729B: Applied Social Data Science
Spring 2023

Logistics

Main instructor: Jóhanna Birnir

Module instructor: Ernesto Calvo

Teaching Assistant and workshop coordinator (with Harriet Goers): Jordan Dewar

Day and time: Tuesdays 10:00am - 12:45pm

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Challenges of our time with transparent evidence based approaches when the data science methods we use and teach are outdated as soon as we teach them?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. The course learning components consist of faculty 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.

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,

[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.

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 pilot the two modules that will be taught focus on conducting Natural Language Processing (NLP) of news data on Social and Racial Injustice and collecting and processing social media data for early detection of Treats to Democracy. 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 will work 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.[2] 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

[2]For spring 23 Birnir met with workshop leaders Dewar and Goers to decide on general topics and discuss outside speaker.

archived on Github and maintained by the 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

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 basic methods 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

  Therefore, 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.

- **Modular methods training.**

  Assessment: Rotating student pairs 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. For the substantive, showcased, projects students are expected to work in pairs.

- **Graduate student led methods workshop.**

Implementation: The workshop runs all year and starts in the fall semester. In fall the faculty instructor will work 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 oversee the workshop for the full year. Implementation of this course component will begin in fall 2022 and continue through spring 2023. The graduate student selected outside speaker will, in the pilot face of this course, be scheduled for **spring 2023.** Thereafter, the graduate student led workshop will become an integral part of the reconfigured semester long course but with faculty collaboration for the entire year.

Assessment: The usefulness of the workshop will be assessed with a survey of students at the end of the pilot and annually thereafter. This survey will also gather information that can be used to organize the workshop in the following year.

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

January 31: Data science and resources for learning.

Why data science?

Guest Presenter:

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

Foundational Modules - Birnir

February 7: Storing your work and working together: Github

Required Reading

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

February 14: Workflow organizing and presenting

Quarto and Overleaf (including beamer)

Required Reading

R for Data Science 2e. Chapters 30-32.
https://r4ds.hadley.nz/quarto.html

Recommended Reading

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

February 21: Acquiring the data.

scraping: web data, text files (pdf)

For this module please identify some news sites of interest to you, please also bring some pdf files with data that you would like to extract information from

for your research.

Required reading

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

Recommended reading

https://smltar.com/

February 28th: 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

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

March 7: Machine learning for analysis of text.

Guest presenters:

- Henry Overos: AMAR and the Machine.
- Leo Heinrich Maria: Transformers.

Required reading.

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

Vallejo Vera, Sebastian. "Rage within the Machine: Activation of Racist Content in Social Media." Latin American Politics and Society (Forthcoming)

March 14: Transparency in the social sciences: Making your work publicly available

Building a website on quarto and hosting it on github for the purpose of giving public access to your work

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/

March 21. SPRING BREAK. NO CLASS

March 28: The Cutting Edge in Data Science

Invited guest speaker

- Professor Steven Miller. (http://svmiller.com/)

The event is co-hosted with the graduate student workshop. This year the event is virtual as our guest speaker is in Sweden. The location of the event is 1101 Morrill Hall and begins at 10am EST. Breakfast will be provided. Please see announcement from the graduate student workshop leaders for the zoom-link.

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

April 4: Social Media Research: APIs, Meta-Data, Text-as-data, Cross-Plataform research

packages (including academictwitteR, 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.

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.

April 11: 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.

April 18: 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.

April 25: Working with Text with location

Required reading

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

May 2: Working with Hyperlinks

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.

May 9: Conclusions: Machine learning for your research

Required reading

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

Lones, Michael. 2023 How to avoid Machine Learning Pitfalls. A guide for researchers. http://www.macs.hw.ac.uk/ ml355/

Students are expected to have their websites up an running with their publicly available projects, for a show and tell.