

Biostat 202
Opportunities and Challenges of Complex Biomedical Data:
Introduction to the Science of "Big Data"
Summer 2021

Course Description: The growing availability of large amounts of data — obtained either through research or electronic capture of everyday activity — has been termed "big data". This course introduces the opportunities and challenges of using biological and health-related "big data" to perform biomedical research. We will distinguish big data from non-big data and explore the phases of data science: obtaining data, cleaning data, visualizing data, analyzing data, and drawing conclusions.

At the conclusion of this course, students will be able to:

- Use software to manipulate and clean "big data";
- Generate effective graphical displays of data;
- Implement different approaches to both supervised (classification and regression) and unsupervised modeling (clustering and data reduction);
- Access public use (and non-public) sources of data such as NHANES.

In addition to practical skills, students will develop a base of knowledge allowing them to:

- Define and describe artificial intelligence, machine learning/data mining, and deep learning;
- Identify structured and unstructured data and choose an appropriate analysis strategy;
- Develop intuition for the potential/limitations of biomedical "big-data";
- Describe the issues that arise when trying to use "big data"-based observational studies to derive causal conclusions.

Course website (<https://courses.ucsf.edu/course/view.php?id=8594>): Please visit the CLE course website for an online syllabus which includes lectures, assignments, and materials for the final project. For public course-related questions (i.e. technical questions, project help), please post to the Course Forum which will serve as a center for troubleshooting and discussion.

Faculty:

Course Director: Aaron Scheffler, PhD*
Assistant Professor
Department of Epidemiology and Biostatistics
aaron.scheffler@ucsf.edu
*Please direct individual course-related questions to the course director.

Lecturer: Aaron Scheffler, PhD
Assistant Professor
Department of Epidemiology and Biostatistics
aaron.scheffler@ucsf.edu

Teaching Assistants: Ntemena Kapula
ntemena.kapula@ucsf.edu

Fanxiu Xiong
fanxiu.xiong@ucsf.edu

Location: Lectures: remote instruction via Zoom
<https://it.ucsf.edu/service/zoom-web-conferencing>
Laboratories: in person instruction (MH 1400)

Course format: Twice weekly lectures introduce the substantive content for each module, which is subsequently reinforced in weekly applied assignments. Weekly computer lab sessions give students the opportunity to learn to use the software, ask questions, and have more interaction with faculty.

Lectures: Mondays and Thursdays: July 22 through August 30. Lectures are divided into two portions, 1. a pre-recorded 60-minute lecture covering the main content, and 2. 45-minute live “in-class” Zoom instruction.

1. Twice weekly pre-recorded lecture posted to CLE. Students are expected to view this lecture prior to in-class instruction.
2. **1:00 to 1:45 PM:** in-class Zoom instruction which will provide a summary of the pre-recorded lecture and give students a chance to ask questions and participate in activities to reinforce content introduced in the pre-recorded lectures.

Zoom link:

(<https://ucsf.zoom.us/j/93139422031?pwd=QXVOZGIUejJpaUZoM3pUelVvU01iZz09>)

Computer Labs: Thursdays: 2:00 to 3:45 PM, July 22 through August 30. You will need to bring a computer to all lab sessions with Orange Data Mining software installed. ***Note: labs will be held in person in MH-1400.*** We encourage you to attend labs in person. If you cannot attend in person lab for reasons related to COVID-19, you are welcome to attend a once weekly office hour to clarify any questions surrounding written laboratory instructions.

Office Hours: Thursdays: 4:00 to 5:00 PM, July 22 through August 30. Office hours are held via Zoom and are meant to answer outstanding questions related to lecture or lab.

Zoom link:

(<https://ucsf.zoom.us/j/93139422031?pwd=QXVOZGIUejJpaUZoM3pUelVvU01iZz09>)

All students are expected to attend or watch all the lectures. Computer labs are an integral part of the course. All students are expected to turn in their weekly assignments through the assigned online portal. All course materials and handouts will be posted on the course's online syllabus.

Software: The software package Orange Data Mining is used in this course. Each student is expected to install Orange on their own computer prior to the first computer lab. This software is available free for students of the course. Instructions on how to obtain the free software and install it will be provided by the course director prior to the course start date.

Grading: 55 % - Weekly take-home assignments
 45 % - Project
 * 3 'components' - 15 %
 * Final project write-up - 30%

Computer lab assignments are due the week after their assignment. In addition, students will hand in a final project consisting of three components and a final write-up (see below for more information). Students must hand in all computer lab assignments/homework problem sets (even if late), complete a satisfactory final project, and receive at least 80% of the total number of points assigned during the quarter to

receive a Satisfactory (if taking Satisfactory/Unsatisfactory) or B (if taking for a letter grade) in the course.

Students not in full-year TICR Programs who satisfactorily pass all course requirements will, upon request, receive a Certificate of Course Completion.

Submitting assignments: All assignments are submitted through CLE. To submit homework and lab assignments, please download the Microsoft Word document corresponding to the assignment and place your answers in line with the questions using a colored, bolded, or underlined text. Project components and the final project should be submitted as written reports with clear organization and formatting.

- Assignment 1 (due 7/29/2021)
- Assignment 2 (due 8/05/2021)
- Assignment 3 (due 8/12/2021)
- Assignment 4 (due 8/19/2021)
- Assignment 5 (due 8/26/2021)

Final project: All students will be required to submit a final project in which they manipulate, clean, and analyze data emanating from a large data source. Students will be given a choice of datasets and guidelines for performing the project. Students may use their own data set at the course director's discretion on a case-by-case basis. The final project will be completed over the course of the class, including three separate components culminating with a final written report.

- Component 1: specify the question, access, and clean the data set (due 8/08/21)
- Component 2: describe the data (due 8/15/21)
- Component 3: analyze the data (due 8/22/21)
- Final project: combine Components 1 through 3 (due 9/02/21)

Lecture schedule:

<u>Date</u>	<u>Location</u>	<u>Topic</u>
Thursday, July 22	Zoom	Lecture 1: Introduction to the course
Monday, July 26	Zoom	Lecture 2: Getting data from large databases
Thursday, July 29	Zoom	Lecture 3: Managing data and data storage

Monday, August 2	Zoom	Lecture 4: ML1, Introduction and supervised learning [regression]
Thursday, August 5	Zoom	Lecture 5: ML2, Supervised learning [classification]
Monday, August 9	Zoom	Lecture 6: ML3, Supervised learning, cont. [classification]
Thursday, August 12	Zoom	Lecture 7: ML4, Cross-validation, ensemble methods, and feature importance
Monday, August 16	Zoom	Lecture 8: ML5, Unsupervised learning [clustering/data reduction]
Thursday, August 19	Zoom	Lecture 9: Causal inference in “Big Data”
Monday, August 23	Zoom	Lecture 10: Data visualization/storytelling
Thursday, August 26	Zoom	Lecture 11: Case studies in “Big Data”
Monday, August 30	Zoom	TBD

TICR Professional Conduct Statement: As a participant in the TICR Program, and a clinical researcher, you are expected to maintain the highest ethical standards. The UCSF School of Medicine Statement of Principles includes three statements that the TICR Program would like you to affirm:

1. I will maintain the highest standards of academic honesty.
2. I will neither give nor receive aid in examinations or assignments *unless such cooperation is expressly permitted by the instructor* (Our policy: mutual collaboration on labs and assignments is fine. The project should be your own work).
3. I will conduct research in an unbiased manner, report results truthfully, and credit ideas developed and work done by others.

We ask that you also affirm these two additional statements relating to collaboration, which is allowed on homework assignments in many TICR courses:

1. I will not use answer keys from prior years.

2. I will write answers in my own words, and, when collaboration is permitted, acknowledge collaborators when answers are jointly formulated.

Unprofessional conduct may result in verbal or written warnings; failure to achieve credit for a particular assignment or examination; overall failure in a course; a letter to your Dean; and/or dismissal from the TICR Program. For the most serious infractions, course failure and notification of the Dean will be recommended. When the appropriate level of disciplinary action is unclear or is in contention, instructors and students may appeal to the TICR Disciplinary Committee for a binding decision.

UCSF Student Disability Services: UCSF is committed to ensuring access to graduate education for all students. Early communication with the relevant administrators is critical to a successful partnership in arranging accommodations. We encourage students with disabilities to disclose to the Office of Student Disability Services. SDS is the appropriate and confidential office for seeking accommodations, and will coordinate communications and procedures with you and the graduate faculty and programs. For complete information and services, please visit UCSF Student Disability Services website (<https://sds.ucsf.edu/>). Please contact the course director if any further accommodation or coordination is needed.