729B: Applied Social Data Science
Fall 2023 - DRAFT

Logistics

Main instructor: Jóhanna Birnir (jkbirnir@umd.edu)

Module instructor: Ernesto Calvo (ecalvo@umd.edu)

Teaching Assistant and workshop coordinator Leo Bauer (leobauer@umd.edu)

Day and time Thursdays:

- September 7th - October 12th: 1:00pm - 3:45pm
- October 19th and November 2nd - December 7th 3:30pm - 6:15pm
- October 26th see workshop announcement for time and location

Location TYD 1136

Data Science: Life long collaborative learning

How do we prepare to tackle the Grand Challenges of our time with transparent evidence based approaches when the data science methods we use and teach are outdated as soon as we teach them? Our answer is an approach to continuous learning of methods that can be sustained throughout a career, with an emphasis on transparency. Correspondingly, the course learning components consist of faculty and student led (rotating) modular training in current data analysis with emphasis on transparency and institutionalization of a graduate student led methods workshop. [1] The workshop will train graduate students to collaboratively further their methods skills and to survey and host external presenters to inform participants of cutting edge methodological advances.

[1] Modular class components may be taught by different people in different years and updated in tandem with the development of new data science approaches.

Learning by doing: Self-directed methods training.

Data science is a rapidly evolving field where new data and methods are continuously developed, updated and shared. To contribute to solving current and future Grand Challenges students must learn not only the standard analytical tools for causal inference but also how to continually learn new methods, to find, curate, analyze, and share new data, and how to make their data and methods publicly available for the greatest transparency in service of scientific validation. To train students in continuous self directed and collaborative learning approaches this class employs a two part experiential approach of scholarship in practice.

- **Modular training in current methods.** By modular methods training we are referring to self contained course components that can be taught in a shorter time than the traditional 15 week course format. Instead of a class being built around the instruction of a single method of analysis the class consists of modules that can be rotated. As such, modular methods training is sufficiently flexible to be varied and updated from year to year based on advances in the field, the expertise of the instructor of a given module, and on student demand. At the same time all modules teach a common foundation for continuous self directed methods training. Common components within each module include the development of a strategy to find and gather the most appropriate data for the task at hand, the selection and use of appropriate methods (including programming language, software and packages, algorithms and statistics, and troubleshooting code) for the analysis of that data, and the storing and transparent sharing of the data and methods. In this phase the two modules that form teh centerpiece of the training anchor on Natural Language Processing (NLP) and network analyses. The instructors will divide the research process by creating discrete tasks, each with distinct methodological challenges, ensuring that the distinct modules give graduate students control of the full research process: the choice of platform, software, packages, and functions, along with effective troubleshooting and how to capture social media and newspaper data. They will also teach students how to store data, work with repositories on Github, and how to disseminate their findings through public websites. The modular instruction will consist of hands on training as students complete all steps of the analysis alongside the instructor on separate cases of their choosing.

- **Institutionalization of a graduate student led methods workshop.** The second component emphasizes collaborative methods updating and participation in creating a public information good. To this end the lead instructor works with a graduate student lead who, in consultation with the student body, selects current methods topics to be presented by students and outside experts in the workshop. The faculty adviser and the graduate student lead will also work with student presenters in the program to develop professional presentations on methods that the students

are working to master. These presentations will be recorded and archived on Github and maintained by the Graduate Student Association with the assistance of faculty to be made available to successive cohorts of graduate students. The faculty instructor will also work with the graduate student lead, in consultation with other graduate students in the program, to find and host an outside presenter of cutting edge methods from industry or academia.

Course Requirements

In person seminar attendance and active participation throughout the semester is required. This is an applied class where several components build on one another. Therefore, it is vital that students attend all classes. If you have to miss a class for any reason you must clear this with the instructor and make sure that you keep up with the material presented in the session. Because classes are run like workshops with presentations followed by individual assisted completion of tasks take breaks as you need and please inform the instructors of any special needs or sensitivities that should be taken into account in the format of the class.

- **Pre-requisites.**

  We expect that students will take this class in the first semester of their second year when they have mastered some technical skills and have started to think about

  - a: the methods they would like to use in their own research
  - b: how to best work with others, present and publicise their work

  Specifically, we expect that students have basic proficiency in R. Familiarity with tidy is a plus but not required. No prior familiarity with any of the methods topics covered is expected. If students need a refresher on R we recommend a short introduction to R, for example, at DataCamp.com.

- **Modular methods training.**

  Assessment: Students will create their own mini-projects as they work their ways through the basic modules. Additionally, the students will apply at least one of the data gathering methods and show some proficiency in applying some of the analytical methods to one substantive project of their choosing. The substantive projects are expected to showcase mastery of methods that will vary depending on the modules. Finally, students are expected to showcase their substantive project on their own websites hosted on Github. Students can work individually but are encouraged to collaborate in pairs that may rotate throughout the semester.

- **Graduate student led methods workshop.**

  Implementation: The workshop runs all year and starts in the fall semester. In fall the faculty instructor works with the graduate student leader of the workshop in soliciting modular workshop proposals for components to be peer developed for a monthly workshop offered throughout the year. The instructor will assist the graduate student lead who oversees the workshop for the full year. The instructor works with the graduate student lead to invite one outside speaker for the fall semesters. Students in the class are expected to attend the graduate student led workshop for the duration of the class and hopefully beyond the class.

Academic Conduct

It is assumed that all students are familiar with and adhere to the code of academic integrity. For the relevant policies see: gradschool.umd.edu

Diversity

The University of Maryland and the Department of Government and Politics values diversity. Diversity refers to differences in race, ethnicity, culture, gender, sexual orientation, religion, age, abilities, class, nationality, and other factors. We are committed to creating a respectful and affirming climate in which all students, staff, and faculty are inspired to achieve their full potential. We believe that actively fostering an affirming environment strengthens our department as a whole. A department that values and celebrates diversity among its students, staff and faculty is best able to develop the strengths and talents of all members of the department community.

I invite you, if you wish, to tell us how you want to be referred to both in terms of your name and your pronouns (he/him, she/her, they/them, etc.). The pronouns someone indicates are not necessarily indicative of their gender identity. Visit trans.umd.edu to learn more. Additionally, how you identify in terms of your gender, race, class, sexuality, religion, and dis/ability, among all aspects of your identity, is your choice whether to disclose (e.g., should it come up in classroom conversation about our experiences and perspectives) and should be self-identified, not presumed or imposed. I will do my best to address and refer to all students accordingly, and I ask you to do the same for all of your fellow Terps.

Schedule

Readings are to be found on our class space on ELMS.
Tutorials are also on ELMS and updated versions will be made available on www.johannabirnir.com

Introductions

September 7: Data science and resources for learning.

Why data science?

local resources for data science

- The graduate student methods workshop and resources maintained by the GSA (see https://github.com/gsa-gvpt/gvpt-methods)

Is your computer up to it? Make sure you have recent R and R studio updated on your computer.

Assignment

If you experienced problems running the script on the dataset provided in class work to resolve this issue before the following class.

Foundational Modules - Birnir

September 14: Storing your work and working together: Github; Wiriting: Latex

Required Reading

Familiarize yourselves with:
https://happygitwithr.com/index.html
Familiarize yourselves with
https://overleaf.com/

Assignment 1 (a and b)

1) Assignment 1a: Create a free account on github if you don't already have one. Identify and clone to your local computer one tutorial from a GSA repo. Add a document relevant to your learning the method and reconcile the repo with a github repo that you have created for the purpose of learning the skill in the tutorial you cloned. Upload a screenshot of the repo with your added document to ELMS.

2) Assignment 1b: Create a free account on Overleaf if you don't already have one. Take one of your existing documents and style the text using either the assignment or paper template from the GVPT Methods Workshop Github under introLaTeX/Templates. Upload the pdf to ELMS.

September 21: Transparency in the social sciences: Preparing your work for public distribution and replication. Workflow and webpage creation in quarto and hosting on github.

The objective of this module is to get you started on the final project for the class. Your final project consists of displaying and making publicly available on a webpage that you have built in quarto and hosted on your github account, the data and source codes for a project that graphs analysis of text related to your research. The project should incorporate some elements from the social network analysis introduced in the second part of the class.

Required reading

Familiarize yourselves with
https://quarto.org/docs/websites/ R for Data Science 2e. Chapters 30-32.
https://r4ds.hadley.nz/quarto.html

Recommended Reading

Familiarize yourselves with:
https://quarto.org/

Assignment2

Screenshot your landing webpage and upload to ELMS.

Substantive module 1: Text as data - Birnir

September 28: Working with text as data

corpuses, dictionaries, basic analysis (Frequencies, Comparisons, Wordclouds, Sentiment)

Required reading

Text Mining with R: A Tidy Approach
https://www.tidytextmining.com/.
Chapters 1-6

Recommended reading

Stephen Wolfram. 2023. What Is ChatGPT Doing ... and Why Does It Work? Wolfram Media.

R for data science 2e
https://r4ds.hadley.nz/

Assignment 3

Use one of the texts in the class (or another text of your choosing) to create and clean a corpus, tokenize the words and produce a graphical representation of your choice (frequency, wordcloud etc). Upload to ELMS.

October 5: Collecting text as data, webscraping, pulling data from reddit, extracting data from pdfs

Required reading

R for data science 2e. Chapter 26.
https://r4ds.hadley.nz/webscraping.html

Recommended reading

https://smltar.com/

Assignment 4

Scrape all email addresses from the GVPT Faculty webpage and store them in a dataframe. Upload your codefile to ELMS.

Bonus: Find a way to scrape the names as well and store them in a separate column in the same dataframe so that names correspond with email addresses.

October 12: Machine learning in research

Required reading

Leo Breiman. 2001. Statistical Modeling: The Two Cultures *Statistical Science*. 16(3):pp. 199-215

Green, Jon, and Mark H. White II. Machine Learning for Experiments in the Social Sciences. 2023. Cambridge University Press, Elements Series in Experimental Political Science.

Overos et al. Working paper. AMAR and the Machine: Assisted Text Analysis for Coding of Cross-Sectional Time Series Data.

Recommended reading

Stephen Wolfram. 2023. What Is ChatGPT Doing ... and Why Does It Work? Wolfram Media.
See also: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Bradley Boehmke Brandon Greenwell. Hands on Machine Learning with R. https://bradleyboehmke.github.io/HOML/
Lones, Michael. 2023 How to avoid Machine Learning Pitfalls. A guide for researchers. http://www.macs.hw.ac.uk/ ml355/

Assignment 5

Use one of the datasets provided (or a dataset of your choosing) to run a classification model as demonstrated in class. Experiment with tuning. Graph 2 results (table or plot) and upload on ELMS along with your code.

Substantive Module 2: Machine learning for social network analysis. Calvo.

October 19: Social Media Research: APIs, Meta-Data, Text-as-data, Cross-Plataform research

packages (including httr, igraph, quanteda, rdd)

Required reading

Salganik, M. J. (2019). Bit by bit: Social research in the digital age. Princeton University Press. Chapters 1 and 2.

Wu, P. Y., Tucker, J. A., Nagler, J., & Messing, S. (2023). Large language models can be used to estimate the ideologies of politicians in a zero-shot learning setting. arXiv preprint arXiv:2303.12057.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? Psychological science, 26(10), 1531-1542.

Assignment 6

Connect to ChatGPT's API and return a dataset of dyadic comparisons between news media outlets. Upload your code to ELMS.

October 26: The Cutting Edge in Data Science

Invited guest speaker

- Professor Andrew Halterman. (https://andrewhalterman.com/)

- October 25, 2:30-4pm

- for ZOOM link swee announcement.

The event is co-hosted by the methods field and the graduate student workshop. The event is virtual and will take place on October 25th from 2:30-4.

November 2: Working with Dashboard Data

Required reading

Aruguete, N., Calvo, E., & Ventura, T. (2021). News by popular demand: Ideological congruence, issue salience, and media reputation in news sharing. The International Journal of Press/Politics, 19401612211057068.

Assignment 7

Create a corpus of news articles from Buzzsumo data. Upload your code to ELMS.

November 9: Annotating text and creating datasets with Large Language Models

Required reading

Törnberg, P. (2023). Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. arXiv preprint arXiv:2304.06588.

Optional reading

Heseltine, M; von Hohenberg, B. C. (2023) Large Language Models as a Substitute for Human Experts in Annotating Political Text. Mimeo.

Assignment 8

Use LLM to annotate a corpus of text. Upload

November 16: Working with Networks

Required reading

Imai, K. (2018). Quantitative social science: an introduction. Princeton University Press. Chapter 5.

Feld, S. L. (1991). Why your friends have more friends than you do. American journal of sociology, 96(6), 1464-1477.

Eom, Y. H., & Jo, H. H. (2014). Generalized friendship paradox in complex networks: The case of scientific collaboration. Scientific reports, 4(1), 1-6.

Assignment 9

Network visualization of Members of Congress using Twitter data. Upload your visualization.

November 23: No class Thanksgiving

November 30: Working with Events (and time)

Required reading

Calvo, E., Waisbord, S., Ventura, T., & Aruguete, N. (2023). Winning! Adjudication and Dialogue in Social Media. PlosOne (Forthcoming)

Lansdall-Welfare, T., Dzogang, F., & Cristianini, N. (2016, December). Changepoint analysis of the public mood in UK Twitter during the Brexit referendum. In 2016 IEEE 16th international conference on data mining workshops (ICDMW) (pp. 434-439). IEEE.

Final Assignment 10

Use the skills gained in this class to produce a data collection and analysis of your choosing with a graphical display of the results. Upload this graphical display to your website. During the class students will show and explain their publicly available projects to their class colleagues.

Conclusions: Birnir and Calvo

December 7: Student presentations