I310D - Introduction to Human-Centered Data Science

Semester: Fall 2024
Time / Venue: Monday / Wednesday 9:30AM - 11:00AM, UTC 1.130

Instructor: Dr. Abhijit Mishra (he/his)
Email: abhijitmishra@utexas.edu

TA:Sai Nikhil Pillai
Email: saipillai@utexas.edu
Canvas: https://utexas.instructure.com/courses/1394982

Office Hours
Abhijit Mishra: Monday 12:00PM-2:00PM and Wednesday 12:00PM-2:00PM (https://utexas.zoom.us/j/8979599959) or by appointments.

TA Office Hours: TBD

Communication and Asking for Help
Please ask all questions that are applicable to the entire class on Canvas, so that others may benefit from the discussion. Only use email for questions unique to individual circumstances; in those cases, please address all questions to both abhijitmishra@utexas.edu and Shaunak (saipillai@utexas.edu)

Course Description

I310D- Introduction to Human-Centered Data Science is a survey course that introduces students to the theory and practice of data science through a human-centered lens, with emphasis on how design choices influence algorithmic results. Students will gain comfort and facility with fundamental principles of data science including (a) Programming for Data Science with Python (b) Data Engineering (c) Database Systems (d) Machine Learning and (e) Human centered aspects such as privacy, bias, fairness, transparency, accountability, reproducibility, interpretability, and societal implications.

Each week's class divided into two segments: (a) **Theory and Methods**, a concise description of theoretical concept in data science, and (b) **Tutorial**, a hands-on session on applying the theory just discussed to a real-world task on publicly available data. We will use Python for programming and cover Python basics in the beginning of the course. For modules related to databases, we will use PostGre SQL.

Prerequisites

I301: Introduction to Informatics is a prerequisite/co-requisite for this course.

Learning Outcomes

Following the completion of the course, students should be able to:

- Explain the lifecycle and requirements for conducting human-centered data science
- Recognize appropriate data-driven strategies to apply in a diverse array of circumstances based on the problem statement, conditions and constraints
- Identify existing or potential ethical, technical, or logical issues within data science applications, and reflect on their societal impact.

Instruction Modality

Class meetings will be in person, with some exceptions and dependent on the state of the COVID-19 endemic. Under extraordinary circumstances, if we are unable to meet in person, classes will be held virtually via Zoom. Classes will be a mixture of lecture and hands-on sessions. As of now, there are **no plans to record lectures**.

Accommodations for Students with Disabilities

The university is committed to creating an accessible and inclusive learning environment consistent with university policy and federal and state law. Please let me know if you experience any barriers to learning so I can work with you to ensure you have equal opportunity to participate fully in this course. If you are a student with a disability, or think you may have a disability, and need accommodations please contact Services for Students with Disabilities (SSD). Please refer to SSD's website for contact and more information: http://diversity.utexas.edu/disability/. If you are already registered with SSD, please deliver your Accommodation Letter to me as early as possible in the semester so we can discuss your approved accommodations and needs in this course.

Required Materials

There is no required textbook for this course; all assigned readings will be available online at no cost. You may find the readings in the course outline section below. Additional materials/resources may be added to canvas prior to each class. **Slides and lecture notes will be provided one week in advance.**

Required Devices

This course requires students to bring their laptop computers, although it is device agnostic. Students may be required to install Python, SQL and Jupyter notebooks.

Class Participation (Attendance through quiz and tutorials)

Students are expected to attend every class, actively participate in discussions, and complete the lab tutorial at the end of each session. Tutorials can be polished and submitted by 11:59 PM on the same day.

While attendance will not be explicitly recorded, I may occasionally administer quizzes during the theory period. Students present in class will be provided with an access code and 10 minutes to complete the quiz. In-class participants will have multiple opportunities to submit the quiz and earn full points. Those absent "without an excused absence" will receive a ZERO for the quiz. **Given that all attendees have the potential to earn full points, quiz dates will not be announced in advance.**

Assignments and Project

1. **Written Assignments: THREE** Reading Reflections
2. **Coding Assignment 1** - Python programming for Data Science (Python)
3. **Coding Assignment 2** - Data Collection and Curation (Python)
4. **Coding Assignment 3 -** Database operations (SQL)
5. **Coding Assignment 4** - Data Bias (Python)
6. **Course Project (group-based)**
    1. Project Group Formation
    2. Proposal and Planning
    3. Project interim progress discussion
    4. Final Presentation
    5. Final Report

There are no traditional examinations within this course. Instead, your assessment will encompass written assignments, coding projects, and collaborative group assignments.

Note: All take-home assignments will be posted on Mondays (after lectures) and will be due on Wednesdays in the following week (10 days of turnaround time).

Late Work and Missed Work

In an effort to accommodate any unexpected personal events, I have enacted **a grace policy of two days** for this course. You do not have to utilize this policy, but if you find yourself struggling with unexpected personal events, I encourage you to email me as soon as possible (in advance of the due date) to notify me that you are using our grace policy. You may either have a two-day grace period for one assignment, or you may have 2 one-day extensions for two different assignments. The only absences that will be considered excused are for religious holidays or extenuating circumstances due to an emergency. If you plan to miss class due to observance of a religious holiday, please let us know at least two weeks in advance. You will not be penalized for this absence, although you will still be responsible for any work you will miss on that day if applicable. In the event of an unexcused absence, we do not guarantee the opportunity to make up missed in-class work, but one may be granted. Check with us for details or arrangements.

I310D Grading Policies

Course grades will be made up of the following components. Final letter grades will be awarded according to the grade cutoffs below, including pluses and minuses.

Grade Component	Percentage
Class participation and attendance (i.e., all graded hands on exercises + in class quizzes)	20%

Grade Component	Percentage
Reading Reflections	10%
Final project	30%
Four Coding Assignments	40%

Grade Breaks

Grade	Cutoff
A	94%
A-	90%
B+	87%
B	84%
B-	80%
C+	77%
C	74%
C-	70%
D+	67%
D	64%
D-	60%
F	< 60%

Course Outline

All instructions, assignments, readings, rubrics and essential information will be on the Canvas website. Check the site regularly and use it to ask questions about the course schedule. Changes to the schedule may be made at my discretion and if circumstances require. For example, we might want to slow down, speed up or drop certain topics depending on student input. It is your responsibility to note these changes when announced.

WEEK 1. Introduction to Human-Centered Data Science (Aug 26 - Aug 28)

Lecture: Introduction, Course Syllabus, Other Concentration Courses in HCDS
Lecture: Data Science Applications, Problems and Solutions

References:

1. Provost, Foster, and Tom Fawcett. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data* 1.1 (2013): 51-59
2. Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, 50(3), 1-42.

Written Assignment 1: Reading Reflection (Posted on Canvas)

WEEK 2. Introduction to Python Programing for Data Science (Sept 4)

Lecture: Programming for Data Science, implementing algorithms - Python Basics-I
Tutorial (Same day): Writing programs using Python in Jupyter notebooks

References:

1. A beginner's guide to algorithmic thinking: **https://learntocodewith.me/posts/algorithmic-thinking**
2. Python Data Structures: https://www.geeksforgeeks.org/python-data-structures/
3. **Geeksforgeeks Tutorial:** https://www.geeksforgeeks.org/data-science-tutorial/#pyt

WEEK 3. Basics of Data Processing and Manipulation (Sept 9 - Sept 11)
Lecture: Data Processing Requirements, Data and Data Formats, Python Advanced Datatypes, Dealing with Dataframes with Pandas
Tutorial: Data operations with Advanced Data Types and Pandas

References:

1. Data Processing with Python (https://www.geeksforgeeks.org/data-science-tutorial/#dat)
2. Pandas Tutorial: https://www.w3schools.com/python/pandas/default.asp (Section: Basic)

Coding Assignment 1: Python Programing for Data Processing (Posted on Canvas)

WEEK 4. Data Engineering I: Data Collection and Management (Sept 16-Spet 18)

Lecture: Data engineering lifecycle, Data pipelines, Resource generation with Crowdsourcing Data pipeline (ETL)
Tutorial: ETL example with Python based web-scraping

References:

1. What is crowdsourcing: https://www.clickworker.com/about-crowdsourcing/
2. Gray, Mary L. and Suri, Siddharth. What it's really like to be one of the ghost workers on Amazon's Mechanical Turk. *Fast Company*, May 2019.

3. Amos, David. 2022. A Practical Introduction to Web Scraping in Python. https://realpython.com/python-web-scraping-practical-introduction/

Written Assignment 2: Reading Reflection (Posted on Canvas)

WEEK 5. Data Engineering II: Data Storytelling, Visualization (Sept 23-Sept 25)

Lecture: Overview of Exploratory Data Analysis, Descriptive Statistics, Data visualization through plots and chats, Libraries and tools, Visualization rules
Tutorial: EDA with data-prep library
References:

1. Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387), 531-554.
2. Python matplotlib tutorial by Jay Parmar: https://blog.quantinsti.com/python-matplotlib-tutorial/
3. Comparing the Five Most Popular EDA tools: https://towardsdatascience.com/comparing-five-most-popular-eda-tools-dccdef05aa4c

Coding Assignment 2: Data Curation and Analysis (Posted on Canvas)

WEEK 6. Database Design I: Relational and Non-relational Databases, Introduction to SQL (Sept 30 - Oct 2)

Lecture: Databases and types, Relational Databases, Non-relational Databases, Introduction to SQL, Example Queries
Tutorial: PostgreSQL setup and Basic SQL Query execution

References:

1. Anderson, B. (2022, June 12). SQL vs. NoSQL databases: What's the difference? IBM. Retrieved January 7, 2023, from https://www.ibm.com/cloud/blog/sql-vs-nosql

Project: Group Formation (Posted on Canvas)

WEEK 7. Database Design II: More on SQL and Database Operations (Oct 7 - Oct 9)

Lecture: Data Definition and Data Manipulation Languages, Examples of DDL and DML in SQL, Data Definition Language (DDL) - CREATE, ALTER and DROP Tables, Data Manipulation Language (INSERT, UPDATE, DELETE statements), Retrieving data through SELECT statement, the WHERE clause, basics of database joining operations

Tutorial: Executing DDL , DML and Selection queries in PostGre SQL. Inner and outer join examples

References:

1. Handout DDL and DML : https://www.eecs.yorku.ca/~papaggel/courses/eecs3421/docs/lectures/05-dml-ddl-views-indexes.pdf
2. Handout SQL DML : https://course.ccs.neu.edu/cs5200f17/ssl/lectures/lecture_03_sql_1.pdf
3. SQL Basics and Joins: https://ocw.mit.edu/courses/1-204-computer-algorithms-in-systems-engineering-spring-2010/ca8f491f5f7f57335b87a52700e54849_MIT1_204S10_lec03.pdf

Coding Assignment 3: Database Operations (Posted in Canvas)

WEEK 8. Machine Learning I: Introduction to ML and Deep Learning (Oct 14 - Oct 16)

Lecture: What is machine learning, Predictive analysis using machine learning, ML Applications, Type of learners, Classification and Regression, ML Algorithms, Brief Introduction to Neural Network and Deep Learning

Tutorial:, Hands-on: ML on classification data

References:

1. Raschka, S. (2020, August 5). Chapter 1: Introduction to Machine Learning and Deep Learning. Sebastian Raschka, PhD. Retrieved January 7, 2023, from https://sebastianraschka.com/blog/2020/intro-to-dl-ch01.html
2. Introduction to Machine Learning with Scikit Learn https://scikit-learn.org/stable/tutorial/basic/tutorial.html

WEEK 9. Machine Learning II: Unsupervised methods (Oct 21 - Oct 23)

Lecture: Introduction to unsupervised methods, unsupervised ML examples Clustering, k-means algorithm, Time series data and Anomaly detection

Tutorial: Anomaly detection example through K-means clustering with Python

References:

1. K-means clustering blog: https://neptune.ai/blog/k-means-clustering
2. Sarkar, D. (D. J. (2018, November 17). A comprehensive hands-on guide to transfer learning with real-world applications in Deep learning. Medium. Retrieved January 7, 2023, from https://towardsdatascience.com/a-comprehensive-hands-on-guide-to-transfer-learning-with-real-world-applications-in-deep-learning-212bf3b2f27a

Project: Proposal and Planning Document (Posted on Canvas)

WEEK 10. Human Centered Design - Ethics, Privacy and Consent, Research Methods (Oct 28 - Oct 30)

Lecture: Privacy and Fair Information Practices, Sensitive and Private Data, Data Anonymization Techniques, Differential Privacy, Data legislation, Research Methods
Lecture: Quantitative vs Qualitative methods, Mixed methods, Ethnography

References:
1. S. Barocas and H. Nissenbaum (2014). Big Data's End Run Around Consent and Anonymity. In Privacy, Big Data and the Public Good. Cambridge University Press
2. Lim, Yish. (2019). Doing Data the Right Way. Medium, Towards Data Science, 9 Jan. 2019,
3. Wang, Tricia. Why Big Data Needs Thick Data. Ethnography Matters, 2016.

Written Assignment: Reading Reflection 3 (Posted on Canvas)

WEEK 11. Bias, Fairness, and Transparency (Nov 4 - Nov 6)

Lectures: Bias and Fairness Data bias, Algorithmic bias, Evaluation Bias, Fairness and Accountability, Example
Tutorial: Data bias example using Python

References:
1. Ingold, David and Soper, Spencer. (2016). Amazon Doesn't Consider the Race of Its Customers. Should It?. *Bloomberg*
2. Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. (2018). Machine Bias: Risk Assessment in Criminal Sentencing. *Propublica*, May 2018.

Coding Assignment 4: Data Bias (Posted on Canvas)

WEEK 12. Research and Reproducibility, Open Source Project Development (Nov 11 - Nov 13)

Lectures: Introduction to Open Research, Reproducibility and Replicability, Best Practices in Research, OSS Development Philosophy, Data and Code Licenses

Tutorial: Git basics and project creation and collaboration using GIT

References:
1. Kitzes, J., Turek, D., & Deniz, F. (Eds.). (2018). The practice of reproducible research: case studies and lessons from the data-intensive sciences. *Univ of California Press* Chapter 2: Assessing Reproducibility and Chapter 3: The Basic Reproducible Workflow
2. Gienow, Michelle. (2018). Tutorial: Git for Absolutely Everyone. https://thenewstack.io/tutorial-git-for-absolutely-everyone/

Project: Progress Update Document (Posted on Canvas)

WEEK 13. Algorithmic Auditing, Model Interpretability, Collaboration in Data Science (Nov 18- Nov 20)

Lecture: Algorithmic Auditing, Importance of Interpreting, Interpretability vs Explainability, LIME, Interpretability in Neural Nets

Lecture: Working in organizations and communities, Working across disciplines, Handling people's data, Human AI collaborations

References:

1. Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). Local Interpretable Model-Agnostic Explanations: An Introduction. O'Reilly.
2. The Algorithmic Auditing Trap: https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d

WEEK 14. Fall Break (NO CLASSES HELD)

WEEK 15. Course Summary, Data Science Career Options, Final Project Presentation -I (

Lecture: Data Science Career Options, Course Overview
Presentation: Final Project Presentation 1

Project: Project slides (Posted on Canvas)

WEEK 16. Final Project Presentation and Group Activities (Cont...)
Presentation: Final Project Presentation 2

WEEK 17. Offline Activities
Final Project Report Submission - Dec 12, 2024

Mantra for student success : Navigating the I310D HCDS course

- Achieve higher attendance, aiming for 100% to maximize exposure and engagement during lectures and practical exercises.
- Submit practicums and assignments promptly, recognizing that minor errors can be overlooked while focusing on continuous improvement.
- Prioritize transparency by appropriately citing tools, resources, and data sources, showcasing your commitment to ethical and accountable work.
- Approach in-class quizzes with a clear understanding and well-organized thoughts, leveraging your conceptual clarity to excel.
- If programming presents challenges, embrace deliberate practice to strengthen your skills and confidently navigate technical aspects. Repeatedly and extensively seek help from the instructor / TA during office hours.
- Embrace iteration as you prepare presentations, ensuring impactful task. demonstrations, comprehensive analyses, and well-structured reports.

- Recognize that success in the course is a result of these concerted efforts, culminating in your growth as a proficient and accomplished data practitioner.

Academic Integrity

Students who violate University rules on academic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since such dishonesty harms the individual, all students, and the integrity of the University, policies on academic dishonesty will be strictly enforced. For further information, please visit the Student Conduct and Academic Integrity website at http://deanofstudents.utexas.edu/conduct.

AI Tools Usage Policy

The utilization of AI-powered tools, including platforms like ChatGPT, Google Gemini, Meta LLaMa, DALL-E, or ANY other small/large language/image/audio/video generative models, to create content such as text, code, images, multimedia, or any related materials intended for assignments, quizzes, or projects that contribute directly to the evaluation of grades within this course is **strictly proscribed**. Exceptions to this rule apply only if the incorporation of such systems aligns with the specified objectives of the assignment or project. Breaching this policy may result in the initiation of proceedings related to student misconduct.

Should there be any suspicion surrounding the content submitted by a student, suggesting the involvement of an AI tool, I retain the authority to request clarification from the student. This clarification may be sought through email communication or arranged verbal discussions in the form of one-on-one meetings. In the event of any inconsistencies between the provided explanations and the submitted solutions, I reserve the right to instigate misconduct proceedings against the concerned student. Upon enrolling in this course, students inherently express their agreement to adhere to this policy as well as any forthcoming policies described below.

Course Material Sharing Policy

Unauthorized sharing or distribution of lecture notes, slides, or examination questions is **strictly prohibited** without prior permission from the instructors. Failure to adhere to this policy may result in the initiation of legal actions. In the event that class should be recorded, class recordings are reserved only for students in this class for educational purposes and are protected under FERPA. The recordings should not be shared outside the class in any form. Violation of these restrictions by a student could lead to Student Misconduct proceedings.

Religious Holy Days

By UT Austin policy, you must notify me of your pending absence as far in advance as possible of the date of observance of a religious holy day. If you must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, you will be given an opportunity to complete the missed work within a reasonable time after the absence.

Names and Pronouns

Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, culture, religion, politics, sexual orientation, gender, gender variance, and nationalities. I will gladly honor your request to address you by your chosen name and by the gender pronouns you use. Class rosters are provided to the instructor with the student's chosen (not legal) name, if you have provided one. If you wish to provide or update a chosen name, that can be done easily at this page, and you can add your pronouns to Canvas.

Basic Needs Security

Any student who faces challenges securing their food or housing and believes this may affect their performance in the course is urged to contact the Dean of Students for support. UT maintains the UT Outpost which is a free on-campus food pantry and career closet.

Mental Health Support

I urge students who are struggling for any reason and who believe that it might impact their performance in the course to reach out to me if they feel comfortable. This will allow me to provide any resources or accommodations that I can. If immediate mental health assistance is needed, call the Counseling and Mental Health Center (CMHC) at 512-471-3515 or you may also contact Bryce Moffett, LCSW (iSchool CARE counselor) at 512-232-2983. Outside CMHC business hours (8a.m.-5p.m., Monday-Friday), contact the CMHC 24/7 Crisis Line at 512-471-2255.

Land Acknowledgement

I would like to acknowledge that we are meeting on the Indigenous lands of Turtle Island, the ancestral name for what now is called North America. Moreover, I would like to acknowledge the Alabama-Coushatta, Caddo, Carrizo/Comecrudo, Coahuiltecan, Comanche, Kickapoo, Lipan Apache, Tonkawa and Ysleta Del Sur Pueblo, and all the American Indian and Indigenous Peoples and communities who have been or have become a part of these lands and territories in Texas.

Title IX Reporting

Title IX is a federal law that protects against sex and gender-based discrimination, sexual harassment, sexual assault, unprofessional or inappropriate conduct of a sexual nature, dating/domestic violence and stalking at federally funded educational institutions. UT Austin is committed to fostering a learning and working environment free from discrimination in all its forms. When unprofessional or inappropriate conduct of a sexual nature occurs in our community, the university can:

1. Intervene to prevent harmful behavior from continuing or escalating.
2. Provide support and remedies to students and employees who have experienced harm or have become involved in a Title IX investigation.
3. Investigate and discipline violations of the university's relevant policies.

Beginning January 1, 2020, Texas Senate Bill 212 requires all employees of Texas universities, including faculty, report any information to the Title IX Office regarding sexual harassment, sexual assault, dating violence and stalking that is disclosed to them.

Texas law requires that all employees who witness or receive any information of this type (including, but not limited to, writing assignments, class discussions, or one-on-one conversations) must be reported. **I am a Responsible Employee and must report any Title IX related incidents** that are disclosed in writing, discussion, or one-on-one. Before talking with me, or with any faculty or staff member about a Title IX related incident, be sure to ask whether they are a responsible employee. If you would like to speak with someone who can provide support or remedies without making an official report to the university, please email advocate@austin.utexas.edu. For more information about reporting options and resources, visit http://www.titleix.utexas.edu/ , contact the Title IX Office via email at titleix@austin.utexas.edu, or call 512-471-0419.

Although graduate teaching and research assistants are not subject to Texas Senate Bill 212, they are still mandatory reporters under Federal Title IX laws and are required to report a wide range of behaviors we refer to as unprofessional or inappropriate conduct of a sexual nature, including the types of conduct covered under Texas Senate Bill 212. The Title IX office has developed supportive ways to respond to a survivor and compiled campus resources to support survivors.