CptS 475/575: Data Science
About me

• Name: Assefaw Gebremedhin
• Office: EME B43
• Webpage: www.eecs.wsu.edu/~assefaw
• Joined WSU: Fall 2014
• Research interests: algorithmic data science, network science, high performance computing, bioinformatics
• Lab: Scalable Algorithms for Data Science Laboratory (https://scads.eecs.wsu.edu)
• NSF CAREER project: Fast and Scalable Combinatorial Algorithms for Data Analytics www.eecs.wsu.edu/~assefaw/fascada

Teaching at WSU:
• CptS 483: Intro to Data Science  (Fall 2015, 2016, 2017)
• CptS 591: Elements of Network Science (Spring 2015, 2016, 2017, 2018)
• CptS/STAT 424: Data Analytics Capstone (Starting Spring 2019)

CptS 475/575 Fall 2018
• Lectures: MWF, 9:10—10, College Hall 220
• Office Hours: Wed. 10:30am—12pm
• Teaching Assistant: Helen Catanese
• Email: helen.catanese@wsu.edu
• Office: Dana 115
About Data Science Class of 2018
(What I know so far)

• CptS 475: 20 enrolled
• CptS 575: 40 enrolled, 10 waitinglist (50 total)
• Class room capacity: 65

• CptS 475 breakdown (all BS or BA)
  • CS: 18
  • Math: 2

• CptS 575 breakdown (the 50 enrolled/waiting)
  • 32 MS
    • CS: 31
    • EE: 1
  • 18 PhD
    • CS: 8
    • EE: 4
    • CE: 1
    • Math: 2
    • Economics: 1
    • Geology: 1
    • Operations Research: 1
Course websites

- Public course site: [https://scads.eecs.wsu.edu/index.php/data-science](https://scads.eecs.wsu.edu/index.php/data-science)
  - Syllabus
  - Overview of schedule (updated after every lecture)
  - Resources
- OSBLE+: [https://plus.osble.org](https://plus.osble.org)
  - Lecture material
  - Assignments
  - Announcements and posts
  - Submissions and feedback
  - Currently: ~ 50 added users; ~ 20 whitelisted (be sure to respond to invitation ASAP)
Course Description

• Data Science is the study of the generalizable extraction of knowledge from data
• Data science requires integrated skill set spanning
  • Computer science
  • Mathematics & Statistics
  • Domain expertise
    • art of problem formulation to engineer effective solutions
• Purpose of this course: introduce basic principles, tools, and general mindset
• Emphasis on breadth rather than depth; and on synthesis of concepts
• Primarily uses the statistical computing language R
Expectation

• Basic knowledge of algorithms and reasonable programming experience (equivalent to completing CptS 223)
• Familiarity with basic linear algebra
• Basic probability and statistics

• Deficiencies can to a degree be overcome with extra effort
Topics

1. Introduction: What is Data Science?
2. Statistical Learning and Intro to R
3. Exploratory Data Analysis and the Data Science Process
4. Linear Regression
5. Classification
   • Overview, Logistic regression, k-Nearest Neighbors
6. Resampling Methods
   • Cross-validation, The bootstrap
7. Unsupervised Learning
   • Principal Components Analysis, K-means clustering, Hierarchical clustering
8. Data Wrangling
   • Data transformation and manipulation (dplyr); Data tidying (tidyr)
9. Data Visualization
10. Time Series Data Mining
    • Distance measures, transformations, algorithms, tools (Matrix Profile, SAX)
11. Tree-based Methods
    • Decision Trees; Bagging, Random Forest and Boosting
12. Intro to Deep Learning
13. Data Science and Ethics
A few things

• Pre-course survey
  • Your background
  • Level of familiarity with R, Python, MathLab
  • Topics you are excited about
  • Other topics you wish to see covered
  • Complete and submit on OSBLE

• R tutorial

• (Python tutorial)

• Tutorial generally preferred time
Course work and assessment: CptS 475

- **Assignments (30%)**
  - Total of 4 to 5 assignments spread through the semester
  - Completed and submitted individually
  - Each of the assignments carries equal weight

- **Semester Project (30%)**
  - Team of two or three
  - Option between choosing from a given list OR propose own project
  - Guidelines will be provided

- **Exam (30%)**
  - Late midterm
  - Designed to cover most material AND complement assignments and semester project

- **Class participation (10%)**
  - Attendance
  - Active participation
Course work and assessment: CptS 575

• Assignments (25%)
  • Total of 4 to 5 spread through the semester
  • Completed and submitted individually
  • Each of the assignments carries equal weight

• Semester Project (30%)
  • Team of two or three
  • Option between choosing from a given list OR propose own project
  • Guidelines will be provided

• Survey Paper (15%)
  • Submitted individually
  • Further explore a specific topic related to the course content
  • Topic to be chosen in consultation with instructor

• Exam (20%)
  • Late midterm
  • Designed to cover most material AND complement assignments and semester project

• Class participation (10%)
  • Attendance
  • Active participation
## Weekly Schedule

<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>Statistical Learning Overview, R</td>
<td>Survey due, Assignment 1 out</td>
</tr>
<tr>
<td>03 (Sep 03)</td>
<td>Exploratory Data Analysis, R</td>
<td>Assignment 1 due</td>
</tr>
<tr>
<td>04 (Sep 10)</td>
<td>Linear Regression</td>
<td>Assignment 2 out</td>
</tr>
<tr>
<td>05 (Sep 17)</td>
<td>Classification I</td>
<td>Assignment 2 due</td>
</tr>
<tr>
<td>06 (Sep 24)</td>
<td>Classification II</td>
<td>Assignment 3 out</td>
</tr>
<tr>
<td>07 (Oct 01)</td>
<td>Resampling Methods</td>
<td>Assignment 3 due</td>
</tr>
<tr>
<td>08 (Oct 08)</td>
<td>Unsupervised Learning, Project setup</td>
<td>Project proposal out</td>
</tr>
<tr>
<td>09 (Oct 15)</td>
<td>Data Wrangling</td>
<td>Assignment 4 out</td>
</tr>
<tr>
<td>10 (Oct 22)</td>
<td>Data Visualization</td>
<td>Assignment 4 due, Project proposal due</td>
</tr>
<tr>
<td>11 (Oct 29)</td>
<td>Tree-based Methods</td>
<td>Assignment 5 out</td>
</tr>
<tr>
<td>12 (Nov 05)</td>
<td>Time Series Data Mining</td>
<td>Assignment 5 due</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></td>
</tr>
<tr>
<td>16 (Dec 03)</td>
<td>Project presentations</td>
<td>Final project report due on Dec 12</td>
</tr>
</tbody>
</table>
Learning Outcomes

• Describe what Data Science is and the skill sets needed
• Describe the Data Science Process
• Use R to carry out basic statistical modeling and analysis
• Carry out exploratory data analysis (to gain insight)
• Apply basic machine learning algorithms for predictive modeling
• Correctly apply cross-validation to assess model performance
• Apply unsupervised learning methods to discover patterns, trends and anomalies in data
• Use effective data wrangling approaches to manipulate data
• Create effective visualization of data (to communicate or persuade)
• Reason around ethical and private issues in data science and apply ethical practices
• Work effectively in teams on data science projects
• Apply knowledge gained in the course to carry out a project and write technical report
Books

- No required textbook
- Lecture notes (slides) and reading material will be made available on the OSBLE+ page
- References
  - Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013. (Freely available online)
Policies

- Conduct in class
  - Silence personal electronics
  - Arrive on time and remain throughout the class

- Correspondence
  - Happens via OSBLE+

- Attendance
  - Required. Make sure absences are cleared with me

- Missing or late work
  - Max 48 hrs with 10% penalty per 24 hrs

- Academic Integrity
  - Strongly enforced

- Consult syllabus for more details
Welcome again to the course!
I am excited to have you in the class,
and I look forward to your participation and to
a great semester!