SJSU SAN JOSÉ STATE UNIVERSITY

College of Science · Computer Science

Processing Big Data: Tools and Techniques

Section 02 & 03

CS 131

Spring 2025 In Person 3 Unit(s) 01/23/2025 to 05/12/2025 Modified 01/29/2025

Contact Information

Instructor: Quang Duy Tran

Email: quangduy.tran@sjsu.edu

Note: Email is my main way of contact. Do not send Canvas Message, as I don't check it too often.

Office Hours

Monday, Wednesday, 3:00 PM to 4:00 PM

In Person: DH 282

Zoom: https://sjsu.zoom.us/j/84386899435 -- Passcode: 562340

Note: Office hours is by appointment only (for both in-person and zoom). Notify the instructor via Email to set up the meeting.

Course Information

Section 02

Monday, Wednesday, 12:00 PM to 1:15 PM, Science Building 311

Section 03

Monday, Wednesday, 1:30 PM to 2:45 PM, Duncan Hall 318

Course Description and Requisites

In-depth study of essential tools and techniques for processing big data over the UNIX operating system and/or other operating systems. On UNIX, it includes using grep, sed, awk, join, and programming advanced shell scripts for manipulating big data.

Prerequisite(s): CS 46B or BIOL 123B with a grade of "C-" or better. Allowed Declared Majors: Computer Science BS, Data Science BS, MS Bioinformatics (MS BI).

Letter Graded

Classroom Protocols

- Attendance in this **in-person course** is not mandatory, but regular participation in lectures strongly contributes to student success. While I don't track attendance, students are responsible for keeping up with course material and seeking help if they encounter difficulties or fall behind.
- Lecture recordings will be available as a supplementary reference and should not be replied upon as a substitute for attending class. The instructor cannot guarantee the availability or quality of recordings due to potential technical issues with internet connectivity, equipment, or other factors. Regular in-person attendance is necessary for success in this course.
- When emailing me, please always **start your email subject line with "CS131: XXXXX"** to get my attention. (XXXXX: Subject, for example: CS131: HW1 Question)
- **Plagiarism/Cheating will not be tolerable:** 'F' will be given to your FINAL COURSE GRADE and will be reported to the Department and the University. If academic dishonesty is suspected, I reserve the right to review all previously submitted homework assignments and exams. I may also require students to meet during office hours to explain their work and demonstrate their understanding. (Please be noted: obtaining HW solutions from someone or giving/showing your HW solutions to someone is also treated as plagiarism/cheating.)
- Students are responsible for all materials distributed/posted on Canvas and discussed in the class. I also reserve the right to make announcements in class that may not appear on Canvas.

Program Information

Diversity Statement - At SJSU, it is important to create a safe learning environment where we can explore, learn, and grow together. We strive to build a diverse, equitable, inclusive culture that values, encourages, and supports students from all backgrounds and experiences.

Course Learning Outcomes (CLOs)

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

- Analyze, manipulate and process large-scale data.
- Develop shell scripts for use in data-intensive applications.
- Build data analysis pipelines, automate tasks, make analyses reproducible and shareable.
- Solve big data challenges with the application of machine learning algorithms.
- Apply data science solutions to datasets from example domains, such as biology, business, finance.

- Perform big data analyses efficiently, document and reproduce analyses.

Course Materials

There are no required books for this class. All the necessary material will be available on the class Canvas web page.

Course Requirements and Assignments

Assignments

Assignments must be turned in on time; late submission will not be accepted with the exception of medical emergencies or similar exceptional circumstances that must be discussed in advance with the instructor. All assignments are due at midnight on the announced due date. Extension may be provided if necessary.

Note: Students are expected to work on their own homework with no collaborations. AI tools are prohibited and shall not be used under any circumstances. If solution is taken from online source, students must provide the source in the homework. Violate these rules will be considered as cheating/plagiarism, and will be reported to the Department and University.

Mid-Terms and Final Exams

Midterms may be held in-class via Canvas (students will need a laptop), take home, or both. It is up to the instructor to decide which format of the exam will take place. In-class exams includes conceptual questions in a form of essay or multiple-choice format. Take home exams includes programming, and is due one week since the exams start.

If you are unable to attend any one of the exams, arrangements may be made only if you have a legitimate reason. You need to inform your instructor ahead of time and have written documentation available. If you are unable to attend the exam due to illness or emergency, you also need to inform your instructor before the exam and bring documentation afterwards to request a make-up exam.

There is no Final Exam. Final Exam will be replaced by Team Project.

Team Project

Projects will be carried out in groups of 2 members. Every team will write a final report and present their work at the end of the semester. Topics will be determined in class by the instructor.

Grading Information

Criteria

The components of the final grade will be distributed as follows:

- Homework Assignments: 25% (5-6 programming assignments)
- Midterm 1: 25%
- Midterm 2: 25%
- Team Project: 25%
  - Report + Code: 15%
  - Oral Presentation: 10%

Breakdown

Digit number grades will be assigned according to the following policy:

92 ~ 100 ---- A

90 ~ 91 ---- A-

88 ~ 89 ---- B+

82 ~ 87 ---- B

80 ~ 81 ---- B-

78 ~ 79 ---- C+

72 ~ 77 ---- C

70 ~ 71 ---- C-

68 ~ 69 ---- D+

62 ~ 67 ---- D

60 ~ 61 ---- D-

0 ~ 59 ---- F

University Policies

Per University Policy S16-9 (PDF) (http://www.sjsu.edu/senate/docs/S16-9.pdf), relevant university policy concerning all courses, such as student responsibilities, academic integrity, accommodations, dropping and adding, consent for recording of class, etc. and available student services (e.g. learning assistance, counseling, and other resources) are listed on the Syllabus Information (https://www.sjsu.edu/curriculum/courses/syllabus-info.php) web page. Make sure to visit this page to review and be aware of these university policies and resources.

Course Schedule

This schedule is subject to change. Any change will be communicated via Canvas with fair notice.

Week	Date	Topics, Readings, Assignments
1	01/27	Introduction
1	01/29	Introduction
2	02/03	Machine Learning fundamentals
2	02/05	Machine Learning fundamentals
3	02/10	Theory: Machine Learning for Big Data: Supervised approach
3	02/12	Lab: Machine Learning for Big Data: Supervised approach
4	02/17	Theory: Machine Learning for Big Data: Supervised approach
4	02/19	Lab: Machine Learning for Big Data: Supervised approach
5	02/24	Theory: Machine Learning for Big Data: Supervised approach
5	02/26	Lab: Machine Learning for Big Data: Supervised approach
6	03/03	Lab: Data Preprocessing
6	03/05	Midterm 1 – Week
7	03/10	Midterm 1 – Week
7	03/12	Theory: Machine Learning for Big Data: Unsupervised approach
8	03/17	Lab: Machine Learning for Big Data: Unsupervised approach
8	03/19	Theory: Machine Learning for Big Data: Unsupervised approach
9	03/24	Lab: Machine Learning for Big Data: Unsupervised approach
9	03/26	High-class Imbalance in Big Data: Tools and techniques

10	03/31	SPRING BREAK
10	04/02	SPRING BREAK
11	04/07	High-class Imbalance in Big Data: Tools and techniques
11	04/09	High-class Imbalance in Big Data: Tools and techniques
12	04/14	High-class Imbalance in Big Data: Tools and techniques
12	04/16	Midterm 2 – Week
13	04/21	Midterm 2 – Week
13	04/23	TBD
14	04/28	TBD
14	04/30	Project Presentation
15	05/05	Project Presentation
16	05/07	Project Presentation
16	05/12	Project Presentation