



CSCI 8360: Data Science Practicum

Lecture 1

Dr. Shannon Quinn

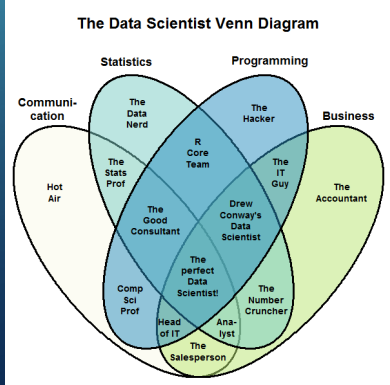
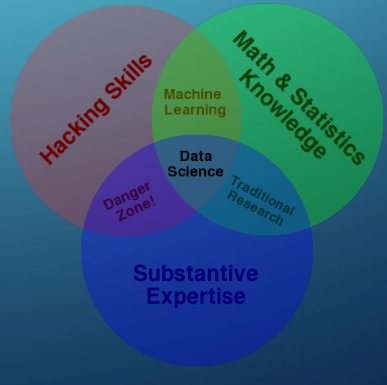
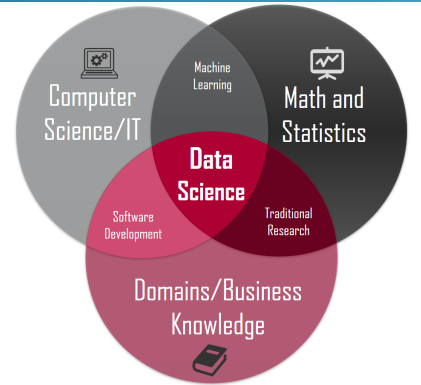
The background of the slide is a teal-to-blue gradient. It features a decorative pattern of white circuit board traces and nodes, primarily located in the corners and along the left and right edges. The main content is centered in the lower half of the slide.

Part I: Lightning Overview

CSCI 8360 DATA SCIENCE PRACTICUM

Data Science

- What is it?
- Why is it important?
- How does one learn it?



CSCI 8360: What Is It?

What this class is **NOT**

- Introduction to *Machine Learning*
- Introduction to *Distributed Systems*
- Introduction to *Software Engineering*

What this class **IS**

- Hands-on data science
- Team-based problem solving
- “Kaggle in the Classroom”

CSCI 8360 Requirements

- Thorough understanding of machine learning and statistics
 - (or teammates who can bring you up to speed very quickly)
- Good software engineering skills
 - (working on teams)
- An ability to learn fast
 - (definition of “graduate student”)

CSCI 8360 Links

- Course website
 - <https://dsp-uga.github.io/sp18>
 - Lectures and assignments will be posted here
- Slack
 - <https://eds-uga-csci8360.slack.com>
 - This is where **all course communication** will happen
- GitHub
 - <https://github.com/dsp-uga/>
 - All team development will happen here (**part of your grade**)
- AutoLab
 - <https://autolab.cs.uga.edu>
 - Project submissions for grading and evaluation
- Google Compute Platform (GCP)
 - Everyone will get credits

Course Outline

- 4 Projects (+ a pseudo-project), each 2-3 weeks
- 1 Final Project
- Lecture every Wednesday
- Office hours Tuesday/Thursday
- No exams!
- Attendance

The slide features a teal-to-blue gradient background. In the four corners, there are decorative white line-art patterns resembling circuit board traces and nodes.

Part II: Administrative Details

CSCI 8360 DATA SCIENCE PRACTICUM

Lectures, Revisited

- Location: **Boyd 208**
- Time
 - Today, 11:00am – 12:15pm
 - Next Tuesday (January 9), 11:00am – 12:15pm
 - Every Wednesday, 11:15am – 12:05pm
 - **NO OTHER LECTURE TIMES** (unless announced in Slack)

Office Hours, Revisited

- Location: **Boyd 208**
- Time: Tuesdays / Thursdays, 11:00am – 12:15pm
- (yep, when we'd otherwise have lecture, so I know you can't possibly have conflicts)
- Happy to set up appointments if you need them

Slack, Revisited

- Slack: free team messaging platform
- Web-based and mobile apps
- Teams can set up private direct chats to coordinate
- Can also send individual DMs
- **All course announcements will be made here**

The screenshot shows a Slack workspace for 'EDS@UGA: CSC...'. The left sidebar lists channels: # announcements (selected), # lounge, and # techprobs. The main area shows the # announcements channel with a pinned message from 'spq' at 4:02 PM: 'Course website, which effectively functions as the syllabus: <https://dsp-uga.github.io/sp18>'. Above this is a message from 'spq' at 3:10 PM: '@everyone Final projects due tonight!'. There is also a photo of a workshop presentation with a quote: 'I'm currently at the IEEE BigData conference in D.C. running a workshop on open science in big data, and Jeremy freeman--the guy who made codeneuro from P4--just gave our class a shout-out in his keynote talk :)'. The photo shows a person at a podium in front of a screen displaying a workshop agenda.

GitHub, Revisited

- Most popular code repository in the world
- Uses the *git* concurrent versioning system (itself an open source project)
- Lots of useful team-based tools (issue tracker, wiki, GUI)
- **All projects will be sourced in the DSP-UGA GitHub organization**



A screenshot of a GitHub repository page for the organization 'dsp-uga' and repository 'sp18'. The page shows the repository name, navigation tabs (Code, Issues, Pull requests, Projects, Wiki, Insights, Settings), and repository statistics (4 commits, 1 branch, 0 releases, 1 contributor, MIT license). A commit history table is visible, showing a commit by 'magsol' with files 'docs', 'LICENSE', and 'README.md'. The README content is partially visible, starting with 'Spring 2018: CSCI 8360' and 'Spring 2018 rendition of CSCI 8360 Data Science Practicum.'

AutoLab, Revisited

- Assignment submission and autograder
- Also has leaderboards!
- **All project outputs will be submitted to AutoLab for ranking**

Google Cloud Platform

- (comparable to Amazon Web Services, or AWS)
- Spin up elastic compute resources on-demand
- Every student gets \$50 in credits (usable across ALL services)
- “Cloud Dataproc” contains APIs for specifically spinning up Spark and Hadoop clusters
- **Details to come**



The slide features a teal-to-blue gradient background. In the corners, there are decorative white line-art patterns resembling circuit boards or data paths, with small circles at the end of the lines.

Part III: Projects

CSCI 8360 DATA SCIENCE PRACTICUM

Project Overview

- Solving real-world machine learning problem
 - Classify a large corpus of documents
 - Convex optimization over a huge dataset
 - Dimensionality reduction over a high-dimensional matrix
 - etc.
- Each project varies in length from 2 to 3 weeks
 - “Introductory” Project 0 out **next Tuesday**, will be only 1 week long
 - Project 1 (P1) out the following Tuesday (Jan 16), will be 2.5 weeks long

Project Requirements: Teams

- Teams (2-3 people per team)
 - Assigned *completely randomly* (by me)
 - Will change for each project
 - (you can form your own teams for the final project)
- Each team member should have a **clear, well-defined role**
 - Not everyone has to be a coder!
 - But 1 person should not be carrying the whole project



Project Requirements: Code

- Use good coding practices
 - Documentation (in code, in GitHub wiki, in README, in commit comments)
 - Well-organized structure (should be easy for me to understand)
- Use organizational GitHub account
 - <https://github.com/eds-uga>
- Recommended additional practices
 - License your code with a permissive open license (<https://choosealicense.com/>)
 - Add a continuous integration module (<https://travis-ci.org/>)
 - Implement unit testing for your code
 - Create a website for your project (see GitHub documentation; makes this easy)



Project Requirements: AutoLab

- Submit to AutoLab before the deadline
 - One submitter per team (can submit as many times as you like)
 - Unless otherwise specified, submission will always be a text file with your code's predictions on a test dataset
 - If your submission is correctly formatted, your performance should show up on the leaderboard in short order
- AutoLab submission **shuts down after the deadline**

Project Requirements: Lightning Talks

- Wednesdays *after* a project deadline, all teams will give a lightning talk (~4-6 minutes long)
- Talks will outline the problem, the team's approach, their results, and any other discussion points
- Creativity welcome—code examples, live demos, interactive slides, etc
- One person from each team will speak

Project Grading

- Everyone starts at an 85% (solid B)
- Grading split into three categories
 1. Theory (the approach you use as implemented by the code)
 2. Engineering (everything around the implementation)
 3. “Extras”
- Go above and beyond—extra points
- Shortcomings (approach is flawed or too simple, code poorly documented, one person did almost all of the project, poor leaderboard ranking) reduce points
- Grading report will be issued to each team shortly after the project deadline

Final Project

- Also team-based, 2-3 people (but you choose your own teams)
- Includes proposal + final write-up + final presentation components
- Presentations will happen the week of April 16 (last two weeks of classes)
- More details to come!

A decorative background featuring a light blue to dark blue gradient. The corners are adorned with white circuit board traces and nodes, creating a technical and digital aesthetic.

Part IV: The Next Step

CSCI 8360 DATA SCIENCE PRACTICUM

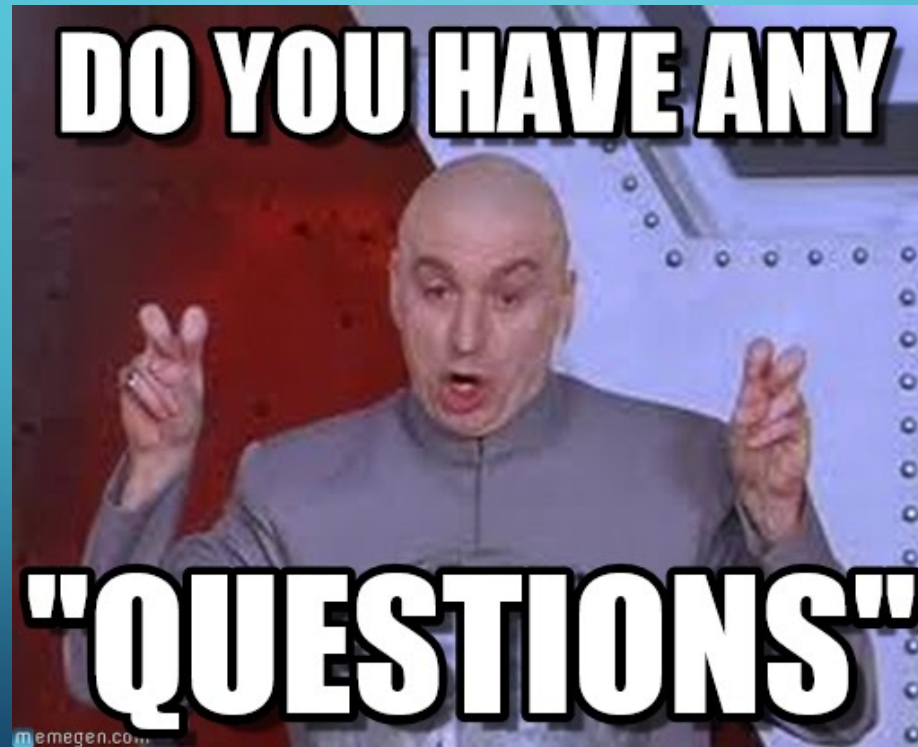
Project -1 (P-1)

1. Email me (squinn@cs.uga.edu) with your preferred email to send a Slack invite. Join the Slack team.
2. Send me your GitHub username (create an account if you don't have one). Join the GitHub "Data Science Practicum" team.
3. Start looking at Apache Spark (for Project 0 next week).

Next week: Project 0

- The only individual project of the semester
- Mainly to familiarize you with Apache Spark (used for Projects 1 and 2, possibly for 3 and 4 as well), AutoLab, GitHub, and Slack
- Won't count for a grade, but **is required**

QUESTIONS?



Finally...

- What large-scale problems do you want to work on?
- Yes, this an opportunity to suggest Projects. If you have an idea, send me:

1. The problem to be solved (optimization, dimensionality reduction, classification, etc)
2. How the solutions should be evaluated
3. Training and validation datasets



Your idea
could be
featured as a
full project!