Team `EInstEIn` DREAM Challenge Writeup (1.1 mPower): Gait Features to Predict Parkinson's Disease (PD)

Research Question

Parkinson's Disease (PD) is a debilitating idiopathic neurodegenerative disorder. In the mPower study [2, 7], walking data was collected on iPhones using an open protocol that varied over time. Can features be extracted from the motion data collected via Apple's ResearchKit and specifically the CoreMotion sensors to identify a protocol record as being performed by a participant with PD?

1 Background

1.1 Protocol

An unmonitored participant turns on the iPhone app, walks in a "straight line" for an amount of time (*outbound*), turns around, and walks back (hopefully) on the same path (*return*). In one version of the protocol, the participant would rest between the outbound task and return task; in another version, the participant would rest after the outbound and return tasks (*rest*).

We make an assumption the participant is walking on a flatish surface which may help with dead-reckoning drift problems based on the accelerometer data.

The summary of the protocol has been explained as follows:

```
[Larsson]
We had to do a little digital archeology - there have been 29 different versions of the app. And
     although most of them used overlapping instructions we have found at least three different
     ones:
   1. The one in the video - walk for 20 steps (but timer is also running for 30s) turn around and
        wait for 30s then walk back
   2. Walk for 20 steps (but timer is also running for 30s) turn around walk back for 20 steps
        (also with 30 s timer) followed by standing for 30s
   3. Walk for 30s "back and forth" then turn around 360 degrees then stand still for 30s


It is possible that there are small changes in between these ones as well.
```

1.2 Protocol Deviations: Relation to PD

In the theoretical frame, this difference in protocols may matter. It is know that PD sufferers many times have a hard time "getting started" and once started have a hard time "stopping." With this in mind the protocol: outbound-rest-return may inhibit mobility whereas the protocol: outbound-return-rest may encourage mobility.

In fact, the entire mPower design is anchored to a adaptation heuristic rather than scientific methodology. Rather than asking "what gait characteristics could we extract from wearable sensors that would better help understand PD?" the frame was "can we build an iPhone app that can extract some data about motion that we can somehow use to assess PD gait?"

1.3 Concern with mPower DREAM Challenge

One could consider this a constrained-optimization problem, but given that the initiating principle investigator has left Sage Networks to work for Apple there are a few other concerns to explicitly state [3, 5, 4].

Conflict of Interest

It is worth noting that the mPower protocol only used iOS and not the Android platform. Now the initiatiating PI works for Apple.

2 Data Manipulation

2.1 Harvesting and Tabulating

Using a single CORE and the Synapse Login, I have accessed and downloaded the training data. The Synapse caching system took about 80GB of data, and the reorganization for my use and the creating of new `R` objects necessitated an additional 100GB of storage (mostly redundant). I created a folder for each participant (an hash of the healthCode), and therein I create a subfolder for each recordId. For each record, I collect the appropriate JSON files and organize into a timeseries panel. I cache this original object as a raw list, (`rlist`).

```
str(rlist);
List of 3
 $ outbound:List of 3
  ..$ accel     :'data.frame':     2443 obs. of 4 variables:
  .. ..$ y        : num [1:2443] -0.752 -0.787 -0.858 -0.893 -0.92 ...
  .. ..$ timestamp: num [1:2443] 1156 1156 1156 1156 1156 ...
  .. ..$ z        : num [1:2443] 0.0453 0.0699 0.1285 0.1772 0.2414 ...
  .. ..$ x        : num [1:2443] -0.725 -0.653 -0.598 -0.542 -0.512 ...
  ..$ deviceMotion:'data.frame':    2362 obs. of 6 variables:
  .. ..$ attitude      :'data.frame': 2362 obs. of 4 variables:
  .. .. ..$ y: num [1:2362] -0.498 -0.49 -0.482 -0.475 -0.469 ...
  .. .. ..$ w: num [1:2362] 0.691 0.689 0.688 0.686 0.684 ...
  .. .. ..$ z: num [1:2362] -0.00292 -0.00849 -0.01385 -0.01935 -0.02516 ...
  .. .. ..$ x: num [1:2362] 0.524 0.533 0.542 0.551 0.558 ...
  .. ..$ timestamp     : num [1:2362] 1156 1156 1156 1156 1156 ...
  .. ..$ rotationRate  :'data.frame': 2362 obs. of 3 variables:
  .. .. ..$ x: num [1:2362] 0.763 0.866 0.895 0.795 0.632 ...
  .. .. ..$ y: num [1:2362] 0.27589 0.33625 0.29125 0.17494 -0.00668 ...
  .. .. ..$ z: num [1:2362] -2.45 -2.46 -2.42 -2.32 -2.14 ...
  .. ..$ userAcceleration:'data.frame': 2362 obs. of 3 variables:
  .. .. ..$ x: num [1:2362] -0.0971 -0.1349 -0.1545 -0.1151 -0.0617 ...
  .. .. ..$ y: num [1:2362] 0.305 0.263 0.17 -0.07 -0.146 ...
  .. .. ..$ z: num [1:2362] -0.0264 -0.0264 -0.0792 -0.1128 -0.1267 ...
  .. ..$ gravity       :'data.frame': 2362 obs. of 3 variables:
  .. .. ..$ x: num [1:2362] -0.685 -0.667 -0.648 -0.631 -0.614 ...
  .. .. ..$ y: num [1:2362] -0.728 -0.744 -0.759 -0.774 -0.787 ...
  .. .. ..$ z: num [1:2362] 0.0452 0.0491 0.0538 0.0588 0.0638 ...
  .. ..$ magneticField :'data.frame': 2362 obs. of 4 variables:
  .. .. ..$ y      : int [1:2362] 0 0 0 0 0 0 0 0 0 0 ...
  .. .. ..$ z      : int [1:2362] 0 0 0 0 0 0 0 0 0 0 ...
  .. .. ..$ x      : int [1:2362] 0 0 0 0 0 0 0 0 0 0 ...
  .. .. ..$ accuracy: int [1:2362] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 ...
  ..$ pedometer  :’data.frame’:      3 obs. of 6 variables:
  .. ..$ floorsAscended : int [1:3] 0 0 0
  .. ..$ floorsDescended: int [1:3] 0 0 0
  .. ..$ endDate        : chr [1:3] "2015-03-09T15:40:18-0400" "2015-03-09T15:40:21-0400"
      "2015-03-09T15:40:23-0400"
  .. ..$ startDate      : chr [1:3] "2015-03-09T15:39:59-0400" "2015-03-09T15:39:59-0400"
      "2015-03-09T15:39:59-0400"
  .. ..$ numberOfSteps : int [1:3] 16 18 21
  .. ..$ distance      : num [1:3] 13 14.4 16.5
 $ return :List of 3
# [...] and so on
 $ rest   :List of 3
# [...] and so on
  ..$ pedometer  : list()
```

2.2 Aligning Design Points and Pedometer Calculations

Thereafter, the data is aligned based on the timestamps to further flatten into panel form (`tlist`):

```
str(tlist);
List of 3
 $ outbound:List of 3
  ..$ accel       :’data.frame’:      2443 obs. of 4 variables:
  .. ..$ timestamp: num [1:2443] 1156 1156 1156 1156 1156 ...
  .. ..$ x        : num [1:2443] -0.725 -0.653 -0.598 -0.542 -0.512 ...
  .. ..$ y        : num [1:2443] -0.752 -0.787 -0.858 -0.893 -0.92 ...
  .. ..$ z        : num [1:2443] 0.0453 0.0699 0.1285 0.1772 0.2414 ...
  ..$ deviceMotion:’data.frame’:    2362 obs. of 14 variables:
  .. ..$ timestamp: num [1:2362] 1156 1156 1156 1156 1156 ...
  .. ..$ dax      : num [1:2362] 0.524 0.533 0.542 0.551 0.558 ...
  .. ..$ day      : num [1:2362] -0.498 -0.49 -0.482 -0.475 -0.469 ...
  .. ..$ daz      : num [1:2362] -0.00292 -0.00849 -0.01385 -0.01935 -0.02516 ...
  .. ..$ daw      : num [1:2362] 0.691 0.689 0.688 0.686 0.684 ...
  .. ..$ drx      : num [1:2362] 0.763 0.866 0.895 0.795 0.632 ...
  .. ..$ dry      : num [1:2362] 0.27589 0.33625 0.29125 0.17494 -0.00668 ...
  .. ..$ drz      : num [1:2362] -2.45 -2.46 -2.42 -2.32 -2.14 ...
  .. ..$ dux      : num [1:2362] -0.0971 -0.1349 -0.1545 -0.1151 -0.0617 ...
  .. ..$ duy      : num [1:2362] 0.305 0.263 0.17 -0.07 -0.146 ...
  .. ..$ duz      : num [1:2362] -0.0264 -0.0264 -0.0792 -0.1128 -0.1267 ...
  .. ..$ dgx      : num [1:2362] -0.685 -0.667 -0.648 -0.631 -0.614 ...
  .. ..$ dgy      : num [1:2362] -0.728 -0.744 -0.759 -0.774 -0.787 ...
  .. ..$ dgz      : num [1:2362] 0.0452 0.0491 0.0538 0.0588 0.0638 ...
  ..$ pedometer  :’data.frame’:      3 obs. of 16 variables:
  .. ..$ floorsAscended      : int [1:3] 0 0 0
  .. ..$ floorsDescended     : int [1:3] 0 0 0
  .. ..$ endDate             : chr [1:3] "2015-03-09T15:40:18-0400" "2015-03-09T15:40:21-0400"
      "2015-03-09T15:40:23-0400"
  .. ..$ startDate           : chr [1:3] "2015-03-09T15:39:59-0400" "2015-03-09T15:39:59-0400"
      "2015-03-09T15:39:59-0400"
  .. ..$ numberOfSteps       : int [1:3] 16 18 21
  .. ..$ distance            : num [1:3] 13 14.4 16.5
  .. ..$ deltaDistancePerSecond: num [1:3] 0.682 0.473 1.05
  .. ..$ deltaStepsPerSecond : num [1:3] 0.842 0.667 1.5
  .. ..$ deltaDistance      : num [1:3] 12.96 1.42 2.1
  .. ..$ deltaSteps         : num [1:3] 16 2 3
  .. ..$ deltaTime          : num [1:3] 19 3 2
  .. ..$ distancePerSecond  : num [1:3] 0.682 0.654 0.687
  .. ..$ stepsPerSecond     : num [1:3] 0.842 0.818 0.875
  .. ..$ diffU              : num [1:3] 19 22 24
  .. ..$ endU               : num [1:3] 1.43e+09 1.43e+09 1.43e+09
  .. ..$ startU             : num [1:3] 1.43e+09 1.43e+09 1.43e+09
 $ return :List of 3
# [...] and so on
 $ rest :List of 3
# [...] and so on
  ..$ pedometer  : list()
```

Above you will also note that the pedometer data was manipulated to capture (delta) distance per second and (delta) steps per second. The full timestamp was available, so it may be possible to match the partial timestamps in the other data files to the pedometer data. Since the pedometer data is not as robust, we can use it for approximate confirmation as we try to count steps. Below, we will describe some features extracted from this data to capture the PD characteristic of *shuffling*.

2.3 Smoothing Design Points

It is well documented that the accelerometer for the iPhone provides spurious data at times. In addition the time measurements are not precise. For these reasons, I have chosen to smooth the data and create equal-distant design points.

The data was provided in 10ms increments. So if I sample to about 100ms, I will have about 10 observations per redesign frame. I have written the code to allow for customization in the degree of granularity (`scale`). For the cumulative observations with the scale, I define a median value. I emphasize that this approach creates equal and consistent design points (e.g., 100ms, 200ms, 300ms, ...).

Further research could address optimal granularity, but with this setup (100 ms), I can capture 10Hz motion. Most human-body motion is in the 2-4Hz range, and tremors tend to also be below 5Hz (citations needed).

```
   slist = scaleToTimeIncrement(tlist,designpoint);
```

In addition, the number of points used to create the new design frame are also recorded within the new object (`slist`):

```
> str(slist)
List of 3
 $ outbound:List of 2
  ..$ accel     :'data.frame':      238 obs. of 5 variables:
  .. ..$ timestamp: num [1:238] 1156300 1156400 1156500 1156600 1156700 ...
  .. ..$ x      : num [1:238] -0.725 -0.602 -0.746 -0.565 -0.117 ...
  .. ..$ y      : num [1:238] -0.752 -0.859 -0.59 -0.676 -0.666 ...
  .. ..$ z      : num [1:238] 0.0453 0.2115 -0.0253 0.0706 0.1497 ...
  .. ..$ points : num [1:238] 1 8 9 10 9 9 9 10 9 9 ...
# [...] and so on
```

If the above smoothing feature doesn't have enough data (the first few observations, I truncate those elements from the panel) using a (`pareto=0.8`) Pareto parameter. In the example above the first value has

1 data point so it is discarded, the second has 8, which is exactly at the 0.8 threshold and is kept. All values thereafter (9,10,9,9,9,10, ...) are also kept.

This unifies the userAcceration data with the deviceMotion data with common timestamps, which I merge into a common object (`mlist`):

```
  mlist = mergeListsAccelDeviceMotion(slist);
```

2.4 Orienting

Next, I use the deviceMotion$gravity coordinates to re-orient the userAcceleration coordinates *at every* `scale` *time interval.* At each iteration, I compute an angle difference from the previous adjusted coordinate. I will use the angle value as “acceleration angles” in further feature analysis.

```
olist = orientToGravity(mlist);
str(olist);
List of 3
 $ outbound:'data.frame':    235 obs. of 5 variables:
  ..$ timestamp: num [1:235] 1156600 1156700 1156800 1156900 1157000 ...
  ..$ x        : num [1:235] 0.398 0.203 0.145 0.231 0.151 ...
  ..$ y        : num [1:235] 0.474 0.101 0.179 -0.165 -0.033 ...
  ..$ z        : num [1:235] 0.631 0.655 0.856 1.862 0.913 ...
  ..$ angle    : num [1:235] 129.84 27.75 8.15 16.86 3.79 ...
 $ return  :'data.frame':    299 obs. of 5 variables:
  ..$ timestamp: num [1:299] 1180800 1180900 1181000 1181100 1181200 ...
  ..$ x        : num [1:299] 0.0644 0.0573 0.0625 0.0557 0.0488 ...
  ..$ y        : num [1:299] -0.284 -0.278 -0.276 -0.282 -0.278 ...
  ..$ z        : num [1:299] 0.961 0.962 0.963 0.96 0.964 ...
  ..$ angle    : num [1:299] 74.03 0.525 0.335 0.546 0.512 ...
 $ rest    :'data.frame':    298 obs. of 5 variables:
  ..$ timestamp: num [1:298] 1211300 1211400 1211500 1211600 1211700 ...
  ..$ x        : num [1:298] 0.216 0.215 0.215 0.212 0.22 ...
  ..$ y        : num [1:298] -0.258 -0.254 -0.253 -0.253 -0.252 ...
  ..$ z        : num [1:298] 0.944 0.945 0.947 0.948 0.943 ...
  ..$ angle    : num [1:298] 65.844 0.206 0.104 0.153 0.515 ...
```

2.5 Dead reckoning

Finally, I can perform the first integral to get velocity and the second integral to get position. Rather than using the Synapse *mPower* approach ($\frac{1}{2}at^2$), I choose to do a numerical approach using the trapezoidal method.

I also am able to stitch the separate events (outbound, rest, return) based on the relative timestamps to determine the order of the record’s protocol. I will use this as a control feature (I originally intended to use this to compute a compliance feature). Ultimately, due to the lack of compliance/control, I capture this information but at this time do not use it in my feature selection.

```
  morder = determineOrder(mlist);
    # should be outbound, rest, return [compliance]
```

where default=0 (outbound-rest-return) and other=1 (outbound-return-rest) or other=-1 (rest-outbound-return).

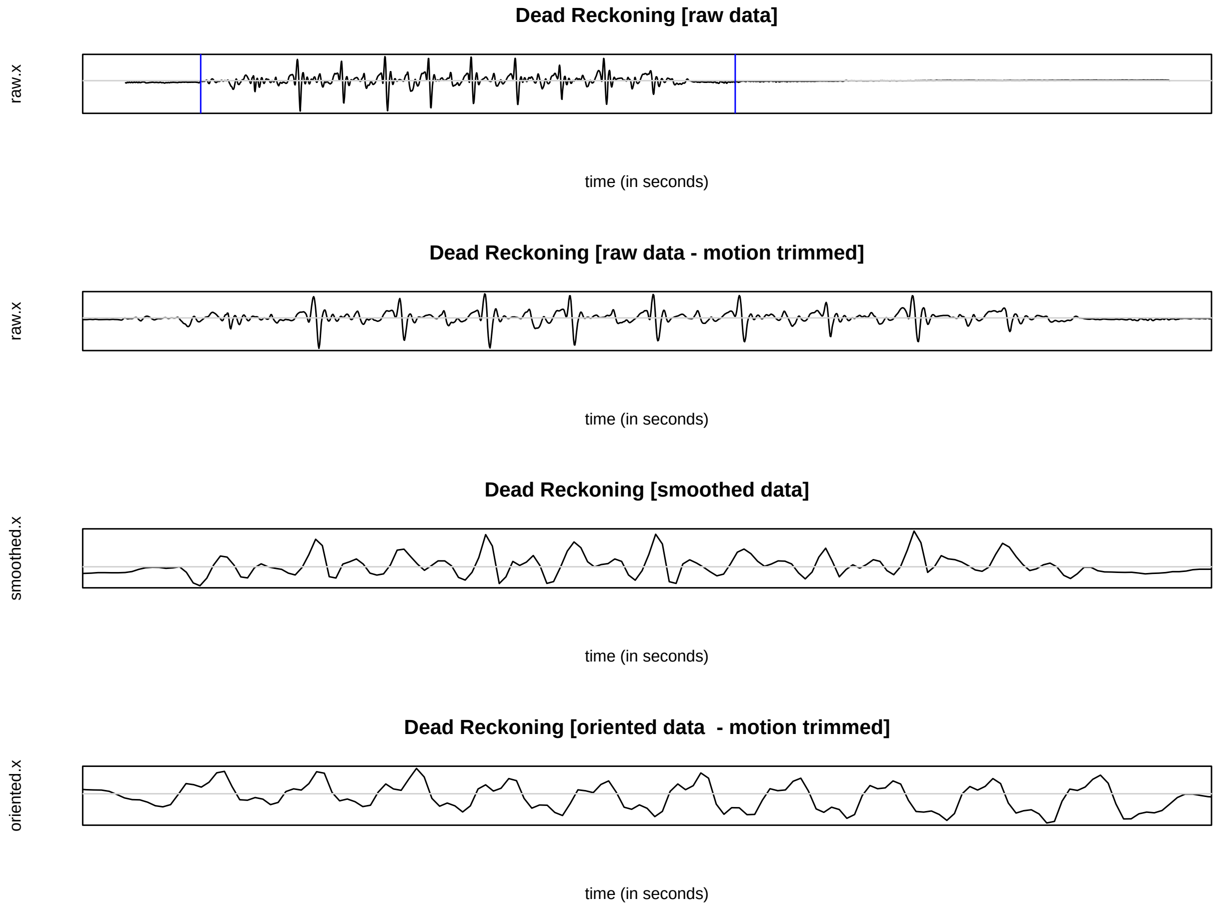

Figure 1: Overview of the data cleansing protocols at the dead-reckoning stage. In the top graph, the raw data is reckoned and a motion-region is determined (vertical lines), which is trimmed in the second graph. With this time-trimmed window, the data is smoothed using equal design points. In the final graphy the data is oriented to gravity. This final graph appears to have several sinusoidal functions that could be modeled using fourier transforms. From such transformations, a full gait cycle could possibly be modeled (e.g., PD participants should have small 'swing phase' in their gait cycle). However, the variables immediately extracted will be based on frequency of extreme changes in this graph.

2.6 Recycle and Update with Kalman or Other Filter

With the pedometer data (possibly stitched in) and a knowledge of the gait cycle (e.g., flat-floor hypothesis or moments of zero velocity), we could potentially recycle and update the panel data using dead-reckoning approaches to prevent drift with some filtering approach. This is tabled until a future time. The idea would be to utilized bounding constraints of human biomechanics to prevent drift.

3 Feature Extraction for Walking Exercise

It is worth noting at the beginning of this section that features to distinguish PD sufferers from healthy nonPD controls will generally be related to variability in walking characteristics. Since the data is inherently skewed, and dead-reckoning may lead to some systematic error (due to drift), variablility will be measured using a median, an interquartile range (IQR), a MAD (median absolute deviation), and/or a MAD-adjusted sample variance ($\hat{\sigma}$).

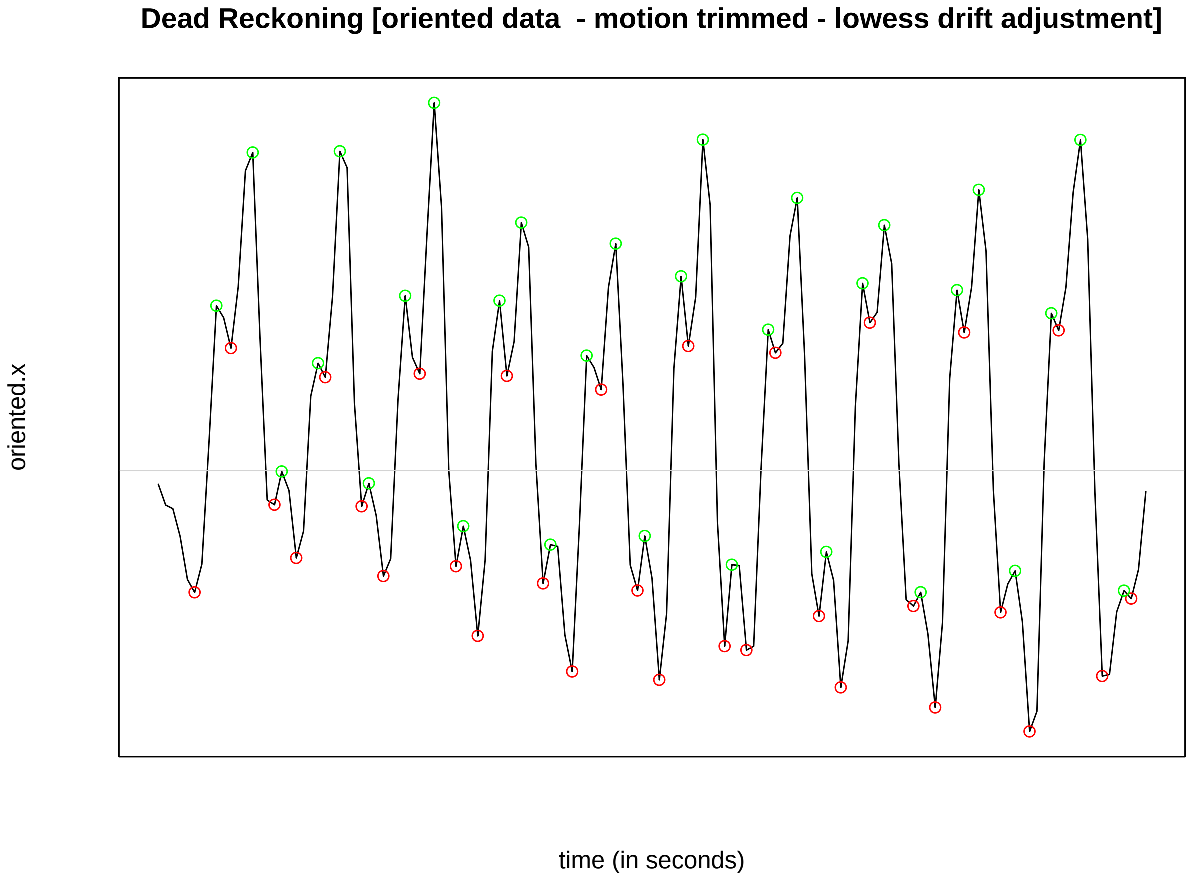

Figure 2: Identifying `turningpoints` from cleansed position data (e.g., "x" dimension). Frequency and volatility of turning points will be used to assess the overall motion of the participant. These points represent the minor cycle (frequency and amplitude). I further scan these values to identify a major cycle with frequency and amplitude features. I create a graph of each dimension (x,y,z) and each movement segment: OUTBOUND, RETURNWALK, RESTING and store in the RECORD folder.

3.1 Pedometer Data

In the PD literature, it has been observed that those suffering from PD have a unique cadence. Since data on height (which would correlate with typical stride length) is not available, we intentionally normalize stride information within an activity sample (e.g., outbound is different from return).

PD sufferers will have more frequent steps per second and travel a shorter overal distance per second. Together, this is how we can define the *cadence*. Deviations of cadence are determined below:

```
dps = range(tlist$outbound$pedometer$distancePerSecond);
   dpsMin = dps / min(dps);
      dpsMinScore = abs(dpsMin[2]-dpsMin[1]);
   dpsMax = dps / max(dps);
      dpsMaxScore = abs(dpsMax[2]-dpsMax[1]);
dps;
   dpsMin;
   dpsMinScore;
   dpsMax;
   dpsMaxScore;

sps = range(tlist$outbound$pedometer$stepsPerSecond);
   spsMin = sps / min(sps);
      spsMinScore = abs(spsMin[2]-spsMin[1]);
   spsMax = sps / max(sps);
      spsMaxScore = abs(spsMax[2]-spsMax[1]);
sps;
   spsMin;
   spsMinScore;
   spsMax;
   spsMaxScore;
```

3.2 Hip rotation

The bio-mechanics of hip rotation are well established [1, 6]. In this study, because the location of the phone in relationship to the hip is unknown (front or back pocket, left or right side, highly affective or non-highly affective side for PD patients, or possibly in the participants hand or bag), the degree of the rotation is not possible to calculated. However, the range of motion, normalized to the individual record, is what may be captured.

The PD feature of interest is *rigidity*. The hypothesis is that in general hip rotation will be lower; the PD participant will have a lower range of angular motion during the walk.

```
te = getMAD(diff(olist$outbound$angle));
   str(te);
te = getMAD(diff(olist$return$angle));
   str(te);
te = getMAD(diff(olist$rest$angle));
   str(te);
```

These results may be protocol-contingent. I would hypothesize that for the outbound-return-rest protocol PD participants will warm up, and the MAD from return will be higher than outbound. Conversely, for outbound-rest-return, the rest phase had the PD patients ‘cool down’ so there would not be any significant difference between outbound and return.

For nonPD participants, I would hypothesis than in either scenario, the first walking exercise would have a higher MAD than the second MAD exercise as they find a rhythm and stride on the known return path.

I hypothesis that the range for PD sufferers to nonPD control participants will be most noticeable in the rest phase. Natural sway and weight shifting will be prevalent with nonPD participants. The PD rigidity will be apparent during rest.

3.3 Medication

PD participants are "on" if they just took medication to control mobility issues. They are "off" if they are coming down from a medication dosage. The could also be somewhere in between. Control participants are, in general, probably more ambulatory in general, and there is a group of participants that did not answer.

```
medTimepoint
```

I leverage this information to create a pseudo-continuous scale to use as a medication feature: 0="off" medication, 1=between, 2="on" medication, 3=[blank/unknown], 5=no medication.

4 Submission

4.1 Imputation

A common practice in Statistics is to cleanse the data set removing bad records. Thereafter, it is common to create a hold-out sample. In machine learning, the model is determined with the training sample and the hold-out sample is generally defined as the test sample. Different terminology for similar methodological protocols.

In this challenge, the data set was not cleansed, so I also have to address bad records. To do so, I did a three-tier step-wise approach based on median values of relevant subsets of the data.

The first step is within-subject imputation. That is, if at least one complete record for the subject exists, the information will be extracted from the collection of records to create a within-subject median score for any missing values.

The second step is within-medication imputation. This only occurs if the previous imputation method could not be implemented. The collection of records derived with be from those within the same medication group (and possibly further subsetted based on age / gender). [NOT used]

The third and final step is a within-collection imputation. Again, a median will be used but this group will be the entire collection of good records.

The iterative process is nested meaning if record #5 was imputed using a first-step approach that update is included in a record #23 imputation using a third-step approach.

If any part of a record was imputed, the feature *imputed* will be 1; 0 otherwise. [NOT implemented yet]

4.2 Design-point Experiment

Based on the outlined data processing/smoothing methodologies described above, the sets of features submitted will used fixed design points: observations every 100ms.

For practicality, all features are included in one submission and the "best features" in training are included in a second submission using hierarchical stepwise selection approach in function:

```
stepwiseFeatureSelection
```

where we try to remove multicollinearity issues. Future work could expand on interaction effects and so on using other nonhierarchical approach. Sensitivity analysis is important. An ideal model may be a function of what data is available. If all data, do X; if only outbound, do Y; if outbound and rest, do Z. Future work may need to consider HMM or some other approach (PCA, interaction effects) because several features are positive individually, but once a few are included, not much more can be extracted.

Outliers are also dampened based on the IQR, things too far afield are assigned a truncated min or max value.

```
dampenOutliers
```

Benchmarks

The baseline ‘random’ standard in the ‘training’ sample is skewed to PD (ROC = 0.607064). Further testing is to determine improvements from randomness. The target ‘gold’ standard in the ‘training’ sample is limited to under 100% (ROC = 0.917341). Further testing is to determine improvements that approach this standard.

Features: PEDOMETER

For all 20 PEDOMETER features, the improvement was not much better (+ 0.004702) than randomness (ROC = 0.611766).

I report each of the 20 features with their delta ROC as an isolated ensemble run.

Main Effects, first pass (delta ROC: positive values improved over randomness).

```
[19] mindeviationDSPS       [4] medianDS      [2] medianDSPS
            0.02347             0.02082             0.01952
[15] mindeviationDDPS [16] maxdeviationDDPS [13] mindeviationDPS
            0.01755             0.01500             0.00779
        [3] medianDD       [6] medianSPS         [11] iqrDPS
            0.00469            -0.00270            -0.00318
[20] maxdeviationDSPS     [1] medianDDPS [14] maxdeviationDPS
           -0.00334            -0.00490            -0.00503
 [18] maxdeviationSPS        [12] iqrSPS [17] mindeviationSPS
           -0.00619            -0.01126            -0.01262
         [7] iqrDDPS       [5] medianDPS         [8] iqrDSPS
           -0.01537            -0.01893            -0.01966
          [10] iqrDS           [9] iqrDD
           -0.02100            -0.03329
```

With the hierarchical selection ‘c(19,4,15,20)’, the improvement was a bit better (+ 0.054549) than randomness (ROC = 0.6616127628). There is clear multicollinearity in the PEDOMETER features, so this approach may remove such bias and stops when the most relevant factors are found.

```
[19] mindeviationDSPS      # (0.6305339)
                     # delta-steps-per-second
                     # smallest number of steps relative
                     #    normed to largest number of steps (max = 1)

[4] medianDS            # +(0.02514) # maybe do this one twice?, 16 then shows up?
                     # delta-steps

[15] mindeviationDDPS      # +(0.00517)
                     # delta-distance-per-second
                     # smallest distance relative
                     #    normed to largest distance (max = 1)

[20] maxdeviationDSPS      # +(0.00077)
                     # largest number of steps relative
                     #    normed to smallest number of steps (min = 1)
```

For fun, I try a model with an initial seed of ‘c(19,4,4);’ thereby suggesting that using a feature twice may alter the decision tree location. The hierarchical model identification shows a different pathway to success, with slight marginal gains. Currently, I have to treat the scoring model as a black box, but it is evident that if their were a nested-model approach or some other tweaks to the ensemble, improvements may additionally occur. Result: 'c(19,4,4,15,16)' with delta ROC +(0.0574759) an improvement over the above non-nested selection model by +(0.0029269). This suggests that the hierarchical approach is a good starting point but a nested approach will achieve the best gains (in the current black-box setup; I am certain those with more experience with ensembles can tweak the black box as well).

Features: MOTION

I define MOTION features as features extracted from nonPedometer data sources.

I report each of the 195 features with their delta ROC as an isolated ensemble run.

Main Effects, first pass (delta ROC: positive values improved over randomness).

```
               [151] RESTINGminorHzMAX
                                 0.04974
                [163] RESTINGmajorCyclesMAX
                                 0.04968
                [148] RESTINGminorCyclesMAX
                                 0.04920
                [142] RESTINGsubExtremesMAX
                                 0.04812
             [141] RESTINGsubExtremesMEDIAN
                                 0.04580
                  [166] RESTINGsubLengthMAX
                                 0.04409
             [147] RESTINGminorCyclesMEDIAN
                                 0.04318
               [165] RESTINGsubLengthMEDIAN
                                 0.04142
                [140] RESTINGsubExtremesMIN
                                 0.03653
             [162] RESTINGmajorCyclesMEDIAN
                                 0.03481
                  [164] RESTINGsubLengthMIN
                                 0.03393
                [146] RESTINGminorCyclesMIN
                                 0.03206
[123] RETURNWALKmajorCycleTimeMedianMEDIAN
                                 0.03166
                 [150] RESTINGminorHzMEDIAN
                                 0.03063
   [183] RESTINGmajorCycleTimeMedianMEDIAN
                                 0.02968
                [161] RESTINGmajorCyclesMIN
                                 0.02967
                  [143] RESTINGsubDomainMIN
                                 0.02837
      [174] RESTINGmajorCycleTimeMADMEDIAN
                                 0.02736
                    [149] RESTINGminorHzMIN
                                 0.02609
      [184] RESTINGmajorCycleTimeMedianMAX
                                 0.02480
```

```
                  [89] RETURNWALKminorHzMIN
                                 0.02364
               [95] RETURNWALKoverallMaxMIN
                                 0.02327
         [185] RESTINGmajorCycleTimeIQRMIN
                                 0.02279
   [122] RETURNWALKmajorCycleTimeMedianMIN
                                 0.02264
      [182] RESTINGmajorCycleTimeMedianMIN
                                 0.02257
                   [181] RESTINGmajorHzMAX
                                 0.02244
                 [17] usefulSecondsResting
                                 0.02215
 [139] RETURNWALKnAmplitudesContractionMAX
                                 0.02122
                 [145] RESTINGsubDomainMAX
                                 0.02112
     [190] RESTINGmajorCycleTimeMADMAX.1
                                 0.02104
                [121] RETURNWALKmajorHzMAX
                                 0.02095
                 [91] RETURNWALKminorHzMAX
                                 0.02079
         [175] RESTINGmajorCycleTimeMADMAX
                                 0.02076
   [124] RETURNWALKmajorCycleTimeMedianMAX
                                 0.02049
             [120] RETURNWALKmajorHzMEDIAN
                                 0.02023
                        [11] walkingPOSMED
                                 0.02018
                         [8] restingACCMED
                                 0.01964
         [187] RESTINGmajorCycleTimeIQRMAX
                                 0.01949
      [99] RETURNWALKoverallAmplitudeMEDIAN
                                 0.01876
    [189] RESTINGmajorCycleTimeMADMEDIAN.1
                                 0.01826
```

```
          [119] RETURNWALKmajorHzMIN
                           0.01805
      [38] OUTBOUNDoverallAmplitudeMIN
                           0.01730
          [180] RESTINGmajorHzMEDIAN
                           0.01728
[63] OUTBOUNDmajorCycleTimeMedianMEDIAN
                           0.01706
 [128] RETURNWALKmajorCycleTimeMADMIN.1
                           0.01689
 [199] RESTINGnAmplitudesContractionMAX
                           0.01680
        [144] RESTINGsubDomainMEDIAN
                           0.01638
   [186] RESTINGmajorCycleTimeIQRMEDIAN
                           0.01499
      [96] RETURNWALKoverallMaxMEDIAN
                           0.01415
    [98] RETURNWALKoverallAmplitudeMIN
                           0.01401
           [60] OUTBOUNDmajorHzMEDIAN
                           0.01387
    [188] RESTINGmajorCycleTimeMADMIN.1
                           0.01377
      [93] RETURNWALKoverallMinMEDIAN
                           0.01357
           [34] OUTBOUNDoverallMinMAX
                           0.01339
           [35] OUTBOUNDoverallMaxMIN
                           0.01236
   [76] OUTBOUNDnAmplitudesExpansionMAX
                           0.01227
     [170] RESTINGmajorAmplitudeIQRMIN
                           0.01178
              [61] OUTBOUNDmajorHzMAX
                           0.01165
[48] OUTBOUNDmajorAmplitudeMedianMEDIAN
                           0.01068
             [179] RESTINGmajorHzMIN
                           0.01003
                   [9] restingACCMAD
                           0.00979
     [160] RESTINGoverallAmplitudeMAX
                           0.00875
       [84] RETURNWALKsubDomainMEDIAN
                           0.00869
   [196] RESTINGnAmplitudesExpansionMAX
                           0.00850
         [97] RETURNWALKoverallMaxMAX
                           0.00765
                  [10] restingACCIQR
                           0.00715
```

```
          [173] RESTINGmajorCycleTimeMADMIN
                                 0.00613
                [157] RESTINGoverallMaxMAX
                                 0.00590
                  [19] usefulSecondsReturn
                                 0.00564
[136] RETURNWALKnAmplitudesExpansionMAX
                                 0.00559
     [113] RETURNWALKmajorCycleTimeMADMIN
                                 0.00558
          [158] RESTINGoverallAmplitudeMIN
                                 0.00534
      [100] RETURNWALKoverallAmplitudeMAX
                                 0.00460
     [62] OUTBOUNDmajorCycleTimeMedianMIN
                                 0.00433
                      [13] walkingPOSIQR
                                 0.00425
                [154] RESTINGoverallMinMAX
                                 0.00379
[191] RESTINGmajorCycleOutOfBoundsMIN.1
                                 0.00367
     [127] RETURNWALKmajorCycleTimeIQRMAX
                                 0.00361
             [90] RETURNWALKminorHzMEDIAN
                                 0.00355
              [85] RETURNWALKsubDomainMAX
                                 0.00285
                [152] RESTINGoverallMinMIN
                                 0.00162
[73] OUTBOUNDmajorCycleOutOfBoundsMAX.1
                                 0.00125
         [50] OUTBOUNDmajorAmplitudeIQRMIN
                                 0.00108
   [178] RESTINGmajorCycleOutOfBoundsMAX
                                 0.00087
   [58] OUTBOUNDmajorCycleOutOfBoundsMAX
                                 0.00067
                [37] OUTBOUNDoverallMaxMAX
                                 0.00027
     [74] OUTBOUNDnAmplitudesExpansionMIN
                                 0.00022
             [94] RETURNWALKoverallMinMAX
                                -0.00023
         [102] RETURNWALKmajorCyclesMEDIAN
                                -0.00076
            [88] RETURNWALKminorCyclesMAX
                                -0.00083
                      [12] walkingPOSMAD
                                -0.00098
              [83] RETURNWALKsubDomainMIN
                                -0.00136
          [156] RESTINGoverallMaxMEDIAN
                              -0.00211
           [92] RETURNWALKoverallMinMIN
                              -0.00249
 [77] OUTBOUNDnAmplitudesContractionMIN
                              -0.00267
[198] RESTINGnAmplitudesContractionMEDIAN
                              -0.00274
  [167] RESTINGmajorAmplitudeMedianMIN
                              -0.00278
   [39] OUTBOUNDoverallAmplitudeMEDIAN
                              -0.00330
[193] RESTINGmajorCycleOutOfBoundsMAX.1
                              -0.00351
   [159] RESTINGoverallAmplitudeMEDIAN
                              -0.00359
  [176] RESTINGmajorCycleOutOfBoundsMIN
                              -0.00447
  [171] RESTINGmajorAmplitudeIQRMEDIAN
                              -0.00489
          [153] RESTINGoverallMinMEDIAN
                              -0.00501
  [47] OUTBOUNDmajorAmplitudeMedianMIN
                              -0.00530
                [59] OUTBOUNDmajorHzMIN
                              -0.00536
 [195] RESTINGnAmplitudesExpansionMEDIAN
                              -0.00540
            [155] RESTINGoverallMaxMIN
                              -0.00566
  [64] OUTBOUNDmajorCycleTimeMedianMAX
                              -0.00580
                    [14] restingPOSMED
                              -0.00595
     [65] OUTBOUNDmajorCycleTimeIQRMIN
                              -0.00607
                [31] OUTBOUNDminorHzMAX
                              -0.00609
     [172] RESTINGmajorAmplitudeIQRMAX
                              -0.00621
      [40] OUTBOUNDoverallAmplitudeMAX
                              -0.00623
           [22] OUTBOUNDsubExtremesMAX
                              -0.00642
  [66] OUTBOUNDmajorCycleTimeIQRMEDIAN
                              -0.00724
 [79] OUTBOUNDnAmplitudesContractionMAX
                              -0.00844
          [36] OUTBOUNDoverallMaxMEDIAN
                              -0.00888
     [55] OUTBOUNDmajorCycleTimeMADMAX
                              -0.00891
```

```
      [169] RESTINGmajorAmplitudeMedianMAX
                                  -0.00898
              [43] OUTBOUNDmajorCyclesMAX
                                  -0.00954
   [114] RETURNWALKmajorCycleTimeMADMEDIAN
                                  -0.00964
              [20] OUTBOUNDsubExtremesMIN
                                  -0.00971
            [80] RETURNWALKsubExtremesMIN
                                  -0.00984
              [28] OUTBOUNDminorCyclesMAX
                                  -0.01007
            [33] OUTBOUNDoverallMinMEDIAN
                                  -0.01027
          [105] RETURNWALKsubLengthMEDIAN
                                  -0.01033
              [26] OUTBOUNDminorCyclesMIN
                                  -0.01057
                       [15] restingPOSMAD
                                  -0.01067
 [78] OUTBOUNDnAmplitudesContractionMEDIAN
                                  -0.01126
           [21] OUTBOUNDsubExtremesMEDIAN
                                  -0.01170
           [27] OUTBOUNDminorCyclesMEDIAN
                                  -0.01174
                  [29] OUTBOUNDminorHzMIN
                                  -0.01179
   [107] RETURNWALKmajorAmplitudeMedianMIN
                                  -0.01209
       [70] OUTBOUNDmajorCycleTimeMADMAX.1
                                  -0.01215
             [104] RETURNWALKsubLengthMIN
                                  -0.01226
   [168] RESTINGmajorAmplitudeMedianMEDIAN
                                  -0.01226
     [56] OUTBOUNDmajorCycleOutOfBoundsMIN
                                  -0.01239
           [101] RETURNWALKmajorCyclesMIN
                                  -0.01251
[129] RETURNWALKmajorCycleTimeMADMEDIAN.1
                                  -0.01262
               [46] OUTBOUNDsubLengthMAX
                                  -0.01286
      [49] OUTBOUNDmajorAmplitudeMedianMAX
                                  -0.01354
               [23] OUTBOUNDsubDomainMIN
                                  -0.01380
   [75] OUTBOUNDnAmplitudesExpansionMEDIAN
                                  -0.01389
            [82] RETURNWALKsubExtremesMAX
                                  -0.01401
     [111] RETURNWALKmajorAmplitudeIQRMEDIAN
                                 -0.01406
                  [44] OUTBOUNDsubLengthMIN
                                 -0.01430
            [87] RETURNWALKminorCyclesMEDIAN
                                 -0.01450
               [45] OUTBOUNDsubLengthMEDIAN
                                 -0.01461
        [125] RETURNWALKmajorCycleTimeIQRMIN
                                 -0.01465
                 [32] OUTBOUNDoverallMinMIN
                                 -0.01475
               [24] OUTBOUNDsubDomainMEDIAN
                                 -0.01478
        [115] RETURNWALKmajorCycleTimeMADMAX
                                 -0.01508
  [117] RETURNWALKmajorCycleOutOfBoundsMEDIAN
                                 -0.01511
   [133] RETURNWALKmajorCycleOutOfBoundsMAX.1
                                 -0.01515
           [52] OUTBOUNDmajorAmplitudeIQRMAX
                                 -0.01540
            [81] RETURNWALKsubExtremesMEDIAN
                                 -0.01544
   [108] RETURNWALKmajorAmplitudeMedianMEDIAN
                                 -0.01583
        [54] OUTBOUNDmajorCycleTimeMADMEDIAN
                                 -0.01591
     [57] OUTBOUNDmajorCycleOutOfBoundsMEDIAN
                                 -0.01617
                  [25] OUTBOUNDsubDomainMAX
                                 -0.01620
     [126] RETURNWALKmajorCycleTimeIQRMEDIAN
                                 -0.01620
    [137] RETURNWALKnAmplitudesContractionMIN
                                 -0.01626
                        [16] restingPOSIQR
                                 -0.01634
      [130] RETURNWALKmajorCycleTimeMADMAX.1
                                 -0.01635
   [72] OUTBOUNDmajorCycleOutOfBoundsMEDIAN.1
                                 -0.01664
        [51] OUTBOUNDmajorAmplitudeIQRMEDIAN
                                 -0.01669
                         [7] walkingACCIQR
                                 -0.01675
           [53] OUTBOUNDmajorCycleTimeMADMIN
                                 -0.01688
               [86] RETURNWALKminorCyclesMIN
                                 -0.01691
[132] RETURNWALKmajorCycleOutOfBoundsMEDIAN.1
                                 -0.01713
```

```
   [131] RETURNWALKmajorCycleOutOfBoundsMIN.1
                                 -0.01717
                [41] OUTBOUNDmajorCyclesMIN
                                 -0.01720
             [42] OUTBOUNDmajorCyclesMEDIAN
                                 -0.01736
                         [5] walkingACCMED
                                 -0.01746
         [68] OUTBOUNDmajorCycleTimeMADMIN.1
                                 -0.01755
     [134] RETURNWALKnAmplitudesExpansionMIN
                                 -0.01755
             [103] RETURNWALKmajorCyclesMAX
                                 -0.01759
        [110] RETURNWALKmajorAmplitudeIQRMIN
                                 -0.01782
    [116] RETURNWALKmajorCycleOutOfBoundsMIN
                                 -0.01805
        [194] RESTINGnAmplitudesExpansionMIN
                                 -0.01828
    [118] RETURNWALKmajorCycleOutOfBoundsMAX
                                 -0.01865
      [69] OUTBOUNDmajorCycleTimeMADMEDIAN.1
                                 -0.01876
                         [6] walkingACCMAD
                                 -0.01931
   [192] RESTINGmajorCycleOutOfBoundsMEDIAN.1
                                 -0.01934
      [197] RESTINGnAmplitudesContractionMIN
                                 -0.01961
        [112] RETURNWALKmajorAmplitudeIQRMAX
                                 -0.02002
 [138] RETURNWALKnAmplitudesContractionMEDIAN
                                 -0.02004
               [106] RETURNWALKsubLengthMAX
                                 -0.02029
     [71] OUTBOUNDmajorCycleOutOfBoundsMIN.1
                                 -0.02081
           [67] OUTBOUNDmajorCycleTimeIQRMAX
                                 -0.02145
   [135] RETURNWALKnAmplitudesExpansionMEDIAN
                                 -0.02186
     [109] RETURNWALKmajorAmplitudeMedianMAX
                                 -0.02248
                 [30] OUTBOUNDminorHzMEDIAN
                                 -0.02365
                 [18] usefulSecondsOutbound
                                 -0.02434
    [177] RESTINGmajorCycleOutOfBoundsMEDIAN
                                 -0.02761
```

Similarly, for all 195 MOTION features, I report the hierarchical selections. First, I report a 'fastignore=TRUE' as a seed for another deeper-dive hierarchical selection of only the positive results from the first iteration of the main effects described above; that is, all positive values were included in the next pass, with an initial seed of 'c(151,97)'. From this setup, I report a delta ROC of (+ 0.10250992) with the final features 'c(151,97,121,99,96)'. I do a similar slow pass on the initial seed, and report a delta ROC of (+ 0.1026788) with the final features 'c(151,97,121,99,18)'.

I note that for these variables the trailing uppercase values MIN,MEDIAN,MAX represent a dimension(x,y,z) with the least, average, and most value (respectively) for a given variable. In this setup, I am not overally concerned with the direction (x,y,z) allowing for some variability in the phone orientation. [Todo, go back and allow x,y,z with MIN,MEDIAN,MAX as separate features.]

I do some analysis on subfeature collections of MOTION: ANGLES (5:16) [delta ROC +0.037361 with c(11,9,8,10)], OUTBOUND (18,20:79) [delta ROC +0.0621645 with c(38,62,53,40,33,52)], RETURNWALK (19,80:139) [delta ROC +0.060899 with c(123,98)], RESTING (17,140:199) [delta ROC +0.0614532 with c(151,190,142,148)].

We create a best MERGE of PEDOMETER and MOTION, yet see no gains over the MOTION scores above.

Our best delta ROC of (+ 0.1026788) with the final features 'c(151,97,121,99,18)'.

My submission consists of my feature files (csv), this writeup (pdf) and a link to a github repository.

https://github.com/MonteShaffer/mPowerEI/

Helper files
.../mPowerEI/example/config.txt
.../mPowerEI/example/R-manualSynapseInclude.txt

A vignette notebook outlining the features can be found in the github repository:
.../mPowerEI/example/R.notebook.mPowerEI
.../mPowerEI/example/R.notebook.Rmd
.../mPowerEI/example/R.notebook.nb.html

About EInstEIn

The Team Leader, Monte J. Shaffer has a Ph.D. in the social sciences. In addition, with his undergrad in Math/Physics, MBA in Market Research and his M.S. in Statistics (where he consulted for various bio-informatic projects and learned to use R), he is able to provide a unique range of hard and soft science skills to address many difficult problems. He is the founder of a boutique research institute Entrepreneurial Innovation (EI) and derived from the research institute's acronym the team name: **Team EInstEIn**. He consulted with drone hobbyists to understand key features of the 'dead-reckoning' problem. He consulted with physical therapists to understand gait cycles and specifically PD-specific gait characteristics. The research institute (Entreprenuerial Innovation) has several key projects designed specifically to improve QoL (quality of life) for PD patients and to develop wearables specific to PD (rather than this iPhone approach). An important aim of the wearables is to have a better understanding of mobility deterioration over time, and distinguish PD degeneration from natural aging processes (the Shaffer-Gatto Functional Mobility Index). In time, such an approach would enable the research community to find the biomarkers needed to find the cure. As a data scientist, the organization of such big wearable data for research consumption needs to have data-management protocols to protect the data, but more importantly, it needs to be organized that in minutes the researchers can import the data into their preferred programming interface (R, python, SAS, Matlab, gauss, stata) and be off and running.

References

[1] Ambres, Oscar and Gracian Trivino (2012). Gait quality monitoring using an arbitrarily oriented smartphone. In *International Workshop on Ambient Assisted Living*, pp. 224–231. Springer.

[2] Bot, Brian M, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E Ray Dorsey, et al. (2016). The mPower study, Parkinson disease mobile data collected using ResearchKit. *Scientific data* 3, 160011.

[3] Clover, Juli (2016). Apple Hires Sage Bionetworks President Stephen Friend for Health-Related Projects. https://www.macrumors.com/2016/06/23/apple-researchkit-hire-stephen-friend/. [Online; accessed September 14, 2017].

[4] Densford, Fink (2017). CEO Cook says healthcare is "big area for Apples future". http://www.massdevice.com/ceo-cook-says-healthcare-big-area-apples-future/. [Online; accessed September 14, 2017].

[5] Enriquez, Jof (Enriquez). Apple Taps Sage Bionetworks Co-Founder Stephen Friend For "Health-Related Projects". https://www.meddeviceonline.com/doc/apple-taps-sage-bionetworks-co-founder-stephen-friend-for-health-related-projects-0001. [Online; accessed September 14, 2017].

[6] Niijima, Arinobu, Osamu Mizuno, and Tomohiro Tanaka (2014). A study of gait analysis with a smartphone for measurement of hip joint angle. In *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pp. 1–4. IEEE.

[7] Wilbanks, John and Stephen H Friend (2016). First, design for data sharing. *Nature biotechnology* 34 (4), 377–379.