PS 300 - Data Science Methods & Machine Learning

Fall 2025

Taught by: Fabricio Vasselai **(**vasselai@illinois.edu)
When: Mondays & Wednesday, 5:00pm-6:20pm.
Classroom: 70B Wohlers Hall.
Office Hours: (in person) Tuesday 11:50am-12:50pm at 426 David Kinley Hall.
(remotely) variable (usually either T, Th or Fr, late afternoon), by appointment.

Course Overview

This course is designed as a gentle, applied introduction to key computational concepts and methods in Data Science and Machine Learning, geared towards Social Scientists. The rapid increase in the variety and sophistication of tools at our disposal is transforming the Social Sciences, making them increasingly computational. The goal of the course is to provide students, even those with very little quantitative background, with the knowledge and skills necessary to 1. understand and apply the basic versions of each of a variety of Data Science techniques; 2. be able to later pursue further training in case they decide to learn some of those methods in greater depth. The course assumes that students have already been introduced to the R programming language, but we will spend two weeks doing a fast-paced but thorough review of base R and programming logic (free online R tutorials will also be made available). Next, students will learn critical Data Science techniques, such as advanced data handling and cleaning, data visualization and web-scraping, as well as how computer hardware works. In the second part of the course, I introduce Monte Carlo simulations and bootstrapping, as well as the basics of text analysis, network analysis, and spatial data. The third and final part of the course focuses on introducing students to Machine Learning, covering key topics such as k-nearest neighbors, decision trees, random forests, nonlinear regression, and deep neural networks.

Prerequisites

Students should only take this course if they:

- Successfully took PS 230 or equivalent (do reach out to me to discus, in case you are in doubt);
- Have been introduced to linear regression in a university course before;
- Have been introduced to reading and writing R code.

In the first two weeks, the course will do a fast paced (but as detailed as possible) review of R. Besides, additional optional DataCamp online courses on R will be made freely available to those enrolled, in case students want to further review / improve their R skills.

Computers in class

The course will be taught in a Computer Lab Classroom, equipped with a computer per student. Yet, students can also bring their own laptops if they so prefer. Note that ATLAS Share is a university program that provides long term loaner computers to LAS undergraduate students. We will be using the RStudio software to write code in the R programming language, which will both be already installed in the computers present in the classroom and other university computers as well. However, if you also plan on using your own personal computer or laptop (which I highly recommend, if you can), you should go ahead and install

the R statistical software and RStudio IDE. I can help troubleshoot basic installation issues, but major roadblocks with installation will need to be addressed with help of LAS ATLAS.

Course Structure

With few but noticeable exceptions, class sessions will be very hands-on, focusing mostly on implementing techniques using R. Heavily commented code used by the instructor will be always made available to you, but you should come prepared to follow in your computer (your own, or the classroom's) what the instructor does in class. This is absolutely essential.

All course content and readings will be organized on Canvas under "Modules" on the left-hand side of the screen. Students should read all of the eventually assigned readings listed for a given class session ***ahead of class time*** and be prepared for in-class discussion.

There is no specific *textbook* for this course.

Communications

Strictly all regular course-related communication will happen via Piazza, (which means not via email), a very handy on-line questions-and-answers platform. In Piazza, students can ask questions anonymously, both publicly and privately (although the latter should be saved only for discussing personal matters or when the question being asked contains part of a solution for a homework). This way, we centralize our communication, and it makes it so that all questions and doubts you might have end up reaching all colleagues. Another neat feature of Piazza is that students can also answer to students. I **will** consider your Piazza activity (asking & answering questions) in your participation grade.

In the rare circumstance that you should feel the need to send me a direct email, start the subject of the email with "[PS300]" (not just "PS300"), otherwise there will be a high likelihood that I miss your email.

Always expect 2 business days for me to answer to any form of communication (that is about 48 hours, not including weekends).

Evaluation & Grading

DataCamp courses	***15%***
First Home Exercises	***12.5%***
First In-Person Exercises	***12.5%***
Second Home Exercises	***12.5%***
Second In-Person Exercises	***12.5%***
Third Home Exercises	***12.5%***
Third In-Person Exercises	***12.5%***
Participation	***10%***

Note: I will only permit students from dropping out of the course before their Third Home Exercises.

DataCamp courses
In some weeks, students will have to complete DataCamp courses / course chapters online, that will be assigned beforehand. These must be completed always before class time. They will not be graded for accuracy, just for completion.

First, Second and Third Home Exercises
There will be 3 lists of graded home exercises, focusing mostly on the applied programming of topics covered before their due dates.

First, Second and Third In-Person Exercises
There will be 3 lists of in-person exercises, focusing mostly on the applied programming of topics covered before their due dates. These in-person exercises will be direct extensions of the home exercises, happening shortly after the home exercises are due.

Participation
A big part of this class involves in-class discussion of the concepts and challenges discussed in class, as well as of the assigned readings.

Submitting Late Assignments

Due to the nature of the assignments (where there is an in-person set of exercises that immediately follow-up the submission of the home set of exercises), this course has a **very** strict lateness policy for assignment submission. In detail, here is the late policy.

DataCamp courses
Considering their online nature, their short length and their purpose of having students practice code before class, DataCamp courses should never be completed late, under any circumstances. If they are completed late:

- Up to 48 hours late, I will discount half of its corresponding points;
- DataCamp courses completed more than 48 hours late will be given zero points.

First, Second and Third Home Exercises
In case home exercises are submitted late:

- up to 6 hours late, there will be no penalty;
- more than 6 but less than 24 hours late will incur in a -10 points penalty (out of 100);
- more than 24 but less than 48 hours late will incur in a -20 points penalty (out of 100);
- assignments submitted more than 48 hours late will not be accepted*.

* of course, unless a student has academic accommodations **previously** approved by the Division of Disability Resources and Educational Services. In that case, I will follow the guidelines from the specific accommodations the student was approved for (see below). But these must be discussed in advance.

Plagiarism and Academic Honesty

According to the Student Code, 'it is the responsibility of each student to refrain from infractions of academic integrity, from conduct that may lead to suspicion of such infractions, and from conduct that aids others in such infractions.' I expect you to be familiar with and understand the university's policies on academic honesty for this course. Please consult the Student Code for more information (https://studentcode.illinois.edu/article1/part4).

- Any plagiarism (of existing work, of others or even of one's own past work, will result in a zero grade and will be reported to the student's home department and to the University authorities.
- Any copying of others' work will result in a zero grade and will be reported to the student's home department and to the University authorities.
- Unless otherwise explicitly stated in my assignment instructions, **absolutely any usage of ChatGPT or other AI / Generative AI / Large Language Model tools is strictly prohibited**. Again, using any of those will result in a zero grade and will be reported.
- When in doubt, ask! It is far better to check with us prior to submitting an assignment than waiting.

Please note: it is my work, ethical and even legal responsibility as an instructor to uphold the academic integrity policy of the University. This is really for real. I have zero flexibility with these things and will go above and beyond to make sure students that are eventually caught infringing the Student Code do face all consequences admissible by the University, state and federal statutes.

Note: unrelatedly, here is a small experiment just to verify how many of you are reading the syllabus with attention. If you've read this far, go to our Canvas website and submit your all-time favorite meme (or link to a song/video you love, in case you don't have a favorite meme) in the Assignment called "Testing, 123!". Let's see how many students catch this!

Statement on diversity, inclusion, and disability

The University of Illinois is committed to diversity and rigorous inquiry from multiple perspectives. The Political Science department shares this commitment and seeks to foster productive learning environments based upon inclusion, open communication, and mutual respect for a diverse range of identities, experiences, and positions.

The University of Illinois is committed to ensuring equitable access to our academic programs and services. Students with disabilities who have been approved for the use of academic accommodations by Division of Disability Resources and Educational Services (DRES) and need a reasonable accommodation(s) to participate fully in this course should follow the procedures established by DRES for using accommodations. Timely notifications are required in order to ensure that accommodations can be implemented. Please meet with the instructor to discuss access needs in this class after completing the DRES procedures for requesting accommodations.

- Email: disability@illinois.edu
- Phone: (217) 333-1970

Mental Health

Significant stress, mood changes, excessive worry, substance/alcohol misuse or interferences in eating or sleep can have an impact on academic performance, social development, and emotional wellbeing. The University of Illinois offers a variety of confidential services including individual and group counseling, crisis intervention, psychiatric services, and specialized screenings which are covered through the Student Health Fee. If you or someone you know experiences any of the above mental health concerns, it is strongly encouraged to contact or visit any of the University's resources provided below. Getting help is a smart and courageous thing to do for yourself and for those who care about you.

- Counseling Center (217) 333-3704
- McKinley Health Center (217) 333-2700
- National Suicide Prevention Lifeline (800) 273-8255
- Rosecrance Crisis Line (217) 359-4141 (available 24/7, 365 days a year)

Tentative Course Schedule

Note: *This schedule is still subject to change. Check on Canvas for updates as the course progresses.*

Class	Date	Topics
		Part I - R programming skills
1	**Aug 25**	Class presentation and R review (programming concepts & R syntax)
2	**Aug 27**	R review (basic syntax and vectors)
Sep 01 (no class)		*Labor Day*
3	**Sep 03**	R review (vectors, matrices, data.frames)
4	**Sep 08**	R review (if-else conditions & loops)
5	**Sep 10**	R review (apply family of commands & custom functions)
6	**Sep 15**	Data Cleaning and Data Handling with base R
7	**Sep 17**	Data Cleaning and Data Handling with data.table
8	**Sep 22**	How computers work & Code Performance
9	**Sep 24**	Parallel Computing in R
Due Sep 28 (Sunday)		*First home exercises*
Oct 1		*First in-person exercises*
		Part II - Data Science basics
10	**Sep 29**	Pseudo-Random Numbers and Intro to Monte Carlo
11	**Oct 06**	Applied Monte Carlo methods
12	**Oct 08**	Web Scraping (for HTML & automatized downloads)
13	**Oct 13**	Web Scraping (for CSS & XML)
14	**Oct 15**	Processing and visualizing Text data
15	**Oct 20**	A primer on Text Analysis and Text Mining
16	**Oct 22**	Handling and visualizing Network data
17	**Oct 27**	A primer on Network Analysis
18	**Oct 29**	Handling and visualizing Geo-spatial data
Due Oct 31 (Friday)		*Second home exercises*
Nov 03		*Second in-person exercises*

Part III - Intro to Machine Learning		
19	**Nov 05**	Intro to Machine Learning
20	**Nov 10**	SML - k-Nearest Neighbors & Decision Trees
21	**Nov 12**	SML - Random Forests
22	**Nov 17**	SML - Linear regression
23	**Nov 19**	SML - Non-linear regressions
Nov 24 & 26 (no class)		*Thanksgiving*
24	**Dec 01**	SML - Deep Neural Networks - part 1
25	**Dec 03**	SML - Deep Neural Networks - part 2
Due Dec 05 (Friday)		*Third home exercises*
Dec 08		*Third in-person exercises*
26	**Dec 10**	SML wrap-up: Model selection & Cross-Validation