CptS 475: Data Science
Syllabus

Course Information
Credit Hours: 3
Semester: Fall 2018
Meeting times and location: MWF, 9:10–10:00, CUE 319
Course website: https://scads.eecs.wsu.edu/index.php/data-science/

The course website will be used to post relevant course material—including this syllabus—and course related resources. Additionally, the online portal OSBLE+ (https://plus.osble.org) will be used for posting lecture material, assignments, announcements, and messages; and for handling student submissions and instructor feedbacks.

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

Office hours: Wednesdays 10:30–12pm, or by appointment.
Teaching Assistant: Helen Catanese
Email: helen DOT catanese AT wsu DOT edu
Office: Dana 115
Office Hours: Wednesdays 2–4pm

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, using primarily the statistical computing language R.

Topics to be covered include: the data science process, exploratory data analysis, data wrangling, linear regression, classification, clustering, principal components analysis, data visualization, 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.
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).

Course Work

The course consists of several elements: lectures (three times a week, 50 min each) that incorporates labs as needed; a set of assignments; a substantial semester project; and an exam. Below is how the coursework and assessment is broken down.

- **Assignments (30%)**: There will be a total of 4 to 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 30% of final grade.

- **Semester Project (30%)**: 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 presentation (a poster and short oral presentation in class). General guidelines for how to prepare a report will be provided by the instructor.

- **Exam (30%)**: There will be one exam designed to a) cover most of the material in class and b) complement the assignments and semester project.

- **Class Participation (10%)**: Attendance and active class participation (in discussions, in-class exercises and thought experiments) is required. It will count to 10% 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

<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 4); Intro to R (week 2); Most subsequent topics throughout the semester; Linear Regression (week 6); Classification (week 7); Tree-based methods (week 11); Deep Learning (week 13); Cross-validation (week 8); Unsupervised Learning (week 9); Time Series Data Mining (week 12); Data Visualization (week 10)</td>
<td>Assignments; Exam; Assignments; Exam; Project; Assignments; Project; Exam; Assignments; Project; Exam</td>
</tr>
<tr>
<td>• Describe the Data Science Process</td>
<td>EDA (week 4); Data Wrangling (week 5)</td>
<td>Assignments; Project; Assignments; Project</td>
</tr>
<tr>
<td>• Use R to carry out statistical modeling and analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Carry out exploratory data analysis</td>
<td>EDA (week 4)</td>
<td>Assignments; Project</td>
</tr>
<tr>
<td>• Use effective data wrangling approaches to manipulate data</td>
<td>Data Wrangling (week 5)</td>
<td>Assignments; Project</td>
</tr>
<tr>
<td>• Apply machine learning algorithms for predictive modeling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Apply effective resampling methods to assess model performance</td>
<td>Cross-validation (week 8)</td>
<td>Exam; Project</td>
</tr>
<tr>
<td>• Apply learning methods to discover patterns, trends and anomalies in data</td>
<td>Unsupervised Learning (week 9); Time Series Data Mining (week 12)</td>
<td>Assignments; Project; Exam</td>
</tr>
<tr>
<td>• Create effective visualization of data (to communicate or persuade)</td>
<td>Data Visualization (week 10)</td>
<td>Assignments; Project</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>
Detailed Topics and Course Outline

1. Introduction: What is Data Science?
   - Big Data and Data Science hype – and getting past the hype
   - Landscape of perspectives
   - Skill sets needed

2. Intro to R
   - R basics
   - R graphics
   - R Markdown

3. Overview of Machine Learning
   - Supervised Learning (canonical examples and real world applications)
   - Unsupervised Learning (canonical examples and real world applications)

4. Exploratory Data Analysis and the Data Science Process
   - Basic tools (plots, graphs and summary statistics) of EDA
   - Philosophy of EDA
   - The Data Science Process

5. Data Wrangling
   - Data transformation and manipulation (dplyr)
   - Data “tidying” (tidyr)

6. Linear Regression
   - Simple linear regression
   - Multiple linear regression
   - Extensions of the linear model
   - Lab session on linear regression

7. Classification
   - Overview of classification
   - Logistic regression
   - K-Nearest Neighbors (KNN)
   - Lab session on classification methods

8. Resampling Methods
   - Cross-validation
   - The Bootstrap

9. Unsupervised Learning
   - Principal Components Analysis (PCA)
   - K-means clustering
   - Hierarchical clustering
   - Lab session on PCA and clustering

10. Data Visualization
    - Telling story with data
    - Choosing tools to visualize data
    - Visualizing patterns over time
    - Visualizing proportions
- Visualizing relationships
- Visualizing text information

11. Tree-based Methods
   - Decision Trees
   - Bagging, Random Forests and Boosting

12. Time Series Data Mining Basics
   - Examples of areas where time series data arise
   - Distance measures
   - Algorithms (motif discovery, anomaly detection, segmentation, classification, clustering)
   - Tools: Matrix Profile and SAX

13. Intro to Deep Learning
   - What is deep learning?
   - Overview of deep networks
   - TensorFlow

14. 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 books will be used as a starter to guide the discussions, but they will be supplemented with instructor-developed lecture notes and reading assignments from other sources. The lecture notes and reading material will be made available on the OSBLE+ page of the course as the course proceeds.

  The book is freely available online at: http://www-bcf.usc.edu/~gareth/ISL/.
  (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.pdf. The material is intended for a modern theoretical course in computer science.)
  The book is freely available online at: http://www.mmds.org/#ver21.
Weekly Schedule

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

<table>
<thead>
<tr>
<th>Week</th>
<th>Topics</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 (Aug 20)</td>
<td>What is Data Science?</td>
<td>Pre-course survey out</td>
</tr>
<tr>
<td>02 (Aug 27)</td>
<td>intro to R</td>
<td>Survey due, Assignment 1 out</td>
</tr>
<tr>
<td>03 (Sep 03)</td>
<td>Overview of Machine Learning</td>
<td>Assignment 1 due</td>
</tr>
<tr>
<td>04 (Sep 10)</td>
<td>Exploratory Data Analysis</td>
<td>Assignment 2 out</td>
</tr>
<tr>
<td>05 (Sep 17)</td>
<td>Data Wrangling, Project set up</td>
<td>Assignment 2 due, Assignment 3 out</td>
</tr>
<tr>
<td>06 (Sep 24)</td>
<td>Linear Regression</td>
<td>Assignment 3 due, Project proposal out</td>
</tr>
<tr>
<td>07 (Oct 01)</td>
<td>Classification</td>
<td>Assignment 4 out</td>
</tr>
<tr>
<td>08 (Oct 08)</td>
<td>Resampling Methods</td>
<td>Project proposal due</td>
</tr>
<tr>
<td>09 (Oct 15)</td>
<td>Unsupervised Learning</td>
<td>Assignment 4 due</td>
</tr>
<tr>
<td>10 (Oct 22)</td>
<td>Data Visualization</td>
<td>Assignment 5 out</td>
</tr>
<tr>
<td>11 (Oct 29)</td>
<td>Tree-based Methods</td>
<td>Assignment 5 due</td>
</tr>
<tr>
<td>12 (Nov 05)</td>
<td>Time Series Data Mining</td>
<td></td>
</tr>
<tr>
<td>13 (Nov 12)</td>
<td>Intro to Deep Learning</td>
<td>Mid-term Exam</td>
</tr>
<tr>
<td>14 (Nov 19)</td>
<td>Thanksgiving break</td>
<td></td>
</tr>
<tr>
<td>15 (Nov 26)</td>
<td>Ethics, Wrap-up</td>
<td>In-class Exercise</td>
</tr>
<tr>
<td>16 (Dec 03)</td>
<td>Project presentations</td>
<td>Final project report due on Dec 12</td>
</tr>
</tbody>
</table>

Table 1: Tentative week-by-week 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
- arriving on time and remaining throughout the class

Correspondence

All class related correspondence with the instructor will be made via OSBLE+. I will check messages sent to my Inbox on a regular basis, and will do my best to respond promptly. Students are encouraged to choose their OSBLE+ settings so that they get emails notifications when messages are sent or posted.

Attendance

Regular attendance is required. While students may miss class for urgent reasons, 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 the OSBLE 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 and will be reported to the Office of Student Standards and Accountability. Cheating is defined in the Standards for Student Conduct WAC 504-26-010 (3). You can learn more about Academic Integrity on the WSU campus at http://conduct.wsu.edu. Please also read this link carefully: EECS Academic Integrity Policy (http://www.eecs.wsu.edu/~schneidj/Misc/academic-integrity.html). Use these resources to ensure that you do not inadvertently violate WSU’s standard of conduct.

Safety on Campus
Washington State University is committed to enhancing the safety of the students, faculty, staff, and visitors. It is highly recommended that you review the Campus Safety Plan (http://safetyplan.wsu.edu/) and visit the Office of Emergency Management web site (http://oem.wsu.edu/) for a comprehensive listing of university policies, procedures, statistics, and information related to campus safety, emergency management, and the health and welfare of the campus community.

WSU Classroom Safety
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).

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.

**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.

**Weather Policy**

For emergency weather closure policy, consult: http://alert.wsu.edu.

**Changes**

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