# STAT40180/STAT40620
# Data Programming.
# Lab 9: Input/output and string manipulation

There are four files in the Blackboard folder named `file1.txt`, `file2.txt`, `file3.txt` and `file4.txt`. Save them somewhere on your computer and set your working directory to that location.

1. The file `file1.txt` contains a 100 by 200 matrix of random digits. However, it has a missing value (denoted by -99) and some random text at the top of the file. Each value is also delimited by a dollar sign. Read in the data using `read.delim`. Which row and column contain the missing value? (Hint: use `which`)

2. The file `file2.txt` contains 40 lines of random length text. Read in the data using `scan` so that the text is stored in a vector. When extracted, every third letter from each element of this vector makes up part of a sentence. What is that sentence?

3. The file `file3.txt` contains the html webpage output from the internet movie database top 250 movies (`http://www.imdb.com/chart/top`). Load in the data using `scan`. Which actor appears most in the top 250 movies Morgan Freeman, Jack Nicholson or Brad Pitt? (Hint: `grep` will be useful.)

4. A set of spam data are available at: `https://web.stanford.edu/~hastie/ElemStatLearn/datasets/spam.data`. Each row here represents an email and each of the first 48 columns give the percentage occurrence of a particular word in the email. There are then 9 further columns about the email (which we will ignore). The final column gives whether the data were considered spam or not (`1 = yes`, `0 = no`). The list of 48 words is given in the file `file4.txt`. Load in the spam data and `file4.txt`. Using the first 48 columns (i.e. the word frequencies) and the last column (whether spam or not) of the `spam` data and the word list, find out which word has the greatest difference in average frequency between spam and non-spam emails.

5. Use the function `getURL` in the library `RCurl` followed by `readHTMLTable` in the library `XML` to download the table of the winning team, beaten finalist etc of each All-Ireland Senior Football Championship final from: `https://en.wikipedia.org/wiki/List_of_All-Ireland_Senior_Football_Championship_finals` The table should be the 3rd tag in the resulting list when you run `readHTMLTable`. Which county has won the most All Ireland football finals? Produce a bar plot to illustrate the `Winner` variable. Which county is the most common finalist?