Stanford University
Summer 2025
DATASCI 112: Principles of Data Science
Syllabus

Instructor: Alex Dekhtyar
Email: *dekhtyar@calpoly.edu*
Office: Data Science

Teaching Assistants (TAs)	**Course Assistants (CAs)**
Jack Krew (jkrew@stanford.edu) - Head TA Allison Xu (anxu@stanford.edu) - Head TA Reese Mae Feldmeier (rfeld@stanford.edu) Disha Gandwani (disha123@stanford.edu) Rahul Kanekar (rkanekar@stanford.edu) Sarah Zhao (smxzhao@stanford.edu)	Ryder Fried (ryfried@stanford.edu) Luke Duthie (lduthie@stanford.edu) Mark Leschinsky (mark28@stanford.edu) Sean Nesamoney (nsean23@stanford.edu)

Class Schedule:

Day	Class	Section	Time	Instructor	Location
Monday	lecture	ALL	10:30 - 11:45	Alex	200-002
Tuesday	sections	DIS 02 DIS 03 DIS 04 DIS 05 DIS 06	4:30 - 5:20 4:30 - 5:20 10:30 - 11:20 10:30 - 11:20 9:30 - 10:20	Allison,Luke Reese Mae, Sean Disha, Jack Mark, Rahul Sarah, Ryder	160-123 ART 360 ART 360 STLC 118 160-123
Wednesday	lecture	ALL	10:30 - 11:45	Alex	200-002
Thursday	sections	DIS 02 DIS 03 DIS 04 DIS 05 DIS 06	4:30 - 5:20 4:30 - 5:20 10:30 - 11:20 10:30 - 11:20 9:30 - 10:20	Luke, Allison Sean, Reese Mae Jack, Disha Rahul, (Mark) Ryder, Sarah	160-123 ART 360 ART 360 STLC 118 160-123
Friday	lecture	ALL	10:30 - 11:45	Alex	200-002

Office Hours Schedule:

Day	Time	Who	Where
Monday	12:00-1:00	Alex	TBA
Wednesday	12:00-1:00	Alex	TBA
TBA	TBA	TAs/CAs	TBA

Note: Each TA/CA for the class will schedule one hour of office hours weekly. The full schedule (days, times, room assignments) will be finalized mid-Week 1 of class and when ready, shared with you on the course web page and Slack.

Overview

Data Science is a multidisciplinary field of study covering a large variety of topics related to acquisition, maintenance, querying, analyzing and visualizing the data. This course serves as a gentle introduction into the field of Data Science both for those of you who want to pursue an in-depth study of Data Science and related topics (Machine Learning, Data Management, Statistics), as well as those of you who want to better understand what it means to work with data, and apply these skills to your own disciplines and areas of study.

Learning Objectives. This course has the following learning objectives. After completion of the course you will be able to

- Acquire and process tabular, textual, or hierarchical data.
- Uncover patterns by summarizing and visualizing data.
- Apply machine learning to answer real-world prediction problems.

Communication

Course Web Page. The course web page is on Canvas. The course web page will contain all materials distributed in class either in hardcopy or electronically, including all lecture notes, assignments and Jupyter/Colab notebooks.

Course web page: **https://canvas.stanford.edu/courses/210872/**

Course Slack. You all should have received an invitation to the course Slack workspace. Besides email communication, the course Slack workspace is the main means of communication between the instructor/TAs/CAs and you, ***in both directions.*** The Slack workspace has several thematic channels (#general, #assignments, #office-hours, #announcements, etc.) accessible to all of you, as well as channels for individual discussion sections which are primary conduits of information between you and the TAs teaching your recitations.

For questions regarding the course content, please use public Slack channels set up for the specific topics your questions are about. This way your questions will be viewed by everyone on the course team (myself, TAs, CAs). From 8am to 9pm on weekdays, at least one TA/CA is going to be monitoring the Slack channels for questions, and they will respond when they see a question posted. ***Please reserve PMs*** to the course instructor, the head TAs, or your discussion section leaders only for matters best discussed in private.

You are expected to join the course Slack workspace, and are responsible for monitoring it for any announcements posted by the instructor or the TAs - all such information constitutes official course communication.

Course Policies

Textbook. The course does not have a designated textbook, however, the course content relies on the following E-book developed by Dr. Dennis Sun for this course:

- Dennis Sun, *Principles of Data Science, https://github.com/dlsun/pods*

The link to this book is available both on the course web page and on our Slack. Most of Jupyter/Colab notebooks used in class are modified versions of Dr. Sun's original notebooks, or inspired by the notebooks in the book.

Prerequisites. I fully recognize that this class has a wide range of students with a wide range of expertise and skills. However, **we expect all of you to have basic knowledge of programming in general, and Python**, the programming language in which this course is taught, specifically. While this course introduces multiple new Python constructs, ***this is not an Introduction to Python*** course, and if you do not know Python, your experience in this course will be difficult. In addition, this course assumes basic familiarity with **introductory Statistics concepts**. We will be using some of these concepts throughout the course, and while we'll provide some refreshers, if you are completely unfamiliar with them, your experience in the course will be difficult.

Grading

Course Participation	20%
Labs	30%
Midterms	30%
Project	20%

Grading Policies.
Numeric scores on assignments are reflections of *your ability to complete them without making mistakes*. Very few of us are infallible, and even fewer of us are infallible 100% of the time. As a result, points will be taken off on assignments whenever you make mistakes, or whenever you do something *other* than what the specific assignment is asking you to do.

Letter grades for the course are a reflection of *how well you mastered the learning objectives of the course* (see above), and how well you are able to continue using the knowledge and skills gained in the course in your future pursuits. Can you achieve full command of the learning objectives of the course having had points taken off for mistakes on some of the assignments? The answer is ***yes, you can***. In fact, I express sincere hope here that ***all of you will.***

The grading policies, procedures, and responses to regrade requests for the course are developed in line with the two observations made above. You have the right to know why your score on an assignment was reduced, but the reason why we *want you* to know it is primarily to ensure that you understand the mistakes (omissions, misinterpretations, etc.) you made, and learn how to avoid them in the future. Regrade requests may be granted in case of a *bona fide* grading error; additionally, your scores may be adjusted, if we made an arithmetical error tallying up the results. However, all grading deductions for all assignments will be part of well-developed and detailed grading rubrics that will be applied consistently to all submissions, and requests to adjust scores for ***acknowledged/validated mistakes/omissions/misinterpretations*** on the

assignment that violate the grading rubrics **will not be granted.** *Missing points on assignments is NOT a sign that you are failing the class, it is a sign that you have things to learn in it.*

Most of the assignments will be submitted through GradeScope. GradeScope submission instructions are available on Canvas. GradeScope also has the functionality to request a regrade. All regrade requests will be handled through this feature.

Course Participation (20%)

- **Lecture participation** is highly encouraged. We will not be doing roll calls during lecture periods, but we might use short surveys to be completed during the lecture period. While learning your actual responses is the key objective of such activities, these surveys can also be used to establish the pattern of lecture attendance.
- **Discussion section participation** is **mandatory.** Each discussion session involves work on a Colab/Jupyter notebook that needs to be completed and submitted by a designated time (end of class, or some time after the class is over). Attendance will be checked every class.
    - If you are unable to attend the discussion session, ***please inform your*** **TA/CA** as soon as you are able to (email, or Slack PM). You can download the notebook for the day and complete it independently, and submit it by the time designated by the TA/CA in their communication with you.
    - You can do this up to **four (4) times** during the course.
    - Any additional missed recitation sessions will result in a decrease of the participation score at the rate of 2% per each missed recitation session up to a total of 12%.

  Please, also read the **Summer Session/High School Conflicts** section below.

Lab Assignments (30%)

Throughout the course, you will be given several lab assignments. Each lab assignment is an exploration of a specific dataset, primarily to be done in a pursuit of one or more specific goals related to the content covered in class. All lab assignments will be distributed in a form of Colab/Jupyter Notebook with instructions, and some instructor code to get you started.

Completed notebooks (with the student contact information cell at the top filled out) shall be submitted via Gradescope by the due date and time indicated for each assignment. No extensions will be granted. Late assignments are accepted only in the cases of documented medical/family emergencies, and with advance warning of the instructor (via email or Slack).

Midterms (30%)

The course will have three midterms tentatively scheduled for Weeks 3 and 6, as well as the last day of classes. Midterms administered during our lecture time will have the duration of 75

minutes. Midterms administered in session will have the duration of 50 minutes. Each exam is worth 10% of the grade. The exact nature of each midterm will be announced at least a week ahead of time. We expect all exams to be paper-and-pencil, but we reserve the right to schedule a live coding exam if we see need in it.

Study guides for paper-and-pencil exams will be shared with you ahead of any paper-and-pencil exam to give you an idea of the kind of questions you might see. We might share solutions to some study guide problems, but we may also choose to leave some problems in the study guide for you to solve on your own.

If scheduled, live coding exams will be structured similarly to the Colab/Jupyter notebooks used in discussion sections and assigned as labs, scaled to fit the expectations of live coding for 40 minutes or 65 minutes (10 minutes in each live coding exam are reserved to startup and winding down, including accessing the exam notebooks at the beginning of the exam, and submitting them at the end of the exam) depending on whether the exam is administered in the discussion section or in lecture.

Project (20%)

The course project is the culminating experience in this course and will be due the last official day of class (**Friday, August 15).** The project shall be completed by a small team of 3-4 students attending the same discussion section (you will be given some time during some discussion section meetings to work with your team). Sometime around Week 4 of the course, you will receive the project specification documentation. You will get one week to form a project team, decide on the nature of the project, and submit a project proposal.

Your project proposal will be evaluated in an expedient manner, the feedback will be provided to you by the instructor (and in some cases - by the TAs), and your final project needs to incorporate the provided feedback (mostly, the instructor is concerned with feasibility of your proposed projects, and with the chances of finding interesting insight, and he will caution you of any issues he perceives with the proposed work). . Once approved, you will have an intermediate milestone for your project at the end of Week 6 (or early Week 7) structured around data collection and preparation for analysis.

At the end of the course, each team shall submit **one set of deliverables** consisting of all the code used for the project, any data files necessary to make the code run, and a **written report** (format TBD) describing the project and its results. The project evaluation rubric will be shared with you when the project assignment is released.

Course Content Overview. Below is a short overview of the course. A more detailed schedule is posted on the course web page and will be updated as needed throughout the quarter.

Week	Course Content	Assignments
1	Introduction, data science and process Work with data in Python and Jupyter Data Operations	Lab 1
2	Analysis of individual variables Categorical and Numeric variables Data Visualization	Lab 2
3	Analysis of interactions of two variables Data Visualization (continued)	Lab 3 Midterm 1
4	Distance metrics and Text Analysis	Lab 4 Project Proposal
5	Predictive Modeling Supervised Learning: Regression	Midterm 2
6	Model Evaluation, metrics, training and testing models	Lab 5 Project Milestone
7	Supervised Learning: Classification	Lab 6
8	Unsupervised Learning: Clustering Review	Midterm 3 Project Due

Technology. The course uses Python, and specifically, most, if not all programming in this course will be done via **Python notebooks.** We use Google Colab in the course to create, maintain, demo, and run the notebooks, but you can also install Anaconda on your own computers and use Anaconda's Jupyter Labs server that will run on your own machine. All notebooks distributed by the instructor and the TAs will be equally runnable in each of these two environments (Colab, Jupyter Labs).

Google Colab environment is available via **any computer** capable of running a web browser, and requires relatively little in terms of resources on your own computing devices. It is integrated into the Google Drive suite of tools, and allows you to create, maintain, and run Jupyter Notebooks from your Google Drive. To use Google Colab you **must have a Google account (gmail account).**

Jupyter Labs environment available from the Anaconda package (the course web site contains instructions on how to download and install it) runs as a web service on your computing device, and can also be accessed via a web browser (you will be accessing a URL on your own machine that looks like "localhost:<some number>"). It allows you to work on your notebooks without using a Google/Gmail account, but all computations will be performed by your computing device, and the efficiency (time to completion) of these computations will be determined by how powerful it is. Additionally, your ability to complete certain tasks may be impacted by how much memory your computing device has. While this will not be a major factor for running the vast majority of notebooks shared with you by the instructor and the TAs (these are designed to use relatively small by modern standards datasets), this might impact your ability to complete certain tasks when working on the course project.

Academic Integrity

This course is participating in the proctoring pilot overseen by the Academic Integrity Working Group (AIWG). The purpose of this pilot is to determine the efficacy of proctoring and develop effective practices for proctoring in-person exams at Stanford. To find more details on the pilot or the working group, please visit the AIWG's webpage.

While we provide the information about the Stanford University Honor Code for your reference below, we note that Stanford is in the process of revising the Honor Code and the policies around exam proctoring to better fit with the modern use of computing technology. The Proctored Exam Pilot Program involves in-person proctoring of your exams by the instructor and/or TAs/CAs for the course.

Stanford University Honor Code

The Stanford University Honor Code is a part of this course. It is Stanford's statement on academic integrity first written by Stanford students in 1921. It articulates university expectations of students and faculty in establishing and maintaining the highest standards in academic work. It is agreed to by every student who enrolls and by every instructor who accepts appointment at Stanford. The Honor Code states:

1. The Honor Code is an undertaking of the students, individually and collectively
    a. that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;

b. that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.

2. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.

3. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

Penalties for violation of the Honor Code can be serious (e.g., suspension, and even expulsion). So re-read the Honor Code, understand it, and abide by it.

Sharing of Materials. Any materials linked to on the course web page can be retained and shared with you as you see fit. At the same time, from time to time I may share some materials with the class via means that prevent outside access to them - for example by posting a file to a Slack channel. Such materials are meant for you in the context of the class, but you **are not allowed** to publicly share them outside of the class, or with individuals not enrolled in the class (these are primarily solutions to various exams and exercises).

Collaboration on assignments. Discussing things you do not understand out loud with others is one of the best ways of eventually understanding them. As a result, you are welcome to discuss your assignments with your classmates as part of your work. **However,** you are **solely responsible** for the work that you submit. Your submitted assignments - code and written parts, ***must be solely authored by you.*** The line between what is allowable collaboration and what isn't is fine, but distinct: do not use other people's code or words - submitting any work that is not your own is ***taking credit for another person's work, and it prohibited by Stanford's academic integrity policies.***

Soliciting any input on your assignments from people not involved in the course is **not allowed.**

AI-generated Code and Content. The nature of the assignments in this course and the learning objectives of the course are incompatible with the use of AI-generated answers as part of your work. As such, they are **prohibited.** We are aware that Google Colab now has built-in AI to help with development. For all of the course assignments, there should be no need to engage it. We are well aware that many of the individual pieces of code needed to complete the assignments can be filled in by the Colab built-in AI, as well as other AI-powered software development

assistants. Our goal though is not to train AI models, but rather to teach **you** how to perform the tasks.

Accommodations

Students with Documented Disabilities. Students who may need academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE). Professional staff will evaluate the request with required documentation, recommend reasonable accommodations, and prepare an Accommodation Letter for faculty dated in the current quarter in which the request is being made.

OAE accommodation requests in DATASCI 112 will be handled with the help of College of Humanities and Sciences staff, who are working to consolidate the scheduling of the accommodations and make the process more convenient for students and faculty. Full instructions on how to request accommodations will be placed on the course Canvas page together with the links to the relevant on-line forms. Students who need OAE accommodations should initiate the accommodation process as soon as possible as it takes time to properly schedule and staff the accommodations. For your information, the OAE is located at 563 Salvatierra Walk (phone: 723-1066, online at http://oae.stanford.edu).

High School/Summer Session Conflicts. The Summer Session office includes the following language in the handbook:

- *Stanford Summer Session is an in-person program where students are expected to be on campus for the duration of the program.*
- *Summer Session high school students are expected to attend every class session of their courses during the program. Missing class without cause (i.e., properly documented medical or family emergency) is grounds for removal from the program.*

Pursuant to this, you are expected to be present in class, including your mandatory participation in Discussion Sections. No accommodations for remote course attendance will be granted. The Discussion Section attendance policy allows you up to four **excused** absences without Course Participation penalty. Excused absences must be documented and supporting documentation shall be submitted to the course instructor **by the end of the first week of class.** In case of emergencies, -please contact the instructor as soon as you know that you will be missing a discussion section.

If you are a high school student, in case of a conflict between the course schedule during the last week of this course (August 11 - August 15) and **required high school attendance,** you can request an alternate exam time. This course does not have a final exam, but it will have Midterm #3 taking place in discussion sections on **Thursday, August 14.** If you are requesting an alternate date due to a required high school attendance conflict, please submit the request to me

in writing (email *dekhtyar@stanford.edu)* by the end of Week 1 of class. Your request shall provide the dates of mandatory attendance, and link to the academic calendar of your high school. Optionally, you can suggest times, during the weeks of August 11-15 and August 18-22 during which you will be available.

Important Note.

Data Science does not exist in vacuum, rather our work on data analysis is contextualized by the data we work with, what this data means, when and how it was collected, and many other factors. Throughout the course, we will look at numerous historic datasets which, if collected today, would have had different structure. Whenever appropriate we will discuss cultural assumptions made by the data collectors, in terms of information collected and NOT collected, and we may discuss how it affects the analysis that we perform with the data. This *may* create a level of discomfort among some of you (a typical example would be a use of *binary gender* or *sex* in some historic datasets featured in the class). The data used in the course was selected with the explicit purpose of availing us to conversations about how information and data changes over time in line with the change of public perceptions and cultural expectations. If you are uncomfortable with using any of the datasets used in the class for any reason, please reach out to the instructor to discuss it.

Similarly, in my conversation with the class (as well as in TAs conversations in their recitation sections) you will often hear stories, pop culture references, puns, and other attempts to be illustrative in discussing the content of the class. Our intention is never to be offensive, but misstatements and unintentional offense can happen. If something one of us says, does, or writes ever bothers you, please let me know quickly so that I/we can correct the mistake, apologize, and, most importantly, avoid the offensive behavior in the future. Thank you.