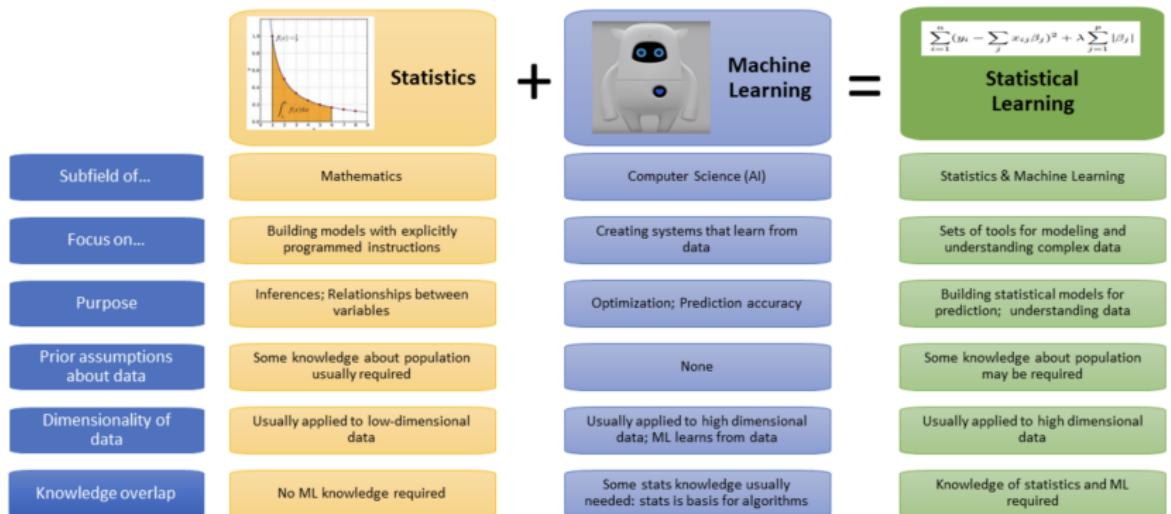


What is Statistical Learning?

“Statistical learning refers to a vast set of tools for understanding data.”

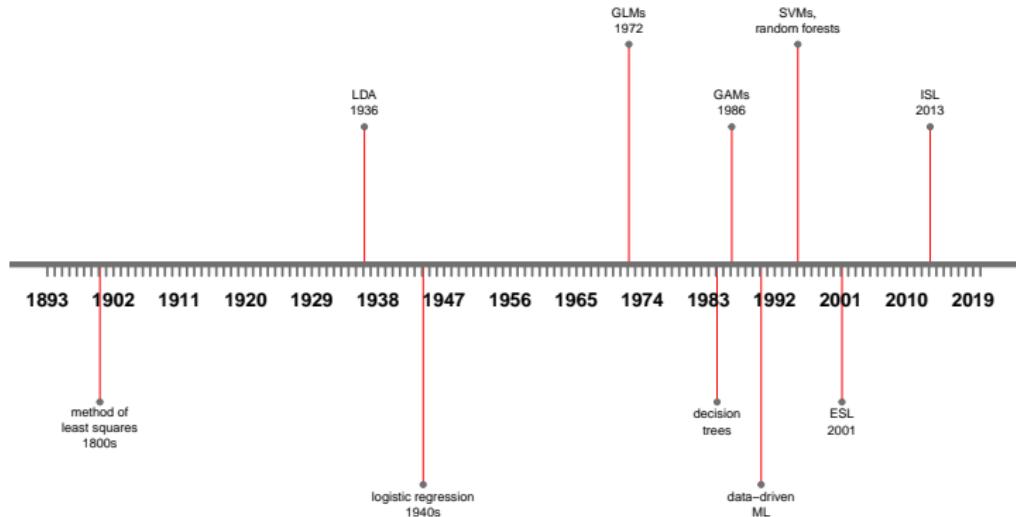
— *ISL*

Statistical vs. Machine Learning?



Musio image: Akawikipic [CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>)]

Timeline of Statistical Learning



Elements of Statistical Learning (2001)

- ▶ One of the first reference textbooks on fundamentals of statistical machine learning
- ▶ More comprehensive than ISL in terms of number of approaches and depth
- ▶ More technical details & in-depth mathematical treatment

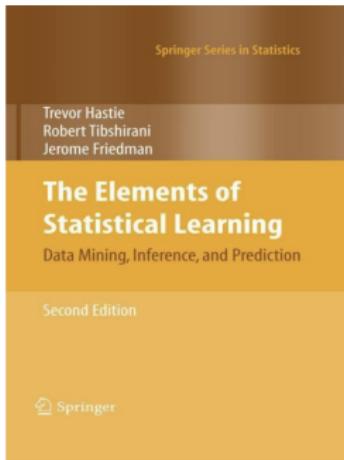


Figure 1: Elements of Statistical Learning

Introduction to Statistical Learning (2013)

- ▶ More of a hands-on introduction to computational aspects of statistical learning with real-world data.
- ▶ Uses R as 'the language of choice for academic statisticians' (also a bridge across disciplines!)

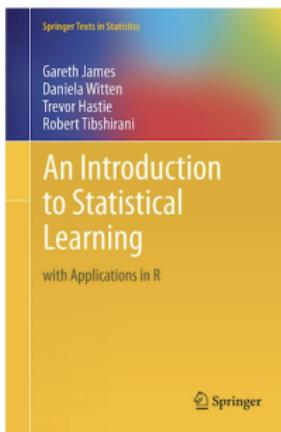


Figure 2: Introduction to Statistical Learning

Statistical Learning Approaches

Supervised methods

- ▶ Regression (observe input & quantitative output)
- ▶ Classification (observe input & categorical output)

Unsupervised methods

- ▶ Data are not labeled (observe only input)
- ▶ Goal may be learning relationships among variables and data structure (e.g. how many population clusters are there?)

Supervised Learning: Regression or Classification?

- ▶ Predicting whether or not a patient has cancer based on expression patterns of 5,000 genes?
- ▶ Predicting quantity of a certain protein, given expression patterns of 5,000 genes?
- ▶ Estimating plant height according to variation at 5,000 positions in the genome?
- ▶ Determine whether pixels in a satellite image indicate an agricultural field, river, or forest?
- ▶ Determine how many different types of bacterial communities have been sampled using microbiome sequence data from 10 tissues?

Another definition...

Consider a quantitative response, Y , and p different predictors, X_1, X_2, \dots, X_p .

We assume some relationship between Y and $X = (X_1, X_2, \dots, X_p)$, which can be written in the general form:

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

where f is a fixed but unknown function of X_1, X_2, \dots, X_p and ϵ is a random error term which is independent of X and has mean zero.

What is Statistical Learning?

“Statistical learning is a set of tools for estimating $f(X)$. ”
— ISL

How should we choose what tool to use to estimate $f(X)$?

Why estimate f ?

1. Prediction

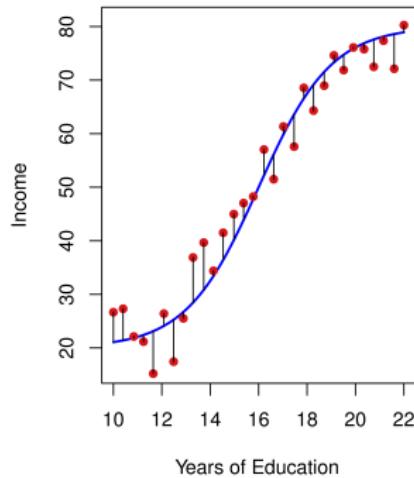
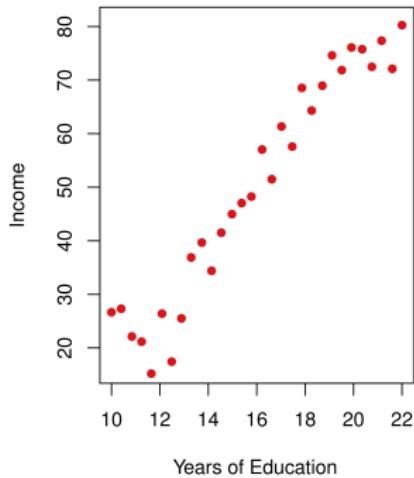


Figure 3: ISL Fig. 2.2

Linear model fit to Income data

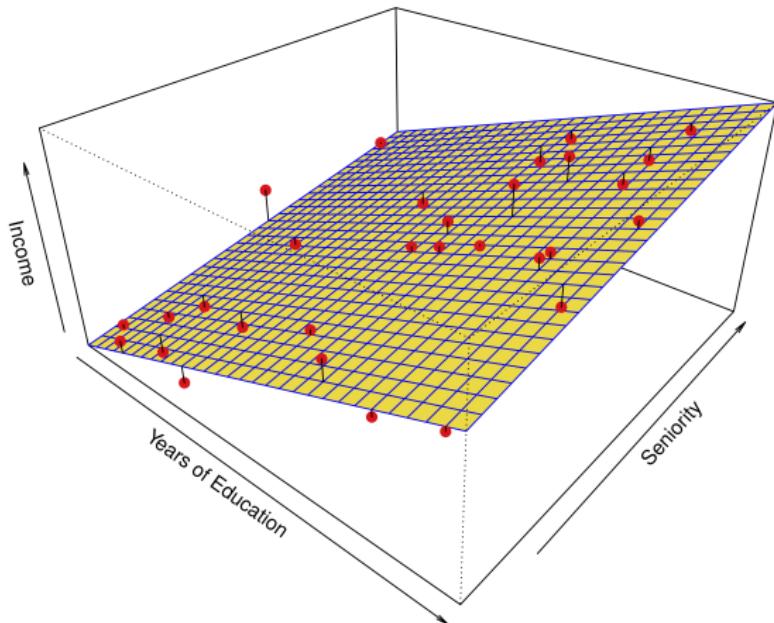


Figure 4: Fig 2.4

linear model fit by least squares

Non-parametric model fit to data

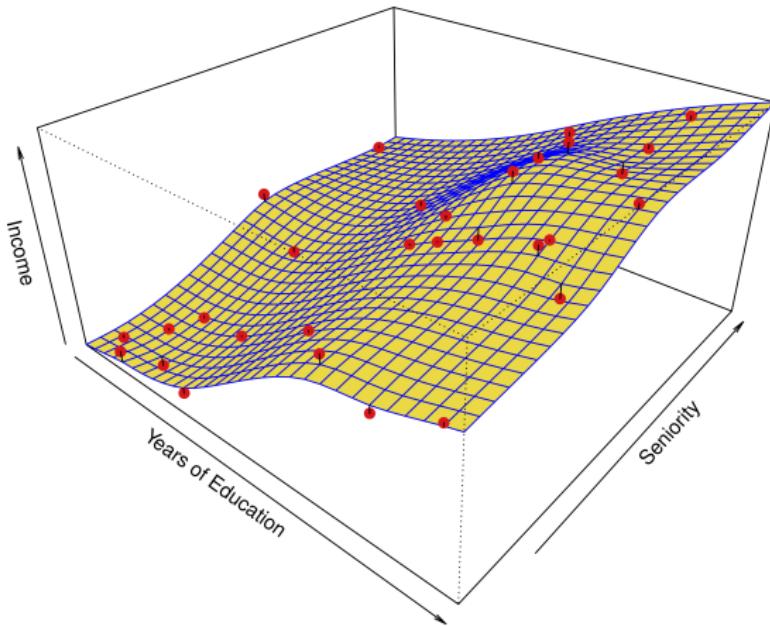


Figure 5: Fig 2.5

fit w/thin-plate spline

Considering our prediction of Y :

$$\hat{Y} = \hat{f}(X)$$

The expected value of the squared error is then:

$$E(Y - \hat{Y})^2 = [f(X) - \hat{f}(X)]^2 + \text{Var}(\epsilon)$$

We can choose different modeling techniques to improve our estimate of f , and minimize the reducible error.

Why estimate f ?

2. Inference

- ▶ Which predictors are associated with the response?
- ▶ What is the relationship between the response and each predictor?
- ▶ Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is it more complicated?

Trade-offs: Flexibility vs. interpretability

- ▶ Whether we are interested in inference vs. prediction influences how we choose to estimate f .
- ▶ It determines whether we might choose a **parametric** vs. **non-parametric** approach.



Figure 6: Fig. 2.7

Parametric methods

- ▶ Assume a specific form for f .
 - ▶ e.g. the linear model is an example of a parametric model.

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

- ▶ Fit/train the model
 - ▶ e.g. using ordinary least squares (Chp. 3) or other methods (Chp. 6).
- ▶ Instead of having to estimate an arbitrary p -dimensional function, we only need to estimate $p + 1$ coefficients $\beta_0, \beta_1, \dots, \beta_p$.

Non-parametric methods

- ▶ Do not make explicit assumptions about the form of f .
- ▶ More flexible, can provide a better fit to f .

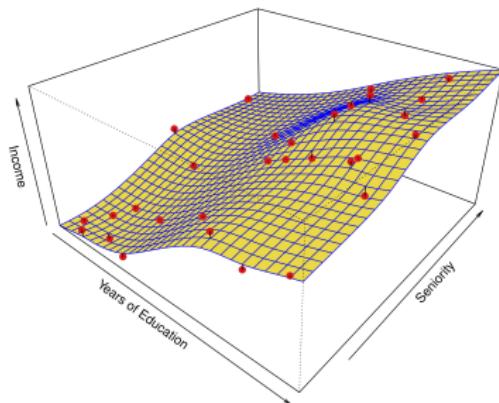


Figure 7: Fig 2.5

fit w/thin-plate spline

Non-parametric methods

- ▶ Typically need larger sample size compared to parametric approach to get accurate estimate of f .
- ▶ Can suffer from overfitting (model learns ‘noise’ in the training data & performs well on training data but poorly on test data)
- ▶ More difficult to interpret

Parametric vs. non-parametric methods

Linear methods (Part I)

Chp. 3: Linear methods for regression

Chp. 4: Linear methods for classification

Chp. 5: Resampling

Chp. 6*: Alternatives to fitting with least squares

Non-linear methods (Part II)

Chp. 7: Non-linear methods w/single & multiple input variables

Chp. 8: Tree-based methods

Chp. 12: Unsupervised methods

Chp. 9: Support vector machines

Ch. 10*: Deep Learning

*Lectures optional for MBS students, see schedule

Measuring the Quality of Fit

- ▶ In regression, one commonly used measure is the mean squared error (MSE).
- ▶ The MSE is a measure of how close the predicted response is to the true response for an observation.

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

- ▶ We want to choose the method that gives the lowest MSE on a *testing* dataset, not the lowest MSE on a *training* dataset

Measuring the Quality of Fit

- ▶ No guarantee that the model with the lowest training MSE will also have the lowest test MSE.
- ▶ Note characteristic ‘U-shape’ of test MSE (red).

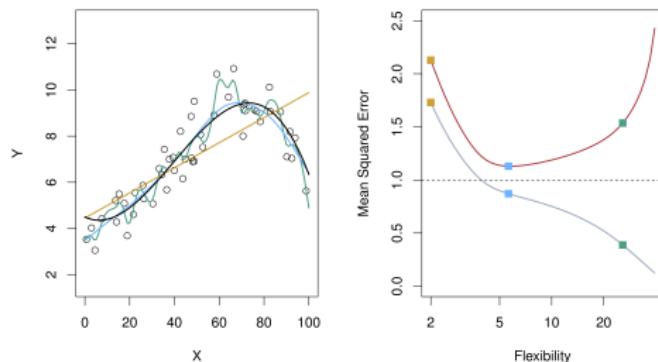


Figure 8: Fig 2.9

Left: Three estimates of f based on simulated data (open circles); linear regression (orange line), and two different smoothing splines (blue and green curves) Right: Training MSE (grey curve) and testing MSE (red curve)

Measuring the Quality of Fit

- ▶ Depending on the true shape of f , more interpretable methods like linear regression may provide a very good fit to the data.

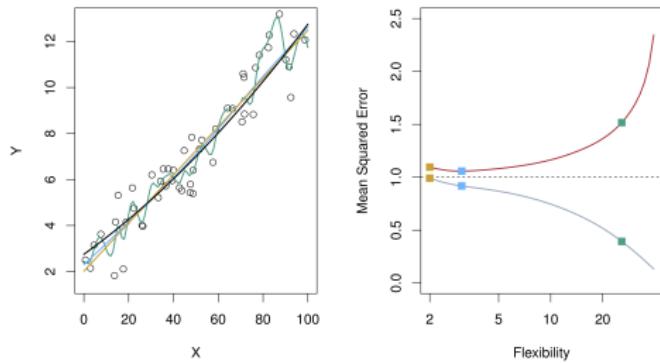


Figure 9: Fig 2.10

ISL Fig. 2.10

How do we minimize the expected test MSE?

Imagine repeatedly estimating f using \hat{f} using a large number of training sets, and testing each of the \hat{f} at a given value, x_0 .

The expected test MSE at x_0 is given by:

$$\mathbb{E}(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon)$$

Average across all x_0 in test set to get overall expected test MSE.

Trade-off: Bias vs. Variance

To minimize the average test error, we need to select a statistical learning method that simultaneously has low variance and low bias.

Trade-off: Bias vs. Variance

Variance is the amount by which \hat{f} changes from one training set to another. Ideally, \hat{f} does not change a lot between different training sets.

$$\text{Var}(\hat{f}(x)) = \mathbb{E}[(\hat{f}(x) - \mu)^2]$$

$$\text{where } \mu = \mathbb{E}[\hat{f}(x)]$$

More flexible methods generally have higher variance.

Trade-off: Bias vs. Variance

Bias refers to the error introduced by approximating a complicated problem by a much simpler model.

$$\text{Bias}(\hat{f}(x)) = \mathbb{E}(\hat{f}(x)) - f(x)$$

More flexible methods generally have lower bias.

Trade-off: Bias vs. Variance

The 'best' method (low variance AND low bias) depends on the true form of f .

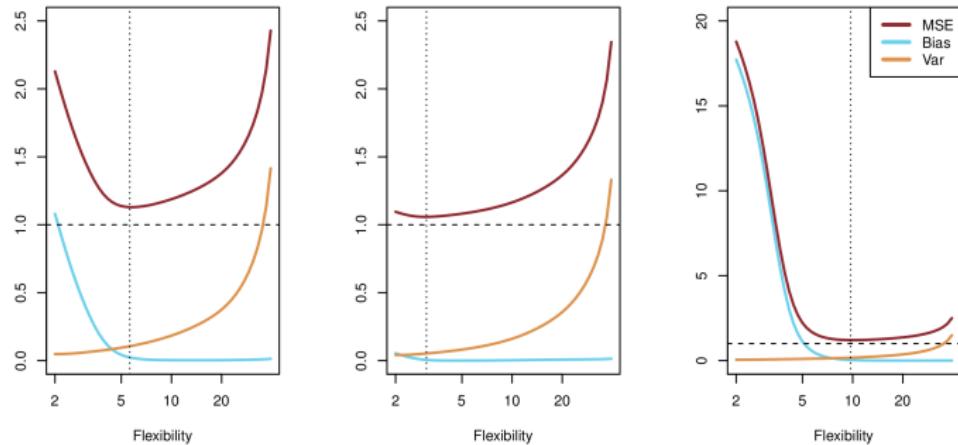


Figure 10: Fig 2.12

Model Accuracy in Classification

- ▶ The test error rate is minimized on average by a simple classifier (the Bayes classifier) that *assigns each observation to the most likely class, given its predictor values.*
- ▶ Test error is given by

$$\text{Avg}(I(y_0 \neq \hat{y}_0))$$

where $I(y_0 \neq \hat{y}_0) = 1$ if $y_0 \neq \hat{y}_0$ but 0 otherwise.

Model Accuracy in Classification

With 2 classes, Bayes decision boundary corresponds to predicting class one if $Pr(Y = 1|X = x_0) > 0.5$ and class two otherwise

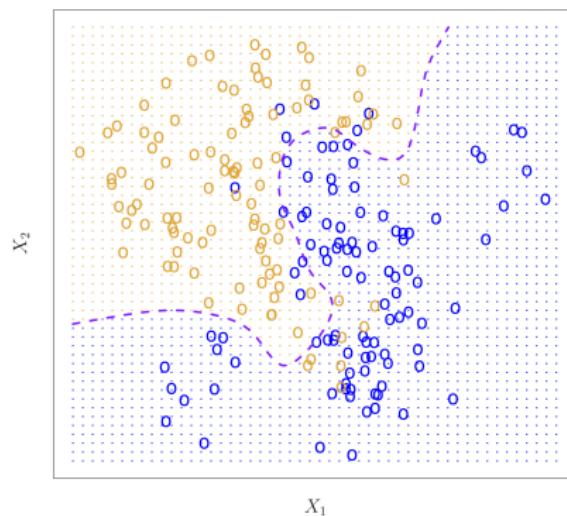


Figure 11: Fig 2.13

ISL Fig. 2.13; Purple dashed line: Bayes decision boundary

Non-parametric classification: KNN

In reality, we never know the conditional probability of Y given X ! The Bayes Classifier boundary is therefore an unattainable ‘gold standard’. But, we can estimate it using various methods (logistic regression, KNN classification, LDA, QDA).

K-Nearest Neighbors classification

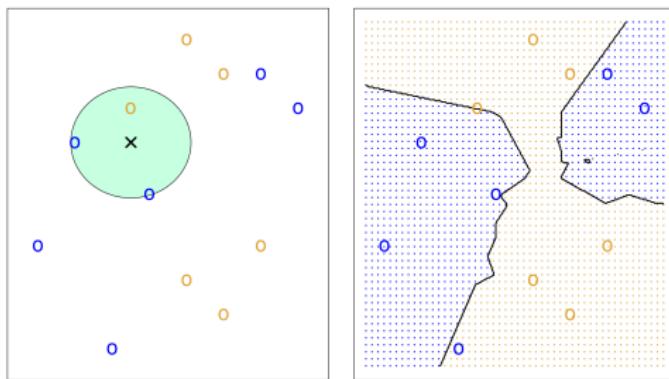


Figure 12: Fig 2.14

ISL Fig. 2.14. KNN decision boundary for $K=3$

Non-parametric classification: KNN

Choosing appropriate level of flexibility

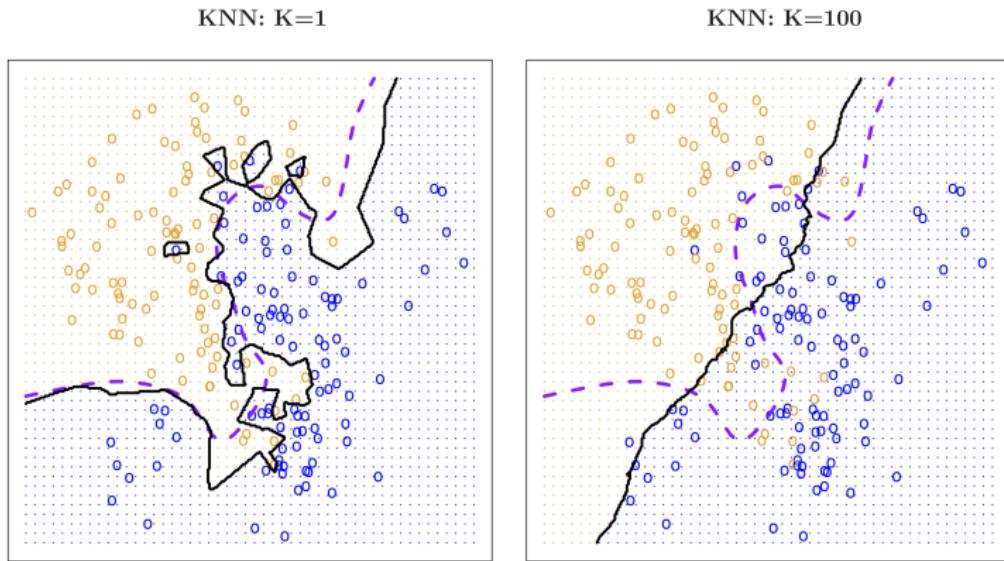


Figure 13: Fig 2.16

Non-parametric classification: KNN

Bias-variance trade-off

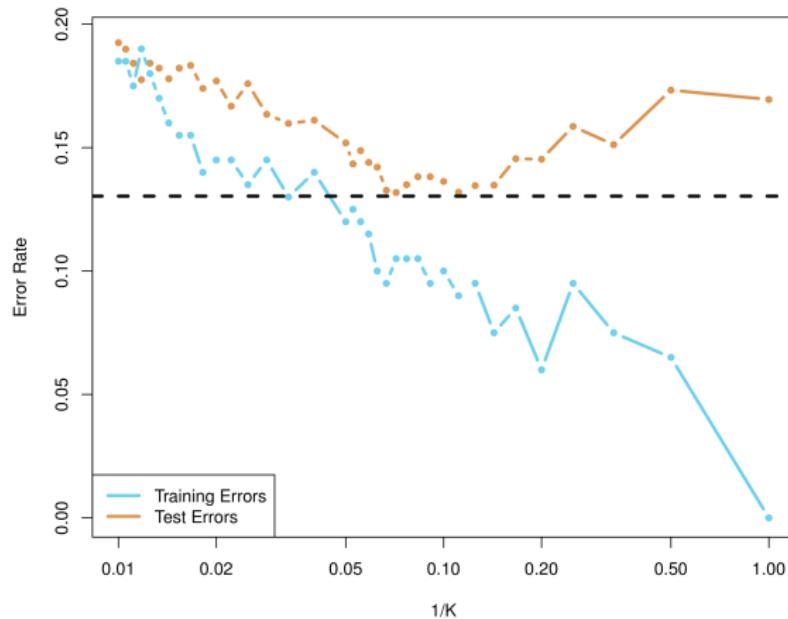


Figure 14: Fig 2.17

In summary:

- ▶ Statistical learning refers to a vast set of tools for understanding data, and estimating $f(X)$.
- ▶ How we choose $f(x)$ is shaped by a trade-offs.
- ▶ Whether we are more interested in accuracy vs. interpretability determines whether we might choose a more flexible model and whether the model is characterized by relatively higher bias vs. variance.