San José State University
Computer Science Department

CS185C: Command line and systems programming for data-intensive applications, Fall 2021

Course and Contact Information

Instructor:	William “Bill” Andreopoulos
Office Location:	Online (former MacQuarrie Hall 416)
Email:	william.andreopoulos@sjsu.edu Please use Canvas Messaging and the Discussion Forum
Class Days/Time:	Tuesday and Thursday 17:45-19:00
Classroom and Office Hours:	Online via Zoom

Faculty Web Page and Canvas Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on Canvas Learning Management System course login website at http://sjsu.instructure.com. You are responsible for regularly checking with the Canvas messaging system to learn of any updates. You should modify the Canvas settings for notifications of announcements and discussion forum postings to be sent to you.

Course Description

An in-depth study of essential skills for mastering the UNIX operating system, including processing data using grep, sed, awk, join. We will cover programming advanced shell scripts for manipulating data and performing system administration tasks.

Prerequisites

This course is offered to MS Bioinformatics or upper-level undergraduate CS students at San Jose State University. MS-BI students must have completed BIOL 123B with a grade of C- or better. CS students must have completed CS 146 with a grade of C- or better.

Course Learning Outcomes (CLO)

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

1. Use the UNIX/Linux command line to manipulate and process large-scale data.
2. Perform system administration tasks, modify user accounts and groups, control jobs and customize the environments in UNIX/Linux.
3. Develop shell scripts for system maintenance tasks and for use in data-intensive applications.

4. Build data pipelines with shell scripting and document them to make analyses reproducible and shareable.
5. Understand how data analysis on the command line compares with use of graphical user interface and web-based tools and compare the benefits and drawbacks of each approach.
6. Deal with challenging problems in data science that require use of the UNIX/Linux shell and command-line tools.
7. [for students who come from a biology background] Handle biological datasets; learn how to run bioinformatics analyses efficiently, how to document and reproduce analyses, how to use cloud computing for data-intensive problems.

Recommended Texts/Readings

Beginner: UNIX Command Line: A Complete Introduction. William Shotts Jr.
Moderate: Linux Command Line and Shell Scripting Bible. Blum and Bresnahan
Advanced: UNIX Power Tools. Jerry Peek, Tim O'Reilly, and Mike Loukides.

Other good readings:
Advanced Programming in the UNIX Environment. W. Richard Stevens, Stephen A. Rago. 3rd Edition, 2013, Addison-Wesley.

A copy of my slides will be available to the students enrolled in the class.

Additional handouts will be provided through Canvas.

Other technology requirements / equipment / material

Practice of command-line operations will be done on IBM's LinuxOne computing cloud. Instructions to subscribe for a free student account will be provided.

Course Requirements

SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class, participating in course activities, completing assignments, and so on.

Reading assignments: Readings will regularly be assigned for the next class (see schedule). Slides will be posted under the Canvas modules before the next class.

Hands-On Worksheets:
We will have a number of hands-on worksheets. A worksheet will be due weekly. Please refer to Canvas for detailed instructions and deadlines. You need to submit the worksheets by their closing time on the due date. There will be no makeup on worksheets. No worksheet will be re-opened after its closing date. As this is a fast-paced course, it is essential that you submit the worksheets in a timely fashion in order to keep up.

The purpose of the hands-on worksheets is to develop your understanding of the material and skills in using the command-line tools. Many of the hands-on worksheets

will involve data manipulation or use of bioinformatics tools. The hands-on worksheets will involve learning how to use command line tools for performing bioinformatics analysis and manipulating data. Students will use IBM's LinuxOne computing cloud for practice. We will take time at the beginning of each class to discuss any difficulties students have in completing the worksheets from previous classes.

Homework assignments: Assignments will be assigned for each module of the course. The assignments will be similar to worksheets. All assignments should be submitted on the corresponding assignment page in Canvas by 11:59 P.M. on the due date. The programming assignments cumulatively will be worth 40% of your grade.

More information will be given at the time of the first assignment. There will be a penalty for late submission 5% for every 3 days up to 15 days; after 15 days no submission will be accepted and the submission page will be closed. Never email your assignments, always upload to Canvas.

All assignment solutions that you submit must be completely your own work (i.e., your solution cannot be copied from another source, such as other students, the internet, etc.). While it is fine to discuss the worksheet/assignment solutions with other students, solutions submitted on Canvas should reflect your own efforts. Oral examination might be requested. All homework should be submitted on Canvas, not by e-mail.

iClicker participation during class: The iClicker questions are in the form of multiple choice and true-false questions. All students are expected to participate with iClicker. Credit is given for participation and it is not necessary to get the correct answer to get credit. Please install iClicker on your phone (app) or laptop (iclicker.com) following these instructions: http://www.sjsu.edu/ecampus/teaching-tools/iclicker/

Exams:
Midterm Exam One: Monday, October 5, 2021.
Midterm Exam Two: Monday, November 2, 2021.
Final Exam: Monday, December 14, 2021.
The midterm exams are each one hour and fifteen minutes long. The final exam is two hours and fifteen minutes long.

The exams will contain multiple choice questions, true/false and short answer questions. Exams are *open book, open notes*, and comprehensive. The exams should be done individually and are not group work. No make-up exams except in case of verifiable emergency circumstances.

Discussion Forum on Canvas

Students should use the Canvas Discussion Forum for all issues about the course. Regular participation is recommended. The instructor will open a graded thread for each module. Students must ask at least 1 original question and provide at least 1 original answer in each graded thread. Discussion Forum participation counts for 5% of the course grade.

Determination of Grades

The course grade is based on:

40% Five Assignments
14% Weekly Worksheets
1% iClicker participation
5% Discussion Forum Participation
20% Two midterms
20% Final

Grade	*Points*	*Percentage*
A plus	*960 to 1000*	*96 to 100%*
A	*930 to 959*	*93 to 95%*
A minus	*900 to 929*	*90 to 92%*
B plus	*860 to 899*	*86 to 89 %*
B	*830 to 859*	*83 to 85%*
B minus	*800 to 829*	*80 to 82%*
C plus	*760 to 799*	*76 to 79%*
C	*730 to 759*	*73 to 75%*
C minus	*700 to 729*	*70 to 72%*
D plus	*660 to 699*	*66 to 69%*
D	*630 to 659*	*63 to 65%*
D minus	*600 to 629*	*60 to 62%*

Communication with the instructor

Questions for the instructor may be asked during Zoom class meetings or office hours, or at any time via the Canvas Discussion Forum or Canvas messaging. Announcements of general interest will be posted under Announcements on Canvas. Questions about worksheet-specific code should be asked during class meeting time, not by email.

Class Attendance

Class attendance (via Zoom) is highly recommended. Classes will be recorded as Zoom screencasts and posted on Canvas. Students are responsible for all material presented in all classes.

Regrading Procedure

Grades assigned are final, unless there was an error in the grading. Students may request regrading by filling out a Regrade Request form on Canvas. Regrading may result in a lower grade.

Classroom Protocol

Students should be muted when not speaking, and must be dressed appropriately when their camera is on.

Add/Drop Policy

For those wishing to add this course, the deadline is January 26, 2021. The last day to drop a course without a “W” grade is February 8, 2021. To drop after this date, a Late Drop petition will be required. According to University and Department guidelines, dropping after February 8, 2021, requires a serious and compelling reason to drop a course. Grades alone do not constitute a reason to drop a course. Students who stop attending without officially dropping will be issued a “U” at the end of the semester, which is counted as an F in calculations of GPA.

Students are responsible for understanding the policies and procedures about add/drop, grade forgiveness, etc. Refer to the current semester's Catalog Policies section at http://info.sjsu.edu/static/catalog/policies.html. Add/drop deadlines can be found on the current academic year calendars document on the Academic Calendars webpage at http://www.sjsu.edu/provost/services/academic_calendars/. The Late Drop Policy is available at http://www.sjsu.edu/aars/policies/latedrops/policy/. Students should be aware of the current deadlines and penalties for dropping classes. Information about the latest changes and news is available at the Advising Hub at http://www.sjsu.edu/advising/.

Consent for Recording of Class and Public Sharing of Instructor Material

University Policy S12-7, http://www.sjsu.edu/senate/docs/S12-7.pdf, requires students to obtain instructor's permission to record the course. Common courtesy and professional behavior dictate that you notify someone when you are recording him/her. You must obtain the instructor's permission to make audio or video recordings in this class. Such permission allows the recordings to be used for your private, study purposes only. The recordings are the intellectual property of the instructor; you have not been given any rights to reproduce or distribute the material.

Course material developed by the instructor is the intellectual property of the instructor and cannot be shared publicly without his/her approval. You may not publicly share or upload instructor-generated material for this course such as exam questions, lecture notes, hands-on exercises or homework solutions without instructor consent.

Academic Integrity

Your commitment as a student to learning is evidenced by your enrollment at San Jose State University. The University Academic Integrity Policy S07-2 at http://www.sjsu.edu/senate/docs/S07-2.pdf requires you to be honest in all your academic course work. Faculty members are required to report all infractions to the office of Student Conduct and Ethical Development. The Student Conduct and Ethical Development website is available at http://www.sjsu.edu/studentconduct/. Instances of academic dishonesty will not be tolerated. Cheating on exams or plagiarism (presenting the work of another as your own, or the use of another person's ideas without giving proper credit) will result in a failing grade and sanctions by the University. For this class, all assignments are to be completed by the individual student unless otherwise specified. If you would like to include your assignment or any material you have submitted, or plan to submit for another class, please note that SJSU's Academic Integrity Policy S07-2 requires approval of instructors.

- Anyone caught cheating (including sharing answers with others during exams) in the class will receive a failing grade on the exam or assignment, in addition to other sanctions that are permitted by the University, including but not limited to the filing of a report with the Dean of Student Services and expulsion from the University.

Campus Policy in Compliance with the American Disabilities Act

If you need course adaptations or accommodations because of a disability, or if you need to make special arrangements in case the building must be evacuated, please make an appointment with me as soon as possible, or see me during office hours. Presidential Directive 97-03 at http://www.sjsu.edu/president/docs/directives/PD_1997-03.pdf requires that students with disabilities requesting accommodations must register with the Accessible Education Center (AEC) at http://www.sjsu.edu/aec to establish a record of their disability.

In 2013, the Disability Resource Center changed its name to be known as the Accessible Education Center, to incorporate a philosophy of accessible education for students with disabilities. The new name change reflects the broad scope of attention and support to SJSU students with disabilities and the University's continued advocacy and commitment to increasing accessibility and inclusivity on campus.

University Policies

Per University Policy S16-9, university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. will be available on Office of Graduate and Undergraduate Programs' Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/

CS185C: Command line and systems programming for data-intensive applications, Fall 2021

The schedule is subject to change with fair notice.

Course Schedule

Classes	Topic
08/24	Introduction to the Bash shell command line, passwords, permissions, ssh/sftp/scp with keys
08/31	Home directories, terminal setup and environment variables, shell prompt setup, pathnames, file system (dirs, links, move, copy)
09/07	Shell interpretation of user input, wildcards, aliases, symbolic and hard links, editing, pagers, which
09/14	Job control, finding files (-exec), dealing with many files, data pre-processing, wc, uniq, sort, top/htop
09/21	Saving and restoring work with screen and tmux, tar/zip, history, curl
09/28	**Midterm 1**
10/05	Pipes and pipeline concept for data analytics tasks, jobs vs. processes, input/output redirection, gnu parallel

10/12	Awk, sed, grep, join, diff, with bioinformatics/data analytics examples
10/19	Awk, sed, grep, join, cut, tr, regular expressions with bioinformatics/data analytics examples
10/26	**Midterm 2**
11/02	Shell scripting, quotas, disk space, counting inodes and files
11/09	Shell scripting, nslookup, traceroute, crontabs, sudo, task automation
11/16	Case studies of reproducible data processing with containers (Docker, Singularity)
11/23	Case studies of workflow tools (Snakemake, Airflow, Nextflow, Clara Parabricks, Luigi, WDL, CWL, Galaxy)
11/30	A comprehensive data pipeline workflow for RNA-Seq and amplicon analysis
12/07	Workflow managers in HPC clusters (Slurm, Torque) to process large amounts of data. Final exam review
12/14	**Final exam**