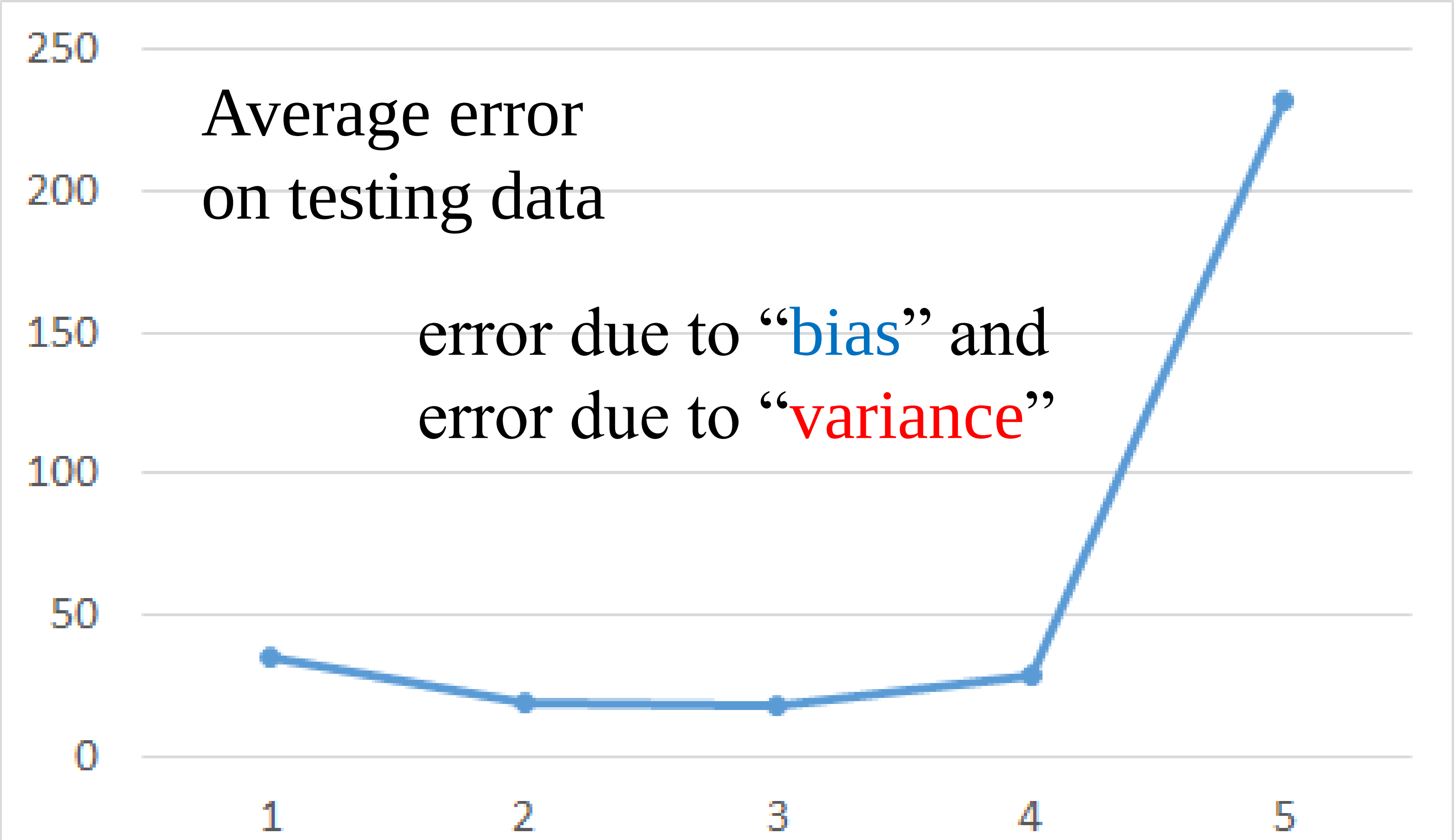
Where does error come from?

2020/11/04

Review



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

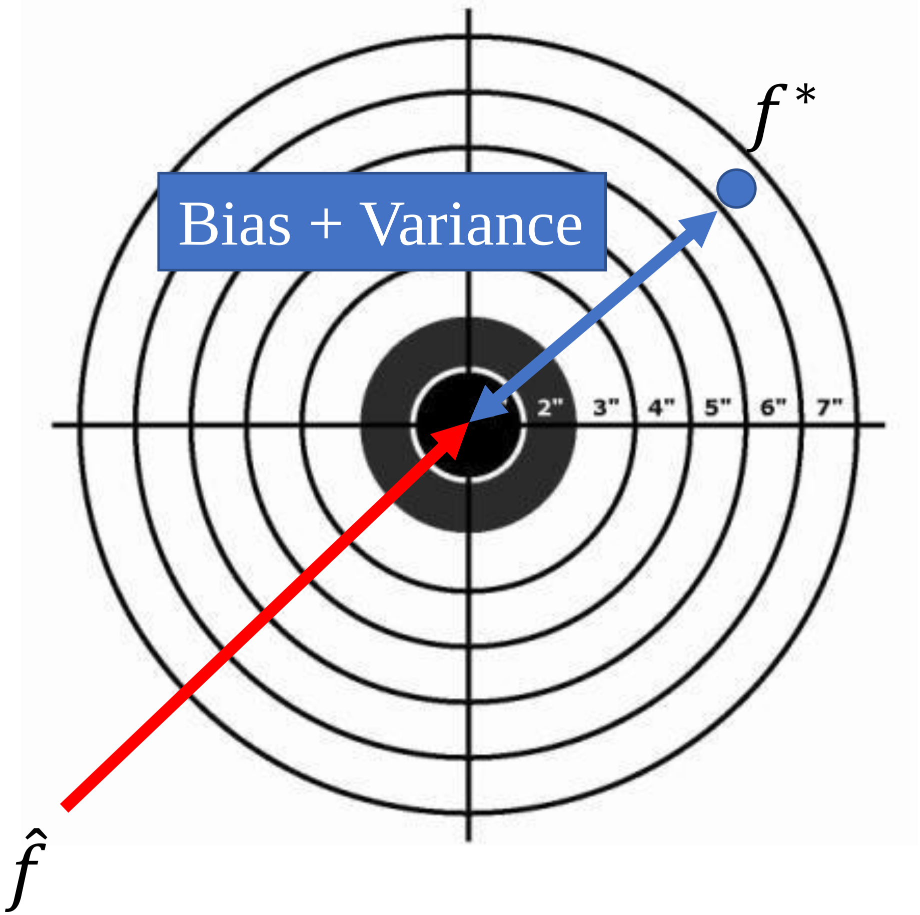
Estimator

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

only Niantic knows $\hat{f}$

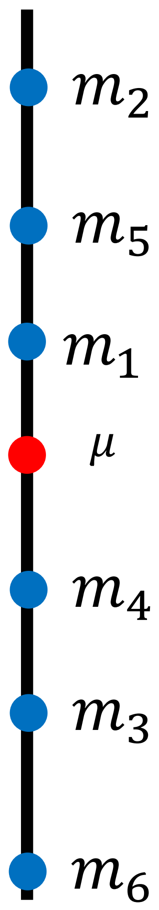
from training data, we find f^*

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



Bias and Variance of Estimator

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: $\{x^1, x^2, \dots, x^N\}$

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

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

Bias and Variance of Estimator

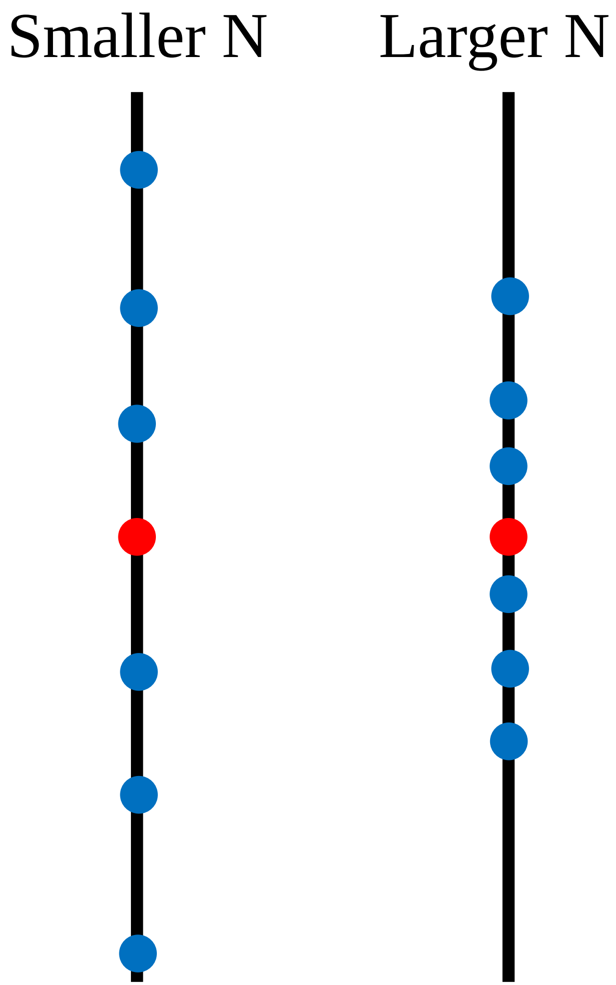
- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of mean μ
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

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

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

Variance depends on the number of samples

unbiased



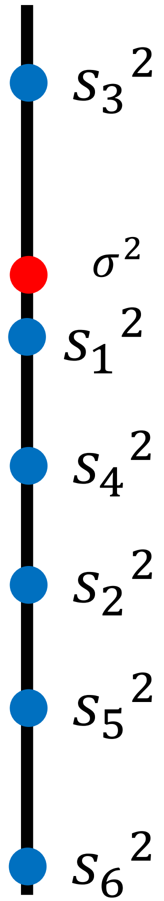
Bias and Variance of Estimator

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

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

biased estimator

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



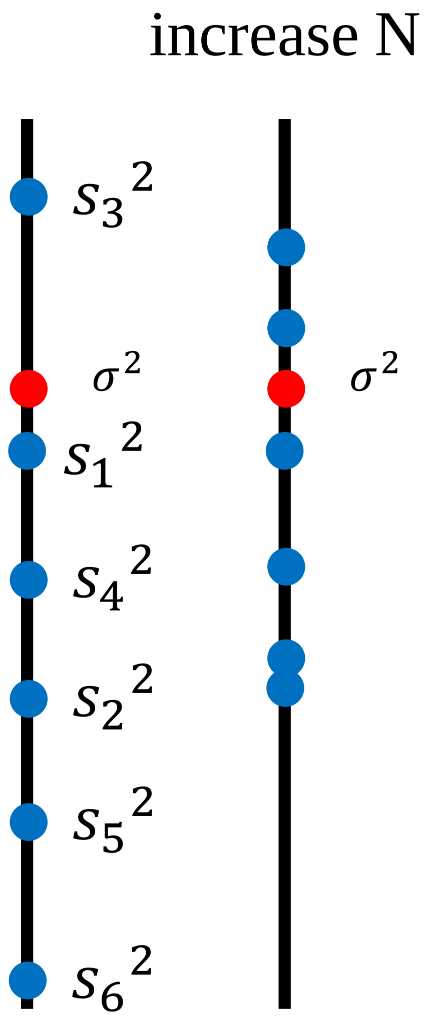
Bias and Variance of Estimator

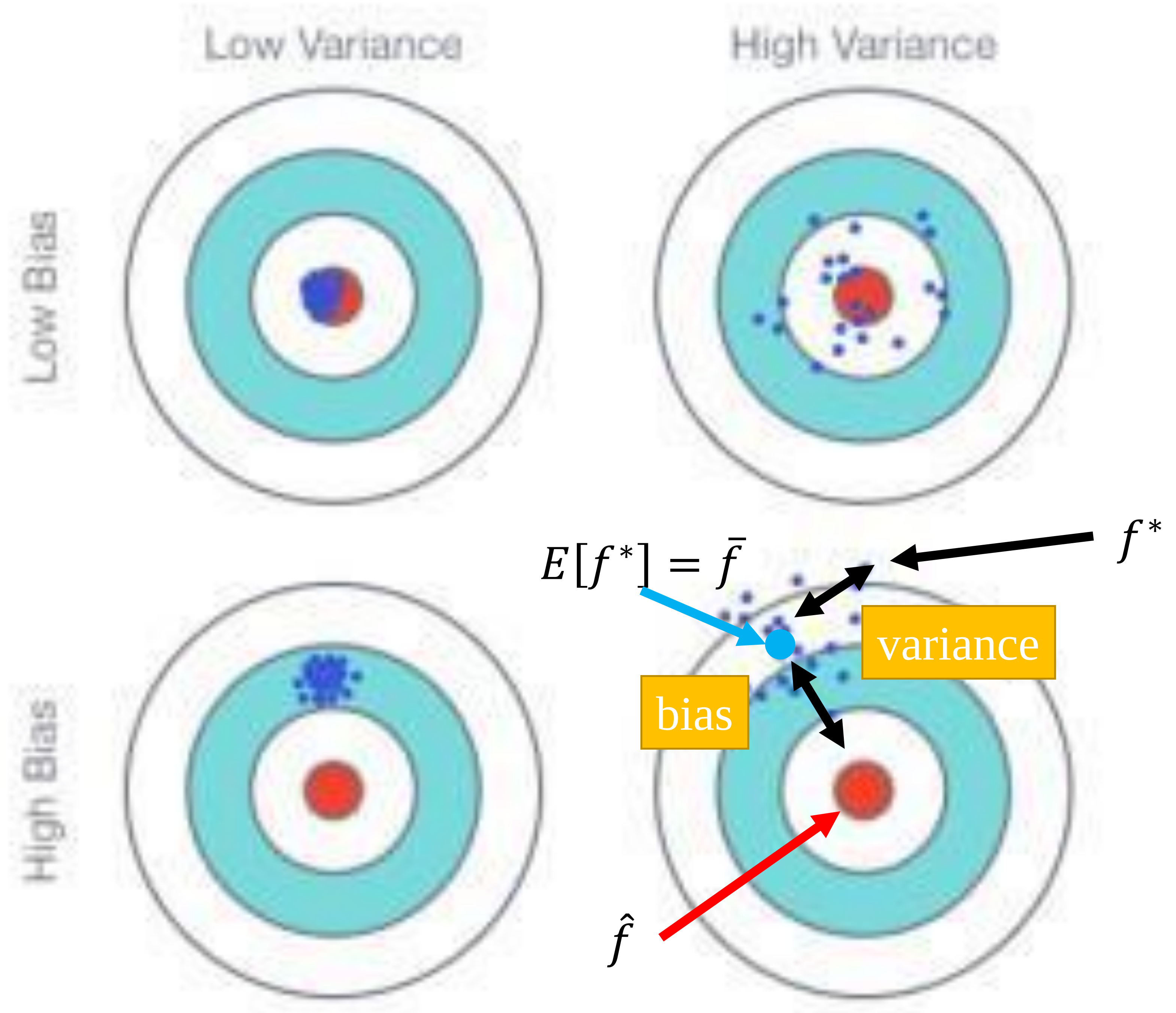
- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of mean μ
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

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

biased estimator

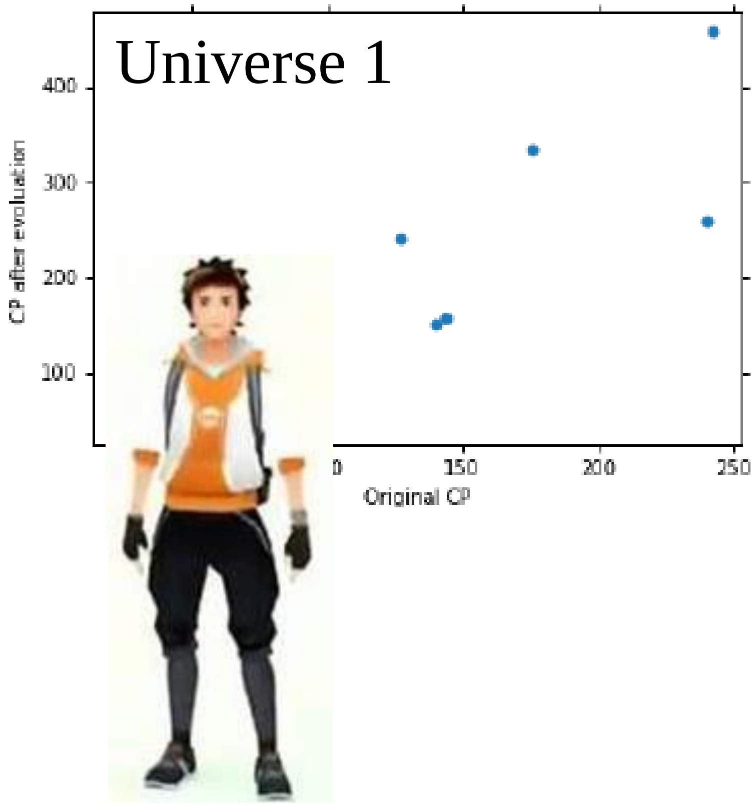
$$E[s^2] = \frac{N-1}{N}\sigma^2 \quad \neq \sigma$$

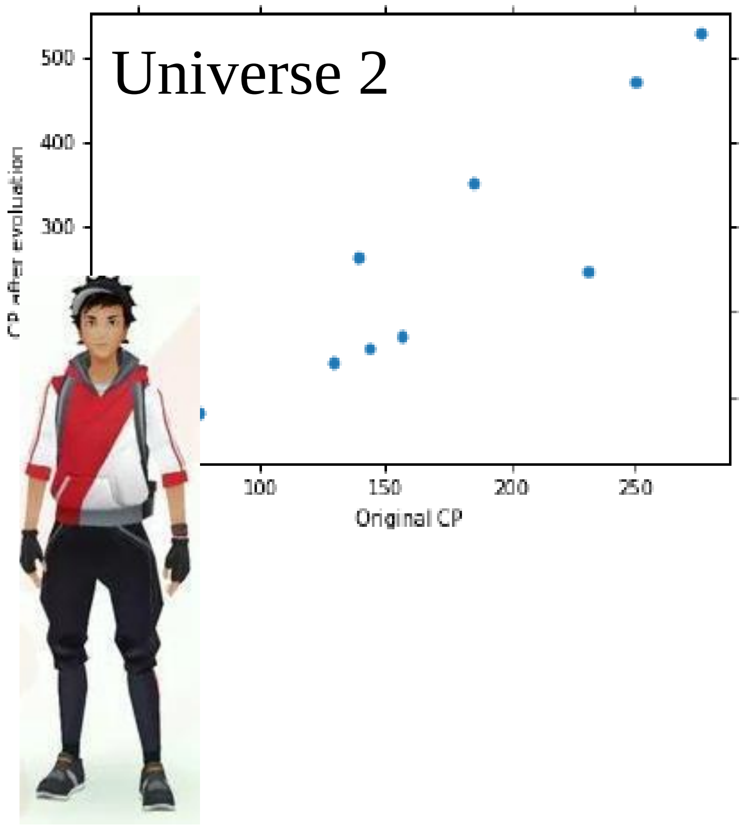


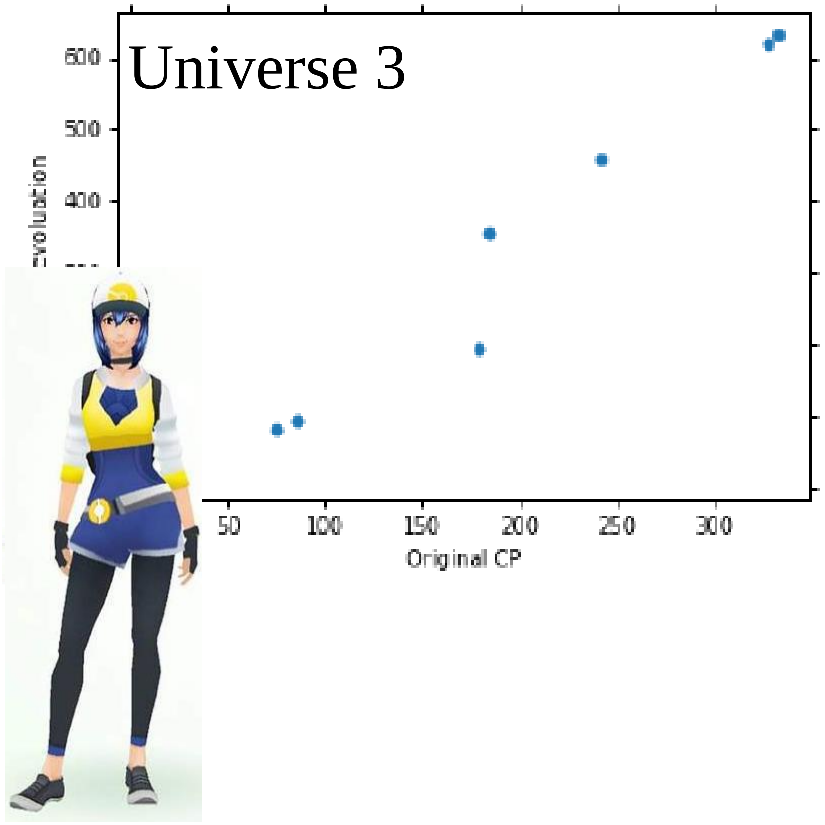

Low Variance
High Variance
Low Bias
High Bias
$E[f^*] = \bar{f}$
f^*
variance
bias
$\hat{f}$

Parallel Universes

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

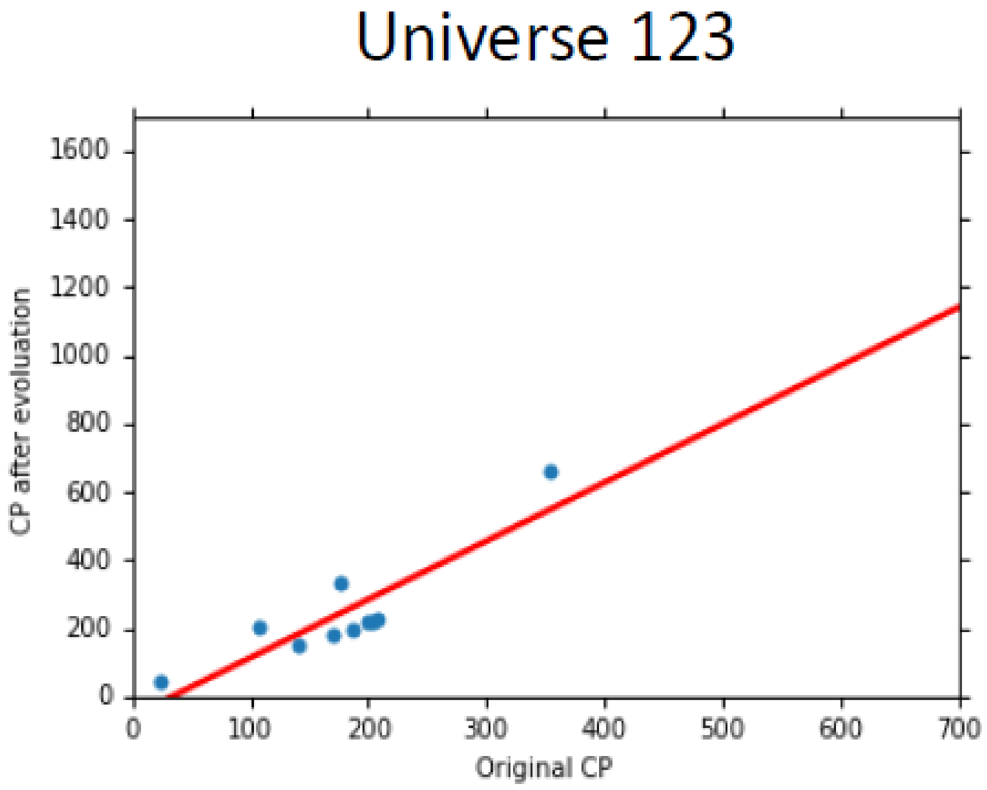




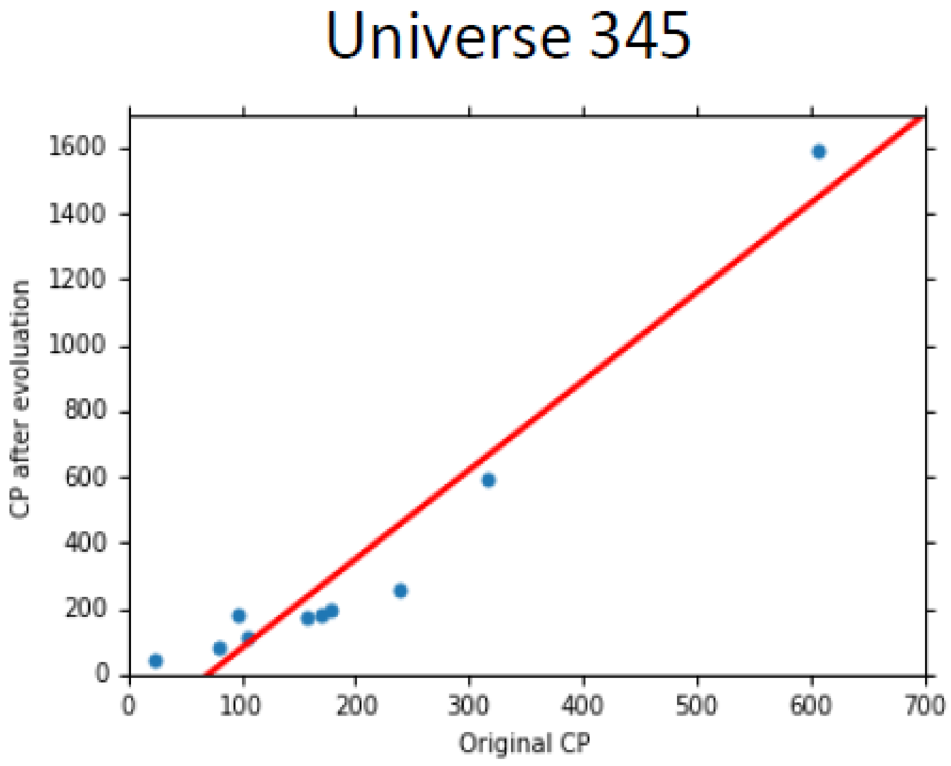


Parallel Universes

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



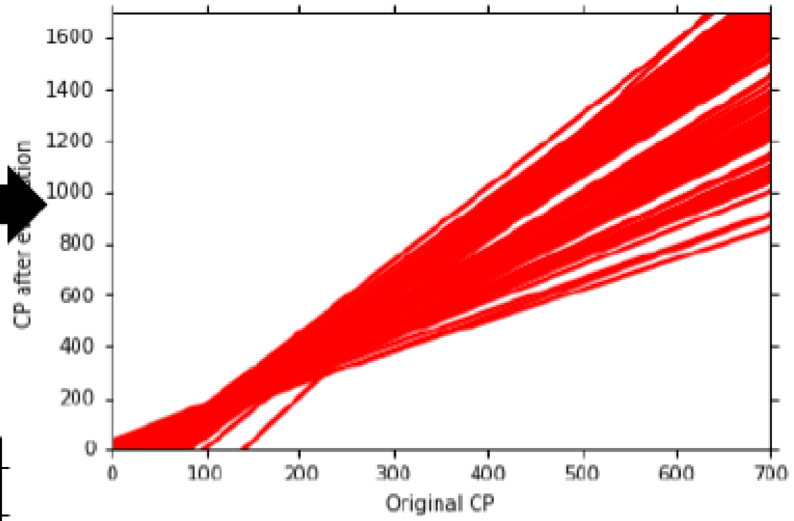
$$y = b + w \cdot x_{cp}$$



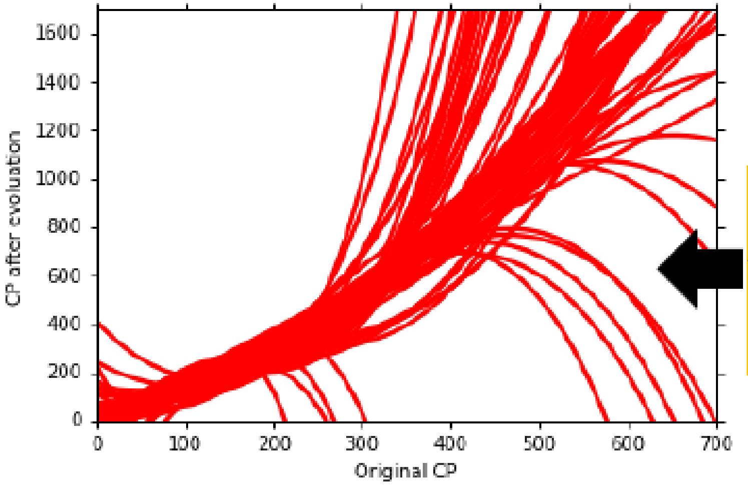
$$y = b + w \cdot x_{cp}$$

f^* in 100 universes

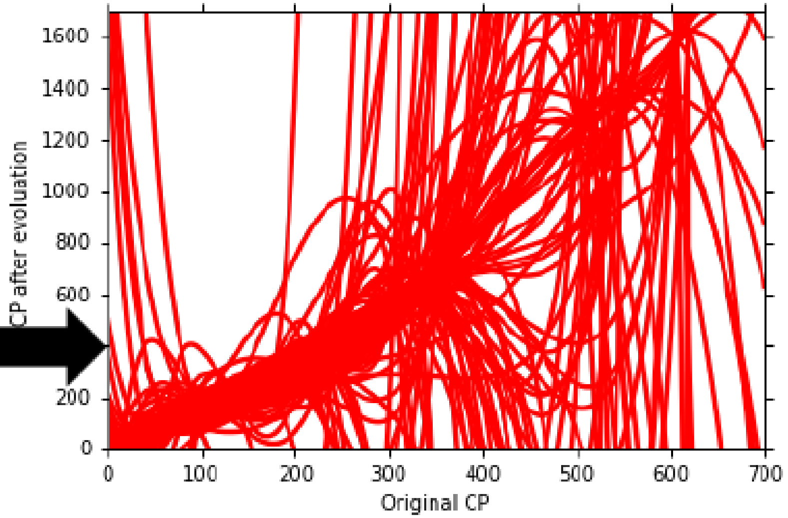
$$y = b + w \cdot x_{cp}$$



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

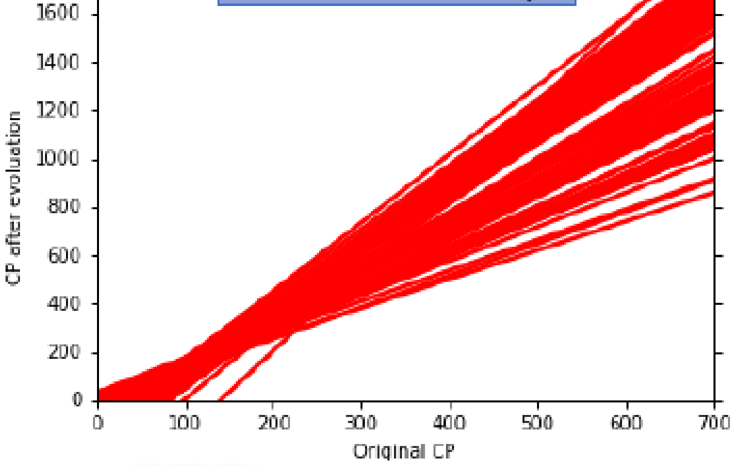


$$y = b + w_1 \cdot x_{cp} + w_2 \cdot \left(x_{cp}\right)^2 + w_3 \cdot \left(x_{cp}\right)^3 + w_4 \cdot \left(x_{cp}\right)^4 + w_5 \cdot \left(x_{cp}\right)^5$$

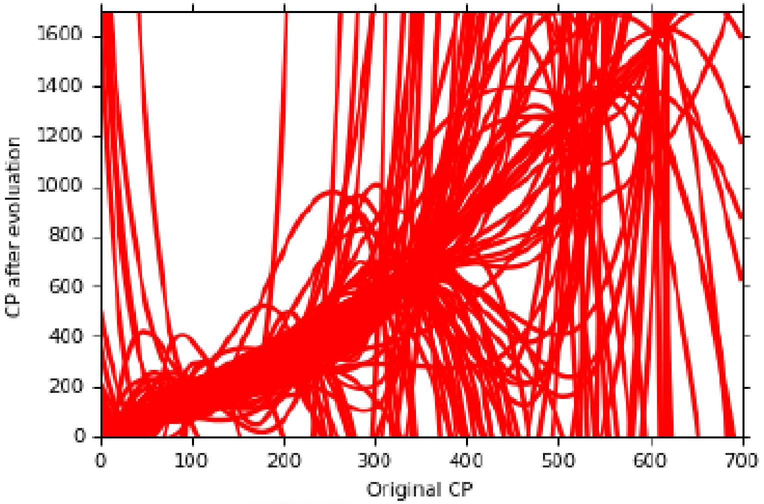


Variance

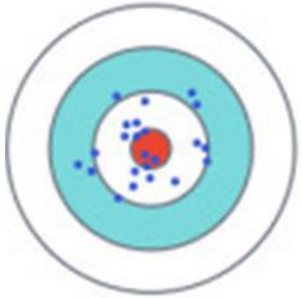
$$y = b + w \cdot x_{cp}$$



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



small variance



large variance

simpler model is less influenced by the sampled data

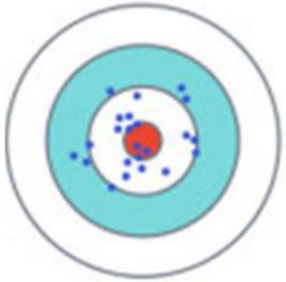
consider the extreme case f(x)=c

Bias

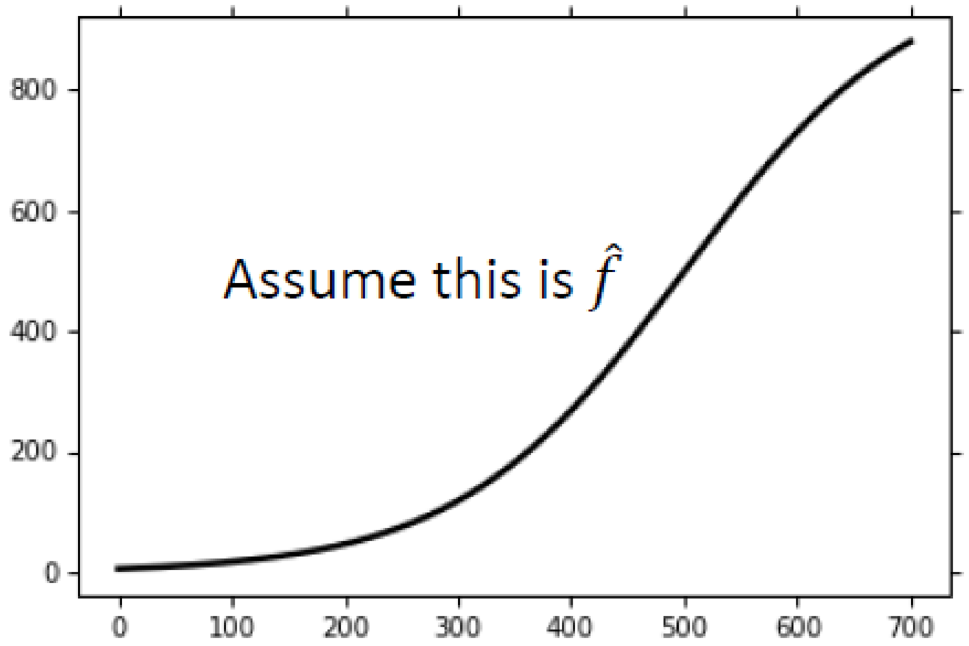
$$E[f^*] = \bar{f}$$

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

Large Bias



Small Bias

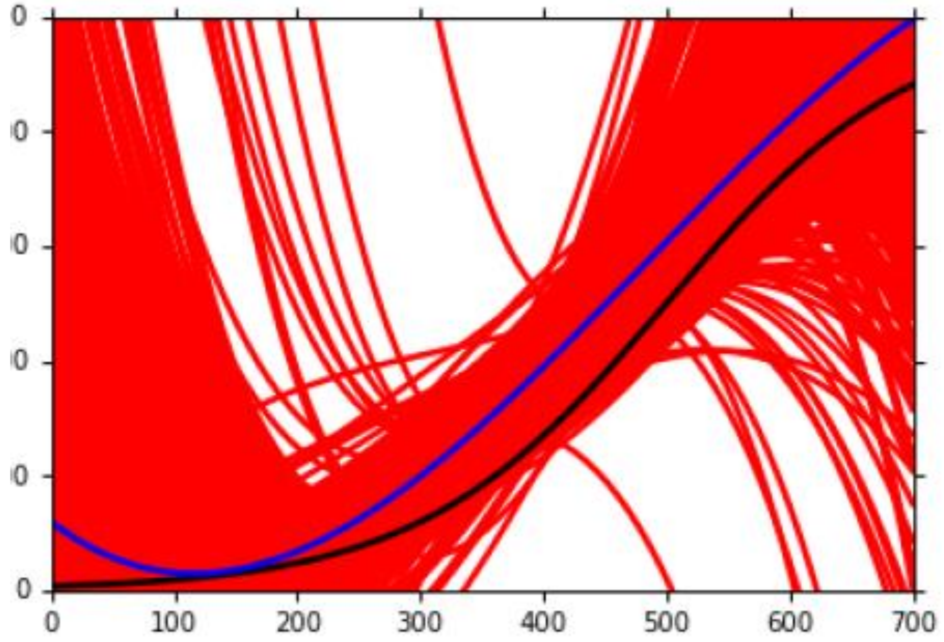


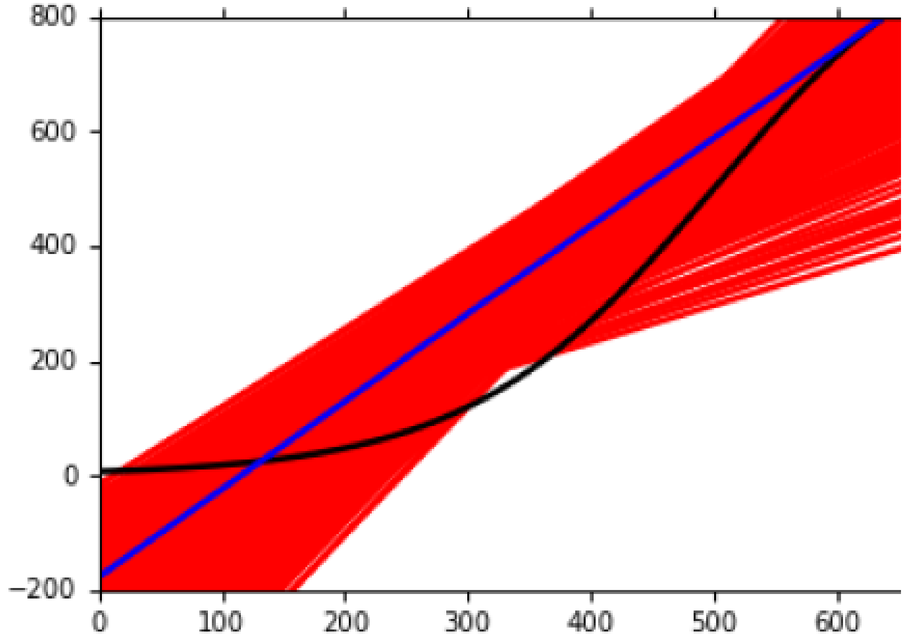
black curve: the true function $\hat{f}$

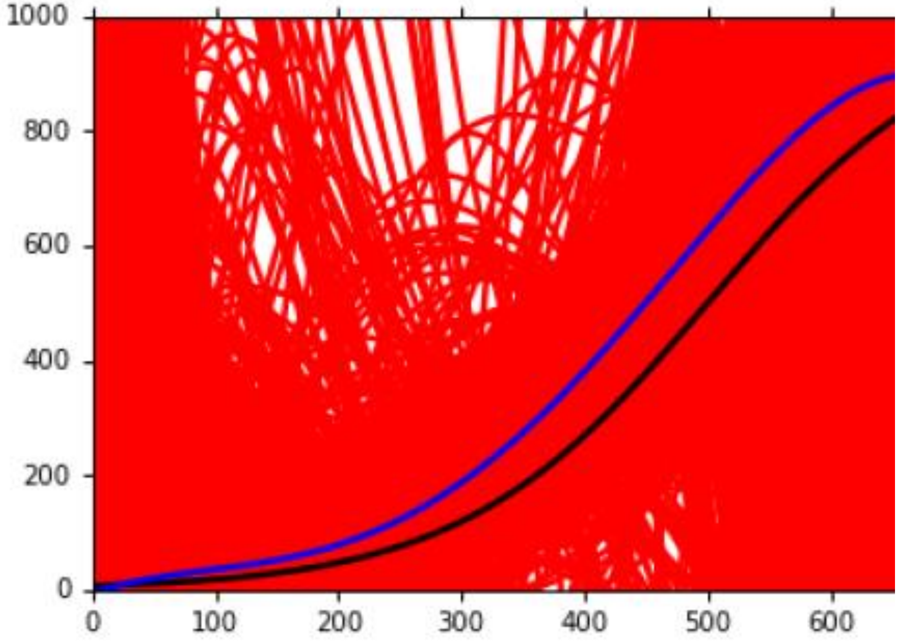
red curve: 5000 $\hat{f}$

blue curve: the average of 5000 f^*

$$= \bar{f}$$

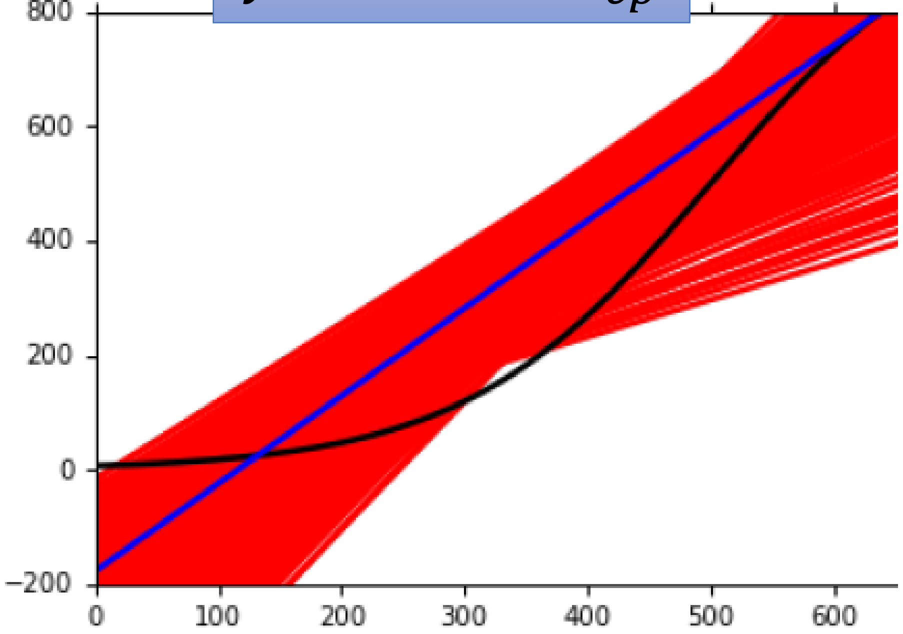


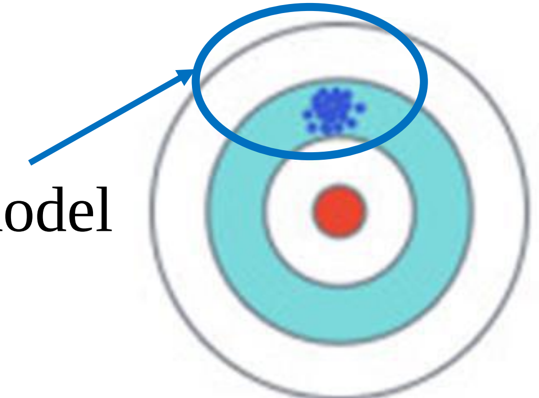




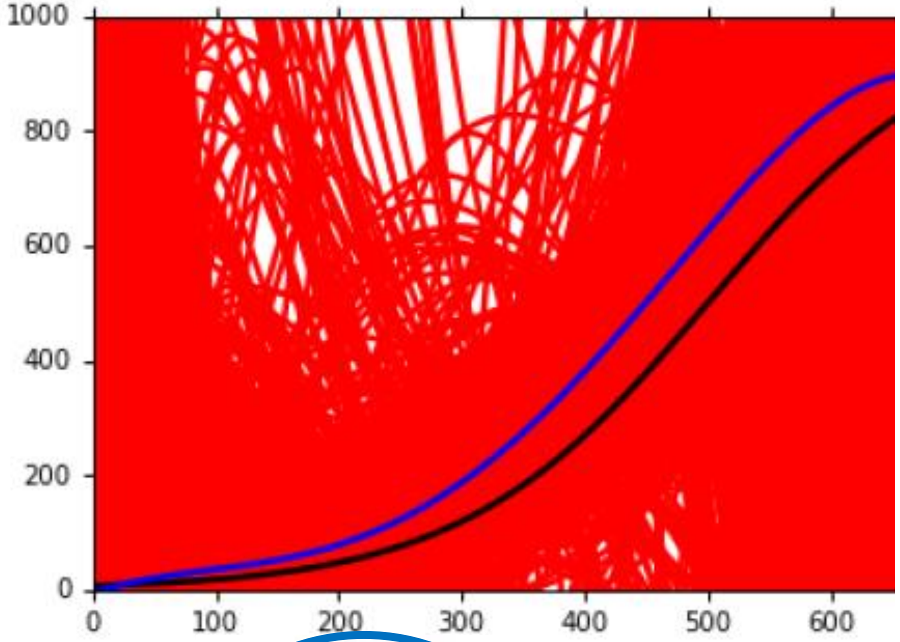
Bias

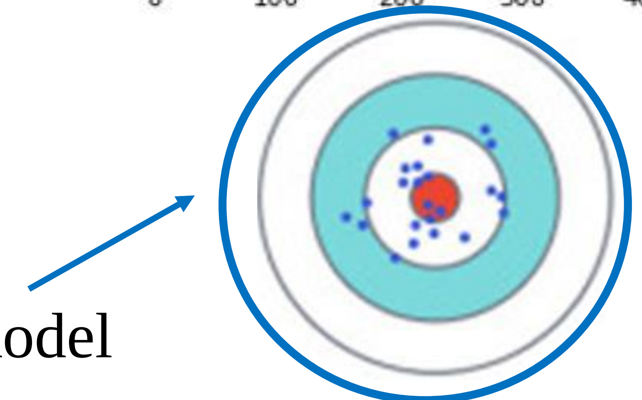
$$y = b + w \cdot x_{cp}$$



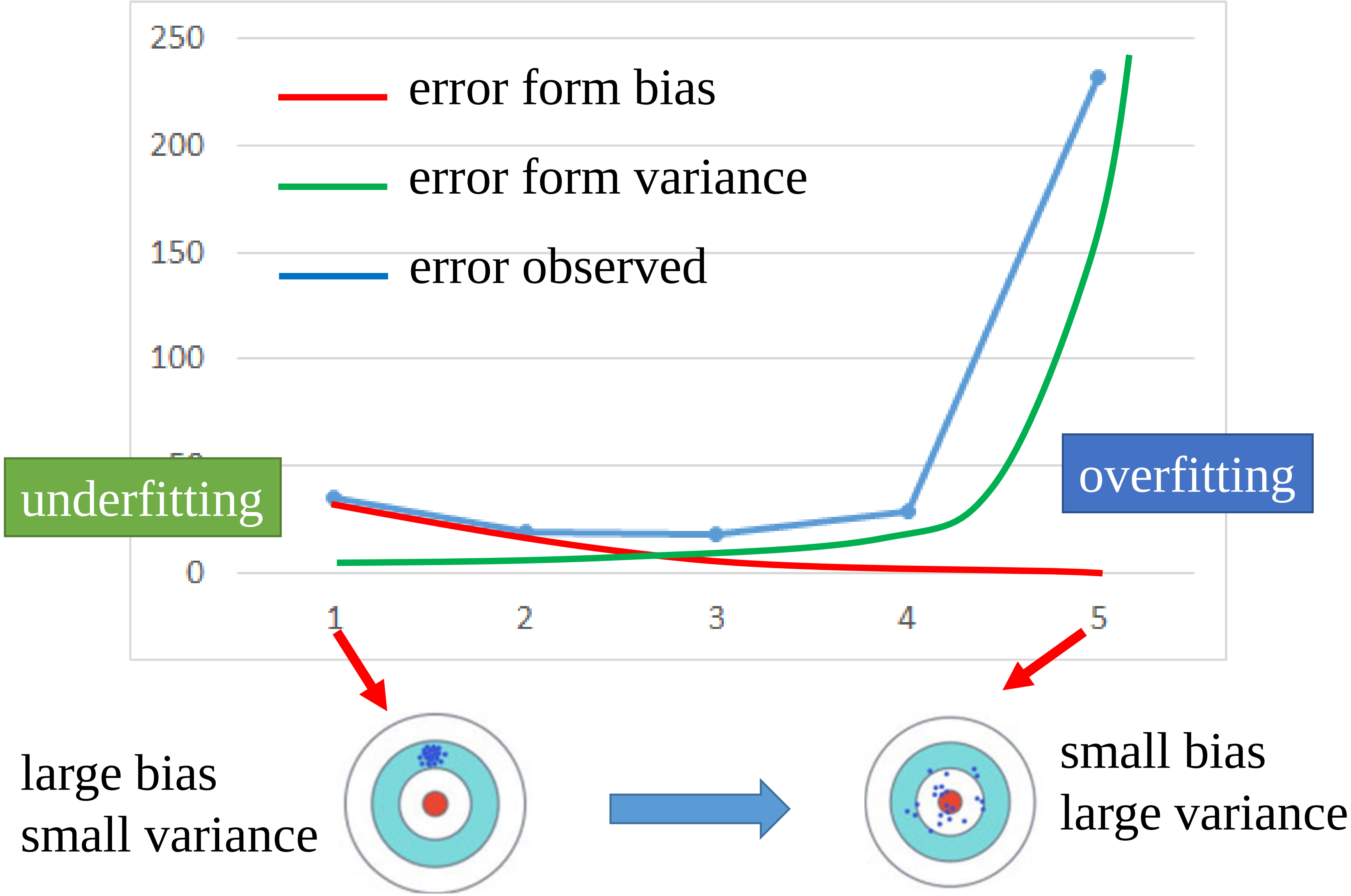


$$y = b + w_1 \cdot x_{cp} + w_2 \cdot \left(x_{cp}\right)^2 + w_3 \cdot \left(x_{cp}\right)^3 + w_4 \cdot \left(x_{cp}\right)^4 + w_5 \cdot \left(x_{cp}\right)^5$$





bias v.s. variance



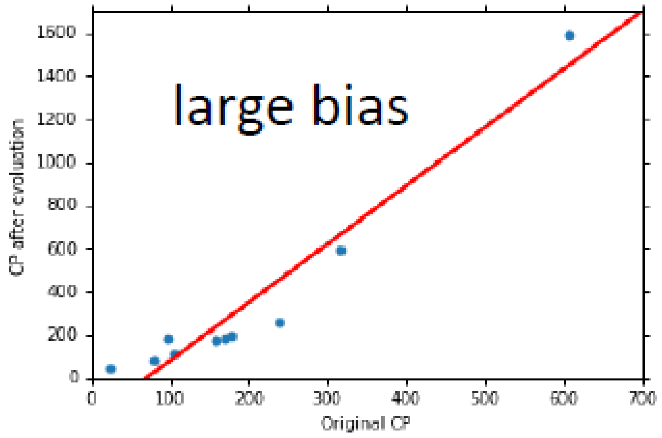
What to do with large bias?

- diagnosis:
  - If your model cannot even fit the training examples, then you have large bias.
  - If you can fit the training data, but large error on testing data, then you probably have large variance

underfitting

overfitting

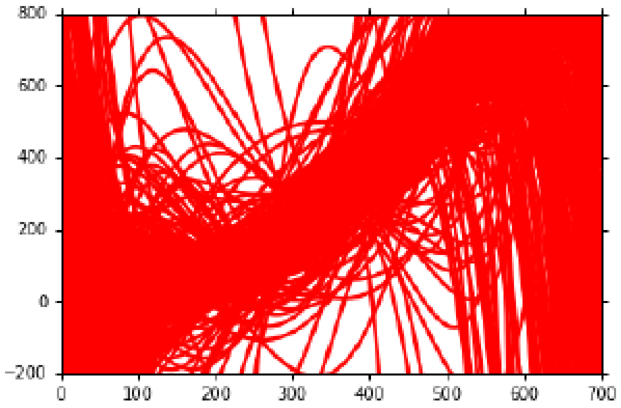
- For bias, redesign your model:
  - add more features as input
  - a more complex model



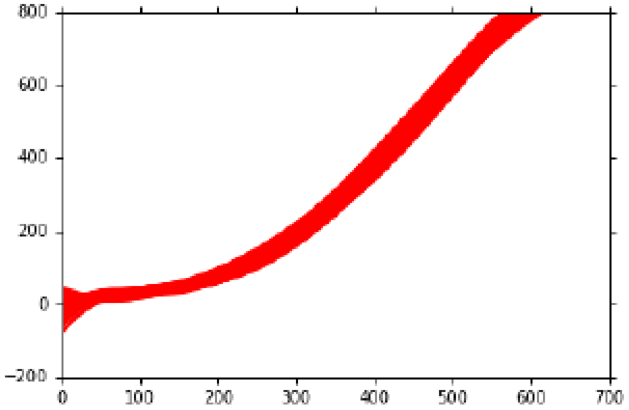
What to do with large variance?

- More data

very effective, but
not always practical

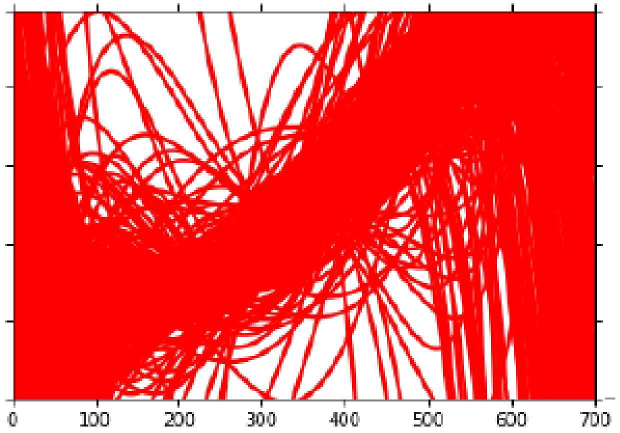


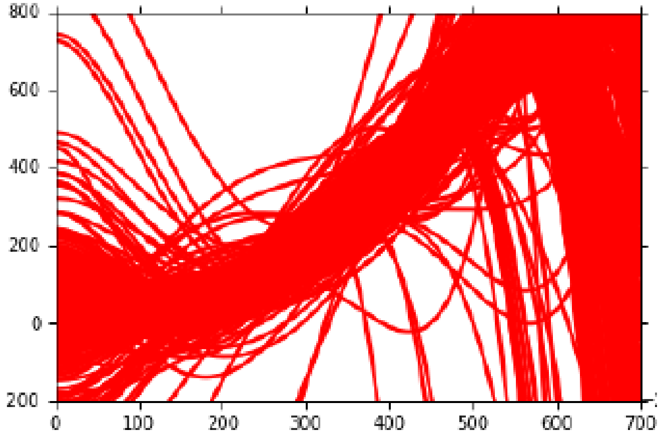
10 examples

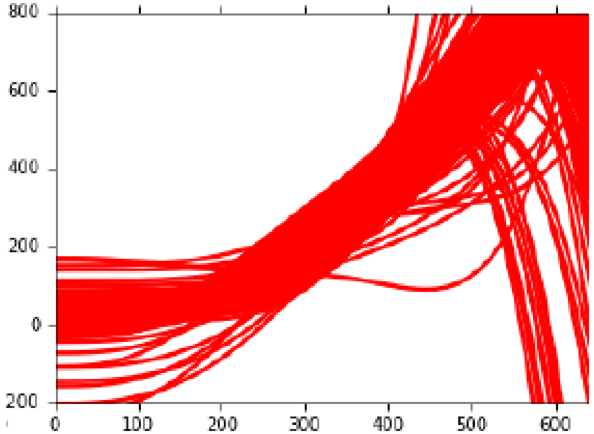


100 examples

- Regularization

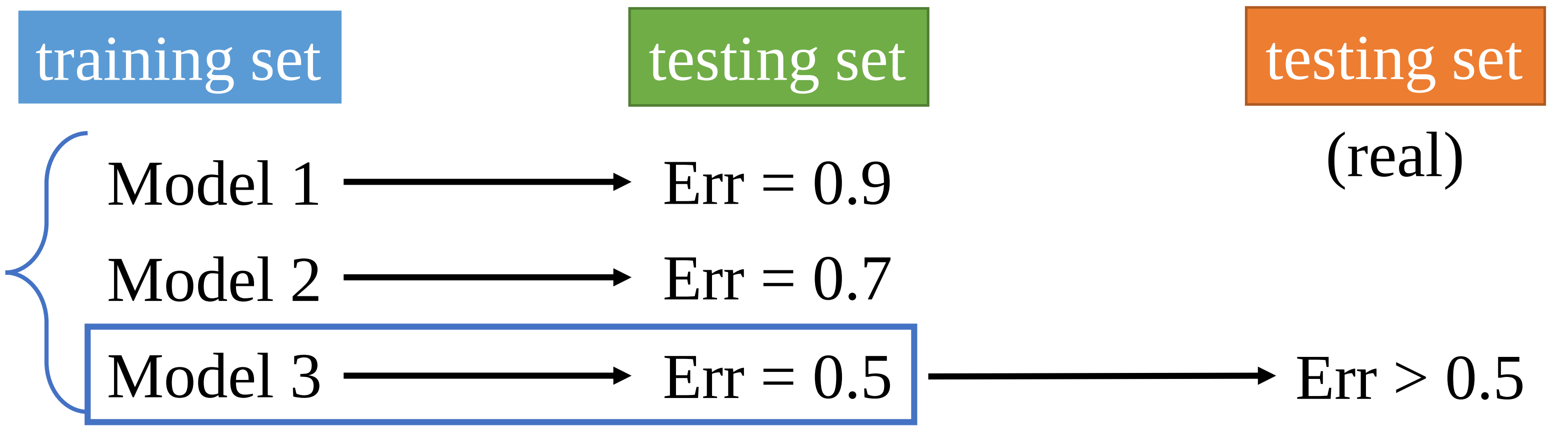






Model Selection

- 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:



Kaggle

public | private

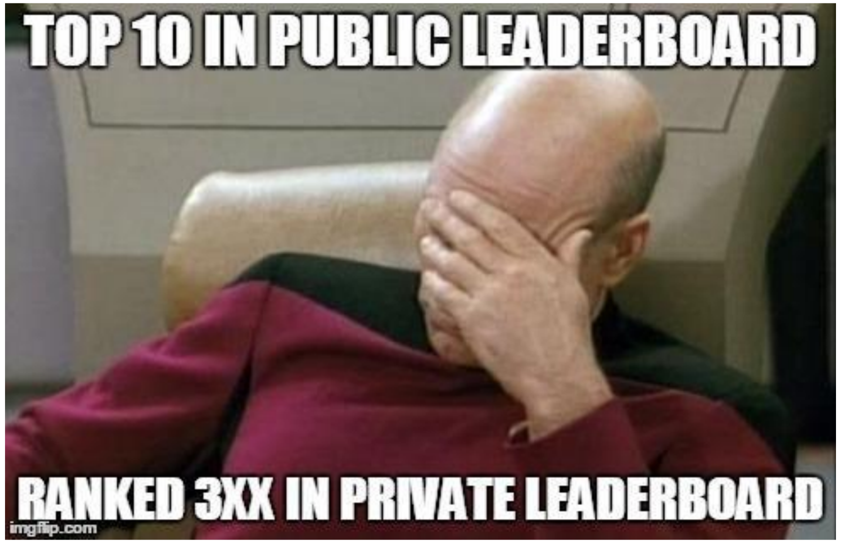
training set | testing set | testing set

Model 1 → Err = 0.9

Model 2 → Err = 0.7

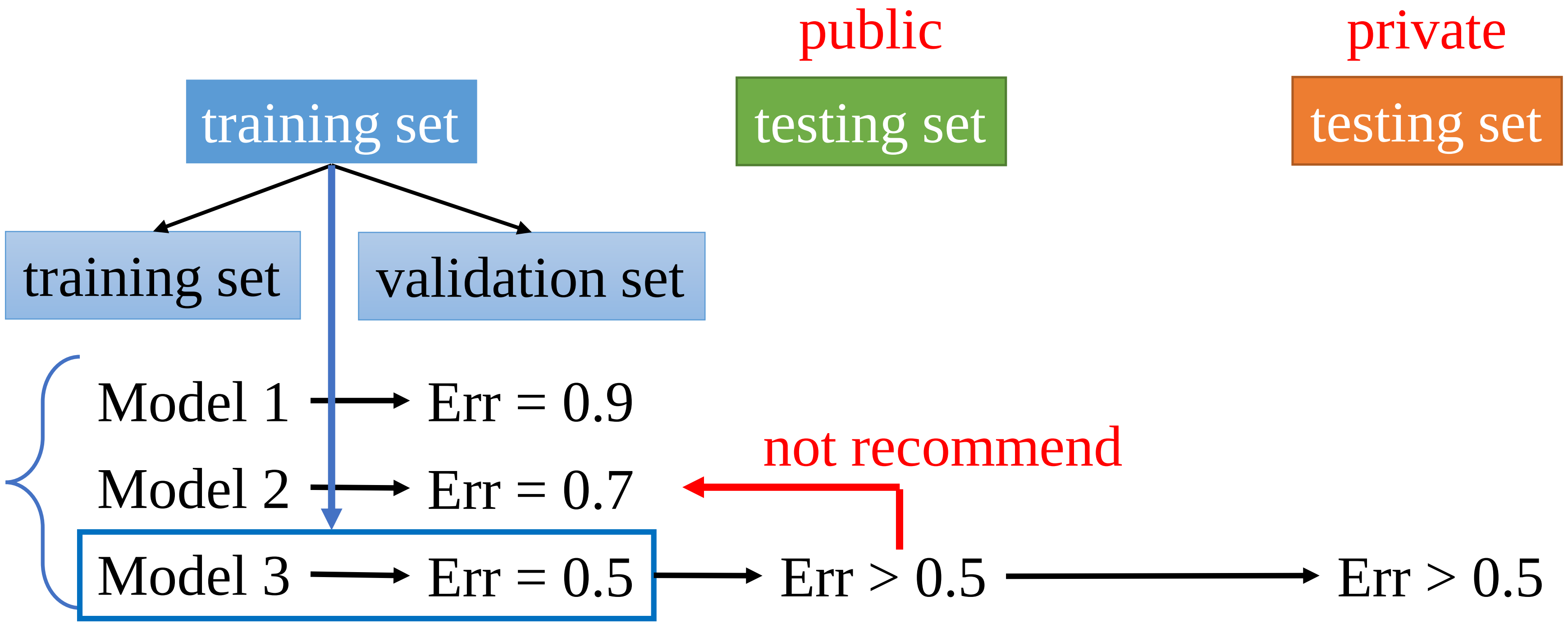
Model 3 → Err = 0.5 → Err > 0.5

I beat baseline! | No, you don’t



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

Cross Validation



using the results of public testing data to tune your model
you are making public set better than private set.

N-fold Cross Validation