CptS 475: Data Science
Syllabus, Fall 2024

Course Information

Credit Hours: 3
Semester: Fall 2024
Meeting Times: MWF, 9:10–10:00 am
Meeting Location: Bryan Hall 305
Learning Management System: Canvas will be used for the management of this course, including for posting of lecture materials, assignments, announcements, and messages. It will also be used for handling student submissions and instructor and Teaching Assistant feedbacks.

Instructor Information

Instructor: Assefaw Gebremedhin
Office: EME B43
Email: assefaw DOT gebremedhin AT wsu DOT edu
Webpage: https://eecs.wsu.edu/~assefaw/
Instructor Office Hours: Wednesdays 12–1:30 pm
Office hours will be conducted in-person, but I will also make a zoom option available in case you can't make it in person. Zoom meeting information is posted on the Canvas site of the course.

Teaching Assistant: Funso Oje
Email: olufunso DOT oje AT wsu DOT edu
Office: Dana 115
TA Office Hours: Tuesdays and Thursdays 1–2:30 pm
TA Office hours will be conducted in person in Dana 115, but a zoom option is also made available in case you can't make it in person. Zoom meeting information is posted on the Canvas site of the course.

Course Description

Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning computer science, mathematics, statistics, and domain expertise along with a good understanding of the art of problem formulation to engineer effective solutions. The purpose of this course is to introduce students to this rapidly growing field and equip them with some of its fundamental principles and tools as well as its general mindset. The course will use the programming languages R (primarily) and Python.

Topics to be covered include: the data science process, exploratory data analysis, data wrangling, data visualization, linear regression, classification, clustering, principal component analysis, time-series data mining, deep learning, and data and ethics.

The focus in the treatment of these topics is on breadth, rather than depth, and emphasis is placed on integration and synthesis of concepts and their application to solving problems. Necessary theoretical abstractions (mathematical and algorithmic) are introduced as and when needed.

Audience

The course is suitable for upper-level under-graduate students in computer science, engineering, applied mathematics, the sciences, business, and related analytic fields. The course is offered conjoint with a 500-level graduate course (CptS 575).

Prerequisites

Students are expected to: (i) have taken an introductory course in statistics and probability, (ii) have fundamental knowledge of algorithms and good programming experience (equivalent to completing a data structures course such as CptS 223), and (iii) have fundamental knowledge of linear algebra (e.g. eigenvalue/vector computation).

Coursework

The course consists of several elements: lectures (three times a week, 50 min each); a set of assignments; a substantial semester project; one mid-term exam and no final exam. Below is how the coursework and assessment is broken down.

- **Assignments (35%)**. There will be a total of 5 assignments spread through the semester. Each assignment will have one major topic of emphasis. Assignments are to be completed and submitted individually. Each assignment will carry equal weight. Together all assignments account for 35% of final grade.
- **Semester Project (40%)**. Students, working individually or in a team of two, will complete a semester project. A project could take one of several forms: analyzing an interesting dataset using existing methods and software tools; developing a new method; building your own data product; or creating a visualization of a complex dataset.
  Students will be given an opportunity to choose from a list of projects the instructor provides or propose their own project. Guidelines for what constitutes a project will be provided by the instructor. A project will culminate in a written report and a short presentation in class. General guidelines for how to prepare a report will be provided by the instructor. Students are expected to follow the guidelines. Similarly, guidelines for how to prepare and deliver good presentations will be provided by the instructor, and students are expected to follow the guidelines.
- **Exam (23%)**. There will be one mid-term exam designed to complement the assignments and the semester project. The exam is tentatively scheduled to take place in the week of Nov. 4. Final date will be decided later after consulting with the class.
- **Class Participation (2%)**. Active class participation—in discussions during lectures, surveys, and other online discussions, including responding to Participation Question of The Day—is required. Class Participation will count towards 2% of the final grade.

Expectations for Student Effort

For each hour of lecture equivalent, students should expect to have a minimum of two hours of work outside class.

Grading

Letter grades will be given according to the following ranges:

A (93–100%), A- (90–92.99%), B+ (87–89.99%), B (83–86.99%), B- (80–82.99%), C+ (77–79.99%), C (70–76.99%), C- (67–69.99%), D (60–66.99%), F (less than 60%).

Learning Outcomes and Assessment

Student Learning Outcomes. *By the end of the course, students should be able to:*	**Course Topics/Dates.** *The following topics/dates will address this outcome:*	**Evaluation.** *This outcome will be evaluated primarily by:*
• Describe what Data Science is and the skill sets needed	What is Data Science? (week 1);	Assignments; Exam
• Describe the Data Science Process	EDA and the Data Science Process (week 3)	Assignments; Exam; Project
• Use R (or Python) to carry out statistical modeling and analysis	Intro to R (week 2); Most subsequent topics throughout the semester	Assignments; Project
• Carry out effective exploratory data analysis	EDA (week 3)	Assignments; Project
• Use effective data wrangling approaches to manipulate data	Data Wrangling (week 4, week 5)	Assignments; Project
• Create effective visualization of data (to communicate or persuade)	Data Visualization (week 5, week 6)	Assignments; Project
• Apply machine learning algorithms for predictive modeling	Linear Regression (week 7); Classification (weeks 8 and 9); Deep Learning (week 13, week 15)	Assignments; Project; Exam
• Apply effective methods to assess model performance	Cross-validation (week 10)	Exam; Project
• Apply learning methods to discover patterns, trends and anomalies in data	Unsupervised Learning (week 11); Time Series Data Mining (week 12)	Assignments; Project; Exam
• Reason around ethical and privacy issues in data science conduct, and apply ethical practices	Data and Ethics (week 15)	In-class exercise
• Work effectively in teams on data science projects		Project
• Apply knowledge gained in the course to carry out a project and write a technical report		Project

Detailed Topics and Course Outline

1. Introduction: What is Data Science?
   - Big Data and Data Science; Landscape of perspectives; Skill sets needed
2. Intro to R
   - R basics; R graphics; R Markdown
3. Exploratory Data Analysis and the Data Science Process
   - Basic tools of EDA; Philosophy of EDA; The Data Science Process
4. Data Wrangling
   - Data transformation and manipulation (dplyr); Relational data; Data "tidying" (tidyr)
5. Data Visualization
   - Principles of data visualization; Use of color and design considerations; Web mapping; data visualization and story-telling
6. Overview of Machine Learning
   - Supervised Learning (canonical examples and real world applications);
     Unsupervised Learning (canonical examples and real world applications)
     Reinforcement Learning
7. Linear Regression
   - Simple linear regression; Multiple linear regression; Extensions of the linear model
8. Classification
   - Overview of classification; Logistic regression; Linear Discriminate Analysis; Naive Bayes classifier; K-Nearest Neighbors (KNN); Decision Trees and Random Forest
9. Resampling Methods
   - Cross-validation; The Bootstrap
10. Unsupervised Learning
    - Principal Component Analysis (PCA); K-means clustering; Hierarchical clustering
11. Time Series Data Mining Overview
    - Examples of areas where time series data arise; Distance measures; Algorithms (motif discovery, anomaly detection, segmentation, classification, clustering).
12. Intro to Deep Learning
    - What is deep learning? The perceptron; Activation functions; Building neural networks; Training neural networks; Regularization; Software packages for DL; Convolutional neural networks
13. Data Science and Ethical Issues
    - Discussions on privacy, security, ethics; A look back at Data Science

Weekly Schedule

See Table 1 for a weekly schedule of topics and assignments.

Week	Topics	Assignments/Notes
01 (Aug 19)	What is Data Science?	Assign. 1 out
02 (Aug 26)	Intro to R/Python	Assign. 1 due, Assign. 2 out
03 (Sep 02)	Exploratory Data Analysis	No class 9/2; Assign. 2 due
04 (Sep 09)	Data Wrangling I	Assign. 3 out
05 (Sep 16)	Data Wrangling II, Data Visualization I	Assign. 3 due
06 (Sep 23)	Data Visualization II	Assign. 4 out
07 (Sep 30)	Semester project set-up, Overview of ML	Assign. 4 due, Project proposal out
08 (Oct 07)	Linear Regression	Project proposal due, Assign. 5 out
09 (Oct 14)	Classification I	
10 (Oct 21)	Classification II, Resampling methods	Assign. 5 due
11 (Oct 28)	Unsupervised Learning	Project progress report due
12 (Nov 04)	Time Series Data Mining	Mid-term exam
13 (Nov 11)	Deep Learning (DL)	No class 11/11 (Veterans Day)
14 (Nov 18)	DL II, Ethics, Course wrap-up	In-class exercise
15 (Nov 25)	**Thanksgiving break**	
16 (Dec 02)	Project presentations	
17 (Dec 09)		Final project report due on Dec 10

Table 1: Tentative week-by-week schedule of topics and assignments. The date shown in parenthesis is just the Monday of that week.

Books

There is no required "textbook" for this course. Select chapters from the followings references will be used as starting points for discussions, and they will be supplemented with instructor-developed lecture notes and reading assignments from other sources.

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning.* The book comes in two versions, one version in R (ISLR) and the other in Python (ISLP). The second edition of ISLR was published in 2021 and ISLP was published in 2023. The book is freely available online at: https://www.statlearning.com/.
- Hadley Wickham, Mine Centinkaya-Rundel and Garett Grolemund. *R for Data Science*, Second Edition. Freely available on-line at: https://r4ds.hadley.nz/
- Kevin P. Murphy. *Probablistic Machine Learning: An Introduction.* MIT Press, March 2022. Various info including draft PDF files of chapters available at: https://probml.github.io/pml-book/book1.html.
- Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. *Mining of Massive Datasets.* Third Edition, Cambridge University Press. 2021.
  The book is freely available online at: http://www.mmds.org/.
- Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning.* MIT Press, 2016. ISBN 9780262035613. The book is freely available online at: http://www.deeplearningbook.org

Policies

Conduct

Students are expected to maintain a professional and respectful classroom environment. In particular, this includes:

- silencing personal electronics (non-disruptive devices may be used during class)
- arriving on time and remaining throughout the class.

Use of Generative AI Tools and Technologies

The use of generative AI tools and technologies, such as ChatGPT, to create content is permitted, but it must be fully disclosed as follows:

- If the tool is used to generate **code**, the generated code should be first saved in some way (e.g. in a separate file or commited to a repository) before being adapted with the following statement in the comments of the code: "Generative AI was utilized to generate this code" followed by a description of the functionality that the autogenerated code is fulfilling.
- Next, the code can be adapted and saved again (or commited)—the comments of the code should be updated to provide a summary of the adaptation.
- If the tool is used to generate **text**, at the beginning of the submitted document, include a statement that says "Generative AI was utilized to generate a draft of this text", and then indicate in what way you edited the draft to produce the submitted text.

If you find yourself in a situation that is not listed above and are worried about whether or not it might be considered cheating or if you are uncertain about the need to disclose the use of a particular tool, err on the side of caution and email me (via Canvas). I'll give you a definitive answer.

Correspondence

All class related correspondence with the instructor will be made via Canvas.

Attendance

Regular attendance is expected. While students may miss class for urgent reasons, excessive absences that are not cleared with the instructor will factor into the Class Participation portion of the semester grade.

Missing or Late Work

Submissions will be handled via Canvas. Students are expected to submit assignments by the specified due date and time. Assignments turned in up to 48 hours late will be accepted with a 10% grade penalty per 24 hours late. Except by prior arrangement, missing or work late by more than 48 hours will be counted as a zero.

Missed Exam

There will be only one exam in the course, tentatively scheduled to take place in the week of Nov 4. Final date of the exam will be decided at least two weeks prior to the exam date. Date will be picked in consultation with the class to accomodate as much as possible students' other exam schedules and commitmments. Make-up exam is not allowed if the exam is missed unless there is an extra-ordinary circumstance.

Academic Integrity

Academic integrity is the cornerstone of higher education. You are responsible for reading WSU's Academic Integrity Policy, which is based on Washington State law. If you cheat in your work in this class you will:

- Fail the assignment or exam in which the cheating happened.
- Be reported to the Center for Community Standards.
- Have the right to appeal my decision.
- Not be able to drop the course or withdraw from the course until the appeals process is finished.

If you have any questions about what you can and cannot do in this course, ask me.

If you want to ask for a change in my decision about academic integrity, use the form at the Center for Community Standards website. You must submit this request within 21 calendar days of the decision.

Lauren's Promise: WSU's Commitment to Address Discrimination and Harassment

On October 22, 2018, Lauren McCluskey, 21 years old, was murdered by a man she briefly dated on the University of Utah campus, where she was a student. Lauren was raised in Pullman, Washington. Together with her parents, who are professors at WSU, this university community stands firmly behind Lauren's Promise: **WSU will listen and facilitate support and reporting options if someone is threatening you.**

WSU prohibits discrimination and harassment. This includes discriminatory harassment, hate crimes, sexual discrimination, sex-based harassment, stalking, dating violence, domestic violence, sexual assault, and all types of sexual violence.

If you are in immediate danger, call 911.

If you have experienced or have witnessed discriminatory behavior, you can contact the WSU Compliance and Civil Rights (CCR) and/or the WSU Title IX Coordinator. CCR can provide information on reporting options, including confidential resources available to you, and facilitate supportive measures. To contact CCR:

Online: Online Reporting Form

Email: ccr@wsu.edu

Phone: 509-335-8288

For more information, see the WSU Policy Prohibiting Discrimination and Harassment (Executive Policy 15), WSU Standards of Conduct for Students (Chapter 504-26 WAC), and the WSU Notice of Nondiscrimination.

Reasonable Accomodation

Students with disabilities or chronic medical or psychological conditions can request reasonable accommodations. If you need reasonable accommodations to fully participate in your courses, please go to your campus' Access Center/Services website (see links below). Follow the procedures to request accommodations. You may also contact your campus office to schedule an appointment with an Access Advisor.

The Access Center/Services will notify your instructors of your requested accommodations, but you may need to communicate with your instructors about how some of your accommodations will work (by email, Zoom, or in person).

Contact an Access Advisor on your campus:

- Pullman, WSU Global Campus, Everett, Bremerton, and Puyallup: 509-335-3417, http://accesscenter.wsu.edu, or email access.center@wsu.edu
- Spokane: 509-358-7816, https://spokane.wsu.edu/studentaffairs/access-resources/, or email spokane.access@wsu.edu
- Tri-Cities: 509-372-7352, http://www.tricity.wsu.edu/disability/, or email tricities.AccessServices@wsu.edu
- Vancouver: 360-546-9238, https://studentaffairs.vancouver.wsu.edu/access-center, or email van.access.center@wsu.edu

Emergencies on Campus

To receive emergency alerts on your phone or by email, click on the link to the page of your campus that you can access via this university syllabus website: https://syllabus.wsu.edu/university-syllabus/. These alerts may include information about active shooter situations and severe weather.

In case of an active shooter, follow these ideas: "Run, Hide, Fight".

In any emergency, remain ALERT by observing and paying attention to WSU emergency alerts. ASSESS your specific situation, and ACT to ensure your own safety and the safety of others if you are able.

Other Policies in the University Syllabus

Students are responsible for reading and understanding all university-wide policies and resources pertaining to all courses (for instance: accommodations, care resources, policies on discrimination or harassment), which can be found in the university syllabus website: https://syllabus.wsu.edu/university-syllabus/.

Academic Dates and Deadlines

Students are encouraged to refer to the academic calendar often to be aware of critical deadlines throughout the semester. The academic calendar can be found at http://registrar.wsu.edu/academic-calendar.

Changes

This syllabus is subject to change. Updates will be posted on the course website.