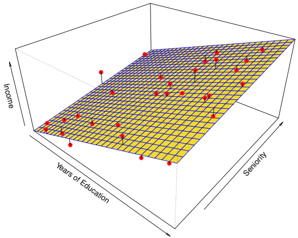
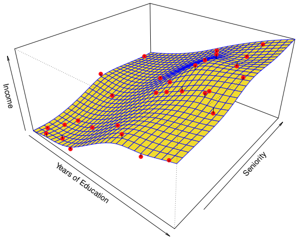
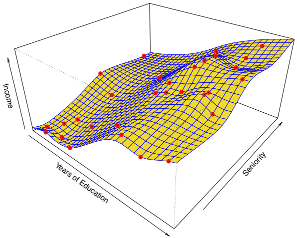
Example: If We Don't Know True $f(X)$, How Do We Pick?



Least Flexibility



Medium Flexibility



Most Flexibility

Source: James, Witten, Hastie & Tibshirani (2021)

Trade-Offs

- Why not just use more flexible (non-parametric) methods that match the data better?

1. Trade-Off 1: model flexibility vs. interpretability
    - Non-parametric models fit a wider range of possible patterns of $f(X)$
    - But parametric models are easier to interpret (& explain)
        - β_1 is the average change in Y for a 1 unit increase in X_1, holding all else equal vs. ....
2. Trade-Off 2: bias vs. variance trade-off
    - Even if first reason isn't relevant, simpler models are *often* more accurate!
    - Intuition: hard to fit a more flexible model w/out overfitting
    - In other words, there's a trade-off between under vs. overfitting a model

Some Definitions

- There are two types of data

1. Training data - observations of X & Y we use to teach our method to estimate $f(X)$

$$Train = \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$$

2. Test data - *new* observations of X & Y that we use to asses our estimate of $f(X)$

$$Test = \{(x_1, y_1), (x_2, y_2), \ldots (x_m, y_m)\}$$

Training vs. Test MSEs

- Suppose we fit a model $\hat{f}(X)$ to some training data by minimizing

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

- How do we figure out how well the model does vs. other possible models?
  - Can't use MSE^{Train} b/se it's biased towards overfitting
- $\Rightarrow$ Compute average MSE using (new) test data

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

What Determines MSE^{Test}?

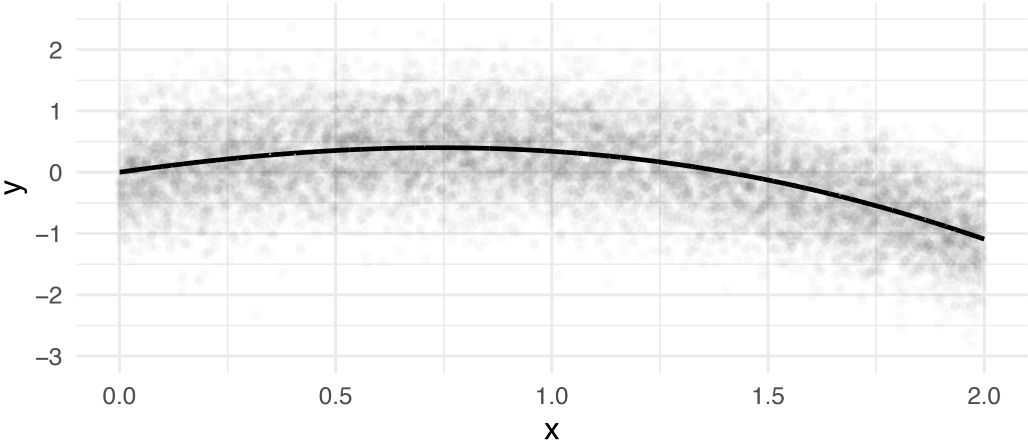
- For a given model fit on training data, let (x_0, y_0) be a test data observation
- The expected test MSE is

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = E\left(f(x_0) + \varepsilon - \hat{f}(x_0)\right)^2$$
$$= \underbrace{\underbrace{var\left(\hat{f}(x_0)\right)}_{\text{Training Variance}} + \underbrace{\left(Bias\left(\hat{f}(x_0)\right)\right)^2}_{\text{Model Bias}}}_{\text{Reducible}} + \underbrace{var(\varepsilon)}_{\text{Irreducible}}$$

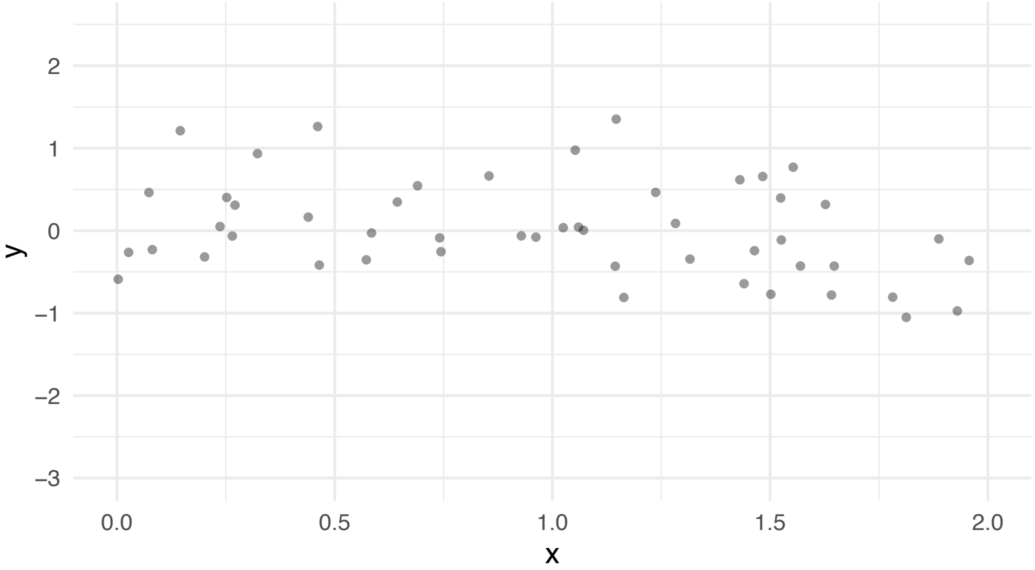
- Intuition: expected test MSE is determined by variability from
  - Training dataset ($var\left(\hat{f}(x_0)\right)$)
  - How (flexibly) we choose to model Y ($Bias\left(\hat{f}(x_0)\right)$)
  - And the irreducible error ($var(\varepsilon)$)

Consider a data generating process ($\mathcal{P}$)

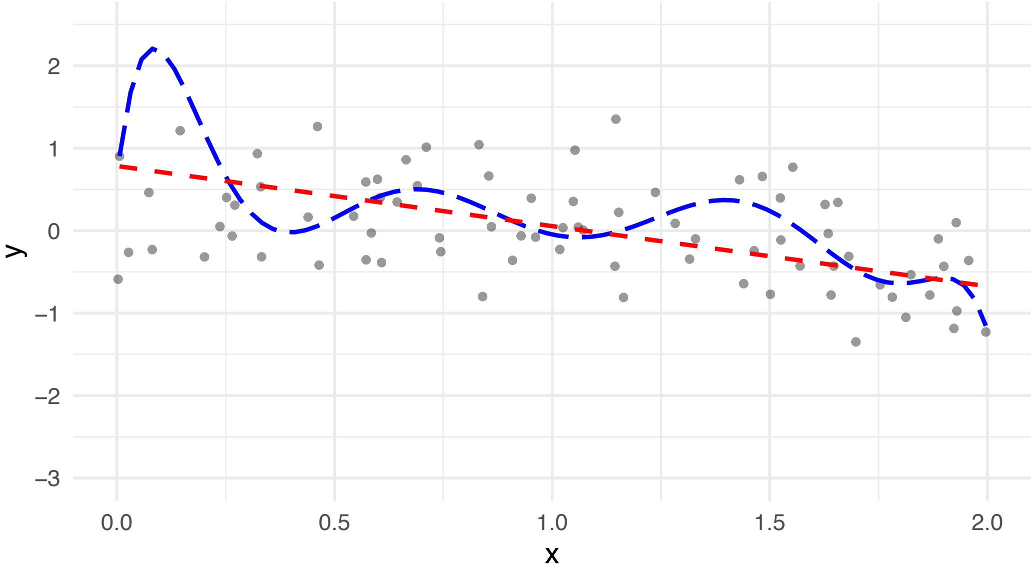
$$y_i = sin(x_i) - .5 \times x_i^2 + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon)$$



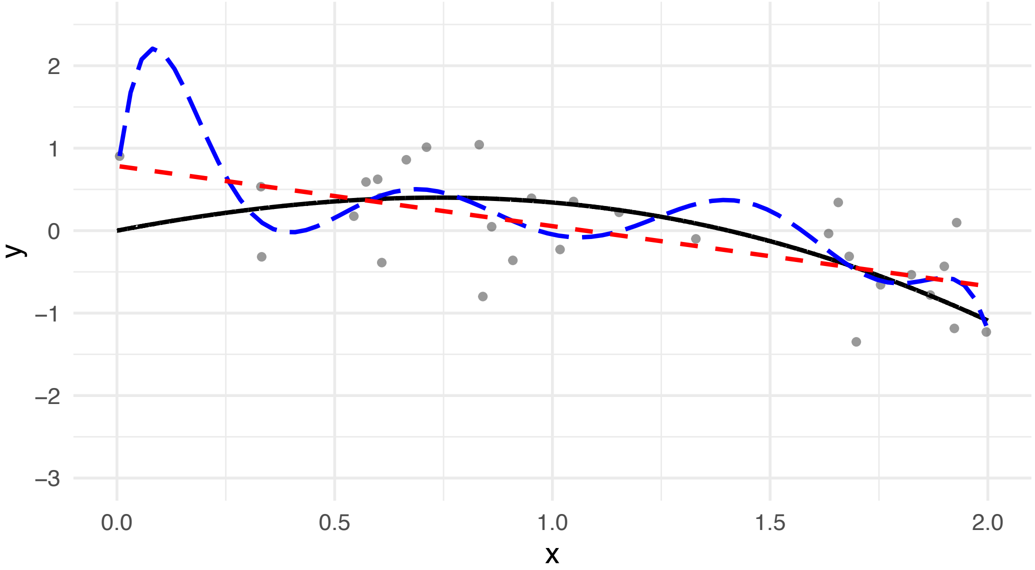
We take a random sample from $\mathcal{P}$



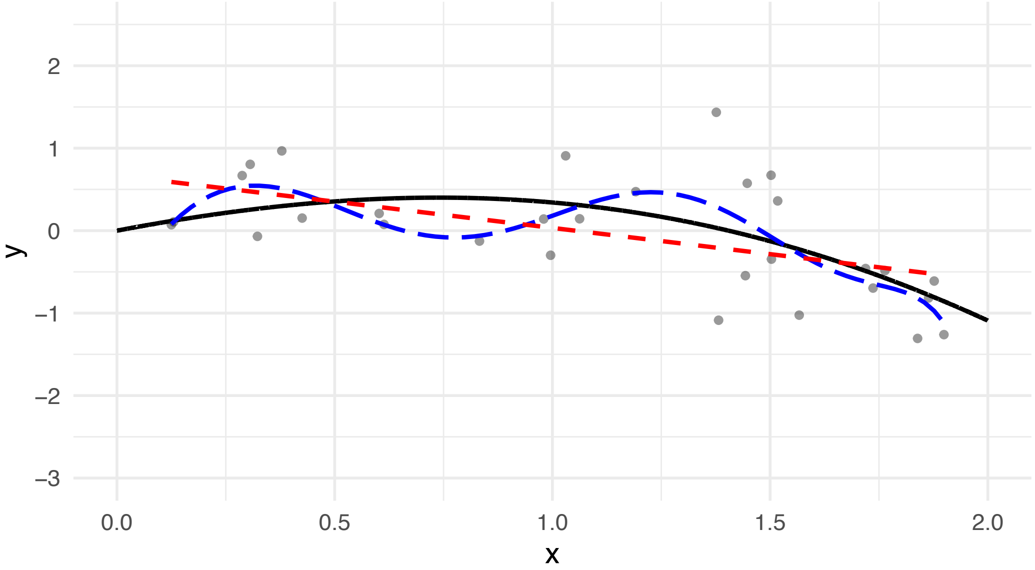
Fit two models: linear and an 8th degree polynomial



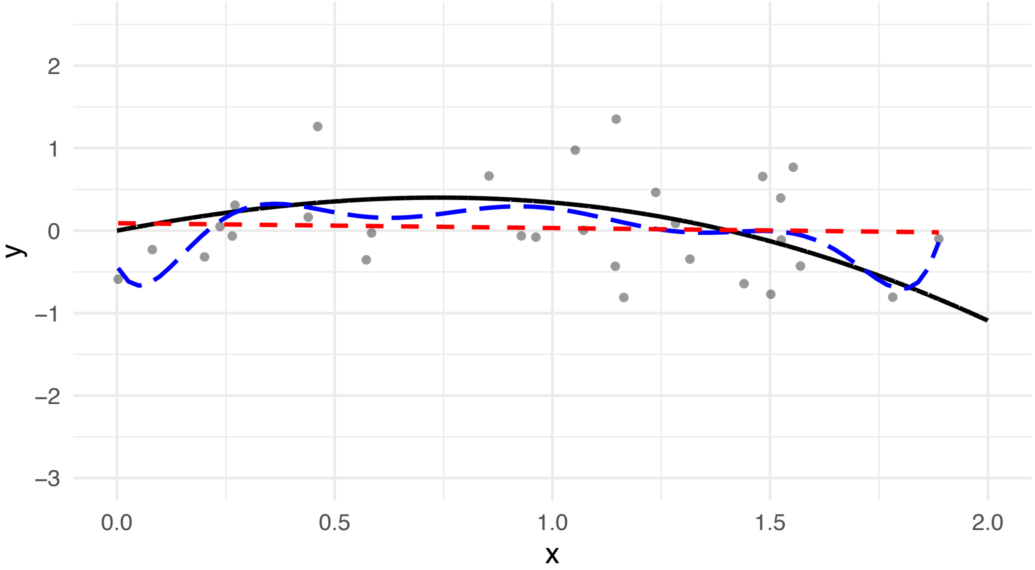
Fit two models: linear and an 8th degree polynomial



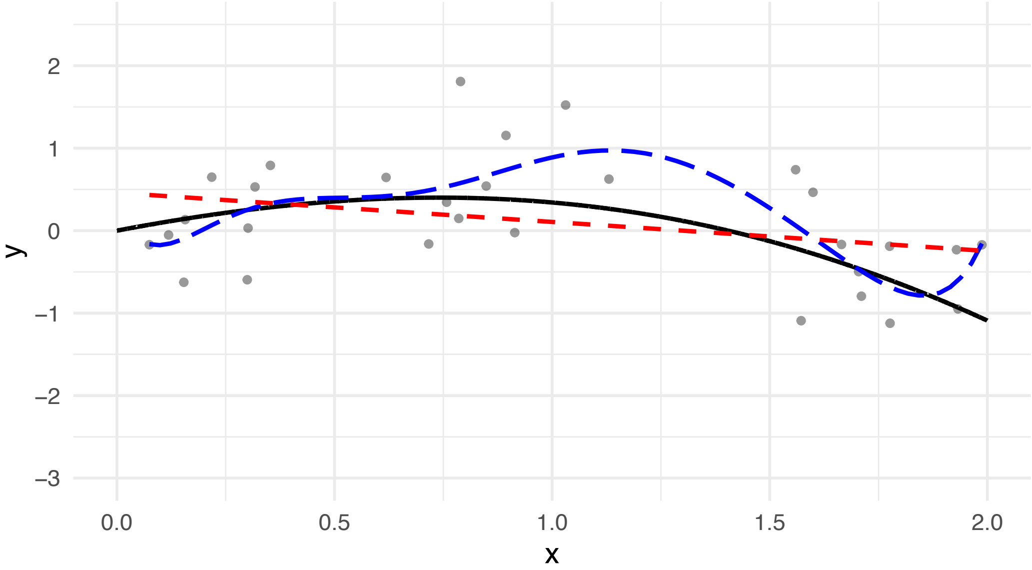
. . . and sample from $\mathcal{P}$ again



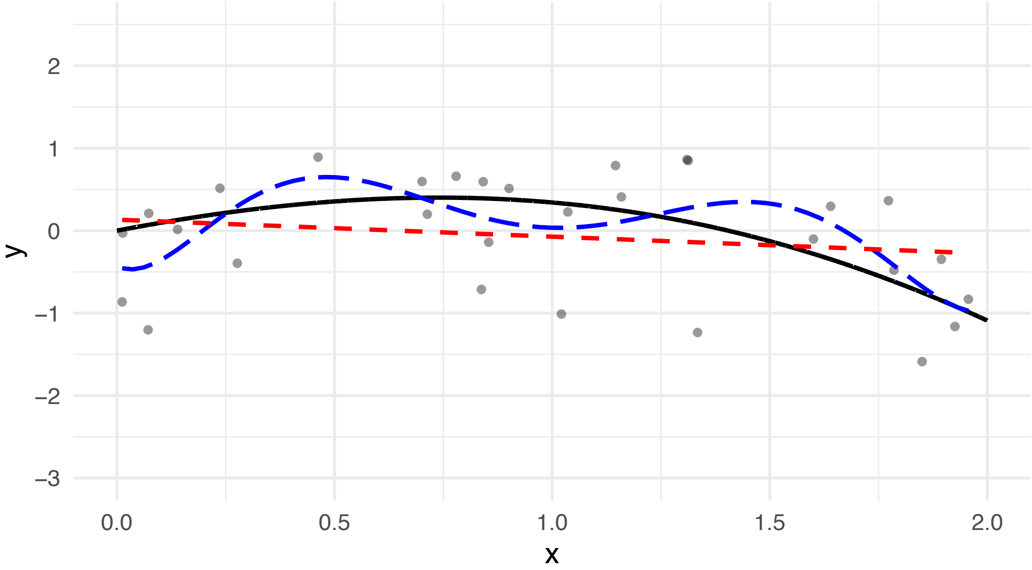
. . . and sample from $\mathcal{P}$ again



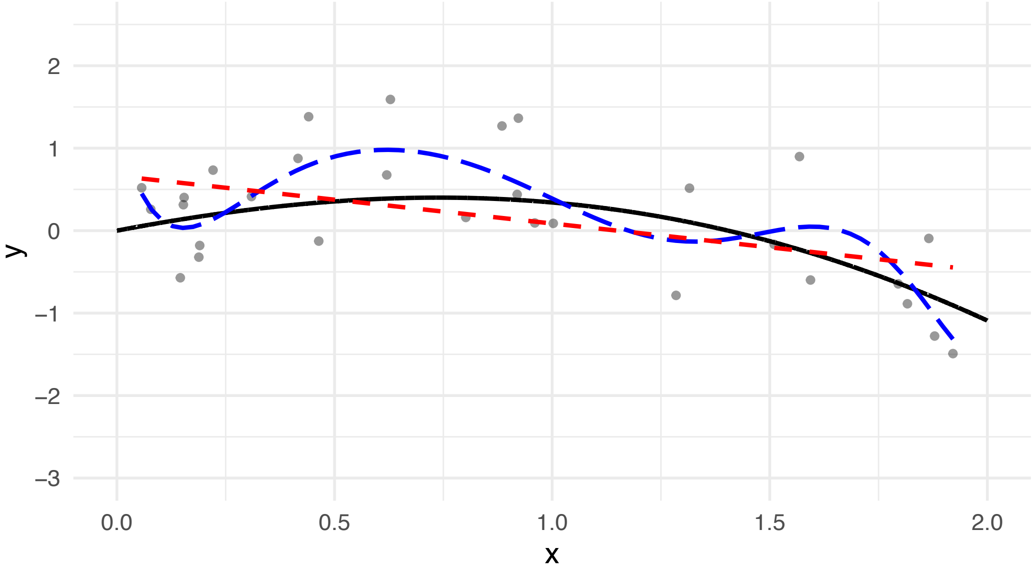
. . . and sample from $\mathcal{P}$ again



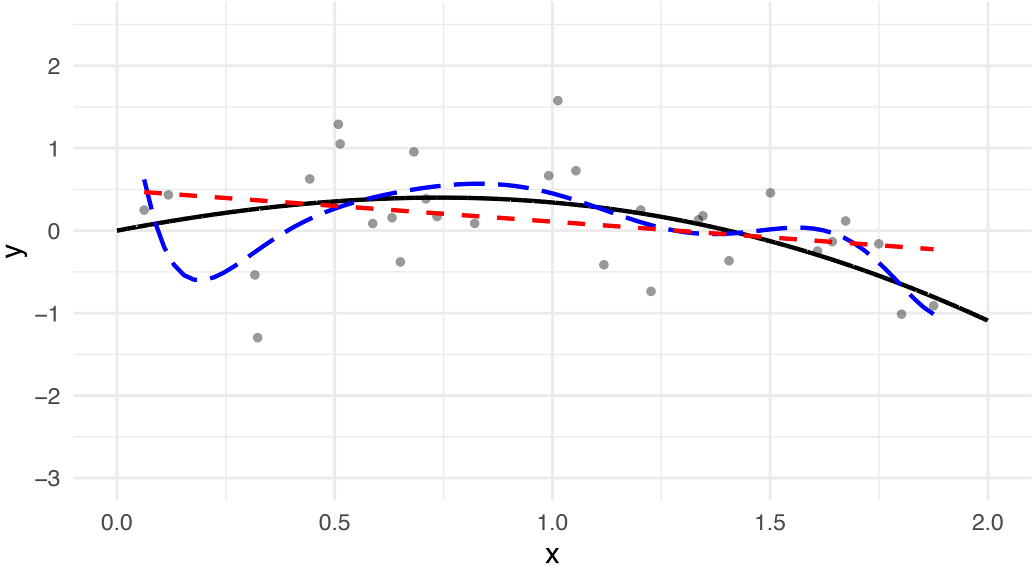
... and sample from $\mathcal{P}$ again



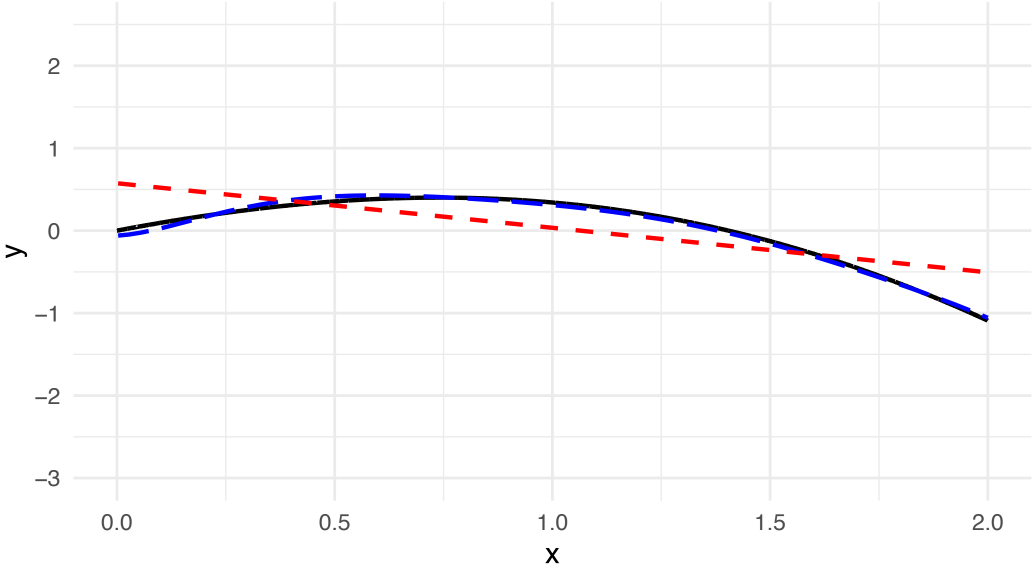
. . . and sample from $\mathcal{P}$ again



. . . and sample from $\mathcal{P}$ again



If we did this many times and averaged the functions



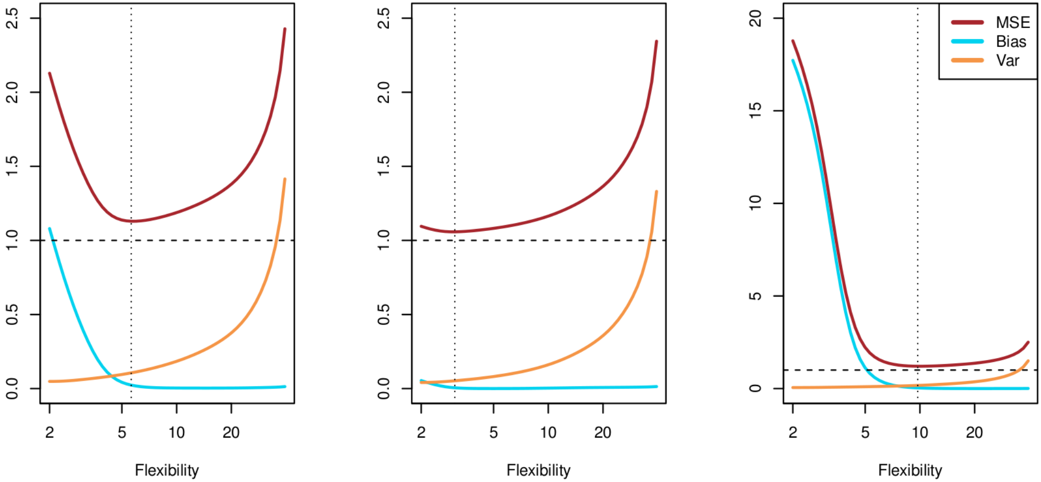
What Determines MSE^{Test}?

- For a given model fit on training data, let (x_0, y_0) be a test data observation
- The expected test MSE is

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = E\left(f(x_0) + \varepsilon - \hat{f}(x_0)\right)^2$$
$$= \underbrace{\underbrace{var\left(\hat{f}(x_0)\right)}_{\text{Training Variance}} + \underbrace{\left(Bias\left(\hat{f}(x_0)\right)\right)^2}_{\text{Model Bias}}}_{\text{Reducible}} + \underbrace{var(\varepsilon)}_{\text{Irreducible}}$$

- Intuition: expected test MSE is determined by variability from
    - Training dataset ($var\left(\hat{f}(x_0)\right)$)
    - How (flexibly) we choose to model Y ($Bias\left(\hat{f}(x_0)\right)$)
    - And the irreducible error ($var(\varepsilon)$)

Bias vs. Variance Trade-Off for Three Examples



Source: James, Witten, Hastie & Tibshirani (2021)

Bias vs. Variance Trade-Off in Words

- U-shaped test MSE curves b/se two competing factors in determining model accuracy
  - Bias - error that is introduced by simplifying a complex, real-life problem w/ a model
    - Model flexibility $\uparrow \Rightarrow$ bias $\downarrow$
    - The more flexible/complex a model, the less bias it will generally have
  - Variance - how much $\hat{f}(X)$ would change by if you had a different training data set
    - Model flexibility $\uparrow \Rightarrow$ variance $\uparrow$
    - Generally, the more flexible a model is, the more variance it has
- $\Rightarrow$ Choosing flexibility based on average test error results in a bias-variance trade-off

Implications of Bias vs. Variance Trade-Off

- What's the bias-variance trade-off mean for doing ML?
  - Recall: trade-off between under vs. overfitting a model
- No guarantee method w/ smallest training MSE will have smallest test MSE
  - Generally, more flexible methods have lower training MSEs $\Rightarrow$ will "fit" or explain the training data very well
  - But test MSE may be higher for a more flexible method than a simpler approach!!!