Syllabus MAS405 Spring 2021

Contents

```
## Last run: 2021-03-22
```

Data Management

Dave Zes, Slichter 3877, dave.zes@stat.ucla.edu

Office Hours: 6:00-6:45pm, Thursdays; 5:00-5:45pm, Fridays; and by appointment

Most Importantly

Everything is subject to change — all natural phenomenon is a function of time. Please make a habit of checking CCLE regularly.

I will continue to update this syllabus over the next 2 weeks.

General/Goals

I feel I can best convey our objectives by making clear what we are going to avoid. I DO NOT want this to be a peppering of tactical, esoteric operations in a cloistered, isolated, or boxed-in environment. The inevitable result of which is that a few weeks after the quarter ends you start forgetting what you did, the environment ceases to be maintained, and, ultimately, when someone asks you to connect to a DB, or a friend asks for help with their website, you are uncertain, and ultimately have to "start over".

Not cool.

In our time together, you will design and deploy numerous databases with different data sets, and perform analytics with these databases. You will access my databases and those of your classmates, perform analytics, present results. Moreover, you will have concise example code, and the database resources you create are yours for as long as you wish to keep them.

Having practical facility with relational databases is one of THE MOST EMPOWERING skills anyone can develop.

Our maxim here is **simplicity**. We can do a lot with a little — much better than little reward after lots of work.

This is not a "coding" course, however we're going to be enjoying a great deal of coding. After all, computer instructions — AKA "code" — is really just a way to get a machine to do what we want.

Much of our time together will be "follow the leader" (not to be confused with so-called "monkey see monkey do"). I.e., while this is like a "workshop" — it is critical that it be clear that while we will tactically execute computational operations, we have a lucid understanding of our motivation for each step.

Objectives

- Data management
- Relational databases
- AWS RDS
- SQL — namely MySQL
- Data sharing and basics of data administration
- Object structures XML and JSON
- XPath queries
- Web scraping
- Serialization and object digests
- Static analytics reporting
- Real-time web-based analytics
- Database connectivity for professional reports/articles/analytics
- Crafting compelling stories from data

Challenges

Data enablement involves a constantly evolving platform, security, and privacy landscape.

No guard rails at first. We are going to start by jumping right into data enablement. This will allow us to quickly attain some really important milestones. To achieve these early milestones, we are not going to be thinking about adhering to "production" standards – doing so would be too great a millstone. Needless to say, we will not be dealing with PII or personal data.

Technology

Will will be using Zoom for our remote sessions. As a UCLA student, you have access to this resource under UCLA's license without charge.

We will mostly be using R.

If you have **no** familiarity with R, our TA or myself will hold a workshop.

The resources below we will learn as we go.

We will use some PHP for web based data handling. There will likely be some Python and SageMaker. And, of course, lots of SQL using MySQL.

MySQL Workbench is a nice GUI app that's free and allows simplification for some operations on our DBs (e.g., granting permissions to other users).

We will be using AWS for our implementations. We will cover enablement of this aspect in class.

We're going to lean heavily on AWS (Amazon Web Services) for our work. This will provide a uniform "platform" for our adventures. AWS is the real deal; all your work will be uniquely yours, and you will have total dominion over it long after we've finished the quarter.

I will be providing a codebase with simple examples.

Others TBD

Last comment for this section:

Many students are becoming too comfortable using the R app and/or RStudio. The R Application is nice as many people like the point-and-click GUI. RStudio is really nice for drafting reports and articles (like this syllabus was rendered using RStudio). However, for heavy computational work, I want to **strongly** encourage you to run R commandline via Terminal.

Textbooks

Required

- Data Management for Researchers by Kristin Briney – PELAGIC.

Totally optional

- Learning MySQL by Seyed M.M. "Saied" Tahaghogbi & Hugh E. Williams — O-REILLY

My comment on this text: This is truly big league. But, I think it's serious overkill for our purposes.

- Database Concepts, 7th Ed. by David M. Kroenke and David J. Auer — Pearson.

My comment on this text: My first reaction is that it must be good if it's authored by two people named David. This is really a nice book and it's inexpensive used. But they spend quite a bit of time working with high-level software, e.g., Microsoft SQL Server, MS Access.

A number of insightful and informative articles will be provided.

Campuswire

We will be using Campuswire for class-wide communication and collaboration. The system is catered to getting you help fast and efficiently from classmates, the TA, and myself. Please email any non-personal questions you might have about the course or the content to Campuswire rather than to me or the TA. I will only answers questions via Campuswire (Personal questions are exceptions to this policy. For example, if you want to make an appointment to see me, if have questions about your grade, etc. please contact me or the TA directly).

Invitations to Campuswire will be sent out during Week 1.

Website

Our class business will be conducted through the CCLE course management system: http://ccle.ucla.edu. The first time you log in there will be special instructions to get you started. I recommend you log in immediately. All homework and announcements will be made through this site, so you should check it daily. Grades will be posted on my.ucla.edu.

DO NOT rely on the "course total" grade calculation provided on CCLE. It is not likely something I will be managing over the course of the quarter.

Collaboration

First, collaboration is built into the curriculum for this course. You'll be accessing my/classmates published data and me/classmates will be accessing the data you publish.

Moreover, collaboration is hugely important for this course — and I'm not just saying that because emphasizing collaboration is fashionable. If you are struggling with something, your classmates and I need to know; or, if you know a better way of doing something (e.g., a more efficient SQL query, or a better web graphic), we're counting on you to share it . . .

. . . **However**, online quizzes are to be completely individually.

Weekly Assignments

We are going to attempt to complete the (more or less weekly) Assignments in Lecture. Anything we cannot complete in Lecture you will complete either on your own, or with classmates, or possibly during office hours.

No-Excuse Policy

Assignments must be submitted on time in readable format. The lowest assignment score will be dropped. The purpose of this is not to give you a chance to ignore an Assignment, but is there so that if something beyond your control prevents you from turning in the homework, you will not be penalized.

Final Project

Full data "pipeline" – your DB to final PDF executive report

Grading

Your final grade will be determined by

- Graded Assignments: 60%

- Final Project: 40%

- I always reserve the right to add points to students who have excellent engagement, show good citizenship, or any of the good stuff that makes graduate-level learning rewarding and enjoyable.

- Quizzes may be a surprise paper, or online quiz during Lecture.

Classroom Environment

In this class, I want each of you to feel welcome to contribute to class discussions. Doing so requires a thoughtful environment in which each of us participates both as listener and speaker. By attending class, you are agreeing to listen with respect to the speaker, and choose your words so that they reflect respect for your classmates. As listeners, we agree to give the speaker the benefit of the doubt by assuming that their words are not meant to cause insult.

Please feel free to send emails about your concerns any time and visit during office hours. If you need help badly, please email me to set up an appointment. Remember, questions about content should be sent to the Campuswire site.

Academic Integrity

As a student and member of the University community, you are here to get an education and are, therefore, expected to demonstrate integrity in your academic endeavors. All students must uphold University of California Standards of Student Conduct as administered by the Office of the Dean of Students. Students are subject to disciplinary action for several types of misconduct, including but not limited to: cheating, multiple submissions, plagiarism, prohibited collaboration, facilitating academic dishonesty, or knowingly furnishing false information. You may have assignments or projects in which you work with a partner or with a group. For example, you are welcome, and even encouraged, to work with others to solve homework problems. Even though you are working together, the assignment you submit for a grade must be IN YOUR OWN WORDS, unless you receive specific instructions to the contrary. For more information about academic integrity, please go to www.deanofstudents.ucla.edu.

Collaboration Policy for Assignments   For Assignments, read and adhere to Level 1: https://turing.bowdoin.edu/dept/collab.php

Schedule

% latex table generated in R 4.0.3 by xtable 1.8-4 package % Mon Mar 22 18:50:26 2021

Week 1	2021-03-29	Monday	
	2021-03-30	Tuesday	
	2021-03-31	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-04-01	Thursday	
	2021-04-02	Friday	
	2021-04-03		
	2021-04-04		
Week 2	2021-04-05	Monday	
	2021-04-06	Tuesday	
	2021-04-07	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-04-08	Thursday	
	2021-04-09	Friday	
	2021-04-10		
	2021-04-11		
Week 3	2021-04-12	Monday	
	2021-04-13	Tuesday	
	2021-04-14	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-04-15	Thursday	
	2021-04-16	Friday	
	2021-04-17		
	2021-04-18		
Week 4	2021-04-19	Monday	
	2021-04-20	Tuesday	
	2021-04-21	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-04-22	Thursday	
	2021-04-23	Friday	
	2021-04-24		
	2021-04-25		
Week 5	2021-04-26	Monday	
	2021-04-27	Tuesday	
	2021-04-28	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-04-29	Thursday	
	2021-04-30	Friday	
	2021-05-01		
	2021-05-02		
Week 6	2021-05-03	Monday	
	2021-05-04	Tuesday	
	2021-05-05	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-05-06	Thursday	
	2021-05-07	Friday	
	2021-05-08		
	2021-05-09		
Week 7	2021-05-10	Monday	
	2021-05-11	Tuesday	
	2021-05-12	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)

	2021-05-13	Thursday	
	2021-05-14	Friday	
	2021-05-15		
	2021-05-16		
Week 8	2021-05-17	Monday	
	2021-05-18	Tuesday	
	2021-05-19	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-05-20	Thursday	
	2021-05-21	Friday	
	2021-05-22		
	2021-05-23		
Week 9	2021-05-24	Monday	
	2021-05-25	Tuesday	
	2021-05-26	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-05-27	Thursday	
	2021-05-28	Friday	
	2021-05-29		
	2021-05-30		
Week 10	2021-05-31	Monday	Memorial Day Holiday
	2021-06-01	Tuesday	
	2021-06-02	Wednesday	Discussion (5:00pm-5:50pm) ; Lecture (6:00pm-8:50pm)
	2021-06-03	Thursday	
	2021-06-04	Friday	
Final Project Presentations	TBD		