DS2001, Spring 2019
Northeastern University, Boston | New College of Humanities, London
Data, Ethics and Culture Program

Class Times & Locations	Wednesday - 4.00-5.40 pm NCH – The Archive
Instructor	Ieke de Vries Violence and Justice Research Laboratory School of Criminology and Criminal Justice i.devries@northeastern.edu
Office Hours:	Mondays from 14:00 to 15:00 – The Archive Thursdays from 09:00 – 11:00 – *Room changes every week; please consult your time-schedule for updates or send me an email.*

Required Textbooks:
Matt Salganik (2017). Bit by Bit: Social Research in the Digital Age. Princeton, NJ: Princeton University Press. It is available here. I also encourage you to buy and read the entire book when you have a chance.
Additional readings can be found online and/or will be posted on the Blackboard web page for this course.

Course Description: The amount of information that we are now able to store increases rapidly and so do the computational techniques to access and analyze this information. Familiarity with computational techniques is of increased importance in this digital era. The course introduces students to the fundamental concepts and programming techniques that are needed to access and analyze data. It offers students the opportunity to develop research questions and answer them with techniques learned in the course. How can we predict public ratings of products, rentals, neighborhoods and so on, utilizing the text in reviews? Can we understand a sentiment based on specific words utilized? How do we predict where crime events occur when combining publicly available datasets? When does a tweet about a social matter cause other people to retweet? How does information travel through networks? There are a lot of really interesting questions that we can now answer given that information is stored digitally. However, the digital era is also an era of uncertainty. Data and techniques are rapidly developing but our laws, regulations and principles to safeguard ethics might not be developing as quickly. The course challenges students to think critically about ethics and ethical dilemmas when answering important questions utilizing computational techniques.

Course Learning Goals: The practicum aims to help students familiarize with the fundamental techniques of programming in the social sciences and humanities such that:

- Students understand the concepts behind programming and can apply the acquired techniques to answer questions relevant to the social sciences and humanities;
- Students understand how to write scripts to read in and describe different data formats;
- Students can apply computational techniques to visualize data patterns;
- Students understand the possible social implications of computational techniques;
- Students are aware of ethical challenges and can contribute to the debate on ethical challenges associated with computational techniques and big data in the social sciences and humanities;

- Students can engage in discussions about cross-cultural similarities and differences in programming, data access and ethics;
- Students know how to answer questions about social problems utilizing the techniques acquired in the class.

Course prerequisites: No pre-requisites are required. The practicum accompanies the lectures of DS2000 and includes exercises to practice with the concepts introduced in DS2000 on a weekly basis. The practicum will also address issues introduced in other courses included in the 2019 Data, Ethics, and Culture Program.

Practicum format: The practicum is scheduled for Wednesdays from 4.00 pm to 5.40 pm. Typically, we will start the class with an open and active discussion about the ethical and social implications of techniques covered that specific week, followed by brief hands-on exercises extending on the course material in DS2000. The practicum seeks to advance your programming skills and domain knowledge so that you can use computational techniques to answer important questions about the social world. To that end, every practicum will have applied reading and coding exercises to help you become familiar with processing and analyzing social science material through Python. The expectation is that you come prepared to every class.

Course Requirements: This practicum will be taught in the open source programming language Python. We will begin by writing Python code in the open source text editor Atom. Part way through the course we will introduce the programming environment Jupyter. Prior to the first day of class, students should install Anaconda for Python 3.7 and the text editor Atom on their laptops (https://www.continuum.io/downloads ; https://atom.io/). Anaconda includes Python, the necessary Python packages, and Jupyter. **Laptops are necessary and students must bring their laptop to class every day. When you do not have a laptop, please contact me so that we can work on a solution.**

Course Activities and Assignments: There are weekly quizzes, reading assignments and coding exercises to be submitted every practicum. Most assignments are individual assignments though students are encouraged to give each other feedback and help each other out during and beyond the practicum. At the beginning of the semester, students will form teams of two or three to work collaboratively on the final project. If preferred, you can also work individually on your final project.

Practicum Grading: It is the student's responsibility to become familiar with the assigned readings prior to each class and to be aware of the project deadlines. Your final grade will be calculated based on the following:

Milestones and Expectations	*Due Dates*	*%*
Practicum Exercises (Reading and Coding exercises)	Weekly	**40%**
Final Project and Presentation		**60%**
Submit group preferences	**Week 7 (Feb. 27)**	-
Submit project proposal	**Week 10 (March 20)**	20%
Submit final project	**Week 14 (Apr. 15)**	20%
Final presentations	**Week 14 (Apr. 17)**	20%
Final Grade	-	**100%**

Weekly Practicum Exercises: There are three types of exercises to be submitted:

1) *Quizzes*: Biweekly, a practicum starts with a 10-minute quiz practicing your knowledge on the material taught in the DS2000 lectures. ***Grades on quizzes are part of DS2000.***
2) *Reading exercises.* The course will teach you programming and analytical skills in Python but also emphasizes the application to social science and humanities questions. Short readings will be assigned each week to increase your familiarity with the role of computational methods in the social sciences. Through Blackboard, you will be presented with a brief question about the reading and you should submit your reading responses based on the literature. Most of the readings are assigned from Matt Salganik's book *Bit by Bit* (see above). ***All reading responses are due by 4pm on Wednesdays and can be submitted in the Blackboard Discussion Environment. Weeks 1 and 10 do not have reading exercises assigned.***
3) *Coding exercises:* There will be weekly coding exercises to get practice with every week's new concepts and techniques. ***Coding exercises are to be submitted by the end of each practicum.***
   Note: see the first **Practicum Handout** for how to test and submit practicum assignments.

Final Project: The final project exists of a project proposal, Jupyter notebook, and final presentation. See the **Project Guidelines** for further details and below some starter points.

Project Proposal: The goal of the final project is to creatively combine the techniques you learned in the course to **preliminarily** explore a question related to the humanities or social sciences that has either not been addressed before, or explore an old question in new ways. Note that this is an introductory course, and I will present a lot of material throughout the semester, so your final project will by necessity be only a preliminary exploration of a research question and I do understand you do not know the full universe of questions that have already been explored.

Through this project you should show that you understand (a) what types of questions are interesting or important to humanists and/or social scientists, (b) what types of questions can be best answered using computational or digital techniques, (c) what types of techniques and evidence are appropriate to best answer your question, and (d) that you can think about how to present your findings and analysis in a reproducible way and in a way that supports, and persuades others of, your (preliminary) conclusion.

The project proposal will be a Jupyter notebook detailing a preliminary plan for your final project. You should include the following in your project proposal:

1. Identify a general question related to the humanities or the social sciences that you plan to address in your final project. You should outline why this is an interesting or important question and describe why computational methods are necessary and/or helpful in exploring this question. If possible, explain how others have answered/attempted to answer this question using different methods.
2. Identify the data or collection of material you will use to explore this question, and briefly describe why the data/material is appropriate. Additionally, describe how you will collect the data/material and whether or not it will need to be cleaned prior to analysis. If possible, import your data and provide a glimpse of its format.
3. Describe the techniques you expect, or would like, to use to analyze the data/material and explore your question. Why these techniques and not others? What kind of evidence will these techniques produce, and how will this help you answer the question and persuade

others of your answer? If you already have some preliminary analyses, include these as well.
4. Briefly discuss any data visualization or interpretive techniques you will, or would like to, use to present your findings and convince others of your interpretation.
5. If your project will be a team project, detail a planned division of labor and a shared contract about your collective expectations (I encourage you to revisit this document well before the final project is due, to check-in with each about whether the shared contract is working). The best teams will include people from different disciplinary backgrounds so you can leverage each other's specialized knowledge, e.g. a computer scientist and a historian.

Final Project: Keeping the above goals in mind, the final Jupyter notebook should include the following:

1. 1-2 cells describing the question or puzzle you are exploring, why it is interesting or important, (briefly) how others have attempted to answer this question, and how you are improving on these answers. If no one has addressed this question, explain why you think this is the case. In other words, what are you doing that's different than what others have done?
2. 2-4 cells describing the data or material you are using to explore the question and how you collected the data or material. These cells should include summary statistics of the data/material. If appropriate, describe what your data or material are representative of.
3. 2-10 cells containing the analysis and visualization, or a visualization and possible steps toward a more sophisticated analysis. These cells should contain a description of the planned analysis process and why it is appropriate for your question and data/material, followed by code implementing either some of the techniques or at least provides some summary descriptions/visualizations of your data or material, the output from the calculations or the summary descriptions of your data or material.
4. 1-2 cells detailing your interpretation of the output, and broader conclusions about the social world that you draw from your exploration, or that you would hope to draw if you carried the project further. Support your interpretation with evidence from your analysis. End with suggestions for further analyses and other data or material that could help us continue to explore your question.
5. If you worked in a team, revisit the planned division of labor and your shared contract that you submitted in the project proposal, and briefly describe whether you think you collectively lived up to the contract and why. If you would like to send me private thoughts about your team you can do so as well.

Presentation: The final presentation allows you to present your work to fellow students. An effective and persuading presentation should include:

1. Your research question and an appealing introduction to the importance of answering your research question.
2. Background about the extent to which prior research or studies have answered this question.
3. Methods: Which data are you utilizing, how have you accessed the data and what does the data represent (visualize some basic descriptive information about the data). Make also sure to include a brief discussion of possible ethical issues related to your methods. Which analytical strategies have you utilized to answer your research question?
4. Findings: Utilizing the analytical strategies from step 3, present your findings in an effective way. Use as many visualizations as possible to summarize your findings. Present your code only when it is needed to explain the steps that have brought you to the findings.

5. Conclusion: How would you answer your research question?
6. Limitations and next steps: What were challenges in your research and how could these challenges inform next steps in computational social science research?
7. Social implications: Given the findings and while keeping in mind the limitations, how does your research contribute to our understanding of our social world. What are possible policy implications?

Course Resources:

Lecture Notes: The lecture notes of DS2000 are available through this url. Additional lecture notes utilized in the practicum will be made available through Blackboard.

Other Resources: There are a couple of interesting sources in the area. NCH has a list of exciting events that can be relevant to this course or the Data, Ethics and Culture Program more generally. Keep an eye out also for the new world-first Center for Data Ethics and Innovation in London. And given that you are in the area, you might be interested in seeing interesting research updates by the Oxford Internet Institute. When you are back in Boston, the Digital Scholarship Group at Northeastern University is a great resource. They offer digital data collections, a quiet space to work and a wealth of services.

Blackboard and Communication: A discussion board for this course is available on Piazza. This discussion board can serve as a platform to ask course-related questions and help fellow students. Start a new thread for new errors or questions. Everyone is encouraged to answer each other's questions in a respectful way. Make sure to also read previous posts in case a question you may have is already answered by someone else.

Office Hours: You are also encouraged to come to my office hours. Email can be used for quick logistical questions or to inform me about a planned absence.

Class Policies:
All Policies as described in the DEC Student Academic Handbook apply to this course as well. Below are a few key policies. Please contact me if you have any questions about these policies.

Class Discussion and Participation: Participation in class, in the form of discussion, helping and collaborating with other students (e.g. by demonstrating code) is essential. Every week builds on the skills learned in previous weeks and it is therefore critical to attend every class. Learning Python is like learning a foreign language, the best way to learn is to keep using it all the time.

Attendance, Early Departures and Absences: Registers will be taken for all classes. ***Please note that class attendance is also necessary for you to submit the weekly reading and coding exercises.*** If you know you are going to miss a class, you should notify me, preferably at least four days in advance in order to be able to make up the points for the weekly practicum exercises. As indicated in the Student Academic Handbook, you should also notify the Program Director (Peter Maber), by email, as soon as you are aware that you will be missing a class for any reason. Evidence for an excused absence may be requested by me or the Program Director. No early departures or absences are permitted unless previously discussed with me. Systematic tardiness or early departures (defined as being late for class or leaving class early more than two times) will lead to a deduction from your grade.

Late and Missing Assignments
I must be notified in advance if you anticipate missing an assignment for a valid reason (e.g. serious emergencies). Documentation may be requested, and I reserve the right to approve or deny any such requests. Missing assignments should be turned in at a later point in time to be agreed upon with me. Late assignments without a valid reason that was notified to the instructor in advance results in a grade deduction as indicated in the Student Academic Handbook.

Respect
The course involves students from different disciplines and with varying familiarity to programming and analyses. This demands an open attitude, respect and a willingness to help fellow students. It is important to give everybody the space to talk and address our comments at the ideas and not the person. Disrespect will absolutely not be tolerated.

Sports-Related Absences Policy
All student-athletes are required to notify me at the beginning of the course about all sports-related absences.

Students with Disabilities
Any student who may require special accommodations for this course should notify me as soon as possible. You may need to register with the university's Disability Resource Center (DRC). The DRC can provide students with services such as note-takers and extended time for taking exams. The DRC is located in 20 Dodge Hall and can be reached at 617-373-2675 or through www.northeastern.edu/drc/.

Academic Integrity Policy:
Please see the DEC Student Academic Handbook for all policies on academic honesty that apply to all courses you take as part of the Data, Culture and Ethics program.

- All students must follow Northeastern University's procedures regarding academic integrity. Commitment to the principles of academic integrity is essential to the mission of Northeastern University. Northeastern University expects students to complete all examinations, tests, papers, creative projects, and assignments of any kind according to the highest ethical standards as set forth in the Northeastern University Student Handbook. It is the student's responsibility to become familiar with his/her rights and responsibilities.
- A detailed explanation of what constitutes academic cheating, plagiarism, and facilitating academic dishonesty, and how such cases are handled by Northeastern University can be found on the website of the Office of Student Conduct and Conflict Resolution (OSCCR), http://www.northeastern.edu/osccr/academic-integrity-policy/. I have summarized a few points below (but this is by no means exhaustive and you should carefully read the Student Academic Handbook and the above URL):
    - Cheating includes handing in the same paper for more than one course without explicit permission from the instructors.
    - Cheating includes storing notes in a portable electronic devise for use during an examination.
    - Plagiarism can occur accidentally or deliberately. It is defined as using as one's own words, ideas, data, code, or other original academic material of another without providing proper citation or attribution. Forgetting to document ideas or materials taken from another source does not exempt one from plagiarizing.

    - Participation in academically dishonest activities includes misrepresenting oneself or one's circumstances to an instructor.
    - Facilitating academic dishonesty is defined as intentionally or knowingly helping or contributing to the violation of any provision of the Northeastern University Student Handbook.
        - This includes doing academic work for another student.
        - This includes making available previously used academic work for another individual who intends to resubmit the work for credit.
- In this course, cheating or plagiarizing on an assignment, as defined by Northeastern University's Academic Integrity Policy, will result in receiving a "0" on that assignment, meaning it may result in a failing grade for the course. This conduct will also be reported to the Office of Student Conduct and Conflict Resolution (http://www.northeastern.edu/osccr/).
- Please consult the "Avoiding Plagiarism" on the NU Library Website: http://library.northeastern.edu/get-help/research-tutorials/avoid-plagiarism.
- If you have any questions of whether you should be citing to a source or paraphrasing a source in a different way, please let me know. Often situations implicating academic integrity can be avoided as they come about due to confusion regarding appropriate citations.

Title IX – Gender Discrimination and Sexual Violence: Title IX of the Education Amendments of 1972 is a federal law that prohibits discrimination based on gender, which includes sexual harassment and sexual assault. Title IX prohibits sex discrimination in all university programs and activities, including, but not limited to; student services, academic programs, class assignment, grading, athletics, admissions, recreation, recruiting, financial aid, counseling and guidance, discipline, housing, and employment. Title IX also prohibits retaliation against people for making or participating in complaints of sex discrimination. If you have questions or concerns regarding discrimination based on gender, sexual harassment, or sexual assault, you can get help at the Northeastern Office for Gender Equity and Compliance, http://www.northeastern.edu/titleix/ Title IX-related concerns will be reported to the NU Title IX office.

Grading Scale (in %):

A	94 – 100	B+	87 - < 90	C+	77 - < 80	D+	67 - < 70	F < 60
A-	90 - <94	B	82 - < 87	C	72 - < 77	D	62 - < 67	
		B-	80 - < 82	C-	70 - < 72	D-	60 - < 62	

Class Schedule and Topical Outline

Subject to change at the discretion of the instructor. Changes will be announced in class and through Blackboard. It is your responsibility to be aware of such changes.

Part 1	Hello, Python!
Part 2	Intro to Programming
Part 3	Working with data
Part 4	Advanced working with data

Date	Week	Topics	Subtopics	Readings
Part 1 – Hello, Python!				
Jan. 9	1	Hello, World!	Course Structure; Why Python; Types of errors; Values, Data Types; Console I/O; formatted strings; statements and expressions; functions; **Exercise: "Hello, World!"**	No readings
Part 2 – Intro to programming				
Jan. 16	2	Looping and Conditional Statements	Boolean variables/expressions; if + for **Exercise: conditional statements**	Salganik (2017). Introduction.
Jan. 23	3	Looping and Conditional Statements	while loops, range() function **Exercise: loop over data, aggregate results**	Salganik (2017). Chapter 2. Observing Behavior. Parts 2.1 – 2.3.
Jan. 30	4	Functions	Creating functions; tuples; list comprehensions; variable scope **Exercise: practice with functions**	Salganik (2017). Chapter 2. Observing Behavior. Parts 2.4 – 2.5.

Feb. 6	5	Functions	Modules; __main__; useful string and list functions **Exercise: more practice with functions**	Salganik (2017). Chapter 3. Asking Questions. Parts 3.1-3.4; 3.7.
Part 3 – Working with data				
Feb. 13	6	Files	Reading and processing a file of data and outputting results **Exercise: working with files**	Salganik (2017). Ethics. Parts 6.1 – 6.3.
Feb. 18 – Feb. 23 [Reading week]				
Feb. 27	7	Dictionaries	Reading and processing data utilizing dictionaries **Exercise: working with dictionaries**	Safiya Noble (2018). *Algorithms of Oppression: "Introduction"* New York University Press. (Blackboard)
Feb 27 Due in Class: Submit group preferences				
March. 6	8	OOP	Object-Oriented Programming **Exercise: design and implement a data object**	Cathy O'Neil (2016). *Weapons of Math Destruction: "Introduction."* Crown Publishing Company. (Blackboard)
March 6 In Class: Finalize groups				
Part 4 – Advanced working with data				
March. 13	9	Hello, Jupyter!	Jupyter Notebooks; Matplotlib and Other Visualizations **Exercise: "Hello, Notebook!"**	Kieran Healy and James Moody (2014). Data Visualization in Sociology. *American Review of Sociology* 40: 105-28 (Blackboard)
March. 20	10	Navigating through data structures	Navigating documentation; Csv modules and API **Exercise: Analyzing and Visualizing Real Data**	No readings; work on project proposal

March 20 – 3pm: Submit proposal				
March. 27	11	Case Study	Bonus! Case Study **Exercise: API call**	DiMaggio, Naga and Blei (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* 41(6): 570-606.
March 27: Proposal feedback				
Apr. 3	12	Data frames	Pandas **Exercise: working with pandas**	Lazer et al. 2009. Computational Social Science. *Science 323* (5915): 721-723.
Apr. 10	13	Machine learning	Intro to machine learning **Exercise: predicting**	Jordan & Mitchell (2015). Machine learning: Trends, perspectives, and prospects. *Science* 349: 255-260. Optional: Google Uses Searches to Track Flu's Spread in The New York Times
Apr. 17	14	**Final Presentations**		