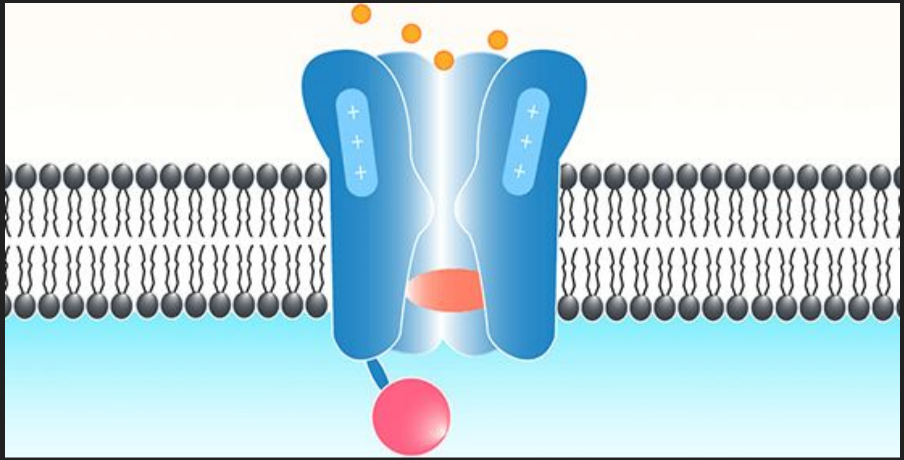
Predicting Ion Channels Opening

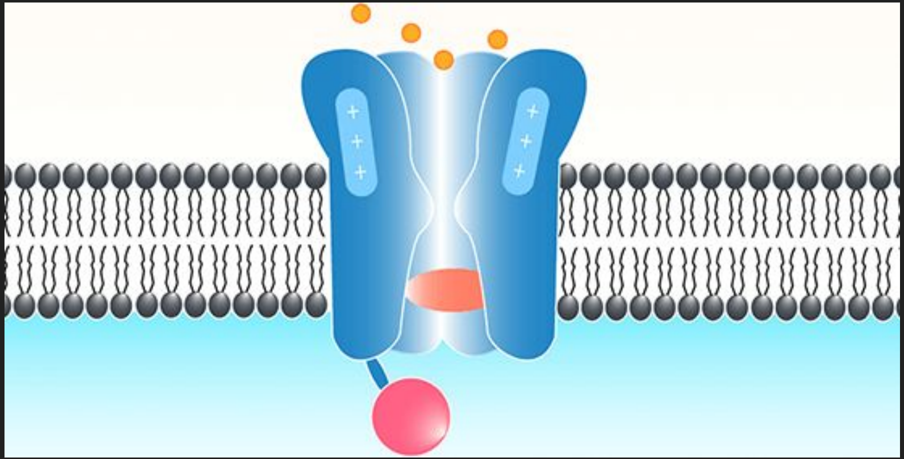
Jarod Carroll

The Ion Channel



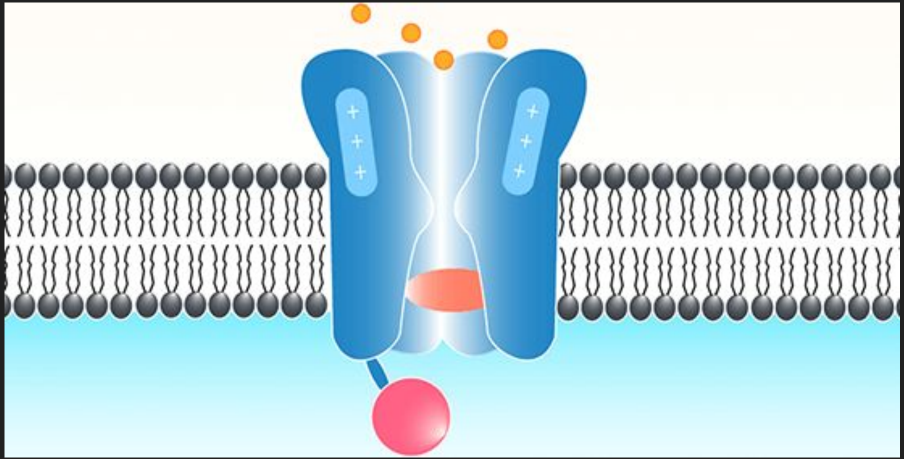
The Ion Channel

- Allows ions to pass through it



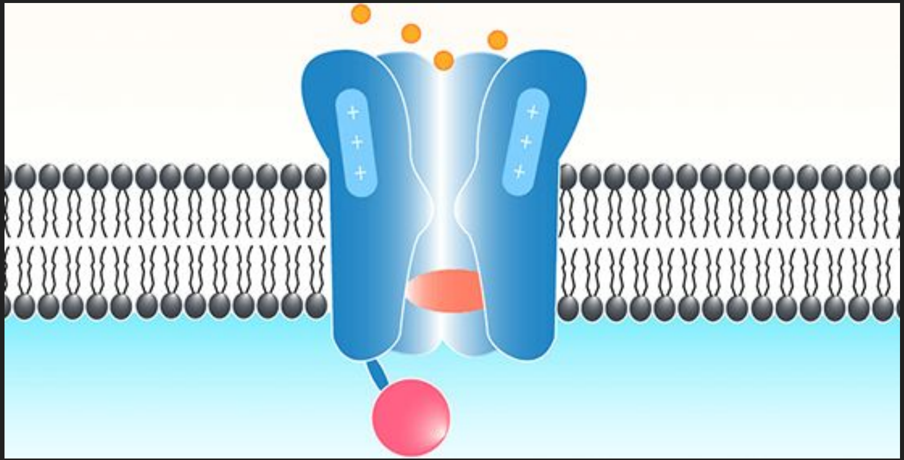
The Ion Channel

- Allows ions to pass through it
- Can change between states (e.g. open and closed)



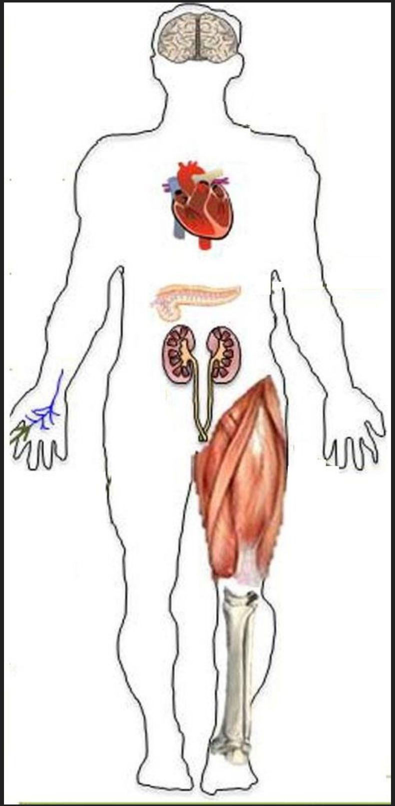
The Ion Channel

- Allows ions to pass through it
- Can change between states (e.g. open and closed)
- Responsible for almost all electrical activity in the body



Channelopathies

- Ion channels don’t function properly
- Leads to diseases like epilepsy, cystic fibrosis, heart arrhythmia
- We need to study these channels but how?

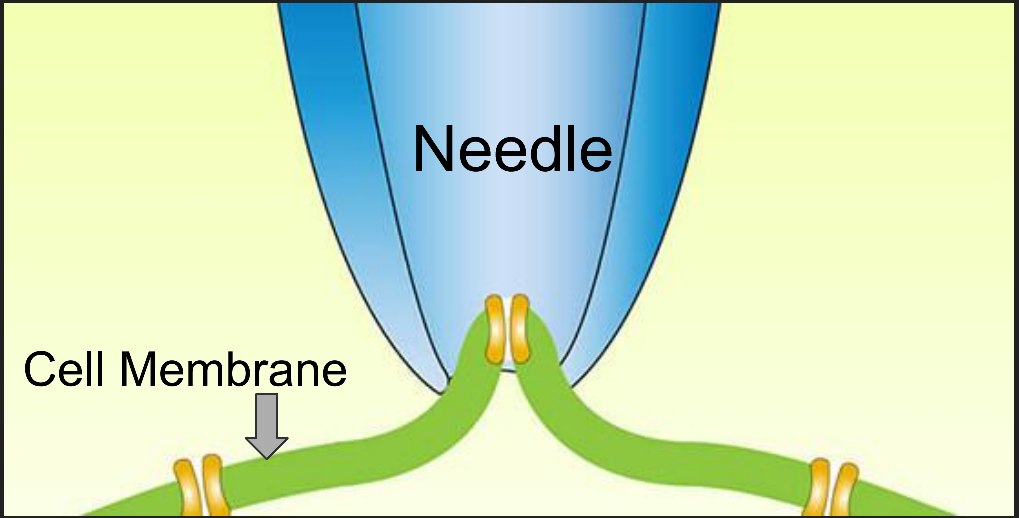


Patch Clamp

Allows for the capture and recording of a patch of the cell membrane.

Ideally, one or two channels, but can get many channels.

Multiple channel data discarded. This is a huge time/labor waste



Goal and Why it’s important

Be able to predict the number of open channels from patch clamp data.

Allows for the study of ion channels.

Being able to interpret multichannel data can increase the ability to study ion channels.

Process

- Get the data
- Process and transform it
- Break the data into its different waveforms
- Train a model
- Transform the test data
- Make predictions and get a score from Kaggle

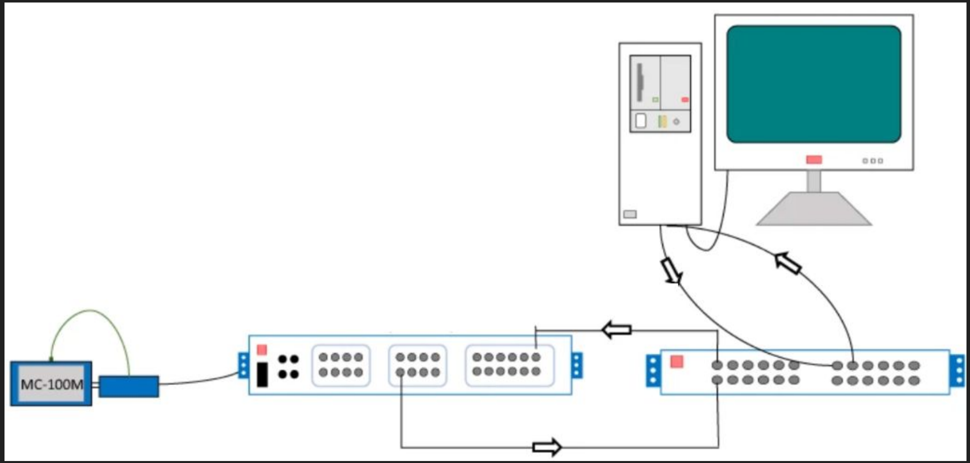
Where does the Data come from?

Downloaded from Kaggle https://www.kaggle.com/c/liverpool-ion-switching/data

Generated by a computer. Indistinguishable from real data.

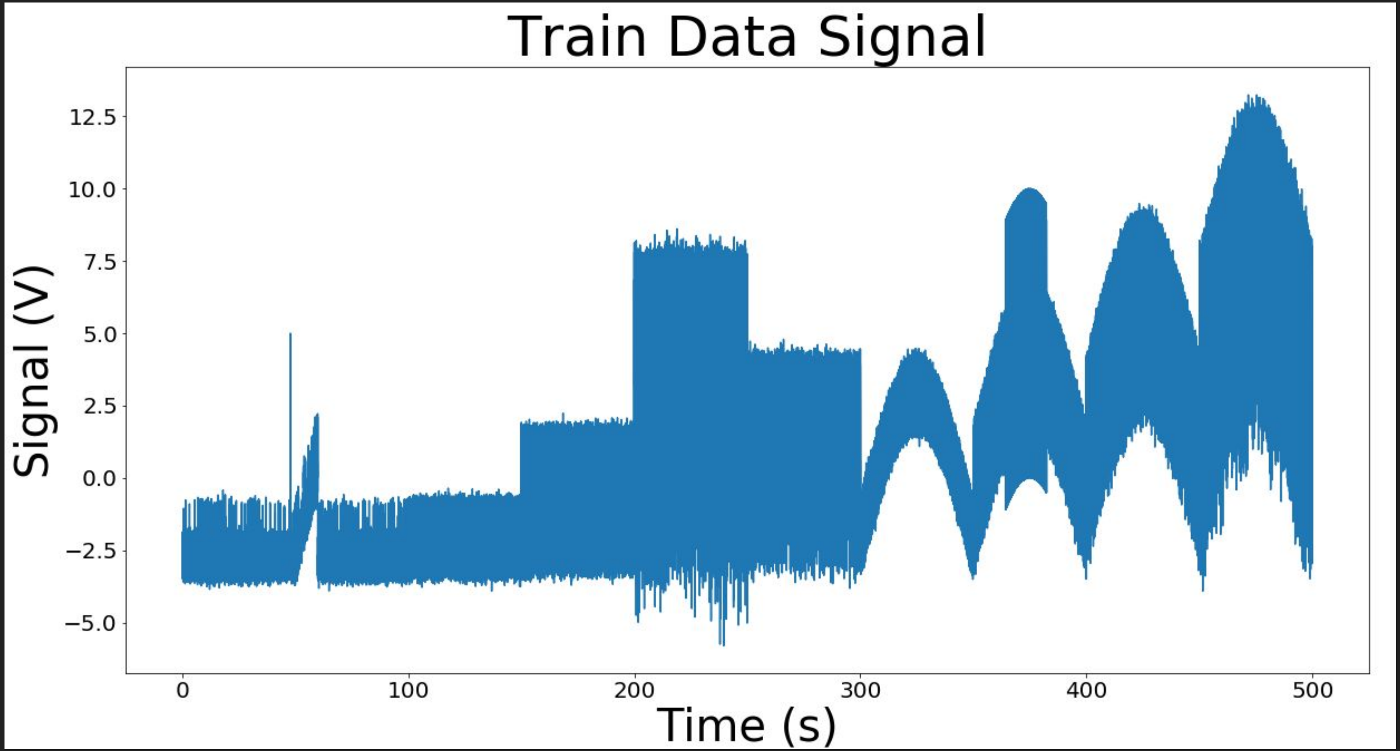
Properties of the data

- 10 batches
- 50 seconds per batch
- 10 kHz data (10k points per s)



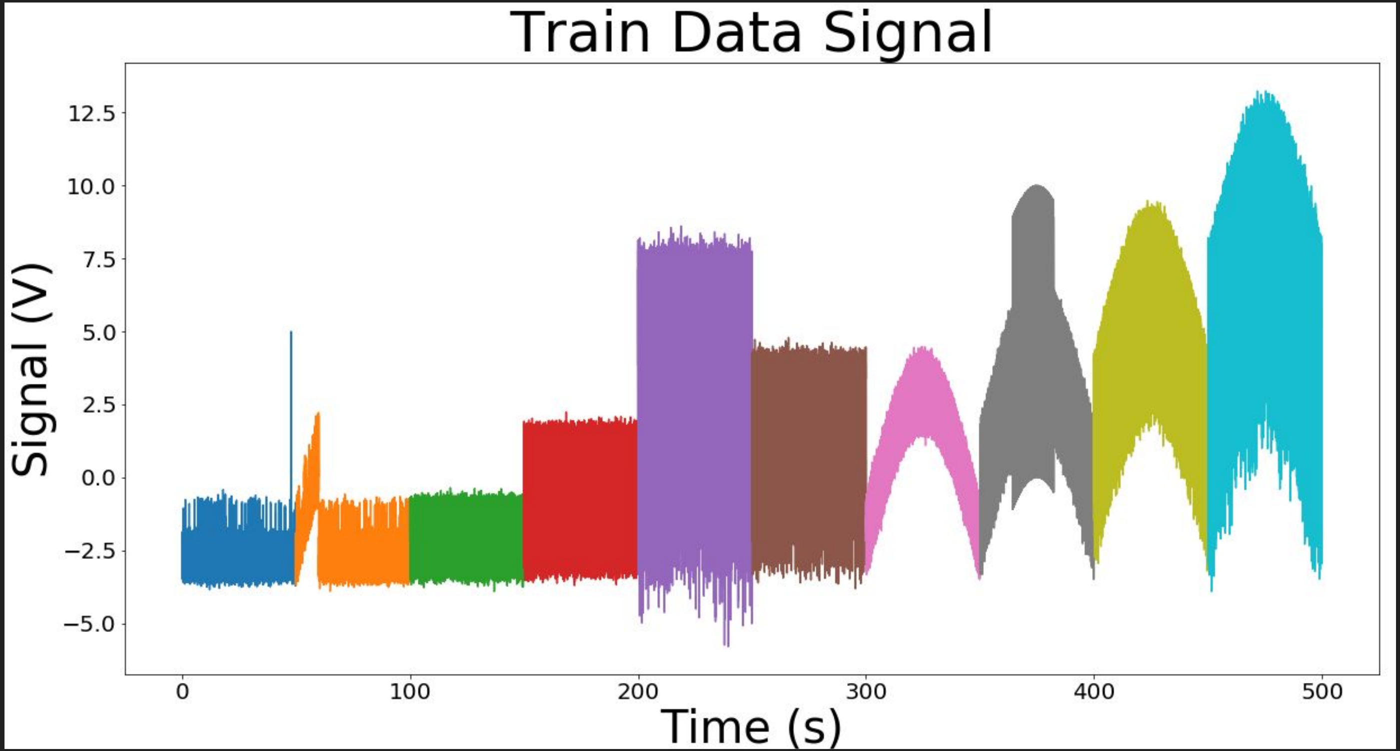
Inspecting the data

10 distinct batches

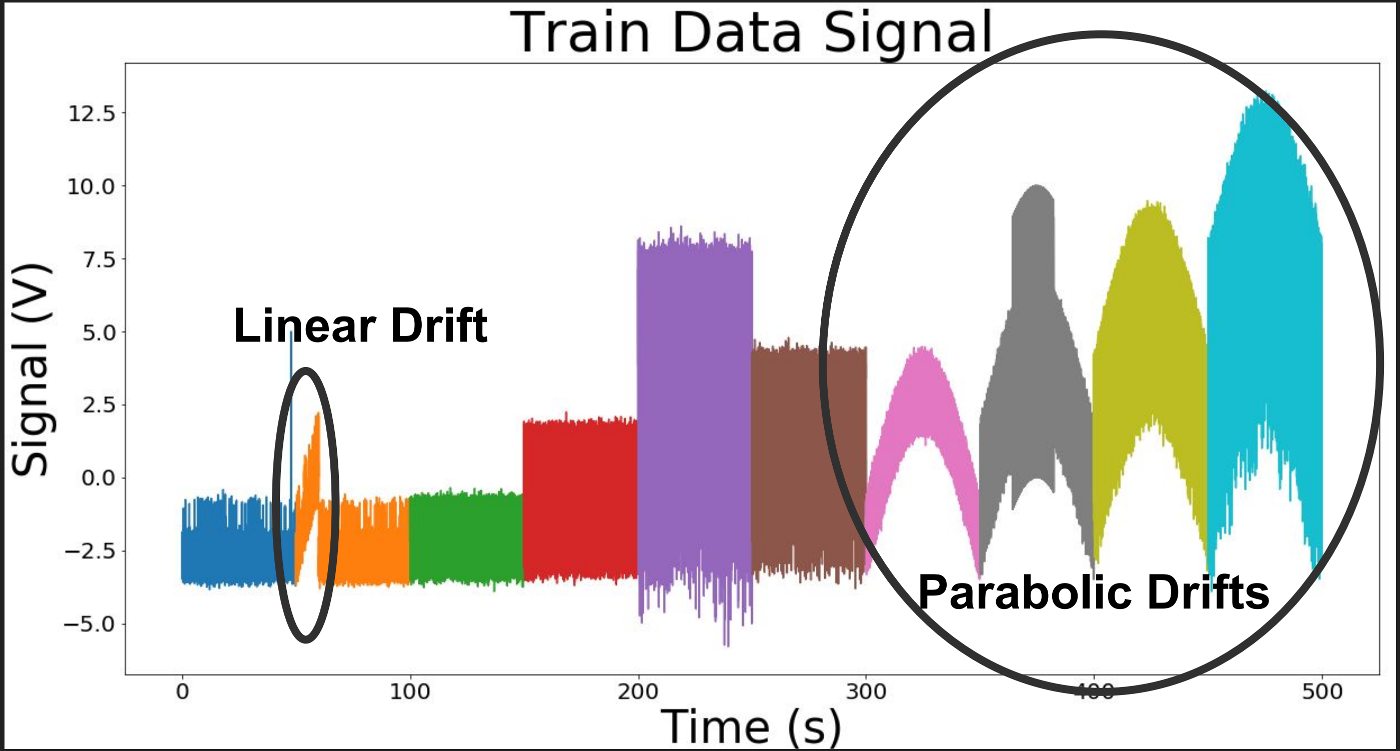


10 Sets of Data

(This was part of the details of the data set)



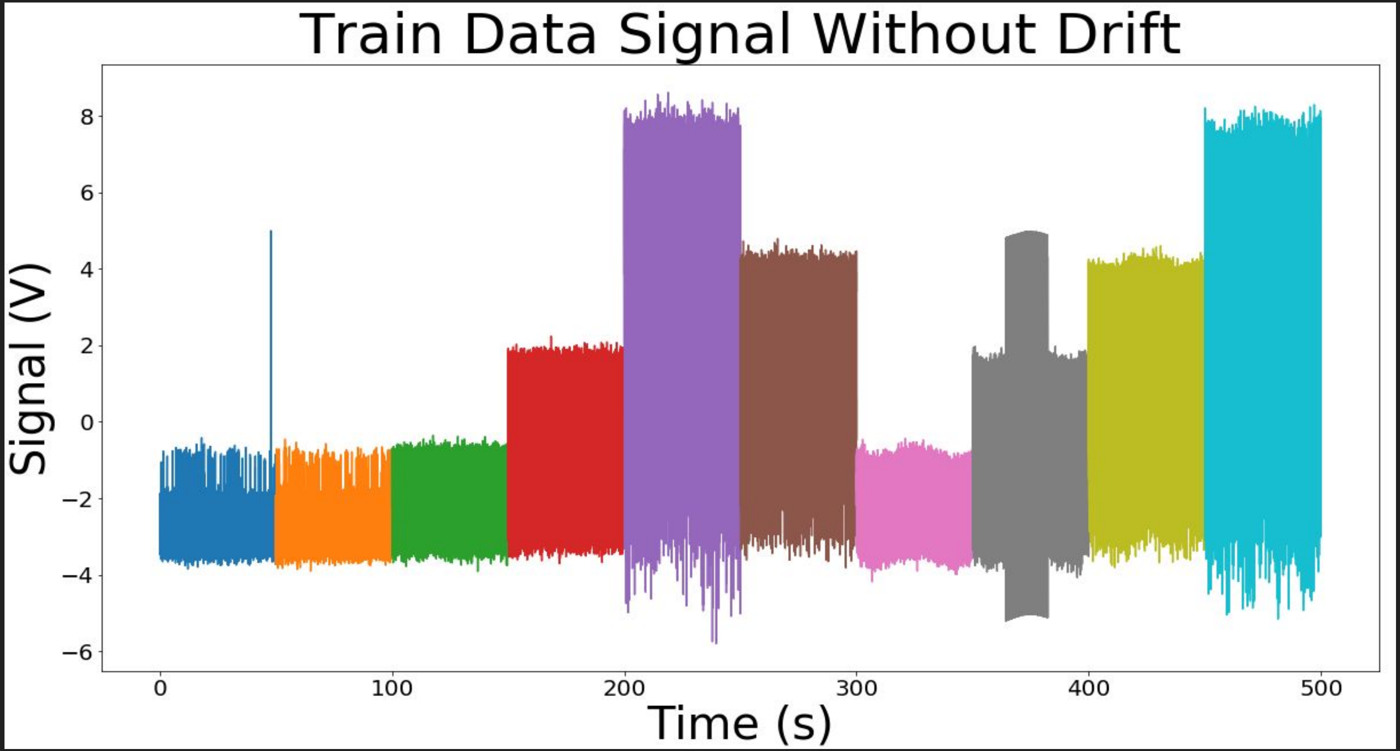
Drift



Train Data Signal
Linear Drift
Parabolic Drifts
Signal (V)
Time (s)
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
0
100
200
300
400
500

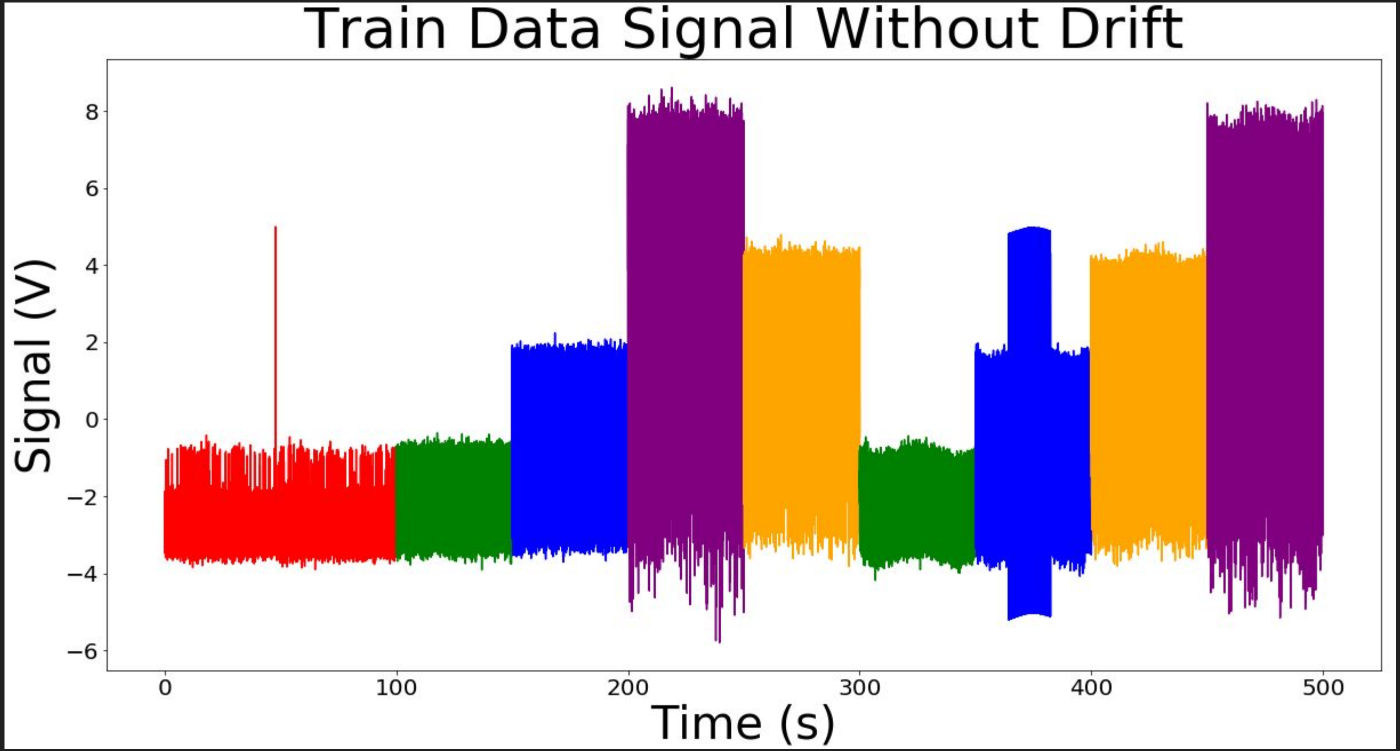
Remove Drifts

This was done by taking sections with drift and subtracting off an approximating function



Different Waveforms

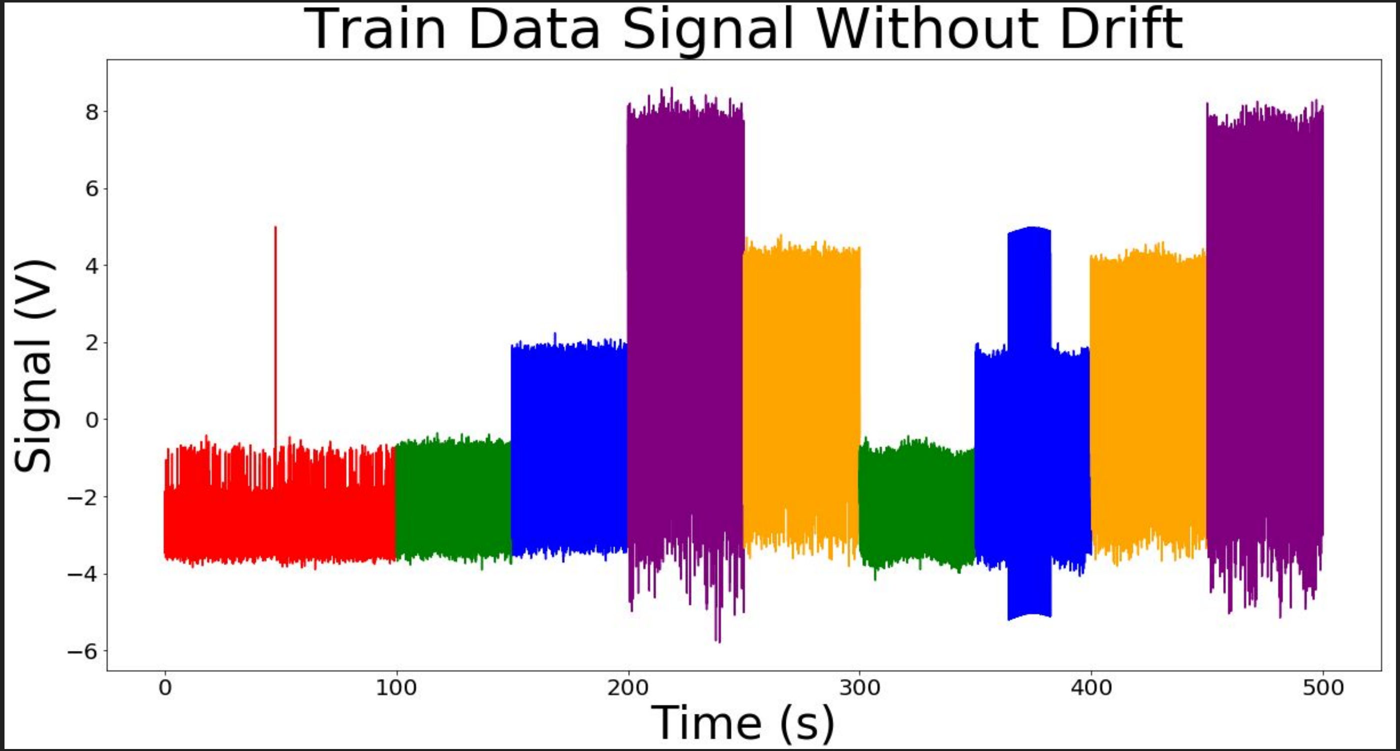
Matching colors = similar waveforms



Different Waveforms

Matching colors = similar waveforms

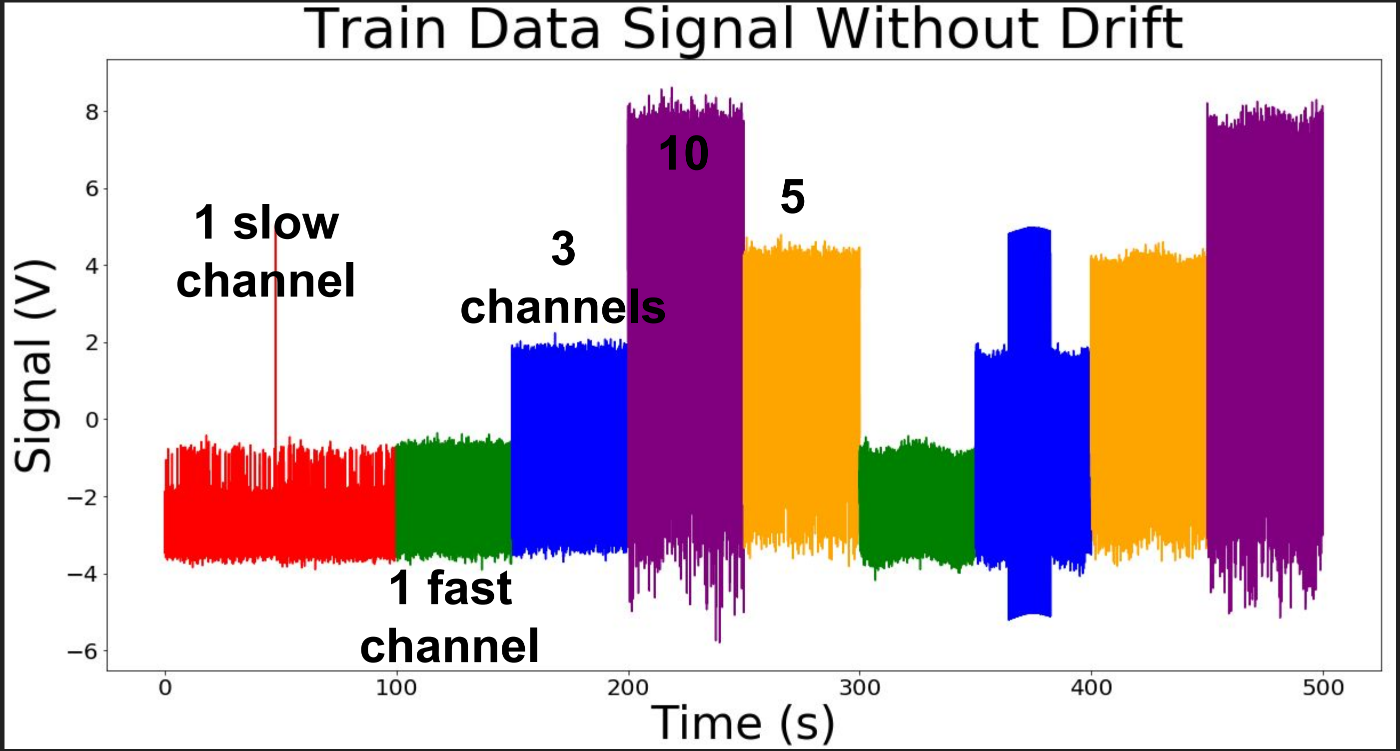
Distinguished by looking at differences in the signal data



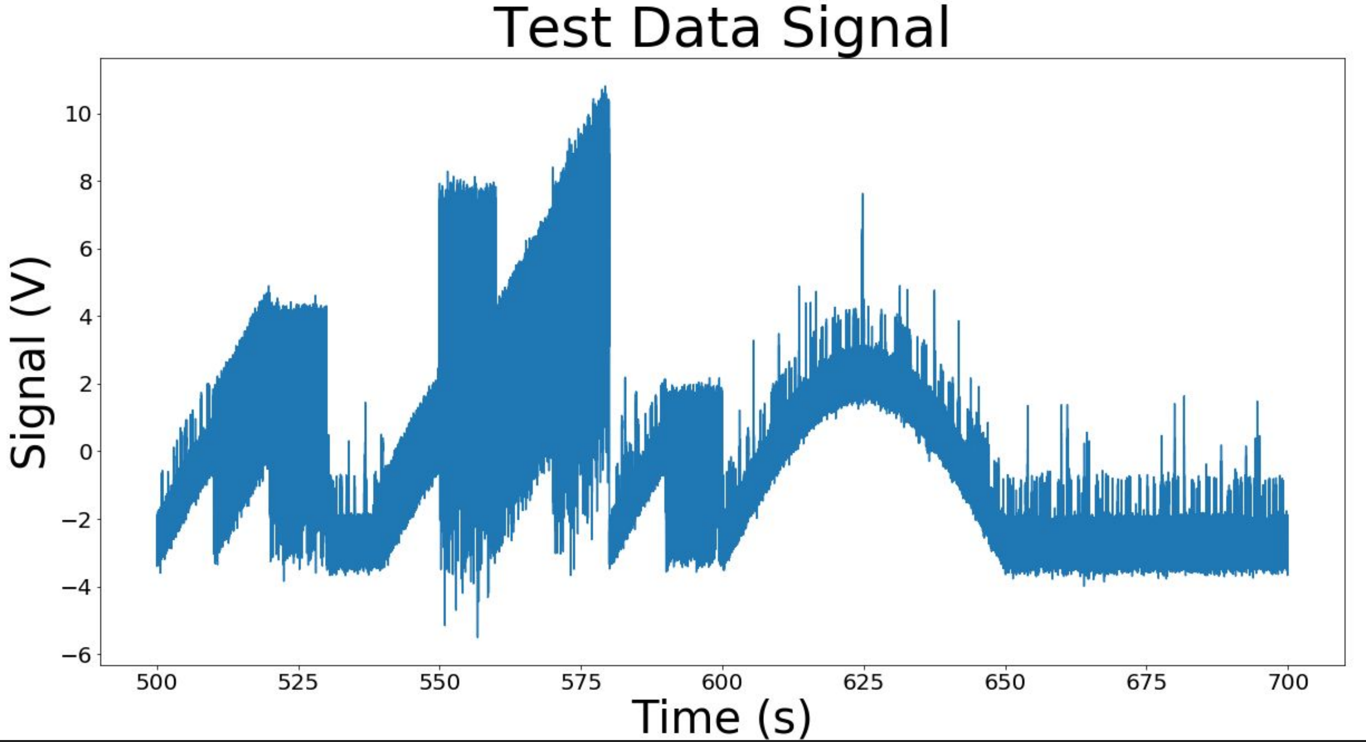
Different Waveforms

Matching colors = similar waveforms

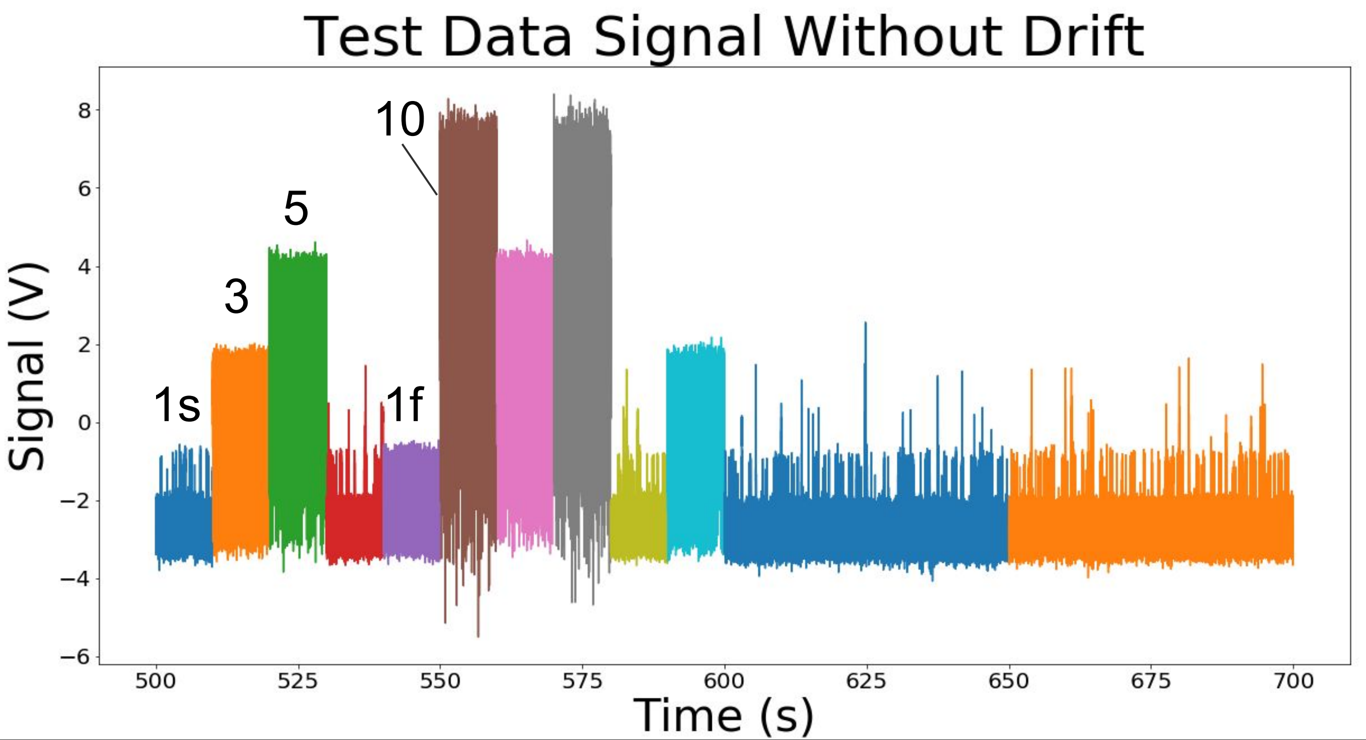
Distinguished by looking at differences in the signal data



Test Data Drift



Test Data Drift



5 Models

Make a model for each waveform after the drifts are removed

Apply the the model that corresponds to the test waveform

Submit on Kaggle to get a macro F1 score

- Macro F1
  - Average of the F1 score for each class (a class is each number of ion channels open [0,10])
  - Example of f1 for class 9
    - False positive: guessing class 9 when it is not actually class 9
    - False negative: guessing not class 9 when it is actually class 9
  - Do this for all classes and take the unweighted average

Poor models

- Without breaking up the data and without removing the drift:
  - Dummy Model (guess majority class)
    - F1: 0.160
  - Random Forest
    - F1: 0.210
- After breaking up the data but leaving in the drifts
  - Dummy Model
    - F1: 0.278
  - Neural Network
    - F1: 0.630

Better Models

- Remove drift by using each waveforms rolling mean
  - Random Forest
    - F1: 0.86
  - Neural Network
    - F1: 0.88
- Remove Drift Using function (better)
  - Random Forest
    - F1: 0.92
  - Neural Network
    - F1: 0.92

Best Models so far

Sliding Window

- Each point is the signal 10 time points before to 10 time points after
- A neural network was trained using this sliding window data

$$\boxed{\mathbf{0.940}}$$

What Next?

Optimize the model more.

Think of new ways to feature engineer and to make models.

Try to get an F1 score above 0.940

Thank you!

https://github.com/Jarodc33

www.linkedin.com/in/jarod-carroll-762a361a8