CSE 519 -- Data Science (Fall 2024)
Prof. Steven Skiena
http://www.cs.stonybrook.edu/~skiena/519

Date/Time/Place: Tuesday-Thursday, 9:30-10:50AM, in 145 Engineering Hall
Office Hours: Tuesday-Thursday, 11:00AM-12:30PM in 251 New Computer Science, and by appointment

Overview: Data Science is an emerging discipline at the intersection of statistics, machine learning, data visualization, and mathematical modeling. This course is designed to provide a hands-on introduction to Data Science by challenging students to build predictive models.

Caveat: I am experimenting with more changes to course policies this semester in an attempt to serve you better, so please be patient and understanding if some may have to change again because of course size, technology issues, teaching support, or epidemiological factors. As I write this, the total course registration is roughly 170 students, a little smaller than usual but beyond the threshold where it is necessary to transition to "large course mode" of fewer project choices and more intensive peer grading.

I note that tools like ChatGPT have improved considerably even since my previous iteration of the course. My official policy is given below, but I will trust in student's openness as we better understand what they are capable of.

Textbook: We will use my book "*The Data Science Design Manual*", Springer-Verlag, 2017. The book website is http://www.data-manual.com. If you like the book I would appreciate it if you let Amazon know that.

Problem Wiki: We have set up a solution Wiki at http://data-manual.com/datawiki, where students can enter answers to problems from the textbook. *Caveat Emptor* (let the buyer beware) because I never look at the posted solutions and have no reason to believe they are correct.

Recommended Readings: Two other books may be useful to read and consult:

- Python for Data Analysis, by Wes McKinney, O'Reilly Media, 2013 -- This book is a nuts and bolts guide to data wrangling with Python, including such tools/libraries as Pandas, NumPy, and IPython. You will be expected to use these tools in doing your projects.
- The Signal and the Noise: Why so many predictions fail but some don't, by Nate Silver, Penguin Press, 2012 -- This popular book focuses on how effectively data can be used to make predictions in domains like sports, science, economics, and politics. This is *exactly* what we are trying to do, and Silver's book is an excellent model to build on.

Course Projects: Roughly one third of the course grade will come from a course project. Students will typically work in small groups (2-3 people) on independent research projects. I will distribute a list of 5-7 possible projects about six weeks into the semester. The project list

will be kept short this year to facilitate management of such a large course.

Video: In Fall 2014, I ran a version of this course as a ``TV reality show'' in data science, resulting in a professionally-edited episode for each of the eight team projects. Students are encouraged to watch the videos at www.quant-shop.com very early in the semester for insight into common pitfalls in data science.

Exams: There will be one midterm exam and also a final exam to encourage students to review the material at the end of the course. Both will be largely (but not exclusively) based on the exercises from the textbook. Both exams will be written and in person for all students.

Grading: Grades will be assigned according to the following scale:

- Individual Homeworks -- (1) Python data manipulation (10%), (2) experimental analysis (10%), and (3) machine learning (10%) (total 30% of course grade)
- Short quizzes (5% of course grade)
- Midterm exam (15% of course grade)
- Group project (total of 30% of the course grade)
    - Project proposal (10% of grade)
    - Final project report (20% of grade)
- Final exam (20% of course grade)

I reserve the right to modify the distribution formula should this prove necessary.

Positive Class Experience Policies: The class of 2018 exhibited negative and unprofessional behavior with respect to attendance, social media interaction, whining, and grade grubbing. To crack down, I introduced a strike system, where students receive strikes for each infraction. The first strike results in a penalty equal to 10% of the total semester grade, with subsequent strikes losing 20%, 30%, and 40% of the total semester grade, respectively. Observe that no more points remain after the fourth strike, ensuring an F for the semester. Students can earn strikes for:

- Each and every negative/unprofessional comment on Piazza, peer grading or in class.
- Each and every request for a regrade or an extension of a HW or project assignment.
- Poor attendance.

I am pleased that I did not have to award any strikes for the past three offerings, and hope it will be the same this year.

Lectures: I will give formal lectures during almost all class periods, which will be filmed by Zoom. Videos of all lectures from my Fall 2021 course (and earlier courses) are available on my YouTube channel and http://www.data-manual.com.

Rules of the Road

1. The course web page is http://www.cs.stonybrook.edu/~skiena/519. All course handouts and notes will be posted there.
2. For this semester, we will be using Google Classroom. https://classroom.google.com/c/NzAyMTc1MDczOTM1?cjc=fpjvfxp with code fpjvfxp. Make sure you are accessing it with your stonybrook.edu email, not cs.stonybrook.edu!

   On the homepage, click Add  Join class. Enter the code and click Join. You will be receiving and submitting your assignments through this platform as well as any online quizzes.
3. Sign up for the course discussion board https://piazza.com/stonybrook/fall2024/cse519 (NOTE CHANGE) I will be expecting positive and informative interactions this semester. For this semester, we will hope to have the quizzes open the evening before each lecture, and remain open for the first few minutes of class. These quizzes will be hosted on Google Classroom and automatically will be posted at the appropriate time. Submissions after the deadline will be deleted. Try to complete the quiz before class.
4. CSE 519 is intended as an introduction to Data Science, a survey course. If you have already taken more advanced courses this may not be the best use of your time. Look at the textbook and lecture notes to make an informed decision.
5. The course size for CSE 519 is much larger than reasonable for a project-based course. Policies like peer grading, restricted project selection, and regular quizzes are part of my attempts to deal with this. Please be patient and understanding.
6. I do like a lot of interaction and discussion to happen in lecture: indeed the class is large enough that I would be most happy if it ran almost as a recitation section. Read through the textbook in addition to the slides -- it is fair for me to have textbook material on the quizzes and exams.
7. Unfortunately, I will have somewhat more travel this semester than usual, with the dates marked on the syllabus. I apologize, and will do my best to make good alternate plans.
8. I expect a substantial amount of effort on the project. Some of this work may prove mundane (such as data entry), but I will be expecting each group/student to do what they have to do to get the job done.
9. Each team member will grade the performance of other team members at the end of each part of the project, to make sure credit is awarded fairly among the group. Larger project teams need to do better projects to earn the same grade as small teams.
10. A few students may take this course in the hopes of working with me on their 523/524 projects. I make decisions about masters students only after they can send me their

Fall grades in all their classes. Do not contact me about your possible interest before then. Students I advise must prove they can follow advice, and this is part of the test.

11. I try to make lectures fun through jokes and analogies, but always fear saying something that may offend someone in the class. I want everyone to feel comfortable in my classroom. If anything I say bothers you, please come by and tell me so. I will apologize, and then do my best to understand the issue to avoid doing so again.
12. Text/Code-Assistance Policy. Use of code generating/completion tools (ChatGPT, GitHub CoPilot, etc…) is allowed if and only if: (A) the tool is available to all students and (B) you include comments to (1) mark the code segment ("#BEGIN" and "#END") as well as (2) reference the tool name, web address, and the prompt used to get the code:

    ```
    #BEGIN[TOOL][WEBSITE ADDRESS]"prompt"
    Code…
    #END[TOOL]
    ```

    For Example:

    ```
    #BEGIN[ChatGPT GPT-4][https://chat.openai.com/auth/login]"Write python to bubble sort a list of strings"
    for i in range(0,len(lst)-1):
        for j in range(len(lst)-1):
            if(lst[j]>lst[j+1]):
                temp = lst[j]
                lst[j] = lst[j+1]
                lst[j+1] = temp
    return lst
    #END[ChatGPT]
    ```

    You should include the reference even if you alter the code or even if the algorithm returned was in another language. The point is that you are referencing something that informed your code. You should include the entire prompt of what you provided the tool (use multiple lines of comments if necessary). You should not need to use such a tool and it is recommended you attempt the assignment without it.
13. Each student must pursue his or her academic goals honestly and be personally accountable for all submitted work. Representing another person's work as your own is always wrong. Faculty is required to report any suspected instances of academic dishonesty to the Academic Judiciary. Faculty in the Health Sciences Center (School of Health Technology & Management, Nursing, Social Welfare, Dental Medicine) and School of Medicine are required to follow their school-specific procedures. For more comprehensive information on academic integrity, including categories of academic dishonesty please refer to the academic judiciary website at http://www.stonybrook.edu/commcms/academic_integrity/index.html

14. The Department of Computer Science academic dishonesty policy is fully described in https://www.cs.stonybrook.edu/sites/default/files/drupalfiles/basicpage/GraduateAcademicDishonesty.pdf
15. If you have a physical, psychological, medical, or learning disability that may impact your course work, please contact the Student Accessibility Support Center, Stony Brook Union Suite 107, (631) 632-6748, or at sasc@stonybrook.edu. They will determine with you what accommodations are necessary and appropriate. All information and documentation is confidential.
16. Students who require assistance during emergency evacuation are encouraged to discuss their needs with their professors and the Student Accessibility Support Center. For procedures and information go to the following website: https://ehs.stonybrook.edu/programs/fire-safety/emergency-evacuation/evacuation-guide-disabilities and search Fire Safety and Evacuation and Disabilities.
17. Stony Brook University expects students to respect the rights, privileges, and property of other people. Faculty are required to report to the Office of Student Conduct and Community Standards any disruptive behavior that interrupts their ability to teach, compromises the safety of the learning environment, or inhibits students' ability to learn. Faculty in the HSC Schools and the School of Medicine are required to follow their school-specific procedures. Further information about most academic matters can be found in the Undergraduate Bulletin, the Undergraduate Class Schedule, and the Faculty-Employee Handbook.
18. Many grad students experience enormous pressure – professional, personal and financial. If you or anyone you know is feeling anxious or just needs to talk, help is available. Be aware of the Student Counseling Center on campus: (631) 632-6720 or stop in at the Student Health Center at 1 Stadium Road. Visit https://www.stonybrook.edu/caps/
    1-800-GRAD-HLP is a national crisis-line staffed by trained counselors who understand your unique pressures and know how to help. Use it anytime, 24-7. Your life matters. For more information, see http://gradresources.org/crisisline/.