# CodeBook

## Objectives

The purpose of this project is to demonstrate the ability to collect, work with, and clean a data set. The goal is to prepare tidy data that can be used for later analysis.

The data linked to from the course website represent data collected from the accelerometers from the Samsung Galaxy S smartphone. A full description is available at the site where the data was obtained:

[http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones]

Here are the data for the project:

[https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip]

## Data Processing

### Step 1: Merge the training and the test sets to create one data set

The data downloaded was read in with names corresponding to the source files, including: *x_test, y_test, subject_test, x_train, y_train, subject_train, features* and *activity*. The test and training data set were combined to create a new data set, namely *merged_dat*, with the size of 10299x563. Where:
* First column: Subject identification (ranging from 1 to 30) represents the identification of the subject carried out the expriment.
* Second column: Activity label (ranging from 1 to 6) numeric representation of the activity labels.
* Column 2-563: 561 features with time and frequency domain variables.

### Step 2: Extracts only the measurements on the mean and standard deviation for each measurement

The first and second column names in the *merged_dat* data set were changed to *SubjectID* (range from 1 to 30) and *Activity* (range from 1 to 6) respectively to conveniently index and extract the variables related to mean and standard deviation. A variable *mean_sd_index* was created, including variable names containing measuments on mean and standard deviation. The frequency mean was not taken into consideration, thus there are 68 variable names in total that refer to mean and standard deviationfor each measurement. The variable *mean_sd_index* was then used to extract corresponding values from *merged_dat*, whose result is stored in *mean_sd* data set.

### Step 3: Use descriptive activity names for activity measurements

By indexing the *activity* variable, the numeric representation of Activity in *merged_sd* data set was changed to character representation, corresponding to defined Activity labels. Specifically:
Activity
1 WALKING
2 WALKING_UPSTAIRS
3 WALKING_DOWNSTAIRS
4 SITTING
5 STANDING
6 LAYING

### Step 4: Appropriately labels the data set with descriptive variable names

By creating a temporary variable *col_names* containing the data set names, some modifications have been made for better description. Concretely:
- *f* and *t* were changed to *frequency* and *time* respectively, representing frequency domain and time domain signals.
- *mean* and *std* were uppercased for better readability.
Finally, the *col_names* was assigned back to the *mean_sd* data set to apply the changes to column names.

### Step 5: From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject

By applying **aggregate()** function, the *mean_sd* data set was extracted, grouping by *SubjectID* and *Activity* and applying the function *mean* to each variable. The result was stored in the new variable *tidy_dataset*, which then was applied with *order* fucntion to reorder the rows i.e. sorted by *SubjectID* first, then each Subject will have corresponding Activity observations. Finally, the output was written to the file *tidyDataset.txt*.

## Variables

- x_train = training set with 7352 observations of 561 variables.
- y_train = training lables with 7352 obsevations of 1 variable.
- x_test = test set with 2947 observations of 561 variables.
- y_test = test lablels with 2947 observations of 1 variable.
- subject_train = 7352 observations representing the identity of the person performed the experiment (in the training set), ranging from 1 to 30.
- subject_test = 2947 observations, representing the identity of the person performed the experiment (in the test set), ranging from 1 to 30.
- features = a 516-feature vector with time and frequency domain variables.
- actvity = activity lable, ranging from 1 to 6
- merged_train = merged set of x_train, y_train and subject_train (column-binded).
- merged_test = merged set of x_test, y_test and subject_test (column-bided).
- merged_dat = merged set of training and test sets, yeilded from row-binding the merged_train and merged_test
- mean_sd = data set containing measurements on mean and standard deviation only (not taking frequency mean into consideration).
- tidy_dataset = final tidy dataset, with average of each variable for each activity and subject, with 180 observations of 68 variables.

## Variable Names Description

The features selected for this database come from the accelerometer and gyroscope 3-axial raw signals tAcc-XYZ and tGyro-XYZ. These time domain signals (prefix 't' to denote time) were captured at a constant rate of 50 Hz. Then they were filtered using a median filter and a 3rd order low pass Butterworth filter with a corner frequency of 20 Hz to remove noise. Similarly, the acceleration signal was then separated into body and gravity acceleration signals (tBodyAcc-XYZ and tGravityAcc-XYZ) using another low pass Butterworth filter with a corner frequency of 0.3 Hz.

Subsequently, the body linear acceleration and angular velocity were derived in time to obtain Jerk signals (tBodyAccJerk-XYZ and tBodyGyroJerk-XYZ). Also the magnitude of these three-dimensional signals were calculated using the Euclidean norm (tBodyAccMag, tGravityAccMag, tBodyAccJerkMag, tBodyGyroMag, tBodyGyroJerkMag).

Finally a Fast Fourier Transform (FFT) was applied to some of these signals producing fBodyAcc-XYZ, fBodyAccJerk-XYZ, fBodyGyro-XYZ, fBodyAccJerkMag, fBodyGyroMag, fBodyGyroJerkMag. (Note the 'f' to indicate frequency domain signals).

These signals were used to estimate variables of the feature vector for each pattern:
'-XYZ' is used to denote 3-axial signals in the X, Y and Z directions.

- tBodyAcc-XYZ
- tGravityAcc-XYZ
- tBodyAccJerk-XYZ
- tBodyGyro-XYZ
- tBodyGyroJerk-XYZ
- tBodyAccMag
- tGravityAccMag
- tBodyAccJerkMag
- tBodyGyroMag
- tBodyGyroJerkMag
- fBodyAcc-XYZ
- fBodyAccJerk-XYZ
- fBodyGyro-XYZ
- fBodyAccMag
- fBodyAccJerkMag
- fBodyGyroMag
- fBodyGyroJerkMag

The set of variables that were estimated from these signals are:

- mean(): Mean value
- std(): Standard deviation