The Social Determinants of COVID Attitudes

Joe Dean - Degenerate Hessians

November 18, 2020

Introduction

Following the news on COVID-19 often creates anxiety, not just because of the virus itself but often because we know we could be doing better. The human element of fighting the virus cannot be ignored. Rather than simply raising the volume on our speech, though, how can we look to data to target our approach and tailor it to the realities of what people think and believe? While we are modest about our own contributions, we believe this project can help refine communication around the virus by highlighting what people really believe.

Problem Description

What percentage of people hold an accurate scientific view of COVID? To begin to answer such a question, we need to define what a "scientific view" even means. Armed with that knowledge, we can then make targeted approaches to educate others about what really works against COVID. Alternatively, while it may be the case that we need to do a better job educating some, it may also be the case that different types of non-scientific messaging could be more effective in marshalling others in the fight against COVID. In any case, we need to know who to send what message.

Broad Approach

Fortunately, c3.ai's COVID data lake provided rich resources to dig into beliefs about the virus. Our key idea was to combine the survey and location data with data from the Robert Wood Johnson foundation (RWJF) and to create belief profiles. The RWJF county health rankings data is rich with features on health outcomes as well as social and environment factors that affect health.

Technical Details of Approach

Matching the data

To match the survey data to the RWJF data we had to use the population data from the location API in the c3 datalake and Housing and Urban Development ZIP-to-FIPS crosswalks. Because survey data was only mapped to the zip-3 level, we mapped it to the highest population county. This fuzzy matching was not perfect, and the 82% match rate we achieved ultimately meant some data dropping out like Puerto Rico.

Belief Profiles

To summarize beliefs about the virus, we took the 12 TRUE/FALSE questions from the survey and looked for subgroups of common answers by using cluster analysis. Since the data consists of binary responses, popular algorithms such as k-means which rely on Euclidean distance were not appropriate. Instead we implemented a Bernoulli mixture model in C++ and R which models the answers as realizations of a set of Bernoulli distributions and converges on the parameters via expectation-maximization.

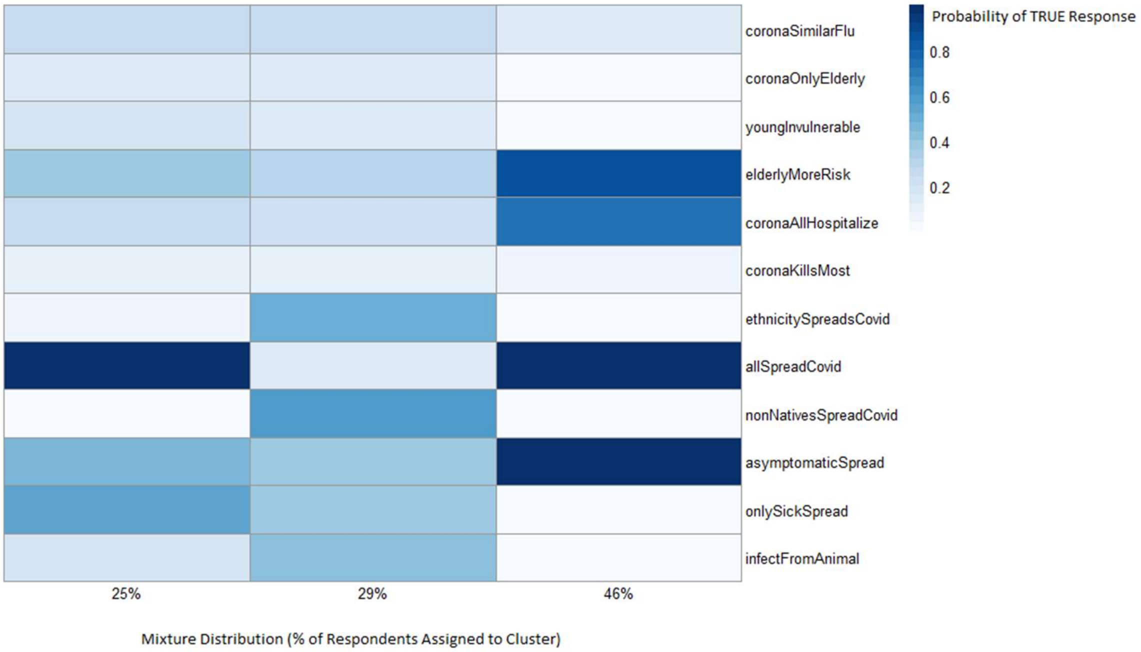
The resulting cluster profiles are illustrated below. Each column represents a cluster and each row a survey response. Darker cells indicate a higher probability of a TRUE response for that question. After some experimentation we found three clusters best explained the data:

1. **Skeptical** -- We see their profile on the left. This is roughly 25% of respondents. They believe all can spread COVID but don't acknowledge asymptomatic spread or the risk to elderly.

2. **Suspicious** -- This group, representing about 29% of respondents, largely believes animals, non-natives, and ethnicity are primary factors in the spead of the virus.

3. **Scientific** -- This third cluster seen on the right (46% of respondents) represents scientific beliefs about the virus and its properties. Asymptomatic spread is acknowledged as is the risk of all for being hospitalized.

Figure 1: Belief Profiles from Bernoulli Mixture



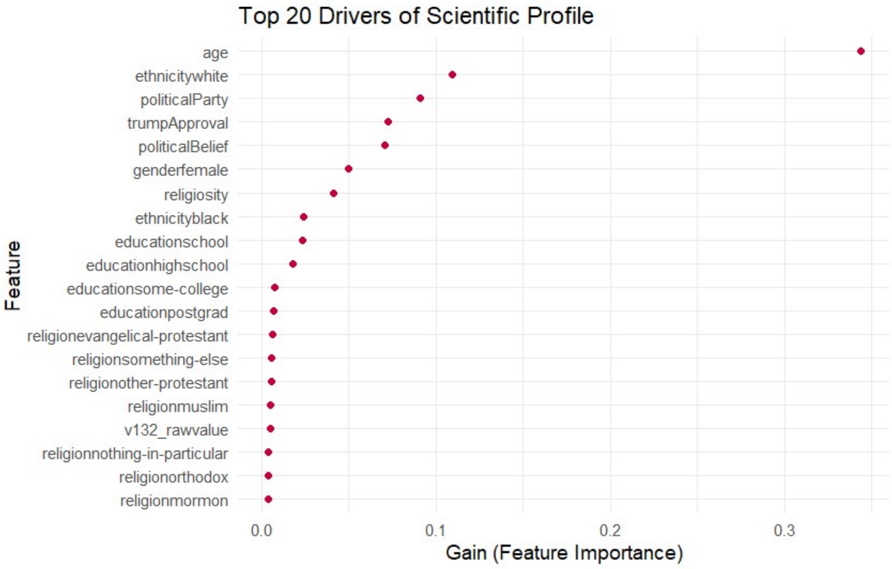
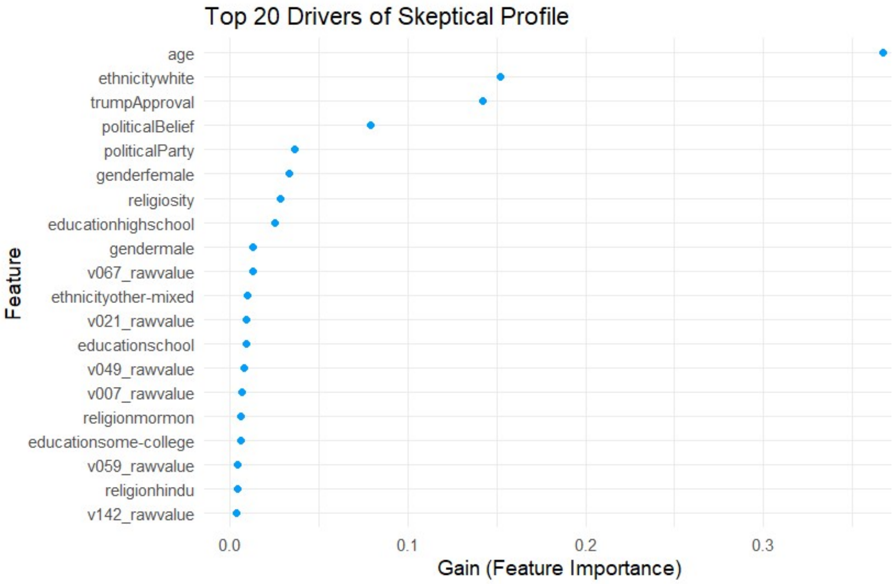
Feature Creation and Driver Analysis

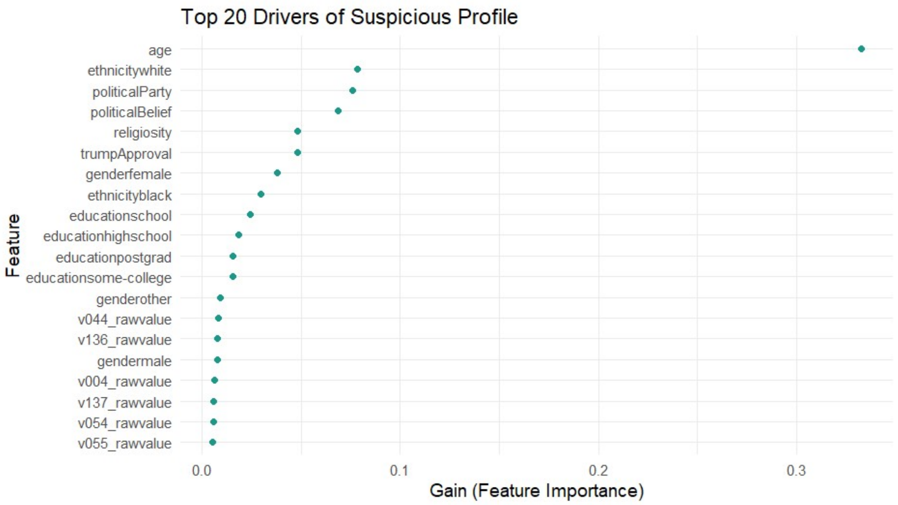
We tried standardizing the RWJF data, imputing means to the missing values, and performing a principal components analysis to create new uncorrelated features. We used these features mainly for exploratory data analysis.

We used the xgboost implementation of gradient boosting in R to perform the driver analysis. We chose this method for a number of reasons: 1) it easily finds variable interactions; 2) it allows for non-linear relationships of continuous values; 3) it converges quickly on large datasets; 4) it is robust to correlated features. We used 10-fold cross-validation and measured the AUC of the test sets. The main idea was to start with base demographic features and see if adding new ones raised the test AUC. This would indicate whether the variables added additional predictive value. We focused on finding drivers of the scientific belief set and the suspicious belief set, as these are the extremes.

The table below shows variable importance results for predicting membership to each profile:

Table 1: Predictive Drivers of Beliefs

Cluster	**AUC (10-fold CV)**	**Variable Importance (Both positively and negatively correlated variabels)**
Scientific	.80	 *v132 = Access to exercise opportunities
Skeptical	.645	 *v067 = Driving alone to work value

Suspicious	.726	

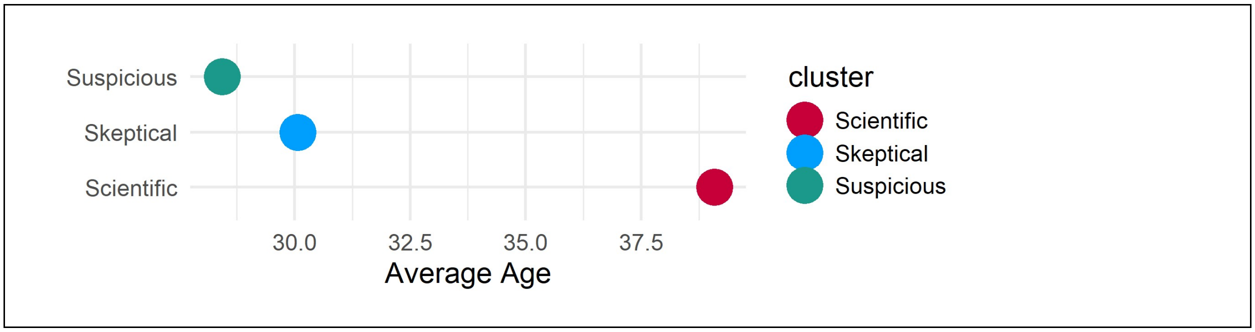
Results

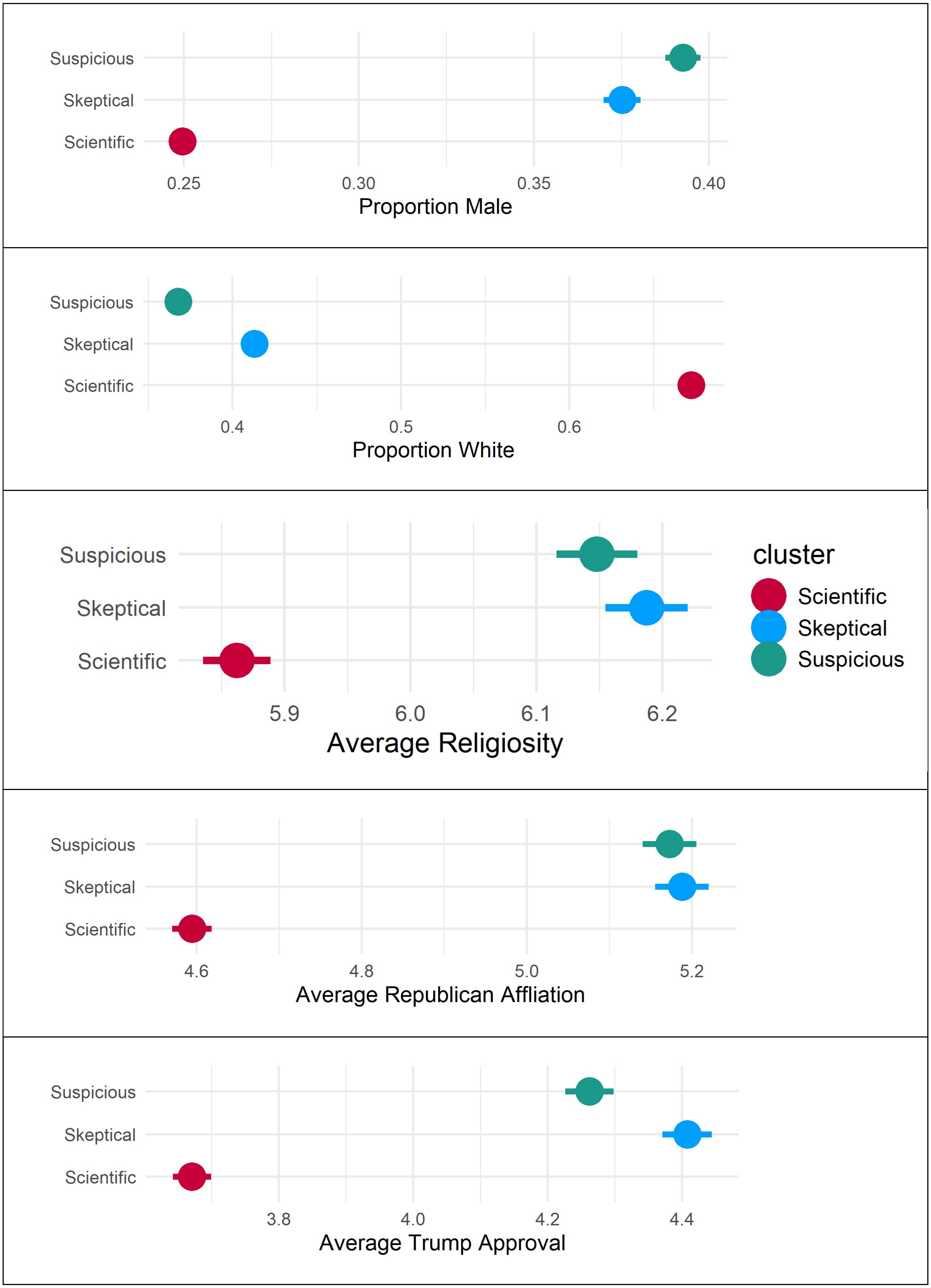
We found demographic features to be the most predictive by far--with age dominating. Gender and ethnicity also play a big role. While we had hoped many other health outcome data points would feature, the data simply did not support this hypothesis after accounting for demographics. However, the predictive power was good for the scientific and suspicious clusters based on demographic alone (.80 and .73 AUC respectively).

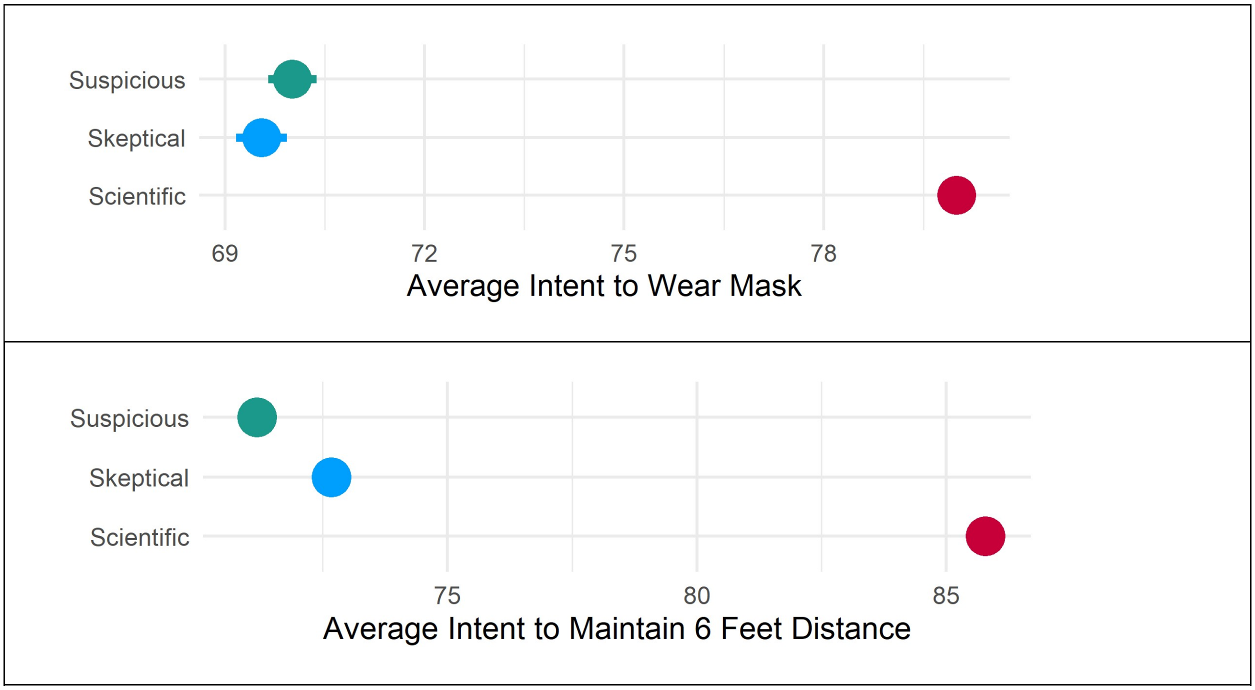
We also see the non-scientific clusters split on demographics and politics. This suggests that there is not one homogeneous group of incorrect beliefs but separate profiles that need to be addressed in different ways.

In the plots below, we show the average of the most salient features by cluster. These features were not used to create the clusters—rather they tell us who they are and what their behavioral intentions are with taking precautions against COVID-19.

Figure 2: Demographics and Intentions by Belief Group




Suspicious
Skeptical
Scientific
0.25
0.30
0.35
0.40
Proportion Male
Suspicious
Skeptical
Scientific
0.4
0.5
0.6
Proportion White
Suspicious
Skeptical
Scientific
5.9
6.0
6.1
6.2
Average Religiosity
cluster
Scientific
Skeptical
Suspicious
Suspicious
Skeptical
Scientific
4.6
4.8
5.0
5.2
Average Republican Affliation
Suspicious
Skeptical
Scientific
3.8
4.0
4.2
4.4
Average Trump Approval



Impact

We show that less than half the population holds a scientific view of COVID. While health outcomes data did not provide extra value, we demonstrate that demographic features provide most of what is needed for predictive accuracy. Technical innovations include use of a Bernoulli mixture model for analyzing TRUE/FALSE survey responses and defining the scientific/skeptical/suspicious belief profiles.

We believe with updated survey data even more can be done to create targeted communications around COVID-19.