**Project Goals:**

Now that you have the necessary background to carry out the project, let's get started!

There will be no more lectures, reading quizzes, concept questions or clickers. Starting next class, you will work on a 6-week project during class time. The project will be done in teams of 2 or 3 students. Larger teams won't be allowed and you need special permission to work by yourself.

The goal for the project is to build a data pipeline using a relational database. A data pipeline is an automated process of collecting, transforming, storing, and querying data. The pipeline will have a web scraper that extracts data from a set of web pages. The extracted web data will be reformatted, cleansed, and stored into a relational database. The transformed data is queried from the database through a simple command-line interface.

The data pipeline will be developed in Python and SQL. The database will be constructed in MySQL. If you have not written a web scraper before, you should carefully read the assigned chapters from our textbook "Web Scraping with Python" to learn how to extract data from various kinds of web pages. The book also has a chapter on storing data into a MySQL database from Python. Code samples from the book are available on GitHub. Feel free to use the samples as a starting point to get your job done.

You will form teams and propose a project. Once your proposal is approved, you will design the data model, create the database, and develop the data pipeline for your project. You will retrieve the queryable data using a simple command-line interface. Once all of these requirements are met, you may optionally choose to develop a web interface for extra credit.

**Tools:**

You must utilize MySQL server 5.7 for the database management system. You may also use MySQL Workbench 6.3 to graphically interact with the database.

You must utilize SQL and Python 3.4 for developing the data pipeline. You must use Beautiful Soup 4 for traversing and parsing the web data.

You must utilize PyMySQL 0.7.2 for connecting to and interacting with MySQL from Python.

You must also utilize GitHub’s version control to manage your source code. This allows us to see exactly who contributed what portions of code and is an easy way to see how your code is progressing.

Finally, you may also use LucidChart for diagramming your database schema.

**Logistics:**

Starting next class, we will be meeting in GDC instead of CLA. On Mondays, we will meet in the Faculty Lounge at **GDC 6.302**. On Wednesdays, we will meet in **GDC 4.304** and in the adjacent

bridge, **GDC 4.100**. These rooms have tables and chairs so you can sit with your team and work together on the project.

Your team will need a laptop and you should bring one to the class meetings. In fact, there should be a minimum of one laptop per team. If your team doesn't have a working laptop, you should come talk to us asap.

**Progress Reports:**

In lieu of concept questions, you will submit a progress report at the end of each class. The progress report should list which team members attended class, what you accomplished during the class meeting, what you hope to accomplish during the next class meeting (or outside of class), and any issues you are facing. You should only submit one progress report per team. The progress reports will be done via a Google Form which will be shared during class time. If you do not submit this report by the end of class, your team will not earn participation points for that class. These participation points will represent a significant portion of your final project grade.

**Instruction & Support:**

Eric, Daniel and I will be supporting you with your project activities. During class meetings, we will make the rounds, check-in with your team, answer any questions, and help you if you get stuck. We ask that you prepare your questions and bring them up when we make our rounds.

Due to the high number of teams, we are going to specialize in different areas: Eric and Daniel will address any issues specific to GitHub, Python, and Beautiful Soup. I will address any issues specific to the database design, SQL, and MySQL. The three of us will support you on any issues related to the MySQL connector for Python.

To emphasize, if you have an outstanding issue that you need our help with, speak to us when we make our rounds. If your issue has not been addressed by the end of class time, make note of it in your progress report. We will be reading your progress report carefully and will follow-up with you on any issues raised in your report.

**Office Hours:**

Starting on Wednesday, my office hours will be held in **GDC 4.314**. My office hours will start after class at ~7:45pm and will end once the last person leaves. Eric and Daniel will continue to hold their office hours at the same time & location as before.

**Milestones & Important Dates:**

*Note: For email-based submissions, be sure to follow the subject format carefully. All due dates are before the specified class time, unless otherwise noted. Only the latest on-time submission will be counted. If your only submission is a late one, the penalty is 50% of the milestone grade.*

*All milestones below must be submitted to receive a non-zero project grade.*

**M1.** Project Teams due **Wednesday 03/23**. List the members of your team. At this point, each team member should have created a GitHub account. **One person from your team should email the professor and both TAs, carbon-copying your team members, the following: your team name (please avoid bad taste), followed by the full names and emails of all team members, along with their GitHub IDs. Your email should have the following subject, replacing <TeamName> with your team name: [CS327E][M1][<TeamName>].**

**M2.** Project Proposals due **Monday 03/28**. A short description of your project that includes screenshots of the web pages you will use, the data you will extract from those web pages, any other datasets you will use (if applicable), the transformations you will make to the data, and the queries you plan to support. Also, list the planned responsibilities for each team member and any challenges/concerns you foresee toward accomplishing the project. **One person from your team should email the professor and both TAs, carbon-copying your team members, the proposal as a PDF attachment. Your email should have the following subject, replacing <TeamName> with your team name: [CS327E][M2][<TeamName>][Proposal].**

**M3.** Computer Setup due **Monday 04/04**. Install Python 3.4, Beautiful Soup 4, and PyMySQL 0.72 and optionally MySQL Workbench 6.3. Please see Chapter 1 from our Web Scraping book for instructions on how to install Beautiful Soup 4. Please see Chapter 5 for instructions on how to install PyMySQL using pip. Once installation is complete, run a test to ensure that you can read a web page using Beautiful Soup and another test to ensure that you can connect to MySQL from Python.

Finally, one member should create a private code repo under the cs327e-spring2016 organization for this project. The name of the repo should be the name of your team. Each team member should be joined as an owner on that repo. Commit a simple “Hello World” Python3 program to your repo.

**Each person from your team should email the professor and both TAs a screenshot of your BeautifulSoup test, your database connection test, as well as your “Hello World” GitHub repo link via email. Your email should have the following subject, replacing <TeamName> with your team name: [CS327E][M3][<TeamName>][Setup].** We will automatically verify that your team repo is set up.

**M4.** Data model for your project due **Monday 04/11**. The data model should consist of a conceptual diagram for your database schema, a data dictionary that describes the important entities and attributes, and a listing of the most important queries that your query interface will support. Remember that you can diagram the schema using LucidChart (instead of by hand). **One person from your team should email the professor and both TAs, carbon-copying your team members, the data model as a PDF attachment. Your email should have the following subject, replacing <TeamName> with your team name: [CS327E][M4][<TeamName>][Data Model].**

**M5.** Develop the data pipeline and query interface due **Monday 04/25**. Your team will have 2 weeks to do the core development. Since two weeks go by very quickly, you must be sure to assign specific tasks and deadlines to each team member via the GitHub Issue Tracker. That said, all team members should understand how the entire codebase works and should be able to answer any questions about it. Also, as you develop the code, remember to check in changes

to the repo on a regular basis. **One person from your team should email the professor and both TAs, carbon-copying your team members, the following: a link to your team GitHub Repo, as well as the SHA1 commit hash of the code you wish to submit. Your email should have the following subject, replacing <TeamName> with your team name: [CS327E][M5][<TeamName>][Demo Code].** You may continue to work (and check in code) after this. However, we'll be grading M5 based on this commit to be fair to other teams who have an earlier demo.

**M6.** Project Demos. Present a live-demo of your project. Each team will be given a 10-minute time slot to present their project. The presentation will be held in my office, GDC 4.314, during class times between Monday 04/25 and Wednesday 05/04. When you are not presenting, you can continue to develop your project. We understand that teams demo-ing earlier will likely have a less-refined project. That is okay, but all team members should be prepared to answer any questions that we have about the project. **Before your demo, one person from your team should email the professor and both TAs, carbon-copying your team members, your presentation as a PDF attachment. Your email should have the following subject, replacing <TeamName> with your team name: [CS327E][M6][<TeamName>][Presentation].** If you are demo-ing a different version of code than the one you submitted in M5, be sure to include the SHA1 commit hash of the code in your email. If you are presenting on the first week, you won't be expected to demo a finished product.

**M7.** Final Report & Submission due **Friday 05/06 5:00PM**. Write a 5-10 page description of your team's work. Include screenshots of the web pages you used and the conceptual diagram(s) and data dictionary for the database design. Describe the data pipeline you built and any technical challenges you faced during the development process. Note any improvements you made to the code since the demo and any general lessons learned. (Note that the page count is exclusive of any screenshots that you include). **One person from your team should email the professor and both TAs, carbon-copying your team members, the final report as a PDF attachment. Your email should have the following subject, replacing <TeamName> with your team name: [CS327E][M7][<TeamName>][Final Report & Submission]. Before the due date, be sure all code is checked into GitHub!** You will also submit a self-evaluation and an evaluation for each team member. The evaluations will be done via a Google Form which will be shared at a later date.

**Final Grade:**

Your grade on the project will be determined by how well you do in the following areas:

1. Functionality
2. Project Demo
3. Final Report
4. Milestones
5. Teamwork

**References:**

Ryan Mitchell's Web Scraping with Python: Collecting Data from the Modern Web, 2015.

Mark Pilgrim's Dive into Python 3. Free online version: DiveIntoPython3.org
PyMySQL docs: https://pypi.python.org/pypi/PyMySQL
PyMySQL mailing list: https://groups.google.com/forum/#!forum/pymysql-users
GitHub Issue Tracker: https://guides.github.com/features/issues/
LucidChart: https://www.lucidchart.com/
What Google Learned From Its Quest to Build the Perfect Team: http://tinyurl.com/zev3s78