NYU | ROBERT F. WAGNER GRADUATE SCHOOL OF PUBLIC SERVICE

PADM-GP 2505 Big Data Analytics for Public Policy

Spring 2019

Instructor Information

- Julia Lane
- Email: jil4@nyu.edu
- Office Address: NYU-Wagner, 295 Lafayette Street, New York, NY 10012-9604
- Office Hours: by appointment.
- Daniela Hochfellner
- Email: daniela.hochfellner@nyu.edu
- Office Address: NYU-Wagner, 295 Lafayette Street, New York, NY 10012-9604
- Office Hours: Fridays 2:00pm - 4:00pm.

Course Information

- Class Meeting Times: Fridays, 9:00 – 10:40 am
- Class Location: GCASL_379
- Lab Meeting Times: Fridays, 11:00 – 12:00 pm
- Lab Location: GCASL_379

Course Description

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.

Course and Learning Objectives

- 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.

- Learn 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, Web scraping exercises, 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 ducumentation
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

We expect you to be prepared for class discussions and to keep up with what we have done in the prior classes. You are expected to attend every class session, to arrive prior to the starting time, to remain for the entire class, and to follow basic classroom etiquette, including (unless otherwise directed) using electronic devices except where necessary to follow along with the lecture or lab. Attendance will only be taken once: at the very beginning of every class.

The NYU Classes site for this course will contain the lecture slide, additional reading materials, and assignments. Furthermore, notifications and updates will be sent out through NYU classes. If you have questions about class material that you do not want to ask in class, or topics that are not directly related to the content covered in class please reach out to us and make an appointment or use the discussion board on NYU classes. The discussion board is the preferred method of asking questions, as others may benefit from the answers being available on NYU Classes. As a corollary to this, please try to answer your classmates questions.

Required Readings

This is a graduate course so we'll 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 is providing reference to additional readings. For each of the Session the required readings are different chapters outlined in the Syllabus of 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

Requirements and Presentation

In addition to lectures this course will include computer exercises. All computations will be done in Python. You don't need to be an expert in coding with Python, however, to get along with the course material basic Python skills are necessary, or comparable experience in R, Stata or SAS (syntax). There is a Intro to Python session, but you should be taking the Intro to Python for Data Science by Data Camp online class (free) before the semester starts if you haven't worked with Python before. Furthermore, this class will teach you empirical methods to work big data. To understand the basic statistical concepts behind the algorithms we will learn this class requires basic knowledge in statistics and multivariate regression modeling.

Course Structure

The course will be structured in weekly sessions, whereas each session is combined with required lab time. The sessions will consist of lectures and computing exercises, the required lab will give you time to work on your assignments, ask questions, or discuss specific interests or problem sets in more detail with the instructors. The class takes place every Friday from 9:00am-10:40am, followed by lab time from 11:00am-12:00pm. Both sessions are mandatory.

- Session 1
    - Date: 02/01/2019
    - Topic: Intro

- Session 2
    - Date: 02/08/2019
    - Topic: Python
    - Assignment Posted: Assignment 1
- Session 3
    - Date: 02/15/2019
    - Topic: Big Data
- Session 4
    - Date: 02/22/2019
    - Topic: APIs
    - Assignment Posted: Assignment 2
    - Assignment Submission: Assignment 1
- Session 5
    - Date: 03/01/2019
    - Topic: Record Linkage
    - Assignment Posted: Assignment 3
- Session 6
    - Date: 03/08/2019
    - Topic: Machine Learning
    - Assignment Submission: Assignment 2
- Session 7
    - Date: 03/15/2019
    - Topic: Machine Learning
    - Assignment Posted: Assignment 4
    - Assignment Submission: Assignment 3
- SPRING BREAK – NO CLASS
- Session 8
    - Date: 03/29/2019
    - Topic: Presentations
- Session 9
    - Date: 04/05/2019
    - Topic: Biases ML
- Session 10
    - Date: 04/12/2019
    - Topic: Text Analysis
    - Assignment Submission: Assignment 4
- Session 11
    - Date: 04/19/2019
    - Topic: Visualization
    - Assignment Posted: Assignment 5
- Session 12
    - Date: 04/26/2019
    - Topic: Inference
- Session 13
    - Date: 05/03/2019
    - Topic: Privacy
    - Assignment Submission: Assignment 5

- Session 14
    - Date: 05/10/2019
    - Topic: Presentations

Detailed Course Overview

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

- Organizational details for class/housekeeping
- Tutorial on how to define and scope a 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
- Get to know research projects of the class

LAB

- Overview of the computing environment and project space
- Intro into Linux

Readings

- Chapter 1 of textbook
- Linux/Unix common terminal commands
- Measuring the impact of R&D Spending, Nature
- Watching the players, not the scoreboard, Nature
- Wrapping it up in a person, Examining the earnings and employment outcomes for PhD recipients

Session 2: Python for Data Analytics

- Python/Pandas basics: Python basics needed for all data analyses done in this class
- 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

- Python exercises and Introduction to data being used in class

Readings

- Wes McKinney, Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2012, pp. 466
- Python for Economists

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

Session 3: Big Data and Policy Research

- Introduction into the big data landscape
- Compare big data and survey data, and data commonly being used in Social Sciences
- Advantages/Disadvantages of different data sources for research

LAB

- Project work/ Identifying additional datasets for projects

Session 4: Intro into APIs

- Retrieving data from the web. Intro into APIs
- The goal is to become familiar with different types of APIs (GET- and POST- based HTTP APIs), different formats of requests, and how to learn a given API
- Learn the tools used to interact with network based APIs: Understand and use the tools for talking directly with APIs over HTTP connection, introduce libraries that abstract the details of the API and present a simplified programmatic interface

LAB

- Making raw HTTP API requests, Using pre-packaged API client libraries, practical considerations

Readings

- Chapter 2 of textbook
- Ryan Mitchell, Web Scraping with Python, O'Reilly Media, 2015
- Python's requests & Beautiful Soup libraries (for web scraping & APIs)

Session 5: Record Linkage

- Theory and Principles of record linkage
- Pre-processing needed before linking records: How to parse string fields

LAB

- Record Linkage exercise, Introduction into regex

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

More Resources (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
- Regex: link to PDF
- Python regular expressions
- Online regular expression tester

Sessions 6 - 7: Machine learning

- 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

LABS

- Machine Learning modeling with Python, exercises
- Project work

Readings:

- Chapter 6, textbook
- Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning Data Mining, Inference, and Prediction. Springer, 2009.
- 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

Session 08: Presentations of Preliminary Results

- Students will be presenting their first results. We will discuss and provide feedback.

Session 09: Text Analysis

- Introduction in text analysis: Information retrieval, clustering and text categorization, text summarization, machine translation

- How to transform a corpus of text into a matrix on which NLP can be applied?
- Learn how to implement topic modeling
- Document tagging and evaluation of document tagging

LAB

- Text Analysis with Python, exercises

Readings

- Chapter 7 of textbook
- Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. O'Reilly, 2009

Session 10: Biases in Machine Learning

- Address biases in machine learning techniques and their consequences for public policy, for example how race biases can lead to unfair treatment of ethnic minorities in public policy.

LABS

- Guided project work

Session 11: Information Visualization

- Theory of information visualization
- Communication tool
- Choosing a chart type
- Labeling and information overload
- Color consideration

LAB

- Visualizing analytical results with Python, exercises

Readings

- Chapter 9 of textbook

Session 12: Inference/Errors

- How to deal with inference and the errors associated with big data
- Problems of Big data and the errors resulting from it
- The total error paradigm: Traditional models and their implication for big data research

LAB

- Guided project work

Readings

- 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 13: Privacy, Confidentiality, and Ethics

- Recognize where and understand why ethical issues can arise when applying analytics to policy problems starting with collection and moving through the management, sharing, and analysis of data
- Plan, execute, and evaluate a research project along privacy concerns and ethical obligations
- Key technical, ethical, policy, and legal terms and concepts that are relevant to a normative assessment of novel analytic techniques and tools for mitigating or managing the ethical concerns.

LAB

- Guided 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
- The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Ethical Principles and Guidelines for the Protection of HumanSubjects of Research [The Belmont Report], Washington, DC: Department of Health, Education, and Welfare, April 18, 1979.

Session 14: Presentations of Final Results

- Students will be presenting their final results. We will discuss and provide feedback.

Evaluation

During the class students will be assigned their own small research project they work on during the entire semester. There will be a midterm presentation and final presentation of the project

results. At the end of the semester (05/13/2019) 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 consist of a problem set based on the method learned in class. The assignments constitute 50% of the grade:

- Assignment 1: Python Code Lab 10%
- Assignment 2: API exercise 10%
- Assignment 3: Preparing Data for Record Linkage 10%
- Assignment 4: Machine Learning Model for research 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 at least 24-hours prior to the beginning of the following session. 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. (E.g., if you dont 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. Lets say, 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 discussion board. If you dont get a solution, escalate it to us.

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, regular attendance & contributive participation in class will constitute 10% of the final grade.

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.