DATASCI W210

Synthetic Capstone

Course Description

In the Capstone class, students combine technical, analytical, interpretive, problem-solving, and strategic thinking dimensions to design and execute a full end-to-end data science project. Students will develop their technical and non-technical skills as data scientists who focus on real-world and impactful applications and situations. The final project provides a learning opportunity and "sandbox" to integrate all of the core skills and concepts learned throughout the MIDS program and to provide necessary experience and hands-on tools in formulating and implementing an impactful and compelling project. Students are evaluated on their ability to work in a dynamic team environment to collaborate, co-develop, and communicate their work in both written and oral forms. The capstone project is completed as a group project, and each project will leverage open, pre-existing, secondary data. The project experience is supplemented by readings and case discussions throughout the course. Through immersive skill development in technical, strategic, organizational, and communication areas, the course prepares students for long-term professional success in the rapidly-evolving field of data science.

Learning Objectives

Upon Completion of the course, the student will be able to:

1. Identify and articulate a problem space to address through application of data driven methods, approaches and practices that includes an understanding of stakeholders, social contexts, potential impact, and potential obstacles.

2. Demonstrate an ability to integrate and synthesize knowledge and skills gained through other courses in the program (critical technical, analytical, strategic thinking, problem-solving, communication, influencing, and management skills) in developing and implementing a capstone project that addresses a data science problem.

3. Effectively engage in a process of teamwork, feedback from peers, instructors and experts, and informed iteration that mirrors the challenges and opportunities of applying data science in a realistic organizational setting.

4. Assess and select data and the data collection methods that best fit the specific outcome or need of a project or problem space.

5. Understand, incorporate, and practice integrated understanding of what it takes to imagine, design, and execute a data science project from start to finish.

6. Demonstrate proficiency in applying technical and analytical skills towards the collection, storage, manipulation, and analysis of data towards problem-solving and project execution.

7. Construct and perform persuasive, informative and understandable written, spoken, or visualized narratives that concisely convey findings, solutions, and applications of data-driven approaches that have been incorporated in project work.

Course Summary

We begin by covering fundamentals for beginning a project, including problem ideation and structuring, selecting tools/approaches, product planning, and managing ethical and legal considerations. We then broaden our focus to execution, including exploratory data analysis, data privacy, model evaluation, and scalability. We conclude with advanced topics like influencing in organizations, building products, and perspectives on the future of data science.

Prerequisites: Students must have completed (or are completing during the same semester) all other courses necessary for degree completion.

Assignments and Grading Policies

The course assignments have been selected to enable project teams to make and measure progress towards the completion of a capstone project. Their weight in the final grade is:

- **5% - Individual Project Proposal**
- **10%** - **Team Process Agreement**, **Team Project Plan**
- **10% - Presentation #1** - Focus on problem and impact
- **15%** - **Presentation #2 –** Focus on technical approach
- **20%** - **Presentation #3** – Focus on demo and results, lessons learned
- **20% - Web-based final deliverable** - examples of past projects can be found here: https://www.ischool.berkeley.edu/programs/mids/capstone
- **20% - Participation and Teamwork** - class discussions and group project participation

Detailed rubrics will be made available for each graded component. Completing the Web-based final deliverable, and uploading it to the MIDS program website, is a requirement for this course. For most projects, the final deliverable will explain the work in detail and showcase its interactive functionality. For projects without interactive functionality, it is sufficient to explain the work and showcase the results. Students are required to share all working code with their instructors and are encouraged to incorporate key examples into their deliverable.

Examples of past projects can be found here:
https://www.ischool.berkeley.edu/programs/mids/capstone.

Upload instructions for the final deliverable can be found here:
https://www.ischool.berkeley.edu/intranet/students/mids/capstone.

A note on participation in live sessions: We believe in the importance of the social aspects of learning: between students, and between students and instructors, and we recognize that knowledge-building is not solely occurring on an individual level, but that it is built by social activity involving people and by members engaged in the activity. Participation, teamwork, and communication are key aspects of this course that are vital to the learning experiences of you and your classmates. Required live session activities towards this end may include the Week 2 presentation of 3 ideas, helping to lead in-class workshops and/or summarizing papers for the class.

To maximize opportunities for live session engagement, we like to remind all students of the following requirements:

- Students are required to join live class sessions from a study environment with video turned on and with a headset for clear audio, without background movement or background noise, and with an internet connection suitable for video streaming.
- Students are expected to engage in class discussions, breakout room discussions and exercises, and to be present and attentive for their and other teams' in-class presentations.
- Students should keep the microphone on mute when not talking to avoid background noise. Do your best to minimize distractions in the background video, and ensure that your camera is on while you are engaged in discussions.

That said, in exceptional circumstances, if students are unable to meet in a space with no background movement and with a good internet connection, they should arrange with the instructors (beforehand if possible) to explain the situation. Sometimes connections and circumstances make turning off video the best option. If this is a recurring issue in students' study environment, they are responsible for finding a different environment that will allow full participation in classes, without distraction to classmates.

Failure to adhere to these requirements will result in an initial warning from the instructor(s), followed by a possible reduction in grades or a failing grade in the course.

Course Data Sets, Software, and Tools

This course allows students wide discretion for selecting problems, analyses, data sets, software, and tools for use on their group projects.

Students may select any problem and approach that meets their interests, affords opportunities for truly impactful results, and is feasible within the timeframe of this course. Students may select any open data sets (publicly available, unrestricted) that suit their focal problem/analysis, and may use any software or tools that are appropriate for meeting their goals. Students must present and defend their choices across these dimensions throughout the course.

Students are required to document their analyses thoroughly. They are also strongly encouraged to build an end-to-end analysis pipeline covering data sourcing, cleaning/preparation, transformation, processing, visualization, interpretation, etc.

Readings and Course Pack

There is no textbook for this course. Readings are drawn from various relevant books, articles, reports, and academic papers and are made available either in the course pack or online.

Syllabus

Week 1: Intro to Capstone

Topics

- Extended intros of instructors and students
- Preview of course structure
- Expectations around presentations, workshops, deliverables, grading policies
- Project idea brainstorming

Required Reading

- Review prior capstone projects on Berkeley website
- Review asynch for week 1

Week 2: Idea Generation

Topics

- Students present ideas for discussion
- Past capstone projects
- Techniques for idea generation

Required Reading

- Davenport and Kim, Keeping Up with the Quants, 2013, HBR Press Ch. 2, “Framing the Problem”

Assignment due

- Prepare 3 potential project ideas, post on ISVC Course Wall.
- Prepare to present one project idea in class (3-5 minutes).

- Write up project idea (presented in class) as Project Proposal, incorporating feedback from previous class, using project template, post for voting (by Friday of week 2).

Week 3: Product Principles

Topics

- Opportunity assessment and Customer segmentation
- Gathering user requirements
- Minimum viable products in data science

Required Reading

- Kagan, M. (2018). In *Inspired: How to create products customers love*. Sunnyvale, CA: SVPG Press. The full book is highly recommended and skimmable, but if you are short on time focus on the Introduction and Chapters 3-5, 11-14, 19, and 21-22

Assignment due

- Select project proposals by each student per voting process (by Monday of Week 3).
- Review asynch for week 3.

Week 4: Ethical Considerations

Topics

- Ethical and legal considerations in data science
- Discussion of ethical risks of AI and data science using examples from the news (bias, equity, effects on the workplace…)
- Identifying and managing ethical and legal risks in projects

Required Reading

- E. Ntoutsi, et. al., "Bias in data-driven artificial intelligence systems—An introductory survey," Wiley Online Library, 03 February 2020, https://onlinelibrary.wiley.com/doi/full/10.1002/widm.1356
- Kahneman, D. (2011). Chapters 1-3, 21, and 34. In *Thinking, Fast and Slow* (2nd ed.). New York

Assignment due

- Due by class: completed Team Process Agreement and Team Project Plan using template provided by instructors.
- Review asynch for week 4.

Week 5: Group Presentation 1

Topics

- Deliver 1st presentation

Assignment due

- Prepare Presentation 1 focusing on why problem was selected, approach, potential challenges

Week 6: Selecting and Exploring Data Sources

Topics

- Selecting data sources
- Walk through example of exploratory data analysis
- Feature selection and engineering

Required Reading

- Mawer, et. al "The Value of Exploratory Data Analysis," March 2017, https://svds.com/value-exploratory-data-analysis/
- Review asynch for week 6

Week 7: Privacy and Digital Identity

Topics

- Importance of privacy and digital identity
- Privacy regulations (U.S. and global)
- Technical approaches to preserving privacy
- Existing and emerging data-driven products in data privacy and digital identity

Required Reading

- Dwork, Cynthia, Nitin Kohli, and Deirdre Mulligan. 2019. "Differential Privacy in Practice: Expose Your Epsilons!". Journal of Privacy and Confidentiality 9 (2) Background: https://www.ischool.berkeley.edu/research/publications/2019/differential-privacy-practice-expose-your-epsilons
- Restoring Trust in a Digital World, https://www.mastercard.us/content/dam/mccom/en-us/issuers/digital-identity/digital-identity-restoring-trust-in-a-digital-world-final-share-corrected.pdf

Week 8: Storytelling with Data and Influencing as a Data Scientist

Topics

- How people and organizations process information and make decisions
- Challenges to building a data-driven organization, along with potential solutions
- Tactical guidance for effective influencing and consensus building

Required Reading

- Cole Nussbaumer, "Storytelling with Data," https://www.youtube.com/watch?v=4g2g1vS157Q
- Patil, D. J., & Mason, H. (2015). *Data driven*. Available from http://www.oreilly.com/data/free/data-driven.csp

Assignment due

- Students to prepare elevator pitch about their projects and about themselves as data scientists
- Team peer evaluation #1

Week 9: Technical Model Evaluation

Topics

- Data sets
- Experimental framework
- Metrics
- Effectiveness vs. efficiency

Required Reading

- Srivastava, T. "11 important model evaluation metrics for machine learning everyone should know", https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

Week 10: Group Presentation 2

Topics

- Deliver 2nd presentation

Assignment due

- Prepare presentation 2 focusing on technical approach and preliminary result

Week 11: From Prototype to Product, and Beyond

Topics

- Technical debt in data science/machine learning products
- Scalability in data science organizations
- Deploying your data science product

Required Reading

- D. Sculley, et. al., "Hidden Technical Debt in Machine Learning Systems," 2015, https://papers.neurips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf
- Robinson, J. "How Facebook Scales Machine Learning," Feb 3, 2019, https://medium.com/@jamal.robinson/how-facebook-scales-artificial-intelligence-machine-learning-693706ae296f
- (Optional) "The Ultimate Guide to Deploying Machine Learning Systems," May 30, 2020, https://mlinproduction.com/deploying-machine-learning-models/

Week 12: Possible Futures of Data Science and Wrap-Up

Topics

- Recent data science–related trends and possible futures
- Perspectives on the future of data science in terms of technology, application areas, and the profession
- Recap and review
- Goodbye and parting thoughts

Assignment due

- Students to summarize recent data science developments for discussions

Week 13: Dry runs for final presentation

Topics

- Dry runs of final presentation
- Team group time
- Course evaluations

Weeks 14: Final Group Presentations

Topics

- Deliver third presentation.

Assignment due

- Prepare third and final presentation, show demo, web deliverable, and present lessons learned
- Submit final deliverable, including presentation slides, website delivery and project summary. Upload instructions for the project summary be found here: https://www.ischool.berkeley.edu/intranet/students/mids/capstone.
- Team peer evaluation #2