HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

BST 209
Collaborative Data Science in Healthcare
Summer 2
2-3:30pm ET

Instructor Information
Faculty

Heather Mattie, PhD
Instructor of Data Science
Department of Biostatistics
Harvard T.H. Chan School of Public Health
Email: hemattie@hsph.harvard.edu
Office: Building 1 room 421A
Phone: 617-432-5308

Leo Anthony Celi, MD, MS, MPH
Associate Professor (Part-time), Harvard Medical School
Principal Research Scientist, MIT
Email: lceli@mit.edu
Office: 45 Carleton Street, E25-505
Cambridge, MA 02142
Phone: 617-324-1556

Tom Pollard, PhD
MIT Lab for Computational Physiology
Email: tpollard@mit.edu

Teaching Assistants

Jeff Joseph, SM
Harvard T.H. Chan School of Public Health
Email: jmjoseph@hsph.harvard.edu
Office hour: 1-2pm 7/27 – 7/30

Meg Salvia
PhD Candidate
Harvard T.H. Chan School of Public Health
Email: msalvia@g.harvard.edu
Office hour: 1-2pm 8/2, 8/3, 8/5, 12-1pm 8/4

Credits
2.5 credits

Course Description
The first two weeks of this course focus on methods for learning from data in order to gain useful predictions and insights. Through real-world examples of wide interest, we introduce methods for five key facets of an investigation:

1) data wrangling/cleaning in order to construct an informative, manageable data set;
2) software engineering skills for accessing data as well as organizing data analyses and making these analyses sharable and reproducible;
3) exploratory data analysis to generate hypotheses and intuition about the data;
4) inference and prediction based on statistical tools with a focus on machine learning;
5) communication of results through visualization, stories, and interpretable summaries.

During the last week of the course, with the help of the instructors and TAs, student teams will choose a clinically relevant question and complete a group project that includes parsing the question into a study design and methodology for data analysis and interpretation, with an emphasis on the data curation that is required before any analysis can be performed. The Medical Information Mart for Intensive Care (MIMIC) database or the eICU Collaborative

Research Database will be used for each project. Students are expected to be familiar with R and RStudio before enrolling in this course.

- **Pre-Requisites and Co-Requisites**
  Students should have completed BST 206, 207 and 208, or be taking BST 207 and 208 concurrently. Students are also expected to have R and RStudio installed on their laptops and to be familiar with programming in R. The instructors highly recommend completing an online introduction to R course (via edX, Coursera, Data Camp, Udacity, etc.) before enrolling in this course if the student hasn't been exposed to R. A brief review of programming in R will take place during the first session.

Learning Objectives
Upon successful completion of this course, you should be able to:

- Wrangle data in a meaningful way
- Visualize data
- Perform exploratory data analyses
- Perform analyses using machine learning
- Communicate analysis results
- Collaborate on data science projects with diverse teams
- Make your work reproducible

Course Readings:

- Required:
  - An Introduction to Statistical Learning in R
  - MIMIC IV Database documentation
  - eICU Database documentation
- Recommended:
  - R for Data Science
  - Data Visualization with R
  - Reproducible Research with R and RStudio
  - The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences

Course Structure

Classroom Participation:
This course includes in-class coding work and a group project component. Students are expected to be active participants. This includes attending all classes and being ready to offer suggestions and insights during group discussions.

Canvas Course Website: Course announcements will be posted on the course Canvas site.

Technical Information: Students are expected to have R and RStudio installed on their laptops before the first session.

Grading, Progress and Assessment
Class attendance and thoughtful participation are important and will be reflected in part in the final grade. Please notify an instructor or TA of an absence before the class. Because the course will be held remotely, if students are unable to attend the live lectures, they may view the lecture recordings and email a short summary and any questions to the teaching team as a substitute for class attendance and participation.

The final grade for this course will be based on:

- Attendance and participation 60%
- Group project 40%
    - Project proposal (10%)
    - Project abstract (10%)
    - Project notebook (10%)
    - Project presentation (10%)

Class attendance and participation
Class attendance is mandatory. Students will be excused from class in the event of a family emergency, medical issue, religious observance, or other extenuating circumstance, and should contact the instructors or TAs to inform them of their absence. A maximum of one absence is allowed without penalty to the class attendance grade and overall grade. A 10% deduction in class attendance grade will be taken for every additional missed class.

Because the course will be held remotely, if students are unable to attend the live lectures, they may view the lecture recordings and email a short summary and any questions to the teaching team as a substitute for class attendance and participation. Students who are unable to attend live lectures will also be put into small groups to work on the project asynchronously with an assigned TA.

Group Project
Students will be assigned to a team of 2-3 individuals. Each team will be assigned a data scientist teaching assistant and given code to pull data from a database and a research question prompt. Each team must first work on a project proposal including:

1. A rationale/motivation for the work
    a. This should include a specific scientific question(s), hypotheses and knowledge gap the work will fill
2. The data to be used and why it is appropriate
    a. The final sample size for the project
    b. The extent of missingness and how missing data will be handled
    c. The outcome of interest
    d. The predictors/covariates to be used
    e. Any data limitations
3. Analysis(es) that will be performed
    a. Include why they are appropriate, e.g. if the outcome is binary then logistic regression may be an appropriate method
    b. How the analysis(es) will answer the research question(s)

Teams will receive feedback and guidance about their proposal the day after it is submitted. The next deliverable is a completed project abstract. Teams will receive feedback and guidance the day after it is submitted. Each team will be responsible for submitting an R Markdown notebook that includes abstract, introduction, methods, results, discussion and references sections with

commented code interspersed. Visuals and schematics should be included. In addition, the code used must be submitted in a reproducible format, and all project documents uploaded to the course Canvas site and GitHub repository. Each group must also give a 5-minute presentation on the last day of class.

Harvard Chan Policies and Expectations

Inclusivity Statement

Diversity and inclusiveness are fundamental to public health education and practice. Students are encouraged to have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact me if you have any concerns or suggestions.

Bias Related Incident Reporting

The Harvard Chan School believes all members of our community should be able to study and work in an environment where they feel safe and respected. As a mechanism to promote an inclusive community, we have created an anonymous bias-related incident reporting system. If you have experienced bias, please submit a report here so that the administration can track and address concerns as they arise and to better support members of the Harvard Chan community.

Title IX

The following policy applies to all Harvard University students, faculty, staff, appointees, or third parties: Harvard University Sexual and Gender-Based Harassment Policy.
Procedures For Complaints Against a Faculty Member
Procedures For Complaints Against Non-Faculty Academic Appointees

Academic Integrity

Each student in this course is expected to abide by the Harvard University and the Harvard T.H. Chan School of Public Health School's standards of Academic Integrity. All work submitted to meet course requirements is expected to be a student's own work. In the preparation of work submitted to meet course requirements, students should always take great care to distinguish their own ideas and knowledge from information derived from sources.

Students must assume that collaboration in the completion of assignments is prohibited unless explicitly specified. Students must acknowledge any collaboration and its extent in all submitted work. This requirement applies to collaboration on editing as well as collaboration on substance.

Should academic misconduct occur, the student(s) may be subject to disciplinary action as outlined in the Student Handbook. See the Student Handbook for additional policies related to academic integrity and disciplinary actions.

Accommodations for Students with Disabilities

Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact Colleen Cronin ccronin@hsph.harvard.edu in all cases, including temporary disabilities.

Religious Holidays, Absence Due to

According to Chapter 151c, Section 2B, of the General Laws of Massachusetts, any student in an educational or vocational training institution, other than a religious or denominational training

institution, who is unable, because of his or her religious beliefs, to attend classes or to participate in any examination, study, or work requirement on a particular day shall be excused from any such examination or requirement which he or she may have missed because of such absence on any particular day, provided that such makeup examination or work shall not create an unreasonable burden upon the School. See the student handbook for more information.

Grade of Absence from Examination
A student who cannot attend a regularly scheduled examination must request permission for an alternate examination from the instructor in advance of the examination. See the student handbook for more information.

Final Examination Policy
No student should be required to take more than two examinations during any one day of finals week. Students who have more than two examinations scheduled during a particular day during the final examination period may take their class schedules to the director for student affairs for assistance in arranging for an alternate time for all exams in excess of two. Please refer to the student handbook for the policy.

Course Evaluations
Constructive feedback from students is a valuable resource for improving teaching. The feedback should be specific, focused and respectful. It should also address aspects of the course and teaching that are positive as well as those which need improvement.

Completion of the evaluation is a requirement for each course. Your grade will not be available until you submit the evaluation. In addition, registration for future terms will be blocked until you have completed evaluations for courses in prior terms.

Schedule at a glance

Session	Date	Topic(s)
1	7/26	Introduction to course Brief review of R, RStudio and RMarkdown
2	7/27	Data wrangling and curation
3	7/28	Data wrangling and curation
4	7/29	Visualization and visualization principles Exploratory data analysis
5	7/30	Visualization and visualization principles Exploratory data analysis
6	8/2	Introduction to machine learning • Terminology and notation • Training, validation and test sets • Confusion matrix • Prevalence, sensitivity, specificity, other performance metrics • Supervised vs unsupervised Classification vs regression
7	8/3	Logistic regression k-nearest neighbors (kNN)
8	8/4	Decision trees Random forest
9	8/5	PCA

		k-means clustering
10	8/6	Begin group project
11	8/9	Continue group project
12	8/10	Continue group project
13	8/11	Continue group project
14	8/12	Continue group project
15	8/13	Group project presentations

Course Schedule & Assessment of Student Learning

Session topics	Objectives	Readings	Activities/ Assignments
Session 1: July 26			
1. Introduction to course **2. Brief review of R, RStudio and RMarkdown** **3. Reproducible data science**	**Upon successful completion of this session, you should be able to:** 1. Write R code in RMarkdown 2. Describe the fundamentals and importance of reproducible research 3. Create a fully reproducible research project	**Required** 1. Reproducible Research with R and RStudio, chapters 1-3 **Recommended** 2. The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences	In-class coding examples
Session 2: July 27			
1. Data wrangling and curation **2. Exploratory data analysis**	**Upon successful completion of this session, you should be able to:** 1. Organize and clean datasets in preparation for analysis 2. Perform an effective exploratory data analysis including visualizations and summary statistics	**Required** 1. R for Data Science, chapters 2 and 5 **Recommended** 2. Reproducible Research with R and RStudio, chapters 4,5,7	In-class coding examples
Session 3: July 28			
1. Data wrangling and curation	**Upon successful completion of this session, you should be able to:**	**Required** 1. R for Data Science, chapters 2 and 5	In-class coding examples

2. Exploratory data analysis	1. Organize and clean datasets in preparation for analysis 2. Perform an effective exploratory data analysis including visualizations and summary statistics	**Recommended** 2. Reproducible Research with R and RStudio, chapters 4,5,7	
Session 4: July 29			
1. Visualization and visualization principles	**Upon successful completion of this session, you should be able to:** 1. Create visually pleasing and informative visualizations using ggplot2	**Required** 1. R for Data Science, chapter 3 2. Data Visualization with R, chapters 2-5	In-class coding examples
Session 5: July 30			
1. Visualization and visualization principles	**Upon successful completion of this session, you should be able to:** 1. Create visually pleasing and informative visualizations using ggplot2	**Required** 1. R for Data Science, chapter 3 2. Data Visualization with R, chapters 2-5	In-class coding examples
Session 6: Aug 2			
1. Introduction to machine learning terminology and notation **2. Training, validation and test sets** **3. Confusion matrix** **4. Prevalence, sensitivity, specificity, other metrics** **5. Supervised vs unsupervised** **6. Classification vs regression**	**Upon successful completion of this session, you should be able to:** 1. Define machine learning terms 2. Split a dataset into training, validation and tests sets 3. Calculate various machine learning model performance metrics 4. Categorize a machine learning method as either supervised or unsupervised, classification or regression	**Required** 1. Introduction to Statistical Learning in R (ISLR), chapter 2	In-class coding examples

Session 7: Aug 3			
1. Logistic regression	**Upon successful completion of this session, you should be able to:** 1. Fit a logistic regression model for prediction in R 2. Assess model performance	**Required** 1. Introduction to Statistical Learning in R (ISLR), chapter 4, sections 1-3	In-class coding examples
Session 8: Aug 4			
1. Decision trees **2. Random forest**	**Upon successful completion of this session, you should be able to:** 1. Make predictions using decision trees and random forests in R 2. Assess the performance of fit decision trees and random forests	**Required** 1. Introduction to Statistical Learning in R (ISLR), chapter 8	In-class coding examples
Session 9: Aug 5			
1. Principle components analysis (PCA) **2. k-means clustering**	**Upon successful completion of this session, you should be able to:** 1. Code PCA and k-means clustering in R 2. Identify when to use PCA 3. Assess the performance of a k-means clustering model	**Required** 1. Introduction to Statistical Learning in R (ISLR), chapter 10	In-class coding examples
Session 10: Aug 6			
1. Introduction to group projects **2. Begin work on group projects**	**Upon successful completion of this session, you should be able to:** 1. Access the MIMIC IV and eICU databases 2. Brainstorm clinically effective research questions to be answered with available data	**Required** 1. MIMIC IV Database documentation 2. eICU Database documentation	Group project work

	3. Effectively collaborate on a group project		
Session 11: Aug 9			
1. Continue work on group projects	**Upon successful completion of this session, you should be able to:** 1. Effectively collaborate on a group project	**Required** 1. MIMIC IV Database documentation 2. eICU Database documentation	Group project work Deliverable: group project proposal
Session 12: Aug 10			
1. Continue work on group projects	**Upon successful completion of this session, you should be able to:** 1. Effectively collaborate on a group project	**Required** 1. MIMIC IV Database documentation 2. eICU Database documentation	Group project work
Session 13: Aug 11			
1. Continue work on group projects	**Upon successful completion of this session, you should be able to:** 1. Effectively collaborate on a group project	**Required** 1. MIMIC IV Database documentation 2. eICU Database documentation	Group project work Deliverable: group project abstract
Session 14: Aug 12			
1. Continue work on group projects	**Upon successful completion of this session, you should be able to:** 1. Effectively collaborate on a group project	**Required** 1. MIMIC IV Database documentation 2. eICU Database documentation	Group project work Deliverable: group project slides
Session 15: Aug 13			
1. Group project presentations	**Upon Successful completion of this week, you should be able to:** 1. Give an engaging group project presentation		Group project presentations

- Please note, session topics and activities may be subject to change during the course