**Ilya Shafir**
UCSD AI/ML Bootcamp - March 2024 Cohort

# 20.6.1 Capstone Project Proposal

**25th February 2025**

## US Election 2020 Twitter Sentiment Analysis

There are plenty of existing population polls trying to predict the outcome of a Presidential election. It would be of interest to perform ML sentiment analysis of a sample of past Twitter feeds (before it became X) to see how accurate it would be. Of course it begs the question if the sample population in Twitter has a diverse representation. On the other hand, the same can be said of the traditional polls and whether their samples are also diverse.

I plan to use the US Election 2020 Sample Dataset from Kaggle for this analysis. It consists of two CSV files totaling about 1.7 million rows, one with Tweets focused on Joe Biden, and the other on Donald Trump. I would take advantage of existing ML tools such as Hugging Face sentiment analysis models to estimate whether Tweets were overall positive, negative, or neutral for each candidate and compare the two to determine the winner.

The dataset needs to be culled of Tweets originating outside of the US. While that doesn't guarantee that non-US citizens in the US are Tweeting (or that citizens outside of the US are Tweeting) it helps narrow down any international opinions that would skew the results. The rest would be just data cleaning, for example, removing hashtags/mentions, URLs, and any extraneous text that is not generally readable and that may not convert to useful tokens.

As for the size of the data as provided, it may be too large for a desktop PC, and perhaps a bit too large to use Google Colab so I'll have to test what a reasonable smaller sample size would be to reduce resource and time usage.