UNIVERSITY OF CALIFORNIA, DAVIS

GRADUATE SCHOOL OF MANAGEMENT

DATA DESIGN & REPRESENTATION

(BAX-422, WINTER 2020)

PRELIMINARY SYLLABUS

Instructor: Jörn Boehnke, Assistant Professor

Office: 3404, Gallagher Hall, Graduate School of Management

E-Mail: jb@ucdavis.edu

Office Hours: Anytime needed. Really. Just reach out via email and we will make it work. In addition, I will arrange to have at least four in-person office hours in SF since all of you are based here.

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
Entity Relationship Modeling for SQL
JavaScript Object Notation (JSON)
MongoDB (NoSQL)

TAs: Chitrabhanu Gupta (cbgupta@ucdavis.edu)
Ehsan Gholami (egholami@ucdavis.edu)
Justin Reafsnyder (jareafsnyder@ucdavis.edu)

Evaluation: 30% Assignments
30% Individual midterm project
30% Final group project
10% Class participation

Late assignments, exams, or 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 are encouraged to form groups (max. size 4) for the homework assignments and final project. The group work will be subject to peer assessment of each member's contribution. One-third of your total group work grade (assignments and final project) will be determined by your 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.

Assignments: There will be 8 one-week homework assignments. Homework assignments are released during 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.

Midterm: The individual midterm project will be posted at the beginning of week 5 and is due the subsequent week.

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 right after the midterm. Successful projects will require programming, 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 Java *or* Python 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:

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: **Academic Honor Code**

All students pledge to adhere to the University of California, Davis' Code of Conduct for all work, cf. http://sja.ucdavis.edu/files/cac.pdf.

Electronics in Class

You are allowed to use your laptop / tablet in class for activity pertaining to the class. You cannot use your laptops for anything else. **If I see you doing anything else** (e.g. Email, Facebook, YouTube, assignments for other classes etc.), **you will be required to hand in your next homework assignment individually using Java.** Moreover, I will personally test run and grade it.

Smartphone / phone use is not allowed in class. If your phone rings or beeps, I may insist that I get to answer it and I will scare the caller away for good! With my permission, you are welcome to step out of class in order to take an important call.

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.