Where does the error come from?

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

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

- Estimator of variance σ^2
  - Sample N points:$\{x^1, x^2, \dots, x^N\}$

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

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

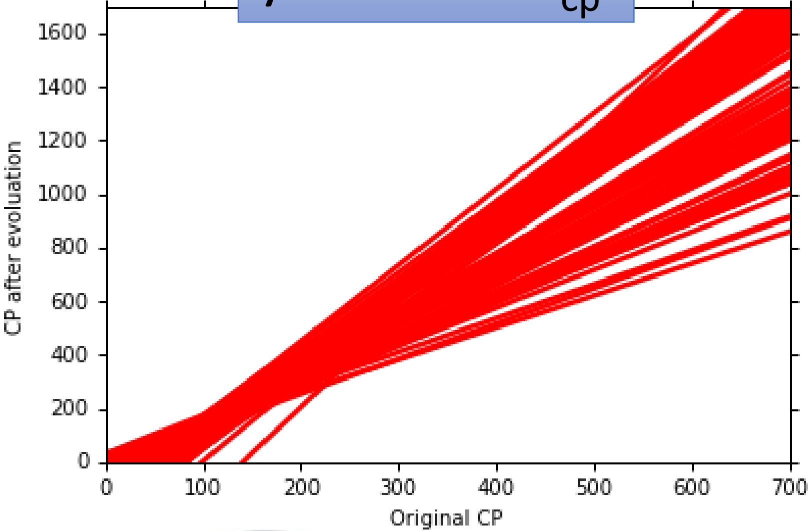
$$Var[m] = \frac{\sigma^2}{N}$$

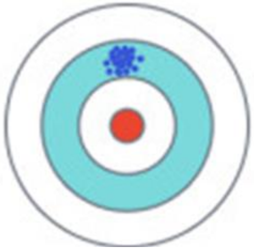
Biased estimator

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

Variance

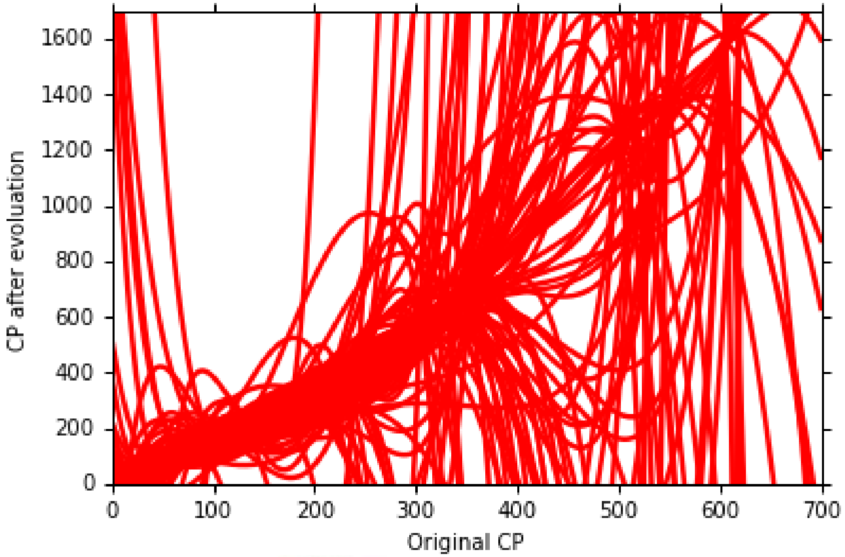
$y = b + w \cdot x_{cp}$

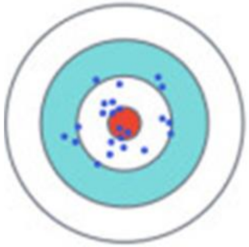




Small Variance

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



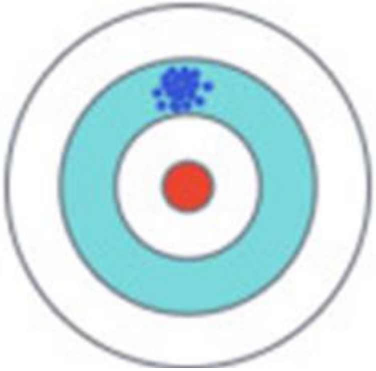


Large Variance

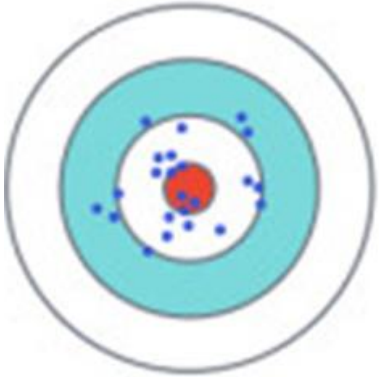
Simpler model is less influenced by the sampled data

Bias

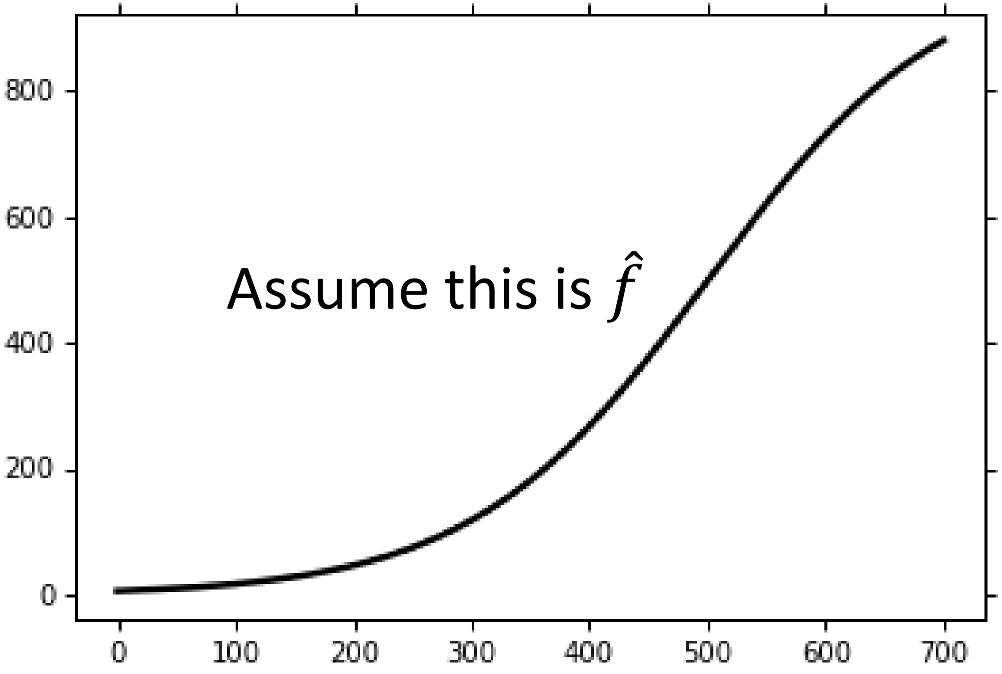
$$E[f^*] = \bar{f}$$



Large Bias



Small Bias



Assume this is $\hat{f}$

What to do with large bias?

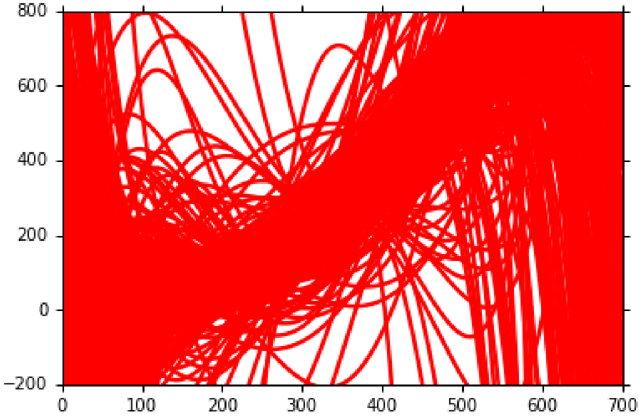
- Diagnosis:
  - If your model cannot even fit the training examples, then you have large bias Underfitting
  - If you can fit the training data, but large error on testing data, then you probably have large variance Overfitting

- For bias, redesign your model:
  - Add more features as input
  - A more complex model

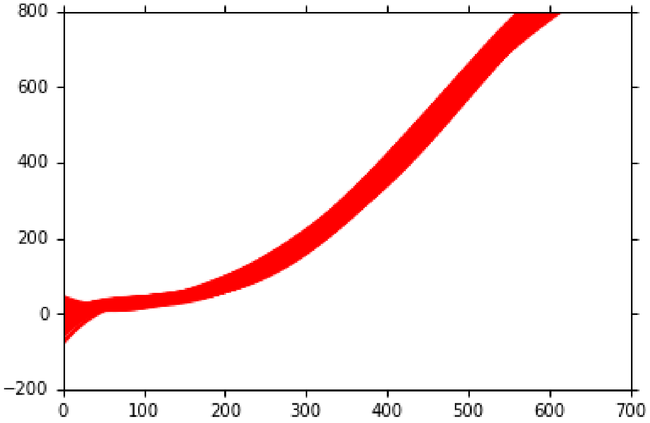
What to do with large variance?

- More data

Very effective, but not always practical



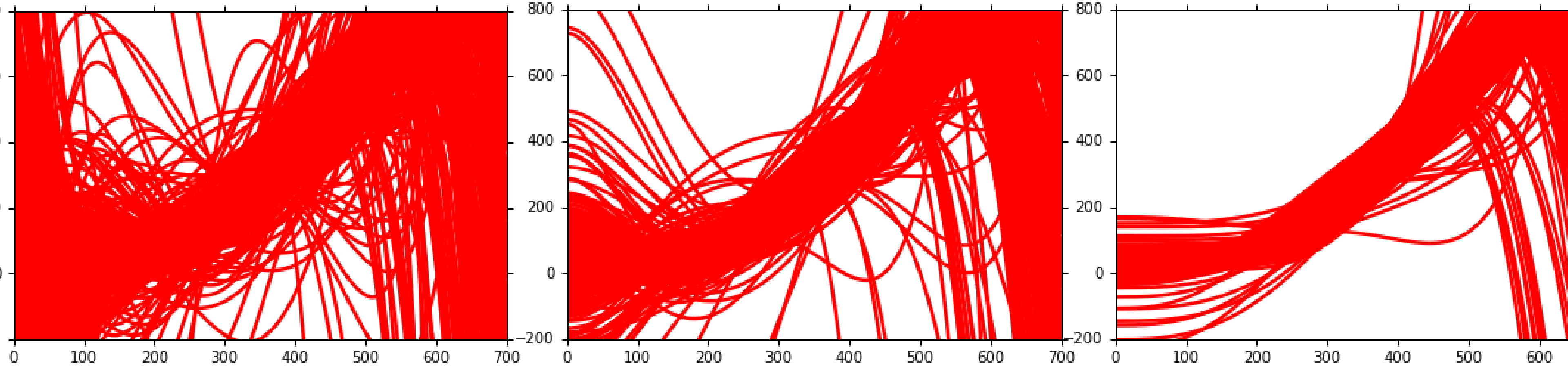
10 examples



100 examples

- Regularization

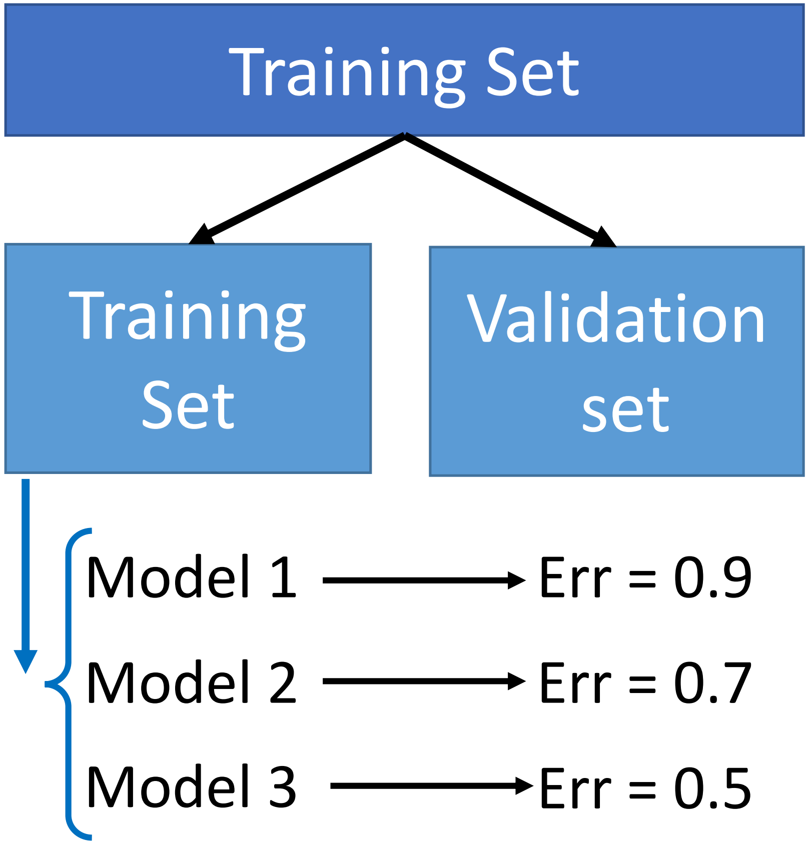
May increase bias



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

Cross Validation



Training Set

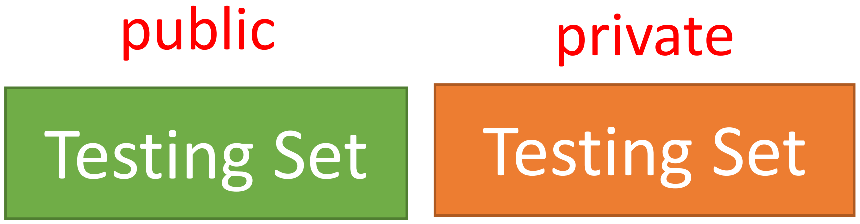
Training Set

Validation set

Model 1 → Err = 0.9

Model 2 → Err = 0.7

Model 3 → Err = 0.5



public

private

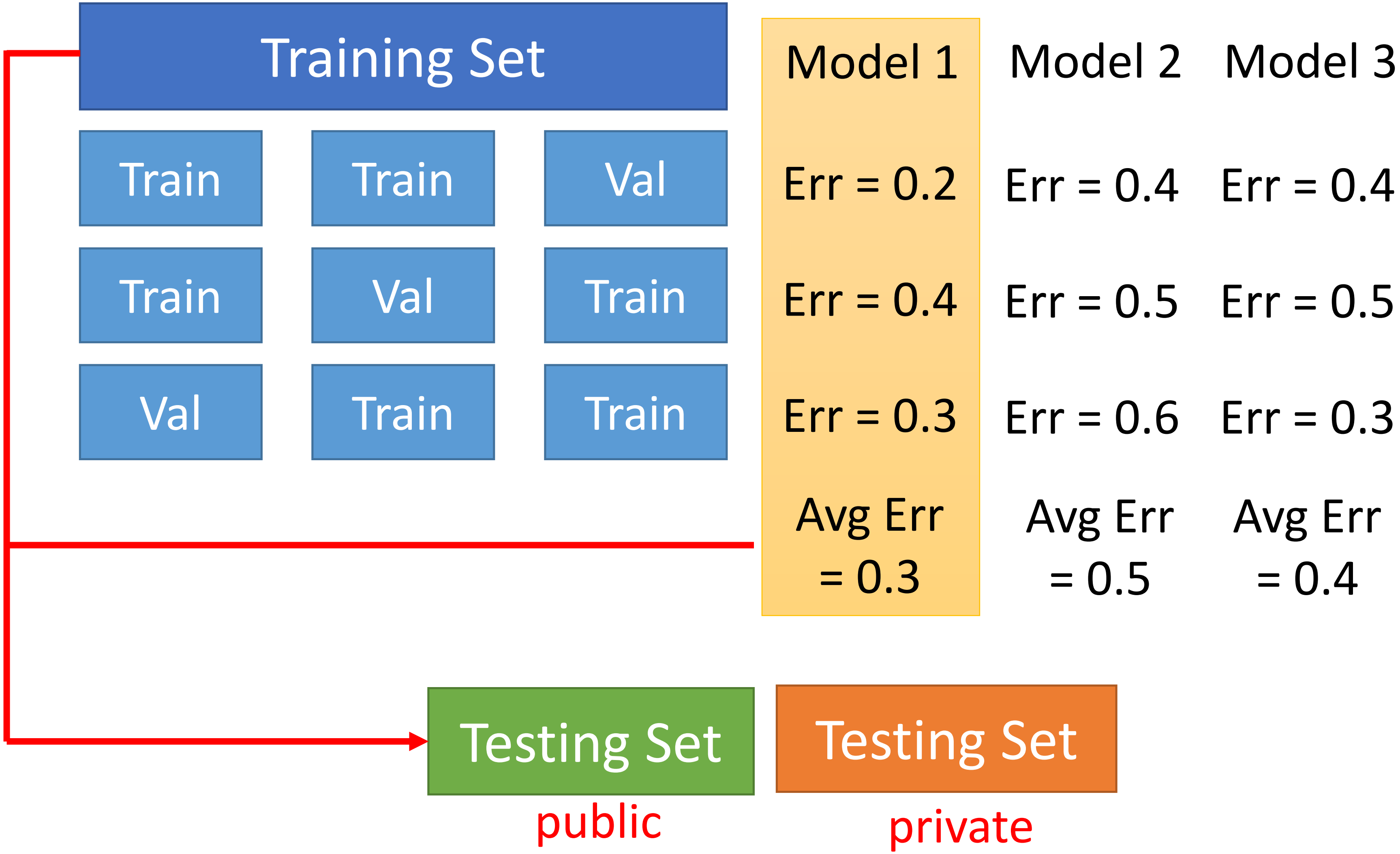
Testing Set

Testing Set

Using the results of public testing data to tune your model

You are making public set better than private set.

N-fold Cross Validation



Training Set

			Model 1	Model 2	Model 3
Train	Train	Val	Err = 0.2	Err = 0.4	Err = 0.4
Train	Val	Train	Err = 0.4	Err = 0.5	Err = 0.5
Val	Train	Train	Err = 0.3	Err = 0.6	Err = 0.3
			Avg Err = 0.3	Avg Err = 0.5	Avg Err = 0.4

Testing Set

public

Testing Set

private