Harvard Extension School
HARVARD DIVISION OF CONTINUING EDUCATION

CSCI E-101

Foundations of Data Science and Engineering

Spring Term 2025

Course Information

CRN: 26190
Section Number: 1
Format: Online
Credit Status: Graduate
Credit Hours: 4

Course Description: Most data scientists spend 20 percent 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. Students cannot earn Harvard Extension School degree credit for CSCI E-101 if it is taken after CSCI E-29.

Prerequisites: CSCI E-7, CSCI E-50, or the equivalent.

Course Overview Video

Instructor Information & Office Hours

Bruce Huang, Ph.D., Ed.D.
Email: bruce_huang@fas.harvard.edu
Additional Information:
https://scholar.harvard.edu/brucehuang/home

Teaching Support Team:

TBD

Section Meetings

In addition to the instructor's scheduled weekly optional live session, there will be weekly live TA/TF discussion sections at which you can get extra help and an Ed Discussion forum on the course website where the TA/TFs and your peers will answer questions.

Each TA/TF of our teaching team will host their own weekly live (via Zoom) discussion section. Our TA/TFs will announce the schedule of their weekly live discussion section at the beginning of the semester. Students can join any or all of the weekly live discussion sections. These TA/TFs discussion sections will be recorded and made available to those unable to join the live discussion sections.

Course Goals / 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.
- Identify the challenges in database management system installation and administration.
- 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, and Tableau.
- Explore the concept of Natural Language Processing (NLP).
- Communicate and present data science projects and results.

Mode of Attendance & Participation Policy

CSCI 101 will be taught using the 'Online' course format in Spring 2025. By definition, 'Online' means there are no scheduled class meetings. Students progress through the course by watching videos on demand and completing assignments and exams. **However, the instructor, Bruce Huang, will host weekly optional attendance live class sessions on Tuesdays.** The day of the week for the optional live class session may change. The weekly optional attendance live class session aims to simulate an actual in-person class session to cover the course topics in a live session setting. The weekly optional attendance live class session will be conducted via Zoom and recorded and made available to those unable to attend the live session. **You can attend the weekly optional attendance live class session over Zoom or watch the recording afterward.**

If you are attending the weekly optional attendance live class session over Zoom:

You should attend Zoom meetings with a functional web camera and microphone with your camera on. You may turn off your camera for occasional interruptions or momentarily for privacy.

You will also need the most up-to-date Zoom client installed on your computer to join the optional attendance live class sessions. Please participate in a safe and appropriate environment with appropriate clothing for class. Participating while traveling or in a car is not permitted. In addition, please do not join the optional attendance live class sessions via mobile phone or web browser.

Pre-recorded Video Lectures:

In addition to the weekly optional attendance live class session, the instructor will provide pre-recorded bite-size videos covering all course topics. These pre-recorded videos will

be published in the Canvas Modules weekly after the weekly optional live class section. The goal of these bite-size videos is to give you an easy way to review individual or selected topics.

Weekly TA Discussion Sections

See the SECTION MEETINGS section for TA Discussion Sections.

Please be sure to review important information on **Student Policies and Conduct**.

Required Course Materials:

- No textbook is required, but students must have:
    1. 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.
    2. Microsoft Excel
    3. MySQL (open source)
    4. MySQL Workbench (open source)
    5. Tableau Desktop (trial or academic license)
    6. Jupyter (open source)
    7. Python 3+ (open source)
    8. Webcam (for Exam Proctoring)

Academic Integrity Policy

You are responsible for understanding Harvard Extension School policies on Academic Integrity and how to use sources responsibly. Violations of academic integrity are taken

very seriously. Visit Using Sources Effectively and Responsibly and the Harvard Guide to Using Sources to review important information on academic citation rules.

AI Technologies. The Extension School's Academic Integrity Policy. prohibits students from representing work as their own that they did not write, code, or create. It is never permissible to submit work generated by machine learning and AI technologies (such as ChatGPT) without proper attribution. Alternatively, your instructor may prohibit all use of AI technologies in their course.

Writing Code. While it may be common practice in non-academic settings to adapt code examples found online or in texts, this is not the case in academia. In particular, you should never copy code produced as coursework by other students, whether in the current term or a previous term; nor may you provide work for other students to use. Copying code from another student or any other source is a form of academic dishonesty, as is deriving a program substantially from the work of another.

Writing code is similar to academic writing in that when you use or adapt code developed by someone else as part of your assigned coursework, you must cite your source. Paraphrasing without proper citation is just as dishonest with programming as it is with prose. A program can be considered plagiarized even though no single line is identical to any line of the source.

Accessibility Services Policy

The Division of Continuing Education (DCE) is committed to providing an accessible academic community. The Accessibility Services Office (ASO) is responsible for providing accommodations to students with disabilities. Students must request accommodations or adjustments through the ASO. Instructors cannot grant accommodation requests without prior ASO approval. It is imperative to be in touch with the ASO as soon as possible to avoid delays in the provision of accommodation.

DCE takes student privacy seriously. Any medical documentation should be provided directly to the ASO if a substantial accommodation is required. If you miss class due to a short-term illness, notify your instructor and/or TA but do not include a doctor's note. Course staff will not request, accept, or review doctor's notes or other medical documentation. For more information, email accessibility@extension.harvard.edu.

Assignments & Grading

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

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

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 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.

Late Assignment Policy

Students are given three late days for assignment submission throughout the semester. All assignments will be due at 11:59 pm ET. Late days are automatically used

after the assignment is due. **You can use up to one late days on each assignment.** For example, you will have one late day left for other assignments if you use one late day for PSET 1 and one late day for PSET 4. There is no need to inform course staff of the intent to use late days separately. Unless express permission is granted, no credit will be given for an assignment submitted late after all free late days are used. No credit will be given to an assignment submitted two or more days after the due date, even if you still have free late days available. **NOTE: Free late days cannot be used on the last assignment (PSET 9).**

While we do not ask for the reasons for using the free late days, the free late days are intended for emergencies such as sick or unavoidable family or personal situations. You will need to contact the Accessibility Office if you need further accomodation.

Additional Assignment / Exam Policies:

You are responsible for understanding and complying 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 the 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.

Publishing or Distributing Course Materials Policy

Students may not post, publish, sell, or otherwise distribute course materials without the written permission of the course instructor. Such materials include, but are not limited to, the following: lecture notes, lecture slides, video, or audio recordings, assignments, problem sets, examinations, other students' work, and answer keys. Students who sell, post, publish, or distribute course materials without written permission, whether for the purposes of soliciting answers or otherwise, may be subject to disciplinary action, up to and including requirement to withdraw. Further, students may not make video or audio recordings of class sessions for their own use without written permission of the instructor.

Canvas Access After End of Term

The Canvas website for this course will remain available to enrolled students for a limited time after the course concludes. **You are encouraged to download coursework and materials you wish to keep *before* the term ends.** See Course Formats & Required Technology for additional information on Canvas access.

Course Schedule

WEEK	TOPIC	ASSIGNMENT
1-M1 1/28/2025	Introduction to Data Science and Statistics Thinking	PSET 0: Preclass Survey Due: 8:00 am ET on 1/27/2025 PSET 1 Statistics Thinking Exercise (Canvas Quiz) Due: Wednesday, 2/5/2025, 11:59 pm ET
2-M2 2/4/2025	Python Basics Refresher Data Types Expressions Controls /User-Defined Functions	Python Practice Exercise – (Nongraded – no submission) PSET 2 Python String Data Manipulation Due: Wednesday 2/12/2025, 11:59 pm ET
3-M3 2/11/2025	Managing Data Concept of Relational Database Managing High-Volume Data Database Management System Installation and Administration	

	SQL Programming Python Integration (Cursor)	
4-M4 2/18/2025	Managing Data SQL Programming Python Integration (Pandas) Python Working with Data Files	PSET 3 Managing Data Exercise 1 Due: Wednesday, 2/26/2025, 11:59 pm ET
5-M5 2/25/2025	Exploratory Data Analysis with DML Exploratory Data Analysis with DML Python Data Cleaning and Missing Data	PSET 4 Managing Data Exercise 2 Due: Wednesday, 3/5/2025, 11:59 pm ET
6-M6 3/4/2025	Exploratory Data Analysis using Python Error Handling in Python Python Data Filtering and Pivot Table Python Data Preparation	PSET 5 Exploratory Data Analysis Exercise Due Wednesday, 3/12/2025, 11:59 pm ET
7-M7 3/11/2025	Midterm Review	
8-M8 3/18/2025	Spring Break Holiday ------- Break ----------	
9-M9 **3/24/2025**	**Midterm Exam** (Live InClass Zoom Proctored Exam – Please make	**NOTE: No make-up exam allowed. You must ensure that you will be available**

	arrangements to be available for the exam session at either 7:00 am ET or 3:00 pm ET)	**to take the midterm-exam in real time on Monday, 3/24/2025 either 7:00 am ET or 3:00 pm ET.**
10-M10 4/1/2025	Storytelling, Visualization, and the Use of Tableau to Extract Data Storytelling Concepts Preattentive Attributes Exploratory and Visualization Visualization Principles Tableau Random Sample Generation using Tableau Remote Database Connectivity Data Extraction and Filtering Graphs, Charts, and Maps	PSET 6 Storytelling and Visualization Exercise Due: Wednesday, 4/9/2025, 11:59 pm ET
11-M11 4/8/2025	Python for Data Engineering Object-Oriented Programming Inheritance, Encapsulation Class and Constructor Data Structure: Array and Linked List Concepts	PSET 7 Python OO Exercise Due: Wednesday, 4/16/2025, 11:59 pm ET
12-M12 4/15/2025	Machine Learning using Python T-TEST and ANOVA using Python	PSET 8 T-TEST and ANOVA Exercise Due: Wednesday, 4/23/2025, 11:59 pm ET
13-M13 4/22/2025	Machine Learning using Python Regression Model Building and Analysis	
14-M14	Classification Model Building and	PSET 9 Machine Learning

4/29/2025	Analysis Python for Natural Language Processing Python Working with Various Data Sources Text Manipulation	Regression, Classification, NLP Exercise Due: Wednesday, 5/7/2025, 11:59 pm ET (No late submission allowed)
15-M15 5/6/2025	Final Exam Review	
16-M16 **5/14/2025**	**Final Exam** (Live InClass Zoom Proctored Exam – Please make arrangements to be available for the exam session at either 7:00 am ET or 3:00 pm ET)	**NOTE: No make-up exam allowed. You must ensure that you will be available to take the final-exam in real time on Wednesday, 5/14/2025, either 7:00 am ET or 3:00 pm ET.**

Final Exam

16-M16 **5/14/2025**	**Final Exam** (Live InClass Zoom Proctored Exam – Please make arrangements to be available for the exam session at either 7:00 am ET or 3:00 pm ET)	**NOTE: No make-up exam allowed. You must ensure that you will be available to take the final-exam in real time on Wednesday, 5/14/2025 either 7:00 am ET or 3:00 pm ET.**