Course Syllabus

Jump to Today ✎ Edit

Basic Information

Instructor

Name: Frank Elavsky, PhD

Email: felavsky@calpoly.edu (mailto:felavsky@calpoly.edu)

Office: 14-228

Office hours

I highly encourage office hours for anyone, any reason. I will take students in the order they arrive and try to give everyone at least 10 to 20 minutes of my time.

Wednesdays: 2-3pm

Thursdays: 3-4pm (after class)

You can always email me (felavsky@calpoly.edu (mailto:felavsky@calpoly.edu)) or post in slack if you have a question or would like to schedule a specific time to chat. When emailing or messaging me, please explain what you would like to chat about.

A little bit of history

DATA 3301 is a result of merging material from two quarter courses: DATA 301: Introduction to Data Science and CSC 466: Knowledge Discovery in Data. These courses appeared in Cal Poly calendar at different times and while both became constituent parts of the Cross-Disciplinary Studies Minor in Data Science, they could be taken in any order. This created multiple points of entry into the disciplines of Data Science and Machine Learning on quarters, which created some challenges downstream, when students were taking some of the more in-depth machine learning coursework.

Merging the two courses enables us to create a single entry point into the rigorous study of Data Science and Machine Learning and guarantee for the upper-division Machine Learning and Data Science courses that all students who take them have exactly the same background. While the merger of the two 10-week courses into a single 15-week course comes with loss of some of the material from CSC 466,

on quarters CSC 466 and DATA 301 shared several topics. The created some efficiency, the result of which is that all content relevant for all downstream courses is preserved in DATA 3301.

Disclaimers and Caveats

The prerequisites for this course are:

- Linear Algebra: MATH 206, MATH 1151 on semesters. We accept MATH 244 as a prerequisite as well, but this requires a permission code.
- Intro to Computing: CSC 101 or CSC 1001 on semesters.
- Intro to Statistics: on quarters - one of STAT 217, STAT 218, STAT 301, or STAT 312. On semesters, these collapse into one of the following: STAT 1110, STAT 1510 (for STAT/MATH and future Data Science majors), or STAT 3210 (for CS/SE majors).

While there are many Data Science programs where the first Data Science course has no prerequisites, we, at Cal Poly felt that putting a fairly strong list of prerequisites enables us to provide more in-depth coverage of the material and prepare everyone interested in Data Science for the courses that follow.

This works well on two levels:

- For most of you this is the first but not last Data Science/Machine Learning course. This enables us to make the follow-up courses in-depth studies of modern machine learning methods.
- For some of you (e.g., if you are pursuing Cross-Disciplinary Studies Minor in Bioinformatics), this course enables you, upon completion, to perform a wide range of data analytical tasks within the confines of your field of study.

This is NOT a database course. CSC 3665, Introduction to Database Systems covers relational database model and SQL. We will discuss work with data in a way that mimics how queries are written for relational databases, but DATA 3301 is not a replacement for a traditional Intro to Databases course.

This is NOT a ”Big Data” course in a sense that we won’t be pushing compute resources to and past their absolute limits to analyze datasets of extraordinary sizes. A lot of the data we use in the course comes from relatively small datasets that are easy to ingest, manipulate and analyze, but that are rich enough for interesting analysis. Work with terabytes of data is reserved for some of our downstream courses such as DATA 4401, DATA 4610, DATA 4620, CSC 4667, and CSC 4888.

This is the first time we are teaching DATA 3301. This semester there are two sections of the course taught by two different instructors. While we strive to present the same material, DATA 3301 course instructors have some leeway in terms of topic choice, topic order, selection of assignments and assessments and other aspects of the course. If you compare notes with a peer who is taking the other section, you might find some material presented differently and at different points in the course. This is ok. We guarantee that any version of DATA 3301 will serve the two core purposes: making all of you comfortable working on data analytical/data science tasks, and getting most of you ready for downstream Machine Learning/Data Science courses.

Diversity, Equity, and Inclusivity

I strongly believe that access to, and support within, computing education is a matter of social justice. Computing touches nearly every facet of most lives in our society and in many of those around the world. Computing should not be practiced by only a small, mostly homongeneous group and should not be defined by that same group. The field of computing is better when it includes and actively supports a diversity of practitioners because this diversity of experiences and of cultural and personal values provides greater perspective on the problems that should be solved to better humanity.

Learning Environment

I believe in a supportive learning environment wherein every person deserves respect, every question deserves an answer, and we all work together toward a goal of improved understanding and life-long learning. All members of this class are expected to contribute to a respectful, welcoming, and inclusive environment for every other member of the class.

There is an old adage that encourages those with questions to ask because others may have the same question; you do not need some imagined quorum in order to validate your question or your request for assistance. If you have a question or need assistance, that alone is sufficient. You belong here and your participation in the course is not only welcomed but encouraged.

And failure is an inherent part of learning and growing. This means that the course materials will be difficult! And for some students who have mostly learned in environments where they excelled, it can be intimidating, disorienting, and isolating facing challenges. But engaging actively with your peers, asking questions, and coming to me outside of class, these are the basic ingredients for overcoming difficulty.

Statistically speaking, students who have friends in a class have a better chance of completing a course and higher grades: so be friendly and respectful to others. Make friends! And students who feel ahead: this is a great chance to further connect with peers. I encourage peer-group office hours, where you collectively come to me with similar issues and we work through materials together.

Personal Well-being

There are times when coursework and non-academic obligations may conflate to create a seemingly unbearable situation. During these times, your coursework may suffer and you may opt to not complete an assignment by the posted date; I understand this and you need not apologize for it. If this happens, do not get discouraged. Contact me as soon as you are able to. I am here to assist you; just ask for help.

If you or someone you know is having difficulty accessing enough healthy food or stable housing, please know that our campus offers a number of free and effective resources.
Visit **www.basicneeds.calpoly.edu (http://www.basicneeds.calpoly.edu/)** for more information.

Accessibility

My research area is accessibility and I work extensively with people with disabilities. I am also open about living with disabilities myself. But mentioning my experience means a couple things: first, that I am more than happy to accommodate you. And by asking me to, you can help improve my teaching style, approach, pacing, or materials in any way. You are welcome to reach out to me directly, in an email or visit during office hours, and let me know what you need. However, it is heavily in your favor to *also* work with the **Disability Resource Center (DRC) (https://drc.calpoly.edu/)** and get formal accommodations whenever possible. *I'm* happy to work with you. But will *every* professor be? It can require a bit of executive function if you aren't used to applying for and requesting accommodations, but it is worth it to at least get set up in case you need it. The DRC exists to help you get what you need and to require professors to do the work that they should be doing. If you want help with any of these things too, including navigating higher education with a disability, let me know. I was a Section 504 kid in high school and have had to go through the hoops to get accommodations in my undergrad, grad school, and places of employment too. I want to help you succeed and feel fully included in your educational journey here at Cal Poly.

Cal Poly Diversity Learning Objectives

The following are Cal Poly's **Diversity Learning Objectives (https://academicprograms.calpoly.edu/content/academicpolicies/diversity_lo)**. Though we may not always directly engage with these in this course, we will, at times, certainly indirectly attend to some of them, with particular emphasis on #5.

All Cal Poly graduates should be able to:

1. Recognize and understand the contributions to knowledge and civilization that have been made by members of diverse cultural and gender groups and other historically marginalized people in the United States and across the world;
2. Understand the history of issues related to diversity, social and economic inequities, and political power in the United States and across the world;
3. Analyze the current social, political, artistic, and/or economic lives of historically marginalized people in the United States and across the world;

4. Analyze the various institutions and structures that create and maintain social, economic, and political inequality in the United States and across the world; and, identify those that offer redress for these issues;
5. Define and describe the various issues related to diversity, equity, and inclusion in their respective disciplines;
6. Critically examine their own personal beliefs, attitudes, and biases about historically marginalized people and cultures in the United States and across the world.

Texts and References

This course does not have an official textbook. At the same time, we will actively use materials from the textbook prepared by Dr. Dennis Sun for DATA 301. The textbook is prepared in a form of a series of Jupyter notebooks that discuss a series of core concepts in the area of data science. It is available from **Dr. Sun's Github page. (https://github.com/dlsun/pods)**

The actual lectures, labs, and notebooks for this course are based mostly on Dr. Sun's textbook, but diverge in places. This course was assembled by Professor Alex Dekhtyar before passing it on to me.

CSC 466 used as its textbook the following book:

- Bing Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, Springer, 2st ed. 2011. ISBN: 978-3642194597.

This book describes many of the topics covered in DATA 3301 is a very accessible, *algorithms-centric* way.

An additional book I can recommend is:

- Mohammed J. Zaki, Wagner Meira Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, 2014, ISBN: 978-0-521-76633-3.

This book is more *optimization-centric* than the Bing Liu book, and is an appropriate textbook for DATA 4610, but having it at hand might be useful.

Activities

Participation (25%)

Participation is the largest part of your grade.

A. Class Attendance and Interaction (5%)

The lectures are for your benefit. You should ask questions when you have them. Use lecture time to discuss general approaches to any project, to clarify the material, and to make sure that you are prepared for success.

B. In-Lecture and In-Lab Reflections (15%)

Lectures and Labs may start or end with a break for writing reflections, by hand, on paper. Reflections are a significant portion of your grade overall (15%) and will sometimes involve group reflections, prompts to answer, or will be unstructured. These are intended to get you to think about activities you have just done (in lab, in a previous assignment, or during a lecture). Expect about 15 of these across the term, each worth up to 1% of your grade.

C. In-Lecture Exercises (5%)

Lectures will include a variety of in-class exercises and example work, with your input, on the board or on computer. Your participation in the live discussion is welcome and encouraged.

Labs (20%)

There will be a lab activities that are in-class and lab assignments that are take-home. There are about 8 take-home lab assignments, each will typically be assigned at the end of lecture on Thursday and due the following Thursday. Both activities and assignments are designed to develop and apply the skills discussed in lectures. Many of these labs will directly support the course project. You ***can*** discuss and work on these lab activities+assignments with your classmates but you will be graded independently.

A. Activities

Many sessions will involve coding during class and often have a reflection accompanying them at the end.

B. Assignments

There will be some take-home lab assignments in this course, which will also have reflections accompanying them (or have a quiz immediately following the assignment deadline).

Project Assignments (15%)

This course has one major project, broken into three progressive assignments. This project intended to be of a significantly greater length and complexity than the lab assignments and is required to be done as a group activity. You cannot propose a project as an individual, so be prepared to join a group. As part of each assignment you will also be asked about your assignment work in the reflection activities in

class, as part of your "participation" grade (so, in a way, your work on your project will influence your participation grade, in addition to the final exam).

A. Proposal (2%)

You must propose your project idea with your group. You will have a short prompt to help guide you in crafting a 2 page proposal. Hand-drawn and low-fidelity sketches, figures, screenshots, and other supplemental figures are highly recommended but do not count towards your 2 page proposal. Before you begin any work, you will receive feedback on the proposal.

Your proposal will be graded based on how well you communicate your ideas, expected challenges, and application of course topics.

B. Plan and Preliminary (3%)

After getting feedback on your proposal, you will need to develop a plan for your group that divides the labor of the project (in whichever way you suggest). You must also perform preliminary work (gathering your dataset(s) and performing descriptive statistics). If your working conjectures require revisiting, you must do so as part of the preliminary report. At this point, you must also schedule a time during finals week to demo your project as a group to me, according to options on a schedule I provide you. After this assignment deadline, I will prioritize group schedules based on the earliest time that this assignment was submitted to me successfully and completely. I will try to accommodate everyone, but see the late/missed work policy below for more details on the final.

C. Completed Project (10%)

At the end of the term, you are expected to submit a completed project by the project deadline.

In-Class Quizzes (20%)

There will be in-class quizzes. The quizzes will cover concepts presented in lecture and material based on the lab assignments. These will be noted in the schedule below, so you can prepare in advance (note that this is subject to small adjustments).

Approx. Schedule of Quizzes

Quiz	Topic	Tentative Week	%
1	Tabular data	2	2
2	Categorial and numerical data	3	2
3	Visualization	4	2
4	Distance/similarity metrics	5	2
5	Regression	7	2

6	Evaluation of machine learning	8	3
7	Classification I	10	2
8	Classification II	12	2
9	Clustering	14	3

Final Exam (20%)

There a single final, in-person oral exam, which is a demo of your group's final project to me. These demos are not in-class presentations but scheduled with me in a 20 minute window during finals week. You will be graded not only on the quality of your demo to me but also in how well each group member handles my questions to you and your group.

The success of your final exam is something that you and your group prepare in advance: you set the content that you want to "focus" on and thus be graded for. What this means is, the more content chosen from the course materials that you choose to apply in your project and demonstrate to me in your final, the better your grade. For example, if you choose data visualization as a focus area, you will need to adequately demonstrate understanding of that topic in your demo by articulating which parts of visualization you considered and how you used principles, tools, and practices effectively.

Grade Breakdown

Activity	**% total**
Participation	25
Labs	20
Project	15
Quizzes	20
Final	20
Total	100

This course has relatively hard problems. Because of this, the traditional 90 = A, 80 = B, 70 = C grading schema does not work. Historically in this course, the A/B cutoff has been around 80-85%, while the B/C cutoff has been around 67-70%.

Course Policies

The following course policies have been established to support your success in this course and are shared here for easy reference.

Communication Policy

My primary role in this course is to support your learning. There are multiple communication methods for this course. Please take advantage of these methods to reach out for assistance so that I can help you.

- **Office Hours** — Office hours are an opportunity for one-on-one or group discussions. These discussions may be about the course material or assignments, about your career plans, about academic requirements, or many other topics. Feel free to stop by if you have questions or wish to talk.
- **Slack** — I maintain and regularly monitor a Slack workspace for questions about the assignments and the course material. General programming and concept questions should be made public so that all students can benefit from and contribute to responses. Questions with code specific to a solution should be asked via direct messages. Treat each other with kindness; I will not tolerate inappropriate posts.
- **Email** — You can also email (felavsky@calpoly.edu) me with any questions, concerns, or personal issues that you may have. Though Slack provides the benefit for shared discussion of course material and is generally preferred for that reason, I respect that email is a more appropriate method for some communications.

You can contact me any day of the week and at any time. I will generally respond to each email and Slack message within 24 hours during weekdays or 48 hours on weekends, but often much more quickly. Please do not hesitate to contact me, but give adequate time for a response. Even though I will likely respond very quickly, you should not rely on my ability to do so very close to a deadline. I do not actively check work-related messages at late or odd hours and cannot promise that I will be readily available in the final minutes or hours before a deadline.

Late Work Policy

There are times when coursework and non-academic obligations conflate to create a seemingly unbearable situation; I understand this and you need not apologize for it. Some of the course assessments are structured to automatically account for the possibilities of such circumstances.

- **Quizzes** — Quizzes happen in class, on paper, and cannot be made up (at least in the format that was given in class). However: I don't want anyone coming to class if you are sick. Please stay home! For quizzes, you will be able to anticipate ahead of time when a quiz is taking place (and you will be

notified of any changes in the schedule). This can hopefully help you to plan. However, you are able to **drop at least 1 missed quiz** at the end of the term. Note that you cannot drop a quiz with a low score (the drop policy only applies to missed quizzes). This encourages you to feel okay missing class if you get sick.

- **Labs** — These are formative activities to develop the skills that you will use in the project and that you will demonstrate on the quizzes (skills that will benefit you going forward). You will be required to hand-write reflections of your lab work at periodic in-person occasions according to specific prompts that I provide, which heavily determine your grade for the lab portion of the class as well as participation. These can be made up during later lab times if you miss them, but must still be hand-written reflections.
- **Assignments** — (labs + project) Each lab or project assignment deadline is intended to coincide with the beginning of the time frame for the next assignment. It is ideal that lab assignments, in particular, do not overlap. As such, you will want to pace your work to meet these deadlines. Start on the assignments early and work on them periodically over the assignment time frame. Each assignment submission needs to be pushed to your repository by the corresponding deadline. I *will* be checking your commit history and times, so don't think that on-time submissions (with late code commits) is a strategy that will work. And in terms of late policies: you are able to submit assignments past the deadline, but there is a 5% late penalty (applied immediately) plus a 5% penalty per additional day (which maxes at 50%). You're always encouraged to turn in assignments, even if late, but I *do* have an eventual cutoff in order to get grades in. There will come a point in the term, generally during finals, where I simply cannot be as flexible with late attendance.
- **Participation —** Unfortunately, most participation cannot be made up. If you miss some classes, you miss out on participation! That being said, the reflections (which are the main part of the grade) can be done in the first class you attend after an absence, just reach out to me if you need this.
- **Final Exam —** The final exam will be scheduled during the week of finals, based on your group's availability. You do not all need to be present! However, **at least 2 group members must be present** and you must discuss ahead of time as a group and agree that some group members can miss the final demo. Members who do not attend in person will be required to submit a hand-written reflection to me. If only one person attends the demo, they will be graded based on the demo, but the rest of the group will fail. If nobody attends the demo, the whole group will fail the final exam.

All of this said, life often necessitates exceptions. Few could have predicted the extent of the exceptions needed at the end of Winter 2020, throughout Spring 2020, and again at the end of Spring 2020. Contact me for any reason to discuss needs particular to your situation.

Collaboration Policy

Talk, share ideas, help each other. But be sure that you are learning and applying the materials from class.

Group activities: not all graded as a group

- Labs are our formative activities ("learn by doing"). And people often learn better by doing things collaboratively. You are allowed and encouraged to work together on lab activities and assignments. However, while you can work collaboratively, you will be evaluated as an individual.
- The project is a group activity and is evaluated as a group activity. You will share a grade.
- The final exam/demo of your project is a group activity but will also include individual evaluations, based on how each group member demonstrates their understanding of the work that was done overall as well as the work they contributed.

Fully individual assessments

- Reflections are individual efforts. For this reason, make sure that you are demonstrating and submitting your own work and that you are learning from the experiences you have, even if they are done as a group.
- Quizzes are assessments of your understanding of the course material and must be the work of only the individual.
- Again, lab assignments are individually graded. The code that you submit must be your own. See the course's AI policy for more information as well.

Further Details

It is fine to talk with others about general approaches used to solve the lab assignments, but each student is to develop their own solution. Students are not to exchange program code in any form (hardcopy or electronically). Using solutions from any other source is forbidden (github is not a resource for studying); in particular, using solutions (either instructor’s or other students’) from previous offerings of this course is not allowed. Using solutions found on the Internet is not allowed. Referring to previous solutions while developing your solution is not allowed. Hiring a "tutor" or another service to implement the assignments for you is not allowed. Leveraging generative AI to write code for you, even in part, is not allowed. Collaboration that goes beyond a high-level discussion of general approaches or pseudo-code will be considered cheating. If you are unsure about what constitutes proper or improper collaboration, consult the instructor for guidance.

What is allowed when collaborating?

- Help with proper syntax and run-time errors.
- Help finding (but not fixing) a bug, but only after a reasonable amount of time spent testing and debugging. Do not accept help until you have tried to find the bug yourself; this is part of learning.
- Discussion of methods for debugging.
- Discussion of general programming techniques.
- Discussion of algorithms and data structures required by the assignment description.
- Discussion of ethics, bigger picture issues, and contextual factors.

Academic Integrity

AI-generated Content Policy

First and foremost, my approach as an instructor is informed by *harm-reduction*. This is a strategy taken by activists and health services/officials during the HIV/AIDS epidemic: telling people that using needles is morally bad/harmful resulted in more shame, and thus more secrecy. And when people use dangerous things in secret, it is far harder to get help when you need it. So in order to be a true instructor, who helps you learn, I adopt a shame-free approach. If you've been using generative AI in my class or before my class, no big deal. Let me know. And if you need help because you recognize you've fallen behind because of using generative AI, then I *especially* want you to let me know.

Students who do better in today's educational landscape, as well as do better in professional environments, are ones who are deeply literate in what AI is, how it works, the shortcomings of it as a tool, and who also possess a deep degree of humility and expertise of their own. The trouble as an educator is that I need to simultaneously inform you all of the downsides of AI for learning while also not actively scaring you away from speaking to me when you need help. My goal is to help you gain that expertise that separates you from someone who just leverages Generative AI without understanding a domain of work. My intention is for my policy to serve that goal:

Use of Generative AI is **strongly *not* encouraged** but *also not forbidden* on any take-home assignments. Therefore, the condition is that all use of Generative AI on take-home assignments must be explicitly disclosed, both in the text of any reports you submit, as well as by placing comments in any code cells/code blocks that were produced with assistance of Generative AI. Failure to do so constitutes violation of the Academic Integrity Policies outlined below.

In this course, the following types of assignments are classified as take-home and use of Generative AI is **permitted**:

- Lab assignments (does not apply to in-class activities).
- Project-related assignments.

This means that only 25% of your grade could be determined by generative AI. However, there is strong research that shows that high performance coupled with short time spent on homework strongly correlates to poor exam and quiz scores. GenAI helps you turn in really good looking homework in very little time spent. Turns out, those markers correlate the most strongly with poor exam scores and individual assessments. I want you to succeed as a student and a human being in this world, so it is important that you understand that generative AI, while technically allowed on take-home work, will likely work against your grade overall.

Use of Generative AI is **prohibited** on any in-class assignments/activities (and using it would be too awkward for most of these anyway, given the state of existing technology):

- Participation (including reflections on your take-home project and lab work)
- In-lab notebook + activities
- Quizzes (which are pencil-and-paper)
- Final (which is primarily a demo and group oral exam)

The quizzes and the final are the true evaluations in this class, as they will demonstrate both your thorough understanding of the course materials as well as your ability to articulate your course-relevant contributions to the project in specific and concrete terms. All content in this course is intended as a "study guide" in service of quizzes and the final.

Real life will require you to have a deep understanding of materials. The idea of a 10x or 100x experience when using generative AI means little if your starting value is 0. Zero times 100 is still 0. You must have fundamentals, ideally even deep specialization, in order to navigate the landscape of today's technical work in conjunction with (and without) generative tooling. Your expertise will compound and magnify your use of tools. But this means that you must gain expertise *first*.

University Policies

Cal Poly's **Academic Integrity policies are found at this link (https://academicprograms.calpoly.edu/content/academicpolicies/Cheating)** . **(http://www.academicprograms.calpoly.edu/academicpolicies/Cheating.htm)**

In particular, these policies define cheating as (684.1) "…obtaining or attempting to obtain, or aiding another to obtain credit for work, or any improvement in evaluation of performance, by any dishonest or deceptive means. Cheating includes, but is not limited to: lying; copying from another's test or examination; discussion of answers or questions on an examination or test, unless such discussion is specifically authorized by the instructor; taking or receiving copies of an exam without the permission of the instructor; using or displaying notes, "cheat sheets," or other information devices inappropriate to the prescribed test conditions; allowing someone other than the officially enrolled student to represent same."

Plagiarism, per University policies is defined as (684.3) "... the act of using the ideas or work of another person or persons as if they were one's own without giving proper credit to the source. Such an act is not plagiarism if it is ascertained that the ideas were arrived through independent reasoning or logic or where the thought or idea is common knowledge. Acknowledgement of an original author or source must be made through appropriate references; i.e., quotation marks, footnotes, or commentary."

University policies state (684.2): "Cheating requires an "F" course grade and further attendance in the course is prohibited." (appeal process is also outlined, see the web site above for details.). Plagiarism, per university policies (684.4) can be treated as a form of cheating, although a level of discretion is given to the instructor, allowing the instructor to determine the causes of plagiarism and effect other means of remedy. It is the obligation of the instructor to inform the student that a penalty is being assessed in such cases.

Course Summary:

Date	Details	Due