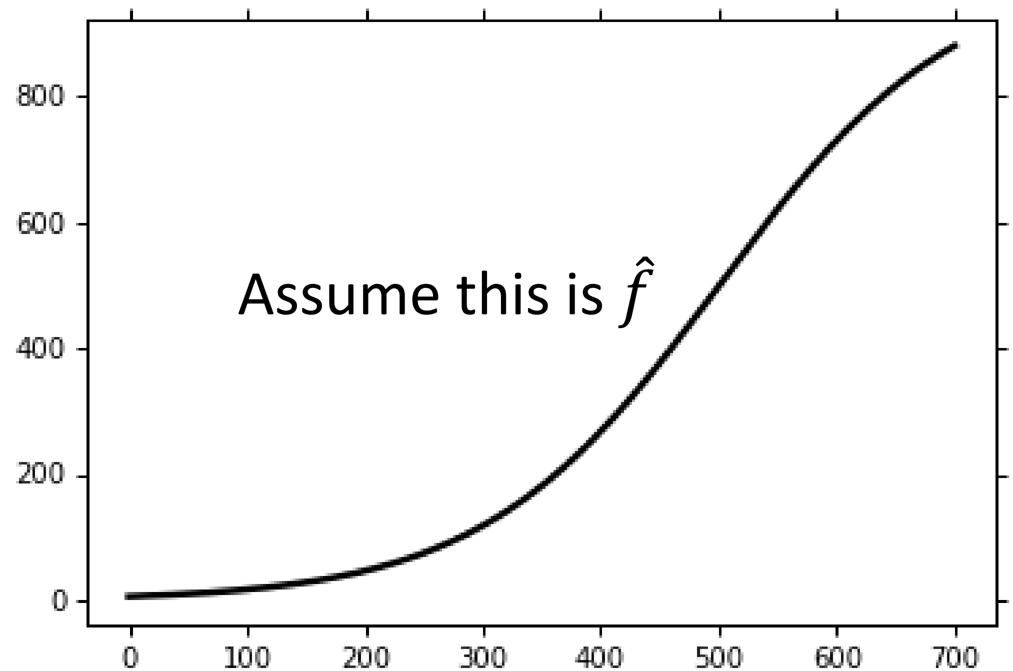
Bias

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

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

Large Bias

Small Bias

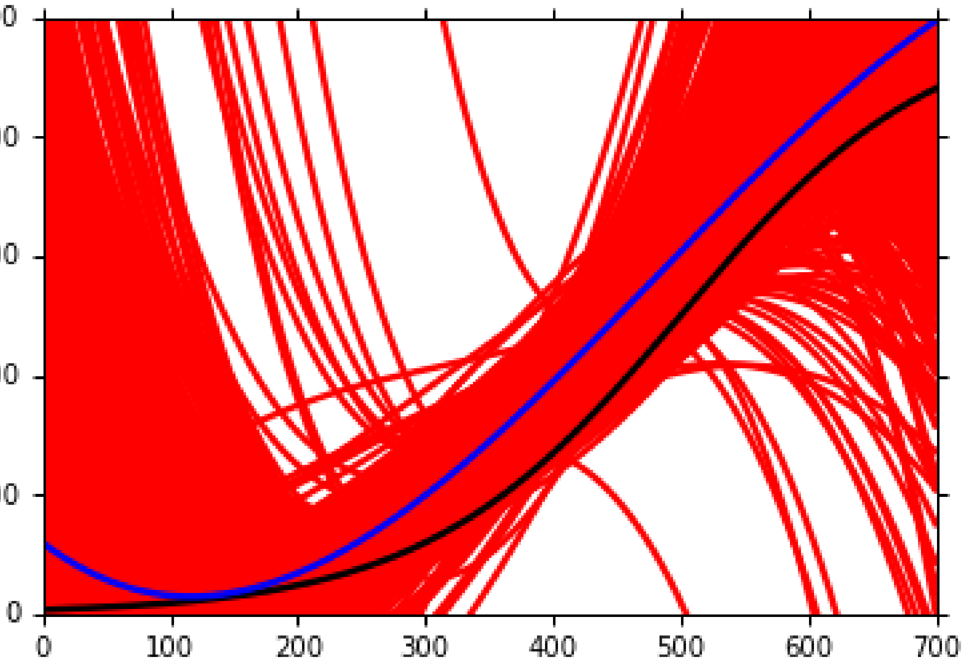


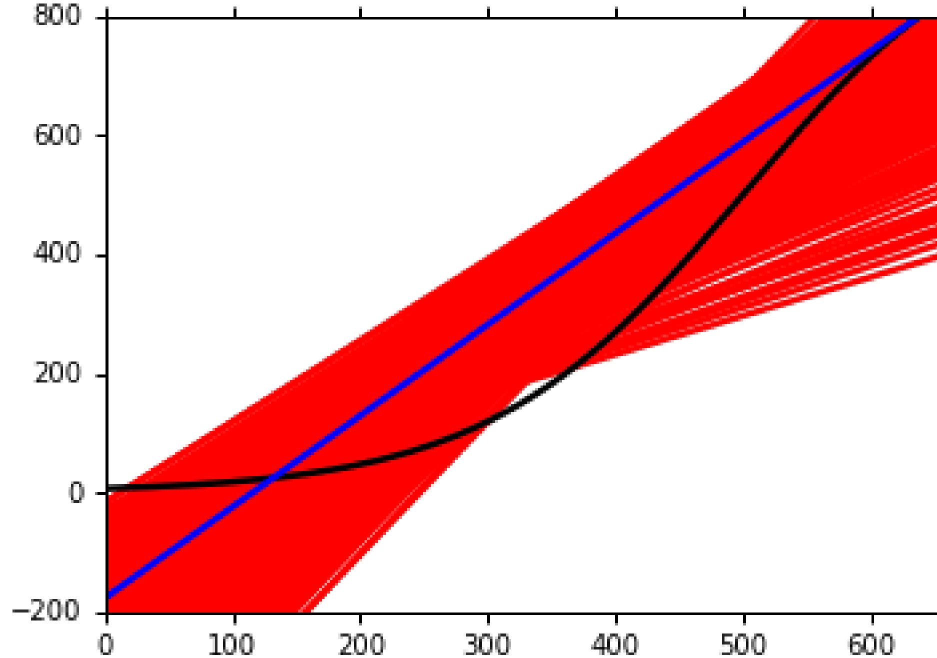
Black curve: the true function $\hat{f}$

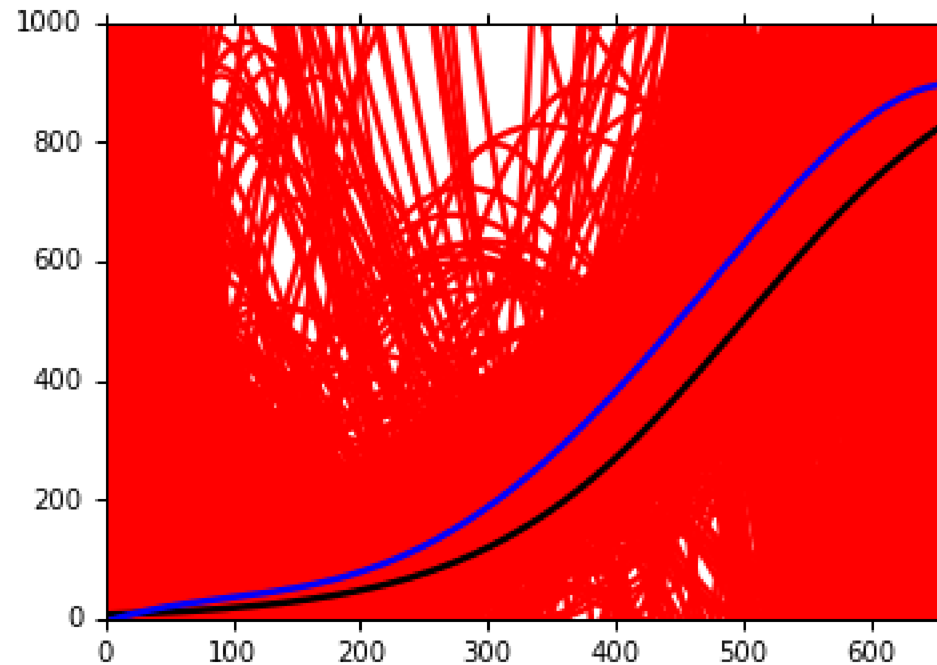
Red curves: 5000 f^*

Blue curve: the average of 5000 f^*

$= \bar{f}$

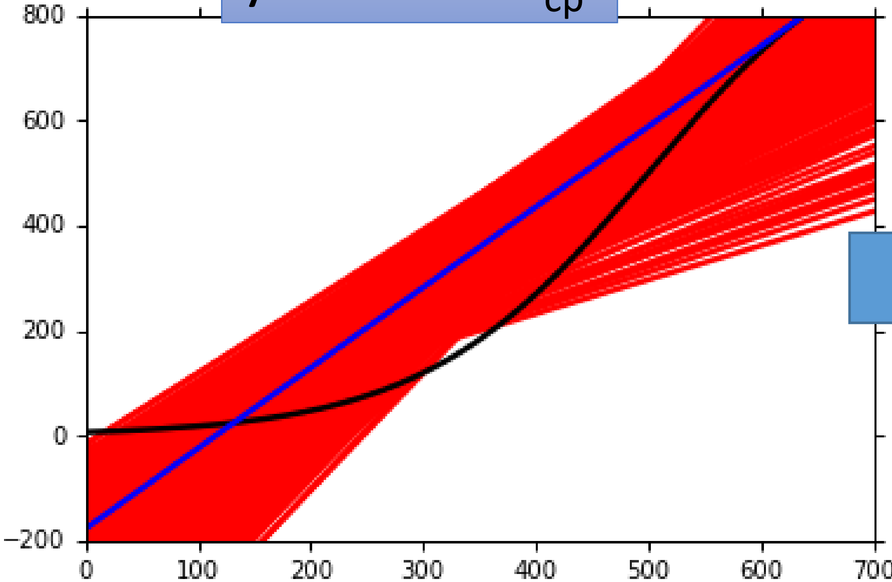




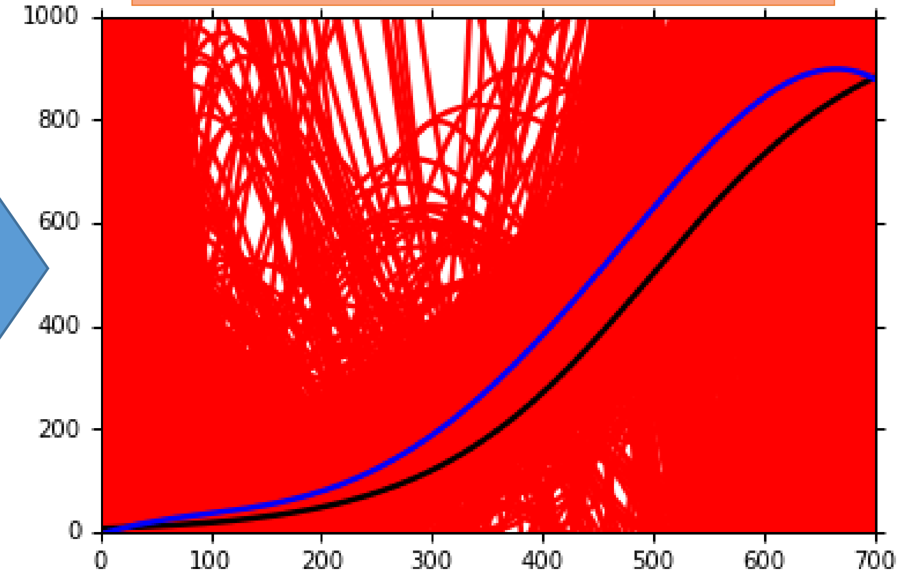


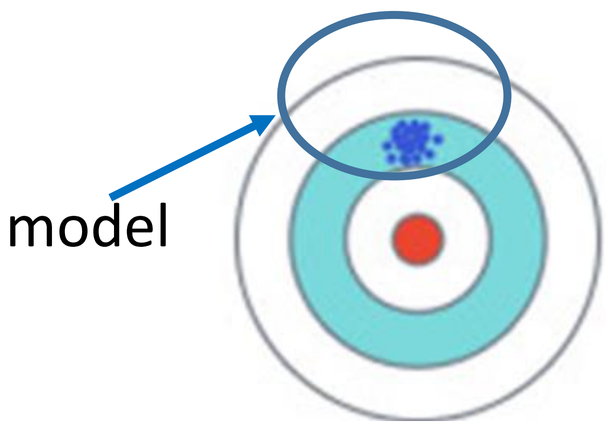
Bias

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

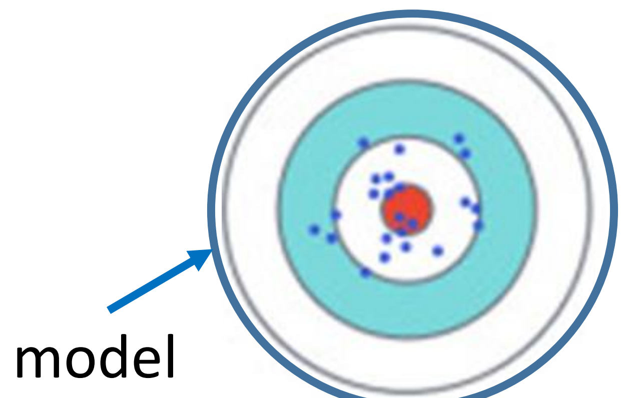


$$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$$



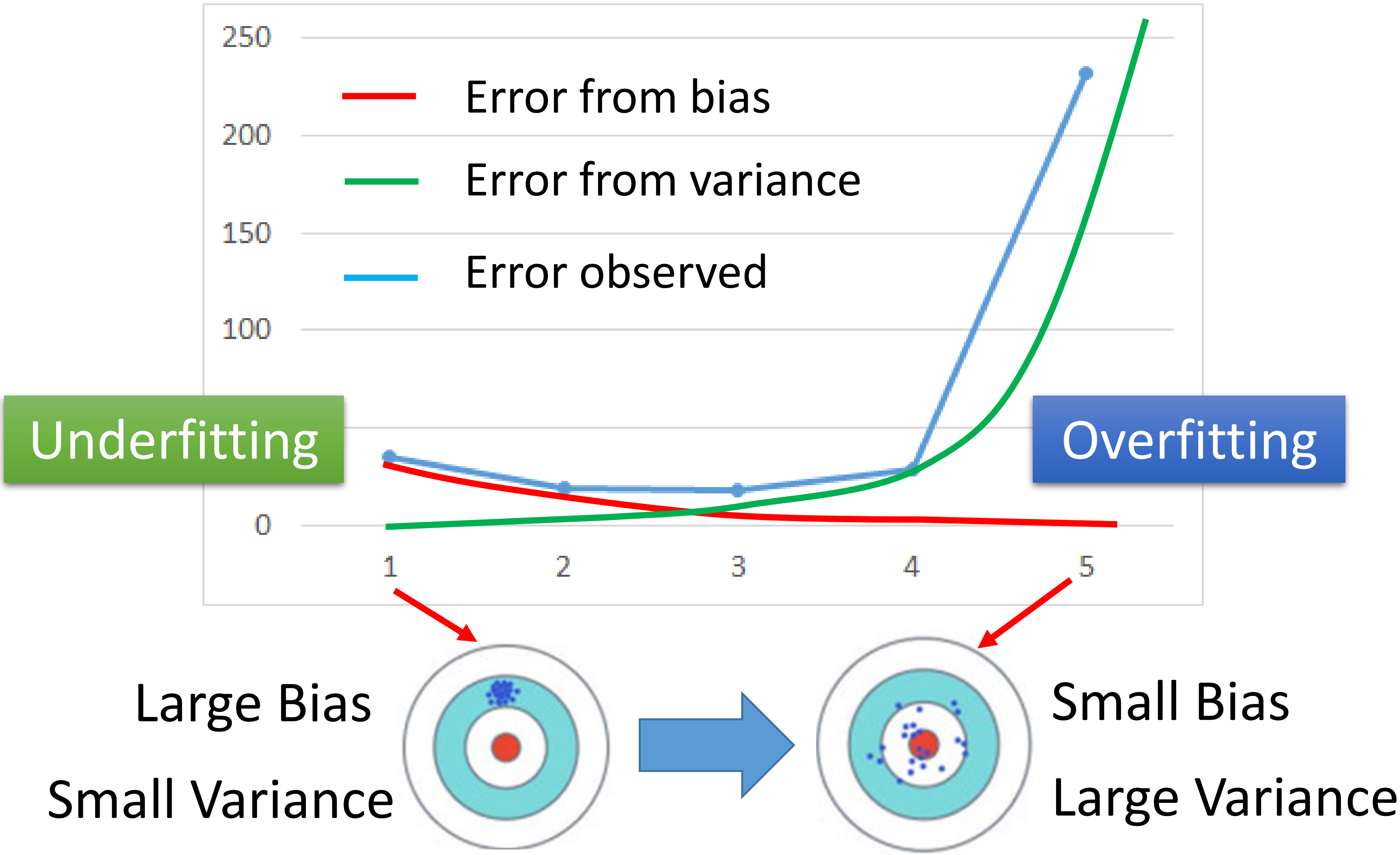


Large Bias



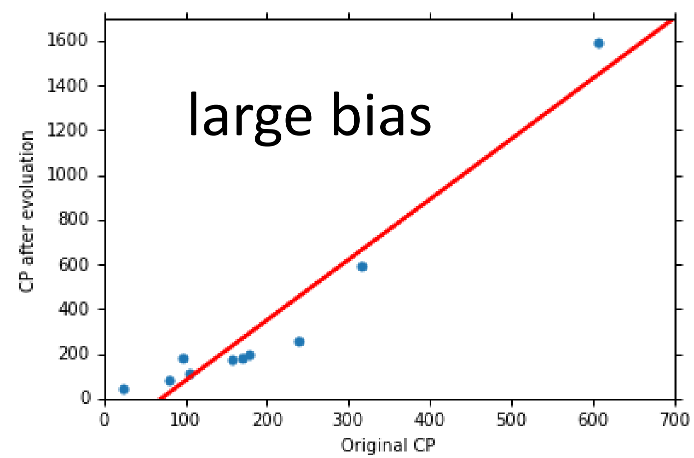
Small Bias

Bias v.s. Variance



What to do with large bias?

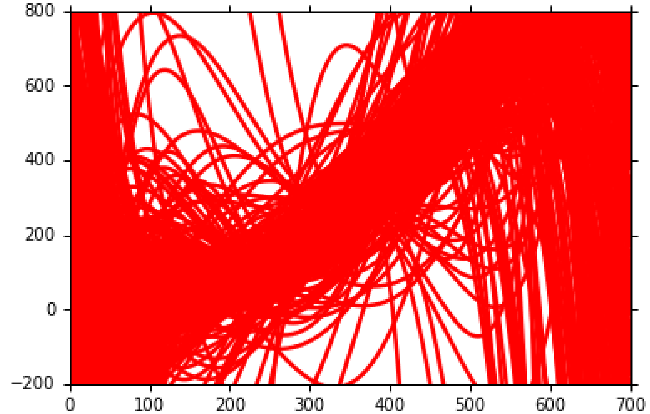
- Diagnosis:
  - If your model cannot even fit the training examples, then you have large bias **Underfitting**
  - If you can fit the training data, but large error on testing data, then you probably have large variance **Overfitting**
- For bias, redesign your model:
  - Add more features as input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



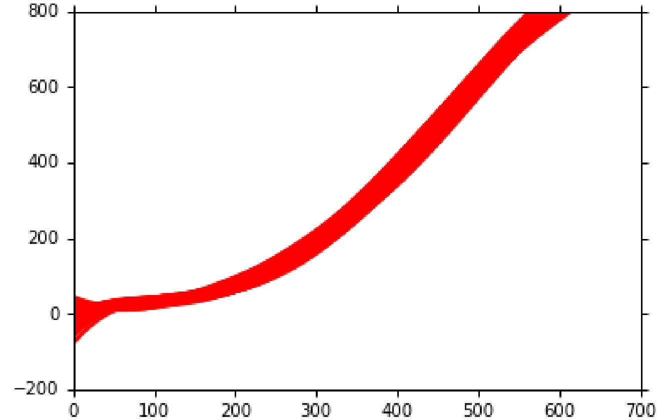
What to do with large variance?

- More data

Very effective,
but not always
practical

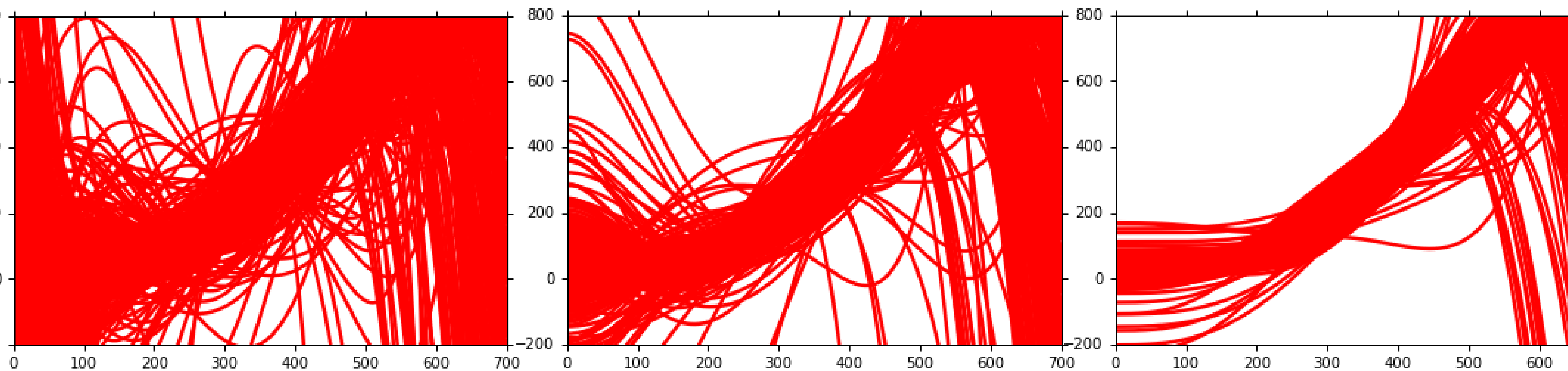


10 examples



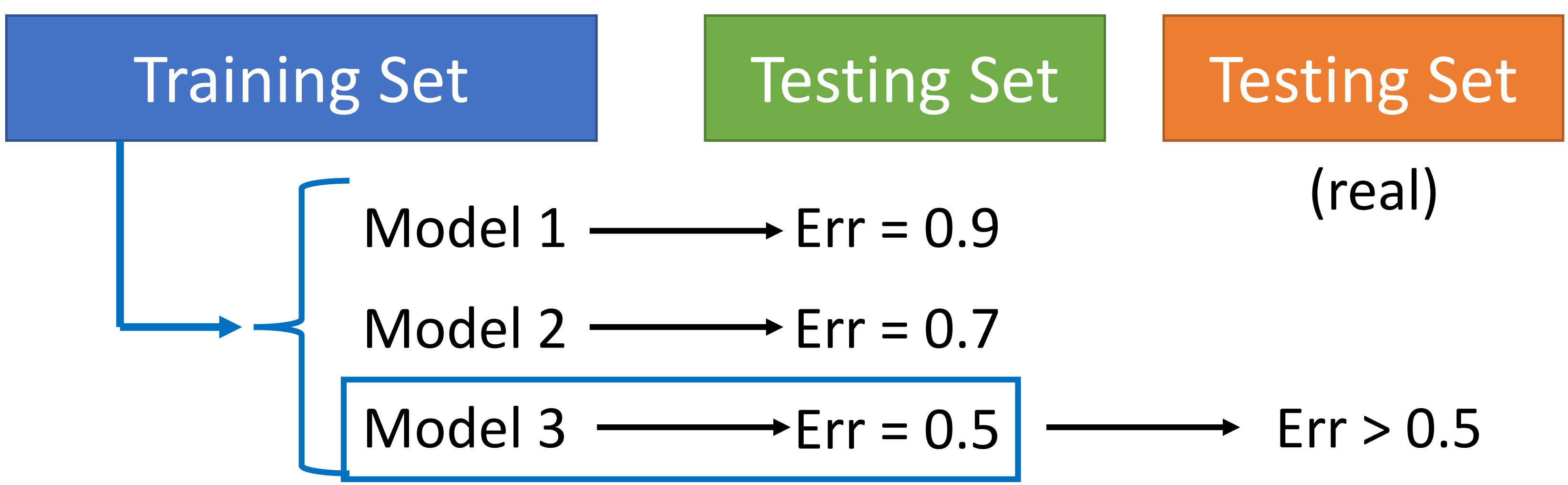
100 examples

- Regularization



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
- What you should NOT do:



Homework

Training Set	public Testing Set	private Testing Set

Model 1 ⟶ Err = 0.9

Model 2 ⟶ Err = 0.7

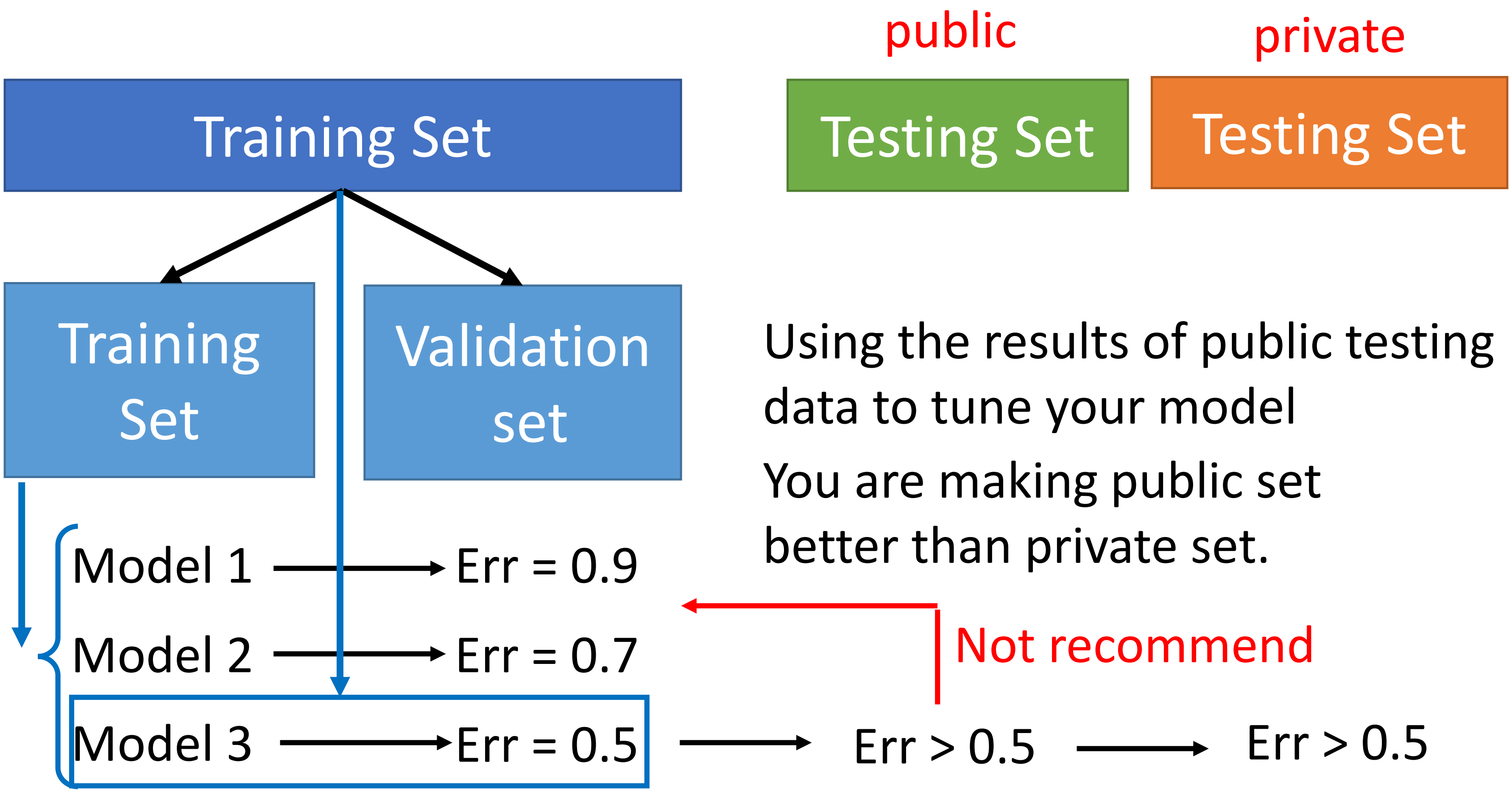
Model 3 ⟶ Err = 0.5 ⟶ Err > 0.5

I beat baseline! No, you don't

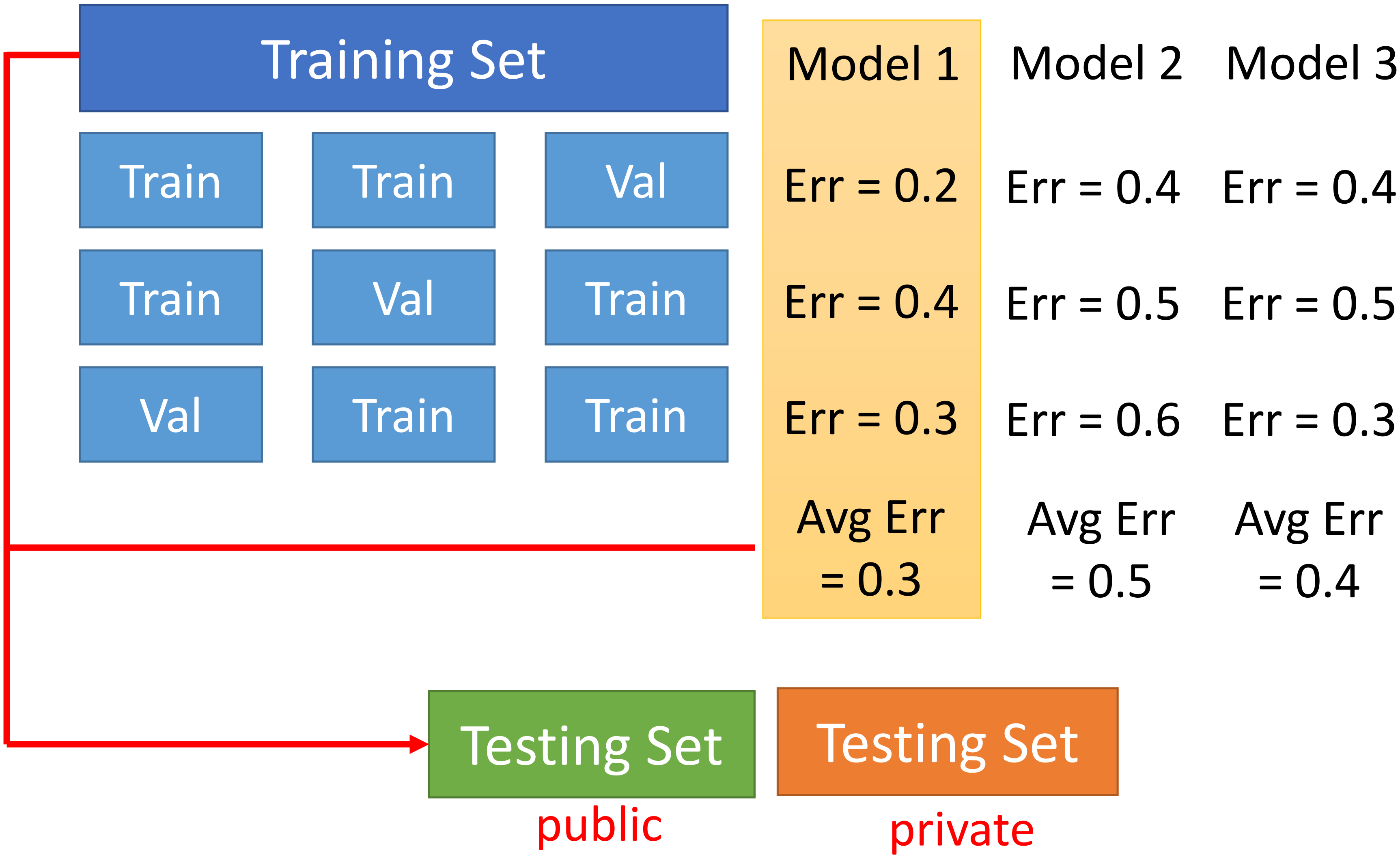
What will happen next Friday?

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

Cross Validation



N-fold Cross Validation



Reference

- Bishop: Chapter 3.2