Rule 1: **Script everything**

# Structure of a Data Analysis

## 1. Steps in a data analysis

* Define the question
* Define the ideal data set
* Determine what data you can access
* Obtain the data
* Clean the data
* Exploratory data analysis
* Statistical prediction/modeling
* Interpret results
* Challenge results
* Synthesize/write up results
* Create reproducible code

## 2. Subsampling the dataset

> library(kernlab)

> data(spam)

> set.seed(3435)

> trainIndicator = rbinom(4601,size=1,prob=0.5)

> table(trainIndicator)

trainIndicator

0 1

2314 2287

> trainSpam = spam[trainIndicator==1,]

> testSpam = spam[trainIndicator == 0,]

> nrow(spam)

[1] 4601