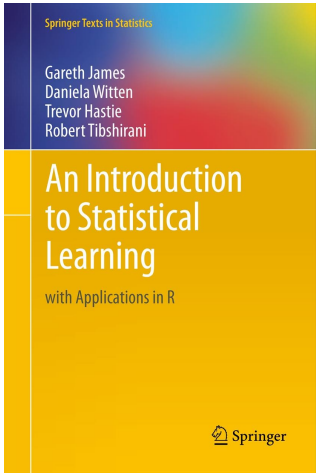
Statistical Learning. Introduction

Jose Ameijeiras Alonso
Departamento de Estatística, An. Mat. e Optimización (USC)

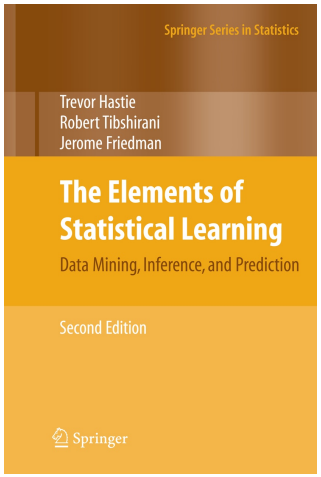
Manuel Mucientes Molina
Departamento de Electrónica e Computación (USC)

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

Statistical learning



For this course, we will mainly follow the book An Introduction to Statistical Learning with application in R, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.



At some point, we will take additional material from the more advanced book The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman.

* The content of these slides is based on material by Beatriz Pateiro.

What is statistical learning?

> Gaining knowledge, making predictions, making decisions or constructing models from a set of data
>
> *Bousquet et al. (2004) Introduction to Statistical Learning Theory.*

- The aim for this session is to explain the basic concepts in statistical learning:
  - supervised learning vs. unsupervised learning
  - regression vs. classification

What is statistical learning?

Example 1: Suppose that we are hired by a client to provide advice on how to improve sales of a particular product.

Data set available at http://www-bcf.usc.edu/ gareth/ISL/data.html

Images from http://www.vecteezy.com/

What is statistical learning?

Example 1: Suppose that we are hired by a client to provide advice on how to improve sales of a particular product.

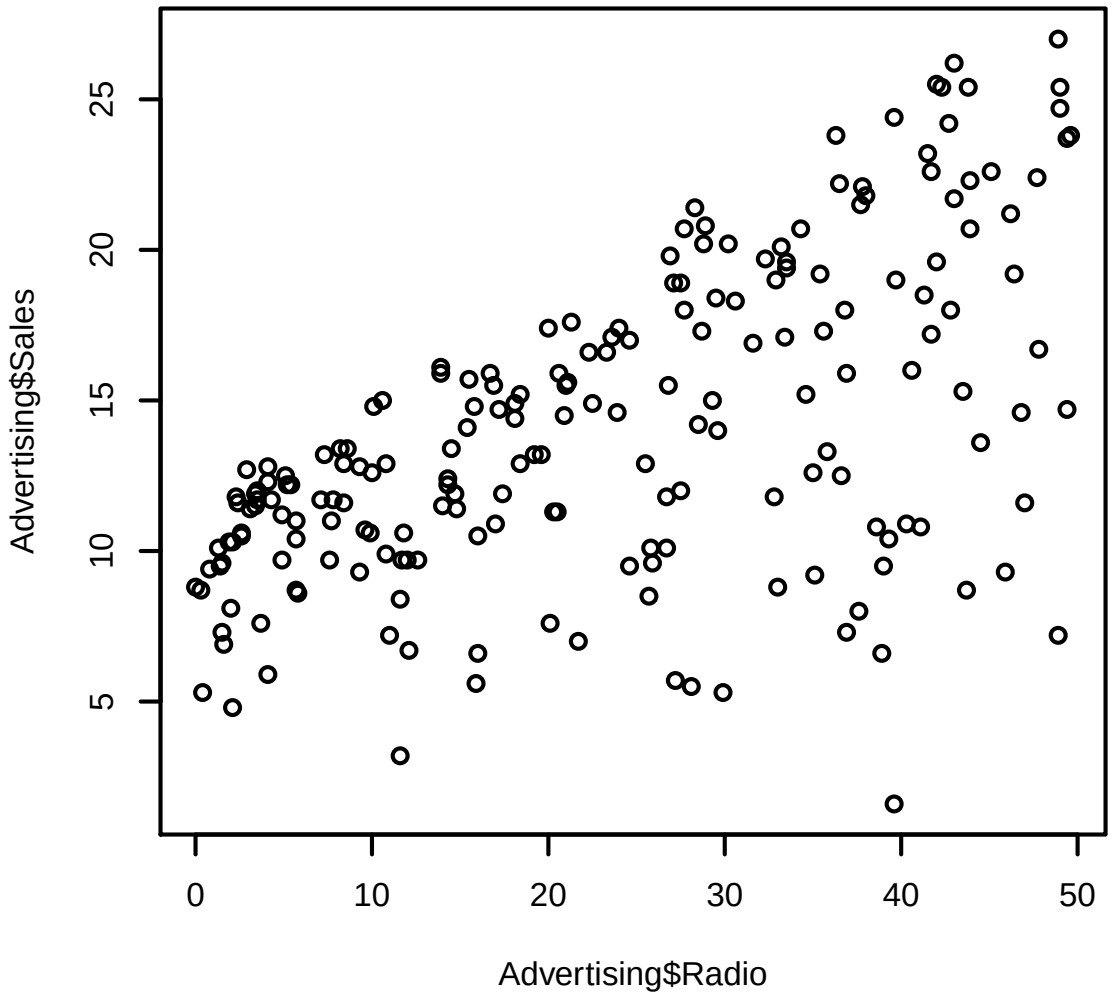
The Advertising data set consists of the sales of that product in different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper

Data set available at http://www-bcf.usc.edu/ gareth/ISL/data.html

Images from http://www.vecteezy.com/

What is statistical learning?

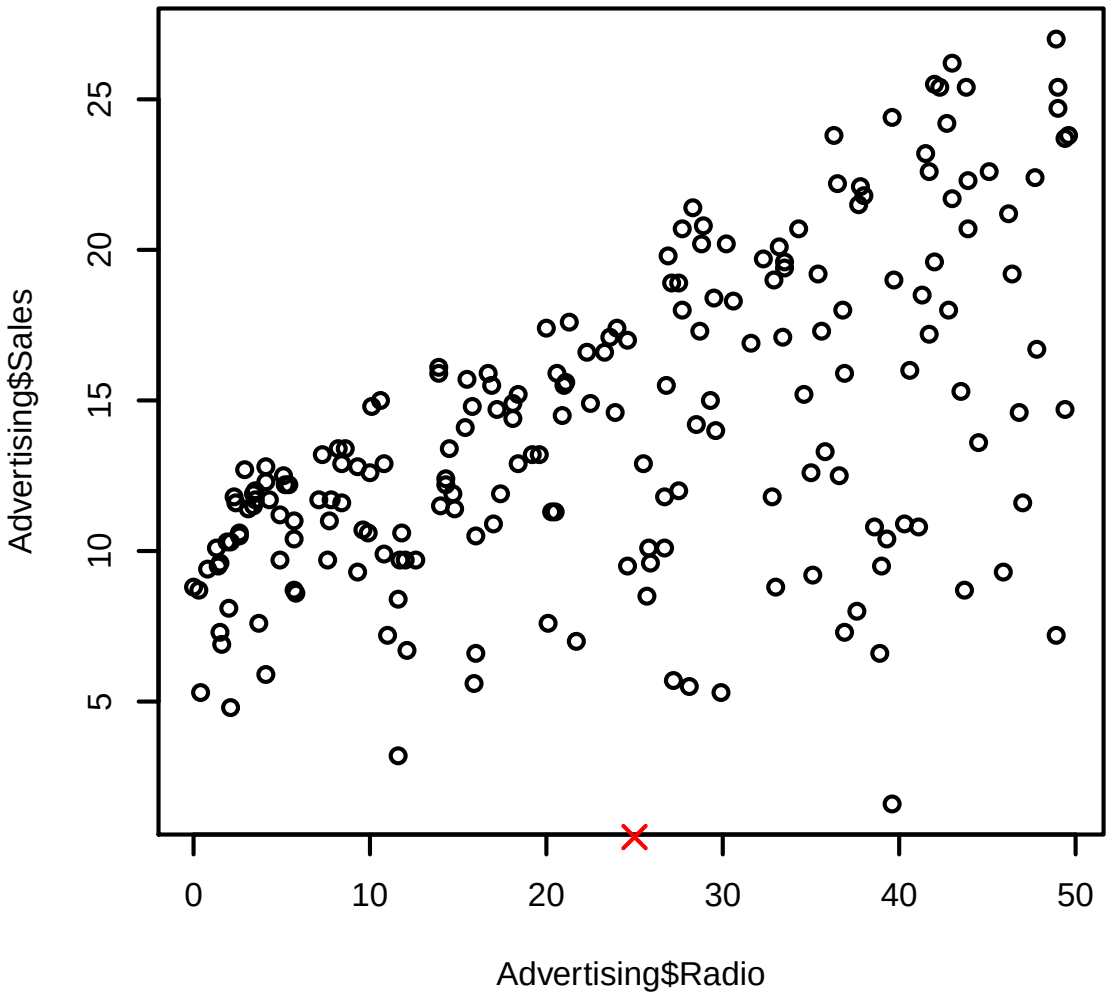
```
> Advertising <- read.csv("datasets/Advertising.csv")
> plot(Advertising$Radio, Advertising$Sales)
```



1. Observe a phenomenon

What is statistical learning?

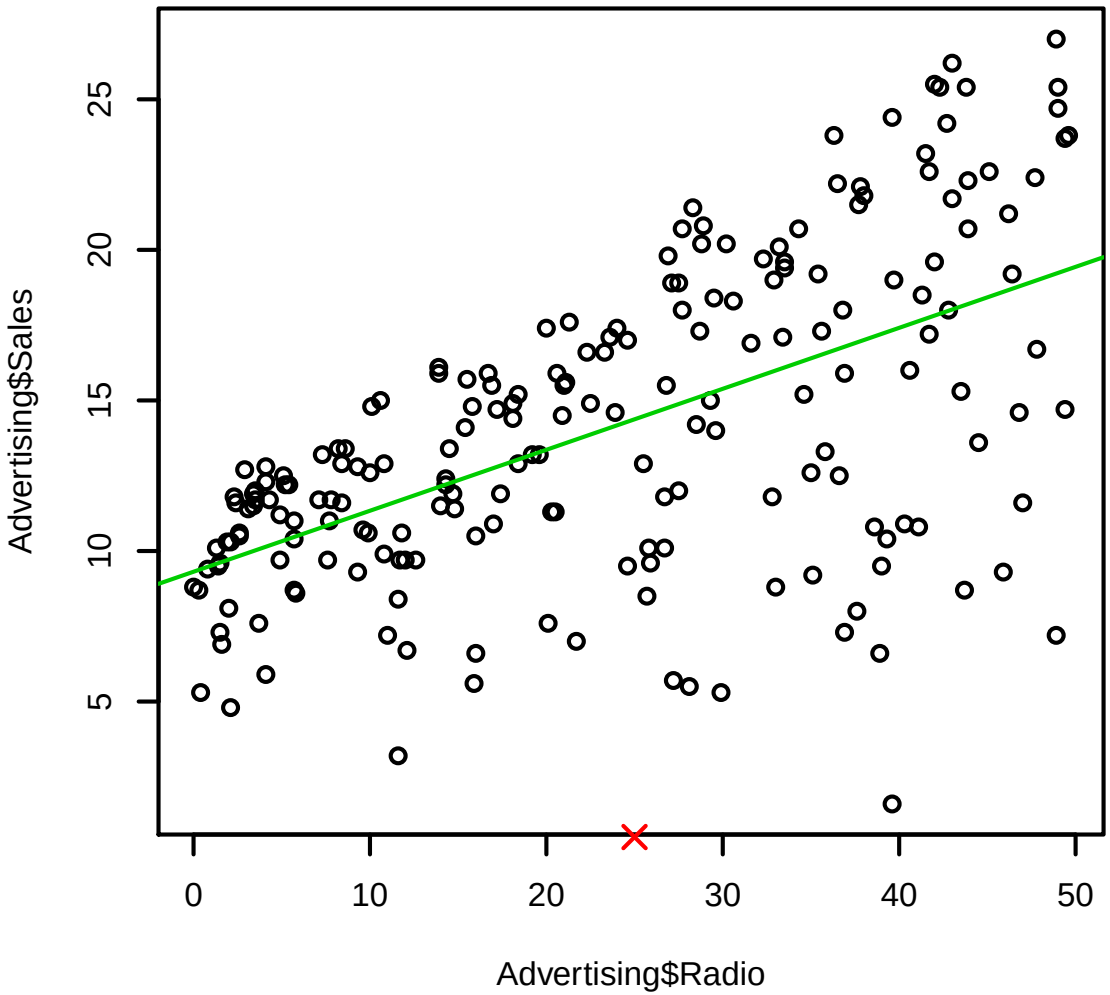
```
> Advertising <- read.csv("datasets/Advertising.csv")
> plot(Advertising$Radio, Advertising$Sales)
```



1. Observe a phenomenon

What is statistical learning?

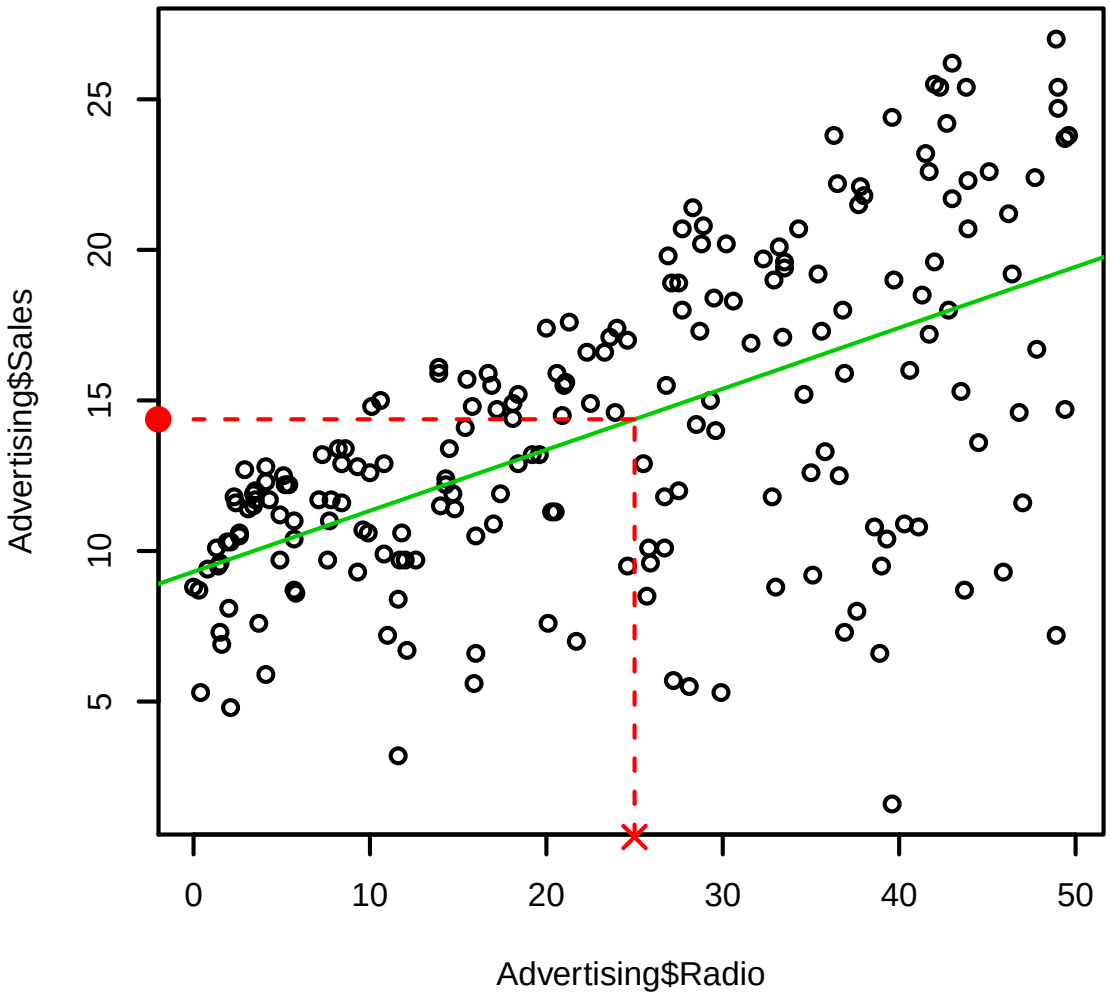
```
> Advertising <- read.csv("datasets/Advertising.csv")
> plot(Advertising$Radio, Advertising$Sales)
```



1. Observe a phenomenon
2. Construct a model for that phenomenon

What is statistical learning?

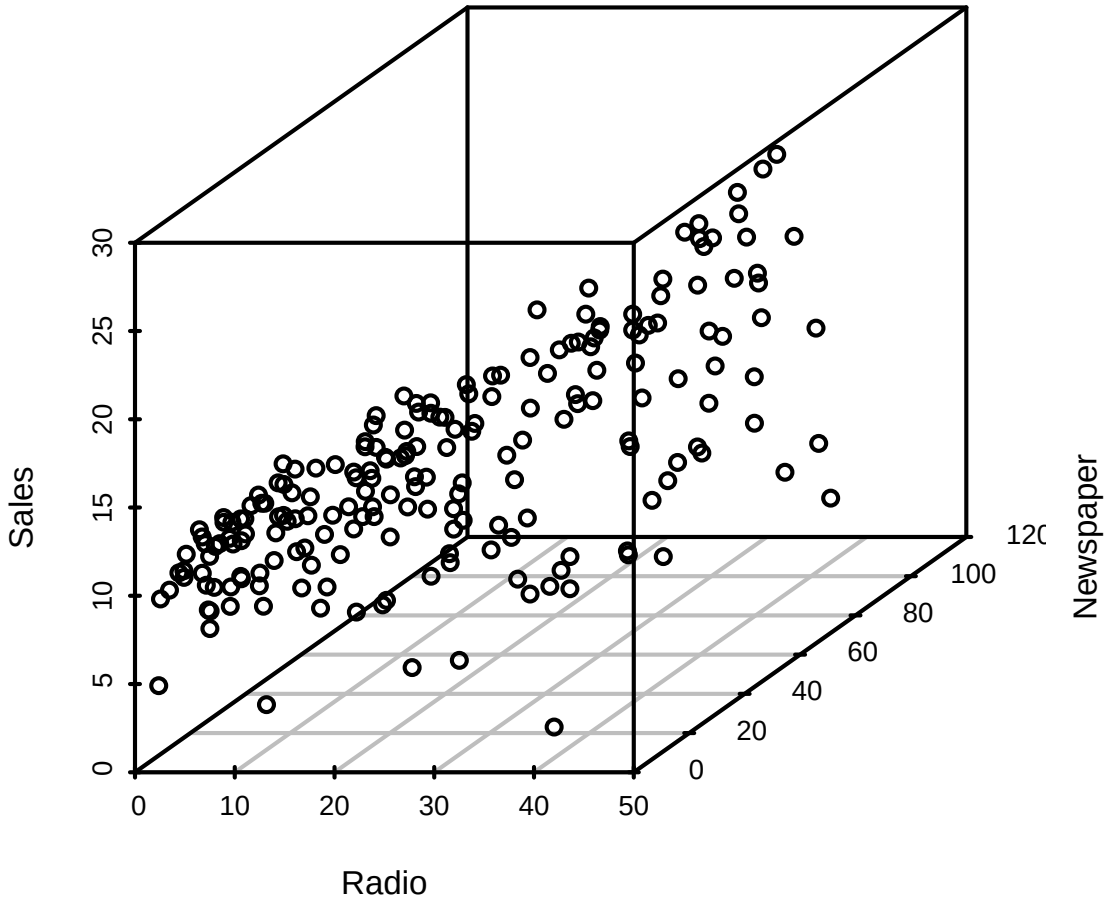
```
> Advertising <- read.csv("datasets/Advertising.csv")
> plot(Advertising$Radio, Advertising$Sales)
```



1. Observe a phenomenon
2. Construct a model for that phenomenon
3. Make predictions

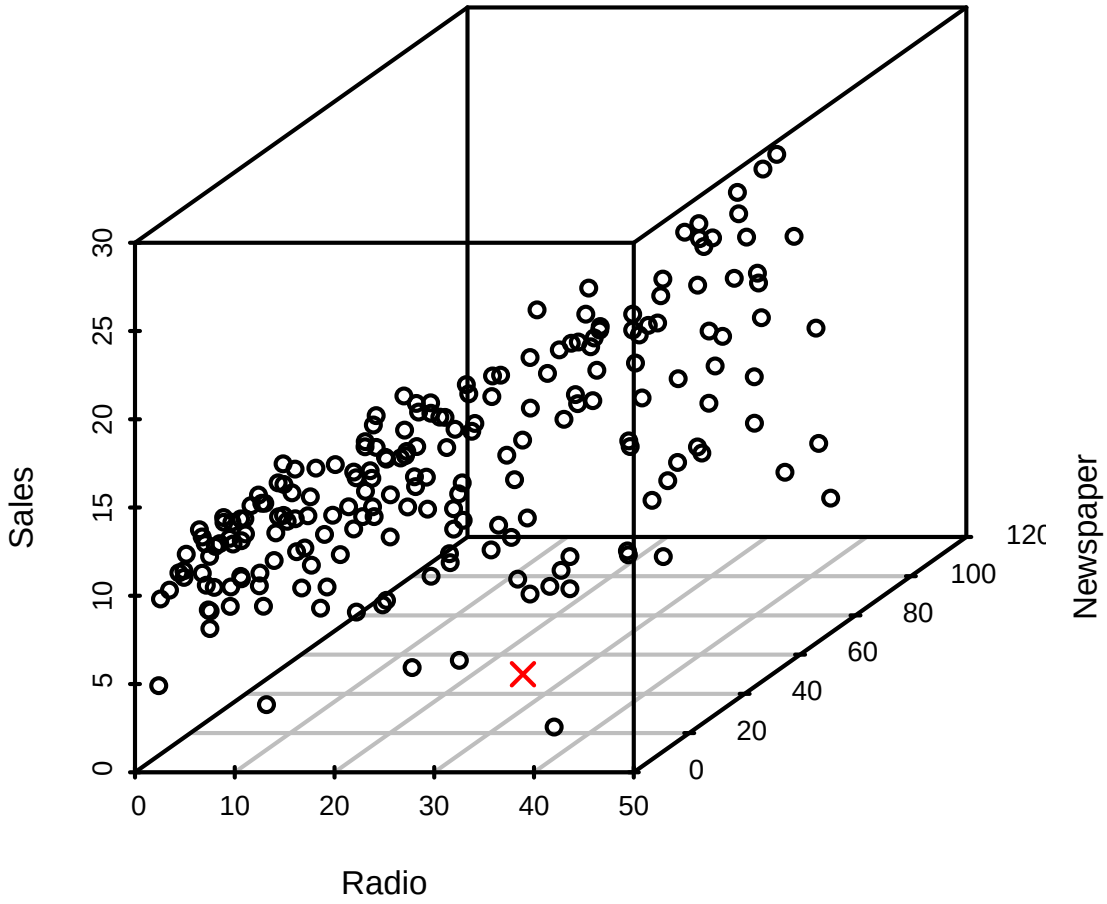
What is statistical learning?

```
> Advertising <- read.csv("datasets/Advertising.csv")
> library(rgl)
> plot3d(Advertising[, c("Radio", "Newspaper", "Sales")])
```



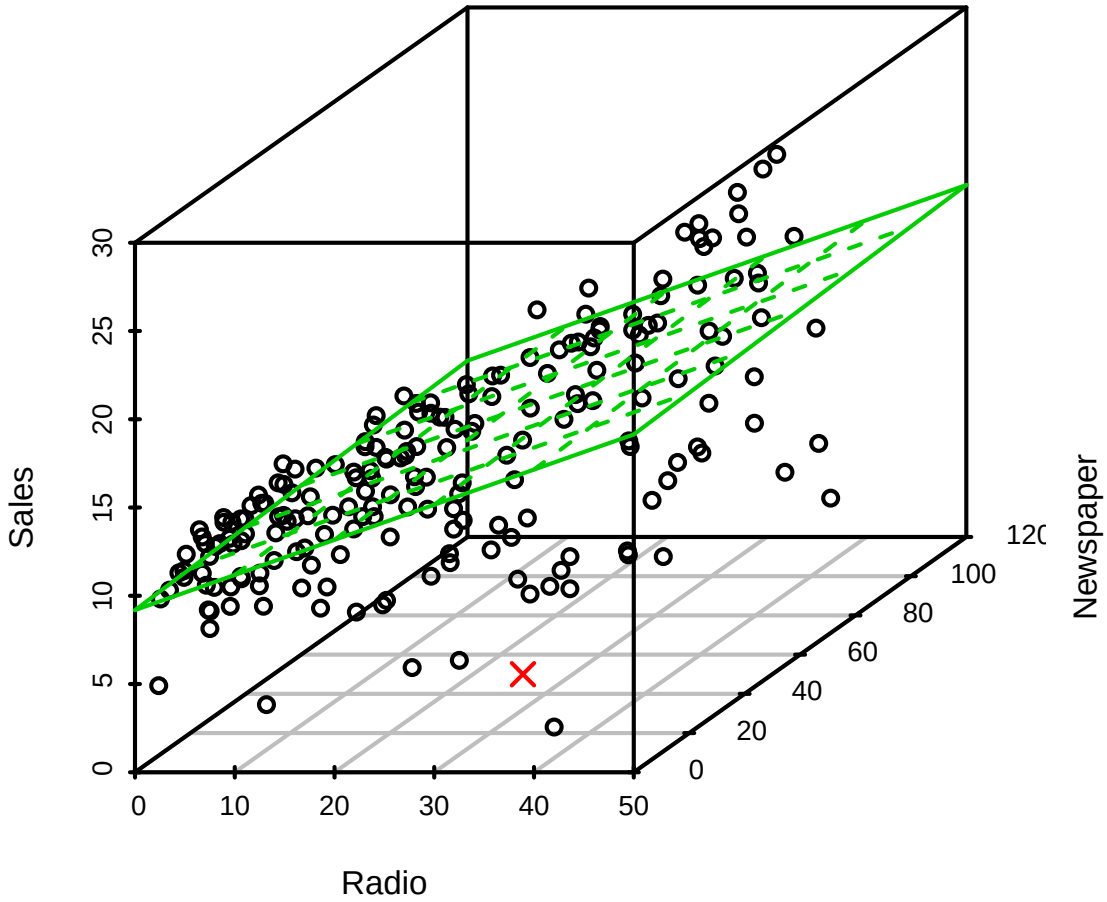
What is statistical learning?

```
> Advertising <- read.csv("datasets/Advertising.csv")
> library(rgl)
> plot3d(Advertising[, c("Radio", "Newspaper", "Sales")])
```



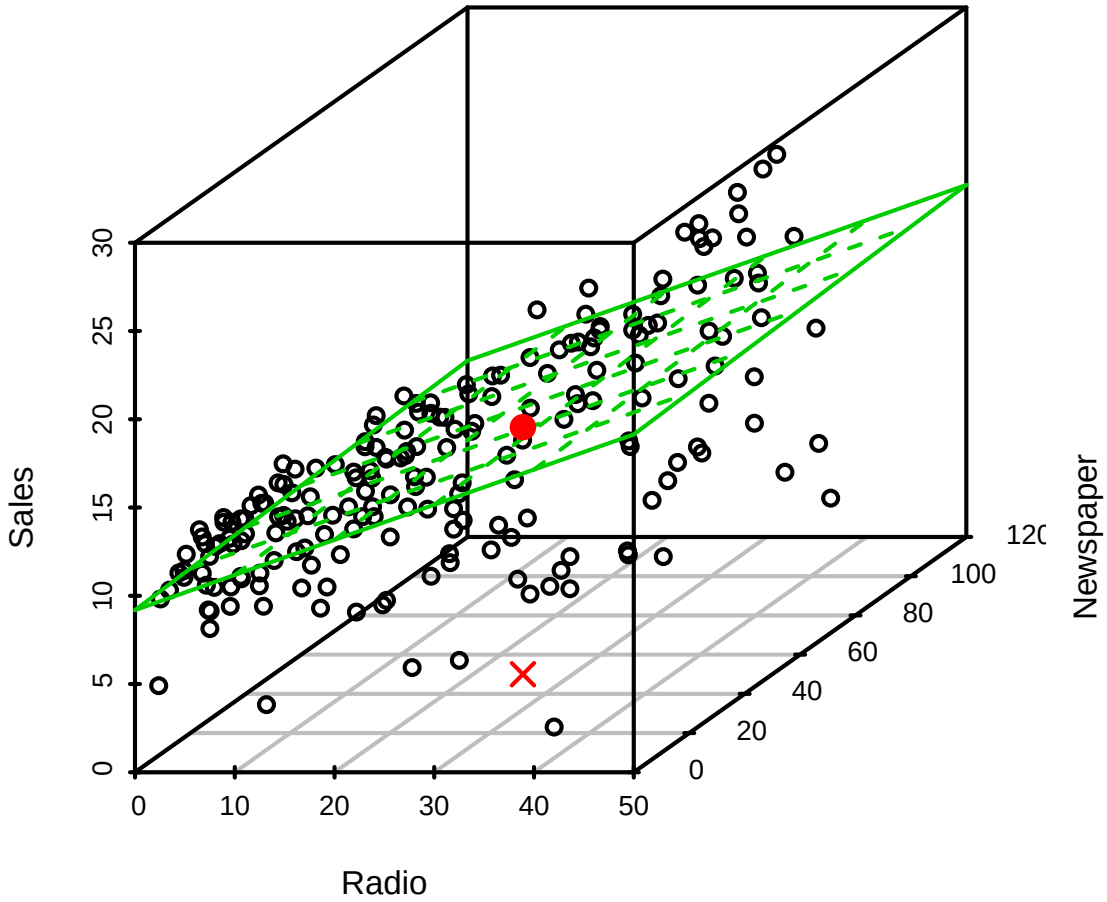
What is statistical learning?

```
> Advertising <- read.csv("datasets/Advertising.csv")
> library(rgl)
> plot3d(Advertising[, c("Radio", "Newspaper", "Sales")])
```



What is statistical learning?

```
> Advertising <- read.csv("datasets/Advertising.csv")
> library(rgl)
> plot3d(Advertising[, c("Radio", "Newspaper", "Sales")])
```



What is statistical learning?

Example 2: Suppose we work in the automotive sector and we want to know what will be the fuel consumption in MPG (miles per gallon) for a new automobile

Data set available in package ISLR

What is statistical learning?

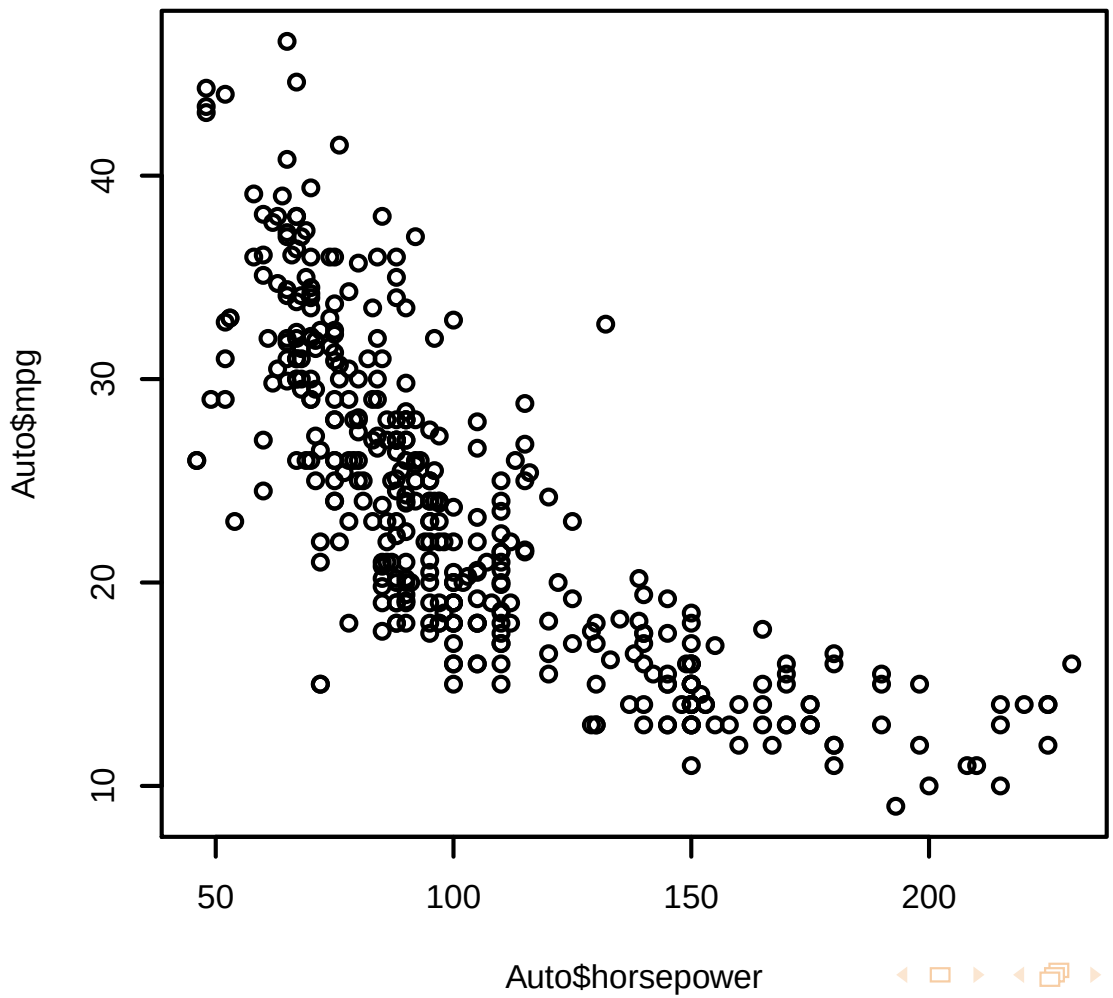
Example 2: Suppose we work in the automotive sector and we want to know what will be the fuel consumption in MPG (miles per gallon) for a new automobile

The Auto data set contains gas mileage, horsepower, and other information for cars.

Data set available in package ISLR

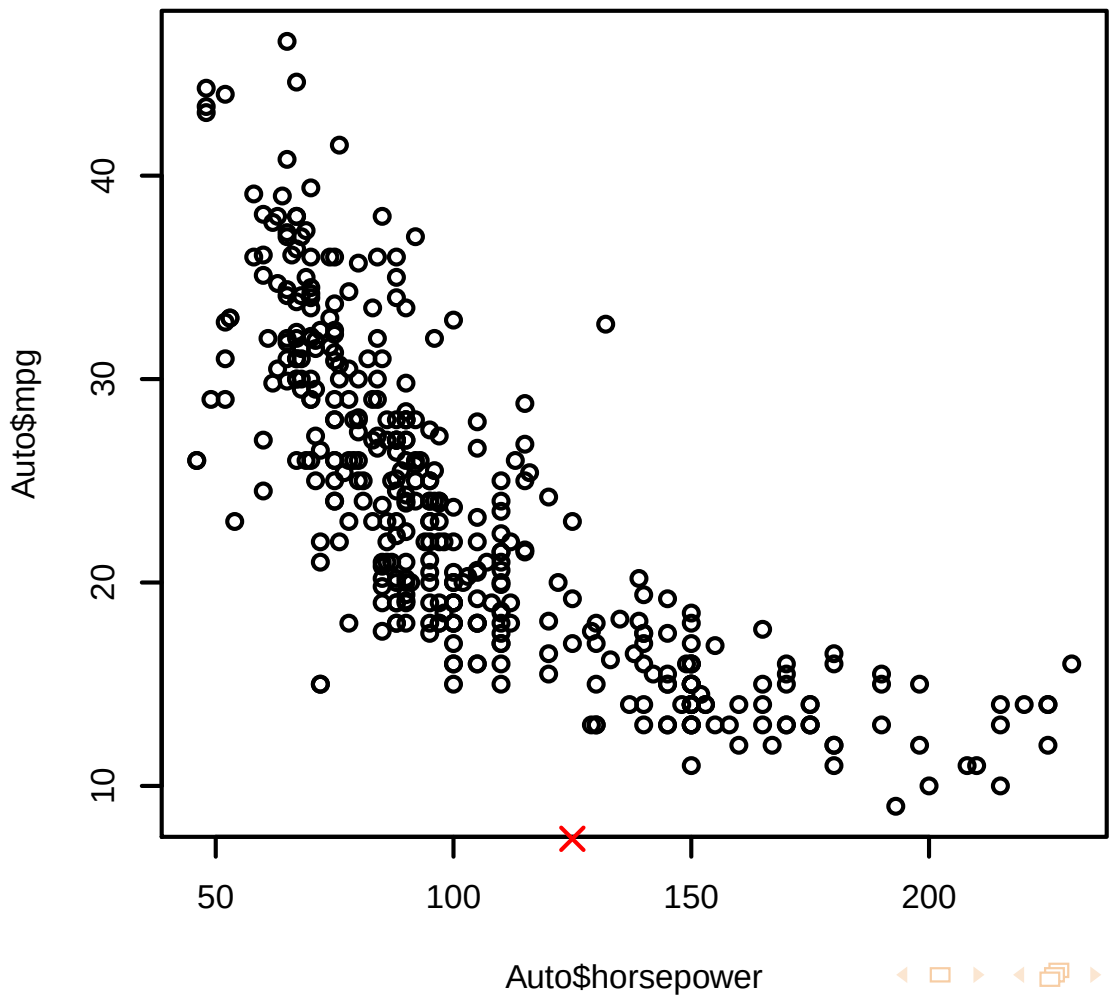
What is statistical learning?

```
> library(ISLR)
> data(Auto)
> plot(Auto$horsepower, Auto$mpg)
```



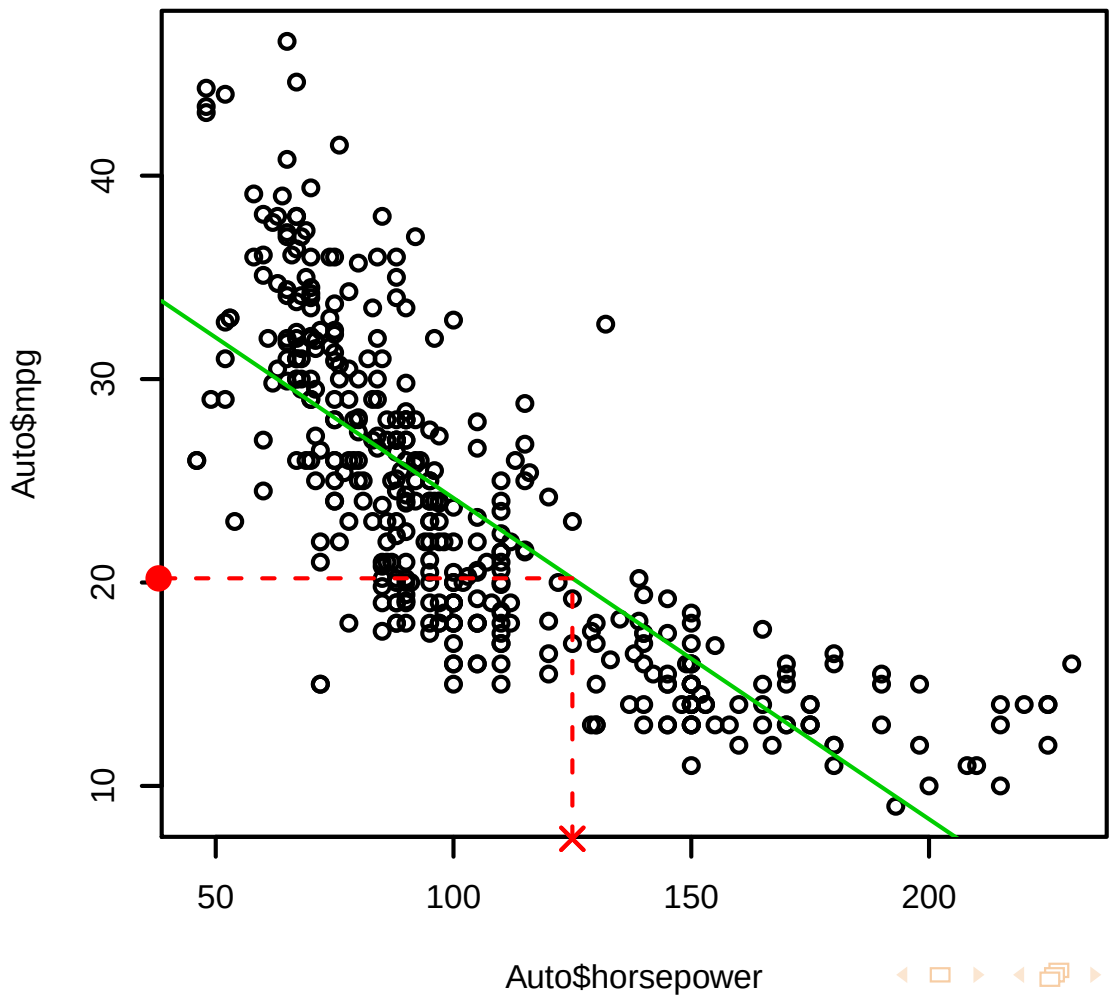
What is statistical learning?

```
> library(ISLR)
> data(Auto)
> plot(Auto$horsepower, Auto$mpg)
```



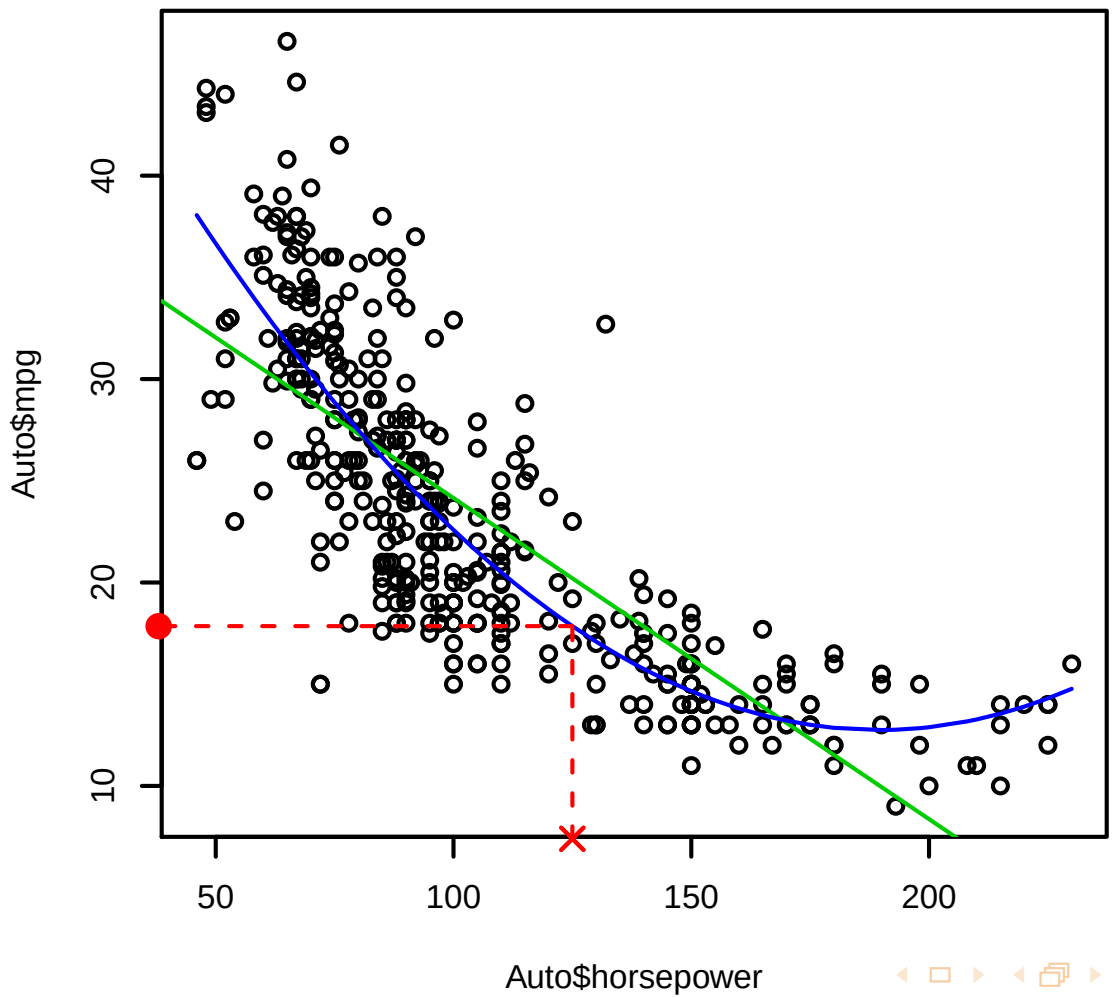
What is statistical learning?

```
> library(ISLR)
> data(Auto)
> plot(Auto$horsepower, Auto$mpg)
```



What is statistical learning?

```
> library(ISLR)
> data(Auto)
> plot(Auto$horsepower, Auto$mpg)
```



What is statistical learning?

Most statistical learning problems fall into one of two categories:

- supervised learning
- unsupervised learning

What is statistical learning?

- Examples 1 and 2 are examples of supervised learning.
- For each observation of the predictor measurement x_i there is an associated response measurement y_i, for $i = 1, \ldots, n$. (“right answers”)

What is statistical learning?

- Outcome measurement Y (output, response, dependent variable)
- Vector of p predictor measurements X (inputs, regressors, independent variables, covariates)
- We have training data $(x_1, y_1), \ldots, (x_n, y_n)$. These are observations (examples, instances) of these measurements.

What is statistical learning?

Objective: We wish to fit a model that relates the response to the predictors, with the aim of:

- prediction: in many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained.
  - We seek to predict Y using X.
- inference: we are often interested in understanding the way that Y is affected as predictor measurements X change.
  - Which predictors are associated with the response?
  - What is the relationship between the response and each predictor?

What is statistical learning?

Example 3: Suppose we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of credit card balance

Data set available in package ISLR

What is statistical learning?

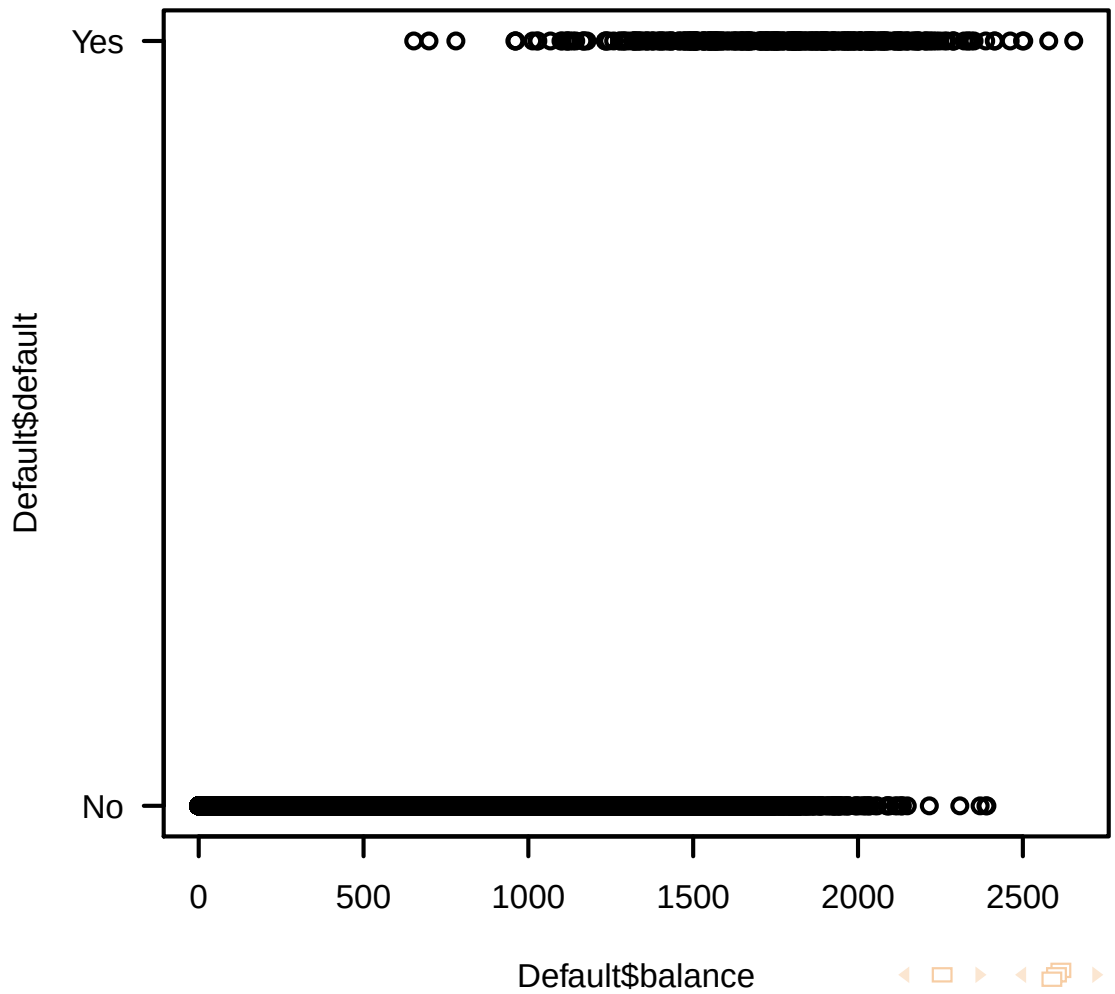
Example 3: Suppose we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of credit card balance

The Default data set contains customer default records for a credit card company.

Data set available in package ISLR

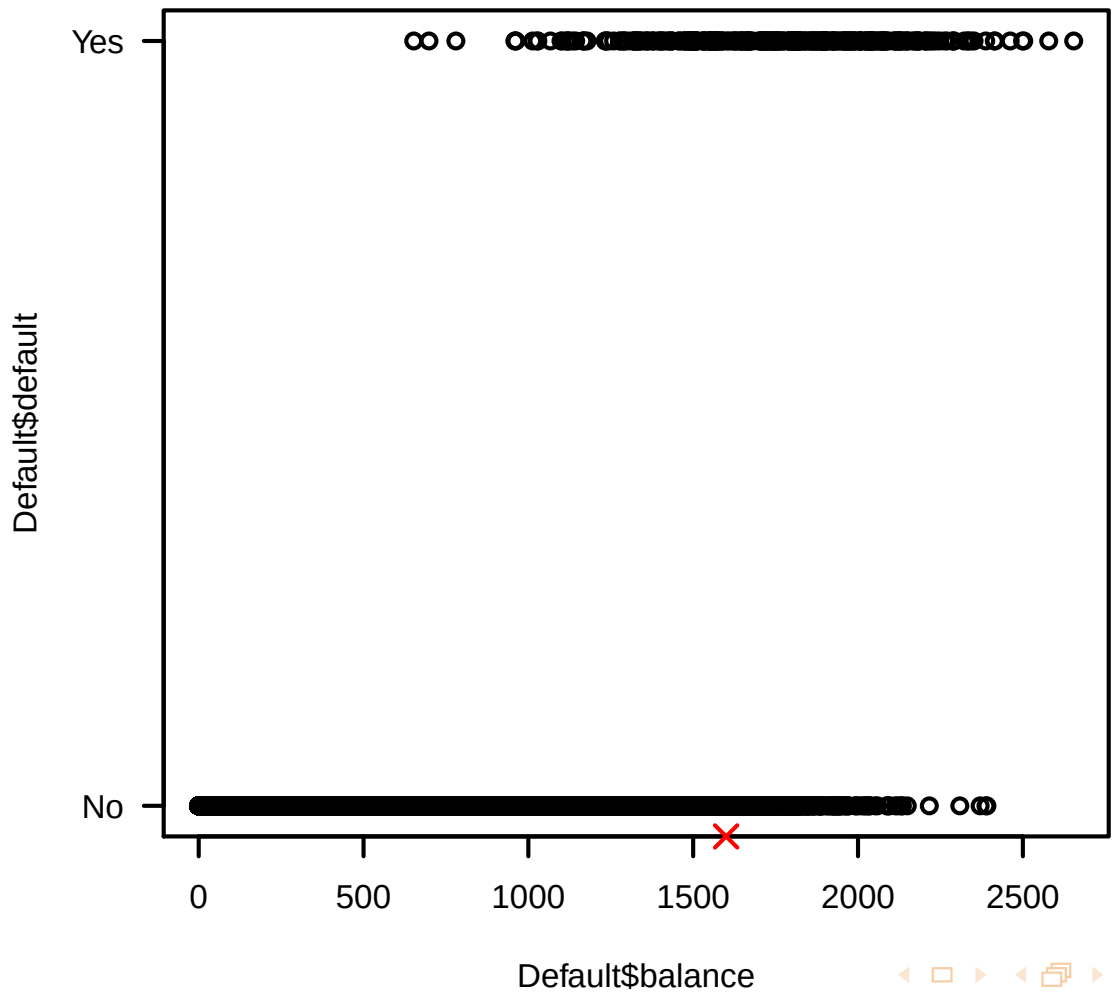
What is statistical learning?

```
> library(ISLR)
> data(Default)
> plot(Default$balance, Default$default)
```



What is statistical learning?

```
> library(ISLR)
> data(Default)
> plot(Default$balance, Default$default)
```



What is statistical learning?

- The example of prediction of default is another example of supervised learning.
- For each observation of the predictor measurement x_i (balance) there is an associated response measurement y_i (default yes or no), for $i = 1, \ldots, n$. (“right answers”)

What is statistical learning?

Variables can be characterized as either **quantitative** or **qualitative**

- Regression problems: problems with a quantitative response.
- Clasiffication problems: problems with a qualitative response.

What is statistical learning?

Example 4: Suppose we collect measurements on the petal length and width of iris specimens

Data set available in package datasets

What is statistical learning?

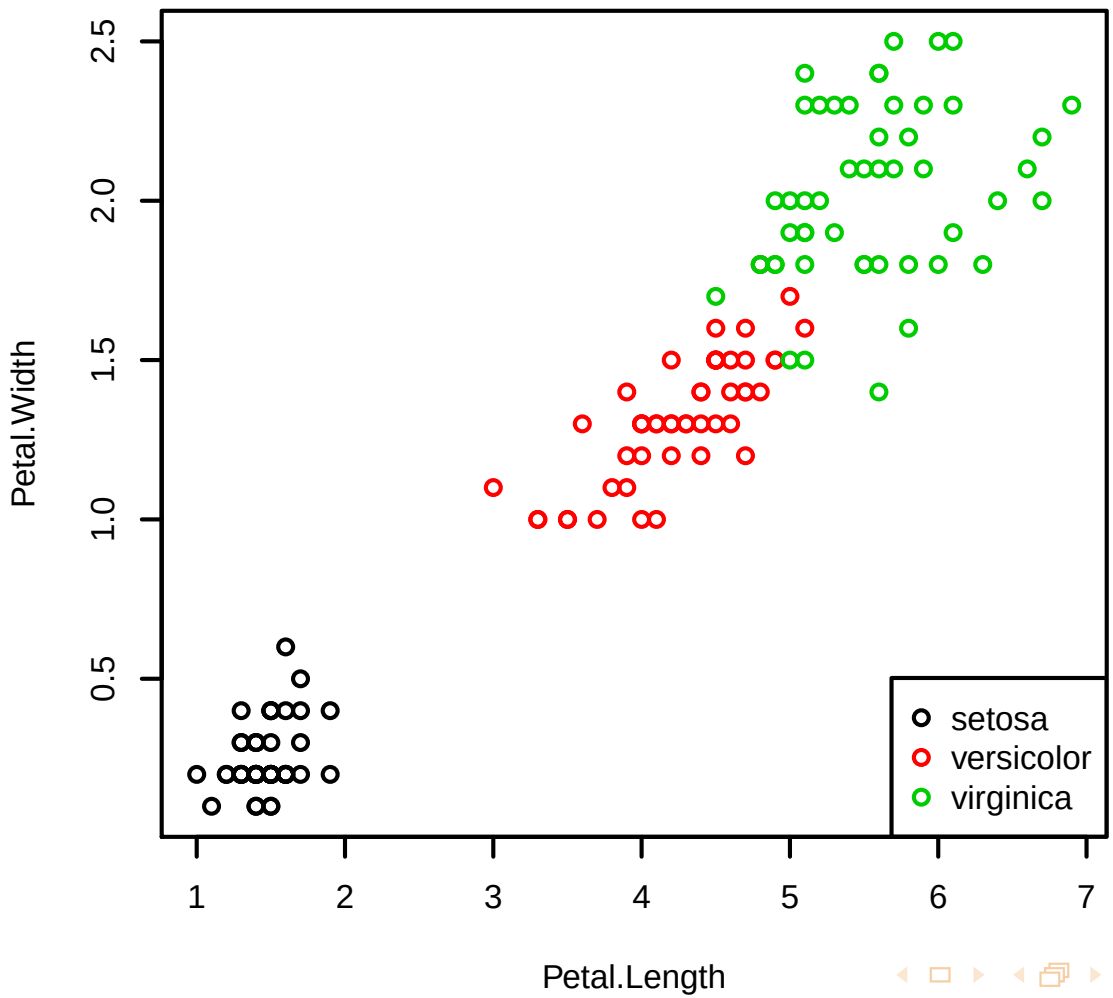
Example 4: Suppose we collect measurements on the petal length and width of iris specimens

The iris data set contains measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris.

Data set available in package datasets

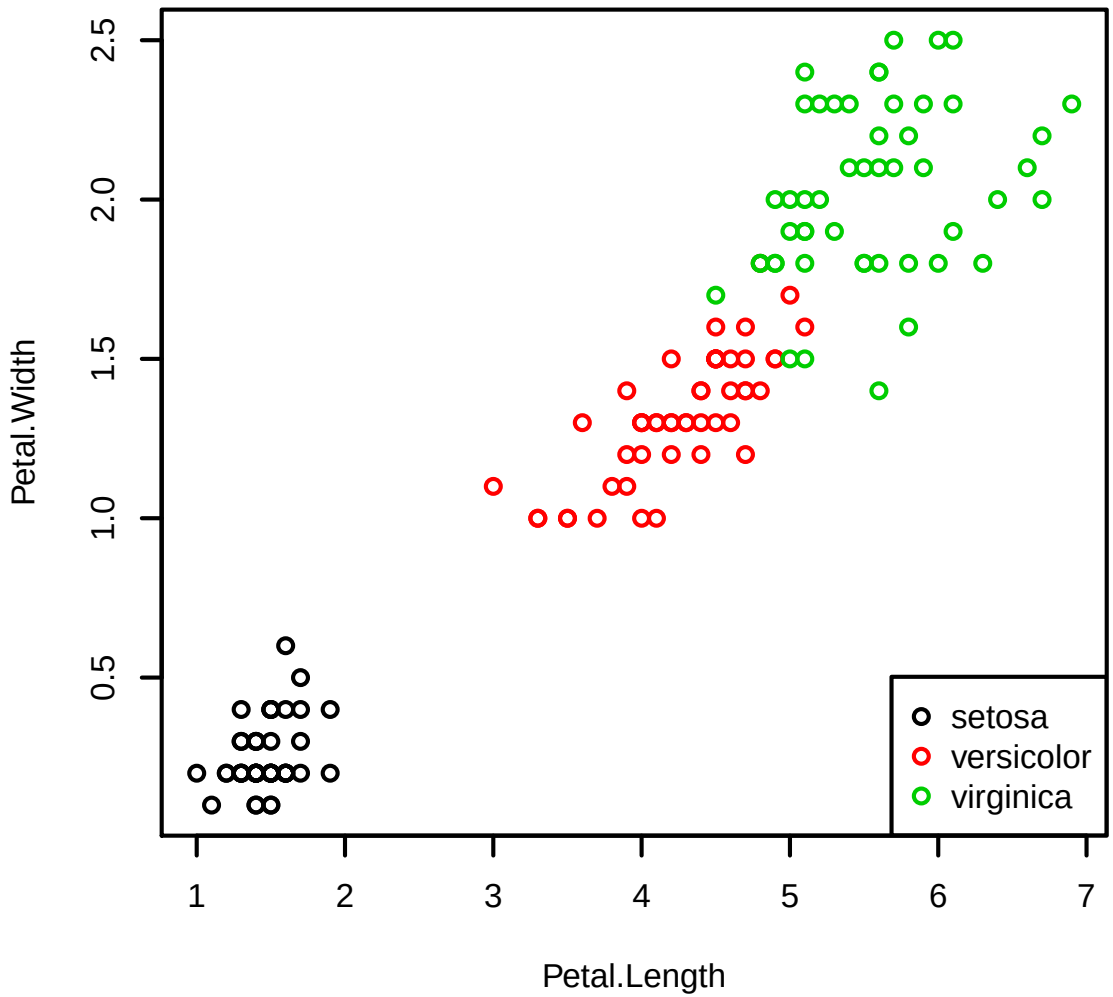
What is statistical learning?

```
> data(iris)
> plot(iris[, 3:4], col = iris$Species)
> legend("bottomright", levels(iris$Species), col = 1:3, pch = 1)
```



What is statistical learning?

Supervised learning: guess the class of an individual flower given the measurements of petals



What is statistical learning?

Unsupervised learning: we are interested in finding groupings among iris onbservations, based on their petal measurements