**INFO 430**

**He Chen, Michelle Chen, Lucas Woo, Henry Lin**

**Needed Data + Source:**

The data that we are staring with is from <https://www.washington.edu/students/timeschd/WIN2019/#AUP>. This source has all the UW classes, credits per class, department, and credit type. We will then combine this with data for student feedback on teacher performance and course ratings. The source of data will be from <https://www.ratemyprofessors.com/>. This will give us the overall quality, level of difficulty, date and comments. Further, we want to combine this with course evaluations from the University of Washington <https://www.washington.edu/cec/toc.html>. We will be getting information ratings for: the course as a whole, the course content, instructor’s contribution, instructor’s effectiveness, instructor’s interest, amount learned, and grading techniques. The last data that we will need is data for metrics such as teaching style, participation required, level of group work required, lectures recorded (Y/N), weight of exams towards grades, etc. This data will be newly created data since there are no sources that track these metrics. As a result, we will be dealing with a pretty substantial amount of data. Considering there’s hundreds of different classes, multiple different teachers for these classes, and potentially many reviews for each teacher, we will likely be dealing with thousands of rows of data. Because we are not taking this to full website implementation, our data will not increase; however, there would be new teachers and new reviews coming in if we did. For right now, we are focusing our scope of just University of Washington reviews, classes, and professors. Similar processes could be carried out in the future to expand it to different universities.

**Methodology on Data Collection:**

For rate my professor, course evaluations, and course details we will have to use Python techniques to scrape the data from the web. For the rest of the data, we will need to make up data since there is no readily available source. After collecting and creating data, we will have to make sure that there are no null values and clean the data further. For example, we will have to convert rate my professors date taken to year and quarter taken. Because we are extracting data from multiple different sources, the challenge with cleaning the data will be converting them all to be compatible and of similar types. Thus, this presents a lengthy process since the data from disparate sources will all have different styles and input types for their data. Further, linking the current reviews to users also presents a massive challenge. All the data from the three sources, as well as the made-up data, will be stored in a central SQL database; each course and its related information will be grouped in rows as shown in the attached ERD. Our system does take user-submitted data as part of the review. Ideally, user input will be matched against the data already stored in our database and inspected on a regular basis to ensure validity. Further, users would have to create a user account linking the account to their university email to ensure validity of reviews. There would also be an option to report the review if the review seems malicious. Those reviews that were reported could be pulled every week and analyzed to determine if the review was a fair review. But since we won’t be taking this project to full website implementation, we won’t need to worry about validating user input since it’s out of scope.