DATA 301: Introduction to Data Science
Spring 2022
Course Syllabus

March 27, 2022

Instructor: Alexander Dekhtyar
email: dekhtyar@csc.calpoly.edu
office: 14-210

What	Day	Time	Location
Lecture	TR	3:10 – 4:30pm	180-262
Lab	TR	4:40 – 6:00pm	180-262

Office Hours

	When	Where
Monday	9:10am - 10:00am	14-212 and Zoom
Tuesday	9:10am - 10:00am	14-212 and Zoom
Friday	9:10am - 11:00am	14-212 and Zoom

Additional appointments can be scheduled by emailing the instructor at *dekhtyar@calpoly.edu*.

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 who want to pursue the Cross-Disciplinary Minor in Data Science, as well as for those who are looking to better understand what it means to work with data.

Texbook

The 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 and can be found at this url:

https://github.com/dlsun/data-science-book

The instructor will provide lecture notes for all concepts covered in the course.

Topics

The following topics will be covered in the course. The specific order of topics may vary.

No.	Topic	Duration (weeks)
1.	Introduction: data science and data science process	1
2.	Data acquisition	1
3.	Data cleaning	1
4.	Data modeling/Feature selection	1-2
5.	Data analysis	3-4
6.	Data visualization	1-2

Additionally, the course will look at a variety of data types, including, but possibly not limited to:

- tabular data
- structured and semi-structured data
- textual data
- temporal data
- geo-spatial data

Grading

Homeworks	0–5%
Labs	45 – 55%
Project	**10 – 15%**
Exams	30-40%

I give relatively hard problems and take points off on exams. Because of this, the traditional 90-A, 80-B, 70-C grading schema does not work in my classes. Historically, the A/B cutoff has been around 80-85%, while the B/C cutoff has been around 67-70%.

Course Policies

Disclaimers

Please be aware of the following:

- The prerequisites for this course are CSC 202 and either STAT 302 or STAT 312.
- This is **NOT** a database course. CSC 365, Introduction to Database Systems covers relational database model and SQL.
- This is **NOT** a machine learning course. CSC 466 and (in part) STAT 419 cover the data analytical methods traditionally referred to as "machine learning". We will see *some* machine learning technicques and will discuss them in action, but this course is not going to give you an in-depth knowledge of machine learning.
- This is **NOT** a "Big Data" course. DATA 401, and, partly, CSC 369, as well as CSC 487 involve workin, in a variety of ways, with large datasets. DATA 301 tends to stick to data that is smaller in nature.
- This is **NOT** a NoSQL course. CSC 369 covers aspects of NoSQL DBMS. So do certain versions of CSC 560.
- This is **NOT** a Hadoop (or MapReduce) course. Again, CSC 369 covers those topics.
- This course has been routinely taught by faculty from Statistics and Computer Science (initially it was also taught by Dr. Brian Granger from the Physics Department). You should expect that while we all will wind up talking about the same topics, each of us will bring the perspective (and, on occasion, biases) of our own field into the course. As such, *you should expect* ***this particular section*** *of DATA 301* to be taught the way a Computer Science course is taught at Cal Poly. If you compare notes with your peers who take a course from a different instructor, you will see some differences in the approaches to the material, assignments, and expectations. This is a **normal and natural** reflection of the fact that data science is not a one-size-fits-all discipline, and it can be taught, even at a very introductory level, in a number of different ways. All of them are equally valid, and you should be getting a reasonable perspective regardless of who the instructor in the course is. But in *this* section you will have to live with the specific idiosyncracies **I** will introduce into the course as the instructor.

Exams

The course will have either

- a midterm and a final exam, or
- two midterms plus project reports due the week of the finals, or
- one midterm, and oral project presentation scheduled for the week of the finals

The specific details of the final examination will be determined closer to the end of the course.

The midterm(s) can include a lab component (i.e., a portion of a midterm exam can be administered as a programming exercise).

Our scheduled final exam time is **Tuesday, June 7, 2022, 4:10am - 7:00pm**.

Homeworks, Labs

The course will primarily use Python as the programming language. The vast majority of students taking this class should have had CSC 101/CSC 202 in Python. We will rely on this in the course.

We will extend your knowledge of Python by stressing its use in data-driven applications. This includes the study of several Python packages traditionally used in data science and/or for data management, as well as the study of various Python coding techniques appropriate for data science applications.

The course will include multiple programming labs. Some of the labs will use the Jupyter Labs environment, and will require you to build Jypyter (iPython) notebooks complete with Python code, explanations, and visualizations of the results. Other labs will have you develop software as you usually do in Computer Science courses and submit it via `handin`.

We will have both individual and pair-programming labs. Overall, this class is about individual work conducted by each student. However, data science is a very collaborative and cross-disciplinary process, and therefore for some assignments, the importance of being able to discuss the nature of the work with a partner cannot be understated.

I typically use paper-and-pencil homeworks are study guides for the exams. Since we will have at least one written exam, I expect that we will have at least on paper-and-pencil homework.

Course Project

A key part of the course is the project. The project will take place over the last four weeks of the course. The project will be done in teams of two people. While the full details of the project will be revealed in due time, the general outline of the project is as follows. Each team will find/collect/use a dataset, ask analytical questions about the data and will conduct the necessary analysis and report the results.

At the end of the quarter we will arrange for project presentations either in the form of posters, or in the form of oral presentations (or both). Additionally, you will write papers describing your project.

Late Submissions

All assignments are due at classtime on the due date: homeworks - at the beginning of the class (with grace period extending to the beginning of the lab period); lab assignments - at the end of the lab period. Any deviations from these rules will be spelled out explicitly in the assignments.

Homework/lab assignments submitted later than indicated above will be considered *late submissions.*

If paper-and-pencil homework solutions are distributed on the due date of the homework, ***late homework submissions will not be accepted***. Otherwise, late homeworks can be submitted during next 24 hours for a 10-30% penalty (the exact amount will depend on the submission time and the specific circumstances). No homework submissions will be accepted afterwards.

Late lab assignment submissions can be turned in before or at the beginning of the next lab period for a 10-30% penalty (the exact amount will depend on the submission time and the specific circumstances[1]). No lab assignment submissions will be accepted after that.

Communication

Slack. The main form of communication is our Slack workspace. You all have received Slack invitations. I will use Slack for day-to-day announcements, interactions with individual students, and for any Q&As that may arise throughout the course.

Mailing List. The mailing list for the course is: data-301-10-2224@calpoly.edu All students enrolled in the class are automatically subscribed to the mailing list. I will use it for important announcements, usually to ensure that everyone, including those not typically active on Slack have access to all relevant course information. Please note, though, that I will mostly use Slack throughout the quarter.

Email. Feel free to email me at any time. While I prefer PMs on Slack as the method of personal communication regarding course content, I will try to respond to any emails as soon as I can.

Web Page

Class web page can be found at

http://www.csc.calpoly.edu/~dekhtyar/DATA301-Spring2022

This is a simple static page with links to all relevant course materials. These materials may also be linked to on Slack, but you will always be able to find them on the course page.

[1]The penalty will be larger if the gap between the two lab periods includes a weekend and smaller otherwise

COVID-19

We'll take it slow. I will start the class masked, as this is the safest. I will examine everyone's levels of comfort in the course during the first week (there will be a survey). Based on the response, I will make decisions on whether the course will continue masked or unmasked throughout the quarter.

We will follow Cal Poly instructions for all our COVID-19 related protocols, and our internal policies are subject to change if the COVID-19 situation in the country and/or on campus worsens.

Academic Integrity

University Policies

Cal Poly's Academic Integrity policies are found at

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 instuctor 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 Policies

All homeworks are to be completed by each student **individually**. Lab assignments are to be completed by the appropriate units (individual, pair, group), and no code/solution-sharing between units is permitted. Students are encouraged to discuss class content among themselves but NOT in a manner that constitutes plagiarism and cheating as defined above (e.g., you can solve together a problem from the textbook that had not been assigned in the homework, but you should solve assigned problems individually).