CptS 575: Data Science
Syllabus, Fall 2021

Course Information
Credit Hours: 3
Semester: Fall 2021
Meeting times and location: MWF, 9:10–10:00am, CARP 102.
Learning Management System: Canvas will be used for the management of this course, including for posting of lecture material, assignments, announcements, and messages, and for handling student submissions and instructor feedbacks.
The public course website will be used to post basic course information, including this syllabus.

Instructor Information
Instructor: Assefaw Gebremedhin
Office: EME B43
Email: assefaw DOT gebremedhin AT wsu DOT edu
Webpage: www.eecs.wsu.edu/~assefaw

Instructor Office Hours: Wednesdays 10:30–12pm, or by appointment.

Teaching Assistant: Shruti Patil
Email: shrutisunil DOT patil AT wsu DOT edu
Office: Dana 115
TA Office Hours: Tuesdays and Wednesdays 3–4:30pm

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, 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 graduate students in computer science, engineering, applied mathematics, the sciences, business, and related analytic fields.

The course is offered conjoint with a 400-level 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 survey paper; a substantial semester project; and one mid-term exam (no final exam). Below is how the coursework and assessment is broken down.

- **Assignments (25%)**. 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 25% 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; developing new data science methods; careful performance evaluation of known methods; 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.

- **Survey Paper (15%)**. Each student, individually, will write a survey paper further exploring a specific topic related to the course content. The topic will be chosen in consultation with the instructor. The background material for the paper may be drawn from journal/conference literature reflecting recent research, books, or reports. The format of the paper will be similar to typical journal/conference survey papers. The length of the paper will depend on the nature of the topic chosen, but will typically be around 10 pages.

- **Exam (17%)**. There will be one mid-term exam designed to complement the assignments and the semester project. The exam is tentatively scheduled for the week of November 8.

- **Class participation (3%)**. Active class participation (in discussions during lectures, surveys, and other online discussions) is required. It will count towards 3% of the final grade.
### Learning Outcomes and Assessment

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>By the end of the course, students should be able to:</td>
<td>The following topics/dates will address this outcome:</td>
<td>This outcome will be evaluated primarily by:</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>Intro to R (week 2); Most subsequent topics throughout the semester</td>
<td>Assignments; Project</td>
</tr>
<tr>
<td>• Carry out exploratory data analysis</td>
<td>EDA (week 3)</td>
<td>Assignments; Project</td>
</tr>
<tr>
<td>• Use effective data wrangling approaches to manipulate data</td>
<td>Data Wrangling (week 4, week 5)</td>
<td>Assignments; Project</td>
</tr>
<tr>
<td>• Create effective visualization of data (to communicate or persuade)</td>
<td>Data Visualization (week 5, week 6)</td>
<td>Assignments; Project</td>
</tr>
<tr>
<td>• Apply machine learning algorithms for predictive modeling</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 effective methods to assess model performance</td>
<td>Cross-validation (week 10)</td>
<td>Exam; Project</td>
</tr>
<tr>
<td>• Apply learning methods to discover patterns, trends and anomalies in data</td>
<td>Unsupervised Learning (week 11); Time Series Data Mining (week 12)</td>
<td>Assignments; Project; Exam</td>
</tr>
<tr>
<td>• Reason around ethical and privacy issues in data science conduct, and apply ethical practices</td>
<td>Data and Ethics (week 15)</td>
<td>In-class exercise</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; Tuft e’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
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.

- Avrim Blum, John Hopcroft and Ravindran Kannan. *Foundations of Data Science*. (Note: this is a book currently being written by the three authors. The authors have made a draft of their notes for the book available online https://www.cs.cornell.edu/jeh/book%20notes%20March%202019.pdf. The material is intended for a modern theoretical course in computer science.)

**Weekly Schedule**

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

**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.

**Correspondence**

All class related correspondence with the instructor will be made via Canvas.
<table>
<thead>
<tr>
<th>Week</th>
<th>Topics</th>
<th>Assignments/Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 (Aug 23)</td>
<td>What is Data Science?</td>
<td>Assign. 1 out</td>
</tr>
<tr>
<td>02 (Aug 30)</td>
<td>Intro to R/Python</td>
<td>Assign. 1 due, Assign. 2 out</td>
</tr>
<tr>
<td>03 (Sep 06)</td>
<td>Exploratory Data Analysis</td>
<td>No class 9/6; Assign. 2 due</td>
</tr>
<tr>
<td>04 (Sep 13)</td>
<td>Data Wrangling I</td>
<td>Assign. 3 out</td>
</tr>
<tr>
<td>05 (Sep 20)</td>
<td>Data Wrangling II, Data Visualization I</td>
<td>Assign. 3 due</td>
</tr>
<tr>
<td>06 (Sep 27)</td>
<td>Data Visualization II, Semester project set up</td>
<td>Assign. 4 out, Project proposal out</td>
</tr>
<tr>
<td>07 (Oct 04)</td>
<td>Overview of ML, Linear Regression I</td>
<td>Assign. 4 due, Survey paper call out</td>
</tr>
<tr>
<td>08 (Oct 11)</td>
<td>Linear Regression II, Classification I</td>
<td>Project proposal due, Assign. 5 out</td>
</tr>
<tr>
<td>09 (Oct 18)</td>
<td>Classification II</td>
<td>Assign. 5 due, Survey paper topic due</td>
</tr>
<tr>
<td>10 (Oct 25)</td>
<td>Resampling methods</td>
<td></td>
</tr>
<tr>
<td>11 (Nov 01)</td>
<td>Unsupervised Learning</td>
<td>Project progress report due</td>
</tr>
<tr>
<td>12 (Nov 08)</td>
<td>Time Series Data Mining</td>
<td>Mid-term</td>
</tr>
<tr>
<td>13 (Nov 15)</td>
<td>Deep Learning (DL)</td>
<td></td>
</tr>
<tr>
<td>14 (Nov 22)</td>
<td>Thanksgiving break</td>
<td></td>
</tr>
<tr>
<td>15 (Nov 29)</td>
<td>DL II, Ethics, Wrap-up</td>
<td>In-class Exercise</td>
</tr>
<tr>
<td>16 (Dec 06)</td>
<td>Project presentations</td>
<td>Survey paper due</td>
</tr>
<tr>
<td>17 (Dec 13)</td>
<td></td>
<td>Final project report due on Dec 13</td>
</tr>
</tbody>
</table>

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

**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 November 8. Exact 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.

**COVID-19 Policy**

Students are expected to abide by all current COVID-19 related university policies and public health directives. These directives may be adjusted to respond to the evolving COVID-19 pandemic. Directives may include, but are not limited to, compliance with WSU’s COVID-19 vaccination
policy, wearing a cloth face covering, physically distancing, and sanitizing common use spaces. All current COVID-19 related university policies and public health directives are located at https://wsu.edu/covid-19/. Students who do not comply with these directives may be required to leave the classroom; in egregious or repetitive cases, student non-compliance may be referred to the Center for Community Standards for action under the Standards of Conduct for Students.

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. Students who violate WSU’s Academic Integrity Policy (identified in Washington Administrative Code (WAC) 504-26-010(4) will fail the assignment implicated, will not have the option to withdraw from the course pending an appeal, and will be reported to the Center for Community Standards.

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). Read and understand all of the definitions of cheating given here http://app.leg.wa.gov/WAC/default.aspx?cite=504-26-010. If you have any questions about what is and is not allowed, ask your course instructor.

If you wish to appeal an instructor’s decision relating to academic integrity, please use the form available at http://communitystandards.wsu.edu. Make sure you submit your appeal within 21 calendar days of the instructor’s decision.

Students with Disabilities

Reasonable accommodations are available for students with documented disabilities or chronic medical conditions. If you have a disability and need accommodations to fully participate in this class, please visit your campus Access Center website (websites listed below) to follow published procedures to request accommodations. Students may also call or email the Access Center to schedule an appointment with an Access Advisor. All disability related accommodations are to be approved through the Access Center. It is a university expectation that students with approved accommodations visit with instructors (in person or via Zoom) within two weeks of requesting their accommodations to discuss logistics.

For more information contact a Disability Specialist on your home campus:

- Pullman or WSU Online: 509-335-3417 http://accesscenter.wsu.edu, Access.Center@wsu.edu
- Spokane: https://spokane.wsu.edu/studentaffairs/access-resources/
- Tri-Cities: http://www.tricity.wsu.edu/disability/
- Vancouver: 360-546-9138; https://studentaffairs.vancouver.wsu.edu/student-wellness-center/access-center

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. See also Academic Regulation 82, available at https://registrar.wsu.edu/academic-regulations/.

Safety and Emergency Notification
Classroom and campus safety are of paramount importance at Washington State University, and are the shared responsibility of the entire campus population. WSU urges students to follow the “Alert, Assess, Act” protocol for all types of emergencies and “Run, Hide, Fight” response for an active shooter incident. Remain ALERT (through direct observation or emergency notification), ASSESS your specific situation, and act in most appropriate way to assure your own safety (and the safety of others if you are able).

Please sign up for emergency alerts on your account at MyWSU. For more information on this subject, campus safety and related topics, please view the FBI’s Run, Hide, Fight video (https://www.fbi.gov/about-us/cirg/active-shooter-and-mass-casualty-incidents/run-hide-fight-video) and visit the WSU safety portal (https://faculty.wsu.edu/classroom-safety).

Full details can be found at https://provost.wsu.edu/classroom-safety/.

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.