MATH/LING 289: Introduction to Data Science, Fall 2018
Tuesday, Thursday 13:30-14:45 PURH G13

Instructor:	**Taylor Arnold**
E-mail:	tarnold2@richmond.edu
Office:	Jepson Hall, Rm 218
Office hours:	to be determined
Course Website:	`https://statsmaths.github.io/stat289-f18`

Overview:

Data science is an interdisciplinary field concerned with extracting knowledge from data and communicating those results to some public audience. Data science needs to be learned *by doing* data science. There are no short-cuts and the process cannot be learned by simply working our way methodically through a textbook of disconnected topics.

Therefore, this course will be taught using a problem-based learning model. As a class, we will be addressing an open ended research question and learning various skills that will assist our inquiry. Both individually and in small groups, students will take responsibility for particular subtasks that drive our research forward. At the end of the semester students will have acquired a toolkit of methods, and the knowledge of how to use them in practice, to address important social, cultural, and scientific questions with data-driven techniques.

We will make heavy use of computing throughout the semester, but *no prior programming experience is required*. Also note that this course has a MATH designation because statistics is currently housed in the mathematics department. The topics of this course, however, do not fall within the traditional disciplinary boundaries of mathematics.

Research Question:

Our object of study for this semester (Fall 2018) concerns the edit history of pages on Wikipedia. Our research for the semester will be guided by drawing on questions such as:

- How is knowledge being represented through the works of an anonymous, decentralized collective of users connected across the internet?
- What role does memory play in the representation of current and semi-current events?
- What kinds of knowledge are privileged, or taken for granted, by citation patterns on the interent?
- How do cultural and linguistic factors play into the structure of Wikipedia pages (by looking at pages across different languages)?
- What role do images and other media play in the structure and development of encyclopedic pages? has this changed over time?

These questions can be studied from a number of disciplinary perspectives. For example, one might draw on methods from from one or more of: psychology, sociology, linguistics, political science, American studies, media studies, and critical theory. In this course we will see how the methods of data science provide a new set of tools that are able to engage with, rather than against, these disciplinary techniques while opening the possibility of producing knowledge through the study of large unstructured data sets.

Grades:

Your final grade will consist of three elements, weighted as follows:

- Class Participation, 20%
- Project Prospectus, 20%
- Final Project and Presentation, 60%

Course expectations and community standards will be discussed, developed and distributed in the first week of the course. This will include policies for class participation, attendance, and late work.

Texts:

Readings for the course, which will help us address and position our research, will be pulled from open access journal articles and the following required text:

> Jemielniak, Dariusz, 2014. *Common Knowledge?: An Ethnography of Wikipedia.* Stanford University Press.

As a reference for technical topics, I will make frequent reference to:

> Wickham, Hadley and Grolemund, Garrett, 2016. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media.

This second text as it is available online (`http://r4ds.had.co.nz/`) for free and in its entirety.

Computing:

We will use the **R** programming environment throughout the semester. It is freely available for all major operating systems. No prior programming experience is assumed or required for this course. We will install the required software on your computer during the first week of the course (alternatively, R is installed in several computer labs across campus) and cover all of the required coding skills throughout the semester.

Data Science Methods:

Given the project-oriented aspect of this course, the exact schedule of topics will depend on and adapt to the natural course of our investigation. Methods that may be covered during the semester include:

- data structures for storing data
- visualization techniques for exploratory analysis
- web scrapping
- document summarization
- sentence parsing
- named entity recognition (NER)
- automated geocoding
- dimensionality reduction
- topic models
- network analysis
- image processing

Strong consideration will be given to the specific of interests and motivations of students enrolled in the course.