# Literature Review

To place this research in context, a brief overview of the current state of the literature follows. First discussed are public perceptions of the acceptable uses of online data and the potential ways these data could be exploited. Next, a theory-driven social science perspective on why digital data may be predictive of political party preference is defined, followed by a brief review of prior studies that have successfully used such data to achieve these means. The benefits of search engine data in particular are then discussed, as well as the specific contribution of this research, particularly regarding the individual-level nature of the data.

## Digital Data: Public Perceptions and Implications for Democracy

Beginning with Obama’s first presidential campaign, data-driven microtargeting has been a major theme for many major political campaigns, featuring prominently in the 2016 Trump and Clinton campaigns, for example (Kruschinski, 2017). Digital data can be used to develop profiles of individuals in an effort to curate and deliver tailor-made messaging that is as efficient as possible for eliciting a desired behavioral response. The Cambridge Analytica scandal brought the issue to the public’s attention, with critics warning of the potential to manipulate voters and erode privacy, and supporters pointing to the potential benefits microtargeting has for mobilizing specific target groups, or those who may be naturally less inclined to vote (Kruschinski, 2017).

Ur and colleagues investigated non-technical users’ attitudes towards Online Behavioral Advertising (OBA), which relies on users’ browsing history to deliver custom ads (both political or commercial). They found that participants felt OBA was both useful and simultaneously “creepy,” expressing concerns about privacy. While participants were aware that contextual targeting was taking place, many were surprised to know that not only could online behavior theoretically be used to tailor advertising, but that this is already common practice. Concerns were particularly pronounced when participants felt that the profiles generated to target them were inaccurate: For example, if they felt stereotyped or that they were receiving ads that weren’t representative of their true interests (Ur, Leon, Cranor, Shay, & Wang, 2012; Dolin et al., 2018).

Several studies have found that individuals lack a fundamental understanding of how such targeting takes place, and that users are even less comfortable with targeted advertising once they gain a fuller understanding of how the data are gathered (Dolin et al., 2018; Ur et al., 2012).

While in theory such data-driven targeting could be a boon to grassroots campaigners and allow organizers to bring more individuals into the political sphere, critics have expressed concern about the ability for the political elite to exert undue influence on the political process. Gurumuthy and Bharthur warn against a future where the unethical use of Big Data and AI “allows political influence to move from public campaigns to private sentiment, a shift that repositions electoral politics from a spectacle that is overt to a script that is covert,” where voter behavior is manipulated such that outcomes no longer reflect informed decision-making or the democratic will (Gurumurthy & Bharthur, 2018).

The opportunity to exploit these new micro-targeting capabilities also exists for foreign powers, not just the elite within a particular country. Indeed, the 2016 U.S. Presidential election saw probable evidence of Russian interference that relied on data from social media to target particularly-relevant constituencies in an effort to bolster the Trump campaign (Jamieson, 2018).

In this context where the availability of digital data is growing at an unprecedented rate, machine learning is becoming an ever-more powerful tool, and the public lacks a general awareness about and comfort with the modern targeting applications, it is critical to examine the extent to which digital data, such as search engine queries, can actually be used to efficiently target individuals to influence political processes.

## Theoretical Basis for Search Engine Query Data as a Predictor of Political Preferences

Prior studies relying on digital data to predict political preferences have been criticized for lacking a solid theoretical basis to qualitatively explain their results. This can lead to poor reproducibility, and increases the chances of incorrectly interpreting results they may be simply due to chance (Lui, Metaxas, & Mustafaraj, 2011; Yasseri, 2016). This paper argues that differences in browsing behavior and linguistic choices are meaningfully related to differences in demographics, which also share strong associations with political preference.

Browsing behavior has been shown to vary on the basis of demographic characteristics, which are in themselves associated with differences in political affiliation. For example, Hu and colleagues describe how women are more likely to seek medical or religious information online than men are (Hu, Zeng, Li, Niu, & Chen, 2007). In turn, gender is also correlated with party identification, with women in the United States being more likely to favor the Democrats (Pew Research Center, 2016). Similar relationships hold true for other characteristics such as age and level of education, which have also been shown to vary with ideology (Hu et al., 2007; Weber & Castillo, 2010).

Weber & Castillo showed similar findings for web search behavior in particular, based on factors such as the length of queries and the web pages visited after a search. They determined that “demographic factors have a measurable influence on search behavior:” For example, queries beginning with the first name “Hal” in low-education areas typically ended the search with the last name “Lindsey,” in contrast to those in higher-education areas where “Higdon” was the more common last name (Weber & Castillo, 2010). Query language has also shown to be able to predict age and gender (Jones & Tomkins, 2007).

Differences in query language varying systematically based on demographic characteristics is not surprising in the larger context of linguistic variety. Indeed, several studies have found that blogger age and gender are inferable on the basis of linguistic choices such as length of a post and the words contained, punctuation, capitalization, and general prose style (Burger & Henderson, 2006; Nowson & Oberlander, 2005). Formal written texts have also been found to vary in a meaningful way on the basis of age and gender (Argamon, Koppel, Fine, Shimoni, & Science, 2003; Koppel, Argamon, & Gan, 2000).

Even smaller strings of text in the form of Tweets have been able to predict demographics and political preference. For example, Democrats and Republicans “tend to use a specific vernacular (‘obamacare’) when discussing issues of interest to both sides (healthcare reform)” (Pennacchiotti & Popescu, 2010). Rao and colleagues also work on Twitter data, but emphasize the importance of sociolinguistic cues. For example, character repetition (e.g., “that’s soooo crazy”), is often indicative of a female writer, as are the use of emoticons or multiple exclamation points (“!!!”). Like Pennacchiotti and Popescu, Rao et al. note particular vocabulary as being particularly illustrative, with certain terms like “dude” or “bro” being strongly associated with younger writers (Rao, Yarowsky, Shreevats, & Gupta, 2009).

Thus, web browsing behavior generally, as well as linguistic decisions even in short text (such as search queries), have been shown to be able to illustrate differences in demographics, which are also clearly associated with differences in political preferences (Pew Research Center, 2016).

In addition to how people search and behave online, the simple condition of whether or not they make a politically-oriented query or the volume of such queries can be a meaningful for explaining the importance of search queries from a theoretical perspective. Yasseri and Bright elaborate:

*“We base this theory on a rational choice approach to explaining voting behavior, which conceptualizes voters as similar to consumers in a market, seeking to vote for the political party who offers them the greatest “pay-off” in terms of policies​. Online information seeking, from this rational choice perspective, can be explained in terms of voters looking for more information about the election: perhaps about practical matters such as how to vote, or perhaps about substantive matters such as which political party might suit them best. Such information seeking is rational in that it increases the chance that the voter will vote for the party which represents them best.”* (Yasseri, 2016)

## Prior Uses of Digital Data for Prediction of Political Preferences

The possibility for harnessing the predictive power of online behavior data is not particularly new, with prior studies utilizing information on Facebook and Twitter posts (including the volume of posts and the sentiment of the related text), web browsing data, and aggregate level Google Trends search query information, with mixed success.

Early work by Tumasjan, Sprenger, Sandner, and Welpe claimed that Tweet volume could be used as an alternative to traditional polling, and that the sentiment of politician’s and parties’ Twitter messages “closely corresponds to political programs, candidate profiles, and evidence from the media coverage of the campaign trail” (Tumasjan, Sprenger, Sandner, & Welpe, 2010). Tweet sentiment analysis has also been found to correlated to presidential job approval polls (O’Connor, Balasubramanyan, Routledge, & Smith, 2010).

However, these works and others have faced criticism for their lack of reproducibility and disregard for sample representativeness. For example, Chung and Mustafaraj applied the same methods employed by Tumasjan et al. (2010) and O’Connor et al (2010) to a new dataset and found it was unable to result in an accurate prediction on a new sample (Chung & Mustafaraj, 2010). Similarly, Schoen, Jungherr, and Ju showed that even using the same case as the Tumasjan et al (2010) study, vastly different results were achieved through the inclusion of a different set of parties or timeframe, both of which appeared to be arbitrary choices in the original paper, thus also placing doubt on the ability of this method to generalize to future elections (Schoen, Jungherr, & Ju, 2012).

The inability for Twitter to consistently predict election outcomes or political preferences is unsurprising given that social media users differ in meaningful ways from the electorate at large. Indeed, even the prevalence of bots and spam accounts should make one question the reliability of such a sample (Gayo-avello, 2011). Additionally, there is likely to be a significant influence of self-selection bias, as those who are active on social media are likely to be the most politically-oriented and perhaps ideologically extreme (Gayo-avello, 2012).

Gayo-avello (2012) outlines several key ways to overcome the challenges perceived in prior studies utilizing social media data. Namely, a credible baseline should be established (as discussed in section XX), the study timeframe should be clearly specified and justified, only data from eligible voters should be included, state of the art sentiment analysis should be employed over simplistic methods, spam should be removed, and bias in the data should be analyzed and acknowledged (Gayo-avello, 2012).

Another avenue of research focused on the predictive power of online browsing history as a whole, thus overcoming several of these issues, in particular due to the fact that internet users generally are more representative of the electorate than users of a particular social media platform.

Comarela, Barford, Christenson, and Crovella (2018) found that web browsing history was able to predict candidate preference at rates in line with modern polling techniques. They focus on a state-by-state and day-by-day analysis, comparing the web browsing data of 100,000 individuals over a 56 day period shortly before the 2016 US Presidential election to statewide polling data, thereby overcoming the common challenge of missing individual-level “ground-truth” labels. Their results showed that domain-level URL visit history was able to predict election results with a comparable accuracy to polling (with linear correlation of 0.94), and that the fine-tooth nature of the method allows for the analysis of the impact of a specific event, such as the release of the “Comey letter” on a day-by-day and state-by-state basis (Comarela, Barford, Christenson, & Crovella, 2018).

Bach and colleagues conducted a similar study, though with the advantage of having survey data on political preference to augment their corpus of browsing data for 2,000 German adults eligible to vote in the 2017 federal election, for four months before and after the vote. However, they found that online browsing behavior was not a strong predictor of self-reported voting behavior in their sample. In particular, their model struggled to identify undecided voters, though performed better for parties at the political periphery, such as the Greens and AfD (Bach et al., 2019).

This addition of individual ground-truth labels is quite rare in most prior studies utilizing digital data for predicting political preferences. Indeed, the few studies that have prioritized predicting individual-level characteristics have primarily relied on differences in linguistic features (Pennacchiotti & Popescu, 2010; Rao et al., 2009), though Hu and colleagues also took an individual focus when analyzing differences in web browsing behavior (Hu et al., 2007).

At the present time no studies that have taken an individual-level approach to using search queries as a predictor of political preferences are known. However, collective applications of search query data both for illustrating current events (“predicting the present” – as coined by Choi & Varian, 2011) as well as forecasting future outcomes. Google Trends data has been successfully used to forecast topics such as unemployment (D’Amuri & Marcucci, 2017), housing prices (Wu & Brynjolfsson, 2015), consumer purchasing behavior (Goel, Hofman, Lahaie, Pennock, & Watts, 2010), and the spread of influenza (Ginsberg et al., 2009).

Google Trends data have also successfully been applied to elections. Stephens-davidowitz showed that search volume for the terms “vote” or “voting” in a particular geographic area was strongly correlated with the electoral turnout in the region in the 2008, 2010, and 2012 US elections (Stephens-davidowitz, 2013). Polykalas, Prezerakos, and Konidaris (2013) applied a similar method to three German elections, also relying on a pre-defined list of keywords that were determined to be relevant for electoral outcomes and measuring their relationship to election results. The algorithm was able to accurately predict the election outcome of all three of the studied elections (Polykalas, Prezerakos, & Konidaris, 2013).

Lui, Metaxas, and Mustafaraj (2011) cast doubt on the applicability of Google Trends data to forecasting elections, however. They argue that such data were not successful at predicting the 2008 and 2010 US elections compared to incumbency, polls, or even chance They point out that this could be due to limitations on simply using search volume for a particular candidate’s name, since this does not adequately illustrate *why* an individual may be searching for a candidate. For example, if a candidate is particularly well-known, they may not be searched for at all, which is actually a good sign for their election prospects. The researchers therefore recommend employing sentiment analysis to get an understanding for the driving forces behind a user’s query (Lui et al., 2011).

## Strengths of Search Query Data

Much of the interest in utilizing digital data to forecast political preferences is to overcome the drawbacks inherent in traditional polling. For example, polls require significant time and monetary resources, and hence “cannot give insight into the short-term dynamics of vote choice, especially on a per-state level.” Results can also be sullied through interviewer effects, word choice, question order, or even reticent respondents (Comarela et al., 2018).

Additionally, self-reported vote forecasts have been shown to often be misleading. Rogers and Aida (2014) examined seven pre-election surveys with post-election vote validation and discovered that many predicted voters do not vote after all, and many who say they won’t vote actually do. Additionally, self-predicted voters differ significantly from actual voters, though there is little difference between self-predicted voters and non-voters, thereby showing that, “Vote self-prediction is “biased” in that it misleadingly suggests that there is no participatory bias” (Rogers & Aida, 2014).

Another concern with polling is that participants may be untruthful about their voting intentions when they hold views they believe to be socially undesirable, such as racial animus, or even one’s intention to vote for a polarizing candidate like Donald Trump (Brownback & Novotny, 2016). There is some evidence to suggest that search query data may be able to combat this issue. For example, Google searches are unlikely to exhibit major social censoring, because users are typically acting alone, and online (Stephens-davidowitz, 2012). Additionally, in a study on user perceptions of web-based information disclosure, participants expressed that they are typically honest when conducting web searches (Conti, Point, York, & Sobiesk, 2007).

Therefore, search query data – if it is able to successfully predict political preferences – may be more accurate than self-reported voting intention, both because it avoids common polling issues, as well as the impact of social desirability bias.

## Contributions of this Research

This research builds on prior studies that have relied on digital data to predict political preferences in several meaningful ways:

1. The unique nature of the dataset means that individual-level search query history is able to be compared to ground-truth, individual-level survey data on political affiliation and voting history. This also ensures that bots are not polluting the data, and the demographics of the sample can be confirmed and compared to those of the electorate at large.
2. While much work has already been done using Google Trends, prior studies have relied exclusively on pre-defined lists of terms, and measured their search volume. However, there is a clear opportunity available in utilizing the complete search history, since perhaps there are specific words, topics, or themes that are unexpectedly predictive of political preferences.
3. Similarly, sentiment analysis will be implemented to augment the keyword-only approach taken by previous researchers. This will help to create a fuller understanding of *why* someone is searching for a particular term (Lui et al., 2011).
4. No recent prior studies have considered contemporary search engines other than Google, despite the fact that users of Bing.com or DuckDuckGo.com may differ in meaningful ways from Google users.

# Dataset

## Survey and Search Query Data

The data for this research come from the YouGov Pulse panel survey *Paying Attention to Attention: Media Exposure and Opinion Formation in an Age of Information Overload,* running from April 2018 through November 2018 over five waves. Each wave was comprised of a nationally-representative sample of US adults. The original sample was **XX**, and once filtered for those with adequate search history data and a response on the outcome measures, the analytic sample is **XX** individuals. Survey questions included a variety of demographics, as well as voting intention before the 2018 midterm elections, and whether and for whom the participant voted for after the election.

In addition to the panel survey data, web tracking data (which includes all search engine queries) was passively collected through YouGov Pulse. Participants consented to installing Reality Mine, software which tracks web browsing history in real time – with the exception of sensitive items such as passwords and financial transactions.

**Add something about the representativeness of the panel, note about missing search queries and differences**

## Target Variable

The goal of this research is to determine if it is possible to predict 1) whether an individual will vote in an upcoming election and 2) if so, for which party for the House of Representatives election (Democrats, Republicans, or Independent).

For Question 1 regarding turnout, the following question was re-coded as binary (with positive responses to option 5 “I definitely voted in the midterm election on November 6” coded as voted, and all others as did not vote), and all NA answers were removed:

Which of the following statements best describes you?

1. I did not vote in the election this November
2. I thought about voting this time, but didn’t
3. I usually vote, but didn’t this time
4. I attempted to vote but did not or could not
5. I definitely voted in the midterm election on November 6

Those who responded that they voted were further asked:

For whom did you vote for the U.S. House of Representatives?

1. The Republican candidate in my congressional district
2. The Democratic candidate in my congressional district
3. The Independent candidate in my congressional district
4. I did not cast a vote for the U.S. House

If the respondent answered with option 4 “I did not cast a vote for the U.S. House” or the response was coded NA, these responses were removed from the dataset. Thus, the remaining three options for Question 2 were that the respondent voted for the Republican, Democrat, or Independent.