# LECTURE 1 INTRODUCTION TO ANALYTICS

LEK HSIANG HUI

# OUTLINE

Introduction to Analytics

Supervised Learning

Unsupervised Learning

Applications of Analytics

# INTRODUCTION TO ANALYTICS

# DATA

Companies are generating a LOT of data
e.g. sales, transaction, customer data, etc

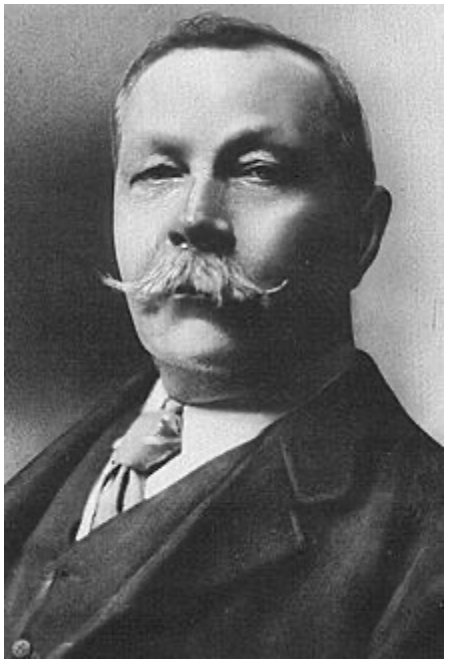
# MORE DATA


Data is produced not only by the companies but also by others about the companies

> Data! Data! Data! I can't make bricks without clay!

# Sir Arthur Conan Doyle

*Writer of the famous detective story Sherlock Holmes*



> The goal is to turn data into information, and information into insight.

**Carly Fiorina**

*Former CEO of HP*

# WHAT IS ANALYTICS?

**Analytics is the use of**

- Data
- Information Technology
- Statistical Analysis
- Mathematical or computer-based models

to help in **decision making**

# DECISION MAKING

## Decision Making

- Process of choosing two or more possible actions for the purpose of attaining a goal
- Want to provide a scientific/systematic explanation for the decision made
- Heavily influenced by many different disciplines: law, psychology, computer science, statistics, economics, operation research, etc

# HOW TO MAKE DECISION MAKING?

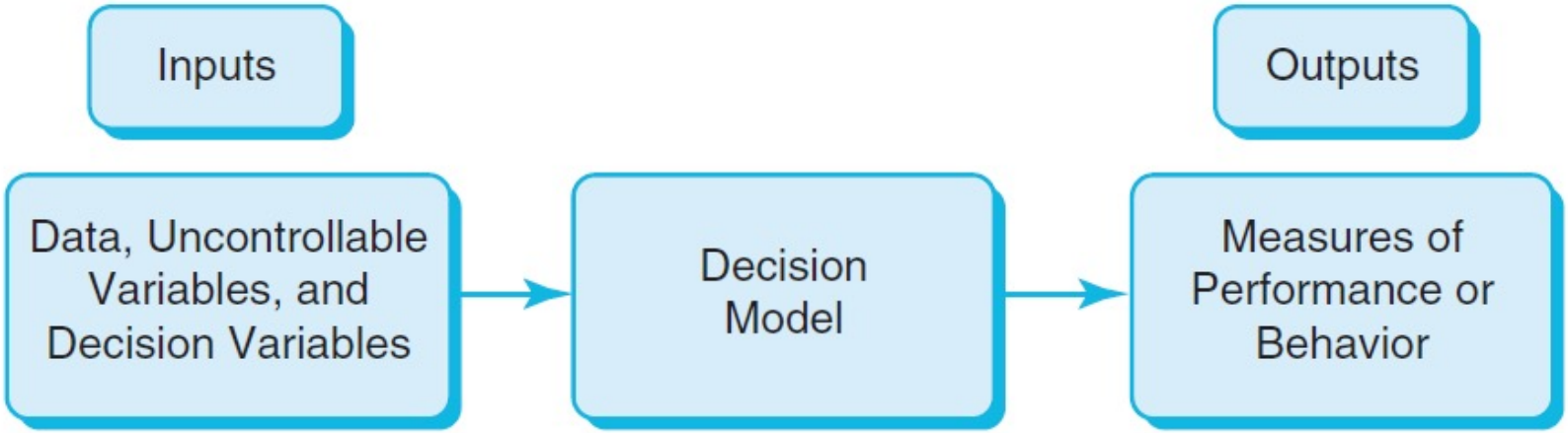
**Decision Making**

- Make **prediction** based on the data
  - E.g. Based on past experience (i.e. historical data) predict how many patients we are expecting today
- To do this, we build **decision models** using historical data
- The models can then allow us to make prediction of future data instances

# WHAT IS A DECISION MODEL?

**Decision Model is a model used to understand, analyze, or facilitate decision making**

- Can be in the form of a mathematical formula or software



# TYPES OF DECISION PROBLEM (LEARNING PROBLEMS)

**Regression (Supervised)**

- Stock price prediction

**Classification (Supervised)**

- Weather forecast (sunny, rainy, cloudy, etc)

**Clustering (Unsupervised)**

- Group Weibo/Twitter users based on their interest

What's the difference between regression & classification?

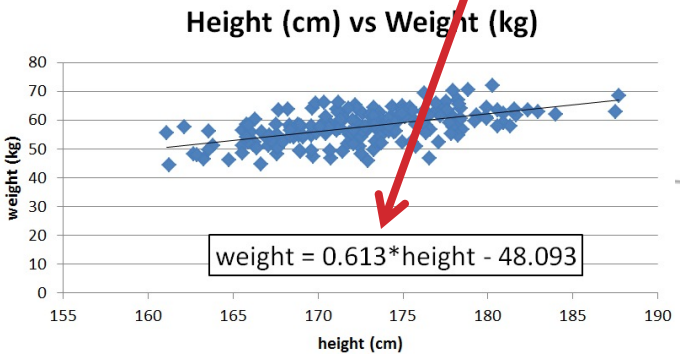
# SUPERVISED LEARNING

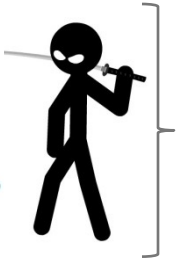
# REGRESSION

## Regression (Supervised)

- Using existing data instances to learn a **model** for predicting subsequent instances
- Example:
  - Assume we have a list of human height and weight

| Height(cm) | Weight(kg) |
|---|---|
| 167.0812 | 51.25136008 |
| 181.6608 | 61.91077208 |
| 176.276 | 69.41318376 |
| 173.2788 | 64.56428528 |
| 172.1866 | 65.4533256 |
| 174.498 | 55.9278936 |
| 177.292 | 64.17873208 |
| 177.8254 | 61.89716432 |
| 172.466 | 50.97013304 |
| 169.6212 | 54.73494664 |
| 168.8846 | 57.8103004 |
| 171.7548 | 51.77239088 |
| 173.482 | 56.97569112 |
| 170.4848 | 55.54687632 |
| 173.4312 | 52.65749528 |

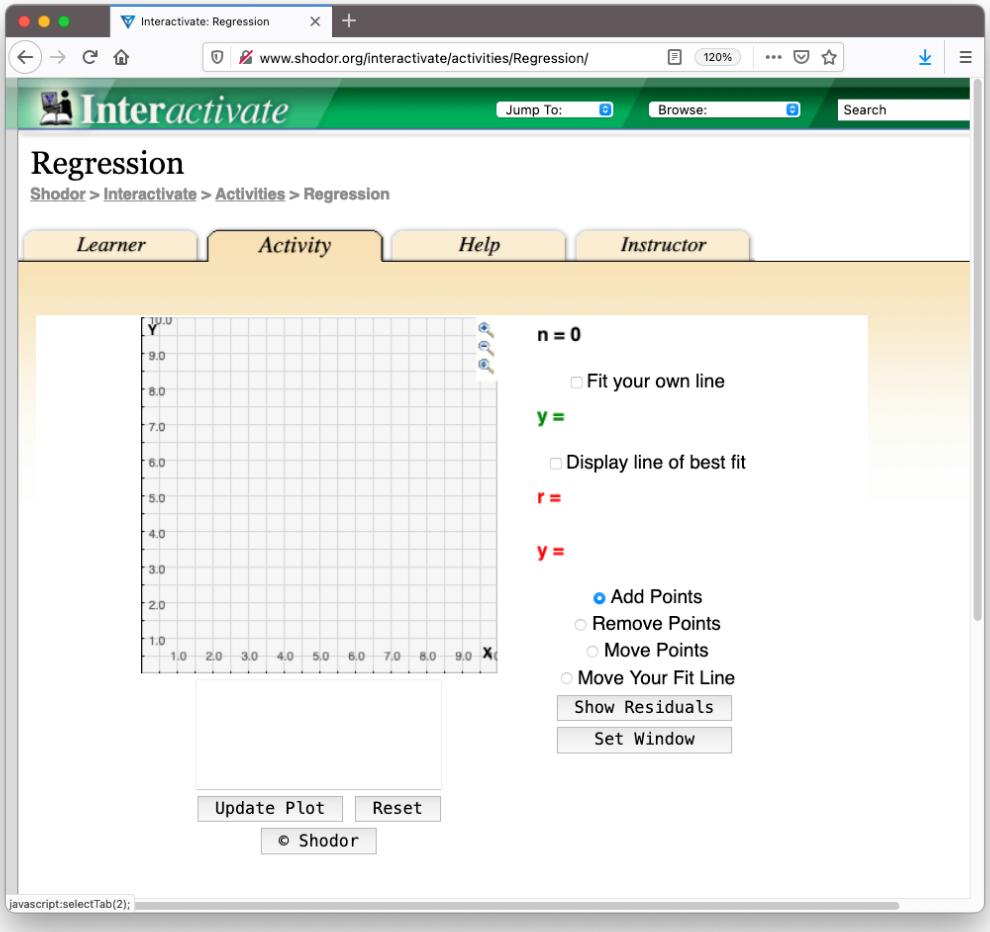


height = 170cm

weight = ?

# REGRESSION DEMO




http://www.shodor.org/interactivate/activities/Regression/

# REGRESSION EXAMPLE

**Dataset:**

- https://www.kaggle.com/mohansacharya/graduate-admissions
- Prediction of Graduate Admissions from an Indian perspective
- 500 instances
- Admission_Predict_Ver1.1.csv

# GRADUATE ADMISSION

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Serial No. | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of Admit |
| 2 | 1 | 337 | 118 | 4 | 4.5 | 4.5 | 9.65 | 1 | 0.92 |
| 3 | 2 | 324 | 107 | 4 | 4 | 4.5 | 8.87 | 1 | 0.76 |
| 4 | 3 | 316 | 104 | 3 | 3 | 3.5 | 8 | 1 | 0.72 |
| 5 | 4 | 322 | 110 | 3 | 3.5 | 2.5 | 8.67 | 1 | 0.8 |
| 6 | 5 | 314 | 103 | 2 | 2 | 3 | 8.21 | 0 | 0.65 |
| 7 | 6 | 330 | 115 | 5 | 4.5 | 3 | 9.34 | 1 | 0.9 |
| 8 | 7 | 321 | 109 | 3 | 3 | 4 | 8.2 | 1 | 0.75 |
| 9 | 8 | 308 | 101 | 2 | 3 | 4 | 7.9 | 0 | 0.68 |
| 10 | 9 | 302 | 102 | 1 | 2 | 1.5 | 8 | 0 | 0.5 |
| 11 | 10 | 323 | 108 | 3 | 3.5 | 3 | 8.6 | 0 | 0.45 |
| 12 | 11 | 325 | 106 | 3 | 3.5 | 4 | 8.4 | 1 | 0.52 |
| 13 | 12 | 327 | 111 | 4 | 4 | 4.5 | 9 | 1 | 0.84 |

Identifier (unique for all instance) → Not useful for modeling

Undergraduate GPA
(out of 10)

Research Experience
(0 or 1)

Chance of Admin
(0 to 1)

# GRADUATE ADMISSION

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Serial No. | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of Admit |
| 2 | 1 | 337 | 118 | 4 | 4.5 | 4.5 | 9.65 | 1 | 0.92 |
| 3 | 2 | 324 | 107 | 4 | 4 | 4.5 | 8.87 | 1 | 0.76 |
| 4 | 3 | 316 | 104 | 3 | 3 | 3.5 | 8 | 1 | 0.72 |
| 5 | 4 | 322 | 110 | 3 | 3.5 | 2.5 | 8.67 | 1 | 0.8 |
| 6 | 5 | 314 | 103 | 2 | 2 | 3 | 8.21 | 0 | 0.65 |
| 7 | 6 | 330 | 115 | 5 | 4.5 | 3 | 9.34 | 1 | 0.9 |
| 8 | 7 | 321 | 109 | 3 | 3 | 4 | 8.2 | 1 | 0.75 |
| 9 | 8 | 308 | 101 | 2 | 3 | 4 | 7.9 | 0 | 0.68 |
| 10 | 9 | 302 | 102 | 1 | 2 | 1.5 | 8 | 0 | 0.5 |
| 11 | 10 | 323 | 108 | 3 | 3.5 | 3 | 8.6 | 0 | 0.45 |
| 12 | 11 | 325 | 106 | 3 | 3.5 | 4 | 8.4 | 1 | 0.52 |
| 13 | 12 | 327 | 111 | 4 | 4 | 4.5 | 9 | 1 | 0.84 |

GRE Scores
(out of 340)

TOEFL Score
(out of 120)

Uni. Rating
(out of 5)

Statement of Purpose (SOP) & Letter of Recommendation (LOR) Recommendation Strength
(out of 5)

# TYPES OF DATA

Target/Response
Dependent Variable

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Serial No. | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of Admit |
| 2 | 1 | 337 | 118 | 4 | 4.5 | 4.5 | 9.65 | 1 | 0.92 |
| 3 | 2 | 324 | 107 | 4 | 4 | 4.5 | 8.87 | 1 | 0.76 |
| 4 | 3 | 316 | 104 | 3 | 3 | 3.5 | 8 | 1 | 0.72 |
| 5 | 4 | 322 | 110 | 3 | 3.5 | 2.5 | 8.67 | 1 | 0.8 |
| 6 | 5 | 314 | 103 | 2 | 2 | 3 | 8.21 | 0 | 0.65 |
| 7 | 6 | 330 | 115 | 5 | 4.5 | 3 | 9.34 | 1 | 0.9 |
| 8 | 7 | 321 | 109 | 3 | 3 | 4 | 8.2 | 1 | 0.75 |
| 9 | 8 | 308 | 101 | 2 | 3 | 4 | 7.9 | 0 | 0.68 |
| 10 | 9 | 302 | 102 | 1 | 2 | 1.5 | 8 | 0 | 0.5 |
| 11 | 10 | 323 | 108 | 3 | 3.5 | 3 | 8.6 | 0 | 0.45 |
| 12 | 11 | 325 | 106 | 3 | 3.5 | 4 | 8.4 | 1 | 0.52 |
| 13 | 12 | 327 | 111 | 4 | 4 | 4.5 | 9 | 1 | 0.84 |

Predictors
Independent Variables

# TYPES OF DATA

Datasets with values for both predictors & response are also known as **labeled** data (also known as **training data**)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Serial No. | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of Admit |
| 2 | 1 | 337 | 118 | 4 | 4.5 | 4.5 | 9.65 | 1 | 0.92 |
| 3 | 2 | 324 | 107 | 4 | 4 | 4.5 | 8.87 | 1 | 0.76 |
| 4 | 3 | 316 | 104 | 3 | 3 | 3.5 | 8 | 1 | 0.72 |
| 5 | 4 | 322 | 110 | 3 | 3.5 | 2.5 | 8.67 | 1 | 0.8 |
| 6 | 5 | 314 | 103 | 2 | 2 | 3 | 8.21 | 0 | 0.65 |
| 7 | 6 | 330 | 115 | 5 | 4.5 | 3 | 9.34 | 1 | 0.9 |
| 8 | 7 | 321 | 109 | 3 | 3 | 4 | 8.2 | 1 | 0.75 |
| 9 | 8 | 308 | 101 | 2 | 3 | 4 | 7.9 | 0 | 0.68 |
| 10 | 9 | 302 | 102 | 1 | 2 | 1.5 | 8 | 0 | 0.5 |
| 11 | 10 | 323 | 108 | 3 | 3.5 | 3 | 8.6 | 0 | 0.45 |
| 12 | 11 | 325 | 106 | 3 | 3.5 | 4 | 8.4 | 1 | 0.52 |
| 13 | 12 | 327 | 111 | 4 | 4 | 4.5 | 9 | 1 | 0.84 |

# TYPES OF DATA

Datasets with only values for predictors are also known as **unlabeled** data (also known as **testing data**)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Serial No. | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of Admit |
| 2 | 401 | 304 | 100 | 2 | 3.5 | 3 | 8.22 | 0 | ? |
| 3 | 402 | 315 | 105 | 2 | 3 | 3 | 8.34 | 0 | ? |
| 4 | 403 | 324 | 109 | 3 | 3.5 | 3 | 8.94 | 1 | ? |
| 5 | 404 | 330 | 116 | 4 | 4 | 3.5 | 9.23 | 1 | ? |
| 6 | 405 | 311 | 101 | 3 | 2 | 2.5 | 7.64 | 1 | ? |
| 7 | 406 | 302 | 99 | 3 | 2.5 | 3 | 7.45 | 0 | ? |
| 8 | 407 | 322 | 103 | 4 | 3 | 2.5 | 8.02 | 1 | ? |
| 9 | 408 | 298 | 100 | 3 | 2.5 | 4 | 7.95 | 1 | ? |
| 10 | 409 | 297 | 101 | 3 | 2 | 4 | 7.67 | 1 | ? |
| 11 | 410 | 300 | 98 | 1 | 2 | 2.5 | 8.02 | 0 | ? |

We want the generated model to predict the response values

# SUPERVISED LEARNING

**This also illustrates the idea of**
**Supervised Learning**

- Teach the machine to do prediction with examples and the corresponding expected prediction

# HANDS-ON: REGRESSION

# HANDS-ON: REGRESSION

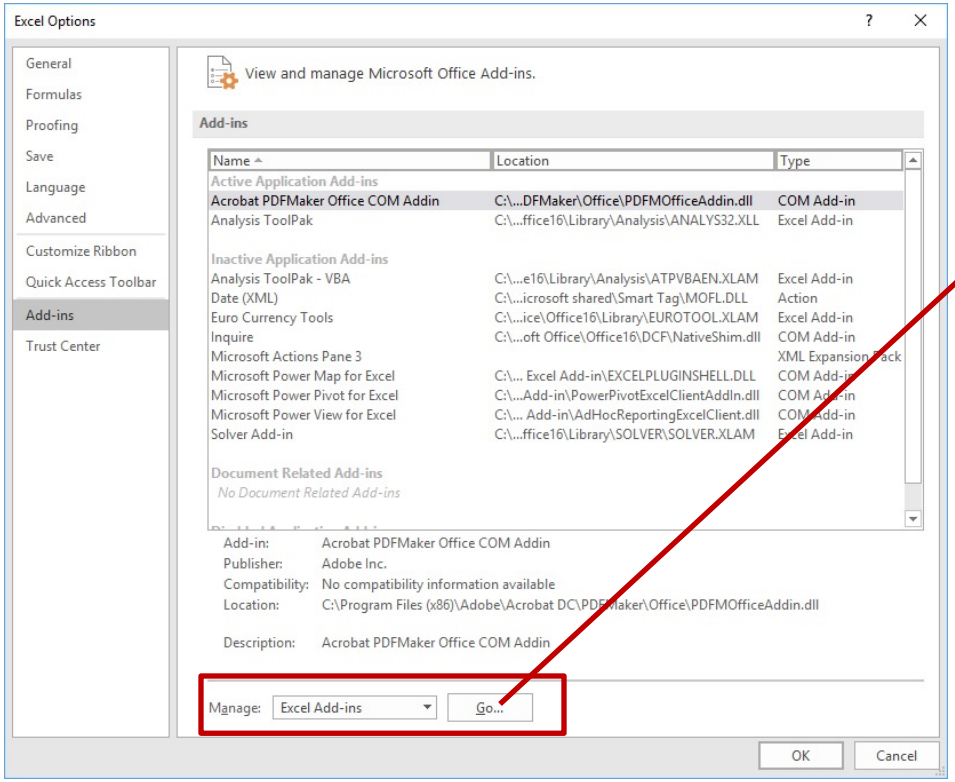
**Dataset:**

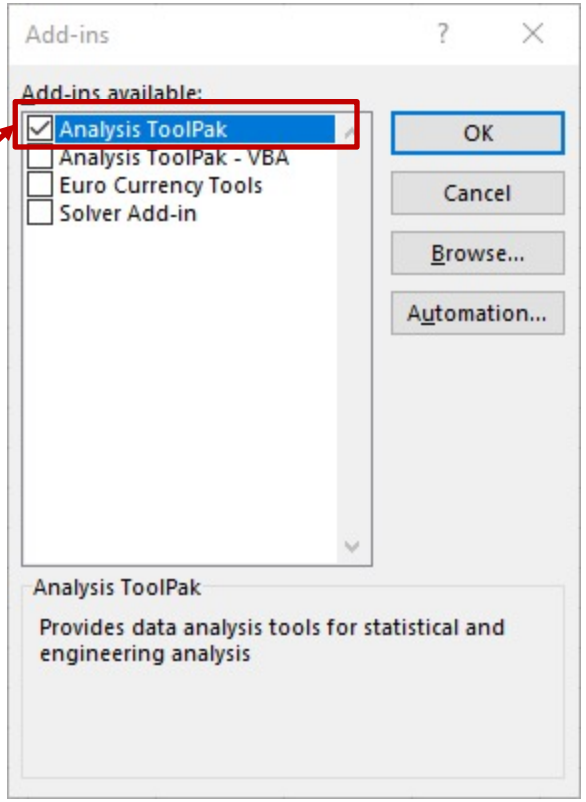
- https://www.kaggle.com/mohansacharya/graduate-admissions
- Prediction of Graduate Admissions from an Indian perspective
- 500 instances
- **Manually divided into:**
  - **400 instances training data (Admission_Predict_Ver1.1.train.csv)**
  - **100 instances testing data (Admission_Predict_Ver1.1.test.csv)**

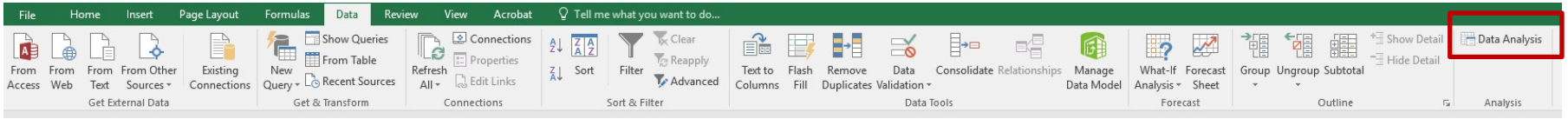
**Software:**

- Excel (with the **Analysis ToolPak**)
- WEKA (https://www.cs.waikato.ac.nz/ml/weka/)

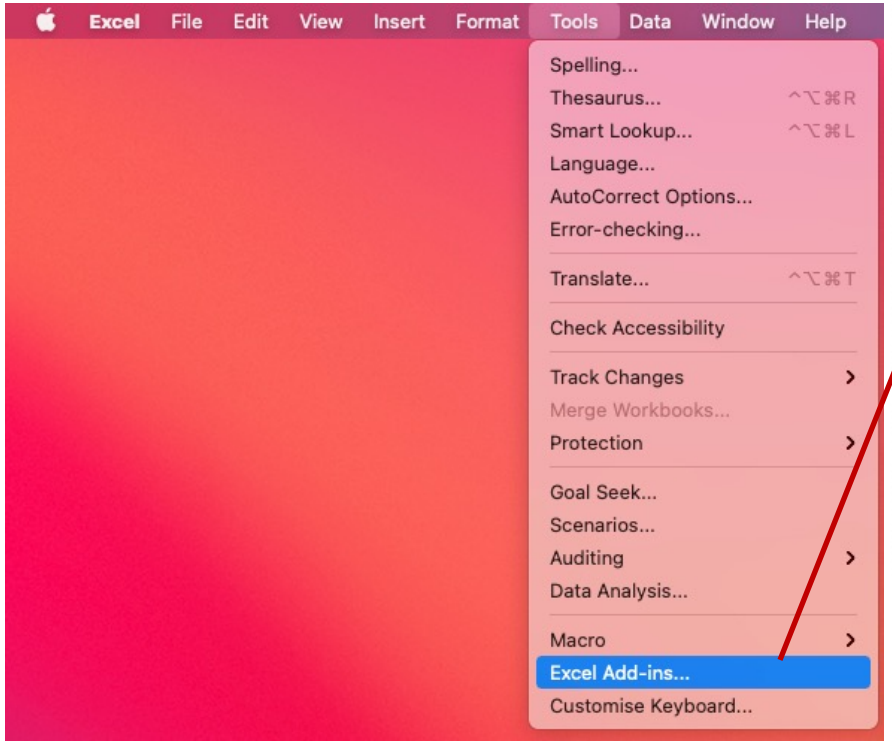

# INSTALLING EXCEL ANALYSIS TOOLPAK (WIN)

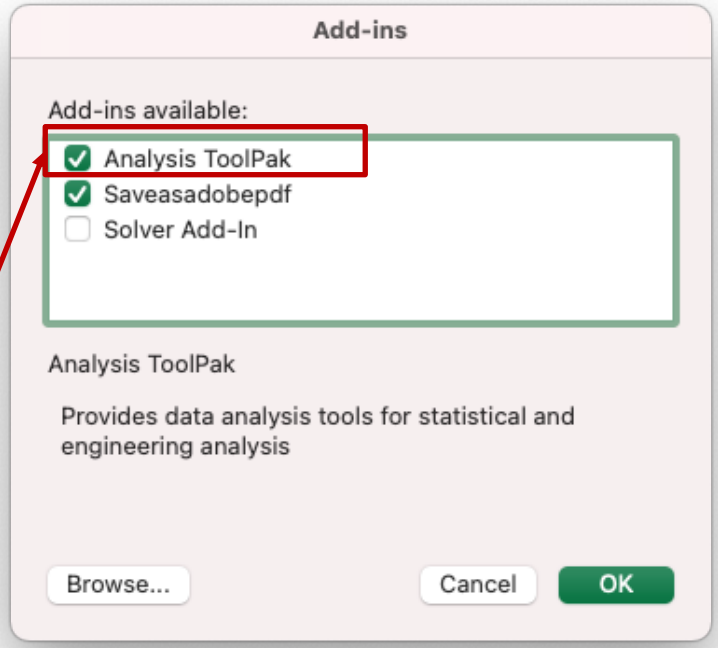


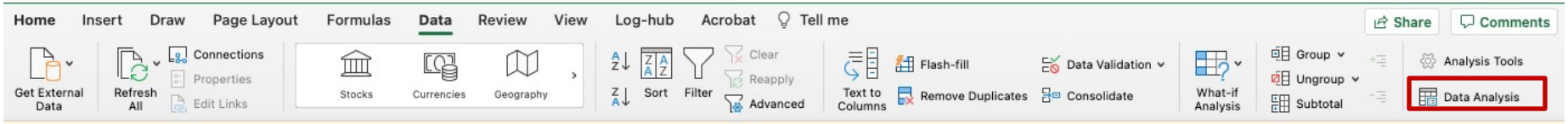



# INSTALLING EXCEL ANALYSIS TOOLPAK (MAC)

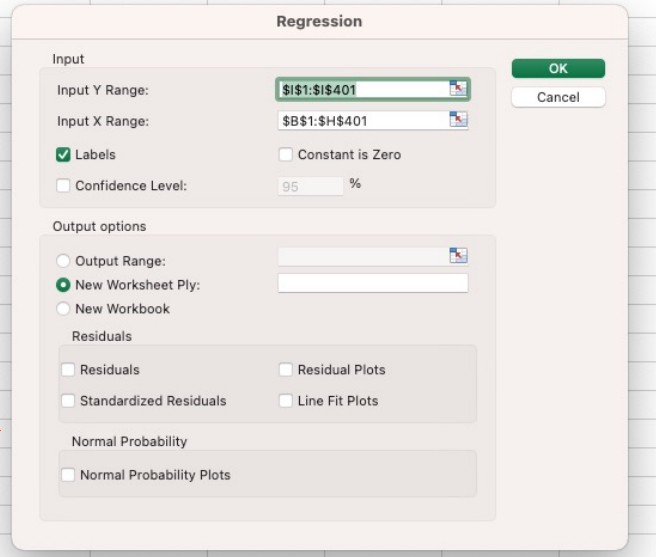


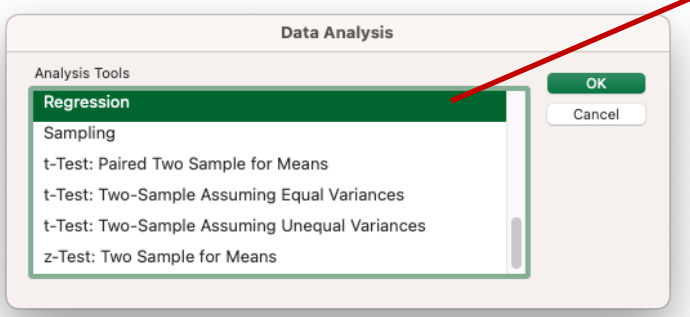



# GENERATING REGRESSION MODEL (EXCEL)

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Serial No. | GRE Score | TOEFL Score | University Ra | SOP | LOR | CGPA | Research | Chance of Admit |
| 2 | 1 | 337 | 118 | 4 | 4.5 | 4.5 | 9.65 | 1 | 0.92 |
| 3 | 2 | 324 | 107 | 4 | 4 | 4.5 | 8.87 | 1 | 0.76 |
| 4 | 3 | 316 | 104 | 3 | 3 | 3.5 | 8 | 1 | 0.72 |
| 5 | 4 | 322 | 110 | 3 | 3.5 | 2.5 | 8.67 | 1 | 0.8 |
| 6 | 5 | 314 | 103 | 2 | 2 | 3 | 8.21 | 0 | 0.65 |
| 7 | 6 | 330 | 115 | 5 | 4.5 | 3 | 9.34 | 1 | 0.9 |
| 8 | 7 | 321 | 109 | 3 | 3 | 4 | 8.2 | 1 | 0.75 |
| 9 | 8 | 308 | 101 | 2 | 3 | 4 | 7.9 | 0 | 0.68 |
| 10 | 9 | 302 | 102 | 1 | 2 | 1.5 | 8 | 0 | 0.5 |
| 11 | 10 | 323 | 108 | 3 | 3.5 | 3 | 8.6 | 0 | 0.45 |
| 12 | 11 | 325 | 106 | 3 | 3.5 | 4 | 8.4 | 1 | 0.52 |
| 13 | 12 | 327 | 111 | 4 | 4 | 4.5 | 9 | 1 | 0.84 |
| 14 | 13 | 328 | 112 | 4 | 4 | 4.5 | 9.1 | 1 | 0.78 |
| 15 | 14 | 307 | 109 | 3 | 4 | 3 | 8 | 1 | 0.62 |
| 16 | 15 | 311 | 104 | 3 | 3.5 | 2 | 8.2 | 1 | 0.61 |
| 17 | 16 | 314 | 105 | 3 | 3.5 | 2.5 | 8.3 | 0 | 0.54 |
| 18 | 17 | 317 | 107 | 3 | 4 | 3 | 8.7 | 0 | 0.66 |
| 19 | 18 | 319 | 106 | 3 | 4 | 3 | 8 | 1 | 0.65 |
| 20 | 19 | 318 | 110 | 3 | 4 | 3 | 8.8 | 0 | 0.63 |




Admission_Predict_Ver1.1.train.csv

# GENERATING REGRESSION MODEL (EXCEL)

| | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
|---|---|---|---|---|---|---|---|---|---|
| 17 | Intercept | -1.259432478 | 0.124730747 | -10.097209 | 1.841E-21 | -1.5046574 | -1.0142076 | -1.5046574 | -1.0142076 |
| 18 | GRE Score | 0.001737412 | 0.000597897 | 2.9058702 | 0.0038701 | 0.0005619 | 0.0029129 | 0.0005619 | 0.0029129 |
| 19 | TOEFL Score | 0.002919577 | 0.001089532 | 2.6796622 | 0.0076802 | 0.0007775 | 0.0050616 | 0.0007775 | 0.0050616 |
| 20 | University Rating | 0.005716658 | 0.004770425 | 1.1983539 | 0.2315032 | -0.0036622 | 0.0150955 | -0.0036622 | 0.0150955 |
| 21 | SOP | -0.003305169 | 0.005561643 | -0.5942792 | 0.5526682 | -0.0142395 | 0.0076292 | -0.0142395 | 0.0076292 |
| 22 | LOR | 0.022353127 | 0.005541485 | 4.0337793 | 6.599E-05 | 0.0114584 | 0.0332479 | 0.0114584 | 0.0332479 |
| 23 | CGPA | 0.118939454 | 0.012219435 | 9.7336294 | 3.382E-20 | 0.0949156 | 0.1429633 | 0.0949156 | 0.1429633 |
| 24 | Research | 0.024525106 | 0.007959756 | 3.0811379 | 0.0022076 | 0.008876 | 0.0401743 | 0.008876 | 0.0401743 |

-1.259432478 + 0.001737412 * **GRE** + 0.002919577 * **TOEFL** + 0.005716658 * **Uni_Rating** - 0.003305169 * **SOP** + 0.022353127 * **LOR** + 0.118939454 * **CGPA** + 0.024525106 * **Research**

# PREDICTING TESTING DATA (EXCEL)

Admission_Predict_Ver1.1.test.csv

-1.259432478 + 0.001737412 * **GRE** + 0.002919577 * **TOEFL** + 0.005716658 * **Uni_Rating** - 0.003305169 * **SOP** + 0.022353127 * **LOR** + 0.118939454 * **CGPA** + 0.024525106 * **Research**

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Serial No. | GRE Score | TOEFL Score | University Rating | SOP | LOR | CGPA | Research | Chance of Admit (Actual) | Chance of Admit (Predicted) |
| 2 | 401 | 304 | 100 | 2 | 3.5 | 3 | 8.22 | 0 | 0.63 | 0.605305387 |
| 3 | 402 | 315 | 105 | 2 | 3 | 3 | 8.34 | 0 | 0.66 | 0.654940123 |
| 4 | 403 | 324 | 109 | 3 | 3.5 | 3 | 8.94 | 1 | 0.78 | 0.782207991 |
| 5 | 404 | 330 | 116 | 4 | 4 | 3.5 | 9.23 | 1 | 0.91 | 0.862802581 |
| 6 | 405 | 311 | 101 | 3 | 2 | 2.5 | 7.64 | 1 | 0.62 | 0.575424919 |
| 7 | 406 | 302 | 99 | 3 | 2.5 | 3 | 7.45 | 0 | 0.52 | 0.516349434 |

Predicted values

# PERFORM REGRESSION IN WEKA

**Limitations in Excel**

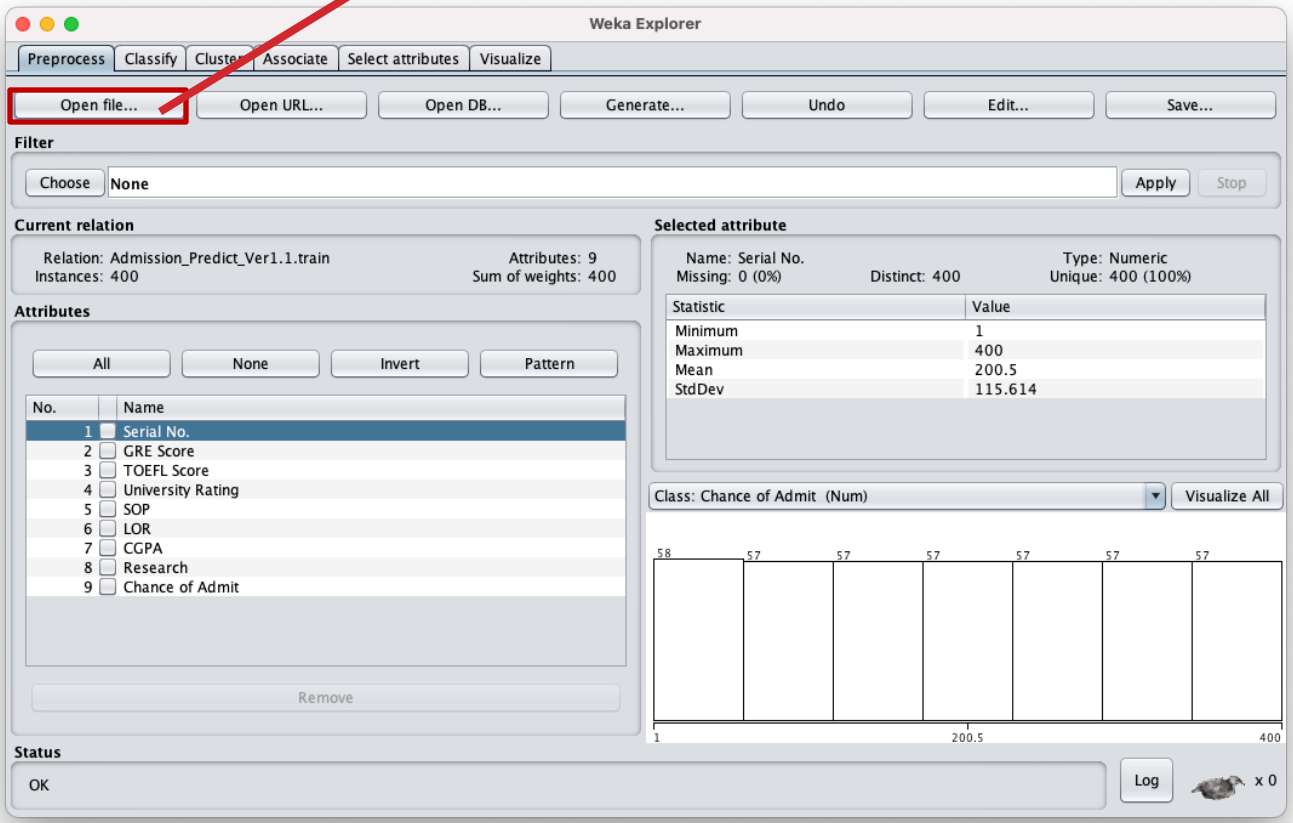
- Very tedious to do it properly in Excel
- Regression tool in Analysis Toolpak has a limit on the number of independent variables

**WEKA**

- Open source machine learning software with a graphical user interface for trying out different machine learning techniques and for running experiments

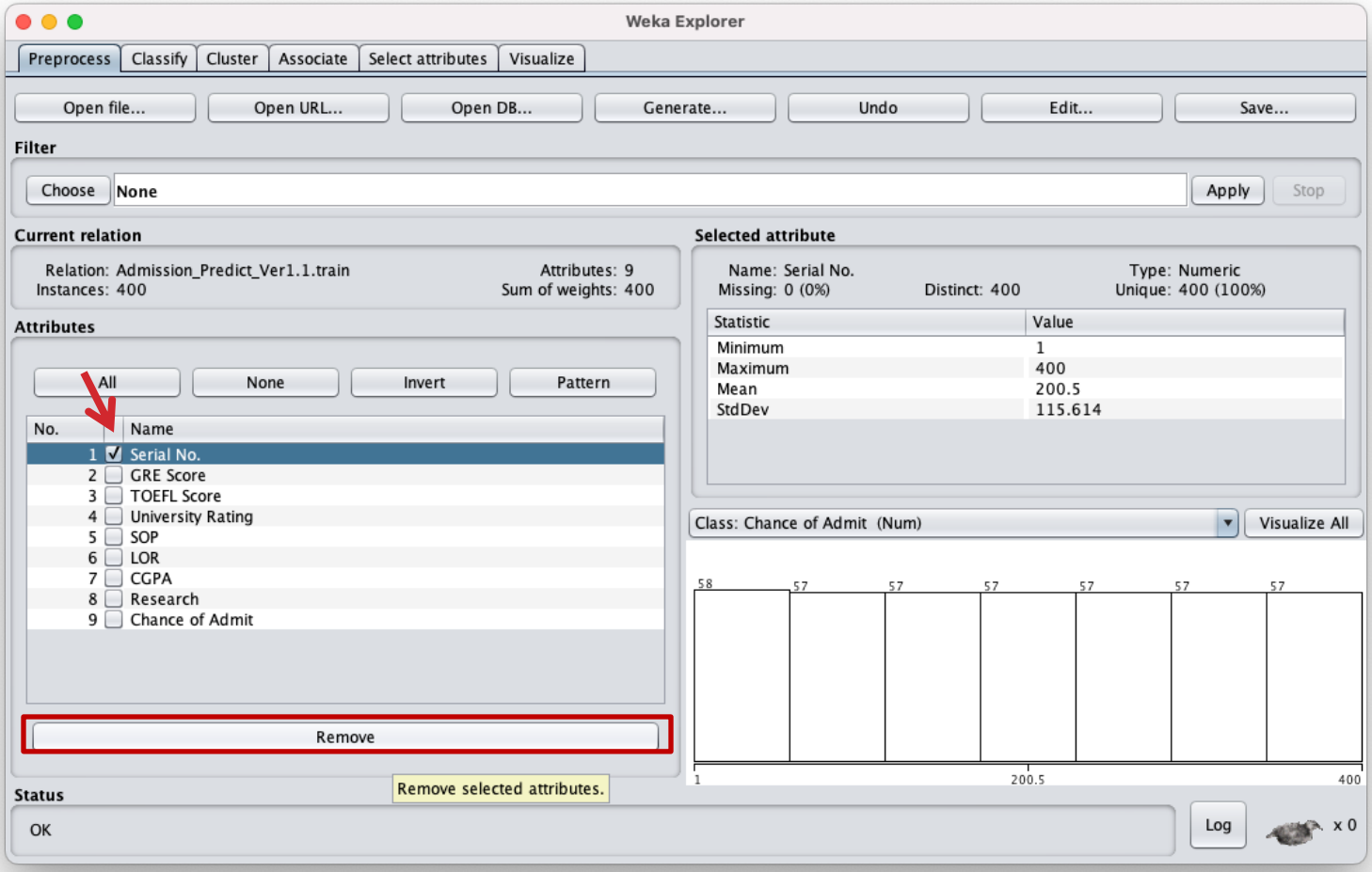
# PERFORM REGRESSION IN WEKA


Admission_Predict_Ver1.1.train.csv

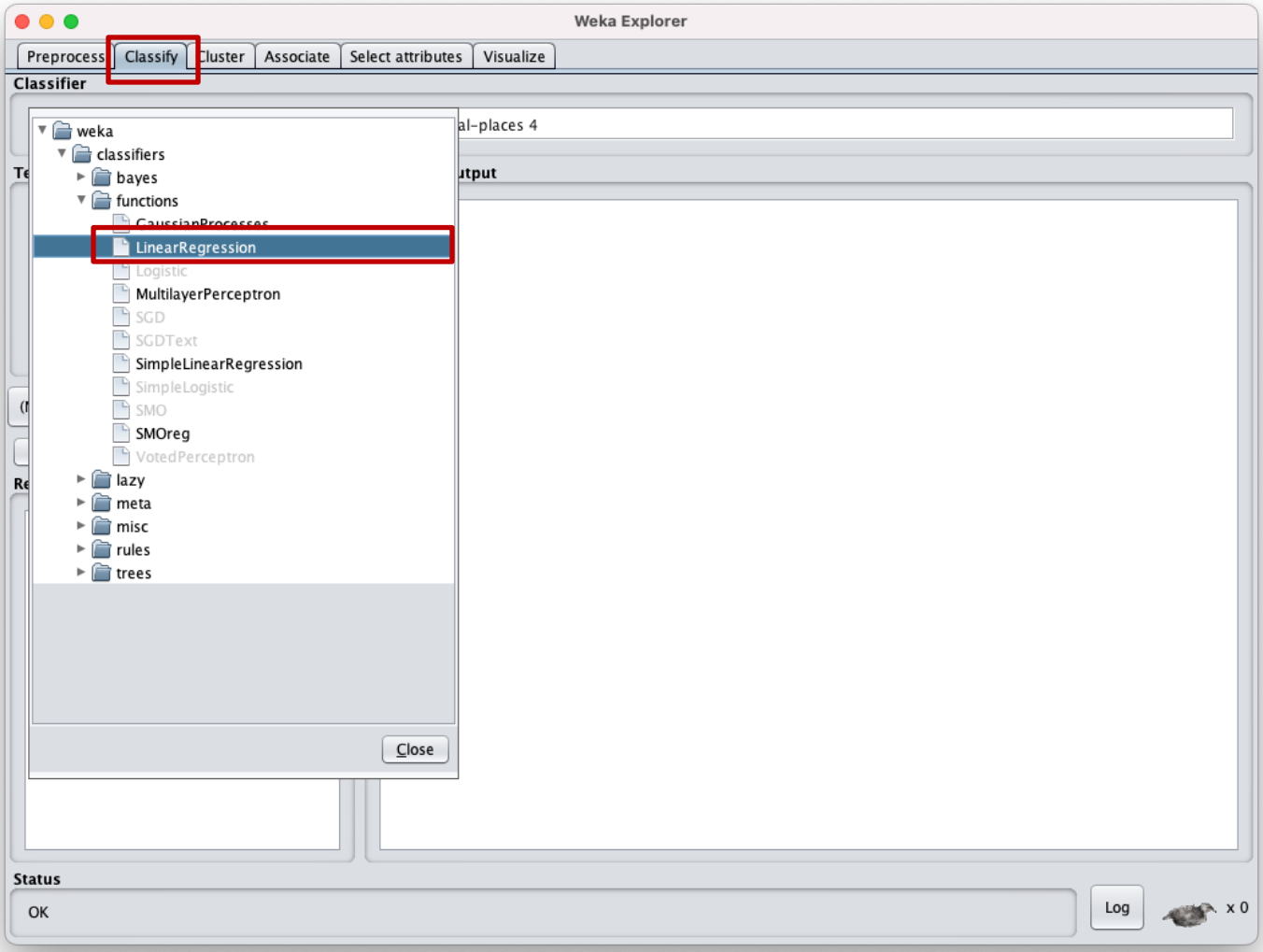


# PERFORM REGRESSION IN WEKA

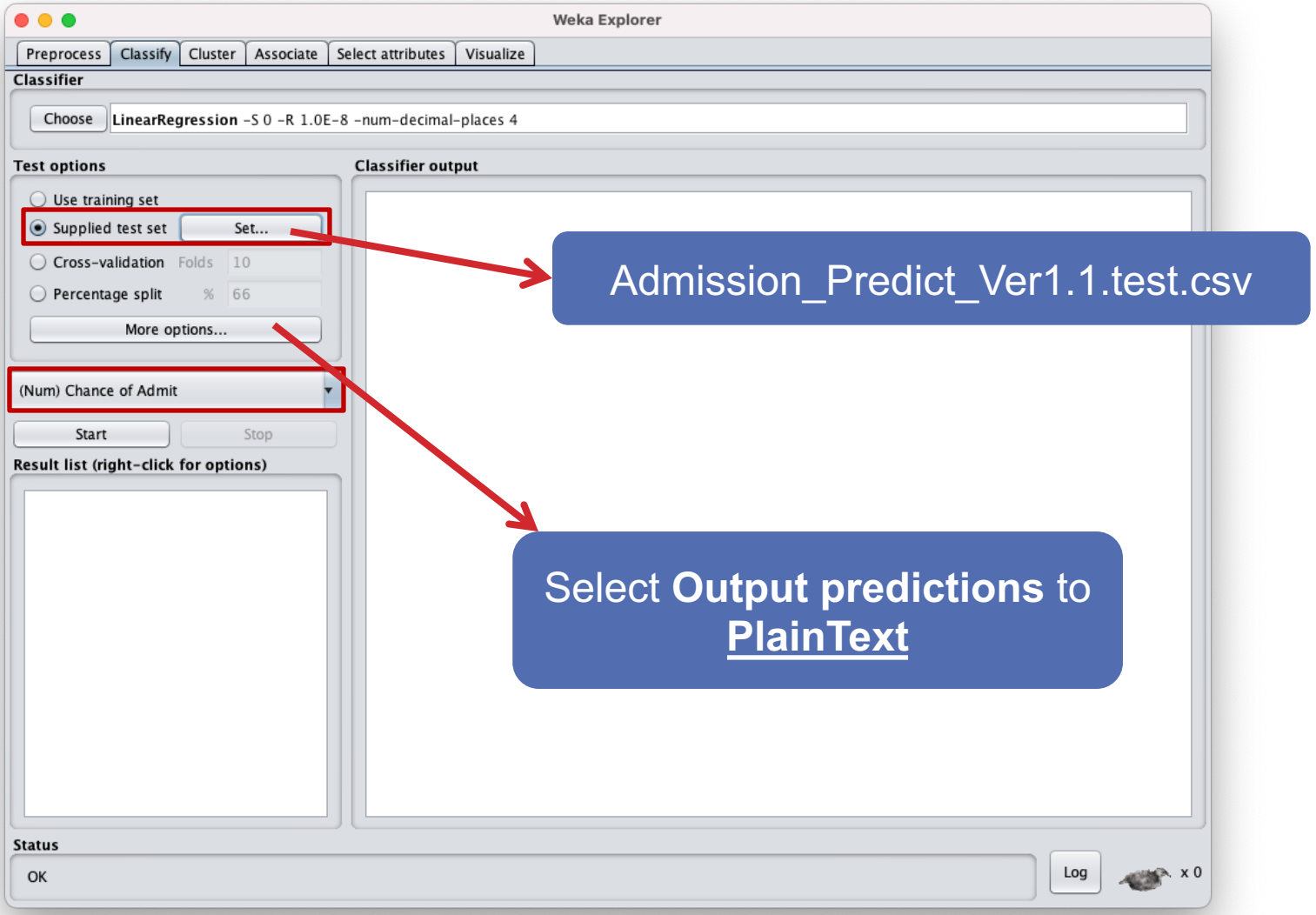
Remove Serial No. attribute



# PERFORM REGRESSION IN WEKA

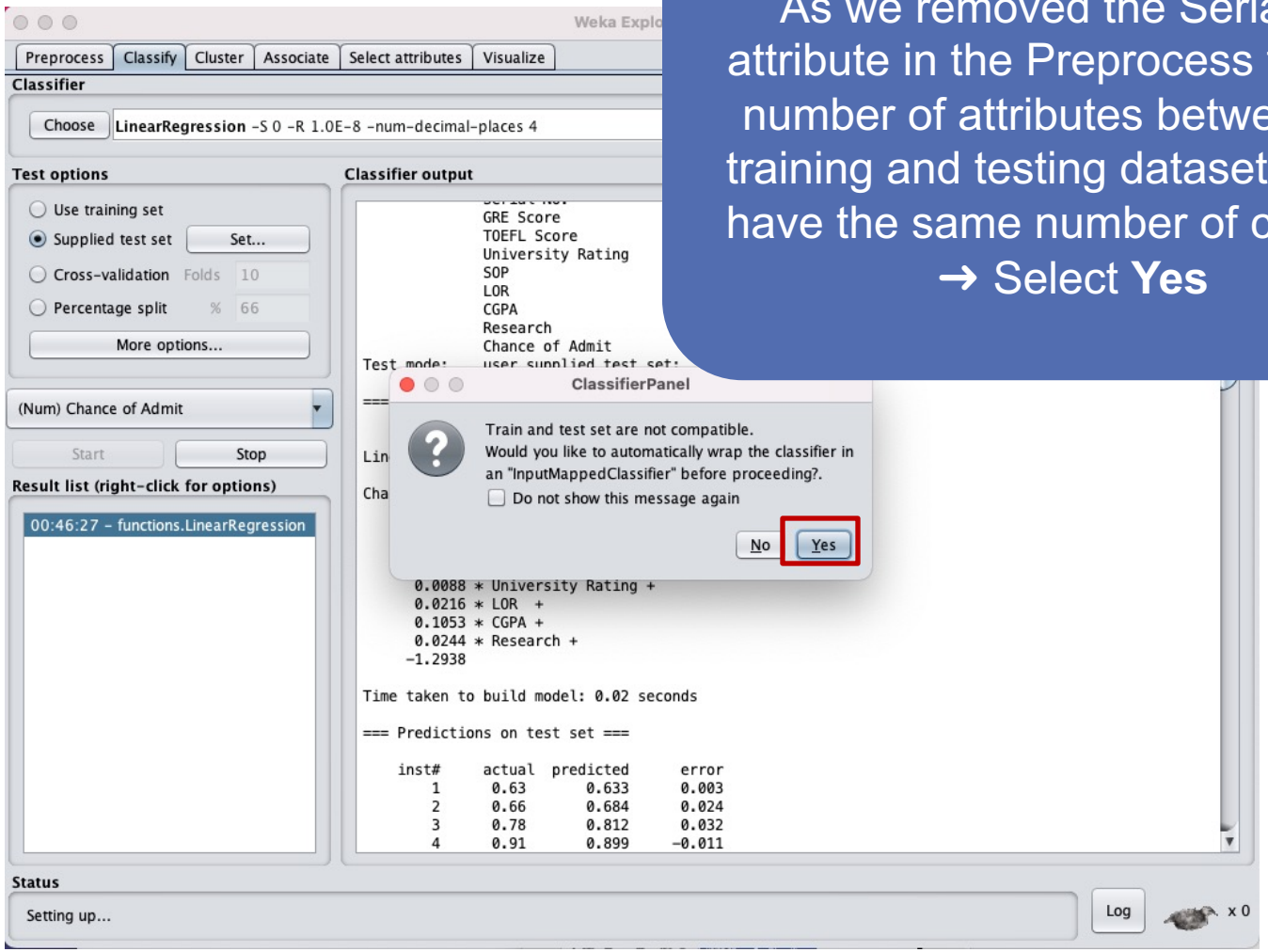


# PERFORM REGRESSION IN WEKA



# PERFORM REGRESSION IN WEKA

As we removed the Serial No attribute in the Preprocess tab, the number of attributes between the training and testing dataset will not have the same number of columns → Select **Yes**



# PERFORM REGRESSION IN WEKA

```
Linear Regression Model

Chance of Admit  =

      0.0018 * GRE Score +
      0.003  * TOEFL Score +
      0.0228 * LOR  +
      0.121  * CGPA +
      0.0246 * Research +
     -1.2985
Attribute mappings:

Model attributes                    Incoming attributes
----------------                    -------------------
(numeric) GRE Score                 --> 2 (numeric) GRE Score
(numeric) TOEFL Score               --> 3 (numeric) TOEFL Score
(numeric) University Rating         --> 4 (numeric) University Rating
(numeric) SOP                       --> 5 (numeric) SOP
(numeric) LOR                       --> 6 (numeric) LOR
(numeric) CGPA                      --> 7 (numeric) CGPA
(numeric) Research                  --> 8 (numeric) Research
(numeric) Chance of Admit           --> 9 (numeric) Chance of Admit


Time taken to build model: 0.02 seconds

=== Predictions on test set ===

    inst#     actual  predicted        error
        1       0.63      0.609       -0.021
        2       0.66      0.659       -0.001
        3       0.78      0.784        0.004
        4       0.91      0.862       -0.048
        5       0.62      0.568       -0.052
        6       0.52       0.51        -0.01
        7       0.61       0.64         0.03
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correlation coefficient                  0.9476
Mean absolute error                      0.0336
Root mean squared error                  0.043
Relative absolute error                 29.9675 %
Root relative squared error             31.8241 %
Total Number of Instances              100
```

Overall performance
(there are other metrics – will be elaborated more in the regression topic)

# PERFORM REGRESSION IN WEKA

```
Linear Regression Model

Chance of Admit  =

      0.0018 * GRE Score +
      0.003  * TOEFL Score +
      0.0228 * LOR  +
      0.121  * CGPA +
      0.0246 * Research +
     -1.2985
Attribute mappings:

Model attributes                    Incoming attributes
----------------------------       ----------------
(numeric) GRE Score                --> 2 (numeric) GRE Score
(numeric) TOEFL Score              --> 3 (numeric) TOEFL Score
(numeric) University Rating        --> 4 (numeric) University Rating
(numeric) SOP                      --> 5 (numeric) SOP
(numeric) LOR                      --> 6 (numeric) LOR
(numeric) CGPA                     --> 7 (numeric) CGPA
(numeric) Research                 --> 8 (numeric) Research
(numeric) Chance of Admit          --> 9 (numeric) Chance of Admit


Time taken to build model: 0.02 seconds

=== Predictions on test set ===

    inst#     actual  predicted      error
        1       0.63      0.609     -0.021
        2       0.66      0.659     -0.001
        3       0.78      0.784      0.004
        4       0.91      0.862     -0.048
        5       0.62      0.568     -0.052
        6       0.52       0.51      -0.01
        7       0.61       0.64       0.03
```

```
=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds

=== Summary ===

Correlation coefficient                  0.9476
Mean absolute error                      0.0336
Root mean squared error                  0.043
Relative absolute error                 29.9675 %
Root relative squared error             31.8241 %
Total Number of Instances              100
```

Notice that not all predictors are used (due to attribute selection)

Possible to turn off (and use all predictors)

# STATISTICAL LEARNING

**What is Statistical Learning?**

- Suppose $Y_i$ = quantitative response, $X_i = (X_{ij}, .., X_{ip})$ for observation $i = 1, .., n,$ and predictor $j = 1, .., p$
- We assume there is some relationship between $Y$ and $X$
- This can be modeled as

$$Y_i = f(X_i) + \varepsilon_i$$

- where $f$ = unknown function and $\varepsilon$ = random error with mean zero

# STATISTICAL LEARNING

$$Y_i = f(X_i) + \varepsilon_i$$

**Example**

- Suppose $Y_i$ = quantitative response, $X_i = (X_{ij}, .., X_{ip})$ for observation $i = 1, .., n,$ and predictor $j = 1, .., p$

| Height(cm) | Weight(kg) |
|---|---|
| 167.0812 | 51.25136008 |
| 181.6608 | 61.91077208 |
| 176.276 | 69.41318376 |
| 173.2788 | 64.56428528 |
| 172.1866 | 65.4533256 |
| 174.498 | 55.9278936 |
| 177.292 | 64.17873208 |
| 177.8254 | 61.89716432 |
| 172.466 | 50.97013304 |
| 169.6212 | 54.73494664 |
| 168.8846 | 57.8103004 |
| 171.7548 | 51.77299088 |
| 173.482 | 56.97569112 |
| 170.4848 | 55.54687632 |
| 173.4312 | 52.65749528 |

$X_{ij}$ = predictor

$Y_i$ = quantitative response

**Predicting the weight using the height**

# WHY ESTIMATE F?

$$Y_i = f(X_i) + \varepsilon_i$$

**2 main reasons:**

- Prediction
- Inference

# PREDICTION

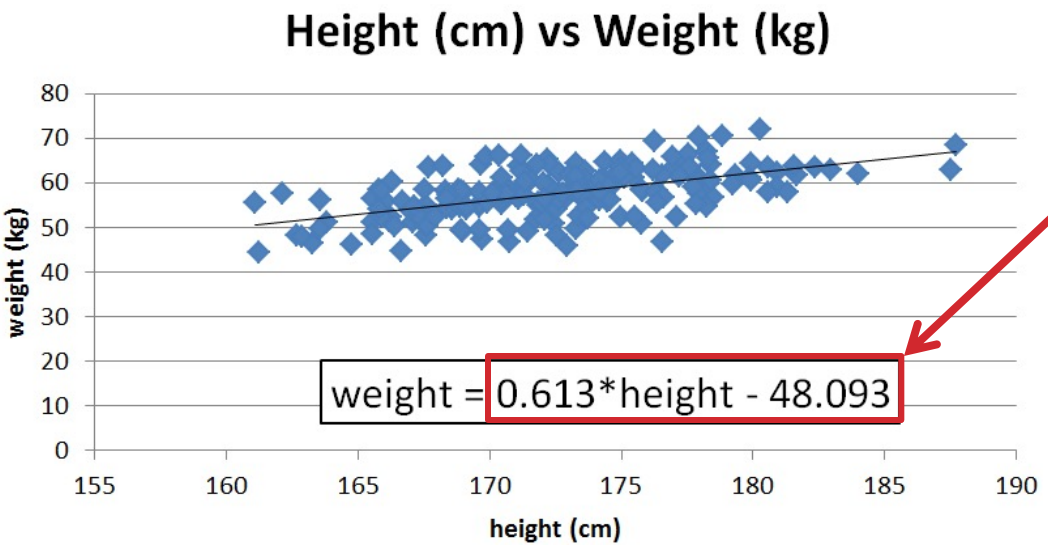
Mean of 0

$$Y_i = f(X_i) + \varepsilon_i$$

Prediction of Y

Estimate for *f*

$$\hat{Y}_i = \hat{f}(X_i)$$




Predicting the weight using the height

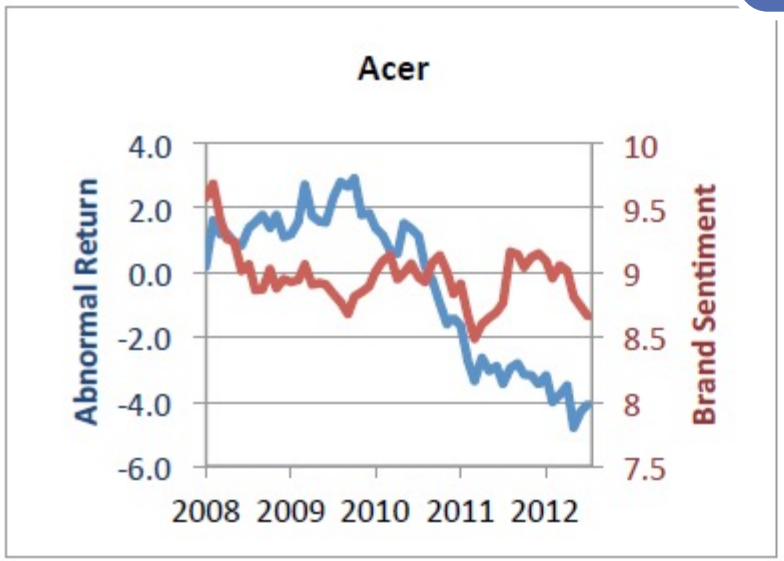
# INFERENCE

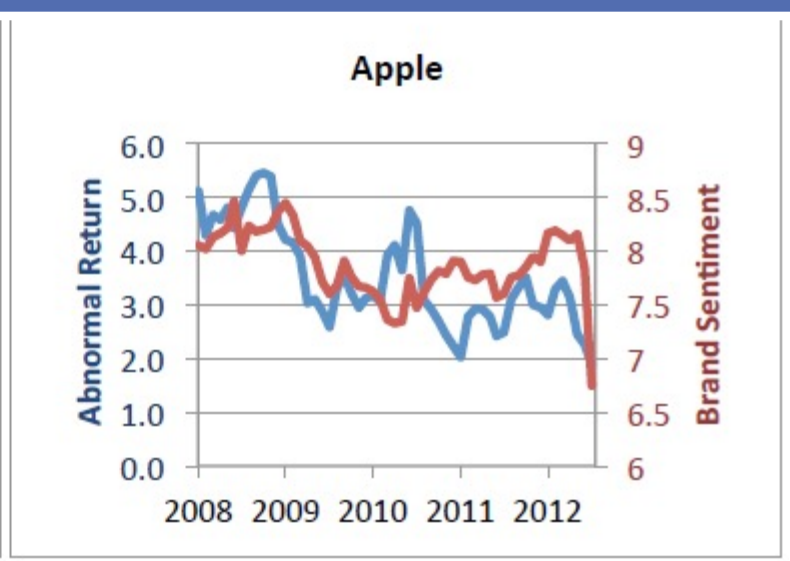
**Want to find the relationship between response *Y* and predictors $X_{1, \ldots} X_p$**

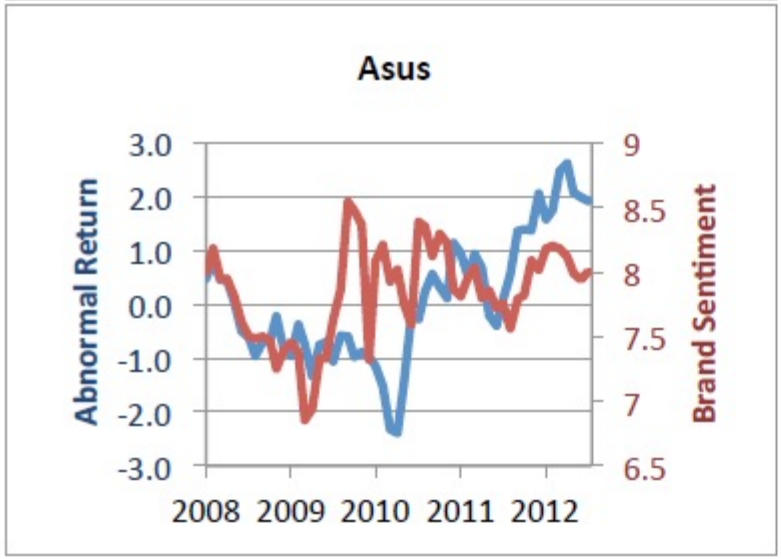
- Which predictor will affect the response?
- Which predictor is more significant than the other predictors?
- Is the relationship positive or negative?
- Can the relationship between Y and X be modeled using a linear function?

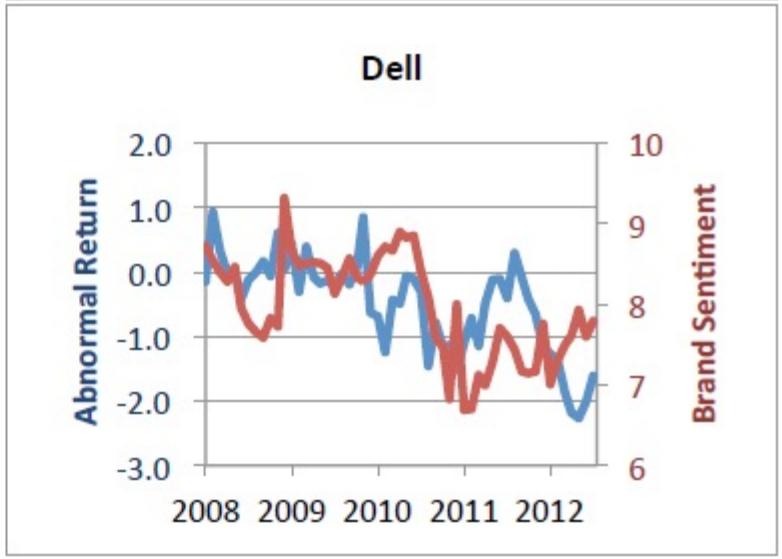
# INFERENCE

Is there a correlation between Brand Sentiment and Stock Price Movements?



# HOW TO ESTIMATE F?

**Assume we have a set of training data:**

$$\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$$

**Apply a statistical learning method and feed in the training data to train a model / estimate *f***

# RECALL: TYPES OF DECISION PROBLEM (LEARNING PROBLEMS)

**Regression (Supervised)**

- Stock price prediction

**Classification (Supervised)**

- Weather forecast (sunny, rainy, cloudy, etc)

**Clustering (Unsupervised)**

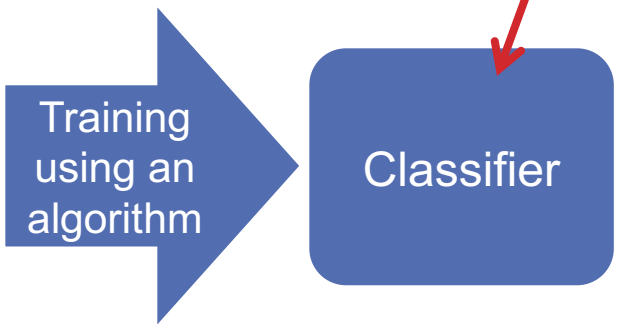
- Group Facebook users based on their interest

# CLASSIFICATION

**Classification (Supervised)**

- Using existing data instances (labeled) to learn a model for predicting subsequent instances (unlabeled)
- Example:
  - Assume that you have some attributes of the current weather, and we need to decide whether to go out to play

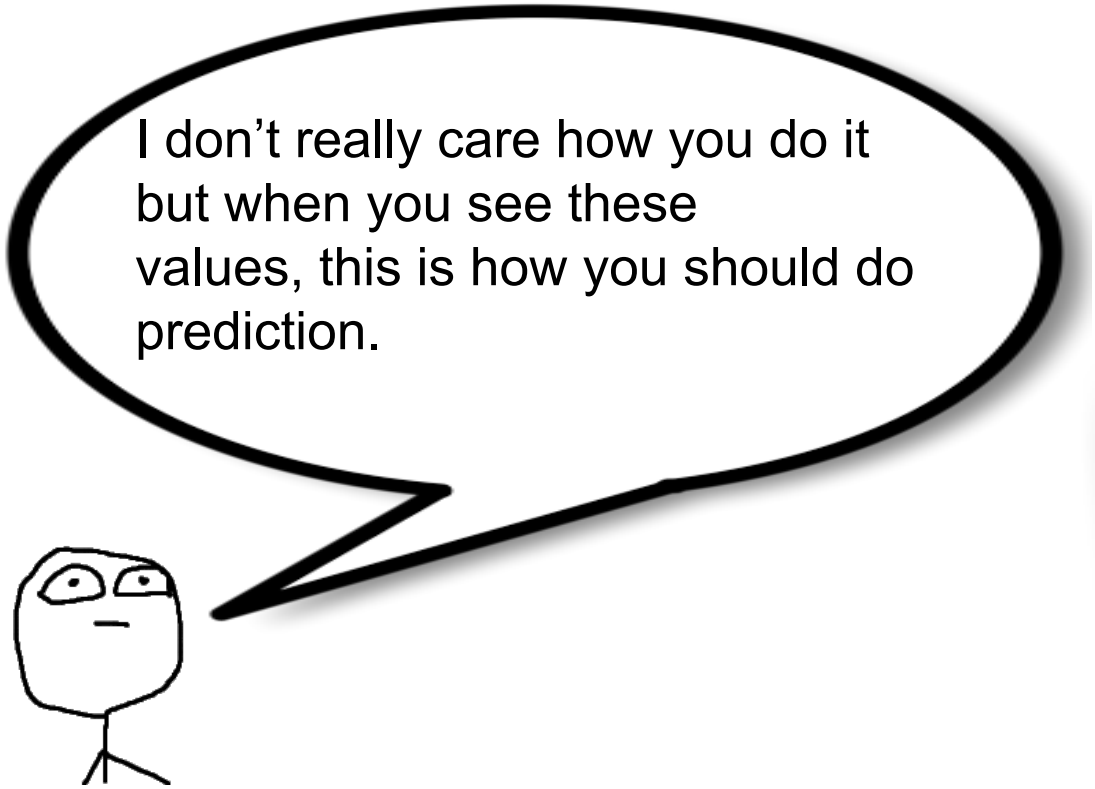
| No. | 1: outlook Nominal | 2: temperature Nominal | 3: humidity Nominal | 4: windy Nominal | 5: **play** Nominal |
|---|---|---|---|---|---|
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |



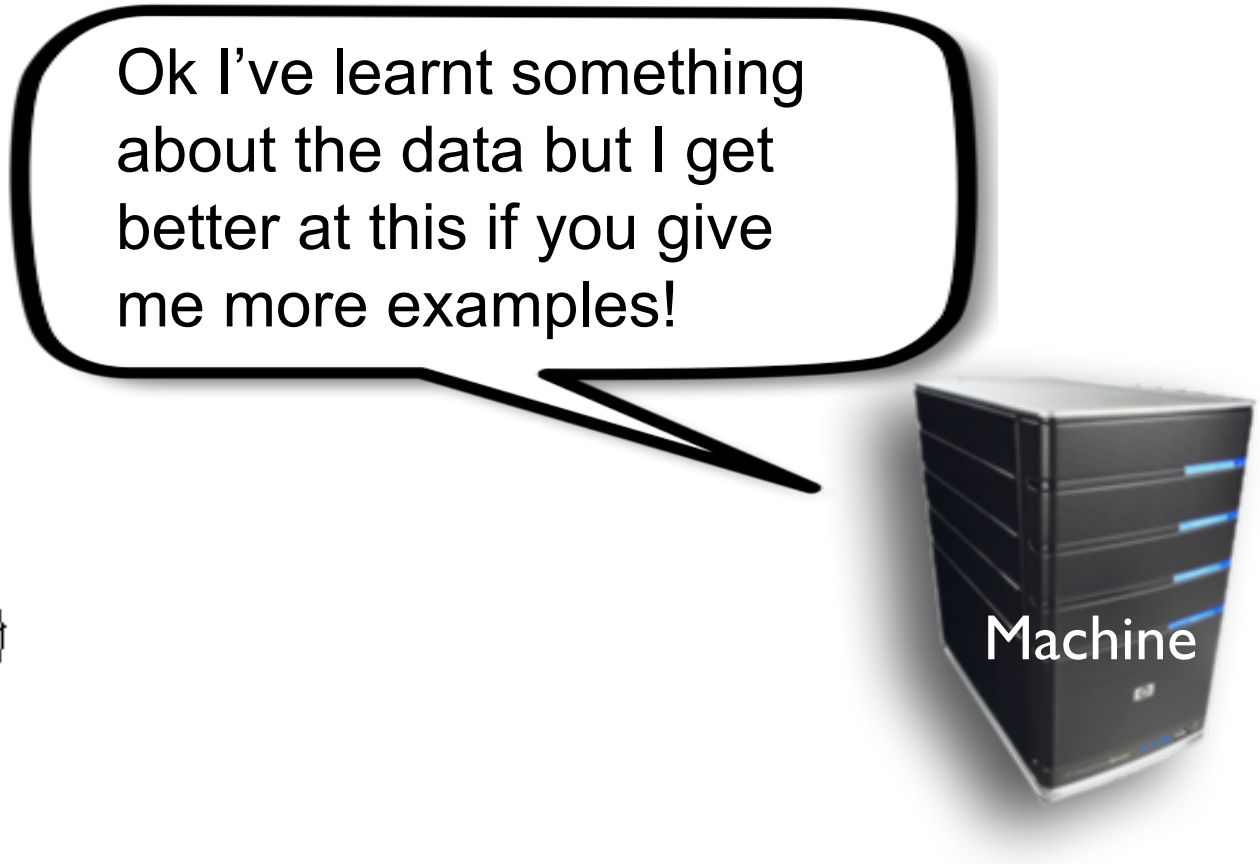

outlook: rainy
temperature: mild
humidity: high
windy: TRUE
play: ?

# UNSUPERVISED LEARNING

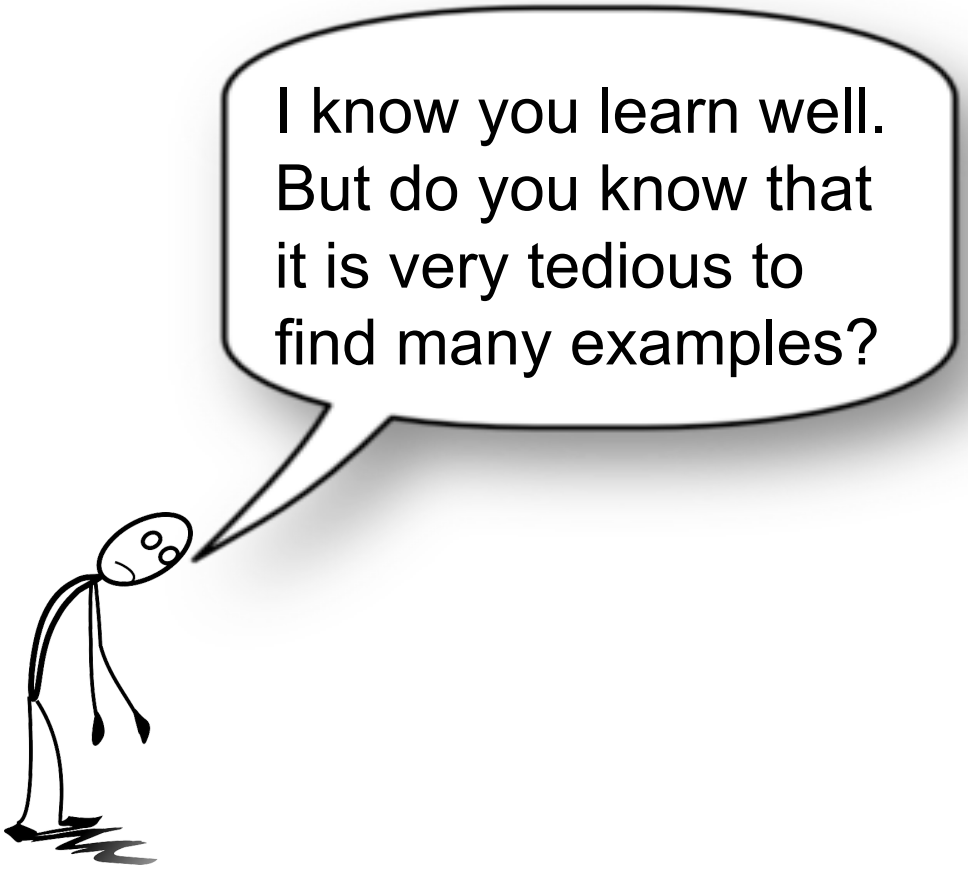
# MACHINE LEARNING




# MACHINE LEARNING



# CHALLENGES OF SUPERVISED LEARNING




# UNSUPERVISED

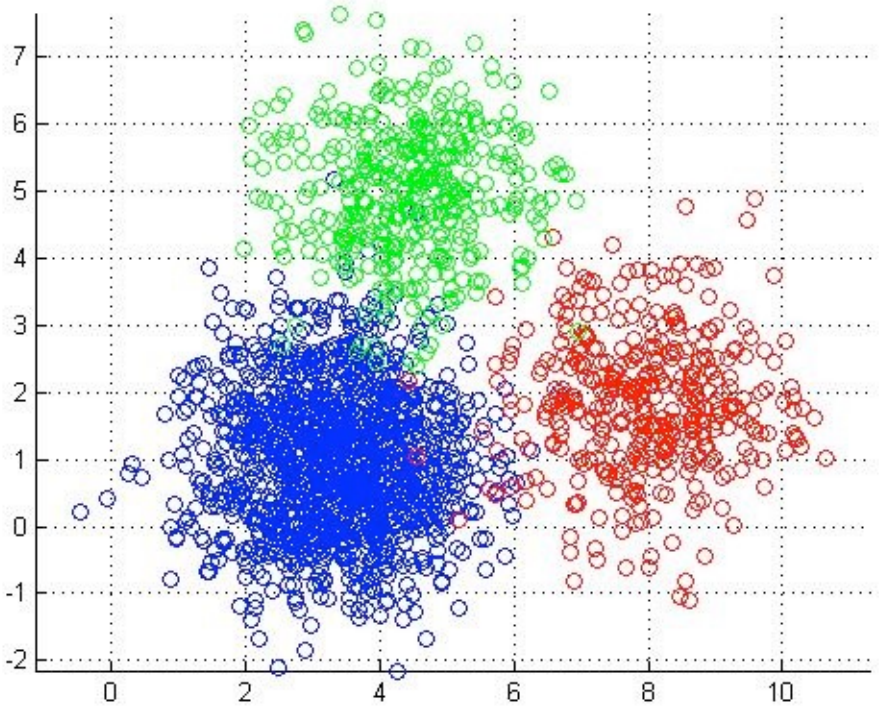
**Unsupervised**

- Assume we only have unlabeled data and we want to either:
  - label the data instances
  - or group up data instances into subsets sharing common characteristics
- One commonly application of unsupervised approach is clustering

- What do you think is the accuracy of unsupervised approaches compared to supervised?
- Why are unsupervised approaches important?

# CLUSTERING EXAMPLE

**Points are clustered together because they are closer to each other**

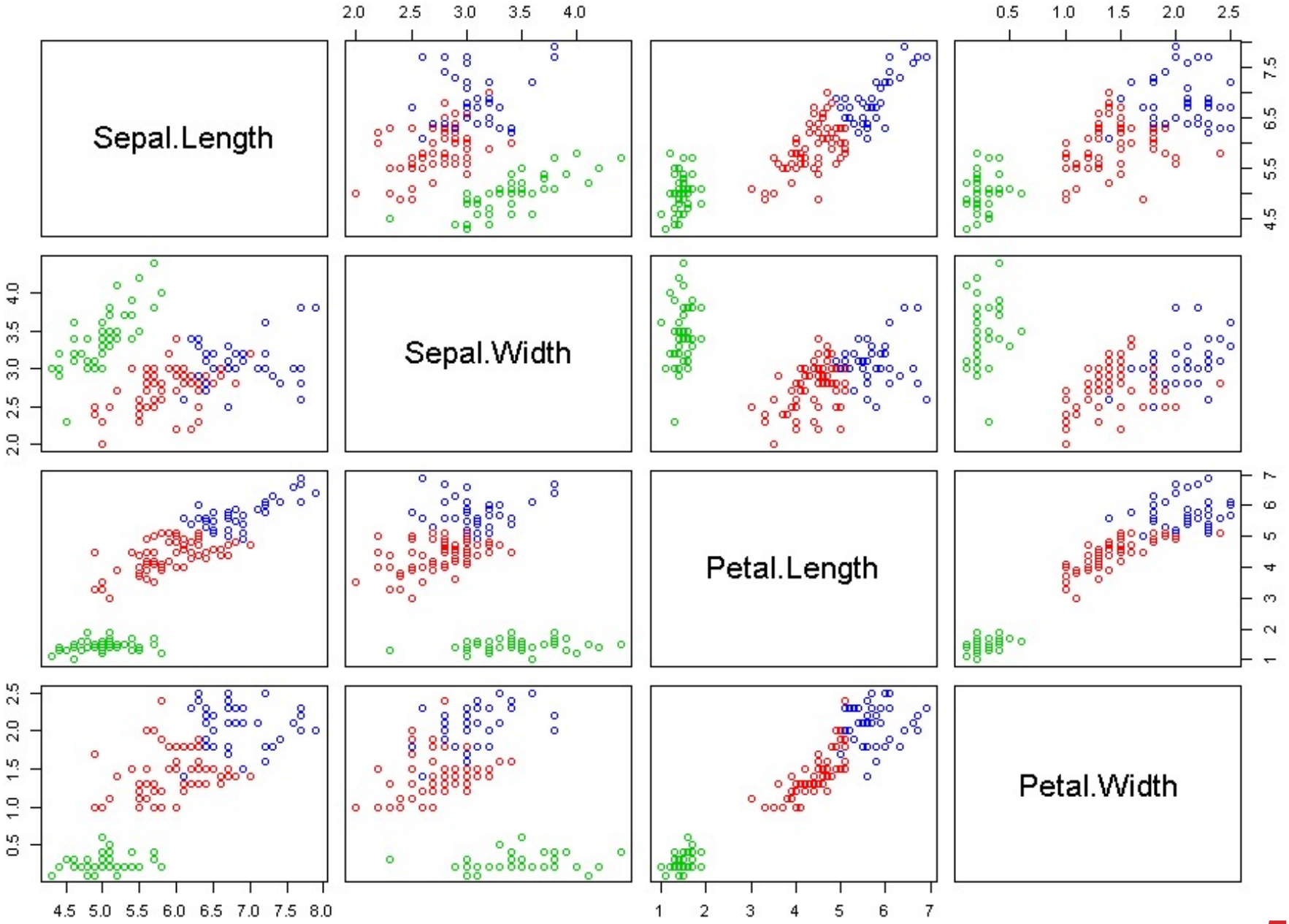


# UNSUPERVISED


## Clustering Example (Iris flower)

| Sepal Length | Sepal Width | Petal Length | Petal Width | Species |
|---|---|---|---|---|
| 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 4.9 | 3 | 1.4 | 0.2 | setosa |
| 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 6.2 | 2.2 | 4.5 | 1.5 | versicolor |
| 5.7 | 2.6 | 3.5 | 1 | versicolor |
| 5.5 | 2.4 | 3.8 | 1.1 | versicolor |
| 5.5 | 2.4 | 3.7 | 1 | versicolor |
| 7.2 | 3.6 | 6.1 | 2.5 | virginica |
| 6.5 | 3.2 | 5.1 | 2 | virginica |
| 6.4 | 2.7 | 5.3 | 1.9 | virginica |
| 6.8 | 3 | 5.5 | 2.1 | virginica |




Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
2.0 2.5 3.0 3.5 4.0
0.5 1.0 1.5 2.0 2.5
4.5 5.5 6.5 7.5
2.0 2.5 3.0 3.5 4.0
1 2 3 4 5 6 7
0.5 1.0 1.5 2.0 2.5
4.5 5.0 5.5 6.0 6.5 7.0 7.5 8.0
1 2 3 4 5 6 7

# APPLICATIONS OF ANALYTICS

# APPLICATIONS OF ANALYTICS

## Customer Relationship Management

- Maximize return on marketing campaigns
- Improve customer retention (churn analysis)
- Maximize customer value (cross-, up-selling)
- Identify and treat most valued customers

## Banking & Other Financial

- Automate the loan application process
- Detecting fraudulent transactions
- Maximize customer value (cross-, up-selling)
- Optimizing cash reserves with forecasting

# APPLICATIONS OF ANALYTICS

## Retailing and Logistics

- Optimize inventory levels at different locations
- Improve the store layout and sales promotions
- Optimize logistics by predicting seasonal effects
- Minimize losses due to limited shelf life

## Manufacturing and Maintenance

- Predict/prevent machinery failures
- Identify anomalies in production systems to optimize the use manufacturing capacity
- Discover novel patterns to improve product quality

# APPLICATIONS OF ANALYTICS

## Brokerage and Securities Trading

- Predict changes on certain bond prices
- Forecast the direction of stock fluctuations
- Assess the effect of events on market movements
- Identify and prevent fraudulent activities in trading

## Insurance

- Forecast claim costs for better business planning
- Determine optimal rate plans
- Optimize marketing to specific customers
- Identify and prevent fraudulent claim activities

# APPLICATIONS OF ANALYTICS

**Homeland security and law enforcement**

**Travel industry**

**Healthcare**

**Medicine**

**Entertainment industry**

**Sports**

**Etc**

# WHAT'S NEXT?

## Cross-Industry Standard Process for Data Mining (CRISP-DM)