

# MODELING FRAMEWORK



# Outline

- Machine Learning
- Prediction vs. Inference
- Accuracy vs. Interpretability
- Regression and Classification problems
- Overfitting
- Bias-Variance Tradeoff
- The Model
- Inferential Statistics

# Artificial Intelligence

- Early AI used Rules-Based systems to mimic human intelligence
- Consider a Rules-Based system for approving Loan Applications
  - *If FICO > 680 and Credit History > 3 years, then Approve*
  - *If FICO > 680 and Credit History < 3 years, then Request more information*
  - *If FICO < 680, then Reject*
- This works if the number of rules is small.
  - *Rules-based system worked for Chess but not for Go.*

# Machine Learning

- Machine learning is a family of techniques where these rules are determined from data and then can be applied to previously unseen situations.
- It has been argued, *machine learning* is really about *learning from data*
- See an interesting illustration in this [clip from the movie Groundhog Day](#).

# Machine Learning vs. Rules-based System

- Is it a Dog or a Cat?



# Machine Learning

Draws from many disciplines

- Math and Statistics
  - *Draw inferences*
  - *Estimate models*
- Computer science
  - *Algorithms for enabling analytical techniques*
  - *Efficient, scalable computing*
- Application Domain: Finance, Geography, Genomics, Marketing, Physics,..

# Machine Learning (ML) vs. Traditional Econometrics

1. ML is focused on the best out-of-sample predictions. Traditional econometric methods are aimed at deriving unbiased estimators. Latter do not perform well out-of-sample.
2. ML approaches are useful even in the absence of *a priori* theory. For instance, consider the image recognition problem where the goal is to recognize the object in a picture and data are pixels. There is no model for how the pixels combine to make an image of, say, a dog or a house.
3. ML methods can handle an extremely large number of variables. For e.g., predicting whether a user will click on an ad is extremely complex and affected by several factors.
4. ML methods apply feature selection and optimization to achieve scale and efficiency.

Source: Daria Dzyabura and Hema Yoganarasimhan (2018) "Machine Learning" chapter for Handbook of Marketing Analytics: Methods and Applications in Marketing, Public Policy, and Litigation Support , editors Dominique Hanssens and Natalie Mizik, January, 2018

# Machine Learning

- Predictors (also known as Inputs, Features, or Independent Variables)
  - Denoted as  $X$
- Outcome (also known as Output, Response, or Dependent Variable)
  - Denoted as  $Y$
- $Y = f(X) + \varepsilon$
- Machine Learning is a set of approaches for using data to determine the functional relationship ( $f$ ) between predictor(s) ( $X$ ) and outcome ( $Y$ )

# Machine Learning

- Supervised Learning
  - *Data is labeled including an outcome and predictor(s)*
  - *Goal is to map predictors to outcome*
  - *E.g., regression, trees*
- Unsupervised Learning
  - *Data is unlabeled*
  - *Goal is to find interesting patterns among variables without the use of any outcome variable.*
  - *E.g., cluster analysis, factor analysis, market basket analysis*
- This course will review only Supervised Learning methods

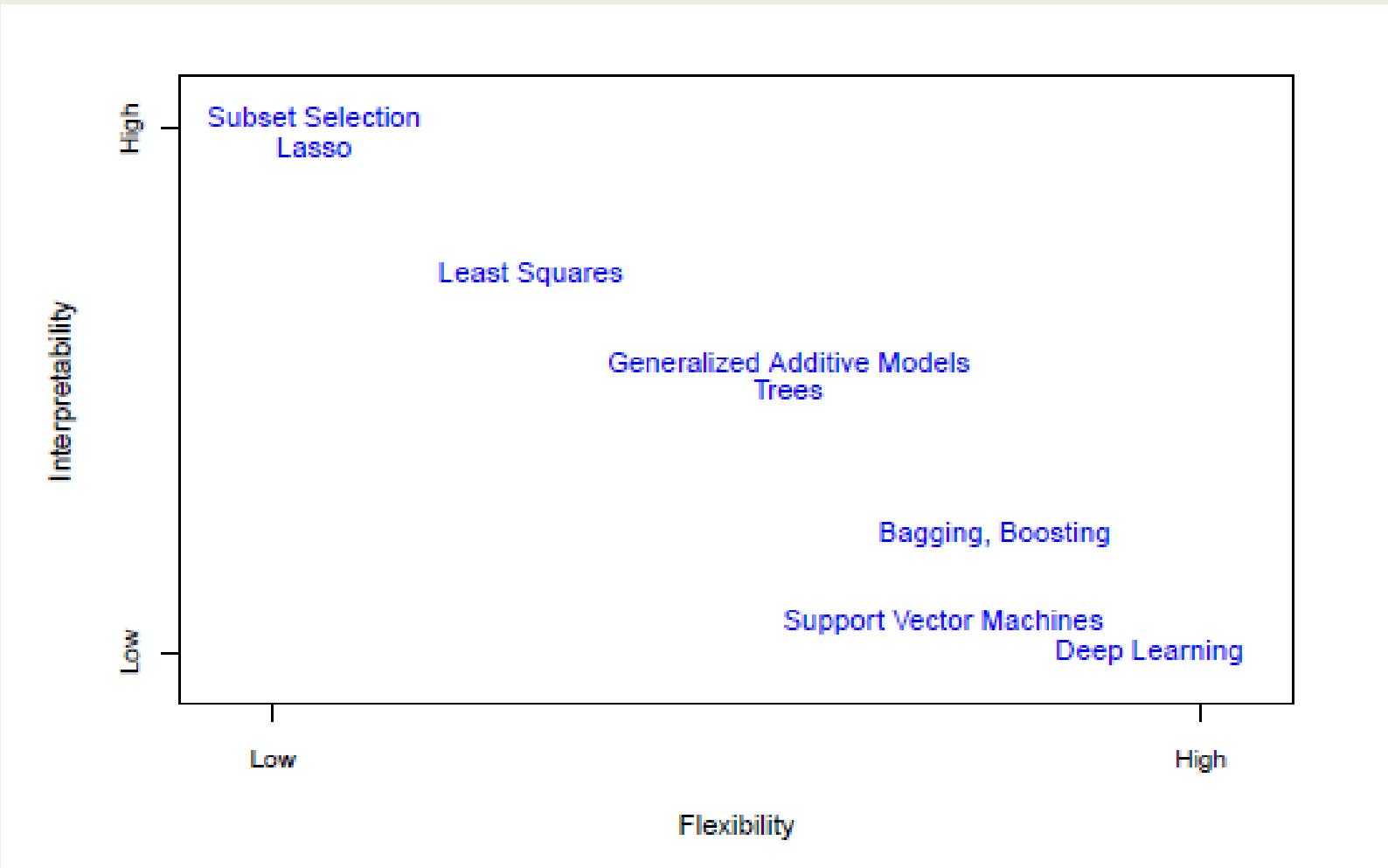
# Supervised Learning

- Consider the model
  - $\text{House\_Sale\_Price} = f(\text{Area}, \text{Age}, \text{Number\_of\_Bathrooms}, \text{Month\_of\_Listing})$
- Prediction
  - Goal is to generate accurate predictions of House\_Sale\_Price
  - Prediction Error ( $\varepsilon$ ) = Reducible Error + Irreducible error ( $\text{Var}(\varepsilon)$ )
  - Techniques discussed in this class aim at estimating  $f$  with the aim of minimizing the reducible error
- Inference
  - Determine predictors associated with House\_Sale\_Price
  - Determine nature of relationship (e.g., valence, i.e., positive or negative; functional form such as linear or non-linear)

# Prediction vs. Inference

- Many problems are predominantly interested in only one of the two goals.
  - *New product development (Inference): Which product features influence sales and by how much?*
  - *Customer Targeting (Prediction): Using demographics and online behavior, predict which customers will click on the link in an email?*
  - *Of course, there are a few situations where both are of interest*
- Techniques that favor one don't do so well at the other
  - *Models with the lowest prediction errors are generally hard to interpret*
  - *Flexible models are generally better for predictions while restrictive methods are better for explaining phenomena*

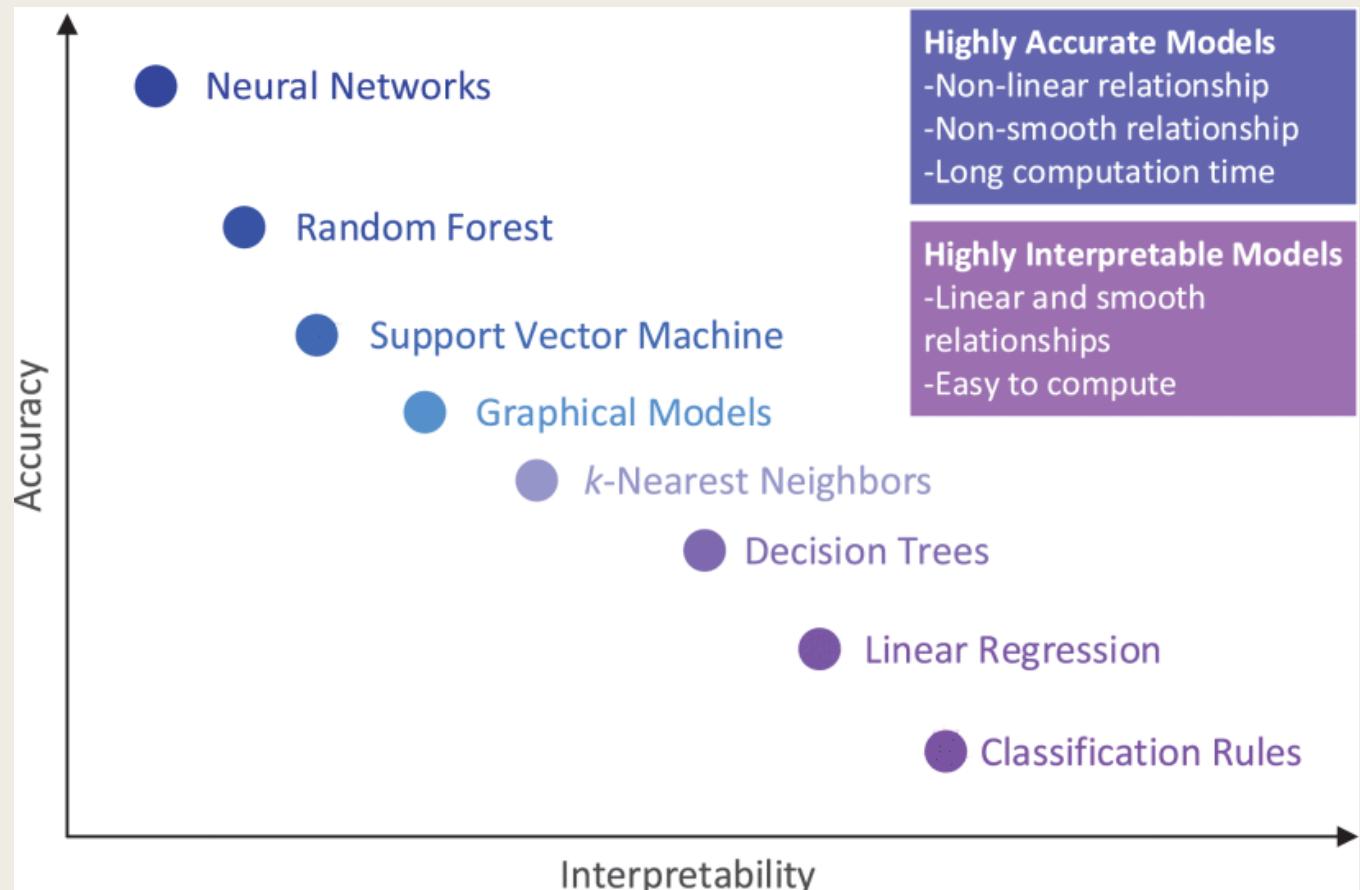
# Prediction vs. Inference



Source: Introduction to Statistical Learning by James et al (2021)

# Interpretability

- A model is interpretable if its structure reveals the relationship between predictors and outcome. E.g. Linear regression, GAMs, logistic regression, decision trees.
- On the other hand, models like XGBoost or neural networks use millions of parameters making it impossible to map the relationship between predictors and outcome.
- *Models that are good at prediction tend to be poor at interpretation and vice versa.*



Source

# Why bother with Interpretability?

- To Inculcate Trust in models. Models that are easy to understand are more likely to be adopted.
- To debug predictions. When models commit errors in prediction, one can go back and identify the source. This is particularly important in high-risk domains such as loan applications, parole recommendations, or hiring decisions. Consider the following
  - *Algorithm used by Dutch Tax authorities wrongly accused people of social benefits fraud. The errors were attributed to the use of citizenship in applications, specifically those with dual citizenship were more likely to flagged as defrauders.* ([vice.com](https://www.vice.com))
  - *COMPAS: An important factor in Judges' sentencing decisions is the criminal defendant's likelihood of recidivism, or chance of re-offending. The probability is computed by algorithms. Are they getting it right?* ([Propublica](https://www.propublica.org), 2016)
  - *Healthcare Risk: Black patients assigned the same level of risk by one widely used algorithm were sicker than White patients. A deeper dive into explanation revealed this was because the algorithm uses health costs as a proxy for health needs.* ([Science](https://www.science.org)).
- Compliance. Laws may prohibit the use of certain predictors in certain domains. For instance, demographics cannot be used in hiring or housing decisions.

*“Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”*  
([Rudin, 2019](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7009150/))

# Estimation Approaches

- Parametric methods
  - *Make an assumption of the functional form of relationship between predictors and outcome*
  - *Use training data to estimate parameters of equation*
  - *E.g., Linear regression*
- Non-parametric methods
  - *Does not make any assumption about the functional form of relationship*
  - *Can fit a wider range of shapes for  $f$*
  - *But, needs a very large number of observations*
  - *E.g., splines*

# Regression vs. Classification Problems

- Depends on nature of the outcome variable
- Regression problem: Outcome variable is numeric
  - *Least squares linear regression*
- Classification problem: Outcome variable is categorical
  - *Logistic regression*
- While some techniques can address only one i.e., regression or classification problems, others can address either. The latter include trees, forests, and boosting.

# Regression vs. Classification Problems

## Model performance metrics

### Regression Problems

- Estimate Predictions

### Classification Problems

- Decision Predictions
  - *Group 1 or group 2; High or Low*
  - *Often involves categorizing a probability outcome into class predictions*

# Regression vs. Classification Problems

Regression Problems

Predictor1	Predictor2	Predictor3



Outcome
232.32
134.54
67.45
129.46
162.89

Classification Problems

Predictor1	Predictor2	Predictor3



Outcome
Not Buy
Buy
Buy
Buy
Not Buy

# Regression Problems

## Model Performance Metrics

- Measures of error
  - *Mean Squared Error (mse)*
  - *Root Mean Squared Error (rmse)*
  - *Mean Absolute Error (mae)*
  - *Mean Absolute Percentage Error (mape)*
- Measure of explained variance
  - $R^2$

Predictor1	Predictor2	Predictor3	Outcome
			232.32
			134.54
			67.45
			129.46
			162.89

# Classification Problems

## Model Performance Metrics

- Class-probability based metrics
  - *Log-likelihood*
  - *Gini*
  - *Entropy*
- Accuracy-based metrics
  - *Accuracy*,
  - *Misclassification rate (= 1-accuracy)*,
- *Accuracy focused on type of error. All errors are bad, some are worse.*
  - *Precision*
  - *Recall (or Sensitivity), and*
  - *Specificity*
- Performance independent of cutoff value
  - *Area under the ROC curve (AUC)*

Predictor1	Predictor2	Predictor3	Outcome
			Not Buy
			Buy
			Buy
			Buy
			Not Buy



# OVERFITTING

# Model Accuracy

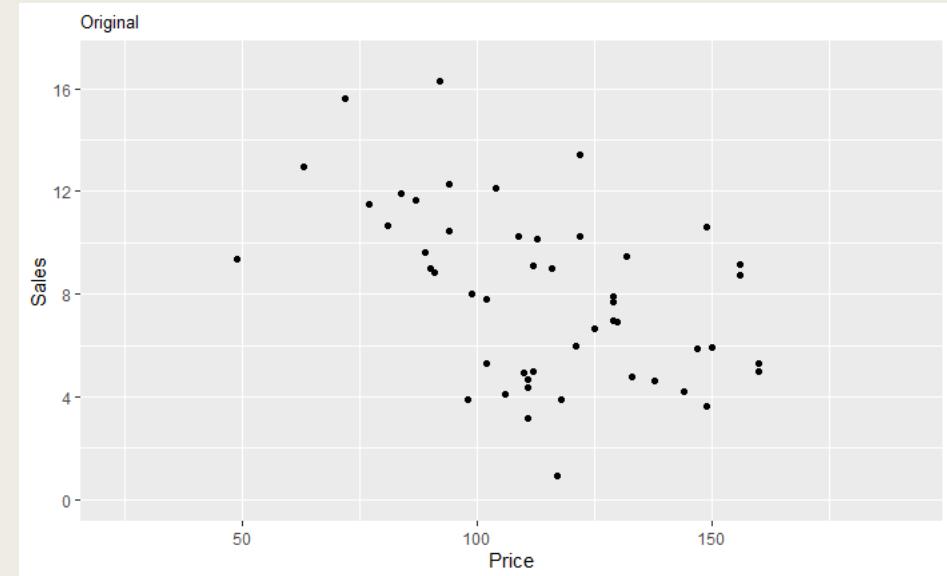
- Performance of a model is determined by comparing model predictions to true values.
- Performance can only be judged based on the data the researcher has, i.e., the data used to train the model.
- *But, in most cases the researcher is interested in performance of the model in the real world, i.e., on data not used to train the model.*

# Real World is..... different

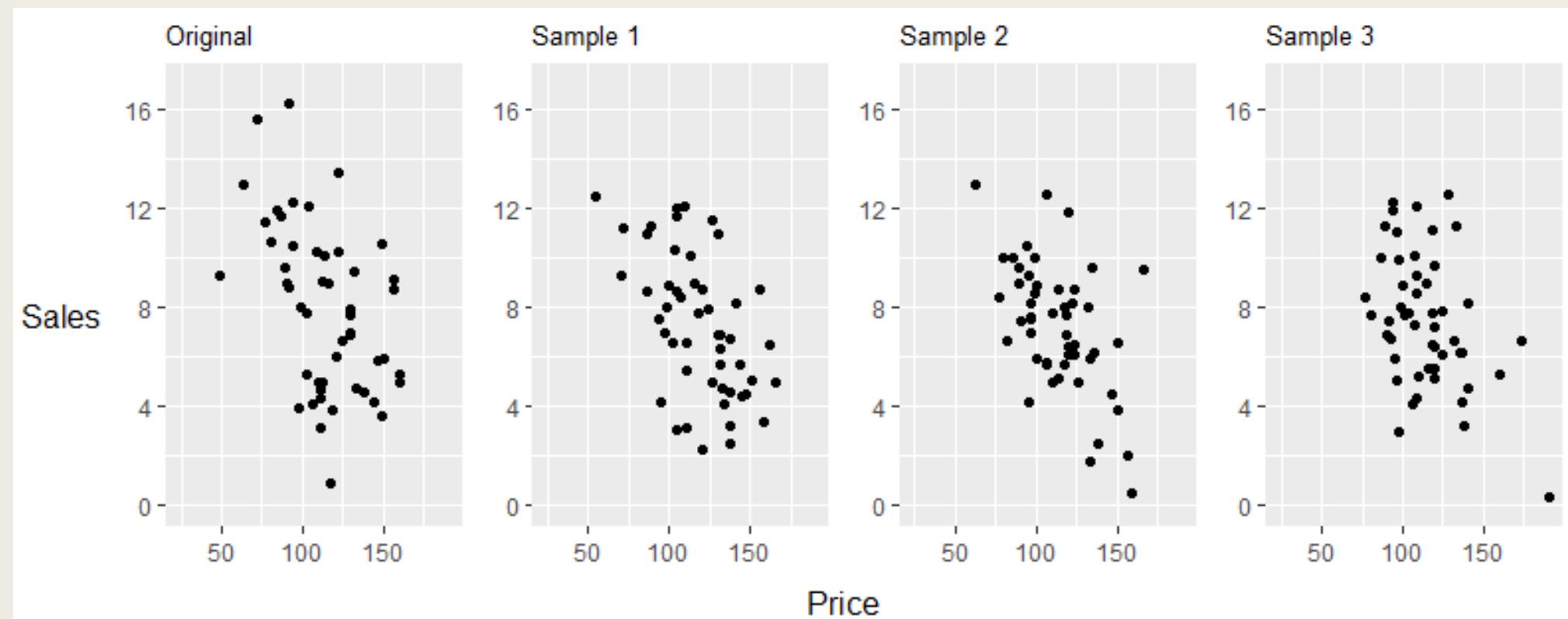
- A student who practices hard for a standardized test sees his scores improving rapidly. The actual exam is a bit of a shocker as his score is significantly lower.
- Car performance tuned on test tracks often falters on real roads.
- Self-driving cars in Australia ([The Guardian](#))
- Self-driving car trained for highways failed on local roads ([Electrek, 2016](#))
- Anti-Tank Dogs ([Wikipedia](#))
- Million dollar prize winning Netflix algorithm only worked for DVDs.

# To Illustrate

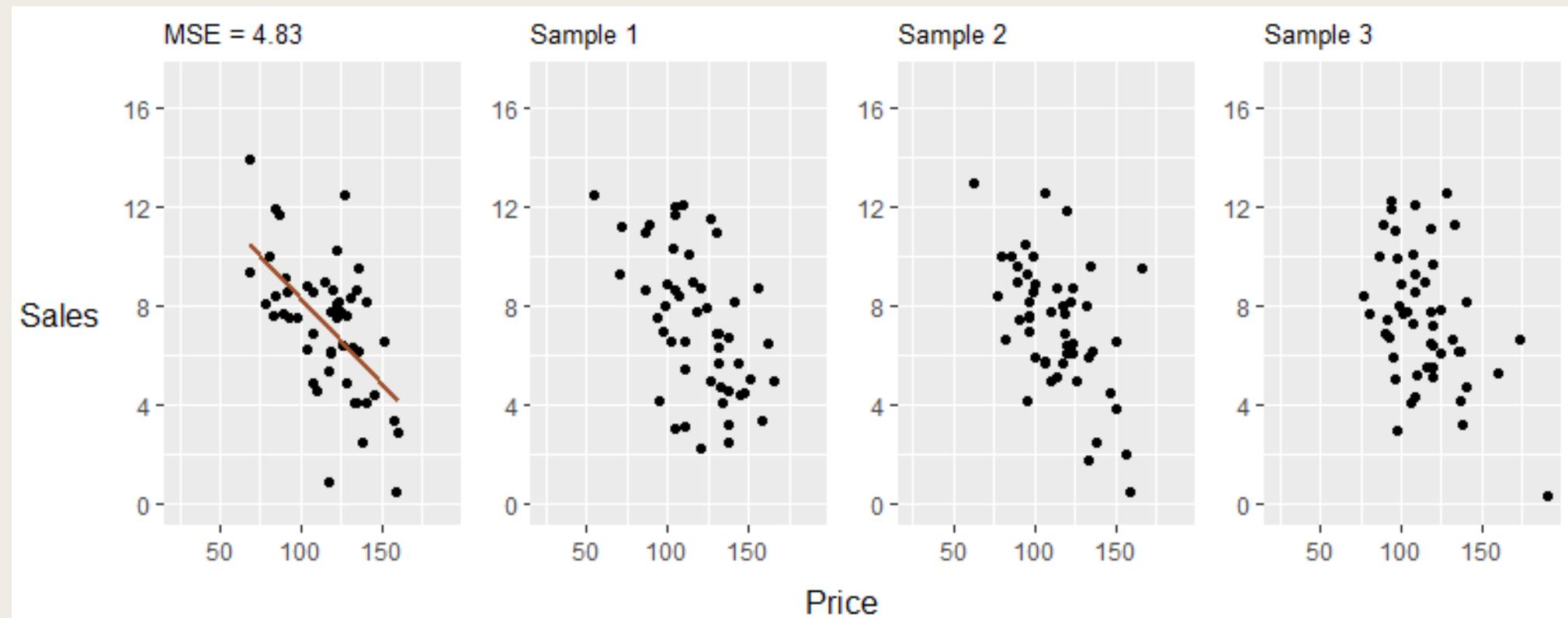
- Here is a random sample of data ( $n=50$ ) of Carseats data. Data is represented as a scatterplot between Sales and Price.



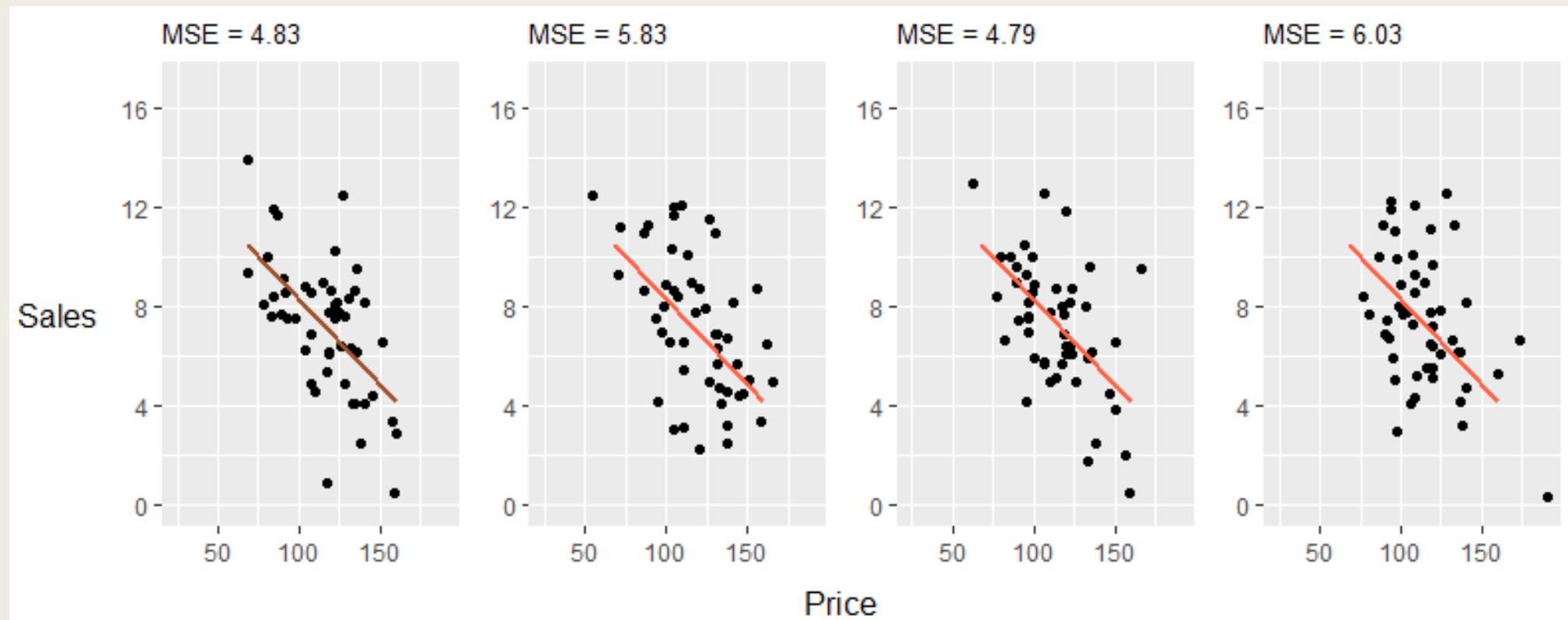
# Original Sample vs. Random Samples



# Train Simple Model

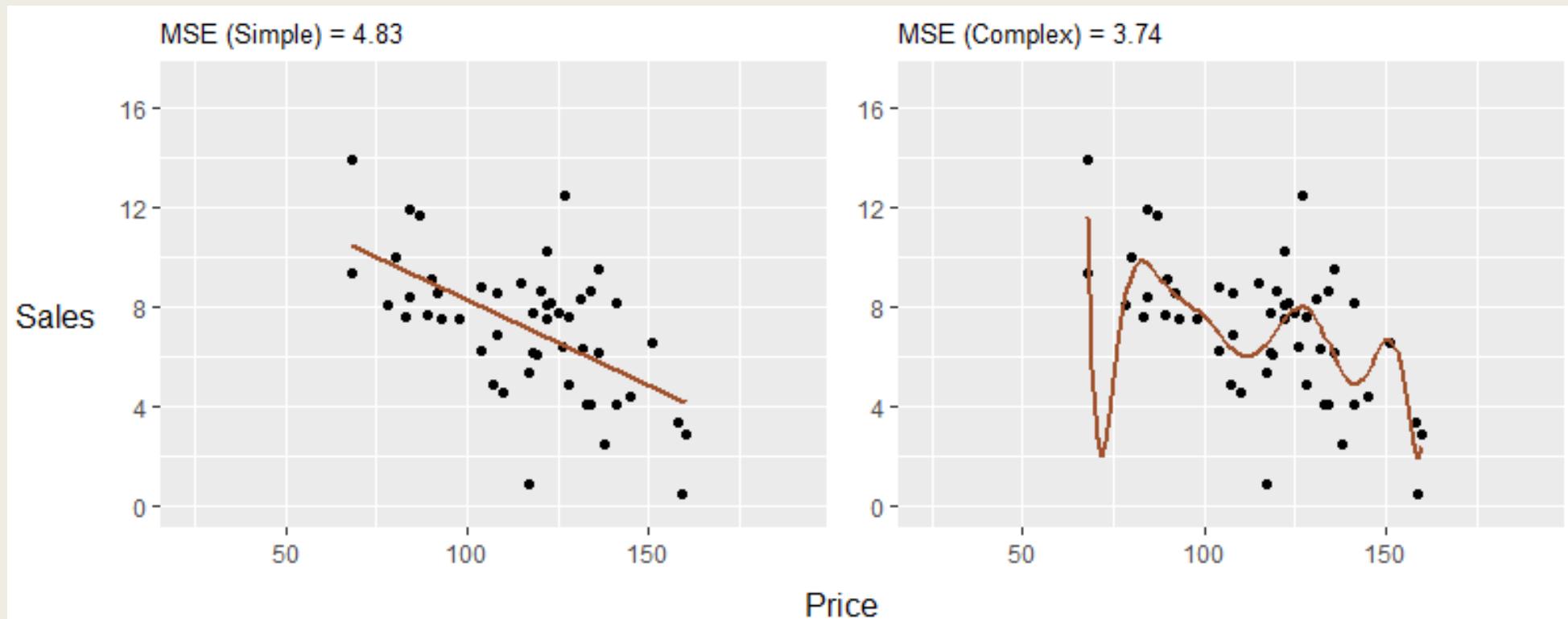


# Apply Trained Model to New Data



# BIAS-VARIANCE TRADEOFF

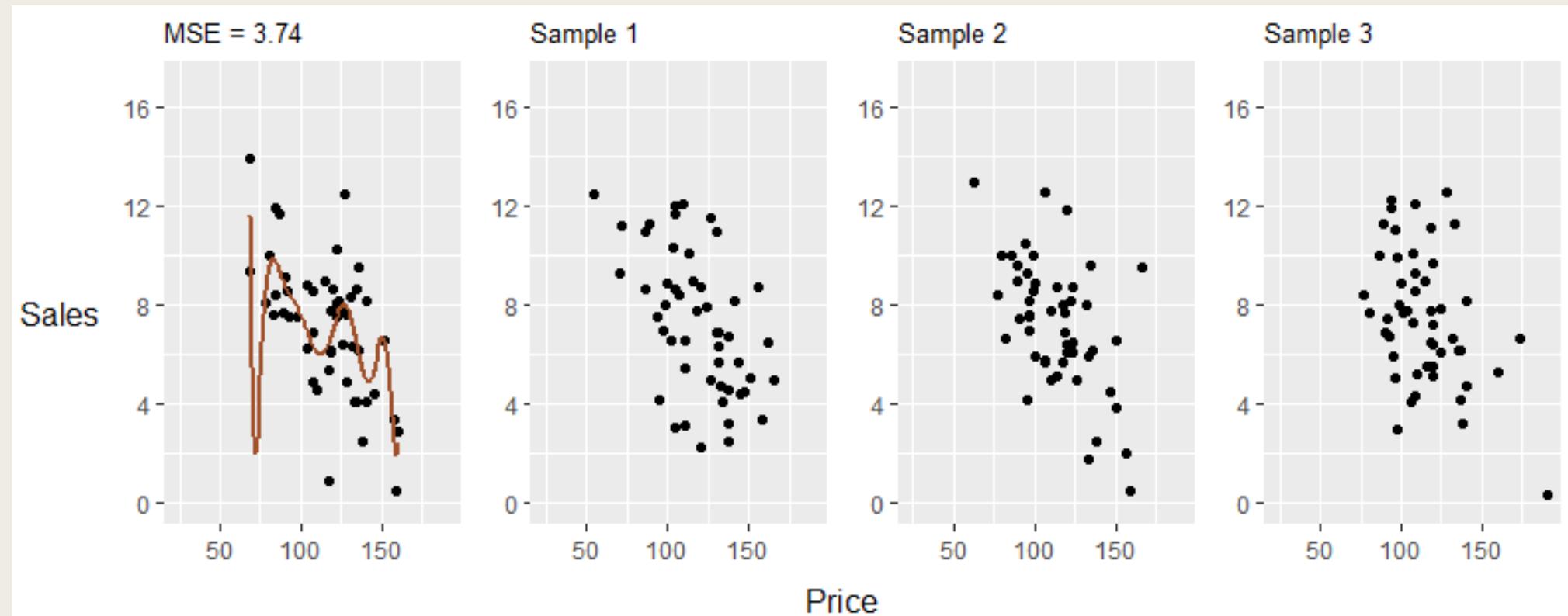
# Which Model is better, Simple or Complex?



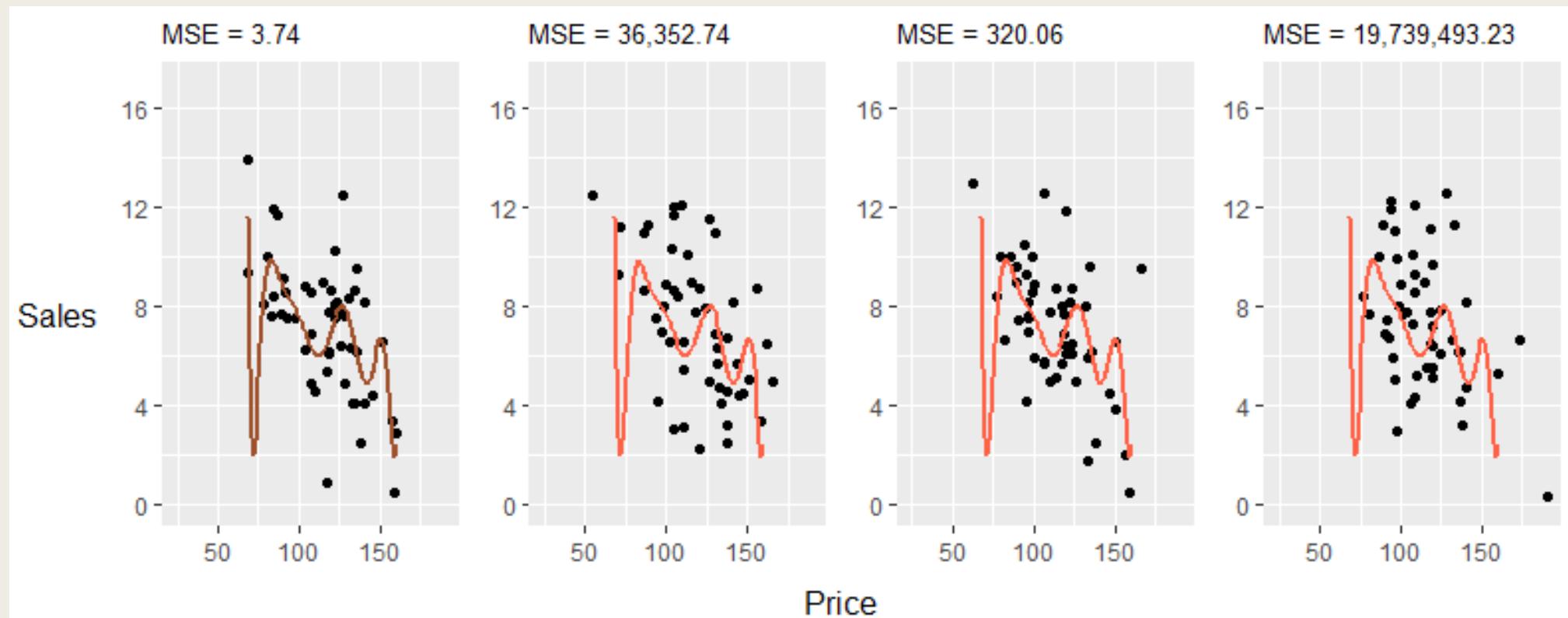
# Going from Simple Models to Complex Models

- As model complexity increases,
  - *models perform better on the sample used to train the model*
  - *but they also perform worse on datasets not used to train the model*
- The extent to which the model performs well on the data used to build it versus data not used to build it is called *Overfitting*.
- Overfitting is seen when in-sample performance far exceeds out-of-sample performance.
- Overfitting is exacerbated by model complexity
- This is the classic Bias-Variance tradeoff
- Let us review this issue.

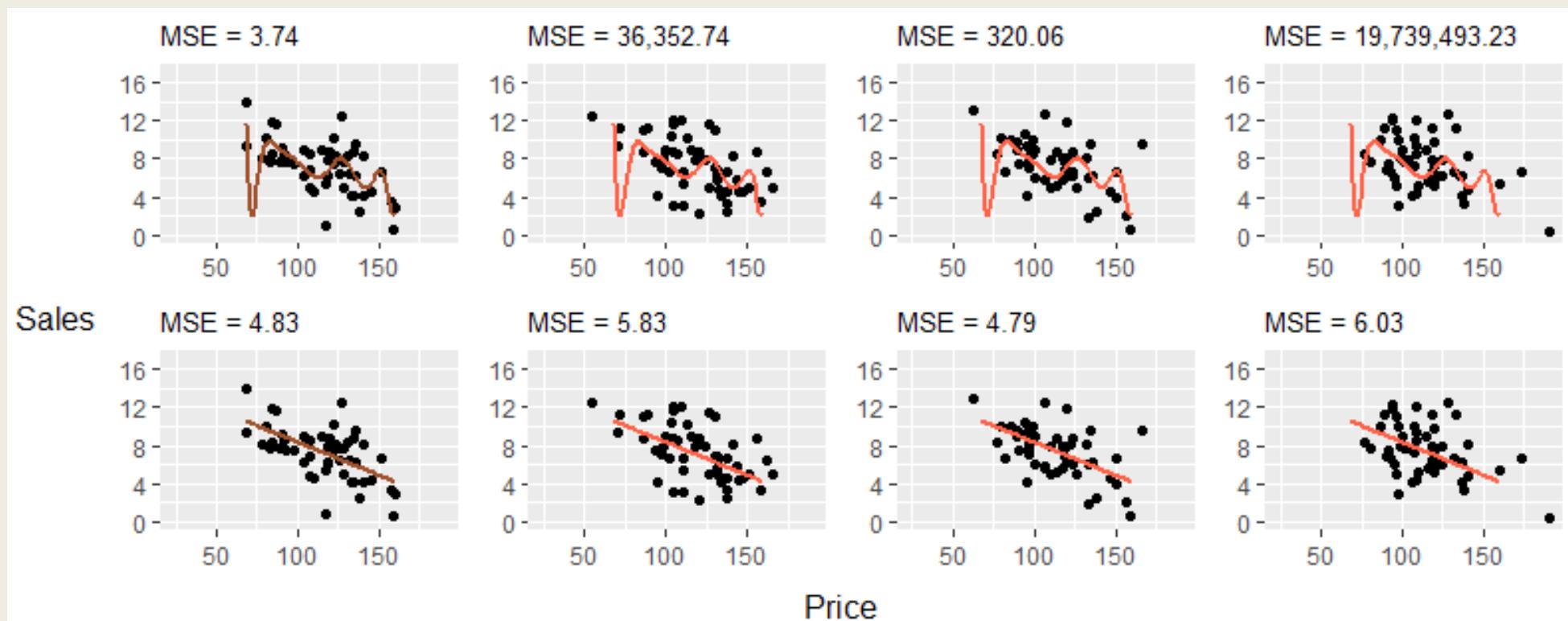
# Train Complex Model



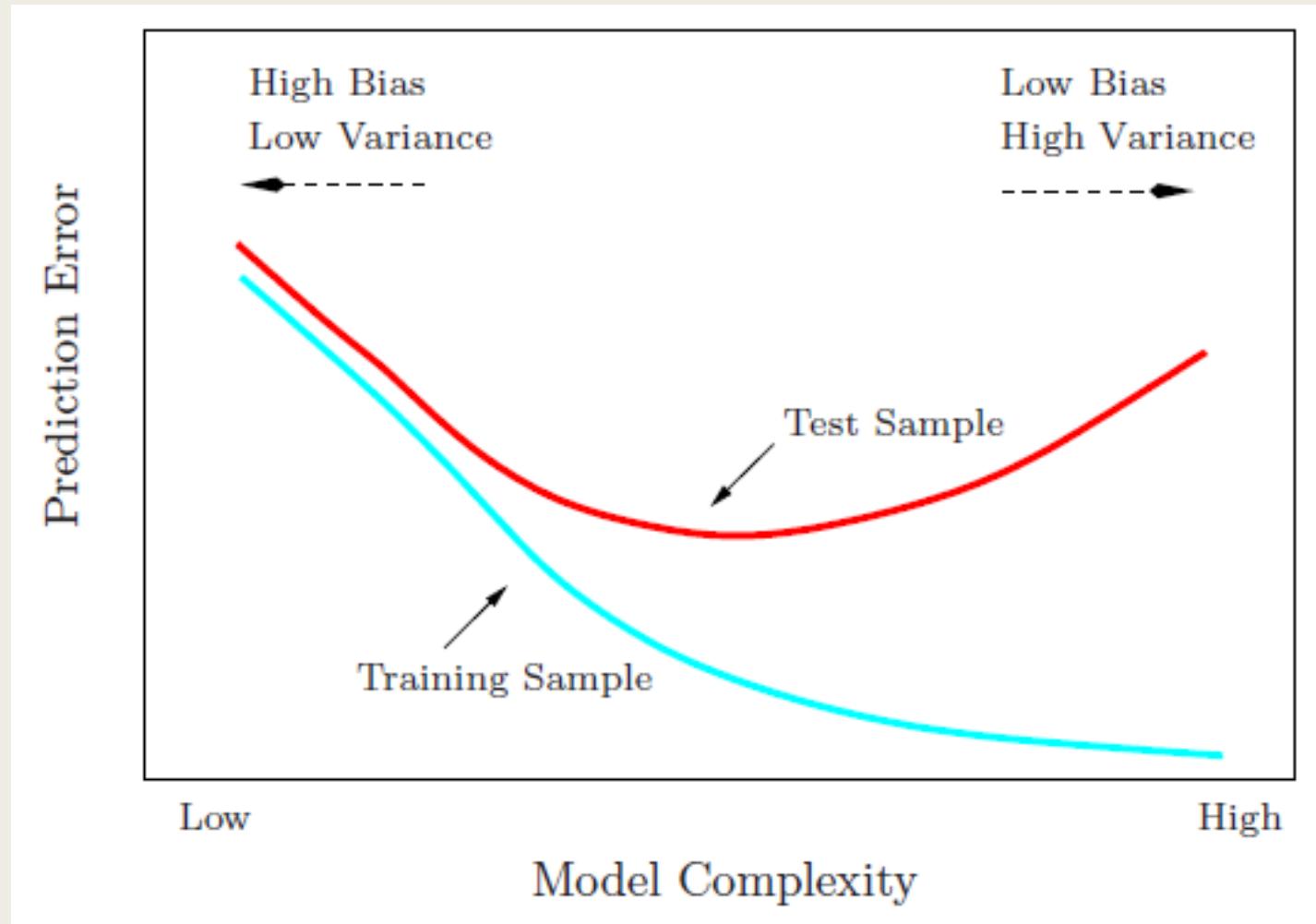
# Apply Complex Model to New Data



# Overfitting in Simple and Complex Models

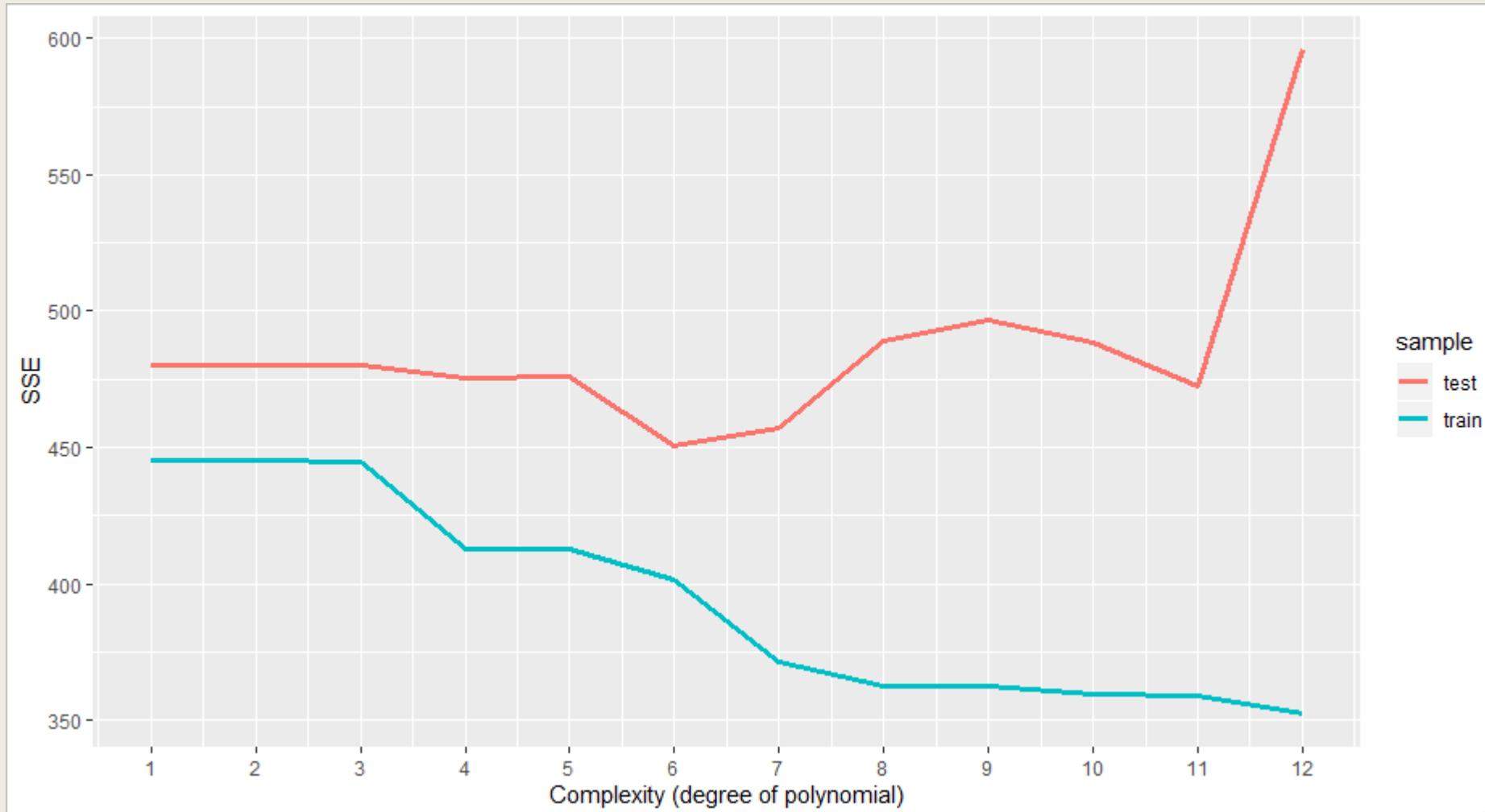


# Training vs. Test Set



# Training vs. Test Set

Prediction Accuracy vs Complexity: Sales =  $f(\text{Price}^d)$ , where d is degree

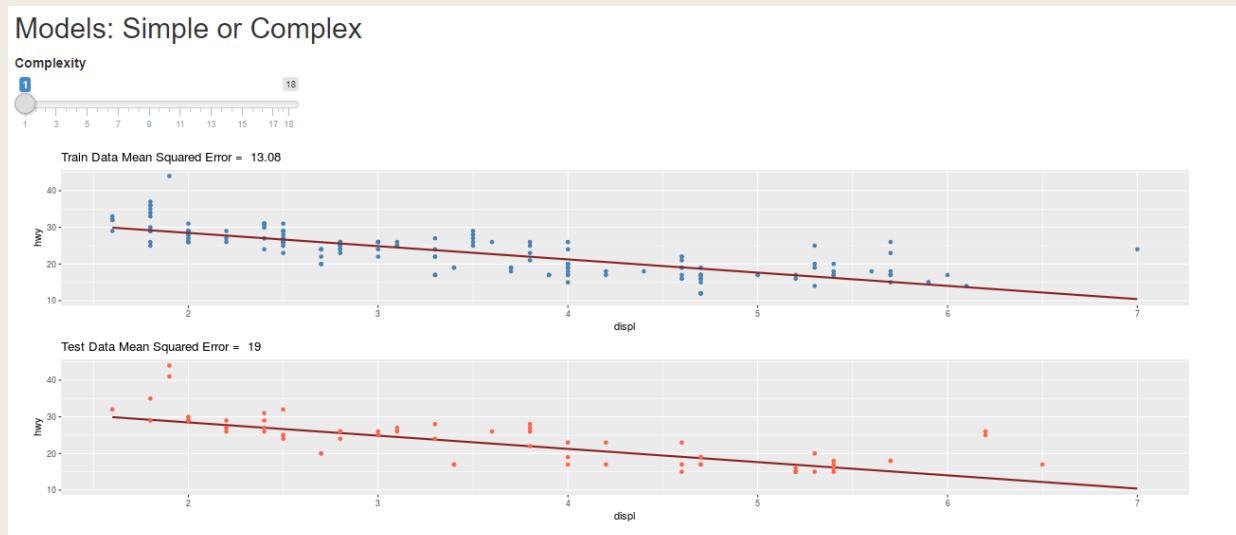


# Training vs. Test Set

- Researcher is generally interested in developing a model that performs well out-of-sample.
- In practice, we only have training data, therefore not possible to assess performance out-of-sample.
- Also, as noted in foregoing illustration, in-sample performance is a poor proxy for out-of-sample performance.

# Training vs. Test Set

- Here is an [interactive chart](#) to examine the effects of complexity on train and test set performance.
- Complexity is reflected by the degree of a polynomial regression model
- Model uses displ to predict hwy (highway gas mileage) for different degrees of displ.



# Train and Test Samples

- One solution is to split the sample into two parts: train and test.
  - *Other solutions such as cross-validation will be discussed later.*
- Estimate the model on train set and evaluate using the test set.
- Performance of model on test set can be used as an indication of out-of-sample performance.
- Note:
  - *train sample is also referred to as estimation sample*
  - *test sample is also known as validation or holdout sample*

**Data**



<b>Data</b>										
<b>Train</b>										
<b>Test</b>										

# Train and Test Samples

## Factors to Consider

- Size of train and test sample
  - *If data is sufficiently large, a 50:50 split may be done*
  - *Generally, train sample is larger than test sample, with the split being 60:40 or 70:30. These are heuristics not rules.*
- Method of split
  - *Non-random approaches: Only used in very specific situations. E.g., time-series data.*
  - *Random approaches*
    - Simple random sampling: Designed to make train and test sample as similar as possible.
    - Stratified sampling: Applies random sampling within subgroups.
      - *On outcome: Random sampling while ensuring the distribution or proportion of outcome is the same across samples*
      - *On predictors: Same idea as above but for specific predictors such as gender or location*

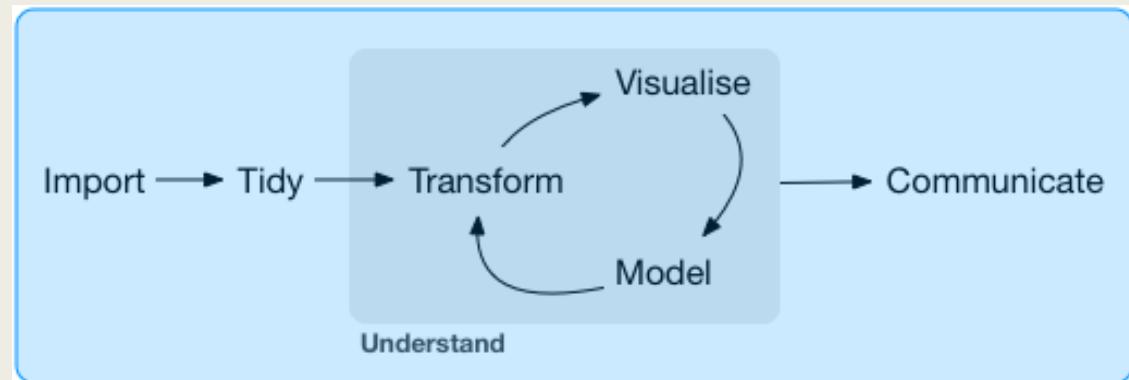
# THE MODEL

# The “Best” Model

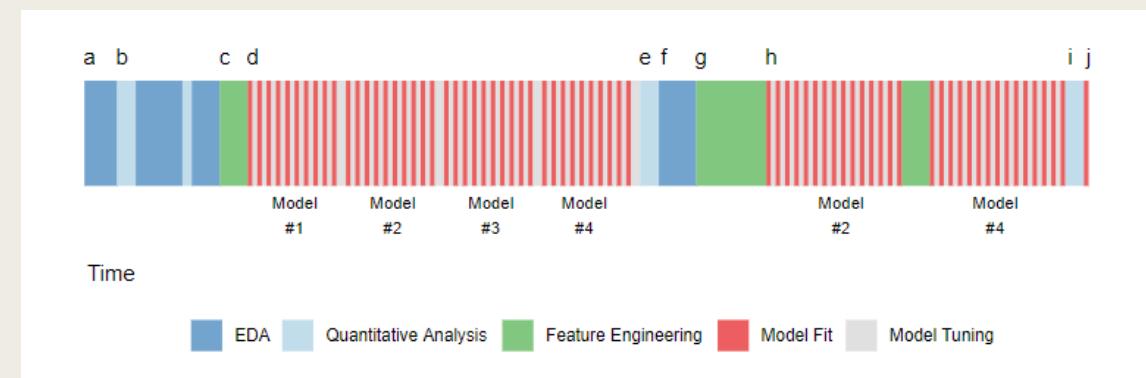
- *The [No Free Lunch Theorem](#) shows that under certain assumptions*
  - *No single predictive model can be declared to be the best*
- While certain models work with certain data characteristics (e.g., missing values), they may fail with different data characteristics
- Rather than seeking a silver bullet, analysts, should examine the problem or data at hand, before deciding on the models to use.

# Road to the Best Model

- Modeling process is iterative, not linear
- Predictive analysis is much more than just fitting a single model to tidy data



Source: [R for Data Science](#)



Source: Kuhn and Johnson (2019)

# INFERENTIAL STATISTICS

# Inferential Statistics

- Population
  - *Collection of all units for the study*
- Sample
  - *Subset of the population*
- Sample is used to draw inferences about the population
- Most studies are based on a sample

# Process of inferential statistics

- Generate a hypothesis about the population, null hypothesis ( $H_0$ ) and an alternative hypothesis ( $H_1$ ) such that the two cover the Universe of possibilities
- Select a statistical technique to generate a test statistic. Test statistic often follows a well known distribution such as t, F, or  $\chi^2$ .
- Choose a level of significance (e.g.,  $\alpha = 0.01$ ) to reflect tolerance for Type I error, i.e., rejecting  $H_0$  when in fact it is true.
- Gather data and calculate value of test statistic
- Determine the probability (p-value) of obtaining the test statistic assuming null hypothesis is true.
- If  $p < \alpha$ , reject  $H_0$

# Illustration

Consider the Linear Model: Sales =  $b_0 + b_1 * \text{AdSpend}$

- Hypotheses, being tested (although not always explicitly stated)
  - $H_0: b_1 = 0$
  - $H_1: b_1 \neq 0$
  - *If coefficient of AdSpend ( $b_1$ ) in the population is 0, one would conclude AdSpend does not drive Sales*
- Test statistic: t value for coefficient of AdSpend
- Level of Significance ( $\alpha$ ) = 0.01
  - *Values used tend to be 0.1, 0.05, 0.01, 0.001 but whatever the threshold, it should be set before looking at the data*
- Gather data and calculate value of test statistic
- Translate t value into p-value. Let's say  $p = 0.002$ . This means if  $b_1$  is 0 then there is only a 0.2% chance of obtaining the sample data.
- Since the chance ( $p=0.002$ ) is below our threshold ( $\alpha = 0.01$ ), one would reject the null hypothesis and conclude that the coefficient of AdSpend is not zero. In other words, AdSpend influences Sales.

# In Practice

- Desirable results are generally in  $H_1$ , so analysts generally seek to reject  $H_0$  in favor of  $H_1$ .
- p-value does not reflect strength of effect
- p-value is sensitive to sample size. With large samples, even very small effects are statistically significant
- Statistical significance does not imply practical significance.
- On the other hand, before one can examine practical significance, it is imperative that the results are statistically significant.

# Conclusion

- In this module, we reviewed
  - *machine Learning*
  - *goals of prediction vs. inference*
  - *assessing model accuracy*
  - *problem of overfitting*
  - *splitting the data to estimate test error*
  - *the iterative modeling process*
  - *inferential statistics to determine significance of results*