WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Syllabus
STAT 679: Computing in Data Science & Statistics
Spring 2021, 3 Credits

Description This course provides a survey of some of the tools and frameworks that are currently popular among data scientists and statisticians working in both academia and industry. Our focus will be on complementing the tools that students are already familiar with from their previous courses on R. The course will begin with an accelerated introduction to the Python programming language and brief introductions to object-oriented and functional programming. We will then cover some of the scientific computing platforms available in Python, including numpy, scipy and scikit-learn, as well as visualization using matplotlib. We will then turn to discussing collecting data from the web both by scraping and using APIs. The course will conclude with a brief survey of distributed computing platforms, focusing on the Hadoop platform and Google Cloud Platform.

Prerequisites There are no strict prerequisites for this course, but I will assume that your background is equivalent to having taking STAT605 or equivalent. In practice, that means being comfortable using R, familiarity with the UNIX/Linux command line, and prior experience with distributed computing resources (e.g., using the Slurm scheduler or similar). Students who have no prior programming experience are discouraged from enrolling.

Instructor

Keith Levin, `kdlevin@wisc.edu`
Instructor office hours: Wednesdays, 10am to 11am and 9pm to 10pm on BB Collaborate, or by appointment. All times are Madison local time.

Meetings

Lecture: lecture videos will be released weekly on Canvas and on the course webpage, `http://pages.stat.wisc.edu/~kdlevin/teaching/Spring2021/STAT679/index.html`. Owing to remote instruction and the fact that students are enrolled from all corners of the globe this semester, there are no required meeting times of this course. Students are encouraged to attend office hours to ask questions about the readings, lectures and homework assignments.

Textbook, Readings & Online Resources

There is no physical textbook required for this course. In the first half of the course, we will make frequent reference to Allen B. Downey's *Think Python*, available at `http://greenteapress.com/wp/think-python-2e/` and to Charles Severance's *Python for Informatics*, available at `https://www.py4e.com/book`. Other required readings will be made available as we cover relevant material, and supplemental readings will be suggested for those who are interested in learning more.

All class resources will be made available on the course web page, `http://pages.stat.wisc.edu/~kdlevin/teaching/Spring2021/STAT679/index.html`. and on the course Canvas page. Please contact the instructor if any resources are missing from either of these websites. The instructor will make an effort to post slides and demo code alongside lecture videos.

Course Topics

- **Introduction to Python**. Data types. Programming patterns. Classes and objects. Functional programming.
- **Visualization with** `matplotlib`. Basic plotting.
- **Scientific computing in Python**. Introduction to `numpy`, `scipy` and `scikit-learn`.
- **Processing Structured Data**. Regular expressions. Markup languages. Databases and SQL.
- **Retrieving Data with APIs**. HTTP request methods. Installing and using APIs.
- **Big data and distributed processing**. Basics of parallel/cloud computing. The MapReduce framework. Hadoop and Spark.
- **Specifying and training models with** `TensorFlow`. Basics of Google TensorFlow. Function graphs. Symbolic differentiation.

Grading, Homeworks & Late Days

Grades will be based on cumulative performance on a set of approximately twelve homeworks, though the exact number of homework assignments is subject to change. Each homework assignment is worth a given number of points, and grades will be based on a percentage out of the total possible points. Assignments later in the semester will be worth more points, on average, than those earlier in the semester. I reserve the right to curve these scores in the event of skewed class performance. Students may contest their grade on an assignment up to two (2) weeks from the day that an assignment's grades are released, after which grades may not be changed. In order to comply with the registrar's grading schedule, students may not contest any grade more than one (1) week past the grading of the final homework. Homework due dates are strict, and you may turn in work late only with the use of "late days", of which you have seven (7) to use over the course of the semester. For each late day you spend, you may extend the deadline of a homework by up to 24 hours. You may spend multiple late days per homework. Once you have turned in your homework you may not spend more late days to turn in your homework again after the deadline (you may, of course, turn in multiple versions of your homework assignment through Canvas prior to the deadline). The purpose of this late day policy is to give you a way to deal with unexpected circumstances (e.g., illness, family emergencies, job interviews) without having to come to me. Of course, if dire circumstances arise (e.g., long-term illness that causes you to miss multiple weeks of lecture), please speak with me as promptly as possible. **Note:** owing to the university grading schedule, you may not use late days to extend any deadline beyond Friday, May 7th.

Key Dates First lecture: Monday, January 25, 2021
Last lecture: Friday, April 30, 2021
Last homework due: Friday, May 7 by 11:59 p.m. (this due date may not be changed using late days).

Ethics and class policies

Academic misconduct includes such actions as copying code from the web or from your fellow students, providing code to your fellow students, looking up solutions online, turning in assignments from other classes or previous iterations of this course, and hiring others to complete your work for you. You are welcome to discuss homeworks with your classmates, but the work that you turn in must be yours and yours alone, and you must disclose the names of those collaborated with in your homework.
From the Office of Student Conduct and Community Standards:

> [A]cademic misconduct is behavior that negatively impacts the integrity of the institution. Cheating, fabrication, plagiarism, unauthorized collaboration, and helping others commit these previously listed acts are examples of misconduct which may result in disciplinary action.

See `https://conduct.students.wisc.edu/academic-misconduct/` for more information.
Violations of these or other university ethical standards surrounding academic honesty will be met with serious consequences and disciplinary action. At a minimum, cheating on an assignment will result in a 0 for that assignment and the incident will be reported to the appropriate office. At the instructor's discretion, depending on the circumstances, an additional full letter grade may be deducted from the student's final grade in the course.

Accommodations for Students with Disabilities

The University of Wisconsin-Madison supports the right of all enrolled students to a full and equal educational opportunity. The Americans with Disabilities Act (ADA), Wisconsin State Statute (36.12), and UW-Madison policy (Faculty Document 1071) require that students with disabilities be reasonably accommodated in instruction and campus life. Reasonable accommodations for students with disabilities is a shared faculty and student responsibility. Students are expected to inform faculty [me] of their need for instructional accommodations by the end of the third week of the semester, or as soon as possible after a disability has been incurred or recognized. Faculty [I], will work either directly with the student [you] or in coordination with the McBurney Center to identify and provide reasonable instructional accommodations. Disability information, including instructional accommodations as part of a student's educational record, is confidential and protected under FERPA.