



Collecting digital trace data

Shuning Lu (Ph.D.)
Assistant Professor
Department of Communication
North Dakota State University



Today's goals

Morning:

- Overview of digital trace data
- Data collection tutorial

Afternoon:

- Group activity & Report



What is digital trace data?

Social media

Geo-spatial data

News websites

Digital archives

Discussion: other kinds of data?



Strengths and weaknesses of digital trace data

Strengths: Always-on, unobtrusive, capture relationships

Weaknesses: proprietary, non-representativeness, drift, algorithmic confounding, unstructured, other bias

How to leverage the strengths?

How to alleviate the weakness?



Ways of collecting digital trace data

Downloading existing dataset

Screen/web scraping

API



Downloading existing dataset

<https://datasetsearch.research.google.com>

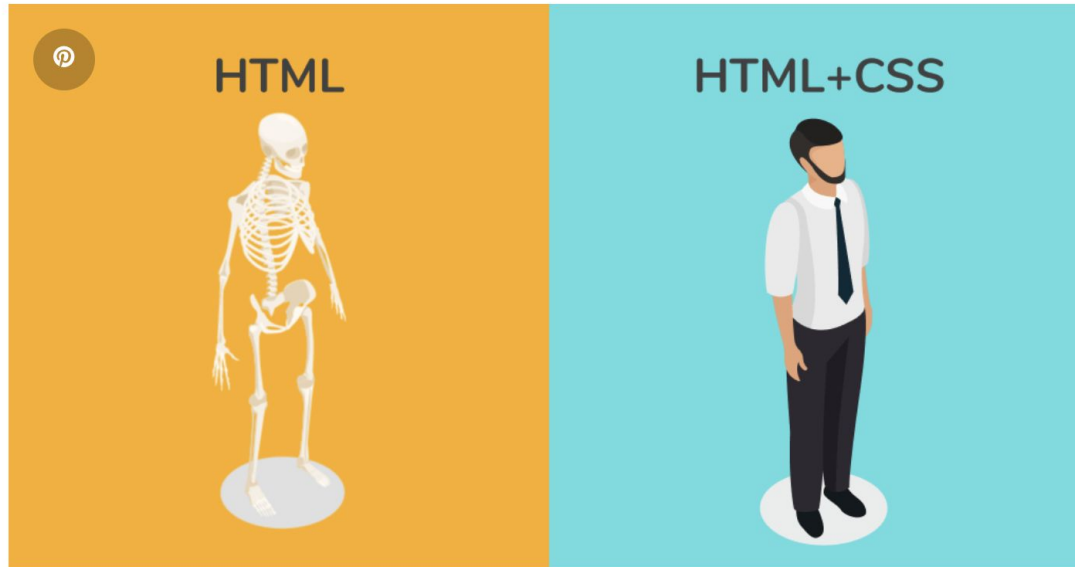
Macro-level data: World Bank, OECD, government, etc.

Traditional micro-level data: GSS, Pew Research, etc.

Public accessible platform data: Ad data, legal request data, special topics, etc.

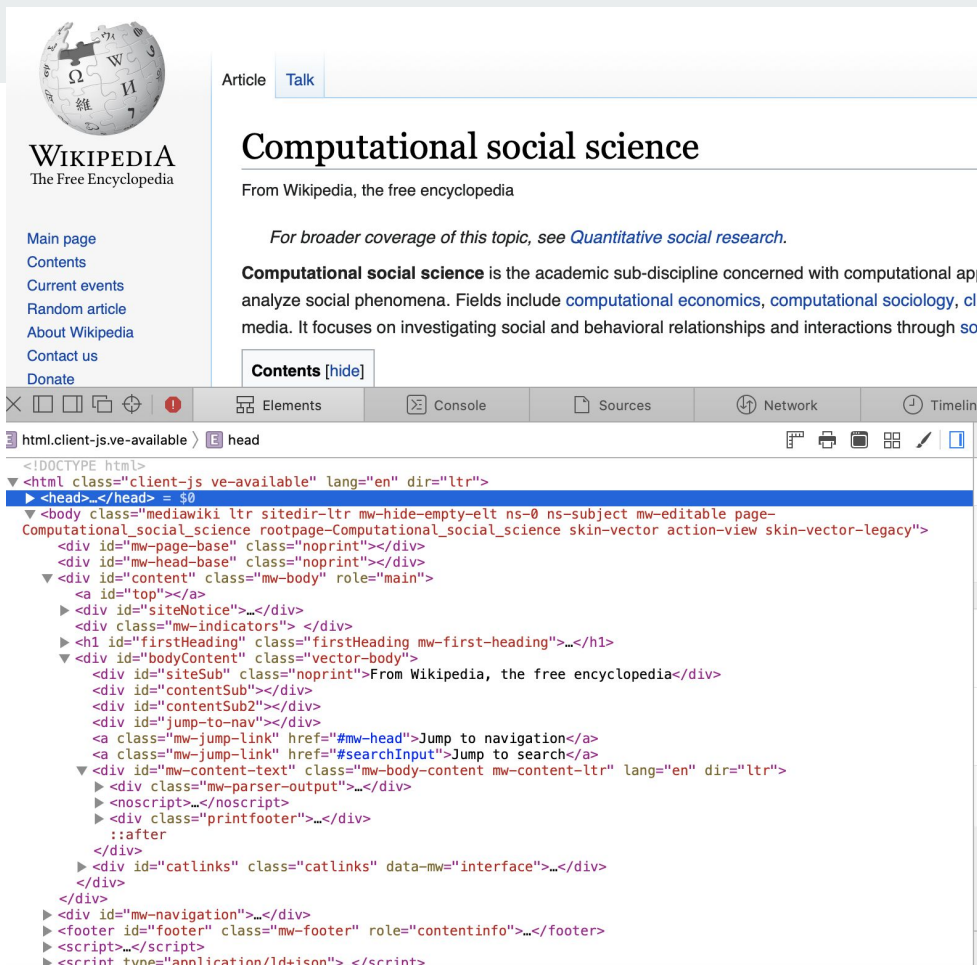
Research data: GitHub, Open Science Framework, Harvard Dataverse, ICPSR, etc.

Screen/web scraping



Developer tool

Inspect html/css codes



The screenshot displays a web browser window with the Wikipedia article "Computational social science" open. The browser's developer tools are active, showing the HTML structure of the page. The page content includes a Wikipedia logo, navigation links, and the main article text. The developer tools panel is expanded to the "Elements" tab, showing the DOM tree with the following structure:

```
<!DOCTYPE html>
<html class="client-js ve-available" lang="en" dir="ltr">
  <head>...</head>
  <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable page-Computational_social_science rootpage-Computational_social_science skin-vector action-view skin-vector-legacy">
    <div id="mw-page-base" class="noprint"></div>
    <div id="mw-head-base" class="noprint"></div>
    <div id="content" class="mw-body" role="main">
      <a id="top"></a>
      <div id="siteNotice">...</div>
      <div class="mw-indicators"> </div>
      <h1 id="firstHeading" class="firstHeading mw-first-heading">Computational social science</h1>
      <div id="bodyContent" class="vector-body">
        <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia</div>
        <div id="contentSub"></div>
        <div id="contentSub2"></div>
        <div id="jump-to-nav"></div>
        <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
        <a class="mw-jump-link" href="#searchInput">Jump to search</a>
        <div id="mw-content-text" class="mw-body-content mw-content-ltr" lang="en" dir="ltr">
          <div class="mw-parser-output">...</div>
          <noscript>...</noscript>
          <div class="printfooter">...</div>
          ::after
        </div>
        <div id="catlinks" class="catlinks" data-mw="interface">...</div>
      </div>
    </div>
    <div id="mw-navigation">...</div>
    <div id="mw-footer" class="mw-footer" role="contentinfo">...</div>
  </body>
</html>
```




Screen/web scraping workflow

Install/load R package: Rvest, tidyverse, ggplot

Choosing webpage

Extracting content

Basic cleaning

Exploratory data analysis



Tool time: web/screen scraping

Code along!



Using API

Application Programming Interface: customized requests of data from the server/owner

Strengths: efficiency, personalization, automation, integration, broad scope, structured

Weaknesses: availability, query limitation, affordability, maintenance



Available APIs

Twitter API

Reddit API

NYT API

...



API workflow

Read API documentation

Sign up for a developer account

Once approved, set up an App

Save your credentials (keys and tokens)

Set up user authentication information

Get your R environment set up

Start your first API request

Data wrangling



Twitter Academic API

Launched in Jan. 2021

Full history of Twitter content

10 million tweets per month (academic research access)



Tool time: Twitter API

Code along!



Group activity Day 2

Summary

An open-ended group exercise to collect digital trace data, formulate research questions, and/or incorporate a hybrid research design.

Activity

- Split into small groups and select person(s) to take notes and report group process/results.
- 13:40-13:50: brainstorm potential research ideas and select one to pursue
- 13:50-14:00: discuss sampling strategy, strengths & weaknesses of the data
- 14:00-15:20: collect data
- 15:20-15:30: reflect on the strengths/limitations of what you have completed and ways to address
- 15:30-16:00: come back together as a large group and discuss projects at the end of the day