**Section 3: Enhanced Data Preprocessing after Initial Evaluation (11/4/2021 - 11/29/2021)**

* **Designed and implemented an Artifact Removal Algorithm based on the Central Limit Theorem and Least-Squares Regression.**
* **Extracted SPEC features for additional reference.**
* **Examined the preprocessed dataset based on customized measurement for the removal of bad subjects.**

**Conclusion: Dataset was well cleaned automatically.**

**11/4/2021:**

Started implementing the data pre-processing process. Used the librosa package to implement the SPEC feature extraction on the data. Resulted in each session per subject with 25,876 features.Talked with Khuong Vo about the results.

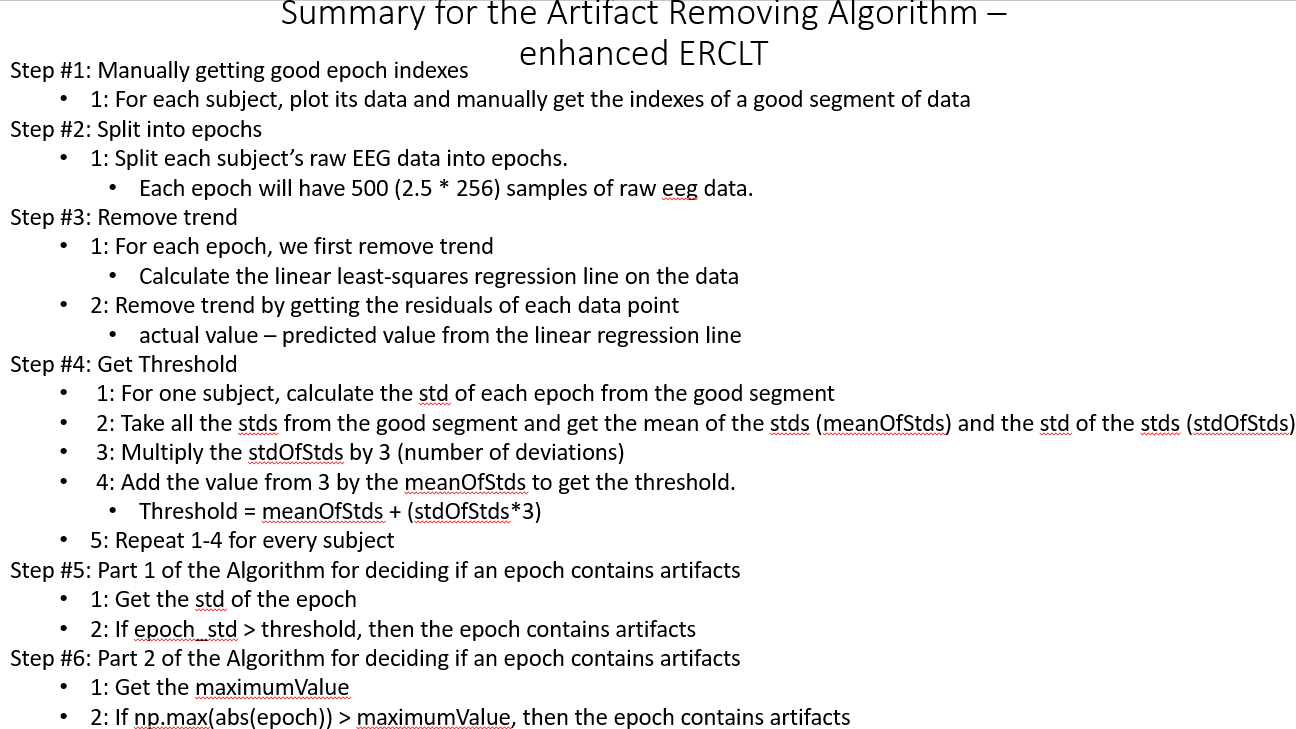
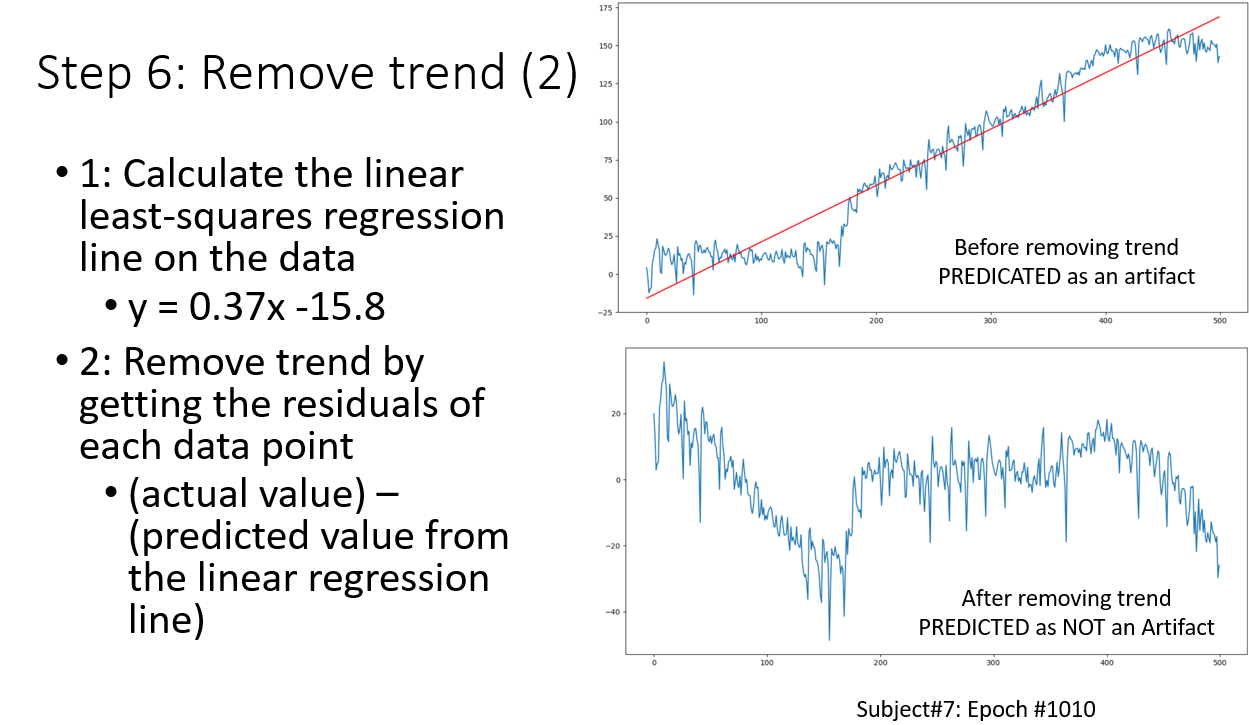
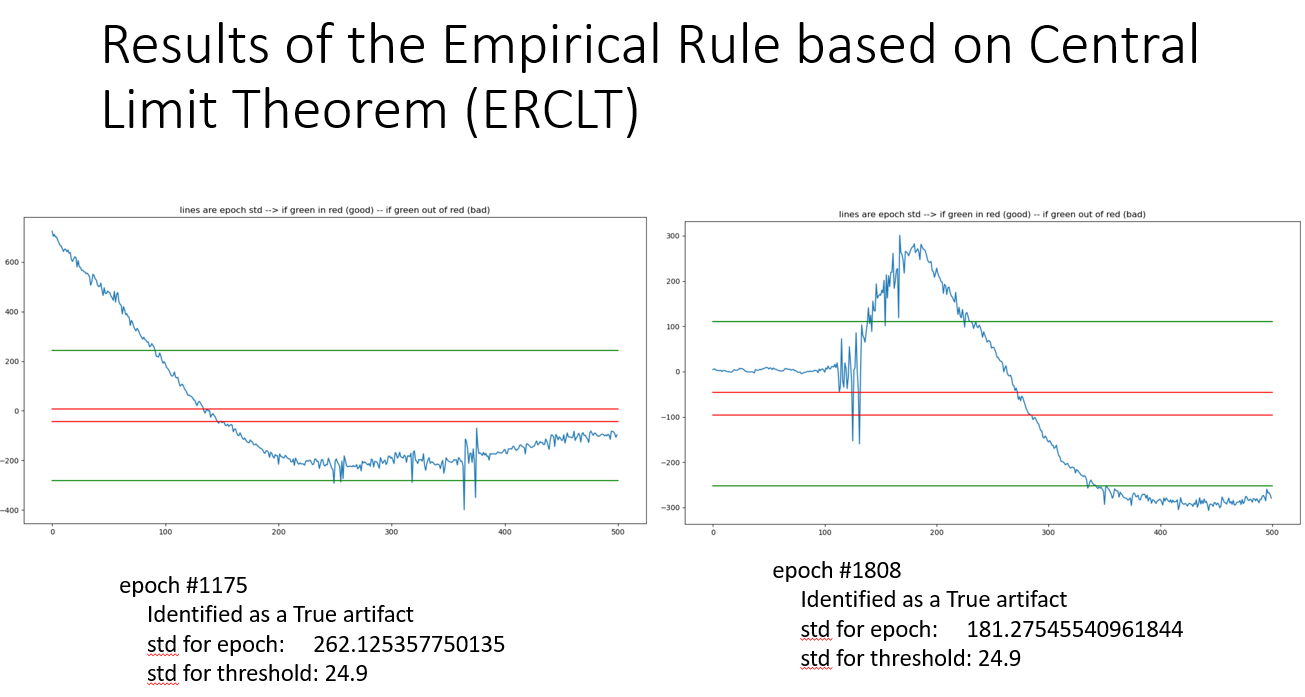
Concluded that I did the feature extraction incorrectly. Should have 224 features for each sample (same as BED paper), not 25,876. I also need to only use the RC Stimulus data, not all the data.

**11/12/2021:**

Revised my approach to extract the SPEC features: The RC Stimulus part of the BED dataset is 2 minutes long. For each 3 second signal, I will produce 224 features, resulting in about 40 samples for each session per subject (same as BED paper). Decided to put away the feature extraction algorithm and train a model on the RO Stimulus of the BED dataset as well as another model on RC and RO combined, to see if there is an

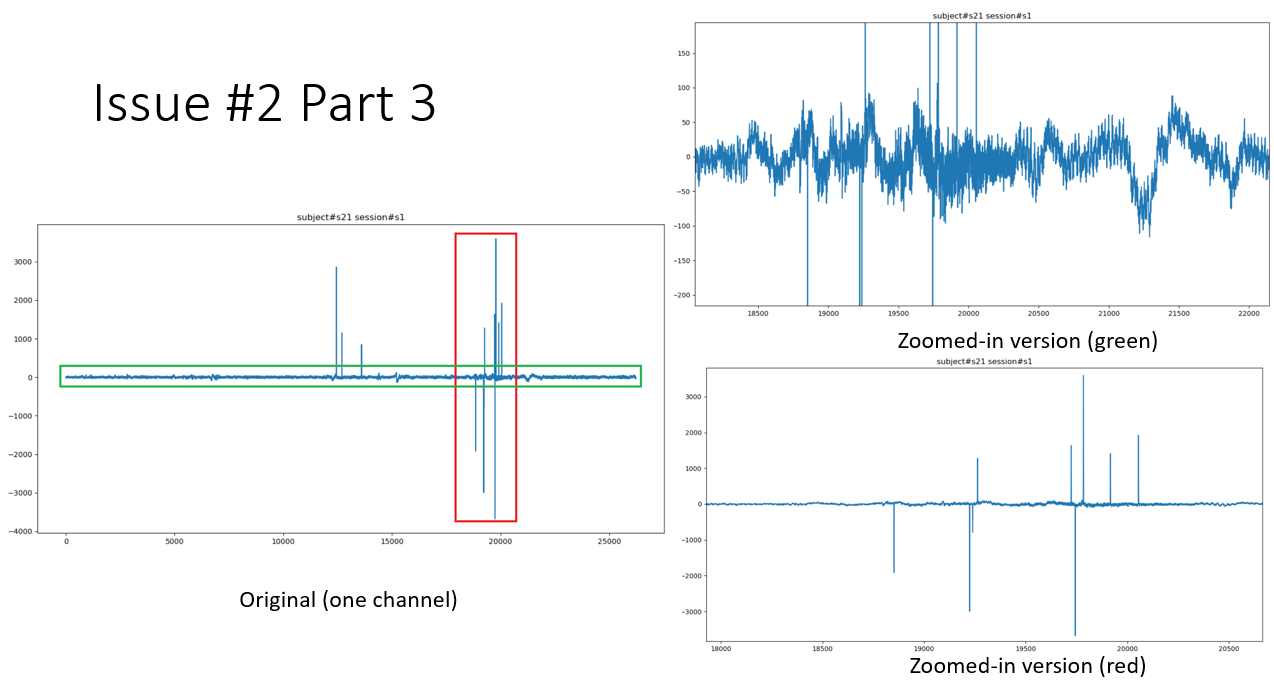
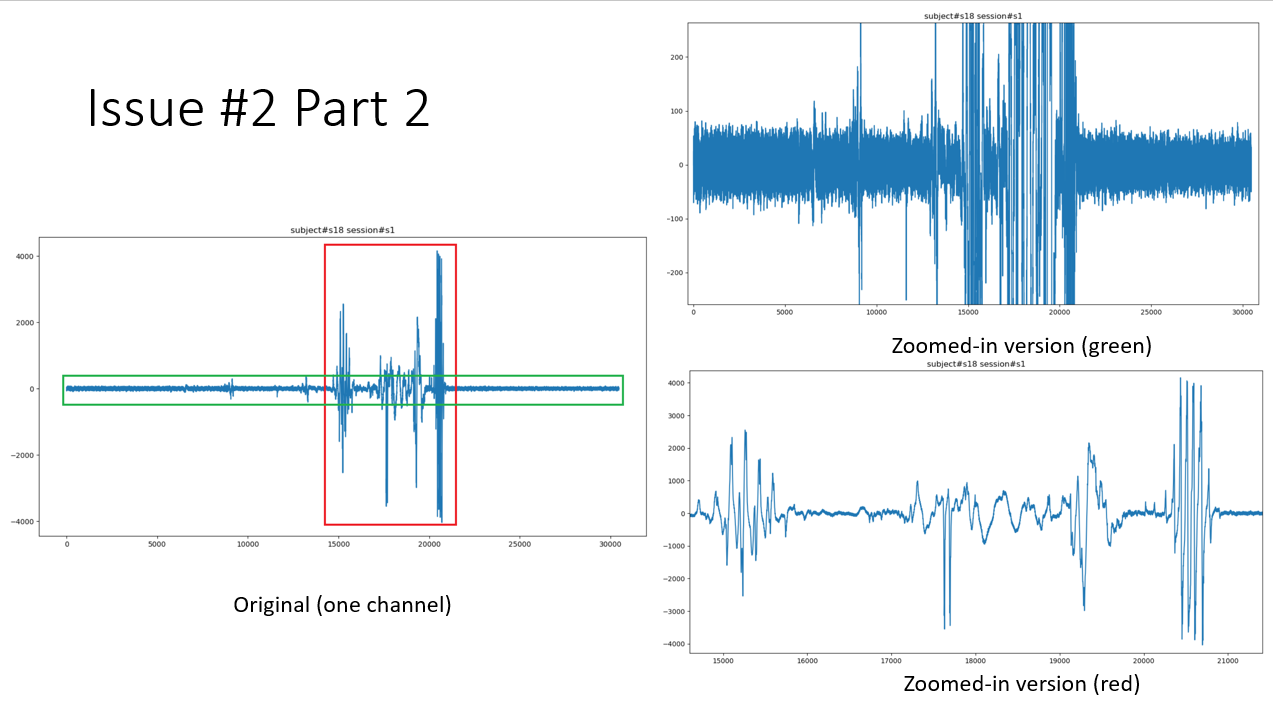
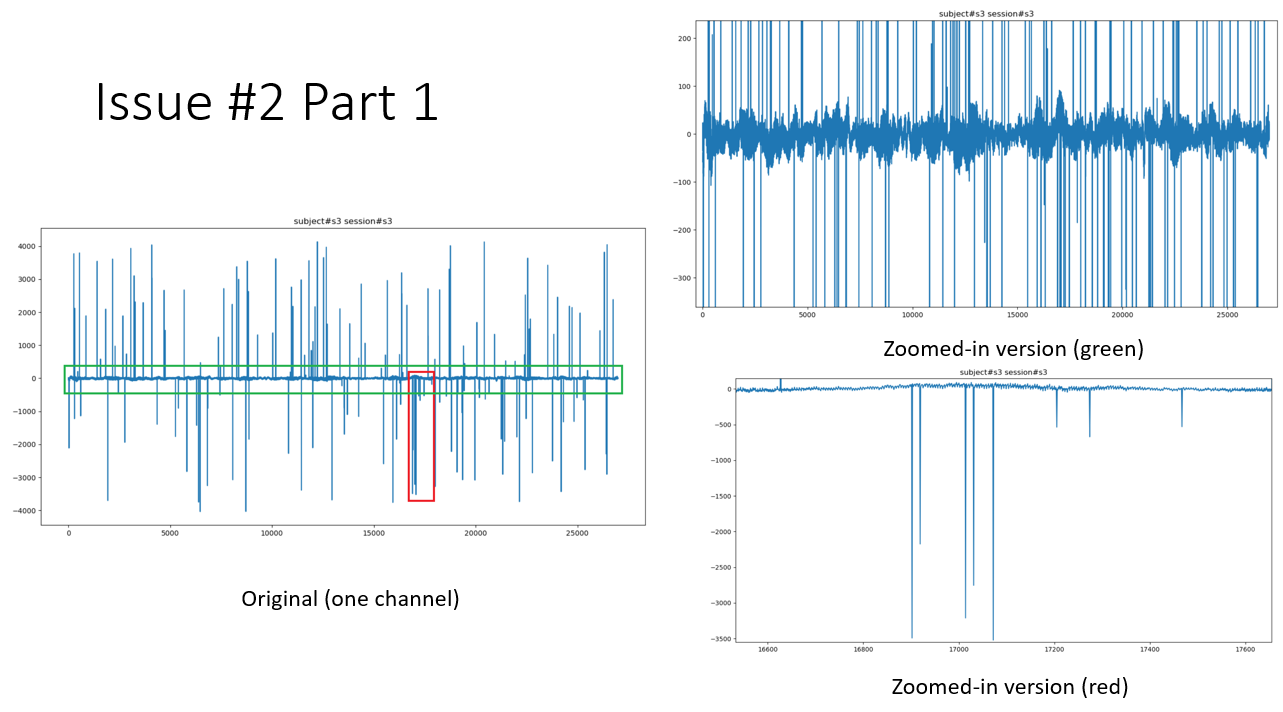
**11/17/2021:**

Designed my implementation of artifact removal using the Central Limit Theorem and Least Squares Regression.



**11/21/2021:**

Visualized the preprocessed dataset from Nee. Noted issues that I saw. Noted issues that I saw.

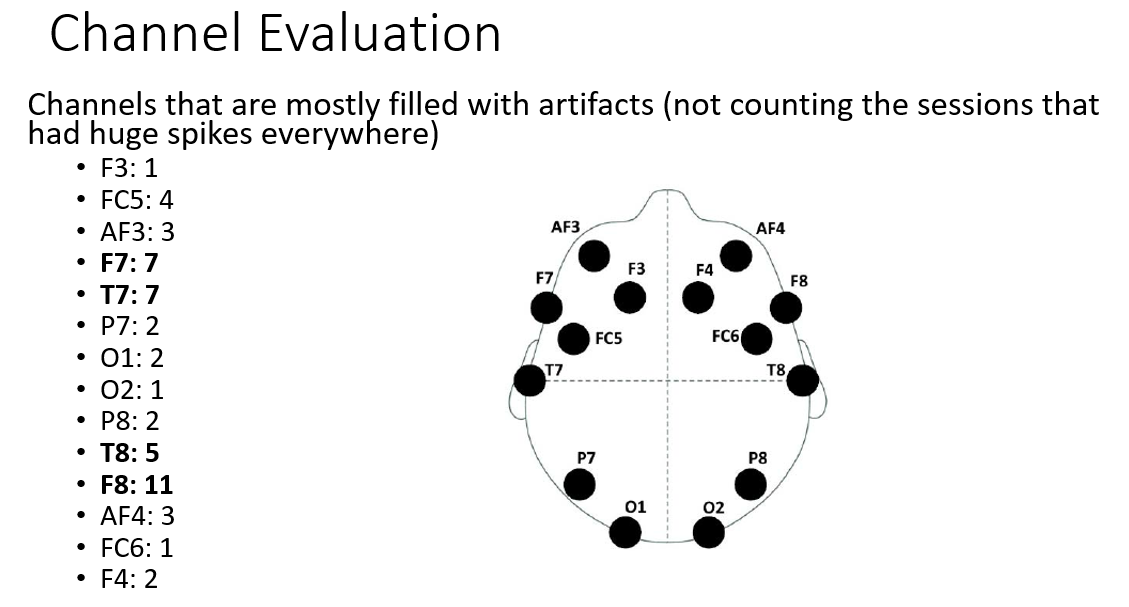
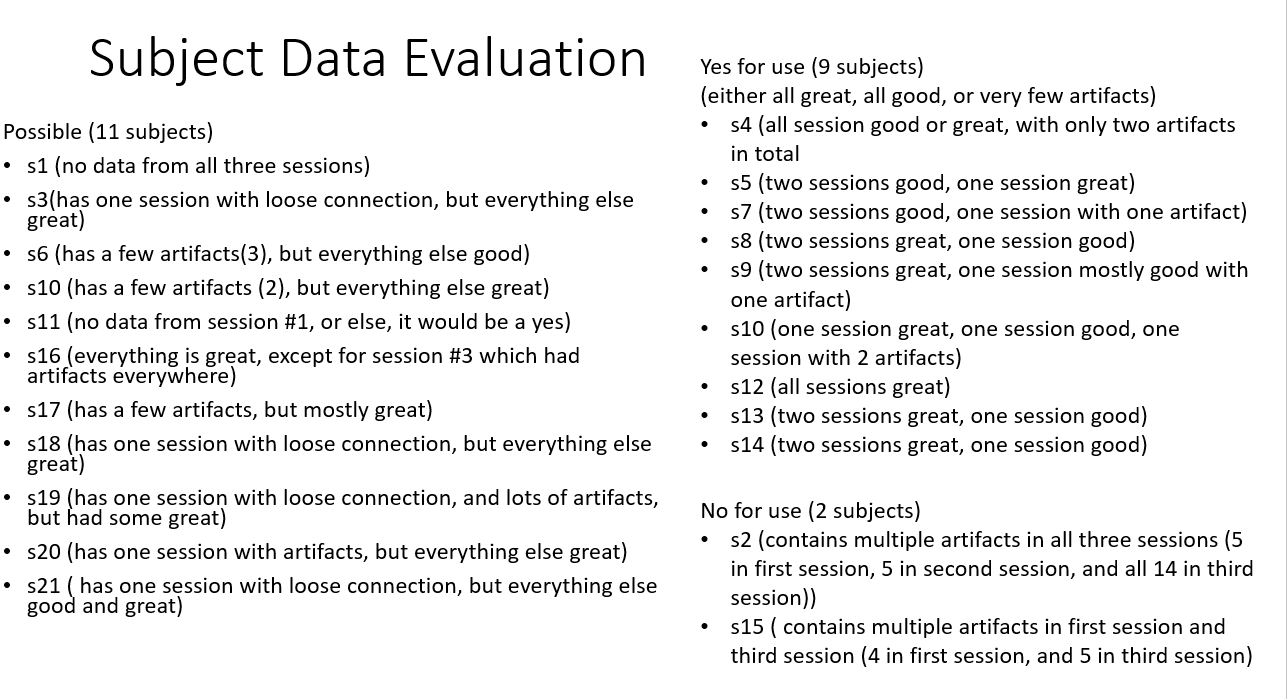
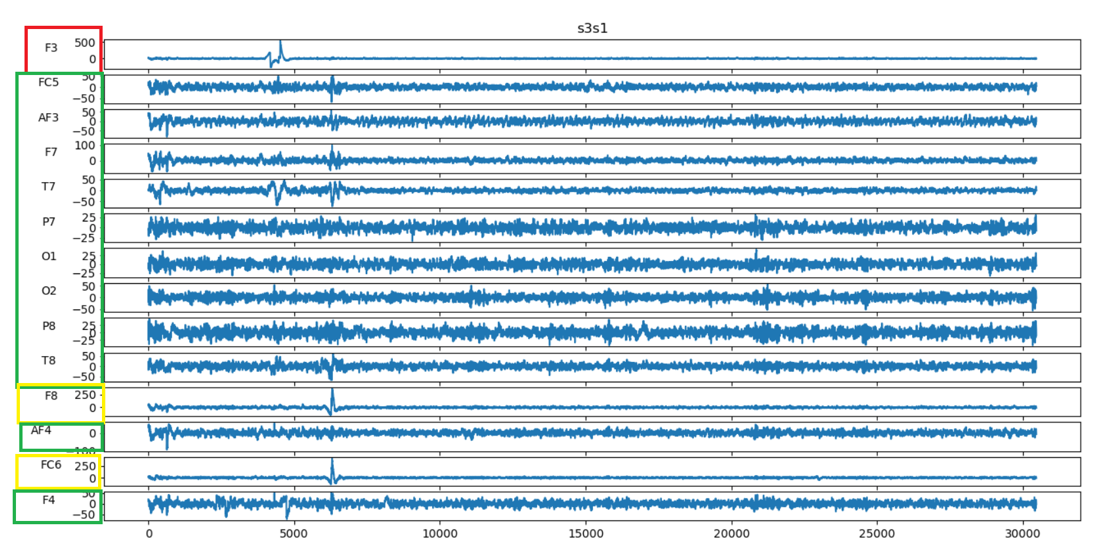


**11/22/2021:**

Talked to Khuong Vo about the issues. Decided that there are still a lot of artifacts. We need to evaluate the subjects to see which subject still has good data or bad data. We might have to remove some subjects, which is okay.

**11/27/2021:**

Did a comprehensive review on each subject and grouped them into the “yes”, “possible”, or “no” subject category. Artifacts are defined as a huge increase (+400) in the EEG’s signal. Also noted the channels that the artifacts came from.



**11/29/2021:**

Talked to my mentor about the subject data evaluation and the channel evaluation. Conclusions: It is great that we will have at least 9 subjects to use. It is surprising that the F7, T7, T8, and F8 channels had the most artifacts as those channels usually do not have a lot of artifacts. Decided that I will start to implement deep learning classifiers. I need to research about them.