Introduction to Data Science: CptS 483–04
Syllabus

Course Information
Credit Hours: 3
Semester: Fall 2016
Meeting times and location: MWF, 12:10–13:00, Sloan 5
Course website: http://www.eecs.wsu.edu/~assefaw/CptS483-04/

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

Office hours: Wednesdays 1:00–2:30pm, or by appointment.

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. This course introduces students to this rapidly growing field and equips them with some of its basic principles and tools as well as its general mindset. Students will learn concepts, techniques and tools they need to deal with various facets of data science practice, including data collection and integration, exploratory data analysis, predictive modeling, descriptive modeling, data product creation, evaluation, and effective communication. The focus in the treatment of these topics will be on breadth, rather than depth, and emphasis will be 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.

Learning Outcomes
At the conclusion of the course, students should be able to:

- Describe what Data Science is and the skill sets needed to be a data scientist.
- Explain in basic terms what Statistical Inference means.
- Use R to carry out basic statistical modeling and analysis.
- Explain the significance of exploratory data analysis (EDA) in data science. Apply basic tools (plots, graphs, summary statistics) to carry out EDA.
- Describe the Data Science Process and how its components interact.
- Use effective data wrangling approaches.
- Apply basic machine learning algorithms (Linear Regression, k-Nearest Neighbors (k-NN), k-means, Naive Bayes) for predictive modeling.
• Identify common approaches used for Feature Generation.
• Identify basic Feature Selection algorithms (Filters, Wrappers, Decision Trees, Random Forests) and use in applications.
• Identify and explain fundamental mathematical and algorithmic ingredients that constitute a Recommendation System.
• Carry out basic social network mining tasks using a suitable network analysis tool.
• Create effective visualization of given data (to communicate or persuade).
• Work effectively in teams on data science projects.
• Reason around ethical and privacy issues in data science conduct and apply ethical practices.
• Apply knowledge gained in the course to carry out a project and write a technical report.

Audience
The course is suitable for upper-level undergraduate or graduate students in computer science, engineering, applied mathematics, the sciences, business, and related analytic fields.

Prerequisites
Students are expected to have basic knowledge of algorithms and reasonable programming experience (equivalent to completing a data structures course such as CptS 223), and some familiarity with basic linear algebra (e.g. solution of linear systems and eigenvalue/vector computation) and basic probability and statistics. If you are interested in taking the course, but are not sure if you have the right background, talk to the instructor. You may still be allowed to take the course if you are willing to put in the extra effort to fill in any gaps.

Course Work
The course consists of lectures (three times a week, 50 min each), and involves a set of assignments (about 4) and a 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 are encouraged to work in teams of two or three for a project. Assignments, on the other hand, are to be completed and submitted individually. Besides the assignments and a project, there will be frequent opportunities for in-class exercises and “thought experiments”.

Grading
Your final grade will be determined based on your performance on each of the following items; the percentages in parenthesis show the weight each item carries to the final grade.

• Class participation (10%)
• Assignments (30%)
• Project (30%)
• Exam (30%)

Letter grades: 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%). Grading scale may be adjusted depending on class average.
Topics and Course Outline:

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

2. Statistical Inference and R
   - Populations and samples
   - Statistical modeling, probability distributions, fitting a model
   - Intro to R

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

4. Three Basic Machine Learning Algorithms
   - Linear Regression
   - k-Nearest Neighbors (k-NN)
   - k-means

5. One More Machine Learning Algorithm and Usage in Applications
   - Motivating application: Filtering Spam
   - Why Linear Regression and k-NN are poor choices for Filtering Spam
   - Naive Bayes and why it works for Filtering Spam

6. Data Wrangling
   - Data cleaning, data reshaping, data integration
   - dplyr, tidyr

7. Feature Generation
   - Motivating application: user (customer) retention
   - Feature Generation (brainstorming, role of domain expertise, and place for imagination)

8. Feature Selection
   - Filters; Wrappers
   - Decision Trees; Random Forests

9. Recommendation Systems
   - Algorithmic ingredients of a Recommendation Engine
   - Dimensionality Reduction
   - Singular Value Decomposition

10. Mining Social-Network Graphs
    - Social networks as graphs
    - Node-level analysis
    - Group-level analysis

11. Data Visualization
    - Basic principles, ideas and tools for data visualization
    - Examples of inspiring (industry) projects
    - Exercise: create your own visualization of a complex dataset

12. Data Science and Ethical Issues
    - Discussions on privacy, security, ethics
    - A look back at Data Science
    - Next-generation data scientists
### Week Topics Assignments

<table>
<thead>
<tr>
<th>Week</th>
<th>Topics</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 (Aug 22)</td>
<td>What is Data Science</td>
<td>Survey out</td>
</tr>
<tr>
<td>02 (Aug 29)</td>
<td>Statistical Inference, R</td>
<td>Survey due, Assignment 1 out</td>
</tr>
<tr>
<td>03 (Sep 05)</td>
<td>Exploratory Data Analysis, R</td>
<td>Assignment 1 due</td>
</tr>
<tr>
<td>04 (Sep 12)</td>
<td>Linear Regression</td>
<td>Assignment 2 out</td>
</tr>
<tr>
<td>05 (Sep 19)</td>
<td>k-nearest Neighbors, k-means</td>
<td>Assignment 2 due</td>
</tr>
<tr>
<td>06 (Sep 26)</td>
<td>Naive Bayes</td>
<td>Assignment 3 out</td>
</tr>
<tr>
<td>07 (Oct 03)</td>
<td>Data Wrangling</td>
<td>Assignment 3 due</td>
</tr>
<tr>
<td>08 (Oct 10)</td>
<td>Feature Generation</td>
<td>Project proposal out</td>
</tr>
<tr>
<td>09 (Oct 17)</td>
<td>Decision Trees, Random Forests</td>
<td></td>
</tr>
<tr>
<td>10 (Oct 24)</td>
<td>Recommendation Systems</td>
<td>Project proposal due</td>
</tr>
<tr>
<td>11 (Oct 31)</td>
<td>Mining Social Networks</td>
<td>Mid-term Exam</td>
</tr>
<tr>
<td>12 (Nov 07)</td>
<td>Visualization</td>
<td>Assignment 4 out</td>
</tr>
<tr>
<td>13 (Nov 14)</td>
<td>Ethics, Look-back</td>
<td>Assignment 4 due</td>
</tr>
<tr>
<td>14 (Nov 21)</td>
<td>Thanksgiving break</td>
<td></td>
</tr>
<tr>
<td>15 (Nov 28)</td>
<td>Project presentations</td>
<td></td>
</tr>
<tr>
<td>16 (Dec 05)</td>
<td>Project presentations</td>
<td>Final project report due</td>
</tr>
</tbody>
</table>

Table 1: Tentative week-by-week schedule of topics and assignments.

### Books

There is no standard one “textbook” for this course. The following book will be used as a primary text to guide much of the discussions, but it will be heavily supplemented with 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.


Additional references and books related to the course:


### Schedule (tentative)

See Table 1 for a tentative 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
- arriving on time and remaining throughout the class

You may use any non-disruptive personal electronics during class.

Correspondence

All class-related correspondence with the instructor will be made via OSBLE+. I will check messages sent to my Inbox or posted to the Dashboard 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.

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 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 website (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.