# Stor 390 Midterm

Eric Rash

02/27/24

## Concern:

The article opens with the sentence, "A major concern arising from ubiquitous tracking of individuals' online activity is that algorithms may be trained to predict personal sensitive information, even for users who do not wish to reveal such information." (citation). From this sentence alone, we can tell that the major concern depicted throughout the paper is going to be data privacy. In the case of this paper the concern arises out of the availability of digital trace data. The issue outlined is that someone may be able to use an algorithm combined with these digital trace data to predict something about someone. In the case of their study, they use data to attempt to predict how someone would vote in an upcoming election. Data privacy is concerning because people do not know how their data is going to be used, nor what it can even be used for. Many apps and websites require you to read their terms and services, but it is well known that nobody every really does. So people are willingly selling their data (more accurately, allowing their data to be sold) to companies who are actively trying to profit off of them at the best. The real intentions of whoever gets access to someone's data cannot be known.

## Methods:

The researchers started by having respondents install add-on to their computers and phones that would report back the complete URL, the domain of the website, the device they used to access the website, and the amount of time spent on the websites. Respondents were able to turn these off at any time, however, the researchers suspect based on the responses that they did not. Data was obtained from July to October. They create 4 different boolean variables based on survey results: *Undecided* - Did they respond that they were undecided in surveys? *Voted* - Did they respond that they voted in surveys? *AFD* - Did they respond that they voted for this party? *Greens* - Did they respond that they voted for this party? They chose the most polarized parties because they believe having more polarization between the two parties will make it easier to classify an individual based on their digital trace data. They have three different blocks of data with different predictive ability. The first includes information about the device and when it is used. The second includes the duration and frequency with which respondents access the 50 most used news domains. The third block of predictors features websites/apps that were used more than 80 times for over a minute. When creating the model they decided on using XGBoost because it is able to filter through predictors to select only the best variables in the model building process. These create various different decision trees, then use the issues from the previous tree to create a new tree that should be an improved model.

## Reference

Bach, L. R., et al. (2019, October 22). Predicting Voting Behavior Using Digital Trace Data. Sagepub. https://journals.sagepub.com/doi/full/10.1177/0894439319882896#table2-0894439319882896