DATA 219 — Foundations for Data Science

Professor: Stephen Davies
Spring semester 2019

Class: MWF 2pm, Trinkle B6

Final exam: Tuesday, May 1st, 3:30pm

Office Hours (in Trinkle B22):

Mon	1–2pm
Tue	2–3pm
Wed	3–5pm
Thu	10–11am

`http://cs.umw.edu/~stephen/data219`

Welcome to Part II of your Data Ninja training! This second half of the DATA 101–219 sequence is designed to sharpen your tools, increase your array of weapons, and introduce you to a new level of more challenging opponents. We'll continue to build on the concepts presented in DATA 101, teaching you how to work with and analyze a variety of different types of data in a principled way. We'll also get our hands dirty early and often with real-world data sets.

By the end of this course, you should be able to handle yourself "in the data wild." No longer will you need things to be handed to you on a silver platter: instead, you'll be able to bust outside the boundaries, confidently wrangle data no matter where it is or what it looks like, and discover on your own what you need. You will thus have a set of skills that are highly sought after by industrial and government organizations across the globe, and that enable you to take your first steps as a professional.

Course Objectives

- To gain further experience and expertise with the various stages of the Data Processing Pipeline. You will learn more about what questions to ask, how to frame them, how to answer them, and how to justify your answers.
- To continue to build your Python programming skills for data analysis. You will write more complicated and longer programs to deal with data in new and interesting ways.
- To encounter, and tame, data "in the wild." You will understand and be able to recognize the different ways data is stored, organized, and transmitted. You'll get practice parsing common formats for data representation: text and binary files, .csv, hierarchical (XML/JSON), screen scraping, and connecting to an API.
- To complete your preparation for upper-level DATA electives. Those courses will deal with more advanced techniques and with domain-specific knowledge from various areas.

They will presuppose a level of fluency with the toolsets and the lower-level operations that this course will give you.

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. 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.

4. Don't cheat. Cheating is heinous, rude, and bad karma. It really makes me mad, and it will also eat away your character like hydrochloric acid if you're not careful. If you ever feel tempted to cheat, in this class or any other, come and talk to me about it. It's not wrong to feel tempted, and we can find other ways out of whatever dilemma you're facing without compromising your moral character.

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.

Late policy

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

Basis for determining mid-semester reports

For midterm progress reports, I will look mostly at your lab scores and 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.

Books

- Davies, S. *Foundations for Data Science: Course Notes.* 2019.
- McKinney, W. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* 2nd edition, O'Reilly Media, 2017.

I have prepared detailed lecture notes that will serve as a textbook of sorts for the semester. You can pick up your inexpensive copy at the UMW Bookstore. They're organized into "lessons" which we will go through pretty much in order. The goal here is for you to not have to frantically scribble down everything I say in lecture (during which I speak at approximately 2,000 words per minute), but can instead pay close attention to thinking through the concepts themselves.

I will make the notes available online as well but you **SHOULD** buy the inexpensive coursepack from the Bookstore as well, bring it to class, and mark it up with your own notes. (I get no royalties for this, by the way.)

Wes McKinney is the horse of the horse's mouth: the creator of the Pandas Python library himself. This excellent book can serve as a valuable reference for most of the basic technology stack we're using this semester (scikit-learn excepted.)

Grading

- 25% — Eight Canvas quizzes, covering recent material and outside-of-class readings. Quizzes are open-notes. There are no makeups for these, but I'll drop your bottom two scores.
- 50% — Skill-building homework assignments in which you'll get hands-on practice with the data techniques we'll learn. On some of these, you must work alone. On others, you may work with a partner. **Each assignment will clearly state whether partner work is permitted.** It is your responsibility to read this and abide by it.
- 25% — Final exam: open-book, open-notes, 3:30–6pm May 1st. (But see below.)

DataFest!

Stay tuned for details! If you attend DataFest with the Data Mavens, I will give you a "grade" for your group's work and presentation. If you're happy with that grade, **that can be your grade for the final exam, and you won't have to take the final.** Otherwise, you can take the final at the end of the semester and I will count the higher of your DataFest "grade" or your exam grade.

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 quizzes and final exam are to be solely your own work. They are all open-notes, but they are all completely closed to other humans.

Some of the homeworks will state that you must work **alone** on them. For others, you will be assigned a **partner**. On the partner ones, you are (of course) free to work with your partner, **but with no one else**.

By the way, if there is ever something you feel tempted to cheat about, 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.

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.

How to reach me

Come to office hours, see me after class, or e-mail me (`stephen@umw.edu`). I'm normally very quick to respond!

How to reach you

I will post announcements to the course website often, so be sure to subscribe to its RSS feed and check it in your feed reader at least once a day! Also, I will occasionally be communicating with you outside of class time via e-mail, so make sure to check your UMW e-mail every day!

Calendar

The calendar for the course, complete with assignment due dates, tests, etc., will be maintained on the course website at `http://cs.umw.edu/~stephen/data219`.

Road map

This topic list is arranged **topically**, not chronologically. We will not proceed in exactly this order!

1. Navigating the analysis environment
   - Spyder
   - Interfacing with files & directories
   - Using the Programming API documentation effectively
2. The NumPy/SciPy library
3. Encountering data "in the wild"
   - The Data Exploration Checklist
   - Data Formats
     - CSV
     - JSON
     - XML/HTML and Beautiful Soup
     - SQLite
     - Plain text
   - Data Wrangling
     - Data type conversion
     - Parsing dates/times
     - Recoding
     - Rescaling and z-scores
     - Screen scraping
     - Data fusion
     - "Tidy" data
   - Interfacing to a Data API
4. Generating synthetic data sets
5. Exploratory Data Analysis, cont.
   - Understanding correlation
   - Scatterplot Matrices
   - Exponential vs. power-law distributions
   - Logarithmic plots and what they reveal
   - Kernel Density Estimates
   - Locally Weighted Regression (LOWESS)
6. Special kinds of data
   - Graph/network data
   - Natural language (and basic text mining)
   - Time series
7. Machine Learning, cont.
   - The `scikit-learn` library
   - Feature selection
   - The "bias-variance tradeoff"
   - The "curse of dimensionality"
   - Classification algorithms
     - Naïve Bayes
     - k-NN
   - Association analysis
   - Clustering algorithms
     - K-means
     - Hierarchical clustering