Data Scientist's Toolbox Study Group

Clifford Anderson-Bergman

cliff.andersonbergman@gladstone.ucsf.edu
Gladstone Institutes

What is Data Science?

- The current hot trending career
- The real world analysis of data
  - Applied use of statistical methods
  - Management/cleaning of data
  - Use and creation of automated data tools

Data Science for UCSF researchers

- Most data scientists need to deal with large piles of data and try to find *some* manner of answering of a potentially vague question, often in an automated manner
- Most UCSF researchers have an extremely laboriously created smaller dataset with a well defined question they want analyzed once so they can publish
- May be less concerned with handling and manipulation of large datasets than traditional data scientists
- Still share many tools for analysis of data!

What to expect from this series

- First course: basic introduction to common tools by data scientists
  - R/Rstudio, for data analysis
  - Github, for management and sharing of data tools
  - Command Line
  - Should be 1-2 hours of work a week, 4 weeks
- Later courses: more detailed explanation of various tools
  - If only interested in learning tools required for UCSF research, may only want to take relevant courses
  - If interested becoming a data scientist, strongly suggest taking all courses
  - Later courses more of a time commitment (4-9 hrs/week, 4 weeks each)

What to expect from this study group

- We are not grading you (Coursera is)
- We are here to help you out with difficulties you may face during the course
- We do expect you to have watched the videos, but we are happy to help explain anything that you did not find clear
- We are happy to help you with small issues related to data analysis outside the course (i.e. "how do you load in this excel file to R?"), but within reason (i.e. not "how do you analyze fMRI data?")
    - Coursera questions will always be prioritized

Summary of Important Topics this Week

- Data Science Tools
  - R/Rstudio
    - Very popular program for data analysis
    - Other alternatives (not for this course): S+, Python, SAS, Stata, Excel, Prism
  - Github
    - Software for keeping track of changes in programming code
    - *Extremely* useful when working on large projects where you must share code with other researchers
  - Command Line
    - MS-DOS like tool for interfacing computer via text commands, rather than GUI
    - Typically used by most serious programmers

Summary of Important Topics this Week

- This 4 week course will very briefly introduce these tools
- After this course, various topics will be covered much more detail in further courses (each 4 weeks long)
    - Some courses will be very relevant for analyzing biological data
    - Some very specifically for data scientists

Summary of Important Topics this Week

- Courses important for UCSF research
  - R Programming
  - Exploratory Data Analysis
  - Statistical Inference
  - Regression Models
- Courses for focussed on traditional data science
  - Getting Data
  - Reproducible Research
  - Practical Machine Learning
  - Developing Data Products