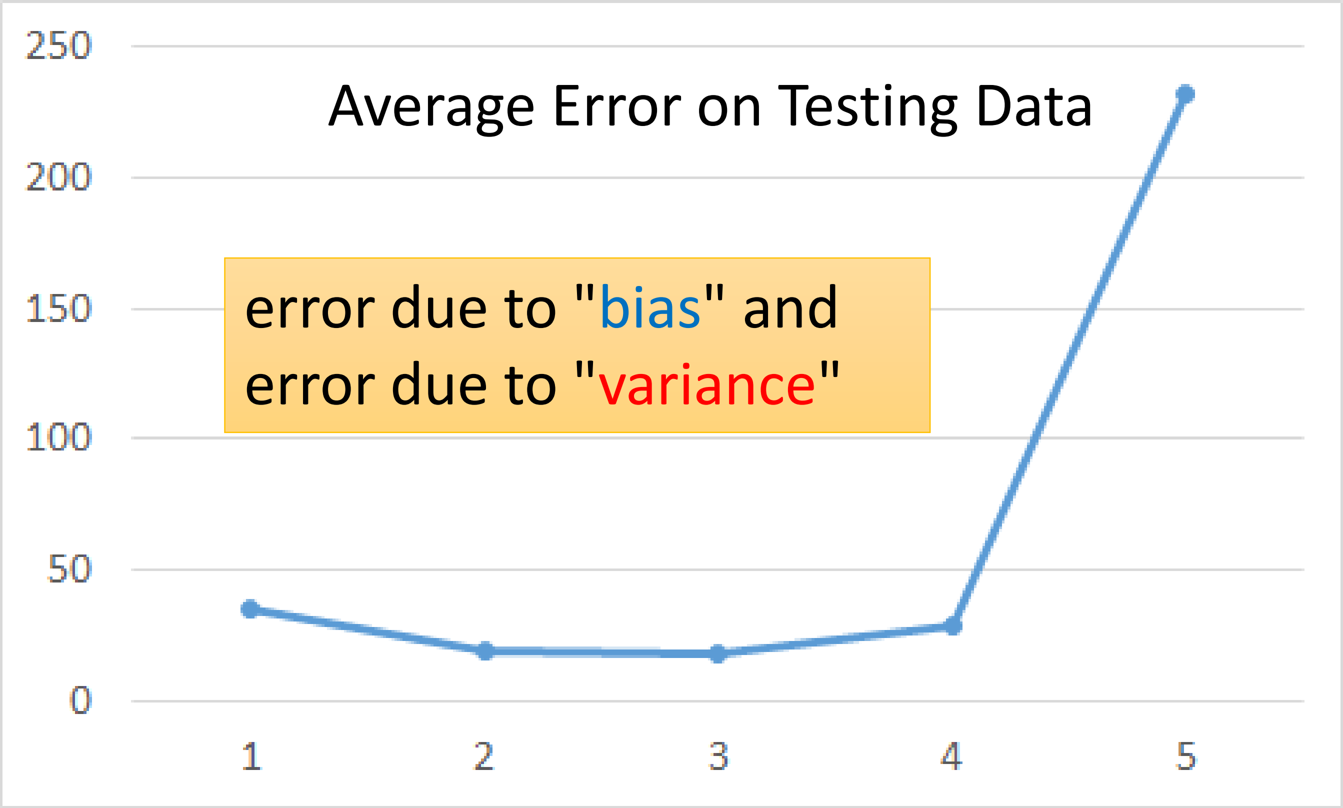
Where does the error come from?

Review



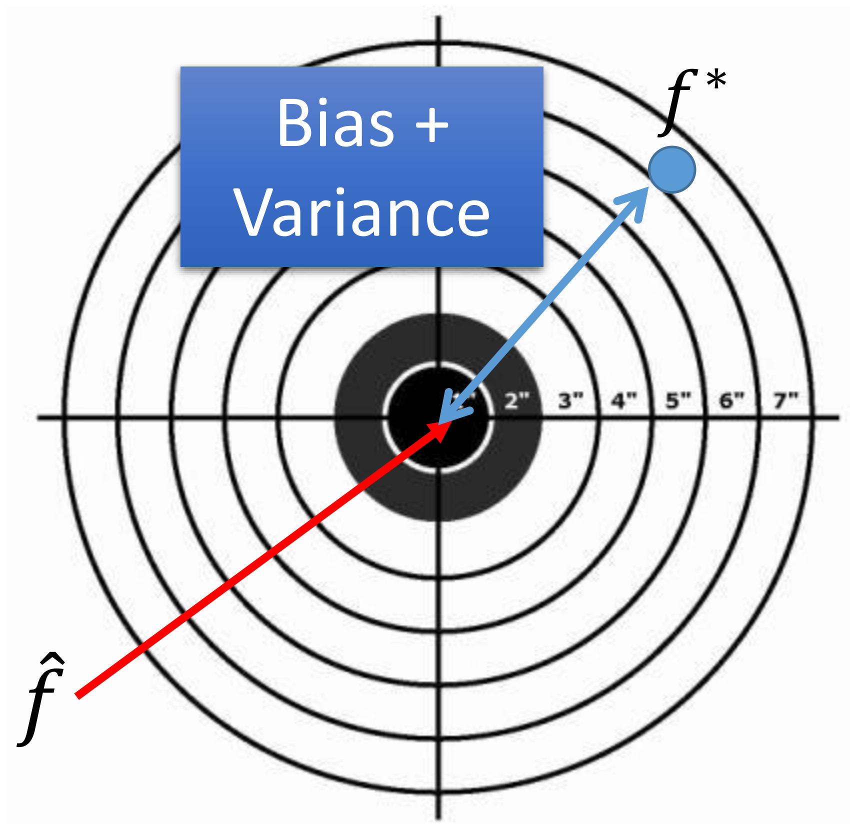
A more complex model does not always lead to better performance on ***testing data***.

f^:表示实际的函数，我们想无限逼近的那个函数
f*：每一次训练所得到的最佳函数
f-：进行多次训练，将所得到的函数取平均得到的函数

Estimator

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

$)$



Only Niantic knows $\hat{f}$

From training data,
we find f^*

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

f*是f^的估计量，估计函数/或者说可以通过许多个f*得到f^

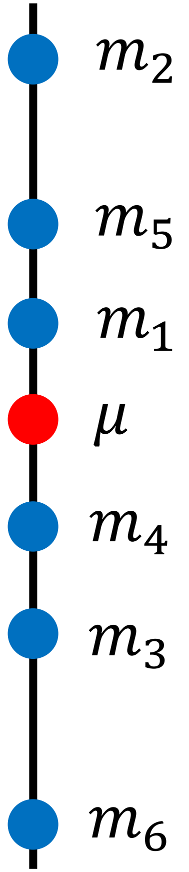
Bias and Variance of Estimator

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

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

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

unbiased



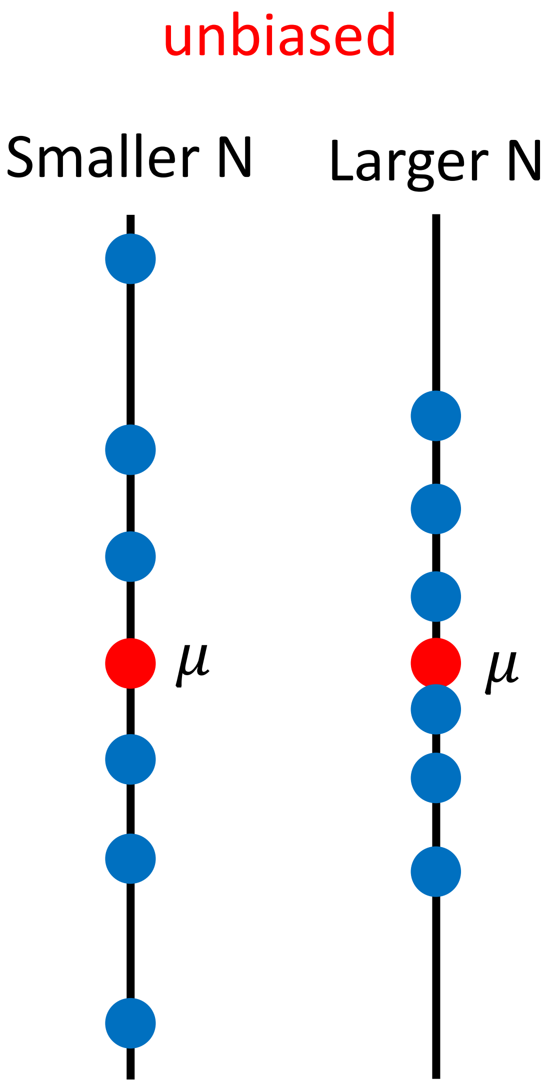
Bias and Variance of Estimator

- 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: $\{x^1, x^2, \ldots, x^N\}$

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

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

Variance depends on the number of samples



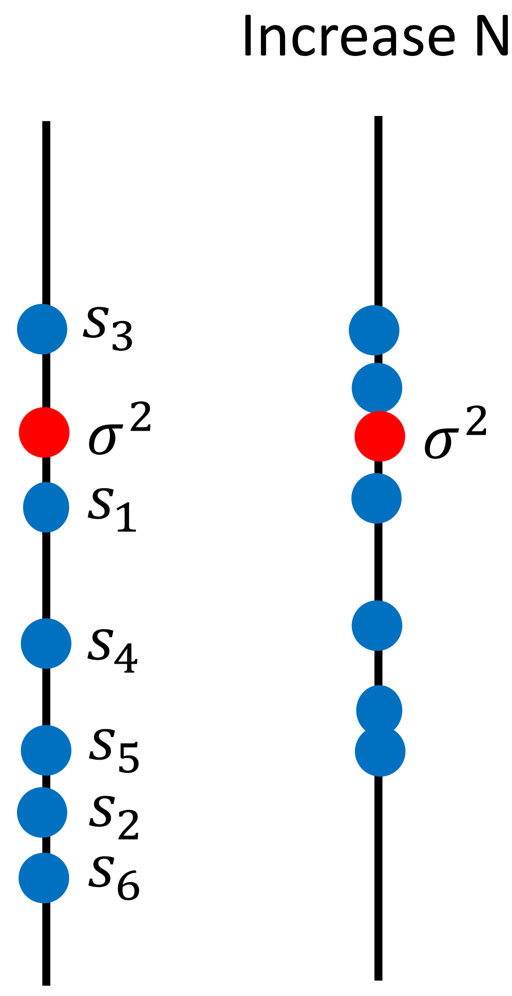
Bias and Variance of Estimator

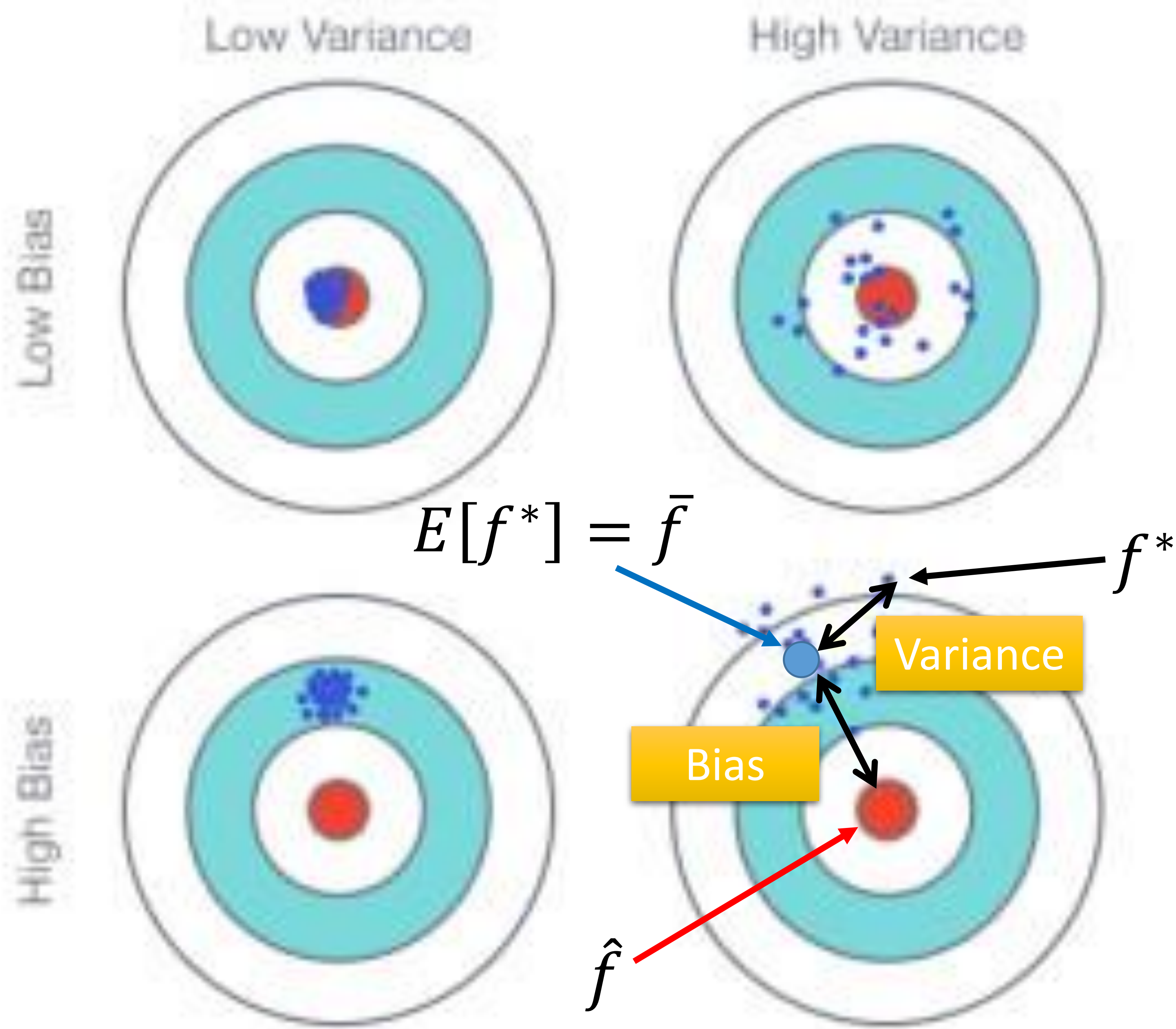
- Estimate the mean of a variable x
  - assume the mean of x is μ
  - assume the variance of x is σ^2
- Estimator of variance σ^2
  - Sample N points: $\{x^1, x^2, \ldots, x^N\}$

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

Biased estimator

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





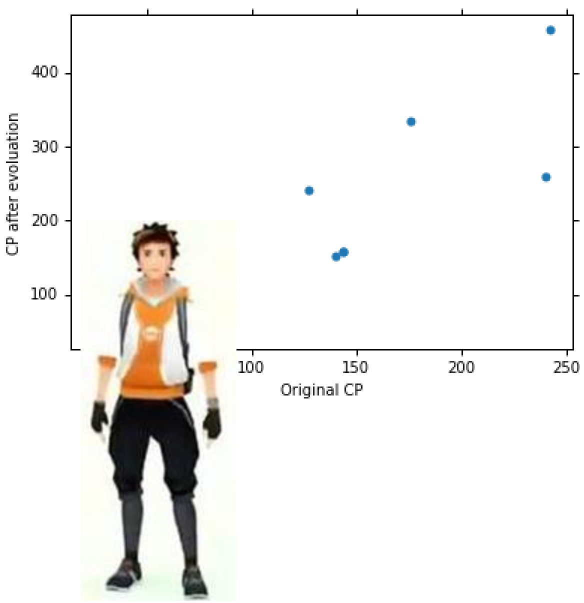
Low Variance
High Variance
Low Bias
High Bias
$E[f^*] = \bar{f}$
f^*
Variance
Bias
$\hat{f}$

Parallel Universes

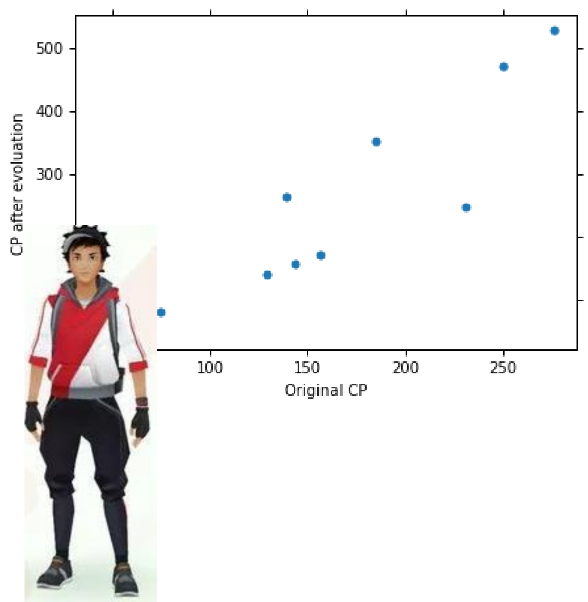
- In all the universes, we are collecting (catching) 10 Pokémons as training data to find f^*

对于同样一个model，用不同的数据训练会得到不同的f*，也就是不一样的函数。

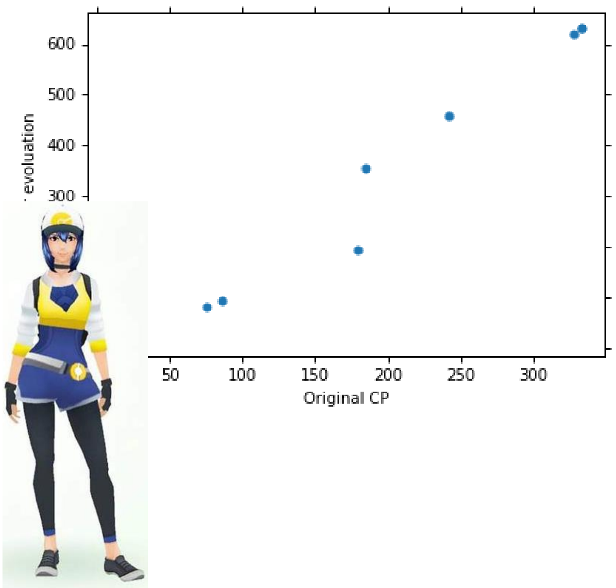
Universe 1



Universe 2



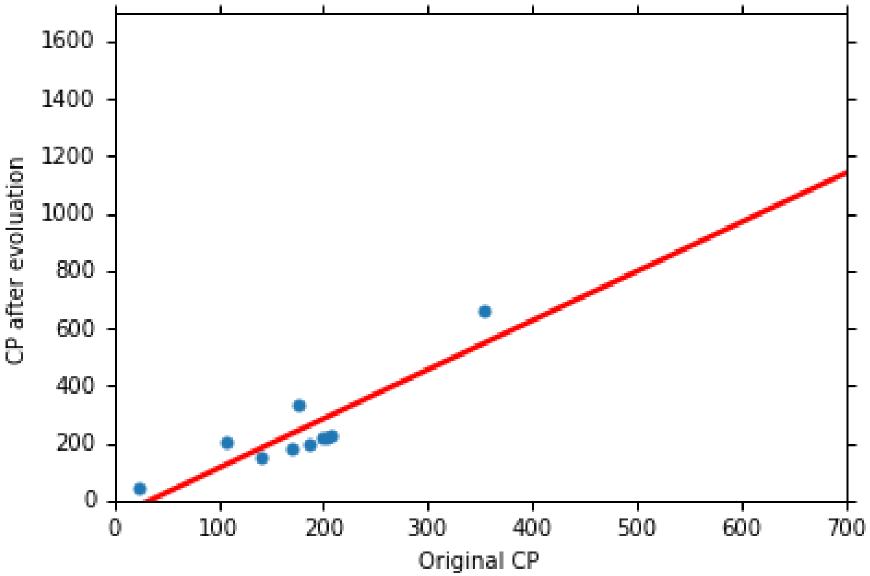
Universe 3



Parallel Universes

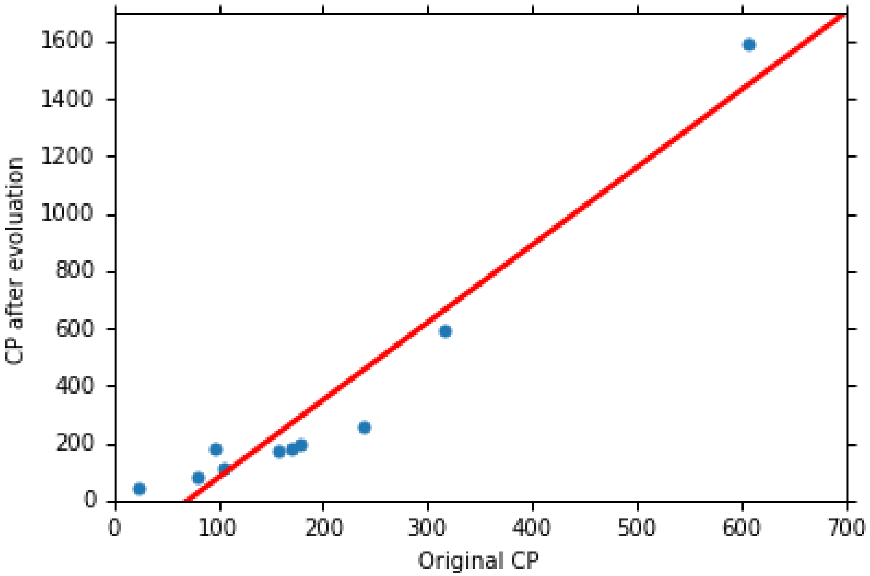
- In different universes, we use the same model, but obtain different f^*

Universe 123



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

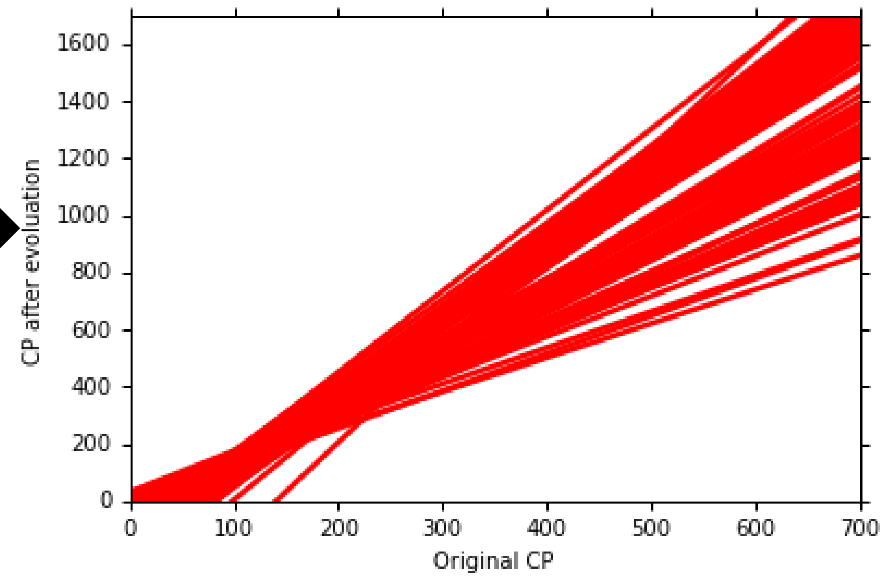
Universe 345



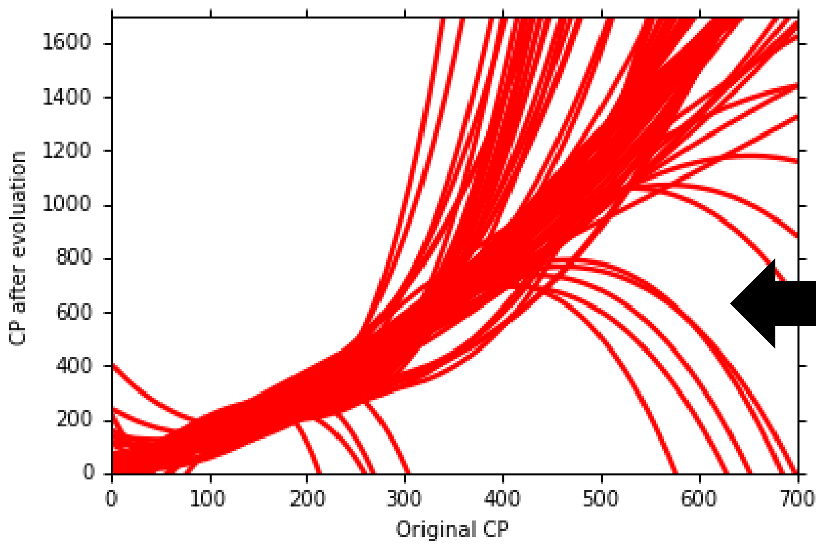
$$y = b + w \cdot x_{cp}$$

f^* in 100 Universes

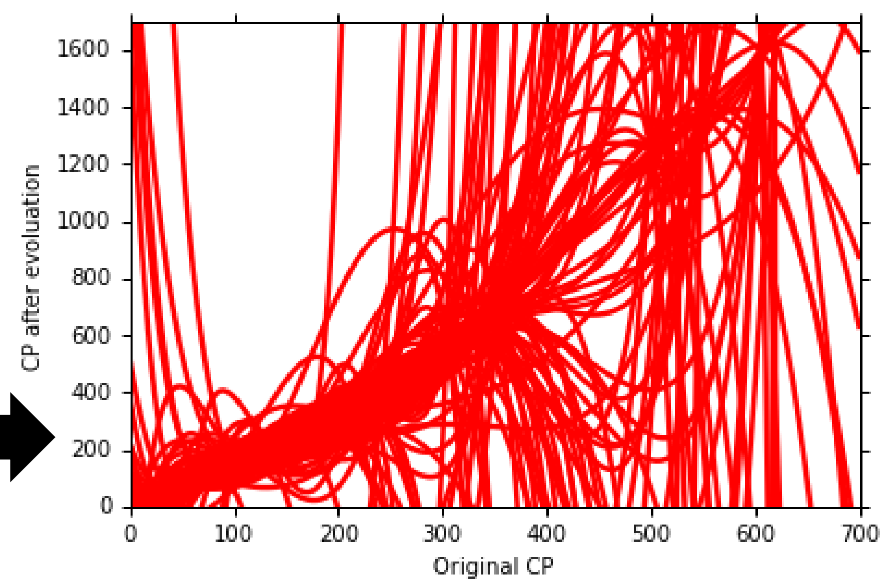
$$y = b + w \cdot x_{cp}$$



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

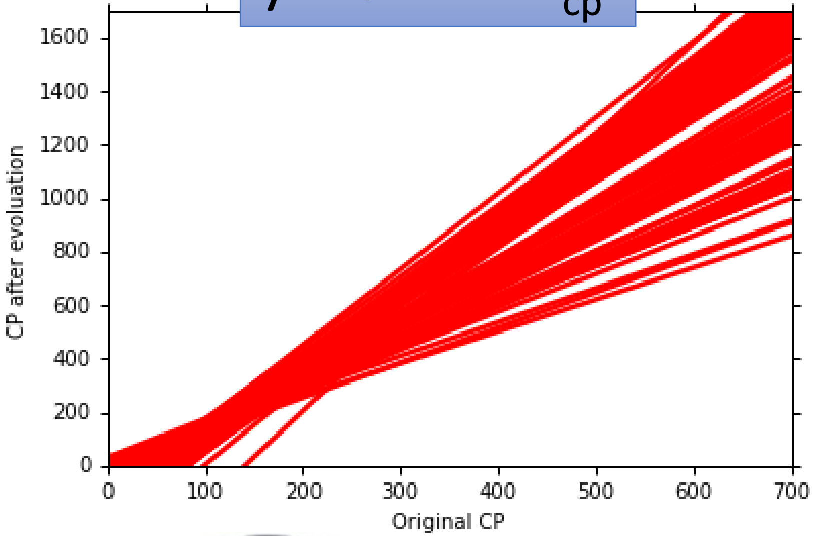


$$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$$

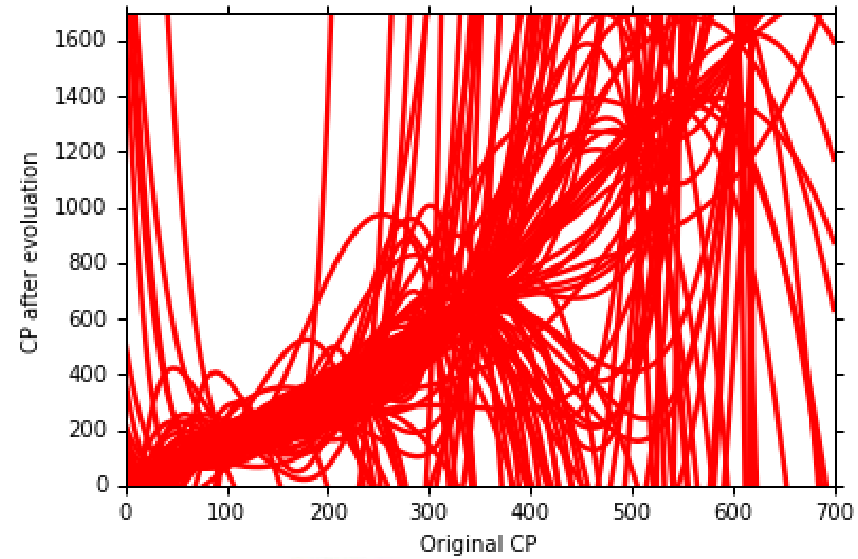


Variance

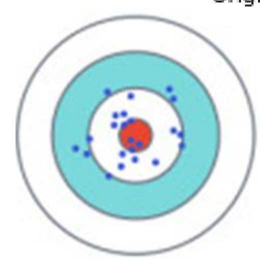
$y = b + w \cdot x_{cp}$



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



Small Variance



Large Variance

Simpler model is less influenced by the sampled data

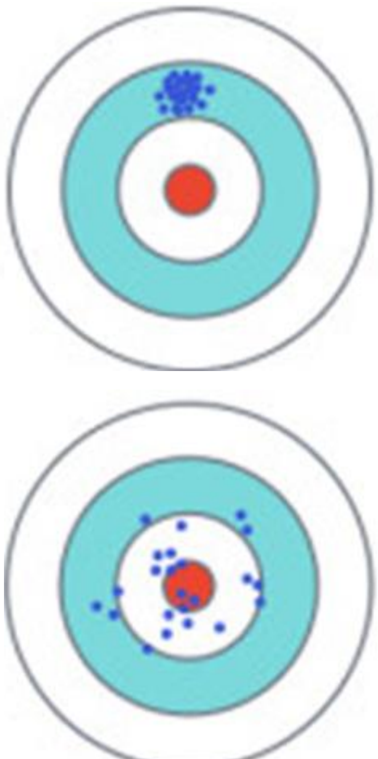
取极端一点，当只有一个变量b时，函数的输出永远是一个常量，与任何的样本都无关。

Consider the extreme case f(x) = 5

Bias

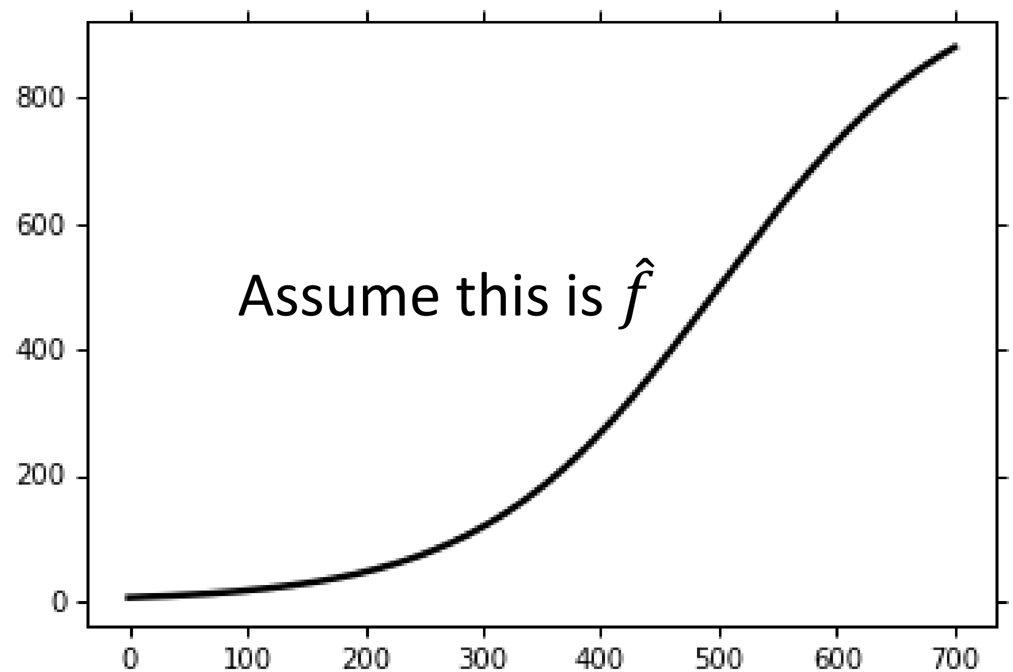
$$E[f^*] = \bar{f}$$

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



Large Bias

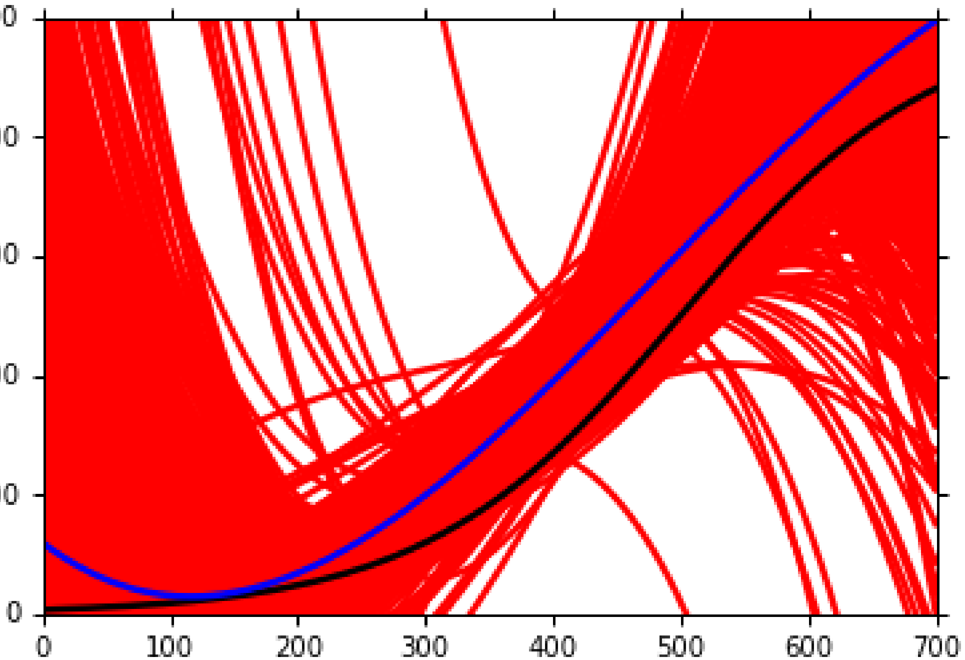
Small Bias

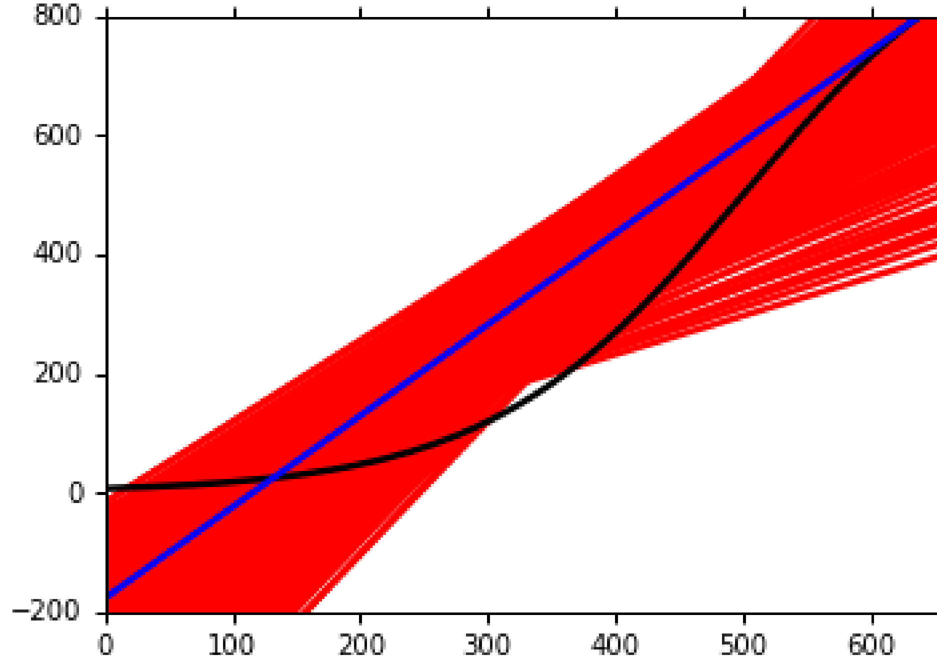


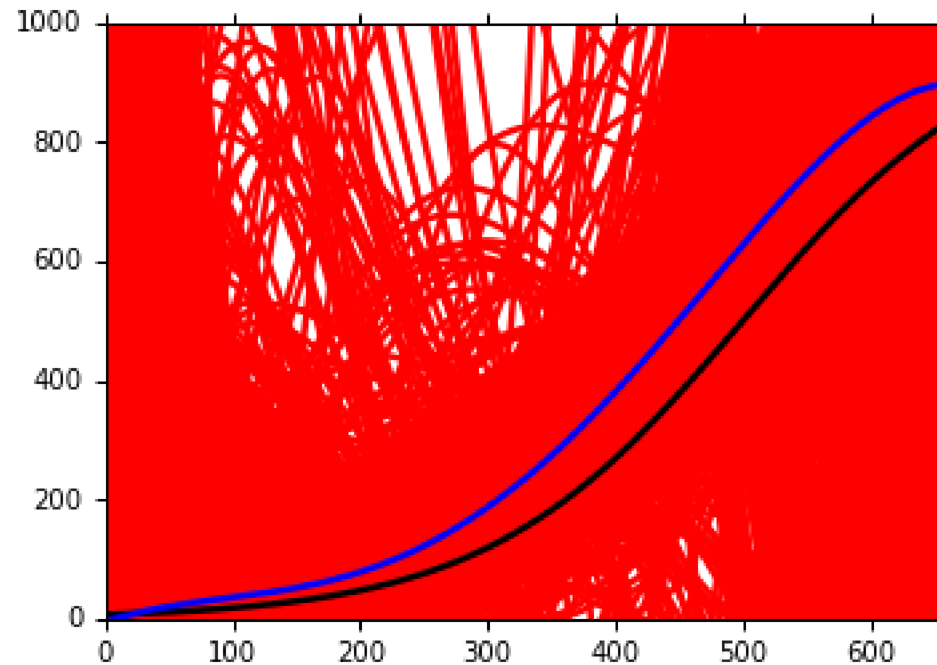
Black curve: the true function $\hat{f}$

Red curves: 5000 f^*

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

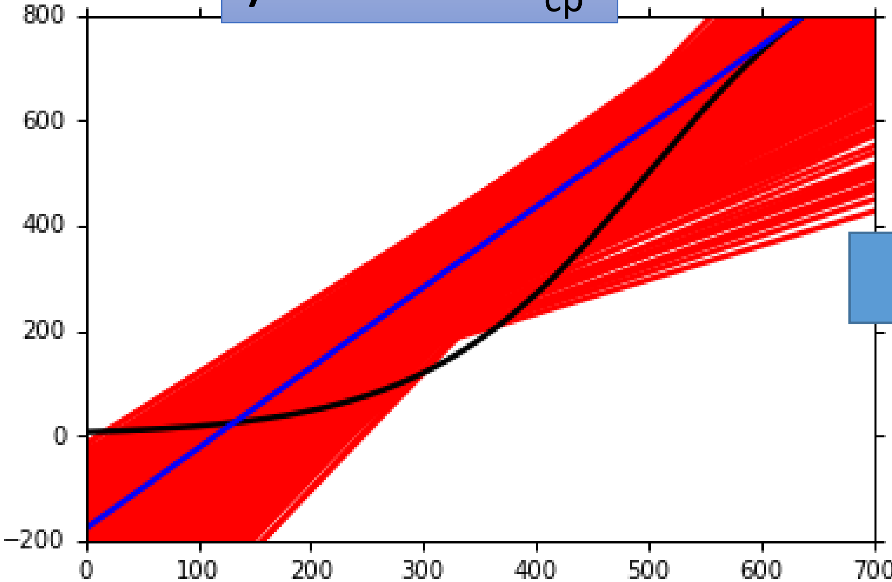




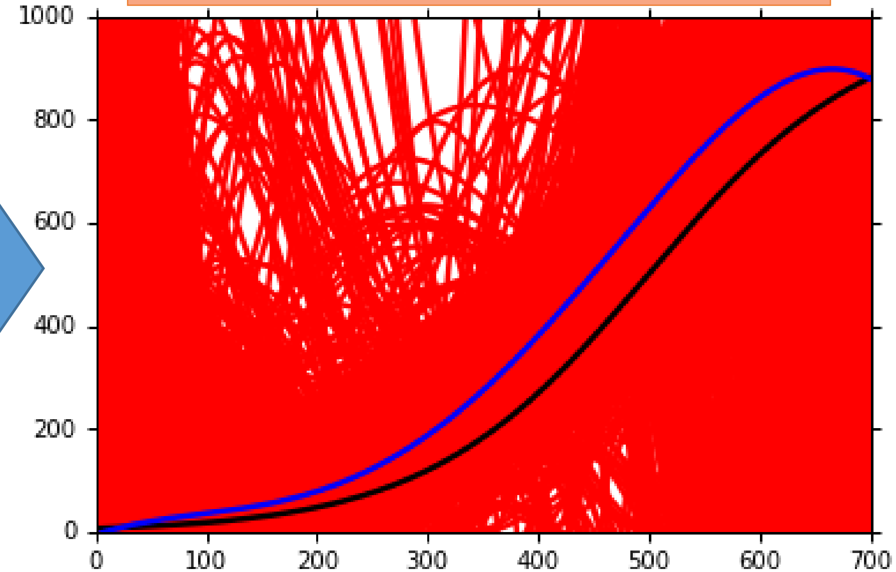


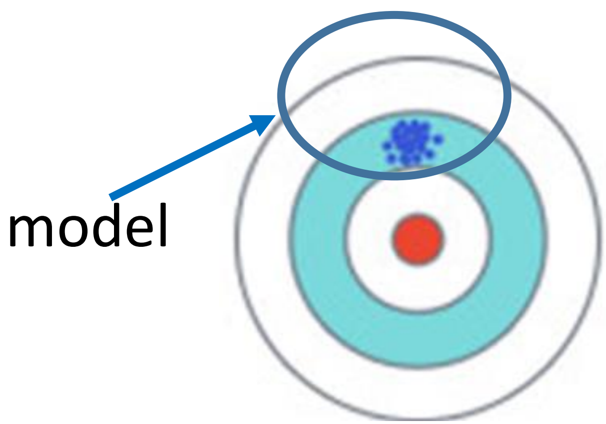
Bias

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

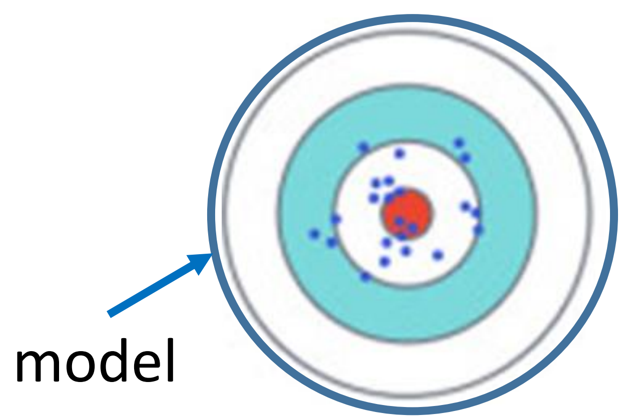


$$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$$



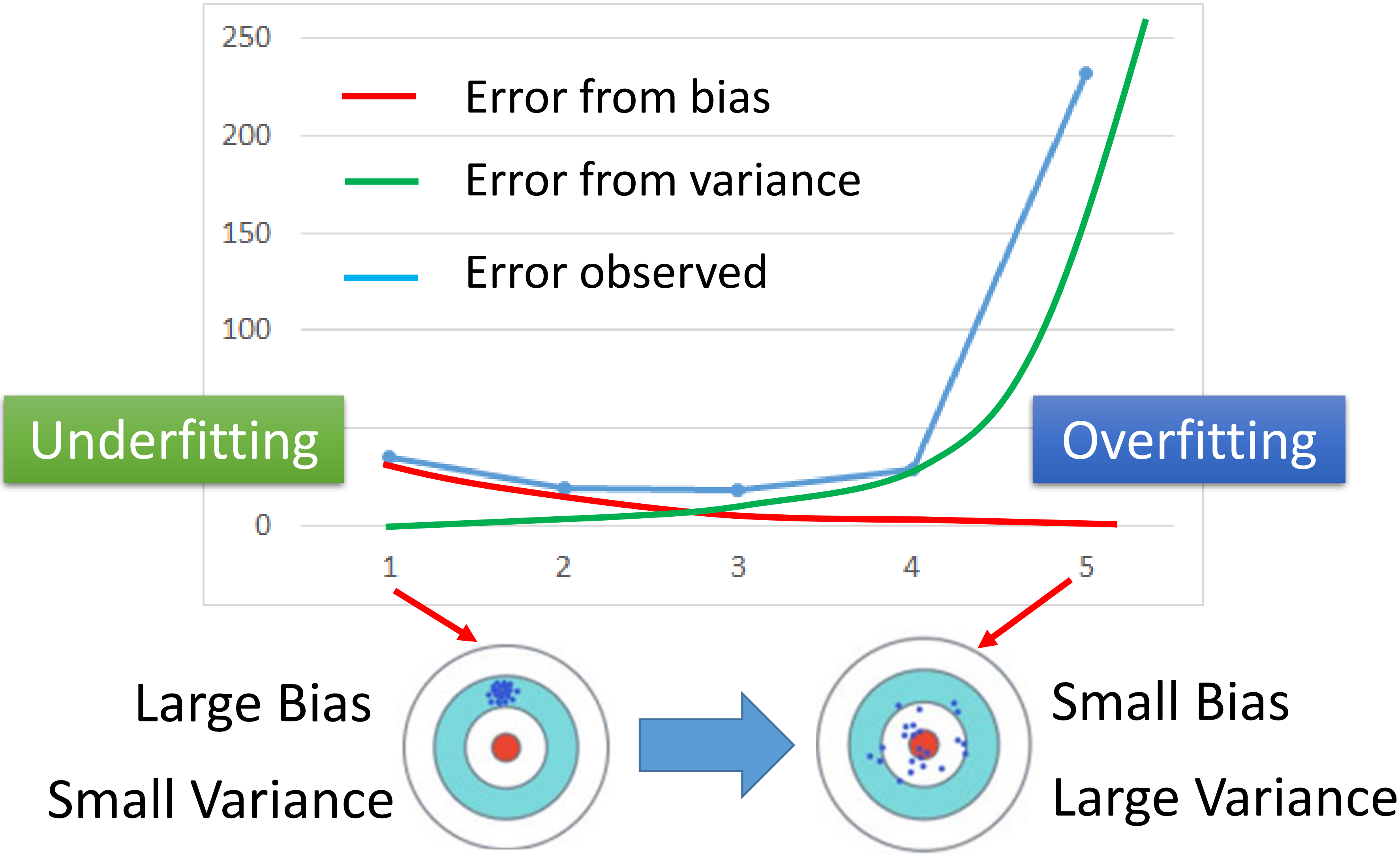


Large Bias



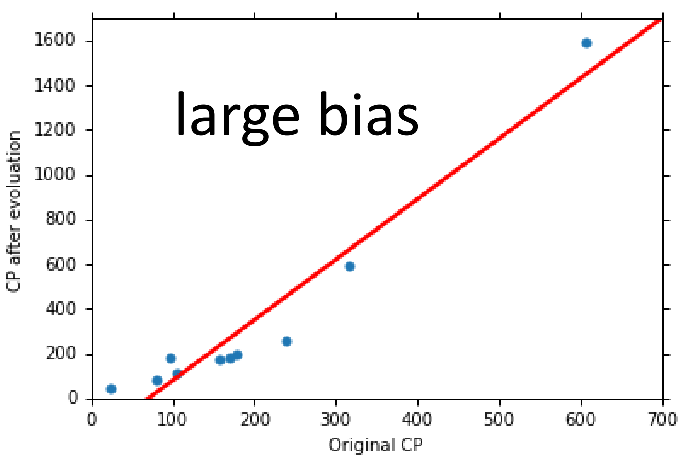
Small Bias

Bias v.s. Variance



What to do with large bias?

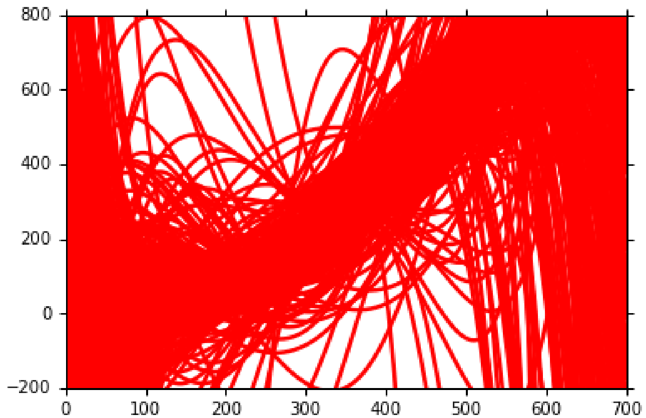
- Diagnosis:
  - If your model cannot even fit the training examples, then you have large bias **Underfitting**
  - If you can fit the training data, but large error on testing data, then you probably have large variance **Overfitting**
- For bias, redesign your model:
  - Add more features as input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



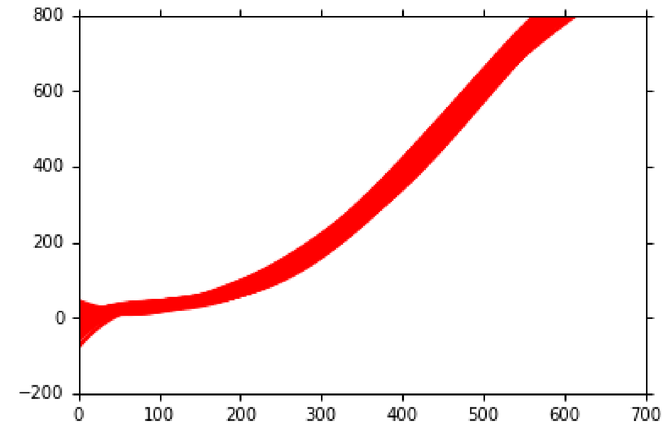
What to do with large variance?

- More data

Very effective, but not always practical



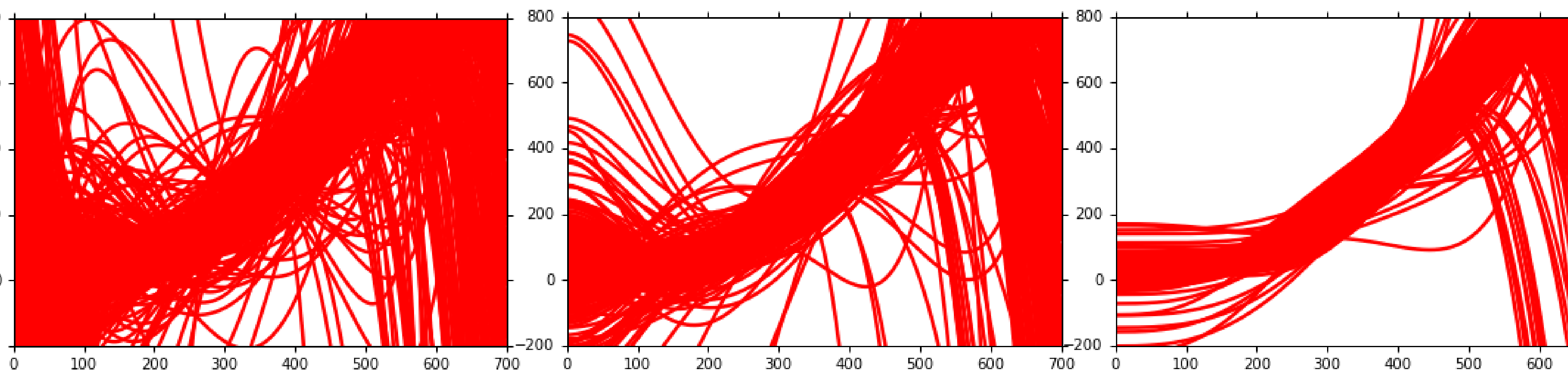
10 examples



100 examples

- Regularization

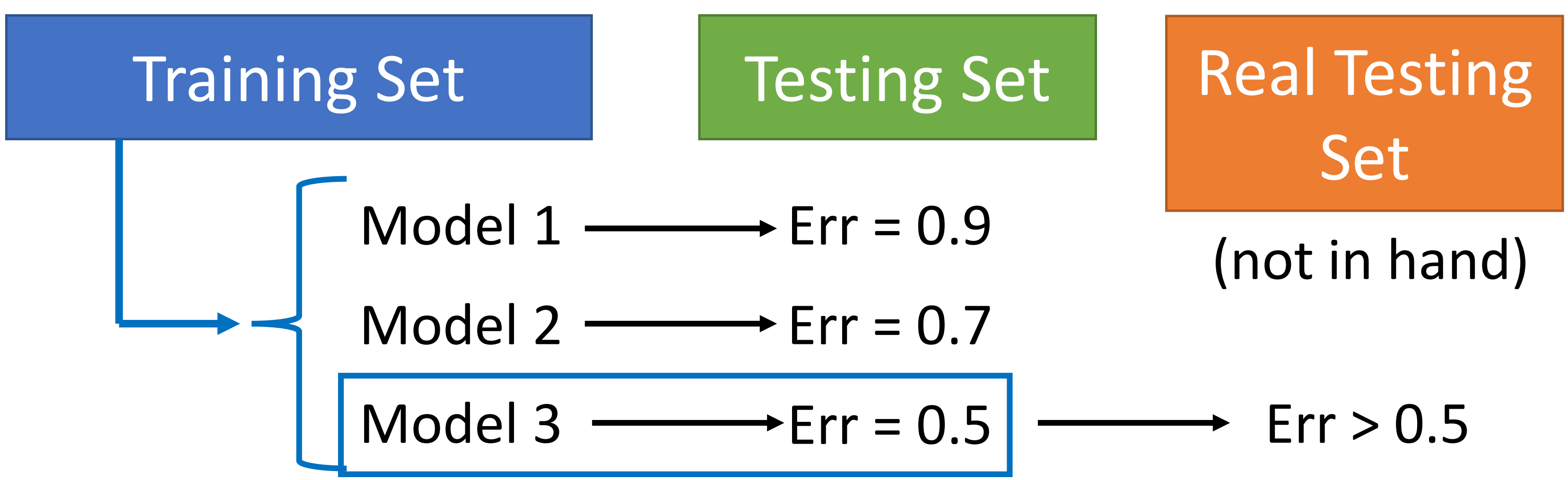
May increase bias



这一部分讲的是如何使用数据集，方法是交叉验证

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

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

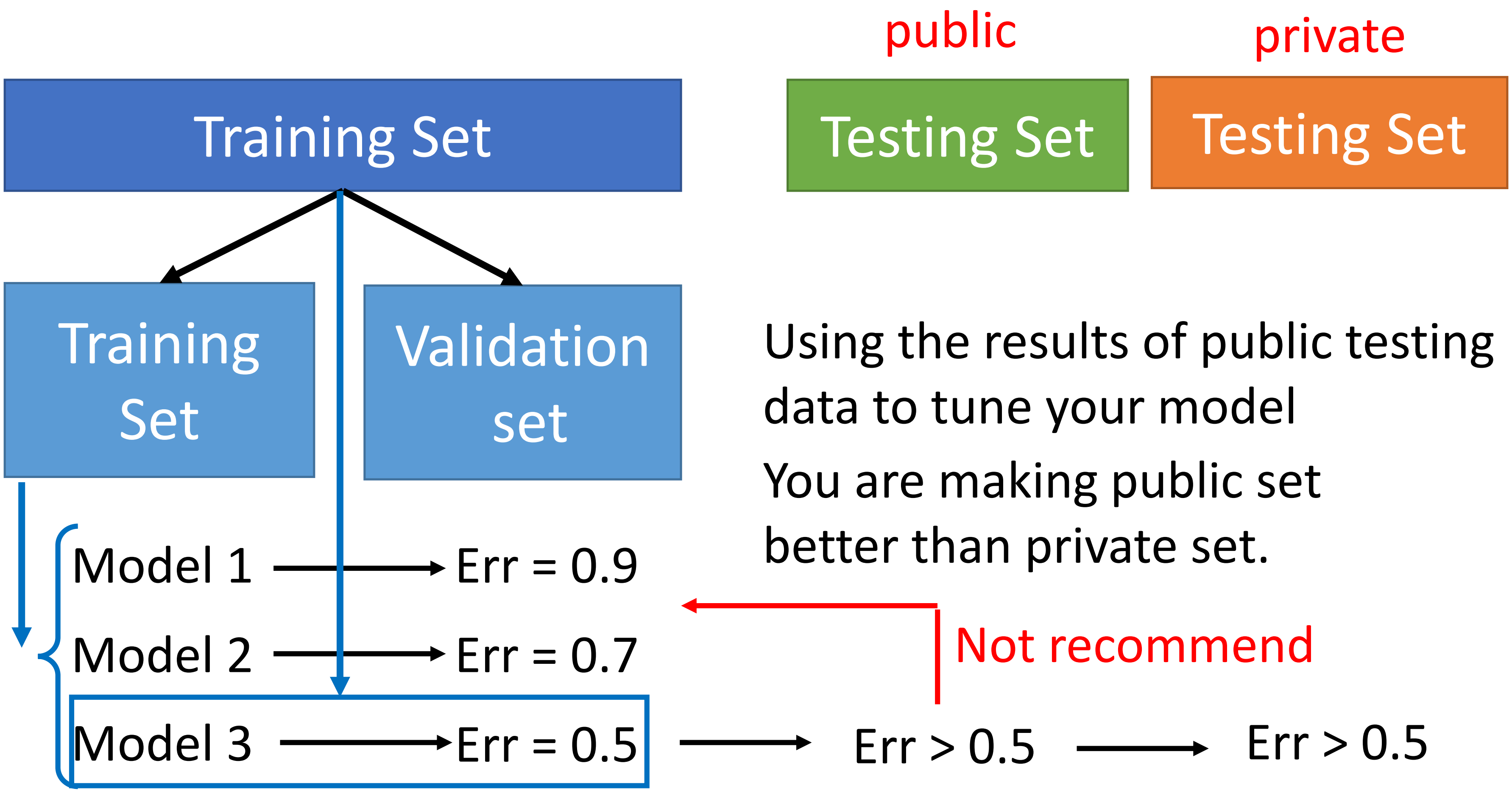
I beat baseline!

No, you don't

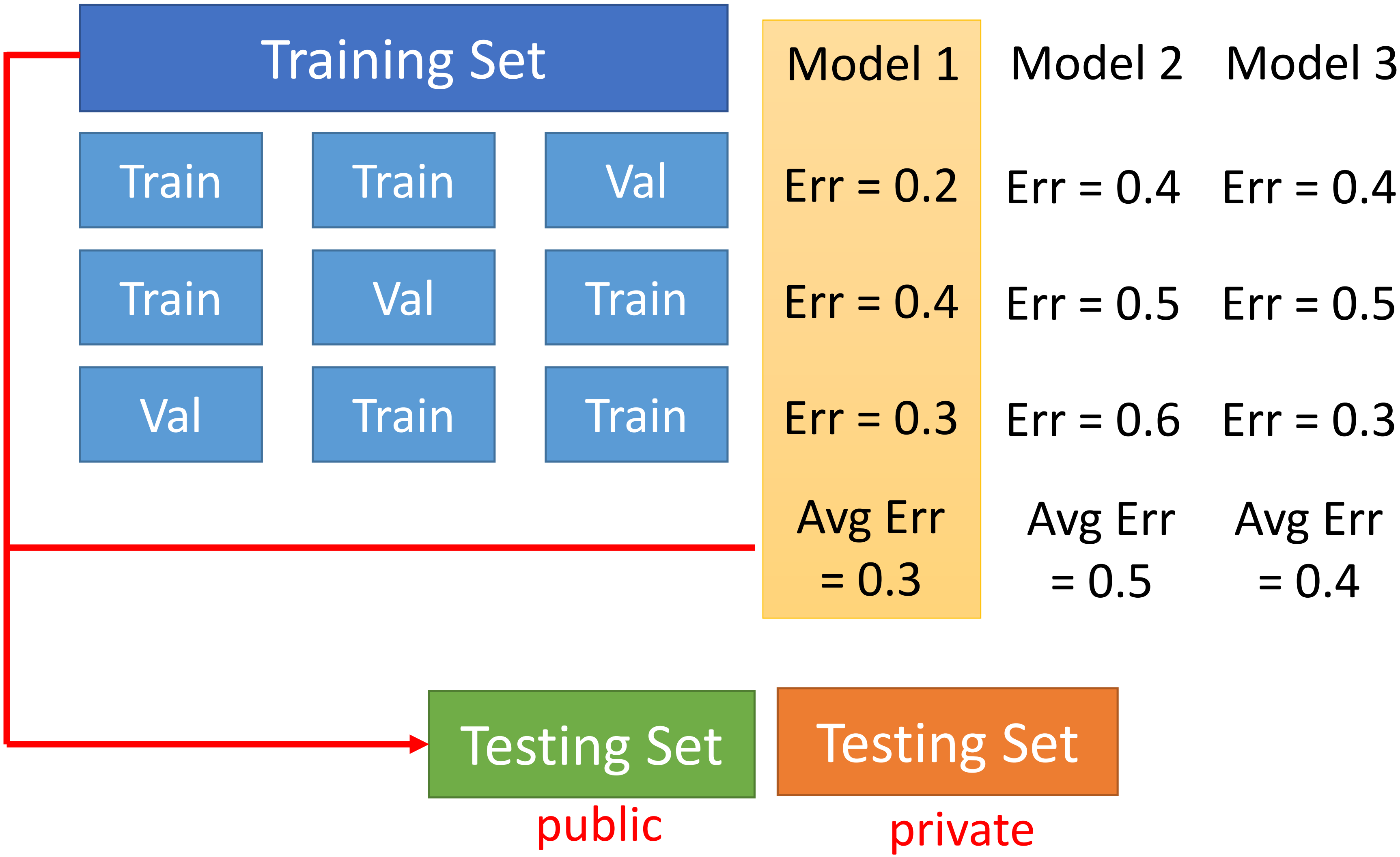
What will happen?

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

个人对于这个PPT的总结：

1 variance（方差）表示的是f*与f^的离散程度。越复杂的函数，f*可选区域更大，它出现的点更分散，也就是离散程度更大，方差也就更大。越简单的函数，f*的可选区域更小，它出现的点更集中，也就是方差更小。

当variance大时，这说明函数过于复杂，也就是说出现 过拟合（oberfitting）：
增加数据量；正则化（正则化会使函数变平滑，可能会增大bias）

2 bias（偏差）表示的是多次实验的f*的中心与f^的距离。bias的大小取决于函数的复杂程度（函数的复杂度来自于特征的数量与特征的次数），函数的复杂程度越低，f*可选的范围就越小，很可能就没有包括f^，所以bias大；当函数复杂程度比较高时，f*的选取范围更大，更有可能包括f^，所以bias小。

当bias大时，这说明函数欠拟合（underfitting）；
重新设计函数：增加特征数，增加特征的次数。

简单的函数，曲线平滑，其bias较大，variance较小；复杂的函数相反。