DSC Capstone Sequence

Lecture 01
2021-2022

Today's Outline

- Introduction to the 180AB (Capstone Sequence)
- Syllabus + Quarter I assignments

Course Resources / Services

- Capstone Website: https://dsc-capstone.github.io
  - Syllabus; assignment descriptions; course schedule; lecture slides
- Canvas
  - assignment submission; administrative announcements; course staff directory (domain mentors and methodology TAs).
  - **Building access code for SDSC**
- Computing Resources: DSMLP Server
- Gradescope: for code assignments and some methodology HW.
- Course communication:
  - Canvas/Piazza for lecture/administrative questions
  - Talk to domain mentor for communication preferences for discussion

Introduction to the Capstone Sequence

Capstone Sequence Course Goals

- Put together DSC skills through the lifecycle of a two-quarter project.
- Learn methodological best practices for large projects:
  - Reproducible and flexibly generic work
  - Effective (visual, oral) communication of work and results
- Starting an investigation with a *question* instead of a method.
- A detail-oriented pursuit of a proposal in a chosen domain.
- Produce and show off work that you are *proud of!*

Challenges of a Data Science Capstone @UCSD

- **Topical Variety**: topics can span almost anything imaginable.
  - How to find mentors that reasonable "cover" this space?
  - How do you consistently evaluate such varied student work?
- **Flipped Background**: traditional capstones *start* with domain
  - Students come in without domain knowledge.
  - Students have a robust "methodological toolkit"
- **Large Size**: ~200 students/year.
  - Not specific to DS, but more challenging because of it!

These questions motivate the structure of the course!

Structure of the Capstone Sequence

- Mentors sponsor **domains of inquiry** that support (multiple) **projects**.
- Students enroll in a domain of inquiry based on their interests.

	DS Methodology (1hr/wk -- Lecture)	**Domain Mentorship (1hr/wk -- Section)**
Quarter One (10 wks)	Best practices in DS project development	Introduction to the domain + project proposal
Quarter Two (10 wks)	Project planning and effective communication	Project execution + (varied) presentation

- Plus 2 "Lab Hours" (unscheduled, or schedule by mentor)

A Closer Look: Methodology (Lecture)

- Methodology portion sets standards for a data science project across a wide variety of domains:
  - Responsible resource usage (remote vs local development; test data)
  - Reproducible research (git; docker; python packages; updateable reports/notebooks)
  - Effective Communication (scientific writing; oral presentations; teamwork)
- All student work for the methodology portion is directly applied to the domain.
  - E.g. The codebase for student projects are graded against these standards.
  - E.g. Students analyze the *writing* of a publication in their chosen domain.

A Closer Look: Domain of Inquiry (Discussion)

- Q1 focuses on learning a domain, through studying work in the field:
  - e.g. replicate a paper
  - Build useful code/tools for Q2
  - Emulate and practice effective scientific communication
- Domain mentors will generally assign tasks and reading
  - Come to section prepared to *discuss* the tasks/reading...
  - Weekly work is *difficult*; your mentors are aresource -- ASK THEM QUESTIONS!
- For questions about the domain, contact your mentor with their preferred mode of communication.

Note: **Mentors largely operate unaware of how the lecture portion is structured**. Give them context, if asking them a question about the domain that relates to lecture!

Discussion Section Structure

- Learn the domain and pursue a proposal *guided* by domain expert.
  - *You* are responsible for learning material and doing data analyses.
  - Come ready each Wednesday to *actively* discuss the material and results.
  - **Coming to section prepared is mandatory and necessary for the success of capstone!**
- Discussions are for engaging contextual questions, data, and conclusions.
  - Sections should *not* deal with coding problems.
  - Data Scientists must translate problems about code into the language of the domain/data.
- Discussions are for:
  - Engaging with the domain and the questions at hand.
  - A place to ask for clarification about the data generation process.
  - Brainstorm with peers about how possible proposals.

What do your projects look like (in Q2)?

- Each project is worked on in groups of ~2-4 (formed by instructors)
- Groups submit project proposals in their domain (w/plan) and give an elevator pitch. Mentors give ok, given their expertise and flexibility.
- The final project artifact consists of:
  - A public github following best practices for DS project development (a developer should be able to extend your work from this code).
  - A public website explaining the project to an intended audience.
  - A pdf report, following best practices in scientific writing.
  - An elevator pitch (e.g. helpful in job interviews).
  - A longer slide-based talk.
- https://dsc-capstone.github.io/projects-2020-2021/

Syllabus and Assignments

Syllabus

Component	% of Grade
Methodology HW	5%
Discussion Section Participation	10%
Domain Q1 Report (2 reports)	50%
Code for Domain Q1 Report	20%
Project proposal	15%

Assignments: Learning the Domain (Q1 Project)

You will write a report that summarizes your work from learning your domain.

Part 1 (Checkpoint):

- Report: Introduction to problem, description of data and/or methods
- Code: Data Processing, initial method implementation

Part 2 (Final):

- Report: results, discussion of shortcomings and possible improvements
- Code: Code that produces the results, using best practices for your project.

Assignments: The Proposal

- Worked on in groups (same as your project).
- Write and submit a proposal, with background research.
- Write a plan/schedule for executing your work.
- Rehearse and deliver a 2-3 minute elevator pitch (general audience)
- Create a skeleton workflow for the project (github repo with boilerplate).

You domain mentor will approve your proposal. Sticking as close to the Q1 work as possible lets you move faster in Q2.

Your group will work on and present the project in Quarter 2!

Assignments: Methodology / Participation

- Methodology HW
  - Short homeworks focused on DS best-practices, applied to your domain.
  - E.g. define a reproducible software environment for your project.
  - E.g. Learn to structure your project
- Participation HW
  - Default assignment (on website) asks you to respond to domain reading/tasks.
  - Due **before** section; meant to prepare you for for discussion in section.
  - Your mentor may create their own questions for you to answer
- Participation (end of Quarter)
  - Mentor will assign a participation grade for discussion at end of quarter, as well

Grading

Assignments are graded by a combination of Domain Mentors and TAs

- Domain Mentors will grade your reports:
  - Your reports should make it clear that your code is reasonably close to correct!
  - Domain Mentors may give you feedback in OH, instead of written feedback.
  - Their feedback will be from the standards of the domain!
- Course TA's will grade your code and other assignments according to a rubric for what's taught in lecture.
- All assignments graded as A/B/C/F only (no plus/minus). Final grade computed by the standard GPA conversion.

Advice

- Work slow and steady. This material is *hard* and you will hit unexpected obstacles.
- Ask Questions. Ask Questions. Ask Questions.
  - Access to a mentor like this is rare at UCSD!
  - Research is deceptively hard -- if you are confused, others likely are too.
  - Domains benefit from discussions and working together!
- Don't be afraid of redoing work. You will rewrite your code many times.
  - It doesn't mean it was wrong the first time; it means you understand it in a different way.