A Project-Driven Introduction to Data Science

Karl Schmitt,
Valparaiso University

Outline

- Course Overview & Context
- Why projects? Why early?
- How to have ‘successful’ projects
- Things I haven’t figured out yet
- Where to find resources
- Questions

Institutional Context

- Valparaiso University:
  - About 3500 undergrads
  - About 800 engineering students (28 faculty)
  - Lutheran (faith-based) institution
  - 15 full-time Math/Stat faculty (14 Tenure-Track)
  - 9 full-time CIS faculty (4* Tenure-Track)
- Data Science Program
  - Housed in Mathematics & Statistics (MST)
  - *Director holds affiliate appointment in Computing and Information Sciences (CIS)
  - Started in Fall 2016
  - Graduate Program in Analytics and Modeling since 2011

Course Overview/Context

- First* course in “Data Science” (Major)
  - Minimal prerequisites
- Offered in SPRING semesters
- Typically taken by:
  - Freshmen/Sophomores (majors or Stat/Math)
  - Junior/Seniors (CS or Engineers)
- 15 Weeks, Offered as 2+3:
  - ‘Lecture’ 2x a week (50 min)
  - Lab/Project Time 2x a week (75 min)
  - All sessions meet in a computer lab

Course Overview/Context: Learning Goals

- Course Goals:
  - Understand the fundamental concepts of data science and knowledge discovery
  - Apply and perform the basic algorithmic and computational tasks for data science
  - Develop and improve analytical thinking for problem formation and solution validation

Course Overview/Context: Learning Goals

- Topical Objectives:
  - Gain an overview of the field of knowledge discovery
  - Learn introductory data mining algorithms
  - Be able to distinguish and translate between data, information, and knowledge
  - Apply algorithms for inductive and deductive reasoning
  - Apply information filtering and validation on real world datasets
  - Understand the social, ethical, and legal issues of informatics and data science
  - Apply data mining, statistical inference, and machine learning algorithms to a variety of datasets

Why Projects?

- Evidenced-based best practice for engaging minorities (and students in general)[1]
- Real data is messy![2]
- Prepares them for realistic workforce experiences
- Portfolio/Resume building

[1]Corbett, C. and Hill, C. 2015. Solving the equation: the variables for women's success in engineering and computing. DC: AAUW. (2015).
[2]Committee on Envisioning the Data Science Discipline: The Undergraduate Perspective et al. 2018. *Data Science for Undergraduates: Opportunities and Options*. National Academies Press.

Why early?

- Provide context for topics/learning in future classes
- Builds important communication and self-management skills
- Evidenced-based best practice for improving retention! (real data)
- Establishes resume/portfolio for early internship applications
- Better prepares them for doing undergraduate research
- (Potentially) provides a unified topic/dataset for many future classes
- It’s fun..
  And why they to be Data Scientists!

HAVING ‘SUCCESSFUL’ PROJECTS

(1) MANAGE EVERYONE'S EXPECTATIONS!

This is hard. But by far, the most important.

(a) Carefully define "success"

- For Students
- For Clients
- *From an Instructor perspective

(b) Established expected project outputs/deliverables early

(2) Start getting projects EARLY

(even earlier than you think you need to)

This year I started in mid-November…

(‘war story’ of issues with this)

(3) Have several mile-stones during semester

Each mile-stone has ‘sub’-deliverables.

I use 5 phases:

1. Proposal & Design Sketch
2. Data Processing & Design Specifications
3. Algorithm Plans (and Problem Revisit)
4. Basic Data Pipeline/Product
5. Final Paper (with 3 sub-stages here)

(4) (Carefully) Assign Teams

Students rank projects.

I assign teams, striving for:

- Balance of grade level/majors
- Mix of (previous) skills needed for projects
- Non-isolation of minorities/gender
- Personalities

Things I changed in Year 2…

- Heavier front-loading of programming assignments
- Required use of GitHub (sorta…)
- Included significant in-class time for working on projects
- Had clients present the outcomes/use of projects at the end of the semester

Things I haven’t figured out:

- Getting students to have more professional engagement with clients
  - Limit number of interactions
  - Better scheduling and student commitment
- Course is still a LOT more work than most introductory courses
  - FOR EVERYONE – Students and faculty
- Where/how (website/wiki) to have students post materials for general sharing
- How much programming to front-load
- How much project-management concepts to front-load

Things I’m changing next year:

- Senior student project managers
  - Former students (as TAs), majors, etc.
- Explicit and limited (5-6) meetings between students and clients
- Getting projects even earlier!
  - Especially the data!
- Extended in-class discussions of each phase/mile-stone expectations
  - Provide exemplars (and failures)
- (Maybe) Start working on projects a little later (last time, Week 2 of course)

Resources

- SIGCSE and SIG-STATED List-Serves
- Projects:
  - Kaggle.com
  - Challenge.gov
  - Riipen.io and Telanto.com
- http://www.teachingdatascience.org/
  - Sort of defunct
- ‘From the Director’s Desk’
  - blogs.valpo.edu/datadesk
  - A blog about data science education and curriculum

Thank you for listening!

Questions?