Syllabus
SoDA 501
Approaches and Issues in Big Social Data
Spring 2024
Monday 9-12, The DataBasement

Instructor: Bruce A. Desmarais

Office:	Sparks B001 (The Databasement)
Email:	bdesmarais@psu.edu
Web Page:	`brucedesmarais.com`
Office Hours:	1:30-2:30 M, 9-10 W & by appt.

Course Description: This seminar is part of the core sequence for students in the Social Data Analytics dual-title PhD and doctoral minor. The primary objective of the seminar is interdisciplinary exposure to, engagement with, and integration of the tools, practices, language, and standards used in the collection and management of data in the component disciplines of the Social Data Analytics field. Each of you is well on your way toward a PhD– formal certification as an "expert"– in one of the component disciplines of Social Data Analytics and has in your coursework and research become well versed in one or more of the many computational, informational, statistical, visual analytic, or social scientific approaches to data, and the issues faced by those approaches. Here, we are interested in trying to integrate your multidisciplinary expertise, particularly in the context of data that are social (about, or arising from, human interaction) and big or intensive (of sufficient scale, variety, or complexity to strain the informational, computational, or cognitive limits of conventional approaches to data collection, management, manipulation, or analysis).

Reading presentations (20% of grade): Each week we will briefly discuss one or more published papers that make core use of the concepts covered in the respective week. The papers are listed in the course schedule below. Each student must give one ten-minute presentation in which they summarize the article and describe the role played by the core concepts covered in the course in the respective application. Students must sign up for their presentation via SignUpGenius by Thursday, 1/11.

Resource/methods tutorial (30% of grade): Each student will be responsible for presenting a detailed tutorial of a resource or methodology that is relevant to the core concepts

presented in the respective class session. A complete draft of the tutorial must be submitted via email to the instructor one week prior to the in-class presentation date. The tutorial should be approximately 30-minutes long. In the event that two students sign up on the same date, they must coordinate to assure that they are not presenting the same method/resource. Students must sign up for their presentation via SignUpGenius by Thursday, 1/11.

Research paper (50% of grade): Students are required to write an original research paper as part of a group. The final draft is due on Wednesday, May 1st. The following list includes a set of project submission phases with deadlines. If you need an extension for whatever reason, please do not hesitate to ask (ideally, ahead of the deadline).

- **Topic**: The paper can address virtually any topic in which students are interested. I simply ask that the plan involve the development or application of innovative social data, broadly conceptualized. The topic of the research project, including a description of the data to be generated and/or collected, should be written up in a 1–2 page document. This is due on 2/5.
- **Pre-registration report**: A pre-registration report is a research design prepared with enough detail that the reader knows exactly what the authors will do, and why. Ideally, the pre-registration report will include a complete draft of the final paper excluding the results. Students should submit a first draft that includes an introduction, theory, research design, and plan for analyzing the data collected. This is due on 3/25.
- **Presentation**: Students will deliver twenty-minute presentations during the final meeting period of the semester (on 4/22).

Grading:

- Application presentation: 20%.
- Resource/methods tutorial: 30%.
- Research paper: 50%.

Grading Scale.

Grade	Lower	Upper
A	93	101
A-	90	93
B+	87	90
B	83	87
B-	80	83
C+	77	80
C	73	77
C-	70	73
D+	67	70
D	63	67
D-	60	63
F	0	60

Course Schedule: The schedule below gives the required reading. The readings listed for a particular date should be read before class time that day. Full citations can be found below in the references section.

Course Introduction

1. 1/08, Course Intro:
    - Review syllabus on Canvas.
    - Create an account on ROAR Collab and login to the portal.
    - Fill out the background form at `https://docs.google.com/forms/d/e/1FAIpQLScCYGh_6TEl6tbXz3IXeEvU0OJNonHJls-q_mFJNmEMUEcMgQ/viewform?usp=sf_link`
2. 1/15, The Big Picture on Big Social Data
    - Mitchell et al. (2021)
    - Poschmann and Goldenstein (2022)
3. 1/22, Experiments (guest lecture by Nicole Kreisberg)
    - Hofman et al. (2021)
    - Carey, Nyhan, Phillips and Reifler (2022)
    - Kreisberg (2023)

4. 1/29, Survey Data (Lijiang Shen)

    - Argyle, Busby, Fulda, Gubler, Rytting and Wingate (2023)
    - Salomon et al. (2021)
    - Guest topic TBD

5. 2/05, Ethics (guest lecture by Amulya Yadav)

    - **Homework:** Complete the Responsible Conduct of Research training in CITI `citi.psu.edu`.
    - Chen and Quan-Haase (2020)
    - Fiesler, Beard and Keegan (2020)
    - Rajtmajer et al. (2022)

6. 2/12, Measurement (guest lecture by Cassandra Tai)

    - Drost (2011)
    - Quinn et al. (2010)
    - TBD, guest paper

7. 2/19, Data from Physical Sensors, Opportunities and Risks (guest lecture by Rick Gilmore)

    - Fowler and Dawes (2008)
    - Avery, Giuntella and Jiao (2022)
    - TBD, guest topic

8. 2/26, Relational databases (guest lecture by Carrie Brown)

    - Chapters 2, 5, & 6 of Badia (2020)
    - Meraz and Papacharissi (2013)

9. 3/11, Digital Social Data (guest lecture by Michael Burnham)

    - Tenkanen et al. (2017)
    - Bail, Argyle, Brown, Bumpus, Chen, Hunzaker, Lee, Mann, Merhout and Volfovsky (2018)

- TBD, guest paper

10. 3/18, Text and Image data (guest lecture by Kevin Munger)
    - Islam and Goldwasser (2021)
    - Torres and Cantú (2022)
    - TBD, guest topic
11. 3/25, Geospatial data (guest lecture by Corina Graif)
    - Andris (2016)
    - De Choudhury, Counts and Horvitz (2013)
    - TBD, guest paper
12. 4/01, Time series and longitudinal data (guest lecture by Sy-Miin Chow)
    - Van Der Donckt et al. (2022)
    - Gangl and Ziefle (2015)
    - TBD, guest paper
13. 4/08, Large Scale Official Data (guest lecture by Johabed Olvera).
    - Gunderson, Cohen, Schiff, Clark, Glynn and Owens (2021)
    - Kim et al. (2018)
    - TBD, guest paper
14. 4/15, Network Data (guest lecture by Cornelius Fritz)
    - Butts (2008)
    - Gile (2011)
    - TBD, guest paper
15. 4/22, Project presentations

Disability Accommodation Statement Penn State welcomes students with disabilities into the University?s educational programs. Every Penn State campus has an office for students with disabilities. Student Disability Resources (SDR) website provides contact information for every Penn State campus (http://equity.psu.edu/sdr/disability-coordinator). For further information, please visit Student Disability Resources website (http://equity.psu.edu/sdr/).

In order to receive consideration for reasonable accommodations, you must contact the appropriate disability services office at the campus where you are officially enrolled, participate in an intake interview, and provide documentation: See documentation guidelines (http://equity.psu.edu/sdr/guidelines). If the documentation supports your request for reasonable accommodations, your campus disability services office will provide you with an accommodation letter. Please share this letter with your instructors and discuss the accommodations with them as early as possible. You must follow this process for every semester that you request accommodations.

Academic Integrity Statement Academic integrity is the pursuit of scholarly activity in an open, honest and responsible manner. Academic integrity is a basic guiding principle for all academic activity at The Pennsylvania State University, and all members of the University community are expected to act in accordance with this principle. Consistent with this expectation, the University's Code of Conduct states that all students should act with personal integrity, respect other students? dignity, rights and property, and help create and maintain an environment in which all can succeed through the fruits of their efforts.

Academic integrity includes a commitment by all members of the University community not to engage in or tolerate acts of falsification, misrepresentation or deception. Such acts of dishonesty violate the fundamental ethical principles of the University community and compromise the worth of work completed by others.

Counseling and Psychological Services Statement Many students at Penn State face personal challenges or have psychological needs that may interfere with their academic progress, social development, or emotional wellbeing. The university offers a variety of confidential services to help you through difficult times, including individual and group counseling, crisis intervention, consultations, online chats, and mental health screenings. These services are provided by staff who welcome all students and embrace a philosophy respectful of clients? cultural and religious backgrounds, and sensitive to differences in race, ability, gender identity and sexual orientation.

Counseling and Psychological Services at University Park (CAPS)
(http://studentaffairs.psu.edu/counseling/): 814-863-0395

Counseling and Psychological Services at Commonwealth Campuses
(http://senate.psu.edu/faculty/counseling-services-at-commonwealth-campuses/)

Penn State Crisis Line (24 hours/7 days/week): 877-229-6400 Crisis Text Line (24 hours/7

days/week): Text LIONS to 741741

Educational Equity/Report Bias Statements Consistent with University Policy AD29, students who believe they have experienced or observed a hate crime, an act of intolerance, discrimination, or harassment that occurs at Penn State are urged to report these incidents as outlined on the University?s Report Bias webpage (http://equity.psu.edu/reportbias/)

References

Andris, Clio. 2016. "Integrating social network data into GISystems." *International Journal of Geographical Information Science* 30(10):2009–2031.
URL: *https://doi.org/10.1080/13658816.2016.1153103*

Argyle, Lisa P, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting and David Wingate. 2023. "Out of one, many: Using language models to simulate human samples." *Political Analysis* 31(3):337–351.

Avery, Mallory, Osea Giuntella and Peiran Jiao. 2022. "Why Don't We Sleep Enough? A Field Experiment Among College Students." *Review of Economics and Statistics* pp. 1–45.

Badia, Antonio. 2020. *SQL for data science: data cleaning, wrangling and analytics with relational databases*. Springer Nature.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M.B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences* 115(37):9216–9221.

Butts, Carter T. 2008. "network: a Package for Managing Relational Data in R." *Journal of statistical software* 24:1–36.

Carey, John, Brendan Nyhan, Joseph B. Phillips and Jason Reifler. 2022. "Partisanship Unmasked? The Role of Politics and Social Norms in COVID-19 Mask-Wearing Behavior." *Journal of Experimental Political Science* pp. 1–14.

Chen, Wenhong and Anabel Quan-Haase. 2020. "Big data ethics and politics: Toward new understandings." *Social Science Computer Review* 38(1):3–9.

De Choudhury, Munmun, Scott Counts and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*. pp. 47–56.

Drost, Ellen A. 2011. "Validity and reliability in social science research." *Education Research and perspectives* 38(1):105–123.

Fiesler, Casey, Nathan Beard and Brian C. Keegan. 2020. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the international AAAI conference on web and social media*. Vol. 14 pp. 187–196.

Fowler, James H and Christopher T Dawes. 2008. "Two genes predict voter turnout." *The Journal of Politics* 70(3):579–594.

Gangl, Markus and Andrea Ziefle. 2015. "The making of a good woman: Extended parental leave entitlements and mothers' work commitment in Germany." *American Journal of Sociology* 121(2):511–563.

Gile, Krista J. 2011. "Improved inference for respondent-driven sampling data with application to HIV prevalence estimation." *Journal of the American Statistical Association* 106(493):135–146.

Gunderson, Anna, Elisha Cohen, Kaylyn Jackson Schiff, Tom S Clark, Adam N Glynn and Michael Leo Owens. 2021. "Counterevidence of crime-reduction effects from federal grants of military equipment to local police." *Nature human behaviour* 5(2):194–204.

Hofman, Jake M. et al. 2021. "Integrating explanation and prediction in computational social science." *Nature* 595(7866):181–188.

Islam, Tunazzina and Dan Goldwasser. 2021. Analysis of Twitter Users' Lifestyle Choices using Joint Embedding Model. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15 pp. 242–253.

Kim, Young Mie et al. 2018. "The stealth media? Groups and targets behind divisive issue campaigns on Facebook." *Political Communication* 35(4):515–541.

Kreisberg, A Nicole. 2023. "Nativity Penalty and Legal Status Paradox: The Effects of Nativity and Legal Status Signals in the US Labor Market." *Social Forces* 101(3):1343–1371.

Meraz, Sharon and Zizi Papacharissi. 2013. "Networked gatekeeping and networked framing on Egypt." *The International Journal of Press/Politics* 18(2):138–166.

Mitchell, Shira et al. 2021. "Algorithmic fairness: Choices, assumptions, and definitions." *Annual Review of Statistics and Its Application* 8:141–163.

Poschmann, Philipp and Jan Goldenstein. 2022. "Disambiguating and specifying social actors in big data: Using Wikipedia as a data source for demographic information." *Sociological Methods & Research* 51(2):887–925.

Quinn, Kevin M. et al. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54(1):209–228.

Rajtmajer, Sarah et al. 2022. A synthetic prediction market for estimating confidence in published work. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36 pp. 13218–13220.

Salomon, Joshua A. et al. 2021. "The US COVID-19 Trends and Impact Survey: Continuous real-time measurement of COVID-19 symptoms, risks, protective behaviors, testing, and vaccination." *Proceedings of the National Academy of Sciences* 118(51):e2111454118.

Tenkanen, Henrikki et al. 2017. "Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas." *Scientific Reports* 7(1):1–11.

Torres, Michelle and Francisco Cantú. 2022. "Learning to see: Convolutional neural networks for the analysis of social science data." *Political Analysis* 30(1):113–131.

Van Der Donckt, Jonas et al. 2022. "tsflex: Flexible time series processing & feature extraction." *SoftwareX* 17:100971.