COSC 480 Data Science
Spring 2017

Time	Lecture MW 2:45–4:00 pm
	Lab A Tu 1:20–3:10 pm
	Lab B Tu 3:30–5:10 pm
Location	314 McGregory (lecture), 315 McGregory (lab)
Instructor	Prof. Michael Hay (303 McGregory, mhay@colgate.edu)
Office hours	TBD (will post on website)

Course Description

This course will focus on fundamental principles that guide the extraction of knowledge from data. Topics include data visualization, data manipulation/wrangling, computational statistics, machine learning, and additional topics as the schedule permits. Students are not expected to have a background in probability and statistics. The course will be project-oriented and students will be expected to work independently, acquiring the necessary background knowledge specific to their project.

The required credit-bearing laboratory COSC 480L must be taken concurrently with COSC 480.

Prerequisites: COSC 301.

Materials & Resources

Course schedule/website: Course information including an up-to-date schedule will be available here: `https://github.com/colgate-cosc480ds/lecture`

Textbook (required): *Data Science From Scratch* by Joel Grus. `http://shop.oreilly.com/product/0636920033400.do` It is available in the Colgate Bookstore. An electronic version is also available through the Colgate Library. The code for the book is available here: `https://github.com/joelgrus/data-science-from-scratch`.

Additional readings (required): Additional readings will be posted online throughout the semester. Unless otherwise specified, you can assume these readings are *required*.

Supplemental readings (optional): Readings marked "supplemental" can be considered optional but you may find them useful for your projects.

Software (optional): The lab computers have all of the software needed for this course installed. These computers are available during lab and open lab hours in the evenings (schedule TBA). If you prefer to work on a different machine, you will be able to replicate the lab environment on your own personal machine provided that you install VirtualBox and Vagrant and then follow the instructions from the first lab.

Piazza (required): We will use Piazza for online discussion. It's accessible via Moodle.

Course Work

Reading: Reading assignments for each lecture will be posted on the schedule. Please complete the reading *before* class.

Lab: There will be a weekly lab assignment. Labs are intended to reinforce concepts from lecture and provide an opportunity to develop foundational skills necessary for data science.

Quizzes: Frequent, short quizzes will be used to check for understanding of concepts.

Capstone Project: You are expected to complete a capstone team project. The project will run the entire semester with milestones (proposal, mid-term reports, etc.) and culminate with a final presentation and report.

Participation in lecture: You are expected to attend class though I will not take attendance. My goal is to make the classroom a fun and supportive learning environment. To achieve this goal, I need your help. Please come to class on time, mentally and physically ready to engage in the learning process (i.e., pay attention, ask questions, answer questions, give your best effort on in-class exercises, etc.).

Participation in lab: Since lab time will be used to work intensively on the lab assignments, it is important that you come prepared and attend every lab session.

Participation on piazza: You are expected to monitor piazza for course announcements. In addition, piazza can be a useful resource to you when you're working on assignments. To that end, I will award bonus points (up to 3 points) on the final grade for those students who were active on piazza, posting but also *answering* questions.

Grading

Lab work is a significant component of the course work. Therefore, you will receive a single grade for the course as a whole and that one grade will be submitted to the registrar for both the course (COSC480) and lab (COSC480L).

An outline of the composition of your final grade is as follows. Grading is on an absolute scale (*i.e.*, no curve).

Coursework	**Portion of grade**
Labs	40%
Capstone project	40%
Quizzes	18% (see note 1)
Participation	2%

Note 1 The lowest quiz grade will be dropped.

Final course grades are determined as as follows. As a general rule, fractions are rounded down (e.g., an 89.9 is a B+, not an A-). A grade of A+ is awarded when the student demonstrates truly exceptional performance and is not simply determined by having a high final course grade.

F	D-	D	D+	C-	C	C+	B-	B	B+	A-	A	A+
< 60	60–62	63–66	67–69	70–72	73–76	77–79	80–82	83–86	87–89	90–92	≥ 93	*

Schedule & Special events

A *tentative* course schedule is available online: `https://github.com/colgate-cosc480ds/lecture`

Please take note of a few key special events. Your attendance at these events is mandatory:

- NASC Colloquium talk: Friday, April 21, 3:30pm
- Project presentations: Wednesday, May 3rd, 3-6pm

Topics

The course focuses on the following concepts:

1. Data management
   (a) Acquisition
   (b) Cleaning / wrangling / munging
   (c) Manipulation (SQL)
   (d) Exploratory data analysis
2. Computational statistics
   (a) Probability
   (b) Statistics
   (c) Hypothesis testing
   (d) Monte carlo, bootstrap
   (e) Linear regression
3. Machine Learning
   (a) Classification
   (b) Linear discriminants
   (c) Nearest neighbors
   (d) Decision trees
   (e) Overfitting
4. Visualization
5. Additional topics
   (a) Clustering
   (b) Record linkage (fuzzy matching)
   (c) Natural language processing
   (d) Streaming data
   (e) Network analysis
   (f) Privacy
   (g) Fairness

Policies

Late work A late submission is one that is not submitted before the deadline. An assignment submitted one second after the deadline is considered late. Lab assignments submitted within the first 24 hours after the deadline can receive a maximum score of 90%; within 24-48 hours, a maximum score of 80%; and so on. Note: the project deadlines, unless otherwise specified, are **firm** deadlines; late work will not be accepted.

Academic honesty You are expected to abide by Colgate's academic honor code: `http://www.colgate.edu/docs/default-source/default-document-library/honor-code8-20.pdf`. The next bullet clarifies what forms of collaboration are permitted in my course.

Collaboration, Plagiarism, and the Difference Between the Two There are two different kinds of working together: collaboration and plagiarism.

Collaboration

- Collaboration is good.
- You are encouraged to collaborate on ideas and program design.
- Programming is often a social effort, and there is much you can learn by talking out the ideas in this class with each other.
- You can by all means talk to each other and share ideas.
- You are permitted to look at each others' programs *under the right circumstances*. For example, what about helping a classmate *debug* their implementation after you completed yours? That can be good for both of you, providing that you are *helping them fix their code* and not simply sharing your solution. Another example, what about turning to a neighbor and saying, "I have no clue how to write this method. What did you do?" This is not collaborating – it's a one way transfer of information from someone who has learned to someone who hasn't. Not good.

Plagiarism

- Plagiarism is bad. DON'T DO IT!
- Any submission should be your work.
- Even if you work with someone else and share ideas, you must still write your own program/solution. If a piece of your submission uses someone else's idea, you must give that person credit (e.g., in program comments).
- Do not simply give your code to other students. I encourage you to work together to help debug your code, but you should do so sitting together.

The following are examples of plagiarism:

- Taking someone else's program, changing comments and variable names, putting your name at the top, and turning it in.
- Finding a similar program online, changing the variables and comments around, putting your name at the top, and turning it in.
- Finding a similar program in a book, changing the variables and comments around, putting your name at the top, and turning it in.

I am compelled by University policy to report instances of plagiarism and cheating. The potential consequences are outlined in the honor code.

[As an illustration of academic honesty, I will tell you that credit for the above description belongs to Dave Musicant, a professor at Carleton College. What you see above is slightly modified from the version on his syllabus. Thanks, Dave!]

Academic Support and Disabilities Services If you feel you may need an accommodation based on the impact of a disability, please contact Lynn Waldman, Director of Academic Support and Disability Services at 315-228-7375 in the Center for Learning, Teaching, and Research.

`http://www.colgate.edu/centers-and-institutes/center-for-learning-teaching-and-research/academic-support-and-disability-services`

Additional resources

Student Resources

Case Library/Informational Literacy and Reference: Use of the stellar library offerings, including the services of the outstanding reference and informational literacy librarians, is something to be made the most of during your time at Colgate. I suggest you get to know the librarians and to use their exceptional and imaginative expertise for assistance in ways that will enrich and enliven your intellectual studies and academic work.

Writing and Speaking Center: Regardless of ability, all writers benefit from having someone else read their essays and offer feedback. At the Writing and Speaking Center, writing consultants can help native and non-native speakers alike with a written paper's focus, development, organization, clarity, citations, or grammar. If you're preparing a speech or oral presentation, speaking consultants can help you organize your content and improve your delivery to an audience. All meetings are private, and you may visit at any stage of your process, from clarifying your initial ideas to reviewing a final draft or practicing an oral performance. See `http://www.colgate.edu/writingcenter` for more information or to reserve an appointment. The center is located in 208 Lathrop Hall. Phone: (315) 228-6085.

Counseling Center: Dawn LaFrance, Director. `http://www.colgate.edu/offices/support/counseling`. The first year can sometimes get bumpy; if you are experiencing emotional and personal difficulties (related to college or not), the Counseling Center offers completely confidential and highly professional services, both for individuals and groups.

ITS: IT Service Desk. Support and expertise related to computer and technology questions and problems, such as Moodle, email, Internet and public access computers on campus. Phone: (228-7111) Location: Third Floor of Case Geyer Library `http://www.colgate.edu/offices-and-services/information-technology`