Chad Leonard

AMERICAN EPILEPSY SOCIETY SEIZURE PREDICTION CHALLENGE

The Kaggle Competition

- **The Challenge Question**: Is it possible to accurately predict when a subject with epilepsy is going to have a seizure?
- **Epileptic States**: There are two epileptic states of interest for this competition – interictal and preictal
  - **Interictal States**: Defined as the state between seizures, or the baseline state.
  - **Preictal States**: Defined as the states prior to a seizure.
- **The Goal**: To predict preictal states.

Null & Alternative Hypotheses

- **Null Hypothesis**: It's not possible to accurately classify preictal brain states in dogs and humans.
- **Alternative Hypothesis**: The preictal brain state in dogs and humans does exist and can be accurately classified with naturally occurring epilepsy.

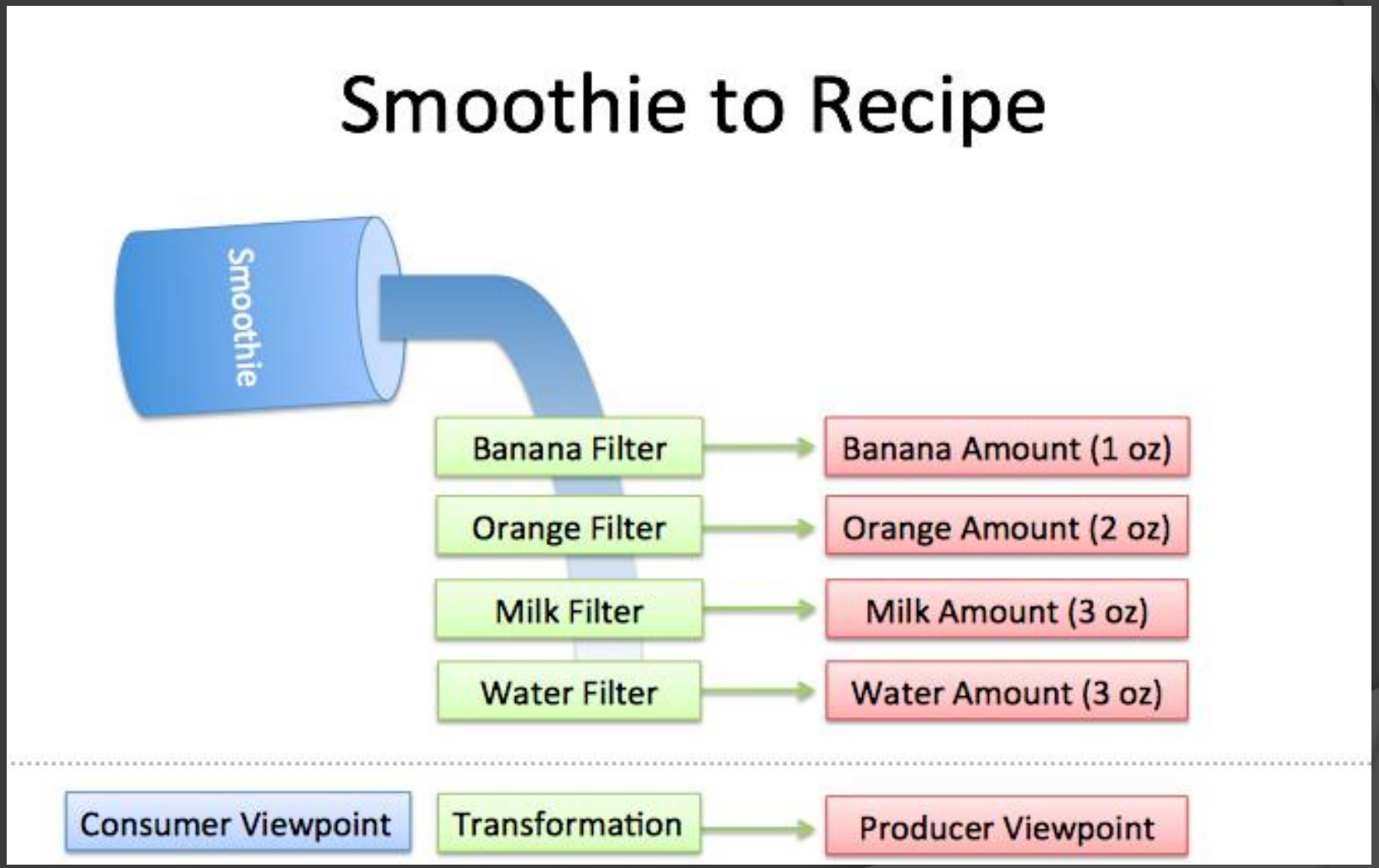
The Raw Data

- **Raw Input Data**
  - There are 7 epileptic subjects (5 dogs and 2 humans) represented in the data provided by Kaggle
  - Each subject has hundreds of files
  - Each file is either an interictal file or a preictal file
  - The files cover 10 minutes worth of brain activity, that's about 239k measurements taken every .002 seconds
  - The channels are the names of the electrodes placed on the subject's head and body. They are also the predictor variables
  - There are 239k measurements per channel

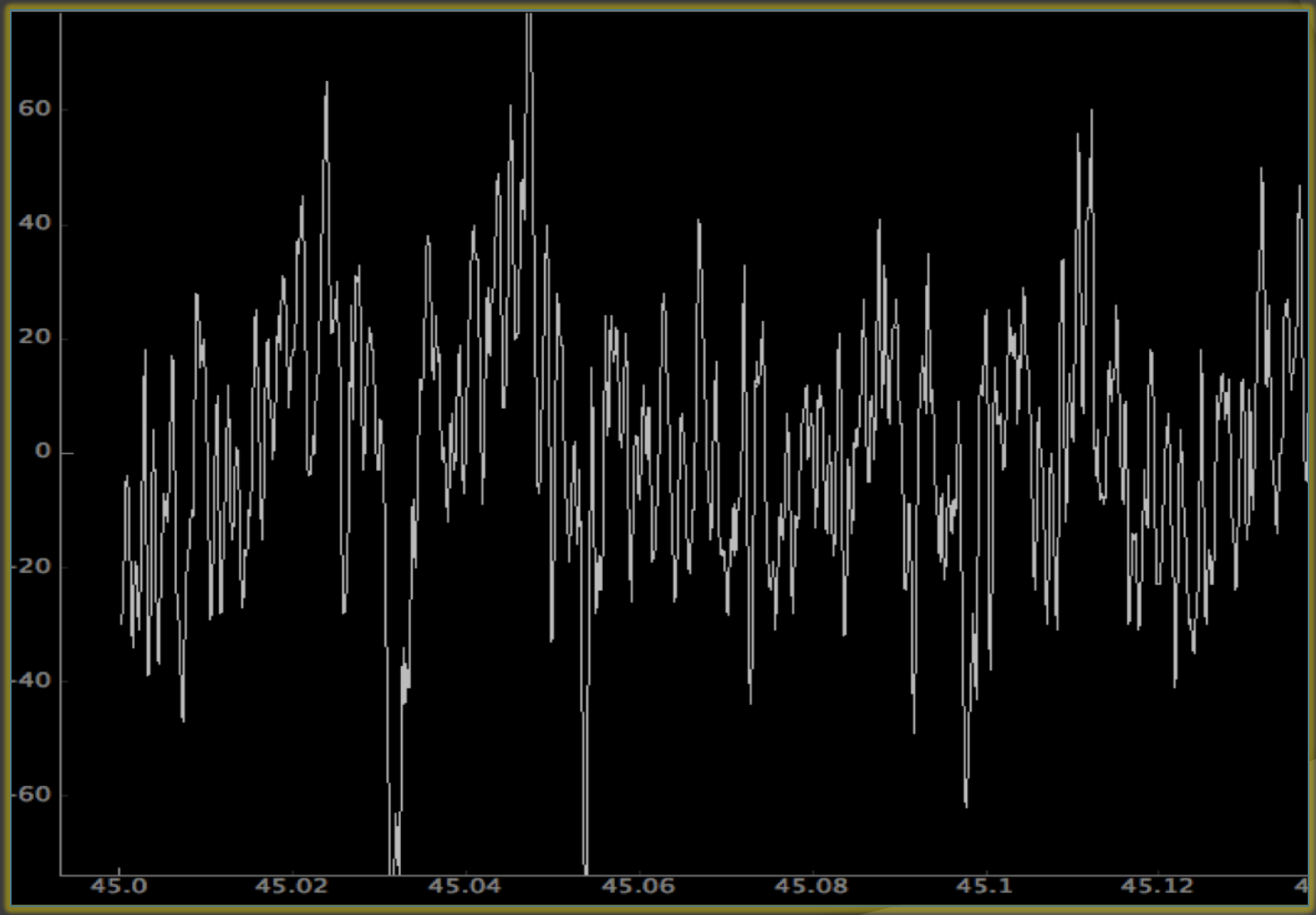
Data Transformation

- **Fast Fourier Transform (FFT)**:
  - **What does the Fourier Transform do?** Given a smoothie, it finds the recipe.
  - **How?** Run the smoothie through filters to extract each ingredient.
  - **Why?** Recipes are easier to analyze, compare, and modify than the smoothie itself.
  - **How do we get the smoothie back?** Blend the ingredients.

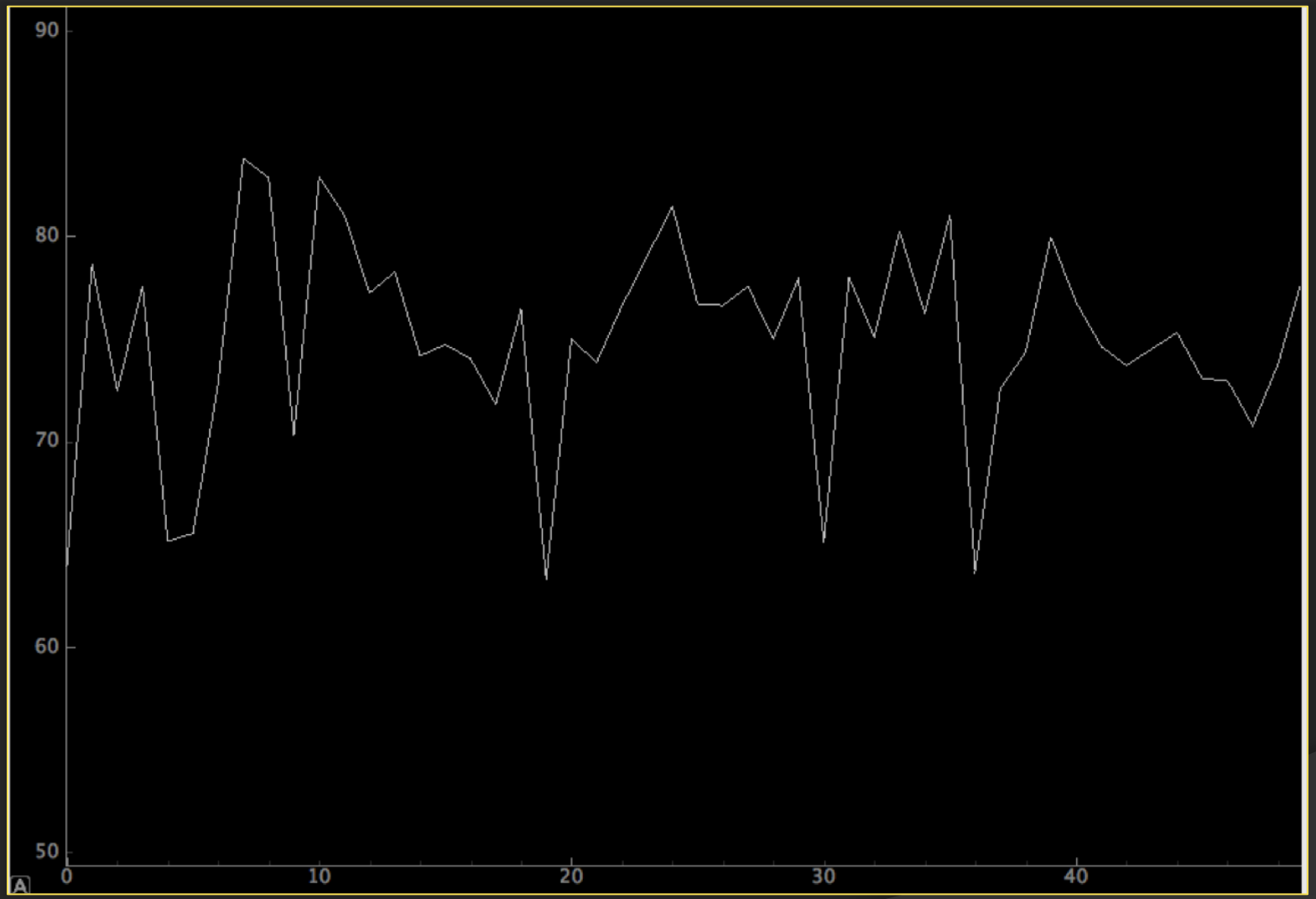
FFT Smoothie to Recipe



Plot of Raw Data–Channel One



Plot of FFT Data – Channel One



The Methodology

- **Read Data**:
  - Read Matlab files into DataFrame
- **Transform Data**:
  - FFT as described above
  - Each file condensed down to a single record
  - “ictal_ind” binary variable created with a value of 1 for a preictal file else 0
  - Subset of the transformed data is used. Of the 239k record only 7200 used.
  - Those 7200 time intervals are divided into 24 segments of 300 time intervals each per channel.
  - The 24 segment are then averaged.
- **Feature Selection**:
  - Logit
    - Choosing channels with p-values >=.05
  - Random Forest
    - Trial and Error from the list created from the Logit Regression logic above.

The Methodology Con't.

- **Machine Learning Algorithms**:
  - **Logistic Regression**
    - Initially used for prediction but proved unstable
    - Eventually it was just used for feature selection only
  - **Random Forest**
    - Used sklearn.ensemble's RandomForestClassifier

Random Forest

- **Random Forest**:
  - Used sklearn.grid_search's GridSearchCV to experiment with various numbers of trees
  - Chose the number of trees that produced the best ROC AUC Curve score on a 5 fold Cross Validation of train/test data
  - After the number of trees was determined, train_test_split was used to divide the data into test and training datasets and modeled
  - The predictions for Kaggle were based on this model

Test Results

Subject	Number of Trees	Random State	ROC AUC Curve
Dog_1	60	1	0.82202
Dog_2	30	1	0.92927
Dog_3	50	1	0.87484
Dog_4	40	1	0.86558
Dog_5	30	1	0.98432
Patient_1	40	1	1.00000
Patient_2	40	1	0.88461

Conclusion

- The winning ROC AUC score was circa .84.
- My ROC AUC score for the Kaggle competition was only .54.
- The ROC AUC values that were recorded for each subject using data from the train/test split were much better.
- Not sure what the cause of the discrepancy between my scores on my test data versus my score on the Kaggle's test data. Possibly too much overfitting or Kaggle's test data is not representative of their training data.

Lessons Learned

- Try to avoid transformations that are not fully understood, i.e. FFT.
- Get something modeled early.
- If in a Kaggle competition, try and make many submissions and use results as feedback.
- Have a better plan for Cross Validation and testing. Incorporate feedback from submissions into testing.