# Literature Review

To place this research in context, a brief overview of the current state of the literature follows. First, public perceptions of the acceptable uses of online data, and the potential ways these data can be exploited are discussed. Next, in an effort to outline a theory-driven social science perspective on why digital data may be predictive of political party preference is defined, as well as 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 subsequent candidates, featuring prominently in the 2016 Trump and Clinton campaigns (Kruschinski, 2017). Digital data can be used to develop profiles of individuals in an effort to develop 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, while supporters point 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. They found that participants felt OBA is both useful as well as “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 less comfortable with targeted advertising once they gain a fuller understanding of how the data were 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 just due to chance (Lui & Metaxas, 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. 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, Zeng, Li, Niu, & Chen, 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 differences. 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 & Popescu, Rao et al. also 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.

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. For example, Yasseri and Bright explain:

*“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)