Interacting with Web Data using R

Web APIs and Web Scraping (including RSelenium)

Spencer Lourens, Ph.D.

3/11/2019

Outline

- Why web data?
- Web APIs
  - Example, code run through
- Scraping static web content
  - css selectors, xpath, demos
  - rvest demo and example
- Scraping dynamic web content
  - Selenium web driver, and RSelenium (Selenium ported into R)
  - Example scraping dynamic content, iteration/scaling up

Why web data?

- In many applications, we may not have access to the data we need for an application of interest
  - Environmental concentration changes over time and space and their effect on health outcomes
  - Demographic data for adjustment and comparison across counties/states
  - Historical data regarding health behaviors and outcomes in a specific time/place
- In some settings, there are previous studies that have the data available and ready for download, but this isn't always the case
- We might see the data online in some spot, but perhaps need to gather many data from many slightly different locations

Why web "scraping"?

- Seems to require a LOT of copy and paste, labor (or time) intensive keystrokes/manipulations
- How long is this going to TAKE!?!?
- Can we somehow automate the process (i.e. automate the boring stuff), so that after we build our software, the data extraction/transformation process becomes "plug and play"?
- This is, in my opinion, the pinacle of web scraping, and a really really fun process to engage in
- This is also the topic of this webinar, extracting and transforming web data, either using an API provided by the data curators, or by scraping the data using "web scraping" techniques
  - **rvest** and **RSelenium** are used for this
- This webinar will be as interactive as possible - I'll do a lot of live demos, so cross your fingers!

Web APIs

- There are thousands of web APIs available over the internet for accessing data for use in our own applications
- API - application programming interface - so basically an interace that allows YOU to program your own application utilizing external resources (data)
- The R programming language comes equipped with a package called **httr** (developed by Hadley Wickham) which allows executing cURL requests (mostly GET and POST requests, and in our case just GET requests) from within R
- Be careful, you don't NEED to recreate the wheel!
  - Someone else might have already developed an R API for the website you're interested in
  - Google search for "r API"

First Example - healthdata.gov API

- asdfsa