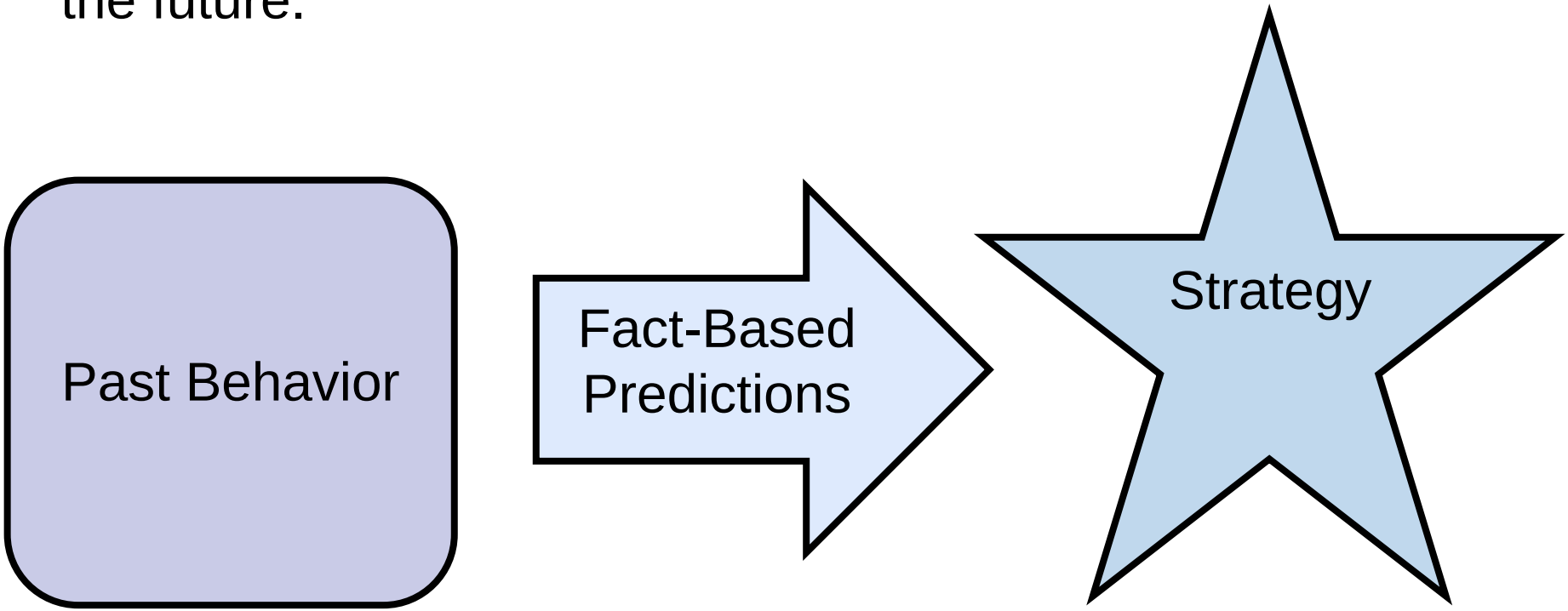
MODEL BUILDING & SCORING FOR PREDICTION

Institute for Advanced Analytics

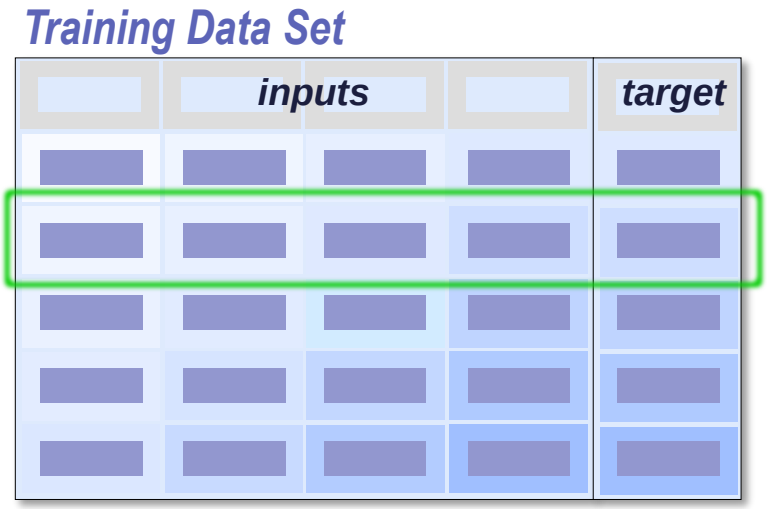
MSA Class of 2020

From Descriptive to Predictive Modeling

Predictive modeling techniques, paired with scoring and good model management, enable you to use your data about the past and the present to make good decisions for the future.



Predictive Modeling Terminology



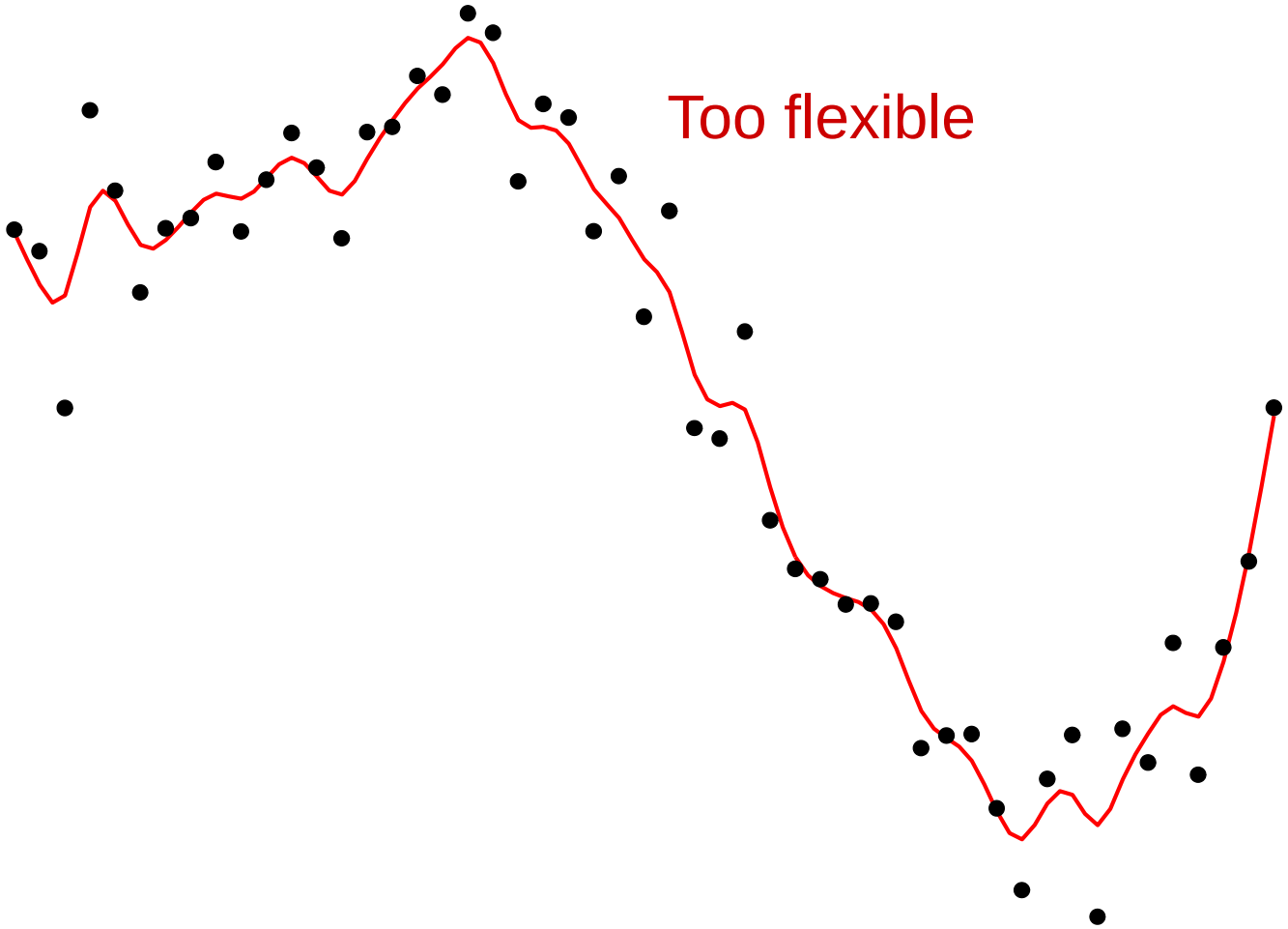
The variables are called *inputs* and *targets.*

The observations in a training data set are known as *training cases*.

Model Complexity

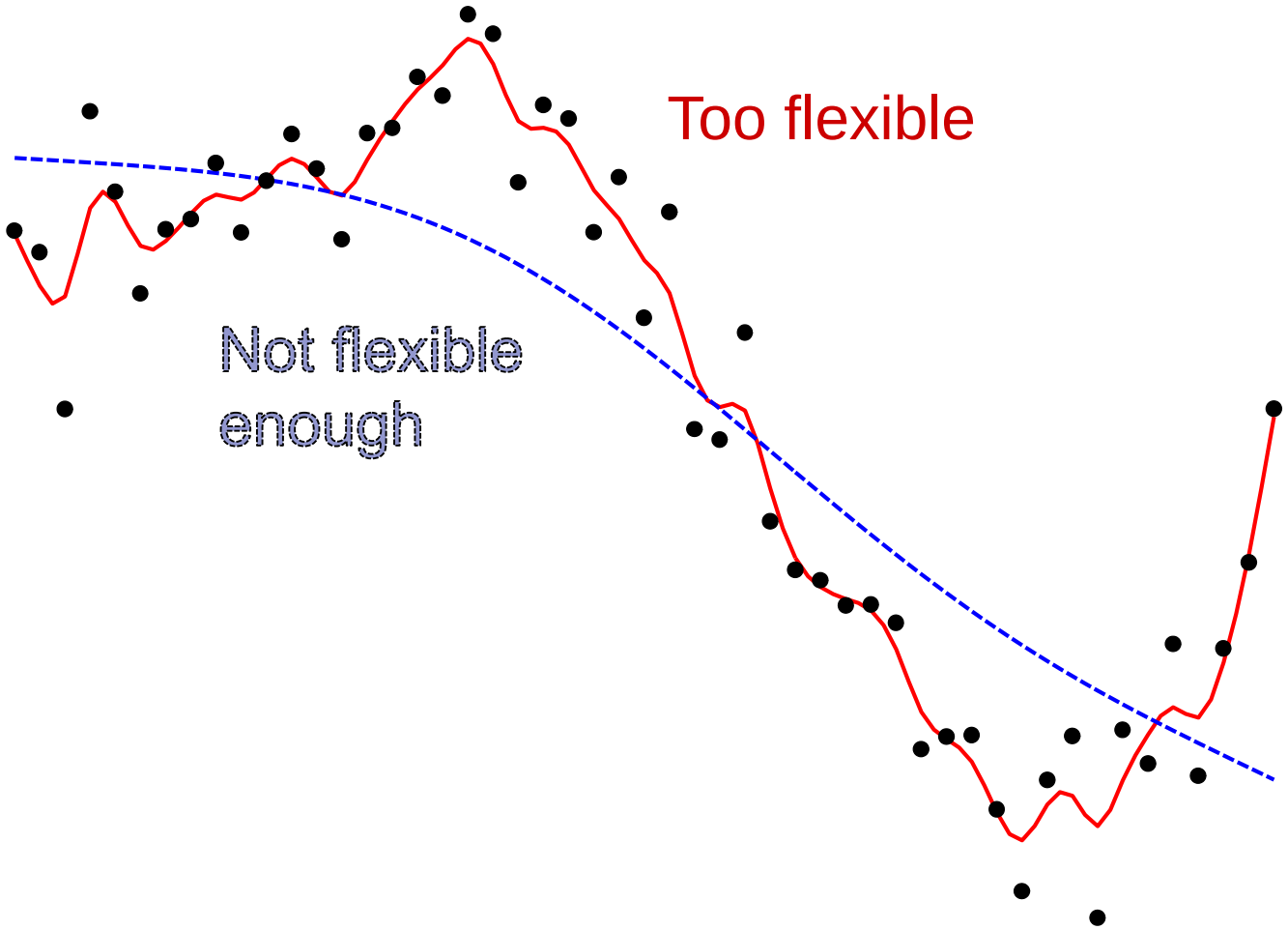
...

Model Complexity



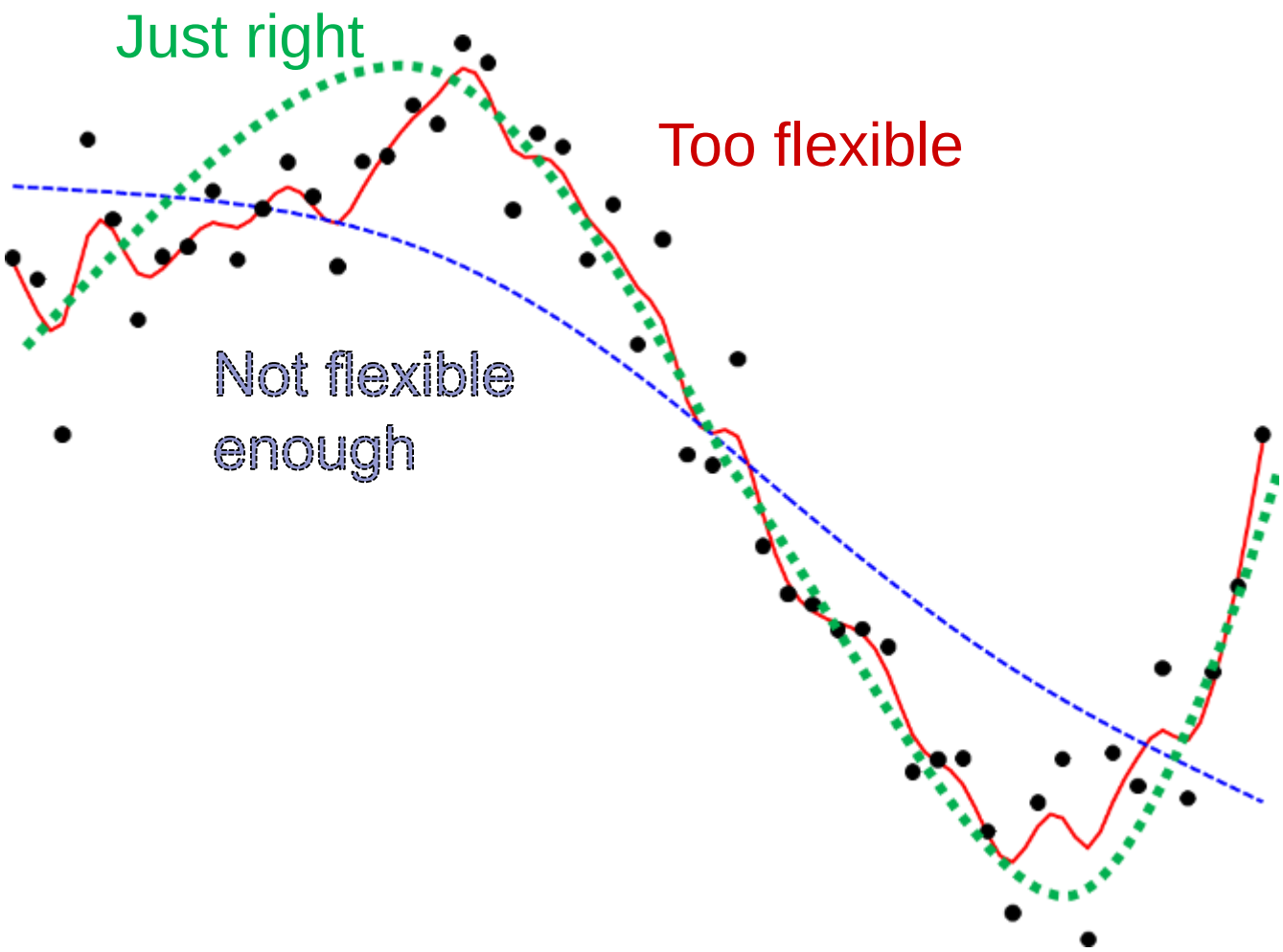
...

Model Complexity



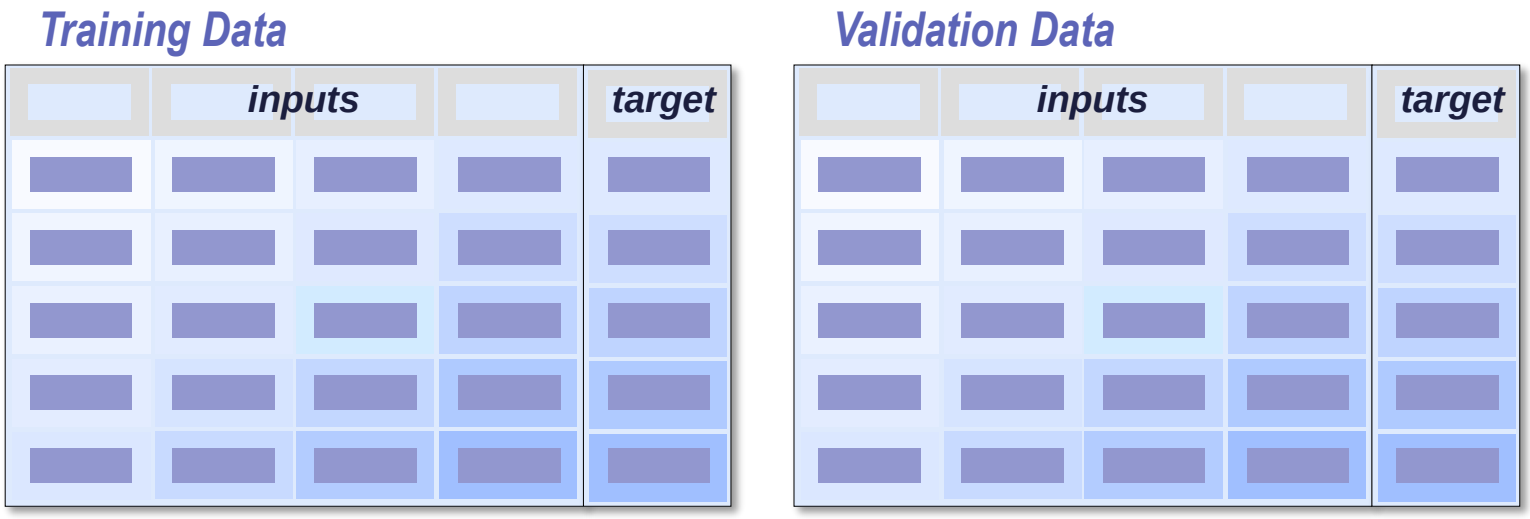
...

Model Complexity



DATA PARTITIONING

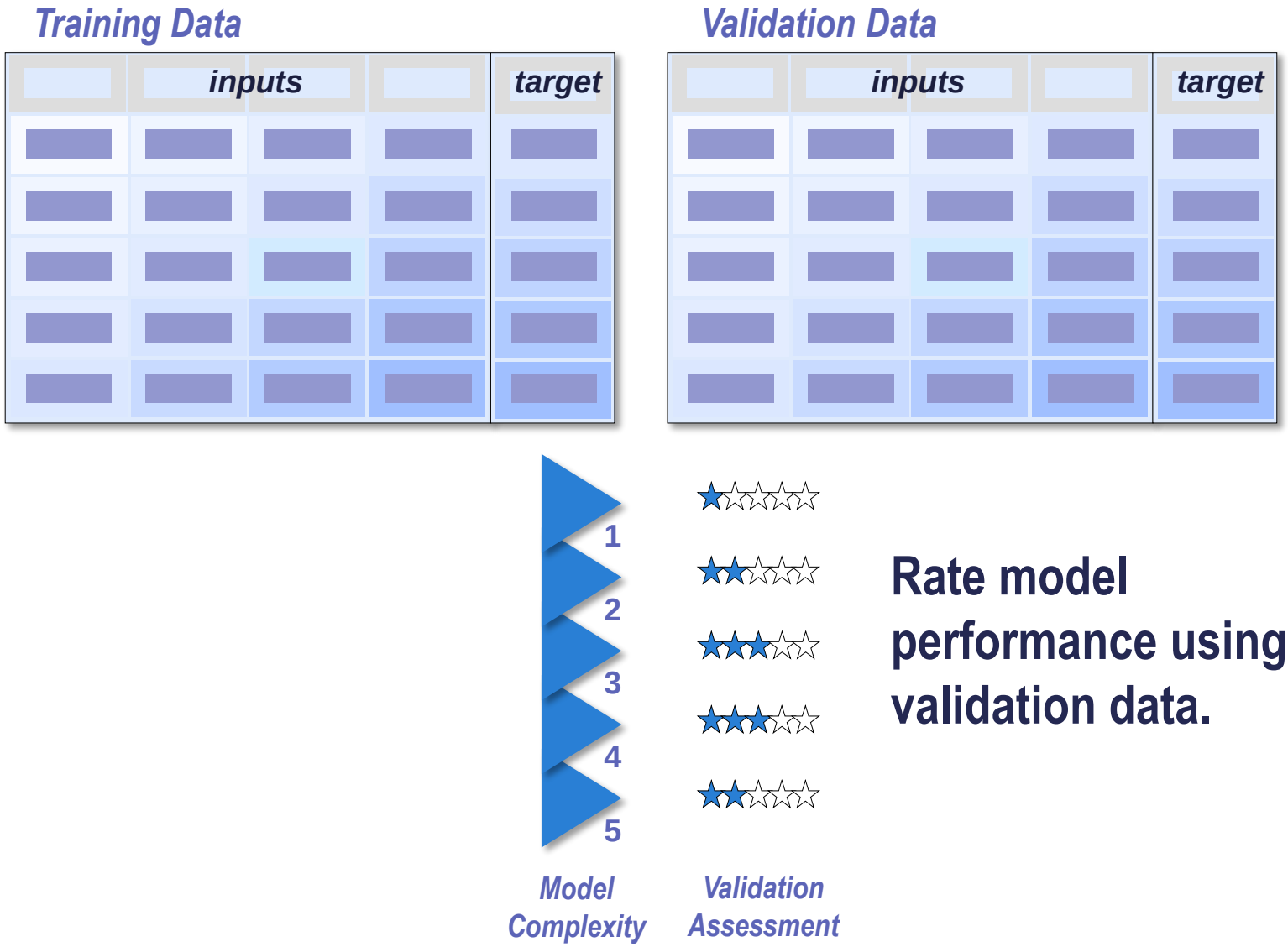
Honest Assessment and Data Partitioning



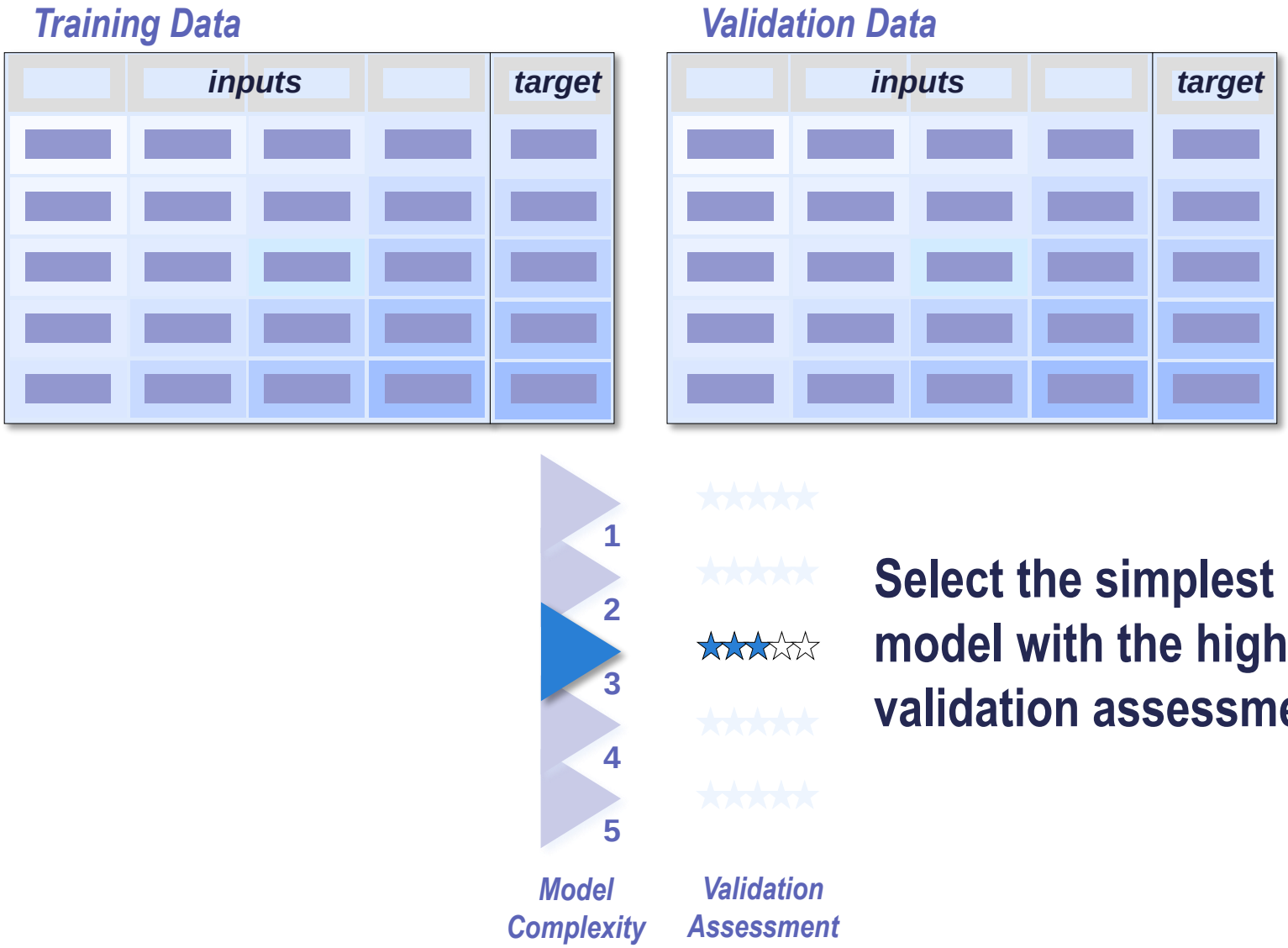
Partition available data into training and validation sets.

The model is fit on the training data set, and model performance is evaluated on the validation data set.

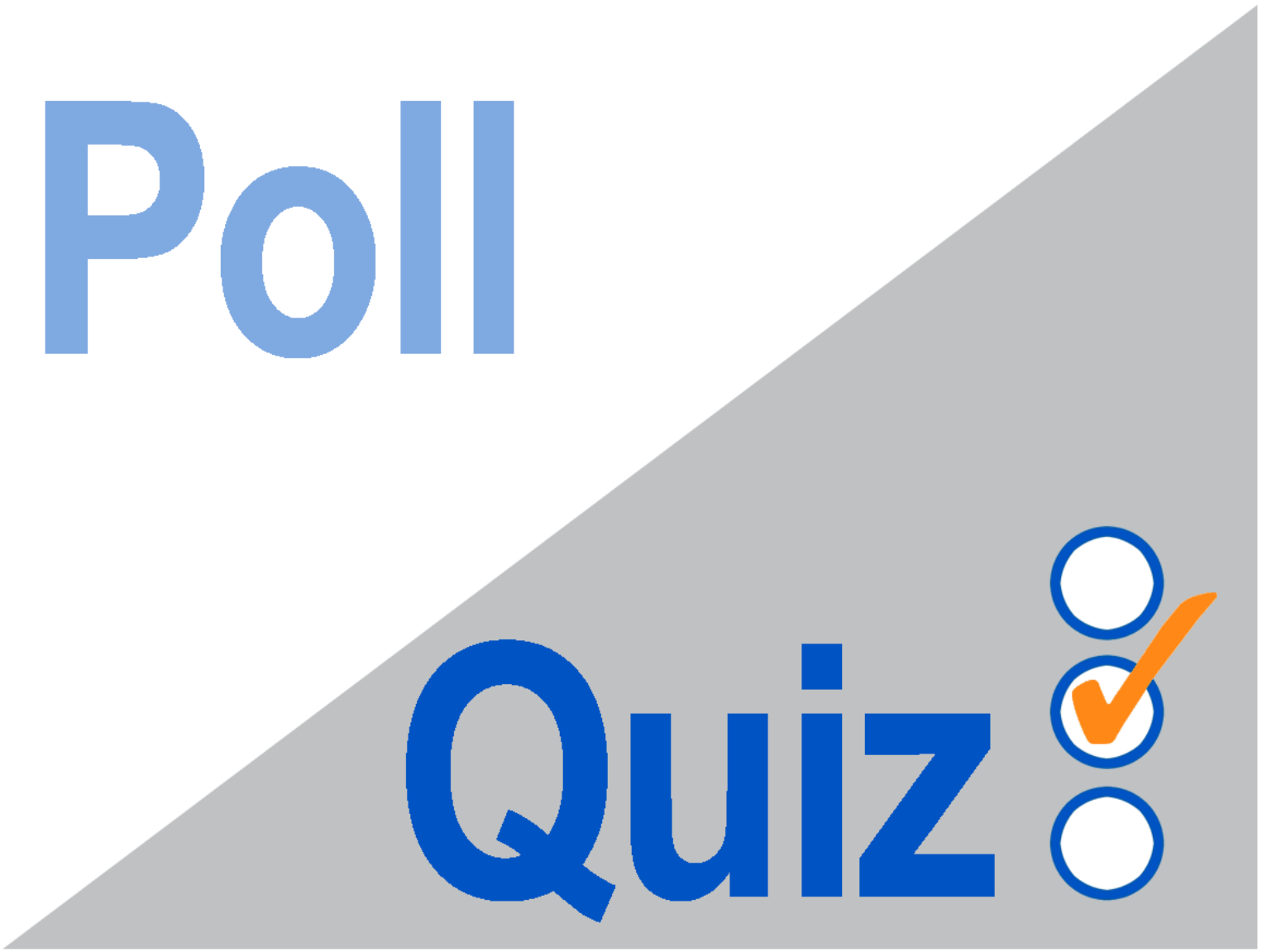
Model Performance Assessment



Model Selection



Poll



Multiple Choice Poll

- When using honest assessment, which of the following would be considered the best model?

a. The simplest model with the best performance on the training data
b. The simplest model with the best performance on the validation data
c. The most complex model with the best performance on the training data
d. The most complex model with the best performance on the validation data

Multiple Choice Poll – Correct Answer

- When using honest assessment, which of the following would be considered the best model?

a. The simplest model with the best performance on the training data

b. The simplest model with the best performance on the validation data

c. The most complex model with the best performance on the training data

d. The most complex model with the best performance on the validation data

Validation Data Set with Data Step

```
data ameshousing3_train ameshousing3_valid;
     set bootcamp.ameshousing3;
     random = RAND("Uniform");
     if random <=0.2 then output ameshousing3_valid;
     else output ameshousing3_train;
run;
```

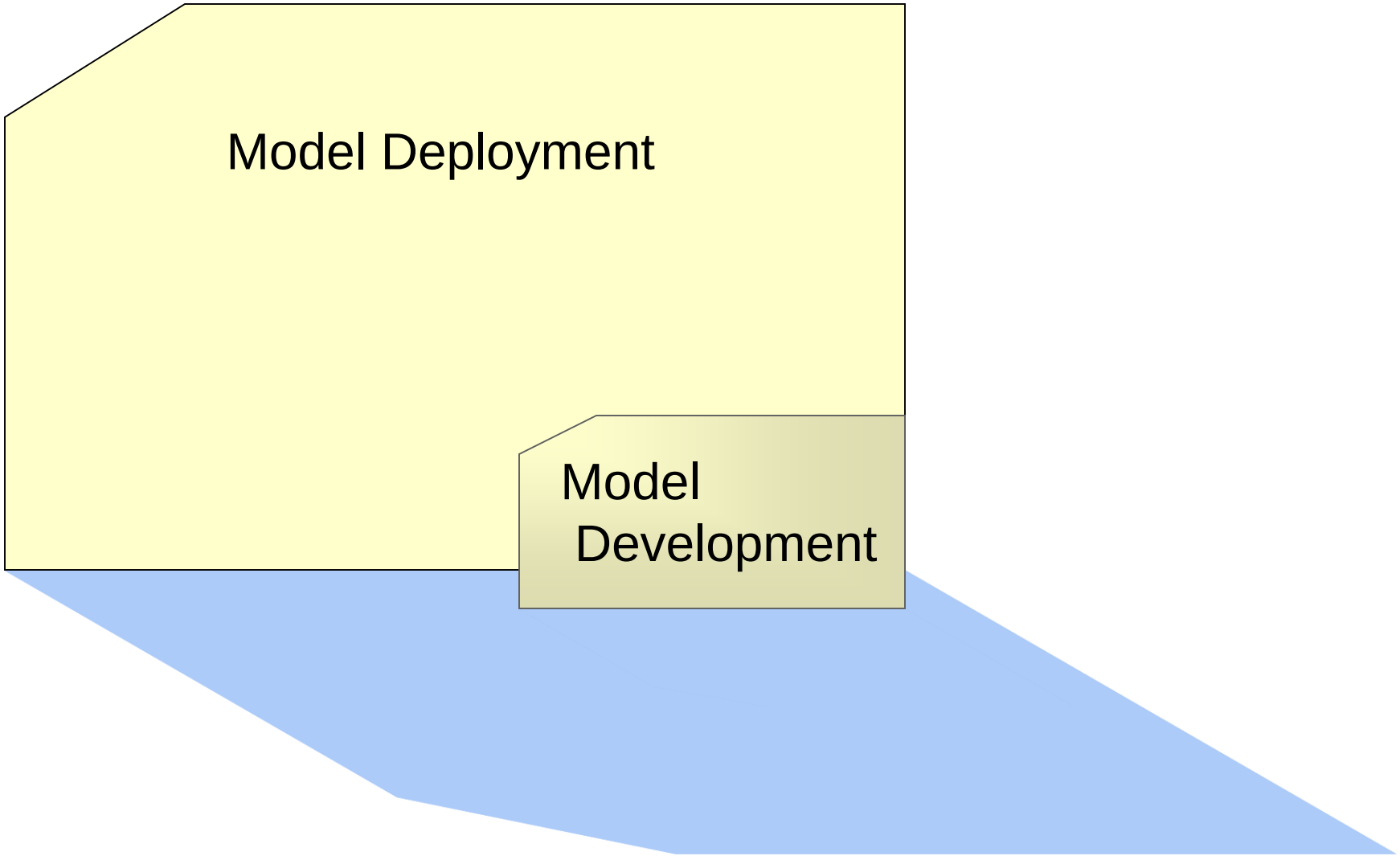
Validation Data Set with PROC's

```
proc surveyselect data=bootcamp.ameshousing3
                  method=srs rate=0.2
                  out=ameshousing3_split outall;
run;

data ameshousing3_train ameshousing3_valid;
     set ameshousing3_split;
     if Selected = 1 then output ameshousing3_valid;
     else output ameshousing3_train;
run;
```

PREDICTION / SCORING

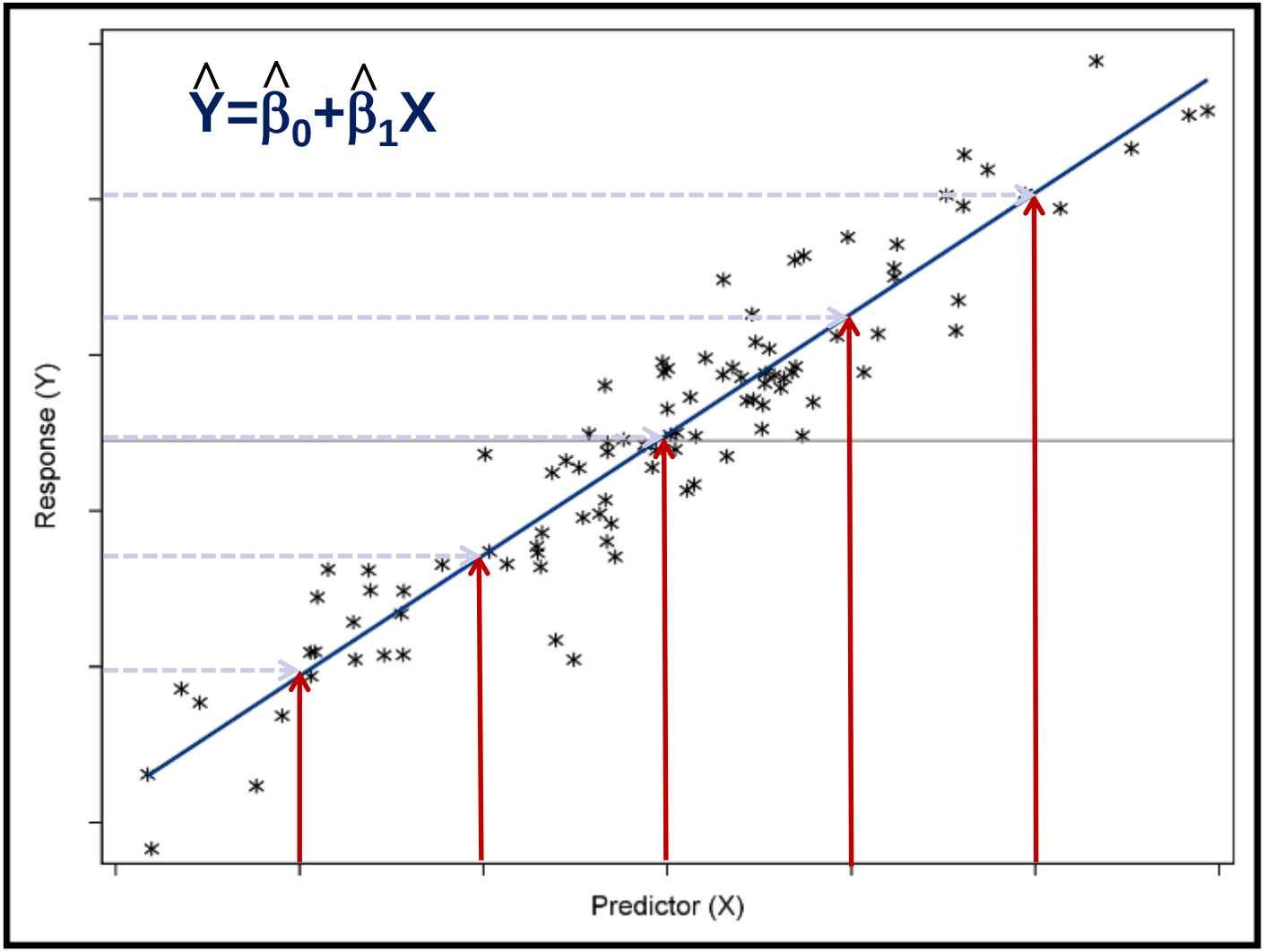
Scoring



Scoring Recipe

- The model results in
  a formula or rules.
- The data require modifications.
  - Derived inputs
  - Transformations
  - Missing value imputation
- The scoring code is deployed.
  - To score, you do not
    rerun the algorithm;
    apply score code (equations) obtained from the final model to the scoring data.

Producing Predicted Values



Model Diagnostic Statistics

1. Mean Absolute Percent Error:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|$$

2. Mean Absolute Error:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right|$$

Model Diagnostic Statistics

1. Mean Absolute Percent Error:

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|$$

→ Problems:
- Overweight of Over-predictions
- Actual of 0

2. Mean Absolute Error:

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}_t\right|$$

→ Problems:
- Not scale invariant

Scoring Training Data Set

```
proc reg data=ameshousing3_train outest=Betas;
     model SalePrice = Basement_Area Lot_Area;
     title "Model with Basement Area and Lot Area";
     output out=Scored predicted = pred;
     store out = model;
run;
quit;

data MAPE_t;
     set Scored;
     AE = abs(pred - SalePrice);
     APE = (abs(pred - SalePrice) / SalePrice)*100;
run;

proc means data=MAPE_t mean;
              var AE APE;
run;
```

Scoring Training Data Set

Model with Basement Area and Lot Area

The MEANS Procedure

Variable	Mean
AE	21049.16
APE	17.1034523

Scoring Validation Data – PROC SCORE

```
proc score data=ameshousing3_valid score=Betas
           out=Scored type=parms;
    var Basement_Area Lot_Area;
run;

data MAPE_v;
    set Scored;
    AE = abs(Predicted - SalePrice);
    APE = (abs(Predicted - SalePrice) / SalePrice)*100;
run;

proc means data=MAPE_v mean;
    var AE APE;
run;
```

Scoring Validation Data – PROC SCORE

Model with Basement Area and Lot Area

The MEANS Procedure

Variable	Mean
AE	21101.76
APE	18.3498292

Scoring Validation Data – PROC PLM

```
proc plm restore=model;
     score data = ameshousing3_valid out = Scored;
run;

data MAPE_v;
     set Scored;
     AE = abs(Model1 - SalePrice);
     APE = (abs(Model1 - SalePrice) / SalePrice)*100;
run;

proc means data=MAPE_v mean;
     var AE APE;
run;
```

Scoring Recipe

- Not all types of models can be scored with PROC SCORE or PROC PLM.
- Other PROC's have SCORE statements to score new observations directly.
- Data Step is another traditional way of scoring observations – SELF STUDY CODE FOLLOWS

Predicted Values with Data Step

```
data ameshousing3_split2;
    set ameshousing3_split;
    if Selected = 1 then SalePrice = .;
run;

proc reg data=ameshousing3_split2;
    model SalePrice = Basement_Area Lot_Area;
    title "Model with Basement Area and Lot Area";
    output out = Scored predicted=pred;
    title 'Sale Price Regression';
run;
quit;
```

Predicted Values with Data Step

```
data Scored;
    set Scored;
    if SalePrice ne . then delete;
run;

data Scored;
    merge Scored ameshousing3_valid;
    keep SalePrice Pred;
run;

data MAPE_v;
    set Scored;
    AE = abs(Pred - SalePrice);
    APE = (abs(Pred - SalePrice) / SalePrice)*100;
run;

proc means data=MAPE_v mean;
    var AE APE;
run;
```