DATA 101 — Introduction to Data Science

Professor: Stephen Davies
Fall semester 2025

Class: MWF 4pm, Farmer 025

Final exam: Monday, Dec. 8th, 3:30pm

Office Hours (James Farmer Hall 044):

Mon	1–4pm
Tue	12–2pm

`http://stephendavies.org/data101`

Welcome to the gateway course for the Data Science program! This course serves two purposes: (1) to give you an overview of what Machine Learning, Artificial Intelligence, and Data Science entail and a taste of what they can achieve, and (2) to help you build some skills with the tools and environments that Data Scientists actually use in practice. You'll see what kinds of problems can be solved, and get your hands dirty with some real data and state-of-the-art tools. By the end of this course, you should know whether Data Science is a field you want to further explore (and whether the Data Science major or minor might be of interest to you).

Student Learning Outcomes

After completing this course, students will be able to:

- Describe what the field of "Data Science" has to offer, and what the professionals who engage in it do every day.
- Write computer programs that use most basic programming primitives: variables, calculation, branching, functions, and loops. This is a subset of what students learn in CPSC 110, but with a focus on data analysis more than algorithmic construction.
- Employ basic data exploration techniques: textually and graphically inspect a data set to get first impressions, calculate and properly interpret summary statistics, identify ranges and outliers, and form possible hypotheses for more formal modeling.
- Recognize the different ways data can be structured and arranged in memory, and to make appropriate selections for organizing new data.
- Encounter, and begin to tame, data "in the wild." This involves dealing with different storage formats, and writing code to perform basic data transformations such as type conversion and scaling.

- Describe the basic framework for supervised machine learning, including the importance of separate training and test sets, and how at least one algorithm works under the hood to perform classification.

Rules of the game

1. There are NO stupid questions. I will never belittle you or make you feel dumb about anything. Your job is not to already know everything before you start, but to roll up your sleeves and work hard to try and learn. Even if your question is, "Stephen, I totally didn't even get that, can you start over and explain it again?", please ask.

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 "the answer" every time. Go ahead and give it a shot, or else just say "**pass**" which is always accepted.)

3. You are free to work with anyone you wish, to consult other sources, to use AI, etc., on the (ungraded) homeworks. For the quizzes and exams, however, you must work entirely on your own, without any help from any unauthorized source (see "Grading" and "The Honor Code and this course" sections, below).

4. Exams and quizzes will be based on both the book readings and the material I cover in class. Make sure you know both!

5. Please, **no laptops during lecture**. I've had students claim that they take notes on their laptop during class, but even if it's true, those things are too big a distraction to you and your fellow students to make it worth it. Just stay tuned in, because I move fast.

Books and materials

- Davies, S. (2025). *The Crystal Ball Instruction Manual, Volume One: Introduction to Data Science,* Version 1.2. Blurb, Incorporated. ISBN 9798319910158.

The *Crystal Ball* series is an Open Educational Resource (OER) and is available electronically in its entirety at `http://stephendavies.org`. I still encourage you to obtain a hardcopy, though, because past experience has indicated that students are much more likely to actually do their required reading if they have a physical copy. You can get one from the UMW Bookstore, or from Blurb.com (cheaper but takes longer to arrive) or Amazon (more expensive but arrives quicker). (Btw, I get ***zero*** royalties for this book, whether you buy it at the bookstore or online.)

Late policy

No late work will be accepted this semester. Get your stuff in on time, there's no excuse not to!

Grading

- 0% — Skill-building **homework** activities, which you are encouraged (in the strongest possible terms) to work on and complete. Warning: they will involve problem-solving, not simply regurgitation of commands or code examples.
- 50% — Weekly Friday **quizzes**, which cover recent material from class, from the homeworks, and from the book. Quizzes are **closed-book**, **closed-notes**, and **timed** (about 20 minutes each). They will begin *at the very start of class*, so don't be late! There are no makeups for these, but I'll drop your bottom two scores.
- 20% — **Midterm exam**: **open-book, open-notes**, take-home, **timed (2 hours)**. Distributed in class Monday, Oct. 6th; due back Friday, Oct. 10th, 4pm.
- 30% — **Final exam**: **open-book, open-notes**, **timed ($2\frac{1}{2}$ hours)**, Monday, Dec. 8th, 3:30pm.

The Honor Code and this course

I strongly believe in UMW's honor code and scrupulously adhere to it. Here are the rules for this course:

The homeworks are ungraded, and I want you to decide how you will learn best. If you think working in a small group will help you learn the material better, you're free to do that. If you think Googling for solutions or partial solutions to problems will help, you're free to do that. If you think using AI (e.g., Gemini, Colab's AI, ChatGPT) will help, you're free to do that. Just remember: **the point is not to complete the exercise and check it off, but to learn what it was meant to teach you.** You're going to hate life if you perfectly complete every homework assignment, but in a way that you didn't learn much as a result. You will be cruelly exposed on the quizzes and exams.

The quizzes and exams are to be **solely your own work**. The quizzes are closed-book and notes, and the final is open-book and notes, but they are both completely closed to other humans (even those not currently at UMW) and closed to AI.

By the way, if there is ever something you feel tempted to cheat about, in this or any other class, please come and talk to me about it. I will not penalize you or think less of you in any way for admitting that you feel tempted — on the contrary, I will think highly of you for having the courage to come forward. If you feel like you need to cheat, the solution is as follows: come talk to me about whatever part of the class you're struggling with and let me teach it to you better until you feel comfortable with the material. Then, there won't be any need to cheat.

Basis for determining mid-semester reports

I'll mostly base your midterm progress reports on your quiz scores, although if I think I've detected that they might be either a false negative or a false positive, I'll make adjustments. A "U" for your mid-semester grade does not mean things are hopeless – instead it's a wake-up call to perhaps change your approach to the class. Please don't hesitate *at all* to come talk to me about this so we can figure out how you can do better!

Calendar

The calendar for the course, complete with homework "due dates," tests, *etc.*, will be maintained on the course website at `http://stephendavies.org/data101`.

Use of Artificial Intelligence (AI) technologies

See section on "The Honor Code and this course", above (p. 3).

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`).

How to reach you

I will post announcements to the course website often, so be sure to check it regularly! Also, I may occasionally be communicating with you outside of class time via e-mail, so make sure to check your UMW e-mail every day!

Road map

This semester will feature the following topical units, but *not strictly in this order*. Instead, we'll be bouncing around between topic areas and integrating them into a complex whole.

The Data Science Field

- Data, info, knowledge, and wisdom
- Computation, statistics, and application
- Exploration vs. prediction vs. inference
- Data-generating processes
- The Data Processing Pipeline
- Triumphs and challenges
- Careers in Data Science

Representing data

- Bits & bytes
- Atomic data types
- Aggregate data types
- Heterogeneity
- Representing data in Python/Pandas
- Representing data in text files
- What data looks like in a program's memory

Interpreting data

- Univariate vs. multivariate
- Scales of measure
- Data vs. metadata
- Missing data
- Quantiles and z-scores
- Outliers
- Normalization and redundancies
- Controlled experiments vs. observational studies
- Confounding factors and randomization
- Types of causality

Processing data

- Calculation
- Branching
- Loops
- Functions

Transforming data

- Filtering
- Recoding and rescaling
- Composite measures
- Grouping and summarization
- Reshaping
- Merging
- Text mining transformations
- Image transformations

Exploring and presenting data

- Exploratory Data Analysis
- Statistical summaries
- Visualizations
  - histograms
  - scatter plots
  - box plots
- Constructing a narrative

Machine learning

- Training vs. test data
- Classification, regression, and clustering
- Model complexity and "overfitting"
- Measuring performance
- Decision Trees: a classification algorithm