Northeastern University

**DS 3000: Foundations of Data Science**

Fall 2023 Course Syllabus (CRNs: 17451, 17959, 18514)

**Professor Info**

| | | | |
|---|---|---|---|
| **Instructor:** | Dr. Eric Gerber | **Office:** | ME 331 |
| **Email:** | e.gerber@northeastern.edu | **Email Prefix:** | [DS 3000]: |

**Links**

| | | | |
|---|---|---|---|
| **Canvas Page:** | Canvas Login | | |
| **Gradescope:** | Gradescope Link | **Piazza:** | Piazza Link |
| **Zoom:** | Zoom Link | | |
| **OH Link:** | OH App Link | **OH App Help Guide:** | OH App Help Guide |

**TL;DR Summary (See following pages for detailed syllabus)**

- The two main websites for you to use are **Canvas** (for course information) and **Gradescope** (for assignment submission).
  - Piazza and Khoury Office Hours websites are your main sources for help.
  - See links to all four sites above.
- You may **always** email the professor to set up meetings outside of regularly scheduled office hours.
- **In-person attendance is mandatory**, however you get three free passes.
- You are expected to have a introductory level grasp of Python before starting the course. If you need to review, there are some review materials on Canvas.
  - All assignments will be completed in Jupyter notebooks. The program you use to open and edit the notebooks (Google Colab, Cantor, or Jupyter Notebook itself) is up to you, but you will always upload the raw .ipynb file to Gradescope as your submission.
- You must bring your laptop to every class.
- **There is no late homework accepted**. Homework submitted more than 24 hours in advance will receive extra credit.
- **Grade Breakdown**: 5% Attendance, 5% In-class Labs, 40% Homework, 15% Out-of-class Quizzes, 35% Final Group Project
- Please see **Page 7** for important statements regarding Academic Integrity, DRC and other student resources, and Title IX information.
- **Finally:** I am here to **support your learning**. If you have any questions about how the rules outlined in this syllabus are meant to do that, please don't hesitate to ask.

**Class Meeting Times/Classrooms:**

- Section 4 (CRN: 17451): TF 9:50 am – 11:30 am (**INV 019**)
- Section 5 (CRN: 17959): T 11:45 am – 1:25 pm, R 2:50 pm – 4:30 pm (**HT 130**)
- Section 6 (CRN: 18514): TF 1:35 pm – 3:15 pm (**HT 130**)

**Communication:** Please check your email, Piazza, and Canvas regularly. **Be aware that attendance is mandatory**, though there is flexibility built into the grading scheme in case you must miss class. See the attendance section below for more details.

**Course Canvas Page:** All sections have been cross-listed to the same Canvas page. All relevant information and assignments will be posted here, and you can access both Gradescope and Piazza from this page, and also keep track of your grades. Log in to Canvas at `http://canvas.northeastern.edu` and find our class; you should automatically be added to the course.

**Office Hours:** To facilitate your learning and engagement, your TAs will be holding virtual office hours. This course will make use of the Khoury Office Hours App, which you will need a Khoury account to access. The schedule for the virtual Office Hours can be found at: `http://khouryofficehours.com/`:

- Log on to the Khoury Office Hours App (link above).
- All TAs currently hosting virtual hours will have set up a Queue, which you can join.
- When it is your turn to be assisted, a Teams Chat will open through which the TA will assist you.

If you want more help understanding how the Khoury Office Hours App works, there is also a Help Guide.

Additionally, the professor is happy to meet students who drop by his office, Meserve Hall 331, on a first-come, first-serve basis. While there is no strict schedule of availability for this, the professor is usually in the office:

- Thursdays 10:30 am – 2:30 pm
- Fridays 12:00 pm – 1:30 pm
  - Note that occasionally the professor will be out to lunch during these times.

You may also **always** email me, or ask after class, to set up a time for us to meet that works best for you.

**Prerequisites:**

- CS 2510 or DS 2500 (strong background in programming assumed)

**Description:**
Data Science (DS) is an interdisciplinary field of inquiry concerned with the study and application of systematically extracting generalizable knowledge from data and using this knowledge to draw useful conclusions. This course is an introductory DS class focusing on the foundations of DS as an emerging field. The course introduces core modern DS technologies and methods that provide a foundation for subsequent DS classes. Therefore, the focus of this class is on the breadth of various DS skills, rather than the depth of the topics covered. As a skills-based course, DS 3000 will cover the use of Python for DS and will introduce some of the widely-used essential Python libraries, such as NumPy, pandas, matplotlib, and scikit-learn. More specifically, this class covers:

- working with applied linear algebra in standard numerical computing libraries (e.g. NumPy)
- loading, processing and integrating data from a variety of structured and unstructured sources using Python libraries (e.g. pandas)
- visualizing data using basic techniques and tools (e.g. matplotlib, seaborn, plotly)
- applying introductory concepts in probability, statistics, and machine learning using Python libraries (e.g. scikit-learn)
- using a standard DS tool (e.g. Jupyter Notebook).

**Materials:** The majority of class time will be spent writing and running computer code. Each student should bring their laptop to each class and be prepared to submit in-class assignments, when they occur, at the conclusion of the class.

**Textbook and References:** There is **NO** required textbook for this course. Material and lecture notes will be posted and updated on Canvas every week. If you are interested in having a free textbook as a resource, the below is quite useful:

- *Python Data Science Handbook*, by Jake VanderPlas.
  https://jakevdp.github.io/PythonDataScienceHandbook/

**Tentative Course Outline:** There is a tentative schedule (updated as needed) on the Canvas page. This is meant to give a rough idea of the topics covered over the course of the semester, and of tentative due dates, but is by no means set in stone.

**Lectures: You are expected to attend every class period, if able.** Be prepared to type and run code in every class and to save your work. There will be a Qwickly Check-In PIN at the beginning of every class that you will have until the end of class to enter via and indicate your attendance. See next page for details.

**Lecture Notes:** All Jupyter Notebooks and relevant in-class material will be posted to Canvas before each class. These materials are meant to help you review, and catch-up should you miss a class. They are not a replacement for attending class.

**Push-up Promise:** To encourage you to attend and pay attention, I will make a promise to you that has worked out for all parties in the past. For every mistake **YOU** catch me make in lecture over the course of the semester, I will do **ONE** push-up at the end of the last day of class. Every tally will occur only after one of you points out a mistake I made (and after we discuss why/how I made the mistake). This not only gives you an incentive to pay close attention in lecture, but gives me an incentive to limit my mistakes. Don't be afraid to point out my mistakes! We all make them, and it really does help learning!

**Attendance (5%):** Attendance is worth 5% of your overall grade, and is measured via Qwickly Check-In. At the beginning of each class, I will write a PIN for the class on the board and you will have until the end of class to enter it in Qwickly to log your attendance.

Life happens and sometimes you will be unable to attend class. All students will be given up to three "no-question" absences. You may not choose when to use them; at the **end** of the semester, your first three missed Qwickly points will be returned to you. Only in **the most extreme** situations (determined at the professor's discretion) will further absences be granted.

**Group Exit Interview:** During finals week, there will be time slots for students to attend a group exit interview with the professor. This will give you another opportunity to recover **one** missed attendance point.

**In-class assignments/Labs (5%):** We will occasionally practice concepts with small group work in-class. These labs are meant to be low-stakes, and will be due **at the end of the class day** (by **11:59 pm**) via Gradescope submission.

**Visitors:** My hope is that lab days will start with a short question and answer session with a visitor who works in the Data Science industry. This is a tentative plan, which will rely on the availability of the visitors. In the past, we have had guest speakers from companies like Meta, the Detroit Tigers, and Finra.

**Homework (40%):** There are four planned homework assignments (worth 10% each) that will typically be assigned and due every three weeks, due **Tuesdays by 11:59 pm**. Assignments will be done through Gradescope. These assignments will have several questions, some of which will be relatively open-ended/project-like. Your TAs and I am very happy to help with homework questions, but you are expected to show independent thought on each assignment. Programming code requires thorough commenting, which must be done independently. Interpretations of generated output must be your own work. If there is evidence that you directly copied answers from **ANY** source, you will receive a score of 0 for the assignment, and further instances of academic misconduct will lead to disciplinary action. Please also see the student guide on Canvas for working with AI tools.

**Early/Late Homework:** There will be no late homework accepted except in **extreme** situations. However, students may receive **5% Extra Credit** if their assignment is submitted **more than 24 hours early: by Monday at 11:59 pm**.

**Quizzes (15%):** There will three out-of-class quizzes done through Gradescope which will assess your understanding of the course material. They will be more like traditional exams in style. Quizzes will be written to take about one hour, though up to two hours will be given to complete the quiz. You may choose when to take the quiz within the window given to best suit your schedule.

**Warning:** Under no circumstances may students observe the quiz before they sign in to Gradescope to officially start their quiz. "Observations" include both observing the quiz documents as well as receiving any sense of the content or difficulty from other students. Sharing this information will result in academic integrity consequences.

**Final Project (35%):** You will be required to conduct a final group project focusing on finding a large dataset of interest, collecting, cleaning and analyzing the data with appropriate statistical techniques, and interpreting the results in an appropriate manner. The following policies are in place to ensure an equitable share of effort is done by all team members on the project:

- To receive any credit for the final project, you must earn an average of 60% on **both** the homework and quizzes separately.

- At the completion of the project, each group member will be asked to describe the contribution of all other group members with a **Statement of Contributions**. With this information, the instructor may adjust grades of individual group members, including failing students who have not made meaningful contributions to the final project.
- There will be a presentation aspect of the project where each group member will be required to present on something of substance.
- If there are life circumstances preventing your adequate participation in the group project, please reach out to me as soon as possible. The earlier in the semester we're aware of a situation, the more options we have to put supports in place and remedy it.

More details on the project will be shared after the first two weeks of the semester.

**Total Grading Policy:** Attendance (5%), Labs (5%), Homework (40%), Quizzes (15%), Final Project (35%).

The grading scale will be the traditional 60-70-80-90 for D-C-B-A, but + and - grades will be given based on cut-offs determined at the end of the semester.

**Important Tentative Dates:**

| | |
|---|---|
| Project Proposals Due, Groups Assigned | October 16 |
| Project Data and Analysis Plan Due | November 6 |
| Thanksgiving Break | November 23-24 |
| Final Project Due | November 30 |

**Academic Integrity:** Under no circumstances may one student view or share their ungraded homework or quiz with another student. Sharing or viewing another students ungraded work will result in a failing course grade. This does not extend to discussion of concepts or ideas, but prohibits any sharing of personal code. Academic dishonesty is not tolerated and in addition to course failure, all violations will be reported to OSCCR: http://www.northeastern.edu/osccr/academic-integrity.

Like every computer scientist, you are encouraged to borrow code you find online, so long as it was not written for this class in any semester. Doing so requires that you attribute credit to the source:

- a quick url link comment (e.g. stackoverflow) will suffice.
- please also see the student guide on Canvas for working with AI tools.

**Disability Resource Center (DRC):** The office is available to assist students who have a legally documented disability or students who suspect that they may have a disability. If you have a disabling condition that may interfere with your ability to successfully complete this course, please contact the Disability Resource Center: http://www.northeastern.edu/drc/.

**Student Resources:** Your health and wellness is more important than any assignment. Please use the resources below if you are struggling, and don't hesitate to ask for my help!

- **Counseling Center (24/7 support): 877-233-9477** or https://www.northeastern.edu/uhcs/

**Title IX:** I am a mandatory reporter under Title IX, which means that I am required to report any and all allegations of discrimination to the Title IX coordinator.

Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance.

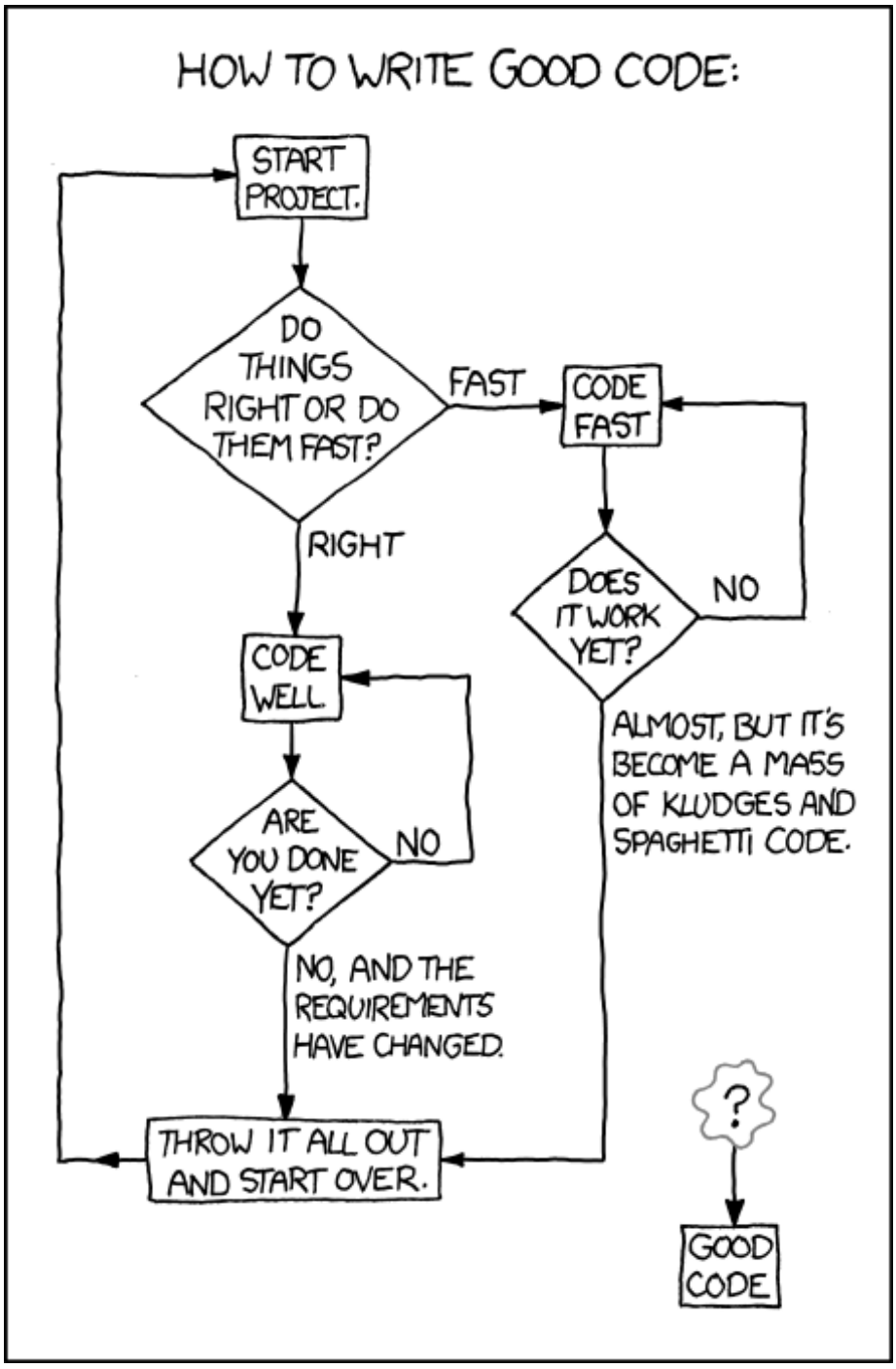
Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty and staff.

If you or someone you know has been a survivor of a Prohibited Offense, confidential support and guidance can be found through University Health and Counseling Services staff and the Center for Spiritual Dialogue and Service clergy members. By law, those employees are not required to report allegations of sex or gender-based discrimination to the University.

Alleged violations can be reported non-confidentially to the Title IX Coordinator within The Office for Gender Equity and Compliance at and/or through NUPD (Emergency 617-373-3333; Non-Emergency 617-373-2121). Reporting Prohibited Offenses to NUPD does **NOT** commit the victim/affected party to future legal action.

**Changes to Syllabus:** This course syllabus is intended as a guide. The instructor reserves the right to revise any part of the syllabus during the course. Any changes will be announced well in advance, during class time and via email.

**Some "Encouragement" (via xkcd.com/844/ and xkcd.com/1838/):**




THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS