# Web Scraping

Dr. Garrett Dancik

# Overview

- Web scraping is the process of retrieving web pages and extracting relevant data from them
- Why?
  - Search engines collect data to index web pages
  - Collecting weather and climate data for research
    - https://www.sciencedirect.com/science/article/pii/S0168169909002348
  - For businesses and consumers to keep track of products
  - For research in economics
    - https://www.aeaweb.org/articles?id=10.1257/jep.30.2.151

# Legal / ethical considerations and guidelines

- Respect each websites terms of service
- Respect each site's *robots.txt* file
  - A web site's robots.txt file lets robots (crawlers and web scrapers) know what behavior is permissible and what is not
  - https://www.promptcloud.com/blog/how-to-read-and-respect-robots-file
  - From eastern: http://www.easternct.edu/robots.txt
  - From Travelocity: https://www.travelocity.com/robots.txt
- Use a web site's Application Programming Interface (API) if available.
- Do not overload the host's server by pausing/sleeping if making multiple requests
- Identify yourself in the HTTP request header

# Steps for web scraping

- Identify a page you want to scrape
- Understand the structure of the page (e.g., by using the Web Inspector)
- Write a script that
    - Retrieves the web page
        - From a URL using the Python requests library
        - From a file using *open*
    - Extracts information from the web page
        - Using Python's BeautifulSoup library