NYU | ROBERT F. WAGNER GRADUATE SCHOOL OF PUBLIC SERVICE

PADM-GP 2505

Advanced Data Analytics and Evidence Building
Spring 2022 Section 1

Instructor Information

- Julia Lane
- Email: jil4@nyu.edu
- Office Address: NYU-Wagner, Room 3038, 295 Lafayette Street, New York, NY 10012
- Office Hours: Wednesday 4-5pm with instructors 4-5pm with instructors Carolyn Gorman csg9413@nyu.edu and Ekaterina Levitskaya el2727@nyu.edu.

Course Time and Location

- Lecture: Tuesdays 2-3:40pm

Course Description and Learning Objectives

The goal of this course is to develop the key data analytics skill sets necessary to inform evidence-based policy. Its design offers hands-on training in how to make sense of and use large scale real world heterogeneous datasets in the context of addressing real world problems. The main learning objectives are to develop a better understanding of how to develop and apply new techniques to analyze social issues using data from a variety of different sources. It is designed for graduate students who are seeking a stronger foundation in data analytics and scoping questions that can be answered with available data. The online textbook provides more information.

Learning Assessment Table

Program Competencies	**Corresponding Course Learning Objective**	**Corresponding Assignment Title**
Foundations of Data Science	Developing and scoping a research question Developing a Theory of Change and Evaluation	Group project and team paper Video responses
Data Exploration, Management, and Curation	Making use of different types of data Managing and structuring heterogeneous data Sharing and documenting data decisions	Data exploration notebook Video responses
Measurement	Combining and linking data from different sources Creating longitudinal cohorts from cross sections Constructing thoughtful and robust measures	Measurement memo Video responses
Analysis and Inference	Text Analysis Supervised and unsupervised machine learning Basics of Evaluation and different metrics Testing for bias	Text Analysis Notebook Machine Learning memo Video responses
Privacy, Confidentiality and Ethics	Principles of confidentiality Application to federal, state and local data Current approaches and challenges	Final project Video responses

Housekeeping

- The NYU Brightspace site for this course will contain the lecture slides, additional reading materials, and assignments. In addition, all lectures are recorded and available on NYU Brightspace after each session. Notifications and updates will be sent out through NYU Brightspace on a regular basis.
- You are expected to attend classes in person and use the lab time to work on your class project. We expect you to meet 1-2 hours outside the class for your group projects.
- Punctuality is **very important**. We realize unforeseen circumstances arise, but please try to be on time. Disruptions affect not only us, but your fellow classmates as well. Please notify the Professor Lane in advance if you will be late or unable to attend.
- Active participation is a part of your overall grade. This means asking questions in chat, responding to questions when asked, posting in the forum, helping classmates by sharing code snippets and helping them to debug code, sharing information you come across that might be interesting for your classmates.
- We expect you to be prepared for class discussions and to keep up with what we have done in prior classes. The open exchange of ideas will be respected by all students. Respectful and inclusive discussion is required.
- Grades on assignments and class projects are non-negotiable.

- Late assignments are accepted. If you submit an assignment after the posted deadline, it will be counted as late and will be penalized (see evaluation section). You can always turn an assignment in early to avoid penalties. There are no make-up assignments.

Readings

This is a graduate course, so we assume that you have the self-motivation and discipline to keep up with the readings on your own. The course is mainly based on one textbook. However, the syllabus provides reference to additional readings, and you will be pointed to more readings during lectures. For each of the sessions the required readings are different chapters outlined in the syllabus of the following book:

Big Data and Social Science: A practical guide to models and tools, 2nd edition, Taylor Francis 2020, Ian Foster, Rayid Ghani, Ron Jarmin, Frauke Kreuter and Julia Lane

Democratizing Our Data: A Manifesto. MIT Press. 2020, Julia Lane

Course Structure

The course will be structured in weekly sessions. The sessions will consist of lectures and applied work. The applied work will range from coding practice to working on group projects. The time can also be used to ask questions, or discuss specific interests or problem sets in more detail with the instructors. The calendar below is not set in stone and is subject to change. Readings should be completed prior to class on the day of the assigned reading. Additional resources can be found on NYU Brightspace.

Prereading ahead of Week 1: Chapter 1 of Democratizing our Data: The Problem, Why it Matters, and What to Do is good context to “why data and evidence

Session 1: Introduction to class work, structure and research projects

- Date: 01/25/2022
- Lecture:
    - Organizational details for class/housekeeping
    - Developing a research question
    - Basics of Theory of Change and Evaluation
- Readings
    - Textbook Chapter 1
    - Lane, J. (2010). Let's make science metrics more scientific. Nature, 464(7288), 488-489.
    - Romer, P. What It Takes To Be a Leader in Both Basic Science and Technological Progress. https://paulromer.net/statement-for-house-budget-comittee/
    - Impact Evaluation in Practice https://www.worldbank.org/en/programs/sief-trust-fund/publication/impact-evaluation-in-practice
- Application
    - Setting up Jupyter + starter notebook (connect to the class folder and read in the data)

- Assignment 1 release: Project scoping memo

Prereading: Ahead of Week 2: *Chapter 4 of Democratizing our Data: A Successful Model*

Session 2: Developing an empirical research question

- Date: 02/01/2022
- Video view https://www.youtube.com/watch?v=_FLJbQwpjFc and ask two questions
- Lecture:
  - The Challenge of Evidence-based Policymaking
  - Scoping a question
  - Finding data sources
  - Rescoping
- Readings:
  - Textbook Chapter 2 and Chapter 4
  - Commission on Evidence-based Policymaking Final Report https://bipartisanpolicy.org/commission-evidence-based-policymaking/
  - Kreuter, Frauke, Rayid Ghani, and Julia Lane. "Change through data: A data analytics training program for government employees." Harvard Data Science Review 1.2 (2019): 1-26.
- Assignment 2 (part 1) release: Data Exploration Notebook (part 1) – Federal Reporter

Session 3: Data exploration and management

- Date: 02/08/2022
- Lecture:
  - Understanding the data generating process
  - Different data collection methods (aside from traditional surveys, APIs and webscraping)
  - Data types and structures
  - Relational databases and schemas
- Readings
  - Textbook Chapter 4
  - Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, and Abe Usher. "Big data in survey research: AAPOR task force report." Public Opinion Quarterly 79, no. 4 (2015): 839-880.
- Assignment 2 (part 2) release: Notebook: Data Exploration (part 2) – Patentsview and API
- Assignment 1 due

Session 4 : Record Linkage

- Date: 02/15/2022
- View Linkage Video and answer questions
- Lecture: Understanding the Problem
  - The conceptual framework
  - Deterministic approaches
  - Probabilistic approaches
  - Bias and ethics issues

- Readings:
    - Textbook Chapter 3
    - Chang, Garner, Owen-Smith, Weinberg A. Linked Data Mosaic or Policy-Relevant Research on Science and Innovation: Value, Transparency, Rigor, and Community https://coleridgeinitiative.org/wp-content/uploads/2021/10/HDSR_datamosaic_for_submission_15Sept21BW-1.pdf
- For Review: Record Linkage Notebook

Ahead of Week 5: *Chapter 2 of Democratizing our Data: The Current State of Play*

Session 5: Measurement

- Date: 02/22/2022
- Video https://www.youtube.com/watch?v=BeqdB48TU6I and ask two questions
- Lecture:
    - Defining analytical datasets
    - Defining input measures
    - Defining output measures
- Readings:
    - Hall, B. H., & Harhoff, D. (2012). Recent research on the economics of patents. Annu. Rev. Econ., 4(1), 541-565.
    - https://www.callingbullshit.org/
        - https://www.callingbullshit.org/case_studies/case_study_rule_of_21_part_1.html
        - https://www.callingbullshit.org/case_studies/case_study_rule_of_21_part_2.html
- Additional readings
    - Card, David. "Origins of the unemployment rate: the lasting legacy of measurement without theory." American Economic Review 101.3 (2011): 552-57.
    - https://www.scientificamerican.com/article/the-u-s-needs-a-national-data-service/
    - Lane, J. (2020). After Covid-19, the US statistical system needs to change. Significance, 17(4), 42-43. https://rss.onlinelibrary.wiley.com/doi/full/10.1111/1740-9713.01428
- Assignment 3 release: Measurement Memo
- Assignment 2 due

Session 6: Text Analysis

- Date: 03/01/2022
- Video: View text analysis video and answer questions
- Lecture:
    - Conceptual framework
    - Introduction to text analysis: Information retrieval, clustering and text categorization, text summarization
    - Learn how to implement topic modeling
    - Application to scientific fields
    - Evaluation
- Readings:
    - Chapter 8 of textbook

- Assignment 4 release: Text Analysis Notebook

Session 7: Visualization

- Date: 03/08/2022
- Video: View Visualization video and answer questions
- Lecture:
  - Basics of visualization
  - Examples of successful visualizations
  - Applications (two notable uses for visualization: data exploration, presentation)
- Readings:
  - Chapter 6 of textbook
  - Tufte and the Challenger Disaster http://williamwolff.org/wp-content/uploads/2013/01/tufte-challenger-1997.pdf
  - The Healing power of data https://theconversation.com/the-healing-power-of-data-florence-nightingales-true-legacy-134649
- For review: Visualization Notebook
- Assignment 3 due

03/15/2022 Spring Break: No classes

Session 8: Midterm project presentations

- Date: 03/22/2022
  - Students present current stage of their project
  - Students provide feedback on projects
- Readings: no readings

Session 9: Applications

- Date: week of 03/29/2022
- Attend one public meeting of either the National AI Research Resources Task force (https://www.ai.gov/nairrtf/) or the Advisory Committee on Data for Evidence Building (https://www.bea.gov/evidence)
- Assignment 5 release: Summary of the meeting and your key takeaways
- Assignment 4 due

Session 10: Machine Learning Models I

- Date: 04/05/2022
- Videos: View machine learning videos (Introduction to ML for public policy and Training and Testing Models) and answer questions
- Lecture:
  - Formulate research questions in a machine learning framework: from transformation of raw data to feeding them into a model
  - How to build, evaluate, compare, and select models
  - How to reasonably and accurately interpret models
- Application
  - Summarize and submit the results of your group meeting, together with any questions about the project or the class, using a Google Form

- Readings:
  - Chapter 7 of textbook
  - Occupational Classifications: A Machine Learning Approach, Akina Ikudo, Joe Staudt, Julia Lane and Bruce Weinberg *Journal of Economic and Social Measurement*, 2019
- Assignment 5 due

Session 11: Machine Learning Models II

- Date: 04/12/2022
- Video: View machine learning videos (Classification with Decision Trees, Evaluating Models, Clustering) and answer questions
- Lecture:
  - Supervised Machine Learning
  - Assessing model fit
- Readings:
  - Athey, Susan. "Machine learning and causal inference for policy evaluation." Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015.
  - Athey, S. (2019). 21. The Impact of Machine Learning on Economics. In The economics of artificial intelligence (pp. 507-552). University of Chicago Press. https://www.nber.org/system/files/chapters/c14009/c14009.pdf
  - https://www.brookings.edu/techstream/the-tensions-between-explainable-ai-and-good-public-policy/
  - Can an algorithm tell when kids are in danger? https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html

- Assignment 6 release: Machine Learning memo
- For review: Machine Learning Notebook

Session 12: Biases, Fairness, and Inference

- Date: 04/19/2022
- Video: View Inference video and answer questions
- Lecture:
  - Address biases in machine learning techniques and their consequences for public policy
  - How to deal with inference and the errors associated with big data
  - Problems of Big data and the errors resulting from it
- Readings:
  - Implicit bias test
  - Chapter 10 and 11 of textbook
  - Paul D Allison. Missing Data, volume 136. Sage Publications, 2001
  - Paul P Biemer. Total survey error: Design, implementation, and evaluation. Public Opinion Quarterly, 74(5):817–848, 2010
  - O'Neil, Cathy. On Being a Data Skeptic, Sebastopol, CA: O'Reilly Media, 2013.

        - Crawford, Kate. "The Hidden Biases in Big Data." Harvard Business Review, April 1, 2013.
    - Additional resource:
        - Dealing with Bias and Fairness in Data Science Systems: A Practical Hands-on Tutorial

Session 13: Privacy, Confidentiality, and Ethics in Research

- Date: 04/26/2022
- Video: View Privacy and Confidentiality video and answer questions
- Lecture:
    - Recognize where and understand why ethical and confidentiality issues can arise when applying analytics to policy problems
    - Plan, execute, and evaluate a research project along privacy concerns and ethical obligations
- Readings:
    - Chapter 12 of textbook
    - Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2014). Privacy, big data and the public good: Frameworks for engagement. Cambridge University Press.
    - Boyd, Danah, and Kate Crawford. "Critical Questions for Big Data." Information, Communication & Society 15, no. 5 (June 2012): 662–679. doi:10.1080/1369118X.2012.678878
- Assignment 6 due

Session 14: Final Project Presentations

- Date: 05/03/2022
    - Students present their final project

Evaluation

Project work

During the class students will work on their own small class research project during the entire semester. The goal of the research project is for students to develop and apply the techniques taught in the class. Groups are expected to summarize the results of their meetings each week, together with any questions about the project or the class using a Google Form. There will be a midterm presentation and final presentation of the project results. At the end of the semester each team will be required to submit a short research paper documenting their project work.

At the end of the semester, each group member will be asked to rate the contribution of the other group members on a scale of 1 to 10. The group grade will be allocated to each individual based on the rating of the other team members.

Assignments

You are required to complete 5 assignments throughout the class. The assignments constitute 25% of the grade: Please submit your assignments on time. Assignments up to 24 hours late will have their grade reduced by 25%; assignments up to one week late will have their grade reduced by 50%. After one week, late assignments will receive no credit. Please turn in your assignment early if there is any uncertainty about your ability to turn it in on time. You are expected to use Python throughout the entire class.

Class preparation and participation

We have prepared videos for almost each class, which you are expected to view prior to the class. There are a set of questions with each video: responses to the videos will constitute 20% of the grade.

.

Active participation in class will constitute 15% of the final grade (examples: participation in class and group discussions, posting on NYU Classes forum, responding to questions when asked, helping classmates by sharing code snippets and helping them to debug code, sharing information you come across that might be interesting for your classmates).

The breakdown of the evaluation activities:

Activity	Proportion of Grade
Project work	**40%**
Group discussion	10%
Midterm presentation	10%
Final presentation	10%
Research memo (10 pages)	10%
Assignments	**30%**
Assignment 1: Project scoping memo	5%
Assignment 2: Data Exploration Notebook (choose one)	5%
Assignment 3: Measurement memo	5%
Assignment 4: Text Analysis Notebook	5%
Assignment 5: Meeting memo	5%
Assignment 6: Machine Learning memo	5%
Class preparation and participation	**30%**
Responses to class videos	15%
Active class participation	15%

General guidance

The statistical package used to work on the assignments and project work is Python. All project and individual assignments should be posted on NYU Classes before the deadline. Answers to the assignments should be well thought out and communicated precisely, as if reporting to your boss, client, or potential funding source. Avoid sloppy language, poor diagrams, irrelevant discussion, and irrelevant program output.

If you prepare and participate in the course you should be able to work on the assignments without major problems. But we all experience problems that we can’t figure out right away. If you get stuck on something while preparing for class or working on the assignments, spend some time Googling to try to find the answer. If you seem to be moving forward, keep going. That search and discovery method will pay off, both in terms of the direct learning about how to do what you need to do, and also in terms of your learning how to find such things out. (if you don’t know what Stackoverflow is, you will learn!).

However, in order to limit frustrations with class work we advise you to start your assignments early enough that if you experience problems without finding an answer, you still have enough time to ask about it.

If you are stuck after 30 minutes, just stop and ask your classmates or post on the forum on NYU classes. All class participants have access to this and can help you with your questions. You will most likely encounter the same problems as your peers. The forum is there for you to ask your peers for advice. If you don't find a solution, escalate it to the instructors.

Plagiarism

All students must produce original work. Outside sources are to be properly referenced and/or quoted. Lifting copy from websites or other sources and trying to pass it off as your original words constitutes plagiarism. Such cases can lead to academic dismissal from the university.

Academic Integrity

Academic integrity is a vital component of Wagner and NYU. All students enrolled in this class are required to read and abide by Wagner's Academic Code. All Wagner students have already read and signed the Wagner Academic Oath. Plagiarism of any form will not be tolerated and students in this class are expected to report violations to me. If any student in this class is unsure about what is expected of you and how to abide by the academic code, you should consult with me.

Henry and Lucy Moses Center for Students with Disabilities at NYU

Academic accommodations are available for students with disabilities. Please visit the Moses Center for Students with Disabilities (CSD) website and click on the Reasonable Accommodations and How to Register tab or call or email CSD at (212-998-4980 or mosescsd@nyu.edu) for information. Students who are requesting academic accommodations are strongly advised to reach out to the Moses Center as early as possible in the semester for assistance.

NYU's Calendar Policy on Religious Holidays

NYU's Calendar Policy on Religious Holidays states that members of any religious group may, without penalty, absent themselves from classes when required in compliance with their religious obligations. Please notify me in advance of religious holidays that might coincide with exams to schedule mutually acceptable alternatives.