{CS, STAT} 387: Data Science II (DSII)
David Rushing Dewhurst
TR 18:00 - 19:15 EST

From the catalogue:

> *Advanced data analysis, collection, and filtering. Statistical modeling, Monte Carlo statistical methods, and in particular Bayesian data analysis, including necessary probabilistic background material. A practical focus on real datasets and developing good habits for rigorous and reproducible computational science.*

Unpacking this – we will cover some (hopefully large) subset of the following topics:

- *Advanced data analysis, collection, and filtering [...] real datasets*:
  - Representations and transformations of numerical and non-numerical data, e.g. text, network, music, chemical and biological, image and video, transaction,...
  - Relational and nonrelational databases in practice, file systems, data persistence and storage
  - Collecting data "in the wild": making usable datasets from multiple messy real-world sources
- *Statistical modeling, Monte Carlo statistical methods*:
  - Discriminative models, maximum likelihood (MLE) and maximum *a posteriori* (MAP) estimation, necessary optimization review, derivation of MLE and MAP models from probabilistic first principles.
  - Overview of nonparametric statistics, nonparametric tests, and signal processing methods. Black-box model and model-free uncertainty estimation and hypothesis testing.
- *and in particular Bayesian data analysis*:
  - Generative models and algorithmic data generating processes. Fundamentals of probabilistic programming, trace-based probabilistic programming languages, model grammars
  - Sampling-based inference, including {rejection, importance} sampling, Markov chain methods including Metropolis-Hastings and Hamiltonian methods, potentially research topics including reversible jump and involutive MCMC
  - Variational inference: analytical results, black-box, variational posterior design, discrete latent variables, variable model dimensions, scaling to large data
- *probabilistic background material*:
  - Probability foundations including options interpretation and estimating probability of outcomes from market data, Bayes's theorem, updating, conjugacy, high-dimensional distributions

  - Basic information theory, derivation of probability distributions from maximum entropy and transformation groups

- *rigorous and reproducible computational science*:

  - Version control including basics of Git and remote hosting, containerization including Docker and Kubernetes
  - Unit and integration testing, continuous integration, continuous delivery and continuous deployment, large-scale software design philosophies in practice
  - Literate programming and when (not) to use it, data security, and information sensitivity

If there is supermajority interest, we may cover additional topics that are not listed above.

Class policies and things to note

- Grading: 40% homework, 10% reading responses, 50% midterm + final project. More details about final projects are below.

- I will not mandate any set of prerequisite classes. However, you should be comfortable with multivariate calculus, linear algebra, and computer programming in one or more modern languages. {CS, STAT} 287 is listed as a prerequisite; if you have not taken this, you should be comfortable with all the material taught in that course.

- When you turn in code, it should work. There should be a file that tells me how to run your code e.g. a `README` file, including all steps necessary to get the desired output. Your code should have “sane” dependencies (e.g., `gcc` is okay, but some strange Fortran compiler from 1968 is not). You should comment your code judiciously and ensure that a non-language expert can at least understand what your code is doing.

- I will not enforce a particular programming language for the course. That being said, most of my examples will be in either Python or Julia, with possible detours into C++ and R. You are not required to know any of these languages to take the course, but I will not spend time teaching them during the course. (This is not to be mean, but to help you: in industry, it is common to be thrown into a large codebase written in a language you have never seen before!)

- Your midterm and final project combined are to write an academic-quality manuscript. Your goal is to post this manuscript to arXiv.org by the end of the course. I will work with students whose manuscripts are of exceptional quality to prepare the manuscripts for submission to conferences or journals. I will provide more information during the first week of class (and throughout the semester).

- I have no idea how I will collect and grade homework. When I figure that out, I’ll let you know.

- Because of the global pandemic, we will be meeting remotely. This is unfortunate. We will have to use Zoom, or Teams, or something – again, when I figure that out, I’ll let you know.

- The best way to contact me is through email: `drd@davidrushingdewhurst.com`