DATA 301: Introduction to Data Science
Spring 2016
Course Syllabus

March 27, 2016

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

What	Day	Time	Location
Lecture	MWF	10:10 – 11:00am	14-246
Lab	MWF	11:10 – 12:00pm	14-301

Office Hours

	When	Where
Monday	2:10pm - 3:00pm	14-210
Wednsday	2:10am - 3:00pm	14-210
Friday	8:30am - 10:00pm	14-210

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

There are a few O'Reilly books on data science/data analysis using Python.

Of those, Dr. Brian Granger, the course instructor for the innaugural version of the course selected the following (upcoming) book as the course textbook:

- Jake VanderPlas, *Python Data Science Handbook*, O'Reilly Media, 2016, early release.

You can purchase the early release E-book at the following URL (the link is available on the course web site):

http://shop.oreilly.com/product/0636920034919.do

If you want additional Data Science related books, the following list is a good start.

- Joel Grus, *Data Science from Scratch: First Principles with Python*, O'Reilly Media, 2015, ISBN: 978-1-4919-0142-7
- Wes McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy and iPython*, O'Reilly Media, 2012, ISBN: 978-1-4493-1979-3.
- Cathy O'Neill, Rachel Schutt, *Doing Data Science: Straight Talk from the Frontline*, O'Reilly Media, 2013, ISBN:978-1-4493-5865-5.
- Eli Bressert, *SciPy and NumPy: Examples to Jumstart Your Scientific Python Programming*, O'Reilly Media, 2012, ISBN: 978-1-4493-0546-8.
- Steven Bird, Ewan Klein, Edward Loper, *Natural Language Processing With Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, 1st Edition, 2009, ISBN: 978-0-596-51649-9.

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 102 and either STAT 302 or STAT 312. (This quarter I will allow STAT 321 to count as well, but moving forward all CS majors must switch to STAT 312, so this will stop being an issue).
- 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".
- This is **NOT** a "Big Data" course. For those of you planning on proceeding with the Data Science Minor, DATA 401 will cover some case studies actually involving massive amounts of data.
- 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, over the past two quarters (Winter'16 and Spring'16) is being taught by faculty from three different programs: Physics, Statistics and Computer Science. 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, 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.

Our scheduled final exam time is **Friday, June 10, 2016, 10:10am - 1:00pm**.

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 in Python. We will rely on this in the course.

The course will include multiple programming labs. Some of the labs will use the Jupyter 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 course of 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

The class have an official mailing lists. The email address for the mailing list is: data-301-10-2164@calpoly.edu All students enrolled in the class are automatically subscribed to the mailing list.

I encourage questions during classtime and questions via email. My answers to email questions may be broadcast to the entire class via the mailing list, if the answer may be relevant to everyone (e.g. a correction in a text of a handout, or a clarification of a homework problem), and may also appear on the web page. The questions can also be posted to the mailing list directly. The mailing list will also be used for all annoucements related to the course. It is your responsibility to read your class-related email. Failure to read email posted to the mailing list cannot be used as an excuse in the class.

Web Page

Class web page can be found at

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

http://www.csc.calpoly.edu/∼dekhtyar/DATA301-Spring2016

Through this page you will be able to access all class handouts including homeworks, project information and lecture notes (should the latter be written).

Strike

As you might know, Cal Poly faculty, as well as the faculty at other CSU campuses are represented for the purposes of collective bargaining by California Faculty Association (CFA), an affilliate union of SEIU, American Association of University Professors (AAUP), National Education Association (NEA) and California Teachers Association (CTA).

During the 2015-2016 academic year CFA has engaged in collective bargaining activities with the CSU adminstration on behalf of the CSU faculty. The CFA has asked for a 5% salary increase for all CSU faculty to help if not offset, then at least somewhat minimize the financial hardships faced by the CSU faculty over the past eight years of no stagnant salaries (complete with a 10% furlough during one of these years).

At present the CFA is at an impasse with the CSU administration regarding the salary negotiations. While this impasse can be resolved at any time, CFA has used its right to call for a concerted action. **As things stand now, CFA has called for a CSU-wide strike to take place during the following days:**

- April 13 — April 15
- April 18 — April 19

We have three classes scheduled during the announced strike days: on April 13 (Wednesday), April 15 (Friday) and April 18 (Monday).

If CFA and the CSU administration reach an agreement on faculty salaries, the strike will be called off and the classes on those days will take place as usual[2].

If **the agreement is not reached** and **the strike is on**, you should expect the following to happen:

- **No classes** on April 13, April 15, and April 18.
- Because the faculty who go on strike must stop **all work-related activities**, if I join the strike I will not respond to any work-related emails, and in fact, I will be unable to check my calpoly.edu email account. (Please note that April 16 and April 17 are **not** strike days, so during these two days, I may be able to check email and respond to it).
- Faculty who go on strike will not be present on campus, except for picket lines.

[2]Except for a possibility of April 15 class being converted to a double lab.

- **You are not on strike** in your capacity as Cal Poly students. What this means is that while I may not be able to hold the classes, there may be assignments that will have to be done by you during the strike days. I promise **no due dates** for course assignments on any of the strike days, but if the strike takes place, one or more assignments may be issued before the strike commences with due dates after the strike is over. The expectation is that you will spend some of your time during the strike days working on these assignments, even if I am unable to answer the questions you might have about them. (We will try to make sure you have sufficient time to resolve any questions related to these assignments prior to the start of the strike).

- Classes will resume as usual on Wednesday, April 20.

I understand how disruptive a week-long strike will be on your education and I apologize in advance for the negative impact on your education the strike might have. At the same time, as far as preventing the strike, the ball is squarely in the court of the CSU administration. If you want the CSU administration to know your thoughts concerning the faculty strike, I would like to encourage you to convey those thoughts - either directly, or through Cal Poly administration.

For more information concerning the strike, please see

http://www.calfac.org/item/faq-cfa-bargaining-possible-strike

and

http://www.calfac.org/strike-411-students

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).