CEE 498DS: Data Science, Fall 2020

THIS SYLLABUS IS SUBJECT TO CHANGE! Please check back throughout the course.

Basic Course Information

- Department: Civil and Environmental Engineering
- Title: CEE 498DS: Data Science
- Credits: 3 for Undergraduates, 4 for Graduate Students
- Semester: Fall 2020
- Meeting time and location: 12-1:20 on Tuesdays and Thursdays in 2312 Newmark
- First day of instruction: 8/25/2020
- Last day of instruction: 12/8/2020

Basic Instructor Information

- Instructor: Prof. Christopher Tessum, PhD
- Office: 3213 Newmark Civil Engineering Laboratory
- Office hours: TBD
- Email: ctessum@illinois.edu
- Website: https://cee.illinois.edu/directory/profile/ctessum
- Names and contact information for teaching assistants: TBD

Description of the course

Welcome to CEE Data Science! This semester, you will learn to leverage data to study civil and environmental engineering problems, identify patterns, and make actionable insights. This course combines training in digital and computer tools—including distributed computing, exploratory data analysis, and statistical modeling and deep learning—with application of those tools to civil and environmental engineering issues.

This course differs from other available machine learning and data science courses in that it focuses on civil and environmental engineering problems and the methods used to solve them. In particular, this course emphasizes working with spatial data, which is common in physical science but less common in data science when applied to other disciplines.

By the end of the semester, you will be able to:

1. Use software tools for data processing and visualization, machine learning, and deep learning to
2. Retrieve, manipulate, and analyze data; and
3. Make inferences and predictions about the (built) environment.

This course will help you to gain the skills and tools necessary to make the most of the great increases in the amount and quality of data related to civil and environmental engineering that is being collected and stored.

Because data science methods are used across a number of different industries and instructional materials are readily available, this course will include readings and video lectures from across the internet. We will focus our face-to-face time on learning aspects of civil and environmental data science that differ from data science as used by other fields, and on applying data science concepts to solving physical problems. This course will be structured around semester-long projects; students will

choose project topics at the beginning of the semester and will apply the concepts learned in the class to their projects as the semester progresses.

Prerequisites

- CEE 202;
- CEE300, 330 or 360; and
- CS 101 or equivalent.

Course Structure

This course is structured as a series of modules, with each module containing recorded lectures, readings, and quizzes to be completed before each class meeting. Class meetings will be held on Zoom to go into further depth on the material that was covered in recorded lectures and readings. Near the beginning or the semester, students will choose a topic for a group or individual project, which the will work on throughout the semester, applying the concepts that we learn in class. Additionally, students will complete homework assignments and a midterm and final exam.

Course Requirements and Assessment Overview

- Grades will be assigned based on several types of deliverables:
  - Mini quizzes and assignments on readings and video lectures, due before class: 20% of total grade
  - Quizzes: 5% of final grade
  - Homework problem sets: 15% of final grade
  - Midterm exam: 5% of final grade
  - Final exam: 15% of final grade
  - Course project: 40% of final grade: 5% each for each of 5 checkpoints, 5% for midterm presentation, and 10% for final presentation.
- Graduate students are expected to register for 4 credits and undergraduates are expected to register for 3 credits. Correspondingly, course projects for graduate students are expected to include a machine learning component that is more complex than linear regression, whereas for undergraduates this is optional.
- Letter grades will be assigned according to the following scale:
  - 97-100: A+
  - 94-96.5: A
  - 90-93.5: A-
  - 87-89.5: B+
  - 84-86.5: B
  - 80-83.5: B-
  - 77-79.5: C+
  - 74-76.5: C
  - 70-73.5: C-
  - 67-69.5: D+
  - 64-66.5: D
  - 60-63.5: D-
  - Below 59.5: F

Learning Resources

- Students are expected to bring have use of a laptop for class.

- There is no required textbook to purchase. Course material will draw from a number of sources across the internet.
- Some supplemental textbooks which students may find useful are:
  - Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython
  - Hands-On Machine Learning with Scikit-Learn & TensorFlow

Policies

Inclusive Environment

The effectiveness of this course is dependent upon the creation of an encouraging and safe classroom environment. Exclusionary, offensive or harmful speech (such as racism, sexism, homophobia, transphobia, etc.) will not be tolerated and in some cases subject to University harassment procedures. We are all responsible for creating a positive and safe environment that allows all students equal respect and comfort. I expect each of you to help establish and maintain and environment where you and your peers can contribute without fear of ridicule or intolerant or offensive language.

Accommodations

To obtain disability-related academic adjustments and/or auxiliary aids, students should contact both the instructor and the Disability Resources and Educational Services (DRES) as soon as possible. You can contact DRES at 1207 S. Oak Street, Champaign, (217) 333-1970, or via email at disability@illinois.edu.

Participation

Active participation in the online learning environment is vital to your success in this course. Depending on your course, you may be asked to engage in online discussions and other interactive learning environments that invite your active participation and involvement with other students and your instructor.

Student Commitment

By registering for this online course, you commit to self-motivated study, participation in online course activities, and timely submission of all assignments. Furthermore, you commit to accessing the course website and checking email at least four days per week (daily for 4-week courses), as well as to devoting at least 6–8 hrs./week (16-week course), 12–16 hrs./week (8-week course), or 24–32 hrs./week (4-week course) to preparing for each module and completing the required assignments and readings.

Deadlines

If you are unable to meet a particular deadline, it is your responsibility to make prior arrangements with the instructor for that given week. Otherwise, work submitted later than 1 day late will receive 10% penalty, and work submitted later than 2 days late will not be considered for grading unless consent has been given by the instructor.

Instructor Responses

Instructor Feedback Turnaround Time

Questions posted to the Course Help Discussion Forum generally will be answered within 48 hours. If possible, students are encouraged to answer questions posted by other students to the Course Help Discussion Forum, rather than waiting for an instructor's response.

Assignments submitted online will be reviewed and graded by the course instructor within 5 business days. Exams, essays, and term papers will be graded within 10 business days. If your instructor is unable to meet this timeline, students will be notified.

Contacting the instructor

For the fastest response response, the best way to contact the instructor is by attending office hours or posting questions to the Course Help Discussion Forum.

The instructor will not respond to phone calls. The instructor will respond to email messages within 48 hours of receiving them unless the instructor notifies you ahead of time of an inability to do so. When sending email, include a subject line that identifies the course number and nature of your question. The instructor may not respond to questions sent to him or her that should be posted in the Course Help Discussion Forum. Please don't be offended if you are asked to forward your question to this location.

Responding to the Discussion Forums

The role of the instructor within the discussion forums is to help facilitate discussion by providing probing questions, asking for clarification, and helping solve conflicts as necessary. The instructor will not respond to every post. You are encouraged to share your thoughts, experiences, and ideas with each other as well.

Academic Integrity

Academic dishonesty will not be tolerated. Examples of academic dishonesty include the following:

- Cheating
- Fabrication
- Facilitating infractions of academic integrity
- Plagiarism
- Bribes, favors, and threats
- Academic interference
- Examination by proxy
- Grade tampering
- Non-original works
- Should an incident arise in which a student is thought to have violated academic integrity, the student will be processed under the disciplinary policy set forth in the Illinois Academic Integrity Policy. If you do not understand relevant definitions of academic infractions, contact your instructor for an explanation within the first week of class.

Giving and receiving advice on projects and homework assignments is acceptable and encouraged. However, it is expected that help be given in general terms and in the form of natural language sentences (for example, English) rather than in the form of mathematical equations, algorithms, computer code, or anything else that could be copied and pasted into the recipient's answer. Similarly, students are encouraged to consult the Internet, but copying and pasting code from the Internet and submitting it for the class is not acceptable. The work that each student submits is expected to be their own, written with their own hand or typed on their own keyboard.

Copyright

Student Content

Participants in University of Illinois courses retain copyright of all assignments and posts they complete; however, all materials may be used for educational purposes within the given course. In group projects, only the portion of the work completed by a particular individual is copyrighted by that individual. The University of Illinois may request that students' materials be shared with future courses, but such sharing will only be done with the students' consent. The information that students submit during a course may, however, be used for the purposes of administrative data collection and research. No personal information is retained without the students' consent.

Non-student Content

Everything on this site and within University of Illinois courses is copyrighted. The copyrights of all non-student work are owned by the University of Illinois Board of Trustees, except in approved cases where the original creator retains copyright of the material. Copyrights to external links are owned by or are the responsibility of those external sites. Students are free to view and print material from this site so long as

- The material is used for informational purposes only.
- The material is used for noncommercial purposes only.
- Copies of any material include the respective copyright notice.
- These materials may not be mirrored or reproduced on non-University of Illinois websites without the express written permission of the University of Illinois Board of Trustees. To request permission, please contact the academic unit for the program.

Student Behavior

Student Conduct

Students are expected to behave in accordance with the penal and civil statutes of all applicable local, state, and federal governments, with the rules and regulations of the Board of Regents, and with university regulations and administrative rules.

For more information about the student code and handbook, see the CITL course policies page.

Netiquette

In any social interaction, certain rules of etiquette are expected and contribute to more enjoyable and productive communication. The following are tips for interacting online via email or discussion board messages, adapted from guidelines originally compiled by Chuq Von Rospach and Gene Spafford (1995):

- Remember that the person receiving your message is someone like you, deserving and appreciating courtesy and respect.
- Be brief; succinct, thoughtful messages have the greatest effect.
- Your messages reflect on you personally; take time to make sure that you are proud of their form and content.
- Use descriptive subject headings in your emails.
- Think about your audience and the relevance of your messages.
- Be careful when you use humor and sarcasm; absent the voice inflections and body language that aid face-to-face communication, internet messages are easy to misinterpret.

- When making follow-up comments, summarize the parts of the message to which you are responding.
- Avoid repeating what has already been said; needless repetition is ineffective communication.
- Cite appropriate references whenever using someone else's ideas, thoughts, or words.

Communications

Daily Contact

Your daily contact should be via the discussion forums in our Learning Management System and via email.

Course Questions

Questions pertaining to the course should be posted in our Course Help Discussion Forum. You can get to this forum from the course home page. Posting questions here allows everyone to benefit from the answers. If you have a question, someone else is probably wondering the same thing. Anyone submitting a question via email will be directed to resubmit the question to the Course Help Discussion Forum. Also, participants should not hesitate to answer questions posed by peers if they know the answers and the instructor has not yet responded. This not only expedites the process but also encourages peer interaction and support.

Personal and Grade-Related Questions

Questions of a personal nature should first be sent to the instructor's email address (listed on the Instructor Information page). When sending email, include a subject that identifies the course number and nature of your question.

Emergencies

If you have an emergency that will keep you from participating in the course, please notify your instructor by using the instructor's email address (listed on the Instructor Information page). Provide callback information in your email (if necessary). You should also notify your program director of any emergencies.

Zoom

Zoom is a tool that allows multiple people to join together simultaneously via a computer to text chat, audio chat, video chat, collaborate on a digital whiteboard, and even share their computer desktops with one another. The instructor's Virtual Office (when available) makes use of Zoom.

Instructor's Virtual Office

Another way to communicate with the instructor is to make use of the Virtual Office hours through the Zoom Interface. The instructor will be available for office hours via Zoom on the dates and during the times listed on the Virtual Office page in the Syllabus.

Announcements

The Announcements forum serves as a way for your instructor and University of Illinois administrators to make announcements within our online learning environment. Announcements posted here will also be sent to your Illinois email address, so be sure to check your email or the Announcements forum at least once a day to see whether any new announcements have been made.

Sexual Misconduct Policy and Reporting

The University of Illinois is committed to combating sexual misconduct. Faculty and staff members are required to report any instances of sexual misconduct to the university's Title IX and Disability Office. In turn, an individual with the Title IX and Disability Office will provide information about rights and options, including accommodations, support services, the campus disciplinary process, and law enforcement options.

A list of the designated university employees who, as counselors, confidential advisors, and medical professionals, do not have this reporting responsibility and can maintain confidentiality, can be found in the Confidential Resources section. Other information about resources and reporting is available at wecare.illinois.edu.

Student Wellness Resources

The University of Illinois strives to promote student success through the support of student psychological and emotional well-being. Please take advantage of the resources listed on the Student Affairs website.

Course Schedule

- **Week 1: Open Reproducible Science.** Students will learn how to structure a computational workflow for scientific analysis, including version control, documentation, data provenance, and unit testing.

    - Before second class: *The Introduction to Earth Data Science Textbook* Section 1
    - In class: Lecture on data science workflow best practices:
        - Git/Github
        - Jupyter
        - Unit testing
        - Software 1.0 and 2.0

- **Week 2: Data science for the physical environment.** Students will learn the types of environmental questions that data science and machine learning can help to answer, and begin to think about topics for course projects.

    - Before class, read/browse through existing and proposed data science projects and brainstorm ideas for course projects.
        - Tackling Climate Change with Machine Learning
        - PANGEO Geoscience Use Cases
        - Kaggle data science competitions
        - Earth Engine Case Studies
        - OpenAQ.org
        - Array of Things
        - CACES air quality data
        - Others from other CEE disciplines (TBD)
    - In class:
        - Students present on potential projects, with class discussion

- **Week 3: Programming review**: Students will refresh their skills in basic Python programming

    - Before class, complete the Google python class and complete python assignment on prarielearn.
    - In class:
        - Python & Jupyter exercises and troubleshooting

- **Week 4: Big data**: Students will explore opportunities and challenges related to large databases

    - Before class, view lectures on numerical computing in Python and complete numerical python assignment in prarielearn.
    - In class:
        - Lecture and demonstration:
            - Cloud / High-performance computing
            - Pangeo
            - Earth engine
        - Practice and discussion
        - Choose project groups and topics

- **Week 5: Spring break**

- **Week 6: Exploratory data analysis (EDA)** Students will learn how to explore and process an unfamiliar dataset.

    - Before class: Watch mlcourse.ai video lectures on exploratory data analysis and visualization and work through accompanying notebooks 1, 2.1 and 2.2
    - In class:
        - Lecture: Statistics review
        - EDA group exercises
    - Students should begin working on EDA for their projects, which will be due in Week 9.

- **Week 7: Geospatial data**: Students will learn about processing spatial data, which is common in physical data science

    - Before class, students should work through the geopandas tutorial and complete a related assignment on prairielearn.
    - In class lecture:
        - raster vs. vector formats
        - joins and boolean operations
    - Spatial statistics homework assigned

- **Week 8: Spatial statistics**: Students will learn how to perform statistical analysis of spatial data.

    - Before class, students should review the PySAL library and notebooks and complete an assignment brainstorming how one or more of these algorithms could be used for their project.
    - In class:
        - Lecture: Spatial statistics (spatial autocorrelation, Modifiable areal unit problem, kriging)
        - Discussion: How spatial statistics can be applied to this semester's student projects

- **Week 9: Mid-way project presentations**: Students should be able to access, characterize, and visualize the data for their projects by this point.

    - Written project EDA report due
    - Oral presentations of EDA results and plan for remainder of project.

- **Week 10–10.5: Supervised learning**: Students will learn what supervised machine learning is and how it can help answer environmental questions

    - Spatial statistics homework due
    - Before class, students should complete the Google Machine Learning Crash Course sections on framing machine learning, gradient descent, optimization, tensorflow, generalization, training and testing, and validation, and the accompanying quiz on prairielearn.
    - In class, we will work through some applications to environmental data and discuss how supervised learning can be applied to student projects.
    - Machine learning homework assigned.

- **Week 11: Unsupervised learning**: Students will learn about basic unsupervised learning algorithms and how they can be used on environmental applications.

    - Before class, view Andrew Ng's lectures on unsupervised learning and clustering, work through the mlcourse.ai workbook, and complete the quiz on prarielearn.
    - In class, we will work through some applications to environmental data and discuss how supervised learning can be applied to student projects.

- **Week 12: Deep learning**: Students will learn about deep learning, the opportunities and drawbacks it presents, and applications to environmental problems.

    - Before class, students should complete the Google Machine Learning Crash Course sections on Introduction to Neural Networks, Training Neural Networks, and Multi-Class Neural Networks and complete the prairielearn quiz.
    - In class:
        - Lecture on hyperparameter optimization and inductive biases
        - Discuss applications to student projects

- **Week 13–14: Projects**: Students will work on their course projects

    - During class time we will work together to troubleshoot student course projects. Students can sign up for time slots where they can present a problem they have encountered and the class will discuss possible solutions.
    - Machine learning homework due

- **Week 15–16: Final exam; final project**: Students should have completed a project where they access and explore a civil or environmental dataset and use it to answer a scientific question.

    - Written report due
    - Oral presentations to class
    - Comprehensive final exam

Modules

Module 1: Open Reproducible Science

Overview

This module covers tools and methods for ensuring your work is correct, understandable, and reproducible.

Objectives

You will learn how to structure a computational workflow for scientific analysis, including version control, documentation, data provenance, and unit testing.

Readings and Lectures

Develop your answers to the discussion questions below while completing the readings and lectures. * Introduction to Earth Data Science Chapter 1 * Introduction to Earth Data Science Chapter 2 * Introduction to Earth Data Science Chapter 3 * Andrej Karpathy: Software 2.0 * NOVA: What Makes Science True?

Discussion

This module includes a discussion section to help you understand by articulating how the module content could be useful in your professional life. Consider the following questions:

- What does it mean to practice open and reproducible science, and how could you apply it to your academic or professional life?
- Although the readings and NOVA video mainly refer to academic science, how could they be relevant to science practiced in industry?
- For the "Software 2.0" essay: What is the author talking about? Instead of trying to understand every detail in the essay (although by the end of the semester you should be able to understand a lot of it), focus on the main message: What is Software 2.0 and what are its implications for how science is carried out? Log in to the module discussion forum and make one initial post and two responses. Refer to the Discussion Forum Instructions and Rubric for instructions how to compose posts to the discussion forum, and how they will be graded. **All posts for this module are due by 0001-01-01 00:00:00 +0000 UTC.**

Quiz

The quiz for Module 1 covers the required readings and lectures and is available here. **The quiz for this module is due by 0001-01-01 00:00:00 +0000 UTC.**

Assessment Instructions and Rubrics

Discussion Forum Instructions and Rubric

This section describes how to participate in the discussion forum, and how your posts will be graded.

Initial Post

In the Discussion Forum for the module, compose an initial post that responds to at least one of the questions above. Your initial post is your opportunity to engage with the prompt in a way that is unique to you. Some ways to accomplish that include:

- Connect with the prompt in a personal way by incorporating personal anecdotes.
- Reflect on any potential biases you may have based on your experiences.
- Consider any potential biases in the information presented in the prompt itself. Be open to different points of view by providing some suggestions of what those might be.

Your initial post must meet the following requirements:

- Include at least **200 words**, excluding any references.
- Use appropriate evidence from the readings and lessons to support your claims and judgments.

Response Posts

Post at least 2 responses in the same thread. Your replies should stimulate more in-depth discussion about the topic. Some ways to accomplish that include:

- Clarify and/or extend your peers' line of thinking.
- Compare/contrast their views on the topic with your own.
- Suggest/question what explanation(s) you think your peers might be missing that could strengthen their arguments.
- End your response with a question to further the dialogue.

Your response posts should meet the following requirements:

- Include at least **50 words**, excluding references.
- Use of appropriate evidence from the readings and lessons to support your claims and judgments.

Submission Directions

- Access the discussion board and begin a new thread.
- Response Post: Select the title of any post to review it and read any replies already submitted. Click Reply next to any post to compose a reply.

Evaluation

This activity is worth 40 points: 20 points for your initial post and 10 points for each response post. Please see the rubric below for detailed information about how your posts will be graded.

Discussion Forum Rubric

Contributions	Description	Initial Post Points Assigned	Response Points Assigned
Provocative	Response goes beyond simply answering the prompt; attempts to stimulate further thought & discussion	20	10
Substantial	Response provides most of the content required by the prompt, but does not require further analysis of the subject	15	7.5
Superficial	Response provides obvious information without further analysis of the concept; lacks depth of knowledge or reasoning	10	5

Contributions	Description	Initial Post Points Assigned	Response Points Assigned
Incorrect	Response does not accurately address the prompt; rambling and/or without consistency	5	2.5
None	No response provided to the prompt within the associated timeframe	0	0