Web Scraping in the Social Sciences

Yasin Shafi

29 January 2026

Conceptual Questions

Please write three to ten sentence explanations for each of the following questions. **You are only required to answer ONE of the two questions below.**

1. In the social sciences, what are two ethical or scientific risks of collecting data via web scraping (e.g., representativeness, privacy, terms of service, measurement error, scraping-induced missingness)? For each risk, briefly describe one practical mitigation strategy you would use in a reproducible workflow.
   Answer:

   Representativeness and sampling bias: Web scraping often captures only publicly visible content, which may not represent the full population of interest. For example, scraping social media posts only captures users who chose to make their content public, potentially missing important demographic groups or perspectives. Scraped data reflects what website operators chose to display, not necessarily a random or representative sample. A practical mitigation strategy is to carefully document the sampling frame in the reproducible workflow — explicitly stating what population your scraped data represents, what's excluded, and how this might bias the findings. We should also consider triangulating with other data sources to validate patterns.

   Terms of service violations and legal/ethical boundaries: Many websites prohibit scraping in their terms of service, and violating these terms can have legal consequences or ethical implications. Some sites explicitly forbid automated data collection, while others may allow it under certain conditions. A key mitigation strategy is to always check and respect robots.txt files, which specify which parts of a site can be scraped. Additionally, implementing rate limiting in code to avoid overwhelming servers, and documenting legal review process in the workflow documentation are good practices. If we are not certain about whether scraping is permitted, we should seek permission or use alternative data sources.

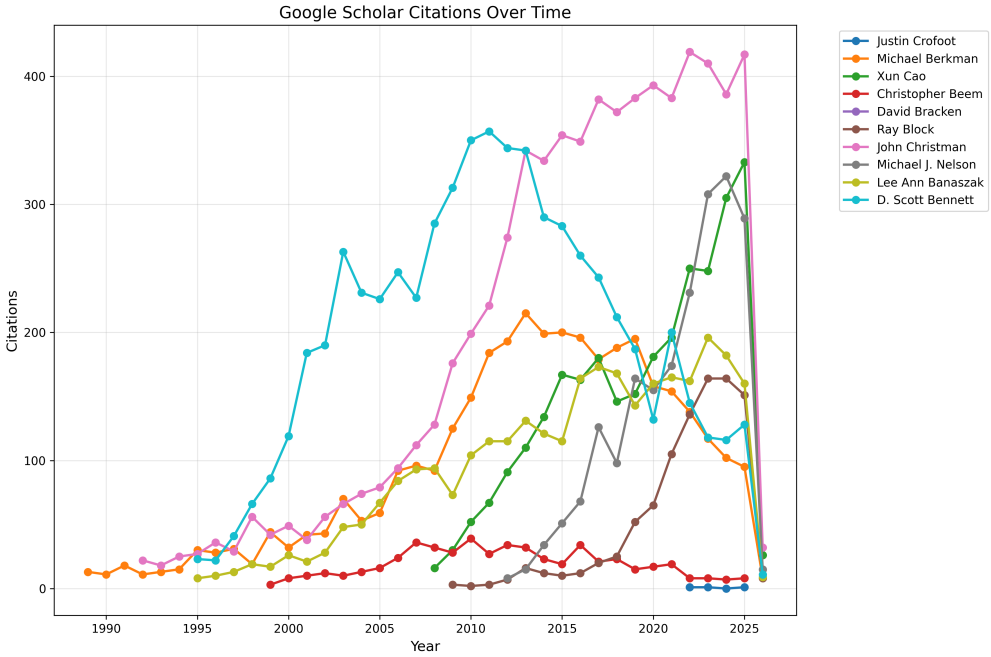
Applied Exercises

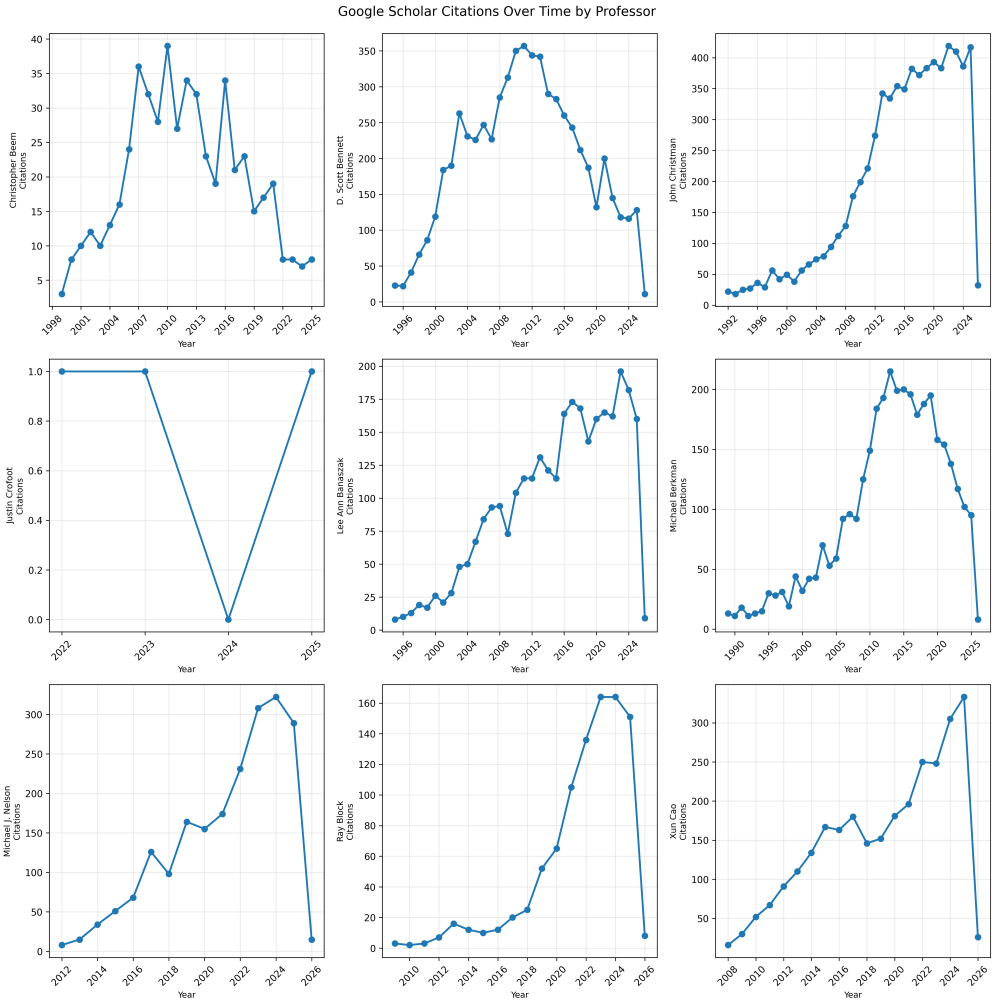
Answer: Work for this week's problem set is stored in: `https://github.com/yasinshafi/soda501/tree/main/03_web_as_data/problem_set`.

Use the code in the week's code tutorial and the lecture slides to answer the following questions.

3. Using **ten** Penn State faculty members from your department(s) or affiliated with SoDA, create a plot of **citations over time** for each professor.
    - Try changing the plot style (e.g., line thickness, points, theme, labels).
    - Your figure should be readable with 10 people (facets are fine).

Answer:

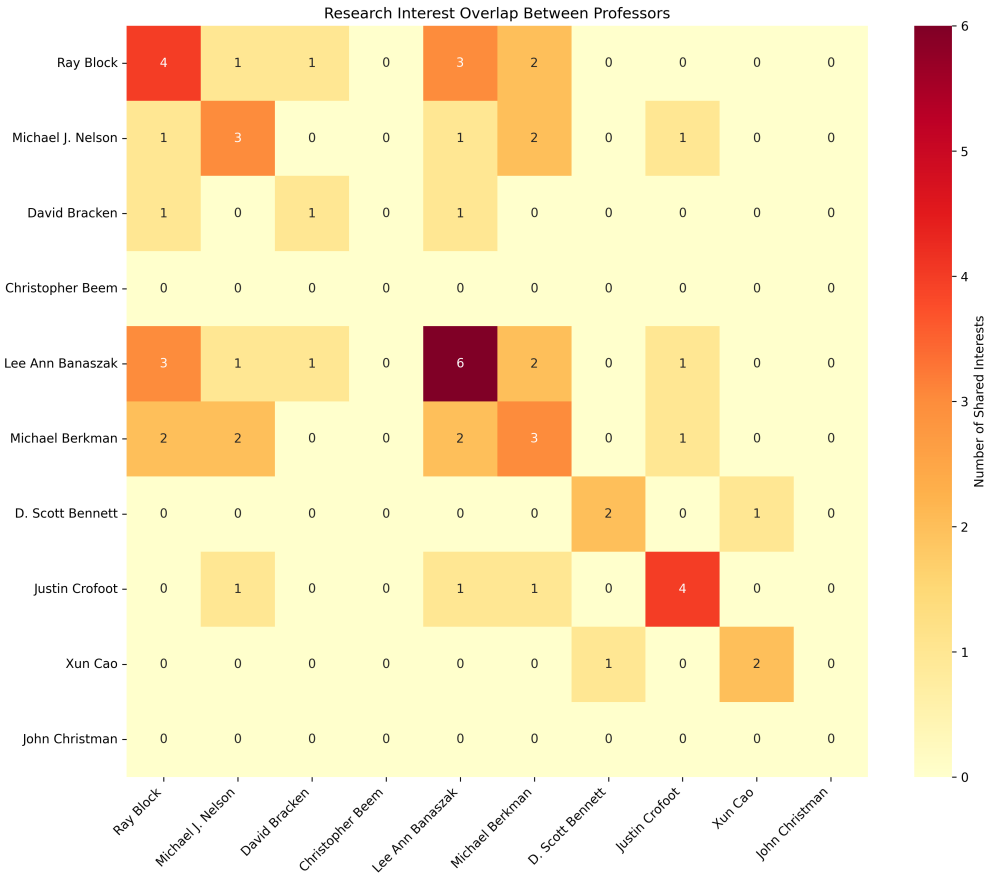




4. Visualize **or** discuss how the work of these professors overlaps.
   - One approach: use `scraped_interests` from PSU profile pages.
   - Another approach (advanced): compare their most common publication keywords.

   Provide at least one visualization **or** a short written discussion (5–10 sentences) describing the main overlap patterns you observe.

Answer:



5. What is the **median citation count (per year)** for each person in the data?
   - Hint: `group_by(name)` + `summarize(median(...))`.
   - Clearly state whether your median is computed over observed years only, or whether you treat missing years as zero (and why).

   Answer: The response is in the python script.

6. **Challenge Question (Optional — if you finish early):** Compute each scholar's **total citations** and **h-index** using `get_profile()`, then compare those across the faculty members.
   - Present a clean comparison (a table and/or a bar chart).
   - Briefly interpret what the comparison does *and does not* tell you (2–4 sentences).