



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





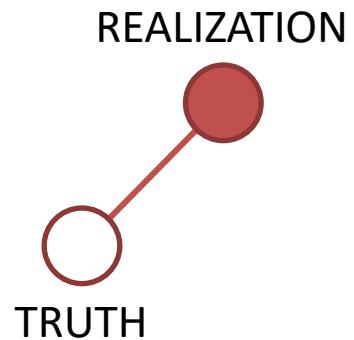
TRUTH

$$y = f(x)$$

TRUTH



NANYANG TECHNOLOGICAL UNIVERSITY | SINGAPORE



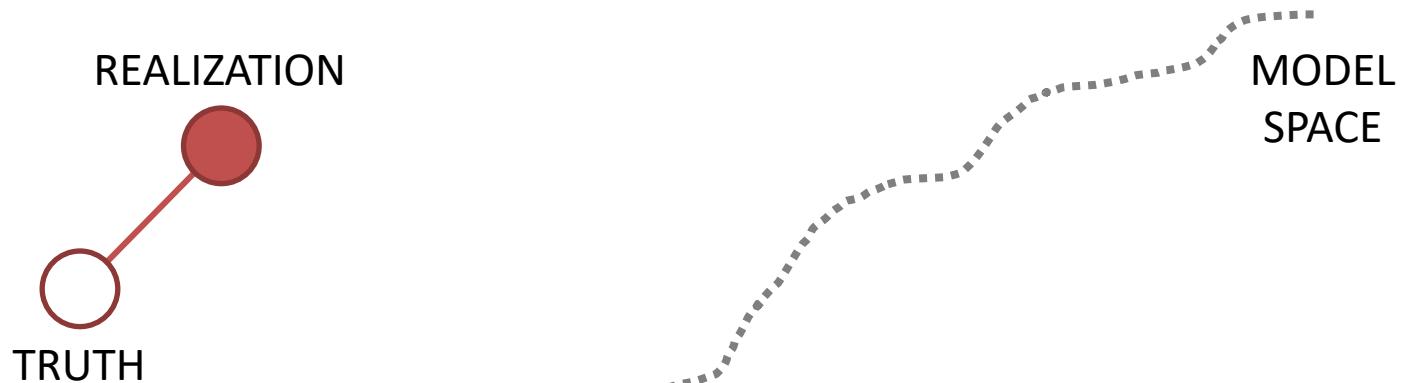
$$y = f(x)$$

TRUTH

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

REALIZATION





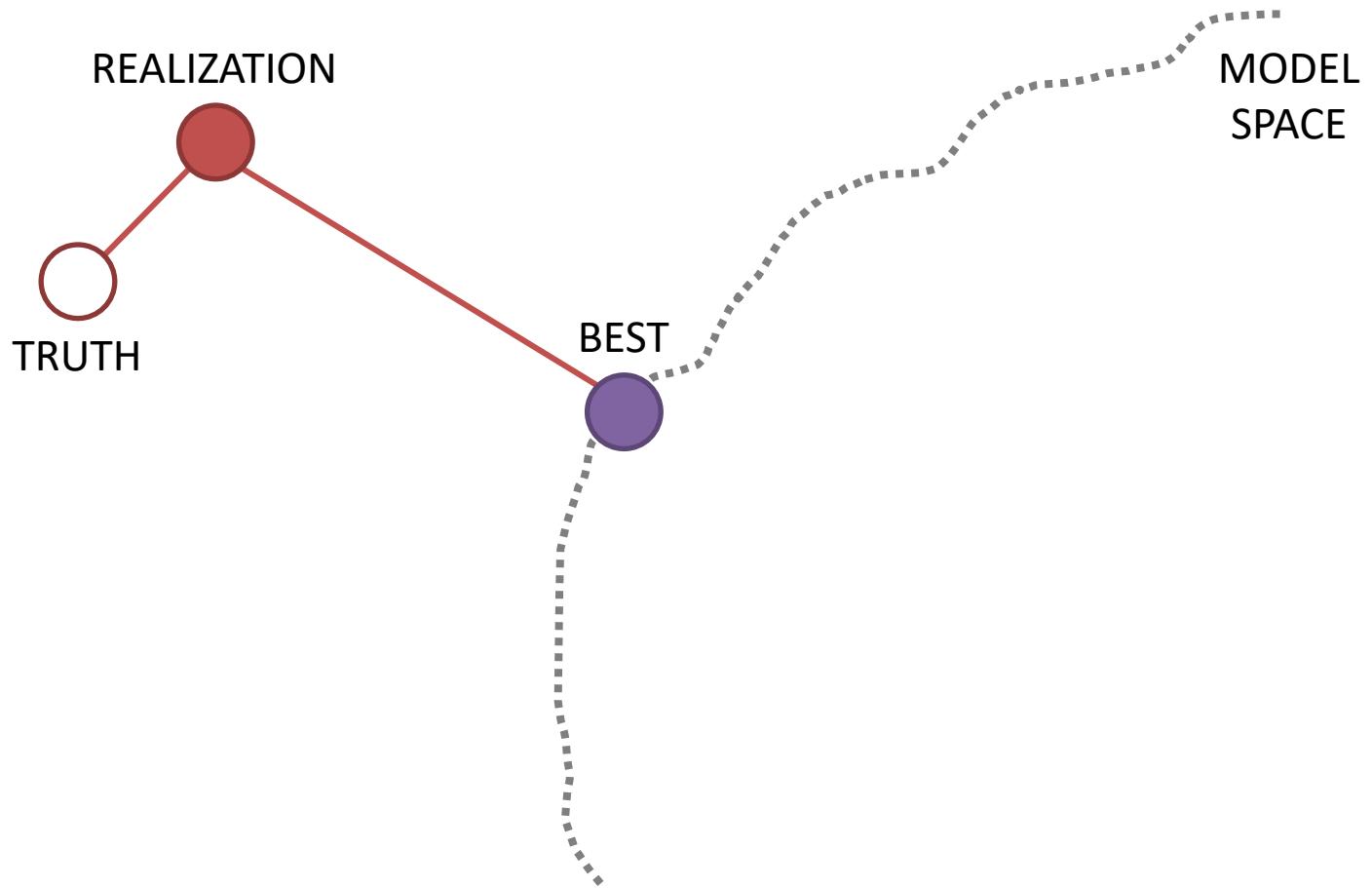
$$y = f(x)$$

TRUTH

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

REALIZATION





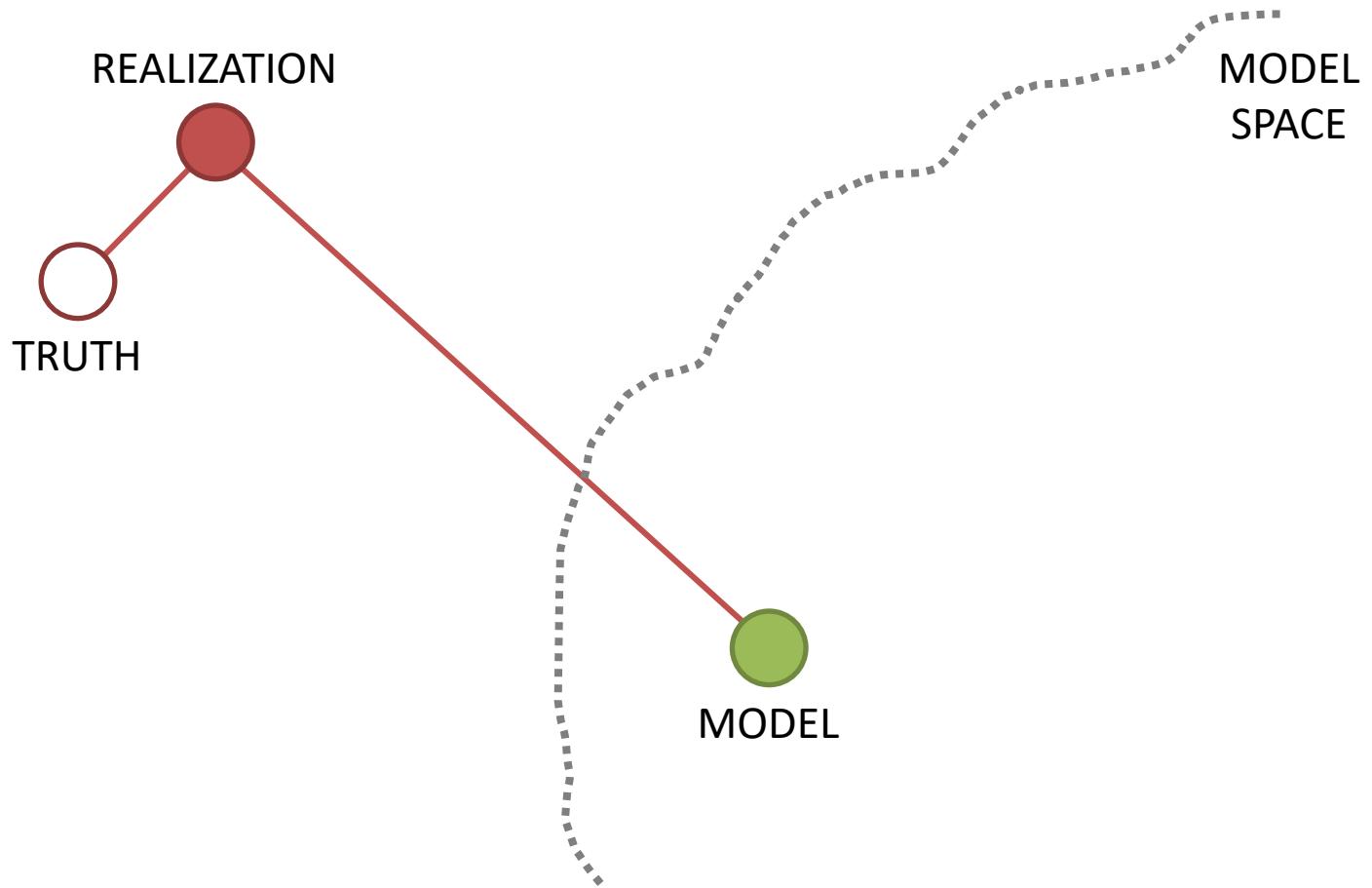
$$y = f(x)$$

TRUTH

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

REALIZATION





$$y = f(x)$$

TRUTH

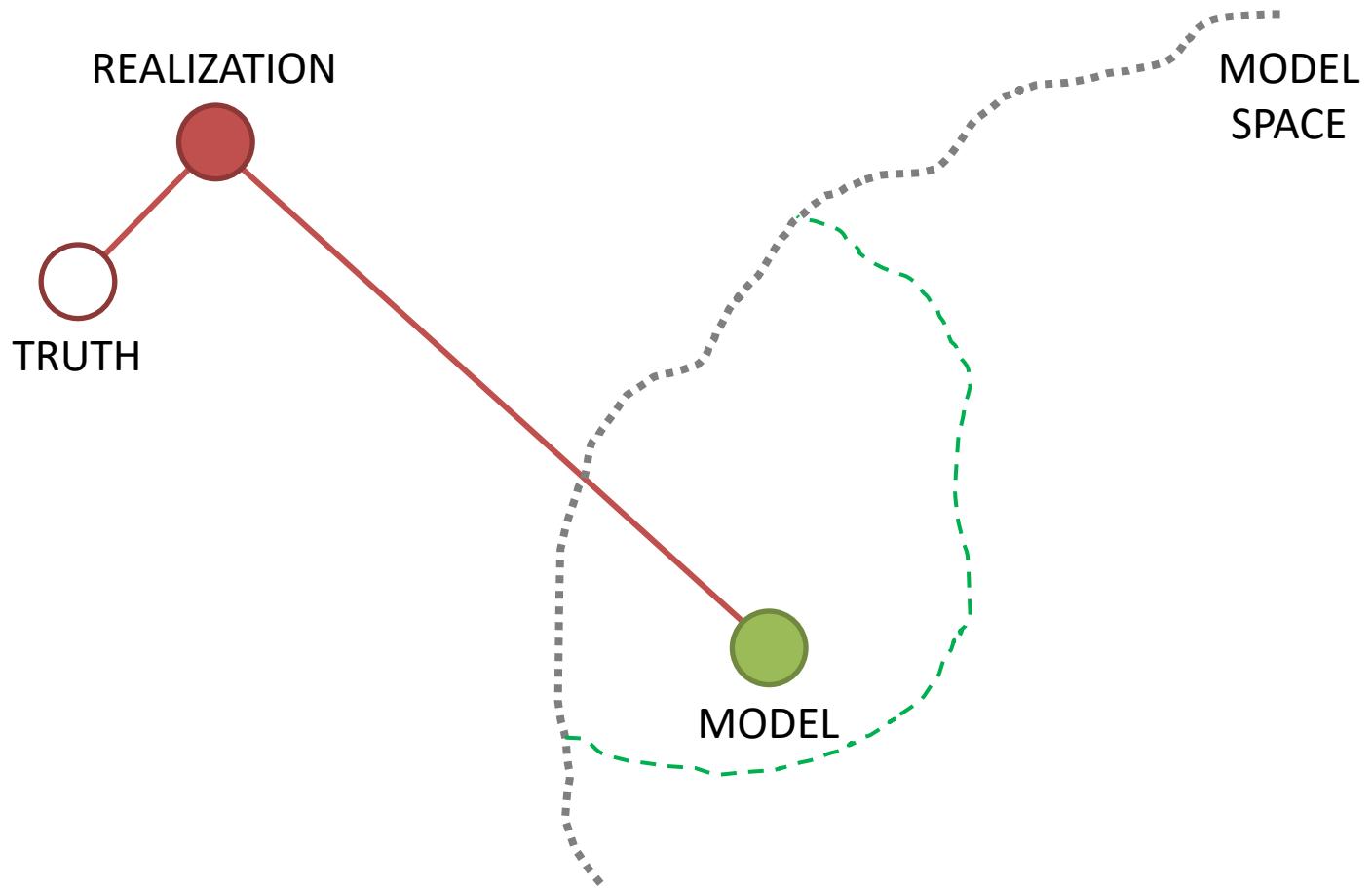
$$y = f(x) + \epsilon$$

REALIZATION

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

MODEL





$$y = f(x)$$

TRUTH

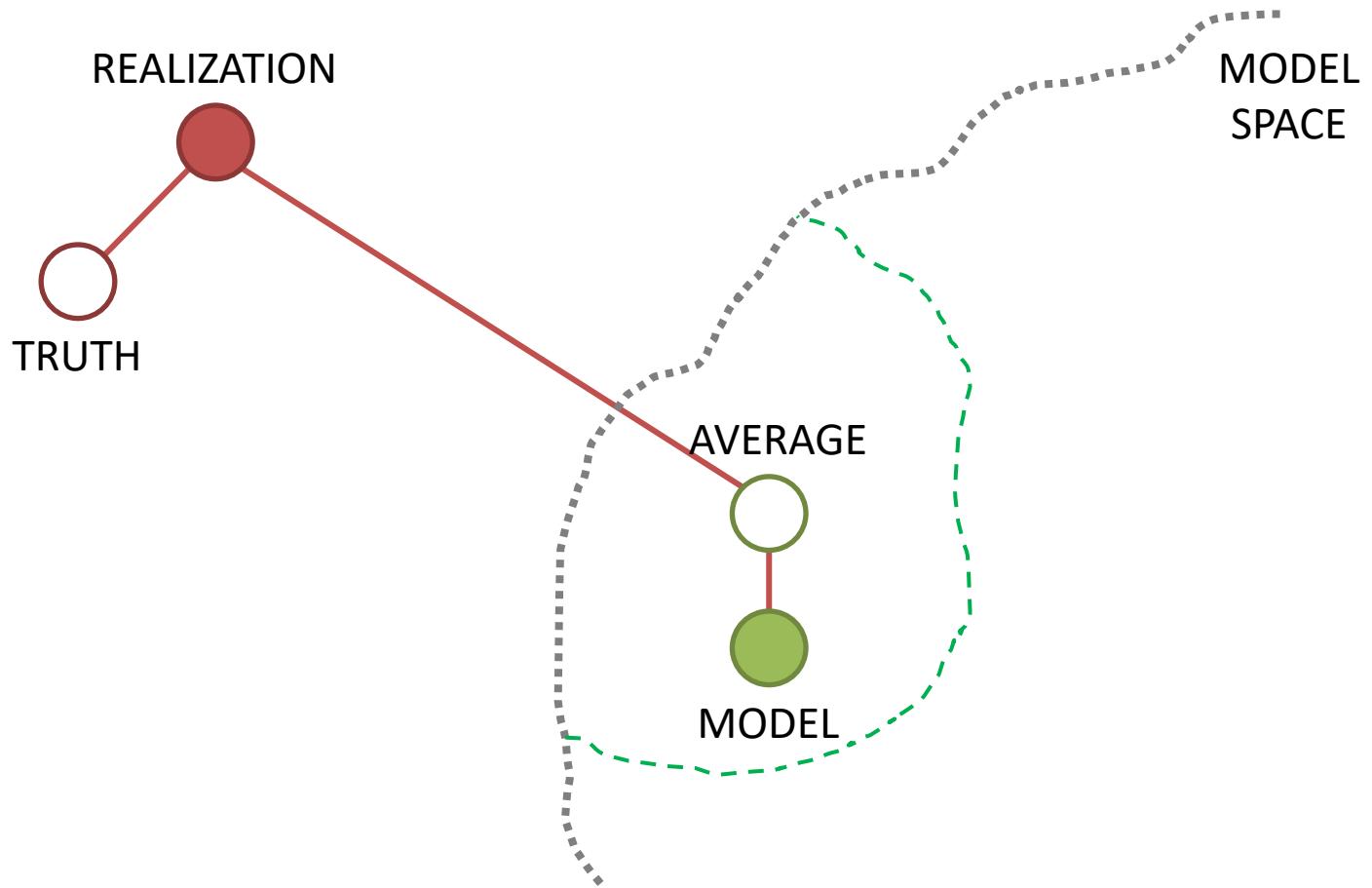
$$y = f(x) + \epsilon$$

REALIZATION

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

MODEL





$$y = f(x)$$

TRUTH

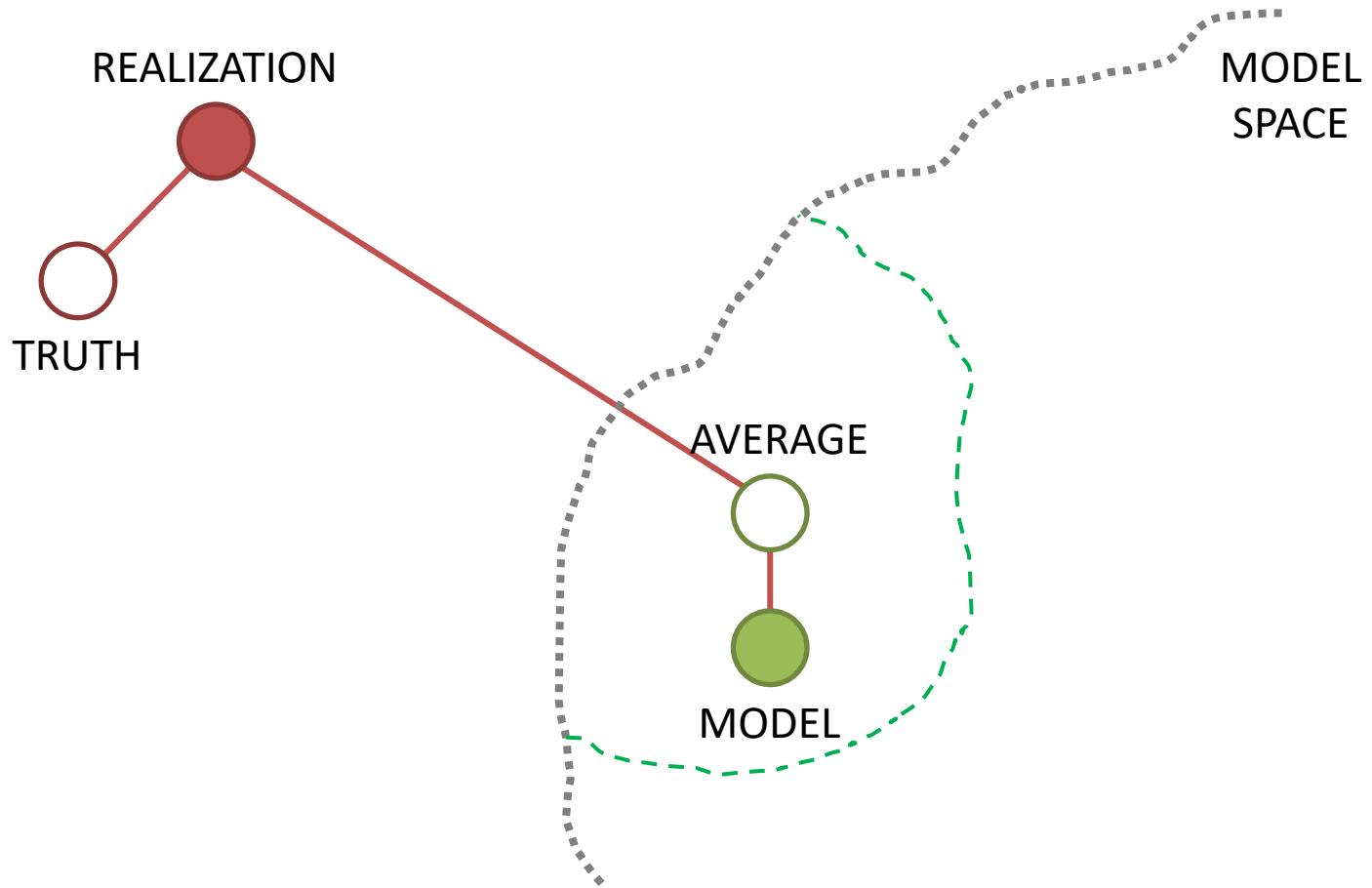
$$y = f(x) + \epsilon$$

REALIZATION

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

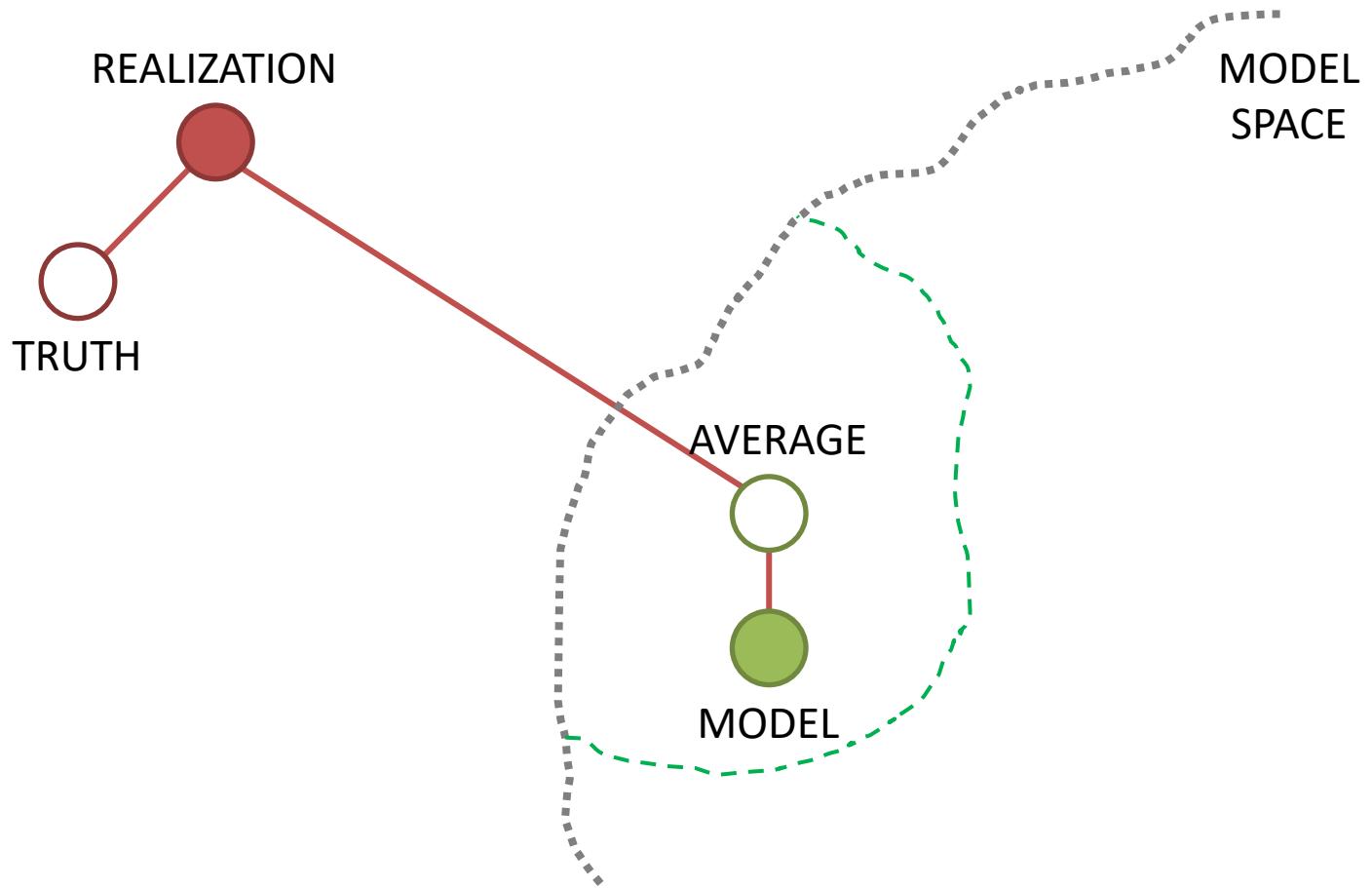
MODEL





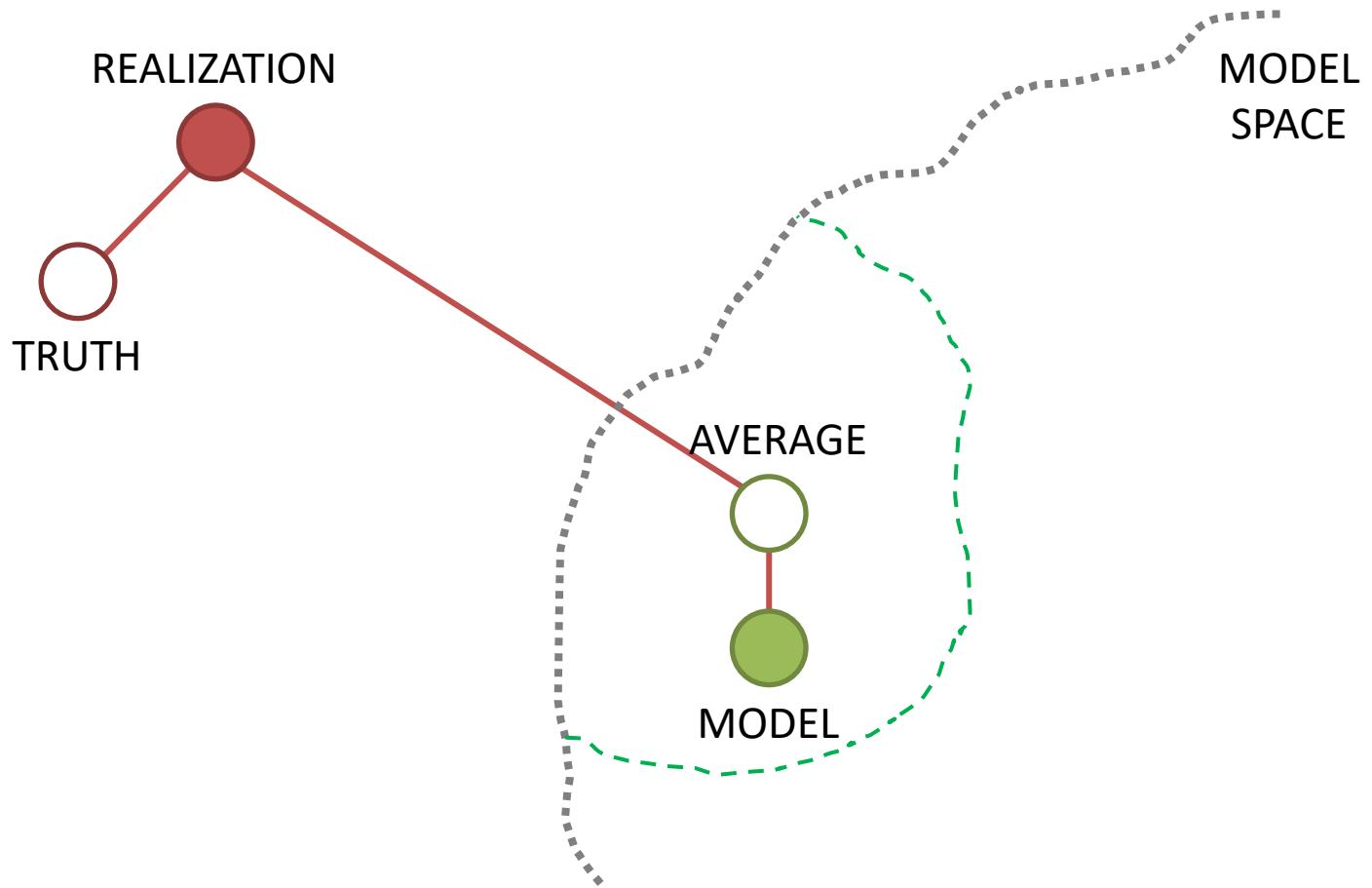
The goal is to minimize the Error, on an average!





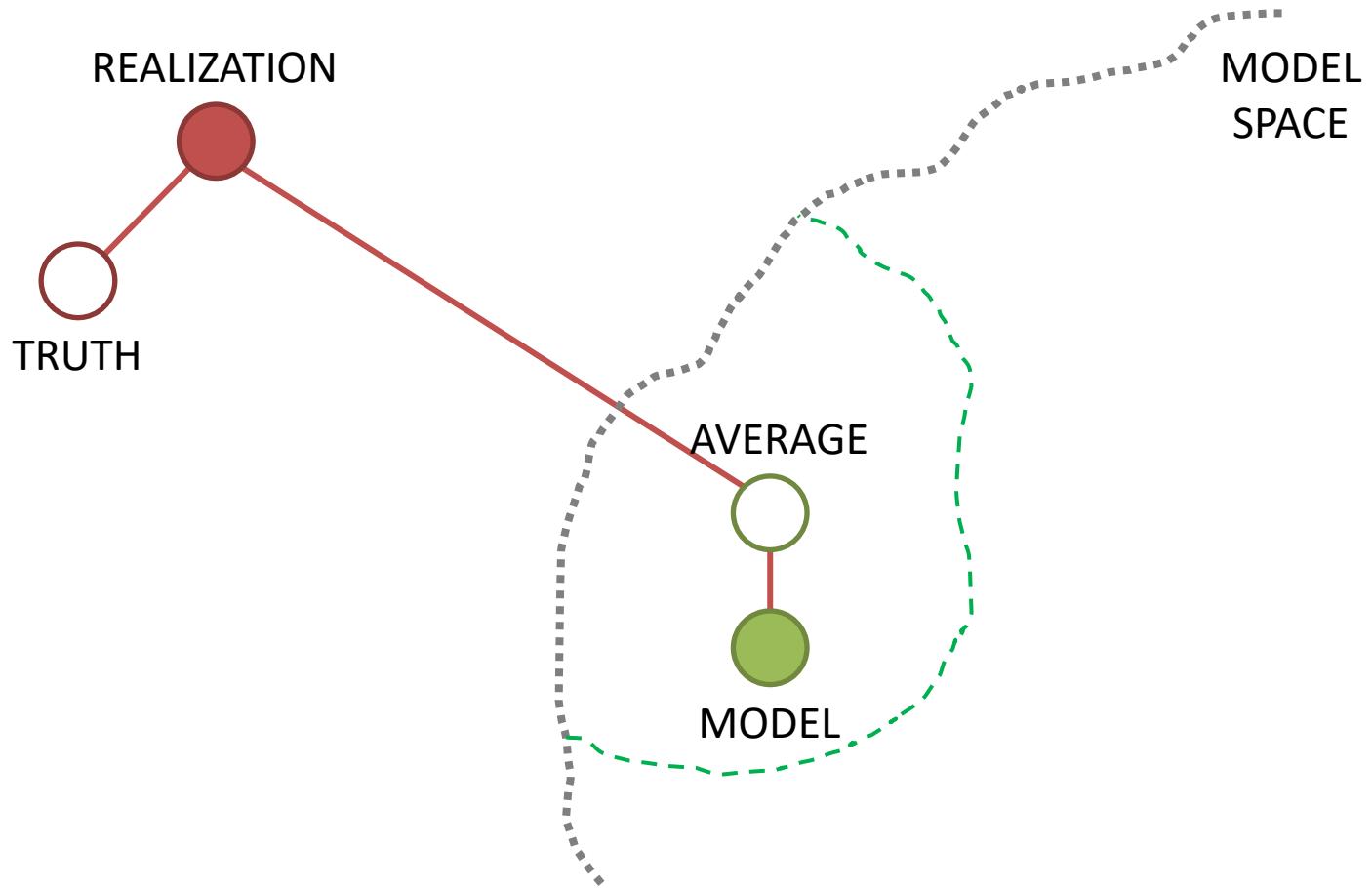
$$E(\text{error}^2) = E[(y - \hat{y})^2] = E \left[\left(f(x) + \epsilon - \hat{f}(x) \right)^2 \right]$$





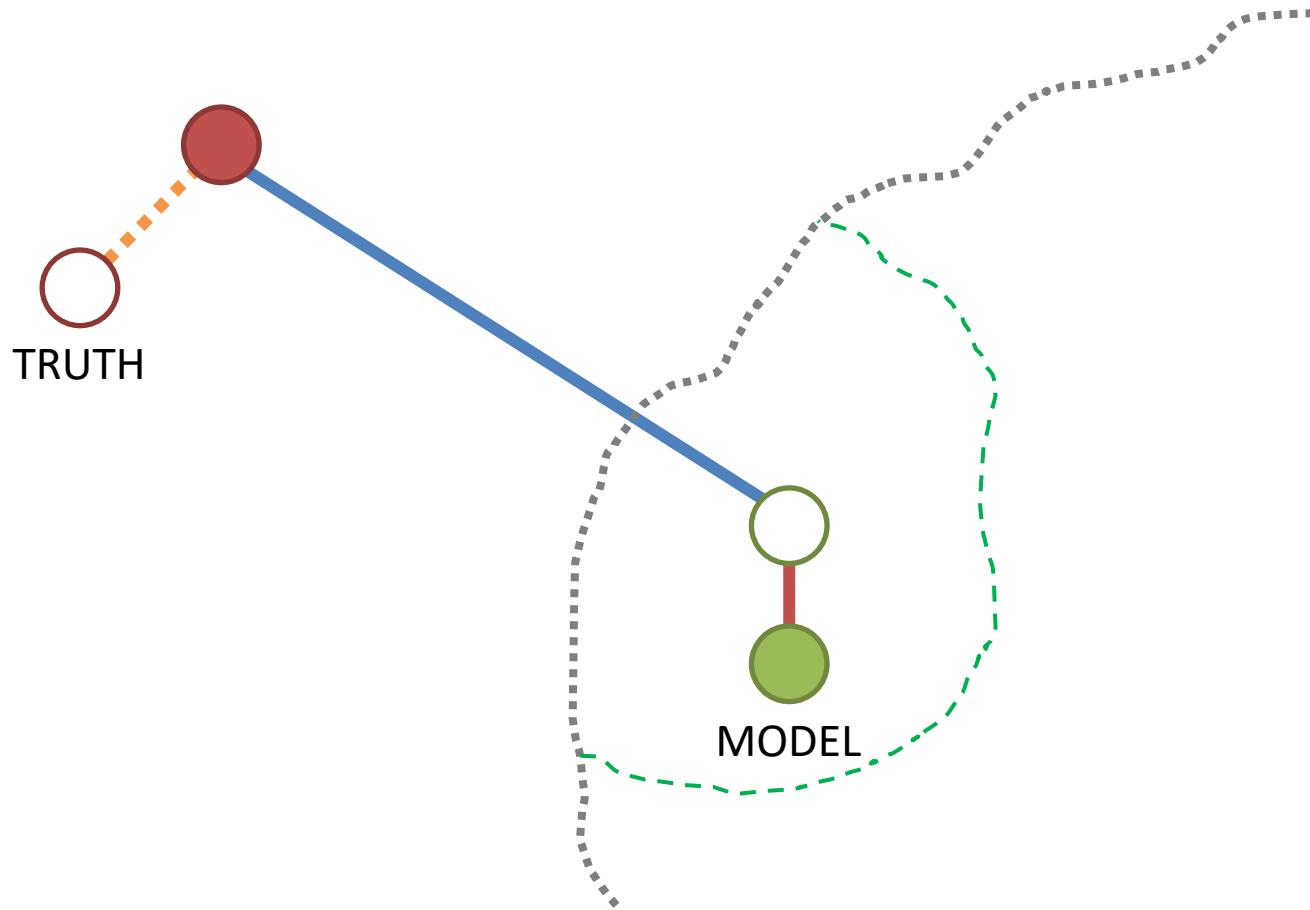
$$E(error^2) = (f - E\hat{f})^2 + E(E\hat{f} - \hat{f})^2 + E(\epsilon)^2$$





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





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



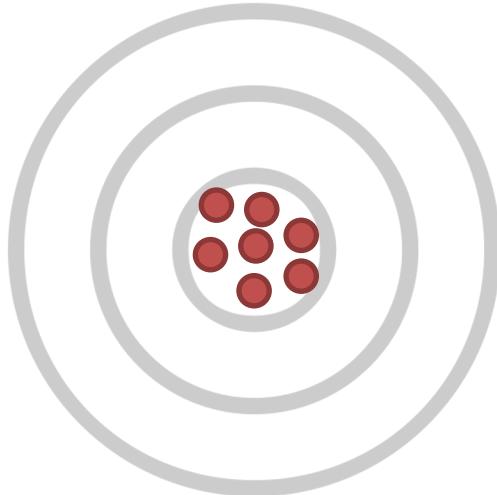
Core principle of Data Science

BIAS-VARIANCE TRADE-OFF

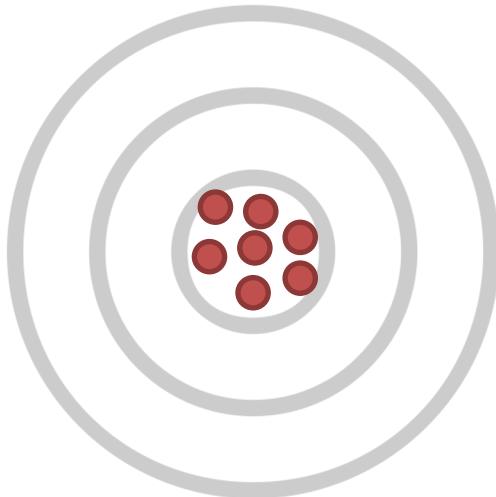




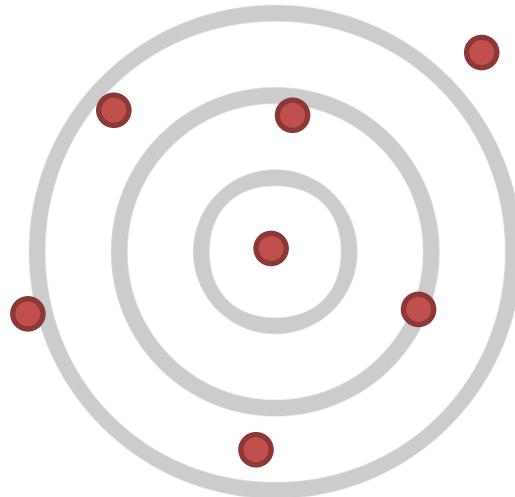
Low Bias
Low Variance



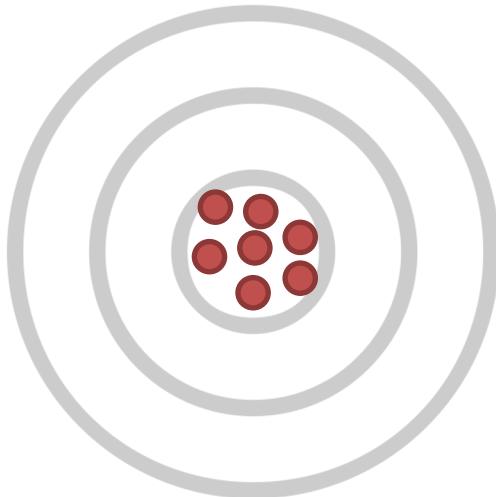
Low Bias
Low Variance



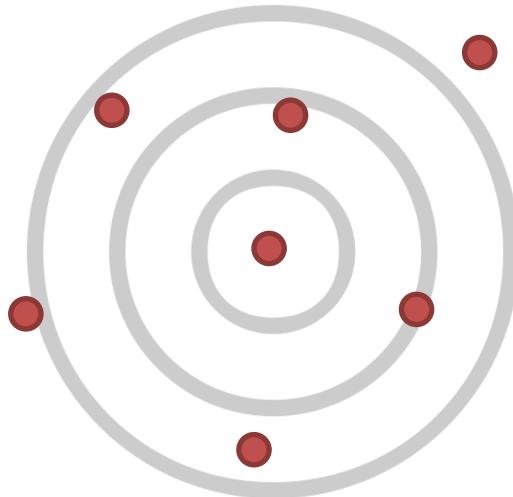
Low Bias
High Variance



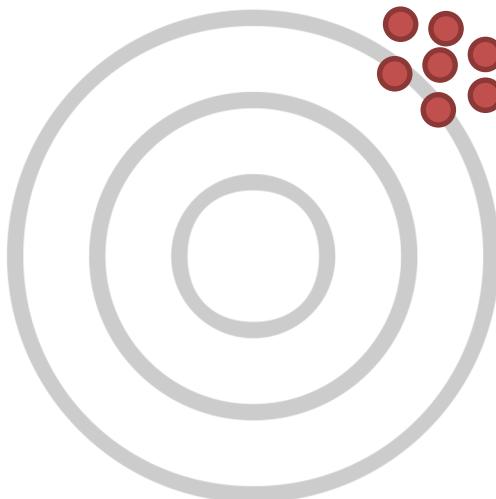
Low Bias
Low Variance



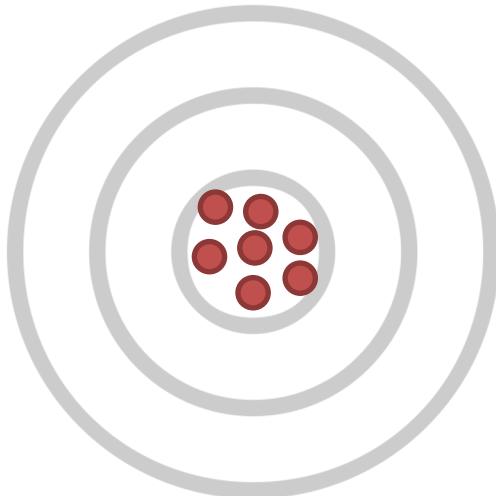
Low Bias
High Variance



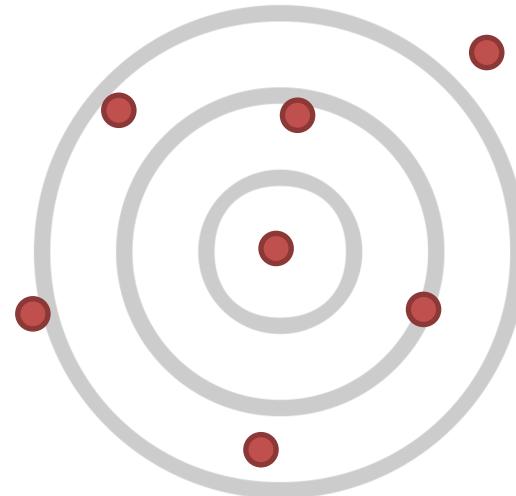
High Bias
Low Variance



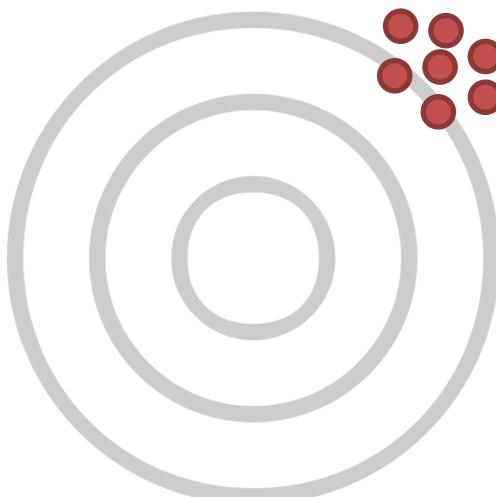
Low Bias
Low Variance



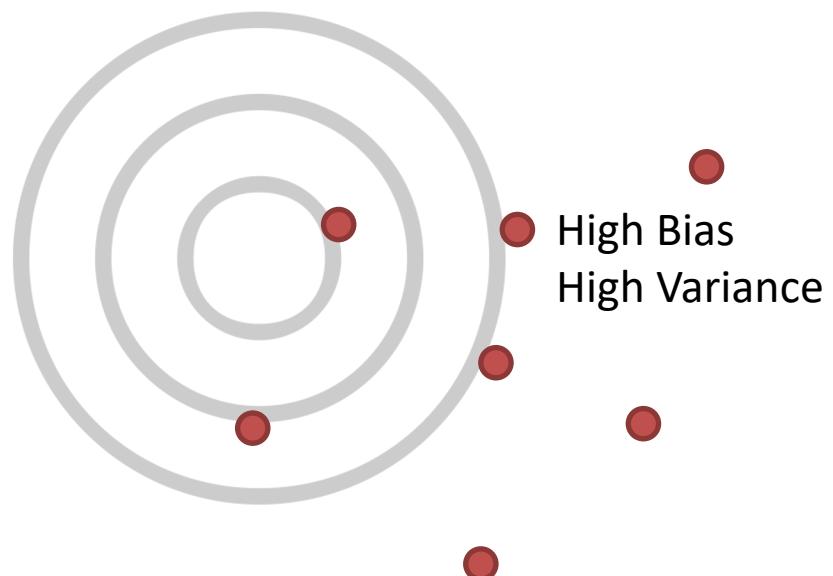
Low Bias
High Variance

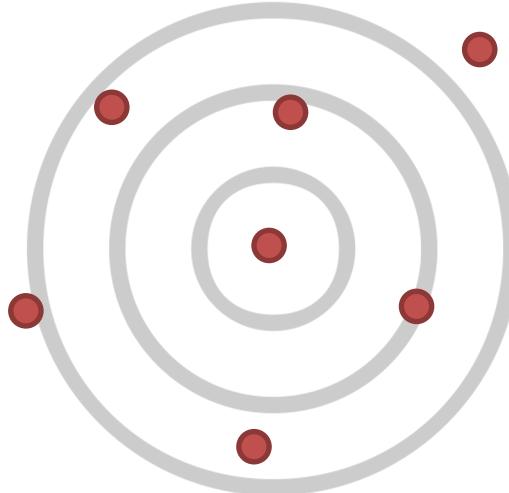


High Bias
Low Variance

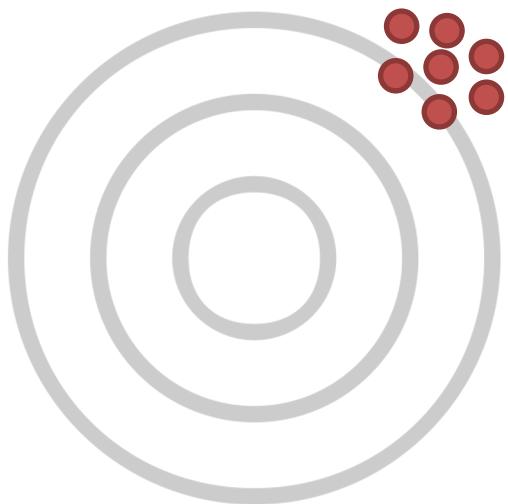


High Bias
High Variance



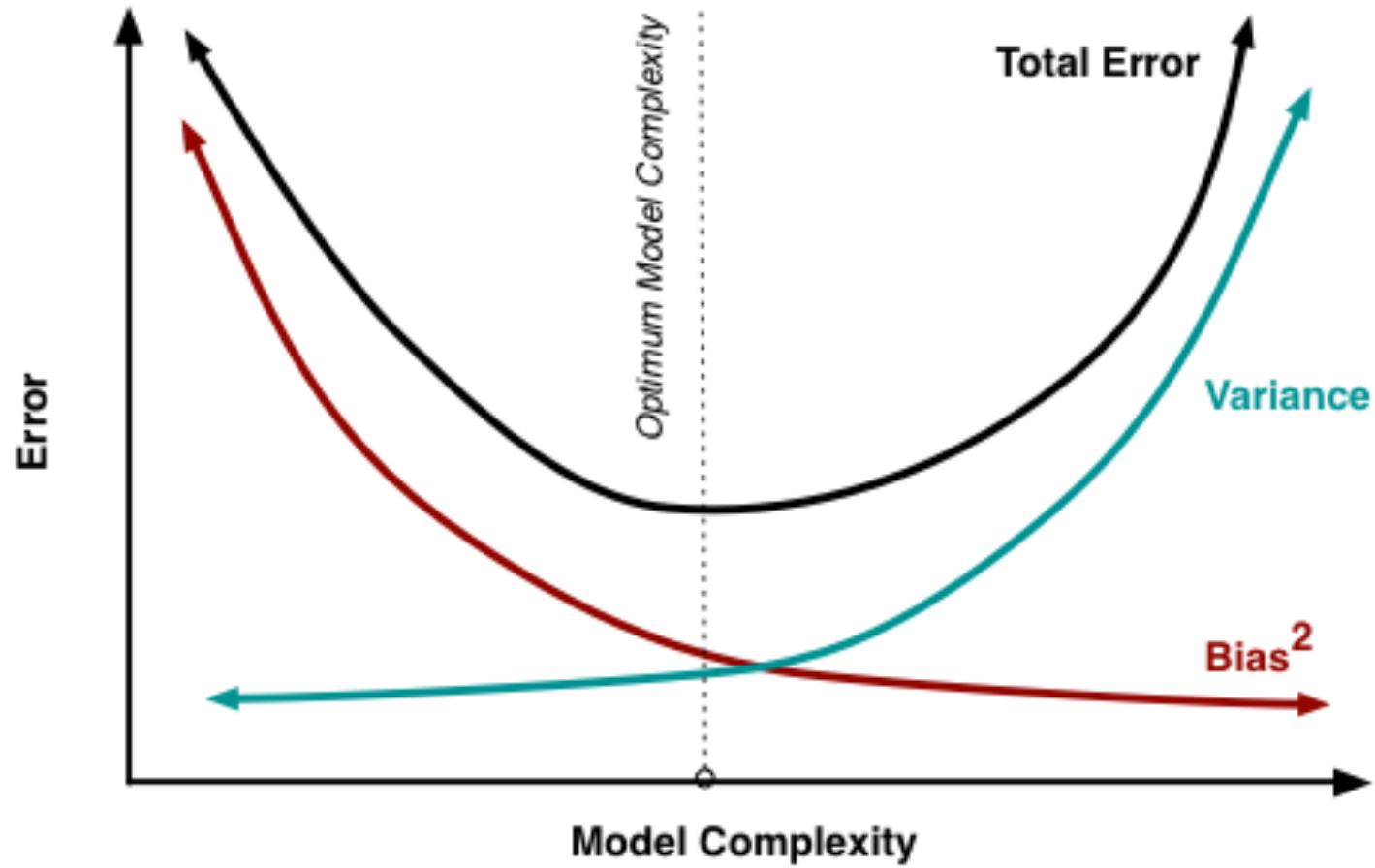


Low Bias
High Variance



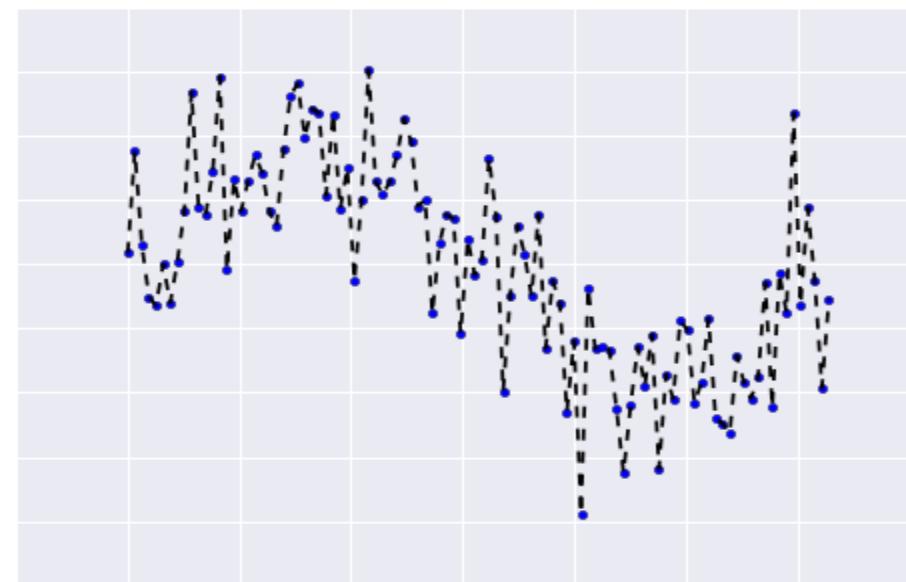
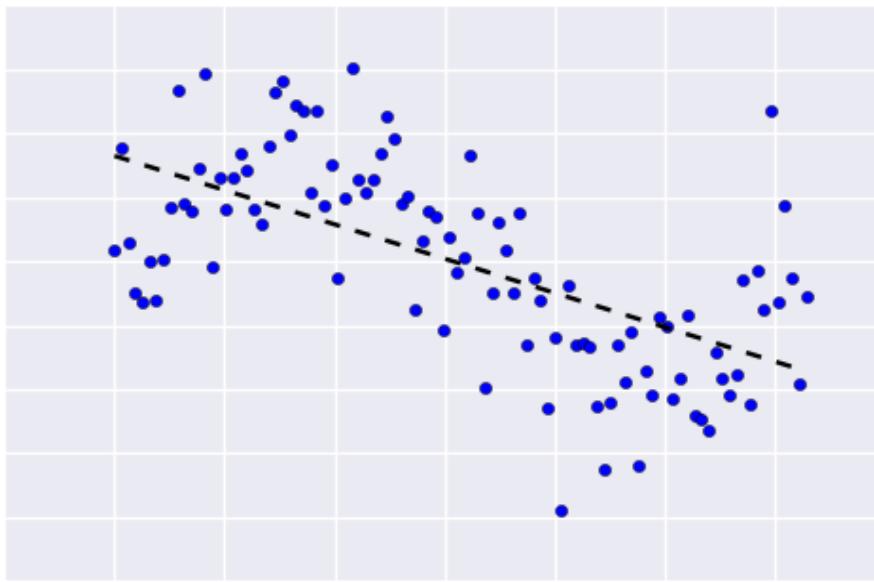
High Bias
Low Variance

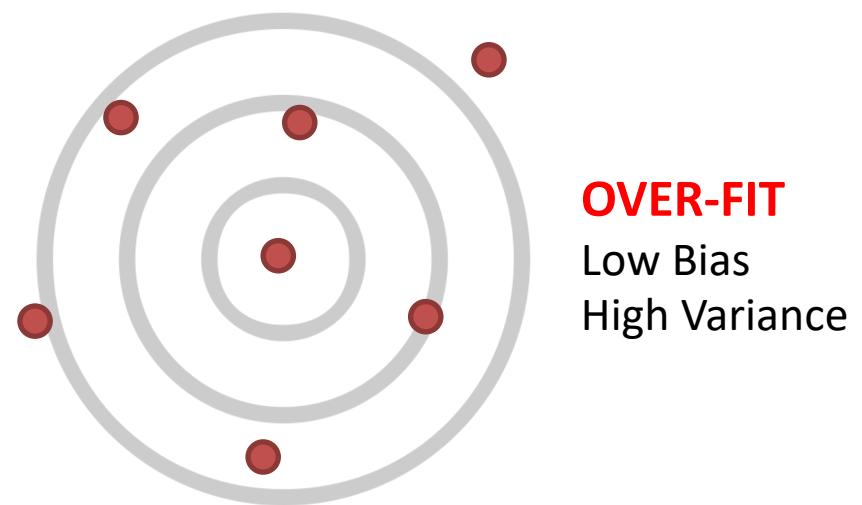
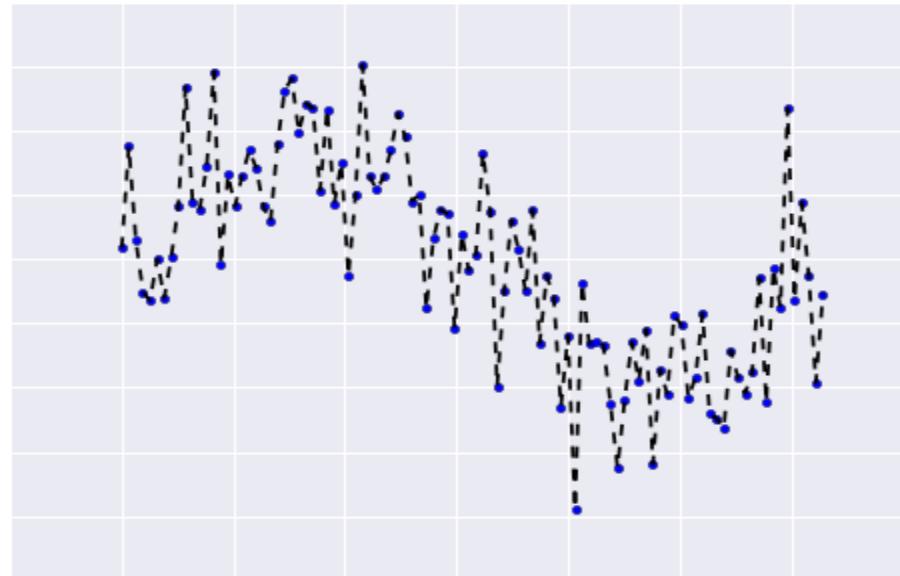
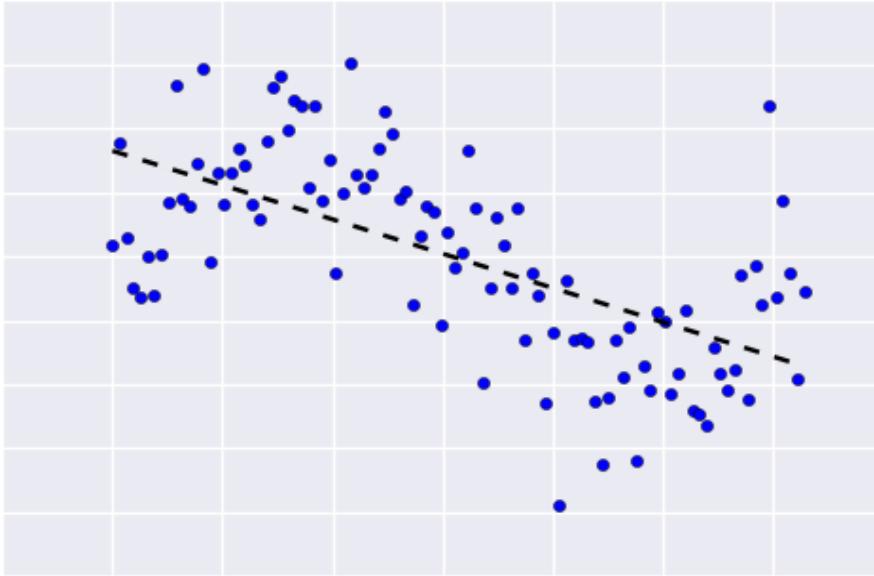


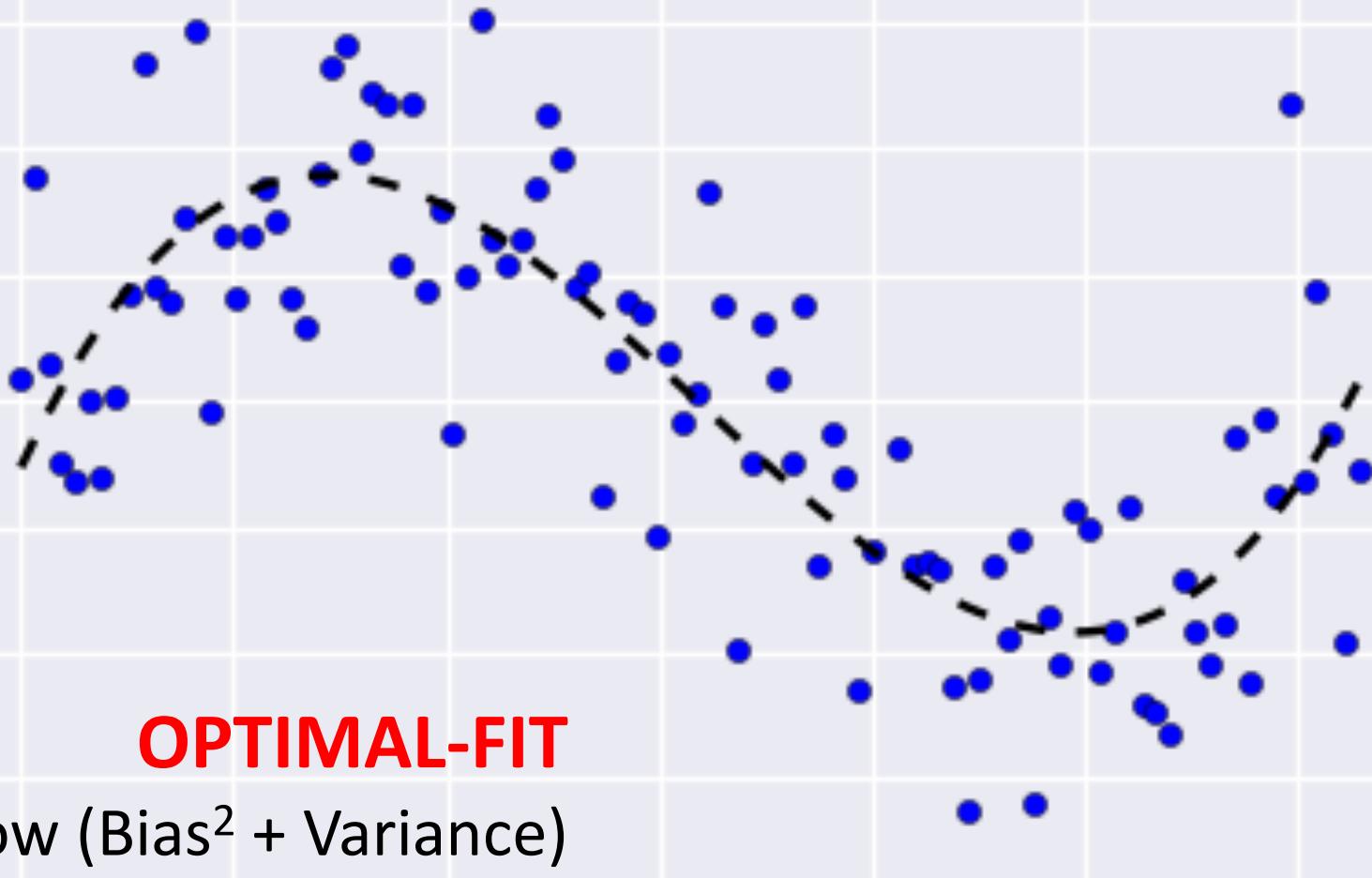


Model Complexity is the Hyper-Parameter to Tune









The essential Data Science tool

CROSS-VALIDATION



Sampled Data

There is a very high chance of Overfitting the given Sample Data



Sampled Data

Training

Test

Test Set should never be used for Model re-training or re-learning



Sampled Data

Training

Test

Training

Validation

Test

Validation Set offers a scope for Model “testing” and re-learning



Sampled Data

Training

Test

Training

Validation

Test

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



General Practice

Training Set : Validation Set :: 70 : 30

Leave-One-Out-Cross-Validation

k-Fold Cross-Validation

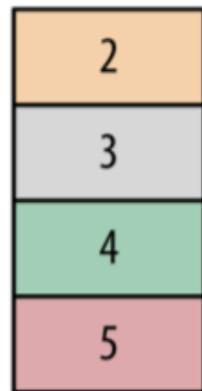




Split the dataset
into k folds/parts

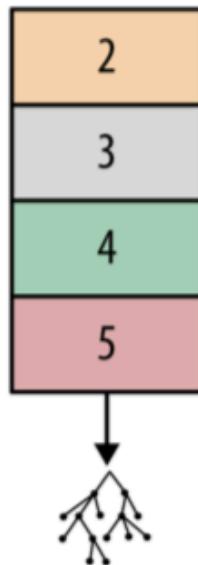


Use $(k-1)$ folds of
data for Training



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

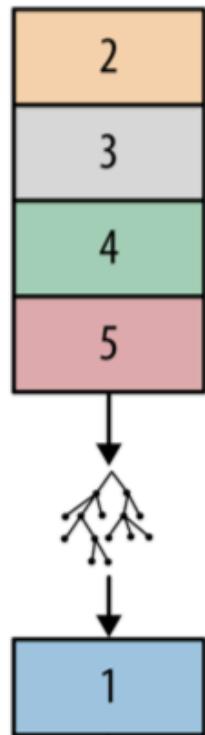
Build the Model



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

Build the Model

Use the remaining
fold for Validation

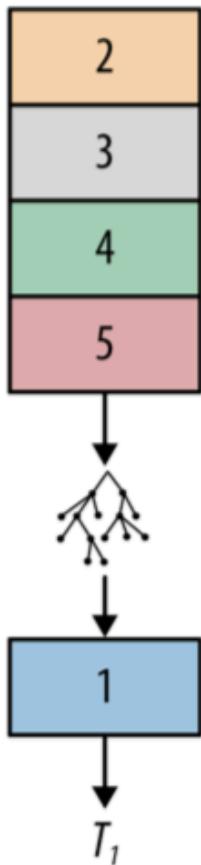


Use $(k-1)$ folds of
data for Training

Build the Model

Use the remaining
fold for Validation

Note Performance



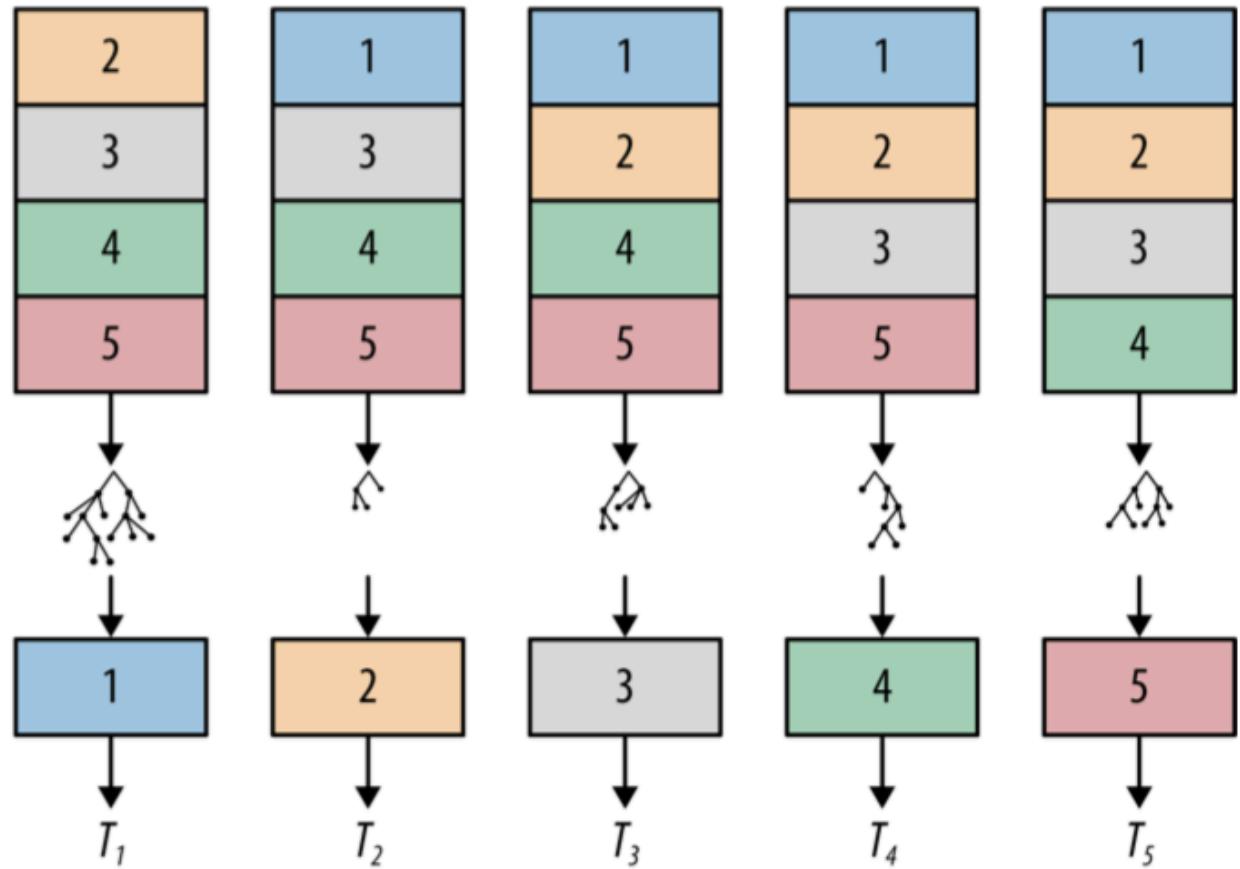
Repeat the process k times

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

Build the Model

Use the remaining fold for Validation

Note Performance



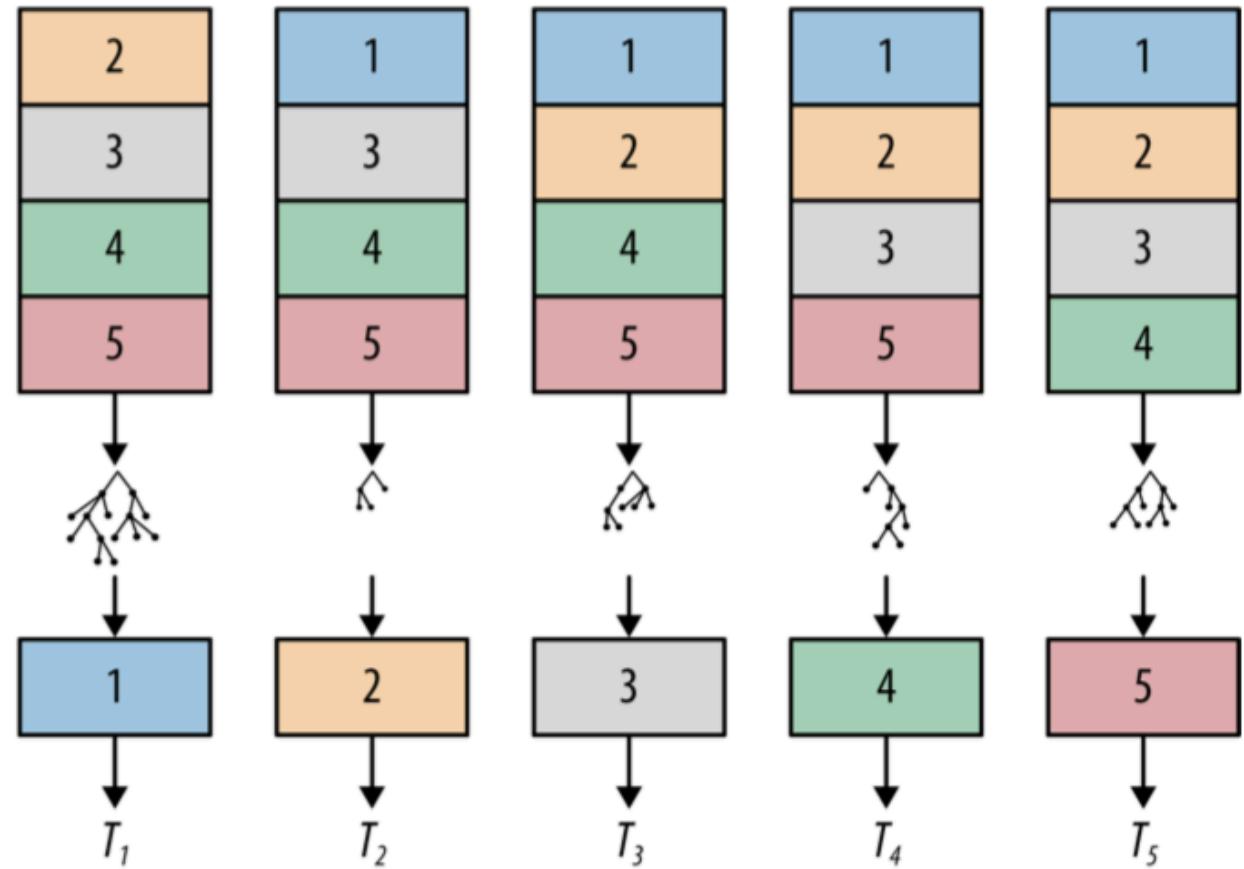
Repeat the process k times

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

Build the Model

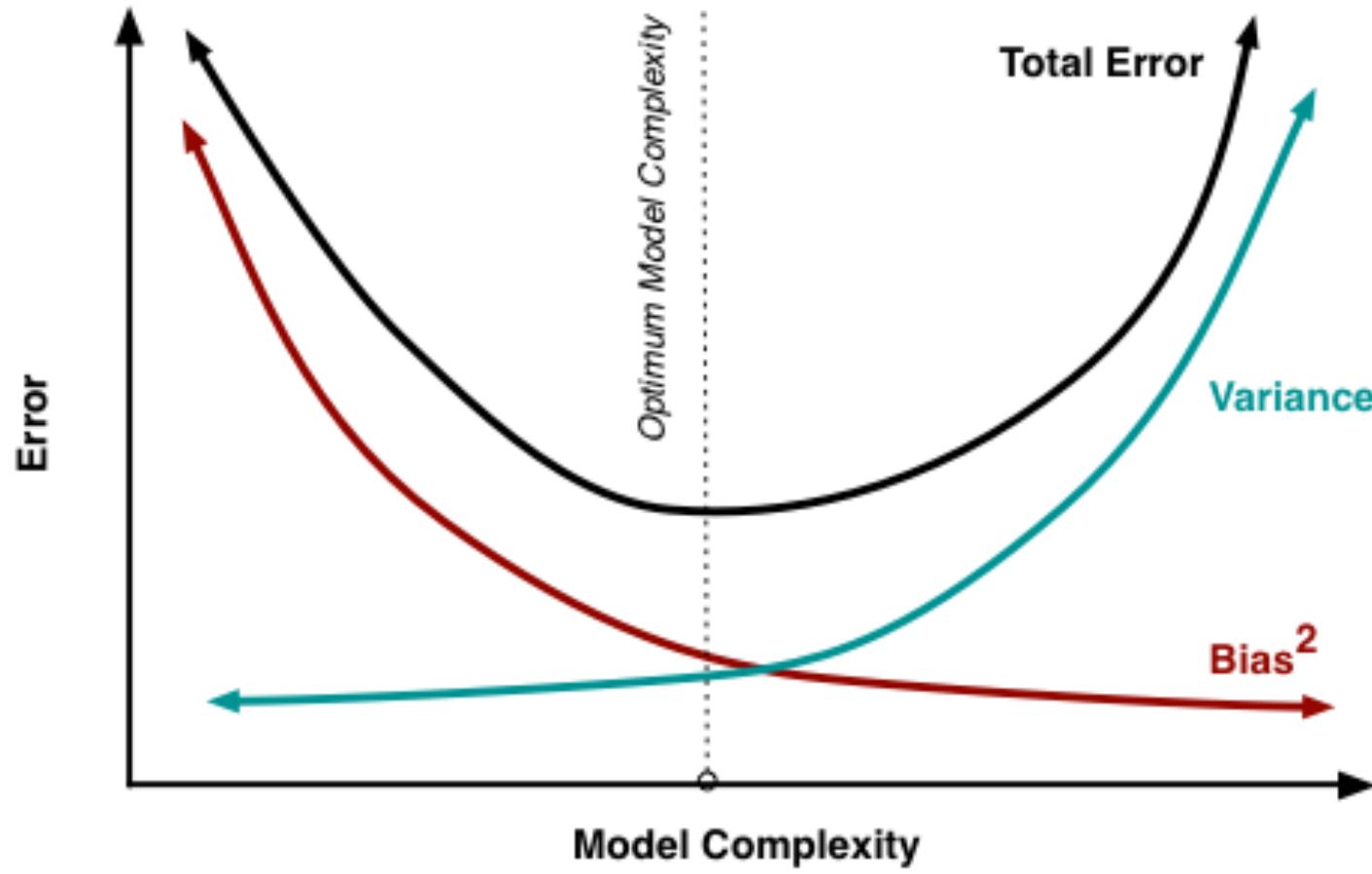
Use the remaining fold for Validation

Note Performance



Note the Mean Performance and its Variance





Understanding optimality of Model Complexity is critical in Data Science!

