# Getting and Cleaning Data Project Code Book

This code book describes the process and variables used to clean up the UCI acceleration data.

## About the Data

* The data can be downloaded from <https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip>

A summary from the data's README.txt is included below.

The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING\_UPSTAIRS, WALKING\_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain. See 'features\_info.txt' for more details.

For each record it is provided:

* Triaxial acceleration from the accelerometer (total acceleration) and the estimated body acceleration.
* Triaxial Angular velocity from the gyroscope.
* A 561-feature vector with time and frequency domain variables.
* Its activity label.
* An identifier of the subject who carried out the experiment.

The dataset includes the following files:

* 'README.txt'
* 'features\_info.txt': Shows information about the variables used on the feature vector.
* 'features.txt': List of all features.
* 'activity\_labels.txt': Links the class labels with their activity name.
* 'train/X\_train.txt': Training set.
* 'train/y\_train.txt': Training labels.
* 'test/X\_test.txt': Test set.
* 'test/y\_test.txt': Test labels.

The following files are available for the train and test data. Their descriptions are equivalent.

* 'train/subject\_train.txt': Each row identifies the subject who performed the activity for each window sample. Its range is from 1 to 30.
* 'train/Inertial Signals/total\_acc\_x\_train.txt': The acceleration signal from the smartphone accelerometer X axis in standard gravity units 'g'. Every row shows a 128 element vector. The same description applies for the 'total\_acc\_x\_train.txt' and 'total\_acc\_z\_train.txt' files for the Y and Z axis.
* 'train/Inertial Signals/body\_acc\_x\_train.txt': The body acceleration signal obtained by subtracting the gravity from the total acceleration.
* 'train/Inertial Signals/body\_gyro\_x\_train.txt': The angular velocity vector measured by the gyroscope for each window sample. The units are radians/second. ##Required libraries *reshape* reshape2

## About the Data Cleaning

The project had essentially five subproblems. The problems are outlined and explained below.

1. Merges the training and the test sets to create one data set.

* To merge these two datasets, the data needed to be appropriately labeled. The addition of the y\_test, subject\_test, y\_train and subject\_train give the X\_test and X\_train data id labels as well as which activity is being performed. The data can then be simply column-bound.

2)Extracts only the measurements on the mean and standard deviation for each measurement.

\*The features file illustrates what the columns in the X\_test and X\_train data actually represent. They are column names. The problem requires the capture of variables reporting means and standard deviations. You can find which variables report mean or standard deviations by using a regular expression to find which column names have either "mean" or "std" in the names in the features data.

1. Uses descriptive activity names to name the activities in the data set

\*This data explicitly comes from the activity labels dataset. I matched the activity label to the subject test and subject train variables.

1. Appropriately labels the data set with descriptive variable names.

\*The data are appropriately labeled, in fact labeled from the names in the features dataset. No creative license was taken.

1. 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.

\*The average from the data in step 4 is calculated by casting the previously melted data and using the mean function as one of the arguments to dcast.