DSC Capstone Sequence

Lecture 01
2020-2021

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).
- Computing Resources: DSMLP Server
- Gradescope: for code assignments and methodology HW.
- Course communication:
  - Canvas for lecture/administrative questions
  - Look up course staff directory for domain mentors' prefered mode.

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": 1hr each methodology and domain OH (unscheduled).
- Scheduled Friday's 9-11 will sometimes be used for class events.

Capstone Domains (2019-2020)

Domain	Data	Methods	Groups
“Fair Policing”	(Local) open gov data	Causal Inference	4
“Conflict on Wikipedia”	Wikipedia edit histories (large XML)	Social Analysis	3
“Quantitative Analysis of Artistic Style”	Image data	Hand-coded features; (interpretable) CNNs	6
“Clustering the Human Genome”	DNA Sequence Data	Clustering and Association Studies	6
“Malware and Heterogeneous Graphs”	Disassembled Java Code (Unstructured)	Machine Learning on Graphs	8

Capstone Projects (sample: 15/28)

Genome-wide association studies on Alzheimer's disease	Analyzing YouTube thumbnail trends	Analyzing the media's effect on San Diego Police traffic stops
Predicting disease through genome wide association studies	Understanding a book by its cover	The effect of predictive policing across communities in Los Angeles
Germ layers and cancer	Quantifying and analyzing ragas in Indian classical music	Engagement and collaboration on Wikipedia over time
Understanding miRNA in pre-Type 1 diabetes	Predicting Malware in Android Apps with Heterogeneous Graph Embeddings	Understanding the spread of COVID-19 through the scope of the media and Wikipedia
Detecting plagiarism via AST similarity	Probabilistic Record Linkage	Improving HinDroid with metapath2vec

Capstone Domains (2020-2021)

VPN-XRAY (Viasat)	Particle physics (Wuerthwein/Duarte)	Graph data analysis (Mishne)	Malware & Graph Learning (Fraenkel)
Evaluating recommender systems (Jemmott)	Cyber-Physical Systems using IOT Devices (R Gupta)	Opioid overdose prevalence analysis (A Gupta)	Conflict/collaboration in online communities (Fraenkel)
The robustness of autonomous vehicles (Silberman)	Large-scale Multiple Testing (Schwartzman)	Spatial-temporal Analyses of Infectious Disease Dynamics (Ma)	The Spread of Misinformation (Eldridge)
COVID-19 & microbiome (Knight)	Text mining and NLP (Shang)	The genetic basis of mental health (Ellis)	Explainable AI (Schultze)
System Usage Reporting (Intel)			

A Closer Look: Methodology

- 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

- A domain of inquiry usually consists of a paper for students to replicate:
  - Introduction to area
  - 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 an invaluable resource -- take advantage!
- For questions about the domain, contact your mentor with their preferred mode of communication.

Note: Mentors largely operate unaware of how the larger sequence 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.
- Discussion 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 (groups 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-2019-2020/

Syllabus and Assignments

Syllabus

Component	% of Grade
Methodology HW	10%
Discussion Section Participation	10%
Domain result replication (3 reports)	30%
Domain result replication (workflow)	20%
Project proposal	30%

Assignments: Learning the Domain (replication)

1. The Data
    a. Report: Introduction (to problem, data and/or method)
    b. Code: Data ingestion code, using best practices.
2. Cleaning/EDA and beginning Methods
    a. Report: Summarize results of EDA; defense of choice of cleaning code.*
    b. Code: Cleaning and EDA code, using best practices.*
3. The Result Replication
    a. Report: Summary of result of the ‘replication’, shortcomings, and possible improvements.
    b. Code: Replication code that produces results, using best practices.

Note: Certain domains may tweak the subjects covered in these checkpoints.

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
  - E.g. Analyze the paper in your domain for effective writing.
- Participation HW
  - Default assignment (on website) asks you to respond to domain reading/tasks.
  - Due **before** section; meant to prepare you for section.
  - Your mentor may create their own questions for you to answer.

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.