Implementing a Data Science Course in Secondary Schools

Statistical Society of Canada, May 2019

rgould@stat.ucla.edu

Outline

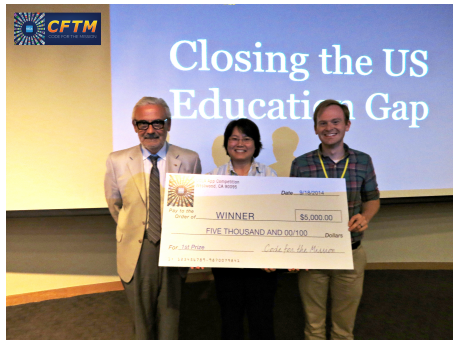
- A brief(ish) overview of the Mobilize Introduction to Data Science curriculum
- Challenges we faced, and expected
- Challenges we faced, but did not expect, although in retrospect we probably should have.

https://www.introdatascience.org/

mobilize

IDS

Introduction to Data Science



Suyen Moncada-Machado, James Molyneux, Amelia McNamara, Terri Johnson, LeeAnn Trusela, Hongsuda Tangmunarunkit, Steve Nolen

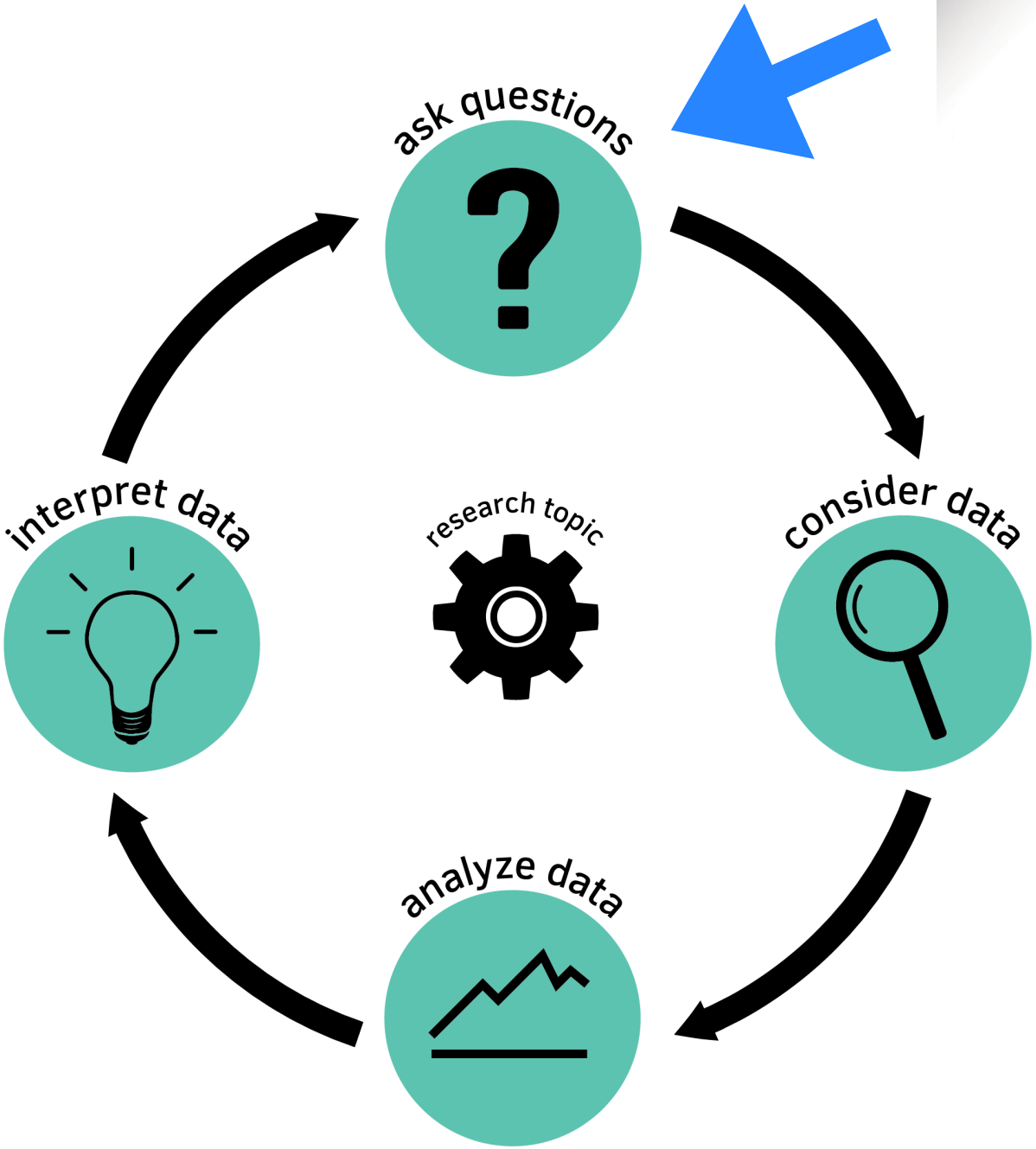
Participatory Sensing

"At its heart, participatory sensing is data collection and interpretation. … Participatory Sensing emphasizes the involvement of citizens and community groups in the process of sensing and documenting where they live, work, and play. It can range from private personal observations to the combination of data from hundreds, or even thousands, of individuals that reveals patterns across an entire city. Most important, Participatory Sensing begins and ends with people, both as individuals and members of communities."—Burke, Estrin, et. al. (2006)

Goals of the Course

- Develop Statistical Thinking (GAISE A/B level)
- Learn to analyze data of a variety of types using R (via Rstudio)
- Understand social implications and issues surrounding data.
- In a nutshell: it is not about computing things, it is about computing things so that you can interpret the data to answer meaningful questions about real-world things.

The Data Cycle



Implementation

- Began as collaboration with Los Angeles Unified School District and 10 teachers.
- LAUSD: 650,000+ students, 20% English learners, 20% not fluent in English, 50% "very low" economic group, 90 languages and 7 major ethnic groups.
- 2019-20 academic year: 16 school districts, 51 schools.
- 2019-20: 121 sections, approx. 4235 students/year
- By Fall 2019 IDS will have prepared 106 teachers to teach data science.

Components

- In-class activities with guidelines for teachers to develop conceptual understandings, terms, methods.
- Computer labs where students learn the R language and practice data analysis (using PS data, open data, data scraped for purposes of this class)
- Participatory Sensing “campaigns” where students collect and analyze data
- Practicums: projects to tie units together

4 Units, each about 9 weeks

- Unit 1: Focus on data
  “ This unit will introduce the idea of ‘data,’ fundamental to the rest of the course”

- Unit 2: Informal inference using randomization paradigm
  “ This unit deepens the “informal” statistical reasoning skills developed in Unit 1 by enriching students' technical vocabulary and developing more precise analytical tools. Most importantly, this unit introduces the formal concept of probability as a tool for understanding that sometimes patterns observed in data are not ‘real.’"

- Unit 3: Data Collection
  “focuses on data collection methods, including traditional methods of designed experiments and observational studies and surveys. It introduces students to sampling error and bias, which cause problems in analysis made from survey data”

- Unit 4: Predictions, Multivariate
  “ This unit will develop modeling skills, beginning with learning to fit and interpret least squares regression lines and learning to use regression to make predictions. Students will learn to evaluate the success of these predictions and so compare models for their predictive accuracy.”

★ Introduction to data, cultural issues, distributions

- Visualization of distributions
- Exploratory data analysis/summary statistics
- Basic probabilities through simulation

★ Informal inference with randomization based testing

- The Normal distribution
- Controlled experiments/random assignment
- Observational studies, confounding factors

★ Survey Sampling/writing questions

★ Humans as sensors to collect data

★ Scraping data from html tables on the internet

- One-variable regression, prediction emphasis

★ Multiple-regression, prediction emphasis

★ Model Eliciting Activity as summary activity

★ Classification and Regression Trees, Clustering with K-means

Lesson 1: Data Trails

Objective:

Students will understand what are data, how they are collected, and possible effects of sharing data.

Materials:

1. *The Target Story* video: https://ids.mobilizingcs.org
2. Data Science (DS) journal (quad-ruled composition book or similar); MUST be available for every lesson
3. *Data Diary* handout (LMR_1.1_Data Diary)
4. *Terms and Conditions* video (https://www.youtube.com/watch?v=ZcjtEKNP05c)

Vocabulary:

data, observations, data trails, privacy

Essential Concepts: Data are a collection of recorded observations; data are gathered by people and by sensors; patterns in data can reveal previously unknown patterns in our world; data play a large, and sometimes invisible, role in our lives.

Challenges (Expected)

- Must meet state standards
- Needs to be part of a pathway to university/college.
  - alternative to algebra II
  - computational pathway to supplement trad. math
- Professional Development

Professional Development

- For the most part, teachers have had no preparation in data analysis, although some (a few) have had a mathematical stats course at some point.
- For the most part, teachers have had little to no background in programming/coding.
- Safely assume P(know R) = epsilon

PD Goals

- Learn to use the technology: Rstudio, participatory sensing
- Learn to analyze data, learn to teach students to analyze data
- Become familiar with the curriculum itself.
- Develop fluency in student-centered, equity-driven pedagogy

PD Schedule

- One 4-day “summer institute” at beginning of summer
- One 3-day institute at end of summer
- Five 6-hour saturday-workshops throughout academic year.
- Year 2 "academies"
- Support community centered around Google group.

Surprise Challenges

- When students discuss real-life situations, the discussion can be less comfortable for a math teacher than when students discuss, say, the quadratic equation. The same can be true for students.

- High School Counselors are much more important than we immediately suspected.

- Assessment is key

    - formative assessment

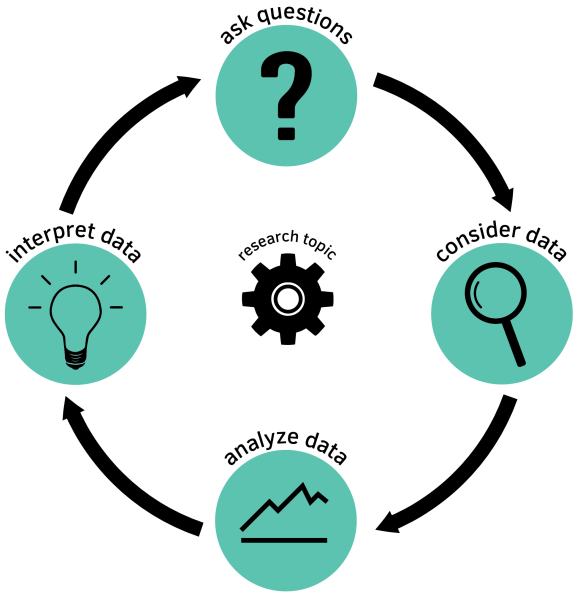
    - summative research assessment: how to measure the synthesis of computational and statistical thinking?

LOCUS

Levels of Conceptual Understanding in Statistics

https://locus.statisticseducation.org/

The Data Cycle



Browse questions by component

- Formulate Questions
- Collect Data
- Analyze Data
- Interpret Results

Continuing Challenges

- There's no research to guide us. What should we expect 14-year olds to understand about data science? 18 year olds? In what order should topics be presented?
- Glimmers from our observations:
  - third-year students do better than 1st or second
  - exposure to geometry helps. (But why??)
  - experience with programming helps a lot.
  - taking the course a second time helps a lot.
  - experienced teachers helps a lot

THANKS!

https://www.introdatascience.org/

rgould@stat.ucla.du