Nicholas Eubank
Assistant Professor

POLISCI NUMBER
TERM & YEAR

Computer Science for Data Science (CS for DS)

1 Course Description

Computer science has much to offer the budding Data Scientist, but all too often computer science programs and texts just aren't designed with data science in mind. Computer science programs are full of useful concepts and topics, but they are often mixed in with lots of material that isn't relevant for non-academics or people not going into software development.

This course is designed to provide an introduction to key computer science concepts of relevance to data scientists in an applied, efficient manner, including:

- Defensive Programming: How to write code that minimizes the likelihood you'll make mistakes and maximizes the likelihood that when you do make mistakes, you'll be able to catch them.
- How computers think about numbers and text: learn why 42.0 doesn't always equal 42.0.
- Parallelization: What is parallel computing, why is it becoming more and more important, and what are it's limitations
- Big Data and The Memory Hierarchy: Why working with big data requires categorically different strategies than smaller datasets.
- Speed: Why are some programming languages fast and others slow, how do I write code that runs quickly, and how to I evaluate the speed of my code?

In addition to these core computer science concepts, students will also be introduced to a set of **ancillary skills and tools** which are often overlooked in courses that emphasis either just the core data science languages (Stata, Matlab, Python, or Julia), or just the statistics of methods like machine learning. We will explore the *purpose* of these tools and why they may be of use to you, as well as doing hands-on exercises to develop comfort with these tools. In particular, we will cover are provided with information on:

- The Terminal
- Git and Github
- How to Get Help Online

2 Prerequisites and Course Fulfillment

This course also requires students be comfortable doing basic data manipulations (e.g. loading CSVs, tabulating data, merging data) in a language like Stata, R, Python, or Julia.

3 Course Schedule

Outlines of Specifics Material Being Covered
for Each Topic Can Be Found At www.csfords.com
All topics will also be paired with hands-on exercises and tutorials!

PART I. The Tools of Data Science No One Taught You

- Week 1: The Terminal
- Week 2: Git and Github
- Week 3: Class 1: Getting help online
- Week 3: Class 2: Jupyter Labs

PART II. Programming, a CS Perspective

- Week 4: Defensive Programming, Decomposition
- Week 5: Data Types
- Week 6: Data Structures

PART III. Computer Architecture For Data Science

- Week 7: Parallelization, Processors, and Amdhel's Law
- Week 8: Big Data and the Memory Hierarchy
- Week 9: Writing Performant Code