# fMRI Dataset from Complex Natural Simulation with Forrest Gump: A Restudy

Chang, Jordeen
jodreen

Daks, Alon
AlonDaks

Luo, Ying
yingtluo

Yu, Lisa Ann
lisaannyu

November 30, 2015

**Abstract**

Most fMRI studies use highly simplified stimulus that are very different from what people experience in everyday life. A High-Resolution 7- Tesla fMRI Dataset from Complex Natural Stimulation with an Audio Movie by Hanke et. al. sought to create a dataset of naturally occurring brain states by exposing participants to a more complex stimulus, the audio description of Forrest Gump. This particular audio description allows for the study of auditory attention and cognition, language and music perception, and retrieval of explicit memory without the effect of visual imagery. Furthermore, the study was conducted on 20 subjects enabling research into brain similarities among individuals when exposed to the same complex stimulus. The goal of our paper is to first reproduce a subset of the analysis conducted by Hanke et. al, then apply machine learning to see if we can predict if a subject was listening to a day or night scene of the movie based on brain state.

## 1 Introduction

The main purpose of the original study was to examine properties of brain response patterns that are supposedly common when people are exposed to audio and movie simulation. We intend to replicate their experiment using the data they gathered from the 20 participants. For example, a BOLD time-series similarity measure (e.g. correlation) is often used to quantify similarities in responses among individuals. Hank et al. recognized that this was a common approach, but they went beyond that and also implemented representational similarity analysis (RSA). To do so, we will create dissimilarity matrices for 18 individuals using the same searchlight mapping approach that they used (Subjects 4 and 10 were not included due to missing data). Doing so will capture 2nd-order isomorphisms in response patterns. Lastly, to access statistical significance, we will transform the representational consistency map into percent rank with respect to the total distribution of the DSM correlations. We'll calculate the mean correlation coefficient and compare our value to theirs.

Before we formally began, we performed basic sanity checks on the data. We downloaded and loaded the files successfully, and we have confirmed that we have data for every test subject. Because there were various versions of the data (e.g. linear vs. nonlinear), we had to first isolate which dataset best suited our needs. The nonlinear alignment had more smoothing than the linear alignment, which made that dataset more appropriate for initial exploratory data analysis. We recognize that the nonlinear alignment is crude, since it probably was conducted with no specific hypothesis about which parts of the brain should be expanded or compressed, but not having the machinery or background to better preprocess the data, we simply used the nonlinear alignment version of the data provided by Hanke et al [1].

Reproducibility is crucial in research, especially when such high volumes of data are involved, because it allows other people to fact check the work. When people collaborate, new insights can be shed and the rate of progress is expedited. For this study, we began by extracting data we see fitting and asking questions that were not addressed by the original study. To answer these questions, we utilized neuroscience concepts, additional packages, and parallel processing techniques. We found that in certain functions, such as parsing through csv files, it would have been easier to hard code in the correct values. However, hard coding would have defeated the reproducibility aspect of our study because our code would break if the data were slightly modified.

# 2 Data

The data is curated and segmented into 20 .TGZ files, where each of the 20 .TGZ files corresponds to one of the 20 subjects in the experiment. Each subject accounts for approximately 16 GBs of data. We verified the usability of the data by inspecting and loading data corresponding to subject 1. We limited our initial exploration to a single subject since downloading each .TGZ takes approximately one hour. The download for each subject includes a lot of other information besides just the fMRI data, such as cardiac and respiratory trace, angiographies, and structural MRI data. Each subject's fMRI data includes several formats: Raw BOLD functional MRI, Raw BOLD functional MRI (with applied distortion correction), Raw BOLD functional MRI (linear anatomical alignment), and Raw BOLD functional MRI (non-linear anatomical alignment). The corrected and aligned versions of the data attempt to eliminate device and scan related noise. Scan data is accessible in nibabel compatible formats (.NII).

In addition to fMRI data for each subject, there was also data about the experiment as a whole, including a German audio description of the stimulus and a csv file with scene information, including the timestamp of the beginning of each scene, the scene location, whether the scene took place during the day or night, and whether the scene was inside or outside.

# 3 Methods

## 3.1 Reproduction

A central theme of the course is exploring reproducible practices while conducting scientific research. For our project we address reproducibility from two perspectives: first, we designed our code and scripts to be modular and portable so that anyone with an appropriate computer can rerun our analysis and reproduce the results we cite in this paper. Second, we attempted to reproduce some of the results cited in the original paper by emulating Hankes technique and comparing values. This section addresses how we tried to reproduce Hankes analysis, challenges we faced, and how our results match up.

The original paper focused on inter-individual response pattern similarity. In other words, their analysis measures brain pattern correlation between subjects who listened to Forrest Gump. Hanke identifies two methods for measuring correlation. The first method takes the BOLD time-series and calculates the voxel-wise pearson correlation and second method employs representational similarity analysis to identify 2nd-order isomorphisms in the response patterns across brains. Since the first technique is simpler, we chose that to be the portion of the analysis we tried to reproduce. Although 20 subjects were scanned, Hanke excluded subjects 4 and 10 from the paper since data was missing. With the 18 remaining subjects there are 153 pairs (18 choose 2) for which voxel-wise correlations are calculated. Hanke calculated correlations on linearly and nonlinearly aligned raw data. Instead of using the papers raw data, which was applied through a bandpass filter, we used a version of the raw data provided by Matthew Brett which was non-linearly aligned and applied through a highpass filter.

The main challenge in reproducing the papers analysis resides in the large amount of data that must be processed. Since subjects were scanned at 7 tesla, the brain images have a very high resolution, and therefore are large. The full two hour scan for each subject is approximately 7.5 GBs, meaning 153 pairs of 2 x 7.5 GB files must be processed. A seemingly trivial calculation of computing correlations between a set of voxels is difficult at this papers scale. On an Intel core i7 (1.7 GHz) machine with 8 GBs of RAM, preprocessing (concatenating runs and converting the 4D image to 2D) takes approximately 30 minutes per subject. Calculating the correlations between a pair of subjects takes 10 minutes (8 minutes to load both subjects time courses into memory and 2 minutes to calculate the pearson correlations for all the voxels). To combat these concerns, we used pythons multiprocessing module to parallelize the reshaping of 4D to 2D arrays in the preprocessing step and parallelize the calculating of correlations between subjects. Additionally, we ran our analysis on Amazon EC2 instances with memory capacities large enough to fully load both two hour time courses for each subject in the pair being considered. By renting multiple EC2 instances, we are able to parallelize the preprocessing work enabling multiple subjects to be preprocessed once.

Since rerunning our preprocessing and correlation code may take several hours depending on ones machine, we defined a UNIX environment variable called STAT159_CACHED_DATA. If this variable is set to True, then our python scripts will look to see if preprocessed versions of the data exist in the /data/processed directory and use those instead of recomputing them. If however one wants to fully recompute everything, merely setting STAT159_CACHED_DATA to False will ensure that all python scripts do not use data files that have already been computing. If the user does not specify STAT159_CACHED_DATA, the default behavior is to use cached files. Enabling the user to easily choose whether or not to recompute intermediary data files is consistent with the reproducible paradigm this course emphasizes.

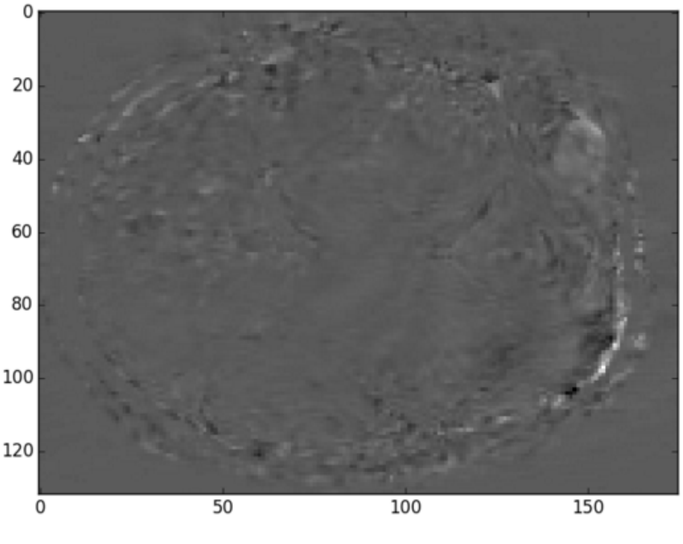
The results for this portion of our paper are still being computed, and be will include in final draft.

## 3.2 Extracting Data

We will not be using all of the data. The data files are very large: for each subject, the download size is approximately 16 GB, and that does not even include unzipping all the file inside. Given that we have twenty subjects, downloading all of this could be a paper in it of itself. and since the focus of this class is mainly fMRI data, we decided to ignore everything else. However, the fMRI data is still not of a trivial size because three versions of the data is included: the raw data, the linear alignment, and the non-linear alignment. We will only be choosing one of them (the one linearly aligned) because using all of them would just be redundant.

We smoothed the data before performing any analyses by applying a Gaussian filter. This filter smooths the signal between adjacent voxels, addressing our assumption that adjacent voxels are not independent. We recognize that this may be crude and perhaps counterproductive, since we want to pick out the voxels with the highest difference in signal between the two conditions, but we need to address our assumption, and smoothing is better than nothing.

Hanke et al. preprocessed the data for us by creating an EPI group template. They aligned each image to the first image for each individual, then aligned all brains using FLIRT. However, we thought it would still be a good idea to model the noise created by linear drift. This subject moved backwards and to the left, but the data does look quite messy. It appears as though the fMRI is picking up signal from outside the subjects head.



We will be only working with the first subject for now for testing purposes, but if time had allowed, we would have ran our code on a couple of other subjects for comparison testing. And to ensure speedy access to other subjects' datasets, our strategy for getting all the data entails each group member spending downloading a different quarter of the overall dataset, and then locally transferring the remaining three-quarters of the data from our hard drives. However, we did find the data online at studyforrest.org, which allowed us to only download the files we needed, as opposed to the entire datafile for each subject.

## 3.3 Exploration of Data

After perusing all the stimulus related data (e.g. data annotating each Forest Gump scene and when a subject was exposed to that scene), we decided to guide part of our analysis to see if we could summarize any interesting trends in fMRI response with respect to these features. Additionally, by investigating a single subject for our initial analysis and using the nonlinearly aligned transformed data, we have limited our projects scope to approximately 5 gigabytes from the roughly 320 gigabytes that were published with the paper. With the data reduced to a manageable state, we have successfully overcome this obstacle. Finally, by introducing a data_path.json file to reference where each raw data file is located, we have designed a clean schema for our python scripts to find and load data in the project repository.

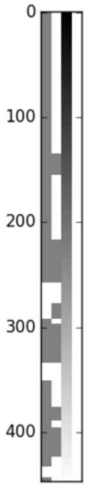
We deviated from the original data analysis in our reproduction. The paper used the raw data which varied largely with every subject, and processed that by standardizing among twenty of them with both linear alignments and non- linear alignments. We decided that we would just use the already processed nonlinearly aligned data since they already provided it, and we would not be able to do it as well as them.

Our initial plan was to first just conduct some basic exploratory data analysis and then further delve into the findings that strike us as intriguing or unusual. We began by writing functions that loaded the .nii files for each subject and calculated and plotted the standard deviations across voxels. After exploring the files provided by Hanke et al., we saw that there exists a csv file of the scenes in the movie with a time-stamp, a brief scene description, whether it takes place in day or night, and if it is inside or outside, as mentioned in the Data section.

We used this information and the fMRI data to select features with the goal of building the best possible predictor of certain aspects of a movie scene given fMRI data. We wanted to be able to accurately predict whether a particular movie scene falls under one of two categories for two separate groups, day or night, and inside or outside. Our plan of action was to 1) select voxels that seemed to differ most between the two groups as features, and 2) use machine learning to classify slices under one of two groups.

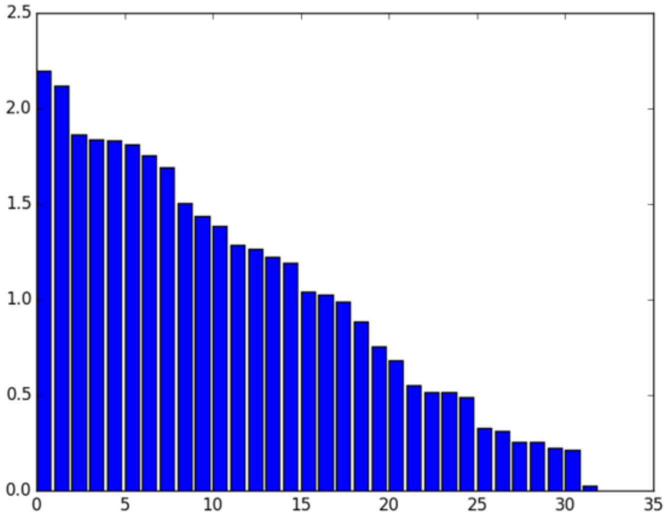
We began with the analysis for differentiating day and night slices, but this analysis can easily be applied to address the question of which voxels pick up on the difference between interior and exterior slices. We can then use machine learning to predict which slices are interior and which are exterior.

To select voxels as features for our classification problem, we first determined which slices took place during the day and which took place during the night, which were inside and which were outside. In order to account for baseline BOLD and any linear drift created by the subject moving in the scanner, we created a design matrix. The design matrix currently includes 4 columns: a column for day vs. night, a column for interior vs. exterior, a column for linear drift, and a column of ones for the intercept.



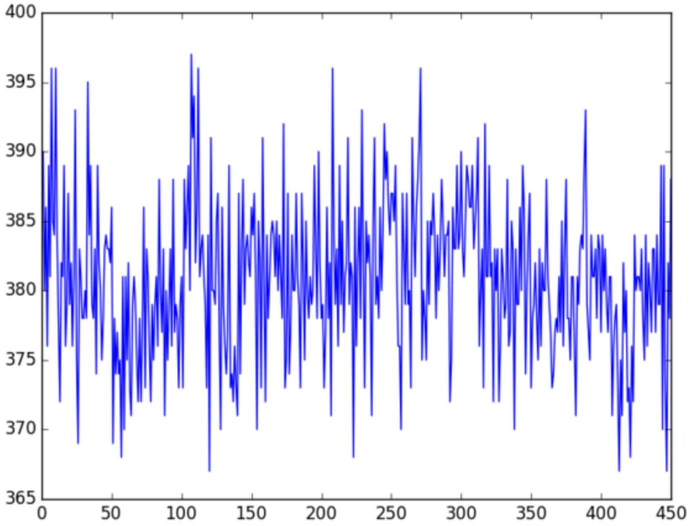
The figure above shows the design Matrix, including day vs. night, interior vs. exterior, linear drift, and a column of ones

We then performed a t-test for each voxel to determine if the signal is significantly different between each of the two groups we were comparing. This gave us an array of t-statistics, one for each voxel, a total of 1108800, as well as an array of betas, telling us the effect size. Since we wanted to pick out the voxels with the biggest change between groups, we cut down that number by selecting the top portion of them. This number of feature selections is rather arbitrary, but should be relatively small so we can create a random forest with it that does not overfit. Part of our analysis is determining the optimal number of features to select. This will be done via cross validation.

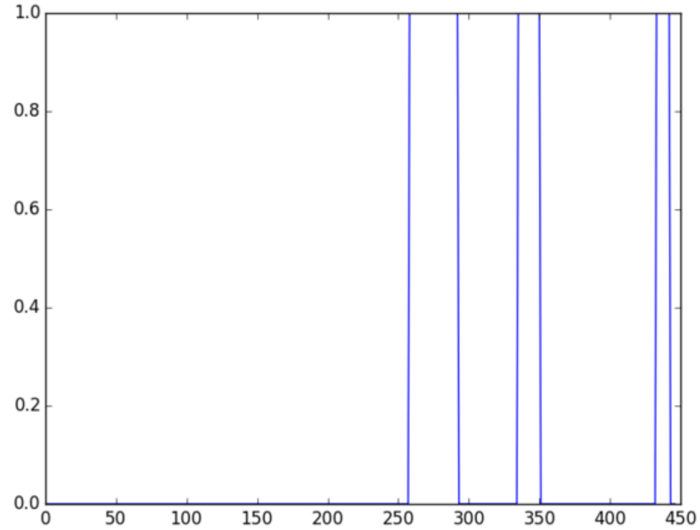


The figure above is the largest 32 t-statistics.

The time course for the voxel with the highest t-statistic looks very noisy. We can compare it to the stimulus (day) time course, but there is no obvious relationship.



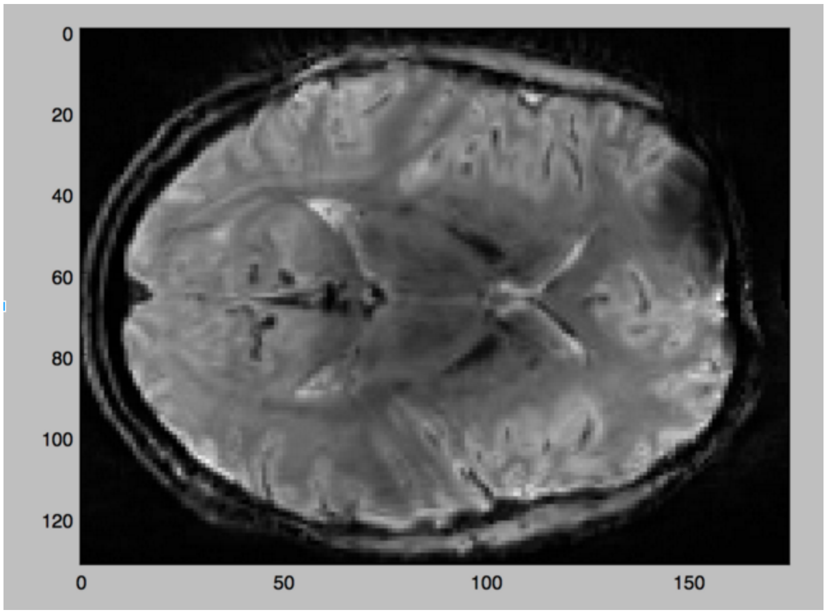
The figure is the time course for the voxel with the highest t-statistic.



The fiture is the day-night time course for Task 1, Run 1. Very few scenes take place at night for this run.

Because we are ultimately interested in feature selection, not in whether certain voxels are significantly different between the two groups, we did not correct for multiple comparisons. If we were interested in the significance of certain voxels, we would apply a Bonferroni correction.

We wanted to know which areas of the brain seemed to be most inclined to pick up on the difference between day and night. Thus, we plotted the betas testing the hypothesis that there is no difference between day and night for each voxel. Unfortunately, this looks very similar to our linear drift image, where there appears to be an artifact.



We can also apply this to the question of whether we can tell the difference between positive, negative, and neutral scenes by examining certain voxels in the brain. Unlike the scenes data, which we can only code as day, night, interior, or exterior, since that is all the information we have, since we are extrapolating the sentiment of the scene from the scenes descriptions ourselves and sentiments can vary along a spectrum from very positive to very negative, we can perform a linear regression with sentiments. We only coded the sentiment for scenes in which a narrator was speaking, and only the time in which he was speaking.

To see if we could predict whether a slice was taken from a day or night scene, we focused on creating a random forest, using voxels selected as the features as the nodes. We first trained on a random 80% of the slices, collapsing day slices and night slices together, so that the distribution of the slices selected was roughly equivalent to the distribution of the entire data, rather than selecting 80% of the day slices and 80% of the night slices. The other 20% of the slices we preserved to be our testing set. We created a decision tree, which we used to predict whether each slice in the testing set took place during the day or the night. In theory, we could use the first six runs as the training set, and see how well they predict the last two runs for a given participant.

### 3.4 Analysis

After perusing all the stimulus related data (e.g. data annotating each Forest Gump scene and when a subject was exposed to that scene), we decided to guide part of our analysis to see if we could summarize any interesting trends in fMRI response with respect to these features. Additionally, by investigating a single subject for our initial analysis and using the linearly aligned transformed data, we have limited our projects scope to approximately 5 GBs from the roughly 320 GBs that were published with the paper. With the data reduced to a manageable state, we have successfully overcome this obstacle. Finally, by introducing a data_path.json file to reference where each raw data file is located, we have designed a clean schema for our python scripts to find and load data in the project repository.

We are deviating from the original data analysis. The paper used the raw data which varied largely with every subject, and processed that by standardizing among twenty of them with both linear alignments and non-linear alignments. We decided that we would just use the already processed linearly aligned data since they already provided it, and we would not be able to do it as well as them.

Our initial plan is to first just conduct some basic exploratory data analysis and then further delve into the findings that strike us as intriguing or unusual. We began by writing functions that loaded the .nii files for each subject and calculated and plotted the standard deviations across voxels. From there, we will continue our analysis by isolating outliers and referencing those data points with the corresponding movie time to look for correlations between certain movie scenes and certain physiological responses.

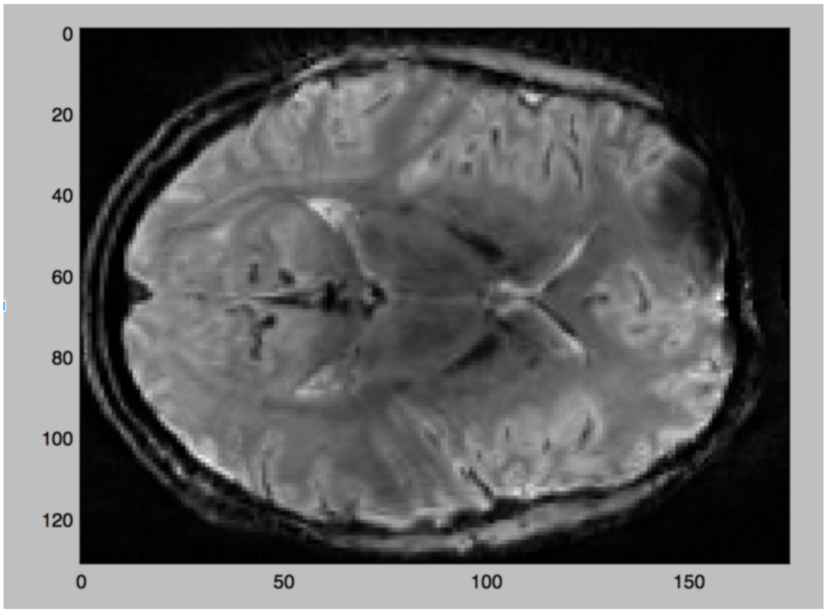
After exploring the files provided for Forrest Gump, we saw that there exists a csv file of the scenes in the movie with a time-stamp, a brief scene description, whether it takes place in day or night, and if it is inside or outside. We will try to see if there is anything interesting in the data based on these characteristics of the movie scenes.

After determining which points in time were during the day and which were during the night, and which were inside and which were outside, we can perform a t-test to determine if the signal is significantly different between each of the two groups for each voxel. We will perform a multiple comparisons test to correct for the number of t-tests we will be running. We will also model each of those voxels that are statistically significant to see what their time courses look like.

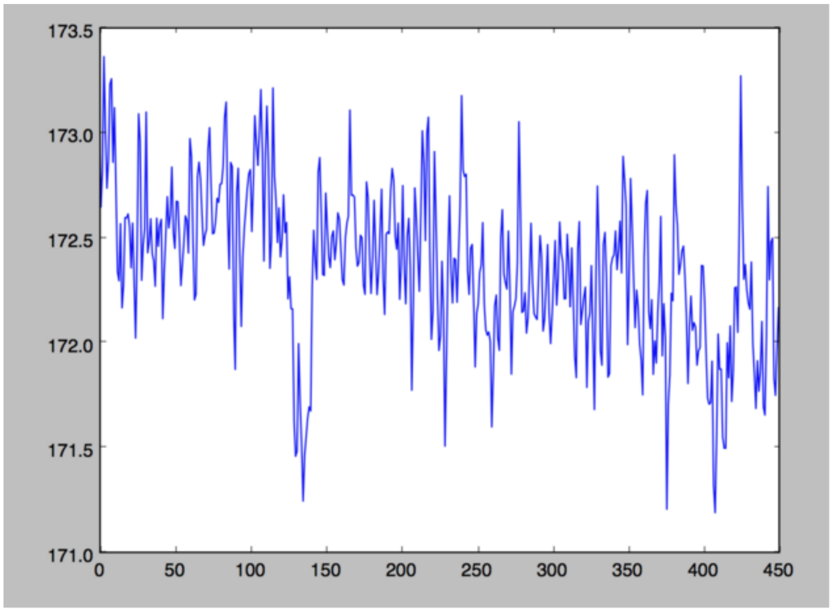
We have a few ideas as to what questions to ask, but are in the process of looking at other ideas. There is an entire website dedicated to studying this dataset, studyforrest.org, and we are researching to see what other people have done.

## 4 Results

We began our exploratory data analysis by writing functions that plot a slice of the brain at any given time and the standard deviations across voxels for individual subjects using Matlibplot. This enabled us, at a quick glance, to gauge the variability of the fMRI data.



It takes a few minutes to calculate the standard deviation for each volume in the 4-D array for Subject 1s Task 1, Run 1. This makes sense, as the 4-D array is (132, 175, 48, 451) in dimension; this means we are calculating and plotting 451 standard deviations in total. From the plot above, we can see that for this particular run and task, the standard deviations across all 451 points is within 171 and 173.5. Though this range may seem small, from the plot we can see that there is quite of variability within that range of 2.5. One possible explanation for this is the varying responses to different movie scenes. Some scenes will evoke different reactions and thus different values per voxels.



One initial direction we had hoped to explore was possible differences in physiological responses to certain movie scenes by gender and age. However, we later had to forgo this idea because doing so required us to process and analyze the entire dataset, something that we could not efficiently do given our limited computing resources. To be more specific, we would have had to comb through around 16 gigabytes worth of data just for one subject. Though these were certainly interesting questions, their focus was too broad for us to be able to fully address them.

So, in order to glean anything remotely useful, we had to narrow down our focus. There are 20 subjects with 8 runs each. We limited our initial analysis to just one run for one subject in order to conduct hypothesis tests. And after parsing the two csv files (scenes.csv and demographics.csv), we then had limited data about the movie scenes and personnel about each subject. We implemented Scene Slicer and Subject classes with attributes that were unique to each scene or subject.

For the Scene Slicer object, its attributes were:

1. path_to_scene_csv: a string that is the absolute path to scene.csv
2. image: the image loaded from the .nii data
3. scene_slices: a list of the scene slices
4. segment_duration: a list containing the time duration of each movie slice
5. scene_desc: a dictionary mapping the start of the scene time to tuple of day or night and interior or exterior
6. scene_keys: a sorted list of keys from the scene_desc dictionary

For a Subject object, it has the attributes:

1. id: an integer that identifies the subject and is unique to each subject, ranging from 1 to 20
2. gender: a string, either m or f, that denotes the subjects gender
3. age_range: a string that denotes the age range of the subject
4. forrest_seen_count: an integer that denotes how many times the subject has seen Forrest Gump prior to the study

Out of the 451 recorded instances of whether a scene was day or night for Subject 1 and Task 1, Forrest Gump had 390 day slices and 61 night slices and 296 interior slices and 155 exterior slices. Subject 1 is a male who is 30-35 years old and had seen the movie 5 times prior to the study.

Though there were much more data in the demographics.csv file than the attributes given to a Subject object, we had deemed those data to be irrelevant for our purposes. For example, there were also data about each subjects music preferences and languages spoken. As we suggested before, there were a lot of directions that we could have taken with all this data, but these directions were too general and broad for us to be able to use them in conjunction with the fMRI data to do any conclusive studies.

The decision tree we created had a prediction accuracy of 83.5%. However, 86.5% of slices took place during the day, so this first tree is performing worse than chance.

# 5 Discussion

## 5.1 Strengths

Making our project reproducible is among our strengths. Our team has a firm understanding of command line and python scripts, along with the importance of grouping common behavior in general functions, such as not merging new functions into master unless those functions are unit tested, to ensure that a) functions work according to specified behavior and b) future changes do not break functions that had been working.

Learning how to work with the data in class was helpful because we have never worked with images in that type of format. We were not familiar with libraries like nibabel and thus it was very helpful to see how we could use them in our analyses. Exercises working with other fMRI data and various calculations we could perform on them was informative and relevant.

## 5.2 Obstacles

The main objective of the original study and paper was to gather raw data, mostly for other researchers to use in their own studies. Consequently, we were not given much direction in terms of what analyses we should initially conduct or reproduce. Also, as previously mentioned, the sheer amount of data we have to work with is almost overwhelming, especially given our limited computing resources. However, we assuaged this issue by using external hard drives to transfer data among group members.

We have differing prior experiences with Github and with research. Those who have more experience with git and Github help those who have less experience set up repos and push and pull correctly. Finding a method of communication was another issue we initially faced: we have been communicating via Facebook, and switching to communicate on Github to directly reference code has been difficult.

As a result, many issues on Github and pull requests have not been seen by other members. We have addressed this issue by reminding one another to look at pull requests on Facebook. We have varying areas of expertise, such as code review and statistical analysis, but we have used that to our advantage by having group members work on their areas of expertise.

Understanding the data - specifically in terms distinguishing the several forms of normalized data - and what conclusions we can draw is difficult, but we understand that that is beyond the scope of our project. Although our main goal is to become familiar with reproducibility and collaboration via the software tools taught in class, feeling shaky on understanding the statistical meaning of our results is disconcerting. For example, a deeper understanding of time series would enable us to perform stronger statistical techniques.

To successfully complete our project we need to keep the data we use well structured and organized, perform insightfully and interpretable exploratory data analysis, and perform hypothesis tests referencing Forest Gump scene features (although these tests are not expected to indicate causal inferences). Finally success also depends on doing these items in a reproducible and collaborative manner. We must strive to automate our analysis routines through command line interfaces and exploit the modern git and github workflow.

### 5.3 Potential Future Studies

Since the 20 participants have differing amounts of exposure to Forrest Gump, with some participants having never seen the movie before and one participant who had seen the movie twelve times before, a future study could examine the difference in voxel activation between participants who had never seen the movie before and those with high amounts of exposure.

Ideally, a future investigator could split the participants into a training set and testing set, and predict which participants have limited familiarity with the movie, defined as seeing the movie 2 or fewer times. However, there are a relatively small number of participants, only 20, since fMRI is expensive, so the predictive ability of that study would be rather limited.

## References

[1] M. Hanke et al., *A high-resolution 7-tesla fmri dataset from complex natural stimulation with an audio movie*, Scientific Data, 1 (2014), pp. 1–18.