# San José State University
# Computer Science Department
# CS271: Topics in Machine Learning, Section 2, Spring 2023

## Course and Contact Information

Instructor(s): William "Bill" Andreopoulos
Office Location: Online (former MacQuarrie Hall 416)
Telephone: (408) 924 5085
Email: william.andreopoulos@sjsu.edu
Office Hours: Wednesday 1:30-3:00 pm in MQH416, Friday 2:00-3:30 pm Online via Zoom
Class Days/Time: Monday and Wednesday 18:00-19:15pm
Classroom: MQH 233

## Course Description

Variable topics in machine learning. Content may include hidden Markov models, principal component analysis, support vector machines, clustering, boosting, random forests, neural networks, and deep learning. Relevant applications will be covered.
This section will study recent advances in machine learning methods with applications to solving sequence analysis problems in molecular biology and natural language processing. The methods examined include word embeddings, vector space representations, language models, and deep learning architectures. A substantial course project is required.

## Course Format

This course adopts an in-person classroom delivery format.

## Faculty Web Page and MYSJSU 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 course messaging system to learn of any updates. You should modify the Canvas settings for notifications of announcements and Slack messages to be sent to you.

## Prerequisites

Graduate standing. Allowed Declared Major: MS in Computer Science, Bioinformatics, Data Science. BIOL 123B and MATH 162 (for bioinformatics majors), or CS157A. Or instructor consent.

## Course Learning Outcomes (CLO)

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

1. Use machine learning and deep learning in bioinformatics sequence analysis to answer biological questions and to generate biological hypotheses.
2. Comprehend the nature, scope and limits of using machine learning and deep learning in the field of bioinformatics.
3. Develop machine learning and deep learning solutions for sequence data.
4. Compare different machine learning algorithms and choose a solution based on suitability for a particular data set.
5. Compare biomolecular analysis with machine learning to analysis with classical bioinformatics tools.

6. Appreciate some of the most challenging problems in life sciences that use machine learning methods, possess insight into how to solve those problems.

### Texts/Readings

We don’t use a specific textbook in this class as there exists a lot of relevant material on bioinformatics found in various references. The reading material will be the slides, references and handouts.

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

Additional handouts will be provided through Canvas.

Major references:

- Data Analytics in Bioinformatics: A Machine Learning Perspective, 1st Edition (2021). by Rabinarayan Satpathy, Tanupriya Choudhury, Suneeta Satpathy, Sachi Nandan Mohanty, Xiaobo Zhang (Editors). ISBN-13: 978-1119785538.
- Haoyang Li, Shuye Tian, Yu Li, Qiming Fang, Renbo Tan, Yijie Pan, Chao Huang, Ying Xu, Xin Gao. Modern deep learning in bioinformatics. Journal of Molecular Cell Biology, Volume 12, Issue 11, November 2020, Pages 823–7.
- Deep learning in bioinformatics. Edited by Xin Gao, Wei Wang. Elsevier Methods. Volume 166, 15 August 2019, Pages 1-120.
- Walsh, Ian; Pollastri, Gianluca; Tosatto, Silvio C. E. (September 2016). “Correct machine learning on protein sequences: a peer-reviewing perspective”. *Briefings in Bioinformatics*. 17(5): 831–840.
- Chicco, D (December 2017). "Ten quick tips for machine learning in computational biology". *BioData Mining*. 10 (35): 35.
- Yang, Yuedong; Gao, Jianzhao; Wang, Jihua; Heffernan, Rhys; Hanson, Jack; Paliwal, Kuldip; Zhou, Yaoqi (May 2018). "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?". *Briefings in Bioinformatics*. 19 (3): 482–494.
- Wang, Sheng; Peng, Jian; Ma, Jianzhu; Xu, Jinbo (January 2016). "Protein secondary structure prediction using deep convolutional neural fields". *Scientific Reports*. 6: 18962.

#### Other technology requirements / equipment / material

Students will use colab.research.google.com and create Jupyter notebooks in Python to ensure their work is shareable and reproducible.

### Course Requirements and Assignments

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. The hands-on worksheets will involve use of bioinformatics tools. The purpose of the hands-on exercises is to develop your understanding of the material and skills in using the tools.

The Hands-On worksheets will involve learning how to use machine learning and deep learning tools with the Python programming language for performing bioinformatics analysis. Students will use colab.research.google.com and create Jupyter notebooks in Python to ensure their work is shareable and reproducible.

**Term Project and In-Class Presentation:** There will be a term project. It is a group project. Each group consists of two students. A list of possible projects will be provided to you by the instructor.
Team Formation is due on Monday, February 6, 2023.
A Progress Report is due on Monday, April 3, 2023 (after Spring Recess).
The final project is due on Monday, May 15, 2023.
The in-class presentation will also take place on May 8-15, 2023.
A grading rubric will be provided.
All homework should be submitted on Canvas, not by e-mail.

**Examinations**

Midterm Exam One: Wednesday, March 8, 2023.
Midterm Exam Two: Wednesday, April 19, 2023.
Final Exam: Wednesday, May 17, 2023, 6-8:15pm.
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.

**Presentation of a research paper:** Each student should present an influential research paper of his/her choice, which is related to their project topic, to one of the classes. Students should sign up in the given spreadsheet for a date to present a paper. The paper, chosen by the student, should either use machine learning/deep learning towards making a biological discovery or introduce a novel tool for Natural Language Processing or text mining or bioinformatics. The presentation should last for no more than 10 minutes followed by Q&A. A grading rubric will be provided.

**Participation during class via Zoom:** The polling questions are in the form of multiple-choice and true-false questions. All students are expected to participate with Zoom polling. Credit is given based on participation and it is not necessary to get the correct answer in polls to get credit. Please contact eCampus at ecampus@sjsu.edu with any questions or issues with the Zoom technology.

## Grading Information

The course grade is based on:

| | |
|---|---|
| Hands-On Worksheets: | 19% |
| Midterms: | 20% |
| Final: | 20% |
| Project: | 30% |
| Zoom participation: | 1% |
| Presentation of a research paper: | 10% |

| *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

Students should follow the correct channels for communication. Questions should preferably be done during the regular class meeting time via Zoom or office hours. For course-related electronic communication students should use the Discord channel:

1) We will be using the course Discord channel for class discussion. The system is catered to getting you help efficiently from classmates, the TA, embedded tutor, and the instructor. Rather than emailing redundant questions to the teaching staff, students should post questions on the Discord channel where the entire class can read and benefit from the responses. The professor may re-post questions that are of general interest to the general channel or discuss them in class. The professor may ask students to reveal their real name if they are making special requests on Discord (e.g. deadline extensions) to prevent abuse.

2) Students are invited to join the office hours.

*Private messages sent to the instructor's other email addresses get lost due to the large volume of emails received.*

The instructor does not write messages after normal business hours, on weekends or holidays.

Reviewing code for the homework and technical trouble-shooting should be done during the office hours.

Never email your entire code for an assignment to the instructor. The instructor will not fix all the bugs in your code. Limit the code you post to 20 lines or less.

Announcements that concern everyone, such as reminders about due dates or class policy, will be posted.

### Class Attendance

Attendance (in-person or 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. If a student wants to request a regrade of a homework or test, please follow instructions on the "Regrade request" page on Canvas. A request for a regrade is not a technique to drum up a few more points. If the course instructor thinks a component was scored too generously the first time, it may be lowered in a regrade. Thus, regrading may result in a lower grade.

### Classroom Protocol

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

Course material developed by the instructor is the intellectual property of the instructor. Students can 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 permission.

### Add/Drop Policy

For those wishing to add this course, the deadline is February 20, 2023. The last day to drop a course without a "W" grade is February 20, 2023. To drop after this date, a Late Drop petition will be required. According to University and Department guidelines, dropping after February 20, 2023, 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/

# CS271: Topics in Sequence-based Machine Learning for Bioinformatics, Spring 2023

The schedule is subject to change with fair notice provided via Canvas announcements.

**Course Schedule**

| Week | Topic |
|---|---|
| **01/25-01/27** | Introduction, overview of unsupervised and supervised ML in bioinformatics |
| **01/30-02/03** | Essentials of machine learning in bioinformatics, NLP and text mining |
| **02/06-02/10** | Sequence classification with Linear and Logistic Regression |
| **02/13-02/17** | Language models using k-mers and word embeddings |
| **02/20-02/24** | Vector space representations: clustering & visualization with PCA, t-SNE, UMAP |
| **02/27-03/03** | Hidden Markov Models and Markov chains |
| **03/06-03/10** | Review for midterm with problem-solving exercises / *Midterm 1* |
| **03/13-03/17** | Sequence classification with Naïve Bayes |
| **03/20-03/24** | Deep Learning introduction, fundamentals and architectures |
| **03/27-03/31** | *Spring recess* |
| **04/03-04/07** | Deep Learning in bioinformatics: CNNs, LSTMs, Recurrent Neural Networks (RNNs), Long Short Term Memory (LSTM) neural networks for sequence modelling |
| **04/10-04/14** | Word embeddings and language models with neural networks, transformers, BERT, transfer learning |
| **04/17-04/21** | Review for midterm with problem-solving exercises / *Midterm 2* |
| **04/24-04/28** | Efficient sequence searching, min-hashing, locality-sensitive hashing, vector quantization |
| **05/01-05/05** | Case studies using deep learning in bioinformatics / Project discussion |
| **05/08-05/12** | Project presentations |
| **05/15-05/17** | Review, wrap-up. Final exam on Wednesday, May 17, 2023, 6-8:15pm |