CSCI 036 Foundations of Data Science Syllabus

Mark Huber

Summary

Data science is the interdisciplinary study of the tools and theory behind using data to extract knowledge. It combines ideas from statistics, computer science, and particular domains in the hard and social sciences in order to make predictions and optimal decisions. In this course you will learn the foundations of data science including the basics of how to structure, visualize, transform, and model data. The primary programming language that we will be using is R, which is both simple to use and was designed around using data. The development environment we will be using is R Studio. Both R and R Studio are open source, and so may be downloaded to your personal laptop or any other computer for free.

Time and Place

The class will meet at 9:00-9:50 AM MWF in Kravis 165, CMC.

Textbook

The textbooks are

- Foundations of Data Science by Mark Huber,
- Data Science for Climate Change by Mark Huber and Branwen Williams.

Both of these books are open source and free to download.

Office hours

I will hold open office hours Monday, Tuesday, Wednesday, and Thursday from 3:00-4:00 PM. These will be held over Zoom.

- Zoom office hours: https://cmc-its.zoom.us/j/155961110 (https://cmc-its.zoom.us/j/155961110).

You are free to pop in anytime to these hours without an appointment. If you cannot make these hours for a particular week, let me know and we can set up an appointment to meet over Zoom or in person.

Email

The best way to reach me is through email at autotomic@gmail.com (mailto:autotomic@gmail.com).

Please begin your subject line with **CSCI 036** (exactly!) so that I can filter your emails from the spam. While I try to check my email often, there might be delays, and so you should not assume that I will answer your emails immediately. Especially the night before homework or a lab is due you might not get a response until the next morning.

Exams

This course has two midterms, each worth 20% of your grade and a final worth 40%. The dates for the exams are:

- Midterm 1: 5 October, 2022
- Midterm 2: 2 November, 2022
- Final: (to be announced)

The date for the final is set by the registrar and cannot be changed. Please bear this in mind when making your travel arrangements.

All of these exams will be given in class, and you will be allowed pencil, paper, and calculator. You will *not* be using your laptop for these problems. For the midterms, you will be allowed to use a sheet of paper (US letter size) both sides to write whatever you would like. For the final, you will be allowed to use two sheets of paper (US letter size) both sides to write whatever you would like.

Assignments

There will be weekly assignments, posted on Friday and due back by the next Friday. While you are welcome to work together on the assignments, the final write up should be your own. In the write up, indicate your calculations and reasoning for all work submitted. For numerical answers, put a box around your answers and use four significant figures for approximation (unless instructed otherwise in the problem statement.)

Homework in this course must be prepared using R Markdown. You can then knit your `.Rmd` file to an `.html` file, which can then be printed to a `.pdf` file. This file will then be submitted using Gradescope. It is up to the user to make sure that their file is printed properly with separate pages, and that the problem answer is marked for each page properly on Gradescope. Submissions that are improperly formatted risk getting 0 out of 10 points.

Each problem in the homework will be worth 1 point, and will receive a score in the interval $[0, 1]$. That is to say, partial credit will be given for partially correct answers.

The lowest homework score of the semester will be thrown out. This is basically to handle an emergency where you are unable to complete an assignment for external reasons that would just sound silly if you tried to email them to me as an excuse. I strongly recommend you save this freebie as long as possible and do not blow off an early assignment, since the later assignments tend to be more difficult. This especially applies to those of you with term papers due in other courses–that free homework assignment can be lifesaver for the last week of the semester. The homework is designed to take 4 to 6 hours each week. Roughly speaking, this involves an hour of looking over notes, an hour or two of solving the problems, an hour in office hours getting strategies for tackling more difficult exercises, and an hour or two on the final write up. Your time needs will vary, and I always recommend starting the assignment as early as possible, especially the computer related parts. Homework comprises 10% of your grade. You are allowed to work together on the homework, but the write up must be your own. The homework is a service provided for your benefit to gauge your understanding of the material.

Just to emphasize: the homework is worth only $1/10$ of your grade. The homework are exercises that allow you to test yourself on your comprehension of the material, and practice your skills in preparation for the tests. I'm often asked what the best way to prepare for the tests. The answer: try to do the homework by yourself, and only after questions arise talk with friends or come in and see me.

Solutions for the homeworks will be posted before each exam to help study.

Labs

Monday and Wednesday will be regular lecture sections. Fridays are designated as *lab sessions*. These labs are intended to be exercises to work through to build understanding and check that you comprehend the material. Sometimes they will move beyond the material in lecture. These labs are primarily computer based, and so you will want to bring a laptop with you to class on these days.

The product of the lab session will be a script, a `.R` file, that will be uploaded to Gradescope and autograded. That means that you will instantly be given a score. You can then continue to work, fixing problems that are incorrect, and resubmitting as often as you wish before the deadline, which will be Friday at midnight.

The labs will be posted at the beginning of the week for those who wish to get a jump on them.

Like the homework you are welcome to work with other students on the problems. Again, the labs are primarily for your benefit to test your understanding of the material. They are, like the homework, only 10% of your grade.

Instant failure

Failure to submit 5 homeworks or labs on time (except under extraordinary circumstances) will result in an F automatically being given for the course. If you can only do one problem half right, turn it in and it will not count against your five missing pieces of work.

Classroom participation

As part of the classroom experience, I will at times ask questions of randomly chosen members of the class. This is not meant to torture students, rather there are several reasons for this approach. First, I need to determine how understandable the lecture is to the class. I understand the material, but it can be difficult without direct questioning to discover how much the class understands. Second, very few students have the experience of speaking computer language at the collegiate level out loud.

Being able to fluently discuss mathematics and code is an ability that can be developed with practice. Homework typically only tests your ability to write code, but not to say it out loud.

Also, by practicing now in a relatively laid back environment, it will be much easier to converse in code and mathematics when it really matters, such as at a job interview or when presenting work at conferences. Finally, it keeps people awake.

Extra credit: Zombie points

The extra credit available in the course is called *Zombie points*, and work as follows. Each time you see an error in any email, text, homework, lab, or other written communication, you can email me with subject line **CSCI 036: Zombie point** pointing out the location of the error. If you are correct and the first to submit, you will receive a Zombie point.

If your Zombie point total is more than the value of any part of your score, then you can bring that homework back from the dead up to the points you have. For example, if you have a homework with score 6 out of 10 and 8 Zombie points at the end of the semester, you can raise that homework up to an 8 using your Zombie points.

If you have 17 Zombie points you can raise any homework or lab up to 10 and still have 7 Zombie points for a second homework or lab, and so on.

If you get enough Zombie points, you can even change a midterm or final, but I’ve only ever had one student get enough to accomplish that.

Grades

Your course numerical grade is

$$0.1 \cdot \text{homework} + 0.1 \cdot \text{lab} + 0.2 \cdot \text{mid1} + 0.2 \cdot \text{mid2} + 0.4 \cdot \text{final}.$$

After calculating your numerical grade, it will be converted to a letter grade as follows.

Score	Grade
93% and up	A
90%-93%	A-
87%-90%	B+
83%-87%	B

80%-83%	B-
77%-80%	C+
73%-77%	C
70%-73%	C-
70% and below	Let’s not find out

Note: I do not round scores, all that does is change the cutoff for points. If your percentage grade is 82.995%, then you will receive a B-.

Tentative schedule

Sometimes lectures will take longer or be shorter than the scheduled time, but this outline provides a rough idea of what topics will be covered when. It also gives the appropriate chapters in the textbooks to read before that day’s lecture.

Date	Note
29-08-2022	Introduction to Data Science (Ch 1 FDS, Ch 1 DSCC)
31-08-2022	R and R Studio (Ch 2 FDS, Ch 2 DSCC)
02-09-2022	Lab #1: Scripts in R
05-09-2022	Labor Day Holiday (no class)
07-09-2022	The tidyverse (Ch 3 FDS, Ch 3 DSCC)
09-09-2022	HW #1 due, Lab #2
12-09-2022	Importing Data (Ch 4 FDS, Ch 4 DSCC)
14-09-2022	Transforming Data (Ch 5 FDS, Ch 5 DSCC)
16-09-2022	HW #2 due, Lab #3
19-09-2022	Graphical grammars (Ch 6 FDS, Ch 6 DSCC)
21-09-2022	Advanced graphical grammars (Ch 7 FDS, Ch 8 DSCC)
23-09-2022	HW #3 due, Lab #4

26-09-2022	Grouping observations (Ch 8 FDS, Ch 7 DSCC)
28-09-2022	Joining datasets as sets (Ch 9 FDS)
30-09-2022	HW #4 due, Lab #5
03-10-2022	Review for Midterm #1
05-10-2022	Midterm #1
07-10-2022	Lab #6
10-10-2022	Joining datasets using keys (Ch 10 FDS, Ch 9 DSCC)
12-10-2022	Shaping data (Ch 11 FDS, Ch 10 DSCC)
14-10-2022	HW #5 due, Lab #7
17-10-2022	Fall Break (no class)
19-10-2022	Strings and regular expressions (Ch 12 FDS, Ch 11 DSCC)
21-10-2022	HW #6 due, Lab #8
24-10-2022	Backslashes (Ch 13 FDS, Ch 12 DSCC)
26-10-2022	Factors (Ch 14 FDS, Ch 13 DSCC)
28-10-2022	HW #7 due, Lab #9
31-10-2022	Review for Midterm #2
02-11-2022	Midterm #2
04-11-2022	Lab #10
07-11-2022	Introduction to SQL (Ch 15 FDS)
09-11-2022	Joining tables in SQL (Ch 16 FDS)
11-11-2022	HW #8 due, Lab #11: working with SQL
14-11-2022	Models (Ch 17 FDS)
16-11-2022	Evaluating models (Ch 18 FDS)

18-11-2022	HW #09 due, Lab #12
21-11-2022	Case study Titanic survival (Ch 19 FDS)
23-11-2022	Thanksgiving (no class)
25-11-2022	Thanksgiving (no class)
28-11-2022	Representating data (Ch 20 FDS)
30-11-2022	Machine learning (Ch 21 FDS)
02-12-2022	HW #10 due, Lab #13
05-12-2022	Ethics in Data Science
07-12-2022	Review for final
09-12-2022	HW #11 due, Lab #14