UNIVERSITY OF CALIFORNIA, DAVIS

GRADUATE SCHOOL OF MANAGEMENT

DATA DESIGN & REPRESENTATION

(BAX-422, WINTER 2024)

PRELIMINARY SYLLABUS (12/14/2023)

Instructor: Jörn Boehnke, Assistant Professor

Office: 3404, Gallagher Hall, Graduate School of Management

E-Mail: jb@ucdavis.edu

First Class: **1/12**

Office Hours: Every Monday 7:15am – 8:15am at
https://ucdavis.zoom.us/j/96665507050?pwd=ckhydkJPYmw0UlZFSmtCbUNVd2FsZz09

The first office hour is 1/8.

Please feel free to reach out to me via email at any point. We can also schedule additional meetings if necessary.

Objectives: After taking this course, you will be able to automatically extract information from most unstructured sources, process it, and structure it for subsequent business use. This includes knowledge in the following topics:

Regular Expressions (RegEx)
Web-Scraping
Web Data Processing
JavaScript Object Notation (JSON)
MongoDB (NoSQL)

TA: Shreya Gundu   (sgundu@ucdavis.edu)

Shreya offers weekly office hours and occasional seminars. Details are TBD.

Please also feel free to reach out to your TA via email at any point.

Evaluation: 15% Assignments

  5% Mini exam week 1 (Monday, 1/8 5pm)

20% Individual project week 5

20% Individual project week 9

  5% Mini exam week 11 (Monday, 3/18 5pm)

15% Final group project

20% Class participation

Late assignments, exams, and projects will not be accepted.

Clerical scoring errors will be corrected without hassle, but for other re-grades you must hand back the work and submit an email request; the entire paper will be subject to re-grading.

Groups: You will be assigned to a group at the beginning of the quarter to discuss homework assignments and work on the final project. *Please note that group work will be subject to a peer assessment at the end. One-half of your total group work grade (assignments and final project) will be determined by the score on the peer evaluations.* I.e., individuals who do not contribute their fair share to the group (as determined by the group) will be penalized.

Class Participation: I hope you actively participate in class. Meaningful class interactions add greatly to the learning experience, and I consider class participation to be an important component of this course and its success. Accordingly, active and meaningful participation in class and on Piazza is part of your grade evaluation. Please be prepared to be cold-called in class.

Assignments: There will be 8 one-week homework assignments. Homework assignments are released after class. *Only the best 7 assignments will be counted towards your grade.* Assignments are submitted through Canvas and should have a clear and concise presentation.

All assignments are individual assignments. Assignments *may be* discussed among your group members but *must be* submitted individually. While your code submission may be similar to that of your group's members, any write-up, code comments, etc. must be your own. Simply copying and pasting the write-up and code comments from other members in your group will yield point deductions. I want each group member to fully understand the code and analytics required in assignments.

Assignments are not to be discussed outside your group. Any violation of this policy will be considered plagiarism and penalized as such.

Classroom Etiquette / Strikes / Extra Assignments: Classes begin on time. You are required to attend every class unless extenuating circumstances arise and make it difficult to do so. If you are unable to attend a class, you are required to inform me about the extenuating situation by email *at least one hour* before the start of class. A late arrival ($\geqslant$ 5 minutes) or skipped class results in *one strike*.

All cell phones must be muted before the start of class. The computer should only be used to take class notes and code / run in-class exercises. All other programs should be shut down before the start of class. Any messaging, web surfing, e-mailing, etc. disturbs the class and is a breach of classroom etiquette. If it comes to my attention, any such behavior results in *one strike*.

Strikes are accumulated throughout this course. Ideally, your strike balance remains zero. However, I recognize that the unexpected sometimes happens, and a total of two strikes or less will remain without consequences. Any strike above two yields an additional individual assignment. I.e., the number of (counted) assignment goes up by one for each strike in excess of two.

Projects: Individual projects will be posted at the beginning of week 5 and week 9 and are due 5 days later.

Individual projects are *not* to be discussed with anyone (including your group members). You are required to *solve them by yourself*. Any violation of this policy will be considered plagiarism and penalized as such.

Mini Exam: There will be two mini exams. The exams will be conducted on Monday 1/8 and 3/18 at 5pm Pacific Time using Respondus LockDown Browser + Monitor. All exams will be closed book exams.

Mini exam 1 will cover the core concepts learned in Fall Quarter's BAX-401 and BAX-441. Mini exam 2 tests your programming and data structure knowledge and is meant to prepare you for coding interviews that you will encounter during job search.

Final: For the final project you will prepare data for a business problem of your choice. The goal of the project is to source data and structure it for business use. Specifically, your data and data design choices ought to add value to the business problem at hand. For this project, think of yourselves as consultants who want to support some company's business decisions. These decisions should be based on data and it is your task to collect and prepare these data for use.

Planning should begin around week 6. Successful projects will require web-scraping, data wrangling, databasing, and managerial skills. They will implement data from a variety of sources (some of them unstructured, primary online sources that require web-scraping), and connect these data into one dataset applying the concepts we covered in class. All data design decisions must be documented, and all code must be attached to your submission. Moreover, this project must be significantly different to any final

project developed for other MSBA classes.

This project is by no means limited to data sourced from primary web content. Please feel encouraged to also utilize secondary sources such as Kaggle, the Census (IPUMS), etc. and other primary sources (e.g., data from someone you know at a company). Please always adhere to the terms and conditions pertaining to the websites and data sources used.

Typically, project reports will involve the following components: (1) Title Page, (2) Executive Summary, (3) Background, Context, and Domain Knowledge: business (scenario) in mind, type of industry, products / services, (4) Introduction of the Data Sources, Description of the Web-Scraping Routine(s), and Explanation of the Dataset / Database Design Choices, (5) Discussion of how the dataset will help answer business relevant questions and allow an efficient use of the data. Please highlight the advantages of the chosen database implementation (over alternative ones), explain the design choices, and lay out the business value created by it, and finally, (6) Summary and Conclusions.
(4) and (5) will showcase what you learned in class and how well you can apply it. Please make sure that these components form the centerpiece of your project report.

The final project report cannot exceed 10 pages (not including the title page, code attachments, references, and appendices). Any necessary tables, figures, visualizations, and text must be contained within the 10 pages. I urge you to ensure that the written report is direct, insightful, and specific to the problem at hand. The report should adhere by the following formatting guidelines: text to no smaller than 11-point font, 1-inch margins on all pages, and all text should be double-spaced. The report should also contain an executive summary, which counts towards the page limit. **All submissions are due on Thursday 11:59pm Pacific of finals week (i.e., Thursday of week 11).** No exceptions.

Software: Please feel free to use either Python *or* Java for the programming assignments and projects in this course. Alternatively, you can use R, but please be aware that R is a *statistical* computing language that has not been developed with these use-cases in mind.

Each topic of the course will require its own set of software. I will live-code a good amount while we move along in class. While I do not expect you to "code-along," I strongly encourage you to review all the code we develop in class and run it on your own machine. The software environment I will used in class contains:

Python + VS Code
or
Java + Eclipse

RegEx compatible text editor (Notepad++, BBEdit, or Sublime Text are good)

Browser with integrated development tools (either Chrome or Firefox are good)

MySQL *or* MariaDB

MySQL Workbench *or* phpMyAdmin

MongoDB

I strongly recommend to replicate this setup to easily execute code developed in class.

Rules: All participants in the course, instructor and students, are expected to follow the UC [Davis Principles of Community](#), which includes affirmation of the right of [freedom of expression](#), and rejection of discrimination. The right to express points-of-view without fear of retaliation or censorship is a cornerstone of academic freedom. A diversity of opinions with respectful disagreement and informed debate enriches learning. However, in this course, any expression or disagreement should adhere to the obligations we have toward each other to build and maintain a climate of mutual respect and caring.

You are expected to take UC Davis's [Code of Academic Conduct](#) as seriously as we do. You were given this code of conduct with explicit explanations of violations (e.g. plagiarism, cheating, unauthorized collaboration, etc.) and your responsibilities in regard to them during orientation, and you signed a statement affirming that you understand it. Academic conduct violations will not be tolerated, and your instructor will not hesitate to turn violators over to Student Judicial Affairs. If you are uncertain about what constitutes an academic conduct violation, please refer to the code linked above, contact your instructor, or refer to the [Office of Student Judicial Affairs](#).

Q&A: We are using Piazza for class discussion (accessible through Canvas). The system is highly catered to getting you help fast and efficiently from classmates, the TAs, and myself. Rather than emailing your questions, try to post on Piazza. If you have any problems, you can email me or team@piazza.com.

Please feel encouraged to answer your classmate's questions; it's a huge help to us, and even if you are wrong everyone learns (we check the answers and clear up confusion). While you can post anonymously, we encourage you to take credit for your questions and answers.

Texts: There is no required textbook: all materials will be available on the class website. The best preparation you can do before lectures is to go through class code and work through examples.

This course is not a programming course. That said, a basic understanding of Java will be helpful to easily follow the code developed in class. *Head First Java: A Brain-Friendly Guide* by Kathy Sierra and Bert Bates is a friendly introduction to Java. If you already know a lot of Java and want to develop a deeper understanding for it, *Effective Java* by Joshua Block (he wrote a large amount of core Java functionality) is one of my favorites.

This course touches on a broad variety of subject and no one book covers all of them. *Mastering Regular Expressions* by Jeffrey Friedl is a solid book on Regular Expressions. Ryan Mitchell's *Web Scraping with Python: Collecting More Data from the Modern Web* is a good companion for web-scraping in Python. Unfortunately, I have not yet found a good book for web-scraping in Java. *Case*Method: Entity Relationship Modelling* by Richard Barker as well as *Data Modeling Essentials* by Simsion Witt are detailed introductions to Entity Relationship Modeling. *NoSQL with MongoDB in 24 Hours* by Brad Dayley uses Java to introduce the reader to MongoDB.

Accommodation: UC Davis is committed to educational equity in the academic setting, and in serving a diverse student body. All students who are interested in learning about how disabilities are accommodated can visit the [Student Disability Center](#) (SDC). If you are a student who requires academic accommodations, please contact the SDC directly at sdc@ucdavis.edu or 530-752-3184. If you receive an SDC Letter of Accommodation, submit it to your instructor for each course as soon as possible, at least within the first two weeks of a course.

Safety and Emergency Preparedness: UC Davis has many resources to help in case of emergency or crisis. While reviewing campus [Emergency Information](#), you may want to register for UC Davis Warn Me and Aggie Alert, which will give you timely information and instructions about emergencies and situations on campus that affect your safety.

If there is an emergency in the classroom or in non-Davis locations, follow the instructions of your instructor.

Disclaimer: Unexpected events might require elements of this syllabus to change. Your instructor will keep you informed of any changes.

All material in the course that is not otherwise subject to copyright is the copyright of the course instructor and should be considered the instructor's intellectual property.