BIS 687 Data Science Capstone

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data, and apply knowledge and actionable insights from data across a broad range of application domains. Its effective practice is as tied with software design and development as much as the statistical skills of analyzing data and communicating those results. This course prepares students to transition from the classroom to the real-world practice at the intersection of biostatistics and data science. Students develop a holistic solution to an analytical problem by proposing study aims, hypotheses, and system design and then develop a robust, reproducible solution addressing said hypotheses. Moreover, as biostatisticians must be able to clearly communicate their findings to fellow statisticians and the domain experts with whom they collaborate, this course provides multiple opportunities for students to present their work orally and in writing. As in statistical practice, there are opportunities for problem-solving and decision-making at the individual and group level.

Objectives

The goal of this course is to bring together the skills students have acquired during their last year-and-a-half to provide a data analysis touching on all aspects of the research cycle. This includes:

1. Either meeting with or acting as an investigator to understand a scientific challenge (or challenges) that can be addressed through data science and statistics.
2. Develop a research plan that includes deliverables to present your work.
3. Develop a research schedule with milestones, decision points, and a project management plan.
4. Receive and integrate critiques from classmates, an investigator, and the instructor.
5. Implementing the research plan and keeping to the proposed schedule.
6. Presenting your work and submitting a final research report.

Office Hours

Michael: Wednesday after class by zoom with appointment.

Materials

Text

The text for this class will be Git Essentials Developer Guide to Git. We will begin using Github when we start creating an R package for the class. So it is recommended that you've read the book by week 5.

Online resources

1. Advanced R Programming
2. The Python Data Science Handbook
3. The Python Data Science Tutorial
4. R for Data Science

5. ggplot2: Elegant Graphics for Data Analysis (Use R!)
6. R Packages
7. Interactive web-based data visualization with R, plotly, and shiny

Class

Classes will meet on Wednesday from 3 - 4:50 PM in LEPH 103.

Topics by week

- Defining a research project (weeks 1-2)
  - Scientific research questions vs. statistical hypotheses.
  - The Specific Aims and Research Strategy for an R21 grant
  - Augmenting your writing using ChatGPT
- Defining a research plan with milestones and deliverables (weeks 3 - 4)
  - Github projects
  - Issues and milestones
- Students present their proposals (weeks 5-6)
- Instructor's choice (weeks 7-8)
- Students provide informal progress reports (weeks 9-10)
- Instructor's choice (weeks 11)
- Final reports and presentation (weeks 12 - 13)

Unsolicited advice and suggestions

- Be able to explain your scientific research questions without using statistics words like "data", "variable" or "association." You will answer the scientific question with statistics. If you conflate the two you will end up asking a statistics question that a scientist may not understand or care about.
- It's easier to have everyone push to the same repository but less correct. Integrating changes with pull requests is better to do with more than one person reviewing the request.
- Meet with your teams weekly to discuss progress and adjust milestones.
- If you put more time into the first few weeks, the rest of the class will go more smoothly.
- A useful tweet on the limits of ChatGPT

Grading

- Proposal and milestone: 20%
- Proposal presentation: 10%
- Proposal critiques: 10%
- Progress presentations: 20%
- Progress critiques: 10%
- Write-up: 20%
- Final presentation: 10%