

DATA 419 — Data Mining (Machine Learning I)

Professor: Stephen Davies

Fall semester 2026

Class: TR 2pm in Farmer 025

Final exam: Thursday, December 10th, 3:30–6pm

Office Hours (James Farmer Hall 044):

Mon 12–2pm

Tue 10–11am

Thu 12–2pm

<http://stephendavies.org/data419>

On the short list for “the technological development that has most changed our 21st-century world” is the exploding family of fields known as Data Mining, Machine Learning, and Artificial Intelligence. You probably don’t need much convincing of this anymore. These techniques power recommendation systems, speech and image recognition, fraud detection, medical diagnosis and drug discovery, weather prediction, supply-chain optimization, autonomous vehicles, scientific discovery — and, increasingly visibly, chatbots and large language models, AI coding assistants, and systems that generate images, audio, and video. I could easily double, triple, or quadruple that list without even having to think too hard, and you probably could too. Machine Learning is all around us.

And then there’s the dark side. Deepfakes and other synthetic media. AI-assisted scams and misinformation. Recommendation systems that steer attention toward toxic or manipulative content. The reinforcement of bias in hiring, lending, policing, and other consequential decisions. Threats to privacy. Systems that confidently produce information that simply isn’t true. Cathy O’Neil’s *Weapons of Math Destruction* chronicles some of the ways data-driven algorithms were already negatively impacting society before the current AI explosion; the stakes have only gotten larger.

Now some of the most spectacular recent advances in AI involve neural networks and other techniques that are mostly the subject of DATA 429, the course you’ll take next semester. This class is a bridge to get there. Here we’ll concentrate on another large and important part of the story: how computers can discover patterns in data, learn from examples, make predictions, identify useful structure, and help us make decisions.

The leaders in this powerful arena, btw, are sometimes called **wizards**, seemingly able to do magic with statistics and computers as their tools. But one of the goals of this course is to discover that it isn’t really magic. We’ll open up the cauldron, look inside, and drink a good, long draught to see what it tastes like.

Course Objectives

- To survey some of the most important applications of Machine Learning (ML), so that you know where the various use cases and technologies “fit” within the field.
- To tie together the two halves of Data Science. Whereas a course like 420 (Modeling & Simulation) is about creating a model of a data-generating process and seeing whether the data it generates matches the real world, 419 is about analyzing real-world data to infer properties about the process that generated it. It’s kind of like 420 in reverse.

- To give you both theoretical understanding and also practical experience working with real ML implementations on real data sets.

Student Learning Outcomes

After completing this course, students will be able to...

- ...identify the different subfields within the field of Machine Learning (ML), and to explain their purposes, assumptions, promises, and limits.
- ...understand – and competently carry out – the various stages in the ML pipeline, from data acquisition all the way to the presentation of results.
- ...both:
 - comprehend the mathematical theory underpinning important ML algorithms, and
 - apply that theory and implement those algorithms concretely in a programming language (Python).
- ...fluently use the key ML-related Python libraries of NumPy, SciPy, matplotlib, and scikit-learn.
- ...explain how data mining (DATA 419) – reasoning from observed data backwards to a plausible model – is the inverse of modeling and simulation (DATA 420) – producing and analyzing the data from a simulated model.

Rules of the game

1. There are absolutely, positively, NO stupid questions!! Your job is not to already know everything before you start the course. Your job is to try hard to learn, and part of that involves asking questions. I'm a nice guy, and I will not ever belittle you, snub you, or make fun of you; and if anyone else does so I will personally break both of their arms.
2. This class will be interactive. When I point at you in class, say your first name, and be prepared to try and answer questions. (Don't worry if you don't know all the answers.)
3. The book we'll be using as a framework this semester is James & Witten's *Introduction to Statistical learning in Python*." It is an excellently-written, mathematically sound, fully open-source text, and can be thought of as the baby version of their much longer *Elements of Machine Learning*," a tome which is considered hands-down the most definitive on the entire subject. It is mandatory, and you will be required to actually read the sections I assign. However, the quizzes and the final exam will not cover anything from the book that I don't specifically cover in class. Also, the quizzes and the final exam *will* definitely contain lecture material beyond what's in the book.

4. Don't skip class. Just don't. It's bad form. I work hard to prepare for class, to make it compelling and relevant. It hurts my feelings when you don't come. Plus you miss out on important stuff, and you'll end up falling behind if you skip lecture. So come every time. Come happy, fresh, excited, ready to think and to participate.
5. **Absolutely no laptops, cell phones, or other devices during class** I've had students claim that they take notes on their laptop during lecture, but even if it's true, those things are way too big a distraction to you and your fellow students to make it worth it. Just stay tuned in, because I move fast.
6. For any Zoom class that may take place this semester, **you must have your webcam on during the entirety of the lecture**. If you don't have a working Webcam, buy one immediately.

The Honor Code and this course

For this course:

- The Canvas quizzes must be taken alone, in a quiet place, without any form of contact with anyone.
- **You must write all your own Python code in its entirety.** You may not copy any part of anyone else's program. That being said, it is okay to work alongside classmates in a study group, and you are allowed to talk through homework assignments, ask others how they solved a particular problem, and so forth. **If you do this, you must name all the people you worked with on your homework submission. If you fail to give this information clearly on your homework submission, it is an Honor Code violation.**
- For the programming assignments specifically, do **not** copy or share code verbatim with anyone, but you can talk general strategy with classmates, or ask classmates how they solved something. You may **not** discuss any assignment with anyone not currently in the class (which includes people not even at UMW, of course).
- Note that the above statement about study groups does **not** include anyone not in the course. You may not discuss the homeworks with former 419 students, other DS minors or CPSC majors, or anyone outside UMW.

Books

- James, G., et al.. (2013). *An Introduction to Statistical Learning, with Applications in Python* ISBN 978-1-4614-7138-7.
- Davies, S. (2017). *A Quick Steep Climb Up Linear Algebra*. ISBN 978-1034259794.

The immortal *ESL* ("Elements of statistical learning") is a positive force in the field of principled Data Mining, and both the depth and the breadth of its content make it very attractive to keep professionals and researchers the world around engaged with the very cutting edge of the discipline. But while a fantastic research work, it's too

broad, and technical, and mathematical for a beginning to encounter the concepts for the first time. So some of the authors from the original ESL book decided to fill a need in the community: the same clear and insightful content, and the same vocabulary and goals, but with 1/4th the ambition. That “littler” work is called ISL (“Introduction to statistical leaning”) and should fit our educational objectives this semester like a glove.

Even ISL is not a book to just slog through reading assignments that your instructor names, however, and then sell it downstream to recoup some of your \$\$ investment. Instead, it’s a book to read, and then re-read; to mark up with notes and highlights, to circle concepts that gave you trouble and stat practice problems that open doors for you. Treat the book as an ongoing catalog reflecting the shifts and fads in the field, and keep it within arm’s length of Data Science job, internship, or project you find yourself on!

I’ll be supplementing ISL as well, by interleaving content from other works (including my own) at the proper time. But at a 50,000-foot view we’ll hold pretty tightly to the story that ISL develops. I’ll point out where I disagree with their their methodology (this won’t happen often).

Another mini-field that we’re going to have to fiddle with a bit is called “Linear Algebra,” a topic you may already know, or think you know. To build skills with vectors and matrices (the two main topics in Linear Algebra) we’ll be learning on select sections of my “Quick Steep Climb Up Linear Algebra” text. Like ESL, it’s actually completely free in its entirety in its digital form. So if you’re okay with digital textbooks, the total you’ll have to spend on course materials this semester is \$0 (which I think is how it should be, and why I give away all my books.)

Late policy

No late work will be accepted this semester. Get your stuff in on time!

Grading

Grading this semester will be based on “experience points” (XP). As you complete activities, you will earn XP towards your final total. XP can never be lost, only gained, but you have to earn what you get (*i.e.*, you don’t “start off with a 100%” and lose points based on mistakes you make).

There will be opportunities to earn XP throughout the course. Some of these will be spontaneous as the mood strikes me. Some you can earn by completing in-class activities. Some may be in response to impressive things I see you do as the semester progresses. The following opportunities, however, are *guaranteed* to be available to you:

Guaranteed XP opportunities:

Activity	Possible XP
Six open-book, open-note, timed Canvas quizzes	30 each
Seven homework assignments	40 each
Two responsive readings	50 each
Final exam (comprehensive)	200
Various and sundry others	varies

Grading levels

Here are the levels you may achieve, together with the grade awarded (if any) and the points necessary to reach!

Level	Total XP	Semester grade
Gandalf	720	A+
Harry Potter	660	A
Elsa	600	A–
Lady Jessica	560	B+
Elphaba	520	B
Professor Albus Dumbledore	480	B–
Violet Sorrengail	440	C+
Merlin	400	C
Shallan Davar	360	C–
Moiraine Damodred	330	D+
Princess Irulan	300	D
Hermione Granger	280	
Raistlin Majere	260	
Saruman	240	
Egwene al’Vere	220	
Countess Margot Fenring	200	
Dr. Strange	180	
Melisandre	160	
Glinda	140	
11	111	
Kvothe	90	
Willy Wonka	60	
Ron Weasley	40	
Alex Russo	10	
Fruit Pie the Magician	0	

Submitting homeworks

Rules for submitting homeworks (whether in Python or in English) will be given when the homework is assigned.

For programs, you'll be emailing me your program code as an attachment, and using a specific subject line to distinguish it from my hordes of other email. Meeting the deadline is a matter of sending your email before time expires.

For responsive readings, **you must submit a hardcopy of your paper – nothing electronic will serve as an adequate substitute.** I'll probably collect these in class.

Also, most of my homeworks are due at “midnight.” Here's what “midnight” means: if a homework is due “at midnight on Thursday,” then it is due after all of Thursday has elapsed, and the clock strikes twelve. (In other words, this is good news: you have all Thursday to work on it.)

Basis for determining mid-semester reports

For midterm progress reports, I look mostly at (a) whether you've been turning assignments in (and preferably on time), and (b) quiz scores. If either or both of these categories are lacking, it's a sign of danger, and I will give you a “U” for your mid-semester grade. Please don't hesitate at all to come talk to me about this so we can figure out how you can do better in the course.

Guidelines for class participation

I believe that students learn best when they participate wholeheartedly in all aspects of the learning process. Hence while your grade will not be partially determined by any “class participation score” *per se*, it is very much to your advantage, and very much recommended, that you join in during class discussions, ask questions, and make comments.

Disabilities

If you have a documented disability, please present me your letter from the Office of Disability Resources and I'll be happy to accommodate you.

Course artificial intelligence (AI) policy

About AI, I have three simple rules. Learn them carefully.

1. It ***is*** permitted to use an AI assistant for writing your programs this semester. In fact, I encourage this practice if it is done responsibly, and you will quite often see me demonstrate in class how I personally do it responsibly. I frankly think that using an AI assistant is “the new way to code,” and anyone who does not learn how to use one effectively is going to find themselves obsolete in nothing flat.

All I ask is that you **submit the logs of all your AI chats along with your assignment**, so that I can see how you're using it. This isn't to scare you or police you; it's for me to get a read on which ways students are using it well and which aspects I need to give more class time to demonstrating.

2. It **is *also*** permitted to use an AI assistant in composing your responsive readings and other writing assignments. Again, I do ask that you **submit your AI chat logs along with your assignment**, but I condone and even encourage the practice. (I'll have more to say in person about the right way to get AI assistants to improve your writing.)
3. It ***is NOT*** permitted to use an AI assistant while taking any quiz or exam, period.

Guidelines for class participation

I believe that students learn best when they participate wholeheartedly in all aspects of the learning process. Hence while your grade will not be partially determined by any "class participation score" *per se*, it is very much to your advantage, and very much recommended, that you come to lecture every single class period, and participate fully in it.

Title IX Statement

UMW faculty are committed to supporting students and upholding the University's *Policy on Sexual and Gender Based Harassment and Other Forms of Interpersonal Violence*. Under Title IX and this Policy, discrimination based upon sex or gender is prohibited. If you experience an incident of sex or gender based discrimination, we encourage you to report it. **While you may talk to me, understand that as a "Responsible Employee" of the University, I must report to UMW's Title IX Coordinator what you share.** If you wish to speak to someone confidentially, please contact the confidential resources below. They can connect you with support services and help you explore your options. You may also seek assistance from UMW's Title IX Coordinator; their contact information can be found below. Please visit <http://diversity.umw.edu/title-ix/> to view *UMW's Policy on Sexual and Gender Based Harassment and Other Forms of Interpersonal Violence* and to find further information on support and resources.

Resources

Ruth Davison, Ph.D.
Title IX Coordinator
Lee Hall, Room 401
540-654-5656
rdavison@umw.edu

Confidential Resources

On-Campus

Talley Center for Counseling Services
Lee Hall 106, 540-654-1053

Student Health Center
Lee Hall 112, 540-654-1040

Off-Campus

Empowerhouse (24-hr hotline)
540-373-9373

RCASA (24-hr hotline)
540-371-1666

Recording Policy

Classroom activities in this course may be recorded by students enrolled in the course for the personal, educational use of that student or for all students presently enrolled in the class only, and may not be further copied, distributed, published or otherwise used for any other purpose without the express written consent of the course instructor. All students are advised that classroom activities may be taped by students for this purpose. Distribution or sale of class recordings is prohibited without the written permission of the instructor and other students who are recorded. Distribution without permission is a violation of copyright law. This policy is consistent with UMW's *Policy on Recording Class and Distribution of Course Materials*.

Accessibility statement

The Office of Disability Resources has been designated by the university as the primary office to guide, counsel, and assist students with disabilities. If you receive services through the Office of Disability Resources and require accommodations for this class, please provide me a copy of your accommodation letter via email or during a meeting. I encourage you to follow-up with me about your accommodations and needs within this class. I will hold any information you share with me in the strictest confidence unless you give me permission to do otherwise.

If you have not made contact with the Office of Disability Resources and have reasonable accommodation needs, their office is located in Seacobeck 005, phone number is (540) 654-1266 and email is odr@umw.edu. The office will require appropriate documentation of disability.

Basic needs security

Learning effectively and engaging wholly in class is dependent upon our basic security and having our fundamental needs met: having a safe place to sleep at night, regular access to nutritious food, and some assurance of safety. If you have difficulty affording groceries or accessing sufficient food to eat every day, or if you lack a safe and stable place to live, please contact Chris Porter, Assistant Dean of Students, at cjporter@umw.edu. Additionally, the Gwen Hale Resource Center is a free resource on campus, providing food, toiletries and clothing to any member of our community. It is open Monday, Tuesday and Friday from 1pm-6pm, on the 5th floor (floor A for Attic) of Lee Hall, or resource@umw.edu. Finally, you are always welcome to talk with me about needs, if you are comfortable doing so. This will enable me to provide any resources I may possess.

How to reach me

Come to office hours, see me after class, or e-mail me (stephen@umw.edu).

Calendar

The official calendar for the course, complete with assignment due dates, tests, *etc.*, will be maintained on the course website at <http://stephendavies.org/data419>. **In any way that the website conflicts with the tentative calendar below, the website is to be considered correct, and the tentative calendar below out of date.**

Week	Book Chapter	Topics	Due
1	1	...Stephen is overseas; work on HWs 0 and 1...	
2	2, 5	Machine Learning (ML) settings and concepts	HW 0
3	—	Mathematical & programming machinery for ML	HW 1
4	3	Linear regression models	
5	4	Linear classification models	HW 2
6	4	Alternative classification models	
7	6	Linear model selection and regularization	HW 3
8	6	Dimensionality reduction	RR 1
9	6, 10.5.1	Tokenization and NLP embeddings	
10	8	Decision trees	HW 4
11	—	Ethical considerations in ML	
12	8	Decision Trees, cont.	HW 5
13	7, 9	Nonlinear methods and SVMs	
14	12	Unsupervised learning: clustering	RR 2
15	—	Looking towards ML II	HW 6

“HW”=Homework, “RR”=Responsive Reading