

Data for Democracy Lab

Michele P. Claibourn
University of Virginia Library

The Data for Democracy Lab is a place to promote the use of data science for the common good. In contrast to growing concerns that data science is deployed *against* citizens (O’Neil 2016), we aspire to a data science that is educative – working to ensure more people have access to the power of data, that is inclusive – emphasizing the value of diverse perspectives, experiences, and knowledge, and that is open – revealing the methods, models, and limits to all and allowing others to adapt, improve, or apply the work.

Our initial efforts include the Public Presidency Project (working with the StatLab Fellows) and the Public Interest Data Lab (a for-credit lab taught through the Data Science Institute, working on a client project for the Charlottesville Department of Social Services).

The Public Presidency Project

The gap between what citizens do and what democratic theory asks of them has rarely been larger. Individuals increasingly live in echo chambers and information bubbles, consuming only news that aligns with their ideological leanings (Jamieson and Cappella 2008; Pariser 2011). Scholarship on the limits of citizen ability to hold elected leaders accountable has grown increasingly pessimistic about this basic civic obligation (Achen and Bartels 2016). The already frequent expressions of distaste for politics among citizens (Hibbing and Theiss-Morse 2002) worsens as the political atmosphere becomes more polarized (McCarty, Poole, and Rosenthal 2016; Campbell 2016) and information channels proliferate. Consequentially, scholars have demonstrated that these attributes – selective attention, limited civic capacity, and a distaste for politics – heighten the likelihood that political leaders attend primarily to the goals and preferences of the relatively affluent (Bartels 2008; Gilens 2012).

In response, there’s a urgent need to imagine and develop new ways for the public to engage political news, one that allows people to monitor the activities and attentions of government from a relatively high level, that encourages the consumption of information from multiple and varied sources, and that lowers the barriers to attentive citizenship in ways that reduce inequalities in time, education, and access. Citizens don’t need to understand everything a president does to form reasonable evaluations (Schudson 1998), but even scanning the headlines in a complex, targeted media environment is becoming a strain.

The Public Presidency Project is designed to extract and present features of presidential attention and activity as represented by a range of diverging news and information sources. Using computational text analyses, we propose to develop a dynamic representation of presidentially-relevant news that citizens and researchers could use to monitor information about the presidency, to compare features across information streams, and to drill down on areas that interest them. The project affords opportunities for advancing new kinds of research in multiple fields – understanding the relation of presidential representations across information channels, developing new computational methods adapted for political media,

extending to additional institutions, levels of government, geographies and more; for creating hands-on educational experiences for undergraduates, graduate students, and staff through fellowships, courses, and lab involvement; and for contributing to the common good.

Libraries have been the vanguard of democratizing information by lowering barriers to access. The Library, in partnership with domain experts, is an ideal environment for researching and developing tools and methods to help citizens at large better use political information and lower the cost of political attentiveness. This project seeks to engage new kinds of work driven by [core library values](#): promoting equal information access and use, enabling democratic governance, and contributing to the public good.

Scope and Methods

The “proof of concept” for the project began with analysis of New York Times and Wall Street Journal coverage from January through August 2017. We have begun a [GitHub repository](#) to document our work and a [project website](#) to blog about research results. The goal is to generate a sampling of sources representing presidential actions, attention, and behavior from diverse perspectives and formats. Intended information sources, in addition to newspapers, including: television news (e.g., Fox News, CNN), Internet news (e.g., Politico, Breitbart), formal presidential statements (e.g., The Daily Compilation of Presidential Documents), and informal presidential communication (e.g., the Trump Twitter Archive).

From each source, we will extract a subset of relevant features. Possibilities include: Issue/policy attention; ideological position; tone or sentiment; topic; policy values; named entities; presidential events (e.g., agent-verb-object). Extracted features, but not the underlying text, will be graphically represented over time, in comparison, or through other visual means and made publicly accessible via a dedicated project website. In addition to their value in summarizing and comparing presidential representations and attention, these data can be used to address research questions about news production, agenda setting, framing, and tone across across and within information sources, policy issues, and events.

Each of these formats presents its own challenges in acquisition, wrangling, and analysis. The next steps include further developing the analysis of newspaper coverage, bringing in the formal presidential statements and the informal presidential twitter feed, and acquiring funds to access television news transcripts. *The NewsBank licensing to acquire news transcripts will be approximately \$10,000.*

Contributors

Michele Claibourn directs the project in the UVA Library. Beginning Fall 2017, we are piloting our StatLab fellowship. Four graduate students will be working as part of the lab to expand, shape, and contribute to the project while learning new methods, content, or collaborative tools. The lab has received support by social science librarians who have aided in database searches, data management librarians advising on data sharing resources, computational and statistical consultants who have helped to troubleshoot methods and

coding, and the Scholars' Lab which has advised on building a public web resource. We welcome involvement from interested research partners.

The Public Interest Data Lab

In partnership with the Data Science Institute, we are introducing a new for-credit "Public Interest Data Lab" this coming spring. The class responds to increasing demand for applied training and experience applying data science to problems in the public interest. The 10-week lab course will provide experience working collaboratively, openly, and reproducibly on data science projects organized by the director. We are working in tandem with the Batten School's Policy Labs to develop an applied lab framework around community engagement

The lab will focus on using open-source tools (GitHub, R) and resources (books and tutorials) and conducting our work as openly as possible, sharing workflows and evaluating the challenges and constraints publicly. Work will involve data wrangling, analysis, and visualization as well as acquisition of domain knowledge, use of reproducible workflows, and presentation and communication of results.

The first project responds to a call from the City of Charlottesville's Department of Social Services for a better understanding of the data collected on participants in the Child Welfare System (CWS). Our primary charge is to help the department understand if there is disproportionate minority contact in the CWS, whether there are differences in duration in the system based on race and ethnicity, if there are differential outcomes, and how children leave the system. The Virginia Department of Social Services has agreed to supply the requested data elements from Charlottesville cases for the past three years in January 2018.

Bibliography

- Achen, C. H. and L. M. Bartels (2016). *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton University Press.
- Bartels, L. M. (2008). *Unequal Democracy: The Political Economy of the New Gilded Age*. Princeton University Press.
- Campbell, J. E. (2016). *Polarized: Making Sense of a Divided America*. Princeton University Press.
- Gilens, M. (2012). *Affluence and Influence: Economic Inequality and Political Power in America*. Princeton University Press.
- Hibbing, J. R. and E. Theiss-Morse (2002). *Stealth Democracy: Americans' Beliefs About How Government Should Work*. Cambridge University Press.
- Jamieson, K. H. and J. N. Cappella (2008). *Echo Chamber: Rush Limbaugh and the Conservative Media Establishment*. Oxford University Press.
- McCarty, N., K. T. Poole, and H. Rosenthal (2016). *Polarized America: The Dance of Ideology and Unequal Riches*. Second. The MIT Press.
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

- Pariser, E. (2011). *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin Books.
- Schudson, M. (1998). *The Good Citizen: A History of American Civic Life*. Free Press.