****

**PADM-GP 2505**

Big Data - Tools and Techniques for Public Policy Practitioners

Spring 2025

## Instructor Information

* Carlos Quirola
* Email: carlos.quirola@nyu.edu
* Office Hours: Tuesdays, 6-7pm, Wagner Building, Office 312A
  + *Subject to change*

## Course Information

* Lecture: Wednesdays, 6:45 PM - 8:25 PM
* Lab: Wednesdays, 8:35 PM - 9:35 PM
* Location: 105 E 17th Street, Room 117

## Course Description

The goal of this course is to equip students with the skills to answer a policy question using data and to leave students with a deliverable that showcases their data skills. We will walk through the steps and develop the skills needed to achieve this goal including: scoping and finding data; gathering and enriching data; modelling and visualizing data; ethical and privacy considerations, and preparation of deliverables.

In support of the above, the course’s content will give students exposure to the various sources for data gathering (APIs, SFTPs), and help them learn how to access those sources using a programming language like Python. In addition, the course’s content will cover techniques to clean data and will illustrate the application of quantitative methods that students have learned in other courses. The course will also provide students with practical items to consider when working with data (such as privacy and ethics best practices), as well as an overview of the most recent and important regulations that govern the use of data. Finally, the course will cover how to write up and present findings in an effective, audience conscious manner.

## Course Structure

The course will be structured in weekly sessions. The sessions will consist of interactive lectures with some time also devoted to class projects and will be supplemented with a weekly lab directly following each class session. The lectures will cover both general theory and specific applications, and the labs will expand on the lectures by discussing relevant coding techniques to implement these methods on the class data in Python. Labs will typically focus on the coding materials based on concepts introduced in the previous week’s lecture.

The class schedule is tentative and is subject to change. Readings and interactive coding materials are expected to be completed before class on the day of the assignment. Additional resources can be found on NYU Brightspace.

While the course will use three (3) problem sets to evaluate students, students will also be expected to work on a semester-long data project that will involve all facets covered in the course. For the semester-long project, students will work with the instructor to build a dataset that can answer a policy question, choosing from a menu of API-based data options that include the Census, NYC Open Data, St. Louis Federal Reserve, US Department of Education Higher Education Institutions, amongst others.

## Readings

The syllabus provides the relevant chapters from the textbook and additional readings for each week. Students may be pointed to more readings during lectures. This course's textbook is available at the NYU Library through the hyperlink below.

[Big Data and Social Science: A practical guide to models and tools](https://ebookcentral.proquest.com/lib/nyulibrary-ebooks/detail.action?docID=4689171), 2nd edition, Taylor Francis 2020, Ian Foster, Rayid Ghani, Ron Jarmin, Frauke Kreuter and Julia Lane.

## Housekeeping

* The NYU Brightspace site for this course will contain the lecture slides, additional reading materials, and assignments. Since instruction will take place in person, lectures will not be recorded. Slides will be available on Brightspace after class for those who wish to review content. Notifications and updates will be sent out through NYU Brightspace regularly.
* Students are expected to attend classes in person and use the lab time to apply the class concepts to their projects. **Students should plan to devote 1-2 hours a week outside of class to developing their class project.**
* Students are not permitted to take notes via laptop or phone. Laptops and phones may only be used for participating in class polls.
* Punctuality is **very important**. Unforeseen circumstances may arise, but all students are generally expected to be on time. Students should notify Professor Feder in advance if they will be late or unable to attend.
* Students are expected to be prepared for class discussions and to keep up with prior class content. The open exchange of ideas will be respected by all. Respectful and inclusive discussion is required.
* Grades are non-negotiable.
* Late submissions are accepted but are counted as late and will be penalized as specified in the Evaluation section. Students can always turn in an assignment early to avoid penalties. There are no make-up assignments.

## Projected Class Schedule and Assignments

*The schedule listed below is subject to change and should only be used as a reference. The most up-to-date information will be available on NYU Brightspace.*

*Data Journey Part 1: Scoping the right project and finding the right data*

**Session 1: Introduction to class work, structure, and the big data dilemma**

* Date: 01/22/2025
* Lecture:
  + Organizational details for class/housekeeping and assignments
  + Tech vs Research: The Big Data Dilemma and Opportunities for Public Policy
  + Data Project Scoping
* Readings:
  + More Data for Africa: <https://www.economist.com/middle-east-and-africa/2024/11/14/to-get-more-capital-africa-needs-more-data>
  + Peru’s Bad Drivers: A Data Gold Mine: <https://www.economist.com/the-americas/2024/08/15/perus-crazy-drivers-offer-a-data-deluge-for-self-driving-cars>
  + How NOT to endanger people with data: https://www.economist.com/britain/2023/08/09/a-big-data-breach-endangers-police-in-northern-ireland
* Lab:
  + Python refresher
  + Visual Studio Code
  + Final Project Overview

**Session 2: The right data and the right questions to ask**

* Date: 01/29/2025
* Lecture:
  + Scoping 101
  + Framing the right question / data objective
  + Types of Data and their attributes (traditional, survey, unstructured)
  + How and where can we find data?
  + Re-scoping
* Readings:
  + Textbook Chapter 1
  + Data Science and Public Policy Lab at Carnegie Mellon University. (2021, November 3). Data Science Project Scoping Guide. <http://www.datasciencepublicpolicy.org/our-work/tools-guides/data-science-project-scoping-guide/>
  + Lane, J. (2010). Let's make science metrics more scientific. Nature, 464(7288), 488-489.
  + Lane, Julia. “A Vision for Democratizing Government Data.” *Issues in Science and Technology* 39, no. 1 (Fall 2022): 84–88. <https://issues.org/democratizing-government-data-lane/>
* Lab:
  + Review Assignment 1 Instructions
  + Scoping Activity
  + Menu of Datasets

**Session 3: Data exploration and management**

* Date: 02/05/2025
* Lecture:
  + Data discovery and exploratory data analysis techniques
  + Understanding the data-generating process
  + Different data gathering methods (traditional surveys, APIs, web scraping, SFTPs, S3 buckets and more)
  + Data types and structures
  + Relational databases and schemas
* Readings:
  + Behrens, J. T., & Yu, C. H. (2003). Exploratory data analysis. *Handbook of psychology*, *2*, 33-64.
  + Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). Exploratory data analysis. In *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. essay, O’Reilly.   
    Chapter 10: <https://r4ds.hadley.nz/eda>
  + Textbook Chapters 2 + 4
  + Python for Economists: https://scholar.harvard.edu/files/ambell/files/python\_for\_economists.pdf
* Lab:
  + Using Python to explore a dataset: techniques for different data types
  + Answer questions about Assignment 1

*Data Journey Part 2: Data Gathering and Enrichment*

**Session 4: Fetching Data - APIs in a Public Policy Mission**

* Due: Assignment 1
* Date: 02/12/2025
* Lecture:
  + Data Gathering Beyond the Spreadsheet
  + APIs
    - What are APIs?
    - How to access data from an API?
    - Methods, endpoints, rate limits
  + Additional ways to gather information
    - SFTPs
    - S3 Buckets
* Readings:
  + Watch this video: <https://www.youtube.com/watch?v=OVvTv9Hy91Q>
  + Web Scraping with Python: https://blog.hartleybrody.com/web-scraping-cheat-sheet/
* Lab:
  + Working Session: Census API

**Session 5: Record Linkage**

* Date: 02/19/2025
* Lecture:
  + Understanding the Problem
  + The conceptual framework
  + Deterministic approaches
  + Probabilistic approaches
  + Bias and ethics issues
* Readings:
  + Textbook Chapter 3
  + Chang, W.-Y., Garner, M., Basner, J., Weinberg, B., & Owen-Smith, J. (2022). A Linked Data Mosaic for Policy-Relevant Research on Science and Innovation: Value, Transparency, Rigor, and Community. Harvard Data Science Review, 4(2). <https://doi.org/10.1162/99608f92.1e23fb3f>
* Lab:
  + Linking records exercise

**Session 6: Cloud Computing and Data Storage**

* Date: 02/26/2025
* Lecture:
  + What is Cloud Computing and how does it fit within our data journey?
  + What are the main sources of cloud computing? (paid and unpaid)
  + Methods to systematically collect new data and store it safely
* Readings:
  + TBD
* Lab:
  + Google Sheets as a Guerrilla Database

*Data Journey Part 3: Modeling Data, Visualizing and Sharing Results (Code)*

**Session 7: Visualization**

* Due: Proposal Memo
* Date: 03/05/2025
* Lecture:
  + Basics of visualization
  + Examples of successful visualizations
  + Applications (two notable uses for visualization: data exploration, and presentation)
* Readings:
  + Textbook Chapter 6
* Lab:
  + Assignment 2 Overview
  + Visualizing data with Python

**Session 8: Text Analysis and Topic Modeling**

* Date: 03/12/2025
* 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
* Reading:
  + Chapter 8 of textbook
* Lab:
  + Sentiment models

**Session 9: Midterm project presentations**

* Date: 03/19/2025
  + Students present current stage of their project
  + Students provide feedback on projects
  + **Note: No lab**
* Reading: N/A

**03/26/2024 Spring Break: No classes**

**Session 10: Machine Learning Models I**

* Date: 04/02/2025
* 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
* Readings:
  + Chapter 7 of the textbook
  + Hao, J., & Ho, T. K. (2019). Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. Journal of Educational and Behavioral Statistics, 44(3), 348–361. <https://doi.org/10.3102/1076998619832248>
* Lab:
  + TBD
  + Questions on Assignment 2

**Session 11: Machine Learning Models II**

* Date: 04/09/2025
* Lecture:
  + Supervised Machine Learning
  + Assessing model fit
  + 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:
  + Chapter 7 of the textbook
  + Athey, S. (2015, August). Machine learning and causal inference for policy evaluation. In *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 5-6).

<https://caii.ckgsb.com/uploads/life/201901/26/1548495020438310.pdf>

* + Agrawal, A., Gans, J., Goldfarb, A., & Athey, S. (2019). The Impact of Machine Learning on Economics. In *The Economics of Artificial Intelligence: An agenda* (pp. 507–547). essay, The University of Chicago Press.  
    <https://www.nber.org/system/files/chapters/c14009/c14009.pdf>
  + Coyle, D. (2020, September 15). *The tensions between explainable AI and good public policy*. Brookings. <https://www.brookings.edu/articles/the-tensions-between-explainable-ai-and-good-public-policy/>
  + Hurley, D. (2018, January 2). Can an algorithm tell when kids are in danger? *The New York Times Magazine*. <https://www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html>

*Data Journey Part 4: Presenting your Results, Ethics and Existing Regulations*

**Session 12: Deploying your Results or Product**

* Due: Assignment 2
* Date: 04/16/2025
* Lecture:
  + Web Apps
  + git and version control
* Reading:
  + Review one of the presentations in Session 1 or 2 the Value of Science conference hosted by the National Center for Science and Engineering Statistics. <https://coleridgeinitiative.org/value-of-science-data-products-and-use-conference>. You will be asked to comment on what you think was effective and what was not effective.
  + Review the Theory of Change in Appendix I - Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource. (2023, January). *National Artificial Intelligence Research Resource Task Force*. Retrieved from <https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf>.
* Lab:
  + From Translation Services to Leadership: the importance of data communication for different audiences
  + gitHub as a means of collaboration, communication and deployment

**Session 13: Presenting your Results to different audiences**

* Date: 04/23/2025
* Lecture:
  + Understanding audiences
  + Audience-driven communication
  + Effective memos
* Readings:
  + TBD
* Lab: Data Communication Activity

**Session 14: Privacy, Confidentiality, and Ethics for Data**

* Date: 04/30/2025
* 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 the textbook
  + Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (2014). Privacy, big data and the public good: Frameworks for engagement. Cambridge University Press.
* Lab:
  + Activity: is my data compliant with international data privacy standards?
  + Project Work: Final Project Assistance

**Session 15: Final Project Presentations & Memo**

* Date: 05/07/2025
* Lecture: N/A, students present final projects
* **Note: No lab**

## Evaluation

Project work

Each student will work on an individual project throughout the semester. The goal of the final project is for students to independently develop and apply the techniques taught in the class to structure a data project from start to finish. Questions or concerns about individual projects can be discussed via email and office hours.

There will be a midterm presentation and a final presentation of the project, accompanied by scoping, mid-progress and final memos. The final memo should document the project objective, methodology, and results obtained throughout the semester, as well as future thinking. Submitting the final paper up to 24 hours late will result in a 25% grade reduction; between 25 and 48 hours late, a 50% reduction; beyond 48 hours late, no credit.

In addition to the final project, students will submit three (3) problem sets throughout the semester.

Class preparation and participation

Preparation and class participation will constitute 25% of the final grade (examples: participation in class discussions, responses in class polls, engagement in peer midterm and final presentations, posting on NYU Classes forum, responding to questions when asked, helping classmates by sharing code snippets and helping them debug code, sharing information that might be interesting for classmates). Class participation and concise, clear communication is essential for this class. Attendance is mandatory, and unexcused absences will be noted in the participation grade.

|  |  |
| --- | --- |
| **Activity** | **Percentage of Grade** |
| **Project work** | **75%** |
| Assignment 1 | 15% |
| Scoping Memo (Final Project) | 5% |
| Midterm Presentation | 10% |
| Mid-progress Memo (Final Project) | 5% |
| Assignment 2 | 15% |
| Final project and documentation | 15% |
| Final Memo | 10% |
| **Class preparation and participation** | **25%** |
| Class participation | 15% |
| Peer feedback on the midterm and final presentations | 10% |

## Project Guidance

All prepared coding materials—in the form of interactive Jupyter Notebooks—will be written in Python. Students should repurpose as much code as possible from the notebooks for their class projects. Stack Overflow and generative AI tools such as ChatGPT (more on this in the Plagiarism section) can also be useful coding resources. Students are permitted to use additional software tools and programming languages to complete their projects, but no support will be provided for non-Python tools and languages.

## 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 original words constitutes plagiarism. Such cases can lead to academic dismissal from the university, as taking credit for other’s writing is a violation of NYU’s Academic Integrity policy.

The use of generative AI (ChatGPT and related tools) is allowed in this class as a coding support tool. Since part of the individual project evaluation for each activity will focus on creative and critical thinking, using generative AI will likely be counterproductive. In particular, copying generative AI output and submitting it as a final report is prohibited. Plagiarism of any form—including copying generative AI output—will not be tolerated. Students in this class are expected to report violations to Professor Feder.

## 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](https://wagner.nyu.edu/portal/students/policies/code). All Wagner students have already read and signed the [Wagner Academic Oath](https://wagner.nyu.edu/portal/students/policies/academic-oath). Students unsure about expectations and how to abide by the academic code should consult with Professor Feder immediately.

## 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 the “Get Started” button to apply for accommodation. Students can also contact CSD directly (212-998-4980 or mosescsd@nyu.edu) for information. Students 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](https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/university-calendar-policy-on-religious-holidays.html) states that members of any religious group may, without penalty, absent themselves from classes when required in compliance with their religious obligations. Please notify Professor Feder in advance of religious holidays that might coincide with presentations to schedule mutually acceptable alternatives.

## NYU’s Wellness Exchange

[NYU’s Wellness Exchange](http://www.nyu.edu/life/safety-health-wellness/wellness-exchange.html) has extensive student health and mental health resources. A private hotline (212-443-9999) is available 24/7 to connect students with a professional who can help them address day-to-day challenges as well as other health-related concerns.