# DATA SCIENCE

# WHAT IS MACHINE LEARNING?

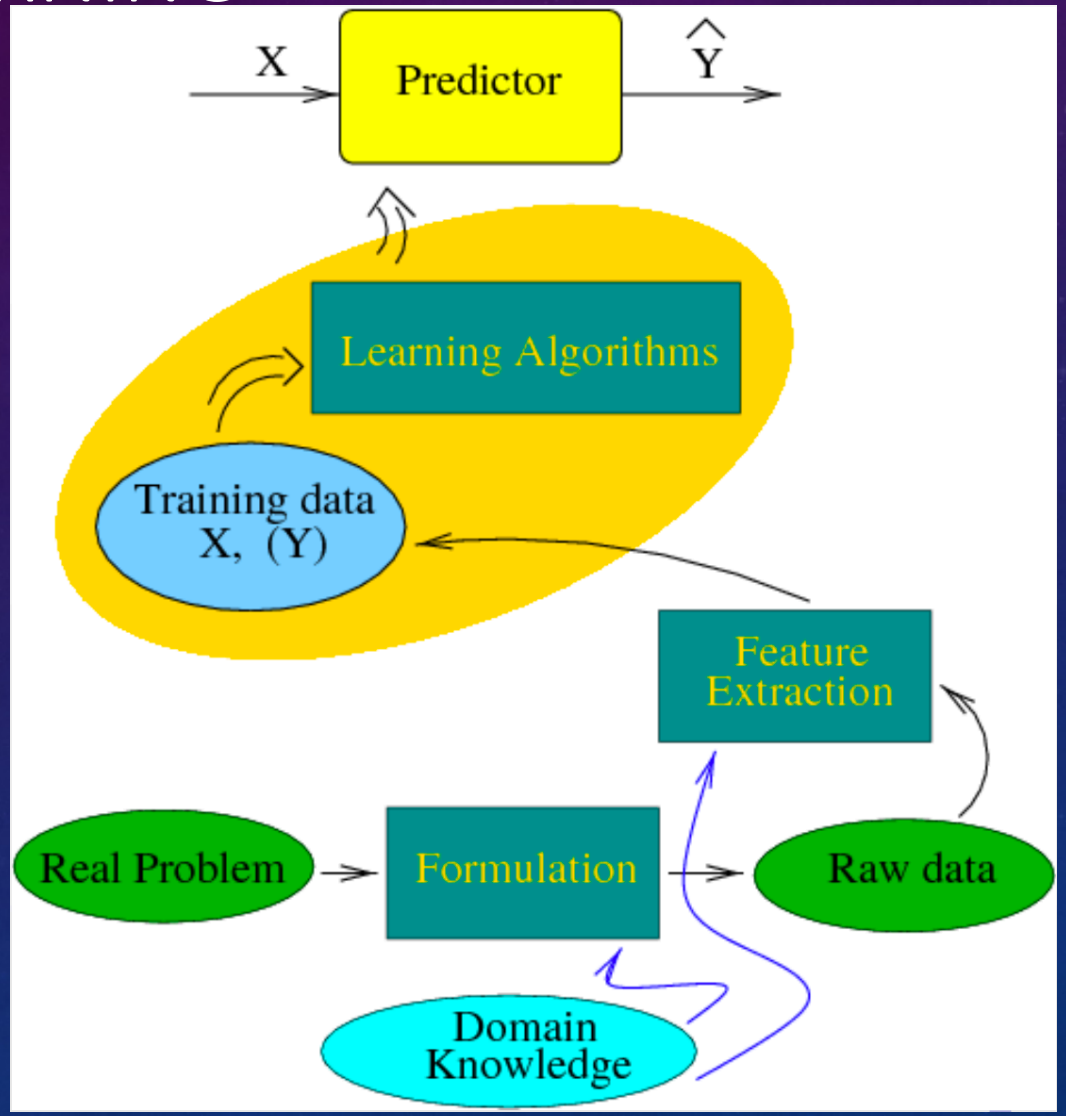
A QUICK OVERVIEW OF WHAT'S TO COME IN THIS SUMMER DATA SCIENCE PROGRAM

ADRIANO ZAMBOM

# WHAT IS MACHINE LEARNING?

- Machine learning is a set of statistical and computer science techniques that use data to recognize patterns and predict future outcomes.
- These predictions are based on models from sample inputs (data).
- This is closely related to artificial intelligence (AI)

# MACHINE LEARNING



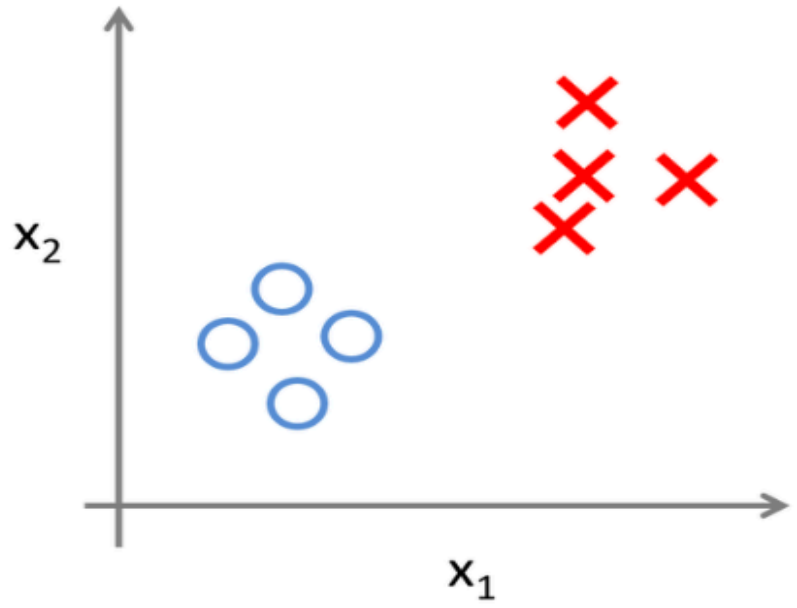
# MACHINE LEARNING

- Let's focus on two aspects: Supervised vs Unsupervised Learning

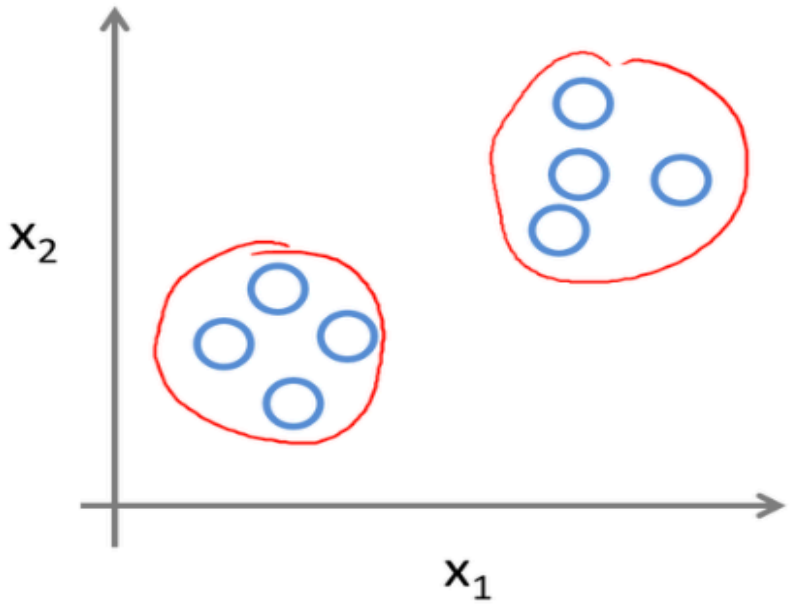
- Supervised Learning: Besides the data on each person (object) we also know an outcome
  - Can we use the data to classify according to the outcome or predict the outcome?

- Unsupervised Learning: Want to learn about the structure of the data
  - Can people be clustered according to their characteristics?

# SUPERVISED LEARNING




# UNSUPERVISED LEARNING




What does the data look like?

| X1 | X2 | Y |
|---|---|---|
| 2 | 4 | 1 |
| 2.8 | 5 | 1 |
| 3.2 | 2.2 | 1 |
| 4 | 4 | 1 |
| 6.8 | 6 | 2 |
| 7.1 | 6.7 | 2 |
| 7.1 | 7 | 2 |
| 9 | 6.5 | 2 |

| X1 | X2 |
|---|---|
| 2 | 4 |
| 2.8 | 5 |
| 3.2 | 2.2 |
| 4 | 4 |
| 6.8 | 6 |
| 7.1 | 6.7 |
| 7.1 | 7 |
| 9 | 6.5 |

# SUPERVISED LEARNING

- Some techniques used:

- logistic regression,
- naive Bayes,
- discriminant analysis
- support vector machines,
- artificial neural networks,
- random forests
- etc

# SUPERVISED LEARNING

- Example: Email spam:

- Goal: predict whether an email is a junk email, i.e., “spam”.
- Raw data: text email messages.
- Input X: relative frequencies of 50 of the most commonly occurring words and punctuation marks in the email message.
- Training data set: 4601 email messages with email type known (supervised learning).

| Email id | Email Type | Rel. Freq: “is” | Rel. Freq: “need” | Rel. Freq: “have” | Rel. Freq: “.” | … |
|---|---|---|---|---|---|---|
| Email 1 | Spam | 10/234 | 10/234 | 19/234 | 20/234 | … |
| Email 2 | Not Spam | 22/420 | 1/420 | 14/420 | 21/420 | … |
| Email 3 | Not Spam | 3/98 | 0/98 | 4/98 | 5/98 | … |
| Email 4 | Spam | 5/121 | 4/121 | 5/121 | 13/121 | … |
| Email 5 | Not Spam | 8/206 | 2/206 | 6/206 | 13/206 | … |

# SUPERVISED LEARNING

- Example: Handwritten digit recognition

- Goal: identify single digits 0 ~ 9 based on images.
- Raw data: images that are scaled segments from five digit ZIP codes.
  - 16 × 16 eight-bit grayscale maps
  - Pixel intensities range from 0 (black) to 255 (white).
- Input data/Training Dataset: Images transformed into 256 dimension vectors, or feature vectors with lower dimensions. We know in our data what number the image corresponds to.

- Challenge: New image comes in, what is the number?

# SUPERVISED LEARNING

- Example: Handwritten digit recognition

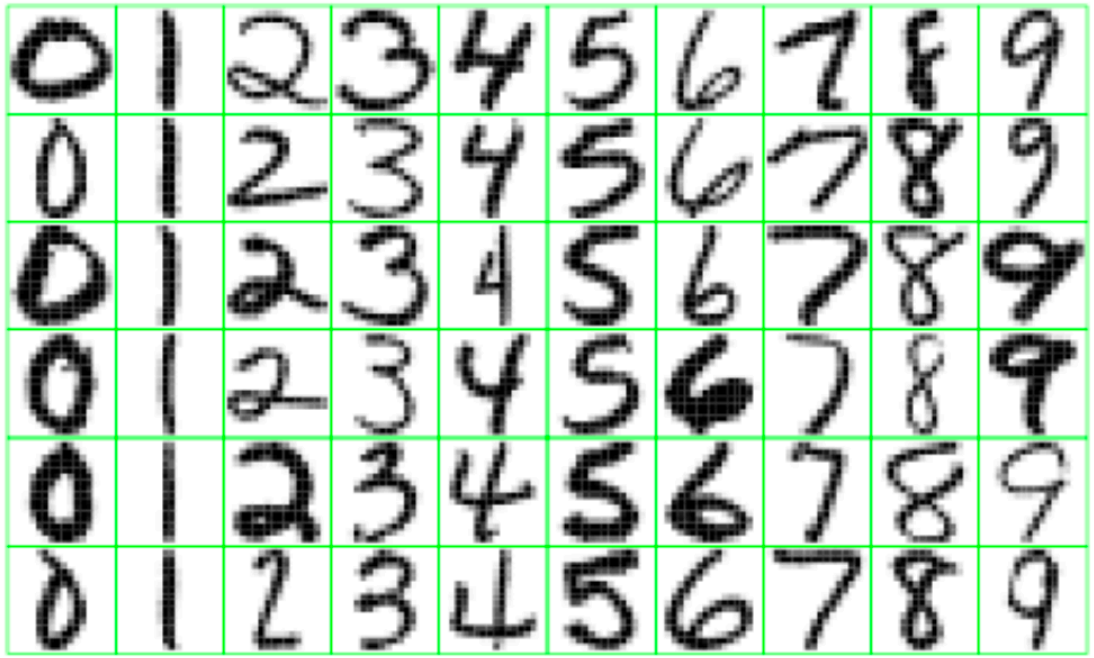


Figure 1.2: *Examples of handwritten digits from U.S. postal envelopes.*

# SUPERVISED LEARNING

- Example: Speech recognition:

- Goal: identify words spoken according to speech signals
- Automatic voice recognition systems used by airline companies
- Automatic stock price reporting
- Raw data: voice recordings from text reading (we know the text).

- Challenge: A new recording comes in, can we identify what he/she said?

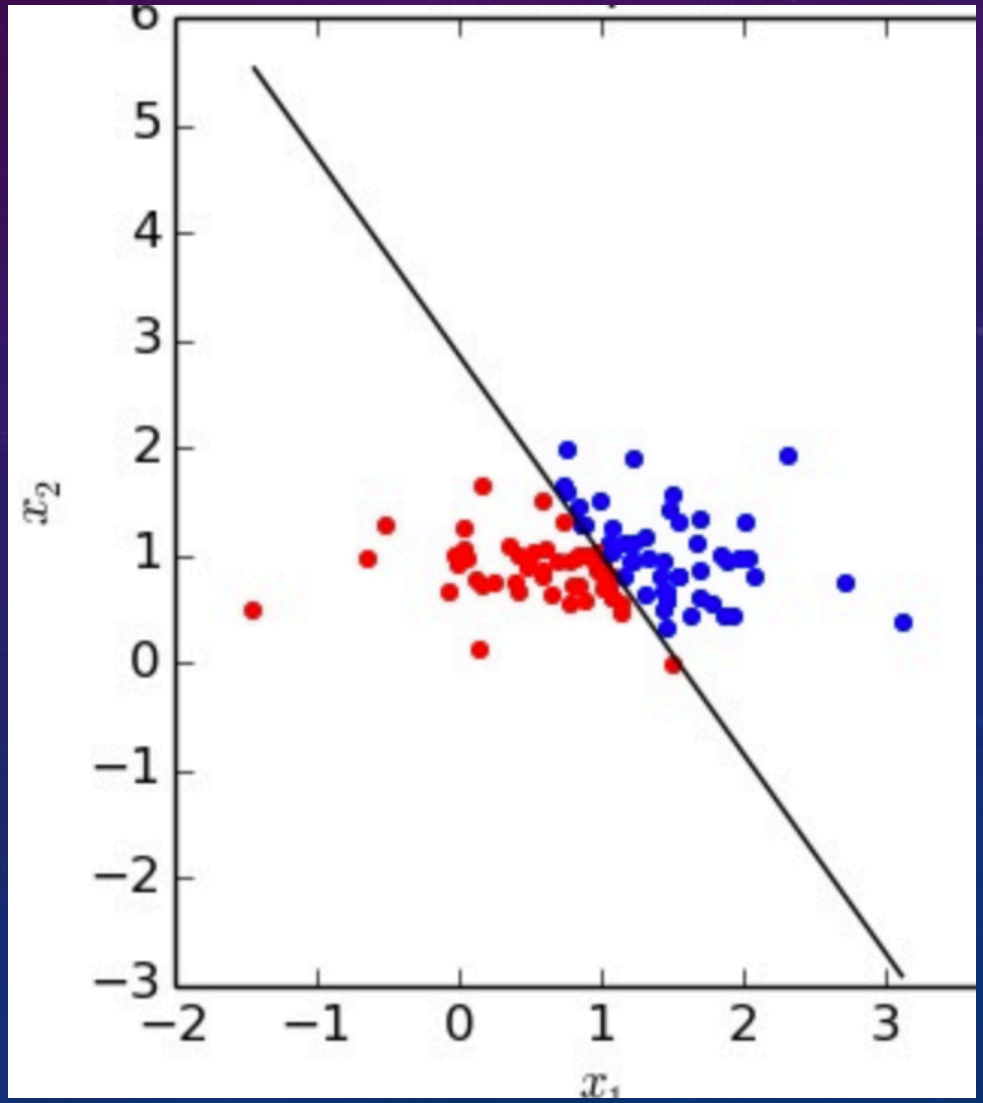
# SUPERVISED LEARNING

- Example: DNA Expression Microarray:

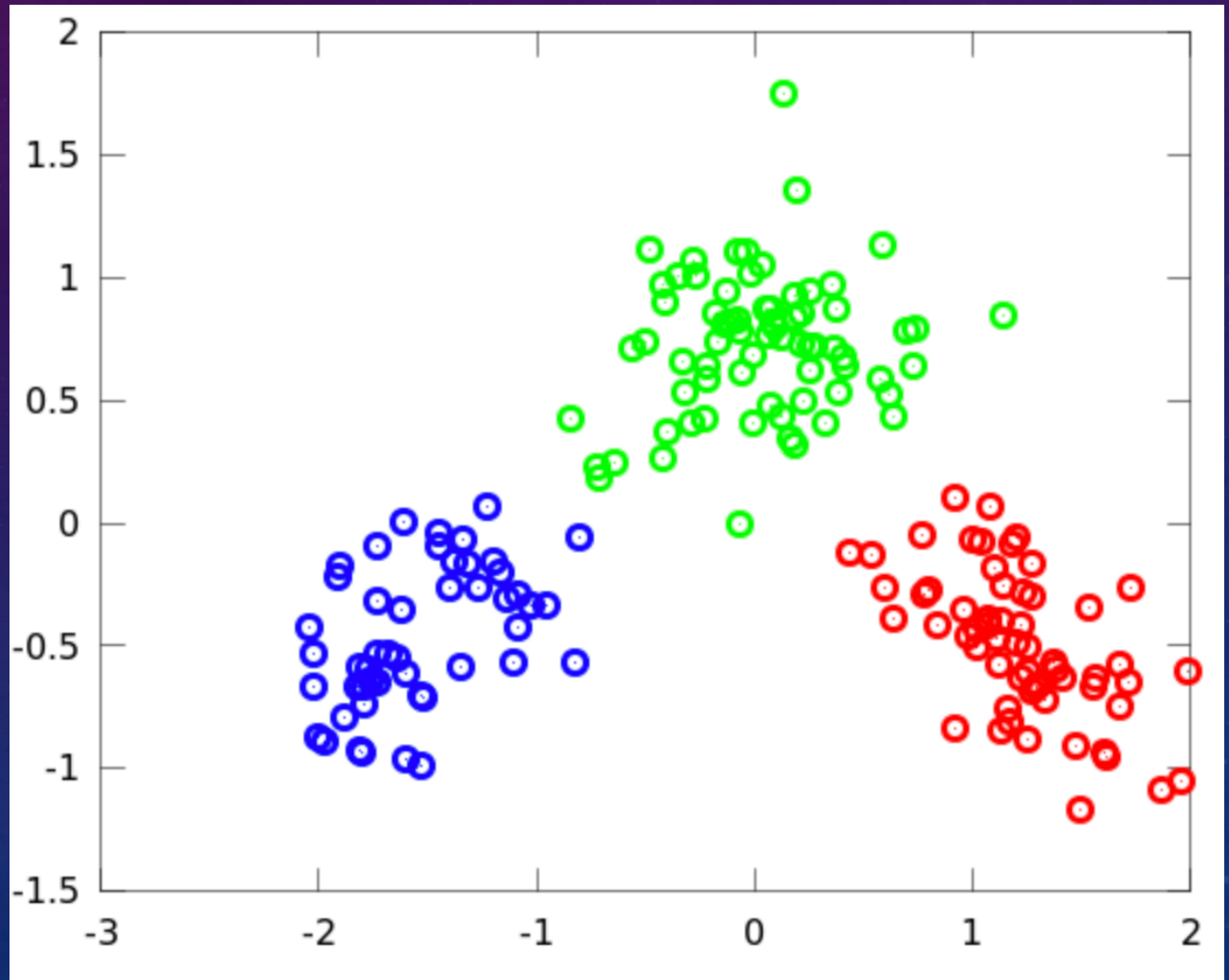
- Goal: identify disease or tissue types
- Raw data: for each sample taken from a tissue of a particular disease type, the expression levels of a large collection of genes are measured.
- Input data: cleaned-up gene expression data
- Example data set: 4026 genes, 96 samples taken from 9 classes of tissues ( we know the classes)

- Challenge: Given a person's gene expression, can we say what type of tissue he/she has (cancer/no cancer)?

# SUPERVISED LEARNING



# SUPERVISED LEARNING

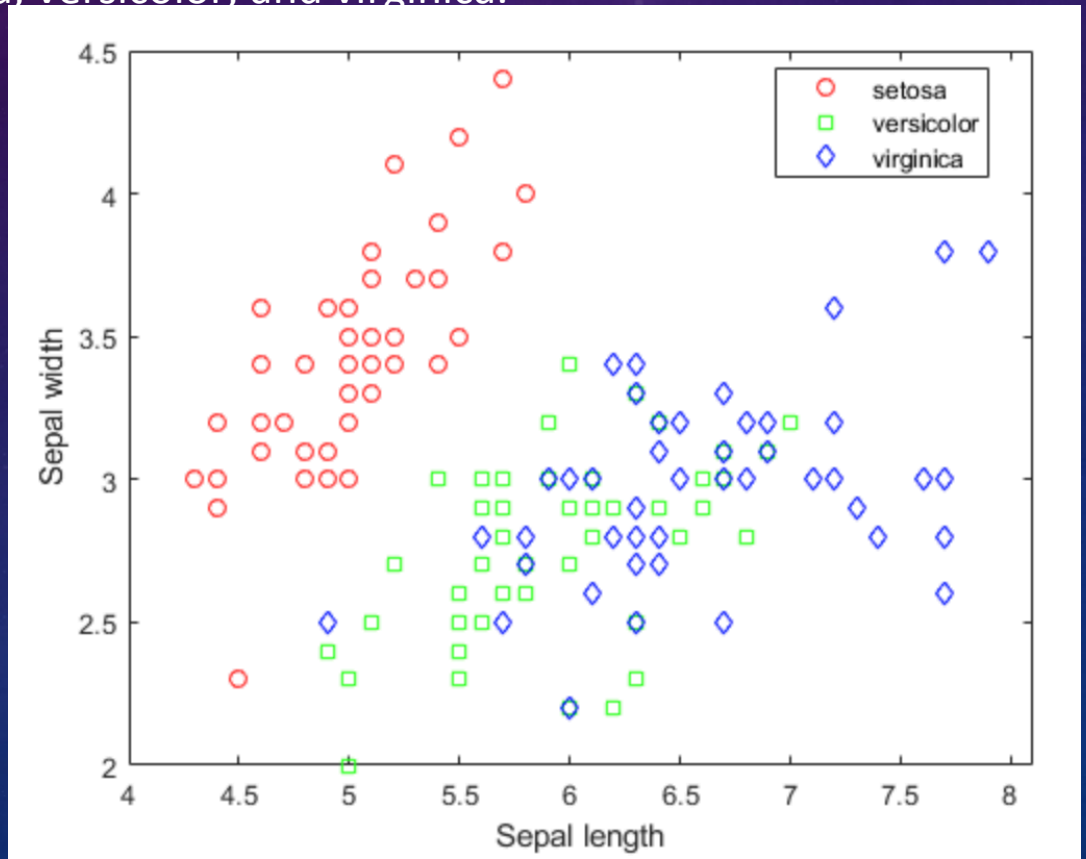


# SUPERVISED LEARNING

This famous (Fisher's or Anderson's) iris data set gives the 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.

The species are Iris setosa, versicolor, and virginica.

Let’s look at this Dataset in R



# SUPERVISED LEARNING

- **Regression:** Your outcome is a continuous variable

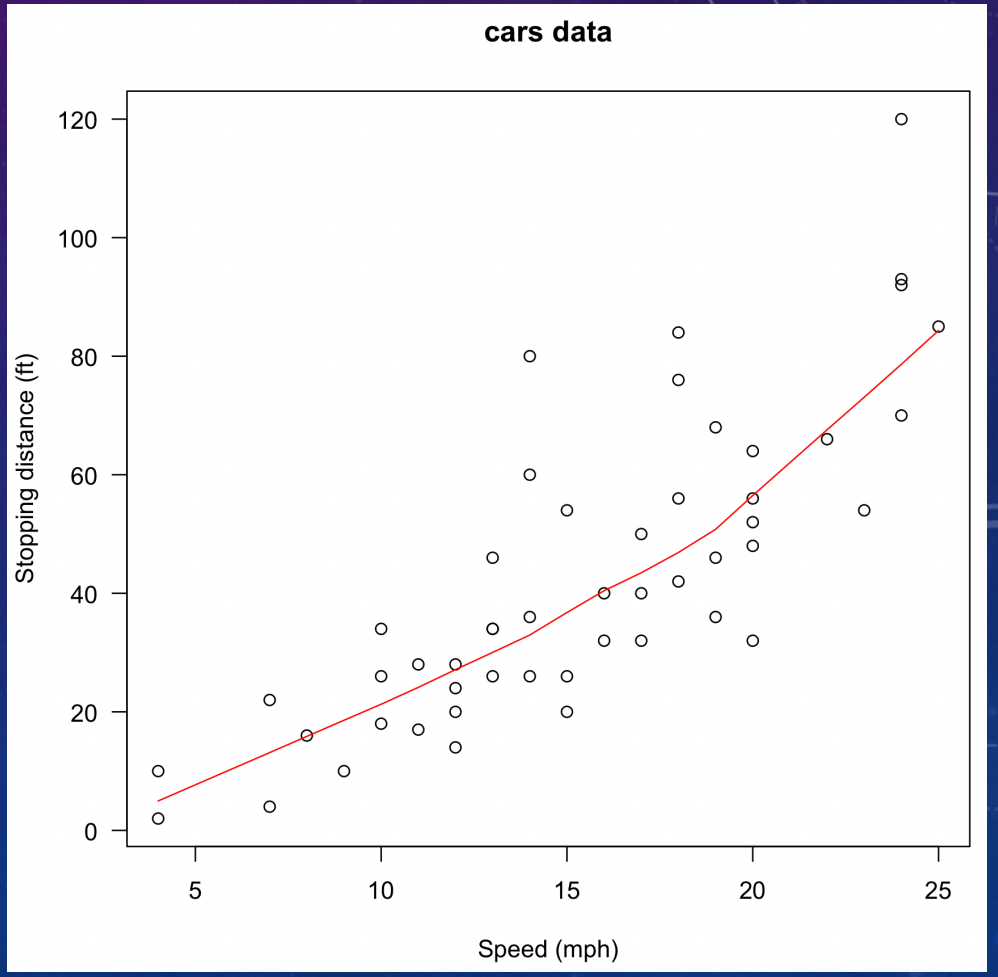
- We have the information X, and the outcome Y.
- For example the outcome Y may be: cholesterol level, height, size, distance, yield, etc

# SUPERVISED LEARNING

- **<u>Regression Example</u>: Speed and Stopping Distances of Cars**

- The data (in R) give the speed of cars
- and the distances taken to stop.
- Note that the data were recorded in the 1920s.

- Let's take a look at this data in R.



# UNSUPERVISED LEARNING

- Usually we want to:

- Cluster or group people or items
- Visually explore their characteristics
- Estimate densities

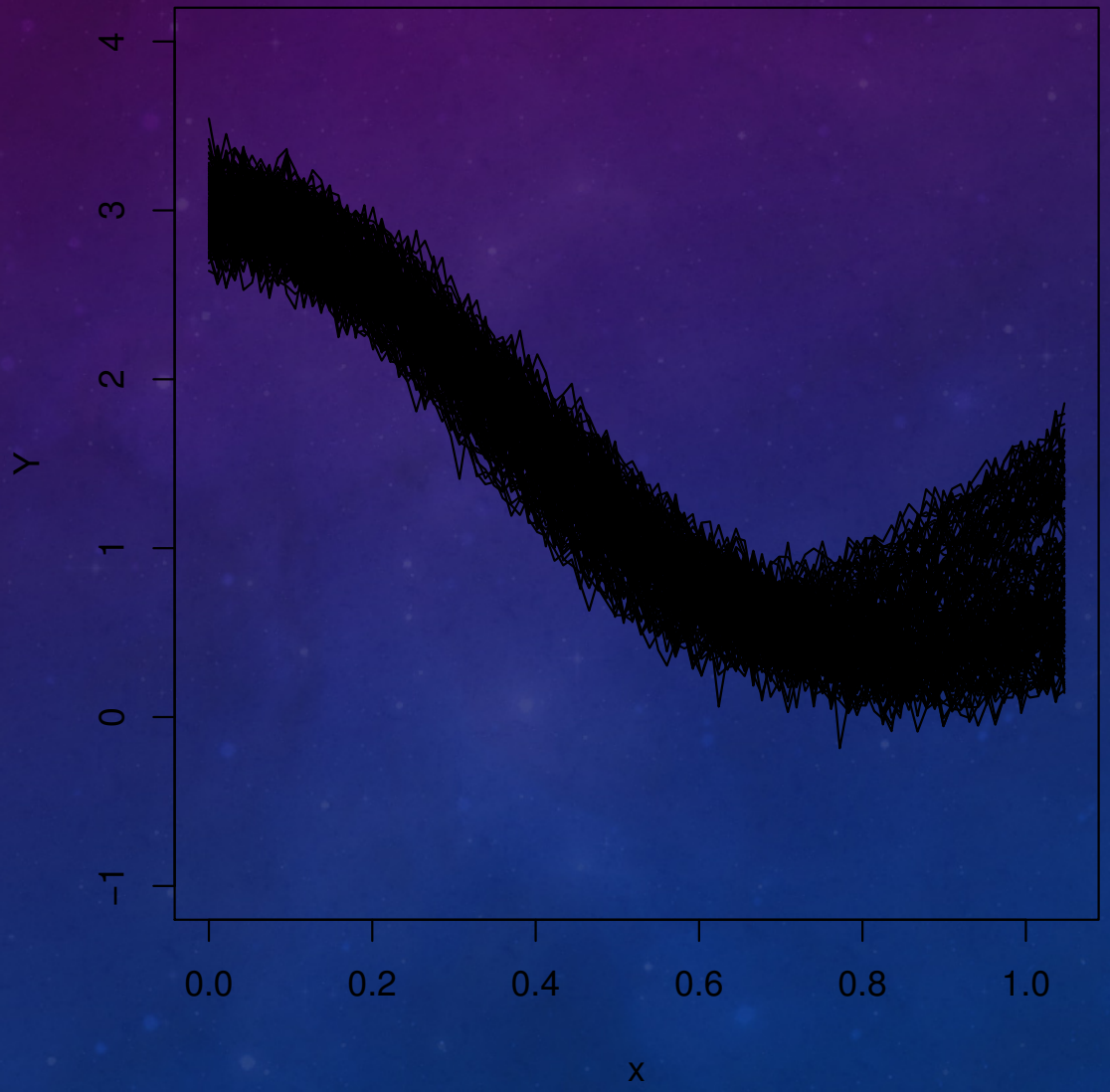
# UNSUPERVISED LEARNING

- Common techniques:

- Dimension reduction
- Principal Component Analysis (PCA)
- Factor Analysis
- Clustering
  - K-means algorithm
  - Hierarchical clustering,
  - Gaussian mixture models,
  - Hidden Markov Models

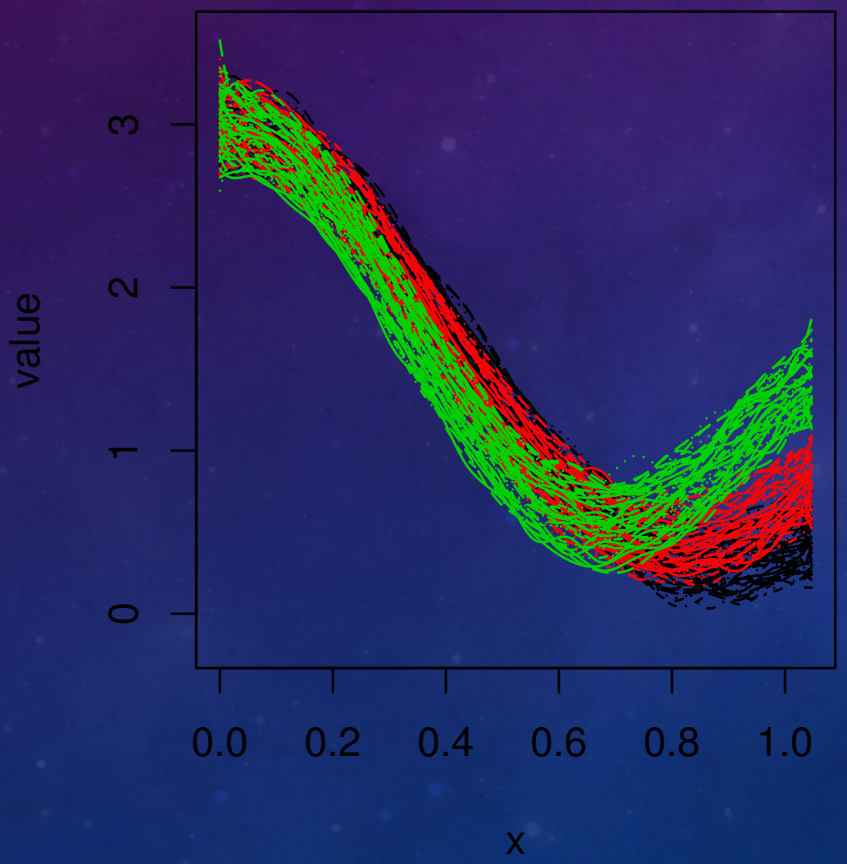
# UNSUPERVISED LEARNING

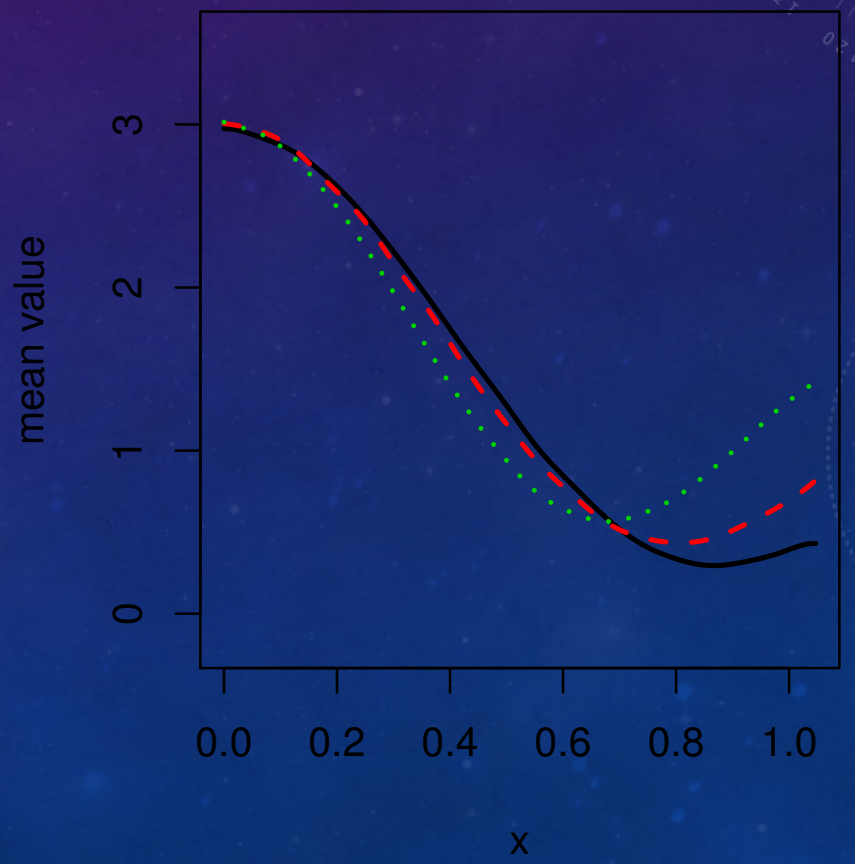
Example: Cluster curves



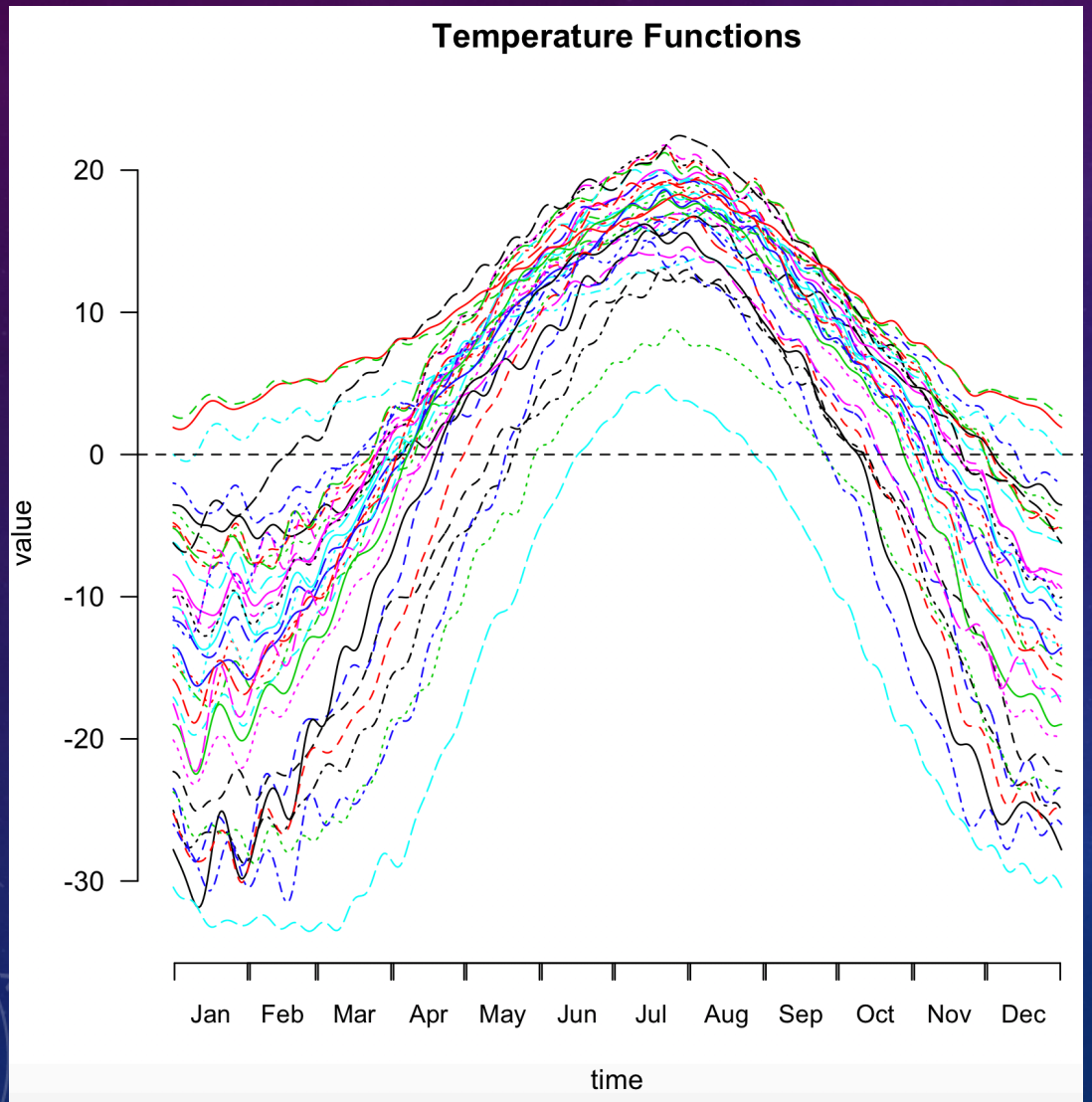
# UNSUPERVISED LEARNING

Example: Cluster curves



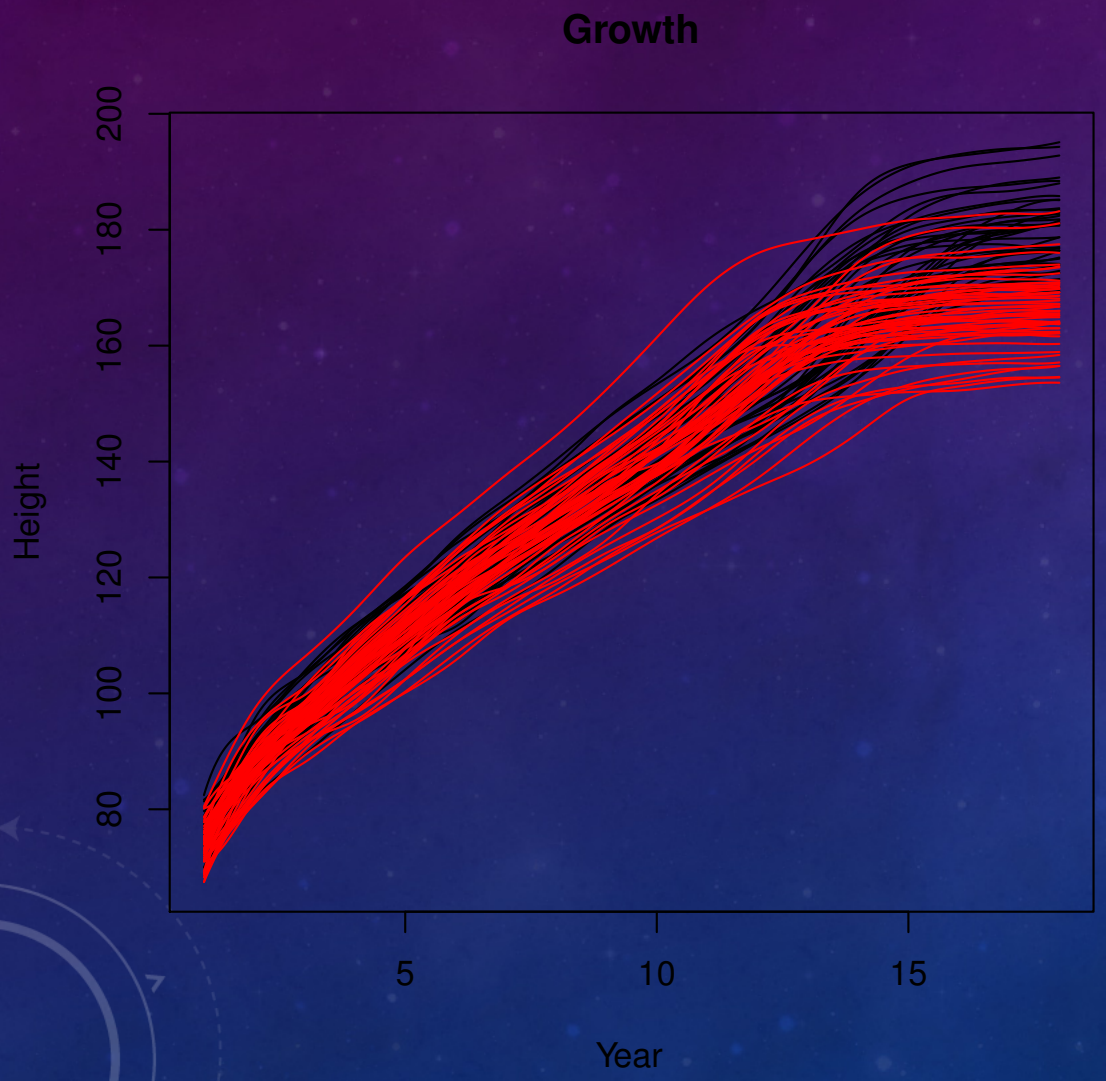



# UNSUPERVISED LEARNING




Let’s take a look at package fda in R

# UNSUPERVISED LEARNING




A list containing the heights of 39 boys and 54 girls from age 1 to 18 and the ages at which they were collected.

# SOME INTERESTING VIDEOS

- https://www.youtube.com/watch?v=qDbpYUbf3e0
- https://www.youtube.com/watch?v=hfO6iRj-GZo

# HOMEWORK

- 1. Take this quiz: link – This is just to see how much of data science you know – Don't worry if you don't know the answers.

- 2. Readings:
- A) Read this carefully: Data Science and Statistics
- B) Read about the questions that come up in interviews: here
- C) Take a look at this Berkeley description of Data Science: here

- 3. Explore Kaggle
- https://www.kaggle.com/
- If you have never visited this website, you should definitely spend some time there.
- Choose a data set that you like. We may use it, analyze it during the next few of weeks