Discussing the History of Ideas in a Data Science Seminar

Paul Anderson
Kelly Bodwin
Alex Dekhtyar
Hunter Glanz
California Polytechnic State University
San Luis Obispo, CA, USA

Foaad Khosmood
Lubomir Stanchev
Dennis L. Sun
Jonathan Ventura
California Polytechnic State University
San Luis Obispo, CA, USA
jventu09@calpoly.edu

ABSTRACT

As one part of an NSF-sponsored Data Science Fellowship at Cal Poly, San Luis Obispo, a group of faculty offered a unique one-unit quarter-long seminar on the history of ideas behind the core principles of Data Science. We present an overview of this seminar, its learning objectives, and outcomes and lessons learned.

1 INTRODUCTION

As part of the NSF HDR-supported Central Coast Data Science Partnership, a group of Cal Poly faculty developed a Data Science (DS) Fellowship program. The program places 12-18 students per year in direct contact with eight faculty members, via individual research apprenticeships, and via a weekly DS Seminar organized for the students. Here we discuss the Fall quarter curriculum of the seminar, a the history of ideas behind DS.

2 OVERVIEW OF THE SEMINAR

The core of the curriculum is eight weeks of discussions of assigned readings.The eight reading assignments were selected to highlight the development of important ideas behind Statistics, Data Science, and Machine Learning. These readings are organized into three sections:

(1) Development of modern statistical thought
   - Week 2: History of hypothesis testing [2].
   - Week 3: Influence of eugenics on statistical thought (and how to approach the darker pages in the history of science) [1].
   - Week 4: Frequentist vs. Bayesian approaches to Statistics

(2) Machine Learning and AI
   (a) Week 5: Statistical (interpretive) vs Algorithmic (predictive) models in Machine Learning.
   (b) How likely is the AI Winter?

(3) Big Data Technologies
   (a) The history of "Future directions" in Database research
   (b) The rise and fall(?) of MapReduce
   (c) Development and influence of the World Wide Web

Each week's reading assignment involves one or two main articles or book chapters, and, sometimes, additional reading materials to provide context for the upcoming discussion. The full list of readings, both main and additional will be included in the poster.

For each discussion, we break the students into groups of six, with different groups each week to enable students to communicate with all of their peers. Within each group, we designate three roles, informally known as the "proponent", the "opponent", and the "summarizer", with two students per group assigned each role.

The proponents present arguments in favor of the core points supported in the assigned readings. The opponents present arguments in favor of the opposing view. The summarizers either present a succinct summary of the discourse, or compare and contrast the opposing points of view.

Each week, the first 25-30 minutes of the 50 minute class are spent in group discussions, with one or two faculty members facilitating the conversations. Prepared remarks are followed by a free-form discussion in which we discussed the actual opinions of the students (and faculty) on the subject matter at hand. Finally each group "reports out" to the entire class on their group's discussion, concluding remarks are given, and the last five minutes of the class are spent previewing the next week's reading assignment.

3 LESSONS LEARNED

The two core objectives of the seminar were (a) exposure of students to important ideas that determined the development of modern data science and machine learning, and (b) engaging students in conversations about data science.

We found that students were hesitant to engage in free-form discussions, but were more willing to engage in structured discussions as described above, where they had a specific role to play/point of view to present, *even if that point of view did not match their own.* We noted that students with various technical backgrounds and different levels of machine learning/data science skills were able to engage in productive discussions with each other. The discussions of ideas behind data science and machine learning provided important context and complemented the growth in technical proficiency of the students during their Fellowship year.

ACKNOWLEDGMENTS

This work was supported by NSF Award #1924008.

REFERENCES

[1] Angela Saini. 2019. *Superior: the return of race science.* Beacon Press.
[2] David Salsburg. 2001. *The lady tasting tea: How statistics revolutionized science in the twentieth century.* Macmillan.

SIGCSE 2023, March 15–18, 2023, Toronto, ON, Canada

ACM ISBN 978-1-4503-9433-8/23/03.
https://doi.org/10.1145/3545947.3576274