COSC 3337 : Data Science I

N. Rizk

College of Natural and Applied Sciences

Department of Computer Science

University of Houston

Outline

- What Is Statistical Learning?
  - Why estimate f?
  - How do we estimate f?
  - The trade-off between prediction accuracy and model interpretability
  - Supervised vs. unsupervised learning
  - Regression vs. classification problems

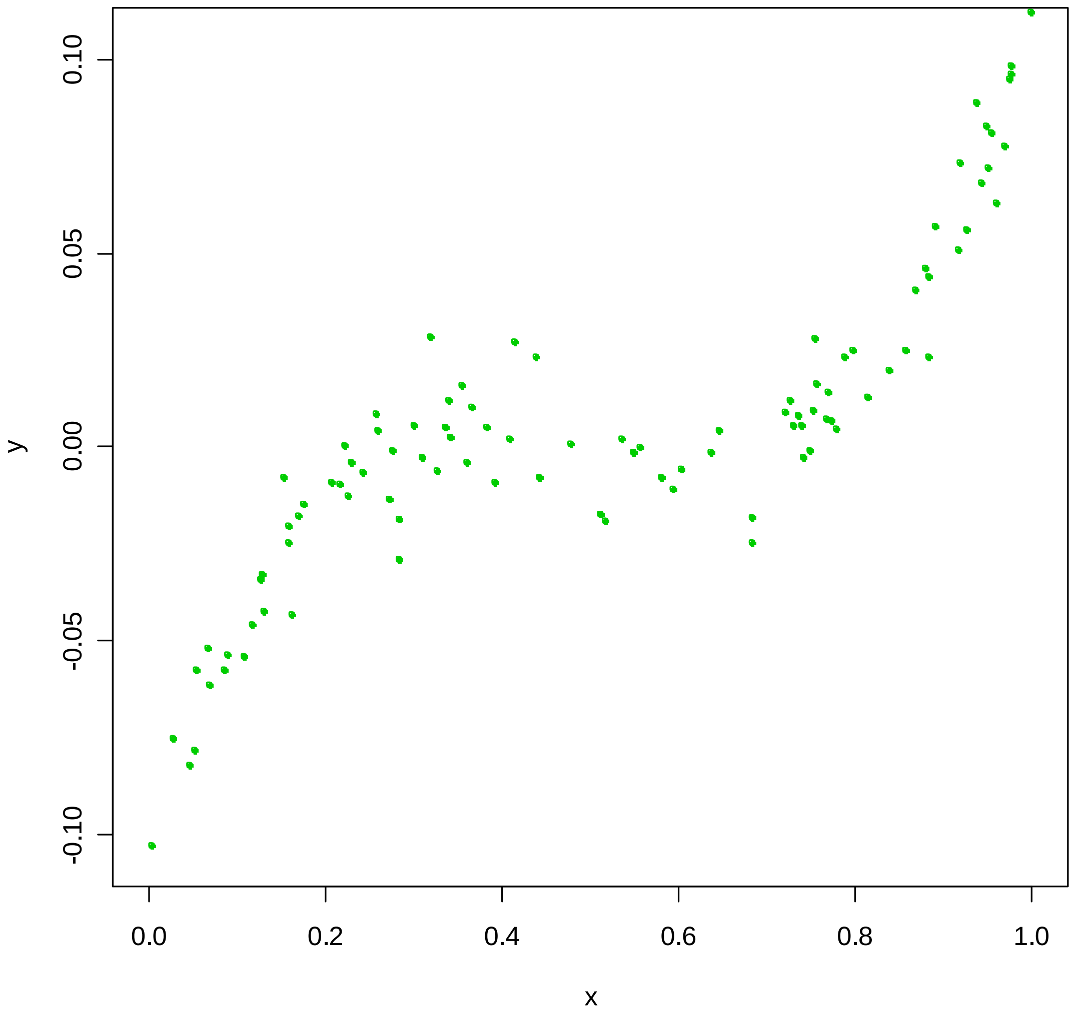
What is Statistical Learning?

- Suppose we observe Y_i and $X_i = (X_{i1}, \ldots, X_{ip})$ for $i = 1, \ldots, n$
- We believe that there is a relationship between Y and at least one of the X's.
- We can model the relationship as

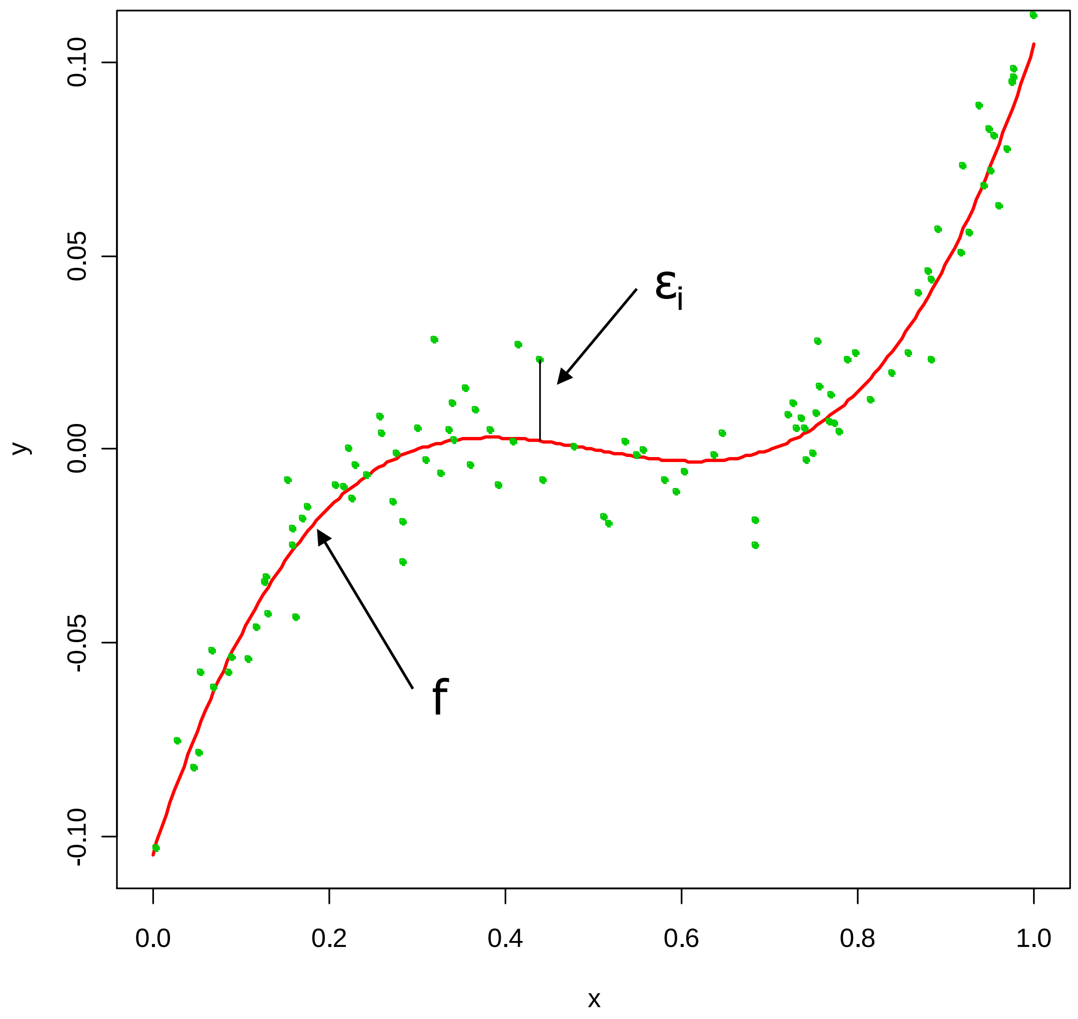
$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

- Where f is an unknown function and ε is a random error with mean zero.

A Simple Example

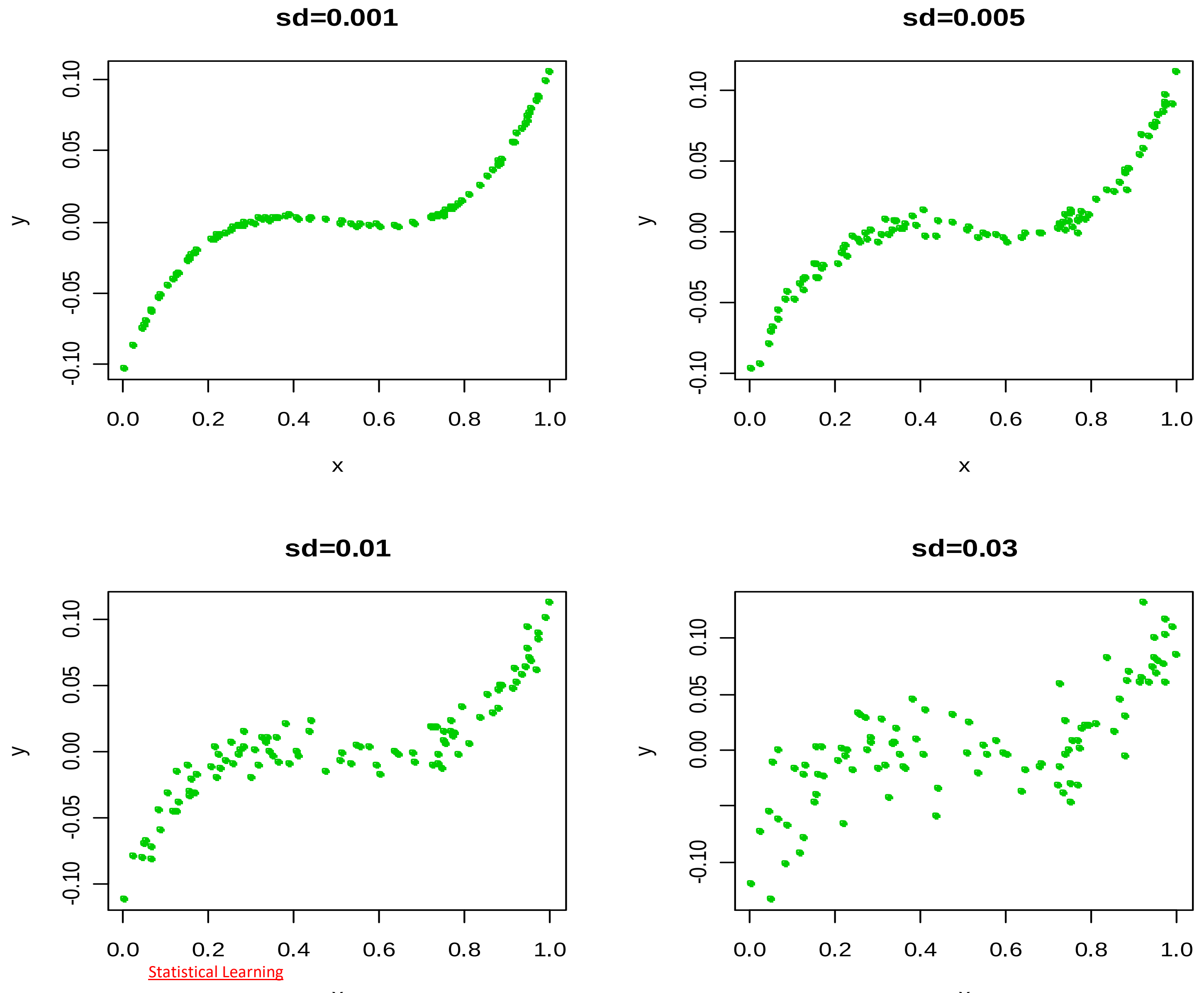


A Simple Example

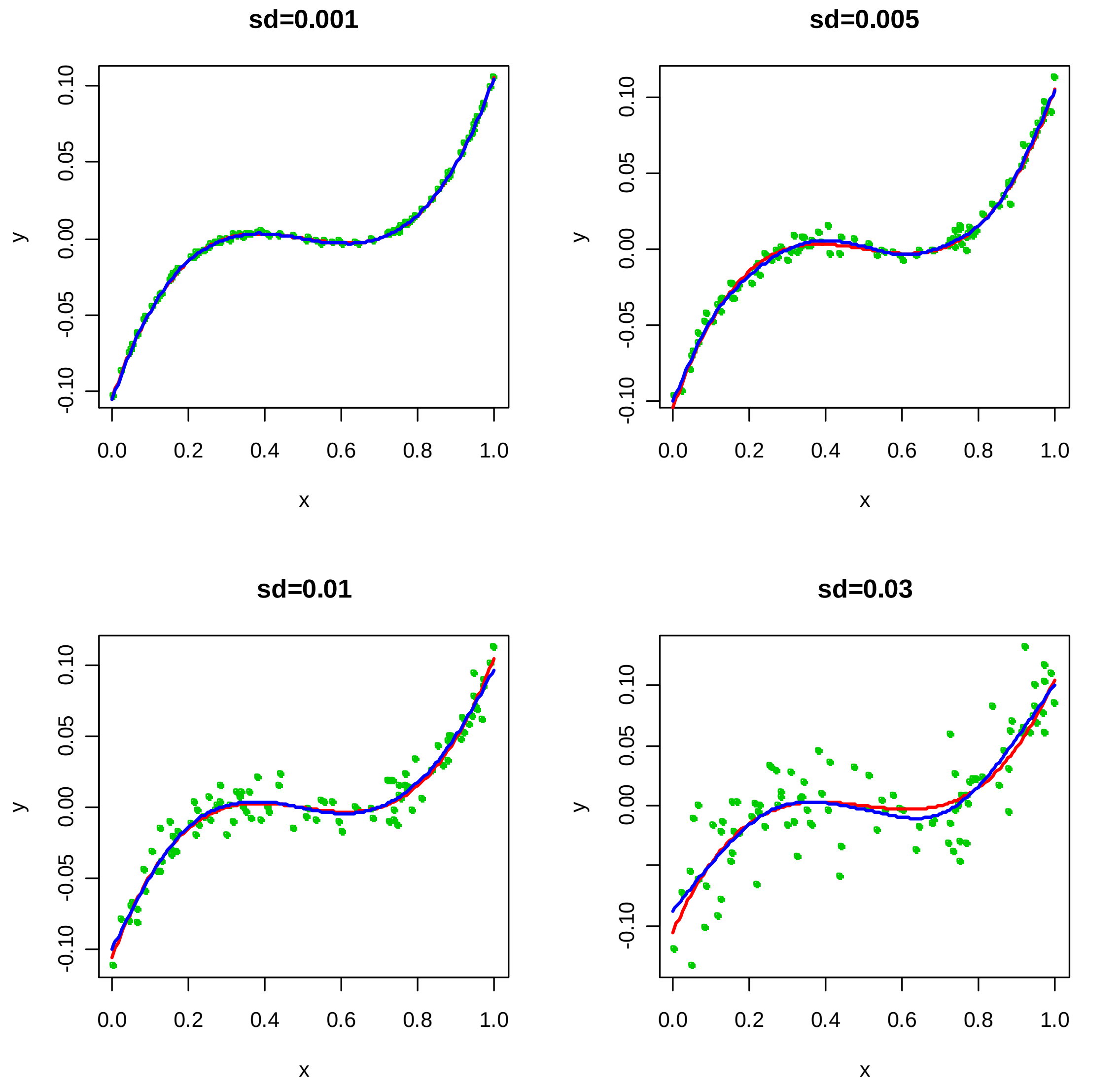


Different Standard Deviations

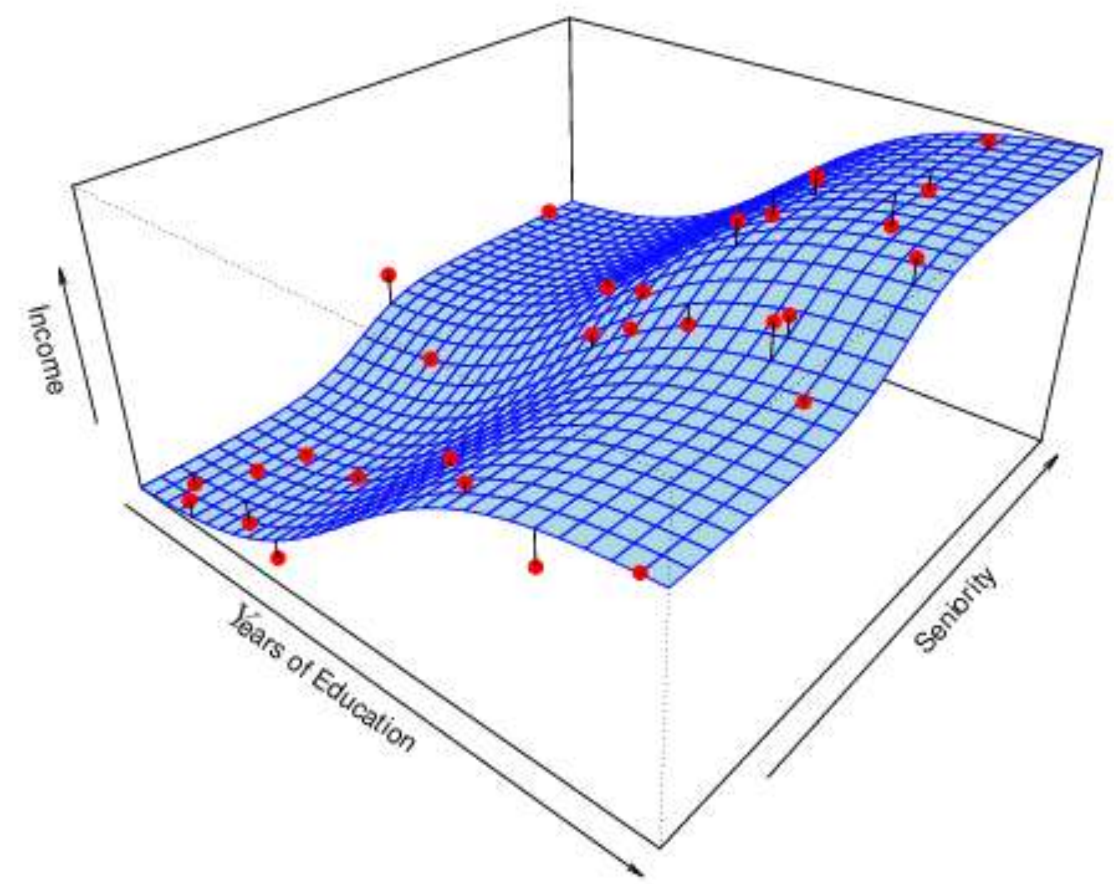
- The difficulty of estimating f will depend on the standard deviation of the ε's.



Different Estimates For f



Income vs. Education Seniority



Why Do We Estimate f?

- Statistical Learning, and this course, are all about how to estimate f.
- The term statistical learning refers to using the data to “learn” f.
- Why do we care about estimating f?
- There are 2 reasons for estimating f,
  - **Prediction** and
  - **Inference**.

1. 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, based on a new value of **X**.

Example: Direct Mailing 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?

2. Inference

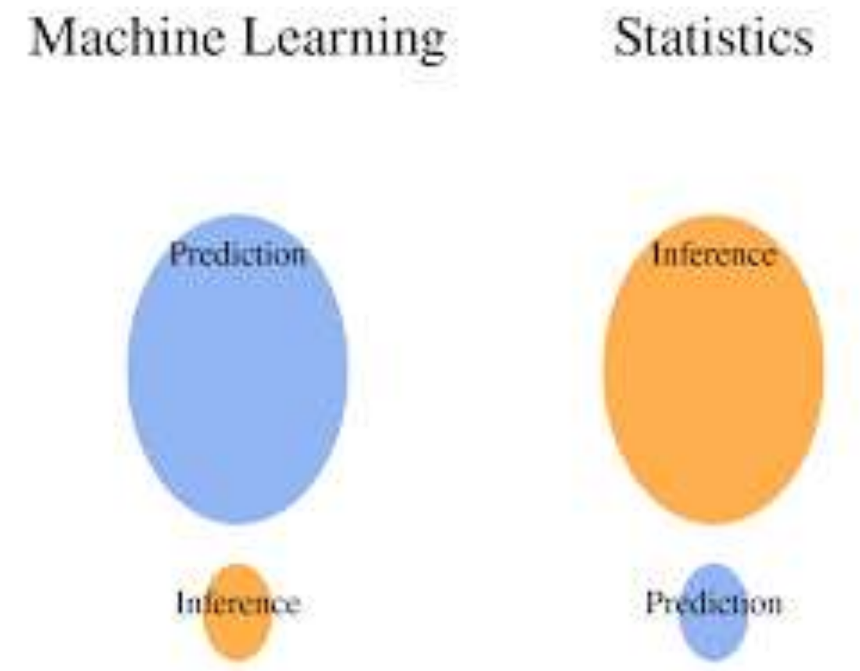
- Alternatively, we may also be interested in the type of relationship between Y and the X's.
- For example,
  - Which particular predictors actually affect the response?
  - Is the relationship positive or negative?
  - Is the relationship a simple linear one or is it more complicated etc.?

Example: Housing Inference

- Wish to predict median house price based on 14 variables.
- Probably want to understand which factors have the biggest effect on the response and how big the effect is.
- For example how much impact does a river view have on the house value etc.

Inference:

Given a set of data you want to **infer** how the output is generated as



Prediction: Given a new measurement, you want to use an existing data set to build a model that reliably chooses the correct identifier from a set of outcomes.

How Do We Estimate f?

- We will assume we have observed a set of **training data**

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

- We must then use the training data and a statistical method to estimate f.
- Statistical Learning Methods:
  - Parametric Methods
  - Non-parametric Methods

Parametric Methods

- It reduces the problem of estimating f down to 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. The most common example is a linear model i.e.

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

Parametric Methods (cont.)

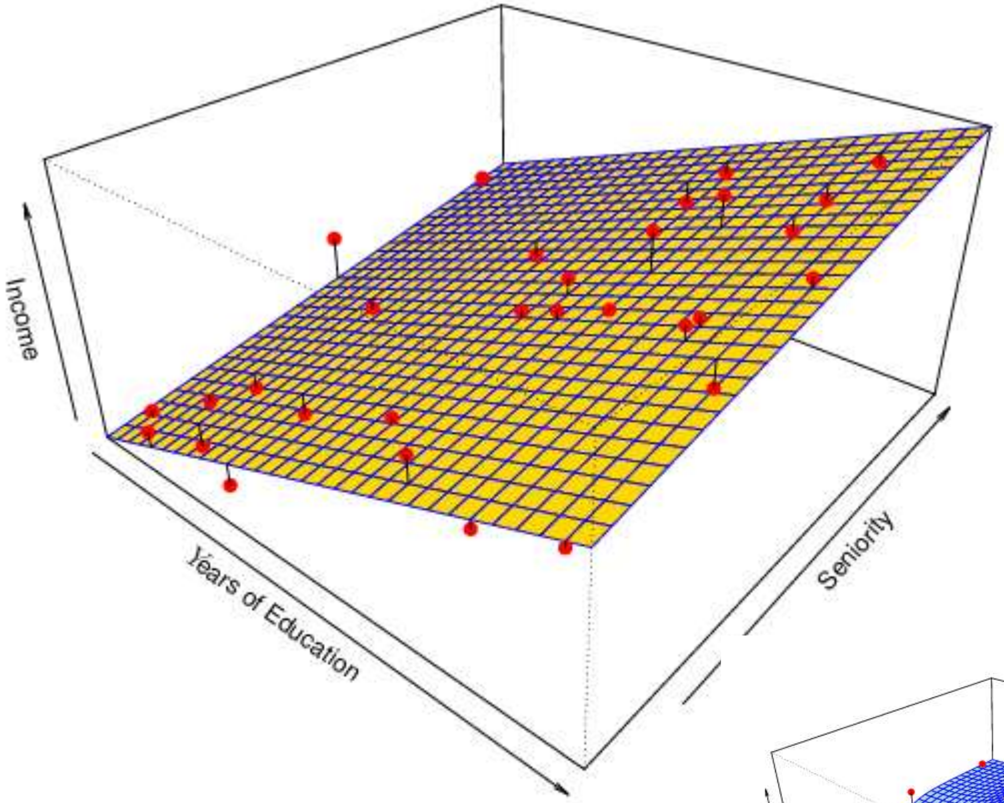
STEP 2:

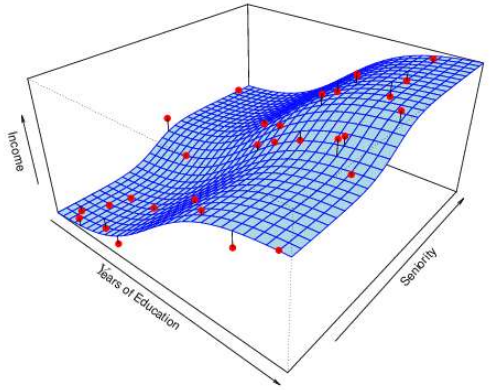
Use the training data to fit the model i.e. estimate *f* or equivalently the unknown parameters such as β_0, β_1, β_2,..., β_p.(Learn the coefficients for the function from the training data.)

- The most common approach for estimating the parameters in a linear model is ordinary least squares (OLS).
- However, this is only one way.
- We will see in the course that there are often superior approaches.

Example: A Linear Regression Estimate

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





$$f = b_0 + b_1 \times Education + b_2 \times Seniority$$

Parametric machine learning algorithms include:

- Logistic Regression
- Linear Discriminant Analysis
- Perceptron
- Naive Bayes
- Simple Neural Networks

Benefits of Parametric Machine Learning Algorithms:

- **Simpler**: These methods are easier to understand and interpret results.
- **Speed**: Parametric models are very fast to learn from data.
- **Less Data**: They do not require as much training data and can work well even if the fit to the data is not perfect.

Limitations of Parametric Machine Learning Algorithms:

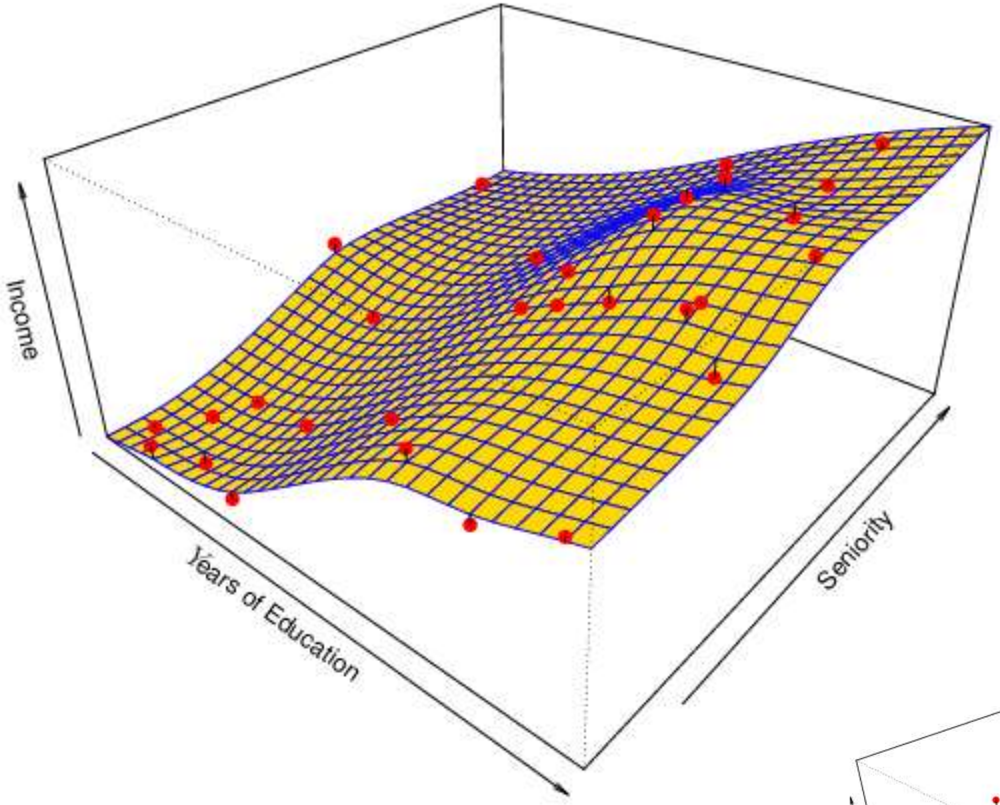
Constrained: By choosing a functional form these methods are highly constrained to the specified form.
Limited Complexity: The methods are more suited to simpler problems.
Poor Fit: In practice the methods are unlikely to match the underlying mapping function.

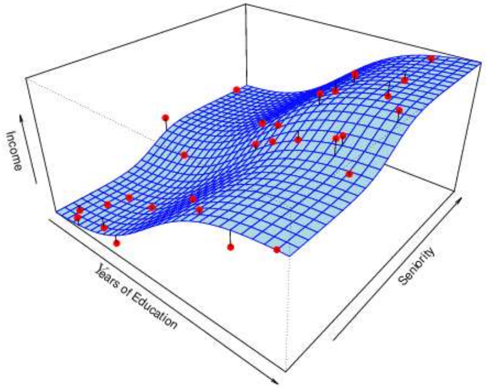
Non-parametric Methods

- They do not make explicit assumptions about the functional form of f.
- Advantages: They accurately fit a wider range of possible shapes of f.
- Disadvantages: A very large number of observations is required to obtain an accurate estimate of f

Example: A Thin-Plate Spline Estimate

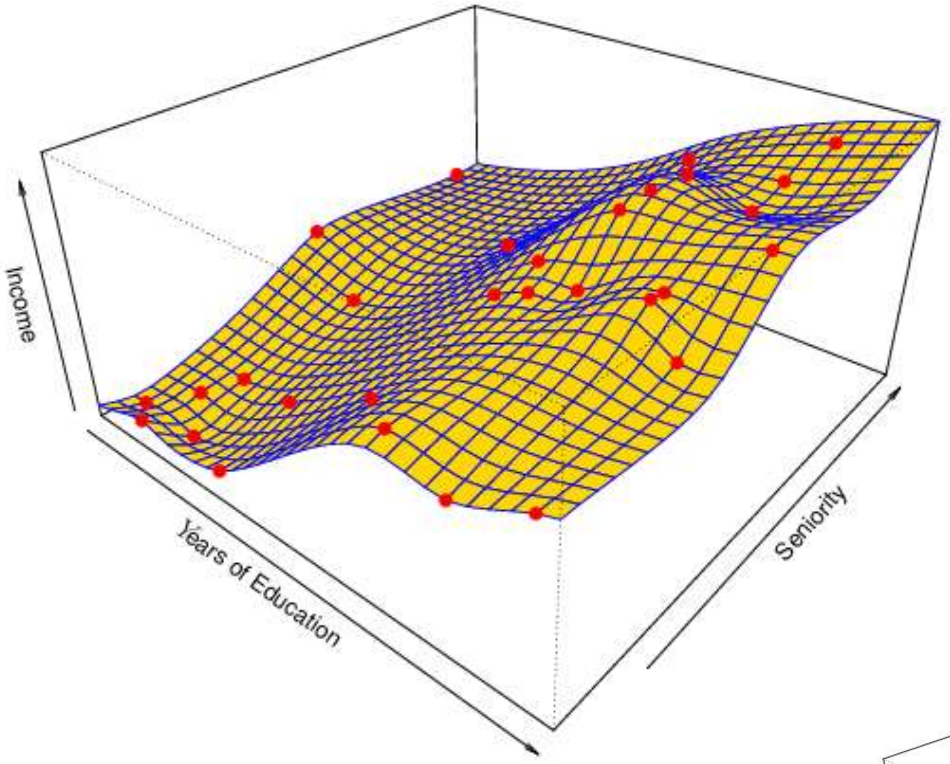
- Non-linear regression methods are more flexible and can potentially provide more accurate estimates.

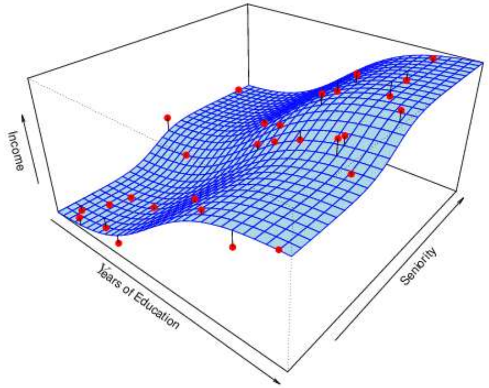




A Poor Estimate

- Non-linear regression methods can also be too flexible and produce poor estimates for *f*.





Some more examples of popular nonparametric machine learning algorithms are:

k-Nearest Neighbors
Decision Trees like CART and C4.5
Support Vector Machines

Benefits of Nonparametric Machine Learning Algorithms:

Flexibility: Capable of fitting a large number of functional forms.
Power: No assumptions (or weak assumptions) about the underlying function.
Performance: Can result in higher performance models for prediction.

Limitations of Nonparametric Machine Learning Algorithms:

More data: Require a lot more training data to estimate the mapping function.
Slower: A lot slower to train as they often have far more parameters to train.
Overfitting: More of a risk to overfit the training data and it is harder to explain why specific predictions are made.

A **parametric** algorithm has a **fixed number of parameters**. A parametric algorithm is **computationally faster**, but makes **stronger assumptions** about the data; the algorithm may work well if the assumptions turn out to be correct, but it may perform badly if the assumptions are wrong. A common example of a parametric algorithm is **linear regression**.

In contrast, a **non-parametric** algorithm uses a **flexible number of parameters**, and the number of parameters often **grows as it learns from more data**. A non-parametric algorithm is **computationally slower**, but makes **fewer assumptions** about the data. A common example of a non-parametric algorithm is **K-nearest neighbor**.

To summarize, the **trade-offs** between parametric and non-

Tradeoff Between Prediction Accuracy & Model Interpretability

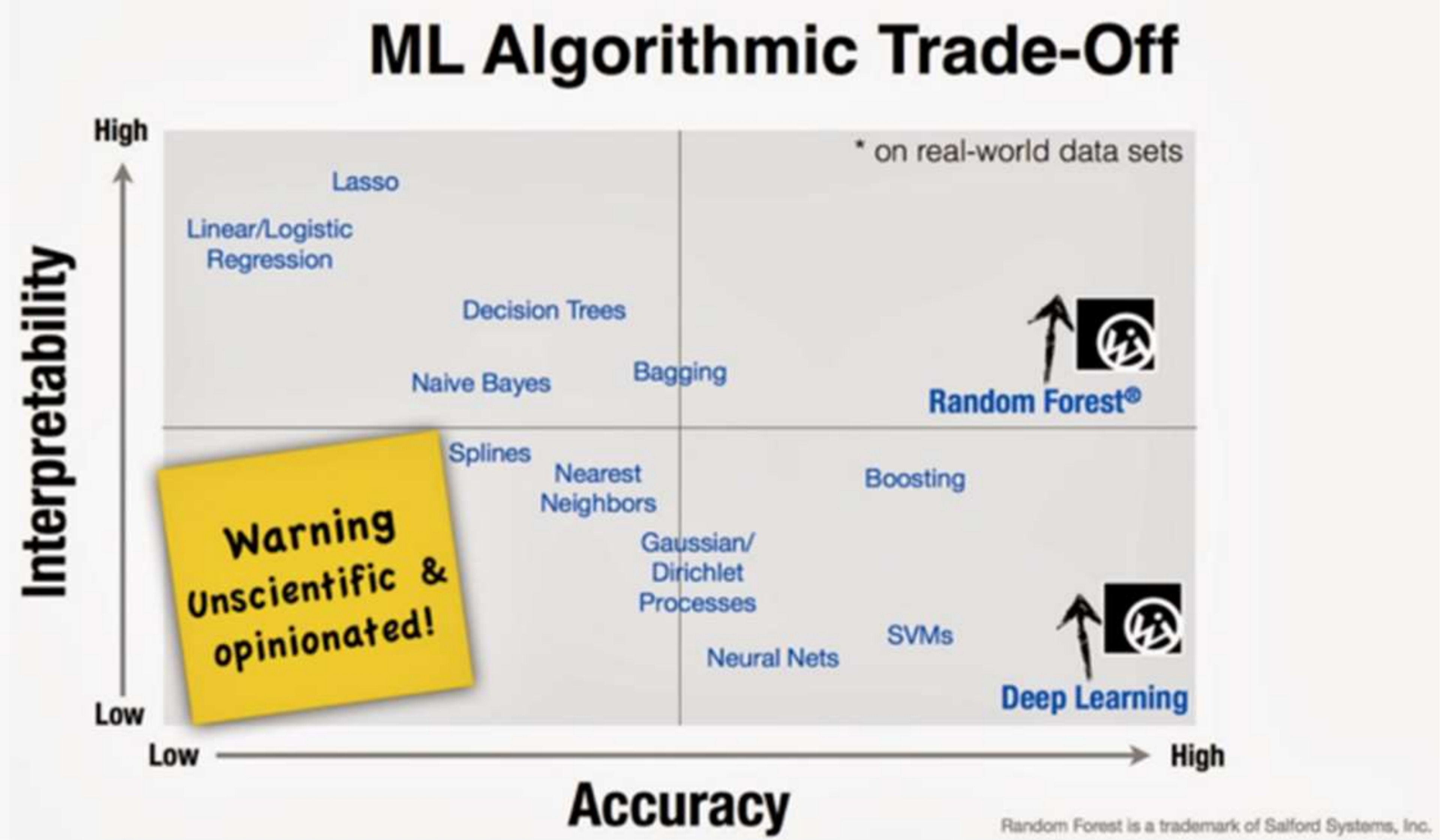
- Why not just use a more flexible method if it is more realistic?
- There are two reasons

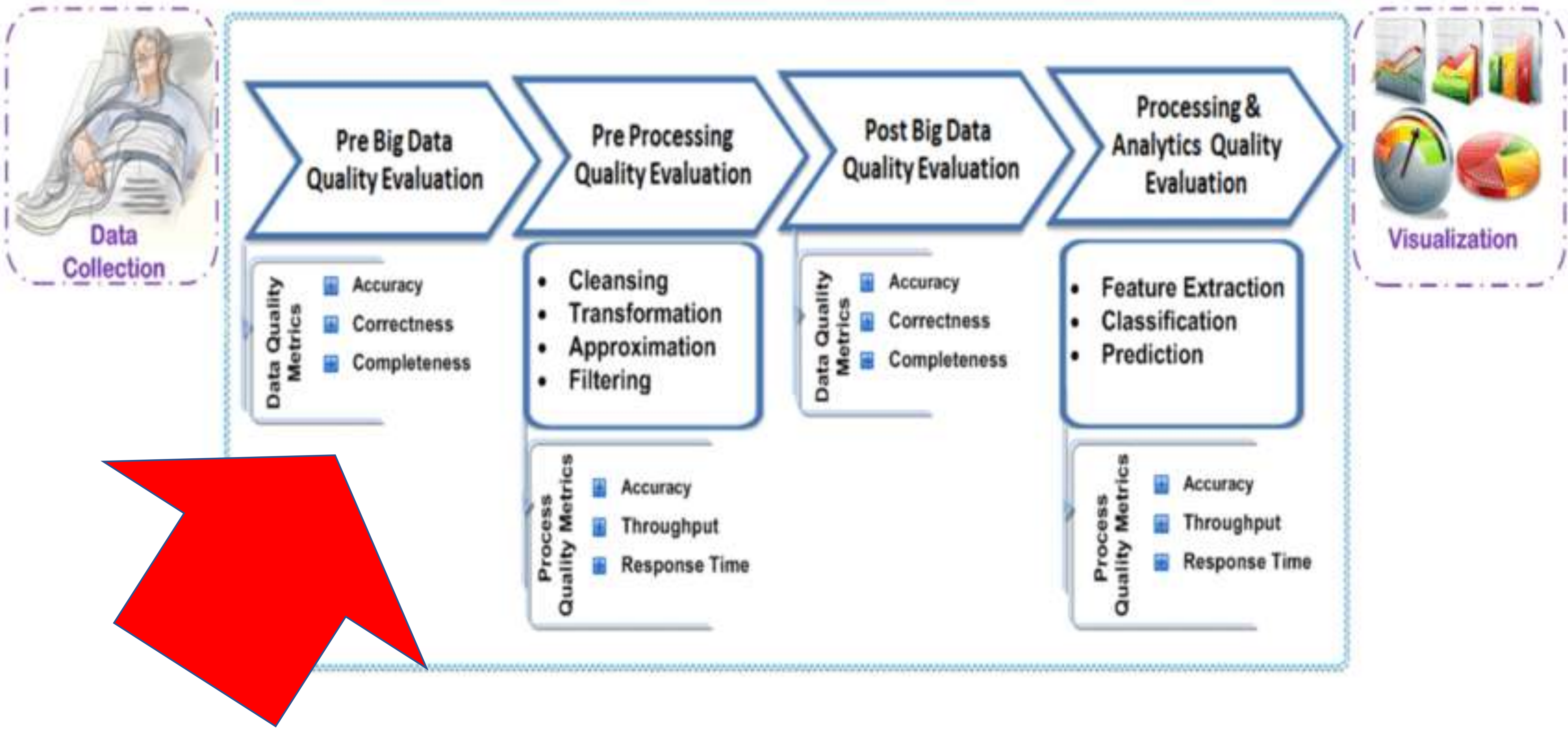
Reason 1:

A simple method such as linear regression produces a model which is much easier to interpret (the Inference part is better). For example, 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, so the first reason is not relevant, 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 a more flexible model.





Data Collection
Pre Big Data Quality Evaluation
Data Quality Metrics
Accuracy
Correctness
Completeness
Pre Processing Quality Evaluation
Cleansing
Transformation
Approximation
Filtering
Process Quality Metrics
Accuracy
Throughput
Response Time
Post Big Data Quality Evaluation
Data Quality Metrics
Accuracy
Correctness
Completeness
Processing & Analytics Quality Evaluation
Feature Extraction
Classification
Prediction
Process Quality Metrics
Accuracy
Throughput
Response Time
Visualization

Supervised vs Unsupervised

There is a bunch of different fruits

Supervised

Based on its color/shape/weight…

- Is that "fruit" an apple?

Unsupervised

- How the different fruits can be classified inside your grocery store?

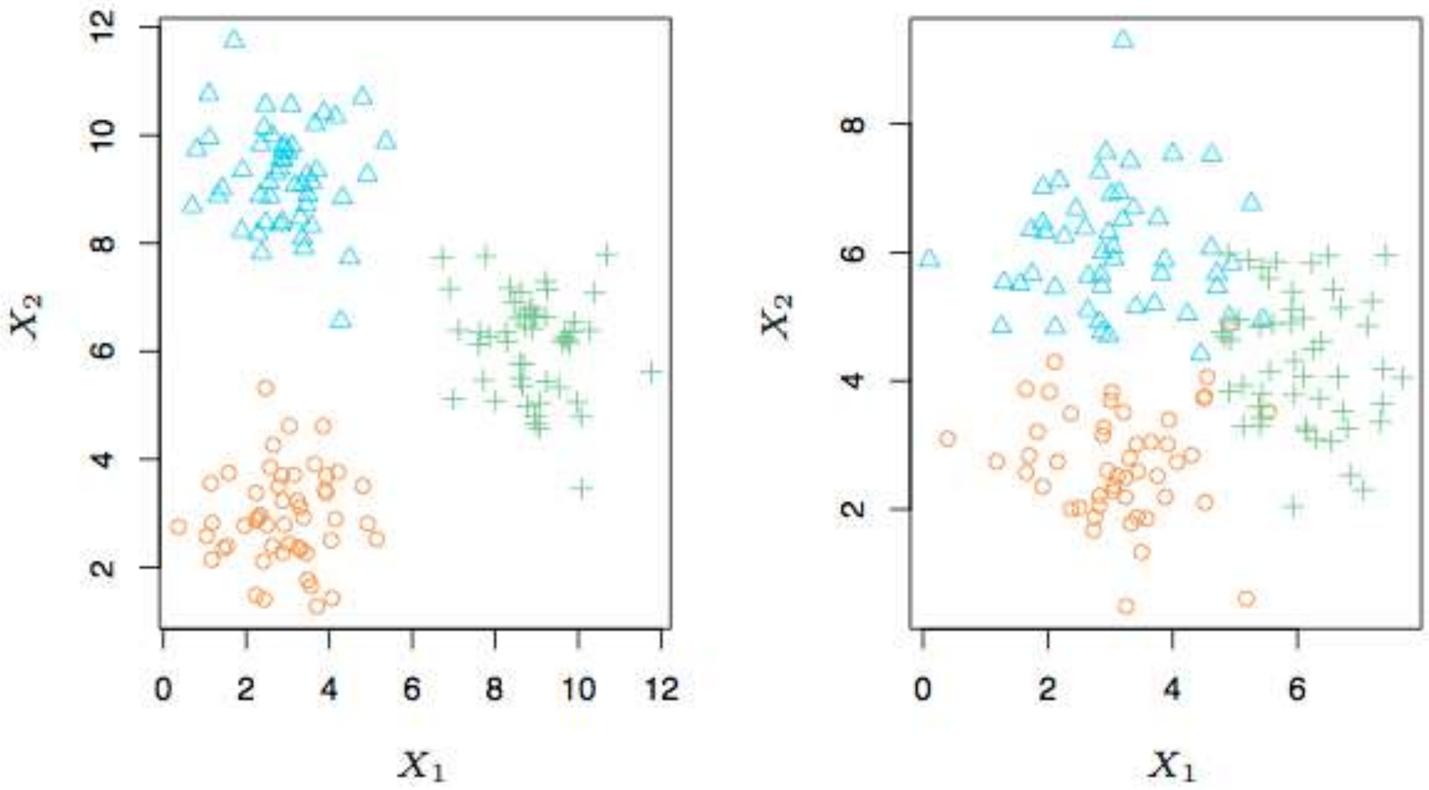
Supervised vs. Unsupervised Learning

- We can divide all learning problems into Supervised and Unsupervised situations
- <u>Supervised Learning:</u>
  - Supervised Learning is where both the predictors, $\mathbf{X}_i$, and the response, Y_i, are observed.

Supervised vs. Unsupervised Learning

- **Unsupervised Learning:**
  - In this situation only the $\mathbf{X}_i$'s are observed.
  - We need to use the $\mathbf{X}_i$'s to guess what Y would have been and build a model from there.
  - A common example is market segmentation where we try to divide potential customers into groups based on their characteristics.
  - A common approach is clustering.

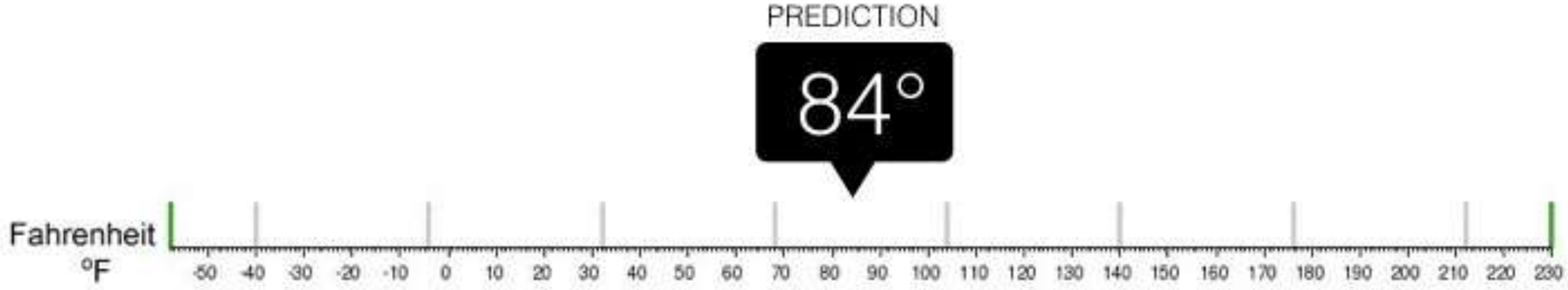
A Simple Clustering Example



Supervised learning

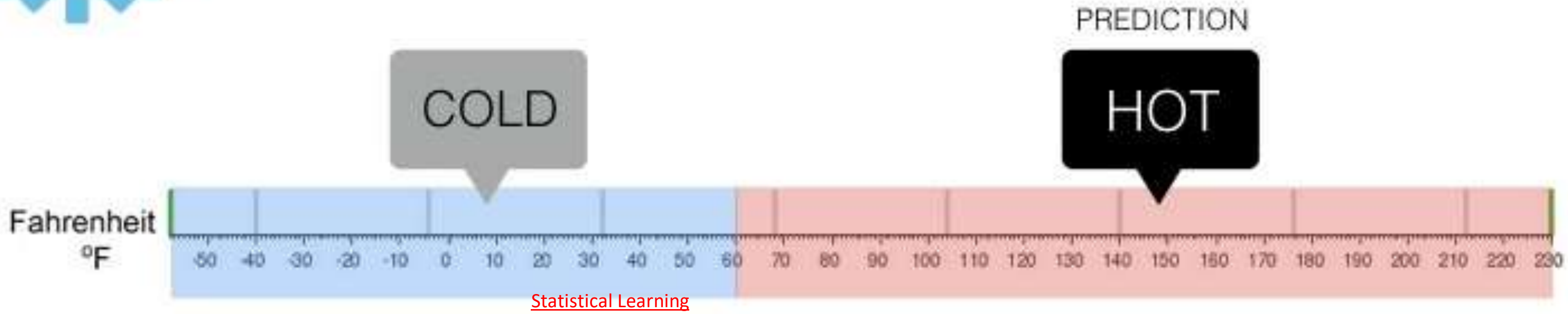
Regression

What is the temperature going to be tomorrow?



Classification

Will it be Cold or Hot tomorrow?



Regression vs. Classification

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

Summary

- prediction accuracy and model interpretability
- Parametric vs non parametric
- Supervised vs. unsupervised learning
- Regression vs. classification problems