Data Science for the Liberal Arts

Kevin Lanning

2018-01-06

Contents

Part I

Part I Introduction

Chapter 1

Preface

This work-in-progress includes the notes for Introduction to Data Science at the Wilkes Honors College of Florida Atlantic University. It's written using the R package Bookdown (Xie, 2017). The source code for the book may be found at, and the class syllabus and schedule may be found at.

Chapter 2

Introduction

Hochster (in Hicks & Irizarry, 2017) describes two broad types of data scientists: Type A (Analysis) data scientists, whose skills are like those of an applied **statistician**, and Type B (Building) data scientists, whose skills lie in problem solving or coding, using the skills of the **computer scientist**. Our course is like those at the universities of North Carolina, British Columbia, Duke, Maryland, Wisconsin, Stanford, BYU, Harvard, Pennsylvania, and UC Berkeley (and will likely draw from all of these) in that it is closer to a Type A than a Type B treatment, one which is closer to Statistics than to Computer Science. But there's more.

2.1 Type C data science

Hochster's view of data science arguably omits a critical component of the field. Data science is driven not just by statistics and computer science, but also by "domain expertise:"

The iconic Venn diagram model of data science suggests what we can call a "Type C data science." It begins with "domain expertise" in your **concentration** in the arts, humanities, social and/or natural sciences, it both informs and can be informed by new methods and tools of data analysis, and it includes such things as **communication** (including writing and the design and display of quantitative data), **collaboration** (making use of the tools of team science), and **citizenship** (serving the public good, overcoming the digital divide, furthering social justice, increasing public health, diminishing human suffering, and making the world a more beautiful place). It's shaped, too, by an awareness of the **creepiness** of living increasingly in a measured, observed world.

The WHC Intro to Data Science course will be a Type C course, or more accurately, a CAB course - equal parts statistics and domain knowledge, with just enough computing to be proficient to use (but not yet build) the software tools at our disposal. We aren't unique here - there are courses (which we may again draw from) with similar goals at schools including Chicago, Georgia Tech, UC Santa Barbara, Princeton, UC Berkeley, at Berlin's Hertie School of Governance, and in Columbia's School of Journalism.

2.2 What will be in the class?

R

In my rough survey of introductory data science courses, I saw a pretty even split between those which begin with Python and those which begin with the statistical programming language R. This difference corresponds, loosely, to the split noted above: Computer science based approaches to data science are frequently grounded

in Python, while stats based approaches are generally grounded in R. Our course, like those for most of the syllabi and courses linked above, will be based in R.

Reproducible science

The course will provide an introduction to some of the methods and tools of reproducible science. We will consider the replication crisis in the natural and social sciences, and then consider three distinct approaches which serve as partial solutions to the crisis. The first of these is training in a notebook-based approach to writing analyses, reports and projects (using R markdown). The second is using public repositories (such as the Open Science Framework and GitHub) to provide snapshots of projects over time. Finally, the third is to consider the place of significance testing in the age of Big Data, and to provide training in the use of descriptive, exploratory techniques of data analysis.

Good visualizations

Part of Type C data science is communication, and this includes not just writing up results, but also designing data displays that incisively convey the key ideas or features in a flood of data. We'll examine and develop data visualizations such as plots, networks and text clouds. More advanced topics may include maps, interactive displays, and animations.

~~All~~ *Some of* the data

It's been argued that in the last dozen years, humans have produced more than 60 times as much information as existed in the entire previous history of humankind. (It sounds like hyperbole, but even if it's off by an order of magnitude it's still amazing). There are plenty of data sources for us to examine, and we'll consider existing datasets from disciplines ranging from literature to economics to public health, with sizes ranging from a few dozen to millions of data points. We will also clean and create new datasets.

~~All~~ *Some of* the meaning

Data matter, can save, enhance, or destroy human lives. (This is a crummy sentence: My pedantic insistence on treating the term "data" as plural rather than singular likely distracted you from the substance of my message. I'll leave it here as a reminder that we can all become better writers). Back to my point: In this class, we'll explore approaches to analyzing the meaning of data in areas including the analyses of simple texts and data journalism.

~~All~~ *Some of* the skills

The skills required to extract meaning from data include an understanding of classical statistical principles (e.g., probability theory and sampling theory), core statistical techniques (regression), and the extension of these core principles and basic techniques to problems in natural language processing, the analysis of social networks, and machine learning and classification.

~~All~~ *Some of* the tools

In addition to R, we'll use a range of other tools: We'll communicate on the Slack platform. We'll write using markdown editors such as Typora. We'll certainly use spreadsheets such as Excel or Google Sheets. We *may* use additional tools for visualizing data such as Gephi and Tableau.

Hands-on computing

We had initially anticipated that the lectures would include discussion, and that the computing part of the class would occur just in the lab. But, in the course of examining syllabi at other schools, it became apparent to me that **there will be computing throughout the course**, not just in the lab.

What will be in the lab?

The labs will have two features. First, they will allow for a project-based approach, focused on the collaborative analysis of problems of your own choosing. Second, these projects will include a deeper dive into some of the topics and problems above. Here are a few examples of how the treatment in the lecture and the lab are likely to differ:

Topic	Lecture	Lab: Deeper dive
Introduction	Stephens-Davidowitz book; Google trends	Examining one or more scholarly papers in data journalism, computational science, or computational social science
Getting data	Extant data	Using APIs to scrape social media
Sharing science	Setting up account on OSF	Using GitHub, Becoming a Repo ~~Man~~ *person*
Exploratory data analysis / Data visualization	Static data displays	Interactive plots (r Shiny), animated displays?, Tableau?
Sampling theory	Test vs training datasets	k-fold cross-validation
Regression and classification	Linear and logistic regression	Machine learning: Robust techniques / regularized regression Supervised prediction, possibly semi-supervised and unsupervised regression
Graph theory and network analysis	Introduction to centrality, community structure, contagion	Network robustness, different approaches to centrality & community structure
Analyzing texts	Word clouds	Text mining, natural language analysis
Generating products	Team project in class	Project in class and poster

Bibliography

Xie, Y. (2017). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.5.