NYU | ROBERT F. WAGNER GRADUATE SCHOOL OF PUBLIC SERVICE

PADM-GP 2505 Big Data Analytics for Public Policy

Spring 2020

Instructor Information

- Julia Lane (pronouns: she/her)
- Email: jil4@nyu.edu
- Office Address: NYU-Wagner, Room 3038, 295 Lafayette Street, New York, NY 10012-9604
- Office Hours: In office and virtual office hours can be made by appointment

- Daniela Hochfellner (pronouns: she/her)
- Email: daniela.hochfellner@nyu.edu
- Office Address: NYU-Wagner, Room 3038, 295 Lafayette Street, New York, NY 10012-9604
- Office hours:
  - Virtual Monday 4pm-6pm and Wednesday 4pm-6pm: I will answer messages you send me via this **form**; I will not be in my physical office during my virtual office hours; I might be on skype during that time as well, my skype name is: frau_nilsson
  - In office Friday 2pm-4pm: You can come and see me in my office anytime during that period. In office hours can be made by appointment if you cannot make the listed hours.

Course Time and Location

- Lecture: Fridays, 9:00 – 10:40am, 60 5th Avenue, Room: 125
- Lab: Fridays, 11:00 – 12:00pm, 7 East 12th Street, Room: 121 (SB)

Course Description and Learning Objectives

The goal of the Big Data Analytics class is to develop the key data analytics skill sets necessary to harness the wealth of newly-available data. Its design offers hands-on training in the context of real microdata. The main learning objectives are to apply new techniques to analyze social problems using and combining large quantities of heterogeneous data from a variety of different sources. The course will explain through lectures and real-world examples the fundamental principles, uses, and appropriate technical details of machine learning, data mining and data science. It is designed for graduate students who are seeking a stronger foundation in data analytics and want to understand the fundamental concepts and applications of data science. After taking this course you will be able to

- Evaluate which data are appropriate to a given research question and statistical need.
- Identify the different data quality frameworks and apply them to public policy problems.
- Use a broad array of basic computational skills required for data analytics, typically not taught in social science, economics, statistics or survey courses.

Learning Assessment Table

Program Competencies or Program Learning Objectives	**Corresponding Course Learning Objective**	**Corresponding Assignment Title (Memo, Team Paper, Exam, etc.)**
Foundations of Data Science	The social science of measurement, formulating research questions, basics of program evaluation, differentiating data sources, "Big Data" - definitions, technical issues, quality frameworks and varying needs, introduction to the data that will be used in this class, case studies, introduction to Python, working with Jupyter notebooks, exploring data visually.	Assignment 1 Midterm Presentation
Data Management and Curation	Introduction to APIs, introduction to characteristics of large databases, building datasets to be linked, linkage in the context of big data, create a big data work flow, data hygiene: curation and documentation.	Assignment 2, Assignment 3, Metadata and code documentation
Data Analysis in Public Policy	What is machine learning, examples, process and methods, fundamentals of record linkage techniques, directed and undirected graphs, different text analytics paradigms, discovering topics and themes in large quantities of text data, mapping your data.	Assignment 4 Final Presentation, Research Memo
Presentation, Inference, and Ethics	Using graphics packages for data visualization, error sources specific to found (big) data, examples of big data analysis and erroneous inferences, inference in the big data context, big data and privacy, legal framework, statistical framework, disclosure control techniques, ethical issues, practical approaches	Assignment 5

Housekeeping

- The NYU Classes site for this course will contain the lecture slides, additional reading materials, and assignments. In addition, all lectures are recorded and available on NYU classes after each session. Labs might or might not be recorded depending on the topic of each lecture. Notifications and updates will be sent out through NYU classes on a regular basis.

- We do encourage you to attend all classes in person and use the practice time given to you to work on your class project. However, we assume that you are capable of setting your own priorities for your educational career and won't penalize students who decide to watch the class recordings and study at home instead of coming to classes in person.
- 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.
- Laptops, cell phones, pda's, I-pods, etc shall be turned off during the lecture. Laptops are not acceptable for the purpose of taking notes. The devices must not be in your view.
- Active participation is a part of your overall grade. We think about active participation as: participate in discussions (in class or on the NYU class forum), share your opinion, ask questions, help classmates in case they struggle (share code snippets and help debug code), share information you come across that might be interesting for your classmates as well, make use of office hours to discuss your progress in class.
- 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.
- Classroom behavior should not interfere with the instructor's ability to conduct the class or the ability of other students to learn from the instructional program. Unacceptable or disruptive behavior will not be tolerated. Students who disrupt the learning environment may be asked to leave class and may be subject to judicial, academic or other penalties. The instructor shall have full discretion over what behavior is appropriate/inappropriate in the classroom.
- 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 section Evaluation). If you have a game, extracurricular activity, have to work, a wedding to go to, etc on the day an assignment is due than you may turn the assignment in early to avoid penalties.
- Make-up assignments will only be allowed under exceptional circumstances. Having another assignment that day is not an exceptional circumstance.

Required 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, Taylor Francis 2016, Ian Foster, Rayid Ghani, Ron Jarmin, Frauke Kreuter and Julia Lane (https://coleridge-initiative.github.io/big-data-and-social-science/)
- Federal Data Strategy – Action Plan 2020 (https://strategy.data.gov/action-plan/)
- Kreuter, F., Ghani, R., & Lane, J. (2019). Change Through Data: A Data Analytics Training Program for Government Employees. Harvard Data Science Review, 1(2). (https://doi.org/10.1162/99608f92.ed353ae3)

Course Structure

The course will be structured in weekly sessions. Usually, but not always, each session will be followed by lab time. The sessions will consist of lectures and computing exercises, the required lab will give you time to work on practicing coding, on your assignments or class project, ask questions, or discuss specific interests or problem sets in more detail with the instructors. Lecture and lab time is combined for topics that require a longer uninterrupted period of time. 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 classes.

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

- Date: 01/31/2020
- Lecture:
  - Organizational details for class/housekeeping
  - How to define and scope an empirical research project
  - Example study: New linked data on research investments: Scientific workforce, productivity, and public value, Lane, Owen Smith, Rosen and Weinberg, Research Policy Volume 44, Issue 9, November 2015, Pages 1659-1671
- Lab:
  - Set up the computing space
- Required Readings: none

Session 2: Big Data and Policy Research

- Date: 02/07/2020
- Lecture:
  - Common data source that are used in Policy Research
  - Advantages and disadvantages of big data vs. classical survey data and administrative data
- Lab:
  - Getting to know the data being used in class
  - Project Scoping: Think about your class project
- Readings:
  - Textbook chapter 1
  - Measuring the impact of R&D Spending
  - Watching the players, not the scoreboard
  - Wrapping it up in a person, Examining the earnings and employment outcomes for PhD recipients
  - https://www.uspto.gov/about-us/news-updates/new-uspto-tool-allows-exploration-40-years-patent-data
- Assignment 1 posted after class

Session 3: Introduction to Python and Jupyter

- Date: 02/14/2020
- Lecture:
  - What is Python and Jupyter?
  - Learn to code: variables, data structures – lists and maps, logic – if then else and loops, functions – calling and writing
- Lab:

    - Coding practice
    - Investigate class data: What data do you need for your research project?
  - Readings:
    - Python for Economists
    - Online tutorials:
  - More Resources for Python/Pandas (not required as readings):
    - Introduction to Python for Econometrics, Statistics and Data Analysis by Kevin Sheppard (free)
    - Python: 1-pager from DataCamp & longer version of general Python notes
    - Pandas
    - Software Carpentry
    - Python Tutorial
    - Wes McKinney, Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2012, pp. 466
  - Assignment 2 posted after class

Session 4: Record Linkage I

  - Date: 02/21/2020
  - Lecture/Lab combo: Record linkage and preprocessing
  - Readings:
    - Chapter 3 of textbook
    - Hernández MA, Stolfo SS 1998, Real-world data is dirty: data cleansing and the merge/purge problem. Data Mining and Knowledge Discovery 2(1), 9-73
    - https://www.beehive.govt.nz/release/final-data-set-enhances-risk-youth-profile
    - https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html

  - More Resources for record linkage (not required as readings):
    - Ivan P. Fellegi and Alan B. Sunter, A Theory for RecordLinkage, Journal of The American Statistical Association Vol. 64, Iss. 328,1969
    - Record linkage by Herzog, Scheuren and Winkler
    - Dunn, H.L. (1946). “Record Linkage”. American Journal of Public Health, 36(12),1412-1416
    - Winkler WE 2009. Record linkage. D Pfeffermann and CR Rao (Hg.) Handbook of Statistics 29A, Sample Surveys: Design, Methods and Applications Amsterdam: Elsevier
    - Gill LE 2001. Methods for Automatic Record Matching and Linkage and Their Use in National Statistics. Norwich: Office of National Statistics

Session 5: Record Linkage II

  - Date: 02/28/2020
  - Lecture/Lab combo: Record linkage with Python
  - Readings:

        - Record linkage: link to manual
        - Regex: link to PDF
        - Python regular expressions
        - Online regular expression tester
    - Submission of Completed Assignment 1 and 2

Session 6: Application Programming Interface (API)

- Date: 03/06/2020
- Lecture:
    - What is an API and how is it being used in policy research?
    - What is json?
- Lab:
    - Making raw HTTP API requests, Using pre-packaged API client libraries
- Readings:
    - Chapter 2 of textbook
    - Python's requests & Beautiful Soup libraries (for web scraping & APIs)
    - Patent API
- Assignment 3 posted after class

Session 7: Text Analysis and Topic Modeling

- Date: 03/13/2020
- Lecture:
    - Introduction in text analysis: Information retrieval, clustering and text categorization, text summarization
    - Learn how to implement topic modeling
- Lab:
    - Text Analysis in Python
- Readings:
    - Chapter 7 of textbook
    - Identifying Food Safety related Research, Julia Lane and Evgeny Klochikhin in *Measuring the Economic Value of Research: The Case of Food Safety,* Kaye Husbands Fealing, Julia Lane, John King, Stanley Johnson Eds, Cambridge University Press, 2018
- Submit completed assignment 3

SPRING BREAK – NO CLASS

Session 8: Midterm project presentations

- Date: 03/27/2020
- Lecture/lab combo:
    - Students present current stage of their project
    - Students provide feedback on projects
- Readings: no readings

Session 9: Machine Learning Models I

- Date: 04/03/2020
- Lecture/lab combo:
    - Formulation 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
- Readings:
    - Chapter 6, textbook
    - James, G., Witten, D., Hastie, T., Tibshirani, R. An Introduction to Statistical Learning. Springer, 2013.
    - Xindong Wu et al. (2008). Top 10 algorithms in data mining. Knowl Inf Syst (2008) 14:1–37
    - Occupational Classifications: A Machine Learning Approach, Akina Ikudo, Joe Staudt, Julia Lane and Bruce Weinberg *Journal of Economic and Social Measurement*, 2019
- More Resources for machine learning (not required as readings):
    - Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer, 2009.
- Assignment 4 posted after class

Session 10: Machine Learning Models II

- Date: 04/10/2020
- Lecture:
    - Examples of ML models in Python
    - Assessing model fit
- Lab:
    - Develop ML model for your class project I
- Readings:
    - ML in Python - Cheatsheet

Session 11: Biases, Fairness, and Inference

- Date: 04/17/2020
- 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
- Lab:
    - Develop ML model in your class project II
- Readings:
    - Implicit bias test
    - Chapter 10 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.

Session 12: Visualizations with Big Data

- Date: 04/24/2020
- Lecture:
    - Theory of information visualization
    - Choosing a chart type and color considerations
    - Labeling and information overload
- Lab:
    - Visualizing analytical results with Python, exercises
- Readings:
    - Chapter 9 of textbook
- Submit completed assignment 4
- Assignment 5 posted after class

Session 13: Privacy, Confidentiality, and Ethics in Research

- Date: 05/01/2020
- 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
- Lab:
    - Project work
- Readings:
    - Chapter 11 of textbook
    - Karr, A., & Reiter, J. P. (2014). Analytical Frameworks for Data Release: A Statistical View. In J. Lane, V. Stodden, H. Nissenbaum, & S. Bender (Eds.), Privacy, Big Data, and the Public Good: Frameworks for Engagement. Cambridge University Press.
    - 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
- Submit completed assignment 5

Session 14: Final Project Presentations

- Date: 05/08/2020
- Lecture/lab combo:
    - Students present their final project
    - Students provide feedback on projects
- Readings: no readings

Evaluation

During the class students will work on their own small class research project during the entire semester. There will be a midterm presentation and final presentation of the project results. At the end of the semester each student has to submit a short research memo documenting their project work. The goal of the research project is to demonstrate the ability to use the techniques learned over time according to academic principles. The project work will constitute 30% of the grade:

- Midterm presentation 10%
- Final presentation 10%
- Research memo (5 pages) 10%

In addition, there will be 5 assignments throughout the class. The assignments constitute 50% of the grade:

- Assignment 1: Project scoping and research agenda 10%
- Assignment 2: Data selection for class project 10%
- Assignment 3: API Python exercises 10%
- Assignment 4: Machine Learning Model for class project 10%
- Assignment 5: Visualization exercise 10%

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 experience problems without finding an answer, you still have enough time to ask about it. If you feel like you have not moved forward after 30 minutes of being stuck, 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 us. During virtual office hours you can submit questions using this form: https://docs.google.com/forms/d/e/1FAIpQLSdgBqi4t-3fGnssZMBXfFIjR68ViccxVkXWC1TOJfaMMFNb_Q/viewform. You can also come to regular office hours on Friday.

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. Metadata documentation of all your code will be 10% of the grade.

Last but not least, active participation in class will constitute 10% of the final grade. We think about active participation as: participate in discussions (in class or on the NYU class forum), share your opinion, ask questions, help classmates in case they struggle (share code snippets and help debug code), share information you come across that might be interesting for your classmates as well, make use of office hours to discuss your progress in class.

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.