Data Science Collaborator Application

7/24/2019

1. Project Lead:

Devin Judge-Lord
Ph.D. Candidate in Political Science (CV here)
University of Wisconsin-Madison
110 North Hall
JudgeLord@wisc.edu
715-204-4287

2. Project Summary:

Every day, unelected bureaucrats make thousands of important decisions affecting government services, individual rights, the distribution of federal grants, and the regulation of commerce. Elected officials, advocacy groups, companies, and even ordinary people often weigh in on these decisions. Yet we lack systematic data on how and when political actors attempt to shape government decisions by lobbying federal agencies directly. Compared to topics like voting and campaign donations, we know little about who contacts federal agencies, what kinds of agency decisions are the target of external pressure, and or whether external pressure affects agency decisions. For example: Are Members of Congress more likely to write letters of support for companies that donate to their campaigns? Do thousands of emails citizens affect how the Environmental Protection Agency regulates pollution?

I address this gap in two related projects. First, for my dissertation, I collect and analyze millions of public comments on proposed regulations. Second, in collaboration with a research team at UW-Madison and Stanford University, I collect and analyze hundreds of thousands of letters and emails received by federal agencies from Members of Congress. These data offer a rare look at political behaviors that are key features of American democracy but much less transparent than voting, legislating, or donating to political campaigns.

These new data help answer questions such as To what extent do legislators advocate for their constituents, donors, or policy goals? Are elected officials more likely to advocate for businesses that fund their campaigns? Which agency decisions attract external attention and which decisions go unnoticed? Does attention from the public or elected officials affect government decisions? If so, who benefits from external pressure on bureaucrats' decisions?

These projects face similar technical challenges and opportunities for broader impact with help from a data scientist:

Continuous integration: As Members of Congress, companies, advocacy groups, and individuals continue to write to federal agencies, both data generating processes are ongoing. Through regular Freedom of Information Act (FOIA) Requests and API calls, we have a continuous flow of new data. These new data occasionally produce errors in our data-cleaning code. The congressional letters project workflow stores data in Google Sheets so that research assistants can code data by hand. I developed methods to assess intercoder reliability among Google sheets coded by different RAs. I see broad potential for scientists, including ourselves, to better build interactive data like Google sheets into reproducible and continuously integrating workflows.

Methods to retrieve and process government data: Both projects required developing functions for API calls (data.gov, regulations.gov, or google's API) and web scraping. Other political scientists have asked to use the R code that I developed to pull data from the regulations.gov API and website. These functions would be accessible to a broader audience as an R package.

Content tagging methods: I have developed content-tagging algorithms to find the names of Members of Congress (in a variety of formats, even when they are partial or misspelled) and the names of organizations that lobby the federal government (including nonprofits, advocacy groups, and companies, which we link to parent company campaign donations). If turned into R packages, these algorithms could be more accessible to political scientists, journalists, and others to identify key political actors in large amounts of texts.

Integrating data sources: I have developed crosswalks linking data on Members of Congress (campaign donations, voting records, committee memberships, etc.), companies (parent and subsidiary companies, campaign donations, industry codes, etc.), and executive branch policies (regulations.gov data, Unified Agenda data, ORIA data, etc.). With the help of a data science collaborator, I hope to package these functions into R packages that allow researchers to augment one source of data with others. For example, after using the function to extract the names of Members of Congress from a text, a second function could augment these names with the unique identifiers that link to each member's voting record, campaign donations, or committee membership, and letter-writing behavior. A similar function could augment a data frame of regulations retrieved from the regulations.gov API with other data about each policy.

3. Data

Collected:

- Metadata on over 76 million public comments on proposed regulations (retrieved from regulations.gov API, stored as .Rdata on SSCC's server)
- 20,000 pdfs containing the text of 18.5 million public comments (stored on SSCC's server)
- Metadata on over 400,000 letters and emails from Members of Congress (collected via FOIA requests, stored as google spreadsheets and .Rdata on GitHub)
- 10,000 pdfs containing the texts of letters from Members of Congress (stored on SSCC's server)

Integrated:

- Metadata on all agency rules (retrieved from the Unified Regulatory and Deregulatory Agenda and Office of Information and Regulatory Affairs, stored on GitHub)
- Voting-based ideology scores for all Members of Congress and unique ICPSR identifiers linking to voting records (retrieved from voteview.com, stored on GitHub)
- Committee assignments and leadership positions for Members of Congress 2007-2018 (stored on GitHub)
- Campaign finance-based ideology scores for Members of Congress and unique DIME and FEC identifiers linking to campaign donations 1986-2014 (stored on GitHub)

To integrate:

- Spatial data on the economies and demographics of congressional districts (US Census, etc.)
- Committee assignments and leadership positions for Members of Congress 2000-2006 (data here)
- 2015-2018 campaign finance data from the DIME database (data here)
- Names of companies, parent companies, subsidiary companies appearing in congressional letters and their campaign donations (crosswalk being compiled here)