# QTM 347 Machine Learning

# Lecture 2: Bias-variance decomposition

Ruoxuan Xiong

Suggested reading: ISL Chapter 2

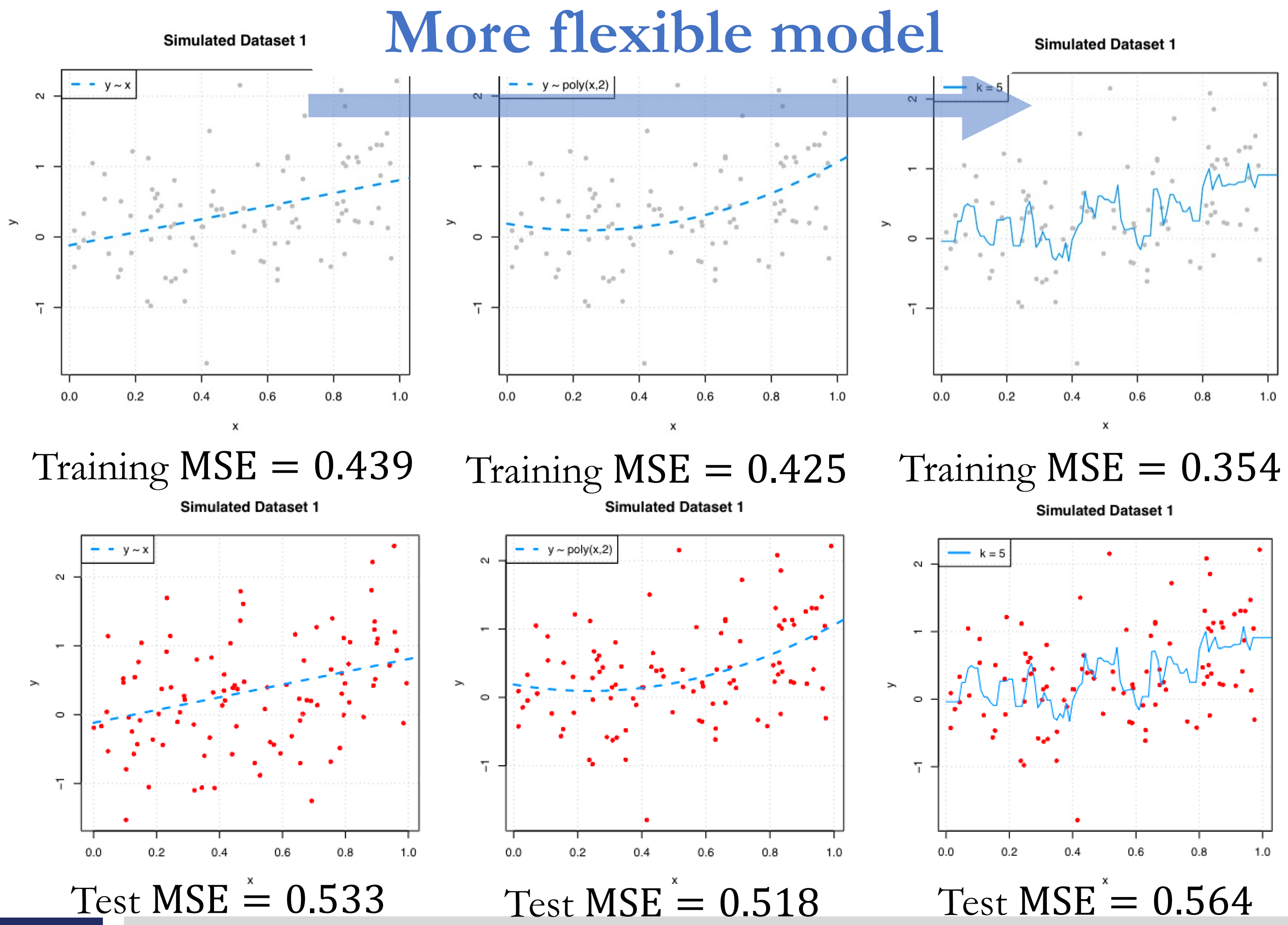
# Lecture plan

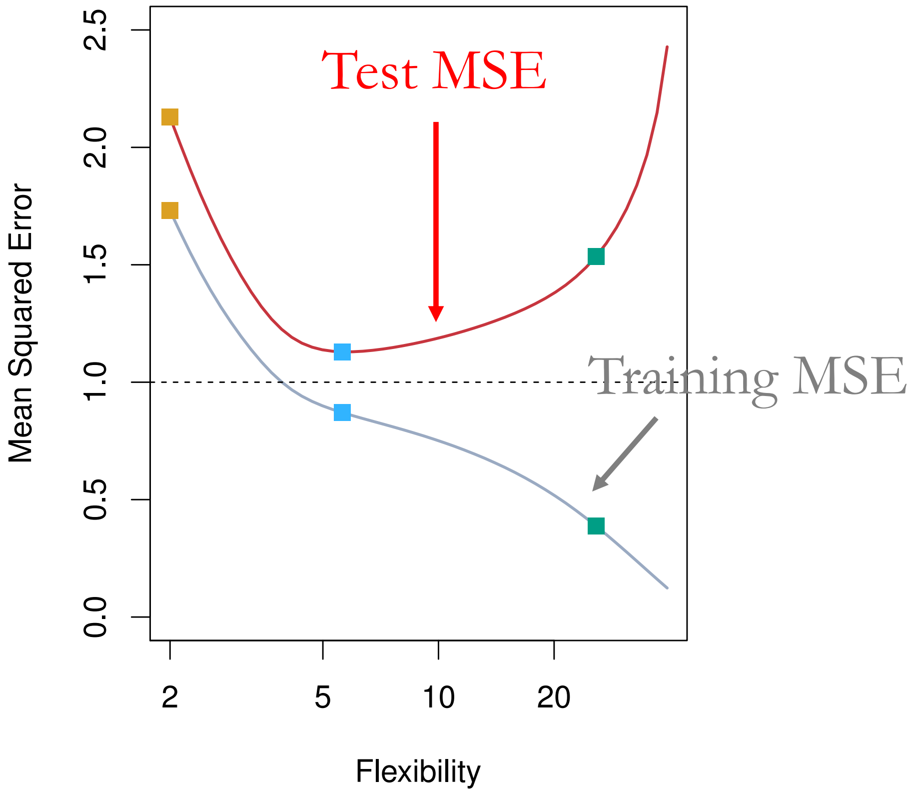
- Bias-variance decomposition for regression problems (MSE)

# Training/Test data and training/test MSE

- **Training data**: the data, $(X_1, Y_1)$, $(X_2, Y_2)$, $\cdots$, $(X_n, Y_n)$, that are used to fit $f$
- **Training MSE**: $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i))^2$
- **Test data**: the data, $(X_1', Y_1'), (X_2', Y_2'), \ldots, (X_m', Y_m')$, that are previous unseen and not used to fit $f$
- **Test MSE**: $\text{MSE} = \frac{1}{m}\sum_{i=1}^{m}(Y_i' - \hat{f}(X_i'))^2$
- We care test MSE more than training MSE
- However, *a low training MSE does not imply a low test MSE*…

# MSE varies with model flexibility



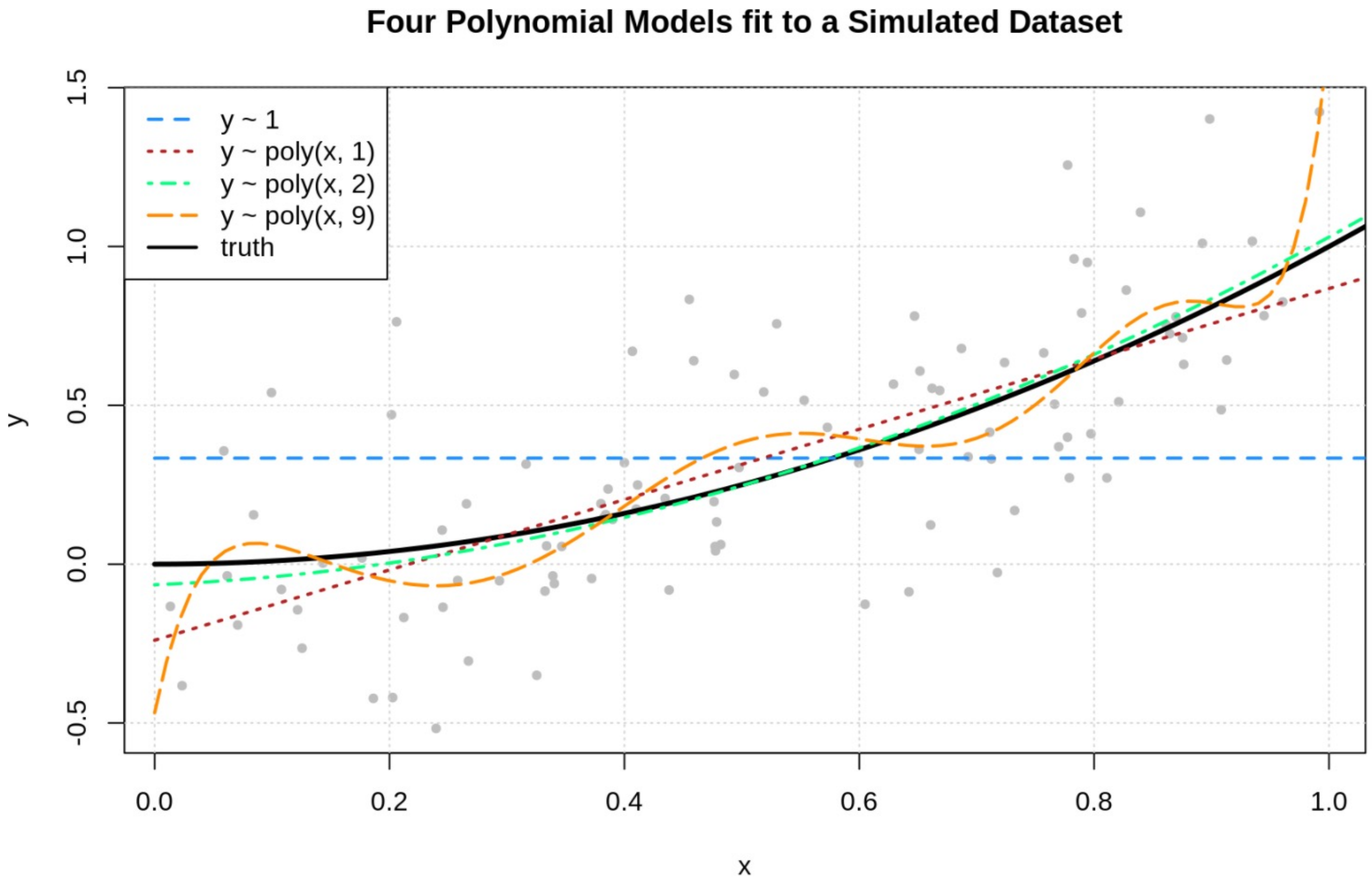
# From now on, let us study why MSE varies with model flexibility

- Suppose the *true regression function* is $f(x) = x^2$ ($x$ is a scalar)

- We consider *training the following models* to learn this function
  - A constant function: $\hat{f}_0(x) = \hat{\beta}_0$
  - A linear function: $\hat{f}_1(x) = \hat{\beta}_0 + x \cdot \hat{\beta}_1$
  - A quadratic function: $\hat{f}_2(x) = \hat{\beta}_0 + x \cdot \hat{\beta}_1 + x^2 \cdot \hat{\beta}_2$
  - A ninth degree polynomial function: $\hat{f}_9(x) = \hat{\beta}_0 + x \cdot \hat{\beta}_1 + \cdots + x^9 \cdot \hat{\beta}_9$

  - The model is *more flexible* if the *polynomial degree is larger*

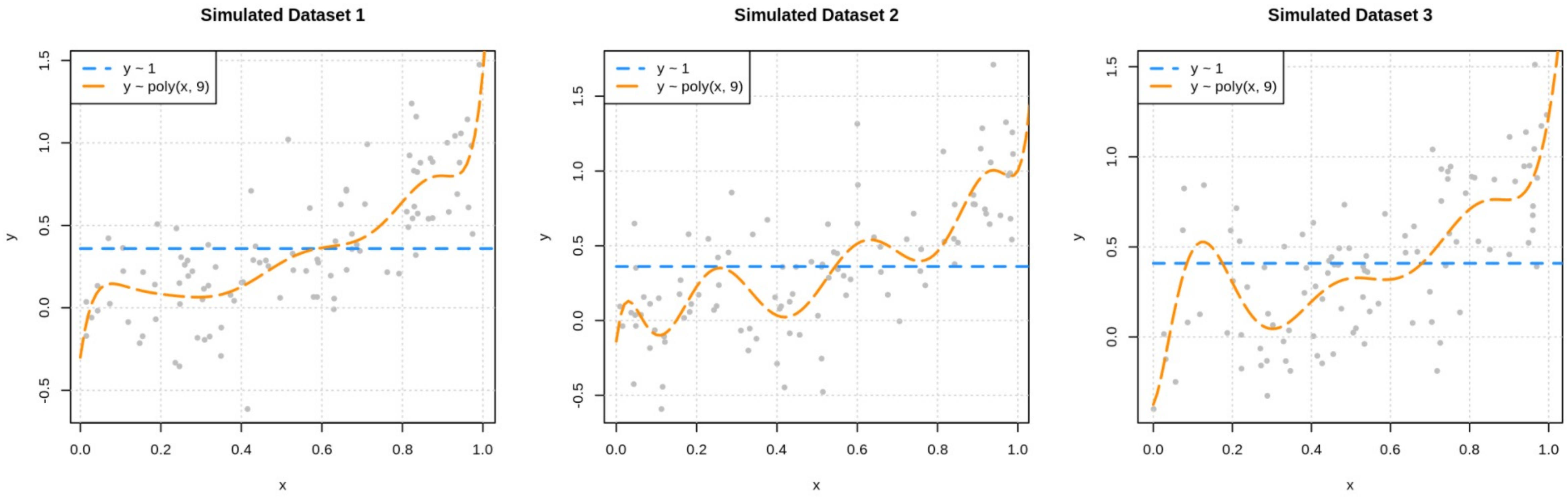
# Four fitted models




- Zero predictor model fits poorly
- Linear model is reasonable
- Quadratic model fits much better
- Ninth degree model seems rather wild

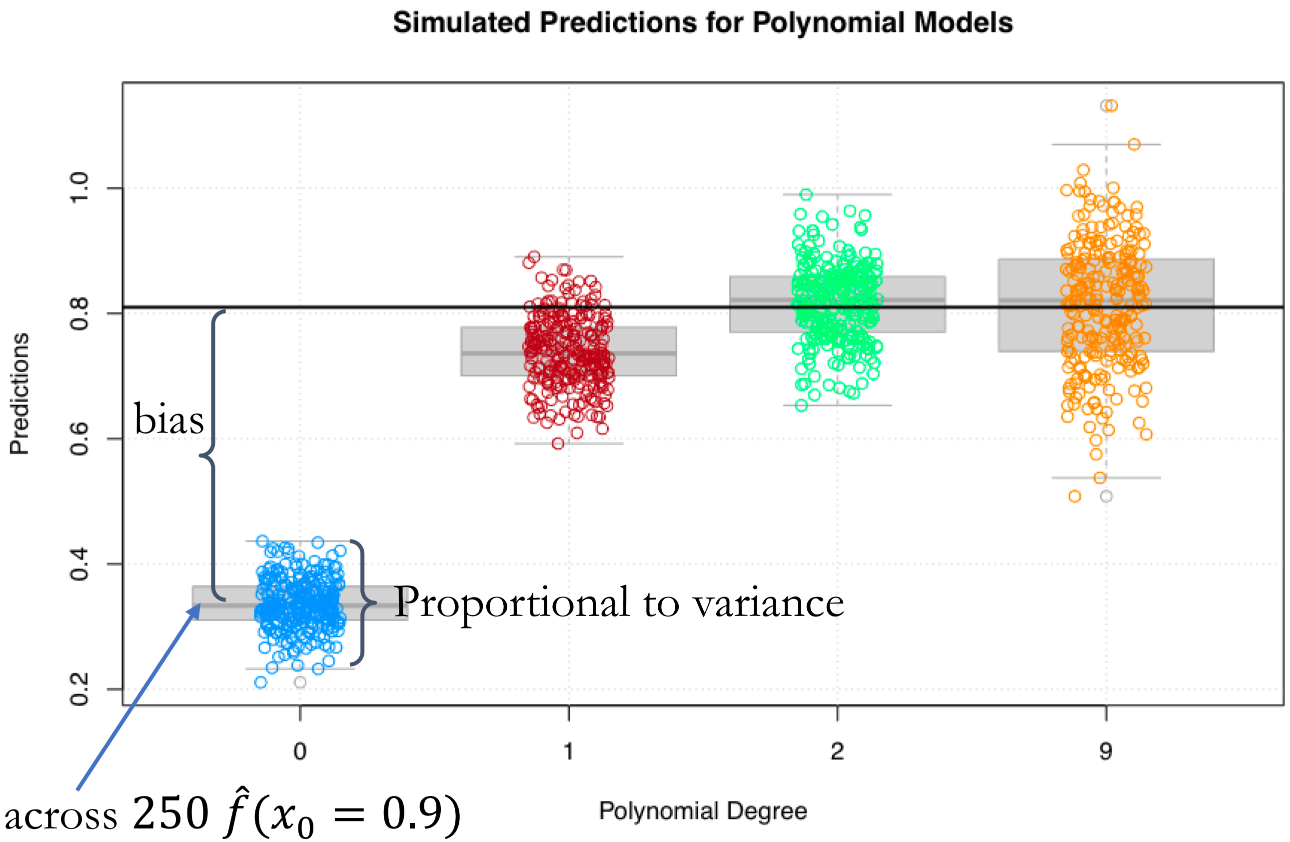
# If we fit models on three additional datasets

- In general, the fitted model depends on the training data
- The zero predictor $\hat{f}_0(x)$ slightly varies, but the ninth-degree polynomial varies $\hat{f}_9(x)$ quite a bit
- We call this the variance of a model: the variance of $\hat{f}_0(x)$ is smaller than the variance of $\hat{f}_9(x)$



# Suppose we are interested in predicting $f(x_0)$

- Test point: $x_0 = 0.9$
- The truth, $f(x_0 = 0.9) = x_0^2 = 0.81$
- We have 250 datasets
- For each dataset, we fit polynomials with degree $0, 1, 2, 9$, and plot $\hat{f}(x_0 = 0.9)$




Average prediction across 250 $\hat{f}(x_0 = 0.9)$

# Suppose we are interested in predicting $f(x_0)$

- The squared bias

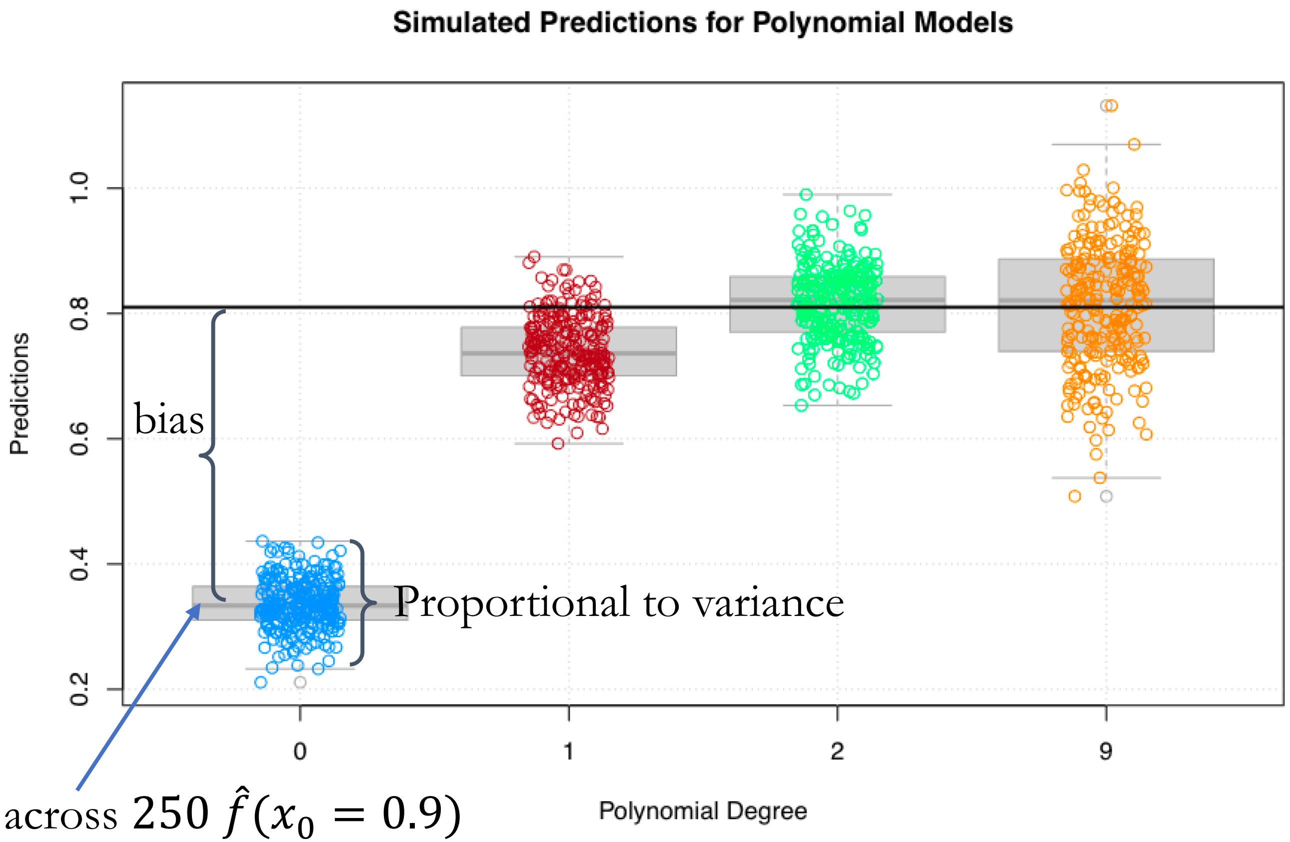
  $$\hat{f}_2(x) \approx \hat{f}_9(x) < \hat{f}_1(x) < \hat{f}_0(x)$$

  - Increasing degree from 2 to 9 does not further reduce bias

- Variance

  $$\hat{f}_0(x) < \hat{f}_1(x) < \hat{f}_2(x) < \hat{f}_9(x)$$

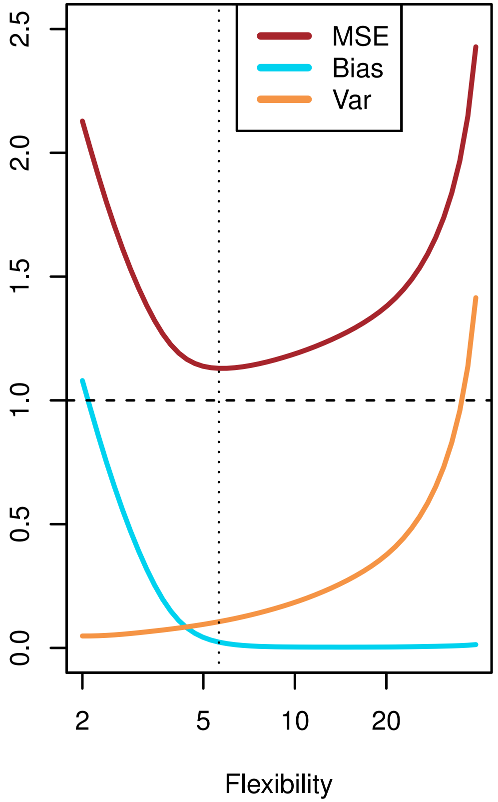
  - Increasing degree increases variance




Average prediction across 250 $\hat{f}(x_0 = 0.9)$

# Combining bias and variance

- MSE depends on both bias and variance




- MSE can be decomposed into variance plus the square of bias
- Next we show how…

# Let us analyze MSE

- Suppose the true function is $f$
- The response follows $Y = f(X) + \varepsilon$, with $\mathrm{E}[\varepsilon] = 0$
- For notation simplicity, let $\mathcal{D} = \{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ be the training data
- $\hat{f}$ *depends on* $\mathcal{D}$
  - If we change the training data $\mathcal{D}$, we get a different estimate $\hat{f}$
- Let $x_0$ be a fixed test point. The MSE at $x_0$ can be decomposed as

$$\mathrm{MSE}(x_0) = \underbrace{E_{Y|X,\mathcal{D}}}_{\text{Expected value over } \mathcal{D} \text{ and } Y|X}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x_0\right] = \underbrace{E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

# Decomposition of MSE

- The MSE at $x_0$ can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X,\mathcal{D}}}_{\text{Expected value over } Y|X \text{ and } \mathcal{D}}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x_0\right] = \underbrace{E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

  - MSE at $x_0$ is also called the expected prediction error for a random $Y$ given a fixed $X = x_0$ and a random $\hat{f}$
  - Two sources of *randomness*:
    - Given that $X = x_0$, $Y$ is random because $Y = f(X) + \varepsilon$ and $\varepsilon$ is random
    - $\hat{f}$ is random because it depends on (randomly sampled) training data $\mathcal{D}$

# Irreducible error

- The MSE at $x_0$ can be decomposed as

$$\mathrm{MSE}(x_0) = \underbrace{E_{Y|X,\mathcal{D}}}_{\text{Expected value over } Y|X \text{ and } \mathcal{D}}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x_0\right] = \underbrace{E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

  - Irreducible error: the variance of $Y$ given that $X = x_0$
    - If $Y = f(X) + \varepsilon$ with $\mathrm{E}[\varepsilon] = 0$ and $\mathrm{V}[\varepsilon] = \sigma_\varepsilon^2$ , then $V_{Y|X}[Y \mid X = x_0] = \sigma_\varepsilon^2$
    - Irreducible error can not be reduced for any $\hat{f}$

# Reducible error

- The MSE at $x_0$ can be decomposed as

$$\text{MSE}(x_0) = \underbrace{E_{Y|X,\mathcal{D}}}_{\text{Expected value over } Y|X \text{ and } \mathcal{D}}\left[\left(Y - \hat{f}(X)\right)^2 \mid X = x_0\right] = E_{\mathcal{D}}\underbrace{\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right]}_{\text{Reducible error}} + \underbrace{V_{Y|X}[Y \mid X = x_0]}_{\text{Irreducible error}}$$

  - Reducible error: the expected squared error by using $\hat{f}(x_0)$ to estimate $f(x_0)$
    - The only thing that is random is $\mathcal{D}$, the training data used to obtain $\hat{f}$ (both $f$ and $x_0$ are fixed)
    - Reducible error can be reduced by using a better $\hat{f}$

# Bias-variance decomposition of reducible error

- *Reducible error* can be decomposed as the *squared bias* and *variance*

$$E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] = \underbrace{\left(f(x_0) - E_{\mathcal{D}}[\hat{f}(x_0)]\right)^2}_{\text{bias}^2(\hat{f}(x_0))} + \underbrace{E_{\mathcal{D}}\left[\left(\hat{f}(x_0) - E_{\mathcal{D}}[\hat{f}(x_0)]\right)^2\right]}_{\text{var}(\hat{f}(x_0))}$$

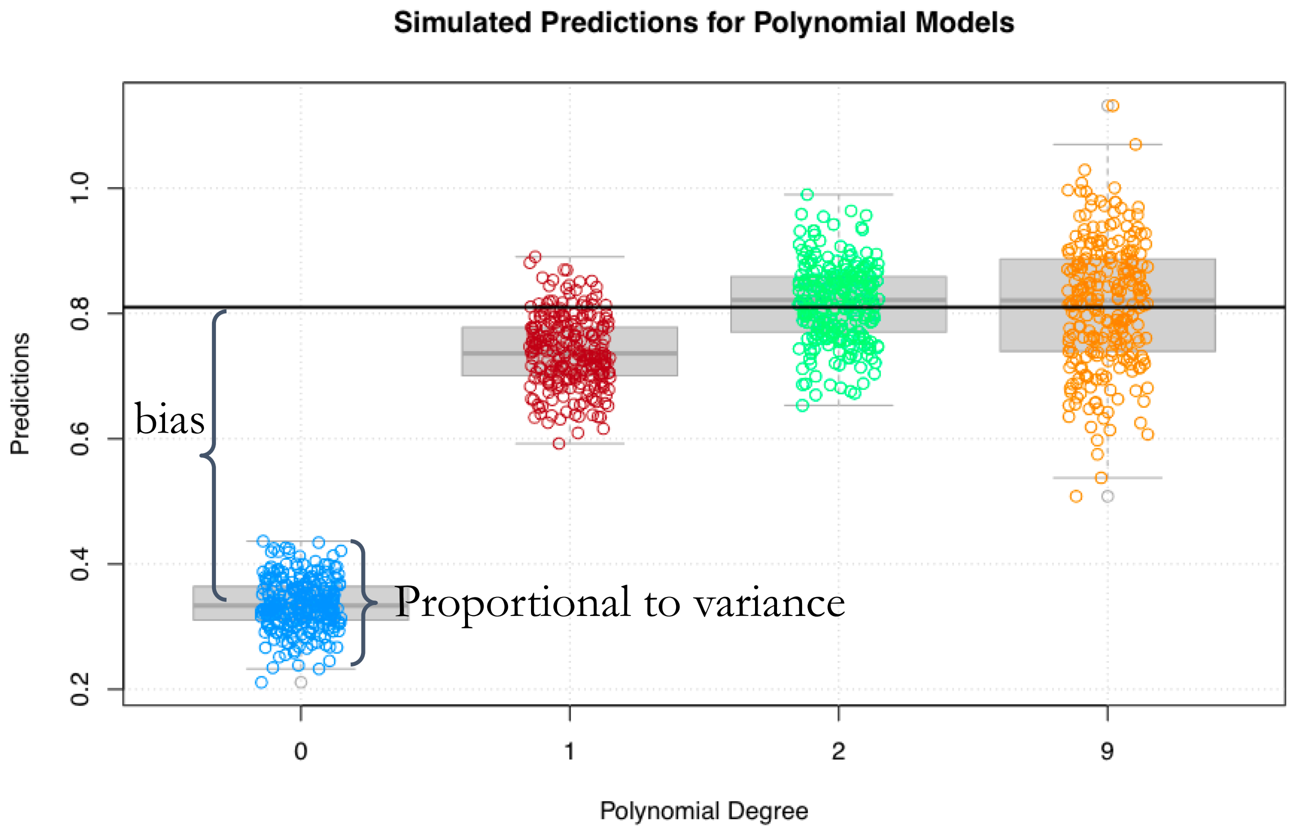
- *Take home exercise: Prove this decomposition*
- Hint: Use the property

$$E_{\mathcal{D}}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] = E_{\mathcal{D}}\left[\left(\underbrace{f(x_0) - E_{\mathcal{D}}[\hat{f}(x_0)]}_{A} + \underbrace{E_{\mathcal{D}}[\hat{f}(x_0)] - \hat{f}(x_0)}_{B}\right)^2\right]$$

$$E_{\mathcal{D}}[(A+B)^2] = A^2 + 2A\underbrace{E_{\mathcal{D}}[B]}_{=0} + E_{\mathcal{D}}[B^2] = A^2 + E_{\mathcal{D}}[B^2]$$

($A$ is nonrandom and $E_{\mathcal{D}}[A^2] = A^2$)

# Recall our toy example



# Summing up the decomposition

- The MSE at $x_0$ can be decomposed as

$$\mathrm{MSE}(x_0) = \underbrace{\mathrm{bias}^2\left(\hat{f}(x_0)\right) + \mathrm{var}\left(\hat{f}(x_0)\right)}_{\text{Reducible error}} + \underbrace{\sigma_\varepsilon^2}_{\text{Irreducible error}}$$

# Visualization of bias-variance decomposition

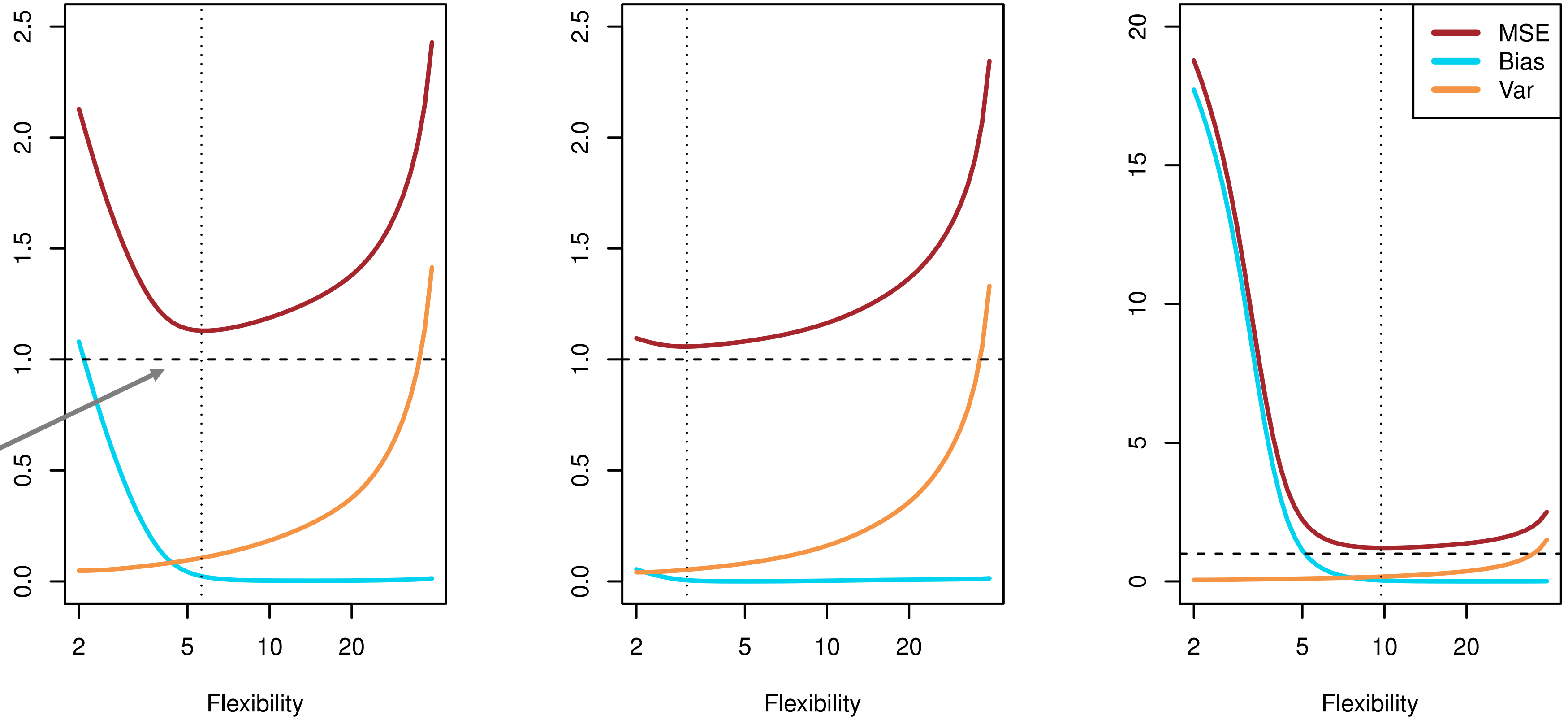
$Y = f(X) + \varepsilon$
$\mathrm{Var}(\varepsilon) = 1$

SNR: Signal-to-noise ratio
Signal measured by $f(X)$
Noise measured by $\varepsilon$

$f$ is complicated, low SNR

$f$ is simple, low SNR

$f$ is complicated, high SNR




Irreducible error