PSTAT 131/231: Introduction to Statistical Machine Learning

Guo Yu

Lecture 2
Bias-Variance Tradeoff

ISL Chapter 2

ESL Chapter 2 (for 231 students)

Office Hours (updated)

Guo Yu (Instructor)	**Tu/Th 5:15 - 6:15 PM**
Hanmo Li (TA)	**M 5:00 - 7:00 PM**
Chau Tran (TA)	**F 3:30 - 5:30 PM**
Eric Bletcher (ULA)	**M 1:00 - 2:00 PM**
Michael La (ULA)	**Tu 10:30 - 11:30 AM**

All (except Chau's) office hours will be held synchronously via Zoom meeting (links available on Gauchospace)

Homework 1 out tomorrow…

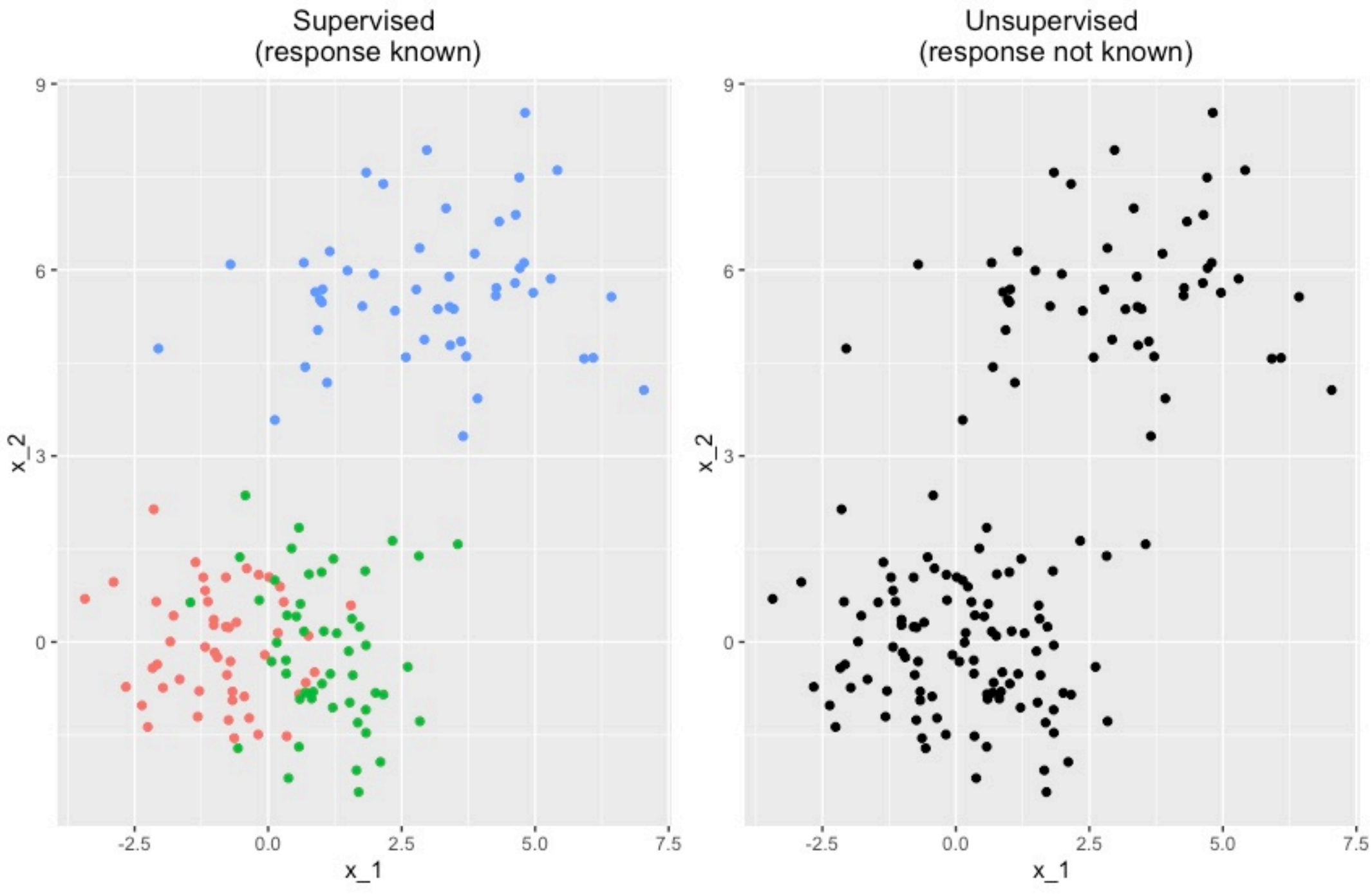
Due **Monday, Oct 11,** 2021 at **23:59**

Data processing / analysis in R

Make sure to attend lab session, starting this Wednesday!

Last time…

Supervised Learning vs Unsupervised Learning



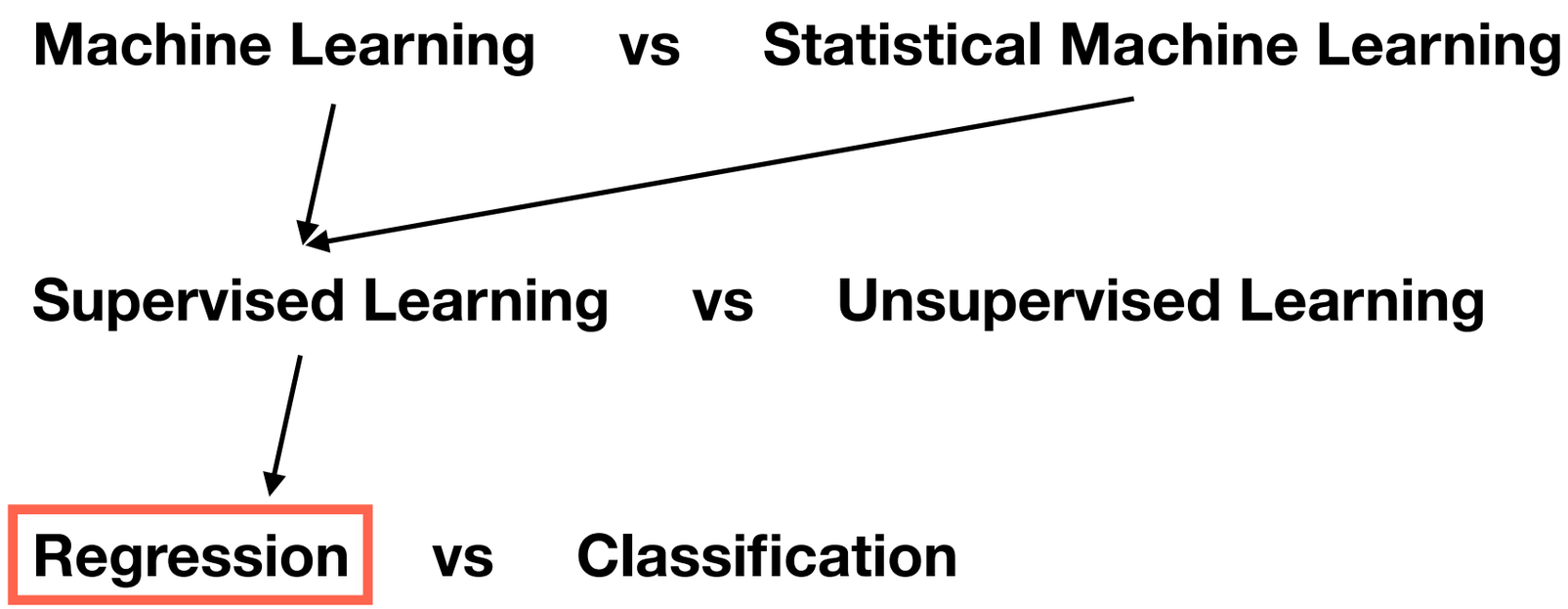
Last time…

Machine Learning vs Statistical Machine Learning

Supervised Learning vs Unsupervised Learning

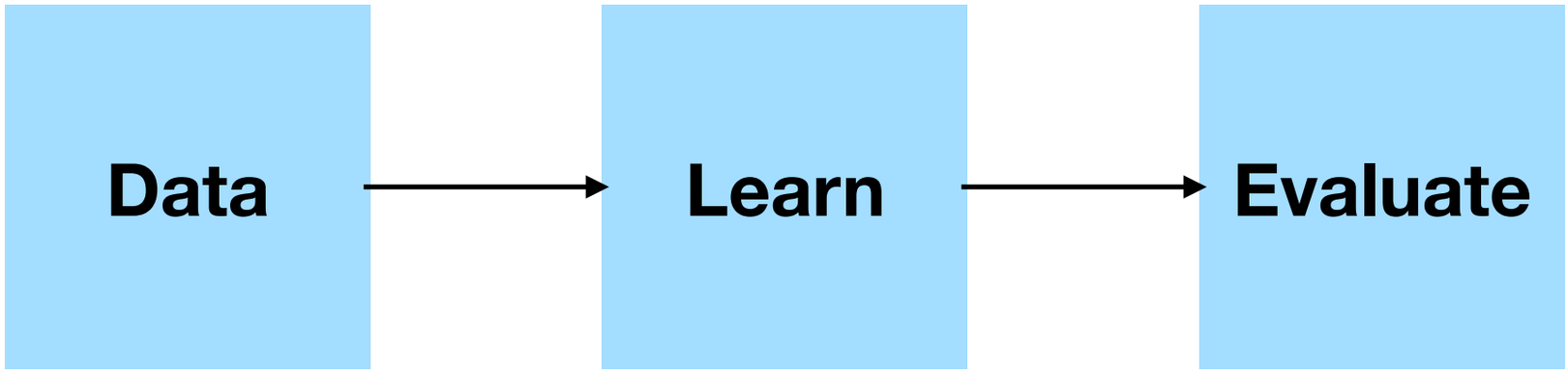
Regression vs Classification

Last time…


Machine Learning vs Statistical Machine Learning
Supervised Learning vs Unsupervised Learning
Regression vs Classification
today's focus

Regression setting

$$Y = f(X_1, \ldots, X_p) + \varepsilon$$



Training set

$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$

$\hat{f}$

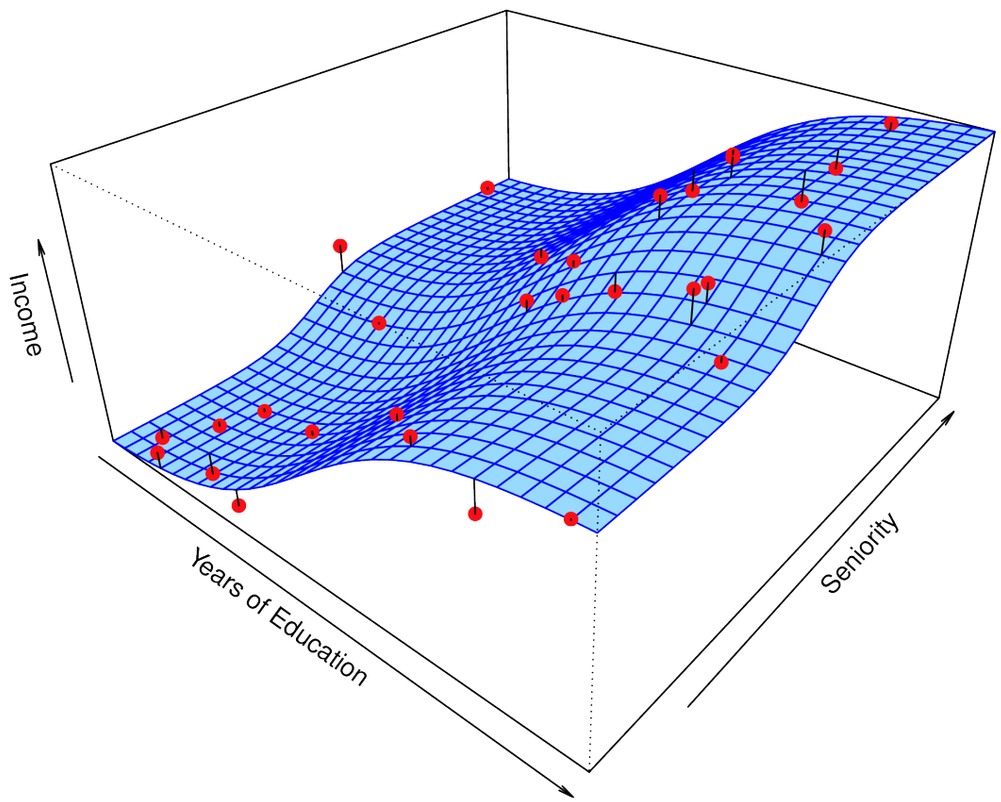
Different statistical learning methods give different $\hat{f}$, i.e., different models

Different $\hat{f}$ can be characterized by its flexibility/complexity

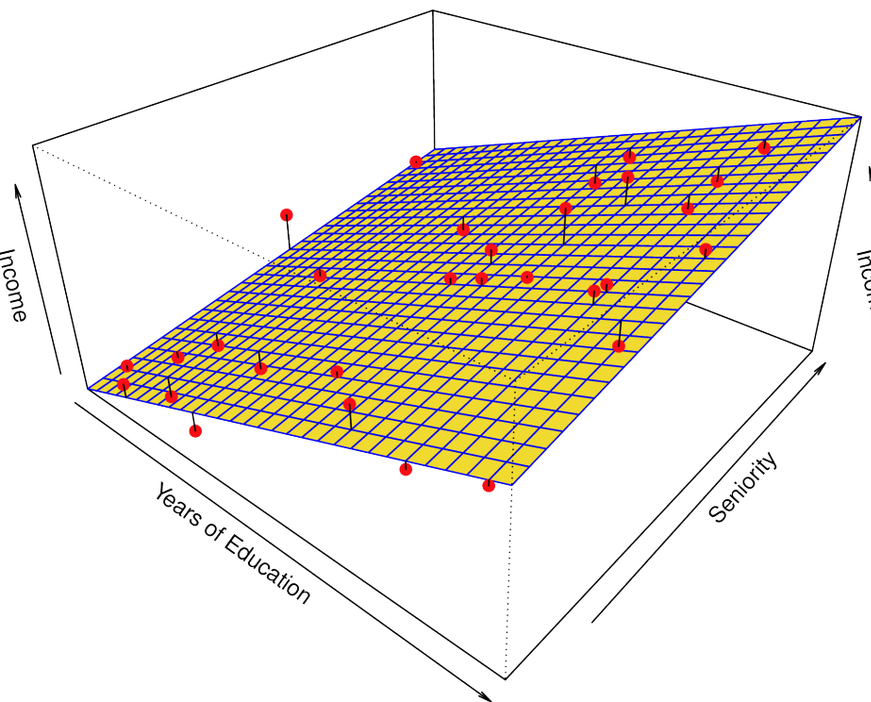
$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

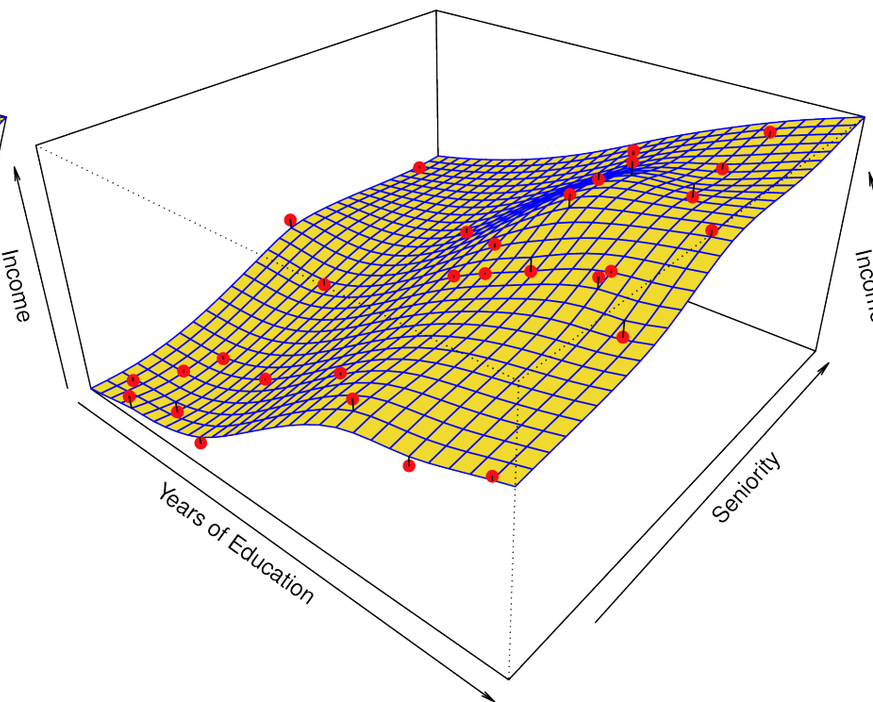
Model flexibility/complexity

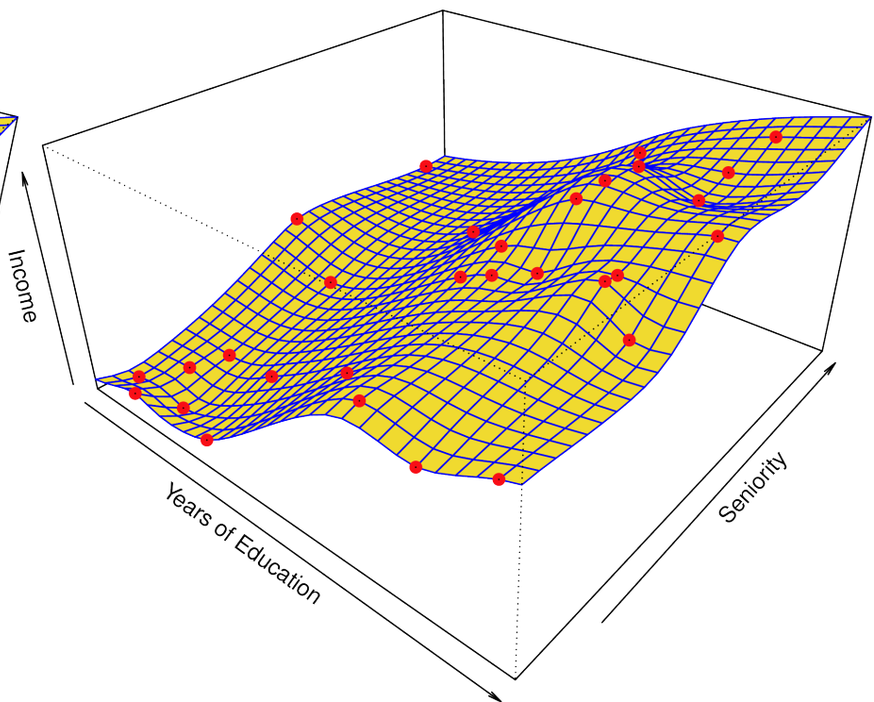
$$Y = f(X_1, X_2) + \varepsilon$$



model flexibility/complexity



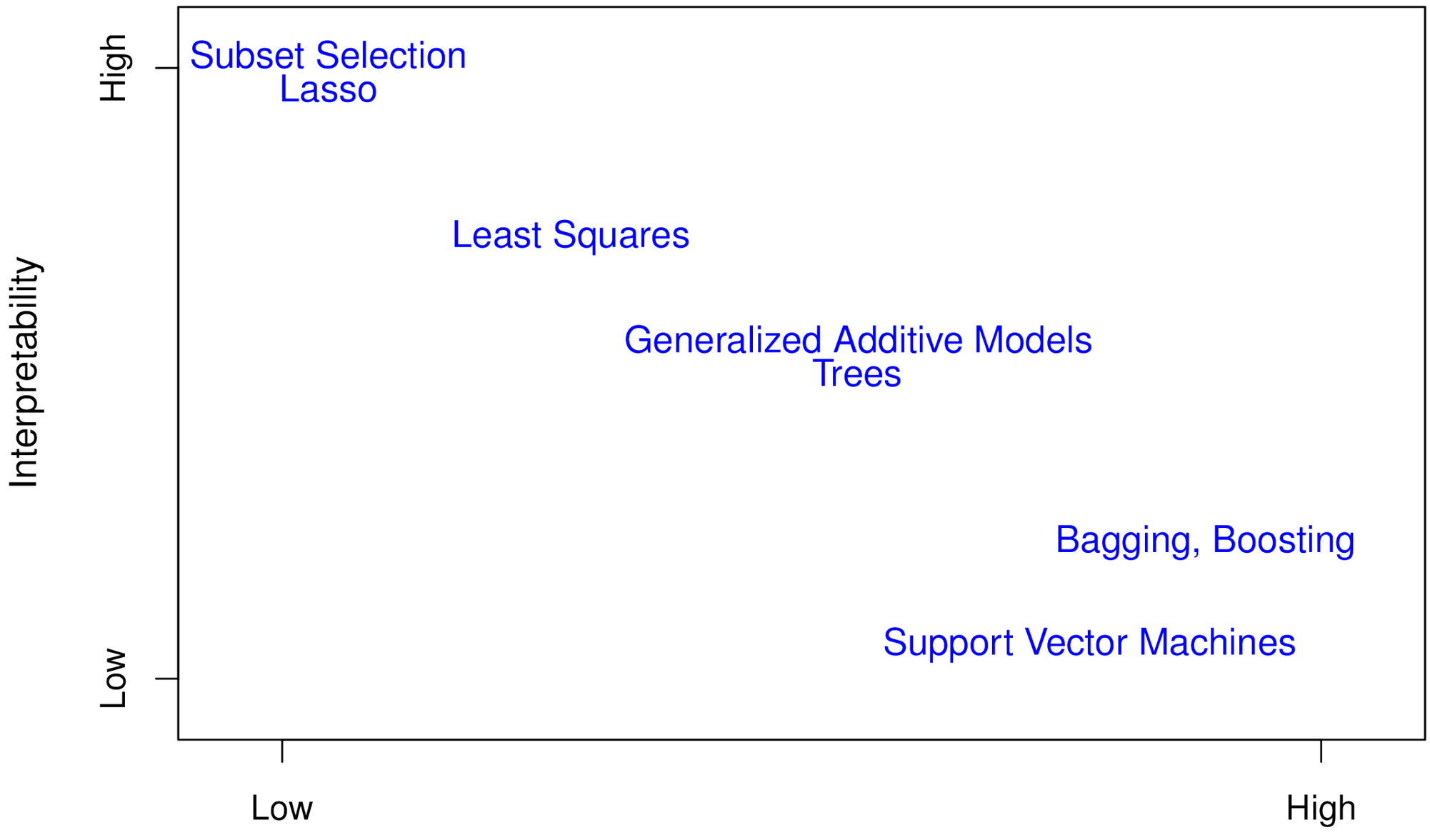




Model flexibility vs interpretability

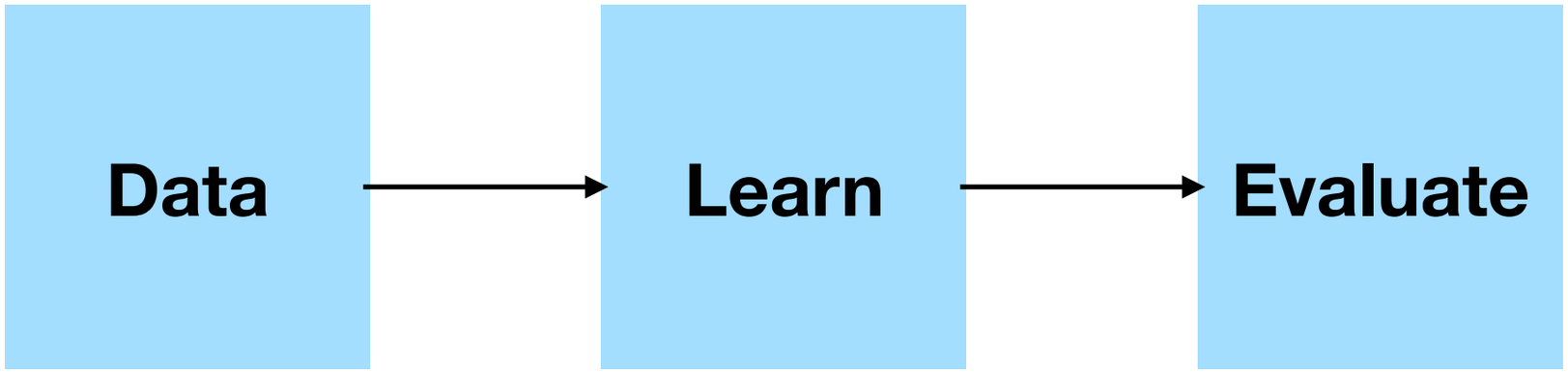
Q: Why should we ever choose a less flexible method over a flexible one?

A: Better interpretability (this slide), and sometimes even smaller error (next)!



Regression setting

$$Y = f(X_1, \ldots, X_p) + \varepsilon$$



Training set

$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$

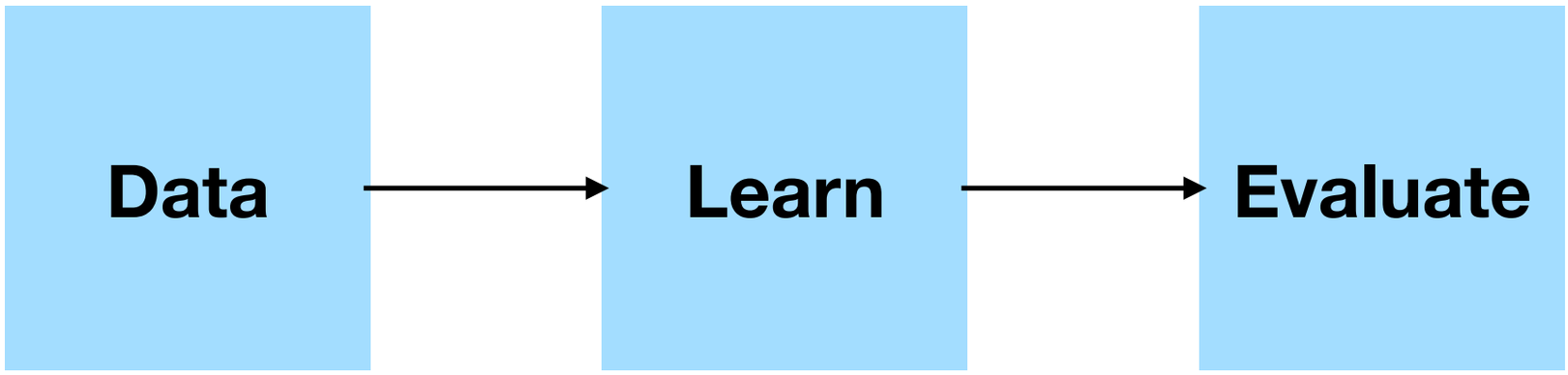
$\hat{f}$

how well does $\hat{f}$ learn?

No single method is the best choice for all datasets

Selecting a good statistical learning method depends on evaluating its performance

Training MSE



Training set

$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$

$\hat{f}$

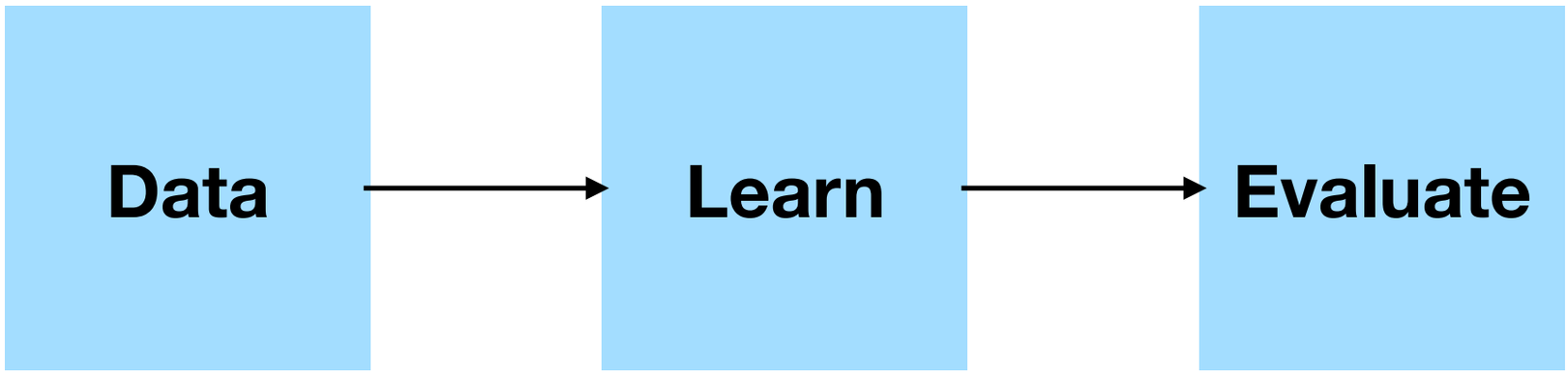
how well does $\hat{f}$ learn?

$$\text{Training MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(\mathbf{x}_i)\right)^2$$

Training Mean Squared Error

prediction that $\hat{f}$ gives for the ith training observation

Test MSE



Training set

$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$

$\hat{f}$

how well does $\hat{f}$ learn?

We are less interested in how $\hat{f}$ performs on training set

Consider test set $\{(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_m, \tilde{y}_m)\}$, not seen or used to train $\hat{f}$

$$\text{Test MSE} = \frac{1}{m} \sum_{i=1}^{m} \left(\tilde{y}_i - \hat{f}(\tilde{\mathbf{x}}_i) \right)^2$$

Test Mean Squared Error

prediction that $\hat{f}$ gives for the ith test data

Training MSE vs Test MSE

Training set

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

Test set

$$\{(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_m, y_m)\}$$

Training MSE

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

Test MSE

$$\frac{1}{m}\sum_{i=1}^{m}\left(\tilde{y}_i - \hat{f}(\tilde{\mathbf{x}}_i)\right)^2$$

$\hat{f}$ **is obtained on the training set!**

Training MSE vs Test MSE

Training set

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

Test set

$$\{(\tilde{\mathbf{x}}_1, \tilde{y}_1), (\tilde{\mathbf{x}}_2, \tilde{y}_2), \ldots, (\tilde{\mathbf{x}}_m, y_m)\}$$

Training MSE

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

Test MSE

$$\frac{1}{m}\sum_{i=1}^{m}\left(\tilde{y}_i - \hat{f}(\tilde{\mathbf{x}}_i)\right)^2$$

We want $\hat{f}$ for which the test MSE is as small as possible

If test set is available…great!

When test set is NOT available…

When test set is not available…

Q: Can we just select $\hat{f}$ that minimizes Training MSE

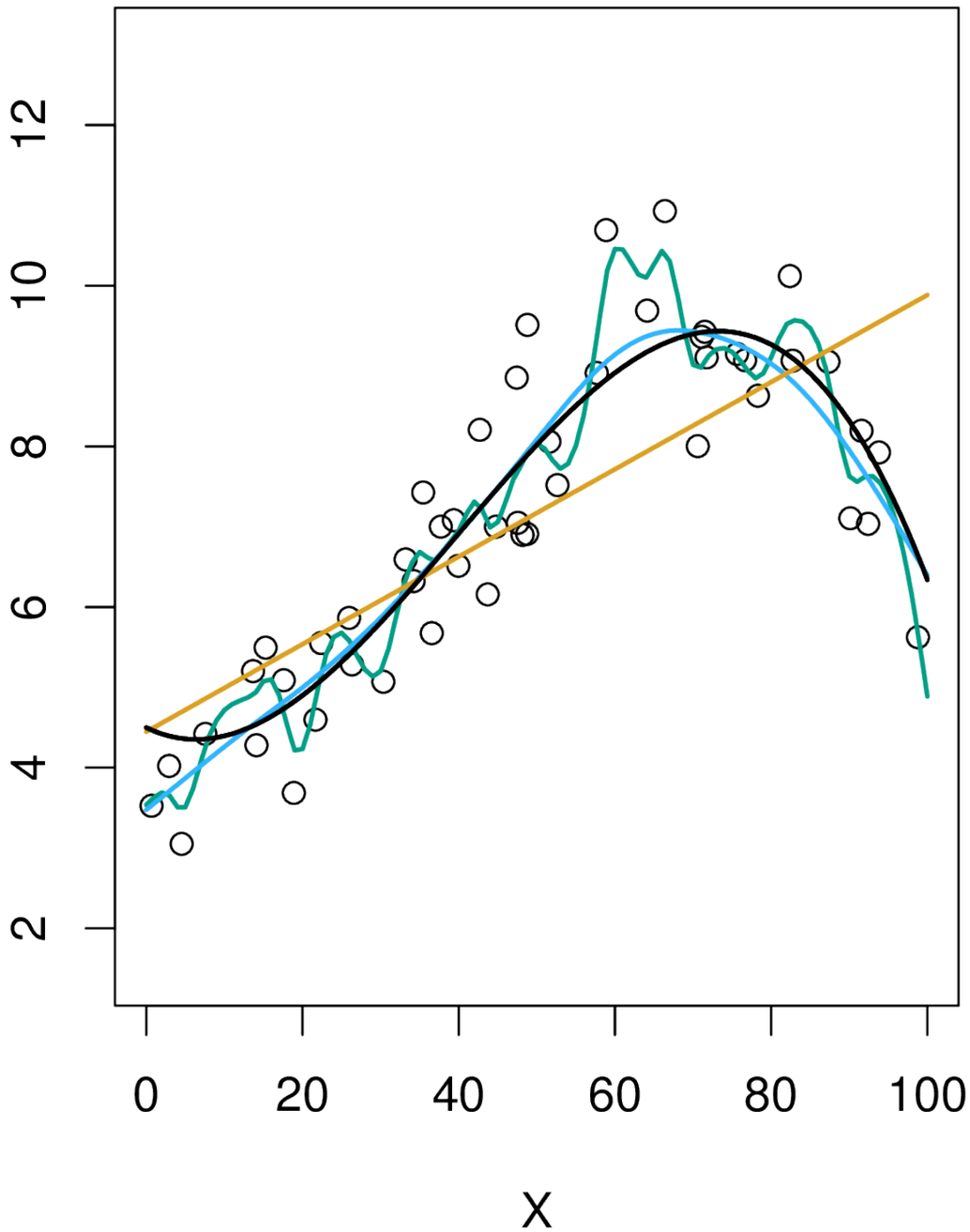
$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(\mathbf{x}_i)\right)^2$$

?

NO!

A method that has lowest training MSE does not necessarily imply that it also has the lowest test MSE

Lowest training MSE $\not\rightarrow$ lowest test MSE



Higher model flexibility → matches data points better

Higher model flexibility → more wiggly curve

Higher model flexibility $\not\rightarrow$ better fit to the true f

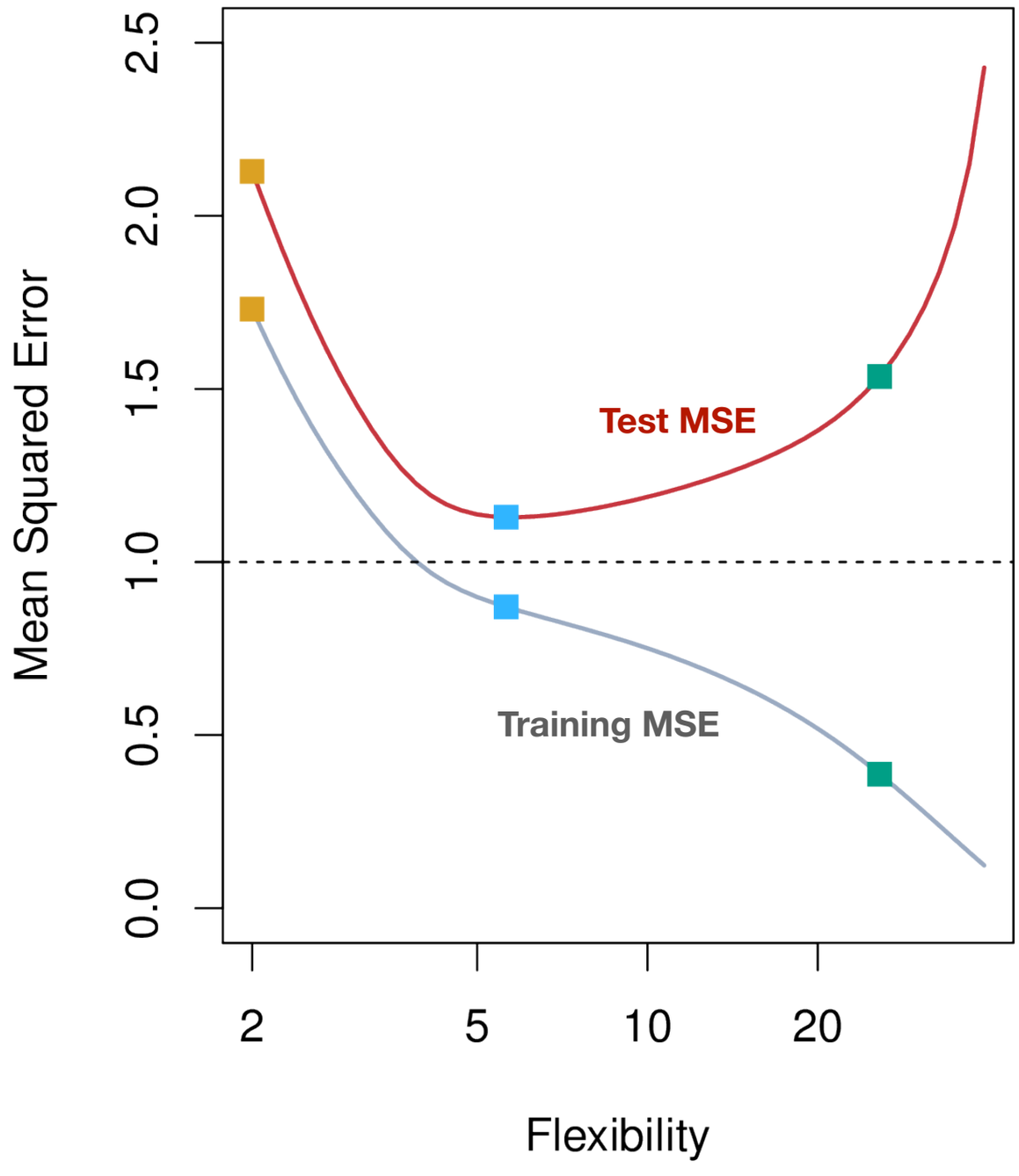
Linear regression **True f** **smoothing spline (simple)** **smoothing spline (complicated)**

level of model flexibility

Lowest training MSE $\not\rightarrow$ lowest test MSE

Higher model flexibility $\rightarrow$ matches data points better (≈lower training MSE)

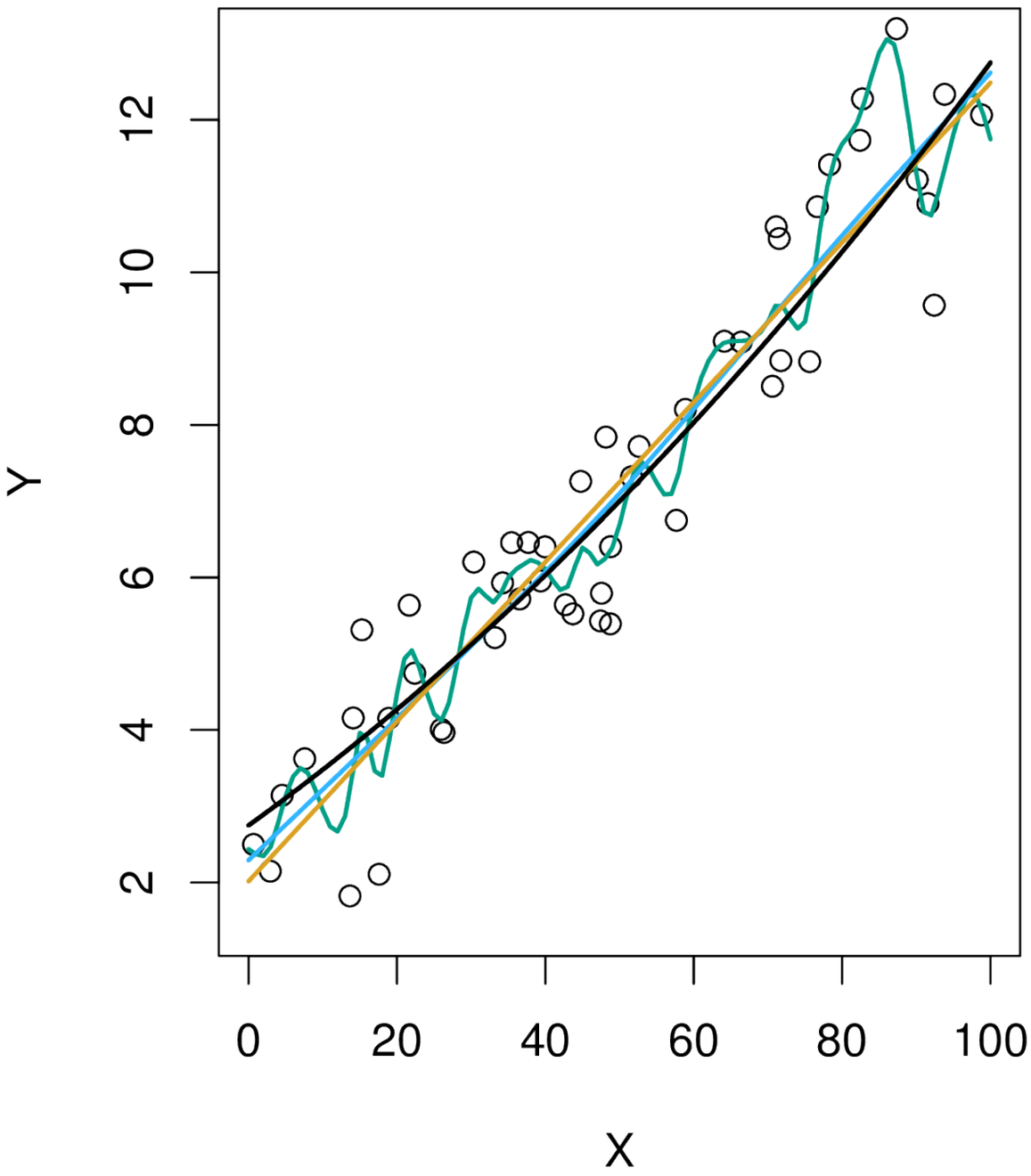
Higher model flexibility $\not\rightarrow$ better fit to the true f (≈lower test MSE)

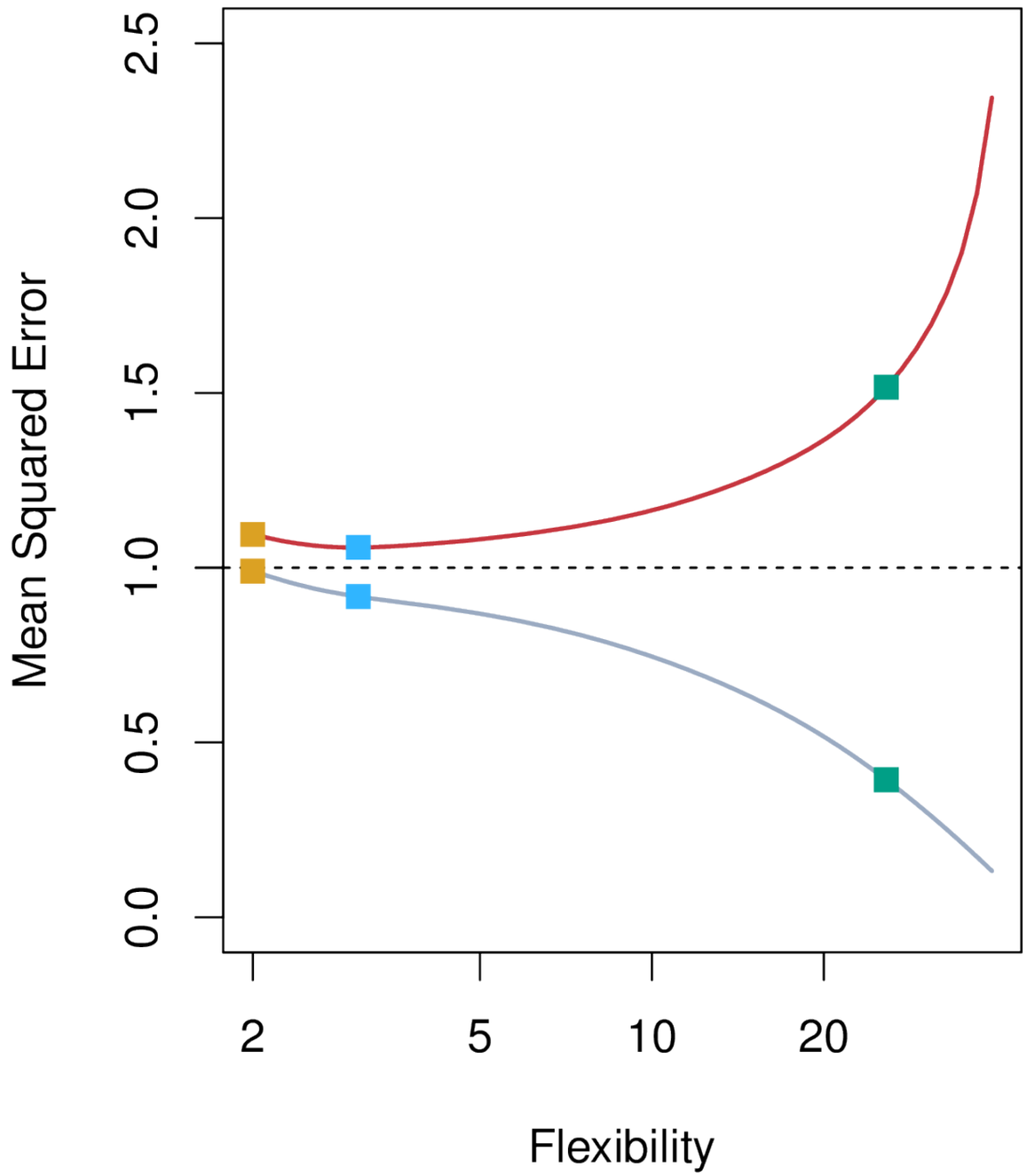


Linear regression True f smoothing spline (simple) smoothing spline (complicated)

level of model flexibility

Lowest training MSE ↛ lowest test MSE





Linear regression True f smoothing spline (simple) smoothing spline (complicated)

level of model flexibility

Overfitting = Large Test MSE + Small Training MSE

Learn too hard to find patterns in the training data

Some of the learned patterns in the training data are just caused by chance

Such patterns don’t necessarily exist in the test data!

One can always “memorize” the correct solutions in some practice questions

It does not necessarily guarantee a very high score in exam!

Some patterns learned in practice don’t necessarily exist in the exam

e.g. choose “C” in multiple choice problem when one doesn’t know the answer…

Estimating test MSE

Computing training MSE is easy… computing test MSE is NOT easy!

What should we do when test set is not available?

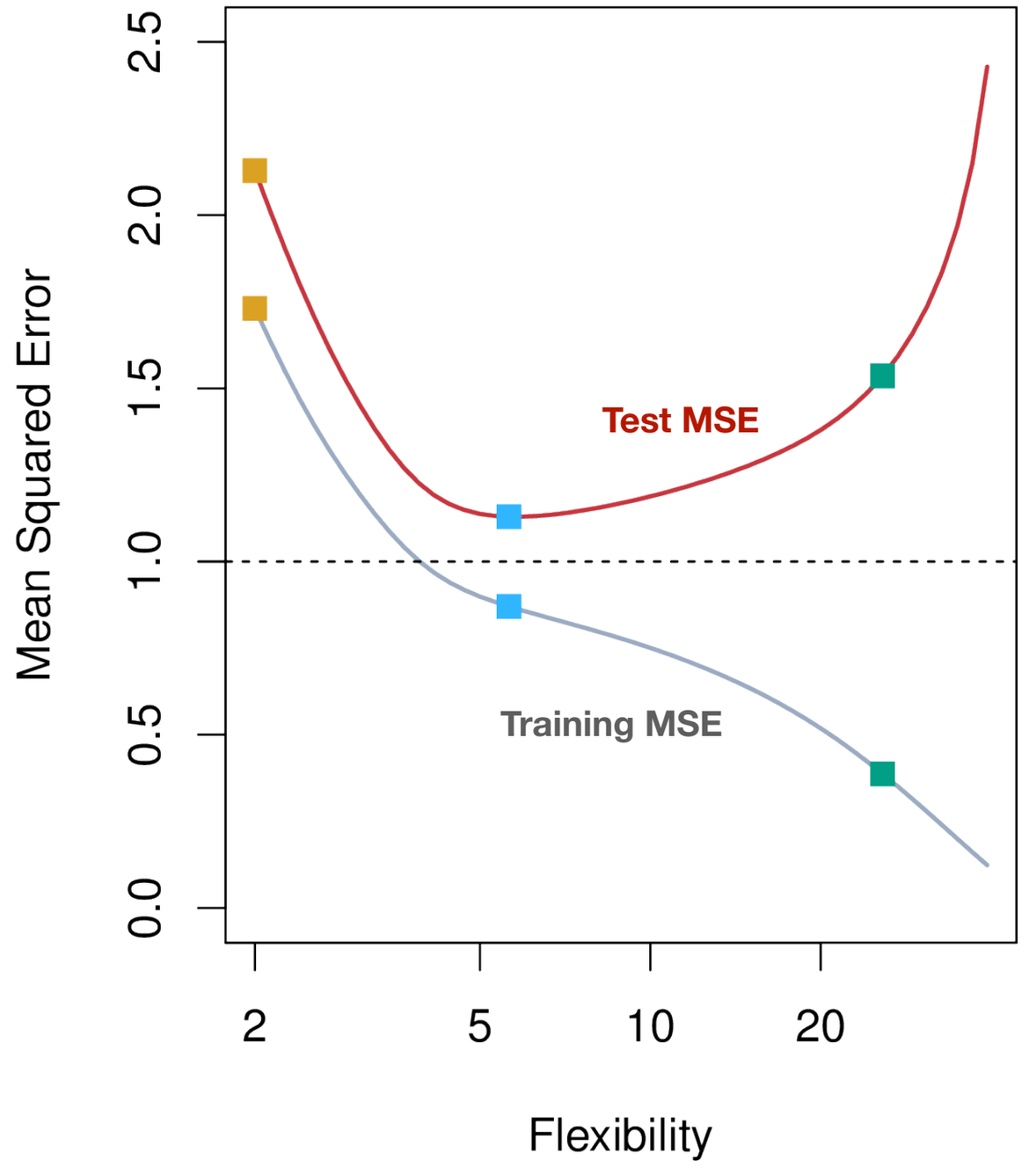
Cross-validation:

a method for estimating test MSE using only training data!

later in the course…

Bias-variance decomposition

Characterize the "U-shape" of Test MSE



Bias-variance decomposition

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

$\hat{f}$ is learned from a training set

$(\mathbf{x}_0, y_0)$ is a test observation, independent of the training set, from the model

$$Y = f(X) + \varepsilon$$

non-random **zero-mean noise**

Bias-variance decomposition

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

Expected test MSE

Irreducible error

$$Y = f(X) + \varepsilon$$

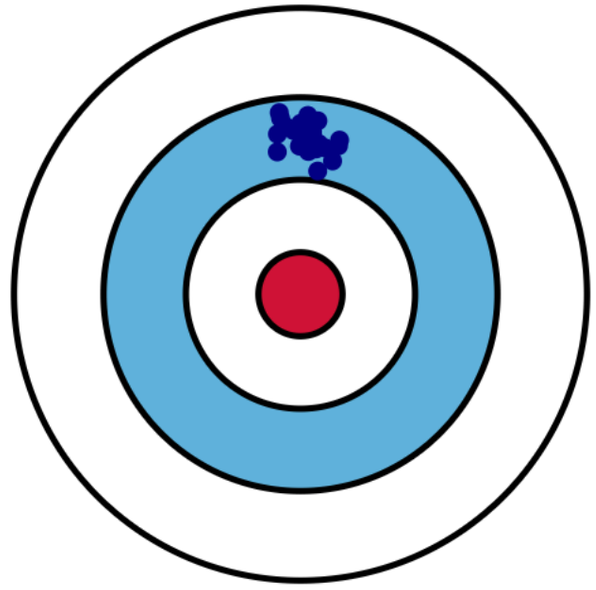
non-random **zero-mean noise**

**(Out of the scope of this course,
but interesting to think about:
What does the “expected” mean here?)**

Bias

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

$$\mathrm{Bias}(\hat{f}(\mathbf{x}_0)) = \mathrm{E}\left[\hat{f}(\mathbf{x}_0)\right] - f(\mathbf{x}_0)$$



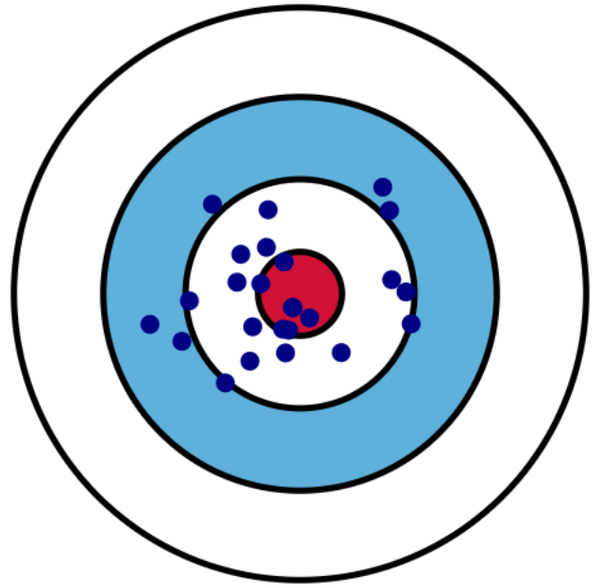
Error that is introduced by approximating a real-life problem

e.g., the real relationship between response and predictors is nonlinear, but we fit a linear model, which causes bias

Variance

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

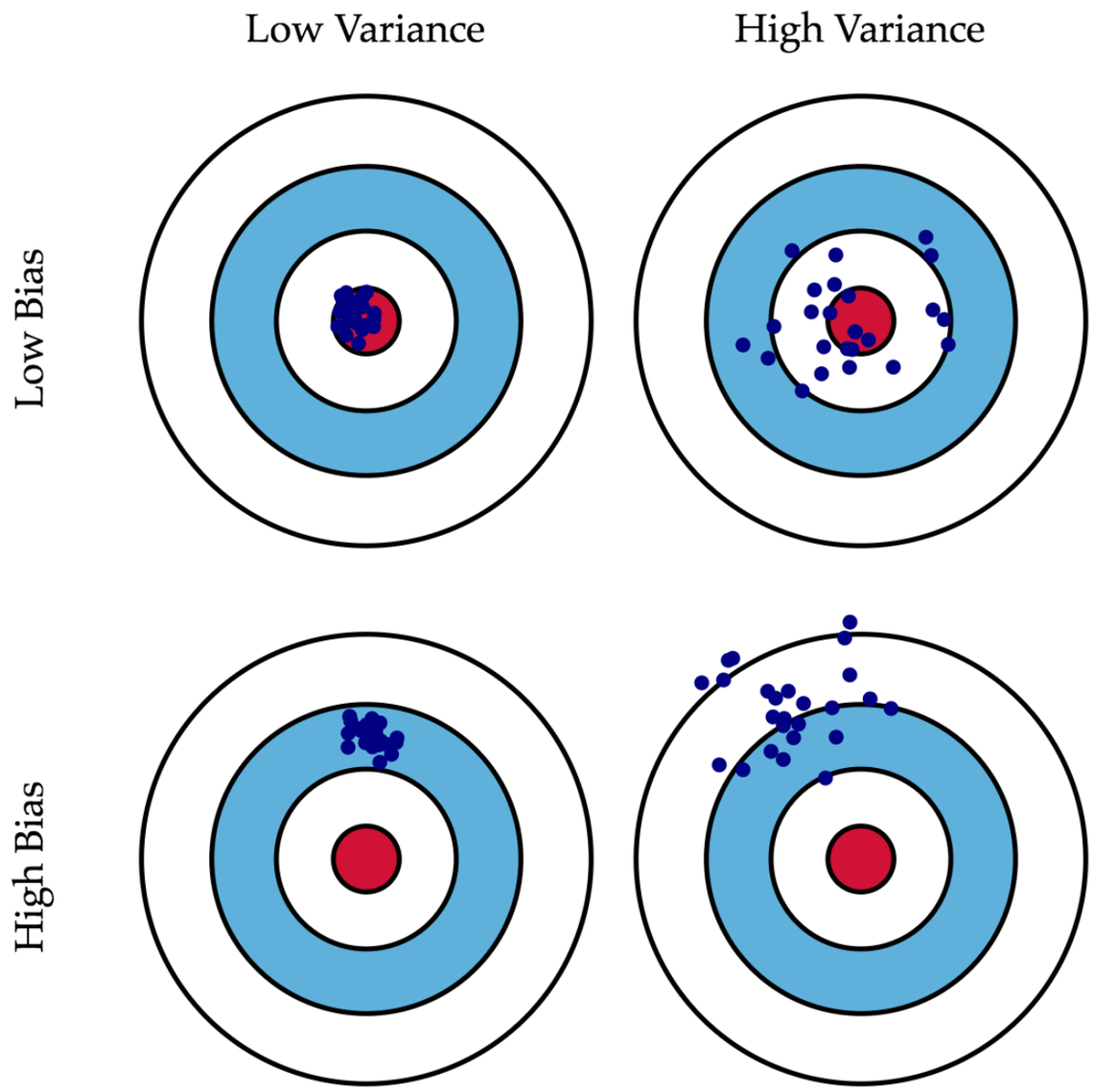
$$\mathrm{Var}(\hat{f}(\mathbf{x}_0)) = \mathrm{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathrm{E}\hat{f}(\mathbf{x}_0)\right)^2\right]$$



the amount by which $\hat{f}$ **change** if we estimated it using a different training set

Bias-variance tradeoff

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

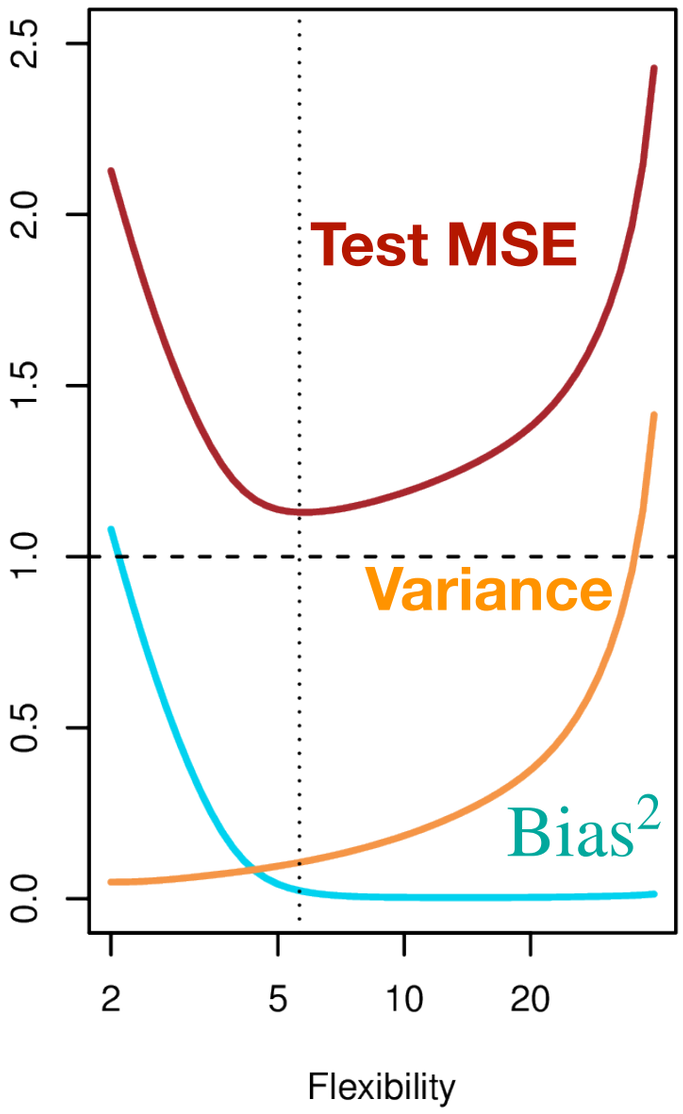


Want to **simultaneously** achieve **low variance** and **low bias**

Bias-variance tradeoff

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

Want to **simultaneously** achieve **low variance** and **low bias**



A simple model → high bias + low variance

A flexible model → low bias + high variance

Challenge: find a method for which both the variance and (squared) bias are low

Bias-variance decomposition

$$Y = f(X) + \varepsilon$$

non-random zero-mean noise

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

Expected test MSE = **Variance** + **Bias^2** + **Irreducible error**

$$= \mathrm{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathrm{E}\hat{f}(\mathbf{x}_0)\right)^2\right] + \left[\mathrm{E}\left[\hat{f}(\mathbf{x}_0)\right] - f(\mathbf{x}_0)\right]^2 + \mathrm{Var}(\varepsilon)$$

If we take

$$\hat{f}(\mathbf{x}_0) = \mathrm{E}\left[Y \,|\, X = \mathbf{x}_0\right]$$

then

$$\mathrm{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathrm{E}\hat{f}(\mathbf{x}_0)\right)^2\right] = 0 \quad \textbf{and} \quad \left[\mathrm{E}\left[\hat{f}(\mathbf{x}_0)\right] - f(\mathbf{x}_0)\right]^2 = 0$$

Bias-variance decomposition

$$Y = f(X) + \varepsilon$$

non-random zero-mean noise

If we take

$$\hat{f}(\mathbf{x}_0) = \mathrm{E}\left[Y \,|\, X = \mathbf{x}_0\right]$$

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

Expected test MSE = **Variance** + **Bias^2** + **Irreducible error**

$$= \underbrace{\mathrm{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathrm{E}\hat{f}(\mathbf{x}_0)\right)^2\right] + \left[\mathrm{E}\left[\hat{f}(\mathbf{x}_0)\right] - f(\mathbf{x}_0)\right]^2}_{\text{minimized!}} + \mathrm{Var}(\varepsilon)$$

$$= \mathrm{Var}(\varepsilon)$$

Irreducible error

Bias-variance decomposition

$$Y = f(X) + \varepsilon$$

non-random zero-mean noise

The "best" we can do

$$\hat{f}(\mathbf{x}_0) = \mathrm{E}\left[Y \,|\, X = \mathbf{x}_0\right]$$

Unknown in practice!!

Because the joint distribution of (X, Y) is unknown in practice

In summary

Focus on regression setting

Method flexibility vs interpretability

Training error vs test error

Bias-variance tradeoff