*Members:* Derek Leung, Felix Yang, Ethan Lee, Chris He

*Data Sources:*

- https://www.kaggle.com/datasets/chrisfilo/onion-or-not (for titles of articles and classification of whether or not it's satire)
- Web Scraping of the article contents from the article title (could grab first url link from a Google Search API and use newspaper3k python package to scrape the content)

## Identifying Newspaper Writing Styles (and Satire)

The contents of various news articles these days can be concerning and deceptive and have different biases when it comes to different politically-sided sites. In addition, there are news articles written as purely satire for a fun read, yet when put side to side with a real newspaper article, it could be taken as real if one doesn't fully understand the background or content at hand. We plan on using various natural language processing techniques and various modeling of our content articles in order to see if we can determine the origin of our various articles and especially if we can determine if an article is satirical or real. This is important in looking towards articles and determining legitimacy of the article when its origin is unfamiliar to the user and could even potentially identify false information.

The above problem will be one of classification and seeing if the various features of some articles are meaningful to determine from where the article is from and whether or not it is satire. We can utilize features of vocabulary, article length, and other various tokenizations of words to implement into a certain model of classification. The models we can compare are k-nearest neighbors, decision trees, and support vector machines; the first two models can be used for both identifying origins/satirical articles with support vector machines being mainly used for determining whether or not an article is satire.

Following our implementation of these machine learning classification models, we will compare and contrast the success to which each was able to identify the origins and writing styles of articles. We will also investigate and discuss potential reasons why certain models ended up outperforming others.