CptS 483:04 Introduction to Data Science
About me

- Name: Assefaw Gebremedhin
- Office: EME B43
- Webpage: www.eecs.wsu.edu/~assefaw
- Joined WSU: Fall 2014
- Research interests: combinatorial scientific computing, network science, data mining, machine learning, 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/STAT 424: Data Analytics Capstone (Planned)
About Data Science Class of 2017
(What I know so far)

• Current enrollment: 30
• By level:
  • Graduate: 12 (7 PhD, 5 MS)
  • Undergraduate: 16 (Senior)
  • Post-bacc undergraduate: 2
• By program:
  • Computer Science: 22
  • Electrical Engineering: 2
  • Computer Engineering: 1
  • Software Engineering: 1
  • Bio and Ag Engineering: 1
  • Mathematics: 1
  • Antropology: 1
  • Biology: 1
Course websites

- Public course site: https://scads.eecs.wsu.edu/index.php/data-science
  - Syllabus
  - Overview of schedule (updated after every lecture)
  - Resources
- OSBLE+: https://plus.osble.org
  - Lecture material
  - Assignments
  - Announcements
  - Posts
  - Submissions and feedback

- Currently: 18 added users; 12 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**
   - K-NN, Logistic regression, Naïve Bayes classifier, Decision Trees
6. **Unsupervised Learning**
   - K-means clustering, Hierarchical clustering, Principal Components Analysis
7. **Data Wrangling**
   - Data cleaning, data reshaping, data integration; dplyr, tidyr
8. **Data Visualization**
9. **Time Series Data Mining**
   - Distance measures, transformations, algorithms, tools (Matrix Profile, SAX)
10. **Recommender Systems and Social Network Mining**
11. **Intro to Deep Learning**
12. **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

• Assignments (30%)
  • About 4 throughout 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
## Weekly Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Topics</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>01 (Aug 21)</td>
<td>What is Data Science</td>
<td>Survey out</td>
</tr>
<tr>
<td>02 (Aug 28)</td>
<td>Statistical Learning, R</td>
<td>Survey due, Assignment 1 out</td>
</tr>
<tr>
<td>03 (Sep 04)</td>
<td>Exploratory Data Analysis, R</td>
<td>Assignment 1 due</td>
</tr>
<tr>
<td>04 (Sep 11)</td>
<td>Linear Regression</td>
<td>Assignment 2 out</td>
</tr>
<tr>
<td>05 (Sep 18)</td>
<td>Classification I</td>
<td>Assignment 2 due</td>
</tr>
<tr>
<td>06 (Sep 25)</td>
<td>Classification II</td>
<td>Assignment 3 out</td>
</tr>
<tr>
<td>07 (Oct 02)</td>
<td>Unsupervised Learning</td>
<td>Assignment 3 due</td>
</tr>
<tr>
<td>08 (Oct 09)</td>
<td>Data Wrangling, Project setup</td>
<td>Project proposal out</td>
</tr>
<tr>
<td>09 (Oct 16)</td>
<td>Data Visualization</td>
<td></td>
</tr>
<tr>
<td>10 (Oct 23)</td>
<td>Time Series Data Mining</td>
<td>Project proposal due</td>
</tr>
<tr>
<td>11 (Oct 30)</td>
<td>Recommender Systems, Social Networks</td>
<td>Mid-term Exam</td>
</tr>
<tr>
<td>12 (Nov 06)</td>
<td>Intro to Deep Learning</td>
<td>Assignment 4 out</td>
</tr>
<tr>
<td>13 (Nov 13)</td>
<td>Ethics, Wrap-up</td>
<td>Assignment 4 due</td>
</tr>
<tr>
<td>14 (Nov 20)</td>
<td>Thanksgiving break</td>
<td></td>
</tr>
<tr>
<td>15 (Nov 27)</td>
<td>Project presentations</td>
<td></td>
</tr>
<tr>
<td>16 (Dec 04)</td>
<td>Project presentations</td>
<td>Final project report due</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
- Apply basic machine learning algorithms for predictive modeling
- Apply unsupervised learning methods to discover patterns, trends and anomalies in data
- Use effective data wrangling approaches to manipulate data
- Identify and explain mathematical and algorithmic ingredients of a recommender system
- Create effective visualization of data
- 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)
  • Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. MIT Press, 2016. (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