# Proposal

## Problem Description

Wearable technologies are getting more and more common and machine learning plays an important role in enabling the machine to recognize a person's activity. While typical machine learning projects seeks to classify the activity performed, this project attempts to go a step further and aims to develop a method to best classify the activity and the person performing the activity. Using the T,P,E framework, the problem can be summarized into:

* **Task** - Predict the activity *and* the person performing the activity
* **Performance** - Number of actions *and* person performing the activity correctly classified
* **Experience** - PAMAP2 data set of labeled IMU readings available from the UCI Machine Learning Repository

The constructed model may be deployed in a health care setting, to monitor a patient's activities. With the ability to differentiate the activities and the person performing it, it also opens up the possibility of sharing the wearable device.

## Approach

The approach that we propose would be to first explore in detail to extract the features most representative of the activities and the subject. Next, 3 different models will be compared in the classification of human activity and the person performing it. The 3 models that will be compared are summarised as follows:

1. **Model 1** - Classify Subject (Person) --> Feed subject back into model to classify action of the subject
2. **Model 2** - Classify Action --> Feed action back into model to the classify subject
3. **Model 3** - Classify both subject and action simultaneously

The most suitable model (in terms of accuracy, precision and recall) to classify an activity that is carried out by a unique individual will be selected. Supervised learning methods will explored and used to construct the model. The model will then be interpreted to extract insights on how are the actions and subjects classified.

Hold-out or k-fold cross validation will be used to avoid over-fitting the model. The final cross validation method will be decided later. Source control will be done using Github.

## Data Set

The PAMAP2 data set available from UCI Machine Learning Repository [(Link)](https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring) consist the data of 9 subjects performing 18 different physical activities such as sitting, lying, walking and running in a controlled environment - all went through the exact same sequence of activities with optional activities performed by some of the subjects. Optional activities performed but will not be in the scope of this project. The subjects (8 males, 1 female – ages: 27-31 years old) who participated were mainly employees or students at the German Research Center for Artificial Intelligence (DFKI). The data collection was done by attaching 3 Colibri wireless inertial measurement units (IMUs) to each subject's body.

The full data set exists in 9 separate .dat files, one for each subject, of 54 columns containing data such as Timestamps, Activity IDs, Heart Rate and IMU readings on the subjects' hand, chest and ankle. Each IMU sensory data (hand, chest and ankle) contains the temperature, 3D acceleration data, 3D gyroscope data, 3D magnetometer data. Orientation data is also included in the data set but invalid in this data set as mentioned in the code book available with the PAMAP2.

## Initial Data Exploration & Preparation

As the data set exists in 9 separate .dat files a script was prepared to read the data and consolidate it into one .txt file for easier processing. As an one-time filtering, activities that are performed by less than 6 subjects and activities performed for only a few seconds (i.e 24 - Rope Jumping) were removed. Optional activities resides in separate data files and are not read. Each of the record was also labeled with the subject performing the activity as part of the initial and consolidation.

A summary of the number of records and columns is provided below.

* Total Records Read From Original Data Set: 2872533
* Total Records Written to Consolidated Data Set: 1893512
* Total Records Filtered: 979021
* Total Columns in Consolidated Data Set: 55

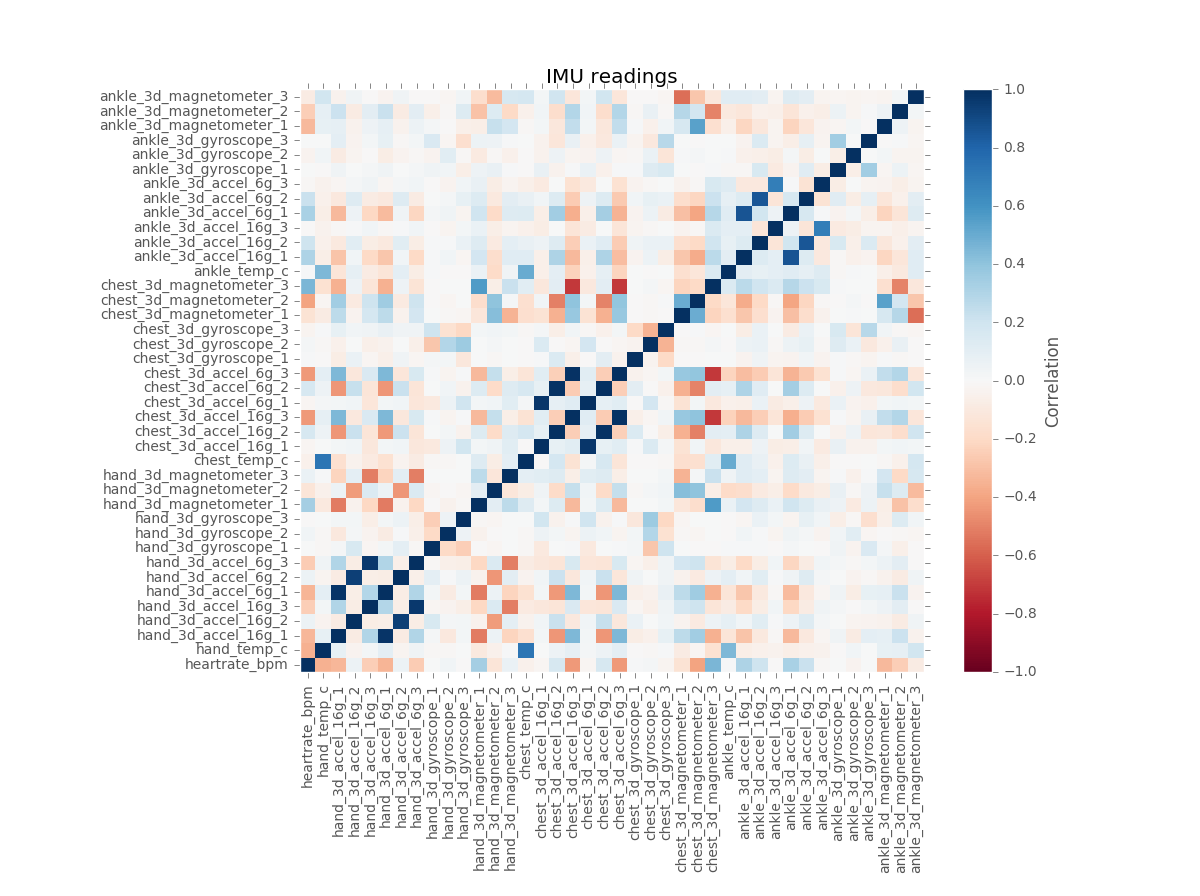
A summary of the data was performed using the Python pandas package and initial exploration of the data set reveals several important points that will provide the basis for training the models:

|  |  |
| --- | --- |
| **Exploration item** | **Remarks & Potential implication** |
| Subject 9 has significantly lower amounts of recorded data than the other subjects | Including the subject will severely distort the analysis as he is an obvious outlier. As such, we exclude subject 9 from our analysis. |
| The data set consists of a significant amount of NaN data | Cannot conduct PCA with missing values. There are missing values mainly because of technical issues and the fact that heart rate and other recorded variables were measured at different frequencies. As such, we have to handle the missing data. |
| Timestamps are available | Including the timestamp in the analysis will cause the model to be biased - it is possible to predict results solely based on timestamps as subjects performed the activities in a sequence. As such, we have to exclude timestamps. |
| Some activities were only carried out by a few subjects (e.g. watching TV was only carried out by 1 subject) | Activities carried out by less than 6 out of 8 subjects should be left out in order to train the model properly. |

Initial exploration also includes plotting the correlation of the variables and performing principal component analysis. The following steps were taken to conduct PCA on the data set as part of exploration.

1. Run an initial PCA on the dataset (excluding the timestamp and activityID)
2. Find out which columns affect the components the most
3. Normalize those columns
4. Re-run PCA with normalized columns and plotting the scree plot

Both correlation and the scree plot are shown below. Further exploration will be needed to look at the data and in particular the principal components and the highly correlated variables. Any implications will then need to be considered when constructing the models.

**Correlation Matrix of the variables** 

**Scree Plot of Principal Components** 