Steve Su, Kristian Montoya

1. What is your project topic? (please give the title and a brief description)

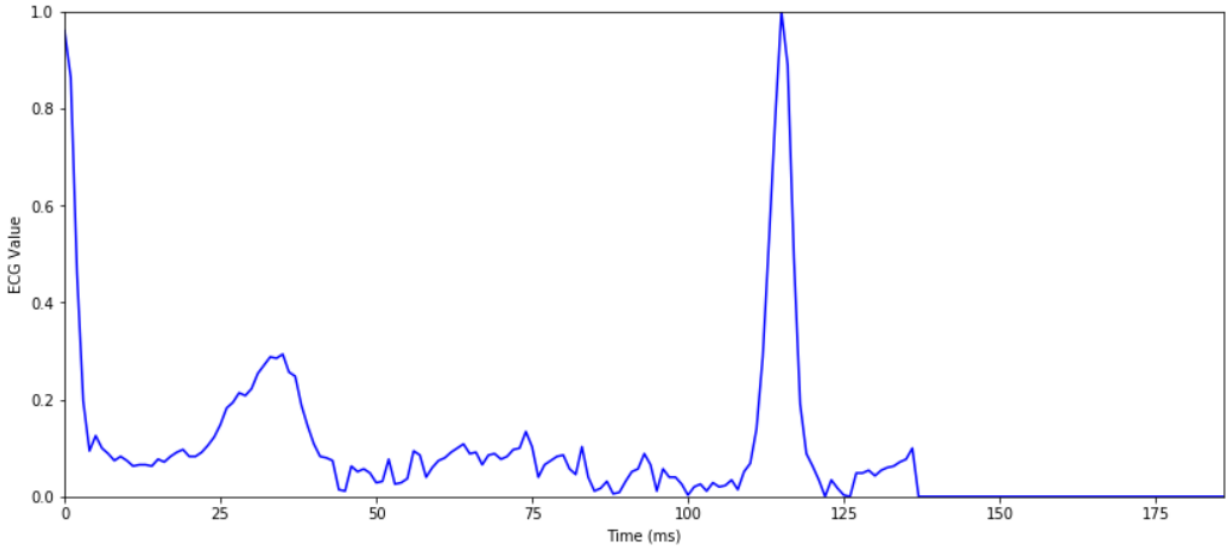
Our project topic covers the analysis of ECG's (electrocardiogram) of patient heartbeats to classify heart conditions. Each instance within our dataset represents an "image" of a patient's ECG over a 187 millisecond time interval where each millisecond is represented with a normalized data point that captures the ECG value at that given point in time. Our goal is to use/test several classification models to find which models provide the highest overall accuracy. Specifically, we want to test RNN vs. CNN (CNN seems to be the most popular model to use with this task/project) to see whether or not RNN would outperform a CNN model since RNN is applicable with timed datasets (such as ECG).

2. What are the data sources (links)? and briefly describe the data. (ex1) number of samples and features for tabular data (ex2) scraped (text/sound/image or other type) data from xyz website that has N training examples

Dataset Shape:

- 87,553 rows: Each instance is a single ECG
- 188 columns: Each column represents a millisecond. The last column is the ECG classification result (0-4)

As stated above, each instance without our dataset represents a single recorded ECG. The individual data points of each instance are normalized values between 0 & 1 that represent a single millisecond of said ECG.



	0	1	2	3	4	5	6	7	8	9	...	177	178	179	180	181	182	183	184	185	186
0	0.960114	0.863248	0.461538	0.196581	0.094017	0.125356	0.099715	0.088319	0.074074	0.082621	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Kaggle source: https://www.kaggle.com/shayanfazeli/heartbeat

Other source: https://arxiv.org/pdf/1805.00794.pdf

Steve Su, Kristian Montoya

3. Have you done some exploratory data analysis/visualization and started initial approach?

Initial approach can include:

1) search for research papers or internet articles or someone else's git to reference

2) you have some ideas and are on a way to implementing those

There are numerous notebooks posted on the original kaggle page that can be used for reference. The top rated notebooks seem to use CNN or SVM models to tackle the project and visualize their data through simple linegraphs (to visualize the ECG) and heatmaps (to present category correlation).

Our visualization (so far) has focused on comparing random sampled ECGs to the average ECG for matching categories. Once we make a confusion matrix for the data, we will implement some heatmaps as well.