Data Analytics for Social Good

An introduction to Data Science for Freshman

Robin Donatello, Chico State
https://www.norcalbiostat.com/
2021-02-19

Outline

- How is Chico approaching Data Science Curriculum?
- Identified Barriers & Proposed Solutions
- Outline & Objectives of a "Data Science for all" exposure course

DS at Chico - The short short vrs

- Similar problems/path as Cal Poly
  - Idea for program started in 2013
  - Proposed cohort hire: 1 ea. Business, Computer Science, Statistics
  - Ended up with 1 hire (me in '14)
- Lots of interest, but few boots on the ground.
- No formal guidance or vision from administration.
  - Good: I could build whatever I wanted
  - Bad: I have no idea where to start or what was possible.
- 1 year researching & building collaborations
  - Chico campus needs assessment - who was interested, and what do they want?
  - How does the curriculum processes work, barriers, restrictions, and available/appropriate "boxes" for such an interdisciplinary program.
- Resulted in an Undergraduate Certificate in DS (F 19)
  - Targeted to current CS & Stats students
  - +8 units for CS, +12 units for Stat
  - Intro DS class has a Calculus 1 pre-req.

See more about the Data Science Initiative at https://www.csuchico.edu/datascience/

Current Barriers

1. Calculus tends to be avoided
2. Late awareness of DS program (Jr/Sr)
3. Non DS faculty want their students to have DS skills, but don't want to teach it.
    - This is actually an opportunity!

Data 185: Analytics for Social Good

- No pre-req course to pull from a broad range of students. Carries GE QR credit as the hook.
- Use the DATA8 model to encourage connector courses
  - DS for smart cities (civil Engineering), Exploring Japanese-American Internment through Digital Sources (History), Immunotherapy of cancer (Biology)
- Use concepts from Data Studies courses at UC Davis and UW to introduce the impacts of Data & Data Science on our global society.
- DATA 185 could be integrated into existing programs as
  - alternative Calculus pre-req for the 300 level "Intro to DS"
  - substitution for Intro to Stats (GE level stats) for certain courses

Other plans to expand DS into the graduate space are also being explored

Analytics vs Science

- Why "Analytics" instead of "Science"?
- IMO, DS is the overarching, large umbrella term that *includes* analytics
- "Data Literacy + Mathematics"
- The focus of the class isn't modeling, predicting, optimizing, but thinking critically with data.
- Balance between
  - letting them eat cake and playing with black boxes, and
  - enough mathematics to qualify for GE B4 Quantitative Reasoning credit
- Aim to describe graphs, tables, up to linear models and classification models as mathematical models
- No formal hypothesis testing, probably light discussion of randomization as a way to talk about the concept of likelihood and uncertainty

DATA 185: Course Description

Data is not neutral, nor are the algorithms that control how the data that governs our lives are used. Interpretations and recommendations made using data are subjective. This course introduces students how to start harnessing the power of data to intelligently cope with the requirements of citizenship, employment, and family to be prepared for a healthy, happy and productive life. [1]

Students will practice collecting and wrangling data into a usable form, visualizing large data sets to discover patterns, representing data in a meaningful way, exploring varying interpretations of the data and results, and discussing potentials for misuse and abuse. This course promotes critical reflection on the ethical, social, cultural, and political dimensions of data as well as providing direct hands on experience with both spreadsheets, and the programming language R.

3 units: 2 hours lecture (online), 2 hours lab ( in person).

[1] Drawn from *Guidelines for Assessment and Instruction in Statistics Education (GAISE) K–12 report* of the American Statistical Association

DATA 185: Course Objectives

The topics and outcomes in this course are "low floor & high ceiling" [2] outcomes. There is great potential for upward growth and exploration, but as this is an introductory course only an introductory level of proficiency is required.

CO1 Explain how the method of data collection, the involvement of stakeholders, and decisions made during data processing can have downstream impacts.

CO2 Interpret and draw inferences from Analytical models while taking into account the socio-political-ethical-economic implications of conclusions being made.

CO3 Use Data Analytical tools and processes to uncover insights in a large data set.

[2] Credit to Paul Bailey for this phrase.

Obj. #1: Get & Process Data

> **CO1** Explain how the method of data collection, the involvement of stakeholders, and decisions made during data processing can have downstream impacts.

- Collect, record and organize (tidy) data in spreadsheets.
- Import, wrangle, and explore data using the statistical programming language R.
- Assess data integrity, identify potential privacy issues, and decide what questions can and cannot be answered with the data.
  - e.g. if the question is about the economic impact of no-kill animal shelters, the data needs to contain geographic information, economic information for those geographies, information about the shelters such as intake, adoption & euthanasia rates.
- Identify stakeholder involvement at each stage in the data lifecycle.
  - what data was collected, from whom, who decided how to transform measures (race/ethnicity categorization 2+), what results to share and using which graphics, and to what audience for what purpose?
- Create a reproducible data processing pipeline that explains all data collection and processing decisions in a literate and transparent manner.
  - Rmarkdown FTW

Obj. #2: Understand impact of data

> **CO2** Interpret and draw inferences from Analytical models while taking into account the socio-political-ethical-economic implications of conclusions being made.

- De-mystify Analytical models using words, numbers, graphs, symbols, and equations.
- Explain the relationship between Analytics and Optimization
  - Analyzing trends vs finding most effective solution
- Identify ways in which Data Analytics can contribute, or harm, the cultural and economic well being of diverse societies in local, national, and global scopes.
  - Weapons of Math Destruction Cathy O'Neil
- Describe an algorithm as a mathematical model, explain why they are created and how they are used, and provide examples of what it means to encode bias into an algorithm.
  - Algorithms of Oppression, Safiya Umoja Noble
- Critically examine and critique data used in reports, news articles and advertisements.
  - Calling Bullshit
  - "The Cause of Autism may have been discovered" @justsaysinmice
  - NYT "What's going on in this graph

Obj. #3: Data Discovery

> **CO3** Use Data Analytical tools and processes to uncover insights in a large data set.

Cumulative Project: Create a literate, reproducible data analysis report that combines words, code, graphs, and equations to tell a story.

- Work on throughout the semester
- Students pick a topic area that they are passionate about
- They find data on the source, submit for review & approval
- Conduct an EDA - using tables, graphs, words
- They ask their own questions about the data.
  - Both as words and as mathematical statements
- Creative license on dissemination of results.
  - YouTube, poster, slide presentation, website, paper report, music video, interactive dance....

Current Build Status:

https://norcalbiostat.github.io/DATA185/

Things I'm still figuring out

- Keeping the course free
  - R Studio Cloud is *amazing* for teaching out of, but we'd need more time than the free tier.
    - Talk Deans into paying for Cloud
    - Talk ITSS into installing & managing RStudio Server with SSO sign on.
  - Alternative possibility - VPN onto campus computers with R Studio+MS One Drive.
    - Downside: Reliance on ITSS for package management.
- Discord vs Zoom+Slack
  - Test run of Discord during March workshop on how to build a website

What questions do you have?