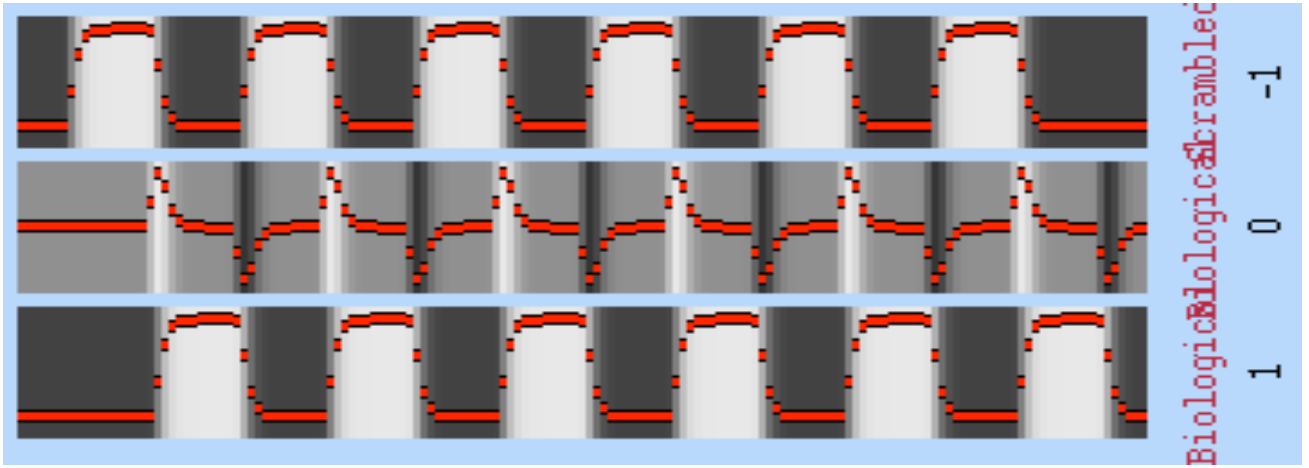
Exploring fMRI Data - Analyzing Brian Activity with Limited Data

Springboard Foundations of Data Science
Capstone Project

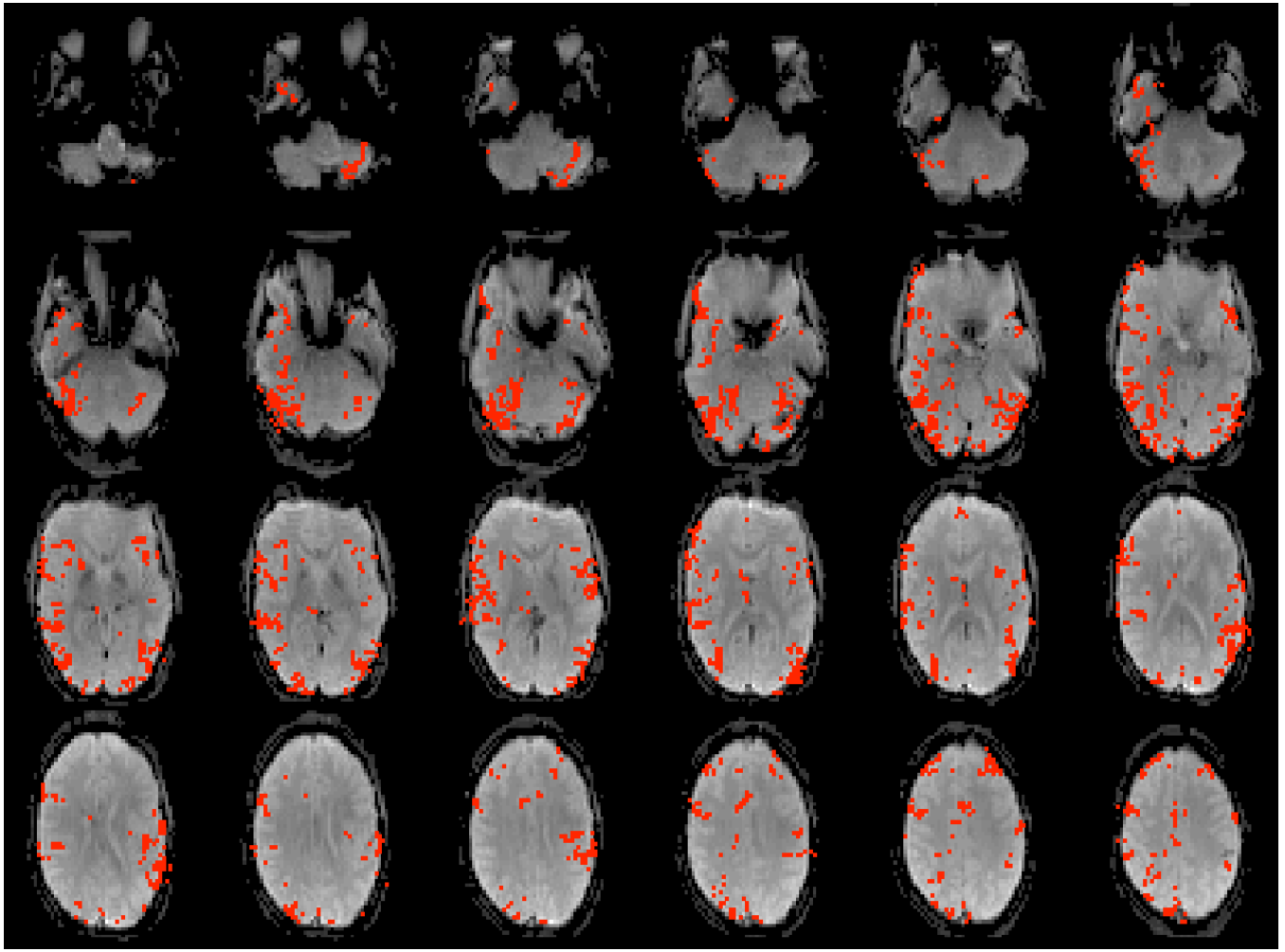
Michael An
Mentor: Joel Bangalan

Functional Magnetic Resonance Imaging (fMRI)

- Use fMRI to noninvasively study brain function
  - Indirectly measure brain function by measuring oxygenated blood flow (more blood flow, more activity)
  - Scan patient brains over time while performing a carefully designed, timed experiment

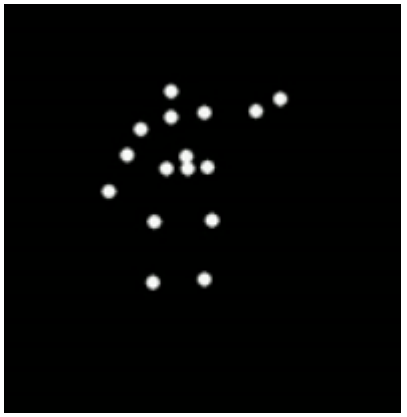


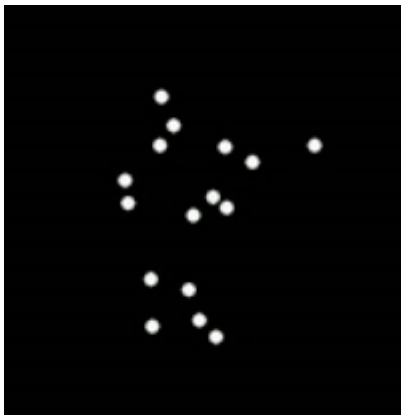
- Assess how blood flow in the brain regions change under different experimental tasks



The Dataset

- Biological Motion vs. Scrambled Motion Task





Advantages of Learning with Less Information

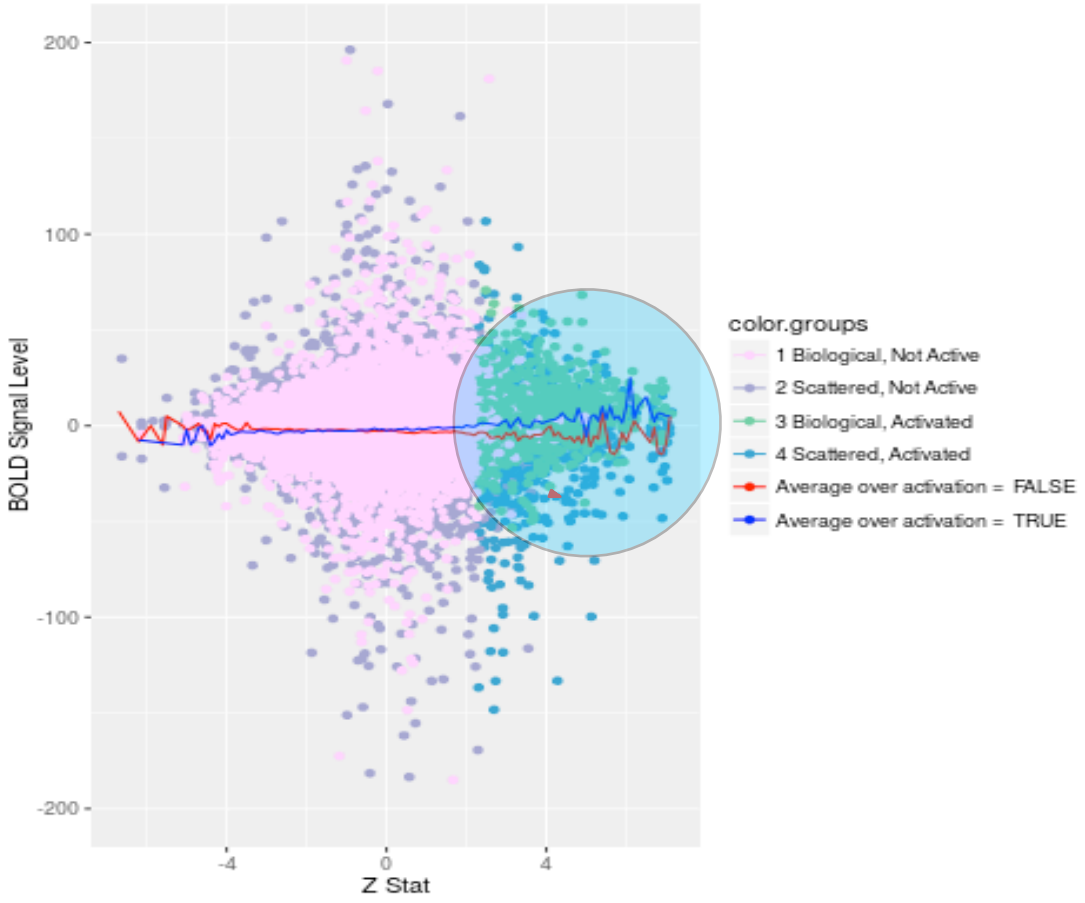
- Patient perspective: Scanning children, patients with disorders such as ADHD, autism, for long periods is difficult
- Data perspective: Modeling with less, but most salient features, reduces dimensionality of data, allows flexible usage of machine learning techniques

- Project is an initial exploration into this possibility of using less scan information (less runs of experiment)

Exploratory Data Analysis

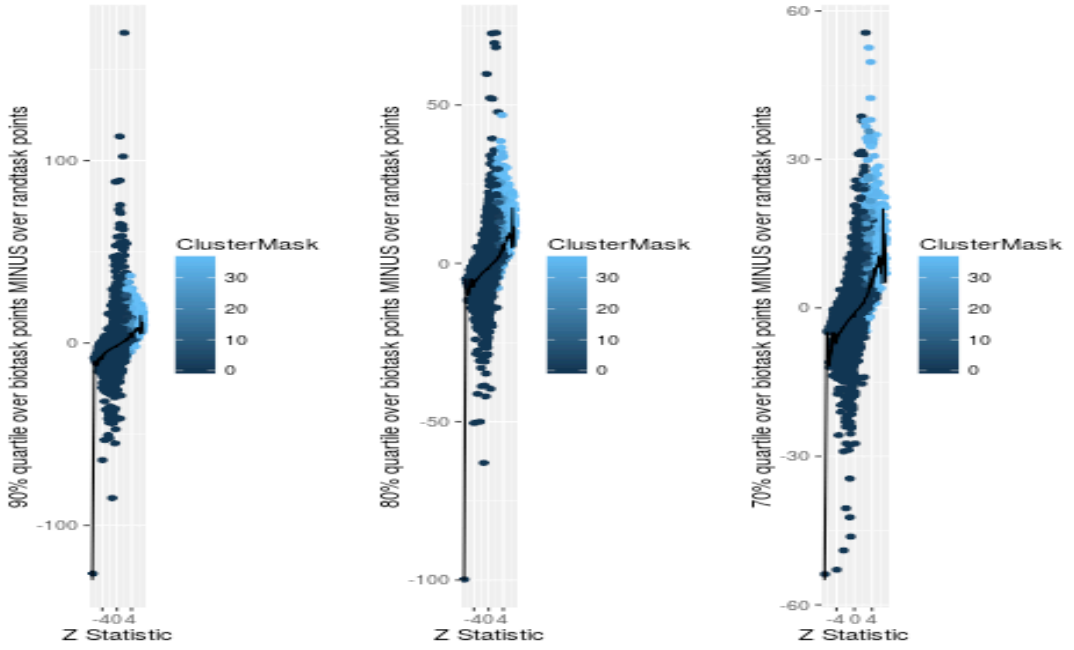
- Goal: Exploring the fMRI time series data to find good subsets of data, functions of data
- Analyze scatterplots over z stats (since z stats directly thresholded to determine activation at each voxel)

- For activated regions: Slight increase in BOLD activity for time points where biological task is being run
  - Suggests that BOLD activity levels themselves can be used to predict activation regions

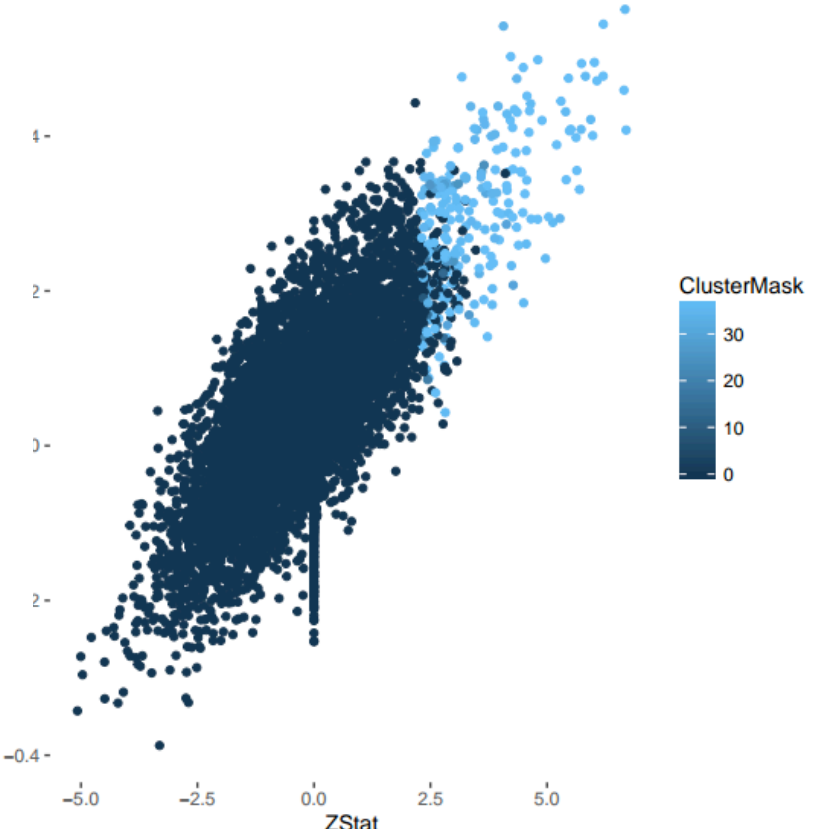


- Functions of data that have best spread of values between active (light blue), inactive (dark blue):
  - 70% quartile of biological task activity minus scrambled task activity, correlation coefficient of time activity to bio. task event paradigm

Quartiles vs Z Statistic



Correlation Coefficient vs Z Statistic



Takeaways from EDA

- Mean centered fMRI time activity gives different information than raw time activity
  - With centering, relative activity differences under the 2 tasks captured
  - Without centering, overall activity differences between active/inactive regions evident
  - Use centered data in analysis: interested in relative difference between the tasks
- 75% quartile, correlation coefficient to event paradigm, show greatest difference between active/inactive regions

Modeling – Part 1

- Logistic Regression model to predict functional activation/ no activation.
- Model 1: Predictors – All time activity, points, 70% quartile, correlation coefficient

- Only 3 time activity points shown to be significant in model (smallest p value = .0014)
- Correlation coefficient much more significant predictor than others ($p < 2e\text{-}16$)
  - May be due to multicollinearity
  - Remove this from model

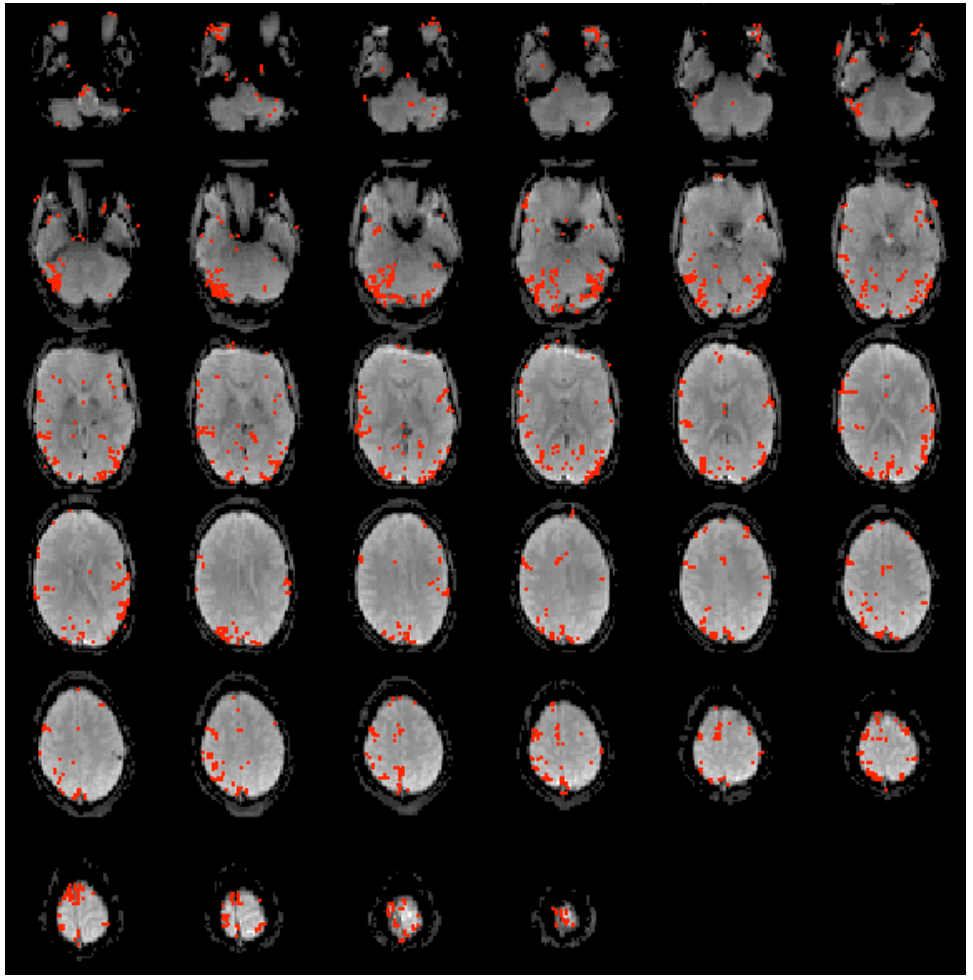
Model #2

- Removing correlation coefficient as predictor
- Now, many times points are significant (45 time points out of 159)
- Very high true positive rate – 99%
- Acceptable true negative rate – 70%

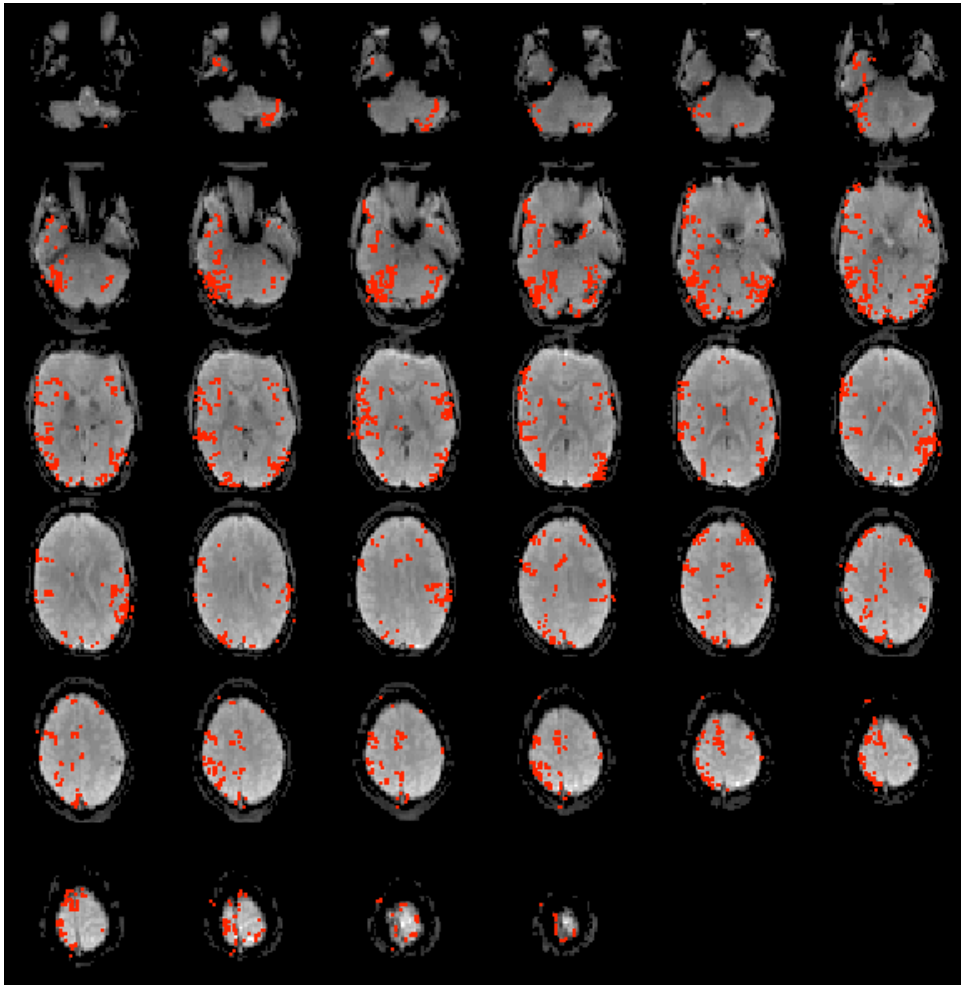
Model #3

- Use only 2 runs of the time series, not all runs
- Most times points are significant (max p value = .027 for significant points, with most p values < 2e-16)
- Similar true positive rate – 99%
- Lower true negative rate – 41%
- Qualitatively still reflect similar patterns

Predicted Activation Map
from Model #3



Precomputed
Group Activation Map



Model #4

- Run PCA over the data, and use select principal components as predictors
  - 48 components in total
  - First 40 components explain 95% variance
  - Would desire using less components though, even if you lose more information
    - try using only first 20
  - True positive rate – 99%
  - True negative rate – 40%
  - Very little degradation in quality compared to using all time points over 2 runs (model #3)

Takeaways

- Qualitatively, using less runs still gives a similar picture of activation
- More false negatives --> miss certain regions. May be removing too many active regions, but still better than many false positives.

Suggestions for What to Do Next

- This analysis is performed over a single subject
  - Train model on a group of subjects, and reassess if this can be used as a robust predictor.
- Train the model on actual “partially collected” data – less runs, incomplete scans, etc. Assess robustness