CEE 498DS: Data Science, Fall 2020

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

Basic Instructor Information

- Instructor: Prof. Christopher Tessum, PhD
- Office: 3213 Newmark Civil Engineering Laboratory
- Office hours: TBD
- email: ctessum@illinois.edu
- 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 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.
- Grades will be assigned according to the attached rubrics.
- Letter grades will be assigned according to the following scale:
  - 90%-100%: A
  - 80%-89%: B
  - 70%-79%: C
  - 60%-69%: D
  - <60%: F

Learning Resources

- Students are expected to bring a laptop to class. Laptops can be loaned from the library.
- There is no required textbook to purchase. Course material will draw from a number of sources across the internet:
  - *The Introduction to Earth Data Science Textbook*, CU Boulder Earth Lab.
  - Video lectures from Google, Udemy, and mlcourse.ai
  - Jupyter notebooks from mlcourse.ai and geopandas
- 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

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