



Machine Learning

Statistical Learning

- What Is Statistical Learning?

- Why estimate f ?
- How do we estimate f ?
- The trade-off between prediction accuracy & model interpretability



- Some important taxonomies (I expect you'll know this by heart!)

- Prediction vs. Inference
- Parametric vs. Non Parametric models
- Regression vs. Classification problems
- Supervised vs. Unsupervised learning
- ...

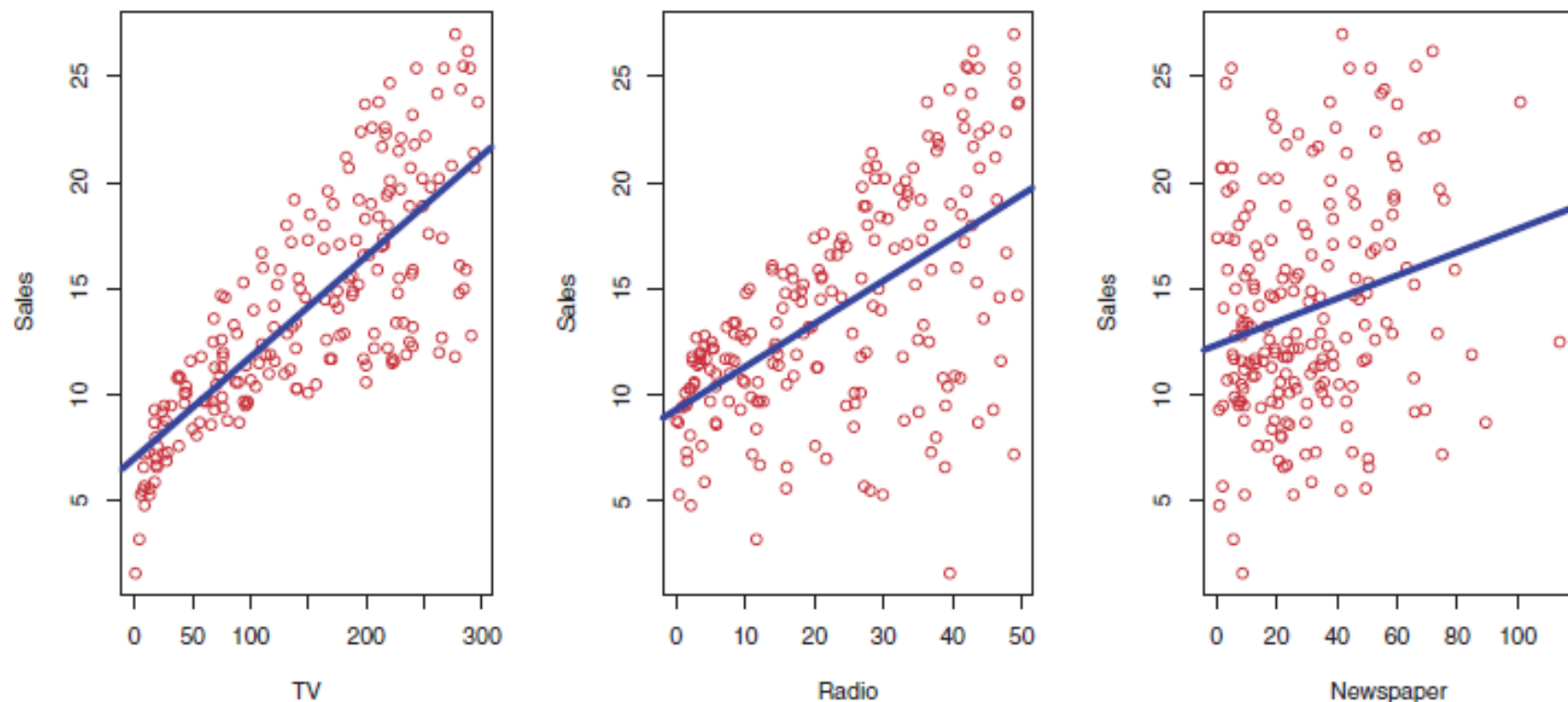


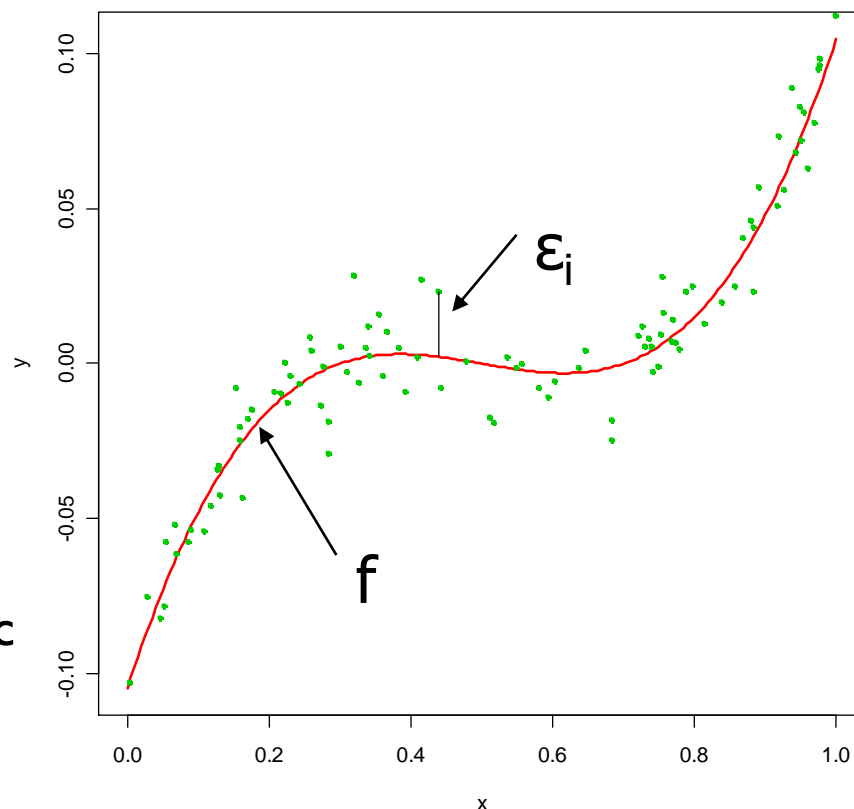
FIGURE 2.1. The **Advertising** data set. The plot displays **sales**, in thousands of units, as a function of **TV**, **radio**, and **newspaper** budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of **sales** to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict **sales** using **TV**, **radio**, and **newspaper**, respectively.

- Suppose we observe Y_i and $X_i = (X_{i1}, \dots, X_{ip})$ for $i = 1, \dots, n$

- Assume a relationship exists between Y and at least one of the observed X 's
- Assume we can model this relationship as

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

- f : unknown function systematic
- ε_i : zero mean random error

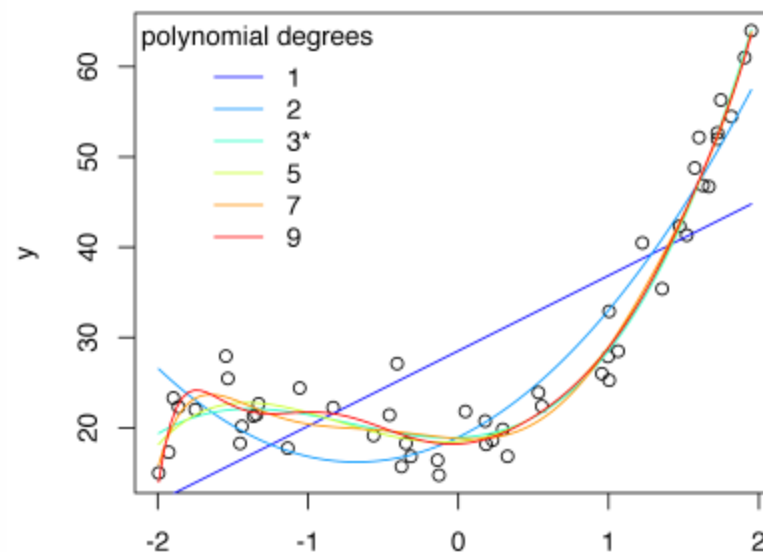


- The term statistical learning refers to using the data to “learn” f

- The error our estimate will have has two components

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

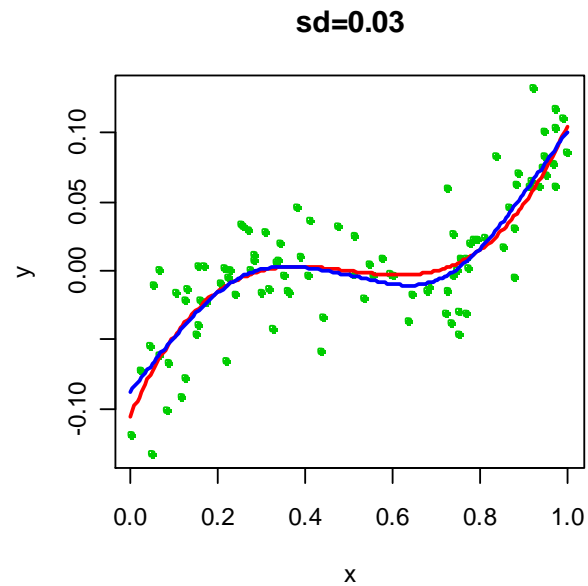
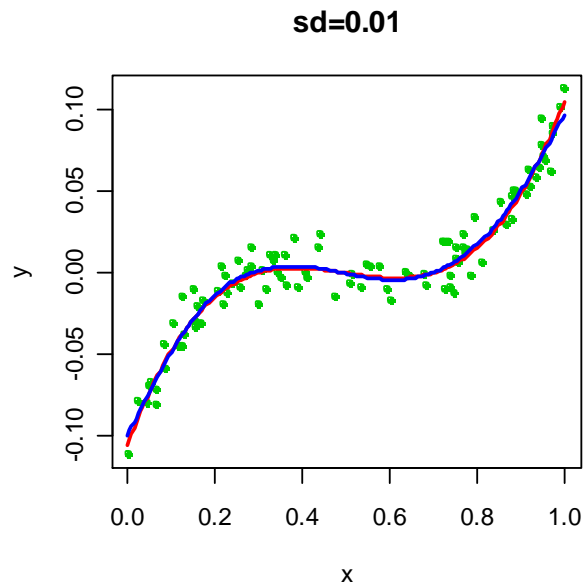
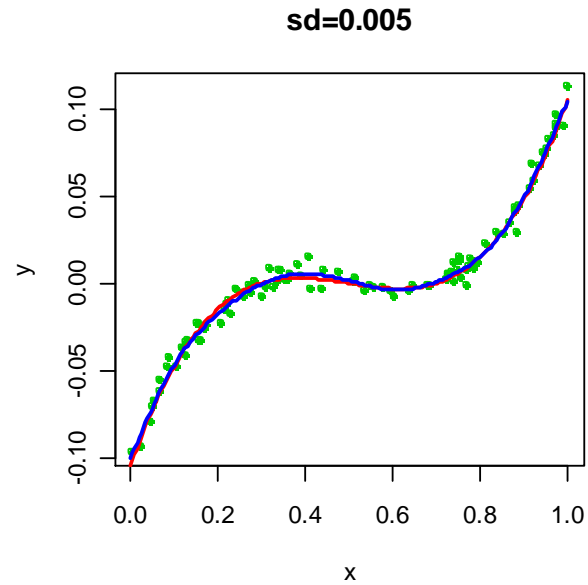
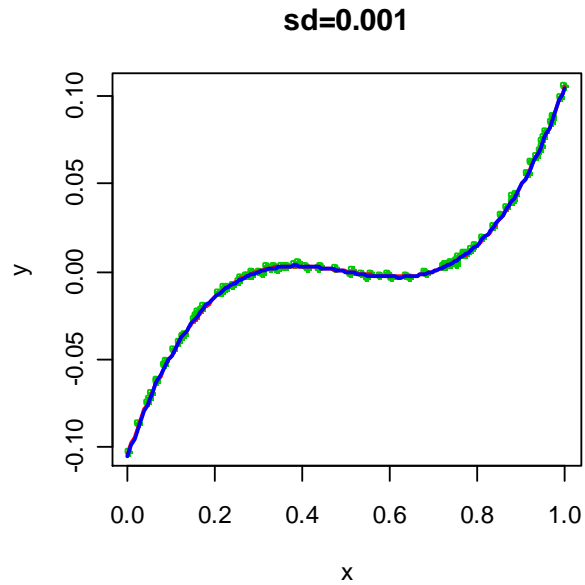
- *Reducible error due to the choice of f (model complexity)*



- *Irreducible error due to the presence of ε_i in the training set*

Because noise matters ...

6



- The error our estimate will have has two components

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

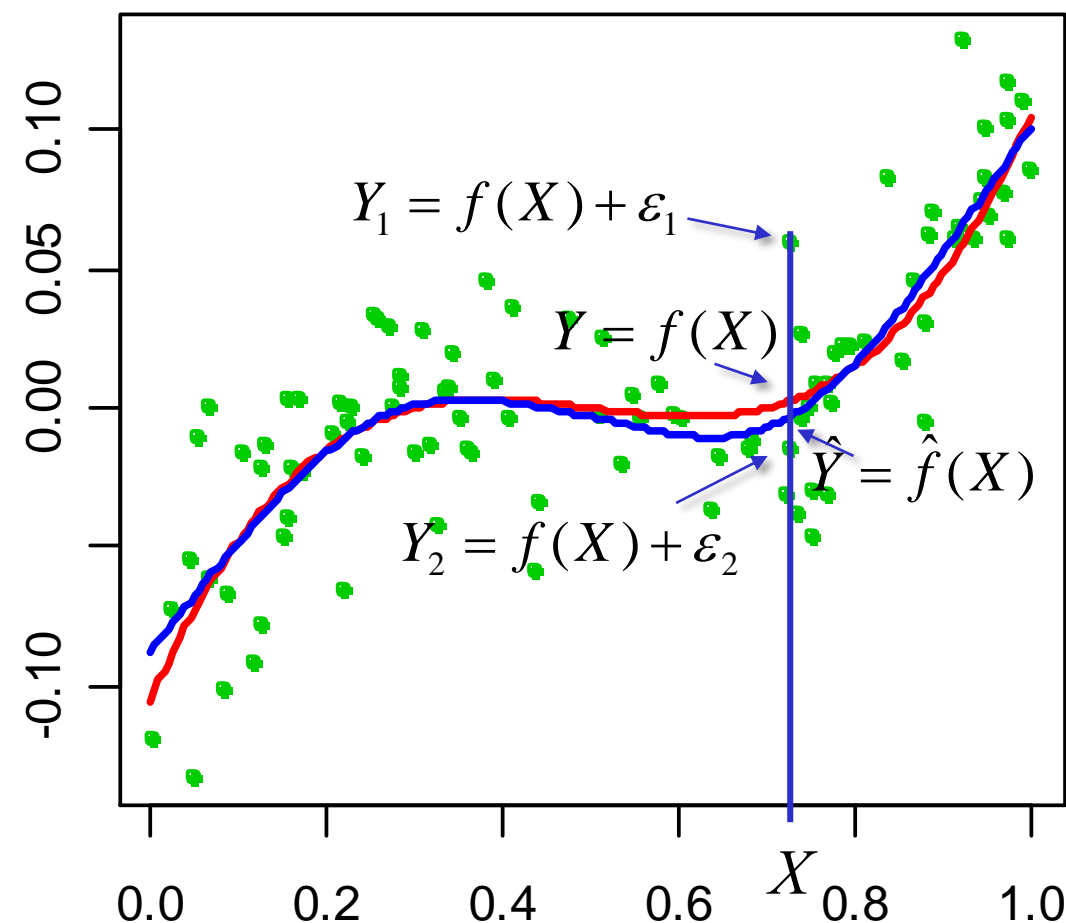
- *Reducible error* due to the choice of f (*model complexity*)
- *Irreducible error* due to the presence of ε_i in the training set

- Let assume \hat{f} and \mathbf{X} fixed for the time being

$$\begin{aligned}\hat{Y} &= \hat{f}(X) \\ E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}\end{aligned}$$

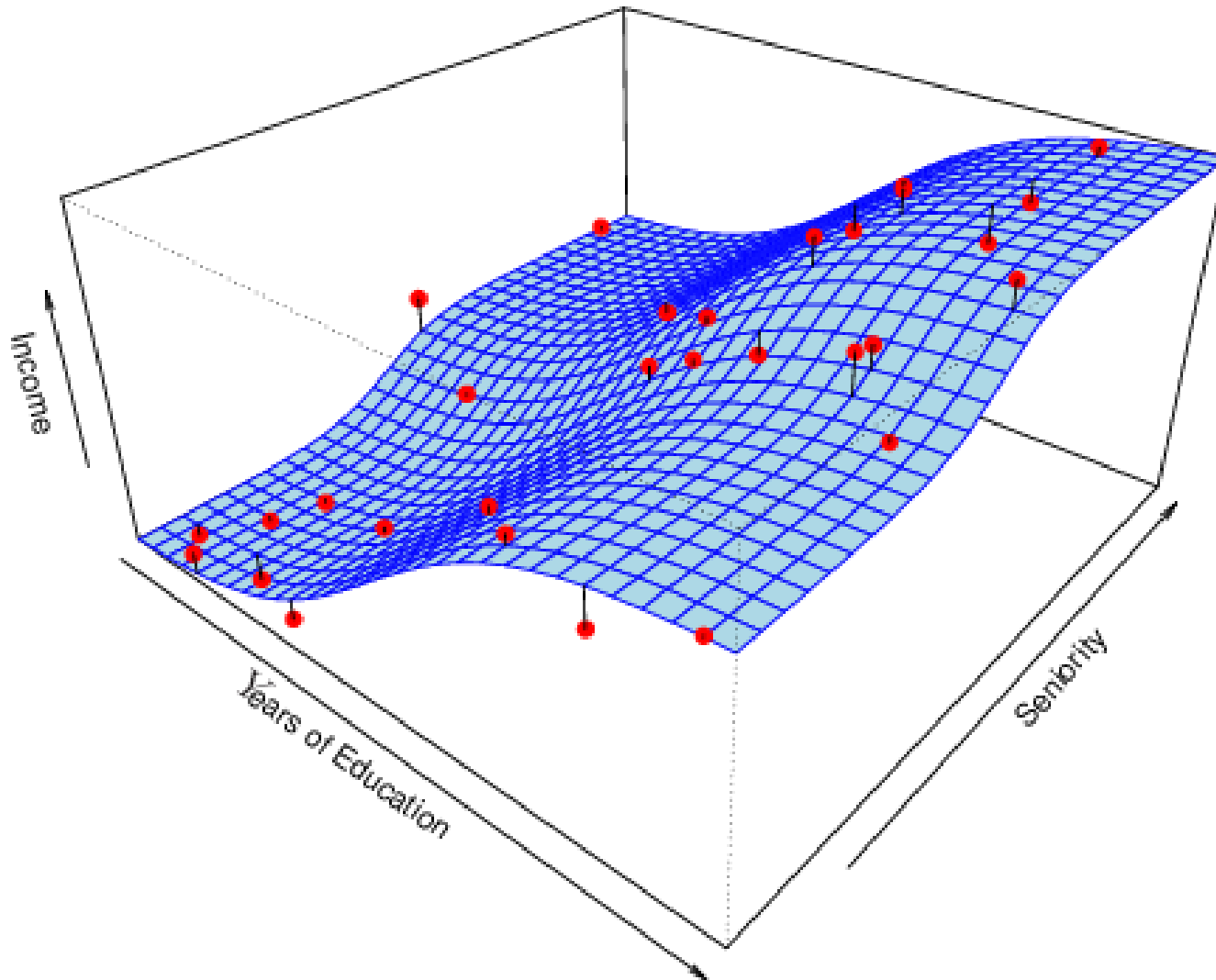
- Can you derive this? $E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2$

$$= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}}$$



$$\begin{aligned} E[(Y - \hat{Y})^2] &= \\ &= E[(f(X) + \epsilon - \hat{f}(X))^2] = \\ &= E[f(X)^2 + \epsilon^2 + \hat{f}(X)^2 - 2 \cdot \epsilon \cdot f(X) \\ &\quad - 2 \cdot \epsilon \cdot \hat{f}(X) - 2 \cdot f(X) \cdot \hat{f}(X)] = \\ &= f(X)^2 + E[\epsilon^2] + \hat{f}(X)^2 + 2 \cdot E[\epsilon] \cdot f(X) \\ &\quad - 2 \cdot E[\epsilon] \cdot \hat{f}(X) - 2 \cdot f(X) \cdot \hat{f}(X) = \\ &= f(X)^2 + E[\epsilon^2] + \hat{f}(X)^2 - 2 \cdot f(X) \cdot \hat{f}(X) = \\ &= (f(X)^2 + \hat{f}(X)^2 - 2 \cdot f(X) \cdot \hat{f}(X)) + E[\epsilon^2] = \\ &= (f(X) - \hat{f}(X))^2 + E[\epsilon^2] - 0 = \\ &= (f(X) - \hat{f}(X))^2 + \text{Var}(\epsilon) \end{aligned}$$

- Function f might also involve multiple variables ...



- There are 2 reasons for estimating f

- Prediction
- Inference



- Prediction

- If we can produce a good estimate for f (and the variance of ε is not too large) we can make accurate predictions for the response, Y/G , based on a new value of X .

- Inference

- We may be interested in the type of relationship between Y/G and the X 's to control/influence Y/G .
 - Which particular predictors actually affect the response?
 - Is the relationship positive or negative?
 - Is the relationship a simple linear one or is it more complicated etc.?

○ Direct Mail Prediction

- Interested in predicting how much money an individual will donate based on observations from 90,000 people on which we have recorded over 400 different characteristics.
- Don't care too much about each individual characteristic.
- Just want to know: For a given individual should I send out a mailing?

○ Medium House Price

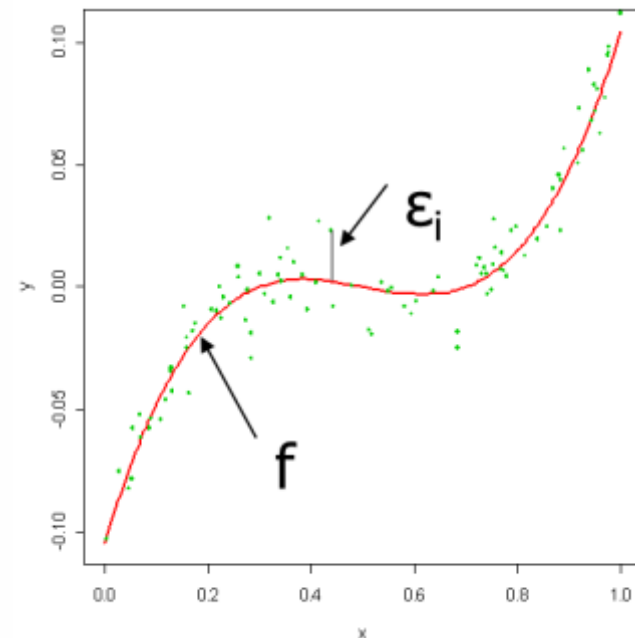
- Which factors have the biggest effect on the response
- How big the effect is.
- Want to know: how much impact does a river view have on the house value

- We have observed a set of training data

$$\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$$

- Use statistical method/model to estimate f so that for any (X, Y)

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



- Statistical methods/models are usually divided in
 - Parametric Methods/Models
 - Non-parametric Methods/Models

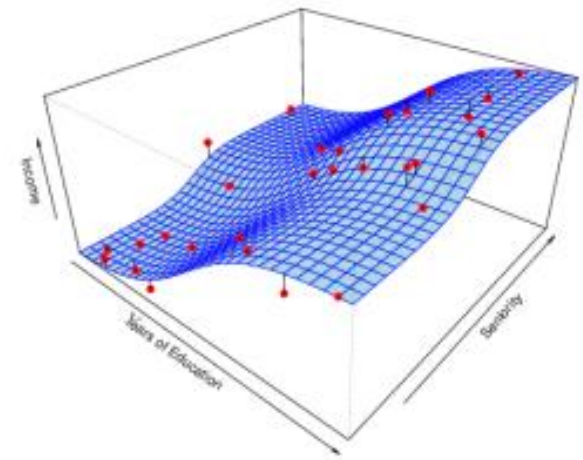
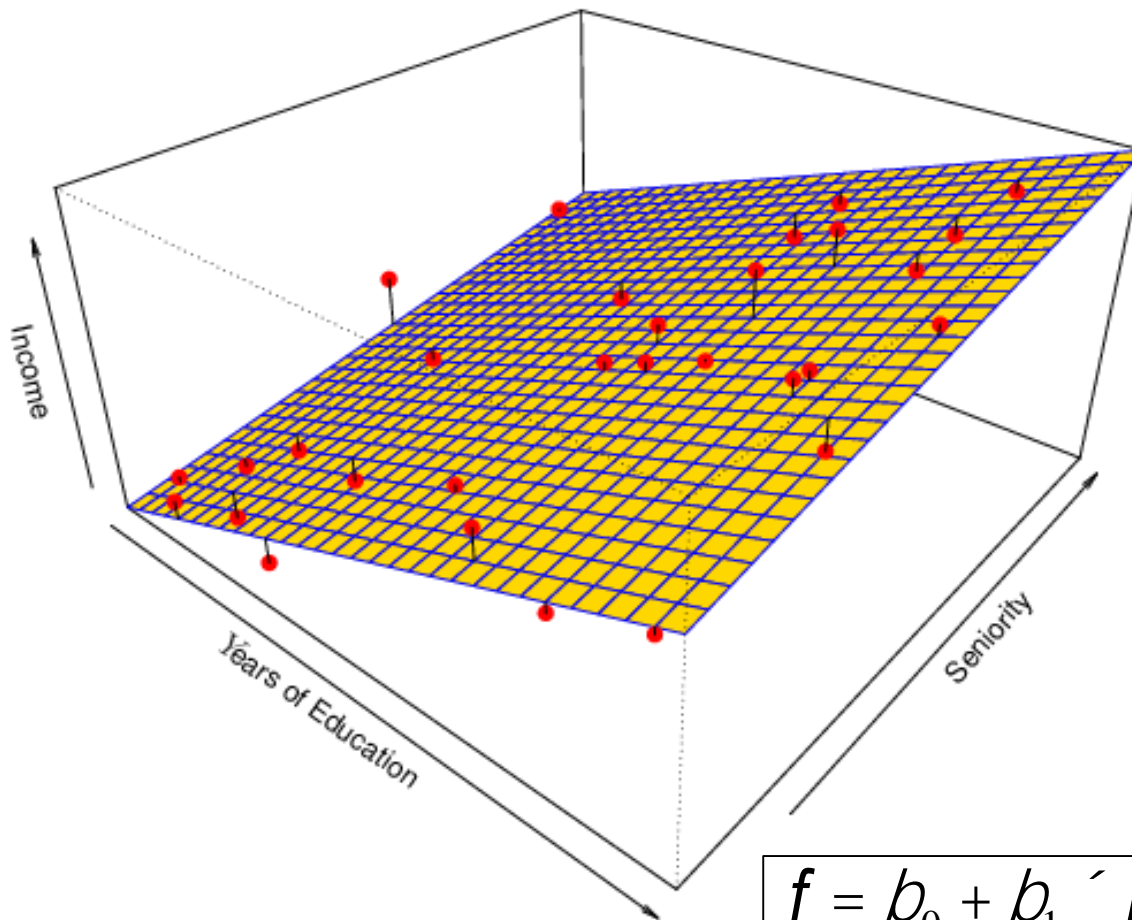
- Parametric methods leverage on an assumption about the model underlining f
 - They reduce the problem of estimating f down to the one of estimating a set of parameters
 - They involve a two-step model based approach
- STEP 1: Make some assumption about the functional form of f , i.e. come up with a model (e.g., a linear model)

$$f(\mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}$$

- STEP 2: Use the training data to fit the model, i.e., estimate f through the unknown parameters

$$\beta_0 \quad \beta_1 \quad \beta_2 \quad \dots \quad \beta_p$$

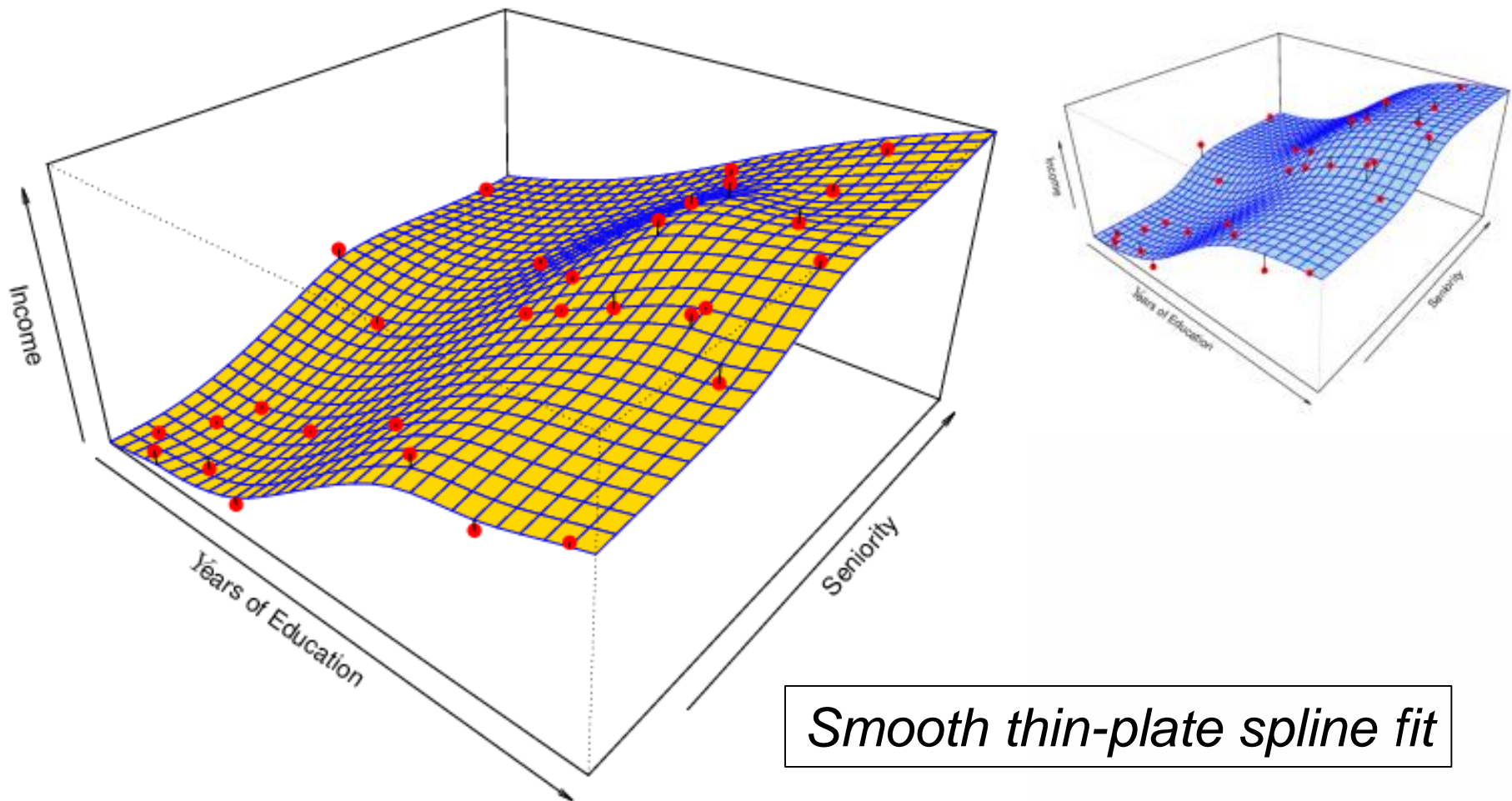
- Parametric methods leverage on an assumption about the model underlining f
 - They reduce the problem of estimating f down to the one of estimating a set of parameters
 - They involve a two-step model based approach
- STEP 1: In this course we will examine far more complicated, and flexible, models for f w.r.t linear ones. In a sense the more flexible the model the more realistic it is.
- STEP 2: The most common approach for estimating the parameters in a linear model is Ordinary Least Squares (OLS), but there are often superior approaches.



$$f = b_0 + b_1 \cdot \text{Education} + b_2 \cdot \text{Seniority}$$

- Even if the standard deviation is low we will still get a bad answer if we use the wrong model.

- Sometimes they are referred as “sample-based” or “instance-based” methods, they do not make explicit assumptions about the functional form of f , they exploit the training data “directly”
- Advantages:
 - They accurately fit a wider range of possible shapes of f
 - They do not require a “training” phase
- Disadvantages:
 - A very large number of observations is required to obtain an accurate estimate of f
 - Higher computational cost at “testing” time
 - They accurately fit a wider range of possible shapes of f .



- Non-parametric regression methods are more flexible thus they can potentially provide more accurate estimates

- Why not just use a more flexible method if it is more realistic?

Reason 1:

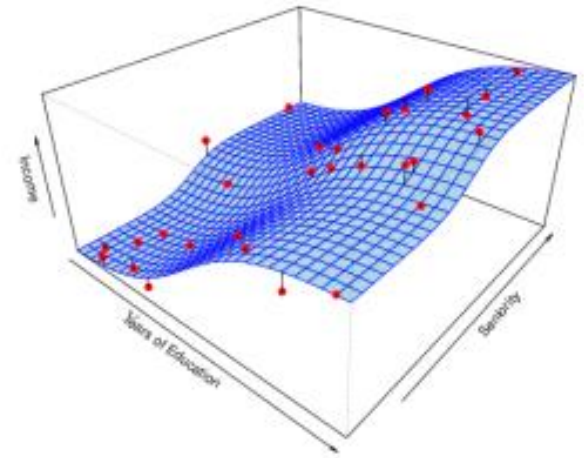
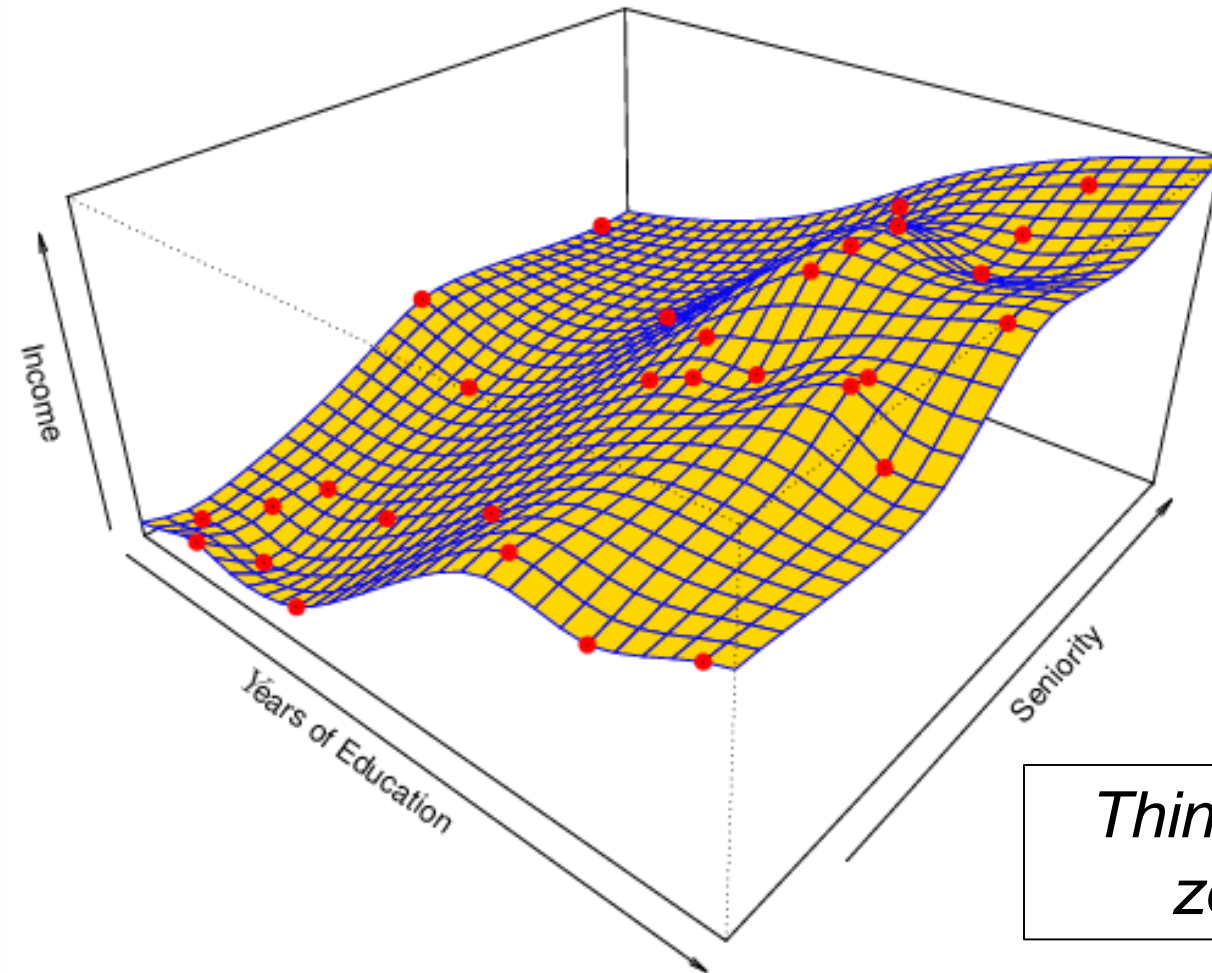
A simple method such as linear regression produces a model which is much easier to interpret (the Inference part is better).

- E.g., in a linear model, β_j is the average increase in Y for a one unit increase in X_j holding all other variables constant.

Reason 2:

Even if you are only interested in prediction, it is often possible to get more accurate predictions with a simple, instead of a complicated, model.

- This seems counter intuitive but has to do with the fact that it is harder to fit properly a more flexible model.



*Thin-plate spline fit with
zero training error*

- Non-parametric regression methods can also be too flexible and produce poor estimates for f

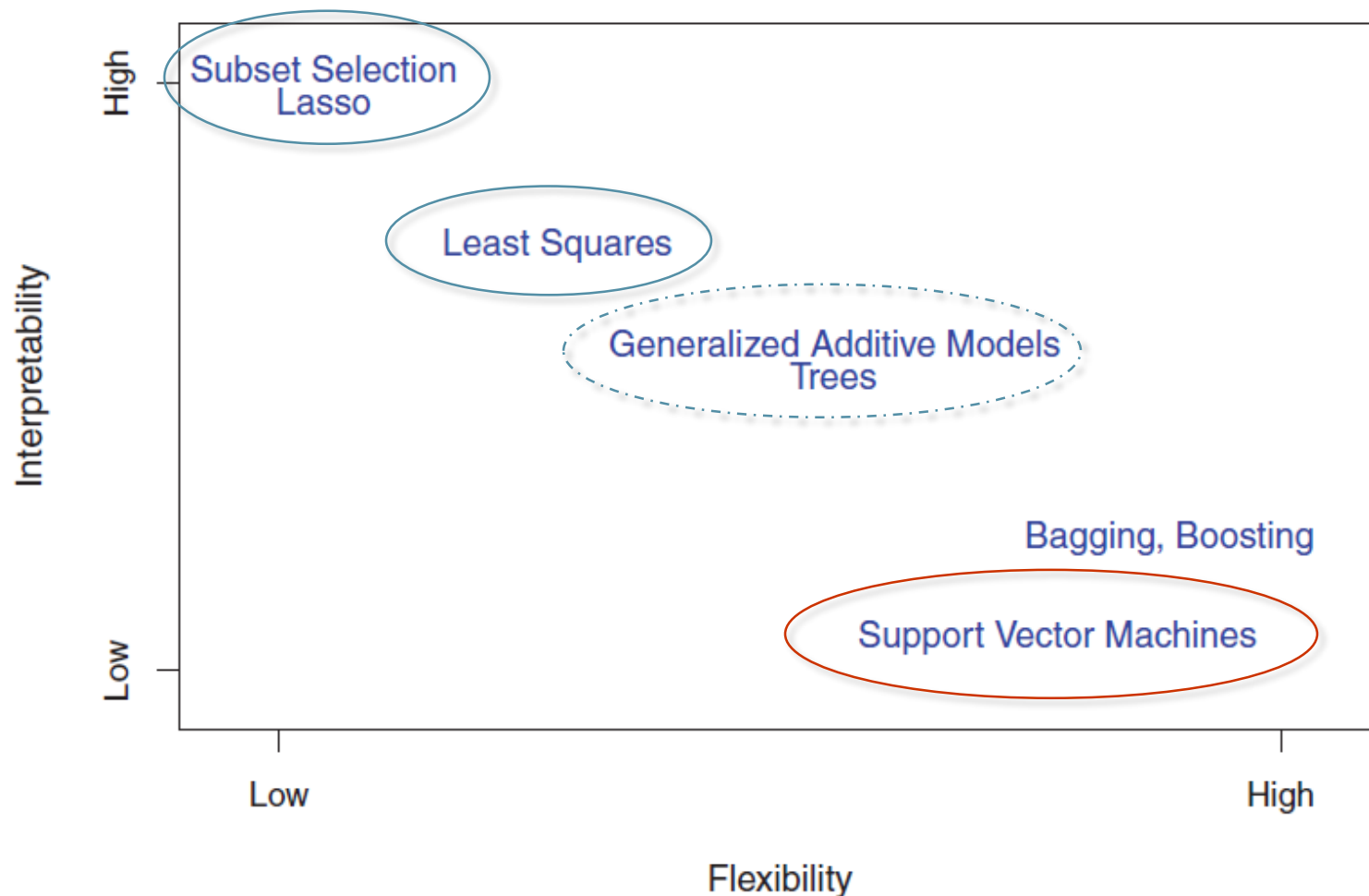
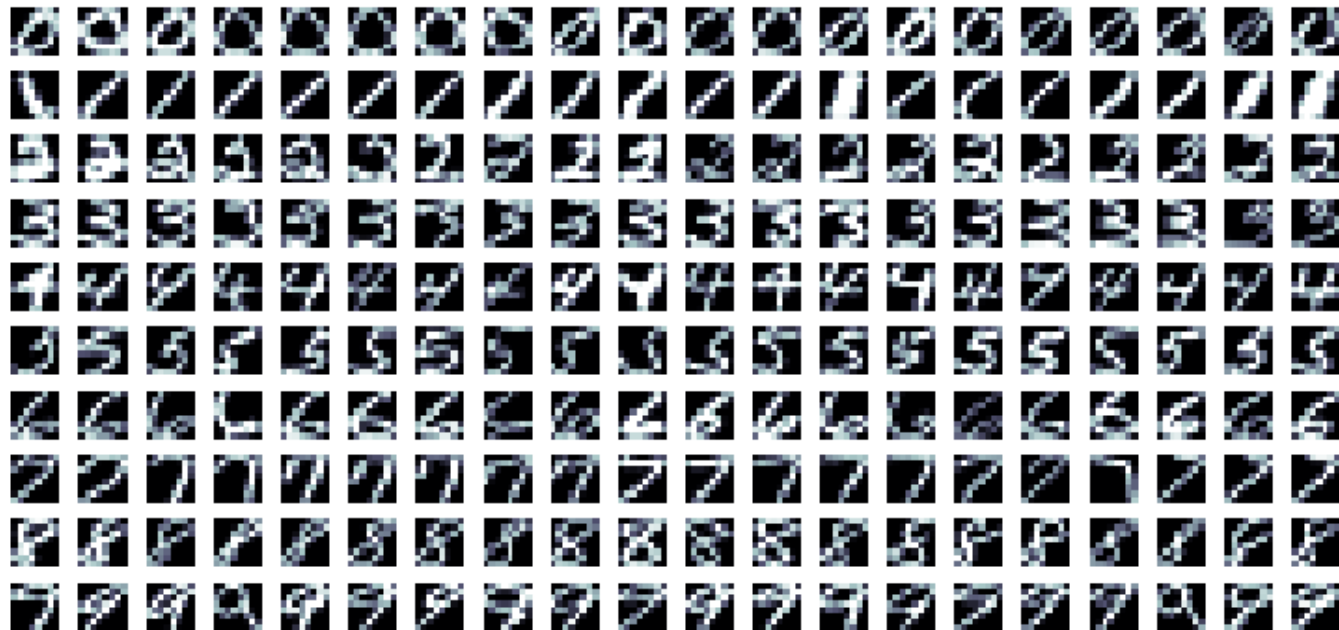
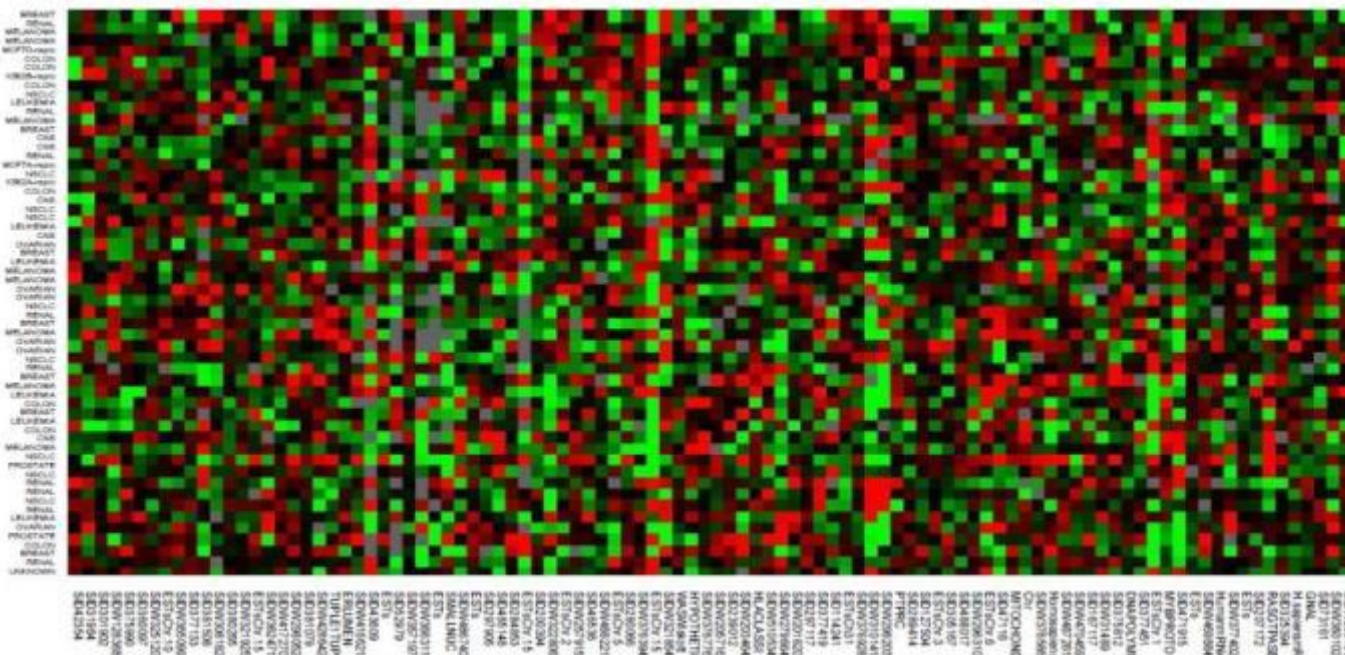


FIGURE 2.7. A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

- Machine Learning makes usually a clear distinction between
 - Supervised Models
 - Unsupervised Models
- Supervised Learning:
 - Supervised Learning is where both the predictors, X_i , and the response, Y_i , are observed.



- Machine Learning makes usually a clear distinction between
 - Supervised Models
 - Unsupervised Models
- Unsupervised Learning:
 - Only the X_i 's are observed and use them to build a high level representation (possibly for modeling some Y)



- Supervised learning problems can be further divided into
 - Regression problems cover situations where Y is continuous/numerical
 - Predicting the value of the Dow in 6 months
 - Predicting the value of a given house based on various inputs.
 - Classification problems cover situations where Y is categorical
 - Will the Dow be up (U) or down (D) in 6 months?
 - Is this email a SPAM or not?

TABLE 1.1. *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between spam and email.*

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

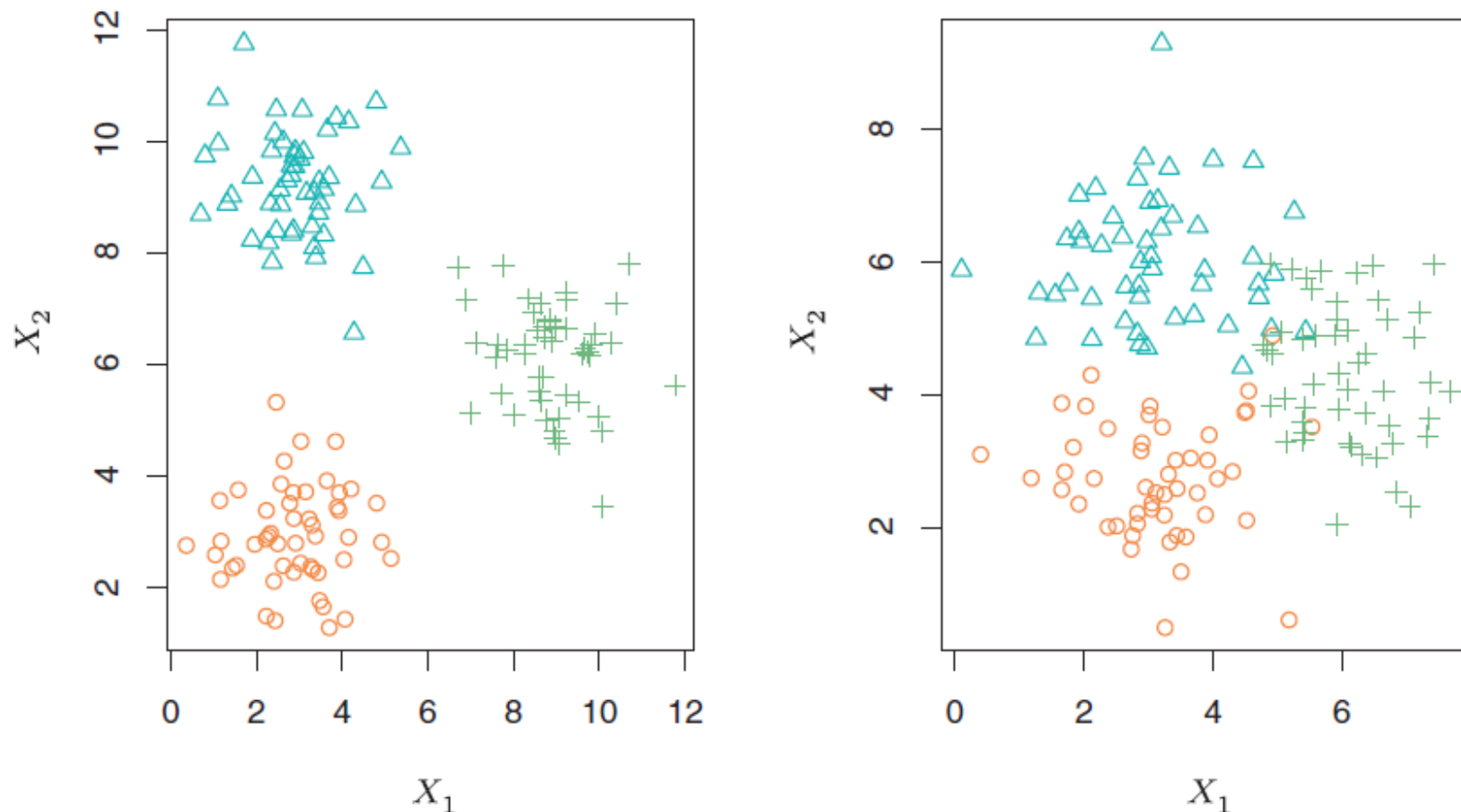
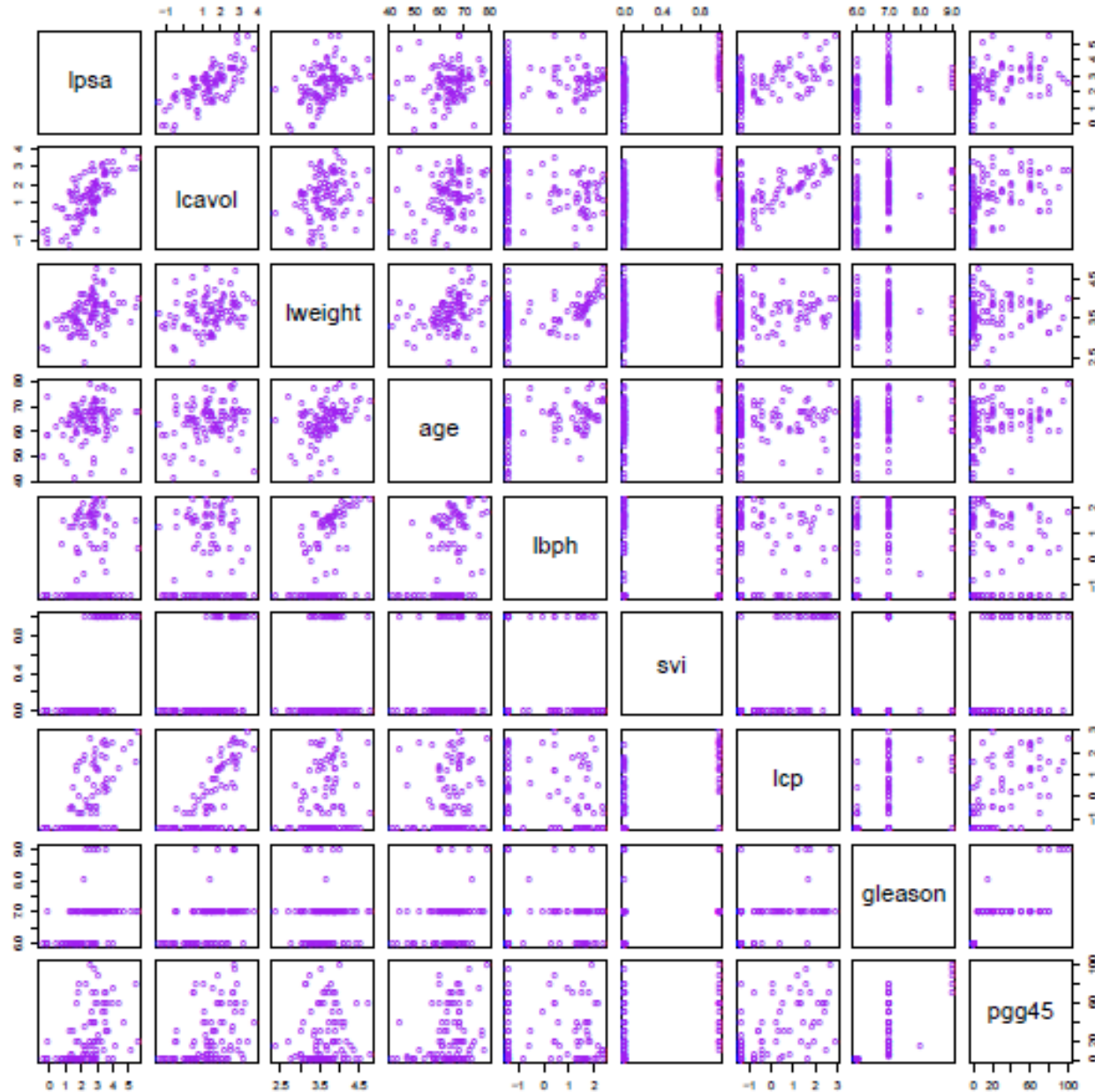


FIGURE 2.8. A clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering task is more challenging.

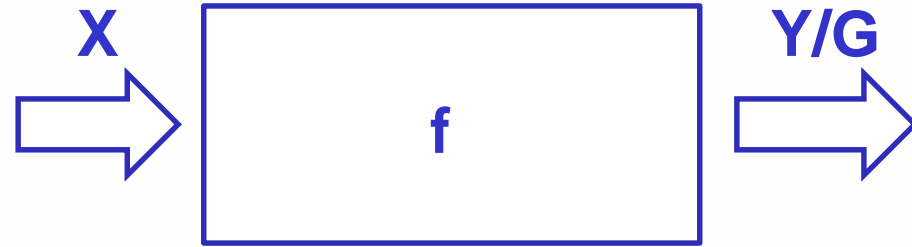
What about higher dimensions?

25



- What Is Statistical Learning?

- Why estimate f ?
- How do we estimate f ?
- The trade-off between prediction accuracy & model interpretability



- Some important taxonomies (I expect you'll know this by heart!)

- Prediction vs. Inference
- Parametric vs. Non Parametric models
- Regression vs. Classification problems
- Supervised vs. Unsupervised learning
- ...