NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Data Science

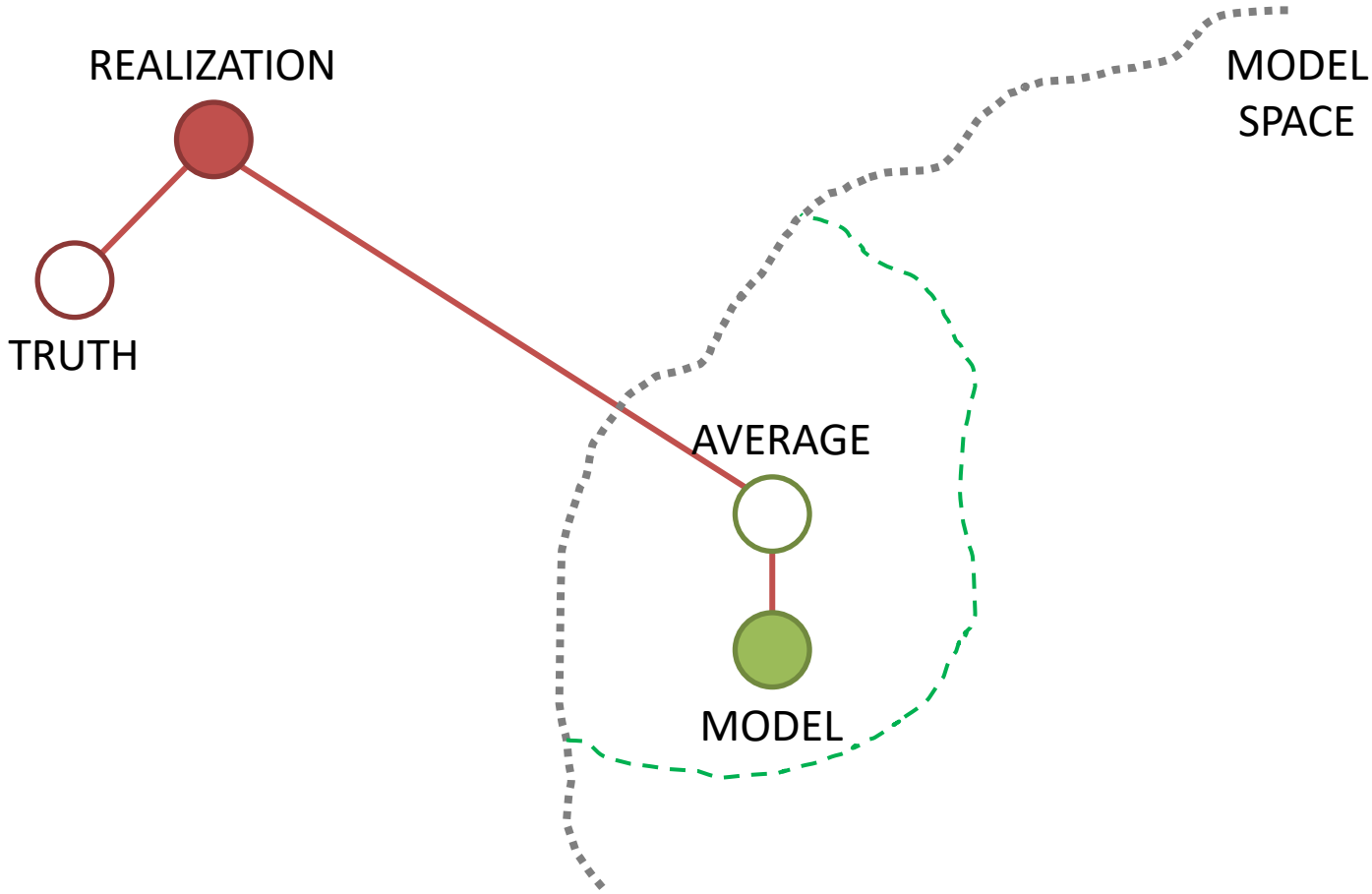
Bias Variance

Sourav SEN GUPTA, Lecturer

School of Computer Science and Engg.
Nanyang Technological University

The main concern of Data Science

ACCURACY OF THE MODEL



$$y = f(x)$$

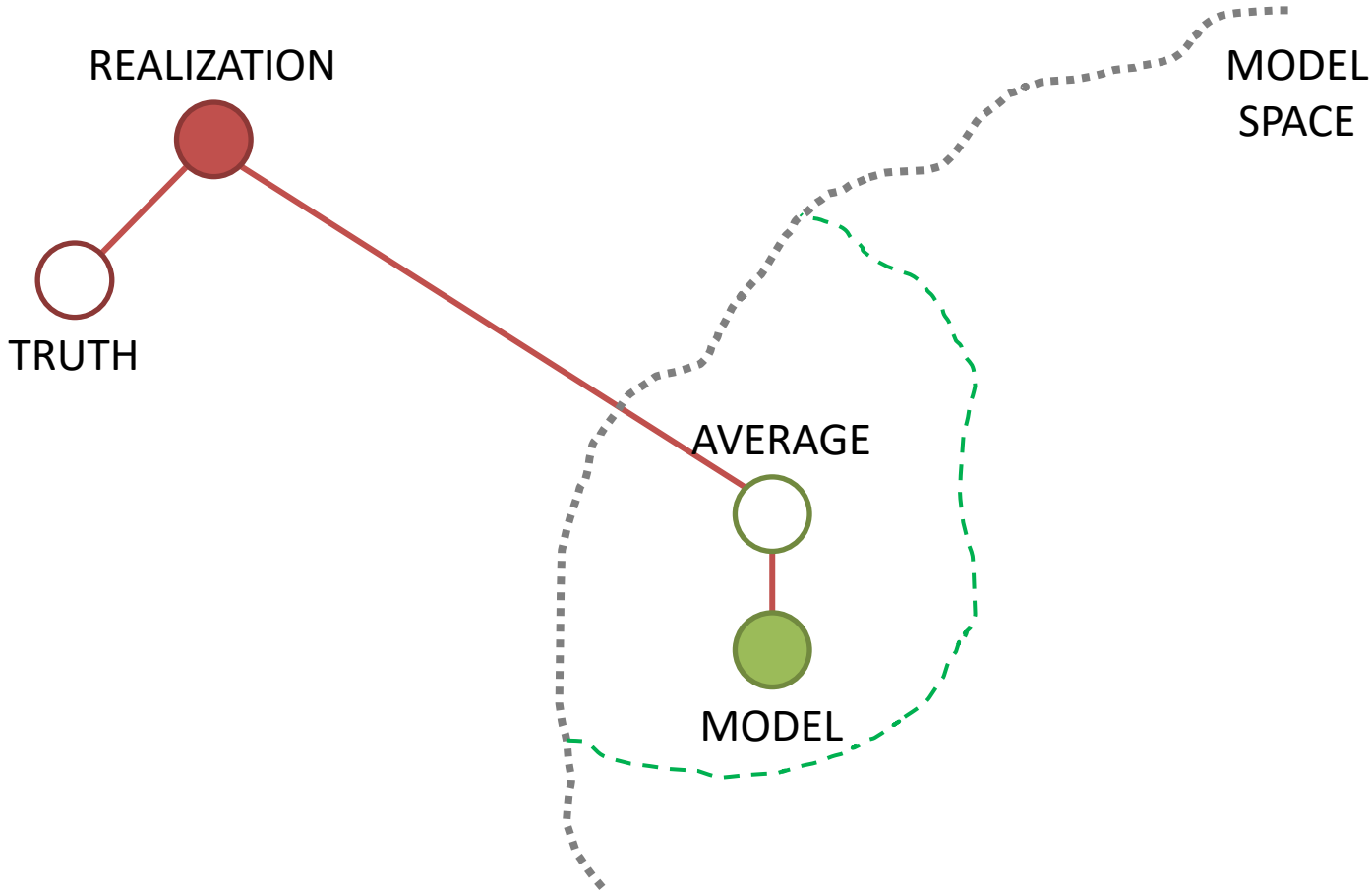
TRUTH

$$y = f(x) + \epsilon$$

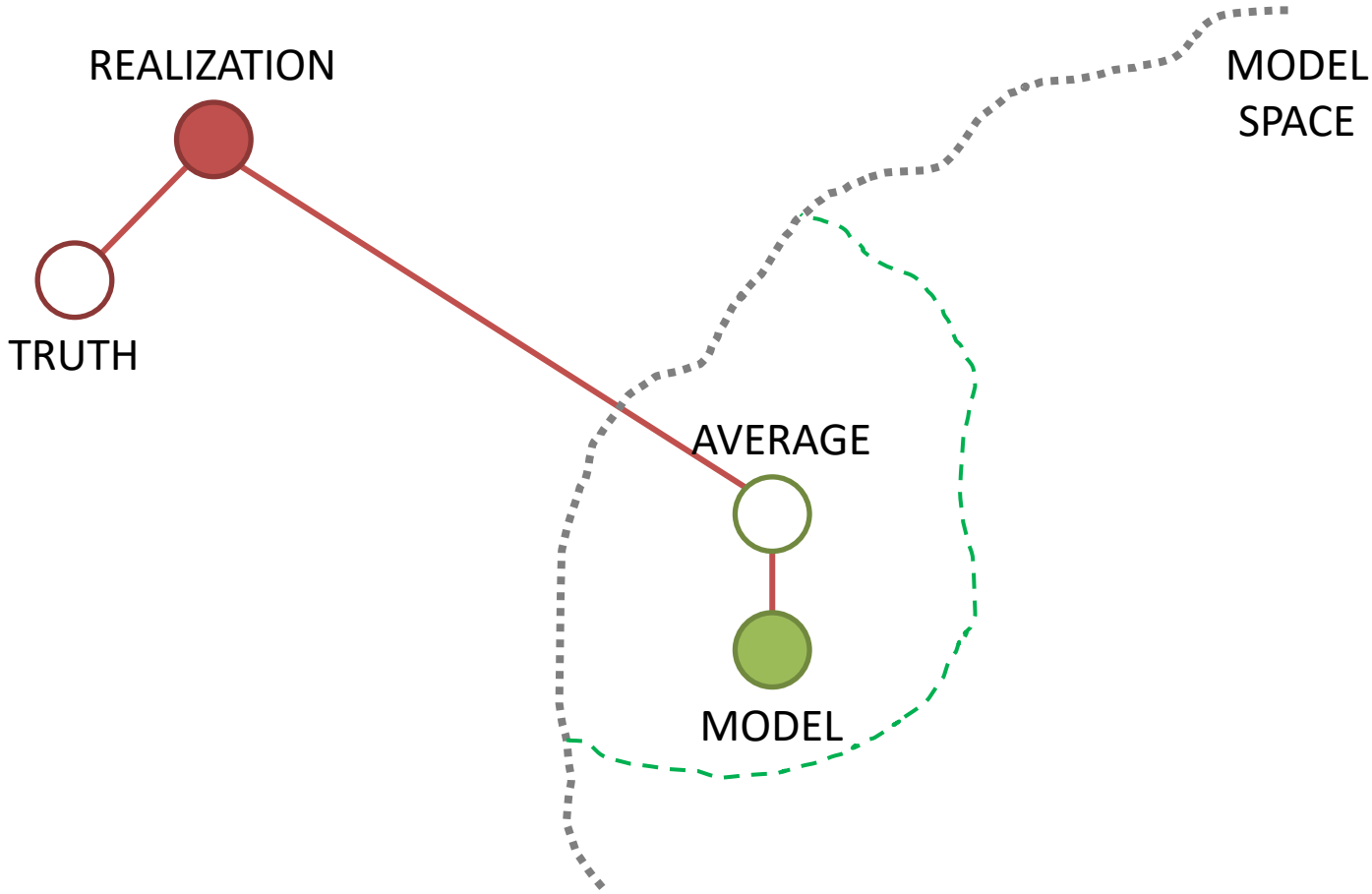
REALIZATION

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

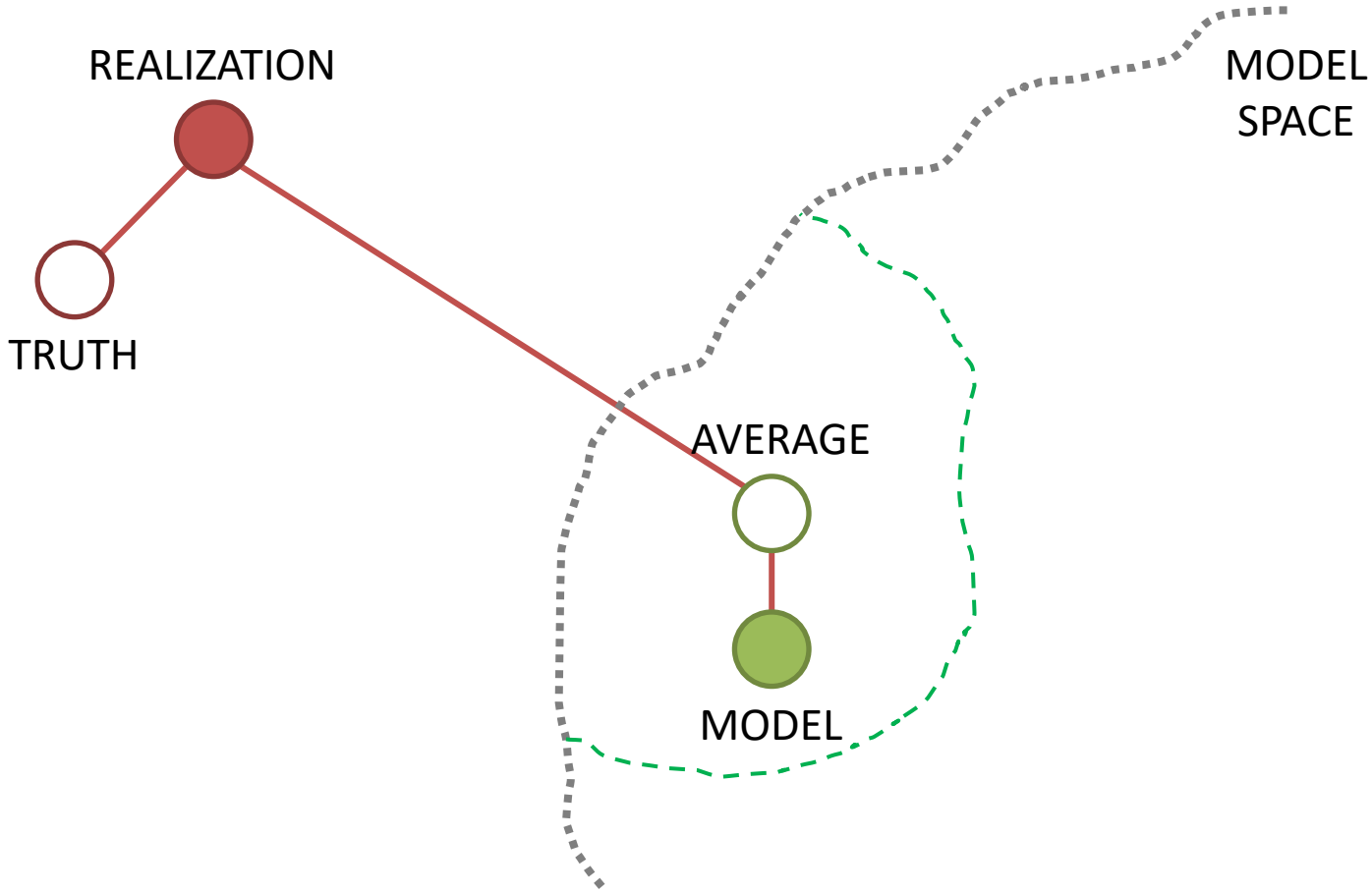
MODEL



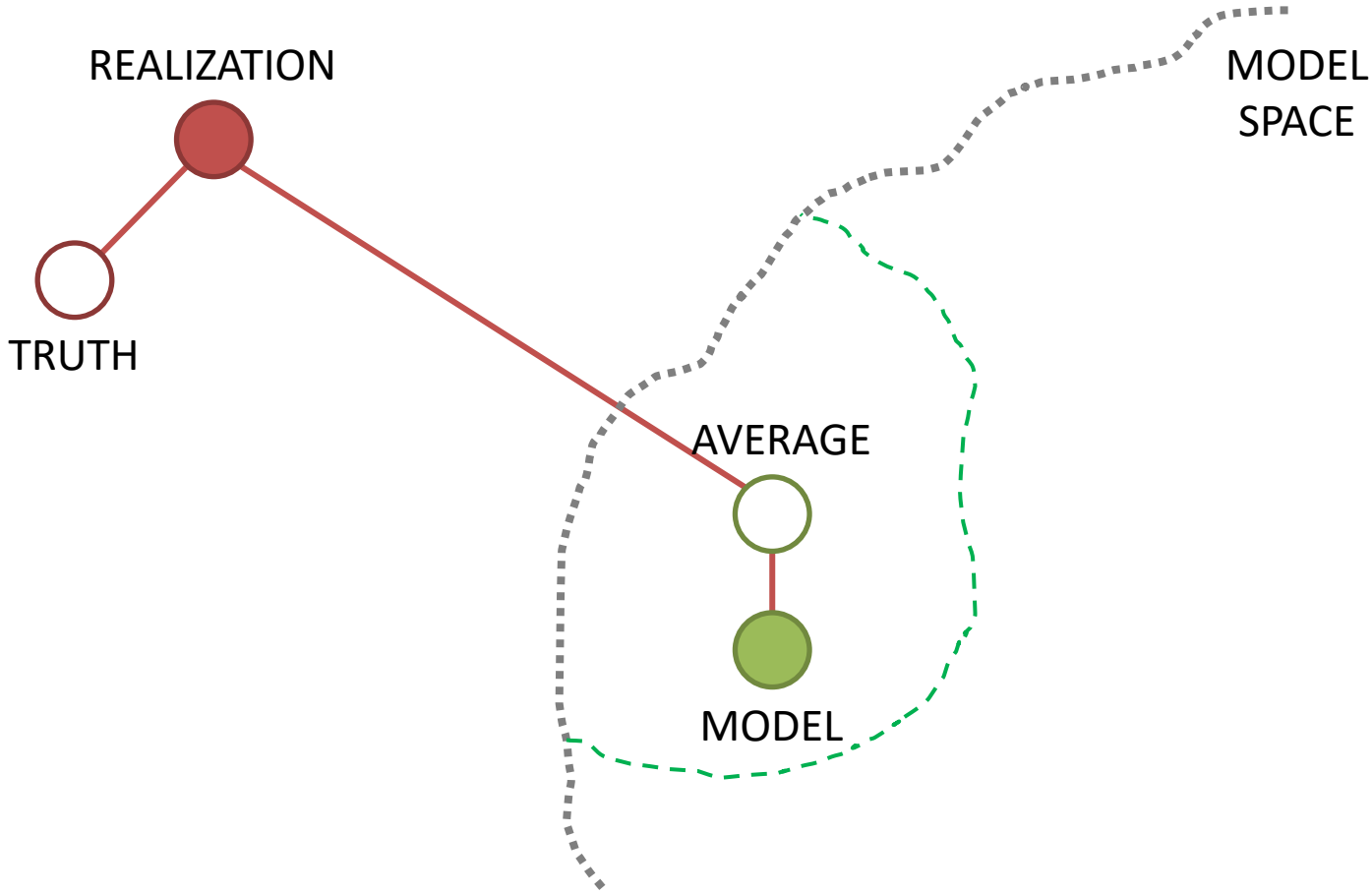
The goal is to minimize the Error, on an average!



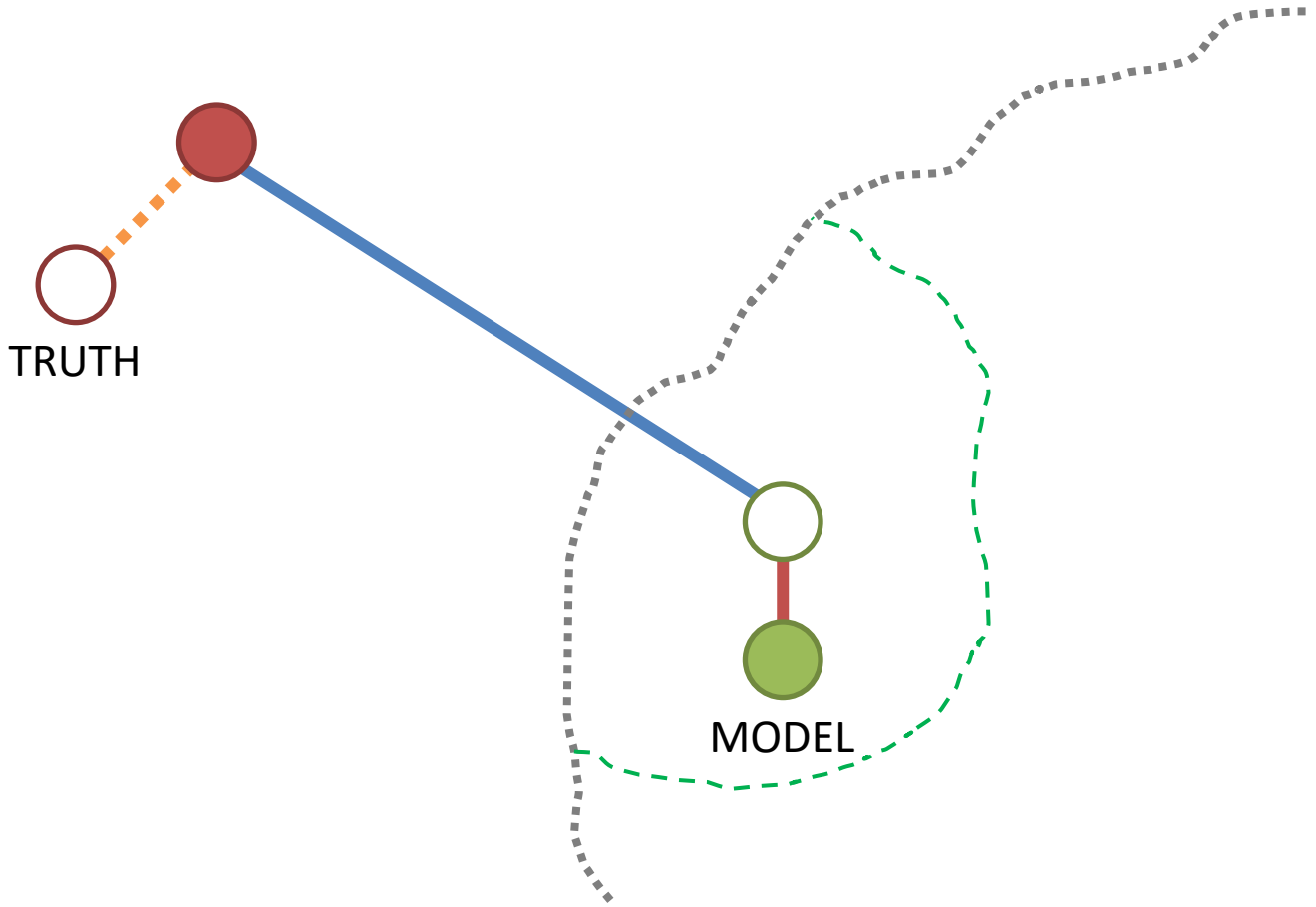
$$E(error^2) = E[(y - \hat{y})^2] = E\left[\left(f(x) + \epsilon - \hat{f}(x)\right)^2\right]$$



$$E(error^2) = \left(f - E\hat{f}\right)^2 + E\left(E\hat{f} - \hat{f}\right)^2 + E(\epsilon)^2$$



$$E(error^2) = Bias^2 + Variance + Irreducible\ Err$$

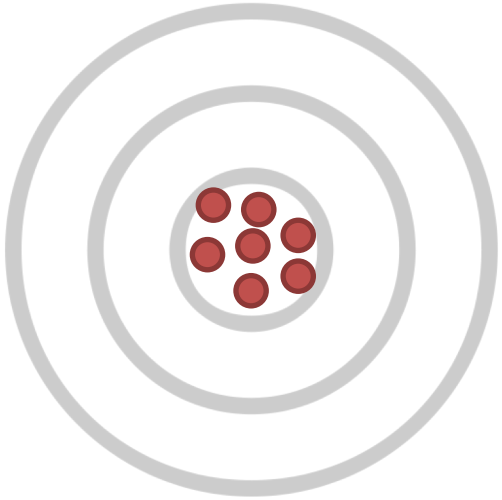


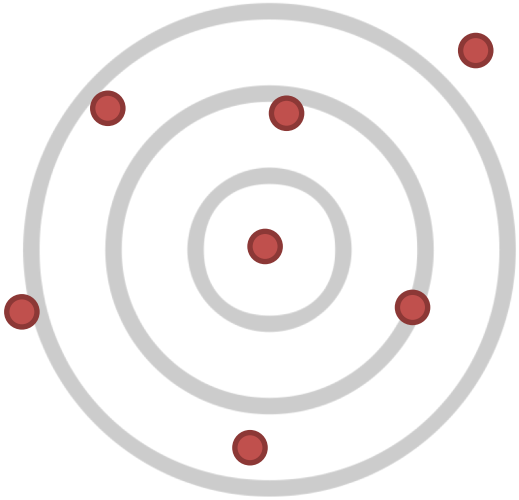
$$E(error^2) = \underline{Bias^2} + \underline{Variance} + \underline{Irreducible\ Err}$$

Core principle of Data Science

BIAS-VARIANCE TRADE-OFF

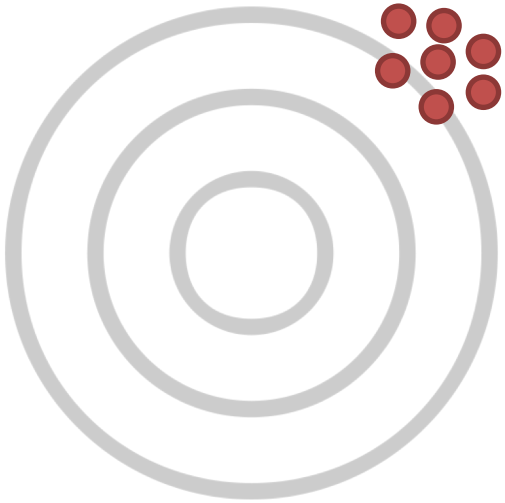
Low Bias
Low Variance

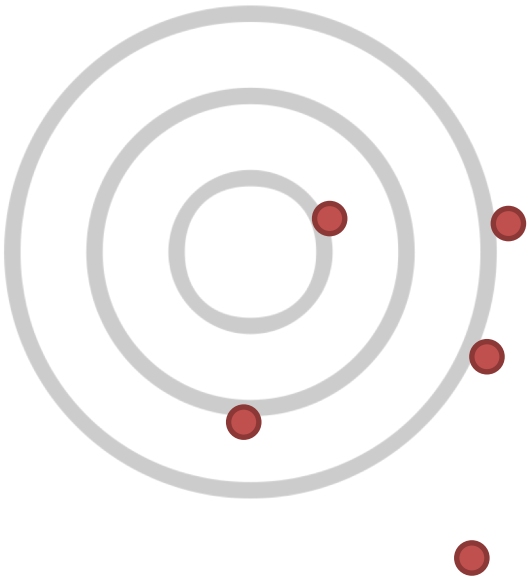




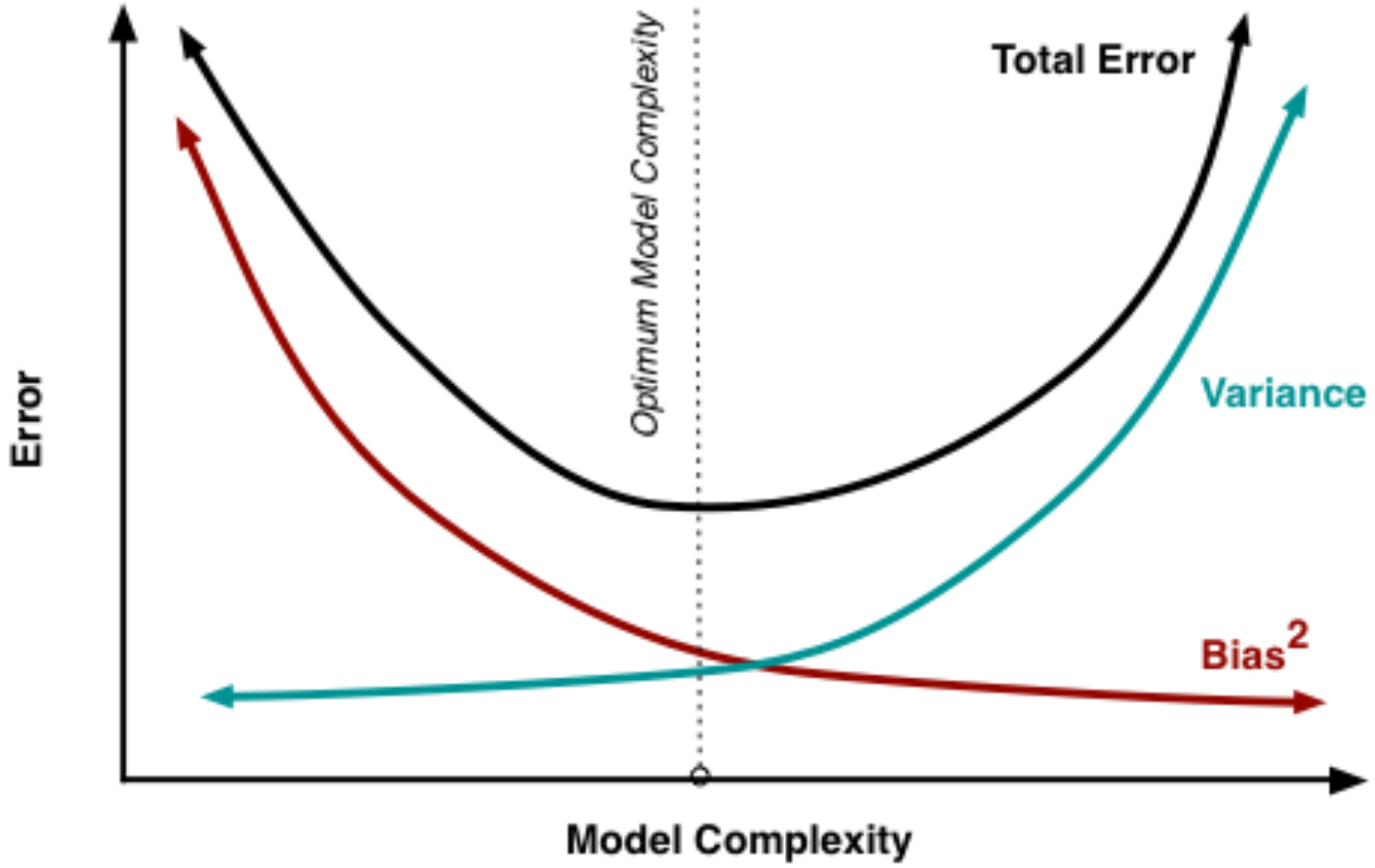
Low Bias
High Variance

High Bias
Low Variance

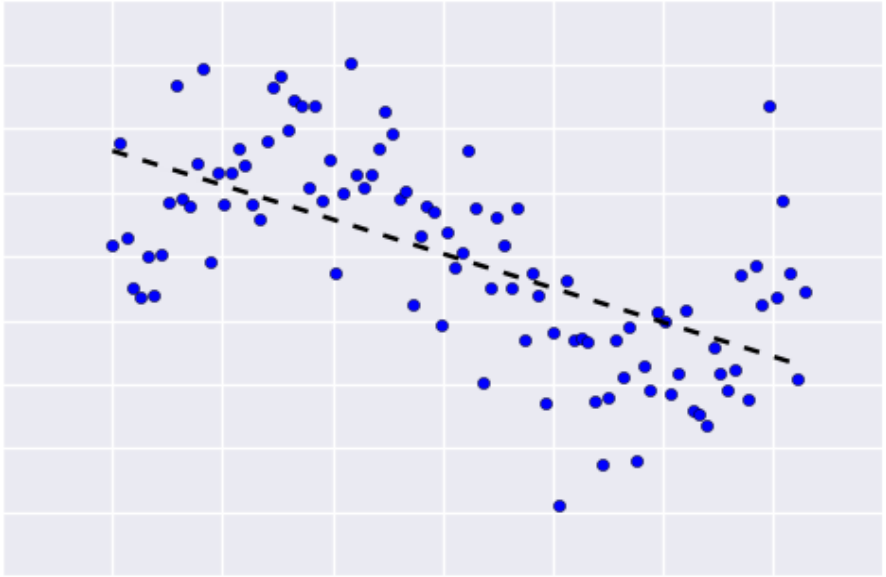


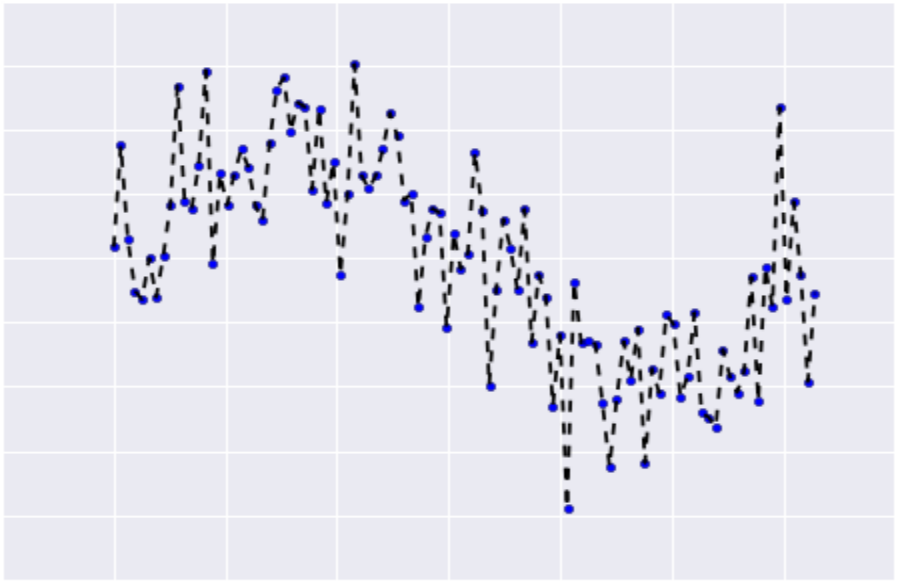


High Bias
High Variance



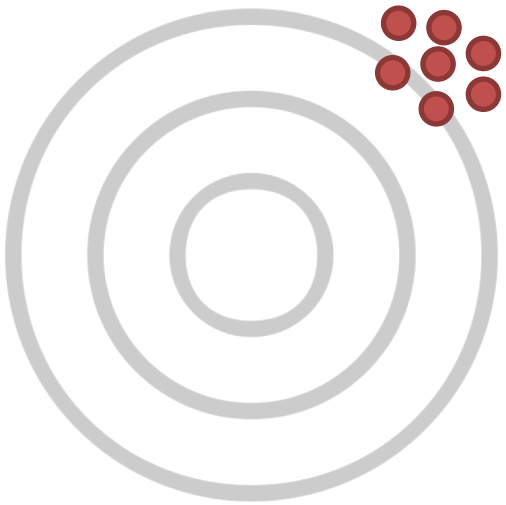
Model Complexity is the Hyper-Parameter to Tune

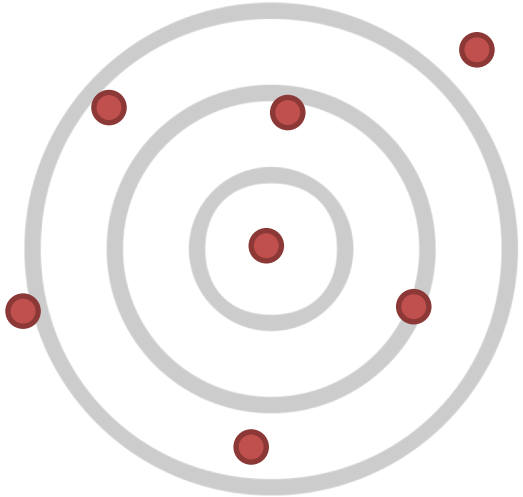




UNDER-FIT

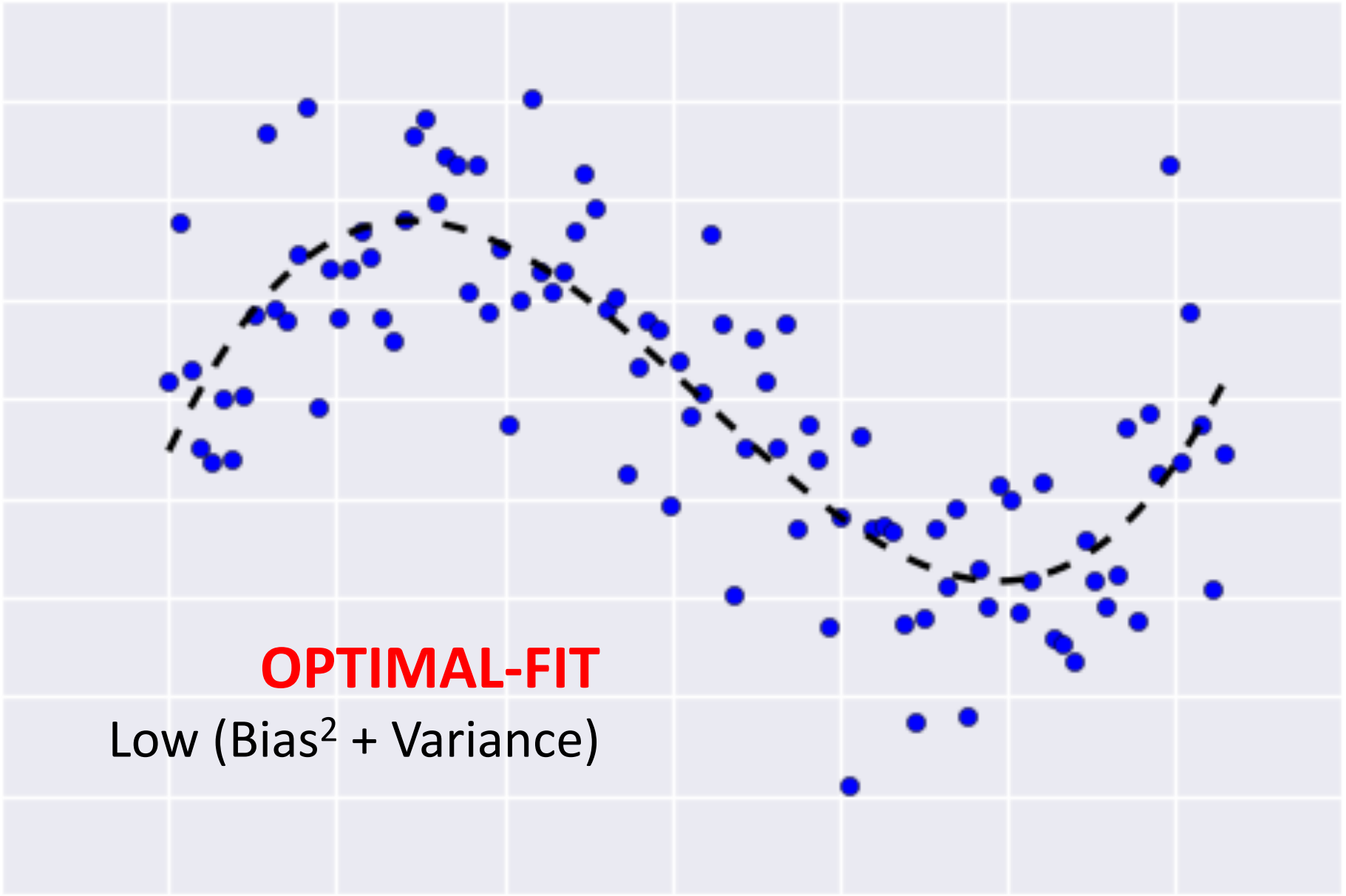
High Bias
Low Variance





OVER-FIT

Low Bias
High Variance

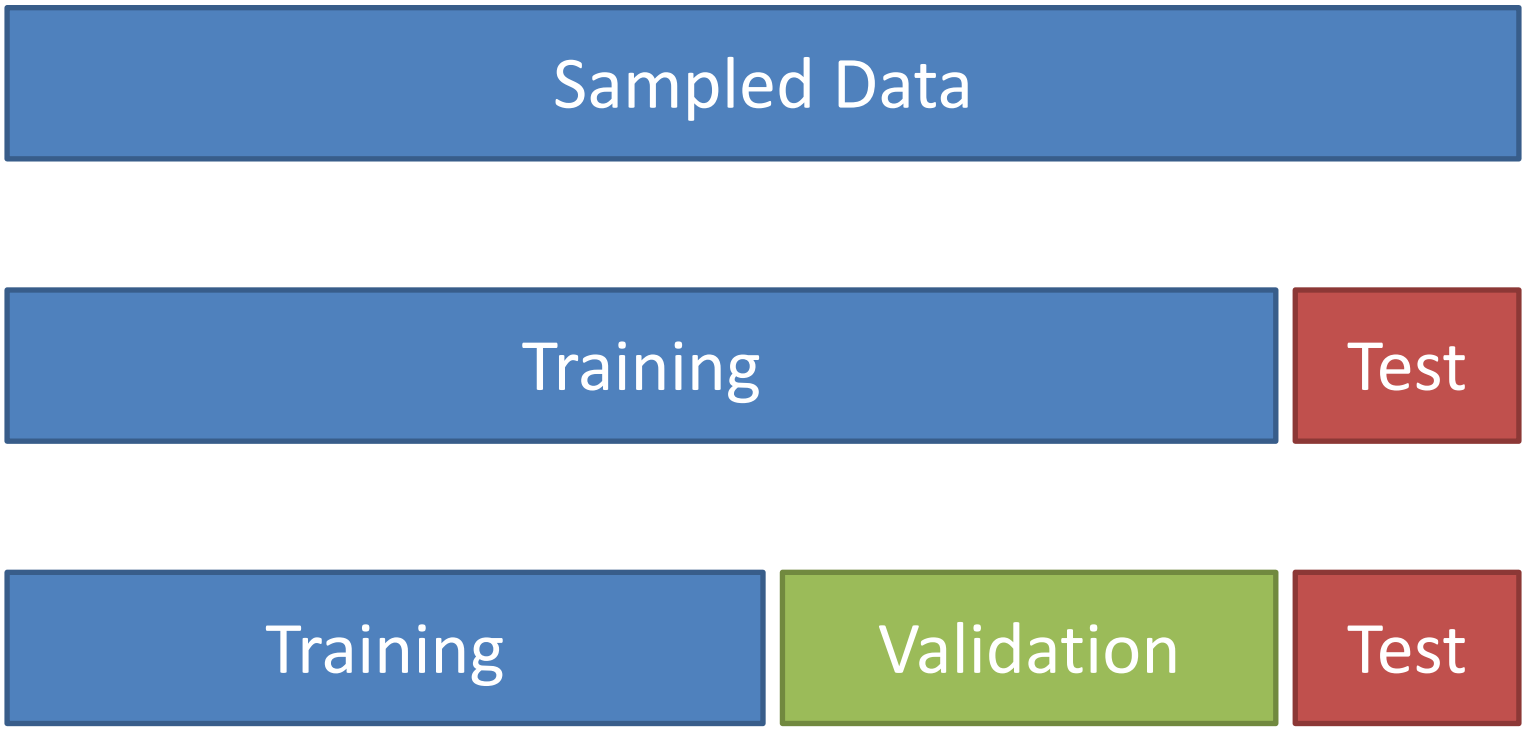


OPTIMAL-FIT

Low (Bias^2 + Variance)

The essential Data Science tool

CROSS-VALIDATION



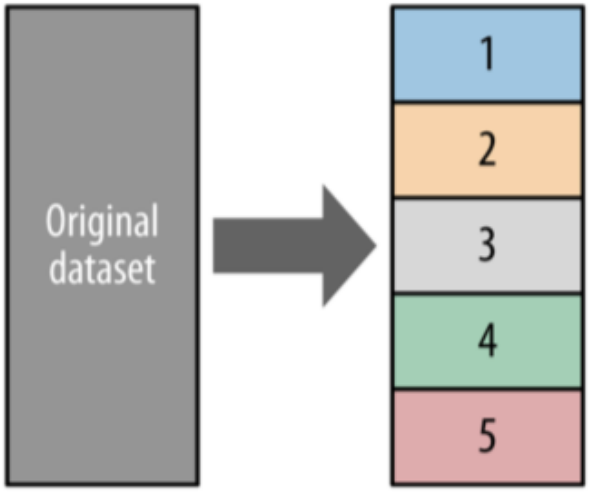
Larger Training Set implies "More Information" in Model Building
Larger Validation Set implies "More Diversity" in Model testing

General Practice

Training Set : Validation Set :: 70 : 30

Leave-One-Out-Cross-Validation

k-Fold Cross-Validation



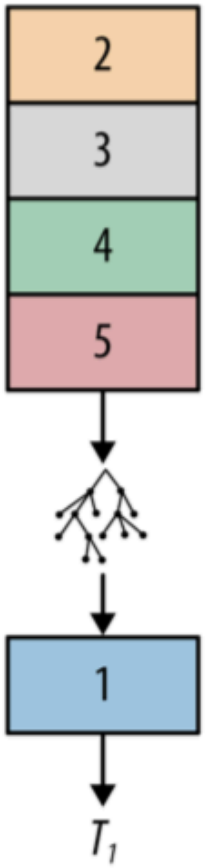
Split the dataset into *k* folds/parts

Use (*k*-1) folds of data for Training

Build the Model

Use the remaining fold for Validation

Note Performance



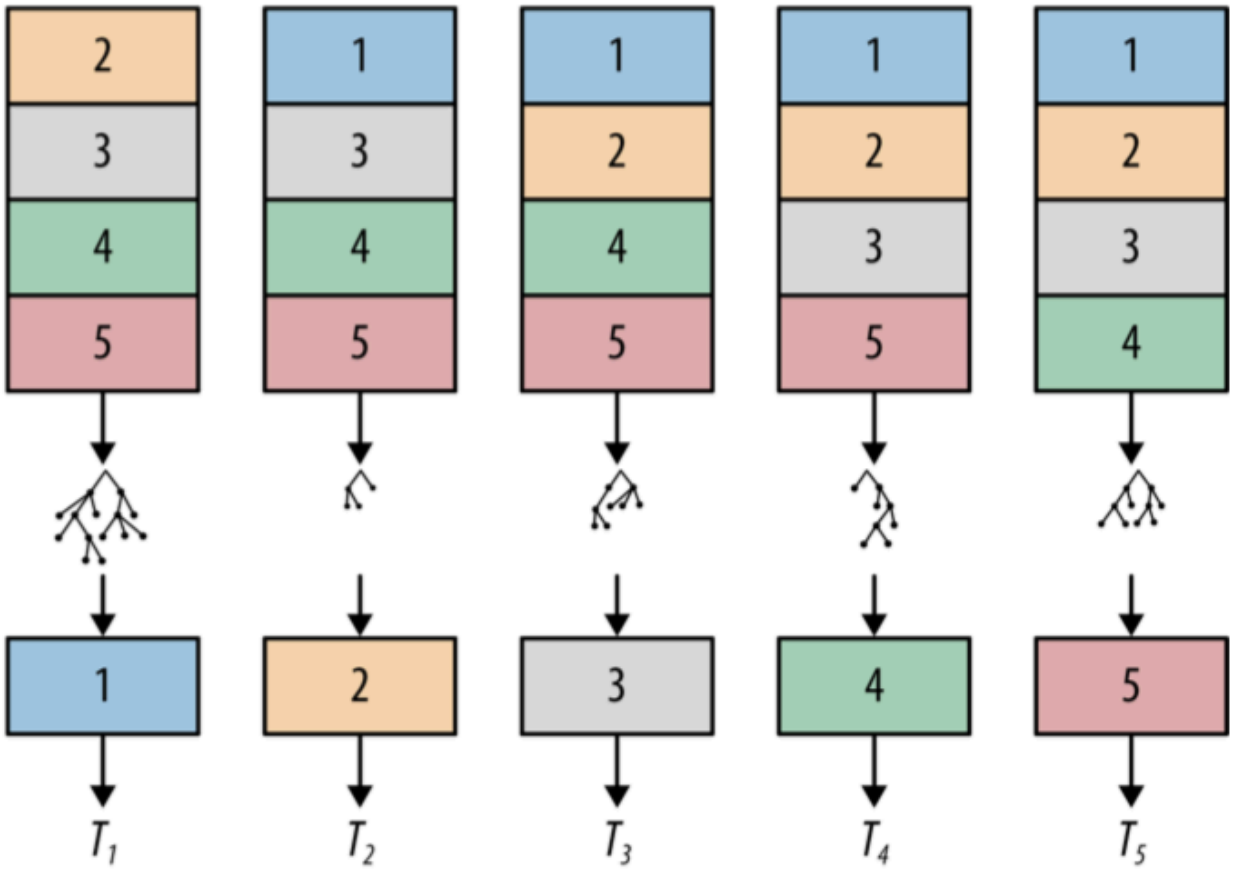
Repeat the process *k* times

Use (*k*-1) folds of data for Training

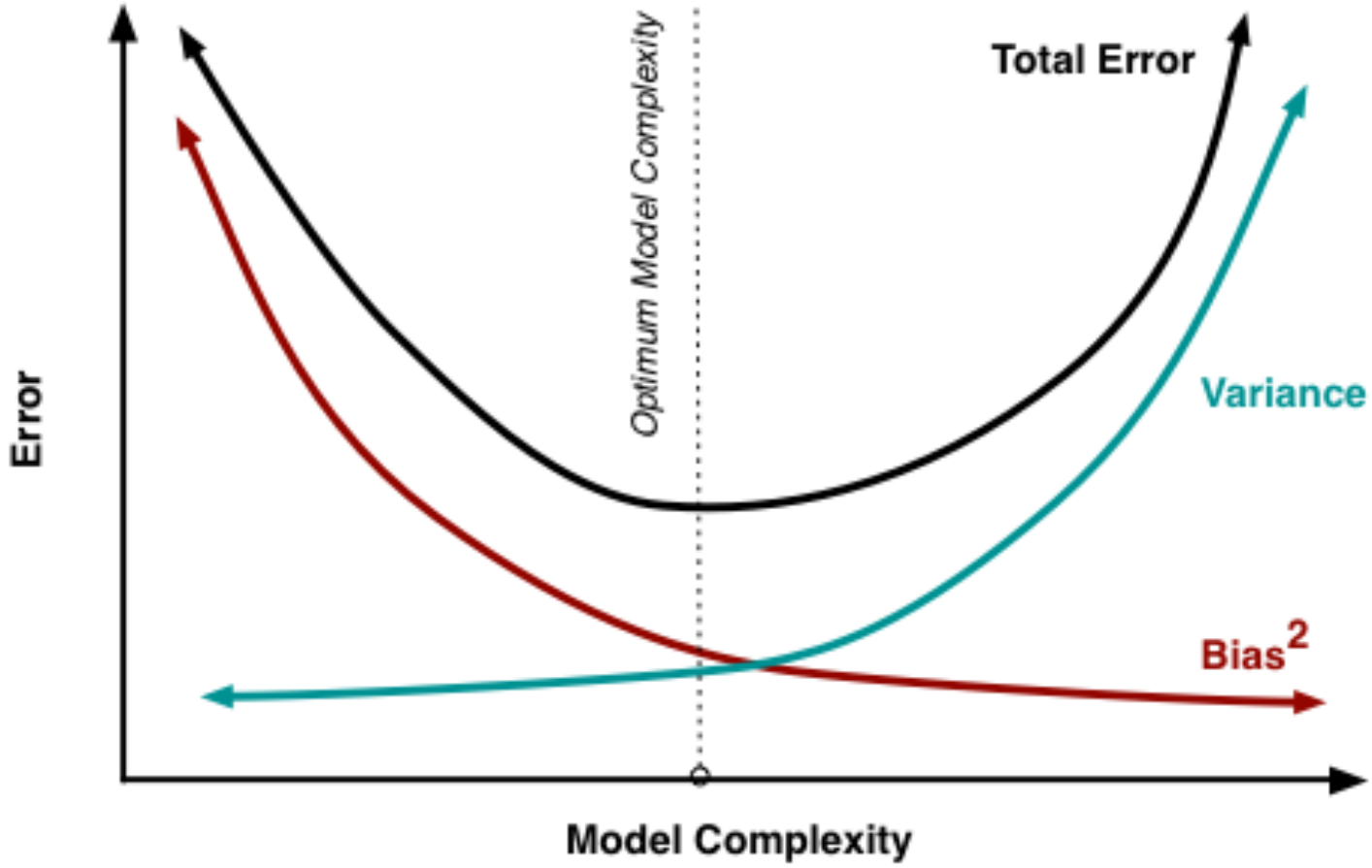
Build the Model

Use the remaining fold for Validation

Note Performance



Note the Mean Performance and its Variance



Understanding optimality of Model Complexity is critical in Data Science!