# Project Proposal

*Birna Ósk Valtýsdóttir – Gylfi Þór Helgason – Helgi Tuan Helgason*

## General Information

Our proposal for this final project in Data Analysis is the Search Engine Manipulation Effect, or SEME.

Today, our everyday lives somehow involve algorithmic systems. We are all familiar with Google search, which personalizes our search results to try and surface more relevant content; Netflix or Amazon, which recommend media and products; and Facebook or Instagram which personalizes each user's news-feed to highlight engaging content.

Scientists and regulators are concerned that these algorithms are harming individuals and society, although there are many cases where algorithms are beneficial to users. Political scientists worry about democracy being threatened, and the filters used for personalization on the web might be increasing political polarization. For some years now, many scientists have been auditing search engines in general, and especially their role in political elections. They distinguish three different kinds of efforts to detect political bias in search platforms: Third-party manipulation, Ranking Bias and Ecosystem Bias.

For this project Ranking Bias will be our focus. We will represent data involving 5 experiments in 3 different studies. The data from this particular research has focused on measuring political bias of search engine algorithms to detect possible search engine manipulation effects on voters or unbalanced ideological representation in search results.

It is imperative for people to understand how these algorithms are being implemented and the data they use if we are going to fully understand the effect they have on people and their choices. Sadly, however, the subtleness of this manipulation technique is so effective that most people do not even have the slightest idea that they're opinions and beliefs are in fact being influenced by these algorithms. Using that as inspiration we found this to be a great subject for this Final Project. Our data and target sets will be undecided voters from both the United States and India. We want to clearly demonstrate how the search engine rankings can shift the voting preferences of undecided voters.

## Description of the data set and its collection

The group has found two datasets to work with for this project, both are in csv format. Our primary dataset comes from five double-blind, randomized controlled experiments, using a total of 4556 subjects of undecided voters, representing diverse demographic characteristics of the voting population of the United States and India. These experiments were performed at the American Institute for Behavioral Research and Technology by Robert Epstein and Ronald E. Robertson. Our secondary dataset contains constituency (state-level) returns for elections to the U.S presidency from 1976 to 2020 and is collected from the Harvard Dataverse, and presented by the MIT Election Data and Science Lab.

# Goal

Our goal for this project is to present this significant data as to how algorithmic systems affect our everyday lives and society, in particular we will demonstrate how Google Search can be biased when it comes to elections, and answer the following question; How much are Search Engines shifting voters' choices and their political preferences? Afterwards we will join the two datasets to compare the results of the 2020 and 2016 US election, and the meaning of SEME in this context.

## Plan for data analysis

Data analysis will be done using Python with numpy, pandas and matplotlib libraries and using statsmodels and scipy for stats. Our main focus will be set on the primary dataset for the data manipulation.
Preprocessing will start with cleaning our datasets, or munging them to be precise. The primary dataset is a bit messy to work with, e.g., it is One-hot encoded, and has column headers for values. We expect the cleaning process to be quite time consuming as it is a large and complex dataset.
We plan to implement some statistics plots for the data. A final decision has not been made on which types of plots we will use but most likely a histogram, scatter plot, and/or some other basic plots we will see fit and sufficiently descriptive.
Statistical tests for this project will most definitely revolve around our hypothesis and its truth value. The hypothesis in our case, is that the power and robustness of the Search Engine Manipulation Effect (SEME) and its biased search rankings, can (or already have) influence a big enough sample of a population to alter the outcome of an election. In accordance with our primary dataset (SEME) we tend to use the voting data for the U.S president election 2020 and 2016 to make an assumption of how it might have in fact turned out differently if search engines and biased algorithms were out of the picture.
As for the model evaluation, we plan to use the Split Sample Method to train our model to predict the chance of people voting differently due to SEME. We also think a Confusion Matrix is a good idea to better demonstrate our observations among other performance measures we see fitted for this domain.

## Expected outcome

The expected outcome is pretty straight forward as we intend our model to illustrate the voting patterns and behavioral aspects before and after subjects were or were not manipulated by the search engines algorithm.

At best we wish that these results and our representation of the matter will bring more awareness to the public of this issue (yes, we dream big). Stating the fact that all search engine- and social media companies (the wealthiest companies in the world) are currently unregulated, the results for this research should be a cause for concern.