CSCI S-101 Foundations of Data Science and Engineering

Faculty: Bruce Huang. Ph.D
Email: bruce_huang@fas.harvard.edu

Teaching Staff (TF/TA):

- **Head TF/TA:** Clement Lee
- TF/TAs: Michael Chang
- TF/TAs: Lucas Chu
- TF/TAs: Kimberly Moon
- TBD

Course Description:

Most data scientists spend 20% of their time building data models and analyzing model results. What do they do with the remaining 80 percent of their time? The answer is data engineering. Data engineering is a subdiscipline of software engineering that focuses on the transportation, transformation, and management of data. This course takes a comprehensive approach to explore data science, which includes data engineering concepts and techniques. Key topics include data management and transformation, exploratory data analysis and Visualization, statistical thinking and machine learning, natural language processing, and storytelling with data, emphasizing the integration of Python, MySQL, Tableau, development, and big data analytics platforms.

Learning Outcomes:

Upon the successful completion of this course, students will be able to:

- Recognize the skills required to perform data science tasks from data acquisition to storytelling with data.
- Demonstrate an understanding of how data science projects are approached.
- Manage data with database management systems and cloud infrastructure.
- Use advanced Python programming techniques to prepare and transform data.
- Apply preattentive attributes and visualization theory in storytelling with data.
- Explore machine learning models to solve business problems.
- Analyze data using Python, MySQL, Tableau, and big data analytics platforms.
- Explore the concept of Natural Language Processing (NLP).
- Communicate and present data science projects and results.

Prerequisites:

CSCI E-7- Introduction to Programming with Python, CSCI E-50 – Intensive Introduction to Computer Science, or equivalent.

Required Course Materials:

- No textbook is required, but students must have:

- Windows, Mac, or Linux laptops with WIFI or Internet access
  - You must have admin/root permission to install and modify the configuration of your computer (Work laptop with access restrictions will not work). You are responsible for having a working operating environment for MySQL, Python, Jupyter Notebook, Excel, Tableau, Webcam, upload and download files.
- Microsoft Excel
- MySQL (open source)
- MySQL Workbench (open source)
- Tableau (instructor will obtain a class license from Tableau)
- Jupyter (open source)
- Python 3+ (open source)
- Webcam (for Exam Proctoring)

Requirements:

Grades for the course will be determined from the following activities:

- Assignments (9) 70%
- Midterm Exam 15%
- Final Exam 15%

Grade System:

A	94% to 100%
A-	< 94% to 90%
B+	< 90% to 87%
B	< 87% to 84%
B-	< 84% to 80%
C+	< 80% to 77%
C	< 77% to 74%
C-	< 74% to 70%
F	< 70% to 0%

If you are taking this course to fulfill the admission course requirement for the master's degree, data science field of study, you will need a B to satisfy the admission course requirement.

Regrade Policy:

Assignment grades recorded in Canvas Gradebook by the TF/TA are the official grades. Students are responsible for making sure that grades have been recorded correctly.

Please adhere to the following policy when making a regrade request:

- If there was a typo or calculation error in your grade, please make your regrade request through Gradescope.
- If you lost points for answers that you believe are correct, please make your regrade request through Gradescope with sufficient details for your TF/TA to review the situation.
- If you lost points for something and do not understand the TF/TA's comments, please speak with your assigned TF/TA during their office hours/ discussion sections, or request a meeting with your TF/TA.
- For any other issues, please get in touch with the instructor.

Students must make regrade requests within one week from the date the grade is posted in Canvas Gradebook. TF/TA may choose to review other parts of the assignment besides the one in question for regrading consideration. There is no guarantee that the revised grade will be greater than the current grade. In some cases, the revised grade may be lower than the original grade due to grading errors discovered during the review. In any case, TF/TA will explain the revised grade. The revised grade will be the final official grade for the assignment.

Attendance/Participation Policy:

This course meets via web conference. Students are encouraged to attend at the scheduled meeting time, but a recorded session will be available to students who cannot attend the live session. All students are expected to participate actively and collaboratively in either the live session, discussion sections, or the Ed Discussion forum.

Late Assignment Policy:

Students are given 3 late days for assignment submission throughout the semester. All assignments will be due at 11:59 pm EST. Late days are automatically used after the assignment is due. For example, all three late days will be used for one assignment that is three days late or three assignments with one day date submission each. There is no need to inform course staff of intent to use late days separately. No credit will be given for an assignment submitted more than 3 days after the submission deadline unless express permission is granted. If a student uses all 3 late days, assignments may still be submitted (up to 3 days late) with a 10% penalty added to the final assignment grade per day.

Accessibility:

Harvard Extension School is committed to providing an accessible academic community. The Accessibility Office offers a variety of accommodations and services to students with documented disabilities. Please visit https://extension.harvard.edu/for-students/support-and-services/accessibility-services/ for more information.

Academic Integrity:

You are responsible for understanding Harvard Extension School policies on academic integrity (https://extension.harvard.edu/for-students/student-policies-conduct/) and how to use sources

responsibly. Not knowing the rules, misunderstanding the rules, running out of time, submitting the wrong draft, or being overwhelmed with multiple demands are not acceptable excuses. To support your learning about academic citation rules, please visit the Resources to Support Academic Integrity (https://extension.harvard.edu/for-students/student-policies-conduct/academic-integrity/) where you will find links to the Harvard Guide to Using Sources (https://usingsources.fas.harvard.edu) and two free online 15-minute tutorials to test your knowledge of academic citation policy. The tutorials are anonymous open-learning tools.

Teaching Fellows/Assistant (TF/TA) Office Hours/Discussion Sections:

In addition to the weekly class meetings, there will be informal TA/TF office hours/discussion sections at which you can get extra help and an Ed Discussion forum on the course website where the teaching staff and your peers will answer questions.

Additional Assignment / Exam Policies:

You are responsible for understanding and comply with assignment and exam specifications (SPECS). In the professional world, programmers need to follow SPECS and seek clarifications when developing products. The teaching staff is here to answer your questions regarding assignment and exam SPECS. You can seek clarification during the TF/TA discussion sections or through Ed Discussion for questions regarding the assignments. You can use the Zoom chat box during the exam. Not knowing the rules, misunderstanding the SPECS, and SPECS are not clear are not acceptable excuses for incorrect answers or not being able to complete the assignment or exam.

WEEK	TOPIC	ASSIGNMENT
1 8/31/2021	**Introduction to Data Science and Statistics Thinking**	PSET 0: Preclass Survey Due: 8:00 am EST on 8/31 Statistics Thinking Exercise (Canvas Quiz) Due: Tuesday, 9/7/2021 11:59 pm EST
2 9/7/2021	**Python Basics Refresher** Data Types Expressions Controls	PSET 1: Python 101 Exercise (Canvas Quiz) Due: Tuesday, 9/14/2021 11:59 pm EST
3 9/14/2021	**Managing Data** Concept of Relational Database Managing High Volume Data Database Management System Installation and administration	

4 9/21/2021	**Managing Data** SQL Programming Python Integration (Cursor)	PSET 2: Managing Data Exercise 1 Due: Tuesday, 9/28/2021 11:59 pm EST
5 9/28/2021	**Managing Data** SQL Programming Python Integration (Pandas)	PSET 3: Managing Data Exercise 2 Due: Tuesday, 10/5/2021 11:59 pm EST
6 10/5/2021	**Exploratory Data Analysis with DML** Exploratory Data Analysis with DML SQL Programming	
7 10/12/2021	**Exploratory Data Analysis with DML** Exploratory Data Analysis with DML SQL Programming Python Integration Python Data Cleaning and Filtering	PSET 4: Exploratory Data Analysis Exercise Due: Tuesday, 10/19/2021 11:59 pm EST
8 10/19/2021	**Storytelling, Visualization, and the use of Tableau to Extract Data** Storytelling Concepts Preattentive Attributes Exploratory and Visualization Tableau	
9 10/26/2021	**Storytelling, Visualization, and the use of Tableau to Extract Data** Tableau Random Sample Generation using Tableau Remote Database Connectivity Data Extraction and Filtering Visualization Principles Graphs, Charts, and Maps Midterm Review	PSET 5: Storytelling and Visualization Exercise Due: Tuesday, 11/2/2021 11:59 pm EST
10 11/2/2021	**Midterm Exam**	Midterm Exam (Class Time Online Proctored via Zoom)
11 11/9/2021	**Python for Data Engineering** Object-Oriented Programming Inheritance, Encapsulation Class and Constructor	PSET 6: Python OO Exercise Due: Tuesday, 11/16/2021 11:59 pm EST
12 11/16/2021	**Machine Learning using Python** Data Structure: Array and Linked List Concepts T-Test and ANOVA using Python	PSET 7: T-Test and ANOVA Exercise Due: Tuesday, 11/23/2021 11:59 pm EST

13 11/23/2021	**Machine Learning using Python and SQL** Regression Model Building and Analysis Classification Model Building and Analysis	PSET 8: Machine Learning Regression and Classification Exercise Due: Tuesday, 11/30/2021 11:59 pm EST
14 11/30/2021	**Python for Natural Language Processing** Text Manipulation	PSET 9: NLP Exercise Due: Tuesday, 12/7/2021 11:59 pm EST
15 12/7/2021	**Final Exam Review**	
16 12/14/2021	**Final Exam**	Final Exam (Class Time Online Proctored via Zoom)