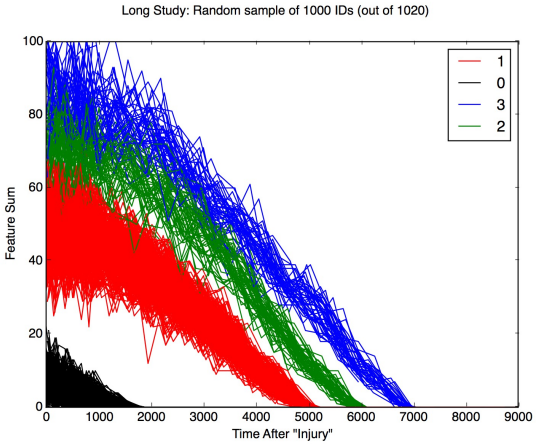
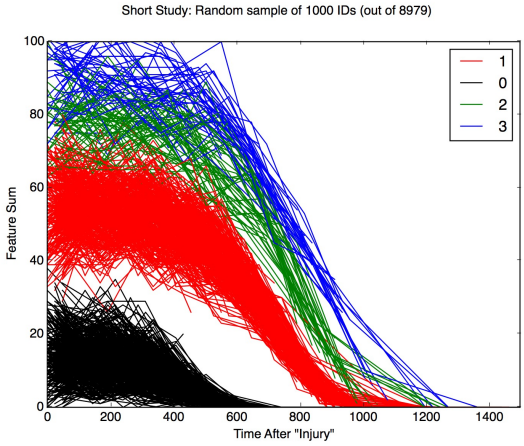
Research progress on simulated data

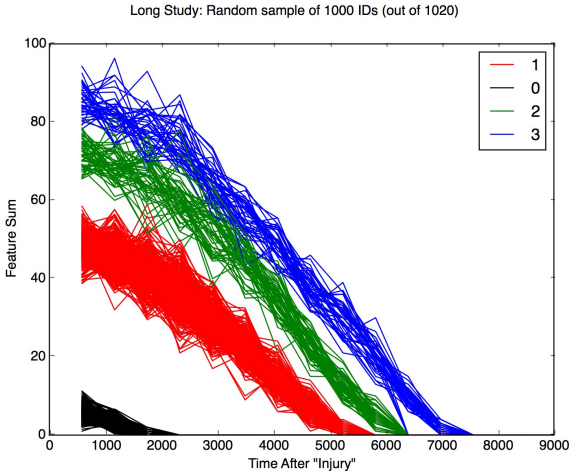
- Plotted sum of features vs time after injury
  - The 4 different classes were visually separable. Two different “study lengths” were detected. Ill call them ‘long’ and ‘short’. Here are their respective plots:

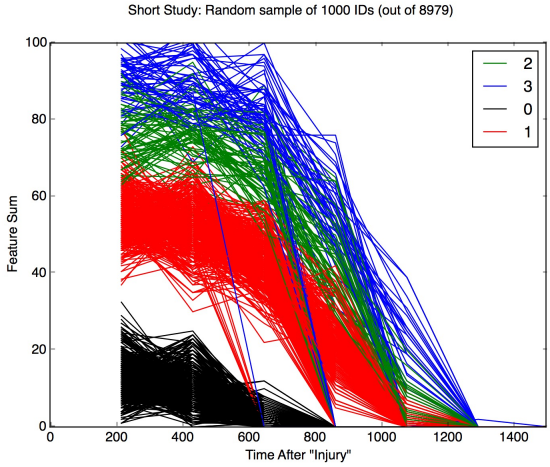




  -

- Bucketing: In order to compare points on the curve, the features sums were aggregated by averaging the points within bucketed time windows.
  - The first and simplest method was to partition the time domain into M equal buckets. Here are the plots for that method:
  - ‘Long’ Study, 16 buckets. ‘Short’ study, 7 buckets





- It turned out that the observations for each patient were more frequent for the early measurements then they became more spread out. Using evenly spaced buckets this resulted in more points being aggregated in the early buckets than the later buckets.

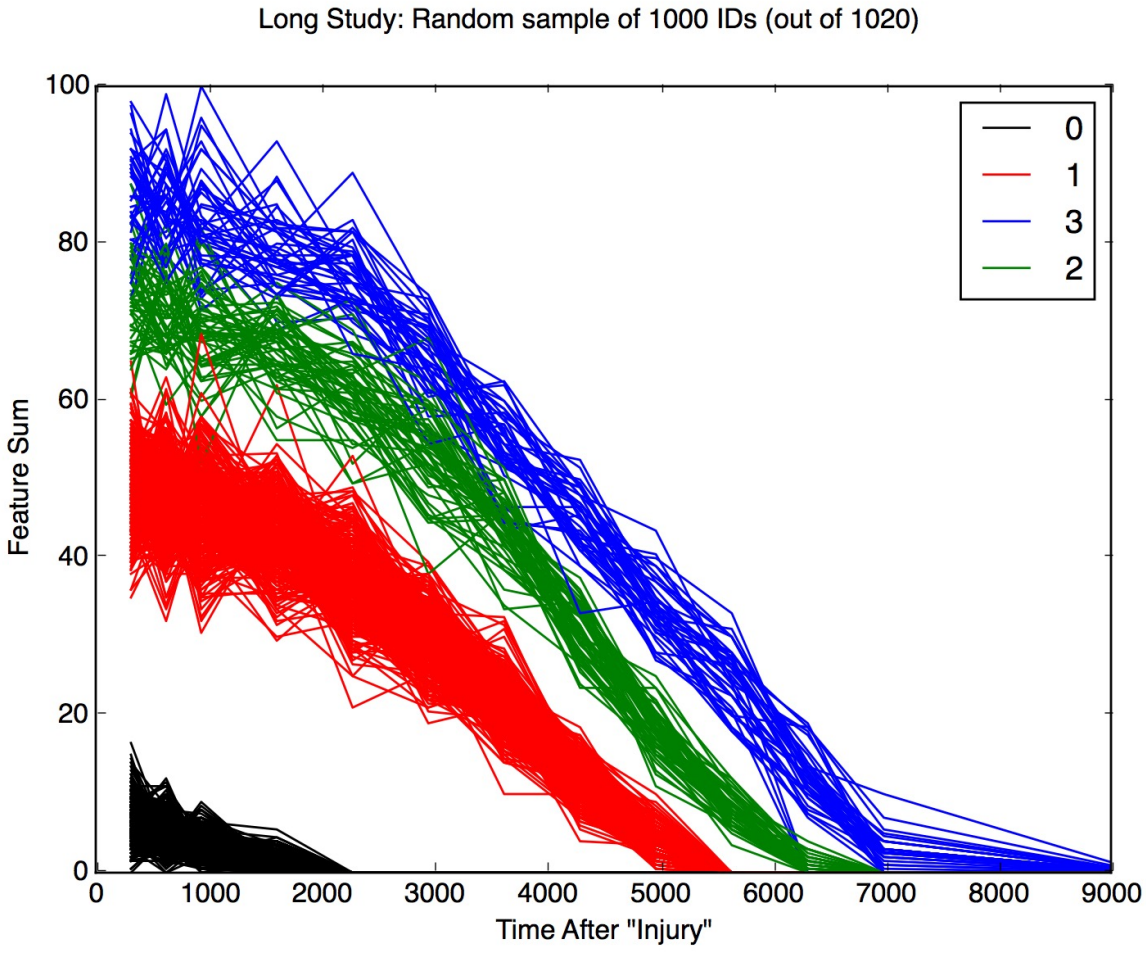
```
Average Compression per Bucket ('Short' Study):
    [ 1.50228311  1.30014478  1.18008687  0.9355162   0.64862457  0.22998107
  0.13442477  0.06002896]

Average Compression per Bucket ('Long' Study):
    [ 4.13333333  2.89705882  1.90098039  1.67352941  1.64901961  1.76960784
  1.79411765  1.68921569  1.69705882  1.76764706  1.52058824  1.19705882
  0.99215686  0.68823529  0.40294118  0.17058824]
```

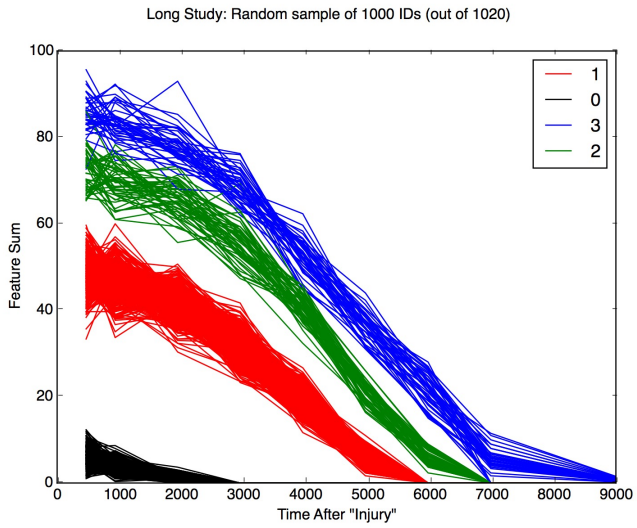
- To optimize the buckets size (or time length of the buckets) I calculated the average length of time between observations then used every second 'average observation time' as a bucket. In theory this would mean that each bucket would contain 2 observations on average. Based on the average 'compression'(observations per bucket) below it appears that this worked as expected.

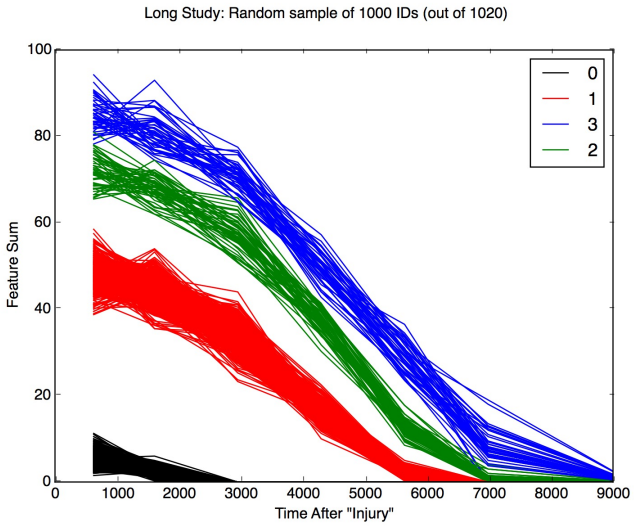
```
Average Compression per Bucket ('Long' Study):
    [ 2.20588235  2.06568627  2.00392157  2.04705882  2.00588235  2.02745098
  2.00686275  2.01568627  2.00098039  1.99607843  1.84901961  1.50980392
  2.20882353]
```

- Below is the plot for this bucketing technique. You will notice that the bucket sizes are smaller to start and get bigger as time after injury increases.
- Unfortunately, this bucketing technique only worked well for the long study. In order for me to make it work well for the short study I will have to figure out a good way to interpolate values for buckets with no observations. This is a coding problem that would take more time than I felt was worth it for this simulated data. However, it is an idea that I may use for the real EEG data.



- In this simulated TBI data the 'long' study trajectories have ~25 observations on average. When these data are aggregated into a smaller number of large buckets the variation at each time point is reduced for each class. (sampleSD = popSD/sqrt(n)) As we know when taking averages the variance decreases as the sample size increases. Also, as per the Central Limit Theorem the distribution of these averages becomes more normal.
    - Here are two plots with there respective compression rates where the number of buckets are decreased. The right plot has 9 buckets the left plot has 7 buckets.






```
Average Compression per Bucket ('Long' Study):
    [ 3.33529412  2.94019608  3.04901961  3.03137255  3.00294118  3.02058824
  2.95294118  2.40196078  2.20882353]
```

```
Average Compression per Bucket ('Long' Study):
    [ 4.27156863  4.05098039  4.03333333  4.02254902  3.99705882  3.35882353
  2.20882353]
```

- You can see that the variance decreases and the trajectories become more distinct when the number of buckets is decreased.
- Now, that the data is bucketed into discrete time intervals the next steps are to:
    - Conduct statistical test (t or z tests) at each point to differentiate the trajectories.
    - Fit polynomial functions to each class. The coefficients for the fitted polynomial or each class can be used as features to train traditional ML models.
        - This could be a great technique applied to EEG data. Coefficients can be calculated from the complexity index trajectories for each calculated metric (i.e. sampEntropy, RQA, etc) that we decided to use.

Conclusion:
This weeks study benefited from the simplicity of the simulated data. Working with trajectory that have many observations (25+) allowed for effective bucketing. Obviously the goal is to generalize these techniques to EEG data where similar curves can be calculated from the Complexity index for each calculated measurement. In a perfect world we would have many patients who all get measured at the same time points in their development. Almost as good would be to have many patients who get many(25+) measurements over a similar time domain

(3-24 months). This way rather than interpolating missing data we can aggregate a plethora of data.
Unfortunately, the reality is that neither of these scenarios will be easy to facilitate. We will probably need to explore different techniques of imputing missing data.
Maybe simply replicating data between labeled patients of the same class would work.
For example:

Time:[3, 6, 9, 12, 15, 18] (months)
id_1: [2, 4, NA, 5, NA, 7]
id_2: [1, 3, 4, 4, 5, 6]
becomes,
id_1: [2, 4, 4, 5, 5, 7]
id_2: [1, 3, 4, 4, 5, 6]

Or we could try basic interpolating:

Time:[3, 6, 9, 12, 15, 18] (months)
id_1: [2, 4, NA, 5, NA, 7]
id_2: [1, 3, 4, 4, 5, 6]
becomes,
id_1: [2, 4, 4.5, 5, 6, 7]
id_2: [1, 3, 4, 4, 5, 6]

Generally, I think finding a good bucketing convention and figuring out a good way to fill in missing data will be important steps to making the data workable. Once it is structured in regular intervals with a reasonable value at each observation then there are several different classification techniques to explore.