NYU | ROBERT F. WAGNER GRADUATE SCHOOL OF PUBLIC SERVICE

# PADM-GP 2505 Big Data Analytics for Public Policy

# Spring 2021

## Contact Information


- Julia Lane
- Email: jil4@nyu.edu
- Office Address: NYU-Wagner, Room 3038, 295 Lafayette Street, New York, NY 10012-9604
- Office Hours: Virtual office hours can be made by appointment

- Ekaterina Levitskaya
- Email: el2727@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 (Zoom link)

- Aidan Feldman
- Email: alf9@nyu.edu
- Office Address: NYU-Wagner, Room 3038, 295 Lafayette Street, New York, NY 10012-9604
- Office hours: Virtual 5:30pm-6:30pm on Wednesdays (Zoom link)


## Course Time and Location

- Lecture: Fridays, 9:00 – 10:40am, Via Zoom
- Lab: Fridays, 11:00 – 12:00pm, Via Zoom

## 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<br>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 workflow, data hygiene: curation and documentation. | Assignments 2 and 3 (Data Exploration: Grants and Patents) |
| 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 (Text Analysis)<br>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 (Visualization) |

## 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. Notifications and updates will be sent out through NYU classes on a regular basis.

- You are expected to attend virtual classes in person and use the lab time to work on your class project.
- Active participation is a part of your overall grade. This means asking questions in chat, 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.
- 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.
- Make-up assignments are not available.

# 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, 2nd edition, Taylor Francis 2020, Ian Foster, Rayid Ghani, Ron Jarmin, Frauke Kreuter and Julia Lane

Additional readings of interest are here:

- 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)
- Yarkoni, Tal, Dean Eckles, James Heathers, Margaret Levenstein, Paul Smaldino, and Julia I. Lane. "Enhancing and accelerating social science via automation: Challenges and opportunities." Harvard Data Science Review, forthcoming https://osf.io/preprints/socarxiv/vncwe/

# 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. In sum,

1. Before class, you will:
    1. Complete the prework.

    2. Watch the video.
    3. Submit responses to the form.
2. In the lecture, the instructor will speak to these responses, and bring up other questions for discussion.
3. During the lab, you will work on your project with your group and get help from the lab instructors.

Between classes, you are expected to work on assignments and spend additional time with your group on your project.

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

- Date: 01/29/2021
- 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:
  - Project Template Week 1 exercise
  - Review directions on setting up instruction space
  - Summarize and submit the results of your group meeting, together with any questions about the project or the class, using a Google Form
- Required Readings: none

**Session 2: Big Data and Policy Research**

- Date: 02/05/2021
- Prework: Review “Introduction to Python and Pandas” (short videos and practice notebook)
- Lecture:
  - Common data sources that are used in policy research
  - Advantages and disadvantages of big data vs. classical survey data and administrative data
- Lab:
  - Get to know the data being used in class
  - Project Template Week 2 exercise
  - Summarize and submit the results of your group meeting, together with any questions about the project or the class, using a Google Form
- Readings:
  - Textbook chapter 1
  - Wrapping it up in a person, Examining the earnings and employment outcomes for PhD recipients
  - Introduction to Python for Econometrics, Statistics and Data Analysis by Kevin Sheppard (free)
  - Python: 1-pager from DataCamp & longer version of general Python notes
- Assignment 1 posted after class

**Session 3: Data Exploration and the Data Generation Process**

- Date: 02/12/2021
- Prework: Review Federal Reporter Data Exploration Notebook (Notebook 1)
- Lecture:
  - Understanding inputs: Grant and award data
  - Understanding outputs: Federal Reporter
- Lab:
  - Coding practice
  - Investigate grants data
  - Project Template Week 3 exercise
  - Summarize and submit the results of your group meeting, together with any questions about the project or the class, using a Google Form
- Readings:
  - Python for Economists
  - Online tutorials
- More Resources for Python/Pandas (not required as readings):
  - 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: Evidence-based Policy Making in Practice**

- Date: 02/19/2021

**Attend Advisory Committee meeting: https://www.bea.gov/evidence**
Send an email with a specific question about the Advisory Committee meeting to Dr. Julia Lane by 5pm of the following Thursday.

Submit completed Assignment 1

**Session 5: Record Linkage**

- Date: 02/26/2021
- Prework:
  - View Linkage Video and answer questions
  - Review the Patentsview Data Exploration Notebook (Notebook 2)
- Lecture: Understanding the Problem
  - Data Preprocessing
- Lab:
  - Project Template Week 5 exercise
  - 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 3 of textbook
  - Dinusha Vatsalan, Peter Christen, Vassilios S. Verykios, A taxonomy of privacy-preserving record linkage techniques, Information Systems Volume 38, Issue 6, 2013, https://doi.org/10.1016/j.is.2012.11.005

- More Resources for record linkage (not required as readings):
  - https://www.kdnuggets.com/
  - https://users.cecs.anu.edu.au/~Peter.Christen/publications/christen2019csic-tutorial-slides.pdf
- Assignment 3 posted after class

**Session 6: Visualization**

- Date: 03/05/2021
- Prework:
  - View Visualization video and answer questions
  - Review Visualization Notebook (Notebook 3)
- Lecture:
  - Visualization and public policy
- Lab:
  - Project Template Week 6 exercise
  - 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 6 of textbook
- Submit completed Assignments 2 and 3

**Session 7: Application Programming Interface (API)**

- Date: 03/12/2021
- Prework: https://www.youtube.com/watch?v=OVvTv9Hy91Q
- Lecture:
  - Using an API in policy research (motivation and examples)
- Lab:
  - Project Template Week 7 exercise
  - Making raw HTTP API requests, Using pre-packaged API client libraries
  - 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 2 of textbook
  - Python's requests & Beautiful Soup libraries (for web scraping & APIs)
  - Patent API
  - https://www.coursera.org/lecture/data-collection-processing-python/introduction-rest-apis-xP6Ek

03/19/2021 Spring Break: No classes

**Session 8: Midterm project presentations**

- Date: 03/26/2021
  - Students present current stage of their project
  - Students provide feedback on projects
- Readings: no readings

**Session 9: Text Analysis and Topic Modeling**

- Date: 04/02/2021
- Prework:
  - View Text Analysis video and answer questions

        - View Text Analysis notebook
    - Lecture:
        - Introduction to text analysis: Information retrieval, clustering and text categorization, text summarization
        - Learn how to implement topic modeling
    - Lab:
        - Project Template Week 9 exercise
        - Text Analysis using Python
        - 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 8 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
    - Assignment 4 (Text Analysis) posted after class

**Session 10: Machine Learning Models I**

- Date: 04/09/2021
- Prework:
    - 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
- Lab:
    - Project Template Week 10 exercise
    - 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

**Session 11: Machine Learning Models II**

- Date: 04/16/2021
- Prework:
    - View machine learning videos (Classification with Decision Trees, Evaluating Models, Clustering) and answer questions
- Lecture:
    - Examples of ML models in Python
    - Assessing model fit
- Lab:
    - Project Template Week 11 exercise
    - Summarize and submit the results of your group meeting, together with any questions about the project or the class, using a Google Form

- Readings:
  - ML in Python - Cheatsheet

Assignment 5 (Visualization) posted after class

**Session 12: Biases, Fairness, and Inference**

- Date: 04/23/2020
- Prework:
  - 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
- Lab:
  - Project Template Week 12 exercise
  - Summarize and submit the results of your group meeting, together with any questions about the project or the class, using a Google Form
- 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.
- Submit completed assignment 4

**Session 13: Privacy, Confidentiality, and Ethics in Research**

- Date: 04/30/2021
- Prework:
  - 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
- Lab:
  - Project work
  - 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 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
    - Submit completed assignment 5;

**Session 14: Final Project Presentations**

- Date: 05/07/2021
    - 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 has to submit a short research memo documenting their project work.

The project work will constitute 40% of the grade:

- Group work summary and questions 10%
- Midterm presentation 10%
- Final presentation 10%
- Research memo (5 pages) 10%

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:

- Assignment 1: Project scoping and research agenda 5%
- Assignment 2: Grant data exploration 5%
- Assignment 3: Patent data exploration 5%
- Assignment 4: Text Analysis 5%
- Assignment 5: Visualization 5%

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. All work is expected to be your own.

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 live or via Zoom chat, 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** | **25%** |
| Assignment 1: Project scoping | 5% |
| Assignment 2: Grant data exploration | 5% |
| Assignment 3: Patent data exploration | 5% |
| Assignment 4: Text Analysis | 5% |
| Assignment 5: Visualization | 5% |
| **Class preparation and participation** | **35%** |
| Responses to class videos | 20% |
| Active class participation | 15% |

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 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 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.