CptS 475: Data Science
Syllabus

Course Information
Credit Hours: 3
Semester: Fall 2020
Meeting times and location: MWF, 9:10–10:00am, via Zoom.

Course management system: The course website will be used to post basic course information, including this syllabus, and course related resources. Otherwise the course management system Blackboard will be used for posting lecture material, assignments, announcements, and messages; and for handling student submissions and instructor feedbacks.

Instructor Information
Instructor: Assefaw Gebremedhin
Email: assefaw DOT gebremedhin AT wsu DOT edu
Webpage: www.eecs.wsu.edu/~assefaw

Instructor Office Hours: Wednesdays 10:30–12pm, or by appointment. Office hours will be conducted via Zoom. Zoom meeting information is posted on the Blackboard site of the course.

Teaching Assistant: Helen Catanese
Email: helen DOT catanese AT wsu DOT edu
TA Office Hours: Tuesdays and Wednesdays 3–4:30pm, via Zoom. Zoom meeting information is posted on the Blackboard 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 basic 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, linear regression, classification, clustering, principal component analysis, data visualization, time-series data mining overview, 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.

Course Delivery
The class will be taught remotely via Zoom and will happen in a synchronous mode of learning. That is to say, the lectures will happen (live) MWF 9:10–10AM. The lectures will be recorded and
the video recording will made available at the course’s Blackboard site for later review. This will be the case for every single lecture.

**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.

**Prerequisites**

Students are expected to: (i) have taken an introductory course in statistics and probability, (ii) have basic knowledge of algorithms and reasonable programming experience (equivalent to completing a data structures course such as CptS 223), and (iii) have some familiarity with basic 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 about 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 in teams of two or three, will complete a semester project. A project could take one of several forms: analyzing an interesting dataset using existing methods and software tools; 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 (5-min) presentation in class. General guidelines for how to prepare a report will be provided by the instructor. Similarly, guidelines for how to prepare and deliver good presentations will be provided by the instructor.

- **Exam (20%)**. There will be one mid-term exam designed to complement the assignments and the semester project. The exam will be take-home and will be submitted electronically via the Blackboard page of the course. The exam is tentatively scheduled for the week of November 9.

- **Participation (5%)**. Active class participation (in discussions during lectures, surveys, and other online discussions) is required. It will count towards 5% of the final grade.
### Learning Outcomes and Assessment

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

### 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%).

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
   - Telling story with data; Choosing tools to visualize data; Visualizing patterns over
time; Visualizing proportions; Visualizing relationships; Visualizing text information;
Ascombe’s quartet; Tufte’s visualization aesthetic.
6. Overview of Machine Learning
   - Supervised Learning (canonical examples and real world applications);
Unsupervised Learning (canonical examples and real world applications)
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
Books
There is no required “textbook” for this course. Select chapters from the followings references will be used as starting points for discussions, but they will be supplemented with instructor-developed lecture notes and reading assignments from other sources. Video recording of the lectures and other lecture notes and reading material will be made available on the OSBLE+ page of the course as the course proceeds.


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

Policies
Conduct
Students are expected to maintain a professional and respectful virtual classroom environment. In particular, this includes:
  - muting when not speaking
  - using the chat feature only for purposes related to the class
  - joining the meeting on time and remaining throughout the class

Correspondence
All class related correspondence with the instructor will be made via email. I will check messages on a regular basis, and will do my best to respond promptly.

Attendance
Regular attendance is expected. While students may miss class for urgent reasons or internet issues, excessive absences that are not cleared with the instructor will factor into the Class Participation portion of the semester grade.
Table 1: Updated week-by-week schedule of topics and assignments. The date shown in parenthesis is just the Monday of that week.

### Missing or late work
Submissions will be handled via the Blackboard page of the course. 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.

### Academic Integrity
Academic integrity is the cornerstone of higher education. As such, all members of the university community share responsibility for maintaining and promoting the principles of integrity in all activities, including academic integrity and honest scholarship. Academic integrity will be strongly enforced in this course. Any student who violates the University’s standard of conduct relating to academic integrity will receive an F as a final grade in this course, will not have the option to withdraw from the course pending an appeal and will be reported to the Office of Student Standards and Accountability.

Cheating includes, but is not limited to, plagiarism and unauthorized collaboration as defined in the Standards of Conduct for Students, WAC 504-26-010(3). You can learn more about Academic Integrity on the WSU campus at [http://conduct.wsu.edu](http://conduct.wsu.edu). If you have any questions about what is and is not allowed in this course, you should ask course instructors before proceeding.
Students with Disabilities
Reasonable accommodations are available for students with a documented disability. If you have a disability and need accommodations to fully participate in this class, please either visit or call the Access Center (Washington Building 217; 509-335-3417) to schedule an appointment with an Access Advisor. All accommodations must be approved through the Access Center. For more information, consult the webpage http://accesscenter.wsu.edu or email at Access.Center@wsu.edu.

COVID-19 Policy
Students are expected to abide by all current COVID-19 related university policies and public health directives, which could include wearing a cloth face covering, physically distancing, self-attestations, and sanitizing common use spaces. All current COVID-19 related university policies and public health directives are located at https://wsu.edu/covid-19/.

Accommodation for Religious Observances or Activities
Washington State University reasonably accommodates absences allowing for students to take holidays for reasons of faith or conscience or organized activities conducted under the auspices of a religious denomination, church, or religious organization. Reasonable accommodation requires the student to coordinate with the instructor on scheduling examinations or other activities necessary for course completion. Students requesting accommodation must provide written notification within the first two weeks of the beginning of the course and include specific dates for absences. Approved accommodations for absences will not adversely impact student grades. Absence from classes or examinations for religious reasons does not relieve students from responsibility for any part of the course work required during the period of absence. Students who feel they have been treated unfairly in terms of this accommodation may refer to Academic Regulation 104 – Academic Complaint Procedures.

Important 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.