CSCI 036 Foundations of Data Science Syllabus (Fall 2024)

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 at Kravis 165, CMC.

Textbook

The textbook is

- Foundations of Data Science (https://s3.us-west-1.amazonaws.com/markhuber-datascience-resources.org/Books/Foundations_of_Data_Science/_book/introduction-to-data-science.html) by Mark Huber,

This book is open source and free to download.

Office hours

I will hold open office hours Monday, Tuesday, Wednesday, and Thursday from 2:50-3:50 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, including the space!) 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.

Grade breakdown

The grade will be: Homework 10%, Labs 10%, Midterm 1 20%, Midterm 2 20%, Final Part 1 20%, and Final Part 2 20%.

Exams

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

- Midterm 1: 04-Oct-2024
- Midterm 2:
- Final: 12-Dec-2024 9:00AM-12:00PM (Thursday)

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.

Sometimes homework will be submitted as an .Rmd file through Gradescope in fashion similar to the labs. Be sure to read the instructions on the homework to see if this is the case for a particular week.

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 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 10-11% 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 about $1/10$ 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.

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 an .Rmd 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 10-11% 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 only extra credit available in the course is called *Zombie points*. Each time you see an error in any posted text, homework, lab, or other written communication on Sakai, 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 for the effort.

These are *not* simple bonus points. Instead, 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, then 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 a few students get enough to accomplish that. One senior graduating student attained 90 Zombie points, and with his other grades meant that he could skip the senior Midterm 3, use his 90 points on it, and still get an A.

Grades

Your course numerical grade is

$$0.10 \cdot \text{homework} + 0.10 \cdot \text{lab} + 0.20 \cdot \text{mid1} + 0.20 \cdot \text{mid2} + 0.20 \cdot \text{final part 1} + 0.20 \cdot \text{final part 2}.$$

Numbers to letters

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, and so topics will have to be moved to other days. That being said, 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
2024-08-26	Introduction to Data Science (Ch 1 FDS)
2024-08-28	R and R Studio (Ch 2 FDS)
2024-08-30	Lab #1: Using scripts and Rmd files

Date	Note
2024-09-02	Labor Day, no lecture
2024-09-04	The tidyverse (Ch 3 FDS)
2024-09-06	HW #1 due, Lab #2
2024-09-09	Importing Data (Ch 4 FDS)
2024-09-11	Transforming Data (Ch 5 FDS)
2024-09-13	HW #2 due, Lab #3
2024-09-16	Graphical grammars (Ch 6 FDS)
2024-09-18	Advanced graphical grammars (Ch 7 FDS)
2024-09-20	HW #3 due, Lab #4
2024-09-23	Grouping observations (Ch 8 FDS) (Video, no lecture)
2024-09-25	Joining datasets as sets (Ch 9 FDS) (Video, no lecture)
2024-09-27	HW #4 due, Lab #5 (No inperson class)
2024-09-30	Joining datasets using keys (Ch 10 FDS)
2024-10-02	Review for Midterm #1
2024-10-04	Midterm #1
2024-10-07	Shaping data (Ch 11 FDS)
2024-10-09	Strings and regular expressions (Ch 12 FDS)
2024-10-11	HW #5 due, Lab #6
2024-10-14	Fall Break
2024-10-16	Backslashes (Ch 13 FDS)
2024-10-18	HW #6 due, Lab #7
2024-10-21	Factors (Ch 14 FDS)
2024-10-23	Representing data (Ch 21 FDS)
2024-10-25	HW #7 due, Lab #8
2024-10-28	Introduction to SQL (Ch 15 FDS)

Date	Note
2024-10-30	Joining tables in SQL (Ch 16 FDS)
2024-11-01	HW #8 due, Lab #9: working with SQL
2024-11-04	Principles of Functional Programming (Ch 23 FDS)
2024-11-06	Review for 2nd midterm
2024-11-08	Midterm #2
2024-11-11	Models (Ch 17 FDS)
2024-11-13	Evaluating models (Ch 18 FDS)
2024-11-15	HW #09 due, Lab #10
2024-11-18	Case study Titanic survival (Ch 19 FDS)
2024-11-20	Machine learning (Ch 20 FDS)
2024-11-22	HW #10 due, Lab #11
2024-11-25	No lecture (Thanksgiving Week)
2024-11-27	No lecture (Thanksgiving Week)
2024-11-29	No lecture (Thanksgiving Week)
2024-12-02	Ethics in Data Science
2024-12-04	Building apps with R
2024-12-06	HW #11 due, Lab #12
2024-12-12-09:00	Final