

# Web Scraping Workshop

Connected\_Politics Lab

Sarah King

# Introduction

- ▶ Web scraping is a technique for efficiently collecting and organizing information from websites. Although these data can be collected manually, automation saves time and is less error-prone.
- ▶ In this workshop, participants will use R to automate the web scraping process. Participants will learn about the general structure of a typical web page and how to use the `rvest` package to select elements, such as text fields and tables, and iteratively extract relevant data.
- ▶ All of the materials for the workshop (slides & R Script) can be found in this GitHub repository (<https://github.com/sarahashleyking/ConnectedPolitics-Scraping-Workshop.git>)

# Outline of Content

- ▶ Introduction of necessary packages
- ▶ Important functions
- ▶ FOR loop
- ▶ HTML: The front-end syntax
- ▶ Selector Gadget
- ▶ Scraping multiple pages/tables/data that is not on the specified page
  - ▶ Brief QTA example
- ▶ Caveats/Conclusion

# Packages

- ▶ tidyverse - The tidyverse is an opinionated collection of R packages designed for data science. Necessary for data cleaning/wrangling.
  - ▶ rvest - (a part of the tidyverse) necessary for the actual web-scraping/crawling

```
library(rvest)
library(tidyverse)
```

# Caveats

- ▶ DDoS attacks.
  - ▶ A distributed denial-of-service (DDoS) attack is a malicious attempt to disrupt the normal traffic of a targeted server, service or network by overwhelming the target or its surrounding infrastructure with a flood of Internet traffic.
  - ▶ `Sys.sleep()`
- ▶ Robots.txt
  - ▶ Washington Post
  - ▶ Twitter
  - ▶ TripAdvisor
  - ▶ `rvest` in concert with `polite`. The `polite` package ensures that you're respecting the robots.txt and not hammering the site with too many requests.