Info 98: Practical Data Science Skills for Internships

Fall 2018 Syllabus

General Information:

Student Team:

- *Student Facilitators*: Jerry Lin
- *Lecturers*: Peter Butler, Alex Nakagawa, Leo Li, Samyak Parajuli, Patrick Chao, Emma Russon, Junseo Park, Jerry Lin and other selected lecturers from DSS and SUSA
- *Teaching Assistants*: Lily Bhattacharjee, Dhruv Jhamb, Gary Xiang, Shichao Han, Muskaan Goyal, Calvin Chen, Tina Ye

Instructor of Record: Professor Paul Laskowski

Lectures: Monday 6:30 PM - 8:00 PM, Barrows 20

Prerequisites: Ability to code in a general-purpose or data-oriented programming language

Textbook: An Introduction to Statistical Learning by James, Witten, Hastie and Tibshirani (physical version not required)

Units: 2

Grading:

- 2 units, P/NP only. 100+ /144 points is required to pass the course.
- There will be four required projects that count for 100 points total.
- These four required assignments are worth 10, 20, 30, and 40 points respectively.
- Reading responses are worth 4 extra credit points each, with 3 in total.
- Attendance counts for roughly 31% of the overall grade.

Permission Numbers: Distributed privately.

Course Piazza: https://piazza.com/berkeley/fall2018/info98/home

Attendance Policy:

- Attendance will be measured via lecture codes submitted as bCourses assignments. Each lecture counts for 4 points.

Late Submission Policy:

- 10% of the total assignment grade will be deducted for every day it is turned in late.

Academic Honesty:

- Plagiarism will not be tolerated. Any evidence of cheating will result in immediate failure of the course.

Course Overview:

Welcome to Data Science Society at Berkeley's very own DeCal: Practical Data Science Skills for Internships! This course is geared towards exposing you to essential Data Science and Statistical Analysis topics that will challenge you, sharpen your skills, and elevate you in the internship game.

In this course, you will learn everything you need to know from the ground up—from an

introduction on how to proficiently use Git/Github, to building essential collaboration skills through programming projects, querying databases with SQL, and finally, Statistical Analysis.

This class is a project-based, Pass/No-Pass 2-unit course with an emphasis on concepts that require minimal background but amount to significant pain points for those learning data science at UC Berkeley. It will consist of 3-4 projects, some collaborative, that will help you build up your portfolio. The course is graded on satisfactory completion of these projects. There will be a few readings, but writing assignments are for extra credit.

This term we will be using **Piazza** for class discussion. The system is highly catered to getting you help fast and efficiently from your classmates, the TAs, and facilitators. Rather than emailing questions to the teaching staff, we encourage you to post your questions on Piazza.
Find our class page at: https://piazza.com/berkeley/fall2018/info98/home

We encourage all interested students of all majors to enroll in this course. The student facilitators, lecturers, and TAs intend to provide you with a comprehensive introduction to data science, with the goal of preparing you for industry. We therefore encourage students who are genuinely enthusiastic and interested in what we have to offer to enroll. Please email datasciencedecal@gmail.com with any questions or concerns, and we will be happy to get back to you!

Schedule:

PART I: COLLABORATION

Lecture	*Date*	*Title*	*Lecturer*
1	9/10	Introduction of group project assignment + Making beautiful documents using LaTeX	Jerry Lin
2	9/17	Using Github	Patrick Chao
3	9/24	Workshop for Git	TA team

Readings, Assignments, and Projects

- LaTeX tutorial by David Xiao (~1 hour to complete, *optional*)
- Interactive Git/Github Tutorial by Code School (~1 hour to complete, *optional*)
- **Sample Resume using LaTeX Assignment** (**10 points**, ~5 hours to complete) ***Due:*** 9/16/18 11:59 PM
- **Design your own DeCal syllabus Group Project** (**20 points**, ~15 hours to complete) ***Due:*** 9/30/18 11:59 PM

PART II: QUERYING DATABASES

Lecture	*Date*	*Title*	*Lecturer*
4	10/1	Intro to PostgreSQL: Basic Commands + Joining	Alex Nakagawa
5	10/8	Filtering, Aggregation, and Voting Project	Alex Nakagawa

Readings, Assignments, and Projects

- SQL Language Tutorial (~5 hours to complete, *optional*)
- PostgreSQL: Up and Running by Regina Obe and Leo Hsu (~8 hours to complete)
- **Instant Runoff Voting Simulation Assignment** (**30 points**, ~15 hours to complete) ***Due:*** 10/14/18 11:59 PM

PART III: MACHINE LEARNING

Lecture	*Date*	*Title*	*Lecturer*
6	10/15	Regular Expressions / Data Cleaning	Leo Li
7	10/22	Linear Regression pt. 1	Peter Butler
8	10/29	Linear Regression pt. 2	Peter Butler
9	11/5	Logistic Regression	Jerry Lin
10	11/19	Bias & Inferential Thinking	Jerry Lin
11	11/26	Cross Validation, Regularization, and Metrics for Performance	Patrick Chao

Readings, Assignments, and Projects

- **Machine Learning Project** (**40 points**, ~25 hours to complete) ***Due:*** 12/1/18 11:59 PM

Project Rubrics:

Plagiarism will not be tolerated. Any evidence of cheating will result in immediate failure of the course.

Late submission policy:

10% of the total assignment grade will be deducted for every day it is turned in late.

LaTeX Resume (out of 10 points)

Good use of formatting, very professional looking	10 - 11 points (extra credit will be given for exemplary submissions)
Good use of formatting, could use a few touches	8 - 9 points
Mediocre use of formatting, obvious lack of effort	5 - 7 points
Submitted something, vaguely resembles resume	1 - 4 points

Design Your Own Decal Syllabus (out of 20 points)

As this is a group project, 70% of your grade will come from the overall quality of the final product. 30% of the grade will come from proof of collaboration (as seen from your github).

Good use of formatting, there is an actual attempt at designing your own syllabus, and the project contains at least 3 branches	17 - 25 points (extra credit will be given for exemplary submissions)
Good use of formatting, there is a vague attempt at designing the syllabus, and/or the project contains fewer than 3 branches	14 - 16 points
Mediocre use of formatting, there is almost no attempt at designing the syllabus and/or there are no more branches in the project	10 - 13 points
Submitted something, vaguely resembles a syllabus	1 - 9 points

Election Assignment (out of 30 points)

Both the winner and margin of error for all types are calculated correctly, code is well put together	28 - 31 points (extra credit will be given for exemplary submissions)
Both the winner and margin of error for all types are calculated correctly, code is not well put together	24 - 27 points
Winner and/or margin of error not calculated correctly, but the code is a decent attempt at calculating the right answer	18 - 23 points
Submitted something, obvious lack of effort	1 - 17 points

Machine Learning Project (out of 40 points)

Outstanding use of techniques taught in class, code is well put together	37 - 45 points (extra credit will be given for exemplary submissions)
Code demonstrates solid understanding of techniques taught in class, code is well put together	30 - 36 points
Code demonstrates limited understanding of the techniques taught in class, code is not well put together	24 - 29 points
Code is a mediocre attempt at executing the techniques taught in class, code is not well put together	17 - 23 points
Submitted something, obvious lack of effort	1 - 16 points

Supervision & Responsibility of Instructor of Record

1. The Instructor of Record for Info 98 - Practical Data Science Skills for Internships will be Prof. Paul Laskowski.
2. The Lecturers will share instructional materials with Instructor of Record, via email/google drive, at least two weeks before the instruction, and expect to get feedback, via email/comments on materials, in a week.
   The Lecturers will then have ~5 days to revise and rehearse for new materials.
3. The Lecturers will schedule in-person meetings with Instructor of Record, at least a week before instruction begins, to seek advice and/or explanation on confusing/uncertain concepts.
4. Our Instructor of Record will also help with creating bCourse site, if necessary. If eventually the Lecturers decide to use Git/GitHub purely, Instructor of Record will be invited to join the repository.
5. The Instructor of Record also holds the responsibility for supervising the awarding of all final grades and for reporting the grades to the Registrar. The Lecturers choose to use the same end-of-semester course evaluation form as that being used by Instructor of Record for his other course(s). The Lecturers may or may not create another evaluation form for the team to improve this DeCal course, in forms of Google Form. If Lecturers would like to do so, they will communicate with Instructor of Record two weeks ahead.
6. Our Instructor of Record is welcomed to do in-class observations, and to give any help/suggestion during break.
7. Our Instructor of Record agrees to help with any other unforeseeable issues if needed, unless he sees his involvements being unappropriated.