Class 6: Introductory Machine Learning Terminology and Bias-Variance Tradeoff

MGSC 310

Prof. Jonathan Hersh

Class 6: Announcements

1. TA Office Hours:
   - Tuesdays: 5:30 – 7
   - Thursdays: 12:30-2
   - Mondays: 5-6:30
2. Quiz 2 posted, due Thursday @ midnight
3. Be sure you are following along with the course reading (ISLR pp 15-36 covered today)
4. Data Analytics Accelerator Info Sesh Oct 5 @ 11am

Class 6: Outline

1. Fun with R ggridges
2. Classification vs Regression
3. Supervised vs Unsupervised Learning
4. Bias and Variance
5. Bias-Variance Tradeoff

Fun with R: Ridgeline plots (upload compiled HTML file when done)

```r
#----------------------------------------------------------
# Fun with R: Ridgeline plots
#----------------------------------------------------------
# load these libraries
library('dplyr')
library('tidyr')
library('tidyverse')
library('ggridges')
library('gganimate')
library('forcats')


#----------------------------------------------------------
# Exercises
#----------------------------------------------------------
# Vignettes are short examples where
# 1. Read about the ggridges in the vignette to understand the basics of the package
#    https://cran.r-project.org/web/packages/ggridges/vignettes/introduction.html

# 2. Load the top 5000 IMDB movies database and create the
#    movies_clean data file using the code from lab 5
#   (be sure to load the libraries at the top )

movies <- read.csv(here::here("datasets", "IMDB_movies.csv"))

movies_clean <-
  movies %>%
  distinct() %>%
  mutate(budgetM = budget/1000000,
         grossM = gross/1000000,
         profitM = grossM - budgetM) %>%
  rename(director = director_name,
         title = movie_title,
         year = title_year) %>%
  relocate(title, year, country, director, budgetM, grossM, imdb_score) %>%
  filter(budgetM < 400)

# 3. Install the 'forcats' package
```

What is classification?

Regression: outcome predicting is **continuous**

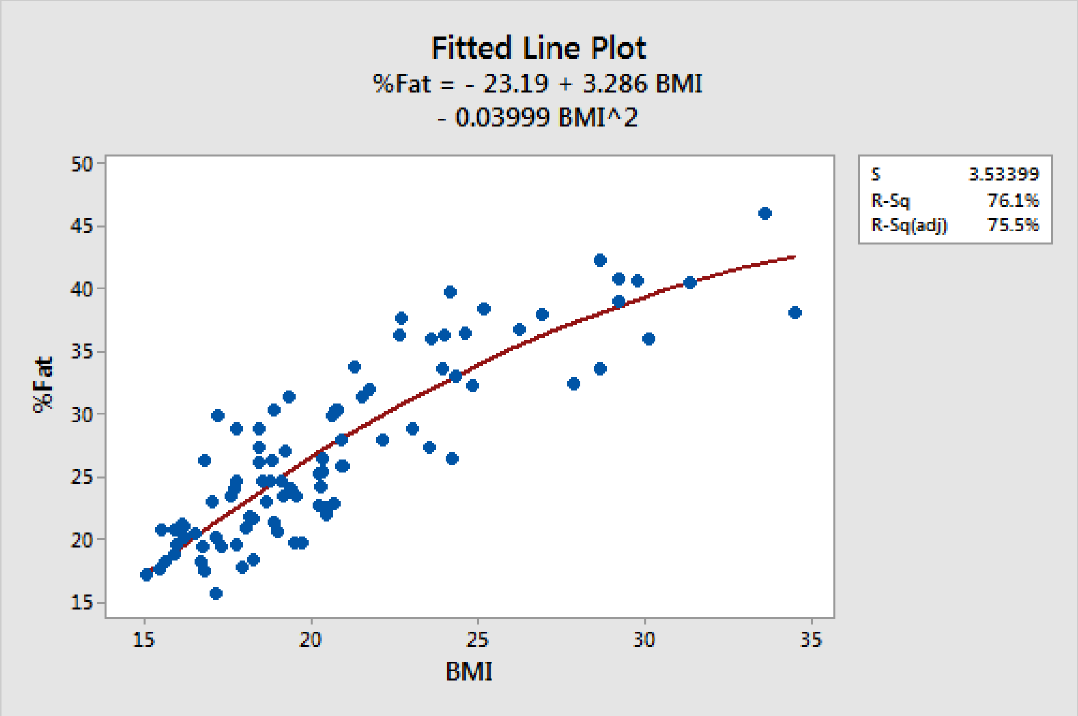
$$y_i \in R$$

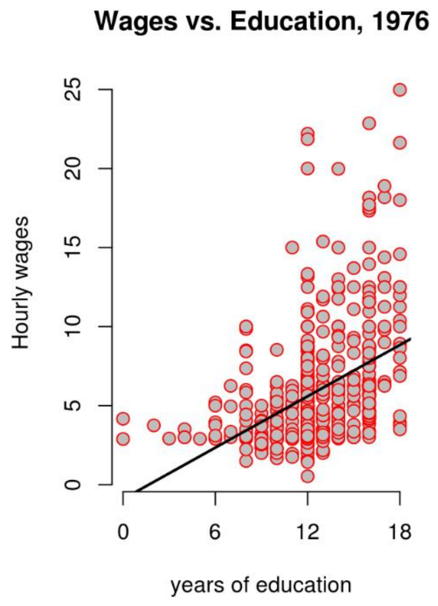
Classification: outcome is **discrete**

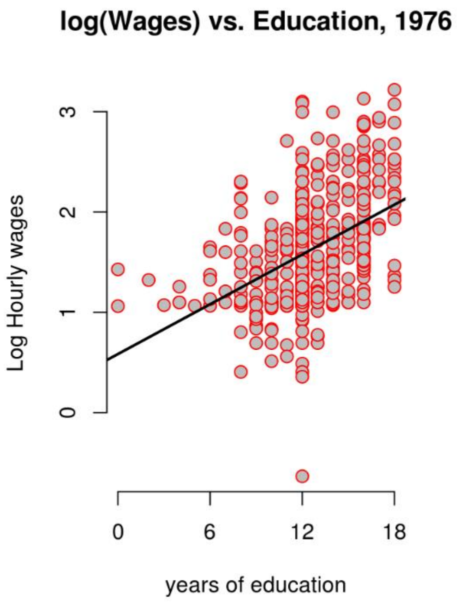
$$y_i \in \{0,1\}$$

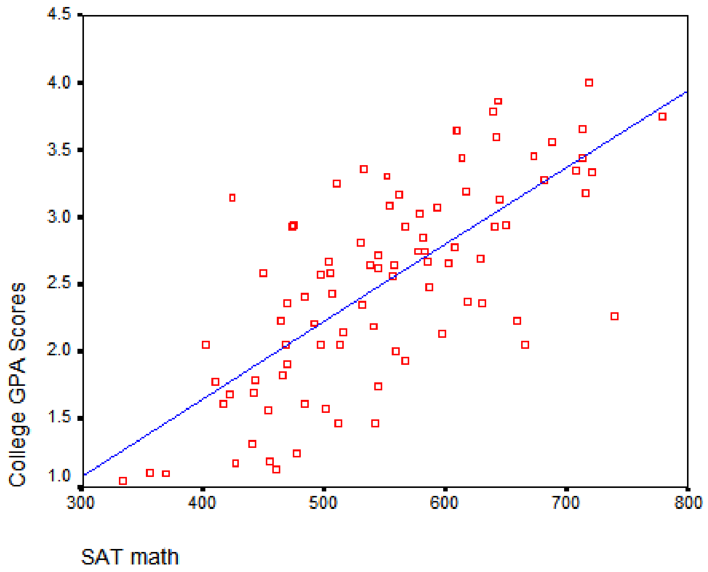
$$y_i \in \{red, black, blue, blond, green\}$$

Regression examples



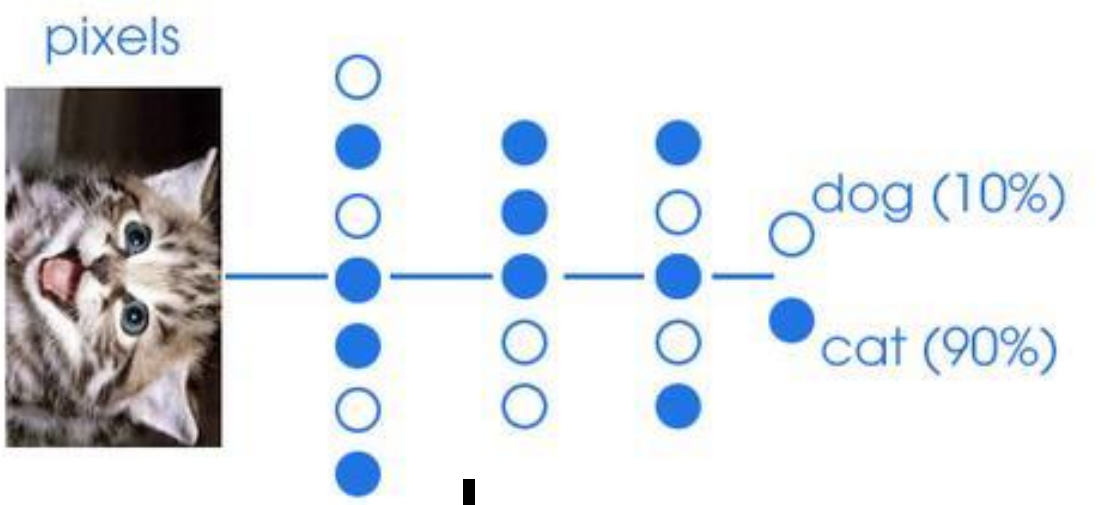


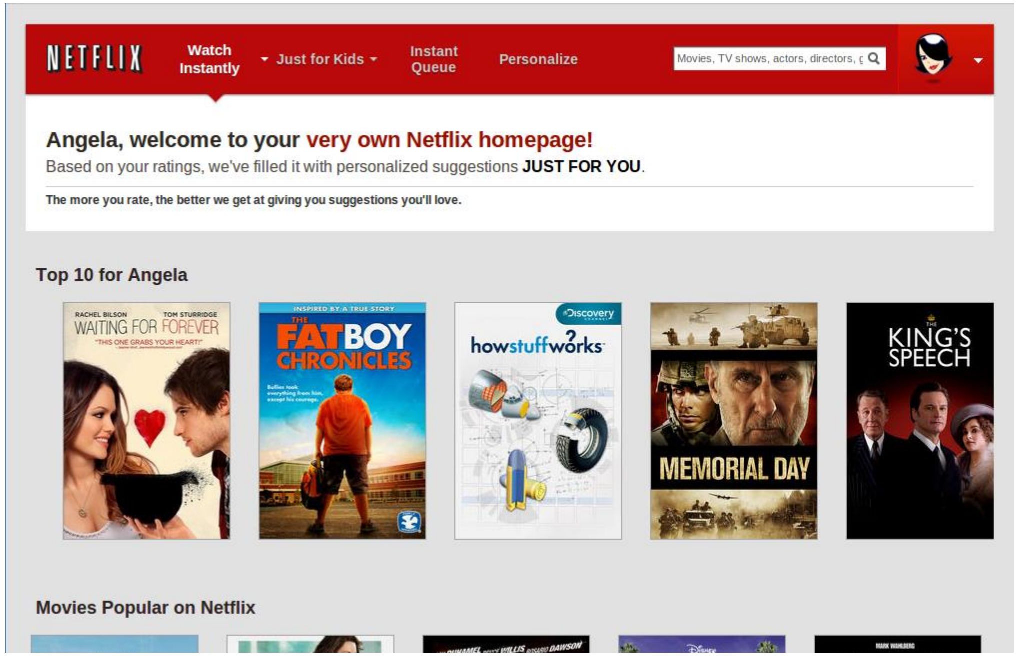


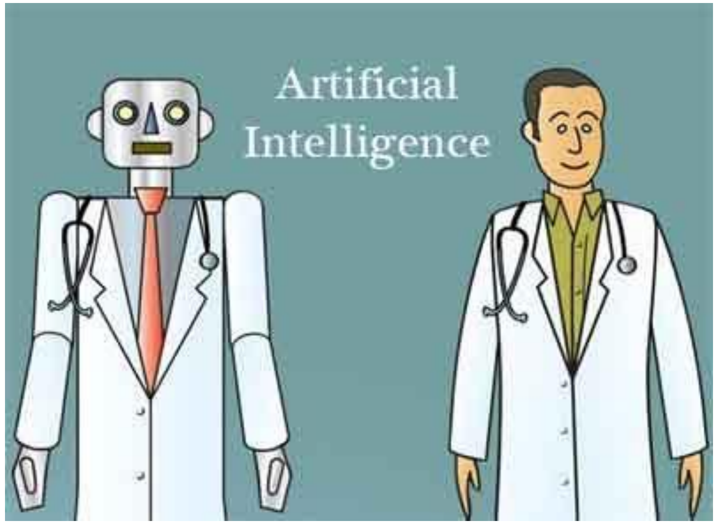


Classification examples

More classification examples







Supervised vs Unsupervised Learning

Supervised Learning:

- For every x_i we observe some y_i
- Ex: random forests to predict loan default (y_i) based on applicant characteristics (x_i)

Unsupervised Learning:

- We only observe x_i
- Ex: clustering loan applicants based on characteristics (x_i)

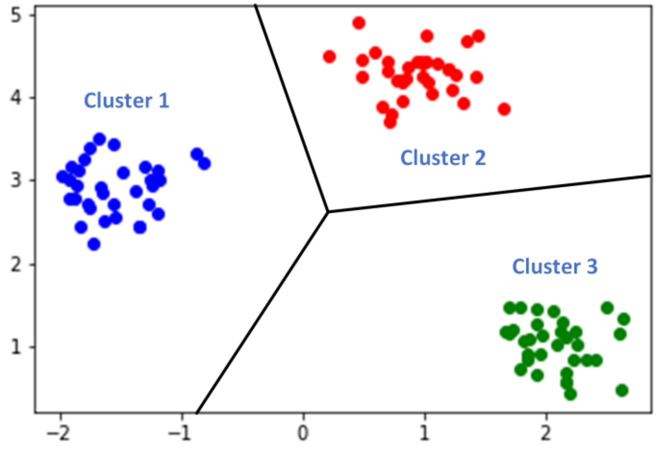
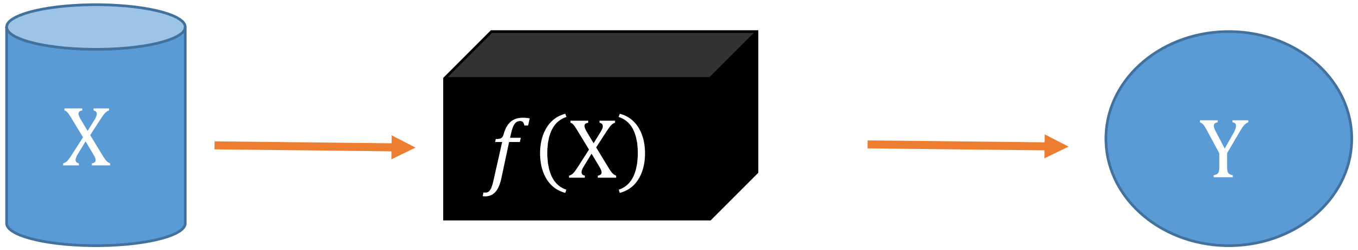


Fig.1. An Example Of Data Clustering

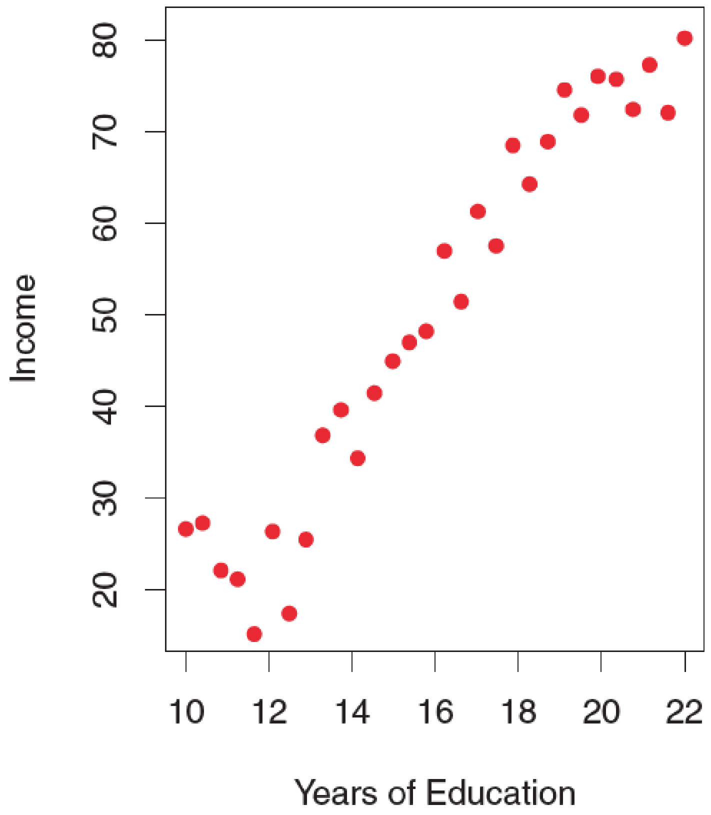
Supervised learning: learning $f(\mathrm{X})$ our predicted out given inputs

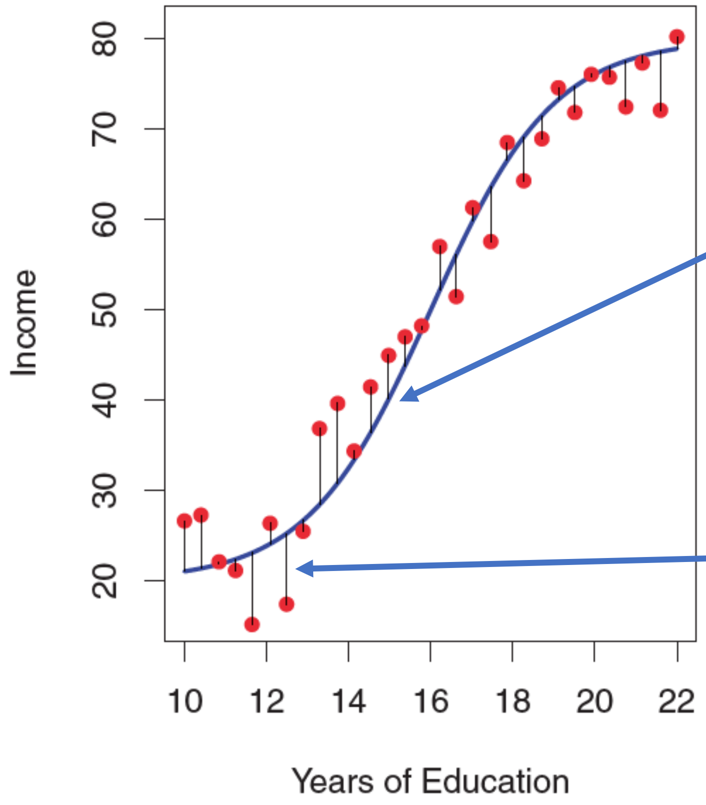
$$\mathrm{Y} = f(\mathrm{X}) + \epsilon$$



ϵ = “epsilon” (unexplained portion)

Example: education and income





Blue line is our Prediction
$\hat{y} = f(x)$

Distance to blue Line are errors
$\epsilon_i = y_i - \hat{y}_i$

“Estimating” $\hat{f}(X)$

- $\mathrm{Y} = f(\mathrm{X}) + \epsilon$ is the true value
- We can only use data to “guess” at $f(\mathrm{X})$
- We call this guess $\hat{f}(X)$

How do we know when we’ve selected a “good” $\hat{f}(X)$?

- We reserve a portion of our data into a “test” set, estimate a model on the other part, and see how our model performs on this test set

Testing Training Data Subsets

Training set: (observation-wise) subset of data used to develop models

Testing/Training Split

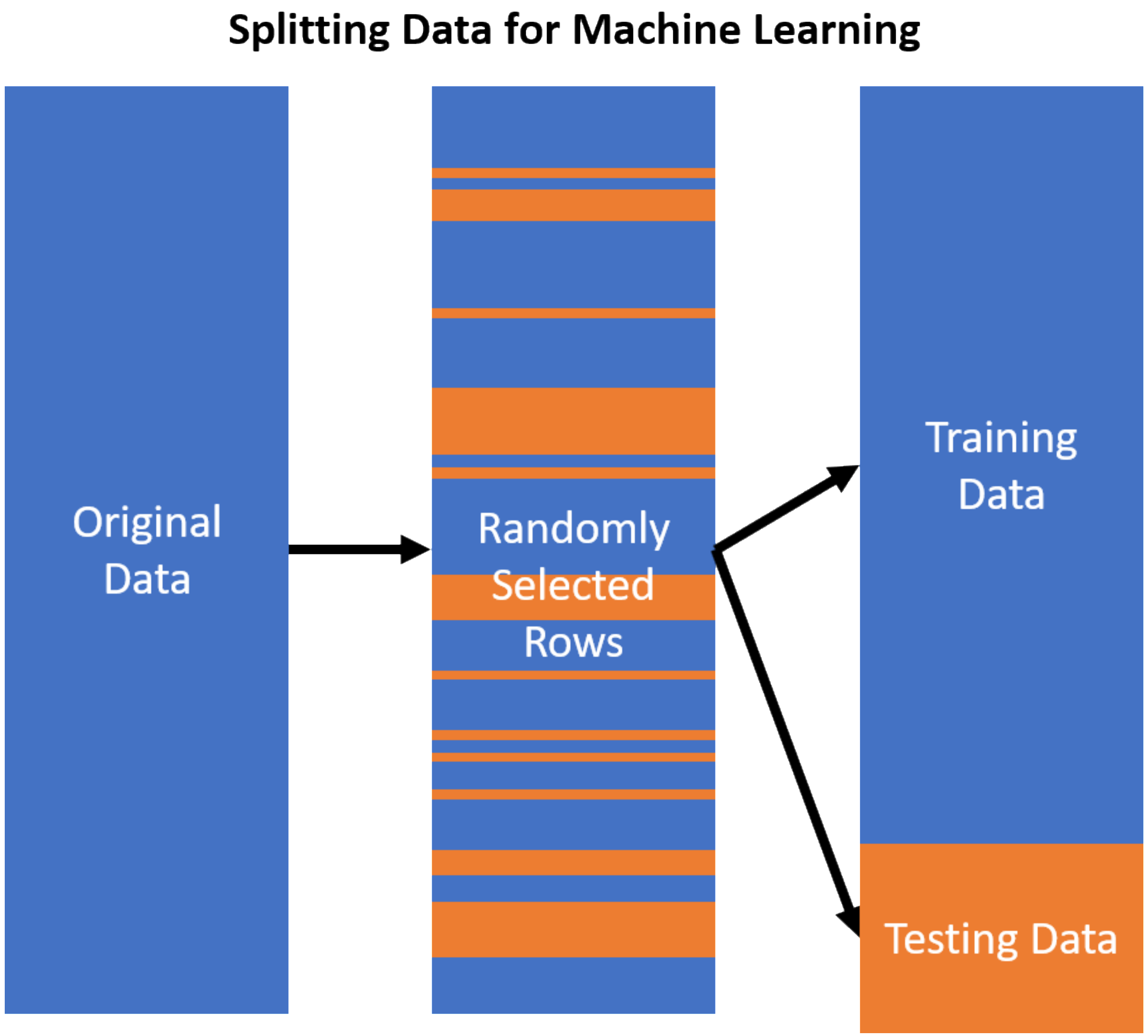
Training set: (observation-wise) subset of data used to develop models

Test set: subset of data used to assess machine learning model performance at end of modeling process

Rule of thumb 75% training 25% test -ish

Randomly Selecting Rows for Test or Training Sets

- Observations are randomly selected into either testing or training splits of the data



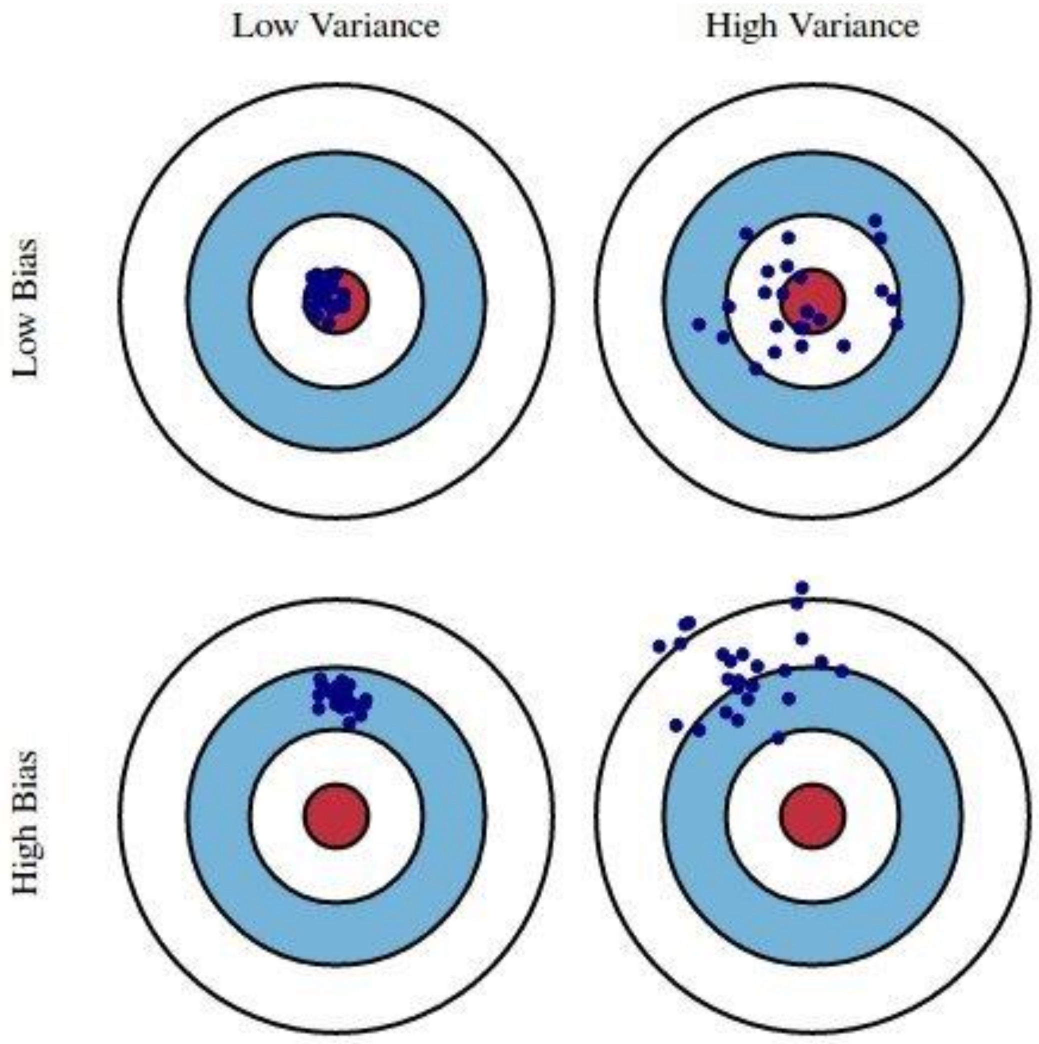
Bias and Variance

Bias: **Tendency of an in-sample statistic to over or under estimate the statistic in the *population***

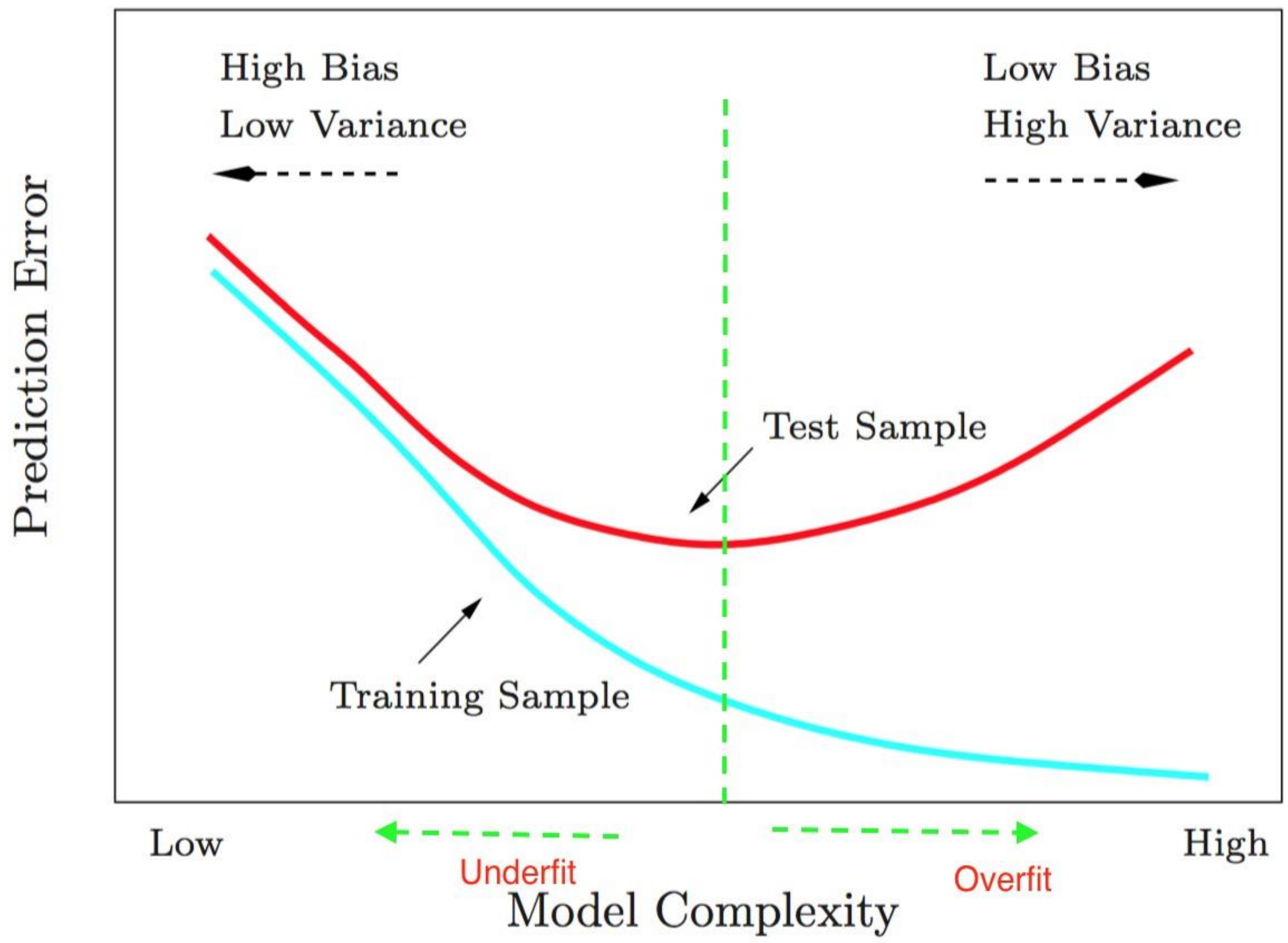
Variance: **Tendency to noisily estimate a statistic**.

E.g., sensitivity to small fluctuations in the training dataset.

Bias-Variance Tradeoff

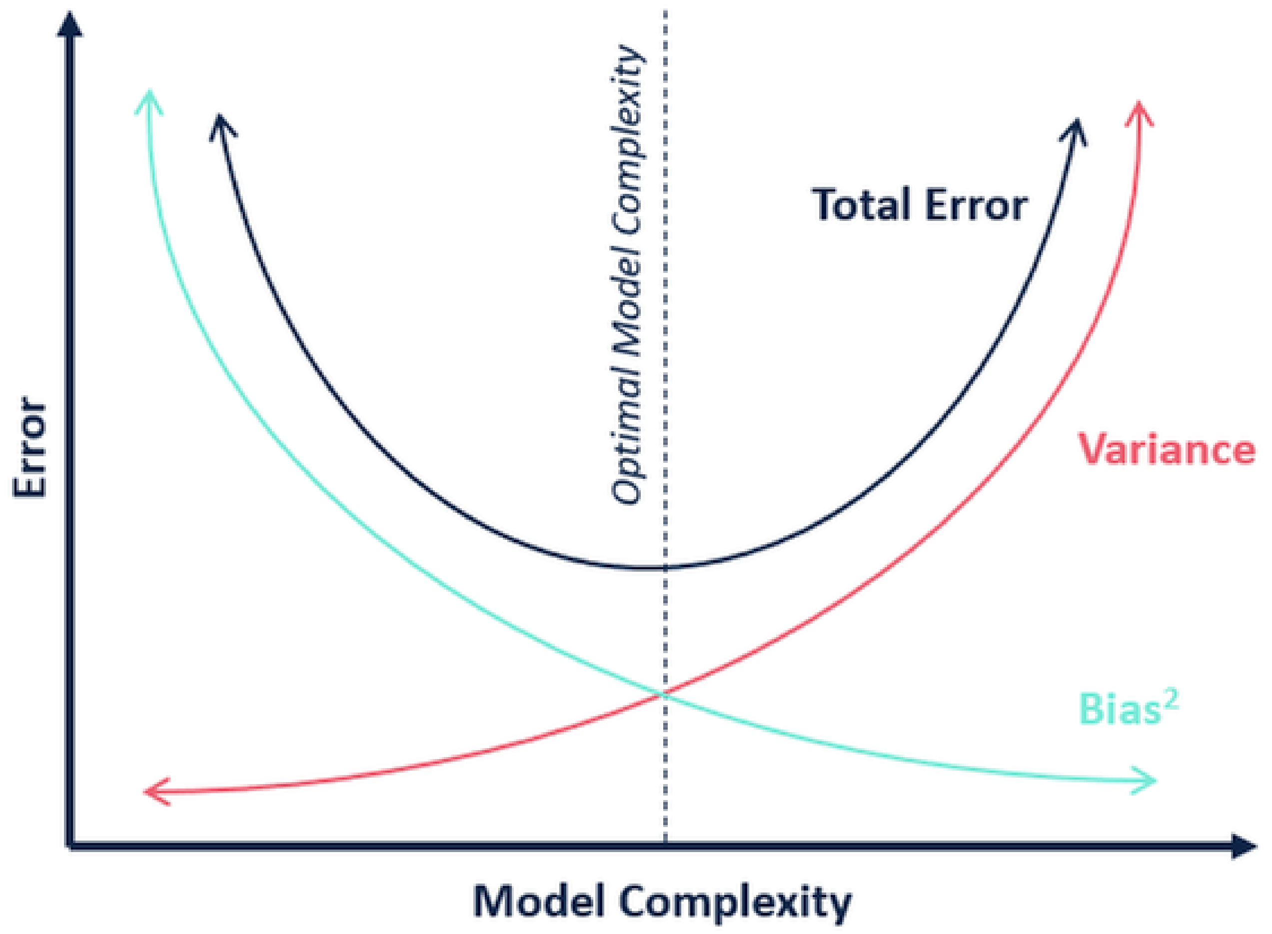


Bias-Variance Tradeoff

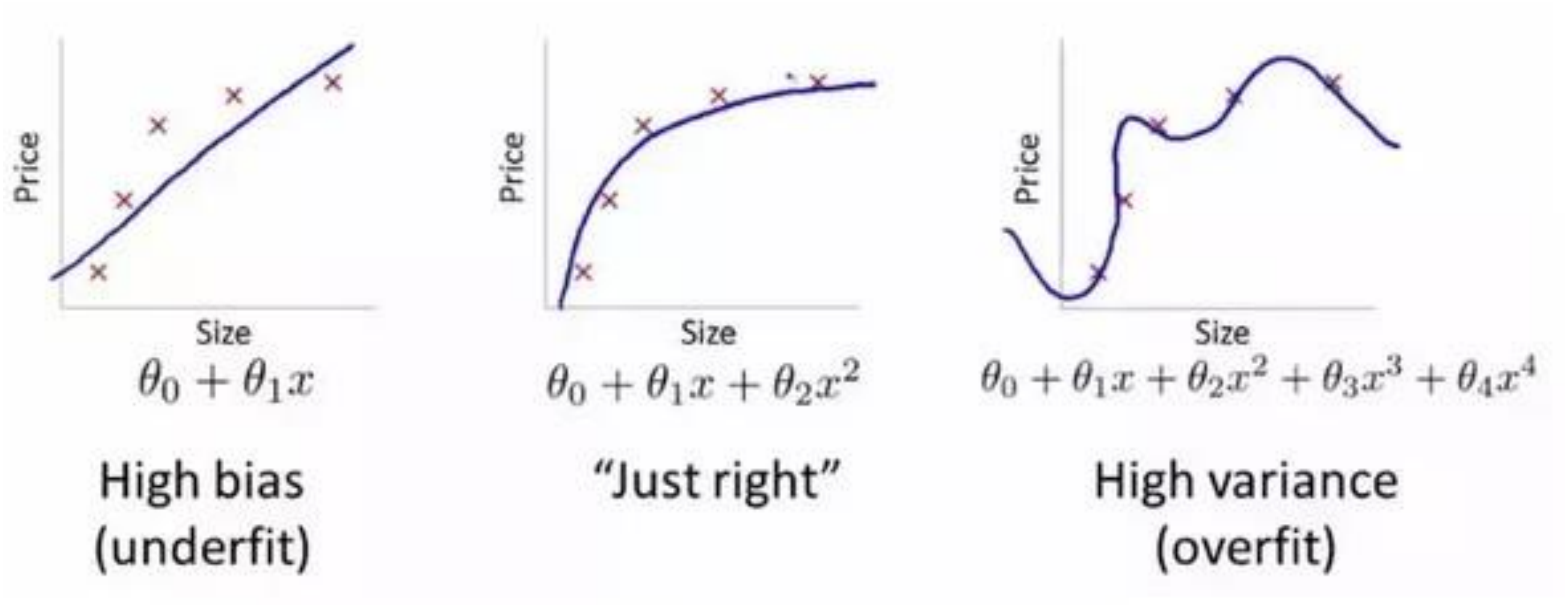


- Error in Training sample (~bias) ↓ as we ↑ model complexity (e.g. number of variables)
- Error in Test sample (~variance) ↑ as we ↑
- Key: finding optimal model complexity

Key: Finding Optimal Model Complexity



Optimal Model Complexity: Neither Underfit Nor Overfit



Mean Squared Error in Practice

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

$\sum$ means we add up anything with i, starting at $i = 1$ to $i = n$

y_i	$\widehat{y}_i$	$y_i - \widehat{y}_i$	$(y_i - \widehat{y}_i)^2$
5	5	0	0
5	7	2	2^2=4
9	8	1	1^2=1
10	1	1	9^2=81
13	13	0	0

Example: Overfitting (True Relationship Linear)

3 Models

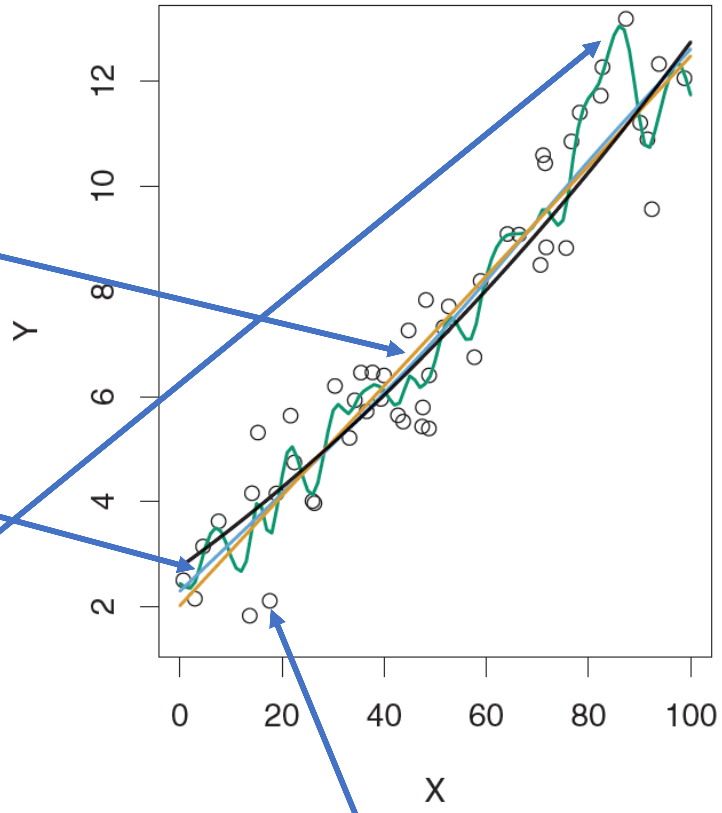
Simple (linear): gold line

Moderate complexity (linear): blue line

Very complicated: green line

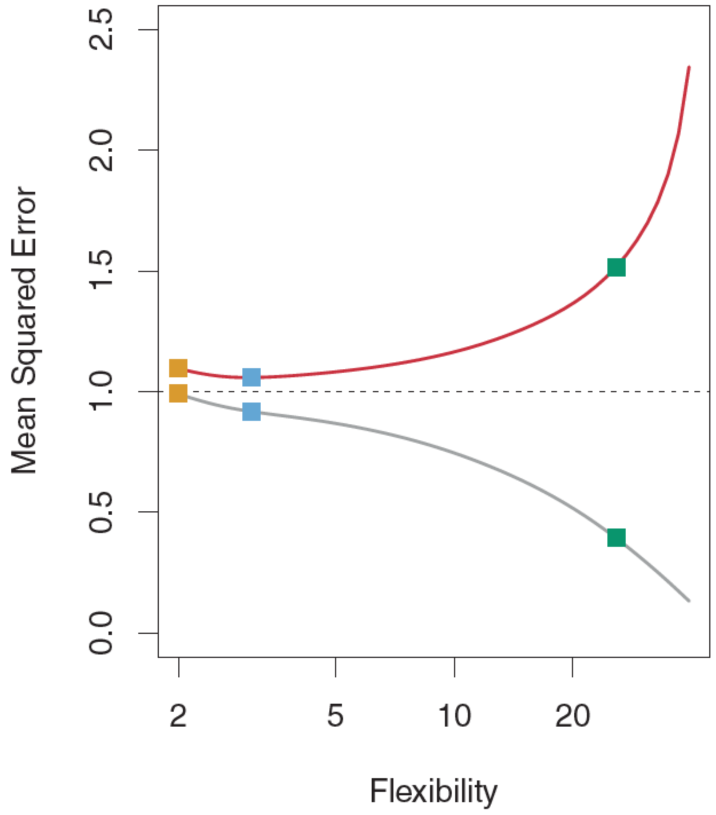
Black line: true relationship

Training Data



Data points (true relationship observed with noise)

Results on Test Data



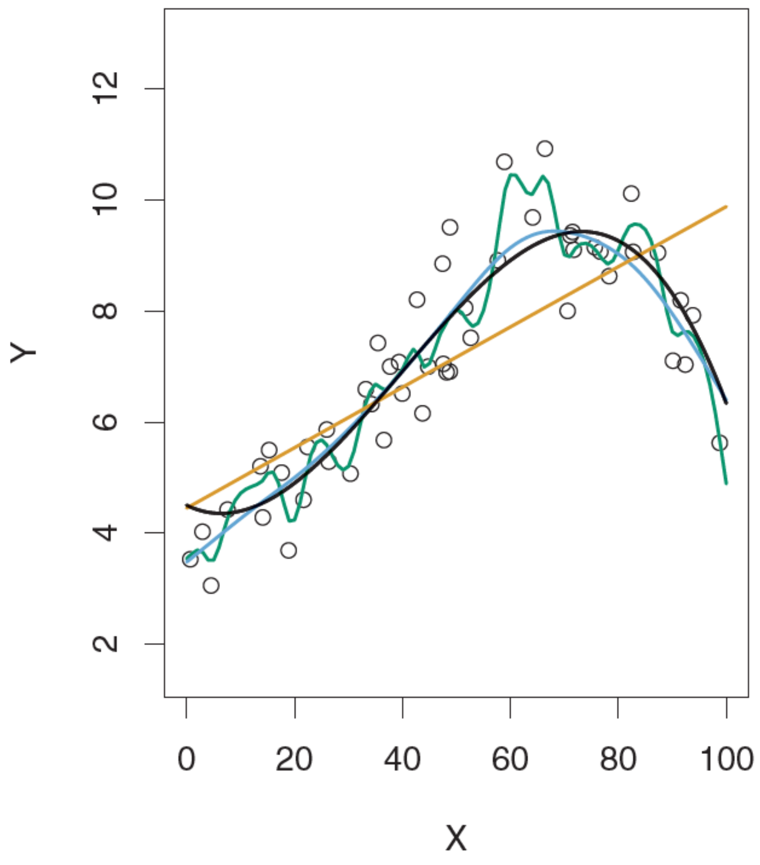
Example: Overfitting (True Relationship Slightly Complicated)

Simple (linear): gold line

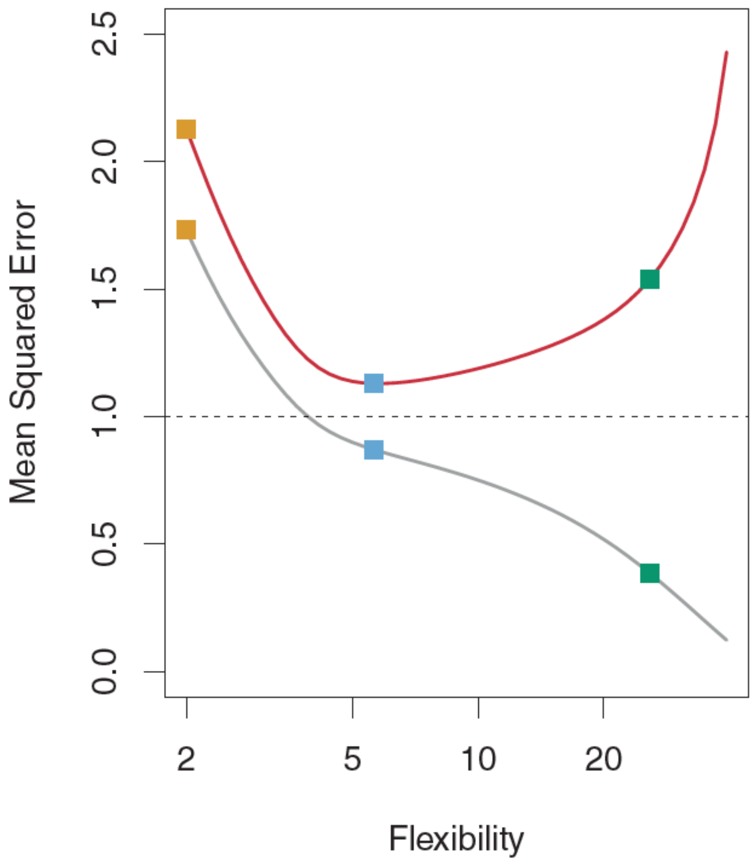
Moderate complexity (linear): blue line

Very complicated: green line

Training Data



Results on Test Data



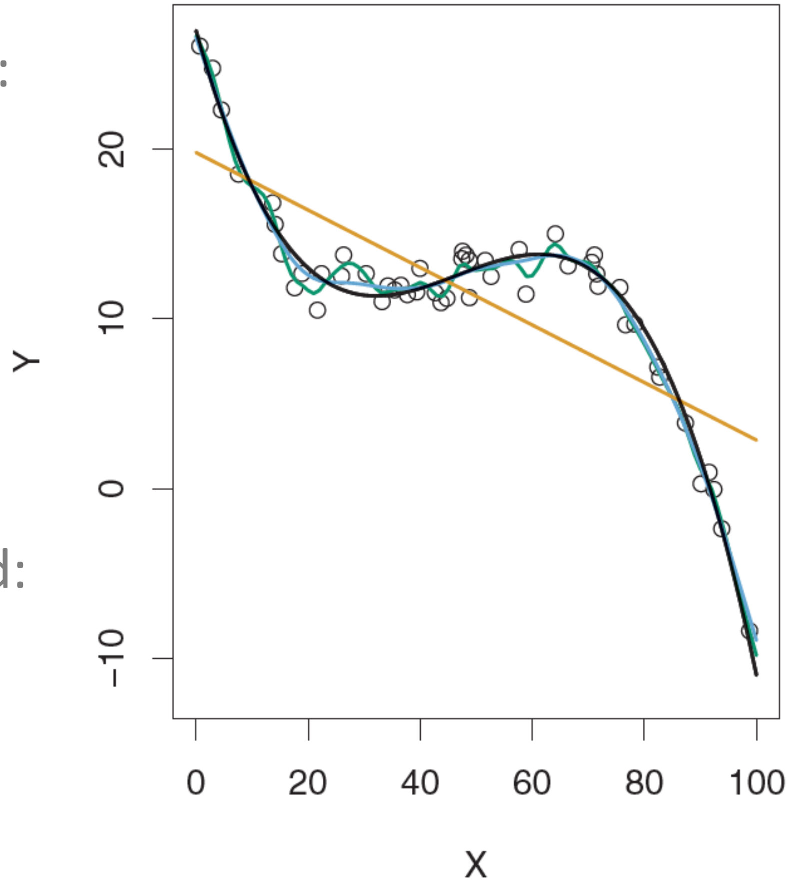
Example: Overfitting (True Relationship Very Complicated)

Simple (linear): gold line

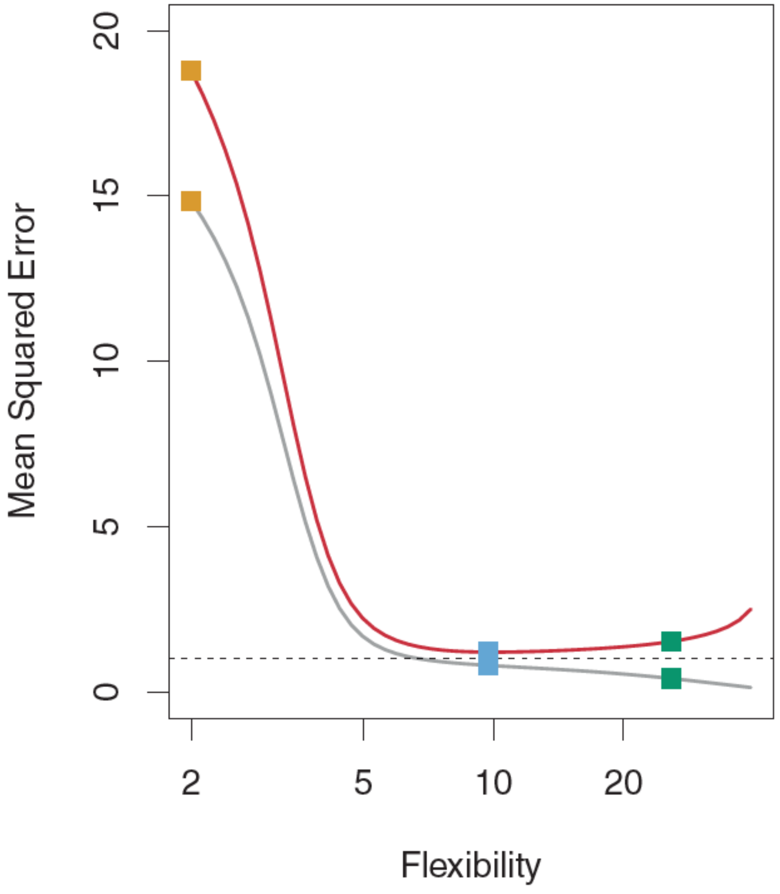
Moderate complexity (linear): blue line

Very complicated: green line

Training Data



Results on Test Data



Class 6 Summary

- **Regression** predicts a continuous outcome
- **Classification** predicts a discrete outcome
- **Supervised** models contain a y_i (target/outcome variable) for every x_i (descriptor variables)
- **Unsupervised** models contain only x_i
- **Training** data is the data we will use to estimate our model parameters
- **Testing** data is the data used to evaluate our model performance
- **Bias:** tendency of an in-sample statistic to over or underestimate the true value
- **Variance:** tendency to noisily estimate that statistic
- **Bias-Variance** tradeoff is the idea that total error is composed of both bias and variance, and we care about minimizing both together