CSCI 8360

Data Science Practicum

Location: Boyd 323 (T/R), Geo/Geo 154 (W)

Dr. Shannon Quinn

Email: squinn@cs.uga.edu
Website: https://dsp-uga.github.io/
Office Location: Boyd GSRC 638A
Office Hours: T/R 9:35 - 10:50am, or by appointment

The course syllabus is a general plan; when (not if) deviations arise, they will be announced.

Course Description: This course covers advanced data science techniques for analyzing large-scale data in distributed environments. Students will develop scalable algorithms in frameworks such as Spark and dask, and use deep learning libraries such as Keras and PyTorch. This course is team-based, involving several mini-projects over the course of the semester.

The course aims to provide students with a hands-on practicum for studying scalable machine learning with industry-grade analytics frameworks. Students will have the opportunity to implement cutting-edge algorithms in large-scale natural language processing, deep convolutional image analysis, deep reinforcement learning, and current cloud compute infrastructure. This course is ideal for students who anticipate working in quantitative fields such as biomedical imaging or natural language processing, or plan to enter a data science position in industry.

It is assumed that students taking this course have thorough knowledge of basic machine learning concepts (classification, clustering, regression, dimensionality reduction, etc) and are proficient in designing software using Java, Scala, or Python. Very little time will be spent introducing new machine learning concepts or explaining basic constructs of programming languages; students will be expected to work independently and formulate solutions with their teammates.

Prerequisites: Any one or more of the following:

- CSCI 6360 Data Science II
- CSCI 6380 Data Mining
- CSCI 6850 Biomedical Image Analysis
- CSCI 8140 Parallel Processing and Computational Science
- CSCI 8850 Advanced Biomedical Image Analysis
- CSCI 8945 Advanced Representation Learning
- CSCI (ARTI) 8950 Machine Learning
- CSCI 8951 Large-Scale Optimization for Machine Learning
- CSCI 8955 Advanced Data Analytics: Statistical Learning and Optimization

It is highly, highly, **HIGHLY recommended** that you take one of the above prerequisites OR have prior knowledge of and experience with machine learning / linear algebra / probability and statistics techniques before you take this course.

Credit Hours: 4

No required textbooks! Recommended texts include:

1. François Challet. *Deep Learning with Python* (1^{st} ed., 2017). ISBN-13: 9781617294433.
2. Sandy Ryza, Uri Laserson, Sean Owen, Josh Willis. *Advanced Analytics with Spark: Patterns for Learning from Data at Scale* (1^{st} ed., 2015) ISBN-13: 978-1491912768.
3. Andrew W. Trask. *Grokking Deep Learning* (1^{st} ed., 2019). ISBN-13: 9781617293702.
4. Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Elements of Statistical Learning* (2^{nd} ed., 2016). ISBN-13: 978-0387848570.
5. Christopher Bishop. *Pattern Recognition and Machine Learning* (1^{st} ed., 2006). ISBN-13: 978-0387310732.

Topical Course Outline

1. Lightning review of basic machine learning, probability, and statistics
2. Theory and practice of distributed computing (MapReduce, Hadoop, Spark, dask)
3. In-memory distributed computing
4. Cloud storage and elastic computing
5. Functional programming and distributed computing
6. Team-based software engineering principles (version control, branching / merging, documentation, tickets / milestones)
7. Deep learning for image segmentation and classification
8. Q-learning and Markov Decision Processes
9. Deep reinforcement learning for long-term agent planning and real-time decision-making
10. Current topics in data science

Grade Distribution:

Project 0	0.0%
Project 1	25.0%
Project 2	25.0%
Project 3	25.0%
Project 4	25.0%
Total	**100%**

Course Policies

- **Announcements**
  - I use either Slack or Discord for disseminating information and making course announcements outside of in-person class meetings. Signing on is **required** for the class, as is checking for new announcements at least once every 24 hours. Beyond that is entirely at your discretion.

- While you can and are more than welcome to directly DM or email me with any questions you have, I **strongly recommend** you post your questions in the `#questions` forum of the Slack/Discord chat. In over 5 years of offering this course, I can count on one hand total the number of times I've been asked questions by students whose answers pertained only to that student. Far, far more often, I get 5-10 DMs asking me the exact same question. If you're struggling with something, chances are someone else is, too!

- **Attendance**
  - Whenever there is a lecture scheduled, please make an effort to attend, especially when there is a guest lecturer. There really aren't many incentives beyond "you'll learn something," since the course has no exams or quizzes, and no grade for attendance. But you should still come to lecture. Who knows; I might say something hilarious, and you'll wish you'd been there when you hear from your classmates how hilarious it was. But alas, the moment will be lost forever; a lesson learned too late.
  - Don't *ask* if you can miss a day of lecture for a doctor's appointment / family event / nap, just *inform* me that you'll be missing lecture on a given day (don't even need to explain why) so that I don't go looking for you. You're adults with your own lives and schedules, and if you need to miss class for any reason, just let me know and then go do whatever you need to do.
  - Since this class is a practicum, lectures will taper off after the first couple weeks to be once per week, with the other class meetings being replaced by Office Hours. This way, I know you're available for Office Hours! So if you have questions, don't hesitate to come by. Many project teams even use this time to meet in-person, and use the opportunity to ask me questions as they come up.

- **Projects**
  - Projects are due by 11:59:59pm on the noted date. Projects submitted after this deadline will lose 25/100 points for every subsequent 24 hour-period they are late. **This takes effect as of 12:00:00am after the deadline**. "`git commit`" timestamps are used to determine time of project submission.
  - Each project is team-based. Teams are randomly assigned at the start of each of the main projects, and are created in such a way as to guarantee that you are matched with new teammates each time.
  - The final project is the only project which is neither randomly assigned nor subject to the unique teammate policy. Students are allowed to form their own final project teams, though these teams **must consist of at least 2 people and no more than 4**.

- **Deliverables**
  1. **Code.** Provided as a repository in the DSP-UGA GitHub organization.
  2. **README file.** A text file with 1) description of the program, 2) answers to any project questions, and 3) listing of any bugs in the code.
  3. **CONTRIBUTORS file.** A text file detailing the specific contributions and roles of the individual team members. Each team member should personally verify their contributions in this document.

4. **Issues.** Use the GitHub "Issues" functionality to assign tasks to teammates and discuss important aspects of the project; this will not only help keep the projects moving, but will also show effective division of labor between teammates.
5. **Documentation.** This takes many forms: from comments in the code, to descriptions in the repository wiki, to examples in the README; from development "paper trail" items such as GitHub issues, milestones, and even the `git commit` log; to project extras, including a project website, demo, or unit testing.
6. **Lightning talks.** Each team is expected to prepare a 5-minute (*strictly enforced*) talk at the end of each project cycle that briefly covers the specifics of their team's approach and lessons learned. Since each team works on the same project, no background on the project is needed; instead, teams should focus their limited time on explaining the approach they took, how they explored their solution, what worked, what didn't, and what they would try next or do differently next time.
7. **Peer review submission.** *This is a requirement of EVERY INDIVIDUAL*, unlike the previous items, of which only one was required for the whole team. In this case, each person is assigned a single team on which to conduct a 20-30 minute peer review.
8. Other deliverables and expectations are enumerated in the peer review submission form and will be discussed in detail ahead of the projects.

Academic Honesty

As a University of Georgia student, you have agreed to abide by the University's academic honesty policy, "A Culture of Honesty," and the Student Honor Code. All academic work must meet the standards described in "A Culture of Honesty" found at: `https://ovpi.uga.edu/academic-honesty/academic-honesty-policy`. Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation. Questions related to course assignments and the academic honesty policy should be directed to the instructor.

- Read "A Culture of Honesty," the UGA academic honesty policies and CS Academic Integrity policies.
- You must not allow others to copy or look at your work.
- You must not give/share your lab/project assignment work to a fellow student.
- Copying significant portions of code from a fellow student or any other source (including internet) is plagiarism and will be dealt with as such.
- Seriously though: I've caught a few students who tried to use previous years' solutions. It's harder than you think to pass off code written by others as your own, especially with machine learning on my side. You can use blocks of code here and there **provided you cite where you got them**, but please assemble the full solution yourselves.
- If in doubt, **cite**.
- If you have questions about an assignment or if you run into problems, **ask me**.
- This course has no exams or quizzes, only projects, so please don't ask about extra credit. There is none.

- All of your coursework must meet the aforementioned policies and rules. Students that violate any of these rules or the UGA Academic Honesty policies will be liable to a penalty. The instructor will strictly enforce Academic Honesty policies and report any violation of the aforementioned policies and rules.
- Content generated by an AI framework (e.g., ChatGPT, Copilot, etc) fall under the same guidelines as content generated by another student or colleague: namely, that you **need to cite it** and explain clearly how it didn't just do all the work for you.

****Highly Tentative** Course Outline**
Subject to change.

Week	Content
Week 1	• Course introduction; overview of data science; introduction to distributed computing • *Project 0 out*
Week 2	• Introduction to Cloud computing • *Project 0 due; Project 1 out*
Week 3	• Introduction to Natural Language Processing
Week 4	• Introduction to Large-Vocabulary Document Classification
Week 5	• Technical Debt in Machine Learning • *Project 1 due; Project 2 out*
Week 6	• Project 1 Lightning Talks
Week 7	• Introduction to Computer Vision
Week 8	• Ethics in Data Science • *Project 2 due; Project 3 out*
Week 9	• Project 2 Lightning Talks
Week 10	• Introduction to Reinforcement Learning • *Final Project assigned*
Week 11	• Guest Lecturer / Current Topics • *Project 3 due*
Week 12	• Project 3 Lightning Talks
Week 13	• Guest Lecturer / Current Topics • *Final Project mid-point check-in*
Week 14	• Guest Lecturer / Current Topics
Week 15	• Course Wrap-Up
Week 16	• Final Project Talks • *Final Project due*